Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3939~3946
e-ISSN: 2548-6861

Optimization of Diabetes Mellitus Classification Using the Random Forest and SMOTE-ENN Methods

Imam Fadhur Rahman[1], Egia Rosi Subhiyakto[2*], Cinantya Paramita[3]
Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang, 50131, Indonesia
111202012509@mhs.dinus.ac.id[1], egia@dsn.dinus.ac.id[2], cinantya.paramita@dsn.dinus.ac.id[3]

Article Info

Article history:

Received 2025-09-30
Revised 2026-07-24
Accepted 2026-08-12

Keyword:

Classification
Diabetes Mellitus
Machine Learning
Random Forest
SMOTE-ENN

ABSTRACT

Diabetes Mellitus is a non-communicable disease (NCD) that has turned into a worldwide health issue with a steadily rising prevalence. Timely identification is essential for minimizing the risk of complications and the financial strain on the healthcare system. This research focuses on creating a precise and dependable diabetes classification model through the Random Forest algorithm by implementing a series of systematic data preprocessing methods. This methodology utilizes a dataset obtained from Kaggle, consisting of 768 samples. The steps taken include addressing missing values using Multiple Imputation by Chained Equations (MICE) , removing outliers with the Z-Score and Interquartile Range (IQR) techniques , selecting features based on ANOVA F-value to identify the eight most significant features , and balancing classes through the Synthetic Minority Over-sampling Technique-Edited Nearest Neighbours (SMOTE-ENN) to correct dataset imbalance. The assessment of the Random Forest model revealed outstanding performance, attaining an accuracy of 94.3% and an Area Under Curve (AUC) value of 0.98. These findings suggest that the model possesses strong discriminative capability to differentiate between diabetic and non-diabetic individuals. This research concludes that the Random Forest algorithm, when backed by suitable data preprocessing, is very efficient and could be utilized in clinical decision support systems as an early tool for diabetes screening.

I. PENDAHULUAN

Penyakit tidak menular (PTM) adalah salah satu dari sekian banyak masalah kesehatan utama di Indonesia. Sekarang, penyakit seperti diabetes melitus, kanker, hipertensi, stroke, dan jantung merupakan penyebab kematian dan kesakitan paling umum, menggantikan penyakit menular yang lebih sering terjadi sebelumnya (Waskito et al., 2020). Gaya hidup modern yang tidak sehat, seperti mengonsumsi rokok dan alkohol, pola makan yang tidak terjaga, kurangnya olahraga, dan stres berlebihan, memperburuk kondisi ini. Selain itu, kurangnya kesadaran masyarakat tentang deteksi dini dan pencegahan PTM membuat banyak kasus baru ditemukan dalam kondisi yang sudah parah. Selain itu, PTM menyebabkan beban ekonomi yang signifikan bagi individu dan sistem layanan kesehatan nasional.[1] International Diabetes Federation (IDF) melaporkan pada tahun 2022 bahwa sekitar 537 juta orang di seluruh penjuruh dunia menderita Diabetes Melitus. Angka ini diproyeksikan akan naik menjadi sekitar 578 juta pada tahun 2030 dan menjadi sekitar 700 juta pada tahun 2045. Di antara kasus ini, sepuluh persen penderita Diabetes Melitus Tipe 2 dan sepuluh persen penderita Diabetes Melitus Gestasional (IDF, 2024).[2]. Pada tahun 2019, diabetes adalah penyebab dari 1,5 juta kematian, dengan 48% dari seluruh kematian terjadi pada orang di bawah usia 70 tahun. Angka kematian akibat diabetes meningkat sebesar 3% berdasarkan kelompok umur antara tahun 2000 dan 2019, dan sebesar 13% meningkat di negara-negara berpenghasilan rendah. Menurut Organisasi Kesehatan Dunia (WHO, 2023).[3]

Dari sepuluh negara di Asia Tenggara, Indonesia memiliki 19,5 juta pasien diabetes tipe II (Dinkes Provinsi Lampung 2020). Data dari Riset Kesehatan Dasar menunjukkan bahwa prevalensi diabetes di Indonesia saat ini 10,9%, dan akan meningkat menjadi 11,7% pada tahun 2023. Selain itu, ada

perbedaan antara kelompok usia produktif (18-59 tahun) dan kelompok lansia di atas 60 tahun karena jumlah responden yang didiagnosa diabetes dan yang mendapatkan pengobatan atau kembali ke fasilitas kesehatan (BKPK, 2023).[4]

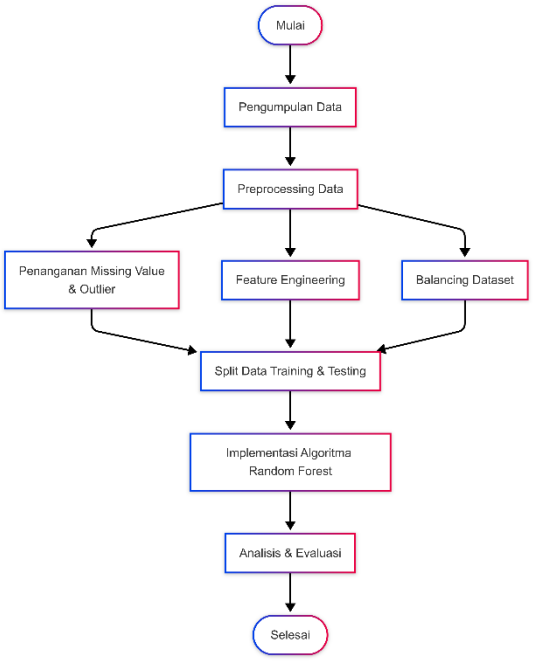
Dalam bidang ilmiah dan teknologi, klasifikasi adalah suatu teknik penting untuk mengorganisasi dan mengelompokkan data atau objek berdasarkan fitur tertentu. Proses ini memudahkan pengambilan keputusan yang lebih efektif dan analisis data yang lebih mendalam. Klasifikasi sering digunakan dalam ilmu komputer dan statistik untuk membuat model prediksi berdasarkan data historis; model ini menemukan dan memprediksi kategori data baru berdasarkan pola yang telah dipelajari dari data sebelumnya.[5] Karena ketidakseimbangan kelas, kecenderungan klasifikasi mungkin memilih kelas mayoritas.[6] Machine learning menganalisis pola dari riwayat medis pasien. Usia, hipertensi, kanker, dan penyakit kardiovaskular adalah beberapa faktor risiko dari riwayat medis pasien yang dapat digunakan untuk mengklasifikasikan diabetes [7] Beberapa algoritma machine learning yang dapat menangani masalah diagnosis penyakit adalah Random Forest, Naive Bayes, K-Nearest Neighbor, dan Logistic Regression. Hasil evaluasi model menunjukkan bahwa penelitian ini berjalan dengan baik; hasilnya menunjukkan akurasi keseluruhan sebesar 74% untuk penggunaan hutan random [8]Penelitian berjudul "Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression" dilakukan pada tahun 2020 oleh Muhamad Ichsan Gunawan, Dedy Sugiarto, dan Is Mardianto. Tujuan dari penelitian ini adalah untuk meningkatkan akurasi prediksi penyakit Diabetes Mellitus dengan menggunakan metode Regresi Logistik dan teknik Pencarian Grid. Penelitian ini menggunakan data dari Database Diabetes Pima Indians. Hasilnya menunjukkan bahwa model Regresi Logistik memiliki akurasi rata-rata sekitar 79%, dengan akurasi sebesar 83,33% ketika diuji dengan data baru. Penelitian lain tentang penyakit diabetes bertujuan untuk membuat model yang dapat memprediksi kemungkinan munculnya diabetes pada pasien dengan ketelitian maksimal (Sisodia & Sisodia, 2018).[9] Decision Tree, Support Vector Machine, Naive Bayes, dan Random Forest adalah metode pengklasifikasi. Eksperimen ini menggunakan Database Diabetes Pima Indians (PIDD), yang diperoleh dari Repositori Pembelajaran Mesin UCI. Hasilnya menunjukkan bahwa akurasi masing-masing algoritme adalah 76,30%, 65,10%, dan 73,82% dari Decision Tree, Support Vector Machine, dan Naive Bayes [10]. Selanjutnya, Sahat Pandapotan Nainggolan et al. melakukan analisis perbandingan akurasi algoritme klasifikasi hutan acak dan peningkatan gradient. [11] . Akurasi gradient boosting sebesar 81% dan random forest sebesar 79% ditemukan dengan 9 variabel. [12]. Penelitian ini mengusulkan rancang bangun pipeline hibrida yang menggabungkan metode imputasi adaptif Multiple Imputation by Chained Equations (MICE), filter ganda Z-Score dan Interquartile Range (IQR), serta teknik penyeimbang SMOTE-ENN. Algoritma Random Forest kemudian digunakan untuk mengoptimasi kerangka kerja pra-pemrosesan ini untuk membuat arsitektur prediksi yang mengejar metrik akurasi tinggi sambil mempertahankan kestabilan.

II. METODE

Sebagai bagian dari metodologi penelitian ini, proses pengumpulan data, pra-pemrosesan, balancing, dan implementasi model pembelajaran kelompok disusun secara sistematis. Setiap langkah dilakukan dengan tujuan meningkatkan keakuratan prediksi risiko diabetes dan menciptakan model yang dapat diandalkan dan membantu dalam pengambilan keputusan kesehatan.

A. Tahapan Penelitian

Untuk mencapai hasil yang optimal dalam penerapan klasifikasi penyakit diabetes menggunakan algoritma rondom forest, metodologi penelitian harus sistematis dan terstruktur. Tahapan penelitian digambarkan secara berurutan dalam diagram alur, yang memungkinkan setiap proses terorganisir dengan baik dan membantu tercapainya tujuan dari penelitian ini:



Gambar 1 tahapan penelitian

B. Pengumpulan Data

Dataset yang digunakan pada penelitian ini di dapat dari media platform Kaggle (mailto:https://www.kaggle.com/code/melikedilekci/diabetes-dataset-for-beginners/input). Dataset tersebut terdapat informasi mengenai berbagai factor yang saling berkaitan dengan penyakit diabetes seperti, usia, jenis kelamin, hipertensi, penyakit jantung, status pernikahan, jenis pekerjaan, tempat tinggal, indeks massa tubuh (BMI), dan status merokok. Dalam penelitian ini, 768 sampel data digunakan, terdiri dari 268 sampel data terkena diabetes dan

500 sampel data tanpa diabetes. Karena fakta bahwa jumlah kasus tanpa diabetes jauh lebih besar dibandingkan dengan kasus yang terkena diabetes, ketidakseimbangan data yang signifikan terjadi karena 65,1% data menunjukkan kasus tanpa diabetes dan 34,9% menunjukkan kasus yang terkena diabetes. Ketidakseimbangan ini dapat menyebabkan model memprediksi kasus tanpa diabetes dengan lebih baik, sehingga sensitivitas mendeteksi diabetes berkurang. Oleh karena itu, agar model dapat mengenali pola dengan lebih akurat, ketidakseimbangan harus diatasi melalui proses yang mencakup preprocessing dan balancing dataset.[13].

C. Preprocessing Data

Pra-Pemrosesan data merupakan Langkah yang penting dalam machine learning karena jika Pra-Pemrosesan diabaikan maka data bisa menjadi tidak terstruktur,mengandung data anomali, atau tidak lengkap yang dapat meerugikan kinerja model dan menghasilkan prediksi yang tidak dapat diandalkan . Dengan menggunakan kumpulan data yang tidak seimbang, penggunaan pra-pemrosesan data dapat mengurangi efek tumpang tindih kelas dan pergeseran kumpulan data dalam klasifikasi yang tidak seimbang. Kami menggunakan berbagai teknik pra-pemrosesan untuk penelitian ini. Ini termasuk pengkodean fitur kategorikal, rekayasa fitur baru, dan penanganan nilai hilang. Ini dilakukan untuk menjamin bahwa model akan menghasilkan hasil yang akurat dan dapat diandalkan. Oleh karena itu, pra-pemrosesan data merupakan dasar yang kuat untuk analisis lebih lanjut dan pembuatan model yang berhasil.

Penanganan Missing Value & Outlier: Data menunjukkan beberapa nilai yang hilang dari fitur Body Mass Index (BMI). Oleh karena itu, untuk mengatasi hilangnya data pada fitur BMI, metode Mutiple Imputation by Chained Equations (MICE) dengan algoritma imputatif itarative digunakan. Algoritma ini mengisi nilai yang hilang secara berulang dan menyelidiki pola hubungan antar fitur dalam dataset. [14] Untuk meningkatkan akurasi proses pengisian data, iterasi atau pengulangan kami dilakukan sepuluh kali. Metode ini memungkinkan estimasi nilai yang hilang menjadi lebih akurat dengan mempertimbangkan korelasi antar elemen dataset. Nilai data yang menyimpang dari distribusi umum dikenal sebagai outlier, yang dapat mengganggu kemampuan model untuk menemukan pola yang valid. Dalam penelitian ini, Z-Score dan Interquartile Range (IQR) digunakan untuk menemukan dan menangani outlier pada fitur BMI dan tingkat glukosa rata-rata. Z-Score adalah teknik yang menghitung deviasi suari nilai dari satuan standar deviasi untuk menunjukkan bahwa distribusi fitur BMI relatif normal. Rumus Z-Score digambarkan sebagai berikut Dengan X sebagai nilai sampel, mu sebagai rata-rata data, dan sigma sebagai standar deviasi Data yang memiliki nilai Z-Score di luar rentang -3 hingga +3 dianggap sebagai outlier dan dikeluarkan dari kumpulan data. Namun, fitur tingkat glukosa Avg yang tidak terdistribusi normal dievaluasi melalui metode IQR. Metode ini menemukan outlier berdasarkan perbedaan antara kuartil ketiga (Q3) dan kuartil pertama (Q1), yang dikenal sebagai IQR (Q3–Q1) dan titik data di luar rentang. dikeluarkan sebagai anomali. Untuk memastikan bahwa model dapat mempelajari pola yang lebih akurat tanpa mengganggu nilai ekstrem, kombinasi kedua teknik ini dipilih untuk mengatasi perbedaan karakteristik distribusi data.

1) Feature Engineering

Untuk meningkatkan kemampuan model untuk memprediksi kemungkinan diabetes dalam penelitian ini Kami mengolah data dengan menggunakan penan pengkodean fitur kategorikal untuk mengubah data non-numerik menjadi format yang dapat dipahami oleh model dan rekayasa fitur untuk menghasilkan fitur baru yang dapat menemukan pola tersembunyi dalam data.

a) Select K base

Proses seleksi fitur digunakan dalam penelitian ini untuk meningkatkan kinerja model klasifikasi dengan mengurangi jumlah fitur yang tidak relevan atau redundan. Untuk klasifikasi dengan target kategorikal, SelectKBest menggunakan fungsi skor ANOVA F-value (f_classif). [15].

Tujuan dari tahap seleksi fitur adalah untuk menghilangkan komponen prediktif yang tidak diperlukan. Menilai signifikansi masing-masing variabel terhadap kemungkinan diagnosis diabetes dilakukan dengan menggunakan skor pembobotan parametrik ANOVA F-value (melalui fungsi f_classif). Tidak ada batasan baku yang memengaruhi keputusan untuk menggunakan delapan atribut yang dipilih; sebaliknya, hasil dari ambang batas yang dihasilkan oleh proses validasi model dan kalkulasi relasi klinis memengaruhi keputusan tersebut. Meskipun fitur seperti tekanan darah dan ketebalan kulit memiliki skor probabilitas F yang rendah, evaluasi menemukan bahwa fitur-fitur ini tetap dipertahankan karena mereka membawa kontribusi medis kecil. Ini membantu pohon keputusan algoritma Random Forest saat membelah node (pemilihan cabang informasi acak).

Perhitungan nilai F statistik digunakan untuk memilih fitur ini. Nilai F yang dihasilkan oleh suatu fitur sebanding dengan pengaruhnya terhadap variabel target. Faktor-faktor dengan skor tertinggi dianggap paling informatif dalam membedakan antara orang yang menderita diabetes dan orang yang tidak menderita diabetes.Pemilihan 8 fitur ini membantu menyederhanakan model tanpa mengurangi kualitas data yang dibutuhkan. Selain itu, seleksi fitur dapat mempercepat proses pelatihan model dan mengurangi risiko overfitting.

Untuk menunjukkan kontribusi masing-masing fitur terhadap target, nilai F setiap fitur ditampilkan dalam bentuk grafik batang horizontal.. Diharapkan bahwa proses modeling menggunakan algoritma Random Forest akan meningkatkan akurasi dan efisiensi dalam klasifikasi penyakit diabetes setelah memilih fitur-fitur tertentu. Keunggulan tersebut sejalan dengan dataset yang kami gunakan yang memiliki banyak fitur kategorikal. [16]

2) Balancing Dataset

Ketidakseimbangan kelas sering terjadi dalam dataset diabetes dapat mempengaruhi keakuratan prediksi, dengan rasio 50,8% kasus survival dan 49,2 % kasus mortalitas. Untuk mengatasi permasalahan ini, kami menggunakan metode penyeimbangan data khusus untuk kasus medis:

a) SMOTE (Synthetic Minority Over-sampling Technique)

Dengan menggunakan proses interpolasi linier antar titik data terdekat di dalam ruang fitur, teknik oversampling SMOTE dapat meningkatkan jumlah sampel kelas minoritas dengan menyeimbangkan distribusi data pada kelas minoritas.[17]. Untuk menghilangkan gangguan data yang dapat mengaburkan pola prediksi diabetes, kombinasi Edited Nearest Neighbours dan TomekLinks digunakan. Teknik ini dipilih karena bisa menemukan dan menghilangkan kasus yang membingungkan model tanpa menghilangkan data klinis penting. [18].

b) Edited Nearest Neighbours

Untuk menyeimbangkan kelas minoritas dalam dataset diabetes.Interpolasi antara pasien yang memiliki karakteristik yang sebanding, terutama dengan parameter medis penting, digunakan untuk membuat sampel sintetis. Karena sampel baru dibuat berdasarkan pola data nyata, SMOTE membantu meningkatkan jumlah kasus kematian tanpa menyebabkan overfitting. ENN membuang titik-titik anomali—baik sampel asli dari kelas mayoritas maupun *noisy synthetic samples* buatan SMOTE—apabila klasifikasinya tidak cocok atau "berlawanan" dengan dominansi *k-Nearest Neighbors* di sekelilingnya. Pembersihan ini menjamin batas garis putusan (*decision boundaries*) model menjadi tegas, akurat, dan tidak terganggu pola ilusi.

c) Pengendalian Outlier

ADASYN digunakan sebagai metode oversampling sekunder. AdaSYN secara adaptif menghasilkan lebih banyak sampel sintetis untuk kasus gagal jantung yang lebih kompleks, berbeda dengan SMOTE. Dengan parameter n_neighbors yang lebih besar, sampel sintetis dapat mencerminkan variasi yang lebih besar dalam karakteristik pasien.

D. Split Data Training & Testing

Setelah proses preprocessing selesai, langkah selanjutnya adalah memisahkan data menjadi 20% data uji dan 80% data latihan dengan menggunakan metode train_test_split dari kumpulan scikit-learn. Pemisah ini dilakukan untuk memastikan bahwa model dapat belajar dari data latih dan menggunakannya untuk data uji. Parameter stratify digunakan untuk melatih distribusi kelas agar seimbang pada data uji.[19]. Strategis memungkinkan model mempelajari pola dengan adil, menghindari kecerundungan prediksi yang tidak diinginkan. Setelah pelatihan selesai, hanya data uji yang dapat digunakan untuk mengevaluasi kinerja model. Hal ini dilakukan untuk menghindari penyebaran data yang dapat mengakibatkan hasil evaluasi yang tidak sah.

E. Implemantasi Algoritma Rondom Forest

Proses penggunaan algoritma Random Forest untuk klasifikasi penyakit diabetes dilakukan secara sistematis. Random Forest membuat banyak pohon keputusan dari subset data dan fitur secara acak, kemudian menggabungkan hasilnya untuk meningkatkan akurasi dan mengurangi overfitting. Model dilatih menggunakan dataset yang telah melalui preprocessing menyeluruh, termasuk penanganan nilai hilang, outlier, dan penyeimbangan data. Untuk mengevaluasi stabilitas dan kinerja model pada berbagai skenario pembagian data. [20].

F. Analisis dan Evaluasi

Analisis dan evaluasi model dilakukan secara menyeluruh dengan penggunaan berbagai metrik standar evaluasi performa. Analisis confusion matrix memberikan gambaran mendalam tentang distribusi prediksi benar dan salah, akurasi untuk menilai ketepatan prediksi secara keseluruhan, recall untuk menilai kemampuan model untuk mendeteksi kasus positif, dan Area Under Curve (AUC-ROC) untuk menilai kemampuan diskriminatif model. [21]

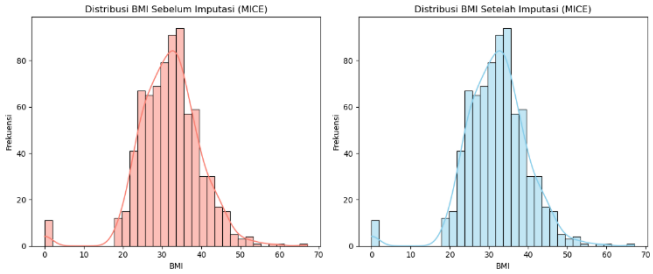
III. HASIL DAN PEMBAHASAN

Pada bagian ini menyajikan hasil penggunaan model dan juga membahas seberapa baik model mengklasifikasikan risiko diabetes.

A. Hasil Klasifikasi

Pada proses klasifikasi menggunakan metode algoritma random forest pengujian ini menghasilkan :

1) Distribusi Nilai BMI

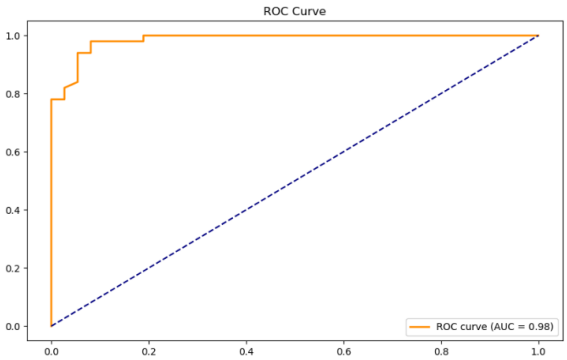


Gambar 1 Distribusi nilai BMI sebelum dan sesudah MICE

Berdasarkan confusion matrix yang ditunjukkan, model menunjukkan kinerja yang sangat baik, dengan tingkat akurasi mencapai 94.3%. Dari 87 data yang diuji, model berhasil memprediksi 82 data, terdiri dari 34 prediksi benar untuk kelas 0 (True Negative) dan 48 prediksi benar untuk kelas 1. Hanya 5 kesalahan yang dilakukan oleh model, yaitu 3 prediksi false positive di mana kelas 0 diprediksi sebagai positif, dan 2

prediksi false negatif di mana kelas 0 diprediksi sebagai negatif.

Secara keseluruhan, model menunjukkan konsistensi yang baik terhadap kedua kelas, dengan nilai 94.4% untuk kelas 0 dan 94.1% untuk kelas 1, yang menunjukkan bahwa ketika model memprediksi suatu kelas, prediksi tersebut memiliki tingkat kepercayaan yang tinggi. Sebaliknya, recall model untuk kelas 0 adalah 91.9% dan untuk kelas 1 adalah 96.0%, yang menunjukkan bahwa model sedikit lebih sensitif dalam mendeteksi kelas 1 daripada kelas 0.



Gambar 2 ROC Curve

Dengan nilai AUC (Area Under Curve) sebesar 0.98, model menunjukkan performa klasifikasi yang sangat baik, seperti yang ditunjukkan oleh ROC Curve. Kurva ROC berwarna oranye terlihat sangat mendekati sudut kiri atas grafik, yang menunjukkan bahwa model memiliki kemampuan yang hampir sempurna untuk membedakan kelas positif dan negatif. Bentuk kurva yang naik tajam di awal dan kemudian mendatar di atas menunjukkan bahwa model dapat mencapai rasio positif asli (sensitivitas) yang sangat tinggi sambil rasio negatif palsu yang sangat rendah.

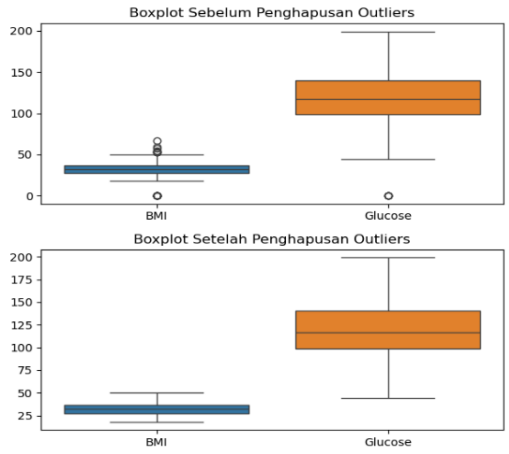
Dengan nilai AUC 0.98, terdapat kemungkinan 98% bahwa model akan memberikan skor yang lebih tinggi untuk sampel positif dibandingkan dengan sampel negatif yang dipilih secara acak. Garis diagonal putus-putus berwarna biru menunjukkan performa random classifier (AUC = 0.5).

TABEL 1
HASIL KLASIFIKASI

Accuracy	Precision	Recall
0,9425	0,9444	0,9189
	0,9412	0,9600

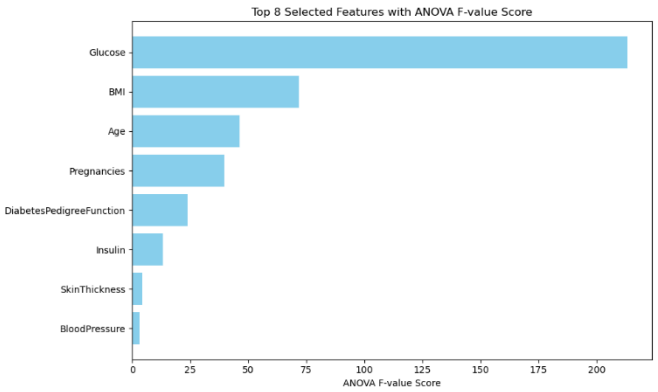
Hasil klasifikasi menunjukkan bahwa model klasifikasi sangat efektif. Untuk menunjukkan kemampuan generalisasi model terhadap data, evaluasi ini menggunakan set data uji, yang tidak digunakan selama proses pelatihan atau tuning. Untuk kategori negatif (diabetes), model precesion mencapai 0,9444, dan recall 0,9189. Sementara untuk kategori positif (diabetes), model mencapai precesion 0,9412, recall 0,9600. Akurasi keseluruhan model mencapai 0,9425 , menunjukan bahwa hampir seluruh prediksi dilakukan dengan tepat.

2) *Boxplot Distribusi*



Gambar 3 Boxplot sebelum dan sesudah penghapusan outlier

Dalam proses klasifikasi penyakit diabetes menggunakan algoritma Random Forest, boxplot menunjukkan variabel BMI dan glukosa sebelum dan setelah penghapusan outlier. Sebelum penghapusan, ada beberapa outlier ekstrim yang tidak wajar. Misalnya, nilai BMI di atas 50 dan nilai glukosa mendekati 200 bahkan 0, yang secara medis tidak wajar dan dapat mengganggu pembelajaran model. Outlier semacam ini dapat menyebabkan pembentukan struktur pohon keputusan yang tidak optimal. Akibatnya, akurasi dan generalisasi model akan menurun. Setelah penghapusan outlier, distribusi kedua variabel menjadi lebih terpusat dan bersih dan rentang nilainya menjadi lebih realistis. Ini memungkinkan Random Forest untuk menemukan pola-pola yang signifikan dalam data. Oleh karena itu, untuk meningkatkan kinerja model klasifikasi diabetes, penghapusan outlier merupakan tahap preprocessing yang sangat penting.

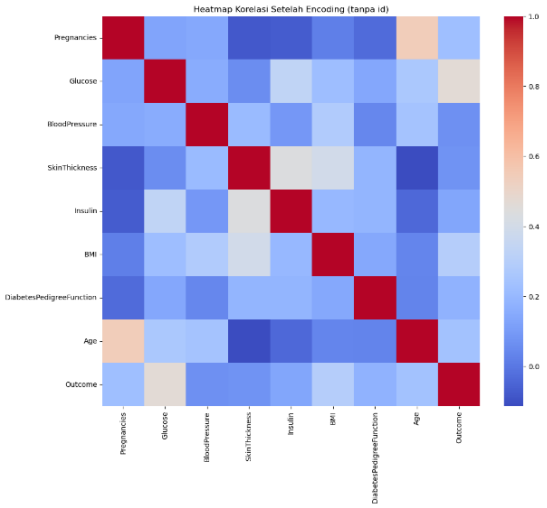


Gambar 4 Seleksi Fitur dengan ANOVA F-Test

Gambar 4 menujukan hasil seleksi fitur menggunakan algoritma Random Forest untuk menghasilkan skor ANOVA F-value terhadap delapan variabel dalam dataset klasifikasi penyakit diabetes. Hasil menunjukkan bahwa variabel glukosa menerima nilai F-score tertinggi, menunjukkan bahwa kadar

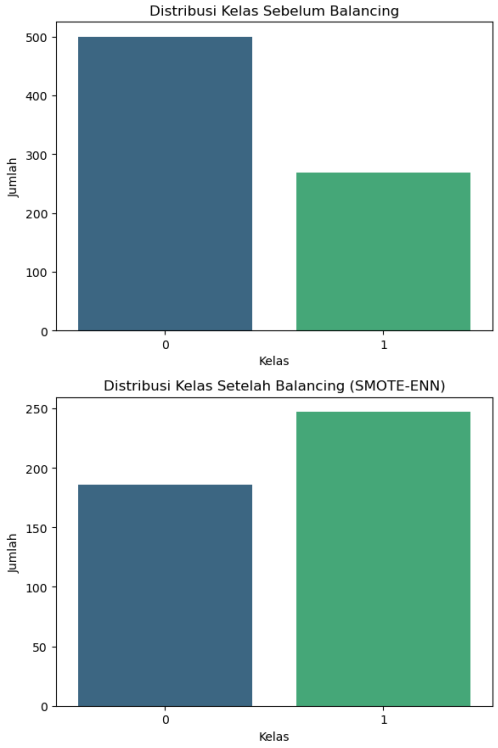
glukosa darah merupakan aspek yang paling berpengaruh dalam membedakan antara orang yang menderita diabetes dan orang yang tidak menderita diabetes. Selanjutnya adalah usia dan BMI, yang keduanya sangat memengaruhi klasifikasi. Fitur tambahan seperti ketebalan kulit dan tekanan darah memiliki F-score yang sangat rendah, yang menunjukkan bahwa kontribusinya terhadap model sangat kecil. Model Random Forest dapat dibangun secara lebih efisien dengan mengidentifikasi fitur yang paling penting, terutama yang memiliki kekuatan diskriminatif tinggi. Ini meningkatkan akurasi model, mempercepat proses pelatihan, dan mengurangi kompleksitas model tanpa mengurangi performa.

3) Heatmap



Gambar 5 Heatmap korelasi antar fitur

Diatas, variabel glukoss menunjukan korelasi paling kuat dengan hasil, yang sesuai dengan fakta bahwa glukosa darah tinggi merupakan indikator utama diabetes. Variabel kehamilan juga memiliki korelasi positif dengan hasil, karena jumlah kehamilan dapat meningkatkan risiko metabolik dan resistensi insulin pada perempuan. Selain itu, ada korelasi antara indeks massa tubuh (BMI) dan usia, di mana indeks massa tubuh yang tinggi dikombinasikan dengan bertambahnya usia secara fisiologis meningkatkan risiko terkena diabetes tipe 2. Namun, variabel lain seperti tekanan darah, ketebalan kulit, insulin, dan diabetes pedigree function menunjukkan korelasi yang lebih lemah terhadap hasil; oleh karena itu, meskipun mereka berkontribusi, pengaruh mereka tidak sebesar variabel utama. Hasil ini menunjukkan bahwa glukosa, kehamilan, BMI, dan usia adalah komponen utama yang dominan dalam klasifikasi penyakit diabetes menggunakan algoritma Random Forest. Ini karena mereka memiliki korelasi yang lebih signifikan dengan label hasil.

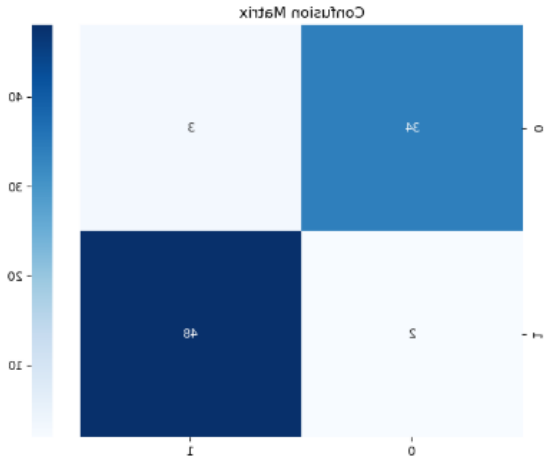


Gambar 6 Perbandingan Distribusi Kelas

Gambar 7 ditunjukan distribusi kelas sebelum dilakukan proses balancing. Data kelas 0 (tidak terkena diabetes) lebih banyak daripada data kelas 1. Kondisi ini dapat menyebabkan bias pada model karena algoritma cenderung "menghafal" pola dari kelas mayoritas dan mengabaikan kelas minoritas. Akibatnya, akurasi algoritma dalam mendeteksi pasien diabetes yang sebenarnya dapat menurun. Untuk menyelesaikan masalah ini, metode SMOTE-ENN—Teknik Sampling Minoritas Sintetis—Edited Nearest Neighbours—digunakan untuk mengimbangi data. Setelah balancing, distribusi kelas menjadi lebih seimbang, dan jumlah data untuk kelas 0 dan kelas 1 hampir sama. Perubahan distribusi ini diharapkan akan membuat algoritma Random Forest lebih baik dalam mengenali pola dari kedua kelas secara adil. Ini akan memungkinkan model untuk mengklasifikasikan pasien dengan diabetes dengan lebih baik daripada yang tidak. [22]

4) Confuison Matrix

Berdasarkan confusion matrix yang ditunjukkan, model menunjukkan kinerja yang sangat baik, dengan tingkat akurasi mencapai 94.3%. Dari 87 data yang diuji, model berhasil memprediksi 82 data, terdiri dari 34 prediksi benar untuk kelas 0 (True Negative) dan 48 prediksi benar untuk kelas 1. Hanya 5 kesalahan yang dilakukan oleh model, yaitu 3 prediksi false positive di mana kelas 0 diprediksi sebagai positif, dan 2 prediksi false negatif di mana kelas 0 diprediksi sebagai negatif.



Gambar 7 Confusion Matrix

Secara keseluruhan, model menunjukkan konsistensi yang baik pada kedua kelas, dengan nilai 94.4% untuk kelas 0 dan 94.1% untuk kelas 1, yang menunjukkan bahwa ketika model memprediksi suatu kelas, prediksi tersebut memiliki tingkat kepercayaan yang tinggi. Sebaliknya, recall model untuk kelas 0 adalah 91.9% dan untuk kelas 1 adalah 96.0%, yang menunjukkan bahwa model sedikit lebih sensitif dalam mendeteksi kelas 1 daripada kelas 0.

IV. KESIMPULAN

Dengan menggunakan berbagai teknik preprocessing dan optimasi yang sistematis, penelitian ini berhasil mengembangkan model klasifikasi penyakit diabetes menggunakan algoritma Random Forest. Berdasarkan hasil eksperimen dan evaluasi model, dapat disimpulkan bahwa tahap preprocessing data sangat berkontribusi terhadap performa model. Ini termasuk penerapan MICE untuk menangani nilai yang tidak ada, penggunaan metode IQR untuk menghilangkan outliers, dan standardisasi fitur numerik. Dengan menggunakan ANOVA F-value, pemilihan fitur berhasil menemukan delapan fitur utama. Sementara itu, SMOTE-ENN berhasil mengatasi ketidakseimbangan kelas dalam dataset.

Dengan akurasi 94.3% dan nilai AUC 0.98, model Random Forest menunjukkan hasil yang sangat baik. Strategi Verified K-Fold Cross Validation memastikan bahwa model konsisten dengan akurasi rata-rata yang stabil. Analisis fitur penting memberikan interpretasi yang signifikan untuk aplikasi kesehatan. Di sisi lain, kemampuan untuk membedakan pasien diabetes dan non-diabetes sangat baik, dengan kurva ROC dengan nilai AUC 0.98. Untuk menyelesaikan masalah seperti nilai yang hilang, anomali outlier, dan ketidakseimbangan kelas, penelitian ini menunjukkan bahwa arsitektur pipeline machine learning medis yang komprehensif dan dapat diandalkan adalah kontribusi ilmiah yang kuat. Arsitektur ini terbukti efektif dalam mengubah data klinis mentah menjadi model pendukung keputusan klinis yang lebih sensitif dan akurat.

Hasil penelitian menunjukkan bahwa algoritma Random Forest dengan preprocessing yang tepat dapat membantu dalam screening awal diabetes. Model ini cocok untuk digunakan dalam sistem pendukung keputusan klinis karena sangat akurat dan stabil. Untuk meningkatkan aplikabilitas klinis, studi kelompok teknik dan validasi yang lebih kompleks diperlukan untuk pengembangan berikutnya.

DAFTAR PUSTAKA

[1] N. N. Rosyidah and E. A. Cahyono, “Diabetes Melitus Tipe 2,” Enfermeria Ciencia, vol. 3, no. 1, pp. 44–63, Feb. 2025, doi: 10.56586/ec.v3i1.74.

[2] W. Ode Nurmila, A. Edy Dawu, P. S. Studi, and I. Teknologi adan KesehatanaAvicenna, “Hubungan Dukungan Keluarga, Pengetahuan dan Sikap Dengan Tingkat Kepatuhan Pasien dalam Pengobatan Diabetes Melitus di Wilayah Kerja Puskesmas Poasia Kota Kendari Tahun 2024,” 2025. doi: https://doi.org/10.69677/avicenna.v4i2.152.

[3] N. Hikmah and P. Yuwono, “Hubungan Efikasi Diri Terhadap Kualitas Hidup Pasien Diabetes Melitus Tipe 2 Di RS PKU Muhammadiyah Gombong,” Jurnal Ilmiah Kesehatan Keperawatan, vol. 21, no. 1, p. 6, Jul. 2025, doi: 10.26753/jikk.v21i1.1449.

[4] E. Elsa, “Hubungan Kepatuhan Diet Terhadap Kadar Glukosa Pada Penderita Diabetes Melitus Tipe II di Wilayah Kerja UPTD Puskesmas Selajambe Tahun 2025,” Abdimas Awang Long, vol. 8, no. 2, pp. 230–237, Jun. 2025, doi: 10.56301/awal.v8i2.1684.

[5] M. R. Maulana, A. Sucipto, and H. Mulyo, “Optimisasi Parameter Support Vector Machine Dengan Particle Swarm Optimization Untuk Peningkatan Klasifikasi Diabetes,” Nov. 2024.

[6] E. R. Subhiyakto et al., “Evaluation of Resampling Techniques in CNN-Based Heartbeat Classification,” Ingénierie des systèmes d information, vol. 29, no. 4, pp. 1323–1332, Aug. 2024, doi: 10.18280/isi.290408.

[7] M. Salsabil, N. Lutvi, and A. Eviyanti, “Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost,” Jurnal Ilmiah Komputasi, vol. 23, no. 1, Mar. 2024, doi: 10.32409/jikstik.23.1.3507.

[8] T. Hidayat, S. S. Anelia, R. I. Pratiwi, N. Salsabila, and D. S. Prasvita, Perbandingan Akurasi Klasifikasi Penyakit Diabetes Menggunakan Algoritma Adaboost-Random Forest Dan Adaboost-Decision Tree Dengan Imputasi Median Dan KNN. 2021. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database

[9] E. Cahya, P. Witjaksana, R. Rohmat Saedudin, and V. P. Widartha, “Perbandingan Akurasi Algoritma Random Forest Dan Algoritma Artificial Neural Network Untuk Klasifikasi Penyakit Diabetes,” Oct. 2021.

[10] W. Nugraha, “Resampling Technique for Handling Class Imbalance in the Classification of Diabetes using C4.5, Random Forest, and SVM,” Aug. 2021, doi: 10.33633/tc.v20i3.

[11] L. Maretva Cendani and A. Wibowo, “Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes,” 2022. doi: https://doi.org/10.14710/jmasif.13.1.42912.

[12] S. P. Nainggolan and A. Sinaga, “Comparative Analysis Of Accuracy Of Random Forest And Gradient Boosting Classifier Algorithm For Diabetes Classification,” Sebatik, vol. 27, no. 1, pp. 97–102, Jun. 2023, doi: 10.46984/sebatik.v27i1.2157.

[13] A. B. Alpiansah and Y. Ramdhani, “Optimasi Fitur dengan Forward Selection pada Estimasi Tingkat Obesitas menggunakan Random Forest,” SISTEMASI, vol. 12, no. 3, p. 860, Sep. 2023, doi: 10.32520/stmsi.v12i3.3125.

[14] Y. Tasya, “Kombinasi Metode Imputasi Mean Dan Multiple Imputation By Chained Equations (Mice) Untuk Penanganan Data Hilang Dan Peningkatan Evaluasi Kinerja Klasifikasi Prediksi Penyakit Diabetes Melitus,” 2023.

[15] F. Kurniawan, "Analisis Pengaruh Seleksi Fitur Anova Terhadap Performa Model Klasifikasi Gaussian Naïve Bayes Pada Dataset Pima Indians Diabetes," 2023.
[16] N. Nur Muttaqin, "Klasifikasi Penyakit Diabetes Menggunakan Metode Random Forest Dan Adaboost," 2024.
[17] C. Paramita, C. S. Simbolon, A. S. Pamungkas, J. M. Triono, E. P. Widi Utomo, and E. R. Subhiyakto, "Analisis Pengaruh SMOTE terhadap Kinerja Model KNN untuk Prediksi Risiko Stroke," Jurnal Informatika: Jurnal Pengembangan IT, vol. 10, no. 4, pp. 978–988, Sep. 2025, doi: 10.30591/jpit.v10i4.8809.
[18] M. A.-Z. Faradeya and E. R. Subhiyakto, "Klasifikasi Penyakit Gagal Jantung Menggunakan Algoritma Naive Bayes," Jurnal Algoritma, vol. 22, no. 1, pp. 115–127, May 2025, doi: 10.33364/algoritma/v.22-1.2178.
[19] M. Fadli and R. A. Saputra, "Klasifikasi Dan Evaluasi Performa Model Random Forest Untuk Prediksi Stroke Classification And Evaluation Of Performance Models Random Forest For Stroke Prediction," vol. 12, Oct. 2023, doi: http://dx.doi.org/10.31000/jt.v12i2.9099.
[20] N. H. Alfajr and S. Defiyanti, "Prediksi Penyakit Jantung Menggunakan Metode Random Forest Dan Penerapan Principal Component Analysis (PCA)," Jurnal Informatika dan Teknik Elektro Terapan, vol. 12, no. 3S1, Oct. 2024, doi: 10.23960/jitet.v12i3S1.5055.
[21] A. Latif and S. Khotimatul Wildah, "Analisis Kinerja Algoritma Ensemble Dalam Prediksi Perilaku Pembelian Pelanggan," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 9, no. 1, pp. 557–563, Dec. 2024, doi: 10.36040/jati.v9i1.12413.
[22] S. Ernawati and I. Maulana, "Meningkatkan Klasifikasi Penyakit Diabetes Menggunakan Metode Ensemble Softvoting Dengan SMOTE-ENN dan Optimasi Bayesian," Evolusi : Jurnal Sains dan Manajemen, vol. 13, no. 1, pp. 71–86, Mar. 2025, doi: 10.31294/evolusi.v13i1.8267.