Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Application of Diabetes Risk Prediction Using Machine Learning Algorithms

Valentino Dikha Rizaldi[1*], Fadhil Widjonarko[2], Dimas Prasetia[3], Muhammad Ifan Rifani Ihsan[4]

[1,2,3,4] *Faculty of Engineering and Informatics, Universitas Bina Sarana Informatika, Indonesia*
15230272@bsi.ac.id [1*], 15230357@bsi.ac.id [2], 15230414@bsi.ac.id [3], ifan.mii@bsi.ac.id [4]

Abstract

Diabetes mellitus is a chronic disease that poses a significant global health burden, requiring effective early detection strategies to reduce complications and mortality. In recent years, machine learning techniques have been widely applied to support medical decision-making, particularly in disease risk prediction. This study aims to compare the performance of several machine learning algorithms for diabetes risk prediction and to implement the best-performing model into a web-based application. The PIMA Indians Diabetes Dataset was used in this study, and data preprocessing was conducted to address class imbalance and improve model performance. Five classification algorithms were evaluated, namely Logistic Regression, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes. Model performance was assessed using accuracy, recall, F1-score, and Area Under the Curve (AUC), with a particular emphasis on recall and F1-score due to their importance in medical screening applications. Experimental results show that the SVM model outperformed the other algorithms, achieving higher recall, F1-score, and AUC values. The selected model was then implemented into a web-based application using the Streamlit framework, enabling users to input clinical parameters and obtain real-time diabetes risk predictions. The results indicate that machine learning models, particularly SVM, can effectively support diabetes risk prediction and demonstrate the potential of integrating predictive models into practical healthcare applications.

Keywords: *Diabetes mellitus; Machine learning; Risk prediction; Streamlit*; *Support vector machine*

1. Introduction

Diabetes mellitus is one of the most critical global health challenges, characterized by a continuous increase in prevalence and mortality rates worldwide. According to reports from the International Diabetes Federation (IDF) and the World Health Organization (WHO), diabetes has become a major contributor to non-communicable disease burden, leading to severe complications such as cardiovascular disease, neuropathy, nephropathy, and premature death [1][2]. In Indonesia, this situation is equally concerning. National health statistics indicate that the prevalence of diabetes among adults has increased significantly in recent years, placing Indonesia among the countries with the highest number of diabetes cases globally [3]. These conditions highlight the urgent need for effective strategies that support early detection and preventive intervention.

Early identification of individuals at high risk of diabetes is essential to reduce disease progression and long-term healthcare costs. Timely intervention through lifestyle modification and medical treatment has been shown to significantly lower the risk of complications and improve patient quality of life [1]. However, traditional screening methods often rely on limited clinical indicators and manual assessment, which may not adequately capture complex interactions among multiple risk factors. As a result, there is growing interest in data-driven approaches that can enhance risk stratification and clinical decision support.

Recent advances in machine learning (ML) have enabled the development of predictive models capable of extracting meaningful patterns from large and complex clinical datasets. ML algorithms can analyze nonlinear relationships among multiple variables and have demonstrated promising performance in medical prediction tasks, including diabetes risk assessment [4][5]. Several studies have applied classification algorithms such as K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, and Support Vector Machine (SVM) to predict diabetes using health-related datasets. While many of these studies report high accuracy, a common limitation is the overreliance on accuracy as the primary evaluation metric [6].

In medical screening contexts, accuracy alone is insufficient, particularly when dealing with imbalanced datasets where the number of non-diabetic cases exceeds diabetic cases. High accuracy does not necessarily indicate good performance in detecting positive cases, which is critical in disease screening. Metrics such as recall, F1-score, and Area Under the Curve (AUC) provide a more reliable assessment of a

model's sensitivity to minority classes and its overall discriminative ability [7]. Failure to prioritize these metrics may result in a high number of false negatives, posing serious clinical risks.

Another important limitation observed in prior research is the gap between model development and real-world implementation. Many machine learning models with strong theoretical performance are not deployed into practical systems that can be used by healthcare practitioners or the general public [8]. Without accessible and user-friendly implementation, the potential benefits of predictive analytics remain largely unrealized.

Based on these challenges, this study aims to develop a diabetes risk prediction model that emphasizes clinically relevant evaluation metrics, particularly recall and F1-score, to improve sensitivity toward diabetic cases. Five machine learning algorithms—Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbors, and Naive Bayes—are evaluated using the PIMA Indians Diabetes dataset. Furthermore, the best-performing model is implemented in a web-based application using Streamlit to provide real-time risk prediction. By combining performance-oriented model evaluation with practical deployment, this research seeks to bridge the gap between analytical modeling and real-world application in diabetes risk assessment.

2. Literature Review

The application of machine learning techniques in healthcare has increased significantly in recent years, particularly for disease risk prediction tasks. Machine learning models are capable of processing large-scale health data and identifying complex patterns that are often difficult to capture using traditional statistical approaches [4][5]. These capabilities make machine learning a promising approach for supporting early disease detection and risk classification, including diabetes mellitus.

Several studies have explored the use of classical classification algorithms for diabetes risk prediction. Logistic Regression is commonly used as a baseline model due to its simplicity, interpretability, and suitability for binary classification problems in medical research [4]. However, Logistic Regression is limited in its ability to model nonlinear relationships among clinical features, which may reduce its predictive performance when applied to complex health datasets.

To overcome these limitations, more advanced algorithms such as Support Vector Machine (SVM) and Random Forest have been widely adopted. SVM is known for its strong performance in high-dimensional feature spaces and its ability to handle nonlinear decision boundaries through kernel functions [7]. Random Forest, on the other hand, is an ensemble-based algorithm that combines multiple decision trees to improve classification robustness and reduce overfitting. Previous studies have shown that Random Forest can achieve competitive results in diabetes risk prediction, particularly when feature interactions are present [6].

In addition to these methods, K-Nearest Neighbors (KNN) has been applied in several diabetes prediction studies due to its simplicity and instance-based learning mechanism. KNN classifies data based on similarity measures and can perform effectively when proper feature scaling and preprocessing are applied [4]. Naive Bayes has also been utilized in medical classification tasks because of its probabilistic framework and computational efficiency. However, its strong assumption of feature independence often limits its performance when applied to real-world clinical data, where features are commonly correlated [5].

A recurring challenge identified in previous studies is the issue of class imbalance in medical datasets. In diabetes prediction datasets, non-diabetic instances typically outnumber diabetic cases, which can bias classification models toward the majority class. This condition may lead to high accuracy values while failing to correctly identify individuals with diabetes [4][6]. To address this problem, data balancing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) have been introduced to improve model sensitivity toward minority classes by generating synthetic samples [4].

Another important aspect highlighted in the literature is the selection of appropriate evaluation metrics for medical prediction tasks. Several studies emphasize that accuracy alone is insufficient for evaluating model performance in healthcare applications, particularly when false negatives may result in serious clinical consequences [6][7]. Metrics such as recall, F1-score, and Area Under the Curve (AUC) provide a more comprehensive assessment of a model's ability to detect positive cases and distinguish between classes. Prioritizing these metrics is therefore essential in the development of reliable and clinically relevant diabetes risk prediction models [7].

Based on the reviewed studies, it can be concluded that machine learning algorithms offer significant potential for diabetes risk prediction. However, careful consideration of algorithm selection, data preprocessing, class imbalance handling, and evaluation metrics is required to ensure model effectiveness and applicability in medical screening contexts. These considerations form the foundation for the methodology proposed in this study.

3. Research Methodology

3.1. Research Design and Framework

This study employs an experimental research design using a comparative machine learning approach to predict diabetes risk. The objective of this methodology is to evaluate and compare the performance of several machine learning algorithms in identifying individuals at risk of diabetes mellitus. The research workflow consists of dataset collection, data preprocessing, model training, model evaluation, and system implementation.

The overall research framework begins with the utilization of a publicly available diabetes dataset, followed by data preprocessing to handle missing values, feature scaling, and class imbalance. Subsequently, multiple machine learning algorithms are trained and evaluated using predefined performance metrics. The best-performing model is then implemented into a web-based application to support real-time diabetes risk prediction.

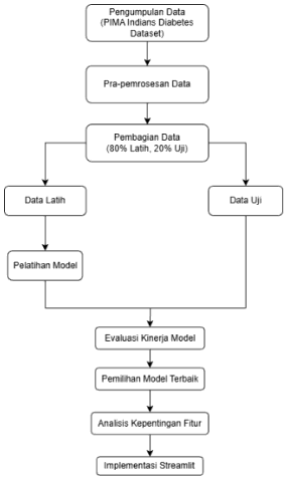

Fig. 1: Research framework

3.2. Dataset Description

The dataset used in this research is the PIMA Indians Diabetes Dataset, which is widely used for benchmarking diabetes prediction models. The dataset consists of medical diagnostic measurements collected from female patients of Pima Indian heritage aged 21 years and above. It contains several clinical attributes related to diabetes risk, with a binary target variable indicating the presence or absence of diabetes.

The dataset includes features such as number of pregnancies, plasma glucose concentration, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. The target variable is labeled as *Outcome*, where a value of 1 indicates diabetes and 0 indicates non-diabetes.

Table 1: Description of dataset attributes

No	Attribute Name	Description	Data Type	Variable Type
1	Pregnancies	Number of times the patient has been pregnant	Float	Discrete
2	Glucose	Plasma glucose concentration measured 2 hours after an oral glucose tolerance test	Float	Continuous
3	BloodPressure	Diastolic blood pressure (mm Hg)	Float	Continuous
4	SkinThickness	Triceps skin fold thickness (mm)	Float	Continuous
5	Insulin	2-hour serum insulin level (ml)	Float	Continuous
6	BMI	Body Mass Index (kg/m²)	Float	Continuous
7	DiabetesPedigreeFunction	Function that assesses the likelihood of diabetes based on family history	Float	Continuous
8	Age	Age of the patient	Float	Continuous
-	Outcome	Class label indicating diabetes status (target variable)	Integer	Categorical (Binary)

3.3. Data Pre-processing

Data preprocessing is a critical step to ensure the quality and reliability of machine learning models. In this study, missing values were identified in several attributes, represented by zero values that are not physiologically plausible. These values were replaced using median-based imputation to minimize the influence of outliers.

Feature scaling was applied to standardize the range of numerical attributes, ensuring that all features contribute equally during model training, particularly for distance-based algorithms such as K-Nearest Neighbors and margin-based models such as Support Vector Machine. Furthermore, the dataset exhibited class imbalance, where non-diabetic instances outnumbered diabetic cases. To address this

issue, the SMOTETomek technique was applied to balance the dataset by combining synthetic oversampling and data cleaning. Finally, the preprocessed dataset was split into training and testing sets to evaluate model performance on unseen data.

3.4. Machine Learning Algorithms

This study evaluates five machine learning algorithms commonly used in medical classification tasks: Logistic Regression, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes.

Logistic Regression is used as a baseline classifier due to its simplicity and interpretability in binary classification problems. Support Vector Machine is employed for its ability to handle high-dimensional data and nonlinear decision boundaries. Random Forest is utilized as an ensemble-based method that improves classification robustness through multiple decision trees. K-Nearest Neighbors classifies instances based on similarity measures, while Naive Bayes applies probabilistic reasoning based on Bayes' theorem with feature independence assumptions.

These algorithms were selected to provide a comprehensive comparison between linear, nonlinear, ensemble, distance-based, and probabilistic classifiers.

3.5. Model Evaluation Metrics

To evaluate model performance, several classification metrics were employed, including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). While accuracy provides a general measure of correctness, it may be misleading in imbalanced datasets. Therefore, recall and F1-score were prioritized to reduce the risk of false negatives, which is critical in medical screening applications.

The evaluation metrics are defined as follows:

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-Score:

$$F1 - Score = 2 \cdot \frac{Precison \cdot Recall}{Precision + Recall} \tag{4}$$

3.6. System Implementation

The best-performing machine learning model was implemented into a web-based application to enable practical usage. The system was developed using Python and the Streamlit framework, allowing users to input clinical parameters and obtain real-time diabetes risk predictions. The application workflow includes data input, preprocessing, model inference, and output visualization.
The user input interface and prediction result interface of the developed application are presented in Fig. 2 and Fig. 3, respectively.

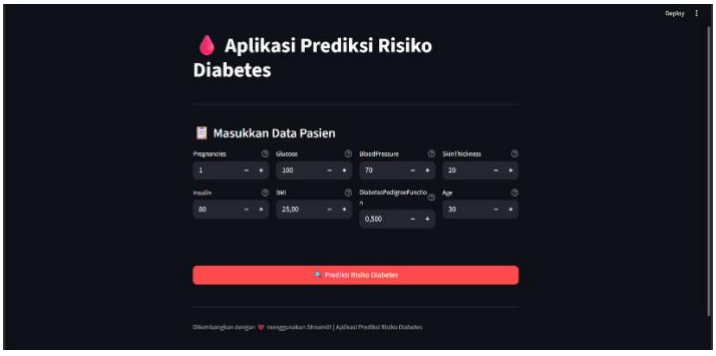

Fig. 2: User Input Interface
This interface allows users to enter the required clinical parameters, such as glucose level, body mass index, and age, before submitting the data for prediction.

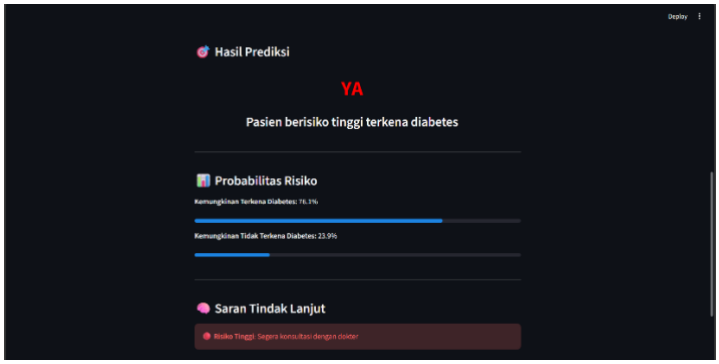

Fig. 3: Prediction Result

The system displays the prediction outcome, including the diabetes risk classification, probability score, and recommended follow-up actions based on the prediction results.

4. Result

4.1. Experimental Setup

This section presents the experimental results obtained from the implementation of five machine learning algorithms for diabetes risk prediction, namely Logistic Regression, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes. The experiments were conducted using the preprocessed PIMA Indians Diabetes Dataset, as described in the previous section.

The dataset was divided into training and testing sets to evaluate model performance on unseen data. Each model was trained using the same dataset configuration to ensure a fair comparison. The evaluation was performed using accuracy, precision, recall, F1-score, and Area Under the Curve (AUC), with particular emphasis on recall and F1-score due to the importance of minimizing false negatives in medical diagnosis tasks.

4.2. Performance Comparison of Machine Learning Models

The performance of each machine learning model was evaluated and compared using standard classification metrics. Table 2 summarizes the evaluation results obtained from the testing dataset. The results indicate variations in performance across different algorithms, highlighting the strengths and limitations of each approach.

Table 2: Performance comparison of machine learning models

Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest (SMT)	75.32	0.62	0.74	0.68	0.87
Support Vector Machine	72.73	0.58	0.80	0.67	0.89
K-Nearest Neighbors	72.08	0.60	0.61	0.61	0.81
Naive Bayes	70.13	0.56	0.74	0.63	0.79
Logistic Regression	69.48	0.55	0.74	0.63	0.82

To further analyze the classification performance of the best-performing model, the confusion matrix of the SVM model is presented in Fig. 4.

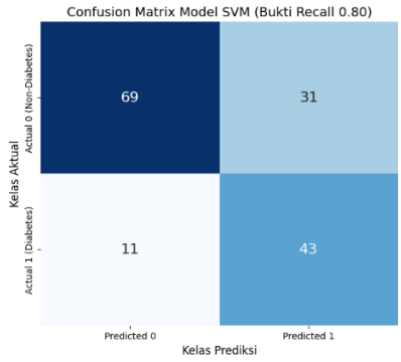

Fig. 4: Confusion matrix of the SVM model for diabetes risk prediction

The confusion matrix illustrates the distribution of true positive, true negative, false positive, and false negative predictions generated by the model.

4.3. ROC Curve and AUC Analysis

To further analyze the discriminative capability of the classification models, Receiver Operating Characteristic (ROC) curves were generated. The ROC curve illustrates the trade-off between the true positive rate and false positive rate at various threshold settings. The Area Under the Curve (AUC) provides a single scalar value representing the overall classification performance.

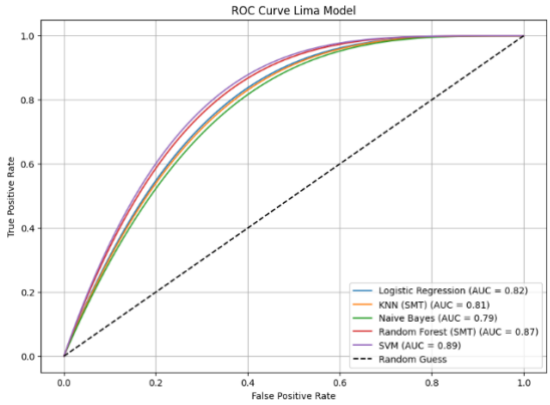

Fig. 5: ROC curves of machine learning models for diabetes risk prediction

The ROC curves show that different models exhibit varying classification performance. The SVM model demonstrates a higher AUC value compared to the other algorithms, indicating superior discriminative capability. This result is consistent with the performance metrics presented in Table 2, where the SVM model achieved higher recall and F1-score values.

4.4. Discussion of Results

This study aimed to compare the performance of several machine learning algorithms for diabetes risk prediction and to implement the best-performing model into a practical web-based application. Based on the experimental results, the SVM model consistently demonstrated superior performance compared to other algorithms, particularly in terms of recall, F1-score, and AUC.

The high recall achieved by the SVM model indicates its effectiveness in identifying individuals at risk of diabetes, which is a critical factor in medical screening applications. Minimizing false negative predictions is essential to reduce the risk of undetected cases that may lead to delayed diagnosis and treatment. The confusion matrix presented in Fig. 4 further confirms the model's ability to correctly classify diabetic and non-diabetic instances.

Random Forest also showed competitive performance, benefiting from its ensemble-based learning approach. However, its performance was slightly lower than that of SVM in terms of recall and AUC. Logistic Regression provided a stable baseline but was outperformed by more complex models. KNN and Naive Bayes demonstrated lower performance, which may be attributed to sensitivity to feature scaling and simplifying probabilistic assumptions, respectively.

Overall, the results indicate that selecting appropriate machine learning algorithms and evaluation metrics plays a crucial role in developing reliable diabetes risk prediction systems. The findings of this study are aligned with previous research that emphasizes the importance of recall, F1-score, and AUC when evaluating classification models in healthcare applications.

5. Conclusion

This study presented a comparative analysis of several machine learning algorithms for diabetes risk prediction using the PIMA Indians Diabetes Dataset and demonstrated the implementation of the best-performing model into a web-based application. Five classification algorithms—Logistic Regression, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors, and Naive Bayes—were evaluated using standard performance metrics.

Based on the experimental results, the SVM model achieved the best overall performance, particularly in terms of recall, F1-score, and Area Under the Curve (AUC), indicating its effectiveness in identifying individuals at risk of diabetes. These metrics are especially important in medical screening applications, where minimizing false negative predictions is critical to support early detection and timely intervention.

The implementation of the trained model into a web-based application provides a practical tool that allows users to input clinical parameters and obtain real-time diabetes risk predictions. This implementation demonstrates the potential of machine learning techniques to support decision-making processes in healthcare settings.

Future work may include the use of larger and more diverse datasets, the application of advanced feature selection or optimization techniques, and further evaluation of the system in real-world clinical environments to enhance model generalizability and reliability.

References

[1] International Diabetes Federation, *Diabetes Atlas*, vol. 11th editi. 2025. [Online]. Available: https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/

[2] World Health Organization, "Diabetes," 2024, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[3] Kementerian Kesehatan Republik Indonesia, "Survei Kesehatan Indonesia (SKI)," 2023. [Online]. Available: https://kemkes.go.id/id/survei-kesehatan-indonesia-ski-2023

[4] S. Sidiq, Alfian, and N. S. Mabrur, "Pengembangan Model Prediksi Risiko Diabetes Menggunakan Pendekatan AdaBoost dan Teknik Oversampling SMOTE," *J. Ilm. Inform. dan Ilmu Komput.*, vol. 4, no. 1, pp. 13–23, 2025.

[5] R. A. Pratama, F. Wabula, H. Ilmandry, L. M. Isabela, M. Raharjo, and R. Sianipar, "Literature Review the Impact of Machine Learning in Modern Industries," *Nian Tana Sikk. J. ilmiah Mahasiswa*, vol. 3, no. 1, pp. 177–182, 2025.

[6] B. Siswoyo and M. I. Nurhafidz, "Penerapan Algoritma Random Forest Untuk Prediksi Risiko Diabetes Berdasarkan Data Kesehatan Pasien," *J. Teknol. Inf. Digit.*, vol. 1, no. 1, pp. 35–38, 2025.

[7] E. Giunchiglia, F. Imrie, M. van der Schaar, and T. Lukasiewicz, "Machine learning with requirements: A manifesto," *Neurosymbolic Artif. Intell.*, vol. 1, pp. 1–12, 2025, doi: 10.3233/nai-240767.

[8] P. R. Sihombing and I. F. Yuliati, "Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 20, no. 2, pp. 417–426, 2021, doi: 10.30812/matrik.v20i2.1174.