Diabetes Mellitus Disease Analysis using Support Vector Machines and K-Nearest Neighbor Methods

Ahmad Rizky Nusantara Habibi [1,*], Ilham Sufiyandi [2], Murni [3], A K M Jayed [4], Arman Mohammad Nakib [5], Abdul Syukur [6], Furizal [7]

[1,2] Department of Computer Science, Universitas Qamarul Huda Badaruddin, Central Lombok 83562, Indonesia
[3] Department of Informatics Engineering, Universitas Muhammadiyah Sorong, Sorong 141010, Indonesia
[4] Department of Computer Science and Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China
[5] School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China
[6] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 10617, Taiwan
[7] Department of Research and Development, Peneliti Teknologi Teknik Indonesia, Sleman 55281, Indonesia

ARTICLE INFO

Article history:

Received October 20, 2024
Revised November 5, 2024
Published January 21, 2025

Keywords:

classification; diabetes melitus; knn; machine learning; svm.

ABSTRACT

Diabetes Mellitus (DM) is a chronic disease characterized by high blood sugar levels and can cause various serious complications if not treated properly. This study aims to analyze the effectiveness of Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) methods in classifying diabetes mellitus patient data. The methodology used includes collecting diabetes datasets, preprocessing data, and applying SVM and KNN algorithms to perform classification. The performance of both methods is analyzed using evaluation metrics such as accuracy, precision, recall, and F1-score. The experimental results show that the SVM method provides more optimal performance in classifying diabetes data compared to KNN, with higher accuracy and lower error rate. This finding indicates that SVM is more suitable for early detection of diabetes mellitus in the dataset used in this study.

Corresponding Author:

Ahmad Rizky Nusantara Habibi, Department of Computer Science, Universitas Qamarul Huda Badaruddin, Central Lombok, Indonesia
Email: ahmadhabibi7159@gmail.com

1. INTRODUCTION

Diabetes Mellitus (DM) is a chronic disease characterized by high blood sugar levels and is one of the Non-Communicable Diseases (NCDs) that risks endangering the body's health. Etymologically, in Greek, “diabetes” means flow or outpouring, while “mellitus” means sugar or honey. Thus, literally, diabetes mellitus can be interpreted as a flow of body fluids containing high levels of sugar.

Every year, many patients lose their lives due to diabetes. Predictions in 2045 also say that there will be an increase in diabetes to 629 million people. Type 1 DM, also known as Insulin-Dependent Diabetes Mellitus (IDDM), is caused by damage to pancreatic beta cells (autoimmune reaction). Only pancreatic beta cells can produce insulin which functions to regulate glucose levels in the body [1]. Based on data from the World Health

Organization (WHO), in 2014, adults aged 18 years and over suffered from diabetes. And in 2019, the number of people with diabetes mellitus worldwide reached 463 million, with 4.2 million deaths. According to the World Health Organization (WHO), diabetes symptoms can occur suddenly. Meanwhile, according to the American Diabetes Association (2011) the symptoms of someone developing diabetes are polyuria (frequent urination), polydipsia (frequent thirst), weight loss, polyphagia (frequent hunger), blurred vision, frequent infections, and wounds that are difficult to heal [2]. Previous research focused on the application of the KNN method to various datasets and other studies focused on the calculation of neighbors in the KNN method, so in this study, the focus of the research is to compare various distance measurement metrics contained in the KNN [3]. Diabetes can be caused by various factors such as age, high blood pressure, high blood sugar levels, obesity, family history, insulin levels, and diet. These factors will be used in this study to build a calcific system that can predict diabetes [4]. Khaled research using K-Nearest Neighbor (KNN) imputation and Tri-ensemble voting model achieved high performance with 97.49% accuracy, 98.16% precision, 99.35% recall, and 98.84% F1 score. The study compared its proposed model with seven other machine learning algorithms under two conditions: without and with KNN imputer. The results show the superiority of this model in handling missing data in diabetes diagnosis, promising early detection and improved patient care quality [5].

2. METHODS

In this research method, there are stages carried out in building this diabetes prediction system using machine learning. Machine learning approaches offer advantages such as high accuracy, adaptability to new language nuances or specific domains, and the ability to process complex sentences and broader contexts [6]. From the start of collecting the data used, then processing the data, then the process of mining data using the support vector machine method and forward selection, then implementing using the python programming language, and evaluating testing with confusion matrix [7].

2.1. Data Collecting

The data used in this study is a dataset in CSV format containing 768 rows of data with 9 columns, namely: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome. The "Outcome" column is a classification label that indicates whether a patient has diabetes (1) or not (0). This dataset is obtained from trusted sources and is specifically designed for diabetes prediction analysis. This dataset is obtained from trusted sources and is specifically designed for diabetes prediction analysis. The data used in this study came from Kaggle, with a focus on the diabetes dataset. The data include variables relevant to the classification of diabetes [8].

2.2. Text Preprocessing

Data pre-processing is done to clean the data [9][10][11][12] and performed to improve the quality of the data before it is used in the model training process. At this stage, the data will be explored by preprocessing by checking duplicate data, missing values, and one-hot-encoding on categorical data [13]. The steps applied include:

- Empty Value Handling: Data that had blank values in numeric columns such as Insulin and Skin Thickness were filled with median values.
- Normalization: All numeric features were normalized using Min-Max Scaling to have a range of values between 0 and 1.
- Data Division: The dataset was divided into training data (70%) and test data (30%) randomly to avoid bias in the model evaluation results.

2.3. Implementation of Support Vector Machine Algorithm

SVM is an unsupervised learning algorithm in machine learning [14]. SVM is a very effective algorithm in advanced machine learning especially in prediction studies [15]. Support Vector Machine (SVM) algorithm was invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. In 1963. In 1992, Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik proposed a way to create a nonlinear classifier by applying a kernel trick to maximum-margin hyperplanes [16]. The Support Vector Machines (SVM) algorithm is implemented with the following procedure:

- Kernel Selection: Radial Basis Function (RBF) kernel is selected because it is able to map non-linear data to higher dimensional space effectively.
- Model Training: The SVM model is trained using the preprocessed training data.
- Parameter Optimization: C and gamma parameters are optimized using Grid Search method to get the best performance.
- Performance Evaluation: Evaluation is done by measuring accuracy, precision, recall, and F1-score metrics on the test data.

2.4. Implementation of K-Nearest Neighbor Algorithm

The KNN classification is a statistically based algorithm that is relatively stable and effective for classification tasks [17]. The K-Nearest Neighbor (KNN) algorithm is implemented with the following steps:

- K Value Selection: The optimal value for K is determined using the cross-validation method.
- Model Training: The KNN model is trained with normalized training data to avoid differences in feature scales.
- Performance Evaluation: Model performance is evaluated using accuracy, precision, recall, and F1-score metrics on test data.

2.5. Evaluation

Classification evaluation methods, such as precision, accuracy, recall, and F1 score, have been widely used and proven effective in analyzing confusion matrices [18].

3. RESULTS AND DISCUSSION

This section presents the evaluation results of the Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) models applied to the diabetes mellitus dataset and analyzes the performance of the two algorithms [19].

SVM:
Accuracy: 76%
Confusion Matrix: [[81, 18], [19, 36]]
Precision, Recall, F1-Score:

- Class 0: Precision = 0.81, Recall = 0.82, F1-Score = 0.81
- Class 1: Precision = 0.67, Recall = 0.65, F1-Score = 0.66

KNN (K=5):
Accuracy: 75.3%
Confusion Matrix: [[80, 19], [19, 36]]
Precision, Recall, F1-Score:

- Class 0: Precision = 0.80, Recall = 0.81, F1-Score = 0.80
- Class 1: Precision = 0.65, Recall = 0.65, F1-Score = 0.65

The following table summarizes the performance evaluation results of SVM and KNN:

Table 1. Performance evaluation table

Model	Accuracy	Precision (0/1)	Recall (0/1)	F1-Score (0/1)
SVM	76%	0.81 / 0.67	0.82 / 0.65	0.81 / 0.66
KNN	75.3%	0.80 / 0.65	0.81 / 0.65	0.80 / 0.65

From the results obtained, it can be concluded that the SVM algorithm has a slightly superior performance compared to KNN in classifying diabetes mellitus data on the dataset used.

3.1. KNN

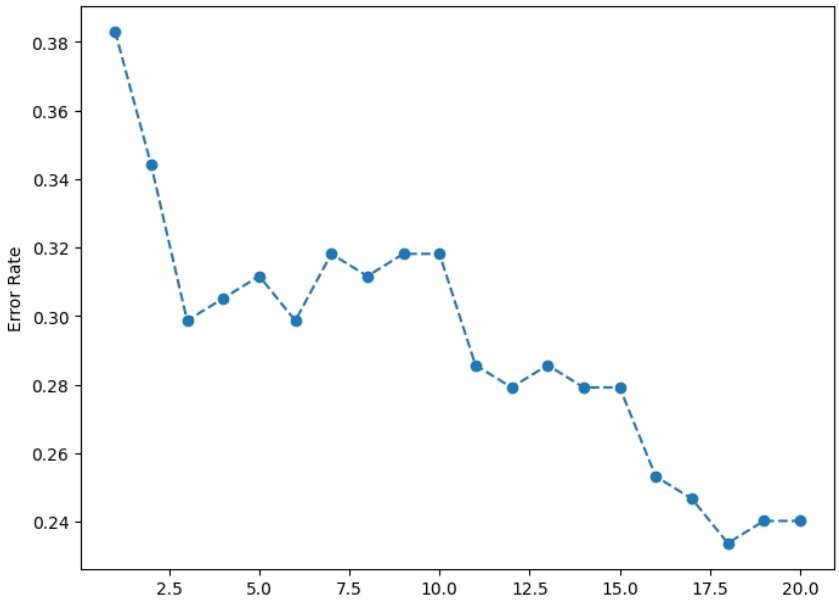

Fig. 1. Method elbow to determine the value of K

The elbow method diagram shown is the result of testing the K-Nearest Neighbor (KNN) algorithm with various values of K. The graph illustrates the relationship between the value of K and the error rate (prediction error rate) in the classification process.

a. Explanation of Graph

X-axis (K value): Shows the number of nearest neighbors used in the KNN model.
Y-axis (Error Rate): Shows the prediction error rate on the test data.

b. Interpretation

At low values of K (e.g. K=1 or K=2), the error rate tends to be high due to overfitting, where the model overfits the training data and thus performs poorly on the test data. As the value of K increases, the error rate starts to decrease and becomes more stable, indicating that the model starts to find a balance between bias and variance. The “elbow” point on the graph indicates the optimal value of K where the error rate starts to stabilize and no longer decreases significantly. In this graph, the optimal point seems to be around K=17 or K=18.

This graph helps in determining the optimal value of *K* by selecting the elbow point, where the error rate achieves a balance between good accuracy and not excessive model complexity. Choosing the optimal *K* value is important to avoid underfitting or overfitting in the KNN algorithm.

3.2. SVM

Confusion matrix is a method to provide information on the results of the classification carried out by the system which is useful for analyzing how well the classifier recognizes tuples from different classes. For example, for a two-class confusion matrix, it will be mentioned as a positive class and a negative class. "True positive refers to a positive class that is appropriately marked by the classifier, while true negative is a negative class that is appropriately marked by the classifier. For false positives are negative classes that are inappropriately marked. Furthermore, false negatives are positive classes that are inappropriately labeled [20].

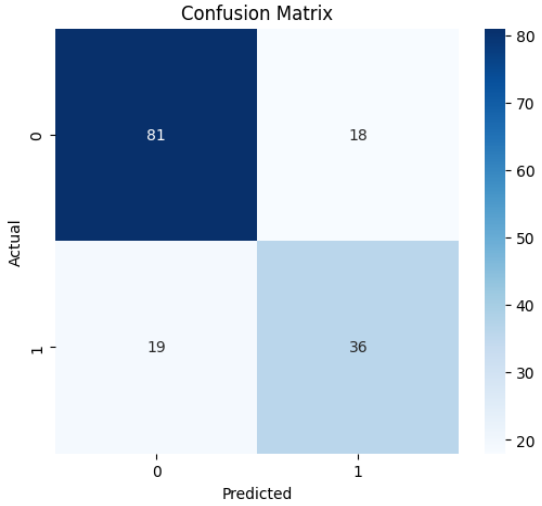

Fig. 2. Confusion matrix in SVM

The confusion matrix on the SVM diagram displayed is used to evaluate the performance of the classification model by comparing the prediction with the actual label on the test data.

a. First Line:
   - 81 (True Negative - TN): A total of 81 samples were correctly classified as non-diabetic (label 0).
   - 18 (False Positive - FP): A total of 18 samples were classified as diabetic (label 1) when they were actually non-diabetic (label 0).

b. Second Row:
   - 19 (False Negative - FN): A total of 19 samples were classified as non-diabetic (label 0) when they were actually positive for diabetes (label 1).
   - 36 (True Positive - TP): A total of 36 samples were correctly classified as diabetic (label 1).

From this confusion matrix, several evaluation metrics can be calculated:
Accuracy: (TP + TN) /(TP + TN + FP + FN) = (81 + 36) / 154 =0.76 or 76%
Precision for positive class: TP / (TP + FP) = 36 / (36 + 18) =0.67
Recall for positive class: TP / (TP + FN) = 36 / (36 + 19) = 0.65
F1-Score for positive class: 2 × ((Precision × Recall) / (Precision + Recall)) = 2 × ((0.67 × 0.6) / (50.67 + 0.65)) = 0.66

The SVM model achieved an accuracy of 76%, demonstrating a balanced performance between precision and recall. This indicates that the model maintains an equilibrium in detecting both positive and negative cases, although a notable error rate persists. The confusion matrix evaluation suggests that while the SVM model is effective, further optimization or parameter adjustments may be necessary to minimize false positives (FP) and false negatives (FN), ultimately enhancing its overall predictive performance.

4. CONCLUSION

Based on the evaluation results that have been carried out, the Support Vector Machines (SVM) algorithm shows superior performance compared to K-Nearest Neighbor (KNN) in classifying diabetes mellitus data on the dataset used. SVM has a higher accuracy rate with a more stable metric evaluation value than KNN. However, the effectiveness of each algorithm can be affected by different dataset characteristics, so the selection of the optimal algorithm must consider the complexity and distribution of the data used.

REFERENCES

[1] Apriyani, H. (2020). Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus. Journal of Information Technology Ampera, 1(3), 133-142.

[2] Oktavia, A., Wijaya, D., Pramuntad,i A., Heksaputra, D. (2024) Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Algoritma K-Nearest Neighbor (K-NN). MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4(3), 812-818. https://doi.org/10.57152/malcom.v4i3.1268
[3] Nasien, D., Darwin, R., Cia, A., Leo Winata, A., Go, J., Charles Wijaya, R., Charles Lo, K. (2024). Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, Dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes. Jurnal Teknik Informatika, 4(1), 11-16. https://doi.org/10.58794/jekin.v4i1.640
[4] Hendro Martono G, Sulistianingsih N. (2024). Perbandingan Matriks jarak pada Algoritma K-NN untuk Prediksi Penyakit Diabetes Comparison of Distance Matrices in the K-NN Algorithm of Predicting Diabetes. JoMI: Journal of Millennial Informatics, 2(1), 1-6.
[5] Patil R, Tamane S, Rawandale S, Patil K. (2022). A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus. International Journal of Electrical and Computer Engineering, 12(1), 524-533. https://doi.org/10.11591/ijece.v12i1.pp524-533
[6] Asno Azzawagama Firdaus et al., "Application of Sentiment Analysis as an Innovative Approach to Policy Making: A review," Journal of Robotics and Control (JRC), vol. 5, no. No. 6, pp. 1784–1798, 2024.
[7] Hovi Sohibul W, Asep Id H, Fajri Rakhmat U. (2022). Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM). Informatics and Digital Expert (INDEX), 4(1), 40–45. https://doi.org/10.36423/index.v4i1.895
[8] Oktaviana, A., Wijaya, D. P., Pramuntadi, A., & Heksaputra, D. (2024). Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Algoritma K-Nearest Neighbor (K-NN). MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4(3), 812–818. https://doi.org/10.57152/malcom.v4i3.1268
[9] A. A. Firdaus, A. Yudhana, and I. Riadi, "Indonesian presidential election sentiment: Dataset of response public before 2024," Data Brief, vol. 52, p. 109993, 2024, doi: 10.1016/j.dib.2023.109993
[10] A. A. Firdaus, A. Yudhana, and I. Riadi, "Public Opinion Analysis of Presidential Candidate Using Naïve Bayes Method," Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, vol. 4, no. 2, pp. 563–570, May 2023, doi: 10.22219/kinetik.v8i2.1686.
[11] A. A. Firdaus, A. Yudhana, and I. Riadi, "DECODE : Jurnal Pendidikan Teknologi Informasi," Decode: Jurnal Pendidikan Teknologi Informasi, vol. 3, no. 2, pp. 236–245, 2023, doi: http://dx.doi.org/10.51454/decode.v3i2.172.
[12] M. M. Dakwah, A. A. Firdaus, Furizal, and R. A. Faresta, "Sentiment Analysis on Marketplace in Indonesia using Support Vector Machine and Naïve Bayes Method," Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI), vol. 10, no. 1, pp. 39–53, 2023, doi: 10.26555/jiteki.v10i1.28070
[13] Yinshan Yu, Mingzhen Shao, Lingjie Jiang, Yongbin Ke, Dandan Wei, Dongyang Zhang,Mingxin Jiang, Yudong Yang. (2021). Quantitative analysis of multiple components based on support vector machine (SVM). Optik. Volume 237, July 2021, 166759. https://doi.org/10.1016/j.ijleo.2021.166759
[14] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," Comput Methods Programs Biomed, vol. 220, p. 106773, Jun. 2022, doi: 10.1016/j.cmpb.2022.106773.
[15] A. A. Firdaus, A. Yudhana, and I. Riadi, "Prediction of Presidential Election Results using Sentiment Analysis with Pre and Post Candidate Registration Data," Khazanah Informatika, Vol. 10, No. 1, 2024, doi: https://doi.org/10.23917/khif.v10i1.4836
[16] Abhilash S, Sukhkirandeep K, Naz Memon, A. Jainul Fathima, Samrat R, Mohammed Wasim B. (2021). Alzheimer's patients detection using support vector machine (SVM) with quantitative analysis. Neuroscience Informatics. Volume 1, Issue 3, November 2021, 100012. https://doi.org/10.1016/j.neuri.2021.100012
[17] Q. Xu, "Application of an Intelligent English Text Classification Model with Improved KNN Algorithm in the Context of Big Data in Libraries," Systems and Soft Computing, p. 200186, Jan. 2025, doi: 10.1016/j.sasc.2025.200186.
[18] A. A. Firdaus, A. Yudhana, and I. Riadi, "Prediction of Indonesian Presidential Election Results using Sentiment Analysis with Naïve Bayes Method," Jurnal Media Informatika Budidarma, vol. 8, no. 1, pp. 41–50, 2024, doi: 10.30865/mib.v8i1.7007.
[19] Saut Dohot S, Yusra Uli R G, Nita S, Salim Butar-Butar H, Riki Marthin S. (2023). Implementation of KNN algorithm in classifying diabetic ulcers in patients with diabetes mellitus. Jurnal Mantik (Manajemen, Teknologi Informatika dan Komunikasi). Vol. 7 No. 2 (2023). https://doi.org/10.35335/mantik.v7i2.3928
[20] Khaled Alnowaiser. (2024). Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model. https://doi.org/10.1109/ACCESS.2024.3359760.