SAINTEK: Jurnal Sains, Teknologi & Komputer
Vol. 1, No. 3, Oktober 2024, pp. 75 ~ 79

Analisis Faktor Risiko Diabetes Melitus Menggunakan Algoritma C4.5: Implementasi pada Aplikasi Orange

Mutiara Akbar Nasution[1], Zaid Ahlun Ulumuddin[2], Anisa Fitri[3]
[1,3]Program Studi Bisnis Digital, Universitas Negeri Medan, Indonesia
[2]Program Studi Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara, Indonesia
Email: mutiaraakbarnst03@gmail.com; zaidahlun93@gmail.com; anisafitri.7212250002@mhs.unimed.ac.id

ABSTRAK

Penelitian ini bertujuan untuk menganalisis faktor risiko diabetes melitus dan membangun model prediksi menggunakan algoritma C4.5 yang diimplementasikan melalui aplikasi Orange. Dataset yang digunakan adalah *Pima Indians Diabetes Dataset* dari UCI Machine Learning Repository yang berisi 768 sampel dengan delapan variabel prediktor dan satu variabel target. Proses analisis meliputi *preprocessing* data, pembagian data latih dan uji (80:20), pembangunan model pohon keputusan, serta evaluasi performa menggunakan metrik akurasi, presisi, *recall*, F1-score, dan AUC. Hasil penelitian menunjukkan bahwa model C4.5 mencapai nilai akurasi sebesar 71,9% dan AUC sebesar 0,674, yang menandakan kemampuan klasifikasi cukup baik. Faktor yang paling berpengaruh terhadap risiko diabetes adalah kadar glukosa darah, indeks massa tubuh (BMI), dan usia. Penelitian ini memberikan kontribusi terhadap pengembangan model prediksi penyakit berbasis data mining yang mudah diterapkan serta dapat mendukung proses pengambilan keputusan dalam deteksi dini diabetes melitus.

Keyword: Data Mining; Algoritma C4.5; Diabetes Melitus; Aplikasi Orange; Faktor Risiko

ABSTRACT

This study aims to analyze the risk factors for diabetes mellitus and build a predictive model using the C4.5 algorithm implemented through the Orange application. The dataset used is the Pima Indians Diabetes Dataset from the UCI Machine Learning Repository containing 768 samples with eight predictor variables and one target variable. The analysis process includes data preprocessing, splitting training and testing data (80:20), building a decision tree model, and evaluating performance using accuracy, precision, recall, F1-score, and AUC metrics. The results show that the C4.5 model achieved an accuracy value of 71.9% and an AUC of 0.674, indicating a fairly good classification ability. The most influential factors on diabetes risk are blood glucose levels, body mass index (BMI), and age. This study contributes to the development of a data mining-based disease prediction model that is easy to implement and can support the decision-making process in the early detection of diabetes mellitus.

Keyword: Data Mining; C4.5 Algorithm; Diabetes Mellitus; Orange Application; Risk Factors

Corresponding Author:
Mutiara Akbar Nasution,
Universitas Negeri Medan,
Jl. William Iskandar Ps. V, Kenangan Baru, Kec. Percut Sei Tuan, Kabupaten Deli Serdang, Sumatera Utara 20221, Indonesia
Email: mutiaraakbarnst03@gmail.com

1. PENDAHULUAN

Diabetes melitus (DM) merupakan salah satu penyakit tidak menular yang menjadi ancaman serius bagi kesehatan masyarakat dunia, termasuk di Indonesia. Berdasarkan hasil *Riset Kesehatan Dasar (Riskesdas)* tahun 2018, prevalensi diabetes mencapai 8,5%, meningkat dari 6,9% pada tahun 2013 (Kemenkes RI, 2018). Peningkatan ini menunjukkan urgensi untuk mengembangkan strategi deteksi dini dan pencegahan yang lebih efektif guna menekan angka kejadian diabetes di masa mendatang.

Dalam konteks tersebut, penerapan teknologi analitik seperti *data mining* menjadi pendekatan yang menjanjikan untuk menganalisis data kesehatan dalam skala besar. *Data mining* memungkinkan peneliti menemukan pola tersembunyi dan hubungan antarvariabel yang mungkin tidak terdeteksi melalui analisis konvensional (Sari & Jasmir, 2024). Salah satu algoritma *data mining* yang banyak digunakan untuk klasifikasi penyakit adalah algoritma C4.5, yaitu metode *decision tree* yang mampu menghasilkan model prediktif yang dapat diinterpretasikan dengan mudah (Hana, 2020). Melalui algoritma ini, faktor-faktor risiko dapat diidentifikasi secara sistematis berdasarkan kontribusi atribut terhadap terjadinya penyakit.

Aplikasi Orange merupakan platform *data mining* berbasis visual yang menyediakan antarmuka interaktif untuk membangun model analisis tanpa perlu menulis kode secara manual (Demšar et al., 2013). Keunggulan Orange terletak pada kemampuannya mengintegrasikan proses *preprocessing*, pembentukan model, evaluasi, serta visualisasi hasil analisis dalam satu ekosistem yang intuitif. Hal ini menjadikan Orange sebagai alat yang efisien dan edukatif, khususnya dalam bidang kesehatan yang memerlukan interpretasi hasil secara visual dan transparan.

Penelitian ini bertujuan untuk menganalisis faktor-faktor risiko diabetes melitus menggunakan algoritma C4.5 yang diimplementasikan melalui aplikasi Orange dengan memanfaatkan *Pima Indians Diabetes Dataset* dari UCI Machine Learning Repository. Variabel yang dianalisis meliputi jumlah kehamilan (*pregnancies*), kadar glukosa darah (*glucose*), tekanan darah (*blood pressure*), ketebalan lipatan kulit (*skin thickness*), kadar insulin (*insulin*), indeks massa tubuh (*BMI*), riwayat keluarga diabetes (*diabetes pedigree function*), dan usia (*age*).

Melalui pendekatan ini, diharapkan dapat diperoleh pemahaman yang lebih komprehensif mengenai variabel yang paling berpengaruh terhadap risiko diabetes melitus. Hasil penelitian diharapkan tidak hanya memberikan kontribusi terhadap pengembangan model prediksi berbasis *data mining*, tetapi juga dapat mendukung tenaga medis dan pembuat kebijakan dalam upaya deteksi dini serta pengendalian diabetes di Indonesia secara lebih efektif dan berbasis data.

2. METODE PENELITIAN

A. Sumber dan Deskripsi Data

Penelitian ini menggunakan dataset *Pima Indians Diabetes* yang diperoleh dari UCI Machine Learning Repository, dan dapat diakses melalui platform Kaggle. Dataset ini terdiri atas 768 sampel data dengan delapan variabel prediktor dan satu variabel target. Variabel prediktor meliputi: jumlah kehamilan (*pregnancies*), kadar glukosa darah (*glucose*), tekanan darah diastolik (*blood pressure*), ketebalan lipatan kulit (*skin thickness*), kadar insulin serum dua jam (*insulin*), indeks massa tubuh (*BMI*), fungsi riwayat keluarga diabetes (*diabetes pedigree function*), dan usia (*age*). Sementara itu, variabel target (*outcome*) merupakan data kategorikal yang menunjukkan status diabetes individu, dengan nilai 0 (tidak menderita) atau 1 (menderita diabetes).

B. Tahap Preprocessing Data

Sebelum proses analisis dilakukan, data menjalani tahap *preprocessing* untuk memastikan kualitas dan konsistensi. Pertama, variabel *outcome* dikonversi dari nilai numerik (0 dan 1) menjadi label kategorikal "Negative" dan "Positive" guna mempermudah interpretasi. Selanjutnya, data dibagi menjadi dua subset: 80% sebagai data latih (*training set*) dan 20% sebagai data uji (*testing set*). Pembagian ini bertujuan untuk menguji performa generalisasi model terhadap data yang belum pernah dilihat sebelumnya.

C. Pembangunan Model dengan Orange

Seluruh proses analisis dilakukan menggunakan aplikasi Orange, sebuah platform *data mining* berbasis antarmuka visual. Dataset dimuat ke dalam Orange melalui widget *File*, kemudian dilakukan penyesuaian variabel dengan widget *Edit Domain*. Pembagian data dilakukan menggunakan *Data Sampler*, dan pembangunan model pohon keputusan dilakukan dengan widget *Tree* menggunakan algoritma C4.5. Dalam tahap ini, parameter seperti *max depth*, *min subgroup size*, dan kriteria pemilihan atribut (*selection measure*) dapat disesuaikan untuk mengoptimalkan hasil model.

D. Evaluasi dan Visualisasi Model

Evaluasi performa model dilakukan dengan menggunakan widget *Test & Score*, yang mengukur metrik-metrik utama seperti akurasi, presisi, *recall*, F1-score, dan AUC (*Area Under Curve*). Visualisasi pohon keputusan diperoleh melalui widget *Tree Viewer* untuk mempermudah interpretasi aturan klasifikasi dan struktur pohon. Hal ini memungkinkan peneliti untuk mengamati cabang-cabang keputusan yang mengindikasikan faktor risiko utama berdasarkan pembagian atribut.

E. Analisis Hasil

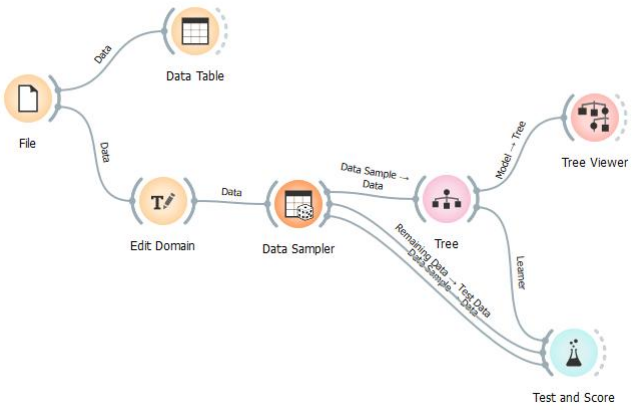
Model yang telah dibangun kemudian dianalisis berdasarkan metrik evaluasi untuk menilai akurasi dan reliabilitasnya dalam memprediksi risiko diabetes. Selain itu, struktur pohon keputusan yang dihasilkan dianalisis lebih lanjut untuk mengidentifikasi atribut-atribut yang paling berkontribusi terhadap klasifikasi

diabetes, sehingga dapat ditarik kesimpulan mengenai faktor-faktor risiko yang paling dominan dalam dataset tersebut.

3. HASIL DAN PEMBAHASAN

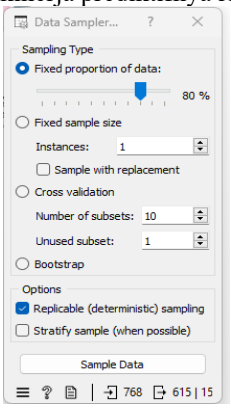
A. Rangkaian Proses Analisis pada Aplikasi Orange

Proses analisis data dalam penelitian ini dilakukan menggunakan aplikasi Orange melalui serangkaian widget yang membentuk alur kerja terintegrasi dari tahap input data hingga visualisasi hasil. Widget *File* digunakan untuk memuat dataset *Pima Indians Diabetes* dalam format *.csv* yang diperoleh dari Kaggle. Selanjutnya, widget *Edit Domain* dimanfaatkan untuk menyesuaikan variabel target, yaitu mengubah nilai *Outcome* dari "0" dan "1" menjadi "Negative" dan "Positive" agar hasil klasifikasi lebih mudah diinterpretasikan.



Gambar 1. Rangkaian Widget

Proses pembagian data dilakukan menggunakan widget *Data Sampler* dengan proporsi 80% untuk *training set* dan 20% untuk *testing set*. Pembagian ini bertujuan agar model dapat diuji menggunakan data yang belum pernah dilihat sebelumnya, sehingga kinerja prediktifnya lebih objektif.



Gambar 2. Pembagian Data

Widget *Tree* menjadi komponen utama dalam proses analisis, di mana algoritma C4.5 diterapkan untuk membangun model *decision tree*. Parameter yang disesuaikan antara lain *beam width*, *min subgroup size*, *max depth*, dan *selection measure*, sementara fitur *pruning* diaktifkan untuk menghindari overfitting. Setelah model terbentuk, evaluasi dilakukan menggunakan widget *Test & Score*, sedangkan struktur pohon divisualisasikan dengan widget *Tree Viewer* untuk memahami aturan klasifikasi yang terbentuk.

Dengan rangkaian widget tersebut, seluruh tahapan analisis — mulai dari *preprocessing*, pembangunan model, hingga evaluasi — dapat dilakukan secara visual, sistematis, dan efisien tanpa memerlukan penulisan kode manual. Pendekatan ini memperlihatkan keunggulan Orange sebagai alat eksplorasi *data mining* yang mudah digunakan oleh peneliti dan tenaga kesehatan.

B. Hasil Evaluasi Model

Model pohon keputusan yang dibangun menggunakan algoritma C4.5 diuji menggunakan data uji (*testing set*) sebanyak 20% dari total sampel. Hasil evaluasi model ditunjukkan pada Tabel 1 berikut:

Tabel 1. Hasil Evaluasi Model C4.5 pada Dataset Pima Indians Diabetes

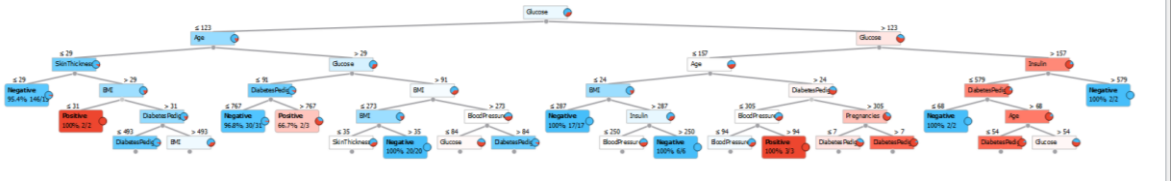
Metrik Evaluasi	Nilai
AUC (*Area Under Curve*)	0.674
CA (*Classification Accuracy*)	0.719
F1-Score	0.717

Metrik Evaluasi	Nilai
Precision	0.716
Recall	0.719
MCC (*Matthews Correlation Coefficient*)	0.397

Nilai *Classification Accuracy* sebesar 71,9% menunjukkan bahwa model mampu mengklasifikasikan data dengan cukup baik. Nilai F1-Score (0,717) serta *Precision* dan *Recall* yang hampir seimbang menandakan bahwa model memiliki kestabilan dalam mendeteksi kasus positif diabetes tanpa banyak kesalahan klasifikasi. Sementara itu, nilai AUC sebesar 0,674 menunjukkan bahwa model memiliki kemampuan diskriminatif yang moderat dalam membedakan antara pasien diabetes dan non-diabetes. Nilai MCC sebesar 0,397 mengindikasikan adanya korelasi positif yang cukup antara hasil prediksi dan kondisi aktual.

C. Visualisasi dan Interpretasi Pohon Keputusan

Visualisasi hasil pemodelan menggunakan widget *Tree Viewer* pada aplikasi Orange menghasilkan struktur pohon keputusan seperti ditunjukkan pada Gambar 3. Pohon keputusan ini dibangun berdasarkan algoritma C4.5 dengan variabel *glucose* sebagai simpul akar (*root node*), menandakan bahwa kadar glukosa darah merupakan faktor utama dalam menentukan klasifikasi diabetes.



Gambar 3. Visualisasi Pohon Keputusan Algoritma C4.5 pada Dataset Pima Indians Diabetes

Struktur pohon menunjukkan bahwa individu dengan nilai *glucose* lebih dari 127,5 diklasifikasikan cenderung Positive (berisiko diabetes). Sebaliknya, jika *glucose* berada di bawah ambang tersebut, model masih melakukan percabangan lanjutan menggunakan variabel lain seperti *BMI*, *Insulin*, *Age*, dan *DiabetesPedigreeFunction* untuk menentukan hasil klasifikasi yang lebih spesifik.

Beberapa pola penting yang dapat diidentifikasi dari pohon keputusan ini antara lain:

- Glucose menjadi penentu utama; nilai *glucose* tinggi secara konsisten berasosiasi dengan kelas *Positive*.
- BMI dan Insulin muncul berulang sebagai simpul sekunder, menunjukkan bahwa keduanya memiliki kontribusi signifikan terhadap risiko diabetes. Individu dengan BMI di atas 30 dan kadar insulin tinggi cenderung termasuk ke dalam kelas *Positive*.
- Age juga berperan penting pada cabang tertentu, di mana pasien berusia di atas 40 tahun lebih sering terklasifikasi sebagai *Positive*.
- Nilai DiabetesPedigreeFunction (riwayat keluarga) menjadi faktor pendukung, terutama pada kasus dengan *glucose* menengah, yang membantu model memperkuat keputusan klasifikasi.

Secara keseluruhan, struktur pohon ini memberikan representasi visual mengenai logika klasifikasi yang digunakan oleh algoritma C4.5. Model mampu mengidentifikasi kombinasi atribut seperti kadar glukosa tinggi, indeks massa tubuh di atas normal, dan usia lanjut sebagai indikator kuat terhadap kemungkinan seseorang menderita diabetes. Hasil ini sejalan dengan literatur medis yang menyatakan bahwa obesitas, resistensi insulin, dan faktor usia merupakan determinan utama dalam perkembangan diabetes melitus tipe 2 (American Diabetes Association, 2019; Aris & Benyamin, 2019).

D. Pembahasan

Temuan ini memperkuat hasil penelitian sebelumnya yang menyatakan bahwa kadar glukosa darah, indeks massa tubuh, dan usia merupakan prediktor kuat terjadinya diabetes melitus (Hana, 2020). Nilai akurasi 71,9% tergolong kompetitif dibandingkan penelitian sejenis yang menggunakan algoritma pohon keputusan pada dataset yang sama, yang umumnya berkisar antara 70–75%.

Kelebihan penelitian ini terletak pada penggunaan aplikasi Orange yang memungkinkan visualisasi hasil analisis secara langsung, sehingga interpretasi pohon keputusan menjadi lebih intuitif dan mudah dipahami bahkan oleh pengguna non-programmer. Namun, keterbatasan penelitian ini adalah belum adanya proses optimasi parameter (misalnya *pruning depth* atau *split criteria*) secara sistematis, serta penggunaan dataset internasional yang mungkin belum sepenuhnya merepresentasikan karakteristik populasi Indonesia.

Penelitian lanjutan disarankan untuk menggunakan dataset lokal yang lebih representatif dan membandingkan performa algoritma C4.5 dengan metode lain seperti *Random Forest*, *Naïve Bayes*, atau *Support Vector Machine* untuk memperoleh model prediksi yang lebih robust.

4. KESIMPULAN

Penelitian ini menunjukkan potensi besar penerapan *data mining* dalam bidang kesehatan, khususnya melalui penggunaan algoritma C4.5 yang diimplementasikan menggunakan aplikasi Orange, untuk mengidentifikasi faktor-faktor risiko diabetes melitus. Model pohon keputusan yang dihasilkan dari dataset *Pima Indians Diabetes* terbukti mampu memprediksi risiko diabetes dengan tingkat akurasi sebesar 71,9% dan nilai AUC sebesar 0,674, yang menunjukkan performa klasifikasi yang cukup baik. Lebih dari itu, hasil visualisasi pohon keputusan memberikan pemahaman mendalam mengenai kontribusi setiap variabel prediktor terhadap kejadian diabetes.

Faktor-faktor seperti kadar glukosa darah, indeks massa tubuh (*Body Mass Index*), usia, dan kadar insulin muncul sebagai atribut paling berpengaruh dalam penentuan risiko diabetes. Hal ini sejalan dengan teori dan bukti medis bahwa kadar glukosa tinggi, obesitas, dan faktor usia merupakan determinan utama dalam perkembangan diabetes melitus tipe 2. Dengan demikian, penelitian ini tidak hanya memberikan model prediksi yang cukup akurat, tetapi juga memperkuat landasan ilmiah untuk memahami hubungan antara variabel klinis dengan risiko penyakit.

Secara praktis, hasil penelitian ini dapat dimanfaatkan oleh tenaga medis dan peneliti kesehatan sebagai alat bantu dalam melakukan deteksi dini serta perencanaan intervensi preventif. Implementasi algoritma C4.5 pada aplikasi Orange juga membuktikan bahwa teknologi *data mining* dapat digunakan secara efektif tanpa memerlukan kemampuan pemrograman lanjutan, sehingga membuka peluang pemanfaatan lebih luas di sektor kesehatan publik dan penelitian akademik.

Meskipun demikian, penelitian ini memiliki keterbatasan pada ukuran dan representasi data yang masih terbatas pada populasi luar negeri. Oleh karena itu, penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih besar dan berasal dari populasi Indonesia agar model yang dihasilkan lebih kontekstual dan akurat. Pengembangan lebih lanjut juga dapat dilakukan melalui perbandingan algoritma lain seperti *Random Forest*, *Naïve Bayes*, atau *Support Vector Machine* untuk memperoleh model prediksi yang lebih optimal.

Dengan pengembangan tersebut, algoritma C4.5 dan pendekatan *data mining* serupa berpotensi menjadi alat analisis yang berharga dalam mendukung sistem deteksi dini, perencanaan kebijakan kesehatan berbasis data, serta pengendalian diabetes secara nasional di Indonesia.

REFERENSI

Aditya, M. F., Pramuntadi, A., Wijaya, D. P., & Wicaksono, Y. (2024). Implementasi metode decision tree pada prediksi penyakit diabetes melitus tipe 2. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(3), 1104-1110. https://doi.org/10.57152/malcom.v4i3.1284

American Diabetes Association. (2019). Diabetes advocacy: Standards of medical care in diabetes. *Diabetes care*, *43*(Supplement_1), S203-S204. https://doi.org/10.2337/dc20-S016

Aris, F., Benyamin (2019). Penerapan data mining untuk identifikasi penyakit diabetes melitus dengan menggunakan metode klasifikasi. *Router Research*, *1*(1), 1-6. https://doi.org/10.29239/j.router.2019.313

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Stajdohar, M., Umek, L., Žagar, L., Žbontar, J., & Zupan, B. (2013). Orange: Data mining toolbox in Python. *Journal of Machine Learning Research, 14*(1), 2349–2353.

Hana, F. M. (2020). Klasifikasi penderita penyakit diabetes menggunakan algoritma decision tree C4. 5. *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, *4*(1), 32-39. https://doi.org/10.47970/siskom-kb.v4i1.173

International Diabetes Federation. (2021). *IDF diabetes atlas* (10th ed.). Author.

Kementerian Kesehatan Republik Indonesia. (2018). *Hasil utama Riskesdas 2018*. Kementerian Kesehatan Republik Indonesia.

Noviandi, N. (2018). Implementasi algoritma decision tree c4.5 untuk prediksi penyakit diabetes. *Indonesian of Health Information Management Journal (INOHIM)*, *6*(1), 1-5. https://doi.org/10.47007/inohim.v6i1.142

Sari, Z. D. R., & Jasmir, J. (2024). Penerapan Data Mining untuk prediksi penyakit diabetes menggunakan algoritma C4.5. *Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM)*, *4*(1), 827-834. https://doi.org/10.33998/jakakom.2024.4.1.1624

Siallagan, R. A. (2021). Prediksi penyakit diabetes mellitus menggunakan algoritma c4.5. *Jurnal Responsif: Riset Sains dan Informatika*, *3*(1), 44-52. https://doi.org/10.51977/jti.v3i1.407

Velu, S. R., Ravi, V., & Tabianan, K. (2023). Machine learning implementation to predict type-2 diabetes mellitus based on lifestyle behaviour pattern using HBA1C status. Health and Technology, 13(3), 437-447. https://doi.org/10.1007/s12553-023-00751-5