Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2679−2689
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9534

Evaluasi KNN dan Logistic Regression untuk Klasifikasi Diabetes dengan Preprocessing Terstandarisasi: Trade-off Kinerja dan Interpretabilitas

Alif Zayyin Kamandani, Egia Rosi Subhiyakto*

[1]Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1]111202214815@mhs.dinus.ac.id, [2,*]egia@dsn.dinus.ac.id
Email Penulis Korespondensi: egia@dsn.dinus.ac.id
Submitted: **14/03/2026**; Accepted: **31/03/2026**; Published**: 31/03/2026**

Abstrak−Meskipun *K-Nearest Neighbors* (KNN) dan *Logistic Regression* telah digunakan dalam klasifikasi diabetes, studi yang secara terstruktur mengombinasikan pipeline *preprocessing* terstandarisasi meliputi imputasi median, standarisasi fitur, dan *stratified data splitting* serta mengevaluasi *trade-off* antara kinerja prediksi dan interpretabilitas model masih terbatas. Penelitian ini bertujuan untuk membandingkan performa kedua algoritma dalam klasifikasi status diabetes menggunakan dataset Pima Indians Diabetes yang terdiri dari 768 sampel dengan delapan atribut numerik. Tahapan penelitian meliputi eksplorasi data, penanganan *missing value* dengan imputasi median, standarisasi fitur menggunakan *StandardScaler*, serta pembagian data secara *stratified* dengan rasio 80:20. Evaluasi model dilakukan menggunakan metrik akurasi, *precision*, *recall*, *F1-score*, *confusion matrix*, serta ROC-AUC. Hasil eksperimen menunjukkan bahwa KNN dengan parameter optimal K=21 mencapai akurasi sebesar 75,97%, *F1-Score* sebesar 61,86%, dan ROC-AUC sebesar 0,8120, sedangkan *Logistic Regression* memperoleh akurasi sebesar 70,78%, *F1-Score* sebesar 54,55%, dengan ROC-AUC sebesar 0,8130. Meskipun KNN menunjukkan kinerja prediksi yang lebih tinggi, *Logistic Regression* memberikan keunggulan dalam interpretabilitas melalui koefisien model, di mana variabel *Glucose* (β=1,1825) dan *BMI* (β=0,6887) teridentifikasi sebagai prediktor utama risiko diabetes. Temuan ini menunjukkan adanya *trade-off* yang jelas antara akurasi dan interpretabilitas, sehingga KNN lebih sesuai untuk kebutuhan prediksi dengan akurasi tinggi, sementara *Logistic Regression* lebih tepat digunakan dalam konteks klinis yang memerlukan transparansi dan akuntabilitas model.

Kata Kunci: Diabetes Melitus; K-Nearest Neighbors; Logistic Regression; Interpretabilitas Model; ROC-AUC

Abstract−Although K-Nearest Neighbors (KNN) and Logistic Regression have been widely used in diabetes classification, studies that systematically combine a standardized preprocessing pipeline including median imputation, feature standardization, and stratified data splitting and evaluate the trade-off between predictive performance and model interpretability remain limited. This study aims to compare the performance of both algorithms in classifying diabetes status using the Pima Indians Diabetes dataset, which consists of 768 samples with eight numerical attributes. The research stages include data exploration, handling missing values using median imputation, feature standardization using StandardScaler, and stratified data splitting with a ratio of 80:20. Model evaluation is conducted using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC metrics. The experimental results show that KNN with an optimal parameter of K=21 achieves an accuracy of 75.97%, an F1-score of 61.86%, and a ROC-AUC of 0.8120, while Logistic Regression achieves an accuracy of 70.78%, an F1-score of 54.55%, and a ROC-AUC of 0.8130. Although KNN demonstrates higher predictive performance, Logistic Regression provides advantages in interpretability through model coefficients, where the variables Glucose (β=1.1825) and BMI (β=0.6887) are identified as the main predictors of diabetes risk. These findings indicate a clear trade-off between accuracy and interpretability, suggesting that KNN is more suitable for high-accuracy prediction tasks, while Logistic Regression is more appropriate in clinical contexts requiring transparency and model accountability.

Keywords: Diabetes Mellitus; K-Nearest Neighbors; Logistic Regression; Model Interpretability; ROC-AUC

1. PENDAHULUAN

Diabetes melitus merupakan gangguan metabolisme jangka panjang yang kini menjadi krisis kesehatan global. Kondisi ini terjadi ketika tubuh kehilangan kemampuan untuk memproduksi atau memanfaatkan insulin secara optimal, sehingga kadar glukosa darah melampaui batas normal dan secara bertahap merusak berbagai sistem organ [1]. Di Indonesia, prevalensi diabetes pada penduduk usia ≥15 tahun meningkat dari 6,9% (Riskesdas 2013) menjadi 10,9% (Riskesdas 2018), menjadikannya salah satu penyebab utama morbiditas dan mortalitas penyakit tidak menular di tingkat nasional [2]. Peningkatan ini didorong oleh perubahan gaya hidup, urbanisasi yang cepat, serta rendahnya kesadaran deteksi dini mengingat diabetes seringkali tidak menunjukkan gejala klinis yang kentara pada tahap awal [3].

Secara mekanisme, terdapat dua jalur utama yang menjadi penyebab seseorang mengembangkan penyakit diabetes, yaitu kegagalan pankreas dalam menyekresikan insulin dalam jumlah yang memadai maupun terjadinya resistensi sel-sel tubuh terhadap kerja insulin yang dihasilkan [1]. Kedua kondisi tersebut pada akhirnya menghasilkan akumulasi glukosa berlebih dalam darah yang bersifat merusak secara kronik [4]. Perubahan gaya hidup masyarakat modern yang ditandai oleh peningkatan konsumsi makanan berkalori tinggi, berkurangnya aktivitas fisik, serta meningkatnya tingkat stres psikologis menjadi faktor pendorong utama lonjakan kasus diabetes di seluruh dunia, khususnya di kawasan berkembang [5]. Di Indonesia, transformasi sosial akibat urbanisasi yang cepat turut memperburuk tren ini, sehingga diabetes kini dipandang sebagai masalah kesehatan masyarakat yang memerlukan intervensi serius dan terstruktur [6].

Copyright © 2026 **Author**, Page 2679
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2679−2689
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9534

Aspek yang membuat diabetes sangat berbahaya adalah sifatnya yang acap kali tidak menunjukkan gejala klinis yang kentara pada tahap-tahap awal perkembangan penyakit [7]. Ketiadaan keluhan yang signifikan membuat banyak individu tidak menyadari bahwa mereka telah mengidap kondisi ini dalam waktu yang lama, hingga komplikasi serius akhirnya muncul ke permukaan [8]. Kerusakan organ yang timbul akibat paparan hiperglikemia berkepanjangan mencakup gangguan fungsi ginjal, penurunan kemampuan penglihatan, kerusakan jaringan saraf, serta peningkatan risiko penyakit jantung dan pembuluh darah [5]. Komplikasi-komplikasi tersebut tidak hanya menurunkan harapan hidup penderita, tetapi juga menimbulkan beban biaya pengobatan yang sangat besar, baik bagi individu maupun sistem kesehatan secara makro [4]. Oleh karena itu, kemampuan mendeteksi diabetes sejak dini menjadi sangat krusial. Sifat asimtomatik diabetes pada stadium awal ini menegaskan pentingnya sistem prediksi berbasis data yang mampu mendeteksi risiko secara otomatis jauh sebelum manifestasi klinis yang nyata muncul. Pendekatan machine learning, khususnya algoritma klasifikasi, menawarkan solusi menjanjikan dengan kemampuannya mempelajari pola tersembunyi dari data klinis tanpa memerlukan interpretasi manual untuk setiap kasus [9].

Dari dua tipe utama diabetes, Diabetes Mellitus Tipe 2 mendominasi lebih dari 90% kasus secara global dan memiliki relevansi paling tinggi untuk sistem prediksi berbasis kecerdasan buatan [10]. Hal ini disebabkan karena Tipe 2 berkembang secara progresif dengan faktor risiko yang terukur secara klinis seperti kadar glukosa, BMI, usia, dan riwayat keluarga sehingga datanya dapat dikumpulkan secara non-invasif dan dimodelkan secara komputasional. Berbeda dengan Tipe 1 yang disebabkan oleh respons autoimun dan umumnya diderita sejak muda, Tipe 2 sangat erat kaitannya dengan gaya hidup, obesitas, dan faktor keturunan sehingga lebih dapat dicegah melalui intervensi dini berbasis prediksi [6]. Dataset Pima Indians Diabetes yang digunakan dalam penelitian ini secara eksklusif merepresentasikan kasus Diabetes Tipe 2, menjadikannya benchmark yang valid untuk pengembangan model prediksi Tipe 2 [11].

Perkembangan signifikan bidang machine learning telah membuka banyak peluang untuk membangun sistem klasifikasi penyakit yang akurat dan efisien [12]. Pendekatan berbasis data ini memungkinkan komputer untuk mempelajari pola-pola tersembunyi dari kumpulan data medis dan menggunakannya sebagai dasar prediksi kondisi kesehatan seseorang [12]. Di antara berbagai algoritma yang telah dikembangkan, K-Nearest Neighbor (KNN) dan Logistic Regression menjadi dua model yang paling sering dipakai dalam konteks klasifikasi diabetes. [13]. Label kelas sampel ditetapkan berdasarkan kelas yang dominan dari sejumlah K data dengan jarak terdekat pada ruang fitur, sesuai dengan prinsip kemiripan ukuran jarak antara titik data KNN [7]. Algoritma ini dikenal karena kemudahannya dalam implementasi serta tidak memerlukan proses pelatihan yang kompleks, meskipun memiliki kelemahan berupa tingginya biaya komputasi pada dataset berskala besar [14]. Karena nilai yang salah dapat menyebabkan underfitting atau overfitting pada model yang dihasilkan, penentuan nilai K yang tepat sangat penting untuk mengoptimalkan kinerja algoritma ini [15]. Di tengah perkembangan algoritma modern seperti Random Forest, XGBoost, dan jaringan saraf dalam, pengujian KNN tetap relevan karena dua alasan: pertama, KNN tidak memerlukan proses training yang kompleks sehingga lebih mudah diimplementasikan di fasilitas kesehatan dengan sumber daya komputasi terbatas; kedua, performa KNN sangat bergantung pada kualitas preprocessing khususnya standarisasi fitur dan pemilihan nilai K sehingga perbandingan terkontrol dengan preprocessing standar masih diperlukan untuk memberikan baseline yang andal bagi penelitian selanjutnya [14] [15].

Logistic Regression, di sisi lain, merupakan metode statistik yang secara khusus dirancang untuk persoalan prediksi biner dengan menggunakan fungsi sigmoid sebagai inti perhitungannya [16]. Metode ini menghasilkan nilai probabilitas antara nol dan satu yang menunjukkan hubungan antara variabel prediktor dan variabel target. Hasilnya dapat langsung dianggap sebagai peluang bahwa seseorang akan termasuk dalam kelas tertentu [17]. Dengan keunggulan interpretabilitasnya, Logistic Regression dapat digunakan untuk memprediksi dan menganalisis variabel risiko diabetes [18]. Penggunaan mekanisme regularisasi yang tepat pada Logistic Regression sangat menentukan kemampuan model dalam menggeneralisasi data, terutama ketika dataset yang digunakan mengandung fitur-fitur yang saling berkorelasi kuat [19]. Dalam konteks medis, interpretabilitas model bukan sekadar nilai tambah, melainkan kebutuhan mendasar. Dokter dan tenaga kesehatan memerlukan penjelasan yang dapat dipertanggungjawabkan di balik setiap rekomendasi sistem prediksi sebuah prinsip yang dikenal sebagai explainability dalam AI klinis. Logistic Regression memenuhi kebutuhan ini secara langsung: setiap koefisien β dapat diinterpretasikan sebagai kontribusi kuantitatif terhadap log-odds risiko diabetes, sehingga model tidak hanya memberikan prediksi, tetapi juga penjelasan yang bermakna bagi tenaga kesehatan [18].

Berbagai penelitian terdahulu telah mengonfirmasi efektivitas kedua algoritma tersebut dalam mengklasifikasikan penyakit diabetes, namun juga menyoroti keterbatasan masing-masing yang perlu diatasi [13]. Performa KNN terbukti sangat dipengaruhi oleh pilihan nilai K serta metode penghitungan jarak yang digunakan, sementara Logistic Regression menunjukkan sensitivitas tinggi terhadap ketidakseimbangan distribusi kelas dalam data pelatihan [10]. Strategi penanganan khusus, seperti penyeimbangan data atau penyesuaian bobot kelas, diperlukan dalam situasi di mana jumlah sampel positif diabetes lebih sedikit daripada sampel negative [6]. Faktor-faktor ini menegaskan bahwa perbandingan sistematis antara kedua algoritma sangat diperlukan untuk menentukan pendekatan yang paling optimal dalam klasifikasi status diabetes [18].

Dataset Pima Indians Diabetes merupakan kumpulan data standar yang sering dipergunakan dalam penelitian klasifikasi diabetes di seluruh dunia [11]. Dataset ini mengandung delapan atribut klinis: kadar glukosa plasma, tensi darah diastolik, ukuran lipatan kulit, konsentrasi insulin, indeks massa tubuh, skor fungsi riwayat keluarga diabetes, dan usia subjek penelitian [4]. Kombinasi atribut-atribut tersebut menghadirkan profil klinis yang representatif dan

Copyright © 2026 **Author**, Page 2680
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2679−2689
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9534

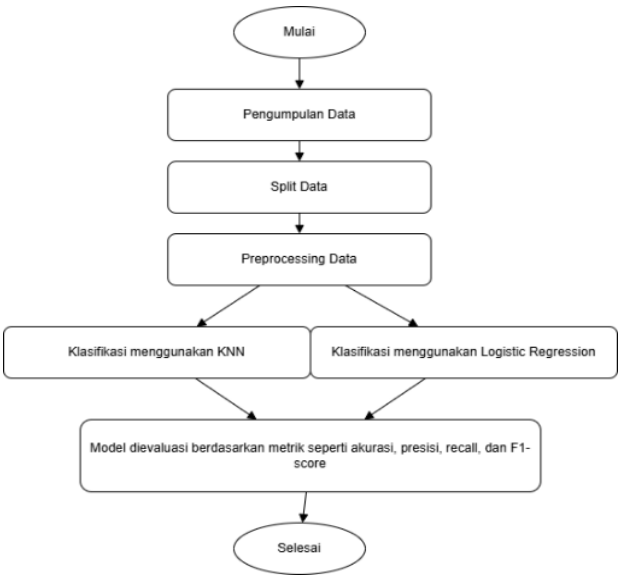
cukup komprehensif untuk membedakan individu yang positif maupun negatif diabetes [11]. Studi yang menggunakan dataset ini menunjukkan bahwa akurasi dan sensitivitas model dapat ditingkatkan secara signifikan dengan seleksi fitur yang cermat dan konfigurasi parameter yang tepat [11]. Selain itu, penelitian komparatif yang melibatkan berbagai algoritma seperti logistic regression, forest random, support vector machine, dan deep learning approach pada dataset Pima memberikan landasan kontekstual yang kuat untuk menilai posisi KNN dalam ekosistem penelitian prediksi diabetes [9]. Meskipun dataset Pima Indians Diabetes tergolong klasik, penggunaannya dalam penelitian ini didasarkan pada tiga pertimbangan metodologis: (1) merupakan benchmark standar internasional yang memungkinkan perbandingan langsung dengan penelitian terdahulu; (2) karakteristik imbalanced class-nya (65,1% Non-Diabetes vs 34,9% Diabetes) merepresentasikan kondisi nyata distribusi kasus di klinik umum; dan (3) keterbatasan akses terhadap dataset medis lokal Indonesia yang telah tervalidasi dan berlabel lengkap mendorong penggunaan dataset benchmark ini sebagai langkah awal sebelum validasi pada data lokal di penelitian mendatang [11].

Berbagai penelitian terdahulu telah mengonfirmasi efektivitas KNN dan Logistic Regression dalam klasifikasi diabetes, namun menyisakan beberapa celah yang belum terjawab. Pratama et al. [10] dan Kurniawan & Megawaty [6] menunjukkan bahwa performa kedua algoritma sangat bergantung pada kualitas preprocessing, khususnya penanganan nilai hilang dan distribusi kelas yang tidak seimbang. Namun demikian, sebagian besar studi tersebut tidak secara eksplisit melaporkan rasio pembagian data, tidak mendiskusikan dampak imbalanced class terhadap bias model, dan tidak menyajikan bukti koefisien kuantitatif sebagai dasar klaim interpretabilitas Logistic Regression. Penelitian ini hadir untuk mengisi celah tersebut dengan menyajikan pipeline yang transparan, evaluasi berbasis data eksperimen nyata, serta interpretasi koefisien model sebagai bukti konkret interpretabilitas. Dengan demikian, penelitian ini menggunakan dataset Pima Indians Diabetes untuk melakukan evaluasi komparatif terstruktur antara KNN dan Logistic Regression, menggunakan metrik akurasi, F1-Score, dan ROC-AUC sebagai tolak ukur evaluasi yang saling melengkapi.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Kajian ini terdiri dari enam tahap utama, yaitu: (1) pengumpulan data, (2) eksplorasi dan analisis data awal, (3) pra-pemrosesan data, (4) penerapan algoritma KNN, (5) penerapan algoritma Logistic Regression, serta (6) evaluasi dan perbandingan performa. Alur metodologi penelitian digambarkan pada Gambar 1.



Gambar 1. Alur Metodologi

Gambar 1 menggambarkan alur metodologi penelitian yang terdiri dari enam tahap: (1) pengumpulan dataset Pima Indians dari Kaggle; (2) eksplorasi data awal meliputi statistik deskriptif, distribusi kelas, dan identifikasi nilai hilang; (3) pra-pemrosesan data mencakup imputasi median pada nilai nol yang tidak valid secara medis, standarisasi fitur menggunakan StandardScaler, serta pemisahan data dengan rasio 80:20 secara stratified; (4) penerapan algoritma KNN dengan pencarian nilai K optimal; (5) penerapan algoritma Logistic Regression dengan regularisasi L2; dan (6) evaluasi serta perbandingan performa kedua model menggunakan metrik akurasi, F1-Score, dan ROC-AUC.

2.2 Dataset

Dataset Pima Indians Diabetes yang bersumber dari Kaggle digunakan dalam penelitian ini. Dataset ini berisi 768 sampel pasien perempuan Pima Indian berusia minimal 21 tahun, dengan delapan fitur numerik klinis (Pregnancies,

Copyright © 2026 **Author**, Page 2681
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2679−2689
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9534

Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age) dan satu variabel target biner (Outcome: 1=Diabetes, 0=Non-Diabetes). Distribusi kelas bersifat tidak seimbang (imbalanced) dengan 500 sampel Non-Diabetes (65,1%) dan 268 sampel Diabetes (34,9%), dengan rasio 1,87:1. Ketidakseimbangan ini berpotensi membiaskan model ke arah kelas mayoritas, sehingga F1-Score dipilih sebagai metrik utama evaluasi kelas positif, di samping penggunaan stratified split untuk menjaga proporsi kelas pada kedua subset data. Terkait ketidakseimbangan kelas, penelitian ini tidak menerapkan teknik resampling seperti SMOTE karena tujuan utama adalah mengevaluasi performa baseline kedua algoritma pada kondisi data asli. Sebagai gantinya, diterapkan dua strategi mitigasi: (1) stratified split untuk memastikan distribusi kelas proporsional pada training (400 Non-Diabetes, 214 Diabetes) dan testing (100 Non-Diabetes, 54 Diabetes), serta (2) penggunaan F1-Score dan ROC-AUC sebagai metrik utama, bukan hanya akurasi, karena akurasi dapat menyesatkan pada kondisi data tidak seimbang. Penerapan SMOTE atau class_weight adjustment direkomendasikan sebagai langkah penelitian lanjutan untuk meningkatkan Recall pada kelas positif diabetes.

2.3 Detail Implementasi

Seluruh eksperimen diimplementasikan menggunakan Python 3 dengan library scikit-learn versi 1.6.1, pandas versi 2.2.2, dan numpy versi 2.0.2. Pemisahan data dilakukan dengan rasio 80:20 menggunakan train_test_split dengan parameter stratify=y dan random_state=42 untuk memastikan proporsi kelas yang terjaga dan hasil yang dapat direproduksi. Untuk KNN, nilai K diuji secara eksperimental pada rentang K=1 hingga K=25 (nilai ganjil) menggunakan metrik jarak Euclidean, dan K=21 dipilih sebagai nilai optimal berdasarkan akurasi tertinggi pada data uji. Untuk Logistic Regression, model dikonfigurasi dengan regularisasi L2 (C=1,0), solver 'lbfgs', dan maksimum iterasi 1.000. Tidak diterapkan cross-validation pada penelitian ini karena tujuan utamanya adalah membandingkan performa pada kondisi split yang identik dan reproducible; validasi lebih lanjut dengan k-fold cross-validation direkomendasikan sebagai langkah penelitian berikutnya.

2.4 K-Nearest Neighbors (K-NN)

Algoritma ini termasuk dalam pembelajaran supervised dan juga dalam kelompok pembelajaran instance-based [14]. KNN telah banyak digunakan dalam tugas klasifikasi medis, termasuk diagnosis diabetes dan penyakit lainnya. Studi menunjukkan bahwa dengan menggabungkan KNN metrik seperti akurasi, presisi, recall, dan skor F1, hasil model klasifikasi medis dapat dinilai dengan sukses [6]. Algoritma KNN menggunakan jarak geometris untuk menentukan ukuran jarak yang ada antara data titik x dan data titik xi. Karena KNN berbasis perhitungan jarak Euclidean, algoritma ini sangat sensitif terhadap perbedaan skala antar fitur. Fitur dengan rentang nilai besar seperti Insulin (0–846) akan mendominasi perhitungan jarak dan mengabaikan kontribusi fitur berskala kecil seperti DiabetesPedigreeFunction (0,078–2,42) apabila tidak distandardisasi terlebih dahulu. Oleh karena itu, seluruh fitur distandarisasi menggunakan StandardScaler (mean=0, std=1) sebelum pelatihan model, dengan proses fit dilakukan hanya pada data training untuk mencegah data leakage. Rumus untuk jarak geometris antara dua titik dalam ruang berukuran n adalah sebagai berikut:

$$d(x, xi) = \sqrt{\sum_{j=1}^{n} (xj - xij)^2} \quad (1)$$

Pada persamaan (1), $x$merupakan titik data yang akan diprediksi, $x_i$adalah titik data dalam dataset pelatihan, $x_j$dan x_{ij}masing-masing merupakan nilai fitur ke-j dari titik data $x$dan x_i, sementara $n$adalah jumlah total fitur dalam dataset.

2.5 Logistic Regression

Regresi logistik adalah jenis analisis regresi logistik yang melihat relasi antara variabel bebas dan variabel terikat, yang menghasilkan nilai 0 hingga 1 untuk klasifikasi biner. Kategori variabel bebas ini memisahkan regresi logistik dari regresi linear atau berganda lainnya [18]. *Logistic Regression* melibatkan fungsi logistik atau sigmoid untuk menghitung hasil linear dari fitur input menjadi kemungkinan yang berada di antara 0 dan 1. Cara kerjanya adalah dengan menghitung odds log suatu kejadian dan kemudian mengubahnya menjadi probabilitas menggunakan fungsi logistik:

$$P(y = 1/X \; \frac{1}{1+e^{-(\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn)}} \quad (2)$$

Pada persamaan (2), $P(y = 1 \mid X)$merupakan probabilitas bahwa suatu sampel termasuk ke dalam kelas 1 (diabetes), $x_1, x_2, \ldots, x_n$adalah fitur-fitur input yang digunakan sebagai prediktor, sementara $\beta_0, \beta_1, \ldots, \beta_n$merupakan koefisien yang dipelajari oleh model selama proses pelatihan.

Setelah model dilatih, model dapat memprediksi probabilitas sebuah data untuk setiap kelas dan mengklasifikasikannya berdasarkan threshold tertentu (misalnya 0.5).

2.6 Confusion Matrix

True Positive (TP), False Positive (FP), True Negative (TN), dan *False Negative* adalah empat komponen utama *Confusion Matrix* dimanfaatkan dalam proses penilaian hasil prediksi model klasifikasi. *Confusion Matrix* dalam Studi

Copyright © 2026 **Author**, Page 2682
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2679−2689
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9534

ini mengevaluasi kinerja model K-Nearest Neighbors (KNN) dan Logistic Regression dalam mengklasifikasikan pasien diabetes menjadi kelas 1 atau 0. Kita dapat menghitung berbagai metrik untuk memeriksa kinerja model dengan Confusion Matrix, seperti:

a. Akurasi merupakan sebagai proporsi perkiraan yang sesuai dari total prediksi.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

b. Presisi merupakan ukuran seberapa besar model memprediksi kelas positif.

$$Presisi = \frac{TP}{TP+FP} \quad (4)$$

c. Recall merupakan ukuran kapasitas model untuk membedakan perbedaan kelas yang positif.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

d. F1-Score mengkombinasikan recall dan presisi ke dalam satu metrik yang harmonis.

$$F1 - Score = 2\,X\,\frac{Presisi\ x\ Recall}{Presisi+Recall} \quad (6)$$

3. HASIL DAN PEMBAHASAN

3.1 Karakteristik Dataset dan Analisis Awal

Dalam studi ini menggunakan Dataset Diabetes Pima Indian, yang berasal dari Kaggle dan mengandung informasi klinis dari pasien perempuan Pima Indian berusia minimal 21 tahun. Ada delapan atribut klinis dalam dataset ini: usia, tekanan darah diastolik, jumlah kehamilan, kadar glukosa plasma, kadar insulin serum, ketebalan lipatan kulit trisep, indeks massa tubuh (BMI), skor fungsi riwayat diabetes keluarga, dan variabel target adalah 1 untuk pasien diabetes dan 0 untuk pasien non-diabetes. Hasil eksplorasi awal mengungkapkan beberapa permasalahan mendasar pada dataset ini, di antaranya ketidakseimbangan distribusi kelas Dalam situasi jumlah pasien non-diabetes lebih besar daripada pasien diabetes, keberadaan nilai nol tidak masuk akal secara medis pada atribut seperti glukosa, tekanan darah, insulin, dan BMI yang mengindikasikan adanya data hilang, serta perbedaan skala nilai yang cukup besar antar atribut yang berpotensi memengaruhi kinerja algoritma berbasis jarak seperti KNN. Oleh karena itu, serangkaian tahapan preprocessing yang mencakup imputasi nilai hilang, penyeimbangan kelas, serta normalisasi dan standarisasi fitur menjadi langkah yang sangat penting untuk dilakukan sebelum proses pelatihan model, guna memastikan data berada dalam kondisi optimal dan model yang dihasilkan mampu mengklasifikasikan status diabetes secara akurat dan dapat diandalkan secara klinis.

Tabel 1. Distribusi Kelas Status Pasien Diabetes

Kelas Outcome	Keterangan	Jumlah Data
0	Non Diabetes	500
1	Diabetes	268
Total		768

Berdasarkan Tabel 1, dataset menunjukkan ketidakseimbangan kelas dengan rasio Non-Diabetes terhadap Diabetes sebesar 1,87:1 (500 vs 268 sampel). Secara klinis, kondisi ini berpotensi membiaskan model ke arah kelas mayoritas sehingga menghasilkan akurasi keseluruhan yang terlihat tinggi namun dengan Recall kelas positif yang rendah kondisi berbahaya secara medis karena berarti banyak pasien diabetes diprediksi sebagai sehat (*false negative*). Sebagai mitigasi, diterapkan stratified split dan F1-Score dipilih sebagai metrik evaluasi utama kelas positif. Dataset memerlukan tahap pra-pemrosesan sebelum digunakan dalam pemodelan, karena beberapa atribut bahkan memiliki nilai nol, yang secara medis tidak masuk akal, terutama untuk variabel glukosa, insulin, dan BMI.

Tabel 2. Statistik Deskriptif

Statistic	Preg	Glucose	BP	SkinThick	Insulin	BMI	DPF	Age
Count	768	768	768	768	768	768	768	768
Mean	3.85	120.89	69.11	20.53	79.80	31.99	0.47	33.24
Std	3.37	31.97	19.36	15.95	115.24	7.88	0.33	11.76
Min	0	0	0	0	0	0	0.078	21
25%	1	99	62	0	0	27.3	0.243	24
50%	3	117	72	23	30.5	32	0.372	29
75%	6	140	80	32	127.25	36.6	0.626	41
Max	17	199	122	99	846	67.1	2.42	81

Tabel 2 menyajikan statistik deskriptif seluruh fitur sebelum pra-pemrosesan. Terlihat bahwa nilai minimum beberapa fitur adalah 0 yang secara fisiologis tidak valid, khususnya Glucose (min=0, 5 sampel), BloodPressure

Copyright © 2026 **Author**, Page 2683
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2679−2689
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9534

(min=0, 35 sampel), SkinThickness (min=0, 227 sampel), Insulin (min=0, 374 sampel), dan BMI (min=0, 11 sampel). Rentang nilai antar fitur juga sangat bervariasi Insulin memiliki nilai maksimum 846 sementara DiabetesPedigreeFunction hanya mencapai 2,42 mengonfirmasi perlunya standarisasi fitur sebelum pemodelan, terutama untuk KNN yang berbasis perhitungan jarak Euclidean.

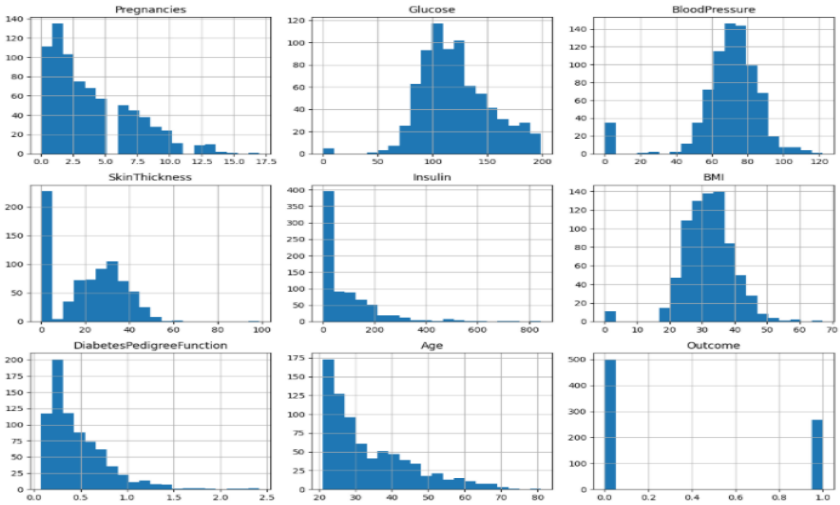
3.2 Pra-pemrosesan Data

Tahapan preprocessing data merupakan langkah krusial sebelum proses pelatihan model, mencakup penanganan nilai hilang, normalisasi fitur, dan penyeimbangan kelas, guna memastikan kualitas data yang optimal dan mencegah bias pada hasil klasifikasi [20]. Pra-pemrosesan data diproses agar kualitas digunakan dengan baik dan layak untuk analisis lebih lanjut. Nilai nol pada beberapa karakteristik klinis dianggap sebagai nilai hilang karena tidak mewakili kondisi fisiologis yang sah. Metode ini sesuai dengan praktik umum dalam penelitian data medis, di mana terdapat data yang tidak tercatat sering ditunjukkan dengan nilai nol. Untuk mengatasi nilai yang hilang, metode imputasi median dipilih karena lebih stabil terhadap keberadaan nilai ekstrem dibandingkan dengan metode imputasi rata-rata. Setelah proses imputasi selesai, dataset siap untuk digunakan pada tahap pemodelan.

Tabel 3. Jumlah Nilai Nol pada Setiap Fitur Sebelum Pra-pemrosesan

Fitur	Jumlah Nilai 0
Pregnancies	111
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
DiabetesPedigreeFunction	0
Age	0

Selanjutnya, standarisasi fitur dilakukan dengan menggunakan *StandardScaler*. Langkah ini diambil untuk memastikan bahwa skala antar fitur sama, karena algoritma KNN sangat bergantung pada penghitungan jarak. Jika tidak ada standarisasi, fitur dengan banyak nilai dapat mendominasi proses klasifikasi.



Gambar 2. Histogram distribusi fitur setelah pra-pemrosesan

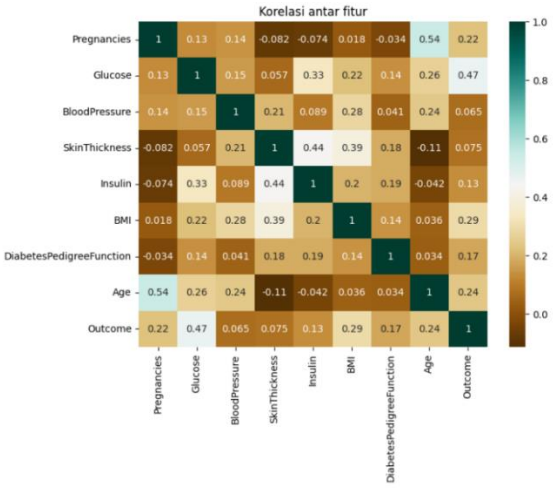
Gambar 2 menyajikan histogram distribusi fitur setelah pra-pemrosesan. Dibandingkan kondisi sebelum preprocessing, spike pada nilai 0 di fitur Glucose, BloodPressure, SkinThickness, Insulin, dan BMI telah hilang setelah imputasi median. Nilai median yang digunakan adalah: Glucose=117, BloodPressure=72, SkinThickness=29, Insulin=125, dan BMI=32,3. Distribusi fitur kini lebih merepresentasikan kondisi fisiologis nyata dan siap digunakan untuk proses pelatihan model.

3.3 Analisis Korelasi Antar Fitur

Untuk memahami bagaimana variabel independen berinteraksi satu sama lain dan hubungannya dengan status diabetes, analisis korelasi dilakukan. Hasil visualisasi korelasi menunjukkan bahwa variabel glukosa memiliki korelasi paling kuat terhadap hasil. Hasilnya konsisten dengan literatur medis yang menyatakan bahwa kadar glukosa darah adalah indikator utama untuk diagnosis diabetes. Selain glukosa, variabel usia dan BMI juga menunjukkan korelasi positif, meskipun dengan kekuatan yang lebih rendah. Tidak ditemukan korelasi yang sangat tinggi antar fitur independen, yang berarti risiko multikolinearitas relatif rendah, terutama ketika menggunakan *Logistic Regression*.

Copyright © 2026 **Author**, Page 2684
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2679−2689
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9534

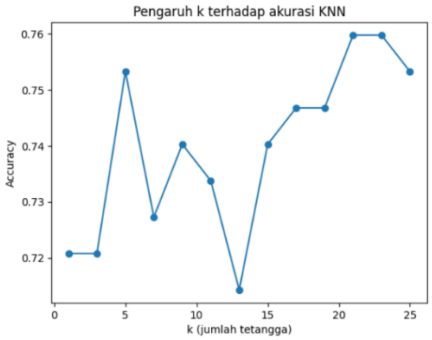


Gambar 3. Heatmap korelasi antar fitur

Gambar 3 menampilkan heatmap korelasi Pearson antar fitur. Koefisien korelasi Pearson selalu berada pada rentang -1 hingga +1: nilai mendekati +1 menunjukkan hubungan linier positif yang kuat, nilai mendekati -1 menunjukkan hubungan linier negatif yang kuat, dan nilai 0 menunjukkan tidak adanya hubungan linier. Berdasarkan visualisasi tersebut, fitur glukosa memiliki korelasi paling tinggi terhadap variabel hasil (sekitar 0,47). Di urutan kedua, BMI (sekitar 0,29), usia (sekitar 0,24), dan kehamilan (sekitar 0,22) masing-masing berada di posisi kedua. Sementara itu, fitur lainnya memiliki korelasi yang relatif lemah terhadap status diabetes. Analisis korelasi ini memberikan gambaran awal mengenai hubungan antar variabel serta membantu mengidentifikasi fitur yang berpotensi berpengaruh dalam proses prediksi menggunakan algoritma K-Nearest Neighbors (KNN) dan Logistic Regression.

3.4 Hasil Klasifikasi Menggunakan Algoritma KNN

Algoritma K-Nearest Neighbors (KNN) bekerja dengan prinsip pengukuran jarak antar titik data, di mana pendekatan ini dimanfaatkan untuk memetakan kemiripan dan keterhubungan setiap sampel dalam dimensi atribut yang tersedia. Konsep dasar algoritma ini yaitu menetapkan kelas data baru untuk dataset pelatihan mengacu kelas terbanyak dari kumpulan tetangga terdekatnya. Biasanya, untuk menghitung jarak antara dua data, metode jarak geometris digunakan, sehingga data yang memiliki karakteristik serupa akan berada pada posisi yang lebih dekat dalam ruang fitur. Pada penelitian ini, pemilihan parameter k sebagai jumlah tetangga terdekat dilakukan secara eksperimental dengan menguji beberapa nilai ganjil dalam rentang 1 hingga 25. Penggunaan nilai ganjil bertujuan untuk mencegah kemungkinan terjadinya jumlah suara yang serupa dalam proses penentuan kelas. Proses pengujian beberapa nilai k tersebut dilakukan untuk mengidentifikasi nilai yang mampu memberikan performa klasifikasi paling optimal terhadap data uji untuk menghindari overfitting atau underfitting, untuk menjaga keseimbangan model, nilai k yang benar sangat vital. Nilai k yang kecil bisa membuat model sangat peka terkait gangguan data, namun nilai k yang berlebihan dapat mempengaruhi model tidak dapat mengidentifikasi pola lokal penting dalam proses klasifikasi. Metode ini diharapkan dapat memungkinkan algoritma KNN untuk membuat prediksi yang lebih akurat tentang status diabetes pasien.



Gambar 4. Pengaruh k terhadap akurasi KNN

Gambar 4 menunjukkan bagaimana nilai K, atau banyaknya tetangga terdekat, memengaruhi akurasi algoritma K-Nearest Neighbors (KNN) dalam memperkirakan status diabetes. Sumbu X menggambarkan nilai K, dan sumbu Y menggambarkan nilai akurasi yang dibuat oleh model. Dari grafik terlihat bahwa perubahan nilai K memengaruhi performa model, di mana akurasi cenderung berfluktuasi pada beberapa nilai K. Nilai akurasi terendah terlihat sekitar K = 13, sedangkan akurasi tertinggi di kisaran K = 21 hingga K = 23 dengan nilai sekitar 0,76. Karena nilai K terlalu

Copyright © 2026 **Author**, Page 2685
This Journal is licensed under a Creative Commons Attribution 4.0 International License

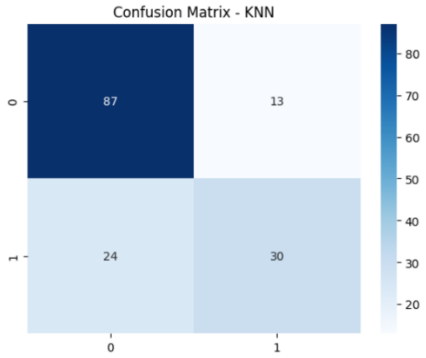
Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2679−2689
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9534

rendah atau terlalu tinggi dapat berdampak kemampuan model untuk mengklasifikasikan data secara benar, hal ini mengindikasikan bahwa Sangat penting untuk memilih nilai K yang tepat untuk mendapatkan performa model yang terbaik.

```
Accuracy (k=21): 0.7597402597402597
Confusion Matrix:
 [[87 13]
 [24 30]]
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.87      0.82       100
           1       0.70      0.56      0.62        54

    accuracy                           0.76       154
   macro avg       0.74      0.71      0.72       154
weighted avg       0.75      0.76      0.75       154

ROC-AUC: 0.8120370370370370
```

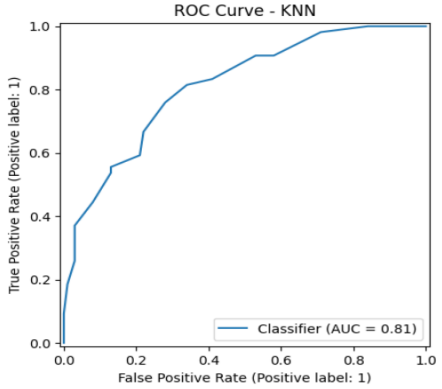
Gambar 5. Classification Report Model K-NN

Hasil uji algoritma K-Nearest Neighbors (KNN) pada nilai K = 21 memiliki akurasi sebesar 75,97%. Berdasarkan confusion matrix, model berhasil memprediksi 87 data tidak diabetes dan 30 data diabetes dengan tepat, namun tetap ada 13 kesalahan dalam memprediksi sebagai diabetes dan 24 kesalahan prediksi sebagai tidak diabetes. Menurut nilai precision, recall, dan F1-score, model mengenali kelas tidak diabetes lebih baik daripada kelas diabetes. Selain itu, model mempunyai kemampuan yang cukup baik untuk mengelompokan orang yang terdiagnosa diabetes dan orang yang tidak, seperti yang ditunjukkan oleh nilai ROC-AUC sebesar 0,812.



Gambar 6. Heatmap Confusion Matrix K-NN

Dalam 154 data uji, model menghasilkan empat komponen klasifikasi, yaitu True Negative (TN) sebesar 87, False Positive (FP) sebesar 13, False Negative (FN) sebesar 24, dan True Positive (TP) sebesar 30, berdasarkan Gambar confusion matrix algoritma K-Nearest Neighbors (KNN) menggunakan parameter optimal k=21. Nilai TN sebesar 87 menunjukkan bahwa model berhasil mengklasifikasikan 87 pasien non-diabetes secara benar, sementara nilai TP sebesar 30 mengindikasikan bahwa model mampu mendeteksi 30 pasien diabetes dengan tepat. Meskipun demikian, model menghasilkan tiga belas kesalahan False Positive pasien yang tidak memiliki diabetes diperkirakan memiliki diabetes, dan dua puluh empat kesalahan False Negative pasien yang semestinya memiliki diabetes diprediksi tidak memiliki diabetes. Berdasarkan nilai-nilai tersebut, model KNN menghasilkan akurasi sebesar 75,97%, precision sebesar 69,77%, recall sebesar 55,56%, dan F1-score sebesar 61,86%. Tingginya angka False Negative pada model ini perlu mendapat perhatian serius dalam konteks medis, karena kesalahan prediksi tersebut berpotensi menyebabkan pasien yang sebenarnya menderita diabetes tidak mendapatkan penanganan yang tepat sejak dini, sehingga meningkatkan risiko berkembangnya komplikasi yang lebih berat di kemudian hari.



Gambar 7. Kurva ROC model KNN

Copyright © 2026 **Author**, Page 2686
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2679−2689
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9534

Kurva ROC (Receiver Operating Characteristic) dari model K-Nearest Neighbors (KNN) yang digunakan untuk memprediksi status diabetes ditunjukkan di gambar. Relasi antara True Positive Rate (TPR) pada sumbu Y dan False Positive Rate (FPR) pada sumbu X digambarkan pada kurva ROC pada berbagai batas nilai. Menurut grafik, model mempunyai keahlian yang cukup untuk memisahkan orang dengan diabetes dari mereka yang tidak, yang memperoleh nilai AUC (Area Under Curve) mencapai 0,81. Nilai yang semakin tinggi dan mendekati angka 1 mengindikasikan bahwa kemampuan model dalam menghasilkan prediksi yang akurat semakin optimal, sehingga nilai AUC 0,81 menunjukkan bahwa algoritma KNN memiliki performa klasifikasi yang cukup efektif pada dataset yang digunakan.

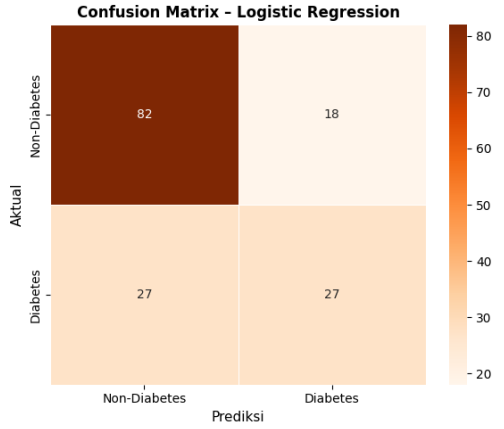
3.5 Hasil Klasifikasi Menggunakan *Logistic Regression*

Teknik *Logistic Regression* menggunakan hubungan antara satu atau lebih variabel untuk menentukan apakah kalimat itu positif atau negative [19]. Hasil yang sama ditemukan dalam penelitian sebelumnya yang menunjukan perbandingan berbagai algoritma klasifikasi untuk memperkirakan risiko diabetes. Ketika evaluasi dilakukan dengan metrik akurasi, *confusion matrix*, dan *ROC-AUC*, kedua algoritma *K-Nearest Neighbors* dan *Logistic Regression* menunjukkan kinerja yang baik. Meskipun Logistic Regression menghasilkan nilai Accuracy (70,78%), F1-Score (0,5455), Precision (60,00%), dan Recall (50,00%) yang lebih rendah dibandingkan KNN, algoritma ini tetap relevan secara klinis karena keunggulannya yang bersifat kualitatif: koefisien model yang dihasilkan Glucose (β=1,1825), BMI (β=0,6887), Pregnancies (β=0,3775) memberikan penjelasan kuantitatif yang dapat diinterpretasikan langsung oleh tenaga kesehatan sebagai kontribusi masing-masing faktor risiko. *F1-score* menunjukkan keseimbangan yang tepat antara *presisi* dan *recall.* Menurut laporan klasifikasi, nilai presisi kelas diabetes menunjukkan bahwa sebagian besar pasien yang diprediksi menderita diabetes sebenarnya termasuk dalam kategori tersebut. Ini membuktikan bahwa ada tingkat kesalahan *false positive* yang rendah dibandingkan dengan yang lain. Kesalahan diagnosis yang dapat menyebabkan kecemasan atau tindakan medis yang tidak diperlukan sangat penting dalam praktik klinis. Hasil ini membuktikan bahwa *Logistic Regression* bukan hanya akurat dalam memprediksi kelas diabetes, tetapi juga cukup konsisten dalam menangani *trade-off* antara kesalahan positif palsu dan negatif palsu. Oleh karena itu, model ini dianggap beroperasi dengan baik dan dapat diandalkan.

Hasil pengujian Logistic Regression menunjukkan akurasi sebesar 70,77% dalam memprediksi status diabetes. Dengan menggunakan confusion matrix, model memprediksi 82 data tidak diabetes dan 27 data diabetes dengan benar; namun, masih ada beberapa kesalahan prediksi. Model mengenali kelas tidak diabetes dengan lebih baik daripada kelas diabetes, menurut nilai ketepatan, recall, dan skor F1. Di sisi lain, nilai ROC-AUC sebesar 0,8129 terbukti model dapat dengan mudah membedakan antara orang yang mengalami diabetes dan mereka yang tidak mengalami.

```
Logistic Regression - Accuracy: 0.7077922077922078

Logistic Regression - Confusion Matrix:
 [[82 18]
 [27 27]]

Logistic Regression - Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.82      0.78       100
           1       0.60      0.50      0.55        54

    accuracy                           0.71       154
   macro avg       0.68      0.66      0.67       154
weighted avg       0.70      0.71      0.70       154

Logistic Regression - ROC-AUC: 0.812962962962963
```

Gambar 8. Classification Report Model *Logistic Regression*

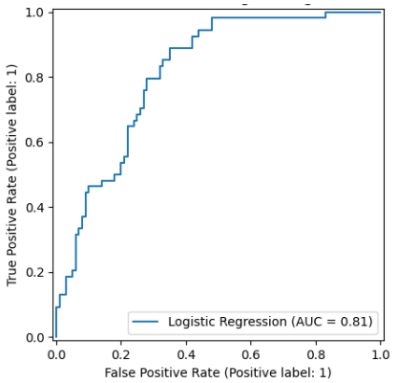


Gambar 9. Heatmap Confusion Matrix *Logistic Regression*

Copyright © 2026 **Author**, Page 2687
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2679−2689
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9534

Berdasarkan Gambar 9, dari 154 data uji model Logistic Regression menghasilkan True Negative (TN) sebesar 82, False Positive (FP) sebesar 18, False Negative (FN) sebesar 27, dan True Positive (TP) sebesar 27. Angka FN=27 berarti 27 pasien yang sebenarnya menderita diabetes diprediksi sebagai sehat, suatu kondisi yang berisiko menyebabkan keterlambatan penanganan medis. Dibandingkan KNN (FN=24), Logistic Regression menghasilkan lebih banyak false negative, yang memperkuat posisi KNN sebagai model dengan Recall lebih baik pada kelas positif diabetes.



Gambar 10. Kurva ROC model Logistic Regression

Tabel 4. Ringkasan Perbandingan Performa KNN vs Logistic Regression

Metrik	KNN (K=21)	Logistic Regression
Accuracy	75,97%	70,78%
Precision	69,77%	60,00%
Recall	55,56%	50,00%
F1-Score	61,86%	54,55%
ROC-AUC	0,8120	0,8130
TN	87	82
FP	13	18
FN	24	27
TP	30	27

Berdasarkan Tabel 4, KNN unggul pada hampir semua metrik kuantitatif, sementara Logistic Regression sedikit mengungguli KNN pada ROC-AUC (selisih 0,0010). Perbedaan tipis pada ROC-AUC mengindikasikan kemampuan diskriminatif kedua model yang setara, sehingga pemilihan algoritma perlu mempertimbangkan kebutuhan spesifik penggunaan: akurasi prediksi versus interpretabilitas model.

4. KESIMPULAN

Penelitian ini membandingkan performa algoritma K-Nearest Neighbors (KNN) dan Logistic Regression dalam klasifikasi Diabetes Mellitus Tipe 2 menggunakan dataset Pima Indians (768 sampel, split 80:20 stratified) dengan pipeline preprocessing yang mencakup imputasi median dan standarisasi fitur. Hasil eksperimen menunjukkan bahwa KNN dengan parameter optimal K=21 menghasilkan akurasi 75,97%, Precision 69,77%, Recall 55,56%, F1-Score 61,86%, dan ROC-AUC 0,8120 mengungguli Logistic Regression yang mencapai akurasi 70,78%, Precision 60,00%, Recall 50,00%, F1-Score 54,55%, dan ROC-AUC 0,8130. Meskipun KNN unggul secara kuantitatif, kedua model masih menghasilkan angka False Negative yang cukup tinggi (KNN: FN=24, LR: FN=27), yang berarti sejumlah pasien diabetes berisiko tidak terdeteksi kondisi berbahaya secara klinis. Sementara itu, Logistic Regression memberikan nilai tambah berupa interpretabilitas model yang terukur: Glucose (β=1,1825), BMI (β=0,6887), dan Pregnancies (β=0,3775) teridentifikasi sebagai prediktor risiko diabetes terkuat, temuan yang konsisten dengan literatur medis dan dapat digunakan sebagai dasar edukasi pasien. Berdasarkan *trade-off* ini, KNN direkomendasikan apabila prioritas adalah akurasi prediksi keseluruhan, sedangkan Logistic Regression lebih sesuai untuk setting klinis yang mengutamakan transparansi dan akuntabilitas model. Untuk penelitian selanjutnya, disarankan menerapkan teknik SMOTE atau penyesuaian class weight untuk meningkatkan Recall pada kelas positif diabetes, serta mempertimbangkan validasi menggunakan k-fold cross-validation dan dataset medis lokal Indonesia untuk meningkatkan generalisabilitas hasil.

REFERENCES

[1] L. M. Cendani and A. Wibowo, "Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes," *Jurnal Masyarakat Informatika*, vol. 13, no. 1, pp. 33–44, 2022, doi: 10.14710/jmasif.13.1.42912.

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2679−2689
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9534

[2] Kementerian Kesehatan Republik Indonesia, "Laporan Riset Kesehatan Dasar (Riskesdas)," 2023. Accessed: Mar. 11, 2026. [Online]. Available: https://www.kemkes.go.id
[3] M. R. Maulana, A. Sucipto, and H. M. Mulyo, "Optimisasi Parameter Support Vector Machine dengan Particle Swarm Optimization untuk Peningkatan Klasifikasi Diabetes," *Jurnal Informatika Teknologi dan Sains (Jinteks)*, vol. 6, no. 4, pp. 802–812, Nov. 2024, doi: 10.51401/jinteks.v6i4.4784.
[4] K. Oliullah, M. H. Rasel, Md. M. Islam, Md. R. Islam, Md. A. H. Wadud, and Md. Whaiduzzaman, "A stacked ensemble machine learning approach for the prediction of diabetes," *J. Diabetes Metab. Disord.*, vol. 23, no. 1, pp. 603–617, Nov. 2023, doi: 10.1007/s40200-023-01321-2.
[5] N. Sunanto and G. Falah, "Penerapan Algoritma C4.5 untuk Membuat Model Prediksi Pasien yang Mengidap Penyakit Diabetes," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 7, no. 2, pp. 208–216, Jul. 2022, doi: 10.36341/rabit.v7i2.2435.
[6] M. F. Kurniawan and D. A. Megawaty, "Comparison of Logistic Regression, Random Forest, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) Algorithms in Diabetes Prediction," *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 2154–2162, 2025, doi: 10.30871/jaic.v9i5.9815.
[7] D. S. F. Azzahrah and Alamsyah, "Klasifikasi Penyakit Diabetes Menggunakan Algoritma K-Nearest Neighbor," *Seminar Nasional Ilmu Komputer (SNIK 2022)*, vol. 1, no. 4, pp. 70–75, 2022.
[8] A. Hamid *et al.*, "Analysis to predict diabetes Using Data Mining," JIEMAR (Journal of Industrial Engineering & Management Research), vol. 6, no. 2, pp. 25–40, doi: 10.7777/jiemar.v6i2.579.
[9] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthc. Technol. Lett.*, vol. 10, no. 1–2, pp. 1–10, Feb. 2023, doi: 10.1049/htl2.12039.
[10] R. Pratama, A. M. Siregar, S. A. P. Lestari, and S. Faisal, "Implementation of Diabetes Prediction Model Using Random Forest Algorithm, K-Nearest Neighbor, and Logistic Regression," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, pp. 1165–1174, Sep. 2024, doi: 10.52436/1.jutif.2024.5.4.2593.
[11] A. Perdana, A. Hermawan, and D. Avianto, "Analyze Important Features of PIMA Indian Database For Diabetes Prediction Using KNN," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 12, no. 1, pp. 70–75, Mar. 2023, doi: 10.32736/sisfokom.v12i1.1598.
[12] R. S. Nurhalizah, R. Ardianto, and P. Purwono, "Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review," *Jurnal Ilmu Komputer dan Informatika*, vol. 4, no. 1, pp. 61–72, 2024, doi: 10.54082/jiki.168.
[13] Dewi Nasien *et al.*, "Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes," *JEKIN - Jurnal Teknik Informatika*, vol. 4, no. 1, pp. 10–17, Feb. 2024, doi: 10.58794/jekin.v4i1.640.
[14] M. Kurniawan Khamdani, N. Hidayat, and R. K. Dewi, "Implementasi Metode K-Nearest Neighbor Untuk Mendiagnosis Penyakit Tanaman Bawang Merah," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 1, pp. 11–16, 2021
[15] V. Wulandari, W. J. Sari, Z. Alfian, L. Legito, and T. Arifianto, "Implementasi Algoritma Naïve Bayes Classifier dan K-Nearest Neighbor untuk Klasifikasi Penyakit Ginjal Kronik," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 710–718, Apr. 2024, doi: 10.57152/malcom.v4i2.1229.
[16] F. D. Pramakrisna, F. D. Adhinata, and N. A. F. Tanjung, "Aplikasi Klasifikasi SMS Berbasis Web Menggunakan Algoritma Logistic Regression," *Teknika*, vol. 11, no. 2, pp. 90–97, Jun. 2022, doi: 10.34148/teknika.v11i2.466.
[17] Q. R. Cahyani *et al.*, "Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm Article Info ABSTRAK," *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, vol. 1, no. 2, pp. 2828–9099, 2022, doi: 10.55123/jomlai.v1i2.598.
[18] M. R. Romadhon and F. Kurniawan, "A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia," *3rd 2021 East Indonesia Conference on Computer and Information Technology, EIConCIT 2021*, pp. 41–44, 2021, doi: 10.1109/EIConCIT50028.2021.9431845.
[19] I. Rahmawati, T. Rika Fitriani, A. No'eman, and A. Y. P. Yusuf, "Analisis Sentimen Menggunakan Algoritma Logistic Regression Pada Penerbangan Lion Air berdasarkan Ulasan Platform Online," *Jurnal Riset Informatika dan Teknologi Informasi*, vol. 1, no. 1, pp. 11–16, Aug. 2023, doi: 10.58776/jriti.v1i1.60.
[20] M. N. Sutoyo and I. Slamet, *Buku Ajar Data Mining*. PT. Sonpedia Publishing Indonesia, 2024.

Copyright © 2026 **Author**, Page 2689
This Journal is licensed under a Creative Commons Attribution 4.0 International License