

Analisis Faktor Risiko Diabetes Menggunakan Algoritma Decision Tree

Muhammad Ibra Fatahilah Akbar¹⁾, Farhan Adiputra Subagya²⁾

¹⁾ Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Jl. Kramat Raya No.98, Senen, Jakarta Pusat 10450.; Telp. (021)8000063
Email: ibrafatahilah.24@gmail.com

Abstrak

Diabetes mellitus merupakan penyakit kronis dengan prevalensi yang meningkat dan konsekuensi kesehatan serius. Penelitian ini bertujuan mengidentifikasi faktor risiko yang berpengaruh terhadap kemungkinan diagnosis diabetes serta mengevaluasi kinerja algoritma Decision Tree dalam klasifikasi risiko menggunakan dataset sekunder berjumlah 768 sampel. Variabel yang dianalisis meliputi jumlah kehamilan, kadar glukosa darah, tekanan darah, ketebalan kulit, kadar insulin, Indeks Massa Tubuh (BMI), riwayat diabetes keluarga, dan usia. Pra-pengolahan dilakukan dengan imputasi nilai nol yang dianggap tidak valid untuk beberapa fitur (Glukosa, Tekanan Darah, SkinThickness, Insulin, BMI) menggunakan median tergrup berdasarkan kelas outcome, serta pembagian data latih:uji 70:30. Model Decision Tree (criterion='gini') dilatih dan dievaluasi menggunakan confusion matrix, precision, recall, dan F1-score. Pada data uji (n=231), model mencapai akurasi 70.1% dengan matriks kebingungan: TN=107, FP=44, FN=25, TP=55. Variabel paling dominan pada struktur pohon adalah kadar glukosa dan BMI. Hasil menunjukkan Decision Tree memberikan model yang interpretatif dan layak sebagai alat bantu deteksi dini, meskipun peningkatan akurasi dapat dicapai melalui penanganan data dan teknik ensemble.

Kata kunci: Decision Tree; Diabetes; Klasifikasi; Machine Learning

Abstract

Diabetes mellitus is a chronic disease with rising prevalence and serious health consequences. This study aims to identify risk factors associated with diabetes diagnosis and to evaluate the performance of a Decision Tree classifier using a secondary dataset of 768 samples. Features include number of pregnancies, plasma glucose concentration, blood pressure, skin thickness, insulin, Body Mass Index (BMI), family diabetes history, and age. Preprocessing involved imputing invalid zero values for selected features using class-wise medians and splitting data into 70% train and 30% test sets. A Decision Tree (criterion='gini') was trained and evaluated via confusion matrix, precision, recall, and F1-score. On the test set (n=231), the model achieved 70.1% accuracy with confusion matrix: TN=107, FP=44, FN=25, TP=55. Glucose and BMI were the most influential features in the tree structure. The Decision Tree offers interpretable rules useful for early detection, but accuracy can be improved using better preprocessing or ensemble methods.

Keywords: Decision Tree; Diabetes; Classification; Machine Learning

1. PENDAHULUAN

Diabetes mellitus merupakan salah satu penyakit tidak menular yang menjadi perhatian utama dalam bidang kesehatan masyarakat. Peningkatan jumlah penderita diabetes setiap tahun menunjukkan perlunya strategi yang lebih efektif dalam melakukan pencegahan dan deteksi dini terhadap penyakit ini. Keterlambatan diagnosis dapat menyebabkan berbagai komplikasi serius, seperti penyakit jantung, gangguan ginjal, dan kerusakan saraf.

Perkembangan teknologi informasi, khususnya di bidang machine learning, memberikan peluang besar dalam membantu proses analisis data kesehatan. Metode klasifikasi dalam machine learning mampu mengidentifikasi pola tersembunyi pada data medis dan digunakan untuk memprediksi risiko suatu penyakit. Salah satu algoritma klasifikasi yang banyak digunakan adalah Decision Tree, karena mampu menghasilkan model yang mudah dipahami dan diinterpretasikan.

Beberapa penelitian sebelumnya menunjukkan bahwa algoritma Decision Tree cukup efektif dalam menganalisis risiko diabetes dengan memanfaatkan variabel kesehatan seperti kadar glukosa darah, indeks massa tubuh, dan usia. Namun, masih diperlukan penelitian yang menekankan pada interpretasi hasil model serta pemanfaatannya sebagai alat bantu pengambilan keputusan.

Berdasarkan hal tersebut, penelitian ini bertujuan untuk menganalisis faktor-faktor risiko yang berpengaruh terhadap diabetes serta mengevaluasi kinerja algoritma Decision Tree dalam melakukan klasifikasi risiko diabetes berdasarkan data pasien.

Selain berdampak pada individu, diabetes juga memberikan beban ekonomi yang signifikan terhadap sistem pelayanan kesehatan. Oleh karena itu, pemanfaatan data kesehatan secara optimal menjadi kebutuhan penting dalam mendukung kebijakan pencegahan berbasis bukti.

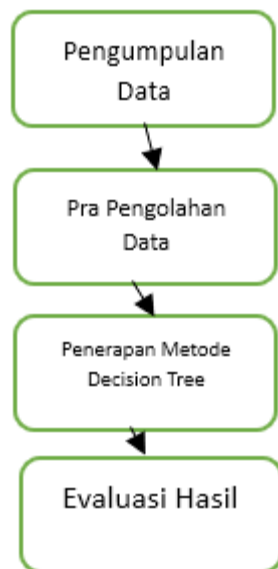
Kontribusi utama penelitian ini terletak pada analisis interpretatif faktor risiko diabetes menggunakan algoritma Decision Tree, sehingga hasil model tidak hanya bersifat prediktif tetapi juga dapat dipahami oleh tenaga kesehatan sebagai dasar pengambilan keputusan.

2. METODE

Penelitian ini menggunakan dataset sekunder yang terdiri dari 768 data pasien. Dataset mencakup delapan variabel independen, yaitu jumlah kehamilan, kadar glukosa darah, tekanan darah, ketebalan kulit, kadar insulin, indeks massa tubuh, riwayat diabetes dalam keluarga, dan usia. Variabel dependen berupa status diabetes pasien yang diklasifikasikan menjadi dua kelas, yaitu tidak diabetes dan diabetes.

2.1. Pra-pengolahan Data

Tahap pra-pengolahan data dilakukan untuk meningkatkan kualitas dataset sebelum proses



Gambar 1. Metode Penelitian

Diagram ini menunjukkan tahapan penelitian yang meliputi pengumpulan dataset, pra-pengolahan data, pembagian data latih dan data uji, proses klasifikasi menggunakan algoritma Decision Tree, serta evaluasi kinerja model.

pemodelan. Nilai nol pada beberapa variabel klinis yang tidak memungkinkan secara medis dianggap sebagai data tidak valid dan ditangani menggunakan metode imputasi. Selanjutnya, dataset dibagi menjadi data latih dan data uji dengan rasio 70% dan 30%.

2.2. Metode Klasifikasi

Algoritma Decision Tree digunakan sebagai metode klasifikasi dalam penelitian ini. Algoritma ini membentuk struktur pohon keputusan berdasarkan atribut yang memiliki kemampuan terbaik dalam memisahkan data. Kriteria pemisahan yang digunakan adalah Gini Index untuk menentukan node terbaik pada setiap tingkat pohon.

2.3. Evaluasi Model

Evaluasi kinerja model dilakukan menggunakan confusion matrix serta metrik evaluasi berupa akurasi, precision, recall, dan F1-score. Metrik ini digunakan untuk menilai kemampuan model dalam mengklasifikasikan data pasien secara tepat.

Variabel	Rata-rata (Mean)	Standar Deviasi	Min	Max
Jumlah Kehamilan	3.84	3.37	0	17
Kadar Glukosa Darah	120.89	31.97	0	199
Tekanan Darah	69.10	19.35	0	122
Ketebalan Kulit	20.53	15.95	0	99
Kadar Insulin	79.79	115.24	0	846
Indeks Massa Tubuh	31.99	7.88	0	67.1
Riwayat Diabetes Keluarga	0.47	0.33	0.078	2.42
Usia	33.24	11.76	21	81

Gambar 2. Statistik deskriptif data pasien

Gambar ini menyajikan nilai minimum, maksimum, rata-rata, dan standar deviasi dari masing-masing variabel yang digunakan dalam penelitian.

Berdasarkan hasil statistik deskriptif, variabel kadar glukosa darah menunjukkan rentang nilai yang cukup luas, yang mencerminkan perbedaan kondisi metabolik pasien dalam dataset. Nilai indeks massa tubuh (BMI) juga memiliki variasi yang signifikan, mengindikasikan bahwa faktor obesitas berpotensi berperan dalam peningkatan risiko diabetes. Selain itu, variasi pada variabel usia dan riwayat diabetes dalam keluarga menunjukkan bahwa risiko diabetes tidak hanya dipengaruhi oleh faktor fisiologis, tetapi juga oleh faktor demografis dan genetik. Temuan ini memperkuat relevansi variabel yang digunakan dalam proses klasifikasi.

2.4. Parameter Model dan Implementasi

Model Decision Tree diimplementasikan menggunakan bahasa pemrograman Python dengan pustaka scikit-learn. Parameter yang digunakan antara lain `criterion='gini'`, tanpa

pembatasan kedalaman maksimum pohon untuk memperoleh struktur aturan yang optimal. Proses eksperimen dilakukan pada lingkungan komputasi standar dengan tujuan memperoleh hasil yang reproduktif.

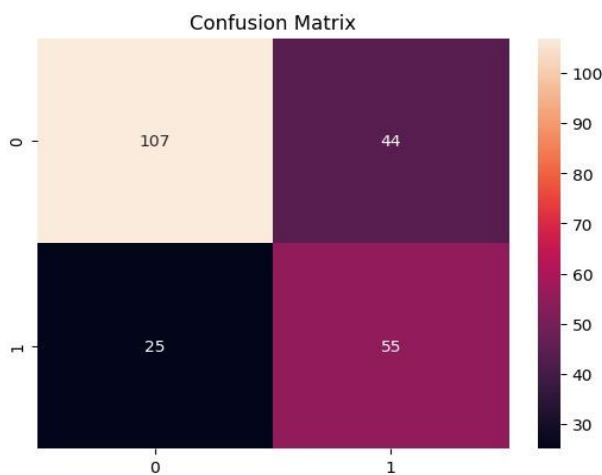
3. HASIL DAN PEMBAHASAN

3.1. Statistik Deskriptif Data

Analisis statistik deskriptif dilakukan untuk mengetahui karakteristik data pasien. Hasil analisis menunjukkan bahwa variabel kadar glukosa darah dan indeks massa tubuh memiliki variasi nilai yang cukup tinggi. Beberapa variabel juga mengandung nilai tidak valid yang telah ditangani pada tahap pra-pengolahan data.

3.2. Hasil Klasifikasi

Hasil pengujian model Decision Tree pada data uji menunjukkan tingkat akurasi sebesar 70%. Confusion matrix menunjukkan bahwa model memiliki kemampuan yang lebih baik dalam mengklasifikasikan pasien tidak diabetes dibandingkan pasien diabetes.



Gambar 1. Confusion matrix

Nilai false positive yang relatif tinggi menunjukkan bahwa model cenderung mengklasifikasikan beberapa pasien tidak diabetes sebagai diabetes. Kondisi ini dapat diterima dalam konteks deteksi dini karena lebih mengutamakan sensitivitas terhadap pasien berisiko dibandingkan kesalahan diagnosis negatif.

Akurasi: 0.7012987012987013

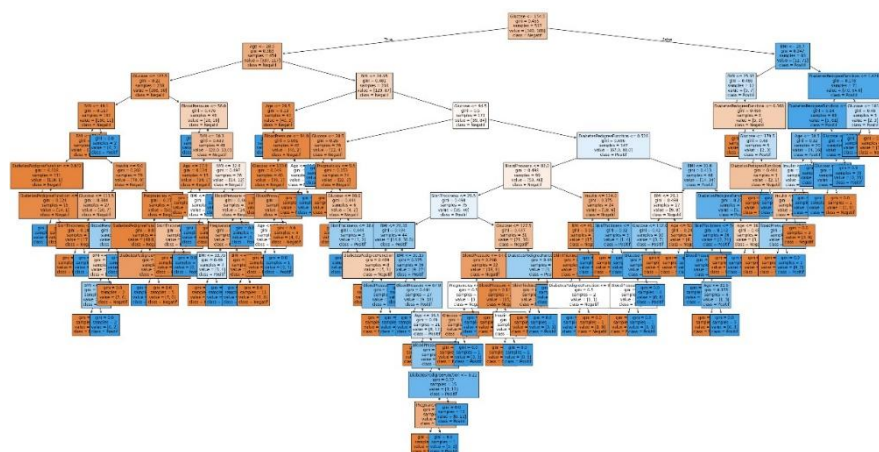
Laporan Klasifikasi:				
	precision	recall	f1-score	support
0	0.81	0.71	0.76	151
1	0.56	0.69	0.61	80
accuracy			0.70	231
macro avg	0.68	0.70	0.69	231
weighted avg	0.72	0.70	0.71	231

Gambar 2. Laporan klasifikasi

Gambar ini menampilkan nilai precision, recall, dan F1-score untuk masing-masing kelas sebagai indikator kinerja model.

3.3. Analisis Pohon Keputusan

Hasil visualisasi pohon keputusan menunjukkan bahwa variabel kadar glukosa darah menjadi faktor utama dalam menentukan klasifikasi risiko diabetes. Variabel indeks massa tubuh dan usia juga memiliki peran penting dalam pembentukan aturan keputusan.



Gambar 3. Visualisasi pohon keputusan

Gambar ini menunjukkan struktur pohon keputusan yang terbentuk berdasarkan variabel kesehatan pasien. Kadar glukosa darah menjadi node utama dalam proses pemisahan kelas, diikuti oleh indeks massa tubuh dan usia sebagai faktor pendukung.

3.4. Pembahasan

Hasil penelitian menunjukkan bahwa algoritma Decision Tree mampu menghasilkan model klasifikasi yang cukup baik serta mudah diinterpretasikan. Temuan bahwa kadar glukosa darah dan indeks massa tubuh merupakan faktor dominan sejalan dengan hasil penelitian sebelumnya. Meskipun demikian, tingkat akurasi model masih dapat ditingkatkan melalui penggunaan metode lain atau penambahan variabel pendukung.

Penelitian ini memiliki beberapa keterbatasan, antara lain penggunaan dataset sekunder dengan distribusi kelas yang tidak seimbang serta penerapan satu algoritma klasifikasi tanpa pembandingan. Selain itu, beberapa variabel penting terkait gaya hidup pasien belum tersedia dalam dataset yang digunakan.

4. SIMPULAN

Algoritma Decision Tree mampu digunakan untuk mengklasifikasikan risiko diabetes dengan tingkat akurasi sebesar 70%. Variabel kadar glukosa darah dan indeks massa tubuh merupakan faktor yang paling berpengaruh dalam menentukan risiko diabetes. Model yang dihasilkan bersifat interpretatif dan berpotensi digunakan sebagai alat bantu deteksi dini penyakit diabetes.

5. REKOMENDASI

Penelitian selanjutnya disarankan untuk menggunakan metode klasifikasi lain atau pendekatan ensemble guna meningkatkan kinerja model. Selain itu, penambahan variabel gaya hidup dan teknik penanganan ketidakseimbangan data dapat dipertimbangkan.

6. DAFTAR PUSTAKA

Bahri, S., Nugroho, A., & Budi, P., 2020. "Penerapan Algoritma Decision Tree untuk Klasifikasi Data Kesehatan". *Jurnal Teknologi Informasi dan Ilmu Komputer* 7(4): 63–70.

Elfaretta, F., Fadli, R., & Siregar, M., 2024. "Klasifikasi Calon Pendorong Darah Menggunakan Algoritma Decision Tree". *MALCOM: Indonesian Journal of Machine Learning and Computer Science* 4(1): 15–23.

Fadlu Rahman, A., Pratama, D., & Hidayat, R., 2024. "Klasifikasi Diagnosa Pasien Menggunakan Algoritma Decision Tree". *Jurnal Teknik Ibnu Sina* 9(2): 101–110.

Hana, F.M., 2020. "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5". *Jurnal Informatika dan Komputer* 5(1): 45–52.

Kaur, H. & Kumari, V., 2019. "Predictive Analysis of Diabetes Using Data Mining Techniques". *Procedia Computer Science* 132: 437–445.

Nuryamin, Y. & Risyda, F., 2025. "Analisis Prediksi Penyakit Diabetes Menggunakan Metode Decision Tree C4.5 dan Naive Bayes". *Jurnal Sistem Informasi dan Komputasi* 5(1): 234–242.

Safitri, L. & Fatah, Z., 2023. "Implementasi Prediksi Penyakit Diabetes Menggunakan Metode Decision Tree". *Jurnal Ilmu Komputer dan Sistem Informasi* 2(2): 125–132.

Siregar, M., Putra, A., & Lubis, R., 2023. "Penerapan Algoritma Decision Tree C4.5 dalam Prediksi Diabetes Mellitus". *Jurnal Sistem Komputer dan Kecerdasan Buatan* 6(1): 55–63.

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". *Proceedings of the Annual Symposium on Computer Application in Medical Care*: 261–265.

World Health Organization, 2022. *Global Report on Diabetes*. Geneva: World Health Organization.