

Penerapan Metode Algoritma C.48 untuk Klasifikasi Penyakit Diabetes

Implementation of the C4.8 Algorithm Method for Diabetes Disease Classification

Donna Oktar Endras Wanto^{a,1*}, Cendra Harmon^{a,2}, Jupron^{a,3}

^{1,2,3}University of Pamulang, Jl. Raya Puspatek, South Tangerang 15310, Indonesia

¹dosen03269@unpam.ac.id*; ²dosen02677@unpam.ac.id; ³dosen02664@unpam.ac.id

*Corresponding author

Received:	Revised:	Accepted:	Available online:
16.09.2025	05.10.2025	19.10.2025	30.10.2025

Abstrak: Penyakit diabetes merupakan salah satu gangguan metabolik kronis yang prevalensinya terus meningkat di berbagai negara, termasuk Indonesia. Deteksi dini sangat penting untuk mencegah komplikasi jangka panjang, dan salah satu pendekatan yang efektif adalah dengan memanfaatkan algoritma klasifikasi berbasis machine learning. Penelitian ini dilakukan secara komputasional menggunakan data sekunder dari Kaggle, dengan tujuan untuk menerapkan algoritma C4.8 dalam mengklasifikasikan penyakit diabetes menggunakan aplikasi WEKA, serta mengevaluasi kinerjanya berdasarkan data medis yang tersedia. Dataset yang digunakan terdiri dari 768 instance dan 9 atribut, termasuk Glukosa, BMI, dan Usia. Proses preprocessing mencakup penanganan nilai hilang, normalisasi menggunakan metode Min-Max, serta penyeimbangan kelas dengan teknik SMOTE. Hasil pelatihan model menunjukkan tingkat akurasi sebesar 84,11%, sementara pada pengujian menggunakan metode 10-fold cross-validation, akurasi menurun menjadi 73,70%, dengan nilai Kappa sebesar 0,414. Hasil visualisasi mendukung temuan bahwa atribut Glukosa merupakan variabel paling dominan dalam membedakan kelas, diikuti oleh BMI dan Usia, sedangkan atribut seperti Tekanan Darah dan Ketebalan Kulit kurang memberikan kontribusi signifikan secara individual. Temuan ini menunjukkan bahwa algoritma C4.8 dapat digunakan sebagai dasar dalam sistem pendukung keputusan untuk diagnosis awal diabetes, meskipun masih terdapat ruang untuk peningkatan performa model melalui penerapan teknik lanjutan.

Kata kunci: diabetes, klasifikasi, algoritma C4.8, WEKA, machine learning, diagnosis dini

Abstract: Diabetes mellitus is a chronic metabolic disorder with a rising prevalence worldwide, including Indonesia. Early detection is crucial to prevent long-term complications, and one effective approach is the use of machine learning-based classification algorithms. This study was conducted computationally using secondary data from Kaggle, aiming to apply the C4.8 algorithm to classify diabetes disease using the WEKA application, and to evaluate its performance based on the available medical data. The dataset consists of 768 instances and 9 attributes, including Glucose, BMI, and Age. Preprocessing involved handling missing values, normalization using the Min-Max method, and class balancing with the SMOTE technique. The training results showed an accuracy of 84.11%, while testing with 10-fold cross-validation yielded a reduced accuracy of 73.70%, with a Kappa value of 0.414. Data visualization supported the finding that Glucose is the most dominant attribute for class differentiation, followed by BMI and Age, whereas attributes such as Blood Pressure and Skin Thickness contributed less significantly on an individual basis. These results indicate that the C4.8 algorithm can serve as a foundation for decision support systems in early diabetes diagnosis, although there is still room for model performance improvement through advanced techniques.

Keywords: diabetes, classification, C4.8 algorithm, WEKA, machine learning, early diagnosis

1. PENDAHULUAN

Diabetes mellitus merupakan penyakit metabolik kronis yang ditandai oleh tingginya kadar glukosa darah secara persisten akibat gangguan sekresi atau kerja insulin. Peningkatan prevalensi diabetes, baik secara global maupun nasional, menjadikan deteksi dini sebagai langkah penting dalam pencegahan komplikasi dan pengendalian penyakit. Dalam konteks ini, teknik klasifikasi berbasis data mining, khususnya algoritma pohon keputusan seperti C4.5 atau J48, telah banyak digunakan dalam penelitian untuk membantu diagnosis awal. Beberapa studi di Indonesia dalam lima tahun terakhir membuktikan efektivitas pendekatan ini. Misalnya, metode C4.5 mampu mencapai akurasi 91,35%, presisi 93,55%, recall 92,06%, dan F1-score 92,80% pada dataset diabetes dari Kaggle[1]. Penelitian lain menunjukkan bahwa algoritma C4.5 memiliki akurasi sebesar 84,62% dan bersaing dengan Naïve Bayes dalam memprediksi diabetes [2].

Namun, sebagian besar penelitian masih terbatas pada penerapan algoritma standar tanpa modifikasi atau optimasi yang mempertimbangkan karakteristik unik data medis, seperti nilai yang hilang (missing values), ketidakseimbangan kelas (class imbalance), dan noise. Di sisi lain, platform WEKA cukup populer dalam implementasi algoritma machine learning di bidang kesehatan, tetapi studi yang secara eksplisit membahas penerapan algoritma C4.8 dalam klasifikasi diabetes masih sangat terbatas. Bahkan, penelitian yang mengkaji klasifikasi diabetes dengan C4.5, tidak menyebutkan varian algoritmik lebih lanjut seperti C4.8.[3]

C4.5 sendiri merupakan algoritma decision tree yang mampu menangani atribut kontinu, nilai hilang, serta menerapkan teknik pruning untuk mencegah overfitting. Algoritma C4.8 dianggap sebagai penyempurnaan dari C4.5 dan dikenal luas melalui implementasi J48 di platform WEKA. Namun demikian, hingga saat ini tidak ditemukan dokumentasi resmi atau publikasi akademik yang secara eksplisit menjelaskan perbedaan algoritmik antara C4.5 dan C4.8. Informasi yang tersedia mengenai C4.8 cenderung berasal dari dokumentasi teknis WEKA atau komunitas pengembang, yang menyebutkan adanya optimasi pada pemilihan atribut, pengaturan threshold, dan teknik pruning lanjutan. Oleh karena itu, klaim-klaim mengenai “varian modifikasi (teoritis atau rekayasa)” dalam C4.8 masih terlalu spekulatif karena tidak didukung oleh referensi ilmiah formal. Hal ini perlu diperhatikan agar tidak menimbulkan kesalahpahaman dalam kajian akademik.

Penelitian ini bertujuan untuk mengevaluasi efektivitas algoritma C4.8 (dalam bentuk implementasi J48 di WEKA) dalam klasifikasi penyakit diabetes, dengan fokus pada penanganan atribut kontinu, missing values, dan ketidakseimbangan kelas. Dataset yang digunakan adalah *Pima Indian Diabetes Dataset* yang diperoleh secara resmi dari UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes>). Proses eksperimen dilakukan menggunakan platform WEKA versi 3.8.6, dengan konfigurasi C4.8 (J48) sebagai berikut: confidence factor = 0.25, minimum number of instances per leaf = 2, subtree raising diaktifkan, collapse tree diaktifkan, dan binary splits dimatikan. Evaluasi dilakukan dengan metode 10-fold cross-validation.

Tahapan preprocessing dalam penelitian ini mencakup: (1) penanganan missing values dengan mean imputation, terutama untuk atribut seperti glukosa, tekanan darah, dan insulin; (2) normalisasi data menggunakan metode Min-Max Scaling untuk menyeragamkan skala antar atribut; serta (3) balancing kelas menggunakan metode SMOTE (Synthetic Minority Over-sampling Technique) untuk mengatasi ketimpangan antara jumlah kelas positif dan negatif. Teknik SMOTE sendiri terbukti efektif dalam meningkatkan performa model, seperti yang juga diterapkan dalam penelitian yang menggabungkan SMOTE dengan Random Forest untuk mendeteksi diabetes.[4]

Alur metodologi yang diterapkan dalam penelitian ini terdiri atas: pengumpulan dataset, preprocessing data (penanganan nilai hilang, normalisasi, balancing), implementasi algoritma (C4.8 dan pembandingan C4.5), evaluasi kinerja model menggunakan metrik akurasi, presisi, recall, F1-score, dan AUC, serta analisis dan penarikan kesimpulan. Dengan pendekatan tersebut, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan model klasifikasi penyakit diabetes yang lebih akurat dan robust terhadap tantangan nyata di data medis, serta memperkaya referensi lokal terkait penerapan algoritma decision tree lanjutan di domain kesehatan.

2. TINJAUAN PUSTAKA

2.1. Kajian Empiris

Algoritma pohon keputusan merupakan salah satu metode klasifikasi yang banyak digunakan dalam bidang medis karena kemampuannya menghasilkan model yang dapat diinterpretasikan dengan baik oleh tenaga medis dan non-teknis. Salah satu algoritma pohon keputusan yang populer adalah C4.5, yang kemudian melahirkan berbagai varian, termasuk C4.8, yang menawarkan peningkatan dari segi akurasi, efisiensi, dan kemampuan menangani data kompleks. Platform *open-source* seperti WEKA turut mendukung pengembangan dan pengujian model ini karena menyediakan fitur visualisasi serta parameter tuning yang fleksibel. Di tengah meningkatnya kebutuhan akan diagnosis dini penyakit kronis seperti diabetes, pengembangan metode klasifikasi yang lebih akurat dan robust menjadi penting, terutama dalam menghadapi tantangan data medis seperti missing value, noise, dan ketidakseimbangan kelas.

Beberapa studi empiris menunjukkan efektivitas metode pohon keputusan dalam klasifikasi penyakit diabetes. Penelitian yang berjudul "*Implementasi Metode Decision Tree pada Prediksi Penyakit Diabetes Melitus Tipe 2*" menggunakan dataset lokal dari Puskesmas dan menghasilkan tingkat akurasi yang tinggi, menandakan bahwa model ini cukup andal dalam diagnosis awal penyakit diabetes [5]. Sementara itu, studi berjudul "*Analisis Prediksi Penyakit Diabetes Menggunakan Metode Decision Tree C4.5 dan Naive Bayes*" melakukan komparasi antara dua metode klasifikasi dan menunjukkan bahwa pemilihan algoritma yang sesuai dapat memengaruhi performa secara signifikan, terutama ketika diterapkan pada dataset publik dan lokal [6].

Penelitian lain yang membandingkan beberapa algoritma pohon keputusan menunjukkan pentingnya modifikasi algoritma untuk meningkatkan performa. Studi di *Medikom* menyajikan komparasi antara C4.5, Random Forest, dan CatBoost yang dilengkapi dengan teknik boosting. Hasilnya memperlihatkan bahwa penguatan model dapat meningkatkan akurasi dalam prediksi diabetes, sekaligus memperkuat argumen bahwa pohon keputusan dasar masih bisa ditingkatkan melalui pendekatan ensemble atau varian algoritmik yang lebih canggih [7]. Hal serupa juga diungkapkan dalam penelitian yang dipublikasikan di Open Journal, yang mengusulkan optimasi decision tree melalui teknik boosting untuk klasifikasi penyakit diabetes dan menghasilkan performa yang lebih stabil dan presisi tinggi [8].

Dari sisi implementasi teknis, platform WEKA banyak dimanfaatkan dalam penelitian klasifikasi medis karena kemudahan penggunaannya serta ketersediaan berbagai algoritma yang dapat langsung diuji. Studi yang dipublikasikan di Ejournal BSI menunjukkan bahwa WEKA secara luas digunakan dalam analisis klasifikasi penyakit umum, termasuk penggunaan algoritma C4.5, dan menjadi alat bantu efektif dalam penelitian di bidang kesehatan [9].

3. METODE

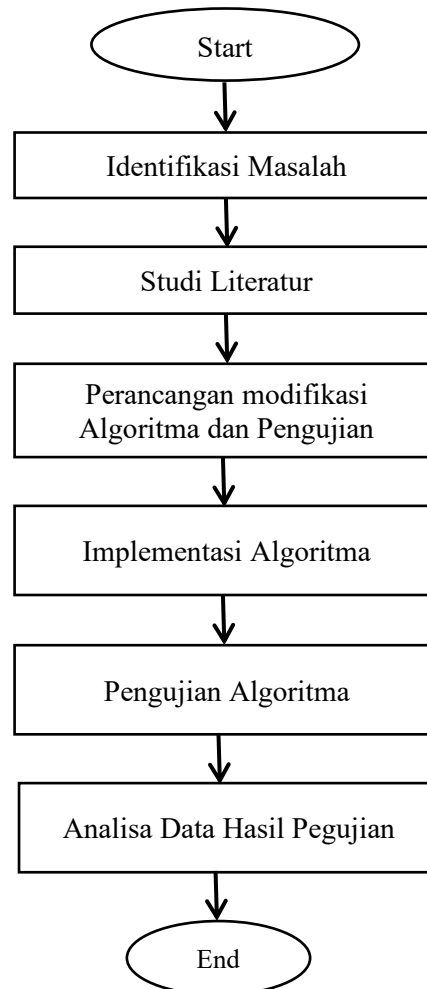
Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen yang bertujuan untuk menguji penerapan algoritma C4.8 dalam klasifikasi penyakit diabetes menggunakan perangkat lunak WEKA. Ruang lingkup penelitian mencakup analisis data pasien diabetes berdasarkan atribut klinis seperti kadar glukosa, tekanan darah, usia, indeks massa tubuh (BMI), dan variabel lain yang relevan, dengan fokus pada peningkatan akurasi klasifikasi penyakit diabetes. Objek penelitian adalah dataset diabetes yang diambil dari platform Kaggle, yang merupakan data sekunder yang telah tersedia secara daring dan dianggap representatif untuk kasus klasifikasi penyakit diabetes. Pendekatan serupa juga digunakan dalam penelitian yang menerapkan algoritma C4.5 pada data diabetes untuk mengidentifikasi faktor-faktor penyebabnya [10].

Definisi operasional variabel dalam penelitian ini terdiri atas variabel bebas, yaitu atribut-atribut klinis pasien seperti kadar glukosa, tekanan darah, usia, dan BMI, serta variabel terikat yaitu status klasifikasi diabetes (positif atau negatif). Populasi dalam penelitian ini mencakup seluruh data pasien yang tercantum dalam dataset, dan sampel data dipilih secara purposive. Pembagian data dilakukan dengan rasio 70:30 antara data latih dan data uji. Pemilihan rasio ini didasarkan pada pertimbangan umum dalam pembelajaran mesin, di mana rasio tersebut dinilai mampu menjaga keseimbangan antara kebutuhan pelatihan model dan validasi performanya. Proses pembagian dilakukan secara acak (randomized) untuk memastikan distribusi kelas yang seimbang dan mencegah bias.

Sebelum proses pelatihan model dilakukan, dataset menjalani tahapan pra-pemrosesan (preprocessing) guna memastikan kualitas dan kelayakan data. Langkah ini mencakup penanganan nilai hilang (missing values) yang ditangani dengan metode imputasi seperti pengisian berdasarkan nilai rata-rata atau median, atau penghapusan data jika nilai hilang terlalu banyak. Selain itu, data numerik dinormalisasi menggunakan metode seperti Min-Max Scaling atau Z-score Standardization untuk menyetarakan skala antar atribut, sehingga tidak ada atribut yang mendominasi proses pembelajaran. Jika terdapat ketidakseimbangan distribusi kelas (class imbalance), dilakukan penyeimbangan data menggunakan teknik seperti SMOTE (Synthetic Minority Over-sampling Technique), sebagaimana juga diterapkan dalam penelitian yang membahas penggunaan algoritma machine learning dalam klasifikasi penderita diabetes [11].

Algoritma C4.8, yang merupakan pengembangan dari algoritma C4.5, digunakan sebagai metode utama klasifikasi. Implementasi dilakukan melalui perangkat lunak WEKA versi 3.8.6. Proses pelatihan model dilakukan menggunakan data latih, dan evaluasi performa dilakukan pada

data uji [12]. Evaluasi kinerja model mencakup metrik seperti akurasi, precision, recall, F1-score, dan area under curve (AUC), untuk memperoleh gambaran komprehensif terhadap efektivitas model dalam mengklasifikasikan data [13]. Hasil yang diperoleh diharapkan dapat menunjukkan seberapa baik algoritma C4.8 dalam mengidentifikasi status diabetes pasien berdasarkan data klinis. Penelitian ini terdiri dari lima tahap utama, yaitu studi literatur, perancangan serta pengujian modifikasi algoritma, implementasi algoritma, pengujian algoritma, dan analisis data hasil pengujian. Keterkaitan antara setiap tahapan tersebut digambarkan secara visual melalui sebuah diagram alir yang ditampilkan pada gambar berikut.



Gambar 1. Diagram Alir Proses Penelitian

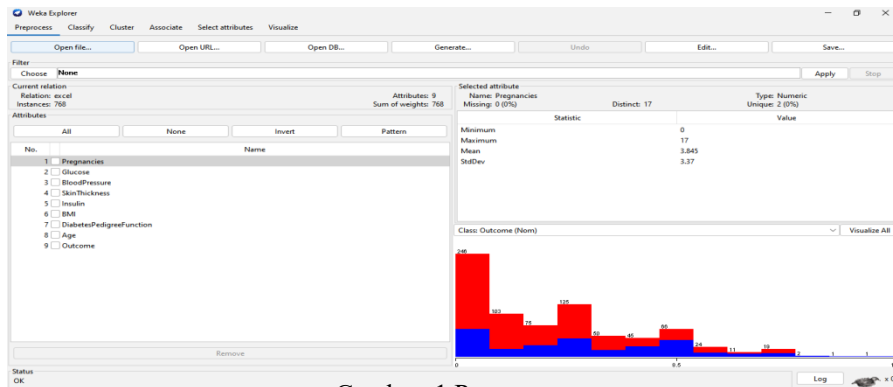
4. HASIL DAN PEMBAHASAN

A. Implementasi Tampilan Weka

1. Preprocess

Eksplorasi awal dataset dilakukan menggunakan perangkat lunak Weka Explorer. Dataset yang digunakan bernama *diabetes* dan terdiri atas 768 baris data (instances) dan 9 atribut (kolom), yaitu: *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, *BMI*, *DiabetesPedigreeFunction*, *Age*, dan *Outcome*. Atribut *Outcome* digunakan sebagai variabel target dalam klasifikasi, dengan nilai 0 menunjukkan pasien tidak menderita diabetes dan nilai 1 menunjukkan pasien menderita diabetes. Pada tahap awal, atribut *Pregnancies* dipilih untuk dianalisis secara deskriptif. Atribut ini memiliki tipe data numerik, dengan tidak terdapat data hilang (0%) dari total 768 data. Berdasarkan statistik deskriptif yang ditampilkan, *Pregnancies* memiliki nilai minimum sebesar 0 dan nilai maksimum sebesar 17, dengan rata-rata (mean) sebesar 3,845 dan standar deviasi sebesar 3,37. Total terdapat 17 nilai unik pada atribut ini, yang menunjukkan adanya variasi jumlah kehamilan pada masing-masing pasien wanita. Distribusi frekuensi nilai *Pregnancies* divisualisasikan melalui histogram berdasarkan kelas *Outcome*. Nilai 0

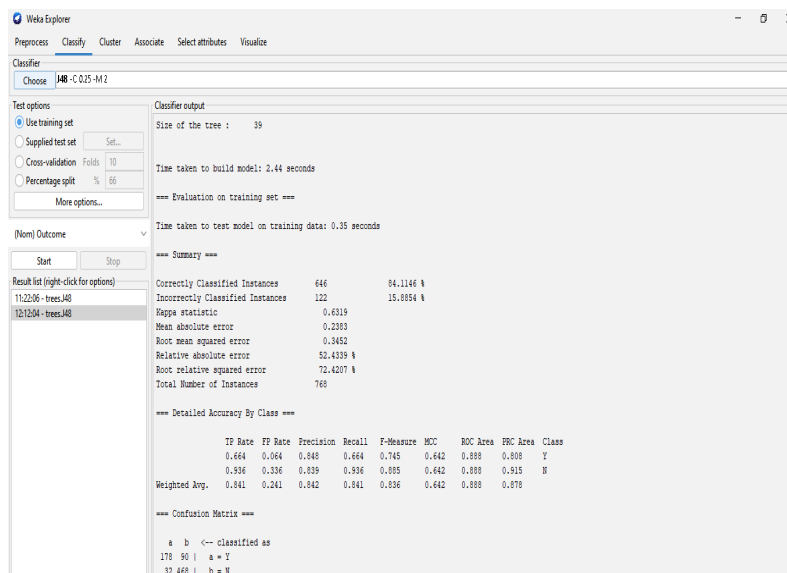
pada atribut *Pregnancies* memiliki jumlah tertinggi, yaitu sebanyak 246 data, disusul oleh nilai 1 sebanyak 135 data. Ini menunjukkan bahwa sebagian besar pasien dalam dataset tercatat tidak pernah mengalami kehamilan. Tidak ditemukan data duplikat atau nilai ekstrem yang tidak sesuai pada atribut ini. Seluruh atribut dalam dataset telah memiliki struktur data yang baik dan siap digunakan dalam proses pemodelan klasifikasi menggunakan algoritma C4.8. Kualitas data yang lengkap dan bersih mendukung validitas proses analisis selanjutnya, serta memastikan bahwa model machine learning dapat dilatih dengan optimal tanpa perlakuan khusus untuk penanganan data hilang.



Gambar. 1 Preprocess

2. Use training set

Model klasifikasi yang dilatih menggunakan algoritma C4.8 pada dataset diabetes berhasil memberikan performa yang cukup baik. Dari total 768 data latih yang digunakan, sebanyak 646 instance atau sebesar 84,11% berhasil diklasifikasikan dengan benar, sedangkan 122 instance atau 15,89% diklasifikasikan secara salah. Hal ini menunjukkan bahwa model memiliki kemampuan generalisasi yang cukup tinggi terhadap data pelatihan. Nilai Kappa Statistic sebesar 0,6319 mengindikasikan tingkat kesepakatan yang cukup kuat antara hasil klasifikasi model dengan label sebenarnya, dan jauh lebih baik dibandingkan dengan klasifikasi acak.



Gambar 2. Data Training

Berdasarkan evaluasi terhadap error, model menghasilkan nilai Mean Absolute Error (MAE) sebesar 0,2383 dan Root Mean Squared Error (RMSE) sebesar 0,3452. Sementara itu, nilai Relative Absolute Error mencapai 52,43%, dan Root Relative Squared Error sebesar 72,42%. Nilai-nilai ini menunjukkan bahwa meskipun akurasi tinggi, masih terdapat penyimpangan atau

kesalahan prediksi yang relatif terhadap model baseline, terutama pada beberapa atribut input yang mungkin tidak memiliki pengaruh dominan terhadap kelas output.

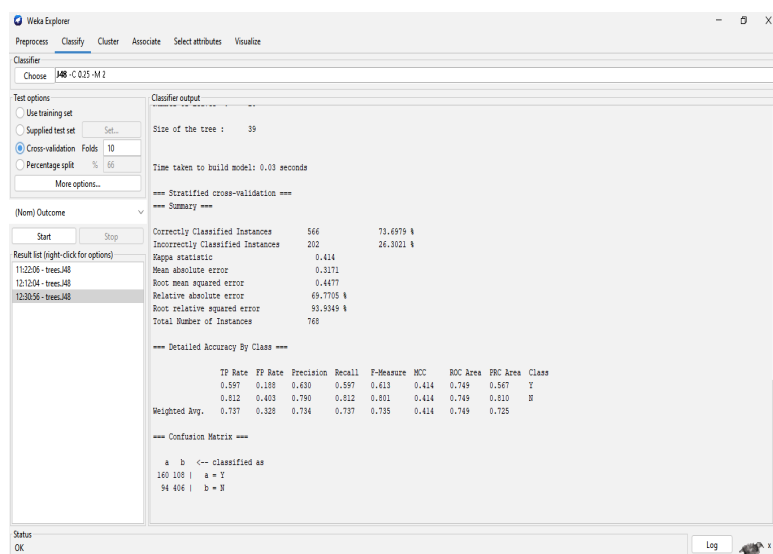
Analisis lebih lanjut dilakukan dengan melihat detail akurasi berdasarkan masing-masing kelas. Untuk kelas "Y" (penderita diabetes), model memiliki nilai True Positive Rate (Recall) sebesar 0,664, yang berarti 66,4% dari seluruh data penderita diabetes berhasil dikenali dengan benar. Precision-nya mencapai 0,848, artinya dari seluruh prediksi "positif" diabetes, 84,8% adalah benar-benar penderita diabetes. F1-score untuk kelas ini berada pada angka 0,745, yang menunjukkan keseimbangan antara Precision dan Recall. Sementara itu, untuk kelas "N" (tidak menderita diabetes), Recall sangat tinggi yaitu 0,936, menunjukkan bahwa hampir seluruh pasien non-diabetes dikenali dengan tepat, meskipun Precision-nya sedikit lebih rendah yakni 0,839. Ini berarti model sangat andal dalam mengenali pasien sehat, meskipun masih terdapat kesalahan klasifikasi terhadap sebagian kecil penderita.

Berdasarkan confusion matrix, dari total 268 pasien yang benar-benar menderita diabetes, sebanyak 178 diklasifikasikan dengan benar sebagai penderita (True Positive), namun 90 lainnya diklasifikasikan keliru sebagai non-diabetes (False Negative). Sementara itu, dari 500 pasien yang tidak menderita diabetes, sebanyak 468 diklasifikasikan dengan benar (True Negative), dan 32 diklasifikasikan keliru sebagai penderita (False Positive). Metrik evaluasi secara keseluruhan menunjukkan nilai Precision, Recall, dan F-Measure yang seimbang, dengan nilai rata-rata tertimbang (weighted average) di atas 0,83 untuk semua parameter tersebut. Selain itu, nilai ROC Area juga tinggi yaitu 0,888, yang menandakan bahwa model memiliki kemampuan klasifikasi yang stabil dan andal.

Secara umum, hasil ini menunjukkan bahwa algoritma C4.8 dapat digunakan secara efektif untuk klasifikasi awal penyakit diabetes. Model mampu mengenali pasien non-diabetes dengan akurasi sangat tinggi, dan cukup baik dalam mengidentifikasi penderita diabetes, meskipun masih terdapat kesalahan klasifikasi yang perlu diperbaiki pada kelas minoritas. Dengan peningkatan pada preprocessing data atau kombinasi model, performa klasifikasi dapat ditingkatkan lebih lanjut.

3. Cross Validation

Hasil evaluasi menggunakan metode 10-fold cross-validation pada model klasifikasi algoritma C4.8 menunjukkan bahwa dari 768 instance, sebanyak 566 data (73,70%) berhasil diklasifikasikan dengan benar, sedangkan 202 data (26,30%) diklasifikasikan secara salah. Tingkat akurasi ini lebih rendah dibandingkan hasil pelatihan penuh (full training), yang menandakan bahwa model mengalami sedikit penurunan performa ketika diuji menggunakan skema validasi silang. Nilai Kappa Statistic sebesar 0,414 mencerminkan tingkat kesepakatan sedang antara prediksi model dan label sebenarnya, dan menunjukkan adanya peluang untuk perbaikan dalam meningkatkan konsistensi prediksi.



Gambar 3. Cross Validation

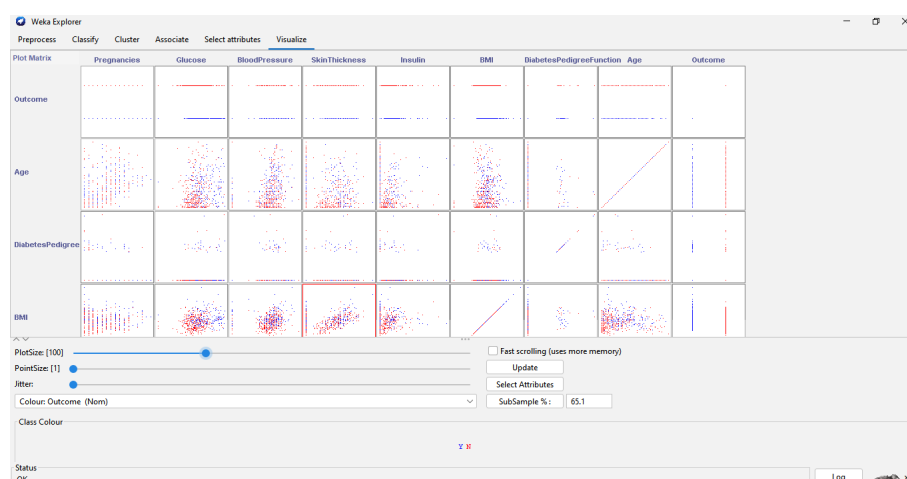
Evaluasi kesalahan model menghasilkan nilai Mean Absolute Error (MAE) sebesar 0,3171 dan Root Mean Squared Error (RMSE) sebesar 0,4477, yang menunjukkan adanya deviasi yang lebih tinggi dibandingkan hasil training penuh. Sementara itu, Relative Absolute Error mencapai 69,77% dan Root Relative Squared Error sebesar 93,93%, yang menandakan bahwa tingkat kesalahan relatif terhadap baseline model cukup tinggi, terutama karena ketidakseimbangan data antara kelas positif dan negatif.

Dari sisi performa per kelas, untuk kelas Y (penderita diabetes), model memiliki True Positive Rate (Recall) sebesar 0,597 dan Precision sebesar 0,630, yang menunjukkan bahwa hanya sekitar 59,7% kasus diabetes berhasil dikenali dengan benar, dan dari seluruh prediksi positif, hanya 63% yang benar-benar positif. Nilai F-Measure untuk kelas ini adalah 0,613, menandakan performa yang sedang. Sementara itu, untuk kelas N (tidak menderita diabetes), model menunjukkan performa lebih baik dengan Recall sebesar 0,812 dan Precision sebesar 0,790, serta F-Measure sebesar 0,801. Ini mengindikasikan bahwa model lebih akurat dalam mengklasifikasikan pasien yang tidak menderita diabetes dibandingkan dengan mendeteksi kasus diabetes.

Confusion matrix menunjukkan bahwa dari 268 pasien penderita diabetes, hanya 160 diklasifikasikan dengan benar (True Positive), sedangkan 108 salah klasifikasi sebagai non-diabetes (False Negative). Sementara dari 500 pasien yang tidak menderita diabetes, 406 diklasifikasikan dengan benar (True Negative) dan 94 salah diklasifikasikan sebagai penderita (False Positive). Ketidakseimbangan dalam performa antar kelas ini menjadi indikasi bahwa model cenderung bias terhadap kelas mayoritas (non-diabetes), dan membutuhkan peningkatan seperti penyesuaian threshold, penyeimbangan data, atau penggunaan algoritma ensemble.

4. Visualisasi

Berdasarkan hasil analisis visual dari *plot sebar (scatter plot matrix)* yang ditampilkan pada tab Visualize di WEKA, atribut Glukosa terbukti menjadi variabel paling dominan dalam membedakan kelas target (*Outcome*) pada dataset diabetes. Hal ini ditunjukkan oleh distribusi titik-titik data yang secara jelas membedakan antara pasien yang terdiagnosis diabetes (ditandai dengan titik Merah) dan pasien non-diabetes (titik Biru). Mayoritas pasien dengan nilai glukosa tinggi tergolong dalam kelas diabetes, sedangkan pasien dengan kadar glukosa rendah sebagian besar termasuk dalam kategori non-diabetes. Selain Glukosa, atribut BMI (Body Mass Index) dan Age (Usia) juga menunjukkan korelasi yang kuat terhadap Outcome. Titik-titik merah cenderung terkonsentrasi pada rentang nilai yang lebih tinggi untuk kedua atribut ini, menandakan bahwa pasien dengan usia lebih lanjut dan indeks massa tubuh lebih tinggi memiliki kemungkinan lebih besar menderita diabetes.



Gambar 4. Visualisasi

Sebaliknya, atribut seperti BloodPressure (Tekanan Darah) dan SkinThickness tidak menunjukkan pola distribusi yang berbeda secara signifikan antara dua kelas, karena titik-titik dari kedua kelas tampak bercampur dan tidak membentuk cluster yang jelas. Hal ini mengindikasikan

bahwa kedua atribut tersebut cenderung kurang informatif secara individual dalam proses klasifikasi, terutama jika digunakan tanpa kombinasi atribut lain. Secara keseluruhan, hasil visualisasi ini memperkuat temuan sebelumnya bahwa atribut Glukosa, BMI, dan Usia harus diprioritaskan dalam proses pembentukan model klasifikasi, karena memiliki kontribusi paling besar dalam memisahkan kelas dengan jelas dan konsisten. Dengan demikian, pemanfaatan ketiga variabel ini secara optimal dapat meningkatkan akurasi dan ketepatan dalam mendeteksi risiko diabetes melalui pendekatan pembelajaran mesin.

5. KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma C4.8 (pengembangan dari C4.5) memiliki kemampuan yang cukup baik dalam mengklasifikasikan data pasien terkait risiko penyakit diabetes berdasarkan atribut-atribut medis. Berdasarkan hasil training model, algoritma C4.8 berhasil mengklasifikasikan data dengan akurasi tinggi, yaitu 84,11%, menunjukkan performa optimal ketika model diuji pada data yang telah dikenalnya. Namun, saat diuji menggunakan 10-fold cross-validation, akurasi model menurun menjadi 73,70%, yang menandakan bahwa model masih memiliki keterbatasan dalam generalisasi terhadap data baru.

Evaluasi detail dari metrik klasifikasi memperlihatkan bahwa model memiliki kemampuan yang lebih baik dalam mengenali pasien non-diabetes (kelas negatif) dibandingkan pasien yang benar-benar menderita diabetes (kelas positif). Hal ini tampak dari nilai recall yang lebih tinggi untuk kelas non-diabetes. Atribut Glukosa terbukti sebagai faktor paling dominan dalam proses klasifikasi, sebagaimana ditunjukkan baik oleh struktur pohon keputusan maupun visualisasi scatter plot di WEKA. Atribut BMI dan Usia juga memberikan kontribusi signifikan, sedangkan atribut seperti BloodPressure dan SkinThickness kurang efektif jika dilihat secara individu.

Secara keseluruhan, algoritma C4.8 dapat digunakan sebagai pendekatan awal dalam sistem pendukung keputusan medis untuk deteksi dini diabetes tipe 2, terutama pada dataset yang bersih dan terstruktur. Namun, performa model masih dapat ditingkatkan dengan pendekatan lain seperti teknik balancing data, optimasi parameter, atau penerapan metode ensemble. Penelitian ini juga menegaskan pentingnya seleksi atribut dalam proses klasifikasi, di mana variabel Glukosa, BMI, dan Usia menjadi kunci dalam membedakan antara pasien diabetes dan non-diabetes.

DAFTAR PUSTAKA

- [1] J. Kesehatan, P. Adminstras, R. S. Fakultas, K. Teknologi, and U. Bumi, "MODEL KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN," vol. 4, pp. 8–12, 2025.
- [2] F. B. Bahtiarullah, A. Homaidi, and F. Santoso, "Perbandingan Performa Algoritma Klasifikasi C4.5 Dan Naïve Bayes Untuk Prediksi Diagnosa Penyakit Diabetes," *E-Link J. Tek. Elektro dan Inform.*, vol. 19, no. 2, p. 155, 2024, doi: 10.30587/e-link.v19i2.8137.
- [3] R. P. Fadhillah, R. Rahma, A. Sepharni, R. Mufidah, B. N. Sari, and A. Pangestu, "Klasifikasi Penyakit Diabetes Mellitus Berdasarkan Faktor-Faktor Penyebab Diabetes menggunakan Algoritma C4.5," *JUPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.)*, vol. 7, no. 4, pp. 1265–1270, 2022, doi: 10.29100/jupi.v7i4.3248.
- [4] N. Nurussakinah, M. Faisal, and I. B. Santoso, "Algoritma Random Forest dan Synthetic Minority Oversampling Technique (SMOTE) untuk Deteksi Diabetes," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 10, no. 2, pp. 221–234, 2025, doi: 10.14421/jiska.2025.10.2.221-234.
- [5] M. F. Aditya, A. Pramuntadi, D. P. Wijaya, and Y. Wicaksono, "Implementasi Metode Decision Tree pada Prediksi Penyakit Diabetes Melitus Tipe 2," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, pp. 1104–1110, 2024, doi: 10.57152/malcom.v4i3.1284.
- [6] Y. Nuryamin and F. Risyda, "JSI (Jurnal Sistem Informasi) Universitas Dirgantara Marsekal Suryadarma Analisis Prediksi Penyakit Diabetes Menggunakan Metode Decision Tree C4.5 Dan Naive Bayes," vol. 5, pp. 234–242, 2025, [Online]. Available: <https://www.kaggle.com/code/melikedilekci/diabetes-dataset-for-beginners>
- [7] Yennimar, W. Leonardi, H. Weide, D. Cantona, and Gani Mores Hutagalung, "Comparison of data mining algorithms (random forest, C4.5, catboost) based on adaptive boosting in predicting diabetes mellitus," *J. Tek. Inform. C.I.T Medicom*, vol. 16, no. 3, pp. 01–12, 2024, [Online]. Available: www.medikom.iocspublisher.org/journalhomepage:www.medikom.iocspublisher.org
- [8] P. Nopedi and A. Saifudin, "Optimasi Decision Tree Menggunakan Teknik Boosting pada Prediksi Penyakit Diabetes," *J. Inform. Univ. Pamulang*, vol. 8, no. 3, pp. 465–474, 2023, doi: 10.32493/informatika.v8i3.35226.

- [9] R. Komarudin, P. V. B. Yudha, Y. I. Maulana, N. Afni, A. Salim, and I. Carolina, "Penerapan Metode Algoritma C4.5 Dalam Klasifikasi Diagnosa Penyakit Umum Menggunakan WEKA," *Bianglala Inform.*, vol. 9, no. 2, pp. 95–102, 2021, doi: 10.31294/bi.v9i2.11537.
- [10] Z. D. R. Sari, J. Jasmir, and Y. Arvita, "Penerapan Data Mining Untuk Prediksi Penyakit Diabetes Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM)," *J. Inform. Dan Rekayasa Komput.*, vol. 4, no. April, pp. 827–834, 2024.
- [11] Hendriyana and I. M. K. Karo, "Klasifikasi Penderita Diabetes Menggunakan Algoritma Machine Learning Dan Z-Score," *J. Teknol. Terpadu*, vol. 8, no. 2, pp. 94–99, 2022.
- [12] Jupron and Sutrisno, "Analysis of Heart Disease Using the Random Forest Method," *J. Inotera*, vol. 10, no. 1, pp. 167–174, 2025, doi: 10.31572/inotera.vol10.iss1.2025.id452.
- [13] S. Sutrisno and Jupron, "Analisa Klasifikasi Penyakit Diabetes Dengan Algoritma Neural Network," *bit-Tech*, vol. 6, no. 3, pp. 303–310, 2024, doi: 10.32877/bt.v6i3.1161.