Penerapan Metode *Oversampling* SMOTE Pada Algoritma Random Forest Untuk Prediksi Kebangkrutan Perusahaan

Application of the SMOTE Oversampling Method to the Random Forest Algorithm for Predicting Company Bankruptcy

Agung Nugroho[1], Elkin Rilvani[2]
[1,2]Teknik Informatika, Universitas Pelita Bangsa
E-mail: [1]agung@pelitabangsa.ac.id, [2]elkin.rilvani@pelitabangsa.ac.id

Abstrak

Kemajuan teknologi informasi berkembang kearah finansial dalam melakukan prediski. Banyak model algoritma prediksi data keuangan telah dikembangkan. Prediksi kebangkrutan merupakan hal yang sangat penting bagi sebuah organisasi atau perusahaan dalam mengambil keputusan yang diperlukan oleh pemodal dan investor. Prediksi kebangkrutan termasuk dalam permasalahan ketidakseimbangan kelas dalam model kalsifikasi karena jumlah data yang termasuk dalam kelas bangkrut jauh lebih sedikit dibandingkan dengan data yang termasuk dalam kelas tidak bangkrut. Tujuan dari penelitian ini adalah untuk menghasilkan model klasifikasi yang baik pada prediksi risiko kebangkrutan di perusahaan. preprocessing data dilakukan untuk melakukan optimasi algoritma klasifikasi dengan metode *oversampling* SMOTE agar menghasilkan model kalsifikasi yang optimal. Algoritma klasifikasi yang digunakan adalah *Random Forest Classification* untuk mendapatkan hasil prediksi kebangkrutan yang optimal. Berdasarkan hasil pengujian yang telah dilakukan didapatkan bahwa performa klasifikasi meningkat sebesar 7,40% setelah dilakukan preprocessing data dengan menggunakan teknik oversampling SMOTE pada algoritma *Random Forest Classifier.*

Kata kunci: *Bankruptcy, Oversampling, SMOTE, Random Forest*

Abstract

Advances in information technology are developing in a financial direction in making predictions. Many financial data prediction algorithm models have been developed. Bankruptcy prediction is something that is very important for organizations or companies in making decisions needed by investors and investors. The bankruptcy prediction is included in the class imbalance problem in the calcification model because the amount of data included in the bankrupt class is far less than the data included in the non-bankrupt class. The purpose of this research is to produce a good classification model for predicting bankruptcy risk in companies. Data pre-processing is carried out to optimize the classification algorithm using the SMOTE oversampling method to produce an optimal calcification model. The classification algorithm used is the Random Forest Classification to obtain optimal bankruptcy prediction results. Based on the results of the tests that have been carried out, it was found that the classification performance increased by 7.40% after pre-processing the data using the SMOTE oversampling technique in the Random Forest Classifier algorithm.

Keywords: Bankruptcy, Oversampling, SMOTE, Random Forest

1. PENDAHULUAN

Kebangkrutan merupakan masalah yang sangat penting dalam keuangan perusahaan[1]. Kebangkrutan merupakan penutupan perusahaan atau disebut juga likuidasi, hal ini merupakan kondisi terjadinya kegagalan dalam masalah keuangan perusahaan yang tidak sebanding dengan jumlah asetnya. Kerugian tersebut dapat berdampak pada pemegang saham, karyawan, pelanggan dan investor[2]. Oleh karena itu prediksi kebangkrutan berperan penting dalam bidang ekonomi

dan keuangan.

Proses prediksi adalah membuat keputusan yang berhubungan dengan operasi bisnis dengan cara yang dapat diandalkan[2]. Tujuan dari prediksi kebangkrutan adalah untuk menciptakan model prediksi yang efektif untuk mengidentifikasi risiko kebangkrutan berdasarkan data keuangan yang dimiliki oleh perusahaan. Hasil dari prediksi dapat digunakan oleh perusahaan untuk melakukan evaluasi dan merencanakan tindakan-tindakan preventif untuk mencegah terjadinya kebangkrutan.

Model prediksi kebangkrutan berbasis *machine learning* telah banyak dilakukan untuk mencari model prediksi yang optimal. Prediksi kebangkrutan termasuk ke dalam masalah ketidakseimbangan kelas pada proses klasifikasi, sehingga membutuhkan optimasi pada algoritma klasifikasi. Klasifikasi adalah proses penguraian data untuk menemukan model yang menerangkan kelas dalam data tersebut[3][4]. Algoritma klasifikasi bertugas memprediksi label kelas kategori dari suatu data, sehingga data tersebut dapat dikategorikan ke dalam salah satu kelas yang sudah ditentukan[5].

Data pada umumnya merupakan dataset yang memiliki ketidakseimbangan dalam distribusi kelas atau disebut sebagai dataset dengan kelas tidak seimbang. Data tidak seimbang merupakan data yang memiliki ketidakseimbangan jumlah sampel yang cukup signifikan antara satu kelas dengan kelas lainnya[6][7]. Pada data kelas minoritas sering terjadi kesalahan pada proses klasifikasi, oleh sebab itu diperlukan teknik optimasi pada data tersebut.

Banyak algoritma klasifikasi telah digunakan dalam mengatasi ketidakseimbangan kelas pada dataset, salah satunya adalah *Random Forest Classifier*[8]. *Random Forest Classifier* memberikan hasil performa yang baik dalam memproses dataset dalam jumlah yang besar[9]. Namun demikian, penelitian sejenis menganjurkan untuk melakukan optimasi pada Random Forest agar mendapatkan model klasifikasi yang lebih optimal[10].

Teknik optimasi pada dataset tidak seimbang telah banyak dikembangkan[11]. Beberapa teknik optimasi diantaranya adalah dengan pendekatan level data[12], pendekatan pada level algoritma dan pendekatan pada keduanya[13]. Dengan pendekatan level data, beberapa teknik digunakan dalam proses optimasi, diantaranya adalah melakukan *oversampling* pada data[14].

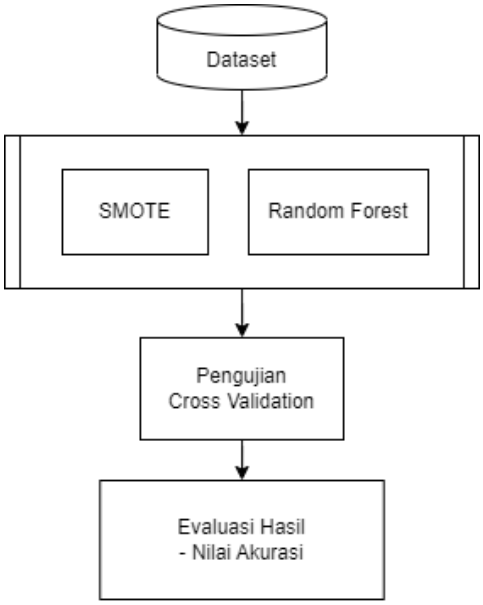
Pendekatan yang digunakan pada penelitian ini adalah *oversampling* SMOTE. SMOTE (*Synthetic Minority Over-sampling Technique*) merupakan metode untuk menyeimbangkan jumlah data dari kelas minor dengan kelas major dengan cara menciptakan data sintetis atau data yang dibuat secara artifisial[15].

Tujuan dari penelitian ini adalah untuk melakukan prediksi risiko kebangkrutan pada perusahaan menggunakan algoritma klasifikasi *Random Forest Classifier* dengan optimasi *oversampling* SMOTE agar proses klasifikasi dapat dilakukan lebih optimal.

2. METODE PENELITIAN

Data yang digunakan pada penelitian ini adalah data yang berasal dari Kaggle berkaitan dengan data *Bankruptcy* dari *Taiwan Economic Journal* tahun *1999–2009*. Data set termasuk ke dalam kategori data tidak seimbang.

Metode penelitian yang digunakan adalah melakukan optimasi pada model algoritma klasifikasi dengan teknik *oversampling* menggunakan metode SMOTE pada algoritma *Random Forest Classifier*. Skema penelitian tergambar pada gambar 1.



Gambar 1. Skema Penelitian

Tahap awal dilakukan pembagian dataset menjadi dua bagian yaitu data test dan data train dengan persentase 30:70 yaitu 30% data test dan 70% data testing. Data testing kemudian dilakukan *pre-processing* untuk menyeimbangkan kelas data dengan menggunakan teknik *oversampling* SMOTE.

SMOTE merupakan sebuah teknik *oversampling* yang digunakan dalam pembelajaran mesin untuk menangani masalah *imbalanced class* pada data[16]. SMOTE bekerja dengan menciptakan data sintetis baru dari kelas minoritas dengan cara menginterpolasi antara titik-titik data yang ada. SMOTE menghasilkan contoh sintetis dari kelas minoritas yang dioperasikan di dalam ruang fitur daripada di dalam ruang data[3]. Teknik ini menambahkan contoh dari kelas minoritas dengan cara mengekstrak sampel data minoritas yang ada dengan menggunakan sampel acak yang diambil dari nilai k tetangga terdekat[17]. Dengan demikian, SMOTE menghasilkan contoh sintetis baru yang dapat memperluas area keputusan dari kelas minoritas. Metode SMOTE bekerja dengan menambahkan jumlah data pada kelas minor hingga sama dengan kelas mayor dengan cara membuat data sintetis[18]. Data sintetis tersebut dibuat berdasarkan nilai k-tetangga terdekat dari kelas minor. Prinsip dari SMOTE adalah untuk menyeimbangkan data kelas dengan cara membangkitkan data sintetis.

Tahap berikutnya dilakukan pengujian menggunakan *cross validation* terhadap algoritma *Random Forest Classifier* pada dataset awal dan *Radom Forest Classifier* dengan data yang sudah seimbang menggunakan *re-sampling* SMOTE.

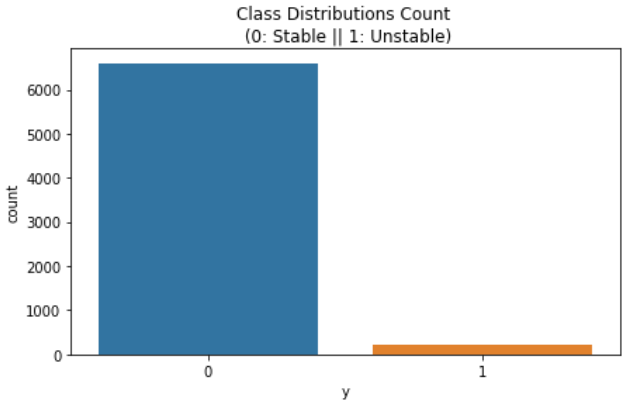
Random Forest Classifier merupakan sebuah algoritma yang menggunakan cara pembagian biner secara bertahap untuk mencapai simpul akhir pada struktur pohon berdasarkan pohon regresi dan klasifikasi[8]. *Random Forest Classifier* menghasilkan banyak pohon yang tidak saling terkait dengan menggunakan subset yang dipilih secara acak melalui teknik bootstrap dari sampel pelatihan dan variabel input di setiap *node*[19]. *Random Forest Classifier* melakukan proses pembagian ke dalam kelas dengan menggunakan pendekatan gabungan dari beberapa pohon dengan mengambil keputusan mayoritas untuk mencapai keputusan akhir[9].

Dataset train pada algoritma *Random Forest Classifier* diformulasikan dengan S = {(xi, yj), i = 1, 2, …, N; j = 1, 2, …, M}, dimana y adalah variabel fitur S dan x adalah sampel. Sedangkan N merupakan jumlah sampel pelatihan, dan disetiap sampel ada variabel fitur M[20]. Sebagai algoritma klasifikasi, *Random Forest Classifier* dapat diterapkan pada dataset dengan jumlah data tidak seimbang dengan hasil yang baik dan waktu eksekusi yang cepat[3].

3. HASIL DAN PEMBAHASAN

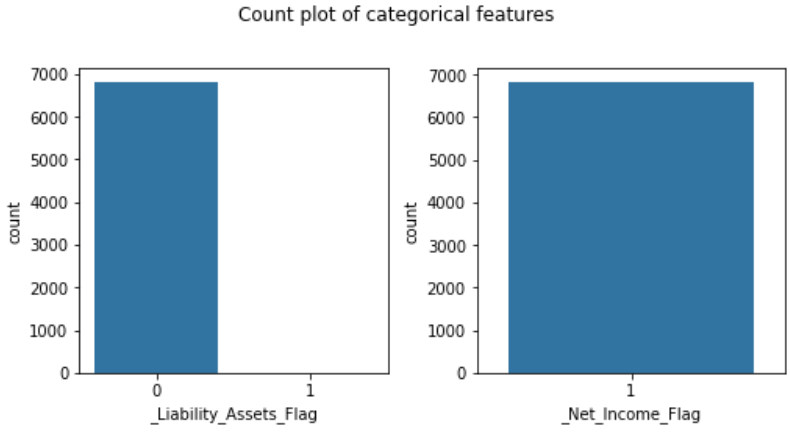
Dataset yang digunakan pada penelitian ini berasal dari *kaggle* yaitu data *bankruptcy*. Proses selanjutnya dilakukan *pre-processing* data dengan menggunakan teknik *oversampling* SMOTE. *Oversampling* SMOTE dilakukan untuk membuat data seimbang.

Dataset bankruptcy terdiri dari 6819 *record* data dan 96 kolom dengan dengan distribusi data *Financially stable* sebesar 96.77% dengan jumlah data sebanyak 6599 *record* dan *Financially unstable* sebesar 3.23% dengan jumlah data sebanyak 220 *record*. Data tersebut termasuk dalam kategori data tidak seimbang, sehingga apabila dilakukan klasifikasi akan menghasilkan nilai akurasi yang rendah disebabkan dominasi satu kelas yang begitu besar yaitu 96.77%. Grafik perbandingan kelas data dapat dilihat pada gambar 2 yang menunjukkan jumlah distribusi data dari masing-masing kelas.



Gambar 2. Distribusi Kelas data

Dari 96 fitur yang ada terdapat satu fitur data yang hanya memiliki nilai 1 yang dapat dilihat pada gambar 3 grafik *categorical features* yaitu kolom "Net income flag". Selain itu distribusi data cenderung merata, yaitu tidak ada bin yang tinggi secara signifikan lebih tinggi daripada bin lainnya. Hal ini menunjukkan bahwa tidak ada nilai yang dominan dalam data tersebut dan setiap nilai muncul dengan frekuensi yang sama.



Gambar 3. Categorical features

Dengan demikian kolom "Net income flag" dihapus dari dataset sebelum dilakukan proses selanjutnya. Tahap berikutnya adalah melakukan re-sampling data training dengan menggunakan teknik *oversampling* SMOTE. Sebelum dilakukan *re-sampling* data dibagi dua menjadi data test dan data train dengan pembagian 70% data train dan 30% data test. Didapatkan 4773 data train dan 2046 data test dengan jumlah kelas data *Financially stable* sebesar 4619 dan jumlah data *Financially unstable* sebesar 154. Setelah dilakukan *re-sampling* dengan

menggunakan SMOTE distribusi kelas menjadi sama yaitu dengan jumlah data sebesar 4619 untuk semua kelas data. Hasilnya ditunjukkan pada tabel 1 yang menyatakan perbandingan data sebelum dan sesudah dilakukan *re-sampling* dengan teknik *oversampling* SMOTE.

Tabel 1. Data train dengan SMOTE

Label (y)	row data	row data with SMOTE
0	4619	4619
1	154	4619

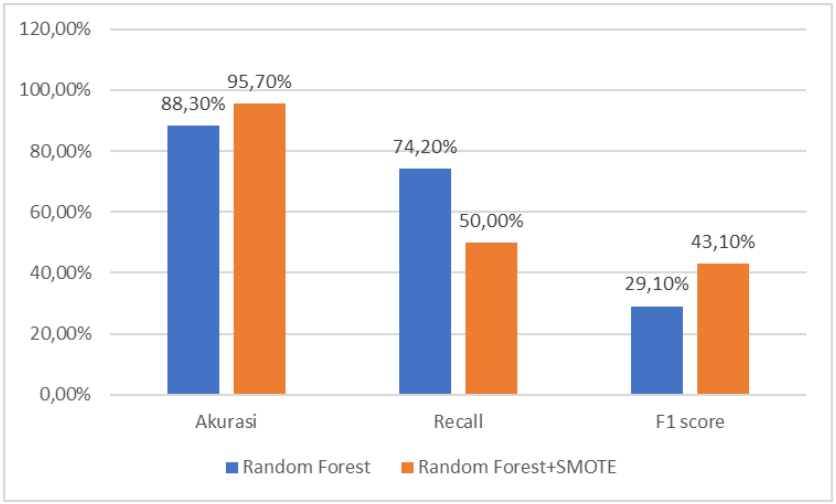
Setelah proses *re-sampling* selanjutnya diperoleh data seimbang yang siap untuk dilakukan proses berikutnya. Tahap berikutnya adalah dilakukan prediksi menggunakan algoritma *Random Forest Classifier* dengan pengujian menggunakan *cross validation*. Pada tahap ini dilakukan dua kali pengujian untuk algoritma *Random Forest Classifier* sebelum dilakukan *re-sampling* data dan sesudah dilakukan *re-sampling* data untuk melihat perbandingan hasil sebelum dan sesudah dilakukan *re-sampling* menggunakan SMOTE.

Tabel 2. Hasil pengujian

Algoritma	Akurasi	Recall	F1 score
Random Forest	88,30%	74,20%	29,10%
Random Forest + SMOTE	95,70%	50,00%	43,10%

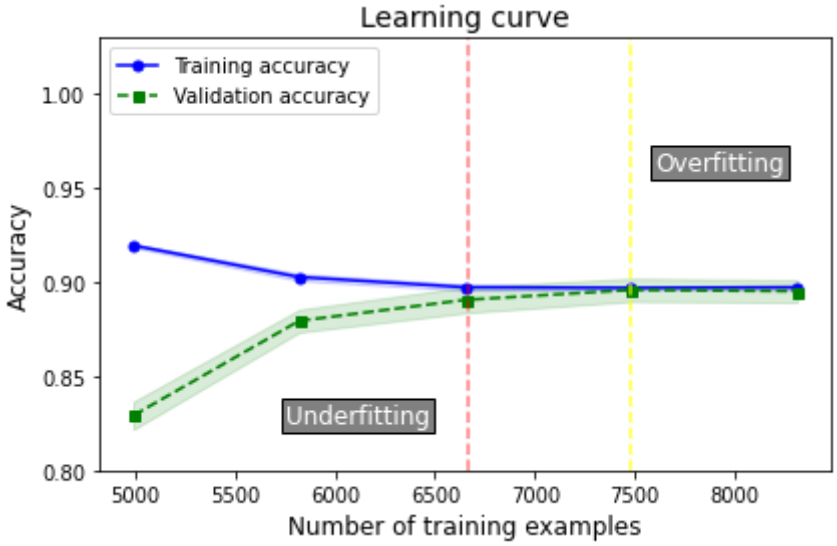
Pengujian pertama dengan dataset awal sebelum dilakukan *re-sampling* menggunakan algoritma *Random Forest Classifier* dengan *cross validation*. Hasil pengujian adalah nilai akurasi diperoleh sebesar 88,30% dengan nilai *recall* sebesar 74,20% dan nilai *F1 Score* sebesar 29,10%.

Pada pengujian kedua menggunakan data hasil *re-sampling* SMOTE menggunakan algoritma *Random Forest Classifier* dengan *cross validation*. Hasil pengujian adalah nilai akurasi diperoleh sebesar 95,70% dengan nilai *recall* sebesar 50,00% dan nilai *F1 Score* sebesar 43,10%.



Gambar 4. Grafik perbandingan akurasi

Pengujian dilakukan sebanyak 10 kali perulangan menggunakan *cross validation*. Tabel 2 dan gambar 4 menunjukkan perbandingan hasil pengujian yang dilakukan.



Gambar 4. Learning curve

Performa hasil training dataset menggunakan algoritma *Random Forest Classifier* dengan *oversampling* SMOTE disajikan pada gambar 5. Grafik tersebut, menjelaskan bahwa model tersebut mengalami *overfitting* pada jumlah data yang cukup kecil. Hal ini dapat dilihat dari perbedaan performa model yang cukup besar antara data train dan data test pada jumlah data yang kecil. Setelah jumlah data mencapai sekitar 5700, perbedaan performa antara data train dan data test mulai menyusut dan kedua kurva tersebut mulai menunjukkan tren yang lebih datar. Hal Ini terlihat bahwa model tersebut sudah cukup baik dalam memprediksi data baru dan tidak mengalami *overfitting* lagi.

4. KESIMPULAN DAN SARAN

Berdasarkan hasil pengujian model yang dilakukan dapat diambil kesimpulan bahwa penggunaan teknik *oversampling* SMOTE dapat meningkatkan nilai akurasi pada algoritma *Random Forest Classifier* untuk prediksi kebangkrutan perusahaan pada dataset tidak seimbang. Terbukti peningkatan akurasi sebesar 7,40% sedangkan *recall* mengalami penurunan sebesar 24,20%. Dari curva *learning* model performa model sudah cukup baik dalam memprediksi data baru dan tidak mengalami *overfitting*. Namun model tersebut belum mencapai puncaknya, dengan demikian masih memungkinkan untuk dilakukan Optimasi dengan metode yang lain agar mendapatkan nilai akurasi yang lebih baik lagi.

DAFTAR PUSTAKA

[1] A. S. Ramadhani and N. Lukviarman, "Perbandingan Analisis Prediksi Kebangkrutan Menggunakan Model Altman Pertama, Altman Revisi, dan Altman Modifikasi dengan Ukuran dan Umur Perusahaan sebagai Variabel Penjelas (Studi Pada Perusahaan Manufaktur yang Terdaftar di Bursa Efek Indonesia)," *J. Siasat Bisnis*, vol. 13, no. 1, Mar. 2011 [Online]. Available: https://journal.uii.ac.id/JSB/article/view/2011

[2] S. D. Saladi and R. Yarlagadda, "An Enhanced Bankruptcy Prediction Model Using Fuzzy Clustering Model and Random Forest Algorithm," *Rev. d'Intelligence Artif.*, vol. 35, no. 1, pp. 77–83, Feb. 2021, doi: 10.18280/ria.350109.

[3] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier, 2012.

[4] S. Umadevi and K. S. J. Marseline, "A survey on data mining classification algorithms," in *2017 International Conference on Signal Processing and Communication (ICSPC)*,

2017, doi: 10.1109/CSPC.2017.8305851.

[5] A. Nugroho and Y. Religia, "Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 3, pp. 504–510, Jun. 2021, doi: 10.29207/resti.v5i3.3067. [Online]. Available: http://jurnal.iaii.or.id/index.php/RESTI/article/view/3067

[6] Haibo He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239. [Online]. Available: http://ieeexplore.ieee.org/document/5128907/

[7] S. M. Abd Elrahman and A. Abraham, "A Review of Class Imbalance Problem," 2013 [Online]. Available: www.mirlabs.net/jnic/index.html

[8] Yoga Religia, Agung Nugroho, and Wahyu Hadikristanto, "Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 187–192, Feb. 2021, doi: 10.29207/resti.v5i1.2813. [Online]. Available: http://jurnal.iaii.or.id/index.php/RESTI/article/view/2813

[9] A. Parmar, R. Katariya, and V. Patel, "A Review on Random Forest: An Ensemble Classifier," 2019, pp. 758–763 [Online]. Available: http://link.springer.com/10.1007/978-3-030-03146-6_86

[10] A. Pedro Duarte Silva, "Optimization approaches to Supervised Classification," *Eur. J. Oper. Res.*, vol. 261, no. 2, Sep. 2017, doi: 10.1016/j.ejor.2017.02.020.

[11] 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.

[12] A. S. Desuky, Y. M. Elbarawy, S. Kausar, A. H. Omar, and S. Hussain, "Single-Point Crossover and Jellyfish Optimization for Handling Imbalanced Data Classification Problem," *IEEE Access*, vol. 10, pp. 11730–11749, 2022, doi: 10.1109/ACCESS.2022.3146424. [Online]. Available: https://ieeexplore.ieee.org/document/9693509/

[13] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Computing Surveys*, vol. 52, no. 4. Association for Computing Machinery, 01-Aug-2019.

[14] O. Heranova, "Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, pp. 443–450, Dec. 2019, doi: 10.29207/resti.v3i3.1275. [Online]. Available: http://jurnal.iaii.or.id/index.php/RESTI/article/view/1275

[15] K. Jiang, J. Lu, and K. Xia, "A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE," *Arab. J. Sci. Eng.*, vol. 41, no. 8, pp. 3255–3266, Aug. 2016, doi: 10.1007/s13369-016-2179-2. [Online]. Available: http://link.springer.com/10.1007/s13369-016-2179-2

[16] A. N. Rais and A. Subekti, "Integrasi SMOTE dan Ensemble AdaBoost Untuk Mengatasi Imbalance Class Pada Data Bank Direct Marketing," *J. Inform.*, vol. 6, no. 2, pp. 278–285, 2019, doi: http://dx.doi.org/10.31294/ji.v6i2.6186. [Online]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/ji/article/view/6186. [Accessed: 11-Nov-2021]

[17] Y. E. Kurniawati, "Class Imbalanced Learning Menggunakan Algoritma Synthetic Minority Over-sampling Technique – Nominal (SMOTE-N) pada Dataset Tuberculosis Anak," *J. Buana Inform.*, vol. 10, no. 2, p. 134, Oct. 2019, doi: 10.24002/jbi.v10i2.2441. [Online]. Available: https://ojs.uajy.ac.id/index.php/jbi/article/view/2441

[18] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, "Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning From Imbalanced Data," *IEEE Access*, vol. 9, pp. 74763–74777, 2021, doi: 10.1109/ACCESS.2021.3080316. [Online]. Available: https://ieeexplore.ieee.org/document/9431216/

[19] W. Zhang, "Machine Learning Approaches to Predicting Company Bankruptcy," *J. Financ. Risk Manag.*, vol. 06, no. 04, pp. 364–374, 2017, doi: 10.4236/jfrm.2017.64026.

[20] J. Chen *et al.*, "A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 4, pp. 919–933, Apr. 2017, doi: 10.1109/TPDS.2016.2603511. [Online]. Available: http://ieeexplore.ieee.org/document/7557062/