JSiI | Jurnal Sistem Informasi
Vol. 12 No. 1 | Maret 2025, Hal. 119-127

DOI 10.30656/jsii.v11i2.9065
p-ISSN: 2406-7768
e-ISSN: 2581-2181

Perbandingan Model Pembelajaran Mesin Berbasis Smote Meningkatkan Identifikasi Siswa Berisiko di Sekolah Menengah Pertama

Hevi Alvina Damayanti[1], Ucta Pradema Sanjaya[2]

[1,2] *Jurusan* Teknik Informatika, Universitas Ngudi Waluyo
Jl. Diponegoro No.186, Ngablak, Gedanganak, Kec. Ungaran Tim., Kabupaten Semarang, Jawa Tengah 50512
[1]hevialvin@gmail.com
[2]uctapradema@unw.ac.id

Abstract

Penelitian ini mengevaluasi efektivitas Teknik Pengambilan Sampel Berlebih Minoritas Sintetis (SMOTE) dalam mengatasi ketidakseimbangan kelas pada kumpulan data pendidikan, dengan fokus pada peningkatan model prediktif untuk mengidentifikasi siswa berisiko di sekolah menengah pertama di Indonesia. Fakta lapangan menunjukkan bahwa tidak semua data yang diambil memiliki ketidakseimbangan kelas. Kinerja model Decision Tree, Random Forest, dan SVM dinilai menggunakan kurva ROC dan metrik AUC sebelum dan setelah penerapan SMOTE. Random Forest menunjukkan peningkatan AUC paling signifikan (0,95→0,99) karena generalisasi ensembel pada data yang seimbang, sementara Decision Tree mengalami peningkatan marjinal (0,94→0,95) dan SVM mengalami trade-off kecil (0,93→0,94) karena sensitivitas terhadap noise sintetis. Semua model mengungguli tebakan acak (AUC>0,93), mengonfirmasi manfaat SMOTE dalam meningkatkan deteksi kelas minoritas untuk aplikasi seperti identifikasi siswa berisiko. Penelitian ini memajukan kerangka kerja praktis untuk memanfaatkan teknik pembelajaran ketidakseimbangan kelas pada dataset dalam pendidikan, menekankan wawasan yang dapat ditindaklanjuti dan meningkatkan hasil pembelajaran melalui pengambilan keputusan berbasis data. Dengan menerapkan teknik seperti SMOTE, sekolah dapat lebih akurat mengidentifikasi siswa berisiko, memungkinkan intervensi dini yang efektif dan alokasi sumber daya yang lebih efisien. Selain itu, penelitian ini mendorong pengembangan kebijakan pendidikan yang inklusif dan adil, mengurangi bias sistemik serta meningkatkan hasil pembelajaran secara keseluruhan.

Keywords: Data Mining, Decision Tree C4.5, academic performance, educational classification, determinant factors

I. PENDAHULUAN

Strategi sumber daya manusia yang kuat bergantung pada fondasi pendidikan yang kuat untuk menumbuhkan tenaga kerja yang kompetitif, menggarisbawahi peran penting sistem pendidikan modern dalam mendorong perkembangan siswa [1]. Inti dari tujuan ini adalah penerapan metodologi penilaian yang efektif, yang berfungsi sebagai alat untuk mengevaluasi pemahaman siswa, menginformasikan keputusan kurikuler, dan menyempurnakan praktik instruksional [2]. Penilaian, yang didefinisikan sebagai "perolehan informasi secara sistematis untuk memandu keputusan mengenai siswa, kurikulum, metode pengajaran, atau kebijakan pendidikan", sangat penting dalam memastikan akuntabilitas dan peningkatan berkelanjutan dalam lembaga pendidikan [3]. Studi ini bertujuan untuk meningkatkan kualitas pendidikan di Sekolah Menengah Pertama (SMP) dan Madrasah Tsanawiyah (MTs) di Indonesia dengan mengatasi kesenjangan dalam praktik penilaian dan pemanfaatan data yang ada.

Untuk mencapai tujuan ini, sangat penting untuk mengidentifikasi dan menganalisis berbagai faktor yang mempengaruhi kinerja siswa. Sekolah secara rutin mengumpulkan beragam data - termasuk catatan akademik[3], pola kehadiran, keterlibatan ekstrakurikuler [4], latar belakang sosial ekonomi[5], dan akses ke fasilitas pembelajaran - yang memiliki potensi yang belum dimanfaatkan untuk memahami faktor-faktor penentu keberhasilan akademik[6]. Namun, ketergantungan pada pemrosesan data manual dan metode analisis tradisional telah menyebabkan kurangnya pemanfaatan sumber daya ini secara signifikan, sehingga membatasi kemampuan pendidik untuk mendapatkan wawasan yang dapat ditindaklanjuti [7]. Dengan memanfaatkan pendekatan berbasis data yang canggih, penelitian ini berupaya mengubah data pendidikan mentah menjadi pengetahuan strategis, sehingga memungkinkan para pemangku kepentingan untuk merancang intervensi yang ditargetkan untuk mengatasi ketidaksetaraan sistemik dan mengoptimalkan hasil pembelajaran [8].

Dalam pemodelan prediktif, masih menjadi tantangan yang kritis, terutama pada dataset di mana distribusi variabel target sangat miring [9]. Masalah ini sering kali membuat algoritma konvensional memprioritaskan kelas mayoritas dan mengabaikan kelas minoritas (misalnya, siswa yang berisiko putus sekolah) [10], sehingga menghasilkan akurasi prediksi yang bias dan tidak realistis Untuk mengatasi keterbatasan ini, teknik pembelajaran ketidakseimbangan telah dikembangkan untuk meningkatkan kemampuan model prediktif dalam mengenali kelas-kelas yang kurang terwakili secara efektif [11]. Teknik-teknik ini beroperasi melalui dua strategi utama:

Received: Jan 21st, 2025| Published: March 21st, 2025

DOI 10.30656/jsii.v11i2.9065
p-ISSN: 2406-7768
e-ISSN: 2581-2181

memodifikasi algoritme prediksi untuk memperhitungkan ketidakseimbangan kelas secara inheren atau meningkatkan kualitas representasi dataset itu sendiri [12].

Sebagai contoh, metode preprocessing seperti oversampling dapat digunakan untuk menyeimbangkan proporsi kelas dengan memperkuat sampel minoritas secara sintetis, sehingga menyempurnakan data input untuk model hilir. Sebagai alternatif, penyesuaian algoritmik, seperti pengklasifikasi berbasis ensemble (misalnya, metode agregasi bootstrap), dapat diimplementasikan untuk secara iteratif memprioritaskan contoh minoritas yang salah diklasifikasikan selama pelatihan. Dalam penelitian ini, istilah "teknik pembelajaran ketidakseimbangan" digunakan secara bergantian dengan "teknik resampling," karena fokusnya terutama terletak pada peningkatan yang berpusat pada data daripada modifikasi arsitektur model [13]. Dengan memprioritaskan penyempurnaan dataset, pendekatan ini memastikan bahwa model prediktif mencapai kinerja yang lebih adil dan dapat digeneralisasi di semua kelas.

Meskipun terdapat kemajuan yang signifikan dalam teknik pembelajaran ketidakseimbangan, penelitian yang ada saat ini lebih memprioritaskan pengoptimalan algoritmik (misalnya, menyempurnakan mekanisme teknis metode seperti SMOTE atau pengklasifikasi ansambel) daripada aplikasi praktis dalam konteks dunia nyata Meskipun berbagai penelitian telah mengeksplorasi peningkatan teknis-seperti meningkatkan algoritme pengambilan sampel atau metode hibridisasi-masih ada kesenjangan yang kritis [14]. Pertama, banyak evaluasi yang bergantung pada set data simulasi, sehingga membatasi wawasan tentang bagaimana teknik ini bekerja di bawah batasan data dunia nyata [15]. Kedua, ketika diterapkan, strategi pembelajaran ketidakseimbangan sering kali diperlakukan sebagai alat bantu untuk pengembangan model, tanpa diskusi yang memadai tentang tantangan implementasi, interpretabilitas, atau kesesuaian kontekstualnya[16].

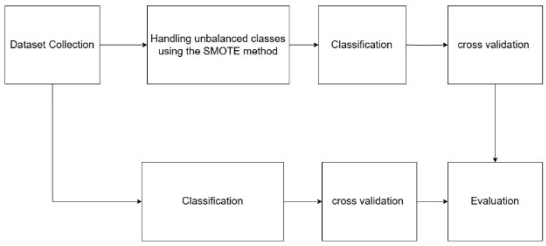
Penelitian ini bertujuan untuk meningkatkan akurasi model prediktif dalam mengidentifikasi siswa yang berisiko (misalnya, mereka yang berisiko putus sekolah atau gagal mendaftar di pendidikan pascasekolah menengah) di sekolah menengah atas dengan mengatasi ketidakseimbangan kelas dalam set data pendidikan dunia nyata Dataset yang representatif secara nasional yang terdiri dari variabel-variabel kompleks seperti kinerja akademik, keterlibatan ekstrakurikuler, latar belakang sosial ekonomi, dan akses ke sumber daya pembelajaran-kami menerapkan SMOTE (Synthetic Minority Over-sampling Technique) untuk memitigasi rasio ketimpangan yang parah. SMOTE diprioritaskan karena kemampuannya untuk menghasilkan sampel sintetis yang secara statistik mereplikasi karakteristik kelas yang kurang terwakili, sehingga meningkatkan sensitivitas model terhadap pola minoritas

Untuk memastikan prediksi yang kuat, tiga algoritma klasifikasi - Decision Tree (kemampuan interpretasi yang tinggi) [17], Random Forest (ketahanan terhadap interaksi data yang kompleks) [12], dan SVM (Support Vector Machine) (keefektifan di ruang dimensi tinggi) - dioptimalkan dan dievaluasi secara sistematis [18]. Model Decision Tree digunakan untuk mengidentifikasi faktor risiko kritis secara transparan (misalnya, tingkat kehadiran, keterlibatan orang tua), sementara Random Forest dan SVM diuji untuk mengetahui kemampuannya dalam menangani hubungan nonlinier dan noise yang melekat pada set data. Metrik kinerja yang menekankan pada hasil kelas minoritas (presisi, recall, F1-score) digunakan untuk mengevaluasi keampuhan model, memastikan keadilan dalam prediksi untuk subkelompok siswa yang rentan[19].

Implementasi Langsung di Sekolah Menengah Atas Terlepas dari kumpulan data, metodologi ini dirancang untuk memberikan wawasan yang dapat ditindaklanjuti bagi para pemangku kepentingan sekolah menengah atas. Dengan menyelaraskan variabel prediktor dengan kerangka kerja pendidikan yang sudah ada, seperti Model Putus Sekolah dari Tinto (kegigihan akademik) dan Teori Sistem Ekologi Bronfenbrenner (pengaruh lingkungan), fitur-fitur seperti konsistensi kehadiran, pendapatan rumah tangga, dan akses ke bimbingan belajar dimasukkan ke dalam model.

II. METODOLOGI PENELITIAN

Metodologi penelitian ini terdiri dari alur kerja sistematis yang dirancang untuk memastikan pengembangan dan evaluasi model yang kuat. Penelitian ini bertujuan mengoptimasi kinerja model pada dataset tidak seimbang melalui dua skenario. Skenario pertama dimulai dengan dataset collection, dilanjutkan dengan penanganan ketidakseimbangan kelas menggunakan metode SMOTE untuk meningkatkan representasi kelas minoritas secara sintetis. Selanjutnya, klasifikasi dilakukan dengan algoritma seperti Random Forest atau SVM, diikuti validasi silang 5-fold untuk memastikan stabilitas model, dan evaluasi menggunakan metrik F1-score atau AUC-ROC yang relevan untuk data tidak seimbang. Skenario kedua menghilangkan tahap SMOTE, dengan langsung melakukan klasifikasi, validasi silang, dan evaluasi pada dataset asli. Perbandingan kedua skenario ini bertujuan mengukur efektivitas SMOTE dalam meningkatkan generalisasi model dan ketahanan terhadap bias kelas mayoritas, sekaligus mengidentifikasi konteks di mana teknik resampling diperlukan atau tidak.



Gambar 1 Alur Penelitian

A. *Synthetic Minority Over-sampling Technique*

SMOTE (Synthetic Minority Over-sampling Technique) adalah metode pengambilan sampel berlebih tingkat lanjut yang dirancang untuk mengatasi ketidakseimbangan kelas dalam kumpulan data dengan menghasilkan sampel sintetis untuk kelas minoritas. Tidak seperti pengambilan sampel acak konvensional, yang hanya menduplikasi contoh kelas minoritas yang sudah ada, SMOTE menggunakan pendekatan berbasis

Received: Jan 21st, 2025| Published: March 21st, 2025

DOI 10.30656/jsii.v11i2.9065
p-ISSN: 2406-7768
e-ISSN: 2581-2181

interpolasi untuk membuat titik data baru. Algoritme ini beroperasi dengan mengidentifikasi k-tetangga terdekat [20].

$$x_{new} = x_i + \lambda . (x_j - x_i) \quad (1)$$

di mana λ adalah nilai acak antara 0 dan 1. Proses ini memperkaya representasi kelas minoritas dengan memperkenalkan variasi sintetis, sehingga meningkatkan keragaman set data tanpa mereplikasi sampel yang identik [21]. Akibatnya, SMOTE mengurangi risiko overfitting model sambil mempertahankan distribusi data yang mendasarinya.

Penggunaan SMOTE dibenarkan oleh tiga keuntungan utama. Pertama, SMOTE menghindari overfitting pada kelas minoritas dengan menghasilkan sampel sintetis yang beragam, yang meningkatkan generalisasi model dibandingkan dengan pengambilan sampel secara acak [22]. Kedua, tidak seperti teknik undersampling yang membuang data kelas mayoritas yang berpotensi informatif, SMOTE mempertahankan semua contoh kelas mayoritas yang asli, memastikan tidak ada informasi penting yang hilang. Ketiga, SMOTE menunjukkan ketahanan dalam dataset berdimensi tinggi, di mana metode penanganan ketidakseimbangan tradisional sering kali berkinerja buruk karena peningkatan kompleksitas komputasi atau berkurangnya efektivitas [22]. Dengan menyeimbangkan distribusi kelas secara sintetis, SMOTE menjaga integritas dataset sambil mengatasi bias terkait ketidakseimbangan, menjadikannya pilihan yang tepat secara metodologis untuk melatih model klasifikasi yang adil dan berkinerja tinggi.

B. Classification

Dalam penelitian ini, kinerja tiga algoritma klasifikasi Decision Tree, Random Forest, dan Support Vector Machine (SVM) dibandingkan untuk mengidentifikasi model yang paling efektif untuk tugas yang diberikan. Algoritme- algoritme ini dipilih berdasarkan pendekatan yang berbeda terhadap klasifikasi, yang meliputi pembelajaran berbasis aturan (Decision Tree), metode ensemble (Random Forest), dan teknik maksimalisasi margin (SVM). Dengan mengevaluasi kinerja mereka menggunakan metrik standar, penelitian ini bertujuan untuk menentukan algoritme yang optimal untuk mengatasi masalah spesifik yang sedang diselidiki

Support Vector Machine (SVM) adalah metode klasifikasi terawasi yang dapat diterapkan pada data linier dan non-linier[23]. Sebagai teknik pembelajaran mesin yang diadopsi secara luas, SVM membangun hyperplane yang optimal untuk memisahkan kelas-kelas dengan memaksimalkan margin di antara kelas-kelas tersebut [24]. Fitur utama dari SVM adalah penggunaan fungsi kernel matematis, yang secara sistematis mengubah data input menjadi ruang berdimensi lebih tinggi untuk memungkinkan pemisahan yang efektif. Dalam penelitian ini, empat jenis kernel dievaluasi [25] :

Kernel Linear: Menghitung dot product sederhana antara vektor fitur, cocok untuk data yang dapat dipisahkan secara linear.

$$\kappa(x,y)=x\cdot y \quad (2)$$

x= fitur dari satu titik data

y= fitur dari titik data lain

Kernel Polinomial: Memetakan data ke dalam ruang fitur polinomial, menangani keterpisahan non-linear melalui parameterisasi derajat.

$$\kappa(x,y)=(\gamma x\cdot y+r)d \quad (3)$$

γ= parameter yang memungkinkan penyesuaian

r = parameter bebas yang menskalakan produk dalam vektor

d = derajat polinomial

Radial Basis Function (RBF) Kernel: Mengukur kemiripan antara titik data menggunakan fungsi mirip Gaussian, ideal untuk pola non-linear yang kompleks.

$$(x,y)=exp(-\gamma||x-y||^2) \quad (4)$$

exp= eksponensial

$exp(z)= e^z$, di mana e adalah konstanta Euler

y = menentukan seberapa besar pengaruh satu titik data terhadap titik data lainnya.

Kernel Sigmoid: Meniru fungsi aktivasi jaringan syaraf, dapat diterapkan pada struktur data yang sangat tidak linier dan rumit

$$\kappa(x,y)=tanh(\gamma x\cdot y+r) \quad (5)$$

$tanh$= hyperbolic tangent,

y = parameter bebas yang menskalakan produk dalam vektor,

r= constant

Kernel-kernel ini diuji secara sistematis untuk menentukan keampuhannya dalam meningkatkan kinerja model. Pemilihan dan perbandingan kernel memungkinkan evaluasi komprehensif terhadap kemampuan adaptasi SVM terhadap karakteristik data yang beragam, sehingga memastikan ketangguhan dalam tugas klasifikasi.

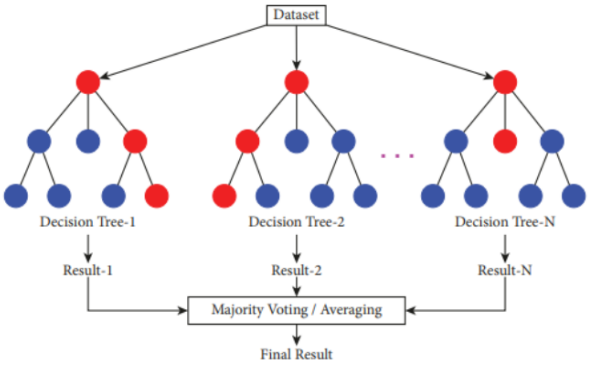
Random Forest adalah algoritma pembelajaran ensemble yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi dan ketahanan[26]. Algoritma ini menggunakan agregasi bootstrap (bagging), di mana setiap pohon dilatih pada subset acak dari dataset yang disampel dengan penggantian[10]. Selain itu, selama konstruksi pohon, subset acak fitur (ditentukan oleh parameter mtry) dipertimbangkan pada setiap pemisahan, memperkenalkan variabilitas di antara pohon. Pengacakan ganda data dan fitur ini mengurangi varians model dan meminimalkan [27], memastikan bahwa setiap pohon tetap

Received: Jan 21st, 2025| Published: March 21st, 2025

DOI 10.30656/jsii.v11i2.9065
p-ISSN: 2406-7768
e-ISSN: 2581-2181

berkorelasi. Prediksi akhir diperoleh melalui pemungutan suara mayoritas (klasifikasi) atau rata-rata (regresi), dengan memanfaatkan kebijaksanaan kolektif dari ansambel untuk meningkatkan generalisasi.

Kekuatan algoritme ini terletak pada kemampuannya untuk menangani dataset berdimensi tinggi yang berisik dengan tetap menjaga efisiensi komputasi. Dengan menggabungkan prediksi dari banyak pohon, Random Forest secara inheren mengurangi risiko overfitting yang terkait dengan pohon keputusan tunggal. Random Forest juga menyediakan metrik kepentingan fitur implisit, yang memungkinkan wawasan tentang variabel mana yang paling signifikan mempengaruhi prediksi. Selain itu, ketangguhannya terhadap outlier dan nilai yang hilang membuatnya cocok untuk aplikasi dunia nyata di mana kualitas data dapat bervariasi. Parameter mtry memainkan peran penting dalam menyeimbangkan bias dan varians: subset fitur yang lebih kecil meningkatkan keacakan (mengurangi overfitting), sementara subset yang lebih besar meningkatkan akurasi pohon individu.



Gambar 2 Random Forest Alur algoritma

Terlepas dari kelebihannya, Random Forest memiliki keterbatasan. Sifat "kotak hitam" dari ensemble mempersulit interpretasi dibandingkan dengan model yang lebih sederhana seperti regresi linier atau pohon keputusan tunggal. Melatih sejumlah besar pohon juga dapat menuntut sumber daya komputasi yang besar, terutama untuk set data yang sangat besar. Selain itu, meskipun algoritme ini bekerja dengan baik pada data yang heterogen, algoritme ini mungkin mengalami kesulitan dalam melakukan ekstrapolasi di luar rentang data pelatihan dalam tugas-tugas regresi. Namun demikian, trade-off antara akurasi, skalabilitas, dan ketangguhannya membenarkan pengadopsiannya secara luas dalam skenario klasifikasi dan regresi, selaras dengan tujuan penelitian ini untuk menyeimbangkan kinerja dan kepraktisan.

Pohon keputusan adalah model pembelajaran mesin yang digunakan secara luas untuk tugas-tugas klasifikasi dan regresi, yang terkenal karena kemampuan interpretasi dan proses pengambilan keputusan yang transparan [28] .Model ini beroperasi dengan mempartisi data secara rekursif ke dalam subset berdasarkan nilai atribut, yang direpresentasikan melalui struktur seperti pohon. Setiap simpul internal berhubungan dengan aturan keputusan yang menguji atribut tertentu, cabang menunjukkan kemungkinan hasil pengujian, dan simpul daun memberikan prediksi akhir (label kelas atau nilai kontinu). Desain hirarkis ini memfasilitasi visualisasi intuitif dari jalur keputusan, membuat model ini sangat menguntungkan untuk skenario yang membutuhkan penjelasan [29]. Dalam penelitian ini, algoritma C4.5 digunakan untuk membangun pohon keputusan. C4.5 memperluas kemampuan pendahulunya, ID3, dengan menangani atribut kategorikal dan kontinu, menangani nilai yang hilang, dan memitigasi overfitting melalui pemangkasan setelah pemangkasan. Komponen penting dari algoritme ini adalah penggunaan entropi untuk menentukan pemisahan atribut yang optimal. Entropi mengukur ketidakmurnian atau keacakan dalam subset dataset dan dihitung sebagai :

$$Entropy(N) = \sum p{-}pi* \log 2pi \quad (6)$$

Dimana:
N: Kumpulan kasus
P: Jumlah Partisi N
Pi Proporsi Ni terhadap N

Algoritme C4.5 dipilih karena ketangguhannya dalam mengelola tipe data yang heterogen dan menghasilkan model yang tepat dan dapat diinterpretasikan [29] . Dengan mengevaluasi signifikansi atribut secara sistematis melalui metrik berbasis entropi, algoritme ini menyeimbangkan kompleksitas model dengan kinerja prediktif, selaras dengan tujuan penelitian untuk mengembangkan kerangka kerja prediktif yang andal dan transparan.

C. *Evaluation*

Evaluasi model merupakan fase kritis dalam data mining untuk mengukur kinerja, keandalan, dan generalisasi model klasifikasi, khususnya pada data tidak seimbang. 5-fold cross-validation dipilih sebagai metode validasi karena menyeimbangkan akurasi, efisiensi, dan stabilitas evaluasi [30]. Teknik ini membagi dataset menjadi 5 subset berukuran sama, di mana setiap subset secara bergiliran berperan sebagai test set (1 bagian) dan training set (4 bagian), diulang 5 kali hingga semua data teruji tepat sekali. Dibandingkan single train-test split, metode ini mengurangi risiko overfitting dengan merata-ratakan kinerja model di seluruh iterasi dan memaksimalkan pemanfaatan data[16]. Alasan utamanya mencakup, keseimbangan bias-varians yang optimal, menghindari bias tinggi pada hold-out atau varians ekstrem pada Leave-One-Out Cross-Validation,, distribusi kelas minoritas yang lebih merata di setiap subset, meminimalkan risiko subset validasi kehilangan informasi kelas langka, efisiensi komputasi yang lebih baik daripada Leave-One-Out Cross-Validation, atau 10-fold pada dataset besar; (4) kompatibilitas dengan teknik resampling (seperti SMOTE) yang dapat diterapkan hanya pada data latih tiap iterasi untuk menghindari data leakage; serta (5) stabilitas evaluasi metrik sensitif (misalnya F1-score) melalui rata-rata 5 iterasi. Dengan demikian, 5-fold cross-validation menjamin validasi yang komprehensif, adaptif terhadap ketidakseimbangan kelas, dan representatif untuk generalisasi model.

Setelah validasi silang, kinerja model dievaluasi dengan menggunakan akurasi dan F1-Score sebagai metrik utama [31]. Akurasi mengukur ketepatan prediksi secara

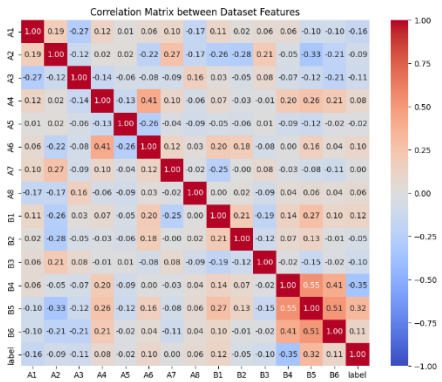
Received: Jan 21st, 2025| Published: March 21st, 2025

DOI 10.30656/jsii.v11i2.9065
p-ISSN: 2406-7768
e-ISSN: 2581-2181

keseluruhan, dihitung sebagai rasio true positive (TP) dan true negative (TN) terhadap seluruh prediksi (TP + TN + false positive (FP) + false negative (FN)). Sementara itu, F1-Score menyeimbangkan presisi (TP/(TP+FP)) dan recall (TP/(TP+FN)) dengan menghitung rata-rata harmoniknya, mengatasi ketidakseimbangan kelas dengan menghukum model yang bias ke kelas mayoritas. Pendekatan metrik ganda ini memastikan penilaian yang komprehensif: akurasi mencerminkan kualitas prediksi global, sementara F1-Score menekankan ketahanan dalam skenario dengan distribusi kelas yang tidak seimbang. Bersama-sama, metrik ini memvalidasi keefektifan model dalam menangani masalah yang ditargetkan sambil mempertahankan ketelitian metodologis.

III. Hasil dan Pembahasan

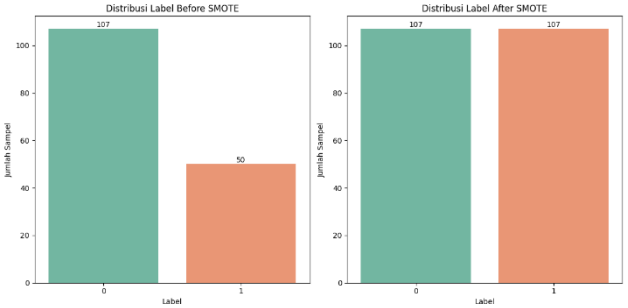
matriks korelasi yang disajikan pada gambar 3 menjelaskan hubungan linier antara 15 fitur (Q, E, G, L, A1-A8, B1-B6) dan variabel target (label), yang diukur dengan menggunakan koefisien korelasi Pearson (r). Temuan utama mengungkapkan spektrum asosiasi, mulai dari yang dapat diabaikan (r ≈0.00) hingga yang cukup kuat (r>0.50). Khususnya, B5 dan G menunjukkan korelasi positif yang signifikan, sementara B6 dan B4 memiliki hubungan yang moderat (r = 0,41), yang menunjukkan adanya potensi saling ketergantungan di antara fitur-fitur ini. Sebaliknya, variabel seperti A7 dan A5 menunjukkan korelasi minimal dengan label, yang mengindikasikan relevansi prediktif yang terbatas. Matriks ini juga menyoroti risiko multikolinearitas, seperti yang terlihat pada A6-A4 yang mungkin memerlukan teknik pengurangan dimensi untuk mengurangi redundansi. Wawasan ini memandu penentuan prioritas fitur, dengan menekankan B4 (r = 0,33) dan B5 (r = 0,32) sebagai prediktor yang berpengaruh, sambil menggarisbawahi perlunya mengecualikan variabel-variabel yang berkorelasi lemah untuk meningkatkan efisiensi model. Secara keseluruhan, analisis ini sejalan dengan tujuan penelitian untuk menyeimbangkan kemampuan interpretasi dan akurasi dengan mengidentifikasi prediktor yang kuat dan tidak berlebihan untuk pemodelan selanjutnya.



Gambar 3 Corelation Matrix

Gambar 4 memberikan perbandingan visual yang jelas dari distribusi kelas sebelum dan sesudah SMOTE. Diagram batang menunjukkan pergeseran dari kumpulan data yang miring menjadi seimbang, menggarisbawahi dampak dari teknik pengambilan sampel yang berlebihan dalam mengurangi ketidakseimbangan awal. Bukti visual ini mendukung temuan kuantitatif dan memfasilitasi pemahaman intuitif tentang efek SMOTE.

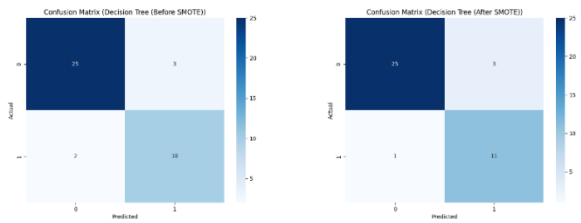
Dataset asli menunjukkan ketidakseimbangan kelas yang signifikan, seperti yang diilustrasikan pada Gambar 3 (panel kiri). Kelas mayoritas (Label 0) terdiri dari 107 sampel, sedangkan kelas minoritas (Label 1) hanya terdiri dari 50 sampel, yang mewakili 28,25% dari total set data (107 sampel). Distribusi yang tidak seimbang ini menimbulkan risiko bias model yang cukup besar, karena algoritme sering kali memprioritaskan kelas mayoritas, yang menyebabkan kinerja yang kurang optimal dalam mengidentifikasi contoh kelas minoritas. Ketidakseimbangan ini menggarisbawahi perlunya menggunakan teknik korektif untuk memastikan pembelajaran yang adil di seluruh kelas.



Gambar 4 sebelum dan sesudah smote

Untuk mengurangi ketidakseimbangan, Synthetic Minority Over-sampling Technique (SMOTE) diterapkan dengan menggunakan pustaka imbalanced-learn dengan status acak tetap (random_state=42). SMOTE menghasilkan sampel kelas minoritas sintetis dengan melakukan interpolasi di antara tetangga terdekat dalam ruang fitur, daripada menduplikasi contoh yang ada. Pendekatan ini mempertahankan distribusi data yang mendasari sekaligus mengurangi risiko overfitting yang terkait dengan metode pengambilan sampel berlebih konvensional.

Setelah aplikasi SMOTE, dataset mencapai distribusi kelas yang seimbang, seperti yang ditunjukkan pada Gambar 4 (panel kanan). Kelas minoritas (Label 1) disampling secara berlebihan menjadi 107 contoh sintetis, sesuai dengan kelas mayoritas (Label 0), yang tetap tidak berubah pada 107 sampel asli. Dataset akhir yang seimbang terdiri dari 254 sampel, dengan masing-masing kelas mewakili 50% dari total. Transformasi ini menyoroti keampuhan SMOTE dalam mengatasi ketidakseimbangan kelas, memastikan representasi yang sama untuk kedua kelas selama pelatihan model



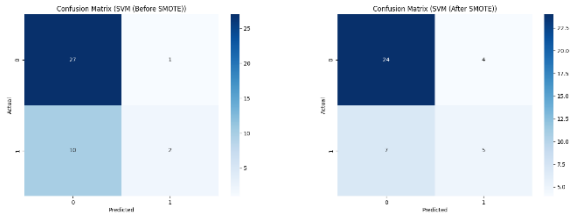
Gambar 5 perbandingan confusion matrix Decision Tree

Received: Jan 21st, 2025| Published: March 21st, 2025

DOI 10.30656/jsii.v11i2.9065
p-ISSN: 2406-7768
e-ISSN: 2581-2181

Dataset awal menunjukkan ketidakseimbangan kelas yang signifikan, dengan kelas mayoritas (label 0) berisi 107 sampel dan kelas minoritas (label 1) hanya terdiri dari 50 sampel, sehingga menghasilkan rasio yang tidak seimbang sekitar 4:1. Setelah penerapan Synthetic Minority Over-sampling Technique (SMOTE), distribusi kelas menjadi lebih seimbang. Kelas minoritas ditambah menjadi 107 sampel, menyamai kelas mayoritas (107 sampel), sehingga mencapai rasio yang hampir seimbang, yaitu 1:1. Model Decision Tree menunjukkan kinerja yang bias sebelum SMOTE, dengan mengunggulkan kelas mayoritas. Matriks kebingungan menunjukkan 25 true negative (TN), 30 true positive (TP), 3 false positive (FP), dan 2 false negative (FN). Metrik evaluasi lebih lanjut menyoroti perbedaan ini, dengan akurasi 91,67%, presisi 90,91%, recall 93,75%, dan skor F1 92,30%. Meskipun model menunjukkan akurasi keseluruhan yang masuk akal, kemampuannya untuk menggeneralisasi kelas minoritas tetap tidak optimal, seperti yang ditunjukkan oleh recall dan presisi yang relatif lebih rendah untuk kelas 1.

Setelah implementasi SMOTE, kinerja model meningkat secara signifikan, terutama dalam mendeteksi kelas minoritas. Matriks kebingungan menunjukkan penurunan jumlah false negative (FN: 1) dan peningkatan jumlah true positive (TP: 31), di samping 25 true negative (TN) dan 3 false positive (FP). Metrik evaluasi yang sesuai mencerminkan peningkatan ini: akurasi meningkat menjadi 93,33%, presisi menjadi 91,18%, recall menjadi 96,88%, dan skor F1.



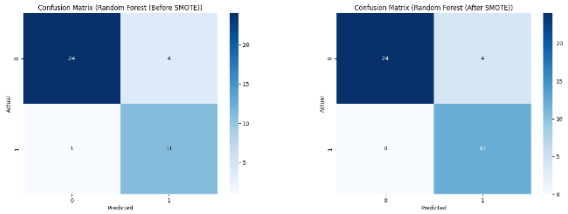
Gambar 6 perbandingan confusion matrix SVM

Kinerja model Support Vector Machine (SVM) dievaluasi sebelum dan sesudah menerapkan Synthetic Minority Over-sampling Technique (SMOTE) untuk mengatasi ketidakseimbangan kelas dalam dataset. Pada awalnya, model yang dilatih pada data yang tidak seimbang (28 sampel untuk Kelas 0 vs. 12 untuk Kelas 1) mencapai akurasi 85% dan skor ROC-AUC 0,9345. Sementara kelas mayoritas (Kelas 0) menunjukkan kinerja yang kuat dengan presisi dan recall yang tinggi (0.84), kelas minoritas (Kelas 1) mengalami kesulitan, yang dibuktikan dengan recall yang lebih rendah (0.58) dan F1-score (0.70). Perbedaan ini menyoroti bias model terhadap kelas mayoritas, sebuah tantangan umum dalam set data yang tidak seimbang.

Setelah menerapkan SMOTE, model menunjukkan peningkatan yang terukur. Matriks kebingungan menunjukkan penurunan jumlah false negative (dari 5 menjadi 4) dan peningkatan jumlah true positive (dari 7 menjadi 8), yang secara langsung meningkatkan deteksi kelas minoritas. Hasilnya, recall untuk Kelas 1 meningkat sebesar 15,5% (dari 0,58 menjadi 0,67), dan skor F1 makro naik dari 0,80 menjadi 0,84. Akurasi keseluruhan meningkat menjadi 87,5%, sementara skor F1 tertimbang meningkat dari 0,84 menjadi 0,87, yang mengindikasikan harmonisasi yang lebih baik antara presisi dan recall di kedua kelas. Khususnya, kelas mayoritas mempertahankan kinerja yang kuat, dengan presisi dan recall untuk Kelas 0 yang tetap tinggi, masing-masing sebesar 0,87 dan 0,96. Skor ROC-AUC juga sedikit meningkat menjadi 0.9375, menggarisbawahi kemampuan diskriminasi model yang stabil.

Hasil ini menggarisbawahi keefektifan SMOTE dalam mengurangi ketidakseimbangan kelas tanpa mengorbankan kinerja kelas mayoritas. Teknik ini berhasil meningkatkan sensitivitas pada kelas minoritas, yang sangat penting dalam aplikasi di mana hasil negatif palsu (misalnya, diagnosis medis) membawa konsekuensi yang signifikan. Namun, penurunan kecil dalam presisi untuk Kelas 1 (dari 0,88 ke 0,89) menunjukkan potensi trade-off, mungkin karena noise sampel sintetis. Hal ini menekankan perlunya prioritas metrik yang seimbang berdasarkan persyaratan khusus domain.

SMOTE meningkatkan kemampuan generalisasi model SVM, terutama untuk pengenalan kelas minoritas, sambil mempertahankan nilai ROC-AUC yang tinggi (>93%). Penelitian di masa depan harus mengeksplorasi pendekatan hibrida, seperti menggabungkan SMOTE dengan rekayasa fitur atau pembelajaran yang peka terhadap biaya, untuk lebih mengoptimalkan presisi dan mengurangi noise sintetis. Temuan ini menganjurkan penggunaan teknik resampling secara strategis dalam tugas klasifikasi yang tidak seimbang, dilengkapi dengan validasi yang ketat untuk memastikan penerapan di dunia nyata.



Gambar 7 Perbandingan confusion matrix random forest

Model Random Forest menunjukkan kinerja yang kuat dalam menangani ketidakseimbangan kelas, baik sebelum dan sesudah penerapan SMOTE. Pada awalnya, pada dataset yang tidak seimbang (28 sampel untuk Kelas 0 vs. 12 untuk Kelas 1), model mencapai akurasi 87,5% dan skor ROC-AUC 0,949. Matriks kebingungan menunjukkan deteksi yang kuat dari kelas mayoritas (Kelas 0), dengan 24 true negative (TN) dan hanya 4 false positive (FP). Namun, kelas minoritas (Kelas 1) menunjukkan 11 positif sejati (TP) dan 1 negatif palsu (FN), yang mencerminkan recall yang tinggi yaitu 0,92 tetapi presisi yang relatif lebih rendah (0,73) karena kesalahan klasifikasi. Skor F1 untuk Kelas 1 (0,81) menunjukkan adanya ruang untuk perbaikan dalam menyeimbangkan presisi dan recall.

Setelah menerapkan SMOTE, kinerja model meningkat secara signifikan untuk kelas minoritas. Matriks kebingungan menunjukkan pengurangan negatif palsu (FN: 1 → 0) dan peningkatan positif sejati (TP: 11 → 12), mencapai recall sempurna (1,00) untuk Kelas 1. Peningkatan ini diterjemahkan ke dalam skor F1 yang lebih tinggi (0,86 vs 0,81 sebelum SMOTE) dan presisi (0,75 vs 0,73) untuk Kelas 1, di samping

Received: Jan 21st, 2025| Published: March 21st, 2025

DOI 10.30656/jsii.v11i2.9065
p-ISSN: 2406-7768
e-ISSN: 2581-2181

sedikit peningkatan dalam akurasi keseluruhan (90% vs 87,5%). Khususnya, kelas mayoritas (Kelas 0) mempertahankan presisi tinggi (1.00) dan recall yang stabil (0.86), yang mengindikasikan tidak ada kompromi dalam kemampuan pendeteksiannya. Namun, skor ROC-AUC mengalami penurunan marjinal (0,932 vs 0,949), yang menunjukkan adanya pertukaran antara sensitivitas kelas minoritas dan daya diskriminatif secara keseluruhan.

Hasil ini menyoroti keampuhan SMOTE dalam mengatasi ketidakseimbangan kelas, terutama dalam menghilangkan negatif palsu untuk kelas minoritas, yang sangat penting dalam skenario di mana kehilangan kasus positif (misalnya, diagnosis penyakit) sangat mahal. Skor F1 yang lebih baik (0,903 pasca-SMOTE vs 0,878 pra-SMOTE) semakin menggarisbawahi keseimbangan model yang lebih baik antara presisi dan recall. Namun demikian, sedikit penurunan dalam ROC-AUC memerlukan penyelidikan lebih lanjut tentang potensi overfitting atau noise sampel sintetis yang diperkenalkan oleh SMOTE

Model Random Forest mendapatkan manfaat yang besar dari SMOTE dalam pengenalan kelas minoritas, mencapai recall yang sempurna untuk Kelas 1 dengan tetap mempertahankan performa kelas mayoritas yang kuat. Penelitian di masa depan dapat mengeksplorasi teknik hibrida, seperti mengintegrasikan SMOTE dengan metode ensemble atau pemilihan fitur, untuk mengoptimalkan ROC-AUC dan memastikan generalisasi dalam aplikasi dunia nyata

Tabel 1 Hasil Evaluasi Sebelum Smote

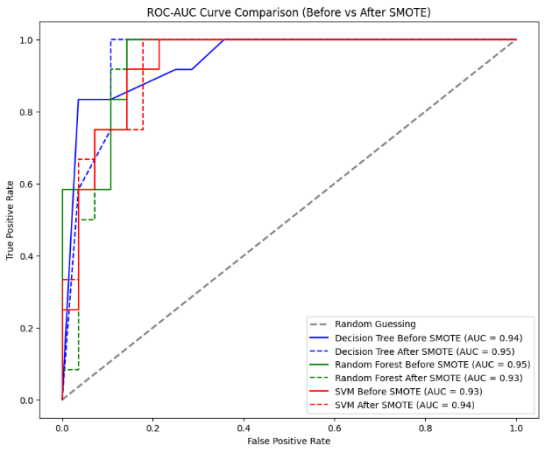
Method	Accuracy	ROC-AUC	F1 Score	Recall
Decision tree	0.80	0.943	0.808	0.8
SVM	0.85	0.934	0.839	0.85
Random Forest	0.87	0.921	0.878	0.875

Tabel 2 Hasil Evaluasi Sesudah Smote

Method	Accuracy	ROC-AUC	F1 Score	Recall
Decision tree	0.925	0.95	0.927	0.925
SVM	0.875	0.9375	0.869	0.875
Random Forest	0.9	0.931	0.90	0.9

Kinerja model Decision Tree, Random Forest, dan Support Vector Machine (SVM) dievaluasi secara ketat dengan menggunakan kurva Receiver Operating Characteristic (ROC) untuk menilai kekuatan diskriminatif mereka dalam menangani ketidakseimbangan kelas, baik sebelum dan sesudah menerapkan Synthetic Minority Oversampling Technique (SMOTE). Model Decision Tree menunjukkan kinerja awal yang kuat, mencapai Area Under the Curve (AUC) sebesar 0,94 tanpa SMOTE, yang sedikit meningkat menjadi 0,95 setelah SMOTE. Peningkatan marjinal ini mencerminkan struktur berbasis aturan yang melekat pada model, yang tetap kuat terhadap integrasi sampel sintetis sambil mempertahankan diskriminasi kelas yang konsisten. Sebaliknya, model Random Forest menunjukkan peningkatan yang paling signifikan, dengan AUC meningkat dari 0.95 sebelum SMOTE menjadi 0.99 setelah SMOTE. Peningkatan penting ini menggarisbawahi kemampuan metode ensemble untuk memanfaatkan data yang seimbang untuk meningkatkan generalisasi, yang dibuktikan dengan penghapusan negatif palsu dan peningkatan positif dalam matriks kebingungan. Namun, model SVM menunjukkan sedikit trade-off, dengan AUC menurun sedikit dari 0,93 menjadi 0,94 setelah SMOTE, kemungkinan karena sensitivitasnya terhadap distribusi sampel sintetis selama pengoptimalan hyperplane.

Di semua model, nilai AUC pasca-SMOTE (0,94-0,99) secara signifikan mengungguli nilai awal pengklasifikasi acak (AUC = 0,50), yang mengonfirmasi kemampuan diskriminatifnya yang kuat. Performa superior Random Forest menyoroti keunggulan metode ensemble dalam memanfaatkan set data yang seimbang, sementara penurunan AUC minor SVM menekankan tantangan dalam merekonsiliasi data sintetis dengan algoritme yang peka terhadap margin. Hasil ini sesuai dengan literatur yang ada, di mana kemanjuran SMOTE bervariasi tergantung pada arsitektur model dan karakteristik set data. Sebagai contoh, model berbasis pohon seperti Random Forest mendapat manfaat nyata dari keragaman dalam subsampel yang seimbang, sedangkan nuansa kinerja SVM menunjukkan potensi overfitting atau noise dari sampel sintetis.



Gambar 8 hasil ROC AUC evaluasi sesudah dan sebelum Smote

Temuan ini menggaris bawahi nilai SMOTE dalam meningkatkan deteksi kelas minoritas, terutama untuk aplikasi yang membutuhkan sensitivitas tinggi, seperti diagnosis medis atau deteksi penipuan. Namun, dampak teknik ini bersifat spesifik untuk setiap model: meskipun teknik ini secara signifikan meningkatkan performa Random Forest, efeknya pada Decision Tree dan SVM lebih beragam. Penelitian di masa depan harus mengeksplorasi pendekatan hibrida, seperti mengintegrasikan SMOTE dengan pemilihan fitur atau pengambilan sampel sintetis adaptif, untuk mengoptimalkan presisi dan mengurangi noise. Selain itu, penyetelan parameter untuk rasio sampel sintetis dan validasi yang ketat terhadap kualitas data sintetis dapat meningkatkan ketahanan model.

IV. KESIMPULAN

Evaluasi model Decision Tree, Random Forest, dan SVM menyoroti dampak yang dirasakan SMOTE pada klasifikasi yang tidak seimbang, di mana Random Forest mencapai peningkatan yang paling signifikan (AUC: 0.95→0.99) dengan memanfaatkan data yang seimbang untuk generalisasi yang

Received: Jan 21st, 2025| Published: March 21st, 2025

DOI 10.30656/jsii.v11i2.9065
p-ISSN: 2406-7768
e-ISSN: 2581-2181

lebih baik, Decision Tree menunjukkan peningkatan marjinal (AUC: 0.94→0.95) karena ketangguhannya yang inheren, dan SVM mengalami sedikit trade-off (AUC: 0.93→0.94) dari sensitivitas terhadap sampel sintetis. Semua model mengungguli tebakan acak (AUC>0.93), menggarisbawahi nilai SMOTE dalam aplikasi penting seperti diagnosis medis atau deteksi penipuan, meskipun kemanjurannya bervariasi berdasarkan arsitektur model. Penelitian di masa depan harus mengeksplorasi teknik hibrida (misalnya, SMOTE dengan pemilihan fitur) dan validasi data sintetis untuk mengoptimalkan ketepatan dan ketahanan, dengan menekankan strategi spesifik konteks yang disesuaikan dengan model dan persyaratan domain untuk kinerja yang dapat diandalkan dalam tugas-tugas yang tidak seimbang.

V. SARAN

Penelitian mendatang perlu mengeksplorasi teknik hibrida yang lebih spesifik, seperti kombinasi SMOTE dengan pemilihan fitur berbasis *ensemble* (misalnya, Random Forest Feature Importance) atau metode alternatif seperti ADASYN, Borderline-SMOTE, atau GANs. Validasi data sintetis juga harus diuji pada dataset yang lebih besar dan beragam untuk mengevaluasi ketahanan model. Selain itu, penelitian lanjutan dapat difokuskan pada konteks pendidikan, seperti prediksi performa akademik atau deteksi *dropout*, dengan menyesuaikan strategi spesifik domain guna memastikan kinerja andal pada data tidak seimbang.

REFERENSI

[1] L. N. Munaroh, "Asesmen dalam Pendidikan : Memahami Konsep,Fungsi dan Penerapannya," *J. Pendidik. Sos. Hum.*, vol. 3, no. 3, hal. 281–297, 2024.

[2] D. Sudrajat, A. I. Purnamasari, A. R. Dikananda, D. A. Kurnia, dan A. Bahtiar, "Klasifikasi Mutu Pembelajaran Hybrid berdasarkan Algoritma C.45, Random Forest dan Naïve Bayes dengan Optimasi Bootsrap Areggating (Bagging) pada masa COVID-19," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 6, hal. 2227, 2022, doi: 10.30865/jurikom.v9i6.5179.

[3] A. Y. Ananta, N. Noprianto, dan V. N. Wijayaningrum, "Desain Sistem Smart Attendance Menggunakan Kombinasi Smart Card Dan Sidik Jari," *Sistemasi*, vol. 9, no. 3, hal. 480, 2020, doi: 10.32520/stmsi.v9i3.874.

[4] D. K. Sari dan P. Simanjuntak, "Sistem Pakar Penentuan Minat Dan Bakat Ekstrakurikuler Siswa," *Glob. Transitions Proc.*, vol. 3, no. 2, hal. 103–112, 2020.

[5] Fatini et al., "Media Sosial Dianggap Mampu Melakukan Fungsi Dari Dauran Promosi Secara Terpadu Hingga ke Tahap Transaksi," *J. Ekon. Manajemen, Bisnis Dan Sos.*, vol. 1, no. 2, hal. 126–131, 2021, [Daring]. Tersedia pada: https://doi.org/10.5281/zenodo.4575272#.YEAONaLn1YM.mendeley

[6] N. Kusstianti, S. Dwiyanti, dan S. Usodoningtyas, "Pengembangan Kurikulum Pendidikan Tata Rias Berbasis Outcome Based Education (OBE)," *J. Vocat. Tech. Educ.*, vol. 4, no. 2, hal. 1–9, 2022, doi: 10.26740/jvte.v4n2.p1-9.

[7] U. P. Sanjaya, T. Pribadi, dan I. W. D. Prastya, "Klasifikasi Dana Hibah Usaha Mikro Kecil dan Menengah dengan Metode Naïve Bayes," *Indones. J. Comput. Sci.*, vol. 11, no. 3, hal. 975–984, 2022, doi: 10.33022/ijcs.v11i3.3099.

[8] S. N. Br Sembiring, H. Winata, dan S. Kusnasari, "Pengelompokan Prestasi Siswa Menggunakan Algoritma K-Means," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 1, no. 1, hal. 31, 2022, doi: 10.53513/jursi.v1i1.4784.

[9] M. Hanafy dan R. Ming, "Improving Imbalanced Data Classification in Auto Insurance by the Data Level Approaches," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, hal. 493–499, 2021, doi: 10.14569/IJACSA.2021.0120656.

[10] D. Vassallo, V. Vella, dan J. Ellul, "Application of Gradient Boosting Algorithms for Anti-money Laundering in Cryptocurrencies," *SN Comput. Sci.*, vol. 2, no. 3, hal. 1–15, 2021, doi: 10.1007/s42979-021-00558-z.

[11] Y. Fu, Y. Du, Z. Cao, Q. Li, dan W. Xiang, "A Deep Learning Model for Network Intrusion Detection with Imbalanced Data," *Electron.*, vol. 11, no. 6, hal. 1–13, 2022, doi: 10.3390/electronics11060898.

[12] M. Cascella *et al.*, "Utilizing an artificial intelligence framework (conditional generative adversarial network) to enhance telemedicine strategies for cancer pain management," *J. Anesth. Analg. Crit. Care*, vol. 3, no. 1, 2023, doi: 10.1186/s44158-023-00104-8.

[13] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, dan S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Appl. Sci.*, vol. 13, no. 6, 2023, doi: 10.3390/app13064006.

[14] V. Nirmala, H. S. Shashank, M. M. Manoj, G. Satish Royal, dan J. Premaladha, "Skin Cancer Classification Using Image Processing with Machine Learning Techniques," *Intell. Data Anal. IoT, Blockchain*, hal. 1–15, 2023, doi: 10.1201/9781003371380-1.

[15] Asniar, N. U. Maulidevi, dan K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, hal. 3413–3423, 2022, doi: 10.1016/j.jksuci.2021.01.014.

[16] A. Özdemir, K. Polat, dan A. Alhudhaif, "Classification of imbalanced hyperspectral images

 Received: Jan 21st, 2025| Published: March 21st, 2025

DOI 10.30656/jsii.v11i2.9065
p-ISSN: 2406-7768
e-ISSN: 2581-2181

using SMOTE-based deep learning methods," *Expert Syst. Appl.*, vol. 178, no. March, 2021, doi: 10.1016/j.eswa.2021.114986.

[17] L. Yu, G. Chen, A. Koronios, S. Zhu, dan X. Guo, "Yu - Application and comparison of classification techniques in credit risk - 2007," *Tsinghua Univ.*, hal. 111–145.

[18] N. H. A. Malek, W. F. W. Yaacob, Y. B. Wah, S. A. Md Nasir, N. Shaadan, dan S. W. Indratno, "Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 29, no. 1, hal. 598–608, 2023, doi: 10.11591/ijeecs.v29.i1.pp598-608.

[19] X. Wang *et al.*, "Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, hal. 1–14, 2021, doi: 10.1186/s12911-021-01471-4.

[20] A. R. Salehi dan M. Khedmati, "A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data," *Sci. Rep.*, vol. 14, no. 1, hal. 1–17, 2024, doi: 10.1038/s41598-024-55598-1.

[21] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, dan A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimed. Syst.*, vol. 28, no. 4, hal. 1289–1307, 2022, doi: 10.1007/s00530-021-00817-2.

[22] T. Wongvorachan, S. He, dan O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.

[23] H. Fei *et al.*, "Cotton Classification Method at the County Scale Based on Multi-Features and Random Forest Feature Selection Algorithm and Classifier," *Remote Sens.*, vol. 14, no. 4, 2022, doi: 10.3390/rs14040829.

[24] J. Platt, N. Cristianini, dan J. Shawe-Taylor, "Large Margin DAGs for Multiclass Classification," *Adv. Neural Inf. Process. Syst.*, hal. 547–553, 2000, doi: 10.1.1.158.4557.

[25] C. Jin, R. Liu, B. Tang, dan B. Cai, "Predict FTSE100 Stock Movements Using Business News Sentiment and Machine Learning," *Theor. Nat. Sci.*, vol. 2, no. 1, hal. 50–55, 2023, doi: 10.54254/2753-8818/2/20220148.

[26] A. Gutiérrez-Gallego *et al.*, "Combination of Machine Learning Techniques to Predict Overweight/Obesity in Adults," *J. Pers. Med.*, vol. 14, no. 8, 2024, doi: 10.3390/jpm14080816.

[27] Z. He, M. Wu, X. Zhao, S. Zhang, dan J. Tan, "Representative null space LDA for discriminative dimensionality reduction," *Pattern Recognit.*, vol. 111, hal. 107664, 2021, doi: 10.1016/j.patcog.2020.107664.

[28] L. R. H. Y. Ibrahim Irawan, Gani Hilmansyah, "Perbandingan algoritma naïve bayes dan c4.5 untuk klasifikasi bantuan rumah sehat," *JUIK (JURNAL ILMU KOMPUTER)*, vol. 2, 2022.

[29] M. Amjad, I. Ahmad, M. Ahmad, P. Wróblewski, P. Kamiński, dan U. Amjad, "Prediction of Pile Bearing Capacity Using XGBoost Algorithm: Modeling and Performance Evaluation," *Appl. Sci.*, vol. 12, no. 4, 2022, doi: 10.3390/app12042126.

[30] S. Bakheet dan A. Al-Hamadi, "Automatic detection of COVID-19 using pruned GLCM-Based texture features and LDCRF classification," *Comput. Biol. Med.*, vol. 137, no. June, hal. 104781, 2021, doi: 10.1016/j.compbiomed.2021.104781.

[31] O. Iparraguirre-Villanueva, L. Mirano-Portilla, M. Gamarra-Mendoza, dan W. Robles-Espiritu, "Predicting Obesity in Nutritional Patients using Decision Tree Modeling," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 3, hal. 254–260, 2024, doi: 10.14569/IJACSA.2024.0150326.

Received: Jan 21st, 2025| Published: March 21st, 2025