

Pengembangan Model Prediksi Risiko Diabetes Menggunakan Pendekatan AdaBoost dan Teknik *Oversampling* SMOTE

Sofian Sidiq^{1*}, Alfian², Nur Shobi Mabur³

^{1,2,3}Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Tangerang, Indonesia

^{1*}sofian.sidiq@gmail.com, ²alfian.alf@ft-umt.ac.id, ³shobimabur@ft-umt.ac.id

Abstrak

Kata Kunci: AdaBoost; Diabetes Mellitus; Model Prediksi; Oversampling; SMOTE

Diabetes mellitus merupakan penyakit kronis yang prevalensinya terus meningkat dan menjadi masalah kesehatan serius di berbagai negara, termasuk Indonesia. Salah satu tantangan utama dalam pengembangan model prediksi risiko diabetes adalah ketidakseimbangan data, di mana jumlah sampel kelas minoritas (penderita diabetes) jauh lebih sedikit dibandingkan kelas mayoritas (bukan penderita diabetes). Penelitian ini bertujuan untuk mengatasi masalah ketidakseimbangan data dengan mengintegrasikan metode AdaBoost dan teknik oversampling SMOTE untuk membangun model prediksi risiko diabetes yang akurat dan andal. Metode AdaBoost dipilih karena kemampuannya dalam meningkatkan akurasi prediksi melalui perbaikan kesalahan secara iteratif, sedangkan SMOTE digunakan untuk meningkatkan representasi kelas minoritas dengan menghasilkan data sintetis. Hasil penelitian menunjukkan bahwa model AdaBoost tanpa SMOTE memiliki akurasi sebesar 81,87% dan nilai ROC-AUC sebesar 0,9031, tetapi performanya cenderung lebih tinggi pada kelas mayoritas dibandingkan kelas minoritas. Setelah menerapkan AdaBoost dengan SMOTE, performa model pada kelas minoritas meningkat signifikan, dengan *precision*, *recall*, dan F1-score yang lebih seimbang antara kedua kelas. Akurasi keseluruhan model meningkat menjadi 82,83%, dan nilai ROC-AUC menjadi 0,9058. Kombinasi AdaBoost dan SMOTE terbukti efektif dalam menangani ketidakseimbangan data, memberikan prediksi yang lebih seimbang antara kelas mayoritas dan minoritas. Penelitian ini berkontribusi dalam pengembangan pendekatan prediktif berbasis *machine learning* untuk mendukung upaya preventif di bidang kesehatan.

Abstract

Keywords: AdaBoost; Diabetes Mellitus; Prediction Model; Oversampling; SMOTE

Diabetes mellitus is a chronic disease with a steadily increasing prevalence, becoming a serious health issue in many countries, including Indonesia. One of the main challenges in developing a predictive model for diabetes risk is data imbalance, where the number of samples in the minority class (diabetes cases) is significantly lower than that in the majority class (non-diabetes cases). This study aims to address the issue of data imbalance by integrating the AdaBoost method with the oversampling technique SMOTE to develop an accurate and reliable diabetes risk prediction model. The AdaBoost method was chosen for its ability to improve prediction accuracy by iteratively correcting errors, while SMOTE was utilized to enhance the representation of the minority class by generating synthetic data. The study's results show that the

AdaBoost model without SMOTE achieved an accuracy of 81.87% and a ROC-AUC score of 0.9031. However, its performance was significantly higher for the majority class compared to the minority class. After applying AdaBoost with SMOTE, the model's performance for the minority class improved significantly, with more balanced precision, recall, and F1-score across both classes. The overall model accuracy increased to 82.83%, and the ROC-AUC score rose to 0.9058. The combination of AdaBoost and SMOTE has proven effective in addressing data imbalance, providing more balanced predictions between the majority and minority classes. This research contributes to the development of machine learning-based predictive approaches to support preventive efforts in the healthcare sector.

1.PENDAHULUAN

Diabetes mellitus merupakan salah satu penyakit kronis yang prevalensinya terus meningkat secara global. Menurut data dari *World Health Organization* (WHO), jumlah penderita diabetes di seluruh dunia telah mencapai angka yang mengkhawatirkan, dengan lebih dari 422 juta orang menderita diabetes pada tahun 2021, dan angka ini diperkirakan akan terus bertambah menjadi 700 juta pada tahun 2045 [1]. Di Indonesia, data dari Survei Kesehatan Indonesia (SKI) 2023 menunjukkan prevalensi diabetes melitus pada penduduk berusia di atas 15 tahun mencapai 11,7%, meningkat dari 10,9% pada tahun 2018 [2]. Peningkatan ini menunjukkan bahwa diabetes menjadi masalah kesehatan yang semakin serius di Indonesia. Pendeteksian dini risiko diabetes memiliki peran yang sangat penting dalam mencegah komplikasi lebih lanjut, seperti penyakit kardiovaskular, gagal ginjal, atau neuropati [3]. Dengan pendeteksian dini, intervensi medis dan perubahan gaya hidup dapat dilakukan lebih awal untuk mengurangi risiko tersebut. Oleh karena itu, pengembangan model prediksi risiko diabetes yang akurat dan andal menjadi salah satu langkah penting dalam mendukung upaya preventif di bidang kesehatan.

Upaya untuk mengembangkan model prediksi risiko diabetes telah dilakukan oleh berbagai penelitian sebelumnya dengan menerapkan beragam metode. Salah satu penelitian menggunakan algoritma *K-Nearest Neighbors* (K-NN) dengan dan tanpa seleksi fitur *Information Gain* [4]. Hasil penelitian tersebut menunjukkan bahwa model K-NN tanpa *Information Gain* memiliki akurasi sebesar 69,11%, sementara model K-NN dengan *Information Gain* mampu mencapai akurasi tertinggi sebesar 72,93% pada nilai $K=17$. Selain itu, penelitian lain memanfaatkan algoritma *Naïve Bayes* untuk klasifikasi diabetes, menghasilkan akurasi sebesar 73,16% dengan nilai *Area Under the Curve* (AUC) pada kurva ROC sebesar 0,818 [5]. Penelitian lain membandingkan algoritma *Naïve Bayes* dengan ID3, yang menunjukkan bahwa *Naïve Bayes* memiliki performa lebih baik dengan akurasi 76% dan nilai AUC sebesar 0,794, dibandingkan ID3 yang menghasilkan akurasi 74% dengan nilai AUC sebesar 0,788 [6]. Selanjutnya, penelitian yang menggunakan *Logistic Regression* dengan optimasi *Grid Search* dilaporkan menghasilkan rata-rata akurasi sebesar 79% [7].

Namun, salah satu tantangan utama dalam pengembangan model prediksi adalah ketidakseimbangan data pada dataset medis [8]. Pada umumnya, dataset medis cenderung memiliki jumlah sampel yang lebih sedikit untuk kelas positif (penderita diabetes) dibandingkan dengan kelas negatif (bukan penderita diabetes). Ketidakseimbangan ini dapat menyebabkan model prediksi cenderung mengabaikan kelas minoritas, sehingga menghasilkan performa yang rendah dalam mendeteksi kasus diabetes. Untuk mengatasi masalah ini, berbagai pendekatan telah dikembangkan, salah satunya adalah penggunaan teknik *oversampling* seperti *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE mampu meningkatkan representasi kelas minoritas dengan cara menghasilkan data sintesis berdasarkan sampel yang ada, sehingga dapat membantu model prediksi untuk lebih fokus pada kelas minoritas [9]. SMOTE juga memiliki keunggulan dalam mengurangi risiko *overfitting* yang sering terjadi pada metode *oversampling* tradisional seperti duplikasi data, karena SMOTE menciptakan data baru yang bervariasi secara statistik [10]. Di sisi lain, algoritma *ensemble* seperti *Adaptive Boosting* (AdaBoost) telah terbukti efektif dalam meningkatkan akurasi prediksi dengan menggabungkan kekuatan dari beberapa model sederhana (*weak learners*) [11]. Keunggulan utama dari AdaBoost adalah kemampuannya untuk secara adaptif memberi bobot lebih besar pada data yang sulit diprediksi,

Sofian Sidiq: *Penulis Korespondensi



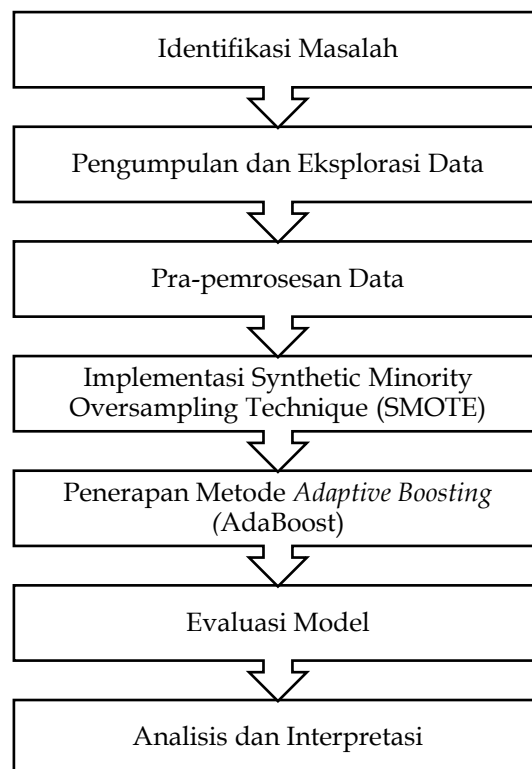
Copyright © 2025, Sofian Sidiq, Alfian, Nur Shobi Maburr.

sehingga iterasi berikutnya akan lebih fokus pada data tersebut [12]. Selain itu, AdaBoost memiliki fleksibilitas untuk digunakan dengan berbagai jenis model dasar, menjadikannya metode yang sangat serbaguna dalam berbagai aplikasi prediksi [13].

Tujuan dari penelitian ini adalah untuk merancang dan membangun model prediksi risiko diabetes dengan mengintegrasikan metode AdaBoost dan teknik *oversampling* SMOTE. Metode AdaBoost dipilih karena kemampuannya dalam meningkatkan akurasi prediksi melalui perbaikan kesalahan pada iterasi sebelumnya. Sementara itu, teknik SMOTE digunakan untuk menangani ketidakseimbangan data yang menjadi kendala utama dalam klasifikasi diabetes, sehingga membantu model menjadi lebih sensitif terhadap kelas minoritas. Kontribusi dari penelitian ini meliputi pengembangan pendekatan integratif yang mampu meningkatkan akurasi model prediksi diabetes, terutama dalam mendeteksi kasus pada kelas minoritas yang sering terabaikan. Selain itu, penelitian ini juga memberikan wawasan baru mengenai efektivitas kombinasi metode AdaBoost dan SMOTE dalam menangani masalah ketidakseimbangan data pada bidang kesehatan.

2.METODE PENELITIAN

Penelitian ini menggunakan metode eksperimen berbasis pembelajaran mesin untuk mengembangkan model prediksi risiko diabetes. Proses pengembangan dilakukan melalui integrasi teknik *oversampling* SMOTE untuk menangani ketidakseimbangan data dan algoritma *ensemble* AdaBoost untuk meningkatkan akurasi prediksi. Adapun tahapan penelitian yang digunakan berdasarkan diagram pada Gambar 1.



Gambar 1. Tahapan Penelitian

Berdasarkan Gambar 1, setiap langkah yang dilakukan dijabarkan secara detail seperti berikut.

A. Identifikasi Masalah

Pada tahap ini, penelitian diawali dengan identifikasi permasalahan utama dalam prediksi risiko diabetes. Fokus utama adalah tantangan ketidakseimbangan data pada dataset medis, di mana kelas positif (penderita diabetes) sering kali memiliki jumlah sampel yang jauh lebih sedikit dibandingkan kelas negatif (bukan penderita diabetes). Ketidakseimbangan ini dapat menyebabkan model prediksi lebih condong pada kelas mayoritas, sehingga akurasi model dalam mendeteksi kasus diabetes menjadi

Sofian Sidiq: *Penulis Korespondensi



Copyright © 2025, Sofian Sidiq, Alfian, Nur Shobi Maburr.

rendah. Selain itu, identifikasi juga mencakup analisis kebutuhan akan pengembangan model yang mampu menangani ketidakseimbangan data dan memberikan prediksi yang lebih akurat, terutama pada kelas minoritas. Masalah ini menjadi dasar pemilihan teknik SMOTE untuk mengatasi ketidakseimbangan dan algoritma AdaBoost untuk meningkatkan performa prediksi.

B. Pengumpulan dan Eksplorasi Data

Tahapan ini melibatkan pengumpulan dataset yang relevan untuk klasifikasi risiko diabetes. Dataset memiliki peran penting dalam proses analisis data dan pengembangan model *data mining* dan pembelajaran mesin, karena berfungsi sebagai sumber utama informasi untuk melatih, menguji, dan mengevaluasi kinerja model [14]. Dataset yang digunakan diambil dari platform Kaggle dengan nama "Diabetes" dan file "diabetes.csv" (<https://www.kaggle.com/datasets/johndasilva/diabetes>) [15]. Dataset ini berisi data rekam medis yang mencakup variabel penting seperti jumlah kehamilan (*Pregnancies*), kadar glukosa (*Glucose*), tekanan darah (*Blood Pressure*), ketebalan kulit (*Skin Thickness*), kadar insulin (*Insulin*), indeks massa tubuh (BMI), skor riwayat genetik diabetes (*Diabetes Pedigree Function*), usia (*Age*), dan *outcome* (*Outcome*, di mana 1 menunjukkan diabetes dan 0 menunjukkan non-diabetes). Dataset ini terdiri dari 2000 entri dengan 9 kolom, dipilih karena fitur-fitur yang tersedia sangat relevan dan memiliki kualitas yang dapat diandalkan untuk mendukung analisis dan prediksi diabetes secara komprehensif.

Setelah dataset diperoleh, eksplorasi data dilakukan untuk memahami karakteristik data. Eksplorasi data yang dilakukan diantaranya distribusi variabel, korelasi antar fitur, serta identifikasi potensi masalah seperti data yang hilang, *outlier*, atau fitur yang tidak relevan. Hasil dari eksplorasi data dibuat dalam bentuk visualisasi dan analisis data. Visualisasi data ini digunakan untuk mempermudah analisis pola dan hubungan antar fitur. Informasi yang diperoleh dari eksplorasi data menjadi landasan untuk menentukan langkah-langkah *preprocessing* data selanjutnya.

C. Pra-pemrosesan Data

Pra-pemrosesan data dilakukan untuk memastikan dataset yang digunakan memiliki kualitas yang baik dan siap untuk proses pelatihan model [16]. Langkah pertama adalah menangani data yang hilang, misalnya dengan mengganti nilai yang hilang menggunakan rata-rata, *median*, atau metode imputasi lain. Selanjutnya, data dinormalisasi agar setiap fitur berada pada skala yang sama, terutama untuk algoritma yang sensitif terhadap perbedaan skala. Untuk data kategorikal, dilakukan *encoding*, seperti *one-hot encoding*, agar data dapat diolah oleh algoritma pembelajaran mesin. Tahapan terakhir adalah pembagian dataset menjadi *training set* dan *testing set* dengan proporsi tertentu (misalnya 80:20), sehingga evaluasi model dapat dilakukan secara objektif menggunakan data yang tidak pernah dilihat model sebelumnya.

D. Implementasi *Synthetic Minority Oversampling Technique* (SMOTE)

SMOTE (*Synthetic Minority Oversampling Technique*) adalah teknik yang digunakan untuk mengatasi masalah ketidakseimbangan data dalam pembelajaran mesin [17]. Ketidakseimbangan ini terjadi ketika salah satu kelas dalam dataset memiliki jumlah sampel yang jauh lebih sedikit dibandingkan kelas lainnya. SMOTE meningkatkan representasi kelas minoritas dengan menghasilkan data sintetis, sehingga membuat dataset menjadi lebih seimbang [18].

SMOTE bekerja dengan prinsip interpolasi, di mana data sintetis dihasilkan dengan membuat sampel baru yang berada di antara dua sampel kelas minoritas yang ada [19]. Prosesnya dimulai dengan memilih sampel dari kelas minoritas, lalu mengidentifikasi beberapa tetangga terdekatnya (*k-nearest neighbors*) menggunakan metrik jarak, seperti jarak *Euclidean*. Setelah itu, salah satu tetangga terdekat dipilih secara acak, dan sampel baru dibuat menggunakan interpolasi linier antara dua titik. Data sintetis ini dihitung dengan persamaan (1).

$$x_{synthetic} = x_i + \lambda \cdot (x_j - x_i) \quad (1)$$

di mana $x_{synthetic}$ merupakan data sintetis yang dihasilkan, x_i menunjukkan data asli dari kelas minoritas yang dipilih secara acak, x_j menunjukkan tetangga terdekat dari x_i (dipilih dari *k-nearest*

neighbors), dan λ merupakan faktor pengali acak dalam interval $[0,1]$, yang menentukan posisi interpolasi antara x_i dan x_j .

E. Penerapan Metode Adaptive Boosting (AdaBoost)

AdaBoost (*Adaptive Boosting*) adalah salah satu algoritma *ensemble* yang dirancang untuk meningkatkan akurasi model prediksi. Algoritma ini menggabungkan beberapa model sederhana atau lemah (*weak learners*), seperti pohon keputusan dengan kedalaman rendah, untuk membentuk model yang lebih kuat dan akurat (*strong learner*) [11]. Keunggulan utama dari AdaBoost adalah kemampuannya untuk secara adaptif memfokuskan pelatihan pada sampel yang sulit diprediksi dengan benar pada iterasi sebelumnya [20].

Cara kerja metode AdaBoost dimulai dengan memberikan bobot awal yang sama kepada semua sampel dalam data pelatihan ($w_i = \frac{1}{N}$, di mana N adalah jumlah sampel pelatihan). Bobot ini mencerminkan tingkat kepentingan masing-masing sampel. Pada setiap iterasi, sebuah *weak learner* dilatih menggunakan dataset berbobot tersebut. Setelah pelatihan, tingkat kesalahan (ε_t) dari *weak learner* dihitung berdasarkan jumlah bobot sampel yang salah diprediksi. Tingkat kesalahan ini dirumuskan melalui persamaan (2).

$$\varepsilon_t = \frac{\sum_{i=1}^N w_i I(y_i \neq h_t(x_i))}{\sum_{i=1}^N w_i} \quad (2)$$

di mana $I(y_i \neq h_t(x_i))$ adalah indikator kesalahan, bernilai 1 jika prediksi salah dan 0 jika benar.

Model yang memiliki tingkat kesalahan rendah akan diberi bobot kekuatan (α_t) yang lebih besar, dihitung dengan persamaan (3).

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1-\varepsilon_t}{\varepsilon_t} \right) \quad (3)$$

Sebaliknya, jika tingkat kesalahan model lebih dari 50% ($\varepsilon_t > 0,5$), model tersebut diabaikan karena dianggap tidak lebih baik dari tebakan acak.

Selanjutnya, bobot sampel diperbarui untuk iterasi berikutnya. Sampel yang salah diprediksi akan mendapatkan bobot lebih besar, sehingga *weak learner* berikutnya akan lebih fokus pada data ini. Pembaruan bobot sampel digunakan persamaan (4).

$$w_i^{(t+1)} = w_i^{(t)} \cdot e^{\alpha_t I(y_i \neq h_t(x_i))} \quad (4)$$

Setelah itu, bobot dinormalisasi agar jumlah total bobot menjadi 1.

Setelah semua iterasi selesai, prediksi akhir dari model *ensemble* dibuat dengan menjumlahkan prediksi setiap *weak learner* yang dibobotkan oleh kekuatannya (α_t). Kombinasi model dirumuskan menggunakan persamaan (5).

$$H(x) = \text{sign}(\sum_{t=1}^T \alpha_t \cdot h_t(x)) \quad (5)$$

Proses ini menghasilkan model prediksi yang lebih akurat dengan memanfaatkan kekuatan dari setiap *weak learner* dan meminimalkan kelemahan mereka secara iteratif. Jika diperlukan, algoritma ini dapat diterapkan untuk berbagai jenis *weak learners*, seperti pohon keputusan sederhana atau regresi logistik.

F. Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa prediksi dengan membandingkan hasil prediksi model terhadap nilai aktual pada data *testing* [21]. Evaluasi dimulai dengan menghitung *confusion matrix*, yang menunjukkan jumlah prediksi benar dan salah untuk masing-masing kelas, termasuk *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN) [22]. *Confusion matrix* berfungsi sebagai dasar analisis kinerja model klasifikasi dengan menyajikan visualisasi komprehensif tentang kemampuan model dalam membedakan antar kelas dan mengidentifikasi pola kesalahan prediksi. Berdasarkan *confusion matrix*, beberapa metrik kinerja dihitung, seperti akurasi, *precision*, *recall*, dan *F1-score*. Selain itu, analisis *Receiver Operating Characteristic* (ROC) *Curve* dilakukan untuk mengevaluasi *trade-off* antara tingkat deteksi positif (*True Positive Rate*) dan tingkat kesalahan positif (*False Positive Rate*). *Area Under the Curve* (AUC) digunakan sebagai indikator kemampuan model

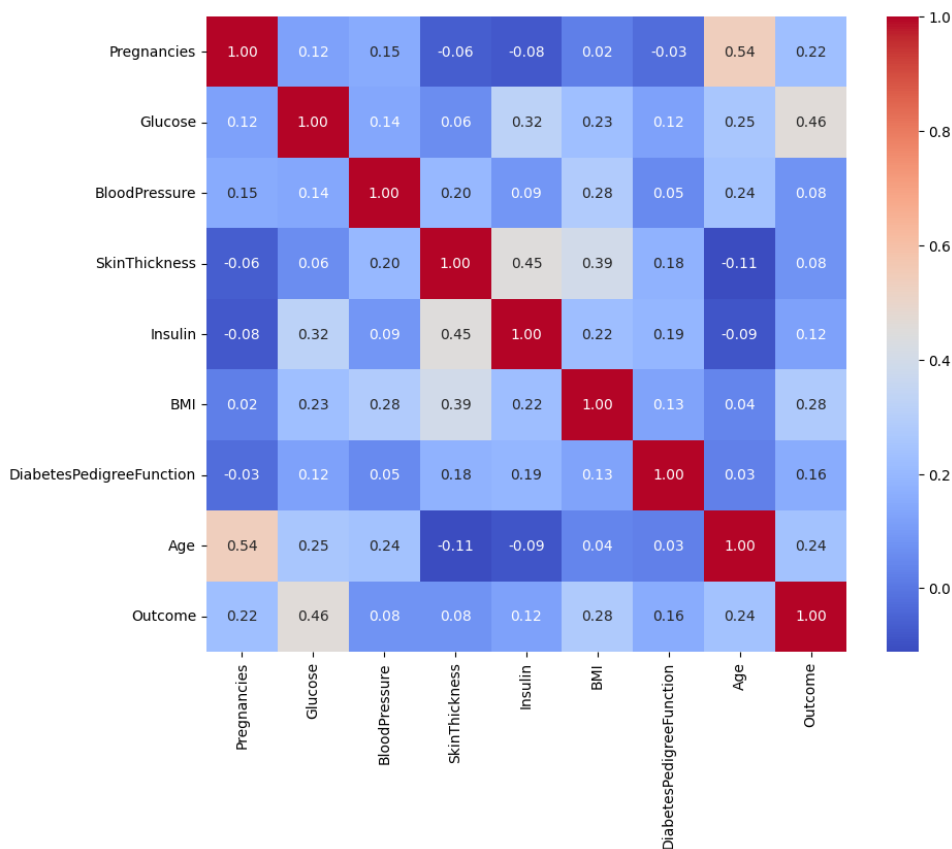
untuk membedakan antara kelas positif dan negatif. Hasil evaluasi dibandingkan dengan model baseline untuk menilai keunggulan pendekatan yang diusulkan.

G. Analisis dan Interpretasi

Tahap terakhir adalah analisis dan interpretasi hasil evaluasi model. Analisis mencakup penilaian efektivitas kombinasi SMOTE dan AdaBoost dalam meningkatkan akurasi dan sensitivitas prediksi pada kelas minoritas. Visualisasi, seperti *confusion matrix*, *ROC Curve*, dan distribusi probabilitas prediksi, digunakan untuk mempermudah interpretasi hasil. Temuan utama penelitian dirangkum sebagai kesimpulan, termasuk keunggulan dan keterbatasan dari model yang dikembangkan.

3.HASIL DAN PEMBAHASAN

Langkah awal dalam membangun model prediksi risiko diabetes adalah menyiapkan dataset yang akan digunakan untuk pelatihan dan pengujian. Dataset dalam penelitian ini diambil dari *platform* Kaggle dengan nama "Diabetes" (<https://www.kaggle.com/datasets/johndasilva/diabetes>), yang terdiri dari 2000 entri dan 9 kolom [15]. Dataset ini mencakup variabel-variabel penting seperti jumlah kehamilan (*Pregnancies*), kadar glukosa (*Glucose*), tekanan darah (*Blood Pressure*), ketebalan kulit (*Skin Thickness*), kadar insulin (*Insulin*), indeks massa tubuh (BMI), skor riwayat genetik diabetes (*Diabetes Pedigree Function*), usia (*Age*), serta hasil akhir diagnosis diabetes (*Outcome*, di mana 1 menunjukkan diabetes dan 0 menunjukkan non-diabetes). Setelah dataset diperoleh, dilakukan eksplorasi awal untuk memahami karakteristik data. Eksplorasi data mencakup analisis korelasi antar variabel, deteksi *outlier*, dan distribusi data target. Analisis korelasi dilakukan untuk memahami hubungan linear antara variabel-variabel dalam dataset, serta untuk mengidentifikasi fitur-fitur yang paling relevan terhadap variabel target. Visualisasi korelasi menggunakan heatmap ditunjukkan pada Gambar 2.



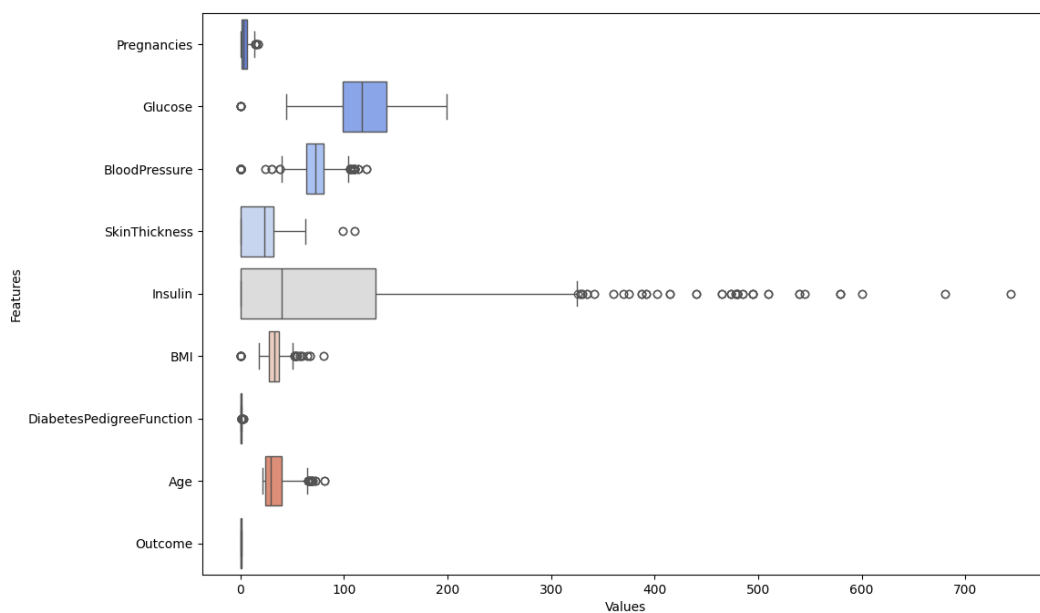
Gambar 2. Heatmap Korelasi Antar Variabel

Pada Gambar 2 memvisualisasikan nilai korelasi dalam bentuk gradasi warna, di mana warna yang lebih terang atau lebih gelap menunjukkan kekuatan hubungan yang lebih besar, baik positif maupun



negatif. Korelasi terbesar dengan variabel *Outcome* (indikator diabetes) adalah *Glucose* (0.46), diikuti oleh *BMI* (0.28) dan *Age* (0.24), menunjukkan bahwa kadar glukosa, indeks massa tubuh, dan usia memiliki hubungan yang signifikan dalam memprediksi diabetes. Variabel lain seperti *Blood Pressure*, *Skin Thickness*, dan *Insulin* memiliki korelasi yang lebih rendah dengan *Outcome* (<0.15), menunjukkan relevansi yang lebih kecil. Selain itu, terdapat korelasi tinggi antar beberapa variabel, seperti antara *Pregnancies* dan *Age* (0.54), yang menunjukkan potensi hubungan erat antara usia dan jumlah kehamilan. Visualisasi ini membantu mengidentifikasi fitur yang paling relevan untuk prediksi dan hubungan antar fitur yang mungkin memengaruhi analisis lebih lanjut.

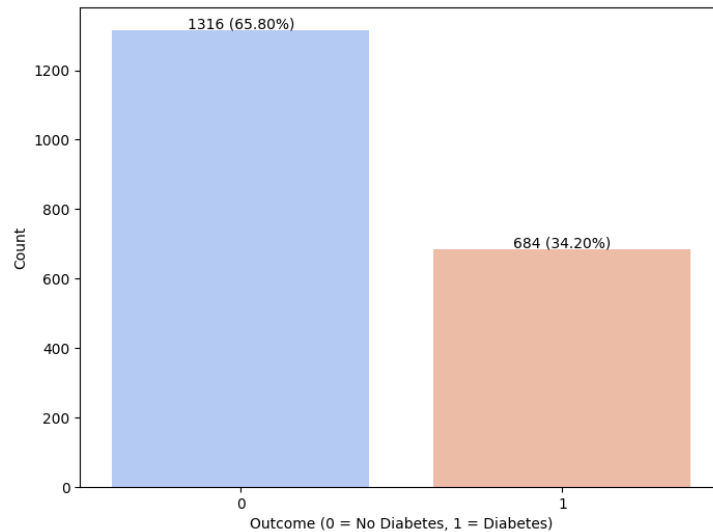
Selanjutnya, dilakukan eksplorasi data untuk mendeteksi keberadaan *outlier*, yaitu nilai-nilai ekstrem yang dapat memengaruhi hasil analisis dan performa model prediksi. Visualisasi deteksi *outlier* bertujuan untuk mengidentifikasi data yang berada di luar rentang normal distribusi, sehingga langkah penanganan seperti penghapusan atau transformasi data dapat dilakukan jika diperlukan. Deteksi *outlier* divisualisasikan menggunakan *boxplot*, yang ditampilkan pada Gambar 3.



Gambar 3. Boxplot Deteksi Outlier

Gambar 3 merupakan *boxplot* yang digunakan untuk mendeteksi *outlier* pada setiap fitur numerik dalam dataset diabetes. Setiap fitur memiliki distribusi data yang ditampilkan melalui median (garis di dalam kotak), *Interquartile Range* (IQR), dan *whisker* (garis vertikal). *Outlier* ditandai dengan titik-titik yang berada di luar *whisker*, yang menunjukkan nilai ekstrem di luar rentang normal. Fitur seperti *Insulin* memiliki banyak *outlier* dengan nilai yang jauh lebih tinggi dibandingkan distribusinya. Selain itu, fitur seperti *Skin Thickness*, *Blood Pressure*, dan *BMI* juga menunjukkan beberapa *outlier*. Visualisasi ini membantu mengidentifikasi nilai-nilai ekstrem yang mungkin memengaruhi performa model prediksi, sehingga perlu dipertimbangkan untuk ditangani dalam proses *preprocessing* data.

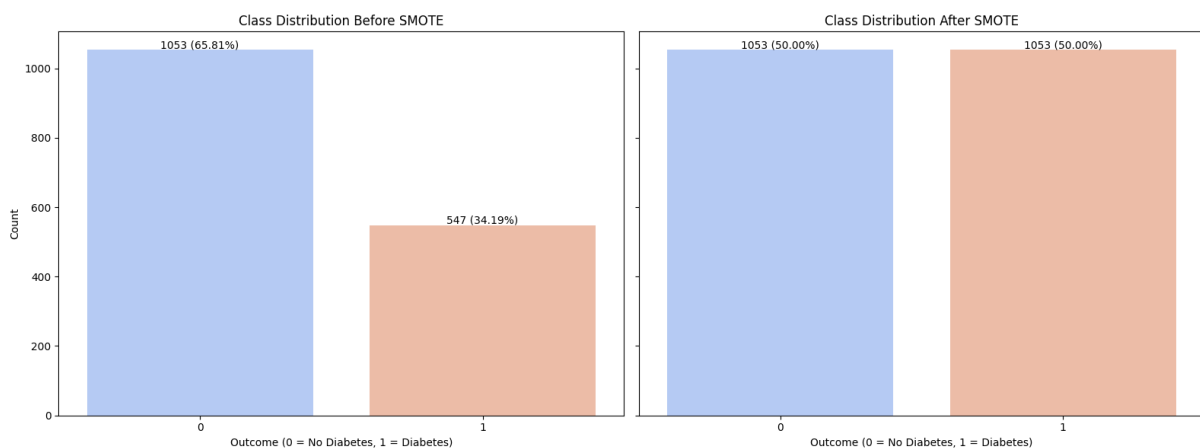
Langkah penting berikutnya dalam eksplorasi data adalah menganalisis apakah distribusi data target berada dalam kondisi seimbang atau tidak seimbang. Analisis ini bertujuan untuk memahami proporsi antara kelas positif dan negatif dalam variabel target, yang sangat penting dalam menentukan strategi pemodelan yang tepat. Jika data target tidak seimbang, langkah-langkah khusus seperti *oversampling* atau *undersampling* perlu dilakukan untuk memastikan model dapat menangani kelas minoritas dengan baik. Untuk menganalisis distribusi tersebut, visualisasi berupa grafik distribusi data target dibuat, seperti yang ditampilkan pada Gambar 4.



Gambar 4. Distribusi Kelas Pada Variabel Outcome

Visualisasi pada Gambar 4 menunjukkan distribusi kelas pada variabel Outcome, di mana kelas 0 (tidak diabetes) mencakup 65,80% dari total data (1.316 sampel), sedangkan kelas 1 (diabetes) mencakup 34,20% dari total data (684 sampel). Distribusi ini mengindikasikan bahwa dataset memiliki ketidakseimbangan kelas, dengan jumlah sampel untuk kelas tidak diabetes jauh lebih banyak dibandingkan kelas diabetes. Ketidakseimbangan kelas dapat mempengaruhi performa model dengan bias prediksi terhadap kelas mayoritas.

Setelah memahami karakteristik data, langkah berikutnya adalah melakukan pra-pemrosesan, yang bertujuan untuk meningkatkan kualitas data agar siap digunakan dalam pelatihan model. Pra-pemrosesan sangat penting untuk memastikan bahwa data yang digunakan tidak hanya bersih dan konsisten, tetapi juga relevan dalam mendukung performa model prediksi yang optimal. Pra-pemrosesan yang dilakukan mencakup berbagai proses, seperti penanganan data yang hilang, normalisasi atau standarisasi fitur, deteksi dan penanganan outlier. Selanjutnya, dilakukan penerapan teknik oversampling SMOTE (Synthetic Minority Oversampling Technique) pada data training untuk menangani ketidakseimbangan kelas. SMOTE menghasilkan data sintetis untuk kelas minoritas, sehingga meningkatkan representasi kelas tersebut tanpa menduplikasi data asli. Hasil penanganan ketidak seimbangan data dengan menggunakan SMOTE ditampilkan pada Gambar 5.

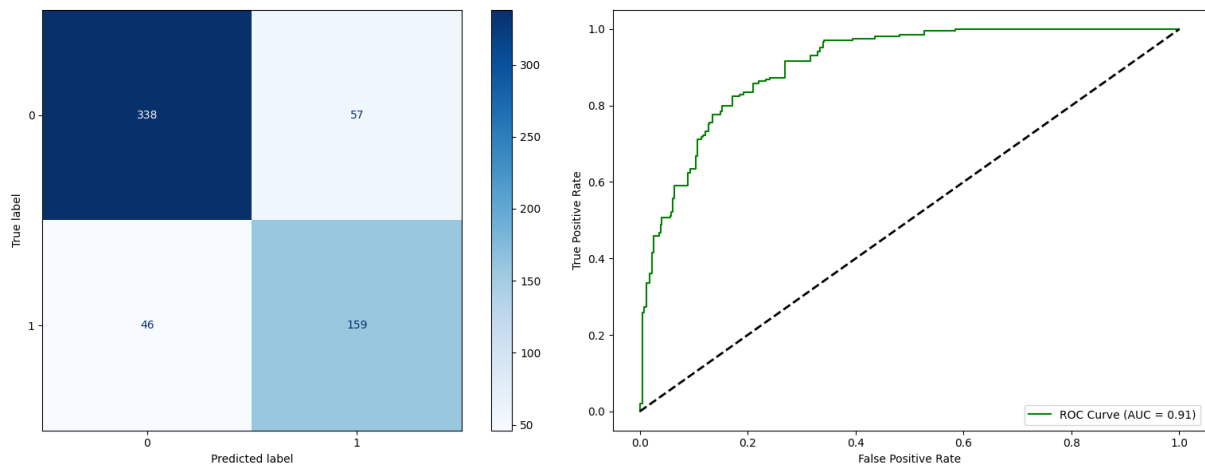


Gambar 5. Distribusi Kelas Target Sebelum dan Sesudah Penerapan SMOTE

Gambar 5 menunjukkan distribusi kelas target sebelum dan sesudah penerapan SMOTE. Sebelum SMOTE, dataset memiliki distribusi yang tidak seimbang, dengan kelas 0 (tanpa diabetes) mencakup 65,81% dari total sampel (1.053 sampel) dan kelas 1 (dengan diabetes) hanya 34,19% (547 sampel).

Setelah SMOTE diterapkan, distribusi kelas menjadi seimbang, dengan masing-masing kelas memiliki 50% sampel (1.053 sampel). Teknik ini bertujuan untuk mengatasi bias model terhadap kelas mayoritas dan meningkatkan performa prediksi pada kelas minoritas.

Setelah melalui tahap pra-pemrosesan, data yang telah diolah digunakan untuk membangun dan melatih model AdaBoost. Proses pembangunan model dimulai dengan memisahkan dataset menjadi fitur (x) dan target (y), di mana variabel *Outcome* dijadikan sebagai target. Selanjutnya, dataset dibagi menjadi *training set* dan *testing set* dengan rasio 80:20 menggunakan teknik *stratified sampling* untuk memastikan proporsi kelas target tetap seimbang antara data pelatihan dan pengujian. Model AdaBoost kemudian dilatih pada data *training* dengan tujuan meningkatkan akurasi prediksi melalui kombinasi iteratif dari beberapa model sederhana (*weak learners*). Pada setiap iterasi, AdaBoost memberikan bobot lebih besar pada sampel yang sulit diprediksi untuk memperbaiki kesalahan dari iterasi sebelumnya, sehingga menghasilkan model *ensemble* yang lebih kuat dan akurat dalam memprediksi risiko diabetes. Model yang telah dilatih dievaluasi menggunakan *confusion matrix* untuk menganalisis distribusi prediksi benar (*True Positive* dan *True Negative*) dan salah (*False Positive* dan *False Negative*) pada setiap kelas, serta *ROC Curve* untuk mengevaluasi *trade-off* antara *True Positive Rate* dan *False Positive Rate*. Visualisasi *confusion matrix* dan *ROC Curve* untuk model prediksi menggunakan AdaBoost dan SMOTE ditampilkan pada Gambar 6.



Gambar 6. *Confusion Matrix* dan *ROC Curve* Pada Model AdaBosst dan SMOTE

Selanjutnya yaitu mengevaluasi model dengan beberapa matrix diantaranya *precision*, *recall*, *F1-score*, *accuracy* dan *ROC-AUC Score*. Untuk memperlihatkan kinerja model inegrasi AdaBoost dan SMOTE maka dilakukan perbandingan dengan model AdaBoost tanpa teknik *oversampling*. Hasil evaluasi perbandingan tersebut disajikan pada Tabel 1.

Tabel 1. Hasil Evaluasi Model AdaBoost Tanpa *Oversampling* dan Adaboost + SMOTE

Model	Kelas	Precision	Recall	F1-Score	Accuracy	ROC-AUC Score
AdaBoost	No Diabetes	0.8396	0.8954	0.8666	0.8187	0.9031
	Diabetes	0.7699	0.6715	0.7173		
AdaBoost + SMOTE	No Diabetes	0.8802	0.8557	0.8678	0.8283	0.9058
	Diabetes	0.7361	0.7756	0.7553		

Tabel tersebut menunjukkan hasil evaluasi performa model AdaBoost tanpa *oversampling* dan AdaBoost dengan SMOTE dalam memprediksi diabetes berdasarkan berbagai metrik evaluasi, seperti *precision*, *recall*, *F1-Score*, akurasi, dan *ROC-AUC Score*. Pada model AdaBoost tanpa SMOTE, kelas *No Diabetes* memiliki *precision* sebesar 0,8396, *recall* sebesar 0,8954, dan *F1-Score* sebesar 0,8666, yang menunjukkan performa prediksi yang baik untuk kelas mayoritas. Namun, pada kelas *Diabetes*, nilai *precision*, *recall*, dan *F1-Score* masing-masing hanya mencapai 0,7699, 0,6715, dan 0,7173. Hal ini menunjukkan bahwa model kurang optimal dalam memprediksi kelas minoritas akibat

ketidakseimbangan data. Akurasi keseluruhan model adalah 0,8187, dengan ROC-AUC Score sebesar 0,9031, yang mencerminkan kemampuan model untuk membedakan antara kedua kelas dengan baik.

Setelah menerapkan AdaBoost dengan SMOTE, performa pada kelas minoritas (*Diabetes*) meningkat secara signifikan, dengan *precision* sebesar 0,7361, *recall* sebesar 0,7756, dan F1-Score sebesar 0,7553. Peningkatan ini menunjukkan bahwa SMOTE berhasil mengatasi ketidakseimbangan data, sehingga model lebih mampu mengenali kelas minoritas. Namun, performa pada kelas mayoritas (*No Diabetes*) sedikit menurun, dengan *precision* sebesar 0,8802, *recall* sebesar 0,8557, dan F1-Score sebesar 0,8678. Akurasi keseluruhan model meningkat menjadi 0,8283, dan ROC-AUC Score sedikit lebih baik di angka 0,9058. Integrasi AdaBoost dan SMOTE meningkatkan kinerja model prediksi dengan mengatasi ketidakseimbangan data melalui peningkatan representasi kelas minoritas. Dengan menambahkan data sintetis yang dihasilkan dari interpolasi antara sampel kelas minoritas yang ada, SMOTE menciptakan dataset yang lebih seimbang. Hal ini memungkinkan model untuk belajar secara lebih efektif dari kedua kelas, sehingga meningkatkan kemampuan model dalam mengenali kelas minoritas, seperti yang terlihat pada peningkatan nilai *recall* dan F1-Score untuk kelas tersebut.

Namun, ada beberapa keterbatasan dari pendekatan ini. Salah satu keterbatasannya adalah risiko *overfitting*, terutama jika data sintetis yang dihasilkan oleh SMOTE terlalu banyak atau tidak sepenuhnya mencerminkan distribusi data asli. Selain itu, penggunaan SMOTE dapat sedikit menurunkan performa pada kelas mayoritas, seperti yang terlihat dari penurunan nilai *recall* pada kelas ini, karena model lebih fokus pada kelas minoritas untuk mencapai keseimbangan. Hal ini mencerminkan adanya *trade-off* antara performa kelas mayoritas dan minoritas. Meskipun demikian, dalam konteks dataset yang tidak seimbang, kombinasi AdaBoost dengan SMOTE memberikan hasil yang lebih seimbang dan menunjukkan peningkatan performa dalam mengenali kelas minoritas tanpa terlalu mengorbankan kinerja pada kelas mayoritas.

4.KESIMPULAN

Penelitian ini berhasil membangun model prediksi risiko diabetes menggunakan metode AdaBoost dan teknik *oversampling* SMOTE untuk mengatasi masalah ketidakseimbangan data. Dataset yang digunakan terdiri dari 2000 sampel dengan distribusi kelas yang tidak seimbang, di mana kelas mayoritas (bukan penderita diabetes) mencakup 65,81% data, sementara kelas minoritas (penderita diabetes) hanya mencakup 34,19%. Hasil evaluasi menunjukkan bahwa model AdaBoost tanpa SMOTE memiliki akurasi sebesar 81,87% dan nilai ROC-AUC sebesar 0,9031. Meskipun model menunjukkan performa yang baik pada kelas mayoritas (bukan penderita diabetes), performanya menurun pada kelas minoritas (penderita diabetes), sebagaimana terlihat dari perbedaan signifikan pada *precision*, *recall*, dan F1-score antara kedua kelas. Hal ini mengindikasikan adanya bias model terhadap kelas mayoritas akibat distribusi data yang tidak seimbang. Setelah menerapkan AdaBoost dengan SMOTE, performa model pada kelas minoritas meningkat secara signifikan. *Precision*, *recall*, dan F1-Score untuk kelas mayoritas dan minoritas menjadi lebih seimbang, mencerminkan peningkatan kemampuan model dalam mengenali kasus diabetes. Selain itu, akurasi keseluruhan model meningkat menjadi 82,83%, dan nilai ROC-AUC sedikit meningkat menjadi 0,9058. Kombinasi AdaBoost dan SMOTE terbukti efektif dalam menangani dataset yang tidak seimbang, meningkatkan kemampuan model dalam mengenali kelas minoritas tanpa mengorbankan kinerja yang berarti pada kelas mayoritas. Penelitian selanjutnya dapat mengeksplorasi teknik SMOTE-ENN, *cost-sensitive learning*, dan validasi silang untuk mengatasi *overfitting*. Selain itu, dilakukan *feature engineering* yang lebih mendalam guna memastikan bahwa data sintetis yang dihasilkan oleh SMOTE lebih representatif terhadap distribusi data asli. Teknik *preprocessing* lanjutan, seperti normalisasi, transformasi *log*, atau *scaling*, juga dapat diterapkan untuk menghasilkan data yang lebih konsisten dan meningkatkan kinerja model.

5.REFERENSI

- [1] E. Retta, Hh. Kusumajaya, and Arjuna, "Faktor-Faktor yang Berhubungan dengan Pemilihan Pengobatan Herbal pada Pasien Diabetes Melitus," *J. Penelit. Perawat Prof.*, vol. 5, no. November, pp. 1541-1552, 2023, doi: 10.37287/jppp.v5i4.1891.
- [2] E. F. Santika, "Prevalensi Diabetes Indonesia Naik Jadi 11,7% pada 2023," Databoks.

- [3] E. Erika, "Meningkatkan Pemahaman Masyarakat Pentingnya Deteksi Dini Diabetes Melitus Melalui Penyuluhan Dan Pengukuran Gula Dan Tekanan Darah," *EJOIN J. Pengabd. Masy.*, vol. 1, no. 7, pp. 685-697, 2023, doi: 10.55681/ejoin.v1i7.1228.
- [4] P. N. Sabrina and A. Komarudin, "Prediksi Penyakit Diabetes Dengan Metode K-Nearest Neighbor (KNN) dan Seleksi Fitur Information Gain," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 8, no. 6, pp. 11320-11326, 2024, doi: 10.36040/jati.v8i6.11364.
- [5] G. A. Putri, A. Trimaysella, and A. Khoiriah, "Penerapan Klasifikasi Data Mining pada Diabetes Menggunakan Metode Naive Bayes," *J. Ilmu Komput. Teknol. Terap.*, vol. 1, no. 14, pp. 1-9, 2024.
- [6] N. Nurdiana and A. Algifari, "Studi Komparasi Algoritma ID3 dan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *Infotech J.*, vol. 6, no. 2, pp. 18-23, 2020, doi: 10.31949/infotech.v6i2.816.
- [7] M. I. Gunawan, D. Sugiarto, and I. Mardianto, "Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Search pada Algoritma Logistic Regression," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 3, p. 280, 2020, doi: 10.26418/jp.v6i3.40718.
- [8] A. Syukron, "Penerapan Metode SMOTE Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung," *J. Teknol. Inf. dan Terap.*, vol. 10, no. 1, pp. 47-50, 2023, doi: 10.25047/jtit.v10i1.313.
- [9] A. Franseda, "Integrasi Metode Decision Tree dan SMOTE untuk Klasifikasi Data Kecelakaan Lalu Lintas," *JUSTIN (Jurnal Sist. dan Teknol. Informasi)*, vol. 08, no. 3, pp. 282-290, 2020, doi: 10.26418/justin.v8i3.40982.
- [10] I. Ayuningtyas and E. U. Kasanah, "Penerapan Synthetic Minority Oversampling Technique (SMOTE) Pada Kasus Dampak Covid-19 Terhadap Penduduk Usia Kerja di Kalimantan Timur," *BESTARI Bul. Statistika dan Apl. Terkini*, vol. 1, no. 1, pp. 1-7, 2021.
- [11] R. Rousyati, A. N. Rais, E. Rahmawati, and R. F. Amir, "Prediksi Pima Indians Diabetes Database Dengan Ensemble Adaboost dan Bagging," *Evolusi J. Sains dan Manaj.*, vol. 9, no. 2, pp. 36-42, 2021, doi: 10.31294/evolusi.v9i2.11159.
- [12] A. Arifin *et al.*, "Klasifikasi Dalam Mendeteksi Penyakit Hipoglikemia Dengan Menggunakan Metode Random Forest dan Adaboost," in *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, Jakarta, 2022, pp. 542-549.
- [13] A. Byna and M. Basit, "Penerapan Metode Adaboost Untuk Mengoptimasi Prediksi Penyakit Stroke Dengan Algoritma Naive Bayes," *J. SISFOKOM (Sistem Inf. dan Komputer)*, vol. 09, no. 03, pp. 407-411, 2020, doi: 10.32736/sisfokom.v9i3.1023.
- [14] R. I. Borman and M. Wati, "Penerapan Data Mining Dalam Klasifikasi Data Anggota Kopdit Sejahtera Bandarlampung Dengan Algoritma Naive Bayes," *J. Ilm. Fak. Ilmu Komput.*, vol. 9, no. 1, pp. 25-34, 2020.
- [15] J. Dasilva, "Diabetes Dataset," Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/johndasilva/diabetes>
- [16] R. I. Borman, R. Napianto, N. Nugroho, D. Pasha, Y. Rahmanto, and Y. E. P. Yudoutomo, "Implementation of PCA and KNN Algorithms in the Classification of Indonesian Medicinal Plants," in *International Conference on Computer Science, Information Technology and Electrical Engineering (ICOMITEE)*, IEEE, 2021, pp. 46-50.
- [17] E. Sutoyo and M. A. Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *JEPIN (Jurnal Edukasi dan Penelit. Inform. Edukasi dan Penelit. Inform.)*, vol. 6, no. 3, pp. 379-385, 2020, doi: 10.26418/jp.v6i3.42896.
- [18] M. P. Pulungan, A. Purnomo, and A. Kurniasih, "Penerapan Smote Untuk Mengatasi Imbalance Class Dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 5, pp. 1033-1042, 2024, doi: 10.25126/jtiik.2024117989.
- [19] G. A. Mursianto, M. Falih, M. Irfan, T. Sakinah, D. Sandya, and P. Prasvita, "Perbandingan Metode Klasifikasi Random Forest dan XGBoost Serta Implementasi Teknik SMOTE pada Kasus Prediksi Hujan," in *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, Jakarta, 2021, pp. 41-50.
- [20] T. H. Saragih, M. Muliadi, M. R. Faisal, and M. A. I. N. R. Said, "AdaBoost Classifier untuk Klasifikasi Tanaman Jarak Pagar," *J. Komputasi*, vol. 9, no. 2, pp. 60-66, 2021, doi: 10.23960/komputasi.v9i2.2865.
- [21] R. I. Borman, F. Rossi, Y. Jusman, A. A. A. Rahni, S. D. Putra, and A. Herdiansah, "Identification of Herbal Leaf Types Based on Their Image Using First Order Feature Extraction and Multiclass SVM Algorithm," in *International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, IEEE, 2021, pp. 12-17.
- [22] R. I. Borman, Y. Fernando, and Y. Egi Pratama Yudoutomo, "Identification of Vehicle Types Using Learning Vector Quantization Algorithm with Morphological Features," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 2, pp. 339-345, 2022, doi: 10.29207/resti.v6i2.3954.