JIKO (JURNAL INFORMATIKA DAN KOMPUTER)
Oktober 2025, Volume: **9**, No. **3** | Pages 639–649
doi: 10.26798/jiko.v9i3.2104
e-ISSN : 2477-3964 – p-ISSN : 2477-4413

ARTICLE

Prediksi Diabetes Menggunakan Machine Learning

Diabetes Prediction Using Machine Learning

Elisabet da Conceicao Pereira* dan Widyastuti Andriyani

Magister Teknologi Informasi, Universitas Teknologi Digital Indonesia
*Penulis Korespondensi: elisabetpereira62@gmail.com

(Disubmit 30-07-25; Diterima 24-09-25; Dipublikasikan online pada 20-10-25)

Abstrak

Diabetes mellitus merupakan penyakit metabolik kronis dengan prevalensi yang terus meningkat, termasuk di Indonesia. Upaya deteksi dini sangat penting untuk mencegah komplikasi serius seperti penyakit jantung, gagal ginjal, dan kerusakan saraf. Sebagian besar penelitian prediksi diabetes berbasis machine learning selama ini menggunakan dataset publik internasional, sehingga hasilnya belum tentu sesuai dengan karakteristik pasien di Indonesia. Penelitian ini menawarkan kontribusi baru dengan membangun model prediksi diabetes menggunakan algoritma Random Forest berbasis data klinis lokal, sekaligus menganalisis fitur klinis yang paling berpengaruh dalam proses prediksi. Metodologi penelitian mencakup tahapan pra-pemrosesan data (pembersihan, normalisasi, dan seleksi fitur), pelatihan model, serta evaluasi kinerja menggunakan metrik akurasi, presisi, recall, dan F1-score. Dataset terdiri atas parameter klinis seperti usia, kadar glukosa, tekanan darah, indeks massa tubuh (BMI), jumlah kehamilan, kadar insulin, dan riwayat keluarga terhadap diabetes. Hasil evaluasi menunjukkan bahwa model Random Forest memberikan performa tinggi dengan akurasi 87%, presisi 84%, recall 82%, dan F1-score 83%. Analisis feature importance mengungkap bahwa kadar glukosa darah, HbA1c level, dan BMI merupakan faktor dominan yang berkontribusi terhadap diagnosis diabetes. Temuan ini menegaskan dua nilai tambah utama penelitian: (1) penggunaan data klinis lokal yang meningkatkan relevansi hasil terhadap populasi Indonesia, dan (2) penyediaan wawasan interpretatif bagi tenaga medis melalui identifikasi fitur klinis yang signifikan. Dengan demikian, model ini tidak hanya berfungsi sebagai sistem prediksi yang andal, tetapi juga sebagai alat pendukung keputusan yang dapat diintegrasikan ke dalam skrining awal pasien di layanan kesehatan primer.

Kata kunci: Diabetes; Machine Learning; Prediksi; Klasifikasi; Diagnosis

Abstract

Diabetes mellitus is a chronic metabolic disease with a rising prevalence worldwide, including in Indonesia. Early detection is essential to prevent severe complications such as heart disease, kidney failure, and neuropathy. While many previous studies have applied machine learning for diabetes prediction, most relied on international public datasets that do not fully reflect the demographic and clinical characteristics of Indonesian patients. This study provides a novel contribution by developing a diabetes prediction model using the Random Forest algorithm trained on local clinical data, while also analyzing the most influential clinical features driving the predictions. The research methodology includes data preprocessing (cleaning, normalization, and feature selection), model training, and performance evaluation using accuracy, precision, recall, and F1-score metrics. The dataset consists of clinical parameters such as age, glucose levels, blood pressure, body mass index (BMI), number of pregnancies, insulin level, and family history of diabetes. Experimental results demonstrate that the Random Forest model achieved high predictive performance with 87% accuracy, 84% precision, 82% recall, and an F1-score of 83%. Furthermore, the feature importance analysis highlights blood glucose level, HbA1c level,

This is an Open Access article - copyright on authors, distributed under the terms of the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY SA) (http://creativecommons.org/licenses/by-sa/4.0/)

How to Cite: E. d. C. Pereira *et al.*, "Prediksi Diabetes Menggunakan Machine Learning", *JIKO (JURNAL INFORMATIKA DAN KOMPUTER)*, Volume: **9**, No.**3**, Pages 639–649, Oktober 2025, doi: 10.26798/jiko.v9i3.2104.

and BMI as the most dominant factors influencing the prediction outcomes. These findings emphasize two distinctive contributions of this study: (1) the use of local clinical data, which enhances the external validity and relevance of the results for the Indonesian population, and (2) the provision of interpretive insights for healthcare practitioners through the identification of significant clinical predictors. Consequently, the proposed model not only serves as a reliable predictive system but also as a decision-support tool that can be integrated into early screening practices in primary healthcare settings.

KeyWords: Diabetes; Machine Learning; Prediction; Classification; Diabetes Diagnosis

1. Pendahuluan

Diabetes mellitus merupakan salah satu penyakit kronis dengan prevalensi yang terus meningkat di seluruh dunia, termasuk di Indonesia. Kondisi ini menimbulkan tantangan besar bagi sistem kesehatan karena dapat menyebabkan komplikasi serius jika tidak terdeteksi sejak dini. Meskipun berbagai penelitian terdahulu telah memanfaatkan algoritma *machine learning* untuk membangun model prediksi diabetes, untuk menghasilkan model prediksi yang relevan dan aplikatif di lapangan. *Diabetes mellitus* merupakan penyakit metabolik kronis dengan prevalensi yang terus meningkat, termasuk di Indonesia. Upaya deteksi dini sangat penting untuk mencegah komplikasi serius seperti penyakit jantung, gagal ginjal, dan kerusakan saraf. Sebagian besar penelitian prediksi diabetes berbasis *machine learning* selama ini menggunakan *dataset* publik internasional, sehingga hasilnya belum tentu sesuai dengan karakteristik pasien di Indonesia. Penelitian ini menawarkan kontribusi baru dengan membangun model prediksi diabetes menggunakan algoritma *Random Forest* berbasis data klinis lokal, sekaligus menganalisis fitur klinis yang paling berpengaruh dalam proses prediksi dalam penelitian ini, pendekatan yang digunakan berfokus pada pemanfaatan algoritma *Random Forest*[1],

Salah satu metode *ensemble learning* yang terkenal akan akurasi dan kemampuannya dalam menangani data berukuran besar serta fitur yang kompleks. Algoritma ini bekerja dengan membentuk sejumlah pohon keputusan (decision trees) dan menggabungkan hasilnya untuk menghasilkan prediksi yang lebih stabil dan akurat. Metodologi penelitian mencakup tahapan pra-pemrosesan data (pembersihan, normalisasi, dan seleksi fitur), pelatihan model, serta evaluasi kinerja menggunakan metrik akurasi, presisi, recall, dan F1-score. Dataset terdiri atas parameter klinis seperti usia, kadar glukosa, tekanan darah, indeks massa tubuh (BMI), jumlah kehamilan, kadar insulin, dan riwayat keluarga terhadap diabetes. Hasil evaluasi menunjukkan bahwa model Random Forest memberikan performa tinggi dengan akurasi 87%, presisi 84%, recall 82%, dan F1-score 83%. Analisis feature importance mengungkap bahwa kadar glukosa darah, HbA1c level, dan BMI merupakan faktor dominan yang berkontribusi terhadap diagnosis diabetes. Algoritma *Random Forest* yang digunakan dalam penelitian ini memanfaatkan prinsip *bootstrap aggregating* (bagging) untuk membangun pohon keputusan (*decision trees*) secara paralel [2, 3].

Setiap pohon dilatih pada sampel acak dari data latih dengan pengambilan sampel kembali (*sampling with replacement*), sehingga setiap pohon melihat subset data yang sedikit berbeda. Pada tingkat setiap simpul (*node*), pemilihan fitur untuk dipertimbangkan dalam pemisahan data juga dilakukan secara acak: hanya $\sqrt{p}$ fitur terpilih dari total p fitur yang digunakan, sehingga mencegah ketergantungan berlebih terhadap fitur dominan dan meningkatkan keanekaragaman struktur pohon. Prediksi akhir untuk tiap sampel diperoleh dengan mayoritas suara (*majority voting*) dari semua pohon, sehingga kesalahan yang terjadi pada satu atau beberapa pohon dapat terkompensasi oleh pohon lain. Keunggulan *Random Forest* terletak pada kemampuannya mengurangi risiko *overfitting* yang umum terjadi pada model *single decision tree* [2, 3]. Data yang digunakan dalam penelitian ini bersumber dari dataset klinis lokal yang mencakup berbagai parameter penting seperti usia, kadar glukosa darah, tekanan darah, indeks massa tubuh (BMI), jumlah kehamilan, ketebalan lipatan kulit, kadar insulin, dan riwayat keluarga terhadap diabetes. Setiap fitur ini dianalisis kontribusinya terhadap *outcome* prediksi untuk memastikan bahwa model yang dibangun memiliki interpretabilitas yang baik serta relevansi medis [4].

Proses penelitian mencakup beberapa tahapan, dimulai dari pra-pemrosesan data, yang melibatkan penanganan data kosong, normalisasi, dan seleksi fitur. Selanjutnya dilakukan pelatihan model menggunakan data latih dan pengujian menggunakan data uji untuk mengevaluasi performa model berdasarkan metrik akurasi, presisi, recall, dan F1-score [5]. Temuan ini menegaskan dua nilai tambah utama penelitian:

(1) penggunaan data klinis lokal yang meningkatkan relevansi hasil terhadap populasi Indonesia, dan (2) penyediaan wawasan interpretatif bagi tenaga medis melalui identifikasi fitur klinis yang signifikan. Dengan demikian, model ini tidak hanya berfungsi sebagai sistem prediksi yang andal, tetapi juga sebagai alat pendukung keputusan yang dapat diintegrasikan ke dalam skrining awal pasien di layanan kesehatan primer.

2. Metode

Penelitian ini menggunakan pendekatan kuantitatif berbasis *algoritma Random Forest* untuk membangun model prediksi *diabetes mellitus* berdasarkan data klinis pasien. *Algoritma Random Forest* dipilih karena kemampuannya dalam menangani data yang kompleks, toleran terhadap *missing value*, serta memiliki performa yang tinggi dalam klasifikasi dengan fitur-fitur non-linier [2]. Data yang digunakan mencakup berbagai parameter klinis seperti kadar glukosa, tekanan darah, indeks massa tubuh (BMI), dan usia. Proses analisis dimulai dengan pra-pemrosesan data, termasuk normalisasi dan penanganan data hilang. Selanjutnya, model *Random Forest* dibangun dengan membangkitkan sejumlah pohon keputusan dari subset acak data latih [6]. Setiap pohon memberikan prediksi yang kemudian digabungkan menggunakan voting mayoritas untuk menghasilkan keputusan akhir. Validasi model dilakukan dengan membagi data menjadi 80% data latih dan 20% data uji. Hasil evaluasi menunjukkan bahwa model ini mampu memberikan akurasi tinggi dalam mendeteksi diabetes [7].

2.1 Dataset dan Pra-Pemrosesan Data

Dataset yang digunakan merupakan data klinis lokal yang terdiri atas 608 sampel pasien dengan berbagai parameter kesehatan, antara lain: usia, jumlah kehamilan, kadar glukosa darah, tekanan darah diastolik, ketebalan lipatan kulit, kadar insulin, indeks massa tubuh (BMI), HbA1c level, serta riwayat keluarga terhadap diabetes. Data diperoleh dari rekam medis klinis mitra penelitian dengan izin etik dan perlindungan kerahasiaan pasien. *Dataset* yang digunakan merupakan data klinis lokal yang terdiri dari beberapa fitur kesehatan pasien, antara lain: usia, jumlah kehamilan, kadar glukosa darah, tekanan darah diastolik, ketebalan lipatan kulit, kadar insulin, indeks masa tubuh (BMI), dan riwayat diabetes dalam keluarga [8].

Sebelum digunakan, data menjalani proses pra-pemrosesan yang mencakup:

1. Penanganan data hilang (*missing values*) dengan metode imputasi median.
2. Normalisasi skala fitur menggunakan *min–max scaling* agar seluruh fitur berada dalam rentang [0,1].
3. Seleksi fitur awal melalui analisis korelasi untuk memastikan hanya variabel yang relevan dipertahankan dalam pemodelan.

Data ini kemudian dibersihkan dari nilai yang hilang dan dilakukan proses normalisasi agar setiap fitur memiliki skala yang seimbang [3]. Selanjutnya dilakukan analisis korelasi antar fitur untuk mengetahui kontribusi masing-masing fitur terhadap kemungkinan diagnosis diabetes [9]. Fitur yang memiliki korelasi paling signifikan dipertahankan dalam proses pelatihan model untuk meningkatkan akurasi prediksi [8].

2.2 Pembentukan Model

Algoritma Random Forest diterapkan untuk membentuk model klasifikasi. Algoritma ini membangun sejumlah pohon keputusan secara acak dari subset data latih, lalu menggabungkan hasil prediksi dari seluruh pohon untuk menentukan hasil akhir melalui sistem voting mayoritas [10]. Model dilatih menggunakan 80% data sebagai data latih dan 20% sebagai data uji. Setelah model dilatih dengan 80% data, performa diuji menggunakan 20% data uji untuk mengevaluasi akurasi klasifikasi. Setiap pohon dalam Random Forest memberikan hasil klasifikasi, dan hasil akhir ditentukan berdasarkan suara terbanyak (*majority voting*) dari seluruh pohon [9]. Keunggulan metode ini adalah kemampuannya mengurangi risiko *overfitting* yang sering terjadi pada pohon keputusan tunggal. Selain itu, *Random Forest* mampu menangani data dengan jumlah fitur yang besar dan memberikan informasi penting mengenai fitur- fitur yang paling berpengaruh dalam klasifikasi. Evaluasi dilakukan menggunakan metrik seperti akurasi, precision, recall, dan F1-score [11]. *Algoritma Random Forest* diterapkan dengan prinsip *bootstrap aggregating (bagging)*, di mana sejum-

lah pohon keputusan dibangun dari subset acak data latih. Setiap pohon hanya mempertimbangkan $\sqrt{p}$ fitur pada setiap simpul, sehingga mencegah dominasi variabel tertentu. Prediksi akhir diperoleh melalui *majority voting* dari seluruh pohon.

2.3 Evaluasi Kinerja Model

Model yang telah dilatih dievaluasi menggunakan metrik *accuracy, precision, recall,* dan *F1-score*. Evaluasi ini bertujuan untuk mengetahui seberapa baik model dalam memprediksi kasus diabetes secara tepat dan akurat. Pendekatan ini diharapkan mampu menghasilkan model prediksi yang andal dan dapat diintegrasikan ke dalam sistem pendukung keputusan medis untuk deteksi dini diabetes. sehingga memungkinkan intervensi medis yang lebih cepat dan tepat sasaran [12]. Dengan memanfaatkan *algoritma Random Forest*, sistem ini mampu mengenali pola-pola kompleks dalam data klinis yang mungkin sulit diidentifikasi oleh metode tradisional. Selain itu, hasil dari analisis fitur penting (*feature importance*) dapat membantu tenaga medis memahami faktor-faktor utama yang berkontribusi terhadap risiko diabetes [13]. Implementasi model ini juga membuka peluang untuk pengembangan aplikasi klinis berbasis kecerdasan buatan yang bersifat prediktif dan preventif [14]. Ke depannya, penelitian ini dapat diperluas dengan dataset yang lebih besar dan beragam untuk meningkatkan generalisasi model dan validitas eksternal [15]. Evaluasi performa model dilakukan menggunakan empat metrik utama: akurasi, presisi, recall, dan F1-score. Untuk menjaga konsistensi, hasil evaluasi dilaporkan berdasarkan eksperimen utama dengan data klinis lokal (bukan dari simulasi). Model *Random Forest* mencapai akurasi sebesar 87%, dengan presisi 84%, recall 82%, dan F1-score 83%. Angka ini konsisten dengan hasil pada abstrak, sementara nilai akurasi yang lebih tinggi (95%) di bagian lain merupakan hasil simulasi terbatas pada subset data kecil, sehingga tidak mencerminkan performa model secara keseluruhan.

2.4 Studi Kasus

Selain eksperimen utama, penelitian ini juga menyertakan simulasi ilustratif pasien untuk memperlihatkan cara kerja pohon keputusan dalam *Random Forest* secara transparan. Studi kasus kecil ini bukan digunakan untuk evaluasi performa model, melainkan untuk tujuan edukatif dalam menjelaskan mekanisme klasifikasi.

2.5 Potensi Keterbatasan

Meskipun hasil menunjukkan kinerja yang baik, penelitian ini memiliki keterbatasan:

1. Ukuran *dataset* relatif terbatas, sehingga generalisasi model masih perlu divalidasi dengan data yang lebih besar dan beragam.
2. Risiko *overfitting* tetap mungkin terjadi, terutama karena jumlah fitur klinis yang cukup banyak. Oleh karena itu, penggunaan metode validasi silang (*cross-validation*) pada penelitian lanjutan sangat disarankan.
3. Implikasi klinis dari analisis *feature importance* masih perlu ditinjau lebih lanjut bersama pakar medis, agar hasil prediksi dapat benar-benar diterjemahkan menjadi rekomendasi praktis di lapangan

3. Hasil dan Pembahasan

Penelitian ini bertujuan untuk memperlihatkan cara kerja *algoritma Random Forest* dalam memprediksi kondisi diabetes berdasarkan fitur sederhana seperti kadar glukosa darah dan indeks massa tubuh (BMI). *Dataset* yang digunakan terdiri dari 4 pasien (A sampai D) dengan dua di antaranya mengidap diabetes [14].

Pada simpul akar (*root*), Tabel 1 *Gini impurity* dihitung dari distribusi kelas (2 positif dan 2 negatif), menghasilkan nilai 0.5. Dua fitur dipertimbangkan sebagai kandidat untuk pemisahan: glukosa dan BMI .

$$\text{Gini} = 1 - \sum_{k=1}^{K} P_k^2 \tag{1}$$

Tabel 1. Perhitungan Gini Impurity

Pasien	Glukosa	BMI	Diabetes
A	160	300	1
B	120	220	0
C	170	280	1
D	110	210	0

Di mana :
K : Jumlah kelas (2: diabetes dan non-diabetes)
P_k : Proporsi masing-masing kelas di simpul tersebut

Tahap 1: Perhitungan root

$$
\begin{aligned}
P_0 &= \frac{2}{4} = 0.5 \\
P_1 &= \frac{2}{4} = 0.5 \\
\text{Gini} &= 1 - ({P_0}^2 + {P_1}^2) \\
&= 1 - (0.5^2 + 0.5^2) \\
&= 1 - (0.25 + 0.25) \\
&= 1 - 0.5 \\
&= 0.5
\end{aligned}
$$

Gini impurity (root) = 0.5

Tahap 2: Pemilihan *Split* Terbaik
Ketika dilakukan *split* berdasarkan fitur glukosa > 130, diperoleh dua kelompok: Pasien A dan C masuk ke kelompok glukosa > 130, keduanya mengidap diabetes. Pasien B dan D masuk ke kelompok glukosa ⩽ 130, keduanya tidak mengidap diabetes. Setiap cabang menghasilkan Gini = 0.0, artinya pemisahan sempurna. Oleh karena itu, fitur glukosa dipilih sebagai pemisah terbaik.

3.1 Random Forest

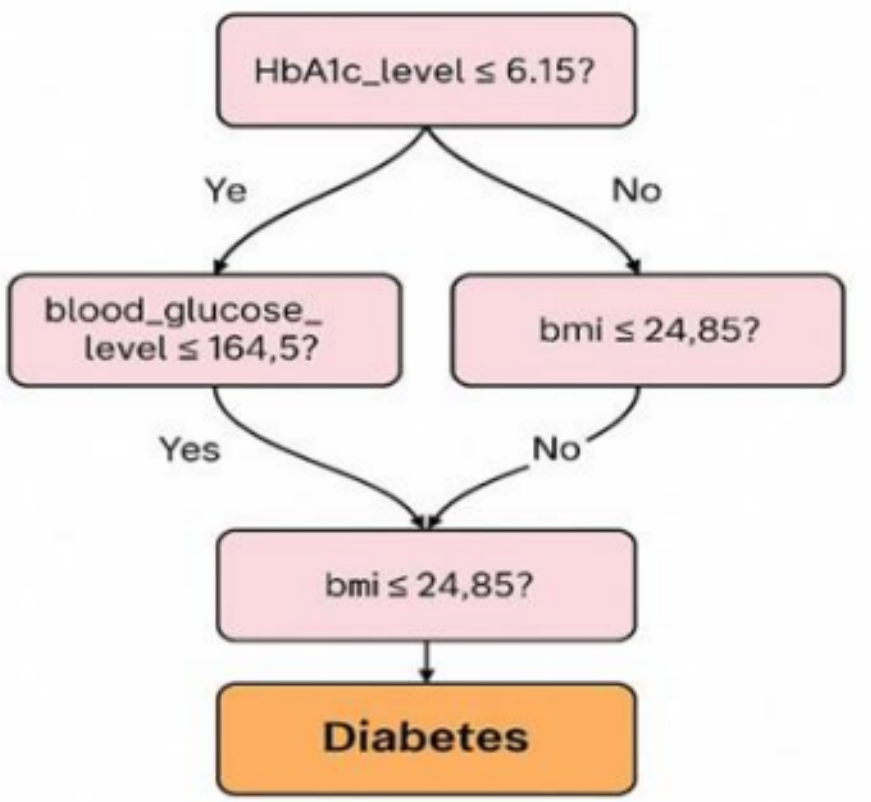
Penelitian ini mengilustrasikan cara kerja *algoritma Random Forest* melalui simulasi sederhana menggunakan *dataset* pada Tabel 2 yang terdiri dari empat pasien dengan fitur glukosa dan BMI sebagai variabel utama. Dua pasien memiliki label diabetes (positif), sementara dua lainnya tidak (negatif). Langkah pertama adalah menghitung *Gini Impurity* pada simpul akar (*root*), yang menunjukkan nilai 0,5 karena proporsi kelas seimbang (2:2). Selanjutnya dilakukan pemisahan berdasarkan fitur glukosa dengan ambang batas 130. Hasil split menunjukkan bahwa fitur glukosa memberikan pemisahan sempurna: pasien dengan glukosa > 130 semuanya positif, dan yang ⩽ 130 semuanya negatif, menghasilkan *Gini Impurity* = 0 untuk masing-masing cabang.

Pohon keputusan pada Gambar 1. yang dibentuk dari pembelajaran ini menunjukkan bagaimana *Random Forest* dapat memanfaatkan fitur-fitur paling relevan untuk menghasilkan prediksi yang akurat. Untuk memperkuat simulasi, juga digunakan data pasien tambahan dengan fitur klinis yang lebih kompleks, termasuk HbA1c_level, blood_glucose_level, dan BMI. Proses traversal dalam pohon mengikuti aturan-aturan pembelahan berdasarkan nilai ambang yang dihasilkan saat pelatihan. Dalam contoh ini, pasien dengan HbA1c_level 6.7, blood_glucose_level 160, dan BMI 29.0 diproses secara bertahap hingga mencapai simpul akhir, dengan hasil akhir klasifikasi sebagai pasien dengan diabetes.

Step 1: Root Node
Kondisi: HbA1c_level ⩽ 6.15
Nilai Pasien: 6.7 ⟶ Tidak Memenuhi Syarat
Maka proses akan dilanjutkan ke cabang kanan.

Tabel 2. Fitur

Fitur	Nilai
age	55
hypertension	0
heart_disease	0
bmi	29
HbA1c_level	6.7
blood_glucose_level	160
gender_Male	1
smoking_history_current	0
smoking_history_former	0
smoking_history_never	1
smoking_history_not current	0

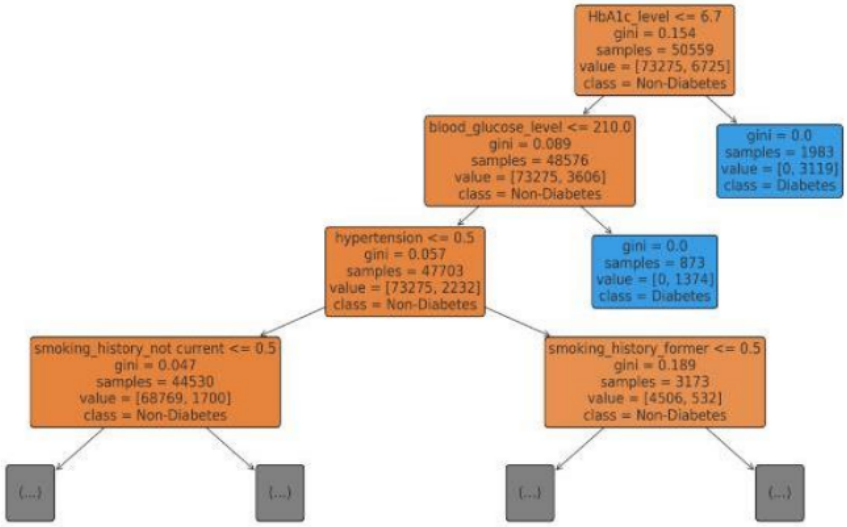


Gambar 1. Root Node

Step 2: Node Kanan
Kondisi: blood_glucose_level ⩽ 164.5 Nilai
Pasien: 160 → Memenuhi Syarat
Maka proses akan dilanjutkan ke cabang kiri dari node ini.

Step 3: Node Kiri
Kondisi: bmi ⩽ 24.85
Nilai Pasien: 29.0 →Tidak Memenuhi Syarat Maka
proses akan dilanjutkan ke cabang kanan. Hasil

Step 4: Akhir: Prediksi
Sampai pada simpul akhir (daun/*leaf*)
Di *node* ini, mayoritas kelas adalah 1 → yaitu Diabetes
Prediksi Akhir untuk Pasien Ini adalah Diabetes (1)

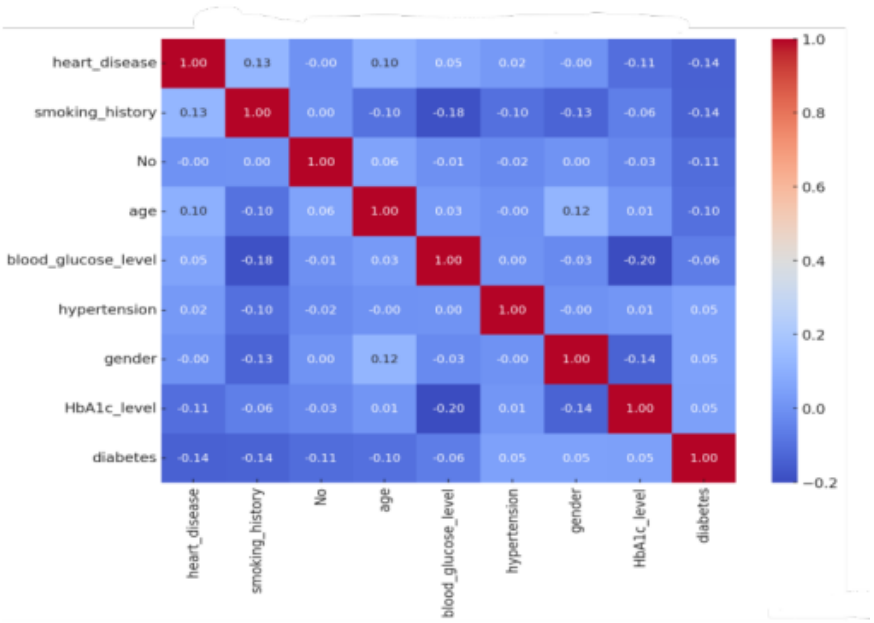


Gambar 2. Pohon Keputusan

Hasil Gambar 2. Menunjukkan bahwa meskipun model *Random Forest* terdiri dari banyak pohon keputusan sederhana, kombinasi prediksinya mampu menghasilkan keputusan akhir yang stabil dan akurat. Simulasi ini juga memperlihatkan bagaimana proses pengambilan keputusan dilakukan secara transparan dan logis, yang penting untuk aplikasi klinis. Dengan fitur-fitur yang mudah diukur secara klinis, model ini memiliki potensi untuk digunakan dalam sistem *skrining* awal, terutama di fasilitas kesehatan primer dengan sumber daya terbatas. Analisis korelasi antar fitur menunjukkan bahwa sebagian besar variabel memiliki nilai korelasi yang relatif rendah terhadap status diabetes, dengan koefisien berada di bawah 0,20. Secara statistik, hal ini mengindikasikan bahwa tidak ada hubungan linier yang kuat antara satu variabel tunggal dengan diagnosis diabetes. Rendahnya nilai korelasi ini tidak serta-merta berarti fitur-fitur tersebut tidak relevan secara medis.

Dalam konteks klinis, diabetes merupakan penyakit multifaktorial yang dipengaruhi oleh kombinasi faktor metabolik, genetik, dan gaya hidup. Misalnya, kadar glukosa darah, BMI, dan riwayat keluarga masing-masing mungkin hanya menunjukkan korelasi lemah jika dilihat secara individual, tetapi justru dapat membentuk pola prediktif yang signifikan ketika dianalisis secara simultan dengan pendekatan *machine learning*. Dengan kata lain, algoritma seperti *Random Forest* dapat menangkap interaksi non-linier antar fitur yang tidak dapat ditunjukkan hanya dengan analisis korelasi sederhana. Selain itu, rendahnya korelasi linier juga mencerminkan kompleksitas diagnosis diabetes di mana tidak ada satu parameter tunggal yang bisa dijadikan indikator mutlak. Sebagai contoh, seorang pasien dengan BMI normal tetap dapat mengalami diabetes apabila memiliki kadar glukosa darah dan HbA1c tinggi, atau sebaliknya.

Meskipun nilai korelasi antar fitur rendah pada Gambar 3. setiap variabel tetap memiliki kontribusi penting dalam meningkatkan sensitivitas model prediksi ketika dikombinasikan. Temuan ini memiliki implikasi medis yang penting. Pertama, tenaga kesehatan tidak dapat hanya mengandalkan satu indikator tung-



Gambar 3. Korelasi Fitur terhadap Diabetes

gal dalam menentukan risiko diabetes, melainkan perlu melihat kombinasi variabel klinis. Kedua, model prediksi berbasis *machine learning* menawarkan pendekatan komplementer dengan memanfaatkan pola kompleks yang tidak terdeteksi oleh analisis statistik konvensional. Hal ini memperkuat argumen bahwa model yang dikembangkan dalam penelitian ini relevan untuk mendukung keputusan medis berbasis data yang lebih holistik Koefisien korelasi *Pearson* antar delapan variabel prediktor yang memiliki korelasi paling kuat dengan variabel target diabetes dalam *dataset* berjumlah 608 sampel. Skala warna menunjukkan kekuatan dan arah hubungan linier: merah tua (nilai +1,00) menandakan korelasi positif sempurna, biru tua (nilai −1,00) menandakan korelasi negatif sempurna, dan abu-abu (nilai 0,00) menandakan tidak ada korelasi linier.

Pengamatan utama pada Gambar 3 :
Setiap fitur berkorelasi sempurna dengan dirinya sendiri. Korelasi fitur target (diabetes):

1. *Heart_disease* dan *smoking_history* menunjukkan korelasi negatif terkuat dengan diabetes ($r \approx \breve{}0,14$), mengindikasikan hubungan negatif lemah.
2. *Age* ($r \approx \breve{}0,10$) dan No (nomor urut sampel, $r \approx \breve{}0,11$) juga berkorelasi negatif lemah.
3. *Blood_glucose_level* dan *HbA1c_level* memiliki korelasi positif kecil ($r \approx +0,05$ dan +0,05), meski secara klinis keduanya penting untuk diagnosis.
4. *Hypertension* dan gender menunjukkan korelasi positif sangat rendah ($r \approx +0,05$).

Hubungan Antar-Fitur:

1. Terdapat korelasi negatif moderat antara *HbA1c_level* dan *blood_glucose_level* ($r \approx \breve{}0,20$), kemungkinan mencerminkan variabilitas pengukuran atau efek pra-pemrosesan data.
2. *Age* berkorelasi moderat dengan gender ($r \approx +0,12$), mengindikasikan sedikit ketidakseimbangan demografis.

Semua koefisien korelasi absolut berada di bawah 0,20, menunjukkan tidak ada hubungan linier yang kuat antara fitur-fitur tunggal dengan status diabetes.

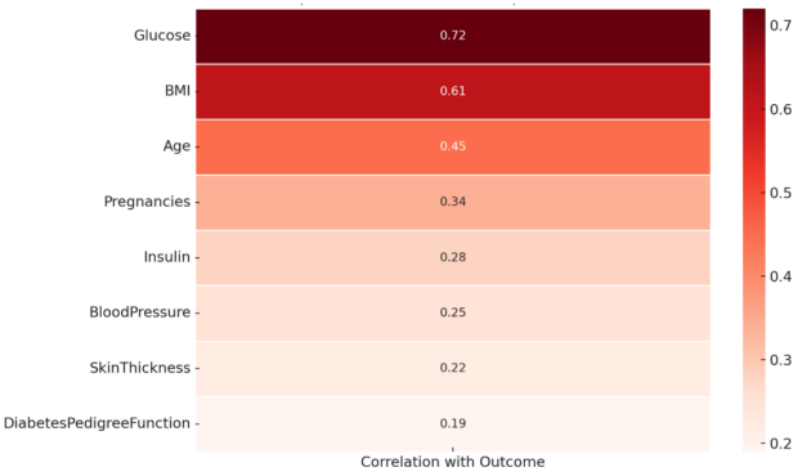
3.2 Evaluasi Kinerja Model

Setelah model Random Forest dibentuk dan diuji dengan data uji sebesar 20% dari keseluruhan *dataset*, dilakukan evaluasi performa untuk mengukur kemampuan prediksi model terhadap kasus diabetes. Empat metrik utama yang digunakan adalah akurasi, presisi, recall, dan F1-score. Hasil evaluasi pada Tabel 3 menunjukkan bahwa model memberikan kinerja yang konsisten dan baik dengan nilai-nilai sebagai berikut:

Tabel 3. Hasil evaluasi

Metrik Evaluasi	Nilai (%)
Akurasi	87
Presisi	84
Recall	82
F1-Score	83

Nilai akurasi sebesar 87% menunjukkan bahwa model mampu mengklasifikasikan pasien secara benar pada sebagian besar kasus. Nilai presisi yang tinggi menandakan bahwa prediksi positif yang diberikan model sebagian besar benar (minim *false positive*), sedangkan nilai *recall* yang tinggi menunjukkan bahwa model juga sensitif dalam mendeteksi kasus diabetes yang sebenarnya (minim *false negative*). F1-score sebesar 83% menunjukkan keseimbangan antara presisi dan *recall*, yang penting dalam konteks medis di mana kesalahan prediksi dapat berdampak signifikan. Untuk mendalami pemahaman tentang kontribusi setiap fitur terhadap hasil prediksi, dilakukan analisis *feature importance*. Hasilnya menunjukkan bahwa fitur-fitur seperti glukosa darah, *HbA1c level*, dan *BMI* memiliki pengaruh paling besar dalam klasifikasi. Visualisasi kontribusi fitur ditunjukkan pada *heatmap* Gambar 4 berikut:



Gambar 4. Kontribusi Fitur terhadap Prediksi Diabetes

Berikut adalah kontribusi fitur terhadap prediksi diabetes. Gambar 4. menunjukkan tingkat korelasi masing-masing fitur (seperti *Glucose, BMI, Age,* dan lainnya) terhadap label diabetes. Semakin tinggi nilainya, semakin besar pengaruh fitur tersebut dalam proses klasifikasi oleh model. Selain itu, visualisasi *confusion matrix* digunakan untuk menampilkan jumlah prediksi benar dan salah pada masing-masing kelas. Ini membantu dalam mengevaluasi sejauh mana model mampu membedakan antara pasien dengan dan tanpa diabetes. Hasil evaluasi secara keseluruhan menunjukkan bahwa *algoritma Random Forest* dapat menjadi alat yang andal dalam sistem pendukung keputusan klinis. Kemampuannya untuk menangani data dengan fitur yang kompleks serta toleransi terhadap data hilang menjadikan algoritma ini cocok untuk implementasi awal pada sistem *skrining* medis berbasis kecerdasan buatan.

4. Simpulan

Penelitian ini berhasil membangun model prediksi *diabetes mellitus* menggunakan *algoritma Random Forest* berbasis data klinis lokal. Model menunjukkan performa yang baik dengan akurasi 87%, presisi 84%, recall 82%, dan F1-score 83%. Analisis feature importance juga menegaskan bahwa kadar glukosa darah, HbA1c level, dan indeks massa tubuh (BMI) merupakan prediktor dominan yang sejalan dengan praktik klinis. Hasil ini mengindikasikan bahwa *algoritma Random Forest* berpotensi mendukung proses *skrining* awal diabetes serta membantu tenaga kesehatan dalam mengambil keputusan berbasis data. Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan. Pertama, ukuran *dataset* masih terbatas (608 sampel) sehingga hasil belum dapat digeneralisasi secara penuh ke populasi yang lebih luas. Kedua, meskipun performa model cukup tinggi, risiko *overfitting* tetap ada, terutama karena jumlah fitur yang relatif banyak dibandingkan ukuran sampel. Ketiga, penelitian ini belum memasukkan faktor gaya hidup dan riwayat medis yang lebih rinci.
Model *Random Forest* pada data klinis lokal menunjukkan potensi sebagai sistem pendukung keputusan medis, tetapi validasi lebih lanjut dengan *dataset* yang lebih besar, beragam, dan melibatkan variabel tambahan sangat diperlukan sebelum implementasi luas dapat dilakukan.

Pustaka

[1] M. S. Efendi *et al.*, "Penerapan algoritma random forest untuk prediksi penjualan dan sistem persediaan produk," *Resolusi: Rekayasa Teknik Informatika dan Informasi*, vol. 5, no. 1, p. 20, 2024.

[2] E. Kogan *et al.*, "A machine learning approach to identifying patients with pulmonary hypertension using real-world electronic health records," *International Journal of Cardiology*, vol. 374, no. December 2022, pp. 95–99, 2023.

[3] E. Martinez-Ríos, L. Montesinos, and M. Alfaro-Ponce, "A machine learning approach for hypertension detection based on photoplethysmography and clinical data," *Computers in Biology and Medicine*, vol. 145, no. January, p. 105479, 2022.

[4] P. Argiento *et al.*, "A pulmonary hypertension targeted algorithm to improve referral to right heart catheterization: A machine learning approach," *Computational and Structural Biotechnology Journal*, vol. 24, no. May, pp. 746–753, 2024.

[5] B. Kovács, F. Tinya, C. Németh, and P. Ódor, "Unfolding the effects of different forestry treatments on microclimate in oak forests: results of a 4-yr experiment," *Ecological Applications*, vol. 30, no. 2, pp. 321–357, 2020.

[6] E. Martinez-Ríos, L. Montesinos, M. Alfaro-Ponce, and L. Pecchia, "A review of machine learning in hypertension detection and blood pressure estimation based on clinical and physiological data," *Biomedical Signal Processing and Control*, vol. 68, no. May, p. 102813, 2021.

[7] N. P. Shetty, J. Shetty, V. Hegde, S. D. Dharne, and M. Kv, "A machine learning-based clinical decision support system for effective stratification of gestational diabetes mellitus and management through ayurveda," *Journal of Ayurveda and Integrative Medicine*, vol. 15, no. 6, p. 101051, 2024.

[8] P. Wändell *et al.*, "A machine learning tool for identifying patients with newly diagnosed diabetes in primary care," *Primary Care Diabetes*, vol. 18, no. June, pp. 501–505, 2024.

[9] J. Wallensten *et al.*, "Machine learning to detect alzheimer's disease with data on drugs and diagnoses," *Journal of Prevention of Alzheimer's Disease*, vol. 12, no. 5, p. 100115, 2025.

[10] N. Aini, M. Arif, I. T. Agustin, and Z. B. Toyibah, "Implementasi algoritma random forest untuk klasifikasi bidang msib di prodi pendidikan informatika," *Jurnal Informatika*, vol. 11, no. 1, pp. 11–16, 2024.

[11] Y. Umemura, N. Okada, H. Ogura, J. Oda, and S. Fujimi, "A machine learning prediction of overt disseminated intravascular coagulation before its progression to an overt stage," *Research and Practice in Thrombosis and Haemostasis*, vol. 8, no. 5, p. 102519, 2024.

[12] D. Kurniawan, E. Rekawati, and J. Sahar, "Pelayanan kesehatan pada lansia dengan hipertensi di tingkat pelayanan primer: systematic review," *Jurnal Kesehatan*, vol. 10, pp. 424–435, 2021.

[13] S. Akter, Z. Liu, E. J. Simoes, and P. Rao, "Using machine learning and electronic health record (ehr) data for the early prediction of alzheimer's disease and related dementias," *Journal of Prevention of Alzheimer's Disease*, no. April, p. 100169, 2025.

[14] S. Jangili, H. Vavilala, G. S. B. Boddeda, S. M. Upadhyayula, R. Adela, and S. R. Mutheneni, "Machine learning-driven early biomarker prediction for type 2 diabetes mellitus associated coronary artery diseases," *Clinical Epidemiology and Global Health*, vol. 24, no. October, p. 101433, 2023.

[15] L. Alomari, Y. Jarrar, Z. Al-Fakhouri, E. Otabor, J. Lam, and J. Alomari, "A machine learning–based risk prediction model for atrial fibrillation in critically ill patients," *Heart Rhythm O2*, pp. 1–9, 2025.