JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9392
Hal 995-1005
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Analisis Perbandingan Metode Random Forest, XGBoost, dan Logistic Regression Untuk Klasifikasi Deteksi Dini Penyakit Diabetes

Novriansyah Afqi Nur Akmal Fauzi[1*], Fikri Budiman[2]

Fakultas Ilmu Komputer, Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1*]111202214701@mhs.dinus.ac.id, [2]fikri.budiman@dsn.dinus.ac.id
Email Penulis Korespondensi: 111202214701@mhs.dinus.ac.id
Submitted **01-12-2025;** Accepted **30-12-2025;** Published **31-12-2025**

Abstrak

Diabetes Mellitus merupakan penyakit kronis yang jumlah penderitanya terus mengalami peningkatan dan memberikan dampak serius terhadap kesehatan masyarakat serta beban ekonomi secara global. Gejala awal yang sering tidak spesifik menyebabkan risiko keterlambatan diagnosis, sehingga diperlukan pendekatan deteksi dini yang memiliki tingkat akurasi tinggi guna mendukung pengambilan keputusan klinis. Penelitian ini bertujuan menganalisis dan membandingkan kinerja tiga algoritma pembelajaran mesin, yaitu Logistic Regression, Random Forest, dan XGBoost, dalam mengklasifikasikan risiko diabetes berdasarkan sejumlah parameter klinis, meliputi usia, indeks massa tubuh (BMI), tekanan darah, kadar glukosa, serta HbA1c. Data penelitian diperoleh dari *Diabetes Prediction Dataset* yang terdiri atas 100.000 record. Proses penelitian mencakup penanganan data hilang, penerapan *One-Hot Encoding* pada variabel kategorikal, normalisasi fitur numerik, serta penyeimbangan distribusi kelas menggunakan metode Synthetic Minority Over-sampling Technique (SMOTE). Evaluasi performa model dilakukan menggunakan metrik *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *ROC-AUC* untuk memperoleh penilaian yang menyeluruh. Hasil pengujian menunjukkan bahwa algoritma XGBoost menghasilkan performa terbaik dengan tingkat akurasi sebesar 96,88% dan nilai *ROC-AUC* mencapai 98,00%. Sementara itu, Random Forest memperoleh akurasi 95,68% dengan nilai *F1-Score* sebesar 74,76%, dan Logistic Regression mencatatkan akurasi 88,96% dengan nilai *recall* tertinggi sebesar 89,12%. Temuan ini mengindikasikan bahwa metode *ensemble learning*, khususnya pendekatan *boosting*, memiliki kemampuan yang lebih baik dalam meningkatkan ketepatan klasifikasi pasien diabetes dan non-diabetes. Kontribusi penelitian ini terletak pada penyajian analisis komparatif berbasis evaluasi multi-metrik yang dapat dijadikan acuan dalam pemilihan model pembelajaran mesin paling efektif untuk pengembangan sistem pendukung keputusan medis pada deteksi dini diabetes.

Kata kunci: Diabetes; XGBoost; Random Forest; Logistic Regression; SMOTE.

Abstract

Diabetes Mellitus is a chronic disease with a continuously increasing prevalence, posing serious challenges to public health and contributing significantly to the global economic burden. The often non-specific nature of early symptoms increases the risk of delayed diagnosis, highlighting the need for accurate early detection approaches to support clinical decision-making. This study aims to analyze and compare the performance of three machine learning algorithms Logistic Regression, Random Forest, and XGBoost in classifying diabetes risk based on several clinical parameters, including age, body mass index (BMI), blood pressure, glucose level, and HbA1c. The dataset used in this research was obtained from the Diabetes Prediction Dataset, consisting of 100,000 records. The research process involved handling missing data, applying One-Hot Encoding to categorical variables, normalizing numerical features, and addressing class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE). Model performance was evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics to provide a comprehensive assessment. The experimental results indicate that XGBoost achieved the best performance, with an accuracy of 96.88% and a ROC-AUC value of 98.00%. Meanwhile, Random Forest attained an accuracy of 95.68% with an F1-Score of 74.76%, while Logistic Regression recorded an accuracy of 88.96% and the highest recall value of 89.12%. These findings suggest that ensemble learning methods, particularly boosting approaches, are more effective in improving the accuracy of diabetes and non-diabetes classification. The primary contribution of this study lies in providing a multi-metric comparative analysis that can serve as a reference for selecting the most effective machine learning model in the development of medical decision support systems for early diabetes detection.

Keywords: Diabetes; XGBoost; Random Forest; Logistic Regression; SMOTE.

1. PENDAHULUAN

Diabetes Mellitus adalah penyakit tidak menular dengan prevalensi yang terus mengalami peningkatan secara global dan telah menjadi salah satu tantangan besar dalam dunia kesehatan masyarakat modern [1], [2]. Berdasarkan IDF Diabetes Atlas edisi ke-10, > 537 juta orang dewasa di dunia hidup dengan diabetes pada tahun 2021, dan jumlah tersebut diperkirakan meningkat hingga mencapai 783 juta kasus pada tahun 2045 [3], [4]. Kondisi ini menunjukkan bahwa diabetes tidak hanya menjadi masalah klinis, tetapi juga permasalahan sosial dan ekonomi global. Laporan World Health Organization (WHO) mengungkapkan bahwasanya penyakit ini merupakan penyebab utama kematian kesembilan di dunia [5], dengan beban ekonomi global yang melampaui US$ 970 miliar pada tahun 2023 [6]. Situasi serupa juga terjadi di Indonesia, di mana prevalensi diabetes mengalami peningkatan dari 6,9% pada tahun 2013 hingga 10,9% pada tahun 2023, menjadikan Indonesia sebagai negara yang memiliki jumlah penderita diabetes paling banyak nomor lima di dunia [7], [8]. Peningkatan yang signifikan ini memperlihatkan pentingnya upaya deteksi dini untuk mencegah komplikasi dan mengurangi beban ekonomi nasional [9], [10]. Permasalahan utama dalam penanganan diabetes adalah keterlambatan serta ketidaktepatan diagnosis pada tahap awal akibat gejala yang tidak spesifik. Kondisi ini menyebabkan banyak pasien baru teridentifikasi setelah mengalami komplikasi, sehingga meningkatkan biaya pengobatan dan menurunkan kualitas hidup.

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9392
Hal 995-1005
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Kemajuan teknologi informasi dan komputasi dalam dua dekade terakhir telah memberikan kontribusi besar terhadap bidang kesehatan. Penerapan Artificial Intelligence (AI) dan Machine Learning (ML) kini menjadi pendekatan yang banyak dimanfaatkan untuk membantu diagnosis penyakit kronis seperti diabetes [11]. Melalui kemampuan algoritma pembelajaran mesin, komputer dapat memahami pola kompleks dalam data medis yang sulit diketahui secara manual oleh tenaga medis. Parameter seperti kadar gula darah, tensi, indeks massa tubuh (BMI), usia, serta riwayat keluarga menjadi faktor penting dalam menentukan risiko diabetes. Sejumlah penelitian telah membuktikan efektivitas pendekatan ini, misalnya Nurussakinah et al. (2025) mengindikasikan bahwasanya algoritma Random Forest mampu mencapai akurasi hingga 97% [12]. Sedangkan, Kurniawati dan Arianto (2023) melaporkan bahwa Logistic Regression masih memiliki performa kompetitif dengan tingkat akurasi sebesar 79,1% [13]. Temuan-temuan ini memperlihatkan potensi besar algoritma pembelajaran mesin dalam mendukung sistem diagnosis berbasis data yang lebih singkat, efisien, dan akurat.

Penelitian-penelitian sebelumnya juga telah mengembangkan dan mengevaluasi karakteristik berbagai model pembelajaran mesin dalam klasifikasi penyakit diabetes. Beberapa di antaranya menyoroti keunggulan dan karakteristik masing-masing algoritma. Random Forest, XGBoost, dan Logistic Regression merupakan tiga algoritma yang paling sering digunakan karena memiliki performa tinggi dan karakteristik saling melengkapi [14]. Random Forest dikenal stabil dan mampu mengatasi masalah overfitting sekaligus menangani data non-linear secara efektif [15]. Di sisi lain, XGBoost unggul dalam efisiensi dan kemampuan boosting adaptif yang dapat meningkatkan akurasi model [16], sedangkan Logistic Regression tetap populer karena modelnya sederhana, mudah ditafsirkan, dan cepat dalam komputasi [17]. Setyawan dan Wakhidah (2025) mencatat bahwa Random Forest memperoleh nilai F1-score sebesar 76%, sementara Susanto dan Cahyana (2025) menemukan bahwa XGBoost dapat mencapai nilai ROC-AUC hingga 0,99 [18], [19]. Walaupun hasil-hasil tersebut menunjukkan performa yang menjanjikan, kebanyakan penelitian masih dilakukan dalam ruang lingkup terbatas dan belum membandingkan ketiga algoritma tersebut secara langsung dengan kondisi eksperimen yang seragam.

Sementara itu, sebagian penelitian terdahulu masih berfokus hanya pada satu atau dua algoritma tanpa melakukan analisis lintas model secara komprehensif [20]. Hingga kini, belum banyak studi yang secara eksplisit membandingkan performa tiga alogaritma Random Forest, XGBoost, dan Logistic Regression menggunakan dataset publik yang sama, seperti Diabetes prediction dataset dari Kaggle dengan menggunakan metrik multi evaluasi yang lebih representatif [21]. Selain itu, metode evaluasi yang digunakan dalam sejumlah penelitian sebelumnya masih terbatas pada metrik tunggal, misalnya akurasi, tanpa mempertimbangkan indikator penting lain seperti Precision, Recall, F1-score, dan ROC-AUC [22]. Kondisi ini menunjukkan perlunya penelitian komparatif yang lebih sistematis untuk menilai keunggulan relatif dari setiap algoritma secara objektif dan terukur

Penelitian ini memiliki tujuan dalam rangka melaksanakan analisis komparatif tiga algoritma utama Random Forest, XGBoost, dan Logistic Regression untuk mengidentifikasi model yang paling optimal dalam klasifikasi diabetes. Perbandingan dilakukan menggunakan dataset publik Diabetes prediction dataset guna megevaluasi seberapa jauh setiap algoritma mampu menghasilkan prediksi yang akurat dan stabil. Sehingga memperoleh pemahaman yang lebih mendalam mengenai efektivitas setiap model dalam konteks diagnosis dini penyakit diabetes. Kontribusi utama penelitian ini terletak pada penyajian analisis perbandingan menggunakan berbagai metrik evaluasi yang diterapkan dalam skema eksperimen yang konsisten. Pendekatan ini memungkinkan penilaian yang lebih objektif terhadap kinerja relatif masing-masing algoritma sekaligus menghasilkan rekomendasi model pembelajaran mesin yang paling sesuai. Hasil analisis diharapkan dapat menjadi dasar ilmiah bagi pengembangan sistem untuk mendukung keputusan medis berbasis Artificial Intelligence yang efisien, akurat, dan mudah diimplementasikan pada praktik kesehatan digital modern [23], [24], [25]. Selain itu, meningkatnya pemanfaatan rekam medis digital dan platform kesehatan berbasis data membuka peluang penerapan model prediksi secara lebih luas dalam pelayanan klinis. Sehingga, penelitian komparatif ini tidak sebatas berkontribusi secara akademik, melainkan turut menawarkan landasan praktis bagi pengembangan sistem deteksi dini diabetes yang responsif, memliki tingkat akurasi presisi tinggi, dan dapat diintegrasikan pada berbagai fasilitas pelayanan kesehatan.

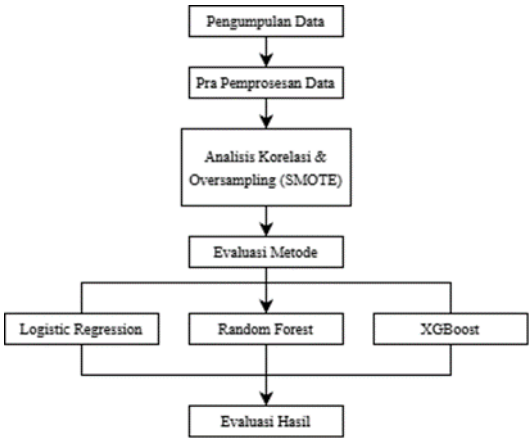
2. METODOLOGI PENELITIAN

2.1 Tahap Penelitian

Pada tahap Penelitian ini mempergunakan pendekatan kuantitatif komparatif dengan menerapkan tiga algoritma pembelajaran mesin, yaitu Logistic Regression, Random Forest, dan XGBoost, guna melaksanakan klasifikasi penyakit diabetes. Dataset yang digunakan adalah *Diabetes Prediction Dataset* dari Kaggle yang terdiri dari 100.000 data dengan sembilan atribut klinis dan satu label target. Data dibagi menggunakan teknik *stratified train-test split* dengan proporsi 80% data latih dan 20% data uji untuk menjaga distribusi kelas tetap seimbang Pendekatan ini dipilih karena mampu memberikan gambaran yang komprehensif mengenai perbandingan performa antara model linier dan non-linier [12], [15]. Logistic Regression digunakan sebagai model baseline karena bersifat interpretatif, stabil, dan mampu memberikan informasi yang jelas mengenai pengaruh setiap fitur terhadap kemungkinan seseorang menderita diabetes [13]. Random Forest dipilih sebab kapasitasnya mengatasi data non-linear, menangani fitur-fitur dengan variasi kompleks, serta tahan terhadap overfitting melalui mekanisme bagging [11]. Sementara itu, XGBoost digunakan karena dikenal unggul dalam efisiensi komputasi dan memiliki akurasi tinggi pada diagnosis penyakit berbasis data medis [17]. Tahapan penelitian dilakukan secara sistematis yang mencakup pengumpulan data, pra-pemrosesan data (penanganan *missing value*, *One-*

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9392
Hal 995-1005
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Hot Encoding, standarisasi fitur, dan analisis korelasi), penyeimbangan kelas menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE), pelatihan model menggunakan *pipeline* Scikit-learn dengan optimasi parameter melalui *GridSearchCV*, serta evaluasi performa model menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan ROC-AUC.Kombinasi ketiga algoritma ini memberikan peluang untuk menilai model dengan karakteristik berbeda secara menyeluruh, baik dari sisi kestabilan, akurasi, maupun kemampuan generalisasi. Kombinasi ketiga algoritma ini memberikan peluang untuk menilai model dengan karakteristik berbeda secara menyeluruh, baik dari sisi kestabilan, akurasi, maupun kemampuan generalisasi. Berikut alur studi yang dilaksanakan oleh peneliti tertera dalam Gambar 1.



Gambar 1. Alur Penelitian

2.2 Pengumpulan Data

Untuk memahami struktur data yang digunakan dalam penelitian ini, dataset yang digunakan dalam penelitian ini adalah Diabetes Prediction Dataset dengan format file diabetes_prediction_dataset.csv yang diperoleh dari repositori publik Kaggle dan telah digunakan secara luas dalam penelitian klasifikasi penyakit kronis [21]. Dataset ini terdiri dari 100000 baris data dan 9 atribut utama, yaitu Gender, Age, Hypertension, Heart Disease, Smoking History, BMI, HbA1c Level, Blood Glucose Level, dan Diabetes sebagai label target. Variabel Diabetes dikategorikan menjadi dua kelas, yaitu non-diabetes (0) dan diabetes (1), sesuai paradigma diagnosis awal pada studi medis [25]. Pembagian data dilakukan dengan proporsi 80% untuk data latih dan 20% untuk data uji, mempergunakan teknik stratified train-test split untuk menjaga distribusi kelas tetap seimbang [12]. Rincian atribut yang digunakan dalam penelitian ini disajikan secara sistematis pada Tabel 1.

Tabel 1. Atribut Dataset Penyakit Diabetes

No	Atribut	Tipe Data	Deskripsi
1	Gender	Categorical (0/1)	Categorical feature indicating gender (0 = Female, 1 = Male).
2	Age	Numerical (years)	Numerical feature representing age of the patient.
3	Hypertension	Numerical (mmHg)	Feature related to patient's health condition.
4	Heart Disease	Numerical (bpm)	Feature related to patient's health condition.
5	Smoking History	Object (class)	Categorical feature indicating smoking history.
6	BMI	Numerical (kg/m²)	Numrical feature representing body mass index (BMI).
7	HbA1c Level	Numerical (percentage %)	Numerical feature representing HbA1c Level.
8	Blood Glucose Level	Numerical (mg/dL)	Numerical feature representing Blood Glucose Level.
9	Diabetes	Categorical (0/1)	Target variable (1 = Diabetes, 0 = No Diabetes)

2.3 Pra Pemprosesan Data

Tahap pra-pemrosesan data dilakukan menggunakan pipeline Scikit-learn agar seluruh tahapan transformasi dijalankan secara otomatis dan berurutan tanpa kebocoran data (data leakage) [13]. Pipeline memastikan bahwa semua transformasi hanya dipelajari dari data latih dan diterapkan pada data uji, dengan demikian hasil evaluasi tetap objektif.

a. Missing Value Imputation

Tidak jarang terdapat data yang tidak lengkap untuk mengatasi data kosong yang berpotensi menurunkan kualitas pelatihan model. Fitur numerik diisi menggunakan rata-rata (mean) dengan SimpleImputer(*strategy='mean'*), sedangkan fitur kategorikal seperti Gender dan Smoking History diisi dengan modus *(most frequent)* menggunakan *SimpleImputer(strategy='most_frequent')* [13]. Pendekatan ini menjaga variasi data tanpa perlu menghapus sampel yang berisi nilai kosong.

b. One-Hot Encoding

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9392
Hal 995-1005
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Data kategorikal diubah menjadi bentuk numerik menggunakan One-Hot Encoding dengan parameter *handle_unknown='ignore'* agar semua kategori dapat diproses tanpa menimbulkan error saat ditemukan label baru [19]. Teknik ini menghasilkan representasi biner (0 dan 1) untuk setiap kategori sehingga model dapat mengenali pola antarvariabel secara lebih akurat.

c. Standarisasi Data

Seluruh fitur numerik kemudian distandarisasi dilakukan dengan *StandardScaler()* agar mempunyai range nol dan deviasi standar satu. Standarisasi ini penting karena beberapa model, seperti Logistic Regression dan XGBoost, sensitif terhadap perbedaan skala antarvariabel. Proses standarisasi mengikuti rumus:

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

dengan x adalah nilai asli fitur, μ merupakan rata-rata, dan σ adalah deviasi standar [17].

d. Analisis Korelasi

Dilakukan juga analisis korelasi antarvariabel numerik menggunakan heatmap correlation matrix untuk mengidentifikasi adanya multikolinearitas, terutama pada fitur yang memiliki hubungan kuat seperti Blood Glucose Level dan HbA1c Level. Fitur dengan korelasi tinggi (r > 0.8) diperhatikan agar tidak memengaruhi stabilitas model linier seperti Logistic Regression [13]. Seluruh tahapan pra-pemrosesan dilakukan menggunakan pipeline Scikit-learn agar transformasi diterapkan secara otomatis dan konsisten, hanya dipelajari dari data latih untuk mencegah data leakage.

2.4 Oversampling Dengan SMOTE

Dataset diabetes cenderung mempunyai distribusi kelas yang tidak seimbang antara penderita diabetes dan non-diabetes. Untuk mengatasi hal ini, maka dari itu diterapkan teknik *SMOTE* pada data latih [24]. Dari data hasil penyeimbangan 80% untuk data latih, dan 20% untuk data uji. Hasil SMOTE total data berjumlah 146400 baris, dengan kelas 0 (normal) berjumlah 73200 baris, kelas 1 (sakit) berjumlah 73200 baris.

2.5 Pelatihan Model

Tahap ini bertujuan melatih model pembelajaran mesin serta mencari parameter terbaik yang memberikan hasil performa optimal. Proses pelatihan dilakukan dengan tiga algoritma utama.

a. Logistic Regression

Logistic Regression adalah model linier untuk klasifikasi biner yang memprediksi probabilitas kelas positif (1) berdasarkan hubungan dengan variabel independen. Algoritma ini banyak digunakan di bidang medis karena interpretatif, stabil, dan efisien secara komputasi [13], [15]. Persamaan dasarnya dituliskan sebagai:

$$P(y = 1 \mid x) = \frac{1}{1+e-(\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn)} \quad (2)$$

dengan $P(y =1| x)$ adalah probabilitas seseorang menderita diabetes, β0 konstanta (intercept), βi koefisien tiap fitur xi, e bilangan eksponensial alami.

b. Random Forest

Random Forest adalah algoritma *ensemble* berbasis *bagging* yang mengombinasikan banyak pohon keputusan independen untuk meningkatkan akurasi dan mengurangi *overfitting*. Prediksi akhir ditentukan melalui mekanisme *majority voting*, dan metode ini efektif dalam klasifikasi diabetes karena mampu menangani hubungan nonlinier serta fitur yang kompleks [11], [14]. Prediksi akhir ditentukan sebagai berikut**:**

$$y^{\wedge} = mode\{h1(x), h2(x), \ldots, hB(x)\} \quad (3)$$

dengan $h_b\ x$ ialah prediksi dari pohon ke-b dan B banyaknya total pohon.

c. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) adalah pengembangan dari Gradient Boosting yang bekerja secara iteratif dengan mengoptimalkan kesalahan prediksi sebelumnya menggunakan pendekatan berbasis gradien. Algoritma ini dikenal efisien secara komputasi dan memiliki akurasi tinggi dalam klasifikasi data klinis [17], [18]. Fungsi objektif XGBoost dirumuskan sebagai:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \quad (4)$$

dengan:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (5)$$

di mana l adalah fungsi kehilangan (loss function), Ω(f) regularisasi untuk mengontrol kompleksitas model, T jumlah leaf, dan w_j bobot setiap leaf [17], [18]. Pencarian parameter terbaik dilaksanakan dengan GridSearchCV dengan

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9392
Hal 995-1005
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Stratified 5-Fold Cross Validation dan metrik optimasi ROC-AUC. Model dengan skor tertinggi di pilih sebagai model terbaik untuk tahap evaluasi [25].

2.6 Evaluasi Model

Tahap evaluasi model memiliki tujuan dalam rangka mengevaluasi seberapa jauh kinerja model pembelajaran mesin mampu melakukan klasifikasi diabetes secara akurat dan konsisten berdasarkan data uji. Evaluasi terhadap data uji menggunakan lima metrik utama, yaitu *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan *ROC-AUC* [12], [24]. Mencakup perhitungan *TP*, *TN, FP*, dan *FN* guna mendapatkan nilai akurasi, yang mendukung analisis seberapa jauh metode memprediksi dengan cara akurat dan mengidentifikasi kesalahan klasifikasi.

a. Accuracy menunjukan tingkat ketetapan keseluruhan model dan dihitung terhadap total data:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

b. Precison mengukur tingkat keakuratan prediksi positif:

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

c. Recall mengindikasikan kemampuan model untuk mendeteksi seluruh kasus positif:

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

d. F1-Score sebagai rata-rata harmonis antara Precision dan Recall:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (9)$$

e. *ROC-AUC* dipergunakan dalam memperhitungkan kemampuan model dalam membedakan antara dua kelas positif dan negatif. Kurva ROC merepresentasikan hubungan antara *TPR* dan *FPR* :

$$\text{TPR} = \frac{\text{TP}}{\text{TP+ FN}} \text{ dengan FPR} = \frac{FP}{FP+TN} \quad (10)$$

f. *AUC* (*Area Under Curve*) merepresentasikan luas area di bawah kurva *ROC*, yang dihitung sebagai integral antara TPR dan FPR:

$$\text{AUC} = \int_0^1 \text{TPR(FPR)d(FPR)} \quad (11)$$

Nilai AUC = 1 menunjukkan model sempurna, Nilai AUC = 0.5 berarti model tidak lebih baik dari tebakan acak.

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Dataset yang dipergunakan dalam penelitian ini ialah *Diabetes Prediction Dataset* dari platform publik Kaggle pada tahun 2023, yang dapat diakses melalui: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset. Dataset ini berisi 100.000 baris data dengan 9 atribut utama yang digunakan untuk memprediksi kemungkinan seseorang menderita diabetes. Keberagaman atribut dalam Tabel 2 memungkinkan model klasifikasi untuk mempelajari pola hubungan antara faktor gaya hidup, kondisi medis, dan indikator biokimia terhadap risiko diabetes. Selain itu, penggunaan variabel numerik dan kategorikal secara bersamaan mencerminkan kondisi data medis nyata yang umum dijumpai dalam sistem pendukung keputusan kesehatan. Berikut perincian dataset dalam tabel mengenai struktur dan karakteristik data yang digunakan, perincian atribut serta contoh nilai pada dataset disajikan pada Tabel 2.

Tabel 2. Sampel Dataset Penyakit Diabetes

No	Gender	Age	Hypertension	Heart Disease	Smoking History	BMI	HbA1c Level	Blood Glucose Level	Diabetes
1	Female	80.0	0	1	never	25.19	6.6	140	0
2	Female	54.0	0	0	No info	27.32	6.6	80	0
3	Male	28.0	0	0	never	27.32	5.7	158	0
4	Female	36.0	0	0	current	23.45	5.0	155	0
…	…	…	…	…	…	…	…	…	…
99997	Female	2.0	0	0	No info	17.37	6.5	100	0
99998	Male	66.0	0	0	former	27.83	5.7	155	0
99999	Female	24.0	0	0	never	35.42	4.0	100	0
100000	Female	57.0	0	0	current	22.43	6.6	90	0

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9392
Hal 995-1005
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

3.2 Pra Pemprosesan

Langkah berikutnya adalah pra-pemrosesan data dilakukan untuk memastikan bahwa data bersih, konsisten, dan dapat diolah secara optimal dalam proses permodelan. Tahapan yang dilakukan meliputi:

a. Missing Value Imputation

Dalam tahap ini untuk mengatasi data kosong pada kolom tertentu. Fitur numerik diisi menggunakan nilai rata-rata (mean imputation*),* sedangkan fitur kategorikal seperti Smoking History dan Gender diisi dengan nilai modus (most frequent strategy). Hasil dari missing value imputation dapat diliat pada Tabel 3, ini memastikan tidak ada data yang hilang sehingga model dapat memanfaatkan seluruh informasi yang tersedia.

Tabel 3. Hasil Missing Value Imputation

No	Gender	Age	Hypertension	Heart Disease	Smoking History	BMI	HbA1c Level	Blood Glucose Level	Diabetes
1	Female	80.0	0	1	never	25.19	6.6	140	0
2	Female	54.0	0	0	No info	27.32	6.6	80	0
3	Male	28.0	0	0	never	27.32	5.7	158	0
4	Female	36.0	0	0	current	23.45	5.0	155	0
5	Male	76.0	1	1	current	20.14	4.8	155	0

Tabel 3 di atas menunjukan hasil proses Missing Value Imputation, Imputasi menggunakan rata-rata untuk fitur numerik dan modus untuk fitur kategorikal, sehingga pola distribusi data tetap terjaga. Hal ini terlihat pada atribut smoking_history yang secara konsisten terisi dengan kategori paling umum (“No Info”) serta nilai numerik seperti age*, BMI*, dan *HbA1c_level* yang tetap berada dalam rentang normal. Dengan demikian, proses imputasi berhasil meningkatkan kualitas data tanpa mengubah karakteristik utama setiap variabel.

b. One-Hot Encoding

Untuk variabel kategorikal, dilakukan *One-Hot Encoding*. Fitur Gender yang awalnya terdiri dari dua kategori (“Male” dan “Female”) diubah menjadi dua kolom biner, yaitu Gender Female dan Gender Male. Dengan demikian, model dapat memahami perbedaan nilai kategorikal tanpa menimbulkan bias numerik. Fitur Smoking History yang memiliki lebih banyak kategori seperti “*never*”, “*current*”, “*former*”, “*not current*”, dan “*ever*” juga dikonversi menjadi kolom biner sesuai masing-masing kategori. Hasil transformasi tersebut tertera dalam Tabel 4, yang menyajikan hasil representasi data setelah proses encoding dilakukan.

Tabel 4. Hasil One-Hot Encoding Pada Fitur Kategorikal

No	Age	Hypertension	Heart Disease	BMI	HbA1c Level	Blood Glucose Level	Gender Female	Gender Male	Gender Other	Smoking History No Info	Smoking History Current	Smoking History Ever	Smoking History Former	Smoking History Never	Smoking History Not Current
1	80	1.0	0.0	27.32	6.5	145.	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
2	19	0.0	0.0	25.18	4.5	126	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0

Dari hasil One-Hot Encoding pada Tabel 4 di atas, diketahui bahwa variabel Gender dan Smoking History berhasil dikonversi menjadi bentuk numerik biner. Variabel Gender diubah menjadi tiga kolom, yaitu Gender Female, Gender Male, dan Gender Other, sedangkan Smoking History diubah menjadi enam kolom, seperti Smoking History Never dan Smoking History Former. Nilai 1 menunjukkan kategori yang dimiliki pasien, sedangkan nilai 0 menandakan tidak termasuk kategori tersebut. Proses ini bertujuan agar model dapat mengolah data kategorikal secara numerik tanpa kehilangan informasi penting dari tiap atribut

c. Standarisasi Data

Pada proses standarisasi menggunakan StandardScaler agar seluruh fitur numerik mempunyai rata-rata 0 dan deviasi standar 1. Proses ini esensial untuk model seperti Logistic Regression dan XGBoost yang sensitif terhadap perbedaan skala fitur. Berikut Tabel 5 adalah perbandingan data standardscaler.

Tabel 5. Data Perbandingan Sebelum dan Setelah Standard Scaler

Atribut	Nilai Sebelum Standard Scaler	Nilai Sesudah Standard Scaler
Age	80.00	1.691070
Hypertension	1.00	3.516004
Heart Disease	0.00	-0.202792

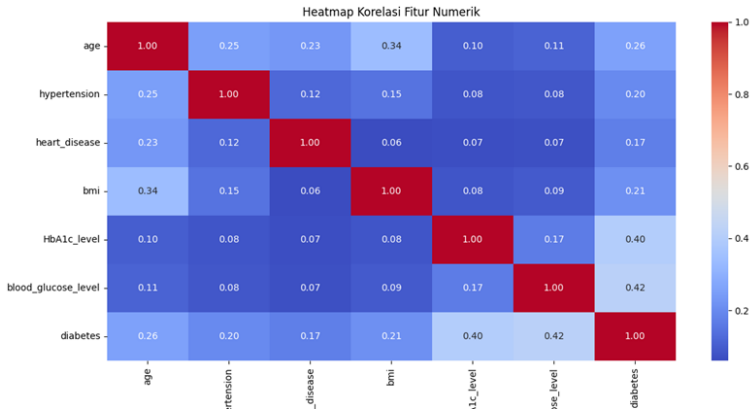
JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9392
Hal 995-1005
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

BMI	27.32	0.000472
HbA1c Level	6.50	0.908249
Blood Glucose Level	145.00	0.170207

Dari Tabel 5 menunjukkan bahwa proses Standard Scaler berhasil menyeragamkan skala pada setiap fitur numerik. Nilai setiap atribut berubah menjadi rentang dengan range mendekati nol dan standar deviasi mendekati satu. Sebagai contoh, atribut Age yang semula bernilai 80 berubah menjadi 1,69 dan BMI dari 27,32 menjadi 0,00047 setelah distandarisasi. Proses ini memastikan bahwa semua fitur mempunyao bobot yang seimbang sehingga model tidak bias terhadap variabel dengan skala besar seperti *Blood Glucose Level*.

d. Analisis Korelasi

Selanjutnya, untuk mengetahui tingkat hubungan antarfitur serta kontribusinya terhadap variabel target dilakukan analisis korelasi. Analisis ini bertujuan untuk mengidentifikasi fitur-fitur yang memiliki keterkaitan paling signifikan terhadap penyakit diabetes. Berikut hasil visualisasi Heatmap Korelasi Fitur Numerik yang menunjukkan kekuatan hubungan antaratribut berdasarkan nilai koefisien korelasi Pearson pada Gambar 2.

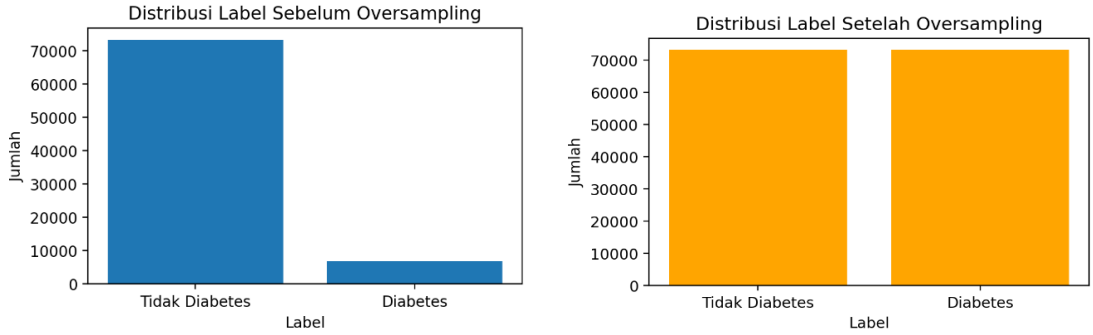


Gambar 2. Korelasi Terhadap Diabetes

Berdasarkan hasil visualisasi pada Gambar 2 heatmap, terlihat bahwa fitur *HbA1c Level* dan *Blood Glucose Level* memiliki korelasi paling kuat terhadap variabel target Diabetes, dengan nilai korelasi masing-masing sejumlah 0,40 dan 0,42. Hal ini mengindikasikan bahwasanya peningkatan kadar glukosa darah dan nilai *HbA1c* cenderung berkaitan langsung dengan kemungkinan seseorang menderita diabetes. Sementara itu, atribut lain seperti Age dan Hypertension memiliki korelasi sedang terhadap Diabetes, sedangkan Heart Disease dan BMI menunjukkan hubungan yang relatif rendah.

e. Balancing Data

Pada tahap ini mengimplementasikan metode SMOTE untuk menyeimbangkan distribusi kelas pada dataset. Proses ini bertujuan mengurangi bias model pada kelas mayoritas dan menaikkan tingkat kemampuan pada pengenalan pola pada kelas minoritas. Sehingga, model dapat melakukan klasifikasi secara lebih objektif dan akurat terhadap kasus diabetes. Perbandingan hasil penyeimbangan kelas tertera dalam dalam Gambar 3 berikut.



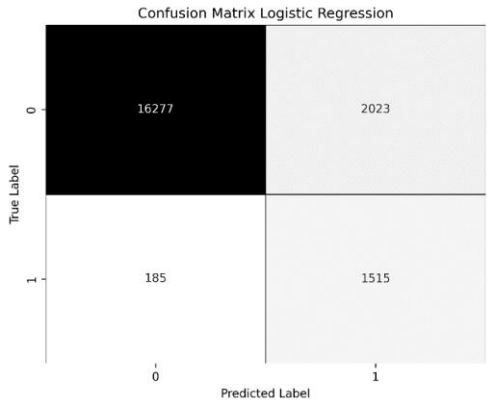
Gambar 3. Distribusi Label Sebelum dan Sesudah Oversampling

Gambar 3 memperlihatkan bahwa distribusi kelas awal tidak seimbang, di mana jumlah pasien non-diabetes jauh lebih besar dibanding dengan pasien diabetes. Setelah penerapan SMOTE, jumlah sampel pada kedua kelas menjadi seimbang, masing-masing sekitar 73.000 data. Penyeimbangan ini penting untuk mencegah bias model terhadap kelas mayoritas serta meningkatkan kemampuan algoritma dalam mengenali pola pada kelas minoritas, sehingga model lebih akurat dalam mendeteksi pasien diabetes.

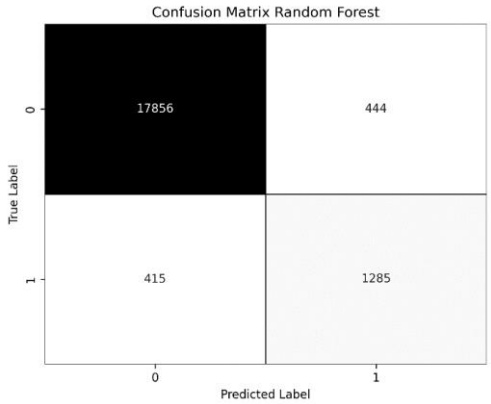
JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9392
Hal 995-1005
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

f. Evaluasi Hasil

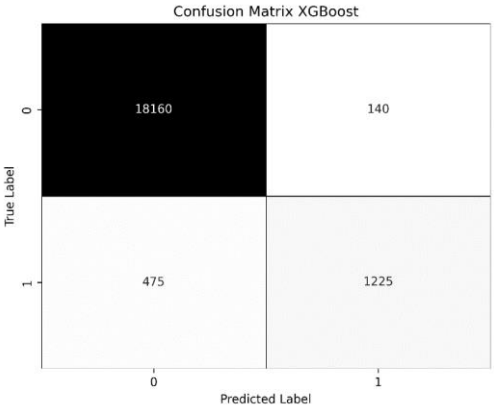
Evaluasi model dilakukan setelah seluruh tahapan pra-pemrosesan data diselesaikan, meliputi imputasi nilai hilang, transformasi fitur kategorikal, standarisasi, serta penyeimbangan kelas.. Studi ini menggunakan tiga algoritma pembelajaran mesin, yaitu *Logistic Regression*, *Random Forest*, dan *XGBoost*. Ketiga model ini dipilih karena memiliki pendekatan berbeda linear, bagging, dan boosting sehingga dapat memberikan gambaran komparatif yang jelas. Untuk menghindari bias hasil pelatihan dan meningkatkan generalisasi model, dilakukan optimasi hiperparameter menggunakan GridSearchCV dengan skema Stratified 5-Fold Cross Validation, sehingga proporsi kelas pada setiap lipatan data tetap terjaga. Teknik ini memungkinkan model diuji secara proporsional pada berbagai subset data, memastikan performa yang stabil dan tidak bergantung pada pembagian data tertentu. Confusion Matrix digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan data secara lebih rinci melalui analisis True Positive, True Negative, False Positive, dan False Negative. Berikut adalah hasil seperti pada Gambar 4, Gambar 5, dan Gambar 6 berikut.



Gambar 4. Confusion Matrix Logistic Regression



Gambar 5. Confusion Matrix Random Forest



Gambar 6. Confusion Matrix XGBoost

Berdasarkan hasil visualisasi Gambar 4 model Logistic Regression menunjukkan jumlah *True Positive* yang cukup tinggi namun masih menghasilkan *False Positive* yang relatif besar, menandakan bahwa model ini cenderung memprediksi kelas positif secara berlebihan. Gambar 5 Model Random Forest memperlihatkan keseimbangan yang lebih baik antara True Positive dan *True Negative*, dengan jusmlah kesalahan klasifikasi yang lebih sedikit. Sementara itu,

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9392
Hal 995-1005
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Gambar 6 model XGBoost menghasilkan distribusi prediksi paling akurat dengan proporsi kesalahan yang sangat rendah, menandakan kemampuannya dalam mengenali pola data yang kompleks.

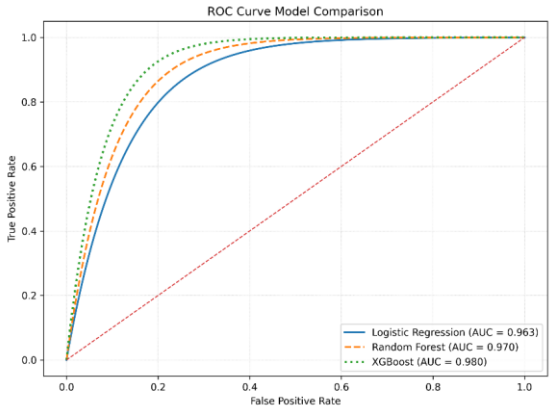
Setelah dilakukan proses validasi dan pengujian, hasil metrik performa dari ketiga model ditampilkan pada Tabel 6 dibawah ini.

Tabel 6. Hasil Evaluasi Model

Algoritma	Accuracy	Precision	Recall	F1-Score
XGBoost	96.88%	89.22%	72.06%	79.73%
Random Forest	95.68%	74.35%	75.18%	74.76%
Logistic Regression	88.96%	42.82%	89.12%	57.85%

Berdasarkan hasil evaluasi yang disajikan dalam Tabel 6, perbandingan kinerja dilakukan menggunakan lima metrik evaluasi utama, yaitu accuracy, precision, recall, F1-score, dan ROC-AUC, guna memperoleh gambaran performa model secara menyeluruh. Model Logistic Regression menunjukkan akurasi sebesar 88,96% dengan nilai recall tertinggi (89,12%), menandakan kemampuan yang baik dalam mengenali pasien diabetes. Namun, precision yang rendah (42,82%) mengindikasikan banyaknya prediksi positif yang keliru, sehingga nilai F1-score (57,85%) menunjukkan ketidakseimbangan antara ketepatan dan sensitivitas. Hal ini berkaitan dengan karakteristik Logistic Regression sebagai model linier yang mengasumsikan adanya hubungan linier antar fitur, sehingga membatasi kemampuannya dalam memodelkan pola nonlinier yang lazim ditemukan pada data medis multivariat. Berbeda dengan itu, Random Forest memberikan hasil yang lebih stabil dengan akurasi 95,68%, precision 74,35%, dan recall 75,18%, mencerminkan keseimbangan yang baik antara kemampuan deteksi dan pengurangan kesalahan klasifikasi. Keunggulan ini diperoleh dari mekanisme *bagging* yang mampu menurunkan variansi dan meningkatkan keandalan prediksi, di mana banyak pohon keputusan dilatih pada subset data yang berbeda dan hasil prediksi ditentukan melalui voting mayoritas, sehingga mampu menurunkan variansi model dan meningkatkan keandalan prediksi pada data yang kompleks. Sementara itu, XGBoost menjadi model dengan performa terbaik, mencapai akurasi 96,88%, precision 89,22%, recall 72,06%, dan F1-score 79,73%, menunjukkan keseimbangan optimal antara presisi dan stabilitas berkat teknik *boosting* bertahap yang secara iteratif mengoreksi kesalahan model sebelumnya melalui proses optimasi berbasis *gradient descent* dan regularisasi, sehingga memungkinkan pemodelan hubungan nonlinier antar fitur secara lebih komprehensif yang mencegah *overfitting*. Dengan demikian, perbedaan kinerja antar model tidak semata-mata ditentukan oleh nilai akurasi akhir, melainkan juga oleh strategi pembelajaran masing-masing algoritma dalam menangkap pola data dan mengelola kompleksitas fitur klinis. Secara keseluruhan, seluruh model menunjukkan performa baik dengan akurasi di atas 85%, namun XGBoost menempati posisi tertinggi, diikuti oleh Random Forest yang konsisten, dan Logistic Regression yang tetap unggul dalam sensitivitas deteksi awal diabetes.

Analisis lanjutan dilakukan menggunakan Receiver Operating Characteristic (ROC) Curve untuk menilai kemampuan diskriminatif model dalam membedakan kelas positif (penderita diabetes) dan kelas negatif (non-diabetes). Kurva ROC menunjukkan hubungan antara TPR dan FPR pada berbagai ambang batas probabilitas. Nilai AUC dipergunakan dalam memperhitungkan luas area di bawah kurva, di mana semakin besar nilainya, semakin baik kapabilitas model dalam membedakan dua kelas. Hasil pengujian tersebut disajikan pada Gambar 7.



Gambar 7. Kurva ROC Curve

Berdasarkan hasil pengujian yang diindikasikan oleh Gambar 7 ketiga model menghasilkan nilai *ROC-AUC* yang sangat tinggi, seluruhnya melebihi 96%, mengindikasikan kapabilitas diskriminatif yang sangat baik dalam membedakan pasien diabetes dan non-diabetes. Model *XGBoost* mencatat nilai tertinggi sebesar 98,00%, menunjukkan performa paling stabil dan efisien dalam klasifikasi, diikuti oleh *Random Forest* dengan 97,00% dan *Logistic Regression* dengan 96,30%. Visualisasi kurva *ROC* memperlihatkan bahwa ketiga model berada sangat dekat dengan sudut kiri atas grafik, menandakan *TPR* yang tinggi serta *FPR* yang rendah, sehingga model mampu mengidentifikasi pasien diabetes dengan kesalahan minimal. Nilai *AUC* yang tinggi pada seluruh model menunjukkan konsistensi performa klasifikasi, namun

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9392
Hal 995-1005
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

XGBoost tetap menjadi model paling unggul karena mekanisme boosting adaptif-nya yang mampu memperbaiki kesalahan secara iteratif dan mengoptimalkan hasil prediksi tanpa mengorbankan akurasi.

3.3 Pembahasan

Penelitian ini membandingkan tiga algoritma pembelajaran mesin, yaitu *Logistic Regression*, *Random Forest*, dan *XGBoost* dalam mengklasifikasikan penyakit diabetes berdasarkan variabel klinis seperti *HbA1c Level, Blood Glucose Level*, *BMI*, dan usia. Ketiga algoritma dipilih karena mewakili pendekatan berbeda, dengan Logistic Regression sebagai model linier, Random Forest sebagai *ensemble* berbasis *bagging*, dan XGBoost sebagai *boosting*. Hasil pengujian menunjukkan bahwa ketiganya memiliki kinerja baik dengan tingkat performa yang berbeda. XGBoost mencatat hasil terbaik dengan akurasi 96,88% dan *AUC* 98,00%, menunjukkan kemampuan tinggi dalam membedakan pasien diabetes dan non-diabetes. *Random Forest* menempati urutan kedua dengan akurasi 95,68% dan *F1-score* 74,76%, menandakan keseimbangan antara *precision* dan *recall*. *Logistic Regression* memiliki akurasi 88,96% dengan recall tertinggi (89,12%), yang menunjukkan sensitivitas tinggi terhadap kasus positif meski dengan *precision* yang rendah. Performa *Logistic Regression* yang lebih rendah disebabkan oleh keterbatasannya dalam menangkap hubungan non-linear antar fitur medis, meskipun tetap unggul dalam interpretasi hasil.

Tahapan pra-pemrosesan data berperan penting dalam peningkatan performa model. Proses imputasi nilai hilang, standarisasi, dan *One-Hot Encoding* memastikan skala data seragam sebelum pelatihan model dilakukan. Teknik (*SMOTE*) diterapkan untuk mengatasi ketidakseimbangan kelas antara pasien diabetes dan non-diabetes, sehingga model lebih peka terhadap kelas minoritas. Penerapan teknik ini terbukti meningkatkan nilai recall pada *XGBoost* dan *Random Forest* tanpa menurunkan akurasi secara signifikan. Selain itu, penggunaan *GridSearchCV* membantu menemukan kombinasi parameter terbaik melalui proses validasi silang, yang meningkatkan stabilitas dan generalisasi model terhadap data baru. Visualisasi kurva *ROC* menunjukkan nilai *AUC* di atas 96% untuk ketiga model, dengan posisi kurva yang mendekati sudut kiri atas, menggambarkan tingkat (*TPR*) yang tinggi dan (*FPR)* yang rendah.

Hasil studi ini linear dengan studi dari Siregar et al. [10] dan Setyawan & Wakhidah [15], yang menunjukkan bahwa model ensemble seperti *XGBoost* dan *Random Forest* lebih unggul dibandingkan model linier dalam klasifikasi penyakit medis. Pendekatan ensemble terbukti efektif dalam menangkap interaksi non-linear antar variabel klinis seperti kadar glukosa dan *BMI* terhadap risiko diabetes. Meskipun demikian, penelitian ini memiliki keterbatasan pada variasi variabel yang masih terbatas pada data klinis dasar. Hasil penelitian ini juga memiliki nilai aplikatif karena model *XGBoost* dapat digunakan sebagai sistem pendukung keputusan medis berbasis data, membantu tenaga medis dalam melaksanakan deteksi dini risiko diabetes dengan cara akurat dan efisien. Sementara itu, *Random Forest* dapat menjadi alternatif yang stabil dan mudah diimplementasikan, dan *Logistic Regression* tetap relevan untuk interpretasi hubungan antar variabel yang penting dalam konteks edukasi kesehatan dan penelitian klinis.

4. KESIMPULAN

Dalam studi ini, tiga algoritma pembelajaran mesin dipergunakan dalam mengklasifikasikan penyakit diabetes berdasarkan data klinis, yaitu *Logistic Regression*, *Random Forest*, dan *XGBoost*. Ketiga model mengindikasikan performa prediksi yang baik, dengan tingkat akurasi keseluruhan berada di atas 85%. Proses optimasi hyperparameter menggunakan *GridSearchCV* terbukti berperan penting dalam meningkatkan kinerja setiap algoritma, terutama pada model berbasis ensemble yang secara umum lebih adaptif terhadap variasi pola data. *XGBoost* menjadi model paling unggul dengan akurasi 96,88%, nilai *AUC* 98,00%, dan F1-score tertinggi sebesar 79.73% yang menunjukkan kemampuannya dalam mengenali hubungan kompleks antarfitur dan menghasilkan prediksi yang stabil. *Random Forest* berada pada posisi berikutnya dengan akurasi 95,68% serta keseimbangan yang baik antara *precision* dan *recall*, sehingga dapat menjadi alternatif andal untuk penggunaan praktis. Sementara itu, Logistic Regression memiliki akurasi lebih rendah, namun tetap memberikan interpretasi yang jelas mengenai kontribusi masing-masing variabel klinis, seperti kadar glukosa, *HbA1c*, *BMI*, dan usia terhadap risiko diabetes. Temuan ini menegaskan bahwa model berbasis ensemble cenderung lebih unggul dalam melakukan klasifikasi medis berbasis data dibandingkan model linier. Selain itu, hasil penelitian diharapkan dapat mendukung pengembangan sistem penunjang keputusan medis dengan basis kecerdasan buatan untuk mendukung tenaga medis dalam melakukan deteksi dini risiko diabetes secara lebih akurat, cepat, dan efisien. Dengan demikian, penerapan algoritma pembelajaran mesin memiliki potensi yang signifikan dalam mendukung deteksi dini risiko diabetes. Pengembangan sistem penunjang keputusan medis berbasis kecerdasan buatan diharapkan dapat membantu tenaga kesehatan dalam pengambilan keputusan klinis secara lebih akurat dan efisien, serta menjadi dasar bagi penelitian lanjutan dengan integrasi variabel klinis yang lebih luas dan pendekatan model yang lebih adaptif guna meningkatkan kualitas pelayanan kesehatan masyarakat.

REFERENCES

[1] D. Care and S. S. Suppl, “Introduction and Methodology: Standards of Care in Diabetes—2024,” *Diabetes Care*, vol. 47, no. December 2024, pp. S1–S4, 2024, doi: 10.2337/dc24-SINT.

[2] World Health Organisation, “Quick facts What are the causes/risk factors for diabetes? What are the symptoms of diabetes?,” 2023.

[3] Magliano DJ and Boyko EJ, *IDF DIABETES ATLAS [Internet]. 10th edition*. 2021. [Online]. Available:

https://www.ncbi.nlm.nih.gov/books/NBK581938/

[4] H. Sun *et al.*, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 183, pp. 1–23, 2022, doi: 10.1016/j.diabres.2021.109119.

[5] K. L. Ong *et al.*, "Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021," *Lancet*, vol. 402, no. 10397, pp. 203–234, 2023, doi: 10.1016/S0140-6736(23)01301-6.

[6] V. Kontis, J. Bentham, C. D. Mathers, J. Rehm, M. Ezzati, and NCD Risk Factor Collaboration, "Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants," *Lancet*, vol. 404, no. 10467, pp. 2077–2093, 2024, doi: 10.1016/S0140-6736(24)02317-1.Worldwide.

[7] S. K. Indonesia, "Dalam angka," 2023.

[8] *IDF Diabetes Atlas*. 2025.

[9] M. Tarigan and E. R. Megawati, "Prediabetes , undiagnosed diabetes , and associated factors in North Sumatra , Indonesia : A community-based study," vol. 31, no. 4, pp. 420–427, 2024.

[10] F. A. Siregar, T. Makmur, and R. Bestari, "Identifying Adult Population at Risk for Undiagnosed Diabetes Mellitus in Medan City , Indonesia Targeted on Diabetes Prevention," vol. 77, no. 6, pp. 455–459, 2023, doi: 10.5455/medarh.2023.77.455-459.

[11] D. C. P. Buani, "Deteksi Dini Penyakit Diabetes dengan Menggunakan Algoritma Random Forest," *EVOLUSI J. Sains dan Manaj.*, vol. 12, no. 1, pp. 1–8, 2024, doi: 10.31294/evolusi.v12i1.21005.

[12] R. Hidayat, D. Mahdiana, and A. Fergina, "Comparative Analysis of Logistic Regression, SVM, Xgboost, and Random Forest Algorithms for Diabetes Classification," *J. Teknol. Sist. Inf. dan Apl.*, vol. 7, no. 1, pp. 281–291, 2024, doi: 10.32493/jtsi.v7i1.38258.

[13] F. Kurniawati and D. Brahma Arianto, "Analisis Implementasi Seleksi Fitur Pada Klasifikasi Diabetes dengan Metode Corellation Matrix dan Algoritma Logistic Regression," *Inform. J. Ilmu Komput.*, vol. 19, no. 3, pp. 157–164, 2023, doi: 10.52958/iftk.v19i3.6019.

[14] N. Nurussakinah, M. Faisal, and I. B. Santoso, "Algoritma Random Forest dan Synthetic Minority Oversampling Technique (SMOTE) untuk Deteksi Diabetes," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 10, no. 2, pp. 221–234, 2025, doi: 10.14421/jiska.2025.10.2.221-234.

[15] N. H. Setyawan and N. Wakhidah, "Analisis Perbandingan Metode Logistic Regression, Random Forest, Gradient Boosting Untuk Prediksi Diabetes," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 10, no. 1, pp. 150–162, 2025, doi: 10.29100/jipi.v10i1.5743.

[16] 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, 2023, doi: 10.46984/sebatik.v27i1.2157.

[17] K. D. K. Wardhani and M. Akbar, "Diabetes Risk Prediction Using Extreme Gradient Boosting (XGBoost)," *J. Online Inform.*, vol. 7, no. 2, pp. 244–250, 2022, doi: 10.15575/join.v7i2.970.

[18] E. R. Susanto and A. Cahyana, "Penerapan Algoritma XGBoost untuk Prediksi Diabetes: Analisis Confusion Matrix dan ROC Curve," *Fountain Informatics J.*, vol. 10, no. 1, pp. 40–50, 2025, doi: 10.21111/fij.v10i1.14311.

[19] M. Salsabil, N. Lutvi, and A. Eviyanti, "Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost," *J. Ilm. Komputasi*, vol. 23, no. 1, pp. 51–58, 2024, doi: 10.32409/jikstik.23.1.3507.

[20] M. Hakeel, "Diabetes Prediction Machine Learning-Based Diabetes Prediction App Using Random Forest Algorithm," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 1, pp. 1370–1376, 2025, doi: 10.36040/jati.v9i1.12654.

[21] R. Sugavanam, S. S. Muralikumar, V. R. Chandran, and S. K. Ponnusamy, "Diabetes Prediction Using Machine Learning," *AIP Conf. Proc.*, vol. 3279, no. 1, pp. 639–649, 2025, doi: 10.1063/5.0263265.

[22] G. Airlangga, "Comparative Analysis of Machine Learning Models for Predicting Diabetes: Unveiling the Superiority of Advanced Ensemble Methods," *G-Tech J. Teknol. Terap.*, vol. 8, no. 2, pp. 1272–1280, 2024, doi: 10.33379/gtech.v8i2.4246.

[23] H. Kurniawan, "Evaluasi Performa Random Forest, XGBoost, dan LightGBM dalam Diagnosis Dini Diabetes Mellitus," *J. JUPITER*, vol. 17, no. 2, pp. 835–844, 2025.

[24] L. Andrade-Arenas and C. Yactayo-Arias, "Comparative Evaluation of Machine Learning Models for Diabetes Prediction: A Focus on Ensemble Methods," *Ing. des Syst. d'Information*, vol. 30, no. 7, pp. 1795–1803, 2025, doi: 10.18280/isi.300712.

[25] N. Katiyar, H. K. Thakur, and A. Ghatak, "Recent advancements using machine learning & deep learning approaches for diabetes detection: a systematic review," *e-Prime - Adv. Electr. Eng. Electron. Energy*, vol. 9, no. May, p. 100661, 2024, doi: 10.1016/j.prime.2024.100661.