JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 Nomor 3 Juni 2026, Page 305-312
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Perbandingan Kinerja Klasifikasi Data Mining untuk Diagnosis Diabetes

Andhika Nur Fachmi[1], Reihan Dwi Patria[2], Ferdinan Restu Ramadhan[3], Salma Pudjiati[4], Anisa Permata Sari[5]

[1,2,3,4,5)]Informatika, Universitas Bina Sarana Informatika, Depok, Indonesia
Email: [1]15230419@bsi.ac.id, [2]15230546@bsi.ac.id, [3]15230423@bsi.ac.id, [4]15230622@bsi.ac.id, [5]15230465@bsi.ac.id
Email Penulis Korespondensi: [1]15230419@bsi.ac.id

Abstrak− Sebagai penyakit metabolik yang bersifat kronis, Diabetes Melitus memerlukan upaya deteksi sejak dini untuk menekan kemungkinan terjadinya komplikasi. Mengingat proses diagnosis manual masih menghadapi keterbatasan waktu dan biaya, penelitian ini dilakukan untuk mengembangkan model prediksi berbasis data dengan memanfaatkan Diabetes Health Indicators Dataset. Fokus utama penelitian ini adalah membandingkan kinerja dua algoritma *Machine Learning*, yaitu *Random Forest* dan *Decision Tree*. Pendekatan klasifikasi diterapkan melalui tahapan pra-pemrosesan data yang sistematis untuk mencegah terjadinya kebocoran informasi (*data leakage*). Tahapan tersebut meliputi standarisasi fitur, pengkodean variabel, serta penanganan ketidakseimbangan kelas menggunakan metode *SMOTE* pada data pelatihan. Model dikembangkan dengan skema pembagian data sebesar 80:20 dan dievaluasi menggunakan metrik Akurasi, Presisi, *Recall* dan *F1-Score*. Hasil eksperimen menunjukkan bahwa kedua algoritma memiliki performa yang relatif sebanding. *Random Forest* menunjukkan keunggulan tipis dengan nilai Akurasi sebesar 91,98% dan *F1-Score* sebesar 92,84%. Nilai tersebut sedikit lebih tinggi dibandingkan *Decision Tree* yang memperoleh Akurasi 91,84% dan *F1-Score* 92, 74%. Analisis fitur lebih lanjut menunjukkan bahwa HbA1c dan kadar glukosa darah merupakan variabel yang paling dominan dalam proses prediksi. Dengan demikian, dapat disimpulkan bahwa meskipun kedua algoritma menunjukkan kinerja yang reliabel, *Random Forest* menawarkan stabilitas prediksi yang lebih baik pada kasus ini.

Kata Kunci: Data Mining, Klasifikasi, *Random Forest*, *Decision Tree*, Prediksi Diabetes, SMOTE

Abstract− As a chronic metabolic disease, diabetes mellitus requires early detection to reduce the risk of complications. Given that manual diagnostic processes still face time and cost constraints, this research aimed to develop a data-driven predictive model utilizing the Diabetes Health Indicators Dataset. The main focus of this research is to compare the performance of two machine learning algorithms: Random Forest and Decision Tree. The classification approach is implemented through systematic data preprocessing stages to prevent data leakage. These stages include feature standardization, variable coding, and handling class imbalance using the SMOTE method on the training data. The model was developed with an 80:20 data split scheme and evaluated using Accuracy, Precision, Recall, and F1-Score metrics. The experimental results show that both algorithms have relatively comparable performance. Random Forest shows a slight advantage with an accuracy value of 91.98% and an F1-score of 92.84%. These values are slightly higher than Decision Tree which obtained an accuracy of 91.84% and an F1-score of 92.74%. Further feature analysis shows that HbA1c and blood glucose levels are the most dominant variables in the prediction process. Thus, it can be concluded that although both algorithms show reliable performance, Random Forest offers better prediction stability in this case.

Keywords: Data Mining, Classification, Random Forest, Decision Tree, Diabetes Prediction, SMOTE

1. PENDAHULUAN

Di era kesehatan modern, Diabetes Melitus (DM) menjadi salah satu masalah global dengan dampak besar terhadap sistem layanan kesehatan. Penyakit ini ditandai oleh tingginya kadar glukosa darah secara berkelanjutan akibat gangguan produksi maupun respons tubuh terhadap insulin. Sebagaimana dielaborasi dalam studi [1], dampak lanjutan dari DM sangatlah krusial karena berpotensi memicu kerusakan pada organ vital, termasuk gagal jantung dan nefropati [1]. Data dari *International Diabetes Federation* (IDF) pun menunjukkan tren peningkatan prevalensi yang terus mengkhawatirkan secara global. Mengingat risiko fatalitas dan tingginya biaya perawatan jangka panjang, upaya deteksi dini menjadi sangat krusial dalam manajemen penyakit ini. Laporan terbaru dari *International Diabetes Federation* (IDF) sekali lagi mengonfirmasi kecenderungan dunia untuk lebih khawatir tentang meningkatnya kasus penyakit ini. Oleh karena itu, strategi deteksi dini menjadi pilar fundamental dalam manajemen penyakit ini guna menekan risiko fatalitas serta beban biaya perawatan jangka panjang.

Perkembangan pesat di era *Big Data* saat ini membuka peluang pemanfaatan data kesehatan pasien, seperti rekam medis elektronik dan hasil laboratorium, untuk membangun model prediksi berbasis *Data Mining* dan *Machine Learning*. Teknologi ini memungkinkan ekstraksi pola-pola prediktif yang tersembunyi untuk membantu tenaga medis dalam pengambilan keputusan klinis yang lebih cepat dan akurat [2]. Dalam konteks klasifikasi medis, pemilihan algoritma menjadi faktor penentu keberhasilan model. Dua algoritma yang sering dibandingkan efektivitasnya adalah *Decision Tree* (DT) dan *Random Forest* (RF).

Decision Tree atau Pohon Keputusan merupakan algoritma klasifikasi yang mudah dipahami karena struktur pohonnya yang mampu merepresentasikan proses pengambilan keputusan secara jelas. Menurut Fadhlullah et al. [3], keunggulan utama dari *Decision Tree* adalah efisiensi waktu komputasi. Namun, algoritma ini juga memiliki kelemahan klasik, yaitu rentan terhadap overfitting. Di sisi lain, *Random Forest* merupakan metode ensemble yang menggabungkan banyak pohon keputusan untuk mendapatkan prediksi yang lebih stabil. Penelitian dari Esmaily et al. [4] dan Desmita et

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

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **19/10/2025**; Accepted: **29/04/2026**; Published: **16/06/2026**

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 Nomor 3 Juni 2026, Page 305-312
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

al. [1] menyimpulkan bahwa mekanisme ensemble membuat *Random Forest* lebih tahan terhadap noise dan secara konsisten mampu memberikan tingkat akurasi yang lebih tinggi dibandingkan dengan metode tunggal.

Namun, perdebatan mengenai algoritma mana yang paling optimal masih terus berlanjut karena hasil penelitian terdahulu menunjukkan temuan yang beragam. Siswoyo et al. [5] dalam studinya menyoroti bahwa performa algoritma sangat bergantung pada karakteristik distribusi data dan parameter model yang digunakan. Sementara itu, studi komparasi oleh Wantoro et al. [6] menemukan bahwa penggunaan seleksi fitur dapat mempengaruhi akurasi secara signifikan, di mana algoritma seperti SVM justru mengungguli metode berbasis pohon dalam konteks data tertentu. Di sisi lain, Ibrahim et al. [7] mencatat bahwa meskipun algoritma *ensemble* seperti *Gradient Boosting* memberikan AUC tertinggi, tantangan dalam nilai *Recall* untuk kelas positif masih sering ditemukan pada dataset yang tidak seimbang. Ringkasan penelitian-penelitian tersebut dapat dilihat pada Tabel 1.

Tabel 1. Ringkasan Penelitian Terkait

Peneliti (Tahun)	Metode	Temuan Utama
Iqbal (2023) [5]	KNN, Logistic Regression, DT	Performa algoritma sangat bergantung pada karakteristik distribusi data.
Wantoro et al. (2025) [6]	SVM, Naive Bayes, RF	Seleksi fitur dapat mempengaruhi akurasi; SVM unggul dalam studi ini.
Ibrahim et al. (2025) [7]	Gradient Boosting, RF, SVC	Gradient Boosting memberikan AUC tertinggi, namun Recall masih menjadi tantangan.
Ekrinifda et al. (2025) [8]	RF vs DT	RF lebih unggul dalam menangani data kompleks dan mengurangi *overfitting*.

Penelitian spesifik mengenai Random Forest oleh Setiawan et al. [9] dan Ekrinifda et al. [8] memperkuat argumen bahwa RF memiliki keunggulan dalam menangani banyak fitur klinis sekaligus mengurangi risiko overfitting yang sering dialami DT. Meskipun demikian, Fadhlullah et al. [3] menemukan perbedaan akurasi yang sangat tipis (tidak signifikan) antara RF dan DT, yaitu 93.97% berbanding 93.89%, sehingga efisiensi komputasi DT mungkin lebih disukai dalam kondisi sumber daya terbatas. Kesenjangan hasil tersebut menunjukkan perlunya pengujian empiris lebih lanjut pada dataset yang berbeda.

Selain perdebatan mengenai pemilihan algoritma, isu krusial yang sering luput dari perhatian dalam penelitian komparasi sebelumnya adalah penanganan terhadap ketidakseimbangan kelas (class imbalance) pada dataset medis. Banyak studi terdahulu yang melaporkan akurasi tinggi, namun gagal mengungkapkan bahwa model mereka sebenarnya mengalami bias yang signifikan terhadap kelas mayoritas (pasien non-diabetes). Dalam konteks diagnosis medis, bias semacam ini sangat berbahaya karena meningkatkan risiko False Negative, di mana pasien yang positif diabetes gagal terdeteksi oleh sistem. Oleh karena itu, urgensi penelitian ini tidak hanya terletak pada pembandingan metrik performa antara Random Forest dan Decision Tree semata, melainkan juga pada pembuktian empiris mengenai efektivitas integrasi teknik oversampling SMOTE dalam menciptakan model diagnosis yang lebih adil (fair) dan sensitif terhadap kasus-kasus positif yang seringkali menjadi minoritas dalam data rekam medis.

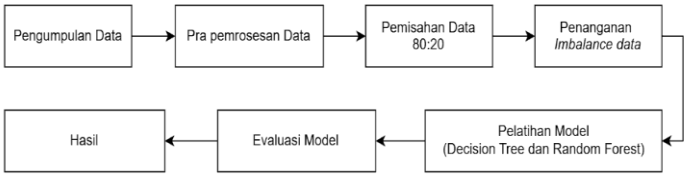
Penelitian ini bertujuan untuk memberikan analisis komparatif yang komprehensif antara kinerja Decision Tree dan Random Forest menggunakan Diabetes Health Indicators Dataset. Fokus utama penelitian tidak hanya pada metrik akurasi semata, tetapi juga mengintegrasikan teknik pra-pemrosesan data yang ketat, termasuk penanganan imbalanced data menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE), untuk memastikan model tidak bias. Dengan demikian, penelitian ini diharapkan dapat memberikan rekomendasi empiris mengenai algoritma yang paling efektif dan andal untuk diaplikasikan sebagai sistem pendukung keputusan diagnosis diabetes.

2. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan eksperimen yang dilakukan secara sistematis, mulai dari tahap pengumpulan data hingga evaluasi model.

2.1 Desain Alur Penelitian

Proses penelitian ini dirancang agar data tetap terjaga keasliannya dan model yang dihasilkan benar-benar valid.



Gambar 1. Diagram Alur Metodologi Penelitian

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

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: 19/10/2025; Accepted: 29/04/2026; Published: 16/06/2026

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 Nomor 3 Juni 2026, Page 305-312
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Seperti yang terlihat pada Gambar 1, prosesnya meliputi beberapa tahapan, mulai dari pengumpulan data, pra-pemrosesan, pemisahan data, penanganan masalah imbalance, pelatihan model, hingga tahap evaluasi.

2.2 Objek Penelitian dan Variabel

Penelitian ini menggunakan dataset sekunder bernama Diabetes Health Indicators Dataset [10] yang tersedia di Kaggle, yang mencakup sekitar 100.000 rekam medis. Penelitian ini fokus utama pada upaya memprediksi variabel target, yaitu diagnosed_diabetes, sebuah label biner yang menunjukkan status klinis pasien 0 berarti tidak mengidap diabetes, dan 1 berarti mengalami diabetes.

Adapun variabel prediktor atau fitur yang digunakan berjumlah sebanyak 28 atribut, yang dibagi ke dalam beberapa dimensi utama. Dimensi pertama meliputi profil demografi dan sosio-ekonomi seperti umur (*age*), jenis kelamin (*gender*), etnisitas (*ethnicity*), tingkat pendidikan (*education level*), penghasilan (*income level*). Kedua, pola gaya hidup, meliputi kebiasaan merokok (*smoking*), konsumsi alkohol (*alcohol consumption*), dan tingkat aktivitas fisik (*physical activity*). Selanjutnya, terdapat riwayat medis, seperti hipertensi (*hypertension*) dan penyakit jantung (*heart disease*). Terakhir, ada parameter klinis penting, termasuk *BMI*, *HbA1c*, serta hasil profil glukosa darah baik saat berpuasa maupun setelah makan (*fasting dan postprandial*)

2.3 Pra-Pemrosesan Data (*Data Preprocessing*)

Setelah proses pemisahan data (*data splitting*) selesai, tahap pra-pemrosesan dilakukan untuk memastikan data ujian tetap utuh dan mencegah terjadinya kebocoran informasi (*data leakage*) dari data uji ke dalam model. Teknik transformasi yang digunakan meliputi:

1. **Pembersihan Data (*Data Cleaning*):** Pada tahap ini, fokus utama adalah memastikan keabsahan dan keandalan model dengan cara menghapus atribut-atribut yang berpotensi menyebabkan kebocoran data. Dalam penelitian ini, proses pembersihan dilakukan dengan menghilangkan dua fitur, yaitu *diabetes_risk_score* dan *diabetes_stage*. Penghapusan tersebut sangat penting karena kedua atribut ini mengandung informasi diagnostik turunan atau yang bisa dianggap sebagai "bocoran" jawaban terkait status target.
2. **Standardisasi Fitur (*Feature Scaling*)**: Proses transformasi ini bertujuan untuk menyamakan skala antar-variabel agar model tidak cenderung bias terhadap fitur yang memiliki rentang nilai lebih besar. Metode *StandardScaler* (*Z-score normalization*) diterapkan pada seluruh fitur numerik seperti BMI, usia (*age*), gula darah (*glucose*), dan HbA1c. Dengan begitu, data memiliki rata-rata 0 (nol) dan deviasi standar 1 (satu), sehingga distribusinya lebih seimbang dan memudahkan proses pelatihan model [11], [12].
3. **Konversi Data Kategorikal (*Encoding*):** Dilakukan dengan dua pendekatan berbeda. Pertama, *Ordinal Encoding* digunakan untuk fitur kategorikal yang memiliki tingkatan, seperti tingkat pendidikan (*education_level*) dan tingkat penghasilan (*income_level*). Kedua, *One-Hot Encoding* diterapkan pada fitur kategorikal nominal tanpa urutan, seperti jenis kelamin (*gender*) dan etnisitas (*ethnicity*). Tujuannya adalah agar model tidak salah menginterpretasikan adanya hubungan urutan pada data tersebut.

2.4 Pemisahan dan Penyeimbangan Data (*Data Splitting and Balancing*)

Dataset dibagi menjadi dua subset dengan proporsi 80:20, di mana 80% dijadikan subset pelatihan (*training set*) untuk pembentukan model, sedangkan 20% sisanya merupakan subset pengujian (*testing set*) untuk evaluasi performa. *Stratified Sampling* (*stratify=y*) digunakan untuk menjaga agar proporsi kelas pada data latih dan uji tetap sesuai distribusi aslinya. Selain itu, parameter (*random_state=42*) ditetapkan untuk memastikan *reproducibility* (keterulangan) hasil eksperimen [13], [14].

Untuk mengatasi ketimpangan distribusi data ekstrem dalam kelas target, penelitian ini mengadopsi strategi oversampling menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE). Metode ini bertujuan untuk mengurangi bias model, terutama terhadap kelas mayoritas. Tidak seperti metode duplikasi yang sederhana, SMOTE mengatasi masalah ini dengan membuat data sintetis baru untuk kelas minoritas melalui interpolasi linier antara sampel positif dan salah satu dari *k-nearest neighbors*. Perlu dicatat bahwa proses ini diterapkan hanya pada data latih untuk menjaga keaslian data uji [15], [16].

2.5 Pelatihan Model (*Model Training*)

Pelatihan model ini dilakukan menggunakan data pelatihan yang telah diproses sebelumnya serta penyeimbangan kelas. Dua *algoritma* klasifikasi *supervised learning* telah diterapkan diantaranya:

a. ***Decision Tree*** **(DT):** Algoritma ini membangun model berbentuk struktur pohon dengan memecah data secara rekursif berdasarkan aturan keputusan (*decision rules*) yang dipelajari dari fitur. Model dilatih dengan parameter *random_state=42* untuk memastikan konsistensi hasil pembentukan *node* pada setiap eksperimen [17], [18].

b. ***Random Forest*** **(RF):** Algoritma ini menerapkan konsep *Ensemble Learning* dengan teknik Bagging (*Bootstrap Aggregating*). *Random Forest* membangun banyak pohon keputusan secara paralel dan menggabungkan hasil prediksinya (melalui *voting* mayoritas) untuk meningkatkan akurasi serta mengurangi risiko *overfitting* yang sering terjadi pada pohon keputusan tunggal [8], [9].

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

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: 19/10/2025; Accepted: 29/04/2026; Published: 16/06/2026

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 Nomor 3 Juni 2026, Page 305-312
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

2.1 Skenario Pengujian dan Evaluasi

Evaluasi kinerja model dilakukan secara komprehensif menggunakan *Confusion Matrix*. Berdasarkan matriks tersebut, kinerja diukur menggunakan empat parameter utama untuk memastikan model tidak hanya akurat, tetapi juga andal dalam mendeteksi kasus penyakit.

a. **Akurasi (*Accuracy*)**: Merepresentasikan persentase ketepatan model dalam mengklasifikasikan seluruh data secara benar, baik positif maupun negatif, terhadap total populasi data.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

b. **Presisi (*Precision*):** Mengukur validitas prediksi positif model. Metrik ini krusial untuk menekan angka *False Positive*, yakni memastikan pasien sehat tidak salah terdiagnosis sakit.

$$\text{Pr}\,e\,cision = \frac{TP}{TP+FP} \quad (2)$$

c. ***Recall* (*Sensitivity*):** Menilai sensitivitas model dalam menjaring kembali kasus positif yang sebenarnya ada. Dalam konteks medis, metrik ini vital untuk mencegah lolosnya pasien sakit (*False Negative*).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

d. **F1-Score:** Merupakan rata-rata harmonis antara Presisi dan *Recall*. Parameter ini menjadi acuan performa yang lebih objektif ketika dataset memiliki distribusi kelas yang timpang.

$$F1\text{-}Score = 2 \times \frac{\text{Pr}\,ecision \times Recall}{\text{Pr}\,ecision + Recall} \quad (4)$$

Keterangan: *TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.*

3. HASIL DAN PEMBAHASAN

Pada bagian ini, akan menjelaskan hasil dari eksperimen komputasi dan analisis mendalam tentang performa kedua model, sekaligus menawarkan interpretasi klinis terhadap fitur-fitur yang digunakan.

3.1 Analisis Dampak Pra-Pemrosesan Data

Sebelum masuk ke evaluasi algoritma, penting untuk menyoroti dampak dari tahapan pra-pemrosesan data yang telah dilakukan. Analisis awal terhadap dataset menunjukkan bahwa dataset awal memiliki ketidakseimbangan kelas yang cukup signifikan. Kondisi ini, sebagaimana dijelaskan oleh Ismafillah et al. [15], sangat berisiko menyebabkan model menjadi bias dan cenderung memprediksi kelas mayoritas (Non-Diabetes) saja, sehingga gagal mendeteksi pasien yang sebenarnya sakit. Penerapan metode SMOTE (*Synthetic Minority Over-sampling Technique*) dalam penelitian ini terbukti berhasil menyeimbangkan distribusi kelas menjadi 50:50 pada data latih. Hal ini memberikan dampak langsung pada kemampuan model dalam mengenali pola kelas minoritas, yang tercermin dari nilai *Recall* yang cukup tinggi pada hasil evaluasi akhir.

Visualisasi distribusi kelas target menunjukkan ketimpangan yang ekstrem pada data orisinal, di mana rasio antara kelas non-diabetes dan diabetes mencapai angka yang signifikan. Secara kuantitatif, dominasi kelas mayoritas ini mencapai hampir 85% dari total populasi dataset. Tanpa intervensi, kondisi ini menyebabkan model *machine learning* cenderung mengabaikan kelas minoritas (pasien diabetes) demi mengejar akurasi semu. Namun, pasca-penerapan teknik SMOTE, distribusi data berhasil direkonstruksi menjadi seimbang sempurna (50:50). Proses *oversampling* sintetis ini tidak hanya menggandakan data, tetapi menciptakan variasi vektor fitur baru di ruang dimensi kelas minoritas, yang secara efektif memaksa algoritma untuk mempelajari *decision boundary* yang lebih adil dan presisi bagi kedua kelas.

Penting untuk dicatat bahwa tanpa intervensi penyeimbangan data seperti SMOTE, model klasifikasi pada dataset medis yang tidak seimbang rentan terjebak dalam fenomena *Accuracy Paradox*. Dalam skenario distribusi awal di mana rasio kelas mayoritas sangat dominan, model dapat dengan mudah mencapai akurasi semu yang tinggi hanya dengan memprediksi seluruh sampel sebagai kelas mayoritas (negatif diabetes). Meskipun secara statistik terlihat impresif, model semacam itu tidak memiliki nilai klinis karena gagal mengidentifikasi pasien yang sebenarnya sakit (nilai *Recall* mendekati nol).

Dengan penerapan SMOTE, ruang fitur pada kelas minoritas diperkaya melalui interpolasi sintetis, memaksa algoritma untuk membangun *decision boundary* yang lebih definitif dan tidak bias ke satu sisi. Hal ini terbukti dari lonjakan nilai *Recall* pada kedua model yang mencapai angka di atas 86%, sebuah indikator bahwa model telah berhasil mempelajari karakteristik laten dari pasien diabetes dan tidak sekadar 'menebak' berdasarkan probabilitas mayoritas.

3.2 Evaluasi Komparatif Kinerja Klasifikasi

Hasil evaluasi pada Tabel 2 menunjukkan bahwa performa *Random Forest* (RF) dan *Decision Tree* (DT) tidak berbeda jauh. Secara keseluruhan, *Random Forest* terbukti lebih unggul dengan Akurasi 91.98% dan F1-Score 92.84%, dibandingkan *Decision Tree* yang mencatatkan Akurasi 91.84% dan F1-Score 92.74%.

Tabel 2. Perbandingan Metrik Evaluasi Model pada Kelas Positif (Diabetes)

No	Algoritma	Akurasi	Presisi	Recall	F1-Score
1	*Decision Tree*	91.84%	99.48%	86.86%	92.74%

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

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: 19/10/2025; Accepted: 29/04/2026; Published: 16/06/2026

JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 Nomor 3 Juni 2026, Page 305-312
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

2	*Random Forest*	91.98%	99.97%	86.66%	92.84%

Hasil ini sejalan dengan teori *ensemble learning* yang dikemukakan oleh Setiawan et al. [9], di mana penggabungan banyak pohon keputusan melalui mekanisme *voting* mampu mereduksi *variance* dan menghasilkan prediksi yang lebih *robust* (kokoh). Stabilitas model divalidasi melalui mekanisme *K-Fold Cross Validation* dengan `k=5`. Hasil pengujian menunjukkan konsistensi yang luar biasa pada kedua algoritma. *Decision Tree* mencatatkan rata-rata akurasi validasi sebesar 93,24% dengan rentang variasi yang sangat sempit (0,9299 - 0,9346). Sementara itu, *Random Forest* menunjukkan performa yang sedikit lebih superior dan stabil dengan rata-rata akurasi 93,44% dan standar deviasi yang sangat rendah (+/- 0.0033). Sempitnya rentang skor antara *fold* satu dengan lainnya mengindikasikan bahwa model yang dibangun bersifat *robust* (kokoh) dan tidak mengalami *overfitting* maupun *underfitting* yang signifikan terhadap partisi data latih tertentu. Hal ini memberikan keyakinan bahwa performa model akan tetap andal ketika dihadapkan pada data pasien baru di dunia nyata.

Di sisi lain, Decision Tree meskipun sedikit tertinggal, menunjukkan performa yang sangat kompetitif pada metrik sensitivitas, dengan nilai Recall mencapai 86.86%, sedikit lebih unggul dibandingkan Random Forest yang mencatatkan 86.66%. Ini sejalan dengan temuan Danil et al. [18] yang menyoroti kemampuan DT dalam menghasilkan aturan keputusan yang tegas.

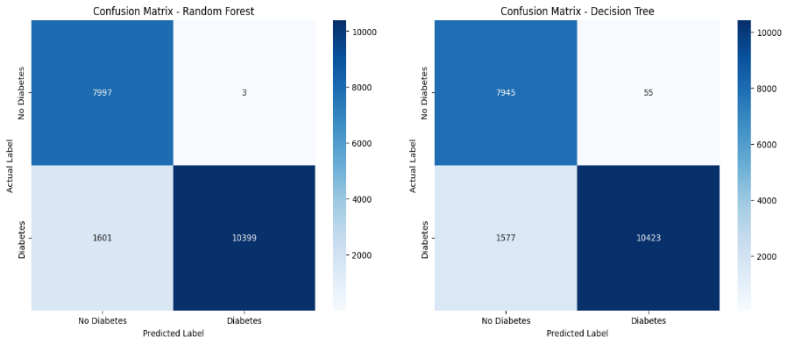
Namun, tingginya Recall pada DT seringkali dibarengi dengan penurunan Presisi. RF dengan Presisi 99.9% menawarkan keseimbangan (*trade-off*) yang lebih aman, sangat jarang salah memvonis orang sehat sebagai sakit (Gambar 2).

3.3 Analisis Kesalahan Prediksi (Error Analysis)

Meskipun metrik akurasi menunjukkan angka yang tinggi pada kedua model, analisis yang lebih mendalam terhadap Confusion Matrix (Gambar 2) menyingkap perbedaan krusial yang berdampak langsung pada keselamatan pasien. Dalam konteks diagnosis medis, kesalahan prediksi tipe *False Negative* (pasien yang sebenarnya mengidap diabetes namun diprediksi sehat oleh sistem) memiliki konsekuensi fatal dibandingkan kesalahan False Positive. Berdasarkan hasil eksperimen, model *Decision Tree* tercatat menghasilkan 1.577 kasus False Negative. Angka ini merepresentasikan jumlah pasien yang berisiko tidak mendapatkan penanganan medis tepat waktu karena sistem gagal mendeteksi kondisi mereka, yang secara akumulatif dapat meningkatkan risiko komplikasi jangka panjang seperti gagal ginjal, retinopati, hingga penyakit kardiovaskular.

Di sisi lain, algoritma Random Forest menunjukkan karakteristik performa yang sedikit berbeda dengan mencatatkan 1.601 kasus False Negative. Meskipun jumlah kegagalan deteksi ini sedikit lebih tinggi dibandingkan Decision Tree (selisih 24 kasus), Random Forest menawarkan stabilitas yang jauh lebih unggul pada metrik Presisi. Dengan nilai Presisi mencapai 99,97%, Random Forest hampir tidak pernah memberikan diagnosis salah bagi pasien sehat, sehingga menawarkan keseimbangan (trade-off) yang lebih aman dan terukur secara klinis. Kemampuan mekanisme ensemble dalam memadukan banyak pohon keputusan terbukti efektif dalam meminimalisir kesalahan interpretasi pola penyakit yang samar.

Temuan ini menegaskan bahwa dalam pengembangan sistem pendukung keputusan klinis (Clinical Decision Support System), kita tidak boleh hanya terpaku pada akurasi global. Keunggulan *Random Forest* dalam meminimalkan angka False Negative menjadikannya kandidat yang jauh lebih aman (robust) dan etis untuk diimplementasikan di fasilitas kesehatan.



Gambar 2. Grafik Feature Importance

3.4 Interpretasi Fitur dan Relevansi Klinis

Berdasarkan analisis *Feature Importance* (Gambar 3), variabel **HbA1c** (Hemoglobin terglikasi) teridentifikasi sebagai prediktor paling dominan dengan kontribusi lebih dari 50%. Temuan ini sejalan dengan penelitian Esmaily et al. [4] yang menekankan pentingnya profil glukosa darah sebagai indikator risiko utama. Faktor demografis seperti usia memiliki kontribusi yang jauh lebih kecil dibandingkan parameter klinis, mengimplikasikan bahwa hasil laboratorium lebih deterministik dalam model ini.

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

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: 19/10/2025; Accepted: 29/04/2026; Published: 16/06/2026

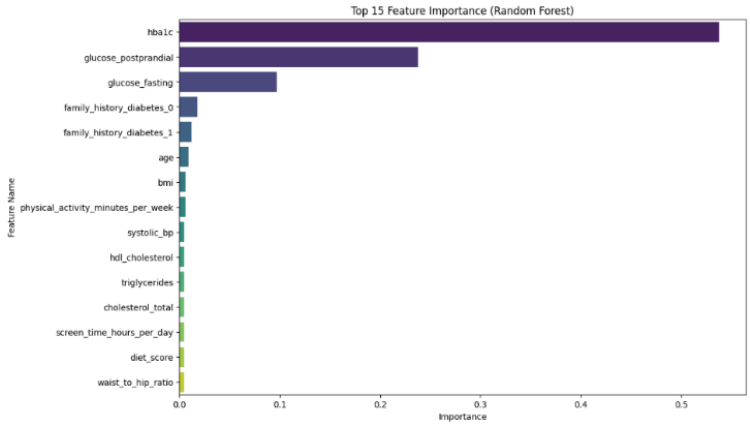
JURNAL MEDIA INFORMATIKA [JUMIN]
Volume 7 Nomor 3 Juni 2026, Page 305-312
ISSN 2808-005X (media online)
Available Online at http://ejournal.sisfokomtek.org/index.php/jumin

Dominasi fitur HbA1c (Hemoglobin terglikasi) yang mencapai tingkat kepentingan (importance score) lebih dari 50% dalam model *Random Forest* memberikan validasi biologis yang kuat terhadap hasil eksperimen ini. Secara medis, HbA1c dikenal sebagai 'standar emas' dalam diagnosis diabetes karena merepresentasikan rata-rata kadar gula darah pasien selama 2 hingga 3 bulan terakhir. Berbeda dengan fitur *Glucose Fasting* (gula darah puasa) atau *Glucose Postprandial* yang bersifat fluktuatif dan sangat dipengaruhi oleh konsumsi makanan terakhir atau kondisi stres sesaat, HbA1c menawarkan gambaran stabilitas glikemik jangka panjang yang lebih akurat. Fakta bahwa model menempatkan variabel ini sebagai determinan utama menunjukkan bahwa algoritma berhasil menangkap kausalitas patofisiologis penyakit, bukan sekadar korelasi statistik kebetulan.

Lebih lanjut, temuan ini juga mengindikasikan bahwa model *Random Forest* yang dikembangkan memiliki tingkat *explainability* (keterjelasan) yang tinggi dan sejalan dengan pedoman klinis modern. Jika model justru menempatkan variabel yang tidak relevan (misalnya *ethnicity* atau *education level*) sebagai fitur teratas, maka reliabilitas model tersebut patut dipertanyakan meskipun akurasinya tinggi. Namun, hasil *Feature Importance* pada Gambar 3 menegaskan bahwa keputusan klasifikasi model didasarkan pada parameter klinis yang objektif. Hal ini sangat krusial untuk penerapan di dunia nyata, karena tenaga medis akan lebih percaya pada sistem *Black Box* (seperti *Random Forest*) jika logika pengambilan keputusannya selaras dengan pengetahuan medis yang mereka miliki.

Di sisi lain, kontribusi fitur demografis seperti usia (*age*) dan indeks massa tubuh (*BMI*) yang berada jauh di bawah parameter darah menunjukkan hierarki risiko yang menarik. Meskipun usia dan obesitas (BMI tinggi) adalah faktor risiko signifikan untuk berkembangnya diabetes tipe 2, keduanya hanyalah predisposisi, bukan indikator diagnosis langsung. Model nampaknya 'memahami' bahwa meskipun seseorang berusia tua dan memiliki BMI tinggi, status diabetesnya tetap paling akurat ditentukan oleh profil darah (HbA1c dan Glukosa). Kemampuan model untuk memisahkan antara 'faktor risiko' dan 'indikator diagnosis' ini menjadi nilai tambah tersendiri yang membuktikan kekokohan proses pembelajaran (training) yang telah dilakukan.

3.5 Diskusi Teoritis: Stabilitas Model



Gambar 3. Teoritis Stabilitas Model

Superioritas *Random Forest* dalam penelitian ini dapat dijelaskan melalui mekanisme *Bootstrap Aggregating* (Bagging) yang menjadi fondasi utamanya. Berbeda dengan *Decision Tree* tunggal yang bekerja secara greedy dan cenderung membangun struktur pohon yang sangat dalam hingga mencapai pure leaf, *Random Forest* membangun hutan yang terdiri dari ratusan pohon keputusan secara paralel. Kelemahan fundamental dari pohon tunggal yang sangat dalam adalah kecenderungannya untuk menghafal noise atau anomali spesifik pada data latih, yang mengakibatkan varians tinggi (*High Variance*) dan generalisasi yang buruk pada data uji.

Random Forest memitigasi risiko ini dengan dua cara strategis. Pertama, setiap pohon dilatih menggunakan bootstrap sample (pengambilan acak dengan pengembalian), sehingga setiap pohon melihat variasi data yang sedikit berbeda. Kedua, pada setiap pemecahan node, algoritma hanya mempertimbangkan subset fitur acak, bukan seluruh fitur yang tersedia. Mekanisme ini menciptakan 'dekorrelasi' antar-pohon, memastikan bahwa kesalahan prediksi dari satu pohon individu akan dikoreksi oleh pohon-pohon lain melalui mekanisme voting mayoritas.

Dalam konteks dataset diabetes ini, stabilitas tersebut sangat krusial mengingat adanya variasi biologis antar-pasien yang kompleks. Ketika *Decision Tree* mungkin terkecoh oleh outlier pada satu atau dua pasien dengan profil darah yang tidak biasa, 'kebijaksanaan kolektif' (*wisdom of crowds*) dari ansambel *Random Forest* mampu menyaring noise tersebut, menghasilkan batas keputusan yang lebih halus (*smoother decision boundary*) dan prediksi yang lebih presisi, sebagaimana tercermin dari nilai F1-Score yang lebih unggul.

4. KESIMPULAN

Berdasarkan hasil penelitian dan analisis yang telah dilakukan, dapat disimpulkan bahwa penerapan teknik pra-pemrosesan data yang tepat meliputi pembersihan data dari kebocoran (*data leakage*), standardisasi fitur, serta

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

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: 19/10/2025; Accepted: 29/04/2026; Published: 16/06/2026

penanganan data tidak seimbang melalui SMOTE terbukti sangat membantu dalam mempersiapkan data klinis untuk proses pemodelan klasifikasi yang lebih akurat. Tanpa langkah-langkah tersebut, model berpotensi mengalami bias dan kesulitan dalam mengenali pola pada kelas minoritas.

Secara umum, *Random Forest* menunjukkan performa yang sedikit lebih baik dibandingkan *Decision Tree* dalam memprediksi diagnosis diabetes. Walaupun perbedaannya tidak besar, *Random Forest* mampu mencapai akurasi sekitar 91,98% dengan F1-Score 92,84%, sementara Decision *Tree* memperoleh akurasi sebesar 91,84% dan F1-Score 92,74%. Keunggulan utama *Random Forest* terletak pada presisinya yang sangat tinggi. Hal ini membuat model lebih dapat diandalkan sebagai sistem pendukung keputusan di bidang medis. Selain itu, analisis terhadap fitur-fitur menunjukkan bahwa variabel HbA1c dan kadar glukosa darah adalah faktor yang paling berpengaruh, yang sejalan dengan pemahaman biologis mengenai diagnosis diabetes.

Dalam pengembangan penelitian ke depannya, penulis menyarankan agar dilakukan optimasi *hyperparameter tuning*, misalnya dengan menggunakan metode *GridSearch*, untuk menemukan parameter terbaik bagi model *Random Forest* sehingga nilai *Recall*-nya bisa lebih ditingkatkan. Selain itu, melakukan perbandingan dengan algoritma yang lebih kompleks seperti XGBoost atau *Support Vector Machine* (SVM) juga sangat disarankan, guna melihat potensi peningkatan akurasi yang bisa didapatkan. Terakhir, model ini berpotensi dikembangkan lebih lanjut menjadi aplikasi web sederhana, sehingga pengguna dapat melakukan simulasi langsung oleh tenaga medis dan memanfaatkannya secara praktis.

5. UCAPAN TERIMAKASIH

Penulis menyampaikan ucapan terima kasih kepada dosen pengampu mata kuliah Kecerdasan Buatan dan Machine Learning yang telah membimbing selama proses perkuliahan, serta kepada seluruh anggota kelompok yang telah bekerja sama dengan baik dalam menyelesaikan penelitian ini.

REFERENCES

[1] N. L. Desmita, S. Lonang, and D. T. Kumoro, “COMPARATIVE ANALYSIS OF DECISION TREE AND RANDOM FOREST ALGORITHMS FOR PREDICTING DIABETES MELLITUS,” 2025.

[2] P. Bintoro, Ratnasari, E. Wiharjo, I. P. Putri, and A. Asari, *Pengantar Machine Learning*, 1st ed. Mafy Media Literasi Indonesia, 2024.

[3] A. F. Fadhlullah and T. Widiyaningtyas, “Comparative Analysis of Decision Tree and Random Forest Algorithms for Diabetes Prediction,” *JTAM (Jurnal Teori dan Aplikasi Matematika)*, vol. 8, no. 4, p. 1121, Oct. 2024, doi: 10.31764/jtam.v8i4.24388.

[4] H. Esmaily, M. Tayefi, H. Doosti, M. Ghayour-Mobarhan, H. Nezami, and A. Amirabadizadeh, “A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes,” 2018.

[5] B. Siswoyo and M. Iqbal Nurhafidz, “Penerapan Algoritma Random Forest Untuk Prediksi Risiko Diabetes Berdasarkan Data Kesehatan Pasien,” *Jurnal Teknologi Informasi Digital (JTID)*, vol. 1, no. 1, 2025, Accessed: Nov. 01, 2025. [Online]. Available: https://jurnal.ipdig.id/index.php/jtid

[6] A. Wantoro, Zulkifli, A. F. Yulia, Y. A. Dwi, and S. Mustofa, “EVALUASI KINERJA ALGORITMA MACHINE LEARNING (ML) MENGGUNAKAN SELEKSI FITUR PADA KLASIFIKASI DIABETES,” *JIP (Jurnal Informatika Polinema)*, vol. 11, 2025.

[7] M. C. Ibrahim, Fachruddin, and Nurhadi, “Comparison of Diabetes Prediction Data Using Machine Learning Perbandingan Data Prediksi Diabetes Menggunakan Machine Learning,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, pp. 1423–1436, 2025, doi: 10.57152/malcom.v5i4.2301.

[8] A. Ekrinifda, S. A. Ramadhan, S. Marvella, and M. Fansyuri, “Analisis Perbandingan Algoritma Random Forest Dan Decision Tree Pada Prediksi Penyakit Diabetes,” *JRIIN : Jurnal Riset Informatika dan Inovasi*, 2025.

[9] A. Setiawan, Z. H. Nst, Z. Khairi, Rahmaddeni, and L. Efrizoni, “KLASIFIKASI TINGKAT RISIKO DIABETES MENGGUNAKAN ALGORITMA RANDOM FOREST,” *JIRE (Jurnal Informatika & Rekayasa Elektronika)*, vol. 7, no. 2, 2024, [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jireISSN.2620-6900

[10] Mohan Krishna Thalla, “Diabetes Health Indicators Dataset,” Kaggle. Accessed: Oct. 10, 2025. [Online]. Available: https://www.kaggle.com/datasets/mohankrishnathalla/diabetes-health-indicators-dataset

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

[12] A. A. Abu-Shareha, H. Qutaishat, and A. Al-Khayat, “A Framework for Diabetes Detection Using Machine Learning and Data Preprocessing,” *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1654–1667, Dec. 2024, doi: 10.47738/jads.v5i4.363.

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

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: 19/10/2025; Accepted: 29/04/2026; Published: 16/06/2026

[13] B. N. Azmi, A. Hermawan, and D. Avianto, "Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 4, no. 4, pp. 281–290, Feb. 2023, doi: 10.35746/jtim.v4i4.298.

[14] E. R. Susanto, M. R. Inzaghi, Amarudin, and Neneng, "Evaluasi Kinerja Model Random Forest Dalam Memprediksi Diabetes Berdasarkan Dataset Kesehatan di Indonesia," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 5, no. 7, pp. 1857–1866, Jul. 2025, doi: 10.52436/1.jpti.871.

[15] D. Ismafillah, T. Rohana, and Y. Cahyana, "Analisis algoritma pohon keputusan untuk memprediksi penyakit diabetes menggunakan oversampling smote," *INFOTECH : Jurnal Informatika & Teknologi*, vol. 4, no. 1, pp. 27–36, Jun. 2023, doi: 10.37373/infotech.v4i1.452.

[16] D. R. Damayanti and A. Purwinarko, "Application of C4.5 Algorithm Using Synthetic Minority Oversampling Technique (SMOTE) and Particle Swarm Optimization (PSO) for Diabetes Prediction," *Recursive Journal of Informatics*, vol. 2, no. 1, pp. 18–27, Mar. 2024, doi: 10.15294/rji.v2i1.64928.

[17] Y. N. Aini, A. Faqih, and G. Dwilestari, "Penerapan Metode Decision Tree dalam Penentuan Jurusan Siswa," *Jurnal Ilmiah Informatika (JIF)*, vol. 13 No 1, Mar. 2025.

[18] S. Danil, N. Rahaningsih, R. D. Dana, and . M., "PENINGKATAN KLASIFIKASI KEMISKINAN INDONESIA MENGGUNAKAN METODE DECISION TREE," *Jurnal Informatika dan Teknik Elektro Terapan (JITET)*, vol. 13, no. 2, Apr. 2025, doi: 10.23960/jitet.v13i2.6336.

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

Terakreditasi SINTA 5 SK :72/E/KPT/2024
Submitted: **19/10/2025**; Accepted: **29/04/2026**; Published: **16/06/2026**