SANATA DHARMA UNIVERSITY PRESS

PROSIDING

SEMINAR NASIONAL INOVASI TEKNOLOGI INFORMASI & KOMUNIKASI

"Optimalisasi Teknologi Kecerdasan Artifisial untuk Mendukung Transformasi Digital dan Masa Depan Otomasi"

25 November 2023
Lembaga Penelitian & Pengabdian kepada Masyarakat
Universitas Sanata Dharma Yogyakarta

LPPM USD

IndoCEISS

SENOVTIK

PROSIDING SEMINAR NASIONAL INOVASI TEKNOLOGI INFORMASI & KOMUNIKASI

"Optimalisasi Teknologi Kecerdasan Artifisial untuk Mendukung Transformasi Digital dan Masa Depan Otomasi"

25 November 2023

Lembaga Penelitian dan Pengabdian kepada Masyarakat
Universitas Sanata Dharma Yogyakarta

IndoCEISS Provinsi D.I. Yogyakarta
dan IndoCEISS Kalimantan Tengah

Sanata Dharma University Press

PROSIDING SEMINAR NASIONAL INOVASI TEKNOLOGI INFORMASI & KOMUNIKASI
"Optimalisasi Teknologi Kecerdasan Artifisial untuk Mendukung Transformasi Digital dan Masa Depan Otomasi"

LPPM Universitas Sanata Dharma, Yogyakarta
IndoCEISS Provinsi DI. Yogyakarta dan IndoCEISS Kalimantan Tengah

BUKU ELEKTRONIK (e-BOOK):
ISBN: 978-623-143-058-8 (PDF)
EAN: 9-786231-430588

Cetakan Pertama, Agustus 2024
vii+215 hlm.; 21x27,9 Cm.

SAMPUL & LAYOUT AKHIR BUKU
Moh. Ali Romli, S.Kom., M.Kom.
Thomas Aquino Hermawan M.

KEPANITIAAN

Pengarah & Penanggung Jawab:
Prof. Dr. Enny Itje Sela, S.Si., M.Kom.
Ketua Panitia:
Muhammad Fachrie, S.T., M.Cs.
Wakil Ketua:
Saucha Diwandari, S.Kom., M.Eng.
Sekretaris:
Prita Haryani, S.Pd., M.Eng.
Anna Dina Kalifia, S.Kom., M.Cs.
Bendahara:
Dr. Anastasia Rita Widiarti, S.Si., M.Kom.
Sie Acara:
Ryan Ari Setyawan, S.Kom., M. Eng.
Fatsyahrina Fitriastuti S.Si., M.T.
Suparyanto, S.T., M.Eng.
Sie Makalah:
Lovandri Dwanda Putra, M.Pd.
Anik Andriani, M.Kom.
Vynska Amalia Permadi, M.Kom.
Sulistyo Dwi Sancoko, S.Si., M.Sc.
Sylvert Prian Tahalea, S.Si., M.Cs.
Moh. Ali Romli, S.Kom., M.Kom.
Sie Publikasi Dekorasi Dokumentasi:
Rr. Yuliana Rachmawati K, S.T., M.T.
Dr. Widyastuti Andriyani, M.Kom.
Edy Prayitno, S.Kom., M.Eng.
Jatmika, S.Si., M.Kom.
Sie. Sarana & Prasarana:
Kharisma, S.T., M.Cs.

DITERBITKAN OLEH

SANATA DHARMA UNIVERSITY PRESS
Lantai 1 Gedung Perpustakaan USD
Jl. Affandi (Gejayan) Mrican, Yogyakarta 55281
Telp. (0274) 513301, 515253; Ext. 51513; Fax (0274) 562383
Website: www.sdupress.usd.ac.id
e-Mail: publisher@usd.ac.id

Sanata Dharma University Press anggota APPTI
(Afiliasi Penerbit Perguruan Tinggi Indonesia)
No. Anggota APPTI: 003.028.1.03.2018

INSTITUSI PENDUKUNG/KERJA SAMA

Lembaga Penelitian dan Pengabdian kepada Masyarakat
Universitas Sanata Dharma Yogyakarta

IndoCEISS Provinsi D.I. Yogyakarta

IndoCEISS Kalimantan Tengah

Kata Pengantar

Puji syukur ke hadirat Tuhan YME atas segala limpahan karunia dan hidayah-Nya, sehingga kegiatan Seminar Nasional Inovasi Teknologi Informasi & Komunikasi (SENOVTIK) ini telah berhasil diselenggarakan dengan baik. Kegiatan tersebut merupakan agenda ilmiah bagi para akademisi maupun mahasiswa untuk mempresentasikan sekaligus mempublikasikan hasil penelitian yang telah dilakukan kepada khalayak, terutama para akademisi, peneliti, maupun mahasiswa yang berkecimpung di bidang Teknologi Informasi dan Kecerdasan Artifisial.

Kegiatan SENOVTIK ini diselenggarakan oleh Lembaga Penelitian dan Pengabdian kepada Masyarkat Universitas Sanata Dharma IndoCEISS Provinsi D.I. Yogyakarta dan IndoCEISS Kalimantan Tengah bekerjasama dengan pada tanggal 25 November 2023 di Yogyakarta. Seminar Nasional Inovasi Teknologi Informasi & Komunikasi "Optimalisasi Teknologi Kecerdasan Artifisial untuk Mendukung Transformasi Digital dan Masa Depan Otomasi" menghadirkan dua *keynote speaker*, yakni Prof. Dr. Suyanto, S.T., M.Sc. dari Universitas Telkom dan Dr. Sri Hartati Wijono, S.Si., M.Kom. dari Universitas Sanata Dharma, yang telah memberikan pemaparan tentang teknologi Kecerdasan Artifisial yang saat ini banyak diterapkan di berbagai bidang. Kemudian, pada sesi '*Call for Paper*', terdapat 37 artikel ilmiah yang lolos hasil seleksi oleh para *reviewer* yang berkompeten di bidangnya. Semua artikel tersebut juga telah dipresentasikan secara daring pada tanggal 23 November 2023. Artikel-artikel tersebut dikelompokkan menjadi dua kategori utama berdasarkan topik risetnya, yakni kategori sistem cerdas dan kategori teknologi web & mobile. Bahasan utama pada kategori sistem cerdas mencakup topik di bidang Machine Learning, Data Mining, Deep Learning, termasuk topik-topik khusus mengenai analisis sentimen dan Text Mining. Kemudian, pada kategori teknologi web & mobile, artikel didominasi oleh topik seputar sistem informasi berbasis web dan juga Android.

Kami berharap agar kegiatan SENOVTIK ini dapat menjadi wadah pembelajaran bagi para mahasiswa dan peningkatan kompetensi bagi para akademisi. Terima kasih kepada seluruh panitia yang telah bekerja keras menyiapkan dan menyelenggarakan kegiatan ini, terutama kepada LPPM Universitas Sanata Dharma dan pengurus IndoCEISS Provinsi DI. Yogyakarta dan IndoCEISS Kalimantan Tengah yang terlibat aktif di dalam kegiatan ini. Kami juga mengucapkan terima kasih kepada seluruh pihak yang telah ikut memberikan sumbangsih pada kegiatan ini, baik secara materil maupun non-materil. Kami juga memohon maaf jika pada penyelenggaraan SENOVTIK ini terdapat berbagai kekurangan. Semoga pada kegiatan mendatang dapat terselenggara dengan lebih baik.

Yogyakarta, Maret 2024

Ketua Panitia Seminar,

Muhammad Fachrie, S.T., M.Cs.

Daftar Isi


SENOVTIK

Analisis Pengaruh *Privilege* terhadap Prestasi Akademik Siswa Menggunakan Regresi Logistik

Nasmah Nur Amiroh
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
nasmah.5200411196@student.uty.ac.id

Neneng Nur Sholihah
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
neneng.5200411216@student.uty.ac.id

Ilmy Eka Handayani
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
ilmy.5200411217@student.uty.ac.id

Siti Nurazila
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
siti.5200411233@student.uty.ac.id

Satriya Adhitama
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
satriya.5200411545@student.uty.ac.id

Afifah Inas Hanifah
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
afifah.5200411042@student.uty.ac.id

Ery Hartati
Departemen Teknik Informatika
Universitas Multi Data Palembang
Palembang, Indonesia
ery_hartati@mdp.ac.id

Abstrak—Privilege adalah hak istimewa yang diperoleh tanpa usaha namun dapat diperoleh melalui berbagai cara. Permasalahan utama yang menyebabkan adanya pro dan kontra di masyarakat mengenai privilege adalah keyakinan bahwa ketidakmampuan seseorang merupakan sesuatu yang memalukan dan perlu diatasi. Jika situasi ini tidak segera diatasi, maka kehidupan masyarakat akan terus menderita. Dalam penelitian ini, metode statistik dan Regresi Logistik digunakan untuk menganalisis korelasi antara privilege yang dimiliki oleh individu dan kinerja mereka dalam pekerjaan atau pendidikan. Untuk mendapatkan koefisien korelasi antara variabel dan atribut, penelitian dilakukan dengan menggunakan berbagai metodologi. Hasil training loss diperoleh nilai 1.4386, sementara hasil testing loss adalah 1.5523, dan hasil MCC yang diperoleh adalah 0.3704 berdasarkan temuan penelitian. Temuan penelitian ini menunjukkan bagaimana pendekatan berbeda yang diambil mencerminkan temuan analisis korelasi antara privilege orang tua dan kinerja di antara siswa. Temuan penelitian ini dianggap sebagai salah satu cara untuk mengatasi stigma dan bagaimana masyarakat memandang masalah ini.

Kata Kunci—Regresi Logistik, Statistik, Prestasi Siswa, Keterlibatan Orang Tua.

I. PENDAHULUAN

Privilege adalah hak khusus, keuntungan, atau kekebalan yang dimiliki dan hanya tersedia untuk orang atau kelompok tertentu [1]. *Privilege* dalam pendidikan merupakan batu loncatan yang dapat menunjang keberhasilan seorang siswa. Akses pendidikan yang lebih tinggi, akses pembelajaran yang berkualitas, serta orang tua yang mendukung dan memiliki latar belakang pendidikan yang baik merupakan beberapa hal yang mendukung atau memberikan *privilege* kepada siswa dalam belajar dan mengenyam pendidikan. [2]. Faktor yang paling penting adalah dukungan orang tua dalam pendidikan siswa. Dukungan orang tua sendiri dapat berupa dukungan finansial dan dukungan mental [3]. Pada umumnya orang tua yang memiliki kesadaran akan pendidikan akan berusaha mendukung pendidikan anaknya dari aspek apapun [4]. Tingkat pendidikan orang tua berpengaruh terhadap keberhasilan prestasi belajar anak. Orang tua yang memiliki tingkat pendidikan yang lebih tinggi juga akan lebih percaya diri dengan kemampuannya dalam membantu anak belajar [5]. Dengan tingkat kepercayaan diri ini, kemampuan akademik siswa dapat meningkat secara signifikan [6]. Namun, hal ini menuai pro dan kontra di kalangan masyarakat. Banyaknya pendapat yang diberikan oleh masyarakat mengenai hal ini menjadi salah satu alasan mengapa penelitian ini dilakukan.

Menurut Pallathadka, et al [7] performa siswa dapat dilihat berdasarkan ketertarikan mereka terhadap suatu mata pelajaran yang mereka sukai pada saat proses ujian nantinya. Pada penelitian ini, pengolahan dan analisis data dilakukan dengan menggabungkan proses penelitian kualitatif dan kuantitatif untuk mendapatkan hasil yang maksimal. Hal ini dibuktikan dengan hasil yang didapatkan dari data yang digunakan dan menggunakan algoritma *machine learning* yang dipilih menghasilkan akurasi yang cukup tinggi, seperti SVM dengan akurasi mendekati 90%. Hal ini juga didukung oleh Hashim, A, et al [8] yang menyatakan bahwa kinerja siswa perlu dilakukan untuk membantu memprediksi keberhasilan siswa dalam belajar, mencegah siswa putus sekolah sebelum ujian akhir, mengidentifikasi siswa yang membutuhkan bantuan tambahan, dan meningkatkan peringkat dan kredibilitas suatu institusi. Selain faktor-faktor tersebut, dukungan yang diberikan oleh orang tua juga sangat berpengaruh terhadap kinerja mahasiswa [9], [10].

Berdasarkan penelitian-penelitian sebelumnya, penelitian mengenai prestasi siswa dengan menggunakan teknik *machine learning* hanya digunakan untuk membandingkan hasil antara metode yang digunakan. Oleh karena itu, selain ingin membuktikan hubungan antara orang tua dengan prestasi siswa, penulis mengusulkan untuk menggunakan beberapa metode statistik secara bersamaan seperti distribusi normal, ANOVA, *Pearson Correlation*, *Chi-Square*, dan menggunakan Regresi Logistik untuk memprediksi dan

mengklasifikasikan hasil prestasi belajar siswa. Regresi Logistik merupakan metode regresi non-linier yang digunakan untuk menggambarkan hubungan non-linier antara variabel dependen Y dengan satu atau lebih variabel independen yang berskala kategorik atau interval. Regresi Logistik digunakan ketika distribusi Y tidak normal, dan variabilitas respon Y tidak konstan, yang tidak dapat dijelaskan oleh model Regresi Linier biasa. [11]. Menurut Hashim, et al [8] Regresi Logistik merupakan metode yang paling akurat dalam penelitiannya yang berhasil memprediksi hasil ujian akhir mahasiswa dengan perbandingan 68,7% lulus dan 88,88% tidak lulus. LİNDBERG & Güven [12] dalam penelitiannya yang memprediksi dan mengklasifikasikan prestasi siswa di Turki menggunakan Regresi Logistik dengan akurasi klasifikasi 49,2% pada siswa yang berprestasi kurang dan 84,4% dalam memprediksi siswa yang berprestasi tinggi. Singh & Alhulail [13] dalam penelitiannya yang melakukan analisis perbandingan terhadap kinerja siswa menyatakan bahwa Regresi Logistik memiliki kinerja terbaik dengan MSE sebesar 0.158611894 dan RMSE sebesar 0.398261088. Oleh karena itu, penelitian ini bertujuan untuk membuktikan hipotesis mengenai masalah korelasi antara prestasi siswa dan keterlibatan orang tua dengan menggunakan Regresi Logistik yang pada proses preprocessingnya diberikan validasi dengan menggunakan metode statistik. Sehingga, dengan hasil yang diperoleh, diharapkan dapat memberikan solusi dan saran terhadap permasalahan dan topik mengenai *privilege* orang tua dalam kesuksesan anaknya di Indonesia.

II. TINJAUAN PUSTAKA

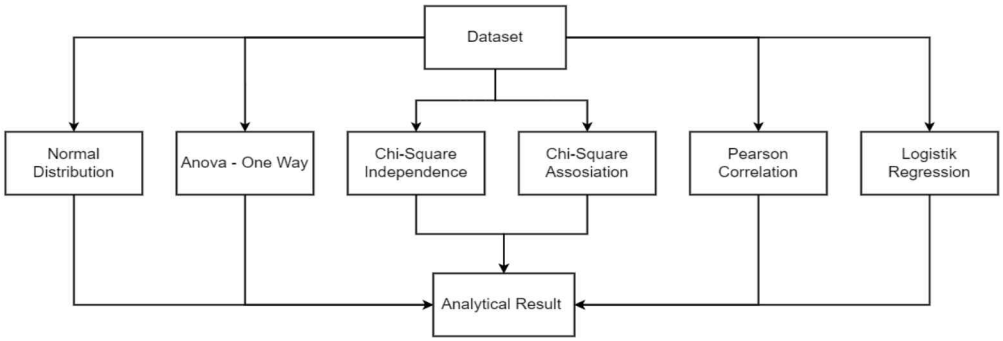
Penelitian oleh Arief et al [14] menggunakan *data ethnicity, family environment, educational level, livelihood, dan income level orang tua* dari 45 siswa. Teknik analisis data menggunakan teknik analisis *univariat*, *bivariat*, dan *multivariat*. Teknik analisis *bivariat* menggunakan uji *Chi-Square ($\chi 2$)* dan teknik analisis *multivariat* menggunakan Regresi Logistik berganda karena variabel dalam penelitian ini bersifat kategorik. Hasil pemodelan multivariat pada Regresi Logistik berganda menunjukkan bahwa secara simultan faktor pendidikan orang tua merupakan faktor yang dominan dan berpengaruh signifikan terhadap hasil belajar siswa. Pendapatan orang tua cenderung berpengaruh lebih rendah terhadap hasil belajar siswa dibandingkan dengan pendidikan orang tua.

Penelitian oleh Lv, dan Han [15] menggunakan data yang bersumber dari China Family Panel Studies (CFPS) pada tahun 2014. Variabel yang digunakan terdiri dari *wage, age, village, degree, wife degree dan family number*. Metode yang digunakan adalah Regresi Logistik dan OLS untuk mengetahui pengaruh pendapatan orang tua terhadap pendidikan anak. Pendapatan keluarga tidak memiliki pengaruh yang signifikan terhadap tingkat pendidikan anak baik pada populasi maupun sampel perkotaan. Namun, pendapatan keluarga di pedesaan memainkan peran yang lebih besar dalam meningkatkan prestasi akademik anak. Oleh karena itu, pendapatan orang tua memiliki dampak yang signifikan terhadap tingkat pendidikan anak, yang diasumsikan meningkat seiring dengan meningkatnya pendapatan.

Penelitian oleh St-Denis [16] bertujuan untuk mengetahui pengaruh pendapatan dan pendidikan orang tua terhadap pencapaian pendidikan anak dengan menggunakan data dari Longitudinal and International Study of Adults (LISA), Wave 3 (2016), yang terbatas pada responden LISA pada kelompok kelahiran 1964 hingga 1980. Metode yang digunakan pada model pertama adalah Regresi Logistik dari pendidikan orang tua dan pendapatan orang tua, secara terpisah dan kemudian digabungkan dalam model *multivariat*. Model kedua adalah model *Regresi Ordinary Least Squared* (OLS) dengan jumlah tahun pendidikan anak sebagai variabel dependen. Variabel ini diturunkan dari variabel kategori kredensial pendidikan tertinggi yang diperoleh. Hasil dari penelitian ini menunjukkan bahwa pendapatan orang tua dan pendidikan orang tua memiliki hubungan yang signifikan dengan pencapaian pendidikan anak. Namun, variabel-variabel tersebut berkorelasi, dan kontribusi pendidikan orang tua lebih besar dibandingkan dengan pendapatan orang tua. Dengan selisih yang relatif besar, kedua karakteristik latar belakang tersebut dimasukkan ke dalam model regresi multivariat.

III. METODOLOGI

Pendekatan yang digunakan dalam penelitian ini adalah metode analisis dan statistik yang mencakup distribusi normal, analisis varians (ANOVA), *Pearson Correlation*, *Chi-Square*, dan terakhir Regresi Logistik. Regresi Logistik digunakan untuk membangun model statistik untuk mencapai tujuan penelitian. [17]. Gambar berikut adalah metodologi yang digunakan dalam penelitian ini.



Gambar 1. Metodologi Penelitian

Metode yang digunakan dapat dilihat pada Gambar 1, penelitian ini menggunakan berbagai macam metode dengan menggunakan dataset yang sama, yaitu dataset maths dari data Kinerja Siswa. Setiap data diolah dengan menggunakan

metode distribusi normal, ANOVA *one-way*, *Chi-Square independence*, *Chi-Square association*, *Pearson Correlation*, dan yang terakhir adalah Regresi Logistik. Dalam hal ini, Regresi Logistik digunakan untuk mencapai tujuan penelitian. Hasil dari setiap proses metode tersebut kemudian akan dianalisis sebagai hasil penelitian.

A. *Dataset Kinerja Siswa*

Dataset Kinerja Siswa meneliti kinerja siswa dalam pendidikan menengah di dua sekolah di Portugal. Data ini mencakup berbagai atribut seperti nilai siswa, rincian demografis, faktor sosial, dan fitur terkait sekolah. Data dikumpulkan melalui kombinasi laporan sekolah dan kuesioner. Ada dua set data terpisah yang secara khusus berfokus pada kinerja dalam dua mata pelajaran yang berbeda: Matematika (math) dan Bahasa Portugis (por). Dalam penelitian kami, hanya menggunakan dataset Matematika dari Dataset Kinerja Siswa. Dataset ini terdiri dari 30 atribut fitur dan 3 atribut target. Dari 33 atribut, penelitian ini menggunakan atribut Medu, Fedu, Mjob, dan Fjob untuk mengetahui pengaruh tingkat pendidikan dan jenis pekerjaan orang tua terhadap prestasi akademik siswa (G1-G3). Tabel 1 merupakan tabel deskripsi atribut yang digunakan dalam penelitian ini.

TABEL 1. DESKRIPSI ATRIBUT

Medu	pendidikan ibu (angka: 0 - tidak ada, 1 - pendidikan dasar (kelas 4 SD), 2 - kelas 5 sampai kelas 9, 3 - pendidikan menengah atau 4 - pendidikan tinggi)
Fedu	pendidikan ayah (angka: 0 - tidak ada, 1 - pendidikan dasar (kelas 4 SD), 2 - kelas 5 sampai kelas 9, 3 - pendidikan menengah atau 4 - pendidikan tinggi)
Mjob	pekerjaan ibu (nominal: 'guru', 'kesehatan' yang berhubungan dengan perawatan, 'layanan sipil' (misalnya administrasi atau polisi), 'di rumah' atau 'lainnya')
Fjob	pekerjaan ayah (nominal: 'guru', 'kesehatan' yang berhubungan dengan perawatan, 'layanan sipil' (misalnya administrasi atau polisi), 'di rumah' atau 'lainnya')
Usia	usia siswa (angka: dari 15 hingga 22)
Ketidakhadiran	jumlah ketidakhadiran di sekolah (angka: dari 0 hingga 93)
Waktu belajar	waktu belajar mingguan (angka: 1 - <2 jam, 2 - 2 hingga 5 jam, 3 - 5 hingga 10 jam, atau 4 - >10 jam)
G1	nilai periode pertama (angka: dari 0 hingga 20)
G2	nilai periode kedua (angka: dari 0 hingga 20)
G3	nilai akhir (angka: dari 0 hingga 20, output)

B. *Distribusi Normal*

Distribusi normal dalam bidang statistika berbentuk seperti lonceng atau genta. Distribusi ini digunakan untuk memperkirakan dan memprediksi peristiwa yang akan terjadi. Selain itu, distribusi ini merupakan distribusi probabilitas kontinu yang penting dalam statistika [18] Perhitungan distribusi normal menggunakan mean (μ) dan standar deviasi (σ) [19].

C. *ANOVA One-Way*

Analisis ANOVA satu arah dilakukan untuk mengevaluasi apakah ada perbedaan dalam rata-rata variabel numerik antara berbagai tingkat variabel kategorikal [20]. Secara sederhana, analisis ini menjawab pertanyaan apakah ada perbedaan yang signifikan antara rata-rata kelompok. Meskipun ANOVA melibatkan beberapa variabel, prinsip dasarnya mirip dengan uji t, yaitu menerjemahkan nilai statistik F dan membandingkannya dengan nilai kritis berdasarkan distribusi [21]. Analisis deskriptif dan uji Analisis Varians (ANOVA) satu arah dilakukan dengan mempertimbangkan berbagai variabel ([22]. Pada ANOVA satu arah, untuk menganalisis variasi dengan tujuan untuk menentukan kemungkinan perbedaan di antara rata-rata kelompok dapat dilakukan dengan membagi variasi total menjadi variasi yang disebabkan oleh perbedaan antar kelompok dan variasi yang disebabkan oleh perbedaan dalam kelompok [23]. Hipotesis untuk melakukan analisis ANOVA adalah:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r \qquad (1)$$
$$H_1: Not\ all\ \mu_1 (i = 1,2, \ldots, r) are\ equal$$

D. *Chi Square - Independence*

Uji *Chi-Square* untuk independensi adalah metode untuk membandingkan dua variabel dalam tabel kontingensi untuk menilai apakah keduanya berhubungan [23]. Secara umum, uji ini digunakan untuk melihat apakah distribusi variabel kategorikal berbeda satu sama lain. Statistik uji *Chi-Square* yang rendah mengindikasikan bahwa data yang diamati sesuai dengan data yang diharapkan, atau terdapat hubungan yang kuat di antara keduanya. Sebaliknya, nilai yang tinggi dari statistik uji *Chi-Square* menunjukkan bahwa tidak ada hubungan yang signifikan. [21]. Hipotesisnya adalah:

$$H_0: kedua\ variabel\ kategorikal\ saling\ independen \qquad (2)$$
$$H_1: kedua\ variabel\ kategorikal\ tidak\ independen$$

Statistik uji *Chi-Square* untuk pengujian ini adalah:

$$x^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - E_{ij})^2}{E_{ij}} \qquad (3)$$
$$l$$

Jumlah elemen dalam sel (i,j), yaitu sel pada baris ke-i dan kolom ke-j (dengan i = 1, 2, ..., r dan j = 1, 2, ..., c), dilambangkan dengan o_{ij}. E_{ij} adalah jumlah yang diharapkan pada sel (i, j) yang didefinisikan dengan :

$$E_{ij} = \frac{R_i C_j}{n} \qquad (4)$$

R_i dan C_j adalah jumlah total untuk baris ke-i dan jumlah total untuk kolom ke-j.

E. *Chi Square - Association*

Chi-Square Association adalah istilah yang merujuk pada uji statistik *Chi-Square* (χ^2) yang digunakan untuk menguji

apakah ada hubungan atau asosiasi antara dua variabel kategorik dalam bentuk tabel kontingensi atau tabel silang. Uji Asosiasi *Chi-Square* digunakan untuk mengetahui apakah distribusi frekuensi dari dua variabel kategorik berhubungan atau tidak. [24].

F. *Pearson Correlation*

Koefisien korelasi Pearson digunakan untuk mengukur hubungan antara dua variabel yang memiliki data kontinu atau numerik [25]. Koefisien korelasi Pearson juga digunakan untuk mengukur tingkat hubungan antara dua fitur, dengan tujuan untuk mengevaluasi adanya multikolinearitas [26]. Korelasi antara variabel dependen dan independen dinilai dengan menggunakan koefisien korelasi Pearson [27]. Koefisien korelasi Pearson (PCC) adalah metode statistik yang umum digunakan untuk mengukur hubungan linier antara variabel x dan y. PCC berkisar antara -1 hingga +1, dan perhitungannya dapat dilakukan dengan menggunakan rumus berikut:

$$R = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right)}} \quad (5)$$

Dalam rumus tersebut, "R" mewakili Koefisien Korelasi Pearson antara variabel x dan y, sedangkan "n" adalah jumlah observasi atau data yang tersedia untuk variabel x dan y [28].

G. *Regresi Logistik*

Regresi Logistik adalah metode umum yang digunakan untuk menganalisis dan menjelaskan hubungan antara variabel yang memiliki dua nilai (biner) [11] dan sejumlah variabel yang digunakan sebagai prediktor [8], [29]. Tujuannya adalah untuk menemukan model yang paling sesuai untuk menjelaskan hubungan antara variabel dependen dan variabel independen [30],[31]. Regresi Logistik berkembang bersamaan dengan Regresi Linier, namun keduanya memiliki perbedaan. Perbedaannya terletak pada jenis variabel respon yang dihadapi, yaitu variabel biner pada Regresi Logistik dan variabel kontinu pada Regresi Linier. [8]. Regresi Linier memiliki persamaan sebagai berikut:

$$y = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_n z_n \quad (6)$$

Dalam hal ini, y adalah variabel yang diminati atau variabel respon, sedangkan $z_1, z_2, z_3, \ldots, z_n$, adalah variabel prediktor yang digunakan. Dengan menerapkan fungsi sigmoid pada persamaan ini, kita dapat memperoleh fungsi logistik [32]. Regresi Logistik memiliki persamaan sebagai berikut:

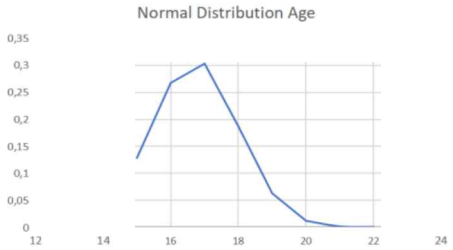
$$\frac{\iota = 1}{[1 + e^{-(\alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_n z_n)}]} \quad (7)$$

Dalam penelitian ini, model yang akan digunakan adalah Regresi Logistik. Karakteristik dari Regresi Logistik adalah kemampuannya dalam membuat prediksi yang bersifat deterministik dan fleksibel dalam menyesuaikan diri dengan berbagai prediksi [33]. Model Regresi Logistik beroperasi dengan memanfaatkan konsep fungsi logit untuk mengidentifikasi probabilitas terjadinya suatu kejadian berdasarkan berbagai nilai dari kelas yang dituju, yang dipengaruhi oleh nilai-nilai prediktor (baik yang berupa variabel kontinu maupun diskrit) yang ada. [30].

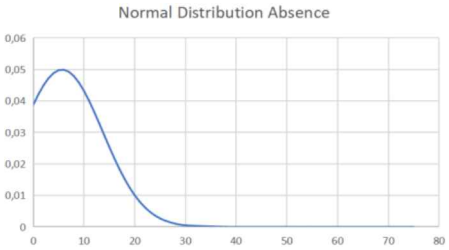
IV. HASIL DAN PEMBAHASAN

A. *Distribusi Normal*

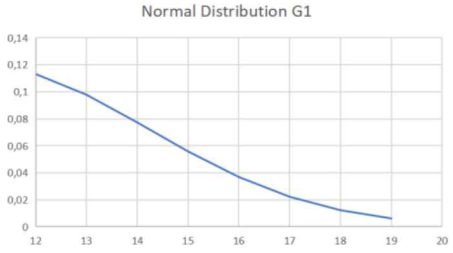
Penelitian ini menggunakan distribusi normal untuk melihat apakah data terdistribusi dengan baik atau tidak. Beberapa variabel yang dilihat adalah *Age, Absence*, G1, G2, dan G3. Distribusi normal dibuat menggunakan excel dengan menghitung nilai rata-rata, standar deviasi, dan kumulatif "FALSE". Gambar 2 merupakan grafik hasil distribusi normal yang diperoleh.



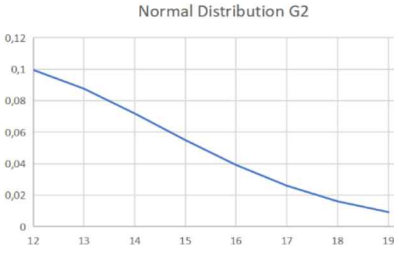
(a)



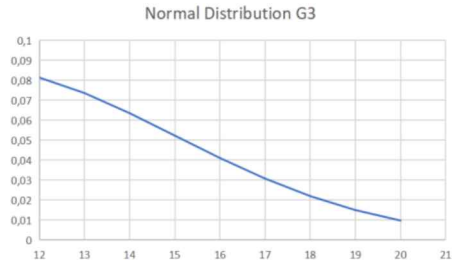
(b)



(c)

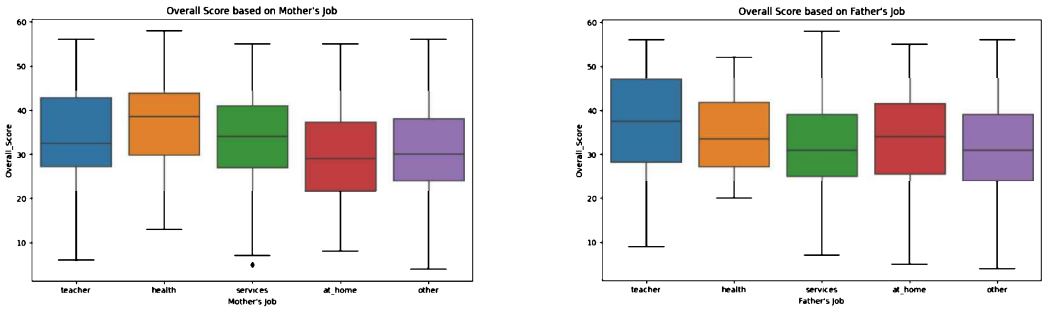


(d)



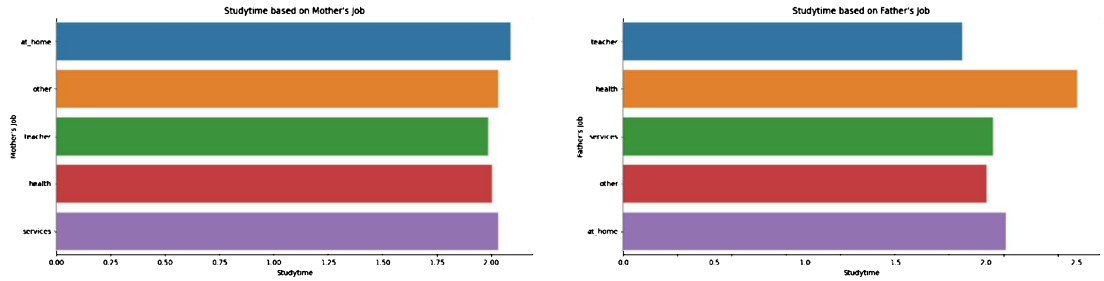
(e)

Gambar 2. (a) Distribusi Normal Usia (b) Distribusi Normal Ketidakhadiran (c) Distribusi Normal G1 (d) Distribusi Normal G2 (e) Distribusi Normal G3



(a) (b)

Gambar 3. (a) Nilai Keseluruhan berdasarkan pekerjaan Ibu (b) Nilai Keseluruhan berdasarkan pekerjaan Ayah



(a) (b)

Gambar 4. (a) Pekerjaan Ibu berdasarkan waktu belajar (b) Pekerjaan Ayah berdasarkan waktu belajar

Berdasarkan kurva di atas, terlihat bahwa variabel *Age*, *Absence*, G1, G2, G3 tidak terdistribusi secara normal. Data yang berdistribusi normal biasanya membentuk kurva seperti lonceng, sedangkan kurva variabel yang diuji tidak demikian. Kurva dari variabel-variabel di atas merupakan kurva lonceng yang tidak sempurna.

B. *Hubungan Pekerjaan Orang Tua dengan Prestasi Akademik*

Dalam penelitian ini, kami menggunakan ANOVA untuk menilai pengaruh pekerjaan ibu (*Mjob*) dan ayah (*Fjob*) terhadap kinerja akademik secara keseluruhan (jumlah nilai G1, G2, dan G3) seperti yang terdapat pada gambar 3. Melalui pendekatan statistik ini, kami bertujuan untuk memahami sejauh mana korelasi antara faktor-faktor ini dan kinerja akademik siswa secara keseluruhan.

Analisis menunjukkan adanya efek utama yang signifikan dari kategori pekerjaan ibu ($F = 4{,}74$, $p = 0{,}00095$) yang mengindikasikan bahwa siswa yang memiliki ibu dengan kategori pekerjaan tertentu mencapai nilai keseluruhan yang lebih tinggi secara signifikan. Khususnya, siswa yang memiliki ibu yang bekerja di sektor kesehatan menunjukkan keunggulan yang berbeda dalam pencapaian akademik dibandingkan dengan siswa yang memiliki ibu yang bekerja di kategori pekerjaan lain. Namun, ketika memperluas analisis ke kategori pekerjaan ayah, hasil ANOVA (F-Statistic = 1,61, P-Value = 0,17) menunjukkan hasil yang kurang meyakinkan atau gagal mencapai signifikansi statistik. Tidak seperti temuan yang kuat untuk pekerjaan ibu, hal ini menunjukkan bahwa pengaruh pekerjaan orang tua terhadap prestasi akademik siswa mungkin lebih jelas ketika mempertimbangkan kategori pekerjaan ibu. Eksplorasi lebih

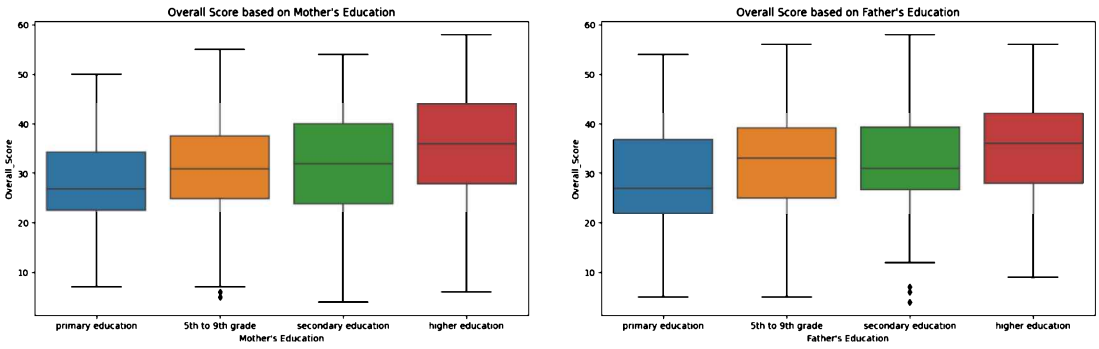
lanjut terhadap berbagai faktor yang berkontribusi diperlukan untuk mendapatkan pemahaman yang lebih baik tentang hubungan yang rumit antara peran orang tua dan keberhasilan siswa. Kami memeriksa korelasi antara waktu belajar (jam belajar mingguan) dan pekerjaan orang tua seperti yang terdapat pada gambar 4. Dalam analisis pekerjaan dan waktu belajar, statistik *Chi-Square* sebesar 5,20 diperoleh dengan nilai P-value sekitar 0,95. Nilai P-value yang tinggi mengindikasikan bahwa tidak ada perbedaan yang signifikan dalam waktu belajar siswa berdasarkan pekerjaan ibu mereka. Hasil ini menegaskan bahwa pilihan pekerjaan ibu tidak memiliki pengaruh yang besar dalam membedakan waktu belajar siswa. Beralih ke analisis F pekerjaan dan waktu belajar, diperoleh statistik *Chi-Square* sebesar 14,79 dengan nilai P-value sekitar 0,25. Meskipun nilai P-value masih di atas tingkat signifikansi yang umum, tidak ada hubungan yang signifikan antara pekerjaan ayah dan jam belajar mingguan siswa. Nilai yang diperoleh menunjukkan bahwa variasi waktu belajar tidak dapat secara signifikan dikaitkan dengan perbedaan pekerjaan ayah.

C. *Hubungan Pendidikan Orang Tua dengan Prestasi Akademik*

Kami melakukan uji varians ANOVA untuk menganalisis hubungan antara tingkat pendidikan ibu (Medu), tingkat pendidikan ayah (Fedu), dan nilai prestasi akademik secara keseluruhan (jumlah nilai dari G1, G2, dan G3) seperti yang terdapat pada gambar 5. Melalui pendekatan statistik ini, kami mengidentifikasi korelasi antara tingkat pendidikan ibu (Medu) dan tingkat pendidikan ayah (Fedu) dengan nilai keseluruhan siswa.

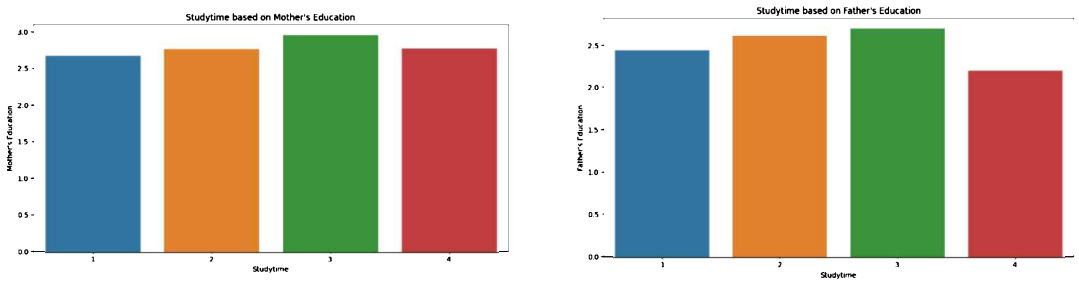
Nilai F-Statistik yang diperoleh dari tingkat pendidikan ibu adalah 9.10 dengan P-Value 7.75e-06 menunjukkan efek utama yang sangat signifikan. Siswa dengan pendidikan ibu yang lebih tinggi mencapai nilai keseluruhan yang lebih baik secara signifikan, menyoroti pengaruh substansial dari pendidikan ibu terhadap keberhasilan akademik. Demikian pula, ANOVA untuk hubungan antara tingkat pendidikan ayah (Fedu) dan kinerja akademik secara keseluruhan menghasilkan efek utama yang signifikan (F-Statistic = 5,32, P-Value = 0,0013). Siswa dengan berbagai tingkat pendidikan ayah menunjukkan nilai keseluruhan yang berbeda secara statistik, yang menekankan dampak pendidikan ayah terhadap prestasi akademik.

Gambar 6 merupakan grafik pemeriksaan korelasi waktu belajar dan tingkat pendidikan orang tua. Dalam pemeriksaan korelasi antara waktu belajar (jam belajar mingguan) dan tingkat pendidikan orang tua, uji *Chi-Square* digunakan. Menganalisis pendidikan ibu (Medu) dan waktu belajar, statistik Chi-kuadrat sebesar 5,20 diperoleh dengan nilai P-value sekitar 0,95. Nilai P-value yang tinggi menunjukkan tidak adanya perbedaan yang signifikan dalam waktu belajar siswa berdasarkan tingkat pendidikan ibu mereka, yang menekankan bahwa tingkat pendidikan ibu yang berbeda-beda tidak memiliki dampak yang besar dalam membedakan waktu belajar siswa. Demikian pula, dalam analisis pendidikan ayah (Fedu), statistik *Chi-square* sebesar 11,84 diperoleh dengan nilai P-value sekitar 0,22. Meskipun nilai P-value tetap berada di atas tingkat signifikansi yang lazim, tidak ada hubungan yang signifikan antara tingkat pendidikan ayah dan jam belajar mingguan siswa. Nilai yang diperoleh menunjukkan bahwa variabilitas waktu belajar tidak dapat secara signifikan dikaitkan dengan perbedaan tingkat pendidikan ayah.



(a) (b)

Gambar 5. (a) Nilai Keseluruhan berdasarkan Pendidikan Ibu (b) Nilai Keseluruhan berdasarkan Pendidikan Ayah



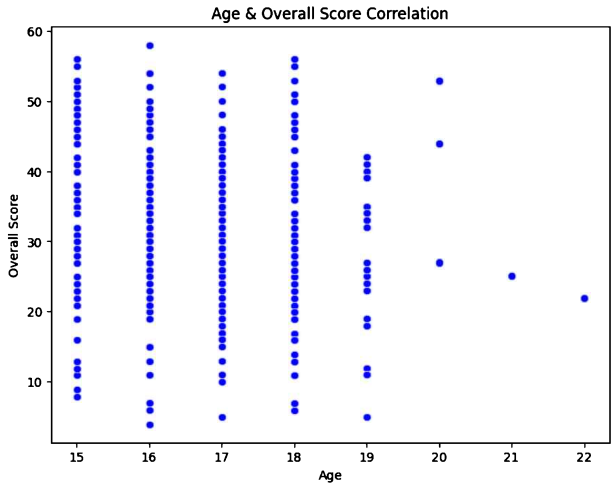
(a) (b)

Gambar 6. (a) Pendidikan Ibu berdasarkan waktu belajar (b) Pendidikan Ayah berdasarkan waktu belajar

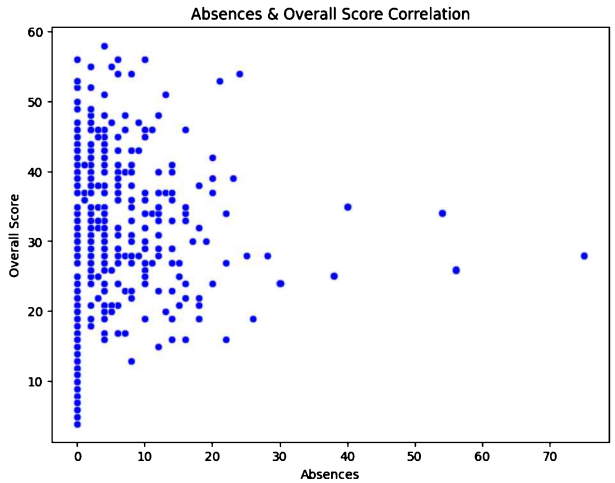
D. *Hubungan Usia dan Ketidakhadiran dengan Prestasi Akademik*

Dalam penelitian ini, kami menggunakan analisis korelasi Pearson sebagai alat untuk menilai hubungan antar variabel. Secara khusus, kami menggunakan metode ini untuk mengeksplorasi korelasi antara usia dan jumlah ketidakhadiran dengan nilai keseluruhan. Tujuannya adalah untuk memahami sejauh mana variabel-variabel ini terkait dengan kinerja akademik siswa secara keseluruhan.

Perhitungan korelasi Pearson antara usia dan *Overall Score* menghasilkan nilai sekitar -0,14. Nilai negatif ini menunjukkan hubungan yang lemah antara usia dan *Overall Score*. Secara sederhana, seiring bertambahnya usia seseorang, nilai akademis mereka cenderung sedikit menurun. Nilai P-value sekitar 0,005 menunjukkan tingkat signifikansi yang cukup rendah. Dengan kata lain, kami memiliki cukup bukti untuk menolak hipotesis nol yang menyatakan bahwa tidak ada korelasi antara usia dan *Overall Score*. Secara keseluruhan, terdapat korelasi yang signifikan secara statistik, namun lemah, antara usia dan kinerja akademik. Meskipun korelasinya lemah, hasil ini menunjukkan tren bahwa seiring bertambahnya usia seseorang ada kecenderungan nilai akademis yang sedikit lebih rendah. Grafik korelasi antara usia dengan kinerja akademik terdapat pada gambar 7.



Gambar 7. Korelasi Usia dan Overall Score



Gambar 8. Korelasi Ketidakhadiran dan Nilai Keseluruhan

Gambar 8 merupakan grafik korelasi antara jumlah ketidakhadiran dan nilai yang diperoleh. Hasil korelasi Pearson antara variabel ketidakhadiran dan nilai keseluruhan adalah sekitar 0.0019, dengan nilai P-value sekitar 0.9694. Nilai korelasi yang sangat kecil (mendekati nol) mengindikasikan bahwa tidak ada hubungan yang kuat antara jumlah ketidakhadiran dan nilai keseluruhan siswa. Nilai P-value yang tinggi, melebihi tingkat signifikansi yang umum (biasanya 0,05), menunjukkan bahwa nilai korelasi ini mungkin terjadi secara kebetulan dan tidak signifikan secara statistik. Dengan kata lain, tidak ada cukup bukti untuk menyatakan bahwa jumlah ketidakhadiran secara signifikan mempengaruhi nilai siswa secara keseluruhan. Dalam konteks ini, dapat disimpulkan bahwa berdasarkan perhitungan korelasi Pearson, tidak ada hubungan yang jelas antara tingkat ketidakhadiran dan prestasi akademik siswa secara keseluruhan.

E. *Analisis Multivariat*

Analisis multivariat dalam penelitian ini dilakukan untuk menyelidiki pengaruh tingkat pendidikan dan pekerjaan orang tua terhadap prestasi akademik siswa. Variabel-variabel tersebut, yaitu tingkat pendidikan ibu (Medu), tingkat pendidikan ayah (Fedu), pekerjaan ibu, dan pekerjaan ayah, secara kolektif diperiksa untuk memahami hubungannya terhadap nilai siswa di G1, G2, dan G3.

TABEL 2. NILAI KONVERSI

Kisaran total G1, G2, dan G3	Kelas
50 - 60	A
40 - 49	B
30 - 39	C
20 - 29	D
11 - 19	E
0 - 10	F

Tabel 2 di atas, mengilustrasikan konversi total nilai di G1, G2, dan G3 ke dalam nilai huruf (A, B, C, D, E, dan F) berdasarkan rentang tertentu. Transformasi ini dilakukan untuk melabeli dataset dengan menjumlahkan ketiga jenis nilai tersebut, sehingga mendapatkan kinerja keseluruhan untuk siswa selama satu semester. Melalui proses pelabelan ini, penelitian ini menggunakan Regresi Logistik multikelas untuk menganalisis hubungan antara tingkat pendidikan orang tua, pekerjaan, dan kinerja akademik siswa.

TABEL 3. HASIL PELATIHAN DAN PENGUJIAN

Akurasi Pelatihan	Kerugian Pelatihan	Akurasi Pengujian	Kerugian Pengujian	PKS
39.21%	1.4386	55%	1.5523	0.3704

Tabel 3 di atas, menyajikan hasil evaluasi model Regresi Logistik multikelas untuk mengklasifikasikan nilai kinerja siswa. Regresi Logistik menghasilkan nilai Koefisien Korelasi Matthews (MCC) sebesar 0,3704. Nilai MCC yang mendekati 0 menunjukkan bahwa tingkat pendidikan dan pekerjaan orang tua tidak secara signifikan mempengaruhi kinerja akademik siswa dalam studi mereka.

V. KESIMPULAN DAN PEKERJAAN DI MASA DEPAN

A. *Kesimpulan*

Keberhasilan siswa dalam belajar tergantung dari diri mereka sendiri dan usaha mereka. Dalam hal ini, *privilege* yang diberikan orang tua hanya sebagai faktor pendukung atau eksternal yang dapat menunjang proses belajar siswa. Dalam penelitian ini, berbagai aspek diuji untuk mengetahui korelasinya dengan prestasi belajar siswa. Namun, hasilnya menunjukkan bahwa usaha dan kemampuan diri sendiri merupakan hal yang paling penting. Hal ini dibuktikan dengan akurasi yang dihasilkan pada hasil training sebesar 39,21% dan testing sebesar 55% dengan menggunakan metode Regresi Logistik. Selain itu, pada percobaan pengujian dengan metode statistik seperti distribusi normal, ANOVA, *Chi-Square*, dan P*earson Correlation* bahwa berbagai aspek tersebut tidak berhubungan satu sama lain. Jadi, dapat disimpulkan bahwa *privilege* orang tua baik materi maupun dukungan fisik dan mental, pekerjaan, dan pendidikan orang tua tidak memiliki hubungan dan keterikatan dalam pembuktian hasil belajar atau prestasi dan keberhasilan belajar mahasiswa.

B. *Pekerjaan di Masa Depan*

Dalam penelitian ini, pengembangan lebih lanjut diharapkan dapat dilakukan dengan melakukan beberapa hal, antara lain:

a) Data yang digunakan dilengkapi dengan berbagai atribut tambahan untuk melihat korelasi yang lebih dalam.

b) Membandingkan pengaruh pendidikan dan pekerjaan orang tua di berbagai wilayah geografis, budaya, atau ekonomi yang berbeda untuk mengidentifikasi perbedaan yang signifikan dalam hasil prestasi belajar anak-anak.

c) Melakukan penelitian lebih lanjut untuk memahami peran faktor psikologis seperti ekspektasi orang tua, motivasi siswa, dan efikasi diri dalam pengaruh pendidikan dan pekerjaan orang tua terhadap prestasi belajar anak.

d) Menggunakan metode lain untuk melihat perbedaan hasil analisis.

REFERENSI

[1] S. B. Prakash, "'Not a Single Privilege Is Annexed to His Character,'" Imp. from Begin., pp. 203–237, 2015, doi: 10.12987/yale/9780300194562.003.0009.

[2] A. Fadlan, "Pengaruh Latar Belakang Ekonomi Keluarga Dan Biaya Pendidikan Terhadap Motivasi Belajar Peserta Didik di SMA Negeri 1 Linggabayu," J. Pamator J. Ilm. Univ. Trunojoyo, vol. 15, no. 1, pp. 81–88, 2022, doi: 10.21107/pamator.v15i1.14064.

[3] I. Naite, "Impact of Parental Involvement on Children's Academic Performance at Crescent International School, Bangkok, Thailand," IOP Conf. Ser. Earth Environ. Sci., vol. 690, no. 1, 2021, doi: 10.1088/1755-1315/690/1/012064.

[4] C. Desforges and A. Abouchaar, "The Impact of Parental Involvement , Parental Support and Family Education on Pupil Achievements and Adjustment : A Literature Review with," Education, vol. 30, no. 8, pp. 1–110, 2003, doi: 10.1016/j.ctrv.2004.06.001.

[5] A. Aprilia, "Pengaruh Tingkat Pendidikan Orang Tua Terhadap Prestasi Belajar Siswa MTSN 4 Lombok Tmur," J. Kaji. Kependidikan Islam, vol. 6, no. 2, pp. 110–122, 2021, doi: 10.22515/attarbawi.v6i2.4672.

[6] J. R. Ramadhan and I. Ichsan, "Pengaruh Pendidikan Orang Tua Terhadap Prestasi Belajar Anak Sekolah Dasar," J. Islam. Educ., vol. 2, no. 2, pp. 69–78, 2021.

[7] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, "Classification and prediction of student performance data using various machine learning algorithms," Mater. Today Proc., vol. 80, no. xxxx, pp. 3782–3785, 2023, doi: 10.1016/j.matpr.2021.07.382.

[8] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," IOP Conf. Ser. Mater. Sci. Eng., vol. 928, no. 3, 2020, doi: 10.1088/1757-899X/928/3/032019.

[9] E. R. Peterson, C. M. Rubie-Davies, M. J. Elley-Brown, D. A. Widdowson, R. S. Dixon, and S. Earl Irving, "Who is to blame? Students, teachers and parents views on who is responsible for student achievement," Res. Educ., vol. 86, no. 1, pp. 1–12, 2011, doi: 10.7227/RIE.86.1.

[10] E. Elstad, K. A. Christophersen, and A. Turmo, "The influence of parents and teachers on the deep learning approach of pupils in Norwegian upper-secondary schools," Electron. J. Res. Educ. Psychol., vol. 10, no. 1, pp. 35–56, 2012, doi: 10.25115/ejrep.v10i26.1483.

[11] F. A. Sutanto, H. Yulianton, and B. Hartono, "Application of Logistic Regression to Analyze Student Performance in Elective Courses," vol. 7, no. 12, pp. 20–25, 2021.

[12] E. niha. LİNDBERG and P. GÜVEN, "The Impact of Parental Involvement and Expectations on Elementary School Students' Academic Achievement," İnönü Üniversitesi Eğitim Fakültesi Derg., vol. 22, no. 1, pp. 809–840, 2021, doi: 10.17679/inuefd.888292.

[13] H. P. Singh and H. N. Alhulail, "Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach," IEEE Access, vol. 10, pp. 6470–6482, 2022, doi: 10.1109/ACCESS.2022.3141992.

[14] S. F. B. Arief, I. Azhari, and N. Nasriah, "Socio-Cultural and Social-Economic Analysis of Parents and Their Influence on Learning Outcomes of Class Vii Students of Smp Negeri 11 Tanjungbalai School Year 2020/2021," J. Ilm. Teunuleh, vol. 2, no. 4, pp. 41–58, 2021, doi: 10.51612/teunuleh.v2i4.74.

[15] H. Lv, "The effects of family income on children's education: An empirical analysis of CHNS data," vol. 02002, pp. 49–54, 2017, doi: 10.24104/rmhe/2017.04.02002.

[16] X. St-denis, "The Relative Role of Parental Income and Parental Education in Child Educational Achievement and Socioeconomic Status Attainment: A Decomposition Approach".

[17] S. L. Prabowo, "The effect of ISO and leadership quality on the sustainable development of academic competence and student performance," J. Soc. Stud. Educ. Res., vol. 13, no. 3, pp. 31–55, 2022.

[18] J. Bayesian, : Jurnal, I. Statistika, D. Ekonometrika, and A. Habibi, "Kajian Simulasi Distribusi Sampling, Teorema Limit Pusat Dan Estimasi Parameter," vol. 3, no. 1, pp. 1–27, 2023.

[19] A. S. Pratikno, A. A. Prastiwi, and S. Ramahwati, "Sebaran Peluang Acak Kontinu, Distribusi Normal, Distribusi Normal Baku, Distribusi T, Distribusi Chi Square, dan Distribusi F," Osf Prepr., vol. 27, no. 3, pp. 1–5, 2020.

[20] M. Tadese, A. Kassa, A. A. Muluneh, and G. Altaye, "Prevalence of dysmenorrhoea, associated risk factors and its relationship with academic performance among graduating female university students in Ethiopia: A cross-sectional study," BMJ Open, vol. 11, no. 3, pp. 1–9, 2021, doi: 10.1136/bmjopen-2020-043814.

[21] "El-Sayed Atlam et al, 2022."

[22] A. Durak and V. Bulut, "Classification and prediction-based machine learning algorithms to predict students' low and high programming performance," Comput. Appl. Eng. Educ., no. August, pp. 1–17, 2023, doi: 10.1002/cae.22679.

[23] "Bernhard O."

[24] A. B. Sakpere, A. G. Oluwadebi, O. H. Ajilore, and L. E. Malaka, "The Impact of COVID-19 on the Academic Performance of Students: A Psychosocial Study Using Association and Regression Model," Int. J. Educ. Manag. Eng., vol. 11, no. 5, pp. 32–45, 2021, doi: 10.5815/ijeme.2021.05.04.

[25] A. N. Pate, S. Neely, D. R. Malcom, K. K. Daugherty, M. Zagar, and M. S. Medina, "Multisite study assessing the effect of cognitive test anxiety on academic and standardized test performance," Am. J. Pharm. Educ., vol. 85, no. 1, pp. 43–54, 2021, doi: 10.5688/ajpe8041.

[26] Y. Ye et al., "Comparison of Machine Learning Methods and Conventional Logistic Regressions for Predicting Gestational Diabetes Using Routine Clinical Data: A Retrospective Cohort Study," J. Diabetes Res., vol. 2020, 2020, doi: 10.1155/2020/4168340.

[27] A. M. Alzahrani, A. Hakami, A. AlHadi, M. A.

Batais, A. A. Alrasheed, and T. H. Almigbal, "The interplay between mindfulness, depression, stress and academic performance in medical students: A Saudi perspective," PLoS One, vol. 15, no. 4, pp. 1–11, 2020, doi: 10.1371/journal.pone.0231088.

[28] Y. Cao et al., "Flash flood susceptibility assessment based on geodetector, certainty factor, and logistic regression analyses in fujian province, china," ISPRS Int. J. Geo-Information, vol. 9, no. 12, pp. 1–22, 2020, doi: 10.3390/ijgi9120748.

[29] F. Ofori, E. Maina, and R. Gitonga, "Using Machine Learning Algorithms to Predict Students' Performance and Improve Learning Outcome: A Literature Based Review," J. Inf. Technol., vol. 4, no. 1, pp. 2616–3573, 2020, [Online]. Available: https://stratfordjournals.org/journals/index.php/Journal-of-Information-and-Techn/article/view/480

[30] D. I. F. S. Engr. Sana Bhutto, Dr. Qasim Ali Arain, Maleeha Anwar, "Through Supervised Machine Learning," Predict. Students' Acad. Perform. Through Supervised Mach. Learn., pp. 1–6, 2020.

[31] S. Amjad, M. Younas, M. Anwar, Q. Shaheen, M. Shiraz, and A. Gani, "Data Mining Techniques to Analyze the Impact of Social Media on Academic Performance of High School Students," Wirel. Commun. Mob. Comput., vol. 2022, 2022, doi: 10.1155/2022/9299115.

[32] M. S. Zulfiker, N. Kabir, A. A. Biswas, P. Chakraborty, and M. M. Rahman, "Predicting students' performance of the private universities of Bangladesh using machine learning approaches," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 3, pp. 672–679, 2020, doi: 10.14569/ijacsa.2020.0110383.

[33] S. Alija, E. Beqiri, A. S. Gaafar, and A. K. Hamoud, "Predicting Students Performance Using Supervised Machine Learning Based on Imbalanced Dataset and Wrapper Feature Selection," Inform., vol. 47, no. 1, pp. 11–20, 2023, doi: 10.31449/INF.V47I1.4519.

Identifikasi Komunitas Topik pada *Academic Citation Network*

Agung Hadhiatma
Departemen Informatika,
Universitas Sanata Dharma
Yogyakarta, Indonesia
agunghad@usd.ac.id

Abstract— Scholarly information such as scientific articles, patents, and books is important for academic research. Scholarly information datasets can be formalized as an academic citation network where nodes can describe the content of scientific articles, and links can represent these references. The academic citation network becomes a complex structure with a massive volume to form Big Data, so it contains high and hidden potential information. Some researchers of scholarly data need network and text analysis. One of the network analyses is graph partition or community detection, while one of the text analyses is topic detection. This research proposes a model that simultaneously performs network and text analyses by combining community detection and topic modeling to identify multi-topic communities. The model generates multi-topic communities represented as communities, *community-topic distributions*, *paper-topic distributions*, *and* topic-word distributions.

Keywords—academic citation network, network analysis, text analysis, multi-topic communities.

I. PENDAHULUAN

Scholarly information merupakan informasi akademik hasil penelitian yang dapat berupa buku, disertasi, *slides*, paten dan artikel ilmiah pada prosiding maupun jurnal. *Scholarly information* disimpan dalam *database* atau dataset artikel ilmiah, contoh Web of Science (WoS), Scopus, Google Scholar, IEEE Xplore, ACM Digital Library dan ACM.

Dataset artikel ilmiah dapat direpresentasikan sebagai *Academic citation network* [1]. *Academic citation network* merupakan (*directed acyclic graph*) yang terdiri atas simpul dan *link*. Simpul dapat diasumsikan dengan artikel ilmiah, pengarang, *venues* dan media publikasi yang lain. Sedangkan *link* adalah referensi atau sitasi yang me-*refer* dari suatu simpul ke simpul lain (misal dari artikel A ke artikel B). Simpul artikel ilmiah dapat mempunyai atribut bertipe teks yaitu: judul, kata kunci, abstraksi dan isi dokumen.

Pencarian, analisis dan penambangan informasi pada *Academic Citation Network* telah menjadi hal yang penting karena potensi informasi yang besar dan kebutuhan untuk memperoleh informasi yang relevan dan cepat. Beberapa penelitian pada dataset artikel ilmiah antara lain: *information retrieval* [2], *academic evaluation* [3], *summarization* [4][5], *topic prediction* [6][7], *academic recommendation* [8] dan lain-lain.

Penelitian pada *Scholarly information* yang direpresentasikan pada *Academic Citation Network* membutuhkan analisis *network* dan analisis teks. Salah satu analisis network yang penting adalah *graph partition* atau *community detection,* sedangkan salah satu analisis teks yang penting adalah *topic detection*. Penelitian ini bermaksud mengusulkan sebuah model yang dapat melakukan sekaligus analisis *network* dan analisis teks. Usulan model dikembangkan dari penggabungan antara *community detection* dengan metode Louvain dan *topic detection* dengan pemodelan LDA yang akan menghasilkan komunitas multi-topik. Motivasi dan alasan untuk meneliti komunitas multi-topik pada *academic citation network* adalah sebagai berikut. Topik terdiri dari kumpulan kata kunci dalam sebuah domain pengetahuan tertentu. Sedangkan komunitas adalah hasil dari *graph partition* atau *clustering graph* yang merupakan sebuah *sub-graph* yang terbentuk berdasarkan kerapatan *link*. Struktur *link* pada komunitas di *academic citation network* dapat menggambarkan mengenai kemiripan topik maupun keterkaitan atau relasi antara satu topik dengan topik yang lain. Komunitas multi-topik adalah sebuah komunitas dengan label multi-topik. Label multi-topik dapat diekstraksi dari komunitas tersebut. Komunitas multi-topik pada artikel ilmiah memungkinkan artikel teks saling terhubung yang dapat membentuk relasi semantik. Komunitas multi-topik yang terbentuk akan dapat digunakan untuk navigasi, seleksi, *filtering* dan pemeringkatan informasi yang dapat digunakan untuk berbagai bidang penelitian, misalnya *information retrieval*, *text mining*, dan rekomendasi artikel ilmiah.

II. DASAR TEORI

Metode yang digunakan untuk deteksi komunitas adalah algoritme Louvain. Metode ini memanfaatkan metode *greedy* untuk optimasi *modularity* dengan strategi menggunakan konsep hierarchical agglomeration [9]. Algoritme ini mempunyai beberapa keunggulan antara lain: kecepatan, akurasi dan kemampuan untuk diset secara multi-skala.

Metode ini hanya mempertimbangkan *non-overlapping nodes*, sehingga $V_i \cap V_j = \emptyset$ untuk semua $i \neq j$ dan setiap *nodes* didefinisikan menjadi anggota dari komunitas. Kualitas dari hasil partisi graph dihitung berdasarkan konsep *modularity* yang didefinisikan dengan Persamaan 1. *Modularity* berfungsi untuk memastikan bahwa kerapatan *link* dalam komunitas lebih besar dari kerapatan *link* antar komunitas.

$$Q(C) = \frac{1}{2m}\sum_{i \in V}\sum_{j \in V}\left(a_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j) \qquad (1)$$

dimana A_{ij} adalah bobot sisi (*edge*) dari simpul (*node)* i ke simpul j, m adalah jumlah sisi pada *graph*, k_i adalah jumlah *degree* dari simpul i, k_j adalah jumlah degree dari simpul j. $\delta(C_i, C_j)$ adalah fungsi Kronecker akan bernilai 1 bila simpul i dan simpul j berada dalam satu komunitas, bila sebaliknya bernilai 0. Pengelompokan simpul dalam satu komunitas ditentukan oleh *modularity gain*. Rumus *modularity gain* pada Persamaan 2.

$$\Delta Q = \left[\frac{\Sigma in + 2\,k_{i,in}}{2m} - \left(\frac{\Sigma tot + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma in}{2m} - \left(\frac{\Sigma in}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \quad (2)$$

Σin adalah jumlah bobot *link* di dalam sebuah komunitas, Σtot adalah jumlah bobot seluruh *link* menuju simpul i, $k_{i,in}$ adalah jumlah bobot *link* dari simpul i ke simpul-simpul yang lain pada satu komunitas, m adalah jumlah bobot dari semua *links* pada *graph*.

Pemodelan topik diselesaikan dengan menggunakan metode pengolahan teks dan model *generative*. Contoh model *generative* adalah: Latent Dirichlet Allocation (LDA). Metode pemodelan topik dengan menggunakan LDA pertama kali diperkenalkan oleh [10]. Metoda LDA melakukan pengelompokan topik dengan mencari distribusi probabilitas topik pada suatu dokumen dan distribusi probabilitas kata pada suatu topik. Langkah Awal dari LDA adalah menentukan jumlah topik, jumlah iterasi, parameter alpha dan beta, kemudian memberikan nilai topik random pada tiap kata. Berikut adalah model LDA yang direpresentasikan pada Algoritme 1, dengan Notasi pada Tabel 1:

Algoritme 1 Generative Topic Model

Input : α, λ, D,
Output : φ_t, θ_d
// t=total numbers of topics
// m=total numbers of documents
// s = total number of word in all documents
1. Choose $\varphi_k \sim$ Dir(λ), where $k \in \{1, \ldots . t\}$,
2. Choose $\theta_d \sim$ Dir(α), where $d \in \{1, \ldots . m\}$
3. **For** each word $w_{d,n}$, $d \in \{1, \ldots, m\}$, $n \in \{1, \ldots, s\}$,
4. Choose randomly a topic $z_{d,n} \sim$ Multinomial(θ_d)
5. Choose randomly a word $w_{mn} \sim$ Multinomial($\varphi_{z_{d,n}}$)
6. **End for**
7. Return φ_k, θ_d

TABEL 1. NOTASI DAN DESKRIPSI

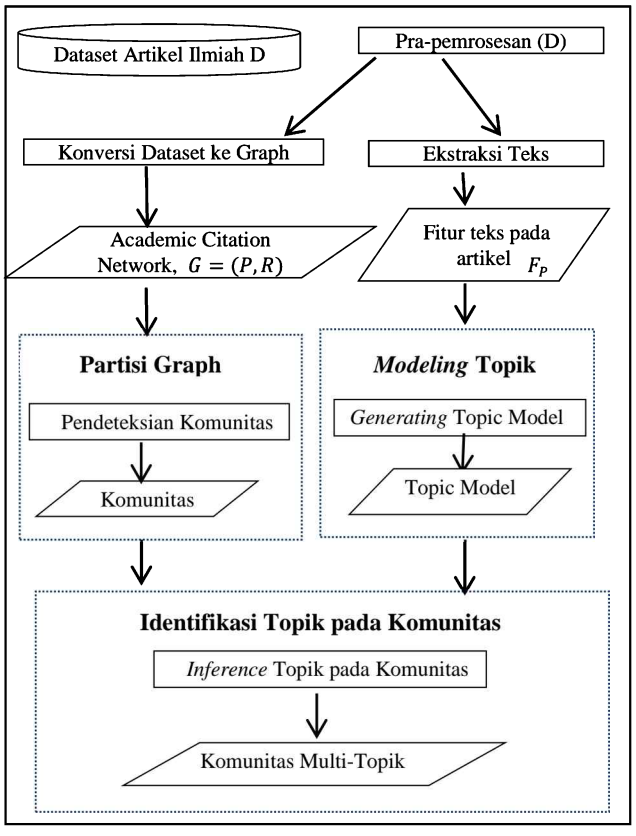
Notasi	Deskripsi
D	Himpunan dokumen
W	Himpunan kata kata pada semua dokumen
A	Parameter Dirichlet prior untuk setiap *document-topic distributions*
β	Parameter Dirichlet prior untuk setiap *topic word distribution*
θ_d	Distribusi topik untuk dokumen ke- d
φ_k	Distribusi kata untuk setiap topik ke- k
$w_{d,n}$	Kata pada dokumen ke-d posisi ke -n
$z_{d,n}$	Penandaan topik pada kata ke-n untuk dokumen ke-d

III. METODOLOGI

Pengembangan model untuk identifikasi komunitas multi-topik pada *academic citation network* ditunjukkan pada Gambar 1. Model identifikasi komunitas multi-topik merupakan penggabungan dari metode *partisi graph* dan pemodelan topik.

Gambar 1 merupakan *flowchart* usulan model untuk identifikasi komunitas-topik pada dataset paper. *Flowchart* tersebut bertujuan untuk mencari komunitas multitopik pada dataset/database artikel ilmiah. Dataset yang digunakan dalam penelitian ini menggunakan dataset DBLB yang sebagian besar diambil dari data ACM Digital Library. *Flowchart* usulan model terdiri dari 4 bagian utama yaitu: 1) pre-processing 2) partisi graph, 3) modeling topik dan 4) identifikasi topik pada komunitas.

Pre-processing merubah dataset artikel ilmiah D menjadi *academic citation network* $G = (P, R)$ dan fitur teks F_P, dimana P adalah himpunan artikel ilmiah, R adalah himpunan referensi, sedangkan fitur teks artikel ilmiah F_P didapatkan dengan model vektor TF-IDF. Artikel ilmiah yang digunakan adalah judul dan abstraksi. Partisi *graph* menggunakan metode *community detection* dengan optimasi *modularity* pada Persamaan 1 dan Persamaan 2 menghasilkan himpunan komunitas artikel ilmiah. Pemodelan topik menggunakan algoritme *generative topic model* LDA (Latent Dirichlet Allocation), seperti pada Algoritme 1 menghasilkan model topik. Sedangkan identifikasi topik pada komunitas yang sudah terbentuk menggunakan metode *inference* topik menghasilkan komunitas multi-topik.



Gambar 1. *Flowchart* Model Identifikasi Komunitas-Topik pada Dataset Artikel Ilmiah.

IV. ANALISA DAN PEMBAHASAN

Berikut adalah analisa dan hasil penelitian mengenai komunitas multi-topik yang dihasilkan dari usulan model sesuai dengan flowchart pada Gambar 1. Dalam penelitian ini, masukan untuk model identifikasi komunitas topik adalah dataset/library artikel ilmiah DBLB (ACM)-Citation-Network V4. Dataset tersebut merupakan sekumpulan artikel ilmiah bidang computer science terdiri dari 1,511,035 artikel

ilmiah dan 2,084,019 sitasi. Proses *pre-processing* juga melakukan *data cleaning*, dimana artikel ilmiah yang mempunyai atribut yang tidak komplit, yaitu yang tidak mempunyai abstraksi dan referensi akan dihapus. Setelah dilakukan pembersihan data, dataset berkurang menjadi 449.498 artikel ilmiah dan 2.047.286 sitasi.

Usulan model menghasilkan komunitas multi-topik yang mempunyai fitur-fitur: himpunan komunitas C, himpunan *community-topic distributions* θ_C , himpunan *paper-topic distributions* θ_{PC} dan himpunan *topic-word distributions* φ_T

Sampel hasil himpunan komunitas C diperlihatkan pada Tabel 2 yang berisi id komunitas ('1','4','7',.., dst), jumlah anggota artikel dalam komunitas dan anggota artikel dalam suatu komunitas (Id artikel). Id komunitas '1' mempunyai jumlah anggota 821 artikel ilmiah dengan sampel beberapa anggota id artikel yaitu '9195', '984349', '175167', …, dst.

TABEL 2. SAMPEL HASIL HIMPUNAN KOMUNITAS C

Id Komunitas	Jumlah artikel pada komunitas	Id artikel sebagai anggota komunitas
'1'	821	'9195', '984349', '175167', '982402', '388932', …, dst
'4'	4672	'11376', '600449', '613092', '613113', …, dst
'7'	1504	'11418', '808029', '499520', '286290', '645835',…dst
'10'	4580	'17965', '1033089', '1032775', '1031167', ,…, dst
'13'	3	'18310', '334116', '349514'
'16'	1684	'18326', '347381', '1113313', '503101', ,…, dst

Setiap komunitas mempunyai distribusi topik θ_C seperti yang terdapat pada tabel 3. Masing-masing komunitas merepresentasikan berbagai topik penelitian. Komunitas multi-topik menggambarkan pengelompokan berbagai artikel ilmiah dengan topik-topik tertentu. Sebagai contoh Id Komunitas '1' adalah komunitas artikel ilmiah yang bersifat multi-topik yang membahas topik utama mengenai topik 4 dan 19 dengan masing-masing bobot 0.661 dan 0.094678. Satu komunitas dapat membahas berbagai topik yang berkaitan. Pada kenyataanya memang banyak penelitian yang terdiri dari multi-topik/multi-bidang.

Masing-masing anggota artikel ilmiah pada setiap komunitas mempunyai *paper-topic distributions* θ_P yang ditunjukkan dalam Tabel 4. Tabel 4 berisi informasi mengenai Id artikel, keanggotaan komunitas, dan distribusi topik pada artikel. Sebagai contoh id artikel '473813' merupakan anggota komunitas 16 yang mempunyai distribusi topik: topik 12 dengan bobot 0.64683, topik 11 dengan bobot 0.12451 dan seterusnya.

Masing-masing topik terdiri dari beberapa kata-kata yang saling mendukung dalam satu domain pengetahuan yang sama. Kata-kata tersebut masing-masing mempunyai nilai bobot membentuk *topic- word distributions* φ_T seperti terlihat pada Tabel 5 dan Tabel 6. Pada percobaan *topic modelling*, jumlah topik di-s*etting* untuk nilai 22 topik yang berbeda. Tabel 5 adalah sampel himpunan φ_T yang berisi contoh untuk topik 11 dan Tabel 6 untuk topik 13. Penentuan jenis topik pada kata-kata yang terbentuk menggunakan analisis domain pengetahuan pada bidang *computer science*. Topik 11 pada Tabel 5 berisi sekumpulan kata-kata dengan domain topik yang berhubungan dengan *storage* dan struktur data. Sedangkan topik 13 pada Tabel 6 berisi sekumpulan kata-kata dengan topik yang berhubungan dengan *information retrieval*.

TABEL 3. SAMPEL HASIL HIMPUNAN COMMUNITY-TOPIC DISTRIBUTIONS θ_C

Id Komunitas	Distribusi komunitas - topik					
'1'	Topik 4 0.661032	Topik 19 0.094678	Topik 11 0.091	Topik 12 0.0608	Topik 20 0.036	Topik 1 0.0187
'4'	Topik 19 0.47063	Topik 4 0.2636	Topik 29 0.1161	Topik 11 0.0359	Topik 12 0.0315	Topik 1 0.0121
'7'	Topik 20 0.75439	Topik 19 0.08559	Topik 1 0.01659	Topik 16 0.01647	Topik 12 0.01487	Topik 2 0.01092
'10'	Topik 2 0.54006	Topik 12 0.25077	Topik 19 0.04654	Topik 4 0.0424	Topik 11 0.03892	Topik 1 0.0266
'13'	Topik 2 0.4478	Topik 12 0.3648	Topik 4 0.0898	Topik 1 0.0538	Topik 11 0.02890	Topik 16 0.0109
'16'	Topik 12 0.5122	Topik 4 0.1315	Topik 7 0.1149	Topi 11 0.1079	Topik 19 0.0478	Topik 1 0.0201

TABEL 4. SAMPEL HASIL HIMPUNAN PAPER-TOPIC DISTRIBUTIONS θ_{PC}

Id Artikel	Id Komunitas	Distribusi artikel-topik					
'473813'	'16'	Topik 12 0.64683	Topik 11 0.12451	Topik 7 0.10107	Topik 22 0.03162	Topik 15 0.03009	Topik 3 0.01798
'20308'	'16'	Topik 12 0.9239	Topik 1 0.04642	-	-	-	-
'472388'	'16'	Topik 12 0.60753	Topik 11 0.19773	Topik 7 0.10645	Topik 10 0.01710	Topik 18 0.06085	-
'1077462'	'14'	Topik 11 0.91825	Topik 15 0.06796	-	-	-	-
'594806'	'14'	Topik 11 0.88118	Topik 1 0.04603	Topik 14 0.0393	Topik 17 0.0195	-	-

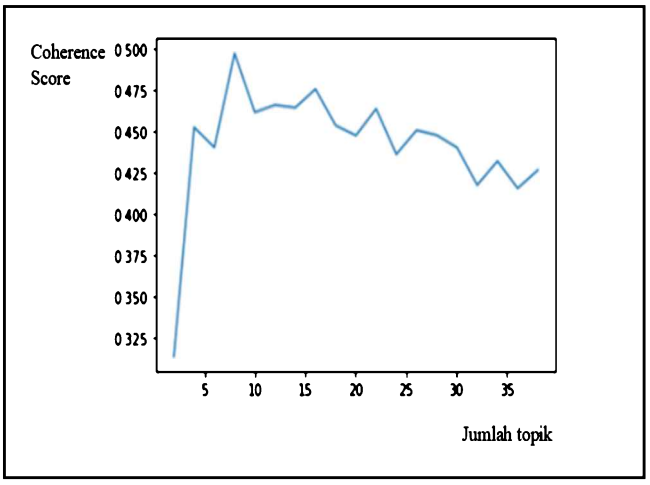
TABEL 5. SAMPEL HASIL HIMPUNAN *TOPIC-WORD DISTRIBUTIONS* φ_T UNTUK TOPIK KE-11

Topik ke 11									
relay	file	disk	index	storage	tree	query	algorithm	memory	datum
0.126362	0.079396	0.061199	0.05951	0.047058	0.045456	0.043671	0.043137	0.042011	0.03979
Hashing	forwarding	hash	time	structure	space	performance	pointer	wsns	scheme
0.037252	0.036608	0.036201	0.035741	0.032419	0.031807	0.030663	0.02992	0.029023	0.028969

TABEL 6. SAMPEL HASIL HIMPUNAN *TOPIC-WORD DISTRIBUTIONS* φ_T UNTUK TOPIK KE-13

Topik ke 13									
query	document	web	search	retrieval	datum	information	user	model	text
0.097248	0.083266	0.068311	0.067179	0.066948	0.065034	0.058248	0.053995	0.051803	0.047448
method	database	classification	system	algorithm	language	semantic	learning	approach	feature
0.046875	0.043042	0.041961	0.040772	0.040434	0.037415	0.036918	0.035647	0.035618	0.034746

Gambar 2 menunjukkan performansi model topik yang diukur dengan *coherence score* terhadap input jumlah topik. Gambar menunjukkan bahwa nilai *coherence score* yang optimum untuk jumlah topik = 7 atau 8. Pada penelitian ini *coherence score* adalah nilai kedekatan kata-kata atau koherensi kata-kata untuk suatu topik pada suatu komunitas yang terbentuk.



Gambar 2. Performansi Model Topic dengan *Coherence Score*

V. KESIMPULAN

1. Usulan model telah dapat menghasilkan himpunan komunitas C, *community-topic distributions* $\boldsymbol{\theta_C}$, *paper-topic distributions* $\boldsymbol{\theta_{PC}}$ dan *topic-word distributions* $\boldsymbol{\varphi_T}$ yang dapat digunakan untuk penelitian ke-depannya dalam bidang *information retrieval*, *text mining*, *paper recommendation* dan lain-lain.

2. Topik yang diekstraksi pada sebuah komunitas artikel ilmiah merupakan sekumpulan kata-kata (*words*) yang dapat mempunyai keterkaitan semantik sesuai dengan domain pengetahuan bidang *computer science*.

REFERENSI

[1] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big Scholarly Data : A Survey," *IEEE Trans. BIG DATA*, vol. 3, no. 1, pp. 18–35, 2017, doi: 10.1109/TBDATA.2016.2641460.

[2] B. Chen, S. Tsutsui, Y. Ding, and F. Ma, "Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval," *J. Informetr.*, vol. 11, no. 4, pp. 1175–1189, 2017, doi: 10.1016/j.joi.2017.10.003.

[3] X. Cai, Y. Zheng, L. Yang, T. Dai, and L. Guo, "Bibliographic Network Representation Based Personalized Citation Recommendation," *IEEE Access*, vol. 7, pp. 457–467, 2019, doi: 10.1109/ACCESS.2018.2885507.

[4] A. Cohan and N. Goharian, "Scientific article summarization using citation-context and article's discourse structure," *Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process.*, no. September, pp. 390–400, 2015, doi: 10.18653/v1/d15-1045.

[5] C. An, M. Zhong, Y. Chen, D. Wang, X. Qiu, and X. Huang, "Enhancing Scientific Papers Summarization with Citation Graph," in *The Thirty-Fifth onference on Artificial Intelligence AAAI*, 2021, doi: 10.48550/arXiv.2104.03057.

[6] K. Asatani, J. Mori, M. Ochi, and I. Sakata, "Detecting trends in academic research from a citation network using network representation learning," pp. 1–13, 2018.

[7] H. Zhu and Y. Mei, "Prediction of online topics ' popularity patterns," *J. Inf. Sci.*, 2020, doi: 10.1177/0165551520961026.

[8] T. Dai, L. Zhu, Y. Wang, H. Zhang, X. Cai, and Y. Zheng, "Joint Model Feature Regression and Topic Learning for Global Citation Recommendation," *IEEE Access*, vol. 7, pp. 1706–1720, 2019, doi: 10.1109/ACCESS.2018.2884981.

[9] V. D. Blondel, J. Guillaume, Renaud Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, pp. 1–12, 2008.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 4–5, pp. 993–1022, 2003, doi: 10.1016/b978-0-12-411519-4.00006-9.

Prediksi Curah Hujan Menggunakan Metode XGboost

Akbar Maulana
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
akbarmaulana55695@gmail.com

Asep Zainal Alfarizi
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
zaenalasep19@gmail.com

Faishal Tirto Nugroho
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
faishaltirto@gmail.com

Rafino Ramdhaniar Prasetyo Putra
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
finoraf01@gmail.com

Julio Ignasius Wangjaya Rumbekwan
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
juliowangjaya@gmail.com

Febrian Sania Putri Vina
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
fputrivina@gmail.com

Abstrak—Cuaca merupakan gambaran dari suatu fenomena alam yang tidak dapat diprediksi seperti hujan dengan intensitas pada saat musim kemarau, udara panas pada saat musim hujan, badai, dan fenomena lainnya. Oleh sebab itu, penelitian ini bertujuan untuk mengklasifikasi prediksi hujan pada hari-hari berikutnya dengan menggunakan salah satu metode klasifikasi yaitu XGBoost yang akan digunakan untuk menentukan kelas mana yang paling optimal. Dari penelitian ini, diharapkan sistem prediksi yang dibuat mendapatkan tingkat akurasi tertinggi dari permasalahan prediksi hujan yang telah diteliti oleh peneliti. Pada penelitian ini, digunakan pembelajaran ensemble yang melibatkan beberapa algoritma untuk mencari nilai rata-rata akurasi. Hasil yang diharapkan dapat menjadi rujukan untuk membangun aplikasi penghitungan cuaca.

Kata Kunci—Klasifikasi, Prediksi, XGboost, Ensemble, Cuaca

I. PENDAHULUAN

Pada masa sekarang ini, faktor cuaca mempengaruhi kehidupan manusia terutama di sektor pertanian, perkebunan, dan penerbangan. Banyak faktor sebagai variabel yang menentukan tentang cuaca dengan melihat kondisi udara pada waktu yang relatif singkat, yang dapat dikelompokkan ke dalam beberapa atribut misalnya tekanan, kecepatan angin, curah hujan, suhu, dan fenomena atmosfer sebagai komponennya. Amril et al (2020) mengungkapkan bahwa permasalahan cuaca yang dilakukan oleh peneliti, prediksi diharapkan memiliki keakuratan yang tinggi terhadap cuaca agar aktivitas manusia tidak terganggu, misalnya sektor pertanian, perkebunan, penerbangan ini sangat tergantung pada kondisi cuaca agar kegiatan tersebut berjalan dengan lancar. Amril et al (2020) mengungkapkan bahwa dengan topik klasifikasi untuk prediksi cuaca menggunakan Ensemble learning, melibatkan beberapa algoritma untuk mencari nilai rata-rata akurasi hasil yang diharapkan bisa menjadi rujukan untuk membangun aplikasi perkiraan cuaca. Hasil akurasi adalah 81,21% dan MSE 18,79%. [1]

Ghaitsa et al (2021) mengungkapkan bahwa penelitian terkait prediksi curah hujan sudah banyak dilakukan menggunakan berbagai metode, diantaranya metode *random forest*, C4.5, dan *classification dan regression trees* (CART). Pada penelitian ini diterapkan algoritma klasifikasi random forest dan XGBoost untuk memperkirakan curah hujan. Random forest melakukan penggabungan pohon/tree dengan melakukan training pada data yang dimiliki. Metode yang digunakan pada penelitian ini sering digunakan karena menghasilkan kesalahan dengan persentase rendah, serta hasil akurasi yang didapat cukup tinggi dalam klasifikasi untuk jumlah data yang sangat besar. [2]

Supriyadi (2019) melakukan penelitian prediksi cuaca menggunakan Deep Learning Long-Short Term Memory (LSTM). Pada penelitiannya menggunakan data yang didapat dari bulan Januari hingga Februari dengan cara mengukur suhu udara, kelembapan udara, kecepatan angin, dan tekanan udara. Data bulan Januari digunakan sebagai data training dan testing, sedangkan data bulan Februari digunakan sebagai pembanding hasil dari training pada bulan Januari. Dari hasil penelitian yang dilakukan menghasilkan RMSE untuk parameter suhu, kelembapan, kecepatan angin, dan tekanan udara masing-masing bernilai 0,576; 2,8687; 2,1963; dan 1,0647. Sedangkan prediksi suhu udara, kelembaban, kecepatan angin, dan tekanan udara untuk 1 hari ke depan (1 Februari 2019) masing-masing sebesar 1,0337; 6,3413; 2,8934; dan 1,4313. [3]

Desmonda et al (2018) menggunakan metode fuzzy time series untuk memprediksi curah hujan. Penelitian ini menghasilkan aplikasi yang dapat mengolah dan menentukan pola data curah hujan serta memprediksi besaran curah hujan. Dari hasil pengujian didapatkan nilai MAPE (Mean Average Percentage Error) yang bervariasi, nilai variasi ini bergantung dari jumlah data dan jumlah interval. Nilai MAPE yang terbaik dari penelitian ini yaitu 0,151% pada penggunaan data curah hujan periode 2015-2017 dengan jumlah interval 401. [4]

Affifah et al (2023) mencoba meneliti curah hujan di Yogyakarta dengan menggunakan algoritma regresi linear berganda. Beberapa variabel yang digunakan dalam penelitian ini yaitu meliputi hal-hal yang berhubungan dengan curah hujan seperti suhu, kelembapan, kecepatan angin dan lama penyinaran matahari. Dilakukan pelatihan dengan menggunakan regresi linier berganda, data yang digunakan dalam penelitian ini yaitu data iklim Yogyakarta pada tahun 2010-2022. Hasil yang diperoleh yaitu R2 score sebesar 12,99%. Prediksi curah hujan di Yogyakarta diperoleh

sebesar 14.42%. Kemudian evaluasi RMSE menghasilkan penyimpangan antara prediksi curah hujan dengan curah hujan sebenarnya sebesar 14.78%. [5]

Pada penelitian ini menggunakan kumpulan data bernama Clean_Weather yang diperoleh dari situs Kaggle.com. Data Clean_Weather adalah data pengamatan cuaca pada tahun 2019-2021. Data yang diperoleh memiliki sebanyak 25 fitur dengan 1.091 total data untuk pengamatan cuaca. Kumpulan data tersebut berupa angka dan huruf yang nantinya akan ditransformasikan menjadi angka.

II. KAJIAN LITERATUR

A. *Kajian Pustaka*

Pada penelitian yang dilakukan oleh Gema Indah Merdekawati dan Ismail (2019) dengan judul "Prediksi Curah Hujan Di Jakarta Berbasis Algoritma Levenberg Marquardt" menghasilkan akurasi sebesar 96%. Hasil tersebut didapatkan dengan menggunakan jaringan syaraf tiruan dengan proses training menggunakan trainlm (levenberg marquardt) sebanyak 1000 epoch. Selanjutnya untuk pengujian Data yang digunakan pada penelitian tersebut terdiri dari 731 record dengan pembagian rasio data sebesar 22:1:1 untuk data latih, data validasi dan data uji. [6]

Penelitian yang dilakukan oleh Syarif et al. (2023) dengan judul "Prediksi Temperatur Cuaca Di Negara Norwegia Menggunakan Metode LSTM" memberikan hasil terbaik dengan RMSE sebesar 1.242 dengan data test sebanyak 12 pengamatan, 48 data latih dan 100 epoch. Selain itu, kesimpulan lain yang didapatkan adalah LSTM efektif dalam menghasilkan prediksi suhu dengan memperhitungkan fluktuasi bulanan dari waktu ke waktu. Meskipun prediksi suhu pada interval yang lebih pendek cenderung lebih fluktuatif, prediksi jangka panjang menunjukkan kecenderungan suhu untuk menjaga kestabilannya dalam jangka waktu yang lebih panjang. [7]

Penelitian yang dilakukan oleh Huda et al. (2023) dengan judul "Prediksi Menggunakan Model Fuzzy Time Series Studi Kasus Curah Hujan di Kabupaten Bandung" menggunakan data presipitasi total bulanan dari tahun 1978 sampai dengan 2020 yang tercatat per tanggal 1 di setiap bulan di Kecamatan Bojongsoang, Kabupaten Bandung. Dari dataset tersebut, data latih yang digunakan dimulai dari tahun 1978 hingga tahun 2017, sedangkan untuk data latih dimulai dari 2018 hingga tahun 2020. Dari penelitian tersebut, didapatkan hasil nilai error MAPE sebesar 0,1917%, MAE sebesar 0,0010, dan RMSE sebesar 2,9829 untuk metode fuzzy time series dengan kelas keanggotaan sebanyak 108. Sedangkan untuk metode SARIMA memberikan hasil nilai error MAPE sebesar 0.0023, MAE sebesar 0.3856, dan RMSE sebesar 0.0032. [8]

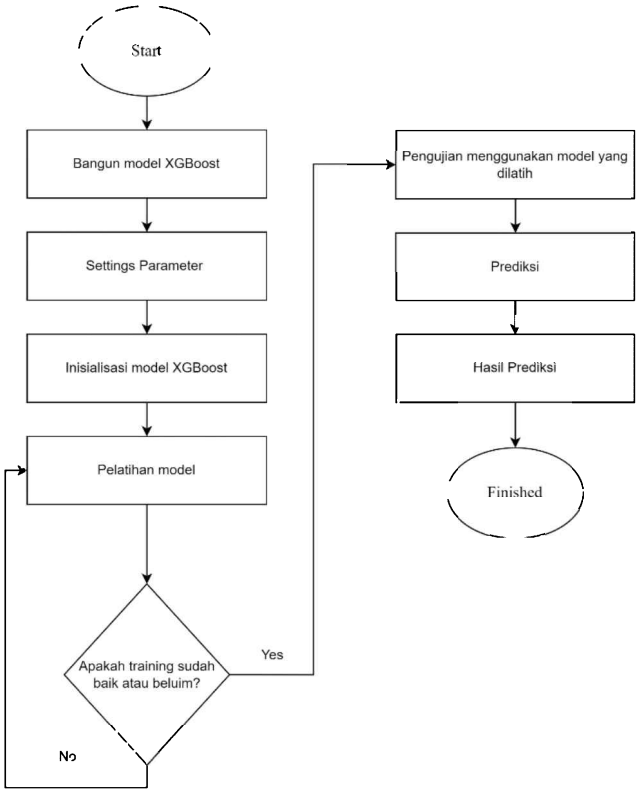
Penelitian yang dilakukan oleh Msy et al. (2021) dengan judul "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir" dengan algoritma CART (Classification and Regression Tree) dengan Teknik data mining CRISP-DM menghasilkan akurasi sebesar 89.4% dengan detail model tersebut mampu memprediksi 110 uji dari 123 data uji. Data yang digunakan pada penelitian tersebut yaitu sebanyak 3.653 record dengan atribut yang terdiri dari suhu rata-rata, suhu min, suhu max, kelembapan, lama penyinaran matahari dan curah hujan. [9]

B. *Dataset Collection*

Pertama, dataset yang digunakan adalah dataset Curah Hujan yang didapatkan dari Kaggle.com. Isi dataset adalah curah hujan dari tahun 2019-2021. Isi datasetnya terdapat temperature suhu, kelembapan, kecepatan angin, dan waktu. Setelah itu, data akan dilakukan pre-processing yaitu menghilangkan data yang missing. Tanda '-' menandakan bahwa tidak adanya hujan sehingga akan dikonversi ke nilai '0'. Peristiwa cuaca khusus (Pck) memiliki beberapa data yang kosong yang berarti tidak adanya kejadian khusus pada hari tersebut. TTU merupakan keadaan dimana hujan terjadi namun dikarenakan sangat kecil maka tidak dapat terukur, sehingga perlu diteliti lagi apakah perlu dituliskan nilai '0' atau nilai yang sangat kecil seperti '0.01'.

C. *XGBoost*

Zhang, Yifan (2023) menyatakan bahwa XGboost menggunakan presentasi rumus Taylor orde dua dan menambahkan suku biasa, yang merupakan kontrol dari kompleksitas pohon dan mencegah overfitting. Untuk menangani nilai yang hilang, algoritma XGboost mencoba memutuskan arah default untuk menangani nilai yang hilang dengan menghitung apakah lebih baik semua nilai yang hilang dimasukkan ke dalam sub-pohon kiri pada node saat ini, atau ke sub-pohon kanan. Bagian yang paling menggunakan banyak waktu dari sebuah weak learner Tunggal adalah proses pemisahan pohon keputusan, yang dapat dipilih secara paralel menggunakan beberapa utas untuk memisahkan titik-titik dengan fitur yang berbeda. [10]



Gambar 1. Alur XGBOOST

Pada gambar 1, merupakan alur bagaimana XGBoost bekerja. Pertama adalah membangun model XGBoost, kemudian setting parameter seperti learning rate η, penalty coefficients γ dan λ. Selanjutnya inisialisasi model XGBoost kemudian melakukan pelatihan model, apabila

hasil training sudah baik maka dilakukan pengujian terhadap model yang sudah dibangun. Terakhir yaitu melakukan prediksi dan akan menghasilkan hasil prediksi.

Untuk dataset E yang ditentukan yang berisi sampel fitur m-dimensi $E = \{(x1, y1)\ (x2, y2) \ldots (xn, yn)\}$, hasil dari model XGBoost untuk input ke-i x_i dapat dinyatakan sebagai:

$$\hat{y}(xi) = y0\ (xi) + \eta - \sum_{k=1}^{Kx} \sum_{j=1}^{Tk} \omega j,k \qquad (1)$$

dimana y_0 ( ) adalah pengklasifikasi dasar, K_x adalah jumlah iterasi, T_k adalah jumlah node daun pohon klasifikasi dan regresi pada iterasi ke-k, $\omega_{j,k}$ adalah nilai pengganti sampel yang sesuai dengan node *ke-j pada* iterasi *ke-k,* η adalah laju pembelajaran. Ketika model XGBoost dilatih, fungsi objektif yang akan dioptimalkan dapat dinyatakan sebagai:

$$L(\varphi) = \sum i\, l(\hat{y}i, yi) + \sum k\, \Omega(\mathrm{f}k), \qquad (2)$$

$$\sum k\, \Omega(\mathrm{f}k) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{Tk} \omega_j^2 \qquad (3)$$

Di mana y_i adalah nilai sebenarnya, $l(\hat{y}_i, y_i)$ adalah fungsi kerugian dari $\hat{y}_i$ dan y_i , $\Omega(\mathrm{f}_k)$ adalah suku biasa, γ dan λ adalah faktor penalti, T adalah jumlah simpul anak dari pohon CART, dan ω_j adalah nilai keluaran untuk simpul anak *ke-j*.

III. METODOLOGI PENELITIAN

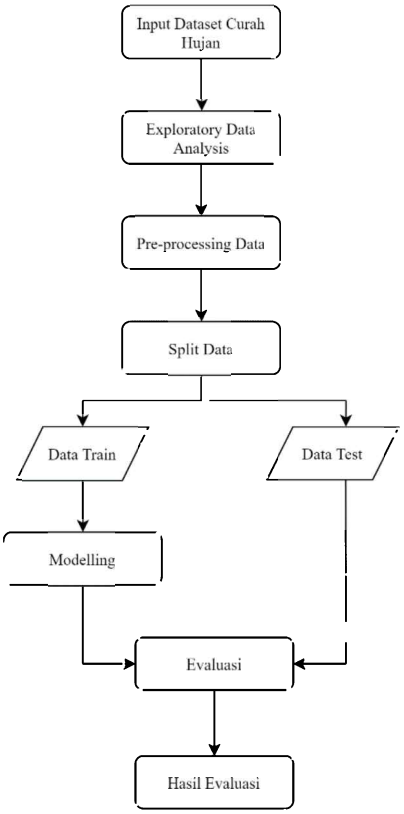
Dalam melakukan penelitian, penulis menggunakan beberapa tahapan metode yang sesuai dengan gambar 2. Tahap pertama yaitu input data, tahap kedua melakukan *preprocessing* data, tahap ketiga melakukan split data, dan yang terakhir yaitu evaluasi terhadap model yang sudah dilakukan percobaan.

A. *Input Dataset*

Pada tahap pertama akan dilakukan penginputan data. Data cuaca yang didapatkan dari kaggle akan diinputkan kedalam program untuk dilakukan pelatihan hingga mendapatkan model untuk memprediksi cuaca. Terdapat total 1.091 data curah hujan, data tersebut adalah data curah hujan dari tahun 2019 hingga 2021.

B. *Pre-processing Data*

Pre-processing adalah tahap untuk merapihkan data. Dataset akan dilakukan pre-processing sehingga data tersebut menjadi lebih baik sehingga kualitas prediksi menjadi lebih baik. Pada preprocessing akan dilakukan tahapan pertama yaitu penghapusan missing value, dengan menghapus missing value data-data yang tidak diperlukan akan dihilangkan dan menyisakan data-data yang penting saja. Selanjutnya yaitu TTU merupakan keadaan dimana hujan terjadi namun dikarenakan sangat kecil maka tidak dapat terukur, sehingga perlu didiskusikan lagi apakah perlu dituliskan '0' atau nilai yang sangat kecil seperti '0.01'. Lalu tanda '-' menandakan bahwa tidak adanya hujan sehingga akan dikonversi ke nilai '0' dan peristiwa cuaca khusus (pck) memiliki beberapa data yang kosong yang berarti tidak adanya kejadian khusus pada hari tersebut.



Gambar 2. Tahapan Metode Penelitian

C. *Split Data*

Setelah dilakukan preprocessing, data akan dibagi menjadi dua data, yaitu data latih dan data test untuk dilakukan pelatihan. Pelatihan ini akan dilakukan untuk mendapatkan model yang dapat digunakan untuk memprediksi curah hujan. Data latih yang digunakan dimulai dari awal data yaitu 1 Januari 2019 hingga 31 Mei 2021, sedangkan untuk data uji dimulai dari 1 Juni 2021 hingga 31 Desember 2021.

D. *Evaluasi*

Pada tahapan ini akan dilakukan pengujian model algoritma regresi XGBoost. dataset akan dilatih hingga mendapatkan model untuk prediksi curah hujan. Evaluasi akan dilakukan dengan cara menampilkan metrik seperti MASE, MRSE, dan MSE. Dengan tahapan evaluasi tersebut, maka dapat diketahui hasil dan kualitas dari prediksi yang sudah dilakukan.

Perhitungan MAE:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |yi - \hat{y}| \qquad (4)$$

Perhitungan MSE:

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (yi - \hat{y})^2 \qquad (5)$$

Perhitungan RMSE:

$$RMSE = \sqrt{MSE} = \sqrt{\sum_{i=1}^{N}(yi - \hat{y})^2} \quad (6)$$

IV. Eksperimen dan Analisis

A. *Eksperimen*

Penelitian ini menggunakan data Clean_Weather yang berasal dari situs Kaggle.com. Dataset tersebut memiliki 25 fitur dengan jumlah data sebanyak 1.091. Penjelasan lebih lanjut mengenai atribut dalam dataset Clean_Weather dapat dilihat pada tabel 1.

Tabel 1 Atribut Variable Dataset

Atribut	Keterangan
Tahun	Tahun
Bulan	Bulan
tgl	Tanggal
temp7	Suhu jam 07.00 WIB
temp13	Suhu jam 13.00 WIB
temp18	Suhu jam 18.00 WIB
temp_avg	Rata-rata suhu
temp_24	Rata-rata Suhu 24 jam
temp_max	Max suhu dalam celcius
temp_min	Min Suhu dalam celcius
CH	Curah hujan dalam mm jam 7 WIB
light_hour	Lama penyinaran matahari dalam jam (08.00-16.00)
light_per	Lama penyinaran matahari (%) (08.00-16.00)
pck	Peristiwa cuaca khusus
tgl.1	Tanggal 1
press	Tekanan udara (mb)
humid7	Rata-rata kelembaban 7 jam
humid13	Rata-rata kelembaban 13 jam
humid18	Rata-rata kelembaban 18 jam
humid_avg	Rata-rata kelembaban
humid_24	Rata-rata kelembaban 24 jam
ws_abg	Kecepatan rata-rata angin (knot)
mod_dir	Arah terbanyak
max_ws	Kecepatan angin terbesar (knot)
dir	Arah angin

a. Pre-processing Data

Setelah dilakukan eksplorasi data, ada beberapa tahapan yang harus dilakukan sebelum dilakukan klasifikasi model, tahapan yang dilakukan yaitu:

1. **Mengisi missing value.** pada visualisasi missing value pada proses eksplorasi data terlihat masih banyak nilai null/missing value dikarenakan data Clean_Weather tidak sepenuhnya bersih dan memiliki banyak noise, lalu dilakukan pengisian missing value dengan menggunakan data modus pada setiap variabelnya. Tahapan terakhir dilakukan pengecekan missing value lagi untuk melihat apakah masih ada missing value atau tidak.
2. **Mengubah tipe data.** Pada tahap ini variabel - variabel fitur yang masih memiliki tipe data objek seperti, 'tgl', 'temp7', 'temp13', 'temp18', 'temp_avg', 'temp_24', 'temp_max', 'temp_min', 'CH', 'light_hour', 'light_per', 'press', 'humid7', 'humid13', 'humid18', 'humid_avg', 'humid_24', 'ws_abg', 'max_ws', dan 'dir' menjadi float berdasarkan jenis data yang ada di dalam variabel tersebut.
3. **Standarisasi**. Proses ini merubah isi dari data dalam variabel sehingga distribusinya akan memiliki nilai rata-rata 0 dan standar deviasi 1.

b. Pembagian data

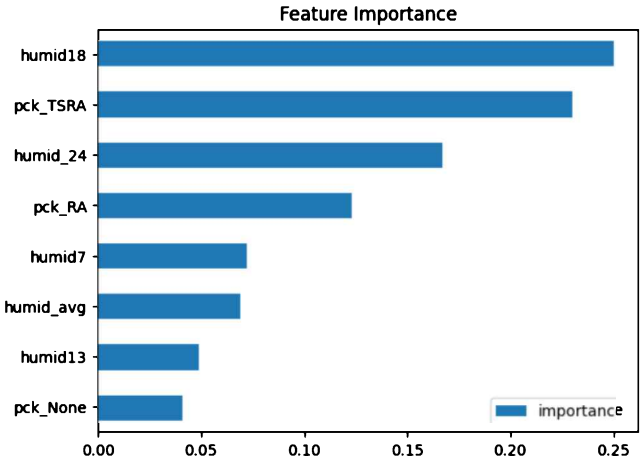
Setelah melalui tahapan pre-processing, proses berlanjut dengan melakukan pembagian data menjadi dua kelompok utama, yaitu data pelatihan dan data pengujian. Keputusan untuk membentuk kedua subset ini ditentukan berdasarkan tanggal pembagian, yang disebut 'split_date', yang telah ditetapkan pada tanggal 1 juni 2021. Melalui pembagian ini, kami menggunakan dua kelompok data yang memiliki peran khusus diantaranya data pelatihan yang digunakan untuk melatih model, sementara data pengujian digunakan untuk menguji sejauh mana model dapat memberikan prediksi yang akurat pada data yang tidak pernah dilihat sebelumnya.

c. Pembuatan Model

Setelah melakukan pre-processing dan merapikan seluruh data, langkah selanjutnya yaitu melakukan pembuatan model. Model prediksi akan dilakukan menggunakan XGBoost Regression untuk memprediksi curah hujan. Parameter yang digunakan pada XGBoost regression yaitu dengan menggunakan parameter n_estimators sebanyak 100 dan melakukan early stopping pada iterasi ke-50. Dari pelatihan yang sudah dilakukan, menghasilkan validation_rmse sebesar 0.83708 dan menghasilkan nilai prediksi sebesar 40,81%.

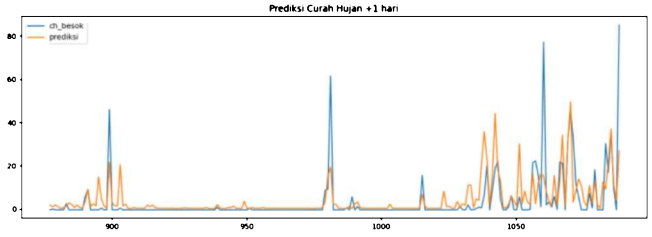
B. *Analisis Hasil Eksperimen*

Dari penelitian yang telah dilakukan, berikut ini adalah hasil yang sudah didapatkan yang diawali dari feature importance. Pada fitur importance dapat diketahui bahwa fitur yang terpenting yaitu humid18 yaitu suhu rata-rata pada jam 18.00 WIB. Berdasarkan Grafik pada gambar 3 terlihat bahwa yang paling menentukan hasil dari prediksi curah hujan ini yaitu adalah fitur yang berhubungan dengan humid yaitu rata-rata kelembapan.



Gambar 3 Feature Importance

Berdasarkan hasil dari prediksi curah hujan yang telah dilakukan, didapatkan model yang akan digunakan untuk memprediksi curah hujan untuk beberapa hari kedepan. Seperti terlihat pada gambar 4 yang menampilkan grafik dari perbandingan prediksi curah hujan, terlihat bahwa grafik prediksi sudah cukup baik dan hasilnya tidak beda jauh dari data pengetesan. Persentase akurasi yang didapatkan dari hasil prediksi yaitu sebesar 40.81%.



Gambar 4 Prediksi Curah Hujan +1 hari

Setelah itu hasil dari metric yang didapatkan dari penelitian ini seperti terlihat pada tabel 2. Pada metric tersebut terlihat nilai yang dihasilkan cukup besar dengan nilai metric yang didapatkan ada pada kisaran 80%, nilai ini cukup besar yang artinya data yang masih tidak sesuai prediksi sangat banyak.

TABEL 2 METRIC DAN HASIL

Metric	Hasil
RMSE	8.791625773374465
MASE	0.7978489030855982
MSE	77.29268373906216

V. KESIMPULAN

Dalam rangka mencapai tujuan prediksi curah hujan untuk tiga hari ke depan, penelitian berhasil mengembangkan alat prediksi cuaca yang menggunakan algoritma XGBoost. Eksplorasi data yang mendalam dilakukan untuk memahami pola dan karakteristik curah hujan, dengan fokus pada identifikasi faktor-faktor krusial seperti suhu, kelembapan, dan tekanan udara yang dapat mempengaruhi hasil prediksi. Keputusan untuk menggunakan XGBoost sebagai model prediksi didasarkan pada kemampuannya yang teruji untuk menangani data yang kompleks dan memberikan hasil yang akurat.

Berdasarkan penelitian yang telah dilakukan, didapatkan hasil bahwa metode XGBoost yang digunakan kurang optimal dalam memprediksi curah hujan, hal ini dikarenakan tingkat akurasi yang didapat dari metode XGBoost adalah 40% dengan MSE (Mean Squared Error) sebesar 77.3%.

Proses pelatihan model melibatkan optimasi parameter dengan tujuan meningkatkan akurasi prediksi serta menghindari fenomena overfitting dan underfitting. Hasil validasi menunjukkan bahwa alat prediksi yang dikembangkan memiliki tingkat akurasi yang tinggi, memberikan kepercayaan bahwa pendekatan menggunakan XGBoost dan fitur-fitur yang dipilih merupakan kombinasi yang efektif untuk prediksi cuaca, terutama dalam konteks curah hujan.

Implementasi alat prediksi ini diharapkan memberikan manfaat yang signifikan, terutama dalam sektor pertanian, di mana pemahaman yang akurat tentang cuaca dapat berkontribusi pada perencanaan tanam dan panen. Selain itu, alat ini memiliki potensi untuk mendukung manajemen bencana dengan memberikan informasi prediktif terkait cuaca ekstrem. Dalam konteks pengambilan keputusan infrastruktur, akurasi prediksi cuaca dapat menjadi faktor penentu untuk perencanaan dan keberlanjutan proyek-proyek tersebut.

Selain memberikan solusi praktis untuk prediksi cuaca, penelitian ini juga memberikan kontribusi terhadap pemahaman kita tentang pentingnya algoritma dalam menangani masalah ekonomi dan manajemen. Kesimpulan ini diperkuat oleh rekomendasi penelitian untuk terus menjelajahi metode pembelajaran mesin lainnya dan integrasi data cuaca yang lebih luas, menciptakan landasan untuk perkembangan lebih lanjut dalam bidang ini. Secara keseluruhan, hasil penelitian ini menggambarkan bahwa penggunaan algoritma, seperti XGBoost untuk memecahkan tantangan prediksi cuaca adalah langkah ilmiah yang efektif, dan pendekatan ini berpotensi menjadi tren utama dalam manajemen ekonomi di masa depan.

REFERENSI

[1] S. Amril Mutoi, Tukino, Sutan Faisal, Ahmad Fauzi and Ilman Kadori, Klasifikasi untuk Prediksi Cuaca Menggunakan Ensemble Learning. Jurnal Pengkajian dan Penerapan Teknik Informatika, Cikarang, vol. 13, no. 2, pp. 138–147, September. 2020.

[2] Mursianto, B. Amany, Isma'il Muhammad Falih, Muhammad Irfan, Tiara Sakinah and D. S. Prasvita, Perbandingan Metode Klasifikasi Random Forest dan XGBoost Serta Implementasi Teknik SMOTE pada Kasus Prediksi Hujan. SENAMIKA, Jakarta-Indonesia, vol. 2, no. 2, pp. 41–50, 2021.

[3] Supriyadi, E. (2021). Prediksi Parameter Cuaca Menggunakan Deep Learning Long-Short Term Memory (LSTM). *Jurnal Meteorologi dan Geofisika*, *21*(2), 55-67.

[4] Desmonda, D., Tursina, T., & Irwansyah, M. A. (2018). Prediksi besaran curah hujan menggunakan metode fuzzy time series. JUSTIN (Jurnal Sistem dan Teknologi Informasi), 6(4), 145-149.

[5] Latifah, A. N., Sidauruk, A., Sulistiyono, M., Satria, B., & Nurcholis, M. T. (2023). Prediksi Curah Hujan Menggunakan Algoritma Regresi Linear Berganda. *Jurnal ICT: Information Communication & Technology*, *23*(1), 39-44.

[6] G. I. Merdekawati and Ismail, "Prediksi Curah Hujan di Jakarta berbasis Algoritma Levenberg Marquardt," Jurnal Ilmiah Informatika Komputer, vol. 24, no. 2, pp. 116–128, Aug. 2019. doi:10.35760/ik.2019.v24i2.2366.

[7] S. Hidayatullah and A. Cherid, "Prediksi Temperatur Cuaca di Negara Norwegia Menggunakan Metode LSTM", simkom, vol. 8, no. 2, pp. 187-198, Aug. 2023.

[8] H. Prasetyo, I. Palupi, and B. Wahyudi, "Prediksi Menggunakan Model Fuzzy Time Series Studi Kasus Curah Hujan di Kabupaten Bandung," Jurnal Penelitian Informatika, vol. 1, no. 1, pp. 8–13, Sep. 2023, doi: https://doi.org/10.25124/logic.v1i1.6405. Available: https://journals.telkomuniversity.ac.id/logic/article/view/6405/2168. [Accessed: Nov. 10, 2023]

[9] M. Hasanah, S. Soim, and A. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir", JAIC, vol. 5, no. 2, pp. 103-108, Oct. 2021.

[10] Y. Zhang, "Stock price Prediction Method based on XGBoost algorithm," in Proceedings of the 2022 International Conference on Bigdata Blockchain and Economy Management (ICBBEM 2022), 2022, pp. 595–603. doi: 10.2991/978-94-6463-030-5_60

Akses dan Kinerja Jaringan *Hotspot* menggunakan *Voucher* Berbayar

Andreas Risky Ardian Kusuma
Departemen Teknik Elektro
Universitas Sanata Dharma
Sleman, Yogyakarta
andreasriskyardiankusuma@gmail.com

Damar Widjaja
Departemen Teknik Elektro
Universitas Sanata Dharma
Sleman, Yogyakarta
ORCID iD: 0000-0001-8598-5731

Abstrak— Penelitian ini bertujuan untuk mengetahui kinerja jaringan *hotspot* saat *voucher* diterapkan untuk membatasi akses data dan kecepatan data *internet* yang telah ditetapkan. Pengukuran *quality of service (QoS)* digunakan untuk mengetahui apakah *voucher* melewati batas dari kecepatan data *internet* yang ditetapkan. *QoS* yang diukur pada penelitian ini adalah *throughput, packet loss, delay, frame error rate, data error rate, data rate, bit error rate, signal strength, signal noise ratio (SNR).* Pengambilan data *QoS* menggunakan Wireshark dan Mikrotik Router OS dan dilakukan dua (2) kali siang dan malam. Penelitian melibatkan 1 *user*, 3 *user*, dan 5 *user* untuk mengakses *browsing, sosial media, dan streaming* selama 10 hari. Hasil pengujian jaringan menunjukkan bahwa *voucher* akan berhenti ketika mencapai batas ukuran data yang telah ditetapkan dan *voucher* tidak dapat melebihi batas kecepatan *upload* maupun *download* yang telah ditetapkan. Parameter kinerja *QoS* pada akses jaringan *hotspot* menggunakan *voucher* berbayar termasuk dalam kategori bagus menurut *TIPHON*.

Kata Kunci—voucher, QoS, hotspot, Mikrotik, Wireshark

I. PENDAHULUAN

Di era sekarang, perkembangan teknologi informasi yang sangat cepat dapat menciptakan hal-hal baru yang dapat membantu dalam kehidupan manusia [1]. Teknologi-teknologi ini dirancang agar dapat memudahkan pekerjaan manusia seperti mengirim *file*, mengirim surat elektronik (*e-mail*) dan masih banyak lagi. Hal ini menyebabkan penggunaan teknologi nirkabel sebagai alat berbagi internet sangat dibutuhkan. Penggunaan jaringan nirkabel yang sering digunakan adalah teknologi *Wireless Fidelity* (Wi-Fi). Wi-Fi sering dijumpai pada tempat umum, warnet, kantor, *coffee shop*, dan masih banyak lagi. Tak jarang ketika ingin akses internet tersebut pengguna diminta untuk memilih paket *voucher* yang sesuai dengan kebutuhan.

Dengan kebutuhan yang tinggi terhadap internet serta didukung dengan teknologi radio akhirnya tercipta teknologi *Wireless Local Area Network* (WLAN) [2]. Hotspot merupakan istilah yang sering dipakai untuk fasilitas WLAN yang tersedia pada tempat-tempat tertentu. Sementara itu, aplikasi yang banyak digunakan melalui *hotspot* salah satunya adalah internet. Dengan *hotspot*, *user* bisa berbagi koneksi internet tanpa kabel.

Mobilitas yang tinggi akan kebutuhan internet membuat penggunaan *hotspot* menjadi tuntutan [3]. Walaupun secara umum koneksi *wireless* masih belum bisa mengalahkan teknologi pendahulunnya (*wired*), namun peningkatan mobilitas yang luar biasa pada penggunaan teknologi *wireless* tersebut perlu dipertimbangkan, misalnya: pertemuan bisnis yang memerlukan koneksi internet dapat dilakukan tidak terbatas di ruangan kerja, tetapi dapat dilakukan di semua *public area.*

Hotspot adalah salah satu standar Wireless Networking tanpa kabel, namun hanya dengan menggunakan komponen yang sesuai dapat terkoneksi ke jaringan WLAN [4]. Pada dasarnya *hotspot* merupakan sistem yang memberikan fitur autentikasi pada *user* yang akan terhubung ke sebuah jaringan. *Username* dan *password* pada *login page* dibutuhkan untuk bisa mengakses jaringan.

Jaringan *hotspot* adalah jaringan dengan beberapa *access point* dalam suatu area atau blok dan dapat saling tersambung [5]. Jaringan *hotspot* memberdayakan pemakaian internet dengan fasilitas internet tersedia selama 24 jam sehari selama sebulan dan biaya yang dikeluarkan murah karena biaya operasional ditanggung bersama yang menggunakan akses jaringan *hotspot.*

Sistem *voucher* merupakan salah satu cara untuk memudahkan *client* dalam penyambungan jaringan internet. Di samping itu, *voucher* juga memudahkan penyedia internet dalam memberikan hak akses internet kepada pelanggan.

Penelitian [6] menyimpulkan mudahnya akses internet menggunakan media *wireless* dengan perangkat *mobile* pribadi. Penelitian [7] menyimpulkan bahwa *voucher* akan habis pada batas waktu. Jika masa berlaku *voucher* sudah habis, maka pengguna tidak akan dapat *login* kembali.

Jaringan *hotspot* harus dapat memberikan layanan yang mudah bagi pengguna maupun bagi penyedia layanan *hotspot.* Kemudahan dalam layanan bagi pengguna adalah kemudahan dalam penggunaan, sedangkan untuk penyedia adalah kemudahan dalam *data record.*

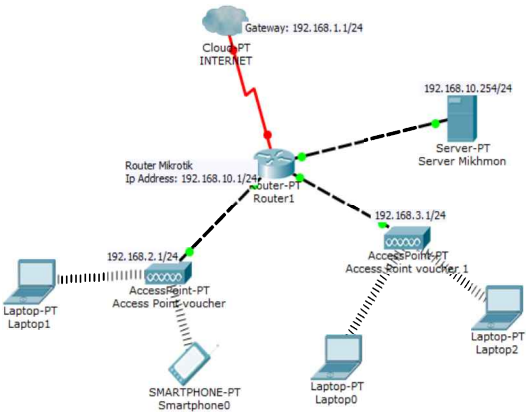
II. METODE PENELITIAN

A. *Perancangan Jaringan*

Gambar 1 merupakan rancangan jaringan yang digunakan untuk implementasi perangkat keras dalam penelitian ini.

Rancangan jaringan pada Gambar 1 memperlihatkan topologi yang digunakan pada penelitian ini. Pada perancangan perangkat keras ini, port *ethernet* 1 pada Router1 terhubung pada Internet dengan *Gateway*: 192.168.1.1/24. Port 2 ethernet Router1 terhubung pada *access point voucher* dengan IP *address* 192.168.2.1/24. Port 3 ethernet Router1 akan terhubung pada *access point voucher 1* dengan IP

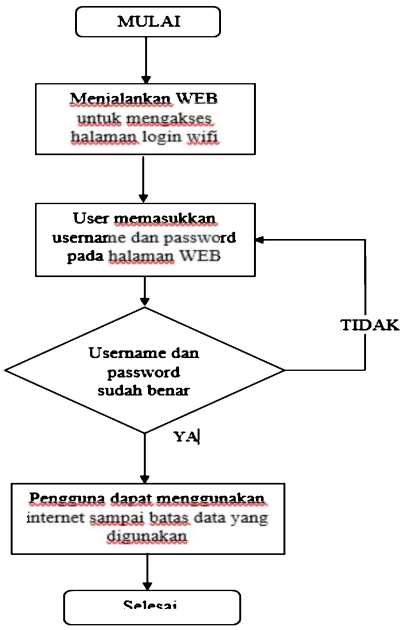
address 192.168.3.2/24. Port 4 ethernet *router* terhubung pada komputer untuk mengakses server Mikhmon dengan IP 192.168.10.254 dan Winbox.



Gambar 1. Rancangan jaringan.

B. *Flowchart*

Gambar 2 menjelaskan *flowchart* untuk proses *user login* ke jaringan internet. Pengguna membeli *voucher* dari penyedia layanan *hotspot* berbayar. Ketika pengguna mendapatkan *voucher*, pengguna masuk ke jaringan melalui *SSID* yang dipancarkan oleh *access point.*



Gambar 2. Flowchart user login.

Kemudian pengguna memasukkan *link* yang terdapat pada *voucher*. Pengguna memasukkan *username* dan *password* yang terdapat pada *voucher*. Jika pengguna salah memasukkan *username* dan *password*, maka pengguna akan mengulang lagi memasukkan *username* dan *password*. Ketika *username* dan *password* yang dimassukkan oleh pengguna benar, pengguna dapat menggunakan layanan *internet* sampai batas data yang ditetapkan.

III. HASIL DAN PEMBAHASAN

A. *Pengujian Voucher*

Pengujian *voucher* dilakuan untuk mengetahui apakah *voucher* yang telah dibuat berdasarkan masa berlaku dapat bekerja seperti yang diharapkan. Pengujian ini dilakukan dengan melakukan *test speed* kode *voucher* yang telah di *generate* oleh Mikhmon. Nilai yang diharapkan dalam penguian *voucher* adalah 1 *Mbps* untuk *download* dan *upload.*

Gambar 3 merupakan tampilan kode *voucher internet* 1 Gb yang akan digunakan dalam pengujian dengan kode *voucher* vnx948. Kode *voucher* tersebut selanjutnya digunakan pada halaman *login page* wifi. Gambar 4 merupakan halaman *login page* wifi setelah memasukkan kode *voucher*.

Gambar 3. *Voucher* Internet 1 GB.

Gambar 4. *User* berhasil memasukkan *voucher.*

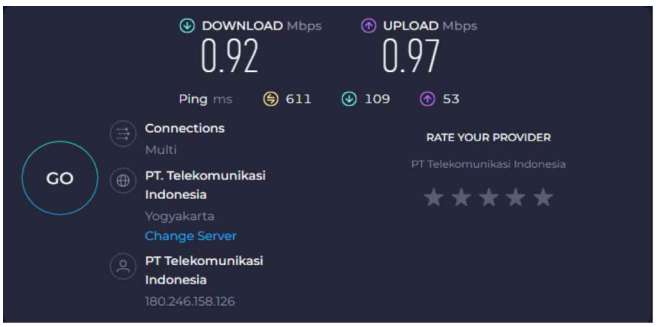
Gambar 4 juga menunjukkan *landing page* ketika *user* telah berhasil memasukkan kode *voucher* yang dimilki. Pada *landing page* tersebut terdapat informasi sisa kuota, alamat *IP address*, kecepatan *upload* dan *download* serta *mac address*

Gambar 5 menunjukkan saat *user* gagal memasukan *voucher*, karena *voucher* yang digunakan telah habis atau dipakai oleh *user* lain (digunakan pada perangkat lain). Keterangan kode *voucher/user* sedang aktif muncul saat *user* gagal masuk kedalam *hotspot*.

Gambar 5. *User* gagal masuk.

Gambar 6 menunjukkan saat paket telah mencapai batas kuotadan muncul pesan kode *voucher/user* sudah mencapai batas kuota. Gambar 7 merupakan tampilan hasil uji coba kecepatan *voucher* harian menggunakan speedtest.net. Hasil uji kecepatan yang didapat sebesar 0,92 Mbps download dan 0,95 Mbps upload. Kecepatan yang didapat sudah mendekati dengan kecepatan yang telah ditetapkan pada saat pembuatan user profile sebesar 1 Mbps untuk download dan upload.

Gambar 6. Paket telah mencapai batas.



Gambar 7. Hasil tes kecepatan internet *voucher*.

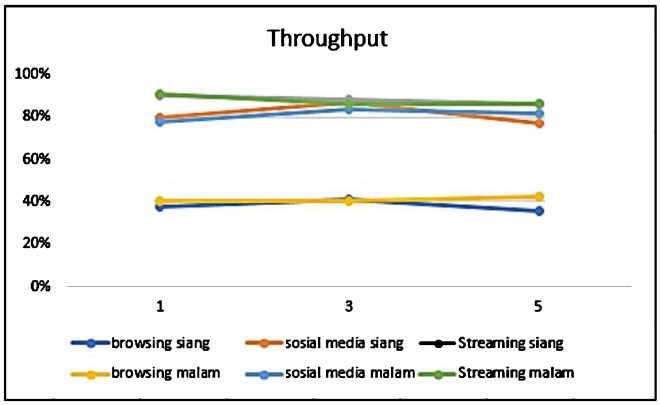
B. *Analisa Hasil Pengukuran QoS*

Pengukuran parameter QoS dilakukan selama 10 hari, 2 kali per hari di waktu siang dan malam. Pengukuran dilakukan saat *user* melakukan *web browsing*, akses media sosial, dan *video streaming*.

- *Throughput*

Gambar 8 menunjukkan prosentase *throughput* yang diukur saat 1 *user*, 3 *user*, dan 5 *user* aktif. *Throughput* saat *user* melakukan *web browsing* sekitar 40% dari bandwidth maksimum yang ditetapkan, baik pada siang maupun malam hari. *Throughput browsing* siang sedikit lebih rendah dari malam. Hal ini terjadi karena jumlah *user* yang mengakses Wifi yang sama di lokasi pengujian di luar jaringan *hotspot* yang diteliti lebih sedikit di malam hari.

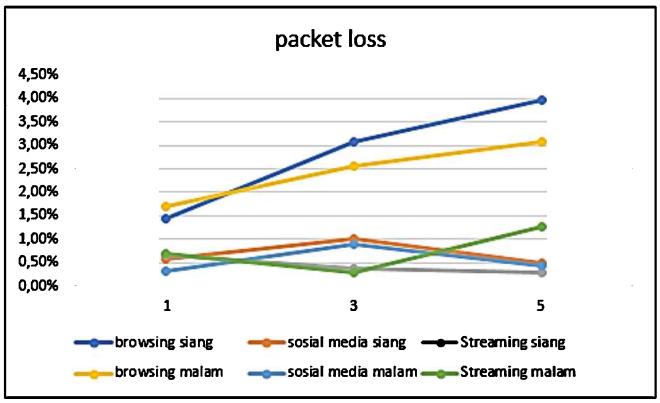
Throughput saat *user* mengakses media sosial dan melakukan *video streaming* cukup tinggi baik di waktu siang maupun malam dengan *throughput* tertinggi adalah *video streaming* di malam hari. Secara keseluruhan, perbedaan jumlah *user* tidak mempengaruhi *throughput* secara signifikan.



Gambar 8. Pengukuran *throughput* untuk 1, 3, dan 5 *user*.

- *Packet Loss*

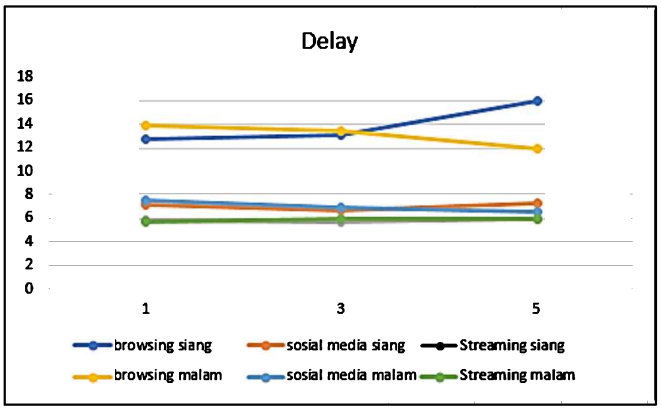
Gambar 9 menunjukkan prosentase *packet loss* yang diukur saat 1 *user*, 3 *user*, dan 5 *user* aktif. Terlihat bahwa semakin banyak *user* yang aktif, prosentase *packet loss* semakin tinggi. *Packet loss* di siang hari lebih tinggi dari pada di malam hari. *Packet loss* yang tinggi karena siang hari adalah jam sibuk sehingga jumlah *user* lebih banyak dari pada malam. *Packet loss* saat *browsing* lebih tinggi dibandingkan saat akses media sosial maupun *video streaming*.



Gambar 9. Pengukuran *packet loss* untuk 1, 3, dan 5 *user*.

- *Delay*

Gambar 10 menunjukkan *delay* (ms) yang diukur saat 1 *user*, 3 *user*, dan 5 *user* aktif.



Gambar 10. Pengukuran *delay* (ms) untuk 1, 3, dan 5 *user*.

Delay saat *web browsing* cukup tinggi dibanding saat akses media sosial dan *video streaming*. Hal ini terjadi karena adanya jeda waktu yang lama antara waktu data dikirim dan waktu data diterima sehingga menyebabkan *buffering* dalam mengakses *web site*. Jumlah *user* dalam jaringan hotspot yang dalam penelitian ini juga tidak banyak berpengaruh terhadap *delay*, kecuali pada saat *user* melakukan *web browsing* di malam hari.

- *Bit Error Rate (BER)*

Pada penelitian ini hasil dari pengukuran BER adalah 0 untuk semua pengujian baik saat 1, 3, dan 5 *user* sedang aktif. Hal ini terjadi karena jaringan *hotspot* yang dirancang telah diimplementasikan di area yang sempit, sehingga *bandwidth* dan *signal strength* terjaga di level maksimum sesuai spesifikasi *router*. *Signal strength* maksimum menyebabkan nilai *Signal to Noise Ratio* (SNR) menjadi sangat besar dan *bit transmission* sangat handal.

IV. KESIMPULAN

Berdasarkan hasil percobaan dan implementasi jaringan *hotspot* menggunakan *voucher* berbayar, kesimpulan yang dapat diambil adalah semua perangkat dalam jaringan yang dibuat dapat bekerja dengan baik sesuai dengan perancangan.

Voucher dapat digunakan seperti pada rancangan. *Voucher* 1 GB akan berhenti ketika mencapai batas data 1 GB. *Voucher* tidak dapat melebihi batas kecepatan *upload* 1 MB maupun *download* 1 MB.

Semua parameter QoS yang diukur menunjukkan bahwa jaringan hotspot yang dirancang bekerja dengan kinerja yang baik. Jumlah user tidak begitu berpengaruh terhadap kinerja jaringan secara keseluruhan.

REFERENSI

[1] Hardiyanti, "Implementasi Jaringan Hotspot Voucher di Warnet Artha dengan Menggunakan Mikrotik, STIMK Musirawas", Jusikom, hal.15-21, Desember 2016

[2] M.S. Ansor, E.P. Nugroho, dan S. Siregar, "Membangun Jaringan Komputer "Hotspot" Management Bandwidth serta Pemasangan Proxy Server di Rumah Makan Sinar Mas Banyumas", Politeknik Tellkom Bandung, 2020.

[3] T.A. Fitria dan A. Prihanto, "Implementasi Generate Voucher Hotspot dengan Batasan Waktu (Time Based) dan Kuota (Quota Based) Menggunakan User Manager di Mikrotik", Universitas Negeri Surabaya, Vol.8, No 02 hal.18-24, 2018.

[4] J.A. Falaq, R. Tulloh, dan M. Iqbal, "Implementasi Jaringan Hotspot Berbayar Berbasis Voucher menggunakan Platform Google Cloud", Universitas Telkom Bandung, Vol.7, No.4 hal.861-876, 2021

[5] C. Kurniawan, "Perancangan Jaringan Hotspot dengan Sistem Voucher menggunakan Mikrotik pada Jaringan Rt/Rw Net", Universitas Muhammadiyah Surakarta, 2014.

[6] N. Hidayat, "Perancangan dan Implementasi Jaringan Hotspot Untuk Akses Internet di SMK Asta Mitra Purwodadi", Skripsi, Jurusan Informatika, Universitas Muhammadiyah Surakarta, 2016.

[7] W.F. Pattipeilohy, "Analisis dan Perancangan User Manager pada Mikrotik Router dengan Sistem Pembelian Kredit Voucher", Jurnal SISFOKOM, Volume 05, No.01, Hal.64-69, 2016.

Analisis Kinerja Metode *Support Vector Regression* (SVR) Dalam Memprediksi Harga Rumah di Depok

Aris Prayogo
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
arisprayoga0806@gmail.com

Helga Raditia Ade
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
helgaraditia0@gmail.com

Aldi Tri Wijaya
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
aldi.5200411242.@student.ac.id

Panji Al Muqsith Prasetyo
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
panjiallazusa@gmail.com

Alfito Herdiansyah
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
hrdnsh6@gmail.com

Alfaeni Syafa Safira
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
alfaenisyafasafira@gmail.com

Abstrak—Rumah merupakan salah satu kebutuhan primer yang dibutuhkan oleh seseorang karena dengan adanya rumah dapat memberikan perlindungan, keamanan, kenyamanan dan juga berfungsi sebagai tempat istirahat. Maka dari itu untuk mendapatkan rumah yang sesuai dengan kriteria tertentu dan juga harga yang lebih murah, sehingga dibutuhkan sistem yang mampu memprediksi harga rumah berdasarkan kriteria tertentu untuk masa mendatang. Sehingga tujuan penelitian ini akan dibangun sistem prediksi harga rumah dengan metode *support vector regression*. Dengan pengambilan data dari kaggle.com khususnya harga rumah di daerah depok dengan variabel yang menentukan harga yaitu kamar_tidur, luas_tanah, luas_bangunan, kamar_mandi, daerah, harga_jm, harga dengan total data 195 record. Agar mendapatkan hasil model yang terbaik penelitian ini melakukan beberapa eksperimen terhadap splitting data yaitu 90%-10%, 80%-20%, 70%-30%, dari eksperimen tersebut untuk hasil terbaik dengan menggunakan splitting data 90%-10% dengan nilai MAE = 378799999,64, *Mean Squared Error* (MSE) = 2,837e+17, dan *Mean Root Squared Error* (RMSE) = 532642094,44.

Kata Kunci—prediksi, SVR, Rumah, Depok

I. PENDAHULUAN

Rumah adalah salah satu komponen kehidupan yang penting karena mampu memberikan perlindungan, keamanan, dan kenyamanan, juga berfungsi sebagai tempat untuk beristirahat. Secara definisi, rumah adalah bangunan atau tempat yang digunakan sebagai tempat perlindungan dari cuaca dan ancaman hewan liar, serta berfungsi sebagai tempat berkumpul, melepaskan kelelahan dan penat setelah beraktivitas di luar, dan sebagai tempat berlangsungnya berbagai kegiatan keluarga [1]. Rumah seringkali dijadikan sebagai simbol pencapaian, penerimaan sosial, dan juga sebagai indikator pertumbuhan jumlah penduduk di perkotaan.

Di dalam dunia properti, penentuan harga rumah merupakan sebuah tugas yang sangat penting dan kompleks. Harga rumah dipengaruhi oleh berbagai faktor, termasuk lokasi, ukuran, kondisi fisik, fasilitas sekitar, dan banyak variabel lainnya. Seperti dalam penelitian yang berjudul "Penentuan Harga Rumah Menggunakan Metode *Tversky* dalam *Reasoning* Berbasis Kasus," disimpulkan bahwa harga rumah tidak hanya bergantung pada faktor lokasi, melainkan juga dipengaruhi oleh faktor-faktor lain, termasuk ketersediaan akses transportasi menuju rumah [2]. Oleh karena itu, pengembangan sistem prediksi harga rumah menjadi sangat relevan dan berharga bagi para pemilik rumah, pembeli, dan agen *real estate*. Sebagian besar pembeli rumah ingin memahami tren harga dan melihat bagaimana mereka dapat memanfaatkan perubahan tersebut. Inilah mengapa penggunaan teknologi dan algoritma pembelajaran mesin seperti *Support Vector Regression* (SVR) dalam prediksi harga rumah semakin penting.

Salah satu algoritma yang telah digunakan secara luas dalam memprediksi harga rumah adalah *Support Vector Regression* (SVR). SVR merupakan pengembangan dari *Support Vector Machine* (SVM) yang melibatkan atribut tambahan untuk menghasilkan prediksi sebagaimana yang dilakukan dalam analisis statistik [3]. SVR adalah algoritma pembelajaran mesin yang mampu menyesuaikan dengan berbagai jenis data dan memiliki tingkat kesalahan yang minim [4]. Algoritma ini digunakan untuk mengidentifikasi pola dan hubungan antara berbagai atribut rumah dengan harga penjualannya. Dengan bantuan SVR, kita dapat mengembangkan model yang mampu memprediksi harga rumah berdasarkan data historis dan atribut rumah yang tersedia.

Pendekatan SVR untuk prediksi harga rumah memiliki sejumlah keunggulan, termasuk kemampuan untuk menangani data non-linear, kemampuan untuk menangani data berdimensi tinggi, dan tingkat akurasi yang tinggi [5]. Oleh karena itu, penggunaan SVR dalam sistem prediksi harga rumah menjadi semakin populer dan diterapkan dalam berbagai skenario, termasuk oleh agen *real estate*, investor, dan pengembang properti.

Pada penelitian ini, kita akan menjelaskan konsep dasar dari *Support Vector Regression* (SVR) dan bagaimana metode ini dapat diterapkan dalam pengembangan sistem prediksi harga rumah. Dengan demikian, pembaca akan mendapatkan wawasan yang lebih baik tentang bagaimana SVR dapat digunakan untuk meningkatkan akurasi dalam memprediksi harga rumah, dan dapat memberikan nilai tambah dalam pengambilan keputusan di pasar properti yang dinamis.

II. KAJIAN LITERATUR

Beberapa hasil penelitian yang pernah dilakukan oleh peneliti sebelumnya yang memiliki bidang dan metode yang sama dengan penelitian yang akan dilakukan, penelitian sebagai berikut.

Penelitian oleh Hasanah dkk [6] dengan judul Analisis Prediksi Harga Rumah di Jabodetabek Menggunakan Multiple Linear Regression. Data yang digunakan diperoleh dari website *kaggle.com*, data yang digunakan merupakan data daftar harga rumah di daerah Jabodetabek. Metode yang digunakan yaitu *Multiple Linear Regression*. Hasil yang didapatkan berdasarkan analisis dari penelitian tersebut menghasilkan tingkat akurasi 85% dengan tingkat error MAE sebesar 375428900.51 dan tingkat error RMSE sebesar 515738165.50.

Penelitian oleh Aeni dkk [7] prediksi jumlah penumpang dan penambahan gerbong kereta api menggunakan metode SVR. Data yang digunakan pada penelitian ini diambil dari PT KAI DAOP 2 Jawa Barat mengenai data jumlah penumpang kereta api Argo Parahyangan periode 2019 dan memiliki 2 kelas yaitu kelas ekonomi premium dan kelas eksekutif. Metode pada penelitian ini menggunakan metode SVR.

Penelitian Oleh Rahmi dan Helma [8] dengan judul Portofolio Optimal dengan Mempertimbangkan Prediksi Return Menggunakan Metode SVR dengan data diperoleh dari Website Bursa Efek Indonesia yang di unduh pada yahoo.finance.co.id dari periode Januari 2022 – Maret 2023. Hasil akhir diperoleh nilai *Squared Error* sebesar 0,9427.

Penelitian Oleh Tillah [9] dengan judul Perbandingan Prediksi Obat Berdasarkan Pemakaian Menggunakan Algoritma *Single Moving* dan SVR. Data diperoleh dari bagian farmasi puskesmas dengan jangka waktu obat Januari 2020 – Juni 2023. Metode yang digunakan pada penelitian ini menggunakan 2 metode yaitu *Single Moving Average* dan *Support Vector Regression.* Hasil akhir menghasilkan nilai mean absolute percentage error sebesar 7,35 % dan 9,52 % sangat baik yaitu metode SVR sedangkan untuk metode SMA mendapatkan hasil 4.10% dan 4,29% yang baik.

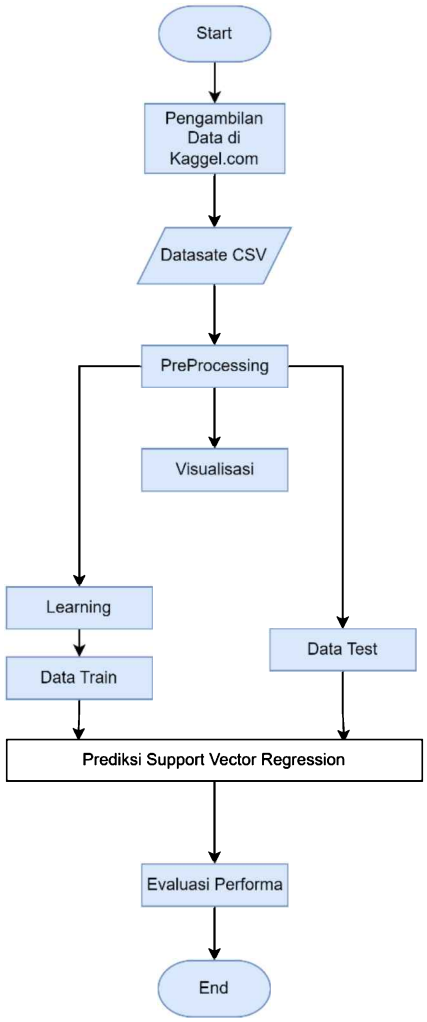
Penelitian oleh Fitri [10] dengan judul Analisis Perbandingan Metode *Regresi Linear*, *Random Forest Regression* dan *Gradient Boosted Trees Regression Method* Untuk Prediksi Harga Rumah. Data diperoleh dari website *kaggle.com* untuk wilayah Jakarta dan daerah Tebet. Metode yang digunakan pada penelitian ini menggunakan tiga metode yaitu metode r*egresi linear, random forest regression* dan *gradient boosted trees regression method*. Hasil tertinggi dari penelitian ini yaitu 81% untuk metode *random forest regression.*

III. METODOLOGI

Perancangan sistem disusun dengan metode penelitian untuk membuat sistem analisis sentimen sesuai tahap-tahapan yang ada pada proses membangun sistem, berikut *flowchart* metodologi penelitian pada Gambar 1.

1. Pengambilan Dataset

   Pada penelitian membutuhkan data untuk pembuatan prediksi harga rumah khususnya di Depok. Untuk memperoleh data tersebut proses pengambilan dilakukan pada website *kaggle.com,* dimana website tersebut merupakan tempat menyediakan ribuan dataset gratis untuk keperluan penelitian. Untuk data yang digunakan pada penelitian memprediksi harga rumah yaitu dataset harga rumah di depok pada tahun 2023.



Gambar 1 Metode Penelitian

2. *Preprocessing* Data

   Preprocessing data yaitu tahapan yang sangat berpengaruh besar terhadap performa model sehingga diharapkan dalam *preprocessing* ini mampu agar dapat menghasilkan nilai *output* yang baik. Tahapan *preprocessing* yang dilakukan yaitu *cleaning* data seperti menghilangkan atribut yang tidak berpengaruh signifikan terhadap performa dalam pembuatan model kemudian menghilangkan data-data yang rusak atau *missing value* agar data lebih lebih optimal.

3. Visualisasi

   Proses Visualisasi merupakan tahapan untuk melakukan representatif grafis pada data, ini memungkinkan penulis untuk memahami pola [11], hubungan dan trend pada data dengan cara yang lebih menarik seperti melalui penggunaan grafik, diagram, maupun visualisasi jenis lainnya dengan visualisasi data ini diharapkan mampu menyajikan data dalam bentuk visualisasi agar mempermudah

dalam memahami isi data dan dapat mempermudah pengambilan keputusan yang lebih baik.

4. Penerapan *Support Vector Regression*

SVR merupakan salah satu pengembangan dari metode SVM yang berfokus pada regresi dengan tujuan memprediksi nilai kontinu atau riil [12]. Prinsip utama yang mendasari SVR adalah upaya meminimalkan batas atas kesalahan generalisasi. Dalam konstruksi model *regresi* SVR, digunakan sekelompok fungsi linier berdimensi tinggi [13]. Salah satu kelebihan utama SVR adalah kemampuannya untuk mengatasi overfitting, sehingga dapat menghasilkan model yang memiliki tingkat kesalahan yang kecil dan akurasi yang tinggi [14]. Hal penting lainnya adalah bahwa kompleksitas komputasi SVR tidak dipengaruhi oleh dimensi ruang *input*, dan model ini unggul dalam hal kemampuan generalisasi dengan tingkat akurasi prediksi yang tinggi [15].

5. Evaluasi Performa

Evaluasi performa pada regresi ini untuk mengukur performa pada penelitian kasus ini dengan metode yang digunakan berupa MSE, MAE dan RMSE, untuk melihat bagaimana untuk melihat keakuratan pada sistem ini.

a. MSE

MSE adalah metrik yang mengukur rata-rata dari kuadrat kesalahan antara prediksi dan nilai aktual. MSE memberi bobot lebih pada kesalahan besar karena kuadratnya. Berikut Rumus *MSE.*

$$MSE = (\Sigma(yi - \hat{y}i)^2) / n$$

MSE =*Mean Squared Error.*
yi = nilai aktual.
$\hat{y}i$ = nilai prediksi.
Σ = simbol untuk menjumlahkan kuadrat kesalahan untuk semua data poin.
n = jumlah total data point.

b. MAE

MAE adalah matrik yang mengukur kesalahan rata-rata antara nilai prediksi dan nilai aktual. MAE adalah pilihan yang baik ketika kita ingin mendapatkan pemahaman yang sederhana tentang sejauh mana model *regresi* kita mendekati nilai aktual. MAE sangat tergantung dengan variasi data, semakin besar variasi datanya maka semakin besar nilai MAE [16].
Rumus MAE:

$$MAE = (\Sigma|yi - \hat{y}i|) / n$$

MAE = *Mean Absolute Error.*
yi = nilai aktual.
ŷi = nilai prediksi.
Σ = simbol untuk menjumlahkan kesalahan absolut untuk semua data poin.
n = jumlah total data point.

c. RMSE

RMSE adalah akar kuadrat dari MSE. Ini mengukur sejauh mana prediksi berbeda dari nilai aktual dalam satuan yang sama dengan data asli. RMSE digunakan untuk klimatologi, peramalan dan analisis *regresi* untuk pembuktian eksperimen [17].

Rumus RMSE:

$$RMSE = \sqrt{MSE}$$

RMSE = Root Mean Squared Error.
MSE = Mean Squared Error.

Dari penggunaan tiga matrik tersebut memiliki tujuan kenapa penelitian ini menggunakan ketiga matrik tersebut untuk model evaluasinya. Dari evaluasi model tersebut yakni MAE, MSE dan RMSE digunakan karena tiga matrik tersebut dalam mengukur nilai *error* atau kesalahan dari sudut pandang yang berbeda sehingga pada penelitian ini menggunakan ketiga matrik tersebut agar mendapatkan gambaran yang lebih lengkap tentang sejauh mana pemahaman tentang kinerja model.

IV. EKSPERIMEN DAN ANALISIS

Bab ini akan menjelaskan semua yang telah dijabarkan sebelumnya pada bab metode penelitian yang terdiri dari, pengumpulan dataset, *preprocessing*, visualisasi data, pemodelan SVR, dan evaluasi model.

1. Pengumpulan dataset

Data yang digunakan dalam penelitian ini diambil secara langsung dari platform Kaggle dengan nama file rumah123-dpk. File ini memiliki tipe file berupa excel (xlsx) dengan jumlah data sebanyak 195 baris dan 7 kolom. Dibawah ini merupakan output dari membaca file dengan *Python* dan memanfaatkan *library pandas.*

	k_tidur	l_tanah	l_bangunan	k_mandi	daerah	harga_jm	harga
0	3	120	110	2	Tapos	1,2 M	1200000000
1	2	72	30	1	Tapos	785 J	785000000
2	3	63	62	2	Cimanggis	770 J	770000000
3	3	90	70	2	Cimanggis	897 J	897000000
4	2	105	68	1	Sukatani	850 J	850000000

Gambar 2. Dataset

Dapat dilihat pada Gambar 2 di atas, dataset ini memiliki 7 kolom dengan deskripsi seperti berikut:

a. k_tidur: Jumlah kamar tidur dalam 1 rumah
b. k_mandi: Jumlah kamar mandi dalam 1 rumah
c. l_tanah: Luas tanah suatu rumah
d. l_bangunan: Luas bangunan suatu rumah
e. daerah: Letak rumah dibangun
f. harga_jm: harga dalam satuan rupiah
g. harga: harga rumah dalam tipe data integer

2. *Pre-processing*

Data yang baru saja dikumpulkan menjadi satu dataset harus melalui tahap *preprocessing*. Ini bertujuan untuk menghindari masalah dalam Pembangunan model

machine learning. Tahapan yang akan dilakukan yaitu penghapusan fitur dan *missing value*.

a. Penghapusan fitur

Jika dataset ditampilkan berupa dataset.info(), *Python* akan mengeluarkan *output* berupa deskripsi tipe data dalam kolom tersebut, nama kolom, dan non-null content.

```
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   k_tidur     195 non-null    int64
 1   l_tanah     195 non-null    int64
 2   l_bangunan  195 non-null    int64
 3   k_mandi     195 non-null    int64
 4   daerah      195 non-null    object
 5   harga_jm    195 non-null    object
 6   harga       195 non-null    int64
dtypes: int64(5), object(2)
```

Gambar 3. Tipe data dalam dataset

Dapat dilihat pada Gambar 3 di atas, terdapat tipe data object yang tidak diperlukan. Fitur ini dapat dihapus tanpa mempengaruhi hasil analisis.

b. Missing Value

Pada tahap ini, *missing value* dilakukan untuk mengecek apakah suatu kolom memiliki baris yang kosong atau tidak ada nilainya. Untuk pengecekan *missing value* dapat menggunakan fungsi dataset.isnull().sum().

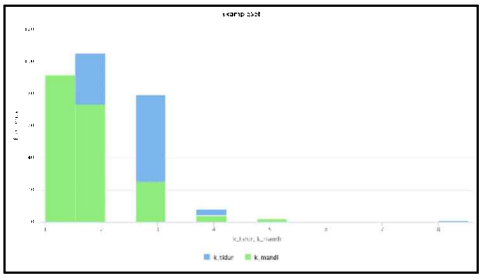
```
k_tidur       0
l_tanah       0
l_bangunan    0
k_mandi       0
harga         0
dtype: int64
```

Gambar 4. Pengecekan Missing Value

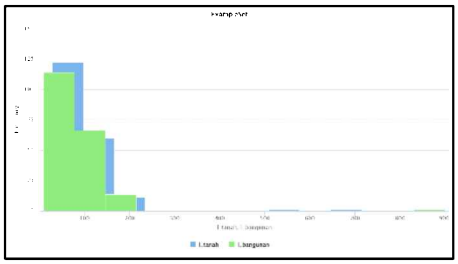
Gambar 4 di atas menunjukkan bahwa hasil *missing value* secara keseluruhan baris dalam dataset ini tidak ada yang hilang atau tidak ada yang kosong di tiap atributnya.

3. Visualisasi Data

Dalam tahap analisis data statistik, diperlukan adanya visualisasi data berupa total tiap baris dalam suatu atribut, dalam visuaslisasi ini ada beberapa yang akan digambarkan. Yaitu untuk atribut l_tanah dan l_bangunan dengan jenis visualisasi bar *chart* untuk melihat distribusi data.

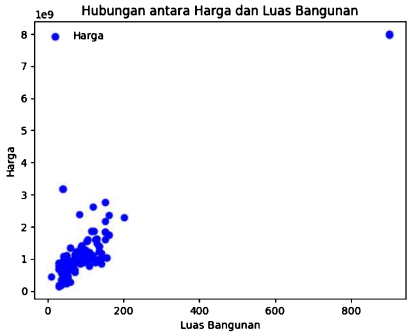


Gambar 5. Persebaran Attribut k_tidur dan k_mandi



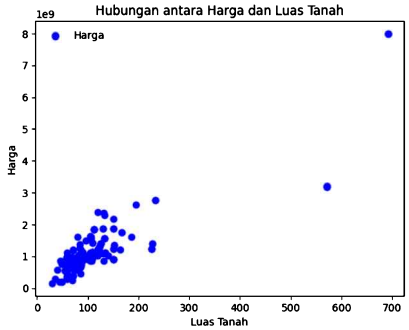
Gambar 6. Persebaran Attribut l_tanah dan l_bangunan

Pada Gambar 5 dan 6 menggambarkan bagaimana frekuensi persebaran attribute pada data k_mandi, k_tidur, l_tanah, l_bangunan yang divisualisasikan menggunkan bar *chart histogram*.



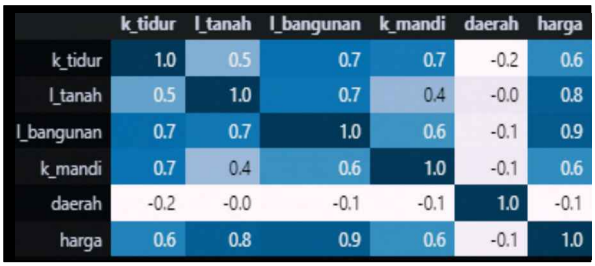
Gambar 7. Korelasi antara Harga dengan Luas Bangunan

Pada Gambar 7 diatas dapat dilihat bagaimana besarnya luas bangunan dapat mempengaruhi harga yang dijual oleh pemilik.



Gambar 8. Korelasi antara Harga dengan Luas tanah

Pada Gambar 8 menunjukkan harga yang signifikan jika berbicara luas tanah. Kedua kolom ini dipilih karena jika melihat langsung data di lapangan, luas tanah dan luas bangunan menjadi penentu utama mengapa sebuah rumah memiliki harga yang tinggi ataupun rendah. Sementara kamar tidur dan kamar mandi tidak menjadikan harga menjadi signifikan. Dapat dibuktikan dengan dengan visualisasi korelasi berikut pada Gambar 9.



	k_tidur	l_tanah	l_bangunan	k_mandi	daerah	harga
k_tidur	1.0	0.5	0.7	0.7	-0.2	0.6
l_tanah	0.5	1.0	0.7	0.4	-0.0	0.8
l_bangunan	0.7	0.7	1.0	0.6	-0.1	0.9
k_mandi	0.7	0.4	0.6	1.0	-0.1	0.6
daerah	-0.2	-0.0	-0.1	-0.1	1.0	-0.1
harga	0.6	0.8	0.9	0.6	-0.1	1.0

Gambar 9. Korelasi antar variabel

Pada Gambar 9 dapat dilihat tabel korelasi antar variabel satu dengan variabel yang lainnya menunjukan bahwa yang paling berpengaruh terhadap harga rumah yaitu l_bangunan dan yang kedua l_tanah kemudian korelasi yang tidak berpengaruh k_mandi dan k_tidur.

4. Pemodelan SVR

SVR adalah bentuk regresi khusus yang menggunakan konsep SVM untuk menangani data yang mungkin tidak memiliki pola linear. Dalam teknik regresi, algoritma tidak memerlukan fitur target seperti pada teknik klasifikasi. Sebagai gantinya, teknik regresi mengandalkan data berupa angka yang tujuan umumnya adalah memprediksi harga. Dalam kasus ini, parameter yang akan digunakan antara lain:

a. test_size = 0.1

b. random_state = 1

c. Kernel SVR linear

d. C = 1.5

e. epsilon = 0.5

Dengan menggunakan test_*size* 0.1 atau 10 % untuk data uji maka data *train* yang digunakan sebesar 90% sehingga model *regresi* memiliki lebih banyak data untuk belajar pola dan hubungan dalam dataset harga rumah. Hal ini dapat membantu model untuk lebih baik dalam menangkap variabilitas dari data. Alasan lainnya adalah berkaitan dengan evaluasi model yang lebih akurat terhadap evaluasi model. Ini membantu memastikan bahwa model tidak hanya mempelajari data pelatihan dengan baik tetapi juga mampu menggeneralisasi pada data yang belum pernah dilihat sebelumnya. Selain itu, rasio 90:10 dapat membantu mengurangi variabilitas hasil karena evaluasi model didasarkan pada set data uji yang lebih besar, yang dapat menghasilkan estimasi kinerja model yang lebih stabil. Ini penting dikarenakan dengan mengurangi variabilitas hasil, kita dapat mendapatkan pemahaman yang lebih baik tentang seberapa baik model dapat berkinerja secara umum. Variabilitas yang tinggi dapat menyulitkan untuk menentukan sejauh mana hasil yang diamati dalam merefleksikan kemampuan sejati model.

5. Evaluasi model

Evaluasi model ini menggunakan MAE, MSE, dan RMSE untuk mendapatkan output yang dihasilkan oleh algoritma SVR.

TABEL 1. HASIL EVALUASI MODEL

Test size	MAE	MSE	MRSE
0.1	378799999.64403474	2.8370760076431197e+17	532642094.43519574
0.2	519192309.9760648	1.5242375421914742e+18	123460015 4.783513
0.3	487737291.8021359	1.1498875120569894e+18	107232808 0.4198823

Tabel 1 di atas menunjukkan hasil pembangunan model yang baik, dilakukan tiga ukuran evaluasi yang digunakan, hasil terbaik dapat diidentifikasi dengan mengamati nilai-nilai yang lebih rendah, yang menunjukkan kesalahan prediksi yang lebih kecil. Berdasarkan hasil evaluasi model regresi yang saya peroleh, dapat disimpulkan bahwa kinerja terbaik terdapat pada ukuran uji sebesar 0,1. Pada ukuran uji ini, nilai *Mean Absolute Error* (MAE) = 378799999,64, nilai MAE didapatkan dari mengukur rata-rata dari selisih *absolut* antara prediksi model dan nilai sebenarnya. *Mean Squared Error* (MSE) = 2,837e+17, nilai MSE memberikan bobot lebih besar pada kesalahan yang besar karena mengkuadratkan selisihnya. *Mean Root Squared Error* (MRSE) = 532642094,44, nilai MRSE di dapatkan dari akar kuadrat dari MSE, memberikan ukuran kesalahan yang lebih mudah diinterpretasikan. Hasil ini menunjukkan bahwa model regresi memberikan prediksi yang lebih akurat dan mendekati nilai sebenarnya pada ukuran uji 0,1. Oleh karena itu, dalam konteks penelitian ini, dapat dianggap bahwa ukuran uji 0,1 menghasilkan kinerja model yang optimal berdasarkan matrik evaluasi yang digunakan.

KESIMPULAN

Dari hasil penerapan hasil prediksi harga rumah di Depok, dapat disimpulkan bahwa tahapan awal dilakukan proses pengolahan data dengan menghilangkan atribute-atribute yang tidak akan berpengaruh terhadap harga rumah. Kemudian dilakukan proses prediksi menggunakan metode SVR dengan hasil yang cukup baik dalam keakuratan prediksi, kemudian dari proses visualisasi data dapat dilihat bahwa korelasi antara luas tanah dan luas bangunan cukup berpengaruh terhadap harga rumah tersebut. Penelitian ini menghasilkan tingkat keakuratan berdasarkan eksperimen yang dilakukan dengan penggunaan pembagian data 90%-10% menjadi eksperimen yang paling bagus model nya yaitu dengan nilai MAE = 378799999,64, *Mean Squared Error* (MSE) = 2,837e+17, dan *Mean Root Squared Error* (MRSE) = 532642094,44.

REFERENSI

[1] E. F. Rahayuningtyas, F. N. Rahayu, and Y. Azhar, "Prediksi Harga Rumah Menggunakan General Regression Neural Network," *J. Inform.*, vol. 8, no. 1, pp. 59–66, 2021, doi: 10.31294/ji.v8i1.9036.

[2] Hendra, Tursina, and R. D. Nyoto, "Case Base Reasoning Penentuan Harga Rumah d engan Menggunakan Metode Tversky (Studi Kasus : Kota Pontianak)," *J. Sist. dan Teknol. Inf.*, vol. 5, no. 2, pp. 75–79, 2017.

[3] F. N. S. Pradana and F. S. Papilaya, "Analisa Prediksi Harga Emas Dengan Kemungkinan Terjadinya Resesi Menggunakan Metode SVR," *SINTECH (Science Inf. Technol. J.*, vol. 6, no. 1, pp. 37–46, 2023, doi: 10.31598/sintechjournal.v6i1.1329.

[4] Z. Rais, "Analisis Support Vector Regression (Svr) Dengan Kernel Radial Basis Function (Rbf) Untuk Memprediksi Laju Inflasi Di Indonesia," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 4, no. 1, pp. 30–38, 2022, doi: 10.35580/variansiunm13.

[5] R. Maharesi, F. Teknologi, I. Jurusan, T. Informatika, and U. Gunadarma, "Penggunaan Support Vector Regression (Svr) Pada Prediksi Return Saham Syariah BEI," *Proceeding PESAT*, vol. 5, pp. 8–9, 2013, [Online]. Available: https://ejournal.gunadarma.ac.id/index.php/pesat/article/view/1180/1041

[6] I. Maula, L. U. Hasanah, and A. Tholib, "Analisis Prediksi Harga Rumah Di Jabodetabek Menggunakan Multiple Linear Regression," *J. Inform. Kaputama*, vol. 7, no. 2, pp. 216–224, 2023, doi: 10.59697/jik.v7i2.135.

[7] U. N. Aeni, A. L. Prasati, and M. Kallista, "Grafik Jumlah Penumpang 2018," vol. 7, no. 2, pp. 4919–4926, 2020.

[8] A. Rahmi, "Portofolio Optimal Dengan Mempertimbangkan Prediksi Return Menggunakan Metode Support Vector Regression (SVR) Program Studi Matematika , Universitas Negeri Padang," vol. 7, no. March, pp. 23745–23753, 2023.

[9] S. Nurfan, H. Tillah, A. Nazir, I. Iskandar, E. Budianita, and I. Afrianty, "Perbandingan Prediksi Obat Berdasarkan Pemakaian Menggunakan Algoritma Single Moving Average dan Support Vector Regression," vol. 7, pp. 1860–1868, 2023, doi: 10.30865/mib.v7i4.6859.

[10] E. Fitri, "Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah," *J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, pp. 58–64, 2023, doi: 10.52158/jacost.v4i1.491.

[11] D. Hartama, "Analisa Visualisasi Data Akademik Menggunakan Tableau Big Data," *Jurasik (Jurnal Ris. Sist. Inf. dan Tek. Inform.*, vol. 3, no. 3, p. 46, 2018, doi: 10.30645/jurasik.v3i0.65.

[12] R. K. D. Olivia Bonita, Lailil Muflikhah and Program, "Prediksi Harga Batu Bara Menggunakan Support Vector Regression (SVR)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 12, pp. 6603–6609, 2018.

[13] A. A. Ghullam, T. Joko, W. Adi, and D. Ph, *SUPPORT VECTOR REGRESSION GROWTH MODEL*. 2020.

[14] R. P. Furi, Jondri, and D. Saepudin, "Peramalan Financial Time Series Menggunakan Independent Component Analysis dan Support Vector Regression (Studi Kasus: IHSG dan Jll)," *S1. Telkom Univ.*, vol. 2, no. 2, pp. 3608–3618, 2015.

[15] R. E. Caraka, H. Yasin, and A. W. Basyiruddin, "Peramalan Crude Palm Oil (CPO) Menggunakan Support Vector Regression Kernel Radial Basis," *J. Mat.*, vol. 7, no. 1, p. 43, 2017, doi: 10.24843/jmat.2017.v07.i01.p81.

[16] F. O. Saputra *et al.*, "JITE (Journal of Informatics and Telecommunication Engineering)," vol. 6, no. January, pp. 538–547, 2023.

[17] M. L. Subiyanto, Y. Amanda, and M. N. Fachrian, "Peramalan Kasus Harian Monkeypox Dunia Dengan Pendekatan Support Vector Regression," pp. 27–36, 2022.

Algoritma K-means untuk Segmentasi Data Nasabah Pemohon Kredit

Axel Frans Silalahi
Departemen Informatika
Universitas Sanata Dharma
Yogyakarta, Indonesia
axelfrans93@gmail.com

Hari Suparwito
Departemen Informatika
Universitas Sanata Dharma
Yogyakarta, Indonesia
shirsj@jesuits.net

Abstrak—Kredit merupakan pendapatan terbesar bank. Akan tetapi bank harus selektif dalam menentukan nasabah yang layak untuk mendapatkan kredit atau tidak. Kemampuan setiap nasabah untuk membayar tagihan kredit sangat beragam, dan tidak dapat dipungkiri bahwa banyak nasabah yang berada dalam bahaya kredit macet, baik karena keterlambatan pembayaran atau bahkan karena tidak dapat membayar tagihan kreditnya. Tujuan studi ini adalah untuk melakukan *clustering* data nasabah pemohon kredit melalui pendekatan *machine learning* menggunakan algoritma *K-means*. Setiap kelompok data nasabah akan dianalisis sehingga dapat disimpulkan kelompok nasabah yang lebih berisiko terkena kredit macet atau kurang berisiko. Melalui metode *K-means* telah didapatkan 2 *cluster* yang paling optimal atau 2 kelompok nasabah dengan hasil validasi menggunakan *Silhouette Coefficient* dengan nilai 0.221. Hasil akhir segmentasi terhadap *cluster* yang terbentuk menunjukkan bahwa *cluster* 0 sebagai *cluster* pertama merupakan kelompok nasabah dengan jumlah 26148 yang lebih berisiko membuat kredit macet. Sedangkan *cluster* 1 sebagai kelompok nasabah yang kedua dengan jumlah data sebanyak 69857 merupakan kelompok nasabah yang kurang berisiko.

Kata kunci—Clustering, Data Mining, K-means, Kredit, Silhouette Coefficient, Segmentasi Nasabah.

I. PENDAHULUAN

Kredit identik artinya dengan pinjam dan meminjamkan uang, sehingga dapat juga dikatakan bahwa kredit mengacu pada penyerahan uang atau tagihan, yang mensyaratkan persetujuan perjanjian pinjaman antara bank dengan pihak lawan yang meminta kepada peminjam untuk membayar kembali. Wajib Utang yang harus dibayar setelah waktu tertentu dan dapat dilunasi. Perusahaan keuangan atau bank sebagai pemberi pinjaman/kreditur tentu harus mengambil risiko, seperti risiko kehilangan uang karena peminjam/debitur gagal melunasi utangnya [1]. Setiap pemberian kredit didasarkan pada kepercayaan bahwa nasabah akan dapat membayar kredit yang telah diberikan, tetapi kepercayaan itu harus dibuktikan dengan data yang telah disetujui. Ini berarti bahwa pemberi pinjaman dapat melihat atau menilai karakteristik dan kemampuan nasabah yang akan diberi pinjaman kredit melalui data yang diberikan oleh nasabah. Perusahaan pasti akan sulit untuk mengidentifikasi semua risiko yang terjadi jika mereka memberikan pinjaman kredit karena banyaknya data nasabah yang memiliki kredit.

Tujuan dari studi ini adalah menemukan pola dari kelompok pelanggan yang lebih berisiko dan kurang berisiko saat diberi pinjaman/kredit. Pendekatan *Machine Learning* menggunakan algoritma *K-means* dipergunakan untuk menemukan pola dari kelompok data tertentu dengan tujuan mengurangi risiko kerugian yang mungkin terjadi.

Dalam penelitian sebelumnya [2], analisis segmentasi terhadap data nasabah yang memiliki kredit pinjaman juga dilakukan dengan menggunakan algoritma *K-means* yang dibandingkan dengan metode *Agglomerative Clustering*, GMM dan DBSCAN. Hasil penelitian menunjukkan bahwa metode *K-means* mendapat nilai terbaik dengan menggunakan nilai *silhouette coefficient* sebesar 0.207 dengan hasil *clustering* sebanyak 3 *cluster*. Segmentasi data dengan 3 *cluster* adalah *customer* dengan penggunaan kartu kredit yang moderat sebanyak 1696 data, *customer* dengan penggunaan kartu kredit paling sedikit sebanyak 1341 data dan *customer* dengan lebih banyak menggunakan kartu kredit dan melakukan pembelian produk lebih sering sebanyak 824 data.

Metode *K-means* juga diterapkan untuk melakukan segmentasi nasabah bank XYZ [3]. Hasil *cluster*ing menunjukkan bahwa 3 *cluster* yang terbentuk yang terdiri dari *cluster* 0, *cluster* 1, *cluster* 2 dengan ciri segmentasi *cluster* 0 adalah nasabah dengan rata-rata jumlah kredit lebih rendah, durasi pendek, dan pelanggan usia paling tua. *Cluster* 1 merupakan *cluster* nasabah dengan rata-rata jumlah kredit tinggi, durasi panjang, pelanggan usia pertengahan. *Cluster* 2 merupakan *cluster* nasabah dengan rata-rata jumlah kredit lebih rendah, durasi pendek, pelanggan usia muda.

Dalam penelitian yang dilakukan oleh Huda Ahsina menggunakan sumber data nasabah kredit bank pada *German Credit Data* oleh Prof. Hofmann. Hasil *cluster* yang diperoleh didasarkan pada metode *elbow*, yaitu 4 *cluster* adalah *cluster* terbaik dari semua kemungkinan *cluster* yang terdiri dari *cluster* 1 dengan jumlah 286 nasabah dengan persentase 28,6%, *cluster* 2 dengan jumlah 130 nasabah dengan persentase 13%, *cluster* 3 dengan jumlah terbesar yaitu 542 nasabah dengan persentase 54,2%, dan *cluster* 4 memiliki jumlah sebaran terendah yaitu 42 nasabah dengan persentase 4,2% [4].

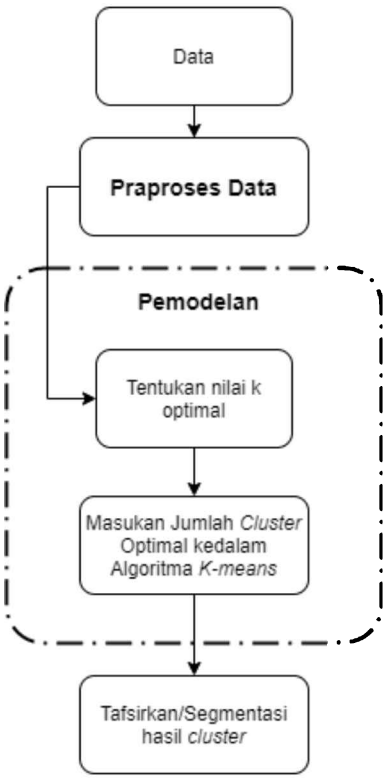
II. METODOLOGI PENELITIAN

A. Tahapan Penelitian

Secara umum, proses atau tahapan penelitian dapat digambarkan seperti diagram alir pada Gambar 1. Ada 4 tahap yaitu: pengumpulan data, praproses data, modeling data dan terakhir adalah tafsiran/segmentasi hasil *cluster*ing.

B. Data

Data yang digunakan dalam penelitian ini merupakan data mentah berjumlah 100000 data dan terdiri dari 74 atribut.



Gambar 1. Tahapan Penelitian

C. *Praproses Data*

Dalam tahapan ini dilakukan pembersihan data (*data cleaning*), seleksi data (*data selection*) dan transformasi data (*data transformation*). Masing-masing proses dijelaskan seperti berikut ini:

- ***Pembersihan data***

Pembersihan data merupakan langkah pertama dalam praproses data yang berfungsi untuk menghilangkan inkonsistensi atau gangguan, seperti data yang hilang (*missing values*), dan data terduplikasi. Dua faktor tersebut dapat mempengaruhi proses penambangan data yang akan menghasilkan data yang tidak tepat dan hasil yang buruk [5]. Melalui proses pembersihan data maka *noise data*, data yang tidak konsisten dan mengandung *missing values* dapat dihilangkan menggunakan *library Pandas pada python*.

Proses pembersihan data dilakukan dengan cara menghapus atribut-atribut yang mengandung *missing values* lebih dari 50% akan dihapus karena data tidak memiliki informasi yang cukup untuk memastikan apakah data tersebut berguna untuk proses *clustering*.

- ***Seleksi Data***

Proses pemilihan atribut dilakukan untuk menghilangkan atribut data yang yang memiliki kemiripan nilai yang sama. Seleksi data akan menggunakan *heatmap* dari *library seaborn* pada *python*.

- ***Transformasi Data***

Proses transformasi seperti data normalisasi diperlukan untuk pengolahan data. *Data transformation* dilakukan untuk mengubah bentuk dan format data [6]. Ada 2 tahap transformasi yang akan digunakan pada penelitian ini adalah data Normalisasi dan *data Labelling*. Proses *Labelling* data dilakukan dengan *label encoder*. *Label encoder* akan mengubah data tipe kategorial menjadi data tipe numerik. Setelah proses pelabelan berhasil dilakukan maka selanjutnya adalah proses normalisasi data, proses ini diperlukan untuk berbagai anomali data dan ketidakkonsistenan input data sebelumnya. *Min-max scalling* adalah metode nomalisasi data. Teknik ini dapat digunakan untuk mengatasi perbedaan atau rentang data yang signifikan dalam dataset. Metode ini bekerja dengan mengubah nilai atribut data yang ada menjadi nilai dengan skala (0,1) tanpa mengubah informasi atribut data.

Proses transformasi yang dilakukan adalah sebagai berikut:

1. Atribut *emp_length*: proses transformasi yang dilakukan pada atribut ini adalah dengan mengubah tahun pada pada atribut ini menjadi angka tetap pada data.
2. Atribut *grade*: terdapat 7 *grade* yang akan diubah menjadi angka 0 sampai 6.
3. Atribut *home_ownership*: terdapat 3 jenis nilai data yang akan diubah menjadi angka 0 sampai 2.
4. Atribut *verification_status*: terdapat 3 jenis nilai data yang akan diubah menjadi angka 0 sampai 2.
5. Atribut *purpose*: terdapat 13 jenis label data pada atribut ini dan akan diubah menjadi angka dari 0 sampai 12.
6. Atribut *initial_list_status*: terdapat 2 jenis label data pada atribut ini yang akan diubah menjadi angka 0 dan 1.
7. Setelah proses pelabelan 1-6 selesai kemudian proses terakhir dari transformasi data adalah tahap Standarisasi atau normalilasi menggunakan perhitungan *MinMaxScaler* dari *library sklearn* pada *python*.

D. *Pemodelan K-means*

Setelah tahap praproses selesai, dilanjutkan dengan tahap pemodelan *clustering* menggunakan metode *K-means clustering*. *K-means* adalah salah satu algoritma *unsupervised learning* paling sederhana tapi akurat untuk memecahkan masalah klasterisasi atau pengelompokan data. Prosedurnya mengikuti cara sederhana dan mudah untuk mengklasifikasikan kumpulan data yang diberikan melalui sejumlah *cluster* tertentu (k *cluster*) [7].

Pemodelan *K-means* dilakukan dengan langkah seperti berikut ini [8]:

1. Proses input atau memasukkan data
2. Memasukkan jumlah K *cluster*.
3. Alokasikan data ke dalam *cluster* secara random.
4. Hitung *centroid*/rata-rata dari data yang ada di masing-masing *cluster*.

Untuk menghitung jarak suatu objek data dengan *centroid* diperlukan rumus yang dinamakan dengan *Euclidean Distance* yaitu sebagai berikut [9]:

$$d(i,j) = \sqrt{(X_{1i} - X_{ij})^2 + (X_{2i} - X_{2j})^2 + \ldots + (X_{kj} - X_{kj})^2} \quad (1)$$

Keterangan:

$d(i,j)$ = Jarak data ke *i* ke pusat *cluster* j
X_{ki} = Data ke-*i* pada atribut data ke-*j*
X_{kj} = Titik pusat ke-*j* pada atribut ke-*k*

5. Alokasikan masing-masing data ke *centroid*/rata-rata terdekat

6. Lakukan kembali langkah 3, apabila masih ada data yang berpindah *cluster* atau apabila perubahan nilai *centroid*.

E. Evaluasi Hasil Clustering

Proses pemilihan k terbaik dilakukan dengan menggunakan metode *silhouette score*. Metode *silhouette* merupakan gabungan dari metode separasi dan kohesi yang digunakan untuk melihat kualitas dan kekuatan *cluster,* seberapa baik atau buruknya suatu obyek ditempatkan dalam suatu *cluster* [10]. Pada tahap evaluasi, metode ini digunakan untuk menguji kualitas *cluster* yang telah didapatkan dan selanjutnya dilakukan keputusan untuk menentukan apakah nilai tersebut dapat digunakan atau tidak dari *data mining* tersebut.

Koefisien *silhouette* yang lebih tinggi adalah *cluster* yang lebih baik. Dalam penelitian ini, evaluasi akan dilihat menggunakan perbandingan nilai dari *Silhouette Coefficient* untuk setiap nilai dari k = 2 hingga k = 9.

F. Segmentasi Cluster

Segmentasi *Cluster* merupakan kegiatan mengelompokkan data menjadi beberapa kelompok berbeda berdasarkan karakteristik yang ditemukan dari karakteristik data yang dipakai. Hasil dari analisis segmentasi ini akan sangat berguna bagi pelaku bisnis seperti perusahaan bank dalam menentukan kebijakan berdasarkan karakter tiap-tiap segmen pelanggan.

III. HASIL DAN PEMBAHASAN

A. Data

Data yang digunakan pada studi ini adalah data kredit dengan proses baca data dilakukan menggunakan fungsi *read_csv* pada bahasa pemrograman *Python*. Data yang terbaca berjumlah 74 atribut dengan jumlah 100000 baris data.

B. Praproses data

- **Pembersihan data**

Akan dilakukan eksplorasi data sebelum memilih apakah data-data yang dipilih layak di hapus (*drop*) atau tidak, dan terdapat banyak kolom atribut yang memiliki nilai kosong atau *missing value* yang lebih dari setengah bahkan mencapai 90% jumlah data dari 100000 data yang tersedia. Akan dilakukan hapus (*drop*) kolom yang mengadung lebih dari 50% data kosong atau *missing value.*

Tabel 1 berikut merupakan atribut-atribut yang akan di hapus berdasarkan jumlah *missing value*.

TABEL 1. *MISSING VALUE*

Atribut	Total Missing Value	Persentase (%)
inq_last_12m	1	100
total_bal_il	1	100
dti_joint	1	100
annual_inc_joint	1	100
tot_coll_amt	1	100
application_type	1	100
tot_cur_bal	1	100
open_acc_6m	1	100
open_il_6m	1	100
open_il_12m	1	100
open_il_24m	1	100
collections_12_mths_ex_med	1	100
mths_since_rcnt_il	1	100
max_bal_bc	1	100
open_rv_24m	1	100
total_cu_tl	1	100
inq_fi	1	100
total_rev_hi_lim	1	100
all_util	1	100
verification_status_joint	1	100
open_rv_12m	1	100
il_util	1	100
policy_code	0.95231	95.231
sub_grade	0.95231	95.231
acc_now_delinq	0.90415	90.415
earliest_cr_line	0.89405	89.405
mths_since_last_record	0.89149	89.149
mths_since_last_major_derog	0.87691	87.691
last_credit_pull_d	0.84607	84.607
emp_title	0.74252	74.252
Title	0.72502	72.502
zip_code	0.69874	69.874
next_pymnt_d	0.64361	64.361
mths_since_last_delinq	0.58078	58.078
desc	0.52956	52.956

Atribut dengan jumlah data yang hilang diatas 50% akan dihapus dan tidak digunakan dalam penelitian ini.

- **Seleksi data**

Proses seleksi dilakukan dengan cara menghapus (*drop*) atribut *loan_amnt, funded_amnt, funded_amnt_inv* karena memiliki nilai kolerasi yang sama seperti tabel 2 dibawah ini.

TABEL 2. *COLLERATION CHECK* UNTUK HAPUS ATRIBUT

	loan_amnt	*funded_amnt*	*funded_amnt_inv*
loan_amnt	1	0.99	0.97
funded_amnt	0.99	1	0.98
funded_amnt_inv	0.97	0.98	1

Dari tabel diatas atribut dengan nilai kolerasi yang hampir mirip dengan nilai kemiripan diatas 0.97 maka dapat disimpulkan bahwa ketiga data ini merupakan data yang hampir sama dan hanya akan diambil salah satu dari 3 atribut data tersebut yaitu atribut *loan_amnt* saja.

- **Transformasi Data**

Proses transformasi data terdiri dari 2 yaitu transformasi menggunakan pelabelan data menggunakan *LabelEncoder* dan normalisasi dengan *MinMax Scaler* dan

TABEL 3. CUPLIKAN DATA HASIL PELABELAN

grade	home_ ownership	verification _status	loan_ status	purpose	initial_ list_ status
1	4	2	5	1	0
2	4	1	0	0	0
2	4	0	5	11	0
2	4	1	5	9	0
1	4	1	1	9	0
0	4	1	5	13	0
2	4	0	1	2	0

Tabel 3 diatas merupakan cuplikan data hasil pelabelan menggunakan *LabelEncoder*. Setiap atribut yang bertipe kategorikal akan diubah kedalam bentuk numerik.

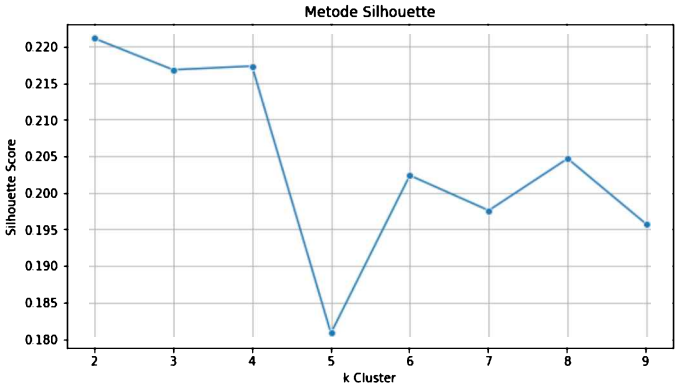
TABEL 4. CUPLIKAN DATA HASIL NORMALISASI

grade	home_ owner ship	verification _status	loan_ status	purpose	initial_ list_ status
0.166	1.0	1.0	0.625	0.076	0.0
0.333	1.0	0.5	0.0	0.0	0.0
0.333	1.0	0.0	0.625	0.846	0.0
0.333	1.0	0.5	0.625	0.692	0.0
0.166	1.0	0.5	0.125	0.692	0.0
0.0	1.0	0.5	0.625	1.0	0.0
0.333	1.0	0.0	0.125	0.153	0.0

Tabel 4 diatas merupakan cuplikan data hasil normalisasi terhadap data yang telah dilabel menggunakan *MinMaxScaler*. Rentang nilai yang jauh pada data akan diperkecil dengan rentang nilai dari 0-1.

C. *Implementasi K-means*

Pemodelan menggunakan *K-means Clustering* akan dibangun dengan kelas *K-means* dari *library scikit-learn (sklearn.cluster)*. Setelah melalui tahap praproses data kemudian dilakukan pemilihan jumlah k yang tepat menggunakan metode *silhouette*. Gambar 2 merupakan implementasi algoritma untuk pemilihan k yang tepat :



Gambar 2. Kurva *Sillhouette Score*

Gambar 2 menunjukkan bahwa *silhouette score* untuk masing masing jumlah *cluster* dari k = 2 hingga k = 9, semakin tinggi score yang didapatkan maka itulah jumlah *cluster* (K) yang terbaik, dilihat bahwa jumlah *cluster* terbaik adalah sebanyak 2 *cluster* dengan *silhouette* score tertinggi.

TABEL 5. Perbandingan nilai *Sillhouette Score*

Nilai *k*	**Nilai Silhouette score**
2	0.221
3	0.216
4	0.217
5	0.180
6	0.202
7	0.197
8	0.204
9	0.195

Dari hasil perhitungan terlihat bahwa nilai k = 2 (jumlah *cluster* = 2) menunjukkan nilai *silhouette* tertinggi atau mendekati 1 yaitu 0.221.

D. *Segmentasi Cluster*

Proses segmentasi merupakan proses penamaan *cluster* atau menentukan komposisi dari *cluster* yang telah terbentuk melalui analisis atribut atau atribut dari data hasil *cluster*. Akan dilakukan perankingan atribut untuk menentukan atribut apa saja yang dapat digunakan sebagai dasar segmentasi atau *profilling cluster*. Proses perankingan untuk mendapatkan 6 atribut terpenting menggunakan aplikasi *Orange*. Hasil perangkingan atribut didasarkan pada nilai *information gain* yang diperoleh. 6 atribut rangking teratas adalah *term, int_rate, loan_amnt, grade, installment, emp_length.* Tabel 6 menunjukkan hasil nilai *information gain*

TABEL 6. HASIL RANKING DATA

Atribut Data	Information Gain
Term	0.839
Int_rate	0.138
Loan_amnt	0.130
Grade	0.101
Installment	0.028
Emp_length	0.013

Selanjutnya akan dilakukan proses segmentasi dengan mencari tahu makna atau hubungan keenam atribut dan pengambilan keputusan atau menyimpulkan kelompok nasabah yang lebih berisiko atau kurang berisiko. Dari hasil analisis hubungan 6 atribut terpilih didapatkan kesimpulan dari 2 kelompok nasabah (2 *cluster* nasabah) sebagai berikut:

***Cluster* 0**

- Merupakan kumpulan nasabah yang didominasi melakukan kredit dengan nilai kredit pinjaman lebih besar.
- *Cluster* dengan total kredit yang lebih rendah dibanding dengan *cluster* 1.
- Merupakan kumpulan nasabah dengan pengalaman kerja yang belum lama.
- Merupakan kumpulan nasabah dengan durasi pengembalian kredit selama 60 bulan.
- Merupakan *cluster* nasabah yang memiliki kredit dengan suku bunga yang rendah.
- Merupakan *cluster* nasabah dengan total cicilan rendah.

***Cluster* 1**

- Merupakan kumpulan nasabah yang didominasi melakukan kredit dengan nilai kredit pinjaman lebih kecil dibanding *cluster* 0.
- Merupakan kumpulan nasabah dengan pengalaman kerja yang lebih lama.
- Merupakan kumpulan nasabah dengan durasi pengembalian kredit selama 30 bulan atau lebih singkat dibanding *cluster* 0.
- Merupakan kumpulan nasabah yang memliki kredit dengan suku bunga yang tinggi.
- Merupakan *cluster* nasabah dengan total cicilan tinggi.

Total jumlah kredit yang tinggi merupakan salah satu alasan mengapa kelompok tersebut menjadi lebih berisiko. Tetapi jumlah total kredit bukan faktor utama karena besarnya eksposur akan mempengaruhi kualitas kredit.

Dengan kata lain, semakin tinggi hutang/kredit maka semakin tinggi pula eksposur kreditnya atau potensi kreditnya. Kualitas atau potensi berisikonya akan lebih tinggi jika dihubungkan dengan bunga kredit yang diberikan. Bunga kredit yang tinggi akan membuat semakin berisiko seorang nasabah mengalami kredit macet.

Suku bunga rendah yang ditetapkan kepada kelompok nasabah *cluster* 0 dikarenakan jumlah kredit yang tinggi dan lama pengembalian kredit. Semakin banyak seseorang melakukan kredit atau berutang maka pihak bank akan memberikan bunga yang rendah tetapi karena jangka waktu pengembalian yang lama membuat total bunga yang akan dikenakan pada nasabah akan semakin tinggi pula sehingga, melihat pengaruh jumlah bunga tinggi yang diberikan kepada nasabah akan membuat nasabah tersebut kesulitan membayar kredit yang telah dipinjam karena selain harus membayar wajib hutangnya nasabah juga dipaksa harus membayar kesepakatan bunga yang telah ditetapkan. Menurut Berampu [11] menyimpulkan bahwa suku bunga kredit berpengaruh signifikan terhadap *Return on Asset* dan jurnal penelitian itu juga menyebutkan bahwa suku bunga tinggi akan memengaruhi kemampuan nasabah untuk mengembalikan kredit dimana suku bunga yang tinggi terindikasi akan menyebabkan kredit bermasalah. Hal inilah yang membuat mengapa bunga kredit yang tinggi pada *cluster* 0 membuat *cluster* ini menjadi lebih berisiko sedangkan *cluster* 1 kurang berisiko.

Jika dilihat dari lama pengalaman kerja, kecenderungan nasabah semakin lama pengalaman kerja semakin sedikit yang memohon kredit. Pengalaman kerja termasuk sebagai faktor kualitas sumber daya manusia [13]. Hasil uji hipotesis pada variabel Kualitas Sumber Daya Manusia terhadap risiko kredit mempunyai nilai signifikan sebesar $0.000 < 0.05$. Hasil tersebut menyimpulkan bahwa variabel Kualitas Sumber Daya Manusia berpengaruh signifikan terhadap risiko kredit. Hal ini ditunjukkan dengan hubungan yang positif antara kualitas suber daya manusia terhadap risiko kredit yang dimana adanya hubungan positif antara kualitas sumber daya manusia terhadap risiko kredit. Sehingga dapat disimpulkan bahwa kualitas kerja seseorang akan memengaruhi kemampuan membayar kredit. Semakin lama seseorang bekerja semakin tinggi kualitas nasabah tersebut dan semakin rendah risiko kredit yang diberikan terhadap bank. Berdasarkan kesimpulan tersebutlah yang membuat mengapa pengalaman kerja pada *cluster* 0 membuat *cluster* ini menjadi lebih berisiko sedangkan *cluster* 1 kurang berisiko.

Penelitian yang dilakukan oleh Olyvia menunjukkan bahwa dari delapan 8 faktor penyebab kredit bermasalah yang diberi nama Faktor Pilihan, Faktor Internal bank, Faktor Internal debitor, Faktor Tingkat keberhasilan, Faktor Manajemen diri, Faktor Kewajiban, Faktor Eksternal dan Karakter debitor dan yang paling dominan menjadi penyebab kredit macet adalah Faktor Pilihan dengan indikatornya yaitu Rentang waktu pembayaran kredit [12]. Faktor pilihan merupakan faktor yang terdiri dari variabel jangka waktu pelunasan *(term)*, suku bunga dan jumlah kredit. Semakin lama durasi pembayaran nya maka semakin berisiko nasabah tersebut tidak mampu membayar kredit atau berisiko bermasalah karena kelonggaran waktu yang diberikan tersebut terlalu lama selama 60 bulan. Sehingga *cluster* 0 merupakan kelompok nasabah yang lebih berisiko.

Berdasarkan penjelasan analisis atribut dari kedua *cluster* dapat disimpulkan bahwa *Cluster* 0 merupakan *cluster* dengan kelompok nasabah yang lebih berisiko memberikan potensi kredit macet. Sedangkan *cluster* 1 merupakan kelompok nasabah yang kurang berisiko.

IV. KESIMPULAN

Hasil penelitian pengelompokan nasabah yang melakukan kredit menggunakan Algoritme *K-means* dan dilakukan segmentasi terhadap 2 *cluster*, diperoleh kesimpulan bahwa algoritme *K-means* mampu dan berhasil melakukan pengelompokan data nasabah menjadi 2 bagian *cluster*. Melalui grafik dan nilai score menggunakan nilai *silhouette coefficient* yang telah berhasil dilakukan menunjukkan nilai k terbaik pada *cluster* 2 dengan nilai 0.221. Setelah terbentuknya 2 *cluster* data, telah dilakukan analisis segmentasi dan hasil analisis untuk segmentasi terhadap 2 *cluster* yang terbentuk adalah *cluster* 0 merupakan kelompok nasabah dengan kriteria kredit yang lebih berisiko dan *cluster* 1 merupakan *cluster* dengan kriteria kurang berisiko.

REFERENSI

[1] Rini Syahril Fauziah and N. H. K. Fadhilah, "The Impact of Credit Risk on The Profitability With Characteristics Bank as Control Variables," *JAK*, vol. 9, no. 2, pp. 145–158, Jul. 2022, doi: 10.30656/jak.v9i2.4346.

[2] F. Defina, S. Alhamdani, A. A. Dianti, and Y. Azhar, "Segmentasi Pelanggan Berdasarkan Perilaku Penggunaan Kartu Kredit Menggunakan Metode *K-m*eans Clustering," MEI, 2021. [Online]. Available: https://www.kaggle.com/arjunbhasin2013/ccdata.

[3] M. Rizky Wijaya and G. Satriyo Wibowo, "Prosiding Seminar Nasional Universitas Ma Chung Customer Segmentation berdasarkan Usia, Jumlah Kredit dan Lama Kredit Nasabah di Bank XYZ menggunakan Model *K-m*eans Clustering," vol. 2021, pp. 101–116, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

[4] N. Huda Ahsina, F. Fatimah, F. Rachmawati, U. Ibn Khaldun Bogor JlKH Sholeh Iskandar Km, and K. Bogor, "Analisis Segmentasi Pelanggan Bank Berdasarkan Pengambilan Kredit Dengan Menggunakan Metode *K-M*eans Clustering," 2022.

[5] M. Bilal, G. Ali, M. W. Iqbal, M. Anwar, M. S. A. Malik, and R. A. Kadir, "Auto-Prep: Efficient and Automated Data Preprocessing Pipeline," *IEEE Access*, vol. 10. pp. 107764–107784, 2022, doi: 10.1109/ACCESS.2022.3198662.

[6] A. P. Joshi and B. V. Patel, "Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process," *Oriental journal of computer science and technology*, vol. 13, no. 0203, pp. 78–81, Jan. 2021, doi: 10.13005/ojcst13.0203.03.

[7] S. S. Nagari and L. Inayati, "IMPLEMENTATION OF CLUSTERING USING *K-m*EANS METHOD TO DETERMINE NUTRITIONAL STATUS," *Jurnal Biometrika dan Kependudukan*, vol. 9, no. 1, p. 62, Jun. 2020, doi: 10.20473/jbk.v9i1.2020.62-68.

[8] A. M. Fadhilah, M. Iwan Wahyuddin, and D. Hidayatullah, "Analisis Faktor yang Mempengaruhi Perokok Beralih ke Produk Alternatif Tembakau (VAPE) menggunakan Metode *K-m*eans Clustering," *Jurnal Teknologi Informasi dan Komunikasi)*, vol. 5, no. 2, p. 2021, 2021, doi: 10.35870/jti.

[9] R. Suwanda, Z. Syahputra, and E. M. Zamzami, "Analysis of Euclidean Distance and Manhattan Distance in the *K-m*eans Algorithm for Variations Number of Centroid K," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jul. 2020. doi: 10.1088/1742-6596/1566/1/012058.

[10] D. M. Saputra, D. Saputra, and L. D. Oswari, "Advances in Intelligent Systems Research," 2020.

[11] L. T. Berampu and W. D. Sari, "Human Resources Transformation in the Digitalization Professional Era in North Sumatera," *Esensi: Jurnal Bisnis dan Manajemen*, vol. 10. no. 2, pp. 135–146, Jan. 2021, doi: 10.15408/ess.v10i2.18477.

[12] Olyvia Darussalam, "Faktor-Faktor Penyebab Kredit Bermasalah Di Pt. Bank Sulut Cabang Utama Manado".

[13] S. Widowati, "Pengaruh Pengendalian Internal, Kualitas Sumber Daya Manusia Dan Pelatihan Kapasitas Usaha Terhadap Risiko Kredit Endang Dwi Retnani Sekolah Tinggi Ilmu Ekonomi Indonesia (STIESIA) Surabaya."

Klasifikasi Pengembalian Sinyal Radar dari Ionosfer Menggunakan *Machine Learning* dengan Metode *Voting Ensemble*

Aziz Prabowo
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
azizprabowo128@gmail.com

Mohammad Bayu P.
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
bajradiatama@gmail.com

Andika Ristianto
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
andikaristianto66@gmail.com

Abstrak—Klasifikasi adalah teknik yang digunakan untuk mengkategorikan objek ke dalam kelompok berdasarkan atributnya. Untuk meningkatkan akurasi dalam klasifikasi, teknik *ensemble learning* digunakan. *Ensemble learning* melibatkan penggunaan beberapa model yang bekerja bersama untuk menghasilkan prediksi yang lebih kuat. Dalam penelitian ini, digunakan berbagai algoritma *Machine Learning*, seperti *Logistic Regression* (LR), *Support Vector Classifier* (SVC), *Multi Layer Perceptron* (MLP), *k-Nearest Neighbors* (KNN), *dan Decision Tree* (DT), untuk mengklasifikasikan data radar ionosfer. Dataset ini digunakan untuk mendeteksi adanya struktur di ionosfer dan dibagi menjadi kelas "baik" dan "buruk" berdasarkan hasil radar. Metode *SMOTEENN* digunakan untuk menangani ketidakseimbangan kelas dalam dataset dengan menggabungkan *oversampling (SMOTE)* dan *under-sampling (ENN).* Hasil eksperimen menunjukkan bahwa *ensemble voting classifier* dengan dan tanpa *SMOTEENN* meningkatkan akurasi dibandingkan dengan algoritma individu. Dalam kasus pemrosesan data dengan *SMOTEENN, ensemble voting classifier* mencapai akurasi yang sangat tinggi (99.73%), sementara tanpa *SMOTEENN*, akurasi tetap lebih tinggi daripada sebagian besar algoritma individu. Kesimpulannya, *ensemble voting classifier* adalah metode yang efektif dalam meningkatkan akurasi dalam tugas klasifikasi, terutama ketika digunakan bersamaan dengan teknik pemrosesan data seperti *SMOTEENN*. Kinerja akhir bergantung pada karakteristik data dan masalah yang dihadapi, tetapi secara umum, *ensemble learning* adalah pendekatan yang kuat untuk meningkatkan prediksi model.

Kata Kunci—Ionosphere, Machine Learning, Ensemble Learning, Voting Classifier, SMOTEENN

I. PENDAHULUAN

Klasifikasi [1] merupakan pemeran utama dalam pengambilan suatu keputusan dalam berbagai permasalahan di dunia. Klasifikasi digunakan ketika terdapat kebutuhan untuk menetapkan objek ke dalam suatu kelas atau grup tertentu yang telah ditentukan berdasarkan atribut objeknya. Klasifikasi merupakan jenis pembelajaran yang diawasi dengan mempelajari cara mengkategorikan data baru menggunakan pengetahuan yang diperoleh dari kumpulan data pelatihan yang diberi label sebelumnya. Algoritma klasifikasi yang umum antara lain *Decision Tree, Neural Network, Logistic Regression, KNN,* dan *Naive Bayes*. Penelitian ini dilakukan untuk mengidentifikasi keberhasilan atau kegagalan suatu model dalam memprediksi apakah sebuah sinyal ionosfer dapat memantul kembali ke bumi atau tidak. Dataset yang digunakan merupakan studi kasus klasifikasi biner dengan dua kelas klasifikasi yaitu "*good*" (sinyal yang dapat memantul kembali ke bumi) dan "*bad*" (sinyal yang tidak dapat memantul kembali).

Ensemble learning [2] adalah pendekatan *machine learning* menggunakan beberapa model yang dilatih atau diterapkan pada kumpulan data dengan tujuan untuk mengatasi masalah yang sama. Tujuannya adalah untuk mengumpulkan beragam prediksi dari model-model ini, kemudian menggabungkannya menjadi satu prediksi yang lebih kuat. Metode ini melibatkan pembuatan kumpulan model, model individual dikembangkan melalui proses pembelajaran yang diterapkan pada suatu masalah tertentu. Model individual ini, yang dikenal sebagai *ensemble*, kemudian digabungkan secara terkoordinasi untuk menghasilkan prediksi akhir. Teknik ini digunakan untuk meningkatkan akurasi prediksi model secara keseluruhan, hasil dari kombinasi model menghasilkan model kolaboratif yang kuat secara keseluruhan.

Bagian II dari jurnal ini menyajikan tinjauan pekerjaan yang terkait, kemudian Bagian III menyajikan detail dari kumpulan data, pemrosesan data dan teknik *machine learning* yang digunakan. Hasil dari masing-masing model beserta akurasinya disajikan pada Bagian IV. Kesimpulan di uraikan pada Bagian V.

II. KAJIAN TERKAIT

Penelitian [3] menggunakan model *ensemble* dengan menggabungkan prediksi tiga algoritma *decision tree*: *CART, CHAID* dan *QUEST* pada dataset *ionosphere*, pada penelitian tersebut menunjukkan penggunaan model *ensemble* memperoleh akurasi yang lebih tinggi dibanding nilai akurasi dari masing masing model.

[4] Penelitian ini menggunakan klasifikasi *ensemble* dalam pendeteksian *ADHD* dengan menggabungkan berbagai algoritma, yang berbasis *KNN: K-Nearest Neighbour* (KNN), *Fuzzy K-Nearest Neighbour* (FKNN), and *Neighbour Weighted K-Nearest Neighbour* (NWKNN). Mereka memperoleh tingkat akurasi yang tinggi sebesar 95% menggunakan klasifikasi *ensemble* dengan nilai k = 10, hasil tersebut lebih tinggi dibandingkan nilai akurasi dari masing-masing model.

Karya Sebelumnya [5] telah secara luas mengeksplorasi penerapan teknik *ensemble learning* dalam konteks pendeteksian intrusi. Mereka mendemonstrasikan bahwa metode *ensemble* dapat meningkatkan akurasi dalam

mendeteksi anomali pada lalu lintas jaringan. Dengan menggabungkan beberapa pengklasifikasi, mereka berhasil mencapai peningkatan signifikan dalam mengidentifikasi anomali. Menggunakan *Naive Bayes*, mereka mencapai tingkat akurasi sebesar 77,4%. Sedangkan, dengan implementasi *ensemble learning*, mampu menghasilkan akurasi yang jauh lebih tinggi, mencapai 96,8%.

III. METODE PENELITIAN

Dalam penelitian ini menggunakan algoritma *Machine Learning* untuk memecahkan masalah yang ditemukan pada data serta di tulis dengan *python* pada google colab. Algoritma *Machine Learning* yang berbeda dapat diterpakan langsung ke dataset dengan menggunakan *pyhton* melalui tahapan *pre-processing* data untuk dinormalisasi agar rentang data yang dimiliki tidak terlalu jauh. Dataset [6] yang digunakan untuk penelitian ini data radar yang bersumber dari *UCI Machine Learning* yang dikumpulkan oleh sistem di Goose Bay, Labrador. Sistem ini terdiri dari 16 antena frekuensi tinggi yang tersusun secara berurutan dan memiliki daya pancar sekitar 6.4 kilowatt. Tujuannya adalah untuk mendeteksi elektron bebas di ionosfer. Keberhasilan radar dapat diukur dari kemampuannya dalam menghasilkan hasil yang menunjukkan adanya beragam struktur di ionosfer dengan label "baik". Sebaliknya, jika sinyal gagal menembus ionosfer, maka dianggap sebagai hasil radar dengan label "buruk". Dataset terdiri dari 35 atribut, 34 atribut merupakan prediktor yang bersifat kontinu dan atribut ke-35 adalah target yang diberi label "baik" atau "buruk" dengan jumlah total 351 baris yang terdiri 126 baris dengan kelas "buruk" dan 225 baris dengan kelas "baik". Tabel 1 menunjukkan tipe atribute dari dataset *ionosphere* yang semua atributnya bertipe rentang dan berisikan nilai numerik sedangkan atribut label bertipe kategorikal dan bernilai "g" yang berarti "*good*" (sinyal yang dapat memantul kembali ke bumi) dan "b" yang berarti "*bad*" (sinyal yang tidak dapat memantul kembali).

TABEL 1. ATRIBUT TIPE DATASET *IONOSPHERE*

Atribut	Tipe	Nilai
Attribute1	*Range*	[0,1]
Attribute2	*Range*	[-1.0, 1.0]
Attribute3	*Range*	[-1.0, 1.0]
…	…	…
Attribute34	*Range*	[-1.0, 1.0]
Attribute35	*Categorical*	g/b

SMOTEENN [7] adalah suatu metode yang menggabungkan teknik penambahan data pada kelas minoritas *(oversampling)* dan pengurangan data pada kelas mayoritas *(under-sampling)* dalam penanganan ketidakseimbangan kelas. *SMOTEENN* menggunakan dua konsep dasar: *SMOTE* untuk *oversampling* dan *Edited Nearest Neighbor* (ENN) untuk *under-sampling*. Prinsip dasarnya adalah berdasarkan pada pendekatan *K-Nearest Neighbors* (KNN). Prosesnya dimulai dengan memilih data acak dari kelas minoritas. Kemudian, menghitung jarak antara data yang dipilih dengan tetangga terdekatnya (K tetangga terdekat). Kemudian akan memilih angka acak antara 0 dan 1, dan mengalikannya dengan jarak yang telah dihitung. Hasilnya akan ditambahkan ke kelas minoritas sebagai data sintetis. Selanjutnya, konsep *Edited Nearest Neighbor* (ENN) digunakan sebagai bentuk *under-sampling*. *ENN* menggunakan *KNN* untuk menentukan kelas mayoritas dari setiap observasi, dan jika kelas mayoritas berbeda dari kelas observasi, maka observasi tersebut akan dihapus dari dataset. *SMOTEENN* memadukan *oversampling* dengan *SMOTE* dan *under-sampling* dengan *ENN*, menjadikannya alat yang efektif dalam menangani masalah ketidakseimbangan kelas dalam data.

Ensemble learning [8] merupakan proses pembelajaran mesin yang di dalamnya terdapat beberapa model, seperti pengklasifikasi serta dibuat dan dikombinasikan secara cerdas untuk menyelesaikan masalah dalam dunia kecerdasan buatan. Teknik *ensemble* ini umumnya digunakan untuk meningkatkan kinerja model atau mengurangi risiko dalam pemilihan model yang kurang tepat. Selain itu, penggunaan pembelajaran *ensemble* juga dapat melibatkan memberikan kepercayaan pada keputusan yang dihasilkan oleh model, memilih fitur-fitur terbaik, menggabungkan data, pembelajaran tambahan, mengatasi perubahan data yang tidak stabil, dan mengoreksi kesalahan dalam prediksi model.

Dalam metode *ensemble voting classifier* [9] terdapat beberapa model klasifikasi yang awalnya dibuat menggunakan data pelatihan sama. Setiap model dasar ini memiliki aturan klasifikasi yang berbeda. Kemudian, setiap model dasar memberikan prediksinya untuk setiap contoh pengujian. Evaluasi model yang digunakan dalam penelitian ini menggunakan teknik *Cross Validation* dengan k split=10. Ini menghasilkan prediksi akhir dengan mempertimbangkan hasil dari model-model yang lebih baik sejumlah kali tertentu.

Pada penelitian ini, kami menggunakan beberapa model, seperti *Logistic Regression* (LR), *Support Vector Classifier* (SVC), *Multi Layer Perceptron* (MLP), *k-Nearest Neighbors* (kNN), dan *Decision Tree* (DT) untuk mengklasifikasikan hasil eksperimen. Model pembentuk *ensemble voting* menggunakan metode *soft* dengan pembobotan yang dilihat dari akurasi tertinggi sebagai acuan untuk mengambil keputusan.

IV. HASIL DAN PEMBAHASAN

Dataset *Ionosphere* dibagi dengan perbandingan masing-masing 80%:20% untuk melatih model dan mengujinya. Model klasifikasi dilatih dengan menggunakan kumpulan data tes kemudian kumpulan data test digunakan untuk mengevaluasi dan generalisasi model. Dari model klasifikasi yang telah terbentuk kemudian digabungkan untuk membangun model *ensemble lerning* dan dilakukan *voting classifier*. Tabel 2 menunjukan hasil akurasi model tanpa *SMOTEENN* didapatkan akurasi rata-rata dari 5 algoritma dasar sebelum dilakukan *ensemble classifier* 90.09%, dan dilakukan *ensemble learning* dengan menggunkan *voting classifier* mendapatkan akurasi 93.46%.

TABEL 2. HASIL AKURASI TANPA *SMOTEENN*

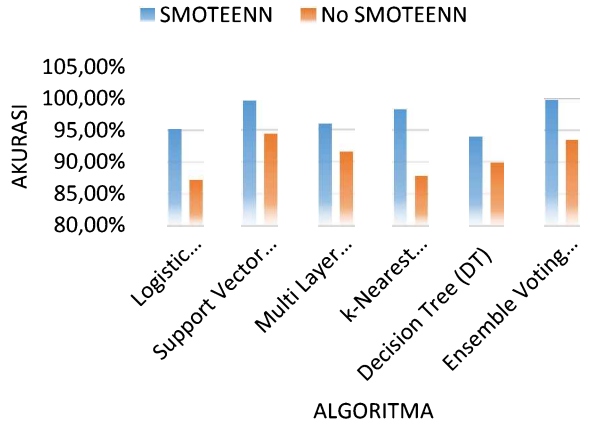
Algoritma	Akurasi
Logistic Regression (LR)	87.19%
Support Vector Classifier (SVC)	94.30%
Multi Layer Perceptron (MLP)	91.47%
k-Nearest Neighbors (kNN)	87.76%
Decision Tree (DT)	89.75%
Ensemble Voting Classifier	93.46%

Proses *SMOTEENN* digunakan karena dataset *Ionosphere* memiliki distribusi kelas tidak seimbang yang dapat mempengaruhi hasil akurasi model. *Resampling method* dengan menggunakan *SMOTEENN* diperlukan untuk dapat mengatasi ketidakseimbangan data. Dari proses dengan metode *SMOTEENN* yang dilakukan, jumlah distribusi kelas berubah menjadi 168 pada kelas b dan 214 pada kelas g. Jarak jumlah antara label berubah menjadi tidak terlalu jauh atau mendekati seimbang sehingga dapat menaikkan hasil akurasi seperti pada Tabel 3.

TABEL 3. HASIL AKURASI DENGAN *SMOTEENN*

Algoritma	Akurasi
Logistic Regression (LR)	95.12%
Support Vector Classifier (SVC)	99.48%
Multi Layer Perceptron (MLP)	95.89%
k-Nearest Neighbors (kNN)	98.20%
Decision Tree (DT)	93.83%
Ensemble Voting Classifier	99.73%

Komparasi hasil akurasi dari model prediksi dengan menggunakan *SMOTEENN* dan tanpa menggunakan *SMOTEENN* dapat dilihat pada gambar 1. Ditunjukan hasil akurasi yang meningkat untuk semua algoritma ketika menggunakan *SMOTEENN* dikarenakan label dari data yang digunakan seimbang sehingga pola-pola yang terbentuk dari model prediksi untuk setiap labelnya menjadi seimbang. Hasil dari *Ensemble Voting Classifier* tanpa *SMOTEENN* sebesar 93.46% meningkat menjadi 99.73% ketika setelah dilakukan *SMOTEENN*.



Gambar 1. Grafik Komparasi *SMOTEENN* dan tanpa *SMOTEENN*

V. KESIMPULAN

Dari percobaan yang dilakukan dapat dilihat perbandingan akurasi antara beberapa algoritma klasifikasi dan *Ensemble Voting Classifier* menggunakan gabungan algoritma sebelumnya pada dua metode pemrosesan data yang berbeda. Berikut adalah kesimpulan dari perbandingan tersebut:

1. Hasil akurasi model dengan metode *SMOTEENN* pada algoritma individu memiliki akurasi yang lebih tinggi dibandingkan dengan tabel akurasi tanpa metode *SMOTEENN*. Hal tersebut menunjukkan bahwa kinerja algoritma meningkat pada pemrosesan data dengan metode *SMOTEENN*.
2. *Ensemble Voting Classifier* pada tabel akurasi dengan metode *SMOTEENN* memiliki akurasi (99.73%) yang lebih tinggi daripada semua algoritma individu. Sehingga penggabungan model dari beberapa algoritma menggunakan *Voting Classifier* menghasilkan hasil yang baik dalam hal akurasi pada tabel akurasi dengan metode *SMOTEENN*.
3. *Ensemble Voting Classifier* pada tabel akurasi tanpa metode *SMOTEENN* juga memiliki akurasi (93.46%) yang lebih tinggi daripada sebagian besar algoritma individu, kecuali SVC (94.30%). Sehingga penggabungan model dari beberapa algoritma menggunakan *Voting Classifier* tanpa metode *SMOTEENN* juga menghasilkan hasil yang baik dalam hal akurasi walaupun tanpa metode *SMOTENN*.

Ensemble Voting Classifier efektif dalam meningkatkan akurasi dibandingkan dengan beberapa algoritma individu dalam kedua metode pemrosesan data dikarenakan pola-pola prediksi yang semakin banyak dari beberapa algoritma dasar. Namun, kinerja akhir tergantung pada data yang digunakan dan karakteristik masalah yang dihadapi seperti pada data imbalance dan data yang telah dilakukan *SMOTEENN*. Dalam kasus pemrosesan data dengan metode *SMOTEENN* dan menggunakan *Ensemble Voting Classifier* memiliki akurasi yang lebih tinggi, sedangkan dengan data yang tidak seimbang sebelum dilakukannya *SMOTEENN* di dapatkan akurasi yang rendah.

REFERENSI

[1] N. O. Aung, "Classification of Radar Returns from Ionoshpere Using NB-Tree and CFS," vol. 2, 2018.

[2] L. G. Kabari and U. C. Onwuka, "Comparison of Bagging and Voting Ensemble Machine Learning Algorithm as a Classifier." [Online]. Available: www.ijarcsse.com,

[3] P. Pushpalata and B. G. Jyoti, "Improving Classification Accuracy by Using Feature Selection and Ensemble Model," 2012.

[4] A. Kusyanti, "Metode Ensemble Classifier Untuk Mendeteksi Jenis Attention Deficit Hyperactivity Disorder (Adhd) Pada Anak Usia Dini," vol. 6, no. 3, pp. 301–308, 2019, doi: 10.25126/jtiik.201961313.

[5] R. Sudiyarno, A. Setyanto, and E. T. Luthfi, "Peningkatan Performa Pendeteksian Anomali Menggunakan Ensemble Learning dan Feature Selection Anomaly Detection Performance Improvement Using Ensemble Learning and Feature Selection," *Citec Journal*, vol. 7, no. 1, 2020.

[6] V. Sigillito, S. Wing, L. Hutton, and K. Baker, "Ionosphere Dataset," Ionosphere. UCI Machine Learning Repository. https://doi.org/10.24432/C5W01B.

[7] A. M. Elsobky, A. El Keshk, and M. G. Malhat, "A Comparative Study for Different Resampling Techniques for Imbalanced datasets," 2023.

[8] B. Mahesh, "Machine Learning Algorithms-A Review," *International Journal of Science and Research*, 2018, doi: 10.21275/ART20203995.

[9] T. Olaleye, A. Abayomi-Alli, K. Adesemowo, O. T. Arogundade, S. Misra, and U. Kose, "SCLAVOEM: hyper parameter optimization approach to predictive modelling of COVID-19 infodemic tweets using smote and classifier vote ensemble," *Soft comput*, vol. 27, no. 6, pp. 3531–3550, Mar. 2023, doi: 10.1007/s00500-022-06940-0.

Analisis Cluster Ulasan Aplikasi MyPertamina pada Google Play Store menggunakan K-Means *Clustering*

Bagas Dwi Santosa
Fakultas Teknik dan Teknologi Informasi
Universitas Jenderal Achmad Yani Yogyakarta
Yogyakarta, Indonesia
bagasdwisantosa87@gmail.com

Ulfi Saidata Aesyi
Fakultas Teknik dan Teknologi Informasi
Universitas Jenderal Achmad Yani Yogyakarta
Yogyakarta, Indonesia
ulfiaesyi@gmail.com

Abstract— **MyPertamina is an application created and introduced as part of efforts to digitize Public Fuel Filling Stations (SPBU). The use of this application has generated various responses from people who use fuel oil (BBM). Therefore, this research takes data from user reviews of the MyPertamina application on Play. Store and App Store. The review data was then cleaned and weighted. Next, clustering was carried out using the K-Means algorithm which obtained 9 optimal clusters. Where 3 clusters described positive reviews and 5 clusters described negative reviews of the MyPertamina application. Positive clusters (clusters 1, 3, and 4) highlight the positive aspects of the application, such as reliability, good service, and ease of purchasing fuel. On the other hand, negative clusters (clusters 0, 2, 5, 6, 7, and 8) raise complaints, such as difficulties in use, frequent login or update problems, as well as problems in the payment process. Overall, the analysis results explain that the MyPertamina application received positive appreciation in several aspects such as reliability, good service and ease of use. However, the findings also highlight several weaknesses, such as technical problems, difficulties in use, and obstacles in the payment process.**

Keywords— MyPertamina, K-means, Clustering

I. PENDAHULUAN

Penggunaan Bahan Bakar Minyak (BBM) merupakan kebutuhan penting bagi masyarakat untuk aktivitas sehari-hari, terutama sebagai bahan bakar transportasi. PT Pertamina adalah Perusahaan yang membuat serta menyuplai bahan bakar bagi keperluan penduduk di Indonesia, PT Pertamina telah memajukan inovasi terbaru dari segi keuangan khususnya cara pembayaran untuk memastikan konsumen mendapatkan kemudahan dalam bertransaksi dalam pembelian produk-produk Pertamina[1]. Salah satu inovasi PT Pertamina adalah layanan aplikasi MyPertamina.

MyPertamina adalah aplikasi yang dibuat dan diperkenalkan sebagai bagian dari upaya digitalisasi Stasiun Pengisian Bahan Bakar Umum (SPBU). Aplikasi MyPertamina merupakan aplikasi layanan digital dari Pertamina dengan berbagai layanan yang ada didalamnya yaitu untuk menemukan lokasi SPBU Pertamina terdekat, pembayaran digital yang sekaligus mendapatkan loyalty point dan sistem pencatatan ketika belanja bensin bulanan agar lebih mudah[2]. Aplikasi MyPertamina bertujuan untuk memastikan penyaluran Bahan Bakar Minyak (BBM) subsidi dilakukan dengan akurat sesuai sasaran yang ditetapkan. Pemerintah, melalui PT Pertamina, berencana menerapkan pembelian BBM bersubsidi melalui penggunaan aplikasi MyPertamina. Hal itu agar penyaluran BBM subsidi bisa tepat sasaran.

Pada penelitian sebelumnya yang telah dilaksanakan, memberikan wawasan terkait dengan analisis tingkat manfaat dari aplikasi MyPertamina. Dari hasil analisis tersebut memberikan kesimpulan bahwa penelitian lebih berfokus pada komentar masyarakat di media sosial Twitter. Aplikasi MyPertamina terpantau hanya mendapatkan nilai 1,3 di Play Store. Rata-rata pengguna juga mengeluhkan aplikasi itu lambat dan menyusahkan[3]. Untuk memahami masalah yang mendasari nilai rendah tersebut, perlu adanya analisis yang memadai terkait dengan ulasan pengguna aplikasi MyPertamina di Google Play Store. Sehingga perlu untuk memahami lebih dalam pandangan serta pengalaman pengguna dalam menggunakan aplikasi MyPertamina. Dalam penelitian ini, penulis melanjutkan eksplorasi lebih lanjut dengan fokus pada analisis ulasan pengguna aplikasi MyPertamina di Google Play Store.

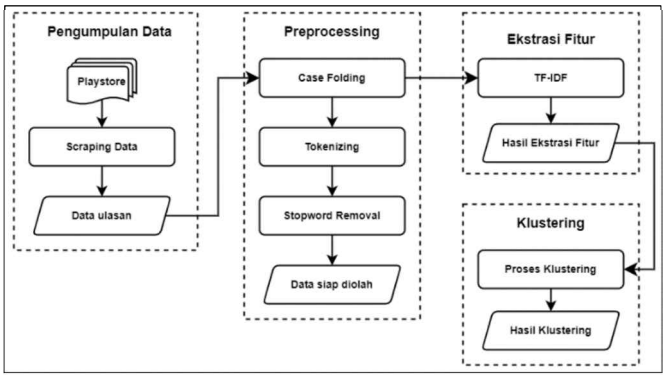
Tujuan dari penelitian ini adalah untuk mendalami ulasan pengguna yang muncul dalam aplikasi MyPertamina. Penulis menerapkan metode K-Means *Clustering* untuk mengelompokkan ulasan-ulasan tersebut menjadi klaster-klaster yang memiliki karakteristik serupa. Dengan metode ini, penulis berharap bisa memberikan pemahaman lebih mendalam tentang pengalaman pengguna aplikasi MyPertamina serta mengidentifikasi pola-pola yang mungkin berguna untuk perbaikan dan pengembangan lebih lanjut dari aplikasi MyPertamina.

II. METODE PENELITIAN

Penelitian ini dilakukan dengan beberapa tahapan yaitu, Pengumpulan data (*Crawling*), Pembersihan data / Pre-processing (*Case folding, Tokenizing, Stopword removal*), Ekstrasi fitur (TF-IDF), dan Klustering menggunakan metode K-Means *Clustering*[4]. Tahapan penelitian terdapat pada gambar 1.

A. Pengumpulan Data

Pengumpulan data dalam penelitian ini menggunakan metode *scraping* data. Data yang diambil adalah ulasan dari pengguna aplikasi MyPertamina di Google Play Store. Data diambil pada rentang waktu 1 Januari – 24 September 2023.



Gambar 1. Tahapan Penelitian

B. *Pre-processing*

Tahap selanjutnya dalam penelitian ini adalah proses pembersihan data atau pre-processing data. Dalam pre-processing terdapat beberapa tahapan seperti, *Case folding* adalah proses dalam pre-processing yang mengubah teks atau karakter menjadi huruf besar maupun huruf kecil digunakan secara konsisten. Dalam penelitian ini teks dari ulasan diubah menjadi huruf kecil secara keseluruhan. Selanjutnya, *Tokenizing* merupakan tahap pemisahan suatu kalimat menjadi pecahan kata tunggal atau token. Selain itu, ini memfilter berdasarkan panjang teks dan menghilangkan karakter tertentu, seperti tanda baca[5]. *Stopword Removal*, yaitu menghilangkan kata yang kurang efektif menggunakan *library* NLTK untuk *filtering* terhadap *dataframe*[6].

C. *Pembobotan Data*

Tahap pembobotan dana menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode TF-IDF dibagi menjadi *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). *Term Frequency* (TF) merupakan proses untuk menghitung jumlah kemunculan kata dalam tiap data ulasan. *Inverse Document Frequency* (IDF) digunakan untuk menghitung kata yang muncul di berbagai data ulasan yang dianggap sebagai kata umum, yang dinilai tidak penting[7]. Formula TF-IDF ditunjukkan seperti pada persamaan (1) dan (2)[8].

$$idf_i = \log\left(\frac{N}{df_i}\right) \quad (1)$$

$$w_{i,j} = tf_{i,j} \times idf_i \quad (2)$$

Keterangan:
w = bobot termpada dokumen (i, j: 1, 2, 3,...n).
n = total keseluruhan dokumen yang ada.
t = term atau kata yang akan dihitung bobotnya. (i: 1, 2, 3,...n).
d = dokumen yang mengandung sekumpulan term (j: 1, 2, 3,...n).
df = frekuensi atau jumlah dokumen yang berisi term (i: 1, 2, 3,...n).
tf = frekuensi atau jumlah termyang muncul dalam sebuah dokumen (i, j: 1, 2 ,3,...n).
idf = inverse frekuensi dokumen, nilai kemunculan term pada kumpulan dokumen (i: 1, 2, 3,...n).

D. *Klustering*

Klustering ialah teknik data mining yang digunakan untuk menganalisis dan mengkaji data untuk menyelesaikan permasalahan dalam pengelompokkan data membagi dari suatu dataset ke dalam subset[9]. Penelitian ini menggunakan metode K-Means *clustering* untuk mengelompokan data ulasan aplikasi MyPertamina di Play Store. Metode K-Means digunakan untuk mengelompokkan data sesuai dengan kemiripannya. Kemiripan ini dihitung dengan jarak *euclidean*[10]. Terdapat beberapa proses metode K-Means yaitu sebagai berikut[11].

- Tentukan nilai *k* sebagai jumlah kluster yang ingin dibentuk.
- Inisialisasi *k* sebagai *centroid* yang dapat dibangkitkan secara *random.*
- Hitung jarak setiap data ke masing-masing *centroid* menggunakan persamaan *Euclidean Distance* yaitu sebagai berikut:

$$d(P,Q) = \sqrt{\sum_{j=1}^{p}(x_j\,(P) - x_j\,(Q))^2}$$

  Dari rumus diatas, dapat dijelaskan $D_{(i,j)}$ yaitu jarak data ke *i* ke pusat cluster *j*, X_{ki} adalah data ke *i* pada atribut data ke *k* dan X_{kj} = titik pusat ke *j* pada atribut ke *k*
- Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan centroidnya.
- Tentukan posisi centroid baru (k).
- Kembali ke langkah 3 jika posisi centroid baru dengan centroid lama tidak sama.

III. HASIL DAN PEMBAHASAN

Pada penelitian ini, *scraping* data dilakukan menggunakan pemrograman python, *library google_play_scraper* serta menggunakan *tools* Google Colaboratory. Jumlah dataset yang berhasil diambil sebanyak 7009 row. Hasil *scraping* dapat dilihat pada table I.

TABEL 1. SAMPLE DATA

No	Ulasan
1.	Aplikasi gak berguna, percuma daftar udah berhasil tapi barcode ditolak waktu ngisi. buat apa bikin barcode
2.	SPBU nya banyak yg ngak bisa pake aplikasi mypertamina, belum siap
3.	Saya puas dengan pelayanannya

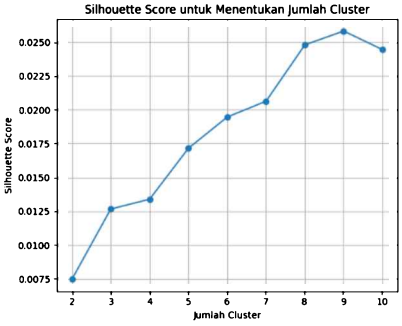
Tahap selanjutnya adalah pre-processing data, pre-processing merupakan tahapan untuk meningkatkan kualitas citra, menghilangkan noise yang ada pada citra, maupun menentukan bagian citra yang akan digunakan dalam tahapan selanjutnya[12]. Hasil beberapa tahapan dalam pre-processing seperti *Case folding, Tokenizing,* dan *Stopword removal* dapat dilihat pada table II.

TABEL 2. CONTOH DATA PRE-PROCESSING

Tahapan	Hasil
Data Asli	Ngluncurin aplikasi kaya gak guna, loading ke pembayan gagal
Case Folding	ngluncurin aplikasi kaya gak guna loading ke pembayan gagal
Tokenizing	[‘ngluncurin’, ‘aplikasi, ‘kaya, ‘gak’, ‘guna’, ‘loading’, ‘pembayaran’, ‘gagal’]
Stopword Removal	[‘luncur’, ‘aplikasi, ‘tidak’, ‘guna’, ‘load’, ‘bayar’, ‘gagal’]

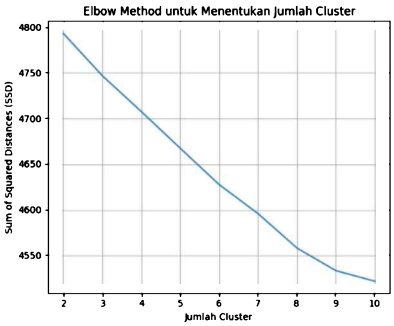
Jumlah data bersih hasil tahap pre-processing sebanyak 4967 *row*, data bersih yang didapat akan dilakukan perhitungan TF-IDF dari setiap kata. Penelitian ini menggunakan pembobotan data dengan metode TF-IDF karena jika hanya menggunakan TF maka menghasilkan data uji yang kurang bagus. Sehingga pembobotan data ini perlu menggunakan IDF.

Dalam penelitian ini, penentuan jumlah kluster yang optimal dilakukan menggunakan metode *silhouette score* dan *elbow method*. *Silhouette score* atau yang sering disebut juga dengan *silhouette coefficient* merupakan metode pengukuran model *machine learning* yang mampu mengukur kualitas dan kekuatan kluster, sehingga dapat dilihat seberapa baik data ditempatkan dalam sebuah kluster[13]. Metode *Elbow* menghitung nilai selisih penurunan nilai *Sum of Square Error*(SSE) yang paling besar dan berbentuk siku[14]. Pengujian dilakukan secara acak dengan pengulangan sebanyak 10 kali jumlah kluster yang diteliti. Hasil dari pengujian *silhoutte score* dan *elbow* method direpresentasikan dalam grafik.



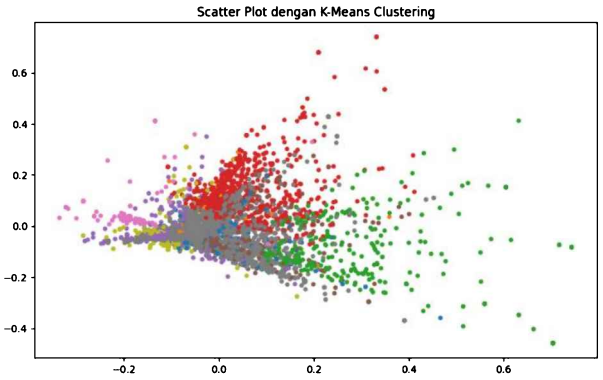
Gambar 2. Pengujian Silhoutte Score

Gambar 2 menampilkan grafik hasil pengujian *silhoutte score,* proses pengujian dilakukan sebanyak 9 kali dan mendapatkan kluster optimal sebanyak 9 kluster dengan nilai *silhoutte* sebesar 0.0257.



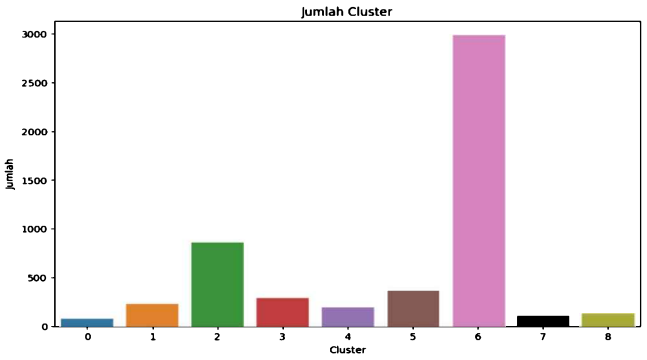
Gambar 3. Pengujian Elbow Method

Gambar 3 menampilkan grafik hasil pengujian *elbow method,* grafik tersebut menunjukkan penurunan yang cukup signifikan pada jumlah kluster 9. Pengujian *elbow method* dan *silhoutte score* sama sama memiliki hasil pengujian di kluster 9. Maka diambil kesimpulan bahwa jumlah K yang optimal adalah k=9. Setelah menentukan jumlah kluster optimal, dilakukan visualisasi data ke dalam 9 kluster yang menunjukkan tingkat homogenitas yang sesuai.



Gambar 4. Sebaran kluster ulasan MyPertamina

Pada gambar 4 memperlihatkan hasil persebaran data dengan menggunakan 9 kluster. Untuk melihat perbandingan jumlah data disetiap kluster bisa dilihat pada gambar 5.



Gambar 5. Grafik jumlah data disetiap kluster

Pada grafik terlihat jelas bahwa terdapat 1 kluster yang memiliki data paling banyak yaitu kluster 6, dan diikuti kluster 2 dan 5. Berdasarkan grafik tersebut perlu dilihat hasil kluster, apakah kluster tersebut berisi ulasan negatif atau ulasan positif.

TABEL 3. SAMPLE DATA KLUSTER 0

Cluster	Hasil
0	aplikasi susah kalau tidak niat jagan bikin tambah ribet
	aplikasi ribet mau pakai updet terus pusing
	parah makin ribet aplikasi
	aplikasi ribet sering error kaya instansi
	aplikasi sampah percuma download tidak guna semua pom bensin metode bayar ribet

Tabel 3 merupakan sampel hasil dari kluster 0, dari sampel tersebut bisa dilihat bahwa kluster 0 berisi keluhan dari pengguna aplikasi MyPertamina seperti ribet, susah dan *error*. Maka dari itu kluster 0 bisa disimpulkan berisi ulasan negatif terhadap aplikasi MyPertamina.

TABEL 4. SAMPLE DATA CLUSTER 1

Cluster	Hasil
1	mantap jadi mudah belum
	mantap aplikasi sering promo
	mantap lebih praktis
	oke mantap banyak promo cashback
	mantap keren pertamina aplikasi

Tabel 4 adalah sampel hasil kluster1, yang berisi kepuasan dalam menggunakan aplikasi MyPertamina seperti mantap, mudah, dan praktis. Maka dari itu kluster 0 dapat disimpulkan sebagai kluster positif.

TABEL 5. SAMPLE DATA CLUSTER 2

Cluster	Hasil
2	aplikasi buka bantu malah repot mau masuk aja susah minta ampun masuk link susah
	tiap mau bayar susah aneh banget
	susah masuk kalo ketika sudah login jangan suka keluar sendiri ribet login ulang kadang susah masuk
	aplikasi tidak jelas susah butuh gimana daftar buat barcod biar guna buat isi bbm spbu malah sama aplikasi tidak jelas malah buat bingung daftar susah anjing
	daftar barcode susah link tidak buka payah

Tabel 5 merupakan sampel kluster 2, yang berisi ulasan keluhan pengguna aplikasi MyPertamina dalam menggunakan aplikasi. Ulasan keluhan tersebut seperti susah bayar, susah masuk, dan susah daftar. Maka bisa disimpulkan kluster 2 berisi ulasan negatif.

TABEL 6. SAMPLE DATA CLUSTER 3

Cluster	Hasil
3	bagus belum semua spbu guna aplikasi beberapa alas utama jaring ada
	sangat bagus sangat bantu tingkat informasi tarik
	layan system bagus kalau tinjau siapa pantas subsidi
	sangat bagus jaga beli guna bisnis
	bagus enak praktis pengguna

Tabel 6 merupakan sample kluster 3, yang berisi ulasan positif. Kluster 3 berisi ulasan seperti bagus, bantu, dan praktis. Yang menjelaskan aplikasi MyPertamina membantu pengguna.

TABEL 7. SAMPLE DATA CLUSTER 4

Cluster	Hasil
4	bagus bantu cuman tak guna karna saldo tak isi
	daftar subsidi tepat sulit cs bantu sama sekali
	sangat bantu mudah
	lebih mudah proses bayar sangat bantu guna bbm
	layan cukup bantu lancar beli bbm subsidi

Tabel 7 merupakan sample kluster 4, yang berisi ulasan positif. Kluster 4 berisi ulasan yang membahas pelayanan aplikasi MyPertamina dalam melakukan pembelian BBM.

TABEL 8. SAMPLE DATA CLUSTER 5

Cluster	Hasil
5	proses daftar sangat sulit bingung mudah akses pihak call center lamban
	pernah daftar baru instal terang no daftar dzolim rakyat bukan mudah
	sudah daftar sudah dapet barcode kenapa tidak ada apa akhir beli dexlite jalan wonosobo jkt
	daftar gagal terus padahal data isi udah benar tetep gagal aneh
	sudah bener login pakek no hp sama pin daftar malah gagal terus

Tabel 8 merupakan sample kluster 5, berisi ulasan negatif. Ulasan tersebut membahas kegagalan dalam melakukan login dan menggunakan barcode.

TABEL 9. SAMPLE DATA CLUSTER 6

Cluster	Hasil
6	masa tiap hari minta update aplikasi parah
	mau isi bensin selalu minta update
	minta update terus aplikasi tidak jelas
	update melulu pas mau dipake
	habis update apl malah makin buruk

Tabel 9 merupakan sample kluster 6, berisi ulasan negatif terhadap aplikasi MyPertamina. Pengguna mengeluhkan aplikasi yang sering membutuhkan update saat digunakan.

TABEL 10. SAMPLE DATA CLUSTER 7

Cluster	Hasil
7	aplikasi bagus kontrol analisa pakai bbm mungkin perlu tambah input km kendara isi bbm serta report avg liter km isi avg liter km total kendara segi operasional kota palembang beberapa spbu belum support untuk support sering temu bisa pakai offline ganggu jadi sayang pas ngisi ga point gara spbu sedang ganggu ga support aplikasi
	ada menu upload bagi stnk kalo foto kamera langsung apk gambar selalu jelek blur segera baik,7
	akhir aplikasi seperti masalah selalu gagal muat saldo mungkin segera baik jadi kendala beli bbm lebih cepat malah lama
	buat apa aplikasi kalau spbu terima bayar lalu aplikasi mypertamina padahal daftar spbu aplikasi tapi tetep tidak mau pake aplikasi bayar cash update lebih parah loading lama padahal jaring g kuota banyak haduh
	kecewa update malah tambah gak jelas stnk sama asli bilang stnk sesuai bukan jadi bagus malah jadi kacau sulit

Tabel 10 merupakan sample kluster 7, berisi ulasan negatif. Bisa dilihat bahwa pengguna aplikasi MyPertamina mengeluhkan permasalahan didalam penggunaan aplikasi tersebut, tetapi pengguna juga memberikan pendapat positif ataupun saran terhadap aplikasi MyPertamina.

TABEL 11. SAMPLE DATA CLUSTER 8

Cluster	Hasil
8	aplikasi payah mau masukin bayar ovo hubung akun link aja dulu hubung pernah coba baik aplikasi hapus aja link aja kalau gak konek my pertamina
	sangat kecewa banyak spbu ga bayar non tunai aplikasi may pertamina percuma dong kalau orang kantor cipta aplikasi may pertamina masih banyak spbu bisa terima bayar non tunai aplikasi may pertamina
	apps tulis pom pake my pertamina pas sana sering banget bilanh tidak gajelas percuma top up
	sangat cocok pakai gak pake udah top up malah gak pake my pertamina mending hapus aja apk nya
	aplikasi si bagus sayang banyak spbu tidak terima bayar pakai my pertamina alas troble

Tabel 11 merupakan sampel kluster 8, yang berisi ulasan negatif. Hal tersebut bisa dilihat dari hasil ulasan kluster 8, para pengguna mengeluhkan permasalahan pembayaran tunai ataupun nontunai dalam aplikasi MyPertamina.

KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, penggunaan metode *silhoutte score* dan *elbow method* mendapatkan 9 kluster yang optimal. 9 kluster tersebut diimplementasikan kedalam algoritma K-Means dan mendapatkan ulasan positif dan negatif.

Dari analisis 9 kluster, terdapat 3 kluster yang berisi ulasan positif yaitu kluster 1, 3, dan 4. Dari setiap kluster positif tersebut berisi topik ulasan yang berbeda beda, seperti kluster 1 yang berisi bahwa aplikasi MyPertamina mantap dan praktis. Kluster 3 yang berisi ulasan bahwa aplikasi MyPertamina dalam segi pelayanan bagus. Dan kluster 4 yang berisi ulasan bahwa MyPertamina mudah digunakan untuk melakukan pembelian BBM.

Kemudian terdapat 5 kluster yang berisi ulasan negatif yaitu 0, 2, 5, 6, 7, dan 8. Kluster 0 yang berisi ulasan keluhan dari pengguna bahwa aplikasi MyPertamina hanya bikin ribet. Kluster 2 yang berisi ulasan dalam menggunakan aplikasi MyPertamina yang susah digunakan. Kluster 5 berisi ulasan bahwa aplikasi MyPertamina sering terjadi gagal login. Kluster 6 berisi ulasan bahwa aplikasi MyPertamina sering terjadi update. Kluster 7 yang berisi keluhan pengunaan aplikasi MyPertamina, tetapi pengguna juga memberikan ulasan positif dan saran. Kluster 8 berisi ulasan terhadap aplikasi MyPertamina bahwa kesusahan dalam melakukan pembayaran seperti tunai ataupun nontunai.

Hasil analisis ini menunjukkan keberhasilan dalam beberapa aspek positif seperti pelayanan dan kemudahan penggunaan, tetapi juga menyoroti beberapa aspek negatif seperti masalah teknis, kesulitan penggunaan, dan masalah pembayaran. Walaupun terdapat aspek positif, dominasi permasalahan yang muncul dalam kluster-kluster negatif tersebut menegaskan bahwa evaluasi secara keseluruhan lebih condong ke arah negatif. Oleh karena itu, penting untuk memberikan fokus yang lebih besar pada penyelesaian masalah yang teridentifikasi guna meningkatkan pengalaman pengguna dan meminimalisir keluhan yang ada.

REFERENSI

[1] R. Maria, R. U. Umayah, S. Mahardinny, D. Kalana, and D. D. Saputra, "Analisis Sentimen Persepsi Masyarakat Terhadap Penggunaan Aplikasi My Pertamina Pada Media Sosial Twitter Menggunakan Metode Na\"\ive Bayes Classifier," *Jurnal Komputer Antartika*, vol. 1, no. 1, pp. 1–10, 2023.

[2] R. Maulana, A. Voutama, and T. Ridwan, "Analisis Sentimen Ulasan Aplikasi MyPertamina pada Google Play Store menggunakan Algoritma NBC," *Jurnal Teknologi Terpadu*, vol. 9, no. 1, pp. 42–48, 2023.

[3] K. Kharisma and U. S. Aesyi, "ANALISIS TINGKAT KEBERMANFAATAN MYPERTAMINA MENGGUNAKAN K-MEANS CLUSTERING," *Journal of Information System Management (JOISM)*, vol. 4, no. 2, pp. 91–96, 2023.

[4] A. B. Saputra, P. W. Cahyo, M. Habibi, and A. Priadana, "Analysis and visualization of BPJS on twitter using K-means clustering," *Int. J. Heal. Sci. Technol*, vol. 3, no. 3, pp. 109–117, 2022.

[5] A. T. J. Harjanta, "Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining," *Jurnal Informatika Upgris*, vol. 1, no. 1 Juni, 2015.

[6] I. M. K. Karo, J. A. K. Karo, Y. Yunianto, H. Hariyanto, M. Falah, and M. Ginting, "Analisis Sentimen Ulasan Aplikasi Info BMKG di Google Play Menggunakan TF-IDF dan Support Vector Machine," *Journal of Information System Research (JOSH)*, vol. 4, no. 4, pp. 1423–1430, 2023.

[7] H. H. Mubaroroh, H. Yasin, and A. Rusgiyono, "ANALISIS SENTIMEN DATA ULASAN APLIKASI RUANGGURU PADA SITUS GOOGLE PLAY MENGGUNAKAN ALGORITMA NAÏVE BAYES CLASSIFIER DENGAN NORMALISASI KATA LEVENSHTEIN DISTANCE," *Jurnal Gaussian*, vol. 11, no. 2, pp. 248–257, 2022.

[8] A. Lahitani, U. S. Aesyi, N. Wulandari, and B. D. Santosa, "Cosine Similarity untuk Mengukur Tingkat Kesadaran pada Topik Software Security Berbasis Teks Komentar di Media Sosial Youtube," *Jurnal Sains dan Informatika*, vol. 8, no. 2, 2022.

[9] P. Alkhairi and A. P. Windarto, "Penerapan K-Means Cluster Pada Daerah Potensi Pertanian Karet Produktif di Sumatera Utara," in *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, 2019.

[10] J. Effendi and M. J. Ramadhan, "Analisa Cluster Aplikasi pada Google Play Store dengan Menggunakan Metode K-Means," in *Annual Research Seminar (ARS)*, 2019, pp. 103–106.

[11] A. Asroni, H. Fitri, and E. Prasetyo, "Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokkan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (Studi Kasus: Fakultas Kedokteran dan Ilmu Kesehatan, dan Fakultas Ilmu Sosial dan Ilmu Politik)," *Semesta Teknika*, vol. 21, no. 1, pp. 60–64, 2018.

[12] N. P. Sutramiani, I. K. G. D. Putra, and M. Sudarma, "Local Adaptive Thresholding Pada Preprocessing Citra Lontar Aksara Bali," *Majalah Ilmiah Teknologi Elektro*, vol. 14, no. 1, pp. 27–30, 2015.

[13] I. B. G. Sarasvananda, R. Wardoyo, and A. K. Sari, "The k-means clustering algorithm with semantic similarity to estimate the cost of hospitalization," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 4, pp. 313–322, 2019.

[14] A. T. Rahman, "Coal trade data clustering using K-means (case study Pt. Global Bangkit Utama)," *ITSMART: Jurnal Teknologi dan Informasi*, vol. 6, no. 1, pp. 24–31, 2017.

Implementasi Pengenalan Wajah dan Geofencing pada Sistem Presensi Karyawan Guna Meningkatkan Keamanan dan Integritas Data

Bagus Trianurdin
Departemen Sistem Informasi
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
bagustrianurdin@outlook.com

Umar Zaky
Departemen Sistem Informasi
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
umar.zaky@staff.uty.ac.id

Abstrak—PT Nusantara Berkah Digital atau lebih dikenal dengan Nutapos merupakan sebuah perusahaan startup digital dibidang teknologi informasi. Saat ini sistem presensi karyawan yang digunakan oleh Nutapos menggunakan Google Spreadsheet yang diakses secara daring oleh semua karyawan yang telah diberikan akses oleh HRD. Namun, sistem presensi tersebut memiliki kelemahan, yaitu integritas data presensi tidak terjamin karena semua karyawan yang diberikan akses ke dalam spreadsheet dapat dengan bebas melihat, mengubah, bahkan menghapus data presensi karyawan lainnya. Selain itu, dapat menimbulkan kecurangan dikemudian hari, salah contohnya adalah karyawan yang melakukan titip abses. Dengan membuat sistem presensi berbasis mobile yang menggunakan metode pengenalan pola wajah dan geofencing dapat menjadi solusi untuk mengatasi permasalahan tersebut. Sistem ini juga teringetrasi dengan platform web yang digunakan oleh Manajer HRD untuk mengelola data presensi dan karyawan. Sistem ini dibengun menggunakan framework Flutter untuk perangkat mobile dan Laravel untuk perangkat web. Setiap karyawan memiliki akun akses individu pada aplikasi mobile yang memungkinkan mereka melihat data pribadi mereka sendiri, sehingga tidak ada lagi kemungkinan bagi mereka untuk mengakses data pribadi karyawan lain. Pengujian yang dilakukan dengan menggunakan blackbox menunjukkan tingkat keberhasilkan proses pada sistem ini sebesar 76,92% dari total sebanyak 13 skenario pengujian.

Kata Kunci—Sistem Informasi, Pengenalan Wajah, Geofencing

I. LATAR BELAKANG

PT Nusantara Berkah Digital atau lebih dikenal dengan Nutapos merupakan sebuah perusahaan startup digital dibidang teknologi informasi. Nutapos didirikan pada tahun 2015 dan selang 2 tahun kemudian tepatnya pada tahun 2017 resmi mendirikan perusahaan dengan nama PT Nusantara Berkah Digital. Setelah 7 tahun berdiri, Nutapos memiliki 2 kantor tepatnya di daerah Bantul, Yogyakarta dan Sidoarjo, Jawa Timur. Dengan jarak kantor yang sangat jauh, tentunya dibutuhkan sistem presensi karyawan yang dapat saling terhubung antar kantor sehingga memudahkan karyawan saat melakukan presensi dan memudahkan HRD (Human Resource Development) saat melakukan rekap data presensi.

Saat ini, proses presensi karyawan di Nutapos masih menggunakan Google Spreadsheet yang diakses secara daring oleh seluruh karyawan yang telah diberikan akses oleh HRD. Karyawan Nutapos yang telah diberikan akses dapat masuk ke dalam spreadsheet tersebut dan langsung melakukan presensi dengan mengedit kolom yang sesuai dengan tanggal presensi dan nama masing-masing karyawan. Namun, sistem presensi menggunakan Google spreadsheet memiliki kelemahan, yaitu integritas data presensi tidak terjamin karena semua karyawan yang diberikan akses ke dalam spreadsheet dapat dengan bebas melihat, mengubah, bahkan menghapus data presensi karyawan lainnya. Selain itu, sistem presensi tersebut dapat menimbulkan kecurangan dikemudian hari, salah contohnya adalah karyawan yang melakukan titip abses, yaitu meminta karyawan lain itu mencatatkan kehadirannya padahal tidak hadir atau karyawan melakukan presensi ditempat selain kantor, seperti rumah atau ditempat lain yang tidak sesuai aturan dan perintah perusahaan.

Membuat HRIS (Human Reources Information System) dapat membantu manajemen presensi menjadi solusi untuk permasalahan di atas. HRIS merupakan sistem yang dapat membantu karyawan dan HRD dalam hal presensi, izin, cuti dan lembur karyawan [3]. Penggunaan HRIS dalam sebuah perusahaan memberikan dukungan yang besar kepada HRD dalam mengelola informasi pegawai dengan lebih efisien, terstruktur, dan akurat. Selain itu, penggunaan HRIS juga berpotensi mengurangi kemungkinan hilang atau rusaknya data [10]. Sistem ini akan diimplementasikan dalam bentuk aplikasi web dan mobile. Perangkat web dapat digunakan sebagai media untuk presensi karena memiliki konektivitas yang tinggi, ringkas dan pengoprasiannya yang mudah, sedangkan perangkat web dapat digunakan sebagai media untuk mengatur administrasi presensi dan karyawan [7]. Sistem presensi pada perangkat mobile akan menggunakan metode pengenalan pola wajah dan teknologi Geofencing. Pengenalan wajah dapat digunakan untuk mengatasi kecurangan titip absen karena pada saat proses presensi akan dilakukan pengecekan wajah dan ketika wajah tidak sesuai dengan pemiliki smartphone, maka proses presensi tidak bisa dilakukan [13]. Sementara, dengan menggunakan teknologi Geofencing dapat diketahui lokasi dan koordinat dari smartphone sehingga dapat diketahui posisi yang pasti di dalam peta. Geofencing sendiri memanfaatkan Global Positioning System (GPS) yang terdapat pada smartphone untuk dapat berjalan [2]. Proses pengenalan wajah pada aplikasi mobile menggunakan Google ML Kit dan Tensorflow Lite.

Aplikasi presensi berbasis web dan mobile menggunakan perangkat yang berbeda, sehingga diperlukan suatu metode untuk menghubungkan atau mengkomunikasikan data antara keduanya. Application Programming Interface (API) merupakan sekumpulan perintah program yang digunakan untuk membangun perangkat lunak dengan menyediakan fungsi dan instruksi dalam bahasa yang dapat dengan mudah dipahami oleh para pengembang [11]. Menggunakan

arsitektur API merupakan pilihan yang tepat untuk mengimplementasikan sistem ini, karena memungkinkan integrasi dengan berbagai perangkat sistem yang beragam. Namun, arsitektur API masih belum memiliki standar yang jelas, sehingga ketika aplikasi menjadi kompleks, manajemennya menjadi lebih sulit. Oleh karena itu, Representatif State Transition (REST) yang merupakan standar arsitektur komunikasi yang biasa diterapkan dalam pengembangan situs website dan layanan berbasis aplikasi hadir untuk mengatasi permasalahan tersebut [4].

Beberapa penelitian terdahulu mengenai topik pengembangan sistem presensi menggunakan face recognition dan GPS, pertama penelitian yang dilakukan oleh [3], yang mengembangkan aplikasi presensi berbasis mobile dengan fitur face recognition dan GPS untuk kantor Balai Desa, Desa Mekarjati, Kabupaten Indramayu, menyimpulkan bahwa sistem presensi dapat berjalan dengan baik dan dapat mengurangi kecurangan yang terjadi. Selanjutnya, penelitian yang dilakukan oleh [9], yang mengembangkan sistem informasi presensi online berbasis mobile dan web menggunakan face recognition dan GPS untuk SMK Muhammadiyah 1 Weleri menyimpulkan bahwa sistem sangat memudahkan user dalam hal ini Guru dan Karyawan dalam melakukan presensi dan membantu Admin dalam merekap data presensi. Berdasarkan hasil dari penelitian sebelumnya, dapat simpulkan bahwa penggunaan sistem presensi yang mengintegrasikan fitur pengenalan wajah (face recognition) bersama teknologi GPS yang terdapat pada smartphone dapat secara efektif meningkatkan efisiensi dalam proses presensi. Selain itu, kombinasi ini dapat mengurangi kemungkinan terjadinya kecurangan selama pelaksanaan proses presensi.

II. METODOLOGI PENELITIAN

A. *Pengumpulan Data*

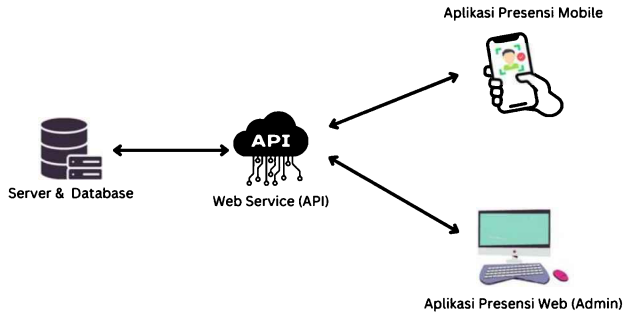
Metode pengumpulan data memiliki tujuan utama yaitu untuk memberikan pemahaman kepada pembaca mengenai berbagai metode yang digunakan dalam penelitian untuk mengumpulkan data yang relevan dengan topik yang sedang diteliti. Dalam subbab ini, akan dibahas dua metode pengumpulan data utama, yaitu wawancara dan observasi, serta studi literatur. Penjelasan mengenai dua metode utama ini akan membantu pembaca memahami secara lebih mendalam bagaimana data dikumpulkan dalam penelitian ini.

1) Wawancara & Observasi: Pengumpulan data awal pada penelitian ini, peneliti memulai tahap ini dengan melakukan wawancara dan observasi. Observasi pertama dilakukan di Kantor PT Nusantara Berkah Digital, yang berlokasi di Bantul, DI Yogyakarta. Tujuan dari observasi ini adalah untuk mengidentifikasi permasalahan yang sedang dihadapi oleh perusahaan. Setelah observasi, langkah selanjutnya adalah melakukan wawancara dengan pejabat terkait di perusahaan. Melalui wawancara ini, peneliti berupaya untuk menghimpun fakta dan data yang diperlukan untuk kelancaran proses penelitian. Pada tahap wawancara ini, peneliti mendapatkan beberapa data karyawan dan data presensi karyawan yang merupakan data hasil rekap selama satu bulan.

2) Studi Literatur: Studi literatur dalam penelitian ini melibatkan pencarian sumber-sumber yang berkaitan dengan permasalahan yang terjadi, seperti artikel, jurnal, buku, paper, dan situs web resmi. Tujuan utama studi literatur adalah untuk memperoleh informasi yang relevan dengan permasalahan yang diteliti serta mengkaji teori-teori yang terkait guna membangun landasan teoritis yang kuat untuk penelitian ini. Dengan demikian, studi literatur menjadi langkah yang penting dalam memahami konteks penelitian dan merumuskan dasar untuk kerangka konseptual.

B. *Arsitektur Sistem*

Sistem presensi yang dikembangkan dalam penelitian ini terdiri dari dua jenis, yaitu berbasis mobile yang akan diakses melalui aplikasi yang diinstal pada perangkat smartphone pengguna dan berbasis web yang dapat diakses melalui peramban web pada perangkat komputer. Kedua sistem ini akan diintegrasikan secara menyeluruh untuk memungkinkan pertukaran data yang efisien di antara keduanya. Oleh karena itu, dalam rangka implementasi sistem ini, pilihan yang paling sesuai adalah mengadopsi arsitektur REST API.



Gambar 1. Arsitektur Sistem

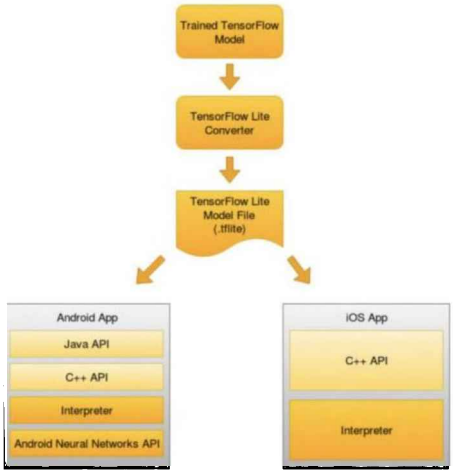
Gambar 1 merupakan serangkaian arsitektur REST API yang diimplementasikan pada penelitian ini. Dengan menggunakan arsitektur tersebut memungkinkan kedua aplikasi yang dibangun, yaitu aplikasi berbasis mobile dan berbasis web dapat dengan mudah terhubung dan berkomunikasi seperti mengirim dan menerima data. Selain itu, arsitektur juga sangat sederhana untuk diimplementasikan ke dalam sebuah sistem karena kemudahan akses data yang dikirim dan diterima. Data yang direpresentasikan oleh REST API berformat text JSON [4], sehingga mobile dan web sebagai klien dapat menerima dan mengirim data dengan mudah.

C. Pengenalan Wajah

Face Recognition atau dalam bahasa Indonesia yaitu Pengenalan Wajah merupakan sistem biometrik atau pengenalan diri yang dapat menganalisis karakter input citra wajah seseorang melalui gambar atau kamera dengan mengukur keseluruhan struktur wajah, jarak antara mata, hitung, mulut dan tepi rahang. Hasil analisis kemudian disimpan ke dalam database untuk kemudian dibandingkan dengan data wajah pembandingnya [14].

Pada penelitian ini, pengenalan wajah dilakukan menggunakan Tensorflow Lite. Menurut [5], Tensorflow merupakan framework atau kerangka kerja buatan Google yang bersifat terbuka (open source) yang dapat digunakan oleh developer untuk membuat, mengambangkan dan melatih model deep learning dengan mudah dengan dan dapat diimplentasikan ke dalam aplikasi web, mobile atau proyek machine learning lainnya. Sedangkan Tensorflow Lite merupakan versi ringan dari Tensorflow yang dirancang untuk perangkat mobile dan perangkat dengan sumber daya terbatas, seperti smartphone, tablet dan perangkat edge.

Pengenalan wajah pada penelitian ini diimplementasikan secara real-time dengan memanfaatkan fitur kamera yang ada pada smartphone. Prosesnya akan melibatkan instruksi kepada pengguna untuk menghadapkan wajahnya ke depan kamera sehingga sistem dapat merekan gambar wajah dari pengguna, kemudian Tensorflow akan memproses gambar tersebut. Namum, perlu diketahui bahwa sistem (komputer) tidak dapat memporses warna dalam gambar secara langsung. Sistem gambar ini dalam format matriks atau vektor. Di sini, sistem akan menerima sebuah gambar RGB. Gambar RGB terdiri dari tiga lapisan warna, yaitu merah, hijau, dan biru. Setiap lapisan ini direpresentasikan oleh sebuah matriks dalam gambar. Unsur-unsur dari setiap matriks tersebut sesuai dengan intensitas warna yang diwakili oleh matriks tersebut pada setiap piksel gambar [12]. Tensorflow sendiri membutuhkan input dari sebuah gambar yang telah dikonversi menjadi bentuk matriks dan akan menghasilkan output dan bentuk yang sama.



Gambar 2. Ekosistem Tensorflow

Tensorflow sendiri memiliki arsitektur yang fleksibel, ini memungkinkan penggunaan yang mudah untuk model pembelajaran mendalam dan jaringan saraf pada CPU, GPU, dan TPU [12]. Gambar 2 merupakan ilustrasi diagram dari ekosistem Tensorflow yang tersedia untuk perangkat mobile melalui Tensorflow Lite, selain itu dijelaskan bahwa TensorFlow membutuhkan model machine learning untuk dapat beroperasi. Untuk menjalankan model tersebut, TensorFlow akan mengubahnya menjadi file model dengan ekstensi tflite yang bisa diproses oleh TensorFlow Lite dan digunakan pada perangkat mobile. Proses konversi ini memungkinkan model dapat berjalan dengan performa yang lebih efisien pada perangkat seluler. Selanjutnya, TensorFlow menyediakan API yang dapat digunakan pada sistem operasi Android dan iOS untuk menjalankan proses machine learning secara efisien dan ringan.

Perlu diketahui, bahwa sebelum dilakukan proses pengenalan wajah, sistem memerlukan cara untuk dapat mendeteksi pola wajah seseorang. Deteksi wajah dilakukan menggunakan ML Kit yang merupakan SDK (Software Development Kit) untuk perangkat mobile yang dikembangkan oleh Google untuk menghadirkan keahlian mechine learning dengan menyediakan beberapa API, seperti Natural Language Processing (NLP), text translation, face detection, object detection, barcode scanning dan lanmark detection sehingga dapat diimplementasikan pada aplikasi Android dan IOS dan memungkinkan pengembang mengintegrasikan fitur-fitur mechine learning tersebut ke dalam aplikasi tanpa perlu mengetahui dan memperdalam tentang cara kerjanya [5].

Pada proses pengenalan pola wajah wajah diperlukan formula atau rumus untuk menghitung jarak terpendek dari dua titik dalam ruang dengan dimensi yang sama. Diambil dari jurnal yang ditulis [8]. Berikut merupakan rumus Euclidean Distance yang digunakan untuk menghitung jarak antara 2 point pada dimensi yang sama:

$$d(x,y) = \sqrt{\sum_{i=1}^{N} (xi - yi)^2} \quad (1)$$

Keterangan:

d	Nilai jarak Euclidean Distance antara x dan y
x, y	: Kedua point dalam Euclidean
i	: Setiap data
n	: Jumlah data
xi, yi	: Kedua point ke-i yang akan dihitung

D. Pengenalan Wajah

Geofencing merupakan teknologi berbasis area yang dapat digunakan oleh sebuah oraganisasi atau perusahaan untuk terhubung dengan karyawan mereka dengan mengirim pesan kepada smarthphone karyawan yang memasuki area atau wilayah geografis yang telah ditentukan sebelumnya sehingga suatu perusahaan mempunyai pagar wilayah berupa latitude dan longitude sebagai prasayat wilayah [15].

Geofencing menggunakan latitude dan lengitude dalam menentukan titik lokasi pada sebuah wilayah seperti yang sudah disebutkan di atas. Dalam kasus sistem presensi ini, ada 2 lokasi yang harus dibandingkan dan dihitung jarak antara keduanya, yaitu lokasi karyawan dan lokasi perusahaan. Untuk menghitung titik lokasi antara lokasi karyawan dan lokasi perusahaan memerlukan formula atau metode. Menurut [6], dalam penetiannya menjelaskan bahwa metode Haversine dapat menghitung jarak antara titik di permukaan bumi menggunakan garis bujur (latitude) dan garis lintang (longitude) sebagai variabel inputan. Berikut merupakan rumus Haversine:

$$a = \sin^2\left(\frac{\Delta lat}{2}\right) + \cos(lat1) \,.\, \cos(lat2) \,.\, \sin^2\left(\frac{\Delta lng}{2}\right) \quad (2)$$

$$d = 2r \,.\, \sin^{-1}(\sqrt{a})$$

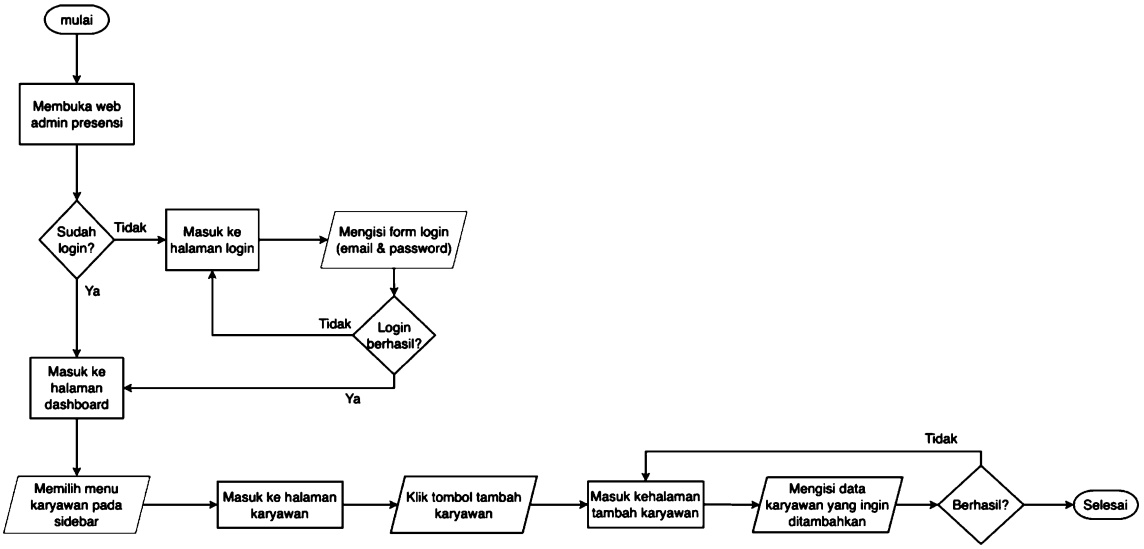
Keterangan:

a	: Jarak
r	: Jari-jari bumi (6356.752 km)
Δlat	: Besaran perubahan latitude = lat2 – lat2
Δlng	: Besaran perubahan longitude = lng2 – lng2

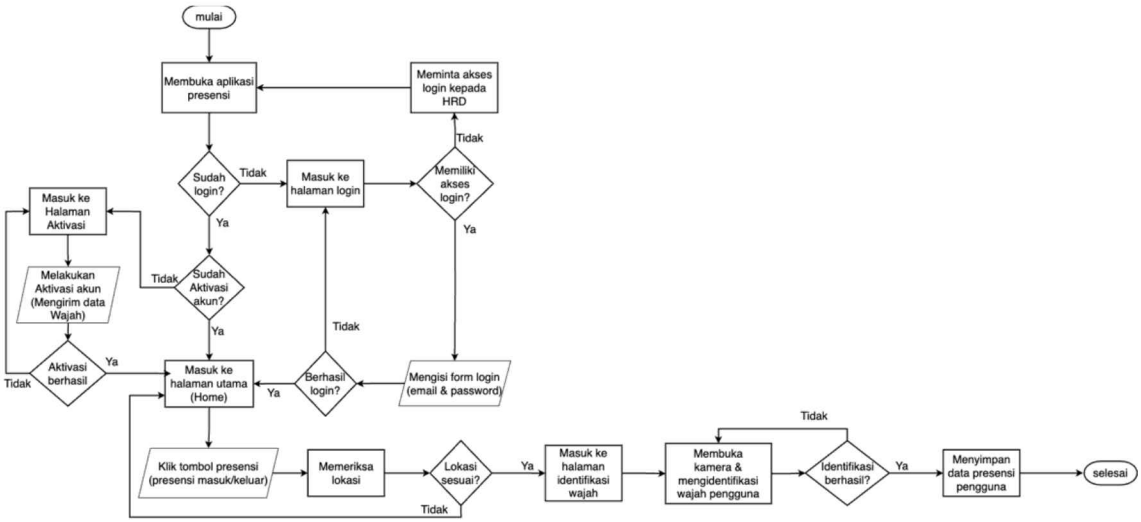
E. Rancangan Sistem

Seperti yang sudah dibahas sebelumnya, penelitian ini mengusulkan sistem presensi dengan menggunakan metode

pengenalan wajah dan teknologi geofencing. Berikut merupakan rancangan proses presensi yang akan dibangun.



Gambar 3. Rancangan Proses Tambah Data Karyawan

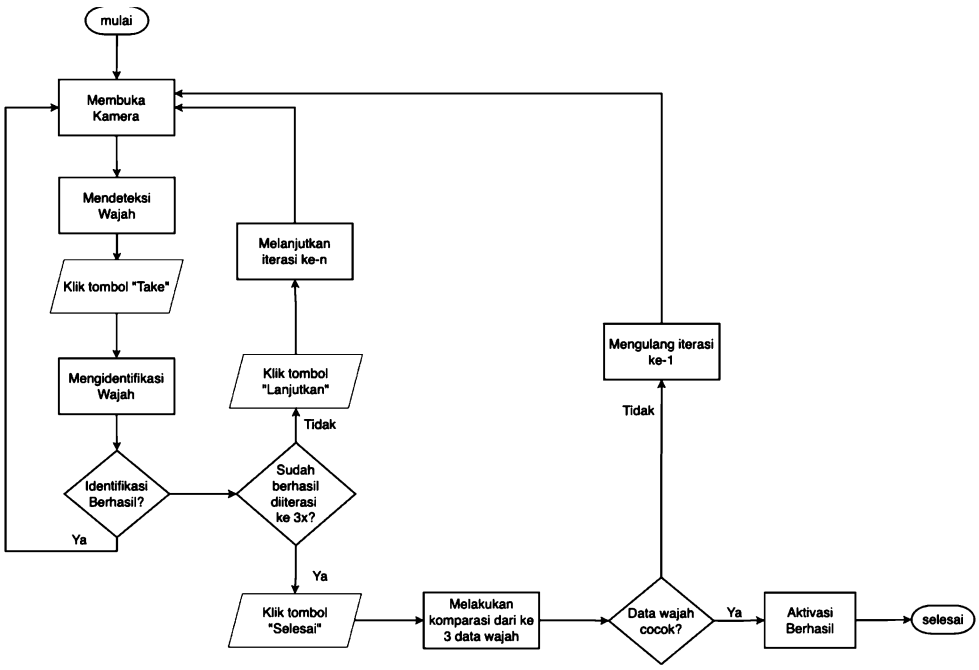


Gambar 4. Rancangan Proses Presensi

Ketika ingin melakukan proses presensi pada smartphone, karyawan memerlukan kredensial akun yang digunakan sebagai identitas karyawan. Gambar 3 merupakan proses dimana Manager HRD yang berperan sebagai Admin dapat mendaftarkan karyawan ke dalam sistem sehingga dapat melakukan proses presensi. Proses pendaftaran sendiri dilakukan menggunakan sistem berbasis web yang dapat diakses melalui peramban. Setelah proses menambahkan karyawan berhasil, Manajer HRD dapat memberikan kredensial akun berupa email dan password kepada karyawan terkait untuk bisa login pada sistem presensi mobile. Gambar 4 merupakan rancangan proses presensi yang dilakukan oleh karyawan.

Gambar 4 terlihat bahwa karyawan memerlukan akses login menggunakan akun yang telah didaftarkan oleh Manajer HRD untuk masuk ke dalam aplikasi dan menggunakan sistem. Saat pertama kali masuk ke dalam sistem, karyawan diwajibkan untuk melakukan aktivasi akun. Dapat dilihat juga pada gambar tersebut terdapat proses aktivasi, dimana pada proses tersebut sistem akan merekam wajah pengguna dan memproses hasil tersebut menggunakan Tensorflow Lite, kemudian menyimpannya ke dalam database dalam bentuk matriks.

Selanjutnya, Gambar 5 menjelaskan bahwa proses aktivasi secara detail dimulai dengan sistem yang akan membuka kamera pada smartphone dan kemudian merekam wajah dari pengguna. Bersamaan dengan itu, sistem akan melakukan proses deteksi pola wajah pengguna, selanjutnya pengguna dapat melakukan klik pada tombol "Take (n)". Proses perekaman wajah dilakukan sebanyak tiga kali, artinya sistem akan menyimpan data wajah ke dalam *database* banyak tiga data yangS disimpan dalam bentuk matriks (array/list). Sebelum disimpan ke dalam *database*, tepat diakhir proses data akan melalui tahap perbandingan dengan ketiga data wajah lainnya untuk memastikan apakah data cocok atau tidak, jika dari ketiga cocok maka data akan disimpan ke dalam database dan jika tidak maka proses aktivasi akan diulang kembali.



Gambar 5. Rancangan Proses Aktivasi

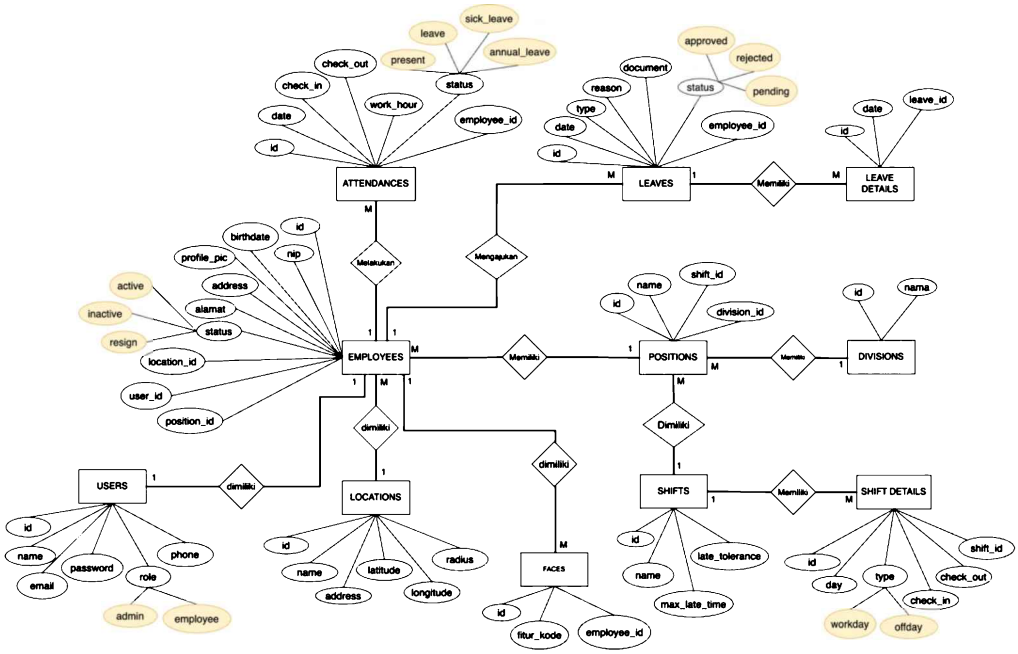
TABEL 1. SAMPEL DATA WAJAH & MATRIKS

No	Foto Wajah	Data Matriks
1		[-0.011336246505379677,0.01282558310776949,-1.2336687177594285e-5,0.0011170025682076812,-0.010225892998278141,-0.012273730710148811,-0.06202023848891258,-0.10159404575824738,-0.01650238409638405,-0.0883495882153511, ...]
2		[-0.014138611033558846,0.011630591936409473,-0.0012514255940914154,0.006366509012877941,-0.01586115173995495,0.0446758046746254,-0.1500401496887207,-0.04219534620642662,-0.09850174188613892, -0.1659150868654251, ...]
3		[-0.007853974588215351,0.013590618036687374,0.0027012510690838 1,-0.0029659175779670477,-0.017467981204390526,0.018725145608186722,-0.10152602195739746,-0.07299498468637466,-0.005309869069606066, 0.11888813227415085, ...]

Seperti yang sudah dijelaskan sebelumnya sekaligus dijelaskan pada Gambar 5, data wajah disimpan ke dalam database dalam bentuk matriks/array/list. Berikut merupakan sampel ketiga data wajah yang berhasil disimpan ke dalam database.

F. Entity Relationalship Diagram

Entity Relationalship Diagram atau disingkat ERD atau Desain Basis Data merupakan sebuah model yang menjelaskan hubungan antara data satu dengan data yang lain dalam sebuah basis data [1]. Basis data diperlukan sebagai tempat untuk menyimpan data pada sebuah sistem, sehingga manajemen data dapat dilakukan dengan mudah. Rancangan basis data untuk sistem presensi pada penelitian ini terdapat pada gambar 6 berikut ini.



Gambar 6. Rancangan Basis Data

III. HASIL DAN IMPLEMENTASI

Setelah tahap perancangan selesai dilakukan, tahap selanjutnya yaitu melakukan implementasi dan pengujian sistem. Dalam bab ini, pengujian sistem dilakukan menggunakan blackbox, dimana blackbox sendiri merupakan pengujian yang dilakukan dengan memberikan beberapa input data dan kemudian sistem akan mengembalikan output sesuai fungsionalitas yang telah ditentukan.

A. *Implementasi Sistem*

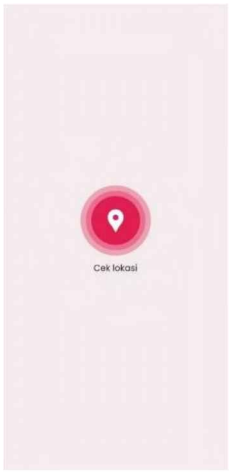
Tahap ini merupakan tahap implementasi sistem berdasarkan perancangan yang telah dilakukan pada tahap sebelumnya. Akan diberikan beberapa sampel hasil implementasi sistem dalam bentuk antarmuka.

1) Proses Aktivasi: Proses aktivasi merupakan proses yang bertujuan untuk merekam data wajah menggunakan kamera depan smartphone, memproses hasil rekaman menjadi sebuah matriks dan kemudian menyimpannya ke dalam database. Gambar 7 merupakan tampilan antarmuka proses aktivasi yang terdapat dalam sistem ini.

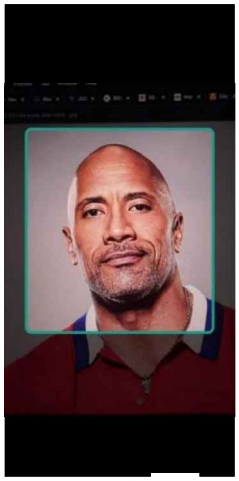
2) Proses Deteksi Lokasi: Proses ini bertujuan untuk mendeteksi lokasi karyawan sebelum melakukan proses presensi, output dari proses ini akan menghasilkan latitude dan longitude yang digunakan untuk melakukan perbandingan jarak dengan lokasi yang ditentukan oleh perusahaan. Tampilan antarmuka proses deteksi lokasi terdapat dilihat pada gambar 8.

3) Proses Presensi: Proses presensi dilakukan sama persis dengan proses aktivasi, yaitu dengan manghadapkan wajah ke kamera smartphone. Tanda kotak berwarna hijau menandakan bahwa sistem telah berhasil mendeteksi wajah karyawan dan kemudian dilanjutkan untuk melakukan pengenalan wajah. Proses presensi berjalan secara otomatis dengan tidak menggunakan tombol untuk melakukan presensi. Ketika sistem berhasil mengenali wajah karyawan, presensi akan otomatis berhasil. Gambar 9 merupakan hasil proses presensi.

Gambar 7. Antarmuka proses aktivasi



Gambar 8. Antarmuka Proses Deteksi Lokasi



Gambar 9. Antarmuka Proses Presensi

B. Pengujian Blackbox

Tahap ini merupakan tahap dimana sistem yang sudah dibuat akan dilakukan uji menggunakan blackbox. Tujuan dari pengujian ini yaitu mendeteksi dan mengidentifikasi potensi kesalahan yang mungkin terjadi pada aplikasi sebelum digunakan oleh user secara umum. Pengujian blackbox yang dilakukan dapat dilihat pada tabel 2.

Dari hasil sebanyak 13 skenario pengujian yang telah dilakukan, terdapat total 10 pengujian yang dinyatakan berhasil dan 3 pengujian yang dinyatakan gagal. Hasil tersebut merupakah hasil yang cukup tinggi, akan tetapi belum sepenuhnya sempurna karena masih terdapat pengujian yang gagal.

TABEL 2. PENGUJIAN BLACKBOX

No	Skenario	Hasil ynag Diharapkan	Hasil Pengujian	Status
1	Klik tombol aktivasi pada halaman aktivasi	Masuk ke dalam proses aktivasi	Sistem mengarahkan ke dalam halaman proses aktivasi.	Berhasil
2	Klik tombol “Next” pada saat aktivasi akun	Menampilkan bottom sheet dan tampil tombol lanjutkan	Sistem menampilakan bottom sheet dan menampilkan tombol lanjutkan.	Berhasil
3	Klik tombol masuk pada halaman home untuk presensi masuk	Masuk ke dalam pengecekan lokasi sebelum proses presensi masuk.	Sistem mengarahkan ke halaman proses pengecekan lokasi atau deteksi lokasi sebelum melakukan presensi masuk.	Berhasil
4	Klik tombol keluar pada halaman home untuk presensi keluar	Masuk ke dalam pengecekan lokasi sebelum proses presensi keluar.	Sistem mengarahakn ke halaman proses pengecekan lokasi atau deteksi lokasi sebelum melakukan presensi keluar.	Berhasil
5	Mematikan lokasi ketika melakukan presensi	Menampilkan pesan “Lokasi tidak aktif”	Sistem menampilkan pesan “Lokasi tidak aktif” ketika pengguna mematikan lokasi pada smartphone.	Berhasil
6	Menyalakan lokasi pada ponsel ketika ingin presensi (masuk/keluar)	Masuk ke dalam proses presensi	Sistem mengarahkan ke halaman proses presensi.	Berhasil
7	Melakukan presensi dengan wajah yang sama	Presensi diterima/berhasil	Sistem berhasil memproses presensi (presensi berhasil)	Berhasil
8	Melakukan presensi di dalam ruangan	Presensi diterima/berhasil	Sistem berhasil memproses presensi (presensi berhasil)	Berhasil
9	Melakukan presensi dengan wajah yang berbeda	Presensi tidak berhasil	Sistem tidak berhasil memproses presensi (gagal) karena data wajah tidak valid.	Berhasil
10	Melakukan presensi menggunakan foto yang dicetak	Presensi tidak berhasil	Sistem berhasil memproses presensi karena sistem menganggap data wajah valid, walaupun data tersebut merupakan foto yang dicetak, tetap diaggap valid oleh sistem.	Gagal
11	Melakukan presensi pada tempat terang	Presensi berhasil	Sistem berhasil memproses presensi.	Berhasil
12	Melakukan presensi pada tempat yang redup	Presensi berhasil	Presensi tidak dapat diproses, karena minim cahaya, sehingga sistem tidak dapat membaca fitur-fitur pada wajah dalam kondisi cahaya redup.	Gagal
13	Melakukan presensi ditempat yang gelap	Presensi berhasil	Presensi tidak dapat diproses karena tempat yang gelap dan tidak ada cahaya, sehingga sistem tidak dapat memproses data presensi.	Gagal

IV. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan di PT Nusantara Berkah Digital, dapat disimpulkan bahwa sistem presensi yang menggabungkan metode pengenalan wajah dan teknologi *geofencing* dapat menjadi solusi atas permasalahan yang terjadi di Perusahaan pada proses presensi. Setiap karyawan memiliki akun akses individu pada aplikasi mobile yang memungkinkan mereka melihat data pribadi mereka sendiri, sehingga tidak ada lagi kemungkinan bagi mereka untuk mengakses data pribadi karyawan lain.

Sementara, dari hasil pengujian yang telah dilakukan dapat disimpulkan sebagai berikut:

1. Sistem presensi yang telah dibuat menunjukkan hasil keberhasilan sebanyak 76,92% pada pengujian blackbox dengan total benyak 13 skenario.
2. Dalam beberapa skenario pengujian terbukti bawah metode pengenalan wajah yang diterapkan pada sistem presensi berfungsi dengan baik karena ketika wajah karyawan dibandingkan dengan wajah karyawan lain, sistem tidak dapat memproses presensi.
3. Dengan diterapkannya geofencing pada sistem presensi juga terbukti bisa membatasi area lokasi yang hanya diizinkan oleh Perusahaan, sehingga kecurangan seperti titip absen dan kecurangan lainnya saat proses presensi dapat dicegah karena karyawan diwajibkan datang ke kantor ketika melakukan presensi.

REFERENSI

[1] Astutik, I. R. I., & Rosid, M. A. (2020). Basis Data (M. Suryawinata, Ed.; 1st ed.). UMSIDA Press.

[2] Budianto, A. (2019). Learning Android and Cyber Counseling (1st ed.). Media Nusa Creative.

[3] Butsianto, S., & Naya, C. (2023). Model Aplikasi Human Resource Management Sistem (HRIS) Dengan Framework UniGui. Bulletin of Information Technology (BIT), 4(1), 81–88. https://doi.org/10.47065/bit.v3i1

[4] Hasanuddin, Asgar, H., & Hartono, B. (2022). Rancang Bangun REST API Aplikasi Weshare Sebagai Upaya Mempermudah Pelayanan Donasi Kemanusiaan. Jurnal Informatika Teknologi Dan Sains, 4(1), 8–14. https://doi.org/10.51401/jinteks.v4i1.1474

[5] Mishra, A. (2020). Machine Learning for IOS Developers. Jhon Wiley & Sons, Inc.

[6] Palupi, R., Yulianna, D. A., & Winarsih, S. S. (2021). Analisa Perbandingan Rumus Haversine Dan Rumus Euclidean Berbasis Sistem Informasi Geografis Menggunakan Metode Independent Sample t-Test. JITU : Journal Informatic Technology And Communication, 5(1), 40–47. https://doi.org/10.36596/jitu.v5i1.494

[7] Permana, I. G. T., Rusdianto, D. S., & Fanani, L. (2019). Pengembangan Sistem Presensi berbasis Lokasi menggunakan Geofence WiFi dan REST API pada Fakultas Ilmu Komputer Universitas Brawijaya. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 3(9), 9305–9313. http://j-ptiik.ub.ac.id

[8] Pribadi, W. W., Yunus, A., & Sartika Wiguna, A. (2022). PERBANDINGAN METODE K-MEANS EUCLIDEAN DISTANCE DAN MANHATTAN DISTANCE PADA PENENTUAN ZONASI COVID-19 DI KABUPATEN MALANG. Jurnal Mahasiswa Teknik Informatika, 6(2), 493–500. https://doi.org/10.36040/jati.v6i2.4808

[9] Putra, Y. W. S., & Adhim, M. F. (2022). Sistem Informasi Presensi Online Menggunakan Teknologi Face Recognition dan GPS. Jurnal Tekno Kompak, 16(1), 149–161. https://doi.org/10.33365/jtk.v16i1.1470

[10] Rohmat, C. L., & Nuriyah, R. (2023). IMPLEMENTASI HUMAN RESOURCE INFORMATION SYSTEM BERBASIS WEBSITE PADA PT LITEDEX DIGITAL INDONESIA. Jurnal Mahasiswa Teknik Informatika, 7(1), 720–726. https://doi.org/10.36040/jati.v7i1

[11] Sari, I. P., Purnama, I., & Ritonga, A. A. (2021). Implementasi API pada Aplikasi Al-Qur'an Berbasis Android dengan Metode UCD. JURNAL MEDIA INFORMATIKA BUDIDARMA, 5(2), 615. https://doi.org/10.30865/mib.v5i2.2913

[12] Singh, A., & Bhadani, R. (2020). Mobile Deep Learning with TensorFlow Lite, ML Kit and Flutter (A. Hoda, Ed.). Packt Publishing Ltd. https://books.google.co.id/books?id=sbTbDwAAQBAJ

[13] Sunarya, F., & Hardyanto, C. (2021). Implementasi Face Recognition Dan Global Positioning System Pada Sistem Presensi Di Desa Mekarjati Kab Indramayu Berbasis Mobile. Jurnal Penelitian Mahasiswa Teknik Dan Ilmu Komputer, 1(2), 52–60. https://doi.org/10.34010/JUPITER.V1I2.6550

[14] Suryansah, A., Habibi, R., & Awangga, R. M. (2020). *Penggunaan Face Recognition untuk akses ruangan* (R. M. Awangga, Ed.). Kreatif Industri Nusantara.

[15] Zhao, J., & Kumar, V. V. (2021). Handbook of Research on Innovations and Applications of AI, IoT, and Cognitive Technologies. IGI Global.

[16] https://doi.org/10.4018/978-1-7998-6870-5

Sistem Informasi Manajemen Panti Asuhan Berbasis Web pada Panti Asuhan Al Dzikro

Baharudin
Abdulloh Mun'im
Fakultas Teknik dan Informatika
Universitas Bina Sarana Informatika
Jakarta, Indonesia
abdullohbahar@gmail.com

Anik Andriani
Fakultas Teknik dan Informatika
Universitas Bina Sarana Informatika
Jakarta, Indonesia
anik.aai@bsi.ac.id

Chriswardana Bayu Dewa
Fakultas Teknik dan Informatika
Universitas Bina Sarana Informatika
Jakarta, Indonesia
chriswardana.chb @bsi.ac.id

Abstrak—Sistem manajemen di panti asuhan yang masih berjalan secara konvensional rentan terhadap resiko kerusakan data, kehilangan data, ketidaksesuaian dalam penginputan salah satunya karena duplikat data, serta membutuhkan waktu lama dalam pengelolaan seperti pencarian data. Tujuan dari penelitian ini adalah membangun sistem informasi manajemen panti asuhan berbasis web. Metode yang digunakan untuk pengembangan sistem informasi tersebut adalah metode Rapid Application Development atau RAD yang cocok digunakan untuk pengembangan perangkat lunak dengan waktu yang singkat. Tahapan pengembangan sistem informasi panti asuhan ini menerapkan tahapan-tahapan dalam siklus metode RAD. Hasilnya diperoleh sistem informasi panti asuhan yang mampu mengelola data takmir atau petugas, data donatur beserta data donasinya, data penggunaan donasi, pengelolaan data aliran kas, dan laporan. Sistem informasi yang dibangun memiliki performa yang lebih baik daripada sistem sebelumnya.

Kata Kunci—sistem informasi, panti asuhan, RAD, website

I. PENDAHULUAN

Panti Asuhan merupakan suatu lembaga yang memiliki kewenangan dan tanggung jawab dalam bergerak di bidang sosial dengan memberikan layanan kepada anak terlantar berupa kesejahteraan sosial [1]. Pengelolaan Panti Asuhan bukanlah hal yang mudah. Terdapat data-data yang sangat penting seperti data anak asuh, data donator, data donasi, data keuangan, dan data surat masuk serta surat keluar. Banyaknya data yang harus dikelola dalam pengelolaan sebuah Panti Asuhan, oleh karena itu diperlukan manajemen yang baik dalam sebuah Panti Asuhan. Namun tidak semua manajemen pengelolaan data sebuah Panti Asuhan sudah dijalankan dengan baik. Salah satu dari sekian banyak Panti Asuhan di daerah Bantul, Yogyakarta adalah Panti Asuhan Al Dzikro. Saat ini pengelolaan data di panti asuhan tersebut masih dilakukan secara konvensional yaitu dengan pencatatan pada buku. Dampaknya pengelolaan data beresiko pada kerusakan data, kehilangan data, dan tidak ada backup. Selain itu dalam penginputan data beresiko terjadi anomali data seperti kesalahan penginputan, data yang duplikat, serta data yang missing value. Selain itu pengelolaan data pada suatu instansi yang dikelola dengan sistem konvensional dapat berdampak pada lamanya waktu dalam penyusunan laporan [2].

Beberapa penelitian terdahulu terkait pengembangan sistem informasi di Panti Asuhan antara lain "Sistem Informasi Pengelolaan Data Panti Asuhan (Studi Kasus Panti Asuhan Maria Visitasi Nebe)" yang membangun sistem informasi untuk pengelolaan data anak asuh, data donatur, data bantuan, data petugas, dan laporan [3]. Pada penelitian tersebut belum tersedia fasilitas untuk mengelola data donasi yang masuk dan yang digunakan serta data aliran keuangan di Panti Asuhan. Penelitian lain berjudul "Rancang Bangun Sistem Informasi Manajemen Panti Asuhan Al-Kahfi Surabaya" mengembangkan sistem informasi berbasis web untuk manajemen panti asuhan. Pada sistem informasi yang dibangun berisi tentang informasi seputar panti asuhan, data anak asuh, data donatur, data donasi, dan artikel-artikel beserta galeri foto seputar kegiatan yang dilakukan panti asuhan Al-Kahfi tersebut. Pada sistem informasi manajemen panti asuhan yang dibangun ini belum menyediakan fitur-fitur yang menyajikan data donasi masuk dan penggunaannya, data aliran kas, dan laporan setiap data yang terlibat dalam sistem informasi yang dibangun [4].

Lemahnya monitoring pada penggunaan data donasi memiliki resiko kesalahan pengelolaan data, resiko kehilangan maupun kerusakan data, resiko kontrol data, serta waktu lama dalam menghitung dan membuat laporan keuangan. Berdasarkan pada latar belakang tersebut, maka penelitian ini bertujuan membangun sistem informasi manajemen Panti Asuhan yang memiliki fitur untuk mengelola data anak asuh, donatur, donasi, penggunaan donasi, sampai pada aliran kas dari Panti Asuhan Al Dzikro.

II. METODE PENELITIAN

A. *Metode Analisis Masalah*

Analisis masalah diperlukan untuk menentukan masalah utama pada sistem yang sudah berjalan, sehingga dapat ditentukan solusi permasalahan yang diterapkan pada sistem informasi yang diusulkan. Metode analisis masalah yang digunakan pada penelitian ini adalah metode PIECES. Tabel 1 menunjukkan hasil analisis masalah menggunakan metode PIECES.

TABEL 1. ANALISIS PIECES

Jenis Analisis	Kelemahan Sistem Lama	Sistem yang Diusulkan
Performance	Pengelolaan data rentan terjadi anomali data saat penginputan data. Penghitungan arus kas, pencarian data	Sistem informasi yang diusulkan terkoneksi dengan basis data dan sistem memberikan

	dan pembuatan laporan butuh waktu lama untuk merekap data.	validasi-validasi untuk antaramuka-antarmuka input data. Sistem yang diusulkan memberikan fasilitas kemudahan untuk perhitungan arus kas, pembuatan laporan.
Information	Informasi data donasi dan penggunaan donasi beserta arus keuangan hanya disampaikan saat rapat pengurus	Sistem informasi manajemen panti asuhan memberikan informasi data donasi beserta arus keuangan yang dapat diakses kapanpun dan dimanapun oleh semua pengurus
Economic	Biaya untuk pencatatan dan perekapan data anak asuh, data donatur, data donasi, data penggunaan donasi, data keuangan cukup besar untuk pembelian buku catatan dan alat tulis yang harus dilakukan secara terus menerus dalam jangka waktu panjang	Alokasi biaya hanya di awal untuk biaya perangkat komputer, wifi, domain dan hosting sistem informasi. Sedangkan perawatan hanya dikeluarkan sekali setahun seperti untuk perpanjangan domain dan hosting
Control	Kesalahan penginputan maupun kesalahan dalam pengolahan data akan sulit dikontrol pada sistem manual.	Sistem informasi yang diusulkan terkoneksi dengan database sehingga mempermudah dalam pengontrolan data.
Efficiency	Perekapan data membutuhkan waktu lama karena harus dicatat di buku rekapan. Selain itu perhitungan arus keuangan serta pembuatan laporan juga membutuhkan waktu lama dalam pengerjaannya	Sistem informasi yang diusulkan memberikan kemudahan dalam pencatatan data dan pengolahan data membutuhkan waktu yang singkat terutama dalam pembuatan laporan karena laporan otomatis tersedia
Services	Pelayanan terkait penyampaian informasi arus keuangan maupun laporan kepada Yayasan harus menunggu direkap oleh petugas.	Pelayanan keuangan dapat didownload sewaktu-waktu bila pengurus atau Yayasan membutuhkan laporan tersebut

B. Metode Pengumpulan Data

Teknik pengumpulan data yang digunakan dalam penelitian ini yaitu observasi dan wawancara. Teknik pengumpulan data yang pertama adalah Observasi yang merupakan salah satu teknik pengumpulan data yang dilakukan dengan cara melihat langsung kegiatan yang dilakukan pengguna sistem [5]. Kegiatan observasi dalam penelitian ini dilakukan dengan mengamati bisnis proses dalam sistem manajemen di Panti Asuhan Al Dzikro yang meliputi manajemen donatur, donasi, dan penggunaan dana. Teknik pengumpulan data yang kedua yaitu wawancara. Wawancara merupakan teknik pengumpulan data yang dilakukan dengan melakukan tanya jawab secara tatap muka dan langsung kepada narasumber [6]. Wawancara dilakukan dengan pengurus panti asuhan Al Dzikro. Hasilnya diperoleh data tentang sejarah, profil, struktur organisasi, dan bisnis proses dari sistem pengelolaan data donasi di panti asuhan Al Dzikro.

C. Metode Pengembangan Perangkat Lunak

Metode Rapid Application Development atau RAD merupakan metode pengembangan perangkat lunak yang digunakan dalam pengembangan sistem informasi manajemen di Panti Asuhan Al Dzikro. Metode RAD merupakan metode yang popular dan cocok digunakan untuk pengembangan perangkat lunak dalam waktu singkat [7]. Metode ini termasuk dalam model pengembangan perangkat lunak dengan proses sekuensial linear dan menekankan pada siklus yang pendek dengan lama waktu sekitar 60 sampai dengan 90 hari. Gambar 1 menjelaskan tentang tahapan-tahapn pada metode RAD yang terdiri dari *Requirement planning, User design, Construction,* dan *Implementation* [8].

1) Requirement planning: Tahap ini merupakan tahap definisi konsep. Tahap ini bertujuan mendefinisikan konsep dari fungsi-fungsi bisnis serta data pada area subjek yang akan didukung oleh sistem yang akan dibangun. Selain itu pada tahap ini juga bertujuan menentukan ruang lingkup dari sistem.

2) User Design: Tahap ini merupakan tahap desain fungsional. Pada tahap ini diterapkan pemodelan sistem data dan pemodelan proses dalam membangun prototipe dari komponen sistem yang penting

3) Construction: Tahap ini merupakan tahap pengembangan. Pada tahap ini dilakukan beberapa kegiatan yaitu penyelesaian pembangunan fisik dari sistem aplikasi, membangun sistem konversi, dan mengembangkan berdasarkan masukan penggunan serta mengimplementasikan rencana kerja.

4) Implementation: Tahap ini disebut juga tahap penerapan. Pada tahap ini dilakukan pengujian pada sistem yang dikembangkan dan pelatihan akhir untuk calon

pengguna, konversi data, serta implementasi sistem informasi yang telah selesai dikembangkan.

III. HASIL DAN PEMBAHASAN

Mengadopsi tahapan-tahapan pada metode RAD, pengembangan sistem informasi Panti Asuhan Al Dzikro diperoleh hasil berikut:

A. *Requirement planning*

Ruang lingkup pada sistem informasi Panti Asuhan Al Dzikro yang dikembangkan terdiri dari antarmuka *back-end*. Pada antarmuka back-end berupa dashboard yang disediakan untuk pengurus Panti Asuhan. Adapun pada dashboard tersebut memiliki beberapa fitur berupa, data anak asuh, data donatur, data donasi, data pengurus, data keuangan, data surat masuk, dan data surat keluar. Sistem ini difokuskan untuk memberikan kemudahan kepada pihak panti asuhan dalam pengolahan atau manajemen data.

Berdasarkan ruang lingkup yang ditetapkan pada sistem informasi panti asuhan yang akan dikembangkan, dilakukan analisis kebutuhan. Analisis terhadap kebutuhan pengguna bertujuan sebagai acuan pengembangan fitur-fitur yang akan dibangun agar sesuai dengan apa yang diharapkan atau dibutuhkan oleh pengguna dan mempermudah dalam perancangan sistem informasi tersebut. Sistem informasi manajemen panti asuhan memiliki empat tipe pengguna, yaitu Admin Yayasan, Admin Keuangan, Admin LKSA, dan Admin Donasi. Hasil analisis kebutuhan pengguna yang dirumuskan berdasarkan empat tipe pengguna yaitu:

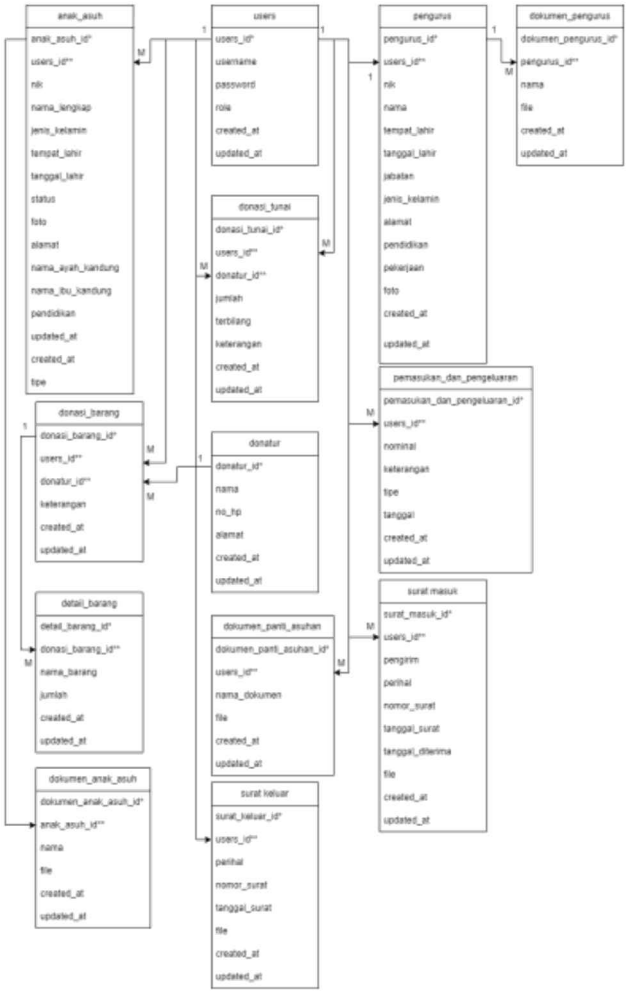
1) Admin Yayasan: Pengguna ini berperan sebagai *super admin* yang memiliki akses utama yaitu mengelola data admin operator yang nantinya berperan sebagai *user* pengelola data manajemen panti asuhan. Admin Yayasan memiliki beberapa hak akses meliputi login ke halaman dashboard admin, mengelola data user atau admin operator yang terdiri dari admin keuangan, admin LKSA, dan admin donasi, mengelola data pengurus, melihat dan mencetak data donasi barang maupun data donasi tunai, melihat dan mencetak data anak asuh, melihat dan mencetak pemasukan dan pengeluaran, mengelola data surat masuk dan keluar.

2) Admin Keuangan: pengguna ini berperan mengontrol arus kas dari panti asuhan, sehingga Admin Keuangan memiliki akses berupa login ke halaman dashboard admin, mengelola data pemasukan dan pengeluaran panti asuhan beserta membuat laporan data keuangan.

3) Admin LKSA: pengguna ini bertanggung jawab pada data anak asuh. Pada sistem informasi manajemen panti asuhan yang dikembangkan Admin LKSA memiliki hak akses berupa login ke halaman dashboard admin, mengelola data anak asuh dan pembuatan laporan data anak asuh.

4) Admin Donasi: pengguna ini bertanggung jawab pada pengelolaan data donatur beserta data donasi. Pada sistem informasi manajemen panti asuhan yang dikembangkan pengguna ini memiliki hak akses berupa login ke halaman dashboard admin, mengelola data donatur dan data donasi baik yang berupa donasi tunai mauapun donasi barang. Admin donasi juga mengelola data penggunaan donasi yang sudah masuk ke panti asuhan.

Selain analisis kebutuhan pengguna dilakukan analisis kebutuhan sistem. Analisis kebutuhan sistem dibagi menjadi Operasional, Kinerja, dan Keamanan. Pada kebutuhan operasional sistem informasi yang diusulkan mampu dijalankan pada sistem operasi Windows, minimal spesifikasi komputer yaitu processor core 2, sistem informasi dapat diakses melalui berbagai macam browser seperti Mozilla Firefox, Microsoft Edge, Google Chrome melalui PC maupun smartphone. Pada kebutuhan kinerja sistem informasi usulan sistem memiliki tampilan yang mudah dipahami oleh pengguna, sistem dapat diakses oleh lebih dari satu pengguna secara bersamaan, sistem hanya dapat diakses oleh empat tipe pengguna yaitu Admin Yayasan, Admin Keuangan, Admin LKSA, dan Admin Donasi. Kebutuhan sistem informasi usulan pada keamanan meliputi tersedianya username dan password yang terinkripsi di basis data yang digunakan pengguna untuk *login* ke dalam sistem informasi.



Gambar 1. Desain basis data panti asuhan

B. *User Design*

Pemodelan sistem data diimplementasikan dalam perancangan basis data. Desain basis data pada sistem informasi panti asuhan digunakan sebagai acuan dalam pengembangan basis data. Gambar 1 menunjukkan desain basis data panti asuhan. Basis data tersebut tersusun dari 13 tabel antara lain tabel user, pengurus, anak_asuh, dokumen_pengurus, donasi_tunai, donasi_barang, donatur,

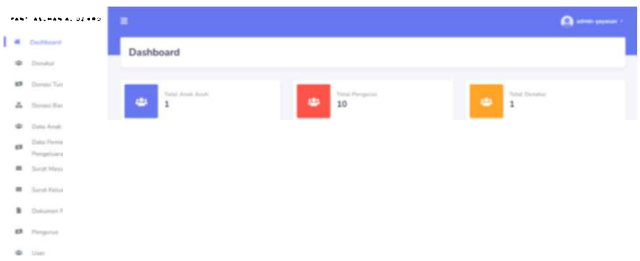
pemasukan_dan_pengeluaran, detail_barang, dokumen_panti_asuhan, dokumen_anak_asuh, surat_masuk, dan surat_keluar.

C. *Construction*

Pembangunan fisik sistem informasi usulan meliputi membangunan antarmuka sistem yang meliputi antarmuka login untuk semua pengguna, antarmuka Admin Yayasan, Admin Keuangan, Admin LKSA, dan Admin Donasi. Gambar 2 menunjukkan halaman login untuk semua pengguna. Pengguna dapat menginputkan *username* dan *password* masing-masing pada antarmuka tersebut.

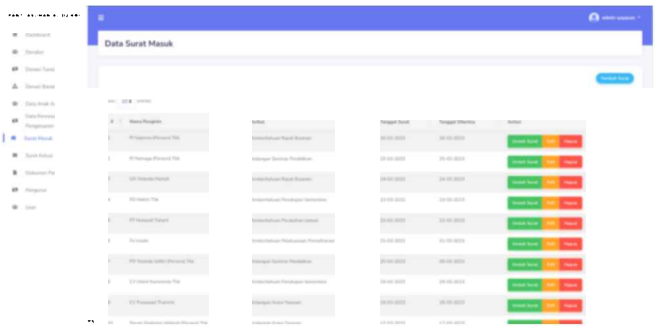
Gambar 2. Halaman login

Gambar 3 menunjukkan halaman admin Yayasan yang merupakan fitur untuk pengurus Yayasan Panti Asuhan Al Dzikro yang berperan sebagai super admin. Fitur-fitur pada dashboard Yayasan memiliki hak akses untuk melakukan monitoring semua data dalam sistem informasi. Fitur-fitur yhang tersedia di dalam dashboard Yayasan antara lain fitur pengelolaan data donatur, data donasi tunai, data donasi barang, data anak asuh, data pemasukan dan pengeluaran, data surat masuk, data surat keluar, dokumen panti asuhan, data pengurus, dan terdapat fitur untuk mengelola data user yang memiliki hak akses mengelola sistem informasi panti asuhan. Fitur pengelolaan user tersebut dapat digunakan untuk menambahkan akun pengurus sebagai admin operator pada sistem informasi tersebut.



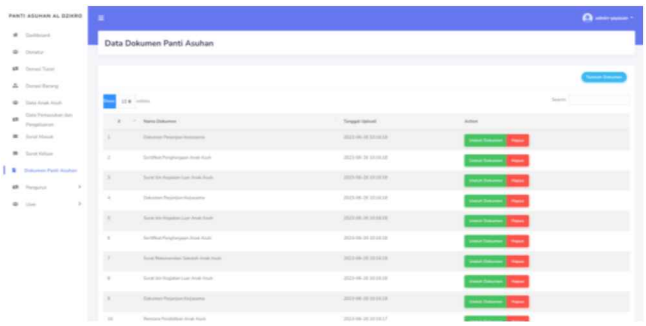
Gambar 3. Halaman dashboard Admin Yayasan

Salah satu fitur yang penting dalam hak akses untuk Admin Yayasan adalah adanya fasilitas untuk mengelola surat masuk dan surat keluar. Gambar 4 menunjukkan antarmuka untuk fasilitas pengelolaan surat pada panti asuhan.



Gambar 4. Halaman pengelolaan surat

Berdasarkan Gambar 4 dapat dilihat pada antarmuka pengelolaan surat meliput input data surat masuk dan surat keluar, edit data surat masuk dan surat keluar, serta hapus data surat masuk dan surat keluar.



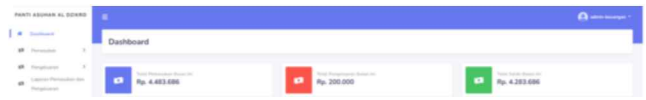
Gambar 5. Halaman pengelolaan dokumen panti asuhan

Selain fitur pengelolaan data surat masuk dan surat keluar terdapat fitur lain yang juga penting yaitu fitur pengelolaan data dokumen panti asuhan. Gambar 5 menunjukkan antarmuka untuk mengelola data dokumen-dokumen panti asuhan. Fasilitas yang disediakan pada antarmuka pengelolaan dokumen panti asuhan berupa fasilitas upload dokumen dan hapus dokumen.

Gambar 6. Halaman pengelolaan akun petugas

Hak akses pengguna sistem informasi panti asuhan perlu diatur untuk mengantisipasi hal-hal seperti pengurus yang non aktif atau mengundurkan diri maupun adanya pengurus

baru yang bertanggung jawab di bagian keuangan atau donasi, maupun bagian LKSA. Fitur pengelolaan akun untuk pengguna atau petugas yang mengelola sistem informasi panti asuhan tersebut ada di bawah kendali Admin Yayasan sebagai super admin. Gambar 6 menunjukkan antarmuka pengelolaan akun petugas meliputi pengaturan data pengurus yang diberi akses sekaligus pengaturan *username* dan *password*.

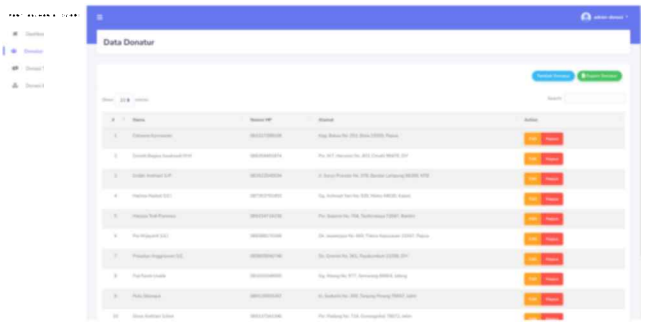


Gambar 7. Halaman dashboard Admin Keuangan

Admin keuangan bertugas mengelola data keuangan. Gambar 7 menunjukkan antarmuka dashboard untuk Admin Keuangan. Fasilitas yang tersedia pada dashboard Admin Keuangan meliputi fasilitas untuk menambah data pemasukan, data pengeluaran, serta data aliran keuangan.

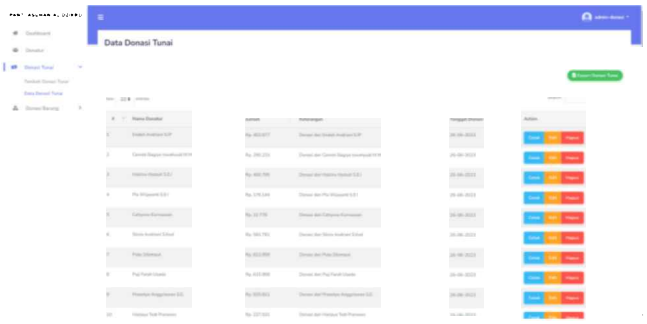
Gambar 8. Halaman data pemasukan dan pengeluaran donasi

Pada antarmuka data keuangan yang masuk dan data keuangan yang digunakan atau dikeluarkan ditunjukkan pada Gambar 8. Terdapat fasilitas untuk lihat data dan export data ke dalam bentuk file Excel yang dapat didownload dan dicetak.



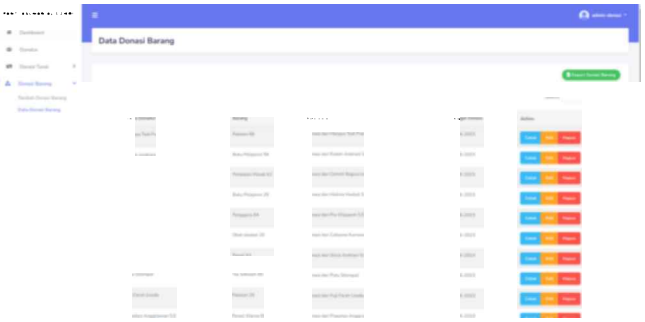
Gambar 9. Halaman dashboard Admin Donasi

Admin Donasi bertugas mengelola data donasi. Donasi yang masuk di panti asuhan meliputi donasi dalam bentuk tunai dan donasi dalam bentuk barang. Gambar 9 menunjukkan dashboard untuk Admin Donasi yang memiliki fasilitas untuk mengelola data donatur, data donasi tunai, dan data donasi barang.



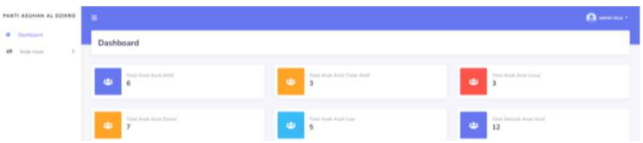
Gambar 10. Halaman data donasi tunai

Pada antarmuka data donasi tunai yang ditunjukkan Gambar 10 memiliki fasilitas untuk menambah data, mengedit data, dan menghapus data.



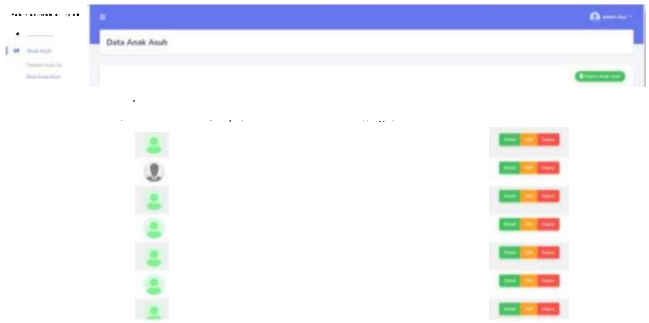
Gambar 11. Halaman data donasi barang

Pada antarmuka data donasi barang yang ditunjukkan Gambar 11 memiliki fasilitas untuk menambah data, mengedit data, dan menghapus data.



Gambar 12. Halaman dashboard Admin LKSA

Admin LKSA memiliki tugas mengelola data anak asuh dan dokumen-dokumen anak asuh. Dashboard Admin LKSA dan fasilitas-fasilitas yang tersedia pada akses Admin LKSA ditunjukkan Gambar 12.



Gambar 13. Halaman data anak asuh

Pada antarmuka data anak asuh yang ditunjukkan Gambar 13 memiliki fasilitas untuk menambah data, mengedit data, dan menghapus data.

D. Implementation

Implementasi sistem informasi usulan dilakukan dengan melakukan pengujian unit oleh pengguna menggunakan metode Black Box Testing. Pengujian terhadap calon pengguna sistem informasi usulan tersebut dilakukan oleh pengurus yayasan, bagian keuangan, bagian donasi, dan bagian LKSA. Tabel 2 menunjukkan kesimpulan data hasil pengujian unit oleh calon pengguna terhadap semua fitur yang tersedia pada sistem informasi manajemen panti asuhan.

TABEL 2. HASIL PENGUJIAN UNIT

Unit	Penguji	Pengujian	Hasil pengujian
Login	– Admin Yayasan – Admin Keuangan – Admin LKSA – Admin Donasi	– Validasi inputan kosong – Validasi username dan password tidak sesuai – Validasi username dan password sesuai	Sesuai harapan
Pengelolaan surat	Admin Yayasan	– Validasi data inputan – Simpan data – Edit data – Hapus data	Sesuai harapan
Pengelolaan Dokumen Panti	Admin Yayasan	– Validasi data inputan – Simpan data – Edit data – Hapus data	Sesuai harapan
Pengelolaan data pengurus	Admin Yayasan	– Validasi data inputan – Simpan data – Edit data – Hapus data	Sesuai harapan
Pengelolaan data keuangan	Admin Keuangan	– Validasi data inputan – Simpan data – Edit data – Hapus data – Export file Excel	Sesuai harapan
Pengelolaan data anak asuh	Admin LKSA	– Validasi data inputan – Simpan data – Edit data – Hapus data	Sesuai harapan
Pengelolaan data donatur	Admin Donasi	– Validasi data inputan – Simpan data – Edit data – Hapus data	Sesuai harapan
Pengelolaan data donasi	Admin Donasi	– Validasi data inputan – Simpan data – Edit data – Hapus data	Sesuai harapan

IV. KESIMPULAN

Sistem informasi manajemen panti asuhan memberikan fasilitas lengkap untuk membantu tugas dan tanggung jawab dari pengurus Yayasan, bagian keuangan, bagian LKSA, dan bagian donasi. Kelebihan dari sistem informasi yang diusulkan antara lain mengurangi resiko adanya kesalahan dalam pengelolaan data dengan adanya validasi-validasi pada setiap proses input data, mengurangi resiko kehilangan data atau kerusakan data karena data yang telah diinputkan tersimpan di dalam basis data, mengurangi kelemahan sistem lama yaitu pada lamanya proses perhitungan arus keuangan dan pembuatan laporan karena data keuangan dapat dilihat kapanpun dan dapat dicetak berdasarkan data keuangan yang telah diinputkan. Hasil pengujian unit dari setiap fitur-fitur yang ada di dalam sistem usulan oleh para calon pengguna menunjukkan hasil yang sudah sesuai harapan.

REFERENSI

[1] S. Akhmad, and Purwaningsih, "Sistem Informasi Manajemen Administrasi Keuangan Panti Asuhan Berbasis Website," Jurnal Responsif, vol. 2, pp. 150–157, 2020.

[2] O. Intan, S. Sri, W. Pipin, "Penerapan Metode PIECES pada Analisis Sistem Informasi Manajemen Apotek," Jurnal Infokes, vol. 11, pp 54-58, 2021.

[3] N. Regina, Rozady. Margaretha, S. Agustinus, "Sistem Informasi Pengolahan Data Panti Asuhan (Studi Kasus Panti Asuhan Maria Visitasi Nebe)," Jurnal In Create, 2023, pp. 46-52.

[4] Y. Andhik, T. Hasyier, A. Afaudin, L. Ammarzain, Fatema, "Rancang Bangun Sistem Informasi Manajemen Panti Asuhan Al-Kahfi Surabaya," Scan, 2020, pp 1-5.

[5] S. E, "Rancang Bangun Sistem Informasi Manajemen Data Mahasiswa dan Dosen Terintegrasi," Journal Reseach and Development, 2020.

[6] P. Titania, and Zulfachmi, "Survey Paper: Perbandingan Metode Pengembangan Perangkat Lunak (Waterfall, Prototype, RAD)." 2021.

[7] A. Anik, and Q. Esty. "Sistem Informasi Penjualan Pada Toko Online Dengan Metode Rapid Application Development (RAD)," 2018.

Algoritma CNN-LSTM untuk Memprediksi Tingkat Pencemaran Udara

Bonifasius Mamerutama
Departemen Informatika
Universitas Sanata Dharma
Yogyakarta, Indonesia
mamerutamab@gmail.com

Hari Suparwito
Departemen Informatika
Universitas Sanata Dharma
Yogyakarta, Indonesia
shirsj@jesuits.net

Abstrak—Pencemaran udara merupakan masalah lingkungan hidup dan menjadi salah satu faktor yang mempengaruhi kualitas hidup manusia. Kualitas udara dan beberapa variabel pendukung menjadi penentu tingkat pencemaran udara di suatu daerah. Studi ini bertujuan untuk memprediksi tingkat pencemaran udara berbasis *Deep Learning* yang terdiri dari CNN dan LSTM. CNN digunakan untuk mengekstrak variabel yang faktor pencemaran udara dan LSTM digunakan untuk pemodelan informasi temporen tidak teratur dalam komponen data deret waktu pencemaran udara. Data diperoleh dari repositori data Jakarta https://data.jakarta.go.id/. Teknik data imputasi dengan algoritma KNN dan MICE digunakan untuk meningkatkan kualitas data. Model CNN-LSTM dengan imputasi KNN pada parameter SO_2 memberikan hasil terbaik pada data testing dengan nilai RMSE 4,474 dan nilai MAPE 11,323%. Hasil prediksi NAQI yang dihasilkan dengan perhitungan parameter pencemar udara memiliki nilai yang baik ditunjukkan dengan nilai RMSE sebesar 10 dan MAPE sebesar 13%.

Kata kunci—cnn-lstm, deep learning, indeks pencemaran udara, prediksi

I. PENDAHULUAN

Pencemaran udara merupakan masalah lingkungan hidup yang serius. Banyak masalah kesehatan yang muncul akibat pencemaran udara seperti iritasi pernapasan, alergi kulit, hingga kanker paru-paru [1]. Berdasarkan situs *AQAir* pada tanggal 2 September 2023, Jakarta menduduki peringkat satu dunia untuk kualitas udara terburuk di kota-kota besar dengan nilai konsentrasi PM_{25} yang bencapai 105,6 microgram per meter kubik. Salah satu faktor yang mempengaruhi tingkat pencemaran udara di Jakarta adalah tingginya kendaraan bermotor di Jakarta [2]. Kementerian Lingkungan Hidup dan Kehutanan Indonesia membuat stasiun pemantauan otomatis kontinu untuk memantau mutu udara. Laporan indeks per variabel udara disampaikan dalam bentuk Indeks Standar Pencemaran Udara (ISPU).

Tujuan studi ini adalah membuat model untuk memprediksi indeks kualitas udara di Jakarta berdasarkan data historis ISPU yang disediakan pemerintah Jakarta pada situs https://data.jakarta.go.id/. Prediksi dilakukan dengan memprediksi setiap variabel pencemar pada data ISPU, kemudian hasil prediksi tersebut akan dihitung nilai indeksnya dengan rumus *National Air Quality Index* (NAQI). Sebelum data dilatih, data tersebut akan melalui tahap praproses dengan imputasi data menggunakan metode *k-Nearest Neighbor* dan *Multiple Imputation by Chained Equations* untuk pengisian data kosong. Model *Convolution Neural Network-Long Short-Term Memory* (CNN-LSTM) digunakan untuk melakukan prediksi indeks kualitas udara. Hasil prediksi dari algoritma tersebut akan diukur menggunakan nilai *Root Mean Squared Error* (RMSE) dan *Mean Absolute Percentage Error* (MAPE) untuk evaluasi kinerja model dari data hasil imputasi yang dilakukan.

Model CNN-LSTM adalah gabungan dari dua jenis arsitektur jaringan syaraf tiruan yang berbeda yaitu CNN dan LSTM. Model ini digunakan untuk memproses dan mengambil informasi dari data sekuensial seperti gambar bergerak, teks, atau data deret waktu *(time series)*. Model LSTM dapat secara efesien menangkap informasi pola pada data deret waktu, sementara model CNN dapat menyaring gangguan data masukan dan mengekstrak fitur yang lebih berharga [3]. Namun, CNN standar baik untuk mengatasi data auto korelasi spasial, model tersebut tidak diadaptasi untuk mengelola ketergantungan temporal yang kompleks dan panjang dengan benar, sedangkan sebaliknya dengan jaringan LSTM meskipun dirancang untuk mengatasi korelasi temporal, model tersebut hanya mengeksploitasi fitur yang disediakan dalam set pelatihan [4].

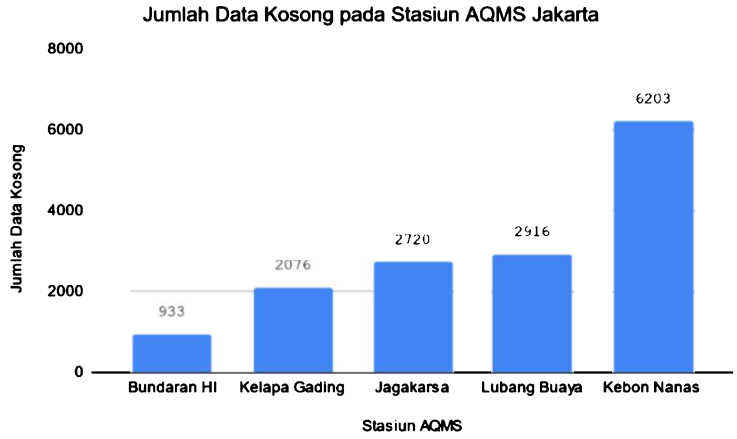
Pada penelitian ini layer dari algoritma CNN akan mengekstrak fitur beberapa variabel yang mempengaruhi tingkat pencemaran udara dan layer dari LSTM akan disesuaikan untuk pemodelan informasi temporan tren tidak teratur dalam komponen data *time series* [5]. Dengan menggunakan CNN dan LSTM secara bersamaan, model prediksi dapat mengoptimalkan keuntungan dari kedua jenis arsitektur, hal tersebut dapat kemungkinan meningkatkan keakuratan prediksi tingkat pencemaran udara.

Beberapa penelitian yang menggunakan model CNN-LSTM diantaranya adalah Alhussein, M (2020) melakukan prediksi pada dataset Australia Individual Household Load dengan algoritma CNN-LSTM dan mendapatkan akurasi MAPE sebesar 40.38% [6]. Kim, T (2019) memprediksi dataset UCI Electricity Consumption dengan CNN-LSTM berhasil mendapatkan nilai MSE sebesar 0.37. Seringkali, data ISPU yang diperoleh tidak lengkap. Hal ini dapat mempengaruhi keakuratan dan reliabilitas prediksi tingkat pencemaran udara. Beberapa penelitian telah dilakukan untuk memprediksi dan mengklasifikasi data ISPU diantaranya adalah Hidayatullah, B. K., Kallista, M., & Setianingsih, C. (2022) memprediksi data ISPU Jakarta dengan rentang waktu antara tanggal 1 Januari 2015 sampai 31 Oktober 2021 dengan algoritma Long Short-Term Memory dengan hasil RMSE pada parameter PM_{10} sebesar 0.00723, parameter SO_2 sebesar 0.05841, parameter CO sebesar 0.05473, parameter O_3 sebesar 0.0446, dan parameter NO_2 sebesar 0.0431.

II. METODOLOGI PENELITIAN

A. Data

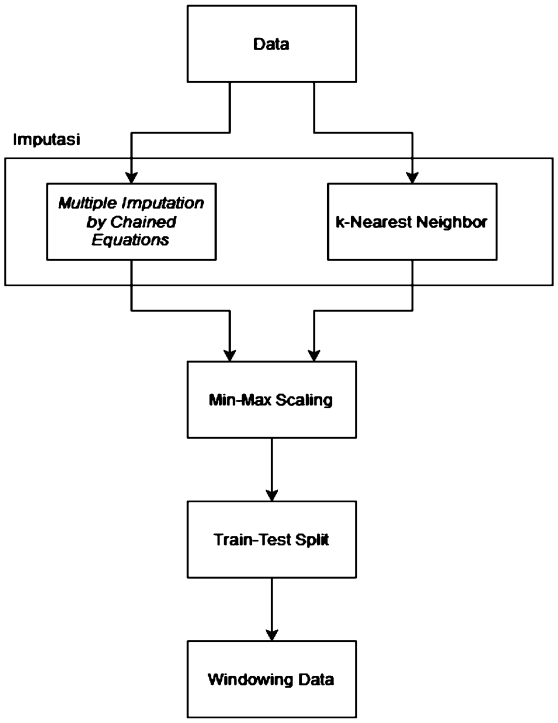
Data yang digunakan pada penelitian ini adalah data ISPU dari tahun 2010 hingga 2021 pada stasiun *Air Quality Monitoring System* (AQMS) di Bundaran HI. Alasan pemilihan lokasi, pertama Bundaran HI karena bundaran hi merupakan pusat kota Jakarta yang didominasi oleh pusat perbelanjaan dan kantoran oleh karena itu bundaran HI sangat ramai dengan kegiatan lalu-lalang transportasi yang menjadi sumber utama tingkat pencemaran udara [7]. Alasan kedua, data ISPU pada lokasi Bundaran HI memiliki data kosong paling sedikit dibandingkan dengan data pada stasiun lainnya. Perbandingan jumlah data kosong pada setiap stasiun AQMS dapat dilihat pada gambar grafik 1 berikut ini.



Gambar 1. Grafik jumlah data kosong pada setiap stasiun AQMS Jakarta

B. Praproses

Praproses dilakukan dengan empat tahapan, yaitu tahap data imputasi, data normalisasi, *train-test data split*, dan *windowing data*. Gambar 2 berikut ini menunjukan 4 data praproses tersebut.



Gambar 2. Praprosess data

Data imputasi menggunakan dua metode *k-Nearest Neighbor* (KNN) dan *Multiple Imputation by Chained Equations* (MICE) yang diterapkan untuk pengisian data kosong. Hasil imputasi data digunakan pada model dan dibandingkan untuk melihat apakah model yang dibuat dapat bekerja dengan baik. Setelah dilakukan data imputasi, data tersebut dinormalisasi dengan teknik *min-max scaling*.

Tahap selanjutnya adalah pembagian data (*data splitting*). Pada penelitian ini, data akan dibagi menjadi data *training* dan *testing*. Data *training* adalah data yang digunakan untuk melatih model yang dibuat, sedangkan data *testing* digunakan untuk mengevaluasi model tersebut. Data *training* yang digunakan adalah data ISPU Bundaran HI dari tahun 2010 hingga tahun 2020, kemudian data ISPU Bundaran HI tahun 2021 digunakan sebagai data *testing*. Total data *training* yang digunakan adalah 4018 baris data dan total data *testing* adalah 365 baris data. Tahap terakhir adalah *windowing data*. Tahap ini membagi nilai x dan y dari data original. Nilai x dan y didapatkan berdasarkan nilai *lookback* yang diberikan. Pada penelitian ini akan digunakan nilai *lookback* sebanyak 30 dengan nilai y = 1. Transformasi data asli menjadi data setelah praproses dapat dilihat pada tabel 1

TABEL 1. TRANSFORMASI DATA

Data Asli dari Parameter PM_{10}	**Data Praproses dari Parameter PM_{10}**	
	Nilai x	**Nilai y**
60, 32, 27, 22, 25, 30, 41, 64, 55, 34, 55, 45, 28, 38, 35, 54, 44, 38, 40, 61, 36, 51, 35, 56, 59, 43, 54, 72, 51, 47, 32	[[0.56], [0.28], [0.23], [0.18], [0.21], [0.26], [0.37], [0.6], [0.51], [0.3], [0.51], [0.41], [0.24], [0.34], [0.31], [0.5], [0.4], [0.34], [0.36], [0.57], [0.32], [0.47], [0.31], [0.52], [0.55], [0.39], [0.5], [0.68], [0.47], [0.43]]	0.28

C. Model CNN-LSTM

Algoritma CNN-LSTM digunakan untuk membuat model prediksi tingkat pencemaran udara. Model CNN-LSTM merupakan gabungan dari dua algoritma *neural network* yaitu *convolution neural network* dan *long short-term memory*.

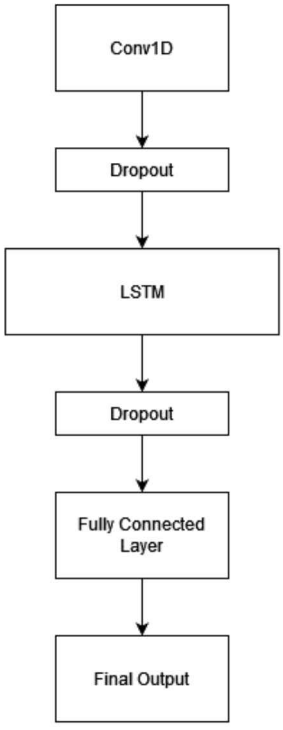
Model tersebut digunakan untuk memproses dan mengambil data sekuensial seperti gambar bergerak, teks, atau data deret waktu. Pada kasus ini, CNN dan LSTM memiliki kemampuan untuk mempelajari secara mendalam pola-pola yang bersifat kompleks pada data deret waktu [8]. Layer pada algoritma CNN akan digunakan untuk mengekstrak fitur beberapa variable yang mempengaruhi tingkat pencemaran udara dan layer dari LSTM akan disesuaikan untuk pemodelan informasi tempran tren tidak teratur dalam komponen data. Model CNN-LSTM terdiri dari dua komponen utama, yaitu komponen *convolutional layer* dan *LSTM layer*. Kemudian ditambahkan komponen *fully connected layer* dan *final output layer* untuk menghasilkan prediksi.

Skenario pengujian yang dilakukan dalam studi ini dapat dilihat pada tabel 2 berikut.

TABEL 2. SKENARIO PENGUJIAN MODEL CNN-LSTM

Layer	Kernels	Keterangan
Convolution 1D	Filters	16/32/64
	Kernel Size	3/4/5
	Activation Function	ReLu
Dropout	Rate	0,2
LSTM	Units	32/64/128
Dense	Units	32/64/128
	Activation Function	ReLu
Dense	Units	1

Pengujian ini menggunakan *optimizer adam* [9] dengan nilai *lookback* 30. Struktur model CNN-LSTM dapat dilihat pada gambar 3.



Gambar 3 Struktur Model CNN-LSTM

Pada struktur model CNN-LSTM dalam gambar 3, terdapat layer dropout. Layer tersebut digunakan untuk menghindari overfitting yang terjadi pada neural network [10].

D. *National Air Quality Index*

National Air Quality Index (NAQI) merupakan gabungan dari sub indeks lima parameter: partikulat, CO, SO_2, oksidan, dan NO_2. Indeks tersebut dapat dituliskan sebagai berikut [11]:

$$NAQI = \sqrt{lc^2 + ls^2 + lp^2 + la^2 + ln^2}$$

dimana,

lc = Indeks pencemaran karbon monoksida

ls = indeks pencemaran belerang dioksida

lp = indeks pencemaran partikulat

la = indeks pencemaran oksida foto kimia

ln = indeks pencemaran oksida nitrogen

E. *Evaluasi Performa Data Model*

Evaluasi performa dari model CNN-LSTM untuk memprediksi indeks tingkat pencemaran udara dalam penelitian ini menggunakan nilai RMSE dan MAPE sebagai kriteria utama metode evaluasi. RMSE digunakan untuk mengetahui besarnya penyimpangan yang terjadi antara prediksi nilai indeks pencemar udara dengan nilai actual indeks pencemar udara hasil observasi [12], sedangkan MAPE digunakan untuk menguji akurasi dari hasil prediksi. Tabel 3 berikut ini menunjukan tingkat prediksi berdasarkan nilai MAPE.

TABEL 3. PARAMETER MAPE

Nilai MAPE	Prediksi
$MAPE \leq 10$	Tinggi
$10 < MAPE \leq 20$	Baik
$20 < MAPE \leq 50$	Layak
> 50	Rendah

Nilai prediksi parameter MAPE: Tinggi, Baik, Layak dan Rendah diperoleh dari penelitian Octavia dan Afandi (2019) [13]

III. HASIL DAN PEMBAHASAN

Library tensorflow digunakan untuk pembuatan model dan *talos* digunakan untuk melakukan uji skenario. Pada proses training model, model yang dibuat akan dioptimasi menggunakan *optimizer Adam*, MSE sebagai *loss function* dan RMSE sebagai metrik akurasi model yang dilatih. Sebelum proses latih, akan diinisiasi terlebih dahulu variable *params* yang merupakan nilai *hyperparameter* yang akan diuji sesuai dengan model *hyperparameter* yang digunakan.

TABEL 4. HYPERPARAMETER MODEL

Hyperparameter	Nilai
Conv_filter	16/32/64
Conv_kernel	3/4/5
dropout	0.2
Lstm_unit	32/64/128
Dense_unit	32/64/128
Activation	ReLu
Optimizer	Adam
Losses	Mean_squared_error

Kemudian dibuat model CNN-LSTM dengan nilai parameter input X_train, y_train, dan *params*. Setelah itu, model akan dilatih dengan total 20 *epoch* dan 128 *batch size*.

Untuk mengetahui hasil prediksi dari model yang telah dibuat, maka perlu dilakukan proses evaluasi model. Data yang digunakan untuk melakukan evaluasi model adalah data testing yakni data periode Januari 2021 - December 2021. Nilai yang akan dihitung adalah RMSE dan MAPE. Setiap variable dari tingkat pencemaran udara mempunyai model yang berbeda untuk menghasilkan nilai terbaik.

A. Hasil Pengujian Data Imputasi KNN

Hasil pengujian skenario dari data imputasi *KNN* dapat dilihat pada tabel 5. Berdasarkan hasil pengujian, dapat dilihat parameter dengan nilai paling baik adalah SO_2 dengan nilai RMSE data training sebesar 0,039, RMSE data evaluasi sebesar 4,474 dan MAPE sebesar 11,365%. Dari keseluruhan nilai akurasi yang dihasilkan, parameter PM_{10} dan SO_2 masuk kedalam range "Baik" dengan hasil perhitungan MAPE diantara 10 hingga 20 persen. Parameter CO, O_3, dan NO_2 masuk kedalam range "layak" dengan hasil perhitungan MAPE diantara nilai 20 hingga 50 persen.

B. Hasil Pengujian Data MICE

Hasil pengujian data imputasi MICE dapat dilihat pada tabel 6. Berdasarkan hasil pengujian dari tabel diatas, dapat dilihat parameter dengan nilai paling baik adalah SO_2 dengan nilai RMSE training sebesar 0,0394, RMSE evaluasi sebesar 4,827 dan MAPE sebesar 11,323%. Dari keseluruhan nilai akurasi yang dihasilkan, parameter PM_{10} dan SO_2 masuk kedalam range "Baik" dan parameter CO, O_3, dan NO_2 masuk kedalam range "Layak".

TABEL 5. HASIL PENGUJIAN DATA IMPUTASI KNN

Params	Hyperparameter				RMSE Data Training	RMSE Data Testing	MAPE Data Testing
	Conv_Filer	*Conv_Kernel*	*Lstm_unit*	*Dense_unit*			
PM_{10}	64	3	128	128	0,1064	10,039	17,431%
SO_2	64	3	128	128	0,039	4,474	11,365%
CO	64	5	128	32	0,0825	4,479	27,385%
O_3	64	3	128	128	0,0892	7,273	31,776%
NO_2	64	3	128	32	0,0556	7,418	27,325%

TABEL 6. HASIL PENGUJIAN DATA IMPUTASI MICE

Params	Hyperparameter				RMSE Data Training	RMSE Data Testing	MAPE Data Testing
	Conv_Filter	*Conv_Kernel*	*Lstm_unit*	*Dense_unit*			
PM_{10}	64	3	128	32	0,106	9,674	17,662%
SO_2	64	3	64	128	0,0394	4,827	11,323%
CO	64	5	128	128	0,0798	4,611	24,865%
O_3	64	4	128	128	0,088	7,180	31,427%
NO_2	64	3	128	32	0,0556	7,503	26,079%

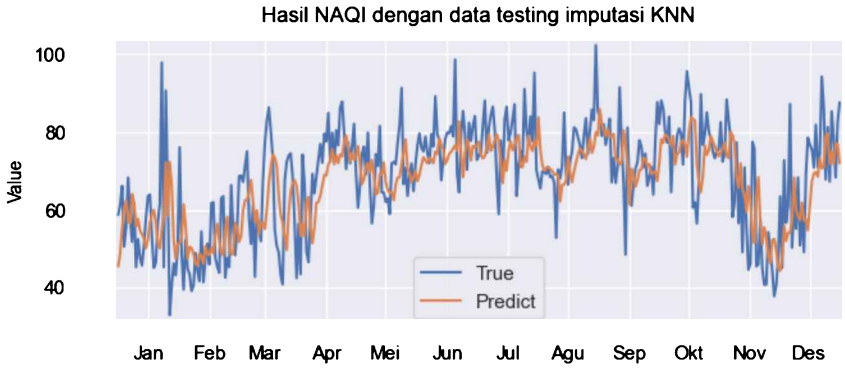
C. Prediksi NAQI

Setelah mendapatkan nilai prediksi dari pengujian skenario terhadap lima parameter pencemar, selanjutnya adalah menghitung nilai NAQI untuk mendapatkan nilai indeks kualitas udara. Hasil perhitungan NAQI dapat dilihat pada gambar 4a dan gambar 4b. Secara visual, hasil prediksi perhitungan NAQI dapat mengikuti pola data historis dengan baik. Berdasarkan nilai NAQI yang tergambar dalam gambar 4a dan gambar 4b didapatkan nilai RMSE dan MAPE pada tabel 7 berikut ini.

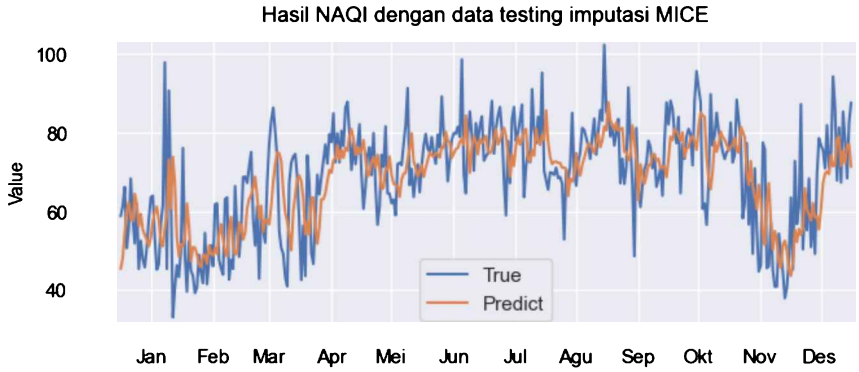
TABEL 7. HASIL EVALUASI PERHITUNGAN NAQI

DATA IMPUTASI	RMSE	MAPE
MICE	10,844	13,405%
KNN	10,924	13,432%

Hasil prediksi nilai NAQI pada data imputasi menggunakan teknik MICE dan KNN masuk ke dalam range "Baik" berdasarkan nilai MAPE yang didapatkan yaitu sebesar 13% kemudian didapatkan nilai RMSE sebesar 10



Gambar 4a. Hasil prediksi NAQI pada data testing imputasi KNN



Gambar 4b. Hasil prediksi NAQI pada data testing imputasi MICE

D. Analisa Tren

Trend dalam kasus penelitian ini merupakan informasi kenaikan atau penurunan indeks kualitas udara dari hari sebelumnya.

TABEL 8. TREND DATA IMPUTASI KNN

Tanggal	Hari	Trend Original	Trend Prediksi	Keterangan
19-4-2021	Senin	Naik	Naik	TRUE
20-4-2021	Selasa	Turun	Turun	TRUE
21-4-2021	Rabu	Naik	Naik	TRUE
22-4-2021	Kamis	Turun	Turun	TRUE
23-4-2021	Jumat	Naik	Naik	TRUE
24-4-2021	Sabtu	Naik	Turun	FALSE
25-4-2021	Minggu	Turun	Naik	FALSE

TABEL 9. TREND DATA IMPUTASI MICE

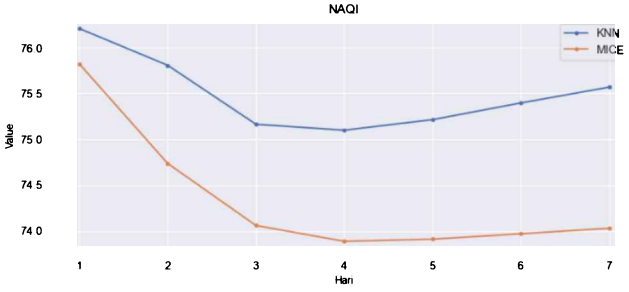
Tanggal	Hari	Trend Original	Trend Prediksi	Keterangan
19-4-2021	Senin	Naik	Naik	TRUE
20-4-2021	Selasa	Turun	Turun	TRUE
21-4-2021	Rabu	Naik	Naik	TRUE
22-4-2021	Kamis	Turun	Turun	TRUE
23-4-2021	Jumat	Naik	Naik	TRUE
24-4-2021	Sabtu	Naik	Turun	FALSE
25-4-2021	Minggu	Turun	Naik	FALSE

Informasi keterangan pada tabel 8 dan tabel 9 merupakan status jika prediksi trend yang dihasilkan oleh model sama dengan prediksi data historis. Berdasarkan grafik pada gambar 4a dan gambar 4b, dapat dilihat bahwa terjadi banyak nilai naik turun pada nilai indeks. Hal tersebut dapat dilihat lebih detail pada tabel 8 dan tabel 9. Pada tabel tersebut memperlihatkan terjadinya indeks naik turun dalam satu minggu. Model dalam penelitian ini berhasil memprediksi trend diawal minggu, tetapi mengalami kesalahan dalam memprediksi diakhir minggu.

Terdapat beberapa faktor yang dapat mempengaruhi, salah satunya adalah banyaknya jumlah kendaraan bermotor di Jakarta [14]. Kendaraan bermotor bertanggung jawab atas penyebaran gas emisi yang tersebar di udara Jakarta. Adanya peraturan-peraturan khusus di Jakarta seperti *Car Free Day* dihari minggu dan ganjil-genap di hari kerja mempengaruhi penyebaran kendaraan di Jakarta [15][16]. Selain itu, kondisi cuaca pada musim tertentu juga mempengaruhi tingkat pencemaran udara. Pada musim panas, tingkat polutan O_3 meningkat. Hal tersebut mempengaruhi nilai indeks tingkat pencemaran udara [17].

E. Prediksi Satu Minggu Kedepan

Prediksi satu minggu kedepan dengan model yang dibuat dapat dilihat pada gambar 5. Hal tersebut membuktikan bahwa model tersebut dapat memprediksi data yang belum diketahui.



Gambar 5. Hasil prediksi satu minggu kedepan

Hasil dari prediksi diatas dapat terlihat perbedaan hasil prediksi dari data hasil imputasi KNN dengan hasil imputasi MICE. Namun kedua data tersebut menghasilkan trend prediksi yang sama. Pada bagian awal prediksi dapat terlihat bahwa trend yang dihasilkan yaitu menurun, kemudian di akhir prediksi trend tersebut menaik.

IV. SIMPULAN

Model CNN-LSTM untuk memprediksi tingkat pencemaran udara pada data ISPU Bundaran HI Jakarta sudah dilakukan. Metode yang digunakan adalah peramalan data deret waktu variable tunggal, dimana prediksi pencemaran udara dilakukan dengan melakukan prediksi pada setiap senyawa pencemar pada data ISPU, kemudian hasil prediksi parameter tersebut dihitung menjadi nilai NAQI yang merupakan nilai indeks pencemar udara. Metode KNN dan MICE digunakan untuk mengisi data kosong pada ISPU. Hasil percobaan menunjukan bahwa CNN-LSTM memiliki akurasi prediksi tinggi dan kinerja yang baik terhadap data hasil imputasi dengan KNN dan MICE. Prediksi parameter terbaik pada hasil prediksi dengan data imputasi tersebut adalah parameter SO_2 pada data imputasi KNN dan data imputasi MICE. Hasil prediksi *NAQI* yang dihasilkan dengan perhitungan parameter pencemar memiliki akurasi prediksi yang tinggi dan kinerja yang baik dengan nilai RMSE sebesar 10 dan akurasi MAPE yaitu "Baik" dengan nilai sebesar 13% pada data hasil imputasi KNN dan MICE. Trend yang dihasilkan oleh model dipengaruhi oleh banyak faktor seperti gas emisi kendaraan bermotor dan cuaca.

REFERENSI

[1] Abidin, J., & Hasibuan, F. A. (2019). Pengaruh Dampak Pencemaran Udara Terhadap Kesehatan Untuk Menambah Pemahaman Masyarakat Awam Tentang Bahaya Dari Polusi Udara.

[2] Ismiyati, I., Marlita, D., & Saidah, D. (2014). Pencemaran Udara Akibat Emisi Gas Buang Kendaraan Bermotor. Jurnal Manajemen Transportasi & Logistik (JMTRANSLOG), 1(3), 241. https://doi.org/10.54324/j.mtl.v1i3.23

[3] Mehtab, S., & Sen, J. (2022). Analysis and Forecasting of Financial Time Series Using CNN and LSTM-Based Deep Learning Models. Dalam J. P. Sahoo, A. K. Tripathy, M. Mohanty, K.-C. Li, & A. K. Nayak (Ed.), *Advances in Distributed Computing and Machine Learning* (Vol. 302, hlm. 405–423). Springer Singapore. https://doi.org/10.1007/978-981-16-4807-6_39

[4] Livieris, I. E., Pintelas, E., & Pintelas, P. (2020). A CNN–LSTM model for gold price time-series forecasting. *Neural Computing and Applications*, *32*(23), 17351–17360. https://doi.org/10.1007/s00521-020-04867-x

[5] Kim, T.-Y., & Cho, S.-B. (2019). Predicting residential energy consumption using CNN-LSTM neural networks. Energy, 182, 72–81. https://doi.org/10.1016/j.energy.2019.05.230

[6] Alhussein, M., Aurangzeb, K., & Haider, S. I. (2020). Hybrid CNN-LSTM Model for Short-Term Individual Household Load Forecasting. *IEEE Access*, *8*, 180544–180557. https://doi.org/10.1109/ACCESS.2020.3028281

[7] Dinas Lingkungan Hidup Daerah DKI Jakarta (2020). Laporan Akhir Pemantauan Kualitas Udara DKI Jakarta. https://lingkunganhidup.jakarta.go.id/files/laporan2020/udara.pdf

[8] Widiputra, H., Adele Mailangkay, & Elliana Gautama. (2021). Prediksi Indeks BEI dengan Ensemble Convolutional Neural Network dan Long Short-Term Memory. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 5(3), 456–465. https://doi.org/10.29207/resti.v5i3.3111

[9] Chang, Z., Zhang, Y., & Chen, W. (2019). Electricity price prediction based on hybrid model of adam optimized LSTM neural network and wavelet transform. *Energy*, *187*, 115804. https://doi.org/10.1016/j.energy.2019.07.134

[10] Ozgur, A., & Nar, F. (2020). Effect of Dropout layer on Classical Regression Problems. *2020 28th Signal Processing and Communications Applications Conference (SIU)*, 1–4. https://doi.org/10.1109/SIU49456.2020.9302054

[11] Hartini, Eko (2020). Indeks Kualitas Udara. https://repository.dinus.ac.id/docs/ajar/Indeks_Kualitas_Udara.pdf

[12] Azka, M. A., Sugianto, P. A., Silitonga, A. K., & Nugraheni, I. R. (2018). Uji Akurasi Produk Estimasi Curah Hujan Satelit Gpm Imerg Di Surabaya, Indonesia. *Jurnal Sains & Teknologi Modifikasi Cuaca*, *19*(2), 83. https://doi.org/10.29122/jstmc.v19i2.3153

[13] Octavia, Y., Afandi, A. N., & Putranto, H. (2019). Studi prakiraan beban listrik menggunakan metode artificial neural network. *TEKNO*, *28*(2), 116. https://doi.org/10.17977/um034v28i2p116-129

[14] Lestari, P., Damayanti, S., & Arrohman, M. K. (2020). Emission Inventory of Pollutants (CO, SO_2, $PM_{2.5}$, and NO_X) In Jakarta Indonesia. *IOP Conference Series: Earth and Environmental Science*, *489*(1), 012014. https://doi.org/10.1088/1755-1315/489/1/012014

[15] Rachman, H. O. (2019). Impact Of Car-Free Day On Air Pollution And Its Multifarious Advantages In Sudirman-Thamrin Street, Jakarta. *International Journal of GEOMATE*, *17*(62). https://doi.org/10.21660/2019.62.8286

[16] Zulkarnain, Ghiffary, A., 2021. Impact of odd-even driving restrictions on air quality in Jakarta. International Journal of Technology. Volume 12(5), pp. 925-934

[17] Qonitan, F., Haidar, F., & Zahra, N. (2022). Overview of the Air Pollution Standard Index and Associated Health Risk in DKI Jakarta during the 2019 Dry Season. *Proceedings of the 1st International Conference on Contemporary Risk Studies, ICONIC-RS 2022, 31 March-1 April 2022, South Jakarta, DKI Jakarta, Indonesia*. Proceedings of the 1st International Conference on Contemporary Risk Studies, ICONIC-RS 2022, 31 March-1 April 2022, South Jakarta, DKI Jakarta, Indonesia, DKI Jakarta, Indonesia. https://doi.org/10.4108/eai.31-3-2022.2321002

Klasifikasi Sentimen Masyarakat Mengenai Kinerja Aplikasi PeduliLindungi Menggunakan *Naïve Bayes*

Bonifatius Choshe Manggala Putra
Departemen Informatika
Universitas Sanata Dharma
Yogyakarta, Indonesia
bchoshechoshe@gmail.com

C. Kuntoro Adi
Departemen Informatika
Universitas Sanata Dharma
Yogyakarta, Indonesia
kuntoroadi@usd.ac.id

Abstrak— Pandemi Covid-19 beberapa waktu yang lalu menyebabkan banyak hal berubah. Aplikasi PeduliLindungi yang dibangun untuk melacak penyebaran Covid-19 sangat membantu masyarakat untuk semakin peduli akan kesehatan diri maupun orang lain. Namun seiring aplikasi ini digunakan oleh Masyarakat, banyak ditemukan penilaian, kepuasan maupun ketidak-puasan masyarakat atas aplikasi ini. Penelitian ini bertujuan untuk melihat bagaimana pendekatan *Naïve Bayes* mampu dengan baik mengklasifikasi sentimen masyarakat terhadap aplikasi PeduliLindungi. Data diambil dari penilaian atau komentar yang terdapat di *Playstore* dan *Twitter*. Dataset yang digunakan berjumlah 100,000 data dengan 50,000 komentar masing-masing dari sentimen positif maupun negatif. Pendekatan *Naïve Bayes* pada penelitian ini mampu melakukan klasifikasi dengan baik dengan nilai akurasi sebesar 88.58%.

Kata kunci—Naïve Bayes, Covid-19, PeduliLindungi, Akurasi.

I. PENDAHULUAN

Pada tanggal 2 Maret 2020, untuk pertama kali *Coronavirus Disease* (Covid-19) dilaporkan masuk ke Indonesia di Depok, Jawa Barat. Dalam waktu yang tidak lama aplikasi PeduliLindungi dikembangkan untuk membantu instansi pemerintah melakukan pelacakan penyebaran Covid-19. Aplikasi ini mengandalkan partisipasi masyarakat untuk saling membagikan data lokasinya saat bepergian agar penelusuran riwayat kontak dengan penderita Covid-19 dapat dilakukan. Pengguna aplikasi ini juga akan mendapatkan pemberitahuan jika berada di keramaian atau berada di zona merah, yaitu area yang di dalamnya ada orang yang terinfeksi positif Covid-19 atau ada pasien dalam pengawasan.

Seiring aplikasi ini digunakan oleh masyarakat luas, ditemukan banyak keluhan terkait kinerja aplikasi ini. Ada beberapa contoh keluhan, misalnya: sertifikat vaksin yang tidak muncul walaupun data sudah sesuai, ada juga sertifikat vaksin dari vaksin 1 dan 2 namun hanya sertifikat vaksin 1 saja yang tercantum pada aplikasi PeduliLindungi, sistem verifikasi yang masih memiliki kekeliruan, dan kesulitan dalam scan *QR Code* untuk *check in* di suatu tempat seperti *minimarket* atau pusat perbelanjaan.

Hal ini mendorong penelitian ini untuk melakukan analisis sentimen terhadap aplikasi PeduliLindungi dengan mengambil data komentar dari *Google Play* dan media sosial *Twitter*.

Berikut beberapa penelitian terkait analisis sentimen. Anggraini, 2020 [1] menggunakan data sentimen terhadap kebijakan kartu pra kerja melakukan klasifikasi menggunakan *Naïve Bayes* menghasilkan akurasi 91.06%

Sa'rony, 2019 [3] melakukan klasifikasi dengan menggunakan metode *Naïve Bayes* terhadap kebijakan pemindahan ibukota Republik Indonesia. Diketahui dari hasil penelitian bahwa prosentase pengguna *Twitter* yang pro terhadap kebijakan pemindahan ibukota sebesar 52% sedangkan sisanya kontra terhadap kebijakan tersebut. Evaluasi sistem menghasilkan akurasi sebesar 94% , presisi sebesar 94,5 %, dan *recall* sebesar 94%.

Utama, 2019 [4] melakukan analisis sentimen terhadap kebijakan kendaraan ganjil genap di tol Bekasi menggunakan algoritma *Naïve Bayes* dengan optimalisasi *Information Gain*. Diketahui dari hasil penelitian bahwa pengguna media sosial yang pro dan kontra terhadap kebijakan ganjil genap di tol Bekasi jumlahnya seimbang dengan persentase masing masing 50%. Dari hasil evaluasi dihasilkan nilai akurasi sebesar 79,55%, presisi sebesar 80,37%, dan sensitivitas atau *recall* sebesar 80,51%.

Bertolak dari beberapa penelitian di atas, penelitian ini akan melakukan analisis sentimen terhadap kinerja aplikasi PeduliLindungi menggunakan algoritma *Naïve Bayes*. Data yang digunakan diambil langsung dari kolom komentar penilaian dari *PlayStore* dan media sosial *Twitter*. Beberapa skenario dilakukan untuk untuk menemukan hasil optimal, misalnya perubahan label positif menjadi 1 dan negatif menjadi -1, melakukan variasi *k-fold* pada *cross-validation* untuk membentuk dan menguji model, serta pengujian variasi akurasi untuk membentuk kurva *Receiver Operating Characteristic* (ROC).

II. METODE PENELITIAN

Sentimen merupakan pendapat yang didasarkan pada perasaan yang mengutarakan secara berlebihan terhadap sesuatu [2]. Sentimen terdapat di dalam pernyataan atau kalimat yang didalamnya terdapat pendapat. Sentimen digunakan untuk mengetahui perasaan yang diberikan terhadap suatu objek.

Preprocessing data penting dilakukan untuk mengubah data mentah dalam bentuk yang mudah dipahami. Langkah *preprocessing* data dilakukan untuk menyelesaikan beberapa masalah seperti *noisy data*, duplikasi data, *missing value*, dan lain-lain.

Terdapat langkah-langkah data *preprocessing* misalnya *filtering data, case folding, stemming, stopword removal, tokenize,* dan *labelling*.

A. *Filtering Data*

Filtering data yaitu digunakan untuk menghilangkan atau menghapus data duplikat dan atribut yang akan digunakan dalam membentuk model.

B. *Labelling Data*

Proses pelabelan data sangat diperlukan untuk menentukan kelas pada suatu ulasan dalam dokumen apakah ulasan tersebut masuk dalam kelas berlabel positif atau negatif.

C. *Case Folding*

Pada tahap ini dilakukan pengubahan kata dengan huruf kapital (*uppercase*), menjadi huruf kecil (*lowercase*).

D. *Tokenize*

Pemecahan kata dalam bentuk token dari suatu rangkaian karakter yang berdasarkan karakter spasi.

E. *Normalize*

Normalisasi kata dilakukan untuk pengecekan dan pengubahan kata yang tidak baku menjadi kata baku berdasarkan kamus yang digunakan.

F. *Stopword Removal*

Stopword removal dilakukan untuk pengecekan dan penghapusan kata kata yang tidak menambah informasi misalnya kata hubung, kata depan, dan kata sambung.

G. *Stemming*

Stemming dilakukan untuk pencarian kata dasar di setiap kata pada *tweet*.

H. *N-Gram*

N-Gram adalah metode dalam mengolah bahasa alami yang digunakan untuk menganalisis teks. Metode ini akan membagi teks menjadi bagian bagian kecil yang disebut dengan *n-grams*. *N-Grams* terdiri dari urutan n kata atau karakter yang berdekatan dalam teks.

Pembobotan kata merupakan proses pemberian nilai atau bobot untuk setiap kata yang terdapat dalam sebuah sentimen [6]. Dalam pencarian informasi peringkat berdasarkan frekuensi kata salah satu metode yang paling banyak digunakan adalah metode TF-IDF (*Term Frequency-Inversed Document Frequency).* Pembobotan dilakukan dengan mengalikan *Term frequency* (TF) - jumlah kata atau term pada dokumen dengan *inverse document Frequency* (IDF) - frekuensi kemunculan kata pada seluruh dokumen.

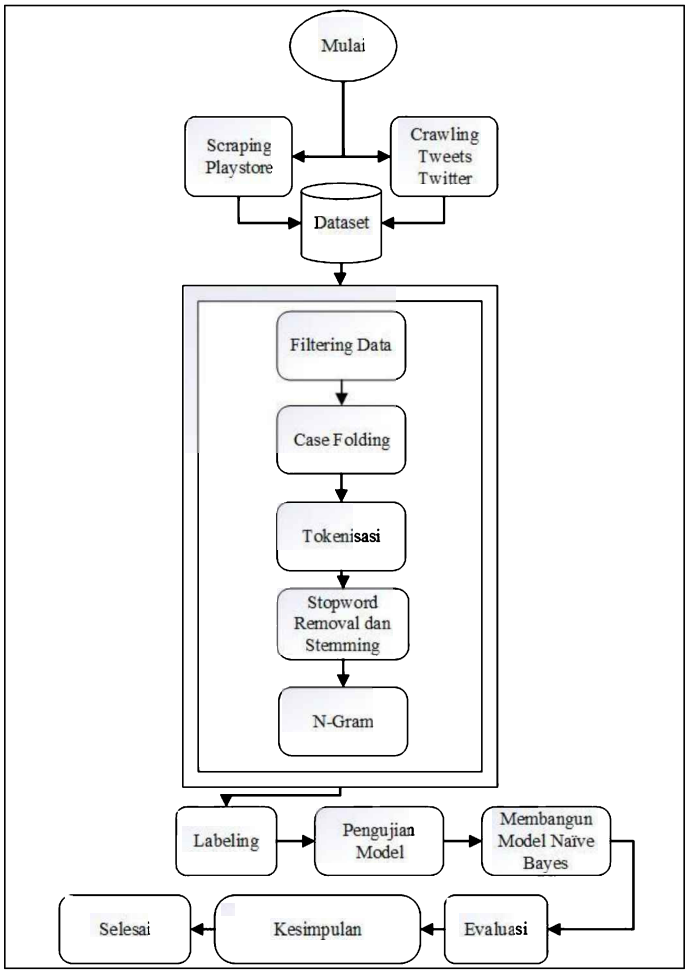
Klasifikasi *Naïve Bayes* adalah klasifikasi yang memiliki sifat *supervised learning* [5]. Kinerja *Naïve Bayes* memiliki kecepatan klasifikasi yang baik [7]. Pada penelitian ini pengujian akan digunakan menggunakan 2 kelas saja yaitu kelas atau sentimen positif dan negatif.

Data diperoleh dari komentar pemberian *Rating* pada laman aplikasi PeduliLindungi di *Google Playstore* dan *Tweeter*. Komentar positif dan negatif masing-masing berjumlah 50,000 data sehingga jumlah data total sebanyak 100,000.

III. PERANCANGAN MODEL

A. *Gambaran Umum Penelitian*

Alur penelitian dalam melakukan klasifikasi sentimen dapat dilihat seperti pada Gambar 1.



Gambar 1. Alur Penelitian Sistem Analisis Sentimen

Dalam penelitian ini:

- Data diambil dari dua sumber yang berbeda yaitu *Google Playstore* dan *tweets Twitter*
- Ada beberapa tahap *preprocessing* data yaitu *filtering* data, *case folding*, *tokenisasi*, *stopword removal*, *stemming*, dan *N-Gram*
- Setelah tahap *preprocessing*, penelitian ini melakukan pelabelan data menggunakan *tools* ***Valence Aware Dictionary and sEntiment Reasoner*** (VADER)
- Dengan variasi k pada *k-fold cross-validation*, penelitian ini kemudian membangun model *Naïve Bayes* dan menguji model menggunakan fitur TF-IDF dari data yang sudah dipersiapkan.
- Kinerja pendekatan *Naïve Bayes* dievaluasi dari perhitungan akurasi yang dihasilkan dalam pengujian sistem.

IV. HASIL MODEL DAN ANALISIS

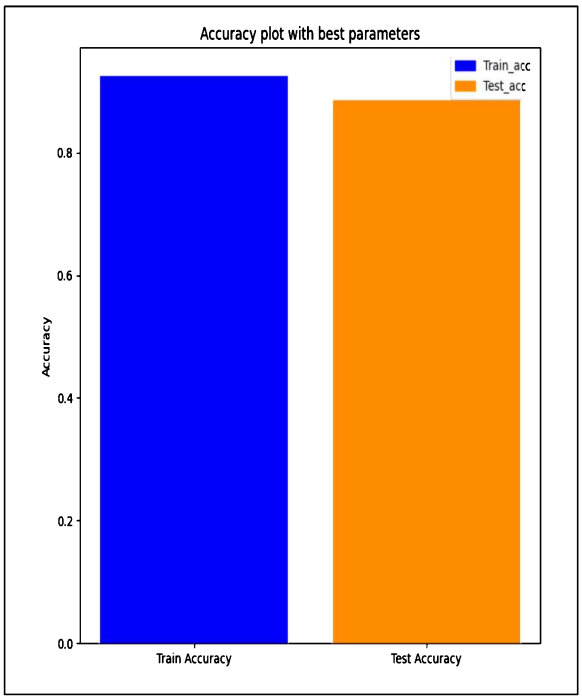
A. *Pengujian Model*

Pengujian model pada penelitian ini memiliki tujuan untuk melihat seberapa baik model dalam melakukan klasifikasi data sentimen yang berasal dari rating *Google Play Store* dan *Twitter* menggunakan metode *Naïve Bayes*.

Beberapa pengujian dengan variasi jumlah k pada *k-fold cross-validation* dilakukan untuk menemukan akurasi, *precision*, *recall*, dan *ROC Curve*.

a) Klasifikasi (Train dan Test Accuracy)

Pengujian klasifikasi terdiri dari hasil akurasi dari data latih dan uji. Penelitian ini menghasilkan nilai akurasi data training sebesar 92.48% dan nilai akurasi data testing sebesar 88.58%. Hasil dari akurasi training dan testing terdapat pada gambar 2 dan tabel 1 berikut ini.



Gambar 2. Grafik hasil akurasi train dan test

TABEL 1. HASIL AKURASI TRAIN DAN TEST

Train Accuracy	Test Accuracy
92.48%	88.58%

b) Confusion Matrix

Tabel *confusion matrix* digunakan untuk mengukur sejauh mana model klasifikasi dapat memprediksi dengan benar kelas-kelas dari data uji. Dengan informasi yang diberikan oleh *confusion matrix*, penelitian dapat membuat keputusan apakah model bekerja dengan baik atau perlu diperbaiki.

Nilai confusion matrix pada penelitian ini terdapat pada tabel 2 dibawah ini.

TABEL 2. HASIL *CONFUSION MATRIX*

True Positive (TP)	19854
True Negative (TN)	2199
False Positive (FP)	1331
False Negative (FN)	1550

Nilai akurasi dihitung dari jumlah (TP + TN) dibagi total data uji (TP + TN + FP + FN).

c) *F1-score*, *Precision* dan *Recall*

F1-score, *precision*, dan *recall* adalah tiga metrik evaluasi yang digunakan untuk mengukur kinerja model klasifikasi. Ketiganya berfokus pada perbedaan antara hasil prediksi model dan kelas aktual data uji.

Precision menggambarkan sejauh mana model berhati-hati dalam membuat prediksi positif. Dengan kata lain, dari semua prediksi positif yang dibuat oleh model, berapa persentase yang memang benar.

Recall mengukur sejauh mana model mampu mengidentifikasi semua kasus positif yang sebenarnya dan menggambarkan sejauh mana model mampu menangkap atau mengingat semua kasus positif.

Nilai *F1* adalah ukuran rata-rata harmonik antara *precision* dan *recall*. Ini berguna saat ingin melihat keseimbangan antara *precision* dan *recall*, terutama jika ada *trade-off* antara keduanya.

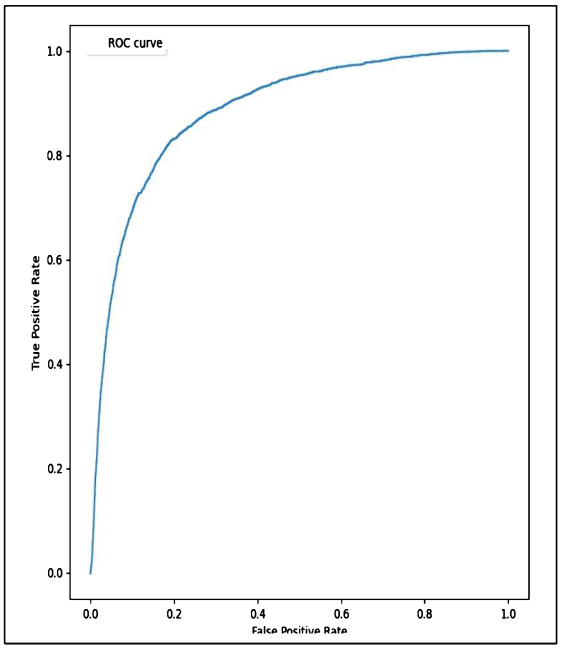
F1-score, *precision*, dan *recall* adalah metrik evaluasi yang penting dalam klasifikasi karena mereka memberikan berbagai aspek kinerja model, terutama dalam kasus ketidakseimbangan antara kelas positif dan negatif. Hasil kinerja *Naïve Bayes* dalam penelitian ini dapat dilihat pada Tabel 3 di bawah ini.

TABEL 3. HASIL *F1-SCORE, PRECISION* DAN *RECALL*

F1-*Score*	76.93%
Precision Score	77.70%
Recall Score	76.23%

d) ROC Curve

Kurva ROC (*Receiver Operating Characteristic*) adalah alat evaluasi yang digunakan untuk mengukur dan memvisualisasikan kinerja model klasifikasi, terutama dalam konteks pemilihan ambang batas yang berbeda untuk pemilihan kelas positif dan negatif. Kurva ROC sering digunakan untuk mengukur sejauh mana model mampu membedakan antara kelas positif dan negatif. Hasil evaluasi menggunakan kurva ROC terdapat pada gambar 3.



Gambar 3. Kurva ROC

B. Pengujian Skenario

Pengujian skenario pada penelitian ini memiliki tujuan untuk melihat seberapa baik akurasi dan ketepatan antara data latih dengan data uji menggunakan metode *Naïve Bayes.*

Pengujian skenario dilakukan dengan melakukan berbagai perbandingan pada setiap *k-fold cross-validation* untuk menemukan hasil yang optimal. Nilai k divariasi antara 3 hingga 11 dan memberikan hasil akurasi sebagaimana dilihat di Tabel 4.

TABEL 4. HASIL AKURASI DENGAN BERBAGAI NILAI K

K-*fold cross validation*	Akurasi
3	88.58%
5	88.57%
7	**88.55%**
9	88.57%
10	88.56%
11	88.56%

Tabel menunjukkan, hasil optimal dengan nilai akurasi tertinggi sebesar 88.58% terjadi ketika sistem melakukan *cross*-validasi dengan nilai k = 3.

V. KESIMPULAN DAN SARAN

A. Kesimpulan Penelitian

Penelitian menunjukkan bahwa pendekatan *Naïve Bayes* mampu melakukan klasifikasi sentimen yang ada dalam aplikasi PeduliLindungi dengan baik. Pemodelan *Naïve Bayes* dengan variasi pengujian k *fold* 3,5,7,9,10,11 memberi hasil akurasi terbaik 88.58% pada nilai k *fold* 3.

B. Saran

- Dari penelitian yang ada, sangat dimungkinkan metode selain *Naïve Bayes* dipergunakan untuk analisis sentimen sehingga didapatkan hasil yang lebih optimal.
- Proses labeling data menggunakan *tools* Vader memberikan hasil yang tidak seimbang. Dalam penelitian ini Vader memberikan hasil label positif sejumlah 14,985 data dan label negatif sejumlah 84,669 data. Metode labeling seperti *text bloob* bisa diuji-cobakan untuk mendapatkan label yang lebih seimbang. Selain itu penyeimbangan data seperti pendekatan *Synthetic Minority Oversampling Technique* (SMOTE) kiranya bisa juga digunakan untuk menemukan hasil yang lebih baik.

REFERENSI

[1] Anggraini, W. P. (2020). Klasifikasi Sentimen Masyarakat Terhadap Kebijakan Kartu . *Faktor Exacta Vol.13 No. 4*.

[2] Bimananda, W. R. (2019). Analisis Text Mining dari Cuitan Twitter Mengenai . *Eigen Mathematics Journal* .

[3] Sa'rony, A. A. (2019). Analisis Sentimen Kebijakan Pemindahan Ibukota Republik Indonesia. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, no. e-ISSN: 2548-964X*.

[4] Utama, H. S. (2019). Sentimen Analisis Kebijakan Ganjil Genap di Tol Bekasi .

[5] Bustami. (2013). Penerapan Algoritma Naïve Bayes untuk Mengklasifikasi Data.

[6] Cheng, J. &. (1999). Comparing Bayesian Network Classifiers.

[7] Hamzah, A. (2012). Klasifikasi Teks dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis. *Seminar Nasional Apikasi Sains & Teknologi (SNAST) Periode III*.

[8] Pramana, S. B. (2018). Data Mining dengan R Konsep Serta Implementasi.

[9] Prayoga, N. D. (2018). Sistem Diagnosis Penyakit Hati Menggunakan Metode Naïve Bayes. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*.

Klasterisasi Perputaran Barang Retail menggunakan Metode *Clustering* K-Means

Candika Silai Prahma Setiadi
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
candika.setiadi@gmail.com - orcid.org/0009-0005-2185-2434

Donny Avianto
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
donny@uty.ac.id - orcid.org/0000-0001-5499-5478

Abstrak—CV ESIA INDORAYA merupakan sebuah perusahaan distributor yang menjual produk seperti pulsa, kuota internet, dan kartu perdana. CV ESIA INDORAYA selama ini mencatat pemenuhan stok barang dan produk secara manual yang menyebabkan terjadinya kesalahan dalam pencatatan data perputaran barang di retail. CV ESIA INDORAYA dalam sistem yang telah berjalan saat ini tidak memiliki kemampuan untuk membentuk kelompok produk yang laku dan tidak laku terjual. Maka diperlukan proses pengolahan data besar dengan menggunakan metode K-Means Clustering. Proses klasifikasi K-Means Clustering ini menghasilkan dua buah *cluster* yang dapat digunakan sebagai acuan retail yang laku dan tidak laku. Sistem yang telah dibuat baik karena Nilai K yang dipakai yaitu K = 2 memiliki SSE yang mengalami penurunan paling besar yaitu SSE = 30.426 x 10^{16}.

Kata kunci—Random Forest, Klasifikasi, Industri Garmen, Karyawan, Produktivitas

I. PENDAHULUAN

Permintaan dan penjualan produk distribusi dari banyaknya retail di area Malang Raya merupakan hal yang sangat penting bagi perusahaan. Pendataan menentukan jumlah minimum produk berupa saldo, kartu perdana dan *voucher* kuota internet yang harus dipenuhi berdasarkan permintaan setiap retail selama ini dicatat secara manual yang menyebabkan sering terjadinya kesalahan pada proses penyortiran data yang telah tersedia dan memakan waktu kurang lebih dua minggu. Data tersebut tidak dapat digunakan untuk mengukur tingkat perputaran barang. Tingkat perputaran barang dapat digunakan untuk mengukur retail mana yang memberi keuntungan lebih besar.

Solusi untuk menyelesaikan permasalahan pada penjelasan di atas salah satunya dengan memanfaatkan algoritma K-Means Clustering. Algoritma yang paling sederhana untuk membentuk *cluster* dari algoritma yang lain adalah Algortitma K-Means. Salah satu kelebihan dari algoritma ini adalah mudah diimplementasikan dan dijalankan, relatif cepat, dapat beradaptasi dengan mudah, dan salah satu algoritma yang paling sering digunakan dalam praktek data *mining*. *Clustering* merupakan sebuah metode yang mengelompokkan beberapa data dengan tujuan data tersebut dikelompokkan dalam beberapa kelompok data mengurangi ruang pencarian yang spesifik dalam merespons suatu *query*.[1]

Penelitian ini bertujuan untuk membuat model program aplikasi yang dapat meng-*cluster* atau membentuk kelompok retail yang memiliki tingkat perputaran barang yang tinggi karena paling diminati, perputaran barang sedang, dan retail dengan perputaran barang rendah. Hasil dari performa perputaran barang memiliki hubungan dengan tingkat laba yang akan didapatkan oleh perusahaan. Dari beberapa bentuk cluster tersebut dapat dianalisis jenis cluster mana yang memiliki perputaran barang tinggi untuk pendistribusian barang pada CV Esia Indoraya.

II. STUDI LITERATUR

Pada penelitian sebelumnya yang dibuat pada tahun 2021 dengan judul penelitian Penerapan Data Mining Metode K-Means Clustering Untuk Analisa Penjualan Pada Toko Fashion Hijab Banten. Pengolahan data oleh peneliti dilakukan menggunakan metode K-means clustering dengan perhitungan secara manual yang bertujuan untuk membentuk beberapa kelompok seperti; pakaian yang sangat laris, laris, dan kurang laris. Pengujian dari data tersebut dilakukan dan diolah menggunakan Rapid Miner. Hasil akhir dari proses pengolahan data terdapat 11 artikel sangat laris, 55 artikel laris, dan 34 artikel untuk kurang laris.[2]

Pada penelitian sebelumnya yang dibuat pada tahun 2019 dengan judul penelitian Penerapan Data Mining Metode Clustering Pada CV. Secom Infotech Menggunakan Algoritma K-Means. Aplikasi yang digunakan untuk melakukan pengujian model adalah aplikasi RapidMiner di mana terdapat dua cluster yang terbentuk. Cluster 0 berjumlah 14 data dan cluster 1 berjumlah 16 data. Metode K-Means clustering ini menggunakan konsep *data mining* untuk mengelompokkan data sesuai atribut[3]

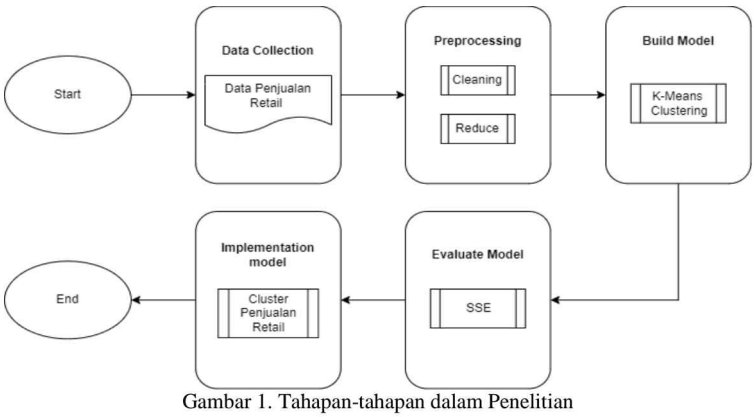
Pada penelitian sebelumnya yang dibuat pada tahun 2019 dengan judul penelitian Klasifikasi Barang Menggunakan Metode Clustering K-Means Dalam Penentuan Prediksi Stok Barang (Studi Kasus: UKM Mar'ah Jilbab Kediri). Pengujian ini memiliki hasil pengelompokan RFM yang tingkat akurasinya sebesar 70%, dam hasil tanpa menggunakan RFM sebesar 76,67%[4]

Pada penelitian sebelumnya yang dibuat pada tahun 2021 dengan judul penelitian Penerapan Algoritma K-Means Dalam Prediksi Penjualan Karoseri. Penelitian tersebut menggunakan data sebanyak 203 data yang dikelompokkan menjadi 2 cluster. Cluster 0 memiliki hasil 93 item penjualan karoseri kurang laris yang kurang diminati oleh konsumen dan cluster 1 memiliki hasil 110 item penjualan karoseri sangat laris yang banyak diminati oleh konsumen.[5]

Pada penelitian sebelumnya yang dibuat pada tahun 2020 dengan judul penelitian Klasterisasi Pola Penjualan Pestisida Menggunakan Metode K-Means Clustering (Studi Kasus Di Toko Juanda Tani Kecamatan Hutabayu Raja). Penelitian ini menggunakan algoritma K-Means untuk mengelompokkan Penjualan Pestisida Sangat Laku sebanyak 53 items, Penjualan Laku sebanyak 21 item dan Penjualan Tidak laku sebanyak 126 items [6].

III. METODOLOGI PENELITIAN

Penelitian ini memiliki beberapa tahap seperti yang terlihat pada Gambar 1. Tahap pertama yaitu data *collection* yang merupakan dataset yang digunakan. *Preprocessing* yang di dalamnya terdapat beberapa proses yaitu *cleaning data* dan *reduce data.* Penelitian ini membuat model dengan memanfaatkan algoritma K-Means. Selanjutnya tahap evaluasi model yang menggunakan *Sum of Square Error (SSE).* Tahap terakhir yaitu *implementation model* yang terdapat hasil dari pengelompokan data perputaran barang retail. Berikut Gambaran Tahapan Penelitian.



Gambar 1. Tahapan-tahapan dalam Penelitian

A. *Data Collection*

Data pada penelitian ini merupakan data perputaran barang retail di CV Esia Indoraya dalam periode Januari 2021 sampai dengan Januari 2022. Data ini sudah berbentuk data RFM (*Recency, Frequency,* dan *Monetary)* yang menjelaskan perputaran barang pada retail tersebut. Sampel dari data ini memiliki jumlah atribut dari data RFM tersebut.

B. *Preprocessing*

Langkah awal yang melibatkan sejumlah kombinasi dan proses yang memerlukan intervensi atau penyesuaian dari pengguna[7]. Pada tahap ini, data awal disiapkan dan dimodifikasi untuk memastikan kualitas, integritas, dan ketersediaan data yang diperlukan dalam analisis atau pemodelan lebih lanjut.

1) Data Cleaning

Data Cleaning atau pembersihan data adalah proses pengelolaan dan pemrosesan data untuk mengidentifikasi, mengoreksi, menghapus kesalahan, atau ketidaksesuaian dalam dataset. Tujuan utama dari pembersihan data adalah memastikan data yang digunakan untuk analisis atau pemodelan adalah akurat dan konsisten[8].

2) Data Reduction

Data Reduction atau reduksi data adalah proses mengurangi volume atau kompleksitas data dan mempertahankan sebagian besar informasi yang relevan. Tujuan dari reduksi data adalah untuk membuat dataset yang lebih kecil dan lebih mudah diolah.

C. *Build Model*

Algoritma K-Means adalah salah satu bentuk algoritma pengelompokan *iterative* yang membuat suatu pembagian dataset ke beberapa *cluster* yang sudah ditentukan parameternya di awal. Algoritma K-Means relatif mudah diterapkan dan dijalankan dengan waktu yang cepat, serta mudah beradaptasi. Dalam pengembangan metode *data mining*, K-means merupakan salah satu dari banyaknya algoritma yang paling berpengaruh dan populer dalam bidangnya. Proses dari Algoritma K-Means awalnya algoritma ini mengambil beberapa komponen dari dataset untuk dijadikan pusat cluster awal. Langkah pengambilan komponen ini pusat cluster dari sekumpulan populasi data diambil secara acak. Selanjutnya, K-Means melakukan pengujian pada setiap komponen pada populasi data dan memberikan tanda spesifik pada komponen tersebut ke salah satu pusat cluster yang telah didefinisikan menurut jarak terdekat antar komponen dengan setiap pusat cluster yang telah ditentukan.

$$V = \frac{\sum_{i-1}^{n} xi}{n} ; i = 1, 2, 3 \dots \dots n \quad (1)$$

$$d(x, y = \|x - y\| = \sqrt{\sum_{x=i}(x_i - y_i)} \quad (2)$$

Keterangan:
V = centroid pada cluster
Xi = objek ke -i
Yi = data y ke -i
n = Jumlah objek yang menjadi anggota cluster

D. *Evaluate Model*

SSE adalah metrik evaluasi yang umum digunakan untuk mengukur kualitas partisi atau kluster dalam algoritma K-means. SSE mengukur jarak kuadrat antara setiap titik data dengan pusat kluster terdekat. SSE dihitung dengan menjumlahkan kuadrat jarak antara setiap titik data dan pusat kluster yang ditetapkan, dan tujuannya adalah untuk meminimalkan nilai SSE. Perbandingan yang didapatkan untuk memvalidasi data yaitu dengan melakukan perhitungan *Sum of Square Error* (SSE) dari nilai cluster.

$$SSE = \sum_{k=1} k \sum_{x_i \in S_k} (X_i - C_k)^2 \quad (3)$$

Keterangan:
V = jumlah kelompok yang digunakan pada algoritma K-Means
Xi = banyaknya data ke- i
Ck = jumlah cluster pada cluster nilai K

E. Implementation Model

Tahap ini merupakan tahap implementasi dari model *clustering* K-Means yang telah dibuat. Model yang sudah dilatih menggunakan metode *clustering* K-Means akan diimplementasikan dalam bentuk *web,* pada implementasi *web* akan memanfaatkan *framework* dari python yaitu Flask. Flask merupakan salah satu jenis dari microframework yang menggunakan bahasa *python*. Fungsi dari Flask ini adalah membuat kerangka kerja aplikasi dan membuat tampilan dari suatu web. Penggunaan Flask dan bahasa pemrograman *python* dapat membantu pengembang membuat sebuah sistem web yang lebih terstruktur serta dapat mempermudah memberikan tampilan yang lebih baik pada integrasi aplikasi.

IV. HASIL DAN PEMBAHASAN

A. Pembahasan

Penelitian ini bertujuan membuat sebuah sistem pengelompokan dengan metode *clustering* K-Means untuk menentukan kelompok mana yang memiliki kelas-kelas dengan perputaran barang yang cepat atau lambat. Cara kerja sistem ini adalah *user* memasukkan jumlah nilai K yang akan menentukan berapa jumlah *cluster* yang akan dibentuk. Setelah itu, akan dilalukan *preprocessing data* dengan melakukan pengurangan jumlah data dan atribut serta membersihkan data yang tidak diperlukan. Dari hasil data yang sudah di *cleaning* akan dilakukan pembuatan model K-Means dan dilakukan evaluasi model dengan menggunakan perhitungan *Sum Square of Error (SSE)* untuk menentukan nilai K mana yang memiliki jumlah penurunan *error* paling tinggi. Tahapan yang dilakukan dalam penelitian ini berupa tahap pengumpulan data, *preprocessing data,* pembuatan model, evaluasi model.

1) Pengumpulan Data

Data yang telah dikumpulkan merupakan data perputaran barang CV Esia Indoraya pada periode Januari 2021 sampai dengan Januari 2022. Data berisikan konten penjualan produk berbentuk data RFM *(Recency, Frequency,* dan *Monetary).* Data yang digunakan dalam penelitian ini hanyalah data dari RFM yang memiliki bentuk angka untuk pembagian kelompok pada data-data tersebut. Data yang digunakan terdapat pada gambar 2 berikut ini.

ID RET	NAMA RET	SE AREA	LEVEL	TSS	R	F	M
36759	PANDUMEDIA 2	SINGOSARI	L7	MALANG BATU	175	32	1,898,700,000
230500	MUJUR SURYA	KLOJEN	L7	MALANG KOTA	179	132	7,605,930,688
193986	GUMB CELL	KARANGPLOSO	L4	MALANG BATU	159	13	147,206,400
4637	MSA CELL	WAGIR	L0	MALANG KABUPATEN	157	27	1,193,000
9201	DEVIS CELL	WAGIR	L1	MALANG KABUPATEN	160	31	1,550,000
27571	SUBUR CELL	WAGIR	L1	MALANG KABUPATEN	155	32	1,241,000
27574	AL BAROKAH CELL	WAGIR	L1	MALANG KABUPATEN	175	41	1,659,000
37011	WD CELL	WAGIR	L1	MALANG KABUPATEN	167	34	5,896,000
37016	ALHUSNA CELL	WAGIR	L3	MALANG KABUPATEN	178	59	50,889,000
37018	SAR CELL	WAGIR	L1	MALANG KABUPATEN	153	20	1,209,000
37316	YAFFA CELL	WAGIR	L1	MALANG KABUPATEN	173	44	2,672,000
37319	PIXEL CELL	WAGIR	L1	MALANG KABUPATEN	163	23	11,981,000
43347	JEHAN 3 CELL	WAGIR	L1	MALANG KABUPATEN	156	31	1,162,000
43723	ARIN CELL	WAGIR	L1	MALANG KABUPATEN	165	24	3,215,000
51970	PINK CELL	WAGIR	L1	MALANG KABUPATEN	175	38	2,243,000
53885	MFM CELL	WAGIR	L1	MALANG KABUPATEN	163	26	1,213,000
54072	I CLOUND CELL	WAGIR	L1	MALANG KABUPATEN	179	28	1,141,000
54073	APP CELL	WAGIR	L2	MALANG KABUPATEN	178	44	6,032,000
54074	PRIE CELL	WAGIR	L1	MALANG KABUPATEN	174	31	1,446,000
54075	YUTONE	WAGIR	L2	MALANG KABUPATEN	174	44	8,467,000

Gambar 2. Data Penelitian

Data ini berjumlah 1932 data dengan beberapa kolom antara lain tentang kolom ID RET untuk ID setiap retail yang tersedia. Kolom kedua NAMA RET menjelaskan tentang nama dari retail yang didaftarkan. Kolom ketiga SE AREA menjelaskan tentang area penjualan dari retail. Kolom keempat LEVEL menjelaskan tentang level dari retail. Kolom kelima TSS menjelaskan tentang area pusat pengambilan produk. Kolom keenam, ketujuh, dan kedelapan menjelaskan tentang data RFM (Recency, Frequency, dan Monetary).

2) Preprocessing Data

Proses prepocessing data, data yang telah tersedia dilakukan proses reduce data dan mengambil bagian kolom RFM (Recency, Frequency, dan Monetary). Proses ini bertujuan untuk mengambil hanya data

yang memiliki komponen untuk mengukur tingkat perputaran barang pada retail. Data yang tidak memiliki kepentingan lebih seperti NAMA RET dan ID RET tidak diinputkan agar memudahkan pengelompokan pada model.

3) Build Model

Pada proses pembuatan model, penulis memanfaatkan metode clustering K-Means untuk membangun model yang akan digunakan pada penelitian. Pada metode K-Means ini, akan memanfaatkan nilai K sebagai inputan jumlah cluster. Penelitian ini akan memanfaatkan beberapa inputan nilai K, selanjutnya untuk mengukur seberapa tepat pusat *cluster* yang telah terbentuk menerapkan *Elbow Method*.

4) Evaluasi Model

Output yang dihasilkan penelitian ini adalah terbentuknya cluster yang dapat digunakan untuk menentukan retail mana yang memiliki perputaran barang yang cepat (laris) dan lambat (kurang laris) melalui analisis nilai monetary (m) pada dataset. Untuk menentukan jumlah cluster paling optimal, maka diperlukan evaluasi cluster dengan SSE yang terdapat pada tabel 1.

TABEL 1. PERCOBAAN NILAI K

No	Nilai K	Nilai SSE
1	K = 2	SSE = 30.426×10^{16}
2	K = 3	SSE = 15.376×10^{16}
3	K = 4	SSE = 9.444×10^{16}
4	K = 5	SSE = 6.113×10^{16}
5	K = 6	SSE = 4.074×10^{16}

Nilai SSE merupakan sebuah nilai penurunan *error* yang menunjukkan kualitas partisi atau klaster pada algoritma K-means. Dalam matriks evaluasi ini, semakin besar nilai SSE maka semakin besar juga kualitas klaster yang telah dibentuk. Berdasarkan pengujian SSE, penurunan nilai *Error* terbesar terdapat pada K = 2 dengan nilai SSE = 30.426×10^{16}. Nilai SSE yang terbentuk merupakan nilai klaster dengan penurunan tingkat *error* yang tinggi, melalui penurunan tersebut semakin tinggi nilai SSE maka semakin baik kluster yang terbentuk. Korelasi nilai SSE dengan perputaran barang yang telah di klaster adalah melalui kualitas dari penurunan error tersebut, dengan semakin besar nilai SSE maka semakin baik kluster yang dibentuk dan hal tersebut berpengaruh pada pembagian kelompok barang laris atau tidak laris. Kluster yang berjumlah dua klaster (K=2) memiliki model optimal karena kualitas dari kluster yang terbentuk memiliki kualitas partisi yang baik. Melalui data yang telah disediakan dan nilai dalam klaster yang dilampirkan, perputaran barang dianalisis melalui nilai kolom pada *Monetary* yang memiliki nilai yang berkaitan dengan keuntungan barang yang dijual (laris) dan memiliki perputaran barang cepat pada konten. Klaster yang terbentuk dapat dikategorikan sebagai klaster perputaran barang cepat (laris) dan klaster perputaran barang lambat (kurang laris).

B. Hasil

Dari hasil penelitian didapatkan bahwa data perputaran barang di retail dengan menganalisis nilai monetary (M) serta menentukan nilai K yang optimal dengan penurunan terbesar yang terdapat pada nilai K = 2 dengan nilai SSE = 30.426×10^{16}. Hasil pengelompokan yang telah dilakukan dapat berfungsi untuk menentukan retail laris dan retail kurang laris. Cluster 0 dengan nilai lebih rendah yaitu 564.163,986 sehingga tergolong kelompok dengan perputaran barang yang lambat (kurang laris) dan Cluster 1 dengan nilai lebih tinggi yaitu 13.090.668,447 tergolong kelompok dengan perputaran barang lebih cepat (laris).

TABEL 2. HASIL PENELITIAN

Cluster	*Recency*	*Frequency*	*Monetary*
0	164.371	24.103	564.163,986
1	174.289	95.447	13.090.668,447

Dari tabel 2 dijelaskan kolom Cluster yaitu 0 dan 1 yang terbentuk sesuai nilai input oleh user. Di kolom selanjutnya adalah Recency, frequency, dan Monetary yang berisi data dari setiap retail yang telah diinput. Hasil tersebut dapat disimpulkan melalu nilai Monetary kelompok mana yang memiliki perputaran barang yang lebih laris.

V. KESIMPULAN DAN SARAN

Model *clustering* K-means yang digunakan pada penelitian ini bertujuan untuk membentuk *cluster* yang mengelompokkan barang dengan tingkat perputaran tinggi , sedang, dan rendah. Kelompok yang telah terbentuk dapat digunakan sebagai sarana untuk menentukan retail mana yang memiliki perputaran barang tinggi dan dapat mengembangkan perusahaan menuju keuntungan yang lebih tinggi.

Sebagai saran pengembangan, penelitian selanjutnya dapat menambahkan data di luar dataset untuk dikelompokkan ke dalam *cluster* yang telah terbentuk dan dikelompokkan sesuai dengan parameter *cluster*. Selain itu, penelitian selanjutnya menerapkan metode pengujian sistem yang lebih baik sehingga nilai K yang terbentuk lebih optimal dan *cluster* yang terbentuk memiliki validasi nilai yang lebih tinggi.

REFERENSI

[1] S. Supriyatna, "Klasisifikasi Data Menggunakan Algoritma K-Means Clustering Dan Naive Bayes Classifier Berdasarkan Analisa Tekstur Metode Gray Level Co-Occurrence Matrix (GLCM)," 2023.

[2] S. Nurajizah and A. Salbinda, "Penerapan Data Mining Metode K-Means Clustering Untuk Analisa Penjualan Pada Toko Fashion Hijab Banten," *Jurnal Teknik Komputer AMIK BSI*, vol. 7, no. 2, 2021, doi: 10.31294/jtk.v4i2.

[3] D. Juliawan, F. Amir, and E. Desi, "Prosiding Seminar Nasional Riset Information Science (SENARIS) Penerapan Data Mining Metode Clustering Pada CV. Secom Infotech Menggunakan Algoritma K-Means," 2019.

[4] N. Salsabila, "Klasifikasi Barang Menggunakan Metode Clustering K-Means Dalam Penentuan Prediksi Stok Barang," 2019.

[5] D. Anggarwati, O. Nurdiawan, I. Ali, and D. A. Kurnia, "Penerapan Algoritma K-Means Dalam Prediksi Penjualan Karoseri," vol. 1, no. 2, pp. 58–62, 2021.

[6] S. Nurajizah and A. Salbinda, "Penerapan Data Mining Metode K-Means Clustering Untuk Analisa Penjualan Pada Toko Fashion Hijab Banten," *Jurnal Teknik Komputer AMIK BSI*, vol. 7, no. 2, 2021, doi: 10.31294/jtk.v4i2.

[7] M. Soni and S. Varma, "Diabetes Prediction using Machine Learning Techniques," 2020. [Online]. Available: www.ijert.org

[8] N. P. A. Widiari, I. M. A. D. Suarjaya, and D. P. Githa, "Teknik Pengolahan Data Cleaning," *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, vol. 8, no. 2, 2020.

Perbandingan Metode *Ensemble Learning* pada Klasifikasi Tingkat Stres Siswa

Dirga Halim Susilo
Departemen Sains Data
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
halimdirga8@gmail.com

Muhammad Fakhri Fakhrurrozi
Departemen Sains Data
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
fahrirozi90@gmail.com

Neni
Departemen Sains Data
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
09neni16@gmail.com

Abstract*—Stress is a feeling that a person faces when feeling pressured, facing a threat, or being in a new situation. Today, stress is not only felt by adults; students can also experience stress. This is due to several factors that make students very vulnerable to stress. A national survey was conducted to identify the factors causing stress in students, revealing several main contributors. The data will then be analyzed to determine the level of stress among students based on the factors they experience. In the analysis process, a comparison is made using several methods, namely decision trees, support vector machines (SVM), logistic regression, KNN, and a voting classifier. The results of the analysis indicate that the Voting Classifier method achieves a better training accuracy of 96%. However, it's important to note that the Decision Tree Model attains 100% training accuracy, which may indicate potential overfitting of the model.

Keywords*—student stress factor, ensemble learning, method comparison

I. PENDAHULUAN

Stres merupakan sebuah perasaan yang dihadapi seseorang ketika merasa tertekan, menghadapi sebuah ancaman ataupun sedang dalam situasi yang baru. Stres ini merupakan respon non-spesifik terhadap sebuah tuntutan yang dirasakan oleh individu baik respon positif ataupun berupa respon negatif. [1]

Setiap fase usia, masing-masing memiliki karakteristik yang berbeda dari fase pertumbuhan diusia lainnya. Begitu pula dengan usia remaja yang pastinya memiliki karakteristik khusus yang membedakan dengan usia anak-anak dan juga dewasa. [2]

Dewasa ini stres tidak hanya dirasakan oleh orang dewasa, seorang siswa juga dapat merasakan stres. Hal tersebut dikarenakan beberapa faktor yang menyebabkan siswa sangat rentang terkena stres. Salah satu faktor yang menyebabkan siswa mengalami stres dari sisi akademik yaitu tuntutan akademik siswa terlalu berat, tugas yang tidak sedikit, mendapatkan nilai yang tidak sesuai atau buruk dan juga lingkungan pergaulan. [3]

Faktor lain yang menyebabkan siswa mengalami stres bisa saja dari faktor lingkungan. Seorang siswa pasti membutuhkan yang namanya teman sebaya untuk mengenali dunia luar selain di dalam lingkungan keluarga. Namun seringkali dari interaksi dengan teman menimbulkan tekanan bagi siswa untuk mengikuti teman sebayanya sehingga mereka akan merasakan stres tersebut. [2]

Dari data-data mengenai faktor tingkat stres pada siswa akan dianalisis menggunakan metode *Ensemble Learning*. *Ensemble Learning* adalah metode pada *Machine Learning* yang dapat digunakan untuk klasifikasi. [4] Dengan adanya teknik *Ensemble Learning* dapat meningkatkan akurasi pada model dan juga dapat mengoptimalkan kinerja pada klasifikasi. [5]

Oleh sebab itu, pada penelitian ini dilakukan analisis dengan metode *Ensemble Learning* menggunakan beberapa model untuk analisis dataset sebagai bahan perbandingan untuk menentukan model yang terbaik. Beberapa model yang banyak digunakan dalam penelitian dan juga akan digunakan pada penelitian ini yaitu model *Decision Tree, Logistic Regression* dan *K-Nearest Neighbor* (KNN).

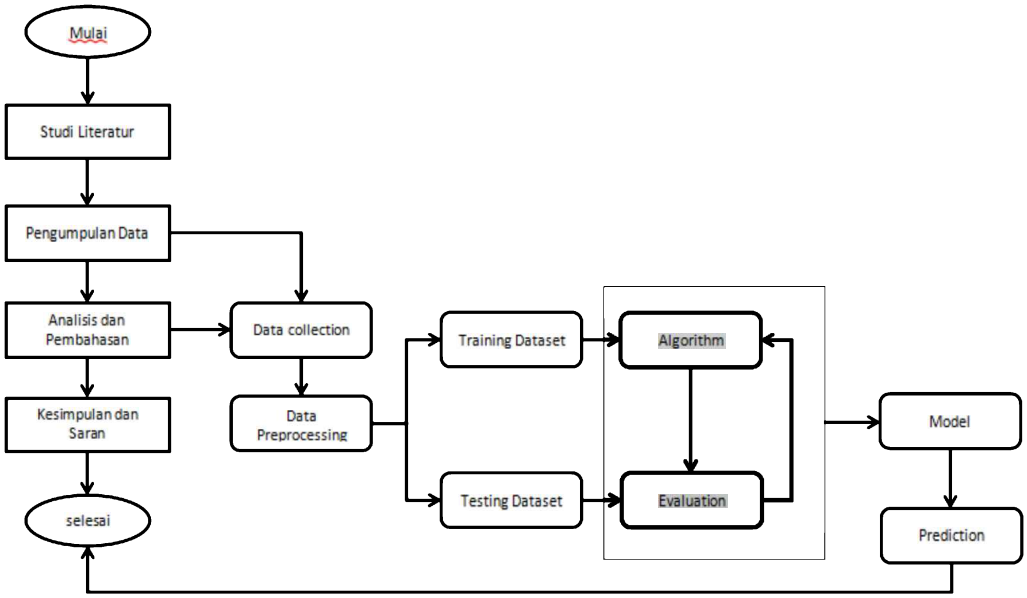
Kontribusi yang dilakukan dalam penelitian adalah untuk menentukan tingkat stres siswa yang disebabkan dari berbagai faktor seperti psikologi, fisiologi, lingkungan, akademik dan sosial.

II. METODOLOGI

Analisis faktor yang menyebabkan siswa stres diawali dengan studi literatur penelitian terdahulu untuk melihat serta menjadikan acuan dalam penelitian. Untuk menghasilkan sebuah analisis yang baik dari sebuah penelitian membutuhkan alur kerja yang baik. Alur penelitian terdapat pada gambar 1.

A. Dataset

Dataset yang dipakai dalam penelitian ini diambil dari *Kaggle Repository* mengenai faktor-faktor stres siswa dilihat dari psikologis, fisiologis, sosial, lingkungan dan akademik. Dataset ini berisi 1100 data yang terdiri dari 20 atribut dan 1 kelas label. Dataset tersebut berkaitan dengan faktor-faktor yang menciptakan dampak paling besar pada stres siswa. Faktor-faktor tersebut dapat berupa gangguan tidur, beban belajar dan bahkan intimidasi. Dataset ini terdiri dari 5 faktor utama yaitu faktor psikologis, faktor fisiologis, faktor lingkungan, faktor akademik dan faktor sosial. Contoh data yang digunakan pada penelitian ini terdapat pada tabel .



Gambar 1. Alur Penelitian

TABEL 1. RINGKASAN DATASET

No.	Atribute Name	Type Atribute
	Faktor Psikologi	
1.	Anxiety_level	Numerik
2.	Self_esteem	Numerik
3.	Mental_healt_history	Numerik
4.	Depression	Numerik
	Faktor Fisiologis	
5.	Headache	Numerik
6.	Blood_pressure	Numerik
7.	Sleep_quality	Numerik
8.	Breathing_problem	Numerik
	Faktor Lingkungan	
9.	Noise_level	Numerik
10.	Living_condition	Numerik
11.	Safety	Numerik
12.	Basic_needs	Numerik
	Faktor Akademik	
13.	Academic_performance	Numerik
14.	Study_load	Numerik
15.	Teacher_student_relationship	Numerik
16.	Future_career_concerns	Numerik
	Factor Sosial	
17.	Social_support	Numerik
18.	Peer_pressure	Numerik
19.	Extracurricular	Numerik
20.	bullying	Numerik
	Output	
21.	Y	Binary

B. Metode Analisis

Dataset yang digunakan akan dianalisis menggunakan metode *Ensemble Learning*. Pada proses analisis dilakukan perbandingan menggunakan beberapa metode yaitu *Decision Tree*, *Support Vector Machine* (SVM), *Logistic Regression*, KNN, dan *Voting Classifier.*

Evaluasi pada model dilakukan dengan melihat akurasi pada setiap model yang dihasilkan, lalu dari hasil setiap pengujian pada masing-masing model yang disajikan akan dilihat model mana yang memiliki performa paling baik. Hasil dari analisis ini nantinya dapat menjadi bahan perbandingan dengan penelitian sebelumnya. [6]

III. HASIL DAN PEMBAHASAN

Penelitian ini dilakukan menggunakan laptop Lenovo ThinkPad dengan sistem operasi Windows 10, serta menggunakan perangkat lunak Visual Studio Code dan juga Jupyter Notebook untuk melakukan analisis dataset. Bahasa pemrograman yang digunakan adalah bahasa Python.

Penelitian ini menggunakan dataset yang diambil dari *Kaggle Repository* tentang faktor stres pada siswa dengan jumlah data sebanyak 1100 baris yang terdiri dari 20 kolom atribut dan 1 kolom label. Dataset ini diambil langsung dari survei untuk mengetahui mengenai faktor-faktor stres yang timbul pada siswa.

Dari dataset yang didapatkan, tidak terdapat *missing value* ataupun *inconsistent* pada dataset. Kemudian dilakukan pembagian data latih dan data uji menggunakan 'train_test_split' dengan rasio 0.8 pada data latih dan 0.2 pada data uji.

Dataset yang telah diuji validasi dilakukan klasifikasi serta perbandingan menggunakan beberapa metode, yaitu *Decision Tree, Support Vector Machine,* dan KNN. Setelah itu, beberapa metode tersebut digabungkan menggunakan metode *Voting Classifier*.

TABEL 2. HASIL AKURASI

Metode	Parameter	Akurasi Pelatihan	Akurasi Pengujian
Decision Tree	criterion='gini', splitter='best', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, random_state=8, min_impurity_decrease=0.0, ccp_alpha=0.0	1.00	0.89
Support Vector Machine	C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=True, tol=0.001, cache_size=200, verbose=False, max_iter=-1, decision_function_shape='ovr', break_ties=False, random_state=8	0.90	0.90
Logistic Regression	penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, random_state=8, solver='lbfgs', max_iter=1000, multi_class='auto', verbose=0, warm_start=False	0.90	0.90
KNN	n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski'	0.90	0.90
Voting Classifier	estimators=['random_forest', 'svm', 'logistic_regression', 'knn'], voting='soft', flatten_transform=True, verbose=False	0.96	0.90

Dari hasil penelitian yang dilakukan pada dataset tingkat stres, ditemukan bahwa akurasi pelatihan tertinggi mencapai 100% pada Model *Decision Tree*, sementara akurasi pengujian tertinggi sebesar 90% terdapat pada tiga Model, yaitu *Support Vector Machine, Logistic Regression*, dan KNN. Setelah dilakukan *Ensemble Learning* dengan metode *Voting Classifier*, ditemukan bahwa akurasi pelatihan mencapai 96%, sementara akurasi pengujian tetap 90%.

Dari tabel 2 dapat diketahui bahwasannya metode *Voting Classifier* mendapatkan akurasi pelatihan yang lebih baik, yaitu sebesar 96%. Hal ini disebabkan oleh Model *Decision Tree* yang memiliki akurasi pelatihan mencapai 100%, yang berpotensi mengalami *overfitting* pada model. Sementara itu, *Voting Classifier* memperoleh akurasi pengujian yang sama dengan tiga model terbaik sebelum dilakukan *Voting Classifier.*

IV. KESIMPULAN

Penelitian ini dilakukan untuk mengidentifikasi model yang paling baik dalam mengklasifikasikan tingkat stres. Dari hasil penelitian yang dilakukan, ditemukan akurasi pelatihan terbaik setelah dilakukan *Ensemble Learning* dengan metode *Voting Classifier*, dengan akurasi sebesar 96%. Sementara itu, ditemukan akurasi pengujian terbaik pada empat Model, yaitu *Support Vector Machine, Logistic Regression*, KNN, dan *Voting Classifier*, dengan akurasi sebesar 90%. Model *Ensemble Learning*, khususnya *Voting Classifier*, dapat menjadi pilihan yang baik dalam mengatasi potensi *overfitting* dari model *Decision Tree*. Hasil penelitian ini dapat digunakan sebagai dasar perbandingan dengan penelitian-penelitian sebelumnya untuk memahami faktor-faktor stres siswa.

REFERENSI

[1] S. A. Musabiq and I. Karimah, "Gambaran Stress Dan Dampaknya Pada Mahasiswa Description Of Stress And Its Impact On Students," *InSight*, vol. 20, no. 2, 2018.

[2] A. Diananda, "Psikologi Remaja Dan Permasalahannya," 2018. [Online]. Available: www.depkes.go.id

[3] M. Barseli, I. Ifdil, and N. Nikmarijal, "Konsep Stres Akademik Siswa," *Jurnal Konseling dan Pendidikan*, vol. 5, no. 3, pp. 143–148, Dec. 2017, doi: 10.29210/119800.

[4] L. Maretva Cendani and A. Wibowo, "Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes," 2022.

[5] R. I. Arumnisaa and A. W. Wijayanto, "SISTEMASI: Jurnal Sistem Informasi Perbandingan Metode Ensemble Learning: Random Forest, Support Vector Machine, AdaBoost pada Klasifikasi Indeks Pembangunan Manusia (IPM) Comparison of Ensemble Learning Method: Random Forest, Support Vector Machine, AdaBoost for Classification Human Development Index (HDI)." [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[6] A. N. Rais and A. Subekti, "Integrasi SMOTE dan Ensemble AdaBoost Untuk Mengatasi Imbalance Class Pada Data Bank Direct Marketing," *JURNAL INFORMATIKA*, vol. 6, no. 2, pp. 278–285, 2019, [Online]. Available: http://ejournal.bsi.ac.id/ejurnal/index.php/ji

Implementasi Regresi Linear untuk Memprediksi Persediaan Barang pada *E-Commerce*

Herlambang Kurniawan
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
herlambang.5200411434@student.uty.ac.id

Muhammad Ilham Triwibowo
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
muhammad.5200411416@student.uty.ac.id

Muhammad Hafidz Ghifary
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
muhammad.5200411415@student.uty.ac.id

Muhammad Satrio Gumilang
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
muhammad.5200411155@student.uty.ac.id

Dwi Nugroho Teguh
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
dwi.5200411177@student.uty.ac.id

Abstrak— *E-commerce* telah menjadi bagian integral dari dunia bisnis *modern*, memungkinkan perusahaan untuk menjual produk dan jasa secara *online* dengan efisien. Ketidaktepatan strategi ketika melakukan analisis terhadap produk penjualan dapat merugikan sebuah perusahaan baik itu dalam jangka panjang maupun dalam jangka pendek, oleh sebab itu perlu analisis yang tepat seperti predksi hasil penjualan agar dapat menghasilkan keuntungan bagi perusahaan. Penelitian ini menginvestigasi penerapan metode regresi linear untuk memprediksi penjualan *e-commerce* pada perusahaan ritel *online* berbasis di Inggris selama satu tahun. Hasil analisis menunjukkan bahwa promosi dan penentuan harga memiliki pengaruh positif dan signifikan terhadap perilaku pembelian impulsif konsumen. Temuan ini memberikan kontribusi berharga dalam memperdalam pemahaman tentang strategi promosi yang efektif dan manajemen stok yang optimal di lingkungan *e-commerce*. Penelitian ini juga mengukuhkan validitas regresi linear sebagai alat prediksi yang handal dalam ranah penjualan *e-commerce*, sambil menyajikan wawasan baru dalam dinamika pasar digital. Hasil. Dari hasil penelitian yang sudah dilakukan diambil 5 hasil terbaik dan hasil terbaik dihasilkan oleh produk *Cream Hanging Heart T-Light Holder* yang terjual di negara *United Kingdom* dengan *MSE* 1429.014 dan *RMSE* 37.802. Sedangkan untuk hasil terendah dari 5 terbaik yaitu dihasilkan oleh produk *World War 2 Gliders Asstd Designs* dengan *MSE* 3888.845 dan *RMSE* 62.361.

Kata Kunci—Regresi Linear, MSE, RMSE, E-commerce

I. PENDAHULUAN

E-commerce adalah proses jual beli barang atau jasa yang dilakukan secara *online* melalui internet. Kehadiran *e-commerce* dapat membantu perusahaan untuk menjangkau pasar yang lebih luas dengan biaya yang lebih rendah [1]. Kemajuan teknologi dan perkembangan sistem transaksi yang awalnya masih bersifat transaksi secara langsung saat ini sudah memasuki tahap dimana transaksi dapat dilakukan secara *online* [2]. Menghasilkan perkiraan penjualan tingkat produk merupakan faktor penting dalam industri ritel karena pengendalian penjualan dan perencanaan produksi memainkan peran penting dalam daya saing setiap perusahaan yang menyediakan barang untuk pelanggannya. Dalam pengelolaan penjualan barang yang dijual, kekurangan barang dapat menyebabkan penurunan keuntungan dan kepuasan pelanggan. Selain itu, kelebihan penjualan barang dapat memaksa toko untuk menjual barang dengan harga lebih rendah sehingga dapat menurunkan keuntungan yang didapat, atau bahkan lebih buruk lagi menyebabkan pengurangan penjualan, tingkat penjualan yang lebih tinggi dari yang dibutuhkan juga meningkatkan biaya pergudangan [3]. Banyaknya jenis produk yang dijual juga dapat menimbulkan manajemen persediaan barang menjadi tidak akurat dikarenakan permintaan konsumen yang berjumlah besar [4].

Oleh karena itu membuat sistem prediksi penjualan sebuah produk merupakan hal yang penting guna meningkatkan keuntungan atau profit bagi perusahaan. Contohnya dengan menggunakan pendekatan *machine learning* dapet menghasilkan sebuah *insight* guna meningkatkan produk hasil penjualan ataupun melakukan strategi khusus pada produk yang kurang terjual [5]. Penelitian tersebut peneliti menggunakan data penjualan pada PT XYZ yang diambil dari tahun 2014-2019. Peneliti menggunakan metode regresi linear untuk memprediksi jumlah unit yang terjual berdasarkan periode penjualan properti. Pada penelitian tersebut didapatkan hasil dengan MSE terendah yaitu sebesar 0,140 pada *property kavling*. Sedangkan untuk properti yang lain memiliki nilai *MSE* di atas 0,5 [6]. Promosi dan harga memiliki pengaruh positif dan signifikan terhadap perilaku *impulse buying* pengguna *Shopee* di Kota Pekanbaru. Hasil analisis regresi linear berganda menunjukkan bahwa peningkatan dalam strategi promosi maupun penurunan harga dapat secara langsung mempengaruhi keputusan pembelian impulsif konsumen. Temuan ini memberikan wawasan berharga bagi perusahaan dalam merancang kampanye promosi yang efektif [7].

Berdasarkan masalah dan penelitian sebelumnya yang sudah dipaparkan, menunjukkan metode regresi ini dapat digunakan dalam memprediksi penjualan barang pada *e-commerce*. Sehingga kami akan mengimplementasikan regresi linear pada dataset transaksi penjualan *e-commerce* (ritel *online*) yang berbasis di Inggris selama satu tahun.

II. KAJIAN PUSTAKA

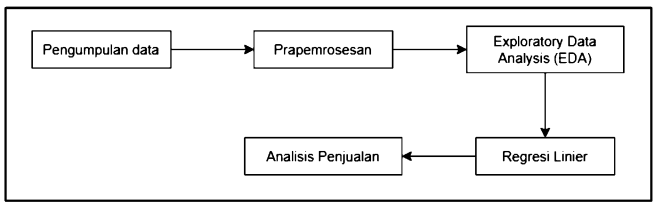
Regresi linear merupakan solusi yang cocok untuk digunakan oleh perusahaan multiproduk karena dengan memperkirakan berbagai kombinasi produk [8]. Hal ini dikarenakan dengan memperkirakan berbagai kombinasi

produk, perusahaan dapat memaksimalkan keuntungan serta memperkirakan jumlah produksi yang tepat [9]. Model prediksi sebelumnya telah digunakan untuk menentukan hasil penjualan pada waktu yang akan datang. Penelitian sebelumnya menggunakan metode regresi linear sederhana untuk memprediksi penjualan produk *unilever*. Data yang digunakan berupa data penjualan 50 produk *unilever* selama 15 bulan. Sebenyak 10 produk terlaris dipilih untuk diprediksi. Hasil prediksi diukur dengan menggunakan *Mean Absolute Percentage Error (MAPE)*. Nilai *MAPE* terendah adalah sebesar 1% untuk produk *Sunsilk Conditioner* dan nilai tertinggi didapat sebesar 10% pada produk *Vixal* [10]. Pada penelitian selanjutnya, dilakukan prediksi data penjualan dari *website shopee seller centre* menggunakan metode regresi linier berganda. Data yang digunakan merupakan data penjualan bulan Mei 2020 sampai April 2022 sebanyak 574 data. Hasil prediksi 30 data testing memperoleh hasil MSE sebesar 5.172.628.212.404, nilai RMSE sebesar 2.274.341,27, dan nilai MAPE sebesar 4,34% [11]. Penelitian menggunakan metode regresi linier untuk memprediksi penjualan pada PT. Eagle Industry Indonesia. Data yang digunakan berjumlah 301 data yang berisi penjualan 3 produk Perusahaan dan memiliki 5 atribut. Metode yang digunakan adalah metode regresi linear dengan *least square method* untuk menentukan persamaannya. Penelitian mendapatkan hasil RMSE sebesar 36241.241 +/- 0.000 dan Squared Error sebesar 1313427569.481 +/- 5882150128.134 [12].

Pada penelitian lain, metode regresi linear dan moving average digunakan untuk memprediksi data penjualan supermarket. Penelitan tersebut membandingkan kinerja kedua metode dalam memprediksi 148 data penjualan supermarket dari bulan Januari sampai Maret tahun 2019 pada kategori kesehatan dan elektronik. Hasil penelitian mendapatkan nilai RMSE dan MSE sebesar 7.106 dan 50.489 menggunakan moving average. Sedangkan metode regresi linear mendapatkan nilai RMSE dan MSE sebesar 7.59 dan 57.603 [13]. Penelitian selanjutnya memprediksi data penjualan supermarket menggunakan metode regresi linear berganda. Data yang digunakan berjumlah sebanyak 245.244 data dalam jangka waktu penjualan dari tahun 2018 sampai tahun 2019. Prediksi dibagi menjadi per-barang dan per-tahun dan mendapatkan hasil nilai RMSE sebesar 40,476 [14]. Penelitian lainnya membandingkan metode regresi linear dan moving average dalam memprediksi data penjualan produk bearing pada CV Mulia Tata Sejahtera dari bulan Januari 2019 sampai Desember 2020. Hasilnya metode regresi linear mendapatkan nilai MSE sebesar 164.650 dan metode moving average mendapatkan nilai MSE sebesar 7.376.044. Dari hasil tersebut metode regresi linear mendapatkan nilai error yang relatif lebih kecil dari metode moving average [15].

III. METODOLOGI PENELITIAN

Analisis yang didapatkan akan berupa hasil prediksi menggunakan model regresi linear membutuhkan beberapa proses agar dataset terbaca dengan baik. Setelah beberapa proses dilakukan, langkah berikutnya adalah menginterpretasi hasil prediksi. Dapat dilakukan evaluasi signifikansi statistik dari masing-masing koefisien regresi untuk memahami seberapa besar pengaruhnya terhadap variabel dependen. Koefisien yang signifikan secara statistik menunjukkan bahwa variabel tersebut memberikan kontribusi yang nyata terhadap prediksi model. Pada gambar 1 terdapat proses penelitian yang dimulai dari pengumpulan data, prapemrosesan, *exploratory data analysis,* lalu melakukan prediksi, dan analisis penjualan untuk hari berikutnya. Berikut penjelasan dari gambar 1 dari proses penelitian yang dilakukan.



Gambar 1. Alur Penelitian

A. Pengumpulan Data

Dalam kasus kumpulan data transaksi penjualan *e-commerce* berbasis di Inggris, data tersebut didapatkan melalui Kaggle. Kaggle adalah platform *online* yang menyediakan kumpulan dataset dan tantangan (*challenges*) untuk para data scientist dan pengembang perangkat lunak. Dataset tersebut berisi transaksi penjualan *e-commerce* (ritail *online*) yang beroperasi di Inggris selama satu tahun. Toko yang berlokasi di London telah menjual berbagai oleh-oleh dan peralatan rumah tangga, baik untuk orang dewasa maupun anak-anak, melalui situs webnya sejak tahun 2007. Pelanggan toko ini berasal dari berbagai belahan dunia dan umumnya melakukan pembelian secara langsung. Selain itu, terdapat juga usaha kecil yang membeli produk dalam jumlah besar dan menjualnya kepada pelanggan lain melalui saluran gerai ritel. Sampel data yang digunakan pada penelitian ini terdapat pada gambar 2 dibawah ini.

	TransactionNo	Date	ProductNo	ProductName	Price	Quantity	CustomerNo	Country
0	581482	12/9/2019	22485	Set Of 2 Wooden Market Crates	21.47	12	17490.0	United Kingdom
1	581475	12/9/2019	22596	Christmas Star Wish List Chalkboard	10.65	36	13069.0	United Kingdom
2	581475	12/9/2019	23235	Storage Tin Vintage Leaf	11.53	12	13069.0	United Kingdom
3	581475	12/9/2019	23272	Tree T-Light Holder Willie Winkie	10.65	12	13069.0	United Kingdom
4	581475	12/9/2019	23239	Set Of 4 Knick Knack Tins Poppies	11.94	6	13069.0	United Kingdom

Gambar 2. Sampel Dataset

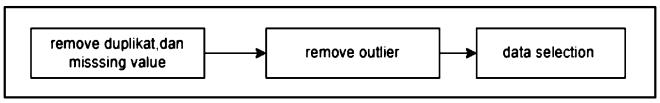
	TransactionNo	Date	ProductNo	ProductName	Price	Quantity	CustomerNo	Country
6511	C581406	12/8/2019	46000M	Polyester Filler Pad 45x45cm	6.19	-240	NaN	United Kingdom
6512	C581406	12/8/2019	46000S	Polyester Filler Pad 40x40cm	6.19	-300	NaN	United Kingdom
90098	C575153	11/8/2019	22947	Wooden Advent Calendar Red	44.25	-1	NaN	United Kingdom
102671	C574288	11/3/2019	22178	Victorian Glass Hanging T-Light	25.37	-1	NaN	United Kingdom
117263	C573180	10/28/2019	23048	Set Of 10 Lanterns Fairy Light Star	14.50	-1	NaN	United Kingdom

Gambar 3. Sampel Missing Value

Pada dataset memiliki 536350 baris data dan 8 kolom yang terdiri dari TranscationNo(kategorik), Date(numerik), ProducNo(kategorik), Product(kategorik), Price(numerik), Quantity(numerik), CustomerNo(kategorik), dan Country(kategorik). Dari gambar 3 dapat diketahui bahwasannya pada dataset tersebut masih terdapat *missing value* pada kolom CustomerNo sebanyak 55 data sehingga nantinya akan dilakukan prapemrosesan untuk menghilangkan *missing value*.

B. Prapemrosesan

Dalam prapemrosesan data untuk dataset transaksi penjualan *e-commerce* berbasis di Inggris, beberapa langkah kritis telah dilakukan. Langkah yang dilakukan berdasarkan gambar 4.



Gambar 4. Alur Prapemrosesan

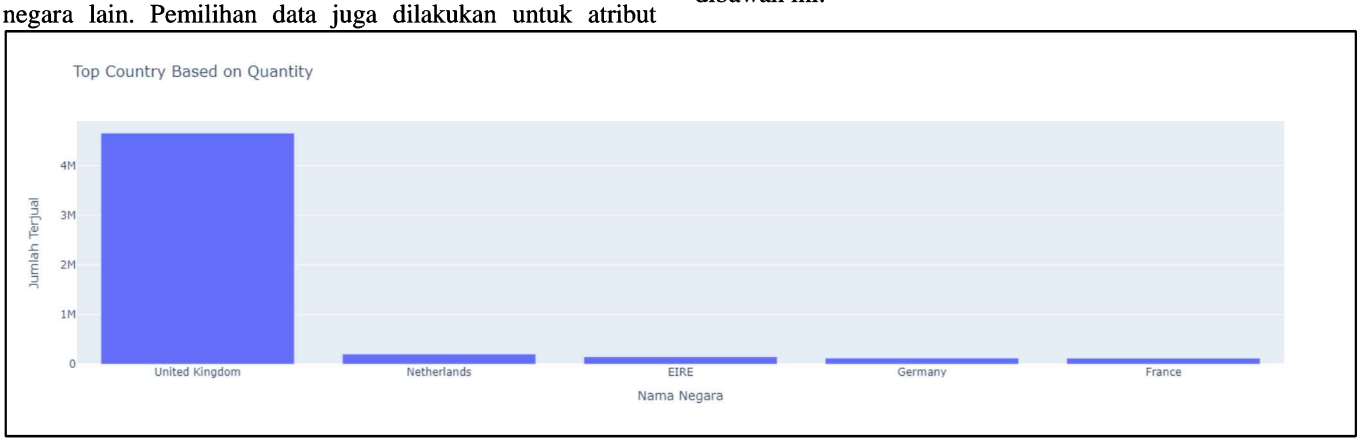
Pertama, duplikat data dihapus untuk memastikan konsistensi dataset dan menghilangkan gangguan dari data ganda. Selanjutnya, penanganan missing value dilakukan untuk memastikan ketepatan analisis dengan mengatasi atau menghapus data yang memiliki nilai yang hilang. Atribut date diformat ulang agar memiliki konsistensi format, memudahkan analisis waktu, dan pengelompokan data berdasarkan tanggal. Langkah berikutnya adalah mengidentifikasi dan menghapus outlier pada atribut Quantity, karena nilai ekstrem dapat mempengaruhi hasil analisis.

Fokus pada negara Inggris (UK) dilakukan dengan memilih hanya data penjualan yang berkaitan dengan negara tersebut. Hal ini bertujuan untuk mendapatkan konteks yang lebih spesifik dan relevan, menghilangkan data dari negara-negara lain. Pemilihan data juga dilakukan untuk atribut ProductName dengan mengambil hanya 5 record ProductName yang paling banyak terjual. Langkah ini membantu mengidentifikasi produk-produk yang paling diminati oleh pelanggan, memfokuskan analisis pada produk yang paling signifikan dalam konteks penjualan. Dengan melakukan prapemrosesan ini, dataset menjadi lebih bersih, terstruktur, dan siap untuk dilibatkan dalam analisis lebih lanjut, mencegah bias, dan mengoptimalkan hasil analisis.

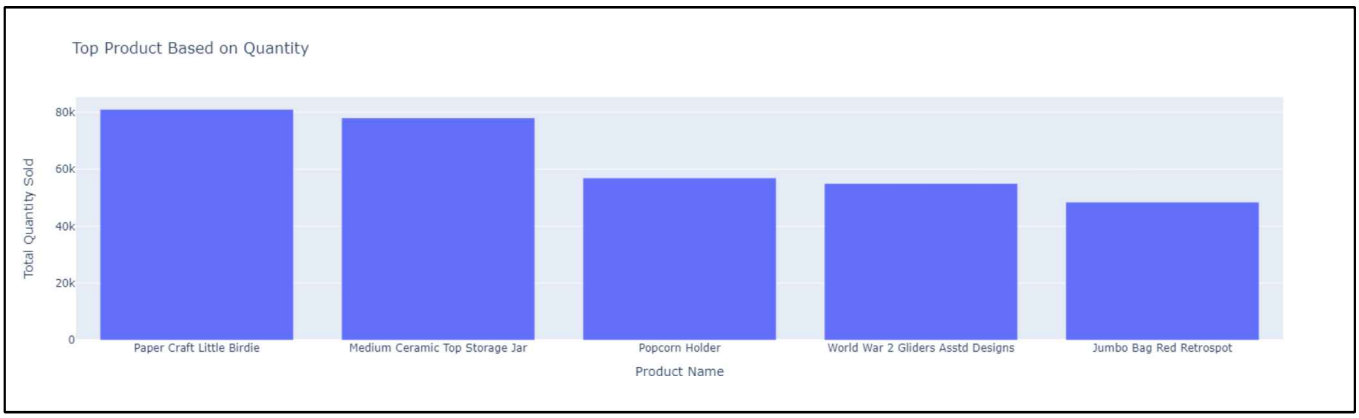
Terakhir, pembuatan kolom baru bernama TotalMoney dengan mengalikan nilai pada kolom Price dengan kolom Quantity. Langkah ini bertujuan untuk menambah dimensi analisis dengan memperoleh informasi tentang total pendapatan yang dihasilkan dari setiap transaksi. Dengan menambahkan kolom TotalMoney, kita dapat melihat secara langsung kontribusi masing-masing transaksi terhadap pendapatan total toko secara keseluruhan. Proses ini memberikan wawasan tambahan yang dapat berguna dalam pemahaman performa penjualan dan produk-produk yang paling menguntungkan dalam dataset tersebut. Dengan demikian, analisis lebih lanjut dapat mencakup aspek keuangan yang lebih mendalam untuk mendukung pengambilan keputusan yang lebih baik.

C. Exploratory Data Analysis (EDA)

Dalam proses *Exploratory Data Analysis (EDA)* ini dilakukan pendekatan analisis data untuk memahami dan menjelajahi data, menemukan pola dan hubungan yang tidak terduga, serta menguji hipotesis awal pada data yang sudah dilakukan pemrosesan sebelumnya. Dari proses ini ditemukan bahwa 5 negara dengan tingkat penjualan jumlah barang terbanyak adalah United Kingdom, Netherlands, EIRE, Germany, dan France. United Kingdom menempati tempat pertama dengan jumlah penjualan barang lebih dari 4+ juta barang. Untuk lebih jelasnya dapat dilihat pada Gambar 5 dibawah ini.



Gambar 5. Top 5 Negara berdasarkan Quantity



Gambar 6. Top 5 Produk berdasarkan Quantity dari Negara United Kingdom

Kemudian pada proses ini dipilih United Kingdom yang kemudian dilakukan EDA pada produknya yang terjual disana. Didapati 5 produk dengan tingkat penjualan tertinggi yaitu *Paper Craft Little Birdie*, *Medium Ceramic Top Storage Jar*, *Popcorn Holder*, *World War 2 Gilders Asstd Design*, dan *Jumbo Bag Red Retrospot*. Paper Craft Little Birdie menempati tempat pertama dengan jumlah penjualan barang lebih dari 80,000+ barang. Untuk lebih jelasnya dapat dilihat pada Gambar 6.

D. Regresi Linear

Metode regresi merupakan sebuah metode statistik yang melakukan prediksi menggunakan pengembangan hubungan matematis antara variabel, yaitu variabel dependen (Y) dengan variabel independen (X). Variabel dependen merupakan variabel akibat atau variabel yang dipengaruhi, sedangkan variabel independen merupakan variabel sebab atau variable yang mempengaruhi. Prediksi terhadap nilai variable dependen dapat dilakukan jika variabel independennya diketahui. Umumnya penjualan atau permintaan suatu produk dinyatakan sebagai variabel dependen yang besar atau nilainya dipengaruhi oleh variabel independen.

Regresi linear menjadi salah satu metode yang dipergunakan dalam produksi untuk melakukan peramalan atau prediksi tentang karakteristik kualitas maupun kuantitas. Hal ini dikarenakan dengan memperkirakan berbagai kombinasi produk, perusahaan dapat memaksimalkan keuntungan serta memperkirakan jumlah produksi yang tepat.

Rumus untuk Regresi Linear dengan metode kuadrat terkecil atau sederhana adalah:

$$a = \frac{(\sum y)(\sum x^2)-(\sum x)(\sum xy)}{n(\sum x^2)-(\sum x)^2} \quad (1)$$

$$b = \frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2} \quad (2)$$

$$y = a + b.x \quad (3)$$

dengan y adalah kuantiti penjualan, x adalah periode penjualan atau bulan penjualan, a adalah konstanta yang menunjukan besarnya nilai y apabila x = 0, dan b adalah besaran perubahan nilai y.

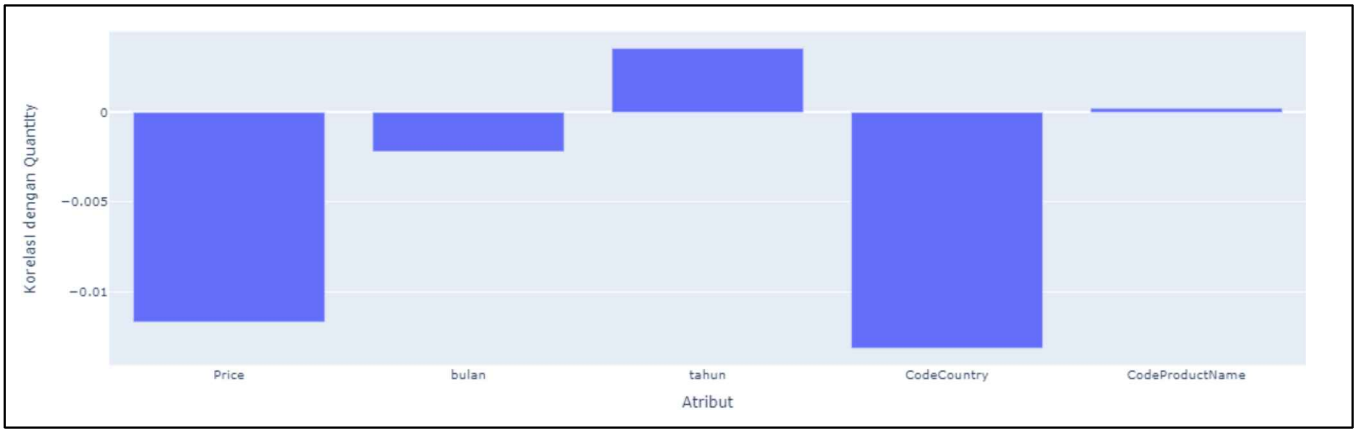
IV. HASIL DAN ANALISIS

Dalam bagian ini, akan diulas evaluasi dan analisis hasil dari penelitian yang telah dilakukan, terfokus pada konteks *e-commerce*. Sehingga dapat diketahui hasil atau strategi yang sesuai untuk meningkatkan keuntungan dalam penjualan.

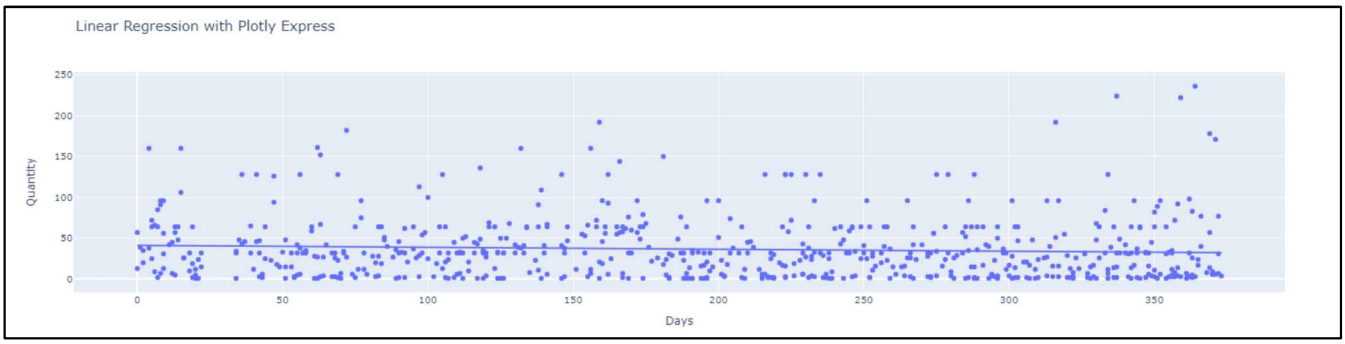
A. Regresi Linear

Sebelum dilakukan regresi linear dilakukan korelasi antar *attribut* terlebih dahulu agar dapat dilakukan regresi linear. Dari gambar 7 diperoleh nilai korelasi antara jumlah barang terjual dan tahun adalah 0,5. Nilai korelasi ini menunjukkan bahwa ada hubungan positif yang kuat antara kedua variabel tersebut. Artinya, semakin tinggi tahunnya, semakin tinggi pula jumlah barang terjual. Nilai korelasi antara jumlah barang terjual dan bulan adalah -0,2. Nilai korelasi ini menunjukkan bahwa ada hubungan negatif yang lemah antara kedua variabel tersebut. Artinya, semakin tinggi bulannya, semakin rendah pula jumlah barang terjual. Nilai korelasi antara jumlah barang terjual dan kode negara adalah 0,01. Nilai korelasi ini menunjukkan bahwa ada hubungan yang sangat lemah antara kedua variabel tersebut. Artinya, kode negara tidak memiliki pengaruh yang signifikan terhadap jumlah barang terjual. Nilai korelasi antara jumlah barang terjual dan kode produk adalah -0,005. Nilai korelasi ini menunjukkan bahwa ada hubungan yang sangat lemah antara kedua variabel tersebut. Artinya, kode produk tidak memiliki pengaruh yang signifikan terhadap jumlah barang terjual. Berdasarkan hasil analisis regresi linear, dapat disimpulkan bahwa faktor utama yang mempengaruhi jumlah barang terjual tersebut adalah tahun. Faktor lain, seperti bulan, kode negara, dan kode produk, memiliki pengaruh yang sangat lemah.

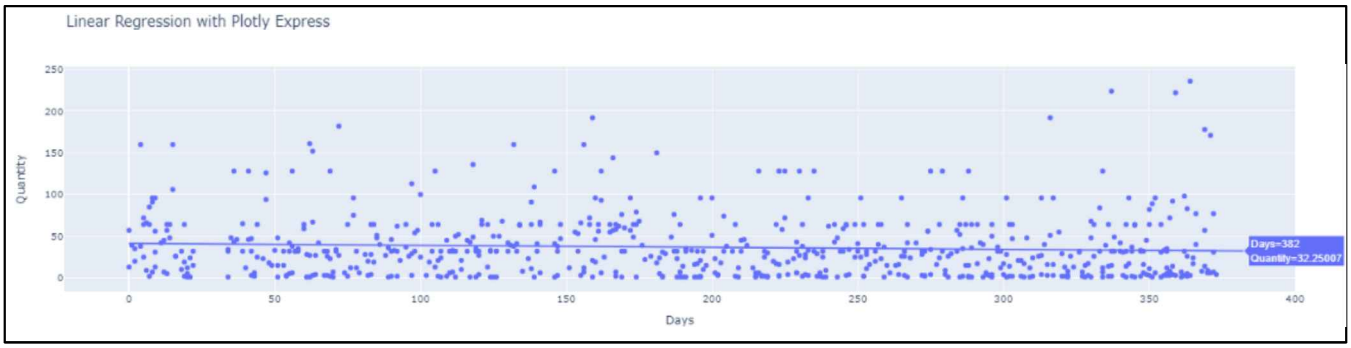
Kemudian dilakukan pelatihan menggunakan regresi linear pada produk *Cream Hanging Heart T-Light Holder* dengan negara yang dipilih yaitu *United Kingdom* dan dihasilkan data seperti Gambar 8 dibawah ini. Banyak *quantity* barang dari 0 hingga 250 dan banyak hari dari 1 hingga 365. Kemudian dilakukan prediksi dengan regresi linear pada produk *Cream Hanging Heart T-Light Holder* dan dihasilkan data seperti Gambar 9 dibawah. Banyak *quantity* sebanyak 0 hingga 250 dan banyak hari dari 0 hingga 400 hari, jadi disini dilakukan prediksi 35 hari kedepan dan *quantity* tersebut diprediksi menurun.



Gambar 7. Korelasi Dengan Attribut Quantity



Gambar 8. Hasil Latih Regresi Linear



Gambar 9. Hasil Prediksi Regresi Linear

B. *Evaluasi dan Analisis*

Kemudian dilakukan perhitungan nilai *MSE* pada data latih untuk menghitung rata-rata dari selisih kuadrat antara nilai prediksi. Kemudian dihasilkan 5 *MSE* terbaik yang dapat dilihat pada Tabel 1 dibawah. Setelah itu dilakukan perhitungan nilai *RMSE* pada data latih untuk menghitung nilai rata-rata dari akar kuadrat dari selisih kuadrat antara nilai prediksi model dan nilai aktual dari data. Dihasilkan *RMSE* dengan hasil seperti Tabel 1 dibawah.

TABEL 1. HASIL EVALUASI

Nama Barang	**MSE**	**RMSE**
Cream Hanging Heart T-Light Holder	1429.014	37.802
Popcorn Holder	2571.672	50.712
Jumbo Bag Red Retrospot	2780.673	52.732
Assorted Colour Bird Ornament	2794.116	52.859

Berdasarkan hasil yang sudah didapatkan penyebab tingginya nilai *MSE* dan *RMSE* adalah karena jumlah *attribut* yang digunakan hanya sedikit. Pada pelatihan model *attribut* yang digunakan hanya *attribut* data dan *quantity*, hal ini karena pada *dataset* yang digunakan memiliki nilai korelasi yang cukup kecil. Sehingga antara *attribut* yang ada dengan *attribut label* yaitu *quantity* tidak memiliki hubungan atau memiliki hubungan yang kecil. Selain dari *attribut* hal lain yang membuat nilai *MSE* dan *RMSE* menjadi tinggi adalah adanya *outlier* atau data pencilan. *Outlier* ini terjadi pada data yang memiliki *quantity* jauh lebih banyak dibandingkan rata-rata *quantity* pada data yang ada. Kemudian penyebab terakhir yang membuat hasil *MSE* dan *RMSE* tinggi adalah penggunaan model *linear regression* yang kurang sesuai dengan studi kasus yang ada.

V. KESIMPULAN DAN PENGEMBANGAN

Setelah dilakukan beberapa proses didapatkan beberapa hasil penelitian yang kemudian diambil 5 hasil terbaik dan hasil terbaik dihasilkan oleh produk *Cream Hanging Heart T-Light Holder* dengan *MSE* 1429.014 dan *RMSE* 37.802. Sedangkan untuk hasil terendah dari 5 terbaik yaitu dihasilkan oleh produk *World War 2 Gliders Asstd Designs* dengan *MSE* 3888.845 dan *RMSE* 62.361. Untuk pengembangan pada penelitian selanjutnya dapat dilakukan dengan mengganti atau menambahkan metode lain serta dapat juga dilakukan penambahan atau pengurangan *attribut* lain.

REFERENSI

[1] Hendrawan, R.A., Nurkasanah, I., Suryani, E., Er, M., Aristio, A.P., Puspita, M. and Saputra, N.A., 2021. Pengembangan eCommerce Multi Kanal untuk UMKM Jajanan & Minuman Produk Lokal di Surabaya. Sewagati, 5(1), pp.94-99.

[2] Siregar, L.Y. and Nasution, M.I.P., 2020. Perkembangan Teknologi Informasi Terhadap Peningkatan Bisnis Online. HIRARKI: Jurnal Ilmiah Manajemen Dan Bisnis, 2(1), pp.71-75.

[3] Zunic, E., Korjenic, K., Hodzic, K. and Donko, D., 2020. Application of facebook's prophet algorithm for successful sales forecasting based on real-world data. arXiv preprint arXiv:2005.07575.

[4] Arisandi, A. and Ependi, U., 2023. Analisis Peramalan Penjualan Produk Pada PT. Enseval Putera Megatrading TBK Menggunakan Metode Regresi Linear Sederhana. JUPITER (Jurnal Penelitian Ilmu dan Teknik Komputer), 15(1b), pp.317-326.

[5] Setyawan, A.R.T., 2022. Implementasi artificial intelligence marketing pada E-commerce: personalisasi konten rekomendasi serta dampaknya terhadap purchase intention. Fair Value: Jurnal Ilmiah Akuntansi dan Keuangan, 4(12), pp.5385-5392.

[6] Ayuni, G.N. and Fitrianah, D., 2019. Penerapan metode Regresi Linear untuk prediksi penjualan properti pada PT XYZ. Jurnal telematika, 14(2), pp.79-86.

[7] Chan, G. F., Akhmad, I., & Hinggo, H. T. (2022). Pengaruh Promosi Dan Harga Terhadap Impulse Buying Pada Pengguna E-Commerce Shopee Di Pekanbaru. ECOUNTBIS: Economics, Accounting and Business Journal, 2(1), 151-159.

[8] Indarwati, T., Irawati, T. and Rimawati, E., 2019. Penggunaan Metode Linear Regression Untuk Prediksi Penjualan Smartphone. Jurnal Teknologi Informasi Dan Komunikasi (TIKomSiN), 6(2).

[9] Ayuni, G.N. and Fitrianah, D., 2019. Penerapan metode Regresi Linear untuk prediksi penjualan properti pada PT XYZ. Jurnal telematika, 14(2), pp.79-86.

[10] Anggrawan, A., Hairani, H. and Azmi, N., 2022. Prediksi Penjualan Produk Unilever Menggunakan Metode Regresi Linear. Jurnal Bumigora Information Technology (BITe), 4(2), pp.123-132.

[11] Syakir, Y., Hermanto, T.I. and Ramadhan, Y.R., 2022. Analisis Marketplace Shopee Untuk Memprediksi Penjualan dengan Algoritma Regresi Linier. J-SAKTI (Jurnal Sains Komputer dan Informatika), 6(2), pp.904-915.

[12] Sunge, A.S. and Zy, A.T., 2023. ANALISIS PREDIKSI PENJUALAN DENGAN METODE REGRESI LINEAR DI PT. EAGLE INDUSTRY INDONESIA. Jurnal Informatika Teknologi dan Sains (Jinteks), 5(3), pp.398-403.

[13] Nafi'iyah, N. and Rakhmawati, E., 2021. Analisis Regresi Linear Dan Moving Average Dalam Memprediksi Data Penjualan Supermarket. Jurnal Teknologi Informasi dan Komunikasi, 12(1), pp.44-50.

[14] Asohi, Y. and Andri, A., 2020. Impelementasi Algoritma Regresi Linier Berganda Untuk Prediksi Penjualan. Jurnal Nasional Ilmu Komputer, 1(3), pp.149-158.

[15] Ubaidilah, U., Herwanto, D. and Nugraha, G.A., 2022. Peramalan Penjualan Bearing di CV. Mulia Tata Sejahtera Menggunakan Metode Single Moving Average Dan Regresi Linier. Jurnal Ilmiah Wahana Pendidikan, 8(19), pp.591-598.

Sistem Penilaian Kinerja Berbasis Laporan Penugasan Karyawan di PT Sinar Palasari Indonesia

Edward Paundra Amasa Exelpatria
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
edward.bhuztan@gmail.com

Muhammad Zakariyah
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
muhammad.zakariyah@staff.uty.ac.id

Enny Itje Sela
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
ennysela@uty.ac.id

Abstrak — PT Sinar Palasari Indonesia belum memiliki sistem informasi khusus untuk mengelola data tugas dan data karyawan. Banyaknya data yang diolah membuat admin seringkali mengalami kesulitan dalam mengelola data karyawan dan data tugas, terutama untuk mengevaluasi kinerja karyawan. Dikembangkannya sistem penilaian kinerja berdasarkan penugasan karyawan diharapkan dapat mempermudah admin dalam mengelola penugasan dan memperoleh informasi terkait kinerja karyawan terhadap tugas yang telah diberikan. Sistem dikembangkan melalui tahapan analisis dan definisi kebutuhan, desain sistem, implementasi dan pengujian unit, integrasi dan pengujian sistem, serta pemeliharaan sistem. Pengujian *black box* yang telah dilakukan menunjukkan hasil bahwa semua fitur telah berjalan sebagaimana mestinya. Pengujian terhadap penerimaan pengguna (*user acceptance test*) juga memberikan hasil yang disetujui oleh pengguna, baik oleh admin dengan persentase sebesar 83,81%, maupun oleh karyawan dengan tingkat persentasee sebesar 78,86%.

Kata Kunci — Sistem Informasi, Manajemen Tugas Karyawan, Android.

I. PENDAHULUAN

PT Sinar Palasari Indonesia adalah perusahaan yang didirikan pada tahun 2008, yang berfokus pada *tower maintenance*. Sebagai bagian dari industri yang sangat kompetitif, keberhasilan perusahaan ini sangat bergantung pada kinerja efektif dan efisien dari setiap karyawan. Oleh karena itu, penilaian kinerja menjadi suatu aspek kritis dalam memastikan bahwa sumber daya manusia perusahaan berkontribusi secara optimal terhadap pencapaian tujuan bisnis.

Era digital yang terus berkembang membuat sistem penilaian kinerja karyawan memainkan peran penting [3], [4]. Sistem informasi dapat mempermudah proses kerja, terutama dalam hal kecepatan dan ketepatan data, maupun pengelolaan data jumlah besar [1]. Penggunaan sistem yang terkomputerisasi memberikan banyak keuntungan, antara lain mempermudah dalam mengakses data, dan memperoleh informasi yang akurat, cepat, dan tepat [2]. Perusahaan telah mengimplementasikan sistem penilaian kinerja berbasis penugasan. Namun, dalam pelaksanaan penilaian kinerja berbasis penugasan, beberapa masalah muncul dan menimbulkan dampak negatif terhadap perusahaan dan karyawan. Beberapa masalah tersebut, seperti ketidakjelasan penugasan, ketidaksetaraan beban kerja, kurangnya umpan balik konstruktif, dan lain sebagainya.

Masalah-masalah tersebut, jika tidak ditangani dengan baik, dapat mengakibatkan dampak merugikan, seperti penurunan produktivitas, peningkatan *turnover* karyawan, dan bahkan dapat mempengaruhi citra perusahaan di mata karyawan dan pelanggan.

Sebagai upaya untuk memperbaiki proses penilaian kinerja yang efektif, aplikasi Android menjadi salah satu solusi alternatif yang cukup bagus [5], [6]. Penerapan sistem penilaian kinerja karyawan berbasis penugasan melalui aplikasi Android diharapkan dapat meningkatkan efisiensi proses penilaian, meningkatkan akurasi penilaian, serta mendorong keterlibatan karyawan dalam mengelola tugas dan tanggungjawabnya. Melalui kemudahan akses dan fleksibilitas yang ditawarkan oleh aplikasi Android, diharapkan PT Sinar Palasari Indonesia dapat mencapai keseimbangan yang lebih baik antara kebutuhan perusahaan dan kepuasan karyawan.

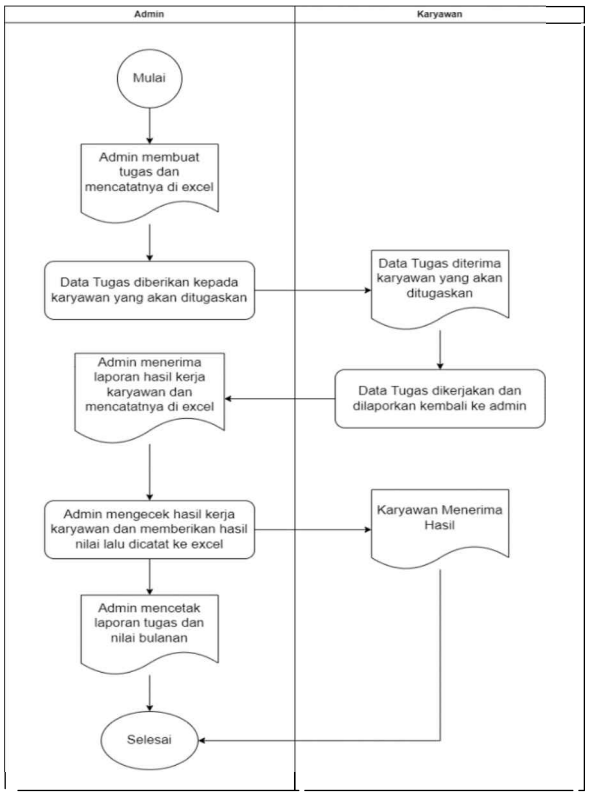
Penelitian ini bertujuan untuk menganalisis efektivitas dan efisiensi sistem penilaian kinerja berbasis penugasan karyawan di PT Sinar Palasari Indonesia, serta mengidentifikasi solusi dan perbaikan yang dapat diimplementasikan untuk mengatasi masalah yang dihadapi. Dengan menginvestigasi secara mendalam masalah-masalah ini, diharapkan penelitian ini dapat memberikan pandangan yang lebih jelas tentang penilaian kinerja di PT Sinar Palasari Indonesia dan memberikan kontribusi pada pengembangan strategi yang lebih efektif untuk manajemen kinerja di perusahaan tersebut.

II. METODE PENELITIAN

A. *Pengumpulan Data*

Penelitian ini menggunakan beberapa tahapan untuk mendapatkan data, diantaranya: observasi, wawancara dan studi pustaka. Pada tahapan observasi, dilakukan penggalian mengenai proses atau alur bisnis penugasan karyawan yang terjadi di PT Sinar Palasari Indonesia. Tahapan ini dilakukan untuk mendapatkan gambaran yang jelas mengenai sistem manajemen tugas karyawan yang sedang berjalan saat ini (Gambar 1).

Selain observasi, pengumpulan data melalui wawancara juga dilakukan agar informasi yang diperoleh lebih detail dan lengkap. Wawancara dilakukan dengan atasan PT Sinar Palasari Indonesia. Melalui wawancara tersebut, didapatkan informasi mengenai sistem manajemen tugas karyawan, informasi mengenai proyek, dan kesulitan apa saja yang dihadapi selama menjalankan proses bisnis. Tahapan studi pustaka dilakukan dengan mencari akar atau sumber yang berkaitan dengan data-data yang akan digunakan untuk



Gambar 1. Alur sistem manajemen tugas karyawan yang berjalan saat ini

melakukan penelitian dalam merancang sistem yang dibutuhkan.

B. *Penilaian Kinerja Karyawan*

Penilaian kinerja karyawan yang digunakan pada sistem, merujuk pada peraturan yang ditetapkan oleh PT Sinar Palasari Indonesia kepada karyawan saat menjalankan tugas yang diberikan. Tabel 1 merupakan rumus penilaian yang digunakan dalam proses penilaian tugas.

TABEL 1. KOMPONEN PENILAIAN KINERJA KARYAWAN

Komponen Penilaian	Deskripsi
Tugas Mulai (TM)	Tanggal yang diberikan oleh admin untuk mulai mengerjakan tugas.
Tugas Deadline (TD)	Tanggal yang diberikan admin sebagai batas terakhir dalam mengerjakan tugas.
Tugas Selesai (TS)	Tanggal saat tugas selesai dikerjakan oleh karyawan.
Nilai Bonus (NB)	Nilai yang diberikan oleh sistem, jika TS kurang dari TD. (Nilai = +10)
Nilai Penalti (NP)	Nilai yang diberikan oleh sistem, jika TS lebih dari TD. (Nilai = -10)
Nilai Tugas (NTgs)	Nilai dasar yang diberikan oleh sistem pada saat TM dengan durasi pengerjaan normal adalah 7 hari. (Nilai = 100)
Lama Pekerjaan (LP)	Waktu yang dibutuhkan oleh karyawan untuk menyelesaikan tugas. Rumus yang digunakan adalah: **LP = TS – TM**
Nilai Total (NTotal)	Nilai akhir yang diberikan oleh sistem kepada karyawan, setelah menyelesaikan tugas yang diberikan. 1. Jika LP ≤ 7, maka: **NTotal = NTgs + (NB x LP)** 2. Jika 7 < LP < 17, maka: **NTotal = NTgs + (NP x LP)** 3. Jika LP ≥ 17, maka: **NTotal = 0**

C. *Metode Pengembangan Sistem*

Metode pengembangan sistem pada penelitian ini menggunakan Metode Waterfall. Metode Waterfall terdiri atas 5 tahapan: Analisis dan Definisi Kebutuhan Sistem, Desain Sistem, Impementasi dan Pengujian Unit, Integrasi dan Pengujian Sistem, dan Pemeliharaan Sistem [7].

1) *Analisis dan Definisi Kebutuhan Sistem*

Proses analisis kebutuhan melibatkan identifikasi data yang digunakan untuk mengembangkan sistem penilaian kinerja berbasis penugasan karyawan. Kemampuan dan fitur-fitur dalam sistem tersebut terbagi menjadi sistem yang berbasis Web (untuk admin) dan berbasis android (untuk karyawan). Sistem yang dikembangkan, dapat mencatat, mengelola, dan memberikan informasi tugas yang dilakukan oleh admin kepada karyawan di PT Sinar Palasari Indonesia. Tabel 2 menunjukkan fitur sistem berbasis web maupun android.

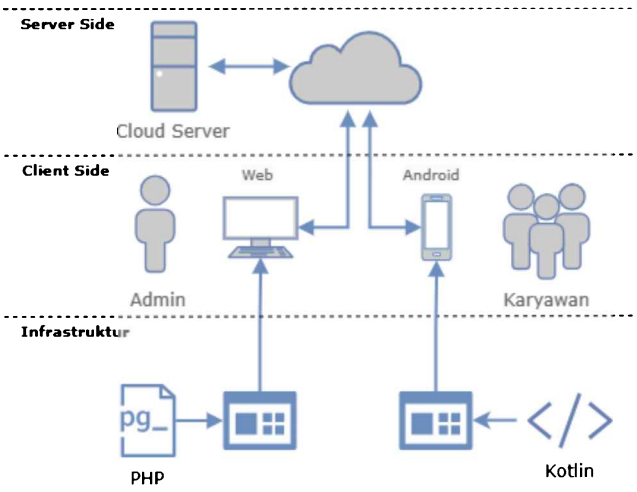
TABEL 2. FITUR UTAMA SISTEM

No	Fitur	Web (Admin)	Android (Karyawan)
1	Login	✓	✓
2	Dashboard	✓	✓
3	Kelola Data Karyawan	✓	
4	Kelola Data Tugas	✓	✓
5	Laporan Bulanan	✓	
6	Penilaian Karyawan	✓	✓

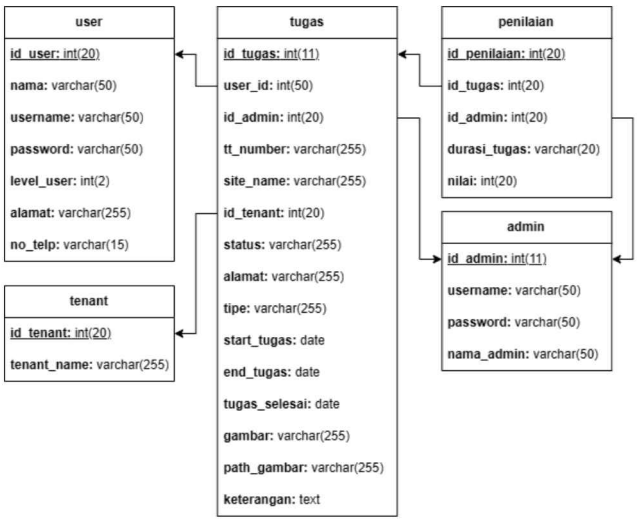
2) *Desain Sistem*

Desain alur sistem digambarkan dalam bentuk pemodelan *logic* dan fisik yang berisikan proses dari sistem yang akan dibuat dan diimplementasikan. Adapun pemodelan *logic* digambarkan dengan menggunakan arsitektur model seperti tampak pada Gambar 2.

Pemodelan fisik berupa relasi antartabel yang merupakan keterkaitan dari tabel-tabel yang ada dalam satu *database*. Pemodelan ini terbentuk melalui skema diagram yang terdiri dari penggabungan *primary key* dan *foreign key*,



Gambar 2. Alur sistem manajemen tugas karyawan yang berjalan saat ini



Gambar 3. Relasi antartabel

yang merupakan visual dari struktur hubungan antarentitas. Hubungan tabel satu dengan yang lainnya dapat dilihat pada Gambar 3.

3) Implementasi dan Pengujian Unit

Unit testing dilakukan menggunakan metode *black box testing* yang bertujuan untuk menemukan kesalahan fungsi dalam program. Proses ini melibatkan masukan/*input*

TABEL 3. KUESIONER PENGUJIAN PENERIMAAN PENGGUNA

No	Item Pertanyaan
Pertanyaan untuk Admin	
1.	Seberapa mudah Anda menavigasi antarmuka sistem penilaian kinerja karyawan?
2.	Seberapa jelas informasi yang disajikan dalam laporan kinerja karyawan?
3.	Sejauh mana sistem ini membantu Anda dalam mengevaluasi kinerja karyawan secara efektif?
4.	Seberapa responsif sistem ini terhadap permintaan Anda?
5.	Sejauh mana Anda merasa sistem ini memperbaiki proses penilaian kinerja karyawan?
6.	Seberapa mudah sistem ini dalam mengelola data karyawan dan laporan kinerja?
7.	Seberapa cocok fitur-fitur yang disediakan dengan kebutuhan Anda sebagai administrator?
Pertanyaan untuk Karyawan	
1.	Seberapa mudah Anda mengakses dan menggunakan sistem penilaian kinerja karyawan melalui perangkat mobile?
2.	Seberapa jelas dan komprehensif informasi yang disajikan dalam laporan kinerja Anda?
3.	Seberapa membantu sistem ini dalam meningkatkan pemahaman Anda tentang kinerja Anda sendiri?
4.	Sejauh mana sistem ini memberikan umpan balik yang berguna untuk meningkatkan kinerja Anda?
5.	Seberapa responsif sistem ini terhadap permintaan atau input yang Anda berikan?
6.	Sejauh mana Anda merasa sistem ini membantu memperbaiki transparansi dalam proses penilaian kinerja?
7.	Seberapa mudah sistem ini digunakan dalam menetapkan dan memantau tujuan kinerja pribadi Anda?

tertentu dan pengamatan terhadap hasilnya. Metode *black box testing* menguji kelayakan masukan dan keluaran sistem. Pengujian melibatkan pengguna yang memberikan *input* kepada sistem yang berjalan dan mengamati *output* yang dihasilkan. Setiap proses diuji untuk memastikan kesesuaian fungsi perangkat lunak.

4) Integrasi dan Pengujian Sistem

Tahap integrasi digunakan untuk menerapkan sistem yang sudah siap, untuk digunakan oleh pengguna akhir (admin/karyawan). Untuk mengevaluasi respon pengguna terhadap sistem yang telah dikembangkan, dilakukan pengujian sistem menggunakan Metode Skala Likert dalam pengujian penerimaan pengguna (*user acceptance test*). Proses pengujian ini melibatkan partisipasi dari beberapa pengguna yang memiliki peran dan tanggung jawab terkait penggunaan sistem, seperti admin dan karyawan.

Pengguna secara langsung menguji *prototipe* dengan menggunakan fitur-fitur yang telah tersedia dalam sistem. Setelah pengujian, pengguna diminta untuk mengisi kuesioner (Tabel 3) yang bertujuan untuk menilai persepsi pengguna terhadap sistem yang diuji. Tujuan dari kuesioner ini adalah untuk mengevaluasi penerimaan pengguna terhadap sistem baru yang telah dikembangkan. Item pada kuesioner ini

Penilaian terhadap responden dengan menggunakan skala *likert,* dilakukan dengan memberikan bobot untuk masing-masing penilaian. Adapun pembobotan untuk masing-masing penilaian seperti ditunjukkan pada Tabel 4 sebagai berikut.

TABEL 4. PEMBOBOTAN PENILAIAN

Penilaian	Bobot
Sangat Tidak Setuju	1
Tidak Setuju	2
Netral	3
Setuju	4
Sangat Setuju	5

5) Pemeliharaan Sistem

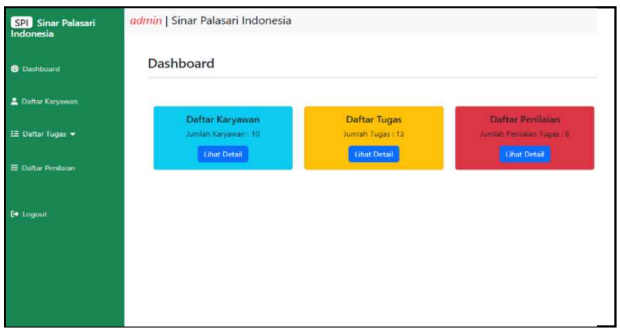
Tahap Operasi dan Pemeliharaan merupakan langkah terakhir dalam pengembangan sistem. Proses ini melibatkan aktivitas menjalankan sistem yang telah diuji dan disetujui oleh pengguna. Pemeliharaan dilakukan jika terdapat kesalahan yang sebelumnya tidak terdeteksi selama uji unit dan uji sistem.

III. HASIL DAN PEMBAHASAN

A. Implementasi Sistem

1) Halaman Dashboard Admin

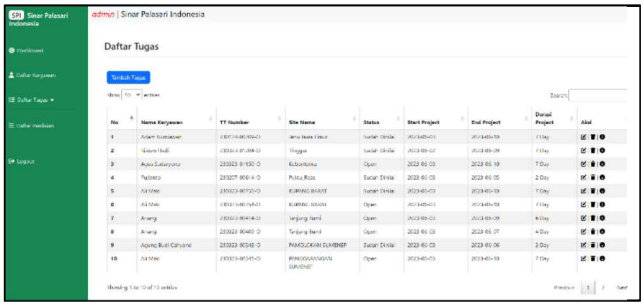
Halaman *dashboard* admin merupakan tampilan utama Web pada sistem manajemen penilaian kinerja karyawan berbasis penugasan karyawan. Halaman ini berisi informasi mengenai daftar karyawan, daftar tugas, dan daftar penilaian pada sistem. Gambar 4 berikut merupakan gambar tampilan halaman *dashboard* admin.



Gambar 4. Halaman dashboard admin

2) Halaman Daftar Tugas

Halaman daftar tugas merupakan tampilan dari daftar-daftar tugas yang ada. Pada halaman ini juga, admin dapat menambah, menghapus, mengubah, dan melihat informasi tugas. Gambar 5 berikut ini merupakan gambar tampilan halaman daftar tugas.



Gambar 5. Halaman daftar tugas karyawan

3) Halaman Detail Nilai Karyawan

Halaman detail nilai karyawan merupakan tampilan detail dari nilai yang sudah diberikan oleh sistem. Pada halaman ini, admin dapat melihat detail nilai dari masing-masing karyawan. Gambar tampilan halaman detail nilai karyawan dapat dilihat pada Gambar 6.

Gambar 6. Halaman detail nilai karyawan

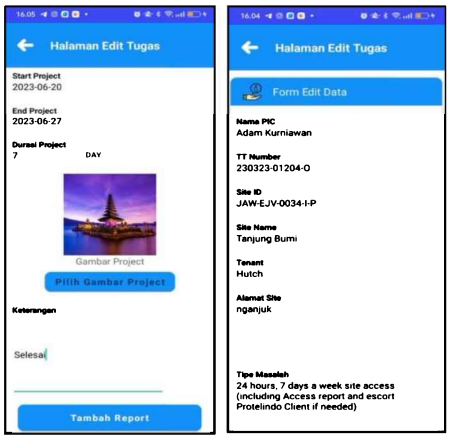
4) Halaman Utama pada Android

Halaman utama karyawan merupakan halaman yang akan menampilkan beberapa menu ketika karyawan berhasil melakukan *login* pada sistem penilaian kinerja karyawan berbasis Android. Gambar 7 menunjukkan halaman utama karyawan.

Gambar 7. Halaman utama karyawan

5) Halaman Pelaporan Tugas pada Android

Halaman pelaporan tugas pada android merupakan halaman yang akan menampilkan detail data tugas yang diberikan oleh admin ke karyawan. Karyawan dapat melaporkan tugas yang telah dikerjakan olehnya melalui halaman ini. Gambar 8 berikut merupakan gambar halaman tugas pada Android.



Gambar 8. Halaman pelaporan tugas karyawan

6) Halaman Nilai Tugas Karyawan pada Android

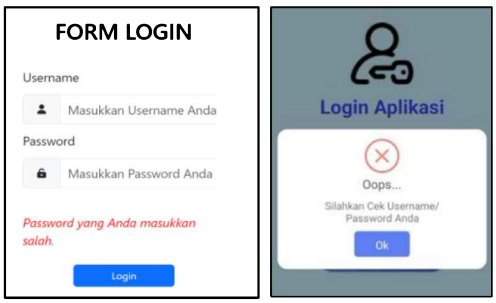
Halaman nilai tugas karyawan merupakan halaman penilaian yang dilakukan sistem atas tugas yang telah diselesaikan oleh karyawan. Halaman tersebut seperti terlihat pada Gambar 9 berikut ini.

Gambar 9. Halaman nilai tugas pada sistem berbasis android

B. *Pengujian Sistem*

1) *Penanganan Kesalahan*

Memasukkan alamat *email* dan *password* yang tidak sesuai dengan data yang tersimpan dalam *database*, dapat menyebabkan terjadinya kesalahan. Sistem akan memberikan pemberitahuan bahwa kombinasi *email* dan *password* yang dimasukkan tidak ditemukan di *database*. Peringatan yang muncul akibat ketidaksesuaian halaman *login* ini memunculkan notifikasi seperti tampak pada Gambar 10.



Gambar 10. Tampilan Penanganan Kesalahan *Login* Pengguna

2) *Perhitungan Penilaian Tugas*

Berdasarkan perhitungan penilaian tugas yang ada di dalam sistem tersebut, maka dapat dicontohkan perhitungan dibawah ini. Berdasarkan Gambar 11, skenario penilaian tugas selesai dapat dilihat bahwa lama pekerjaan adalah 3 hari. Oleh karena itu, perhitungan yang digunakan adalah LP < 7 dengan nilai total sebagai berikut:

$$NTotal = NTgs + (NB\ x\ LP)$$
$$= 100 + (10\ x\ 3)$$
$$= 100 + 30$$
$$= 130$$

Gambar 11. Tampilan Penilaian Tugas Selesai

TABEL 5. HASIL PENGUJIAN UNIT (BLACK BOX TESTING)

Unit	Skenario Pengujian	*Test Case*	Hasil Diharapkan	Hasil Uji
Login	Mengisi *username* dan *password* yang tidak terdaftar di *database*, klik tombol *login*	Username: admin123 Password: 123	Sistem menolak dan menampilkan pesan "Silahkan cek *username* atau *password* anda."	Sistem menolak dan menampilkan pesan "Silahkan cek *username* atau *password* anda."
	Mengisi *username* dan *password* sesuai dengan yang ada di *database*, kemudian klik tombol *login*	Username: edward285 Password: 12345	Sistem menerima dan menampilkan pesan "Berhasil Login."	Sistem menerima dan menampilkan pesan "Berhasil Login."
Kelola Data Karyawan	Mengisi Form tambah karyawan dengan *username* yang ada di dalam *database*, kemudian klik tombol tambah karyawan	Nama: Adam Kurniawan Username: adam123 Password: 12345 Alamat: Surabaya No Hp: 0812101123	Sistem menolak dan menampilkan pesan "Username adam123 Sudah ada!!! Silahkan menggunakan username yang lain."	Sistem menolak dan menampilkan pesan "Username adam123 Sudah ada!!! Silahkan menggunakan username yang lain."
	Mengisi Form tambah karyawan yang tidak ada di database, kemudian klik tombol tambah	Nama: Edward Paundra Username: edward285 Password: 12345 Alamat: Surabaya No Hp: 0827146482	Sistem menerima dan menampilkan pesan "Data User Berhasil Ditambahkan"	Sistem menerima dan menampilkan pesan "Data User Berhasil Ditambahkan"
Kelola Data Tugas	Mengisi Form tambah tugas, kemudian klik tombol tambah tugas	Nama: Edward TT number: 123-1-2-3 Site Id: JAW-EJV-0034-I-P Site Name: Tinggar Tenant: Indosat Status: open Alamat: Bandung Start : 03/07/2023 End: 10/07/2023	Sistem menerima dan menampilkan pesan "Data Tugas Berhasil Ditambahkan"	Sistem menerima dan menampilkan pesan "Data Tugas Berhasil Ditambahkan"
	Tidak melakukan ubah tugas, kemudian klik tombol ubah	Gambar: keterangan:	Sistem akan menolak dan menampilkan pesan "gagal"	Sistem menolak dan menampilkan pesan "gagal"
	Melakukan ubah Gambar dan keterangan, kemudian klik tombol Ubah	Gambar: IMG202307 04015140.jpg keterangan: Selesai	Sistem menerima dan menampilkan pesan "Berhasil"	Sistem menerima dan menampilkan pesan "Berhasil"
Laporan Bulanan	Mengisi tanggal awal dan akhir kemudian klik tombol Cari Semua	klik tombol Cari Semua	Sistem menampilkan semua data laporan pendapatan	Sistem menampilkan semua data laporan pendapatan
	Klik tombol Cetak Laporan	Klik tombol Cetak Laporan	Sistem akan menampilkan halaman baru untuk cetak laporan	Sistem akan menampilkan halaman baru untuk cetak laporan
Penilaian Karyawan	Memilih tabel mana yang akan dinilai, kemudian klik tombol *cheklist*	Klik tombol *cheklist*	Sistem menerima dan menampilkan pesan "Anda Berhasil Memberi Penilaian"	Sistem menerima dan menampilkan pesan "Anda Berhasil Memberi Penilaian"
	Memilih tabel mana yang akan dinilai, kemudian klik tombol silang	Klik tombol silang	Sistem menerima dan menampilkan pesan "Anda berhasil mendecline tugas."	Sistem menerima dan menampilkan pesan "Anda berhasil mendecline tugas."

3) Pengujian Unit (Black Box Testing)

Uji *black box* dilakukan guna mengamati respons sistem terhadap berbagai unit yang diuji. Pengujian dilakukan oleh pengembang sesuai dengan unit yang sedang diuji. Tabel 5 menampilkan hasil dari uji *black box* terhadap sistem penilaian kinerja berdasarkan penugasan karyawan di PT Sinar Palasari Indonesia. Berdasarkan hasil pengujian *black box*, secara keseluruhan sistem telah memenuhi standar pengujian. Hal ini terbukti dengan berhasilnya semua pengujian yang dilakukan pada masing-masing unit pengujian.

4) Pengujian Sistem (User Acceptance)

Pengujian yang dilakukan terhadap pengguna melibatkan 8 responden, yang terdiri atas 3 admin dan 5 karyawan. Hasil pengujian yang dilakukan kepada karyawan, setelah dikonversi dengan pembobotan ditunjukkan pada Tabel 6 berikut.

TABEL 6. HASIL PENGUJIAN USER ACCEPTANCE TERHADAP KARYAWAN

Responden	Pertanyaan						
	1	2	3	4	5	6	7
Karyawan 1	5	4	3	3	5	5	4
Karyawan 2	4	4	3	4	4	3	4
Karyawan 3	5	4	4	4	4	4	4
Karyawan 4	5	4	3	4	5	4	5
Karyawan 5	4	3	4	3	4	3	3
Jumlah	23	19	17	18	22	19	20
Skor (%)	**92**	**76**	**68**	**72**	**88**	**76**	**80**

Berdasarkan Tabel 6, karyawan sangat setuju bahwa sistem penilaian kinerja karyawan dapat diakses dan digunakan dengan mudah (92%). Penilaian terhadap kemampuan sistem untuk membantu dalam meningkatkan pemahaman tentang kinerja karyawan menunjukkan nilai persentase terendah sebesar 68%, namun masih dapat diterima oleh karyawan. Hasil kuesioner menunjukkan bahwa secara keseluruhan pelanggan setuju dengan sistem yang dikembangkan dengan persentase sebesar 78,86%.

Pengujian penerimaan pengguna (*user acceptance*) juga dilakukan terhadap admin. Pengujian yang dilakukan terhadap admin melibatkan 3 responden. Hasil pengujian yang dilakukan ditunjukkan pada Tabel 7.

TABEL 7. HASIL PENGUJIAN USER ACCEPTANCE TERHADAP ADMIN

Responden	Pertanyaan						
	1	2	3	4	5	6	7
Admin 1	4	4	3	5	3	4	4
Admin 2	5	4	4	5	4	5	5
Admin 3	5	4	4	4	4	4	4
Jumlah	**14**	**12**	**11**	**14**	**11**	**13**	**13**
Skor (%)	**93,33**	**80**	**73,33**	**93,33**	**73,33**	**86,67**	**86,67**

Tiga admin terlibat dalam pengujian menggunakan kuesioner, sehingga nilai skala tertinggi adalah 15. Berdasarkan Tabel 7, admin memberikan penilaian yang tinggi terhadap kemudahan navigasi antarmuka sistem penilaian kinerja karyawan dan responsivitas sistem terhadap permintaan. Namun, penilaian terhadap kemampuan evaluasi sistem dan kemampuan perbaikan proses penilaian kinerja karyawan mendapat nilai persentase sebesar 73,33%. Dari hasil kuesioner tersebut, dapat disimpulkan bahwa secara keseluruhan, admin sangat setuju dengan sistem yang dikembangkan dengan persentase sebesar 83,81%.

IV. KESIMPULAN

Berdasarkan hasil penelitian yang dilakukan, aplikasi sistem penilaian kinerja berbasis penugasan karyawan yang dibangun dapat membantu admin dalam melakukan proses pemberian tugas dan penilaian tugas karyawan. Aplikasi ini dilengkapi dengan fitur yang dapat memonitor pekerjaan karyawan sekaligus memberikan penilaian tugas karyawan. Selain itu, karyawan dapat dengan cepat melakukan proses pelaporan pekerjaan yang telah diberikan oleh admin. Hasil pengujian yang telah dilakukan didapatkan persentase penerimaan pengguna sebesar 78,86% (bagi karyawan) dan 83,81% (bagi admin), sehingga layak untuk digunakan. Aplikasi ini diharapkan dapat ditingkatkan dengan menambahkan fitur rekomendasi pemilihan karyawan terbaik.

REFERENSI

[1] H. Karimikia, N. Safari, and H. Singh, "Being useful: How information systems professionals influence the use of information systems in enterprises," *Information Systems Frontiers*, vol. 22, no. 2, pp. 429–453, Apr. 2020, doi: 10.1007/s10796-018-9870-7.

[2] N. Renaldo, Suhardjo, Suharti, Suyono, and Cecilia, "Benefits and Challenges of Technology and Information Systems on Performance," *Journal of Applied Business and Technology*, vol. 3, no. 3, pp. 302–305, Sep. 2022, doi: 10.35145/jabt.v3i3.114.

[3] S. H. Awan, N. Habib, C. Shoaib Akhtar, and S. Naveed, "Effectiveness of Performance Management System for Employee Performance Through Engagement," *Sage Open*, vol. 10, no. 4, p. 215824402096938, Oct. 2020, doi: 10.1177/2158244020969383.

[4] T. Hikmawan and Budi Santoso, "Human Resources Information System to Improve Employee Performance," *Dinasti International Journal of Management Science*, vol. 1, no. 4, pp. 578–584, Mar. 2020, doi: 10.31933/dijms.v1i4.194.

[5] S. Rakshit, N. Islam, S. Mondal, and T. Paul, "Mobile apps for SME business sustainability during COVID-19 and onwards," *J Bus Res*, vol. 135, pp. 28–39, Oct. 2021, doi: 10.1016/j.jbusres.2021.06.005.

[6] N. Angelova, "Mobile Applications for Business," *Trakia Journal of Sciences*, vol. 17, no. Suppl.1, pp. 853–859, 2019, doi: 10.15547/tjs.2019.s.01.140.

[7] T. Tjahjanto, A. Arista, and E. Ermatita, "Information System for State-owned inventories Management at the Faculty of Computer Science," *Sinkron*, vol. 7, no. 4, pp. 2182–2192, Oct. 2022, doi: 10.33395/sinkron.v7i4.11678.

Perancangan Kebutuhan Perangkat Lunak Sistem Informasi Perpustakaan Perguruan Tinggi

Fikko Rafirs Yanuar
Departemen Sistem Informasi
Universitas Pembangunan Nasional
"Veteran" Yogyakarta
Yogyakarta, Indonesia
fikkociparicity06@gmail.com

Riza Prapascatama Agusdin
Departemen Sistem Informasi
Universitas Pembangunan Nasional
"Veteran" Yogyakarta
Yogyakarta, Indonesia
rizapra@upnyk.ac.id

Anggita Erlina Aprilia
Departemen Sistem Informasi
Universitas Pembangunan Nasional
"Veteran" Yogyakarta
Yogyakarta, Indonesia
anggita.erlina93@gmail.com

May Vlawinzky Pelawi
Departemen Sistem Informasi
Universitas Pembangunan Nasional
"Veteran" Yogyakarta
Yogyakarta, Indonesia
mayvlawinzky7@gmail.com

Abstrak— Perpustakaan merupakan salah satu komponen penting dalam lingkungan perguruan tinggi. Dengan adanya perpustakaan membantu mahasiswa dalam melakukan pembelajaran karena perpustakaan menyediakan banyak informasi seperti buku, skripsi, jurnal, dan masih banyak lagi. Maka dibutuhkan sebuah sistem informasi yang bisa membuat kegiatan lebih efektif dan efisien. Hal ini juga berkaitan dengan kemajuan teknologi yang dapat memberikan dampak transformasi terhadap perpustakaan konvensional menjadi perpustakaan digital. Penelitian ini menjelaskan rancangan sistem informasi perpustakaan perguruan tinggi menggunakan metode UML dengan acuan analisis PIECES. Hasil pnelitian ini berupa penembnangan fitur denda pada transaksi pengembalian jika telah mengembalikan pinjaman.

Kata kunci— sistem informasi perpustakaan, Unified Modeling Language, denda

I. PENDAHULUAN

Perkembangan teknologi memberikan kemudahan dalam memperoleh, mengelola, dan menyimpan informasi dengan mudah dan efisien. Saat ini, teknologi telah memberikan dampak yang signifikan terhadap berbagai aspek kehidupan, termasuk dunia pendidikan [1]. Perguruan tinggi adalah lembaga pendidikan formal yang didirikan oleh suatu negara maupun pihak swasta dengan tujuan untuk menyelenggarakan pendidikan yang ditujukan kepada para mahasiswa. Perguruan tinggi juga bertujuan untuk membantu mahasiswa dalam mengembangkan berbagai potensi yang dimiliki dalam diri. Sehingga diperlukan suatu penunjang yang bisa dijadikan sebagai sarana pembelajaran [2].

Salah satu komponen utama bagi perguruan tinggi dalam meningkatkan kualitas pendidikan adalah menyediakan akses terhadap sumber-sumber ilmiah dan literatur yang relevan bagi mahasiswa yaitu melalui perpustakaan. Perpustakaan merupakan salah satu sarana penting dalam pengembangan ilmu pengetahuan dan pendidikan sehingga suatu perguruan tinggi perlu memilikinya. Perpustakaan menyediakan berbagai macam sumber informasi seperti buku, majalah, jurnal dan sebagainya yang dapat dimanfaatkan oleh penggunanya sehingga diperlukan pengelolaan perpustakaan dengan memanfaatkan suatu sistem informasi. Sistem informasi perpustakaan memiliki peran yang sangat penting dalam mendukung pengelolaan koleksi perpustakaan, mengakses informasi, juga pelayanan yang efektif bagi penggunanya [3].

Kondisi yang dihadapi beberapa perpustakaan di perguruan tinggi saat ini yaitu dalam melakukan pendataan, karena masih menggunakan sistem konvensional (manual). Sehingga hal tersebut dapat menyebabkan ketidakefektifan dalam pengelolaan perpustakaan. Dari kendala tersebut, perpustakaan memerlukan sistem yang dapat mengelola informasi sehingga lebih tertata dan terorganisir dengan baik. Sistem tersebut berupa aplikasi berbasis *website* yang dapat mengelola, mendata buku keluar dan buku masuk juga memberikan informasi terkait posisi buku agar memudahkan pengunjung menemukan buku yang dicari [4].

Sebelumnya terdapat beberapa jurnal yang melakukan pembahasan terkait pengembangan sistem informasi perpustakaan. Pada hasil penelitian terdahulu yang dilakukan oleh Akik Hidayat dan Amalyah Nurhasanah dengan menghasilkan suatu sistem yang bisa membantu admin dalam melakukan pengelolaan data buku dan anggota serta dapat melakukan proses transaksi seperti peminjaman dan pengembalian buku yang lebih cepat dan efisien [5]. Terdapat juga penelitian yang dilakukan oleh Saripuddin Muddu dan kawan-kawan menghasilkan suatu sistem yang memudahkan petugas dan anggota dalam mencari data juga pengelolaan data-data di perpustakaan [6]. Penelitian lainnya yaitu oleh Hermawan dan kawan-kawan merancang sistem yang memberi kemudahan kepada semua *user* baik anggota maupun admin dalam melakukan pencarian data buku dan proses peminjaman serta pengembalian buku [7].

Melalui penelitian ini dapat memberikan solusi yang sesuai dengan kebutuhan perpustakaan saat ini dengan menambah fitur lain yang belum dirancang dari beberapa penelitian

sebelumnya. Solusi yang diusulkan dalam merancang perangkat lunak sistem informasi perpustakaan pada penelitian ini yaitu, sistem dapat menampilkan informasi terkait denda yang akan terima oleh mahasiswa jika terlambat mengembalikan buku sesuai dengan jadwal yang telah ditentukan. Dengan demikian diharapkan perangkat lunak yang akan dirancang dapat meningkatkan efisiensi, aksesibilitas dan manajemen informasi di perpustakaan perguruan tinggi [8].

II. METODOLOGI

Penelitian ini berfokus dalam membuat perancangan sistem informasi yang akan di implementasikan di lingkungan perguruan tinggi. Untuk alur penelitian ini hingga tahap implementasi dapat dilihat pada Gambar 1. Untuk lebih lengkapnya dapat dilihat pada penjelasan dibawah ini:

Gambar 1. Kerangka Penelitian

1. Pada tahap awal metodologi yaitu studi literatur yang dilakukan dengan cara pengumpulan data dari berbagai kajian pustaka yang berhubungan dengan penelitian, dari pengumpulan tersebut dilakukan pencatatan informasi yang bermanfaat pada penyusunan teori dan pembahasan.
2. Tahap selanjutnya yaitu analisis masalah yang diperlukan untuk mengetahui masalah yang terjadi pada perpustakaan dari berbagai perguruan tinggi. Analisis tersebut akan diperlukan sebagai acuan dalam perancangan sistem informasi perpustakaan terutama pada alur sistem. Pada tahapan ini dilakukan dengan menganalisa masalah menggunakan metode PIECES.
3. Pada tahap perancangan yaitu diaman pada tahap ini dilakukan perancangan sistem dengan menyesuaikan hasil analisis yang telah dibuat dengan membandingkan hasil penelitian yang terdahulu. Pada perancangan sistem informasi perpustakaan ini meliputi perancangan yang menggunakan diagram UML (*Unified Modeling Language*).
4. Selanjutnya pada tahapan ini sistem yang telah dirancang dan sesuai dengan hasil analisis akan diimplementasikan ke dalam bentuk program, sehingga dapat mempermudah pengguna dalam memahami sistem baru. Tahap implementasi sistem informasi perpustakaan ini merupakan tahap meletakkan sistem agar dapat dioperasikan [9].

Perancangan sistem informasi perpustakaan ini menggunakan diagram-diagram UML (*Unified Modeling Language*) untuk menggambarkan sistem yang akan dirancang. Terdapat empat diagram UML yang dipakai pada penelitian ini untuk mewakili gambaran sistem yaitu *use case diagram, activity diagram, sequence diagram, dan class diagram* [10]. Model UML memiliki tujuan untuk menggabungkan berbagai teknik pemodelan berorientasi objek menjadi lebih terstandarisasi. Sebelum membuat perancangan sistem informasi perpustakaan ini dilakukan analisa permasalahan. Analisa Permasalahan ini dilakukan untuk mengidentifikasi masalah apa saja yang ada pada sistem informasi perpustakaan yang telah ada saat ini. Analisis Permasalahan dilakukan dengan menggunakan metode PIECES.

Metode PIECES *framework* merupakan salah satu kerangka yang digunakan untuk melakukan analisis pada sistem yang bersifat konvensional (manual) atau terkomputerisasi sebelumnya menjadi lebih baik. Metode ini menggunakan proses klasifikasi suatu permasalahan dan menggunakan peluang untuk membuat perancangan sistem baru yang lebih baik. Metode ini dapat menghasilkan suatu usulan baru pada sistem lama sehingga bisa digunakan sebagai bahan pertimbangan pada proses pengembangan sistem yang baru. Pengembangan sistem menggunakan metode PIECES framework ini memiliki jenis analisis yang terdiri dari *Performance, Information, Economic, Control, Efficiency, Service* [11].

III. HASIL DAN PEMBAHASAN

A. *Hasil*

Pada bagian akan menjelaskan tentang hasil yang telah dilakukan berdasarkan dari metodologi. Hasil ini mencangkup analisis dan implementasi dari rancangan yang telah dibuat. Berikut merupakan hasil dari rancangan yang telah dibuat.

1. Analisis PIECES

   Analisis PECES merupakan acuan yang digunakan penelitian ini untuk membuat UML. Analisis ini adalah suatu kerangka kerja untuk menganalisis sebuah sistem yang bersifat konvensional (manual) ataupun terkomputerisasi [12]. Pada Tabel 1 menjelaskan analisis PIECES yang menjadi acuan pembuatan paper ini dengan UML.

TABEL 1. ANALISIS PIECES

Jenis Analisis	Kelemahan Sistem Lama	Sistem yang Diusulkan
Performance	Pendataan yang dilakukan oleh perpustakaan masih menggunakan sistem konvensional (manual) yang rentan terjadinya *human error*.	Sistem berbasis komputer ini memudahkan dalam pengelolaan data yang ada pada perpustakaan.
Information	Informasi denda hanya diberitahukan saat buku dikembalikan	Pada sistem mahasiswa akan mendapatkan informasi jika buku sudah melalui tenggat waktu. Sistem

		akan menginformasikan denda yang perlu dibayarkan saat buku dikembalikan.
Economic	Biaya yang dikeluarkan cukup besar karena untuk keperluan pembukuan, alat tulis, dan mencetak beberapa dokumen seperti pengumuman. Selain itu juga, biaya gaji untuk para petugas perpustakaan.	Dalam jangka pendek, akan memerlukan biaya yang cukup besar pada awal pemasangan atau perilisan sistem. Tetapi dalam jangka panjang akan lebih menghemat karena hanya membayar untuk biaya perawatan sistem.
Control	Sistem manual akan sulit dikontrol terlebih jika terjadinya *human error*. Hal ini bisa terjadi jika lalai mencatat data, kesalahan dalam penulisan sehingga terjadi ketidaksinkronan data.	Adanya sistem akan memudahkan pengontrolan data yang ada karena semua transaksi data yang masuk akan otomatis tersimpan pada sebuah *database*.
Efficiency	Membutuhkan waktu yang banyak untuk bisa merekap dan memasukan data. Terlebih jika pada saat itu kondisi perpustakaan ramai pengunjung atau lebih dari jumlah biasanya. Maka pendataan akan lebih lama.	Sistem mempermudah dan mempercepat pendataan yang ada. Selain itu juga data langsung masuk pada sistem dan bisa diakses oleh para pengunjung perpustakaan.
Services	Pelayanan terhadap pengunjung akan banyak memakan waktu karena selain harus mendata buku keluar dan buku masuk juga harus mengkondisikan lingkungan agar tetap kondusif.	Pelayanan akan lebih baik karena para pengunjung bisa melakukan *self service* sehingga petugas tinggal bertugas mengkondisikan lingkungan perpustakaan.

2. Kebutuhan pengguna
Analisis kebutuhan pengguna merupakan proses yang penting dalam perancangan perangkat lunak dapat dilihat pada Tabel 2. Analisis kebutuhan perangkat lunak bertujuan untuk memahami dan mendokumentasikan kebutuhan pengguna yang harus dipenuhi oleh sistem yang akan dirancang [13].

TABEL 2. ANALISIS KEBUTUHAN PENGGUNA

Jenis Pengguna	Tanggung Jawab	Hak Akses
Mahasiswa	Bertanggung jawab dalam peminjaman, pengembalian, dan denda	• Mengedit akun • *Login* • *Logout* • *Delete* • Melihat buku • Meminjam buku • Mengembalikan buku • Membayar denda
Dosen	-	• Mengedit akun • Login • Logout • Delete • Melihat buku
Kepala Perpustakaan	Bertanggung jawab dalam memantau kinerja perpustakaan	• Login • Logout • Melihat buku • Melihat peminjam • Melihat pengguna aplikasi
Admin Perpustakaan	Bertanggung jawab dalam pendataan dan pengolahan aplikasi buku	• Login • Logout • Menginput buku • Mengedit buku • Menghapus buku • Mengakses database

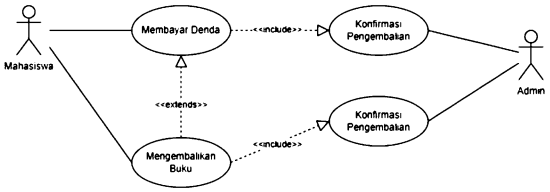
3. Kebutuhan sistem
- Analisis kebutuhan sistem berfokus pada pemahaman menyeluruh tentang apa yang diharapkan dari sistem yang dirancang, dapat dilihat pada Tabel 3. Tujuan utama dari analisis kebutuhan sistem adalah untuk mengidentifikasi, memahami dan mendefinisikan kebutuhan yang harus dipenuhi oleh sistem yang dikembangakan [14].

TABEL 3. KEBUTUHAN SISTEM

Operasional	• Digunakan pada sistem operasi Windows dengan spesifikasi minimal windows 7. • Spesifikasi komputer minimal tipe processor core 2, frekuensi >1,1 GHz

	• Dapat digunakan di *smartphone* baik melalui *browser* maupun aplikasi • Device yang digunakan harus bisa mengakses internet
• Kinerja	• Sistem dapat diakses oleh 50 orang secara bersamaan. • Daya simpan sistem lebih dari 1 TB. • Sistem hanya diakses oleh dosen dan mahasiswa perguruan tinggi. • Sistem memiliki tampilan yang mudah dipahami
• Keamanan	• Akses ke sistem menggunakan *username* dan *password* pengguna • Akses ke *database* menggunakan *password* admin

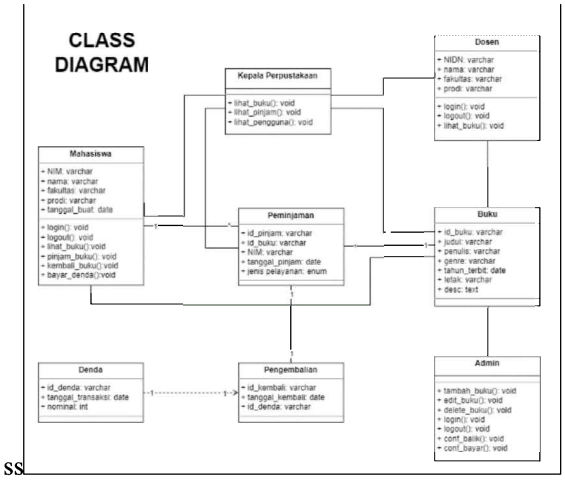
4. *Use Case Diagram*



Gambar 2. *Use Case Diagram*

Use Case Diagram merupakan deskripsi interaksi antara satu aktor atau lebih dengan suatu sistem yang dibangun [15]**.** Pada Gambar 2 *use case diagram* diatas menjelaskan interaksi aktor mahasiswa dengan aktor admin perpustakaan pada kegiatan pengembalian buku dan pembayaran denda. Saat mahasiswa mengembalikan buku sistem akan menginformasikan kepada admin terkait denda. Jika mahasiswa tidak memiliki denda maka buku langsung dikonfirmasi oleh admin. Tetapi, jika mahasiswa memiliki denda maka mahasiswa perlu membayar terlebih dahulu untuk bisa mengembalikan buku dan kemudian admin perpustakaan akan mengonfirmasi pengembalian buku.

5. *Class Diagram*



SS

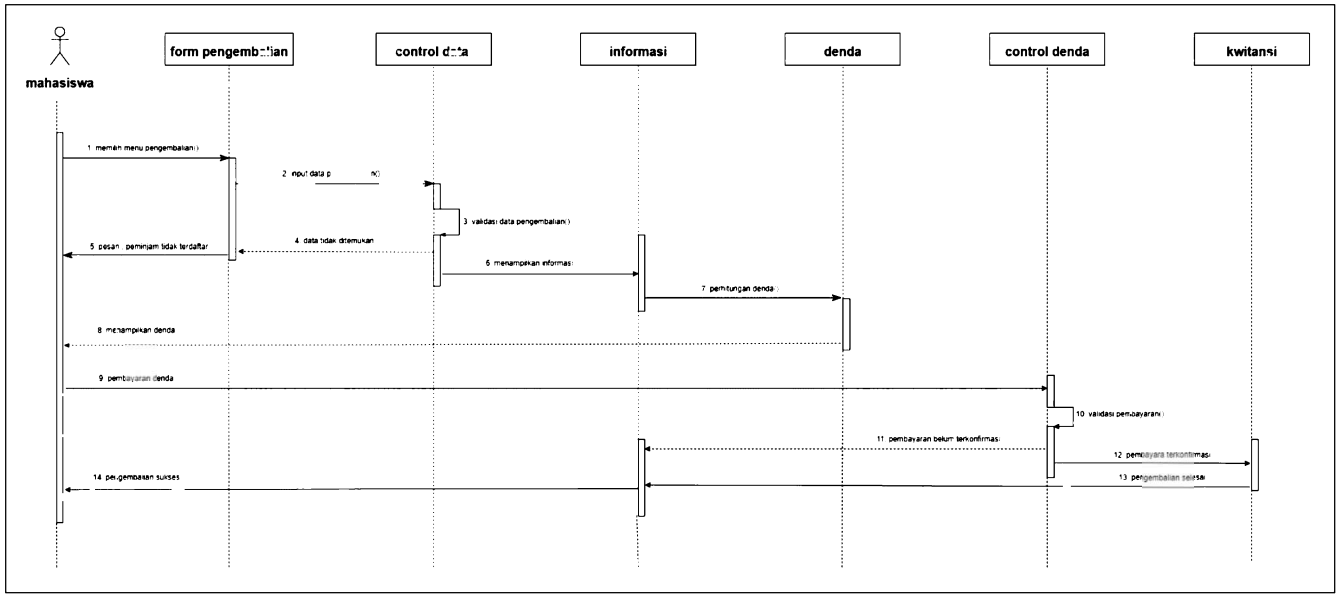
Gambar 3. *Class Diagram*

Class diagram merupakan salah satu bentuk diagram struktur yang menggambarkan struktur secara rinci yang terdapat deskripsi class, atribut, dan hubungan dari setiap objek dengan jelas pada UML [17].
Keterangan :
Terdapat 8 class yaitu dosen, buku, admin, kepala perpustakaan, peminjaman, pengembalian, denda, dan mahasiswa.

a) Pada class dosen terdapat NIDN sebagai primary key, nama, dan tanggal lahir. Dan terdapat fungsi edit akun, login, logout, delete dan lihat buku.
b) Class kepala perpustakaan hanya terdapat fungsi lihat_buku, lihat_pinjam, dan lihat_pengguna.
c) Class Mahasiswa terdapat NIM sebagai primary key, nama, fakultas, prodi, dan tanggal_lahir. Fungsi yang digunakan edit_akun, login, logout, delete, lihat_buku, pinjam_buku, kembali_buku, bayar_denda.
d) Class Peminjaman terdampar id_pinjam sebagai primary key, id_buku dan NIM sebagai foreign key, tanggal_pinjam, dan lama_peminjaman.
e) Class Buku terdapat id_buku sebagai primary key, judul, penulis, genre, tahun_terbit, letak, dan stok.
f) Class admin terdapat fungsi tambah_buku, edit_buku, delete_buku.
g) Class pengembalian terdapat id_kembali sebagai primary key, tanggal_kembali, dan id_denda sebagai foreign key.
h) Class denda terdapat id_denda primary key, tanggal_transaksi, dan nominal menggunakan relasi dependency karena denda tergantung pada pengembalian

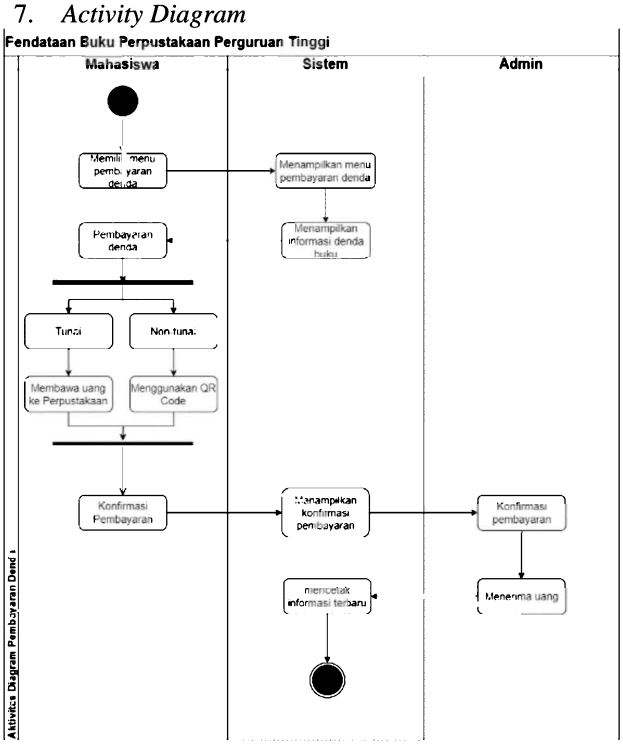
6. *Sequence Diagram*

Gambar 4 *Sequence diagram* merupakan gambaran dari interaksi antara objek di dalam dan di sekitar sistem begitu juga sistem berupa *message* yang menggambarkan interaksi terhadap waktu dengan *user* [16]. Se*quence diagram* mencangkup dimensi horizontal (objek-objek yang terkait) dan dimensi vertikal (waktu). Berikut adalah *sequence diagram* yang dibangun. Pada gambar sequence pengembalian buku berikut menggambarkan bagaimana alur ketika ingin melakukan pengembalian buku hingga dapat melakukan pembayaran denda, ketika *user* (mahasiswa) melakukan proses pembayaran untuk denda yang didapatkan hingga sistem merekam konfirmasi pengembalian telah selesai.



Gambar 4. *Sequence Diagram*

7. *Activity Diagram*

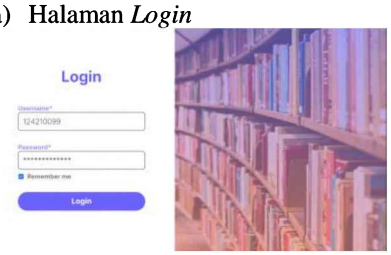


Gambar 5. *Activity Diagram*

Activity diagram adalah sebuah diagram yang menggambarkan suatu konsep alur pada data/kontrol, aksi terstruktur yang dirancang dengan baik pada sistem. [16] Pada Gambar 5. *activity diagram* pembayaran denda aktor (mahasiswa). Merupakan proses pembayaran denda oleh pengguna. Denda ini dibayarkan jika pengguna melewati masa peminjaman. Sistem akan menampilkan besar denda pada tampilan setelah pengembalian, lalu mahasiswa dapat membayarkan denda. Sistem akan mengkonfirmasi pembayaran dan melanjutkan ke proses berikutnya. Apabila belum dibayarkan sistem akan mengirimkan pesan bahwa pembayaran belum dilakukan.

8. *User Interface*

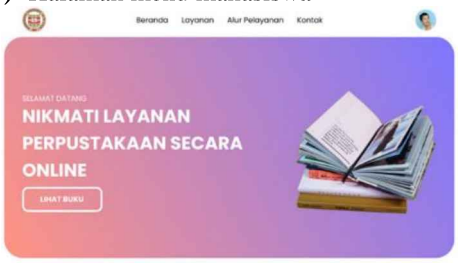
a) Halaman *Login*



Gambar 6. Halaman *login* mahasiswa

Gambar 6 merupakkan tampilan antarmuka halaman login. Halaman login merupakan halaman utama pada *website* sistem informasi perpustakaan. Pada halaman ini *user* perlu memasukkan *username* dan *password* terlebih dahulu sebelum masuk ke dalam sistem *website* perpustakaan.

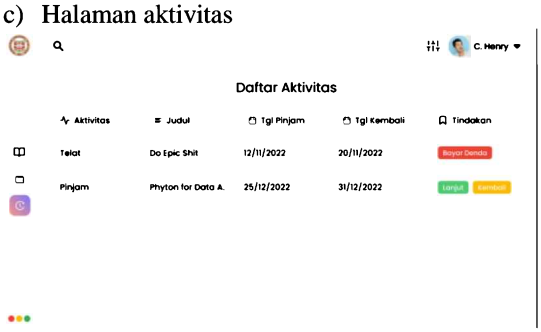
b) Halaman menu mahasiswa



Gambar 7. Halaman menu mahasiswa

Gambar 7 merupakan tampilan antarmuka halaman menu mahasiswa. Pada halaman ini yang berisi informasi untuk melihat buku, beranda, layanan, alur pelayanan, kontak, dan juga terdapat pengaturan profil.

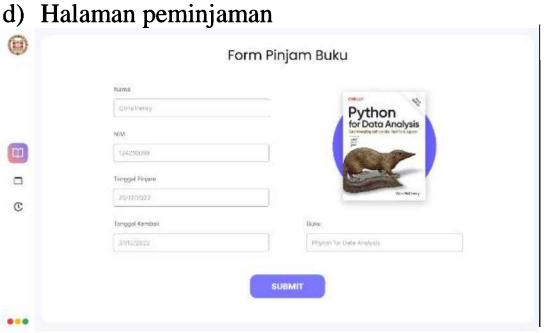
c) Halaman aktivitas



Gambar 8. Halaman aktivitas

Gambar 8 merupakan tampilan antarmuka halaman aktivitas. Pada halaman ini berisi daftar aktivitas *user* pada peminjaman buku perpustakaan, judul buku yang dipinjam, tanggal peminjaman, tanggal pengembalian, serta tindakan untuk pembayaran denda jika *user* terlambat mengembalikan buku perpustakaan.

d) Halaman peminjaman



Gambar 9. Halaman peminjaman

Gambar 9 merupakan tampilan antarmuka halaman peminjaman. Pada halaman ini berisi *form* untuk mengajukan peminjaman buku. Formulir yang akan di *submit* berisi nama, NIM, tanggal pinjam, tanggal kembali, dan judul buku yang akan dipinjam.

e) Halaman pengembalian

Gambar 10. Halaman pengembalian

Gambar 10 merupakan tampilan antarmuka halaman pengembalian buku. Pada halaman ini terdapat *action* untuk melakukan konfirmasi buku telah dikembalikan. Terdapat notifikasi informasi pengembalian buku yang telah berhasil ataupun gagal dilakukan.

f) Halaman denda

Gambar 11. Halaman denda

Gambar 11 merupakan tampilan antarmuka halaman denda. Pada halaman ini berisi informasi pembayaran denda yang terdapat keterangan lama keterlambatan dalam pengembalian serta terdapat total biaya yang harus dibayarkan oleh *user*.

B. *Pembahasan*

Pada penelitian ini menghasilkan sebuah inovasi baru yang terdapat pada sistem perpustakaan ditingkat perguruan tinggi. Sistem ini menerapkan denda yang bisa diakses pada menu mahasiswa. Denda ini akan muncul jika mahasiswa terlambat mengembalikan buku dari tenggat waktu yang telah di tentukan. Selain itu, sistem ini juga memberikan opsi pilihan kepada mahasiswa saat membayarnya. Mahasiswa diberikan dua pilihan metode pembayaran yaitu tunai dan non-tunai. Jika mahasiswa melakukan pembayaran tunai maka mahasiswa bisa langsung melakukan transaksi bersama petugas. Sedangkan, jika non-tunai maka mahasiswa akan diberikan QR kode pembayaran sesuai dengan denda yang diterima. Hal ini menjadi pembeda atau inovasi yang baru dari penelitian penelitian sebelumnya.

IV. KESIMPULAN

Berdasarkan penelitian yang telah dirancang maka dihasilkan suatu Sistem Informasi Perpustakaan Perguruan Tinggi berbasis *website*. Sistem ini dirancang berdasarkan studi kasus beberapa perpustakaan di perguruan tinggi yang masih menerapkan sistem konvensional. Dengan dibangunnya sistem informasi perpustakaan ini diharapkan dapat meningkatkan efektifitas dan efisiensi dalam mengelola data buku, mengelola informasi peminjaman, aktivitas penggunanya, dan terutama dalam pengelolaan denda peminjaman. Informasi yang dihasilkan meliputi data buku dan anggota, serta bisa melakukan transaksi peminjaman dan pengembalian buku. Selain itu sistem informasi perpustakaan ini dapat memberikan informasi denda untuk buku-buku yang balik melebihi batas waktu peminjaman yang dimana terdapat fitur memberikan informasi tarif denda yang harus dibayarkan oleh peminjam apabila telat mengembalikan buku atau melebihi batas waktu peminjaman.

REFERENSI

[1] L. Latifah and N. Ngalimun, "Pemulihan Pendidikan Pasca Pandemi Melalui Transformasi Digital Dengan Pendekatan Manajemen Pendidikan Islam Di Era Society 5.0," *J. Ter. Ilmu - Ilmu Sos.*, vol. 5, no. 1, p. 41, 2023, doi: 10.31602/jt.v5i1.10576.

[2] A. Yahya, "Efektivitas Pembelajaran Daring Terintegrasi di Era Pendidikan 4.0," *J. Teknol. dan Bisnis*, vol. 3, no. 2, pp. 269–280, 2021, doi: 10.37087/jtb.v3i2.103.

[3] R. Senjaya and A. Susinta, "Manajemen perpustakaan digital di era global pada Perpustakaan Kampus Institut Pemerintahan Dalam Negeri," *Unilib J. Perpust.*, vol. 13, no. 2, pp. 56–66, 2022, doi: 10.20885/unilib.Vol13.iss2.art1.

[4] A. Azhar, "Rancang Bangun Aplikasi Data Buku Dan Toko Buku Dengan Metode Location-Based Service Berbasis Android," *J. Real Ris.*, vol. 4, no. 2, pp. 164–173, 2022, doi: 10.47647/jrr.v4i2.644.

[5] A. Hidayat and A. Nurhasanah, "Sistem Informasi Arsip Surat di Fakultas Ekonomi Universitas Siliwangi," *Jumantaka*, vol. 3, no. 1, pp. 221–230, 2019.

[6] S. Muddin, A. Haslindah, R. Manatha, and S. Sartika, "Sistem Informasi Perpustakaan Pada Universitas Islam Makassar Berbasis Web," *ILTEK J. Teknol.*, vol. 15, no. 01, pp. 13–16, 2020, doi: 10.47398/iltek.v15i01.501.

[7] G. Hermawan and S. Wibowo, "Sistem Informasi Masjid Nurul Huda Berbasis Website Di Universitas Pgri Semarang," *Sci. Eng. Natl. Semin. 5*, vol. 5, no. 1, pp. 1–10, 2020, [Online]. Available: http://conference.upgris.ac.id/index.php/sens/article/view/1304/683

[8] S. Dewi and K. P. Sari, "Perancangan Layanan Book Ordering Pada Perpustakaan Universitas XYZ," vol. 11, no. 02, pp. 112–120, 2023.

[9] Pinem S and Maruli Pakpahan V, "Sistem Informasi Perpustakaan Pada Perpustakaan Universitas Efarina Berbasis Web," *J. STMIK Log.*, vol. 2, no. 1, pp. 49–56, 2019.

[10] H. Rajagukguk, M. Raharjo, Y. Isudianto, and H. Rajagukguk, "Berbasis Desktop Pada Lembaga Mimbar Politik," vol. 1, pp. 118–128, 2020.

[11] M. Pangri, S. Sunardi, and R. Umar, "Metode Pieces Framework Pada Tingkat Kepuasan Pengguna Sistem Informasi Perpustakaan Universitas Muhammadiyah Sorong," *Bina Insa. Ict J.*, vol. 8, no. 1, p. 63, 2021, doi: 10.51211/biict.v8i1.1499.

[12] A. Anwardi, A. Ramadona, M. Hartati, T. Nurainun, and E. G. Permata, "Analisis PIECES dan Pengaruh Perancangan Website Fikri Karya Gemilang Terhadap Sistem Promosi Menggunakan Model Waterfall," *J. Rekayasa Sist. Ind.*, vol. 7, no. 1, p. 57, 2020, [Online]. Available: https://jrsi.sie.telkomuniversity.ac.id/JRSI/article/view/380

[13] R. Aditya, V. H. Pranatawijaya, and P. B. A. A. Putra, "Rancang Bangun Aplikasi Monitoring Kegiatan Menggunakan Metode Prototype," *J. Inf. Technol. Comput. Sci.*, vol. 1, no. 1, pp. 47–57, 2021.

[14] N. H. Maulida, "Studi Literatur Penerapan Metode Prototype Dan Waterfall," *Stud. Lit. Penerapan Metod. Prototaype Dan Waterfall Dalam Pembuatan Sebuah Apl. Atau Website*, no. April, pp. 4–6, 2022.

[15] M. Purnasari, Y. Hartiwi, and N. Nurhayati, "Perancangan Sistem Informasi Pengelolaan Dana Masjid Berbasis Web Menggunakan Unified Modeling Language (UML)," *Resolusi Rekayasa Tek. Inform. dan Inf.*, vol. 2, no. 6, pp. 258–264, 2022, doi: 10.30865/resolusi.v2i6.416.

[16] T. Arianti, A. Fa'izi, S. Adam, and Mira Wulandari, "Perancangan Sistem Informasi Perpustakaan Menggunakan Diagram Uml (Unified Modelling Language)," *J. Ilm. Komput. ...*, vol. 1, no. 1, pp. 19–25, 2022, [Online]. Available: https://journal.polita.ac.id/index.php/politati/article/view/110/88

[17] D. Anggoro and A. Hidayat, "Rancang Bangun Sistem Informasi Perpustakaan Sekolah Berbasis Web Guna Meningkatkan Efektivitas Layanan Pustakawan," *Edumatic J. Pendidik. Inform.*, vol. 4, no. 1, pp. 151–160, 2020, doi: 10.29408/edumatic.v4i1.2130.

Analisis Sentimen Twitter tentang Isu Mental Health menggunakan Algoritma *Naïve Bayes* dan SVC

Guntur Firmansyah
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
gunturfirmansyah232@gmail.com

Regina Vannya
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
rvannya19@gmail.com

Agus Ardiyanto
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
Agussembhpar07@gmail.com

Rendi Setya Nugraha
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
rendisetya01@gmail.com

Rizky Fegiyanto
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
Fegix10@gmail.com

Pramadika Egamo
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
pramadika.5200411193@student.uty.ac.id

Abstrak—Kesehatan mental merupakan aspek penting dalam kehidupan sehari-hari, yang dapat dikelompokkan menjadi kesehatan mental yang baik dan buruk. Gangguan kesehatan mental seperti depresi, merupakan salah satu aspek yang sangat berpengaruh dalam menentukan kualitas kesehatan mental seseorang. Keberadaan gangguan kesehatan mental mulai meningkat melalui kasus bunuh diri yang mencapai sekitar 800.000 kasus pada tahun 2019. Penyebab utama seseorang melakukan bunuh diri dilandasi kondisi depresi pada individu sehingga menjadikan depresi sebagai penyakit peringkat ke-4 di dunia menurut data dari Organisasi Kesehatan Dunia (WHO). Penelitian ini bertujuan untuk melakukan eksplorasi terkait isu kesehatan mental melalui analisis sentiment menggunakan data tweet pengguna di platform X atau Twitter. Analisis sentiment dilakukan dengan memberikan label depresi dan tidak depresi pada setiap tweet yang selanjutnya dilakukan tahap pre-processing seperti tokenize, stopword, dan stemming. Terdapat dua algoritma yang digunakan yakni Naïve Bayes Classifier dan SVC guna melakukan klasifikasi sentiment dan membandingkan hasil dari pelatihan model. Hasil pengujian menunjukkan bahwa Naïve Bayes memiliki akurasi sebesar 86.03%, presisi 86%, recall 86%, dan F1-Score 86% sementara SVC memiliki akurasi sebesar 86.40%, presisi 87%, recall 87%, dan F1-Score 87%. Berdasarkan hasil tersebut dapat diketahui bahwa algoritma SVC lebih unggul dibandingkan algoritma Naïve Bayes. Hasil analisis sentiment dan penggunaan wordcloud terkait kata yang sering muncul diharapkan mampu memberikan pemahaman lebih baik terkait kondisi kesehatan mental pengguna X atau Twitter serta dapat menyorot tren isi kesehatan mental dari waktu ke waktu. Penerapan wordcloud pada penelitian ini memberikan informasi bahwa kata yang sering muncul pada kelas tidak depresi adalah thank, love, dan one sedangkan pada kelas depresi adalah depress, treatment, dan miss. Hasil penelitian juga memberikan informasi bahwa jenis gangguan kesehatan mental yang terindikasi pada pengguna platform X atau Twitter adalah depresi.

Kata Kunci—Analisis sentiment, Depresi, Kesehatan mental

I. PENDAHULUAN

Mental Health atau yang sering disebut kesehatan mental merupakan salah satu kondisi yang berkaitan dengan batin seseorang. Kesehatan mental dapat dibagi menjadi dua yaitu kesehatan mental yang baik dan kesehatan mental yang buruk. Kesehatan mental dapat dikatakan baik jika seseorang mampu menjalani hari-hari dengan tenang dan tentram serta membangun relasi yang baik dengan individu lainnya[1]. Sebaliknya, individu yang mengalami gangguan kesehatan mental akan menunjukkan gangguan dalam suasana hati, keahlian berpikir, dan control emosi yang dapat menyebabkan perilaku yang tidak diinginkan[2]. Adapun faktor dari gangguan kesehatan mental dapat berasal dari pengaruh aspek genetik, lingkungan sekitar, dan interaksi keduanya[3]. Gangguan kesehatan mental dapat memberikan efek buruk dalam kehidupan sehari-hari baik dari lingkungan sekitar maupun hubungan dengan orang-orang terdekat. Terdapat beberapa bentuk umum dari gangguan kesehatan mental yakni stress, gangguan kecemasan, dan depresi[4]. Dikutip dari halaman Unicef Indonesia diketahui bahwa faktor risiko yang memengaruhi remaja di Indonesia pada saat ini salah satunya adalah kesehatan mental. Hal tersebut membuktikan bahwa kesehatan mental menjadi faktor penting dalam perkembangan tumbuh kembang remaja khususnya di Indonesia saat ini. Gangguan kesehatan mental seseorang tentu bergantung pada masing-masing individu, khususnya tata cara dalam menangani hal tersebut. Dikutip dari halaman web komnasperempuan sekitar 20% dari 250 juta jiwa penduduk Indonesia belum memiliki layanan untuk menangani gangguan kesehatan mental. Hal tersebut memicu Sebagian individu yang merasa tidak sanggup dalam menangani gangguan kesehatan mental memilih untuk berakhir dengan bunuh diri.

Dalam rentang waktu Januari-Juni 2023, POLRI mencatat adanya 663 insiden bunuh diri di Indonesia. Angka tersebut menunjukkan peningkatan sebesar 36,4% dibandingkan dengan periode yang sama pada tahun 2021 yang mencatat 486 kasus. Provinsi-provinsi dengan jumlah kasus bunuh diri tertinggi adalah Jawa Tengah sebanyak 253 kasus, Jawa Timur sebanyak 128 kasus, Bali 61 kasus, dan Jawa Barat 39 kasus. Kasus bunuh diri tersebut memiliki kecenderungan yang

dipicu oleh gangguan kesehatan mental dengan berbagai faktor dibelakangnya. Berdasarkan kasus tersebut, penderita gangguan kesehatan mental tentu harus diperhatikan dengan baik oleh individu yang ada di lingkungan sekitarnya. Di dalam penelitian ini dilakukan analisis sentiment terhadap kesehatan mental pengguna platform X berdasarkan cuitan tweet individu. Analisis sentiment merupakan sebuah metode yang digunakan untuk mengklasifikasikan teks, baik di dalam dokumen maupun kalimat dalam menentukan sentimen tiap baris data bersifat positif, netral, atau negatif. Metode yang digunakan dalam pengklasifikasian sentiment menggunakan naïve bayes classifier dan SVC. Naïve bayes classifier adalah salah satu algoritma yang menggunakan teori probabilitas dan frekuensi klasifikasi data training dalam melakukan klasifikasi variabel[5]. Algoritma ini termasuk dalam kategori supervised learning yang memerlukan label sebagai dasar pembelajaran berdasarkan kebenaran sebelumnya. Menurut situs web resmi Scikit-learn, Support Vector Classifier (SVC) adalah algoritma yang berasal dari library SVM[6]. Atribut numerik meliputi panjang teks dan jumlah tanda baca, sementara atribut teks mengandung isi seluruh pesan. Hasil akhir dari penelitian ini berupa analisis terhadap banyaknya penderita gangguan kesehatan mental dan jenis mental health yang lebih dominan diidap oleh pengguna di platform X. Penelitian ini diharapkan mampu memberikan pemahaman yang lebih baik terhadap gangguan kesehatan mental serta melakukan pendeteksian perubahan dan tren dalam peningkatan jumlah kasus gangguan kesehatan mental dari waktu ke waktu.

II. KAJIAN LITERATUR

Dalam penelitian ini, peneliti menggali informasi dari penelitian-penelitian sebelumnya sebagai dokumen pembanding yang mempunyai bidang dan topik yang sama dengan penelitian yang akan dilakukan.

Penelitian yang dilakukan oleh Sri Hadianti dan Firman Yosep Tember[7], dengan judul Analisis Sentiment Covid-19 di Twitter Menggunakan Metode Naïve Bayes dan SVM. Penelitian tersebut berfokus pada perbandingan hasil klasifikasi metode Niave Bayes dan SVM, serta mengetahui kecenderungan opini masyarakat di Twitter. Dari hasil pengujian diketahui metode SVM memiliki tingkat akurasi yang paling tinggi dari tingkat akurasi sebesar 54.21% dan didapat pula kecenderungan opini masyarakat di seluruh dunia condong negatif, hal tersebut dapat dilihat dari opini positif sebesar 98 dan negatif sebesar 116. Sedangkan metode Niave Bayes memperoleh tingkat akurasi sebesar 53.27%.

Penelitian yang dilakukan oleh Devi Irawan dkk[8], dengan judul Perbandingan Klasifikasi SMS Berbasis Support Vector Machine, Naïve Bayes Classifier, Random Forest dan Bagging Classifier. Penelitian tersebut melakukan perbandingan dengan beberapa metode untuk klasifikasi SMS, sehingga mendapatkan performance score paling tinggi yaitu metode Bagging Classifier 97.4% sedangkan metode random forrest 96.8%, metode naïve bayes 96.4%, dan terahkir metode support vector machine 86.5%.

Penelitian yang dilakukan oleh Haekal Hilmi Zain dkk[9], dengan judul Perbandingan Model Svm, Knn Dan Naïve Bayes Untuk Analisis Sentimen Pada Data Twitter: Studi Kasus Calon Presiden 2024. Penelitian tersebut berfokus pada pengeksplorasian metode terbaik untuk mengalisis sentimen Twitter terkait calon presiden dalam pemilu 2024 menggunakan tiga algoritma kalsifikasi: SVM, KNN, Naïve Bayes. Melalui tahap pengumpulan data, preporcessing, labelling, word embedding, hyperparameter tuning, dan evaluasi, berdasarkan penelitian tersebut bahwa algoritma SVM memberikan kinerja superios dengan tingkat akurasi total sebesar 0.88, menunjukan konsistensi tingkat dalam presisi dan recall untuk semua kategori sentimen.

Penelitian yang dilakukan oleh Merinda Lestandy dkk[10], dengan judul Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes. Penelitian tersebut memperoleh dataset sebanyak 5000 tweet vaksin COVID-19 dengan pembagian 3800 tweet sentimen positif, 800 tweet sentimen negatif dan 400 tweet sentimen netral. Dataset tersebut dilakukan pre-processing untuk mengoptimalkan pengolahan data. Sehingga hasil pengujian menunjukan RNN (TF-IDF) memiliki akurasi lebih besar yaitu 97,77% dibandignkan Naïve Bayes (TF-IDF) sebasar 80%.

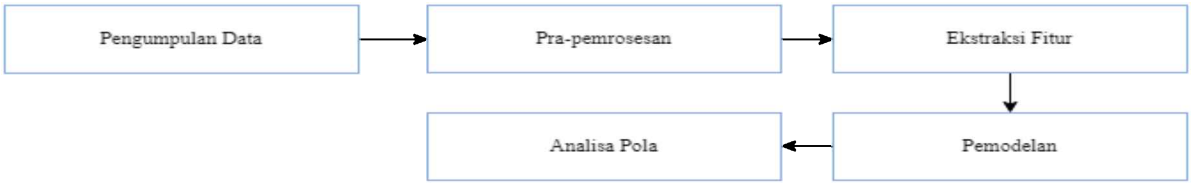
Penelitian yang dilakukan oleh Primandani Arsi dan Retno Waluyo[11], dengan judul Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM). Penelitian tersebut mengusulkan metode SVM untuk diterapkan pada tweets topik pemindahan ibu kota indonesia untuk tujuan klasifikasi sentimen pada media sosial twitter. Teknis klasifikasi dilakukan dengan cara mengklasifikasikan menjadi 2 kelas yakni positif dan negatif. Berdasarkan hasil pengujian yang dilakukan terhadapat tweets sentimen pemindahan ibu kota dari media sosial twitter sebanyak 1.236 tweets (404 positif dan 832 negatif) menggunakan SVM diperoleh akurasi 96,68%.

III. METODOLOGI PENELITIAN

A. Deskripsi Metodologi

Aktivitas pengguna merupakan elemen yang penting terhadap analisis pengguna. Beberapa penelitian telah memanfaatkan algoritma Naïve Bayes dan Support Vector Machine untuk melakukan analisa sentiment. Metodologi disajikan pada Gambar 1. Alur kerja metodologi adalah sebagai berikut:

- Pengumpulan Data. Data berasal dari website Kaggle.
- Pra-pemrosesan. Melakukan Tokenization, Stopword, dan Stemming.
- Ekstraksi fitur. Pengambilan ciri untuk menggambarkan karakteristik data menggunakan TF-IDF Vectorizer.
- Pemodelan. Model yang digunakan yaitu Naïve Bayes dan Support Vector Machine
- Analisa Pola. Pola-pola yang ditemukan pada tahap sebelumnya dipelajari untuk mengetahui kesehatan mental pengguna platform Twitter.



Gambar 1. Metodologi untuk analisis sentiment terhadap pengguna Twitter

B. *Pengumpulan Data*

Proses pengumpulan data dilakukan melalui langkah-langkah yang mencakup pengunduhan dataset dari situs web Kaggle. Dataset yang digunakan dalam penelitian ini dikenal sebagai "Depression: Twitter Dataset + Feature Extraction," yang mengandung sekitar 20.000 data tweet dan terstruktur dalam 11 kolom. Mulai dari kolom post_id, post_created hingga label. Dataset ini memberikan beragam informasi yang relevan dengan penelitian, dapat dilihat contoh sampel dataset yang didapat padat tabel 1.

TABEL 1. SAMPEL DATASET

Sampel Dataset	
Text	***Sentiment***
It's just over 2 years since I was diagnosed with #anxiety and #depression. Today I'm taking a moment to reflect on how far I've come since.	Depresi
@Ashton5SOS I would literally cry for weeks	Depresi
RT @auliicravalho: HAPPY NEW YEAR!!! ðŸŒŸ As this year ends, I am so SO thankful for all of you. ðŸ˜˜ Sending love from Hawai'i!! #HappyNewYear #â€¦	Tidak Depresi
RT @katmcnamvra: plot twist: gabriella and sharpay forgot about troy, fell in love and now live happily ever after. https://t.co/9HDXg18L9e	Tidak Depresi

C. *Pra-premrosesan*

Pra-pemrosesan data adalah suatu langkah awal dalam analisis data yang bertujuan untuk membersihkan dan menyiapkan data mentah agar menjadi lebih siap untuk analisis lebih lanjut. Dalam konteks penelitian ini, tahap pra-pemrosesan dimulai dengan menghilangkan karakter-karakter tertentu seperti angka, Retweet, @, link, dan sebagainya menggunakan ekspresi regular. Lalu mengubah nama label menjadi "Depresi" dan "Tidak Depresi" untuk mempermudah proses analisis. Setelah itu, dilakukan tokenisasi, dimana kalimat-kalimat tweet dibagi menjadi kata-kata dasar untuk memfasilitasi pemahaman dan analisis lebih lanjut. Tahap berikutnya melibatkan proses penghapusan stopword, di mana kata-kata umum yang cenderung tidak memberikan makna signifikan dihilangkan untuk menyederhanakan teks. Terakhir, dilakukan stemming untuk mereduksi kata-kata menjadi bentuk dasarnya, memastikan konsistensi dan keefektifan analisis selanjutnya. Dengan serangkaian langkah pra-pemrosesan ini, data tweet menjadi lebih siap untuk tahap analisis sentimen yang lebih mendalam.

D. *Ekstraksi Fitur*

Dalam bagian Ekstraksi Fitur, metode yang digunakan adalah Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF digunakan untuk mengekstrak fitur-fitur penting dari dataset tweet yang akan membantu menggambarkan kepentingan dan relevansi kata-kata dalam konteks analisis sentimen terkait kesehatan mental. Metode ini memberikan bobot yang tinggi untuk kata-kata yang sering muncul dalam tweet tertentu namun jarang muncul dalam keseluruhan dataset, sehingga mengidentifikasi kata-kata kunci yang dapat mencirikan setiap kategori sentimen dengan lebih baik. Proses ekstraksi fitur ini menjadi langkah krusial dalam mempersiapkan data untuk analisis sentimen selanjutnya.

E. *Pemodelan*

- Naïve Bayes

  Naive Bayes adalah sebuah metode klasifikasi dalam statistika dan pembelajaran mesin yang didasarkan pada Teorema Bayes. Dasar dari Naive Bayes adalah Teorema Bayes, yang menyatakan hubungan antara distribusi probabilitas kondisional. Dengan asumsi fitur-fitur tersebut saling independen, probabilitas kelas dapat dihitung menggunakan rumus Teorema Bayes:

$$P(kelas|fitur) = \frac{P(fitur|kelas).P(kelas)}{P(fitur)} \quad (1)$$

  dimana:

  - P(kelas | fitur) adalah probabilitas kelas setelah mengamati fitur.
  - P(fitur | kelas) adalah probabilitas fitur given kelas.
  - P(kelas) adalah probabilitas prior kelas.
  - P(fitur) adalah probabilitas prior fitur

- Support Vector Machine

  SVM bekerja dengan mencari hyperplane optimal yang memaksimalkan margin dan pada saat yang bersamaan meminimalkan kesalahan klasifikasi. Jika data tidak dapat dipisahkan secara linear, SVM dapat menggunakan fungsi kernel untuk mentransformasikan data ke dalam dimensi yang lebih tinggi, di mana pemisahan linear mungkin dimungkinkan. Rumus dasar untuk Support Vector Machine (SVM) dalam konteks klasifikasi linear adalah sebagai berikut:

$$f(x) = sign(w.x + b) \quad (2)$$

  dimana:

  - f(x) adalah fungsi keputusan.
  - w adalah vektor bobot.

- x adalah vektor fitur input.
- b adalah bias.
- . menyatakan produk dot antara w dan x.
- Sign(.) adalah fungsi tanda yang mengembalikan +1 jika nilai di dalam kurung lebih besar dari nol, -1 jika kurang dari nol, dan 0 jika sama dengan nol.

Jika data tidak dapat dipisahkan secara linear, SVM dapat menggunakan kernel untuk memetakan data ke dalam dimensi yang lebih tinggi. Dalam hal ini, fungsi keputusan menjadi:

$$f(x) = sign(\sum_{i=1}^{N} \alpha i \, \gamma i \, K(Xi, X + b) \quad (3)$$

dimana:

- N adalah jumlah sampel pelatihan.
- αi adalah koefisien Lagrange.
- γi adalah label kelas
- K(Xi,X) adalah fungsi kernel yang menghitung hasil transformasi kernel antara Xi dan X.

F. *Analisa pola*

Analisa pola diperlukan untuk menyaring aturan-aturan atau pola-pola yang menarik dari suatu himpunan dengan menghilangkan aturan-aturan atau pola-pola yang tidak berkaitan. Teknik yang umum digunakan untuk menganalisis pola adalah teknik visualisasi, pendekatan visualisasi dilakukan menggunakan *Wordcloud* untuk memahami pola dan tren dalam dataset tweet terkait kesehatan mental.

IV. EKSPERIMEN DAN ANALISIS

A. *Persiapan dataset dan Pra-pemrosesan*

Pada penelitian ini menggunakan dataset tentang Mental Health yang diambil dari situs kaggle. Dataset tersebut terdiri dari 20000 data tweet twitter mengenai Mental Health dan terbagi menjadi 2 kelas yaitu kelas Depresi dan Tidak Depresi. Dataset tersebut melalui pra pemrosesan teks sebelum akhirnya bisa digunakan dalam penelitian klasifikasi ini. Tahapan dalam pra pemrosesan teks adalah Tokenization, Stopword, dan Stemming. Penerapan pra pemrosesan berguna untuk melakukan pemecahan kalimat menjadi antar kata, menghilangkan kata pendukung seperti *and, or* serta melakukan pemotongan akhiran kata untuk mengurangi variasi kata yang berasal dari inti kata yang sama.

TABEL 2. HASIL PRA PEMROSESAN TEKS

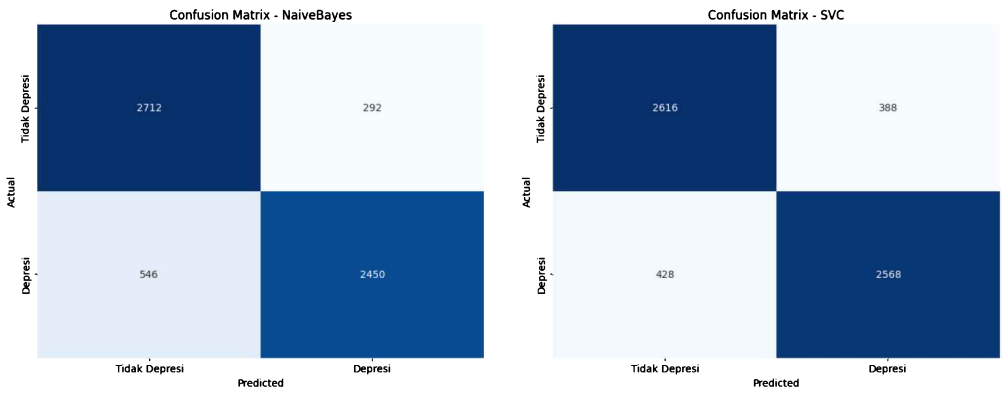
Hasil Pra Pemrosesan Teks	
Text	***Text Clean***
It's just over 2 years since I was diagnosed with #anxiety and #depression. Today I'm taking a moment to reflect on how far I've come since.	year sinc diagnos anxieti depress today take moment reflect far come sinc
Currently in the finding-boxes-of-random-shit packing phase. I think I • € • m a closet hoarder...	current find box random shit pack phase think closet hoarder
RT @ashleiholtson: he says: ur not like other girls he means: gender norms are so engrained in me that u drinking beer while having a vagin • € ℸ	say ur like girl mean gender norm engrain u drink beer vagin

Pada Tabel 2, hasil Pra Pemrosessan Teks bisa dilihat bahwa hasil dari proses Tokenization, Stopword, dan Stemming menghasilkan kata kata tersebut. Setelah proses Pra Pemrosesan kemudian dilanjutkan dalam tahap pembagian data training dan data testing. Pada penelitian ini menggunakan 70% data training dan 30% data testing atau sebanyak 14000 data training dan 6000 untuk data testing. Selanjutnya data tersebut digunakan untuk ekstraksi fitur dan klasifikasi.

B. *Ekstraksi Fitur dan Klasifikasi*

Setelah data tersebut melewati tahap preprocesing dan split data maka didapatkan data yang digunakan untuk proses selanjutnya yaitu ekstraksi fitur dan klasifikasi. Pada proses ekstraksi fitur ini menggunakan TF-IDF Vectorizer yaitu mengubah sekumpulan data menjadi sebuah vector. Setelah didapatkan kumpulan vector tersebut maka dilakukan klasifikasi naïve bayes classifier dan support vector clasiffier. Hasil dari pengujian klasifikasi ini adalah model Naïve bayes mendapatkan akurasi sebesar 86,03%. Sedangkan model SVC memperoleh akurasi sebesar 86,40%.

Selanjutnya adalah confusion matrix, yang mana adalah sebuah tabel perbandingan yang menghitung berapa banyak data yang diprediksi benar atau salah. Hasil confussion matrix dari klasifikasi ini bisa dilihat pada Gambar 2. Hasil Confussion Matrix.



Gambar 2. Hasil Confussion Matrix

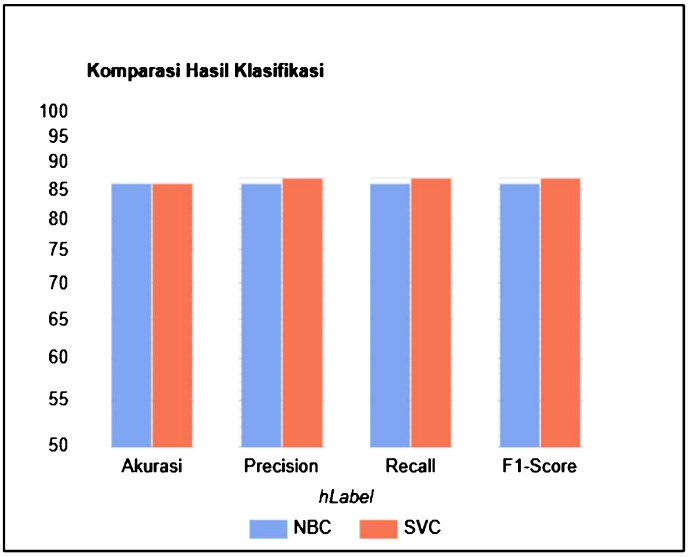
Pada Gambar 2. bisa diketahui bahwa pada confussion matrix Naïve Bayes Label Tidak Depresi yang diprediksi benar adalah 2712 data dan yang diprediksi salah adalah 292 data sedangkan untuk label Depresi yang diprediksi benar adalah 2450 data dan yang diprediksi salah sejumlah 546 data. Kemudian untuk hasil confussion matrix SVC label Tidak depresi yang diprediksi benar adalah 2616 data dan yang salah sejumlah 388 data sedangkan untuk Label Depresi yang diprediksi benar adalah 2568 data dan yang diprediksi salah sejumlah 428 data. Kemudian untuk hasil lebih lanjut tentang classification report atau hasil klasifikasi bisa dilihat pada Tabel 3 Hasil Klasifikasi Per-label.

TABEL 3. HASIL KLASIFIKASI PER-LABEL

	Detail Per Label					
	Accuracy		Precision		Recall	
	Tidak Depresi	Depresi	Tidak Depresi	Depresi	Tidak Depresi	Depresi
NBC	0,90	0,82	0,89	0,83	0,82	0,90
SVC	0,87	0,857	0,87	0,86	0,86	0,87

TABEL 4. HASIL KLASIFIKASI KESELURUHAN

	NBC	SVC
Akurasi	86%	86%
Precision	86%	87%
Recall	86%	87%
F1-Score	86%	87%



Gambar 3. Komparasi Algoritma Klasifikasi

Pada Tabel 3. diatas merupakan hasil dari classification report atau hasil klasifikasi antara dua algoritma klasifikasi yaitu algoritma Naive Bayes dan SVC. Gambar diatas menunjukkan bahwa dalam hasil klasifikasi ini terdapat beberapa parameter sebagai acuan seperti, Akurasi, Precision, Recall dan F1-Score. Hasil klasifikasi tersebut membandingan antara hasil detail per label dan komparasi antar algoritma. Kemudian untuk lebih lanjut hasil total akurasi berdasarkan keseluruhan bisa dilihat pada Tabel 4 Hasil Klasifikasi Keseluruhan.

Pada Tabel 4 merupakan hasil dari classification report atau hasil klasifikasi keseluruhan antara dua algoritma klasifikasi yaitu algoritma Naive Bayes dan SVC. Pada tabel tersebut adalah mengambil Akurasi, Precision, Recall dan F1-Score dengan nilai akurasi antara label Depresi dan Tidak Depresi. Kemudian komparasi antara algoritma naive bayes dan svc bisa dilihat pada Gambar 3. Komparasi Algoritma Klasifikasi.

Pada Gambar 3. dapat disimpulkan bahwa algoritma SVC lebih optimal dibandingkan dengan algoritma Naive bayes ada beberapa pokok kesimpulan yang bisa diambil adalah sebagai berikut.

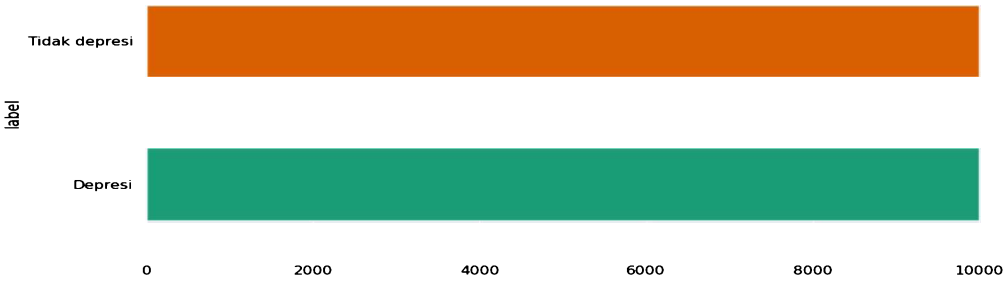
1. Nilai accuracy yang didapatkan dari pengujian Naïve Bayes sebesar 86,03% sedangkan SVC sebesar 86,40%.

2. Nilai precision yang didapatkan dari metode Naïve Bayes dengan nilai 86% dan SVC 87%.
3. Nilai recall dari data diatas didapatkan Naïve Bayes 86% dan SVC 87%.
4. Nilai F1-Score didapatkan dari Naïve Bayes 86% dan SVC 87%.

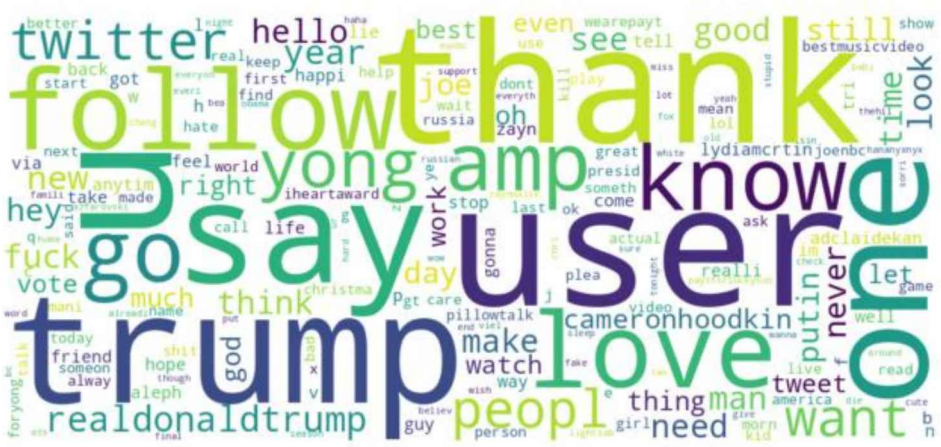
Berdasarkan hasil dari kesimpulan tersebut dapat dikatakan bahwa model klasifikasi SVC memiliki tingkat akurasi, presisi, recall, dan F1-Score yang tinggi dari semua pengujian, dalam mengklasifikasikan data, dibandingkan dengan model klasifikasi Naïve bayes. Walaupun dari segi akurasi hanya selisih sedikit nilai saja namun algoritma SVC tetap dinilai lebih optimal digunakan dalam klasifikasi ini.

C. *Visualisasi*

Pada penelitian ini, visualisasi menggunakan *Wordcloud* untuk menggambarkan kata-kata yang paling sering muncul dalam teks. Ukuran kata dalam visualisasi ini mencerminkan frekuensi kemunculannya, memberikan gambaran tentang kata-kata yang mendominasi dataset. Dari gambar 4 dapat diketahui bahwasannya sebanyak 20.000 data tweet berhasil dikumpulkan, dengan 10.000 data berlabel "Tidak Depresi" dan 10.000 data berlabel "Depresi". Ini menunjukkan bahwa tweet dalam dataset mengenai "Kesehatan Mental" memiliki kecenderungan yang seimbang dan merata antara kedua label tersebut.



Gambar 4. Jumlah *tweets* dari setiap label



Gambar 5. *Wordcloud* sentimen Tidak Depresi

1) *Wordcloud* Tidak Depresi

Pada Gambar 5, terdapat *wordcloud* yang menggambarkan kata-kata dengan sentimen Tidak Depresi dari pengguna Twitter yang berkaitan dengan Kesehatan Mental. Ukuran kata dalam *wordcloud* mencerminkan seberapa sering kata tersebut digunakan, dengan ukuran yang lebih besar menandakan frekuensi yang lebih tinggi. Analisis *wordcloud* menunjukkan bahwa kata yang paling dominan adalah "Thank," diikuti oleh "Love" dan "One." Ini dapat diartikan bahwa pada periode tersebut, tweet-tweet mencerminkan pentingnya memperhatikan dan merayakan kesehatan mental, serta menciptakan kondisi yang positif. Kata-kata yang sering digunakan ini mencerminkan sentimen positif atau ketiadaan depresi dalam konteks tersebut.

Untuk lebih memperkuat pemahaman, perlu diperjelas bahwa kata-kata seperti "Thank" menunjukkan rasa terima kasih atau apresiasi, "Love" mencerminkan kasih sayang, dan "One" mungkin mencerminkan individualitas atau pemahaman diri. Dengan demikian, *wordcloud* ini menangkap nuansa positif dan dukungan terhadap kesehatan mental dari perspektif yang tidak depresi.

2) *Wordcloud* Depresi

Pada Gambar 6, tersedia *wordcloud* yang memvisualisasikan kata-kata dengan sentimen negatif atau terkait depresi. *Wordcloud* ini mencerminkan tantangan dalam kesehatan mental, khususnya depresi, dengan menyoroti pernyataan-pernyataan yang bersifat negative. Analisis *wordcloud* menunjukkan bahwa beberapa kata yang dominan termasuk "Depress," "Treatment," "Miss," "Love," dan "Fuck." Pemilihan kata ini dapat dihubungkan dengan kelas "Depresi" karena mencerminkan berbagai aspek pengalaman emosional yang sering terkait dengan kondisi tersebut. Misalnya, kata "Depress" menjadi pusat perhatian dan dapat dijelaskan sebagai ungkapan langsung tentang

perasaan depresi. Kata-kata lain seperti "Treatment," "Miss," "Love," dan "Fuck" dapat mencerminkan kompleksitas emosi, pengalaman hubungan, dan perasaan yang mendalam yang sering terkait dengan kondisi kesehatan mental ini. Lebih lanjut, analisis kontekstual dari kata-kata ini dapat memberikan pemahaman mendalam tentang mengapa mereka termasuk dalam kelas "Depresi."

Kemunculan kata "love" yang sering muncul di kedua label, baik "Depresi" maupun "Tidak Depresi," dapat dijelaskan dengan melihat konteks penggunaan kata tersebut. Dalam konteks "Tidak Depresi," kata "love" mungkin digunakan untuk menyampaikan dukungan emosional, kasih sayang, atau apresiasi terhadap topik kesehatan mental. Hal ini mencerminkan respons positif dan upaya untuk menciptakan lingkungan online yang penuh empati. Di sisi lain, dalam konteks "Depresi," kemungkinan kata "love" dipakai sebagai ekspresi keinginan akan dukungan emosional, cinta, atau harapan. Dengan demikian, kemunculan kata "love" menunjukkan kompleksitas dan keragaman cara individu menyampaikan dan merespons isu-isu kesehatan mental di platform Twitter.

Gambar 6. *Wordcloud* sentimen Depresi

```
Berapa banyak data yang ingin diinputkan? 2
Masukkan Kalimat Yang akan Diprediksi: I was dizzy and very tired, all the problems made me depressed
Masukkan Kalimat Yang akan Diprediksi: This is a beautiful gift, thank you
Kalimat yang diprediksi ke-0: I was dizzy and very tired, all the problems made me depressed
Kalimat yang diprediksi ke-1: This is a beautiful gift, thank you
Sentimen kalimat ke 0 Adalah depresi
Sentimen kalimat ke 1 Adalah Tidak depresi
```

Gambar 7. Uji coba manual

D. Uji Coba Prediksi

Secara manual, eksperimen dilakukan dengan menggunakan model hasil pelatihan, khususnya dengan algoritma SVC yang mendapatkan akurasi terbaik, pada data pelatihan dan pengujian. Pada Gambar 7. terlihat uji coba yang melibatkan input dua kalimat, satu dengan unsur depresi: "*I was dizzy and very tired, all the problems made me depressed*," dan satu tidak depresi: "*This is a beautiful gift, thank you*." Kedua kalimat ini berhasil diprediksi dengan benar.

V. KESIMPULAN

Berdasarkan hasil klasifikasi menggunakan metode Naïve Bayes dan Support Vector Classifier (SVC), ditemukan bahwa keduanya mencapai tingkat akurasi yang signifikan pada data testing, yaitu 86,03% untuk Naïve Bayes dan 86,40% untuk SVC. Tingkat akurasi yang tinggi ini menunjukkan kemampuan baik dalam melakukan klasifikasi berbasis teks terhadap analisis sentimen pada dataset tweet terkait kesehatan mental. Analisis terhadap kedua label, "Tidak Depresi" dan "Depresi", mengungkapkan dominasi kata-kata positif seperti 'Thank', 'Love', dan 'One' pada sentiment tidak depresi terkait tweets kesehatan mental di Twitter. Sebaliknya, pada sentiment depresi, kata-kata yang mendominasi adalah 'Depress', 'Treatment', 'Miss', 'Love', dan 'Fuck'. Hal ini menunjukkan bahwa individu menggunakan platform Twitter sebagai sarana untuk mengekspresikan perasaan negatif atau terkait depresi pada saat itu.

Hasil ini juga mengindikasikan bahwa jenis gangguan kesehatan mental yang paling banyak didominasi dalam dataset adalah depresi. Dengan demikian, hasil klasifikasi dan analisis sentimen memberikan pemahaman yang lebih mendalam tentang bagaimana masyarakat secara online merespons dan berpartisipasi dalam percakapan mengenai kesehatan mental di platform Twitter.

REFERENSI

[1] T. D. Ariyanti, "Psikoedukasi Untuk Meningkatkan Literasi Kesehatan Mental Pada Remaja," *Jurnal Kesehatan*, vol. 13, 2022.

[2] E. Aprilia and S. Winduwati, "Komunikasi Antarpribadi Caregiver dan Penyintas Gangguan Mental dalam Membangun Hubungan," *Koneksi*, vol. 7, no. 1, 2023, doi: 10.24912/kn.v7i1.15933.

[3] N. Aisyaroh, I. Hudaya, and R. Supradewi, "Trend Penelitian Kesehatan Mental Remaja Di Indonesia Dan Faktor Yang Mempengaruhi: Literature Review," *Scientific Proceedings of Islamic and Complementary Medicine*, vol. 1, no. 1, 2022, doi: 10.55116/spicm.v1i1.6.

[4] I. Nurhafiyah and H. Marcos, "Sistem Pakar Diagnosis Kesehatan Mental Pada Mahasiswa Universitas Amikom Purwokerto," *Komputa : Jurnal Ilmiah Komputer dan Informatika*, vol. 12, no. 1, 2023, doi: 10.34010/komputa.v12i1.8978.

[5] V. Herliansyah, R. Latuconsina, and A. Dinimaharawati, "Prediksi Stunting Pada Balita Dengan Menggunakan Algoritma Klasifikasi Naïve Bayes," *e-Proceeding of Engineering*, vol. 8, no. 5, 2021.

[6] E. Pudjiarti, "Prediksi Spam Email Menggunakan Metode Support Vector Machine Dan Particle Swarm Optimization," *Jurnal Pilar Nusa Mandiri*, vol. 12, no. 2, 2016.

[7] S. Hadianti *et al.*, "Analisis Sentiment Covid-19 Di Twitter Menggunakan Metode Naive Bayes Dan SVM," *Jurnal Teknologi Informasi*, vol. 6, no. 1, 2022.

[8] D. Irawan, E. B. Perkasa, Y. Yurindra, D. Wahyuningsih, and E. Helmud, "Perbandingan Klassifikasi SMS Berbasis Support Vector Machine, Naive Bayes Classifier, Random Forest dan Bagging Classifier," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 10, no. 3, 2021, doi: 10.32736/sisfokom.v10i3.1302.

[9] H. Hilmi Zain, R. Maulana Awannga, and W. Isti Rahayu, "Perbandingan Model Svm, Knn Dan Naïve Bayes Untuk Analisis Sentiment Pada Data Twitter: Studi Kasus Calon Presiden 2024," *JIMPS: Jurnal Ilmiah Mahasiswa Pendidikan Sejarah*, vol. 8, no. 3, 2023.

[10] Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa'ah, "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, 2021, doi: 10.29207/resti.v5i4.3308.

[11] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 8, no. 1, 2021, doi: 10.25126/jtiik.0813944.

Klasifikasi Kematangan Buah Salak Pondoh menggunakan Metode *Support Vector Machine*

Josephine Diva Ayurveda Verol
Departemen Informatika
Universitas Sanata Dharma
Yogyakarta, Indonesia
josephdiva2@gmail.com

Anastasia Rita Widiarti
Departemen Informatika
Universitas Sanata Dharma
Yogyakarta, Indonesia
rita_widiarti@usd.ac.id

Abstrak— Salak pondoh adalah satu dari sekian banyak produk pertanian popular yang berasal dari Sleman, Yogyakarta. Kepopuleran buah ini, didukung oleh produksi berupa panen tahunan yang melimpah. Namun, permasalahan utamanya terletak pada proses pendistribusian atau ekspor, dimana tingkat keberhasilannya dipengaruhi oleh beberapa faktor, diantaranya adalah proses seleksi salak. Pada kenyataannya, pemilihan buah salak pondoh dengan kematangan tertentu dilakukan oleh manusia sehingga resiko kesalahan masih cukup tinggi. Penelitian mengenai kematangan buah salak pondoh pun masih terbatas dan lebih berfokus pada kualitas salak. Proses klasifikasi kematangan buah salak pondoh dimulai dengan akuisisi citra buah salak dengan tiga jenis kematangan, yaitu matang, setengah matang dan mentah. Tahap selanjutnya adalah pengolahan citra yang terdiri dari lima tahapan. Citra hasil pengolahan dengan model warna RGB akan diubah menjadi HSV yang selanjutnya digunakan untuk ekstraksi fitur warna berupa besaran statistik orde pertama. Pengujian model ini dilakukan menggunakan metode k-fold cross validation dengan nilai k sebesar 3. Model hasil pengujian menunjukkan bahwa kernel linear dan fitur campuran dari model warna dan besaran statistik berhasil memperediksi data uji dengan akurasi sebesar 96%. Penelitian ini menunjukkan bahwa penggunaan metode SVM dengan fitur berupa model warna memiliki performa yang sangat tinggi dalam menentukan kematangan buah salak pondoh.

Kata Kunci— Salak pondoh, Support Vector Machine (SVM), klasifikasi, model warna RGB, model warna HSV, k-fold cross validation, confusion matrix

I. PENDAHULUAN

Salak pondoh adalah satu dari sekian banyak produk pertanian popular yang berasal dari Sleman, Yogyakarta. Popularitas salak pondoh dapat dilihat dari jumlah permintaan konsumen yang tinggi dan berasal dari berbagai daerah atau bahkan luar negeri. Kepopuleran buah ini, didukung oleh produksi berupa panen tahunan yang melimpah. Namun, permasalahan utamanya terletak pada proses pendistribusian atau ekspor, dimana tingkat keberhasilannya dipengaruhi oleh beberapa faktor, diantaranya adalah proses seleksi salak [1]. Buah salak pondoh dengan tingkat kematangan sedang atau setengah matang dipilih oleh para pengepul agar bisa matang sempurna selama perjalanan. Maka dari itu, pemilahan buah salak pondoh berdasarkan tingkat kematangannya memiliki peran penting dalam menjaga kualitas salak ketika diterima konsumen. Pada kenyataannya, pemilihan buah salak pondoh dengan kematangan tertentu dilakukan oleh manusia sehingga resiko kesalahan masih cukup tinggi.

Di sisi lain, banyak penelitian yang dilakukan untuk mengklasifikasikan buah salak. Namun sebagian besar penelitian cenderung fokus pada kualitas buah salak, seperti penelitian yang dilakukan oleh Wibawa dan Arif dengan metode ELM mencapai akurasi sebesar 95% dan metode SVM sebesar 97,3% [2]. Klasifikasi menggunakan deep learning khususnya transfer learning dengan arsitektur VGG16 mencapai akurasi 95,83% [3]. Pembelajaran transfer dengan arsitektur Xception digunakan Rismiyati untuk mengklasifikasikan kualitas salak dan mencapai akurasi 94,44% [4]. Sedangkan sistem penentuan kematangan berhasil dibuat dengan mengklasifikasikan citra dompolan salak dan mencapai akurasi 92% untuk algoritma backpropagation dan 93% untuk algoritma K-Nearest Neighbor [5].

Berdasarkan kajian klasifikasi diatas, metode SVM mempunyai nilai akurasi paling tinggi. Metode ini juga telah banyak digunakan untuk mengklasifikasikan kematangan buah. Penggunaan metode SVM untuk menentukan kematangan tomat menggunakan model warna CIELab mencapai akurasi 100% [6]. Klasifikasi lain yang menggunakan karakteristik warna menghasilkan akurasi 92,5% untuk buah sawit [7]. Penggunaan model warna HSV dengan metode ini juga mencapai akurasi yang sangat tinggi dalam mengklasifikasikan kematangan nanas [7]. Konversi model warna RGB menjadi LAB sebagai fitur klasifikasi kematangan jeruk dengan metode yang sama mencapai akurasi 80% [8]. Pada penelitian lain, karakteristik yang digunakan untuk mengklasifikasikan kematangan pepaya adalah warna, tekstur, dan bentuk dengan akurasi mencapai 65% hingga 66% [9]. Dari hasil penelitian dengan menggunakan metode SVM diperoleh nilai akurasi lebih dari 75% ketika menggunakan model warna sebagai ciri untuk mengklasifikasikan kematangan buah.

II. METODE PENELITIAN

Proses klasifikasi kematangan buah salak pondoh memiliki alur seperti pada gambar 1. Proses dimulai dengan akuisisi citra buah salak dengan tiga jenis kematangan, yaitu matang, setengah matang dan mentah. Tahap selanjutnya adalah pengolahan citra yang terdiri dari lima tahapan. Citra hasil pengolahan dengan model warna RGB akan diubah menjadi HSV yang selanjutnya digunakan untuk ekstraksi fitur warna berupa besaran statistik orde pertama. Seluruh fitur dari citra latih akan digunakan untuk menguji model *Support Vector Machine* (SVM). Setelah mendapatkan model, proses selanjutnya adalah pengujian model dengan metode *k-fold cross validation* dan perhitungan akurasi dengan *confusion matrix*.

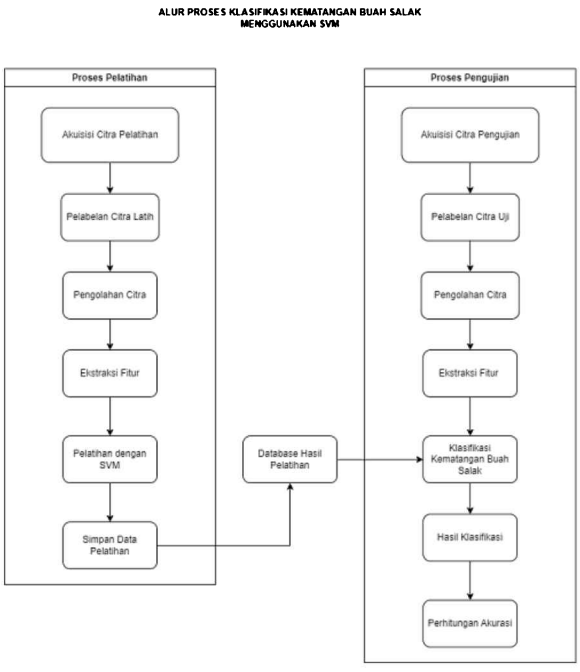
A. *Akuisisi Citra*

Peran pengepul dapat dikatakan cukup besar dalam menjamin nilai jual buah salak pondoh, baik dari segi kualitas maupun dalam pemasaran. Maka dari itu, penelitian ini berfokus pada tahap klasifikasi kematangan di sisi pengepul sehingga data yang digunakan adalah citra buah salak pondoh. Data yang diperoleh merupakan citra buah salak pondoh dengan tingkat kematangan yang berbeda, yaitu mentah, setengah matang dan matang.

Pengambilan citra buah salak pondoh dilakukan dengan menggunakan studio buatan dengan warna latar putih. Citra sebanyak 39 buah didapat dengan menggunakan kamera *smartphone* yang menghasilkan citra sebesar 2128 x 4608 piksel. Penerangan tambahan didapat dari smartpnohe berupa lampu *flash* LED. Proses pengambilan dilakukan sebanyak empat kali, dengan posisi dan jarak objek ke kamera yang sama, yaitu di atas objek sejauh 25 cm dengan perbesaran 1.5 kali.

B. *Pelabelan dan Pemrosesan Citra*

Dalam penelitian ini, pelabelan citra terkait tingkat kematangannya dilakukan dengan merujuk kepada sumber literatur yang membahas kematangan buah salak pondoh. Dimana salak pondoh matang memiliki warna sisik dominan cokelat kekuningan, salak setengah matang bersisik kecoklatan dan salak mentah memiliki sisik rapat berwarna coklat kehitaman [5]. Citra buah salak pondoh berwarna diubah menjadi ukuran 500 x 750 piksel dan citra yang telah diubah ukurannya akan mengalami proses perataan (*smoothing*) menggunakan metode *denoising* dengan filter spasial. Penelitian ini menggunakan metode Gaussian *smoothing* untuk meningkatkan kualitas citra buah salak pondoh. Segmentasi citra buah salak pondoh dilakukan melalui proses pengambangan (*thresholding*) iteratif. Hasil dari langkah ini adalah citra biner, di mana buah salak pondoh muncul sebagai objek dengan piksel bernilai 0 (hitam), sementara latar belakangnya memiliki nilai piksel 1 (putih).



Gambar 1. Bagan alur proses klasifikasi

Operasi morfologi dilakukan untuk mendapatkan bentuk utuh dari buah salak pondoh. Operasi ini terdiri dari erosi, yang digunakan untuk menghilangkan noise putih pada citra, diikuti oleh dilasi, yang bertujuan untuk mengembalikan area objek yang mungkin terpengaruh oleh erosi. Tahap akhir dari pemrosesan citra adalah melakukan pemotongan citra sehingga hanya obyek yang tersisa. Untuk dapat melakukan langkah ini, perlu menentukan indeks piksel yang menjadi batas atau tepi citra yang mengandung obyek.

C. *Ekstraksi Fitur*

Ekstraksi fitur warna yang berupa besaran statistik orde pertama akan dilakukan pada citra yang sudah dipotong (*cropping)*. Pengubahan model warna juga dilakukan untuk mendapatkan besaran statistik dari nilai hue, saturasi dan value dengan mengubah citra RGB menjadi citra HSV. Besaran orde pertama statistik akan berupa besara *mean, variance, skewness* dan *kurtosis.* Penghitungan besaran akan dilakukan untuk warna Red, Green, Blue pada model warna RGB. Sedangkan untuk model warna HSV, akan dihitung besaran pada Hue, Saturation dan Value. Ekstraksi untuk masing-masing model warna akan menghasilkan 24 fitur untuk tiap citra.

D. *Klasifikasi*

Metode SVM yang akan digunakan di penelitian ini adalah pengembangan dari metode yang sudah ada, yaitu SVM multi kelas. Dengan metode ini, seluruh kelas untuk tingkat kematangan salak pondoh dapat dikenali. Proses pelatihan metode ini memiliki dua inputan, yaitu fitur citra dan pemilihan kernel. Terdapat tiga kernel yang digunakan pada penelitian ini, yaitu linear, polinomial dan Gaussian (RBF).

E. *K-Fold Cross Validation*

Validasi silang adalah proses berulang yang digunakan untuk menguji model pembelajaran mesin pada sampel data kecil. Metode ini memiliki satu parameter yang disebut k yang mengacu pada jumlah kelompok atau lipatan di mana sampel data tertentu dibagi. Dalam validasi silang k-fold, sampel data dibagi secara acak menjadi sebanyak k sampel dengan ukuran yang sama. Proses validasi silang dilakukan k secara iteratif, dengan menggunakan masing-masing k sampel tepat satu kali sebagai data validasi dan sisa kelompok sebagai data latih. Hasil prediksi untuk tiap k dapat dirata-ratakan untuk membuat perkiraan tunggal [10].

F. *Confusion Matriks*

Confusion Matrix adalah suatu proses yang digunakan untuk mengukur efektivitas model atau teknik klasifikasi mendalam pembelajaran mesin. Matriks konfusi menampilkan informasi numerik seberapa baik data tersebut diklasifikasikan (*True Positives dan True Negatives*) dan seberapa banyak data yang salah (*False Positives dan False Negatives*).

III. HASIL DAN PEMBAHASAN

A. *Akuisisi Data*

Data yang digunakan dalam penelitian ini terdiri dari citra buah salak pondoh dengan tiga tingkat kematangan yang berbeda, yaitu matang, setengah matang, dan mentah. Dari setiap buah salak pondoh, empat citra dihasilkan dengan berbagai metode pengambilan yang berbeda. Ini berarti bahwa

setiap citra menampilkan satu buah salak pondoh matang dengan empat variasi pengambilan yang berbeda, dan didapatkan total 156 citra asli.

Untuk meningkatkan variasi data, dilakukan proses augmentasi pada citra asli. Hal ini menghasilkan total 1.248 citra, dengan 416 citra untuk setiap tingkat kematangan. Augmentasi mencakup berbagai teknik, seperti membalik citra secara horizontal dan vertikal, membagi citra menjadi dua bagian, memotong bagian tengah citra, dan rotasi citra. Dengan demikian, augmentasi bertujuan untuk meningkatkan keragaman data dan memastikan bahwa dataset mencakup berbagai variasi yang mungkin terjadi dalam proses pemilihan buah salak pondoh.

(Ubah ke update data)

Pada penelitian ini, data latih dan data uji dibagi dengan perbandingan 80%:20%, didapatkan sejumlah 999 data latih dan 249 data uji. Data latih akan dibagi menjadi tiga kelompok data (fold) yang masing-masing berisikan 111 data untuk tiap tingkat kematangan. Untuk data uji, terdapat 83 data untuk tiap tingkat kematangan salak.

B. *Pemrosesan Citra*

Citra salak pondoh asli maupun hasil augmentasi akan diproses untuk menghilangkan noise serta memangkas latar pada citra. Luaran dari proses ini berupa citra salak pondoh utuh sehingga fitur yang dihasilkan dapat mudah dibedakan untuk masing-masing tingkat kematangan. Gambar 2 menampilkan citra salak pondoh matang di tiap tahap pemrosesannya.

- Perubahan Ukuran (*Resize*)

  Citra buah salak pondoh asli maupun hasil augmentasi, akan diubah ukurannya menjadi 500 x 750 piksel.
- Pengurangan Derau: *Gaussian smoothing*

  Pada penelitian ini, pengurangan derau dilakukan dengan fungsi dari *library open source*, yaitu OpenCv (*Open Source Computer Vision Library*). Fungsi cv.GaussianBlur() akan mengoperasikan kernel Gaussian pada citra salak pondoh dengan ukuran 9 x 9.
- Segmentasi: *Iterative Thresholding*

  Segmentasi dilakukan pada penelitian ini untuk dapat membedakan buah salak pondoh dengan latar. Proses ini tidak dapat dilakukan jika batas antara obyek dan latar tidak diketahui. Pengambangan iteratif digunakan untuk mencari batas tersebut, dimana citra inputan akan terlebih dahulu diubah menjadi citra abu. Nilai ambang awal diberikan dari hasil analisis histogram, yaitu 127. Dari nilai tersebut, akan terdapat dua bagian citra yang lebih kecil atau lebih besar ambang awal. Ambang diperbarui dengan menghitung rata-rata seluruh nilai piksel untuk dua kelompok.
- Morfologi: *Opening*

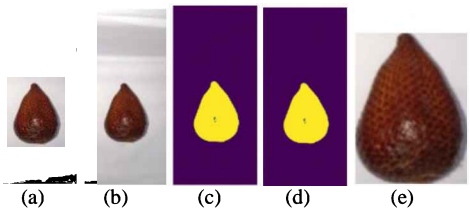
  Implementasi proses morfologi pada penelitian ini menggunakan fungsi cv2.morphologyEx(), dimana besar kernel untuk proses erosi dan dilasi, sebesar 3x3.
- Pemotongan Citra (*Cropping*)

  Pemotongan citra didasarkan pada hasil morfologi, dimana batas antara obyek dan latar sudah diperjelas. Batas tersebut nantinya akan mempermudah proses pemotongan dan menghasilkan citra utuh dari salak pondoh.

C. *Ekstraksi Fitur atau Ciri*

Citra salak pondoh yang sudah diproses merupakan citra berwarna dengan model RGB. Ekstraksi ciri dengan metode statistik orde pertama dengan besaran *mean*, *variance*, *skewness* dan *kurtosis*. Untuk mendapatkan ciri lainnya, citra salak pondoh akan diubah menjadi model warna HSV untuk selanjutnya dilakukan perhitungan besaran statistik orde pertama.

D. *Klasifikasi SVM*



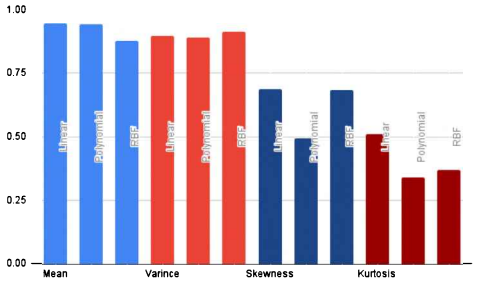
Gambar 2. Perubahan citra salak pondoh untuk tiap pemrosesannya: (a) citra hasil pengubahan ukuran, (b) citra setelah smoothing, (c) citra hasil segmentasi, (d) citra biner hasil morfologi, (e) citra hasil cropping

Hasil ekstraksi fitur akan digunakan untuk melakukan pengujian pada model SVM serta digunakan sebagai data uji akhir untuk menentukan performa model. Selanjutnya, model SVM dibangun dengan tiga kernel berbeda, yaitu kernel linear, kernel polynomial dan kernel Gaussian (RBF). Untuk memastikan bahwa pengujian model dilakukan dengan baik, maka data latih akan dibagi menjadi tiga fold, dengan jumlah data pada masing foldnya sebanyak 333 data. Pada tiap pengujian model, akan dilakukan iterasi untuk menentukan data latih dan data validasi. Maka dari itu, akurasi yang didapatkan merupakan akurasi rata-rata dari iterasi fold.

Pengujian model dibagi menjadi empat skenario: fitur orde pertama statistik, fitur model warna, fitur dari penelitian sebelumnya (mean Red dan mean Value), serta seluruh fitur.

- Pengujian model: fitur orde pertama statistik

  Gambar 3 menunjukkan grafik akurasi pengujian model dengan fitur besaran orde pertama statistika, yaitu *mean*, *variance*, *skewness* dan *kurtosis*

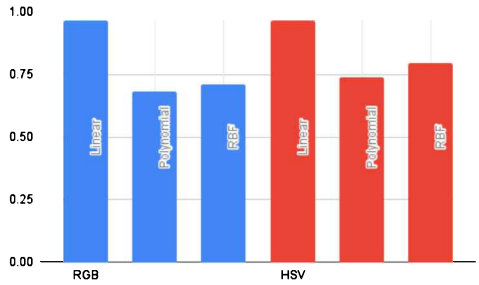


Gambar 3. Grafik rata-rata akurasi pengujian fitur besaran orde pertama statistik

Akurasi rata-rata tertinggi untuk pengujian model dengan fitur orde pertama statistik bernilai 94% pada besaran *mean* untuk kernel linear dan polynomial.

- Pengujian model: fitur model warna

Pada pengujian ini, fitur terbagi menjadi dua bagian, yaitu model warna RGB dan model warna HSV.

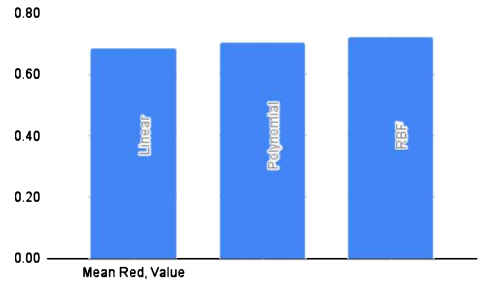


Gambar 4. Grafik rata-rata akurasi pengujian fitur model warna

Gambar 4 menunjukkan grafik rata-rata akurasi dari pengujian, dimana nilai terbesar didapat dari model dengan kernel linear dan fitur warna HSV sebesar 97%.

- Pengujian model: fitur mean Red dan mean Value

Pengujian pada penelitian sebelumnya, akurasi tertinggi didapat dengan menggunakan fitur berupa mean Red dari model warna RGB dan mean Value dari model warna HSV. Dengan tujuan yang sama, kedua besaran tersebut digunakan untuk menguji model dengan metrik evaluasi pada gambar 5. Didapatkan akurasi tertinggi sebesar 72% pada kernel RBF.



Gambar 5. Grafik rata-rata akurasi pengujian fitur mean Red dan mean Value

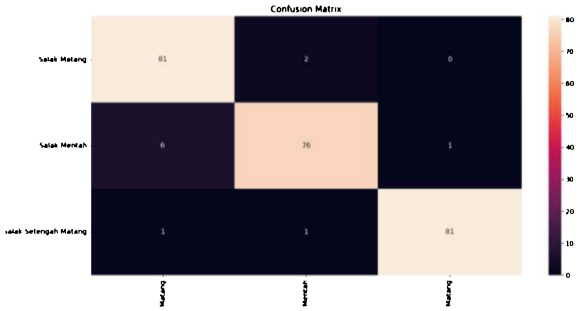
- Pengujian model: seluruh fitur

Pengujian model terakhir akan menggunakan gabungan dari fitur model warna dan besaran orde pertama statistic. **Error! Reference source not found.** menampilkan rata-rata akurasi untuk pengujian ini dan didapatkan akurasi tertingga untuk kernel linear sebesar 98%.

Gambar 6. Grafik rata-rata akurasi pengujian seluruh fitur

Dari seluruh fitur yang telah diujikan pada model, didapatkan akurasi model tertinggi untuk kernel linear dan fitur campuran atau gabungan antara besaran order pertama dan model warna sebesar 98%. Performa model ini kembali diuji dengan data yang berbeda dari proses pengujian sebelumnya. Grafik rata-rata akurasi pengujian seluruh fitur terdapat pada gambar 6.

Data uji berjumlah 249 data dengan jumlah data untuk masing-masing kelas berjumlah 83 data. Dari gambar 7 didapatkan confusion matriks untuk hasil prediksi model. Akurasi didapatkan dengan menjumlahkan seluruh hasil data uji yang berhasil diklasifikasikan dengan benar, yaitu 81 data kelas matang, 76 data kelas mentah dan 81 data kelas setengah matang. Total tersebut akan dibagi dengan total seluruh data uji sebanyak 249 data, sehingga didapatkan akurasi sebesar 96%.



Gambar 7. Confusion Matrix untuk hasil prediksi data uji

IV. KESIMPULAN

Pada penelitian ini, diajukan sistem klasifikasi kematangan buah salak pondoh menggunakan metode SVM. Dari hasil dan pembahasan, didapatkan bahwa penggunaan model warna untuk menentukan tingkat kematangan buah dapat menghasilkan akurasi prediksi yang baik. Augmentasi data dibutuhkan untuk menambah variasi data sehingga akurasi model dapat ditingkatkan. Penelitian ini juga menunjukkan bahwa metode SVM Multi kelas, khususnya dengan kernel linear, dapat digunakan sebagai mesin klasifikasi kematangan buah salak pondoh. Meskipun demikian, tidak menutup kemungkinan bahwa metode-metode lain atau fitur lain dapat digunakan dalam melakukan klasifikasi kematangan buah salak pondoh.

REFERENSI

[1] M. F. Dzulqarnain, S. Suprapto, dan F. Makhrus, "Improvement of Convolutional Neural Network Accuracy on Salak Classification Based Quality on Digital Image," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 2, hlm. 189, Apr 2019, doi: 10.22146/ijccs.42036.

[2] Rismiyati dan A. W. Helmie, "Snake Fruit Classification by Using Histogram of Oriented Gradient Feature and Extreme Learning Machine," 2019.

[3] K. Kualitas Buah Salak dengan Transfer Learning Arsitektur VGG dan A. Luthfiarta, "VGG16 Transfer Learning Architecture for Salak Fruit Quality Classification," *Jurnal Informatika dan Teknologi Informasi*, vol. 18, no. 1, hlm. 37–48, 2021, doi: 10.31515/telematika.v18i1.4025.

[4] A. Luthfiarta, "Transfer Learning with Xception Architecture for Snakefruit Quality Classification," 2022.

[5] P. Rianto dan A. Harjoko, "Penentuan Kematangan Buah Salak Pondoh Di Pohon Berbasis Pengolahan Citra Digital," *2017 IJCCS - Indonesian Journal of Computing and Cybernetics Systems*, vol. 11, no. 2, hlm. 143–154, 2017.

[6] N. Astrianda, "Klasifikasi Kematangan Buah Tomat Dengan Variasi Model Warna Menggunakan Support Vector Machine," *VOCATECH: Vocational Education and Technology Journal*, vol. 1, no. 2, hlm. 45–52, Apr 2020, doi: 10.38038/vocatech.v1i2.27.

[7] A. Septiarini, H. Hamdani, H. Rahmania Hatta, dan A. Ahmad Kasim, "Image-based processing for ripeness classification of oil palm fruit," *2019 5th International Conference on Science in Information Technology (ICSITech)*, hlm. 23–26, 2019.

[8] M. Arief, "Klasifikasi Kematangan Buah Jeruk Berdasarkan Fitur Warna Menggunakan Metode SVM," *Jurnal Ilmu Komputer dan Desain Komunikasi Visual*, vol. 4, no. 1, 2019.

[9] L. A. Wardani, G. Pasek, S. Wijaya, dan F. Bimantoro, "Klasifikasi Jenis Dan Tingkat Kematangan Buah Pepaya Berdasarkan Fitur Warna, Tekstur Dan Bentuk Menggunakan Support Vector Machine (Classification of Types and Levels of Ripeness of Papaya Fruit Based on Color, Texture and Shape Features Using Support Vector Machine)." [Daring]. Tersedia pada: http://jtika.if.unram.ac.id/index.php/JTIKA/

[10] D. Witten, G. M. James, T. Hastie, dan R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R.* . 2013.

Sistem Rekomendasi Indekos menggunakan Pendekatan *Content-Based Filtering*

Kayetanus Jo
Departemen Informatika
Universitas Sanata Dharma
Yogyakarta, Indonesia
jokayetanus@gmail.com

Robetus Adi Nugroho
Departemen Informatika
Universitas Sanata Dharma
Yogyakarta, Indonesia
robertusadi@usd.ac.id

Abstract*— Boarding House is a service that offers rooms or apartments for a certain period of time (usually for a monthly or annual fee). Each boarding house has its own specifications, starting from price, room type, room facilities, bathroom facilities and public facilities. Apart from different boarding house specifications, there are also differences in the preferences or needs of each boarding house service user. In an effort to solve the problems faced, the researcher built a recommendation system by utilizing the Content Based Filtering method to provide recommendations, apart from that the researcher also utilized the precision and reccal methods to calculate the accuracy of the system being built. The data used was taken from the mamikos.com website. The final result of this research is a Boarding House Recommendation System around has average precision value of 50%.

I. PENDAHULUAN

Indekos adalah jasa yang menawarkan kamar untuk jangka waktu tertentu biasanya dikenakan biaya bulanan atau tahunan Sebagian besar pengguna jasa indekos adalah perantau. Indekos merupakan suatu kebutuhan utama bagi para perantau. Perantau sendiri adalah orang yang meninggalkan tempat asal atau kelahirannya dengan tujuan tertentu, bisa untuk bekerja, menuntut ilmu atau hal lainnya. Salah satu contoh perantau adalah mahasiswa. Banyak mahasiswa yang berkuliah di luar kampung halamannya dan menjadikan indekos sebagai tempat tinggal sementara. Mahasiswa cenderung memilih tempat indekos yang dekat dengan kampus, hal tersebut dapat mempermudah mahasiswa dalam menyelesaikan urusannya yang berkaitan dengan kampus, seperti kuliah, bayar uang kuliah, mengerjakan tugas dan hal lainnya.

Dari banyaknya indekos yang berada di sekitar kampus tentunya setiap indekos tersebut mempunyai spesifikasi masing-masing, mulai dari harga, tipe kamar, fasilitas kamar tidur, fasilitas kamar mandi, fasilitas umum dan fasilitas parkir. Selain itu, pemilihan tempat indekos dipengaruhi oleh kesukaan atau kebutuhan oleh setiap penggunanya. Dilihat dari perbedaan tersebut tentunya pemilihan kos tidak semudah seperti yang dibayangkan. Banyaknya pilihan justru akan semakin membuat pemilihan indekos menjadi semakin sulit. Hal inilah yang ingin diselesaikan dalam penelitian ini.

Solusi yang ingin ditawarkan dalam penellitian ini adalah pembuatan sistem rekomendasi untuk memilih indekos bagi mahasiswa. Sistem rekomendasi adalah alat dan teknik perangkat lunak yang memberikan saran berupa item yang berguna bagi pengguna [2].Sistem rekomendasi merupakan suatu sistem yang memberikan rekomendasi produk atau barang kepada pengguna. Sistem rekomendasi ini sudah banyak diterapkan dalam kehidupan sehari-hari, seperti rekomendasi barang, rekomendasi film, mapun rekomendasi lagu [3]. Sistem rekomendasi membantu pengguna untuk menemukan produk atau barang yang sesuai dengan kebutuhan, kesukaan, maupun keinginan pengguna. Sistem rekomendasi memandu pengguna dalam menemukan produk yang cocok dari banyaknya produk yang tersedia [1].

Ada tiga metode yang sering digunakan dalam sistem rekomendasi antara lain *Content-Based Filtering, Collaborative Filtering* dan *hybrid.* Metode *Content-Based Filtering* adalah metode rekomendasi yang didasarkan pada kesamaan produk yang disukai pengguna di masa lalu dengan produk yang belum pernah digunakan oleh pengguna [3]. Metode *Collaborative Filtering (CF)* adalah metode rekomendasi yang didasarkan pada riwayat penilaian oleh pengguna. Metode ini berusaha mencari kesamaan kesukaan antar pengguna yang satu dengan yang lain berdasarkan riwayat penilaian terhadap produk – produk yang pernah digunakan. Metode CF dibagi menjadi dua jenis, yaitu *Item-Based Collaborative Filtering* dan *User-Based Collaborative Filtering*. Metode *hybrid* adalah metode rekomendasi yang dilakukan dengan menggabungkan metode *Content-Based Filtering* dan *Collaborative Filtering* untuk mendapatkan hasil rekomendasi yang lebih baik [1]. Tujuan dari penggabungan ini adalah untuk saling menutupi kekurangan masing-masing metode dengan menggunakan kelebihan dari kedua metode tersebut. Sehingga hasil rekomendasi yang diperoleh menjadi lebih baik [2].

Dalam penelitian-penelitian terdahulu, metode Content-based filetring sudah banyak digunakan. Wijaya dan Alfian (2018) mengimplementasikan metode *content-based filtering* yang digabungkan dengan metode *collaborative filtering* pada kasus sistem rekomendasi laptop. Penggabungkan secara *hybrid* antara metode *collaborative filtering* dan *content-based filtering* dapat menghasilkan sistem rekomendasi laptop yang mampu menutupi kekurangan dari setiap metode yang digunakan. Di dalam percobaannya metode *content-based filtering* memiliki waktu eksekusi lebih cepat dari metode *collaborative filtering* [4].

Fajriansyah dkk (2021) juga berhasil mengimplementasikan metode *content-based filtering* dalam merekomendasikan film. Penelitian ini mencari kemiripan bobot dari *term* pada *bag of words* hasil *pre-processing* sinopsis film dan judul film. Pembobotan dilakukan menggunakan metode *TF-IDF* yang telah dinormalisasi. Kemudian hasil pembobotan akan melalui tahap pencarian kemiripan berdasarkan bobot dan diakhiri dengan *filtering* berdasarkan genre. Pengujian dilakukan dengan melibatkan tiga partisipan dengan total jumlah film sebanyak 4000 judul film. Dari pengujian tersebut didapatkan nilai akurasi (*mean average precision)* sebesar 0.823254 untuk jenis *single query* dan 0.7500556 untuk jenis kueri *multiple seeds query*. Dari

hasil tersebut didapatkan untuk jenis *single query* menghasilkan rekomendasi yang lebih baik daripada jenis *multiple seeds query* [5].

Sementara itu Mngomezulu dan Ajoodha (2022) berhasil menerapkan metode *content-based filtering* dalam merekomendasikan film menggunakan *keywords extractions*. *Content-based filtering* digunakan sebagai model utama dalam penelitian ini, dengan *Term Frequency - Inverse Document Frequency* (TF-IDF) dan *keywords extractions* yang digunakan sebagai kata kunci ekstraktor. Dalam penelitian ini terdapat 244 film yang direkomendasikan untuk digunakan kata kunci dari masing-masing ekstraktor, dengan nilai rata-rata tertinggi sebesar 33% dari film yang direkomendasikan dari masing-masing film identik [6].

Reswara dkk (2023) berhasil menerapkan metode *BERT* dan *cosine similarity* dalam merekomendasikan anim. Sistem ini dirancang untuk memberikan banyak saran judul anim berdasarkan kesamaan mereka. Atribut yang digunakan dalam penelitian ini adalah judul dan *genre* anim, kedua atribut ini digunakan untuk membandingkan kemiripan satu anim dengan anim yang lain. Judul dan *genre* anim akan ditampilkan menggunakan metode *BERT* dan dibandingkan dengan metode *cosine similarity* [7].

Haviana dkk (2023) berhasil menerapkan metode *content-based filtering* dalam studi kasus rekomendasi tempat publikasi jurnal. Sistem rekomendasi ini memberikan tempat yang cocok untuk mempublikasikan jurnal. Hasil akhir dalam penelitian ini mendapat nilai *precision* sebesar 95% dan dapat menyarankan jurnal yang baik dan relavan bagi para peneliti [8].

Hunna dkk (2022) berhasil menerapkan metode *content-based filtering* dalam studi kasus merekomendasikan referensi penelitian yang berkaitan dengan *data mining*. Sistem rekomendasi ini dibuat untuk membantu mahasiswa menemukan referensi dalam pengerjaan tugas akhir. Penelitian ini juga menggunakan algoritma *TF-IDF* untuk menemukan ketersediaan konten yang ada. Sistem pengujian dalam penelitian ini menggunakan *confusion matrix*. Hasil akhir dalam penelitian menadpatkan akurasi sebesar 78% [9].

Anthony dkk (2022) juga berhasil menerapkan metode *content-based filtering* dalam memberikan rekomendasi musik di spotify. Sistem ini memberikan rekomendasi berdasarkan *playlist* pengguna. Hasil akhir dalam penelitian ini berupa perbandingan atara sistem rekomendasi yang dibuat dengan rekomendasi pada spotify, dari hasil perbandingan tersebut diperoleh *cosine similarity* sebesar 80% untuk lagu dan 50 % untuk artisnya [10].

Dari penelitian – penelitian tersebut dapat dilihat dengan jelas penggunaan metode *Content-based Filtering* dalam sistem rekomendasi dapat memberikan hasil yang baik. Oleh karena itu, penelitian ini mencoba menggunakan metode *Content-based Filtering* untuk sistem rekomendasi indekos.

II. METODE PENELITIAN

A. Data

Data yang digunakan dalam penelitian ini merupakan data indekos yang diambil dari website mamikos.com. Adapun jumlah indekos yang diambil adalah 80 indekos di Yogyakarta.

B. Atribut Data

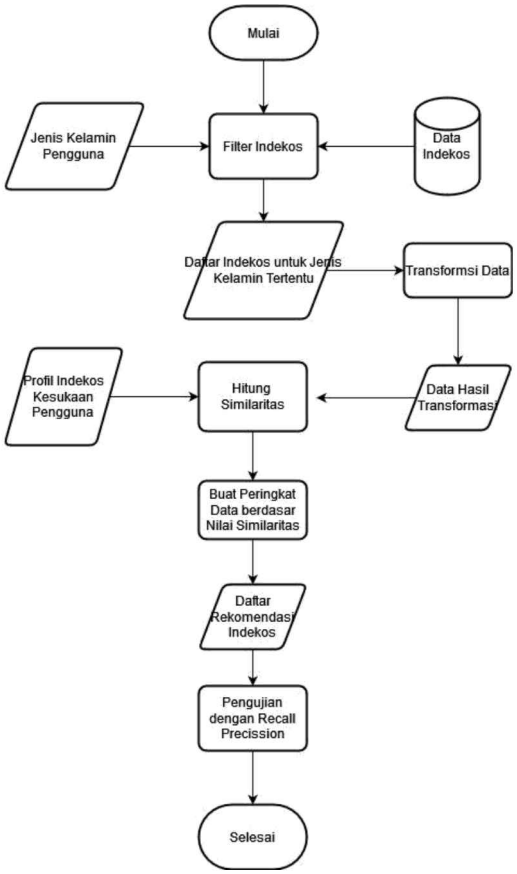
Atribut data yang digunakan dalam penelitian ini dapat di lihat pada tabel 1.

TABEL 1. ATTRIBUT DATA

Atribut	Deskripsi	Keterangan
Id	(int)	ID kos
nama	(string)	Nama kos
harga	(double)	Harga kos
pengguna	(string)	Menerangkan bahwa kos tersebut untuk pria, wanita atau bebas
alamat	(string)	Alamat kos
termasuk_listrik	(string)	Listrik termasuk dalam biaya kos atau tidak
ac	(string)	Ketersediaan fasilitas *ac*
meja	(string)	Ketersediaan fasiltas meja
ventilasi	(string)	Ketersediaan fasilitas ventilasi
cermin	(string)	Ketersediaan fasilitas cermin
jendela	(string)	Ketersediaan fasilitas jendela
kasur	(string)	Ketersediaan fasilitas kasur
bantal	(string)	Ketersediaan fasilitas bantal
guling	(string)	Ketersediaan fasilitas guling
lemari	(string)	Ketersediaan fasilitas lemari
kursi	(string)	Ketersediaan fasilitas kursi
kloset_duduk	(string)	Ketersediaan fasilitas kloset duduk
shower	(string)	Ketersediaan fasilitas *shower*
kamar_mandi	(string)	Ketersediaan fasilitas kamar mandi
wifi	(string)	Ketersediaan fasilitas *wifi*
dapur	(string)	Ketersediaan fasilitas dapur
kulkas	(string)	Ketersediaan fasilitas kulkas
cctv	(string)	Ketersediaan fasilitas *cctv*
parkir_mobil	(string)	Ketersediaan fasilitas parkir mobil
parkir_motor	(string)	Ketersediaan fasilitas parkir motor
parkir_sepeda	(string)	Ketersediaan fasilitas parkir sepeda

C. Alur Penelitian

Sistem pertama – tama akan memfilter daftar indekos berdasarkan jenis kelamin pengguna. Setelah itu sistem akan melakukan transformasi data (lihat tabel 2). Bentuk numerik hasil transformasi ini akan memudahkan proses perhitungan. Selanjutnya sistem akan menghitung tingkat kesamaan (*similarity*) profil indekos yang disukai oleh pengguna dengan data indekos yang sudah mengalami transformasi data (lihat tabel 2). Persamaan untuk menghitung *similarity* dengan *cosine similarity* dapat dilihat pada persamaan 1. Diagram alur penelitian dapat dilihat pada gambar 1.



Gambar 1. Alur Penelitian

$$sim\ (a,b) = \frac{\sum_{i=1}^{n} Ai\ x\ Bi}{\sqrt{\sum_{i=1}^{n} A_i^2}\ x\ \sqrt{\sum_{i=1}^{n} B_i^2}} \qquad (1)$$

Dimana,
a = Indekos A
b = Indekos B
A = Vektor data Indekos A
B = Vektor data Indekos B

Nilai similaritas atau kemiripan yang dihasilkan akan diurutkan seperti tabel 3.

TABEL 2. DAFTAR TRANSFORMASI DATA

Atribut	Deskripsi	Keterangan
harga	harga <= 900k	0
	900k > harga < 1500000k	1
	harga > 1500000	2
termasuk_listrik	tidak	0
	iya	1
ac	tidak	0
	ada	1
meja	tidak	0
	ada	1
ventilasi	tidak	0
	ada	1
cermin	tidak	0
	ada	1
jendela	tidak	0
	ada	1
kasur	tidak	0
	ada	1
bantal	tidak	0
	ada	1
guling	tidak	0
	ada	1
lemari	tidak	0
	ada	1
kursi	tidak	0
	ada	1
kloset_duduk	tidak	0
	iya	1
shower	tidak	0
	iya	1
kamar_mandi	luar	0
	Dalam	1
wifi	tidak	0
	ada	1
dapur	tidak	0
	ada	1
kulkas	tidak	0
	ada	1

TABEL 3. DATA URUT NILAI SIMILARITAS

Nama Kos	Nilai Kemiripan
Kost Sare Pules Tipe Standart Mergangsan Yogyakarta	0.6837634587578276
Kost Ibu Bidan Tipe A Wirobrajan Yogyakarta	0.5345224838248487
Kost Rizki Tipe C Gedong Tengen Yogyakarta	0.5241424183609592
Kost Budi Tipe C Jetis Yogyakarta	0.5241424183609592
Kost Blue House Tipe A Danurejan Yogyakarta	0.5039526306789696
Kost Innepan D'Ambar Tipe A Code Gondomanan Yogyakarta	0.5039526306789696
Kost Bapak Yudi Tipe A Tegalrejo Yogyakarta	0.4558423058385518
Kost Kuncen Fan Malioboro Wirobrajan Yogyakarta	0.40406101782088427
Kost Kuncen Tipe Standart Malioboro Wirobrajan Yogyakarta	0.3903600291794132
Kost Mama O Umbulharjo Yogyakarta	0.3779644730092272
Kost Aldebaran Tipe Premium Tegalrejo Yogyakarta	0.3666793988112845
Kost Jembatan Sardjito Jetis Yogyakarta	0.30304576336566325
Kost Omahe Qoema Syariah 1 Tipe A Ngampilan Yogyakarta	0.2834733547569204

Lalu, sistem akan memilih data yang memiliki nilai similaritas lebih besar atas sama dengan 0.5 untuk direkomendasikan kepada pengguna (lihat tabel 4).

TABEL 4. DATA DENGAN NILAI SIMILARITAS SAMA DENGAN ATAU LEBIH DARI 0.5

Nama Kos	Nilai Kemiripan
Kost Sare Pules Tipe Standart Mergangsan Yogyakarta	0.6837634587578276
Kost Ibu Bidan Tipe A Wirobrajan Yogyakarta	0.5345224838248487
Kost Rizki Tipe C Gedong Tengen Yogyakarta	0.5241424183609592
Kost Budi Tipe C Jetis Yogyakarta	0.5241424183609592
Kost Blue House Tipe A Danurejan Yogyakarta	0.5039526306789696
Kost Innepan D'Ambar Tipe A Code Gondomanan Yogyakarta	0.5039526306789696

III. PENGUJIAN DAN PEMBAHASAN

Berdasarkan data yang diberikan oleh pengguna, didapatkan himpunan R sebagai data indekos yang benar – benar disukai oleh pengguna dari seluruh indekos yang ada di database sehingga R = { Kost Ibu Bidan Tipe A Wirobrajan Yogyakarta, Kost Budi Tipe C Jetis Yogyakarta, Kost Mama O Umbulharjo Yogyakarta, Kost Jembatan Sardjito Jetis Yogyakarta}. Dengan membandingkan hasil rekmendasi yang diberikan oleh sistem dan himpunan R maka didapatkan nilai *Recall* dan *Precission* seperti yang terlihat pada tabel 5.

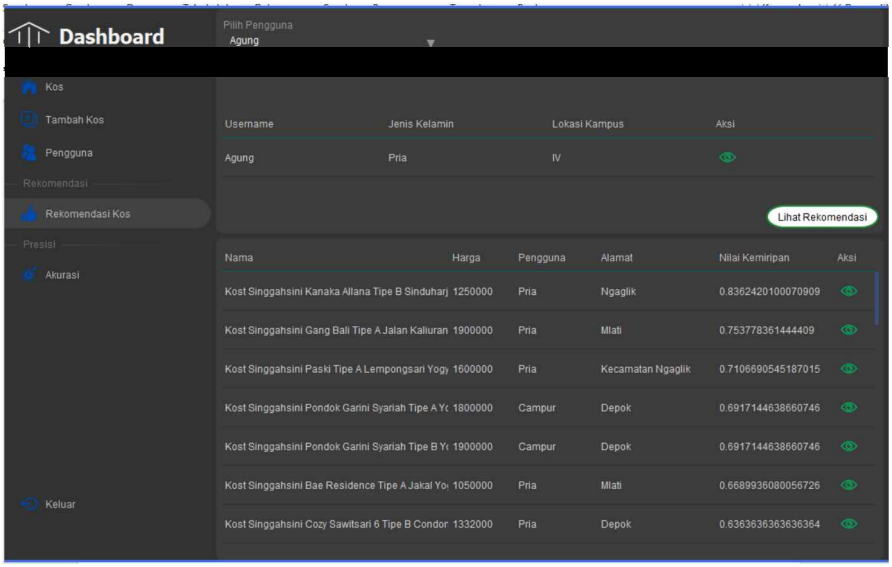
Dari perhitungan *Recall Precission* pada tabel 5, dapat dihitung nilai *Average Precission* = 0,50.

$$Average\ Precision = \frac{0.50 + 0.50}{2} = 0.50$$

Contoh tampilan program dari hasil implementasi Sistem Rekomendasi dengan pendekatan *Content-based Filtering* dapat dilihat pada gambar 2.

TABEL 5. HITUNG RECALL PRECISSION

Peringkat	Nama Kos	Reccal	Presisi
1	Kost Sare Pules Tipe Standart Mergangsan Yogyakarta	0/4 = 0	0/1=0
2	Kost Ibu Bidan Tipe A Wirobrajan Yogyakarta	1/4 = 0,25	1/2 = 0.50
3	Kost Rizki Tipe C Gedong Tengen Yogyakarta	1/4 = 0,25	1/3 = 0,333
4	Kost Budi Tipe C Jetis Yogyakarta	2/4 = 0,50	2/4 = e0.50
5	Kost Blue House Tipe A Danurejan Yogyakarta	2/4 = 0,50	2/5 = 0,40
6	Kost Innepan D'Ambar Tipe A Code Gondomanan Yogyakarta	2/4 = 0,50	2/6 = 0,333



Gambar 2. Tampilan GUI Sistem Rekomendasi

IV. PENUTUP

Berdasarkan penelitian yang dilakukan, peneliti berhasil menerapkan metode *content-based filtering* pada sistem rekomendasi indekos. Penggunaan metode *content-based filtering* ini terbukti efektif dengan didapatnya nilai *average pracision* sebesar 50%, hal ini menunjukkan bahwa metode *content-based filtering* mempunyai potensi dalam memberikan rekomendasi indekos yang relevan. Meskipun demikian, masih ada ruang untuk peningkatan, terutama dalam meningkatkan akurasi dan personalisasi rekomendasi.

Berdasarkan penelitian yang telah dilakukan, penulis menyarankan untuk penelitian selanjutnya, dapat menggabungkan metode *content-based filtering* dengan metode *collaborative filtering*. Hal ini dilakukan untuk meningkatkan performa sistem. Selain itu penulis juga berharap, pada penelitian selanjutnya dapat menggunakan data set yang lebih beragam dan komprehensif sehingga dapat membantu dalam menggeneralisasi hasil penelitian. Penelitian ini berkontribusi pada pemahaman mengenai penerapan sistem rekomendasi indekos atau tempat tinggal dan menawarkan wawasan berharga bagi pengembang aplikasi, pengguna aplikasi dan peneliti dalam bidang ini.

REFERENSI

[1] Prasetya, C. S. D. (2017). Sistem Rekomendasi Pada E-Commerce Menggunakan K-Nearest Neighbor. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) p-ISSN, 2355*, 7699.

[2] Ricci, F., Rokach, L., & Shapira, B. (2010). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1-35). Boston, MA: springer US.

[3] Ritdrix, A. H., & Wirawan, P. W. (2018). *Sistem Rekomendasi Buku Menggunakan Metode Item-Based Collaborative Filtering* (Doctoral dissertation, Universitas Diponegoro).

[4] Wijaya, A. E., & Alfian, D. (2018). Sistem Rekomendasi Laptop Menggunakan Collaborative Filtering Dan Content-Based Filtering. *Jurnal Computech & Bisnis*, *12*(1), 11-27.

[5] Fajriansyah, M., Adikara, P. P., & Widodo, A. W. (2021). Sistem Rekomendasi Film Menggunakan Content Based Filtering. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *5*(6), 2188-2199.

[6] Mngomezulu, M., & Ajoodha, R. (2022, October). A Content-Based Collaborative Filtering Movie Recommendation System using Keywords Extractions. In *2022 International Conference on Engineering and Emerging Technologies (ICEET)* (pp. 1-6). IEEE.

[7] Reswara, C. G., Nicolas, J., Ananta, M., & Kurniadi, F. I. (2023, September). Anime Recommendation System Using Bert and Cosine Similarity. In *2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS)* (pp. 109-113). IEEE.

[8] Haviana, S. F. C., Mulyono, S., Subroto, I. M. I., Sulaiman, N. S., & Yacob, A. (2023, September). Exploiting Content-Based Filtering for Publication Venue Recommendations. In *2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)* (pp. 600-605). IEEE.

[9] Hunna, K. N. M., Renaldi, F., & Santikarama, I. (2022, February). Paper Recommendation for Research References in Data Mining using Content-Based Filtering. In *2022 International Conference on Science and Technology (ICOSTECH)* (pp. 1-6). IEEE.

[10] Anthony, J. T., Christian, G. E., Evanlim, V., Lucky, H., & Suhartono, D. (2022, October). The Utilization of Content Based Filtering for Spotify Music Recommendation. In *2022 International Conference on Informatics Electrical and Electronics (ICIEE)* (pp. 1-4). IEEE.

Klasifikasi Keluarga Miskin menggunakan Algoritma C4.5 dan *Support Vector Machine*

Maria Ina Maram
Fakultas Sains dan Teknologi
Universitas Sanata Dharma
Yogyakarta, Indonesia
inamaran910@gmail.com

Ridowati Gunwan
Fakultas Sains dan Teknologi
Universitas Sanata Dharma
Yogyakarta, Indoensia
rido@usd.ac.id

Abstrak—Salah satu permasalahan social yang terjadi di Indonesia adalah penanganan kemiskinan. Setiap daerah di Indonesia tentunya memiliki penduduk yang berstatus miskin. Sebelum merancang program yang tepat untuk penanganan kemisikinan, perlu terlebih dahulu mendeteksi keluarga mana yang berstatus miskin. Pemerintah di salah satu desa yaitu Kenotan di Nusa Tenggara Timur mengalami kesulitan dalam memetakan penduduk miskinnya, oleh karenanya perlu adanya upaya yang dilakukan yaitu mengelompokan penduduk miskin dan yang tidak. Salah satu metode yang dapat digunakan untuk mengelompokan penduduk miskin adalah algoritma klasifikasi yaitu C4.5 dan Support Vector Machines (SVM). Tujuan dari penelitian ini adalah melakukan klasifikasi pendudukan miskin menggunakan algoritma C4.5 dan SVM, membandingkan tingkat akurasi kedua algoritma tersebut dan mendapatkan faktor apa saja yang mempengaruhi pengelompkan penduduk miskin. Manfaat penelitian ini adalah mendapatkan model pengelompokan terbaik yang dapat digunakan untuk mengelompokan penduduk miskin di desa Kenotan. Hasil penelitian ini mendapatkan tingkat akurasi tertinggi pada model C4.5 dengan akurasi 94.35% dan atribut yang paling berpengaruhi terhadap hasil pengelompokan adalah pekerjaan, pendapatan setiap bulan, jumlah tanggungan, pendidikan terakhir dan umur.

Kata kunci—klasifikasi, penduduk miskin, C4.5, support vector machines, akurasi.

I. PENDAHULUAN

Kemiskinan. merupakan kondisi dimana seseorang atau sekelompok orang tidak mampu memenuhi hak-hal dasarnya untuk mempertahankan dan mengembangkan kehidupan yang bermartabat. Badan Pusat Statistik (BPS) menggunakan konsep kemampuan memenuhi kebutuhan dasar (*basic needs approach*) untuk mengukur kemiskinan. Kemiskinan dipandang sebagai ketidakmampuan dari sisi ekonomi untuk memenuhi kebutuhan dasar makanan dan bukan makanan yang diukur dari sisi pengeluaran. Pendudukan dikategorikan sebagai penduduk miskin jika memiliki rata-rata pengeluaran per kapita per bulan di bawah garis kemiskinan [1]

Pada Maret tahun 2022, BPS mencatat jumlah penduduk miskin di Indonesia mencapai 9.54% dari total penduduk Indonesia. Prosentaase tersebut mengalami penurunan sebesar 0.17% dibandingkan data pada September 2021. Selain data secara keseluruhan di Indonesia, BPS juga mencatat provinsi Nusa Tenggara Timur merupakan provinsi pada urutan ke-3 terbanyak jumlah penduduk miskinnya. Desa Kenotan yang terdiri dari 4 dusun merupakan satu dari 12 desa dan kelurahan yang berada di kecamatan Adonara Tengah kabupaten Flores Timur, memiliki penduduk yang masih di bawah tarif hidupnya. Anak-anak putus sekolah, banyaknya pengangguran dan pekerja serabutan merupakan permasalahan social yang dihadapi di desa Kenotan. Berbagai upaya untuk mengatasi masalah sosial ini sudah dilakukan oleh pemerintah yang kemudian diimplementasikan dalam bentuk kebijakan dan program-program baik yang bersifat langsung maupun tidak langsung seperti Bantuan Langsung Tunai (BLT), Program Jaminan Kesehatan Nasional (JKN-KIS), program Jamkesmas, program IDT, Program Keluarga Harapan (PKH) dan program bantuan social pemerintah lainnya.

Meskipun berbagai upaya telah dilakukan, namun kemiskinan tidak dapat dihilangkan sepenuhnya. Program kemiskinan yang saat ini dilakukan baik yang berasal dari pemerintah maupun non pemerintah umumnya hanya sementara artinya program tersebut akan berjalan selama masih ada anggaran dana setelah dana habis maka selesai juga kegiatan program. Oleh karenanya perlu tersedia data berkaitan dengan kemiskinan untuk menunjang keberhasilan program yang telah direncanakan. Agar program yang telah direncanakan dapat tepat sasaran dan program yang dirancang dapat sesuai dengan permasalahan pada daerah tertentu. Data penduduk miskin yang belum tertata dengan baik menjadi salah satu sebab program pengentasan kemiskinan kurang tepat sasaran. Melihat persoalan tersebut, perlu adanya pengolahan data untuk mengelompokan penduduk yang tergolong miskin agar bantuan-bantuan yang telah dirancang dapat tersalurkan dengan merata berdasarkan data hasil pengelompokan.

Seiring bertambahnya jumlah penduduk maka semakin banyak data yang harus diperhatikan, hal tersebut terjadi juga pada desa Kenotan. Akan tetapi pengolahan data terutama pengelompokan data status keluarga miskin masih dilakukan tidak menggunakan suatu metode atau dengan kata lain masih dilakukan secara manual. Perlu dilakukan cara-cara agar pengelompokan data yang dilakukan lebih cepat dan hasil yang diperoleh akurat. Salah satu yang dapat dilakukan adalah penggunaan metode penambangan data untuk membuat model pengelompokan dan memanfaatkan model untuk memprediksi keluarga yang berstatus miskin atau tidak.

Berbagai penelitian telah digunakan berhubungan dengan penduduk miskin, Pristiwati dkk. melakukan penelitian tentang pengelompokan penerima bantuan beras miskin menggunakan perbandingan metode *K-Nearest Neighbor,* Naïve Bayesian dan C4.5. Jumlah data yang digunakan sebanyak 585 data dengan 33 atribut. Akurasi sebesar 79.00% diperoleh dari hasil pengujian *K-Nearest Neighbor* dengan jumlah tetangga terdekat sebanyak 21, metode Naïve Bayesian memperoleh akurasi 84.10%, sedangkan penggunaan metode C4.5 diperoleh akurasi sebesar 88.36% [2].

Kurnia dkk menggunakan metode *K-Nearest Neighbor* untuk melakukan klasifikasi keluarga miskin dari 100 data

yang dibagi menjadi 4 data latih dan 1 data uji. Atribut yang digunakan adalah no kartu keluarga, jumlah anggota keluarga, pekerjaan kepala keluarga, penghasilan perbulan. Percobaan yang dilakukan menggunakan nilai tetangga terdekat sebanyak 5, 7 dan 9. Akurasi tertinggi diperoleh 90% menggunakan perbandingan 90:10 antra data pelatihan dan data pengujian[3].

Penelitian lain adalah kelayakan penerima bantuan program keluarga harapan (PKH) menggunakan algoritma C4.5 dan Naïve Bayesian. Akurasi yang diperoleh 91.25% menggunakan C4.5, sementara Naïve Bayesian 87.11%, artinya klasifikasi untuk PKH menggunakan C4.5 memiliki tingkat akurasi tinggi [4]

Kasim dan Sudarsono menggunakan metode SVM untuk klasifikasi ekonomi penduduk penerima bantuan. Pengujian dilakukan dengan menggunakan atribut pendidikan terakhir dinding rumah atap rumah luas rumah dan sumber listrik dengan membagi dataset menjadi 80% data latin dan 20% data uji dari pengujian yang dilakukan didapatkan akurasi sebesar 98%. [5]

Penelitian juga dilakukan oleh Qardini dkk dengan membandingkan algoritma C4.5 dan Adaboost pada klasifikasi penerima program bantuan sosial dengan selisih akurasi sebesar 1% yang mana akurasi terbaik didapat pada metode Adaboost yakni 95% sedangkan akurasi dari sistem sebesar 94% [6]

Agustina dkk juga melakukan penelitian menggunakan SVM untuk klasfikasi rumah layak huni dengan akurasi yang didapat sebesar 98.75% [7]

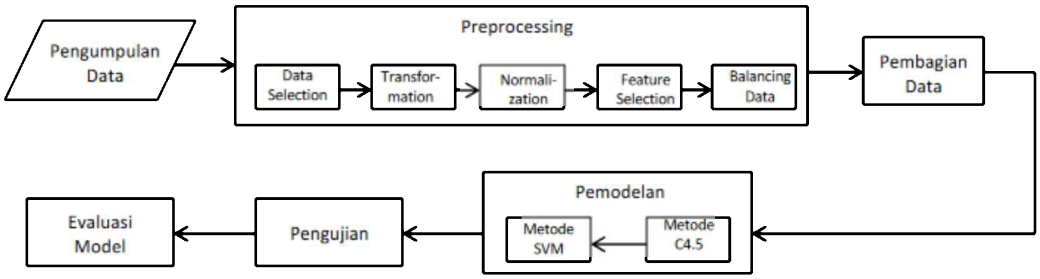
Berdasarkan latar belakang dan berbagai penelitian sebelumnya, terlihat bahwa penanganan keluarga miskin perlu dilakukan dengan tepat menggunakan metode pengelompokan yang sesuai. Algoritma C4.5 dan Support Vector Machine (SVM) dipilih karena dari penelitian sebelumnya menghasilkan akurasi yang tinggi. Selain pengembangan model juga dilakukan proses prediksi berdasarkan masukan dari pengguna menggunakan metode terbaik untuk menentukan termasuk ke dalam keluarga miskin atau bukan. Pemerintah desa Kenotan dapat memfaatkan hasil penelitian ini untuk memprediksi status keluarga di desanya apakah termasuk kedalam status miskin atau tidak.

II. METODE PENELITIAN

Pada bagian ini akan dijelaskan tentang metode penelitian secara umum, bagaimana melakukan pemodelan dan melakukan pengujian. Selanjutnya juga dijelaskan setiap langkah dalam penelitian ini.

A. Gambaran Umum Penelitian

Pada penelitian ini klasifikasi penduduk miskin menggunakan metode C4.5 dan SVM. Tujuan untuk memperlihatkan model klasifikasi dapat memberikan solusi untuk melakukan klasifikasi berdasarkan atribut yang dimiliki. tujuan untuk memperlihatkan bagaimana suatu model klasifikasi data mining dapat memberikan solusi untuk melakukan klasifikasi tingkat kemiskinan berdasarkan atribut yang ada. Tahapan penelitian dimulai dari proses pengumpulan data keluarga di desa Kenotan, dilanjutkan dengan tahapan *preprocessing* yang terdiri dari pemilihan data, transformasi data, normalisasai, pemilihan fitur dan melakukan proses penyeimbangan data. Data yang sudah melewati tahap preprocessing, dibagi kedalam dua kelompok yaitu data yang digunakan untuk membangun model atau data latih dan data untuk menguji model yaitu data uji. Langkah selanjutnya adalah melakukan pemodelan klasfikasi menggunakan algoritma C4.5 dan SVM. Proses selanjutnya adalah melakukan pengujian terhadap kedua model menggunakan data latih dan tahap terakhir adalah melakukan evaluasi model menggunakan data uji. Hasil evaluasi akan mendapatkan model mana yang memberikan hasil akurasi terbaik, model inilah yang nantinya digunakan untuk prediksi. Tahapan penelitian dapat dilihat pada gambar 1.



Gambar 1. Langkah Penelitian

B. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data penduduk yang diperoleh dari desa Kenotan kecamatan Adonara Tengah kabupaten Flores Timur. Data penduduk yang digunakan adalah data penduduk tahun 2022 dalam bentuk file Excel. Data yang diperoleh mempunyai beberapa tribut yakni Jenis Kelamin, Tempat Tanggal Lahir, Umur, Pekerjaan, Pendapatan / Bulan, Tanggungan, Pendidikan Terakhir, NIK, dan KK. Atribut data dapat dilihat pada Tabel 1.

TABEL 1. ATRIBUT DATA

No	Nama Atribut	Tipe Data
1	Jenis Kelamin	Kategorikal
2	Tempat Tanggal Lahir	Date
3	Umur	Numerik
4	Pekerjaan	Kategorikal
5	Pendapatan / Bulan	Numerik
6	Pendidikan Terakhir	Kategorikal
7	Tanggungan	Numerik
8	NIK	Numerik
9	KK	Numerik

C. Pra Pemrosesan Data

Tahap ini merupakan tahap untuk mengubah data awal agar dapat lebih dipahami sebelum digunakan untuk membangun model.

1) Data Selection: bertujuan untuk memilih atribut yang tidak berpengaruh dalam proses pembuatan model. Atribut yang digunakan adalah umur, pekerjaan, jenis kelamin, pendapatan/bulan, tanggungan dan pendidikan terakhir. Sedangkan atribut tempat tanggal lahir, NIK dan KK tidak digunakan karena merupakan atribut yang memiliki nilai unik.

2) Tranformation: Mengubah nilai atribut menjadi bernilai numerik kategorikal. Atribut yang diubah yaitu:

a) Jenis kelamin. Mengubah nilai atribut L menjadi 1 dan P menjadi 2.

b) Pekerjaan. Mengubah setiap nilai atribut menjadi angka.

c) Pendidikan terakhir. Mengubah nilai atribut pendidikan menjadi nilai numerik. Untuk SD menjadi 1, SMP = 2, SMA=3, Diploma=4 dan Sarjana=5.

d) Status penduduk. Nilai variabel Miskin diubah menjadi 1 dan tidak miskin = 2.

3) Feature selection. Atribut yang digunakan untuk proses *featur selection* adalah jenis kelamin, tempat tanggal lahir, umur, pekerjaan, pendapatan/bulan, tanggungan dan pendidikan terakhir. Teknik *feature selection* yang digunakan adalah model perangkingan menggunakan *information gain.*

4) Normalization. Agar nilai variabel memiliki batas atas dan batas bawah dalam batas yang sama. Metode yang digunakan *min max*, batas bawah = 0 dan batas atas=1.

5) Balancing data. Untuk mendapatkan nilai dari variabel kelas seimbang antara mikin dan tidak miskin. Teknik yang digunakan adalah *synthetic majority oversampling technique* (SMOTE).

D. Pembagian Data. Membagi data menjadi data latih dan data uji. Teknik yang digunakan adalah *k-fold cross validation*. Nilai k-fold yang dipilih adalah 3, 5 dan 10.

E. Pemodelan

1) Pemodelan C4.5

Setelah proses pembagian data, data latih digunakan untuk membangun model. Model C4.5 menggunakan algoritma C4.5 yang merupakan pengembangan dari algoritma ID3 (Iterative Dichotomiser) yang pertama kali dikemukakan oleh J Ross Quinlan [7]. Pohon keputusan dibangun dengan cara membagi data secara rekursif sehinga tiap bagian terdiri dari data yang berasal dari kelas yang sama. Secara umum algoritma C4.5 sebagai berikut:

a) Menentukan atribut yang menjadi akar.

Untuk menentukan atribut akar, terlebih dahulu dihitung nilai entropy. Entropy merupakan ukuran keberagaman. Entropy merupakan ukuran keberagaman. Makin tinggi nilai entropy maka nilai dari atribut semakin beragam. Persamaan (1) merupakan rumus Entropy yang digunakan.

$$Entropy\ (S) = \sum_{i=1}^{n} -p_i * log_2 p_i \quad (1)$$

Di mana:
S: Himpunan kasus
n: Jumlah partisi S
p_i: Proporsi dari S_i terhadap S.

Selanjutnya dihitung information Gain. Informastion Gain dihitung dengan menggunakan (2).

$$Gain(S,A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\ (S_i) \quad (2)$$

Dimana:
S: Himpunan kasus
A: Atribut
n: Jumlah Partisi Atribut A
|Si|: Jumlah Kasus pada partisi ke-i
|S|: Jumlah kasus dalam S
p_i : Proporsi dari S_i terhadap S.

Jika algoritma yang digunakan adalah ID3 maka atribut yang memiliki *Information Gain* tertinggi akan dijadikan sebagai atribut akar. Karena C4.5 merupakan pengembangan dari ID3, maka proses pencarian atribut akar dilanjutkan dengan mencari Gain Ratio tertinggi yang diperoleh dari (3).

$$GainRatio(A,S) = \frac{Gain(A,S)}{SplitInfo(A,S)} \quad (3)$$

Persamaan (4) merupakan persamaan yang digunakan untuk mendapatkan Split Information. Sedangkan Gain(A, S) menggunakan (2).

$$SplitInfo(A,S) = -\sum \frac{S_i}{S} log_2 \frac{S_i}{S} \quad (4)$$

b) Membuat cabang untuk tiap-tiap kasus.

Setelah diperoleh atribut akar, Langkah selanjutnya adalah membuat cabang-cabang untuk setiap nilai dari atributnya (kasus).

c) Membagi kasus di dalam cabang.

Setiap nilai kasus (nilai aktribut) dibagi / dimasukan ke dalam cabang-cabang yang telah diperoleh.

d) Ulangi proses b dan c.

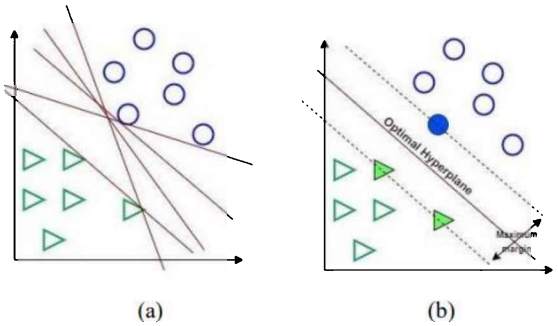
Lakukan perulangan untuk setiap cabang, sampai semua kasus dalam cabang memiliki kelas sama.

2) Pemodelan Support Vector Machine

Support Vector Machine (SVM) merupakan algoritma pembelajaran mesin yang digunakan untuk melakukan proses klasifikasi. SVM diperkenalkan oleh Boser, Guyon dan Vapnik. SVM adalah mesin yang menggunakan vector

sebagai pembeda untuk membagi data ke dalam dua kelompok. SVM bekerja dengan cara mencari *hyperplane* yang terbaik, yaitu dengan margin terbesar. Tujuannya adalah membagi data ke dalam dua kelompok melalui sebuah *hyperplane*. Ilustrasi tentang *hyperplane, margin* dalam SVM dapat dilihat pada Fig. 2.

Gambar 2.(a). merupakan kemungkinan bidang *hyperplane* yang memisahkan data lingkaran dan data kotak, sementara gambar 2.(b) memperlihatkan sebuah *hyperplane* yang sudah sempurna melalui algoritma SVM, mampu memisahkan data kotak dan data lingkaran dengan sempurna. Untuk kasus penelitian ini, data lingkaran dapat diilustrasikan sebagai penduduk miskin dan kotak sebagai penduduk yang tidak miskin.



Gambar 2. Hyperplane (a) Hyperplane yang belum sempurna (b) Hyperplane yang sempurna [8]

F. *Skenario Pengujian Model*

Skenario untuk menguji model klasfikasi adalah:

a. Nilai k-fold yang digunakan adalah 3, 5, dan 10.
b. Menggunakan SMOTE dan tanpa SMOTE
c. Pengujian bergantian untuk 1 atribut terpilih, 2 atribut sampai dengan 5 atribut.
d. Untuk C4.5, diuji untuk minimal *leaf,* yaitu dari 5, 10 dan 50
e. Untuk algoritma SVM, diuji untuk nilai C=0.1, 1 dan 10. Sementara nilai *degree* adalah 2, 3 dan 4.

G. *Evaluasi*

Seluruh skenario pengujian model baik C4.5 dan SVM akan dievaluasi menggunakan nilai akurasi. Nilai akurasi dihitung menggunakan (5)

$$Akurasi = \frac{TP + TN}{TP + FN + FN + TN} \; x \; 100\% \qquad (4)$$

dimana:
TP = jumlah data positif yang terklasifikasi benar,
TN = jumlah data negatif yang terklasifikasi benar.
FN = jumlah data negatif yang terklasifikasi salah,
FP = jumlah data positif yang terklasifikasi salah.

III. IMPLEMENTASI DAN ANALISIS HASIL

Pada bagian ini akan dijelaskan implementasi dari metode penelitian dan juga analisis hasil pengujian. Untuk implementasi model menggunakan Python.

A. Hasil Implementasi Tahap Pra Pemrosesan Data.

Pra pemrosesan data yang dilakukan adalah *data selection, data transformation, data normalization, feature selection* dan *balancing data.* Untuk proses *data selection,* pemilihan data dilakukan dengan melakukan proses 'remove' atribut Tempat Tanggal Lahir, NIK dan KK. Gambar 3 memperlihatkan hasil proses implementasi *data selection.*

	Jenis Kelamin	Umur	Pekerjaan	Pendapatan / Bulan	Tanggungan	Pendidikan Terakhir	Status Penduduk
0	L	69	PETANI	500000	1	SD	TIDAK MISKIN
1	L	69	PETANI	500000	1	SD	TIDAK MISKIN
2	P	67	PETANI	500000	1	SD	MISKIN
3	L	42	PETANI	500000	2	SD	MISKIN
4	L	56	PETANI	500000	2	SD	MISKIN

Gambar 3. Hasil Implementasi *Data Selection*

Proses *data transformation,* mengubah bentuk data dari atribut jenis kelamin, pekerjaan, pendidikan terakhir dan status penduduk ke bentuk yang lebih sesuai, menggunakan *snippet* seperti pada Gambar 4 dan Hasil program dapat dilihat pada Gambar 5. Proses selanjutnya adalah melakukan *feature selection* untuk mendapatkan atribut yang paling mempengaruhi penentuan kelas status penduduk miskin. Implementasi *feature selection* menggunakan fungsi infogain. Gambar 6 merupakan program untuk mendapatkan nilai *information gain*, sementara gambar 7 merupakan hasil *information gain* yang telah diurutkan berdasarkan nilai *information gain* tertinggi. Hasil dari gambar 7 memperlihatkan bahwa jenis kelamin memberikan nilai 0, berarti sama sekali tidak ada pengaruh dari atribut tersebut.

Untuk membuat skala yang sama dari nilai atribut maka dilakukan proses *data normalization*; Implementasi menggunakan fungsi *MinMaxScaler* dengan *range* [0,1]. Gambar 8 memperlihatkan *snippet* penggunaan fungsi MinMaxSclaer dan gambar 9 memperlihatkan sepuluh data teratas setelah dilakukan proses normalisasi.

```
#Transformation
data['Jenis Kelamin'].replace(['L','P'],[1,2], inplace=True)
data['Pekerjaan'].replace(['PETANI','WIRASWASTA','KARYAWAN','GURU','PNS'],[1,2,3,4,5],
                          inplace=True)
data['Pendidikan Terakhir'].replace(['SD','SMP','SMA','DIPLOMA','SARJANA'],[1,2,3,4,5],
                          inplace=True)
data['Status Penduduk'].replace(['TIDAK MISKIN','MISKIN'],[1,2],
                          inplace=True)
data.head(5)
```

Gambar 4. Snippet Data Transformation

	Jenis Kelamin	Umur	Pekerjaan	Pendapatan / Bulan	Tanggungan	Pendidikan Terakhir	Status Penduduk
0	1	69	1	500000	1	1	1
1	1	69	1	500000	1	1	1
2	2	67	1	500000	1	1	2
3	1	42	1	500000	2	1	2
4	1	56	1	500000	2	1	2

Gambar 5. Hasil Proses Tranformasi

```
# Feature Selection - information gain
x = data.drop(['Status Penduduk'], axis = 1)
y = data['Status Penduduk']
rank = infogain(x,y, random_state=42)
pd.set_option('display.max_rows',None)
rank = pd.Series(rank)
rank.index=x.columns
rank.sort_values(ascending=False)
```

Gambar 6. Snippet Information Gain

```
Pekerjaan             0.358955
Pendapatan / Bulan    0.353440
Pendidikan Terakhir   0.191621
Tanggungan            0.106287
Umur                  0.054187
Jenis Kelamin         0.000000
dtype: float64
```

Gambar 7. Hasil Pencarian Atribut Paling Berpengaruh

```
# Normalization
min_max_scaler = preprocessing.MinMaxScaler(feature_range=(0,1))
data = min_max_scaler.fit_transform(data)
dataset = pd.DataFrame({'Jenis Kelamin': data[:, 0], 'Umur': data[:, 1], 'Pekerjaan': data[:, 2],
                        'Tanggungan': data[:, 3], 'Pendapatan / Bulan': data[:, 4], 'Pendidikan Terakhir': data[:,5
dataset.head(10)
```

Gambar 8. Snippet Proses Normalisasi

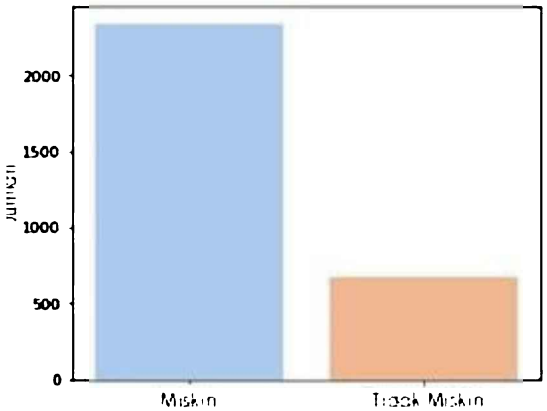
	Jenis Kelamin	Umur	Pekerjaan	Tanggungan	Pendapatan / Bulan	Pendidikan Terakhir	Status Penduduk
0	0.0	0.684211	0.0	0.0	0.0	0.0	0.0
1	0.0	0.684211	0.0	0.0	0.0	0.0	0.0
2	1.0	0.657895	0.0	0.0	0.0	0.0	1.0
3	0.0	0.328947	0.0	0.0	0.2	0.0	1.0
4	0.0	0.513158	0.0	0.0	0.2	0.0	1.0

Gambar 9. Hasil Proses Normalisasi

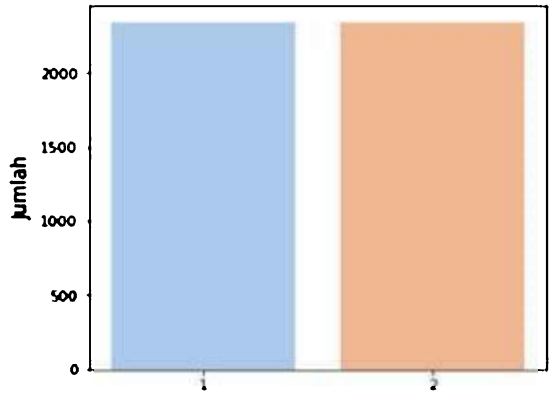
Implementasi proses *balancing data* menggunakan SMOTE. Gambar 10 merupakan *snippet* untuk proses *balancing data,* gambar 11 merupakan data sebelum proses SMOTE dan gambar 12 merupakan hasil setelah proses SMOTE.

```
sm = SMOTE(random_state=42)
x_res, y_res = sm.fit_resample(x, y)
final = pd.concat([x_res, y_res], axis=1)
```

Gambar 10. Snippet SMOTE



Gambar 11. Data Sebelum SMOTE



Gambar 12. Data Setelah SMOTE

B. *Analisis Hasil*

Pengujian yang dilakukan sesuai skenario pengujian yang telah dijelaskan pada bagian sebelumnya. Tabel 2 merupakan tabel hasil pengujian untuk 5 atribut. Berdasarkan hasil pengujian yang telah dilakukan, diketahui bahwa atribut yang paling mempengaruhi status miskin di desa Kenotan adalah pekerjaaan, pendapatan/bulan, pendidikan terakhir, tanggungan dan umur. Parameter pengujian sangat mempengaruhi hasil akurasi, Nilai k-fold mempengaruhi kedua algoritma.

Berdasarkan hasil pengujian yang telah dilakukan juga dapat diketahui bahwa parameter yang digunakan dalam proses pengujian memiliki pengaruh terhadap akurasi yang didapatkan pada masing-masing metode. Penggunaan nilai *k-fold* pada kedua metode berpengaruh pada akurasi.

Untuk metode SVM, semakin tinggi nilai *k-fold,* semakin stabil juga estimasi kinerja kecuali pada percobaan dengan 1 dan 2 atribut akurasi terbaik didapatkan pada *k-fold* terkecil yakni 3. Penggunaan atribut yang terbatas data menjadi sangat sederhana sehingga ketika data dibagi menjadi 3 subset atau k=3 dapat membantu menghasilkan model yang lebih baik. Selain itu juga k=3 membantu menghindari *overfiting* akibat atribut yang terlalu sedikit. Penggunaan parameter nilai C juga berperan dalam mengatur *trade-off* dari keakuratan data yang mana ketika nilai C terlalu tinggi maka akan terjadi *overfiting* sebaliknya ketika nilai C terlalu rendah maka akan terjadi *uderfiting.* Sedangkan penggunaan parameter *degree* berfungsi sebagai nilai yang mengontrol fungsi kernel yang digunakan.

Untuk metode C4.5 pengaruh dari penggunaan nilai *k-fold* sama seperti SVM, akan tetapi pengaruh parameter *minimal sample leaf* lebih dominan daripada pengaruh perubahan nilai *k-fold.* Pengaruh *minimal sample leaf* terlihat saat melakukan pengujian menggunakan 1, 2, 3, dan 4 atribut, akurasi terbaik mengikuti nilai parameter *minimal sample leaf 5* dan 10. Ketika menggunakan nilai parameter yang lebih tinggi terjadi *underfiting* sehingga akurasi yang dihasilkan rendah. Sedangkan pada pengujian 5 atribut dan 6 atribut akurasi terbaik ada pada parameter *minimal sample leaf* = 10. Semakin banyak atribut yang digunakan penggunaan *minimal sample leaf* = 10 dapat mengurangi kompleksitas model dan membantu menghindari *overfiting* maupun *underfiting.*

Pengujian melihat pengaruh dari *balancing* data, terlihat bahwa menggunakan *SMOTE* tidak berpengaruh dalam meningkatkan akurasi, hal ini terlihat pada semua pengujian yang dilakukan, akurasi tertinggi didapat pada data yang tidak dilakukan proses *SMOTE.*

TABEL 2. HASIL PENGUJIAN 5 ATRIBUT

Akurasi C4.5			K-FOLD	Akurasi SVM			
SMOTE	*NO SMOTE*	*Minimal Sample Leaf*		*C*	*Degree*	*SMOTE*	*NO SMOTE*
93.37	92.79	5	3	1,0	2	90.59	92.27
					3	90.78	92.18
					4	90.03	92.35
93.06	94.18	10		1	2	93.31	93.48
					3	93.31	93.31
					4	91.98	92.96
92.11	91.05	50		10	2	93.24	93.14
					3	93.12	92.88
					4	92.74	92.70
93.75	93.57	5	5	0,1	2	91.10	92.36
					3	91.16	92.01
					4	90.40	92.36
93.06	94.00	10		1	2	92.87	93.05
					3	92.80	93.31
					4	91.92	93.14
90.97	93.74	50		10	2	92.80	92.96
					3	93.06	93.31
					4	92.74	92.96
94.19	93.40	5	10	0.1	2	91.10	92.27
					3	91.35	92.36
					4	90.72	92.36
93.50	94.35	10		1	2	93.05	93.23
					3	93.05	93.49
					4	92.49	93.14
91.92	93.74	50		10	2	92.87	93.40
					3	92.99	93.22
					4	92.55	93.22

IV. KESIMPULAN DAN SARAN

A. *Kesimpulan*

Metode C4.5 dan SVM telah berhasil digunakan untuk mengelompokan penduduk miskin di desa Kenotan. Terdapat lima atribut yang mempengaruhi kelompok penduduk miskin yaitu pekerjaan, pendapatan/bulan, jumlah tanggungan, pendidikan terakhir dan umur.

Tinjauan dari sisi penggunaan model klasifikasi, dapat disimpulkan bahwa banyaknya nilai k-fold mempengaruhi akurasi model baik menggunakan metode C4.5 maupun SVM. Akurasi terbaik pada penelitian ini dihasilkan dari metode C4.5 dengan nilai akurasi 94.35%. Parameter yang diperoleh untuk hasil terbaik tersebut adalah data tidak dilakukan proses *balancing data* dengan jumlah *leaf* adalah 10. Untuk penerapan model dengan algoritma SVM diperoleh kesimpulan bahwa akurasi tertinggi 93.49% menggunakan data sebelum *balancing* dengan k-fold sama dengan 10, nilai C sama dengan 1 dan *degree* sama dengan 3. Keefektifan metode C4.5 dan SVM dalam mengklasifikasikan penduduk miskin di desa Kenotan menunjukan bahwa lima atribut utama memiliki pengaruh yang signifikan. Khususnya pada metode C4.5 yang menunjukan akurasi tertinggi, menyarankan potensi yang besar untuk diintegrasikan dalam kerangka kerja pengambilan keputusan desa Kenotan untuk alokasi sumber daya. Tetapi, hasil ini harus dipandang dari konteks keterbatasan *dataset* yang digunakan, yang berpotensi mempengaruhi generalisasi pada hasil penelitian.

B. *Saran*

Penelitian selanjutnya harus mempertimbangkan evaluasi fitur yang lebih serta menggunakan dataset yang lebih beragam untuk memvalidasi model yang ada serta menguji kehandalannya dalam berbagai kondisi. Penelitian ini juga dapat diperluas untuk mencakup perkembangan prototipe aplikasi yang dapat digunakan oleh pemerintah desa untuk membantu upaya identifikasi penduduk miskin secara efektif dan efisien.

REFERENSI

[1] Badan Pusat Statistik, "Sosial dan Kependudukan," https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html.

[2] A. P. Pristiawati, I. Permana, Z. Zarnelly, and F. Muttakin, "Klasifikasi Penerima Bantuan Beras Miskin Menggunakan Algoritma K-NN, NBC dan C4.5," *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, Jun. 2023, doi: 10.47065/bits.v5i1.3617.

[3] F. Kurnia, J. Kurniawan, I. S. Fahmi, and S. Monalisa, "Klasifikasi Keluarga Miskin Menggunakan Metode K-Nearest Neighbor Berbasis Euclidean Distance," 2019.

[4] A. A. Kasim and M. Sudarsono, "Algoritma Support Vector Machine (SVM) untuk Klasifikasi Ekonomi Penduduk Penerima Bantuan Pemerintah di Kecamatan Simpang Raya Sulawesi Tengah," *Seminar Nasional APTIKOM*, 2019.

[5] W. Agustina, M. T. Furqon, and B. Rahayudi, "Implementasi Metode Support Vector Machine (SVM) Untuk Klasifikasi Rumah Layak Huni (Studi Kasus: Desa Kidal Kecamatan Tumpang Kabupaten Malang)," 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[6] Laila Qardini, Andi Seppewali, and Asra Aina, "Decision Tree Dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial," 2021.

[7] J. R. Quinlan, *C4.5: Programs for Machine Learning*. 2014.

[8] D. Dahman, "Support Vector Machine (SVM). Memperkenalkan sebuah algoritma." Accessed: Aug. 20, 2023. [Online]. Available: https://medium.com/sysinfo/support-vector-machine-svm-5d95a7d7a547

Penerapan Pemrosesan Citra dan CNN untuk Klasifikasi Citra Tangan Bahasa Isyarat Indonesia (BISINDO)

Maria Ribka Restu Sukma Ningsih
Departemen Informatika
Universitas Sanata Dharma
Yogyakarta, Indonesia
mariaribkasukmaningsih@gmail.com

Anastasia Rita Widiarti
Departemen Informatika
Universitas Sanata Dharma
Yogyakarta, Indonesia
rita_widiarti@usd.ac.id

Abstract— **This study pertains to the development of a sign language recognition model using image processing technology. Sign language serves as the primary communication mode for individuals with hearing impairments; however, they encounter limitations when interacting with those unfamiliar with sign language. The objective of this research is to create BISINDO recognition to translate BISINDO images into text. The author employed image processing and the Convolutional Neural Network (CNN) technique with the ResNet-50 architecture due to its effectiveness in object recognition tasks in images, particularly in image classification. ResNet-50 is renowned for its ability to tackle complex image classification issues, making it a suitable choice for sign language recognition involving hand image representations with diverse positions and variations. Hand image data were collected from 10 participants representing the BISINDO alphabet and sourced from the Kaggle platform, utilizing the A-Z alphabet (excluding J), attributed to the dynamic nature of these letters portrayed with movements, unadapted for this research. Despite the inclusion of the letter J in the Kaggle dataset, it was intentionally excluded to maintain consistency in the analysis. The study evaluated various scenarios, including image size, the number of epochs, and the utilization of public datasets. The results indicated that the model achieved the highest accuracy without employing public datasets, utilizing a 256x256 image size and 20 and 50 epochs, achieving 100% accuracy. However, a more thorough analysis comparing models utilizing the highest accuracy and epochs from public datasets was conducted. The optimal-performing model incorporated an additional public dataset, employing a 256x256 image size and 50 epochs, achieving 99% accuracy. One limitation of this model is the requirement for participants to face the camera for accurate predictions. Enhancing the dataset volume with greater variations would enhance the model's capability to discern nuances between similar signs.**

Keywords—CNN, ResNet50, BISINDO, Image Classification, Sign Language

I. LATAR BELAKANG

Komunikasi merupakan proses yang kompleks antara dua orang atau lebih untuk mengekspresikan, menafsirkan, dan mengoordinasikan sebuah atau beberapa pesan. Komunikasi juga gagasan untuk berbagi pikiran dengan orang lain. Komunikasi itu sendiri dibagi menjadi tiga jenis, yaitu komunikasi verbal (kata-kata), non-verbal (isyarat), dan gambaran visual[1]. Komunikasi memiliki peran yang sangat penting bagi individu yang mengalami gangguan pendengaran dalam memenuhi kebutuhan sosial mereka. Bahasa Isyarat menjadi sarana utama bagi mereka untuk berkomunikasi, melibatkan ekspresi tangan, ekspresi wajah, dan postur tubuh untuk menyampaikan pesan secara visual dan melalui bahasa tubuh. Meskipun demikian, mereka menghadapi tantangan dalam berinteraksi dengan individu yang tidak memahami Bahasa Isyarat.

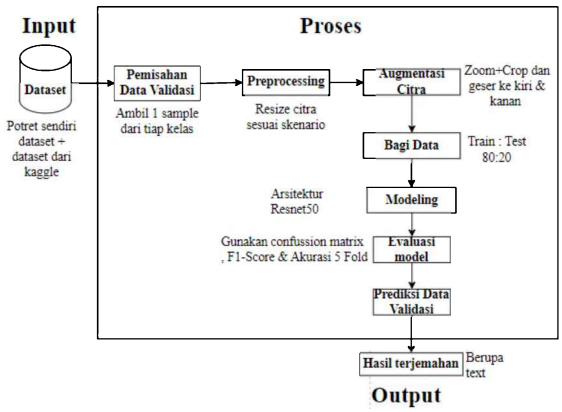
Dalam rangka memfasilitasi komunikasi antara tunarungu dan masyarakat umum, penting untuk mengembangkan model yang dapat melakukan pengenalan bahasa isyarat. Salah satu metode yang digunakan adalah pemrosesan citra, yang merupakan bidang yang mempelajari pengolahan gambar menggunakan teknik dan algoritma komputer.

Pada tahun 2015, sebuah studi melibatkan 100 partisipan yang tunarungu, terdiri dari remaja dan orang dewasa. Hasil studi menunjukkan bahwa sebagian besar dari mereka, sekitar 91%, lebih memilih menggunakan Bahasa Isyarat Indonesia (BISINDO) sebagai bahasa komunikasi sehari-hari mereka. Hanya sebagian kecil, yakni sekitar 9%, yang menggunakan Sistem Isyarat Bahasa Indonesia (SIBI)[2]. Berdasarkan penemuan tersebut, penelitian ini akan difokuskan pada BISINDO.

Pemrosesan citra dapat dimanfaatkan untuk pengolahan gambar sehingga dapat dimanfaatkan untuk menonjolkan ciri dan juga dapat digunakan untuk melakukan augmentasi. Dalam penelitian ini, metode Convolutional Neural Network (CNN) digunakan sebagai algoritma klasifikasi untuk pengenalan bahasa isyarat. CNN adalah jenis jaringan saraf yang efektif dalam mengenali pola kompleks pada citra.

Penelitian yang dilakukan oleh Ego Oktafanda [3] menyimpulkan bahwa dengan dataset yang terbatas, penggunaan metode Convolutional Neural Network (CNN) dengan penerapan teknik transfer learning menggunakan arsitektur ResNet50 mampu mencapai nilai akurasi sebesar 0,95% dalam tugas klasifikasi citra. CNN sangat efektif dalam melakukan klasifikasi objek pada citra[4] selain itu, identifikasi varian kendaraan CNN mencapai nilai akurasi sebesar 73,33%.[5]. dan masih ada beberapa sumber lain yang mengatakan bahwa CNN cukup baik dalam pengenalan data citra sehingga dipilihlah metode ini.

II. METODE PENELITIAN



Gambar 1. Alur Penelitian

Gambar 1 merupakan desain penelitian yang akan memberikan gambaran visual tentang rancangan penelitian. Dengan melihat gambar desain penelitian ini, diharapkan pembaca dapat memahami dengan lebih jelas bagaimana penelitian ini akan dilaksanakan dan bagaimana data akan dikumpulkan serta diolah sampai pada outputnya.

A. *Pengumpulan Data*

Penelitian ini menggunakan citra yang diambil menggunakan kamera ponsel dengan resolusi 3000x4000 piksel. Jarak antara kamera dan objek yang difoto adalah sekitar 50 cm, dan latar belakang dalam foto-foto ini sangat bervariasi. Untuk memastikan stabilitas kamera selama pengambilan gambar, digunakan tripod dengan tinggi sekitar 120 cm.

Ada 10 orang yang terlibat dalam pengambilan dataset, termasuk penulis. Saat pengambilan gambar, setiap partisipan duduk di kursi dan melakukan gerakan tangan yang merepresentasikan abjad dalam Bahasa Isyarat Indonesia (BISINDO) dari huruf A hingga Z, kecuali huruf J. Dengan demikian, masing-masing partisipan menghasilkan 25 foto yang berbeda. Selain dataset yang dibuat oleh penulis dengan menggunakan ponsel dan tripod, penelitian ini juga memanfaatkan dataset publik yang tersedia di platform Kaggle pada link https://www.kaggle.com/datasets/achmadnoer/alfabet-bisindo/data. Dataset tambahan ini digunakan untuk menambah variasi data yang digunakan dalam penelitian dan untuk menguji performa model saat dataset publik ini disertakan dalam analisis. Gambar 2 merupakan contoh data pribadi dan gambar 3 merupakan contoh data publik.

Gambar 2. Data Pribadi

Gambar 3. Data Publik

TABEL 1. DETAIL DATASET

Sumber Dataset	Jumlah Citra Tiap Huruf	Huruf yang Dipakai
Dipotret Sendiri	10	A-Z (Kecuali J)
Kaggle	12	A-Z (Kecuali J)

Dalam pengumpulan dataset, penelitian ini secara khusus menghindari penggunaan huruf J karena huruf tersebut bersifat dinamis (diperagakan dengan gerakan), yang tidak dapat diadaptasi dalam penelitian ini. Meskipun dataset dari Kaggle mencakup huruf J, penelitian ini memilih untuk tidak memanfaatkannya. Keputusan ini diambil untuk menjaga konsistensi dalam analisis.

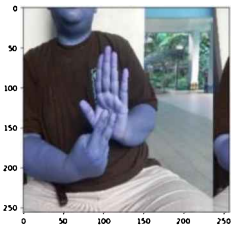
B. *Preprocessing*

Dalam tahap preprocessing, fokus utama adalah membuat ukuran citra sesuai dengan keperluan penelitian sehingga sesuai dengan berbagai skenario pengujian yang berbeda. Penelitian ini menggunakan dua variasi ukuran citra input: 256x256 dan 64x64. Penentuan pilihan ukuran citra ini didasarkan pada hasil penelitian terdahulu oleh Sari [6] yang menunjukkan performansi terbaik diperoleh pada citra berukuran 256x256. Namun, dalam pengamatan hasil pelatihan data dan data validasi, terlihat bahwa citra dengan ukuran 64x64 menunjukkan tingkat stabilitas yang lebih konsisten. Penyesuaian ukuran citra dilakukan sesuai dengan skenario yang sedang dijalankan.

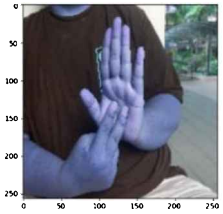
C. *Augmentasi*

Pada penelian ini penulis melakukan beberapa teknik augmentasi hanya pada dataset pribadi untuk menambah variasi data. Augmentasi yang dilakukan adalah:

1. Melakukan zoom dengan nilai [1.0, 1.4, 1.6, 1.8, 1.9] lalu crop dan tambahkan ke array dataset_images dan dataset_labels. Hasil dari proses ini terdapat pada gambar 4.
2. Hasil crop tersebut akan di geser secara horizontal dengan nilai : [-20, -10, 10, 20] dan akan disimpan ke array dataset_images dan dataset_labels. Hasil dari proses ini terdapat pada gambar 5.



Gambar 5. Hasil Augmentasi Penggeseran Citra



Gambar 4. Hasil Zoom dan Crop Citra

D. *Modeling*

Pada tahap pemodelan, penelitian ini memanfaatkan arsitektur ResNet-50, yang termasuk bagian dari arsitektur CNN yang telah terbukti efektif dalam tugas-tugas pengenalan objek pada citra. Pemilihan ResNet-50 sebagai model utama dalam penelitian ini karena arsitektur ini cukup populer untuk klasifikasi gambar [7]. Pada penelitian ini jumlah kelasnya adalah 25 sehingga pada arsitektur resnet50 jumlah kelasnya akan diubah dari 1000 ke 25 dengan menggunakan beberapa variasi epoch dan citra input sesuai skenario. Digunakan pula optimizer adam karena pada penelian terdahulu optimizer tersebut memperoleh akurasi tertinggi dibanding optimizer yang lain[6].

III. HASIL DAN PEMBAHASAN

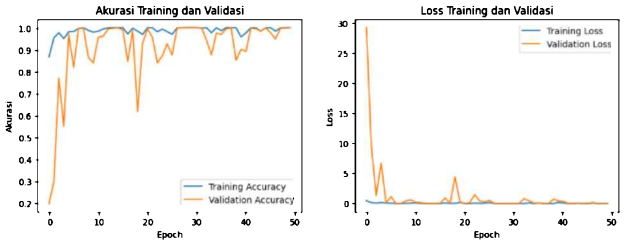
TABEL 2. HASIL PERCOBAAN

Skenario	Citra Input	Epoch	Dataset Publik	F1-Score	Akurasi
1	(256,256,3)	20	Gabung	0.99	0.99
2	(256,256,3)	50	Gabung	0.99	0.99
3	(64,64,3)	20	Gabung	0.99	0.99
4	(64,64,3)	50	Gabung	0.98	0.98
5	(256,256,3)	20	Tidak	1	1
6	(256,256,3)	50	Tidak	1	1
7	(64,64,3)	20	Tidak	0.78	0.78
8	(64,64,3)	50	Tidak	0.99	0.99

Dari tabel 2 yaitu tabel hasil eksperimen, akurasi tertinggi ditemukan pada skenario 5 dan 6, di mana akurasi mencapai 100%, tanpa penggunaan dataset publik tambahan. Namun, hasil ini belum cukup untuk menyatakan bahwa model pada skenario 5 dan 6 adalah yang terbaik. Oleh karena itu, penulis akan melakukan perbandingan dengan skenario-skenario lain yang menggunakan dataset publik dan akurasinya tertinggi.

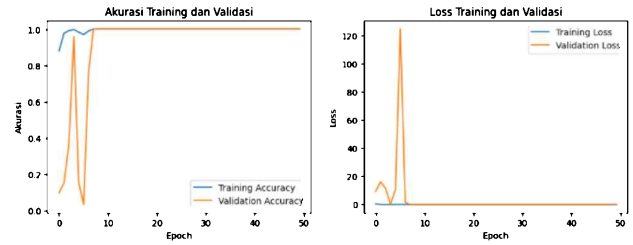
Untuk menjadikan perbandingan lebih efisien dan adil, hanya model dengan jumlah epoch 50 yang akan dibandingkan. Hal ini disebabkan oleh kenyataan bahwa penggunaan epoch memiliki dampak signifikan pada kinerja sistem klasifikasi CNN dengan model ResNet-50, dimana terjadi peningkatan kinerja seiring dengan peningkatan jumlah epoch yang digunakan. [8] . Oleh karena itu, perbandingan akan dilakukan antara skenario 2 dan skenario 6.

Selanjutnya, penulis akan mengamati grafik dan confusion matrix dari kedua model ini untuk mendapatkan pemahaman yang lebih mendalam tentang hasil pelatihan.



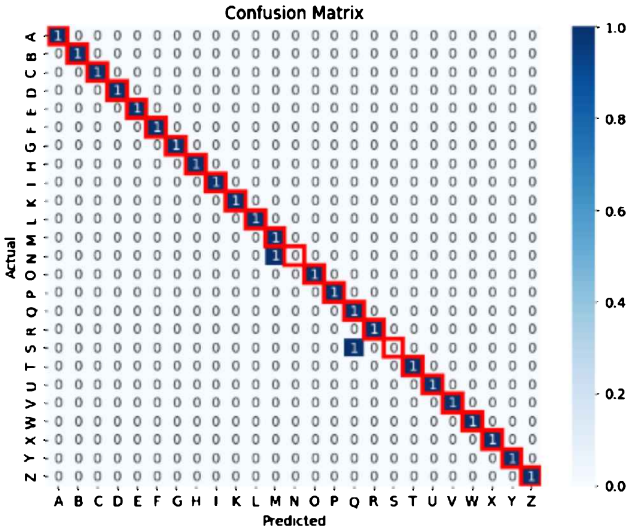
Gambar 6. Grafik Akurasi dan Loss Skenario 2

Pada gambar 6, terlihat bahwa model skenario 2 mengalami naik turun di akurasi dan loss namun tidak begitu signifikan, selain itu semakin lama dilatih nilainya akan menuju stabil.



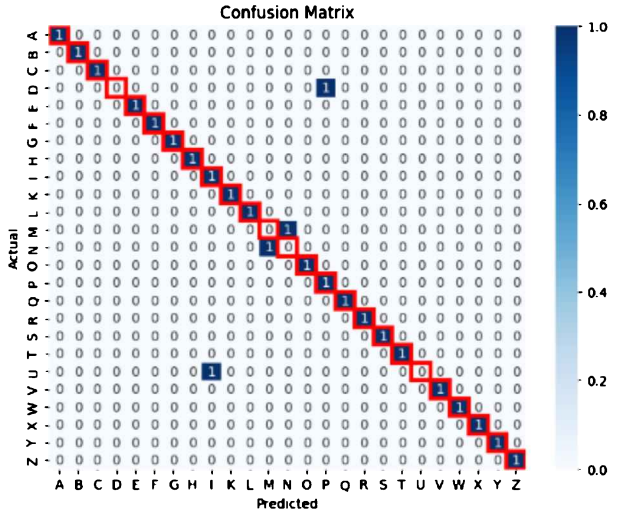
Gambar 7. Grafik Akurasi dan Loss Skenario 6

Sedangkan pada gambar 7, grafik akurasi dan loss model skenario mengalami kenaikan dan penurunan yang sangat signifikan.



Gambar 8. Confussion Matrix Skenario 2

Pada Gambar 8, ditampilkan hasil prediksi data validasi oleh model pada Skenario 2. Secara keseluruhan, model ini mampu memprediksi data validasi dengan sangat baik, dengan hanya dua kesalahan yang tercatat. Kesalahan tersebut terjadi pada huruf 'N', yang salah diprediksi sebagai 'M', dan huruf 'S', yang salah diprediksi sebagai 'Q'.



Gambar 9. Confussion Matrix Skenario 6

Pada gambar 9, ditampilkan pula hasil prediksi data validasi oleh model pada Skenario 6. Secara keseluruhan, model ini juga mampu memprediksi data validasi dengan sangat baik, namun jumlah kesalahan skenario 6 lebih banyak dibanding skenario 2. Untuk memberikan gambaran lebih rinci mengenai jenis kesalahan yang terjadi, tabel berikut membandingkan hasil prediksi antara skenario 2 dan skenario 6 untuk masing-masing huruf:

TABEL 3. PERBANDINGAN HASIL PREDIKSI ANTARA SKENARIO 2 DAN 6

Huruf	Hasil Prediksi Skenario 2	Hasil Prediksi Skenario 6
D	Berhasil	Tidak
M	Berhasil	Tidak
N	Tidak	Tidak
U	Berhasil	Tidak
S	Tidak	Berhasil

Tabel 3 memuat daftar huruf yang tidak berhasil diprediksi oleh skenario 6, yaitu D, U, M, dan N, serta huruf yang tidak berhasil diprediksi oleh skenario 2, yaitu S dan N. Tabel tersebut memberikan informasi tentang bagaimana skenario 2 dan skenario 6 memiliki keberhasilan atau kegagalan dalam memprediksi huruf-huruf tersebut. Dengan membandingkan dua skenario ini, kita dapat melihat bagaimana 2 model ini dapat mengatasi atau gagal mengatasi tantangan yang ada dalam memprediksi masing-masing huruf.

Hasil akhirnya adalah bahwa skenario 2 unggul dalam memprediksi huruf-huruf yang tidak berhasil diprediksi oleh skenario 6. Dengan kata lain, skenario 2 memiliki kinerja yang lebih baik dalam hal ini karena berhasil memprediksi lebih banyak huruf dengan benar dibandingkan dengan skenario 6.

Dalam analisis lebih lanjut terhadap skenario 2, terdapat beberapa kesalahan prediksi yang perlu diperhatikan. Dalam kasus ini, huruf 'N' diprediksi sebagai 'M', dan huruf 'S' diprediksi sebagai 'Q'. Kesalahan ini sebagian besar disebabkan oleh kemiripan bentuk antara huruf tersebut. Kemiripan bentuk huruf menggunakan bahasa isyarat dapat dilihat pada gambar 10, 11, 12, dan 13.

Gambar 10. Huruf M

Gambar 11. Huruf N

Gambar 12. Huruf Q

Gambar 13. Huruf S

IV. KESIMPULAN DAN SARAN

A. *Kesimpulan*

Dari temuan dalam penelitian, dapat disimpulkan bahwa:

1. Hasil eksperimen menunjukkan akurasi & F1-Score tertinggi ditemukan pada skenario 5 dan 6 sebesar 1.0 atau 100% pada model tanpa penggunaan dataset publik dengan citra input berukuran 256x256 dan jumlah epoch 20 dan 50. Namun, untuk menentukan model terbaik, perbandingan dilakukan dengan skenario yang menggunakan data tambahan dan jumlah epoch 50 dengan ukuran citra input yang sama yaitu 256x256 yang merupakan skenario 2.
2. Jumlah epoch terbukti berpengaruh signifikan terhadap performa model. Skenario dengan 50 epoch diprioritaskan untuk analisis komparatif karena menunjukkan peningkatan performa yang konsisten.
3. Skenario 2 menunjukkan superioritas dalam memprediksi beragam isyarat, termasuk isyarat yang mirip, berbanding dengan skenario 6, yang keduanya mencatatkan akurasi yang tinggi.

B. *Saran*

Untuk penelitian selanjutnya pengumpulan data citra tangan dapat diperluas, khususnya untuk isyarat huruf M & N serta S & Q yang sering kali dibingungkan oleh model. Penambahan variasi dalam data dapat dilakukan dengan mengatur posisi tangan, kecerahan, latar belakang, dan sudut pandang pengambilan gambar. Selain itu, variasi data dapat dilakukan melalui penggunaan teknologi seperti \Generative Adversarial Network (GAN) yang dapat memperkaya pemahaman model mengenai nuansa antara isyarat yang serupa. Penelitian selanjutnya pun dapat mengeksplorasi metode untuk meningkatkan generalisasi model terhadap semua isyarat BISINDO dan potensi adaptasi model untuk bahasa isyarat lainnya.

UCAPAN TERIMA KASIH

Ucapan syukur kepada Tuhan Yang Maha Esa untuk segala hal baik dan rahmat yang diberikan sehingga saya dapat menyelesaikan semua ini dengan baik. Terima kasih juga kepada orang tua untuk semua dukungan dan *feedback* baik selama perjalanan ini. Terima kasih sebesar-besarnya kepada Ibu Anastasia Rita Widiarti yang sudah membimbing dan memberikan saya hal yang berharga sampai pada titik ini. Semua doa baik dan segala hal baik semoga menyertai ibu. Tak lupa saya mengucap limpah terima kasih kepada teman-teman yang dengan tangannya memberikan kontribusi yang banyak pada saya. Saya percaya, tanpa kalian, penelitian ini belum tentu bisa selesai. Segala syukur pernah bersama kalian yang terbaik.

REFERENSI

[1] K. S. Verderber, D. D. Sellnow, and R. F. Verderber, *Communicate!* 2017.

[2] R. A. Mursita, "Respon Tunarungu Terhadap Sistem Bahasa Isyarat Indonesia (SIBI) dan Bahasa Isyarat Indonesia (BISINDO) dalam Komunikasi," *Inklusi*, vol. 2, no. 2, pp. 221–232, 2015, [Online]. Available: http://www.change.org/id/petisi/

[3] E. Oktafanda, "Klasifikasi Citra Kualitas Bibit dalam Meningkatkan Produksi Kelapa Sawit Menggunakan Metode Convolutional Neural Network (CNN)," *J. Inform. Ekon. Bisnis*, vol. 4, no. 3, pp. 72–77, 2022, doi: 10.37034/infeb.v4i3.143.

[4] Putra, W. S. Eka, and R. Soelaiman, "Klasifikasi Citra Menggunakan Convolutional Neural Network (CNN) Pada Caltech 101," *J. Tek. ITS*, vol. 5, no. 1, p. 76, 2016, [Online]. Available: http://repository.its.ac.id/48842/

[5] N. Fadlia and R. Kosasih, "KLASIFIKASI JENIS KENDARAAN MENGGUNAKAN METODE CONVOLUTIONAL NEURAL NETWORK (CNN)," *J. Ilm. Teknol. dan Rekayasa*, vol. 24, no. 3, pp. 207–215, 2019, doi: 10.35760/tr.2019.v24i3.2397.

[6] D. H. A. Sari, S. Sa'idah, and N. K. C. Pratiwi, "Klasifikasi Jenis Kulit Wajah Menggunakan Modifikasi Convolutional Neural Network (CNN)," *e-Proceeding Eng.*, vol. 8, no. 6, pp. 3188–3194, 2022.

[7] U. Kulsum and A. Cherid, "Penerapan Convolutional Neural Network Pada Klasifikasi Tanaman Menggunakan ResNet50," *Simkom*, vol. 8, no. 2, pp. 221–228, 2023, doi: 10.51717/simkom.v8i2.191.

[8] F. Nashrullah, S. A. Wibowo, and G. Budiman, "Investigasi parameter epoch pada arsitektur resnet-50 untuk klasifikasi pornografi," *J. Comput. Electron. Telecommun.*, vol. 1, no. 1, pp. 1–8, 2020, doi: 10.52435/complete.v1i1.51.

Implementasi Rantai Markov untuk Prediksi Data Hemoglobin Pasien Pengidap Kanker Payudara

Mikael Raditya Agung Sasmita
Fakultas Sains dan Teknologi
Universitas Sanata Dharma
Yogyakarta, Indonesia
mikaelradityas@gmail.com

Sabina Rossa Adriani Wibowo
Fakultas Sains dan Teknologi
Universitas Sanata Dharma
Yogyakarta, Indonesia
sabinarossa25@gmail.com

Aldiyes Paskalis Birta
Fakultas Sains dan Teknologi
Universitas Sanata Dharma
Yogyakarta, Indonesia
aldiyespaskalisbirta@gmail.com

Anastasia Rita Widiarti
Fakultas Sains dan Teknologi
Universitas Sanata Dharma
Yogyakarta, Indonesia
rita_widiarti@usd.ac.id

Abstrak— Kestabilan pengobatan kanker penting untuk perawatan pasien kanker. Hemoglobin, protein dalam sel darah merah, penting untuk menjaga kestabilan pengobatan kanker. Perubahan tingkat hemoglobin pada pasien kanker disebabkan oleh berbagai faktor, seperti pengobatan kanker, tumor, anemia, atau gangguan produksi hemoglobin. Paper ini menyajikan implementasi rantai markov untuk memprediksi kadar hemoglobin dari seorang wanita pengidap kanker. Data diperoleh dari hasil pengukuran kadar hemoglobin seorang wanita berusia 50 tahun selama 35 kali pengecekan. Dari hasil analisis rantai Markov, distribusi probabilitas jangka panjang dari keadaan-keadaan menunjukkan bahwa dalam jangka panjang, kemungkinan wanita tersebut berada dalam keadaan rendah stabil di sekitar 0,655726 (65,57%), kemungkinan berada dalam keadaan normal stabil di sekitar 0,344274 (34,43%), sementara kemungkinan berada dalam keadaan tinggi adalah 0 (0%). Hal ini mengindikasikan bahwa dengan tingkat kepastian yang tinggi, wanita tersebut akan berada dalam keadaan rendah atau normal pada masa depan.

Kata Kunci—rantai markov, hemoglobin, probabilitas transisi, steady state

I. PENDAHULUAN

Kestabilan pengobatan kanker adalah elemen penting dalam perawatan pasien kanker payudara. Hemoglobin, sebuah protein yang terdapat dalam sel darah merah, memegang peran utama dalam menjaga kestabilan pengobatan kanker. Hemoglobin berfungsi sebagai pembawa oksigen ke seluruh tubuh, termasuk jaringan kanker. Kadar hemoglobin yang optimal penting untuk mendukung fungsi normal sel-sel tubuh dan nutrisi jaringan kanker. Menurut Standar Siloam Hospitals, kadar hemoglobin normal pada wanita dewasa berkisar antara 12-15 gram/dL [1].

Perubahan kadar hemoglobin pada pasien kanker dapat disebabkan oleh berbagai faktor, termasuk jenis pengobatan kanker yang diterapkan, pertumbuhan tumor, kehadiran anemia, atau gangguan dalam produksi hemoglobin. Kadar hemoglobin yang rendah dapat mengakibatkan kelelahan, penurunan energi, penurunan toleransi terhadap pengobatan, dan kekurangan oksigen yang dapat menghambat efektivitas pengobatan kanker itu sendiri.

Manajemen kadar hemoglobin menjadi esensial dalam perawatan pasien kanker. Pengukuran rutin kadar hemoglobin memungkinkan dokter untuk memantau respon pasien terhadap pengobatan, mengidentifikasi kondisi anemia, dan memberikan intervensi yang tepat seperti terapi transfusi darah atau penyesuaian dosis pengobatan.

Studi sebelumnya telah menunjukkan keberhasilan Rantai Markov dalam memprediksi berbagai parameter kesehatan, seperti tekanan darah [2]. Penerapan metode ini juga telah diterapkan luas dalam berbagai disiplin, termasuk pengelolaan stok dalam perusahaan manufaktur [3], pemodelan status pasien di rumah sakit [4], pemodelan kinerja perkerasan jalan [5], perencanaan penjualan [6], serta analisis pergeseran dalam persediaan perusahaan [7]. Penelitian ini melanjutkan tren penerapan Rantai Markov dalam berbagai konteks, mengadaptasinya untuk memprediksi perubahan kadar hemoglobin pada pasien kanker, yang diharapkan akan memberikan kontribusi berharga dalam pemahaman dan peningkatan perawatan individu atau populasi yang relevan di masa depan.

Dalam paper ini, kami akan mengimplementasikan pendekatan Rantai Markov untuk memprediksi kadar hemoglobin pada seorang wanita yang mengidap kanker. Kami akan menjelaskan langkah-langkah yang terlibat dalam proses pengolahan data, mulai dari pengumpulan data kadar hemoglobin dari individu atau populasi yang relevan, analisis data untuk mengidentifikasi pola perubahan, hingga pembangunan model Rantai Markov yang sesuai.

Tujuan utama dari penelitian ini adalah memahami pola perubahan kadar hemoglobin dalam populasi yang dipelajari. Melalui pendekatan ini, diharapkan penelitian ini akan memberikan wawasan yang berharga dalam pemahaman serta memfasilitasi pemantauan yang lebih baik terhadap individu atau populasi yang relevan, sehingga mereka mendapatkan penanganan yang lebih baik untuk kondisi mereka di masa depan.

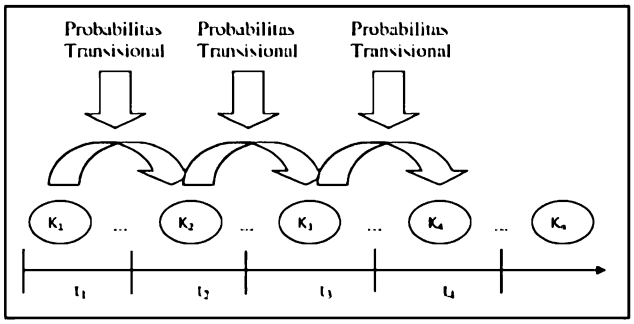
II. METODE PENELITIAN

A. *Metode Pengumpulan Data*

Data hemoglobin diperoleh dari Rumah Sakit Siloam. Subjek penelitian adalah perempuan berusia 50 tahun pengidap kanker. Data mencakup tanggal pengecekan dan kadar hemoglobin dari 35 pengecekan yang mengikuti jadwal pemeriksaan pasien. Pengambilan data mengikuti etika penelitian medis dan menjaga privasi pasien; identitas pasien tidak disertakan dalam data. Pengukuran kadar hemoglobin dilakukan dengan analisis darah rutin oleh staf medis yang berpengalaman. Pengambilan data yang mengikuti jadwal pemeriksaan pasien diharapkan mencerminkan variasi kadar hemoglobin pasien kanker dengan baik.

B. Pemodelan Rantai Markov

Analisis Rantai Markov merupakan suatu teknik matematika yang biasa digunakan untuk melakukan pemodelan bermacam- macam sistem dan proses bisnis. Model Rantai Markov ditemukan oleh seorang ahli Rusia yang bernama A.A. Markov pada tahun 1906, ilustrasi peristiwa yang terjadi dalam Rantai Markov dapat dilihat pada Gambar 1.



Gambar 1. Peristiwa dalam Rantai Markov

Sumber: Kesuma, dkk., 2018

Gambar 1 menggambarkan bahwa untuk setiap waktu *t*, ketika kejadian adalah K_n dan seluruh kejadian sebelumnya adalah $K_1, K_2, \dots, K_{n-1}$ yang terjadi dari proses yang diketahui, probabilitas seluruh kejadian yang akan datang K_n hanya bergantung pada kejadian K_{n-1} dan tidak bergantung pada kejadian-kejadian sebelumnya, yaitu $K_{n-1}, K_{n-2}, \dots, K_1$.

Kejadian-kejadian pada Gambar 1 memiliki sifat berantai. Oleh karena itu teori ini dikenal dengan nama Rantai Markov. Dengan demikian, Rantai Markov menjelaskan pergerakan dari beberapa variabel dalam satu periode waktu di masa yang akan datang berdasarkan pergerakan varibel tersebut di masa kini. Secara matematis persamaan Rantai Markov dapat ditulis sebagai berikut:

$$K_n = P \times K_{n-1} \quad (1)$$

K_n : peluang kejadian ke *n*
K_{n-1} : peluang kejadian ke- *n - 1*
P : probabilitas Transisional

Proses Markov adalah proses stokastik dimana masa lalu tidak mempunyai pengaruh pada masa yang akan datang bila masa sekarang diketahui. Rantai Markov (Markov chain) adalah suatu metode yang mempelajari sifat-sifat suatu variabel pada masa sekarang yang didasarkan pada sifat-sifatnya di masa lalu dalam usaha menaksir sifat-sifat variabel tersebut di masa yang akan datang.
Serangkaian variabel acak dikatakan sebagai Rantai Markov dengan waktu diskrit apabila memenuhi:

$$P\{X_{t+1} = j \mid X_0 = i_0, \dots, X_{t-1} = i_{t-1}, X_t = i\} = P\{X_{t+1} = j \mid X_t = i\} \quad (2)$$

Untuk semua $t \in T$ dan semua *state* $i_0, \dots, i_{t-1}, i, j \in S$

Matriks probabilitas transisi adalah matriks yang berisikan kemungkinan perubahan dari satu keadaan ke keadaan yang lain. Matriks probabilitas transisi satu langkah didefinisikan sebagai:

$$P = \begin{bmatrix} P_{11} & P_{12} & P_{13} & \dots & P_{1m} \\ P_{21} & P_{22} & P_{23} & \dots & P_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{m1} & P_{m2} & P_{m3} & \dots & P_{mm} \end{bmatrix}, \forall i, \sum_{j=1}^{j=m} P_{ij} = 1 \quad (3)$$

dengan P_{ij} menyatakan probabilitas bahwa jika proses tersebut berada pada *state i* maka berikutnya akan beralih ke *state j*. Dimana setiap elemen dari matriks *P* bernilai tak negatif dan jumlah elemen- elemen pada satu baris pada matriks probabilitas transisi harus sama dengan 1.

Probabilitas transisi *n-step* atau peluang trasnsisi *n* langkah $\left(P_{ij}^{(n)}\right)$ adalah peluang bersyarat suatu sistem yang berada pada *state i* dan akan berada pada *state j* setelah proses mengalami n transisi atau perubahan, maka secara matematis dapat dituliskan:

$$P_{ij}^{(n)} = P(X_{t+n} = j \mid X_t = i) \quad (4)$$

Dimana setiap elemen pada metrik $P_{ij}^{(n)}$ bernilai tak negatif (≥ 0) dikarenakan merupakan peluang bersyarat dan jumlah dari setiap barisnya bernilai sama dengan satu. Matriks peluang transisi *n*- langkah dapat dituliskan sebagai berikut:

$$P^{(n)} = \begin{bmatrix} P_{11}^{(n)} & P_{12}^{(n)} & P_{13}^{(n)} & \dots & P_{1m}^{(n)} \\ P_{21}^{(n)} & P_{22}^{(n)} & P_{23}^{(n)} & \dots & P_{2m}^{(n)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{m1}^{(n)} & P_{m2}^{(n)} & P_{m3}^{(n)} & \dots & P_{mm}^{(n)} \end{bmatrix} \quad (5)$$

State atau keadaan pada Rantai Markov dituliskan ke dalam bentuk vektor yang dinamakan vektor *state*. Vektor *state* untuk sebuah pengamatan pada suatu Rantai Markov dengan $X(t)$ *state* adalah sebuah vektor baris *x*. Sebuah vektor *state* dapat dituliskan sebagai:

$$x = [x_1, x_2, \dots, x] \quad (6)$$

Jika P merupakan matriks transisi Rantai Markov dan $x^{(n)}$ adalah vektor *state* pada pengamatan ke-*n*, maka

$$x^{(n)} = P^n x^0 \quad (7)$$

dimana:
$x^{(n)}$: keadaan/ *state* di waktu-n
P^n : matriks peluang transisi
x^0 : keadaan/ *state* di masa sekarang.

Proses Markov akan menuju kondisi stabil atau *steady state* jika setelah seluruh proses Markov berjalan beberapa periode, dan matriks Peluang akan selalu tetap. *Steady state* adalah istilah untuk menandai terjadinya keseimbangan antara dua kekuatan yang saling mencari kondisi yang saling menguntungkan bagi masing-masing. Dalam Rantai Markov, *steady state* menjelaskan bagaimana perubahan-perubahan variabel dalam sistem akan membawa *state* di masa yang akan datang tidak berubah-ubah lagi atau stabil. Secara matematis, jika $x^{(n)} = P^n x^0$ maka kondisi *steady state* terjadi saat $x^{(n\,steady)} = P^n x^{(n\,steady)}$.[8]

III. HASIL DAN PEMBAHASAN

Berdasarkan rancangan analisis dan data yang telah dijelaskan sebelumnya, berikut adalah hasil dan pembahasan dari penelitian mengenai analisis dan prediksi tingkat hemoglobin menggunakan rantai markov:

A. Pembentukan State dan Transisi

Pada tahap ini, state dalam rantai Markov dibentuk berdasarkan rentang nilai hemoglobin yang relevan. Dalam kasus ini, *state* terdiri dari rendah, normal, dan tinggi, dengan rentang nilai yang telah ditentukan. Rentang nilai yang ditetapkan adalah sebagai berikut:

"Rendah" didefinisikan sebagai nilai hemoglobin kurang dari 12 gram/dL (rendah < 12 gram/dL).

"Normal" didefinisikan sebagai rentang nilai hemoglobin antara 12 gram/dL hingga 15 gram/dL (12 gram/dL <= normal >= 15 gram/dL).

"Tinggi" didefinisikan sebagai nilai hemoglobin yang lebih tinggi dari 15 gram/dL (tinggi > 15 gram/dL).

Dengan mengaplikasikan batasan ini, data pengukuran hemoglobin dikategorikan secara jelas ke dalam *state-state* yang sesuai. Hal ini memfasilitasi analisis dalam konteks rantai markov, memungkinkan pemahaman perubahan keadaan berdasarkan data hemoglobin yang diperoleh, dan mendukung kesimpulan serta temuan dalam penelitian ini. Tabel 1 menampilkan hasil pembentukan *state* dan transisi yang terkait dengan kadar hemoglobin tersebut.

TABEL 1. PEMBENTUKAN STATE DAN TRANSISI

Nomor	*Pengecekan Hemoglobin*		
	Tanggal	*Hemoglobin*	*State*
1	21/07/2020	13,6	Normal
2	28/07/2020	13,4	Normal
3	11/08/2020	13,3	Normal
4	18/08/2020	12,8	Normal
5	19/08/2020	11,9	Rendah
6	25/08/2020	12,8	Normal
7	01/09/2020	13,4	Normal
8	08/09/2020	12,9	Normal
9	22/09/2020	12,2	Normal
0	30/09/2020	10,6	Rendah
11	13/10/2020	10,6	Rendah
12	13/10/2020	10,6	Rendah
13	08/12/2020	9,9	Rendah
14	15/12/2020	9,6	Rendah
15	22/12/2020	10,7	Rendah
16	04/01/2021	9,5	Rendah
17	25/01/2021	10,1	Rendah
18	30/03/2021	9,8	Rendah
19	26/04/2021	8,6	Rendah
20	01/05/2021	10,9	Rendah
21	17/05/2021	11,4	Rendah
22	28/05/2021	10,1	Rendah
23	08/06/2021	9,7	Rendah
24	12/06/2021	12,6	Normal
25	18/06/2021	12,5	Normal
26	25/06/2021	12,3	Normal
27	25/06/2021	12,3	Normal
28	14/10/2021	11,9	Rendah
29	06/01/2022	10,9	Rendah
30	06/01/2022	10,9	Rendah
31	31/03/2022	11,6	Rendah
32	02/06/2022	11,9	Rendah
33	28/06/2022	12,1	Normal
34	11/10/2022	12	Normal
35	30/03/2023	10,5	Rendah

B. Peluang Awal Setiap State

Untuk menghitung peluang awal setiap *state*, frekuensi kemunculan setiap *state* dalam data pengukuran kadar hemoglobin yang terkumpul dianalisis. Pada kasus ini, peluang awal *state* rendah adalah 0.5882 (58.82%), peluang awal *state* normal adalah 0.4118 (41.18%), dan peluang awal *state* tinggi adalah 0.0 (0%). Dalam konteks Rantai Markov, vektor *state*, sebagaimana yang ditunjukkan dalam Persamaan (6), adalah representasi keadaan atau status pada suatu waktu t dalam rangkaian pengamatan. Berdasarkan persamaan tersebut, dalam kasus ini, peluang awal setiap *state* direpresentasikan sebagai berikut:

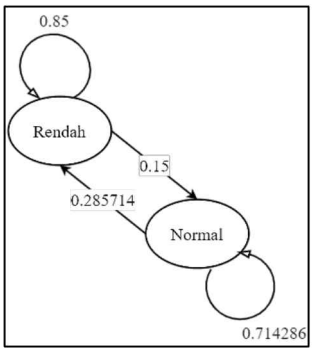
$$x^0 = [0.5882 \quad 0.4118 \quad 0.0] \tag{8}$$

C. Matriks Probabilitas

Matriks probabilitas transisi dibentuk berdasarkan probabilitas transisi yang telah diestimasi dari data. Seperti yang ditunjukkan Persamaan (3), matriks ini merepresentasikan probabilitas perpindahan dari satu state ke *state* lainnya dalam rantai markov. Pada kasus ini, hasil perhitungan matriks probabilitas transisi adalah sebagai berikut:

$$P^{(n)} = \begin{bmatrix} 0.85 & 0.15 & 0.0 \\ 0.285714 & 0.714286 & 0.0 \\ 0.0 & 0.0 & 0.0 \end{bmatrix} \tag{9}$$

Dari matriks probabilitas transisi persamaan (9), dapat divisualisasikan menggunakan gambar diagram transisi. Hasil visualisasi matriks probabilitas transisi dapat dilihat pada Gambar 2.



Gambar 2. Diagram Transisi

D. Analisis Steady State dan Prediksi

Dalam rangka mencari steady state dari Rantai Markov, Persamaan (7) menyatakan bahwa vektor *state* pada pengamatan ke-n, $x^{(n)}$, dapat dihitung dengan mengalikan matriks peluang transisi P^n dengan vektor *state* awal x^0.

Ketika steady state tercapai, akan terjadi $x^{(n\,steady)} = P^n x^{(n\,steady)}$, di mana $x^{(n\,steady)}$ adalah vektor *state* dalam *steady state*, dan P adalah matriks peluang transisi sesuai dengan yang dinyatakan dalam Persamaan (9). Vektor state awal, x^0, diberikan oleh Persamaan (8). Proses perhitungan *steady state* melibatkan perulangan perhitungan menggunakan matriks peluang transisi hingga mencapai situasi di mana $x^{(n)}$ tidak berubah lagi. Pada titik ini, kondisi keseimbangan atau *steady state* dalam Rantai Markov telah tercapai. Hasil penghitungan *steady state* dapat dilihat pada Tabel 2.

TABEL 2. PENGHITUNGAN STEADY STATE

Iterasi	*State*		
	Rendah	*Normal*	*Tinggi*
0	0.588200	0.411800	0.0
1	0.617621	0.382379	0.0
2	0.634224	0.365776	0.0
3	0.643592	0.356408	0.0
4	0.648879	0.351121	0.0
5	0.651863	0.348137	0.0
6	0.653546	0.346454	0.0
7	0.654496	0.345504	0.0
8	0.655032	0.344968	0.0
9	0.655335	0.344665	0.0
10	0.655505	0.344495	0.0
11	0.655602	0.344398	0.0
12	0.655656	0.344344	0.0
13	0.655687	0.344313	0.0
14	0.655704	0.344296	0.0
15	0.655714	0.344286	0.0
16	0.655719	0.344281	0.0
17	0.655722	0.344278	0.0
18	0.655724	0.344276	0.0
19	0.655725	0.344275	0.0
20	0.655726	0.344274	0.0
21	0.655726	0.344274	0.0
22	0.655726	0.344274	0.0
23	0.655726	0.344274	0.0
24	0.655726	0.344274	0.0
25	0.655726	0.344274	0.0
26	0.655726	0.344274	0.0
27	0.655726	0.344274	0.0
28	0.655726	0.344274	0.0
29	0.655726	0.344274	0.0

Hasil perhitungan steady state pada Tabel 2 menunjukkan bahwa vektor state dalam *steady state*, $x^{(n\,steady)}$, telah konvergen ke nilai tertentu setelah sejumlah iterasi. Nilai-nilai dalam vektor *state* adalah 0.655726 untuk state "Rendah" dan 0.344274 untuk *state* "Normal," sementara *state* "Tinggi" tetap pada 0.0. Ini mengindikasikan bahwa sistem telah mencapai kondisi keseimbangan di mana probabilitas berada di antara dua *state* pertama (Rendah dan Normal), sementara probabilitas berada pada 0.0 untuk *state* "Tinggi." Hal ini mengindikasikan bahwa tidak ada kemungkinan bahwa wanita tersebut akan berada dalam *state* tinggi pada masa depan.

Hasil perhitungan *steady state* ini adalah representasi dari kondisi keseimbangan dalam rantai markov, di mana probabilitas transisi antara *state-state* tertentu tidak lagi berubah. Ini adalah informasi yang penting dalam analisis rantai markov, dan hasilnya dapat digunakan untuk membuat prediksi atau evaluasi lebih lanjut terkait dengan prediksi tingkat hemoglobin dilakukan menggunakan pendekatan rantai markov. Dengan memodelkan perpindahan antara *state-state* yang dibentuk berdasarkan rentang nilai hemoglobin, penelitian ini dapat memberikan wawasan tentang perilaku perubahan tingkat hemoglobin pada wanita tersebut.

Dengan memanfaatkan distribusi probabilitas *steady state* ini, penelitian ini juga dapat memberikan prediksi tentang tingkat hemoglobin di masa depan. Misalnya, jika pada saat ini wanita tersebut berada dalam *state* rendah, prediksi menunjukkan bahwa kemungkinan besar tingkat hemoglobinnya akan tetap rendah dalam jangka panjang. Namun, jika pada saat ini wanita tersebut berada dalam *state* normal, prediksi menunjukkan bahwa kemungkinan besar tingkat hemoglobinnya akan tetap normal dalam jangka panjang.

Penelitian ini memberikan dasar yang kuat untuk melakukan analisis dan prediksi tingkat hemoglobin menggunakan pendekatan rantai markov. Namun, perlu diperhatikan bahwa hasil dan prediksi yang diberikan bergantung pada data yang digunakan dan asumsi yang dibuat dalam pembentukan model rantai markov. Oleh karena itu, penelitian ini dapat diperluas dan ditingkatkan dengan mengumpulkan data yang lebih banyak dan melibatkan faktor-faktor lain yang dapat mempengaruhi tingkat hemoglobin.

IV. KESIMPULAN

Pendekatan menggunakan rantai Markov dalam analisis dan prediksi tingkat hemoglobin pada wanita pengidap kanker payudara memberikan wawasan tentang perilaku perubahan tingkat hemoglobin. Hasil analisis *steady state* menunjukkan bahwa wanita tersebut kemungkinan besar akan berada dalam *state* rendah atau normal di masa depan. Prediksi berdasarkan distribusi probabilitas *steady state* juga memberikan gambaran tentang tingkat hemoglobin di masa depan, di mana jika wanita tersebut saat ini berada dalam *state* rendah, tingkat hemoglobinnya kemungkinan besar akan tetap rendah dalam jangka panjang, sedangkan jika berada dalam state normal, tingkat hemoglobinnya kemungkinan besar akan tetap normal dalam jangka panjang. Meskipun penelitian ini memberikan dasar yang kuat, penting untuk mencatat bahwa hasil dan prediksi tergantung pada data yang digunakan dan asumsi yang dibuat dalam model rantai Markov. Oleh karena itu, penelitian ini dapat diperluas dengan melibatkan lebih banyak data dan faktor-faktor lain yang mempengaruhi tingkat hemoglobin.

REFERENSI

[1] V. Lim, W. E. Haryanto and Salvirah, "Mengenal Kadar Normal Hemoglobin dan Fungsinya dalam Tubuh," Siloam Hospital, 8 November 2023. [Online]. Available: https://www.siloamhospitals.com/informasi-siloam/artikel/kadar-hemoglobin-normal. [Accessed 17 November 2023].

[2] F. A. Kurniawan, "Aplikasi Markov Chain Untuk Memprediksi Tekanan Darah," *IncomTech,* vol. 8, no. 2, pp. 103-120, 2018.

[3] U. Wiwi and W. E. Maryati, "Aplikasi Metode Markov Chain untuk Meningkatkan Tingkat Persediaan Bahan Baku yang Optimal," *Optimumm,* vol. 1, no. 1, pp. 80-90, 2000.

[4] S. S, I. S and Sukarna, "Aplikasi Analisis Rantai Markov untuk Mempredikdi Status Pasien Rumah Sakit Umum Daerah Kabupaten Barru," *Online Jurnal of Natural Science,* vol. 3, no. 3, pp. 313 - 321, 2014.

[5] A. Sazali, B. H. Setiadji and B. Haryadi, "Aplikasi Model Rantai Markov Dalam Pengelolaan Jalan," *Rekayasa,* vol. 12, no. 2, pp. 141-150, 2019.

[6] E. Tandelilin, "Aplikasi Rantai Markov Untuk Perencanaan dan Pengendalian Penjualan," *Jurnal Ekonomi dan Bisnis Indonesia,* vol. 4, 1989.

[7] H. Sarjono, Edwin, H. Sentosa and F. Bong, "Analisis Markov Chain Terhadap Persediaan: Studi Kasus Pada CV Sinar Bahagia Group," *Binus Business Review,* pp. 1071-1076, 2011.

[8] Z. M. Kesuma, S. Rusdiana, L. Rahayu and E. Fradinata, Pengantar Biostatistika dan Aplikasinya Pada Status Kesehatan Gizi Remaja, Banda Aceh: Syiah Kuala University Press Darussalam, 2018.

Analisis Akseptabilitas Teknologi Augmented Reality pada Furnitur Rotan menggunakan *Technology Acceptance Model*

Muhammad Nurjaman
Departemen Informatika
Universitas Teknologi Yogyakarta
Sleman, Indonesia
mnurjaman035@gmail.com

Tabia Hanural
Departemen Informatika
Universitas Teknologi Yogyakarta
Sleman, Indonesia
tabiarema@gmail.com

Muhammad Zakariyah
Departemen Informatika
Universitas Teknologi Yogyakarta
Sleman, Indonesia
muhammad.zakariyah@staff.uty.ac.id

Abstrak — Fokus utama pada penelitian ini adalah untuk mengkaji faktor-faktor yang mempengaruhi penerimaan pengguna terhadap penerapan teknologi *Augmented Reality* (AR) pada produk rotan. Pada penelitian ini, *Technology Acceptance Model* (TAM) digunakan sebagai *framework* dasar dalam melakukan pengujian penerimaan pengguna. Data yang digunakan berjumlah 92, yang dikumpulkan dari para pengunjung yang telah menggunaan aplikasi AR rotan selama kegiatan pameran berlangsung. Hasil analisis regresi menunjukkan bahwa faktor-faktor dalam TAM secara signifikan menjelaskan variasi dalam persepsi kemudahan penggunaan (*Perceive Ease of Use*/PEU), kegunaan (*Perceive Usefulness*/PU) dan niat pengguna (*Behavior Intention to Use*/BIU). PEU secara signifikan merupakan faktor yang memengaruhi PU. Demikian halnya dengan PEU dan PU, keduanya terbukti sebagai prediktor BIU untuk mengadopsi AR untuk produk rotan. Temuan ini memiliki implikasi penting bagi pelanggan, industri rotan, dan para peneliti yang bergerak di bidang model penerimaan teknologi AR.

Kata Kunci — Augmented Reality, Technology Acceptance Model, Rotan.

I. PENDAHULUAN

Industri rotan telah lama menjadi bagian penting dari sektor manufaktur di Indonesia. Rotan, sebagai bahan baku utama, memiliki karakteristik yang unik karena kekuatannya, fleksibilitas, dan keindahan alaminya [1]. Produk rotan yang beragam, mulai dari furnitur hingga kerajinan tangan, telah menjadi daya tarik bagi konsumen di seluruh dunia. Industri rotan telah menggunakan berbagai metode pemasaran, termasuk pameran, *showroom*, serta pemasaran konvensional melalui media cetak dan digital [2].

Pameran industri rotan seringkali menjadi salah satu *platform* utama bagi para produsen dan pengekspor untuk memamerkan produk kepada calon pelanggan. Pameran ini tidak hanya menyediakan kesempatan bagi para pelaku industri untuk memperluas jaringan bisnis, tetapi juga memberikan kesempatan bagi konsumen untuk melihat dan mencoba produk secara langsung. Di sisi lain, *showroom* merupakan tempat di mana produk rotan secara langsung dipajang dan ditampilkan kepada konsumen. Pameran dan *showroom* juga sering diintegrasikan dengan platform media sosial dan situs *web* untuk memperluas jangkauan pemasaran.

Pemasaran konvensional melalui media cetak dan digital telah menjadi salah satu pendekatan yang umum di industri rotan. Media cetak, seperti katalog dan brosur, sering digunakan untuk memberikan informasi detail mengenai berbagai produk rotan yang tersedia. Di era digital, pemasaran melalui situs *web*, *platform e-commerce*, dan media sosial juga semakin populer dalam mempromosikan produk rotan kepada pasar global. Penekanan pada visualisasi yang menarik dan deskripsi produk yang komprehensif menjadi kunci untuk menarik minat konsumen dan mempertahankan daya saing di pasar yang semakin kompetitif [3].

Augmented Reality (AR) merupakan teknologi yang menempatkan informasi digital seperti gambar, video, atau model 3D di atas lingkungan dunia nyata. AR mengintegrasikan elemen virtual dengan dunia fisik. Salah satu keunggulan utama AR adalah kemampuannya untuk memperkaya pengalaman pengguna dengan memberikan informasi tambahan yang berguna [4]. Penggunaan AR menawarkan berbagai potensi aplikasi di berbagai bidang, termasuk pendidikan [5], pemasaran [6], permainan [7], dan pelatihan industri [8].

Pada bidang pemasaran, AR telah membuka berbagai peluang untuk menciptakan pengalaman yang lebih menarik dan interaktif bagi konsumen. Perusahaan dapat memperkenalkan produk secara lebih detail, memungkinkan konsumen untuk melihat dan menguji produk secara virtual sebelum melakukan pembelian, dan memberikan informasi tambahan yang berguna tentang produk yang ditawarkan.

Industri rotan telah mengalami perkembangan yang signifikan, khususnya dalam hal inovasi teknologi. Pada era digital saat ini, konsep pemasaran yang inovatif menjadi semakin krusial bagi pertumbuhan dan kelangsungan industri tersebut [9]. Salah satu inovasi terbaru yang menarik perhatian pada bidang pemasaran adalah penggunaan augmented reality (AR). Teknologi ini dapat digunakan untuk memasarkan produk rotan, terutama pada saat diselenggarakannya pameran kerajinan rotan tanpa harus menghadirkan fisik dari produk tersebut. Melalui penggabungan dunia fisik dan virtual, AR telah menunjukkan potensinya untuk meningkatkan interaksi pelanggan, menciptakan pengalaman yang menarik, dan mendorong daya saing di pasar global [10].

Penelitian sebelumnya telah menunjukkan pengaruh faktor-faktor seperti kepuasan, norma sosial, dan faktor psikologis lainnya dalam penerimaan AR [11], [12], [13]. Namun, untuk produk rotan khususnya, penelitian yang mengkaji penerimaan teknologi AR masih terbatas. *Technology Acceptance Model* (TAM) merupakan sebuah kerangka kerja konseptual yang digunakan untuk memahami

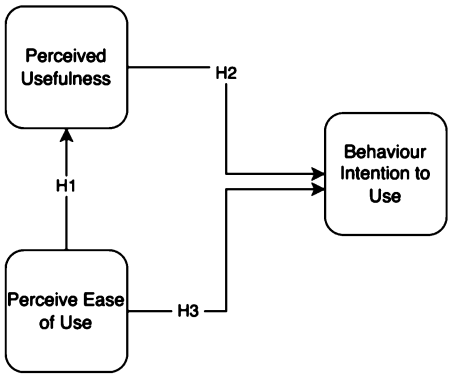
dan menganalisis faktor-faktor yang memengaruhi penerimaan dan adopsi teknologi oleh pengguna. Model ini berfokus pada sikap dan perilaku pengguna terhadap teknologi baru.

Penelitian ini bertujuan untuk mengetahui penerimaan pengguna terhadap penggunaan AR dalam industri rotan dengan menggunakan TAM. Penerapan TAM untuk mengkaji penerimaan teknologi AR pada produk rotan akan memberikan kontribusi yang signifikan dalam memahami dinamika perilaku konsumen dan mengoptimalkan strategi pemasaran dalam industri ini.

II. METODE PENELITIAN

A. *Teknologi Acceptance Model*

Technology Acceptance Model (TAM) yang dibuat oleh Davis [14], merupakan suatu model yang dikembangkan dari *Theory of Reasoned Action* (TRA) untuk menjelaskan bagaimana pengguna menerima teknologi. Model TAM ditunjukkan pada Gambar 1.



Gambar 1. Technology Acceptance Model

Model TAM berasumsi bahwa niat individu dipengaruhi oleh sikap pribadi terhadap penggunaan teknologi *Behaviour Intention to Use* (BIU). Selain itu, di dalam model ini, terdapat dua atribut lain yang memengaruhi sikap seseorang [14]. Atribut tersebut adalah kegunaan yang dirasakan (*Perceive Usefulness*/PU) yang berarti sejauh mana seseorang percaya bahwa menggunakan sistem tertentu akan meningkatkan kinerja pekerjaannya. Atribut kemudahan penggunaan yang dirasakan (*Perceive Ease of Use*/PEU) yaitu sejauh mana individu percaya bahwa menggunakan sistem tertentu akan mudah dilakukan tanpa usaha fisik dan mental yang berlebihan. Selanjutnya, PEU memengaruhi PU, dan PU memengaruhi niat perilaku. Sementara itu, penelitian empiris menunjukkan bahwa TAM dapat digunakan tanpa memasukkan atribut sikap terhadap penggunaan [15], [16].

TABEL 1. HIPOTESIS PENGGUNAAN AR PADA PRODUK ROTAN

Kode	Hipotesis Awal
H1	Kemudahan penggunaan (PEU) yang dirasakan oleh pelanggan produk rotan tidak memiliki efek positif pada kegunaan yang dirasakannya (PU).
H2	Kegunaan (PU) yang dirasakan oleh pelanggan produk rotan tidak memiliki efek positif pada niat untuk menggunakannya (BIU).
H3	Kemudahan penggunaan (PEU) yang dirasakan oleh pelanggan produk rotan tidak memiliki efek positif pada niat perilaku untuk menggunakannya (BIU).

Dari tabel 1 dapat diketahui bahwasannya atribut independen PEU dan PU secara langsung memengaruhi niat perilaku (BIU). Seiring berjalannya waktu, TAM menjadi model kunci untuk memprediksi perilaku manusia dalam hal akseptabilitas teknologi [14].

B. *Hypothesis*

Untuk menguji akseptabilitas aplikasi *Augmented Reality* pada furnitur rotan, diperlukan variabel independen. Oleh karena itu, model TAM pada Tabel 1 merupakan rumusan hipotesis tentang analisis TAM pada aplikasi AR di industri furnitur rotan. Variabel independen terdiri atas Kemudahan penggunaan (PEU) dan Kegunaan (PU), sedangkan variabel dependen adalah niat perilaku untuk menggunakannya (BIU).

C. *Responden*

Studi ini melibatkan 92 peserta, yang terdiri dari 61 laki-laki dan 31 perempuan. Peserta adalah pengunjung yang hadir selama 4 hari di pameran yang diselenggarakan di Jakarta International Expo pada tanggal 9-12 Maret 2023. Survei dilakukan secara daring melalui Google Form untuk menjaga kerahasiaan peserta dan meningkatkan kualitas tanggapan [24].

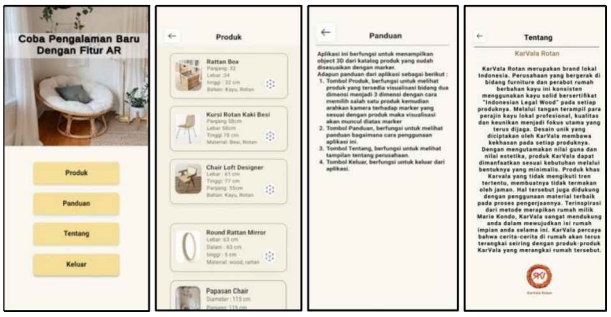
D. *Kuisioner*

Kuesioner untuk penelitian ini terbagi menjadi dua bagian. Bagian pertama difokuskan pada pengumpulan informasi demografis, seperti usia dan jenis kelamin. Bagian kedua terdiri dari 12 item yang menilai 3 komponen yang diajukan dalam model TAM yang diusulkan. Item-item ini disajikan dalam bahasa Indonesia seperti yang ditunjukkan dalam Tabel 2. Responden menggunakan skala Likert lima poin, mulai dari 1 = Sangat Tidak Setuju hingga 5 = Sangat Setuju, untuk mengekspresikan tingkat persetujuan atau ketidaksetujuan terhadap setiap item.

E. *Prosedur*

Tamu yang mengunjungi stan produk rotan dalam pameran diminta untuk mencoba aplikasi *Augmented Reality* furnitur rotan (Gambar 2). Setelah menggunakan aplikasi tersebut, pengunjung diminta untuk berpartisipasi dalam sebuah studi penelitian yang bertujuan untuk memahami faktor-faktor yang memengaruhi penerapan AR di industri rotan, berdasarkan perspektif pengunjung. Bagi pengunjung yang bersedia (selanjutnya disebut sebagai responden), dapat membuka tautan Google Form melalui Kode QR yang telah diberikan.

Responden akan diinformasikan pengenalan singkat yang menjelaskan tujuan dari penelitian dan pentingnya berpartisipasi dalam penelitian tersebut. Responden akan



Gambar 2. Antarmuka Aplikasi Augmented Reality Furnitur Rotan

diminta untuk memberikan persetujuan untuk ikut serta dalam penelitian ini, sebelum melanjutkan ke serangkaian pertanyaan berikutnya. Google Form akan meminta para responden untuk memberikan beberapa informasi dasar mengenai identitas, seperti usia, jenis kelamin, pekerjaan, dan informasi dasar lainnya. Meskipun demikian, data pribadi seperti nama, nomor kontak, dan data pribadi lainnya tidak akan diminta dalam penelitian ini. Mengisi informasi dasar ini penting untuk keperluan analisis data.

Kemudian, responden akan diminta untuk memberikan pendapat mengenai sejauh mana pemahaman terhadap AR dan pengalaman dalam menggunakan teknologi ini. Responden akan diminta untuk berbagi pandangan tentang penggunaan AR pada produk rotan. Pertanyaan-pertanyaan ini akan berfokus pada preferensi, persepsi, dan pengalaman terkait produk anyaman rotan yang terintegrasi dengan teknologi AR. Kuesioner mencakup pertanyaan-pertanyaan yang berkaitan dengan faktor-faktor yang memengaruhi adopsi teknologi AR dalam produk rotan. Informasi ini mencakup pertanyaan mengenai kegunaan dan kemudahan penggunaan teknologi AR, serta niat untuk menggunakannya.

Terakhir, responden akan diarahkan ke bagian kesimpulan untuk memberikan *feedback* (umpan balik) atau komentar tambahan. Setelah menyelesaikan kuesioner, responden dapat menutup Google Form, dan penyelenggara penelitian akan mengumpulkan data tersebut untuk analisis lebih lanjut.

TABEL 2. DESKRIPTIF DAN HASIL UJI RELIABILITAS

Komponen TAM	Mean	SD	Cronbach's α
Perceived Ease of Use (PEU)	**3,23**	**0,77**	**0,859**
Saya mudah mempelajari cara menggunakan aplikasi AR rotan	3,20	1,28	0,868
Saya dapat berinteraksi dengan aplikasi AR rotan dengan jelas	3,32	1,24	0,875
Saya dapat memahami dengan baik cara berinteraksi dengan AR rotan	3,33	1,14	0,873
Saya beranggapan bahwa aplikasi AR rotan merupakan program yang fleksibel	3,08	1,28	0,869
Saya dapat menggunakan aplikasi AR rotan dengan mudah	3,26	1,14	0,871
Perceived Usefulness (PU)	**3,40**	**0,86**	**0,859**
Saya beranggapan bahwa pekerjaan saya menjadi lebih mudah dengan aplikasi AR rotan	3,47	1,20	0,870
Saya dapat meningkatkan efektifitas kerja dengan aplikasi AR rotan	3,52	1,26	0,868
Saya beranggapan bahwa aplikasi AR dapat berguna bagi saya	3,37	1,23	0,868
Saya terbantu dengan mendapatkan informasi furniture melalui AR rotan	3,24	1,20	0,870
Behaviour Intention to Use (BIU)	**3,28**	**0,87**	**0.862**
Saya nyaman menggunakan aplikasi AR rotan	3,24	1,10	0,866
Saya menikmati penggunaan aplikasi AR rotan	3,20	1,28	0,870
Saya beranggapan bahwa aplikasi Ar rotan tidak membosankan	3,39	1,19	0,877

F. Analisis

Data dianalisis menggunakan perangkat lunak SPSS versi 24. Pada awalnya, dilakukan analisis faktor, dan hasilnya mendukung gagasan bahwa dari dua variabel dalam TAM memiliki satu struktur tunggal. Selanjutnya, konsistensi internal variabel TAM dievaluasi menggunakan Koefisien *Cronbach's Alpha*. Setelah penilaian ini, analisis deskriptif dilakukan, termasuk perhitungan standar deviasi untuk tiga variabel dan item-itemnya. Untuk menyelidiki keberadaan korelasi yang signifikan secara statistik di antara variabel TAM, dilakukan analisis Korelasi Pearson. Selanjutnya, tiga hipotesis yang diuraikan dalam bagian sebelumnya diperiksa melalui analisis *regresi linier.*

III. HASIL DAN PEMBAHASAN

A. Analisis Deskriptif dan Hasil Uji Reabilitas

Tabel 2 menunjukkan rata-rata (Mean) dan standard deviasi (SD) untuk tiga komponen dan 12 item individualnya. Nilai rata-rata untuk ketiga komponen berkisar dari 3,23 (Perceive Ease of Us) hingga 3,40 (Perceive Usefuless). Hasil temuan menunjukkan bahwa semua Komponen TAM menunjukkan tingkat moderat.

Seperti yang terlihat di Tabel 2, angka Cronbach's alpha untuk setiap dari tiga variabel TAM semuanya lebih dari 0.80, menunjukkan bahwa kuesioner dengan tiga variabel ini dapat diandalkan dan valid.

B. Hasil Pearson Correlations

Tabel 3 menyajikan gambaran tentang Korelasi Pearson antara semua komponen TAM. Tabel ini menunjukkan bahwa koefisien korelasi untuk ke-tiga komponen TAM adalah signifikan secara statistik, yang mengkonfirmasi 3 hipotesis dalam penelitian ini.

TABEL 3. HASIL UJI KORELASI PEARSON

	PEU	PU	BIU
PEU	1	0,639** (< 0,0001)	0,511** (< 0,0001)
PU		1	0,477** (< 0,0001)
BIU			1

**. Signifikan pada level 0,01 (2-arah).

Selain itu, terdapat hubungan yang signifikan antara semua komponen. Secara khusus, P*erceived Ease of Use* (PEU) memiliki korelasi yang kuat dan signifikan dengan komponen *Perceive Usefulness* (PU) dalam TAM, dengan nilai korelasi 0,639. Sementara itu, PU menunjukkan korelasi yang relatif rendah dengan *Behaviour Intention of Use* (BIU), dengan nilai korelasi 0,477. Secara khusus, PEU berkorelasi kuat dan signifikan dengan komponen PU, serta BIU dengan nilai 0,639 dan 0,511 secara berurutan.

C. Uji Hipotesis

Analisis regresi dilakukan untuk menguji 3 hipotesis yang diusulkan pada penelitian ini (Tabel 1). Hasil analisis regresi terlihat dalam Tabel 4. Berdasarkan tabel tersebut, niat pengunjung (BIU) untuk menggunakan AR pada produk rotan diprediksi oleh kemudahan yang dirasakan (PEU) dengan nilai beta = 0,349, (*p-value* < 0,01) dan kegunaan (PU) dengan nilai beta = 0,254 (*p-value* < 0,05). Bersama-sama, kedua

variabel independen ini berkontribusi sebesar 28,3% dari varians total, terhadap BIU.

Hasil analisis regresi ini juga mengkonfirmasi bahwa, aspek kemudahan yang dirasakan oleh pelanggan produk rotan (PEU), memiliki efek positif pada kegunaan (PU) dengan tingkat kontribusi sebesar 40,2%.

TABEL 4. HASIL UJI REGRESI

Dep. Var.	**Indep. Var.**	**Adjusted R^2**	**F**	**Beta**	**t**	**p-value**
PU	PEU	0,402	62,085	0,639	7,879	**< 0,001****
BIU	PEU	0,283	18,992	0,349	3,024	**0,003****
	PU			0,254	2,199	**0,030***

*. Signifikan pada level 0,05 (2-arah).
**. Signifikan pada level 0,01 (2-arah).

IV. KESIMPULAN

Berdasarkan penelitian pada penerapan TAM dalam pengujian model akseptabilitas, dapat diambil kesimpulan persepsi kemudahan pengguna (*Perceived Ease of Use*/PEU) memiliki nilai *pearson correlation* 0,639 yang menyatakan bahwa pengunjung setuju dengan kegunaan penggunaan aplikasi AR furnitur rotan (*Perceived of Usefulness*/PU). Faktor PEU dan PU secara signifikan berkontribusi sebesar 28,3% terhadap niat pengunjung untuk menggunakan AR (BIU). Hasil analisis statistik deskriptif pada TAM memiliki variabel dengan nilai rata-rata paling kecil yaitu kemudahan penggunaan (PEU) dengan nilai rata-rata 3,23 (moderat). Hal ini dapat digunakan untuk meningkatkan pengembangan AR dengan berfokus pada kemudahan dalam menggunakan AR pada industri rotan.

UCAPAN TERIMAKASIH

Kami mengucapkan terima kasih kepada rekan-rekan penelitian di Universitas Teknologi Yogyakarta (UTY) atas dukungan luar biasa dalam menyelesaikan penelitian kami tentang teknologi Augmented Reality (AR) untuk furnitur rotan. Kami juga berterima kasih atas bimbingan dan kontribusi berharga dari berbagai pihak. Keberhasilan penelitian ini tak terpisahkan dari dedikasi dan kerja keras tim kami. Terima kasih.

REFERENSI

[1] S. Sakinah, B. Afriyansyah, and D. Akbarini, "Etnobotani Rotan Sebagai Bahan Kerajinan Ayaman Oleh Masyarakat Di Kabupaten Bangka Barat," *Al-Kauniyah: Jurnal Biologi*, vol. 12, no. 1, pp. 18–24, Apr. 2019, doi: 10.15408/kauniyah.v12i1.6429.

[2] S. Pemasaran Kerajinan Rotan Desa Sungai Baung Kecamatan Rawas Ulu, K. Musi Rawas Utara, H. Iswarini, P. Pratami Ardina Ningrum, and H. Noval Ibrahim, "Marketing Strategy for Rattan Crafts in Sungai Baung Village, Rawas Ulu District, North Musi Rawas Regency," *Journal of Global Sustainable Agriculture*, vol. 3, no. 2, pp. 25–30, 2023, doi: 10.32502/jgsa.v3i2.6393.

[3] S. Hendricks and S. D. Mwapwele, "A systematic literature review on the factors influencing e-commerce adoption in developing countries," *Data Inf Manag*, p. 100045, Jul. 2023, doi: 10.1016/j.dim.2023.100045.

[4] Z. Du, J. Liu, and T. Wang, "Augmented Reality Marketing: A Systematic Literature Review and an Agenda for Future Inquiry," *Front Psychol*, vol. 13, Jun. 2022, doi: 10.3389/fpsyg.2022.925963.

[5] K. Nistrina, "Penerapan Augmented Reality Dalam Media Pembelajaran," 2021.

[6] S. Puspita Sari *et al.*, "Yumary: Jurnal Pengabdian Kepada Masyarakat Peran Augmented Reality dan Mobile Marketing dalam Meningkatkan Promosi Bisnis (The Role of Augmented Reality and Mobile Marketing in Enhancing Business Promotion) Riwayat Artikel," vol. 3, no. 4, pp. 191–199, 2023, doi: 10.35912/jpm.v3i4.1725.

[7] D. Huamanchahua, G. Diaz-Beltran, A. G. Trinidad-Palacios, and C. M. Pajuelo-Ventocilla, "Implementation of an Augmented Reality Application to Improve Learning in Elementary School Children through Playful Games," in *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, IEEE, Oct. 2022, pp. 0088–0094. doi: 10.1109/UEMCON54665.2022.9965646.

[8] G. Trinanda and S. L. Rahayu, "Aplikasi Augmented Reality Pengenalan Komponen Sepeda Motor Berbasis Android," *Jurnal Teknologi Informasi*, vol. 4, no. 1, 2023, doi: 10.46576/djtechno.

[9] I. Alotaibi, "An Exploratory Study of Augmented Reality Marketing in UAE," in *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, IEEE, Apr. 2021, pp. 271–272. doi: 10.1109/CAIDA51941.2021.9425115.

[10] P. Wibowo Putro, M. Yoga Aditia, A. Eko Sujianto, and U. Islam Negeri Sayyid Ali Rahmatullah Tulungagung, "Jurnal Studi Manajemen dan Bisnis Teknologi Augmented Reality sebagai Strategi Pemasaran Syariah di Era Digitalisasi," *JSMB*, vol. 10, no. 1, pp. 2023–2042, doi: 10.21107/jsmb.v10i1.20442.

[11] N. Marangunić and A. Granić, "Technology acceptance model: a literature review from 1986 to 2013," *Univers Access Inf Soc*, vol. 14, no. 1, pp. 81–95, Mar. 2015, doi: 10.1007/s10209-014-0348-1.

[12] S. Taha, E. Abulibdeh, E. Zaitoun, S. Daoud, and H. G. Rawagah, "Investigating Student Perceptions of Augmented Reality Utilizing Technology Acceptance Model (TAM)," in *2022 International Arab Conference on Information Technology (ACIT)*, IEEE, Nov. 2022, pp. 1–7. doi: 10.1109/ACIT57182.2022.9994196.

[13] J. Jang, Y. Ko, W. S. Shin, and I. Han, "Augmented Reality and Virtual Reality for Learning: An Examination Using an Extended Technology Acceptance Model," *IEEE Access*, vol. 9, pp. 6798–6809, 2021, doi: 10.1109/ACCESS.2020.3048708.

[14] Davis FD., "A technology acceptance model for empirically testing new end-user in-formation systems-theory and results. Massachusetts; 1986.".

[15] L. Yang, S. Bashiru Danwana, and I. F. Yassaanah, "An Empirical Study of Renewable Energy Technology Acceptance in Ghana Using an Extended Technology Acceptance Model," *Sustainability*, vol. 13, no. 19, p. 10791, Sep. 2021, doi: 10.3390/su131910791.

[16] T. Teo and J. Noyes, "An assessment of the influence of perceived enjoyment and attitude on the intention to use technology among pre-service teachers: A structural equation modeling approach," *Comput Educ*, vol. 57, no. 2, pp. 1645–1653, Sep. 2011, doi: 10.1016/j.compedu.2011.03.002.

Penerapan Metode SAW untuk Rekomendasi Pengajuan Daftar Urut Kepangkatan bagi Pegawai Balmon Kota Palangkaraya

Muhammad Yusrif
Departemen Teknik Informatika
STMIK Palangkaraya
Kalimantan Tengah, Indonesia
Yusrief72@gmail.com

Suparno
Departemen Manajemen Informatika
STMIK Palangkaraya
Kalimantan Tengah, Indonesia
Endustong@gmail.com

Muhammad Qomaruzaman Haris
Departemen Ilmu Komputer
Universitas Muhammadiyah Palangkaraya
Kalimantan Tengah, Indonesia
harisqamaruzzaman@yahoo.co.id

Rudi Setiawan
Departemen Sistem Informasi
Universitas Trilogi Jakarta
Jakarta, Indonesia
Rudi@trilogi.ac.id

Abstract— **This research was purposes to provide recommendations for submitting a Daftar Urut Kepangkatan (DUK). DUK is a very important thing in Human Resource Management. In order to ensure objectivity in developing civil servants carrier, it needs to be done based on a career and work performance system. The problem raised in this research was how to implement the SAW (Simple Additive Weighting) method to provide recommendations for submitting DUK at the Center Office of Radio Frequency Spectrum Monitoring Palangka Raya Class II. The aim of this research was to test the SAW method by using 20 data. This data comes from 20 employees' data of the Center Office of Radio Frequency Spectrum Monitoring Palangka Raya Class II. The criteria used to solve this problem are 1). Rank and class (20) Job training (3) Position (5) Years of service (6) Education (7) Age. Based on the results of the implementation, it is known that SAW method can provide good recommendations. So, it can be concluded that the SAW method can be used effectively in providing recommendations for submitting DUK at the Center Office of Radio Frequency Spectrum Monitoring Palangka Raya Class II. This method is able to combine various relevant criteria and provide an objective ranking based on the relative weight of each criterion.**

Keywords— SAW, Daftar Urut Kepangkatan (DUK)

I. LATAR BELAKANG

Daftar Urut Kepangkatan (DUK) sangat penting dalam kepegawaian. Dalam rangka usaha untuk lebih menjamin objektivitas dan pembinaan pegawai negeri sipil berdasarkan sistem karir dan sistem prestasi kerja, maka perlu dibuat Daftar Urut Kepangkatan Pegawai Negeri Sipil. Apabila ada kekosongan jabatan, maka pegawai negeri sipil yang menduduki Daftar Urut Kepangkatan (DUK) yang lebih tinggi, haruslah dipertimbangkan lebih dahulu [5].

Saat ini pengambilan keputusan sering kali didasarkan pada kedekatan seseorang untuk dihalukan, namun dengan metode *Simple Additive Weighting* (SAW) diharapkan mampu meberikan rekomendasi yang adil dan sesuai dengan ketentuan yang berlaku. Selain itu system yang diterapkan saat ini kurang akurat dalam pegelaolaan datanyan karena data tersimpan secara arsip manual/*hard copy.*

Setiap tahun usulan DUK diusulkan dan dibuat oleh bagian kepegawaian untuk mendapatkan kepangkatan ke pangkat yang leih tinggi, hal tersebut akan mengakibatkan kesulitan bagi staf yang harus melakukan penilaian terhadap data. Maka bersama penelitian ini diberikan cara untuk menentukan sebuah DUK sehingga mendapatkan urutan yang tepat dan adil.

Perlunya Sistem Pendukung keputusan untuk menentukan DUK pegawai pada Kantor Balai Monitor Spektrum Frekuensi Radio Kelas II Palangka Raya adalah untuk memberikan rekomendasi kepada pimpinan dalam pengambilan keputusan.

II. METODE PENELITIAN

Penelitian ini mengadopsi pendekatan kuantitatif dengan metode deskriptif-analitik. Pendekatan kuantitatif dipilih karena penelitian ini bertujuan untuk mengukur dan menganalisis hubungan antara variabel-variabel yang telah ditetapkan dalam penilaian kepangkatan. Metode deskriptif-analitik digunakan untuk mendeskripsikan dan menganalisis fenomena atau peristiwa yang terjadi dalam konteks pengajuan daftar urut kepangkatan di Kantor Balai Monitor Spektrum Frekuensi Radio Kelas II Palangka Raya.

Dalam jenis penelitian kuantitatif, data yang dikumpulkan berupa angka atau data berbasis angka yang dapat dianalisis secara statistik. Penerapan Metode SAW dalam rekomendasi pengajuan daftar urut kepangkatan akan melibatkan penghitungan matematis dan analisis statistik terhadap data kriteria yang telah ditentukan. Metode SAW merupakan metode yang menggunakan perhitungan matematis untuk memberikan peringkat pada alternatif berdasarkan kriteria-kriteria yang ada. Pada tahap analisis, data hasil perhitungan akan dianalisis secara statistik untuk menghasilkan peringkat alternatif. Dalam metode ini, penelitian akan mencari alternatif yang memiliki nilai terbobot tertinggi sesuai dengan kriteria yang ditetapkan. Oleh karena itu, analisis data dalam penelitian ini akan fokus pada perhitungan nilai terbobot, normalisasi data, dan perangkingan alternatif.

A. Pengumpulan Data

Penelitian ini menggunakan data karyawan Kantor Balai Monitor Spektrum Frekwensi yang terdiri dari data

pengalaman kerja, kualifikasi pendidikan kinerja. Data dikumpulkan melalui akses catatan berdasarkan data karyawan pada Kantor Balai Monitor Spektrum Frekwensi. Adapun data yang digunakan yaitu: data karyawan, pangkat, jabatan, masa kerja, latihan yang pernah diikuti, pendidikan dan usia. Dengan ketentuan yaitu apabila pegawai/karyawan yang memilki:

- Pangkat paling tinggi maka dia diletakkna nomor urut yang palign tinggi,
- Pelatihan Jabatan: diutamakan bagi yang pernah mengikuti pelatihan.
- Masa Kerja, bagi yang memilki masa kerja lebih lama maka akan dipilih dari masa kerja yang paling lama
- Pendidikan, pendidikan merupakan syarat yang juga harus dipenuhi.
- Usia, usia juga menentukan seseorang diusulkan DUK, bagi yang usia lebih tua maka kemungkinan dapat diberikan kesempatan untuk mendapatkan DUK.

B. Modeling

Pada tahap pemodelan, penelitian ini memanfaatkan metode SAW untuk menentukan rekomendasi seseorang dapat mendapatakan DUK sesuai dengan Urutan berdasarkan syarat2nya. Adapun langkah yang dilakukan dalam model SAW adalah:

1) Persiapan Data Kriteria

TABEL. 1. KRITERIA

Kriteria	Jenis Kriteria	Bobot
Pangkat	Benefit	30%
Latihan Jabatan	Benefit	15%
Jabatan	Benefit	15%
Masa kerja	Benefit	10%
Pendidikan	Benefit	30%
usia	Benefit	10%

Pada tabel 1 kriteria dalam melakukan proses pengurutan pegawai didasarkan pada hirarki atau tingkatan yang paling tinggi dari kriteria berikut :

- **Pangkat**, pegawai dengan pangkat yang lebih tinggi di tempatkan pada pengurutan pertama dalam struktur DUK. Bila terdapat dua atau lebih pegawai yang berpangkat sama, maka penilaian selanjutnya dilihat dari
- **Jabatan,** pegawai dengan jabatan yang lebih tinggi ditempatkan pada pengurutan kedua setelah pangkat dalam struktur DUK. Bila Terdapat dua atau lebih pegawai dengan pangkat yang sama dan diangkat pada waktu bersamaan, maka penilaian dilihat dari siapa yang memangku jabatan yang lebih tinggi. Dan bila jabatan juga sama, maka penilaian di dasarkan pada siapa yang lebih dulu diangkat dalam jabatan tersebut.
- **Masa Kerja**, pegawai dengan masa kerja yang lebih lama ditempatkan pada pengurutan ketiga setelah pangkat dan jabatan dalam struktur DUK. Bila terdapat dua atau lebih pegawai yang memilik pangkat yang sama, dan menduduki jabatan yang sama, maka penilaian dilihat dari yang memiliki masa kerja yang lebih lama dihitung mulai sejak kapan pegawai diangkat.
- **Latihan Jabatan,** pegawai dengan tingkatan mengikuti pelatihan lebih tinggi di tempatkan pada pengurutan keempat setelah pangkat, jabatan, dan masa kerja dalam struktur DUK. Bila terdapat dua atau lebih pegawai yang memilik pangkat, jabatan dan masa kerja yang, maka penilaian dilihat dari siapa yang mengikuti tingkatan jabatan yang lebih tinggi, Tingkatan latihan jabatan dalam struktur kepegawain diatur sesuai dengan undang-undang yang berlaku. Apabila tinkatan latihan jabatan juga sama maka penilaian dilihat berdasarkan nomor urut yang yang dicantumkan pada daftar kelulusan mengikuti latihan jabatan.
- **Pendidikan,** pegawai dengan tingkat pendidikan yang lebih tinggi, ditempatkan pada pengutuan kelima setelah pangkat, jabatan, masa kerja dan latihan jabatan dalam struktur DUK. Bila terdapat dua atau lebih pegawai yang memiliki pangkat, jabatan, dan latihan jabatan yang sama, maka penilaian dilihat dari siapa memiliki tingkat pendidikan yang lebih tinggi. Bila tingkat pendidikan juga sama, maka penilaian dilihat dari siapa yang lebih awal lulus sebagai pegawai.
- **Usia**, pegawai dengan tingkat usia yang lebih tua, ditempatkan pada pengurutan keenam setelah pangkat, jabatan, masa kerja, latihan jabatan, pendidikan dalam struktur DUK. Bila terdapat dua atau lebih pegawai yang memiliki pangkat, jabatan, masa kerja, latihan jabatan dan pendidikan yang sama, mak penilaian dilihat dari mereka yang memiliki umur lebih tinggi atau tua pegawai yang memiliki umur lebih atau yang lebih tua.

2) Pembobotan

Dalam melakukan pengurutan pangkat dan jabat perlu adanya proses perhitungan bobot. Dalam penelitian ini terdapat 4 kriteria dalam melakukan perhitungan bobot. Kriteria pertama pada tabel 2, yaitu kriteria pangkat. Kedua pada tabel 3, yaitu kriteria pernah dan tidak pernahnya seseorang memperoleh jabatan. Ketiga pada tabel 4, yaitu kriteria jabatan. Terakhir pada tabel 5, yaitu kriteria lamanya bekerja. Apabila bobot semakin tinggi maka pangkat dan jabatan seseorang semakin tinggi.

TABEL. 2 . BOBOT KRITERIA PANGKAT

BOBOT	KETERANGAN
1	II C (PENGATUR)
2	II D (PENGATUR TK.I)
3	III A (PENATA MUDA)
4	III B (PENATA MUDA TK.I)
5	III C (PENATA)
6	III D (PENATA TINGKAT I)
7	IV A (PEMBINA)

TABEL. 3. BOBOT KRITERIA JABATAN

BOBOT	KETERANGAN
0	TIDAK PERNAH
1	PERNAH

TABEL. 4. JABATAN

BOBOT	KETERANGAN
1	Pengelola Manajemen Monitoring SFR & Perangkat Informatika
2	Analisis, pengendali pertama (lvl 1)
3	Pengendali lanjutan
4	Analisis Sumber Daya Bidang Monitor SFR (lvl 2)

4	Analisis Sumber Daya Bidang Monitor SFR (lvl 3)
5	Kepala seksi
6	Kelapa Sub
7	Kepala Cabang

TABEL. 5. MASA KERJA

BOBOT	KETERANGAN
1	< =5 TAHUN
2	5-10 TAHUN
3	>10 TAHUN

Setelah dilakukan pembobotan selanjutnya adalah proses perhitungan menggunakan SAW. Adapun yang dilakukan adalah [2]:

1. Melakukan normalisasi data
2. Tahap perhitungan freferensi menggunakan rumus (2):

$$v_i = \sum_{j=i}^{n} w_j r_{i,j} \qquad (2)$$

Dimana:
Vi = Nilai Prferensi (Hasil Perangkingan)
Wj = Bobot persentase setiap kriteria
n = merupakan proses perhitungan ke sekian

Jadi, bentuk penerapan proses perhitungannya, seperti pada rumus (3):

Nilai_= pangkat*30%)+(jabatan*15%)+ (masakerja*15%) +(latihanjabatan*10%)+(pendidikan*20%)+(usia*10 %) (3)

III. HASIL DAN PEMBAHASAN

Hasil yang diperoleh pada penelitian didasarkan pada data pada table 6, dimana dari data tersebut dilakukan langkah-langkah perhitungan SAW maka data akan dikonversi berdasarkan bobot yang telah ditentukan pada setiap kriteria, terlihat seperti pada table 7. Data yang dilakukan proses pembobotan menggunakan metode SAW akan dilakukan normalisasi. Hasil normalisasi data terdapat pada tabel 8.

TABEL. 6 DATA PEGAWAI

NAMA	PANGKAT	JABATAN	MASA KERJA	LAT JAB	PENDIDIKAN	USIA
Rohmudin, S.Sos.,M.M	IV.A	Kepala Balmont	23	pernah	s1	55
Tarson Effendi, S.E	III.D	Analisis Sumber daya lv.1	22	Tidak pernah	S2	57
Hadi Santoso, S.H	III.D	Analisis Sumber daya lv.1	24	Tidak pernah	S2	59
Ifit Citraningtyas, S.T	III.D	Analisis Sumber daya lv.1	12	Pernah	S3	44
Dody Irawan, S.Sos	III.D	Analisis Sumber Daya Bidang Monitor SFR (lvl 2)	23	Tidak pernah	D3	45
Siti Fatimah, S.Sos	III.D	Analisis Sumber Daya Bidang Monitor SFR (lvl 2)	21	Pernah	D3	40

TABEL. 7 PEMBOBOTAN TERHADAP DATA

NO	NAMA	PANGKAT	JABATAN	MASA KERJA	LAT JAB	PENDIDIKAN
Rohmudin, S.Sos.,M.M	7	6	3	1	3	55
Tarson Effendi, S.E	6	2	3	0	4	57
Hadi Santoso, S.H	6	2	3	0	4	59
Ifit Citraningtyas, S.T	6	2	2	1	5	44
Dody Irawan, S.Sos	6	2	2	0	2	45
Siti Fatimah, S.Sos	6	4	3	1	4	40

TABEL. 8 HASIL NORMALISASI DATA

NO	NAMA	PANGKAT	JABATAN	MASA KERJA	LAT JAB	PENDIDIKAN	USIA
1	Rohmudin, s.sos, m.m	1	1	1	1	0,6	0,96
2	Tarson Effendi, Se	0,857143	0,333333	1	0	0,8	1
3	Hadi Santoso, Sh	0,857143	0,333333	1	0	0,8	1,04
4	Ifit Citraningtyas,St	0,857143	0,333333	0,6666667	1	1	0,77
5	Dody Irawan, S. Sos	0,857143	0,333333	0,6666667	0	0,4	0,79

NO	NAMA	PANGKAT	JABATAN	MASA KERJA	LAT JAB	PENDIDIKAN	USIA
6	Siti Fatimah, S.Sos	0	0	0	0	0	0
7	Priska Wina, St	0,857143	0,666667	1	1	0,8	0,7

TABEL. 9 HASIL EKSPERIMEN

NAMA	PREFERENSI	RANGKING
Rohmudin, S.Sos.,M.M	0.98	1
Tarson Effendi, S.E	0.76	2
Hadi Santoso, S.H	0.77	3
Ifit Citraningtyas, S.T	0.76	4
Dody Irawan, S.Sos	0.76	5
Siti Fatimah, S.Sos	0.72	6
Priska Wina, S.T	0.64	7
Khairiansyah, S.H	0.69	8
Azwari Purba, S.	0.56	9
Dedy Adiansyah	0.50	10
Muhamad Yusrif	0.50	11
Domi Wahyu Sih. W, A.Md	0.49	12
Faisal Rizkan, S.T	0.47	13
Aldibangun P. Putro, S.T	0.47	14
Komang Arwana	0.35	15
Yuditha Tallulembang, A.Md	0.33	16
Syahruni	0.39	17
Nur Samroni	0.37	18
Doni Suprihadoko	0.36	19
Kristian Limbong, A.Md	0.23	20

Pada table 9 disajikan hasil dari eksperimen pada metode SAW, pada metode SAW yang diterapkan memperlihatkan bahwa karyawan/pegawai bernama Rohmudin dengan nilai preferensi 0.98 merupakan nilai yang paling tinggi untuk menjadi orang yang berhak untuk diajukan untuk mendapatkan Daftar Urut Kepangkatan. Dimana Rohmudin sendiri memilki data yang relevan dengan syarat-syarat yang ditetapkan yakni seperti pada table. 6.

Dari hasil uji coba perhitungan menggunakan metode SAW maka diperoleh data berdasarkan urutan pengjauan DUK seperti pada table 9, dari hasil tersebut maka nilai freferensi tertinggi akan menjadi rekomendasi yang pertama direkomendasikan untuk mendapatkan DUK.

UCAPAN TERIMA KASIH

Dari temuan dalam penelitian, dapat disimpulkan bahwa:

1. Metode SAW dapat digunakan secara efektif dalam memberikan rekomendasi pengajuan Daftar Urut Kepangkatan di Kantor Balmon. Metode ini mampu menggabungkan berbagai kriteria yang relevan dan memberikan peringkat yang objektif berdasarkan bobot relatif dari masing-masing-masing kriteria.
2. Tingkat kepentingan kriteria yang diperoleh dari responden dapat menjadi acuan dalam menentukan bobot relatif dari setiap kriteria. Dengan demikian, pengambilan keputusan mengenai urutan kepangkatan dapat didasarkan pada preferensi dan kepentingan yang telah diungkapkan oleh pihak terkait.
3. Hasil perangkingan dan rekomendasi yang dihasilkan oleh metode SAW dapat memberikan panduan yang jelas dalam pengajuan DUK. Hal ini dapat membantu dalam meminimalkan bias subjektivitas dan memastikan keadilan serta transparansi dalam proses pengambilan keputusan.

REFERENSI

[1] Dian Sarunggaga, Fembrawati Rinding. (2022). Penerapan Metode Simple Additive Weighting (SAW) dalam pendataan Daftar Urut Kepangkatan (DUK) pada sistem Kepegawaian Kantor DPRD Provinsi Sulawesi Selatan, Universitas DIPA Makasar, 28-40.

[2] Dimas Alexandra & Anita Qoiriah. (2022). Penerapan Metode Simple Additive Weighting (SAW) Dalam Sistem Penilaian Kinerja MitraLembaga Badan Pusat Statistik Jember. Jurusan TeknikInformatika Fakultas Teknik Universitas Negeri Surabaya (Journal of Informatics and Computer Science) Volume 04 Nomor02, Hal. 197.

[3] Kusrini, Konsep dan Aplikasi Sistem Pendukung Keputusan. Yogyakarta: CV ANDI OFFSET, 2007.

[4] Liesnaningsih, Oklawati, Dian Kasoni. 2019. Sistem PendukungKeputusan Penilaian Karyawan Terbaik Menggunakan Metode Simple Additive Weighting Pada PT. Trans Retail Indonesia, Jurnal Teknik Informatika (JIKA), November, Universitas Muhammadiyah Tanggerang, hal. 34-40.

[5] M Afdal Tahir, Andi Patappari (2022), Sistem Informasi Daftar Urut Kepangkatan (DUK) Pegawai Pada Kantor Dinas Pemberdayaan Perempuan Dan Keluarga Berencana Kabupaten Soppeng, Jurnal Ilmiah Sistem Informasi dan Teknik Informatika, Vol 5 (2) Universitas Lamappapoleonro, hal 58

[6] M Afdal Tahir, Andi Patappari (2022), Sistem Informasi Daftar Urut Kepangkatan (DUK) Pegawai Pada Kantor Dinas Pemberdayaan Perempuan Dan Keluarga Berencana Kabupaten Soppeng, Jurnal Ilmiah Sistem Informasi dan Teknik Informatika, Vol 5 (2) Universitas Lamappapoleonro, hal 58.

[7] Risawandi and Rahimi, "Study of the simple Multi-Attribut Rating Technique for Decission Support," *USRST*, vol. 2, no. 6,p. 491-494, 2016.

[8] F. Nashrullah, S. A. Wibowo, and G. Budiman, "Investigasi parameter epoch pada arsitektur resnet-50 untuk klasifikasi pornografi," *J. Comput. Electron. Telecommun.*, vol. 1, no. 1, pp. 1–8, 2020, doi: 10.52435/complete.v1i1.51.

[9] Rizky Eka Putra, Jap Tji Beng, Desi Arisandi. 2020. Sistem PendukungKeputusan Pekerja Terbaik Pada PT. Dwi Karya Kumi Mandiri Dengan Metode Simple Additive Weighting, Jurnal Ilmu Komputerdan Sistem Infomasi, Vol 8(1), Universitas Tarumanegara, hal.151-15

[10] Setiawan, D., & Wibowo, A. (2021). Penerapan metode SAW dalam penentuan urutan kepangkatan pegawai pada Kantor PemerintahanDaerah. Jurnal Informatika Mulawarman, 16(1), 45-52

[11] Turban, E. Aronson, J., & Liang, and R. Decission Support Systems And Intelegence system.US: Prentice-Hall,2005

Analisis Sentimen Masyarakat pada Sosial Media X Terhadap Gibran Rakabuming sebagai Calon Wakil Presiden 2024

Muhammad Zydane Arrosyid
Fakultas Sains & Teknologi
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia

Yessy Yee Nur Ariyanti Sekar P.D.
Fakultas Sains & Teknologi
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia

Luis Fernandes Tokan
Fakultas Sains & Teknologi
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia

Muhammad Al-Fajr Ramadhani
Fakultas Sains & Teknologi
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia

Alvinus Cardova
Fakultas Sains & Teknologi
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia

Edwhin Rantho Rafafi
Fakultas Sains & Teknologi
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia

Abstrak—Pemilihan presiden dan wakil presiden merupakan momen penting dalam pertumbuhan demokrasi suatu negara. pencalonan calon wakil presiden memainkan peran penting dalam membentuk pemerintahan dan kebijakannya. Dalam upaya memahami sentimen masyarakat terhadap calon wakil presiden pada Pemilihan Presiden 2024, fokus penelitian ini akan ditempatkan pada sosok Gibran Rakabuming, yang mungkin menjadi salah satu figur yang menarik perhatian masyarakat dan media dalam konteks pemilihan calon wakil presiden. Tahapan awal pada penelitian ini yaitu pengambilan data dari X dengan kata kunci "Gibran Cawapres" akan terdapat banyak komentar yang memiliki macam-macam komentar dari masyarakat terkait Gibran menjadi Calon Presiden. Penelitian ini menggunakan Teknik *Support Vector Machine* (SVM). *Support Vector Machine* (SVM) adalah sebuah metode pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. SVM mencari fungsi pemisah terbaik untuk memisahkan kelas, dan termasuk dalam pendekatan supervised learning. Penelitian ini memiliki tingkat akurasi yang tinggi, dengan tingkat presisi 96% untuk tweet negatif, algoritme ini kesulitan mengenali tweet positif, dengan tingkat presisi hanya 19%. Hasil ini menunjukkan bahwa pengguna Twitter cenderung memberikan tanggapan negatif terhadap masalah ini. hal tersebut dapat digunakan oleh tim gibran dalam mengubah opini publik dan mendapatkan dukungan lebih luas.

Kata Kunci— supervised learning, pemilu, Support Vector Machine (SVM), Gibran Rakabuming Raka

I. LATAR BELAKANG

Pemilihan presiden dan wakil presiden merupakan momen penting dalam pertumbuhan demokrasi suatu negara. Indonesia, sebagai salah satu negara demokrasi terbesar di dunia, tidak terkecuali dalam hal ini. Dalam lanskap politik Indonesia, pencalonan calon wakil presiden memainkan peran penting dalam membentuk pemerintahan dan kebijakannya

[1]. Peran Relawan Politik Dalam Konstelasi Politik Indonesia. Jurnal Hukum Sasana.Calon wakil presiden merupakan tokoh terkemuka yang diharapkan dapat membantu presiden dalam memenuhi kewajiban negaranya. Oleh karena itu, sangat penting untuk memahami opini dan emosi masyarakat mengenai individu ini.

Opini dan pendapat masyarakat dapat disampaikan dengan berbagai macam media. [2]. Pada bulan januari tahun 2022 pengguna internet di indonesia mencapai 204,7 juta. Salah satu media sosial yang sering dipakai oleh publik untuk beropini yaitu X. X menjadi salah satu cara masyarakat dalam merespon berbagai hal melalui media sosial, baik berupa respon positif maupun negatif [3]. Oleh sebab itu, X memiliki kelebihan dibanding media sosial lain dan X dapat digunakan secara efektif untuk pengumpulan data opini dari masyarakat, namun dalam pengumpulan data setelah dilakukan preprocessing dapat terjadi imbalance data yang dapat mempengaruhi hasil yang diperoleh.

Dalam upaya memahami sentimen masyarakat terhadap calon wakil presiden pada Pemilihan Presiden 2024, fokus penelitian ini akan ditempatkan pada sosok Gibran Rakabuming, yang mungkin menjadi salah satu figur yang menarik perhatian masyarakat dan media dalam konteks pemilihan calon wakil presiden. Gibran Rakabuming Raka adalah seorang figur publik yang terpilih sebagai walikota Solo, Indonesia masa jabatan. Ia adalah putra sulung dari Presiden Indonesia, Joko Widodo [4]. Gibran adalah kandidat dalam Pemilihan Kepala Daerah Kota Surakarta 2020, di mana ia maju sebagai bagian dari partai politik. Ia memenangkan pemilihan dan menjadi walikota Solo [5]. Gibran telah menggunakan media sosial, terutama X dan Instagram, untuk membangun merek pribadinya dan berkomunikasi dengan publik. Ia telah berhasil membangun citra dan menghasilkan opini public melalui kegiatan media sosialnya. Personal branding-nya telah dianalisis menggunakan semiotika dan teori komunikasi relevan lainnya[6][7]. Wali Kota Solo tersebut resmi diusung Partai Golkar menjadi cawapres pendamping Prabowo Subianto pada Sabtu, 21 Oktober 2023. Ia merupakan pemuda pertama di Indonesia yang di usungkan menjadi bakal calon wakil presiden.

Pada konteks inilah penelitian ini akan berfokus. Tujuan utama dari penelitian ini adalah untuk menganalisis sentimen masyarakat terhadap Gibran Rakabuming sebagai calon wakil presiden 2024 berdasarkan data yang diperoleh dari sosial media X. Data yang dianalisis akan mencakup beragam percakapan, komentar, dan pendapat yang terkait dengan

Gibran Rakabuming, sehingga dapat memberikan wawasan mendalam tentang pandangan masyarakat terhadap calon wakil presiden tersebut.

II. PENELITIAN TERKAIT

X adalah media sosial yang ramai digunakan serta menyediakan fitur yang memungkinkan pengguna untuk berbagi pendapat melalui pesan singkat atau sering dikenal dengan tweet [8]. Ulasan dari X dapat diklasifikasikan ke dalam beberapa sentimen. Seperti penelitian yang dilakukan oleh [9] Styawati, S., Hendrastuty, N., & Isnain, A. R. opini masyarakat terkait program Kartu Prakerja diklasifikasikan dalam dua sentimen yaitu positif dan negatif. Dengan metode yang digunakan *Support Vector Machine* (SVM). Metode ini dapat melakukan proses klasifikasi dengan baik. peneliti lain, mengatakan bahwa hasil terbaik untuk mendeteksi sentimen dari X Bahasa Indonesia dapat dicapai dengan menggunakan metode SVM. Menurut, [10] SVM dapat melakukan klasifikasi data opini Film dari Twitter dengan hasil akurasi model 87.5%. Berdasarkan pada penelitian terdahulu, penelitian ini akan menganalisa sentimen masyarakat Pada Sosial Media X Terhadap Gibran Rakabuming Sebagai Calon Wakil Presiden 2024 menggunakan Metode *Support Vector Machine* (SVM).

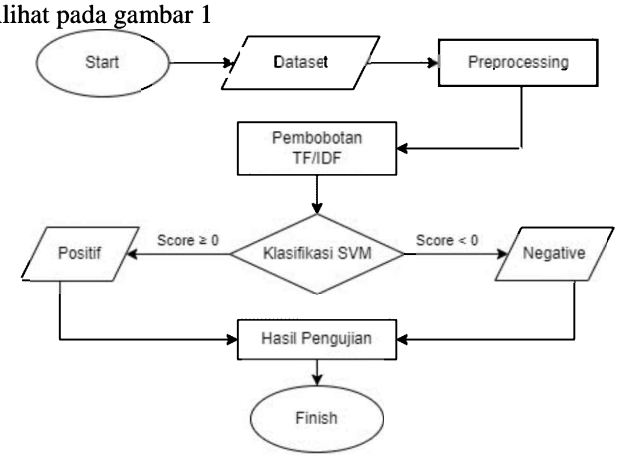
III. METODE PENELITIAN

A. Tahapan Penelitian

Tahapan awal pada penelitian ini yaitu pengambilan data dari Twitter dengan kata kunci "Gibran Cawapres" akan terdapat banyak komentar yang memiliki macam-macam komentar dari masyarakat terkait Gibran menjadi Calon Presiden. Data yang diperoleh kemudian dilakukan preprocessing data. Selanjutnya data dilakukan pelabelan data ketika sudah melalui pra pemrosesan data, pelabelan dilakukan secara manual. Setelah itu, dilakukan pembobotan TF IDF untuk kemudian dilakukan klasifikasi menggunakan algoritma *Support Vector Machine* (SVM). Data akan dikategorikan menjadi dua kelas yaitu positif dan negatif.

B. Skema Pemodelan SVM

Proses klasifikasi data menggunakan metode SVM, dapat dilihat pada gambar 1



Gambar 1. Skema Model SVM

C. Support Vector Machine (SVM)

Teknik *Support Vector Machine* (SVM) bertujuan untuk menemukan fungsi pemisah terbaik di antara fungsi yang ada untuk memisahkan dua macam obyek. Penyelesaian klasifikasi dua kelas dapat menggunakan persamaan berikut :

$$min\ {}^{1}_{2}\boldsymbol{j}^{2} + C\sum\int_{r=1}^{N} \xi r \quad subject\ to{:}\boldsymbol{j}ij + \boldsymbol{m}i \geq tijr, jikaYr = \boldsymbol{i}jij + \boldsymbol{m}i \leq tijr, jikaYr \neq itijr \geq 0 \quad 1)$$

Support Vector Machine (SVM) adalah sebuah metode pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. SVM mencari fungsi pemisah terbaik untuk memisahkan kelas, dan termasuk dalam pendekatan supervised learning. SVM menggunakan metode pembelajaran mesin berupa fungsi linier pada ruang fitur berdimensi tinggi yang menggunakan ruang maya dan dilatih dengan metode pembelajaran mesin berdasarkan teori optimasi yang diturunkan dari teori statistik. SVM merupakan metode pengklasifikasi yang bisa digunakan *scare linear* atau *non-linier*. SVM memiliki keunggulan dalam pengenalan pola yang baik dapat digeneralisasi memakai metode ini[11].

IV. EKSPERIMEN DAN PEMBAHASAN

A. Dataset

Dataset yang digunakan dalam penelitian ini merupakan data yang tersedia secara publik pada media sosial twitter atau X dengan mengakses API milik X. Pengambilan data dilakukan dengan memasukkan kata kunci tertentu. Kata kunci yang digunakan dalam penelitian ini yaitu "Gibran Cawapres" dimana semua *tweet* yang berhubungan dengan Gibran sebagai cawapres akan termuat dalam pencarian ini. Pengambilan dan pengumpulan data dilaksanakan pada tanggal 9 November 2023 yang memuat 500 data *tweet* pengguna terkait Gibran Cawapres. Hasil dari proses File yang dihasilkan adalah file dengan format csv yang kemudian akan diolah dan dilatih menggunakan bahasa pemrograman *python*. Tabel 1 menunjukkan hasil *load data* dari *google colab* yang di unduh dalam format *excel*

TABEL 1. CONTOH HASIL *LOAD DATASET*

Id	*Text*	*Username*
1	@gibran_tweet Yakin mau dukung CAWAPRES BONEKA ? ???????????	msobri99
2	@Gus_Raharjo Gibran Cawapres Boneka Utk seorang yg blm berpengalaman ini smua cm jd ajang coba2 dan lahan titipan masalah dikemudian hari	FaktaNKRI72
3	@Gus_Raharjo Gibran Cawapres Boneka yang gak punya maluu wkwk	lolitaandriani2

B. Preprocessing

Sebelum *dataset* dilatih dengan model, data perlu proses terlebih dahulu agar data bersih sehingga tidak ada *noisy* data pada proses pelatihan model. Tahapan *preprocessing* yang dilakukan pada penelitian ini terdiri dari 3 tahap yaitu

- Menghapus duplikasi

  Proses menghapus duplikasi dilakukan agar load data yang disajikan tidak banyak sehingga penyimpanan dan sumber daya komputasi tidak terlalu besar.

- Case Folding

  Tahap *preprocessing* ini menjadi hal yang sangat krusial karena *case folding* dapat menghindari kebingungan program saat mengenali kata atau frasa dalam teks. *Case folding* yang dilakukan pada penelitian

ini dilakukan dengan teknik *lowercassing* yaitu mengubah semua huruf menjadi huruf kecil.

- *Cleansing Data*

  Tujuan dari cleansing data adalah untuk membersihkan teks dari karakter karakter yang tidak relevan sehingga menghasilkan data yang lebih efektif dan akurat. Pada penelitian ini, tahap *cleansing data* dilakukan dengan menghapus elemen 'http' , tanda atau simbol,atau karakter lain selain huruf a-z atau A-Z.

Perbedaan data sebelum dan setelah di preprocessing dapat dilihat pada tabel 2

TABEL 2. CONTOH HASIL *PREPROCESSING*

Id	Text Sebelum *Preprocessing*	Text Setelah *Preprocessing*
1	@gibran_tweet Yakin mau dukung CAWAPRES BONEKA ? ????????????	yakin mau dukung cawapres boneka
2	@Gus_Raharjo Gibran Cawapres Boneka Utk seorang yg blm berpengalaman ini smua cm jd ajang coba2 dan lahan titipan masalah dikemudian hari	gibran cawapres boneka utk seorang yg blm berpengalaman ini smua cm jd ajang coba2 dan lahan titipan masalah dikemudian hari
3	@Gus_Raharjo Gibran Cawapres Boneka yang gak punya maluu wkwk	gibran cawapres boneka yang gak punya maluu wkwk

C. *Pelabelan Data*

Data yang sudah bersih kemudian di *export* ke file csv untuk proses pelabelan data. Pada penelitian ini proses pelabelan data dilakukan secara manual dengan membagi menjadi dua kelas yaitu positif dan negatif. Pelabelan manual dilakukan berdasarkan opini peneliti terhadap kalimat yang disajikan. Label -1 menunjukkan sentimen negatif sedangkan label 1 menunjukkan hasil positif. Tabel 3 Menampilkan contoh *tweet* yang dilakukan saat eksperimen penelitian.

TABEL 3. CONTOH HASIL PELABELAN DATA

Id	Text Setelah Preprocessing	Label
1	@gibran_tweet Yakin mau dukung CAWAPRES BONEKA ? ????????????	-1
2	@Gus_Raharjo Gibran Cawapres Boneka Utk seorang yg blm berpengalaman ini smua cm jd ajang coba2 dan lahan titipan masalah dikemudian hari	-1
3	@Gus_Raharjo Gibran Cawapres Boneka yang gak punya maluu wkwk	-1

D. *Pembobotan TF IDF*

Kata kata yang sudah melalui proses perataan kata selanjutnya akan dinilai bobotnya dengan menggunakan TF-IDF. TF-IDF berkerja dengan membagi frekuensi kata yang muncul terhadap keseluruhan data. Semakin sering kata tersebut muncul maka bobonya semakin besar

E. *Split Data*

Pada tahap ini akan dilakukan pembagian terhadap keseluruhan dataset sesuai proporsi tertentu. *Dataset* akan dibagi menjadi 2 *dataset*, yaitu data latih dan data uji. Pada penelitian ini *split data* dilakukan dengan membagi data menjadi 2 dataset yaitu data latih sebesar 70% dan data uji sebesar 30%. Hasil proporsi tersebut merupakan hasil akurasi terbaik dari percobaan percobaan yan telah dilakukan. Didapatkan hasil data latih sebanyak 246 baris data dan data uji sebayak 106 data.

F. *Klasifikasi Sentimen*

Proses selanjutnya adalah dengan melatih data kedalam model SVM. Tujuan dari penerapan model SVM adalah klasifikasi atau pengelompokkan data berdasarkan 2 kategori sentimen, yaitu sentimen positif dan negatif.

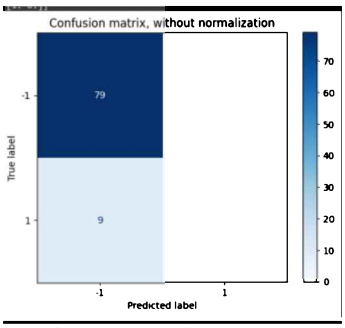
G. *Pengujian dan Hasil Akurasi*

Langkah yang terakhir dari penelitian ini merupakan proses analisis dari sistem yang kita buat untuk mengukur akurasi dari sistem yang dirancang pada penelitian ini. Dalam menentukan hasil akurasi dibutuhkan suatu pengukuran evaluasi yang disebut *confusion matrix*. *Confusion matrix* adalah suatu tabel termasuk sejumlah besar data pengujian yang diprediksi secara kuat dan lemah oleh model klasifikasi yang dipakai dalam penentuan performa sebuah model klasifikasi yang dapat menghasilkan *recall*, *precision* dan *accuracy*. Tabel 4 menunjukkan hasil klasifikasi algoritma SVM terhadap isu Gibran Rakabuming sebagai cawapres.

TABEL 4. HASIL MODEL SVM

	Precision	Recall	F1-Score	Support
Positive	1.0	0.1111	0.19	9
Negative	0.9238	1.0	0.9604	97

Dari tabel 4 dapat diketahui bahwasannya hasil F1-Score *tweet* positif sebesar 19% dan F1-Score negatif sebesar 96%. *Tweet* positif menunjukkan hasil akurasi rendah karena kurangnya data sentimen positif pada saat pengambilan data. Hasil uji *confussion matrix* dapat dilihat pada gambar 2 berikut ini.



Gambar 2. *Confusion Matrix*

Hasil menunjukkan bahwa terdapat 79 *true negatif* dimana sentiment negatif diklasifikasikan sebagai sentimen negatif juga oleh sistem. Dan terdapat 9 *false negatif* dimana sentiment positif diklasifikasikan sebagai sentimen negatif juga oleh sistem.

Jika dibandingkan pada penelitian sebelumnya yang berjudul "Analisis sentimen pemilihan presiden indonesia tahun 2019 di twitter berdasarkan geolocation menggunakan metode naïve bayes" [12] hasil akurasi penelitian ini lebih tinggi.

Berdasarkan eksperimen dan hasil pengujian, didapatkan informasi bahwa isu mengenai "gibran cawapres" pada media sosial X ternyata memiliki sentimen yang cenderung negatif. Informasi tersebut dapat dimanfaatkan oleh tim gibran dalam mengubah opini masyarakat sebelum pemilu 2024.

V. KESIMPULAN

Berdasarkan hasil pada studi yang telah dilakukan, maka dapat diambil beberapa kesimpulan. Evaluasi dan analisis

sistem klasifikasi "Gibran Rakabuming sebagai calon wakil presiden" dengan menggunakan algoritma SVM menghasilkan temuan yang signifikan. Meskipun algoritma ini menawarkan tingkat akurasi yang tinggi, dengan tingkat presisi 96% untuk *tweet* negatif, algoritme ini kesulitan mengenali *tweet* positif, dengan tingkat presisi hanya 19%. Hasil ini menunjukkan bahwa pengguna X cenderung memberikan tanggapan negatif terhadap masalah ini.

Data menunjukkan bahwa dari total 88 tanggapan, 79 merupakan tanggapan negatif benar dan 9 tanggapan negatif palsu. Berdasarkan bukti tersebut, dapat disimpulkan bahwa masyarakat lebih cenderung merespons negatif isu ini di media sosial. Kesimpulan ini menjadi modal berharga bagi tim Gibran dalam menyusun strategi komunikasi pemilu 2024 mendatang. Dengan mengenali prevalensi sentimen pesimistis, tim dapat merancang kampanye yang lebih efektif yang akan mengubah opini publik dan mendapatkan dukungan lebih luas. Informasi ini menjadi landasan untuk menyusun langkah-langkah konkrit dalam mengatasi persepsi negatif yang dapat menghambat persaingan politik.

REFERENSI

[12] Syauket, A. (2022). Peran Relawan Politik Dalam Konstelasi Politik Indonesia. Jurnal Hukum Sasana.

[13] Kemp, S. (2022). Digital 2022: Another Year Of Bumper Growth

[14] Pane, S. F., & Ramdan, J. (2022). Pemodelan Machine Learning: Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Menggunakan Data Twitter. Jurnal Sistem Cerdas, 5(1), 12-20

[15] Andhita, P.R. (2021). Hierarki Pengaruh Dalam Pemberitaan Gibran Sebagai Kandidat Walikota Surakarta Di Solopos.com. Avant Garde.

[16] Putri, S.A. (2021). Personal Branding Pejabat Publik (Analisis Isi Akun Instagram Walikota Solo Gibran Rakabuming Raka). Mediova: Journal of Islamic Media Studies.

[17] Riyanti, J. (2020). Marketing Politik di Media dan Softening News Gibran Rakabuming Raka dalam Pemilihan Wali Kota Solo. CHANNEL: Jurnal Komunikasi.

[18] Wibiyanto, A. (2021). Analisis Pengelolaan Kesan Achmad Purnomo Dan Gibran Rakabuming Menjelang Pilkada Solo 2020. CoverAge: Journal of Strategic Communication.

[19] D. K. Zala, "A Review on Basic Methodology of Twitter Base Prediction System," pp. 447–451, 2018.

[20] Styawati, S., Hendrastuty, N., & Isnain, A. R. (2021). Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine. Jurnal Informatika: Jurnal Pengembangan IT, 6(3), 150-155.

[21] Khairudin, M., Sukendar, A., & Somantri, A. (2023). Analisis Sentimen Film Di Twitter Menggunakan Metode Support Vector Machine. Jurnal Sains dan Sistem Teknologi Informasi, 5(1), 97-102.

[22] Adi, S., & Wintarti, A. (2022). Komparasi Metode Support Vector Machine (Svm), K-Nearest Neighbors (Knn), Dan Random Forest (Rf) Untuk Prediksi Penyakit Gagal Jantung. MATHunesa: Jurnal Ilmiah Matematika.

[23] Wiranto,H & Jay Idoan. Analisis Sentimen Pemilihan Presiden Indonesia Tahun 2019 Di Twitter Berdasarkan Geolocation Menggunakan Naïve Bayesiyan Classification. Jurnal TelKa, 9(2), 115-127

Forecasting Produksi Beras menggunakan LSTM: Menjamin Ketahanan Pangan di Sumatera

Ach. Nur Aqil Wahid
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
nur.5200411387@sudent.uty.ac.id

Cahyo Prakoso
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
cahyo.5200411379@student.uty.ac.id

Muhammad Aulia
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
Muhammad.5200411482@student.uty.ac.id

Fahri Putra Herlambang
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
fahri.5200411389@student.uty.ac.id

Ilham Rafiedhia Pramutighna
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
ilham.5200411490@student.uty.ac.id

Muhammad Rousydi Hunafa
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
muhammad.5200411483@student.uty.ac.id

ABSTRAK

Penelitian ini membahas metodologi untuk meramalkan produksi beras di wilayah Sumatera menggunakan model Long Short-Term Memory (LSTM). Metodologi ini dirancang dengan tujuan sistematis dan efisien, dimulai dari pengumpulan data produksi padi yang komprehensif hingga analisis pola produksi dan proses preprocessing data. Pemodelan dilakukan dengan memanfaatkan arsitektur multivariate LSTM untuk memahami hubungan kompleks antara variabel-variabel seperti curah hujan, suhu, dan jenis tanah yang memengaruhi produksi padi. Evaluasi kinerja model dilakukan dengan metrik seperti Mean Squared Error (MSE) dan Root Mean Squared Error (RMSE) berdasarkan dataset produksi padi dari Kaggle. Eksperimen dan analisis dilanjutkan dengan tahap preprocessing data lanjutan, pembuatan model LSTM, dan evaluasi kinerja model untuk setiap provinsi di Sumatera. Kesimpulannya, metodologi ini memberikan pemahaman mendalam tentang variasi produksi padi di berbagai provinsi Sumatera, memberikan wawasan tentang faktor-faktor yang memengaruhi, dan memberikan dasar bagi pengembangan strategi yang lebih efektif dalam meningkatkan hasil produksi padi di wilayah tersebut. Dengan pendekatan ini, penelitian ini diharapkan memberikan kontribusi signifikan dalam pemahaman dan peningkatan efektivitas produksi padi di Sumatera.

Kata Kunci— Forecasting, LSTM, Beras, Sumatera

VI. PENDAHULUAN

Tingkat ketersediaan pangan yang memadai merupakan indikator penting dalam menilai tingkat kesejahteraan dan keamanan pangan suatu negara [1]. Pulau Sumatra, sebagai salah satu pulau terbesar di Indonesia, memiliki peran yang sangat signifikan dalam sektor pertanian di Indonesia. Dengan lebih dari 50 persen lahan pertanian tersebar di setiap provinsi di pulau ini, Sumatra memainkan peran kunci dalam memasok komoditas pangan utama bagi penduduknya. Komoditas padi menjadi komponen utama dalam produksi pangan, dengan jagung, kacang tanah, dan ubi sebagai komoditas tambahan yang tidak kalah penting [2]. Namun, meskipun Sumatra memiliki potensi besar dalam sektor pertanian, terdapat tantangan yang signifikan dalam mencapai ketersediaan pangan yang memadai di pulau ini. Faktor-faktor seperti fluktuasi iklim, terutama kenaikan suhu dari tahun ke tahun dapat mempengaruhi produksi dan distribusi pangan di wilayah ini. Selain itu, perubahan pola konsumsi masyarakat dan urbanisasi yang terus meningkat juga memberikan tekanan tambahan terhadap sistem produksi pangan di Sumatera [3].

Peneliti terdahulu yang menggunakan Penerapan Jaringan Syaraf Tiruan dengan Algoritma Backpropagation dalam Memprediksi Hasil Panen Gabah Padi. Dalam penelitian tersebut, beberapa model arsitektur digunakan, dan satu arsitektur khusus terpilih karena memiliki tingkat akurasi sebesar 92.9% atau tingkat kesalahan sebesar 7.1% dengan Mean Square Error (MSE) sekitar 0.00094783 [4]. Dengan perbandingan metode machine learning untuk Prediksi Hasil Panen Tanaman Pangan Sumatera. Dalam melakukan penelitiannya, peneliti membandingkan beberapa metode, prediksi, yaitu yaitu Random Forest (RF), Decision Tree (DT), Gradient Boosting (GB), Extra Tree (ET), Support Vector Machine (SVM), dan Artificial Neural Network (ANN). Nilai koefisien R2 untuk produksi padi masing-masing adalah 0,897; 0,893; 0,957; 0,968; 0,928; dan 0,909. Berdasarkan nilai koefisien R2 yang disajikan, model Extra Tree memiliki performa paling baik dalam memprediksi produksi padi dan bahan pokok lainnya, dengan koefisien R2 tertinggi dibandingkan dengan model-model lainnya [5].

Berdasarkan penelitian di atas, untuk memprediksi produksi padi di Pulau Sumatera, diperlukan pendekatan berbasis data yang mampu memodelkan jangka panjang. Dalam hal ini, LSTM (*Long Short-Term Memory*), sebuah jenis Artificial Neural Network (ANN) yang khusus dirancang untuk tugas peramalan jangka panjang, menjadi sangat relevan. Dengan memanfaatkan data dari tahun 1993 hingga tahun 2020, termasuk hasil produksi tahunan dan data cuaca harian seperti curah hujan, kelembaban, dan suhu rata-rata, kita dapat membangun model LSTM yang dapat membantu dalam memprediksi produksi padi di Pulau Sumatera dalam jangka waktu yang lebih luas.

Dengan demikian, model ini dapat menjadi alat penting bagi para pemangku kepentingan dalam sektor pertanian untuk mengambil keputusan yang lebih baik, mengurangi kerentanan terhadap perubahan iklim, dan memastikan ketahanan pangan yang berkelanjutan. Dengan pendekatan berbasis model LSTM ini, kita dapat berkontribusi pada upaya menjaga produksi padi yang stabil di Pulau Sumatera, sambil merespons perubahan

lingkungan yang semakin dinamis dan menghadapi tantangan perubahan iklim global.

VII. KAJIAN LITERATUR

Menurut Herwanto [6], penelitian sebelumnya telah menunjukkan bahwa penggunaan algoritme regresi linear dapat menjadi alat yang efektif dalam memprediksi hasil panen tanaman padi. Regresi linear memungkinkan peneliti untuk menghubungkan hasil panen dengan berbagai faktor yang memengaruhinya, seperti luas lahan, jumlah bibit, dan penggunaan pupuk. Hasil penelitian tersebut menunjukkan bahwa regresi linear dapat memberikan tingkat keandalan yang cukup tinggi dalam prediksi hasil panen padi, yang dapat sangat berguna dalam perencanaan pertanian dan pemenuhan kebutuhan pangan. Dalam penelitian ini, nilai tingkat kecocokan model multiple linear regression sebesar 94,51%, yang mengindikasikan bahwa sebagian besar variasi hasil panen dapat dijelaskan oleh variabel-variabel independen yang digunakan dalam model. Selain itu, hasil akurasi Root Mean Squared Error (RMSE) yang rendah, sebesar 0,432, menunjukkan bahwa model prediksi mendekati akurat. Temuan ini menguatkan gagasan bahwa regresi linear adalah pendekatan yang efektif dalam pemodelan prediksi hasil panen padi dan dapat berkontribusi positif dalam mendukung ketahanan pangan dan pertanian yang berkelanjutan.

Menurut Satria [7], dalam studi literatur terkait prediksi hasil panen tanaman pangan dengan metode *machine learning*, ditemukan bahwa sektor pertanian di Indonesia, terutama di pulau Sumatera, memiliki peran yang krusial dalam pembangunan ekonomi nasional. Dalam konteks perubahan iklim global, penting untuk memahami dampaknya terhadap produksi tanaman pangan. Hasil penelitian sebelumnya menunjukkan bahwa perubahan iklim dapat mengganggu pola tanam, waktu tanam, produksi, dan kualitas hasil pertanian. Oleh karena itu, prediksi hasil panen menjadi semakin penting untuk menjaga ketahanan pangan dan ekonomi di wilayah ini. Metode machine learning, seperti yang digunakan dalam penelitian ini, telah menjadi pendekatan yang efektif untuk memprediksi hasil pertanian dengan mempertimbangkan berbagai faktor, termasuk luas lahan, curah hujan, dan perubahan suhu. Hasil penelitian ini menunjukkan bahwa algoritma Extra Tree muncul sebagai model terbaik dengan tingkat akurasi yang tinggi, yang dapat menjadi landasan penting dalam pengembangan penelitian dan strategi pertanian di pulau Sumatera.

Menurut Pelangi [8], penelitian mereka menggarisbawahi pentingnya sektor pertanian di Provinsi Gorontalo dalam mendukung ekonomi masyarakat lokal serta pemenuhan kebutuhan pangan yang krusial. Mereka mengidentifikasi perubahan luas panen sebagai faktor utama yang memengaruhi hasil produksi tanaman pangan, dengan luas panen yang lebih besar menghasilkan produksi yang lebih tinggi. Penelitian ini mencapai tingkat akurasi tertinggi dalam memprediksi produksi jagung dengan metode K-Nearest Neighbor sebesar 92,83%, sehingga menyimpulkan bahwa aplikasi yang dikembangkan layak digunakan untuk meramalkan hasil produksi tanaman pangan di Provinsi Gorontalo. Penelitian ini memberikan dorongan untuk penelitian lanjutan dalam mengoptimalkan metode tersebut dan mempertimbangkan penambahan variabel untuk meningkatkan akurasi prediksi berdasarkan luas panen.

VIII. METHODOLOGI UNTUK FORECASTING PRODUKSI BERAS DENGAN LSTM

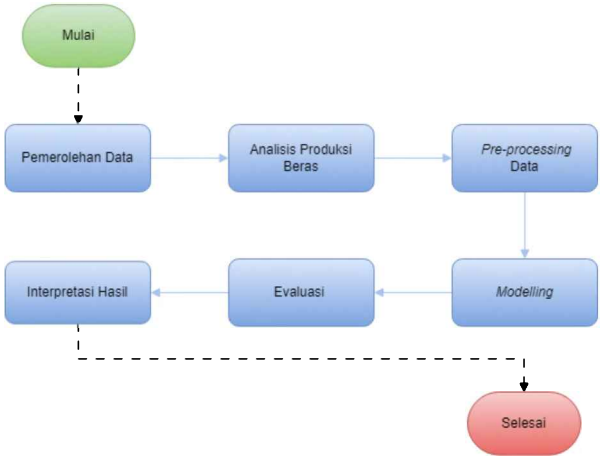
A. *Uraian Metodologi*

Rencana alur metodologi penelitian ini dirancang dengan tujuan untuk menyelesaikan penelitian secara sistematis dan efisien [9]. Berdasarkan hasil analisis yang telah dilakukan, dataset produksi beras menjadi fokus utama. Tahap awal melibatkan pengumpulan data yang komprehensif, termasuk variabel kunci seperti provinsi, tahun, produksi, luas panen, curah hujan, kelembapan, dan suhu rata-rata. Setelah data terkumpul, dilakukan analisis mendalam terhadap produksi padi untuk mengungkap pola dan variasi di berbagai provinsi. Proses *preprocessing* data kemudian diterapkan untuk memastikan kualitas dan kebersihan dataset sebelum dilibatkan dalam proses pemodelan.

Pendekatan LSTM dipilih untuk memodelkan produksi padi, diikuti dengan evaluasi kinerja menggunakan metrik yang relevan. Tahap akhir adalah interpretasi hasil, di mana temuan dari model LSTM dievaluasi untuk memberikan wawasan tentang hubungan antara faktor iklim dan produksi padi di Sumatera. Dengan demikian, metodologi ini diharapkan dapat memberikan kontribusi yang signifikan dalam pemahaman dan peningkatan efektivitas produksi padi di wilayah tersebut. alur metodologi tersebut diuraikan di bawah ini, mengikuti langkah-langkah dari workflow yang telah ditetapkan pada gambar 1:

1) **Pemerolehan Data**
   dataset produksi padi di Sumatera diperoleh dari situs Kaggle dengan variabel kunci mencakup provinsi, tahun, produksi, luas panen, curah hujan, kelembapan, dan suhu rata-rata.
2) **Analisis Produksi Beras**
   Dalam Analisis Produksi Padi, dilakukan analisis awal terhadap dataset untuk memahami pola produksi padi seiring waktu, dengan fokus pada identifikasi variabilitas dan tren di berbagai provinsi Sumatera yang rentan terhadap perubahan iklim.
3) ***Preprocessing* Data**
   Preprocessing Data melibatkan pembersihan dan persiapan data produksi padi Sumatera dengan normalisasi dan penanganan data yang hilang untuk memastikan kualitas dan integritas dataset sebelum memasuki tahap pemodelan.
4) ***Modelling***
   LSTM, varian dari RNN, diciptakan untuk mengatasi kelemahan RNN dalam memprediksi berdasarkan informasi jangka waktu panjang. Dengan desain untuk menyimpan dan mengelola informasi relevan dari waktu ke waktu, arsitektur Multivariate LSTM dipilih untuk memodelkan data produksi padi yang kompleks. Penggunaan Multivariate LSTM dalam produksi padi bertujuan memahami dan memprediksi pola produksi berdasarkan variabel seperti curah hujan, suhu, jenis tanah, dan faktor pertanian lainnya [10].
5) **Evaluasi**
   Evaluasi dilakukan menggunakan data pengujian untuk mengukur kinerja model dengan metrik seperti MSE, RMSE, atau metrik lain yang relevan, berdasarkan dataset produksi padi dari situs Kaggle.

6) **Interpretasi Hasil**
hasil prediksi dan analisis model dievaluasi, dan kesimpulan utama dari analisis hasil serta temuan yang dihasilkan dibahas secara mendalam. Dengan merujuk pada dataset dari Kaggle dan menggambarkan konteks pertanian padi di Sumatera, interpretasi hasil ini memberikan pemahaman lebih baik tentang dampak perubahan iklim terhadap produksi padi di wilayah tersebut.



Gambar 1. Metodologi Peramalan Produksi Padi Dengan LSTM

B. Pemerolehan Data

Dalam rangka mengkaji produksi tanaman padi di Sumatera, dilakukan pengumpulan data melalui unduhan dataset dari situs Kaggle yang terbuka untuk umum. Dataset yang berhasil diunduh mencakup rentang tahun 1993 hingga 2020 dengan kapasitas sebesar 12,73 KB. Informasi yang terdapat dalam dataset ini meliputi 7 atribut yang mendasar, seperti provinsi, tahun, jumlah produksi, luas lahan yang dipanen, curah hujan, kelembaban, serta suhu rata-rata. Hal ini memungkinkan analisis mendalam terkait hubungan dan faktor-faktor yang memengaruhi produksi tanaman padi di wilayah Sumatera.

C. Analisis Produksi Beras

Pulau Sumatera, sebagai salah satu pusat pertanian utama di Indonesia, menyumbang lebih dari 50 persen lahan pertanian di setiap provinsinya, dengan fokus utama pada produksi padi. Data yang diperoleh dari BPS dan BMKG dari tahun 1993 hingga 2020 mencakup variabel kunci seperti Provinsi, Tahun, Produksi, Luas Panen, Curah Hujan, Kelembapan, dan Suhu Rata-Rata. Peneliti menemukan bahwa dataset produksi beras memiliki 3 indikator utama:

1) Indikator Geografis dan Temporal:
   - Provinsi : Menentukan lokasi spesifik data di Pulau Sumatera.
   - Tahun : Menunjukkan periode waktu pengambilan data untuk analisis tren jangka panjang atau perbandingan tahunan.
2) Indikator Produksi Pertanian:
   - Produksi : Mengukur jumlah total padi yang diproduksi dalam satuan tertentu, memberikan indikasi langsung tentang output pertanian.
   - Luas Panen : Menunjukkan luas lahan yang digunakan untuk panen padi, menggambarkan skala kegiatan pertanian.
3) Indikator Lingkungan:
   - Curah Hujan : Mengukur jumlah curah hujan dalam satu tahun, menjadi faktor lingkungan penting yang mempengaruhi pertanian padi.
   - Kelembapan : Menunjukkan tingkat kelembapan rata-rata, memainkan peran dalam pertumbuhan tanaman padi dan meningkatkan pemahaman tentang risiko penyakit.
   - Suhu Rata-Rata : Mengukur suhu rata-rata tahunan, penting untuk siklus pertumbuhan tanaman padi.

D. Preprocessing Data

Proses pra-pemrosesan data untuk mempersiapkan dataset produksi tanaman padi di Sumatera melibatkan beberapa tahap. Dimulai dengan konversi ke *time series*, data diubah ke dalam format yang sesuai dengan pemodelan prediksi berbasis waktu. Langkah berikutnya mencakup filter dan pemisahan dataset, memfokuskan analisis pada data spesifik provinsi yang sedang diteliti. Setelah itu, nilai-nilai dalam dataset disesuaikan agar memiliki rentang antara 0 dan 1 melalui proses skalasi. Skalasi ini diterapkan untuk memastikan keseragaman nilai di semua fitur, sementara penyusunan urutan data latih dan uji memungkinkan model memahami pola dari sejarah data, mendukung kemampuannya meramalkan produksi tanaman padi di masa mendatang. Keseluruhan proses ini dirancang untuk mendukung implementasi model LSTM dalam memprediksi produksi tanaman padi di Sumatera berdasarkan atribut-atribut yang relevan [10][11]. Data kemudian dibagi menjadi dua bagian, di mana 75% digunakan untuk pelatihan dan 25% digunakan untuk pengujian. Dalam konteks pengujian, 25% tersebut terdiri dari data 7 tahun terakhir, yaitu 2014-2020. Berikut adalah rumus normalisasi *min-max scaling* [12]:

$$X' = \frac{(X - min_x)}{(max_x - min_x)} \quad (1)$$

E. Modelling

LSTM, varian unit dari RNN, unggul dalam prediksi dengan kemampuannya menangkap fitur jangka panjang dan mengatasi masalah vanishing gradient. Strukturnya mencakup cell gate, input gate, output gate, dan forget gate. Input gate mengatur masukan ke dalam cell, forget gate mengontrol nilai di dalam cell, dan output gate mengatur penggunaan nilai dalam cell untuk menghitung aktivasi keluaran. Arsitektur LSTM memproses data melalui forget gate dengan fungsi aktivasi sigmoid, memungkinkan pemilihan data yang akan disimpan dalam memory cell. Input gate memiliki dua gate dengan fungsi aktivasi sigmoid dan tanh untuk memperbarui dan menyimpan informasi baru di memory cell. Gate pertama mengubah data dengan aktivasi sigmoid, sementara gate kedua dengan aktivasi tanh membuat vektor nilai baru untuk disimpan dalam memory cell [13] Berikut adalah rumus LSTM [14][15][16][17]:

$$f_t = \sigma(w_f[V_{t-1}, X_t] + b_f) \quad (2)$$

$$m_t = \sigma(w_f[V_{t-1}, X_t] + b_m) \quad (3)$$

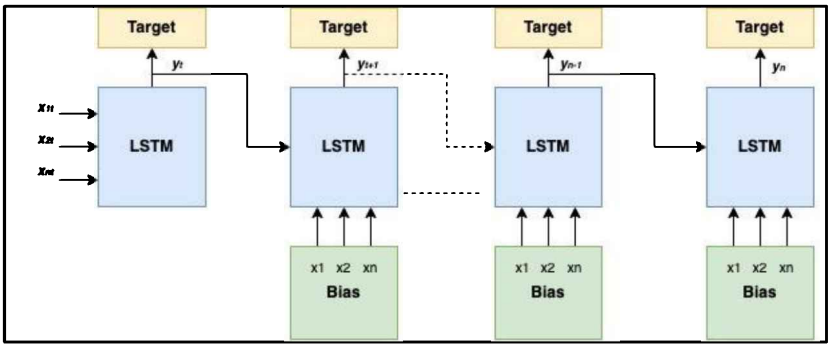
$$N_t = tanh(w_f[V_{t-1}, X_t] + b_n) \quad (4)$$

$$C_t = C_{t-1} f_t + N_t m_t \quad (5)$$

$$Q_t = \sigma(w_q[V_{t-1}, X_t] + b_q) \quad (6)$$

$$Q_t = Q_t tanh(C_t) \quad (7)$$

Multivariate LSTM, sebagai varian LSTM, dirancang untuk menangani permasalahan prediksi pada data multivariabel. Dengan struktur beberapa lapisan LSTM, masing-masing berfungsi untuk memodelkan pola pada fitur-fitur kompleks. Dalam konteks prediksi produksi tanaman padi di Sumatera, Multivariate LSTM dapat memanfaatkan berbagai fitur seperti curah hujan, suhu, jenis tanah, luas panen, dan produksi sebelumnya. Setiap fitur diintegrasikan dan diproses oleh lapisan-lapisan LSTM, memberikan pemahaman yang lebih komprehensif tentang hubungan kompleks antara variabel-variabel tersebut. Dengan menggunakan berbagai jenis gate, termasuk cell gate, input gate, output gate, dan forget gate, Multivariate LSTM mengontrol aliran informasi dan memori dalam jangka waktu yang panjang. Fleksibilitas yang diberikan oleh arsitektur Multivariate LSTM mendukung penanganan data produksi padi yang bersifat multivariabel, meningkatkan akurasi prediksi hasil pertanian. Dibawah ini pada gambar 2 adalah ilustrasi arsitektur jaringan Multivariate LSTM.



Gambar 2. Illustrasi Multivariate LSTM

F. Evaluasi

Dalam menentukan apakah model yang dibangun sudah baik dalam menangani data dibutuhkan yang namanya *evaluation* dalam hal ini digunakan *root mean square error (RMSE)* dan *mean square error (MSE).* Nilai MSE adalah metrik yang mengukur rata-rata dari kuadrat perbedaan antara nilai prediksi model dengan nilai aktual, dimana MSE dihitung dengan menjumlahkan kuadrat perbedaan antara setiap nilai prediksi dan nilai aktual, kemudian diambil rata-rata dari jumlah tersebut. RMSE adalah akar kuadrat dari MSE. Ini memberikan ukuran rata-rata dari kesalahan prediksi dalam satuan yang sama dengan variabel target. RMSE dihitung dengan mengambil akar kuadrat dari MSE. Berikut adalah rumus dari MSE dan RMSE [12][13]:

$$MSE = f(z) = \frac{1}{N}\sum_{i=1}^{N}(fi - yi)^2 \quad (8)$$

$$RMSE = \sqrt{\frac{\sum_{I=1}^{N}(fi - yi)^2}{n}} \quad (9)$$

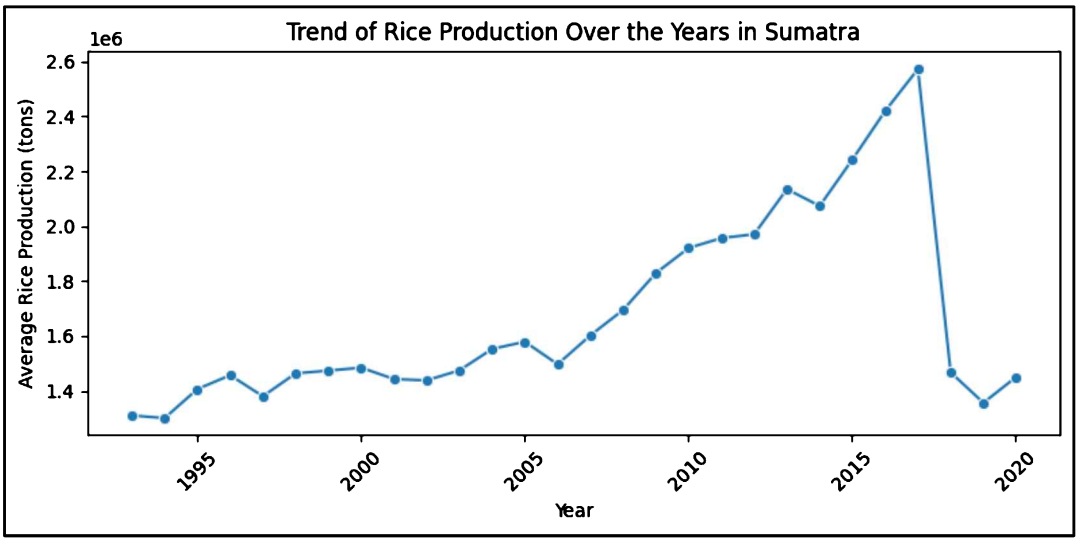
IX. EXPERIMENT AND ANALYSIS

A. Analisis Produksi Beras

Ada beberapa insight yang diperolah pada proses analisis mengenai data produksi tanaman padi di Sumatera, diantaranya :

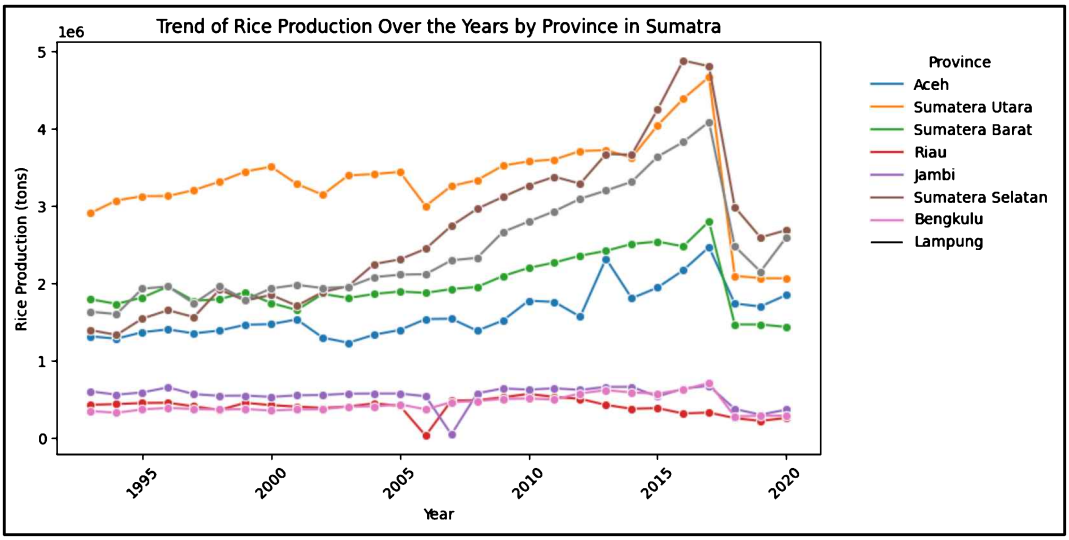
1) Tren Produksi Beras Selama Bertahun-Tahun di Sumatera

   Adanya pergerakan signifikan dalam tren produksi beras menunjukkan kenaikan yang tajam pada harga rata-rata produksi beras (dalam ton) setelah tahun 2015, yang kemudian diikuti dengan penurunan drastis di tahun 2020. Peristiwa penurunan tersebut dipengaruhi oleh beberapa faktor, di antaranya dampak wabah pandemi COVID-19. Situasi tersebut berpotensi mengganggu kinerja para petani dan pemerintah dalam upaya produksi padi serta manajemen hasilnya, menjadi salah satu penyebab utama dari penurunan signifikan dalam produksi dan harga beras pada tahun tersebut [18][19]. Adapun grafik tersebut dapat dilihat pada gambar 3.

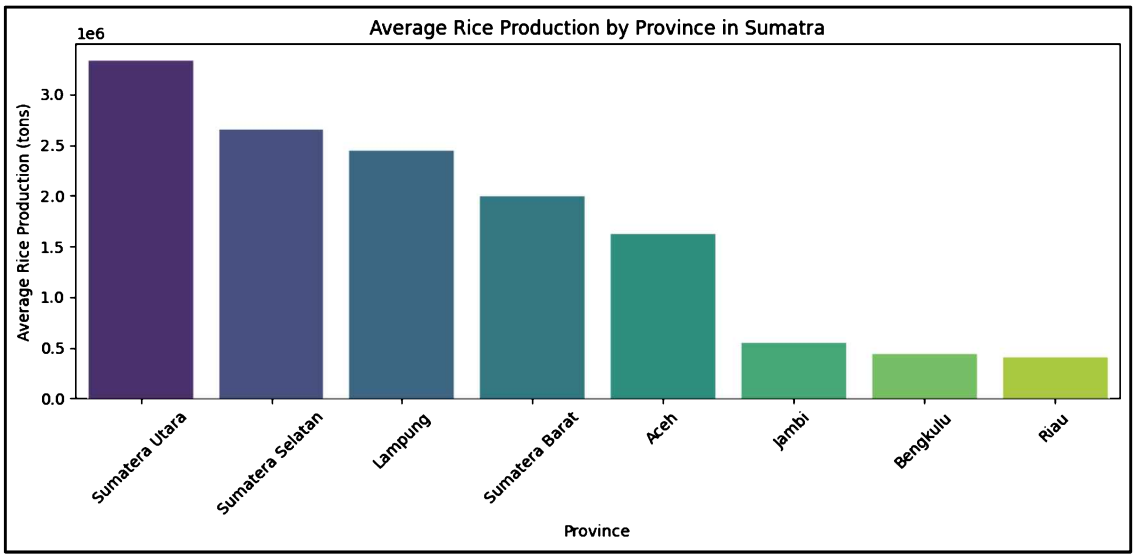


Gambar 3. Trend Produksi Beras Selama Bertahun-Tahun di Sumatra

2) Provinsi dengan rata-rata produksi beras tertinggi
Visualisasi diperluas dengan mengelompokkan tren produksi berdasarkan provinsi. Hal ini memungkinkan kita untuk membandingkan kinerja masing-masing provinsi secara individual [20]. Beberapa provinsi menunjukkan peningkatan produksi yang kuat dan stabil, sementara provinsi lainnya lebih bervariasi. Keragaman tren tersebut menunjukkan bahwa faktor regional memainkan peran penting dalam produksi beras. Sebagai contoh, Sumatera Utara, dengan harga rata-rata padi di atas 3 ton, menunjukkan tingkat produktivitas yang tinggi dalam pengolahan lahan padi, menggambarkan performa yang baik dalam produksi beras di sana. Di sisi lain, Riau, dengan harga rata-rata padi yang paling rendah, memberikan insight bahwa wilayah tersebut memiliki potensi untuk pengembangan pertanian. Analisis ini memberikan peluang untuk fokus pada pengembangan pertanian di Riau guna meningkatkan produktivitas dan kesejahteraan pertanian di wilayah tersebut. Data dari provinsi-provinsi dengan kinerja tinggi dapat memberikan wawasan mengenai praktik pertanian yang efektif. Adapun visualisasi dapat dilihat pada gambar 4 dan gambar 5.



Gambar 4. Provinsi dengan rata-rata produksi beras tertinggi



Gambar *5.* hasil rata-rata (produksi per hektar) di setiap provinsi

3) Perbedaan hasil rata-rata (produksi per hektar) di setiap provinsi

Provinsi di sebelah kiri, dengan batang yang lebih tinggi, merupakan produsen utama, hal ini menunjukkan bahwa provinsi tersebut memiliki kondisi atau praktik pertanian padi yang optimal. Sebaliknya, provinsi-provinsi di sebelah kanan, dengan batasan yang lebih pendek, mempunyai produksi yang lebih rendah. Melalui visualisasi diagram batang yang mewakili rata-rata hasil panen padi per hektar di setiap provinsi, terlihat bahwa Sumatera Barat menonjol dengan hasil panen di atas 5 ton per hektar, mencerminkan efisiensi tinggi dalam produksi padi. Provinsi ini mungkin mengoptimalkan penggunaan lahan melalui kualitas tanah yang baik, teknik pertanian yang unggul, atau faktor-faktor menguntungkan lainnya. Di sisi lain, provinsi-provinsi dengan batas hasil yang lebih rendah mungkin memiliki peluang untuk meningkatkan praktik pertanian atau menghadapi kendala lingkungan yang memengaruhi produktivitas. Informasi ini memberikan wawasan yang sangat penting dalam menilai efektivitas penggunaan lahan untuk produksi padi di seluruh Sumatera, dan memberikan dasar bagi strategi yang berkelanjutan dalam meningkatkan hasil panen padi di wilayah tersebut.

B. Preprocessing

Pada tahapan awal, data dibagi menjadi dilakukan filterisasi berdasarkan provinsi. Terdapat 8 provinsi yang akan dilakukan forecasting analysis secara individu. Tahapan selanjutnya data dibagi menjadi dua bagian, yaitu data latih sebesar 75% dan data uji sebesar 25%. Pada Tabel 1 data uji mencakup beberapa periode waktu pada 7 tahun terakhir, sementara data latih terdiri dari periode 21 tahun sebelumnya.

TABEL 1. SPLIT DATA

Split Data	
Train Data	***Test Data***
1993-2013	2014-2020

Setelah itu, dilakukan normalisasi data numerik menggunakan skala Min-Max (MinMaxScaler). Langkah ini penting agar seluruh fitur memiliki rentang nilai yang seragam, meningkatkan konvergensi model selama proses pelatihan. Langkah selanjutnya melibatkan konversi data ke bentuk sekuensi. Sekuensi input, yang merepresentasikan data historis, diambil dalam jendela waktu sepanjang n past (n=3). Sementara itu, variabel target (y_train dan y_test) menyimpan nilai yang ingin diprediksi, diambil pada beberapa langkah waktu ke depan n_future(n=1). Pada Tabel 2 hasil mengembalikan empat array yang telah diproses, yakni X_train, y_train, X_test, dan y_test dalam bentuk baru.

TABEL 2. KONVERSI BENTUK SEQUENCE

	X_train	***y_train***	***X_test***	***y_test***
New Shape	(18, 3, 4)	(18, 1)	(7, 3, 4)	(7, 1)

Keempat array ini dapat digunakan secara langsung untuk melatih dan menguji model LSTM, membantu model untuk memahami pola dalam data dan membuat prediksi produksi pertanian untuk masa depan berdasarkan sejarah pengamatan yang telah diberikan. Contoh hasil *preprocessing* dapat dilihat pada tabel 3.

TABEL 3. CONTOH HASIL PREPROCESING

		Feature_1	Feature_2	Feature_3	Feature_4
Sequence_1	0	0.000000	0.594530	0.546245	0.000000
	1	0.409487	1.000000	0.875233	0.177665
	2	0.413326	0.554475	0.862818	0.137056
Sequence_2	0	0.409487	1.000000	0.875233	0.177665
	1	0.413326	0.554475	0.862818	0.137056
	2	0.130349	0.279537	0.605214	0.241117
Sequence_3	0	0.413326	0.554475	0.862818	0.137056
	1	0.130349	0.279537	0.605214	0.241117
	2	0.442886	0.816667	0.953445	0.309645
Sequence_4	0	0.130349	0.279537	0.605214	0.241117
	1	0.442886	0.816667	0.953445	0.309645
	2	0.235991	0.727845	0.915580	0.129442

C. *Modelling*

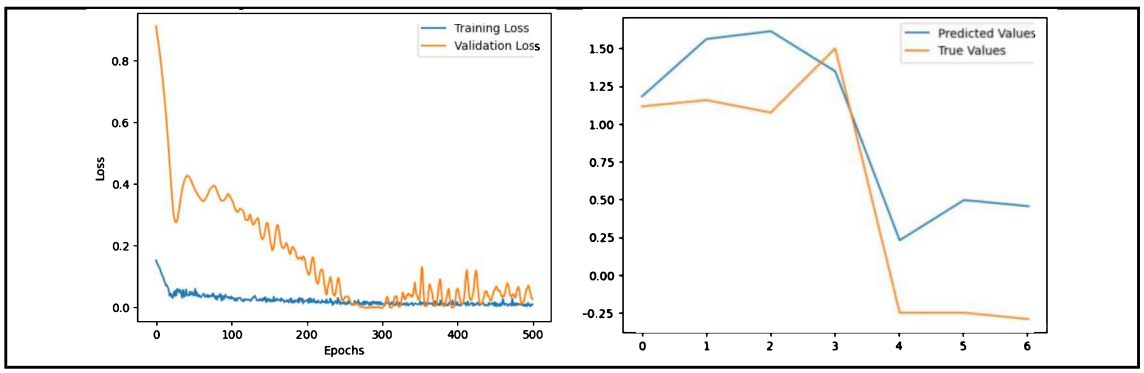
Proses pemodelan yang tercermin dari arsitektur pada Tabel 4 dimulai dengan definisi sebuah Sequential Model, yang merupakan model jaringan saraf tiruan (*neural network*) yang dibangun secara sekuensial. Arsitektur model ini memiliki beberapa lapisan yang dirancang untuk menangkap dan memahami pola dalam data time series. Pertama-tama, tiga lapisan LSTM (*Long Short-Term Memory*) berturut-turut digunakan untuk mengekstraksi informasi jangka panjang dari data sekuens waktu. Setiap lapisan LSTM memiliki jumlah unit yang berbeda dan berperan dalam memahami pola pada berbagai tingkatan kerumitan. Setelah itu, setiap lapisan LSTM diikuti oleh lapisan Dropout, yang berfungsi untuk mencegah overfitting dengan secara acak mengabaikan beberapa unit selama proses pelatihan. Lapisan Dense menyusul, bertanggung jawab untuk menghubungkan hasil dari lapisan LSTM ke unit-unit output. Lapisan Dense terakhir memiliki satu unit yang menghasilkan output akhir model.

TABEL 4. KONFIGURASI MODEL SEQUENTIAL 7

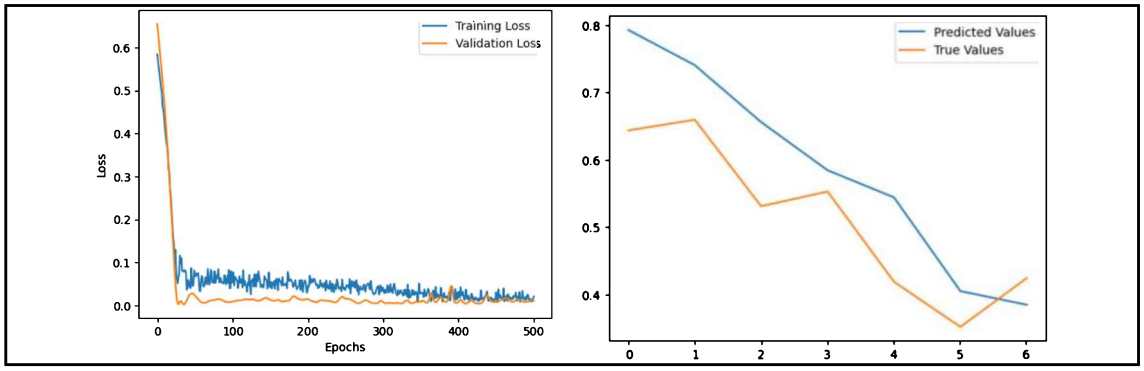
Layer (type)	*Output Shape*	*Parameter*
lstm_21 (LSTM)	(None, 3, 64)	17664
Dropoout_21 (Dropout)	(None, 3, 64)	0
Lstm_22 (LSTM)	(None, 3, 32)	12416
Dropout_22 (Dropout)	(None, 3, 32)	0
Lstm_23 (LSTM)	(None, 16)	3136
dense_14 (Dense)	(None, 32)	544
dropout_23 (Dropout)	(None, 32)	0
Dense_15 (Dense)	(None, 1)	33

D. *Evaluation*

Setelah melatih model LSTM menggunakan data latih, evaluasi dilakukan dengan menganalisis performa model. Grafik menampilkan perbandingan antara loss pada data latih dan validasi selama proses pelatihan. Hasil prediksi model terhadap data uji dievaluasi dengan membandingkan prediksi yang dihasilkan dengan data aktual. Ini memungkinkan untuk melihat seberapa baik model memahami pola yang ada dan sejauh mana kemampuannya dalam memprediksi produksi tanaman padi di Aceh. Dengan memeriksa perbedaan antara prediksi dan data aktual, evaluasi ini memberikan wawasan mengenai performa model dalam menghasilkan prediksi yang akurat yang dapat dilihat pada gambar 6 dan gambar 7 berikut ini.



Gambar 6. Grafik Training Validation Loss & Prediksi Panen Sumatera Barat



Gambar 7. Grafik Training Validation Loss & Prediksi Panen Riau

TABEL 5. TABEL HASIL TESTING SETIAP PROVINSI

No	Nama Daerah	Loss	RMSE
1	Aceh	0.2655	0.5153
2	Sumatera Utara	2.7013	1.6436
3	Sumatera Barat	0.2411	0.4910
4	Riau	0.0093	0.0964
5	Jambi	0.1024	0.3201
6	Sumatera Selatan	0.0583	0.2414
7	Bengkulu	0.2461	0.4961
8	Lampung	0.6725	0.8201

Tabel 5 merupakan hasil akurasi testing di setiap provinsi yang ada di Pulau Sumatra. Setiap provinsi menunjukkan hasil dengan tingkat keakuratan prediksi yang bervariasi. Salah satu yang menjadi sorotan adalah provinsi Sumatera Barat yang memiliki nilai Loss sebesar 0.2411 dan RMSE sebesar 0. 4910, dari hasil testing dengan nilai Loss dan RMSE menunjukkan hasil prediksi yang cukup baik. Kemudian Provinsi Riau menonjol dengan nilai Loss yang sangat rendah (0.0093) dan RMSE (0.0964) yang juga rendah, menandakan prediksi yang sangat akurat.

X. KESIMPULAN

Kesimpulan dari penelitian kami menunjukkan bahwa metodologi yang digunakan untuk meramalkan produksi beras di Sumatera dengan model LSTM memberikan pendekatan yang sistematis dan efisien. Dalam konteks ini, peran penting terdapat pada variabel-variabel tertentu dalam memprediksi produksi beras, yang pada gilirannya dapat membantu merinci faktor-faktor yang memengaruhi hasil ramalan di setiap Provinsi, sebagaimana diuraikan dalam Provinsi Aceh, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, dan Sumatera Utara dengan tingkat keakuratan yang beragam. Provinsi Aceh, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, dan Sumatera Utara. Masing-masing Provinsi menunjukkan hasil dengan tingkat keakuratan yang bervariasi. Salah satunya adalah provinsi Sumatera Barat memiliki nilai Loss sebesar 0.2411 dan RMSE sebesar 0. 4910, dari hasil testing dengan nilai Loss dan RMSE menunjukkan hasil prediksi yang cukup baik. Riau menonjol dengan nilai Loss yang sangat rendah (0.0093) dan RMSE (0.0964) yang juga rendah, menandakan prediksi yang sangat akurat. Secara keseluruhan, model LSTM memberikan hasil tentang pola produksi pertanian di berbagai provinsi, namun dengan tingkat keakuratan yang bervariasi. Hasil percobaan dengan model LSTM membuahkan kesimpulan bahwa metodologi ini dapat memberikan nnilai kontribusi yang signifikan dalam pemahaman dan peningkatan efektivitas produksi padi di Sumatera. Dengan menggunakan hasil dari pendekatan ini, analisis terhadap data-produksi beras dapat memberikan wawasan yang berharga untuk pengambilan keputusan dalam pengembangan strategi pertanian yang berkelanjutan di wilayah tersebut. Dengan demikian, penelitian ini memberikan pemahaman yang mendalam tentang penggunaan data science, khususnya model LSTM, dalam meramalkan dan menganalisis produksi beras di tingkat provinsi di Sumatera, serta memberikan dasar untuk pengembangan strategi pertanian yang berkelanjutan di wilayah tersebut.

REFERENSI

[1] E. Triyanto, H. Sismoro, and A. D. Laksito, "Implementasi Algoritma Regresi Linear Berganda Untuk Memprediksi Produksi Padi di Kabupaten Bantul," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 4, no. 2, pp. 66–75, Jul. 2019, doi: 10.36341/rabit.v4i2.666.

[2] M. Yusuf Nugroho and U. Duta Bangsa Surakarta, "Analisis Faktor-Faktor yang Mempengaruhi Produksi Padi di Sumatera Menggunakan Metode Regresi Linier," 2023.

[3] A. Satria, R. Maulida Badri, I. Safitri, and H. Artikel, "Prediksi Hasil Panen Tanaman Pangan Sumatera dengan Metode Machine Learning," *Digital Transformation Technology (Digitech) | e*, vol. 3, no. 2, 2023, doi: 10.47709/digitech.v3i2.2852.

[4] R. Maiyuriska, "Penerapan Jaringan Syaraf Tiruan dengan Algoritma Backpropagation dalam Memprediksi Hasil Panen Gabah Padi," *Jurnal Informatika Ekonomi Bisnis*, pp. 28–33, Mar. 2022, doi: 10.37034/infeb.v4i1.115.

[5] A. Satria, R. Maulida Badri, I. Safitri, and H. Artikel, "Prediksi Hasil Panen Tanaman Pangan Sumatera dengan Metode Machine Learning," *Digital Transformation Technology (Digitech) | e*, vol. 3, no. 2, 2023, doi: 10.47709/digitech.v3i2.2852.

[6] H. W. Herwanto, T. Widiyaningtyas, and P. Indriana, "Penerapan Algoritme Linear Regression untuk Prediksi Hasil Panen Tanaman Padi," 2019.

[7] A. Satria, R. Maulida Badri, I. Safitri, and H. Artikel, "Prediksi Hasil Panen Tanaman Pangan Sumatera dengan Metode Machine Learning," *Digital Transformation Technology (Digitech) | e*, vol. 3, no. 2, 2023, doi: 10.47709/digitech.v3i2.2852.

[8] K. C. Pelangi, "Prediksi Produksi Tanaman Pangan Di Provinsi Gorontalo Menggunakan Metode K-NN (K-Nearest Neighbor)," vol. 6, no. 2, 2021.

[9] Sudirman, A. P. Windarto, and A. Wanto, "Data mining tools | rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Oct. 2018. doi: 10.1088/1757-899X/420/1/012089.

[10] H. A. Zuhri and N. U. Maulidevi, "Product Review Ranking in e-Commerce using Urgency Level Classification Approach," *Jurnal Online Informatika*, vol. 5, no. 2, pp. 212–216, 2020, doi: 10.15575/join.

[11] M. Diqi, A. Sahal, and F. Nur Aini, "Multi-Step Vector Output Prediction of Time Series Using EMA LSTM," *Jurnal Online Informatika*, vol. 8, no. 1, pp. 107–114, Jun. 2023, doi: 10.15575/join.v8i1.1037.

[12] A. Farhah, A. L. Prasasti, and M. W. Paryasto, "Implementasi Recurrent Neural Network dalam Memprediksi Kepadatan Restoran Berbasis LSTM," *Jurnal Media Informatika Budidarma*, vol. 5, no. 2, p. 524, Apr. 2021, doi: 10.30865/mib.v5i2.2916.

[13] E. Ivan and H. D. Purnomo, "Forecasting Prices Of Fertilizer Raw Materials Using Long Short Term Memory," *Jurnal Teknik Informatika (Jutif)*, vol. 3, no. 6, pp. 1663–1673, Dec. 2022, doi: 10.20884/1.jutif.2022.3.6.433.

[14] I. Ketut, A. Enriko, F. Nizar Gustiyana, R. H. Putra, and K. Kunci, "Komparasi Hasil Optimasi Pada Prediksi Harga Saham PT. Telkom Indonesia Menggunakan Algoritma Long Short Term Memory," 2023, doi: 10.30865/mib.v7i2.5822.

[15] M. N. Ilyas and E. B. Setiawan, "Weight-Based Hybrid Filtering in a Movie Recommendation System Based on Twitter with LSTM Classification," vol. 7, pp. 1838–1849, 2023, doi: 10.30865/mib.v7i4.6668.

[16] A. B. Chopra and V. S. Dixit, "An adaptive RNN algorithm to detect shilling attacks for online products in hybrid recommender system," *Journal of Intelligent Systems*, vol. 31, no. 1, pp. 1133–1149, Jan. 2022, doi: 10.1515/jisys-2022-1023.

[17] S. T. T. Nguyen and B. D. Tran, "Long Short-Term Memory Based Movie Recommendation," *Science & Technology Development Journal - Engineering and Technology*, vol. 3, no. SI1, pp. SI1–SI9, Sep. 2020, doi: 10.32508/stdjet.v3isi1.540.

[18] S. Abbas and Z. A. Mayo, "Impact of temperature and rainfall on rice production in Punjab, Pakistan," *Environ Dev Sustain*, vol. 23, no. 2, pp. 1706–1728, Feb. 2021, doi: 10.1007/s10668-020-00647-8.

[19] B. Das, B. Nair, V. K. Reddy, and P. Venkatesh, "Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India," *Int J Biometeorol*, vol. 62, no. 10, pp. 1809–1822, Oct. 2018, doi: 10.1007/s00484-018-1583-6.

[20] J. Zhang *et al.*, "Effect of drought on agronomic traits of rice and wheat: A meta-analysis," *Int J Environ Res Public Health*, vol. 15, no. 5, May 2018, doi: 10.3390/ijerph15050839.

Klasifikasi Data Penumpang Titanic dengan *Ensemble Learning*: Perbandingan Hasil Voting Classifier

Nuzula Afini
Fakultas Sains dan Teknologi
Universitas Teknologi Yoagyakarta
Yogyakarta, Indonesia
nuzula.afini@gmail.com

Febrina Helmaputri
Fakultas Sains dan Teknologi
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
febrinahrp21@gmail.com

Meylany Putri Maharani
Fakultas Sains dan Teknologi
Universitas Teknologi Yogyyakarta
Yogyakarta, Indonesia
meylanyputrim@gmail.com

Abstrak— Tragedi tenggelamnya kapal Titanic pada tahun 1912 menjadi fokus penelitian yang kaya informasi terkait keselamatan penumpang. Meskipun machine learning telah berhasil dalam berbagai aplikasi, kompleksitas dan ketidakseimbangan data dalam prediksi keselamatan penumpang menantang. Studi ini mengeksplorasi resampling (Majority Methods, SMOTE Oversampling) dan teknik sampling data (Random Sampling, Stratified Sampling) untuk menangani ketidakseimbangan data. Untuk mengatasi tantangan tersebut, diterapkan teknik Ensemble Learning, terutama Hard Voting. Dengan menggunakan berbagai algoritma (Logistic Regression, MLP Classifier, Gaussian Naïve Bayes, dll.), Ensemble Learning meningkatkan akurasi prediksi keselamatan penumpang. Ilustrasi konsep ensemble learning voting diperlihatkan, dengan penelitian ini fokus pada penerapannya pada dataset Titanic. Metode ini memanfaatkan mayoritas suara dari berbagai algoritma yang terlibat untuk mengambil keputusan akhir. Penelitian terkait menunjukkan kesuksesan Ensemble Learning pada berbagai domain. Penelitian ini mencatat langkah penting dalam optimalisasi teknik Ensemble Learning, terutama dalam konteks data tidak seimbang, untuk prediksi keselamatan penumpang. Bagian berikutnya akan membahas teknik, eksperimen, hasil, dan kesimpulan penelitian ini.

Kata Kunci—Ensemble Learning, Tenggelamnya Kapal Titanic, Resampling, Hard Voting, Akurasi Prediksi.

I. PENDAHULUAN

Tenggelamnya kapal Titanic pada tahun 1912 merupakan salah satu tragedi yang berhasil mencuri perhatian dalam sejarah pelayaran dunia. Data penumpang terkait dengan peristiwa tersebut berhasil menawarkan kesempatan studi analisa yang kaya akan informasi, sehingga memungkinkan aplikasi metode pembelajaran mesin untuk menganalisis dan memprediksi faktor-faktor yang memengaruhi keselamatan penumpang. Di sisi lain, *Machine Learning* telah berhasil mencapai kesuksesan yang signifikan dalam meningkatkan performa dalam beberapa kasus berbagai aplikasi, seperti prediksi [1], pengenalan pola [2], klasifikasi [3], terutama dalam analisis data tabular.

Namun, seiring dengan meningkatnya kompleksitas data dan keragaman studi kasus serta kecenderungan data yang tidak seimbang dalam distribusi informasi dalam mengoptimalkan prediksi keselamatan penumpang seringkali satu algoritma *Machine Learning* tidak cukup efektif untuk menghasilkan prediksi yang akurat. Dalam studi kasus serta pembahasan yang akan dianalisa, penelitian ini menyoroti kompleksitas dan ketidakseimbangan data yang digunakan perlu mempertimbangkan penggunaan teknik *Resampling* seperti *Majority Methods* dan *SMOTE Oversampling* untuk menangani ketidakseimbangan, serta teknik sampling data yaitu *Random Sampling* dan *Stratifield Sampling* untuk pembagian *Train-Test Data*.

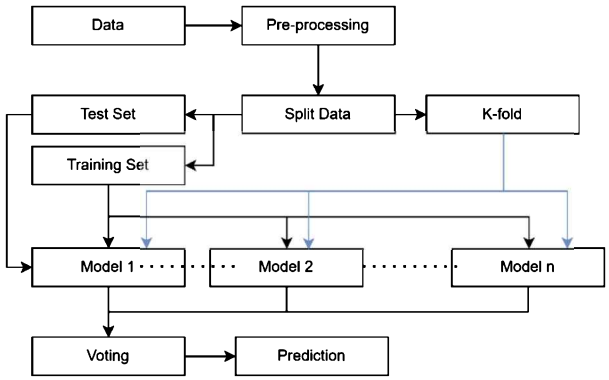
Selain itu untuk mengatasi tantangan, penelitian ini mengeksplorasi cara penerapan *Voting Classifier* dengan salah satu teknik *Ensemble Learning* yang cukup popular, dengan pemungutan suara atau voting yaitu metode *Hard Voting*. Hal ini memungkinkan penggabungan beberapa model untuk meningkatkan efektivitas dan mengatasi kelemahan individu, yang tidak hanya untuk meningkatkan akurasi prediksi kelangsungan hidup penumpang, tetapi juga memperbaiki metode *Voting* yang digunakan. Dalam pendekatan *Voting*, hasil akhir diperoleh dengan mempertimbangkan mayoritas suara dari berbagai model [4]. *Ensemble Learning* melibatkan pelatihan data dengan beberapa algoritma *Machine Learning*.

Algoritma-algoritma yang digunakan dapat memiliki jenis yang sama *(Homogeneous Ensemble Learning)* atau berbeda *(Heterogeneous Ensemble Learning)*. Dalam konteks *Ensemble Learning Voting,* kita mengadopsi pendekatan *Heterogeneous Ensemble Learning* [5]. Output dari pembelajaran berasal dari berbagai metode yang digunakan, dan perbedaan antara metode-metode ini dapat menghasilkan hasil yang berbeda tergantung pada faktor-faktor seperti algoritma yang digunakan [4]. Penelitian terdahulu juga telah mengaplikasikan *Ensemble Learning Voting* dalam berbagai domain, seperti analisis deteksi malware [4] dan deteksi uang kertas palsu [6].

Dalam konteks lebih spesifik, penelitian ini fokus pada pemanfaatan *ensemble learning* untuk meningkatkan performa prediksi. Kami akan menjelaskan bagaimana *ensemble learning* dengan menggunakan algoritma-algoritma yang berbeda dapat membantu mengatasi kelemahan masing-masing algoritma [4]. Dengan melibatkan multiple algoritma dalam proses pembelajaran, kami bertujuan untuk mengurangi risiko terjadinya jenis kesalahan yang sama dan meningkatkan akurasi prediksi.

Pada gambar 1, kami mengilustrasikan konsep *ensemble learning voting*. Voting digunakan sebagai hasil akhir dari prediksi. Dalam penelitian ini, kami akan menerapkan metode *ensemble learning voting* pada dataset Titanic. Kami

akan melatih dataset dengan beberapa algoritma berbeda, dan hasil akhirnya akan diambil dengan mempertimbangkan mayoritas suara dari berbagai algoritma yang terlibat.



Gambar 1. Ilustrasi konsep *ensemble learning voting*

Selain itu, penelitian ini mengeksplorasi metode Ensemble Learning untuk meningkatkan akurasi prediksi. Dalam eksperimen ini, berbagai algoritma dilatih pada dataset Titanic, dan hasil akhir diputuskan berdasarkan akurasi terbaik dari mayoritas suara algoritma-algoritma tersebut, yaitu menggunakan *Logistik Regression, MultiLayer Perceptron Classifier, Gaussian Naïve Bayes, XGB Classifier, Decision Tree, Support Vecktor Machine, Neural Network Perceptron, Random Forest, Stochastic Gradient Decent, Linear SVC,* dan *K Nearest Neighbors.*

Beberapa penelitian terkait menggunakan metode *Ensemble Learning* yaitu terkait pendekatan pemroresan data EKG terhadap deteksi kelelahan berkendaran dengan empat klasifikasi model yang menghasilkan akurasi optimal 98.82% pada set latih dan 81.82% pada set test [5]. Penelitian lainnya yaitu penerapan *Ensemble Learning* pada *chatbot* untuk prediksi jawaban menggunakan 5 metode klasifikasi dan *hard voting* atau *majority voting* dengan hasil akurasi yang didapatkan 86% pada set test pada enam kelas data [6]. Penelitian lainnya terkait tinjaun komprehensif terkait *Ensemble Learning*, dalam penelitian tersebut terdapat dua faktor keberhasilan *Ensemble Learning* yaitu bagaimana algoritma dasar dilatih dan bagaimana algoritma tersebut digabungkan [7].

Penelitian ini menandai langkah penting dalam upaya mengoptimalkan teknik Ensemble Learning dalam situasi data yang tidak seimbang, terutama dalam prediksi keselamatan penumpang dalam peristiwa sejarah seperti tenggelamnya kapal Titanic. Bagian berikutnya menjelaskan struktur *paper* penelitian ini: bagian 2 menyajikan teknik yang digunakan dalam studi penelitian yang akan disajikan kedalam bab metode penelitian, bagian 3 eksperimen penelitian dan hasilnya, dan bagian 4 menutup makalah dan diskusi.

II. METODE PENELITIAN

Pada bagian ini akan membahas sebuah teknik serta metode yang akan digunakan untuk menjelaskan penelitian ini, selain itu penulis akan membahas secara rinci bagaimana kontribusi dari setiap komponen yang terlibat dalam merancang algoritma yang dimaksud, berikut beberapa metode yang digunakan:

A. .Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) merupakan suatu pendekatan analisis data yang dilakukan oleh penulis, bertujuan untuk memahami struktur dan karakteristik dataset sebelum menerapkan model atau statistik lebih lanjut. Pada dataset Titanic, dataset tersebut terdiri atas 2 dokumen dataset CSV yaitu dataset Train dan daset Test. Dataset Train memiliki 891 *instances*, 11 *attribute* fitur dan 1 *attribute class "Survived", sedangkan dalam dataset Test hanya terdapat* 418 *instances*, 11 *attribute* fitur dan tanpa memiliki *attribute class " Survived ".*

Selain itu, data tersebut memiliki *missing value*, dalam dataset Train missing value atau data kosong tersebut terdapat dalam kolom Age sebanyak 177 data, Cabin sebanyak 687 data, serta Embarked sebanyak 2 data. Sedangkan dalam dataset Test data kosong tersebut terdapat dalam kolom Age sebanyak 86 data, Cabin sebanyak 327 data, serta Fare sebanyak 1 data. Distribusi kelas pada data yang digunakan untuk kelas 0 terdapat 549 *intances* dan kelas 1 terdapat 342 *intances*, data kelas 0 menunjukan bahwa penumpang tersebut tidak selamat, dan data kelas 1 menunjukan bahwa penumpang tersebut selamat dari kejadian tersebut, dengan kondisi seperti ini dapat dikategorikan bahwa kelasnya tidak seimbang. Ketidakseimbangan terhadap data ini dapat mempengaruhi kinerja model *Machine Learning*, dan oleh karena itu, memerlukan strategi penanganan yang tepat.

B. Data Preprocessing

Preprocessing data adalah langkah kunci dalam persiapan dataset Titanic. Beberapa tahap preprocessing yang perlu dilakukan adalah:

1. Penanganan Nilai Hilang
   Nilai-nilai yang hilang pada atribut seperti "Age", "Embarked", "Fare" dan "Cabin" akan diisi dengan nilai yang sesuai, seperti rerata usia penumpang atau label yang menunjukkan bahwa data hilang.
2. Konversi Atribut Kategorikal
   Beberapa atribut, seperti "Sex" dan "Embarked," merupakan atribut kategorikal yang perlu diubah menjadi representasi angka untuk digunakan dalam algoritma machine learning.
3. Normalisasi Data
   Beberapa atribut mungkin memerlukan normalisasi, terutama atribut yang memiliki rentang nilai yang berbeda.
4. Pemilihan Atribut
   Dalam beberapa kasus, mungkin diperlukan pemilihan atribut untuk mengidentifikasi atribut yang paling relevan dalam prediksi kelangsungan hidup.
5. Pembagian Data
   Dataset akan dibagi menjadi dua subset, yaitu data pelatihan dan data pengujian, untuk evaluasi model sebagai berikut:
   a. *Stratified Sampling Stratified Sampling* adalah metode pengambilan data yang membagi data menjadi subset homogen, yang disebut strata. Setiap subset dalam metode sampling ini mempertahankan persentase strata atau rasio tiap kelasnya.
   b. *Random Sampling Random Sampling* adalah metode pengambilan sample subset data secaran bebas pada set data. Pada setiap sample memiliki probabilitas yang sama untuk masuk dalam subset.

6. Penggabungan kolom Kolom dalam dataset bisa digabunggkan jika terdapat hubungan korelasi.

7. Imbalance Dataset
   Imbalance dataset adalah keadaan dimana data kelas lain lebih banyak dari kelas lainnya. *Imbalance dataset* tersebut dapat menghasilkan pengambilan kesimpulan yang salah [8]. Berikut cara mengatasi *imbalance dataset:*
   a. *Undersampling* Salah satu pendekatan untuk mengatasi kelas *imbalance* yaitu dengan menghapus data pada kelas yang lebih banyak.
   b. *Oversampling Oversampling* yaitu salah satu pendekatan mengatasi kelas *imbalance* dimana data pada kelas yang lebih kecil ditambahkan, di-*copy,* atau dibuat.
   c. *SMOTE (Synthetic Minority Oversampling Technique) SMOTE* adalah metode *oversampling* yang membuat data sintesis yang merupakan interpolasi dari kelas minoritas [9].

C. Data Modelling

Algoritma yang digunakan untuk perbandingan pada penelitian ini adalah algoritma *Logistic Regression, Multilayer Perceptron, Support Vector Machine, k-Nearest Neighbors,* dan *Decision Tree* yang nantinya akan dilakukan penggabungan. Penggabungan hasil prediksi kelima model algoritma dengan teknik *ensemble learning* menggunakan skema *voting classifier.* Setiap model akan memberikan suara dan suara yang paling banyak akan menjadi hasil prediksi ensemble. Berikut adalah penjelasan algoritma-algoritma yang digunakan:

a.) Logistic Regression

Logistic Regression (LR) adalah salah satu metode paling populer yang digunakan untuk mengklasifikasikan data biner. LR didasarkan pada asumsi bahwa nilai variabel terikat diprediksi dengan menggunakan variabel bebas. Dalam modelnya, Y adalah variabel terikat yang coba kita prediksi dengan mengamati X yang merupakan masukan atau himpunan dari variabel bebas ($x_i, \dots, x_k$). Nilai Y yang sesuai dengan orang yang selamat (Y=1) atau tidak selamat (Y=-1) dan diringkas dengan (X=x). Dari definisi ini, probabilitas bersyarat mengikuti distribusi logistik yang diberikan oleh P(Y = 1|X= x_i). Fungsi ini disebut fungsi regresi yang perlu kita prediksi Y.[10]

b.) Multilayer Perceptron

Multilayer perceptron (MLP) merupakan salah satu jenis ANN yang mempunyai kemampuan menyelesaikan masalah klasifikasi nonlinier dengan akurasi tinggi dan kinerja generalisasi yang baik. MLP telah diterapkan pada berbagai macam tugas seperti pemilihan fitur, pengenalan pola, optimasi, dan sebagainya. MLP dapat dianggap sebagai grafik berarah di mana neuron buatan disajikan dengan node dan tepi berarah dan berbobot menghubungkan node satu sama lain. Node disusun menjadi beberapa lapisan: *input layer*, satu atau lebih *hidden layer*, dan *output layer*. MLP menggunakan *backpropagation* untuk mengklasifikasikan titik data dan dengan menggunakan *backpropagation*, kesalahan disebarkan ke arah belakang untuk menyesuaikan bobot.[10]

c.) Support Vector Machine

Support Vector Machine (SVM) yang dikembangkan oleh Vapnik pada tahun 1995 didasarkan pada prinsip minimalisasi risiko struktural yang menunjukkan kinerja generalisasi yang baik. Dengan SVM, diusulkan untuk menemukan hyperplane pemisah optimal antar kelas dengan berfokus pada vektor pendukung. Hyperplane ini memisahkan data pelatihan dengan margin maksimal. SVM memecahkan masalah nonlinier dengan memetakan titik data ke dalam ruang berdimensi tinggi.[10]

d.) K-Nearst Neighbors

K-Nearst Neighbors (k-NN) adalah salah satu algoritma klasifikasi yang paling umum, paling sederhana dan non parametrik ketika hanya ada sedikit atau tidak ada pengetahuan sebelumnya tentang distribusi data. Dengan menggunakan metrik jarak untuk mengukur kedekatan antara sampel pelatihan dan sampel pengujian, k-NN menetapkan sampel pengujian dengan kelas dari k sampel pelatihan terdekatnya. Dalam hal kedekatan, k-NN sebagian besar didasarkan pada jarak *Euclidean*. Jarak *Euclidean* antara sampel pelatihan $X_1 = (x_{11}, x_{12}, \dots, x_{1N})$ dengan N *features*, sampel uji $X_2 = (x_{21}, x_{22}, \dots, x_{2N})$ dengan N *features* dan m=2 adalah jarak $(X, X_2) = (\sum_{i=1}^{N}(x_{1i} - x_{2i})^m)^{1/m}$. Jika m = 1 jarak tersebut disebut *Manhattan* dan m > 2 jarak tersebut disebut *Minkowski*. [10]

e.) Decision Tree

Decision Tree atau pohon keputusan dengan struktur pembuatannya yang cukup sederhana adalah salah satu pengklasifikasi yang paling banyak digunakan. Pohon keputusan adalah model terstruktur pohon dengan simpul keputusan dan simpul prediksi. Node keputusan digunakan untuk membuat cabang dan node prediksi menentukan label kelas. C4.5 adalah sejenis algoritma pohon keputusan yang membangun pohon keputusan dari data pelatihan dengan menggunakan perolehan informasi. Saat membangun pohon keputusan C4.5 menggunakan pendekatan membagi dan menaklukkan. [10].

f.) Naïve Bayes

Naive Bayes adalah metode statistik dalam klasifikasi yang memungkinkan kita untuk memprediksi probabilitas sebuah data tertentu akan termasuk dalam kelas tertentu dengan menggunakan perhitungan probabilitas [11]. Dalam kasus di dunia nyata, naïve bayes berasumsi secara independent.

g.) Linear Support Vector Classifier

Linear support vector classifier adalah klasifikasi biner yang dilakukan dengan mencari *hyperline* yang dapat memaksimalkan dua kelas margin. Algoritma ini bagus untuk dataset kecil dan bersih.

h.) Perceptron

Perceptron adalah algoritma *supervised learning* untuk melakukan klasifikasi dua kelas, bagian dari algoritma neural network yang bekerja.

i.) Voting

Untuk mendapatkan hasil akurasi klasifikasi, dapat dilakukan penggabungan beberapa algoritma dalam *real-world dataset*. Untuk mendapatkan hasil, dari beberapa algoritma tersebut akan dilakukan pemungutan suara yang disesuaikan dengan aturan [4]. Sebagian pemungutan suara dilakukan dengan pemunguran suara terbanyak, pemungutan suara lainnya dilakukan dengan memberikan bobot pada algoritma.

j.) Gradient Descent Classifier (SGDClassifier)

SGCClassifier adalah algortima klasifikasi yang termasuk dalam linear model. Algoritma Stochastic Gradient

Descent (SGD) digunakan untuk menemukan nilai terbaik dari koefisien dalam model klasifikasi linear. Caranya adalah dengan terus-menerus memperbarui parameter model menggunakan informasi dari gradien (kemiringan) *loss function.*

k.) Random Forest

Random Forest adalah *supervised learning* yang bisa digunakan untuk klasifikasi dan regresi. *Random Forest* akan memilih secara acak data sample, membuat pohon keputusan, memperoleh hasil prediksi dari setiap pohon, selanjutnya memilih opsi pilihan terbaik.

l.) XGBoost

XGboost adalah metode *ensemble learning* yang mengkombinasikan beberapa *decision tree* untuk membuat model prediktif yang kuat. *XGboost* membangun pohon keputusan secara berurutan, dimana pada tiap pohon selanjutnya akan memperbaiki dari system kesalahan pohon sebelumnya *XGBoost* mengoptimalkan *loss function* yang dapat dibedakan untuk meminimalkan kesalah prediksi.

D. Evaluasi Model

Tahap terakhir pada penelitian ini adalah evaluasi terhadap model untuk melihat kualitas serta efektifitas model yang diterapkan pada tahap pemodelan. Serta menentukan apakah model yang terbentuk benar-benar sudah mencapai tujuan untuk diambil keputusan berdasarkan penggunaannya. Evaluasi model pada dataset ini menggunakan *cross-validation* untuk membantu menghindari *overfitting* dan memberikan perkiraan yang lebih stabil dalam kinerja model akibat dari data yang besar. Cross-validation melibatkan pembagian data train menjadi beberapa lipatan (*folds*) dan melatih model pada beberapa kombinasi lipatan dan menguji pada lipatan yang tersisa. Dalam kasus ini menggunakan *10-fold cross-validation.*

E. Evaluasi Model

Tahap terakhir pada penelitian ini adalah evaluasi terhadap model untuk melihat kualitas serta efektifitas model yang diterapkan pada tahap pemodelan. Serta menentukan apakah model yang terbentuk benar-benar sudah mencapai tujuan untuk diambil keputusan berdasarkan penggunaannya. Evaluasi model pada dataset ini menggunakan *cross-validation* untuk membantu menghindari *overfitting* dan memberikan perkiraan yang lebih stabil dalam kinerja model akibat dari data yang besar. Cross-validation melibatkan pembagian data train menjadi beberapa lipatan (*folds*) dan melatih model pada beberapa kombinasi lipatan dan menguji pada lipatan yang tersisa. Dalam kasus ini menggunakan *10-fold cross-validation.*

III. EKSPERIMEN

A. Dataset

Dataset yang digunakan yaitu *Titanic – Machine Learning from Disaster* dari Kaggle [9], yang telah menjadi salah satu dataset terkenal dalam komunitas machine learning. Dataset Titanic berasal dari informasi penumpang kapal RMS Titanic yang tersedia secara historis. Data ini telah dikumpulkan dan disusun untuk berbagai studi terkait prediksi kelangsungan hidup penumpang. Dataset ini telah menjadi sumber belajar yang populer dalam pengembangan model machine learning dan penelitian terkait analisis data.

Dataset Titanic terdiri dari atribut-atribut yang mencakup informasi tentang penumpang, yaitu *Survival, Pclass* (kelas tiket), *Sex, Age, Sibsp* (jumlah saudara atau pasangan), *Parch* (jumlah relasi keluarga), *Ticket, Fare, Cabin,* dan *Embarked*. Atribut yang paling penting dalam penelitian ini adalah "Survived" yang mengindikasikan apakah penumpang selamat (1) atau tidak (0).

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Gambar 2. Dataset sebelum Preprocessing

Dapat terlihat dalam gambar 2, data tersebut memiliki banyak kolom yang tidak sesuai dengan kebutuhan dalam pemrosesan algoritma data. Hasil preprocessing dapat dilihat pada gambar 3 berikut ini.

	PassengerId	Survived	Pclass	Name	Sex	Age	Ticket	Fare	Cabin	Embarked	Relative
288	289	1	2	343	1	42	122	13	8	2	0
697	698	1	3	515	0	25	392	7	8	1	0
298	299	1	1	655	1	38	89	30	2	2	0
208	209	1	3	124	0	16	418	7	8	1	0
483	484	1	3	758	0	63	462	9	8	2	0

Gambar 3. Dataset setelah Preprocessing

Pengecekan kembali pada data setelah dipreprocessing sebagai bentuk penerapan data yang baik untuk algoritma pengolahan selanjutnya. Hasil pengecekan kembali pada dataset yang digunakan terdapat pada gambar 4 berikut ini.

	Data Kosong	Data Duplikat	Data NaNN	Type Data
PassengerId	0	0	0	int64
Survived	0	0	0	int64
Pclass	0	0	0	int64
Name	0	0	0	int64
Sex	0	0	0	int64
Age	0	0	0	int64
Ticket	0	0	0	int64
Fare	0	0	0	int64
Cabin	0	0	0	int64
Embarked	0	0	0	int64
Relative	0	0	0	int64

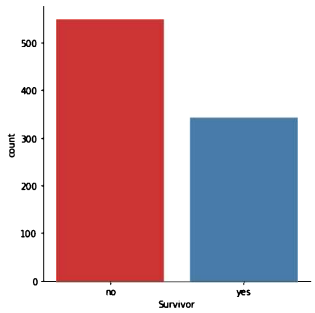
Gambar 4 Dataset setelah Preprocessing

Selain itu, dataset juga berisi beberapa nilai yang hilang, yaitu pada atribut *Age, Embarked,* dan *Cabin*. Untuk detail atribut kami melakukan analisis sebagai berikut:

1) Survival

Pada atribut "Survival", terdapat 549 penumpang tidak selamat dan 342 penumpang selamat. Dapat dianalisa bahwa persentase penumpang selamat lebih rendah dari

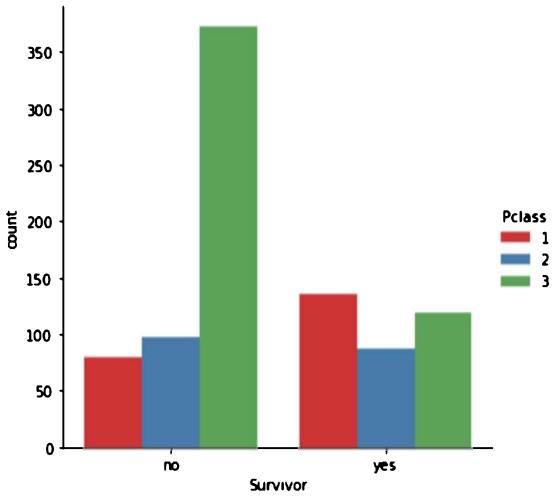
penumpang yang tidak selamat. Grafik jumlah penumpang pada atribut survival terdapat pada gambar 5 berikut ini.



Gambar 5. Distribusi atribut "Survival"

2) Pclass

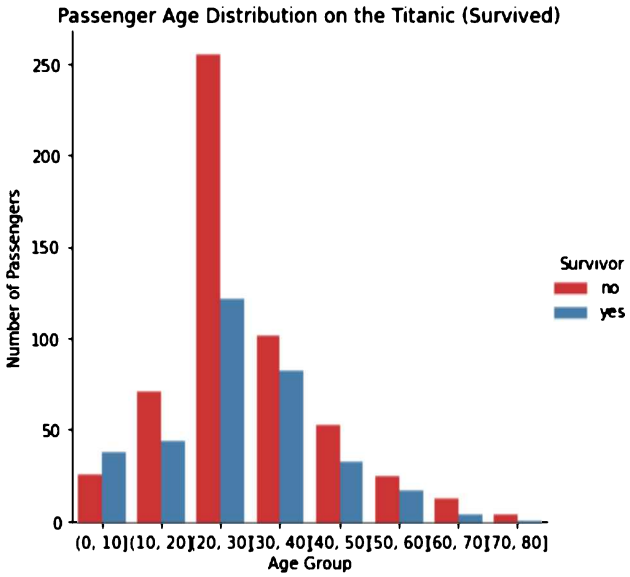
Pada atribut "Pclass" terdapat tiga jenis yaitu 1st *(upper)*, 2nd *(middle)*, dan 3rd *(lower)* seperti yang terlihat pada gambar 6, hal ini yang menandakan sosial-ekonomi status. Dapat dianalisa bahwa status sosial-ekonomi mempengaruhi keadaan, *3rd class* dengan penumpang terbanyak yaitu 491 hanya selamat 24.23% yaitu 119 penumpang. *2nd class* dengan 184 penumpang memiliki persentase penumpang selamat 47.28% yaitu 87 penumpang. *1st class* dengan 216 penumpang memiliki persentase 62.96% yaitu 136 penumpang. Hal ini menandakan semakin tinggi "Pclass" penumpang semakin tinggi kemungkinan penumpang selamat.



Gambar 6. Distribusi atribut "Pclass"

3) Age

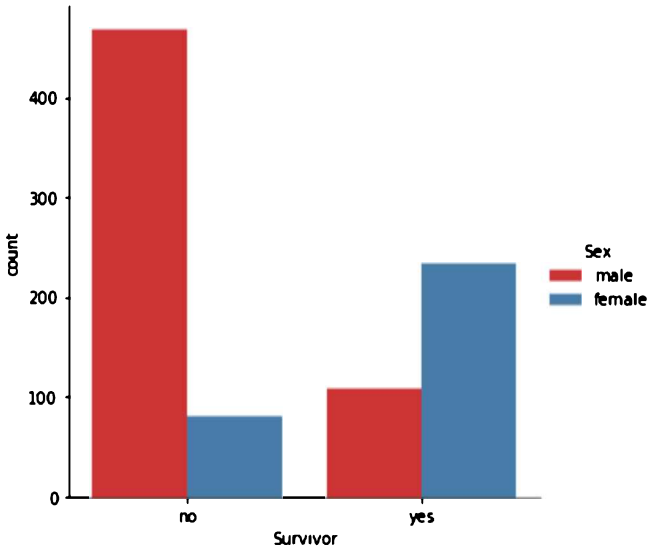
Pada atribut "Age" memiliki range dari 0 sampai 80. Jika dibuat interval tiap 10 tahun, maka interval 20 sampai 30 memiliki angka penumpang selamat dan tidak selamat tertinggi. Sedangkan interval 70 sampai 80 memiliki memiliki angka penumpang selamat dan tidak selamat terendah. Distribusi data pada atribut "Age" terdapat pada gambar 7.



Gambar 7. Distribusi atribut "Age"

4) Sex

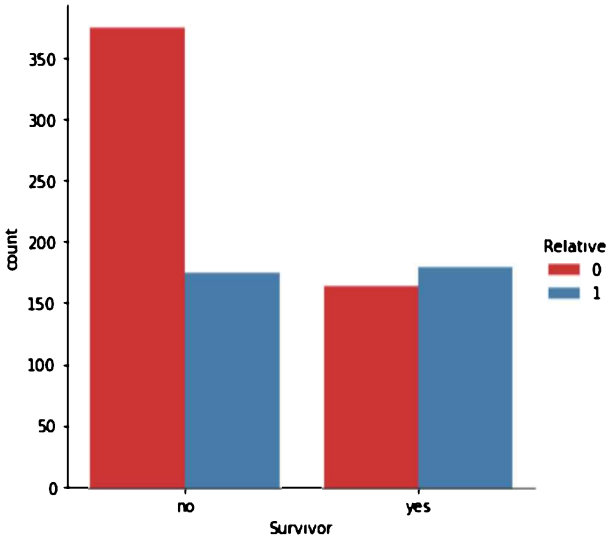
Pada atribut "Sex" terdapat 577 pria dan 314 wanita. Pada penumpang pria didapatkan persentase penumpang selamat sebesar 18.89% yaitu 109 penumpang. Sedangkan pada penumpang wanita didapatkan pesersentase penumpang selamat sebesar 74.2% yaitu 233. Hal ini menunjukkan bahwa wanita didahulukan untuk evakuasi saat kapal mulai tenggelam. Distribusi atribut jenis kelamin terdapat pada gambar 8 berikut ini.



Gambar 8. Distribusi atribut "Sex"

5) Relative

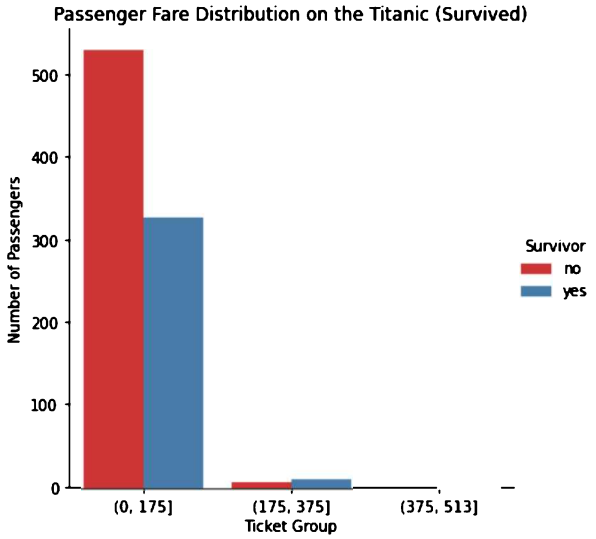
Atribut ini merupakan gabungan dari atribut "Sibsp" dan "Parch" yaitu atribut tentang relasi keluarga. Atribut ini akan memberikan tentang penumpang memiliki relasi atau tidak. Penumpang tanpa relasi 69.65% tidak selamat yaitu 374 penumpang sedangkan 163 penumpang tanpa relasi selamat. Penumpang dengan relasi memiliki persentase selamat 50.56% yaitu 179 penumpang sedangkan 175 penumpang tidak selamat. Distribusi atribut "Relative" terdapat pada gambar 9 berikut ini.



Gambar 9. Distribusi atribut "Relative"

6) Fare

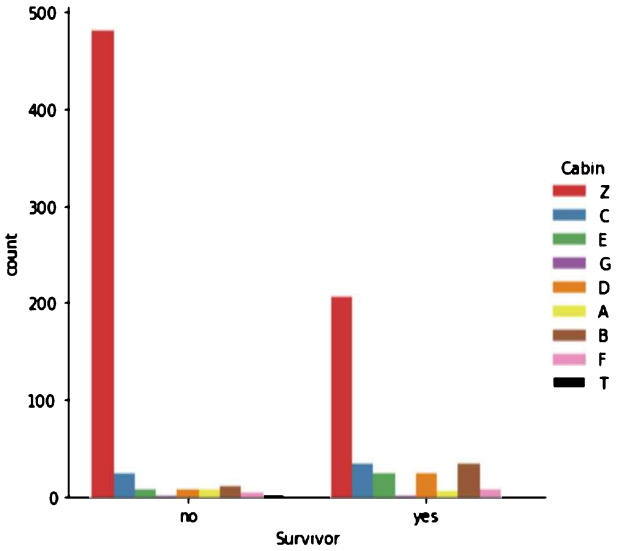
Pada atribut "Fare" terlihat bahwa harga tiket tidak berpengaruh secara signifikan terhadap keselamatan penumpang dan atribut "Pclass". Grafik distribusi atribut "Fare" terdapat pada gambar 10 berikut ini.



Gambar 10 Distribusi atribut "Fare"

7) Cabin

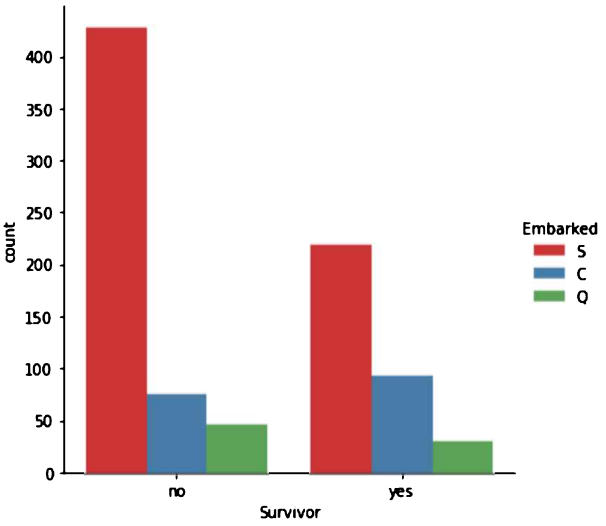
Pada atribut "Cabin" terlihat bahwa Z memiliki penumpang selamat dan tidak selamat tertinggi. Namun, hal tersebut tidak bisa dipastikan karena Z merupakan nilai hilang *(missing value)* yang diberi nilai. Grafik distribusi atribut "Cabin" terdapat pada gambar 11 berikut ini.



Gambar 11. Distribusi atribut "Cabin

8) Embarked

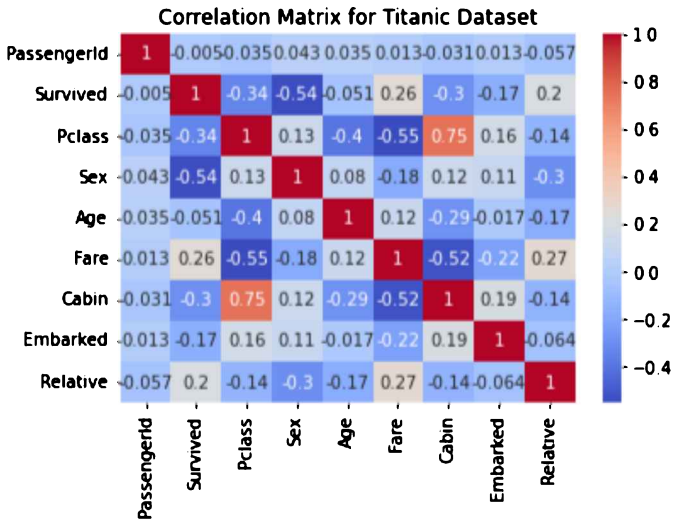
Pada atribut "Embarked" terlihat bahwa penumpang dari Southampton (S) memiliki penumpang selamat dan tidak selamat tertinggi yaitu 219 dan 427. Queenstown (Q) memiliki penumpang selamat dan tidak selamat sebanyak 30 dan 47. Sedangkan dari Cherbourg (C) memiliki penumpang selamat dan tidak selamat sebanyak 93 dan 75. Grafik distribusi atribut "Embarked" terdapat pada gambar 12 berikut ini.



Gambar 12. Distribusi atribut "Embarked"

9) Korelasi antar data

Dalam klasifikasi, korelasi adalah salah satu teknik untuk mencari relasi antar dua variable yang bisa menjadi kunci untuk meningkatkan akurasi model. Grafik hasil perhitungan korelasi data terdapat pada gambar 13.



Gambar 13. Korelasi antar data

B. *Hasil Eksperimen*

Semua algoritma melakukan pembelajaran untuk melakukan klasifikasi untuk mengetahui selamat atau tidak selamatnya penumpang di kapal titanic. Sebelum itu kami menerapkan resampling data untuk data label seperti yang terdapat pada gambar 14 dan 15 berikut ini.

```
label_value = df_upsampled["Survived"].value_counts().head(50).to_frame()
label_value.index
label_value
```

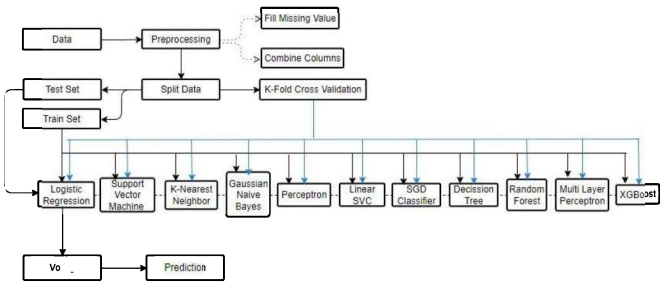
	Survived
1	549
0	549

Gambar 14. Hasil resampling data

```
label_value = undersampled_data["Survived"].value_counts().head(50).to_frame()
label_value.index
label_value
```

	Survived
0	485
1	485

Gambar 15. Hasil resampling data



Gambar 16. Ilustrasi alur eksperimen

Gambar 16 merupakan alur eksperimen dalam penelitian ini. Terdapat beberapa proses utama, pertama data yang telah dikumpulkan dilakukan preprocessing, kemudian kedua dilakukan split data. Ketiga data training akan dilakukan pembelajaran menggunakan beberapa algoritma. Keempat akan dilakukan voting untuk memperoleh hasil klasifikasi terbaik. Terakhir melakukan prediksi pada data baru.

Saat menerapkan algoritma, untuk meningkatkan akurasi, dilakukan beberapa penyesuaian paramater. LR yang diterapkan diatur dengan parameter optimasi masalah "lbfgs". SVM diatur dengan kernel "rbf" dan regulasi parameter 1.0. kNN jumlah k yang dipilih yaitu 5. *Naïve Bayes* menggunakan parameter varians smooting yang harus negative yaitu 1e-9. Linear SVC dengan regulasi parameter 1.0. SGD *classifier* menggunakan parameter *loss function* "hinge" dan kriteria berhernti 1e-3. *Decision tree* memiliki parameter maksimal kedalaman 5. MLP classifier dengan parameter hidden layer (100,50) dan maksimum iterasi 1000. Algoritma dievaluasi menggunakan akurasi *10-fold cross-validation.* Akurasi tiap model dapat diliat pada tabel 1 berikut ini.

TABEL 1. PERBADINGAN AKURASI STRATIFIELD SAMPLING DAN MAJORITY MODEL 1

Stratifield Sampling+Majority	Training Accuracy (%)	Test Accuracy (%)
GaussianNB	89,20	84,27
XGBClassifier	91,44	83,71
Decision Tree	87,10	80,90
Support Vector Machine	79,80	79,78
Random Forest	75,46	71,35
Linear SVC	68,02	65,73
Stochastic Gradient Decent	63,96	65,17
Logistic Regression	70,27	63,48
MLPClassifier	76,16	63,48
K-Nearest Neighbor	69,00	62,36
Perceptron	62,83	61,80

Diketahui hasil akurasi terbaik terdapat pada model pertama, yaitu menggunakan algoritma Gaussian Naïve Bayes, dengan nilai akurasi testing 84,27% teknik Resampling dengan Stratifield Sampling dan Majoritty Oversampling.

TABEL 2. PERBADINGAN AKURASI STRATIFIELD SAMPLING DAN SMOTE MODEL 2

Stratifield Sampling + SMOTE	Training Accuracy (%)	Test Accuracy (%)
GaussianNB	90,34	84,27
XGBClassifier	92,61	83,15
Decision Tree	87,73	81,46
Support Vector Machine	81,48	80.34
Random Forest	78,52	75,28
K-Nearest Neighbor	64,32	65,17
Linear SVC	60,23	62,92
MLPClassifier	75,00	62,92

Stratifield Sampling + SMOTE	Training Accuracy (%)	Test Accuracy (%)
Logistic Regression	71,82	60,67
Stochastic Gradient Decent	52,95	42,70
Perceptron	49,77	38,20

Dari tabel 2 dapat diketahui hasil akurasi terbaik pada model 2, tetap sama menggunakan algoritma Gaussian Naïve Bayes, dengan nilai akurasi testing 84,27% teknik Resampling dengan Stratifield Sampling dan SMOTE.

TABEL 3. PERBANDINGAN AKURASI RANDOM SAMPLING DAN MAJORITY MODEL 3

Random Sampling+Majorrity	Training Accuracy (%)	Test Accuracy (%)
GaussianNB	87,03	93,26
XGBClassifier	89,53	89,89
Decision Tree	85,29	88,76
Support Vector Machine	79,55	87,64
Random Forest	75,19	78,65
Stochastic Gradient Decent	62,34	76,40
Linear SVC	62,97	76,40
K-Nearest Neighbor	67,21	75,28
Perceptron	61,72	74,16
MLPClassifier	69,95	71,91
Logistic Regression	70,07	70,79

Dari tabel 3 dapat diketahui hasil akurasi terbaik pada model 3, menggunakan algoritma Gaussian Naïve Bayes, dengan nilai akurasi testing 93,26% teknik Resampling dengan Random Sampling dan Majoritty Oversampling.

TABEL 4. PERBADINGAN AKURASI RANDOM SAMPLING DAN SMOTE MODEL 4

Random Sampling + SMOTE	Training Accuracy (%)	Test Accuracy (%)
XGBClassifier	91.03	89.89
GaussianNB	88.76	88.76
Decision Tree	87.11	85.39
Support Vector Machine	80.72	84.27
Random Forest	77.94	80.90
Stochastic Gradient Decent	58,08	77,53
Linear SVC	56,08	77,53
K-Nearest Neighbor	64,64	71,91
Perceptron	52,37	71,91
MLPClassifier	77,32	69,66
Random Sampling + SMOTE	Training Accuracy (%)	Test Accuracy (%)
Logistic Regression	72,27	64,04

Dari tabel 4 dapat diketahui hasil akurasi terbaik pada model 4, menggunakan algoritma XGB Classifier, dengan nilai akurasi testing 89,89% teknik Resampling dengan Random Sampling dan SMOTE.

TABEL 5. PERBADINGAN AKURASI STRATIFIELD SAMPLING DAN MAJORITY MODEL

Voting (11 Algoritma)	Train	Test
Model 1	86,96	81,46
Model 2	88,07	80,89
Model 3	81,92	84,26
Model 4	85,56	84,26

Dari tabel 5, penelitian ini mendapatkan jawaban bahwa metode terbaik yang digunakan adalah menggunakan penerapan teknik resampling dengan Random Sampling dan SMOTE sampling.

IV. KESIMPULAN DAN SARAN

A. *Kesimpulan*

Pada penelitian ini dilakukan klasifikasi untuk mengetahui penumpang Titanic selamat atau tidak pada kolom Survived. Pertama, dilakukan *pre-processing* terhadap data hilang dan melakukan investigasi terhadap fitur atau kolom yang akan digunakan, selanjutnya dilakukan *modelling* dengan sebelas algoritma, yaitu . Kemudian algoritma tersebut digabungkan untuk menjadi satu kesatuan model, serta penelitian tersebut, mendapatkan jawaban jika metode terbaik yang digunakan adalah menggunakan penerapan teknik resampling dengan Random Sampling dan SMOTE sampling dengan hasil akurasi 93,26% dengan nilai resampling SMOTE label diseimbangkan menurun mejadi 485 perclassnya.

B. *Saran*

Untuk pengembangan selanjutnya, penulis dapat mempertimbangkan beberapa langkah agar paper menjadi lebih komprehensif dan memberikan kontribusi yang lebih signifikan kepada pembaca. Beberapa saran yang dapat dipertimbangkan adalah:

1. **Pemaparan Metode Ensemble Lainnya:** Merinci informasi mengenai pendekatan ensemble learning lainnya, seperti stacking, bagging, atau boosting. Melakukan perbandingan antara ensemble learning voting yang telah diulas dengan metode-metode alternatif akan memberikan pemahaman yang lebih mendalam.
2. **Optimasi Parameter:** Mendiskusikan aspek optimasi parameter pada metode ensemble learning. Penulis dapat menjelaskan cara mencari parameter optimal untuk meningkatkan kinerja model, terutama dalam konteks stacking yang melibatkan beberapa model dasar.
3. **Evaluasi Performa yang Mendalam:** Melakukan evaluasi performa yang lebih komprehensif dengan

mempertimbangkan berbagai metrik evaluasi, seperti precision, recall, F1-score, dan area under the ROC curve (AUC-ROC). Pendekatan ini akan memberikan wawasan yang lebih holistik tentang dampak ensemble learning terhadap performa model.

Dengan menjelajahi aspek-aspek tersebut, penulis selanjutnya dapat meningkatkan kedalaman dan kelengkapan paper, menjadikannya sumber referensi yang lebih berbobot dan relevan bagi pembaca yang memiliki minat dalam dunia ensemble learning dan machine learning secara keseluruhan.

REFERENSI

[1] Y. Yennimar, R. E. Manihuruk, and E. L. Br Hotang, "Prediction Models with Machine Learning Against Student Success in Online Learning," *SinkrOn*, vol. 6, no. 1, pp. 62–68, 2021, doi: 10.33395/sinkron.v6i1.11095.

[2] K. T. Informatika and U. Hasanuddin, "Machine Learning Pengenalan Citra Digital Berbasis Software As A Service (Saas) The Machine Learning Of Digital Image Recognition Base On Software As A Service (Saas) Andi Lukman," 2013.

[3] T. V Ramachandra, T. Mondal, and B. Setturu, "Relative performance evaluation of machine learning algorithms for land use classification using multispectral moderate resolution data," *SN Appl. Sci.*, 2023, doi: 10.1007/s42452-023-05496-4.

[4] E. M. Detection, R. K. Shahzad, and N. Lavesson, "Comparative Analysis of Voting Schemes for," no. August, pp. 98–117, 2012.

[5] S. T. U. D. Dwivedi, "A comparative study of heterogeneous ensemble methods for the identification of geological lithofacies," *J. Pet. Explor. Prod. Technol.*, vol. 10, no. 5, pp. 1849–1868, 2020, doi: 10.1007/s13202-020-00839-y.

[6] R. S. Khairy, "The Detection of Counterfeit Banknotes Using Ensemble Learning Techniques of AdaBoost and Voting," vol. 14, no. 1, pp. 326–339, 2021, doi: 10.22266/ijies2021.0228.31.

[7] E. Ekinci, S. İ. Omurca, and N. Acun, "A Comparative Study on Machine Learning Techniques using Titanic Dataset".

[8] Fitri Handayani and Feddy Setio Pribadi, "Implementasi Algoritma Naive Bayes Classifierdalam Pengklasifikasian Teks OtomatisPengaduan dan Pelaporan Masyarakat melaluiLayanan Call Center 110," *J. Tek. Elektro*, vol. 7, no. 1, pp. 1–6, 2015.

[9] Will Cukierski, "Titanic - Machine Learning from Disaster," *Kaggle*, 2012.

Analisis Segmentasi Kepribadian Pelanggan Menggunakan K Medoids dan Random Forest untuk Menentukan Strategi Pemasaran

Rendy Wenda Dwi Kurniawan
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
rendy.5200411461@student.uty.ac.id

Nazar Iqbal Bimantoro
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
naza.5200411473@student.uty.ac.id

Muhammad Latif Ma'ruf
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
muhammad.5200411496@student.uty.ac.id

Panji Rangga Adzan Fajar Fakharudin
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
panji.5200411475@student.uty.ac.id

Febiansyah Annaufal Ahnaf Fauzi
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
febriansyah.5200411468@student.uty.ac.id

Muhammad Irsyad Indra Fata
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
muhammad.5200411492@student.uty.ac.id

Abstrak— Strategi pemasaran adalah suatu rencana yang melibatkan pemikiran dan perencanaan yang matang untuk mencapai tujuan pemasaran suatu perusahaan atau organisasi. Kepribadian Pelanggan merupakan pemahaman tentang karakteristik, preferensi, dan perilaku unik dari pelanggan atau konsumen. Ini mencakup faktor seperti kepribadian, nilai, gaya hidup, preferensi produk, dan interaksi pelanggan dengan merek atau produk tertentu. Memahami customer personality adalah kunci untuk menyusun strategi pemasaran yang lebih efektif, personal, dan sesuai dengan kebutuhan pelanggan, sehingga memungkinkan perusahaan untuk memenuhi harapan pelanggan dengan lebih baik dan membangun hubungan yang lebih erat. Algoritma K-Medoids membantu dalam mengatasi data yang mungkin memiliki outlier atau noise yang dapat mempengaruhi hasil pengelompokan. Random Forest adalah sebuah algoritma machine learning yang dapat digunakan untuk klasifikasi. Algoritma ini memanfaatkan ensemble learning dengan menggabungkan sejumlah pohon keputusan (decision trees) untuk meningkatkan akurasi prediksi. Hasil eksperimen dan analisis, diperoleh kesimpulan pada pengklusteringan menggunakan k-medoids berhasil menghasilkan 4 kluster dengan jumlah data 352 untuk "Need Attention", 965 "At Risk", 474 "Potential Loyal", dan 444 data masuk ke dalam kategori "Loyal". Hasil dari klasifikasi menggunakan Random Forest pada data yang telah tersegmentasi mendapatkan akurasi sebesar 100% dari 1788 data untuk proses *training* dan *testing* 99,3% dari 447 data. Dari hasil eksperiman dan analisis ini dapat dipahami terkait karakteristik dari suatu proses pembelian serta diharapkan mampu membantu proses pengambilan keputusan, yaitu mengenai strategi pemasaran yang lebih optimal dan sesuai dengan personalitas pelanggan.

Kata Kunci—K-Medoids, Random Forest, Segmentasi Pelanggan

I. PENDAHULUAN

Strategi pemasaran adalah suatu rencana yang melibatkan pemikiran dan perencanaan yang matang untuk mencapai tujuan pemasaran suatu perusahaan atau organisasi. Ini mencakup berbagai aspek, mulai dari menentukan target pasar yang tepat, mengidentifikasi kebutuhan pelanggan, hingga pengembangan produk, penetapan harga yang sesuai, promosi, dan distribusi. Dengan merinci langkah-langkah seperti penetapan tujuan, analisis pasar, segmentasi pasar, pengembangan produk, harga, promosi, dan distribusi, strategi pemasaran memberikan panduan untuk mengarahkan upaya pemasaran dengan tujuan mencapai hasil yang diinginkan. Dengan demikian, strategi pemasaran merupakan alat penting dalam mencapai kesuksesan bisnis dengan cara mengoptimalkan pengeluaran dan sumber daya yang tersedia untuk mencapai tujuan pemasaran yang ditetapkan. Perusahaan harus mampu mengutamakan stategi pemasaran berorientasi pelanggan. Ini berguna untuk mengelola hubungan yang baik dengan pelanggan agar kepuasan pelanggan tercapai dan mendapatkan loyalitas dari pelanggan [1]. Kepribadian Pelanggan merupakan pemahaman tentang karakteristik, preferensi, dan perilaku unik dari pelanggan atau konsumen. Ini mencakup faktor seperti kepribadian, nilai, gaya hidup, preferensi produk, dan interaksi pelanggan dengan merek atau produk tertentu. Memahami customer personality adalah kunci untuk menyusun strategi pemasaran yang lebih efektif, personal, dan sesuai dengan kebutuhan pelanggan, sehingga memungkinkan perusahaan untuk memenuhi harapan pelanggan dengan lebih baik dan membangun hubungan yang lebih erat. Penerapan strategi pemasaran yang mempertimbangkan kepribadian pelanggan menghadapi beberapa permasalahan yang perlu diperhatikan. Pertama, keanekaragaman kepribadian pelanggan menciptakan tantangan dalam mengidentifikasi dan memahami preferensi yang unik. Perubahan dalam kepribadian pelanggan juga perlu diperhitungkan, sehingga strategi pemasaran harus dinamis. Selain itu, terdapat risiko kesalahan dalam memprofilkan kepribadian pelanggan yang bisa mengarah pada alokasi sumber daya yang tidak efisien. Privasi dan etika juga menjadi perhatian penting dalam pengumpulan data kepribadian. Personalisasi pesan pemasaran harus seimbang, dan perusahaan perlu menginvestasikan dalam teknologi dan analitik yang canggih. Terakhir, segmentasi pelanggan yang lebih halus membutuhkan waktu dan upaya ekstra. Namun, dengan pemahaman yang mendalam tentang pelanggan dan pemecahan masalah ini, strategi pemasaran yang mempertimbangkan kepribadian pelanggan dapat menjadi cara yang efektif untuk meningkatkan keterlibatan dan membangun hubungan yang lebih kuat. Dalam penyusunan strategi pemasaran yang berorientasi pelanggan, segmentasi

pelanggan dapat dimanfaatkan untuk mengidentifikasi karakteristik pelanggan ataupun personaliti pelanggan dan mengelompokkan pelanggan berdasarkan perilaku atau kebiasaan yang dilakukan pelanggan [2].

Data transaksi memiliki potensi untuk digunakan dalam menentukan strategi pemasaran, salah satunya segmentasi pelanggan. Segmentasi pelanggan perlu dilakukan untuk mengelompokkan pelanggan yang memiliki kesamaan karakteristik. Pengolahan data transaksi penjualan dengan komputasi tepat dapat memberikan manfaat bagi perusahaan agar pemasaran lebih efektif dan efisien [3]. Dengan memanfaatkan *data mining*, data dapat diolah untuk mengetahui pola dan dapat diidentifikasi informasi yang tersembunyi dalam data tersebut. Dengan pemahaman yang lebih baik tentang preferensi dan perilaku pelanggan dalam setiap segmen, perusahaan dapat merancang strategi pemasaran yang lebih sesuai dan personal, yang dapat meningkatkan retensi pelanggan, memaksimalkan penjualan, dan memperbaiki kinerja bisnis secara keseluruhan. Segmentasi berdasarkan RFM (*Recency, Frequency, Monetary)* adalah alat yang bermanfaat untuk meningkatkan efisiensi upaya pemasaran dan membantu perusahaan fokus pada segmen pelanggan yang paling berharga [4].

Teknik *clustering* dapat digunakan dalam menentukan segmentasi pelanggan dengan menganalisis kelompok data untuk mengetahui karakteristik dari kelompok pelanggan yang terbentuk. Kemudian, dilakukan proses klasifikasi data yang telah memiliki label dari hasil *clustering.* K-Medoids adalah salah satu algoritma pengelompokan (*clustering*) dalam analisis data yang digunakan untuk mengidentifikasi kelompok data yang serupa berdasarkan jarak atau kesamaan antara data. Bedanya dengan algoritma K-Means adalah bahwa K-Medoids menggunakan titik-titik data aktual (medoids) sebagai pusat kelompok, sedangkan K-Means menggunakan nilai rata-rata. Medoids adalah titik dalam kelompok yang memiliki jarak rata-rata yang lebih kecil ke anggota kelompok lainnya. Algoritma K-Medoids membantu dalam mengatasi data yang mungkin memiliki outlier atau noise yang dapat mempengaruhi hasil pengelompokan [5]. Random Forest adalah sebuah algoritma machine learning yang dapat digunakan untuk klasifikasi. Algoritma ini memanfaatkan ensemble learning dengan menggabungkan sejumlah pohon keputusan (decision trees) untuk meningkatkan akurasi prediksi. Setiap pohon dalam Random Forest dipelatih secara independen dan menghasilkan prediksi, kemudian hasil dari berbagai pohon digabungkan untuk menghasilkan prediksi akhir. Random Forest dikenal dengan kemampuannya dalam mengatasi overfitting dan memproses data yang besar dengan baik, menjadikannya salah satu algoritma populer dalam machine learning [6]. Pada penelitian ini, dilakukan penerapan algoritma K-Medoids dan Random Forest pada dataset untuk menentukan segmentasi kepribadian pelanggan. Tujuan dari penelitian ini adalah untuk mengetahui tipe dan karakteristik pelanggan untuk memberikan rekomendasi strategi pemasaran untuk perusahaan .

II. KAJIAN PUSTAKA

2.1 Segmentatsi Pelanggan

K. Kumar, *et al.* [7], menyatakan bahwa segmentasi pelanggan merupakan metode partisi atau pengelompokan pelanggan berdasarkan berbagai karakteristik. Pelanggan disegmentasi berdasarkan karakteristik mereka yang berbeda seperti usia, jenis kelamin, pendapatan tahunan. Pelanggan dapat disegmentasi secara manual dan dengan bantuan alat digital yang lebih efisien berdasarkan dengan karakteristik data baik diberi label maupun tidak dibandingkan dengan proses manual karena jumlahnya yang sangat besar. Dalam penelitian ini dihasilkan suatu kesimpulan yaitu model yang dirancang menggunakan k-means berhasil menerapkan segmentasi pelanggan berdasarkan berbagai parameter yang digunakan, selain itu proses ini melalui 2 tahap yaitu analisis data sebagai gambaran keseluruhan pemrosesan data pelanggan, dan representasi algoritma clustering untuk detailnya informasi.

R. W. Br. S. Berahmana, *et al.* [8], menyatakan bahwa Segmentasi pelanggan adalah proses membagi pelanggan menjadi beberapa kelompok berdasarkan data masa lalu dengan tuntutan, karakteristik, dan fungsi yang sama. Analisis segmentasi pelanggan terhadap data transaksi perusahaan dilakukan untuk mencari pelanggan yang menguntungkan. Pada penelitian ini menggunakan Customer Relationship Management (CRM) sebagai pendukung segmentasi pelanggan dan k-means sebagai pembuatan modelnya dengan klustering. Selain itu data data yang digunakan dikelompokan keadalam nilai, frekeuensi, dan moneter untuk menganalisis perilaku pelanggan seperti seberapa baru pelanggan membeli (Saat Ini), seberapa sering pelanggan membeli (Frekuensi), dan berapa banyak uang yang dikeluarkan pelanggan dalam melakukan transaksi (Moneter). Data yang digunakan data transaksi tahunan mulai bulan januari 2023 sampai dengan desember 2018 dengan total data 334.641. Hasil penelitian menunjukkan bahwa K-Means mempunyai tingkat validitas yang paling baik dibandingkan K-Medoids dan DBSCAN, dimana Hasil Indeks Davies-Bouldin sebesar 0,33009058, dan hasil Indeks Silhouette sebesar 0,912671056. Berdasarkan pengujian sejumlah cluster berbeda yang diuji menggunakan Indeks Davies Bouldin dan Silhouette Index, jumlah cluster terbaik adalah 2 cluster.

2.2 *K-Medoids*

A. A.D Sulistyawati, *et al.* [9], menyatakan bahwa k-medoids merupakan algoritma teknik *clustering* yang digunakan untuk mengelompokkan objek kedalam *cluster* dengan objek yang serupa atau sama. Algoritma ini mempunyai kelebihan yaitu tidak sensitif terhadap outlier, dapat mengurangai *noise*, serta k-medoids lebih unggul dalam melakukan klasterisasi dataset heterogen/campuran, pemilihan *cluster*, kompleksitas antar ruang *cluster*, dan waktu eksekusi dibandingkan menggunkan k-means. Dalam penelitiannya menyajikan bahwa hasil segmentasi pelanggan dengan menerapkan algoritma k-medoids menghasilkan *cluster* optimal sebanyak 3 pada dataset Perum BULOG, 3 segmen yang optimal dalam penelitian ini yaitu *Lost Customer* (pelanggan yang pernah melakukan transaksi namun tidak melakukan transaksi pada 2 bulan terakhir dan sebagai pelanggan musiman) , *Core Customer* (pelanggan yang sering melakukan transaksi dengan jumlah nominal pembelian terbesar), dan *New Customer* (pelanggan yang baru melakukan transaksi).

C. Oktarina, *et al.* [10], menyajikan perbandingan *clustering* data twitter menggunakan k-means dan k-medoids. Dalam penelitian tersebut peneliti menggunakan data twitter yang di ambil dari tanggal 26 februari 2019 (dengan jumlah tweet sebanyak 2.900), 8 maret 2019 (dengan jumlah tweet sebanyak 15.600), dan 17 maret 2019 (dengan jumlah tweet

sebanyak 2.900). Hasil yang ditunjukkan pada penelitian ini adalah angka optimal dari *cluster* yang ditunjukkan dari k-means adalah 6 *cluster* dan k-medoids adalah 7 *cluster.*

2.3 *Random Forest*

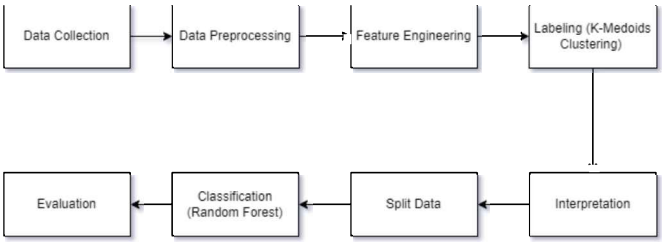
Th. D. Wismarini, *et al.* [11], menyatakan bahwa *Random Forest* merupakan salah satu teknik pembelajaran mesin dapat diandalkan. Klasifikasi, regresi, dan tugas-tugas lainnya dapat dilakukan dengan menggunakan algoritma pembelajaran mesin ini. Untuk menghasilkan prediksi yang lebih akurat dan stabil, algoritma ini menggunakan beberapa model pembelajaran mesin, terutama pohon keputusan. Teknik *Random Forest* merupakan pengembangan dari pendekatan *Decision Tree*. Setiap *Decision Tree* dilatih pada sampel yang berbeda, dan setiap atribut ditugaskan ke pohon yang dipilih secara acak dari subset atribut. Salah satu manfaat dari pemodelan *Random Forest* adalah akurasinya yang tinggi. Hal ini dapat didapatkan, karena dalam memprediksi hasil akhir *Random Forest* dibuat dengan menggabungkan hasil dari beberapa pohon keputusan yang dibangun secara acak.

III. METODOLOGI

3.1 Metodologi

Strategi pemasaran merupakan elemen penting dalam sebuah segmentasi. Beberapa penelitian memanfaatkan strategi pemasaran saat akan memasarkan sebuah produknya. Salah satu metode yang dapat digunakan yaitu Metode K-medoids yaitu teknik dalam analisis klaster atau clustering yang digunakan untuk mengelompokkan data ke dalam beberapa kelompok berdasarkan kemiripan karakteristik tertentu. Dalam konteks pemasaran, penggunaan metode K-medoids dapat membantu dalam segmentasi pasar dan memungkinkan perusahaan untuk mengidentifikasi kelompok pelanggan dengan preferensi atau perilaku serupa [12]. Metodologi disajikan pada Gambar 1. Alur kerja metodologi adalah sebagai berikut:

1. Data collection. Data berasal dari Kaggle yang berjumlah 2.235 data.
2. Pre-processing. Tahap ini merupakan tahap awal dalam mempersiapkan data, tujuan tahap ini adalah untuk membersihkan dan merapikan data mentah supaya dapat digunakan dengan baik pada proses yang selanjutnya.
3. Feature engineering. Melakukan features selection, remove outlier dan visualisasi RFM sebelum melakukan clusteering.
4. Labeling (K-Medoids Clustering). Algoritma K-Medoids yang digunakan bertujuan untuk clustering data yang sebelumnya belum ada label atau cluster di dalamnya.
5. Interpretation. Melakukan interpretasi pada data untuk menemukan jejaring hubungan yang saling berkaitan.
6. Split data. Data yang sudah di proses sebelumnya dibagi menjadi 2 data yaitu data latih dan data uji.
7. Classification (Random Forest). Untuk mengklasifikasikan data, peneliti menggunakan algoritma dari Random Forest karena setelah diuji dengan membandingkan algoritma decision tree hasil terbaik menggunakan random forest.
8. Evaluation. Pada tahap ini merupakan tahap pembuktian hasil klasikasi dengan menampilkan *Precision, Recall, F1-Score*, akurasi.



Gambar 1. Metodologi

3.2 Pelabelan Dataset

Penelitian ini menggunakan pengelompokan data ke dalam kelompok berdasarkan kemiripan karakteristik tertentu guna meminimalkan total jarak antara semua titik data dalam kelompok dengan medoids atau pusat dari kelompok tersebut [13].

Rumus K-Medoids

$$d_{ij} = \sqrt{\sum_{a=1}^{p} (x_{ia} - x_{ja})^2} = \sqrt{(x_i - x_j)'(x_i - x_j)}$$

3.3 Analisis Pola

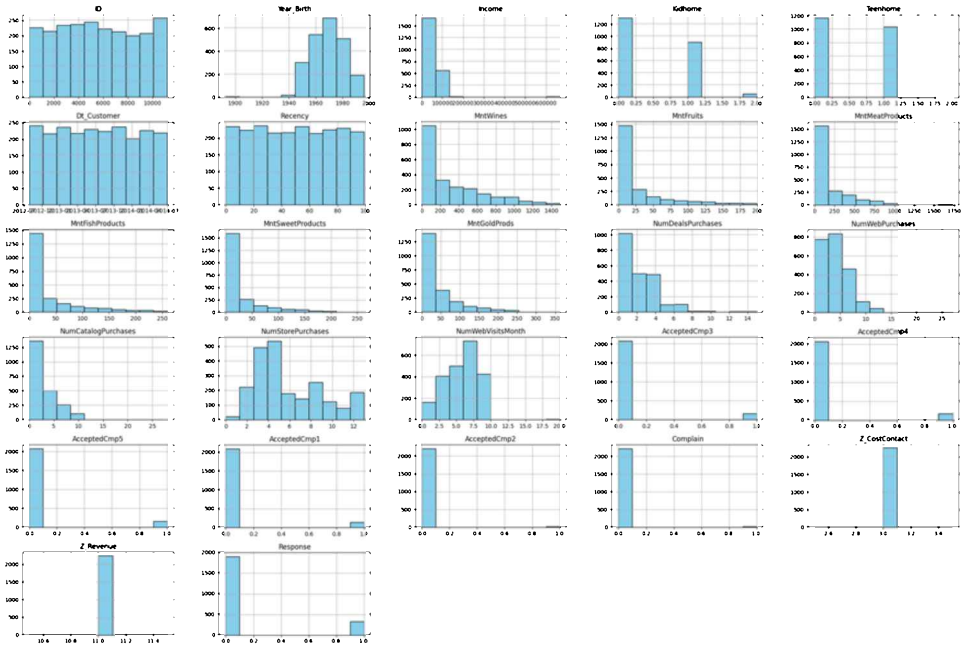
Pendekatan pemasaran yang berfokus pada segmentasi pelanggan berdasarkan karakteristik kepribadian dan perilaku mereka. Melalui analisis pola ini, perusahaan dapat mengidentifikasi dan memahami preferensi, nilai, dan kebutuhan unik dari setiap kelompok pelanggan. Misalnya, kelompok yang cenderung ekstrovert dan berorientasi pada status mungkin lebih tertarik pada promosi yang menekankan eksklusivitas dan prestise. Di sisi lain, kelompok yang lebih introvert dan berfokus pada nilai mungkin lebih menanggapi strategi pemasaran yang menekankan keuntungan praktis dan penawaran khusus. Dengan memahami dan menyesuaikan strategi pemasaran sesuai dengan kepribadian pelanggan, perusahaan dapat meningkatkan relevansi dan efektivitas kampanye mereka, mengarah pada retensi pelanggan yang lebih baik dan peningkatan hasil bisnis secara keseluruhan.

IV. HASIL DAN PEMBAHASAN

Pada bab eksperimen dan analisis ini akan dipaparkan serangkaian proses sebagai lanjutan dari tahapan metodologis terkait segmentasi yang didasarkan pada personalitas pelanggan. Eksperimen dilakukan dengan pengolahan data yang menggunakan pendekatan statistik seperti clustering dan classification. Setiap hasil yang didapatkan disajikan secara grafis atau tabel dan dilakukan analisis serta interpretasi untuk memberikan wawasan mendalam terhadap relevansi hasil terhadap hipotesis. Dari hasil eksperiman dan analisis ini diharapkan mampu membantu proses pengambilan keputusan, yaitu mengenai strategi pemasaran yang lebih optimal dan sesuai dengan personalitas pelanggan.

4.1 Data *Preprocessing* dan Eksplorasi

4.1.1 Persebaran Jumlah Data

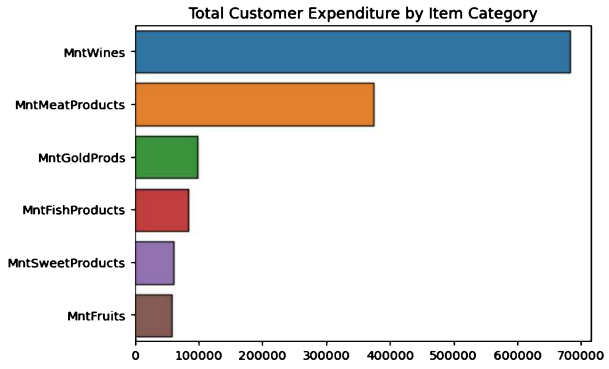


Gambar 2. Persebaran Data

Gambar 2 merupakan grafik dari persebaran seluruh data berdasarkan atribut yang ada pada dataset tersebut. Pada grafik tersebut terdapat 27 atribut yang mencakup aspek profil, penawaran, produk dan mepembelian produk, .

4.1.2 Total Pengeluaran Pelanggan berdasarkan Produk Penjualan

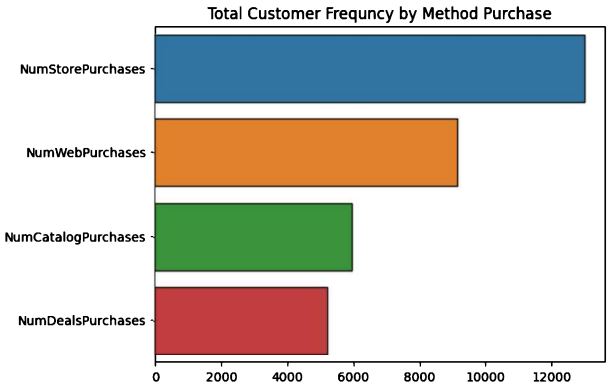
Pada gambar 3 merupakan grafik yang berisi dengan total pengeluaran pelanggan berdasarkan dengan kategori item atau produk penjualan. Produk dengan nama MntWines mempunyai total pengeluaran berkisar 650.000 sampai dengan 700.000. Adapun produk dengan total pengeluaran pelanggan terendah yaitu MntFruits dengan total pengeluaran berkisar antara 40.000 sampai dengan 100.000.



Gambar 3. Total Frekuensi Pengeluaran berdasarkan Produk Penjualan

4.1.3 Total Frekuensi Pelanggan berdasarkan Metode Pembelian

Pada gambar 4 merupakan sebuah grafik yang berisi dengan total frekuensi belanja yang dilakukan oleh pelanggan berdasarkan dengan metode pembelian yang digunakan. Total frekuensi belanja pelanggan tertinggi yaitu menggunakan metode NumStorePurchase dengan total frekuensi lebih dari 12.000 total belanja. Adapun frekuensi belanja pelanggan terendah terdapat pada penggunaan metode NumDealsPurchases dengan total frekuensi berkisar antara 4.000 sampai dengan 6.000 total belanja.

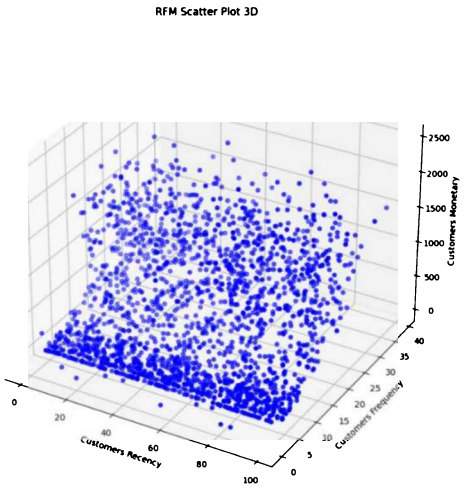


Gambar 4. Total Frekuensi berdasarkan Motode Pembelian

4.1.4 Persebaran Data Sebelum Dilakukan Segmentasi

Berdasarkan data yang telah dipaparkan dalam penjelasan maupun hasil visualisasi di atas, dapat dilihat persebaran dari masing masing variabel dalam bentuk scatter plot 3d. Dalam scatter plot 3D, data direpresentasikan sebagai titik-titik dalam ruang tiga dimensi, dengan setiap sumbu mewakili satu

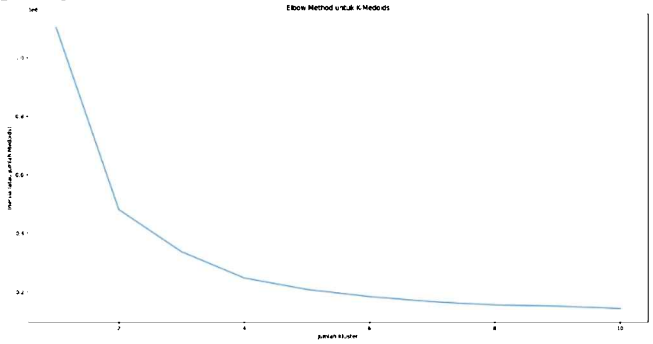
variabel. Hal ini diperlukan untuk analisis data sehingga dapat mengidentifikasi struktur, korelasi, atau pola dari tiga variabel secara langsung, yang dalam penelitian ini yaitu recency, frequency, dan monetary. Untuk hasil visualisasi scatter plot 3d dari data personalitas pelanggan dapat dilihat pada gambar 5.



Gambar 5. Scatter Plot

4.2 Data Labelling dan *Interpretation*

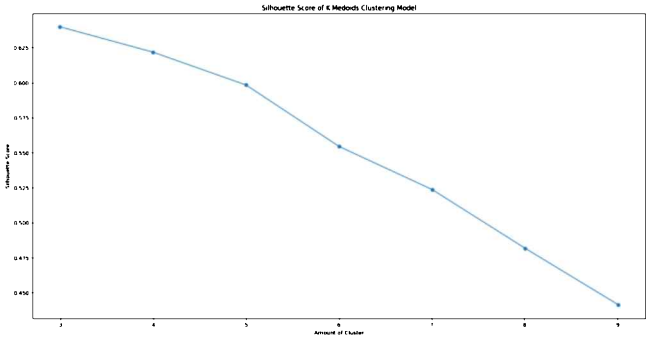
Implementasi Elbow Method dengan K-Medoids, Dalam proses *labeling* menggunakan K-medoids sebagai algoritma clustering, hal yang terlebih dahulu dilakukan adalah penggunaan elbow method untuk membantu menentukan jumlah cluster yang optimal. Pada pengkombinasian elbow method dengan k-medoids, jumlah cluster yang optimal ditandai dengan adanya siku yang terbentuk karena perbedaan inertia yang signifikan. Jumlah cluster yang optimal sebagai hasil dari pengkombinasian elbow method dengan k-means adalah 2, 3, dan 4. Untuk mengetahui jumlah cluster yang terbaik, perlu dilakukan evaluasi model lebih lanjut yang akan dibahas pada bagian berikutnya. Hasil visualisasi elbow method yang dikombinasikan dengan k-medoids dapat dilihat pada gambar 6.



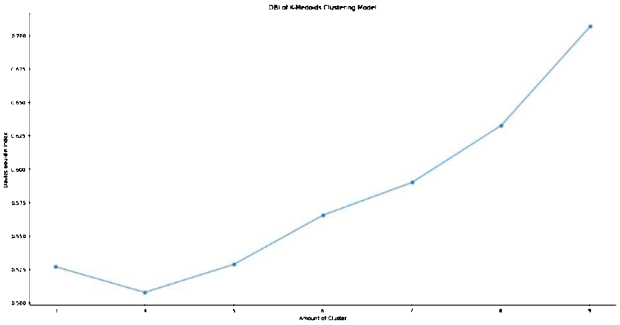
Gambar 6. Elbow Method dengan K-Medoids

4.3 Model Evaluasi Model

Proses permodelan menggunakan algoritma k-medoids telah dilakukan dengan jumlah cluster optimal yang diperoleh melalui elbow method. Dari hasil permodelan tersebut, perlu dilakukan pengujian terhadap hasil clustering yang didapatkan. Dalam penelitian ini proses evaluasi model dilakukan menggunakan parameter Silhouette Coefficient dan Davies-Bouldin Index. Dalam penilaiannya, jika nilai silhouette coefficient mendekati 1 maka hasil clustering dengan penggunaan jumlah cluster itu lebih baik. Sedangkan pada Davies-Bouldin Index, hasil clustering yang didapat dianggap lebih baik jika bernilai mendekati 0. Hasil visualisasi dari penerapan silhouette coefficient dan davies bouldin index pada masing masing model dapat dilihat pada gambar 7 dan gambar 8.



Gambar 7. Silhouette Coefficient

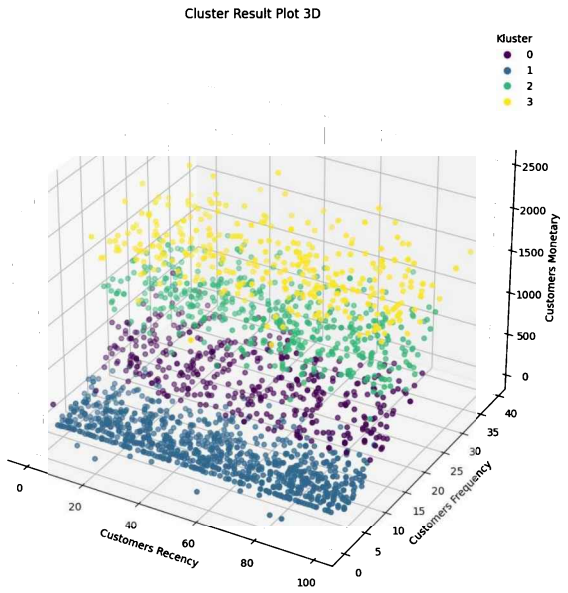


Gambar 8. Davies Bouldin Index

Berdasarkan hasil visualisasi dari silhouette coefficient score dan davies-bouldin index, didapatkan beberapa kesimpulan. Dalam penerapan metode k-medoids yang telah dilakukan evaluasi model menggunakan silhouette coefficient, jumlah cluster yang menghasilkan pengelompokkan data yang paling baik adalah 3 sedangkan 4 menjadi alternatif selanjutnya. Pada evaluasi menggunakan davies-bouldin index, jumlah cluster yang menghasilkan pengelompokkan data yang paling baik pada model dari algoritma k-medoids adalah sama yaitu 4. Hal ini karena dalam penentuan menggunakan davies bouldin index yang terbaik adalah tidak hanya melihat titik terendah, akan tetapi perlu mempertimbangkan siku yang dihasilkan dari model tersebut.

4.4 Hasil Segmentasi

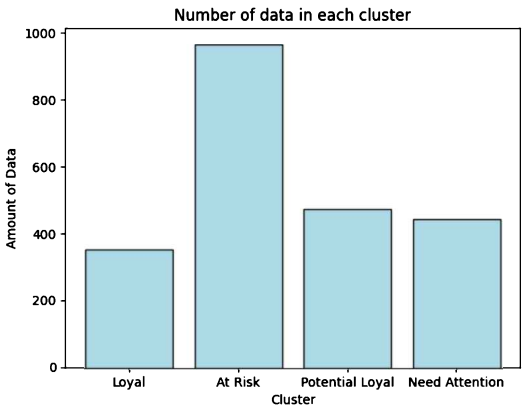
Berdasarkan evaluasi model yang dilakukan menggunakan silhouette coefficient dan juga davies bouldin index terhadap kinerja algoritma k-medoids yang menunjukkan bahwa cluster optimal yang digunakan adalah sejumlah 4. Untuk hasil dari segmentasi menggunakan algoritma k-medoids dengan jumlah cluster 4 dapat dilihat persebaran dari masing masing variabel melalui visualisasi scatter plot 3d yang terdapat pada gambar 9.



Gambar 9. Plotting Hasil Clustering

4.5 Interpretasi

Dari segmentasi pelanggan menggunakan algoritma k-medoids menjadi 4 *cluster,* dilakukan proses interpretasi untuk memudahkan dalam pemanfaatan dan juga pengambilan keputusan terkait strategi pemasaran. *Cluster* "At Risk" terdiri dari individu dengan frekuensi pembelian rendah dan nilai moneter yang menurun, menunjukkan adanya risiko potensial dalam mempertahankan keterlibatan mereka. Kemudian, *cluster* "Need Attention" mencakup pelanggan dengan frekuensi pembelian yang rendah namun nilai moneter yang relatif tinggi, mengindikasikan potensi untuk meningkatkan interaksi pelanggan dalam segmen ini dengan memberikan perhatian khusus pada kebutuhan dan preferensi yang dimiliki. Selanjutnya, *cluster* "Potential Loyal" menunjukkan frekuensi pembelian yang lebih tinggi dan nilai moneter yang memadai, menandakan potensi untuk membangun loyalitas melalui pengalaman yang positif. Sementara itu, *cluster* "Loyal" terdiri dari pelanggan setia dengan frekuensi pembelian dan nilai moneter yang tinggi, mencerminkan hubungan yang solid dan berkelanjutan. Untuk hasil interpretasi dari segmentasi yang dilakukan menggunakan algoritma k-medoids dapat dilihat pada gambar 10.



Gambar 10. Jumlah Data setiap Cluster

Berikut merupakan hasil dari proses *Clustering* pada tabel 1:

TABEL 1. HASIL CLUSTERING DAN INTERPRETASI

Cluster	Jumlah Data	Karakteristik
1	352	Need Attention
2	965	At Risk
3	474	Potential Loyal
4	444	Loyal

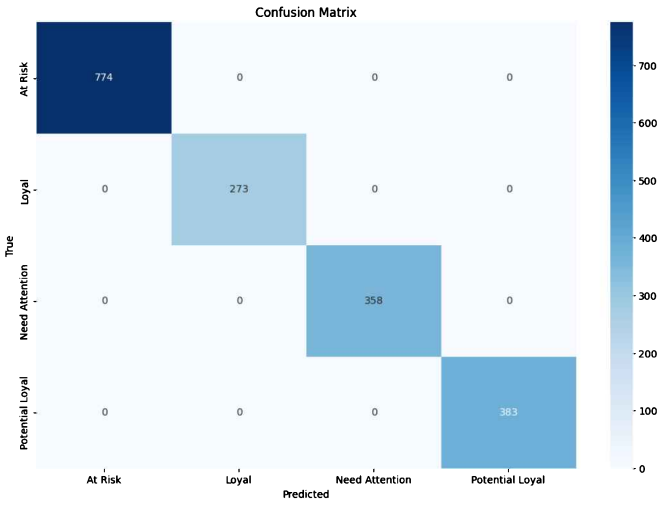
Dari gambar 9 dan tabel 1 diperoleh bahwa pada cluster need attention terdapat 352 jumlah data. Pada cluster at risk terdapat 965 jumlah data. Hal ini yang menjadikan cluster risk menjadi cluster dengan jumlah data terbanyak daripada cluster yang lain. Kemudian pada cluster pontential loyal dan cluster loyal terdapat 474 jumlah data dan 444 jumlah data.

4.6 Klasifikasi

Berdasarkan hasil dari proses pembuatan label menggunakan algoritma K-Medoids maka akan dilakukan proses selanjutnya, yaitu proses pembagian data yang akan digunakan untuk klasifikasi dengan membagi dataset menjadi 80% data training dan 20% data testing. Dalam proses ini, algoritma *Random Forest* dimanfaatkan sebagai metode klasifikasi. Evaluasi kinerja model dilakukan melalui metrik akurasi, presisi, *recall*, dan *F1-score*, yang memberikan pemahaman mendalam tentang kemampuan model dalam mengklasifikasikan pelanggan berdasarkan pola pembelian dan nilai moneter. Sehingga hasil penelitian ini dapat memberikan dasar untuk membantu penentuan strategi pemasaran yang lebih terfokus dan efektif berdasarkan karakteristik dari setiap kelompok pelanggan yang diidentifikasi. Adapun hasil dari C*onfusionMatrix Training* terdapat pada gambar 11 dan tabel 2.

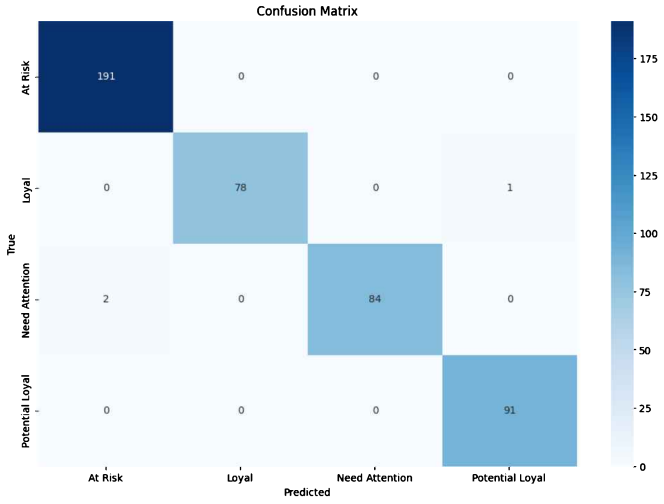
TABEL 2. CLASSIFICATION REPORT TRAINING PROSES

	Presisi	*Recall*	*F1-Score*	*Support*
At Risk	1.00	1.00	1.00	774
Need Attention	1.00	1.00	1.00	358
Potential Loyal	1.00	1.00	1.00	383
Loyal	1.00	1.00	1.00	273



Gambar 11. Confusion Matrix Training

Berdasarkan tabel 2 dan gambar 11 diatas didapatkan hasil training pada data training dengan akurasi sebesar 100% dari 1788 data training.



Gambar 12. Confusion Matrix Testing

TABEL 3. CLASSIFICATION REPORT TESTING PROSES

	Precision	Recall	F1-Score	Support
At Risk	0.99	1.00	0.99	191
Need Attention	1.00	0.98	0.99	79
Potential Loyal	0.99	1.00	0.99	86
Loyal	1.00	0.99	0.99	91

Dari hasil pengujian klasifikasi menggunakan algoritma *Random Forest* yang dilakukan oleh penulis yang terdapat pada gambar 12 dan tabel 3 mendapatkan akurasi sebesar 99,3% dari 447 data testing.

V. KESIMPULAN

Berdasarkan penelitian ini, dapat disimpulkan dari beberapa analisis yang telah dilakukan terhadap data personalitas pelanggan yang digunakan berhasil tersegmentasi dengan kinerja yang baik menggunakan algoritma K-Medoids. Hasil segmentasi menghasilkan 4 kluster dengan jumlah data 352 untuk "Need Attention", 965 "At Risk", 474 "Potential Loyal", dan 444 masuk ke dalam kategori "Loyal". Hasil dari klasifikasi menggunakan Random Forest pada data yang telah tersegmentasi mendapatkan akurasi sebesar 100% dari 1788 data untuk proses *training* dan *testing* 99,3% dari 447 data. Dari hasil eksperiman dan analisis ini dapat dipahami terkait karakteristik dari suatu proses pembelian serta diharapkan mampu membantu proses pengambilan keputusan, yaitu mengenai strategi pemasaran yang lebih optimal dan sesuai dengan personalitas pelanggan.

REFERENSI

[1] W. A. Triyanto, "Algoritma K-Medoids Untuk Penentuan Strategi Pemasaran Produk," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 6, no. 1, p. 183, 2015, doi: 10.24176/simet.v6i1.254.

[2] P. Pln, "Pengklasteran Data Pelanggan Dari Aplikasi Mobile Untuk Penentuan Strategi Pemasaran," vol. 05, 2023.

[3] Y. Suhadi and A. Samsudin, "Optimalisasi Strategi Segmentation, Targeting dan Positioning Dalam Meningkatkan Penjualan pada Pitstop Kopi Gresik," *Al-Kharaj J. Ekon. Keuang. Bisnis Syariah*, vol. 5, no. 6, pp. 2826–2836, 2023, doi: 10.47467/alkharaj.v5i6.3664.

[4] G. Purnama, T. H. Pudjiantoro, and P. N. Sabrina, "Segmentasi Pelanggan Menggunakan K-Medoids Berdasarkan Model Length, Recency, Frequency, Monetary (LRFM)," *SNIA (Seminar Nas. Inform. dan Apl.*, vol. 5, pp. 29–34, 2021, [Online]. Available: https://snia.unjani.ac.id/web/index.php/snia/article/view/240.

[5] A. R. Mulyawan, D. Gunawan, H. Basri, S. Alfarizi, and N. Ichsan, "Penerapan K-Medoids Clustering Dan Silhouette Method Untuk Strategi Pemasaran Program Donasi Pada Lembaga Amil Zakat," *Inf. Syst. Educ. Prof. J. Inf. Syst.*, vol. 8, no. 1, p. 107, 2023, doi: 10.51211/isbi.v8i1.2468.

[6] R. Rindiyani, A. Primadewi, M. Maimunah, and A. H. Purwantini, "Klasifikasi Penjualan berdasarkan Platform pada UMKM Omah Branded Menggunakan Random Forest," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 5, p. 1520, 2022, doi: 10.30865/jurikom.v9i5.4949.

[7] O. Access and K. Kumar, "Applying K-Means Clustering for Customer," no. 07, pp. 3491–3498, 2022.

[8] R. W. Sembiring Brahmana, F. A. Mohammed, and K. Chairuang, "Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 11, no. 1, p. 32, 2020, doi: 10.24843/lkjiti.2020.v11.i01.p04.

[9] A. A. D. Sulistyawati and M. Sadikin, "Penerapan Algoritma K-Medoids Untuk Menentukan Segmentasi Pelanggan," *Sistemasi*, vol. 10, no. 3, p. 516, 2021, doi: 10.32520/stmsi.v10i3.1332.

[10] C. Oktarina, K. A. Notodiputro, and I. Indahwati, "Comparison of K-Means Clustering Method and K-Medoids on Twitter Data," *Indones. J. Stat. Its Appl.*, vol. 4, no. 1, pp. 189–202, 2020, doi: 10.29244/ijsa.v4i1.599.

[11] T. D. Wismarini, H. Murti, and K. Nugroho, "Sales Conversion Optimization Analysis Using the Random Forest Method," vol. 8, no. 4, pp. 2699–2705, 2023.

[12] A. T. Fadilah, "Implementasi Algoritma K-Means Clustering Untuk Targeting ADS Studi Kasus : Data Pelanggan Asuransi Mobil BMW Watson Analytics," pp. 31–41, 2023.

[13] I. S. Afari, "K-Medoids Customer Segmentation Algorithm by Utilizing Customer Relationship Management," *J. Comput. Scine Inf. Technol.*, vol. 9, pp. 89–93, 2023, doi: 10.35134/jcsitech.v9i2.69.

Chatbot Multibahasa *Retrival-Based* dan Rekomendasi *Content-Based* untuk Pelayanan Pelanggan Kedai Kopi dengan Pendekatan Algoritma Word2Vec, LSTM, dan Cosine Similarity

Rizki Aldiansyah
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
rzzkalldi@gmail.com

Enny Itje Sela
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
ennysela@uty.ac.id

Moh. Ali Romli
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
ali.romli@uty.ac.id

Sylvia Jane Annatje Sumarauw
Departemen Matematika
Universitas Negeri Manado
Manado, Indonesia
sylviasumarauw@unima.ac.id

Abstrak—Layanan pelanggan di kedai kopi merupakan faktor penting untuk meningkatkan kepuasan pelanggan. Namun, tantangan muncul ketika ada perbedaan bahasa antara pelanggan dan pegawai kedai kopi, serta kurangnya kesempatan untuk memberikan rekomendasi menu kepada pelanggan. Untuk mengatasi hal itu, diusulkan pengembangan sistem chatbot multibahasa yang mampu memberikan layanan pelanggan dan rekomendasi menu melalui model *Word2Vec* dan LSTM dalam pengembangan chatbot, serta sistem rekomendasi content-based untuk rekomendasi minuman. Data latih yang digunakan berjumlah 1600 berbahasa Indonesia dan bahasa Inggris dengan berbagai konteks pertanyaan pengguna, yang terbagi menjadi 8 jenis pertanyaan. Hasil pengujian menunjukkan akurasi klasifikasi pertanyaan mencapai 98% untuk bahasa Indonesia dan 80% untuk bahasa Inggris, sehingga chatbot ini efektif mengenali dan memahami pertanyaan pelanggan. Diharapkan sistem ini dapat meningkatkan layanan pelanggan dan membantu pegawai dalam memberikan pelayanan yang lebih efisien, meningkatkan pengalaman pelanggan, dan memperkuat citra positif kedai kopi.

Kata Kunci—Layanan Pelanggan, Chatbot Multibahasa, Word2Vec, LSTM, Sistem Rekomendasi Content-Based.

I. PENDAHULUAN

Industri kedai kopi di Jawa Barat mengalami pertumbuhan pesat dalam beberapa tahun terakhir, khususnya antara 2019 hingga 2021. Data dari Open Data Jabar mengungkapkan bahwa jumlah kedai kopi di provinsi ini meningkat sebesar 19,07%, mencapai total 11.510 kedai pada tahun 2021. Bandung, sebagai pusatnya mencatat 1292 kedai. Persaingan di industri ini semakin ketat, terutama dalam aspek harga dan pelayanan [1]. Oleh karena itu, kualitas pelayanan menjadi fokus penting untuk kepuasan pelanggan [2].

Salah satu kedai kopi terdampak adalah M Commpany Coffee. Mereka berupaya mengatasi persaingan dengan menghadapi tantangan memberikan informasi tentang kedai dan rekomendasi minuman sesuai preferensi pelanggan, baik lokal maupun internasional. Penelitian yang dilakukan, mengusulkan pengembangan sistem chatbot yang mampu memberikan informasi dan rekomendasi minuman yang menjadi signature pada kedai tersebut. Chatbot ini juga akan mendukung komunikasi dalam bahasa Indonesia dan bahasa Inggris.

Sebelumnya, penelitian sejenis dilakukan oleh Chandra [3] dengan judul "Perancangan Chatbot Menggunakan *Dialogflow Natural Language Processing* (Studi Kasus: Sistem Pemesanan pada Coffee Shop)". Namun, metode yang digunakan pada penelitian sebelumnya yaitu *Dialogflow* memiliki keterbatasan dalam memahami konteks dan variasi makna percakapan, terutama dalam interaksi yang kompleks. Chatbot yang dihasilkan juga hanya berbahasa Indonesia dan fokus pada informasi menu dan transaksi.

Penelitian ini akan mengembangkan chatbot dengan metode *retrieval-based* yang menggabungkan algoritma *Word2Vec* dan LSTM. Chatbot ini akan mampu memberikan informasi yang lebih luas dan rekomendasi minuman menggunakan pendekatan content-based dengan algoritma *cosine similarity*. Kemampuan ini memungkinkan chatbot untuk mengenali konteks pertanyaan pengguna. Selain itu, chatbot ini akan berkomunikasi dalam dua bahasa, yaitu bahasa Indonesia dan bahasa Inggris. Implementasi chatbot berbasis web di M Company Coffee diharapkan dapat meningkatkan efisiensi dan kualitas layanan pelanggan serta menjaga daya saing di industri yang kompetitif.

II. LANDASAN TEORI

A. *Chatbot Retrieval-Based*

Chatbot adalah program komputer yang dirancang untuk berkomunikasi menggunakan bahasa manusia, menyerupai interaksi antara manusia [4]. Bisa berupa percakapan teks atau suara, merespon pengguna berdasarkan respon yang telah diprogram atau kecerdasan buatan. Interaksi dengan

chatbot mencapai manfaat emosional, rasional, dan psikologis yang serupa dengan interaksi manusia. Perkembangan chatbot telah mengalami evolusi teknologi, terbagi menjadi dua pendekatan, salah satunya adalah pendekatan pembelajaran mandiri (AI) [5]. Pendekatan yang digunakan adalah model Generative (mampu menjawab pertanyaan dari kumpulan jawaban) dan model *Retrieval-Based* (mengambil respon terbaik dari kumpulan respon yang sudah ada). Chatbot dengan model *Retrieval-Based* hanya memberikan respon yang telah diprogram sebelumnya dan tidak menghasilkan respon baru [6]. Model ini cocok untuk percakapan yang sempit, sesuai dengan pernyataan bahwa chatbot lebih sesuai untuk percakapan terstruktur. Model *Retrieval-Based* membantu memberikan jawaban yang relevan dengan pertanyaan terkait.

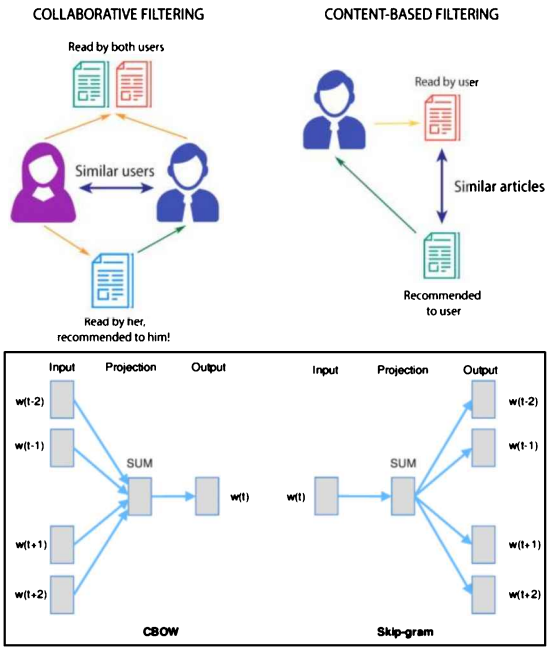
B. *Text Preprocessing*

Proses *preprocessing* data teks merupakan langkah penting dalam analisis data yang bertujuan untuk membersihkan dan mempersiapkan data agar dapat digunakan secara optimal dalam analisis lebih lanjut. Tujuannya adalah menghilangkan noise dan menjaga konsistensi data untuk meningkatkan kualitas analisis [7]. Tahapan proses *preprocessing* meliputi *case folding*, penghapusan angka dan tanda baca, tokenisasi, normalisasi kata, *stemming, stopwords removal, padding, indexing*, dan *word embedding*. Tahapan ini, data menjadi siap untuk dianalisis lebih lanjut dengan berbagai metode seperti klasifikasi teks, analisis sentimen, dan pemrosesan bahasa alami lainnya. Proses ini merupakan langkah awal yang krusial dalam menjalankan analisis data teks dengan akurat dan efisien[8].

C. *Word2Vec*

Word2Vec adalah algoritma *word embedding* berbasis *neural network* yang mengkonversi kata-kata dalam teks menjadi vektor dengan dimensi tertentu, merupakan teknik penting dalam *Natural Language Processing* (NLP) [9]. Model Word2Vec memiliki dua arsitektur utama, yaitu *Continuous Bag of Words* (CBOW) dan Skip-gram. Pada CBOW, model memprediksi kata yang menjadi target dari konteksnya, sementara Skip-gram menggunakan kata yang menjadi target untuk memprediksi konteksnya [10]. Skip-gram lebih efektif dalam menangkap informasi kata-kata langka.

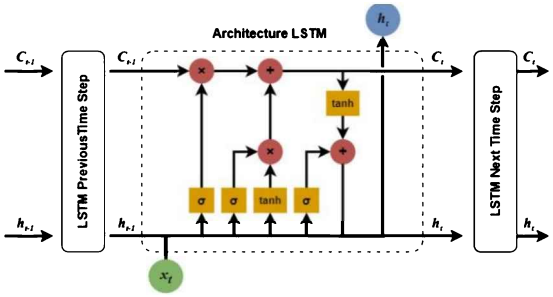
Untuk efisiensi pelatihan pada data besar, terutama Skip-gram yang kompleks, digunakan metode *Negative Sampling* dengan fungsi sigmoid untuk mengurangi kompleksitas perhitungan [11]. Parameter dalam metode ini mengontrol jumlah kata-kata negatif yang dievaluasi, mempengaruhi *trade-off* antara efisiensi pelatihan dan akurasi. Berikut arsitektur *Word2Vec* & Skip-gram ditunjukkan pada Gambar 1.



Gambar 1. Arsitektur *Word2Vec* CBOW & Skip-gram

D. LSTM

Long-Short Term Memory (LSTM) adalah modifikasi dari *Recurrent Neural Network* (RNN) yang menambahkan interaksi tambahan pada setiap modulnya untuk mengatasi masalah vanishing dan *exploding gradient* dalam arsitektur RNN [12]. RNN merupakan jenis jaringan saraf tiruan yang dapat memproses data sekuensial, dengan output berupa state internal yang mencerminkan hasil proses sebelumnya atau proses berikutnya [13]. Namun, RNN memiliki masalah pada perubahan jangkauan nilai yang dapat menyebabkan gradient menghilang atau meledak saat bergerak ke lapisan berikutnya dalam arsitektur LSTM dibangun untuk mengatasi masalah ini [14].



Gambar 2. Arsitektur LSTM

Upaya mengatasi masalah RNN, LSTM menggunakan tiga gerbang: *input gate, forget gate*, dan *output gate*. Gerbang-gerbang ini berfungsi untuk mengatur aliran informasi dan memecahkan masalah gradient yang ditemui dalam RNN tradisional. LSTM juga memanfaatkan data historis dan saat ini dengan menggunakan sel memori, dimana informasi yang relevan dari masa lalu disimpan dan diakses melalui operasi gerbang terbuka atau tertutup. Arsitektur sel LSTM diilustrasikan dalam Gambar 2 [15].

E. *Recommend System Content-Based*

Sistem rekomendasi merupakan alat yang merekomendasikan item atau konten yang relevan berdasarkan preferensi dan perilaku pengguna. Terdapat dua

pendekatan utama dalam sistem rekomendasi, *Content-Based Filtering* (CBF) dan *Collaborative Filtering* (CF). Algoritma *content-based filtering* dirancang untuk merekomendasikan produk berdasarkan akumulasi pengetahuan pengguna serta rekomendasi didasarkan pada kesamaan atribut atau karakteristik item dengan preferensi pengguna, tanpa mempertimbangkan informasi dari pengguna lain. Sebagai contoh, jika pengguna menyukai beberapa restoran dengan atribut serupa seperti lokasi dan jenis makanan, sistem akan merekomendasikan restoran lain yang memiliki atribut serupa [16]. Pada teknik *collaborative filtering* berdasarkan penilaian pengguna sebelumnya. Sistem ini tidak memerlukan banyak fitur produk untuk bekerja. Alur rekomendasi dari metode *collaborative filtering* dan *content-based filtering* dapat dilihat pada Gambar 3.

Sumber: https://dqlab.id/pendekatan-machine-learning-collaborative-filtering

Gambar 3. Rekomendasi *Colaborative Filtering* dan CBF

Salah satu metode yang umum digunakan dalam CBF adalah *Cosine Similarity,* yang mengukur kesamaan antara dua vektor berdasarkan sudut kosinus di antara mereka. Metode ini sangat berguna dalam membandingkan kesamaan antara atribut pada data berbasis teks. Konteks rekomendasi *content-based*, metode ini digunakan untuk menghitung kesamaan antara atribut item (misalnya restoran) dengan preferensi pengguna, untuk memberikan rekomendasi item yang cocok dengan preferensi [17].

III. METODOLOGI PENELITIAN

Metodologi adalah sekumpulan metode ataupun tata cara yang lebih terperinci mengenai tahap-tahap melakukan sebuah penelitian untuk menyelesaikan suatu masalah. Pada bagian metodologi penelitian dijelaskan secara singkat mengenai langkah-langkah yang dilakukan dalam pengerjaan penelitian ini. Mulai dari memahami permasalahan, mempelajari kondisi dan proses bisnis saat ini, menganalisa kebutuhan data dan informasi sistem yang akan datang, hingga merancang sistem usulan.

A. *Sumber Data*

Data yang digunakan dalam penelitian ini diperoleh dari M Company Coffee. Data latihan model dikumpulkan dalam dua bahasa, yaitu Bahasa Indonesia dan Bahasa Inggris. Data dalam Bahasa Indonesia diperoleh melalui wawancara dan observasi, sedangkan data dalam Bahasa Inggris dihasilkan melalui terjemahan. Semua data tersebut diatur dalam format JSON untuk kemudahan pemanggilan respons. Contoh data diubah menjadi format tabel dan ditunjukkan pada Tabel 1.

TABEL 1. SAMPEL DATASET PELATIHAN MODEL LSTM

Input text	Intents
" Rsvp di m company coffee gimana ya?"	rsvp
"Ok bye"	bye
"Mcobot bisa ngapain saja?"	about_mcobot
"Apa minuman spesial yang tersedia di coffeeshop ini?"	Recommendation

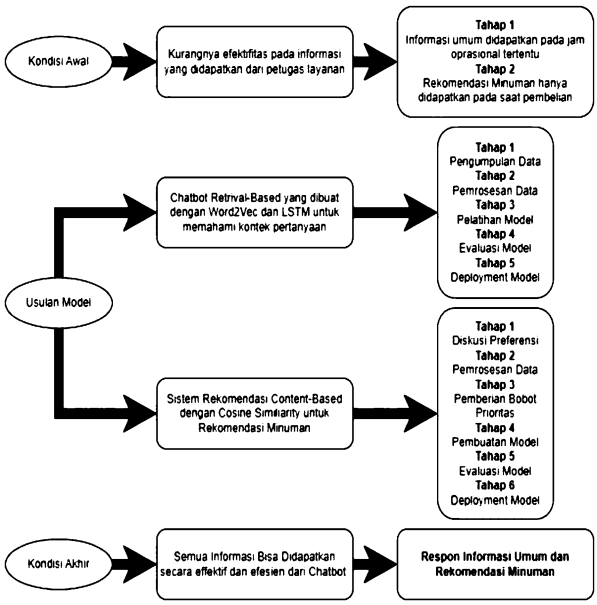
Selain data teks, data preferensi menu minuman juga digunakan dalam penelitian ini. Preferensi ini diperoleh melalui diskusi dengan pihak M Company Coffee untuk mengumpulkan preferensi terhadap menu minuman. Tabel 2 menunjukkan contoh data preferensi minuman yang terdiri dari 8 fitur yang membedakan minuman berdasarkan jenis, suhu, rasa (manis, pahit, asam, gurih), serta penggunaan susu dan sirup buah. Data ini menjadi bagian penting dalam pelatihan model.

TABEL 2. SAMPEL DATASET REFERENSI MINUMAN

Nama	Jenis	Suhu	Manis	Asam	Pahit	Susu
Affogato	1	1	3	2	3	1
Charcoal	1	1	4	3	2	1
BLACKPINK	1	1	4	1	2	1
B.O.T	1	1	3	3	2	1

B. Kerangka Penelitian

Dalam penelitian yang dilakukan, dibuat kerangka penelitian dengan tujuan untuk memudahkan dalam penyusunan persiapan penelitian yang dilakukan, agar lebih terstruktur. Kerangka penelitian dapat dilihat pada Gambar 4.



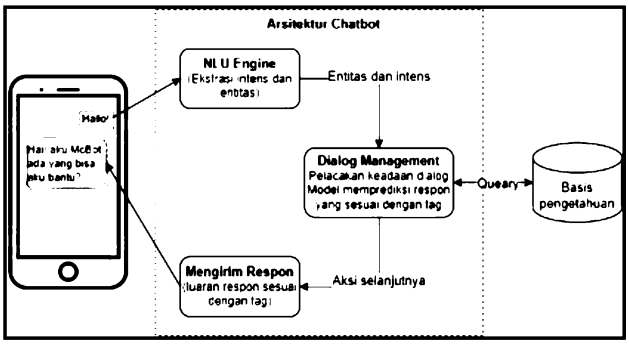
Gambar 4. Kerangka Penelitian

Gambar 4 merupakan alur penelitian yang dilakukan yang terbagi dalam tiga tahapan alur. Pertama kondisi awal dengan melakukan identifikasi pokok permasalahan pada bagian petugas layanan dalam mendapatkan informasi hanya didapatkan pada jam operasional dan informasi menu hanya pada saat melakukan transaksi saja. Tahapan alur yang kedua merupakan usulan model /solusi terkait sistem yang akan digunakan dengan menerapkan sistem *chatbot retrieval based* dengan pendekatan *Word2Vec* dan LSTM untuk memahami konteks pertanyaan dari pelanggan, serta sistem rekomendasi berbasis *content based* dengan *cosine similarity* untuk rekomendasi menu dan minuman. Tahapan ketiga dengan penerapan model yang di usulkan dapat menjadi solusi yang efektif dalam peningkatan pelayanan pada M Company Coffee.

C. Analisis Sistem

Pembangunan chatbot akan melibatkan algoritma *Word2Vec* untuk ekstraksi fitur, serta algoritma LSTM untuk

klasifikasi pertanyaan. Selain itu, metode rekomendasi sistem content-based dengan algoritma *cosine similarity* digunakan untuk rekomendasi minuman. Arsitektur chatbot terdiri dari beberapa tahapan, dimulai dari input pertanyaan pengguna pemahaman entitas dan intent, hingga pengambilan dan pengiriman respon yang sesuai, dapat dilihat pada Gambar 5.



Gambar 5. Arsitektur Chatbot

Fitur-fitur yang akan dikembangkan meliputi pemilihan bahasa, rekomendasi minuman, serta aturan masukan untuk rekomendasi minuman. Prosesnya melibatkan persiapan dan pemrosesan data, ekstraksi fitur, klasifikasi intent pertanyaan, pemilihan dan pengiriman respon, serta rekomendasi minuman. Keluaran dari sistem termasuk model *Word2Vec* dan LSTM, serta respon chatbot dan rekomendasi minuman.

Kebutuhan masukan terdiri dari teks pertanyaan pengguna, dengan peraturan khusus untuk rekomendasi minuman. Kebutuhan proses mencakup tahapan dari persiapan data hingga rekomendasi minuman. Sedangkan keluaran sistem mencakup model *Word2Vec* dan LSTM, respon chatbot, dan rekomendasi minuman. Sistem chatbot yang diusulkan, diharapkan pelayanan pelanggan pada M Coffee Company dapat menjadi lebih efisien dan responsif.

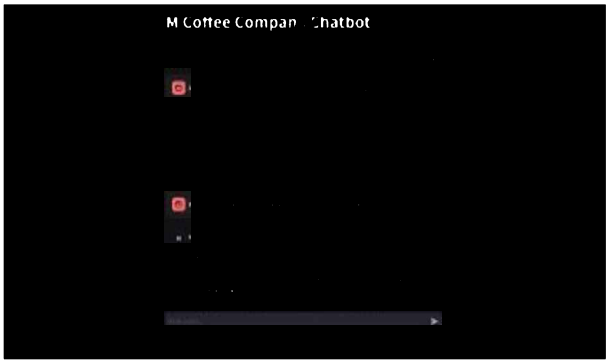
IV. HASIL DAN PEMBAHASAN

A. Hasil Uji Coba Chatbot

Percobaan dilakukan dengan memberikan pertanyaan dalam bahasa Indonesia dan bahasa Inggris setelah memilih bahasa yang cocok dengan respons chatbot. Percobaan pertama menggunakan bahasa Indonesia untuk bertanya tentang lokasi. Hasilnya, seperti yang terlihat pada Gambar 6, chatbot awalnya kesulitan memahami pertanyaan pengguna. Namun, setelah pertanyaan diubah dengan konteks yang serupa, chatbot dapat memberikan respons yang tepat.

Gambar 6. Uji Coba Bahasa Indonesia

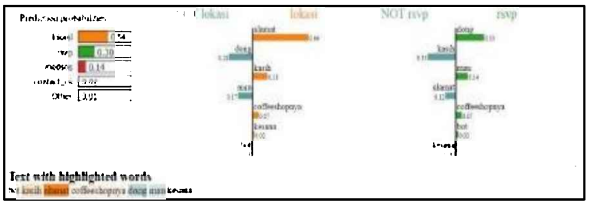
Selain percobaan dalam bahasa Indonesia, eksperimen dilakukan dengan menggunakan bahasa Inggris, seperti yang terlihat pada Gambar 7. Percobaan ini, chatbot berhasil memahami konteks pertanyaan dari pengguna dan memberikan respon yang sesuai.



Gambar 7. Uji Coba Bahasa Inggris

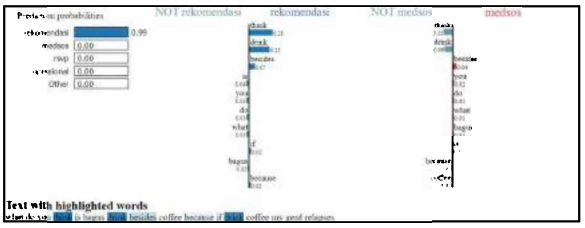
B. Pembuktian Hasil Chatbot

Pada Gambar 8, input pertama pada percobaan chatbot berbahasa Indonesia, terdapat kata-kata dengan kontribusi negatif yang signifikan, seperti "dong" dan "mau". Kata-kata ini memiliki pengaruh negatif yang kuat terhadap label "rsvp". Kehadiran beberapa kontribusi negatif yang signifikan menyebabkan probabilitas untuk label "lokasi" turun dibawah ambang batas, yang menyebabkan chatbot memberikan respons bahwa ia tidak mengerti. Hal ini terjadi karena probabilitas klasifikasi untuk label "lokasi" menjadi kurang dari 54%.



Gambar 8. Interpretasi Model Bahasa Indonesia

Pada Gambar 9, tingkat kepercayaan model terhadap label "rekomendasi" sangat tinggi (sekitar 99%) karena beberapa kata seperti "think", "drink", "besides", "if", dan "because" memberikan kontribusi positif. Kontribusi negatif memiliki nilai rendah, berbeda dengan bahasa Indonesia di mana kontribusi negatif memiliki nilai tinggi. Terdapat juga label "medsos" dalam hasil prediksi, tetapi probabilitasnya sangat rendah karena kontribusi positif sangat sedikit dibandingkan dengan kontribusi negatif yang lebih besar.



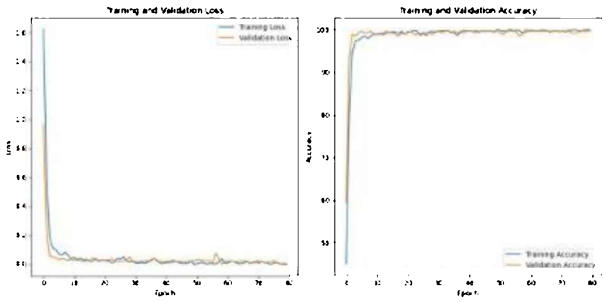
Gambar 9. Interpretasi Model Bahasa Inggris

C. *Pembahasan*

Penelitian ini melakukan serangkaian percobaan pada model berbahasa Indonesia dan Inggris dengan variasi dimensi *Word2Vec* yang digunakan dalam pelatihan, yaitu 100 dan 200 dimensi. Hasil pengujian direpresentasikan melalui *Classifier Report* yang memberikan informasi tentang *f1-score, precision, recall*, dan akurasi. Tabel 3 menunjukkan bahwa model dengan Word2Vec dimensi 100 memiliki akurasi yang lebih baik dibandingkan dengan model dimensi 200, Ini menunjukkan bahwa penggunaan *Word2Vec* dengan 100 dimensi sudah cukup, tidak perlu terlalu kompleks.

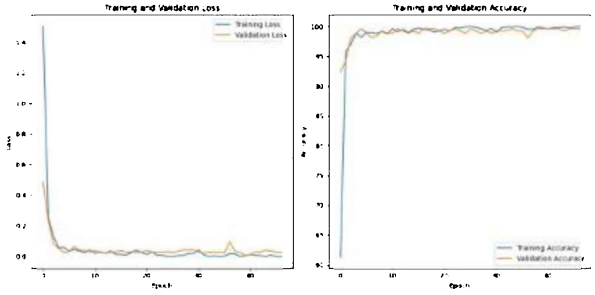
TABEL 3. *CLASSIFIER REPORT* MODEL INDONESIA

Model Indonesia	Word2Vec 100			Word2Vec 200		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Macro Avg	0.99	0.99	0.99	0.98	0.98	0.98
Weighted Avg	0.98	0.98	0.98	0.98	0.97	0.98
Akurasi	0.98			0.97		



Gambar 10. Pelatihan Model Indonesia Dimensi 100

Gambar 10 merupakan percobaan uji pelatihan dengan model Bahasa Indonesia dengan dimensi 100, selain dari *Classifier Report*, pengamatan juga menunjukkan bahwa pelatihan model dengan *Word2Vec* dimensi 100 berlangsung selama 80 epoch, sementara pelatihan model dengan dimensi 200 berlangsung antara 45 hingga 50 epoch. Perbedaan ini muncul karena pelatihan model dimensi 100 memiliki kondisi berhenti jika *loss* telah mencapai ambang batas yang ditentukan, sebagai upaya untuk menghindari *overfitting* pada model.



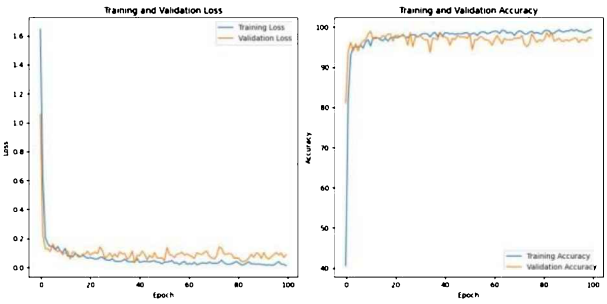
Gambar 11. Pelatihan Model Indonesia Dimensi 200

Percobaan kedua dapat ditunjukkan pada Gambar 11 dilakukan pada model berbahasa Inggris dengan mengubah dimensi *Word2Vec* menjadi 100 dan 200. Hasilnya dijelaskan dalam *Classifier Report* pada Tabel 4. Seperti pada model berbahasa Indonesia, model dengan *Word2Vec* dimensi 100 menunjukkan akurasi yang lebih tinggi dibandingkan dengan model dimensi 200.

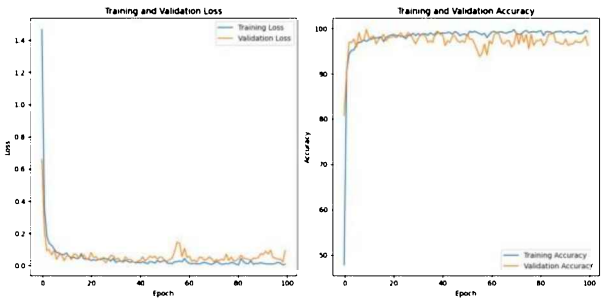
TABEL 4. *CLASSIFIER REPORT* MODEL INGGRIS

Model Inggris	Word2Vec 100			Word2Vec 200		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Macro Avg	0.79	0.79	0.78	0.79	0.78	0.76
Weighted Avg	0.81	0.80	0.79	0.80	0.78	0.77
Akurasi	0.80			0.78		

Hasil pelatihan model berbahasa Inggris menggunakan *Word2Vec* dimensi 100 terlihat pada Gambar 12, menunjukkan performa model yang baik dengan sedikit fluktuasi dalam *loss* dan akurasi.



Gambar 12. Pelatihan Model Inggris Dimensi 100



Gambar 13. Pelatihan Model Inggris Dimensi 200

Namun, model dengan *Word2Vec* dimensi 200 Gambar 13 mengalami perubahan yang mencolok pada epoch 50 hingga 60, yang mengindikasikan adanya potensi *overfitting*. Namun, pada epoch selanjutnya, model ini kembali menuju performa optimal.

V. SIMPULAN

Berdasarkan hasil penelitian dalam proyek ini, dapat ditarik beberapa kesimpulan penting:

1. Penggunaan *Chatbot Relative-Based* Multilingual yang menggabungkan algoritma *Word2Vec* dan LSTM mampu memberikan respons yang efektif dalam bahasa Indonesia dan Inggris.
2. Metode Rekomendasi *Content-Based* dengan algoritma *Cosine Similarity* berhasil digunakan untuk mengukur kemiripan antara preferensi minuman dan preferensi pengguna.
3. Model yang menggunakan *Word2Vec* dengan dimensi100 menunjukkan kinerja yang baik dengan akurasi dan *f1-score* yang tinggi. Hasil ini terlihat pada akurasi model bahasa Indonesia sebesar 98% dan model bahasa Inggris sebesar 80%.

REFERENSI

[24] Dewi, L. And Putri, S. H. (2022), Service Quality , Customer Value , And Price To Consumer Satisfaction At Kopi Kenangan Coffee Vol. 1, No. 6 Pp. 987–992.

[25] Luqmanulhakim, F., Abidin, Z., Kusumaningrum And Rastri (2022), Komunikasi Persuasif Antara Pelanggan Dan Barista Di Coffee Shop Vol. 16, No. 9 Pp. 7395–7406.

[26] Chandra, A. Y., Kurniawan, D. And Musa, R. (2020), Perancangan Chatbot Menggunakan Dialogflow Natural Language Processing (Studi Kasus: Sistem Pemesanan Pada CoffeeShop) Jurnal Media Informatika Budidarma, Vol. 4, No. 1 P. 208.

[27] Rosita, A. (2022), Chatbot Design Using Artificial Intelligence With Natural Language Processing To Increase Customer Vol. 23, No.1 Pp. 4159–4168.

[28] Dosovitsky, G., Pineda, B. S., Jacobson, N. C., Chang, C., Escoredo, M. And Bunge, E. L. (2020), Artificial Intelligence Chatbot For Depression: Descriptive Study Of Usage Jmir Formative Research, Vol. 4, No. 11 Pp. 1–13.

[29] Akkineni, H., Lakshmi, P. V. S. And Sarada, L. (2022), Design And Development Of Retrieval- Based Chatbot Using Sentence Similarity Lecture Notes In Networks And Systems, Vol. 244, No. January Pp. 477–487.

[30] Chai, C. P. (2023), *Comparison Of Text Preprocessing Methods* Natural LanguageEngineering, Vol. 29, No. 3 Pp. 509–553.

[31] [8] Rifaldi, D., Fadlil, A. And Herman (2023), Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet Mental Health Jurnal Pendidikan Teknologi Informasi, Vol. 3, No. 2 Pp. 161–171.

[32] Qorina, E. S. (2020), Analisis Perbandingan Metode Fast Text Dan Word2vec Pada Query Kesamaan Semantik Sistem Temu Kembali Informasi Sirah Nabawiyah Skripsi Oleh : Etna Syirfa Qorina 2020 M / 1442 H.

[33] Gu, J. K., Li, G., Vo, N. D. And Jung, J. J. (2022), Contextual Word2vec Model For Understanding Chinese Out Of Vocabularies On Online Social Media International Journal On Semantic Web And Information Systems, Vol. 18, No. 1 Pp. 1–14.

[34] Gunawan, Y., Young, J. C. And Rusli, A. (2022), Fasttext Word Embedding And Random Forest Classifier For User Feedback Sentiment Classification In Bahasa Indonesia Ultimatics :Jurnal Teknik Informatika, Vol. 13, No. 2 Pp. 101–107.

[35] Alghifari, D. R., Edi, M. And Firmansyah, L. (2022), Implementasi Bidirectional Lstm Untuk Analisis Sentimen Terhadap Layanan Grab Indonesia Bidirectional Lstm Implementation For Sentiment Analysis Against Grab IndonesiaServices Vol. 12 Pp. 89–99.

[36] Yuliana Romadhoni (2022), Klasifikasi KalimatPerbincangan Masyarakat Terhadap Pandemi Covid-19 Pada Twitter Dengan Metode Long Shortterm Memory No. 8.5.2017 Pp. 2003–2005.

[37] Wiranda, L. And Sadikin, M. (2019), *Penerapan Long Short Term Memory Pada Data Time Series Untuk Memprediksi Penjualan Produk Pt. Metiska Farma* Jurnal Nasional Pendidikan Teknik Informatika (Janapati), Vol. 8, No. 3 Pp. 184–196.

[38] Silvanie, A. And Subekti, R. (2022), *Aplikasi Chatbot Untuk Faq Akademik Di Ibi-K57 Dengan Lstm Dan Penyematan Kata* Jiko (Jurnal Informatika Dan Komputer), Vol. 5, No. 1 Pp. 19–27.

[39] Putri, D. A., Pramesti, D., I, D. And Santiyasa,W. (2022), *Penerapan Metode Content-Based Filtering Dalam Sistem Rekomendasi Video Game* Jnatia, Vol. 1, No. 1 Pp. 229–234.

[40] Raharjo, P. N., Handojo, A. And Juwiantho, H. (2022), Sistem Rekomendasi Content Based Filtering Pekerjaan Dan Tenaga Kerja Potensial Menggunakan Cosine Similarity Jurnal Invra, Vol. 10, No. 2 Pp. 1–6.

Optimalisasi Produktivitas Karyawan dengan Prediksi Random *Forest Classification*

Rizky Diar Panuntun
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
Rizky.5200411499@student.uty.ac.id

Candika Silai Prahma Setiadi
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
candika.5200411516@student.uty.ac.id

Syahrul Gunawan Ramdhani
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
syahrul.5200411508@student.uty.ac.id

Adie Gunawan Alwani
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
adie.5200411486@student.uty.ac.id

Roy Fasti
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
roy.5200411463@student.uty.ac.id

Abstrak—Industri garmen, sebagai peran utama dalam menanggapi perubahan global, menghadapi tantangan efisiensi produksi dan kepuasan karyawan. Produktivitas karyawan menjadi elemen kunci dalam menjaga daya saing industri ini. Penelitian ini memanfaatkan konsep Machine Learning (ML) dalam kerangka Artificial Intelligence (AI), terutama menggunakan algoritma Random Forest Classifier (RFC) untuk memahami dan memprediksi faktor-faktor apa saja yang mempengaruhi produktivitas karyawan. Melalui pengumpulan dan analisis data, penelitian ini bertujuan mengidentifikasi hubungan antara jam kerja, istirahat, dan tingkat kepuasan dengan produktivitas. Hasil penelitian ini dapat memberikan wawasan berharga bagi pengambil keputusan di industri garmen. Dengan pembagian data training sebesar 80% dan data testing sebesar 20%, penelitian ini mencapai tingkat akurasi sebesar 85%, menggambarkan keberhasilan dalam memprediksi faktor-faktor yang memengaruhi produktivitas karyawan.

Kata kunci—*Random Forest*, Klasifikasi, Industri Garmen, Karyawan, Produktivitas

Abstract— The garment industry, as a primary player in responding to global changes, faces challenges in production efficiency and employee satisfaction. Employee productivity is a key element in maintaining the competitiveness of this industry. This research leverages the concept of Machine Learning (ML) within the framework of Artificial Intelligence (AI), specifically employing the Random Forest Classifier (RFC) algorithm to understand and predict the factors influencing employee productivity. Through the collection and analysis of data, the study aims to identify the relationships between working hours, breaks, and satisfaction levels with productivity. The results of this research can provide valuable insights for decision-makers in the garment industry. With an 80% training data and 20% testing data split, the study achieves an accuracy rate of 85%, demonstrating success in predicting factors affecting employee productivity.

Keywords—Random Forest, Classification, Garment Industry, Employees, Productivity

I. PENDAHULUAN

Industri garmen adalah salah satu contoh utama dari perubahan yang terjadi dalam dunia industri saat ini. Di tengah perubahan global yang terus berlanjut, industri ini memiliki peran sentral dalam memenuhi permintaan tinggi di seluruh dunia terhadap produk garmen[1]. Proses produksi garmen masih sangat terkait dengan tenaga kerja manusia, dengan banyak tahapan yang dilakukan secara manual. Oleh karena itu, efisiensi kinerja produksi dan pengiriman oleh karyawan di perusahaan garmen menjadi faktor penting dalam menjaga daya saing industri ini. Pentingnya produktivitas karyawan dalam industri garmen tidak dapat diabaikan. Dalam menjaga kinerja tim kerja tetap optimal, para pengambil keputusan di perusahaan garmen dihadapkan pada tantangan yang perlu diatasi. Mereka memerlukan alat dan pengetahuan untuk melacak, menganalisis, dan memprediksi kinerja produktivitas tim kerja di pabrik-pabrik mereka. Hal ini tidak hanya berdampak pada efisiensi operasional, tetapi juga berpotensi mempengaruhi kepuasan pelanggan dan profitabilitas perusahaan.

Penelitian ini bertujuan untuk mengatasi tantangan-tantangan yang dihadapi dalam industri garmen dengan memanfaatkan metode *machine learning*. *Machine Learning* adalah suatu konsep dalam kerangka *Artificial Intelligence* (AI) yang terkait dengan pengembangan teknik, metode, dan algoritma yang memungkinkan sistem untuk memperoleh pengetahuan dari data yang tersedia[2]. Dalam *Machine learning* ada beberapa algoritma dimana pada penelitian ini, akan menggunakan algoritma *machine learning* yang dikenal dengan sebutan *Random Forest Classifier* (RFC), RFC merupakan cara untuk mengelompokkan data ke dalam kelas-kelas berdasarkan karakteristik[3].

Melalui pengumpulan dan analisis data yang relevan, penelitian ini bertujuan untuk meramalkan sejauh mana produktivitas karyawan di industri garmen dapat dipengaruhi oleh faktor-faktor kunci seperti jam kerja, jumlah istirahat yang mereka ambil, dan tingkat kepuasan mereka terhadap pekerjaan mereka. Peneliti melakukan penyelidikan hubungan dua faktor atau lebih dan variabel target dengan metode *Machine Learning* (ML)[4]. Dengan pendekatan ini, peneliti berharap penelitian ini akan memberikan wawasan

berharga kepada para pengambil keputusan di industri garmen.

Diharapkan bahwa hasil dari penelitian ini akan memberikan panduan bagi industri garmen dalam menghadapi tantangan globalisasi dan persaingan yang semakin ketat. Melalui pendekatan berbasis data dan *machine learning*, penelitian ini diharapkan dapat menjadi landasan bagi perusahaan garmen untuk meningkatkan produktivitas, efisiensi, dan kualitas produk mereka. Selain itu, peneliti juga berharap bahwa penelitian ini akan mendukung pertumbuhan berkelanjutan dalam industri garmen di era modern yang terus berubah.

II. Studi Literatur

Pada penelitian yang dilakukan oleh P. Rani dkk, pada tahun 2018 dengan judul penelitian *Traffic Accident Detection Using Random Forest Classifier*. Penelitian tersebut menggunakan tiga algoritma, yaitu: *Artificial Neural Network* (ANN), *Random Forest* (RF), dan *Support Vector Machine* (SVM). Dari penelitian ini menghasilkan Evaluasi menunjukkan bahwa RF mampu untuk mendeteksi kecelakaan dengan akurasi 92% sementara ANN mampu untuk mendeteksi dengan akurasi 90% dan SVM dengan akurasi 89%[5].

Pada penelitian yang dilakukan oleh Soumya S., dan Pramod K.V. pada tahun 2021 dengan judul penelitian *A Decision Support System for Heart Disease Prediction Based Upon Machine Learning*. Penelitian tersebut menggunakan lima algoritma yaitu *Naive Bayes* (NB), *Support Vector Machine* (SVM), *Logistic Regression* (LR), *Random Forest* (RF), dan *Adaboost*. Pada penelitian ini *Random Forest* menghasilkan akurasi terbaik sebesar 86.60% sedangkan NB menghasilkan akurasi 83.55%, SVM menghasilkan akurasi 84.46%, LR menghasilkan 85.07%, dan *Adaboost* menghasilkan akurasi 86.59%[6].

Pada penelitian yang dilakukan oleh Yassine Al Amrani, Mohamed Lazaar dan Kamal Eddine El Kadiri pada tahun 2020 dengan judul penelitian *Sentiment analysis of malayalam tweets using machine learning techniques*. Penelitian tersebut menggunakan tiga algoritma yaitu *Naive Bayes* (NB), *Support Vector Machine* (SVM) dipisah menjadi dua yaitu *(Linear Kernel* dan RBF), dan *Random Forest* (RF). Pada penelitian ini *Random Forest* dengan *Unigram* dengan *Sentiwordnet* termasuk kata negasi, mendapatkan akurasi tertinggi 95.6% sedangkan NB menghasilkan akurasi 94.4%, SVM *with Linear Kernel* menghasilkan akurasi 94.5%, dan SVM *with* RBF menghasilkan 94.8[7].

Pada penelitian yang dilakukan oleh Kailong Liu dkk, pada tahun 2018 dengan judul penelitian *Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis*. penelitian tersebut menggunakan tiga algoritma yaitu *Random Forest* (RF), *Support Vector Machine*, dan penggabungan algoritma *Random Forest Support Vector Machine* (RFSVM). Pada penelitian ini RFSVM menemukan bahwa *precision, recall* dan *F-measure* adalah 83.4%, 83.4%, dan 83.4%[8].

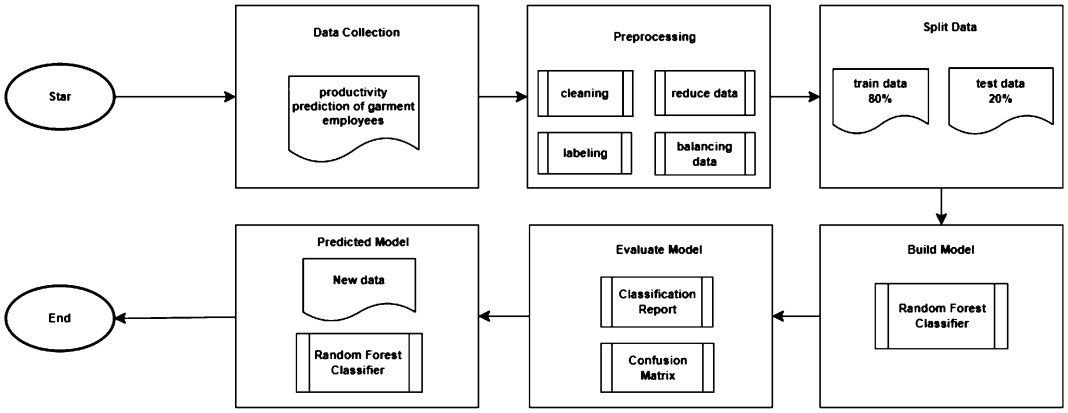
Pada penelitian yang dilakukan oleh Gina L. O'Neil dkk, pada tahun 2021 dengan judul penelitian *Feature Analysis and Modelling of Lithium-ion Batteries Manufacturing based on Random Forest Classification*. penelitian tersebut menggunakan *Decision Tree* (DT), *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), dan *Random Forest* (RF). pada penelitian tersebut *mass load* RF memiliki *macroF1* 90.1% sedangkan pada *porosity* RF memiliki *macroF1* 66.4%[9].

Pada penelitian sebelumnya yang dilakukan oleh Daniel Marpaung, S Surmarno dan Indra Gunawan pada tahun 2020 dengan judul Prediksi Produktivitas Kelapa Sawit di PTPNIV dengan Algoritma Backpropagation. Penelitian tersebut melakukan penelitian terhadap prediksi produktivitas kelapa sawitpada PTPN IV kebun Dolok Sinumbah dengan memanfaatkan algoritma Backpropagation. Adapun akurasi yang dihasilkan pada penelitian tersebut sebesar 92%[10].

Perbedaan dari penelitian sebelumnya dari penelitian yang akan dikakukan terletak pada metode yang digunakan. Penelitian ini menggunakan metode *Random Forest Classifier* untuk memprediksi produktivitas karyawan pada industri garmen. Metode *Random Forest Classifier* digunakan untuk mengklasifikasikan karyawan kedalam jenis produktif dan tidak produktif berdasarkan atribut yang berpengaruh.

III. Metodologi Penelitian

Pada penelitian ini ada beberapa tahap yaitu data *collection* yang merupakan *dataset* yang digunakan. *Preprocessing* yang didalamnya terdapat beberapa proses yaitu (*cleaning reduce data*, *labelling*, dan *balancing data*). *Split* data yang didalamnya terdapat 80% *train* data dan 20% *test* data. *Build* model yang digunakan pada penelitian ini adalah *Random Forest Classifier*. Selanjutnya tahap evaluasi model yang menggunakan *classification report* dan *confusion matrix*. tahap terakhir yaitu *predicted* model yang terdapat *new data* dan *Random Forest Classifier*. Berikut ilustrasi Tahapan Penelitian terdapat pada gambar 1.



Gambar 1. Metodologi Penelitian

A. *Data Collection*

Data pada penelitian ini merupakan data *productivity prediction of garment employees*, data diperoleh pada sistus Kaggle.com yang merupakan situs yang bisa mendapatkan *dataset* berbagai ragam jenis data. Pada penelitian ini pengumpulan data mencakup lima belas atribut dan informasi yang diperlukan untuk memahami faktor-faktor yang mempengaruhi produktivitas para karyawan. Pada penelitian ini data yang didapatkan peniliti berjumal sebayak 1197 dan attribut dari *dataset* sebanyak 15 atribut.

B. *Preprocessing*

Langkah awal yang melibatkan sejumlah kombinasi dan proses yang memerlukan intervensi atau penyesuaian dari pengguna[11]. Pada tahap ini, data awal disiapkan dan dimodifikasi untuk memastikan kualitas, integritas, dan ketersediaan data yang diperlukan dalam analisis atau pemodelan lebih lanjut.

1) Data Cleaning

Data Cleaning atau pembersihan data adalah proses pengelolaan dan pemrosesan data untuk mengidentifikasi, mengoreksi, menghapus kesalahan, atau ketidaksesuaian dalam *dataset*. Tujuan utama dari pembersihan data adalah memastikan data yang digunakan untuk analisis atau pemodelan adalah akurat dan konsisten[12].

2) Data Reduction

Data Reduction atau reduksi data adalah proses mengurangi volume atau kompleksitas data dan mempertahankan sebagian besar informasi yang relevan. Tujuan dari reduksi data adalah untuk membuat *dataset* yang lebih kecil dan lebih mudah diolah.

3) Labeling

Labeling adalah proses untuk menetapkan kategori atau klasifikasi tertentu kepada setiap entitas atau sampel dalam *dataset*. Proses ini dapat menjadi kunci dalam pengembangan model pembelajaran mesin yang bersifat *supervised learning*.

4) Balancing Data

Balancing data adalah proses untuk mengatasi ketidakseimbangan dalam distribusi kategori atau kelas pada *dataset*. Ketidakseimbangan ini terjadi ketika jumlah sampel dalam satu kelas jauh lebih banyak atau lebih sedikit dibandingkan dengan kelas lainnya[13]. Situasi ini dapat mempengaruhi kinerja model terutama dalam tugas klasifikasi. *Balancing data* bertujuan untuk memastikan bahwa model dapat menghasilkan prediksi yang baik.

C. *Split Data*

Train data dan *test data* dengan rasio 80% dan 20%. Data pelatihan digunakan untuk melatih model atau melakukan analisis, sementara data pengujian digunakan untuk menguji kinerja model atau melakukan evaluasi. Rasio 80% dan 20% mengindikasikan proporsi data yang digunakan dalam masing-masing tahap, yang umum digunakan dalam praktik pembelajaran mesin dan analisis data untuk memastikan model yang baik dan evaluasi yang andal[14].

D. *Build Model*

Random Forest Classifier adalah salah satu algoritma *machine learning* yang digunakan untuk tugas klasifikasi. Algorima ini memanfaatkan konsep *ensemble learning*, yang menggabungkan hasil dari beberapa model pembelajaran mesin untuk meningkatkan akurasi dan kestabilan prediksi[15]. Dalam konteks klasifikasi, *Random Forest Classifier* membagi *dataset* menjadi subset acak dan membangun sejumlah pohon keputusan independen. Untuk pembuatan setiap pohon keputusan dapat menggunakan persamaan (1).

$$f(x) = sum_{i=1}^{m} w_i h_i(x) \quad (1)$$

Setiap pohon keputusan dibangun menggunakan subset acak dari fitur-fitur yang tersedia, pemilihan fitur secara acak membantu mencegah pohon keputusan menjadi serupa antara satu dengan yang lain. Kemudian hasil dari setiap pohon keputusan akan digabungkan untuk prediksi akhir. Untuk membentuk *Random Forest* dapat menggunakan persamaan (2).

$$F(x) = frac1M\ sum_{i=1}^{M} f_i(x) \quad (2)$$

E. *Evaluate Model*

Dalam penelitian ini evaluasi model *Random Forest Classifier* menggunakan *classification report* dan *confusion matrix*. *Classification report* adalah sebuah laporan yang memberikan evaluasi kinerja dari suatu model klasifikasi berdasarkan nilai dari *confusion matrix*. Sedangkan *confusion matrix* adalah sebuah tabel yang digunakan untuk mengevaluasi kinerja suatu model klasifikasi[16]. *Confusion matrix* membandingkan hasil prediksi model dengan nilai

sebenarnya (*ground truth*) dari *dataset*. *Confusion matrix* menghasilkan metrik evaluasi model, seperti *Accuracy*, *Precision*, *Recall*, dan *F1-score*.

Accuracy atau akurasi adalah nilai proporsi dari keseluruhan prediksi yang benar. Akurasi dapat dihitung menggunakan persamaan (3).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Precision adalah ukuran sejauh mana model memprediksi sebagai positif oleh model dan berapa banyak yang benar-benar positif. *Precision* dapat dihitung menggunakan persamaan (4).

$$Precision = \frac{TP}{(TP + FP)} \quad (4)$$

Recall adalah ukuran sejauh mana model dapat menemukan prediksi yang seharusnya positif. *Recall* dapat dihitung menggunakan persamaan (5).

$$Recall = \frac{TP}{(TP + FN)} \quad (5)$$

F1-score adalah nilai rata-rata dari *precision* dan *recall*. *F1-score* dapat dihitung menggunakan persamaan (6).

$$f1 - score = 2 * \frac{recall * precision}{recall + precision} \quad (6)$$

F. Predicted Model

Tahap ini merupakan tahap implementasi dari pelatihan model yang peneliti lakukan. pada tahap ini peneliti akan memanfaatkan data baru dan model yang telah dilatih. Data baru ini berfungsi untuk menguji model yang telah dilatih, data baru ini belum memiliki attribut target label sehingga target label akan didapatkan dari model yang telah dibangun. Selain itu, manfaat dari tahap ini juga berfungsi untuk menguji implementasi model berhasil atau tidak.

IV. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk memprediksi produktivitas karyawan pada sebuah Perusahaan garmen menggunakan teknik *Random Forest Classifier*. Dalam penelitian ini, fokus diberikan pada sembilan atribut yang dipilih dengan cermat, yaitu: *team, smv, wip, over_time, incentive, idle_time, idle_men, no_of_style_change,* dan *no_of_workers*. Pemilihan atribut ini didasarkan pada analisis korelasi dan hubungan yang dianggap signifikan terhadap produktivitas karyawan.

Prediksi dilakukan dengan membagi karyawan menjadi dua kategori utama: produktif dan tidak produktif. Model *Random Forest Classifier* dipilih sebagai alat utama untuk melakukan klasifikasi ini. Penggunaan model ini diharapkan dapat menghasilkan prediksi yang akurat berdasarkan atribut-atribut yang telah dipilih.

Dalam upaya menentukan kategori produktif dan tidak produktif, pendekatan yang diambil dalam penelitian ini melibatkan perhitungan nilai ambang batas atau *threshold*. Nilai ambang batas ini diambil dari selisih antara atribut *actual_productivity* dan *targeted_productivity* yang dijadikan sebagai target label. Pendekatan ini memberikan fleksibilitas untuk menyesuaikan kriteria kategorisasi produktivitas sesuai dengan karakteristik *dataset*. Dengan menggunakan selisih antara *actual* dan *targeted productivity*, model dapat membedakan kategori karyawan yang melebihi, mencapai, atau kurang dari target produktivitas. Penggunaan nilai ambang batas ini memberikan cara yang intuitif dan fleksibel untuk memahami sejauh mana produktivitas karyawan memenuhi atau melampaui target yang telah ditetapkan. Atribut yang digunakan untuk melakukan klasifikasi menggunakan *Random Forest Classifier* dapat dilihat pada tabel 1.

TABEL 1. ATRIBUT YANG DIPILIH DENGAN REPRESENTASI

No	Variabel	Representasi
1	team	x_1
2	smv	x_2
3	wip	x_3
4	over_time	x_4
5	incentive	x_5
6	idle_time	x_6
7	idle_men	x_7
8	no_of_style_change	x_8
9	no_of_workers	x_9
10	productivity	y

Model *Random Forests* yang dikembangkan untuk klasifikasi produktivitas telah berhasil dievaluasi menggunakan data pengujian. Hasil evaluasi model menunjukkan tingkat akurasi sebesar [85%], yang menggambarkan sejauh mana model dapat dengan tepat memprediksi kelas biner "produktif" dan "tidak produktif" pada *dataset*. Hasil evaluasi dapat dilihat menggunakan *Classification Report* pada tabel 2.

TABEL 2. CLASSIFICATION REPORT

Classification Report

	precision	recall	f1-score	support
0.0	0.84	0.85	0.84	173
1	0.85	0.84	0.85	177
accuracy			0.85	350
macro avg	0.85	0.85	0.85	350
weighted avg	0.85	0.85	0.85	350

Confusion matrix menyajikan rincian performa model, termasuk jumlah *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN). Visualisasi *confusion matrix* memberikan pandangan yang lebih rinci tentang area di mana model dapat melakukan perbaikan. Hasil dari *confusion matrix* dapat dilihat pada tabel 3.

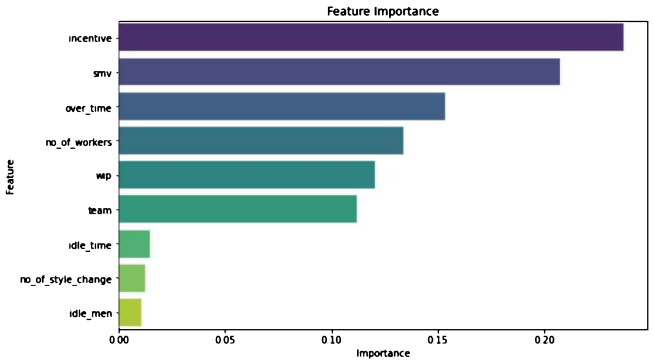
.

TABEL 3. CONFUSION MATRIX

Actual	*Not Productive*	147	26
	Productive	28	149
		Not productive	*Productive*
		Predicted	

Tingkat akurasi yang tinggi menunjukkan bahwa model mampu membuat prediksi dengan tepat, dan karenanya, dapat digunakan sebagai alat yang efektif untuk memprediksi produktivitas berdasarkan fitur-fitur yang ada dalam *dataset*.

Analisis *feature importance* menyoroti fitur-fitur yang paling berpengaruh dalam membuat keputusan klasifikasi. Hasilnya menunjukkan bahwa *incentive*, *smv, over_tim, no_of_workers, wip, team, idle_time, no_of_style_change* dan *idle_men,* yang dapat memberikan wawasan penting terkait faktor-faktor yang memengaruhi produktivitas.



Gambar 2. Featur Importance

Dari gambar 2 diatas dapat dilihat atribut yang berpengaruh dalam prediksi yang dibuat oleh model, dari hasil diatas kita dapat melihat kontribusi *relative* dari setiap fitur terhadap hasil klasifikasi. Berikut adalah penjelasan pengaruh dari setiap atribut.

1) Incentive

Fitur ini memiliki kontribusi paling tinggi terhadap keputusan klasifikasi. Besarannya *incentive* mungkin memberikan dorongan pada pekerja untuk meningkatkan produktivitas, kebijakan *incentive* tertentu dapat memotivasi tim atau individu untuk mencapai target produktivitas yang lebih tinggi.

2) SMV

SMV (*Standard Minute Value*) memiliki kontribusi yang signifikan terhadap keputusan klasifikasi. SMV mengukur standar waktu yang diperlukan untuk menyelesaikan suatu pekerjaan, variasi SMV dapat mencerminkan perbedaan kompleksitas atau tingkat kesulitan pekerjaan.

3) Over Time

Durasi *over time* juga memiliki dampak yang cukup besar pada keputusan klasifikasi. Durasi *over time* mencerminkan tambahan waktu kerja di luar jam kerja normal, *over time* mungkin terkait dengan kebutuhan untuk menyelesaikan pekerjaan lebih cepat atau meningkatkan produksi.

4) No of Workers

Jumlah pekerja juga memiliki pengaruh yang signifikan terhadap hasil klasifikasi. Jumlah pekerja dapat mempengaruhi produktivitas secara langsung, kehadiran lebih banyak pekerja dapat meningkatkan kemampuan untuk menyelesaikan pekerjaan dengan cepat.

5) WIP

Work in Progress (WIP) memiliki dampak yang cukup besar pada hasil klasifikasi. *Work in Progress* mencerminkan tingkat pekerjaan yang sedang berlangsung, WIP yang tinggi mungkin menunjukkan adanya pekerjaan yang harus diselesaikan.

6) Team

Tim atau kelompok kerja juga memberikan kontribusi yang cukup signifikan. Anggota tim atau kelompok kerja dapat saling mendukung dan berkolaborasi, kinerja tim dapat mempengaruhi produktivitas secara keseluruhan. Perbedaan dalam tim dapat mempengaruhi keputusan klasifikasi.

7) Idle Time

Meskipun memiliki kontribusi yang lebih rendah, *idle time* masih memiliki dampak pada hasil prediksi, waktu yang tidak produktif dapat memberikan kontribusi pada hasil klasifikasi. Waktu yang tidak produktif atau *idle time* dapat menunjukkan gangguan dalam alur kerja, *idle time* yang rendah mungkin menunjukkan efisiensi yang lebih baik. Meskipun memiliki bobot lebih rendah, model masih memberikan perhatian pada *idle time* karena melihatnya sebagai faktor yang dapat memengaruhi produktivitas.

8) No of Style Change

Jumlah perubahan gaya atau model memiliki dampak yang lebih rendah pada keputusan klasifikasi, namun masih memberikan kontribusi terhadap hasil prediksi. Jumlah perubahan gaya atau model dapat mencerminkan variasi dalam jenis pekerjaan yang dilakukan. Model memberikan bobot yang lebih rendah pada fitur ini, namun masih memandangnya sebagai faktor yang dapat mempengaruhi hasil produksi

9) Idle Men

Idle men memiliki kontribusi yang lebih rendah dibandingkan fitur lainnya, meskipun demikian, tetap memainkan peran dalam pembuatan keputusan. *Idle men* mencerminkan jumlah pekerja yang tidak aktif, jumlah pekerja yang tidak aktif dapat mempengaruhi kecepatan produksi. Meskipun memiliki bobot yang lebih rendah, model masih memberikan perhatian pada *idle men* karena melihatnya sebagai faktor yang dapat memengaruhi produktivitas, terutama jika jumlah pekerja yang tidak aktif signifikan.

Berdasarkan hasil *feature importance*, kita dapat menyimpulkan bahwa faktor-faktor seperti insentif, SMV, durasi *over time*, jumlah pekerja, WIP, dan tim memiliki dampak besar terhadap hasil klasifikasi. Model cenderung memprioritaskan fitur-fitur ini dalam membuat prediksi

karena melihatnya sebagai indikator yang kuat untuk produktivitas.

V. KESIMPULAN DAN SARAN

Dengan menggunakan model *Random Forest Classifier*, penelitian ini bertujuan untuk memberikan wawasan yang lebih baik tentang faktor-faktor yang mempengaruhi produktivitas karyawan dalam konteks industri garmen. Analisis hasil yang mendalam diharapkan dapat memberikan rekomendasi dan solusi untuk meningkatkan produktivitas karyawan, yang pada gilirannya dapat mendukung efisiensi operasional perusahaan.

Sebagai saran pengembangan, pertimbangkan untuk melakukan optimasi parameter model atau eksplorasi teknik pembelajaran mesin yang lebih canggih. Selain itu, penambahan *dataset* yang lebih besar dan relevan dapat memperkaya pemodelan, dan validasi eksternal dapat dilakukan untuk memastikan generalisasi model terhadap *dataset* yang tidak digunakan selama pelatihan.

REFERENSI

[1] I. Fauzi, F. Ikhsan, and N. Triristina, "Dampak Garmen Impor Bekas Terhadap Daya Beli Produk Garmen Lokal," *NiCMa: National Conference Multidisplinary*, vol. 1, no. 1, 2021.

[2] Y. Ling Goh, C. Long Ng, and R. Ling Leh Bin, "Multiple Linear Regression," *International Journal of Advanced Natural Sciences and Engineering Researches*, vol. 4, no. 4, pp. 163–168, 2023, doi: 10.59287/ijanser.2023.7.4.644.

[3] A. Nugroho, I. Asror, and Y. F. A. Wibowo, "Klasifikasi Tingkat Kualitas Udara DKI Jakarta Berdasarkan Open Government Data Menggunakan Algoritma Random Forest," *eProceedings of Engineering*, vol. 10, No. 2, 2023.

[4] L. Hickman and M. Akdere, "Developing intercultural competencies through virtual reality: Internet of Things applications in education and learning," in *2018 15th Learning and Technology Conference, L and T 2018*, Institute of Electrical and Electronics Engineers Inc., May 2018, pp. 24–28. doi: 10.1109/LT.2018.8368506.

[5] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *J Reliab Intell Environ*, vol. 7, no. 3, pp. 263–275, Sep. 2021, doi: 10.1007/s40860-021-00133-6.

[6] S. S. and P. K.V., "Sentiment analysis of malayalam tweets using machine learning techniques," *ICT Express*, vol. 6, no. 4, pp. 300–305, Dec. 2020, doi: 10.1016/j.icte.2020.04.003.

[7] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, "Random forest and support vector machine based hybrid approach to sentiment analysis," in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 511–520. doi: 10.1016/j.procs.2018.01.150.

[8] K. Liu, X. Hu, H. Zhou, L. Tong, W. D. Widanage, and J. Marco, "Feature Analyses and Modelling of Lithium-ion Batteries Manufacturing based on Random Forest Classification," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2102.06029

[9] G. L. O'Neil, J. L. Goodall, and L. T. Watson, "Evaluating the potential for site-specific modification of LiDAR DEM derivatives to improve environmental planning-scale wetland identification using Random Forest classification," *J Hydrol (Amst)*, vol. 559, pp. 192–208, Apr. 2018, doi: 10.1016/j.jhydrol.2018.02.009.

[10] D. Marpaung, S. Sumarno, and I. Gunawan, "Prediksi Produktivitas Kelapa Sawit di PTPN IV dengan Algoritma Backpropagation," *Kajian Ilmiah Informatika & Komputer*, vol. 1, no. 2, 2020.

[11] M. Soni and S. Varma, "Diabetes Prediction using Machine Learning Techniques," 2020. [Online]. Available: www.ijert.org

[12] N. P. A. Widiari, I. M. A. D. Suarjaya, and D. P. Githa, "Teknik Pengolahan Data Cleaning," *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, vol. 8, no. 2, 2020.

[13] H. T. Nainggolan, B. Hananto, and B. T. Wahyono, "Klasifikasi Sentimen Menggunakan Algoritma K-Nearest Neighbour (Studi Kasus: Magang Merdeka Belajar)," *Informatik : Jurnal Ilmu Komputer*, vol. 19, no. 1, 2023, doi: 10.52958/iftk.v19i1.4777.

[14] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13. Institute of Electrical and Electronics Engineers Inc., pp. 6308–6325, 2020. doi: 10.1109/JSTARS.2020.3026724.

[15] A. R. Putri, R. Purnamasari, and Edwar, "Perbandingan Metode Klasifikasi Pemetaan Tutupan Lahan Menggunakan Algoritma Machine Learning Pada Citra Satelit Dengan Google Earth Engine," *e-Proceeding of Engineering*, vol. 8, no. 6, 2022.

[16] Lutfia Afifah, "Apa itu Confusion Matrix di Machine Learning?," https://ilmudatapy.com/apa-itu-confusion-matrix/.

Implementasi Metode *K-Nearest Neighbors* dalam Memprediksi Harga Mobil Bekas

Robi Ardiansyah
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
robi.5200411477@student.uty.ac.id

Sulthan As Shiddiq
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
shiddiq.5200411019@student.uty.ac.id

Risky Devandra Hartana
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
risky.5200411476@student.uty.ac.id

Muhammad Raka N. Fathansyach
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
muhammad.5200411472@student.uty.ac.id

Bina Sukma Adicahya
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
bina.5200411466@student.uty.ac.id

Abstrak—Studi kasus ini bertujuan untuk mengembangkan model prediksi harga mobil bekas dengan menerapkan algoritma *K-Nearest Neighbors* (KNN). Dataset yang digunakan mencakup 9 kolom, termasuk model mobil, tahun pembuatan, harga, jenis transmisi, jarak tempuh, jenis bahan bakar, pajak jalan raya, konsumsi bahan bakar dalam mil per galon (*mpg*), dan ukuran mesin. Data telah melalui tahap pembersihan yang mencakup penghapusan duplikat dan penyelarasan kolom. Proses analisis ini bertujuan untuk memahami hubungan antar fitur dan mengidentifikasi pola yang dapat mendukung prediksi harga mobil bekas. Metode KNN digunakan karena kemampuannya dalam menangani dataset berdimensi tinggi dan menghasilkan prediksi yang dapat diandalkan. Hasil studi ini diharapkan dapat memberikan wawasan berharga bagi industri otomotif dan para konsumen dengan memberikan perkiraan harga yang akurat berdasarkan karakteristik mobil tertentu.

Kata Kunci—Prediksi, Harga mobil, K-Nearest Neighbors, Data science, Otomotif.

I. PENDAHULUAN

Pasar mobil bekas mengalami dinamika yang kompleks dan dipengaruhi oleh berbagai faktor seperti model, tahun pembuatan, transmisi, jarak tempuh, jenis bahan bakar, pajak jalan raya, konsumsi bahan bakar, dan ukuran mesin [1]. Dalam era teknologi dan mobilitas yang semakin berkembang, keputusan pembelian mobil bekas menjadi semakin krusial bagi konsumen. Meningkatnya kebutuhan untuk mengambil keputusan yang lebih cerdas dalam pembelian mobil bekas mendorong pengembangan model prediksi harga. Keberhasilan industri otomotif dalam memberikan solusi yang akurat dan dapat diandalkan akan memberikan kepercayaan kepada konsumen, sehingga penelitian ini mencoba untuk mengisi kekosongan tersebut.

Penelitian ini merespon kebutuhan tersebut dengan menggabungkan dataset yang telah dibersihkan dan algoritma *K-Nearest Neighbors* (KNN). Pendekatan ini diharapkan dapat menyederhanakan proses penentuan harga mobil bekas, mengatasi ketidakpastian yang seringkali dihadapi pembeli, dan meningkatkan kepercayaan mereka dalam mengambil keputusan pembelian [2]. Melalui pemanfaatan teknologi dan analisis data, penelitian ini tidak hanya memberikan solusi untuk industri otomotif dalam menyusun strategi pemasaran yang lebih efektif, tetapi juga memberikan panduan berharga bagi konsumen yang tengah mencari mobil bekas dengan nilai terbaik.

Dengan memahami faktor-faktor yang mempengaruhi harga mobil bekas, penelitian ini berpotensi membuka wawasan baru dalam industri otomotif. Hasil dari studi ini diharapkan dapat memberikan kontribusi positif dalam mendukung pertumbuhan pasar mobil bekas, memajukan teknologi prediksi harga, dan memberikan manfaat konkret bagi konsumen serta pelaku industri.

II. DASAR TEORI

A. *Pengenalan Metode K-Nearest Neighbors (KNN)*

K-Nearest Neighbour (KNN) adalah algoritma dalam bidang data mining yang menghitung jarak antara data yang akan diprediksi dengan data latih yang sudah ada, lalu mengambil mayoritas label tetangga terdekat [3]. *K-Nearest Neighbour* dikenal karena tingkat efisiensinya yang tinggi, dan sering memberikan tingkat akurasi yang tinggi, terutama dalam tugas pengklasifikasian. Tujuan utama dari algoritma ini adalah mengklasifikasikan objek baru dengan membandingkan atributnya dengan sampel pelatihan. Selain digunakan untuk analisis klasifikasi, K-NN juga telah diterapkan dalam beberapa dekade terakhir untuk keperluan prediksi. Cara kerja *K-Nearest Neighbor* (KNN) mendasarkan pada prinsip klasifikasi berdasarkan kedekatan spasial atau jarak antara data yang diuji dengan data pelatihan yang sudah ada [4]. Kemendekatan atau keterjauhan lokasi dapat diukur menggunakan salah satu metrik jarak yang telah ditetapkan, seperti jarak *Euclidean* atau jarak *Minkowski* [5]. Definisi KNN dapat ditulis sebagai persamaan (1).

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} \left(a_r(x_i) - a_r(x_j)\right)^2} \quad (1)$$

B. *Penerapan KNN dalam Prediksi Harga*

Sejumlah penelitian sebelumnya telah menggunakan metode KNN dalam konteks prediksi harga, terutama terkait mobil bekas. Penelitian-penelitian ini menunjukkan bahwa KNN dapat berhasil digunakan untuk memprediksi harga berdasarkan atribut seperti model, tahun pembuatan,

transmisi, jarak tempuh, dan jenis bahan bakar [6]. Penerapan KNN dalam prediksi harga memberikan pemahaman yang lebih baik tentang bagaimana faktor-faktor ini berkontribusi terhadap nilai jual mobil.

C. *RMSE (Root Mean Square Error)*

Kriteria yang digunakan untuk mengevaluasi kualitas model setelah membangun suatu model adalah *root mean square error* (RMSE). RMSE berfungsi sebagai indikator pemilihan model berdasarkan kesalahan hasil estimasi, yang mencerminkan seberapa besar perbedaan antara hasil estimasi dan nilai yang seharusnya diestimasi [7]. Skor ini akan digunakan untuk menilai model terbaik. Definisi RMSE dapat dirumuskan sebagai persamaan (2).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (2)$$

III. IMPLEMENTASI K-NEAREST NEIGHBOR

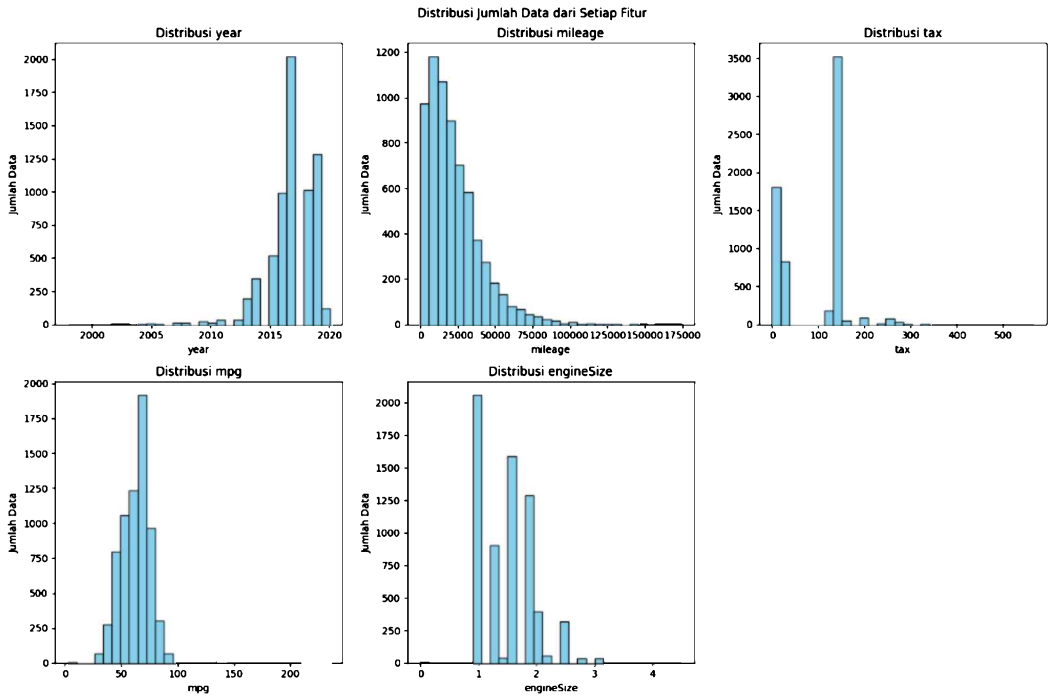
A. *Data*

Penelitian ini memanfaatkan data mobil bekas merek Toyota di *United Kingdom* dengan beberapa variabel termasuk model, tahun produksi, transmisi, jarak tempuh, jenis bahan bakar, pajak jalan raya, konsumsi bahan bakar per mil (mpg), dan ukuran mesin. Variabel ini mencakup aspek vital dari mobil bekas, seperti performa, efisiensi, dan biaya operasional. Data mobil bekas tersebut dapat dilihat pada tabel 1. Pada tabel 1 merupakan sebaran data setiap fitur yang ada dalam dataset. Fitur tersebut terdiri dari tahun, Jarak, pajak, mpg, dan ukuran mesin. Fitur atau variable tersebut digunakan untuk melatih model KNN untuk memprediksi harga mobil.

Pada gambar 2 merupakan sebaran data setiap fitur yang ada dalam dataset. Fitur tersebut terdiri dari tahun, Jarak, pajak, mpg, dan ukuran mesin. Fitur atau variable tersebut digunakan untuk melatih model KNN untuk memprediksi harga mobil.

TABEL 1. DATA

No.	model	year	transmission	mileage	fuelType	price	tax	mpg	engineSize
1	GT86	2016	Manual	24089	Petrol	16000	265	362	20
2	GT86	2017	Manual	18615	Petrol	15995	145	362	20
3	GT86	2015	Manual	27469	Petrol	13998	265	362	20
4	GT86	2017	Manual	14736	Petrol	18998	150	362	20
5	GT86	2017	Manual	36284	Petrol	17498	145	362	20
6	GT86	2017	Manual	26919	Petrol	15998	260	362	20
7	GT86	2017	Manual	10456	Petrol	18522	145	362	20
8	GT86	2017	Manual	12340	Petrol	18995	145	362	20
9	GT86	2020	Manual	516	Petrol	27998	150	332	20
10	GT86	2016	Manual	37999	Petrol	13990	265	362	20



Gambar 1. Sebaran Data Fitur

B. Pengolahan data

1. Pembersihan data

Pemrosesan data dilakukan untuk mengubah data yang akan digunakan sebagai data pengujian, sehingga data tidak lagi memiliki masalah di dalamnya. Ini berarti data akan diperiksa dan dibersihkan hingga tidak mengandung nilai yang hilang.

2. Pemilihan Fitur

Pemilihan fitur ini bertujuan untuk mengambil informasi penting dari data, seperti tahun pembuatan, jarak tempuh, pajak, efisiensi bahan bakar, dan kapasitas mesin mobil. Hal ini dilakukan agar dimensi data dapat direduksi dan fitur-fitur yang relevan dapat digunakan dalam pembelajaran mesin dengan algoritma *K-Nearest Neighbors* (KNN), meningkatkan akurasi prediksi harga mobil bekas.

3. Pembagian Dataset

Untuk memprediksi model yang dibuat, diperlukan dataset yang akan dibagi menjadi dua bagian, yaitu data latih dan data uji. Pembagian dilakukan dengan proporsi 80%-20%, di mana 80% data digunakan sebagai data latih, sementara 20% sisanya digunakan sebagai data uji. Penggunaan parameter *random state* dengan nilai 70 bertujuan untuk memberikan konsistensi pada pembagian data, sehingga memastikan hasil yang sama setiap kali proses pembagian dilakukan. Hal ini penting agar mendapatkan evaluasi model yang konsisten dan dapat diandalkan selama pengujian.

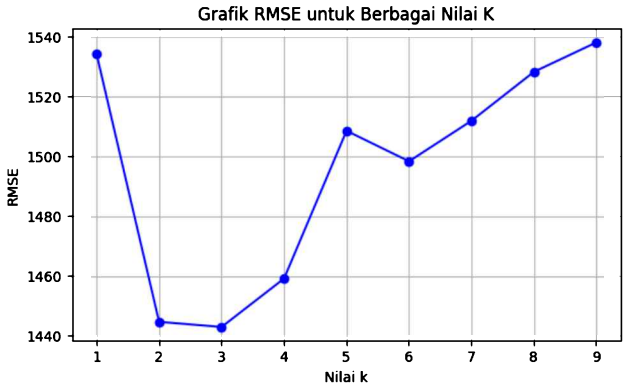
4. Normaliasi Data

Dalam proses persiapan data untuk analisis, normalisasi adalah langkah penting untuk memastikan bahwa semua fitur memiliki skala yang seragam. Pada penelitian ini, kami menggunakan metode normalisasi data dengan *StandardScaler* dari pustaka *scikit-learn*. Normalisasi dilakukan untuk mengubah distribusi nilai fitur sehingga memiliki rata-rata nol dan deviasi standar satu [8]. *StandardScaler* bekerja dengan mengurangkan rata-rata dari setiap nilai fitur dan kemudian membagi hasilnya dengan deviasi standar. Ini memastikan bahwa setiap fitur memiliki skala yang seragam, menghindari dominasi fitur tertentu dalam proses pembelajaran model. Proses ini memberikan keuntungan tambahan saat menggunakan algoritma yang sensitif terhadap skala, seperti *K-Nearest Neighbors* (KNN). Dengan normalisasi yang baik, model KNN dapat bekerja lebih efisien dan memberikan hasil prediksi yang lebih akurat, membantu kita memahami pengaruh masing-masing fitur terhadap harga mobil bekas dengan lebih baik.

5. Pengujian KNN

Pengujian *K-Nearest Neighbors* (KNN) dilakukan untuk mengevaluasi kinerja model yang telah dikembangkan dalam memprediksi harga mobil bekas. Berbagai metrik evaluasi digunakan untuk mengukur sejauh mana model mampu mengeneralisasi pola dari data latih ke data uji. Dalam konteks ini, fokus utama adalah pada penggunaan *Root Mean Squared Error* (RMSE) untuk mengukur tingkat deviasi antara harga mobil bekas yang diprediksi dan nilai aktualnya [9]. Pemilihan Jumlah Tetangga (k) Sebelumnya, eksperimen dilakukan untuk menentukan jumlah tetangga terdekat (k) yang optimal. Pengujian dilakukan dengan memvariasikan nilai k dan mengukur dampaknya terhadap RMSE [10]. Hasil pengujian ini membantu memilih nilai k yang memberikan keseimbangan terbaik antara ketepatan prediksi dan kecanggihan model. Evaluasi Keseluruhan Model Model KNN dievaluasi secara menyeluruh menggunakan data uji. Hasil prediksi dibandingkan dengan harga mobil bekas yang sebenarnya, dan RMSE dihitung sebagai indikator kinerja utama. Visualisasi juga digunakan untuk membandingkan sebaran harga aktual dan harga yang diprediksi, memberikan pemahaman yang lebih dalam tentang kecenderungan model. Analisis Pengaruh

Fitur Pengujian tidak hanya berfokus pada evaluasi keseluruhan model, tetapi juga pada analisis pengaruh masing-masing fitur terhadap prediksi harga. Visualisasi *scatter plot* digunakan untuk mengilustrasikan hubungan antara fitur-fitur seperti tahun produksi, jarak tempuh, pajak, efisiensi bahan bakar, dan ukuran mesin dengan harga mobil bekas. Analisis ini memberikan wawasan tentang sejauh mana masing-masing fitur memengaruhi keakuratan model dan memberikan informasi berharga untuk penyesuaian model di masa depan.



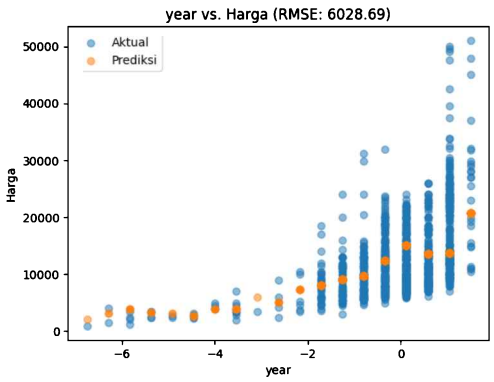
Gambar 2. Uji Coba KNN

Pada gambar 2 pengujian dilakukan dengan menerapkan beberapa nilai k mulai dari 1 sampai 9.

Hasil menunjukkan k=3 menjadi hasil terbaik dalam uji coba karena menghasilkan nilai rmse paling rendah yakni 1443,2065599.

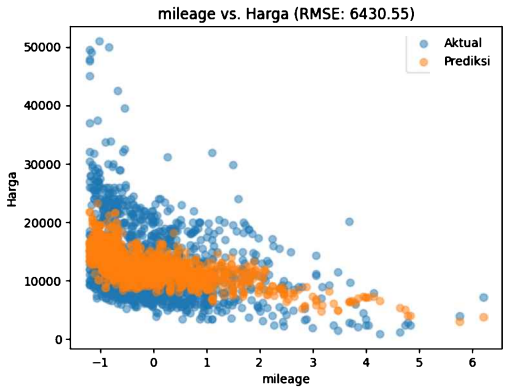
6. Hasil Pengujian

Pada tahap pengujian model *K-Nearest Neighbors* (KNN), hasil evaluasi menunjukkan performa model dalam memprediksi harga mobil bekas. Berdasarkan analisis RMSE untuk masing-masing fitur, diperoleh nilai sebagai berikut:



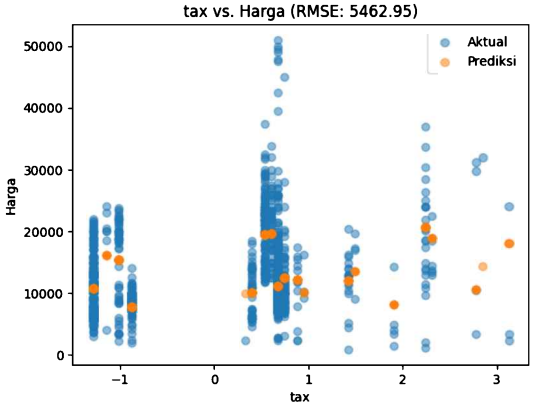
Gambar 3. RMSE Fitur Tahun

Pada gambar 3 merupakan Fitur *Year*: Model memiliki nilai RMSE yang relatif rendah untuk fitur *Year*, dengan nilai RMSE sebesar 6028.69. Ini mengindikasikan bahwa model cenderung dapat memprediksi harga mobil bekas berdasarkan tahun produksinya dengan tingkat akurasi yang memuaskan.



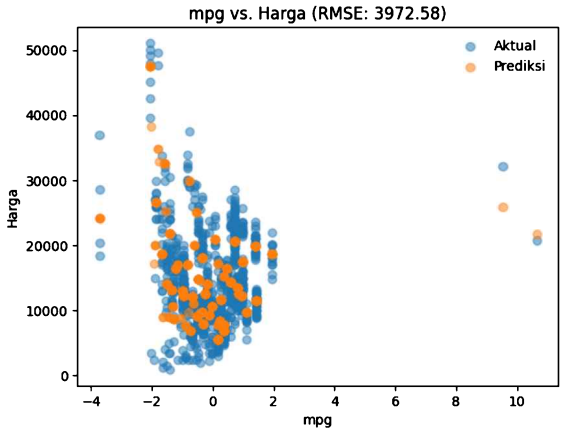
Gambar 4. RMSE Fitur Jarak Tempuh

Pada gambar 4, fitur *Mileage*: Pengujian menunjukkan bahwa prediksi harga berdasarkan jarak tempuh memiliki nilai RMSE sebesar 6430.55. Meskipun nilai ini cukup besar, model masih memberikan gambaran yang layak tentang pengaruh jarak tempuh terhadap harga mobil bekas.



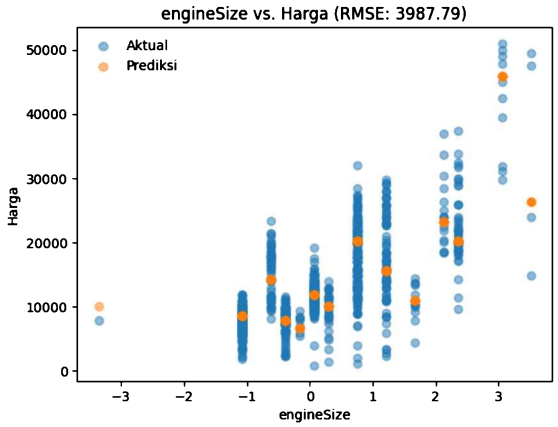
Gambar 5. RMSE Fitur Pajak

Pada gambar 5, fitur *Tax*: Dalam hal prediksi berdasarkan pajak, model menunjukkan nilai RMSE sebesar 5462.95. Meskipun demikian, model masih mampu memberikan estimasi harga yang relatif akurat berdasarkan besaran pajak mobil bekas.



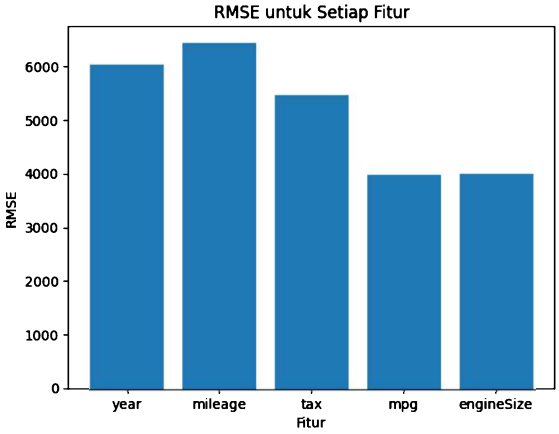
Gambar 6. RMSE Fitur mpg

Pada gambar 6, fitur *MPG* (Efisiensi Bahan Bakar): Evaluasi menunjukkan bahwa model memiliki nilai RMSE sebesar 3972.58 saat memprediksi harga berdasarkan efisiensi bahan bakar (*MPG*). Meskipun ada tingkat ketidakpastian, model tetap memberikan informasi berguna tentang pengaruh efisiensi bahan bakar. terhadap harga.



Gambar 7. RMSE Fitur Ukuran Mesin

Pada gambar 7, fitur *EngineSize*: Hasil pengujian menunjukkan nilai RMSE sebesar 3987.79 dalam memprediksi harga berdasarkan ukuran mesin. Meskipun ada variasi, model memberikan estimasi harga yang dapat diandalkan dengan mempertimbangkan fitur ini.



Gambar 8. Sebaran RMSE Fitur

Pada gambar 8, merupakan sebaran dari setiap fitur yang ada dalam dataset penelitian, pada gambar tersebut terlihat bahwa ukuran mesin memiliki nila RMSE terendah.

IV. KESIMPULAN

Kesimpulan dari penelitian ini menunjukkan bahwa model K-Nearest Neighbors (KNN) efektif dalam memprediksi harga mobil bekas, dengan fitur 'EngineSize' sebagai variabel yang paling signifikan dalam menentukan harga. Nilai Root Mean Squared Error (RMSE) yang rendah mengukuhkan keakuratan model, khususnya dengan RMSE sebesar 3987.79. Oleh karena itu, ukuran mesin mobil menjadi faktor utama yang mempengaruhi prediksi harga mobil bekas. Relevansi fitur *EngineSize* menyoroti pentingnya mempertimbangkan ukuran mesin dalam menilai nilai jual atau beli mobil bekas. Temuan ini memberikan dasar kuat untuk pengembangan strategi pemasaran dan keputusan investasi yang lebih cerdas di industri mobil bekas. Meskipun demikian, penelitian menekankan bahwa faktor-faktor lain juga memiliki peran, mendorong perlunya penelitian lebih lanjut. Saran penelitian meliputi eksplorasi faktor tambahan seperti merek, transmisi, dan kondisi kendaraan, serta analisis spesifik merek dan model. Penambahan data historis, eksplorasi model lainnya, teknik feature engineering, pengumpulan data yang lebih luas, dan evaluasi performa model pada rentang harga tertentu juga diusulkan sebagai langkah-langkah untuk meningkatkan pemahaman dan prediksi yang lebih akurat dalam konteks pasar mobil bekas.

REFERENSI

[1] S. Kosasi and S. Pontianak, "Perancangan Sistem Informasi Penjualan Berbasis Web Dalam Memasarkan Mobil Bekas The Information System Design of Web-Based Sales Second-hand Cars," *Citec Journal*, vol. 3, no. 1, 2015.

[2] D. Budilaksana, M. Sukarsa, A. Agung, K. Agung, and C. Wiranatha, "Implementing kNearest Neighbor Methods to Predict Car Prices."

[3] J. Homepage, S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, "IJCIT (Indonesian Journal on Computer and Information Technology) Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa," 2021.

[4] S. Sistem, P. Predikat, P. M. Mustakim, and G. Oktaviani, "Algoritma K-Nearest Neighbor Classification," *Jurnal Sains, Teknologi dan Industri*, vol. 13, no. 2, pp. 195–202, 2016, [Online]. Available: http://ejournal.uin-suska.ac.id/index.php/sitekin

[5] R. Bahtiar, "Implementasi Data Mining Untuk Prediksi Penjualan Kusen Terlaris Menggunakan Metode K-Nearest Neighbor." [Online]. Available: https://jurnal.publikasitecno.id/index.php/jim203

[6] K. Samruddhi and R. Ashok Kumar, "Used Car Price Prediction using K-Nearest Neighbor Based Model," *International Journal of Innovative Research in Applied Sciences and Engineering*, vol. 4, no. 2, pp. 629–632, Aug. 2020, doi: 10.29027/ijirase.v4.i2.2020.629-632.

[7] N. Aris, A. Nugroho, B. Sudarsono, and L. M. Sabri, "Analisis Kesesuaian Penggunaan Lahan Terhadap RTRW Menggunakan Sistem Informasi Geografis (Studi kasus : Kec.Pedurungan dan Kec.Tembalang,Kota Semarang)," 2021.

[8] N. Saputra, T. B. Adji, and A. E. Permanasari, "ANALISIS SENTIMEN DATA PRESIDEN JOKOWI DENGAN PREPROCESSING NORMALISASI DAN STEMMING MENGGUNAKAN METODE NAIVE BAYES DAN SVM Oleh," 2015.

[9] A. Budianto, R. Ariyuana, and D. Maryono, "PERBANDINGAN K-NEAREST NEIGHBOR (KNN) DAN SUPPORT VECTOR MACHINE (SVM) DALAM PENGENALAN KARAKTER PLAT KENDARAAN BERMOTOR," *Jurnal Ilmiah Pendidikan Teknik dan Kejuruan*, vol. 11, no. 1, p. 27, Nov. 2019, doi: 10.20961/jiptek.v11i1.18018.

[10] I. Colanus and R. Drajana, "Prediksi Jumlah Produksi Coconut Oil Menggunakan k-Nearest Neighbor dan Backward Elimination."

Penerapan Metode *Ensemble Learning Hard Voting* dalam Klasifikasi *Credit Card Fraud*

Shabrina Aurelia
Departemen Sains Data
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
shaaurel1107@gmail.com

Frisca Damayanti
Departemen Sains Data
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
frisca23@gmail.com

Maharani Yulianti
Departemen Sains Data
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
maharaniyt@gmail.com

Abstrak--Pada era digital, kini penggunaan kartu kredit sebagai alat pembayaran elektronik telah menjadi hal umum dalam transaksi keuangan sehari-hari. Namun, penggunaan kartu kredit juga membawa resiko tinggi terhadap penipuan. Salah satu pendekatan yang paling efektif dalam deteksi penipuan kartu kredit adalah menggunakan metode *Ensemble Learning* dengan menggabungkan hasil dari beberapa model pembelajaran mesin (*machine learning*) untuk mencapai kinerja yang lebih baik daripada yang dicapai oleh model tunggal. Penelitian ini menggunakan tiga model Machine Learning, yaitu *Decision Tree*, *Naïve Bayes,* dan *K-Nearest Neighbor* yang menghasilkan akurasi masing-masing 99%,95%, dan 99%. Kemudian ketiga model tersebut digabungkan dengan metode *hard voting* untuk memprediksi penipuan dengan akurasi pada data uji sebesar 98,54% dan akuirasi data latih sebesar 99,21%. Berdasarkan hasil akurasi tersebut, dapat diartikan bahwa *ensemble machine learning* dengan tiga gabungan model tersebut dapat melakukan kinerja yang baik pada klasifikasi penipuan kartu kredit.

Kata Kunci: Penipuan kartu kredit, Klasifikasi, *Ensemble Learning*.

1. PENDAHULUAN

Pada era digital yang semakin berkembang, penggunaan kartu kredit sebagai salah satu bentuk alat pembayaran elektronik telah menjadi hal yang umum dalam transaksi keuangan sehari-hari. Meskipun memberikan kenyamanan yang tak terbantahkan dalam berbelanja dan membayar tagihan, penggunaan kartu kredit juga membawa resiko tinggi terhadap penipuan dan aktivitas criminal terkait transaksi kartu kredit. Oleh karena itu, penting untuk mengembangkan metode yang efektif dalam mendeteksi dan mencegah penipuan kartu kredit.

Penelitian dalam deteksi *fraud* kartu kredit telah menjadi subjek utama perhatian dalam dunia keamanan finansial. Penipuan kartu kredit dapat merugikan bagi pihak perusahaan penerbit kartu kredit maupun pemegang kartu, oleh karena itu, perlu adanya metode yang canggih dan efisien untuk mengidentifikasi aktivitas penipuan tersebut.

Salah satu pendekatan yang paling efektif dalam deteksi *fraud* kartu kredit adalah menggunakan metode *Ensemble Learning*. *Ensemble Learning* adalah teknik yang menggabungkan hasil dari beberapa model pembelajaran mesin (*machine learning*) untuk mencapai kinerja yang lebih baik daripada yang dicapai oleh model tunggal. Pendekatan ini telah terbukti efektif dalam berbagai masalah klasifikasi, termasuk deteksi *fraud* kartu kredit.

Artikel ini bertujuan untuk menyelidiki dan mengevaluasi metode *Ensemble Learning* dalam konteks deteksi *fraud* kartu kredit. Peneliti akan menjelaskan konsep dasar *Ensemble Learning*, jenis-jenis *Ensemble Learning* yang umum digunakan, dan bagaimana metode ini dapat diterapkan dalam pengembangan sistem deteksi *fraud* yang andal.

Selain itu, kami akan menggunakan dataset transaksi kartu kredit yang telah ada untuk menguji dan membandingkan kinerja berbagai model *Ensemble Learning*. Hasil dari penelitian ini diharapkan dapat memberikan panduan yang berguna bagi lembaga keuangan dan perusahaan penerbit kartu kredit dalam mengembangkan sistem deteksi *fraud* yang lebih baik dan efisien.

Artikel ini disusun dalam struktur yang terorganisir, dimulai dengan tinjauan literatur tentang *fraud* kartu kredit dan *Ensemble Learning*, diikuti oleh penjelasan metodologi yang digunakan, hasil eksperimen, dan diskusi. Akhirnya, peneliti akan menyimpulkan temuan penelitian dan memberikan rekomendasi untuk penelitian mendatang.

Deteksi penipuan kartu kredit adalah tantangan penting dalam industri keuangan yang memerlukan upaya terus-menerus untuk mengembangkan metode yang lebih canggih. Dalam beberapa tahun terakhir, pendekatan menggunakan *Ensemble Learning* telah menjadi fokus utama penelitian dalam upaya peningkatan akurasi dalam mendeteksi aktivitas penipuan. Dalam bagian ini kami akan mengulas beberapa penelitian terkait yang telah mengaplikasikan metode *Ensemble Learning* dalam deteksi penipuan kartu kredit.

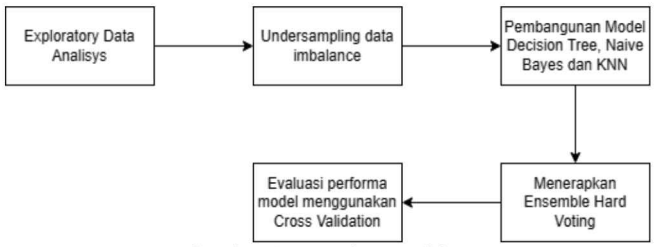
Penelitan pertama yang dilakukan oleh [1] melakukan penelitian terkait klasifikasi *credit approval.* Penelitian tersebut menggunakan algoritma *random forest.* Penelitian lainnya [2] melakukan klasifikasi penipuan rekening bank dengan menggunakan pendekatan *Ensemble Learning*. Penelitian tersebut bertujuan untuk membandingkan algoritma *extreme gradiant boosting (xgoost)* dan *random forest*. Hasilnya algoritma *xgboost* menghasilkan nilai akurasi yang lebih baik dari *random forest*. Pada penelitian lainnya, [3] melakukan klasifikasi penipuan kartu kredit dengan algoritma *random forest* menunjukkan hasil akurasi yang sangat baik.

Dataset *credit card fraud* sering ditemukan memiliki ketidakseimbangan kelas pada dataset tersebut. [1] menyatakan keberadaan distribusi kelas yang tidak seimbang dapat memengaruhi performa dari suatu algoritma klasifikasi, karena suatu algoritma klasifikasi bekerja dengan mengansumsikan distribusi kelas pada dataset relatif seimbang dan biaya kesalahan klasifikasi yang sama. Hal tersebut juga pastinya dapat menyebabkan resiko terjadinya kesalahan klasifikasi terhadap dataset, sehingga kinerja suatu algoritma menjadi tidak maksimal [1]. Maka diperlukan suatu

metode untuk menyelesaikan permasalahan distribusi kelas yang tidak seimbang pada suatu dataset. Untuk menangani distribusi kelas yang tidak seimbang, terdapat metode yang dapat digunakan, yaitu pendekatan pada level data dan pendekatan pada level algortimik. Pendekatan level data digunakan untuk memperbaiki kecondongan distibusi kelas pada dataset. Dalam pendekatan level data terdapat dua metode yang sering dipakai yaitu teknik *resampling* maupun teknik sintetis data. Pendekatan level algoritmik bekerja dengan menyesuaikan operasi algoritma yang ada untuk membuat suatu *classifier* lebih kondusif terhadap klasifikasi kelas minoritas atau memodifikasi maupun penggabungan (*ensemble*) dari beberapa algoritma. Pada penelitian ini dimaksudkan untuk mengatasi permasalahan ketidakseimbangan kelas pada dataset agar proses klasifikasi lebih optimal.

2. METODE PENELITIAN

Penelitian ini dilakukan melalui beberapa tahap, yaitu (1) *Exploratory Data Analysis*, (2) *Undersampling*, (3) Model Pembangunan, (4) *Ensemble Learning*, dan (5) *Cross-Validation*. Ilustrasi tahap penelitian terdapat pada gambar 1 berikut ini.



Gambar 1. Metode Penelitian

A. *Exploratory Data Analysis (EDA)*

Exploratory Data Analysis (EDA) adalah tahap awal yang dilakukan oleh peneliti dengan tujuan memahami dataset yang digunakan. EDA memberikan pemahaman tentang atribut, karakteristik, dan pola dalam dataset yang sangat penting dalam penanganan penipuan kartu kredit. Pada tahap EDA, peneliti mulai menggali lebih dalam dataset terkait *Credit Card Fraud* sehingga diketahui bahwa dataset yang digunakan memiliki 1.000.000 *instances*, 7 *attribute* fitur dan 1 *attribute class*. Selain itu, data tersebut tidak memiliki *missing value* dengan artian tidak ada data yang kosong. Distribusi kelas pada data yang digunakan untuk kelas 0 terdapat 912597 *intances* dan kelas 1 terdapat 87403 *intances*, maka dapat dikategorikan bahwa kelasnya tidak seimbang. Ketidakseimbangan ini dapat mempengaruhi kinerja model *Machine Learning*, sehingga memerlukan strategi penanganan yang tepat.

B. *Undersampling*

Analisis EDA menunjukkan adanya ketidakseimbangan yang signifikan antara jumlah transaksi penipuan dan non-penipuan. Untuk mengatasi masalah ini, peneliti menerapkan teknik *undersampling* untuk mengurangi jumlah sampel dari transaksi non-penipuan sehingga setiap kelas memiliki jumlah yang sama. Hal ini membantu mencegah model cenderung memprediksi kelas mayoritas dan memperkuat kemampuan model dalam mendeteksi penipuan.

C. *Pembangunan Model*

Setelah penanganan ketidakseimbangan dataset, peneliti membagi dataset menjadi dua bagian yaitu data latih dan data uji. Peneliti juga melakukan proses *tuning* untuk menemukan parameter terbaik pada setiap model *Machine Learning* dan melakukan pengujian kinerja model dengan melihat akurasi pada semua model klasifikasi untuk diambil 3 akurasi tertinggi. adapun model *Machine Learning* yang peneliti gunakan adalah sebagai berikut:

a. *Decision Tree*

Metode Decision Tree merupakan suatu model yang dapat memprediksi kategori data dengan cara mempelajari aturan penentuan kategori berdasarkan fitur-fitur yang dimiliki oleh data (Ceballos, Ochiai, Masuma, & Tomii, 2019). Berdasarkan tipe kategori datanya, *decision tree* dibedakan menjadi dua jenis yaitu *classification tree* dan *regression tree*. *Classification tree* memiliki kategori berupa data diskrit berhingga, sedangkan *regression tree* memiliki kategori berupa data diskrit berhingga atau data kontinu [4].

b. *Naïve Bayes*

Metode Naïve Bayes merupakan metode yang memanfaatkan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes. *Naïve Bayes* merupakan metode pengklasifikasian yang sangat sederhana dengan mengasumsikan klasifikasi atribut. Dengan metode *Naive Bayes* terlebih dahulu mencari nilai probabilitas dan *likelihood* maksimum dari setiap atribut untuk masing-masing kelas [5].

c. *K-Nearest Neighbor*

K-Nearest Neighbors (K-NN) adalah suatu metode yang menggunakan algoritma supervised dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada K-NN. Tujuan dari algoritma ini adalah mengklasifikasi objek baru berdasakan atribut dan sampel latih [6].

D. *Ensemble Learning*

Setelah melatih beberapa model, peneliti menggabungkan hasil dari ketiga model dengan menggunakan *ensemble learning*. Teknik ini digunakan untuk mengkombinasikan hasil dari beberapa model berbeda untuk memperoleh prediksi yang lebih akurat.

Selanjutnya, peneliti menggunakan teknik *Voting Classifier* untuk mengumpulkan prediksi dari semua model dan memutuskan hasil akhir berdasarkan mayoritas suara. Dengan kata lain, jika mayoritas model memprediksi transaksi sebagai penipuan, maka itu akan dianggap sebagai penipuan.

E. *Cross-Validation*

Pada tahap terakhir, peneliti menggunakan prosedur *cross-validation* untuk mengukur kinerja model secara lebih objektif. Adapun prosedur *cross-validation* yang digunakan oleh peneliti yaitu *K-Fold Cross Validation* dengan membagi *dataset* menjadi beberapa "lipatan" *(fold)* dan menjalankan proses pelatihan dan pengujian beberapa kali dengan menggunakan *fold* yang berbeda sebagai data uji setiap kali. Ini membantu kami

menghindari *overfitting* dan memastikan kinerja model yang konsisten.

Proses ini adalah langkah kunci dalam evaluasi model karena mengukur sejauh mana model dapat bekerja dengan baik pada data yang belum pernah dilihat sebelumnya.

Semua langkah dalam metode penelitian ini berkontribusi pada upaya mendeteksi penipuan kartu kredit dengan akurasi dan keandalan yang tinggi sehingga membuat langkah-langkah dalam penelitian ini penting untuk mengatasi masalah keamanan transaksi keuangan.

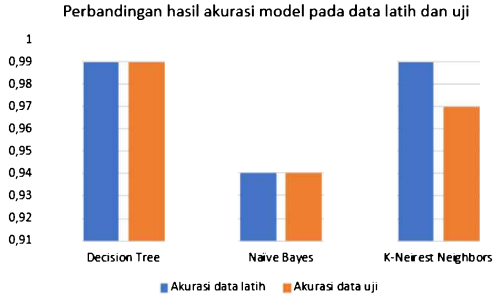
3. HASIL PENELITIAN

Metriks evaluasi kinerja yang digunakan untuk mengukur tingkat keberhasilan pada model machine learning yang digunakan dengan melihat besaran akurasi pada data latih dan data uji. Model Machine Learning ini dibangun menggunakan proses tuning, hal ini dilakukan supaya model yang digunakan dibangun menggunakan parameter terbaik sehingga kinerja model yang dihasilkan juga baik, tabel 1 merupakan penjelasan parameter yang digunakan pada masing-masing model, terdapat *decision tree*, *naïve bayes*, dan *k-nearest neighbors*, model ini dipilih berdasarkan tingkat akurasi tertinggi dalam klasifikasi data card fraud.

TABEL 1. PARAMETER UNTUK MASING-MASING MODEL

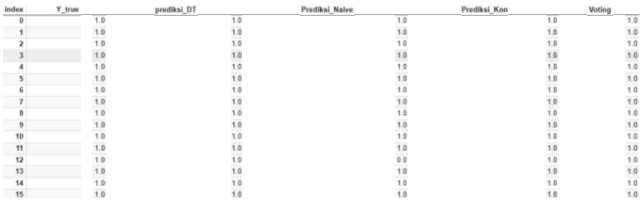
Model	Parameter
Decision Tree	criterion': 'entropy', 'max_depth': None, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'best'
Naïve Bayes	var_smoothing': 1e-09
K-Nearest Neighbors	metric': 'manhattan', 'n_neighbors': 3

Gambar 2 merupakan visualisasi hasil akurasi dimana Model *Decision Tree* mendapat akurasi 99% untuk data latih, 95% untuk *Naïve Bayes*, dan 99% untuk *K-Nearest Neighbor*, sedangkan pada data uji, hasil akurasi yang didapat sebesar 99% untuk *Decision Tree*, 94% unutk *Naïve Bayes* dan 97% untuk *K-Nearest Neighbors* hasil akurasi pada data latih lebih tinggi atau bahkan sama dengan data uji, hal ini disebabkan karena ukuran set latih lebih besar dibandingkan dengan set uji, besaran akurasi tersebut cukup baik untuk dilakukan proses *Ensemble Learning*.



Gambar 2. Perbandingan Hasil Akurasi Model Pada Data Latih dan Data Uji

Selanjutnya adalah proses *Ensemble* dengan menggabungkan ketiga model tersebut menggunakan metode hard voting. metode hard voting bekerja dengan mengambil hasil prediksi terbanyak dalam beragam model yang digunakan tanpa melihat bobot dari masing-masing hasil prediksi. Berikut merupakan tabel dari hasil prediksi ketiga model dan prediksi menggunakan *ensemble hard voting*, dengan nilai pada kolom Y_true adalah data target, prediksi_DT adalah data hasil prediksi menggunakan model *Decision Tree*, Prediksi_Naive adalah hasil prediksi menggunakan metode *Naïve Bayes*, prediksi_Knn merupaan hasil dari prediksi menggunakan metode *K-Nearest Neighbors* dan voting merupakan hasil voting dengan metode *hard voting*. Hasil dari prediksi menggunakan hard voting terdapat pada gambar 3 berikut ini.



Gambar 3. Prediksi Menggunakan Ensemble Hard Voting

Untuk mengukur kinerja pada model *ensemble*, dilakukan pelatihan dan pengujian pada dataset *card fraud* menggunakan *cross-validation* dengan nilai k = 10, dengan metrik kinerja berupa *fit time* untuk mengetahui waktu yang dibutuhkan model untuk dilatih dimana semakin rendah nilai *fit time*, maka efektivitas waktu pelatihan semakin baik dimana rata-rata dari waktu yang dibutuhkan ke sepuluh *fold* adalah 0,656 atau 65%. Dan akurasi pada data uji serta data latih pada masing-masing lipatan, jika semakin tinggi akurasi model maka kinerja model pada dataset tersebut semakin baik. Didapatkan rata-rata rata rata akurasi test sebesar 98,54% dan akurasi pada set latih sebesar 99,21%.

Tabel 2. Hasil Pelatihan dan Pengujian Menggunakan Cross-Validation

Fold	fit_time	test_accuracy	train_accuracy
1	0,724579	0,972828	0,991832
2	0,651261	0,971798	0,991851
3	0,64723	0,970883	0,991775
4	0,653878	0,969967	0,991718
5	0,667665	0,97134	0,991705
6	0,651525	0,999371	0,992411
7	0,636059	0,999714	0,992449
8	0,635189	0,999714	0,992512
9	0,635189	0,999371	0,99243
10	0,658295	0,999485	0,992392

KESIMPULAN

Penipuan kartu kredit adalah betuk pecurian identitas yang paling umum, hal ini terjadi karena seseorang mengambil informasi pribadi untuk melakukan pembayaran. *Credit card fraud detection* ini memiliki 1.000.000 data dengan 95 atribut dengan 1 atribut target. Pendekatan prediksi ini disesuaikan dengan gabungan antara model *Naïve Bayes*, *Decision Tree*, dan *K-Nearest Neighbors* dengan teknik *undersampling* mengunakan *cross-validation* dengan nilai *k-fold* 10 untuk membagi dataset menjadi set latih dan uji untuk menghasilkan model yang kuat dan dapat diandalkan peneliti juga menggunakan proses tuning supaya menemukan parameter terbaik untuk algoritma yang digunakan sehingga ketiga model tersebut layak untuk digabung menjadi model *ensemble*. Di mana model *ensemble* dapat bekerja dengan baik untuk memprediksi hasil voting pada penipuan kartu kredit, hal ini dilihat berdasarkan hasil akurasi yang didapatkan sebesar 98,54% pada data latih dan 99,21% pada data uji.

REFERENSI

[1] N. Widjiyati, "Implementasi Algoritme Random Forest Pada Klasifikasi Dataset Credit Approval," *J. Janitra Inform. dan Sist. Inf.*, vol. 1, no. 1, pp. 1–7, 2021, doi: 10.25008/janitra.v1i1.118.

[2] A. Maghfiroh, Y. Findawati, and U. Indahyanti, "Klasifikasi Penipuan pada Rekening Bank menggunakan Pendekatan Ensemble Learning," *Build. Informatics, Technol. Sci.*, vol. 4, no. 4, pp. 1883–1891, 2023, doi: 10.47065/bits.v4i4.3212.

[3] T. S. Lestari and D. A. N. Sirodj, "Klasifikasi Penipuan Transaksi Kartu Kredit Menggunakan Metode Random Forest," *J. Ris. Stat.*, vol. 1, no. 2, pp. 160–167, 2022, doi: 10.29313/jrs.v1i2.525.

[4] R. Latifah, E. S. Wulandari, and P. E. Kreshna, "Model Decision Tree Untuk Prediksi Jadwal Kerja Menggunakan Scikit-Learn," *J. Univ. Muhammadiyah Jakarta*, pp. 1–6, 2019.

[5] R. Y. Hayuningtyas, "Penerapan Algoritma Naïve Bayes untuk Rekomendasi Pakaian Wanita," *J. Inform.*, vol. 6, no. 1, pp. 18–22, 2019, doi: 10.31294/ji.v6i1.4685.

[6] Y. Yahya and W. Puspita Hidayanti, "Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Efektivitas Penjualan Vape (Rokok Elektrik) pada 'Lombok Vape On,'" *Infotek J. Inform. dan Teknol.*, vol. 3, no. 2, pp. 104–114, 2020, doi: 10.29408/jit.v3i2.2279.

Analisis Pengaruh Spesifikasi Terhadap Harga Handphone menggunakan Algoritma KNN dan Linear Regresi

Yuana Inka Dewi Br Sinulingga
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
inkasnlg54321@gmail.com

Arieska Restu Harpian Dwika
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
arieskarestu02@gmail.com

Tatas Handharu Sworo
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
tatashandharu2002@gmail.com

Putri Marceliana Aryanto
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
putrifikri017@gmail.com

Amalia Rizki Wulandari
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
amaliarizkii1803@gmail.com

Alexander Romian Simarmata
Departemen Informatika
Universitas Teknologi Yogyakarta
Yogyakarta, Indonesia
romian37@gmail.com

Abstrak— Ponsel telah menjadi bagian integral dari kehidupan sehari-hari, dengan beragam fitur dan spesifikasi yang mempengaruhi pilihan konsumen dan harga akhir. Memahami faktor-faktor yang mepengaruhi harga ponsel adalah kunci dalam membuat keputusan pembelian yang tepat. Penelitian ini memperkenalkan sebuah sistem prediksi harga ponsel menggunakan metode KNN dan Regresi Linear dengan tujuan menganalisis pengaruh spesifikasi terhadap harga. Hasil analisis menunjukkan bahwa KNN memberikan prediksi yang lebih akurat, terutama pada rasio pembagian data 7:3, menghasilkan nilai RMSE sebesar 0,250 dan MAE sebesar 0,124. Temuan ini mengindikasikan bahwa pemilihan atribut yang cermat, seperti RAM, secara signifikan meningkatkan kemampuan prediktif model KNN. Penelitian ini memberikan wawasan penting bagi konsumen dan produsen terkait elemen-elemen yang paling berpengaruh terhadap harga ponsel.

Kata Kunci—knn, linear regresi, prediksi, spesifikasi, harga

I. PENDAHULUAN

Ponsel atau smartphone merupakan suatu alat komunikasi yang digunakan dalam kehidupan sehari-hari pada zaman sekarang. Melalui ponsel kita dapat mengakses suatu informasi dengan sangat mudah, ditambah dengan semakin majunya teknologi dapat memberikan banyak fitur yang bermanfaat [1]. Ponsel juga merupakan alat komunikasi jarak jauh yang fiturnya dapat digunakan seperti melakukan chat, voice call, dan video call. Selain itu, ponsel dapat digunakan sebagai sarana hiburan serta berperan sebagai media layanan informasi [2]. Seperti yang kita ketahui di era sekarang ini semakin maju teknologi maka perkembangan ponsel semakin beragam, mulai dari fitur, spesifikasi maupun penampilan. Karena tidak semua ponsel memiliki kualitas yang dapat mendukung kebutuhan konsumen, maka sebelum membeli ponsel konsumen harus mengetahui detail spesifikasi, fitur maupun penampilan yang dapat menunjang aktivitas harian konsumen [3]. Spesifikasi headphone mencakup berbagai aspek seperti kuaitas kamera, ketebalan pnsel, ukuran RAM penyimpanan. Selain itu masih banyak aspek lain yang mempengaruhi harga sebuah handphone termasuk kualitas suara, tipe driver, desain fisik, kenyamanan pemakaian, serta fitur tambahan seperti noise cancellation, konektivitas nirkabel, dan lainnya. Memahami spesifikasi ini menjadi kunci dalam memastikan bahwa headphone yang dibeli sesuai dengan kebutuhan pengguna, mengingat perbedaan dalam preferensi dan tujuan penggunaan. Di samping itu, harga ponsel juga dapat mempengaruhi spesifikasi, fitur maupun penampilan dari sebuah ponsel tersebut. Tidak jarang juga harga dijadikan sebagai patokan dalam pembelian maupun penjualan sebuah ponsel yang mana setiap pelanggan pasti akan memikirkan spesifikasi ponsel sebelum melakukan pembelian [4]. Sehingga untuk mengatasi permasalahan tersebut diperlukan sebuah sistem prediksi harga handphone yang mampu membantu konsumen dalam pengambilan keputusan pada saat ingin membeli handphone sesuai dengan keinginan dan anggaran yang telah kita sediakan sebelumnya.

Ada banyak metode yang dapat dilakukan untuk melakukan prediksi terhadap suatu harga. Salah satu teknik yang dapat digunakan adalah teknik data mining. Namun tidak semua algoritma data mining memiliki performa yang baik dalam melakukan prediksi harga handphone. Oleh karena itu, pada penelitian ini akan dibahas mengenai analisis perbandingan tingkat performa algoritma prediksi K-Nearest Neighbors (KNN), Linear Regresi. Analisis perbandingan yang dimaksud adalah perbandingan tingkat akurasi yang dihitung menggunakan nilai Root Mean Squred Error (RMSE) dan Mean Absolute Error (MAE) dari kEDUA algoritma tersebut.

KNN merupakan metode yang menggunakan algoritma terawasi, dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas kategori pada k-NN yang bertujuan untuk mengklasifikasikan objek baru berdasarkan atribut dan sampel latih. Arhami & Nasir (2020) mengatakan bahwa K-Nearest Neighbor (KNN) merupakan salah satu metode prediksi untuk menentukan label (class) dari suatu objek baru berdasarkan class yang mayoritas dari K-Neighbor dalam sekumpulan data latih [5]. KNN berbasis kepada kesamaan yang dimiliki objek didasarkan pada jarak antara objek yang akan ditentukan dengan objek yang telah ada sebelumnya.Menurut penelitian yang dilakukan oleh (Utari et

al., 2020). KNN mampu melakukan prediksi dengan baik. Pengujian dari algoritma KNN dengan memberikan titik latih. Pada sistem prediksi akan ditemukan sejumlah K objek yang terdekat dengan titik uji. Support Vector Regression (SVR) merupakan sebuah pengembangan metode dari Support Vector Machine (SVM) yang berisi tambahan atribut regresi yang dimana akan memberikan kalkulasi prediksi seperti penggambaran sebuah statistik [6]. Linear regresi merupakan metode statistik yang digunakan untuk memodelkan dan mengukur hubungan antara variabel independen seperti harga & promosi, dan dependen seperti penjualan [7].

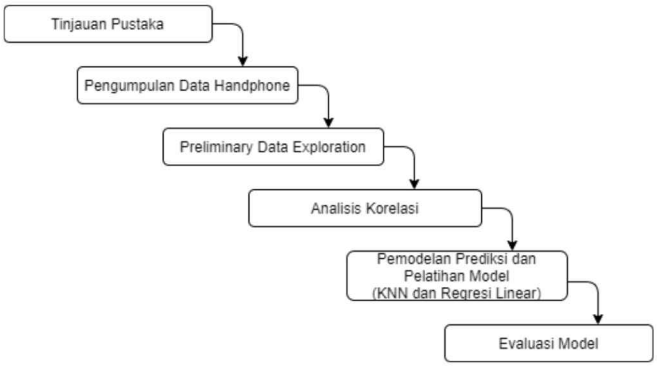
Penelitian mengenai prediksi harga telah dilakukan menggunakan metode *K-Nearest Neighbor dan Regresi Linear* untuk melakukan prediksi harga emas [8]. Dimana hasil dalam penelitian ini adalah dengan menambahkan metode Regresi Linear pada metode K-NN dapat meningkatkan tingkat prediksi dari algoritma tersebut, hal ini ditunjukkan dengan semakin kecilnya nilai RMSE yang diperoleh, adapun nilai RMSE yang diperoleh adalah sebesar 5,807%. Selain itu, penelitian mengenai prediksi harga dengan menggunakan metode KNN berbasis Forward Selection untuk melakukan prediksi harga komiditi lada [9]. Hasil penelitian ini menunjukkan bahwa algoritma K-Nearest Neighbor berbasis forward selection memberikan kinerja yang terbaik dibandingkan dengan KNN berbasis backward elimination dan SVM berbasis seleksi atribut dengan nilai RMSE Lada Hitam sebesar 1559,741 dan Lada Putih sebesar 6328,376. Selain itu, penelitian yang menggunakan metode regresi linear untuk melakukan prediksi penjualan telah dilakukan. Dimana dengan menggunakan algoritma regresi linear mendapatkan nilai Root Mean Squared Error (RMSE):36241.241 +/- 0.000 dan Squared Error: 1313427569.481 +/- 5882150128.134 [10]. Penelitian serupa mengenai prediksi penjualan dengan metode Regresi Linear mampu menunjukkan hasil nilai MAPE terkecil sebesar 1% pada produk Sunsilk Conditioner,dan nilai MAPE terbesar pada produk Vixal sebesar 10% [11].

Berdasarkan permasalahan yang telah disebutkan dan penelitian-penelitian sebelumnya, maka penulis tertarik untuk melakukan penelitian yang berjudul "Analisis Pengaruh Spesifikasi Terhadap Harga ponsel Menggunakan Algoritma KNN dan Linear Regresi". Dalam penelitian ini, peneliti akan membuat sebuah model dari masing masing algoritma yang bertujuan untuk memprediksi harga ponsel berdasarkan spesifikasi teknis yang dimiliki oleh setiap ponsel. Tujuan utama dari penelitian ini adalah untuk mengetahui fitur apa saja yang paling mempengaruhi harga ponsel dan membandingkan kinerja kedua algoritma tersebut dalam melakukan prediksi harga ponsel berdasarkan spesifikasi ponsel dengan keseluruhan atribut atau dengan seleksi atribut dengan tujuan untuk menemukan algoritma yang dapat memberikan prediksi secara akurat.

II. METODOLOGI PENELITIAN

A. *Tahapan Penelitian*

Sebelum memulai sistem penelitian, langkah-langkah yang telah direncanakan sebelumnya menjadi sangat penting. Tahapan penelitian ini menggunakan diagram alur untuk memudahkan penjelasan mengenai langkah-langkah yang dijalankan. Gambar 1 menampilkan rangkaian langkah yang digunakan dalam Analisis Pengaruh Spesifikasi Terhadap Harga ponsel Menggunakan Algoritma KNN dan Linear Regresi.



Gambar 1. Tahapan Penelitian

1) Tinjauan Pustaka
Pada tahap ini, lakukan tinjauan pustaka untuk memahami konsep-konsep terkait dalam analisis harga handphone. Tinjauan pustaka ini dapat memberikan pemahaman mendalam tentang faktor-faktor yang telah diidentifikasi dalam penelitian sebelumnya dan mungkin mempengaruhi harga handphone

2) Pengumpulan Data
Setelah melakukan tinjauan pustaka, tahap selanjutnya ialah mengumpulkan dataset yang mencakup data harga handphone dan variabel-variabel lain yang ingin Anda korelasikan dan prediksi

3) Eksplorasi awal data
Eksplorasi data awal untuk memahami distribusi, outliers, dan karakteristik dasar dari setiap variabel. Jika sudah selesai melakukan eksplorasi data pada tahap awal langkah selanjutnya adalah melakukan data preparation. Data preparation adalah suatu proses penyiapan data bersih untuk diolah dalam penelitian [12]. Data preparation adalah sebutan lain untuk data pre-processing. Preprocessing adalah sebuah langkah penting dalam proses penambangan data. Data yang akan digunakan dalam proses penambangan data tidak selalu dalam kondisi terbaik untuk diproses [13].

4) Seleksi Atribut
Seleksi atribut dilakukan dengan mengalisis hubungan korelasi antar variabel terutama menghitung menghitung hubungan antarvariabel dengan harga handphone, dan mengidentifikasi korelasi yang signifikan.

5) Pemodelan Prediksi dan Pelatihan Model
Selanjutnya, pada tahap pemodelan prediksi, pilih model prediksi berdasarkan karakteristik data dan tujuan penelitian, dan bagi dataset menjadi data pelatihan dan data uji. Kemudian lakukan pelatihan terhadpa modelnya. Adapun model yang digunakan adalah model KNN dan Regresi Linear

6) Evaluasi
Pada tahap evaluasi ini, dilakukan analisis hasil prediksi hasil dari kedua metode yang digunakan. Tahap ini diperlukan untuk menentukan algoritma mana yang paling baik dalam penelitian [14]. Evalusi dilakukan dengan mengukur kinerja model pada data uji menggunakan metrik seperti *Root Mean Squared Error* (RMSE) atau *Mean Absolute Error* (MAE).

B. Data yang digunakan

Data yang akan digunakan bersumber Kaggle, yang terdiri dari dua bagian: data latih (*train*) dan data uji (testing), masing-masing terdiri dari 2000 baris data dan 1000 baris data. Data ini terdiri dari 21 kolom atau atribut. Tabel 1 merupakan contoh data train yang akan digunakan dalam melakukan sistem prediksi harga handphone.

TABEL 1. CONTOH TABEL ATRIBUT YANG DIGUNAKAN

Komponen	Nilai				
battery_power	842	1021	563	615	1821
blue	0	1	1	1	1
clock_speed	2,2	0,5	0,5	2,5	1,2
dual_sim	0	1	1	0	0
fc	1	0	2	0	13
four_g	0	1	1	0	1
int_memory	7	53	41	10	44
m_dep	0,6	0,7	0,9	0,8	0,6
mobile_wt	188	136	145	131	141
n_cores	2	3	5	6	2
pc	2	6	6	9	14
px_height	20	905	1263	1216	1208
px_width	756	1988	1716	1786	1212
ram	2549	2631	2603	2769	1411
sc_h	9	17	11	16	8
sc_w	7	3	2	8	2
talk_time	19	7	9	11	15
three_g	0	1	1	1	1
touch_screen	0	1	1	0	1
wifi	1	0	0	0	0
price_range	1	2	2	2	1

Keterangan:

- ID : Pengenal
- battery_power : Total energi yang dapat disimpan baterai (mAh)
- blue : Memiliki bluetooth atau tidak
- clock_speed : Kecepatan mikroprosesor mengeksekusi instruksi
- dual_sim : Memiliki dukungan dual sim atau tidak
- fc : Mega piksel Kamera Depan
- four_g : Memiliki 4G atau tidak
- int_memory : Memori Internal dalam Gigabyte
- m_dep : Kedalaman Seluler dalam cm
- mobile_wt : Berat ponsel
- n_cores : Jumlah inti prosesor
- pc : Mega piksel Kamera Utama
- px_height : Tinggi Resolusi Piksel
- px_width : Lebar Resolusi Piksel
- ram : Memori Akses Acak dalam Megabyte
- sc_h : Tinggi Layar ponsel dalam cm
- sc_w : Lebar Layar ponsel dalam cm
- talk_time : Waktu paling lama untuk satu kali pengisian daya baterai
- three_g : Memiliki 3G atau tidak
- touch_screen : Memiliki layar sentuh atau tidak
- wifi : Memiliki wifi atau tidak
- price_range : Rentang Harga

III. HASIL DAN PEMBAHASAN

Bab ini membahas bagaimana implementasi beberapa metode untuk mengetahui metode manakah yang digunakan dalam penelitian ini yang memberikan hasil yang tepat atau akurat. Sebelum memulai implementasi, Anda harus mempersiapkan beberapa hal berikut:

A. Implementasi Dan Hasil

1) Kebutuhan Non Fungsional

a) Kebutuhan perangkat lunak (software)

Berikut ini adalah kebutuhan *software* yang dibutuhkan, yaitu:

1. *Operating System Windows 10* digunakan sebagai platform dasar untuk menjalankan dan mendukung operasional perangkat lunak penelitian.
2. *RapidMiner* digunakan sebagai alat untuk melakukan analisis data dan pemodelan prediktif karena rapidminer memiliki kemampuan analisis yang kuat dan antarmuka pengguna yang ramah.
3. *Draw.io* digunakan sebagai perangkat lunak untuk membuat diagram visual yang mendukung dokumentasi dan komunikasi penelitian.

b) Kebutuhan perangkat keras (hardware)

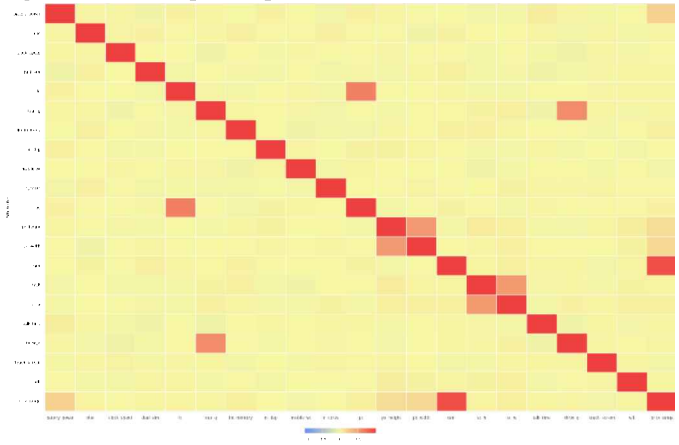
Berikut ini adalah kebutuhan *hardware* yang dibutuhkan, yaitu:

1. *AMD Ryzen R7-3750H*
2. *Radeon Graphics*
3. RAM 8GB DDR4
4. 512GB SSD M.2 2242 PCIe NVMe 3.0

B. Hubungan korelasi antar atribut spesifikasi harga handphone

Gambar 2 merupakan hubungan korelasi antar atribut spesifikasi handphone. Dalam konteks ini, kita mengevaluasi sejauh mana spesifikasi tertentu pada handphone berhubungan dengan kisaran harga yang dimilikinya sehingga

kita dapat menentukan variabel mana saja yang akan digunakan dalam proses prediksi.



Gambar 2. Hubungan Korelasi antar Variabel

TABEL 2. NILAI KORELASI ATRIBUT SPESIFIKASI DENGAN PRICE RANGE

Atribut Spesifikasi	Price_range
battery_power	0,201
blue	0,201
clock_speed	-0,21
dual_sim	0,017
fc	0,022
four_g	0,015
int_memory	0,044
m_dep	0,001
mobile_wt	-0,030
n_cores	0,004
pc	0,034
px_height	0,149
px_width	0,166
ram	0,917
sc_h	0,023
sc_w	0,039
talk_time	0,022
three_g	0,024
touch_screen	-0,030
wifi	0,019
price_range	1

Tabel 2 merupakan korelasi antara atribut spesifikasi (seperti battery power, clock speed, ram, dll.) dengan price_range pada handphone, kita dapat mengetahui beberapa informasi penting yaitu:

1. Ram (ram: 0,917):
   Korelasi yang sangat tinggi (0,917) dengan price_range menunjukkan bahwa kapasitas RAM memiliki pengaruh yang sangat kuat terhadap penentuan kisaran harga handphone. Semakin tinggi kapasitas RAM, kemungkinan handphone masuk dalam kategori harga yang lebih tinggi.

2. Px_width (px_width: 0,166) dan Px_height (px_height: 0,149):
   Resolusi layar (px_width dan px_height) memiliki korelasi positif yang cukup kuat dengan price_range. Handphone dengan layar resolusi tinggi cenderung memiliki kisaran harga yang lebih tinggi.

3. Battery_power (battery_power: 0,201):
   Korelasi positif (0,201) dengan price_range menunjukkan bahwa kapasitas baterai memiliki pengaruh positif pada kisaran harga. Handphone dengan baterai yang lebih besar cenderung masuk dalam kategori harga yang lebih tinggi.

4. Clock_speed (clock_speed: -0,21):
   Korelasi negatif (-0,21) dengan price_range menunjukkan bahwa kecepatan clock memiliki pengaruh negatif pada kisaran harga. Handphone dengan clock speed yang lebih tinggi cenderung masuk dalam kategori harga yang lebih rendah.

5. Mobile_wt (mobile_wt: -0,030) dan Touch_screen (touch_screen: -0,030):
   Korelasi negatif yang kecil dengan price_range menunjukkan bahwa berat handphone dan keberadaan touch screen tidak memiliki pengaruh signifikan terhadap kisaran harga.

6. Variabel Lainnya:
   Atribut lain seperti blue, dual_sim, fc, four_g, int_memory, m_dep, n_cores, pc, sc_h, sc_w, talk_time, three_g, wifi memiliki korelasi yang tidak begitu tinggi dan mungkin memberikan pengaruh yang lebih kecil pada penentuan kisaran harga.

Kita dapat mengetahui bahwa RAM, resolusi layar, dan kapasitas baterai merupakan faktor-faktor utama yang mempengaruhi harga handphone. Kecepatan clock, berat handphone, dan keberadaan touch screen memiliki pengaruh yang kurang signifikan. Perlu diingat bahwa korelasi tidak menyiratkan hubungan sebab-akibat, dan faktor lain di luar variabel tersebut juga dapat berkontribusi pada penentuan harga handphone.

C. Hasil Implementasi Penggunaan Model KNN dan Regresi Linear Menggunakan Rapidminer

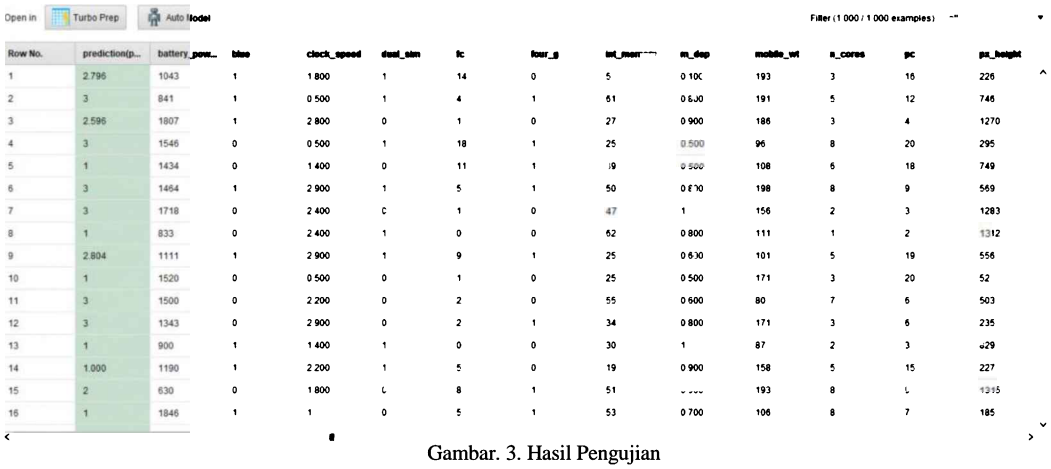
Tabel III dan Tabel IV merupakan hasil implementasi akurasi dari metode KNN dan Regresi Linear dengan menggunakan aplikasi RapidMiner didasarkan pada data spesifikasi harga ponsel. Berikut ini adalah penjelasan proses prediksi akurasi dari metode KNN dan Regresi Linear yang diolah menggunakan aplikasi RapidMiner.

TABEL 3. HASIL PERFORMANCE MENGGUNAKAN SELURUH ATRIBUT

Metode		Split Data		
		Ratio (9:1)	*Ratio (8:2)*	*Ratio (7:3)*
KNN	***RMSE***	0,265	0,251	0,250
	MAE	0,137	0,124	0,124
Regresi Linear	***RMSE***	0,334	0,319	0,321
	MAE	0,274	0,266	0,267

TABEL 4. HASIL PERFORMANCE MENGGUNAKAN ATRIBUT YANG TELAH DISELEKSI

Metode		Split Data		
		Ratio (9:1)	*Ratio (8:2)*	*Ratio (7:3)*
KNN	***RMSE***	0,263	0,250	0,252
	MAE	0,135	0,123	0,125
Regresi Linear	***RMSE***	0,333	0,319	0,320
	MAE	0,273	0,266	0,267



Gambar. 3. Hasil Pengujian

D. Evaluasi Hasil Pengujian

Gambar 3 merupakan hasil dari pengujian prediksi harga hanphone menggunakan model algoritma KNN dengan Menggunakan Perbandingan 7:3.

IV. KESIMPULAN

Penelitian ini menunjukkan bahwa dalam memprediksi harga ponsel, model KNN secara konsisten melampaui Regresi Linear dalam hal akurasi, sebagaimana dibuktikan oleh nilai RMSE dan MAE yang lebih rendah. Terutama, seleksi atribut dengan fokus pada atribut penting seperti RAM telah terbukti meningkatkan kinerja prediktif model KNN. Meskipun variasi rasio pembagian data (9:1, 8:2, 7:3) tidak menunjukkan perbedaan signifikan dalam hasil, praktik terbaik tampaknya menggunakan rasio 7:3. Temuan ini menekankan pentingnya seleksi atribut yang cermat dalam model prediksi harga dan menawarkan wawasan yang bisa diaplikasikan oleh konsumen dan produsen untuk keputusan yang lebih tepat. Namun, penelitian ini memiliki keterbatasan seperti ukuran sampel dan lingkup atribut, yang bisa mempengaruhi generalisasi temuan. Penelitian masa depan dapat melibatkan dataset yang lebih beragam dan menguji model tambahan untuk menentukan pendekatan terbaik dalam memprediksi harga ponsel.

REFERENSI

[1] S. H. Putra dan B. T. Putra, "Klasifikasi Harga Cell Phone menggunakan Metode K-Nearest Neighbor (KNN)," *Prosiding Annual Research Seminar*, vol. 4, no. 1, hlm. 242–245, 2018.

[2] A. Karim, F. Nurhadi, I. K. O. Setiawan, I. A. Rizky, dan R. B. Manurung, "Pengaruh Normalisasi Data Pada Klasifikasi Harga Ponsel Berdasarkan Spesifikasi Menggunakan Klasifikasi Naive Bayes dan Multinominal Logistic Regression," *Jurnal Rekayasa Elektro Sriwijaya*, vol. 4, no. 1, hlm. 8–16, 2022.

[3] A. Nugraha, Y. H. Chrisnanto, dan R. Yuniarti, "Prediksi Sentimen Pada Sosial Media Twitter Mengenai Produk Smartphone Menggunakan Algoritma K-NN Classification," *Seminar Nasional Sains & Teknologi Informasi (SENSASI)*, vol. 2, no. 1, Agu 2019, [Daring]. Tersedia pada: http://prosiding.seminar-id.com/index.php/sensasi/issue/archivePage|251

[4] V. Wanika Siburian dan I. Elvina Mulyana, "Prediksi Harga Ponsel Menggunakan Metode Random Forest," *Annual Research Seminar (ARS)*, vol. 4, no. 1, hlm. 144–147, 2018.

[5] M. Arhami dan M. Nasir, *DATA MINING Algoritma dan Implementasi*, 1 ed. Yogyakarta: ANDI, 2020.

[6] F. N. S. P. Mawan Pradana dan F. S. Papilaya, "Analisa Prediksi Harga Emas Dengan Kemungkinan Terjadinya Resesi Menggunakan Metode Svr," *SINTECH (Science and Information Technology)*, vol. 6, no. 1, hlm. 37–46, 2023, [Daring]. Tersedia pada: https://doi.org/10.31598

[7] A. Supriyadi Sunge dan A. Turmudi Zy, "Analisis Prediksi Penjualan dengan Metode Regresi Linear di PT. Eagle Industry Indonesia," *JINTEKS (Jurnal Informatika Teknologi dan Sains)*, vol. 5, no. 3, hlm. 398–403, 2023.

[8] P. B. Utomo, E. Utami, dan S. Raharjo, "Implementasi Metode K-Nearest Neighbor Dan Regresi Linear Dalam Prediksi Harga Emas," *Informasi Interaktif*, vol. 4, no. 3, hlm. 155–159, 2019, [Daring]. Tersedia pada: http://e-journal.janabadra.ac.id/

[9] M. Nanja dan Purwanto, "Metode K-Nearest Neighbor Berbasis Forward Selection Untuk Prediksi Harga Komoditi Lada," *Jurnal Pseudocode*, vol. 2, no. 1, hlm. 53–64, 2015, [Daring]. Tersedia pada: www.ejournal.unib.ac.id53

[10] Miftahuljannah, A. Supriyadi Sunge, dan A. Turmudi Zy, "Analisis Prediksi Penjualan Dengan Metode Regresi Linear Di Pt. Eagle Industry Indonesia," *Jurnal Informatika Teknologi dan Sains (Jinteks)*, vol. 5, no. 3, hlm. 398–403, 2023.

[11] A. Anggrawan, Hairani, dan N. Azmi, "Prediksi Penjualan Produk Unilever Menggunakan Metode Regresi Linear," *Jurnal Bumigora Information Technology (BITe)*, vol. 4, no. 2, hlm. 123–132, 2022, doi: 10.30812/bite.v4i2.2416.

[12] J. W. Iskandar dan Y. Nataliani, "Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 6, hlm. 1120–1126, Des 2021, doi: 10.29207/resti.v5i6.3588.

[13] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, dan Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 2, hlm. 393–399, Apr 2021, doi: 10.29207/resti.v5i2.3008.

[14] S. Howay dan Suhirman, "Comparison of SVM, NBC, and KNN Classification Methods in Determining Students' Majors at SMK N02 Manokwari," *Journal of Computer Science and Technology Studies*, vol. 5, no. 1, hlm. 15–23, 2023, doi: 10.32996/jcsts.

PROSIDING

SEMINAR NASIONAL INOVASI TEKNOLOGI INFORMASI & KOMUNIKASI

"Optimalisasi Teknologi Kecerdasan Artifisial untuk Mendukung Transformasi Digital dan Masa Depan Otomasi"

SANATA DHARMA UNIVERSITY PRESS
Jl. Affandi, (Gejayan) Mrican, Yogyakarta 55281
Phone: (0274)513301; Ext.51513
Web: *sdupress.usd.ac.id*; E-mail: *publisher@usd.ac.id*

ISBN 978-623-143-058-8 (PDF)

Sains & Teknologi