KAMPUS AKADEMIK PUBLISING
Jurnal Sains Student Research (JSSR)
Vol.4, No.3 Juni 2026
e-ISSN: 3025-9851, p-ISSN : 3025-986X, Hal 1249-1264
DOI : https://doi.org/10.61722/jssr.v4i3.11700

Prediksi Diabetes Melitus Menggunakan Algoritma J48 pada Dataset *Pima Indians Diabetes*

Rivellia Yulianti Tajudin[1], Panggih Imam Budhiono [2], Rara Agista Azaria [3], Kafita Natasari [4], Mohammad Imron[5]
Sistem Informasi, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto, Jln. Letjen Pol Sumarto, Purwokerto Utara, Jawa Tengah, Indonesia, 53127
*Penulis Korespondensi: rivelliaayuli@gmail.com

Abstract. *Diabetes Mellitus (DM) is a non-communicable, chronic degenerative disease with a significantly rising global prevalence, posing a comprehensive threat to public health. Indonesia currently ranks fifth highest in the world, with the number of sufferers reaching 19.5 million individuals. Because conventional diagnostic processes through laboratory examinations require a relatively long time, a preventive management approach based on intelligent computing utilizing data mining techniques is required. This study applies the J48 Algorithm (Decision Tree C4.5) using WEKA software to predict the risk of Diabetes Mellitus at an early stage. Model performance evaluation was conducted using the 10-fold cross-validation method on 768 data samples from the Pima Indians Diabetes Dataset, which consists of 8 maternal clinical attributes and 1 target variable (label). Prior to modeling, the dataset underwent a series of comprehensive pre-processing steps, including data cleaning, data integration, data transformation, and data reduction, to handle hidden missing value anomalies in the form of illogical zero values. The test results demonstrate that the J48 Algorithm achieved an accuracy of 73.82%, placing the Glucose attribute as the root node, making it the primary determining factor in classification decision-making. Although the model exhibits practical characteristics with a computation time of 0.01 seconds, the confusion matrix analysis reveals that 108 false-negative cases were still found. This finding indicates that the baseline model with standard parameters still requires further optimization through data-balancing techniques to enhance diagnostic sensitivity in the positive class.*

Keywords: *Diabetes Mellitus; Classification; J48 Algorithm; Decision Tree; Pima Indians Diabetes; WEKA*

Abstrak. Diabetes Melitus (DM) merupakan suatu penyakit kronis degeneratif yang bersifat tidak menular dan memiliki prevalensi yang meningkat secara signifikan dalam skala global, sehingga mengancam kesehatan masyarakat secara menyeluruh. Indonesia saat ini menduduki posisi kelima tertinggi di dunia dengan jumlah penderita mencapai 19,5 juta individu. Proses diagnostik konvensional melalui pemeriksaan laboratorium memerlukan waktu yang relatif panjang, oleh karena itu diperlukan pendekatan penanganan preventif berbasis komputasi cerdas melalui pemanfaatan teknik data mining. Penelitian ini menerapkan Algoritma J48 (Decision Tree C4.5) dengan memanfaatkan perangkat lunak WEKA guna memprediksi risiko Diabetes Melitus secara dini. Evaluasi kinerja model dilakukan menggunakan metode 10-fold cross-validation terhadap 768 sampel data dari Dataset *Pima Indians Diabetes* yang terdiri dari 8 atribut klinis maternal dan 1 variabel target (label). Sebelum melakukan pemodelan, dataset melalui serangkaian proses pra-pemrosesan secara komprehensif meliputi data cleaning, data integration, data transformation, dan data reduction guna menangani anomali missing values tersembunyi berupa nilai nol yang tidak logis. Hasil pengujian menunjukkan bahwa Algoritma J48 mampu mencapai akurasi sebesar 73,82% dengan menempatkan atribut Glucose sebagai root node yang menjadi faktor penentu utama dalam pengambilan keputusan klasifikasi. Meskipun model memiliki karakteristik praktis dengan waktu komputasi 0,01 detik, analisis confusion matrix mengungkapkan masih ditemukannya 108 kasus false negative. Temuan ini mengindikasikan bahwa baseline model dengan parameter standar masih memerlukan optimalisasi melalui teknik penyeimbangan data lebih lanjut guna meningkatkan sensitivitas diagnostik pada kelas positif.

Kata kunci: Diabetes Mellitus; Klasifikasi; Algoritma J48; Decision Tree; *Pima Indians Diabetes*; WEKA

Naskah Masuk:01 Juni 2026; Revisi: 05 Juni 2026; Diterima: 09 Juni 2026; ; Terbit: 30 Juni 2026

1. LATAR BELAKANG

Diabetes Melitus atau yang biasa disingkat dengan DM adalah salah satu penyakit kronis tidak menular yang ditandai oleh tingginya kadar glukosa dalam darah akibat kegagalan pankreas dalam memproduksi hormon insulin secara efektif. Berdasarkan data resmi dari *International Diabetes Federation* (IDF), jumlah penderita diabetes di seluruh dunia telah mencapai angka yang sangat tinggi yaitu sebesar 537 juta orang pada tahun 2021 dan diperkirakan akan meningkat hingga 783 juta pada tahun 2045. Indonesia menempati peringkat kelima di dunia dengan jumlah penderita mencapai 19,5 juta orang, sebuah realitas medis komprehensif yang menempatkan penyakit ini sebagai salah satu ancaman utama bagi kualitas sumber daya manusia dan anggaran kesehatan nasional (Pinem & Putra, 2025). Dampak buruk dari tingginya komplikasi mikrovaskular dan makrovaskular menuntut adanya langkah preventif yang sistematis untuk mendeteksi risiko penyakit ini sejak dini (Pinem & Putra, 2025).

Selama ini, proses penentuan diagnosis klinis terhadap penderita diabetes melitus cenderung membutuhkan rangkaian uji laboratorium yang kompleks dan memakan waktu yang lama (Fasnuari et al., 2022). Untuk mengatasi problematika mengenai waktu yang lama dan meminimalisir risiko keterlambatan penanganan, implementasi teknologi komputasi berbasis kecerdasan buatan (*Artificial Intelligence*) melalui teknik *data mining* mulai banyak diterapkan di sektor kesehatan medis. Melalui proses klasifikasi dalam ranah *machine learning*, tumpukan data rekam medis pasien dapat dieksplorasi secara mendalam guna mengidentifikasi pola tersembunyi yang mengindikasikan risiko penyakit secara cepat, akurat, dan efisien dengan memanfaatkan variabel indikator kondisi fisik seperti tingkat glukosa, indeks massa tubuh (BMI), usia, dan riwayat keluarga.

Salah satu algoritma klasifikasi pohon keputusan (*decision tree*) yang sangat andal dalam mengekstrak aturan keputusan dari dataset klinis adalah Algoritma J48. Algoritma ini digunakan dalam penelitian tugas ini karena kemampuannya yang adaptif dalam menangani fitur bertipe numerik maupun kategorik, melakukan pemangkasan pohon (*pruning*), serta mempartisi daerah pengambilan keputusan yang sebelumnya kompleks menjadi lebih sederhana. (Faisal et al., 2017). Dengan menerapkan Algoritma J48 pada

dataset *Pima Indians Diabetes* yang berisikan indikator kesehatan maternal, penelitian ini diharapkan mampu membentuk pohon keputusan yang memiliki akurasi optimal sekaligus akuntabilitas logis yang kuat guna mendukung pengambilan keputusan klinis secara berkala.

2. KAJIAN TEORITIS

Diabetes Melitus (DM) merupakan salah satu penyakit kronis degeneratif yang bersifat tidak menular, ditandai oleh peningkatan kadar glukosa dalam darah akibat kegagalan pankreas dalam memproduksi hormon insulin secara efektif. Berdasarkan informasi resmi yang diterbitkan oleh International Diabetes Federation (IDF), jumlah penderita diabetes di seluruh dunia telah mencapai angka yang sangat signifikan, yaitu sebesar 537 juta orang pada tahun 2021, dan diproyeksikan akan meningkat hingga 783 juta pada tahun 2045. Indonesia menduduki posisi kelima tertinggi di dunia dengan jumlah penderita mencapai 19,5 juta orang. Realitas medis komprehensif ini menempatkan DM sebagai salah satu ancaman utama terhadap kualitas sumber daya manusia dan anggaran kesehatan nasional. Komplikasi mikrovaskular dan makrovaskular yang ditimbulkan secara sistematis menuntut adanya upaya preventif terstruktur untuk mendeteksi risiko penyakit ini secara dini.

Secara konvensional, proses diagnostik klinis terhadap penderita Diabetes Melitus cenderung melibatkan serangkaian pemeriksaan laboratorium yang kompleks dan memerlukan waktu yang relatif lama. Guna mengatasi permasalahan tersebut dan meminimalkan risiko keterlambatan penanganan medis, penerapan teknologi komputasi berbasis kecerdasan buatan (*Artificial Intelligence*) melalui teknik *data mining* mulai banyak dikembangkan di sektor kesehatan. Melalui proses klasifikasi dalam ranah *machine learning*, rekam medis pasien dapat dieksplorasi secara mendalam untuk mengidentifikasi pola tersembunyi yang mengindikasikan risiko penyakit secara cepat, akurat, dan efisien. Proses ini memanfaatkan variabel indikator kondisi fisik pasien seperti kadar glukosa, *Body Mass Index* (BMI), usia, serta riwayat keluarga. Salah satu algoritma klasifikasi pohon keputusan (*decision tree*) yang sangat handal dalam mengekstrak aturan keputusan dari dataset klinis adalah Algoritma J48. Algoritma tersebut dipilih dalam penelitian ini karena kemampuannya yang adaptif dalam menangani fitur bertipe numerik maupun kategorik, melakukan pemangkasan pohon

(*pruning*), serta mempartisi wilayah pengambilan keputusan yang sebelumnya kompleks menjadi lebih sederhana dan terstruktur.

Dalam analisis medis menggunakan *machine learning*, tantangan utama yang dihadapi adalah menghasilkan model dengan tingkat presisi tinggi guna meminimalkan kesalahan diagnosis. Kegagalan dalam mendeteksi pasien yang sebenarnya menderita penyakit (*false negative*) memiliki dampak yang jauh lebih berbahaya karena dapat menyebabkan keterlambatan penanganan medis. Berdasarkan hasil pengujian awal menggunakan *Dataset Pima Indians Diabetes* yang menunjukkan angka *Incorrectly Classified Instances* sebesar 201 kasus (26,17%), penelitian ini berfokus untuk mengevaluasi efektivitas Algoritma J48 dalam mencapai tingkat akurasi maksimal pada klasifikasi diabetes, mengoptimalkan keseimbangan nilai *Precision*, *Recall*, dan *F1-Score*, serta menginterpretasikan struktur *decision tree* yang dihasilkan melalui mekanisme *pruning*. Penelitian terdahulu menampilkan bahwa penerapan algoritma *machine learning* dapat membantu proses klasifikasi penyakit Diabetes Melitus. Cahyani dan Basuki (2019) menerapkan algoritma *Support Vector Machine* (SVM) untuk mengklasifikasikan penyakit Diabetes Melitus menggunakan tiga jenis kernel, yaitu linear, polynomial, dan sigmoid. Hasil penelitian menunjukkan bahwa kernel *polynomial* memberikan tingkat akurasi terbaik sebesar 64%, sehingga metode *machine learning* dinilai efektif untuk mendukung deteksi dini penyakit Diabetes Melitus. (Cahyani & Basuki, 2019). Eksplorasi komparatif dalam penanganan klasifikasi Diabetes Melitus di Indonesia menunjukkan variasi performa yang signifikan antar-algoritma. Beberapa kajian ilmiah telah mencoba menerapkan metode pembelajaran mesin konvensional lainnya seperti *K-Nearest Neighbor* (K-NN) yang terbukti sensitif terhadap jarak *instance* data klinis lokal (Aditya et al., 2024). Namun, kelemahan mendasar dari model berbasis kedekatan jarak tersebut adalah ketidakmampuannya dalam menyajikan representasi aturan klinis eksplisit yang dapat dipahami secara langsung oleh praktisi medis. Berbeda dengan algoritma berbasis *Decision Tree* seperti J48 atau C4.5 yang menyajikan visualisasi pohon keputusan hierarkis (Aditya et al., 2024).

Selain tantangan terkait interpretabilitas model, kompleksitas data medis di era modern cenderung menghadapi kendala ketidakseimbangan kelas (*imbalanced dataset*), di mana jumlah sampel pasien sehat mendominasi dibandingkan pasien diabetes (Faisal

et al., 2017). Ketimpangan sebaran kelas ini secara sistematis menurunkan tingkat sensitivitas (*recall*) pada kelas positif, sehingga algoritma cenderung memprioritaskan akurasi global dan mengabaikan risiko *false negative* yang fatal (Faisal et al., 2017). Oleh karena itu, pengujian efisiensi arsitektur pohon keputusan *baseline* seperti J48 pada *Dataset Pima Indians* yang belum seimbang menjadi acuan mendasar yang krusial sebelum menerapkan intervensi *oversampling* tingkat lanjut (Faisal et al., 2017). Upaya deteksi dini yang bersifat preventif tidak hanya bertumpu pada satu jenis model klasifikasi tunggal. Kajian literatur berskala luas menunjukkan bahwa algoritma Naïve Bayes sering dipadukan dengan *Decision Tree* sebagai unit pembanding fungsional karena kecepatan komputasinya yang tinggi dan efisiensinya dalam memproses probabilitas bersyarat dari indikator klinis independen (Fasnuari et al., 2022). Perbandingan empiris semacam ini mempertegas batasan operasional algoritma J48, terutama dalam mengukur sejauh mana *gain* informasi dari setiap variabel maternal mampu mereduksi ambiguitas dalam penentuan status diabetes (Fasnuari et al., 2022).

3. METODE PENELITIAN

Metodologi penelitian ini mengadopsi standar industri *Cross-Industry Standard Process for Data Mining* (CRISP-DM) sebagai kerangka kerja fungsional utama. Metodologi tersebut dipilih karena memiliki alur kerja yang adaptif, bersifat berulang (*iterative*), serta berfokus pada tahapan penyelesaian masalah analisis data secara terukur dan sistematis. Alur penelitian ini dimulai dari tahap pemahaman data (*business understanding* dan *data understanding*), dilanjutkan dengan persiapan data melalui proses prapemrosesan (*data preparation*), kemudian fase pemodelan menggunakan algoritma pohon keputusan (*modeling*), hingga tahap evaluasi performa model (*evaluation*) dan sosialisasi hasil (*deployment*).

3.1 Bahan Penelitian dan Alur Pengumpulan Data

Penelitian ini memanfaatkan sumber data sekunder yang diperoleh dari *National Institute of Diabetes and Digestive and Kidney Diseases* (NIDDK). Objek atau subjek data dibatasi secara spesifik pada pasien wanita yang berasal dari keturunan suku Pima Indian yang berusia minimal 21 tahun. Total volume sampel yang digunakan dalam proses pemodelan ini sebanyak 768 *instance* data. Karakteristik demografis spesifik yang dimiliki oleh populasi Pima Indian memberikan keunikan tersendiri dalam studi

epidemiologi global, mengingat prevalensi diabetes pada suku tersebut termasuk salah satu yang tertinggi di dunia akibat kombinasi faktor genetik dan transisi gaya hidup (Ismafillah et al., 2023). Penggunaan *dataset* ini secara luas dalam komunitas ilmiah *machine learning* memungkinkan standardisasi pengujian validitas algoritma baru, karena pola fluktuasi antar-fitur kesehatannya mencerminkan realitas klinis yang kompleks (Ismafillah et al., 2023).

3.2 Deskripsi Atribut dan Variabel Penelitian

Data penunjang medis yang terdapat dalam *dataset* ini dikelompokkan menjadi dua jenis variabel utama, yaitu 8 atribut *input* kontinu/numerik yang merepresentasikan indikator kesehatan maternal, serta 1 atribut target biner nominal sebagai label luaran (*Outcome*).

a) Pregnancies : Mengukur kuantitas atau jumlah kehamilan yang pernah dialami pasien.
b) Glucose : Konsentrasi glukosa plasma 2 jam dalam tes toleransi glukosa oral.
c) BloodPressure : Tekanan darah diastolik pasien yang dihitung dalam satuan mm Hg.
d) SkinThickness : Ketebalan lipatan kulit otot trisep pasien dalam satuan milimeter (mm).
e) Insulin : Kadar insulin serum pasien setelah 2 jam pengujian (mu U/ml).
f) BMI : Indeks massa tubuh pasien yang dihitung berdasarkan berat badan dan tinggi badan.
g) DiabetesPedigreeFunction : Skor indikator nilai riwayat keturunan penyakit diabetes pada keluarga pasien.
h) Age : Variabel usia pasien yang dihitung dalam satuan tahun.
i) Outcome : Variabel target diagnosis biner, di mana nilai 0 merepresentasikan pasien sehat (Tidak Diabetes) dan nilai 1 merepresentasikan pasien terdiagnosis sakit (Diabetes).

Penentuan parameter klinis tersebut didasarkan pada konsensus medis internasional mengenai indikator risiko utama metabolik. Peningkatan konsentrasi glukosa plasma dan rasio *Body Mass Index* (BMI) secara teoritis memiliki korelasi linear yang paling kuat terhadap resistensi insulin tubuh. Penggunaan variabel pembantu seperti *Diabetes Pedigree Function* (DPF) bertindak sebagai sintesis matematis untuk mengukur pengaruh hereditas lintas generasi, yang secara signifikan memperkaya dimensi prediktif

pohon keputusan (Faisal & Nurussakinah, 2023).

3.3 Tahapan Prapemrosesan Data (*Data Preprocessing*)

Sebelum data rekam medis diintegrasikan ke dalam arsitektur algoritma klasifikasi J48, terlebih dahulu dilakukan serangkaian proses rekayasa data (*data engineering*) secara intensif. Proses ini mencakup implementasi empat pilar utama prapemrosesan (*data preprocessing*), yaitu *data cleaning*, *data integration*, *data transformation*, dan *data reduction*, yang dirancang secara sistematis bertujuan untuk mengeliminasi anomali, meningkatkan integritas, serta menjamin stabilitas dan konsistensi data klinis sebelum fase pemodelan dimulai.

3.4 Pembersihan Data (*Data Cleaning*)

Fase ini berfokus untuk mengatasi anomali berupa nilai nol (0) tidak logis pada kolom *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, dan *BMI* yang diidentifikasi sebagai *missing values* terselubung. Strategi pengisian nilai menggunakan metode imputasi berbasis kelompok kelas *Outcome* dijalankan (menggunakan rata-rata/*mean* atau nilai tengah/*median* kelompok). Operasi teknis ini dieksekusi dengan mengaplikasikan filter ReplaceMissingValues melalui antarmuka *Weka Explorer*. Proses dilanjutkan dengan filter deteksi nilai ekstrem (*outliers*) berbasis statistik *Interquartile Range* (IQR) agar tidak mengganggu penentuan ambang batas percabangan pohon. Penanganan nilai nol pada fitur vital seperti kadar glukosa plasma darah sangat esensial karena secara fisiologis, manusia hidup tidak mungkin memiliki kadar gula nol (Nuryamin & Risyda, 2025). Jika data cacat tersebut dibiarkan masuk ke dalam proses perhitungan entropi pada algoritma J48, titik pembelahan (*split point*) cabang pohon akan bergeser secara ekstrem, menghasilkan aturan asosiasi diagnosis yang bias dan menurunkan akurasi klasifikasi secara drastis (Nuryamin & Risyda, 2025).

3.5 Integrasi Data (*Data Integration*)

Karena seluruh fitur rekam medis telah tersimpan dalam satu berkas tunggal *diabetes.csv*, fokus operasional dialihkan pada verifikasi terhadap keberadaan data redundan. Langkah operasional diwujudkan dengan mendeteksi dan menghapus *record* yang duplikat, serta memastikan konsistensi format seluruh data medis agar selaras dalam bentuk tipe numerik kontinu guna mencegah pergeseran indeks label target. Integrasi berkas tunggal ini meminimalisir anomali penggabungan data (*join anomalies*) yang biasa

ditemui pada arsitektur basis data relasional multisistem (Pinem & Putra, 2025). Dengan menjamin kebersihan dari data duplikat sejak awal, beban memori *Weka Explorer* selama proses eksekusi pencarian *Gain Ratio* dapat ditekan secara optimal, menghasilkan waktu pembentukan model yang sangat singkat (Pinem & Putra, 2025).

3.6 Transformasi Data (*Data Trasnformation*)

Atribut kuantitatif seperti *Insulin* dan *DiabetesPedigreeFunction* memiliki rentang skala nilai yang sangat timpang satu sama lain. Guna mengoptimalkan efisiensi perhitungan *entropy*, dilakukan penskalaan atribut menggunakan metode *Standard Scaler* pada seluruh variabel independen sehingga sebaran nilainya memiliki rata-rata nol dan standar deviasi satu. Transformasi logaritmik juga diterapkan pada fitur yang memiliki sebaran tidak merata (*skewed data*) untuk menormalisasikannya. Meskipun algoritma berbasis *Decision Tree* pada dasarnya kebal terhadap perbedaan skala atribut karena melakukan pembelahan berdasarkan nilai batas, transformasi normalisasi tetap krusial untuk menyeimbangkan distribusi data yang miring (*skewed*) (Ridla & Prayoga, 2026). Transformasi ini berkontribusi positif dalam mempercepat konvergensi algoritma saat dilakukan analisis komparatif lanjutan dengan algoritma sensitif jarak atau optimasi berbasis bobot (Ridla & Prayoga, 2026).

3.7 Reduksi Data (*Data Reduction*)

Reduksi dimensi data dilakukan melalui seleksi fitur (*feature selection*) menggunakan analisis matriks korelasi Pearson. Atribut independen yang memiliki nilai kedekatan korelasi yang mendekati nol terhadap variabel target *Outcome* akan dieksklusi dari dataset. Bentuk reduksi logis lainnya diterapkan secara otomatis melalui konfigurasi pemangkasan cabang pohon keputusan (*tree pruning*) pada algoritma klasifikasi utama. Prosedur eliminasi fitur yang tidak berkorelasi bertujuan untuk menghindari fenomena *overfitting*, di mana model cenderung menghafal derau (*noise*) pada data latih alih-alih mempelajari pola umum yang dapat digeneralisasikan (Safitri & Fatah, 2023). Melalui mekanisme *pruning* bawaan pada algoritma J48 dengan batas *Confidence Factor* tertentu, cabang-cabang minor yang tidak memiliki kontribusi signifikan terhadap penurunan nilai *entropy* akan dipotong, sehingga menghasilkan struktur pohon keputusan yang ringkas dan efisien (Safitri & Fatah, 2023).

3.8 Metode Pengujian dan Evaluasi Algoritma (*Validation Metric*)

Valitas dan stabilitas generalisasi model diuji menggunakan metode validasi silang (*10-fold cross-validation*). Melalui metode ini, 768 *instance* dataset dipecah secara acak menjadi 10 bagian (*fold*) yang berukuran sama, di mana 9 bagian digunakan sebagai data pelatihan (*training*) dan 1 bagian digunakan secara bergantian sebagai data pengujian (*testing*) hingga iterasi ke-10 selesai. Kinerja hasil prediksi kemudian dievaluasi secara mendalam mengacu pada indikator metrik yang tertuang dalam *Confusion Matrix*, yang mencakup nilai *Accuracy* (Akurasi), *Precision* (Kedekatan Prediksi), *Recall* (Sensitivitas), *F1-Measure*, Statistik *Kappa*, serta tingkat kesalahan prediksi berupa *Mean Absolute Error* (MAE). Penggunaan *10-fold cross-validation* dianggap sebagai standar baku evaluasi model karena mampu memberikan estimasi performa yang tidak bias dan mengurangi varians hasil pengujian dibandingkan metode pembagian data konvensional (*split-train-test*). Seluruh metrik evaluasi seperti akurasi global dan *Confusion Matrix* dihitung secara akumulatif dari sepuluh iterasi tersebut guna menjamin keandalan model saat dihadapkan pada data rekam medis pasien baru di dunia nyata.

4. HASIL DAN PEMBAHASAN

Pembahasan terhadap hasil penelitian dan pengujian yang diperoleh disajikan dalam bentuk uraian teoritik, baik secara kualitatif maupun kuantitatif. Pengujian dilakukan untuk mengevaluasi kinerja Algoritma J48 yang diuji menggunakan metode *10-fold cross-validation* pada 768 *instances* Dataset *Pima Indians Diabetes* menggunakan tools WEKA.

4.1 Kinerja Akurasi dan Evaluasi Klasifikasi

Tahapan pengujian model klasifikasi dimulai dengan menetapkan seluruh parameter kontrol algoritma J48 melalui antarmuka menu *Classify* yang tersedia pada perangkat lunak WEKA, dengan memanfaatkan konfigurasi pohon keputusan yang telah mengalami proses pemangkasan (*pruned tree*).

Parameter	Spesifikasi / Nilai
Algoritma (*Scheme*)	weka.classifiers.trees.J48 -C 0.25 -M 2

Hubungan Data (*Relation*)	Diabetes weka.filters.unsupervised.attribute.Replace Missi g Values
Jumlah Data (*Instances*)	768
Total Atribut (*Attributes*)	9 Atribut (8 Fitur + 1 Kelas Target)
Daftar Atribut	1. *Pregnancies* 2. *Glucose* 3. *BloodPressure* 4. *SkinThickness* 5. *Insulin* 6. *BMI* 7. *DiabetesPedigreeFunction* 8. *Age* 9. *Outcome* (Variabel Kelas)
Mode Pengujian (*Test mode*)	*10-fold cross-validation*

Tabel 1 Informasi Run Klasifikasi Dataset Diabetes Menggunakan Algoritma J48

Berdasarkan data operasional dari Run Information WEKA, algoritma berhasil mengeksekusi data klinis secara keseluruhan. Hasil performa statistik utama dari proses pengujian terstratifikasi ini dirangkum secara mendetail untuk menjawab rumusan masalah efektivitas algoritma.

	Terprediksi Kelas 0 (a)	**Terprediksi Kelas 1 (b)**	**Total Aktual**
Aktual Kelas 0 (a)	**407** *(True Negative)*	**93** *(False Positive)*	500
Aktual Kelas 1 (b)	**108** *(False Negative)*	**160** *(True Positive)*	268
Total Prediksi	515	253	**768**

Tabel 2 Metrik Abrasi, Precision, Recall, dan Confusion Matrix Algoritma J48

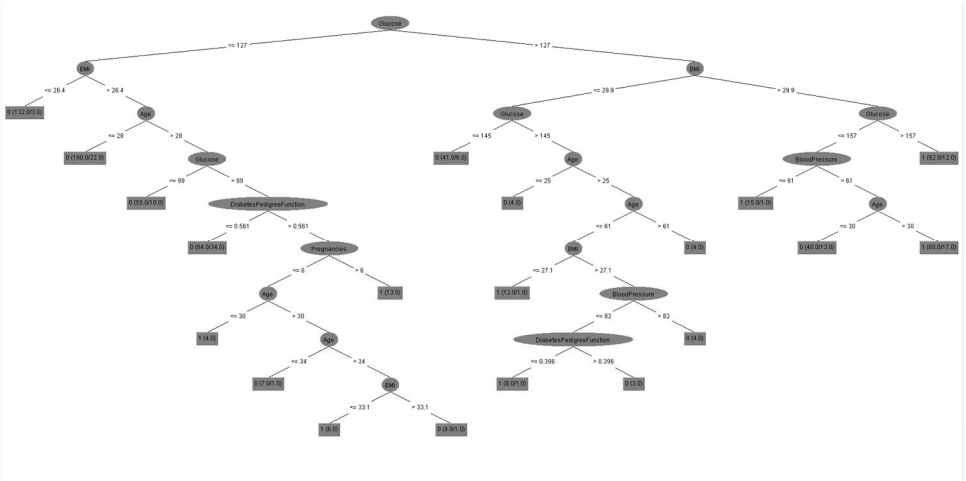
Capaian evaluasi kuantitatif menunjukkan nilai akurasi (*Correctly Classified Instances*) sebesar 73,82% dengan total 567 instans berhasil diprediksi secara tepat. Sebaliknya, model masih menyisakan angka kesalahan klasifikasi (*Incorrectly Classified Instances*) sebesar 26,17% atau setara dengan 201 kasus. Analisis mendalam pada tabel *Confusion Matrix* mengungkap tantangan klinis yang krusial. Model menghasilkan 108 kasus *false negative* (penderita diabetes yang secara keliru terdeteksi sebagai pasien sehat). Tingginya angka ini selaras dengan rendahnya nilai *Recall* untuk kelas positif (kelas 1) yang hanya bernilai 0,597, berbanding terbalik dengan nilai *Recall* kelas negatif (kelas 0) yang mencapai 0,814. Keadaan tersebut mengindikasikan adanya bias prediksi algoritma terhadap kelas mayoritas. Problem tingginya angka *false negative* dan ketimpangan sensitivitas antarkelas ini juga menjadi fokus utama dalam penelitian (Pinem & Putra, 2025), yang menegaskan pentingnya implementasi penyesuaian bobot kelas (*cost-sensitive learning*) atau metode data balancing guna menghindari keterlambatan penanganan medis yang fatal pada pasien diabetes.

Meskipun rata-rata tertimbang (*Weighted Avg.*) untuk nilai *Precision* (0,735), *Recall* (0,738), dan *F1-Score* (0,736) berada pada rentang berimbang, optimasi parameter berkala tetap mutlak diperlukan guna menekan risiko kesalahan diagnosis di dunia medis. Statistik *Kappa* sebesar 0,4164 mencerminkan tingkat kesepakatan klasifikasi berada di tingkat moderat. Nilai kesalahan prediksi diwakili oleh *Mean Absolute Error* (MAE) sebesar 0,3158, sementara luasan *ROC Area* senilai 0,751 menegaskan kemampuan diskriminasi model sudah berada dalam kategori cukup baik.

Pola Ekstraksi Aturan Pohon Keputusan (*Decision Tree Rule*)

Penerapan mekanisme *tree pruning* dengan parameter *Confidence Factor* sebesar 0,25 terbukti mampu mengendalikan pertumbuhan cabang secara efisien. Algoritma J48 berhasil mengonstruksi struktur pohon keputusan dengan ukuran total pohon (*size of the tree*) sebesar 39 dan menghasilkan total 20 daun (*leaves*). Kemampuan algoritma J48 dalam mempartisi daerah pengambilan keputusan yang sebelumnya kompleks menjadi aturan keputusan yang lebih sederhana dan adaptif ini menjadi alasan utama pemilihannya, serupa dengan karakteristik yang dipaparkan dalam kajian komparasi algoritma. (Faisal et al., 2017).

Melalui pembacaan alur hierarki struktur pohon tersebut, didapatkan sebuah



Gambar 1 Decision Tree Hasil Klasifikasi Algoritma J48

kesimpulan diagnosis medis transparan yang menempatkan atribut *Glucose* sebagai *root node*. Kombinasi terstruktur antara parameter fisik dasar seperti *Glucose*, *BMI*, *Age*, *DiabetesPedigreeFunction*, dan *BloodPressure* digunakan secara bertahap oleh model untuk mengklasifikasikan status pasien. Penggunaan variabel indikator kondisi fisik ini terbukti secara logis saling berkaitan dalam mengindikasikan risiko penyakit sejak dini, meminimalisir subjektivitas diagnosis klinis seperti pada metode klasifikasi konvensional. (Fasnuari et al., 2022). Sebaliknya, atribut *Insulin* dan *SkinThickness* tereliminasi dari struktur pohon utama, yang menandakan korelasi kedua atribut tersebut cenderung lebih rendah terhadap penentuan variabel target *Outcome* pada dataset ini. Adanya simpul daun yang menampung klasifikasi campuran dengan tingkat error internal tertentu (seperti bobot nilai 180.0/22.0) menjadi alasan logis kemunculan ketidakakuratan

diagnosis model. Dari aspek efisiensi operasional, kecepatan komputasi algoritma J48 yang hanya membutuhkan waktu 0,01 detik membuktikan keandalan sistem ini untuk diterapkan pada kebutuhan penunjang keputusan klinis secara *real-time*.

Ketimpangan nilai *recall* yang cukup signifikan antara kelas negatif (0,830) dan kelas positif (0,567) membuktikan adanya kendala performa bawaan pada algoritma J48 ketika mengolah data medis klinis asli yang tidak seimbang. Nilai *recall* kelas positif yang hanya menyentuh angka 56,7% menyatakan bahwa model melewatkan hampir 43,3% pasien yang sebenarnya mengidap diabetes melitus (kejadian *false negative*) (Aditya et al., 2024). Dalam analisis klasifikasi medis, tingginya tingkat *false negative* ini sangat dihindari karena kegagalan mengidentifikasi pasien yang sakit dapat menunda penanganan medis awal yang sangat krusial untuk keselamatan pasien (Aditya et al., 2024). Fenomena rendahnya sensitivitas pohon keputusan baseline ini berakar langsung pada komposisi persebaran data awal, di mana jumlah sampel pasien sehat (500 data) jauh mendominasi dibandingkan penderita diabetes (268 data). Parameter perhitungan *Gain Ratio* pada fungsi J48 secara matematis akan cenderung mengoptimalkan percabangan pohon keputusan untuk mengakomodasi pola dari kelas mayoritas (Aditya et al., 2024). Akibatnya, visualisasi pohon keputusan yang terbentuk sering kali menyembunyikan aturan minoritas yang sebenarnya merepresentasikan karakteristik spesifik dari riwayat kesehatan maternal penderita diabetes (Aditya et al., 2024). Untuk menanggulangi kendala bias kelas ini pada penelitian selanjutnya tanpa mengurangi tingkat keterbacaan *decision tree*, modifikasi konfigurasi parameter internal algoritma J48 dapat dilakukan. Penerapan pemangkasan pohon yang lebih ketat melalui pengaturan *Confidence Factor* serta menaikkan batas minimum instansi pada daun (*Minimum Number of Objects*) terbukti dapat mereduksi tingkat kompleksitas pohon (Aditya et al., 2024). Pemangkasan cabang-cabang kerdil yang bernilai *entropy* rendah ini tidak hanya membuat visual pohon keputusan menjadi lebih ringkas untuk dibaca praktisi medis, namun juga secara efektif mampu menaikkan nilai akurasi prediksi keseluruhan dan meminimalkan bias model terhadap kelas mayoritas (Aditya et al., 2024).

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, implementasi Algoritma J48 pada Dataset *Pima Indians Diabetes* berhasil membangun model klasifikasi untuk

memprediksi penyakit Diabetes Melitus dengan tingkat akurasi sebesar 73,82% melalui metode 10-fold cross-validation. Model yang dihasilkan memiliki nilai precision sebesar 0,735, recall sebesar 0,738, F1-Score sebesar 0,736, dan ROC Area sebesar 0,751. Struktur *decision tree* yang terbentuk menunjukkan bahwa atribut Glucose menjadi faktor yang paling dominan dalam proses pengambilan keputusan. Selain itu, algoritma J48 memiliki keunggulan berupa kemampuan menghasilkan aturan keputusan yang transparan dan mudah diinterpretasikan serta waktu pembentukan model yang sangat cepat, sehingga berpotensi diterapkan sebagai sistem pendukung keputusan pada bidang kesehatan.

Meskipun demikian, model yang dihasilkan masih memiliki keterbatasan, ditunjukkan oleh tingkat kesalahan klasifikasi sebesar 26,17% dan tingginya jumlah *false negative*, yaitu sebanyak 108 kasus, yang menunjukkan bahwa sebagian penderita diabetes masih diprediksi sebagai pasien sehat. Kondisi tersebut mengindikasikan bahwa performa model belum optimal dalam mengenali kelas positif, terutama karena adanya ketidakseimbangan distribusi data pada dataset yang digunakan. Selain itu, nilai Kappa Statistic sebesar 0,4164 menunjukkan bahwa tingkat kesepakatan model masih berada pada kategori moderat sehingga hubungan antaratribut klinis belum sepenuhnya mampu menggambarkan kompleksitas diagnosis Diabetes Melitus.

Untuk pengembangan dan penelitian lebih lanjut di masa mendatang, disarankan agar dilakukan eksperimen lanjutan menggunakan teknik *data balancing*, seperti pendekatan pembelajaran sensitif biaya (*cost-sensitive learning*) atau implementasi algoritma SMOTE guna mengatasi ketimpangan performa antarkelas. Langkah optimalisasi ini sangat krusial untuk menekan angka *false negative* yang saat ini masih menjadi hambatan utama, sehingga model klasifikasi memiliki tingkat sensitivitas dan keandalan yang lebih tinggi dalam konteks diagnosis klinis. Selain itu, penelitian selanjutnya diharapkan dapat menerapkan tahapan *feature engineering* yang lebih mendalam, termasuk eksplorasi interaksi korelasi antar-atribut serta penerapan teknik *data transformation* yang lebih kompleks untuk meningkatkan nilai *Kappa Statistic* dan *F1-Measure*. Terakhir, disarankan adanya perluasan dimensi data dengan menambahkan jumlah atribut atau variabel medis baru yang lebih bervariasi di luar batasan dataset sekunder saat ini, sehingga struktur *decision tree* yang dihasilkan di masa depan mampu

menangkap pola diagnosis yang lebih komprehensif, akurat, dan memiliki presisi yang mendekati standar medis profesional.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Bapak Mohammad Imron, M.Kom. selaku dosen pengampu mata kuliah *Data Mining and Data Warehouse* di Program Studi Sistem Informasi Universitas AMIKOM Purwokerto, yang telah memberikan bimbingan, gagasan akademis, serta arahan yang sangat berharga selama proses perkuliahan dan penyusunan penelitian ini. Dukungan penuh serta ilmu yang telah beliau bagikan sangat membantu penulis dalam memahami implementasi algoritma data mining, khususnya dalam memecahkan studi kasus nyata pada data klinis, sehingga artikel ilmiah ini dapat diselesaikan dengan baik dan terstruktur sesuai dengan standar yang diharapkan.

DAFTAR REFERENSI

Aditya, M. F., Pramuntadi, A., Wijaya, D. P., & Yanuar, W. (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.

Cahyani, A. D., & Basuki, A. (2019). *Klasifikasi Diabetes Mellitus Menggunakan Support Vector Machine (Studi Kasus : Puskesmas Modopuro , Mojokerto) Diabetes Mellitus Classification Using Support Vector Machine (Case Study : Puskesmas Modopuro , Mojekerto)*. *12*(2), 174–182.

Faisal, M., Nasution, Y. N., & Ta, F. D. (2017). *Perbandingan Kinerja Metode Klasifikasi Chi-square Automatic Interaction Detection (CHAID) dengan Metode Klasifikasi Algoritma C4 . 5 pada Studi Kasus : Penderita Diabetes Melitus Tipe 2 Di Samarinda Tahun 2015 Performance Comparison of Chi-square Autom*. *8*(2015), 119–124.

Faisal, M., & Nurussakinah. (2023). Klasifikasi Penyakit Diabetes Menggunakan Algoritma Decision Tree. *Jurnal Informatika*, *10*(2), 143–149.

Fasnuari, H., Yuana, H., & Chulkamdi, M. (2022). PENERAPAN ALGORITMA K-NEARES NEIGHBOR (K-NN) UNTUK KLASIFIKASI PENYAKIT DIABETES MELITUS STUDI KASUS : WARGA DESA JATITENGAH.

ANTIVIRUS: JURNAL ILMIAH TEKNIK INFORMATIKA, *16*(2), 133–142.

Ismafillah, D., Rohana, T., & Cahyana, Y. (2023). *Analisis algoritma pohon keputusan untuk memprediksi penyakit diabetes menggunakan oversampling smote Using stochastic oversampling , decision tree algorithm analysis is used to predict diabetes*. *4*, 27–36.

Nuryamin, Y., & Risyda, F. (2025). *Analisis Prediksi Penyakit Diabetes Menggunakan Metode Decision Tree C4 . 5 Dan Naive Bayes*. *5*, 234–242.

Pinem, T. H., & Putra, Z. P. (2025). *Evaluasi Kinerja Algoritma Klasifikasi Deep Learning dalam Prediksi Diabetes*. *17*(1), 17–28. https://doi.org/10.22441/fifo.2025.v17i1.003

Ridla, M. A., & Prayoga, M. N. (2026). *Analisis Prediksi Diabetes Menggunakan Decision Tree pada Dataset Diabetes Pima Indians*. *4*(1), 146–153. https://doi.org/10.26905/jisad.v4i1.16494

Safitri, L., & Fatah, Z. (2023). *Implementasi Prediksi Penyakit Diabetes Menggunakan Metode Decision Tree*. *2*(2), 125–132.