
STUDI PERBANDINGAN ALGORITMA MACHINE LEARNING : SUPPORT VECTOR MACHINE, DECISION TREE DAN RANDOM FOREST DALAM KLASIFIKASI PENYAKIT DIABETES**Muhammad Shodiq*¹**¹Informatika Medis Universitas Muhammadiyah Lamongan
¹shodiqmuhammad13@gmail.com**Agus Priyono²**²Informatika Medis Universitas Muhammadiyah Lamongan
²aguspriyono2012@gmail.com**Neni Purwati³**³Informatika Medis Universitas Muhammadiyah Lamongan
²nenipurwati@umla.ac.id**ABSTRACT**

Hyperglycemia, or elevated blood glucose levels, is a primary indicator of diabetes mellitus, a chronic metabolic disorder whose prevalence continues to rise globally according to reports from the World Health Organization (WHO). Early detection of diabetes risk is crucial for preventing severe long-term complications. This study aims to evaluate and compare the performance of three machine learning algorithms Support Vector Machine (SVM), Decision Tree, and Random Forest in classifying diabetes based on health indicators and lifestyle patterns. The dataset used was obtained from Kaggle, with preprocessing stages including handling missing values and normalization. Model performance was assessed using accuracy, precision, recall, and F1-score. The results showed that the SVM algorithm achieved the highest accuracy of 74.89%, Precision of 72.37, Recall of 80.52 and F1-Score of 72.22 followed by Decision Tree with 73.39%, Precision of 71.34, Recall of 78.20 and F1-Score of 74.61 and Random Forest accuracy with 72.41%, Precision of 70.90, Recall of 76.02 and F1-Score of 73.37. This research is expected to be a reference for developing an early medical screening system to detect the risk of diabetes intelligently and accurately.

Keywords: Diabetes, Classification, Machine Learning, Support Vector Machine, Decision Tree, Random Forest

ABSTRAK

Hiperglikemia, atau peningkatan kadar glukosa darah, merupakan indikator utama diabetes melitus, suatu gangguan metabolisme kronis yang prevalensinya terus meningkat secara global menurut laporan dari Organisasi Kesehatan Dunia (WHO). Deteksi dini risiko diabetes sangat penting untuk mencegah komplikasi jangka panjang yang parah. Studi ini bertujuan untuk mengevaluasi dan membandingkan kinerja tiga algoritma pembelajaran mesin, yaitu Support Vector Machine (SVM), Decision Tree, dan Random Forest, dalam mengklasifikasikan diabetes berdasarkan indikator kesehatan dan pola gaya hidup. Dataset yang digunakan diperoleh dari Kaggle, dengan tahapan pra-pemrosesan termasuk penanganan nilai yang hilang dan normalisasi. Kinerja model dinilai menggunakan akurasi, presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa algoritma SVM mencapai akurasi tertinggi sebesar 74,89%, Presicion sebesar 72,37, Recall sebesar 80,52 dan F1-Score sebesar 72,22 diikuti oleh Decision Tree dengan 73,39%, Presicion sebesar 71,34, Recall sebesar 78,20 dan F1-Score sebesar 74,61 dan akurasi Random Forest dengan 72,41%, Presicion sebesar 70,90, Recall sebesar 76,02 dan F1-Score sebesar 73,37. Penelitian ini diharapkan dapat menjadi referensi untuk mengembangkan sistem skrining medis dini guna mendeteksi risiko diabetes secara cerdas dan akurat.

Kata Kunci: Diabetes, Klasifikasi, Machine Learning, Support Vector Machine, Decision Tree, Random Forest

1. PENDAHULUAN

Hiperglikemia adalah kondisi ketika kadar glukosa darah meningkat melampaui batas normal, yakni 80–90 mg/dl saat puasa atau 140–160 mg/100 ml pada kondisi non-puasa. Laporan World Health Organization (WHO) pada tahun 2023 mencatat bahwa lebih dari 422 juta orang di seluruh dunia hidup dengan diabetes melitus. Data International Diabetes Federation (IDF) tahun 2019 turut menunjukkan bahwa sekitar 483 juta individu berusia 20–79 tahun, atau 9,3% populasi global, menderita diabetes. Prevalensi meningkat lebih tinggi pada kelompok usia 65–79 tahun dengan jumlah mencapai 111,2 juta orang atau 19,9%. IDF juga memproyeksikan bahwa jumlah penyandang diabetes akan bertambah menjadi 578 juta pada tahun 2030 dan mencapai 700 juta pada tahun 2045 (Alfreyzal *et al.*, 2024).

Diabetes melitus merupakan gangguan metabolik kompleks yang ditandai oleh hiperglikemia kronis (Liren Li). Penyakit ini termasuk salah satu penyebab utama kematian di dunia dan memberikan dampak luas pada hampir seluruh keluarga secara global. Kondisi ini umumnya terjadi akibat kurangnya produksi insulin oleh pankreas. Meskipun diabetes dapat dikelola secara efektif apabila dideteksi dan ditangani sejak dini, hingga kini belum tersedia terapi kuratif jangka panjang (Zafar *et al.*, 2025).

Secara klinis, pasien diabetes sering mengalami polifagia, polidipsia, gangguan sendi, penurunan berat badan, serta penurunan imunitas. Deteksi dini melalui pengembangan model prediktif dinilai berperan penting untuk mencegah komplikasi berat dan memungkinkan intervensi klinis yang cepat. Upaya ini turut mendorong para ilmuwan di berbagai negara untuk mengembangkan sistem kesehatan berbasis teknologi yang mendukung diagnosis dini melalui identifikasi faktor risiko diabetes. Deteksi pada tahap awal secara signifikan meningkatkan peluang keberhasilan pengelolaan dibandingkan deteksi pada stadium lanjut (Zafar *et al.*, 2025).

Diabetes terbagi menjadi dua tipe utama, yaitu tipe 1 dan tipe 2. Diabetes tipe 1 bersifat lebih jarang dan dipengaruhi oleh faktor genetik. Sebaliknya, diabetes tipe 2 memiliki prevalensi yang jauh lebih tinggi dan dipengaruhi oleh kombinasi faktor genetik, lingkungan, serta gaya hidup, meskipun mekanisme biologis kemunculannya belum sepenuhnya dipahami (Chang *et al.*, 2022). Faktor demografis seperti usia, jenis kelamin, serta indeks massa tubuh (BMI), bersama dengan penggunaan obat-obatan seperti insulin, sulfonilurea, beta-blocker, angiotensin-converting enzyme inhibitor (ACEI), dan angiotensin II receptor blocker (ARB), turut berkontribusi terhadap perkembangan maupun pengelolaan diabetes (Li *et al.*, 2025).

Dalam beberapa tahun terakhir, kemajuan teknologi telah mempercepat adopsi pembelajaran mesin di bidang perawatan kesehatan, khususnya untuk

prediksi penyakit. Model pembelajaran mesin menunjukkan performa yang stabil dalam memprediksi serta mengklasifikasikan insiden diabetes (Keigley, Ayanian and Dugani, 2026). Analitik kesehatan berperan penting dalam mendeteksi dan mendiagnosis penyakit secara dini sehingga dapat meningkatkan kualitas layanan dan hasil klinis. Model pembelajaran mesin memungkinkan identifikasi pola dalam data dan menghasilkan prediksi yang berguna untuk diagnosis, prognosis, dan penentuan terapi. Dengan kemajuan algoritma dan teknologi, efektivitas model ini dalam mendukung pengobatan pasien semakin meningkat (Chang *et al.*, 2022). Oleh karena itu, solusi berbasis pembelajaran mesin dapat digunakan untuk peningkatan prediksi diabetes, sehingga memungkinkan perawatan yang lebih dini (Ivanovic *et al.*, 2026).

Penelitian sebelumnya, terkait hubungan antara paparan lingkungan perkotaan dan diabetes menggunakan random forest, Penelitian ini menunjukkan bahwa faktor sosial-ekonomi lingkungan seperti rendahnya nilai rumah, tingginya proporsi imigran dan suhu permukaan yang lebih tinggi konsisten dengan peningkatan risiko diabetes tipe 2. Analisis menunjukkan bahwa pengaruh faktor lingkungan lain bervariasi menurut metode, dengan LASSO memiliki akurasi prediksi terbaik, sementara Random Forest juga mengidentifikasi pola penting terutama pada variabel berpengaruh lemah (Ohyanan *et al.*, 2022). Selain itu, algoritma random forest memiliki akurasi 82% untuk klasifikasi penyakit diabetes (Chang *et al.*, 2022).

Penelitian lain, prediksi diabetes melitus menggunakan algoritma decision tree. Data diambil dari Puskesmas Mlati II Kabupaten Sleman. Hasil penelitian menunjukkan bahwa akurasi decision tree sebesar 92% (Aditya and Pramuntadi, 2024). Sedangkan penelitian lain pada prediksi penyakit diabetes menggunakan support vector machine radial basis function (SVM-RBF) menghasilkan nilai akurasi sebesar 91% (Wafa *et al.*, 2022).

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membandingkan model prediksi diabetes menggunakan SVM, Decision Tree dan Random Forest. Selanjutnya, model-model ini dibandingkan untuk menentukan model yang paling efektif dalam hal ini dengan mengevaluasi akurasi prediksinya, bersama dengan metrik kinerja lainnya seperti presisi, recall, dan skor F1. Diharapkan penelitian ini akan berkontribusi pada pengembangan sistem prediksi penyakit diabetes yang tidak hanya mencapai kinerja tinggi tetapi juga transparan dan mudah diinterpretasikan.

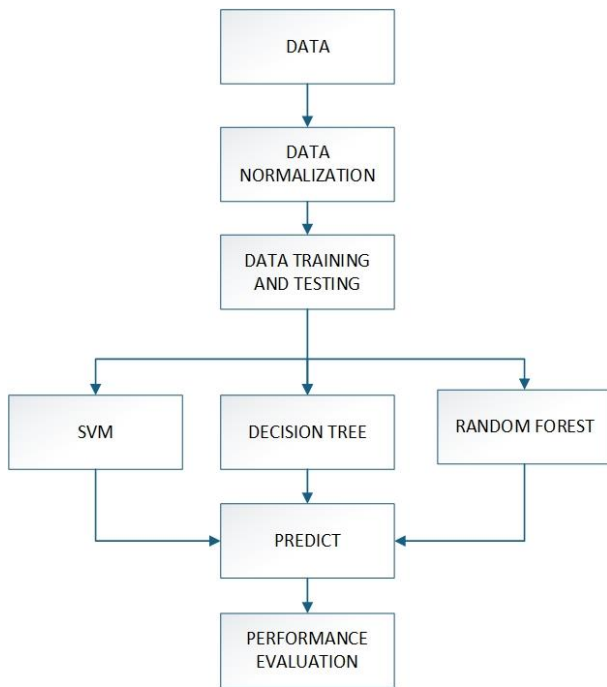
2. METODE PENELITIAN

Dalam penelitian ini, beberapa tahapan dilakukan untuk mengklasifikasikan penyakit diabetes. Tahapan

tersebut mencakup pengumpulan dataset diabetes, pra-pemrosesan data, pembuatan model SVM, Decision Tree dan Random Forest, serta pelatihan dan pengujian model untuk mengklasifikasikan apakah seseorang mengalami diabetes atau tidak. Selanjutnya, kinerja model dievaluasi dan dilakukan analisis perbandingan antar algoritma pembelajaran mesin SVM, Decision Tree dan Random Forest. Alur kerja penelitian ini diilustrasikan pada Gambar 1, yang menyajikan diagram alur metodologi dalam penelitian ini.

A. Data Diabetes

Dalam penelitian ini, data penyakit diabetes diperoleh dari kaggle. Data set ini berisi serangkaian informasi pasien dengan 18 variabel yang relevan untuk menganalisis kondisi seseorang termasuk dalam diabetes atau tidak. Gambar 2 menunjukkan dataset dari penyakit diabetes dari catatan informasi pasien.



Gambar 1. Alur Penelitian

Tabel 1. Dataset Penyakit Diabetes

Age	Sex	High Chol	Chol Chck	BMI	Smoker	HeartDiseaseorAttck	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	GenHlt	MentHlt	PhysHlt	DiffWalk	Stroke	HighbP	Diabetes
4	1	0	1	26	0	0	1	0	1	0	3	5	30	0	0	1	0
12	1	1	1	26	1	0	0	1	0	0	3	0	0	0	1	1	0
13	1	0	1	26	0	0	1	1	1	0	1	0	10	0	0	0	0
11	1	1	1	28	1	0	1	1	1	0	3	0	3	0	0	1	0
8	0	0	1	29	1	0	1	1	1	0	2	0	0	0	0	0	0
1	0	0	1	18	0	0	1	1	1	0	2	7	0	0	0	0	0
13	1	1	1	26	1	0	1	1	1	1	1	0	0	0	0	0	0
6	1	0	1	31	1	0	0	1	1	0	4	0	0	0	0	0	0
3	0	0	1	32	0	0	1	1	1	0	3	0	0	0	0	0	0
6	1	0	1	27	1	0	0	1	1	0	3	0	6	0	0	0	0
12	0	1	1	24	1	1	1	1	1	0	3	0	4	0	0	1	0
4	1	0	1	21	0	0	1	1	1	0	1	0	0	0	0	0	0
7	1	1	1	27	0	0	1	1	1	0	2	0	0	0	0	1	0
10	1	0	1	58	0	0	0	1	1	0	3	3	3	0	0	1	0
10	0	1	1	29	1	0	1	1	0	0	1	0	0	1	0	0	0
10	0	0	1	18	1	0	1	1	0	0	3	0	0	0	0	0	0
9	0	0	1	30	0	0	1	0	1	0	2	0	0	0	0	0	0
10	1	0	1	30	1	0	1	1	1	0	1	0	0	0	0	0	0

Age	Sex	HighChol	CholCheck	BMI	Smoker	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	GenHlt	MentHlt	PhysHlt	DiffWalk	Stroke	HighbP	Diabetes
8	0	0	1	20	0	0	1	1	1	0	2	0	0	0	0	0	0
7	1	0	1	26	0	0	0	1	1	0	3	0	15	0	0	1	0
6	0	0	1	22	0	0	1	1	1	0	1	0	0	0	0	0	0
10	1	0	1	29	1	0	1	1	1	0	5	0	30	0	0	1	0
...
13	0	1	1	25	0	1	0	1	0	0	5	15	0	1	0	1	1
11	0	1	1	18	0	0	0	0	0	0	4	0	0	1	0	1	1
9	0	1	1	25	0	1	1	1	0	0	2	0	0	0	0	1	1

Berdasarkan Gambar 2, dataset penyakit diabetes sebanyak 70.962 dataset yang terdiri dari 18 variabel/atribut, antara lain :

1. Diabetes sebagai variabel target dengan nilai 0 mewakili tidak diabetes dan 1 berarti diabetes.
2. Faktor demografi seperti usia, BMI dan jenis kelamin (0 mewakili perempuan dan 1 mewakili laki-laki).
3. Riwayat dan kondisi medis seperti HighBP (status hipertensi, jika 0 berarti tidak hipertensi dan 1 memiliki hipertensi), HighChol (status kolesterol, 0 berarti normal/rendah dan 1 berarti tinggi), CholCheck (riwayat pemeriksaan kolesterol 5 tahun terakhir, jika 0 berarti belum pernah cek dan 1 berarti sudah pernah cek), heartdiseaseorattack (riwayat penyakit jantung koroner, 0 artinya tidak pernah dan 1 artinya pernah) dan stroke nilai 0 artinya tidak pernah dan 1 artinya pernah.
4. Gaya hidup seperti smoker nilai 0 artinya perokok pasif/ tidak perokok dan 1 artinya perokok aktif, HvyAlcoholConsump artinya 0 tidak pernah konsumsi alkohol dan 1 diartikan konsumsi berat, PhysActivity artinya jika nilai 0 tidak pernah olahraga dan nilai 1 artinya aktif olahraga dan fruits serta veggies artinya jika nilai 0 tidak konsumsi buah/ sayur 1 kali sehari dan jika nilai 1 maka diartikan konsumsi buah/ sayur per hari.
5. Status kesehatan umum seperti GenHlt memiliki skala 1 (sangat baik) hingga 5 (sangat buruk), kemudian kesehatan mental, kesehatan fisik dan difficulty walking yang berarti pasien memiliki masalah dalam kesulitan berjalan, nilai 0 untuk tidak sulit dan nilai 1 untuk mengalami kesulitan.

Kombinasi antara riwayat medis (Hipertensi, Kolesterol), gaya hidup (Olahraga, Merokok, Diet), dan kondisi fisik (BMI, Usia) memiliki korelasi statistik yang sangat kuat dalam dunia medis untuk memetakan risiko diabetes seseorang.

B. Normalisasi Data

Dataset yang telah disusun selanjutnya diproses melalui tahap pra-pemrosesan untuk mengidentifikasi pola dasar dan mengatasi potensi masalah dalam data sebelum pelatihan model. Pada tahap ini juga diterapkan teknik normalisasi guna menyelaraskan skala setiap variabel, meminimalkan bias akibat perbedaan rentang nilai, serta meningkatkan stabilitas dan kinerja algoritma pembelajaran mesin. Dengan demikian, data yang dihasilkan menjadi lebih bersih, terstruktur, dan siap digunakan untuk proses pelatihan model secara optimal.

C. Pelatihan dan Pengujian

Dalam penelitian ini, studi perbandingan menggunakan 3 algoritma machine learning yang dilakukan dengan dua tahap yaitu tahap pelatihan dan pengujian. Dalam tahap pelatihan digunakan 80% data trainingset sedangkan 20% nya digunakan untuk data testing.

D. Model Machine Learning

1. Support Vector Machine

Support Vector Machine (SVM) memakai konsep dasar dari teori-teori komputasi sebelumnya, SVM baru dirumuskan secara utuh pada periode tersebut. Metode ini pada awalnya merupakan linear classifier, kemudian dikembangkan untuk menangani masalah non-linear melalui penggunaan kernel trick yang memetakan data ke ruang berdimensi lebih tinggi (Harahap, Muflikhah and Rahayudi, 2018). SVM merupakan metode yang banyak digunakan di berbagai bidang, seperti bioinformatika, pengenalan gambar, identifikasi suara, dan deteksi pola visual. Teknik klasifikasi ini mampu bekerja pada data linier maupun non-linier. Proses klasifikasi dilakukan dengan membangun sebuah hyperplane pada ruang berdimensi tinggi yang secara optimal memisahkan data ke dalam dua kelas, sehingga dapat diterapkan baik untuk tugas klasifikasi maupun regresi namun, SVM membutuhkan waktu perhitungan yang lambat saat memproses dataset yang besar (Chang *et al.*, 2022).

2. Decision Tree

Decision Tree merupakan metode klasifikasi berbasis aturan yang membentuk model prediksi dalam struktur pohon keputusan. Setiap simpul internal

merepresentasikan pengujian terhadap suatu atribut, sementara setiap cabang menunjukkan hasil dari pengujian tersebut. Decision Tree memiliki kelebihan serta keterbatasan dalam mengolah dataset dengan tingkat kompleksitas berbeda (Ramadhan and Veri, 2025) (Dritsas and Trigka, 2022).

3. Random Forest

Random Forest (RF) merupakan metode ensemble learning yang membangun sejumlah pohon keputusan melalui pemilihan acak subset prediktor dan observasi pada setiap iterasi. Prediksi dari seluruh pohon kemudian digabungkan untuk menghasilkan satu model yang lebih stabil. Untuk menilai kontribusi masing-masing variabel, digunakan permutation importance, sedangkan nilai Shapley dimanfaatkan untuk melihat arah dan kekuatan asosiasi variabel dalam model. Random Forest mampu menangkap hubungan yang bersifat nonlinier maupun non-aditif. Selain itu, perkembangan terbaru meningkatkan kemudahan dalam melakukan penyetelan parameter serta memperjelas interpretasi model (Dritsas and Trigka, 2022) (Ohanyan *et al.*, 2022).

E. Prediksi dan Evaluasi

Tahap evaluasi ini digunakan untuk mengukur seberapa akurat kinerja model. Keakuratan model klasifikasi seperti Decision Tree sering dievaluasi menggunakan metrik seperti accuracy, precision, recall, dan F1-score (Dritsas and Trigka, 2022; Ramadhan and Veri, 2025; Soladoye *et al.*, 2025). Berikut persamaan yang dapat digunakan pada proses evaluasi :

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (1)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (2)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (3)$$

$$\text{F1-score} = (2 * \text{Precision} * \text{recall}) / (\text{Precision} + \text{recall}) \quad (4)$$

Keterangan :

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

Selain itu, untuk mengukur keakuratan model algoritma juga dapat menggunakan Confusion matrix. Confusion matrix merupakan tabel yang menggambarkan jumlah prediksi benar dan salah yang dihasilkan model dalam mengklasifikasikan suatu data. Matriks ini berfungsi sebagai alat evaluasi yang penting untuk menilai kinerja model klasifikasi. Di dalamnya terdapat empat komponen utama, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) (Baswedan, 2024).

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini sesuai dengan pada alur penelitian yang digambarkan pada Gambar 1. Tahap pertama pengumpulan dataset penyakit diabetes yang diperoleh dari kaggle. Kemudian tahap kedua yaitu melakukan normalisasi dataset dengan menangani nilai data yang tidak lengkap dan mengubah variabel string menjadi variabel numerik.

Selanjutnya, dataset dibagi menjadi 2 yaitu 80% data latih dan 20% data uji. Data latih digunakan untuk mempelajari pola dan karakteristik dari data untuk menghasilkan klasifikasi penyakit diabetes. Untuk data uji digunakan untuk mengevaluasi kinerja algoritma dalam proses klasifikasi penyakit diabetes.

Tahap keempat, melakukan klasifikasi penyakit diabetes dengan 3 algoritma machine learning. Berikut ini hasil kinerja data latih dari algoritma SVM, Decision Tree, Random Forest Tabel 2.

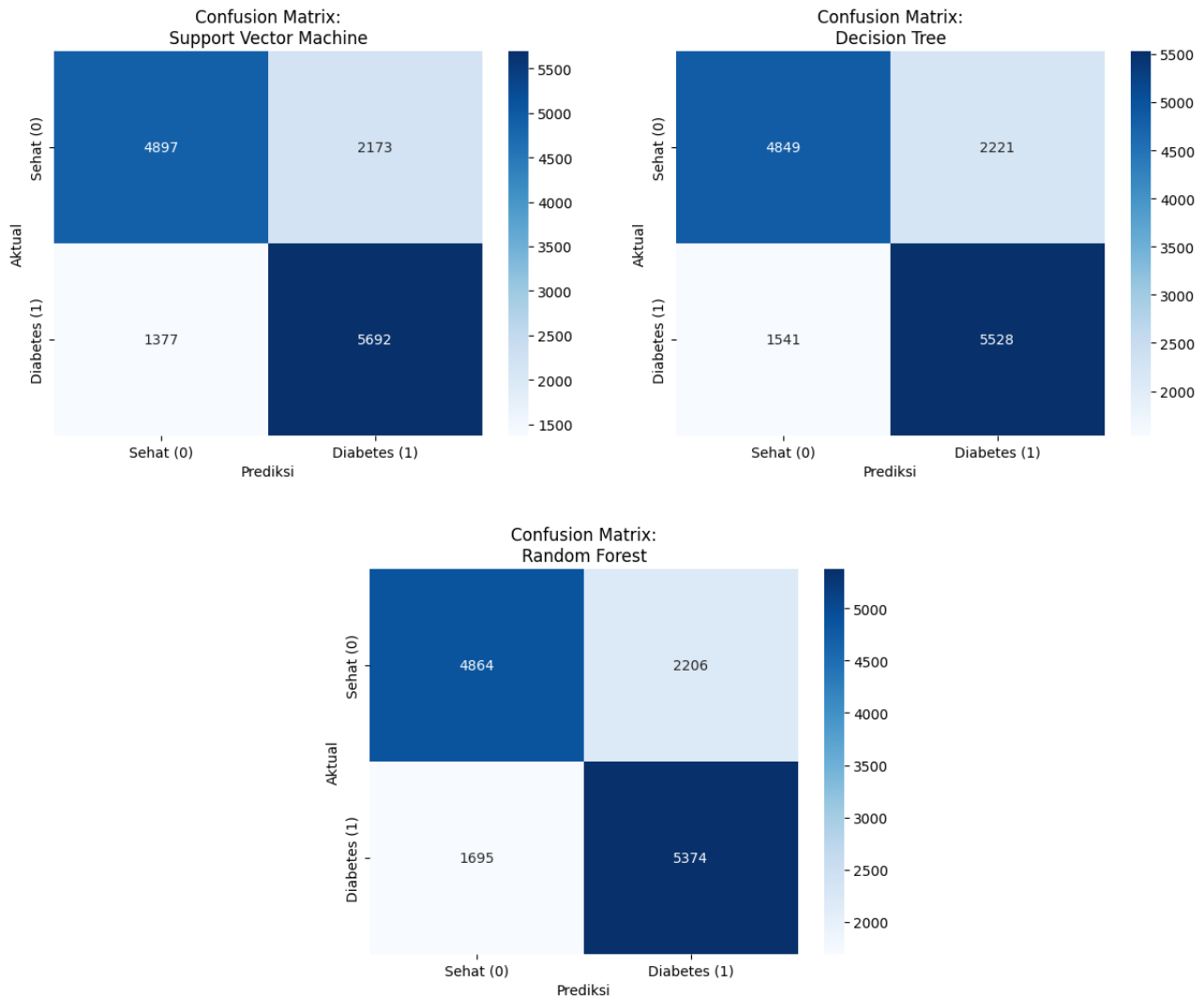
Tabel 2. Kinerja Data Latih

	SVM			Decision Tree			Random Forest		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
0	0.78	0.69	0.73	0.76	0.69	0.72	0.74	0.69	0.71
1	0.72	0.81	0.76	0.71	0.78	0.75	0.71	0.76	0.73
accuracy			0.75			0.73			0.72
Macro avg	0.75	0.75	0.75	0.74	0.73	0.73	0.73	0.72	0.72
Weighted avg	0.75	0.75	0.75	0.74	0.73	0.73	0.73	0.72	0.72

Berdasarkan Tabel 2, Hasil kinerja model menunjukkan bahwa Support Vector Machine (SVM) memberikan performa paling unggul dibandingkan Decision Tree dan Random Forest, dengan akurasi sebesar 0,75. Pada kelas 0, SVM mencapai precision 0,78, recall 0,69, dan f1-score 0,73, sedangkan pada kelas 1 nilai precision dan recall masing-masing sebesar 0,72 dan 0,81, sehingga menghasilkan f1-score 0,76. Model Decision Tree berada sedikit di bawah SVM dengan akurasi 0,73, di mana performanya pada kedua kelas relatif stabil namun tidak melampaui SVM. Sementara itu, Random Forest menunjukkan akurasi

terendah yaitu 0,72, dengan nilai precision, recall, dan f1-score yang konsisten tetapi sedikit lebih rendah dibandingkan model lainnya. Ditinjau dari macro average dan weighted average, SVM tetap menunjukkan nilai tertinggi (0,75), sedangkan Decision Tree dan Random Forest berturut-turut berada pada kisaran 0,73 dan 0,72.

Selanjutnya, pada Tabel 2, semua metrik evaluasi dihitung menggunakan confusion matrix, yang mewakili hasil prediksi dari 3 algoritma yang digunakan, seperti pada Gambar 2.



Gambar 2. Confusion Matrix Model Klasifikasi

Berdasarkan Gambar 2, Hasil confusion matrix menunjukkan perbandingan kinerja tiga model klasifikasi, yaitu Support Vector Machine (SVM), Decision Tree, dan Random Forest, dalam membedakan kelas sehat (0) dan diabetes (1). Pada model SVM, sebanyak 4.897 data sehat berhasil diklasifikasikan secara benar, sementara 2.173 data sehat salah diprediksi sebagai diabetes. Selain itu, SVM menghasilkan 1.377 kesalahan prediksi pada data diabetes yang diklasifikasikan sebagai sehat, dan 5.692 data diabetes teridentifikasi dengan benar. Model ini memperlihatkan kemampuan yang lebih baik dalam mendeteksi diabetes, terlihat dari jumlah kesalahan false negative yang paling rendah di antara ketiga model.

Pada model Decision Tree, jumlah prediksi benar untuk data sehat sedikit lebih rendah dibandingkan SVM, yaitu 4.849, dan kesalahan false positive meningkat menjadi 2.221. Selain itu, Decision Tree

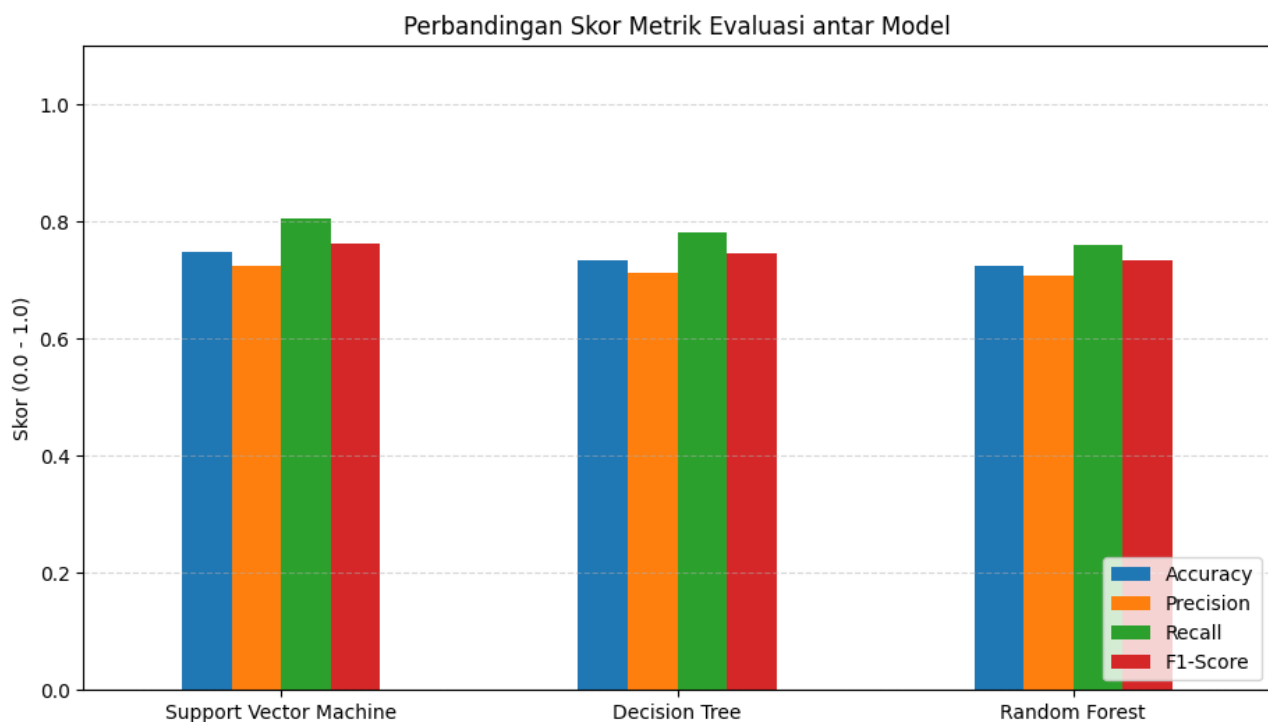
menghasilkan 1.541 false negative, yang menunjukkan bahwa model ini lebih banyak gagal mendeteksi pasien diabetes dibandingkan SVM. Meskipun demikian, 5.528 data diabetes berhasil diprediksi dengan benar oleh model ini.

Sementara itu, Random Forest menunjukkan performa yang berada di antara SVM dan Decision Tree untuk data sehat, dengan 4.864 prediksi benar dan 2.206 false positive. Namun, model ini menghasilkan jumlah false negative terbesar, yaitu 1.695 kasus, sehingga paling banyak melewatkan pasien yang sebenarnya menderita diabetes. Meski demikian, Random Forest masih mampu mengidentifikasi 5.374 data diabetes dengan benar.

Selanjutnya, dilakukan perbandingan performa algoritma klasifikasi menggunakan data uji, seperti ditampilkan pada Tabel 3 dan divisualisasikan pada Gambar 3.

Tabel 3. Evaluasi Performa

	Accuracy (%)	Precision	Recall	F1-Score
SVM	74.89	72.37	80.52	76.22
Decision Tree	73.39	71.34	78.20	74.61
Random Forest	72.41	70.90	76.02	73.37



Gambar 3. Grafik Perbandingan Evaluasi Performa

Pada Tabel 3, Berdasarkan hasil evaluasi performa model, Support Vector Machine (SVM) menunjukkan kinerja paling unggul dibandingkan Decision Tree dan Random Forest pada seluruh metrik yang digunakan. SVM memperoleh akurasi sebesar 74,89%, yang mengindikasikan bahwa model ini mampu memprediksi kelas dengan benar pada hampir tiga perempat dari keseluruhan data. Selain itu, nilai precision sebesar 72,37% menunjukkan bahwa sebagian besar prediksi positif yang dihasilkan SVM adalah benar, sementara nilai recall yang mencapai 80,52% menggarisbawahi kemampuan SVM dalam menangkap sebagian besar sampel yang benar-benar termasuk kelas positif (diabetes). Keseimbangan antara precision dan recall tercermin pada nilai F1-Score sebesar 76,22%, yang merupakan nilai tertinggi di antara ketiga model.

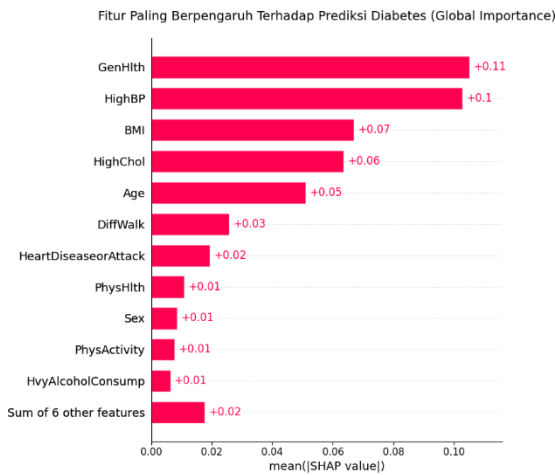
Model Decision Tree menempati posisi kedua dengan akurasi 73,39%, sedikit lebih rendah dibandingkan SVM. Precision model ini sebesar 71,34%, sedangkan recall mencapai 78,20%, menunjukkan bahwa meskipun performanya masih cukup baik, kemampuan model dalam mendeteksi kasus positif tidak sekuat SVM. Nilai F1-Score Decision Tree sebesar 74,61%, yang juga lebih rendah dari SVM,

menandakan bahwa keseimbangan antara kemampuan menghindari kesalahan prediksi positif dan menangkap kasus sebenarnya masih belum optimal.

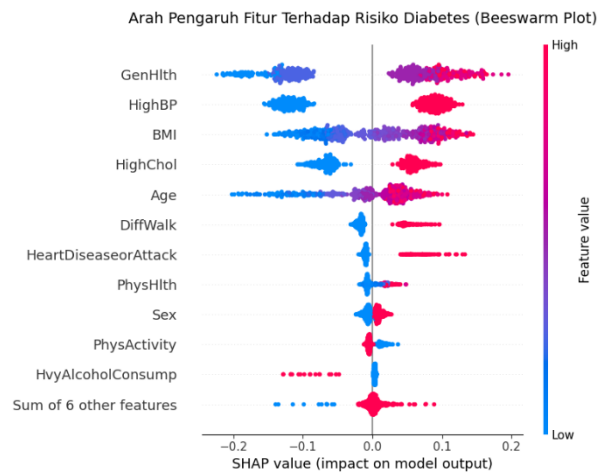
Sementara itu, Random Forest menunjukkan performa paling rendah di antara ketiga model, dengan akurasi 72,41%, precision 70,90%, recall 76,02%, dan F1-Score 73,37%. Meskipun Random Forest merupakan metode ensemble yang umumnya lebih stabil, hasil ini menunjukkan bahwa model tersebut kurang efektif dalam menangani pola data pada kasus ini dibandingkan dua model lainnya.

Dengan demikian dapat dikatakan bahwa SVM merupakan model paling efektif untuk klasifikasi penyakit diabetes pada dataset yang digunakan, dengan performa lebih konsisten dan kemampuan deteksi yang lebih baik dibandingkan Decision Tree dan Random Forest.

Kemudian untuk mendeskripsikan variabel/ atribut yang paling berkontribusi terhadap hasil keputusan atau klasifikasi model, penelitian ini menggunakan grafik SHAP (SHapley Additive exPlanations). Berikut visualisasi grafik SHAP seperti pada Gambar 4(a) dan Gambar 4(b).



Gambar 4(a). Peringkat Kontribusi Fitur



Gambar 4(b). Pengaruh Fitur Terhadap Klasifikasi

Berdasarkan Gambar 4(a), grafik tersebut menunjukkan urutan fitur yang paling berpengaruh secara global terhadap model prediksi diabetes berdasarkan nilai mean (SHAP value). Semakin besar nilai SHAP rata-rata, semakin besar pula kontribusi fitur tersebut dalam memengaruhi keputusan model. Dari hasil analisis, terlihat bahwa GenHlth (kesehatan umum) merupakan faktor paling dominan (+0.11), diikuti oleh HighBP (tekanan darah tinggi) dan BMI, yang masing-masing memiliki kontribusi substansial dalam meningkatkan risiko prediksi diabetes.

Sementara Gambar 4(b) menunjukkan seberapa besar setiap fitur memengaruhi prediksi risiko diabetes. Titik merah berarti nilai fitur tinggi, sedangkan titik biru berarti nilai rendah. Jika titik berada di kanan, artinya fitur itu meningkatkan prediksi risiko diabetes. Jika di kiri, berarti menurunkannya. Dari grafik terlihat bahwa GenHlth, tekanan darah tinggi (HighBP), BMI, kolesterol tinggi (HighChol), dan usia (Age) adalah faktor yang paling berpengaruh. Sementara fitur lain seperti aktivitas fisik atau konsumsi alkohol hanya memberi pengaruh kecil.

4. KESIMPULAN

Berdasarkan hasil analisis komparatif terhadap tiga algoritma machine learning, dapat disimpulkan bahwa Random Forest memberikan performa klasifikasi penyakit diabetes yang paling optimal dengan tingkat akurasi dan F1-score tertinggi karena kemampuannya memitigasi overfitting melalui pendekatan ensemble learning. Di sisi lain, proses normalisasi data sangat penting dalam mengoptimalkan performa algoritma machine learning Support Vector Machine (SVM), Decision Tree dan Random Forest. Hasil penelitian menunjukkan bahwa algoritma SVM memiliki kinerja terbaik dari Decision Tree dan Random Forest dengan nilai akurasi sebesar 74.89%, presisi sebesar 72.73, recall sebesar 80.52 dan F1-Score memiliki nilai 76.22.

5. DAFTAR PUSTAKA

Aditya, M.F. and Pramuntadi, A. (2024). Implementation of Decision Tree Method for Diabetes Mellitus Type 2 Prediction Implementasi Metode Decision Tree pada Prediksi Penyakit Diabetes Melitus Tipe 2, 4(July), pp. 1104–1110.

Alfreyzal. et al. (2024). Jurnal Kesehatan, Volume 13 No. 1, April 2024 Edukasi Kesehatan Pada Keluarga Diabetes Melitus Dengan Masalah Keperawatan Pemeliharaan Kesehatan Tidak Efektif, 13(1).

Baswedan, A. (2024). Analisis sentimen terhadap bakal calon presiden 2024 dengan algoritma multinomial naïve bayes dan oversampling smote, 7(5), pp. 3257–3264.

Chang, V. et al. (2022). Healthcare Analytics An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators, Healthcare Analytics, 2(October), p. 100118. Available at: <https://doi.org/10.1016/j.health.2022.100118>.

Dritsas, E. and Trigka, M. (2022). Stroke Risk Prediction with Machine Learning Techniques. Harahap, E.H., Muflikhah, L. and Rahayudi, B. (2018). Implementasi Algoritma Support Vector Machine (SVM) Untuk Penentuan Seleksi Atlet Pencak Silat, 2(10), pp. 3843–3848.

Ivanovic, V. et al. (2026). International Journal of Medical Informatics Artificial intelligence methods in gestational diabetes mellitus prediction: A systematic literature review, International Journal of Medical Informatics, 206 (September 2025), p. 106158. Available at: <https://doi.org/10.1016/j.ijmedinf.2025.106158>.

Keigley, A.S., Ayanian, S. and Dugani, S.B. (2026). American Journal of Medicine Open Machine Learning Models to Evaluate County-Level

Incidence of Diagnosed Diabetes and Sociodemographic Factors, *American Journal of Medicine Open*, 15(September 2025), p. 100132. Available at: <https://doi.org/10.1016/j.ajmo.2026.100132>.

Li, L. et al. (2025). Diabetes Research and Clinical Practice A machine learning model for hypoglycemia risk prediction in hospitalized patients with diabetes : development and validation, 230(July). Available at: <https://doi.org/10.1016/j.diabetes.2025.112993>.

Ohyanan, H. et al. (2022). Associations between the urban exposome and type 2 diabetes : Results from penalised regression by least absolute shrinkage and selection operator and random forest models, *Environment International*, 170(October), p. 107592. Available at: <https://doi.org/10.1016/j.envint.2022.107592>.

Ramadhan, P. and Veri, J. (2025). Eksplorasi Algoritma Decision Tree untuk Penentuan Siswa Berprestasi, 7(3). Available at: <https://doi.org/10.32877/bt.v7i3.2210>.

Soladoye, A.A. et al. (2025). Current Research in Translational Medicine Enhancing Alzheimer ' s disease prediction using random forest : A novel framework combining backward feature elimination and ant colony optimization, *Current Research in Translational Medicine*, 73(4), p. 103526. Available at: <https://doi.org/10.1016/j.retram.2025.103526>.

Wafa, H.S. et al. (2022). Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM). 1, pp. 40–45.

Zafar, M.M. et al. (2025). From Data to Diagnosis : A Novel Deep Learning Model for Early and Accurate Diabetes Prediction, 1, pp. 1–40.