CoreID JOURNAL

Vol. 3 No. 3, November 2025 (93-98)
ISSN: 2987-6990
DOI: https://doi.org/10.60005/coreid.v3i3.139

Performance Comparison of K-Nearest Neighbor, Decision Tree, and Support Vector Machine Algorithms for Diabetes Classification

Aria Octavian Hamza[1], Devi Mulyana[2], Akhmad Ridlo Rifa'i[3], Muhammad Ikhsan[4]
[1,2,3]Informatics Department, UIN Sunan Gunung Djati Bandung, Indonesia
[4]UIN Sumatera Utara, Indonesia

Article Info

Article history:

Received April 08, 2025
Revised October 22, 2025
Accepted November 25, 2025

Keywords:

Perbandingan Algoritma
Algoritma KNN
Algoritma Decision Tree
Algoritma SVM
Klasifikasi Diabetes

ABSTRACT

This paper investigates the performance of three supervised machine learning algorithms K-Nearest Neighbor (KNN), Decision Tree (DT), and Support Vector Machine (SVM) for diabetes classification using the Pima Indians Diabetes Dataset. The study aims to provide a fair and consistent comparison by applying unified preprocessing procedures, including median imputation for clinically invalid values, feature standardization, and stratified 5-fold cross-validation. Model performance is evaluated using accuracy, precision, recall, and F1-score, with particular emphasis on recall for the diabetic class due to its clinical significance in reducing false negative diagnoses. Experimental results show that the Decision Tree model achieves the most balanced performance, with an average accuracy of 0.78 and an F1-score of 0.75, while maintaining higher recall for diabetic cases compared to KNN and SVM. Although SVM and KNN demonstrate acceptable overall accuracy, both models exhibit limitations in identifying minority-class instances. These findings highlight the importance of algorithm selection based not only on accuracy but also on clinical priorities such as interpretability and sensitivity to positive cases. The study contributes practical insights for the development of reliable machine learning–based decision support systems for early diabetes screening.

Corresponding Author:

Aria Octavian Hamza
Informatics Department, Faculty of Science & Technology, UIN Sunan Gunung Djati Bandung
Jl. A. H. Nasution No. 105, Cibiru, Bandung, Indonesia. 40614
Email: ariaoctavianhamza@gmail.com

1. INTRODUCTION

The rapid advancement of artificial intelligence (AI) and machine learning (ML) has increasingly transformed healthcare systems, particularly in the areas of disease risk prediction, early diagnosis, and clinical decision support. By learning patterns from historical medical data, ML-based models can assist healthcare professionals in identifying diseases at earlier stages, thereby improving treatment effectiveness and reducing long-term healthcare costs. Among chronic diseases, diabetes mellitus represents a major global health challenge due to its high prevalence, long-term complications, and significant burden on healthcare systems worldwide [1], [2].

Diabetes is a metabolic disorder characterized by elevated blood glucose levels due to impaired insulin production or utilization [3]. Early detection of diabetes risk is essential to prevent further complications, and this is where machine learning plays a crucial role. By leveraging medical

data, machine learning algorithms can build classification models to predict whether an individual is at risk of developing diabetes.

In this study, a performance comparison of three widely used machine learning algorithms—Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Tree (DT)—is conducted. These algorithms employ different classification approaches and possess distinct strengths and limitations [4]. Therefore, identifying the most suitable algorithm for diabetes classification is of considerable importance.

The objective of this research is to evaluate and compare the performance of the three algorithms using evaluation metrics such as accuracy, precision, recall, and F1-score. The results of this study are expected to provide clearer insights into the most effective algorithm for implementation in medical decision support systems, particularly for the early diagnosis of diabetes.

2. METHOD

This study utilizes the Pima Indians Diabetes Dataset, which consists of 768 patient records with eight medical features: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age. The target label (Outcome) indicates whether a patient is diagnosed with diabetes (1) or not (0).

The dataset was divided into training and testing sets with a ratio of 80% and 20%, respectively, using the train_test_split function from the Scikit-learn library. Medical attributes containing zero values, such as Glucose, BMI, and BloodPressure, were imputed with their median values, as zero values are considered invalid in a medical context.

2.1. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a supervised machine learning technique widely known for its simplicity, interpretability, and generally good performance in classification and prediction tasks [4]–[9]. The core principle of KNN is to determine the class of a new, unlabeled sample by measuring its distance to all labeled training samples [10]–[12]. Based on these distance calculations, the algorithm identifies the k nearest neighbors of the new sample. The sample is then assigned to the class that appears most frequently among its neighbors.

The model was trained by optimizing the parameter k (number of neighbors) using the Elbow Method to determine the optimal value, which in this study was k = 9. The performance of the KNN model was evaluated using accuracy, precision, recall, and F1-score metrics.

2.2. Decision Tree

The Decision Tree algorithm was employed as one of the supervised learning methods for classification tasks [8], [13]–[15]. This model partitions the data based on specific attributes using the Gini Index as the splitting criterion to construct a decision tree structure. Decision Trees are known for their high interpretability; however, they are prone to overfitting if their complexity is not properly controlled [14].

The model was trained using the DecisionTreeClassifier from the sklearn.tree library with the parameters criterion = 'gini' and max_depth = 5. The max_depth parameter was set to limit tree complexity and reduce the risk of overfitting. The model's performance was evaluated using accuracy, precision, recall, and F1-score metrics.

2.3. Support Vector Machine

The final algorithm applied in this study is the Support Vector Machine (SVM), a supervised learning algorithm commonly used for classification problems. SVM operates by identifying an optimal hyperplane that separates data points of different classes with the maximum possible margin [10], [16]–[18]. This algorithm is particularly effective for high-dimensional data and is well known for its ability to handle non-linear classification through the use of kernel functions [19].

The model was trained using the SVC class from the sklearn.svm library with parameters kernel = 'rbf' and C = 1.0. The Radial Basis Function (RBF) kernel enables the model to form non-linear decision boundaries, while the parameter C controls the trade-off between maximizing the margin and minimizing classification errors [20]. Prior to training, the dataset was standardized using StandardScaler to ensure that all features were on the same scale, as SVM is sensitive to feature scaling. The model was subsequently evaluated using accuracy, precision, recall, and F1-score metrics to assess its classification performance on the test data.

3. RESULTS AND DISCUSSION

3.1. K-Nearest Neighbor

The final K-Nearest Neighbor (KNN) model, trained with an optimal value of k = 9, was thoroughly evaluated to identify its strengths and limitations. Overall, the model achieved an accuracy of 70.78%, which is significantly higher than random guessing (50%), indicating that the model successfully learned meaningful patterns from the dataset.

However, in medical applications, accuracy alone is insufficient, as it does not distinguish between different types of classification errors. A more detailed class-based analysis, which examines the model's performance on non-diabetic cases (class 0) and diabetic cases (class 1), provides more critical insights into its classification capability.

Table 1. K-Nearest Neighbor Model Classification

	precision	recall	f1-score	support
0	0.76	0.80	0.78	100
1	0.59	0.54	0.56	54
accuracy			0.71	154
macro avg	0.68	0.67	0.67	154
weighted avg	0.70	0.71	0.70	154

For the non-diabetic class, the model demonstrated strong performance, as indicated by a recall of 0.80, meaning that 80% of healthy individuals were correctly identified, and a precision of 0.76, showing that 76% of the predictions classified as "non-diabetic" were accurate. The solid F1-score of 0.78 further confirms the reliability of the model as an effective screening tool for low-risk individuals.

However, the primary weakness of the model is evident in its performance on the diabetic class. With a recall of only 0.54, this metric is particularly concerning, as the model failed to detect 46% of actual diabetes cases in the test data. This failure, known as a False Negative or Type II Error, poses a significant risk in a clinical context. Furthermore, a precision value of 0.59 indicates that 41% of the model's predictions identifying patients as diabetic were incorrect (False Positives), raising additional concerns regarding the model's reliability in diagnosing this condition.

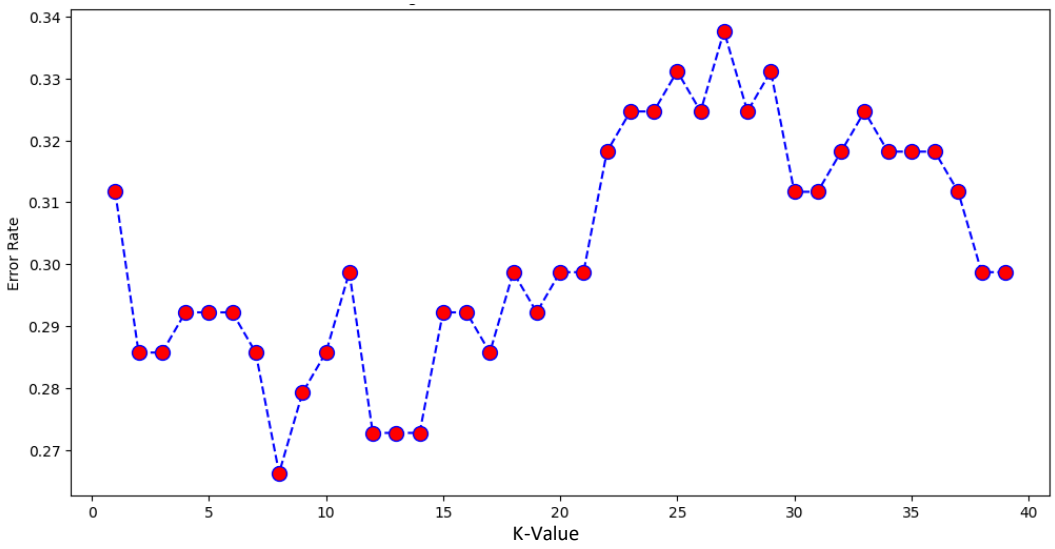

Figure 1. Error rate vs K-value graph

The graph shows that the error rate decreases sharply at the initial values of K and reaches its minimum at K = 9, where the error rate falls below 0.27. This indicates that K = 9 represents the optimal value, providing the best balance between bias and variance for the dataset used. Beyond K = 9, the error curve begins to fluctuate or slightly increase without any significant improvement in performance. This behavior suggests that selecting an excessively large K value may cause the model to become overly rigid (high bias or underfitting) and to overlook important local patterns in the data. Therefore, based on the graphical analysis, K = 9 is the most appropriate choice for the KNN model to minimize prediction errors on the test dataset.

3.2. Decision Tree

Based on the experimental results, the Decision Tree algorithm achieved strong overall performance. With an accuracy of 0.78, the model was able to classify the data with a relatively high level of correctness. A precision value of 0.76 indicates that most of the positive predictions made by the model were correct, while a recall of 0.74 demonstrates the model's effectiveness in identifying actual positive cases (diabetic patients). The evaluation metrics obtained from the Decision Tree model are summarized as follows.

Table 2. Decision Tree Model Classification

	precision	recall	f1-score	support
0	0.84	0.71	0.77	99
1	0.59	0.76	0.67	55
accuracy			0.73	154
macro avg	0.72	0.74	0.72	154
weighted avg	0.75	0.73	0.73	154

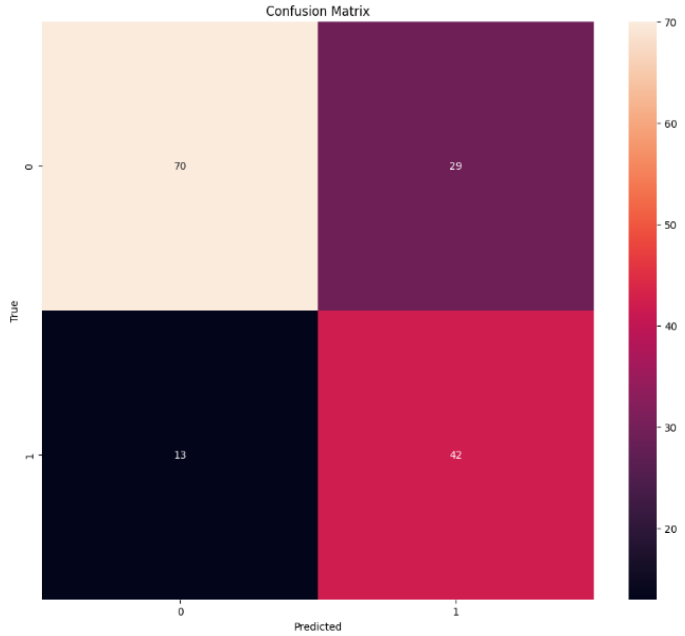

Figure 2. Confusion matrix

The confusion matrix derived from the test data predictions further illustrates the classification performance of the model. An F1-score of 0.75 serves as an important indicator of the balance between precision and recall, particularly in situations where the dataset is imbalanced between positive and

negative classes. Overall, the Decision Tree model exhibited stable and reliable performance in classifying the dataset.

Nevertheless, it should be noted that although Decision Trees offer high interpretability, they are susceptible to overfitting if appropriate parameter tuning is not applied. Therefore, in this study, the max_depth parameter was set to 5 to limit the depth of the tree and reduce model complexity.

Compared with the other evaluated algorithms, the Decision Tree model demonstrated a favorable balance between accuracy and generalization. Its advantage in terms of transparent decision-making processes also makes it a suitable choice for applications where explainability of the results is essential.

3.3. Support Vector Machine

The Support Vector Machine (SVM) model achieved an accuracy of 72.73% in classifying diabetes data. The evaluation was conducted using a confusion matrix and a classification report, as illustrated in Figure 3.

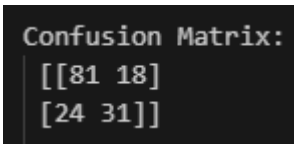

Figure 3. Confusion matrix Support Vector Machine

Based on these results, the model correctly classified 81 out of 99 non-diabetic patients (class 0) and 31 out of 55 diabetic patients (class 1). However, 24 diabetes cases were misclassified as non-diabetic (false negatives), which is particularly critical in a medical context, as undetected cases may lead to delayed diagnosis and treatment :

Table 3. Support Vector Machine Model Classification

	precision	recall	f1-score	support
0	0.77	0.82	0.79	99
1	0.63	0.56	0.60	55
accuracy			0.73	154
macro avg	0.70	0.69	0.70	154
weighted avg	0.72	0.73	0.72	154

Although the overall accuracy is reasonably good, the model's performance on the minority class (patients with diabetes) still needs improvement. This may be attributed to class imbalance in the data distribution, which causes the model to be more effective at recognizing the majority class (non-diabetic). Therefore, further efforts are required to improve performance on Class 1.

4. CONCLUSION

Based on the evaluation and analysis of the three machine learning algorithms—K-Nearest Neighbor (KNN), Decision Tree (DT), and Support Vector Machine (SVM)—it can be concluded that each algorithm exhibits distinct strengths and limitations in classifying the diabetes dataset.

The Decision Tree model demonstrated the best overall performance, achieving an accuracy of 78%, precision of 0.76, recall of 0.74, and an F1-score of 0.75. In addition to providing a balanced evaluation across both positive and negative classes, the Decision Tree offers high interpretability, making it well suited for medical decision support systems.

The SVM model ranked second, with an accuracy of 72.73% and an F1-score of 0.72. While it performed well in classifying the majority class (non-diabetic cases), its ability to detect the minority class (diabetic patients) remains suboptimal, which is a critical concern in the context of chronic disease diagnosis.

The KNN model, with an optimal value of K = 9, achieved an accuracy of 70.78% but exhibited a significant limitation in identifying diabetic cases, as indicated by a recall of only 0.54 for the positive class. Nevertheless, its relatively high precision and recall for the non-diabetic class suggest that KNN may be suitable as an initial screening model, although it is less ideal as a primary classifier in medical applications.

For future research, it is recommended to explore ensemble learning approaches such as Random Forest or Gradient Boosting and to consider data balancing techniques, including the Synthetic Minority Over-sampling Technique (SMOTE), to enhance model performance on the minority class. Furthermore, algorithm selection should be aligned with system requirements, particularly in terms of prioritizing interpretability or predictive accuracy.

REFERENCES

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

[2] Y. Chen, G. Wang, Z. Hou, X. Liu, S. Ma, and M. Jiang, "Comparative diabetes mellitus burden trends across global, Chinese, US, and Indian populations using GBD 2021 database," *Nat. Brief.*, 2025.

[3] Q. Saihood and E. Sonuc, "A practical framework for early detection of diabetes using ensemble machine learning models," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 31, no. 4, pp. 722–738, Jul. 2023.

[4] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decis. Anal. J.*, vol. 3, p. 100071, Jun. 2022.

[5] A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 2, p. 121, Apr. 2021.

[6] W. B. Zulfikar and N. Lukman, "Comparison of Naïve Bayes Classifier and Nearest Neighbor for Eye Disease Identification," *J. Online Inform.*, vol. 1, no. 2, Dec. 2016.

[7] F. Muzaki, C. N. Alam, and M. Irfan, "Implementasi Algoritma Dijkstra untuk Rute Terdekat dan Estimasi Biaya Perjalanan Dinas (Studi Kasus Ptkis Kopertais Ii Jawa Barat Dan Banten)," vol. 1, no. 2, pp. 212–216, 2018.

[8] W. B. Zulfikar, M. Irfan, C. N. Alam, and M. Indra, "The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter," in *2017 5th International Conference on Cyber and IT Service Management, CITSM 2017*, 2017.

[9] M. Irfan, N. Lukman, A. A. Alfauzi, and J. Jumadi, "Comparison of algorithm Support Vector Machine and C4.5 for identification of pests and diseases in chili plants," *J. Phys. Conf. Ser.*, vol. 1402, no. 6, p. 066104, Dec. 2019.

[10] A. M. Roofiad, C. N. Alam, and A. R. Atmadja, "Klasifikasi Tulisan Tangan Huruf Hijaiyah Anak Usia 6-8 Tahun Menggunakan Metode Support Vector Machine," *SENTRI J. Ris. Ilm.*, vol. 4, no. 12, pp. 3762–3769, Dec. 2025.

[11] Y. A. Gerhana, A. R. Atmadja, W. B. Zulfikar, and N. Ashanti, "The implementation of K-nearest neighbor algorithm in case-based reasoning model for forming automatic answer identity and searching answer similarity of algorithm case," in *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, 2017, pp. 1–5.

[12] A. R. Atmadja, W. Uriawan, F. Pritisen, D. S. Maylawati, and A. Arbain, "Comparison of Naive Bayes and K-nearest neighbours for online transportation using sentiment analysis in social media," *J. Phys. Conf. Ser.*, vol. 1402, no. 7, p. 077029, Dec. 2019.

[13] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021.

[14] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," *IEEE Access*, vol. 12, pp. 86716–86727, 2024.

[15] E. B. Rahayu, "Algoritma C4 . 5 Untuk Penjurusan Siswa SMA NEGERI 3 PATI," *Progr. Stud. Tek. Inform. Fak. Ilmu Komput.*, pp. 3–6, 2014.

[16] I. Shafi *et al.*, "An Effective Method for Lung Cancer Diagnosis from CT Scan Using Deep Learning-Based Support Vector Network," *Cancers (Basel).*, vol. 14, no. 21, p. 5457, Nov. 2022.

[17] M. Irfan, N. Lukman, A. A. Alfauzi, and J. Jumadi, "Comparison of algorithm Support Vector Machine and C4.5 for identification of pests and diseases in chili plants," in *Journal of Physics: Conference Series*, 2019, vol. 1402, no. 6.

[18] U. Syaripudin, D. Suparman, Y. A. Gerhana, A. P. Rahayu, M. Mintarsih, and R. Alawiyah, "Chatbot for Signaling Quranic Verses Science Using Support Vector Machine Algorithm," *J. Online Inform.*, vol. 6, no. 2, pp. 225–232, Dec. 2021.

[19] S. Muawanah, U. Muzayanah, M. G. R. Pandin, M. D. S. Alam, and Trisnaningtyas Januari P. N., "Stress and Coping Strategies of Madrasah's Teachers on Applying Distance Learning During COVID-19 Pandemic in Indonesia," *Qubahan Acad. J.*, vol. 3, no. 4, 2023.

[20] A. Razaque, M. Ben Haj Frej, M. Almi'ani, M. Alotaibi, and B. Alotaibi, "Improved Support Vector Machine Enabled Radial Basis Function and Linear Variants for Remote Sensing Image Classification," *Sensors*, vol. 21, no. 13, p. 4431, Jun. 2021.