Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma KNN

Zaehol Fatah[1,] Asvic Baid Anbala[2*]
[1]Zaehol Fatah, [2]Prodi Teknologi Informasi, fak. SAINTEK, Universitas Ibrahimy, Situbondo, Jawa Timur
Email: [1]zaeholfatah@gmail.com, [2]*asvicbaidanbala@gmail.com

Abstrak

Penyakit diabetes mellitus merupakan salah satu Deabetes lronis meningkat setiap tahun di banyak negara, termasuk Indonesia. Sangat penting untuk mendeteksi diabetes dini untuk mencegah komplikasi serius dan meningkatkan kualitas hidup pasien. Berdasarkan data medis pasien diabetes, penelitian ini menggunakan algoritma K-NN untuk mengklasifikasikan pasien diabetes. kumpulan data csv. Penelitian meliputi tahapan preprocessing data seperti penanganan nilai hilang, normalisasi, pembagian data (split data), serta implementasi algoritma K-NN pada RapidMiner Studio. Berdasarkan hasil pengujian dengan nilai k tertentu, diperoleh akurasi sebesar 71,43%, weighted mean recall sebesar 66,50%, dan weighted mean precision sebesar 68,41%. Hasil Ini menunjukkan bahwa algoritma K-NN dapat mengklasifikasikan dengan baik. data penderita diabetes, meskipun masih terdapat ruang untuk peningkatan performa melalui pengaturan parameter k dan metode optimasi jarak.

Kata Kunci: Klasifikasi, Diabetes Mellitus, K-Nearest Neighbor, Data Mining, RapidMiner

Abstrack

Diabetes mellitus is a long-term condition that is becoming prevalent across the world, especially in Indonesia. To avoid serious problems and enhance patients' quality of life, early identification of possible diabetes is essential. The K-Nearest Neighbor (K-NN) method was used in this investigation. was used to classify diabetes patients based on medical data obtained from the diabetes.csv dataset. The research process included data preprocessing steps such as handling missing values, normalization, data splitting, and implementing the K-NN algorithm using RapidMiner Studio. Based on the test results with a specific k value, the obtained accuracy was 71.43%, weighted mean recall 66.50%, and weighted mean precision 68.41%. These findings show that the K-NN algorithm does a respectable job at identifying diabetic data; however, there is still an opportunity for improvement: parameter tuning and distance optimization methods.

Keywords: Klasifikasi, Diabetes Mellitus, K-Nearest Neighbor, Data Mining, RapidMiner

PENDAHULUAN

Kemajuan dalam sistem informasi kesehatan telah mendorong adopsi Rekaman Medis Elektronik (RME) atau EMR sebagai sarana utama dalam penyimpanan dan pengelolaan data pasien secara digital. Penerapan RME memungkinkan informasi kesehatan terdokumentasi secara lebih terstruktur, terintegrasi, dan mudah diakses oleh tenaga medis, sehingga dapat meningkatkan mutu pelayanan kesehatan sekaligus mengurangi potensi terjadinya kesalahan diagnostik(Fatah 2025). Selain berperan sebagai basis data digital pasien, RME juga menyediakan himpunan data berukuran besar yang dapat diolah untuk berbagai keperluan analitis, seperti identifikasi pola penyakit, peramalan tingkat risiko, dan penilaian efektivitas penanganan medis. Ketika dikombinasikan dengan metode data mining, RME menjadi sumber data yang strategis dalam pengembangan model prediktif berbasis algoritma kecerdasan

buatan, khususnya untuk mendukung deteksi dini penyakit kronis seperti diabetes mellitus(Indian and Non-reservation 2018).

Perkembangan teknologi informasi telah membawa kemajuan pesat dalam pengolahan dan analisis data medis(Segara, Irwan, and Nasution 2025). Salah satu penyakit yang menjadi fokus penelitian dalam bidang data mining adalah diabetes mellitus, yaitu penyakit metabolik kronis Hal ini disebabkan oleh gangguan produksi atau penggunaan insulin. data dari WHO(Ii 2023), Di seluruh dunia, jumlah penderita diabetes terus berkembang setiap tahun dan diproyeksikan akan mencapai lebih dari 853 juta orang pada tahun 2050(Federasi Diabetes Internasional (IDF) 2025).

Kondisi ini menuntut adanya sistem pendukung keputusan yang mampu melakukan deteksi dini terhadap risiko diabetes secara cepat dan akurat(Fitriyadi 2025). Algoritma K-NN, algoritma klasifikasi berbasis jarak, adalah salah satu metode yang dapat digunakan. sederhana namun efektif dalam memetakan pola data(Bakri et al. 2025).

Dengan menggunakan data medis sebagai tekanan darah, glukosa, umur, dan indeks massa tubuh,dan variabel lainnya, algoritma K-NN dapat digunakan untuk memprediksi apakah seseorang berpotensi menderita diabetes atau tidak. Penelitian ini dilakukan untuk mengimplementasikan algoritma K-NN menggunakan perangkat lunak RapidMiner Studio dan menganalisis tingkat akurasi serta performa algoritma dalam klasifikasi penderita penyakit diabetes berdasarkan dataset yang tersedia(Ardianto and Rushendra 2025).

Rumusan masalah dalam penelitian ini berfokus pada bagaimana proses klasifikasi penderita penyakit diabetes dapat dilakukan menggunakan algoritma k-Nearest Neighbors (K-NN) pada platform RapidMiner Studio, serta sejauh mana tingkat akurasi dan performa algoritma tersebut dalam mengidentifikasi kondisi diabetes berdasarkan dataset yang digunakan. Penelitian ini juga mempertanyakan efektivitas K-NN dalam menghasilkan prediksi yang reliabel, sehingga diperlukan pengujian menyeluruh terhadap akurasi, presisi, dan tingkat recall dari model yang dibangun.

Sejalan dengan rumusan masalah tersebut, tujuan penelitian ini adalah untuk mengimplementasikan algoritma K-NN dalam proses klasifikasi penderita penyakit diabetes serta menganalisis performa model melalui evaluasi akurasi, presisi, dan recall yang dihasilkan dari penerapannya pada RapidMiner Studio. Analisis ini diharapkan mampu memberikan gambaran empiris mengenai kemampuan algoritma K-NN dalam memodelkan data kesehatan, khususnya dalam mendukung proses diagnosis awal diabetes.

Banyak pihak mendapatkan manfaat besar dari penelitian ini. Peneliti menggunakan penelitian ini untuk memperluas pemahaman mereka dan meningkatkan keterampilan mereka dalam menggunakan algoritma data mining dalam proses klasifikasi data kesehatan. Hasil kajian ini dapat membantu masyarakat memahami pentingnya pemanfaatan teknologi analisis data untuk deteksi dini penyakit, terutama penyakit kronis. Sementara itu, bagi pengembang sistem, penelitian ini dapat menjadi dasar konseptual dan teknis dalam perancangan serta pengembangan aplikasi prediksi diabetes berbasis kecerdasan buatan yang lebih efektif dan akurat.

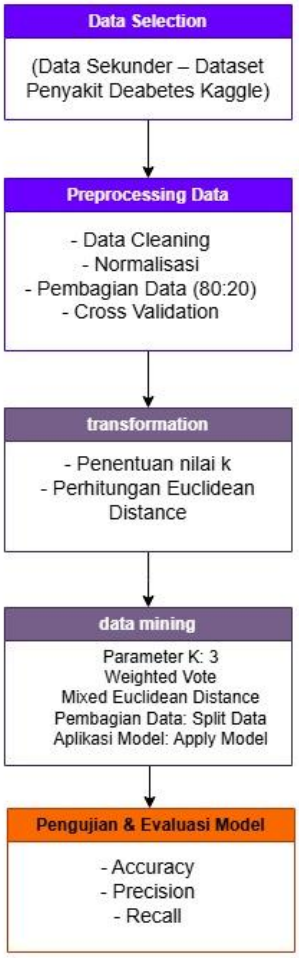
METODE

Metode Analisa Penelitian

Analisis deskriptif kuantitatif dengan pendekatan data mining yang berbasis klasifikasi penderita penyakit diabetes. Pendekatan ini bertujuan untuk mengidentifikasi pola hubungan antara atribut medis dengan status diabetes pasien berdasarkan dataset yang digunakan(Arrohman and Fatah 2024).

Dalam penelitian ini, proses penemuan pengetahuan dalam basis data (Knowledge Discovery in Databases / KDD) diimplementasikan melalui lima tahapan utama yang disesuaikan dengan

alur pemodelan pada RapidMiner Studio. Dataset yang digunakan merupakan dataset diabetes.csv yang diunduh dari platform Kaggle, berisi atribut klinis seperti kadar glukosa, tekanan darah, usia, jumlah kehamilan, indeks massa tubuh (BMI), serta variabel diagnosis diabetes. Dataset tersebut selanjutnya diproses melalui gambar tahapan penelitian berikut.



Gambar 1. Tahapan Penelitian

Penelitian ini menggunakan data pasien dengan sejumlah atribut yang berperan sebagai variabel input (independen) dan satu variabel output (dependen).

Tabel 1: Variabel Penelitian

No	Nama Variabel	Keterangan	Jenis Data
1	Pregnancies	Jumlah kehamilan (khusus untuk pasien wanita)	Numerik
2	Glucose	Konsentrasi glukosa plasma setelah 2 jam (mg/dL)	Numerik
3	BloodPressure	Tekanan darah diastolik (mm Hg)	Numerik
4	SkinThickness	Ketebalan lipatan kulit (mm)	Numerik
5	Insulin	Kadar insulin serum (µU/ml)	Numerik
6	BMI	Indeks massa tubuh (kg/m²)	Numerik
7	Diabetes Pedigree Function	Nilai fungsi riwayat diabetes keluarga	Numerik
8	Age	Usia pasien (tahun)	Numerik
9	Outcome	Status diabetes (1 = Penderita, 0 = Tidak)	Kategorikal

Metode Pengembangan Sistem Studi ini menggunakan metode Waterfall, yaitu metode pengembangan sistem yang bersifat berurutan dan terstruktur. Metode ini dipilih karena sederhana serta sesuai untuk sistem dengan kebutuhan yang telah ditentukan sejak awal. Tahapan dalam metode ini meliputi analisis kebutuhan untuk mengidentifikasi data dan sistem yang digunakan, perancangan alur proses serta model K-Nearest Neighbor (KNN), implementasi sistem di RapidMiner Studio, pengujian menggunakan Confusion Matrix untuk menilai akurasi model, serta pemeliharaan sistem agar tetap optimal. Dengan metode ini, proses pengembangan dapat berjalan sistematis dan menghasilkan model klasifikasi yang akurat serta mudah dikembangkan.

HASIL DAN PEMBAHASAN

Berdasarkan hasil pengujian seuai dengan tahapan penelitian yaitu:

1. Data Selection (Pemilihan Data)

Tahap awal dilakukan dengan membaca dataset diabetes.csv menggunakan operator Read CSV pada RapidMiner. Pada tahap ini dilakukan

pemilihan atribut yang relevan untuk proses klasifikasi, yaitu variabel-variabel medis yang memiliki kontribusi terhadap penentuan status diabetes. Selanjutnya, operator Set Role digunakan untuk menetapkan atribut label, yakni variabel "Outcome", sebagai target klasifikasi.

2. Preprocessing (Pembersihan Data)

Pada tahap preprocessing dilakukan pembersihan data untuk memastikan kualitas dataset sebelum digunakan dalam pemodelan. Proses ini mencakup penanganan missing value menggunakan operator Replace Missing Values, sehingga tidak terdapat nilai kosong yang berpotensi mengganggu proses pembelajaran model. Pembersihan ini penting untuk menjaga konsistensi dan meminimalkan distorsi pada hasil klasifikasi.

3. Transformation (Transformasi Data)

Setelah data dibersihkan, tahap selanjutnya adalah transformasi data melalui proses Normalisasi menggunakan operator Normalize. Transformasi ini bertujuan untuk menyeragamkan skala seluruh atribut numerik agar perhitungan jarak pada algoritma K-NN tidak didominasi oleh variabel dengan skala besar. Normalisasi menjadi langkah krusial mengingat algoritma K-NN mengandalkan perhitungan distance measure untuk menentukan kedekatan antar titik data.

4. Data Mining (Pemodelan Data dengan K-NN)

Pada tahap pemodelan, algoritma k-Nearest Neighbors (K-NN) diterapkan menggunakan operator k-NN pada RapidMiner. Berdasarkan parameter yang digunakan dalam proses, nilai k ditetapkan sebesar 3, dengan metode pemungutan suara (weighted vote) serta distance measure tipe Mixed Euclidean Distance. Dataset sebelumnya dibagi menjadi dua bagian menggunakan operator Split Data, yakni data latih dan data uji. Model yang terbentuk kemudian diaplikasikan melalui operator Apply Model untuk menghasilkan prediksi status diabetes.

5. Evaluation (Evaluasi Model)

Tahap evaluasi model dilakukan menggunakan operator Performance, yang menghasilkan metrik evaluasi meliputi accuracy, precision, dan recall, serta tampilan confusion matrix.

Dan implikasinya Menghasilkan algoritma K-NN dengan nilai yang diperoleh nilai sebagai berikut:

- Accuracy : 71.43%
- Weighted Mean Recall : 66.50%
- Weighted Mean Precision : 68.41%

Dan juga diperoleh hasil evaluasi kinerja(Teknologi et al. 2024) seperti yang ditunjukkan pada tabel confusion matrix berikut:

Tabel 2: Confusion Matrix

Prediksi / Aktual	True 1 (Positif Diabetes)	True 0 (Negatif Diabetes)
Pred. 1 (Positif)	27	17
Pred. 0 (Negatif)	27	83

Tabel di atas menjelaskan bahwa, data positif yang diprediksi benar positif (TP)=27, data negatif yang diprediksi salah negatif (FP)=17, dan data negatif yang diprediksi benar negatif (TN)=83.

Hasil

Dari hasil yang telah di proses, penjelasannya sebagai berikut:

1. Akurasi

Dengan nilai akurasi 71.43%, model K-NN dapat mengklasifikasikan data dengan tepat. sebesar 71.43% dari total data uji. Dengan kata lain, dari 154 data uji, sekitar 110 data diprediksi dengan benar.

2. Precision

Nilai rata-rata tertimbang (weighted mean precision) sebesar 68.41% berarti bahwa dari seluruh data yang diprediksi sebagai "penderita diabetes", sekitar 68% merupakan prediksi yang benar. Precision yang tidak terlalu tinggi menunjukkan bahwa masih banyak data negatif yang diklasifikasikan salah sebagai positif (false positive).(Desmita, Lonang, and Kumoro 2025).

3. Recall

Nilai weighted mean recall sebesar 66.50% mengindikasikan bahwa model berhasil mendeteksi sekitar 66% dari seluruh data penderita diabetes yang sebenarnya positif. Ini menunjukkan adanya beberapa kasus penderita diabetes yang tidak terdeteksi (false negative).

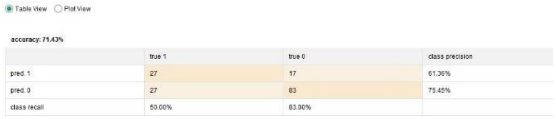
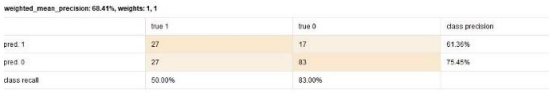


Table View Plot View

accuracy: 71.43%

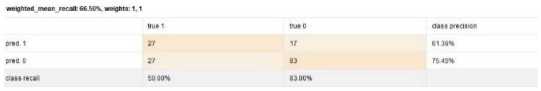
	true 1	true 0	class precision
pred. 1	27	17	61.36%
pred. 0	27	83	75.45%
class recall	50.00%	83.00%	

Gambar 3: Hasil Akurasi


weighted_mean_precision: 68.41%, weights: 1, 1

	true 1	true 0	class precision
pred. 1	27	17	61.36%
pred. 0	27	83	75.45%
class recall	50.00%	83.00%	

Gambar 4: Hasil Weighted Mean Precision


weighted_mean_recall: 66.50%, weights: 1, 1

	true 1	true 0	class precision
pred. 1	27	17	61.36%
pred. 0	27	83	75.45%
class recall	50.00%	83.00%	

Gambar 5: Hasil Weighted Mean Recall

Hasil pengujian menunjukkan bahwa algoritma KNN sangat akurat. 71.43%, yang dapat dikategorikan cukup baik untuk kasus klasifikasi dua kelas (penderita dan tidak penderita diabetes).

Untuk meningkatkan hasil klasifikasi, penelitian lanjutan dapat mempertimbangkan:

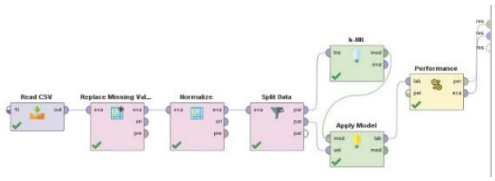
1. Optimasi nilai parameter k menggunakan metode cross-validation.
2. Menerapkan teknik feature selection untuk memilih atribut paling relevan.
3. Membandingkan kinerja K-NN dengan algoritma lain seperti Decision Tree atau Naïve Bayes.

Pembahasan

A. Implementasi Sistem

Pada tahap implementasi, sistem klasifikasi penderita penyakit diabetes dibangun menggunakan aplikasi RapidMiner Studio. Dataset yang digunakan merupakan data penderita diabetes yang diambil dari Kaggle dalam format CSV.

Langkah implementasi dilakukan melalui beberapa tahapan sebagai berikut:



Gambar 2 Perancnagan SItem Dengan RapidMiner

1. Read CSV

Tahap awal membaca dataset menggunakan operator Read CSV untuk memuat data ke dalam lingkungan RapidMiner.

2. Replace Missing Values

Dilakukan pembersihan data dengan mengganti nilai-nilai yang kosong (missing values) agar tidak mempengaruhi proses klasifikasi.

3. Normalize

Seluruh atribut numerik dinormalisasi ke dalam jangkauan nilai 0–1. penting karena algoritma K-NN sangat sensitif terhadap perbedaan skala data.

4. Split Data

Data pelatihan dan data pengujian membentuk 80% data set. digunakan untuk pelatihan dan 20% untuk pengujian.

5. K-NN (K-Nearest Neighbor)

Dengan nilai parameter k tertentu (misalnya k=5). Proses ini akan menghitung jarak antara data latih dan data uji dengan menggunakan metrik jarak geometris.

6. Apply Model

Model hasil training diterapkan ke data testing untuk menghasilkan prediksi terhadap status diabetes.

7. Performance Evaluation

Evaluasi kinerja dilakukan dengan operator Performance (Classification) untuk memperoleh nilai accuracy, precision, dan recall serta confusion matrix.

SIMPULAN (PENUTUP)

Dengan tingkat akurasi sebesar 71,43%, precision sebesar 68,41%, dan recall sebesar 66,50%, hasil implementasi dan pengujian sistem klasifikasi penderita diabetes aplikasi algoritma KNN mengkonfirmasi bahwa algoritma ini cukup

efisien dalam mengenali pola data medis dan membedakan antara orang yang menderita diabetes dan orang yang tidak.

Implementasi sistem ini membuktikan bahwa K-NN dapat digunakan sebagai metode pendukung dalam proses analisis data kesehatan, khususnya dalam membantu identifikasi awal risiko diabetes berdasarkan parameter seperti kadar glukosa, tekanan darah, usia, dan indeks massa tubuh(Informatika et al. 2022).

Untuk pengembangan di masa mendatang, penelitian ini dapat ditingkatkan dengan melakukan optimasi nilai k, penerapan metode normalisasi dan seleksi fitur yang lebih baik, serta perbandingan dengan algoritma lain seperti Tree of Choice, Naive Bayes, atau Support Vector Machine (SVM) agar diperoleh hasil klasifikasi yang lebih akurat dan efisien.

UCAPAN TERIMA KASIH

Penulis menyampaikan apresiasi yang mendalam kepada orang tua atas doa dan dukungan berkelanjutan, kepada dosen pembimbing atas kontribusi pemikiran, arahan, dan pendampingan akademik, serta kepada rekan-rekan seperjuangan atas sinergi dan semangat kolektif selama proses penyelesaian karya ini.

DAFTAR PUSTAKA

Ardianto, Muhammad Rezanur, and Rushendra Rushendra. 2025. “Prediksi Penyakit Diabetes Berdasarkan Perbandingan Klasifikasi Metode K-Nearest Neighbor, Naïve Bayes, Dan Decision Tree Menggunakan Rapid Miner.” 10(2):973–85.

Arrohman, Supri, and Zaehol Fatah. 2024. “Gudang Jurnal Multidisiplin Ilmu Prediksi Diabetes Menggunakan Algoritma Klasifikasi K-Nearest Neighbors (K-NN) Pada Perempuan Indian Pima.” 2:220–26.

Bakri, Safa Nadia, Lailan Sofinah Harahap, Universitas Islam, Negeri Sumatera, Teknologi Informasi, Universitas Muhammadiyah, Sumatera Utara, Kota Medan, Struktur Daerah, and Kota Medan. 2025. “Analisis Klasifikasi Algoritma K-Nearest Neighboar (K-NN) Pada Struktur Daerah Di Kota Medan.” 182–93.

Desmita, Nindri Lia, Syahrani Lonang, and Danang Tejo Kumoro. 2025. “COMPARATIVE ANALYSIS OF DECISION TREE AND RANDOM FOREST ALGORITHMS FOR PREDICTING DIABETES MELLITUS.” 8.

Fatah, Zaehol. 2025. *TIK Dan Masyarakat*. PT Penamuda Media. Federasi Diabetes Internasional (IDF). 2025. “Fakta Dan Angka Diabetes Menunjukkan Meningkatnya Beban Diabetes Global Bagi Individu, Keluarga, Dan Negara. Atlas Diabetes Terbaru Dari Federasi Diabetes Internasional (IDF) (2025).”

Fitriyadi, Farid. 2025. “Optimasi Integrasi Adaptif Metode Naive Bayes Dan Information Gain Untuk Prediksi Komplikasi Diabetes Adaptive Integration Optimization of Naïve Bayes and Information Gain for Diabetes Complication Prediction.” 14:2990–3006.

Ii, B. A. B. 2023. “No Title.” (Dm).Indian, American, and Reservation Versus Non-reservation. 2018. “HHS Public Access.” 33(1):102–9. doi: 10.1111/jrh.12178.Assessing.Informatika, Jurnal, Dan Rekayasa, Komputer Jakakom, Asih Asmarani,

M. Ilham Permana, Annisa Putri, M. Rizky Wijaya, Errissya Rasywir, Despita Meisak Yovi, and Asih Asmarani. 2022. “Implementasi Algoritma K-Nearst Neighbor Untuk Memprediksi Penyakit Diabetes Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM).” 2(September):231–39.

Segara, Khoirunnisya Gita, Muhammad Irwan, and Padli Nasution. 2025. “Perkembangan Teknologi Informasi Di Indonesia : Tantangan Dan Peluang.” 3(1).

Simorangkir, Anastasya. 2023. "SPESIFIKASI KEBUTUHAN PERANGKAT LUNAK." (October). Teknologi, Sains, Universitas Ibrahimy, Sains Teknologi, and Universitas Ibrahimy. 2024. "Gudang Jurnal Multidisiplin Ilmu Implementasi Metode K-Nearest Neighbor (K-NN) Pada Klasifikasi Stunting Balita." 2:282–88.