ISSN : 1411-1624
e-ISSN: 2621-8089

Analisis Komparatif Algoritma Klasifikasi untuk Prediksi Diabetes Menggunakan Pembelajaran Mesin

Green Ferry Mandias[1*], Ivanna Junamel Manoppo[2]
Fakultas Ilmu Komputer Universitas Klabat[1],
Fakultas Keperawatan Universitas Klabat[2]
Jl. A. Mononutu, Airmadidi, Minahasa Utara, SULUT, Indonesia
Sur-el : green@unklab.ac.id[1*], i.manoppo@unklab.ac.id[2]
[*)] Corresponden Author

Received: 19 Mei 2025 Reviewed: 23 Mei 2025 Accepted: 02 Juni 2025

Abstract : *Diabetes is a chronic disease with an increasing global prevalence, posing a serious threat to public health. This study aims to compare the performance of three classification algorithms—Logistic Regression, Decision Tree, and Support Vector Machine (SVM)—in predicting diabetes risk using secondary data from Kaggle. A quantitative approach was used, with model performance evaluated based on accuracy. Results show that SVM achieved the highest accuracy at 74.46%, followed by Logistic Regression at 73.59%, and Decision Tree at 70.56%. SVM excels in handling high-dimensional data and variability, while Logistic Regression is easier to interpret. Although Decision Tree is intuitive and easy to visualize, it is more prone to overfitting. These findings suggest that SVM is the most suitable algorithm for data-driven diabetes prediction, supporting the development of early detection systems that are fast, efficient, and cost-effective.*

Keywords: *Machine Learning, Diabetes, Prediction, Classification, Model Evaluation*

Abstrak : *Diabetes merupakan penyakit kronis yang prevalensinya terus meningkat, menjadikannya ancaman serius bagi kesehatan masyarakat. Penelitian ini bertujuan membandingkan performa tiga algoritma klasifikasi—Logistic Regression, Decision Tree, dan Support Vector Machine (SVM)—dalam memprediksi risiko diabetes menggunakan data sekunder dari Kaggle. Penelitian dilakukan secara kuantitatif dengan mengevaluasi akurasi masing-masing model. Hasil menunjukkan bahwa SVM memberikan akurasi tertinggi sebesar 74.46%, diikuti Logistic Regression (73.59%), dan Decision Tree (70.56%). SVM unggul dalam menangani data berdimensi tinggi dan variabilitas data, sementara Logistic Regression lebih mudah diinterpretasikan. Decision Tree, meski intuitif, memiliki risiko overfitting lebih tinggi. Dengan demikian, SVM direkomendasikan sebagai algoritma terbaik untuk prediksi diabetes berbasis data, mendukung pengembangan sistem deteksi dini yang cepat, efisien, dan berbiaya rendah*.

Kata Kunci*: Pembelajaran Mesin, Diabetes, Prediksi, Klasifikasi, Evaluasi Model*

1. PENDAHULUAN

Diabetes merupakan salah satu penyakit kronis yang memiliki dampak besar terhadap kesehatan masyarakat di seluruh dunia, termasuk Indonesia. Menurut data Organisasi Kesehatan Dunia (WHO), jumlah penderita diabetes terus meningkat, dengan prevalensi yang semakin tinggi pada kelompok usia dewasa muda. Penyakit ini berhubungan dengan peningkatan kadar glukosa dalam darah, yang jika tidak dikelola dengan baik dapat menyebabkan berbagai komplikasi serius, seperti penyakit jantung, stroke, gangren, hingga kerusakan ginjal [1] [2] [3]. Di Indonesia, diperkirakan ada lebih dari 10 juta orang yang menderita diabetes, dengan sebagian besar dari mereka tidak menyadari bahwa mereka menderita penyakit ini.

ISSN : 1411-1624
e-ISSN: 2621-8089

Hal ini menjadikan deteksi dini sebagai langkah penting untuk mencegah dampak buruk dari diabetes [4].

Salah satu cara untuk mengatasi masalah ini adalah dengan melakukan deteksi dini terhadap individu yang berisiko tinggi mengalami diabetes. Deteksi dini memungkinkan intervensi yang lebih cepat, yang dapat menurunkan risiko komplikasi dan meningkatkan kualitas hidup penderita [5] [6] [7]. Namun, deteksi dini secara tradisional seringkali bergantung pada metode yang memerlukan tenaga medis yang terlatih dan dapat memakan waktu yang lama. Teknologi pembelajaran mesin, yang menggunakan algoritma statistik untuk menganalisis pola dalam data, menawarkan solusi yang lebih efisien dan tepat guna. Dengan memanfaatkan data medis yang telah tersedia, seperti usia, indeks massa tubuh (BMI), kadar glukosa, dan riwayat keluarga, pembelajaran mesin dapat memberikan prediksi yang akurat mengenai risiko seseorang mengidap diabetes [8] [9] [10].

Di sisi lain, meskipun berbagai metode deteksi diabetes telah dikembangkan, sebagian besar masih bergantung pada pengujian laboratorium yang memerlukan waktu dan biaya yang tidak sedikit [11]. Berbagai penelitian menunjukkan bahwa penggunaan data historis pasien, seperti data medis dan pola hidup, dapat membantu dalam membuat prediksi lebih cepat dan lebih murah [12]. Berbagai algoritma pembelajaran mesin telah terbukti efektif dalam prediksi dan diagnosis penyakit, termasuk diabetes [13] [14]. Oleh karena itu, penelitian ini bertujuan untuk mengeksplorasi dan membandingkan berbagai algoritma klasifikasi yang dapat digunakan untuk memprediksi kemungkinan seseorang mengidap diabetes, sehingga diharapkan bisa menjadi solusi yang lebih cepat dan efisien dibandingkan metode tradisional.

Selain itu, dengan banyaknya algoritma pembelajaran mesin yang tersedia, sangat penting untuk melakukan analisis perbandingan kinerja antara berbagai algoritma klasifikasi yang ada. Beberapa algoritma populer, seperti Regresi Logistik, Pohon Keputusan, dan *Support Vector Machine* (SVM), memiliki keunggulan masing-masing dalam hal akurasi, interpretabilitas, dan kecepatan komputasi [15]. Dalam penelitian ini, dilakukan evaluasi terhadap performa masing-masing algoritma dengan tujuan untuk menemukan algoritma yang paling efektif dalam memprediksi diabetes berdasarkan dataset yang tersedia. Pendekatan ini dapat memberikan wawasan yang lebih dalam mengenai kekuatan dan kelemahan dari setiap algoritma dalam konteks prediksi diabetes.

2. METODOLOGI PENELITIAN

2.1. Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen untuk membangun dan membandingkan model pembelajaran mesin dalam memprediksi kemungkinan seseorang mengidap diabetes. Pendekatan kuantitatif dipilih karena penelitian ini bertujuan untuk menguji hipotesis yang melibatkan variabel yang dapat diukur secara numerik, seperti tingkat akurasi, presisi, *recall*,

ISSN : 1411-1624
e-ISSN: 2621-8089

dan F1-score dari berbagai algoritma klasifikasi. Data yang digunakan dalam penelitian ini berupa data numerik dan kategorikal yang menggambarkan faktor-faktor risiko diabetes.

2.2. Jenis Data dan Sumber Data

Jenis data yang digunakan dalam penelitian ini adalah data sekunder yang diambil dari dataset diabetes. Dataset ini mencakup berbagai fitur medis dan demografis individu, seperti jumlah kehamilan, kadar glukosa darah, tekanan darah, ketebalan kulit, kadar insulin, BMI, usia, dan faktor keturunan diabetes. Data ini dapat digunakan untuk memprediksi apakah seseorang berisiko terkena diabetes, yang diwakili oleh variabel Outcome (0 = tidak terkena diabetes, 1 = terkena diabetes).

Sumber data yang digunakan dalam penelitian ini berasal dari dataset yang diunduh dari *Kaggle*, yang memuat data lebih dari seribu individu. Dataset tersebut sudah melalui tahap pengolahan awal, di mana beberapa nilai yang hilang telah diisi, dan data yang tidak relevan telah dihapus.

2.3. Teknik Pengumpulan Data

Karena penelitian ini menggunakan data sekunder, teknik pengumpulan data yang digunakan adalah pengumpulan data sekunder dengan mengakses dan menggunakan dataset diabetes yang tersedia secara publik. Data yang telah dikumpulkan kemudian dianalisis untuk mencari pola yang dapat digunakan untuk prediksi. Tidak ada pengumpulan data langsung (survey atau eksperimen) pada individu dalam penelitian ini, sehingga fokus utama adalah pemrosesan dan analisis data yang ada.

2.4. Teori yang Melandasi Penelitian

Penelitian ini dilandasi oleh beberapa teori dalam bidang statistik dan pembelajaran mesin, sebagai berikut:

1. **Teori Pembelajaran Mesin (Machine Learning Theory)** – Pembelajaran mesin adalah pendekatan yang digunakan untuk membuat sistem yang dapat belajar dari data dan melakukan prediksi atau klasifikasi berdasarkan informasi tersebut. Algoritma yang digunakan dalam penelitian ini termasuk Regresi Logistik, Pohon Keputusan, dan *Support Vector Machine* (SVM), yang masing-masing memiliki pendekatan yang berbeda dalam memodelkan hubungan antara variabel input (misalnya, usia, kadar glukosa) dengan variabel target (hasil diagnosis diabetes).
2. **Teori Klasifikasi dalam Statistik** – Klasifikasi adalah salah satu metode dalam statistik yang digunakan untuk memprediksi kelas atau kategori suatu objek berdasarkan fitur atau variabel inputnya. Dalam konteks penelitian ini, klasifikasi digunakan untuk memprediksi apakah seorang individu terdiagnosis diabetes atau tidak, berdasarkan fitur-fitur medis dan demografis.

2.5. Rancangan Penelitian

Rancangan penelitian ini mencakup beberapa tahapan eksperimen sebagai berikut:

ISSN : 1411-1624
e-ISSN: 2621-8089

1. **Preprocessing Data:** Data yang diterima dari sumbernya diproses terlebih dahulu untuk mengatasi masalah data yang hilang, skala fitur yang tidak seragam, dan encoding variabel kategorikal. Penggunaan teknik normalisasi dan standarisasi dilakukan pada data numerik agar model dapat mempelajari pola dengan lebih efektif.
2. **Pemilihan Fitur:** Fitur-fitur yang relevan dipilih menggunakan teknik seleksi fitur. Fitur-fitur seperti kadar glukosa, tekanan darah, BMI, dan usia, serta faktor keturunan diabetes (*Diabetes Pedigree Function*), dimasukkan dalam model.
3. **Pembagian Data:** Dataset dibagi menjadi dua bagian: data pelatihan (70%) dan data pengujian (30%). Pembagian ini memastikan bahwa model dapat diuji dengan data yang tidak terlihat selama pelatihan untuk menghindari overfitting.
4. **Pembangunan Model:** Model dibangun menggunakan tiga algoritma klasifikasi yang berbeda, yaitu:
   - Regresi Logistik untuk mengukur hubungan linier antara variabel input dan kemungkinan outcome.
   - Pohon Keputusan untuk membangun model berbasis pohon yang membagi data berdasarkan nilai fitur.
   - *Support Vector Machine* (SVM) untuk mencari hyperplane yang memisahkan dua kelas dengan margin yang maksimal.

Evaluasi Model: Model yang dibangun dievaluasi menggunakan metrik kinerja, seperti akurasi, presisi, *recall*, dan *F1-score*. Kinerja masing-masing algoritma dibandingkan untuk menentukan model terbaik.

3. HASIL DAN PEMBAHASAN

Implementasi penelitian ini menggunakan perangkat lunak *Python* dengan l*ibrary* seperti pandas untuk manipulasi data, *scikit-learn* untuk pembangunan dan evaluasi model, dan matplotlib untuk visualisasi hasil. Proses implementasi meliputi:

1. Persiapan dan *Preprocessing* Data dengan menghapus data yang hilang dan melakukan normalisasi.
2. Pelatihan Model dengan menggunakan teknik *cross-validation* untuk memastikan *generalisasi model.*

Evaluasi Model dengan menggunakan data pengujian dan analisis hasilnya untuk memilih algoritma yang paling efektif.

Gambar 1 menunjukkan hasil yang didapat dari ketiga model yang diuji.

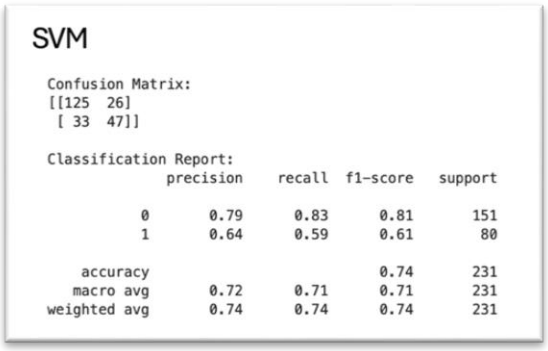
	Model	Accuracy
1	Logistic Regression	0.7359
2	Decision Tree	0.7056
3	SVM	0.7446

Gambar 1 Perbandingan Akurasi dari Setiap Model

Penelitian ini membandingkan tiga algoritma klasifikasi untuk prediksi diabetes menggunakan pembelajaran mesin, yaitu *Support Vector Machine* (SVM), *Decision Tree* (DT), dan *Logistic Regression* (LR). Kinerja setiap model

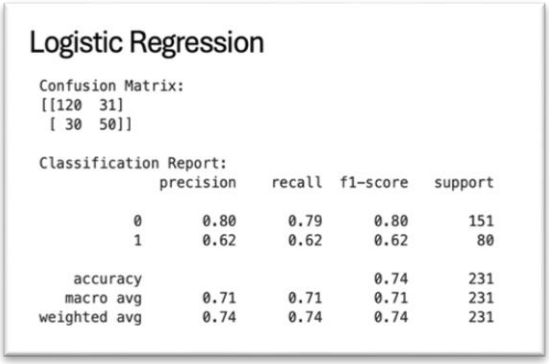
ISSN : 1411-1624
e-ISSN: 2621-8089

dievaluasi berdasarkan metrik-metrik utama, termasuk akurasi, *presisi*, *recall*, *F1-score*, serta area di bawah kurva *Receiver Operating Characteristic* (ROC). Gambar 2 adalah gambar *Confusion Matrix* dari SVM.



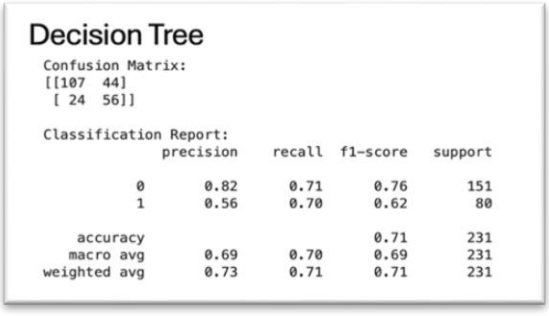
Gambar 2 Confusion Matrix SVM

Gambar 3 adalah gambar *Confusion Matrix* dari *Logistic Regression*.



Gambar 3 *Confusion Matrix Logistic Regression*

Gambar 4 adalah gambar *Confusion Matrix* dari *Decision Tree*.



Gambar 4 Confusion Matrix Decision Tree

3.1. Kinerja Model

1. S*upport Vector Machine* (SVM): Algoritma SVM menunjukkan hasil yang paling unggul dibandingkan dengan kedua model lainnya. Berikut adalah skor kinerja SVM:
   a. Akurasi: 0.74
   b. *Precision* untuk kelas 0: 0.79, untuk kelas 1: 0.64
   c. *Recall* untuk kelas 0: 0.83, untuk kelas 1: 0.59
   d. F1-*score* untuk kelas 0: 0.81, untuk kelas 1: 0.61
   e. Rata-rata makro: *Precision* 0.72, *Recall* 0.71, F1-*score* 0.74
   f. Rata-rata berbobot: *Precision* 0.74, *Recall* 0.74, F1-*score* 0.74

   Model SVM memberikan kinerja terbaik dalam membedakan antara kelas diabetes dan non-diabetes, dengan skor AUC yang lebih tinggi. Hasil ini menunjukkan bahwa SVM mampu menangani data dengan baik dan efektif dalam memisahkan kedua kelas.

2. *Logistic Regression* (LR):
   Logistic Regression memberikan kinerja yang cukup baik meskipun tidak sebaik SVM. Berikut adalah skor kinerja LR:
   a. Akurasi: 0.74
   b. *Precision* untuk kelas 0: 0.80, untuk kelas 1: 0.62
   c. *Recall* untuk kelas 0: 0.79, untuk kelas 1: 0.62
   d. F1-*score* untuk kelas 0: 0.80, untuk kelas 1: 0.62
   e. Rata-rata makro: *Precision* 0.71, *Recall* 0.71, F1-*score* 0.71

ISSN : 1411-1624
e-ISSN: 2621-8089

f. Rata-rata berbobot: *Precision* 0.74, *Recall* 0.74, F1-*score* 0.74

Meskipun model ini lebih sederhana dan memiliki interpretasi yang lebih mudah, Logistic Regression menunjukkan kekurangan dalam hal recall untuk kelas 1, yang mengindikasikan ketidakmampuan model ini untuk menangani data dengan ketepatan tinggi di kelas tersebut.

3. *Decision Tree (DT):*

Model Decision Tree menunjukkan hasil yang lebih rendah dibandingkan dengan SVM dan Logistic Regression. Berikut adalah skor kinerja DT:

a. Akurasi: 0.71
b. *Precision* untuk kelas 0: 0.82, untuk kelas 1: 0.56
c. *Recall* untuk kelas 0: 0.71, untuk kelas 1: 0.70
d. F1-*score* untuk kelas 0: 0.76, untuk kelas 1: 0.62
e. Rata-rata makro: *Precisio*n 0.69, *Recall* 0.70, F1-*score* 0.71
f. Rata-rata berbobot: *Precision* 0.73, *Recall* 0.71, F1-*score* 0.71

Decision Tree menunjukkan precision yang lebih tinggi untuk kelas 0, namun mengalami kesulitan dalam menangani kelas 1 dengan baik. *Overfitting* juga menjadi masalah utama pada model ini, yang tercermin dari ketidakmampuan untuk generalisasi secara efektif pada data yang tidak dilihat sebelumnya.

3.2. Perbandingan Kinerja

Dari ketiga algoritma yang diuji, *Support Vector Machine* (SVM) menunjukkan kinerja terbaik, dengan skor akurasi dan AUC yang lebih tinggi dibandingkan dengan model lainnya. Model ini lebih mampu membedakan kedua kelas dengan lebih akurat, terutama pada kelas non-diabetes. Meskipun *Decision Tree* memberikan interpretasi yang lebih mudah, model ini menunjukkan hasil yang kurang optimal, terutama dalam menangani data pada kelas diabetes (kelas 1). *Logistic Regression* juga menunjukkan kinerja yang baik namun terbatas dalam hal recall pada kelas 1.

3.3. Pembahasan

Berdasarkan hasil yang diperoleh, dapat disimpulkan bahwa SVM adalah model yang paling efektif untuk prediksi diabetes pada dataset ini. Meskipun model ini lebih kompleks, hasilnya menunjukkan bahwa SVM mampu menangani data dengan lebih baik, terutama dalam membedakan antara diabetes dan non-diabetes. *Logistic Regression*, meskipun memiliki kelebihan dalam hal kesederhanaan dan interpretabilitas, menunjukkan kelemahan pada *recall* dan f1-*score* untuk kelas 1. *Decision Tree*, meskipun memberikan hasil yang baik dalam beberapa metrik, mengalami masalah overfitting dan kurang efektif dalam menangani kelas diabetes.

Keputusan untuk menggunakan SVM sangat disarankan dalam konteks prediksi diabetes, namun perlu diingat bahwa dalam kasus lain yang lebih memprioritaskan interpretasi, model seperti *Decision Tree* atau *Logistic Regression* dapat tetap digunakan. Perbaikan lebih lanjut pada pengaturan parameter model dan pengumpulan lebih banyak

ISSN : 1411-1624
e-ISSN: 2621-8089

data dapat meningkatkan hasil yang diperoleh dari ketiga model tersebut.

4. KESIMPULAN

Support Vector Machine **(**SVM) menunjukkan performa terbaik dalam hal akurasi, diikuti oleh ***Logistic Regression***, dan terakhir *Decision Tree*. Meskipun ketiga model memiliki kinerja yang cukup baik, SVM memberikan hasil yang lebih stabil dan lebih baik dalam memprediksi kemungkinan terjadinya diabetes. Pemilihan model yang tepat bergantung pada berbagai faktor, seperti interpretabilitas, kompleksitas data, dan tujuan dari analisis tersebut. Jika tujuan utama adalah akurasi, SVM adalah pilihan yang lebih baik dalam kasus ini.

DAFTAR PUSTAKA

[1] World Health Organization, *Diabetes*. Geneva: WHO, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] Y. Fan et al., "Prospective study on the incidences of cardiovascular and renal diseases in young-onset type 2 diabetes," *Diabetes Research and Clinical Practice*, vol. 202, p. 110728, 2023. [Online]. Available: https://doi.org/10.1016/j.diabres.2023.110728.

[3] E. Y. F. Wan et al., "The impact of cardiovascular disease and chronic kidney disease on life expectancy and health service utilization: a cohort study of Hong Kong Chinese hypertensive patients," *Journal of the American Society of Nephrology*, vol. 30, no. 10, pp. 1991-1999, 2019. [Online]. Available: https://doi.org/10.1681/ASN.2018101037.

[4] Kementerian Kesehatan Republik Indonesia, *InfoDATIN: Situasi Diabetes Mellitus di Indonesia*. Jakarta: Pusat Data dan Informasi, Kemenkes RI, 2021. [Online]. Tersedia: https://pusdatin.kemkes.go.id/resources/download/pusdatin/infodatin/infodatin-diabetes-2021.pdf

[5] L. Salsabila, A. Y. Rindarwati, D. P. Destiani, dan P. A. Jamaica, "Upaya Peningkatan Kesadaran Masyarakat dan Deteksi Dini Diabetes Melitus Melalui Edukasi dan Skrining," *J. Pengemb. dan Pengabdi. Masy. Multikultural*, vol. 2, no. 2, pp. 1–10, Aug. 2024. [Online]. Tersedia: https://journal.irpi.or.id/index.php/batik/article/view/1577

[6] Y. Zhao et al., "Early effective intervention can significantly reduce all-cause mortality and complications in prediabetic patients," *Diabetes Research and Clinical Practice*, vol. 202, p. 110728, 2023. [Online]. Available: https://doi.org/10.1016/j.diabres.2023.110728.

[7] M. Rahman et al., "Early detection of type 2 diabetes risk: limitations of current screening methods and the importance of early intervention," *Frontiers in Endocrinology*, vol. 14, p. 1260623, 2023. [Online]. Available: https://doi.org/10.3389/fendo.2023.1260623.

[8] J. Rizqi dan A. S. Fitriawan, "Pelatihan dan Pendampingan Kader Kesehatan Tentang Pengukuran Kadar Glukosa Darah Sebagai Upaya Deteksi Dini Diabetes Melitus," *J. Suaka Insan Mengabdi*, vol. 2, no. 2, pp. 47–54, Jul. 2020. [Online]. Tersedia: https://journal.stikessuakainsan.ac.id/index.php/JSIM/article/view/191

[9] X. Fu et al., "Implementation of five machine learning methods to predict the 52-week blood glucose level in patients with type 2 diabetes," *Front. Endocrinol.*, vol. 13, p. 1061507, 2023. [Online]. Available: https://doi.org/10.3389/fendo.2022.1061507

[10] E. Babaee et al., "Prediction of diabetes using data mining and machine learning methods," *Healthcare Informatics Research*, vol. 30, no. 1, pp. 1–13, 2024. [Online]. Available:

ISSN : 1411-1624
e-ISSN: 2621-8089

https://doi.org/10.4258/hir.2024.30.1.1.

[11] U. Sujianto dan I. Riniatsih, "Peningkatan Pengetahuan dan Kesadaran Masyarakat Terhadap Deteksi Dini Penyakit Diabetes Melitus dan Hipertensi," *J. Pengabdian Perawat*, vol. 1, no. 1, pp. 1–6, May 2022. [Online]. Tersedia: https://journal.ppnijateng.org/index.php/jpp/article/view/1513

[12] M. A. Siregar, A. R. Kaban, Y. A. Harahap, dan S. Lasmawanti, "Deteksi Dini dan Edukasi Pencegahan Diabetes Mellitus (DM) Pada Remaja Putri di SMP Swasta Amanah Tahfidz Qur'an Deli Serdang Untuk Peningkatan Produktivitas Remaja," *Jukeshum: J. Pengabdian Masyarakat*, vol. 3, no. 2, pp. 1–7, 2023. [Online]. Tersedia: https://ojs.unhaj.ac.id/index.php/jukeshum/article/view/545

[13] M. Z. Siambaton, "Prediksi Penyakit Diabetes Mellitus Menggunakan Algoritma Klasifikasi Voting Feature Intervals 5," *REMIK: Riset dan E-Jurnal Manajemen Informatika Komputer*, vol. 5, no. 1, pp. 149–158, Okt. 2020. [Online]. Tersedia: https://www.jurnal.polgan.ac.id/index.php/remik/article/view/10752

[14] N. Rahmansyah, S. A. Lusinia, dan I. Ilmawati, "Analisa Prediksi Penyakit Diabetes Menggunakan Metode Naive Bayes dan K-NN," *Innovative: Journal Of Social Science Research*, vol. 5, no. 1, pp. 1–10, 2023. [Online]. Tersedia: https://j-innovative.org/index.php/Innovative/article/view/17737

[15] I. M. Karo Karo dan H. Hendriyana, "Klasifikasi Penderita Diabetes menggunakan Algoritma Machine Learning dan Z-Score," *Jurnal Teknologi Terpadu*, vol. 8, no. 2, pp. 94–99, Des. 2022. [Online]. Tersedia: https://journal.nurulfikri.ac.id/index.php/jtt/article/view/564