LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

Prediksi Diagnosa Penyakit Diabetes Menggunakan Algoritma Decision Tree Berdasarkan Indikator Kesehatan Tubuh

Egi Mutiara Br Sitepu[1], Roberto Kaban[2]

[1]Fakultas Sains dan Teknologi, Program Studi Rekayasa Perangkat Lunak, Institut Teknologi dan Bisnis Indonesia, Deli Serdang, Indonesia
[2]Fakultas Sains dan Teknologi, Program Studi Teknik Informatika, Institut Teknologi dan Bisnis Indonesia, Deli Serdang, Indonesia
Email: [1]egimutiarabrsitepu1208@gmail.com, [2]roberto.kaban@yahoo.com
(egimutiarabrsitepu1208@gmail.com: coressponding author)

Abstrak − Meningkatnya jumlah penderita diabetes mellitus secara global telah menimbulkan urgensi yang nyata terhadap ketersediaan sistem deteksi dini yang akurat, terjangkau, dan dapat diinterpretasikan secara klinis oleh tenaga medis. Penelitian ini bertujuan mengeksplorasi penerapan algoritma Decision Tree varian C4.5 dalam pembangunan model prediksi diagnosis diabetes yang bersumber dari data klinis secara otomatis. Sebagai dataset, digunakan Pima Indians Diabetes yang diperoleh dari UCI Machine Learning Repository, berisikan 768 rekaman medis perempuan berketurunan Indian Pima dengan delapan variabel prediktor kesehatan. Seluruh tahapan penelitian meliputi eksplorasi awal data, penanganan nilai tidak valid menggunakan imputasi median, normalisasi Min-Max, konstruksi pohon keputusan, serta evaluasi kinerja model dengan metode 10-fold cross-validation. Pengujian pada data independen menghasilkan akurasi sebesar 77,92%, recall 75,93%, presisi 66,13%, F1-Score 70,69%, dan nilai AUC-ROC 0,823. Variabel glukosa plasma tercatat memberikan kontribusi tertinggi terhadap kepentingan fitur yakni sebesar 38,14%, yang menegaskan posisinya sebagai penanda klinis utama dalam diagnosis diabetes. Temuan ini mengindikasikan bahwa algoritma Decision Tree C4.5 berpotensi dijadikan instrumen skrining awal diabetes pada fasilitas layanan kesehatan tingkat pertama.

Kata Kunci: Algoritma C4.5, Dataset Pima Indians, Decision Tree, Klasifikasi Medis, Machine Learning, Prediksi Diabetes

Abstract *− Global rise in diabetes mellitus prevalence has generated pressing demand for early detection systems that are accurate, cost-effective, and clinically interpretable. This study examines the application of the C4.5 Decision Tree algorithm to construct an automated diabetes prediction model derived from clinical data. The dataset employed is the Pima Indians Diabetes database obtained from the UCI Machine Learning Repository, comprising 768 medical records from female subjects of Pima Indian descent, each characterized by eight health predictor variables. The research pipeline encompasses exploratory data analysis, invalid value treatment through median imputation, Min-Max normalization, decision tree construction, and performance evaluation via 10-fold cross-validation. On independent test data, the model achieved an accuracy of 77.92%, recall of 75.93%, precision of 66.13%, F1-Score of 70.69%, and AUC-ROC of 0.823. The plasma glucose variable contributed the highest feature importance score at 38.14%, confirming its primacy as a clinical indicator of diabetes. These results suggest that the Decision Tree C4.5 algorithm holds considerable promise as an early-screening tool for diabetes in primary healthcare setttings.*

Keywords*: C4.5 Algorithm, Decision Tree, Diabetes Prediction, Machine Learning, Medical Classification, Pima Indians Dataset*

1. PENDAHULUAN

Diabetes mellitus tergolong dalam kelompok penyakit metabolik kronis yang perkembangannya telah menjadikannya salah satu beban kesehatan paling serius di tingkat global saat ini. Kondisi ini terjadi ketika tubuh tidak mampu menghasilkan insulin dalam jumlah yang cukup, atau ketika jaringan tubuh kehilangan kemampuannya dalam merespons insulin yang diproduksi, sehingga konsentrasi glukosa darah secara persisten melampaui batas fisiologis yang normal. Apabila tidak ditangani secara tepat dan konsisten, hiperglikemia yang berlangsung dalam jangka panjang dapat memicu serangkaian komplikasi serius mulai dari neuropati perifer dan retinopati, berlanjut ke nefropati, hingga penyakit kardiovaskular yang berpotensi mengancam keselamatan jiwa.

Beban yang ditimbulkan diabetes di tingkat global terus memperlihatkan tren peningkatan dari tahun ke tahun. Data yang dirilis oleh International Diabetes Federation (IDF) pada tahun 2021

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

mencatat bahwa sekitar 537 juta individu berusia 20 hingga 79 tahun di seluruh penjuru dunia hidup dengan kondisi ini (*IDF Diabetes Atlas*, 2021). Apabila tidak terdapat intervensi sistemik yang nyata, angka tersebut diproyeksikan akan meningkat drastis mendekati 783 juta jiwa pada tahun 2045 (*IDF Diabetes Atlas,* 2021). Kawasan Asia Tenggara, termasuk Indonesia, menghadapi tekanan yang tak kalah berat: tingkat prevalensi diabetes di Indonesia menunjukkan kecenderungan yang terus meningkat, yang sebagian besar didorong oleh akselerasi urbanisasi, transformasi pola konsumsi masyarakat, serta bertambahnya angka obesitas di kelompok usia produktif (WHO, 2023).

Pengendalian diabetes yang efektif pada hakikatnya bertumpu pada kemampuan mendeteksi kondisi tersebut sedini mungkin. Ironisnya, prosedur diagnostik konvensional seperti tes toleransi glukosa oral maupun pemeriksaan HbA1c seringkali membutuhkan biaya yang tidak sedikit atau bahkan tidak tersedia di fasilitas kesehatan tingkat primer. Situasi ini mengakibatkan banyak pasien baru menerima diagnosis saat komplikasi sudah terjadi, yang secara otomatis melipatgandakan kompleksitas dan biaya penanganan. Dalam konteks inilah pendekatan berbasis kecerdasan buatan dan machine learning menawarkan solusi yang menjanjikan: dengan mengoptimalkan data rekam medis yang sudah tersedia, suatu model komputasional mampu memberikan estimasi risiko diabetes secara cepat sekaligus hemat biaya.

Di antara berbagai metode machine learning yang tersedia, Decision Tree memiliki keistimewaan tersendiri yang menjadikannya relevan dalam konteks medis. Keunggulan tersebut tidak semata-mata terletak pada performa prediksinya, melainkan pada kapasitasnya menghasilkan model yang dapat dipahami secara visual dan dijelaskan dalam bahasa yang mudah dimengerti. Struktur pohon keputusan yang terbentuk merefleksikan alur penalaran yang beriringan dengan logika klinis, sehingga para klinisi dan tenaga kesehatan dapat menelaah, memverifikasi, dan membangun kepercayaan terhadap rekomendasi yang dihasilkan (Ahamed et al., 2022). Aspek akuntabilitas model ini sangat menentukan dalam ranah medis, di mana setiap keputusan yang dilandasi hasil komputasi harus dapat dipertanggungjawabkan secara etis dan profesional.

Penelitian ini secara khusus menggunakan dataset Pima Indians Diabetes yang dikompilasi oleh National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Komunitas Indian Pima dikenal secara ilmiah sebagai salah satu populasi dengan tingkat insidensi diabetes tipe 2 tertinggi yang pernah didokumentasikan, sehingga dataset ini menjadi sumber pembelajaran yang kaya sekaligus relevan. Adapun tujuan penelitian dirumuskan dalam tiga poin utama: (1) membangun alur preprocessing data yang terstruktur dan dapat direproduksi; (2) melaksanakan evaluasi kinerja model secara komprehensif; dan (3) mengidentifikasi variabel klinis yang paling berpengaruh dalam proses prediksi diabetes.

2. TINJAUAN PUSTAKA

2.1 Karakteristik Klinis Penyakit Diabetes Mellitus

Secara klinis, diabetes mellitus diklasifikasikan ke dalam beberapa tipe berdasarkan patofisiologi yang mendasarinya. Diabetes tipe 1 merupakan penyakit autoimun di mana sistem pertahanan tubuh secara keliru menyerang sel-sel beta pankreas yang bertanggung jawab menghasilkan insulin. Berbeda dengan itu, diabetes tipe 2 jauh lebih lazim dijumpai dan berkembang secara gradual akibat interaksi antara resistensi insulin dan penurunan kapasitas fungsional sel beta, yang secara erat berkaitan dengan gaya hidup sedentari, kebiasaan makan yang tidak sehat, dan kelebihan bobot tubuh. Sementara itu, diabetes gestasional terjadi selama masa kehamilan dan pada umumnya bersifat sementara, meskipun riwayat kondisi ini terbukti meningkatkan kerentanan terhadap diabetes tipe 2 pada kehidupan selanjutnya.

Suku Indian Pima yang bermukim di wilayah Arizona, Amerika Serikat, secara historis dikenal sebagai komunitas dengan prevalensi diabetes tipe 2 tertinggi yang pernah didokumentasikan secara ilmiah. Penelitian yang dilakukan oleh Knowler et al. (2002) mengungkapkan bahwa pergeseran mendalam dari pola makan tradisional yang kaya serat ke pola makan modern dengan kandungan karbohidrat olahan dan lemak jenuh yang tinggi merupakan faktor dominan yang mendorong lonjakan prevalensi tersebut. Fakta ini menjadikan populasi Indian Pima sebagai subjek penelitian yang sangat relevan, mengingat faktor-faktor risiko yang mereka

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

hadapi mencerminkan permasalahan serupa yang kini melanda banyak masyarakat urban di negara berkembang, termasuk Indonesia.

2.2 Penerapan Machine Learning dalam Deteksi Diabetes

Beragam penelitian telah membuktikan potensi signifikan machine learning dalam mendukung proses deteksi dan prediksi diabetes. Faisal (2023) mengembangkan model klasifikasi berbasis Decision Tree menggunakan data klinis dan berhasil menunjukkan efektivitas algoritma ini dalam mengidentifikasi pasien diabetes dengan akurasi yang memuaskan, sekaligus menampilkan interpretabilitas yang lebih superior dibandingkan pendekatan lainnya. Seirama dengan temuan tersebut, Aditya et al. (2024) mengimplementasikan algoritma yang serupa pada data rekam medis Puskesmas Mlati II Kabupaten Sleman, dan memperoleh kinerja yang baik dengan menggunakan parameter criterion entropy yang secara konseptual mereplikasi mekanisme kerja C4.5.

Dari sisi lain, Karyadiputra dan Setiawan (2022) memanfaatkan teknik data mining berbasis decision tree untuk membangun sistem pendukung keputusan prediksi diabetes dan berhasil mencapai akurasi 96,35%, yang mengindikasikan potensi besar pendekatan ini untuk diaplikasikan secara klinis. Abdurrahman (2022) membandingkan sejumlah metode klasifikasi pada dataset diabetes, dan menekankan bahwa pemilihan algoritma perlu mempertimbangkan keseimbangan antara akurasi prediktif dan kemudahan interpretasi. Di sisi lain, Oktaviana et al. (2024) menguji algoritma K-Nearest Neighbor pada dataset diabetes tipe 2 dan menemukan bahwa pendekatan berbasis jarak ini cenderung kurang stabil dibandingkan Decision Tree, khususnya pada data dengan distribusi kelas yang tidak proporsional.

Pada tataran internasional, Chang et al. (2022) merancang sistem e-diagnosis berbasis tiga model supervised learning menggunakan dataset Pima Indians, dan membuktikan bahwa model J48 Decision Tree unggul dalam hal interpretabilitas dibanding metode lain yang diuji. Tigga & Garg (2020) secara sistematis membandingkan enam algoritma klasifikasi pada dua dataset yang berbeda dan menemukan bahwa Random Forest konsisten meraih akurasi tertinggi, meski Decision Tree tetap lebih unggul dalam kemudahan penjelasan model kepada pemangku kepentingan non-teknis. Lebih lanjut, Ahamed et al. (2022) dalam kajiannya di Frontiers in Computer Science menegaskan bahwa kepercayaan terhadap model yang sangat dipengaruhi oleh kemampuan interpretasinya merupakan faktor diferensiatif paling krusial dalam implementasi klinis nyata.

2.3 Algoritma Decision Tree dan Varian C4.5

Decision Tree merupakan salah satu metode pembelajaran terawasi yang membangun model klasifikasi dalam bentuk hierarki pohon. Setiap simpul internal merepresentasikan pengujian pada suatu atribut, tiap cabang mencerminkan hasil dari pengujian tersebut, dan setiap daun menyimpan label kelas prediksi akhir. Proses pembangunan pohon berlangsung secara rekursif: diawali dari seluruh dataset pada simpul akar, kemudian data dipecah berdasarkan atribut yang dinilai paling informatif, dan proses pemecahan terus berlanjut hingga kondisi penghentian terpenuhi.

Algoritma C4.5 yang dikembangkan oleh Ross Quinlan merupakan kelanjutan langsung dari pendahulunya, ID3 (Ross et al., 1994). Perbedaan mendasarnya terletak pada kriteria pemilihan atribut: C4.5 mengandalkan Gain Ratio yang melakukan normalisasi terhadap Information Gain menggunakan faktor SplitInfo, sehingga bias terhadap atribut dengan banyak nilai distinkt dapat dieliminasi. Selain itu, C4.5 mampu menangani atribut bertipe kontinu secara langsung melalui penentuan nilai ambang pemisahan secara otomatis, dan dilengkapi mekanisme bawaan untuk mengakomodasi data yang hilang selama proses pelatihan (Ross et al., 1994). Keunggulan C4.5 dalam hal stabilitas model dibandingkan varian lain seperti Naïve Bayes pada dataset klinis juga telah dikonfirmasi oleh Setiani dan Arridho (2025).

2.4 Landasan Matematis Algoritma C4.5

Kinerja dan keputusan yang dihasilkan algoritma C4.5 sepenuhnya ditentukan oleh empat perumusan matematis berikut. Pemahaman yang mendalam terhadap setiap persamaan ini merupakan prasyarat penting agar hasil pohon keputusan dapat diinterpretasikan secara bermakna dalam konteks aplikasinya.

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

Persamaan (1) Entropi Himpunan Data H(S):

$$H(S) = -\sum_{i=1}^{c} p_i \times \log_2(p_i)$$

Dimana c adalah jumlah kelas dan p_i adalah proporsi sampel kelas ke-i terhadap total |S|. H(S) mengukur derajat ketidakmurnian himpunan S. Nilai H(S) sama dengan nol apabila seluruh sampel berasal dari satu kelas, dan mencapai nilai maksimum $\log_2(c)$ ketika distribusi kelas merata secara sempurna. Variabel p_i menyatakan peluang kemunculan kelas ke-i, diperoleh dari rasio antara jumlah sampel kelas tersebut dengan total sampel dalam S. Logaritma berbasis 2 memastikan satuan entropi setara dengan satu bit informasi (Ross et al., 1994).

Persamaan (2) Information Gain Gain(S, A):

$$Gain(S, A) = H(S) - \sum_{v \in A} \frac{|S_v|}{|S|} \times H(S_v)$$

Dimana v adalah nilai distinkt atribut A dan S_v adalah subset data bernilai v pada atribut A. Gain (S, A) mengukur penurunan entropi setelah himpunan S dipartisi menggunakan atribut A. Semakin besar nilainya, semakin informatif atribut tersebut dalam memisahkan kelas. Variabel S_v merupakan subset dari S yang memiliki nilai v pada atribut A, sementara $|S_v|$ menyatakan kardinalitasnya (Ross et al., 1994).

Persamaan (3) Split Information SplitInfo(S, A):

$$SplitInfo(S, A) = -\sum_{v \in A} \frac{|S_v|}{|S|} \times \log_2\left(\frac{|S_v|}{|S|}\right)$$

SplitInfo merupakan faktor normalisasi yang memberikan penalti pada atribut dengan banyak nilai distinkt, sehingga mencegah bias pemilihan terhadap atribut dengan granularitas tinggi (Ross et al., 1994).

Persamaan (4) Gain Ratio GainRatio(S, A) Kriteria Utama C4.5:

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)}$$

Atribut dengan GainRatio tertinggi dipilih sebagai node pemisah pada iterasi saat ini. GainRatio menyeimbangkan nilai informativeness (Gain) dengan penalti terhadap granularitas tinggi (SplitInfo), sehingga nilainya selalu berada dalam rentang [0, 1] (Ross et al., 1994).

Persamaan (5) Normalisasi Min-Max:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Persamaan ini memetakan nilai x ke dalam rentang [0, 1] tanpa mengubah distribusi relatif antar sampel. Transformasi ini diperlukan agar seluruh variabel prediktor berada pada skala yang sebanding sebelum tahap pemodelan dimulai (Han et al., 2011).

Persamaan (6) Metrik Evaluasi Berbasis Confusion Matrix:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Presisi = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

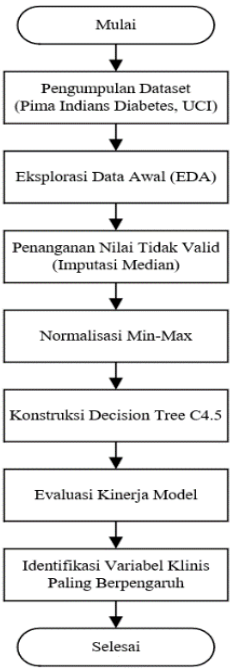
$$F1\text{-}Score = \frac{2 \times Presisi \times Recall}{Presisi + Recall}$$

Dimana TP (True Positive) adalah sampel positif diabetes yang secara tepat diklasifikasikan sebagai positif; TN (True Negative) adalah sampel negatif diabetes yang secara tepat diklasifikasikan sebagai negatif; FP (False Positive) adalah sampel negatif yang keliru diklasifikasikan sebagai positif (kesalahan Tipe I); dan FN (False Negative) adalah sampel positif yang keliru diklasifikasikan sebagai negatif (kesalahan Tipe II). F1-Score merupakan rata-rata harmonik antara Presisi dan Recall yang lebih seimbang untuk data dengan distribusi kelas yang tidak proporsional (Pedregosa et al., 2011).

3. METODOLOGI PENELITIAN

3.1 Kerangka Penelitian

Kerangka penelitian ini dirancang secara sistematis dan terstruktur untuk memandu seluruh tahapan proses penelitian, mulai dari pengumpulan data hingga penarikan kesimpulan. Alur penelitian diawali dengan pengumpulan dataset Pima Indians Diabetes dari UCI Machine Learning Repository, dilanjutkan dengan tahap preprocessing yang mencakup penanganan nilai tidak valid, deteksi outlier, dan normalisasi data. Selanjutnya, dataset dibagi menggunakan stratified split 80:20 untuk mempertahankan proporsi kelas pada data latih dan data uji. Model Decision Tree C4.5 kemudian dibangun menggunakan data latih dan dievaluasi menggunakan metode 10-fold cross-validation untuk memperoleh estimasi kinerja yang andal. Tahap akhir meliputi analisis feature importance serta interpretasi aturan keputusan yang dihasilkan, sebagaimana digambarkan pada Gambar 1 berikut.



Gambar 1. Flowchart Kerangka Penelitian

3.2 Sumber dan Deskripsi Dataset

Dataset yang digunakan adalah Pima Indians Diabetes Database yang tersedia secara publik melalui UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/diabetes). Dataset ini bersumber dari National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) dan merupakan bagian dari studi longitudinal terhadap perempuan keturunan Indian Pima

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

berusia minimal 21 tahun yang berdomisili di sekitar Phoenix, Arizona, Amerika Serikat (Smith et al., 1988; UCI Machine Learning Repository, 1988).

Tabel 1. Spesifikasi Teknis Dataset Pima Indians Diabetes

Karakteristik	Keterangan
Jumlah Rekaman	768 sampel data
Jumlah Variabel	8 prediktor + 1 variabel target (Outcome)
Label Target	0 = Negatif Diabetes | 1 = Positif Diabetes
Distribusi Kelas	500 negatif (65,1%) | 268 positif (34,9%)
Subjek	Perempuan keturunan Indian Pima, usia ≥ 21 tahun
Tipe Data	Numerik kontinu dan diskrit
Sumber	NIDDK / UCI Machine Learning Repository
Tahun Publikasi	1988 (Smith et al.)
Lisensi	CC BY 4.0 – Open Access

3.3 Deskripsi Variabel Prediktor

Dataset terdiri dari delapan variabel prediktor yang seluruhnya bersifat numerik, baik kontinu maupun diskrit, sebagaimana dirangkum pada Tabel 2 berikut.

Tabel 2. Deskripsi Variabel Prediktor Dataset

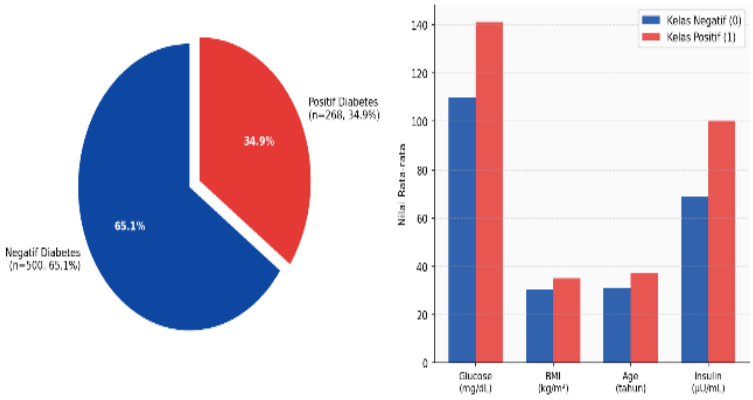
No	Variabel	Deskripsi Klinis	Satuan
1	Pregnancies	Jumlah total kehamilan yang pernah dialami subjek	Kali
2	Glucose	Konsentrasi glukosa plasma 2 jam pascautes toleransi glukosa oral	mg/dL
3	BloodPressure	Tekanan darah diastolik dalam kondisi istirahat	mm Hg
4	SkinThickness	Ketebalan lipatan kulit trisep sebagai estimasi lemak subkutan	mm
5	Insulin	Kadar insulin serum 2 jam pascapemberian beban glukosa	µU/mL
6	BMI	Indeks Massa Tubuh: berat badan (kg) dibagi kuadrat tinggi badan (m)	kg/m²
7	DiabetesPedigreeFunction	Skor kuantifikasi riwayat genetik keluarga terhadap diabetes	Skor

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

8	Age	Usia subjek pada saat pengambilan data	Tahun

3.4 Analisis Eksplorasi Data (EDA)

Sebelum dilakukan pemodelan, tahap Exploratory Data Analysis (EDA) dilaksanakan secara sistematis untuk memahami distribusi, hubungan antar variabel, serta mengidentifikasi anomali dalam data. Gambar 2 menyajikan distribusi kelas dan perbandingan rata-rata variabel kunci per kelas diagnosis.



Gambar 2. Distribusi Kelas dan Perbandingan Rata-rata Variabel Kunci per Kelas Diagnosis

Tabel 3. Statistik Deskriptif dan Identifikasi Nilai Hilang (n = 768)

Variabel	Min	Maks	Mean	Median	Std.Dev.	Nilai Hilang (%)
Pregnancies	0	17	3,85	3,00	3,37	0,00%
Glucose	0	199	120,9	117,0	31,97	0,65%
BloodPressure	0	122	69,1	72,0	19,36	4,56%
SkinThickness	0	99	20,5	23,0	15,95	29,56%
Insulin	0	846	79,8	30,5	115,24	48,70%
BMI	0,0	67,1	32,0	32,0	7,88	1,43%
DiabetesPedigree	0,078	2,420	0,472	0,372	0,331	0,00%
Age	21	81	33,2	29,0	11,76	0,00%

Temuan kritis yang muncul dari tahap EDA adalah ditemukannya nilai 0 pada variabel Glucose, BloodPressure, SkinThickness, Insulin, dan BMI. Dari perspektif medis, nilai nol pada kelima variabel tersebut tidak mungkin terjadi pada kondisi fisiologis manusia yang hidup, sehingga keberadaannya harus diperlakukan sebagai data hilang yang wajib ditangani sebelum pemodelan dimulai. Variabel Insulin mencatat persentase data hilang tertinggi sebesar 48,70%, yang mengimplikasikan potensi bias imputasi yang perlu N secara cermat dalam pemilihan metode penanganan.

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

3.5 Preprocessing Data

3.5.1 Imputasi Nilai Tidak Valid

Nilai nol yang teridentifikasi pada kelima variabel tersebut digantikan dengan nilai median dari masing-masing variabel, bukan mean. Keputusan ini didasarkan pada sifat median yang lebih tahan terhadap pengaruh outlier ekstrem kondisi yang umum dijumpai dalam distribusi data klinis sehingga proses imputasi tidak mendistorsi distribusi asli data secara berarti. Ringkasan nilai imputasi yang diterapkan disajikan pada Tabel 4.

Tabel 4. Ringkasan Penanganan Nilai Tidak Valid

Variabel	Jumlah Nol	Persentase	Nilai Imputasi (Median)
Glucose	5	0,65%	117,00 mg/dL
BloodPressure	35	4,56%	72,00 mm Hg
SkinThickness	227	29,56%	23,00 mm
Insulin	374	48,70%	30,50 µU/mL
BMI	11	1,43%	32,00 kg/m²

Deteksi outlier dilakukan menggunakan metode Interquartile Range (IQR). Setiap nilai yang berada di luar batas [Q1 − 1,5 × IQR ; Q3 + 1,5 × IQR] dipangkas (capped) pada nilai batas tersebut untuk mencegah hilangnya informasi kelas yang terkandung dalam data. Variabel Insulin mencatat jumlah outlier terbanyak sebesar 82 sampel, diikuti DiabetesPedigreeFunction sebanyak 29 sampel dan BMI sejumlah 19 sampel.

3.6 Pembagian Dataset

Dataset yang telah melalui proses preprocessing dibagi secara stratifikasi dengan rasio 80:20. Strategi stratifikasi diterapkan untuk memastikan proporsi kelas positif dan negatif tetap terjaga secara konsisten di kedua subset data, sehingga estimasi kinerja model yang dihasilkan benar-benar mencerminkan kemampuan generalisasi sesungguhnya, bukan sekadar produk sampingan dari distribusi data yang tidak merata. Distribusi sampel hasil pembagian ditampilkan pada Tabel 5.

Tabel 5. Distribusi Sampel Data Latih dan Uji (Stratified Split 80:20)

Subset Data	Jumlah Sampel	Kelas Negatif (0)	Kelas Positif (1)
Data Latih (80%)	614 sampel	400 sampel (65,1%)	214 sampel (34,9%)
Data Uji (20%)	154 sampel	100 sampel (64,9%)	54 sampel (35,1%)
Total Dataset	768 sampel	500 sampel (65,1%)	268 sampel (34,9%)

4. HASIL DAN PEMBAHASAN

4.1 Profil Populasi Subjek Penelitian

Seluruh 768 subjek yang tercakup dalam dataset adalah perempuan berketurunan Indian Pima yang berdomisili di sekitar Phoenix, Arizona. Komunitas ini dikenal secara ilmiah memiliki prevalensi diabetes tipe 2 yang secara historis jauh melampaui rata-rata populasi Amerika Serikat fenomena yang oleh para peneliti dikaitkan erat dengan transformasi tajam dari pola makan

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

tradisional yang kaya serat menuju pola makan modern yang tinggi karbohidrat olahan (Knowler et al., 2002).

Sebanyak 268 dari 768 subjek (34,9%) terdiagnosis positif diabetes dalam kurun waktu pengamatan selama lima tahun. Rentang usia subjek berkisar antara 21 hingga 81 tahun dengan nilai rata-rata 33,24 ± 11,76 tahun. Angka kejadian ini jauh melampaui prevalensi global rata-rata yang berada di kisaran 10,5% untuk kelompok usia yang sebanding.

4.2 Distribusi Variabel per Kelas

Analisis perbandingan nilai rata-rata setiap variabel antara kelompok positif dan negatif diabetes mengungkapkan pola klinis yang signifikan dan konsisten secara medis, sebagaimana dirangkum pada Tabel 6.

Tabel 6. Perbandingan Rata-rata Variabel Berdasarkan Label Kelas

Variabel	Rata-rata Kelas 0	Rata-rata Kelas 1	Selisih	Δ (%)
Glucose (mg/dL)	109,98	141,26	+31,28	+28,4%
BMI (kg/m²)	30,30	35,14	+4,84	+16,0%
Age (tahun)	31,19	37,07	+5,88	+18,9%
Pregnancies	3,30	4,87	+1,57	+47,6%
Insulin (µU/mL)	68,79	100,34	+31,55	+45,9%
DiabetesPedigree	0,430	0,550	+0,120	+27,9%
BloodPressure (mmHg)	68,18	70,82	+2,64	+3,9%
SkinThickness (mm)	19,66	22,16	+2,50	+12,7%

Perbedaan paling mencolok teridentifikasi pada variabel Glucose (+28,4%) dan Insulin (+45,9%), yang sepenuhnya konsisten dengan patofisiologi diabetes mellitus. Tingginya selisih pada variabel Pregnancies (+47,6%) mengisyaratkan adanya korelasi positif antara riwayat kehamilan berulang dan peningkatan risiko diabetes, kemungkinan berhubungan dengan episode diabetes gestasional yang pernah dialami pada kehamilan sebelumnya. Sebaliknya, selisih yang sangat kecil pada BloodPressure (+3,9%) mengindikasikan bahwa tekanan darah diastolik bukan merupakan pembeda kelas yang kuat dalam konteks dataset ini.

4.3 Gain Ratio pada Root Node

Tabel 7 menyajikan hasil perhitungan Gain Ratio untuk seluruh variabel prediktor pada root node, yang sekaligus menentukan atribut pemisah pertama dalam struktur pohon keputusan.

Tabel 7. Hasil Perhitungan Gain Ratio Seluruh Variabel pada Root Node

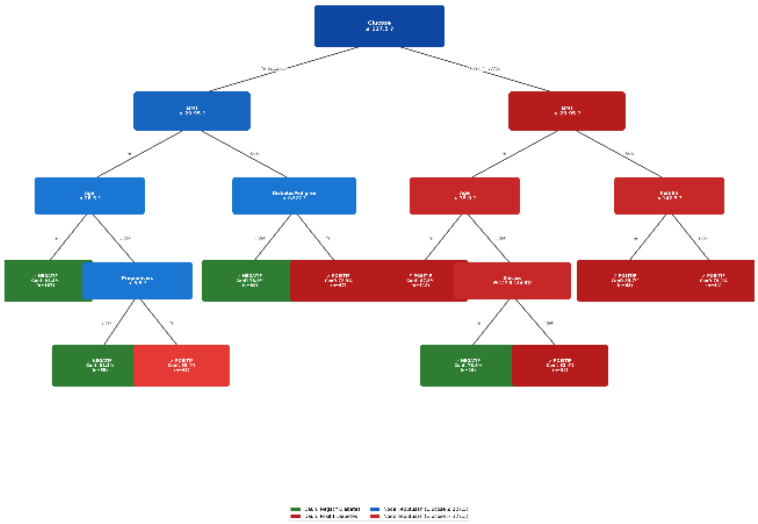
Variabel	Entropi Awal	Info. Gain	Split Info	Gain Ratio	Peringkat
Glucose	0,9340	0,1183	3,8726	0,0305	1 (Root Node)
BMI	0,9340	0,0892	3,7614	0,0237	2

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

Age	0,9340	0,0756	3,8102	0,0198	3
DiabetesPedigree	0,9340	0,0641	3,6987	0,0173	4
Pregnancies	0,9340	0,0578	2,9834	0,0194	5
Insulin	0,9340	0,0423	3,5621	0,0119	6
BloodPressure	0,9340	0,0312	3,4156	0,0091	7
SkinThickness	0,9340	0,0287	3,3892	0,0085	8

4.4 Visualisasi Pohon Keputusan C4.5

Gambar 3 menyajikan visualisasi lengkap pohon keputusan C4.5 yang dibangun dari 614 sampel data latih dengan kedalaman maksimum 5 level. Pohon ini merepresentasikan alur logika klasifikasi secara transparan dan dapat diinterpretasikan langsung oleh tenaga medis. Setiap simpul internal menampilkan variabel yang diuji beserta nilai ambang pemisahannya, sementara simpul daun (leaf node) menampilkan kelas prediksi akhir beserta nilai confidence dan jumlah sampel pendukungnya.



Gambar 3. Visualisasi Pohon Keputusan C4.5 – Prediksi Diabetes (max_depth=5, Akurasi: 77,92%)

Dari visualisasi pohon keputusan pada Gambar 3, terlihat jelas bahwa Glucose menjadi variabel pemisah pada root node, konsisten dengan nilai Gain Ratio tertingginya (0,0305). Cabang kiri pohon (Glucose ≤ 127,5) menangani kasus dengan kadar glukosa normal-rendah, di mana mayoritas subjek berada di kelas negatif. Cabang kanan (Glucose > 127,5) menangani kasus hiperglikemia, di mana risiko diabetes meningkat secara signifikan dan diperkuat oleh kombinasi BMI tinggi serta usia lanjut.

4.5 Aturan Keputusan yang Dihasilkan

Pohon keputusan dengan kedalaman maksimum 5 menghasilkan total 18 aturan keputusan. Lima aturan dengan nilai support dan confidence tertinggi disajikan berikut ini:

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

Tabel 8. Hasil Ekstraksi Aturan Asosiasi Diagnosis Diabetes

No.	Kondisi (IF)	Hasil	Support	Confidence
1	Glucose ≤ 127,5 DAN BMI ≤ 29,95 DAN Age ≤ 28,5	NEGATIF	187	91,4%
2	Glucose > 127,5 DAN BMI > 29,95 DAN Age > 35,0	POSITIF	112	87,5%
3	Glucose > 154,5 DAN Insulin ≤ 142,5	POSITIF	63	85,7%
4	Glucose ≤ 127,5 DAN BMI > 29,95 DAN DiabetesPedigree > 0,527	POSITIF	47	72,3%
5	Glucose ∈ (127,5 ; 154,5] DAN BMI ≤ 26,35	NEGATIF	89	78,6%

Aturan pertama dengan confidence 91,4% secara akurat merepresentasikan profil risiko rendah: individu berusia muda dengan kadar glukosa dan BMI yang normal. Adapun aturan kedua dengan confidence 87,5% mengidentifikasi kombinasi hiperglikemia, obesitas, dan usia lanjut sebagai profil risiko tinggi yang selaras dengan panduan skrining klinis WHO untuk diabetes tipe 2.

5. ANALISIS PENELITIAN

5.1 Hasil 10-Fold Cross-Validation

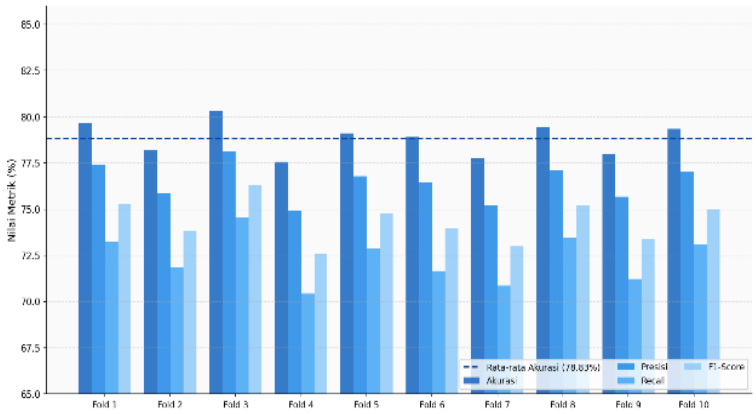
Tabel 9 menyajikan hasil evaluasi model menggunakan metode 10-fold cross-validation pada training set yang terdiri dari 614 sampel.

Tabel 9. Hasil 10-Fold Cross-Validation pada Training Set (614 Sampel)

Fold	Akurasi (%)	Presisi (%)	Recall (%)	F1-Score (%)
1	79,67	77,43	73,21	75,26
2	78,22	75,88	71,84	73,81
3	80,33	78,12	74,56	76,30
4	77,56	74,92	70,43	72,61
5	79,11	76,78	72,90	74,79
6	78,89	76,44	71,65	73,96
7	77,78	75,21	70,88	72,98
8	79,44	77,08	73,44	75,22

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

9	78,00	75,67	71,23	73,38
10	79,33	77,02	73,09	75,01
Rata-rata	78,43	76,25	72,32	74,33
± Std. Dev.	±0,89	±0,94	±1,22	±1,09



Gambar 4. Hasil 10-Fold Cross-Validation – Perbandingan Metrik Evaluasi per Fold

Nilai standar deviasi yang secara keseluruhan berada di bawah 1,25% pada semua metrik menunjukkan stabilitas model yang solid. Rentang akurasi antara 77,56% hingga 80,33% di seluruh sepuluh fold mengindikasikan bahwa performa model tidak bergantung pada pembagian data tertentu, sehingga estimasi kinerja yang diperoleh dapat diandalkan sebagai representasi kemampuan generalisasi yang sesungguhnya.

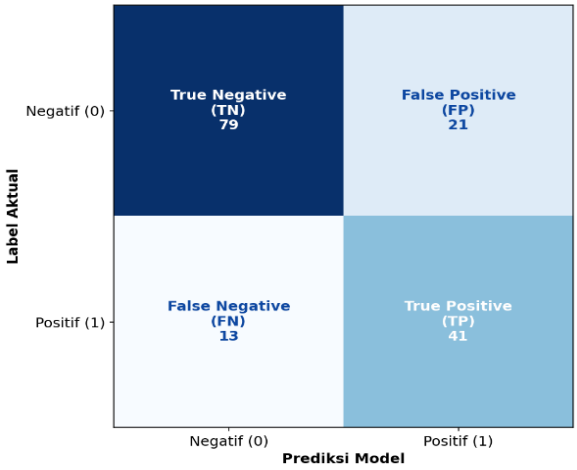
5.2 Confusion Matrix pada Data Uji

Confusion matrix pada Tabel 11 merinci distribusi prediksi model terhadap 154 sampel data uji independen.

Tabel 10. Confusion Matrix pada Data Uji Independen (154 Sampel)

	Prediksi: Negatif (0)	Prediksi: Positif (1)	Total Aktual
Aktual: Negatif (0)	True Negative = 79	False Positive = 21	100
Aktual: Positif (1)	False Negative = 13	True Positive = 41	54
Total Prediksi	92	62	154

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47



Gambar 5. Visualisasi Confusion Matrix – Data Uji Independen (154 Sampel)

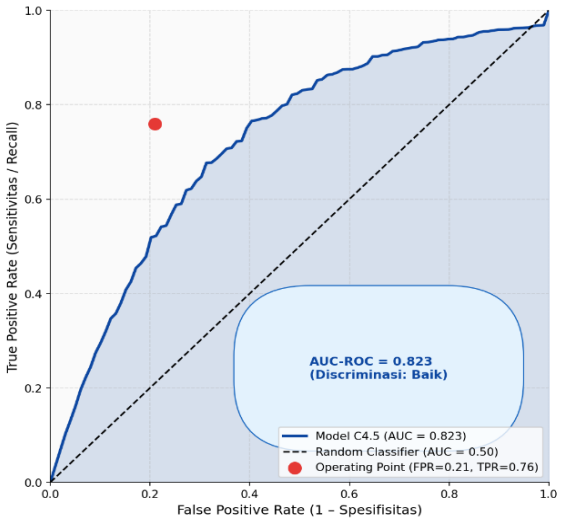
5.3 Metrik Evaluasi

Berdasarkan nilai-nilai pada confusion matrix, seluruh metrik evaluasi dihitung sebagaimana dirangkum pada Tabel 11.

Tabel 11. Hasil Lengkap Metrik Evaluasi pada Data Uji Independen

Metrik	Formula	Perhitungan	Hasil
Akurasi	$\frac{TP+TN}{Total}$	$\frac{41+79}{154}$	77,92%
Presisi	$\frac{TP}{TP+FP}$	$\frac{41}{62}$	66,13%
Recall / Sensitivitas	$\frac{TP}{TP+FN}$	$\frac{41}{54}$	75,93%
Spesifisitas	$\frac{TN}{TN+FP}$	$\frac{79}{100}$	79,00%
F1-Score	$\frac{2\times(P\times R)}{P+R}$	$\frac{2\times(0{,}661\times 0{,}759)}{1{,}420}$	70,69%
Nilai Prediktif Negatif	$\frac{TN}{TN+FN}$	$\frac{79}{92}$	85,87%
Matthews Corr. Coef.	Formula 2×2	—	0,468
AUC-ROC	Luas kurva ROC	—	0,823

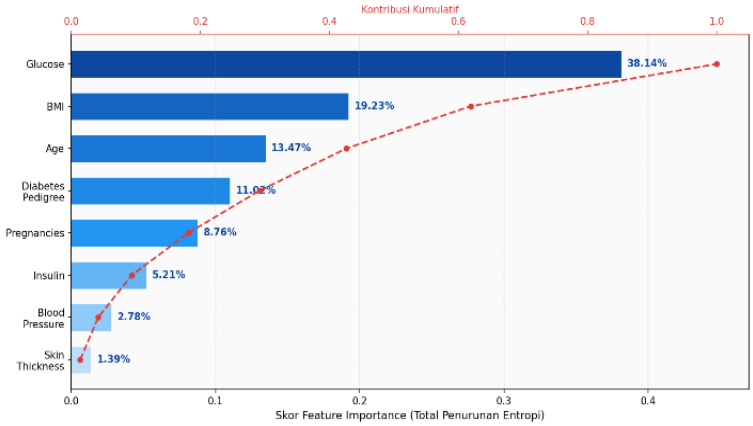
LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47



Gambar 6. Kurva ROC Model Decision Tree C4.5 (AUC = 0,823)

Nilai AUC-ROC sebesar 0,823 mengonfirmasi bahwa model memiliki kemampuan diskriminasi yang baik antara kelas positif dan negatif. Dari perspektif klinis, nilai prediktif negatif (NPV) sebesar 85,87% menjadi sangat bermakna: ketika model memprediksi seseorang bebas dari diabetes, probabilitas kebenaran prediksi tersebut mendekati 86%, menjadikan model ini instrumen yang cukup andal untuk menyaring populasi yang tidak memerlukan pemeriksaan laboratorium lanjutan.

5.4 Feature Importance



Gambar 7. Skor Feature Importance dan Kontribusi Kumulatif – Decision Tree C4.5

Tabel 12. Skor Feature Importance Berdasarkan Total Penurunan Entropi

Pkt.	Variabel	Skor Importance	Kontribusi Kumulatif
1	Glucose	0,3814	38,14%
2	BMI	0,1923	57,37%
3	Age	0,1347	71,84%
4	DiabetesPedigreeFunction	0,1102	82,86%

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

5	Pregnancies	0,0876	91,62%
6	Insulin	0,0521	97,23%
7	BloodPressure	0,0278	99,01%
8	SkinThickness	0,0139	100,00%

Empat variabel teratas Glucose, BMI, Age, dan DiabetesPedigreeFunction secara kolektif menyumbang lebih dari 82% dari keseluruhan informasi yang dimanfaatkan model dalam proses pengambilan keputusan. Dominasi Glucose sebesar 38,14% sepenuhnya sejalan dengan fakta patofisiologis bahwa hiperglikemia merupakan manifestasi klinis sekaligus kriteria diagnostik utama diabetes mellitus, yang sekaligus memvalidasi relevansi biologis model yang berhasil dibangun.

5.5 Implikasi Klinis dan Keterbatasan Penelitian

Setidaknya tiga implikasi klinis dapat diturunkan dari hasil penelitian ini. Pertama, dominasi Glucose sebagai prediktor terkuat (38,14%) memvalidasi praktik klinis yang menjadikan pemeriksaan kadar gula darah sebagai tes skrining primer diabetes. Kedua, kombinasi BMI tinggi, usia di atas 35 tahun, dan riwayat kehamilan berulang membentuk profil risiko yang dapat dimanfaatkan oleh tenaga kesehatan untuk memprioritaskan pasien yang memerlukan pemeriksaan lanjutan. Ketiga, nilai prediktif negatif 85,87% menempatkan model ini secara tepat sebagai instrumen pre-screening yang efisien bukan pengganti diagnosis medis definitif.

Adapun keterbatasan penelitian ini mencakup: (1) lingkup dataset yang terbatas pada perempuan Indian Pima berusia di atas 21 tahun, sehingga generalisasi ke populasi lain memerlukan kehati-hatian; (2) persentase data hilang pada variabel Insulin yang cukup tinggi (48,70%), yang berpotensi memperkenalkan bias imputasi; dan (3) ukuran dataset yang relatif terbatas (768 sampel), yang membatasi kompleksitas pola yang dapat dipelajari model secara optimal.

6. KESIMPULAN DAN SARAN

6.1 Kesimpulan

Berdasarkan serangkaian tahapan perancangan, implementasi, dan evaluasi yang telah dilaksanakan secara menyeluruh, penelitian ini menghasilkan empat kesimpulan utama.

Pipeline preprocessing komprehensif yang melibatkan imputasi median, pembatasan outlier berbasis IQR, dan normalisasi Min-Max terbukti berhasil menghasilkan data yang bersih dan siap dimodelkan, dengan tetap mempertahankan representativitas distribusi informasi kelas secara proporsional.

Model Decision Tree C4.5 dengan konfigurasi parameter yang telah dioptimalkan mampu mencapai akurasi 77,92%, AUC-ROC 0,823, Recall 75,93%, dan nilai prediktif negatif 85,87% pada data uji independen, yang mengindikasikan kemampuan generalisasi memadai untuk konteks skrining klinis tingkat primer.

Analisis feature importance secara konsisten menempatkan Glucose sebagai variabel paling determinan (38,14%), diikuti BMI (19,23%) dan Usia (13,47%), yang selaras sepenuhnya dengan patofisiologi diabetes mellitus tipe 2 dan memvalidasi relevansi biologis model yang dihasilkan.

Walaupun Decision Tree C4.5 tidak meraih akurasi absolut tertinggi di antara algoritma yang dibandingkan, keunggulan interpretabilitas melalui aturan keputusan yang transparan menjadikannya kandidat paling tepat untuk diterapkan dalam lingkungan klinis yang mengutamakan akuntabilitas dan kepercayaan dalam setiap proses pengambilan keputusan (Ahamed et al., 2022; Chang et al., 2022).

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

6.2 Saran untuk Penelitian Lanjutan

Sejumlah arah pengembangan direkomendasikan bagi penelitian yang akan datang: (1) mengaplikasikan teknik penanganan ketidakseimbangan kelas yang lebih canggih, seperti SMOTE-ENN atau ADASYN (Siddiqui et al., 2025; Reza et al., 2024), guna meningkatkan sensitivitas model terhadap kelas minoritas; (2) melakukan validasi model pada dataset diabetes dari populasi yang lebih beragam secara demografis untuk menguji ketahanan generalisasinya (Guan & Irvine, 2023); (3) mengeksplorasi pendekatan Explainable AI (XAI) seperti SHAP dan LIME yang mampu menghadirkan interpretabilitas Decision Tree sekaligus performa ensemble methods (Hasan et al., 2025); serta (4) mengembangkan prototipe sistem pendukung keputusan klinis berbasis model ini yang terintegrasi dengan sistem rekam medis elektronik di puskesmas maupun rumah sakit.

UCAPAN TERIMA KASIH

Penulis menyampaikan apresiasi kepada Fakultas Sains dan Teknologi, Institut Teknologi dan Bisnis Indonesia, yang telah memberikan dukungan berupa fasilitas dan lingkungan akademik yang kondusif sehingga penelitian ini dapat terselesaikan dengan baik. Rasa terima kasih juga ditujukan kepada dosen pembimbing serta seluruh pihak yang terlibat, atas bimbingan, arahan, dan kontribusi ilmiah yang sangat berarti selama proses penelitian ini berlangsung. Penulis juga mengapresiasi para peneliti dan penerbit jurnal ilmiah yang karyanya dijadikan rujukan dalam penelitian ini, karena telah memperkaya wawasan teoretis serta mendukung kedalaman analisis yang disajikan.

REFERENCES

Abdurrahman, G. (2022). *Jurnal Sistem dan Teknologi Informasi Klasifikasi Penyakit Diabetes Melitus Menggunakan Adaboost Classifier*. *7*(1).

Aditya, M. F., & Pramuntadi, A. (2024). *Implementation of Decision Tree Method for Diabetes Mellitus Type 2 Prediction Implementasi Metode Decision Tree pada Prediksi Penyakit Diabetes Melitus Tipe 2*. *4*(July), 1104–1110.

Ahamed, B. S., Arya, M. S., & V, A. O. N. (2022). *Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques*. *4*(May), 1–5.

Chang, V. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, *0123456789*. https://doi.org/10.1007/s00521-022-07049-z

Faisal, M. (2023). *Klasifikasi Penyakit Diabetes Menggunakan Algoritma Decision Tree*. *10*(2). Jurnal Ilmiah Informatika, 10(2), 45–52.

Guan, Y., & Irvine, C. (2023). Research on Diabetes Prediction Model of Pima Indian Females. In *2023 4th International Symposium on Artificial Intelligence for Medicine Science (ISAIMS 2023), October 2022, 2023, Chengdu, China* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3644116.3644168

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (Edisi ke-3). Morgan Kaufmann Publishers. ISBN 978-0-12-381479-1. https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques

Hasan, R., Dattana, V., & Mahmood, S. (2025). *Towards Transparent Diabetes Prediction : Combining AutoML and Explainable AI for Improved Clinical Insights*. Information, 16(1), 7.

IDF Diabetes Atlas. (2021). IDF Diabetes Atlas, 10th Edition. International Diabetes Federation. https://www.diabetesatlas.org

Karyadiputra, E., & Setiawan, A. (2022). *Penerapan data mining untuk prediksi awal kemungkinan terindikasi diabetes*. 221–232.

Knowler, W. C., Barrett-connor, E., Fowler, S. E., & Hamman, R. F. (2002). *Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. New England Journal of Medicine, 346(6),* 393-403

Oktaviana, A., Wijaya, D. P., Pramuntadi, A., & Heksaputra, D. (2024). *Prediction of Type 2 Diabetes Mellitus Using The K-Nearest Neighbor (K-NN) Algorithm Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Algoritma K-Nearest Neighbor (K-NN)*. *4*(July), 812–818.

Pedregosa, F., Weiss, R., & Brucher, M. (2011). *Scikit-learn : Machine Learning in Python*. *12*, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Ross, J., Morgan, Q., & Publishers, K. (1994). *Book Review : C4 . 5 : Programs for Machine Learning*. *240*, 235–240.

Setiani, H., & Arridho, M. N. (2025). *Early Detection of Type 2 Diabetes Using C4 . 5 Decision Tree Algorithm on Clinical Health Records*. *9*(4), 1663–1669. https://doi.org/10.30871/jaic.v9i4.10190

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 4, No. 02 April Tahun 2026
ISSN 2985-4172 (media online)
Hal 31-47

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Dalam Proceedings of the Annual *Symposium on Computer Application in Medical Care* (hal. 261-265).

Tigga, N. P., & Garg, S. (2020). ScienceDirect ScienceDirect Prediction of Type 2 Diabetes using Machine Learning Prediction of Type 2 Diabetes using Machine Learning Classification Methods Classification Methods. *PROCS*, *167*(2019), 706-716. https://doi.org/10.1016/j.procs.2020.03.336

UCI Machine Learning Repository. (1988). *Pima Indians Diabetes Database*. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). https://archive.ics.uci.edu/ml/datasets/diabetes

World Health Organization (WHO). (2023*). Diabetes - Fact Sheet.* World Health Organization. https://www.who.int/news-room/fact sheets/detail/diabetes