HIJM
HORIZON: Indonesian Journal of Multidisciplinary

p-ISSN: 2987-7776
e-ISSN: 2987-7180
Vol. 4, No. 3, 2026

EVALUASI ALGORITMA RANDOM FOREST DAN KNN DALAM MEMPREDIKSI RISIKO DIABETES BERDASARKAN FITUR KLINIS

Siti Jamilah Br Tarigan[1], Alyiza Dwi Ningtyas[2], Arif Hamied Nababan[3], Devanta Abraham Tarigan[4], Dini Rizqi Dwikunti Siregar[5]
[1]Institut Teknologi dan Bisnis Indonesia, Jl. Binjai-Stabat, Deli Serdang, Sumatera Utara, Indonesia
[2, 3, 4]Politeknik Negeri Medan, Jl. Almamater No.1, Padang Bulan, Medan, Sumatera Utara, Indonesia
[4]Universitas Syiah Kuala, Jl. Teuku Nyak Arief, Banda Aceh, Aceh, Indonesia
Email: jamilahtarigan@gmail.com

Article History

Received: 03-06-2026

Revision: 23-06-2026

Accepted: 27-06-2026

Published: 30-06-2026

Abstract. This study aims to demonstrate the performance of the Random Forest and K-Nearest Neighbors (KNN) algorithms in predicting diabetes risk based on numerical clinical data. The study used a dataset of 757 samples with eight clinical features, namely the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin, body mass index (BMI), familial diabetes predisposition function, and age. The data was divided into 80% training data and 20% testing data, with data scale adjustments to support the classification process. The evaluation results showed that Random Forest produced better performance with an accuracy of 73.7% and an F1-Score of 0.623, compared to KNN with an accuracy of 72.4% and an F1-Score of 0.604. Comparison of classification results showed that Random Forest was able to provide more consistent predictions in distinguishing groups at risk of diabetes from healthy groups. The contribution of this study is to provide an empirical evaluation of the description of two classification algorithms commonly used on numerical clinical data and show that Random Forest is more suitable for the development of a decision support system for diabetes risk prediction. This research can be the basis for the development of more accurate prediction models through the use of broader datasets and other machine learning methods.

Keywords: Diabetes; Random Forest; K-Nearest Neighbors; F1-Score; Confusion Matrix; Clinical Prediction

Abstrak. Penelitian ini bertujuan mengevaluasi kinerja algoritma Random Forest dan K-Nearest Neighbors (KNN) dalam memprediksi risiko diabetes berdasarkan data klinis numerik. Penelitian menggunakan dataset 757 sampel dengan delapan fitur klinis, yaitu jumlah kehamilan, kadar glukosa, tekanan darah, ketebalan kulit, insulin, indeks massa tubuh (BMI), fungsi predisposisi diabetes keluarga, dan usia. Data dibagi menjadi 80% data pelatihan dan 20% data pengujian, dengan penyesuaian skala data untuk mendukung proses klasifikasi. Hasil evaluasi menunjukkan bahwa Random Forest menghasilkan performa lebih baik dengan akurasi 73,7% dan F1-Score 0,623, dibandingkan KNN dengan akurasi 72,4% dan F1-Score 0,604. Perbandingan hasil klasifikasi menunjukkan Random Forest mampu memberikan prediksi yang lebih konsisten dalam membedakan kelompok berisiko diabetes dan kelompok sehat. Kontribusi penelitian ini adalah memberikan evaluasi empiris terhadap perbandingan dua algoritma klasifikasi yang umum digunakan pada data klinis numerik serta menunjukkan bahwa Random Forest lebih sesuai untuk pengembangan sistem pendukung keputusan prediksi risiko diabetes. Penelitian ini dapat menjadi dasar pengembangan model prediksi yang lebih akurat melalui penggunaan dataset yang lebih luas dan metode pembelajaran mesin lainnya.

Kata Kunci: Diabetes; Random Forest; K-Nearest Neighbors; F1-Score; Confusion Matrix; Prediksi Klinis

How to Cite: Tarigan, S. J. B., Ningtyas, A. D., Nababan, A. H., Tarigan, D. A., & Siregar, D. R. D. (2026). Evaluasi Algoritma *Random Forest* dan KNN dalam Memprediksi Risiko Diabetes Berdasarkan Fitur Klinis. *HORIZON: Indonesian Journal of Multidisciplinary*, 4 (3), 3325-3338. http://doi.org/10.54373/hijm.v4i3.6400

PENDAHULUAN

Diabetes mellitus merupakan salah satu penyakit kronis non-komunikabel yang menjadi masalah kesehatan masyarakat secara global. Menurut laporan Organisasi Kesehatan Dunia (World Health Organization), jumlah penderita diabetes terus meningkat setiap tahun dan diperkirakan menjadi salah satu penyebab utama kematian di banyak negara dalam dekade mendatang(Phongying & Hiriote, 2023). Deteksi dini diabetes sangat penting untuk mencegah komplikasi serius seperti gagal ginjal, gangguan kardiovaskular, dan neuropati. Prediksi risiko diabetes secara akurat berbasis data rekam medis dapat membantu tenaga kesehatan melakukan intervensi lebih awal sehingga menurunkan angka morbiditas dan mortalitas. Oleh karena itu, perkembangan teknologi pembelajaran mesin (machine learning) menjadi peluang strategis dalam pengembangan sistem deteksi dini diabetes berbasis komputasi(Ruby et al., 2023).

Meskipun banyak penelitian telah dilakukan untuk memprediksi diabetes menggunakan machine learning, tantangan utama masih terdapat pada perbandingan kinerja algoritma klasik seperti Random Forest dan K-Nearest Neighbors (KNN) dalam konteks data klinis numerik. Kedua algoritma ini memiliki karakteristik berbeda: Random Forest merupakan metode ensemble yang memanfaatkan banyak pohon keputusan sehingga cenderung stabil dan toleran terhadap outlier, sedangkan KNN merupakan metode non-parametrik yang mengandalkan kedekatan antar contoh data untuk membuat klasifikasi. Perbedaan karakteristik ini memunculkan persoalan: algoritma mana yang memberikan prediksi risiko diabetes yang paling akurat dan stabil ketika diterapkan pada dataset numerik klinis. Persoalan ini menjadi fokus utama penelitian karena efektivitas prediksi berpengaruh langsung pada kualitas rekomendasi medis yang dihasilkan oleh sistem pendukung keputusan.

Penelitian terkait prediksi diabetes berbasis machine learning dalam lima tahun terakhir menunjukkan peningkatan penggunaan algoritma klasifikasi untuk menganalisis data klinis. Beberapa studi pada dataset Pima Indians menunjukkan bahwa Random Forest memiliki performa yang baik karena mampu meningkatkan generalisasi model melalui pendekatan ensemble dan menangani kompleksitas variabel medis secara efektif(Laila et al., 2022; Parhusip et al., 2026; Tasin et al., 2022). Namun, penelitian lain menunjukkan bahwa performa algoritma seperti KNN, Support Vector Machine (SVM), Random Forest, dan XGBoost dapat berbeda bergantung pada optimasi parameter dan karakteristik dataset yang digunakan. KNN tetap memiliki potensi kompetitif setelah dilakukan penyesuaian parameter dan pengolahan data yang tepat(Kurniawan & Megawaty, 2025; Pratama et al., 2024). Oleh karena itu, evaluasi komparatif antara Random Forest dan KNN masih diperlukan untuk menentukan algoritma yang lebih efektif dalam prediksi risiko diabetes berbasis data klinis numerik.

Penelitian lain memfokuskan pada analisis fitur penting untuk prediksi diabetes menggunakan metode KNN, dengan menilai kontribusi setiap fitur klinis terhadap akurasi prediksi. Hasil menunjukkan bahwa fitur seperti glukosa, indeks massa tubuh (BMI), tekanan darah, usia, dan fungsi pedigri diabetes memiliki pengaruh besar dalam menentukan klasifikasi diabetes pada model KNN. Temuan ini menegaskan pentingnya analisis fitur dalam model prediksi diabetes yang berbasis data numerik klinis(Diranisha et al., 2024).

Selain itu, pengembangan sistem prediksi diabetes tidak hanya terbatas pada algoritma klasifikasi sederhana tetapi juga pada integrasi teknik penyeimbangan data. Salah satu penelitian mengimplementasikan Random Forest dengan Synthetic Minority Oversampling Technique (SMOTE) untuk mengatasi ketidakseimbangan kelas dalam dataset diabetes, yang sering menjadi masalah dalam data medis dan dapat memengaruhi efektivitas model prediksi standar. Hasil evaluasi menunjukkan bahwa SMOTE dapat membantu memperbaiki akurasi prediksi, terutama ketika distribusi kelas relatif tidak seimbang(Fadli et al., 2025).

Berdasarkan kajian literatur di atas, masih terdapat celah penelitian terkait evaluasi komprehensif antara algoritma Random Forest dan KNN pada dataset klinis diabetes yang bersifat numerik. Beberapa studi hanya membandingkan dengan algoritma lain seperti Decision Tree atau SVM tanpa fokus langsung pada perbandingan antara Random Forest dan KNN. Selain itu, implementasi metode penyeimbangan kelas atau teknik optimasi parameter sering kali tampil sebagai variabel tersendiri dalam penelitian tetapi belum ditelaah secara terintegrasi dalam satu kerangka evaluasi komprehensif antara kedua algoritma ini(Santos, 2024). Oleh karena itu, penelitian ini bertujuan untuk mengevaluasi secara objektif kinerja algoritma Random Forest dan KNN dalam memprediksi risiko diabetes berdasarkan fitur klinis yang bersifat numerik. Penelitian ini menggunakan dataset numerik yang mencerminkan kondisi kesehatan pasien, seperti kadar glukosa darah, tekanan darah, indeks massa tubuh, usia, dan rekam medis lainnya yang relevan terhadap risiko diabetes. Fokus utama penelitian adalah membandingkan performa kedua algoritma ini berdasarkan metrik evaluasi seperti akurasi, presisi, recall, dan F1-score.

Urgensi penelitian ini didukung oleh meningkatnya prevalensi diabetes yang menjadi salah satu tantangan kesehatan global. Berdasarkan laporan International Diabetes Federation (IDF), jumlah penderita diabetes terus mengalami peningkatan dan diperkirakan mencapai lebih dari 800 juta orang dewasa di dunia pada tahun 2024. Di Indonesia, diabetes juga menjadi salah satu penyakit tidak menular dengan jumlah kasus yang terus meningkat, sehingga diperlukan pendekatan deteksi dini yang lebih efektif. Sistem prediksi berbasis machine learning berpotensi membantu tenaga medis dalam mengidentifikasi individu dengan risiko tinggi

diabetes secara lebih cepat dan akurat. Selain itu, penelitian ini berkontribusi dalam mengevaluasi kemampuan algoritma machine learning klasik pada data klinis numerik sebagai dasar pengembangan sistem pendukung keputusan kesehatan berbasis data.

Solusi yang ditawarkan dalam penelitian ini adalah melakukan evaluasi komparatif terhadap algoritma Random Forest dan KNN untuk prediksi risiko diabetes menggunakan data klinis numerik. Kebaruan penelitian ini terletak pada analisis perbandingan langsung kedua algoritma dengan pendekatan evaluasi yang konsisten pada dataset yang sama, sehingga dapat mengidentifikasi karakteristik performa masing-masing metode dalam menangani variabel klinis diabetes. Selain menghasilkan nilai kinerja model, penelitian ini memberikan bukti empiris mengenai algoritma yang lebih stabil dan sesuai untuk mendukung pengembangan sistem prediksi risiko diabetes berbasis data. Hasil penelitian diharapkan menjadi referensi dalam pemilihan metode machine learning yang efektif untuk aplikasi sistem pendukung keputusan kesehatan. Kontribusi penelitian mencakup penyajian analisis komparatif yang sistematis antara Random Forest dan KNN untuk prediksi risiko diabetes berdasarkan fitur klinis numerik, penilaian performa model berdasarkan berbagai metrik evaluasi untuk memungkinkan rekomendasi algoritma optimal, serta penentuan kontribusi fitur individual terhadap model prediksi melalui analisis fitur penting untuk meningkatkan interpretabilitas model. Penelitian ini tidak hanya memberikan kontribusi ilmiah dalam bidang machine learning dan kesehatan tetapi juga membekali praktisi data dan tenaga kesehatan dengan pendekatan prediktif yang dapat mendukung deteksi dini diabetes secara lebih efektif.

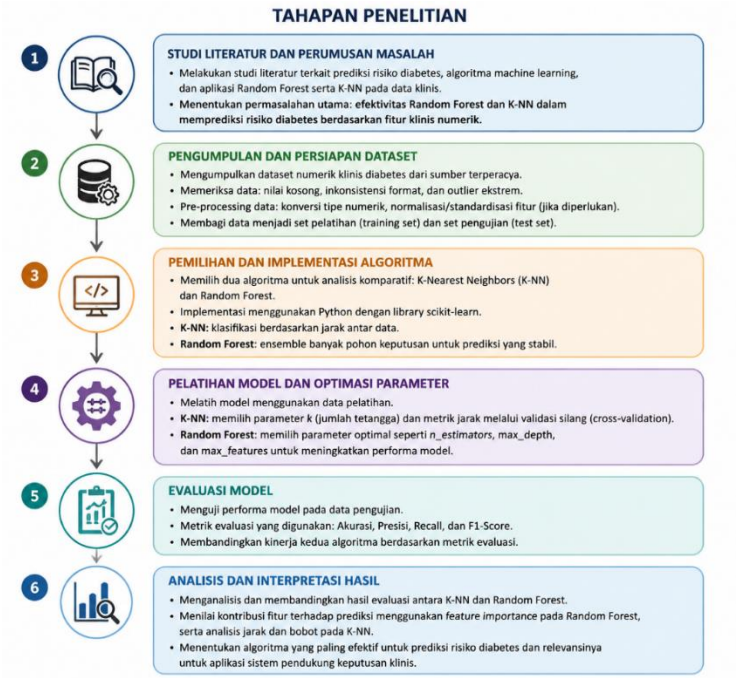
METODE

Tahapan penelitian ini dirancang secara sistematis agar setiap proses dapat berjalan sesuai tujuan penelitian. Secara umum, tahapan penelitian meliputi:

- Studi literatur dan perumusan masalah; penelitian dimulai dengan studi literatur terkait prediksi risiko diabetes, algoritma machine learning, dan aplikasi Random Forest serta K-NN pada data klinis. Dari literatur, ditentukan permasalahan utama yaitu efektivitas kedua algoritma dalam memprediksi risiko diabetes berdasarkan fitur klinis numerik.
- Pengumpulan dan persiapan dataset; dataset numerik klinis diabetes dikumpulkan dari sumber terpercaya. Data ini kemudian diperiksa untuk memastikan tidak ada nilai kosong, inkonsistensi format, atau nilai outlier ekstrem yang dapat memengaruhi hasil prediksi.
- Pemilihan dan implementasi algoritma; dua algoritma dipilih untuk analisis komparatif, yaitu K-Nearest Neighbors (K-NN) dan Random Forest. Implementasi dilakukan dengan pemrograman Python menggunakan library scikit-learn. Algoritma K-NN memanfaatkan

jarak antar data untuk klasifikasi, sedangkan Random Forest menggunakan ensemble banyak pohon keputusan untuk menghasilkan prediksi yang lebih stabil.

- Pelatihan model dan optimasi parameter; setiap algoritma dilatih menggunakan data pelatihan. Untuk K-NN, parameter penting seperti jumlah tetangga (k) dan metrik jarak dipilih melalui validasi silang (cross-validation). Untuk Random Forest, parameter seperti jumlah pohon (n_estimators), kedalaman pohon maksimum (max_depth), dan jumlah fitur maksimum dipilih untuk mengoptimalkan performa model.
- Evaluasi model; setelah proses pelatihan selesai, kedua model diuji menggunakan data pengujian sebesar 20% dari total dataset. Performa Random Forest dan KNN dibandingkan berdasarkan hasil klasifikasi terhadap data risiko diabetes. Evaluasi dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score untuk menentukan algoritma dengan kinerja terbaik. Hasil evaluasi tersebut digunakan sebagai dasar analisis perbandingan kemampuan kedua algoritma dalam memprediksi risiko diabetes berdasarkan fitur klinis numerik.
- Analisis dan interpretasi hasil; hasil pengujian kedua algoritma dibandingkan berdasarkan nilai evaluasi yang diperoleh pada data pengujian. Pada Random Forest, analisis tambahan dilakukan dengan mengidentifikasi fitur yang memiliki kontribusi terbesar terhadap hasil prediksi. Sementara itu, pada KNN dilakukan evaluasi berdasarkan hasil klasifikasi dari data uji setelah proses penyesuaian skala fitur. Hasil perbandingan tersebut digunakan untuk menentukan algoritma dengan performa terbaik dalam prediksi risiko diabetes berbasis data klinis numerik.



Gambar 1. Tahapan penelitian

Secara keseluruhan, tahapan penelitian dirancang untuk memastikan bahwa proses pengolahan data, implementasi algoritma, dan evaluasi hasil dilakukan secara sistematis dan dapat menghasilkan temuan yang valid serta dapat direplikasi.

Dataset

Dataset yang digunakan dalam penelitian ini adalah Pima Indians Diabetes Dataset, yang bersifat numerik dan klinis. Dataset ini memuat informasi rekam medis pasien yang relevan untuk prediksi risiko diabetes. Dataset terdiri dari 757 sampel dengan 8 fitur numerik dan 1 kolom target (Outcome). Fitur-fitur tersebut meliputi:

Tabel 1. Dataset diabetes

No	Nama Atribut	Deskripsi
1	Pregnancies	Jumlah Kehamilan
2	Glucose	Kadar Glukosa Darah
3	BloodPressure	Tekanan Darah
4	SkinThickness	Ketebalan Kulit Tricep
5	Insulin	Kadar Insulin Serum
6	BMI	Indeks Massa Tubuh
7	DiabetesPedigreeFunction	Fungsi Predisposisi Diabetes Keluarga
8	Age	Usia Pasien

Algoritma K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN) merupakan algoritma klasifikasi non-parametrik yang bekerja berdasarkan prinsip kedekatan atau jarak antar titik data. Tahapan K-NN meliputi(Depari et al., 2025; Sudestra et al., 2026):

- Menentukan nilai k, yaitu jumlah tetangga terdekat yang akan digunakan dalam klasifikasi.
- Menghitung jarak antara data baru dengan semua data pada set pelatihan. Umumnya digunakan rumus Euclidean distance:

  $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ (1)

  di mana x adalah vektor fitur data baru, $y$adalah vektor fitur data pelatihan, dan n adalah jumlah fitur.
- Menentukan k tetangga terdekat berdasarkan jarak terkecil.
- Klasifikasi berdasarkan mayoritas kelas tetangga terdekat:

  $\hat{y} = mode\ \{y1, y2, \ldots yk\}$ (2)
- Prediksi kelas data baru diberikan sesuai mayoritas kelas tetangga.
- Kelebihan K-NN adalah sederhana dan efektif pada dataset kecil hingga menengah, namun sensitif terhadap skala fitur dan outlier, sehingga perlu preprocessing berupa normalisasi atau standardisasi fitur numerik.

Algoritma Random Forest

Random Forest adalah algoritma ensemble learning berbasis pohon keputusan. Konsep utamanya adalah membangun banyak pohon keputusan (decision trees) pada subset data yang berbeda, lalu melakukan voting mayoritas untuk menentukan klasifikasi akhir. Tahapan Random Forest(Shamriz Nahzat & Yağanoğlu, 2021; Shetty et al., 2026):

- Bootstrap Sampling: Mengambil sampel acak dengan pengembalian (sampling with replacement) dari dataset pelatihan untuk setiap pohon.
- Membangun Pohon Keputusan. Untuk setiap node pohon, pilih subset fitur acak untuk menentukan split terbaik. Gunakan Gini Impurity atau Entropy untuk menentukan split optimal. Gini Impurity dirumuskan sebagai:

  $Gini = 1 - \sum_{i=1}^{C} p_i^2$ (3)

  di mana p_i adalah proporsi kelas i di node dan C adalah jumlah kelas.
- Membangun Banyak Pohon: Ulangi proses bootstrap dan pembentukan pohon untuk n pohon.
- Voting Mayoritas: Setiap pohon memberikan prediksi kelas, dan kelas yang paling banyak dipilih menjadi prediksi akhir:

  $\hat{y} = mode\ \{y1, y2, \ldots yk\}$ (2)
- Evaluasi Model: Random Forest mampu menangani data numerik dengan baik, mengurangi risiko overfitting dibanding pohon tunggal, serta memberikan analisis feature importance untuk mengetahui kontribusi setiap fitur terhadap prediksi.

Evaluasi Model

Setelah model klasifikasi dijalankan menggunakan algoritma K-Nearest Neighbors (K-NN) dan Random Forest, evaluasi performa dilakukan untuk menilai akurasi dan efektivitas prediksi risiko diabetes. Dua metrik utama yang digunakan adalah Confusion Matrix dan F1-Score(Siregar et al., 2023).

Confusion Matrix

Confusion Matrix adalah tabel yang digunakan untuk menilai performa model klasifikasi dengan membandingkan prediksi model dengan label sebenarnya. Matriks ini berisi empat komponen utama (Tan et al., 2023) yaitu *True Positive (TP):* jumlah data positif yang benar diprediksi positif, *(2) True Negative (TN)*: jumlah data negatif yang benar diprediksi negatif,

(3) *False Positive (FP):* jumlah data negatif yang salah diprediksi positif, dan (4) *False Negative (FN):* jumlah data positif yang salah diprediksi negatif.

Tabel 2. Consufison Matrix

	Prediksi Positif	**Prediksi Negatif**
Data Positif	TP	FN
Data Negatif	FP	TN

Matriks ini membantu memahami jenis kesalahan yang dilakukan model dan seberapa baik model membedakan antara kelas positif dan negatif.

F1-Score

F1-Score adalah metrik evaluasi yang menggabungkan precision dan recall menjadi satu nilai tunggal untuk menilai keseimbangan antara keduanya. F1-Score sangat berguna ketika dataset memiliki distribusi kelas yang tidak seimbang(Mushtaq et al., 2022).

- Precision (Presisi): proporsi prediksi positif yang benar-benar positif.

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

- Recall (Sensitivitas): proporsi data positif yang berhasil diprediksi sebagai positif.

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

- F1-Score: rata-rata harmonik antara precision dan recall.

  $F1 = 2 * \frac{Precision*Recall}{Precision+Recall}$ (6)

Nilai F1-Score berkisar antara 0 hingga 1, di mana nilai yang mendekati 1 menunjukkan performa model yang baik dalam memprediksi kelas positif dengan akurat dan lengkap.

HASIL

Hasil Pengujian Model

Bagian ini menyajikan hasil pengujian model klasifikasi risiko diabetes menggunakan algoritma K-Nearest Neighbors (K-NN) dan Random Forest. Evaluasi dilakukan menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan visualisasi Confusion Matrix. Dataset pengujian terdiri dari 20% sampel dari total dataset diabetes yang telah diproses.

Pengujian Algoritma K-NN

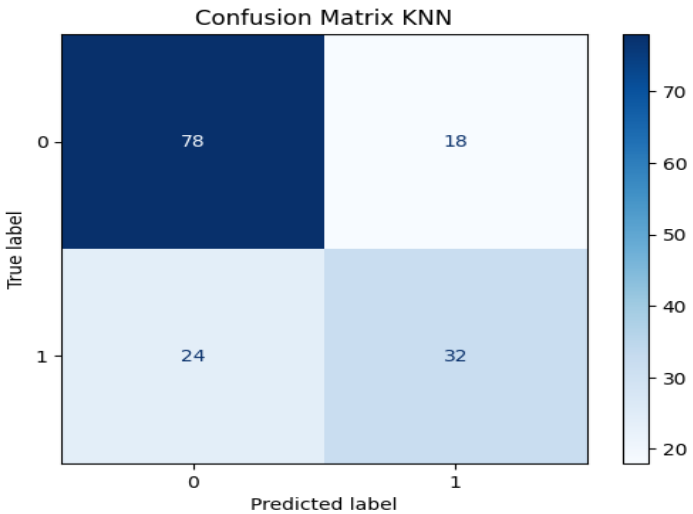
Pada subbab ini, dijelaskan hasil pengujian model K-Nearest Neighbors (K-NN) dalam memprediksi risiko diabetes menggunakan dataset numerik klinis. Model K-NN dikembangkan berdasarkan prinsip kedekatan antar sampel, di mana klasifikasi setiap data uji

ditentukan oleh mayoritas kelas dari k tetangga terdekat. Subbab ini akan menyajikan hasil evaluasi model termasuk akurasi, presisi, recall, F1-Score, serta distribusi prediksi yang ditampilkan melalui Confusion Matrix. Evaluasi ini bertujuan untuk menilai seberapa baik K-NN dalam membedakan pasien dengan risiko diabetes dan pasien sehat berdasarkan fitur klinis numerik. Adapun hasil dari proses pengujian terhadap Algoritma K-NN dapat dilihat berikut:

Tabel 3. Hasil Pengujian Algoritma K-NN

Metric	Hasil Nilai
Akurasi	72.4%
Precision	0,64
Recall	0,57
F1-Score	0,604

Tabel 3 menampilkan evaluasi kinerja model klasifikasi pada dataset diabetes menggunakan algoritma K-Nearest Neighbors (K-NN). Nilai akurasi sebesar 72,4% menunjukkan bahwa sekitar 72 dari 100 sampel dapat diprediksi dengan benar oleh model, baik untuk pasien dengan risiko diabetes maupun pasien sehat. Precision sebesar 0,64 menandakan bahwa dari seluruh prediksi positif yang dihasilkan model, 64% benar-benar merupakan pasien dengan diabetes. Sementara itu, recall sebesar 0,57 menunjukkan bahwa model berhasil mendeteksi 57% dari total pasien diabetes yang sebenarnya ada dalam data pengujian. Nilai F1-Score sebesar 0,604 menggambarkan keseimbangan antara presisi dan recall, yang menunjukkan bahwa model memiliki performa cukup baik dalam memprediksi risiko diabetes dengan memperhitungkan baik akurasi prediksi positif maupun kemampuan mendeteksi pasien diabetes secara tepat. Hasil ini mengindikasikan bahwa K-NN mampu memberikan prediksi yang moderat akurat, namun masih terdapat peluang untuk peningkatan performa, misalnya melalui optimasi parameter k atau teknik feature scaling yang lebih cermat.



Gambar 2. Confusin Matrix Algoritma K-NN

Gambar 2. Confusion Matrix untuk model K-Nearest Neighbors (KNN) menunjukkan distribusi prediksi model terhadap label sebenarnya pada dataset pengujian diabetes. Dari matriks ini terlihat bahwa model berhasil mengklasifikasikan 78 pasien sehat dengan benar sebagai negatif (True Negative) dan 32 pasien diabetes dengan benar sebagai positif (True Positive). Namun, model juga melakukan beberapa kesalahan prediksi, yaitu 18 pasien sehat diprediksi sebagai diabetes (False Positive) dan 24 pasien diabetes diprediksi sebagai sehat (False Negative). Hal ini menunjukkan bahwa meskipun KNN mampu mengenali sebagian besar pasien dengan tepat, terdapat sejumlah kesalahan yang mempengaruhi akurasi keseluruhan. Confusion Matrix ini memberikan gambaran visual yang jelas mengenai kinerja model, memperlihatkan kekuatan model dalam mendeteksi kasus positif dan negatif, serta menjadi dasar perhitungan metrik evaluasi lain seperti Precision, Recall, dan F1-Score.

Pengujian Algoritma Random Forest

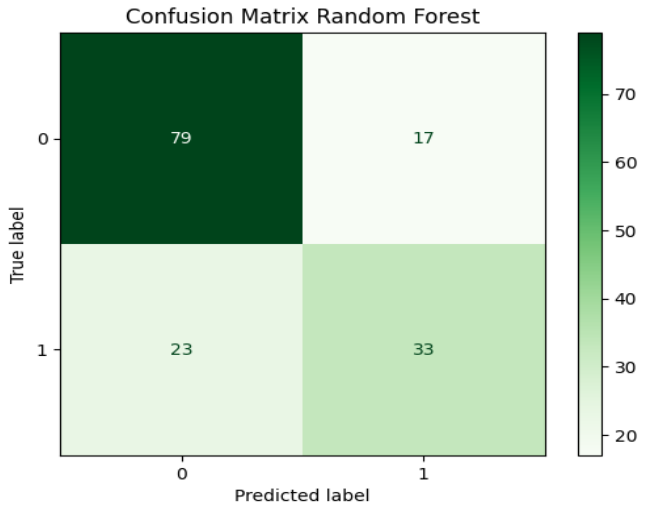
Subbab ini membahas hasil pengujian model Random Forest untuk prediksi risiko diabetes. Random Forest menggunakan banyak pohon keputusan yang bekerja secara ensemble untuk menghasilkan prediksi akhir melalui voting mayoritas. Model ini dirancang untuk menangani kompleksitas fitur numerik pada dataset klinis dan meminimalkan kesalahan prediksi. Pada bagian ini, disajikan hasil evaluasi kinerja model, termasuk akurasi, presisi, recall, F1-Score, serta distribusi prediksi melalui Confusion Matrix. Analisis ini akan menunjukkan kemampuan Random Forest dalam memberikan prediksi yang lebih stabil dan akurat dibanding K-NN, serta memberikan dasar perbandingan performa antar algoritma. Adapun hasil dari proses pengujian terhadap Algoritma Random Forest dapat dilihat berikut:

Tabel 4. Hasil pengujian *random forest*

Metric	Hasil Nilai
Akurasi	73.7%
Precision	0,66
Recall	0,59
F1-Score	0,623

Tabel tersebut menampilkan hasil evaluasi kinerja model Random Forest dalam memprediksi risiko diabetes pada dataset pengujian. Nilai akurasi sebesar 73,7% menunjukkan bahwa model mampu memprediksi dengan benar sekitar 74 dari setiap 100 pasien, baik yang sehat maupun yang berisiko diabetes. Precision sebesar 0,66 menandakan bahwa dari seluruh prediksi positif yang diberikan oleh model, 66% benar-benar merupakan pasien dengan diabetes. Recall sebesar 0,59 menunjukkan bahwa model berhasil mendeteksi 59% dari total

pasien diabetes yang sebenarnya ada dalam dataset pengujian. Nilai F1-Score sebesar 0,623 menggambarkan keseimbangan antara presisi dan recall, menunjukkan bahwa model memiliki kemampuan yang baik dalam memprediksi risiko diabetes dengan mempertimbangkan kedua aspek tersebut. Hasil ini menunjukkan bahwa Random Forest memberikan performa yang lebih stabil dan sedikit lebih unggul dibanding K-NN pada dataset klinis numerik, sehingga dapat dijadikan dasar untuk pengambilan keputusan prediktif dalam sistem pendukung klinis. Adapun hasil dari proses pengujian dapat dilihat berdasarkan dengan gambar Confusin Matrix berikut:



Gambar 3. Confusion *matrix random forest*

Gambar 3. Confusion Matrix untuk model Random Forest menunjukkan distribusi prediksi model terhadap label sebenarnya pada dataset pengujian diabetes. Dari matriks ini terlihat bahwa model berhasil mengklasifikasikan 79 pasien sehat dengan benar sebagai negatif (True Negative) dan 33 pasien diabetes dengan benar sebagai positif (True Positive). Namun, terdapat sejumlah kesalahan prediksi, yaitu 17 pasien sehat diprediksi sebagai diabetes (False Positive) dan 23 pasien diabetes diprediksi sebagai sehat (False Negative). Hal ini menunjukkan bahwa meskipun Random Forest mampu mengenali sebagian besar pasien dengan tepat, masih ada kesalahan prediksi yang harus diperhatikan. Confusion Matrix ini memberikan gambaran visual yang jelas mengenai kemampuan model dalam membedakan pasien sehat dan pasien diabetes, serta menjadi dasar perhitungan metrik evaluasi lain seperti Precision, Recall, F1-Score, dan Akurasi, yang menunjukkan performa keseluruhan model.

DISKUSI

Berdasarkan evaluasi, Random Forest menunjukkan performa sedikit lebih unggul dibanding KNN. Hal ini sejalan dengan karakter algoritma Random Forest yang memanfaatkan ensemble banyak pohon keputusan, sehingga lebih stabil dan toleran terhadap variasi nilai fitur numerik, sedangkan KNN mengandalkan jarak antar sampel yang membuatnya lebih sensitif terhadap skala fitur dan nilai ekstrem.

Tabel 5. Perbandingan hasil pengujian

Model	Akurasi	Precision	Recall	F1-Score
KNN	72.4%	0,64	0.57	0,604
Random Forest	73.7%	0,66	0.59	0,623

Berdasarkan hasil evaluasi pada Tabel 5, Random Forest menunjukkan performa yang lebih konsisten dibandingkan KNN dalam memprediksi risiko diabetes pada dataset pengujian. Keunggulan ini menunjukkan bahwa pendekatan ensemble pada Random Forest mampu menangkap pola hubungan antar fitur klinis secara lebih baik, terutama ketika terdapat variasi karakteristik data pasien. Sebaliknya, performa KNN cenderung dipengaruhi oleh distribusi data karena proses klasifikasi bergantung pada kedekatan antar sampel. Hasil klasifikasi juga menunjukkan bahwa Random Forest lebih mampu menyeimbangkan kemampuan dalam mengenali pasien berisiko diabetes dan pasien tanpa risiko diabetes. Hal ini mengindikasikan bahwa Random Forest memiliki stabilitas yang lebih baik untuk diterapkan pada data klinis numerik. Sementara itu, KNN tetap memberikan hasil yang kompetitif, tetapi sensitivitasnya terhadap skala fitur dan pola persebaran data menjadi faktor yang dapat memengaruhi konsistensi prediksi. Kontribusi fitur *(feature importance)* pada Random Forest menyoroti variabel seperti Glucose, BMI, dan Age sebagai faktor paling berpengaruh terhadap prediksi risiko diabetes. Pada KNN, kontribusi fitur terlihat melalui pengaruh nilai fitur terhadap perhitungan jarak antar sampel, yang juga menentukan klasifikasi mayoritas tetangga. H

Hasil penelitian menunjukkan bahwa Random Forest memiliki performa yang lebih stabil dibandingkan KNN dalam prediksi risiko diabetes berbasis fitur klinis numerik. Meskipun KNN mampu memberikan hasil prediksi yang kompetitif, penerapannya membutuhkan pengaturan parameter dan penyesuaian skala data yang lebih optimal. Temuan ini menunjukkan bahwa pemilihan algoritma memiliki peran penting dalam pengembangan sistem prediksi kesehatan berbasis data. Dalam konteks klinis, model dengan performa yang lebih konsisten seperti Random Forest berpotensi digunakan sebagai alat bantu untuk melakukan identifikasi awal pasien berisiko diabetes, sehingga tenaga kesehatan dapat menentukan prioritas pemeriksaan dan intervensi lebih dini. Namun, implementasi lebih lanjut tetap

memerlukan validasi menggunakan dataset yang lebih besar dan beragam agar model dapat diterapkan secara lebih luas.

KESIMPULAN

Kesimpulan dari penelitian ini menunjukkan bahwa algoritma Random Forest dan K-Nearest Neighbors (KNN) keduanya mampu memprediksi risiko diabetes berdasarkan dataset klinis numerik dengan performa yang memadai, namun Random Forest menunjukkan hasil yang lebih stabil dan akurat. Berdasarkan metrik evaluasi, Random Forest memperoleh akurasi sebesar 73,7% dan F1-Score 0,623, sedangkan KNN memperoleh akurasi 72,4% dan F1-Score 0,604. Confusion Matrix untuk kedua algoritma menunjukkan bahwa Random Forest memiliki jumlah True Positive dan True Negative lebih tinggi serta kesalahan prediksi lebih sedikit dibanding KNN, mengindikasikan kemampuan model ensemble dalam menangani kompleksitas fitur numerik dan outlier. Hasil ini menjawab permasalahan utama penelitian, yaitu menentukan algoritma yang lebih efektif untuk memprediksi risiko diabetes pada dataset numerik klinis.

REKOMENDASI

Keterbatasan penelitian ini antara lain ukuran dataset yang relatif terbatas dan masih adanya kesalahan prediksi pada kedua algoritma, terutama pada pasien diabetes yang terklasifikasi sebagai negatif. Pengaruh fitur lain yang mungkin relevan secara klinis tidak termasuk dalam dataset ini, sehingga potensi informasi tambahan belum dimanfaatkan secara penuh. Perbaikan pada penelitian selanjutnya dapat dilakukan dengan menambahkan dataset yang lebih besar dan beragam, menggabungkan teknik feature selection atau engineering, serta mengeksplorasi algoritma lain seperti XGBoost atau metode deep learning untuk meningkatkan akurasi prediksi. Penelitian ini berhasil memberikan dasar empiris mengenai kinerja Random Forest dan KNN untuk prediksi risiko diabetes serta memberikan rekomendasi algoritma yang dapat diterapkan dalam sistem pendukung keputusan klinis.

REFERENSI

Depari, A. D. S., Tania, K. D., & Sevtiyuni, P. E. (2025). Penerapan Metode Machine Learning Dan Teknik SMOTE untuk Prediksi Diabetes. *Jurnal Sistem Komputer & Informatika (JSON)*, *7*(2), 436–447. https://doi.org/https://doi.org/10.30865/json.v7i2.9032

Diranisha, V., Triayudi, A., & Komalasari, R. T. (2024). Implementation of KNN Algorithm and Random Forest Algorithm in Identifying Diabetes. *SAGA Journal of Technology and Information System*, *2*(2), 234–244. https://doi.org/10.58905/saga.v2i2.253

Fadli, M. B., Purnama, I., & Rohani. (2025). Komparasi Perbandingan Algoritma C4.5, Naive Bayes, KNN, Random Forest untuk Prediksi Penyebab Penyakit Diabetes. *Building of Informatics, Technology and Science (BITS)*, *7*(3), 2118–2126. https://doi.org/https://doi.org/10.47065/bits.v7i3.8683

Kurniawan, M. F., & Megawaty, D. A. (2025). Comparison of Logistic Regression, Random Forest, SVM, and KNN Algorithms in Diabetes Prediction. *Journal of Applied Informatics and Computing*, *9*(5), 2154–2162. https://doi.org/10.30871/jaic.v9i5.9815

Laila, U. e, Mahboob, K., Khan, A. W., Khan, F., & Taekeun, W. (2022). An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study. *Sensors*, *22*(4). https://doi.org/10.3390/s22145247

Mushtaq, Z., Ramzan, M. F., Ali, S., Baseer, S., Samad, A., & Husnain, M. (2022). Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques. *Mobile Information Systems*, *16*. https://doi.org/https://doi.org/10.1155/2022/6521532

Parhusip, J., Stevans, G., Afandy, M. A., Rafliansyah, M., & Andriansah, T. (2026). Analisis Perbandingan Akurasi Algoritma K-NN, Decision Tree, dan Random Forest dalam Klasifikasi Penyakit Diabetes. *Jurnal Media Informatika*, *7*(1), 26–32. https://doi.org/10.55338/jumin.v7i1.7835

Phongying, M., & Hiriote, S. (2023). Diabetes Classification Using Machine Learning Techniques. *Computation*, *11*(5). https://doi.org/10.3390/computation11050096

Pratama, R., Amril, Lestari, S. A. P., & Faisal, S. (2024). Implementation of Diabetes Prediction Model Using Random Forest, KNN, and Logistic Regression. *Jurnal Teknik Informatika*, *5*(4), 1165–1174.

Ruby, A. U., Chandran, J. G. C., Jain, T. J. S., Chaithanya, B. N., & Patil, R. (2023). RFFE – Random Forest Fuzzy Entropy for the Classification of Diabetes Mellitus. *AIMS Public Health*, *10*(2), 422–442. https://doi.org/10.3934/publichealth.2023030

Santos, V. de S. (2024). Comparison and selection of machine learning algorithms for diabetes prediction: An exploratory quantitative study based on medical data analysis. *Seven Editora*, *1*, 737–765. https://doi.org/10.56238/sevened2024.007-053

Shamriz Nahzat, & Yağanoğlu, M. (2021). Diabetes Prediction Using Machine Learning Classification Algorithms. *Ejosat*, *24*, 53–59. https://doi.org/https://doi.org/10.31590/ejosat.899716

Shetty, R., Paul, A., Kumar, A. P., & Rai, S. (2026). Early Prediction of Diabetes Using Interpretable Machine Learning Framework. *Discover Artificial Intelligence*, *400*. https://doi.org/https://doi.org/10.1007/s44163-026-01152-z

Siregar, H. A., Raditya, M. Z., Yesa, A. N., & Permana, I. (2023). Comparison of Classification Algorithm Performance for Diabetes Prediction Using Orange Data Mining. *Indonesian Journal of Data and Science*, *4*(3), 176–182. https://doi.org/https://doi.org/10.56705/ijodas.v4i3.103

Sudestra, I. M. A., Gama, A. W. O., Prathama, G. H., Paramartha, I. G. N. D., & Hakimi, M. (2026). Comparative Performance of Machine Learning Algorithms for Diabetes Prediction. *Journal of Technology and Informatics (JoTI)*, *8*(1), 50–61. https://doi.org/https://doi.org/10.37802/joti.v8i1.1195

Tan, K. R., Seng, J. J. B., Kwan, Y. H., Chen, Y. J., Zainudin, S. B., Loh, D. H. F., Liu, N., & Low, L. L. (2023). Evaluation of Machine Learning Methods Developed for Prediction of Diabetes Complications: A Systematic Review. *Journal of Diabetes Science and Technology*, *17*(2), 474–489. https://doi.org/https://doi.org/10.1177/19322968211056917

Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2022). Diabetes Prediction Using Machine Learning and Explainable AI Technique. *Healthcare Technology Letters*, *10*(2), 1–10. https://doi.org/10.1049/htl2.12039