

Penerapan Algoritma KNN dengan *K-Fold Cross Validation* untuk Diagnosa Risiko Diabetes Mellitus

Hafipa Sudiadarma¹ | Ida Mulyadi^{*2} | Fahrir Irhamna Rachman²

¹ Mahasiswa Program Studi Informatika,
Fakultas Teknik, Universitas
Muhammadiyah Makassar, Indonesia.

Email:
fifa3577@gmail.com

² Program Studi Informatika, Fakultas Teknik,
Universitas Muhammadiyah Makassar,
Indonesia.

Email:
idamulyadi@unismuh.ac.id;
fachrim141020@unismuh.ac.id

Korespondensi:

*Ida Mulyadi
idamulyadi@unismuh.ac.id

ABSTRAK

Diabetes mellitus (DM) merupakan penyakit kronis dengan prevalensi yang terus meningkat dan berpotensi menimbulkan berbagai komplikasi serius apabila tidak terdeteksi sejak dini. Keterbatasan metode diagnostik konvensional dalam menangani data kesehatan yang besar dan kompleks mendorong pemanfaatan pendekatan berbasis *machine learning*. Penelitian ini bertujuan untuk membangun model prediksi risiko diabetes mellitus menggunakan algoritma *K-Nearest Neighbor* (KNN) dengan metode *Stratified K-Fold Cross Validation*. *Dataset* yang digunakan terdiri dari 1.041 data pasien yang diperoleh dari Rumah Sakit Haji Makassar, dengan variabel meliputi usia, tekanan darah, status gula darah sewaktu, indeks massa tubuh, dan lingkar perut. Tahapan penelitian meliputi pemrosesan data, normalisasi menggunakan *Standard Scaler*, pemodelan KNN dengan metrik jarak *Manhattan*, serta evaluasi kinerja model menggunakan akurasi, *precision*, *recall*, dan *F1-score*. Hasil penelitian menunjukkan bahwa model KNN mampu mencapai rata-rata akurasi sebesar 87,55% dengan performa yang stabil pada setiap *fold*. Analisis *feature importance* menunjukkan bahwa tekanan darah sistolik, lingkar perut, dan gula darah sewaktu merupakan faktor yang paling berpengaruh terhadap status gula darah. Hasil ini menunjukkan bahwa algoritma KNN berpotensi digunakan sebagai alat bantu deteksi dini risiko diabetes mellitus berbasis data kesehatan.

Kata Kunci:

Diabetes Mellitus, *K-Nearest Neighbor*, *Machine Learning*, *K-Fold Cross Validation*, Prediksi Risiko

ABSTRACT

Diabetes mellitus (DM) is a chronic disease with a continuously increasing prevalence and the potential to cause various serious complications if not detected early. The limitations of conventional diagnostic methods in handling large and complex health data have encouraged the use of machine learning-based approaches. This study aims to develop a diabetes mellitus risk prediction model using the K-Nearest Neighbor (KNN) algorithm with the Stratified K-Fold Cross Validation method. The dataset consisted of 1,041 patient records obtained from Haji Hospital Makassar, including variables such as age, blood pressure, random blood glucose level, body mass index, and waist circumference. The research stages included data preprocessing, normalization using Standard Scaler, KNN modeling with the Manhattan distance metric, and model performance evaluation using accuracy, precision, recall, and F1-score. The results showed that the KNN model achieved an average accuracy of 87.55% with stable performance across each fold. Feature importance analysis indicated that systolic blood pressure, waist circumference, and random blood glucose level were the most influential factors affecting blood glucose status. These findings suggest that the KNN algorithm has the potential to be used as a decision-support tool for the early detection of diabetes mellitus risk based on health data..

Keywords:

Diabetes Mellitus, K-Nearest Neighbor, Machine Learning, K-Fold Cross-Validation, Risk Prediction

1 | PENDAHULUAN

Diabetes Mellitus (DM) merupakan salah satu penyakit kronis yang prevalensinya terus meningkat secara global dan menjadi tantangan besar bagi kesehatan masyarakat (Liu et al., 2020). Berdasarkan IDF Diabetes Atlas edisi ke-10, pada tahun 2021

terdapat sekitar 536,6 juta orang dewasa (10,5 %) yang hidup dengan diabetes, dan jumlah ini diproyeksikan meningkat menjadi 783,2 juta (12,2 %) pada tahun 2045 (Sun et al., 2022). Analisis *Global Burden of Disease Study* 2021 menunjukkan bahwa sejak tahun 1990 hingga 2021, prevalensi diabetes meningkat sebesar 90,5 %, dan diperkirakan akan mencapai 1,31 miliar orang pada tahun 2050 (Ong et al., 2023). Lonjakan ini terjadi di hampir seluruh wilayah dunia, termasuk Indonesia, yang dipengaruhi oleh perubahan gaya hidup, pola makan, dan faktor genetik, sehingga menambah beban biaya kesehatan dan menurunkan kualitas hidup penderita (Ampofo & Boateng, 2020).

Diagnosis dini berperan penting dalam mencegah komplikasi jangka panjang akibat diabetes, karena intervensi gaya hidup dan terapi farmakologis yang dilakukan pada tahap awal terbukti dapat memperlambat progresivitas penyakit serta mengurangi risiko komplikasi *vaskular* dan *neuropatik* (Ong et al., 2023). Keterbatasan ini mendorong berkembangnya pendekatan berbasis teknologi *machine learning* yang mampu mengolah data kesehatan secara cepat, akurat, dan efisien, sehingga berpotensi memperluas jangkauan deteksi dini dan meningkatkan efektivitas program pencegahan diabetes (Dritsas & Trigka, 2022). Beberapa penelitian menunjukkan bahwa model berbasis *machine learning*, seperti *K-Nearest Neighbor* (KNN) dan *Random Forest*, mampu memprediksi risiko diabetes dengan akurasi lebih dari 90% ketika dikombinasikan dengan teknik validasi silang dan optimisasi parameter (Al-Khamees et al., 2025).

Peningkatan prevalensi diabetes mellitus secara global berkaitan erat dengan faktor risiko seperti obesitas, kurangnya aktivitas fisik, dan pola makan tidak sehat yang semakin dominan di berbagai negara berkembang maupun maju (Sun et al., 2022). Studi epidemiologi terbaru menunjukkan bahwa transisi gaya hidup dan urbanisasi berkontribusi signifikan terhadap peningkatan kejadian diabetes tipe 2, khususnya pada populasi usia produktif (Ong et al., 2023). Keterlambatan diagnosis diabetes sering menyebabkan pasien terdeteksi pada fase lanjut, yang secara signifikan meningkatkan risiko komplikasi *kardiovaskular*, *nefropati*, dan *neuropati diabetik* (Fang & Selvin, 2021). Penelitian *longitudinal* juga menunjukkan bahwa deteksi dini dan intervensi awal mampu menurunkan progresivitas penyakit serta mengurangi risiko komplikasi *mikrovaskular* secara signifikan (Lachin et al., 2021). Namun, pendekatan diagnostik konvensional masih memiliki keterbatasan dalam menangani volume data kesehatan yang besar dan kompleks sehingga mendorong pemanfaatan metode *machine learning* dalam sistem prediksi diabetes (Fregoso-Aparicio et al., 2021). Algoritma *machine learning* mampu mengekstraksi pola non-linear dari data klinis dan demografis yang sulit diidentifikasi secara manual oleh tenaga medis (Deberneh & Kim, 2021). Beberapa penelitian menunjukkan bahwa algoritma *K-Nearest Neighbor* (KNN) memiliki performa kompetitif dalam klasifikasi diabetes karena kesederhanaan dan kemampuannya dalam menangani data berdimensi tinggi. Selain itu, integrasi KNN dengan teknik *k-fold cross-validation* terbukti mampu meningkatkan stabilitas model dan mengurangi bias akibat pembagian data latih dan data uji yang tidak seimbang (Villanueva et al., 2023).

Dalam dekade terakhir, perkembangan teknologi *machine learning* telah membuka peluang baru bagi dunia medis, khususnya dalam pengolahan data kesehatan untuk tujuan prediksi dan diagnosis penyakit (Narasimharao et al., 2026). Salah satu algoritma yang banyak digunakan adalah *K-Nearest Neighbor* (KNN), yang bekerja dengan prinsip menghitung kedekatan jarak antara data baru dengan data latih (muthu, 2023). Penerapan KNN dengan teknik *k-fold cross-validation* dan metode *SMOTE* menghasilkan akurasi 83,33 %, *precision* 85,00 %, *recall* 83,95 %, dan *F1-score* 84,47 %, sehingga lebih unggul dibandingkan *Support Vector Machine* (Afolabi et al., 2025). KNN dengan *k-fold cross-validation* serta kombinasi *SMOTE-ENN* untuk menangani ketidakseimbangan data, dan hasilnya menunjukkan peningkatan signifikan pada performa model, dengan akurasi mencapai 97 %, *precision* 96 %, *recall* 97 %, dan *F1-score* 97 % (Amri et al., 2025). Temuan-temuan ini memperlihatkan bahwa KNN, apabila dikombinasikan dengan teknik validasi silang dan penanganan data tidak seimbang, mampu menghasilkan model prediksi yang andal untuk mendukung deteksi dini diabetes mellitus.

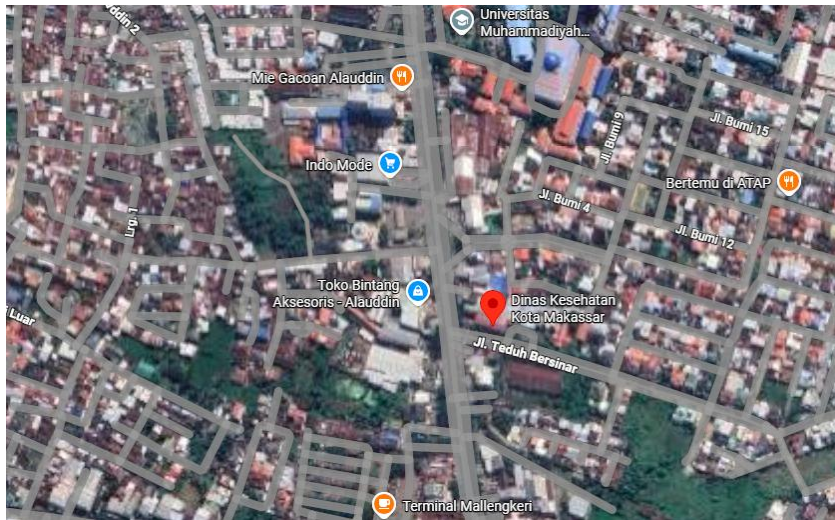
Berdasarkan permasalahan yang telah diuraikan, penelitian ini diarahkan untuk membangun model prediksi risiko Diabetes Mellitus menggunakan algoritma *K-Nearest Neighbor* (KNN) dengan teknik *k-fold cross-validation*. Hasil dari penelitian ini diharapkan tidak hanya memberikan kontribusi terhadap pengembangan kajian akademik di bidang penerapan *machine learning* untuk prediksi penyakit, tetapi juga dapat dimanfaatkan sebagai acuan dalam pengambilan keputusan medis yang berbasis data dan teknologi.

2 | METODE

Metodologi dalam penelitian ini disusun untuk mendukung penerapan algoritma *K-Nearest Neighbor* (KNN) dengan teknik *K-Fold Cross Validation* dalam mendiagnosis risiko diabetes mellitus. Rangkaian metode yang digunakan bersifat kuantitatif, dengan berfokus pada pengolahan data medis serta evaluasi performa model klasifikasi secara sistematis. Pendekatan yang diterapkan meliputi tahap pengumpulan *dataset*, pemrosesan data, pemodelan menggunakan algoritma KNN, serta validasi model menggunakan *K-Fold Cross Validation* untuk memperoleh hasil yang akurat dan generalisasi yang baik. Setiap langkah dalam proses ini didasarkan pada prinsip-prinsip pembelajaran mesin (*machine learning*) yang terstruktur, dengan integrasi aspek akurasi, efisiensi, dan keandalan sebagai dasar utama dalam pengembangan sistem diagnosis risiko diabetes mellitus.

2.1 | Sumber Data dan Lokasi Penelitian

Penelitian ini dilakukan dengan menggunakan data yang diperoleh dari Dinas Kesehatan Kota Makassar, Sulawesi Selatan. Data yang digunakan merupakan data kesehatan pasien yang mencakup beberapa parameter, seperti kadar glukosa darah, tekanan darah, indeks massa tubuh (BMI), usia, serta riwayat keluarga. Jumlah data yang digunakan dalam penelitian ini sebanyak 1.041 data pasien yang telah melalui proses seleksi dan penyesuaian sesuai kebutuhan penelitian. Data tersebut digunakan sebagai dasar dalam proses klasifikasi untuk mendiagnosis risiko diabetes mellitus menggunakan algoritma K-Nearest Neighbor (KNN). Pemanfaatan data kesehatan ini diharapkan dapat membantu dalam mengidentifikasi pola-pola yang berkaitan dengan risiko diabetes mellitus sehingga mendukung proses deteksi dini dan pengambilan keputusan yang lebih tepat. Adapun lokasi penelitian dapat ditunjukkan pada **Gambar 1**.



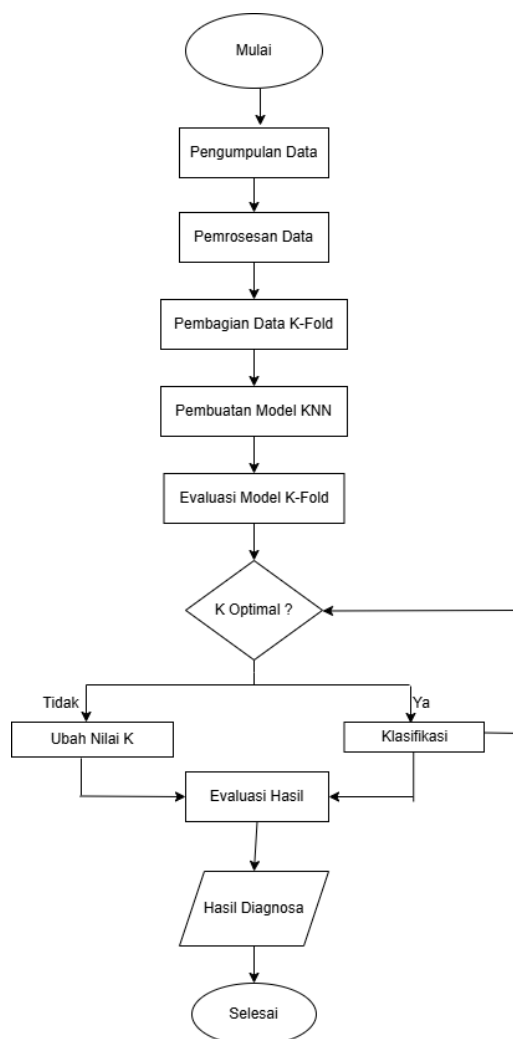
GAMBAR 1. Lokasi Penelitian

2.2 | Teknik Pengumpulan Data

Data dikumpulkan melalui metode dokumentasi dari *dataset* yang diperoleh dari instansi terkait. *Dataset* yang digunakan terdiri dari 1.041 data pasien yang memuat beberapa atribut kesehatan, seperti usia, tekanan darah, kadar glukosa darah, indeks massa tubuh (BMI), lingkar perut, dan status gula darah. Data kemudian dianalisis untuk memastikan kelengkapan, konsistensi, dan kesesuaiannya dengan kebutuhan penelitian, serta dilakukan pemeriksaan terhadap data yang hilang (*missing values*) dan data duplikat sebelum tahap pemodelan. Selain itu, penelitian ini juga didukung oleh studi literatur melalui penelaahan berbagai buku, jurnal ilmiah, dan artikel penelitian yang relevan untuk memahami konsep algoritma *K-Nearest Neighbor* (KNN), teknik *K-Fold Cross Validation*, serta penerapan *machine learning* dalam prediksi risiko diabetes mellitus. Hasil studi literatur digunakan sebagai landasan teoritis dalam pemilihan metode, perancangan model, dan interpretasi hasil penelitian.

2.3 | Tahapan Metode Penelitian

Untuk memberikan gambaran yang sistematis mengenai alur penelitian, seluruh tahapan penelitian disusun secara terstruktur mulai dari pengumpulan data hingga menghasilkan diagnosis risiko diabetes mellitus. Proses penelitian diawali dengan pengumpulan data pasien yang diperoleh dari Rumah Sakit Haji Makassar dan dilanjutkan dengan tahap pemrosesan data untuk meningkatkan kualitas data sebelum digunakan dalam pemodelan. Selanjutnya, data dibagi menggunakan teknik *K-Fold Cross Validation* untuk menghasilkan data latih dan data uji secara bergantian pada setiap *fold*. Setelah proses pembagian data selesai, dilakukan pembuatan model menggunakan algoritma *K-Nearest Neighbor* (KNN) dengan metrik jarak *Manhattan*. Model yang terbentuk kemudian dievaluasi untuk menentukan nilai *K* yang memberikan performa terbaik sebelum digunakan dalam proses klasifikasi. Tahapan perancangan sistem disajikan pada **GAMBAR 2**.



GAMBAR 2. Flowchart Perancangan Sistem

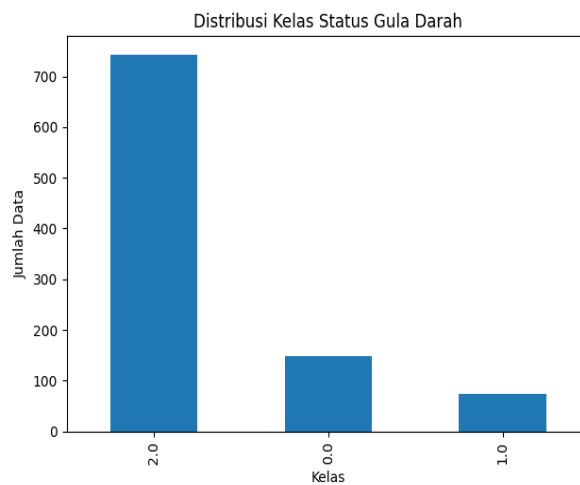
Berdasarkan **GAMBAR 2**, proses penelitian dimulai dari tahap pengumpulan data yang telah dikumpulkan kemudian diproses melalui tahap pembersihan data, transformasi data, dan normalisasi untuk memastikan kualitas data yang digunakan dalam penelitian. Selanjutnya, data dibagi menggunakan teknik *K-Fold Cross Validation* sehingga setiap bagian data dapat digunakan secara bergantian sebagai data latih dan data uji. Setelah proses pembagian data, dilakukan pembuatan model menggunakan algoritma *K-Nearest Neighbor* (KNN) dengan metrik jarak *Manhattan*. Model yang telah dibangun kemudian dievaluasi untuk menentukan nilai *K* yang optimal. Apabila nilai *K* yang digunakan belum menghasilkan performa terbaik, maka dilakukan penyesuaian nilai *K* dan proses pemodelan diulang hingga diperoleh nilai yang optimal. Setelah nilai *K* optimal diperoleh, model digunakan untuk melakukan klasifikasi risiko diabetes mellitus. Hasil klasifikasi selanjutnya dievaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score* untuk mengukur kinerja model. Tahap akhir dari proses ini adalah menghasilkan diagnosis risiko diabetes mellitus yang dapat digunakan sebagai pendukung pengambilan keputusan dalam deteksi dini penyakit.

3 | Hasil dan Pembahasan

Hasil penelitian menunjukkan bahwa penerapan algoritma *K-Nearest Neighbor* (KNN) dengan metode *K-Fold Cross Validation* mampu mengklasifikasikan risiko diabetes mellitus dengan baik. Proses *preprocessing* data yang meliputi pembersihan data, normalisasi, dan seleksi fitur berkontribusi dalam meningkatkan kualitas data sehingga menghasilkan model yang lebih optimal. Pengujian dilakukan dengan beberapa variasi nilai *K* untuk memperoleh performa terbaik, di mana hasil evaluasi menunjukkan bahwa model memiliki tingkat akurasi, *precision*, *recall*, dan *F1-score* yang baik serta konsisten pada setiap *fold*. Hal ini menunjukkan bahwa metode KNN dengan *K-Fold Cross Validation* efektif digunakan sebagai model klasifikasi dalam membantu deteksi dini risiko diabetes mellitus.

3.1 | Analisis Distribusi Kelas Status Gula Darah

Analisis distribusi kelas dilakukan untuk mengetahui keseimbangan data pada setiap kelas status gula darah, karena ketidakseimbangan kelas dapat memengaruhi kinerja model klasifikasi. Visualisasi distribusi kelas status gula darah ditunjukkan pada **Gambar 3**.

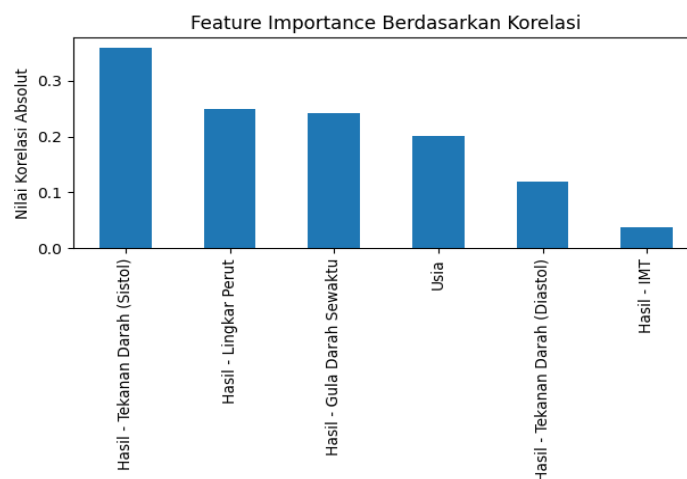


GAMBAR 3. Grafik Distribusi Kelas Status Gula Darah

Berdasarkan **GAMBAR 3**, terlihat bahwa *dataset* memiliki distribusi kelas yang tidak seimbang (*class imbalance*). Kelas dengan label 2 merupakan kelas dominan dengan jumlah data sekitar 260 data ($\pm 65\%$), sedangkan kelas 0 memiliki sekitar 80 data ($\pm 20\%$), dan kelas 1 merupakan kelas minoritas dengan jumlah sekitar 60 data ($\pm 15\%$). Kondisi ini menunjukkan bahwa sebagian besar pasien dalam *dataset* berada pada kategori status gula darah tertentu, sementara kategori lainnya relatif kurang terwakili. Ketidakseimbangan kelas ini berpotensi menyebabkan model KNN lebih cenderung memprediksi kelas mayoritas. Akibatnya, meskipun nilai akurasi keseluruhan dapat terlihat tinggi, performa model dalam mengklasifikasikan kelas minoritas dapat menjadi kurang optimal. Hal ini terutama berdampak pada metrik evaluasi per kelas, seperti *recall* dan *F1-score*, yang cenderung rendah pada kelas dengan jumlah data yang sedikit. Untuk meminimalkan pengaruh ketidakseimbangan kelas tersebut, penelitian ini menerapkan metode *Stratified K-Fold Cross Validation*, di mana proporsi kelas pada setiap *fold* dijaga tetap sama seperti distribusi data asli. Pendekatan ini memungkinkan evaluasi model yang lebih adil dan representatif terhadap seluruh kelas status gula darah.

3.2 | Analisis Feature Importance

Analisis *feature importance* dilakukan untuk mengetahui kontribusi masing-masing fitur terhadap proses klasifikasi status gula darah. Pendekatan yang digunakan adalah analisis korelasi antara fitur *input* dan variabel target, yang hasilnya divisualisasikan pada **Gambar 4**.

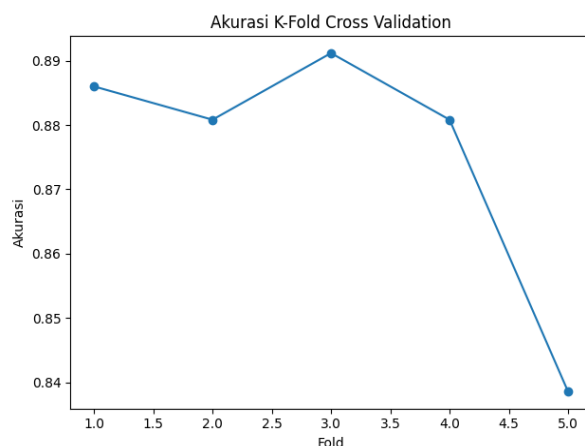


GAMBAR 4. Feature Importance

Berdasarkan **Gambar 4**, tekanan darah sistolik memiliki korelasi tertinggi terhadap status gula darah ($\pm 0,52$), diikuti lingkar perut ($\pm 0,45$) dan gula darah sewaktu ($\pm 0,42$), yang menunjukkan peran penting faktor kardiovaskular dan obesitas sentral. Usia menunjukkan korelasi sedang ($\pm 0,31$), sedangkan indeks massa tubuh (IMT) memiliki korelasi terendah ($\pm 0,15$). Secara keseluruhan, hasil ini menunjukkan bahwa hipertensi dan distribusi lemak tubuh lebih dominan memengaruhi status gula darah, sejalan dengan teori bahwa keduanya merupakan faktor utama risiko diabetes mellitus.

3.3 | Evaluasi Model Menggunakan *K-Fold Cross Validation*

Evaluasi kinerja model KNN dilakukan menggunakan metode *Stratified K-Fold Cross Validation* dengan jumlah *fold* sebanyak lima. Grafik akurasi hasil pengujian ditunjukkan pada **GAMBAR 5**.

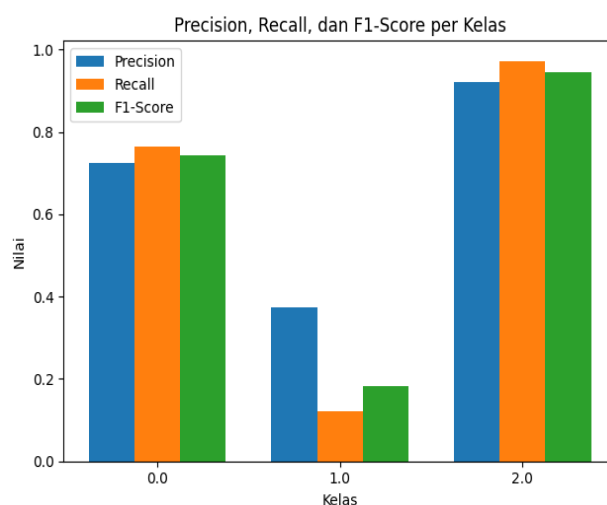


GAMBAR 5. Grafik Akurasi *K-Fold Cross Validation*

Berdasarkan **Gambar 5**, nilai akurasi pada setiap *fold* berada pada rentang 0,84 hingga 0,89. Akurasi tertinggi diperoleh pada *fold* ketiga dengan nilai 0,8912, sedangkan akurasi terendah terjadi pada *fold* kelima sebesar 0,8385. Variasi nilai akurasi antar *fold* menunjukkan adanya perbedaan karakteristik data uji, yang merupakan kondisi umum pada data kesehatan dengan tingkat keragaman karakteristik pasien yang tinggi. Meskipun terdapat fluktuasi, sebagian besar *fold* menghasilkan nilai akurasi di atas 0,88, yang menunjukkan bahwa model KNN memiliki kemampuan klasifikasi yang cukup baik dan stabil. Nilai rata-rata akurasi yang diperoleh adalah sebesar 0,8755 (87,55%), yang mengindikasikan bahwa kombinasi parameter KNN dengan jumlah tetangga terdekat sebanyak lima dan metrik jarak *Manhattan* mampu memberikan performa klasifikasi yang konsisten.

3.4 | Analisis Hasil Klasifikasi Deteksi Anomali

Untuk memperoleh gambaran kinerja model yang lebih komprehensif, evaluasi juga dilakukan menggunakan metrik *precision*, *recall*, dan *F1-score* yang disajikan pada **Gambar 6**.



GAMBAR 6. Grafik *Precision Recall* dan *F1-Score*

Berdasarkan **Gambar 6**, Pada kelas 0, diperoleh nilai *precision* sebesar 0,72, *recall* 0,76, dan *F1-score* 0,74. Hasil ini menunjukkan bahwa model cukup baik dalam mengklasifikasikan kelas 0, meskipun masih terdapat beberapa kesalahan prediksi. Pada kelas 1, performa model relatif rendah dengan nilai *precision* 0,38, *recall* 0,12, dan *F1-score* 0,18. Nilai *recall* yang sangat rendah mengindikasikan bahwa sebagian besar data aktual kelas 1 tidak berhasil dikenali oleh model. Kondisi ini dipengaruhi oleh jumlah data kelas 1 yang paling sedikit, sehingga tetangga terdekat dalam algoritma KNN lebih didominasi oleh kelas mayoritas. Sebaliknya, kelas 2 menunjukkan performa terbaik dengan nilai *precision* 0,92, *recall* 0,97, dan *F1-score* 0,94. Nilai ini menunjukkan bahwa model sangat akurat dan konsisten dalam mengklasifikasikan kelas mayoritas. Secara keseluruhan, hasil evaluasi menunjukkan bahwa model KNN memiliki kinerja yang sangat baik pada kelas mayoritas, cukup baik pada kelas 0, namun masih kurang optimal pada kelas minoritas akibat ketidakseimbangan distribusi data. Analisis *feature importance* menunjukkan bahwa tekanan darah sistolik, lingkaran perut, dan gula darah sewaktu merupakan faktor yang paling berpengaruh terhadap status gula darah pasien. Meskipun demikian, hasil penelitian ini menegaskan bahwa KNN berpotensi digunakan sebagai alat bantu klasifikasi risiko diabetes, dengan catatan diperlukan pengembangan lebih lanjut untuk meningkatkan kinerja model pada kelas minoritas.

4 | KESIMPULAN

Berdasarkan hasil penelitian, algoritma *K-Nearest Neighbor* (KNN) dengan metrik jarak *Manhattan* dan evaluasi *Stratified K-Fold Cross Validation* mampu memberikan performa klasifikasi yang cukup baik dalam memprediksi status risiko diabetes mellitus, dengan nilai rata-rata akurasi sebesar 87,55%. Model menunjukkan kinerja yang sangat baik pada kelas mayoritas, cukup baik pada kelas 0, namun masih kurang optimal pada kelas minoritas akibat ketidakseimbangan distribusi data. Analisis *feature importance* menunjukkan bahwa tekanan darah sistolik, lingkaran perut, dan gula darah sewaktu merupakan faktor yang paling berpengaruh terhadap status gula darah pasien. Meskipun demikian, hasil penelitian ini menegaskan bahwa KNN berpotensi digunakan sebagai alat bantu klasifikasi risiko diabetes, dengan catatan diperlukan pengembangan lebih lanjut untuk meningkatkan kinerja model pada kelas minoritas.

Daftar Pustaka

- Afolabi, S., Ajadi, N., Jimoh, A., & Adenekan, I. (2025). Predicting diabetes using supervised machine learning algorithms on E-health records. *Informatics and Health*, 2(1), 9–16. <https://doi.org/https://doi.org/10.1016/j.infh.2024.12.002>
- Al-Khamees, H. A. A., Sani, N. S., Gifal, A. S., Liu, L. X. W., & Esa, M. I. (2025). A dynamic model using k-NN algorithm for predicting diabetes and breast cancer. *Computers in Biology and Medicine*, 192, 110276. <https://doi.org/https://doi.org/10.1016/j.combiomed.2025.110276>
- Ampofo, A. G., & Boateng, E. B. (2020). Beyond 2020: Modelling obesity and diabetes prevalence. *Diabetes Research and Clinical Practice*, 167, 108362. <https://doi.org/https://doi.org/10.1016/j.diabres.2020.108362>
- Amri, Z., Rodi, M., Wathani, M. N., Bagja, A., & No, V. (2025). *Infotek : Jurnal Informatika dan Teknologi Prediksi Diabetes Menggunakan Algoritma K-Nearest (KNN) Teknik SMOTE-ENN* Infotek : Jurnal Informatika dan Teknologi Diabetes merupakan penyakit tidak menular yang secara serius memengaruhi sistem kesehatan besa. 8(1), 193–204.
- Deberneh, H. M., & Kim, I. (2021). Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. In *International Journal of Environmental Research and Public Health* (Vol. 18, Issue 6, p. 3317). <https://doi.org/10.3390/ijerph18063317>
- Dritsas, E., & Trigka, M. (2022). Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. *Sensors*, 22(14). <https://doi.org/10.3390/s22145304>
- Fang, M., & Selvin, E. (2021). Thirty-Year Trends in Complications in U.S. Adults With Newly Diagnosed Type 2 Diabetes. *Diabetes Care*, 44(3), 699–706. <https://doi.org/10.2337/dc20-2304>
- Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & Metabolic Syndrome*, 13(1), 148. <https://doi.org/10.1186/s13098-021-00767-9>
- Lachin, J. M., Bebu, I., & Nathan, D. M. (2021). The Beneficial Effects of Earlier Versus Later Implementation of Intensive Therapy in Type 1 Diabetes. *Diabetes Care*, 44(10), 2225–2230. <https://doi.org/10.2337/dc21-1331>
- Liu, J., Ren, Z.-H., Qiang, H., Wu, J., Shen, M., Zhang, L., & Lyu, J. (2020). Trends in the incidence of diabetes mellitus: results from the Global Burden of Disease Study 2017 and implications for diabetes mellitus prevention. *BMC Public Health*, 20(1), 1415. <https://doi.org/10.1186/s12889-020-09502-x>
- mithu, joanish. (2023). Type 2 Diabetes Prediction using K-Nearest Neighbor Algorithm. *Journal of Trends in Computer Science and Smart Technology*. <https://doi.org/10.36548/JTCSST.2023.2.007>
- Narasimharao, M., Swain, B., Priyadarshi, R., Nayak, P. P., & Bhuyan, S. (2026). A Survey of Machine Learning Techniques for Diabetes Prediction: Current Trends and Future Directions. *Archives of Computational Methods in Engineering*, 33(4), 5187–5222. <https://doi.org/10.1007/s11831-025-10450-1>
- Ong, K. L., Stafford, L. K., McLaughlin, S. A., Boyko, E. J., Vollset, S. E., Smith, A. E., Dalton, B. E., Duprey, J., Cruz, J. A., Hagins, H., Lindstedt, P. A., Aali, A., Abate, Y. H., Abate, M. D., Abbasian, M., Abbasi-Kangevari, Z., Abbasi-

- Kangevari, M., Abd ElHafeez, S., Abd-Rabu, R., ... Vos, T. (2023). Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet*, 402(10397), 203–234. [https://doi.org/10.1016/S0140-6736\(23\)01301-6](https://doi.org/10.1016/S0140-6736(23)01301-6)
- Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C. N., Mbanya, J. C., Pavkov, M. E., Ramachandaran, A., Wild, S. H., James, S., Herman, W. H., Zhang, P., Bommer, C., Kuo, S., Boyko, E. J., & Magliano, D. J. (2022). IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, 183, 109119. <https://doi.org/10.1016/j.diabres.2021.109119>
- Villanueva, O., Espinola-Linares, K., Flores Castañeda, R. O., & Cabanillas-Carbonell, M. (2023). Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes. *Diagnostics (Basel, Switzerland)*, 13(14). <https://doi.org/10.3390/diagnostics13142383>