

Perbandingan Algoritma Machine Learning untuk Klasifikasi Diabetes Menggunakan Feature Selection dan Hyperparameter Tuning

Fajar Maula Hidayat^{1*}, Hafidz Sanjaya²

^{1,2} Program Studi Ilmu Komputer, Universitas Yayasan Pendidikan Imam Bonjol Majalengka

*fajarmaulahidayat@univypib.ac.id

Abstrak

Diabetes mellitus merupakan salah satu penyakit kronis yang terus mengalami peningkatan setiap tahunnya sehingga diperlukan deteksi dini untuk mendukung penanganan penyakit secara lebih cepat dan tepat. Penelitian ini bertujuan membandingkan performa beberapa algoritma machine learning dalam klasifikasi diabetes menggunakan feature selection dan hyperparameter tuning. Dataset yang digunakan adalah Pima Indians Diabetes Dataset yang diperoleh dari platform Kaggle. Penelitian dilakukan melalui tahapan preprocessing data, feature selection menggunakan SelectKBest, pembagian data training dan testing, hyperparameter tuning menggunakan GridSearchCV, serta evaluasi model menggunakan accuracy, precision, recall, F1-score, ROC-AUC, dan cross validation. Algoritma yang dibandingkan meliputi Logistic Regression, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbor (KNN), dan Naive Bayes. Hasil penelitian menunjukkan bahwa algoritma KNN memperoleh performa terbaik dengan nilai accuracy sebesar 74,02%, precision sebesar 63,46%, recall sebesar 61,11%, F1-score sebesar 62,26%, dan ROC-AUC sebesar 79,60%. Penelitian ini menunjukkan bahwa integrasi preprocessing data, feature selection, hyperparameter tuning, dan cross validation mampu menghasilkan proses evaluasi model yang lebih komprehensif dalam klasifikasi diabetes.

Kata kunci : cross validation, diabetes, feature selection, hyperparameter tuning, klasifikasi, machine learning.

Abstract

Diabetes mellitus is a chronic disease with an increasing prevalence worldwide, requiring early detection to support faster and more accurate disease management. This study aims to compare the performance of several machine learning algorithms for diabetes classification using feature selection and hyperparameter tuning. The dataset used was the Pima Indians Diabetes Dataset obtained from the Kaggle platform. The research consisted of data preprocessing, feature selection using SelectKBest, training and testing data splitting, hyperparameter tuning using GridSearchCV, and model evaluation using accuracy, precision, recall, F1-score, ROC-AUC, and cross validation. The evaluated algorithms included Logistic Regression, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbor (KNN), and Naive Bayes. The results showed that the KNN algorithm achieved the best performance with an accuracy of 74.02%, precision of 63.46%, recall of 61.11%, F1-score of 62.26%, and ROC-AUC of 79.60%. The findings indicate that integrating data preprocessing, feature selection, hyperparameter tuning, and cross validation provides a more comprehensive evaluation process for machine learning models in diabetes classification.

Keywords : cross validation, diabetes, feature selection, hyperparameter tuning, classification, machine learning

1. Pendahuluan

Diabetes mellitus merupakan salah satu penyakit kronis yang menjadi perhatian utama dalam bidang kesehatan global karena dapat menyebabkan berbagai komplikasi serius, seperti

gangguan kardiovaskular, kerusakan ginjal, gangguan saraf, hingga kebutaan apabila tidak ditangani dengan baik. Menurut *World Health Organization* (WHO), jumlah penderita diabetes terus mengalami peningkatan setiap tahunnya

dan menjadi salah satu penyebab utama kematian di dunia [1]. Selain itu, *International Diabetes Federation* (IDF) memperkirakan prevalensi diabetes akan terus meningkat akibat pola hidup tidak sehat, kurangnya aktivitas fisik, dan faktor keturunan [2]. Di Indonesia, diabetes juga termasuk salah satu penyakit tidak menular dengan jumlah kasus yang tinggi sehingga diperlukan upaya deteksi dini untuk mendukung penanganan yang lebih cepat dan tepat [3], [4], [5].

Perkembangan teknologi informasi dan kecerdasan buatan mendorong pemanfaatan machine learning dalam pengembangan sistem deteksi dini penyakit, termasuk diabetes [6], [7], [8]. Berbagai algoritma *machine learning* seperti *Logistic Regression*, *Support Vector Machine* (SVM), *Random Forest*, *K-Nearest Neighbor* (KNN), dan *Naive Bayes* telah banyak diterapkan untuk klasifikasi diabetes dengan tingkat performa yang bervariasi [8], [9], [10], [11]. Beberapa penelitian menunjukkan bahwa algoritma *Random Forest*, SVM, dan KNN mampu memberikan performa yang baik, terutama setelah dikombinasikan dengan tahapan *preprocessing* dan optimasi model [9], [12].

Meskipun berbagai penelitian telah menunjukkan bahwa penerapan *machine learning* mampu memberikan hasil yang baik dalam klasifikasi diabetes, masih terdapat beberapa keterbatasan yang perlu diperhatikan. Sebagian penelitian masih menggunakan parameter *default* tanpa

proses optimasi sehingga performa klasifikasi belum optimal [6], [13], [14]. Selain itu, penerapan *feature selection* pada beberapa penelitian belum dilakukan secara maksimal sehingga masih terdapat kemungkinan penggunaan fitur yang kurang relevan [10], [15], [16], [17]. Permasalahan lain yang juga sering ditemukan adalah kurang optimalnya tahapan *preprocessing* serta penanganan ketidakseimbangan data, yang dapat memengaruhi performa model klasifikasi [7], [12], [18], [19]. Selain itu, sebagian besar penelitian masih berfokus pada penerapan salah satu teknik optimasi secara terpisah, sehingga belum banyak penelitian yang mengintegrasikan *preprocessing*, *feature selection*, *hyperparameter tuning*, dan *cross validation* dalam satu alur eksperimen untuk membandingkan performa beberapa algoritma klasifikasi diabetes secara komprehensif.

Berdasarkan permasalahan tersebut, penelitian ini membandingkan performa beberapa algoritma *machine learning* untuk klasifikasi diabetes dengan menerapkan *preprocessing* data, *feature selection* menggunakan *SelectKBest*, serta *hyperparameter tuning* menggunakan *GridSearchCV*. Algoritma yang dibandingkan meliputi *Logistic Regression*, *Support Vector Machine* (SVM), *Random Forest*, *K-Nearest Neighbor* (KNN), dan *Naive Bayes*. Evaluasi model dilakukan menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC* untuk menentukan algoritma dengan performa terbaik.

Penelitian ini menggunakan Pima Indians Diabetes Dataset, yang telah banyak digunakan sebagai *benchmark* dalam penelitian klasifikasi diabetes, sehingga hasil penelitian dapat dibandingkan secara lebih objektif dengan penelitian sebelumnya. Kontribusi penelitian ini terletak pada penerapan satu alur eksperimen yang mengintegrasikan *preprocessing* data, *feature selection*, *hyperparameter tuning*, dan *cross validation* untuk mengevaluasi serta membandingkan performa beberapa algoritma *machine learning* secara komprehensif.

2. Tinjauan Pustaka

2.1. Penelitian Terkait

Penelitian mengenai klasifikasi diabetes menggunakan *machine learning* telah banyak dilakukan dengan berbagai algoritma dan pendekatan optimasi. Beberapa penelitian terdahulu yang relevan dengan penelitian ini dijelaskan sebagai berikut.

1. Sudestra et al. membandingkan beberapa algoritma *machine learning* seperti *Random Forest*, *Support Vector Machine* (SVM), dan *K-Nearest Neighbor* (KNN) untuk prediksi diabetes.

Hasil penelitian menunjukkan bahwa setiap algoritma memiliki performa yang berbeda bergantung pada karakteristik dataset yang digunakan [9].

2. Phongying dan Hirrote menerapkan beberapa algoritma *machine learning* untuk klasifikasi

diabetes dan menunjukkan bahwa algoritma *Decision Tree*, KNN, *Random Forest*, dan SVM mampu memberikan performa klasifikasi yang cukup baik pada dataset diabetes [11].

3. Nugraha dan Sasongko menerapkan *hyperparameter tuning* menggunakan *Grid Search* untuk meningkatkan performa algoritma klasifikasi. Hasil penelitian menunjukkan bahwa optimasi parameter mampu meningkatkan hasil klasifikasi dibandingkan penggunaan parameter *default* [6].

4. Ariansyah dan Hermanto menunjukkan bahwa penerapan *feature selection* dan *hyperparameter tuning* pada algoritma *Support Vector Machine* mampu meningkatkan performa klasifikasi diabetes [14].

5. Anggreini et al. menerapkan algoritma *Random Forest* dengan *hyperparameter tuning* dan memperoleh peningkatan performa model pada klasifikasi diabetes [12].

6. Samsuri menerapkan metode *Forward Selection* pada algoritma *Naive Bayes* dan menunjukkan bahwa pemilihan fitur yang relevan dapat meningkatkan akurasi klasifikasi diabetes [15].

7. Saxena et al. menunjukkan bahwa penerapan *feature selection* mampu mengurangi penggunaan fitur yang kurang relevan sehingga dapat meningkatkan performa model *machine learning* dalam klasifikasi diabetes [10].

8. Kaliappan et al. menambahkan bahwa

penerapan *feature selection* dan *Explainable Artificial Intelligence* (XAI) dapat meningkatkan interpretabilitas model machine learning pada klasifikasi diabetes [16].

9. Khairunnisa et al. menerapkan *preprocessing* data dan *hyperparameter tuning* menggunakan *Optuna* pada algoritma *Extreme Gradient Boosting* untuk meningkatkan performa prediksi diabetes [18].

10. Sardiarinto et al. menunjukkan bahwa penerapan *preprocessing* data, *SMOTE*, dan *feature selection* mampu meningkatkan kualitas klasifikasi diabetes menggunakan algoritma *Random Forest* [19].

Berdasarkan penelitian-penelitian tersebut, dapat diketahui bahwa *machine learning* memiliki potensi yang baik dalam klasifikasi diabetes. Namun, sebagian besar penelitian masih berfokus pada penerapan salah satu teknik optimasi, seperti *feature selection*, *hyperparameter tuning*, atau *preprocessing* secara terpisah. Belum banyak penelitian yang mengintegrasikan tahapan *preprocessing* data, *feature selection*, *hyperparameter tuning*, dan *cross validation* dalam satu alur eksperimen untuk membandingkan performa beberapa algoritma *machine learning* pada Pima Indians Diabetes Dataset. Oleh karena itu, penelitian ini mengusulkan perbandingan beberapa algoritma *machine learning* menggunakan *feature selection*

dan *hyperparameter tuning* untuk memperoleh model klasifikasi diabetes yang lebih optimal.

2.2. Landasan Teori

1. Machine Learning

Machine learning merupakan salah satu cabang *artificial intelligence* yang memungkinkan sistem komputer mempelajari pola dari data historis untuk menghasilkan prediksi atau keputusan secara otomatis. Dalam bidang kesehatan, *machine learning* telah banyak dimanfaatkan untuk mendukung proses diagnosis, klasifikasi, dan prediksi berbagai penyakit, termasuk diabetes, karena mampu mengidentifikasi pola yang sulit dikenali melalui analisis konvensional [6].

[7], [8]. Pada penelitian ini digunakan lima algoritma *machine learning*, yaitu *Logistic Regression*, *Support Vector Machine* (SVM), *Random Forest*, *K-Nearest Neighbor* (KNN), dan *Naive Bayes*. Kelima algoritma tersebut dipilih karena telah banyak digunakan dalam penelitian klasifikasi diabetes serta menunjukkan performa yang baik pada berbagai karakteristik dataset [9], [10], [11], [12].

2. Feature Selection

Feature selection merupakan teknik yang digunakan untuk memilih fitur-fitur yang paling relevan terhadap variabel target sehingga dapat meningkatkan efisiensi proses pembelajaran model, mengurangi kompleksitas data, serta meminimalkan pengaruh fitur yang kurang penting. Pada penelitian ini, proses *feature*

selection dilakukan menggunakan metode *SelectKBest* untuk memperoleh atribut yang memiliki kontribusi terbesar terhadap proses klasifikasi diabetes [10], [15], [16], [17]. Penerapan teknik *feature selection* juga telah banyak digunakan pada berbagai penelitian *machine learning* karena mampu meningkatkan performa model melalui pemilihan atribut yang paling berpengaruh sekaligus mengurangi kompleksitas komputasi. Selain itu, penelitian yang dipublikasikan pada Jurnal Infotek menunjukkan bahwa pemilihan metode *feature selection* yang tepat dapat meningkatkan kualitas proses klasifikasi dan prediksi pada berbagai kasus *machine learning* [20], [21].

3. Hyperparameter Tuning

Hyperparameter tuning merupakan proses optimasi parameter model yang dilakukan sebelum proses pelatihan agar diperoleh konfigurasi parameter terbaik. Optimasi *hyperparameter* dapat meningkatkan kemampuan model dalam melakukan klasifikasi dibandingkan penggunaan parameter *default*. Penelitian ini menggunakan metode *GridSearchCV* karena mampu mengevaluasi berbagai kombinasi parameter secara sistematis untuk memperoleh konfigurasi model yang optimal [6], [12], [13], [14].

4. Evaluasi Kinerja Model

Evaluasi kinerja model bertujuan untuk mengetahui kemampuan masing-masing

algoritma dalam melakukan klasifikasi diabetes. Pada penelitian ini, evaluasi dilakukan menggunakan *confusion matrix*, *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*. Kombinasi metrik evaluasi tersebut digunakan agar performa model tidak hanya dinilai berdasarkan tingkat akurasi, tetapi juga kemampuan model dalam mengidentifikasi kelas positif secara tepat dan konsisten [9], [11], [12], [15], [19]

3. Metode Penelitian

3.1. Desain Penelitian

Penelitian ini menggunakan metode eksperimen dengan pendekatan kuantitatif untuk membandingkan performa beberapa algoritma *machine learning* dalam klasifikasi diabetes menggunakan *feature selection* dan *hyperparameter tuning*. Metode eksperimen dipilih karena penelitian berfokus pada proses pengujian dan evaluasi performa model *machine learning* berdasarkan hasil klasifikasi yang diperoleh dari dataset diabetes [6], [9], [11].

3.2. Sumber Data

Sumber data yang digunakan dalam penelitian ini adalah Pima Indians Diabetes Dataset yang tersedia pada platform Kaggle dan telah banyak digunakan sebagai *benchmark* dalam penelitian klasifikasi diabetes berbasis *machine learning* [22]. Dataset terdiri atas beberapa atribut medis, yaitu *Pregnancies*, *Glucose*, *BloodPressure*,

SkinThickness, *Insulin*, *BMI*, *Diabetes Pedigree Function*, *Age*, dan *Outcome* sebagai label klasifikasi. Dataset tersebut digunakan sebagai data utama dalam proses pelatihan dan pengujian model.

3.3. Instrumen Penelitian

Instrumen penelitian berupa perangkat lunak yang digunakan dalam proses pengolahan dan analisis data, yaitu bahasa pemrograman Python beserta pustaka pendukung seperti *Pandas*, *NumPy*, *Scikit-learn*, dan *Matplotlib*. Seluruh proses eksperimen dilakukan menggunakan lingkungan komputasi yang mendukung implementasi algoritma *machine learning*, *feature selection*, *hyperparameter tuning*, serta evaluasi performa model.

3.4. Teknik Pengumpulan Data

Teknik pengumpulan data dilakukan melalui studi dokumentasi dengan memanfaatkan Pima Indians Diabetes Dataset yang diperoleh dari platform Kaggle. Dataset digunakan sebagai data sekunder tanpa melibatkan proses pengambilan data secara langsung dari objek penelitian. Seluruh data kemudian dipersiapkan untuk proses eksperimen klasifikasi menggunakan algoritma *machine learning*.

3.5. Tahapan Penelitian

Tahapan penelitian diawali dengan proses

preprocessing data yang meliputi pengecekan *missing value*, penanganan nilai abnormal, dan normalisasi data [7], [12], [18], [19]. Penanganan nilai abnormal dilakukan pada beberapa atribut medis yang memiliki nilai nol tidak realistis, yaitu *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, dan *BMI*, dengan mengubah nilai tersebut menjadi *missing value* kemudian dilakukan imputasi menggunakan nilai rata-rata.

Setelah proses *preprocessing* selesai, dilakukan *feature selection* menggunakan metode *SelectKBest* untuk memilih fitur yang paling relevan terhadap klasifikasi diabetes [10], [15], [16], [17]. Selanjutnya dataset dibagi menjadi data *training* dan data *testing* menggunakan metode *train_test_split* dengan proporsi 80% untuk data *training* dan 20% untuk data *testing*.

Penelitian ini menggunakan lima algoritma *machine learning*, yaitu *Logistic Regression*, *Support Vector Machine* (SVM), *Random Forest*, *K-Nearest Neighbor* (KNN), dan *Naive Bayes* [9], [10], [11]. Untuk meningkatkan performa model dilakukan *hyperparameter tuning* menggunakan *GridSearchCV* pada algoritma *Random Forest*, SVM, dan KNN guna memperoleh parameter optimal dari masing-masing model [6], [12], [13], [14]. Proses *hyperparameter tuning* difokuskan pada ketiga algoritma tersebut karena memiliki sensitivitas parameter yang lebih tinggi dibandingkan *Logistic Regression* dan *Naive Bayes*, sehingga optimasi parameter diharapkan

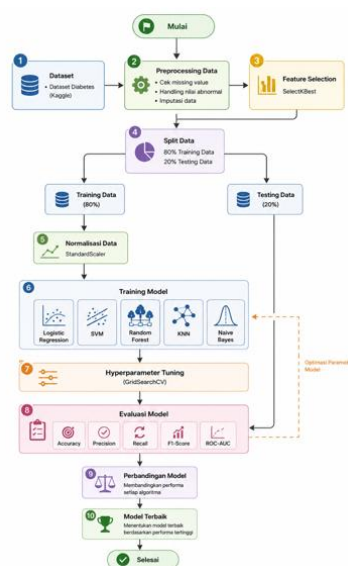
dapat memberikan peningkatan performa yang lebih signifikan.

3.6. Teknik Analisis Data

Analisis data dilakukan dengan mengevaluasi performa masing-masing algoritma menggunakan *confusion matrix*, *accuracy*, *precision*, *recall*, *F1-score*, *ROC-AUC*, dan *cross validation* [6], [9], [12], [19]. Penggunaan beberapa metrik evaluasi bertujuan untuk memperoleh gambaran yang lebih komprehensif mengenai kemampuan model dalam melakukan klasifikasi diabetes. Selanjutnya dilakukan perbandingan hasil evaluasi untuk menentukan algoritma yang memiliki performa terbaik.

3.7. Flowchart Penelitian

Tahapan penelitian yang dilakukan pada penelitian ini ditunjukkan pada Gambar 1.



Gambar 1. Flowchart tahapan penelitian

Berdasarkan flowchart pada Gambar 1, penelitian diawali dengan pengumpulan Pima Indians Diabetes Dataset, kemudian dilanjutkan dengan proses *preprocessing* data, *feature selection*, pembagian data *training* dan *testing*, normalisasi data, pelatihan model, *hyperparameter tuning*, evaluasi model, serta perbandingan performa masing-masing algoritma untuk menentukan model terbaik dalam klasifikasi diabetes.

4. Hasil dan Pembahasan

4.1 Hasil Penelitian

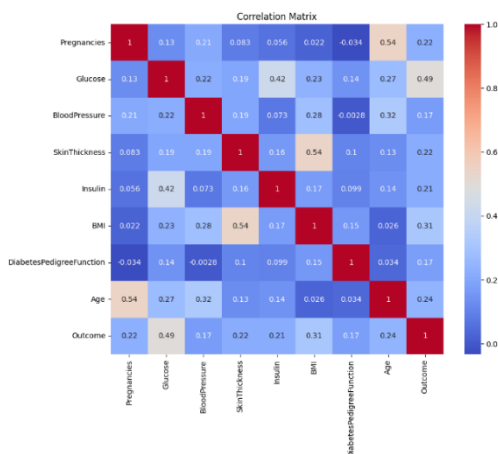
1. Hasil Preprocessing Data

Tahap awal penelitian dilakukan melalui proses *preprocessing* data untuk meningkatkan kualitas dataset sebelum digunakan dalam proses pelatihan model *machine learning*. Tahapan *preprocessing* meliputi pengecekan *missing value*, penanganan nilai abnormal, imputasi data, dan normalisasi data [7], [12], [18], [19].

Pada penelitian ini ditemukan beberapa atribut medis yang memiliki nilai nol yang tidak realistis, yaitu *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, dan *BMI*. Nilai-nilai tersebut diubah menjadi *missing value* kemudian dilakukan imputasi menggunakan nilai rata-rata agar kualitas data menjadi lebih baik. Selanjutnya dilakukan normalisasi data menggunakan *StandardScaler* untuk menyamakan skala antar atribut sehingga proses pelatihan model dapat

berjalan lebih optimal.

Hasil *preprocessing* menunjukkan bahwa dataset telah siap digunakan pada tahap *feature selection* dan pelatihan model *machine learning*. Tahapan ini juga membantu mengurangi potensi bias data serta meningkatkan kualitas dataset sebelum dilakukan proses klasifikasi. Hubungan antar atribut pada dataset diabetes ditunjukkan pada Gambar 2.



Gambar 2. Correlation matrix dataset diabetes

Berdasarkan Gambar 2, atribut *Glucose* dan *BMI* menunjukkan hubungan yang relatif lebih tinggi terhadap variabel *Outcome* dibandingkan atribut lainnya. Hal tersebut mengindikasikan bahwa kedua atribut memiliki kontribusi yang cukup penting dalam proses klasifikasi diabetes.

2. Hasil Feature Selection

Setelah proses *preprocessing* selesai dilakukan, tahap berikutnya adalah *feature selection* menggunakan metode *SelectKBest* untuk memilih fitur yang paling relevan terhadap klasifikasi diabetes [10], [15], [16], [17]. Penerapan *feature*

selection bertujuan untuk mengurangi penggunaan fitur yang kurang relevan sehingga *model machine learning* dapat bekerja secara lebih efisien.

Hasil *feature selection* menunjukkan bahwa atribut *Glucose*, *BMI*, *Age*, *Pregnancies*, dan *DiabetesPedigreeFunction* merupakan fitur-fitur yang memiliki kontribusi paling besar terhadap proses klasifikasi diabetes. Penggunaan *feature selection* membantu meningkatkan efisiensi proses pelatihan maupun pengujian model dengan memfokuskan proses klasifikasi pada atribut yang paling relevan.

3. Hasil Pelatihan dan Evaluasi Model

Penelitian ini menggunakan lima algoritma *machine learning*, yaitu *Logistic Regression*, *Support Vector Machine (SVM)*, *Random Forest*, *K-Nearest Neighbor (KNN)*, dan *Naive Bayes* [9], [10], [11]. Untuk meningkatkan performa model dilakukan *hyperparameter tuning* menggunakan *GridSearchCV* pada algoritma *Random Forest*, *SVM*, dan *KNN* [6], [12], [13], [14].

Evaluasi model dilakukan menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*. Hasil evaluasi masing-masing algoritma ditunjukkan pada Tabel 1.

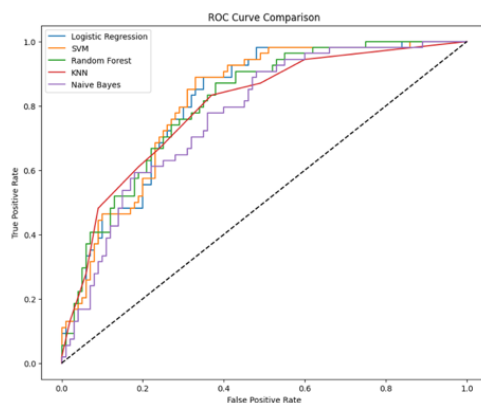
Tabel 1. Hasil Evaluasi Model Machine Learning

Model	Accur acy	Preci sion	Recall	F1- Score	ROC- AUC
KNN	0.740	0.634	0.611	0.622	0.796
	260	615	111	642	019

Model	Accur acy	Preci sion	Recall	F1- Score	ROC- AUC
Random	0.727	0.642	0.500	0.562	0.800
Forest	273	857	000	500	370
SVM	0.707	0.600	0.500	0.545	0.813
	792	000	000	455	796
Logistic	0.694	0.577	0.481	0.525	0.811
Regress	805	778	481	253	667
ion					
Naive	0.694	0.557	0.629	0.591	0.769
Bayes	805	377	630	304	630

Berdasarkan hasil evaluasi pada Tabel 1, algoritma *K-Nearest Neighbor* (KNN) memperoleh performa terbaik dengan nilai *accuracy* sebesar 74,02%, *precision* sebesar 63,46%, *recall* sebesar 61,11%, *F1-score* sebesar 62,26%, dan *ROC-AUC* sebesar 79,60%. Hasil tersebut menunjukkan bahwa kombinasi *preprocessing* data, *feature selection*, *hyperparameter tuning*, dan *cross validation* mampu mendukung peningkatan performa model *machine learning* dalam klasifikasi diabetes.

Perbandingan performa model berdasarkan nilai *ROC-AUC* ditunjukkan pada Gambar 3.



Gambar 3. ROC curve comparison

4. Hasil Cross Validation

Untuk mengetahui kestabilan performa model, penelitian ini juga menerapkan *cross validation* pada proses evaluasi [6], [14], [19]. Penerapan *cross validation* bertujuan untuk mengurangi bias akibat pembagian data *training* dan *testing* sehingga hasil evaluasi model menjadi lebih stabil.

Hasil *cross validation* menunjukkan bahwa algoritma KNN memiliki performa yang relatif konsisten pada beberapa *fold* pengujian. Hal ini menunjukkan bahwa model yang dihasilkan memiliki kemampuan generalisasi yang baik dalam melakukan klasifikasi diabetes

4.2 Pembahasan

Hasil penelitian menunjukkan bahwa tahapan *preprocessing* memberikan kontribusi terhadap peningkatan kualitas dataset sebelum digunakan pada proses pelatihan model *machine learning*. Penanganan nilai abnormal, proses imputasi, serta normalisasi data menggunakan *StandardScaler* membantu menghasilkan data yang lebih representatif sehingga model dapat mempelajari pola data secara lebih optimal [7], [12], [18], [19]. Temuan ini sejalan dengan penelitian Khairunnisa et al. [18] dan Sardiarinto et al. [19] yang menunjukkan bahwa kualitas *preprocessing* berpengaruh terhadap performa model klasifikasi. Penerapan *feature selection* menggunakan metode *SelectKBest* juga memberikan manfaat dalam memilih atribut yang paling relevan

terhadap klasifikasi diabetes. Hasil penelitian menunjukkan bahwa atribut *Glucose*, *BMI*, *Age*, *Pregnancies*, dan *DiabetesPedigreeFunction* merupakan atribut yang memberikan kontribusi terbesar terhadap proses klasifikasi. Temuan ini mendukung penelitian Saxena et al. [10], Samsuri [15], serta Kaliappan et al. [16] yang menyatakan bahwa *feature selection* mampu meningkatkan efisiensi model sekaligus mengurangi penggunaan atribut yang kurang relevan.

Selain itu, penerapan *hyperparameter tuning* menggunakan *GridSearchCV* memberikan konfigurasi parameter yang lebih optimal pada algoritma *Random Forest*, *SVM*, dan *KNN*. Hasil tersebut konsisten dengan penelitian Nugraha dan Sasongko [6], Anggreini et al. [12], serta Ariansyah dan Hermanto [14] yang menunjukkan bahwa optimasi *hyperparameter* mampu meningkatkan performa model dibandingkan penggunaan parameter *default*.

Berdasarkan hasil evaluasi, algoritma *K-Nearest Neighbor* (*KNN*) memperoleh performa terbaik dibandingkan algoritma lainnya dengan nilai *accuracy* sebesar 74,02%. Hasil tersebut sejalan dengan beberapa penelitian sebelumnya yang menunjukkan bahwa algoritma *KNN* memiliki performa yang baik pada klasifikasi diabetes [9], [10], [11]. Meskipun nilai *accuracy* yang diperoleh belum menjadi yang tertinggi dibandingkan beberapa penelitian lain, perbedaan hasil tersebut dapat dipengaruhi oleh karakteristik dataset,

tahapan *preprocessing*, strategi pembagian data, metode *feature selection*, serta konfigurasi *hyperparameter* yang digunakan.

Secara keseluruhan, penelitian ini menunjukkan bahwa integrasi *preprocessing data*, *feature selection* menggunakan *SelectKBest*, *hyperparameter tuning* menggunakan *GridSearchCV*, dan *cross validation* mampu menghasilkan proses evaluasi yang lebih komprehensif terhadap lima algoritma *machine learning* pada **Pima Indians Diabetes Dataset**. Kontribusi penelitian ini tidak hanya terletak pada perbandingan performa beberapa algoritma, tetapi juga pada penerapan satu alur eksperimen yang terintegrasi sehingga dapat menjadi referensi bagi penelitian selanjutnya dalam pengembangan model klasifikasi diabetes.

5. Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan bahwa penerapan *machine learning* dengan kombinasi *preprocessing data*, *feature selection*, *hyperparameter tuning*, dan *cross validation* mampu menghasilkan proses klasifikasi diabetes yang lebih optimal. Penelitian ini membandingkan lima algoritma *machine learning*, yaitu *Logistic Regression*, *Support Vector Machine* (*SVM*), *Random Forest*, *K-Nearest Neighbor* (*KNN*), dan *Naive Bayes* menggunakan *Pima Indians Diabetes Dataset*.

Hasil penelitian menunjukkan bahwa algoritma *K-Nearest Neighbor* (KNN) memperoleh performa terbaik dengan *accuracy* sebesar 74,02%, *precision* sebesar 63,46%, *recall* sebesar 61,11%, *F1-score* sebesar 62,26%, dan *ROC-AUC* sebesar 79,60%. Integrasi *preprocessing* data, *feature selection* menggunakan *SelectKBest*, *hyperparameter tuning* menggunakan *GridSearchCV*, dan *cross validation* menghasilkan proses evaluasi model yang lebih komprehensif serta menjadi kontribusi utama penelitian ini.

Penelitian ini masih memiliki keterbatasan pada penggunaan dataset dengan jumlah atribut dan data yang terbatas serta belum menerapkan metode *ensemble learning* maupun *deep learning*. Oleh karena itu, penelitian selanjutnya disarankan menggunakan dataset yang lebih besar dan beragam, menerapkan *feature engineering*, teknik penanganan *imbalanced dataset*, serta mengeksplorasi metode *ensemble learning* atau *deep learning* untuk meningkatkan performa klasifikasi diabetes

6. Daftar Pustaka

- [1] World Health Organization, "Diabetes," WHO.
- [2] The Diabetes Atlas, "589 million adults (20-79 years) are living with diabetes worldwide," International Diabetes Federation.
- [3] Aji Muhawarman, "Wamenkes Dante Soroti Lonjakan Diabetes pada Remaja: Gaya Hidup jadi Faktor Utama," Kementerian Kesehatan Republik Indonesia.
- [4] Aji Muhawarman, "Pemerintah Kebut Skrining dan Diagnosis Diabetes untuk 50 Juta Anak Sekolah," Kementerian Kesehatan Republik Indonesia.
- [5] Aji Muhawarman, "Pemerintah Perkuat Deteksi Dini Melalui CKG, dan Layanan Perawatan Diabetes di Indonesia," Kementerian Kesehatan Republik Indonesia.
- [6] W. Nugraha and A. Sasongko, "Hyperparameter Tuning on Classification Algorithm with Grid Search," SISTEMASI, vol. 11, no. 2, p. 391, May 2022, doi: 10.32520/stmsi.v11i2.1750.
- [7] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad, and M. Husnain, "Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques," Mobile Information Systems, vol. 2022, pp. 1–16, Mar. 2022, doi: 10.1155/2022/6521532.
- [8] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," Mater. Today Proc., vol. 80, pp. 3200–3203, 2023, doi: 10.1016/j.matpr.2021.07.196.
- [9] I. M. A. Sudestra, A. W. O. Gama, G. H. Prathama, I. G. N. D. Paramartha, and M. Hakimi, "Comparative Performance of Machine Learning Algorithms for Diabetes Prediction," Journal of Technology and Informatics (JoTI), vol. 8, no. 1, pp. 50–61, Apr. 2026, doi: 10.37802/joti.v8i1.1195.
- [10] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods," Comput. Intell. Neurosci., vol. 2022, pp. 1–11, Apr. 2022, doi: 10.1155/2022/3820360.
- [11] M. Phongying and S. Hiriote, "Diabetes Classification Using Machine Learning Techniques," Computation, vol. 11, no. 5, p.

- 96, May 2023, doi: 10.3390/computation11050096.
- [12] N. L. Anggreini, A. Yuliana, D. S. Ramdan, and W. Al-Dayyeni, "Improving Diabetes Prediction Performance Using Random Forest Classifier with Hyperparameter Tuning," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 4, pp. 1847–1860, Aug. 2025, doi: 10.52436/1.jutif.2025.6.4.4755.
- [13] I. F. Ashari and M. C. Untoro, "Hyperparameter Tuning Feature Selection with Genetic Algorithm and Gaussian Naïve Bayes for Diabetes Disease Prediction," *Jurnal Telematika*, vol. 17, no. 1, pp. 17–23, Oct. 2022, doi: 10.61769/telematika.v17i1.488.
- [14] N. Ariansyah and D. Harmanto, "Klasifikasi Risiko Diabetes Menggunakan Support Vector Machine Berbasis Feature Selection dan Hyperparameter Tuning," *Riau Jurnal Teknik Informatika*, vol. 5, no. 1, pp. 60–67, Mar. 2026, doi: 10.30606/rjti.v5i1.4301.
- [15] A. A. F. Samsuri, "Application Of Naïve Bayes Algorithm Using Forward Selection Features In Diabetes," *Antivirus: Jurnal Ilmiah Teknik Informatika*, vol. 19, no. 1, pp. 83–91, 2025.
- [16] J. Kaliappan et al., "Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets," *Front. Artif. Intell.*, vol. 7, Aug. 2024, doi: 10.3389/frai.2024.1421751.
- [17] A. A. Alhussan et al., "Classification of Diabetes Using Feature Selection and Hybrid Al-Biruni Earth Radius and Dipper Throated Optimization," *Diagnostics*, vol. 13, no. 12, p. 2038, Jun. 2023, doi: 10.3390/diagnostics13122038.
- [18] R. D. Khairunnisa, W. Witanti, and A. Komarudin, "Penerapan Algoritma Machine Learning Untuk Prediksi Penyakit Diabetes Militus Menggunakan Metode Extreme Gradient Boosting," *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 11, no. 2, pp. 984–995, 2026.
- [19] V. M. Vadya, Sardiarinto, and Eko Saputro, "Aplikasi Prediksi Diabetes Berbasis Android Menggunakan Algoritma Random Forest dengan SMOTE dan Feature Selection," *Evolusi: Jurnal Sains dan Manajemen*, vol. 13, no. 2, pp. 84–93, Sep. 2025, doi: 10.31294/evolusi.v13i2.10100.
- [20] Y. Yuliani, "Algoritma Random Forest Untuk Prediksi Kelangsungan Hidup Pasien Gagal Jantung Menggunakan Seleksi Fitur Bestfirst," *Infotek: Jurnal Informatika dan Teknologi*, vol. 5, no. 2, pp. 298–306, Jul. 2022, doi: 10.29408/jit.v5i2.5896.
- [21] R. Ahmad, "Survei Teknik Pemilihan Fitur Untuk Sistem Deteksi Intrusi Berbasis Machine Learning," *Infotek: Jurnal Informatika dan Teknologi*, vol. 8, no. 1, pp. 317–323, Jan. 2025, doi: 10.29408/jit.v8i1.28657.
- [22] UCI Machine Learning Repository, "Pima Indians Diabetes Database," UCI Machine Learning Repository. Accessed: Apr. 27, 2026. [Online]. Available: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>.