Jurnal RAMATEKNO, Vol. 5 No. 2, November 2025, hal. 150 – 158 ISSN 2828-2108
Makalah dikirim 1 Oktober 2025; Revisi 20 November 2025; Diterima 27 November 2025

PERBANDINGAN PERFORMA SMOTE DAN SMOTE-ENN PADA DATA TIDAK SEIMBANG

[1]Anak Agung Gde Wahyu Sukma Erlangga, [2]Pande Putu Ode Juliantara KW, [3]Adnan
[1,2,3]Teknologi Informasi, STMIK Bandung Bali, Denpasar
[1]wahyujungde@gmail.com, [2]odepande21@gmail.com, [3]adnandantong@gmail.com

Abstrak

Permasalahan ketidakseimbangan kelas pada data klasifikasi biner dapat menyebabkan model bias terhadap kelas mayoritas dan menurunkan performa dalam memprediksi kelas minoritas. Penelitian ini bertujuan untuk mengevaluasi pengaruh metode penyeimbangan data Synthetic Minority Over-sampling Technique (SMOTE) dan kombinasi SMOTE dengan Edited Nearest Neighbor (SMOTE-ENN) terhadap performa model klasifikasi. Dataset yang digunakan adalah Bank Marketing dengan variabel target berupa keputusan nasabah dalam menerima tawaran produk deposito. Metode yang digunakan meliputi pengumpulan data, preprocessing data, splitting data, resampling data, modeling dan evaluasi performa. Tiga algoritma machine learning yang digunakan dalam penelitian ini, yaitu Logistic Regression, Naïve Bayes, dan Random Forest. Evaluasi model dilakukan menggunakan metrik accuracy, sensitivity, specificity, dan G-Means. Hasil penelitian menunjukkan bahwa penerapan SMOTE meningkatkan sensitivity model terhadap kelas minoritas, namun cenderung menurunkan specificity. SMOTE-ENN memberikan hasil yang lebih seimbang antara sensitivitas dan spesifisitas karena mampu mengurangi data outlier dan overlapping. Model Random Forest dengan data hasil SMOTE-ENN menghasilkan performa paling seimbang secara keseluruhan berdasarkan keempat metric evaluasi. Penelitian ini menyimpulkan bahwa pemilihan metode penyeimbangan data yang tepat berpengaruh signifikan terhadap kualitas klasifikasi pada data tidak seimbang.

Kata kunci: *class imbalance, SMOTE, SMOTE-ENN, klasifikasi, machine learning*

Abstract

The problem of class imbalance in binary classification data can cause models to be biased toward the majority class and reduce performance in predicting the minority class. This study aims to evaluate the impact of the Synthetic Minority Over-sampling Technique (SMOTE) data balancing method and the combination of SMOTE with Edited Nearest Neighbor (SMOTE-ENN) on classification model performance. The dataset used is Bank Marketing, with the target variable being the customer's decision to accept a deposit product offer. The methods used include data collection, data preprocessing, data splitting, data resampling, modeling, and performance evaluation. Three machine learning algorithms were used in this study: Logistic Regression, Naïve Bayes, and Random Forest. Model evaluation was conducted using the accuracy, sensitivity, specificity, and G-Means metrics. The results of the study show that the application of SMOTE improves the model's sensitivity to the minority class but tends to reduce specificity. SMOTE-ENN provides a more balanced result between sensitivity and specificity because it can reduce outliers and overlapping data. The Random Forest model with SMOTE-ENN data produces the most balanced overall performance based on the four evaluation metrics. This study concludes that the selection of the appropriate data balancing method significantly affects the quality of classification in imbalanced data.

.

Keywords*: class imbalance, SMOTE, SMOTE-ENN, classification, machine learning*

.

1. PENDAHULUAN

Ketidakseimbangan kelas atau *class imbalance* merupakan fenomena di mana terdapat kelas yang mendominasi pada data yang disebut dengan kelas mayoritas, sedangkan kelas yang menempati ruang sempit disebut dengan kelas minoritas. Ketidakseimbangan kelas dapat menyebabkan model lebih bergantung pada kelas mayoritas dalam proses klasifikasi [1] , sehingga akan menghasilkan ketidakadilan dan bahkan generalisasi yang buruk [2] . Dengan demikian, hanya kelas mayoritas saja yang dapat diprediksi dengan mudah jika dibandingkan dengan kelas minoritas [3] . Selain itu, model juga bisa mengalami *overfitting* apabila dilatih dengan data tidak seimbang [4].

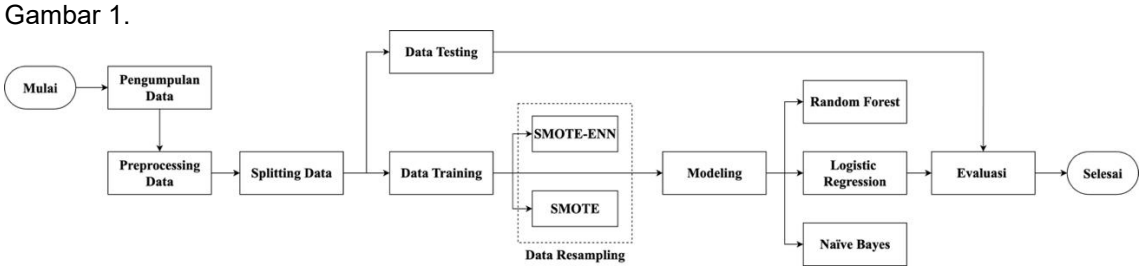
Untuk menyeimbangkan kelas dapat digunakan teknik *oversampling* atau *undersampling*. Namun, *overfitting* dapat terjadi apabila *oversampling* yang dilakukan terlalu banyak [5]. Di sisi yang lain penggunaan undersampling dapat mengakibatkan informasi penting dari datasets menjadi hilang [6], [7]. Untuk menangani itu kita bisa menggunakan teknik SMOTE (*Synthetic Minority Oversampling Technique*). SMOTE merupakan teknik *oversampling* yang bekerja dengan menyintesis data baru melalui interpolasi acak antara k tetangga terdekat dengan kelas minoritas [8], sehingga juga dapat mengurangi *overfitting* [9].

Sebelumnya SMOTE telah digunakan dalam penelitian yang mengangkat permasalahan keseimbangan kelas, salah satunya adalah penelitian [10] , dimana SMOTE berhasil meningkatkan performa model untuk memprediksi *lumpy skin disease* sebanyak 1-2%. Selain itu, penelitian yang dilakukan [11] yang menunjukan bahwa metode SMOTE dapat membantu meningkatkan kinerja dari algoritma machine learning dan algortima yang menghasilkan hasil terbaik adalah Random Forest dengan *accuracy* 89,72%. Lalu, penelitian [12] juga menemukan bahwa SMOTE dapat menaikan performa *accuracy* secara signifikan pada data penyakit diabetes yang diuji menggunakan algoritma ANN. Akan tetapi, SMOTE memiliki suatu kekurangan, yaitu dapat menyebabkan tumpang tindih kelas (*class overlapping*) [13] yang dapat memengaruhi performa model.

Oleh sebab itu, penelitian ini mengusulkan untuk mengkombinasikan SMOTE dengan teknik *undersampling* yaitu ENN (*Edited Nearest Neighbor*), dimana ENN yang berbasis *neighbourhood* dapat mendeteksi dan mengeliminasi instansi mayoritas di wilayah *overlap* antar kelas [14]. Data yang akan digunakan pada penelitian ini adalah data Bank Marketing dari UCI Machine Learning, dimana nantinya data yang telah diseimbangkan akan dilatih menggunakan beberapa algoritma, diantaranya Logistic Regression, Naïve Bayes, dan Random Forest. Lalu, akan dilakukan pengujian dengan *accuracy*, *sensitivity*, *specificity* dan *g-means* untuk mengetahui perbandingan performanya.

2. METODE PENELITIAN

Metode penelitian ini terdiri dari pengumpulan data, *preprocessing* data, *splitting* data, *resampling* data, *modeling*, dan evaluasi. Adapun skema dari penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. *Metode penelitian*

.

2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data Bank Marketing yaitu data *telemarketing* atau kampanye pemasaran langsung yang didapatkan dari *website* UCI Machine Learning dengan keseluruhan data berjumlah 41188 data. Data ini terdiri dari 20 fitur dan sebuah variabel dependen. Adapun setiap fitur dari data ini dapat dilihat pada Tabel 1.

Tabel 1. *Fitur dataset bank marketing*

Atribut	Informasi
age	Umur dari nasabah
job	Jenis pekerjaan dari nasabah
marital	Status perkawinan dari nasabah
education	Pendidikan terkahir dari nasabah
default	Memiliki kredit atau tidak
housing	Memiliki kredit perumahan atau tidak
loan	Mimiliki pinjaman pribadi atau tidak
contact	Jenis komunikasi yang digunakan
month	Bulan kontak terakhir tahun ini
day_of_week	Hari kontak terkahir minggu ini
duration	Durasi kontak
campaign	Jumlah kontak yang telah dilakukan selama kampanye ini terhadap nasabah terkait.
pdays	Jumlah hari yang berlalu setelah klien terakhir dihubungi dari kampanye sebelumnya
previous	Jumlah kontak yang dilakukan sebelum kampanye ini untuk klien terkait
poutcome	Hasil dari kampanye pemasaran sebelumnya
emp.var.rate	Tingkat variasi pekerjaan - indikator triwulanan
cons.price.idx	Indeks harga konsumen - indikator bulanan
cons.conf.idx	Indeks kepercayaan konsumen - indikator bulanan
euribor3m	Tingkat triwulan Euribor – indikator harian
nr.employed	Jumlah karyawan - indikator triwulanan
y	Klasifikasi apakah nasabah berlangganan deposito berjangka atau tidak

2.2 *Preprocessing* Data

Pada tahap ini akan dilakukan beberapa *preprocessing* pada data diantaranya penanganan *missing value*, melakukan *encoding* terhadap fitur kategorikal, memisahkan atribut dan variabel target serta melakukan standarisasi agar rentang nilai dari data menjadi seragam.

.

2.3 *Splitting* Data

Pada proses data akan dibagi menjadi data latih dan data uji, dimana pada penelitian ini data akan dibagi menjadi 80% data latih dan 20% data uji. Data latih digunakan untuk melatih model sedangkan data uji akan digunakan untuk menguji model yang dihasilkan dari pelatihan.

2.4 Resampling Data

Data latih yang sudah kita dapatkan dari proses sebelumnya akan diseimbangkan dengan teknik SMOTE. SMOTE merupakan teknik *oversampling* yang menyeimbangkan kelas dengan cara menghasilkan data sintesis dengan memilih data acak pada minoritas dan di antara titik terpilih dan tetangganya akan dihasilkan data sintetik menggunakan metode interopolasi [5], [11]. Adapun persamaan (1) merupakan persamaan dari SMOTE.

$$x_{syn} = x_i + rand(0,1) \times (x_{knn} - x_i) \qquad (1)$$

Selain itu, data juga akan diresampling menggunakan teknik gabungan dari SMOTE-ENN, dimana Dimana ENN adalah teknik *undersampling* yang akan menghapus sampel yang memiliki label kelas yang berbeda dengan mayoritas dari tetangga terdekatnya sebanyak k [6], [15], [16].

2.5 Modeling

Pada tahap ini data yang telah diseimbangkan baik menggunakan SMOTE maupun SMOTE-ENN dan data *original* akan dilatih menggunakan algoritma *machine learning* yaitu Logistic Regression, Naïve Bayes, dan Random Forest. Ketiga algoritma digunakan karena merupakan yang umum digunakan dalama menangani masalah klasifikasi. Logistic Regression merupakan biasa digunakan untuk melakukan klasifikasi biner atau *binary classification*, dimana algoritma ini berguna untuk memperkirakan probabilitas posterior dari setiap kelas [17] . Naïve Bayes, meskipun mengasumsikan independensi kondisi antar fitur [18] , namun algoritma ini masih termasuk dalam top 10 algoritma *machine learning* untuk melakukan klasifikasi [19] . Sedangkan Random Forest merupakan algoritma klasifikasi yang terdiri dari sekumpulan *decision tree* (pohon keputusan), dimana cocok untuk menangani data yang memiliki banyak fitur atau data yang berdimensi tinggi [20].

2.6 Evaluasi

Untuk mengetahui performa dari model yang telah dihasilkan sebelumnya, maka model harus dievaluasi. Pada penelitian ini model akan dievaluasi menggunakan *accuracy*, *sensitivity*, *specificity* dan *g-means*. Adapun persamaan untuk keempat *metrics* evaluasi itu dapat dapat dilihat pada persamaan (2), (3), dan (4), dimana nilainya diambil dari *confusion matrix* yaitu *true positive* (TP), *false positive* (FP), *true negative* (TN) dan *false negative* (FN).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

$$Sensitivity = \frac{TP}{TP + FN} \qquad (3)$$

$$Specificity = \frac{TN}{TN + FP} \qquad (4)$$

$$G - Means = \sqrt{sensitivity \times specificity} \qquad (5)$$

.

3. HASIL DAN PEMBAHASAN

Pada bagian ini akan dijelaskan bagaimana *output* yang dihasilkan dari penelitian ini dengan tujuan untuk melihat efek penggunaan teknik SMOTE dan juga SMOTE-ENN terhadap peforma model yang dihasilkan pada setiap algoritma *machine learning* yang diuji. Pertama-tama akan dilakukan *preprocessing* terhadap *datasets bank marketing* yang telah kita ambil dari *website* UCI Machine Learning, sehingga data dapat dibagi menjadi data latih dan data uji. Pada penelitian ini data dibagi dengan proporsi 80% data latih dan 20% data uji, dimana jumlahnya dapat dilihat pada pada Tabel 2.

Tabel 2. *Hasil data splitting*

Jenis Data	Jumlah Data
Data Latih	32950
Data Uji	8238

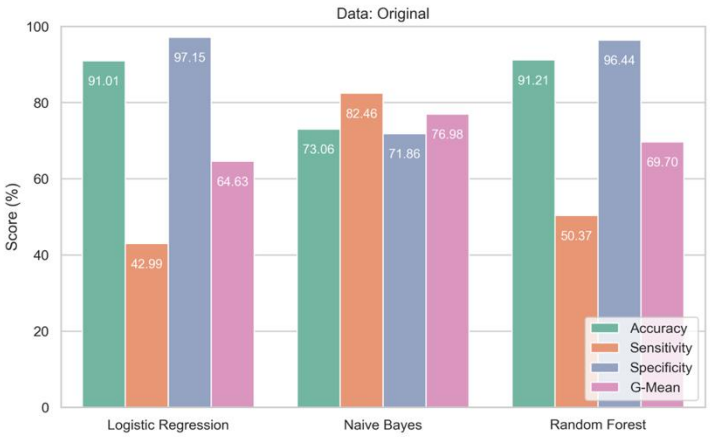
Selanjutnya, untuk mengetahui bagaimana perbandingan performa model, maka akan dilakukan tiga jenis pengujian. Pada pengujian pertama akan menggunakan data latih original tanpa proses *resampling* untuk setiap algortima. Lalu, untuk pengujian kedua dan ketiga akan digunakan data latih yang telah diseimbangkan baik dengan teknik SMOTE dan kombinasi SMOTE-ENN.

Tabel 3. *Perbandingan kelas sebelum dan sesudah resampling*

Datasets	Tidak Berlangganan	Berlangganan
Original	29245	3705
SMOTE	29245	29245
SMOTE-ENN	24597	28179

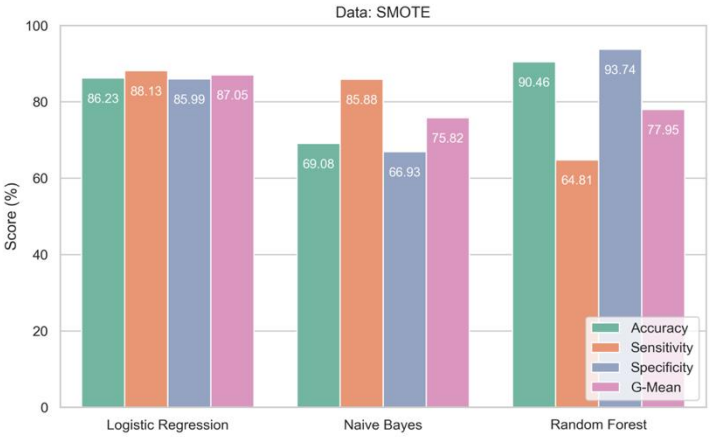
Tabel 3 menunjukan bahwa kelas yang tidak seimbang pada *datasets original* dapat benar-benar diseimbangkan dengan bantuan dari teknik SMOTE. Selanjutnya, pada SMOTE-ENN dapat kita lihat terdapat data-data yang dihapus oleh ENN sehingga kelasnya tidak secara presisi memiliki jumlah yang sama, akan tetapi perbedaan *ratio* kelas yang rendah tidak akan mempengaruhi performa model secara signifikan Prati dkk., dalam [21].

Pada Gambar 2, hasil percobaan pertama menggunakan data latih original menunjukan bahwa algoritma Logistic Regression dan Random Forest mimiliki *accuracy* yang tinggi (91.01% dan 91.21%). Namun, keduanya memiliki performa yang rendah dalam mengenali kelas minoritas yang ditunjukan melalui *sensitivity* yang rendah (42.99% dan 50.37%), serta *G-Means*-nya yang hanya sebesar 64.63% dan 69.70%. Sebaliknya, algortima Naïve Bayes menunjukan performa yang lebih seimbang yang dapat dilihat dari *sensitivity* dan *specificity*-nya yang tidak terlalu berbeda jauh dan memiliki *g-means* tertinggi (76.98%) dibandingkan dengan algoritma lainnya, meskipun memiliki akurasi yang lebih rendah (73.06%). Hal ini menegaskan bahwa *accuracy* saja tidak cukup dalam kasus klasifikasi data tidak seimbang.



Gambar 2. *Hasil pengujian data original*

Selanjutnya, Gambar 3 menunjukan hasil percobaan kedua menggunakan data latih yang telah diseimbangkan menggunakan SMOTE. Terlihat bahwa terjadi penurunan *accuracy* pada semua algoritma, akan tetapi *sensitivity* meningkat secara signifikan pada seluruh model terutama pada algoritma Logistic Regression (88.13%) dan Random Forest (64.81%). Hal itu bisa menjadi indikasi bahwa SMOTE dapat meningkatkan performa model dalam mengenali kelas minoritas. *G-means* yang meningkat (87.05% dan 77.95%) juga menjadi indikasi bahwa kemampuan model dalam memprediksi sudah lebih seimbang. Namun, peningkatan *sensitivity* juga disertai dengan penurunan *specificity*, menandakan adanya potensi *overlapping* antar kelas akibat penambahan data sintetis yang terlalu dekat dengan batas kelas.

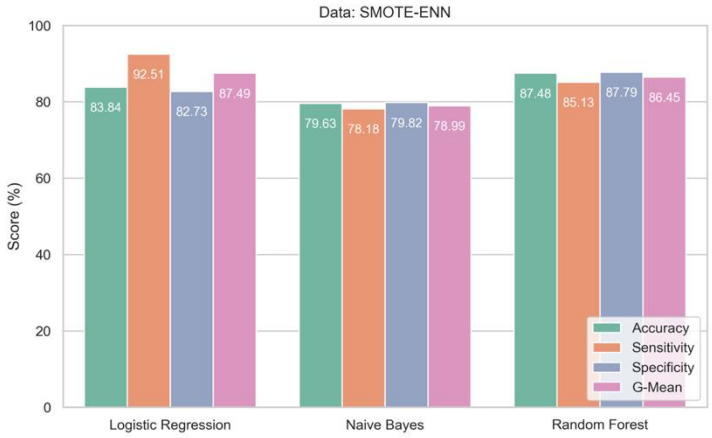


Gambar 3. *Hasil pengujian SMOTE*

Pada percobaan terakhir yang mengkombinasikan teknik SMOTE dan ENN menghasilkan hasil terbaik yang ditunjukan Gambar 4. Terlihat terjadi peningkatan keseimbangan prediksi antar kelas dibandingkan pada percobaan kedua. Logistic Regression tercatat memiliki sensitivity (92.51%) dan *g-means* (87.49%) tertinggi, yang dapat menjadi indikasi bahwa model memiliki kemampuan yang baik dalam mengenali kelas minoritas tanpa mengorbankan *specificity* secara signifikan. Naïve Bayes menunjukan performa yang stabil dengan sensitivity dan specificity yang relatif seimbang (78.18% dan 79.82%), serta *g-means*

.

sebesar 78.99%. Sementara itu Random Forest tetap memiliki akurasi yang tinggi (87.48%) dengan peningkatan *sensitivity* menjadi 85.13% dan *g-means* sebesar 86.45%.



Gambar 4. *Hasil pengujian SMOTE-ENN*

Beradasarkan tiga pengujian yang telah dilakukan, kombinasi dari SMOTE-ENN dapat meningkatkan kembali performa dari algortima machine learning, sehingga dapat melakukan prediksi dengan lebih seimbang baik untuk kelas mayoritas dan minoritas. Temuan ini juga dapat menjadi bukti bahwa ENN memiliki peran yang penting dalam membersihkan data yang berpotensi overlapping atau noise, yang mungkin muncul dari penerapan SMOTE. Bisa kita katakan, ENN dapat membantu dalam memperjelas batas antar kelas sehingga algoritma dapat mempelajari data dengan lebih baik.

4. KESIMPULAN

Penelitian ini telah menguji dan mengevaluasi performa tiga algortima klasifikasi yaitu, Logistic Regression, Naïve Bayes, dan Random Forest pada datasets bank marketing original, diseimbangan dengan SMOTE, dan diseimbangkan dengan SMOTE-ENN. Adapun beberapa kesimpulan yang dapat diambil sebagai berikut:

1. Teknik atau algortitma SMOTE efektif dalam meningkatkan sensitivity model terhadap kelas minoritas, namun memiliki potensi dalam menghasilkan data yang tumpang tindih (overlapping), sehingga malah bisa mengaburkan batas antar kelas.
2. Kombinasi SMOTE-ENN terbukti mampu meningkatkan performa model secara keseluruhan, dengan *g-means* tertinggi oleh algoritma Logistic Regression, tetapi secara keseluruhan Random Forest memiliki performa paling seimbang untuk setiap *metric* pengujian.
3. Temuan ini juga memberikan bukti bahwa ENN berperan penting dalam membersihkan data dari data yang berpotensi merupakan data *noise* dan *overlap*, sehingga dapat memperbaiki kualitas data hasil SMOTE dan meningkatkan kemampuan generalisasi model.

Dengan demikian, penelitian ini menekankan pentingnya strategi penanganan data tidak seimbang yang tepat, tidak hanya dari sisi pemilihan algoritma tetapi juga dalam teknik resampling yang dipilih untuk menanganinya. Kemudian saran untuk penelitian selanjutnya, dapat dilakukan *hyperparameter tuning* pada algoritma yang digunakan ataupun menguji algoritma metode menggunakan algoritma *machine leaerning* lainnya.

DAFTAR PUSTAKA

[1] S. Mutmainah, “Penanganan Imabalance Data pada Klasifikasi Kemungkinan Penyakit Stroke,” Yogyakarta, 2021. [Daring]. Tersedia pada: https://library.uii.ac.id/osr
[2] O. Wu, “Rethinking class imbalance in machine learning,” *arXiv preprint arXiv:2305.03900*, 2023.
[3] C. Kaope dan Y. Pristyanto, “The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance,” *Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 2, hlm. 227–238, 2023, doi: 10.30812/matrik.v22i2.2515.
[4] W. A. Kurniawan dan A. Salam, “Penggunaan Feature Space SMOTE Untuk Mengurangi Overfitting Akibat Imbalance Dataset,” *Jurnal Transformatika*, vol. 22, no. 2, hlm. 140–149, Jan 2025, doi: 10.26623/transformatika.v22i2.8305.
[5] P. Kaur dan A. Gosain, “Comparing The Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise,” dalam *Advances in Intelligent Systems and Computing*, Springer Verlag, 2018, hlm. 23–30. doi: 10.1007/978-981-10-6602-3_3.
[6] H. Guo, X. Diao, dan H. Liu, “Embedding undersampling rotation forest for imbalanced problem,” *Comput Intell Neurosci*, vol. 2018, 2018, doi: 10.1155/2018/6798042.
[7] A. Guzmán-Ponce, R. M. Valdovinos, J. S. Sánchez, dan J. R. Marcial-Romero, “A New Under-Sampling Method to Face Class Overlap and Imbalance,” *Applied Sciences (Switzerland)*, vol. 10, no. 15, Agu 2020, doi: 10.3390/app10155164.
[8] F. Wang, M. Zheng, X. Hu, H. Li, T. Wang, dan F. Chen, “FIAO: Feature Information Aggregation Oversampling for imbalanced data classification,” *Appl Soft Comput*, vol. 161, hlm. 111774, 2024, doi: https://doi.org/10.1016/j.asoc.2024.111774.
[9] H. Cai, S. Shen, Q. Lin, X. Li, dan H. Xiao, “Predicting the Energy Consumption of Residential Buildings for Regional Electricity Supply-Side and Demand-Side Management,” *IEEE Access*, vol. 7, hlm. 30386–30397, 2019, doi: 10.1109/ACCESS.2019.2901257.
[10] S. Suparyati, Emma Utami, dan Alva Hendi Muhammad, “Applying Different Resampling Strategies In Random Forest Algorithm To Predict Lumpy Skin Disease,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, hlm. 555–562, Agu 2022, doi: 10.29207/resti.v6i4.4147.
[11] N. Santoso, W. Wibowo, dan H. Himawati, “Integration of Synthetic Minority Oversampling Technique for Imbalanced Class,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 1, hlm. 102–108, Jan 2019, doi: 10.11591/ijeecs.v13.i1.pp102-108.
[12] C. B. Handoko dan C. S. K. Aditya, “Penerapan Teknik SMOTE Dalam Mengatasi Imbalance Data Penyakit Diabetes Menggunakan Algoritma ANN,” *Smart Comp: Jurnalnya Orang Pintar Komputer*, vol. 14, no. 1, Jan 2025, doi: 10.30591/smartcomp.v14i1.7045.
[13] N. P. Y. T. Wijayanti, E. N. Kencana, dan I. W. Sumarjaya, “SMOTE: Potensi dan Kekurangannya Pada Survei,” *E-Jurnal Matematika*, vol. 10, no. 4, hlm. 235, Nov 2021, doi: 10.24843/mtk.2021.v10.i04.p348.
[14] P. Vuttipittayamongkol dan E. Elyan, “Neighbourhood-based undersampling approach for handling imbalanced and overlapped data,” *Inf Sci (N Y)*, vol. 509, hlm. 47–70, 2020, doi: https://doi.org/10.1016/j.ins.2019.08.062.
[15] M. Bach, A. Werner, dan M. Palt, “The proposal of undersampling method for learning from imbalanced datasets,” dalam *Procedia Computer Science*, Elsevier B.V., 2019, hlm. 125–134. doi: 10.1016/j.procs.2019.09.167.
[16] Z. Xu, D. Shen, T. Nie, dan Y. Kou, “A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data,” *J Biomed Inform*, vol. 107, Jul 2020, doi: 10.1016/j.jbi.2020.103465.
[17] Y. Li, N. Adams, dan T. Bellotti, “A Relabeling Approach to Handling the Class Imbalance Problem for Logistic Regression,” *Journal of Computational and Graphical Statistics*, vol. 31, no. 1, hlm. 241–253, 2022, doi: 10.1080/10618600.2021.1978470.
[18] R. Rachman, R. N. Handayani, dan I. Artikel, “Klasifikasi Algoritma Naive Bayes Dalam Memprediksi Tingkat Kelancaran Pembayaran Sewa Teras UMKM,” *JURNAL INFORMATIKA*, vol. 8, no. 2, 2021, [Daring]. Tersedia pada: http://ejournal.bsi.ac.id/ejurnal/index.php/ji

.

[19] X. Wu *dkk.*, “Top 10 algorithms in data mining,” *Knowl Inf Syst*, vol. 14, no. 1, hlm. 1–37, 2008, doi: 10.1007/s10115-007-0114-2.
[20] D. Ghosh dan J. Cabrera, “Enriched Random Forest for High Dimensional Genomic Data,” *IEEE/ACM Trans Comput Biol Bioinform*, vol. 19, no. 5, hlm. 2817–2828, 2022, doi: 10.1109/TCBB.2021.3089417.
[21] F. Thabtah, S. Hammoud, F. Kamalov, dan A. Gonsalves, “Data imbalance in classification: Experimental evaluation,” *Inf Sci (N Y)*, vol. 513, hlm. 429–441, 2020, doi: https://doi.org/10.1016/j.ins.2019.11.004.

.