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 KNearest Neighbor untuk Data Tidak Seimbang dan Data yang diseimbangkan dengan Metode Adaptive Synthetic pada Dataset LCMS Tanaman Keladi Tikus

Agustina Putri Monika[1],Felisia Elvira Paska Risti[2], Iwan Binanto[3*], Nesti F. Sianipar[4]

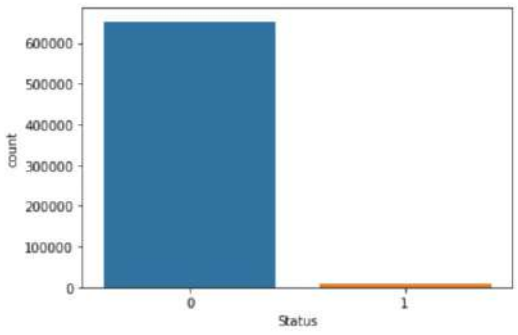
[1,2,3]*Informatika, Universitas Sanata Dharma, Yogyakarta*
[4]*Biotechnology Department, Faculty of Engineering, Bina Nusantara University*
[1]agustina.putrimonica@gmail.com
[2]elvirafelisia@gmail.com
[3]*iwan@usd.ac.id
[4]nsianipar@binus.edu

***Abstrak*-- Tujuan dari penelitian ini adalah membandingkan efektivitas penggunaan teknik data seimbang (balance) menggunakan Adaptive Synthetic (ADASYN) dengan metode klasifikasi K-Nearest Neighbors (KNN), Random Forest (RF), dan Gaussian Naive Bayes (GNB), serta membandingkan performa metode klasifikasi tersebut pada data yang tidak seimbang (unbalanced). Penelitian ini menggunakan data Keladi Tikus yang berasal dari penelitian sebelumnya. Dalam penelitian ini, dilakukan dua eksperimen terpisah. Pertama, ADASYN diterapkan untuk mendapatkan dataset yang seimbang, kemudian model KNN, RF, dan GNB dilatih dan diuji pada dataset tersebut. Kedua, dataset yang tidak seimbang digunakan, model KNN, RF, dan GNB kembali dilatih dan diuji pada dataset tersebut. Hasil penelitian menunjukkan bahwa KNN dan GNB menunjukkan kinerja yang kurang memuaskan baik pada data seimbang menggunakan ADASYN maupun pada data tidak seimbang. Ini menunjukkan bahwa kedua algoritma tersebut tidak efektif saat digunakan bersamaan dengan teknik oversampling ADASYN pada dataset Keladi Tikus. Namun, RF terbukti memiliki ketahanan baik pada data yang seimbang maupun tidak seimbang. Pada data yang seimbang, RF mencapai akurasi hingga 0.985, namun masih memiliki kekurangan dalam menguji data positif, terlihat dari hasil recall dan f1-score yang rendah. Sementara itu, pada data yang tidak seimbang, akurasi RF sedikit menurun menjadi 0.896, namun dataset tersebut memberikan hasil yang lebih seimbang dalam menguji data positif maupun negatif. Hal ini terbukti dari hasil yang relatif seimbang, yaitu presisi sebesar 0.881, recall sebesar 0.917, dan f1-score sebesar 0.898.**

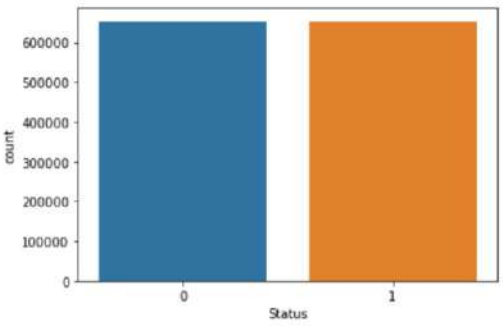
Katakunci: **Klasifikasi, Imbalance Data, Keladi Tikus, LCMS, ADASYN, Adaptive Synthetic**

I. PENDAHULUAN

Saat ini, klasifikasi data yang tidak seimbang telah menjadi salah satu masalah utama dalam pembelajaran mesin. Himpunan data dianggap tidak seimbang jika pada data pelatihan satu kelas memiliki dominasi yang sangat besar dibandingkan dengan kelas lainnya. Bahkan di beberapa kasus pengklasifikasi multi-kelas, ketidakseimbangan data ini menghasilkan representasi data yang rendah, dan pada akhirnya data ini cenderung diabaikan [1].



Gambar 1. Data tidak seimbang



Gambar 2. Data seimbang

Dalam dua dekade terakhir, penanganan data tidak seimbang telah menjadi fokus penelitian yang penting untuk menghasilkan data yang akurat [2], [3]. Salah satu data tidak seimbang didapat pada penelitian Binanto et al. [4] yang

merupakan data LCMS dari tanaman Keladi Tikus hasil penelitian Sianipar et. al.[5] Data ini tidak seimbang karena target biner yang menyatakan senyawa anti kanker dan senyawa biasa sangat kontras seperti terlihat pada gambar 1.

Data yang tidak seimbang dapat menyebabkan ketidakakuratan dalam klasifikasi menggunakan *machine learning*. Namun, terdapat algoritma klasifikasi yang dapat mengatasi masalah ini, yaitu Random Forest [6].

Salah satu pendekatan untuk menyeimbangkan data tidak seimbang adalah melalui algoritma *oversampling*, seperti algoritma ADASYN. Algoritma ADASYN dipilih karena dapat menghasilkan sampel sintetis pada kelas minoritas secara adaptif dengan tingkat kepentingan yang lebih rendah, dan juga implementasinya relatif mudah[7].

Dalam penelitian ini, akan dilakukan eksperimen pada data yang tidak seimbang dan data yang seimbang menggunakan algoritma klasifikasi Random Forest, KNN, dan Gaussian NB menggunakan data tanaman keladi tikus. Data yang tidak seimbang akan dilakukan penyeimbangan menggunakan teknik *oversampling* ADASYN.

Gambar diatas menunjukkan jumlah kelas klasifikasi tidak seimbang dan seimbang yang akan digunakan dalam studi. Tujuan studi ini adalah membandingkan performa algoritma Random Forest, KNN, dan Gaussian NB pada data yang telah seimbang menggunakan ADASYN, serta pada data yang tidak seimbang. Hasil penelitian ini diharapkan dapat membantu praktisi dan peneliti dalam memilih algoritma yang paling cocok untuk karakteristik data yang seimbang dan tidak seimbang.

II. STUDI PUSTAKA

A. *Ketidakseimbangan Kelas*

Ketika melakukan penelitian menggunakan suatu data, seringkali menemukan ketidakseimbangan pada data kelas yang ingin di klasifikasikan. Hal ini dapat menyebabkan hasil yang didapatkan oleh algoritma menjadi buruk[8]. Pada kasus yang digunakan oleh Bagga et al., memberikan contoh bawha jika ada 10 transaksi deteksi *fraud* di antara 10.000 transaksi *non-fraud* maka akan memberi akurasi hingga 99,999% karena model akan memprediksi seluruh transaksi menjadi *non fraud* walaupun sebenarnya terdapat data yang terdeterksi *fraud*[9].Sehingga penting bagi kita untuk memahami alasan di balik ketidakseimbangan dalam data atau kelas karena hal ini bisa berdampak besar terhadap demokrasi data dan bisa menyebabkan masalah yang serius[10].

B. *ADASYN*

ADASYN (Adaptive Synthetic Sampling) adalah metode *oversampling* yang bertujuan untuk mengatasi ketidakseimbangan data dalam tugas klasifikasi. Ide dasarnya yaitu memberikan bobot yang berbeda pada contoh-contoh kelas berdasarkan tingkat kesulitan. Kelas minoritas akan diisi dengan sample sintetis sehingga menghasilkan sampel kelas yang relatif seimbang [11]. Seperti penelitian yang dilakukan oleh Magnolia, Cindy et al., Pengujian dengan data imbalance menunjukkan akurasi dan f1-score berturut-turut 0.7 dan 0.7. Sedangkan untuk data balance dengan ADASYN menunjukan akurasi dan f1-score berturut-turut 0.9 dan 0.9 [12]. Proses ADASYN [13]:

a. Mengevaluasi tingkat ketidakseimbangan dari semua kelas

$$d = \frac{m0}{m1}, d \in (0,1) \quad (1)$$

b. Menghitung jumlah sampel yang akan dihasilkan,

$$G = (m1 - m0) * \beta,\ \beta \in [0,1] \quad (2)$$

mewakili tingkat ketidakseimbangan yang diharapkan setelah pembangkitan data. Jika β = 1, berarti sampel kelas sepenuhnya seimbang setelah pembangkitan data.

c. Untuk setiap sampel dari kelas minoritas, cari K tetangga terdalam dalam ruang berdimensi n.

$$Ii = \frac{\Delta i}{k\ (i = 1,2,\ldots,m)} \quad (3)$$

Δi = jumlah sampel yang termasuk ke dalam kelas mayoritas dan juga merupakan tetangga terdekat ke-k dari sampel xi. Ii ∈ [0,1].

d. Normalisasikan nilai Ii dengan

$$Ii = \frac{Ii}{\sum Ii} \quad (4)$$

Ii menjadi distribusi probabilitas dengan total ∑Ii = 1.

e. Hitung jumlah sampel xi dalam kelas minoritas yang akan dihasilkan, gi = Ii × G.

f. Pilih sampel dari k tetangga terdekat dalam kelas minoritas. Buatlah sampel sintetis baru Sz, di mana $Sz = (x + x) \times \lambda$, dengan λ adalah bilangan acak antara 0 dan 1.

g. Ulangi langkah f sebanyak i kali untuk mendapatkan sampel ke-i.

C. *KNN*

Algoritma KNN memiliki ketergantungan dengan jumlah k yang digunakan dalam menemtukan hasil datanya. Menurut penelitian dari Andre, Jeremy et al., menggunakan jumlah k yang berbeda dari rentang 1 sampai 30 bilangan ganjil didapatkan jumlah k optimal menggunakan k = 23[14]. Ide di balik algoritma KNN adalah untuk mengklasifikasikan data individu berdasarkan mayoritas dari tetangga terdekatnya [11]. Dalam fase pengambilan keputusan, algoritma ini mengembangkan cakupan tetangga terdekat menggunakan nilai k. Hal itu membuat algoritma KNN memperoleh lebih banyak informasi [12]. Algoritma KNN:

a. Menentukan parameter K (jumlah tetangga)

b. Menghitung jarak dimana salah satunya dapat menggunakan *Euclidean*. Rumus *Euclidean* :

$$d = \sqrt{((x2 - x1)^2 + (y2 - y1)^2 + \cdots + (xn - xn - 1)^2)} \quad (5)$$

(x1, y1, … xn-1) = koordinat yang akan di klasifikasi
(x1, y1, … xn-1) = koordinat tetangga terdekat

c. Urutkan data pada langkah kedua dengan urutan *ascending* atau menaik

d. Ambil mayoritas kelas berdasarkan jumlah K yang sudah ditentukan pada langkah a, maka objek dapat di prediksi

D. *Gaussian Naïve Bayes*

GaussianNB adalah sebuah model klasifikasi yang menggunakan algoritma *Naïve Bayes* dengan konsep probabilitas maksimum. Model ini digunakan untuk mempresiksi probabilitas keanggotaan suatu kelas. Fitur-fitur dari data direpresentasikan dalam bentuk vektor dan label kelas diberikan pada *instance* menggunakan model probabilistik. Dalam model ini, setiap fitur dalam data pelatihan dianggap saling bebas. Salah satu keterbatasanya ketika jumlah fitur bertambah, maka model akan memiliki bias pada tabel probabilitas [15]. Algoritma Gaussian NB :

a. Menghitung probabilitas prior (P(C)) :

P(C) = jumlah sampel dengan kelas C/ total sampel

b. Hitung probabilitas *likelihood* setiap fitur :

$$P(X|C) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) * e^{\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)} \quad (6)$$

X = nilai fitur yang diamati
μ = rata-rata fitur kelas C
σ^2 = varians untuk fitur kelas C

c. Hitung probabilitas posterior dengan rumus Teorema Bayes

$$P(C|X) = (P(X|C) * P(C))/P(X) \quad (7)$$

P(C|X) = probabilitas posterior kelas C
P(X|C) = probabilitas *likelihood* fitur X dalam kelas C
P(C) = probabilitas prior kelas C
P(X) = probabilitas fitur X

d. Hasil klasifikasi baru didapat dari instance tertinggi

E. *Random Forest*

Random forest adalah salah satu algoritma pembelajaran terawasi. Cara kerjanya dengan membangun sejumlah besar pohon keputusan saat proses pelatihan dan kemudian mengambil hasil kelas yang paling sering muncul dari pohon-pohon tersebut[16]. Random forest memiliki ketahanan terhadap data tidak seimbang, seperti pada penelitian Syukron, Ahmad et al,. mengenai penilaian kartu kredit mengatakan bahwa random forest dapat meningkatkan kinerja efektif akurasi di data tidak Seimbang.[17]

F. *Evaluasi Performa*

Dalam melakukan evaluasi performa menggunakan beberapa matrix evaluasi, yaitu :

$$\text{accuracy} : \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$\text{precision} : \frac{TP}{TP+FP} \quad (9)$$

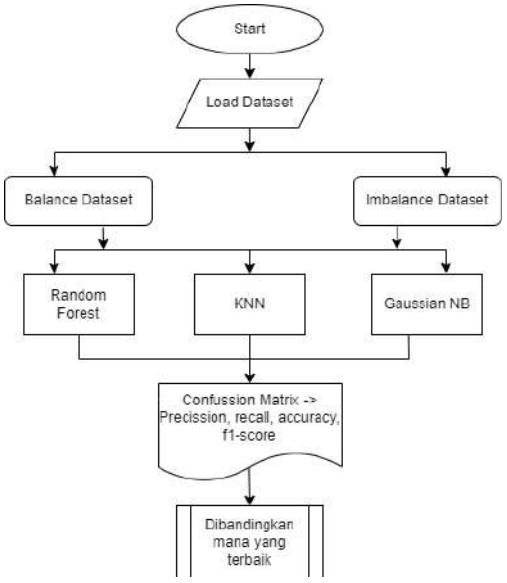
$$\text{recall} : \frac{TP}{TP+FN} \quad (10)$$

$$\text{f1-score} : 2 * \frac{precision*recall}{precision+recall} \quad (11)$$

TP (True Positive) : Jumlah prediksi positif yang benar
FP (False Positive) : Jumlah prediksi positif yang salah
TN (True Negative) : Jumlah prediksi negaif yang benar
FN (False Negative) : Jumlah prediksi negatif yang salah

III. METODE PENELITIAN

Langkah pertama adalah mengumpulkan data dan kemudian membaginya menjadi dua kategori, yaitu data yang seimbang dan data yang tidak seimbang. Data seimbang diperoleh dari data yang tidak seimbang dan diberlakukan teknik *oversampling* ADASYN untuk menghasilkan data sintetis pada kelas minoritas. Tujuannya adalah menjaga proporsi kelas yang realistis.



Gambar 3. Metode Penelitian

Selanjutnya, kedua jenis data, yaitu data seimbang dan tidak seimbang, dibagi menjadi subset pelatihan dan pengujian. Algoritma K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), dan Random Forest dilatih menggunakan subset pelatihan. Evaluasi kinerja algoritma dilakukan dengan

menggunakan metrik evaluasi yang relevan, seperti akurasi, presisi, recall, dan F1-score.

Dengan menggabungkan hasil eksperimen dan analisis tersebut, kesimpulan dapat ditarik mengenai efektivitas algoritma KNN, GNB, dan Naive Bayes dalam menghadapi data yang seimbang dan tidak seimbang dengan menggunakan teknik *oversampling* ADASYN. Pendekatan metodologi ini memberikan panduan sistematis untuk membandingkan kinerja algoritma dalam berbagai kondisi data dan mengevaluasi manfaat penggunaan teknik *oversampling* ADASYN dalam mengatasi ketidakseimbangan data.

IV. HASIL DISKUSI

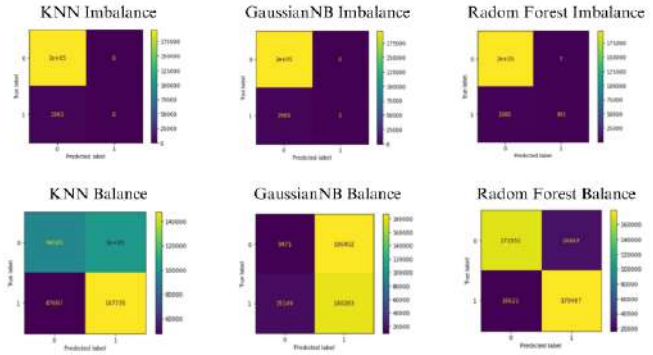
Pengujian yang dilakukan pada data seimbang dan tidak seimbang memiliki parameter yang sama untuk masing-masing algoritma. Hal ini dilakukan untuk memastikan agar perbandingan yang dilakukan benar-benar mencerminkan perbedaan dalam ketidakseimbangan data bukan karena perbedaan parameter.

TABEL 1. *HYPERPARAMETER* PADA *METHOD* KLASIFIKASI

Classifier	HyperParameters
ADASYN	sampling_strategy = 'minority'
KNN	n_neighbors=int(math.sqrt(df.shape[0])), metric='euclidean', random_state = 42, test_size=0.3
Random Forest	n_estimators=100, max_features="sqrt", random_state=42, test_size=0.3
Gaussian NB	random_state=42, test_size=0.3

Classifier	HyperParameters
ADASYN	sampling_strategy = 'minority'
KNN	n_neighbors=int(math.sqrt(df.shape[0])), metric='euclidean', random_state = 42, test_size=0.3
Random Forest	n_estimators=100, max_features="sqrt", random_state=42, test_size=0.3
Gaussian NB	random_state=42, test_size=0.3

Setelah memastikan seluruh parameter di setiap algoritma sama, maka program dapat dijalankan dan memunculkan hasil dari *confussion matrix.*



Gambar 4. *Confussion Matrix balance & imbalance*

Pada data yang sudah balance, *confussion matrix* yang dihasilkan sudah cukup baik meskipun di algoritma KNN dan Gaussian menghasilkan akurasi yang cukup rendah.

Namun masalah muncul di data *imbalance*. Pada data *imbalance*, perbedaan signifikan pada kelas klasifikasi di dataset sangat berpengaruh, kecilnya data status yang bernilai 1 membuat tidak banyak data yang dapat dijadikan sampel. Bahkan dari seluruh sampel yang ada, tidak terdapat sampel yang diprediksi sebagai nilai positif. Hal ini tentu saja berpengaruh di perhitungan presisi, recall dan f1-score yang memerlukan nilai *True Positive*(TP) dan *False Positive*(FP). Pada akhirnya presisi menghasilkan nilai tak terdefinisi sehingga hasilnya di set menjadi 0. Begitu pula pada *recall* dan *f1-score* yang hasil perhitungannya bergantung pada presisi dan recall. Hasil 0 ini dapat dilihat pada bagian KNN dan Gaussian NB.

TABEL 2. PARAMETER *ZERO_DIVISION*

	Imbalance Data			Imbalance Data dengan Zero_division		
	Random Forest	KNN	Gaussian NB	Random Forest	KNN	Gaussian NB
Precission	0.996	0.000	0.000	0.996	1.000	1.000
Recall	0.336	0.000	0.000	0.336	0.000	0.000
F1-Score	0.503	0.000	0.000	0.503	0.000	0.000
Accuracy	0.990	0.985	0.985	0.990	0.985	0.985

Permasalahan tersebut dapat coba ditangani dengan menambahkan parameter *zero_division* yang nilainya di set 1. Parameter *zero_division* ini digunakan untuk mengontrol perilaku saat terjadi pembagian 0. Hal ini akan memberikan nilai presisi yang tinggi meskipun secara logika nilai presisi yang ada ditambahkan dari parameter *zero_division*. Hasil *recall* dan *f1-score* yang masih 0 dapat mengidikasikan bahwa algoritma KNN dan Gaussian NB buruk untuk menangani dataset LCMS keladi tikus.

TABEL 3. HASIL BALANCE DAN IMBALANCE DATA

	Imbalance Data			Balance Data(Oversampling ADASYN)		
	Random Forest	KNN	Gaussian NB	Random Forest	KNN	Gaussian NB
Precission	0.996	1.000	1.000	0.881	0.046	0.492
Recall	0.336	0.000	0.000	0.917	0.093	0.922
F1-Score	0.503	0.000	0.000	0.898	0.062	0.641
Accuracy	0.990	0.985	0.985	0.896	0.958	0.485

Hasil yang didapatkan oleh data seimbang dan tidak seimbang terlihat bahwa algoritma random forest memiliki nilai akurasi tertinggi diabandingkan algoritma KNN dan Gaussian NB, terutama di data yang *imbalance*/tidak seimbang. Hanya saja, meskipun memiliki nilai akurasi tinggi, random forest di *imbalance* data memiliki nilai rendah di bagian recall dan f1-score yang menunjukkan bahwa kemampuan untuk menunjukkan kelas minoritas(positif) rendah. Berbeda dengan random forest di *balance* data yang memiliki akurasi lebih rendah namun untuk hasil presisi, recall dan f1-score memiliki nilai yang lebih seimbang.

Sedangkan untuk algoritma KNN dan Gaussian NB menunjukkan performa yang kurang memuaskan dalam hal akurasi pada data yang sudah seimbang maupun belum seimbang. Hal ini mengindikasikan bahwa kedua algoritma tersebut tidak efektif ketika digunakan bersamaan dengan teknik *oversampling* ADASYN pada dataset yang ada.

V. KESIMPULAN

Melalui data yang sudah diuji coba menunjukkan bahwa random forest merupakan algoritma yang tahan terhadap data *imbalance*. Hal ini ditunjukkan dengan tingginya nilai akurasi dan presisi, hanya saja kemampuan untuk mendeteksi di kelas minoritas masih rendah. Untuk itu, baik jika dilakukan penyeimbangan data menggunakan metode *oversampling* ataupun *undersampling* ataupun bisa mencoba merubah beberapa bagian dalam parameter. Pada percobaan yang dilakukan dengan metode *oversampling* ADASYN menunjukkan akurasi 0.896, hal ini cukup membuktikan bahwa ketika dilakukan teknik *oversampling* random forest dengan dataset yang ada masih tergolong baik ditambah dengan hasil recall, f1-score dan presisi yang seimbang.

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."

[2] N. Rout, D. Mishra, and M. K. Mallick, "Handling imbalanced data: A survey," in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2018, pp. 431–443. doi: 10.1007/978-981-10-5272-9_39.

[3] G. Rekha, A. K. Tyagi, N. Sreenath, and S. Mishra, "Class Imbalanced Data: Open Issues and Future Research Directions," in *2021 International Conference on Computer Communication and Informatics, ICCCI 2021*, Institute of Electrical and Electronics Engineers Inc., Jan. 2021. doi: 10.1109/ICCCI50826.2021.9402272.

[4] I. Binanto, H. L. H. S. Warnars, N. F. Sianipar, and W. Budiharto, "Webscraping Data Labeling System on Liquid Chromatography-Mass Spectrometry of Rodent Tuber for Efficiency of Supervised Learning Preprocessing," *ICIC Express Letters, Part B: Applications*, vol. 13, no. 1, pp. 107–114, 2022, doi: 10.24507/icicelb.13.01.107.

[5] N. F. Sianipar and R. Purnamaningsih, "Enhancement of the contents of anticancer bioactive compounds in mutant clones of rodent tuber (Typhonium flagelliforme Lodd.) based on GC-MS analysis," *Pertanika J Trop Agric Sci*, vol. 41, no. 1, pp. 305–320, 2018.

[6] T. M. Khoshgoftaar, A. Fazelpour, D. J. DIttman, and A. Napolitano, "Alterations to the Bootstrapping Process within Random Forest: A Case Study on Imbalanced Bioinformatics Data," in *Proceedings - 2015 IEEE 16th International Conference on Information Reuse and Integration, IRI 2015*, Institute of Electrical and Electronics Engineers Inc., Oct. 2015, pp. 342–348. doi: 10.1109/IRI.2015.59.

[7] W. Hidayat, M. Ardiansyah, and A. Setyanto, "Pengaruh Algoritma ADASYN dan SMOTE terhadap Performa Support Vector Machine pada Ketidakseimbangan Dataset Airbnb," *Edumatic: Jurnal Pendidikan Informatika*, vol. 5, no. 1, pp. 11–20, Jun. 2021, doi: 10.29408/edumatic.v5i1.3125.

[8] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss." [Online]. Available: https://github.com/kaidic/LDAM-DRW.

[9] S. Bagga, A. Goyal, N. Gupta, and A. Goyal, "Credit Card Fraud Detection using Pipeling and Ensemble Learning," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 104–112. doi: 10.1016/j.procs.2020.06.014.

[10] A. Kulkarni, D. Chong, and F. A. Batarseh, "Foundations of data imbalance and solutions for a data democracy," in *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*, Elsevier, 2020, pp. 83–106. doi: 10.1016/B978-0-12-818366-3.00005-8.

[11] Y. Li, W. Xu, W. Li, A. Li, and Z. Liu, "Research on hybrid intrusion detection method based on the ADASYN and ID3 algorithms," *Mathematical Biosciences and Engineering*, vol. 19, no. 2, pp. 2030–2042, 2021, doi: 10.3934/MBE.2022095.

[12] C. Magnolia, A. Nurhopipah, and A. Kusuma, "Penanganan Imbalanced Dataset untuk Klasifikasi Komentar Program Kampus Merdeka Pada Aplikasi Twitter," *Edu Komputika*, vol. 9, no. 2, pp. 105–113, 2022.

[13] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. N. Davis, "DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values," *IEEE Access*, vol. 7, pp. 102232–102238, 2019, doi: 10.1109/ACCESS.2019.2929866.

[14] J. A. Septian, T. M. Fachrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," *Journal of Intelligent System and Computation*, vol. 1, no. 1, pp. 43–49, 2019, doi: 10.52985/insyst.v1i1.36.

[15] A. Bansal and A. Singhrova, "Performance Analysis of Supervised Machine Learning Algorithms for Diabetes and Breast Cancer Dataset," in *Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 137–143. doi: 10.1109/ICAIS50930.2021.9396043.

[16] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Comput Sci*, vol. 162, no. Itqm 2019, pp. 503–513, 2019, doi: 10.1016/j.procs.2019.12.017.

[17] A. Syukron and A. Subekti, "Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit," *Jurnal Informatika*, vol. 5, no. 2, pp. 175–185, 2018, doi: 10.31311/ji.v5i2.4158.