JRIIN: Jurnal Riset Informatika dan Inovasi
Volume 2, No. 10 Maret 2025
ISSN 3025-0919 (media online)
Hal 1847-1852

IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBORS (KNN) UNTUK PREDIKSI PENYAKIT DIABETES PADA PEREMPUAN

Maulana Fansyuri[1], Dian Nurul Iman[2], Gideon Triman Harefa[3], Shahrudin[4], Arijal Pratama[5], Muhammad Rizki Rahmatullah[6]

[123456]Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia

Email: [1]iyann.zen@gmail.com

Abstrak--Diabetes melitus adalah penyakit kronis yang dapat menyebabkan komplikasi serius seperti gagal ginjal, kebutaan, dan penyakit jantung. Deteksi dini risiko diabetes sangat penting untuk mencegah komplikasi dan meningkatkan kualitas hidup pasien. Penelitian ini menggunakan dataset dari Kaggle yang terdiri dari 768 data dengan 9 atribut, termasuk kadar glukosa dan BMI. Algoritma K-Nearest Neighbor (KNN) dipilih karena kesederhanaan dan efektivitasnya dalam prediksi. Proses penelitian dilakukan menggunakan RapidMiner untuk preprocessing, pemodelan, dan evaluasi, dengan pembagian dataset sebesar 80% untuk pelatihan dan 20% untuk pengujian. Hasil evaluasi menunjukkan akurasi model mencapai 80,73%, dengan precision 74,79% dan recall 67,54% untuk kelas positif. Meskipun akurasi yang tinggi, rendahnya nilai recall menunjukkan bahwa beberapa kasus diabetes tidak terdeteksi, sehingga memerlukan optimasi lebih lanjut. Penelitian ini menyoroti potensi KNN dalam deteksi dini diabetes sebagai alat bantu keputusan bagi tenaga medis. Rekomendasi untuk meningkatkan performa model termasuk penggunaan dataset yang lebih besar dan penerapan teknik pembelajaran mesin yang lebih canggih. Diharapkan hasil penelitian ini dapat berkontribusi pada pengembangan teknologi prediktif di bidang kesehatan.

Kata Kunci: Data Mining, KNN, Prediksi Diabetes, RapidMiner.

Abstract--Diabetes mellitus is a chronic disease that can cause serious complications such as kidney failure, blindness and heart disease. Early detection of diabetes risk is very important to prevent complications and improve the patient's quality of life. This research uses a dataset from Kaggle which consists of 768 data with 9 attributes, including glucose levels and BMI. The K-Nearest Neighbor (KNN) algorithm was chosen because of its simplicity and effectiveness in prediction. The research process was carried out using RapidMiner for preprocessing, modeling and evaluation, with a dataset divided by 80% for training and 20% for testing. The evaluation results show that the model accuracy reached 80.73%, with a precision of 74.79% and a recall of 67.54% for the positive class. Despite the high accuracy, the low recall value indicates that some cases of diabetes are not detected, requiring further optimization. This research highlights the potential of KNN in early diabetes detection as a decision aid for medical personnel. Recommendations for improving model performance include the use of larger datasets and the application of more advanced machine learning techniques. It is hoped that the results of this research can contribute to the development of predictive technology in the health sector.

Keywords: Data Mining, KNN, Diabetes Prediction, RapidMiner.

1. PENDAHULUAN

Diabetes melitus merupakan penyakit metabolik yang ditandai oleh peningkatan kadar glukosa darah akibat gangguan fungsi insulin. Penyakit ini terbagi menjadi dua jenis utama, yaitu diabetes tipe 1 yang disebabkan oleh kelainan autoimun, dan diabetes tipe 2 yang terkait dengan resistensi insulin dan gaya hidup yang tidak sehat. Penyakit ini dapat menyebabkan komplikasi serius seperti kerusakan ginjal, kebutaan, amputasi anggota tubuh, hingga kematian (American Diabetes Association, 2020)

Perubahan gaya hidup masyarakat modern, seperti pola makan yang tidak sehat, kurangnya aktivitas fisik, dan tingkat stres yang tinggi, berkontribusi terhadap peningkatan prevalensi diabetes, terutama di negara berkembang seperti Indonesia. Berdasarkan data dari World Health Organization (WHO), Indonesia berada di peringkat kelima dunia dengan jumlah penderita diabetes terbanyak,

JRIIN: Jurnal Riset Informatika dan Inovasi
Volume 2, No. 10 Maret 2025
ISSN 3025-0919 (media online)
Hal 1847-1852

yang mencapai sekitar 8,3 juta orang pada tahun 2021 (WHO, 2021). Data ini menunjukkan pentingnya upaya deteksi dini untuk mencegah komplikasi diabetes yang lebih parah.

Teknologi berbasis data dan algoritma prediksi, seperti algoritma K-Nearest Neighbor (KNN), dapat menjadi solusi untuk memprediksi risiko diabetes sejak dini. KNN adalah algoritma klasifikasi yang sederhana namun efektif, yang mengandalkan data historis untuk memprediksi kelas dari data uji. Dalam konteks ini, KNN dapat digunakan untuk memprediksi kemungkinan seseorang mengidap diabetes berdasarkan sejumlah faktor risiko yang ada.

2. METODE

Penelitian ini menggunakan pendekatan metodologis yang terdiri dari beberapa tahapan sistematis untuk memprediksi risiko diabetes pada perempuan. Tahap awal adalah observasi untuk mengidentifikasi faktor-faktor risiko yang mempengaruhi perkembangan diabetes, termasuk aspek biologis, lingkungan, dan gaya hidup (American Diabetes Association, 2020). Setelah itu, data dikumpulkan dari dataset "Diabetes Dataset" di Kaggle, yang terdiri dari 768 baris dan 9 kolom informasi penting (Kaggle, 2023). Proses selanjutnya adalah preprocessing data untuk meningkatkan kualitas data melalui pembersihan, penanganan data hilang, dan normalisasi. Setelah data siap, algoritma K-Nearest Neighbor (KNN) diterapkan untuk menganalisis dan memprediksi status diabetes, dengan membagi data menjadi 80% untuk pelatihan dan 20% untuk pengujian. Model dievaluasi berdasarkan akurasi, precision, dan recall untuk menilai performanya dalam mendeteksi diabetes (Asmarani, 2022; WHO, 2021).

2.1 Data Mining

Data mining adalah proses analisis data yang bertujuan untuk menemukan pola, tren, dan informasi berharga dari kumpulan data besar. Dengan menggunakan teknik statistik, algoritma pembelajaran mesin, dan metode analisis data, data mining memungkinkan pengambilan keputusan yang lebih baik dan lebih cepat berdasarkan data yang ada (Han J. Kamber M. & Pei J, 2011). Proses ini melibatkan beberapa langkah, termasuk pengumpulan data, preprocessing, eksplorasi, dan penerapan algoritma untuk menemukan informasi yang relevan (Kargupta H. & Joshi A, 2000).

Dalam konteks kesehatan, data mining merupakan alat yang sangat berharga dengan berbagai penerapan yang signifikan. Salah satu penggunaan utamanya adalah untuk memprediksi kemungkinan terjadinya penyakit dengan menganalisis data pasien (Delen D. Walker G. & Kadam A, 2013). eknik klasifikasi dan regresi memungkinkan profesional kesehatan mengidentifikasi individu yang berisiko tinggi terkena penyakit tertentu, seperti diabetes atau penyakit jantung (Sharma S. & Gupta S, 2016). Selain itu, data mining membantu dalam analisis faktor risiko, mengungkapkan hubungan antara gaya hidup, riwayat kesehatan keluarga, dan kondisi kesehatan saat ini (Raghupathi W. & Raghupathi V., 2014). Teknik clustering juga memungkinkan pengelompokan pasien berdasarkan kesamaan karakteristik, sehingga perawatan dapat disesuaikan dengan kebutuhan spesifik masing-masing individu (Kaufman, 1990).

2.2 Algoritma K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) adalah algoritma klasifikasi yang tidak memerlukan model atau pelatihan data secara eksplisit. KNN bekerja dengan cara mencari "tetangga" atau data yang serupa dari data uji dan mengklasifikasikan data berdasarkan mayoritas kelas dari tetangga terdekat tersebut (Cover, 1967). Setiap data dalam dataset diwakili sebagai titik dalam ruang multidimensi, dan jarak antar titik dihitung menggunakan rumus Euclidean atau metrik lainnya (Zhang, 2016). Nilai K dalam KNN merujuk pada jumlah tetangga yang diperhitungkan dalam klasifikasi, di mana pemilihan nilai K dapat mempengaruhi performa model secara signifikan (Keller J. M. Givens J. R. & Gray M. L, 1985).

Rumus jarak Euclidean yang digunakan dalam perhitungan adalah:

JRIIN: Jurnal Riset Informatika dan Inovasi
Volume 2, No. 10 Maret 2025
ISSN 3025-0919 (media online)
Hal 1847-1852

$$d = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2}$$

Keterangan:

- d: Jarak antara dua data
- x_1, x_2 : Data uji dan data pelatihan
- p: Dimensi data

2.3 Rapid Miner

RapidMiner merupakan salah satu platform perangkat lunak yang populer dalam analisis data dan penerapan data mining, yang menawarkan berbagai alat untuk melakukan eksplorasi data, pemodelan, dan evaluasi. Dalam konteks implementasi algoritma K-Nearest Neighbors (KNN) untuk prediksi penyakit diabetes pada perempuan, RapidMiner menyediakan antarmuka yang intuitif dan user-friendly, memungkinkan pengguna untuk mengakses dan memproses data tanpa memerlukan keterampilan pemrograman yang mendalam (Asmarani, 2022).

Platform ini dilengkapi dengan berbagai fitur, termasuk preprocessing data, pemilihan fitur, dan visualisasi hasil, yang sangat penting dalam tahap persiapan data untuk analisis (Han J. Kamber M. & Pei J, 2011). RapidMiner mendukung berbagai jenis algoritma pembelajaran mesin, termasuk KNN, yang dapat diimplementasikan dengan mudah melalui drag-and-drop interface. Dengan menggunakan RapidMiner, peneliti dapat dengan cepat membangun model prediksi, melakukan validasi, serta menganalisis performa model untuk memperoleh wawasan yang lebih baik mengenai faktorfaktor yang mempengaruhi risiko diabetes pada perempuan (Zhang, 2016).

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Penelitian ini dilakukan melalui serangkaian tahapan yang dimulai dengan identifikasi masalah dan berlanjut pada implementasi algoritma KNN untuk memprediksi risiko diabetes berdasarkan dataset yang telah dikumpulkan. Tahapan penelitian secara rinci adalah sebagai berikut:

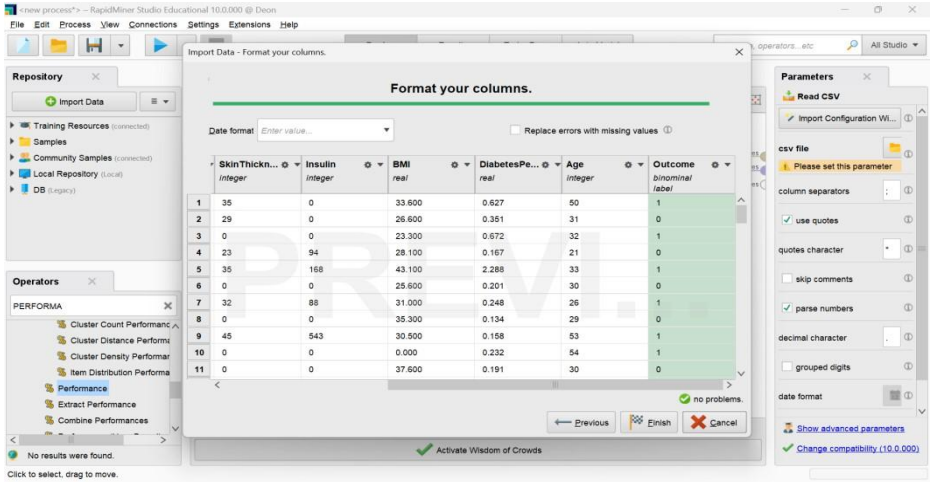
1. Observasi dan Identifikasi Masalah
   Penelitian ini dimulai dengan identifikasi faktor-faktor risiko yang dapat mempengaruhi perkembangan penyakit diabetes. Faktor-faktor ini meliputi aspek biologis, lingkungan, dan gaya hidup yang berkontribusi terhadap peningkatan kadar gula darah. Mengetahui faktor-faktor ini memungkinkan pemilihan teknik prediksi yang tepat, salah satunya adalah algoritma K-Nearest Neighbor (KNN) (Asmarani, 2022).
2. Pengumpulan Data
   Dataset yang digunakan dalam penelitian ini diperoleh dari situs web Kaggle, dengan judul "Diabetes Dataset". Dataset ini berisi 768 baris data dengan 9 kolom. Berikut adalah penjelasan singkat mengenai kolom-kolom yang terdapat dalam dataset tersebut:

Kolom	Deskripsi
Pregnancies	Jumlah kehamilan yang dialami individu. Mempengaruhi risiko diabetes gestasional (Delen D. Walker G. & Kadam A, 2013)

JRIIN: Jurnal Riset Informatika dan Inovasi
Volume 2, No. 10 Maret 2025
ISSN 3025-0919 (media online)
Hal 1847-1852

Glucose	Kadar glukosa dalam darah (mg/dL). Kadar di atas 200 mg/dL dapat menunjukkan diabetes (American Diabetes Association, 2020)
BloodPressure	Tekanan darah (mmHg). Hipertensi terkait dengan risiko diabetes (Han J. Kamber M. & Pei J, 2011)
SkinThickness	Ketebalan lipatan kulit. Menunjukkan lemak subkutan dan risiko obesitas (Keller J. M. Givens J. R. & Gray M. L, 1985)
Insulin	Kadar insulin dalam darah (µU/mL). Resistensi insulin sering terjadi pada diabetes tipe 2 (Zhang, 2016)
BMI	Indeks Massa Tubuh. BMI di atas 30 menunjukkan obesitas, faktor risiko diabetes (Raghupathi W. & Raghupathi V., 2014)
DiabetesPedigreeFunction	Nilai yang menunjukkan riwayat diabetes dalam keluarga. Menunjukkan pengaruh genetik (Kargupta H. & Joshi A, 2000)
Age	Usia individu dalam tahun. Usia yang lebih tua berhubungan dengan peningkatan risiko diabetes (Sharma S. & Gupta S, 2016)
Outcome	Variabel target yang menunjukkan apakah individu memiliki diabetes (1) atau tidak (0) (International Diabetes Federation, 2021)

3.2 Transformasi Data (Transformation)



Gambar 1. *Dataset*

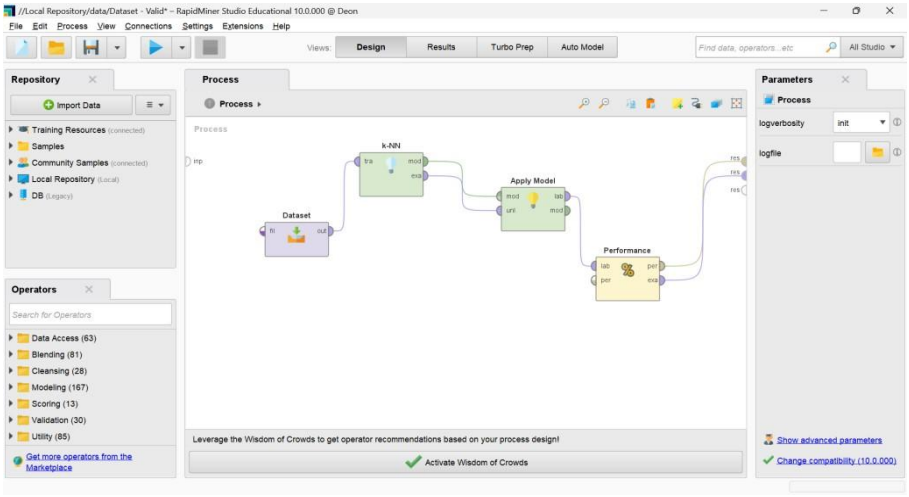
Pada Gambar 1 merupakan proses transformasi data yang akan diberikan label yang dibutuhkan algoritma K-Nearest Neighbor.

3.3 Implementasi Algoritma KNN

Data yang telah dikumpulkan dibagi menjadi dua bagian, yaitu 80% untuk data pelatihan dan 20% untuk data pengujian. Pada tahap ini, peneliti mengimplementasikan algoritma KNN

JRIIN: Jurnal Riset Informatika dan Inovasi
Volume 2, No. 10 Maret 2025
ISSN 3025-0919 (media online)
Hal 1847-1852

menggunakan Aplikasi RapidMiner untuk menganalisis data dan memprediksi status diabetes berdasarkan atribut yang ada. Berikut gambar hasil design di Aplikasi Rapid Miner :



Gambar 2. *Design*

Dari gambar diatas dapat dijelaskan secara singkat gambar tersebut :

a. Dataset berisi dataset yang digunakan dalam penelitina tersebut yang dimana berisi data “diabetes perempuan”
b. K-NN model yang digunakan untuk melakukan pengujian pada dateset tersebut yaitu “diabetes perempuan”
c. Apply Model berfungsi menerapkan model yang telah dilatih untuk memprediksi label dari data baru atau melakukan transformasi data dengan menjalankan model preprocesing.
d. Performance digunakan untuk mengukur kinerja model berdasarkan prediksi yang dilakukan.

3.4 Hasil Penelitian

Berdasarkan hasil analisis kami menggunakan dataset dengan judul “Diabetes Dataset” kami melakukan pengujian model dengan metode algoritma K-Nearest Neighbor (KNN) setelah melakukan proses pengolahan data. Tujuan dari penelitian ini untuk membantu dalam diagnosis dini diabetes, sehingga intervensi dapat dilakukan lebih cepat dan efektif dan juga untuk menilai akurasi analisis data pasien diabetes mellitus dan memprediksi apakah gejala pasien dapat menunjukkan apakah mereka positif atau negatif menderita diabetes. Beriku gambar hasil analisis kami dengan menggunakan aplikasi RapidMiner.

accuracy: 80.73%

	true 1	true 0	class precision
pred. 1	181	61	74.79%
pred. 0	87	439	83.46%
class recall	67.54%	87.80%	

Gambar 3. *Hasil*

Dari hasil gambar diatas dapat kita lihat akurasi model tersebut sebesar 80.73% dapat dikatakan model K-Nearest Neighbor (KNN) dianggap sebagi model yang cukup baik. Berdasarkan model tersebut berhasil mengidentifikasi 181 dinyatakan sebagai pasien yang benar benar menderita diabetes dan Model juga berhasil mengidentifikasi 439 dinyatakan sebagai pasien yang benar benar

JRIIN: Jurnal Riset Informatika dan Inovasi
Volume 2, No. 10 Maret 2025
ISSN 3025-0919 (media online)
Hal 1847-1852

tidak menderita diabetes. Dapat juga dilihat class recall 1 sebesar 67.54% dan class precision sebesar 74.79% sementara class recall 0 sebasar 87.80% dan class precision sebesar 83.46%.

4. KESIMPULAN

Analisis menggunakan model k-Nearest Neighbors (k-NN) untuk klasifikasi diabetes menunjukkan akurasi 80,73%, mencerminkan performa baik dalam deteksi. Namun, metrik precision dan recall mengungkapkan tantangan: model mengidentifikasi 181 pasien diabetes (True Positives - TP), tetapi terdapat 61 False Positives (FP) dan 87 False Negatives (FN). Precision untuk kelas positif adalah 74,79%, menandakan risiko kesalahan diagnosis, sementara recall hanya 67,54%, menunjukkan model belum efektif mendeteksi semua pasien diabetes.

Meskipun k-NN memiliki potensi sebagai alat diagnosis, perlu pengoptimalan untuk meningkatkan recall dan mengurangi pasien yang tidak terdeteksi. Penelitian ini juga menunjukkan bahwa algoritma KNN dapat memprediksi diabetes dengan akurasi 66,67%, yang masih memerlukan penelitian lebih lanjut dengan dataset lebih besar dan teknik pra-pemrosesan untuk meningkatkan kinerja. Hasil penelitian ini diharapkan dapat mendukung pengembangan teknologi prediksi diabetes yang lebih efektif dalam deteksi dini dan pencegahan penyakit.

REFERENCES

American Diabetes Association. (2020). *Classification and Diagnosis of Diabetes: Standars of Medical Care in Diabetes.* Diabetes Care, 43(Suppl 1), S14-S31.

Asmarani, A. e. (2022). *Implementasi Algoritma K-Nearest Neighbor Untuk Memprediksi Penyakit Diabetes.* JAKAKOM, Volume 2, Nomor 2.

Cover, T. (1967). *Nearest neighbor pattern classification.* IEEE Transactions on Information Theory, 13(1), 21-27. .

Delen D. Walker G. & Kadam A. (2013). *Predicting breast cancer survivability: A comparison of three data mining methods.* Artificial Intellegence in Medicine, 59(3), 169-184.

Han J. Kamber M. & Pei J. (2011). *Data Mining: Concepts and Techniques.* Morgan Kaufmann.

International Diabetes Federation. (2021). *"IDF Diabetes Atlas.".* Retrieved from [https://www.diabetesatlas.org].

Kaggle. (2023). *Diabetes Prediction Dataset.* etrieved from https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset. .

Kargupta H. & Joshi A. (2000). *Data Mining in health care* . Journal of Healthcare: Promise and potential. Information Management, 14(2), 44-54.

Kaufman, L. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley.

Keller J. M. Givens J. R. & Gray M. L. (1985). *A Fuzzy K-nearest neighbor algorithm.* IEEE Transactions on Systems, Man, and Cybernetics, 15(4), 580-585.

Raghupathi W. & Raghupathi V. (2014). *Big data analytics in healthcare: Promise and potential.* Helath Information Science and System, 2(1), 3.

Sharma S. & Gupta S. (2016). *Data Mining techniques for health care.* International Journal of Computer Applications, 139(8), 1-5.

WHO. (2021). *Global Diabetes Observatory. World Health Organization.* Retrieved from https://www.who.int/diabetes.

Zhang, Y. (2016). *K-Nearest Neighbor (KNN) Algorithm for Classification.* International Journal of Computer Applications, 140(5), 25-29. .