JOMLAI: Journal of Machine Learning and Artificial Intelligence
Vol. 5, No. 2, Juni 2026, pp. 49~56
ISSN: 2828-9102 (Print) | 2828-9099 (Online)
DOI: 10.55123/jomlai.v5i2.5429

` Model Prediksi Diabetes Berdasarkan BMI, HbA1c, Usia, dan Glukosa Darah Menggunakan KNN

A Diabetes Prediction Model Based on BMI, HbA1c, Age, and Blood Glucose Using KNN

Avelina Garcia Wong[1], Devin Hernando[2], Marcelyn Wijaya[3], Thomas Herpin[4], Vinola Lorencia[5], Ade Maulana[6*]
[1,2,3,4,5,6]Universitas Pelita Harapan Indonesia, Medan, Indonesia

Article Info

Genesis Artikel:

Diterima, 27 April 2026
Direvisi, 01 Juni 2026
Disetujui, 15 Juni 2026

Kata Kunci:

Prediksi
Machine Learning
Diabetes
K-Nearest Neighbors
KNN

ABSTRAK

Diabetes merupakan penyakit kronis yang prevalensinya terus meningkat secara global dan dapat menyebabkan komplikasi serius apabila tidak terdeteksi sejak dini. Penelitian ini bertujuan untuk membangun model klasifikasi yang sederhana namun akurat dalam memprediksi risiko diabetes menggunakan algoritma K-Nearest Neighbors (KNN). Model dikembangkan dengan memanfaatkan empat parameter klinis utama, yaitu indeks massa tubuh (BMI), kadar gula darah, kadar hemoglobin terglikasi (HbA1c), dan usia. Dataset yang digunakan diperoleh dari platform Kaggle dan terdiri atas 50.066 entri. Data tersebut diproses melalui tahapan prapengolahan, yang mencakup normalisasi, penyeimbangan data, dan seleksi fitur, sebelum dibagi menjadi tiga bagian: 70% untuk pelatihan, 20% untuk pengujian, dan 10% untuk validasi. Hasil evaluasi menunjukkan bahwa model KNN dengan nilai k=4 mampu mencapai akurasi sebesar 83% pada data validasi dan 82% pada data pengujian. Meskipun hasil tersebut menunjukkan performa yang stabil dan cukup baik, penelitian selanjutnya disarankan untuk mengeksplorasi metode lain guna meningkatkan keseimbangan prediksi antar kelas.

Keywords:

Prediction
Machine Learning
Diabetes
K-Nearest Neighbors
KNN

ABSTRACT

Diabetes is a chronic disease with a globally increasing prevalence and poses a serious health risk if not detected early. This study aims to develop a simple yet accurate classification model to predict diabetes risk using the K-Nearest Neighbors (KNN) algorithm. The model utilizes four key clinical parameters: body mass index (BMI), blood glucose level, glycated hemoglobin (HbA1c), and age. The dataset used in this study was obtained from Kaggle and consists of 50,066 records. The data underwent preprocessing stages including normalization, class balancing, and feature selection before being split into training (70%), testing (20%), and validation (10%) sets. Experimental results demonstrate that the KNN model with k=4 achieves an accuracy of 83% on the validation set and 82% on the test set. Although the model shows stable and reasonably good performance, further improvement is needed to achieve better class balance in prediction. Future work may involve exploring more advanced machine learning techniques to enhance predictive capabilities and ensure fair classification across different risk groups.

Penulis Korespondensi:

Ade Maulana,
Program Studi Sistem Informasi,
Universitas Pelita Harapan,
Email: ade.maulana@lecturer.uph.edu

1. PENDAHULUAN

Diabetes merupakan salah satu penyakit tidak menular paling berbahaya, dengan jumlah penderita global mencapai sekitar 900 juta orang [1] dan menjadi penyebab 1,6 juta kematian setiap tahunnya [2]. Hingga kini, belum ditemukan obat yang benar-benar menyembuhkan diabetes, sehingga langkah preventif menjadi sangat penting. Salah satu tantangan utama adalah keterlambatan diagnosis dan kurangnya pemahaman masyarakat tentang faktor risiko [3]. Penelitian menunjukkan bahwa gaya

hidup, kondisi kesehatan [4], umur dan kelebihan indeks massa tubuh (BMI) [5], berperan besar dalam mengganggu kerja hormon insulin dan meningkatkan risiko diabetes [6]. Bertambahnya usia juga meningkatkan risiko diabetes diakibatkan oleh penurunan fungsi sel β pankreas dan sensitivitas insulin. Sebuah studi menunjukkan hubungan signifikan antara usia dan kadar gula darah (*p-value* = 0,004) [7], menjadikan usia faktor penting yang memengaruhi diabetes. Diabetes tidak hanya mengganggu regulasi gula darah, tetapi juga meningkatkan risiko komplikasi serius seperti penyakit jantung [5]. Di Indonesia, prevalensi diabetes juga meningkat, dengan sekitar 19,5 juta penduduk usia 20–79 tahun tercatat mengidap diabetes pada 2021 [8], menunjukkan bahwa diabetes menjadi ancaman nyata bagi sistem kesehatan nasional.

Salah satu pendekatan yang banyak digunakan saat ini untuk membantu prediksi penyakit adalah *machine learning*, yaitu metode komputasi yang mampu mempelajari pola dari data historis. Dalam penelitian ini, penulis menggunakan algoritma *K-Nearest Neighbor* (KNN) sebagai metode klasifikasi untuk memprediksi risiko diabetes berdasarkan empat parameter utama: *Body Mass Index* (BMI), kadar gula darah, usia, dan HbA1c. KNN merupakan algoritma klasifikasi yang sederhana namun cukup efektif, karena bekerja dengan cara membandingkan data baru dengan sejumlah data tetangga terdekat yang telah diketahui labelnya, lalu menentukan klasifikasi berdasarkan mayoritas label tersebut. Algoritma ini juga dinilai cocok untuk data kesehatan karena fleksibel dan mudah diimplementasikan.

Sejumlah penelitian sebelumnya telah menunjukkan bahwa KNN mampu memberikan hasil yang kompetitif dalam klasifikasi diabetes. Misalnya, penelitian oleh [9] berhasil memperoleh akurasi hingga 85% dengan K=1 dalam memprediksi potensi diabetes menggunakan data dari *Kaggle*. Penelitian lain [10] juga menyimpulkan bahwa nilai K yang optimal dalam klasifikasi diabetes dapat menghasilkan akurasi hingga 76,56%, tergantung pada karakteristik dataset yang digunakan. Sementara itu, penelitian [11] mencatat akurasi 66,6% pada implementasi KNN menggunakan perangkat lunak Weka dengan data yang lebih sederhana. Selain itu penelitian oleh [12] dengan menggunakan atribut BMI untuk memprediksi diabates memperoleh akurasi sebesar 67,8%.

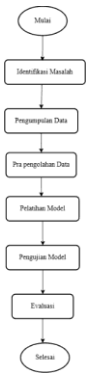
Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengembangkan model klasifikasi diabetes yang sederhana namun akurat, dengan memanfaatkan empat variabel utama yang relevan dan mudah diakses dari data pasien, yaitu BMI, HbA1c, kadar gula darah, dan usia. Diharapkan, hasil dari model ini dapat digunakan sebagai alat bantu dalam deteksi dini risiko diabetes, sehingga pengobatan dapat dilakukan lebih cepat dan risiko komplikasi dapat diminimalkan.

2. METODE PENELITIAN

Metode penelitian yang digunakan dalam studi ini bersifat eksperimental, dengan menerapkan algoritma machine learning K-Nearest Neighbor (KNN) untuk proses klasifikasi data, sebagaimana digambarkan pada Gambar 1. Tahapan penelitian meliputi identifikasi masalah, pengumpulan data, praproses data, pelatihan model, serta pengujian dan evaluasi. Data yang diperoleh dari sumber terkait kemudian diproses melalui beberapa tahap, yaitu pembersihan (*cleaning*), normalisasi, dan pemilihan atribut (*feature selection*). Setelah itu, algoritma KNN digunakan untuk melakukan klasifikasi berdasarkan jarak terdekat antara data baru dengan sejumlah tetangga terdekat dalam data latih. Hasil klasifikasi kemudian dievaluasi untuk mengukur kinerja model, khususnya dalam hal akurasi. Penjelasan secara rinci mengenai setiap tahapan dalam penelitian ini akan disajikan pada subbagian berikutnya.

2.1. Identifikasi Masalah

Diabetes menunjukkan tren peningkatan baik secara global maupun nasional, sementara hingga kini belum ada obat yang benar-benar menyembuhkannya. Deteksi dini sangat penting untuk mencegah komplikasi serius, namun sering terhambat oleh rendahnya kesadaran terhadap faktor-faktor risiko seperti BMI, umur, kadar gula darah, HbA1C dan lain-lain. Penelitian ini bertujuan untuk mengoptimalkan penggunaan faktor-faktor tersebut dalam membangun model klasifikasi berbasis data yang akurat untuk mengidentifikasi individu dengan risiko tinggi terkena diabetes agar deteksi dini dan pencegahan diabetes di masyarakat dapat dilakukan.



Gambar 1. *Research Metodology*

2.2. Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari *platform* Kaggle. Dataset ini berisi sebanyak 50.066 data yang mencakup informasi mengenai kondisi medis dan gaya hidup yang berkaitan dengan risiko diabetes. Setiap data pasien memiliki sembilan atribut, yaitu *gender*, *age*, *hypertension*, *heart_disease*, *smoking_history*, bmi, *HbA1c_level*, *blood_glucose_level*, dan diabetes. Label diabetes digunakan untuk menunjukkan status pasien, dengan nilai 0 untuk pasien tanpa diabetes dan 1 untuk pasien dengan diabetes.

Tabel 1. Atribut

No.	Atribut	Keterangan
1	*gender*	Jenis Kelamin
2	*age*	Umur pasien
3	*hypertension*	Tekanan darah tinggi pasien
4	*heart_disease*	Riwayat penyakit jantung
5	*smoking_history*	Riwayat merokok
6	bmi	*Body Mass Index*
7	*HbA1c_level*	Ukuran kadar gula darah rata-rata
8	*blood_glucose_level*	Jumlah glukosa yang terdapat dalam darah
9	diabetes	Kelas *output* (target)

2.3. Prapengolahan

Jumlah atribut yang berlebihan dapat memicu fenomena *curse of dimensionality* pada KNN [13]. Untuk menguranginya, kami menerapkan seleksi fitur berbasis korelasi Pearson dalam dua tahap, yaitu pertama *univariate filtering* untuk memilih atribut yang paling berkorelasi kuat dengan target [14] , kedua *redundancy filtering* untuk membuang atribut yang saling berkorelasi tinggi satu sama lain [15]. Berikut adalah hasil yang diperoleh ketika kami menghitung masing-masing korelasi Pearson pada atribut.

Tabel 2. Korelasi Pearson pada atribut.

diabetes	1.000000
HbA1c_level	0.598542
blood_glucose_level	0.526804
age	0.398069
bmi	0.305259
gender	0.279547
smoking_history	0.075431
hypertension	0.028299
heart_disease	0.015739

Dari sembilan atribut yang tersedia, empat dipilih berdasarkan relevansinya dalam memprediksi risiko diabetes, karena memiliki nilai korelasi Pearson diatas 3. Empat atribut tersebut adalah HbA1c, kadar gula darah, umur, dan bmi. Prevalensi diabetes meningkat seiring bertambahnya usia—3,0 % pada kelompok 20–44 tahun, 14,5 % pada 45–64 tahun, dan mencapai 24,4 % pada usia ≥ 65 tahun menunjukkan bahwa usia merupakan faktor determinan penting dalam kejadian diabetes [16]. Selain itu, kenaikan *Body Mass Index* (BMI) sebagai indikator adipositas berbanding lurus dengan peningkatan risiko diabetes [17]. Seseorang dapat didiagnosis diabetes jika memiliki kadar gula darah puasa ≥126 mg/dL, gula darah 2 jam setelah makan ≥200 mg/dL, atau HbA1c ≥6,5% [18]. HbA1c ini sering digunakan sebagai atribut yang paling penting dalam model prediksi berbasis data, karena menggambarkan kondisi glikemik seseorang secara kuantitatif dan langsung berhubungan dengan diagnosis medis.

Untuk menyeimbangkan data antara kelas diabetes (kelas minoritas) dan kelas non-diabetes (kelas mayoritas), kami menggunakan teknik SMOTETomek. Teknik ini menggabungkan SMOTE *(Synthetic Minority Oversampling Technique)* untuk menambah data sintetis pada kelas minoritas, dan *Tomek Links* untuk menghapus data yang tumpang tindih antar kelas, sehingga distribusi data menjadi lebih seimbang dan bersih.

Tabel 3. Data kelas diabetes dan non-diabetes sebelum SMOTETomek

Diabetes	39.116
Tidak Diabetes	1.890

Setelah menggunakan SMOTETomek, jumlah data di kedua kelas menjadi hampir seimbang, yaitu 36.129 data untuk kelas diabetes dan 35.279 data untuk kelas non-diabetes. Ini menunjukkan bahwa SMOTETomek berhasil mengurangi ketidakseimbangan data, sehingga model yang dibuat nantinya bisa belajar dengan lebih adil dan tidak bias ke salah satu kelas saja.

Tabel 4. Data kelas Diabetes dan non diabetes sesudah SMOTETomek

Diabetes	36.129
Tidak Diabetes	35.279

Prapengolahan dilakukan untuk memastikan data siap digunakan oleh model KNN. Tahapan dilakukan menggunakan Google Colab dan *library* seperti *pandas* dan *numpy*.

Tabel 5. Langkah prapengolahan

No.	Tahap	Deskripsi
1.	Hapus duplikat	Data identik dihapus untuk menghindari bias.
2.	Ubah fitur menjadi angka	Mengubah fitur jenis kelamin dan riwayat merokok menjadi angka.
3.	Normalisasi data	Hapus usia dibawah 10 tahun, Batasi BMI antara 10-50, Batasi *level* HbA1c antara 20-100, Batasi kadar gula darah antara 50-155
4.	Seimbangkan data	Menyeimbangkan data antara kelas diabetes dan non-diabetes dengan SMOTETomek
5.	Seleksi fitur/atribut	Memilih fitur /atribut usia, BMI, kadar gula darah, *level* HbA1c berdasarkan korelasi Pearson
6.	Hapus fitur/atribut	Setelah memilih fitur/atribut yang cocok , fitur/atribut lainnya dihapus

2.4. Pelatihan Model

Pelatihan model dalam penelitian ini dilakukan dengan membagi dataset menjadi tiga bagian menggunakan 70% untuk data latih, 20% untuk data pengujian, dan 10% untuk data validasi. Pembagian ini dipilih karena secara umum dianggap optimal dalam pengembangan model *machine learning* untuk menjaga keseimbangan antara pelatihan dan evaluasi performa model, serta telah digunakan dalam studi terdahulu seperti oleh [19]. Pemilihan nilai *k* yang optimal sangat memengaruhi efektivitas algoritma KNN dalam klasifikasi diabetes untuk mencapai akurasi yang tinggi dalam prediksi diabetes [20]. Oleh karena itu, penentuan nilai *k* harus mempertimbangkan nilai tersebut cukup besar untuk mengurangi pengaruh *nois*e, namun masih cukup kecil agar model tetap sensitif terhadap pola lokal dalam data. Pemilihan nilai $k = 4$ didasarkan pada hasil akurasi terbaik yang diperoleh. Meskipun lebih rentan terhadap hasil yang seri, model tetap mampu menentukan kelas mayoritas dengan baik.

2.5 Pengujian

Pada tahap pengujian, sebanyak 20% dari total data digunakan. Proporsi ini dipilih karena cukup representatif untuk mengevaluasi kinerja model terhadap data yang belum pernah digunakan dalam proses pelatihan. Dataset yang digunakan dalam pengujian memiliki distribusi kelas yang seimbang, yaitu antara data pasien dengan diabetes dan non-diabetes. Keseimbangan ini merupakan faktor penting dalam pengujian model klasifikasi, karena dapat mencegah model menjadi bias terhadap salah satu kelas [21]. Model yang dilatih dan diuji dengan data yang tidak seimbang berisiko memiliki kecenderungan untuk hanya mengenali kelas mayoritas, sehingga menurunkan kualitas prediksi pada kelas minoritas.

2.6 Evaluasi

Untuk mengevaluasi kinerja model klasifikasi *K-Nearest Neighbors* (KNN), digunakan berbagai metrik evaluasi yang berasal dari *confusion matrix*, seperti akurasi, presisi, *recall*, *f1-score*, serta grafik ROC dan AUC. *Confusion matrix* memberikan gambaran mengenai distribusi model terhadap label sebenarnya, dengan membedakan *true positive*, *false negative*, *false positive*, *dan false negative*.
Akurasi menunjukkan seberapa banyak prediksi yang benar dari keseluruhan data.

$$Akurasi= TP+TNTP+TN+FP+FN\ Akurasi = \frac{TP+TN}{TP+TN+FP+FN}$$

Presisi mengukur ketepatan prediksi untuk kelas positif.

$$Presisi = \frac{TP}{TP + FP}$$

Recall mengukur seberapa banyak dari total kasus positif yang berhasil dikenali dengan benar oleh model.

$$Recall = \frac{TP}{TP + FN}$$

Di mana:

- TP (*True Positive*): jumlah kasus diabetes yang diprediksi dengan benar
- TN (*True Negative*): jumlah kasus non-diabetes yang diprediksi dengan benar
- FP (*False Positive*): jumlah kasus non-diabetes yang salah diklasifikasikan sebagai diabetes
- FN (*False Negative*): jumlah kasus diabetes yang salah diklasifikasikan sebagai non-diabetes

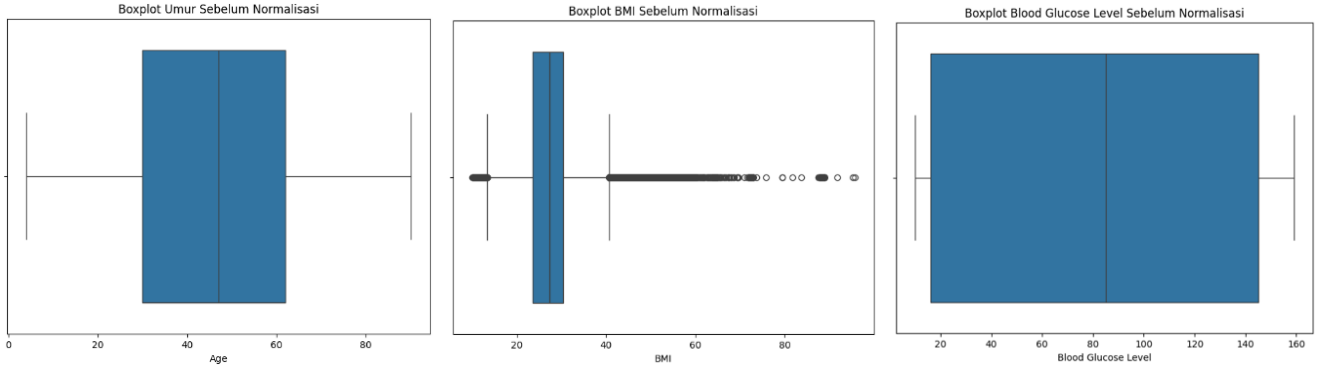
3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan dataset diabetes yang sudah melalui prapengolahan dengan jumlah total 50.066 data. Dataset tersebut memiliki distribusi yang seimbang dengan 25.566 catatan diklasifikasikan sebagai penderita diabetes dan 24.500 catatan diklasifikasikan sebagai bukan penderita diabetes. Berikut ditampilkan beberapa sampel data dari dataset diabetes prapengolahan yang digunakan dalam penelitian.

Tabel 6. Dataset diabetes.

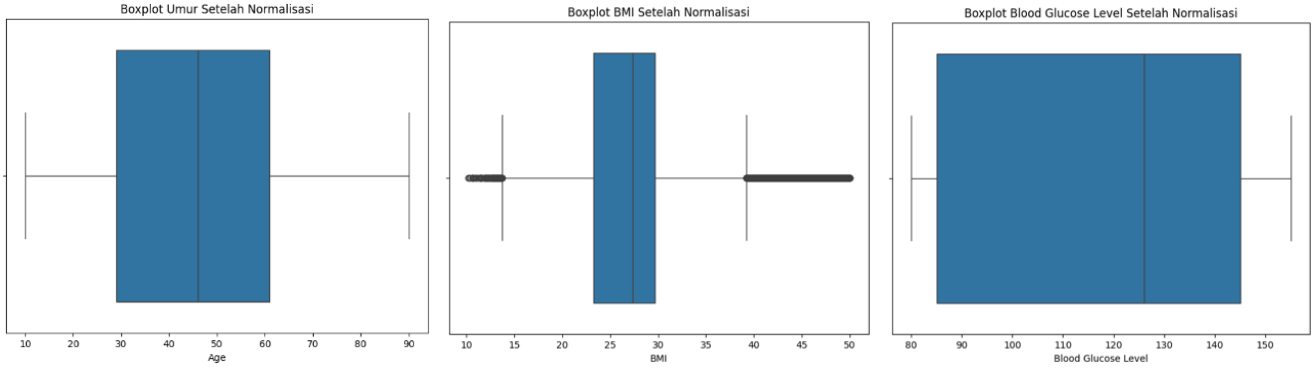
Gender	Age	Heart_disease	bmi	HbA1c_level	blood_glucose_level	Diabetes
0	28	0	27.32	57	158	0
1	36	0	23.45	50	158	0
0	76	1	20.14	48	155	0
…	…	…	…	…	…	…
1	15	0	24.35	61	126	1
0	79	0	24.48	68	159	1

Sebelum dilakukan proses normalisasi dan pembersihan data, penting untuk memahami distribusi nilai dari masing-masing fitur yang digunakan dalam model. Visualisasi ini membantu mengidentifikasi nilai-nilai ekstrem serta pola distribusi awal yang dapat memengaruhi performa model KNN. Pada tahap ini, tiga *boxplot* (gambar 2) ditampilkan untuk menunjukkan penyebaran data umur, BMI, dan kadar gula darah sebelum dilakukan normalisasi. Visualisasi ini membantu melihat rentang nilai dari masing-masing data serta mendeteksi nilai-nilai yang terlalu jauh dari rata-rata.



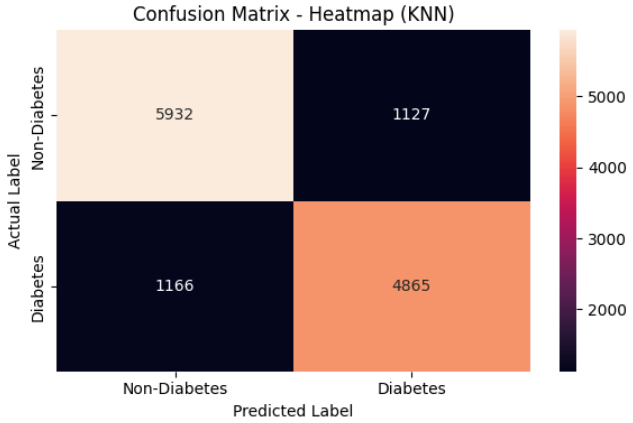
Gambar 2. Box Plot Umur (kiri), BMI (tengah), dan Blood Glucose (kanan) Sebelum Normalsasi.

Setelah data dibersihkan dan dinormalisasi, nilai-nilai pada fitur seperti umur, BMI, dan kadar gula darah menjadi lebih teratur dan masuk akal. Data umur yang terlalu kecil, seperti di bawah 10 tahun, sudah dihapus agar sesuai dengan data sebenarnya. Nilai BMI yang terlalu tinggi juga dibuang supaya datanya tidak terlalu menyimpang. Begitu juga dengan kadar gula darah, hanya data dengan nilai antara 50 sampai 155 mg/dL yang digunakan. Dengan data yang lebih rapi seperti ini, model KNN bisa bekerja lebih baik karena datanya sudah bersih dan seimbang.



Gambar 3. Box Plot Umur (kiri), BMI (tengah), dan Blood Glucose (kanan) Setelah Normalsasi.

Setelah memastikan tidak ada nilai yang hilang pada data, dilakukan transformasi data kategori menjadi format numerik serta penyesuaian skala data. Selanjutnya, data dibagi menjadi tiga subset, yaitu 70% untuk pelatihan, 20% untuk pengujian, dan 10% untuk validasi. Proses klasifikasi dilakukan menggunakan algoritma K-Nearest Neighbor (KNN) dengan parameter k = 4, yang berarti klasifikasi ditentukan berdasarkan empat data terdekat pada setiap titik data. Model kemudian dievaluasi menggunakan confusion matrix, yang selanjutnya digunakan untuk menghitung nilai akurasi, sebagaimana ditunjukkan pada gambar berikut.



Gambar 4. Confusion Matrix.

Model KNN menunjukkan performa yang baik dengan akurasi sebesar 82%. Nilai precision, recall, dan f1-score untuk kedua kelas (0 dan 1) berada pada rentang 0,81 hingga 0,84, menandakan model mampu mengenali dan mengklasifikasikan kedua kelas secara cukup seimbang. Konsistensi model dalam mengklasifikasikan data tercermin dari nilai rata-rata macro dan weighted average sebesar 0,82, meskipun terdapat perbedaan jumlah data antar kelas. Dengan hasil ini, model KNN layak digunakan untuk tugas klasifikasi, terutama pada dataset yang tidak terlalu kompleks dan tidak memerlukan pemisahan kelas yang sangat presisi.

Tabel 7. Hasil pengujian.

	precision	recall	f-1 score	support
0	0.84	0.84	0.84	7059
1	0.81	0.81	0.81	6031
accuracy			0.82	13090
macro avg	0.82	0.82	0.82	13090
weighted avg	0.82	0.82	0.82	13090

Pada data validasi yang merupakan 10% dari total dataset, model mencapai akurasi sebesar 0,82, menunjukkan kestabilan kinerja meskipun diuji pada data yang berbeda dari pelatihan dan pengujian. Metrik lain seperti precision, recall, dan f1-score juga konsisten dengan hasil pada data pengujian. F1-score untuk kelas 0 dan 1 masing-masing adalah 0,84 dan 0,82, menandakan kemampuan model dalam membedakan kedua kelas dengan baik. Precision kelas 0 sebesar 0,84 menunjukkan mayoritas prediksi untuk kelas ini tepat, sedangkan recall kelas 1 sebesar 0,81 mengindikasikan model cukup baik dalam mengenali data yang memang termasuk kelas tersebut. Berdasarkan hasil ini, dapat disimpulkan bahwa model KNN tidak hanya efektif pada data pelatihan dan pengujian, tetapi juga andal dalam memproses data validasi.

Tabel 8. Hasil validasi.

	precision	recall	f1- score	support
0	0.84	0.85	0.84	3530
1	0.82	0.81	0.82	3016
accuracy			0.83	6546
macro avg	0.83	0.83	0.83	6546
weighted avg	0.83	0.83	0.83	6546

Model menunjukkan performa yang baik dan stabil dengan akurasi sebesar 0,82 pada data pengujian dan 0,83 pada data validasi. Hasil ini mengindikasikan bahwa model mampu mengklasifikasikan data dengan benar sebesar 82% dari total data. Nilai f1-score mencapai 0,84 untuk kelas 0 dan 0,81 untuk kelas 1. Precision untuk kelas 0 sebesar 0,84 menunjukkan bahwa sebagian besar prediksi pada kelas ini tepat, sedangkan recall untuk kelas 1 sebesar 0,81 menandakan kemampuan model dalam mengenali mayoritas data yang memang termasuk kelas tersebut. Dengan hasil yang konsisten pada data validasi, model ini terbukti dapat menggeneralisasi dengan baik dan dapat diandalkan untuk mengklasifikasikan data baru.

4. KESIMPULAN

Penelitian ini berhasil mencapai tujuan utama, yaitu mengembangkan model klasifikasi yang sederhana namun akurat untuk memprediksi risiko diabetes berdasarkan empat variabel utama: Body Mass Index (BMI), HbA1c, usia, dan kadar glukosa darah, dengan menggunakan algoritma *K-Nearest Neighbors* (KNN). Model KNN yang dibangun menunjukkan performa yang cukup baik, dengan akurasi sebesar 83% pada data validasi dan 82% pada data pengujian. Hasil ini menunjukkan bahwa pendekatan KNN dapat digunakan secara efektif sebagai alat bantu dalam deteksi dini risiko diabetes.

Dibandingkan dengan beberapa penelitian sebelumnya, akurasi model dalam studi ini tergolong lebih tinggi, yang mengindikasikan bahwa pemilihan variabel serta proses prapengolahan data telah dilakukan dengan tepat. Meskipun demikian, model ini masih memiliki keterbatasan, terutama dalam hal prediksi negatif (non-diabetes), yang cenderung kurang akurat. Hal ini kemungkinan disebabkan oleh distribusi data yang tidak seimbang, di mana jumlah data penderita diabetes lebih dominan dibandingkan data non-penderita.

Ke depannya, pengembangan model dapat diarahkan pada peningkatan keseimbangan data, penyesuaian parameter algoritma, atau penerapan metode klasifikasi lain untuk memperbaiki kinerja prediksi pada kedua kelas. Dengan demikian, model yang dihasilkan tidak hanya akurat, tetapi juga seimbang dalam mengenali individu dengan dan tanpa risiko diabetes.

UCAPAN TERIMA KASIH

Kami mengucapkan terima kasih kepada Universitas Pelita Harapan yang telah memberikan dukungan dalam bentuk fasilitas dan pembimbingan selama proses penelitian ini. Apresiasi juga kami sampaikan kepada dosen pengampu mata kuliah yang telah memberikan arahan dan saran berharga yang sangat membantu dalam penyusunan jurnal ini. Data yang digunakan dalam penelitian ini bersumber dari *platform* Kaggle, yang menyediakan dataset diabetes yang digunakan dalam pengujian model klasifikasi.

REFERENSI

[1] "Diabetes." Accessed: Apr. 09, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] L. E. Wagenknecht *et al.*, "Trends in incidence of youth-onset type 1 and type 2 diabetes in the USA, 2002–18: results from the population-based SEARCH for Diabetes in Youth study," *Lancet Diabetes Endocrinol*, vol. 11, no. 4, pp. 242–250, Apr. 2023, doi: 10.1016/S2213-8587(23)00025-6.

[3] D. Wahyu, H. ∗ Jurusan, I. Kesehatan, I. Keolahragaan, D. Disetujui, and _____ D., "FAKTOR-FAKTOR YANG BERHUBUNGAN DENGAN KEPATUHAN DALAM PENGELOLAAN DIET PADA PASIEN RAWAT JALAN DIABETES MELLITUS TIPE 2 DI KOTA SEMARANG," 2017. [Online]. Available: http://journal.unnes.ac.id/sju/index.php/jhealthedu/

[4] V. Durlach *et al.*, "Smoking and diabetes interplay: A comprehensive review and joint statement," *Diabetes Metab*, vol. 48, no. 6, Nov. 2022, doi: 10.1016/J.DIABET.2022.101370.

[5] D. Mohajan and H. K. Mohajan, "Visceral Fat Increases Cardiometabolic Risk Factors Among Type 2 Diabetes Patients," *Innovation in Science and Technology*, vol. 3, no. 5, pp. 26–30, Sep. 2024, doi: 10.56397/IST.2024.09.03.

[6] L. Ismail, H. Materwala, and J. Al Kaabi, "Association of risk factors with type 2 diabetes: A systematic review," *Comput Struct Biotechnol J*, vol. 19, p. 1759, Jan. 2021, doi: 10.1016/J.CSBJ.2021.03.003.

[7] S. Rahayu and Stik. Jayakarta PKP DKI Jakarta, "HUBUNGAN USIA, JENIS KELAMIN DAN INDEKS MASSA TUBUH DENGAN KADAR GULA DARAH PUASA PADA PASIEN DIABETES MELITUS TIPE 2 DI KLINIK PRATAMA RAWAT JALAN PROKLAMASI, DEPOK, JAWA BARAT," 2020.

[8] "Indonesia - International Diabetes Federation." Accessed: Apr. 09, 2025. [Online]. Available: https://idf.org/our-network/regions-and-members/western-pacific/members/indonesia/

[9] M. S. Fiqri and H. Dwi Bhakti, "KLASIFIKASI POTENSI PENYAKIT DIABETES MELLITUS TIPE II PADA PASIEN MENGGUNAKAN ALGORITME KNN (K-NEAREST NEIGHBOR)," 2024.

[10] N. Azizah, M. R. Firdaus, R. Suyaningsih, F. Indrayatna, and U. Padjadjaran, "Penerapan Algoritma Klasifikasi K-Nearest Neighbor pada Penyakit Diabetes," 2023, [Online]. Available: http://prosiding.snsa.statistics.unpad.ac.id

[11] A. Asmarani *et al.*, "Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM) Implementasi Algoritma K-Nearst Neighbor Untuk Memprediksi Penyakit Diabetes," 2022. [Online]. Available: https://ejournal.unama.ac.id/index.php/jakakom

[12] A. Gunawan and I. Fenriana, "Design of Diabetes Prediction Application Using K-Nearest Neighbor Algorithm," *bit-Tech*, vol. 6, no. 2, pp. 110–117, Dec. 2023, doi: 10.32877/bt.v6i2.939.

[13] N. Kouiroukidis and G. Evangelidis, "The Effects of Dimensionality Curse in High Dimensional kNN Search," in *2011 15th Panhellenic Conference on Informatics*, 2011, pp. 41–45. doi: 10.1109/PCI.2011.45.

[14] G. Manikandan and S. Abirami, "An efficient feature selection framework based on information theory for high dimensional data," *Appl Soft Comput*, vol. 111, p. 107729, 2021, doi: https://doi.org/10.1016/j.asoc.2021.107729.

[15] P. Shen, X. Ding, W. Ren, and S. Liu, "A stable feature selection method based on relevancy and redundancy," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1742-6596/1732/1/012023.

[16] Centers for Disease Control and Prevention, "Laporan Statistik Diabetes Nasional." Accessed: Apr. 21, 2025. [Online]. Available: https://www.cdc.gov/diabetes/php/data-research/index.html

[17] S. Klein, A. Gastaldelli, H. Yki-Järvinen, and P. E. Scherer, "Why does obesity cause diabetes?," *Cell Metab*, vol. 34, no. 1, pp. 11–20, 2022, doi: https://doi.org/10.1016/j.cmet.2021.12.012.

[18] American Diabetes Association, "Glycemic targets: Standards of medical care in diabetes-2020," *Diabetes Care*, vol. 43, pp. S66–S76, Jan. 2020, doi: 10.2337/dc20-S006.

[19] H. Gupta, H. Varshney, T. K. Sharma, N. Pachauri, and O. P. Verma, "Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction," *Complex & Intelligent Systems*, vol. 8, no. 4, pp. 3073–3087, 2022.

[20] A. F. Lubis *et al.*, "Classification of Diabetes Mellitus Sufferers Eating Patterns Using K-Nearest Neighbors, Na\"\ive Bayes and Decission Tree," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 2, no. 1, pp. 44–51, 2024.

[21] Q. Wei and R. L. Dunbrack, "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics," *PLoS One*, vol. 8, no. 7, Jul. 2013, doi: 10.1371/journal.pone.0067863.