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

Penerapan Metode Random Forest, Gaussian NB, Dan KNN Terhadap Data Unbalance dan Data Balance Menggunakan Random Over Sampling Untuk Klasifikasi Senyawa Keladi Tikus

Gabriel Advent Batan[1], Malvino Jordhan Keytimu[2], Flora Lebonna Katumbo[3], Iwan Binanto [4*], Nesti F. Sianipar[5]

Jurusan Informatika, Universitas Sanata Dharma

[5]*Biotechnology Department, Faculty of Engineering, Bina Nusantara University*

[1]bie.ritan.112@gmail.com

[2]keytimujordhan964@gmail.com

3floralebonnakatumbo@gmail.com

[4*]iwan@usd.ac.id

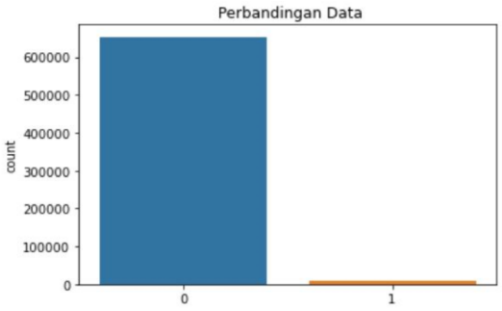
[5]nsianipar@binus.edu

***Abstrak*– Penelitian ini bertujuan untuk melakukan klasifikasi dataset senyawa keladi tikus menggunakan metode Random Forest, Gaussian NB, dan KNN pada dataset yang tidak seimbang dan seimbang dengan menggunakan Random Over Sampling. Metode penelitian melibatkan penerapan metode Random Forest, KNN, dan Gaussian NB pada dataset asli (data tidak seimbang) untuk pemodelan data training dan pengujian menggunakan data uji. Kinerja algoritma diukur dengan menggunakan confusion matrix, dan metode KNN dievaluasi dengan K-fold cross validation. Hasil penelitian menunjukkan bahwa sebelum data diseimbangkan, akurasi pada data tidak seimbang mencapai rata-rata 80%, namun parameter-parameter lainnya memiliki nilai yang rendah. Setelah menerapkan Random Over Sampling, akurasi meningkat untuk Random Forest, tetapi terjadi penurunan akurasi rata-rata pada KNN dan Gaussian NB. Penurunan ini disebabkan oleh penduplikatan pada kelas minoritas dan pengaruh nilai k yang terlalu besar pada KNN. Analisis waktu komputasi menunjukkan bahwa Random Forest dan KNN memerlukan waktu yang lebih lama daripada Naïve Bayes. Berdasarkan hasil penelitian ini, dapat disimpulkan bahwa algoritma Random Forest adalah yang terbaik untuk melakukan klasifikasi pada dataset senyawa keladi tikus, baik pada data tidak seimbang maupun data seimbang.**

***Kata kunci*– Random Over Sampling, Random Forest, Gaussian NB, KNN.**

I. PENDAHULUAN

Tanaman Keladi Tikus (*Typhonium Flagelliforme Lodd*) telah lama digunakan oleh masyarakat Indonesia sebagai obat tradisional. Dalam tanaman ini terdapat sejumlah zat aktif yang memiliki efek antikanker, seperti *triterpenoid*, *alkaloid*, *polifenol*, *Ribosome Inactivating Protein* (RIP), dan fitol [1], [2]. Setelah dilakukan ekstraksi data dari tanaman ini, ditemukan bahwa hasil ekstraksi data tersebut mengalami ketidakseimbangan data, di mana terdapat kelas minoritas dan kelas mayoritas yang sangat signifikan. Berdasarkan gambar 1, dapat dilihat bahwa jumlah data kelas mayoritas mencapai 653398, sedangkan jumlah data di kelas minoritas hanya mencapai 9830. Ketidakseimbangan data ini dapat



Gambar 1. Diagram hasil perbandingan data unbalance

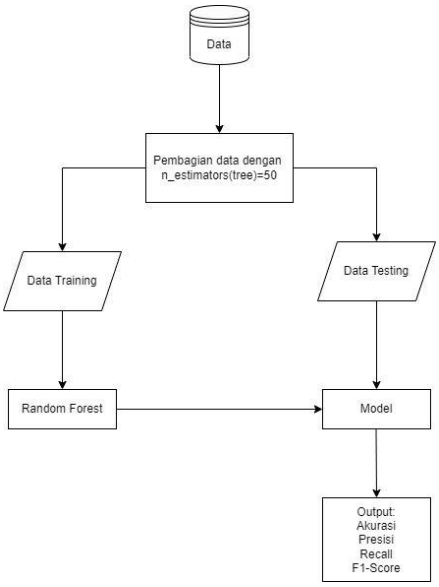
mengakibatkan data sulit diprediksi [3], [4] dan mempersulit pencarian yang cocok [5] pada saat hendak diklasifikasi menggunakan beberapa algoritma seperti *K-Nearest Neighbor*, *Gaussian Naïve Bayes*, dan *Random Forest*.

Salah satu cara untuk mengatasi masalah tersebut adalah dengan melakukan *resampling* dengan metode yang digunakan adalah *Random Over Sampling* (ROS), agar rasio ketimpangan antar kelas dapat dikurangi [4] dengan cara menduplikat data dari kelas minoritas [6], [7] sehingga berjumlah sama dengan kelas mayoritas. Dengan menerapkan algoritma ROS, data menjadi seimbang dan hasil prediksi dapat menjadi lebih baik ketika menerapkan algoritma-algoritma klasifikasi.

II. STUDI PUSTAKA

A. *Random Forest*

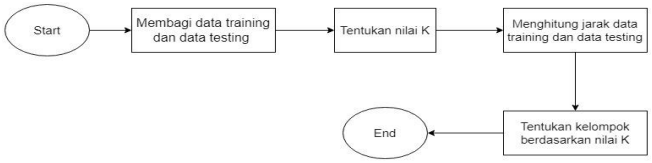
Random forest adalah metode *ensambel learning* yang pertama kali diusulkan oleh Breiman pada tahun 2001 yang merupakan kombinasi dari pohon klasifikasi [8], [9]. *Random Forest* telah banyak digunakan baik untuk klasifikasi dan regresi [8] sehingga metode ini cocok dalam hal klasifikasi [9]. Flowchart dari algoritma ini dapat dilihat pada gambar 2. *Random forest* cocok digunakan untuk memodelkan data dengan dimensi tinggi karena mampu mengatasi nilai yang hilang dan dapat menangani data dengan tipe kontinu, kategorikal, serta biner [10]. Selain memiliki tingkat akurasi prediksi yang tinggi, *Random Forest* juga merupakan metode yang efisien, dapat diinterpretasikan, dan bersifat non-parametrik untuk berbagai jenis dataset [10].



Gambar 2. Flowchart random forest

B. *K – Nearest Neighbour*

K-Nearest Neighbor (KNN) adalah suatu metode yang tergolong dalam *supervised learning*, yang berarti ia memerlukan data *training* untuk mengklasifikasikan objek berdasarkan jarak terdekatnya. Prinsip kerja KNN adalah mencari jarak terdekat antara data yang akan dievaluasi dengan nilai k atau tetangga terdekat dalam data pelatihan [11], [12].

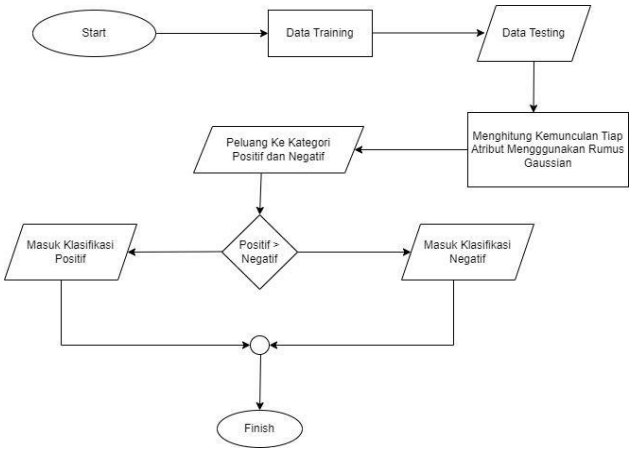


Gambar 3. Flowchart K – Nearest Neighbour

K – Nearest Neighbor (KNN) biasanya digunakan di dalam klasifikasi dan regresi data berdasarkan nilai k yang dimasukan. Untuk dapat diproses, KNN membutuhkan data *training* dan data testing [13]. Sehingga nantinya akan dihitung jarak antara data *training* dan data testing kemudian dari hasil yang didapatkan, akan diambil data sejumlah nilai k yang dimasukan. Untuk menghitung jarak, KNN menggunakan perhitungan jarak euclidean $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ sehingga flowchartnya dapat dilihat seperti pada gambar 3.

C. *Gaussian Naïve Bayes*

Algoritma *Gaussian Naïve Bayes* merupakan algoritma yang menggunakan prinsip probabilitas untuk memprediksi kelas dari suatu objek berdasarkan fitur yang dimiliki [11]. Algoritma ini dapat menghitung probabilitas setiap kelas untuk objek tertentu dengan mengalikan probabilitas setiap fitur dengan

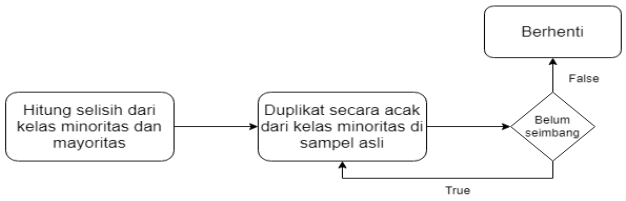


Gambar 4. Flowchart algoritma Gaussian Naive Bayes

mengalikan probabilitas kelas kategori dengan nilai probabilitas tertinggi kemudian dipilih sebagai kelas prediksi. Alur algoritma ini dapat dilihat pada gambar 4.

D. *Random Over Sampling*

Teknik atau algoritma *Random Over Sampling* (ROS) bekerja dengan cara menduplikat atau mereplikasi data sampel pada kelas minoritas sehingga berjumlah sama seperti pada kelas mayoritas [4], [7], [14]. Berdasarkan cara kerja ROS tersebut,



Gambar 5. Flowchart algoritma Random Over Sampling

kelompok mencoba membuat flowchart seperti pada gambar 5 di mana rumus untuk menghitung selisih adalah $Selisih = mayor - minor$.

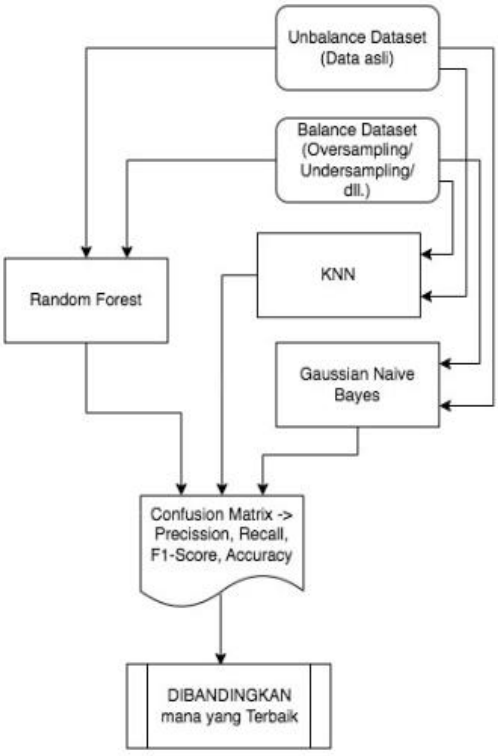
III. METODE PENELITIAN

Pada metode yang diusulkan, tahapan awal yaitu dengan menerapkan metode Random Forest, KNN dan Gaussian NB pada dataset asli (*Unbalance* dataset). Namun sebelum itu akan dilakukan *preprocessing* terhadap data Keladi Tikus yang dapat dilihat pada gambar 6. Adapun *preprocessing* yang dilakukan berupa penghapusan baris dan kolom dari dataset. Penghapusan atau memfilter data ini dilakukan terhadap data yang tidak

	Retention Time	Intensity	m/z	Real_m/z	NamaSenyawa	RumusSenyawa	Status
0	5.022	281952	60.057205	60.03236	Urea	CH4N2O	0
1	5.022	198032	60.246963	60.03236	Urea	CH4N2O	0
2	5.022	43064	66.888519	67.04220	Pyrrol	C4H5N	0
3	5.022	63172	67.268036	67.04220	Pyrrol	C4H5N	0
4	5.022	57852	68.280075	68.03745	Imidazole	C3H4N2	0
...	...	...	...	...	...	...	...
663223	2401.000	456	1188.677734	1188.42798	Man2XylManGlcNAcFucGlcNAc	C45H76N2O34	0
663224	2401.000	191	1191.460938	1191.38086	Foetoside C	C58H94O25	0
663225	2401.000	191	1193.358398	1193.40686	Man6GlcNAc-I	C44H75NO36	0
663226	2401.000	191	1195.192871	1195.35303	Furostane base -2H + O-Hex, O-Hex-dHex-dHex-dHex	C57H94O26	0
663227	2401.000	419	1195.572266	1195.35303	Furostane base -2H + O-Hex, O-Hex-dHex-dHex-dHex	C57H94O26	0

Gambar 6. Tabel Data Keladi Tikus

relevan atau tidak diperlukan dalam proses klasifikasi sehingga data dapat menjadi subset yang lebih relevan dan fokus pada variabel yang penting. Setelah itu dilakukan pemodelan data pada data *training* dengan tiap-tiap metode sebagai pembanding dan diuji dengan data uji melalui proses algoritma tiap metode. Untuk mengukur kinerja algoritma pengklasifikasian data seimbang dan tidak seimbang terdapat



Gambar 7. Diagram alir program

beberapa ketentuan dengan menggunakan *confusion matrix* untuk metode KNN yang diperoleh dari proses validasi menggunakan *K-fold cross validation* yaitu 814-fold dimana hasil tersebut diperoleh dari hasil akar jumlah dataset, lalu *test size* dengan menggunakan rasio 70:30 (0.3), dan *random state* sebesar 42. Setelah semua algoritma metode selesai dijalankan maka dapat kita peroleh hasil akurasi, presisi, f1 score, recal dan waktunya, dimana akan dibandingkan dengan hasil data yang telah diresampling.

Tahapan berikutnya dilakukan *resampling* dengan menggunakan metode *Random Under Sampling* bertujuan agar dataset asli menjadi lebih seimbang. Lalu dari hasil dataset yang telah diseimbangkan, dikombinasikan menjadi data training dan data uji dimana nantinya diuji pada tiap-tiap metode. Lalu menjalankan semua metode sesuai dengan algoritma tiap metode dan diperoleh hasil akurasi, presisi, f1 score, recall dan waktunya. setelah memperoleh hasil baik dari data *balance* maupun *unbalance* dapat kita bandingkan kualitas metode mana yang terbaik berdasarkan nilai akurasi, presisi, f1 score, recal dan waktunya. Secara singkat alur dari program dapat dilihat pada gambar 7.

IV. HASIL DAN DISKUSI

Data yang sudah dilakukan *preprocessing* (dapat dilihat pada gambar 8) akan dilakukan pengujian berdasarkan diagram alur pada gambar 7. Awalnya data *unbalance* akan dilakukan

	Retention Time	Intensity	m/z	Real_m/z
0	5.022	281952	60.057205	60.03236
1	5.022	198032	60.246964	60.03236
2	5.022	43064	66.888519	67.04220
3	5.022	63172	67.268036	67.04220
4	5.022	57852	68.280075	68.03745
...	...	...	...	...
663223	2401.000	456	1188.677734	1188.42798
663224	2401.000	191	1191.460938	1191.38086
663225	2401.000	191	1193.358398	1193.40686
663226	2401.000	191	1195.192871	1195.35303
663227	2401.000	419	1195.572266	1195.35303

Gambar 8. Hasil preprocessing data Keladi Tikus

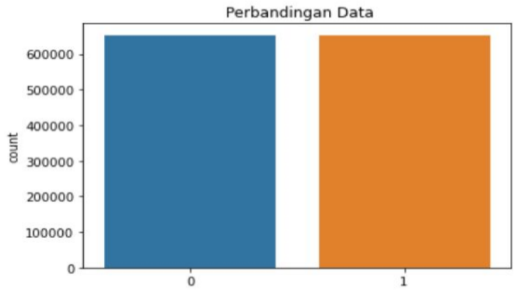
pemodelan data dan pengujian data menggunakan ketiga algoritma tersebut. Data awalnya dibagi menjadi data training dan data test dengan perbandingan 70:30. Dan nilai k yang digunakan pada algoritma KNN bernilai 814 dengan *random state* sebesar 42. Berdasarkan pembagian data tersebut diperoleh hasil seperti pada tabel 1. Pada tabel tersebut dapat dilihat bahwa akurasi yang dihasilnya hampir menyentuh angka 1 di mana dapat dipastikan rata – rata akurasi dari ketiga

TABEL I.
HASIL PENGUJIAN TERHADAP DATA UNBALANCE

Hasil Pengujian Terhadap Data Unbalance			
	Random Forest	**K – Nearest Neighbour**	**Gaussian Naïve Bayes**
Precision	0,993	0	0
Recall	0,342	0	0
F1 – Score	0,508	0	0
Accuracy	0,99	0,985	0,985
Time	50s	60s	0,16s

algoritma tersebut adalah 98%, namun untuk nilai perhitungan yang lain bernilai 0 untuk algoritma KNN dan Gaussian NB. Nilai 0 yang didapatkan pada perhitungan *precision, recall, dan f1-score* dikarenakan adanya eror berupa '*UndefinedMetricWarning'*. Eror ini memberikan peringatan bahwa *precision* tidak ditentukan dan disetel ke 0 karena tidak ada sampel yang diprediksi. Ini terjadi ketika model tidak memprediksi kelas tertentu. Peringatan ini biasanya ditampilkan dalam kasus di mana ada ketidakseimbangan kelas dalam data dan satu kelas atau lebih yang memiliki jumlah sampel yang sangat sedikit. Dalam situasi ini, model mungkin tidak dapat memprediksi kelas langka dan akurasinya tetap tidak terdefinisi karena pembaginya adalah 0. Selain itu, nilai *f1-score* yang mencerminkan sejauh mana model dapat

diklasifikasi secara tepat bernilai 0 yang berarti bahwa terdapat kesalahan pada data yang ingin diujikan. Kesalahan ini dikarenakan terjadinya ketidakseimbangan data seperti yang sudah dibahas sebelumnya.



Gambar 9. Diagram hasil data diseimbangkan

Setelah dilakukan pengujian terhadap data *unbalance*, data akan dijadikan *balance* dengan algoritma ROS. Hasil dari data *balance* yang dapat dilihat dari gambar 9. Kemudian dilakukan pemodelan data dan pengujian data dan mendapatkan hasil seperti pada tabel 2. Pada tabel 2, dapat dilihat bahwa terjadi penurunan akurasi terhadap algoritma KNN dan Gaussian NB di mana pada pengujian sebelumnya mencapai rata – rata 98% sedangkan pada data *balance* hanya mencapai rata – rata 54%. Namun jika diperhatikan untuk nilai *f1-score* terjadi peningkatan yang menandakan bahwa data menjadi lebih baik daripada sebelumnya dan data dapat dilakukan klasifikasi.

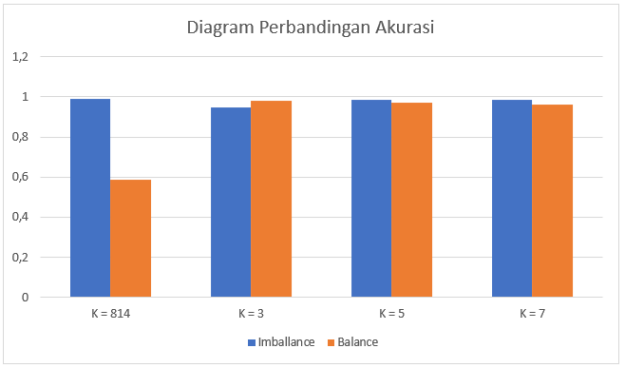
TABEL II.
HASIL PENGUJIAN TERHADAP DATA BALANCE

Hasil Pengujian Terhadap Data Balance			
	Random Forest	**K – Nearest Neighbour**	**Gaussian Naïve Bayes**
Precision	0,999	0,583	0,501
Recall	1	0,683	0,986
F1 – Score	1	0,629	0,664
Accuracy	1	0,598	0,501
Time	97s	130s	0,33s

TABEL III.
PERBANDINGAN HASIL PENGUJIAN DATA

	Unbalance Data			**Balance Data**		
	Random Forest	**KNN**	**Gaussian NB**	**Random Forest**	**KNN**	**Gaussian NB**
Precision	0,993	0	0	0,999	0,583	0,501
Recall	0,342	0	0	1	0,683	0,986
F1 – Score	0,508	0	0	1	0,629	0,664
Accuracy	0,99	0,985	0,985	1	0,598	0,501
Time	50s	60s	0,16s	97s	130s	0,33s

Terjadi penurunan pada kedua algoritma ini dikarenakan pada data sudah dilakukan penduplikatan pada kelas minoritas. Selain itu, penurunan akurasi juga dipengaruhi oleh nilai k pada KNN yang terlalu besar yang dapat menyebabkan terjadi overfitting dan underfitting pada model data yang digunakan. Juga hal ini akan menyebabkan waktu komputasi yang lebih besar daripada sebelumnya. Karena dari itu, disarankan untuk algoritma KNN sebaiknya menggunakan nilai k yang kecil



Gambar 10. Diagram perbanfingan akurasi tiap K

seperti nilai 3 atau 5 atau 7 yang mana dapat dilihat pada perbandingan di tabel 4 atau pada gambar 10 yang mendapatkan akurasi yang bagus pada data *balance*. Hal ini sekaligus membuktikan bahwa data *balance* tersebut sudah menghasilkan akurasi yang bagus dan ketika diujicobakan dengan data asli, dapat menghasilkan akurasi yang tinggi juga.

TABEL IV.
PERBANDINGAN NILAI K

	K = 3		**K = 5**		**K = 7**	
	Unbalance	**Balance**	**Unbalance**	**Balance**	**Unbalance**	**Balance**
Precision	0 948	0.979	0.985	0.969	0.985	0.960
Recall	0.138	0,960	0.261	0.942	0.472	0,926
F1 – Score	0.012	1	0.006	0.970	0.006	1
Accuracy	0.022	0.980	0012	15.94s	0.011	0.962
Time	7,50s	16.63s	8.37s	0.969	8.24s	15.79s

Sedangkan terjadi penurunan terhadap algoritma Gaussian NB dikarenakan algoritmanya sendiri kurang cocok jika melakukan klasifikasi terhadap data yang besar. Dari kedua hasil pengujian data sebelumnya dapat dibandingkan dari tabel 3. Kita dapat melihat perbedaan yang cukup signifikan terhadap algoritma KNN dan Gaussian NB. Sekilas dapat disimpulkan bahwa data *unbalance* merupakan data yang bagus, namun hal ini dapat dipatahkan ketika dimasukan data asli. Karena ketika data yang asli dimasukan akan menghasilkan akurasi yang rendah karena data yang dilatih merupakan data yang tidak seimbang. Sebaliknya, dengan data *balance,* walaupun mempunyai akurasi yang rendah, ketika dimasukan atau diujikan dengan data asli maka akan menghasilkan akurasi yang tidak turun drastis dan bahkan bisa mencapai akurasi yang tinggi.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Ketidakseimbangan data sudah sering terjadi pada pemodelan atau klasifikasi. Hal ini dapat menyebabkan terjadinya salah pemodelan. Sehingga pada penelitian ini telah dilakukan perbandingan mengenai akurasi, precision, recall dan F1-Score untuk data *unbalance* dan juga untuk data *balance* menggunakan random oversampling. Untuk menentukan algoritma terbaik yang dapat digunakan pada data yang didapat saat penelitian menunjukan bahwa algoritma yang lebih unggul yaitu Random Forest. Random Forest memiliki akurasi yang sangat tinggi untuk data *unbalance* maupun data yang *balance* sebesar 99% dan memiliki waktu komputasi yang cukup

singkat. Berdasarkan hasil penelitian ini juga menghasilkan kesimpulan bahwa implementasi algoritma Random Forest dan KNN memerlukan waktu komputasi yang lebih lama dibanding dengan algoritma Naïve Bayes.

B. Saran

Hasil penelitian ini dapat dikembangkan lebih lanjut pada penelitian-penelitian lainnya seperti menggunakan metode *random forest* dapat memproses data yang lebih besar dengan rasio ketidakseimbangan yang lebih tinggi sehingga dapat memberi nilai akurasi yang tinggi dengan waktu komputasi yang cukup singkat dan jika terdapat dataset yang banyak sebaiknya tidak menggunakan algoritma *Gaussian Naive Bayes* dikarenakan menghasilkan nilai akurasi, *precision* dan *f1-score* yang cukup rendah meskipun waktu komputasi sangat singkat. Juga menerapkan metode *ensemble learning* lainnya dengan *base classifier* yang berbeda dan beragam.

REFERENSI

[1] K. Pangesti Yudi Harhari, A. Medawati, D. *Hambat Ekstrak Etanol Daun Keladi Tikus*, P. Studi Pendidikan Dokter Gigi Fakultas Kedokteran dan Ilmu Kesehatan, G. Universitas Gadjah Mada, and D. Biomedis Program Studi Pendidikan Dokter Gigi Fakultas Kedokteran dan Ilmu Kesehatan Universitas Muhammadiyah Yogyakarta, "Daya Hambat Ekstrak Etanol Daun Keladi Tikus (Typhonium flagelliforme Lodd.) terhadap Proliferasi Sel Kanker Lidah Manusia (Sp-c1) secara In Vitro."

[2] N. F. Sianipar, R. Purnamaningsih, and Rosaria, "*Bioactive compounds of fourth generation gamma-irradiated Typhoniumflagelliforme Lodd. mutants based on gas chromatography-mass spectrometry*," in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics Publishing, Sep. 2016. doi: 10.1088/1755-1315/41/1/012025.

[3] R. Perangin-Angin, E. Julia, G. Harianja, and I. K. Jaya, "*Pendekatan Level Data Untuk Menangani Ketidakseimbangan Data Menggunakan Algoritma K-Nearest Neighbor*," 2020. [Online]. Available: https://sci2s.ugr.es/keel/datasets.php

[4] R. Dwi Fitriani, H. Yasin, D. Statistika, and F. Sains dan Matematika, "*Penanganan Klasifikasi Kelas Data Tidak Seimbang Dengan Random Oversampling Pada Naive Bayes (Studi Kasus: Status Peserta Kb Iud Di Kabupaten Kendal),"* vol. 10, no. 1, pp. 11–20, 2021.

[5] 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*."

[6] Institute of Electrical and Electronics Engineers, *ICEE 2012 : 20th Iranian Conference on Electrical Engineering : [Tehran, Iran, May 15-17, 2012].* [IEEE], 2012.

[7] A. Syukron and A. Subekti, "*Penerapan Metode Random Over-Under Sampling dan Random Forest untuk Klasifikasi Penilaian Kredit,*" *JURNAL INFORMATIKA*, vol. 5, no. 2, 2018.

[8] A. Syukron and A. Subekti, "*Penerapan Metode Random Over-Under Sampling dan Random Forest untuk Klasifikasi Penilaian Kredit,*" *JURNAL INFORMATIKA*, vol. 5, no. 2, 2018.

[9] Y. Azhar, G. A. Mahesa, and Moch. C. Mustaqim, "*Prediction of hotel bookings cancellation using hyperparameter optimization on Random Forest algorithm,*" *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 1, pp. 15–21, Jan. 2021, doi: 10.14710/jtsiskom.2020.13790.

[10] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "*Random Forests and Decision Trees,*" 2012. [Online]. Available: www.IJCSI.org

[11] R. N. Devita, H. W. Herwanto, and A. P. Wibawa, "*Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia*," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 4, p. 427, Oct. 2018, doi: 10.25126/jtiik.201854773.

[12] R. Perangin-Angin, E. Julia, G. Harianja, and I. K. Jaya, "*Pendekatan Level Data Untuk Menangani Ketidakseimbangan Data Menggunakan Algoritma K-Nearest Neighbor*," 2020. [Online]. Available: https://sci2s.ugr.es/keel/datasets.php

[13] "*Perbandingan Metode Naïve Bayes dan KNN (K-Nearest Neighbor) dalam Klasifikasi Penyakit Diabetes*."

[14] Institute of Electrical and Electronics Engineers, *ICEE 2012 : 20th Iranian Conference on Electrical Engineering : [Tehran, Iran, May 15-17, 2012].* [IEEE], 2012.