JDMIS: Journal of Data Mining and Information Systems
Vol. 4, No. 1, 2026, p. 40~45
ISSN: ISSN: 2986-3473 | p-ISSN 2986-5271
DOI: 10.54259/jdmis.v4i1.7107

Prediksi Diabetes Berbasis Decision Tree Dengan Menggunakan Dataset Pima Indians Diabetes

Yustri Insani[1], Marcel Filemon Naibaho[2], Sardo Pardingotan Sipayung[3]
[1,2,3]Universitas Katolik Santo Thomas, Medan, Indonesia
Email: [1]berasayustriinsani@gmail.com, [2]marcel01052005@gmail.com, [3]pinsarsiphom@gmail.com

ABSTRAK

Diabetes melitus merupakan penyakit kronis yang ditandai dengan meningkatnya kadar glukosa dalam darah dan dapat menimbulkan berbagai komplikasi serius apabila tidak ditangani sejak dini. Penelitian ini bertujuan untuk memprediksi penyakit diabetes menggunakan algoritma *Decision Tree* dengan *dataset* Pima Indians Diabetes. Tahapan penelitian meliputi pengolahan data, pembentukan model Decision Tree menggunakan kriteria entropy, serta evaluasi kinerja model. Hasil penelitian menunjukkan bahwa model mencapai akurasi sebesar 76,62%. Pengujian melalui *confusion matrix* menghasilkan 83 sampel *True Negative,* 35 sampel *True Positive,* 16 sampel *False Positive,* dan 20 sampel *False Negative.* Atribut Glukosa ditemukan sebagai faktor paling dominan dalam diagnosis, diikuti oleh BMI dan *Age.* Model yang dihasilkan mampu membentuk aturan keputusan yang jelas dan mudah dipahami sehingga dapat digunakan sebagai sistem pendukung keputusan dalam diagnosis awal diabetes.

Kata Kunci: Diabetes, Pohon Keputusan, Pembelajaran Mesin, Klasifikasi, Kemampuan interpretasi

ABSTRACT

Diabetes mellitus is a chronic disease characterized by increased blood glucose levels and can lead to various serious complications if not treated early. This research aims to predict diabetes using the Decision Tree algorithm with the Pima Indians Diabetes dataset. The research stages include data processing, forming a Decision Tree model using the entropy criterion, and evaluating model performance. The results show that the model achieved an accuracy of 76.62%. Testing through a confusion matrix produced 83 True Negative samples, 35 True Positive samples, 16 False Positive samples, and 20 False Negative samples. The Glucose attribute was found to be the most dominant factor in the diagnosis, followed by BMI and Age. The resulting model is able to form clear and easy-to-understand decision rules so that it can be used as a decision support system in the early diagnosis of diabetes.

Keywords: *Diabetes, Decision Tree, Data Mining, Machine Learning, Classification, Interpretability*

Penulis Korespondensi:

Marcel Filemon Naibaho
Email: marcel01052005@gmail.com

Article Info

Diterima: 28 Januari 2026
Direvisi: 2 Februari 2026
Disetujui: 2 Februari 2026

1. PENDAHULUAN

Data mining merupakan teknik pemrosesan data yang bertujuan mengungkap pola tersembunyi dalam *dataset. Output* dari proses ini bisa dimanfaatkan untuk membuat keputusan strategis ke depannya. Teknik ini juga sering disebut sebagai pengenalan pola [1].

Diabetes melitus adalah penyakit tidak menular yang telah menjadi isu kesehatan dunia dengan tingkat kejadian yang terus naik. Organisasi Kesehatan Dunia (WHO) menggambarkan diabetes sebagai kondisi kronis yang disebabkan oleh peningkatan kadar gula darah, yang dapat menimbulkan komplikasi berat jika tidak segera diatasi [2]. Di Indonesia, penyakit ini berkontribusi besar terhadap peningkatan angka mortalitas. Walaupun diagnosis tradisional melalui pemeriksaan laboratorium memberikan

hasil yang tepat, pendekatan ini membutuhkan waktu lama dan biaya tinggi. Oleh karena itu, teknik *data mining dan machine learning* semakin populer untuk mempercepat dan menyederhanakan proses diagnosis.

Database Diabetes Suku Pima Indian memiliki signifikansi khusus karena mencerminkan situasi dunia nyata, sehingga memungkinkan para peneliti untuk menilai efektivitas model pembelajaran mendalam dalam meramalkan risiko diabetes pada populasi yang sangat rentan [3]. Berbagai kajian sebelumnya telah menggunakan algoritma *machine learning* untuk meramalkan diabetes. Algoritma *Decision Tree* dikenal efektif dalam klasifikasi karena menghasilkan model yang mudah dijelaskan dan dipahami [4].

Beberapa penelitian sebelumnya telah menerapkan algoritma *machine learning* untuk prediksi diabetes [5]. Whitten et al. menjelaskan bahwa data mining memungkinkan penggalian pola dari data kesehatan dalam jumlah besar untuk mendukung pengambilan keputusan medis [6]. Quinlan memperkenalkan algoritma Decision Tree sebagai metode klasifikasi yang mampu menghasilkan model keputusan yang bersifat interpretatif dan mudah dipahami [7]. Algoritma C4.5 adalah metode yang bisa diterapkan untuk membuat pohon keputusan. Pohon keputusan atau *Decision Tree,* termasuk salah satu teknik yang relatif sederhana bagi manusia untuk memahaminya. Ini merupakan model prediktif yang memanfaatkan bentuk pohon atau hierarki. Inti dari pohon keputusan adalah mengonversi data menjadi struktur pohon dan seperangkat aturan keputusan [8]. Penelitian ini juga menunjukkan bahwa algoritma *machine learning* seperti *Decision Tree, Naïve Bayes,* dan *Support Vector Machine* dapat digunakan untuk memprediksi diabetes dengan tingkat akurasi yang cukup baik [9]. Meskipun demikian, sebagian penelitian sebelumnya lebih berfokus pada peningkatan akurasi model tanpa membahas secara mendalam interpretasi hasil dan faktor-faktor utama yang mempengaruhi keputusan model. Selain itu, masih terdapat keterbatasan dalam penelitian yang mengaitkan hasil klasifikasi secara langsung dengan kondisi medis yang mudah dipahami oleh pengguna non-teknis. Oleh karena itu, diperlukan penelitian yang tidak hanya menekankan pada performa model, tetapi juga pada interpretabilitas dan kejelasan aturan keputusan yang dihasilkan.

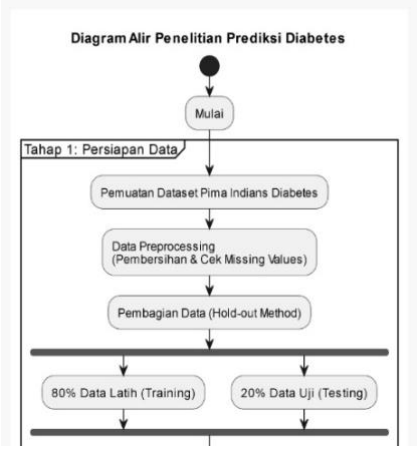
Dataset Pima Indians Diabetes merupakan *dataset* standar yang banyak digunakan dalam penelitian prediksi diabetes karena bersifat valid dan telah diuji secara luas dalam berbagai studi sebelumnya [10]. Namun, pemanfaatan *dataset* ini dengan fokus pada analisis struktur pohon keputusan dan identifikasi atribut paling berpengaruh masih relatif terbatas.

Berdasarkan uraian tersebut, beberapa peneliti sebelumnya berfokus pada penerapan berbagai algoritma *machine learning* untuk meningkatkan akurasi prediksi diabetes. Terdapat studi terbatas yang mengkaji interpretasi model *Decision Tree* secara mendalam untuk mengidentifikasi faktor-faktor utama penyebab diabetes berdasarkan *dataset* standar. Meskipun penelitian sebelumnya telah banyak menerapkan berbagai algoritma *machine learning* untuk prediksi diabetes dengan tingkat akurasi yang baik, sebagian besar studi tersebut cenderung berfokus pada optimasi performa model tanpa membahas secara mendalam interpretasi hasil serta faktor utama yang memengaruhi keputusan model. Oleh karena itu, penelitian ini hadir untuk mengisi celah tersebut dengan menekankan pada analisis struktur pohon keputusan menggunakan algoritma *Decision Tree,* sehingga tidak hanya menghasilkan prediksi yang akurat, tetapi juga memberikan transparansi mengenai alur logika diagnosis melalui visualisasi yang jelas. Tujuan penelitian ini adalah menghasilkan model prediksi yang akurat, mudah dipahami, serta dapat digunakan sebagai sistem pendukung keputusan dalam diagnosis awal penyakit diabetes.

2. METODE PENELITIAN

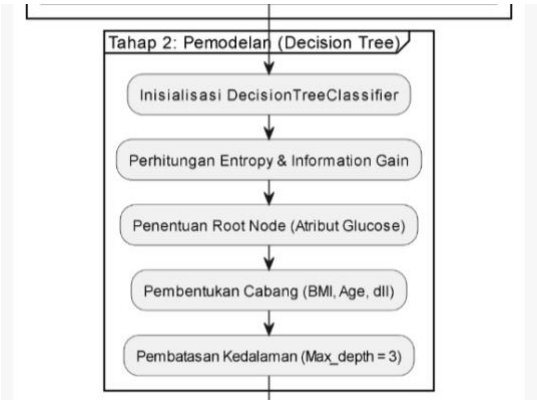
Metode penelitian yang digunakan dalam penelitian ini terdiri dari beberapa tahapan. Sesi ini menyediakan ulasan padat tentang inovasi yang telah dicapai dalam penggunaan teknologi. Orang yang hidup dengan diabetes menjadi target utama pemberitahuan dari model klasifikasi yang diusulkan, yang juga menyertakan kontribusi ke dalam kumpulan data diabetes. Tahap pertama adalah pengumpulan data, yaitu menggunakan *dataset* Pima Indians Diabetes yang berisi data medis pasien seperti *Glucose,* BMI, Age, dan atribut lainnya [11]. Dataset ini dipilih karena bersifat valid dan telah banyak digunakan dalam penelitian sebelumnya. Tahap kedua adalah *preprocessing* data yang meliputi pembersihan data dan penyesuaian format data agar dapat diproses oleh algoritma *Decision Tree.* Tahap selanjutnya adalah pembagian data menjadi data latih dan data uji untuk membangun dan menguji model Kajian ini memanfaatkan *dataset* sekunder Pima Indians Diabetes yang diunduh dari platform Kaggle. Dataset mencakup 768 entri data dengan 8 variabel medis sebagai prediktor independen dan 1 variabel target *(Outcome).* Komposisi kelas dalam *dataset* menunjukkan 500 sampel non-diabetes (65,1%) dan 268 sampel diabetes (34,9%) [12]. Sifat data ini mengindikasikan ketidakseimbangan ringan yang sering terjadi pada kasus kesehatan.

Prosedur dalam penelitian ini dirancang secara terstruktur melalui empat tahapan utama untuk menjamin keandalan model klasifikasi yang dihasilkan. Tahap awal difokuskan pada persiapan data, mulai dari pemuatan *dataset* Pima Indians Diabetes hingga proses *preprocessing* untuk menjamin kualitas data, yang dilanjutkan dengan pembagian *dataset* menggunakan metode *hold-out* (80% data latih dan 20% data uji) [13]. Selanjutnya, dilakukan proses pemodelan inti dengan menginisialisasi *Decision Tree Classifier* menggunakan kriteria entropi dan pembatasan kedalaman pohon (max_depth=3). Seluruh rangkaian alur kerja sistematis mulai dari persiapan hingga evaluasi kinerja model digambarkan secara mendetail dalam diagram alir penelitian.



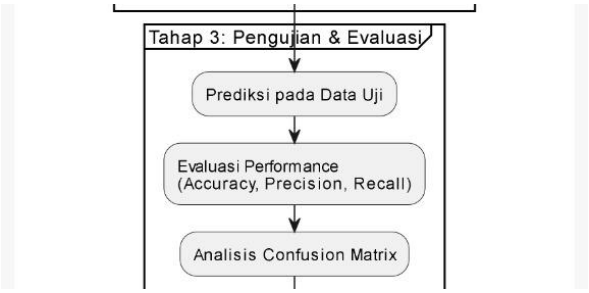
Gambar 1. Tahap 1 Persiapan data

Implementasi algoritma *Decision Tree* pada penelitian ini menghasilkan sebuah model visual yang merepresentasikan logika pengambilan keputusan diagnosa diabetes secara transparan. Berdasarkan hasil pemodelan terhadap 614 sampel data latih, atribut Glukosa teridentifikasi sebagai akar utama *(root node)* yang memiliki pengaruh paling signifikan dalam membedakan kelas pasien. Struktur pohon ini membagi data berdasarkan ambang batas tertentu, seperti nilai glukosa ≤ 127.5, serta pengaruh variabel pendukung lainnya seperti Age (Usia) dan BMI. Visualisasi ini memungkinkan tenaga medis untuk memahami pola klasifikasi diagnosa melalui serangkaian percabangan kondisi yang mudah diinterpretasikan.



Gambar 2. Tahap 2 Pemodelan

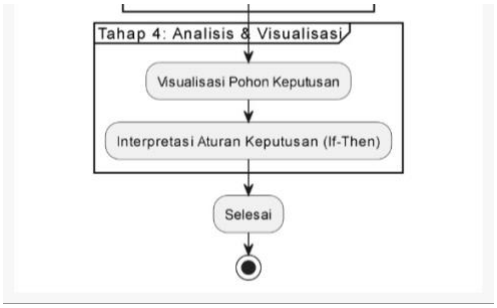
Setelah model berhasil dibangun pada tahap sebelumnya, langkah krusial berikutnya adalah mengukur sejauh mana kemampuan algoritma dalam melakukan prediksi secara akurat. Tahapan ini melibatkan penggunaan data uji *(testing data)* yang belum pernah dikenali oleh model untuk melihat performa klasifikasi di kondisi nyata. Evaluasi dilakukan secara komprehensif menggunakan metrik *Accuracy, Precision,* dan *Recall* guna memastikan model tidak hanya unggul secara angka, tetapi juga handal dalam mendeteksi kelas diabetes. Rincian langkah-langkah dalam proses pengujian dan evaluasi kinerja model ini disajikan secara sistematis.



Gambar 3. Tahap 3 Pengujian & Evaluasi

Tahap akhir dari metodologi ini adalah mentransformasi hasil komputasi algoritma ke dalam bentuk yang dapat diinterpretasikan oleh manusia, khususnya tenaga medis. Proses ini difokuskan pada pembuatan visualisasi pohon keputusan untuk mengidentifikasi variabel klinis yang paling berpengaruh terhadap risiko diabetes. Melalui interpretasi aturan keputusan (if-then rules), hasil prediksi yang kompleks diubah menjadi informasi yang transparan dan mudah dipahami sebagai dasar

pengambilan keputusan medis. Alur kerja pada tahap final yang mencakup visualisasi dan interpretasi aturan hasil klasifikasi ini diperlihatkan secara sistematis.



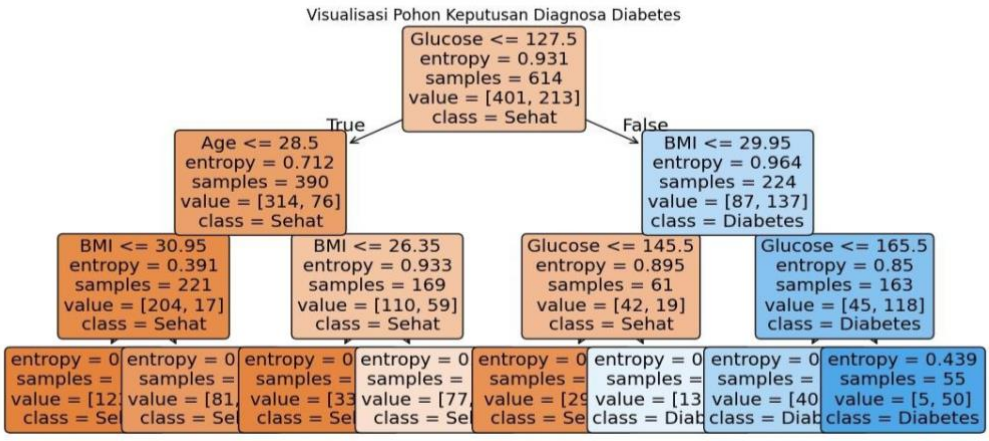
Gambar 4. Tahap 4 Analisis & Visualisasi

Pembagian Data (*Data Splitting)* menerapkan teknik *hold-out,* di mana *dataset* dipisah menjadi 80% set pelatihan *(training set)* dengan 614 sampel dan 20% set pengujian *(testing set)* dengan 154 sampel. Pemisahan dilakukan secara acak untuk mencegah bias pada evaluasi model.

Pembangunan Model: Menerapkan algoritma *Decision Tree* dengan entropi sebagai basis penghitungan pemisahan simpul. Visualisasi dan analisis, mengonversi model matematis menjadi representasi pohon keputusan untuk menemukan variabel yang paling signifikan. Eksperimen ini dikembangkan dengan bahasa pemrograman Python 3.14 [14]. *Library* utama yang diterapkan yaitu Pandas yaitu Untuk pengelolaan dan eksplorasi struktur data. Scikit-Learn (Sklearn) Untuk penerapan algoritma *Decision Tree*, pemisahan data *hold-out,* serta kalkulasi metrik evaluasi. Matplotlib Untuk penyajian visual hasil pohon keputusan dalam format grafik [15]. Untuk menilai performa model secara obyektif, diterapkan beberapa indikator evaluasi yang berasal dari *Confusion Matrix,* yakni Akurasi *(Accuracy)* Menghitung proporsi total ramalan yang tepat (sehat dan diabetes) terhadap seluruh data pengujian. Presisi *(Precision)* Menilai ketepatan model dalam meramalkan kelas positif (diabetes) relatif terhadap total hasil yang diramalkan positif. Sensitivitas (*Recall)* Mengukur kapasitas model untuk mendeteksi semua sampel yang benar-benar positif diabetes [16]. Indikator ini amat penting di bidang medis untuk mengurangi *false-negative.*

3. HASIL DAN PEMBAHASAN

Dari hasil pengolahan *dataset* Pima Indians Diabetes dengan algoritma *Decision Tree*, fitur Glukosa teridentifikasi sebagai elemen paling signifikan dalam menentukan diagnosis diabetes. Fitur ini dipilih sebagai simpul akar karena memiliki nilai *Information Gain* tertinggi dibandingkan yang lain. Secara klinis, hal ini menegaskan bahwa tingkat gula darah adalah penanda utama untuk mendeteksi gangguan metabolisme tubuh. Pohon keputusan yang dihasilkan dari proses *training* dibatasi pada kedalaman maksimal tiga level (max_depth=3) untuk mempertahankan kemudahan pemahaman dan menghindari *overfitting.* Visualisasi struktur logika tersebut disajikan pada gambar berikut.



Gambar 5. Hasil Pohon Keputusan Diagnosa Diabetes

Berdasarkan Gambar 5, logika klasifikasi dapat diuraikan melalui aturan keputusan *(decision rules)* berikut:

1. Simpul Akar *(Glucose):* Jika kadar glukosa pasien ≤ 127.5, model cenderung mengklasifikasikan sebagai "Sehat", meskipun masih mempertimbangkan usia dan BMI.
2. Faktor Usia *(Age):* Di kelompok glukosa rendah, individu dengan usia ≤ 28.5 tahun secara konsisten dikategorikan sebagai Sehat. Namun, untuk pasien yang lebih tua, BMI menjadi faktor tambahan.

3. Faktor Obesitas (BMI) Di kelompok dengan glukosa tinggi (> 127.5), BMI di atas 29.85 menjadi pemicu utama diagnosis Diabetes. Ini menunjukkan bahwa gabungan gula darah tinggi dan berat badan berlebihan adalah prediktor risiko yang sangat kuat.

Setelah menjelaskan mekanisme klasifikasi berdasarkan aturan keputusan yang telah diuraikan sebelumnya, langkah berikutnya melibatkan evaluasi tingkat akurasi aturan tersebut dalam mengantisipasi data yang belum dikenal. Proses pengujian ini dilakukan dengan menggunakan 154 sampel data pengujian yang tidak pernah terlibat dalam proses pelatihan model. Hasil evaluasi tersebut kemudian divisualisasikan dalam bentuk matriks penilaian untuk membandingkan *output* prediksi model dengan kondisi aktual pasien di dunia nyata. Penilaian kinerja ini krusial untuk memastikan bahwa model tidak hanya menawarkan kemudahan interpretasi visual, tetapi juga menunjukkan tingkat kepercayaan yang dapat dipertanggungjawabkan secara medis, sebagaimana tercermin dalam hasil analisis *Confusion Matrix* berikut.

Tabel 1. Confusion Matrix Hasil Prediksi Diabetes

	Prediksi: Sehat (0)	**Prediksi: Diabetes (1)**
Aktual: Sehat (0)	83 *(True Negative)*	16 *(False Positive)*
Aktual: Diabetes (1)	20 *(False Negative)*	35 *(True Positive)*

Berdasarkan data pada Tabel 1, dapat disimpulkan performa objektif dari model tersebut. Dari keseluruhan 99 sampel pasien yang sebenarnya tidak menderita diabetes, model mampu mengidentifikasi 83 di antaranya dengan tepat. Di sisi lain, dari 55 sampel pasien yang benar-benar terdiagnosis diabetes, model berhasil mengenali 35 sampel secara benar. Meskipun demikian, masih ada 20 kasus *False Negative,* yakni individu dengan diabetes yang diprediksi sebagai sehat, yang merupakan aspek krusial dalam penilaian klinis. Untuk menilai kinerja model secara komprehensif sesuai dengan kriteria pembelajaran mesin standar, data dari matriks tersebut diubah menjadi persentase metrik evaluasi dalam tabel berikut.

Tabel 2. Hasil Penilaian Performa Model

Indikator Evaluasi	**Hasil Perhitungan**
Akurasi (*Accuracy*)	76,62%
Presisi (*Precision*)	0,69
Sensitivitas (*Recall*)	0,64
F1-Score	0,66

Evaluasi kinerja yang tercermin dalam Tabel 2 mengungkapkan bahwa model mencapai tingkat akurasi sebesar 76,62%. Angka presisi 0,69 mengindikasikan bahwa ketika model mengklasifikasikan pasien sebagai penderita diabetes, keakuratan prediksinya adalah 69%. Adapun nilai *Recall* yang mencapai 0,64 menunjukkan kemampuan model untuk mendeteksi 64% dari total kasus diabetes dalam *dataset* pengujian. Walaupun masih ada peluang untuk meningkatkan sensitivitas agar mengurangi jumlah *false-negative,* secara umum model ini telah memenuhi harapan sebagai alat bantu keputusan awal yang jelas dan mudah dipahami oleh profesional kesehatan.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat di simpulkan bahwa pendekatan *supervised learning* dengan algoritma *Decision Tree* mampu digunakan secara efektif dalam memprediksi penyakit diabetes menggunakan *dataset* Pima Indians Diabetes. Proses pembentukan model melalui tahapan pengumpulan data, *preprocessing,* pembagian data menggunakan metode *hold-out,* serta pembangunan model berbasis entropi menghasilkan struktur pohon keputusan yang jelas dan mudah dipahami. Evaluasi model pada 154 sampel pengujian menunjukkan performa yang kuat, dengan tingkat akurasi mencapai 76,62%. Lebih lanjut, analisis melalui Matriks Kebingungan mengungkapkan 83 sampel Negatif Benar, 35 sampel Positif Benar, 16 sampel Positif Palsu, serta 20 sampel Negatif Palsu. Model ini juga mencatat nilai presisi sebesar 0,81 untuk kategori non-diabetes dan 0,69 untuk kategori diabetes.

Model yang dikembangkan mampu menghasilkan aturan klasifikasi yang sederhana dan mudah diinterpretasi, sehingga cocok sebagai alat bantu keputusan untuk diagnosis awal diabetes. Atribut Glukosa teridentifikasi sebagai prediktor utama, sebagaimana tercermin dari posisinya sebagai simpul akar dalam struktur pohon keputusan. Hal ini menawarkan keuntungan transparansi bagi profesional kesehatan untuk memahami keterkaitan logis antara indikator risiko medis. Secara umum, studi ini menegaskan potensi algoritma Pohon Keputusan sebagai instrumen deteksi dini yang dapat dijelaskan dengan baik.

Sebagai prospek pengembangan dan studi lanjut ke depannya, disarankan penerapan teknik evaluasi yang lebih mendalam seperti validasi silang *k-fold* guna meningkatkan stabilitas hasil. Penelitian lanjutan juga dapat membandingkan efektivitasnya dengan algoritma klasifikasi lainnya, seperti *Random Forest* atau *Support Vector Machine,* serta mengintegrasikan variasi atribut medis yang lebih luas. Transformasi model ini menjadi aplikasi komputer yang dapat diakses langsung oleh masyarakat akan berkontribusi besar dalam upaya pencegahan dan pengelolaan penyakit diabetes.

Kedua, dalam penelitian berikut, bisa dilakukan perbandingan performa antara algoritma *Decision Tree* dan algoritma klasifikasi lainnya seperti *Random Forest, Support Vector Machine,* atau *Naive Bayes.* Dengan perbandingan ini, diharapkan akan terlihat metode yang paling tepat untuk memprediksi penyakit diabetes berdasarkan karakteristik data yang digunakan.

Ketiga, jika jumlah data dan atribut medis yang lebih komprehensif ditambahkan—seperti riwayat keluarga, tekanan darah, dan kadar kolesterol—ini mungkin akan meningkatkan akurasi dan ketepatan model prediksi. Dengan data yang lebih bervariasi, model yang dihasilkan bisa lebih baik mencerminkan kondisi medis pasien secara keseluruhan.

Terakhir, penelitian ini bisa dikembangkan menjadi sistem atau aplikasi komputer yang dapat langsung digunakan oleh tenaga medis atau masyarakat sebagai alat bantu untuk diagnosis awal. Dengan pengembangan ini, diharapkan hasil penelitian dapat memberikan kontribusi signifikan dalam usaha pencegahan dan pengendalian penyakit diabetes.

REFERENSI

[1] R. Y. Averina and I. G. N. J. A. Widagda, "肖沉 1, 2, 孙莉 1, 2Δ, 曹杉杉 1, 2, 梁浩 1, 2, 程焱 1, 2," *Tjyybjb.Ac.Cn*, vol. 27, no. 2, pp. 635–637, 2021.

[2] J. B. Cole and J. C. Florez, "Genetics of diabetes mellitus and diabetes complications," *Nat. Rev. Nephrol.*, vol. 16, no. 7, pp. 377–390, 2020, doi: 10.1038/s41581-020-0278-5.

[3] A. Mousa, W. Mustafa, and R. B. Marqas, "A Comparative Study of Diabetes Detection Using The Pima Indian Diabetes Database," *J. Univ. Duhok*, vol. 26, no. 2, pp. 277–288, 2023, doi: 10.26682/suod.2023.26.2.24.

[4] B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 1, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[5] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[6] F. Ardyansyah, E. Daniati, and A. Ristyawan, "Pemanfaatan Data Mining untuk Analisis Keputusan," *Agustus*, vol. 8, pp. 2549–7952, 2024.

[7] S. Pewekar, M. Tirkey, A. Mallik, R. Shaikh, and S. A. Wagle, "Diabetes Prediction Using Machine Learning," *Lect. Notes Electr. Eng.*, vol. 1196 LNEE, no. 8, pp. 67–76, 2024, doi: 10.1007/978-981-97-7862-1_5.

[8] A. H. Nasrullah, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris," *J. Ilm. Ilmu Komput.*, vol. 7, no. 2, pp. 45–51, 2021, doi: 10.35329/jiik.v7i2.203.

[9] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved naive Bayes classification algorithm for traffic risk management," *EURASIP J. Adv. Signal Process.*, vol. 2021, no. 1, 2021, doi: 10.1186/s13634-021-00742-6.

[10] O. Y. Inonu, K. Magda, and A. Amarudin, "Analisis Kinerja Algoritma Random Forest Dengan Model Machine Learning Pada Dataset Penyakit Diabetes," *Expert J. Manaj. Sist. Inf. dan Teknol.*, vol. 15, no. 1, p. 1, 2025, doi: 10.36448/expert.v15i1.4312.

[11] Merdin Shamal Salih, "Diabetic Prediction based on Machine Learning Using PIMA Indian Dataset," *Commun. Appl. Nonlinear Anal.*, vol. 31, no. 5s, pp. 138–156, 2024, doi: 10.52783/cana.v31.1008.

[12] M. Kahn, "Diabetes," UCI Machine Learning Repository. [Online]. Available: https://doi.org/10.24432/C5T59G

[13] E. O. Manhitu, Y. P. K. Kelen, and D. Chrisinta, "Implementasi algoritma k-nearest neighbor untuk klasifikasi omset usaha mikro di kabupaten timor tengah utara," *Zo. J. Sist. Inf.*, vol. 7, no. 1, pp. 304–316, 2025.

[14] Putri and Nur, "Penggunaan Bahasa Python Untuk Analisis Dan Visualisasi Data Penduduk Di Desa Sumberjo, Nganjuk," *J. Pengabdi. Kpd. Masy.*, vol. 3, no. 3, pp. 206–217, 2023, [Online]. Available: https://jurnalfkip.samawa-university.ac.id/karya_jpm/index

[15] A. S. Saabith, T. Vinothraj, and M. Fareez, "A Review on Python Libraries and Ides for Data Science," *Int. J. Res. Eng. Sci. ISSN*, vol. 09, no. 11, pp. 36–53, 2021, [Online]. Available: www.ijres.org

[16] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.