8th Seminar Nasional Teknik Elektro, Informatika dan Sistem Informasi (SINTaKS)
26 Agustus 2023 - Yogyakarta, Indonesia
ISSN: 2961-7464

Perbandingan Algoritma Klasifikasi Random Forest, Gaussian Naive Bayes, dan K-Nearest untuk Data Tidak Seimbang dan Data yang diseimbangkan dengan metode Random Undersampling pada dataset LCMS Tanaman Keladi Tikus

Eliana Tangkelobo[1], Wisye Mayaut[2], Hanjian Listanto[3], Iwan Binanto[4*], Nesti F. Sianipar[5]

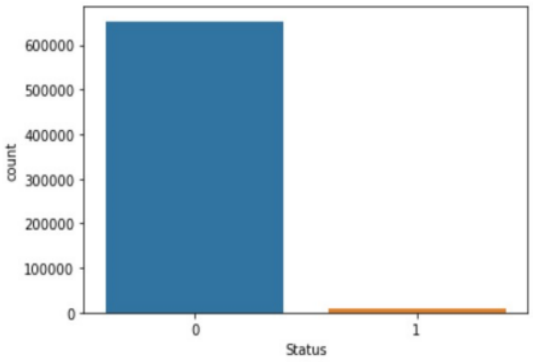
[1,2,3,4]*Informatika, Universitas Sanata Dharma, Yogyakarta*
[5]*Biotechnology Department, Faculty of Engineering, Bina Nusantara University*
[1]elyanatangkelobo05@gmail.com
[2]Wisyemayaut29@gmail.com
[3]ahan.cf.7@gmail.com
[4*]iwan@usd.ac.id
[5]nsianipar@binus.edu

***Abstrak*— Dalam penelitian ini, dua percobaan terpisah dilakukan untuk mengklasifikasikan dataset keladi tikus menggunakan algoritma KNN, Naive Bayes, dan Random Forest. Pertama, dilakukan pengujian dengan menggunakan dataset yang belum seimbang, sedangkan kedua, dilakukan Random Undersampling (RUS) untuk mendapatkan dataset yang seimbang sebelum melakukan pengujian. Hasil pengujian dibandingkan menggunakan confusion matrix. Confusion matrix memberikan informasi lebih rinci tentang performa model dalam mengklasifikasikan sampel pada setiap kelas. Hasil eksperimen menunjukkan bahwa penggunaan Random Undersampling secara signifikan meningkatkan kinerja model klasifikasi pada ketiga metode yang diuji. Pada dataset yang telah diseimbangkan menggunakan Random Undersampling, algoritma Random Forest menunjukkan performa yang jauh lebih baik dalam hal akurasi, presisi, recall, dan F1-score yang mencapai 99% jika dibandingkan dengan Gaussian Naive Bayes dan KNN. Dan juga pada dataset yang belum seimbang, Random Forest tetap menunjukkan hasil akurasi yang lebih tinggi dibandingkan dengan KNN dan Gaussian NB, meskipun perbedaannya tidak sebesar pada dataset yang telah diseimbangkan.**

***Kata kunci*— Algoritma, Imbalance Data, Balance Data, Undersampling, KNN, Random Forest, Gaussian Naive Bayes, Klasifikasi**

I. PENDAHULUAN

Data yang tidak seimbang belakangan ini mendapat perhatian sebagai masalah serius untuk melakukan penelitian atau sampai pada kesimpulan yang bermanfaat. Ketidakseimbangan kelas adalah salah satu masalah dengan pembelajaran mesin dan penambangan data. Kumpulan data yang tidak seimbang telah menghambat efektivitas penambangan data dan algoritma pembelajaran mesin (dalam hal akurasi keseluruhan, pengambilan keputusan). Masalah ketidakseimbangan kelas muncul dalam klasifikasi ketika ada lebih banyak sampel secara signifikan dalam setidaknya satu kelas daripada sampel di kelas lainnya. Kelas yang memiliki sampel terbanyak disebut sebagai kelas mayoritas, dan kelas lainnya disebut sebagai kelas minoritas. [1] Real Data pada umumnya jarang seimbang dalam berbagai situasi. Satu kelas sering melebihi jumlah kelas lain dalam koleksi, yang menyebabkan masalah data kelas yang tidak seimbang atau unbalance. Salah satunya data yang tidak seimbang yaitu ditemukan pada penelitian Binanto [2] yang merupakan salah satu data LCMS yang berasal dari tanaman Keladi Tikus dari hasil penelitian [3] Data yang tidak seimbang mengenai tanaman Keladi Tikus ini divisualisasikan seperti pada gambar 1.



Gambar 1. Visualisasi Data Tidak Seimbang Tanaman Keladi Tikus

Apabila dilakukan klasifikasi pada data keladi tikus yang tidak seimbang ini dengan menggunakan machine learning maka hasil yang diberikan tidak akan akurat. Namun, ada algoritma klasifikasi yang dapat memberikan hasil yang cukup akurat terhadap data yang tidak seimbang seperti tanaman Keladi Tikus ini, yaitu algoritma klasifikasi Random Forest [4]. Data tanaman keladi tikus yang tidak seimbang ini akan

diseimbangkan menggunakan machine learning dengan algoritma undersampling.

Dalam algoritma random undersampling, Algoritma ini menggunakan teknik-teknik yang berfokus pada pengoptimalan statistik selain akurasi untuk mengatasi masalah ketidakseimbangan kelas. Pengklasifikasi ini mampu memberikan model yang lebih informatif dengan mengoptimalkan metrik selain akurasi yang lebih sesuai untuk masalah ketidakseimbangan kelas [5]. Data yang sudah seimbang tersebut akan dilakukan klasifikasi menggunakan machine learning dengan menggunakan Gaussian Naive Bayes (GNB) dan K-Nearest Neighbors (KNN) karena dapat menghasilkan klasifikasi yang lebih baik penggunaan yang lebih efektif, mudah dipelajari, tahan dari noise data, dan efisien apabila data pelatihan yang ditemukan pada dataset tersebut sangat banyak [6]. Penelitian ini bereksperimen kepada data yang tidak seimbang dan data seimbang dengan menggunakan random undersampling untuk menyeimbangkan datanya kemudian menggunakan pendekatan Random Forest, KNN dan Gaussian NB sebagai algoritma klasifikasi machine learning.

II. STUDI PUSTAKA

A. *Random Undersampling*

Algoritma Undersampling merupakan salah satu metode yang dipilih dalam metode penelitian ini karena algoritma ini telah menjadi salah satu standar pendekatan untuk meningkatkan kinerja klasifikasi untuk menangani data yang tidak seimbang. Selain itu, Algoritma Undersampling memiliki kinerja prediktif yang tinggi, kompleksitas waktunya dibatasi oleh ukuran instance kelas minoritas. Teknik ini menemukan jumlah kasus minoritas yang sebanding dan mengelompokkannya selama fase pelatihan atau training. Untuk memastikan bahwa setiap kelas memiliki jumlah objek yang sama, undersampling secara acak menghapus objek dari kelas yang menjadi mayoritas [7].

B. *Random Forest*

Random Forest merupakan metode yang menggunakan beberapa Pohon Keputusan, di mana setiap pohon keputusan diuji menggunakan individu dan atribut yang dianggap akurat. Metode ini memiliki beberapa keunggulan, termasuk kemampuan untuk meningkatkan kinerja saat ada data yang kadaluarsa, tahan terhadap outlier, dan efektif dalam memperluas kumpulan data. Selain itu, Random Forest juga memiliki proses pemilihan fitur terbaik yang dapat meningkatkan kinerja model klasifikasi. Dengan menambahkan fitur tambahan, Random Forest mampu bekerja secara efektif dengan data yang besar menggunakan parameter yang kompleks. Dibandingkan dengan algoritma lainnya. Algoritma Random Forest dapat bekerja lebih baik dalam menghasilkan hasil klasifikasi apabila jumlah atribut yang sama dan juga kumpulan data tersebut besar, dan memiliki lebih banyak instance [8]**.**

Random Forest memiliki beberapa keunggulan, termasuk kemampuannya untuk meningkatkan akurasi ketika ada data yang hilang, serta ketahanannya terhadap outlier. Selain itu, metode ini efisien dalam penyimpanan data. Salah satu fitur penting dalam Random Forest adalah proses seleksi fitur, yang memungkinkannya untuk memilih fitur terbaik dan meningkatkan performa model klasifikasi. Dengan fitur seleksi ini, Random Forest dapat efektif bekerja dengan data yang besar dan menggunakan parameter yang kompleks. Namun, terkadang Random Forest dapat menghasilkan nilai yang tidak diharapkan dan tidak mampu memprediksi rentang nilai respons pada data latih dengan akurasi yang tinggi [9].

Salah satu penelitian yang membahas tentang keunggulan dan tingkat keberhasilan algoritma Random forest dalam mengklasifikasikan dataset yang telah diseimbangkan menggunakan metode Random Undersampling ditemukan pada penelitian Eko Saputro [18].

C. *K-Nearest Neighbors*

Algoritma K-Nearest Neighbors (KNN) adalah sebuah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Algoritma ini mengukur jarak antara objek data dalam ruang fitur dan mencari k-nearest neighbors atau tetangga terdekat dari suatu objek data. Kemudian, dengan menggunakan mayoritas voting atau rata-rata bobot tetangga terdekat, algoritma KNN memprediksi label atau nilai target dari objek data yang diberikan. Algoritma ini terbukti efektif dalam menyelesaikan berbagai masalah klasifikasi. Pendekatan ini relatif mudah dan simpel, di mana jumlah tetangga digunakan untuk mengeliminasi instance dari kelas mayoritas. Meskipun sederhana, uji klasifikasi menggunakan pendekatan KNN menghasilkan kinerja yang lebih baik dan lebih akurat [11].

Berikut ini adalah langkah-langkah sederhana dalam menjelaskan algoritma K-Nearest Neighbors (KNN):

1. Tentukan jarak (biasanya Euclidean) antara setiap contoh set pelatihan T, T, dan xi.
2. Pilih k tetangga terdekat.
3. Kelas yang paling umum di antara k tetangga terdekat digunakan untuk mengkategorikan dan melabeli instance xi. Jarak antara tetangga juga dapat digunakan untuk mempengaruhi pilihan klasifikasi.
4. Buat prediksi: Gunakan label atau nilai target yang dipilih sebagai prediksi untuk objek data yang diberikan.

Salah satu penelitian yang membahas tentang keunggulan dan tingkat keberhasilan algoritma KNN dalam mengklasifikasikan dataset yang telah diseimbangkan menggunakan metode Random Undersampling dan metode Random Oversampling ditemukan pada penelitian oleh Reasianta Perangin-angin [19].

D. *Gaussian Naïve Bayes*

Algoritma Naïve Bayes adalah metode klasifikasi yang menggunakan probabilitas sederhana dan didesain dengan asumsi bahwa kelas-kelas yang ada saling independen satu sama lain. Dalam klasifikasi Naïve Bayes, penekanan utama adalah pada estimasi probabilitas. Pendekatan ini memiliki keuntungan di mana klasifikasi akan menghasilkan nilai error yang lebih kecil ketika digunakan pada data set yang besar, Menurut Han dan Kamber, klasifikasi Naïve Bayes telah

terbukti memiliki tingkat akurasi dan kecepatan yang tinggi saat diterapkan pada basis data dengan jumlah yang besar. Selain itu, Algoritma Naïve Bayes ini sering digunakan sebagai metode klasifikasi karena pada prosesnya, Algoritma tersebut cenderung lebih cepat dan juga mudah untuk diterapkan serta dapat memberikan hasil yang stabil dan tidak dipengaruhi oleh data pencicilan atau outlier dalam dataset [12].

Salah satu penelitian yang membahas penggunaan algoritma Gaussian NB dalam mengklasifikasikan dataset yang telah diseimbangkan menggunakan metode Random Undersampling ditemukan pada penelitian oleh Jus Prasetya [21].

E. Confusion Matrix

Confusion matrix merupakan sebuah metode yang digunakan untuk menilai performa model klasifikasi dengan memperbandingkan hasil prediksi model dengan label kelas yang sebenarnya pada dataset yang diberikan. Confusion matrix memberikan informasi mengenai jumlah prediksi yang tepat dan salah yang dilakukan oleh model, serta jenis kesalahan yang terjadi [14].

TABEL 1
TABEL CONFUSION MATRIX [14]

		Kelas Prediksi	
		1	**0**
Kelas Sebenarnya	**1**	TP	FN
	0	FP	TN

Confusion matrix terdiri dari sebuah matriks berukuran 2x2 (untuk klasifikasi biner) atau ukuran yang lebih besar (untuk klasifikasi multikelas). Matriks ini terdiri dari empat bagian yang merepresentasikan kombinasi hasil prediksi dan label sebenarnya, yaitu:

- True Positive (TP): Jumlah sampel yang diprediksi dengan benar sebagai positif oleh model.
- True Negative (TN): Jumlah sampel yang diprediksi dengan benar sebagai negatif oleh model.
- False Positive (FP): Jumlah sampel yang diprediksi keliru sebagai positif oleh model (kesalahan tipe I atau false alarm).
- False Negative (FN): Jumlah sampel yang diprediksi keliru sebagai negatif oleh model (kesalahan tipe II atau missed detection) [15].

Perhitungan metriks evaluasi seperti akurasi (accuracy), recall, presisi (precision), dan F1-score dapat dilakukan menggunakan nilai-nilai yang terdapat dalam confusion matrix [20]. Berikut adalah rumus-rumus perhitungan metrik tersebut:

1) Akurasi (Accuracy)
   Akurasi mengukur seberapa baik model dapat melakukan prediksi yang benar secara keseluruhan.

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2) Recall (Sensitivitas atau True Positive Rate)
   Recall mengukur seberapa baik model dalam mengenali atau menemukan sampel positif yang sebenarnya.

$$\frac{TP}{TP+FN} \quad (2)$$

3) Presisi (Precision)
   Presisi mengukur seberapa baik model dalam mengklasifikasikan sampel sebagai positif dengan benar.

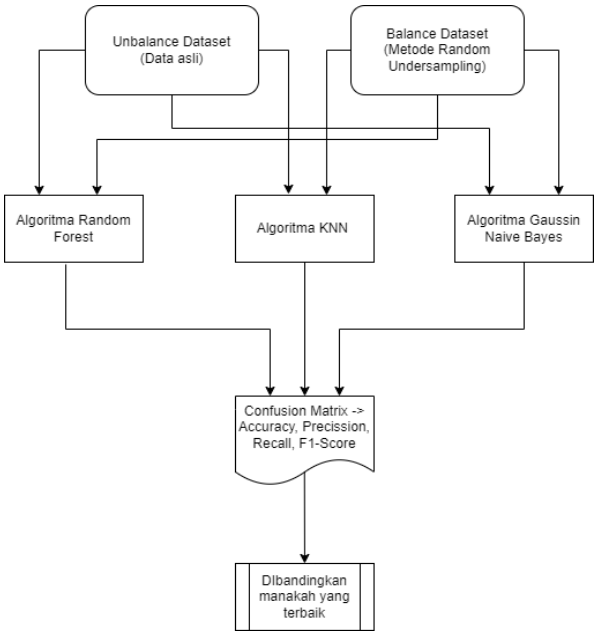
$$\frac{TP}{TP+FP} \quad (3)$$

4) F1-Score
   F1-score adalah rata-rata harmonik antara recall dan presisi. Metrik ini memberikan keseimbangan antara recall dan presisi.

$$\frac{2*(Presisi*Recall)}{Presisi+Recall} \quad (4)$$

III. METODE PENELITIAN

Tahapan penelitian ini meliputi langkah-langkah berikut: dimulai dengan memilih dataset yang tidak seimbang (imbalance data) untuk mengevaluasi kinerja beberapa algoritma klasifikasi, seperti Random Forest, KNN, dan Gaussian Naïve Bayes. Selanjutnya, dilakukan pengolahan ulang data (reprocessing) untuk mencapai keseimbangan. Untuk mengatasi masalah ketidakseimbangan kelas dalam dataset, penelitian ini menggunakan pendekatan level data dengan menggunakan teknik resampling.



Gambar 2. Kerangka Penelitian

Algoritma resampling yang digunakan adalah random undersampling (RUS). Setelah itu, data yang telah seimbang

tersebut akan digunakan untuk mengukur kembali kinerja algoritma klasifikasi, yaitu Random Forest, KNN, dan Gaussian Naïve Bayes. Hasil evaluasi kinerja menggunakan data yang seimbang dari beberapa algoritma klasifikasi tersebut akan dibandingkan. Kerangka metode penelitian ditunjukkan pada gambar 2.

IV. HASIL DAN DISKUSI

A. *Data Spliting*

Penelitian ini menggunakan dataset keladi tikus yang merupakan imbalance data. Dataset yang dipakai dalam penelitian ini berjumlah 663228 baris dengan 6 kolom. Pembagian data menjadi data pelatihan (training) dan data pengujian (testing) adalah langkah yang dikenal sebagai Data Splitting [14]. Data training sebesar 70% dan data testing sebesar 30%. Parameter yang digunakan dalam penelitian ini adalah test_size = 0.3 dan random_state = 42. Tabel 1 memberikan contoh visualisasi dari pembagian data secara umum.

TABEL 2
ILUSTRASI PEMBAGIAN DATA SECARA UMUM

Dataset	
Data Training	Data Testing
70%	30%

Berikut adalah label yang terdapat dalam dataset yang kami gunakan

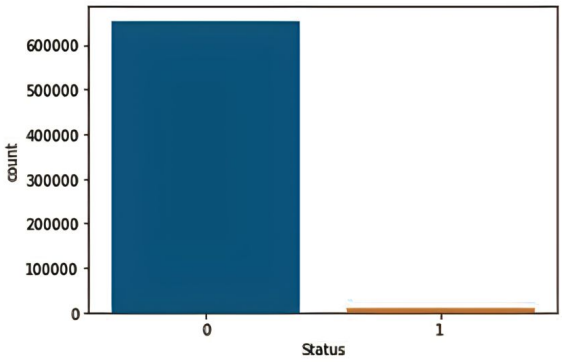
TABEL 3
INFORMASI DATASET

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 663228 entries, 0 to 663227
Data columns (total 7 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Retention Time  663228 non-null  float64
 1   Intensity       663228 non-null  int64
 2   m/z             663228 non-null  float64
 3   Real_m/z        663228 non-null  float64
 4   NamaSenyawa     663228 non-null  object
 5   RumusSenyawa    662916 non-null  object
 6   Status          663228 non-null  int64
dtypes: float64(3), int64(2), object(2)
memory usage: 35.4+ MB
```

Kami mengimplementasikan undersampling menggunakan library imblearn, sementara implemetasi algoritma klasifikasi menggunaan library sklearn. sementara implementasi algoritma klasifikasi menggunakan library sklearn.

B. *Pembahasan*

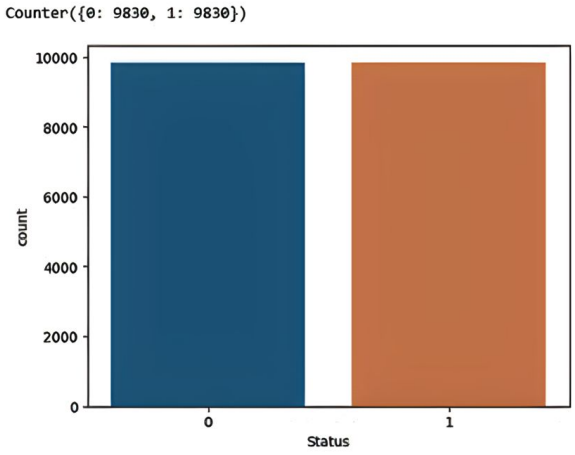
Penelitian ini melakukan perbandingan antara implementasi langsung algoritma klasifikasi dengan imbalance data dan implementasi algoritma klasifikasi dengan balance data yang telah mengalami proses pengambilan sampel (sampling) terlebih dahulu. Berikut ini adalah hasil dari penelitian yang telah dilakukan.



Gambar 3. Data sebelum diundersampling

Pada gambar 3 diatas menunjukkan dataset yang imbalance. Dalam diagram visualisasi, sebelum melakukan random undersampling, terdapat perbedaan yang signifikan dalam jumlah data antara kelas minoritas dan kelas mayoritas. Pada kelas minoritas yang ditandai dengan angka 1, memiliki hanya 9.830 data. Sementara kelas mayoritas, ditandai dengan angka 0 memiliki 653.398 data.

Oleh sebab itu, dilakukan upaya untuk menyeimbangkan data dengan menggunakan algoritma undersampling yang melakukan penghapusan objek secara acak dari kelas mayoritas sehingga jumlah data menjadi setara dengan kelas minoritas [7].



Gambar 4. Data sesudah seimbang

Berdasarkan diagram visualisasi diatas, data pada kelas mayoritas (0) telah diseimbangkan dengan menggunakan algoritma RandomUndersampling. Dimana, datanya sudah sama dengan kelas minoritas (1) yaitu 9.830 data.

Hasil implementasi algoritma klasifikasi langsung ke dataset imbalance tanpa resampling Random Undersampling dengan parameter (test_size = 0.3, random_state = 42). Pada algoritma KNN, nilai K yang dipakai sebagai parameter adalah akar dari jumlah baris data sebagai berikut :

TABEL 4
HASIL EKSPERIMEN IMBALANCE DATA

HASIL EKSPERIMEN			
	Imbalance Data (Data Asli)		
	Random Forest	KNN	Gaussian NB
Accuracy	0.990	0.985	0.985
Recall	0.332	0.000	0.000
Precission	0.996	0.000	0.000
F1-Score	0.496	0.000	0.000
Process Time (s)	81.231	0.90672	0.16096

Pada imbalance data (data yang tidak seimbang), sebelum melakukan Random Undersampling, dalam percobaan pada tabel diatas menggunakan metode Gaussian Naive Bayes menghasilkan nilai akurasi yang tinggi sebesar 98%. Namun, precision, recall, dan F1-score memiliki nilai 0% dengan waktu proses 0.1 detik.

Demikian pula, pada metode KNN sebelum Random Undersampling, tingkat akurasi mencapai 98%, tetapi precision, recall, dan F1-score memiliki nilai yang 0% dengan rata-rata waktu proses 0.9 detik.

Berbeda dengan Random Forest, nilai akurasi dan precision yang didapatkan mencapai 99% namun nilai recall dan F1-score antara 30%-50% dengan waktu proses lebih dari 1 menit.

TABEL 5
HASIL EKSPERIMEN BALANCE DATA

HASIL EKSPERIMEN			
Balance Data (Data Seimbang) dgn Metode RUS			
	Random Forest	KNN	Gaussian NB
Accuracy	0.991	0.529	0.506
Recall	0.999	0.600	0.500
Precission	0.978	0.520	0.992
F1-Score	0.991	0.525	0.660
Process Time (s)	2.128	0.01847	0.00435

Dalam kondisi imbalance data, di mana jumlah sampel dalam setiap kelas tidak terdistribusi secara merata, terdapat beberapa faktor yang dapat menyebabkan terjadinya perbedaan antara nilai akurasi yang tinggi dan nilai recall, presisi, serta F1 score yang rendah atau bahkan 0%. Hal ini disebabkan oleh ketidakseimbangan kelas dalam dataset. Ketika terdapat perbedaan yang signifikan dalam jumlah sampel antara kelas mayoritas dan kelas minoritas dalam dataset, algoritma pembelajaran mesin cenderung untuk memprediksi dengan benar kelas mayoritas sebagian besar waktu. Dalam hal ini, nilai akurasi akan menjadi tinggi karena kelas mayoritas diprediksi dengan benar, namun nilai recall, presisi, dan F1 score untuk kelas minoritas akan menjadi rendah karena algoritma cenderung mengabaikan atau salah mengklasifikasikan sampel dari kelas minoritas [16][17].

Implementasi algoritma klasifikasi setelah dilakukan Random Undersampling pada dataset dengan parameter (test_size = 0.3, random_state = 42). Pada algoritma KNN, nilai K yang dipakai sebagai parameter adalah akar dari jumlah baris data.

Pada data seimbang atau data yang sudah dilakukan Random Undersampling, berdasarkan tabel diatas, dalam percobaan dengan metode Gaussian Naive Bayes akan menghasilkan nilai akurasi sebesar 50%, precision sebesar 50%, recall sebesar 99%, dan F1-score sebesar 66% dengan ata-rata waktu proses 0.004 detik.

Hasil implementasi dengan metode KNN pada data yang seimbang dalam percobaan menghasilkan nilai akurasi sebesar 52%, precision sebesar 52%, recall sebesar 60%, dan F1-score sebesar 52% dengan waktu proses 0.01 detik.

Hasil implementasi algoritma Random Forest pada data yang seimbang dalam percobaan menghasilkan nilai akurasi sebesar 99%, precision sebesar 97%, recall sebesar 99% dan F1-score 99% dengan waktu proses lebih dari 2 menit.

V. KESIMPULAN

Pada data yang seimbang, dimana jumlah sampel dalam setiap kelas dataset terdistribusi secara merata, memiliki nilai akurasi, presisi, recall, dan F1-score yang setara atau mendekati nilai yang sama dianggap lebih baik. Hal ini menunjukkan adanya keseimbangan yang baik dalam melakukan klasifikasi untuk kedua kelas. Dalam konteks ini, nilai-nilai evaluasi yang setara pada data yang seimbang menunjukkan keseimbangan yang baik dalam kinerja klasifikasi untuk kedua kelas. Ini menjadi indikasi bahwa algoritma mampu memperlakukan kedua kelas dengan baik dan menghasilkan hasil klasifikasi yang dapat diandalkan pada data seimbang.

Pada penelitian ini, terdapat dataset tidak seimbang yang telah diseimbangkan menggunakan metode Random Undersampling dan telah diuji coba menggunakan 3 algoritma berbeda yaitu Random Forest, Gaussin Naive Bayes, dan K-Nearest Neighbors. Sehingga berdasarkan hasil penelitian ini dapat disimpulkan bahwa Random Forest menunjukkan performa yang jauh lebih baik dalam hal akurasi, presisi, recall, dan F1-score yang mencapai 99% jika dibandingkan dengan Gaussian Naive Bayes dan KNN. Namun, pemilihan algoritma yang tepat tergantung pada prioritas dan kebutuhan spesifik pengguna. Jika fokus utama adalah mencapai performa tinggi dengan nilai akurasi, recall, presisi, dan F1-score yang setara, maka Random Forest mungkin merupakan pilihan yang lebih baik. Meskipun prosesnya membutuhkan waktu yang lebih lama, hasilnya dapat memberikan kinerja yang lebih baik dalam memprediksi kelas pada data yang seimbang. Namun, jika terdapat batasan waktu yang ketat dan kebutuhan akan waktu proses yang cepat, KNN bisa menjadi pilihan yang lebih baik. Meskipun KNN memiliki kinerja yang sedikit lebih rendah, namun waktu prosesnya jauh lebih singkat.

REFERENSI

[1] A. R. B. Alamsyah, S. Rahma, N. S. Belinda, and A. Setiawan, "A R B Alamsyah et al SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data Case Study: IFLS 5," pp. 305–314, 2021, [Online]. Available:https://proceedings.stis.ac.id/icdsos/article/download/240/29/2098

[2] I. Binanto, H. L. H. S. Warnars, N. F. Sianipar, and W. Budiharto, "Anticancer Compound Identification Model of Rodent Tuber's Liquid Chromatography-Mass Spectrometry Data," *ICIC Express Lett.*, vol. 16, no. 1, pp. 9–16, 2022, doi: 10.24507/icicel.16.01.9.

[3] N. F. Sianipar, R. Purnamaningsih, D. L. Gumanti, Rosaria, and M. Vidianty, "Analysis of gamma irradiated fourth generation mutant of rodent tuber (typhonium flagelliforme lodd.) Based on morphology and rapd markers," *J. Teknol.*, vol. 78, no. 5–6, pp. 41–49, 2016, doi: 10.11113/jt.v78.8636.

[4] A. S. More and D. P. Rana, "Review of random forest classification techniques to resolve data imbalance," *Proc. - 1st Int. Conf. Intell. Syst. Inf. Manag. ICISIM 2017*, vol. 2017-Janua, pp. 72–78, 2017, doi: 10.1109/ICISIM.2017.8122151.

[5] T. Ryan Hoens and N. V. Chawla, "Imbalanced datasets: From sampling to classifiers," *Imbalanced Learn. Found. Algorithms, Appl.*, pp. 43–59, 2013, doi: 10.1002/9781118646106.ch3.

[6] G. A. Rosso, "OPTIMASI TEKNIK KLASIFIKASI MODIFIED K NEAREST NEIGHBOR MENGGUNAKAN ALGORITMA GENETIKA," *William Blake Context*, no. September, pp. 184–191, 2019, doi: 10.1017/9781316534946.021.

[7] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," no. December 2013, 2012.

[8] S. Bangabash and S. Panda, "Machine Learning - Managerial Perspective," 2019.

[9] J. Sun, W. Du, and N. Shi, "A Survey of kNN Algorithm," pp. 1–10.

[10] S. Boriah, "Similarity Measures for Categorical Data : A Comparative Evaluation Similarity Measures for Categorical Data : A Comparative Evaluation," no. June 2014, 2008, doi: 10.1137/1.9781611972788.22.

[11] R. E. Putri, Suparti, and R. Rahmawati, "Perbandingan Metode Klasifikasi Naãve Bayes Dan K-Nearest Neighbor Pada Analisis Data Status Kerja Di Kabupaten Demak Tahun 2012," *J. Gaussian*, vol. 3, no. 4, pp. 831–838, 2014.

[12] M. Kamber and J. Han, *Data Mining: Concepts and Techniques : Concepts and Techniques*. 2018.

[13] C. K. Aridas, S. Karlos, V. G. Kanas, N. Fazakis, and S. B. Kotsiantis, "Uncertainty Based Under-Sampling for Learning Naive Bayes Classifiers under Imbalanced Data Sets," *IEEE Access*, vol. 8, pp. 2122–2133, 2020, doi: 10.1109/ACCESS.2019.2961784.

[14] G. Gumelar *et al.*, "Kombinasi Algoritma Sampling dengan Algoritma Klasifikasi untuk Meningkatkan Performa Klasifikasi Dataset Imbalance," *Pros. Semin. Nas. Sist. Inf. dan Teknol. ke 5*, pp. 250–255, 2021.

[15] L. Qadrini, "Undersampling dan K-Fold Random Forest Untuk Klasifikasi Kelas Tidak Seimbang," vol. 4, no. 4, pp. 1967–1974, 2023, doi: 10.47065/bits.v4i4.3141.

[16] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," *Proc. - 4th Int. Conf. Nat. Comput. ICNC 2008*, vol. 4, pp. 192–201, 2008, doi: 10.1109/ICNC.2008.871.

[17] J. Gaussian, "3 1,2,3," vol. 10, pp. 11–20, 2021.

[18] E. Saputro and D. Rosiyadi, "Penerapan Metode Random Over-Under Sampling Pada Algoritma Klasifikasi Penentuan Penyakit Diabetes," *Bianglala Inform.*, vol. 10, no. 1, pp. 42–47, 2022, doi: 10.31294/bi.v10i1.11739.

[19] R. Perangin-angin, E. J. G. Harianja, and I. K. Jaya, "Pendekatan Level Data untuk Menangani Ketidakseimbangan Data Menggunakan Algoritma K-Nearest Neighbor," *J. TIMES*, vol. IX, no. 1, pp. 22–32, 2020, [Online]. Available: https://ejournal.stmik-time.ac.id/index.php/jurnalTIMES/article/view/615

[20] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 697–711, 2021, [Online]. Available: http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/369

[21] J. Prasetya, "Leibniz : Jurnal Matematika," vol. 2, pp. 11–22, 2022.