ANALISA PERBANDINGAN TEKNIK OVERSAMPLING SMOTE PADA IMBALANCED DATA

Cosmas Haryawan[1], Yosef Muria Kusuma Ardhana[2]

[1,]Program Studi Sistem Informasi, Universitas Teknologi Digital Indonesia, [2,]Program Studi Bisnis Digital, Universitas Teknologi Digital Indonesia

Jl. Raya Janti Karang Jambe No. 143, Kab. Bantul, DIY
[1] cosmas@utdi.ac.id, [2] yosefmurya@gmail.com

Abstract

Imbalanced data is when data has an unbalanced ratio between one class and another, so there is a majority class and a minority class. Making predictions on imbalanced datasets is difficult because classifiers tend to detect majority classes rather than minority classes. The resampling technique is the most effective in solving this imbalanced data problem. One of the categories of resampling techniques is oversampling. Metode oversampling among them are SMOTE and K-Means SMOTE. The use of oversampling will improve the classification measurement results. This study used wine data with 11 features and one target attribute. It aimed to compare the measurement results between synthetic data from SMOTE and K-Means SMOTE with the measurement results of real data use under balanced data conditions. Imbalanced data is created by randomly deleting one of the classes with thresholds of 25%, 50%, 60% and 75%. The results showed that compared to the use of real data, the use of K-Means SMOTE tends to produce higher values for accuracy, sensitivity and specificity while SMOTE, although obtaining better values than K-Means SMOTE, several imbalance conditions have higher values than the use of real data.

Keywords : *imbalanced data, K-Means SMOTE, oversampling, SMOTE*

Abstrak

Data tidak seimbang atau lebih sering disebut *imbalanced data*, adalah kondisi pada saat data memiliki rasio yang tidak seimbang antara satu kelas dengan kelas yang lain, sehingga terdapat kelas mayoritas dan kelas minoritas. Sulit untuk membuat prediksi pada dataset yang tidak seimbang karena pengklasifikasi cenderung mendeteksi kelas mayoritas daripada kelas minoritas. Teknik resampling menjadi salah satu yang paling efektif dalam menyelesaikan permasalahan *imbalanced data* ini. Salah satu kategori dari teknik resampling adalah *oversampling*. Metode *oversampling* diantaranya adalah SMOTE dan K-Means SMOTE. Penggunaan *oversampling* akan meningkatkan hasil *measurement* klasifikasi. Penelitian ini menggunakan data wine yang memiliki 11 fitur serta 1 atribut target dan bertujuan untuk membandingkan hasil *measurement* antara penggunaan data sintetis hasil SMOTE dan K-Means SMOTE dengan hasil *measurement* penggunaan data nyata dalam kondisi data seimbang. Pembuatan data *imbalance* dilakukan dengan menghapus secara random salah satu kelas dengan ambang 25%, 50%, 60% dan 75%. Hasil penelitian menunjukkan bahwa dibandingkan penggunaan data nyata, penggunaan K-Means SMOTE cenderung menghasilkan nilai lebih tinggi untuk akurasi, sensitivitas dan spesifisitas sedangkan SMOTE meskipun memperoleh nilai yang lebih baik dibandingkan K-Means SMOTE tetapi juga terdapat beberapa kondisi *imbalance* yang memiliki nilai lebih tinggi dibandingkan penggunaan data nyata.

Kata kunci : *imbalanced data, K-Means SMOTE, oversampling, SMOTE*

1. PENDAHULUAN

Data tidak seimbang atau lebih sering disebut *imbalanced data*, adalah kondisi pada saat data memiliki rasio yang tidak seimbang antara satu kelas dengan kelas yang lain, sehingga terdapat kelas mayoritas (dengan data yang banyak) dan kelas minoritas (dengan data sedikit) [1]. Sulit

untuk membuat prediksi pada dataset yang tidak seimbang karena pengklasifikasi cenderung mendeteksi kelas mayoritas daripada kelas minoritas. Oleh karena itu, keluaran dari klasifikasi akan menjadi bias [2]. Beberapa metode digunakan untuk mengatasi hal tersebut. Metode resampling menjadi salah satu yang paling efektif dalam menyelesaikan permasalahan imbalanced data ini [3]. Dalam metode resampling terdapat teknik oversampling dan teknik undersampling, Teknik oversampling memberikan hasil yang lebih baik dibandingkan teknik undersampling [4]. Salah satu metode oversampling yang banyak digunakan adalah SMOTE [3].

Akurasi menjadi salah satu hal yang penting dalam measurement teknik klasifikasi dalam data mining. Kondisi data tidak seimbang sering menyebabkan tingkat akurasi menjadi rendah. SMOTE bekerja dengan menambahkan data sintetis pada kelas minoritas untuk membuat data menjadi seimbang [5]. Penerapan metode oversampling SMOTE cenderung meningkatkan akurasi menjadi lebih tinggi bahkan dapat mencapai dua kali lipat akurasi tanpa SMOTE [6].

Penelitian mengenai penggunaan K-Means SMOTE untuk peningkatan akurasi pernah dilakukan untuk klasifikasi [7]. Hasil penelitian menunjukkan bahwa K-Means SMOTE secara rata-rata mampu mengeliminasi 55% hasil false positive. Masih di tahun yang sama, SMOTE juga digunakan oleh Siringoringo digabungkan dengan teknik klasifikasi KNN pada permasalahan credit card [8], ini menghasilkan peningkatan rata-rata G-MEAN dari 53,4% menjadi 81% dan F-Measure dari 38,7% menjadi 81,8%.

Permasalahan Credit card dengan K-Means SMOTE juga dilakukan [9] pada tahun 2021 pada teknik klasifikasi KNN, logistics, SVM, random forest, dan tree. Hasil penelitian menunjukkan bahwa K-Means SMOTE meningkatkan Nilai AUC dari 0.765 menjadi 0.929

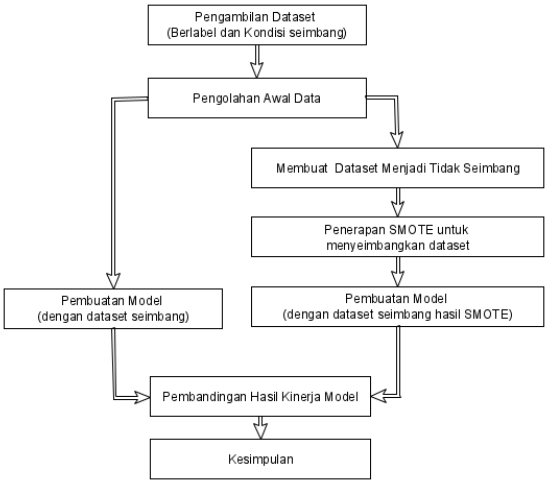
Pada Tahun 2021 [10] juga memanfaatkan K-Means SMOTE untuk memperbaiki *imbalanced data* pada data kelulusan mahasiswa Universitas Bumigora. Hasil yang diperoleh dibandingkan dengan penggunaan data *imbalance*, terjadi peningkatan akurasi sebesar 5.4% dan sensitivitas sebesar 39%.

Berdasarkan hal tersebut maka perlu diketahui perbandingan akurasi dan *measurement* yang lain antara dampak penambahan data sintetis atau data bangkitan dari teknik *oversampling* SMOTE maupun K-Means SMOTE dengan hasil *measurement* proses yang menggunakan data nyata.

2. METODOLOGI PENELITIAN

2.1. Skema Alur Penelitian

Tahapan Penelitian ditunjukkan oleh Gambar 1. Langkah awal adalah mengambil dataset yang berlabel dan berada dalam kondisi seimbang. Data tersebut kemudian dilakukan pengolahan data awal / *preprocessing* yang diperlukan untuk memastikan bahwa data dan semua atribut yang ada didalamnya siap untuk dilakukan proses pengolahan berikutnya.



Gambar 1. Tahapan Penelitian

Selanjutnya, data yang sudah melalui tahap *preprocessing* ini diduplikasi sehingga total terdapat 5 data. Data pertama dibiarkan dalam apa adanya (dalam kondisi seimbang) yang akan digunakan sebagai data kontrol, sedangkan ke-empat data yang lain dijadikan data tidak seimbang dengan mengurangi salah satu kelas yang ada secara random dan dijadikan kelas minoritas. Pengaturan ambang ketidak-seimbangan untuk masing-masing data adalah sebesar 75%, 60%, 50% dan 25%. Kemudian data yang tidak seimbang tersebut dijadikan data yang seimbang menggunakan teknik oversampling SMOTE dan K-Means SMOTE.

Langkah selanjutnya adalah membuat model untuk klasifikasi. Support Vector Machine (SVM) adalah salah satu teknik data mining yang memiliki kinerja bagus dalam melakukan klasifikasi seperti yang pernah dilakukan oleh [11] dan [12]. Akurasi yang tinggi dari SVM untuk melakukan klasifikasi juga dibuktikan oleh [13] dalam penelitiannya. Di dalam penelitian ini, SVM digunakan untuk membuat model klasifikasi baik dari data kontrol maupun dari data hasil *oversampling*. Berdasar model yang ada selanjutnya dilakukan evaluasi kinerja dan dibandingkan hasilnya antara data kontrol dengan data hasil *oversampling*.

2.2. Pengumpulan Data

Dataset diperoleh dari situs open data kaggle.com berupa data kualitas wine. Data penelitian ini berbentuk file csv (*comma separated value*). Data yang digunakan memiliki total 12 atribut, dengan 11 atribut sebagai fitur yaitu : *'fixed acidity', 'volatile acidity', 'citric acid','residual sugar', 'chlorides', 'free sulfur dioxide','total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol'* dan 1 atribut adalah target yaitu: *'quality'*. Semua atribut fitur memiliki tipe data numerik, sedangkan atribut target bertipe string yang berisi label 'GOOD' dan 'BAD'. Banyak data yang terdapat dalam data penelitian ini adalah 1488 data, dengan kondisi label kelas yang seimbang (balance) yaitu GOOD dan BAD masing-masing sebanyak 744 data. Tidak terdapat data yang memiliki nilai NULL untuk semua atribut.

Untuk persiapan pengolahan data, maka atribut target yang sebelumnya bertipe string perlu di-encode menjadi numerik menggunakan pustaka sklearn, sehingga isi label berubah menjadi 0 dan 1. Dalam hal penelitian ini, label BAD menjadi angka 0 dan label GOOD menjadi angka 1

2.3. Analisa Data

Analisa data dilakukan dengan melakukan evaluasi kinerja dari metode klasifikasi yang dipakai. Evaluasi kinerja dilakukan pada semua kondisi data baik seimbang maupun tidak seimbang dengan evaluasi kinerja pada data seimbang digunakan sebagai data kontrol atau pembanding.

Evaluasi kinerja ini dilakukan dengan menggunakan *confusion matrix*. Tabel 1 menunjukkan *confusion matrix* untuk kelas biner, sesuai yang digunakan dalam penelitian ini.

TABEL 1. *CONFUSION MATRIX*

Class	Predictive Positive	Predictive Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Hasil penelitian ini dimasukkan ke *confusion matrix* untuk kemudian di lakukan perhitungan dengan menggunakan persamaan 1 untuk akurasi, persamaan 2 untuk sensitivitas dan persamaan 3 untuk spesifisitas

TP (*True Positive*) adalah banyak data benar pada target yang terklasifikasi benar pada sistem, TN (*True Negative*) adalah banyak data salah pada target yang terklasifikasi salah pada sistem, FP (*False Positive*) adalah banyak data salah pada target yang terklasifikasi benar pada sistem dan FN (*False Negative*) adalah banyak data benar pada target yang terklasifikasi salah pada sistem [14].

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

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

$$Spesifisitas = \frac{TN}{TN+FP} \quad (3)$$

2.4. Pembuatan Imbalanced Data

Berdasar data yang sudah seimbang sebelumnya maka dibentuk 4 data baru yang merupakan modifikasi dari data tersebut. Keempat data dibuat menjadi data yang memiliki kelas yang tidak seimbang sebesar 25%, 50%, 60% dan 75%. Proses ini dilakukan dengan mengurangi kelas BAD secara random sesuai yang diperlukan.

Total data penelitian yang digunakan ada 5 data, dengan data seimbang menjadi data kontrol. Perbandingan tiap kelas untuk masing-masing data dapat dilihat di Tabel 1.

TABEL 2. PERBANDINGAN KELAS 5 DATA PENELITIAN

Kondisi	GOOD	BAD
Balance	744	744
Imbalance 25%	744	559
Imbalance 50%	744	374
Imbalance 60%	744	300
Imbalance 75%	744	189

3. HASIL DAN PEMBAHASAN

3.1. Hasil Penelitian

Proses *Oversampling* dilakukan dengan 2 metode yaitu SMOTE dan K-Means SMOTE, sedangkan untuk parameter Kernel, nilai C, dan nilai gamma pada teknik SVM dijalankan dengan tuning parameter. Tuning parameter dilakukan dengan mencari nilai akurasi tertinggi dengan beberapa variasi nilai C dan gamma, serta variasi jenis kernel. Untuk setiap pengolahan data, 80% data yang ada digunakan sebagai data pelatihan dan 20% digunakan sebagai data pengujian

Hasil penelitian pengujian kinerja SMOTE dengan berbagai variasi kondisi dan hasil kontrol menggunakan data seimbang ditunjukkan oleh

Tabel 3. Sedangkan hasil penelitian pengujian kinerja K-Means SMOTE dengan berbagai variasi kondisi dan hasil kontrol menggunakan data seimbang ditunjukkan oleh Tabel 4.

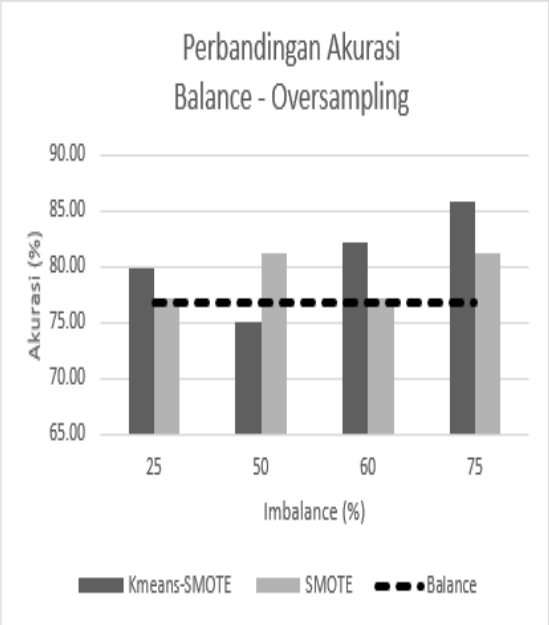
TABEL 3. KINERJA SMOTE

Kondisi	Metric (dalam %)		
	Akurasi	Sensitivitas	Spesifisitas
Balance	76.85	82.05	71.13
Imbalance 25%	77.18	84.08	69.50
Imbalance 50%	81.21	85.21	77.56
Imbalance 60%	77.18	85.81	68.67
Imbalance 75%	81.21	85.71	76.93

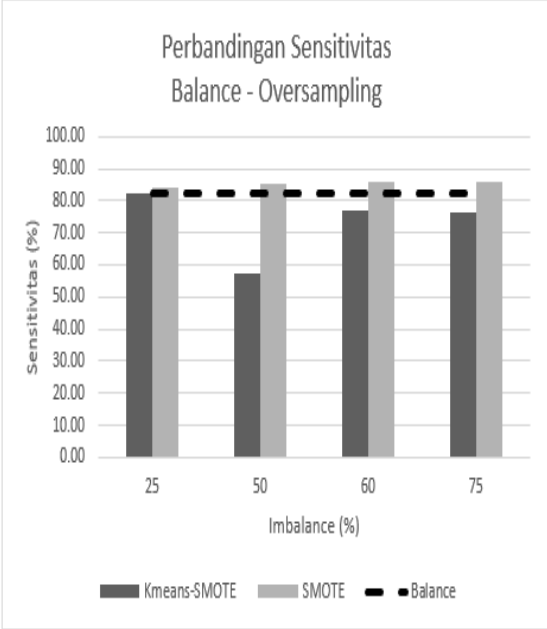
TABEL 4. KINERJA K-MEANS SMOTE

Kondisi	Metric (dalam %)		
	Akurasi	Sensitivitas	Spesifisitas
Balance	76.85	82.05	71.13
Imbalance 25%	77.18	84.08	69.50
Imbalance 50%	81.21	85.21	77.56
Imbalance 60%	77.18	85.81	68.67
Imbalance 75%	81.21	85.71	76.93

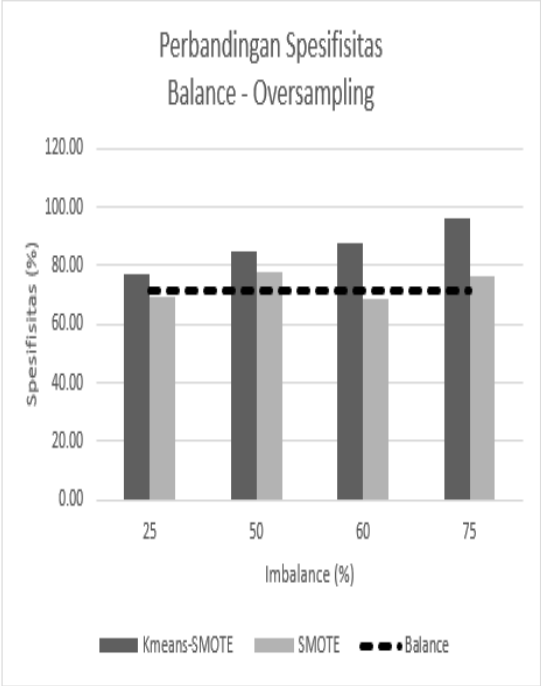
Perbandingan Hasil perhitungan kinerja untuk akurasi, sensitivitas dan spesifisitas antara data kontrol dengan data bangkitan *oversampling* ditunjukkan berturut-turut oleh gambar 2, gambar 3 dan gambar 4.



Gambar 2. Perbandingan Akurasi



Gambar 3. Perbandingan Sensitivitas



Gambar 4. Perbandingan Spesifisitas

3.2. Pembahasan

Hasil penelitian menunjukkan bahwa data kontrol yang merupakan data dalam kondisi seimbang, mendapatkan hasil akurasi sebesar 76.85%, Sensitivitas sebesar 82.05%, dan Spesifisitas sebesar 71.13%.

Berdasarkan Tabel 3, terlihat bahwa data bangkitan yang dibuat dengan teknik *oversampling* SMOTE, memperoleh hasil akurasi yang mendekati data kontrol pada kondisi

imbalance 25% dengan akurasi 77.18% dan *imbalance* 60% yang juga memiliki akurasi 77.18%. Namun, pada *imbalance* 50% dan 75% diperoleh hasil akurasi yang sedikit melebihi data kontrol yaitu sebesar 81.21% atau terdapat peningkatan 5.68% dari data kontrol. Hasil Sensitivitas untuk teknik oversampling SMOTE, diperoleh hasil sensitivitas yang hampir sama untuk semua kondisi imbalance yaitu sekitar 84.08% hingga 85.71%. Angka ini sedikit melebihi nilai sensitivitas data kontrol yang sebesar 82.05% dengan tingkat kelebihan sebesar 2.4% hingga 4.58% dari data kontrol. Hasil Spesifisitas untuk teknik oversampling SMOTE, hampir mirip dengan nilai akurasi, yaitu mendekati data kontrol untuk *imbalance* 25% dan 60%, masing-masing memiliki nilai spesifisitas 69.50% dan 68.67%, (penurunan 2.28% hingga 3.46% dari data kontrol) sedangkan pada *imbalance* 50% dan 75% lebih tinggi dari data kontrol yaitu masing-masing 77.56% dan 76.39%. Angka ini cukup tinggi dan mencapai peningkatan 9.05% dari data kontrol.

Data bangkitan yang dibuat menggunakan teknik K-Means SMOTE, memiliki karakter yang berbeda dengan hasil teknik SMOTE seperti terlihat pada Tabel 4. Untuk akurasi, data hasil bangkitan dengan *imbalance* 50% memperoleh hasil 75% yang berarti sedikit lebih rendah dibanding data kontrol, tetapi untuk *imbalance* 25%, 60% dan 75% hasil akurasinya jauh lebih tinggi dibanding data kontrol, yaitu masing-masing 79.87%, 82.21% dan 85.91%. Peningkatannya mencapai 11.79% dari data kontrol. Hasil Sensitivitas untuk teknik *oversampling* K-Means SMOTE memperoleh hasil yang bagus pada imbalance 25% yaitu sebesar 82,58%, nilai ini sangat mendekati nilai data kontrol yang sebesar 82.05%, akan tetapi pada *imbalance* 50%, nilai sensitivitasnya jauh dibawah data kontrol karena hanya mencapai 57.50%. Imbalance 60% dan 75% memiliki sensitivitas 76.87% dan 76.62%. Hasil spesifisitas teknik *oversampling* K-Means SMOTE memperoleh hasil yang jauh di atas data kontrol, baik untuk imbalance 25%, 50%, 60% maupun 75% yang masing-masing memperoleh hasil 76.92%, 84.72%, 87.42%, dan 95.83%. Ini berarti hasil data bangkitan memiliki peningkatan spesifisitas antara 8.75% hingga 34.74% dari data kontrol.

Apabila dilihat dari grafik yang ditunjukkan oleh Gambar 2, secara umum K-Means SMOTE menghasilkan nilai akurasi lebih tinggi dibandingkan dengan SMOTE. Hal ini sesuai dengan yang disampaikan oleh [15] bahwa performance akurasi K-Means SMOTE lebih baik dari SMOTE. Tetapi jika dibandingkan dengan data kontrol, maka terlihat bahwa hasil akurasi SMOTE lebih mendekati data nyata meskipun hanya pada imbalance 25% dan 50% sedangkan K-Means SMOTE cenderung memiliki selisih cukup banyak dibandingkan data nyata. Pada perbandingan nilai sensitivitas, SMOTE lebih baik dibandingkan K-Means SMOTE di semua kondisi *imbalance* seperti ditunjukkan Gambar 3. SMOTE juga lebih baik dibandingkan K-Means SMOTE pada perbandingan nilai spesifisitas seperti yang terlihat pada Gambar 4.

4. KESIMPULAN DAN SARAN

4.1. Kesimpulan

Penelitian ini menghasilkan suatu perbandingan hasil measurement antara dampak penambahan data sintetis/data bangkitan dengan teknik *oversampling* SMOTE dan K-Means SMOTE dibandingkan penggunaan data nyata. Teknik *oversampling* SMOTE memiliki akurasi yang mendekati data nyata pada kondisi *imbalance* 25% dan 60% sedangkan K-Means SMOTE cenderung memiliki akurasi yang jauh melebihi data nyata, kecuali pada *imbalance* 50%. Hasil sensitivitas teknik *oversampling* SMOTE juga lebih baik dibandingkan K-Means SMOTE di semua kondisi *imbalance*. Hasil spesifisitas teknik *oversampling* SMOTE lebih mendekati data nyata di semua kondisi *imbalance* dibandingan dengan teknik *oversampling* K-Means SMOTE yang memiliki sensitifitas jauh di atas data nyata pada semua kondisi *imbalance*.

Berdasar hasil tersebut dapat dikatakan bahwa untuk pengukuran akurasi, sensitivitas dan spesifisitas, penggunaan SMOTE lebih mendekati data nyata dibandingkan penggunaan K-Means SMOTE, akan tetapi perlu lebih hati-hati juga dalam penerapan teknik *oversampling* SMOTE ini, karena pada kondisi-kondisi imbalance tertentu terdapat hasil-hasil pengukuran yang ternyata lebih tinggi dibandingkan dengan data nyata.

4.2. Saran

Penelitian ini hanya menggunakan 1 teknik data mining yaitu SVM untuk melakukan klasifikasi, sehingga perlu dilakukan berbagai penelitian akibat dari penggunaan teknik *oversampling* SMOTE maupun K-Means SMOTE pada teknik data mining yang lain. Selain itu perlu juga dilakukan penambahan variasi ambang *imbalanced* data agar dapat lebih mengetahui perubahan dampak tiap ambang secara lebih detail.

5. UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Teknologi Digital Indonesia yang telah memberikan sarana dan prasarana sehingga penulis dapat menyelesaikan penelitian ini dengan baik.

Daftar Pustaka:

[1] M. Anis and M. Ali, "Investigating the Performance of Smote for Class Imbalanced Learning: A Case Study of Credit Scoring Datasets," *Eur. Sci. Journal, ESJ*, vol. 13, no. 33, p. 340, 2017, doi: 10.19044/esj.2017.v13n33p340.

[2] Y. R. Chen, J. S. Leu, S. A. Huang, J. T. Wang, and J. I. Takada, "Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets," *IEEE Access*, vol. 9, pp. 73103–73109, 2021, doi: 10.1109/ACCESS.2021.3079701.

[3] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.

[4] S. Mundra *et al.*, "Classification of imbalanced medical data: An empirical study of machine learning approaches," *J. Intell. Fuzzy Syst.*, vol. 43, no. 2, 2022, doi: 10.3233/JIFS-219294.

[5] N. Matondang and N. Surantha, "Effects of oversampling SMOTE in the classification of hypertensive dataset," *Adv. Sci. Technol. Eng. Syst.*, vol. 5, no. 4, 2020, doi: 10.25046/AJ050451.

[6] A. R. Safitri and M. A. Muslim, "Improved Accuracy of Naive Bayes Classifier for Determination of Customer Churn Uses SMOTE and Genetic Algorithms," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 70–75, 2020, doi: https://doi.org/10.52465/joscex.v1i1.5.

[7] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Ny).*, vol. 465, 2018, doi: 10.1016/j.ins.2018.06.056.

[8] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan k-Nearest Neighbor," *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018, doi: https://doi.org/10.29207/resti.v3i2.945.

[9] Y. Chen and R. Zhang, "Erratum: Research on Credit Card Default Prediction Based on k -Means SMOTE and BP Neural Network (Complexity (2021) 2021:13 (6618841) DOI: 10.1155/2021/6618841)," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/9865171.

[10] Hairani, "Peningkatan Kinerja Metode Svm Menggunakan Metode Knn Imputasi Dan K-Means-Smote Untuk Klasifikasi Kelulusan Mahasiswa Universitas Bumigora," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 4, pp. 713–718, 2021, doi: 10.25126/jtiik.202183428.

[11] A. S. Handayani, S. Soim, T. E. Agusdi, Rumiasih, and A. Nurdin, "KLASIFIKASI KUALITAS UDARA DENGAN METODE SUPPORT VECTOR MACHINE," *JIRE, J. Inform. dan Rekayasa Elektron.*, vol. 3, no. 2, 2020.

[12] L. Mutawalli, M. T. A. Zaen, and W. Bagye, "KLASIFIKASI TEKS SOSIAL MEDIA TWITTER MENGGUNAKAN SUPPORT VECTOR MACHINE (Studi Kasus Penusukan Wiranto)," *J. Inform. dan Rekayasa Elektron.*, vol. 2, no. 2, p. 43, Dec. 2019, doi: 10.36595/jire.v2i2.117.

[13] D. Mustafa Abdullah and A. Mohsin Abdulazeez, "Machine Learning Applications based on SVM Classification A Review," *Qubahan Acad. J.*, vol. 1, no. 2, 2021, doi: 10.48161/qaj.v1n2a50.

[14] C. Haryawan and M. M. Sebatubun, "IMPLEMENTATION OF MULTILAYER PERCEPTRON FOR STUDENT FAILURE PREDICTION," *JUTI J. Ilm. Teknol. Inf.*, vol. 18, no. 2, p. 125, Jul. 2020, doi: 10.12962/j24068535.v18i2.a990.

[15] S. Sarkar, A. Pramanik, J. Maiti, and G. Reniers, "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data," *Saf. Sci.*, vol. 125, p. 104616, May 2020, doi: 10.1016/j.ssci.2020.104616.