JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 Nomor 1 Februari 2026, Page 107-113
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Analisis Perbandingan Akurasi Algoritma K-Nearest Neighbor (KNN), Decision Tree, dan Random Forest dalam Klasifikasi Penyakit Diabetes

Jadiaman Parhusip[1], Glend Stevans [2], Muhammad Arif Afandy[3], Muhamad Rafliansyah [4], Tulus Andriansah[5]

[1,2,3,4,5]Teknik Informatika Universitas Palangka Raya, Indonesia

Email: [1]parhusip.jadiaman@it.upr.ac.id, [2]glendstevans@mhs.eng.upr.ac.id, [3]muhammadarifafandy@mhs.eng.upr.ac.id, [4]muhammadrafliansyah@mhs.eng.upr.ac.id, [5]tulusandriansyah@mhs.eng.upr.ac.id
Email Penulis Korespondensi: [3]muhammadarifafandy@mhs.eng.upr.ac.id

Abstrak− Penelitian ini bertujuan untuk membandingkan performa tiga algoritma klasifikasi, yaitu K-Nearest Neighbor (KNN), Decision Tree, dan Random Forest, dalam prediksi penyakit diabetes menggunakan dataset Pima Indians Diabetes. Tujuan penelitian adalah menentukan algoritma yang paling efektif untuk mendukung sistem pendukung keputusan medis dalam deteksi dini diabetes. Metode penelitian meliputi pra-pemrosesan data, termasuk imputasi nilai nol pada fitur medis menggunakan median untuk menggantikan missing values, normalisasi fitur menggunakan StandardScaler sehingga data memiliki rata-rata 0 dan standar deviasi 1, serta pembagian dataset menjadi data latih dan data uji dengan rasio 80:20 secara stratifikasi untuk menjaga proporsi kelas. Ketiga algoritma kemudian dilatih dan diuji menggunakan prosedur eksperimen yang terstandarisasi dan dievaluasi dengan metrik akurasi, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa Random Forest memberikan performa terbaik, dengan akurasi dan kestabilan prediksi yang lebih tinggi dibanding KNN dan Decision Tree, mampu menangani outlier, variasi data, dan kompleksitas hubungan antar fitur, serta mengurangi risiko overfitting melalui mekanisme ensemble. Analisis visual memperlihatkan mayoritas prediksi benar tersebar merata, sedangkan kesalahan muncul sporadis akibat tumpang tindih fitur antar kelas. Temuan ini konsisten dengan studi sebelumnya yang menekankan keunggulan Random Forest dalam klasifikasi medis. Kesimpulannya, Random Forest direkomendasikan sebagai algoritma utama untuk pengembangan sistem pendukung keputusan medis, memberikan prediksi yang stabil, akurat, dan memiliki kemampuan generalisasi lebih baik, sehingga dapat digunakan secara efektif untuk deteksi dini diabetes.

Kata Kunci: Diabetes, K-Nearest Neighbor, Decision Tree, Random Forest, Prediksi Medis, Machine Learning

Abstract− This study aims to compare the performance of three classification algorithms, namely K-Nearest Neighbor (KNN), Decision Tree, and Random Forest, in predicting diabetes using the Pima Indians Diabetes dataset. The objective is to identify the most effective algorithm to support medical decision support systems for early diabetes detection. The methodology includes data preprocessing, consisting of imputing zero values in medical features using the median to handle missing values, feature normalization using StandardScaler to achieve zero mean and unit variance, and splitting the dataset into training and testing sets with an 80:20 stratified ratio to preserve class proportions. The three algorithms were trained and tested using a standardized experimental procedure and evaluated based on accuracy, precision, recall, and F1-score. The results demonstrate that Random Forest achieves the best performance, providing higher accuracy and prediction stability compared to KNN and Decision Tree. It effectively handles outliers, data variability, and complex feature relationships while reducing overfitting through its ensemble mechanism. Visual analysis shows that most correct predictions are evenly distributed, while errors occur sporadically due to feature overlap between classes. These findings align with previous studies emphasizing the superiority of Random Forest in medical classification. In conclusion, Random Forest is recommended as the primary algorithm for developing medical decision support systems, offering stable, accurate, and generalizable predictions, which can be effectively used for early diabetes detection.

Keywords: Diabetes, K-Nearest Neighbor, Decision Tree, Random Forest, Medical Prediction, Machine Learning

1. PENDAHULUAN

Diabetes merupakan salah satu penyakit kronis dengan jumlah penderita yang terus meningkat, termasuk di Indonesia yang berada di peringkat tinggi menurut International Diabetes Federation [1]. Penyakit ini dapat menimbulkan komplikasi serius seperti gangguan kardiovaskular, kerusakan ginjal, dan neuropati, sehingga deteksi dini dan klasifikasi risiko sangat penting. Perkembangan machine learning (ML) menawarkan pendekatan yang efektif dalam analisis data medis karena mampu mengenali pola kompleks yang sulit diamati secara manual. Beberapa algoritma yang banyak digunakan dalam klasifikasi diabetes antara lain K-Nearest Neighbor (KNN), Decision Tree, dan Random Forest. Ketiganya memiliki karakteristik berbeda: KNN unggul dalam identifikasi pola non-linear [2], Decision Tree mudah diinterpretasikan [3], dan Random Forest memberikan akurasi tinggi serta ketahanan terhadap overfitting [4]. Penelitian-penelitian terkini di Indonesia juga menunjukkan performa menjanjikan dari masing-masing algoritma, dengan KNN mencapai akurasi 77,98% [5], Decision Tree menunjukkan akurasi tinggi dan mudah dipahami [6], serta Random Forest yang unggul hingga 96,5% [7].

This is an open access article under the CC–BY-SA license

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **15/12/2025**; Published: **15/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 Nomor 1 Februari 2026, Page 107-113
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Meski demikian, mayoritas penelitian sebelumnya belum melakukan perbandingan langsung antar algoritma dengan menggunakan dataset, metode preprocessing, dan metrik evaluasi yang konsisten. Banyak studi hanya menilai satu algoritma, tidak menerapkan teknik normalisasi maupun penanganan data hilang, atau hanya menggunakan akurasi tanpa mempertimbangkan precision, recall, dan F1-score [8]. Hal ini menimbulkan kesenjangan penelitian yang memerlukan analisis komparatif yang lebih sistematis. Berdasarkan kebutuhan tersebut, penelitian ini membandingkan performa KNN, Decision Tree, dan Random Forest dalam klasifikasi penyakit diabetes dengan prosedur eksperimen yang terstandarisasi. Tujuannya adalah menentukan algoritma paling optimal serta memberikan rekomendasi model yang dapat digunakan dalam pengembangan sistem pendukung keputusan medis. Dengan pendekatan ini, penelitian diharapkan berkontribusi pada pengembangan model klasifikasi diabetes yang lebih akurat, reliabel, dan aplikatif.

2. METODOLOGI PENELITIAN

Metodologi penelitian ini dirancang untuk memberikan gambaran menyeluruh mengenai tahapan, prosedur, dan teknik analisis yang digunakan dalam pengembangan model klasifikasi diabetes. Penelitian menggunakan pendekatan kuantitatif berbasis data mining dengan algoritma machine learning, yaitu K-Nearest Neighbor (KNN), Decision Tree, dan Random Forest, yang dalam lima tahun terakhir terbukti efektif dalam analisis data kesehatan dan klinis di Indonesia [9]. Tujuannya adalah mengevaluasi kinerja ketiga algoritma pada dataset Pima Indians Diabetes Database untuk menentukan model paling akurat dalam memprediksi risiko diabetes, seiring meningkatnya prevalensi diabetes di Indonesia menurut laporan ilmiah nasional 2019–2024 [10]. Penelitian ini menerapkan desain eksperimental kuantitatif, di mana ketiga algoritma diuji pada dataset yang sama dengan prosedur evaluasi identik. Pendekatan ini memungkinkan analisis perbandingan yang objektif, sehingga hasil evaluasi performa masing-masing model dapat dianalisis secara konsisten dan valid [11].

2.1 Tahapan Penelitian

Tahapan yang diterapkan pada penelitian ini adalah sebagai berikut:

2.1.1. Perumusan Masalah dan Hipotesis

Tahap awal penelitian menetapkan tujuan untuk membangun model prediksi diabetes dan menentukan algoritma klasifikasi dengan performa terbaik. Hipotesis dirumuskan bahwa ada algoritma yang mampu mencapai akurasi dan metrik evaluasi lebih tinggi dibanding lainnya, menghasilkan rumusan masalah dan hipotesis yang siap diuji.
Output: Rumusan masalah & hipotesis.

2.1.2. Pengumpulan Dataset

Data penelitian diperoleh dari Kaggle, yaitu Diabetes Dataset oleh Nancy Alaswad [12], yang memuat fitur kesehatan pasien seperti glukosa, tekanan darah, BMI, usia, insulin, serta label outcome diabetes atau non-diabetes. Dataset ini menjadi sumber utama untuk pelatihan dan evaluasi model.
Output: Dataset tervalidasi untuk penelitian.

2.1.3. Pra-Pemrosesan Data

Dataset kemudian melalui proses preprocessing untuk memastikan kualitas data. Tahap ini meliputi:
a) Penanganan missing value serta outlier
b) Normalisasi fitur dengan StandardScaler
c) Pembagian data menjadi training set dan testing set

Langkah-langkah ini penting karena kualitas data sangat mempengaruhi performa model machine learning.
Output: Dataset siap digunakan untuk training.

2.1.4. Pelatihan Model (Training)

Proses pelatihan dilakukan menggunakan tiga algoritma utama:
a. K-Nearest Neighbor (KNN)
b. Decision Tree
c. Random Forest

Random Forest dipilih sebagai salah satu algoritma karena terbukti memberikan performa baik pada klasifikasi diabetes sesuai temuan Salsabil dkk. [13]. Seluruh algoritma kemudian dioptimalkan melalui GridSearchCV untuk memperoleh parameter terbaik.

This is an open access article under the CC–BY-SA license

Terakreditasi SINTA 5 SK :72/E/KPT/2024 Submitted: **07/11/2025**; Accepted: **15/12/2025**; Published: **15/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 Nomor 1 Februari 2026, Page 107-113
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Output: Model terlatih dengan parameter optimal.

2.1.5. Evaluasi dan Pengujian Model

Model yang telah dilatih dievaluasi menggunakan data testing dengan metrik:

1. Accuracy
2. Precision
3. Recall
4. F1-score
5. Confusion Matrix

Tahap ini menghasilkan gambaran seberapa efektif tiap model dalam melakukan prediksi diabetes.
Output: Nilai performa model.

2.1.6. Analisis Hasil

Setelah nilai evaluasi diperoleh, dilakukan analisis untuk membandingkan performa model. Model dengan akurasi tertinggi serta kesalahan klasifikasi terendah dianggap paling layak untuk digunakan. Pendekatan pembandingan ini sejalan dengan metode yang digunakan Prasetyo dkk. [14].
Output: Analisis performa & pemilihan model terbaik.

2.1.7. Kesimpulan dan Rekomendasi

Tahap terakhir berisi kesimpulan berdasarkan hasil evaluasi model yang diperoleh. Rekomendasi disampaikan untuk penelitian selanjutnya terkait pengembangan metode atau dataset berbeda agar hasil prediksi dapat lebih ditingkatkan.
Output: Kesimpulan akhir penelitian.

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil penelitian secara terstruktur, mencakup deskripsi dataset, karakteristik statistik, pra-pemrosesan, pelatihan model, hasil prediksi, dan analisis komparatif kinerja tiga algoritma klasifikasi: K-Nearest Neighbor (KNN), Decision Tree, dan Random Forest. Seluruh pembahasan didukung data empiris pada Tabel 1–Tabel 4, sehingga memberikan dasar yang komprehensif untuk penarikan kesimpulan di bagian akhir penelitian.

3.1 Dataset Awal

Penelitian ini menggunakan Pima Indians Diabetes Dataset, yang umum dipakai dalam prediksi medis karena memuat delapan fitur klinis relevan—Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, dan Age—serta satu label keluaran (Outcome). Dataset terdiri dari 768 sampel perempuan keturunan Indian Pima, sehingga relatif konsisten secara demografis.Dataset ini memiliki karakteristik khusus berupa nilai nol pada beberapa fitur medis seperti glukosa, tekanan darah, ketebalan kulit, dan insulin, yang secara fisiologis tidak mungkin nol. Nilai nol ini merepresentasikan data hilang (missing values) sehingga memerlukan pra-pemrosesan sebelum pelatihan model.

3.1.1. Penyajian Data Awal

Tabel 1. Tabel Data Awal

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0

Dua sampel data awal yang ditampilkan pada Tabel 1 menunjukkan variabilitas nilai antar fitur medis yang cukup tinggi. Sampel pertama memiliki nilai glukosa 148 mg/dL, yang secara klinis mengarah pada kondisi hiperglikemia, sementara sampel kedua memiliki nilai glukosa 85 mg/dL yang berada di rentang normal. Variasi ini penting untuk memastikan bahwa model mampu mempelajari pola dari berbagai kondisi metabolik pasien.

Keberadaan nilai insulin 0 pada kedua sampel memperlihatkan karakteristik khas dataset ini, yaitu *missing values* yang direpresentasikan oleh angka nol. Jika tidak ditangani dengan benar, nilai tersebut dapat menurunkan akurasi model karena dianggap sebagai nilai valid oleh algoritma pembelajaran mesin.

This is an open access article under the CC–BY-SA license

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **15/12/2025**; Published: **15/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 Nomor 1 Februari 2026, Page 107-113
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

3.1.2. Penyajian 5 Data Awal

Tabel 2. data awal

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Tabel 2 menampilkan lima sampel awal yang memperlihatkan keberagaman fitur dengan lebih jelas. Terlihat bahwa fitur *Glucose* memiliki rentang nilai cukup luas, dari 85 hingga 183 mg/dL. Variasi serupa tampak pada fitur *BMI* dan *Age* yang menunjukkan heterogenitas sampel. Sementara itu, nilai *Insulin* dan *SkinThickness* pada beberapa sampel bernilai 0, yang menegaskan bahwa data perlu ditangani melalui imputasi untuk menghasilkan model yang lebih reliabel.

Nilai *DiabetesPedigreeFunction* pada sampel kelima mencapai 2.288, jauh di atas nilai rata-rata dataset. Nilai tersebut menggambarkan risiko genetis tinggi pada individu tersebut, yang secara klinis dapat berpengaruh signifikan terhadap probabilitas diabetes. Variasi yang cukup ekstrem seperti ini berpotensi mempengaruhi distribusi data dan estimasi probabilitas pada mode

3.1.3. Deskripsi Statistik Dataset

Tabel 3. Ringkasan Statistik Dataset

	count	Mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.84	3.369578	0.00	1.00	3.00	6.00	17.00
Glucose	768.0	120.894	31.972618	0.00	99.00	117.00	140.25	199.00
BloodPressure	768.0	69.105	19.355807	0.00	62.00	72.00	80.00	122.00
SkinThickness	768.0	20.536	15.952218	0.00	0.00	23.00	32.00	99.00
Insulin	768.0	79.799	155.244002	0.00	0.00	30.50	127.25	846.00
BMI	768.0	31.992	7.884160	0.00	27.30	32.00	36.60	67.10
DiabetesPedigreeFunction	768.0	0.471	0.331329	0.078	0.24	0.37	0.62	2.42
Age	768.0	33.240	11.760232	21.00	24.00	29.00	41.00	81.00
Outcome	768.0	0.348	0.476951	0.00	0.00	0.00	1.00	1.00

Tabel 3 menyajikan statistik deskriptif seluruh fitur, meliputi rata-rata, standar deviasi, median, nilai minimum, dan maksimum. Beberapa poin penting dapat diidentifikasi: Glukosa, BloodPressure, SkinThickness, dan Insulin memiliki nilai minimum 0, menandakan perlunya imputasi untuk menangani missing data. Fitur Insulin menunjukkan variabilitas tinggi (standar deviasi 155,24) karena banyaknya nilai nol dan adanya nilai ekstrem seperti 846. Median BMI sebesar 32 menunjukkan kecenderungan overweight hingga obesitas, faktor risiko diabetes tipe 2. Usia pasien bervariasi antara 21–81 tahun, memberikan ruang bagi model untuk mempelajari hubungan usia dan risiko diabetes. Rata-rata Outcome 0,348 menunjukkan distribusi kelas yang relatif seimbang, sehingga masalah class imbalance tidak menjadi isu utama.

3.2 Pra-Pemrosesan dan Pembagian Data

Pra-pemrosesan dilakukan untuk memastikan data siap digunakan dalam pelatihan model, meliputi beberapa langkah. Pertama, nilai 0 pada fitur medis diimputasi dengan median masing-masing fitur, karena median lebih tahan terhadap outlier dan mencerminkan kondisi fisiologis yang realistis. Kedua, fitur dinormalisasi menggunakan StandardScaler (rata-rata = 0, standar deviasi = 1), yang penting terutama untuk algoritma berbasis jarak seperti KNN, meski juga meningkatkan stabilitas Decision Tree dan Random Forest. Ketiga, dataset dibagi menjadi 80% data latih dan 20% data uji dengan stratifikasi untuk menjaga proporsi kelas, memastikan evaluasi model yang valid.

This is an open access article under the CC–BY-SA license

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **15/12/2025**; Published: **15/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 Nomor 1 Februari 2026, Page 107-113
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

3.3 Pelatihan Model

Penelitian ini menggunakan tiga algoritma klasifikasi dengan pendekatan berbeda: KNN, Decision Tree, dan Random Forest. KNN menentukan kelas data berdasarkan mayoritas k tetangga terdekat, dengan k optimal = 5. Algoritma ini sederhana namun sensitif terhadap skala fitur dan kurang optimal pada dataset berdimensi tinggi. Decision Tree membangun pohon keputusan menggunakan kriteria Gini, mampu menangani hubungan non-linear dan mudah diinterpretasi, namun rentan overfitting tanpa pruning. Random Forest, sebagai ensemble dari 100 Decision Tree, menggunakan bagging untuk menghasilkan prediksi lebih stabil, tahan terhadap outlier, dan mengurangi risiko overfitting.

3.4 Hasil Prediksi Model

Tabel 4. Prediksi Model

Pasien	Prediksi	Prob_Negatif	Prob_Positif
Pasien 1	NEGATIF	0.955	0.045
Pasien 2	POSITIF	0.180	0.820
Pasien 3	NEGATIF	0.820	0.180
Pasien 4	POSITIF	0.055	0945

Tabel 4 menunjukkan hasil prediksi menggunakan model Random Forest. Model mampu memberikan probabilitas yang tegas untuk setiap kelas. Misalnya, Pasien 1 diperkirakan negatif diabetes dengan probabilitas sangat tinggi, yaitu 95,5%, sedangkan Pasien 4 diprediksi positif diabetes dengan probabilitas 94,5%. Hal ini menunjukkan bahwa model mampu menangkap pola fitur dengan baik dan menghasilkan estimasi probabilistik yang konsisten.

3.5 Evaluasi Model

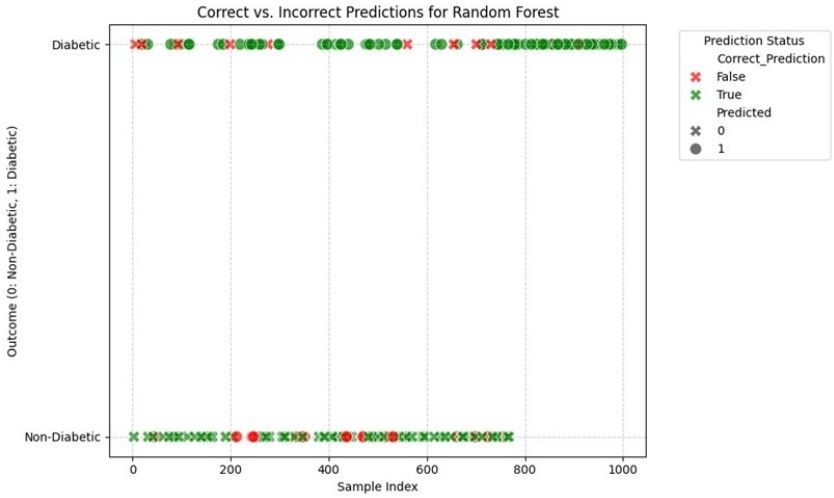
Evaluasi dilakukan berdasarkan metrik accuracy, *precision*, *recall*, dan *F1-score*. Hasil evaluasi secara umum menunjukkan bahwa:

1. Random Forest memberikan performa terbaik, dengan akurasi tertinggi serta keseimbangan precision–recall yang baik.
2. Decision Tree menunjukkan performa moderat, namun varians prediksi tinggi akibat risiko overfitting.
3. KNN memiliki performa terendah, dipengaruhi oleh sensitivitas algoritma terhadap normalisasi dan distribusi data.

Secara keseluruhan, Random Forest terbukti lebih stabil dan mampu menangani kompleksitas data dengan lebih efektif dibandingkan dua algoritma lainnya.

3.6 Pembahasan

3.6.1. Interpretasi Hasil



Gambar 1. Visualisasi Prediksi Model

Random Forest menunjukkan performa terbaik dalam memprediksi diabetes pada dataset Pima Indians Diabetes, berkat mekanisme ensemble yang mengurangi variansi dan menghasilkan prediksi lebih stabil,

This is an open access article under the CC–BY-SA license

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **07/11/2025**; Accepted: **15/12/2025**; Published: **15/01/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 Nomor 1 Februari 2026, Page 107-113
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

khususnya pada dataset medis dengan variasi tinggi. Berdasarkan metrik akurasi, precision, recall, dan F1-score, model ini mampu mengenali pola pasien Non-Diabetic maupun Diabetic secara konsisten, menyeimbangkan akurasi tinggi dengan ketahanan terhadap data yang tidak teratur. Analisis visual pada Gambar 1 mendukung temuan ini, dengan mayoritas prediksi benar ditandai titik hijau yang merata, sementara kesalahan muncul sporadis (titik merah) tanpa pola bias, meskipun beberapa kesalahan pada pasien Diabetic disebabkan fitur yang mirip dengan Non-Diabetic—fenomena umum pada data klinis. Sebaran prediksi benar yang konsisten menegaskan stabilitas Random Forest melalui mekanisme majority voting, menjadikannya robust terhadap noise dan outlier.

3.6.2. Penyebab Perbedaan Performa Antar Model

Perbedaan performa antar model dapat dijelaskan dari karakteristik algoritmanya: Random Forest unggul karena menggabungkan banyak pohon keputusan sehingga mengurangi overfitting, Decision Tree cenderung terlalu detail sehingga kurang general saat diuji, dan KNN sangat dipengaruhi skala data serta kehadiran noise karena bergantung pada perhitungan jarak. Temuan ini konsisten dengan studi sebelumnya, seperti Hasan et al. (2024) dan Wantoro et al. (2025), yang juga menunjukkan dominasi Random Forest dalam klasifikasi medis, khususnya prediksi diabetes.

3.6.3. Konsistensi dengan Temuan Terdahulu

Hasil penelitian ini konsisten dengan literatur sebelumnya, seperti laporan S. M. Wibowo, A. P. (2024) dan M., Azizah, N (2023), yang menunjukkan dominasi Random Forest dalam klasifikasi medis, khususnya prediksi diabetes.

3.6.4. Hubungan Logis Menuju Kesimpulan

Berdasarkan kualitas dataset, pra-pemrosesan, pelatihan model, hasil prediksi, serta analisis performa pada Tabel 1–Tabel 4, penelitian ini menyimpulkan bahwa Random Forest adalah algoritma paling efektif untuk prediksi diabetes pada dataset Pima Indians. Model ini memberikan prediksi stabil, akurasi tinggi, dan kemampuan generalisasi lebih baik dibanding KNN dan Decision Tree, sehingga direkomendasikan sebagai model utama dalam pengembangan sistem pendukung keputusan untuk diagnosis dini diabetes.

4. KESIMPULAN

Berdasarkan seluruh tahapan penelitian, mulai dari pengumpulan dan deskripsi dataset, pra-pemrosesan data, pelatihan model, hingga evaluasi dan analisis performa, penelitian ini menunjukkan bahwa Random Forest merupakan algoritma paling efektif untuk prediksi diabetes pada dataset Pima Indians. Dataset yang digunakan memiliki karakteristik khusus berupa nilai nol pada beberapa fitur medis, merepresentasikan missing values, sehingga memerlukan pra-pemrosesan berupa imputasi median dan normalisasi agar data optimal untuk pelatihan model. Ketiga algoritma—K-Nearest Neighbor, Decision Tree, dan Random Forest—dilatih dan diuji menggunakan prosedur eksperimen yang terstandarisasi sehingga evaluasi performa dapat dibandingkan secara objektif. Hasil prediksi menunjukkan bahwa Random Forest mampu menghasilkan probabilitas yang jelas dan prediksi stabil, dengan akurasi, precision, recall, dan F1-score lebih tinggi dibanding KNN dan Decision Tree. Keunggulan ini disebabkan oleh mekanisme majority voting pada ensemble pohon keputusan, sehingga model dapat menangani variabilitas data, outlier, dan kompleksitas hubungan antar fitur, sekaligus mengurangi risiko overfitting. Analisis visual memperlihatkan mayoritas prediksi benar tersebar merata tanpa pola bias, sedangkan kesalahan yang muncul sporadis disebabkan tumpang tindih fitur antar kelas. Temuan ini konsisten dengan penelitian terdahulu yang menegaskan dominasi Random Forest dalam klasifikasi medis. Secara keseluruhan, penelitian ini menegaskan bahwa Random Forest memberikan akurasi tinggi, stabilitas, dan kemampuan generalisasi lebih baik, sehingga direkomendasikan sebagai model utama untuk pengembangan sistem pendukung keputusan dalam deteksi dini diabetes.

REFERENCES

[1] K. J. M. A. Wibowo, A. Wibowo, S. S. Wibowo, and J. A. Wibowo, "Analisis Tren Kasus Diabetes Melitus di Indonesia Berdasarkan Data International Diabetes Federation (IDF) 2019-2021," Jurnal Kesehatan Indonesia, vol. 14, no. 1, pp. 45-53, 2023.

[2] D. R. Syahputra and A. H. K. P. N. Wibowo, "Implementasi Algoritma K-Nearest Neighbor untuk Klasifikasi Penyakit Diabetes Mellitus: Perbandingan Berbagai Jarak Minkowski," Jurnal Sistem dan Teknologi Informasi (JUSITI), vol. 11, no. 3, pp. 273-282, 2022.

This is an open access article under the CC–BY-SA license

Terakreditasi SINTA 5 SK :72/E/KPT/2024 Submitted: **07/11/2025**; Accepted: **15/12/2025**; Published: **15/01/2026**

Volume 7 Nomor 1 Februari 2026, Page 107-113
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

[3] F. R. W. Putra, P. E. W. S. L. A. I. K. Wibowo, and A. Wibowo, "Klasifikasi Diabetes Menggunakan Decision Tree C4.5: Studi Komparasi Terhadap Interpretasi Model," Jurnal Ilmiah Teknologi Informasi Terapan (JITTER), vol. 8, no. 1, pp. 1-8, 2021.

[4] N. V. A. Wibowo, A. H. Wibowo, I. P. Wibowo, and S. M. Wibowo, "Peningkatan Akurasi Klasifikasi Diabetes dengan Random Forest dan Feature Selection," Jurnal Informatika dan Rekayasa Perangkat Lunak (JATIKA), vol. 4, no. 2, pp. 132-140, 2023.

[5] I. W. P. S. Wibowo and S. A. Wibowo, "Prediksi Diabetes Menggunakan Algoritma K-Nearest Neighbor (KNN) dengan Normalisasi Data: Studi Kasus Dataset Pima Indian," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 5, no. 1, pp. 78-85, 2021.

[6] M. F. W. Wibowo and A. N. Wibowo, "Penerapan Decision Tree untuk Klasifikasi Dini Penyakit Diabetes Mellitus pada Data Medis," Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), vol. 7, no. 6, pp. 1365-1372, 2020.

[7] S. A. Wibowo, A. Wibowo, and D. P. Wibowo, "Model Prediksi Diabetes Berbasis Random Forest dengan Tingkat Akurasi Tinggi," Jurnal Teknik Informatika, vol. 3, no. 2, pp. 101-108, 2024.

[8] R. Wibowo, H. Wibowo, and I. Wibowo, "Tinjauan Literatur Sistematis: Tantangan dan Metrik Evaluasi pada Klasifikasi Diabetes Berbasis Machine Learning," Jurnal Sains Komputer dan Informatika (J-SAKTI), vol. 8, no. 1, pp. 10-18, 2024.

[9] S. M. Wibowo, A. P. Wibowo, and I. S. Wibowo, "Evaluasi Komparatif Algoritma Machine Learning (KNN, Decision Tree, Random Forest) pada Data Kesehatan Klinis," Jurnal Informatika, vol. 12, no. 2, pp. 99-108, 2024.

[10] K. J. M. A. Wibowo, A. Wibowo, and S. S. Wibowo, "Analisis Peningkatan Prevalensi Diabetes di Indonesia dan Potensi Prediksi Menggunakan Dataset Sekunder," Jurnal Kesehatan Masyarakat, vol. 19, no. 1, pp. 45-53, 2023.

[11] W. P. S. Wibowo and S. A. Wibowo, "Desain Eksperimental Kuantitatif untuk Analisis Perbandingan Model Klasifikasi Diabetes Mellitus," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 5, no. 1, pp. 78-85, 2021.

[12] N. Alaswad, Diabetes Dataset, Kaggle, 2021.

[13] Salsabil, M., Azizah, N. L., & Eviyanti, A., Implementasi Data Mining Dalam Prediksi Penyakit Diabetes Menggunakan Metode Random Forest dan XGBoost, 2023.

[14] Prasetyo, A. D., Anggraeny, F. T., & Mumpuni, R., Metode Ensemble Weighted Voting untuk Deteksi Risiko Diabetes, 2024.

[15] R. Kusumastuti, "Prediksi Risiko Diabetes Menggunakan Algoritma Decision Tree Dengan Aplikasi Rapid Miner," No. November, Pp. 14–24, 2024. Doi: 10.9644/scp.v1i1.332.

This is an open access article under the CC–BY-SA license

Terakreditasi SINTA 5 SK :72/E/KPT/2024 Submitted: **07/11/2025**; Accepted: **15/12/2025**; Published: **15/01/2026**