JOURNAL OF COMPUTER SCIENCE AND INFORMATICS ENGINEERING
VOL. 05, NO. 3, JULY 2026: 467-480
Doi: https://doi.org/10.55537/cosie.v5i3.1838

e-ISSN 2827-8356
p-ISSN 2827-9999

Perbandingan SVM dan Random Forest dengan RFE untuk Prediksi Diabetes pada BMI Tinggi

Comparison of SVM and Random Forest with RFE for Diabetes Prediction

Anggi Vandryan*[1], Putry Wahyu Setyaningsih[2],
[1,2]Sistem Informasi, Universitas Mercu Buana Yogyakarta
[1]221210048@student.mercubuana-yogya.ac.id, [2]putryws@mercubuana-yogya.ac.id

Received: June 19, 2026 | Revised: July 20, 2026 | Accepted: July 31, 2026

Abstrak

Individu dengan *Body Mass Index* (BMI) tinggi merupakan kelompok paling rentan terhadap *Diabetes Mellitus* tipe 2 akibat resistensi insulin yang dipicu akumulasi lemak visceral, namun sebagian besar pemodelan *machine learning* yang ada masih dilakukan pada populasi umum tanpa mempertimbangkan karakteristik risiko spesifik tersebut. Penelitian ini menganalisis dan membandingkan performa *Support Vector Machine* (SVM) dan *Random Forest* (RF) khusus pada individu dengan BMI ≥ 25, sekaligus mengevaluasi dampak penerapan *Recursive Feature Elimination* (RFE) terhadap peningkatan performa kedua model. *Dataset Pima Indians Diabetes* dari UCI *Machine Learning* Repository digunakan sebagai sumber data, yang setelah filtering BMI menghasilkan 662 data. Tahap preprocessing mencakup imputasi median untuk nilai tidak valid, normalisasi menggunakan *StandardScaler*, dan seleksi fitur dengan RFE sebanyak empat fitur per model. Pembagian data menggunakan rasio 70:30 dan evaluasi model dilakukan menggunakan *accuracy, precision, recall, F1-score*, serta *confusion matrix*. Hasil menunjukkan RF-RFE sebagai model terbaik dengan *accuracy* 78%, *recall* kelas diabetes 68%, dan *F1-score* 72%, meningkat signifikan dibandingkan RF tanpa RFE (accuracy 74%, recall 60%). Kombinasi *Random Forest* dan RFE terbukti paling efektif dalam menekan *False Negative* yang krusial dalam konteks deteksi dini diabetes secara klinis.

Kata kunci: Body Mass Index, Diabetes Mellitus, Random Forest, Recursive Feature Elimination, Support Vector Machine

Abstract

Individuals with a high Body Mass Index (BMI) are among the most vulnerable groups to developing Type 2 Diabetes Mellitus due to insulin resistance caused by visceral fat accumulation. However, most existing machine learning models have been developed using general population data without considering the specific characteristics of high-risk individuals. This study aims to analyze and compare the performance of Support Vector Machine (SVM) and Random Forest (RF) algorithms in predicting diabetes risk among individuals with a BMI ≥ 25, while also evaluating the impact of Recursive Feature Elimination (RFE) on improving model performance. The Pima Indians Diabetes Dataset from the UCI Machine Learning Repository was used as the data source. After filtering records based on BMI, a total of 662 instances were included in the analysis. The preprocessing stage consisted of median imputation for invalid values, feature normalization using StandardScaler, and feature selection using RFE to select four features for each model. The dataset was divided into training and testing sets using a 70:30 ratio. Model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrix metrics. The results indicate that the RF-RFE model achieved the best performance, with an accuracy of 78%, a recall of 68% for the diabetes class, and an F1-score of 72%, representing a significant improvement over the RF model without RFE (74% accuracy and 60% recall). The combination of Random Forest and Recursive Feature Elimination proved

e-ISSN 2827-8356
p-ISSN 2827-9999

to be the most effective approach for reducing false negatives, which is particularly important in the context of early clinical detection of diabetes.

Keywords: Body Mass Index, Diabetes Mellitus, Random Forest, Recursive Feature Elimination, Support Vector Machine

1. PENDAHULUAN

Indonesia tercatat sebagai salah satu negara dengan beban diabetes tertinggi di dunia, dan kondisi ini diperkirakan akan semakin berat seiring meningkatnya angka obesitas di populasi urban. Secara *global*, *World Health Organization* (WHO) mencatat bahwa lebih dari 422 juta jiwa saat ini hidup dengan diabetes, dengan angka kematian yang secara langsung dikaitkan pada kondisi ini mencapai 1,6 juta jiwa setiap tahunnya [1]. Proyeksi IDF bahkan memperkirakan pada 2045 jumlah penderita akan bertambah 12,2% menjadi 642,7 juta orang, dan Indonesia diprediksi masuk ke dalam lima negara dengan jumlah penderita terbanyak di dunia. Di antara berbagai faktor risiko yang telah diteliti secara luas, *Body Mass Index* (*BMI*) yang tinggi — khususnya pada kategori *overweight* hingga *obese* dengan *BMI* ≥ 25 — menjadi prediktor yang paling konsisten dan terukur, karena akumulasi lemak *visceral* pada individu dengan *BMI* tinggi memicu resistensi *insulin* yang secara langsung berkontribusi pada timbulnya *Diabetes Mellitus* tipe 2. Individu dalam kelompok ini terbukti memiliki risiko 2 – 4 kali lebih besar untuk mengembangkan diabetes dibanding individu dengan berat badan normal [2]. Oleh karena itu, pengembangan sistem deteksi awal yang mampu mengidentifikasi risiko diabetes secara efektif dan akurat pada kelompok berisiko tinggi menjadi suatu kebutuhan mendesak. Seiring dengan perkembangan pesat teknologi informasi, pendekatan *data mining* dan *machine learning*, berbagai teknik klasifikasi telah dikembangkan untuk mendukung deteksi penyakit berbasis data kesehatan [3] .

Meskipun berbagai penelitian tentang klasifikasi diabetes telah dilakukan, sebagian besar masih bersifat umum tanpa memfokuskan pemodelan pada sub-populasi dengan faktor risiko spesifik. Pada penelitian [4] tersebut membandingkan SVM dan *Random Forest* menggunakan dataset diabetes dari *Sylhet Diabetic Hospital* dengan 520 data dan 16 fitur, memperoleh akurasi RF tertinggi sebesar 98,08%. Namun, penelitian tersebut tidak memperhitungkan karakteristik BMI sebagai filter populasi. Penelitian [5] menerapkan RFE pada beberapa *algoritma tree-based* menggunakan *Pima Indians Diabetes Dataset*, memperoleh akurasi tertinggi 77,27% pada *Random Forest* dan *Gradient Boosting*, serta membuktikan bahwa RFE efektif mereduksi dimensi dan mengurangi *overfitting.* Akan tetapi, penelitian tersebut juga tidak membatasi populasi berdasarkan BMI. Penelitian yang dilakukan [6] mengklasifikasikan diabetes menggunakan *Categorical Boosting* dengan mempertimbangkan faktor risiko, namun tidak menggunakan RFE sebagai teknik seleksi fitur. Permasalahan utama yang dijawab dalam penelitian ini adalah bagaimana performa metode SVM dan RF dalam mengklasifikasikan risiko diabetes, serta sejauh mana pengaruh penerapan teknik *feature selection Recursive Feature Elimination* (RFE) terhadap peningkatan performa model.

Melalui penelitian ini, dua hal utama yang ingin dijawab adalah: seberapa besar perbedaan performa antara SVM dan RF ketika diterapkan secara spesifik pada populasi BMI tinggi, dan apakah penerapan RFE secara komparatif pada kedua model dapat meningkatkan kemampuan deteksi kasus diabetes secara signifikan — khususnya dalam menekan angka False Negative yang berdampak langsung pada keselamatan pasien [7]. Berdasarkan gap penelitian yang telah diidentifikasi, penelitian ini secara khusus membatasi populasi pada individu dengan BMI ≥ 25 untuk mencerminkan kondisi kelompok berisiko tinggi yang lebih representatif secara klinis[8]. Sebagian besar penelitian sebelumnya juga belum secara eksplisit mengkaji pengaruh RFE terhadap kedua model secara komparatif pada populasi dengan BMI tinggi. Sebagian besar penelitian juga masih berfokus pada klasifikasi diabetes secara umum tanpa mempertimbangkan

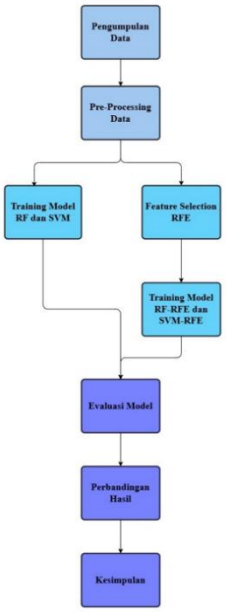
e-ISSN 2827-8356
p-ISSN 2827-9999

faktor risiko seperti *Body Mass Index* (BMI). Kebaruan penelitian ini bukan sekadar penerapan kombinasi SVM, RF, dan RFE, melainkan pada temuan bahwa dampak seleksi fitur melalui RFE bersifat tidak seragam antar algoritma ketika diterapkan secara khusus pada populasi berisiko tinggi (BMI ≥ 25). Penelitian-penelitian sebelumnya umumnya menerapkan RFE pada satu algoritma saja atau pada populasi umum, sehingga belum terungkap apakah efektivitas RFE bersifat spesifik-algoritma atau general. Dengan membandingkan peningkatan performa RF-RFE dan SVM-RFE secara langsung pada populasi yang sama, penelitian ini memberikan bukti empiris bahwa RF memperoleh manfaat yang lebih besar dari RFE dibandingkan SVM pada konteks populasi berisiko tinggi, sebuah temuan yang relevan sebagai pertimbangan pemilihan strategi seleksi fitur pada pengembangan sistem deteksi dini diabetes berbasis machine learning di masa mendatang.

2. METODE PENELITIAN

2.1 Tahapan Penelitian

Penelitian yang dilakukan menggunakan pendekatan kuantitatif berbasis eksperimen komputasional untuk menganalisis dan membandingkan performa metode klasifikasi dalam memprediksi risiko *Diabetes Mellitus*. Pendekatan eksperimen dipilih karena memungkinkan pengujian dan perbandingan sistematis antara dua model *machine learning* dalam kondisi yang terkontrol. Proses penelitian dilakukan secara bertahap meliputi: (1) pengumpulan data, (2) *pre-processing* data, (3) *training model Support Vector Machine dan Random Forest* (4) *Feature Selection* menggunakan *RFE* dan (5) evaluasi performa model menggunakan *confusion matrix*. Setiap tahapan disusun secara sistematis untuk mendukung mengembangkan model prediksi awal yang akurat dan dapat direproduksi. Penelitian ini menggunakan skema pembagian data tunggal (*single hold-out split*) dengan rasio 70:30 dan *random_state* tetap untuk menjaga konsistensi dan reproduktibilitas antar eksperimen SVM dan RF. Pendekatan ini dipilih mengingat ukuran dataset yang relatif terbatas (662 data setelah filtering BMI ≥ 25), di mana skema k-fold cross-validation berpotensi menghasilkan subset data uji yang sangat kecil pada setiap fold sehingga estimasi metrik evaluasi menjadi kurang stabil. Meskipun demikian, penulis menyadari bahwa validasi silang seperti *stratified k-fold cross-validation* akan memberikan estimasi performa yang lebih robust dan mengurangi potensi bias akibat pembagian data yang kebetulan menguntungkan salah satu model, sehingga hal ini diusulkan sebagai arah penelitian selanjutnya.



Gambar 1 Tahapan Penelitian

e-ISSN 2827-8356
p-ISSN 2827-9999

2.2 Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah *Pima Indians Diabetes Dataset*, sebuah dataset publik yang diperoleh *dari UCI Machine Learning Repository* dan telah menjadi salah satu benchmark paling umum digunakan dalam penelitian klasifikasi diabetes. Dataset ini berisi rekaman medis 768 perempuan keturunan suku *Pima Indian* yang berusia minimal 21 tahun, dengan delapan atribut independen dan satu atribut dependen berupa label Outcome. 8 atribut yang digunakan sebagai fitur prediktor mencakup: *Pregnancies* (jumlah kehamilan yang pernah dialami), *Glucose* (konsentrasi glukosa plasma dalam satuan mg/dL), *BloodPressure* (tekanan darah *diastolik* dalam *mmHg*), *SkinThickness* (ketebalan lipatan kulit trisep dalam milimeter), *Insulin* (kadar insulin serum dua jam dalam μU/mL), *BMI* (indeks massa tubuh dalam kg/m²), *DiabetesPedigreeFunction* (fungsi yang mencerminkan riwayat diabetes dalam keluarga), dan *Age* (usia dalam tahun). Atribut dependen berupa nilai biner *Outcome*, di mana nilai 0 menunjukkan individu tidak menderita diabetes dan nilai 1 menunjukkan individu menderita diabetes.

2.3 Pre-Processing Data

Tahap *pre-processing* data ini memiliki beberapa langkah yaitu pembersihan data, seperti nilai yang tidak valid seperti nilai nol pada beberapa atribut dan digantikan dengan nilai median agar tidak memengaruhi hasil klasifikasi. Selanjutnya, dilakukan filtering berdasarkan nilai Body Mass Index (BMI). Normalisasi dilakukan menggunakan metode *StandardScaler* untuk menyamakan skala antar fitur sehingga model dapat bekerja lebih optimal [10].

2.4 Feature Selection

Pada penelitian ini menggunakan teknik *Recursive Feature Elimination* (RFE) untuk melakukan seleksi fitur. Metode ini bekerja secara bertahap dan berulang dengan melatih model menggunakan seluruh fitur yang tersedia pada setiap iterasi. Selanjutnya, fitur yang memiliki kontribusi paling rendah, berdasarkan nilai bobot atau *feature importance*, akan dihapus dari proses pemodelan. Tahapan ini terus dilakukan hingga diperoleh kombinasi fitur yang paling optimal dalam menghasilkan performa klasifikasi terbaik. Keunggulan RFE dibandingkan metode seleksi fitur berbasis filter yaitu pada kemampuannya dalam mempertimbangkan hubungan dan hubungan antar fitur sesuai dengan karakteristik model yang digunakan. Dengan demikian, fitur yang dipilih tidak hanya memiliki korelasi statistik yang tinggi, tetapi juga benar-benar berpengaruh terhadap performa model klasifikasi. Pada penelitian ini, implementasi RFE dilakukan secara terpisah pada masing-masing algoritma, yaitu SVM-RFE menggunakan estimator SVM dengan *kernel linear* C = 1.0 (default), sedangkan RF-RFE menggunakan *estimator* diset secara default *Random Forest*, sehingga fitur yang dihasilkan menyesuaikan karakteristik dari setiap algoritma [11].

2.5 Support Vector Machine

SVM berperan sebagai salah satu dari dua model klasifikasi yang dibandingkan. Cara kerja SVM bertumpu pada pencarian bidang pemisah (hyperplane) dengan margin terlebar yang secara geometris memaksimalkan jarak antara kedua kelas. Ketika data yang diperoleh lalu tidak dipisahkan secara linear di ruang aslinya, fungsi kernel digunakan untuk mentransformasi data ke dimensi yang lebih tinggi. Pada penelitian ini digunakan kernel RBF untuk pelatihan utama, sedangkan kernel linear dipakai khusus untuk SVM-RFE karena mekanisme seleksi fitur berbasis bobot koefisien hanya tersedia pada model linear. SVM menggunakan parameter default scikit-learn: Kernel RBF, C = 1.0, gamma = “scale” [12].

2.6 Random Forest

Random Forest bekerja dengan prinsip kebijaksanaan kolektif: daripada mengandalkan satu pohon keputusan yang rentan overfitting, metode ini membangun ratusan pohon dari subsampel data yang berbeda-beda, lalu menentukan keputusan akhir melalui voting mayoritas.

e-ISSN 2827-8356
p-ISSN 2827-9999

Keberagaman antar pohon inilah yang membuat RF secara alami lebih stabil dan tahan terhadap noise dibandingkan decision tree Tunggal. Random_state = 42, n_estimators = 100, criterian = 'gini', max_depth = None, min_samples_split = 2, min_samples_leaf = 1 [13]. Dalam konteks penelitian ini, sifat RF yang mampu menangani hubungan non-linear antar fitur menjadi keunggulan kunci ketika berhadapan dengan data klinis yang kompleks.

2.7 Evaluasi Model

Evaluasi model dilakukan dengan menggunakan beberapa metrik untuk memperoleh gambaran yang lebih komprehensif dalam mengukur kinerja serta membandingkan performa antar model. Selain itu juga digunakan untuk menunjukkan perbandingan antara label aktual dan label hasil prediksi model, sehingga dapat memberikan informasi terkait tingkat kesalahan dan ketepatan model dalam melakukan klasifikasi.

Table 1. Struktur Confusion Matrix

	Prediksi Positif	Prediksi Negatif
Aktual Positif (P)	True Positif (TP)	False Negative (FN)
Aktual Negatif (N)	False Positif (FP)	True Negative (TN)

Dalam bidang medis, pengambilan keputusan yang tepat sangat berperan penting dalam menjaga keselamatan pasien. Ketika model kecerdasan buatan digunakan sebagai alat bantu dalam proses diagnosis penyakit, khususnya pada kasus yang kompleks seperti diabetes, diperlukan pemahaman yang mendalam mengenai cara kerja serta evaluasi model tersebut. Keempat metrik ini dipilih karena masing-masing mengukur aspek performa yang berbeda dan saling melengkapi, terutama dalam konteks data tidak seimbang seperti yang digunakan dalam penelitian ini[14].

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil penelitian yang dilakukan untuk mengukur dampak RFE terhadap kemampuan prediksi awal diabetes. Perbandingan dilakukan antara dua algoritma SVM dan RF dalam dua kondisi: sebelum dan sesudah seleksi fitur diterapkan. Evaluasi dilakukan menggunakan pendekatan evaluasi model. Tujuannya untuk menilai seberapa efektif RFE dalam meningkatkan efisiensi dan evaluasi model.

Table 2 Dataset Pima Indians Diabetes

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

e-ISSN 2827-8356
p-ISSN 2827-9999

3.1 Pre-Processing Data

Tahapan pre-processing meliputi data cleaning, filtering data berdasarkan nilai BMI, normalisasi data, serta pembagian data latih dan data uji. Setelah dataset awal diperiksa, ditemukan bahwa sejumlah atribut mengandung nilai nol yang secara medis tidak mungkin terjadi pada kondisi nyata. Nilai nol pada atribut seperti SkinThickness, Insulin, dan atribut lainnya tidak mencerminkan kondisi yang valid seseorang tidak mungkin memiliki kadar glukosa nol dan tetap hidup, misalnya. Nilai-nilai tersebut hampir pasti merupakan data yang hilang (missing values) yang direpresentasikan dengan angka nol dalam dataset aslinya. Untuk mengatasi hal ini, setiap nilai nol pada kelima atribut tersebut digantikan dengan nilai median dari masing-masing kolom. Median dipilih sebagai estimator pengganti karena beberapa atribut terutama Insulin memiliki distribusi yang sangat miring ke kanan dengan kehadiran outlier ekstrem, sehingga median memberikan estimasi pusat distribusi yang lebih representatif dibandingkan nilai rata-rata yang mudah terdistorsi oleh nilai ekstrem tersebut.

Tahap berikutnya adalah filtering data berdasarkan BMI, yang menyaring dataset dari 768 menjadi 662 rekaman dengan hanya mempertahankan individu ber-BMI ≥ 25. Setelah itrau, seluruh atribut dinormalisasi menggunakan StandardScaler sehingga setiap fitur memiliki kelas rata-rata mendekati nol dan standar deviasi sebesar satu. Tahap terakhir adalah pembagian dataset menjadi data latih dan data uji menggunakan proses *hold-out*. Dataset dibagi menjadi perbandingan 70% data latih dan 30% data uji. Tabel 3 menampilkan hasil dataset setelah tahap pre-processing.

Table 3. Dataset Setelah Pre-Processing

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	30.5	33.6	0.627	50	1
1	85	66	29	30.5	26.6	0.351	31	0
1	89	66	23	94.0	28.1	0.167	21	0
0	137	40	35	168.0	43.1	2.288	33	1
5	116	74	23	30.5	25.6	0.201	30	0

3.2 Seleksi Fitur Menggunakan Recursive Feature Elimination (RFE)

Metode ini digunakan untuk menentukan fitur-fitur yang paling relevan dalam proses klasifikasi sehingga model dapat bekerja secara lebih optimal. Penerapan RFE bertujuan untuk mengurangi dimensi dataset dengan menghilangkan fitur yang memiliki kontribusi rendah terhadap performa model, sehingga model dapat lebih fokus dalam mengenali pola data yang penting. Proses seleksi fitur menggunakan RFE dilakukan secara bertahap dan berulang (iteratif). Pada tahap awal, seluruh fitur pada dataset digunakan dalam proses pelatihan model. Selanjutnya, model akan menghitung tingkat kontribusi masing-masing fitur berdasarkan nilai bobot (*weight*) atau *feature importance* yang dihasilkan selama proses pelatihan. Setelah proses pelatihan selesai dilakukan, fitur yang memiliki kontribusi paling rendah akan dieliminasi dari dataset. Model kemudian dilatih kembali menggunakan fitur yang masih tersisa. Proses ini dilakukan secara

e-ISSN 2827-8356
p-ISSN 2827-9999

berulang hingga diperoleh kombinasi fitur yang paling optimal sesuai dengan jumlah fitur yang telah ditentukan. Pada penelitian ini, jumlah fitur yang dipilih ditetapkan sebanyak 4 fitur utama untuk masing-masing model klasifikasi.

Implementasi RFE dilakukan secara terpisah pada metode SVM dan RF karena kedua algoritma memiliki mekanisme yang berbeda dalam menentukan tingkat kepentingan fitur. Pada model SVM-RFE, proses seleksi fitur menggunakan estimator SVM *dengan kernel linear,* sehingga pemilihan fitur didasarkan pada nilai koefisien *(coefficient weight*) yang dihasilkan oleh hyperplane. Sementara itu, pada model RF-RFE proses seleksi fitur dilakukan menggunakan *estimator Random Forest* yang menentukan tingkat kepentingan fitur berdasarkan nilai *feature importance* dari setiap decision tree.

Table 4 Penerapan Recursive Feature Selection pada SVM dan RF

NO	Fitur	SVM-RFE	RF-RFE
1.	Glucose	√	√
2.	BMI	√	√
3.	Pregnancies	√	–
4.	SkinThickness	√	–
5.	BloodPressure	–	√
6.	DiabetesPedigreeFunction	–	√
	Total Fitur Dipilih	4	4

Berdasarkan hasil proses seleksi fitur, model SVM-RFE menghasilkan fitur terpilih berupa *Pregnancies*, *Glucose*, *SkinThickness*, dan *BMI*. Sedangkan model RF-RFE menghasilkan fitur *Glucose*, *BloodPressure*, *BMI*, dan *Age*. Perbedaan fitur yang terpilih menunjukkan bahwa setiap algoritma mempunyai karakteristik yang berbeda dalam mengenali pola data pada dataset. Penerapan RFE pada penelitian ini berhasil mengurangi jumlah fitur dari 8 atribut menjadi 4 atribut utama tanpa menurunkan performa model. Selain itu, proses seleksi fitur juga membantu mengurangi kompleksitas model, meminimalkan noise pada dataset, serta meningkatkan performa klasifikasi, khususnya pada model *Random Forest*.

3.3 Pengujian Model

3.3.1 Implementasi Support Vector Machine

Table 5 perbandingan Evaluasi Model SVM tanpa RFE dan SVM-RFE

Kelas	Matrix	SVM (tanpa RFE)	SVM-RFE(dengan RFE)
Kelas 0 (Non-Diabetes)	Precision	76%	77%
	Recall	87%	91%
	F-1 Score	81%	83%

e-ISSN 2827-8356
p-ISSN 2827-9999

Kelas	Matrix	SVM (tanpa RFE)	SVM-RFE(dengan RFE)
	Precision	77%	82%
Kelas 1 (Diabetes)	Recall	60%	61%
	F-1 Score	67%	70%
Accuracy Keseluruhan		76%	78%

Berdasarkan hasil pengujian pada Tabel 5, penerapan *RFE* memberikan peningkatan yang konsisten pada performa model *SVM* di hampir seluruh *metrik evaluasi*. *Accuracy* keseluruhan naik yang menunjukkan bahwa pemilihan empat fitur terbaik oleh *SVM-RFE* — yaitu *Pregnancies*, *Glucose*, *SkinThickness*, dan *BMI* berhasil membantu model memfokuskan diri pada pola yang lebih relevan dibandingkan saat seluruh delapan fitur digunakan sekaligus. Peningkatan yang paling menonjol terlihat pada precision kelas diabetes Ini berarti ketika model *SVM-RFE* menyatakan seseorang menderita diabetes, tingkat kepercayaan prediksi tersebut jauh lebih tinggi dibandingkan model *SVM* tanpa seleksi fitur.

3.3.2 Implementasi Random Forest

Table 6 perbandingan Evaluasi Model RF tanpa RFE dan RF-RFE

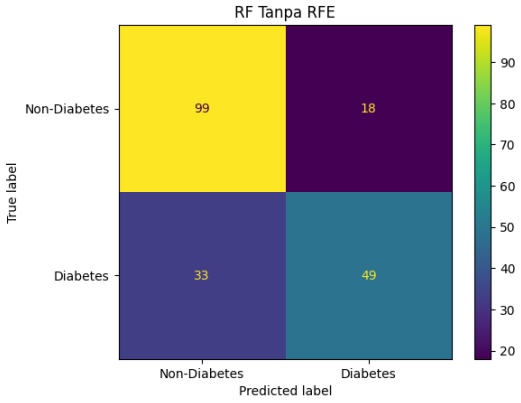
Kelas	Matrix	RF (tanpa RFE)	RF-RFE(dengan RFE)
	Precision	75%	79%
Kelas 0 (Non-Diabetes)	Recall	85%	85%
	F-1 Score	80%	82%
Kelas	Matrix	RF (tanpa RFE)	RF-RFE(dengan RFE)
	Precision	73%	77%
Kelas 1 (Diabetes)	Recall	60%	68%
	F-1 Score	66%	72%
Accuracy Keseluruhan		74%	78%

Berbeda dari *SVM*, model *Random Forest* menunjukkan respons yang jauh lebih signifikan terhadap penerapan *RFE*. Dari Tabel 6, terlihat bahwa hampir seluruh metrik mengalami peningkatan yang berarti setelah seleksi fitur dilakukan. Peningkatan *recall* sebesar delapan poin persentase menjadi temuan yang paling penting dari seluruh eksperimen ini. Artinya, dengan hanya menggunakan empat fitur terpilih *Glucose*, *BloodPressure*, *BMI*, dan *Age* model *RF-RFE* mampu mendeteksi lebih banyak pasien diabetes yang benar dibandingkan saat semua fitur tersedia. Hal ini menarik karena secara intuitif seseorang mungkin berasumsi bahwa lebih

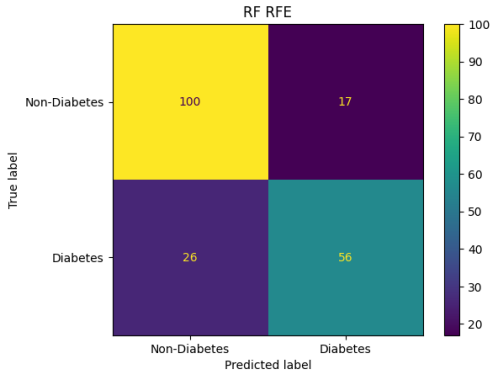
e-ISSN 2827-8356
p-ISSN 2827-9999

banyak informasi selalu menghasilkan prediksi yang lebih baik, padahal justru sebaliknya terdapat fitur-fitur yang tidak relevan atau redundan dapat menambahkan noise yang mengaburkan pola penting, terutama pada model berbasis pohon keputusan.

3.3 Confusion Matrix



Gambar 2 Confusion Matrix Random Forest



Gambar 3 Confusion Matrix RF menggunakan RFE

Berdasarkan pada gambar 2 dan 3, penerapan metode RFE menunjukkan peningkatan performa klasifikasi dibandingkan model tanpa RFE. Pada model *Random Forest* tanpa RFE, diperoleh nilai *True Positive* sebanyak 49 dan *False Negative* sebanyak 33. Setelah penerapan RFE, nilai *True Positive* meningkat menjadi 56, sedangkan *False Negative* menurun menjadi 26. Hal ini menunjukkan bahwa model menjadi efektif dalam mendeteksi kasus diabetes setelah dilakukan seleksi fitur. Selain itu, pada kelas non-diabetes juga terjadi sedikit peningkatan performa. Nilai *True Negative* meningkat dari 99 menjadi 100, sedangkan *False Positive* menurun dari 18 menjadi 17. Penurunan nilai *False Positive* dan *False Negative* menunjukkan bahwa penggunaan RFE mampu membantu model dalam mengurangi kesalahan klasifikasi. Secara keseluruhan, hasil *confusion matrix* menunjukkan kombinasi RF dan RFE memberikan performa yang lebih optimal dalam klasifikasi risiko *Diabetes Mellitus*.

JOURNAL OF COMPUTER SCIENCE AND INFORMATICS ENGINEERING
VOL. 05, NO. 3, JULY 2026: 467-480
Doi: https://doi.org/10.55537/cosie.v5i3.1838
e-ISSN 2827-8356
p-ISSN 2827-9999

Gambar 4 Correlation Matrix Antar Fitur

Gambar 4 menyajikan korelasi matrix yang memvisualisasikan hubungan linear antara seluruh atribut dalam dataset, termasuk hubungannya dengan variabel target *Outcome*. Beberapa pola menarik dapat dibaca langsung dari matriks ini. Fitur *Glucose* memiliki korelasi tertinggi terhadap *Outcome* sebesar 0,44 — sebuah temuan yang selaras dengan pengetahuan medis bahwa kadar glukosa darah adalah penanda utama gangguan metabolisme glukosa yang menjadi inti dari penyakit diabetes. *BMI* dan *Age* mengikuti dengan nilai korelasi masing-masing sebesar 0,22 dan 0,23, yang menegaskan bahwa indeks massa tubuh dan pertambahan usia turut memainkan peran dalam meningkatkan risiko diabetes. *Pregnancies* memiliki korelasi 0,21 yang juga cukup bermakna, mengingat kehamilan berulang diketahui dapat mempengaruhi sensitivitas insulin pada wanita. Sementara itu, *BloodPressure* dan *SkinThickness* menunjukkan korelasi yang lebih lemah terhadap *Outcome*, yang konsisten dengan fakta bahwa kedua fitur ini tidak terpilih oleh *SVM-RFE* namun terpilih oleh *RF-RFE* mengindikasikan bahwa *RF* mampu memanfaatkan interaksi non-linear dari fitur-fitur ini yang tidak tertangkap oleh ukuran korelasi linear sederhana.

Berdasarkan hasil pengujian yang telah dilakukan, penerapan RFE memberikan dampak positif terhadap performa kedua model, dengan peningkatan yang lebih signifikan pada model *Random Forest* dibandingkan SVM. Evaluasi model dilakukan menggunakan metrik *accuracy, precision, recall*, serta didukung dengan analisis *confusion matrix* untuk melihat tingkat kesalahan klasifikasi secara lebih detail. Hasil perbandingan performa kedua model tersebut dapat dilihat dalam bentuk tabel untuk mempermudah analisis. Berdasarkan Tabel 5, penerapan metode *Recursive Feature Elimination* (RFE) memberikan dampak yang berbeda pada setiap algoritma klasifikasi yang digunakan. Pada model SVM, terjadi peningkatan performa yang ditunjukkan oleh naiknya nilai accuracy dari 76% menjadi 78% serta precision dari 77% menjadi 82%. Hasil tersebut menunjukkan bahwa RFE mampu membantu model dalam memilih fitur yang lebih relevan sehingga prediksi terhadap kelas diabetes menjadi lebih akurat. Meskipun demikian, peningkatan recall pada kelas diabetes hanya mengalami kenaikan yang kecil, yaitu dari 60% menjadi 61%, sehingga kemampuan model dalam mendeteksi seluruh kasus diabetes masih belum maksimal. Di sisi lain, model *Random Forest* mengalami peningkatan performa yang lebih signifikan setelah penerapan RFE. Nilai accuracy meningkat dari 74% menjadi 78%, precision meningkat dari 73% menjadi 77%, recall dari 60% menjadi 68%. Peningkatan nilai recall menunjukkan bahwa model menjadi lebih efektif dalam mengenali pasien yang benar-benar menderita diabetes. Dalam bidang medis, peningkatan recall memiliki peranan penting karena dapat mengurangi kemungkinan kesalahan dalam mendeteksi pasien positif diabetes.

Sebagaimana ditunjukkan pada Gambar 4, selain meningkatkan performa klasifikasi, RFE juga menghasilkan subset fitur yang berbeda antara kedua model. Pada model SVM-RFE, fitur yang terpilih adalah Pregnancies, Glucose, SkinThickness, dan BMI. Sementara itu, pada

e-ISSN 2827-8356
p-ISSN 2827-9999

RF-RFE fitur yang dominan adalah Glucose, BloodPressure, BMI, dan Age. Perbedaan ini terjadi karena kedua model memiliki mekanisme penentuan *feature importance* yang berbeda: SVM-RFE menggunakan bobot koefisien dari hyperplane linear, sedangkan RF-RFE menggunakan Gini impurity atau mean decrease impurity dari setiap pohon keputusan. Meskipun berbeda, kedua model sepakat menempatkan Glucose dan BMI sebagai fitur paling relevan. Temuan ini konsisten dengan literatur medis yang menyatakan bahwa kadar glukosa merupakan prediktor terkuat diabetes, sementara BMI mencerminkan kondisi obesitas yang erat kaitannya dengan resistensi insulin. Perbedaan pada fitur Age (dipilih RF) dan Pregnancies (dipilih SVM) menunjukkan bahwa kedua model menangkap aspek risiko yang saling melengkapi, yaitu risiko berbasis usia dan risiko berbasis riwayat reproduksi.

Table 7 Perbandingan Penelitian Ini dengan Penelitian Terkait

Aspek Perbandingan	Penelitian ini	Penelitian [15]	Penelitian [16]	Penelitian [17]
Dataset	Pima Indians Diabetes Dataset	Pima Indians Diabetes Dataset	Pima Indians Diabetes Dataset	Pima Indians Diabetes Dataset
Jumlah data awal	768 data	768 data	768 data	768 data
Filter populasi BMI	BMI ≥ 25 → 662 data	Tidak difilter	Tidak difilter	Tidak difilter
Algoritma klasifikasi	Support Vector Machine + Random Forest	Decision Tree, Random Forest, Gradient Boosting, XGBoost	Decision Tree (entropy, max_depth=3)	Random Forest, SVM, XGBoost
Seleksi fitur (RFE)	✓ RFE pada SVM & RF	✓ RFE pada semua model	Tidak menggunakan RFE	Tidak menggunakan RFE
Jumlah fitur terpilih	4 fitur (dari 8)	2–7 fitur (bervariasi per model)	8 fitur (semua digunakan)	8 fitur (semua digunakan)
Fitur terpilih utama	Glucose, BMI, Pregnancies / BloodPressure / SkinThickness / DiabetesPedigreeFunction	Glucose, BMI (Gradient Boosting); +Age, BloodPressure, dll. (RF)	Glucose, BMI, Age (root node)	Glucose, BMI (*feature importance* tertinggi)
Pembagian data	70% latih : 30% uji	80% latih : 20% uji	80% latih : 20% uji	80% latih : 20% uji
Accuracy terbaik	78% (RF-RFE & SVM-RFE)	77,27% (RF-RFE & GB-RFE)	76,62% (Decision Tree)	76% (Random Forest)
Model terbaik	RF-RFE	RF-RFE & GB-RFE (setara)	Decision Tree	Random Forest

Hasil penelitian ini sejalan [15] yang dimana menerapkan RFE pada beberapa algoritma tree-based classifier menggunakan dataset Pima Indians Diabetes Database. Pada penelitian tersebut, Random Forest dengan RFE mencapai akurasi tertinggi sebesar dengan jumlah fitur optimal sebanyak 5 hingga 6 fitur dari 8 fitur awal, yang mengonfirmasi bahwa RFE efektif dalam meningkatkan performa Random Forest.

e-ISSN 2827-8356
p-ISSN 2827-9999

Perbandingan lebih lanjut dilakukan terhadap penelitian [16] yang menggunakan algoritma Decision Tree berbasis entropi dengan pembatasan kedalaman pohon maksimal tiga level (max_depth=3) pada dataset Pima Indians Diabetes tanpa filtering BMI. Penelitian tersebut memperoleh 20 kasus False Negative dari 55 sampel diabetes pada data uji. Meskipun penelitian ini menekankan keunggulan interpretabilitas melalui visualisasi struktur pohon keputusan berupa aturan if-then yang mudah dipahami tenaga medis, performa deteksi kasus diabetes yang dihasilkan masih lebih rendah dibandingkan model RF-RFE dalam penelitian ini yang mencapai recall 68% dengan jumlah False Negative yang lebih kecil. Perbedaan ini menunjukkan bahwa meskipun Decision Tree unggul dalam hal transparansi logika klasifikasi, penggunaan ensemble method seperti Random Forest yang dikombinasikan dengan seleksi fitur RFE terbukti lebih efektif dalam meminimalkan kesalahan deteksi kasus positif diabetes — aspek yang paling krusial dalam konteks klinis. Selain itu, kedua penelitian sepakat menempatkan Glucose, BMI, dan Age sebagai fitur paling berpengaruh terhadap risiko diabetes, yang memperkuat validitas fitur-fitur terpilih pada model RF-RFE penelitian ini.

Penelitian [17] Random Forest memperoleh akurasi tertinggi sebesar 76% dengan nilai AUC 0,82 dan berhasil mengidentifikasi 32 kasus True Positive dengan 22 False Negative, sementara SVM menghasilkan akurasi lebih rendah sebesar 73% dengan True Positive hanya 28 dan False Negative sebanyak 26 dari total data uji. Kedua model tersebut dilatih menggunakan seluruh 8 fitur tanpa seleksi fitur dan tanpa pembatasan populasi berdasarkan BMI. Jika dibandingkan secara langsung, model RF-RFE dalam penelitian ini menghasilkan akurasi yang lebih tinggi sebesar 78% dengan False Negative yang lebih kecil, yaitu 26 dari data uji yang berfokus pada populasi BMI ≥ 25, sementara model SVM-RFE juga mencapai akurasi 78% — melampaui SVM tanpa RFE milik Setiawan et al. yang hanya 73%. Perbedaan performa ini secara konsisten menunjukkan bahwa penerapan RFE memberikan peningkatan nyata pada kedua algoritma, terutama karena RFE mampu mengeliminasi fitur yang tidak relevan sehingga model dapat lebih fokus mengenali pola risiko diabetes. Selain itu, penelitian juga mengonfirmasi bahwa Glucose dan BMI merupakan fitur paling berpengaruh pada model *Random Forest* dan XGBoost, yang selaras dengan hasil seleksi fitur RF-RFE penelitian ini di mana Glucose dan BMI secara konsisten terpilih sebagai dua dari empat fitur utama. Temuan ini semakin memperkuat validitas pendekatan yang diterapkan dalam penelitian ini, sekaligus menunjukkan bahwa fokus pada populasi berisiko tinggi (BMI ≥ 25) tidak mengurangi relevansi fitur-fitur tersebut, melainkan justru mempertajam kemampuan model dalam mendeteksi kasus diabetes secara lebih akurat.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa kombinasi metode RF dengan Recursive Feature Elimination (RFE) merupakan kombinasi yang paling efektif dalam mengklasifikasi risiko Diabetes Mellitus pada individu dengan BMI ≥ 25. Dibandingkan dengan penelitian-penelitian sebelumnya yang menggunakan dataset yang sama namun tanpa filter populasi dan tanpa seleksi fitur, model RF-RFE dalam penelitian ini mampu menghasilkan akurasi yang kompetitif sebesar 78% sekaligus peningkatan recall kelas diabetes yang lebih signifikan — dari 60% menjadi 68% — sehingga jumlah False Negative dapat ditekan dari 33 menjadi 26 kasus. Keunggulan ini tidak terlepas dari dua kontribusi utama penelitian ini yang belum ada pada studi sebelumnya: pertama, pembatasan populasi pada individu dengan BMI ≥ 25 yang menjadikan pemodelan lebih representatif secara klinis karena berfokus pada kelompok yang memiliki risiko dua hingga empat kali lebih tinggi terhadap diabetes; kedua, penerapan RFE secara komparatif pada dua algoritma sekaligus yang memungkinkan analisis dampak seleksi fitur terhadap masing-masing model secara terukur. Meskipun tingkat akurasi yang dihasilkan tidak setinggi beberapa penelitian yang menggunakan teknik augmentasi data seperti SMOTE, hal ini justru mengindikasikan bahwa model yang dibangun lebih mencerminkan kondisi data yang sesungguhnya dan tidak bergantung pada manipulasi distribusi kelas. Dalam konteks medis, kemampuan model untuk meminimalkan False Negative jauh lebih bernilai dibandingkan sekadar

e-ISSN 2827-8356
p-ISSN 2827-9999

memaksimalkan akurasi keseluruhan, mengingat kegagalan mendeteksi pasien yang benar-benar positif diabetes dapat menunda penanganan dini dan berujung pada komplikasi yang lebih serius.

4. KESIMPULAN

Penelitian ini menganalisis serta membandingkan performa metode SVM dan RF dalam klasifikasi risiko Diabetes Mellitus pada individu dengan BMI ≥ 25, serta mengevaluasi pengaruh penerapan RFE terhadap peningkatan performa kedua model. Hasil eksperimen menunjukkan RF-RFE sebagai model dengan performa terbaik, mencapai *accuracy* 78% dan *F1-score* 72% pada kelas Diabetes, meningkat signifikan dibandingkan RF tanpa RFE (*accuracy* 74%, *recall* 60%). Kontribusi utama penelitian ini bukan terletak pada penerapan RFE itu sendiri — yang telah banyak dieksplorasi pada penelitian sebelumnya — melainkan pada bukti empiris bahwa efektivitas RFE bersifat spesifik-algoritma, bukan seragam, ketika diterapkan pada populasi berisiko klinis tinggi (BMI ≥ 25). Random Forest terbukti memperoleh manfaat yang jauh lebih besar dari seleksi fitur dibandingkan SVM, ditunjukkan oleh peningkatan *recall* kelas Diabetes sebesar 8 poin persentase pada RF (60%→68%) dibandingkan hanya 1 poin persentase pada SVM (60%→61%), serta peningkatan *accuracy* keseluruhan dua kali lebih besar pada RF (+4 poin) dibandingkan SVM (+2 poin). Temuan ini relevan sebagai pertimbangan pemilihan strategi seleksi fitur pada pengembangan sistem deteksi dini diabetes berbasis *machine learning* di masa mendatang, khususnya dalam menekan angka *False Negative* yang berdampak langsung pada keselamatan pasien. Implementasi lengkap disediakan sebagai fondasi yang dapat direproduksi untuk penelitian lanjutan. Untuk penelitian selanjutnya, ada beberapa hal yang layak dieksplorasi lebih jauh. Pertama, pengujian dengan teknik validasi silang seperti *k-fold cross-validation* untuk memperoleh estimasi performa yang lebih *robust*. Kedua, eksplorasi teknik penanganan ketidakseimbangan kelas yang tidak bergantung pada data sintetis, seperti *class weighting* atau *threshold tuning*. Ketiga, perluasan dataset ke populasi yang lebih beragam secara demografis agar model yang dihasilkan memiliki kemampuan generalisasi yang lebih baik. Keempat, penambahan metrik AUC-ROC sebagai pelengkap evaluasi untuk mengukur kemampuan diskriminasi model secara lebih komprehensif.

Terima Kasih

Para penulis menyatakan tidak ada konflik kepentingan dan juga tidak menerima pendanaan khusus dari lembaga manapun, baik dari sektor pemerintah, komersial, maupun nirlaba.

Penulis mengucapkan terima kasih kepada Program Studi Sistem Informasi, Universitas Mercu Buana Yogyakarta, atas dukungan fasilitas akademik selama penelitian ini dilakukan.

DAFTAR PUSTAKA

.

[1] A. Firgiawan, F. N. Andriyansah, R. N. Ramadhan, S. Sumanto, and I. Budiawan, "Analisis Perbandingan Algoritma Random Forest, SVM, dan Logistic Regression untuk Menentukan Model Terbaik Prediksi Penyakit Diabetes," Jurnal Teknik Informatika dan Teknologi Informasi, vol. 5, no. 3, pp. 113–130, des. 2025, doi: 10.55606/jutiti.v5I3.6213.

[2] I. Dzaki Rif, Y. N. Hasneli, and G. Indriati, "Gambaran Komplikasi Diabetes Mellitus pada penderita Diabetes Mellitus," Jurnal Keperawatan Profesional (JKP), vol. 11,no.1,pp.1-18, 2023, doi: 10.33650/jkp.v11i1.5540

[3] C. Z. V. Junus, T. Tarno, and P. Kartikasari, "Klasifikasi menggunakan Metode Support Vector Machine dan Random Forest untuk deteksi awal risiko Diabetes Melitus," Jurnal Gaussian, vol. 11, no. 3, pp. 386–396, jan. 2023, doi: 10.14710/j.gauss.11.3.386-396.

e-ISSN 2827-8356
p-ISSN 2827-9999

[4] N. A. Maori and N. Azizah, “Jurnal Informatika dan rekayasa perangkat lunak penerapan Recursive Feature Elimination (RFE) pada Tree-Based Classifier untuk identifikasi risiko Diabetes,” vol. 6, no. 2, pp. 465–471, sep, 2024, doi: 10.36499/jinrpl.v6i2.11283

[5] N. L. Sabili, R. Fajri, and M. Umbara, “Klasifikasi penyakit Diabetes menggunakan Algoritma Categorical Bossting dengan faktor risiko Diabetes,” Jurnal Mahasiswa Teknik Informatika (JATI), vol.8,no.6,pp.11391-11398,nov. 2024. doi: 10.36040/jati.v8I6.11447.

[6] N. Sulistianingsih, S. A. A. Yusuf, A. Anggreni, F. N. Sari, M. P. Juniarta, and J. Permatasari, “Analisis pengaruh Recursive Feature Elimination terhadap kinerja model prediksi dini Diabetes Mellitus di rs pku muhammadiyah Bima,” JTIM : Jurnal Teknologi Informasi dan Multimedia, vol. 7, no. 3, pp. 548–560, jul. 2025, doi: 10.35746/jtim.v7I3.774.

[7] R. K. Hapsari, A. H. Salim, L. F. Oktavian, and A. R. Fitra, “Classification of Diabetes Mellitus using Decision Trees,” Jurnal Eltikom : Jurnal Teknik Elektro, Teknologi Informasi dan Komputer, vol. 9, no. 2, pp. 107–114, dec. 2025, doi: 10.31961/eltikom.v9I2.1461.

[8] M. Buyukkececi and M. C. Okur, “A Comprehensive review of Feature Selection and Feature Selection stability in Machine Learning,” Gazi University Journal Of Science, vol.36,no.4,pp.1506-1520,dec.2023, doi: 10.35378/gujs.993763.

[9] A. Citra Mawani, Rusdah, L. Li Hin, and D. Anubhakti, “Deteksi dini gejala awal penyakit Diabetes menggunakan Algoritma Random Forest,” Idealis:Indonesia Journal Information System, vol.6,no.2,jul. 2023,doi: 10.36080/idealis.v6i2.3018

[10] K. Syaban and Mardiawati, “Evaluasi model Ensemble Learning pada identifikasi faktor risiko Diabetes Mellitus,” Jurnal Teknologi dan Informasi, vol. 15, no. 2, pp. 121–130, sep. 2025, doi: 10.34010/jati.v15I2.16238.

[11] L. S. Gunawan and C. Dewi, "Optimalisasi model deteksi dini Diabetes dengan teknik Features Selection," JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 10, no. 4, pp. 4312–4323, dec. 2025, doi: 10.29100/jipi.v10I4.7006.

[12] F. Refindha, A. Harianto, Z. Alawi, and I. A. Sa'ida, "Pengaruh komposisi split Data pada akurasi klasifikasi penderita Diabetes menggunakan Algoritma Machine Learning," Jurnal Sistem Informasi dan Informatika (SIMIKA), vol. 8, no. 1, jan. 2025, doi: 10.47080/simika.v8i1.3663

[13] Pratama, A. M. Siregar, S. A. P. Lestari, and S. Faisal, "Implementation of Diabetes prediction model using Random Forest Algorithm, K-Nearest Neighbor, and Logistic Regression," Jurnal Teknik Informatika (JUTIF), vol. 5, no. 4, pp. 1165–1174, sep. 2024, doi: 10.52436/1.jutif.2024.5.4.2593

[14] Y. Insani, M. S. P. Sipayung, and M. F. Naibaho, “Prediksi Diabetes berbasis Decision Tree dengan menggunakan Dataset Pima Indians Diabetes article info,” JDMIS: Journal of Data Mining and Information System, vol. 4, no. 1, pp. 40–45, 2026, doi: 10.54259/jdmis.v4I1.7107

[15] A. Setiawan, D. Mulia Hutabalian, R. Irnanda, H. Fredi, “Evaluasi perbandingan kinerja model Machine Learning untuk prediksi Diabetes: Studi kasus Xgboost, Random Forest, dan SVM”, vol 12,no.2,dec. 2024, doi: 10.56689/infokom.v12i2.2350