UPGRADE: Jurnal Pendidikan Teknologi Informasi
Volume 2, No. 1, Agustus, 2024, (1-10)
E-ISSN: 3021-7076
DOI: https://doi.org/10.30812/upgrade.v2i1.4148

Algoritma *Synthetic Minority Oversampling Technique* dan C5.0 dalam Mengatasi Ketidakseimbangan Data pada Klasifikasi Kelulusan Siswa

Moch Anjas Aprihartha*, Zulhandi Putrawan, Dicky Zulhan, Fatma Ahardika Nurfaizal
Universitas Dian Nuswantoro, Semarang, Indonesia

*Email Korespondensi: anjas.aprihartha@dsn.dinus.ac.id

Genesis Artikel: **Diterima**: 19 Juni 2024 **Diterbitkan**: 26 Juli 2024

ABSTRACT: *Supervised Learning algorithms are used to predict and classify certain attributes, but the main problem is the uneven distribution of data between classes, which can lead to overfitting. To overcome this, minority class augmentation using the Synthetic Minority Oversampling Technique (SMOTE) is required. This research aims to provide a practical solution to overcome data imbalance with SMOTE in the case of students who do not pass all subjects to reduce the risk of overfitting. This research method is experimental research with a quantitative approach using secondary data from students' subject graduation. The data analysis technique of SMOTE results was tested with the C5.0 algorithm, and state variations of 1 to 100 were used to ensure a random selection of training and testing data in each iteration. The results show that testing the original data with the C5.0 algorithm produces inconsistent accuracy, recall, and specificity plots while testing the data processed with SMOTE shows stable plots close to 100%. This means the SMOTE data performs better in the C5.0 algorithm than the original data. The effectiveness of the SMOTE technique and the C5.0 algorithm can contribute to researchers who face similar problems. The implications of the results of this study can also be used as a reference in making applications to detect student graduation and facilitate teachers in making decisions.*

Keyword: C5.0, Classification, Data Imbalanced, SMOTE, Supervised Learning

ABSTRAK: Algoritma *Supervised Learning* digunakan untuk memprediksi dan mengklasifikasikan atribut tertentu, namun masalah utama adalah distribusi Data yang tidak merata antar kelas yang dapat menyebabkan overfitting. Untuk mengatasi ini, diperlukan augmentasi kelas minoritas menggunakan teknik *Synthetic Minority Oversampling Technique* (SMOTE). Tujuan penelitian ini memberikan solusi praktis untuk mengatasi ketidakseimbangan Data dengan SMOTE pada kasus siswa yang tidak lulus semua mata pelajaran, guna mengurangi risiko overfitting. Metode penelitian ini adalah penelitian eksperimental dengan pendekatan kuantitatif menggunakan Data sekunder dari kelulusan mata pelajaran siswa. Teknik analisis Data hasil SMOTE diuji dengan algoritma C5.0, dan variasi state 1 hingga 100 digunakan untuk memastikan pemilihan Data training dan testing secara acak di setiap iterasi. Hasil penelitian menunjukkan bahwa uji Data asli dengan algoritma C5.0 menghasilkan plot akurasi, *recall*, dan spesifisitas yang tidak konsisten, sedangkan uji Data yang diolah dengan SMOTE menunjukkan plot yang stabil mendekati 100%. Artinya, data SMOTE memberikan performa yang lebih baik pada algoritma C5.0 dibandingkan Data asli. Efektivitas teknik SMOTE dan algoritma C5.0 dapat berkontribusi bagi peneliti yang menghadapi masalah serupa. Implikasi hasil penelitian ini juga dapat dijadikan acuan dalam membuat aplikasi untuk mendeteksi kelulusan siswa guna mempermudah guru dalam mengambil keputusan.

Kata Kunci: C5.0, *Data Imbalanced*, Klasifikasi, SMOTE, *Supervised Learning*

Cara Sitasi:
Aprihartha, M.A., Putrawan, Z., Zulhan, D., & Nurfaizal, F.A. (2024). Algoritma *Synthetic Minority Oversampling Technique* dan C5.0 dalam mengatasi ketidakseimbangan data pada klasifikasi kelulusan siswa. *UPGRADE: Jurnal Pendidikan Teknologi Informasi, 2*(1), 1-10. https://doi.org/10.30812/upgrade.v2i1.4148

PENDAHULUAN

Supervised Learning merupakan komponen dalam kecerdasan buatan yang melibatkan atribut luaran yang telah ditentukan selain penggunaan atribut masukan. Algoritma *Supervised Learning* diaplikasikan dalam memprediksi dan mengklasifikan atribut yang telah ditentukan. Tingkat keakuratan dan kesalahan klasifikasinya serta ukuran kinerja lainnya bergantung pada jumlah atribut yang telah ditentukan yang diprediksi dengan benar atau sebaliknya (Berry et al., 2019). Masalah krusial yang dialami pada algoritma *Supervised Learning* mengacu pada kumpulan data dengan distribusi sampel data yang tidak merata antarkelas yang berbeda. Fenomena ini disebut ketidakseimbangan data. Ketidakseimbangan data terjadi apabila terdapat kejadian langka dalam proses mengumpulkan data. Umumnya, sampel di kelas minoritas membawa informasi yang lebih penting dan model klasifikasi yang baik sangat diperlukan dalam kebanyakan kasus (Zhang et al., 2024). Namun, model yang didasarkan pada desain yang berorientasi pada akurasi biasanya menunjukkan kinerja yang tidak diinginkan pada kelas minoritas. Hal ini mengakibatkan model akan rentan mengalami *overfitting*. Guna menghindari masalah *overfitting* dan menanggani kejadian langka maka diperlukan augmentasi kelas minoritas dengan implementasi teknik *Synthetic Minority Oversampling Technique* (SMOTE) (Rahim et al., 2023).

Teknik SMOTE berbeda dengan menyalin objek yang sudah ada. Menduplikasi sampel dapat mengakibatkan representasi yang berlebihan titik data tertentu, mengurangi keragaman dan memperpanjang durasi pelatihan (Xiaolong et al., 2019). Menduplikasi kelas minoritas dapat mengakibatkan *overfitting* (Alex et al., 2024; Aryanti et al., 2023). Sedangkan, penghapusan sampel secara acak dari kumpulan data dapat menghilangkan informasi penting (Alex et al., 2024). SMOTE merupakan teknik yang menciptakan sampel baru dengan interpolasi acak antara sampel kelas minoritas dan k tetangga terdekat (Wang et al., 2024). Penelitian terkait permasalahan keseimbangan data dengan SMOTE telah digunakan pada penelitian sebelumnya.

Beberapa peneltian sebelumnya yang dilakukan oleh oleh Xiao et al. (2024) yang menggunakan algoritma *stacking* pada SMOTE untuk memprediksi tingkat ledakan batuan di empat lokasi tambang emas. Basis data yang dikumpulkan diambil sampelnya secara berlebihan dengan teknik SMOTE untuk menyeimbangkan kategori data ledakan batuan. Hasil penelitian diperoleh akurasi mencapai 100%. Selanjutnya, penelitan yang dilakukan oleh Gamel et al. (2024) dalam mengatasi ketidakseimbangan data dalam prediksi diagnosis transformator daya dengan memanfaatkan SMOTE untuk menghasilkan data sintetis dan menyeimbangkan data. Hasil menunjukkan kombinasi SMOTE dan *Deep Neural Network* (DNN) mencapai akurasi 98,22 pada data *training* dan 94,6% pada data testing. Kemudian, penelitian Widodo et al. (2024) yang mengekplorasi kinerja *Intrusion Detection System* (IDS) menggunakan SMOTE dalam berbagai algoritma klasifikasi. Hasil menunjukkan penerapan SMOTE dapat meningkatkan kinerja secara keseluruhan, dengan akurasi, presisi, *recall*, dan f1-*score*.

Sementara itu, penelitian oleh Putro and Setiadi (2023) yang meneliti tentang proses pengambilan keputusan pada tingkat kelulusan siswa sekolah dasar di Kecamatan Juai menggunakan algoritma klasifikasi *Decision Tree* C4.5. Hasil penelitian diperoleh akurasi sebesar 96,67%. Data yang digunakan pada riset tersebut ditemukan jumlah kelas yang tidak lulus dan kelas lulus memiliki perbedaan yang jauh signifikan. Hal ini dapat terjadi karna jumlah siswa yang tidak lulus disekolah tersebut sangat sedikit. Dalam proses klasifikasi data, performa model *Decision Tree* dapat menjadi tidak akurat pada kelas minoritas dengan kelas mayoritas (Vebriyanti et al., 2024). Berdasarkan analisis terhadap beberapa penelitian sebelumnya, dibutuhkan penangganan khusus dalam mengatasi ketidakseimbangan data melalui metode SMOTE untuk menghasilkan performa model yang lebih baik. Oleh karena itu, kebaruan penelitian ini adalah menangani ketidakseimbangan data dengan teknik SMOTE yang mana penelitian tersebut belum banyak dilakukan oleh peneliti sebelumnya.

Data yang digunakan pada penelitian ini diperoleh dari Putro and Setiadi (2023). Data hasil SMOTE akan diuji dengan algoritma C5.0, algoritma pengembangan dari C4.5, kemudian hasil uji akan dibandingkan dengan data asli untuk melihat perbedaan performa model antarkeduanya. Tujuan penelitian ini untuk memberikan solusi praktis tentang mengatasi ketidakseimbangan data dengan teknik SMOTE pada kejadian langka seperti terdapat siswa yang tidak lulus keseluruhan mata pelajaran, sehingga dapat menurunkan kejadian *overfitting* pada model klasifikasi. Implikasi hasil penelitian ini diharapkan dapat

dijadikan acuan dalam membuat aplikasi berbasis mendeteksi kelulusan siswa agar mempermudah guru dalam mengambil keputusan.

METODE

A. Metode dan Variabel Penelitian

Metode penelitian ini adalah penelitian eksperimental dengan pendekatan penelitian kuantitatif Metode ini mengukur dan menganalisis variabel pada data numerik ataupun kategorik. Jenis sumber data yang menjadi acuan penelitian ini adalah data sekunder yang berasal dari penelitian yang dilakukan oleh Putro and Setiadi (2023) sehingga peneliti mengkaji tentang algoritma C4.5 dalam klasifikasi kelulusan siswa sekolah dasar. Siswa dikatakan lulus sekolah apabila nilai rata-rata keseluruhan mata pelajaran yang diuji diatas 70 (Putro and Setiadi, 2023). Adapun variabel yang digunakan dalam penelitian tersebut disajikan dalam Tabel 1.

Tabel 1. Variabel dan Jenis Data

Variabel	Jenis Data
Status	Kategorik: 1. Lulus 2. Tidak Lulus
Nilai Pendidikan Agama	Numerik
Nilai PKN	Numerik
Nilai Bahasa Indonesia	Numerik
Nilai Matematika	Numerik
Nilai IPA	Numerik
Nilai IPS	Numerik
Nilai SBK	Numerik
Nilai Penjaskes	Numerik
Nilai Bataqu	Numerik
Nilai BSB	Numerik

B. *Z-Score Normalization*

Z-Score Normalization atau standarisasi merupakan teknik yang paling sering digunakan dalam analisis statistik dalam pembelajaran mesin (Kim et al., 2024). Rumus *Z-Score Normalization* untuk setiap variabel dapat dilihat pada Persamaan 1. Berkaitan dengan itu, x_i adalah data asli ke-i variabel ke-j, $\bar{x}$ adalah rata-rata dari variabel ke-j, dan sd(x_j) adalah standar deviasi variabel ke-j.

$$x_{ij'} = \frac{x_{ij} - \bar{x}_j}{sd(x_j)} \tag{1}$$

C. *Synthetic Minority Oversampling Technique* (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) merupakan teknik dalam mengatasi ketidakseimbangan data dengan cara meningkatkan jumlah data sampel pada kelas minoritas. Data yang ditambahkan adalah data sintetik yang dihasilkan dari replikasi data sampel kelas minoritas. Proses menghasilkan data sintetik melibatkan perhitungan jarak Euclidean dengan k tetangga terdekat dari setiap sampel kelas minoritas menggunakan persamaan 2. Berkaitan dengan itu, Widodo et al. (2024) menyatakan bahwa dalam membangkitkan data sintetis dapat menggunakan persamaan 3.

$$d(X, \hat{X}) = \sqrt{\sum_{i=1}^{n} (x_i - \hat{h}_i)^2} \tag{2}$$

Dipublikasi oleh Universitas Bumigora

$$X_{baru} = X_i + (\hat{X}_i - X_i) \times \delta \tag{3}$$

D. Algoritma C5.0

Algoritma C5.0 merupakan algoritma pohon keputusan yang diklaim lebih efisien dibandingkan C4.5 dalam hal memori dan waktu komputasi (Abidin et al., 2023). Algoritma C5.0 memperhitungkan *entropy* sebagai ukuran kemurnian dalam membentuk pohon keputusan. *Entropy* dapat dihitung menggunakan persamaan 4. S adalah himpunan data, c adalah pada jumlah kelas, $p_i = n_i/|s|$ adalah proporsi nilai dalam kelas i, n_i adalah jumlah data pada kelas ke-i, dan $|s|$ adalah jumlah data pada himpunan S.

Algoritma menghitung perubahan homogenitas yang akan dihasilkan dari pemisahan pada setiap kemungkinan fitur yang disebut sebagai *information gain* Lantz (2019). *Information gain* dapat dihitung dengan persamaan 5. Selain itu, penggunaan dengan *entropy* (S_2) adalah *entropy* pada partisi hasil pemisahan yang dapat dihitung menggunakan persamaan 6. $p'_j = (|s_j|)/|s|$ adalah proporsi jumlah sampel variabel independen A terhadap total sampel pada himpunan S, $p_{ij} = n_{ij}/(|s_j|)$ adalah probabilitas S bersyarat A. $|s_j|$ adalah jumlah data dengan variabel A. Menurut (Nurkholis et al., 2021) nilai gain rasio variabel A dihitung dengan persamaan 7 sedangkan dengan $Split(A) = \sum_{j=1}^{v} p_{j'} log_2(p_j)$

$$entropy(S) = \sum_{i=1}^{c} -p_i log_2(p_i) \tag{4}$$

$$IG(A) = entropy(S) - entropy(S_2) \tag{5}$$

$$entropy(S_2) = \sum_{j=1}^{v} p_{j'} \sum_{i=1}^{c} -p_{ij} log_2(p_{ij}) \tag{6}$$

$$GR(A) = \frac{IG(A)}{Split} \tag{7}$$

E. Uji Performa Model

Confusion matrix merupakan alat yang digunakan dalam mengevaluasi efektifitas kemampuan model klasifikasi dan mendapatkan data terkait kesalahan klasifikasi yang dihasilkan (Sano et al., 2023). *Confusion matrix* mengandung informasi berupa total terklasifikasi valid pada kelas positif (TP), total terklasifikasi valid pada kelas negatif (TN), total terklasifikasi tidak valid pada kelas positif (FP), total terklasifikasi tidak valid pada kelas negatif (FN). Sehubungan dengan itu, Aprihartha et al., (2024) menyampaikan bentuk *confusion matrix* dapat dilihat pada Tabel 2. Selain itu, Aprihartha et al. (2024) juga menyampaikan rumus dalam mengukur evaluasi kinerja model dapat dilihat pada Tabel 3.

Tabel 2. Confusion Matrix

	Kelas Prediksi: Ya	**Kelas Prediksi: Tidak**
Kelas Observasi: Ya	TP	FN
Kelas Observasi: Tidak	FP	TN

Tabel 3. Ukuran Evaluasi Model

Ukuran	**Rumus**
Akurasi	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall	$\frac{TP}{TP+FN}$
Spesifisitas	$\frac{TN}{FP+TN}$

HASIL DAN PEMBAHASAN

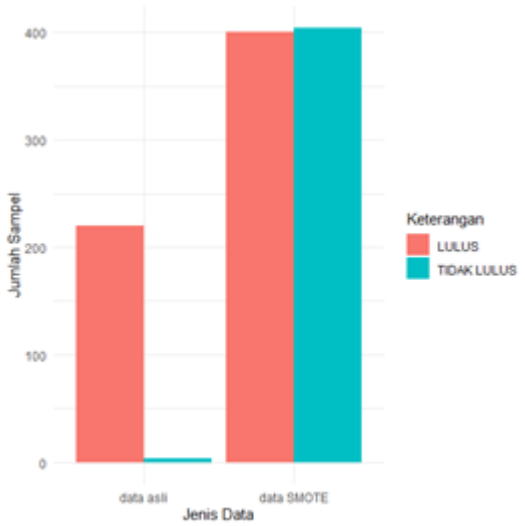
Sebelum analisis lebih lanjut dengan algoritma SMOTE dan C5.0. Data terlebih dahulu dimodifikasi dengan transformasi *Z-Score* menggunakan persamaan 1. Lima data teratas hasil transformasi *Z-Score* ditampilkan pada Tabel 4.

Tabel 4. Data Hasil Transformasi *Z-Score*

No	PA	PKN	B. Indo	...	Penjaskes	Bataqu	Status
1	-0,270	0,458	-0,738	...	-0,451	0,393	Lulus
2	1,412	2,514	2,728	...	0,932	0,623	Lulus
3	0,023	-0,457	-0,041	...	-0,727	0,412	Lulus
4	0,170	1,168	1,859	...	-0,404	0,454	Lulus
5	0,046	0,617	0,331	...	0,056	0,374	Lulus

B. Hasil Implementasi SMOTE

Teknik dasar SMOTE adalah membuat data sintesis baru berdasarkan data asli. Setiap data akan diperhitungkan tetangga terdekatnya dengan menggunakan jarak Euclidean. Apabila diketahui terdapat data yang saling berdekatan satu sama lain maka akan dibuat data baru diantara data tersebut.



Gambar 1. Frekuensi Data Asli dan Data SMOTE

Pada Gambar 1 memperlihatkan frekuensi data asli dan data yang diolah dengan SMOTE. Pada data asli, frekuensi kategori siswa yang lulus lebih mendominasi dibandingkan siswa yang tidak lulus. Setelah digunakan teknik SMOTE, jumlah sampel untuk kedua kategori menjadi meningkat dan seimbang. Frekuensi kategori yang lulus sebanyak 400 siswa dan tidak lulus sebanyak 404 siswa.

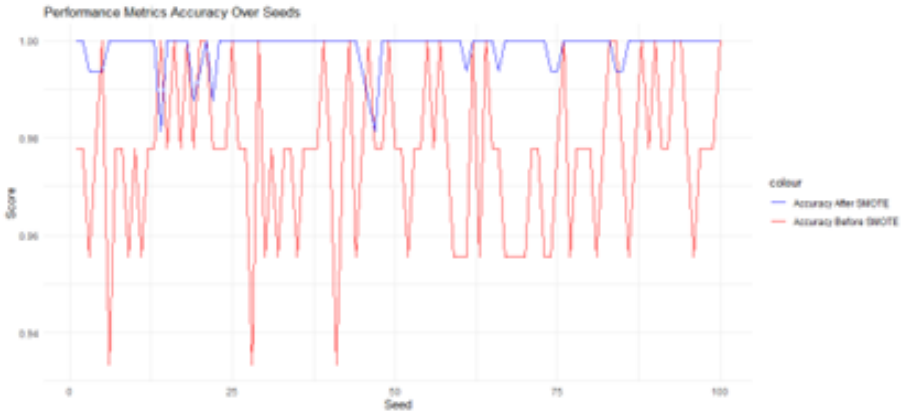
C. Perbandingan Performa Model C5.0

Studi ini menerapkan algoritma C5.0 dalam klasifikasi data kelulusan siswa. Persentase data *training* dan data *testing* berturut-turut yaitu 80% dan 20%. Pengambilan sampel untuk data *training* menggunakan teknik *simple random sampling*, dimana setiap data memiliki peluang yang sama terambil sebagai data *training*. Pengujian dilakukan dengan dua jenis dataset, yaitu data asli dan data yang diolah dengan SMOTE. Selain itu, penelitian ini melibatkan variasi state 1 sampai 100 untuk memastikan data *training* dan data testing dipilih secara acak disetiap iterasi. Ini bertujuan untuk melihat konsistensi kemampuan model dalam klasifikasi data dari *state* 1 sampai 100. Hasil uji pada salah satu *state* pada ditunjukkan pada Tabel 5.

Tabel 5. Performa Model C5.0 pada State ke-*i*

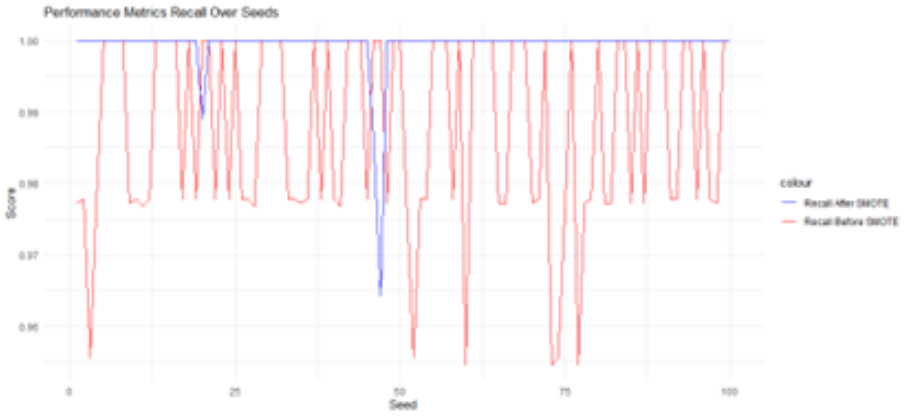
Pengukuran	Data Asli (%)	Data SMOTE (%)
Akurasi	97,7	100
Recall	97,7	100
Spesifisitas	0,00	100

Merujuk pada Tabel 5 memperlihatkan hasil uji data asli yang diperoleh akurasi 97,7%, artinya model C5.0 dapat memprediksi dengan tepat siswa yang lulus dan tidak lulus dengan tingkat ketelitian 97,7%. *Recall* sebesar 97,7% berarti model dapat memprediksi dengan tepat siswa yang lulus dengan tingkat ketelitian 97,7% dan spesifisitas sebesar 0,00% berarti model sama sekali tidak dapat memprediksi siswa yang tidak lulus. Hal ini disebabkan model tidak mendapatkan informasi terkait siswa yang tidak lulus. Sementara itu, data SMOTE memberikan performa sempurna yang ditunjukkan dengan akurasi, *recall*, dan spesifisitas berturut-turut 100%. Hasil uji performa model C5.0 pada 100 *state* disajikan pada Gambar 2, Gambar 3, dan Gambar 4.



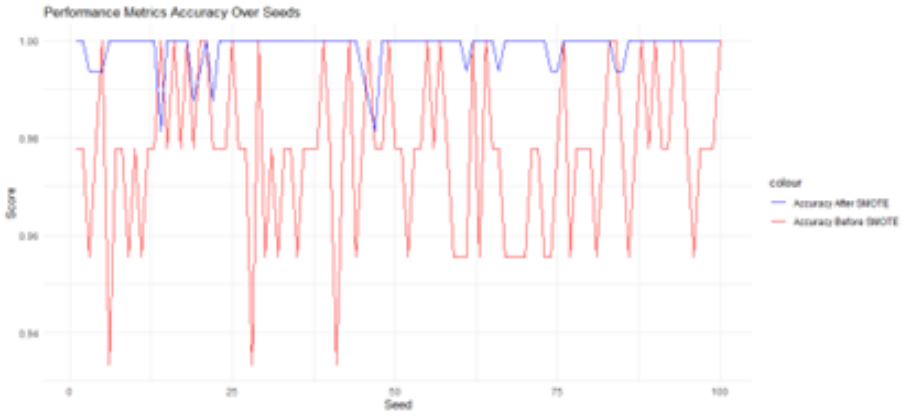
Gambar 2. Performa Akurasi Data Asli dan Data SMOTE

Gambar 2 ditunjukkan akurasi pada dua jenis dataset berbeda yang diolah dengan algoritma C5.0. Plot garis berwarna merah adalah akurasi data asli sedangkan warna biru adalah akurasi data dengan SMOTE. Pada plot akurasi data asli memperlihatkan lebih banyak fluktuasi yang mengindikasikan akurasi setiap *state* berubah-ubah dikisaran 0,93 (93%) sampai 1 (100%). Sedangkan plot akurasi data SMOTE memperlihatkan garis cenderung stabil dan tinggi, akurasi setiap *state* rata-rata konsisten pada 1 (100%). Algoritma C5.0 pada data SMOTE memperlihatkan konsistensi akurasi yang lebih baik dalam setiap state dibandingkan data asli.



Gambar 3. Performa *Recall* Data Asli dan Data SMOTE

Gambar 3 ditunjukkan plot garis *recall* pada data asli (merah) dan data SMOTE (biru). Plot garis pada data asli memperlihatkan *recall* pada setiap *state* tidak stabil. Kemampuan model mengidentifikasi kategori lulus berada pada rentang 0,95 (95%) sampai 1 (100%). Sebaliknya plot garis pada data SMOTE memperlihatkan *recall* cenderung stabil pada 1 (100%) yang berarti kemampuan model setiap *state* berhasil mengklasifikasikan kategori lulus dengan tepat sebesar 100%. Model C5.0 dengan data SMOTE memberikan stabilitas *recall* lebih baik dibandingkan data asli.



Gambar 4. Performa Spesifisitas Data Asli dan Data SMOTE

Gambar 4 memperlihatkan plot garis spesifisitas data asli (merah) dan data SMOTE (biru). Pada data asli menunjukkan garis terlihat turun tajam ke nilai 0 dibeberapa *state*. Ini dikarenakan model sangat buruk dalam mengenali data kategori tidak lulus akibat dari data yang tidak seimbang. Sebaliknya, pada data SMOTE menunjukkan garis yang relatif konsisten mendekati 1 (100%). Ini berarti setelah data asli diseimbangkan dengan teknik SMOTE, model dapat memprediksi dengan sangat baik data kategori tidak lulus. Dengan demikian, berdasarkan ukuran akurasi, *recall*, dan spesifisitas maka data yang diolah dengan SMOTE memberikan performa yang sangat baik bagi algoritma C5.0 dibandingkan dengan data aslinya.

Temuan penelitian ini menunjukkan bahwa hasil penelitian dengan algoritma SMOTE-C5.0 memberikan peningkatan kualitas dari akurasi model dibandingkan dengan algoritma C4.5 pada data asli karena mampu mengatasi kelemahan-kelemahan seperti menagani masalah keseimbangan data dan efisiensi dalam hal memori dan komputasi waktu. Temuan penelitian tersebut sesuai dengan penelitian oleh Prasetya (2021), peneliti melakukan eksperimen klasifikasi kanker serviks dengan membandingkan data asli dan data SMOTE pada algoritma *Random Forest* dan k-*Neareast Neighbor* (KNN). Hasil penelitian diperoleh metode SMOTE-*Random Forest* dan SMOTE-KNN terbukti lebih superior dibandingkan metode klasifikasi tanpa SMOTE. Selain itu, dalam penelitian Kotb and Ming (2021) menerapkan berbagai macam pengklasifikasi untuk menilai kinerja keluarga SMOTE dalam mengatasi masalah ketimpangan data pada kasus kegagalan pembayaran premi di perusahaan asuransi. Pengklasifikasi terdiri dari Regresi Logistik, *Classification and Regression Tree* (CART), C4.5, C5.0, *Support Vector Machine (SVM), Bagged CART, Random Forest, Adaboost, Stochastic Gradient Boosting, Extreme Gradient Boosting, Naive Bayes, KNN, dan Neural Network*. Dalam prosesnya menggunakan validasi model *Random Hold-Out*. Temuan yang diperoleh dengan menggunakan ukuran kinerja model klasifikasi menunjukkan bahwa algoritma *Machine Learning* tidak bekerja dengan baik pada data tidak seimbang. Masalah data yang tidak seimbang harus diatasi dengan oleh teknik SMOTE untuk meningkatkan kinerja pengklasifikasi. Selain itu, pada uji Friendman menegaskan metode SMOTE *hybrid* lebih baik dibandingkan metode SMOTE lainnya, khususnya SMOTE-TOMEK yang memiliki kinerja lebih baik dibandingkan pendekatan sampel ulang lainnya. Dengan demikian, kebaruan penelitian ini adalah penggunaan algoritma SMOTE-C5.0 telah terbukti meningkatkan akurasi model dibandingkan penggunaan algoritma C4.5. Hal ini menunjukkan bahwa penanganan masalah ketidakseimbangan data menggunakan SMOTE dapat menghasilkan model yang lebih akurat.

Dipublikasi oleh Universitas Bumigora

KESIMPULAN

Berdasarkan hasil dan pembahasan diperoleh beberapa kesimpulan. Data asli terdeteksi kelas yang tidak seimbang dengan kelas lulus sebanyak 220 amatan dan kelas tidak lulus sebanyak 4 amatan. Hasil uji dengan algoritma C5.0 menunjukkan plot garis akurasi dan *recall* memberikan hasil yang tidak konsisten pada setiap *state*. Plot garis spesifisitas menunjukkan garis yang turun tajam ke nilai 0 dibeberapa *state*. Ini dikarenakan model tidak dapat memprediksi dengan baik data kelas tidak lulus. Setelah dilakukan SMOTE, jumlah kelas lulus menjadi 400 amatan dan kelas tidak lulus menjadi 404 amatan. Hasil uji menunjukkan plot garis akurasi, *recall*, dan spesifisitas relatif stabil mendekati 1 (100%). Ini berarti data yang diolah dengan SMOTE memberikan performa yang sangat baik bagi algoritma C5.0 dibandingkan dengan data aslinya.

Algoritma SMOTE-C5.0 memberikan peningkatan kualitas dari akurasi model dibandingkan dengan algoritma C4.5 pada data asli, karena mampu mengatasi kelemahan-kelemahan seperti menagani masalah keseimbangan data dan efisiensi dalam hal memori dan komputasi waktu. Oleh sebab itu, kebaruan penelitian ini adalah penggunaan algoritma SMOTE-C5.0 telah terbukti meningkatkan akurasi model dibandingkan penggunaan algoritma C4.5. yang menunjukkan bahwa penanganan masalah ketidakseimbangan data menggunakan SMOTE dapat menghasilkan model yang lebih akurat.

Guna mengoptimalkan hasil penelitian ini, bagi penelitian kedepannya disarankan dapat dikembangkan dengan varian dari SMOTE atau algoritma lain seperti *Random Oversampling* atau *Random Undersampling*. Selain itu, teknik klasifikasi yang digunakan hanya metode C5.0. Penelitian selanjutnya dapat dibandingkan dengan algoritma lain seperti Regresi Logistik, KNN, *Naive Bayes*, *Random Forest*, SVM, dll atau dikombinasikan dengan *ensemble learning* seperti *bagging* dan *boosting* untuk meningkatkan performa model klasifikasi.

UCAPAN TERIMA KASIH

Terima kasih kepada Tuhan YME atas rahmat-Nya. Terima kasih pula kepada Arif Wicaksono Gegadang Putro dan Tedy Setiadi atas data yang telah disediakan dalam papernya. Tidak lupa ucapan terima kasih pada pihak yang telah memberikan bantuan dan dukungan sehingga penelitian ini dapat dilakukan hingga tuntas.

DEKLARASI

Taksonomi Peran Kontributor

Semua penulis berkontribusi sama sebagai kontributor utama dari makalah ini. Semua penulis membaca dan menyetujui makalah akhir.

Pernyataan Pendanaan

Penelitian ini tidak menerima hibah khusus dari lembaga pendanaan di sektor publik, komersial, atau nirlaba.

DAFTAR PUSTAKA

Abidin, Z., Nurhana, E., Permata, P., and Ulum, F. (2023). Analisis perbandingan Algoritma Decision Tree C4.4 dan C5.0 pada data karyawan berpotensi promosi jabatan. *Jurnal Teknoinfo*, 17(2):567–582. https://doi.org/10.33365/jti.v17i2.2702.

Alex, S. A., Nayahi, J. J. V., and Kaddoura, S. (2024). Deep Convolutional Neural Networks with genetic Algorithm-based Synthetic Minority Over-Sampling Technique for improved imbalanced data classification. *Applied Soft Computing*, 156:111491. https://doi.org/10.1016/j.asoc.2024.111491.

Aprihartha, M. A., Astutik, F., and Sulistianingsih, N. (2024). Comparison of Naïve Bayes, CART, dan CART Adaboost methods in predicting tire product sales. *Jurnal Matematika, Statistika dan Komputasi*, 20(3):596–605. https://doi.org/10.20956/j.v20i3.33187.

Aryanti, R., Misriati, T., and Sagiyanto, A. (2023). Analisis sentimen aplikasi primaku menggunakan Algoritma Random Forest dan SMOTE untuk mengatasi ketidakseimbangan data. *Journal of Computer System and Informatics (JoSYC)*, 5(1):218–227. https://doi.org/10.47065/josyc.v5i1.4562.

Berry, M. W., Mohamed, A., and Yap, B. W. (2019). *Supervised and unsupervised learning for data science*. Springer Nature.

Gamel, S. A., Ghoneim, S. S. M., and Sultan, Y. A. (2024). Improving the accuracy of diagnostic predictions for power transformers by employing a hybrid approach combining SMOTE and DNN. *Computers and Electrical Engineering*, 117:109232. https://doi.org/10.1016/j.compeleceng.2024.109232.

Kim, Y.-S., Kim, M. K., Fu, N., Liu, J., Wang, J., and Srebric, J. (2024). Investigating the Iimpact of data normalization methods on predicting electricity consumption in a building using different Artificial Neural Network Models. *Sustainable Cities and Society*, page 105570. https://doi.org/10.1016/j.scs.2024.105570.

Kotb, M. H. and Ming, R. (2021). Comparing SMOTE family techniques in predicting insurance premium defaulting using machine learning models. *International Journal of Advanced Computer Science and Applications*, 12(9).

Lantz, B. (2019). *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.

Nurkholis, A., Alita, D., Sucipto, A., Chanafy, M., and Amalia, Z. (2021). Hotspot classification for forest fire prediction using C5. 0 Algorithm. In *2021 International Conference on Intelligent Cybernetics Technology Applications (ICICyTA)*, pages 12–16. IEEE. https://doi.org/10.1109/ICICyTA53712.2021.9689085.

Prasetya, J. (2021). *Perbandingan analisis klasifikasi SMOTE Random Forest dan SMOTE K-Nearest Neighbors pada data tidak seimbang*. Universitas Gadjah Mada.

Putro, A. and Setiadi, T. (2023). Penerapan klasifikasi Decision Tree (C4.5) untuk memprediksi kelulusan siswa sekolah dasar di Kecamatan Juai. *Jurnal Format*, 12(2):151–157.

Rahim, A. M. A., Pratiwi, I. Y. R., and Fikri, M. A. (2023). Klasifikasi penyakit jantung menggunakan metode Synthetic Minority Over-Sampling Technique dan Random Forest Clasifier. *Indonesian Journal of Computer Science*, 12(5). https://doi.org/10.33022/ijcs.v12i5.3413.

Sano, A. V. D., Stefanus, A. A., Madyatmadja, E. D., Nindito, H., Purnomo, A., and Sianipar, C. P. M. (2023). Proposing a visualized comparative review analysis model on tourism domain using Naïve Bayes classifier. *Procedia Computer Science*, 227:482–489. https://doi.org/10.1016/j.procs.2023.10.549.

Vebriyanti, L. M. L., Martha, S., Andani, W., and Rizki, S. W. (2024). Analisis kelayakan kredit menggunakan Classification Tree dengan teknik Random Oversampling. *Euler: Jurnal Ilmiah Matematika, Sains dan Teknologi*, 12(1):1–8. https://doi.org/10.37905/euler.v12i1.24182.

Wang, F., Zheng, M., Hu, X., Li, H., Wang, T., and Chen, F. (2024). FIAO: Feature Information Aggregation Oversampling for imbalanced data classification. *Applied Soft Computing*, 161:111774. https://doi.org/10.1016/j.asoc.2024.111774.

Widodo, A. O., Setiawan, B., and Indraswari, R. (2024). Machine learning-based intrusion detection on multi-class imbalanced dataset using SMOTE. *Procedia Computer Science*, 234:578–583. https://doi.org/10.1016/j.procs.2024.03.042.

Xiao, P., Liu, Z., Zhao, G., and Pan, P. (2024). Novel stacking models based on SMOTE for the prediction of rockburst grades at four deep gold mines. *Underground Space*. https://doi.org/10.1016/j.undsp.2024.03.004.

Xiaolong, X. U., Wen, C. H. E. N., and Yanfei, S. U. N. (2019). Over-Sampling Algorithm for imbalanced data classification. *Journal of Systems Engineering and Electronics*, 30(6):1182–1191. https://doi.org/10.21629/JSEE.2019.06.12.

Zhang, Z., Tian, H. P., and Jin, J. S. (2024). Multiple adaptive over-sampling for imbalanced data evidential classification. *Engineering Applications of Artificial Intelligence*, 133:108532. https://doi.org/10.1016/j.engappai.2024.108532.