MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (1) 2026, p. 01-09
E-ISSN : XXXX-XXXX
©2026. The Author(s). This is an open acces article under cc-by-sa

Analysis of diabetes disease using k-nearest neighbor (KNN) and naive bayes methods

Noe Uriel Kurnianto [a,1*], Lucky Amadya [a,2]

[a] Informatics Study Program, STMIK Amikom Surakarta, Sukoharjo, Indonesia
[1] noe.10369@mhs.amikomsolo.ac.id; [2]lucky.10379@mhs.amikomsolo.ac.id
* corresponding author

Article Info

Article History:

Received: Jan 12, 2026
Revised: Jan 20, 2026
Accepted: Feb 03, 2026

KeyWords :
Classification, Diabetes Mellitus, K-Nearest Neighbor, Machine Learning, Naïve Bayes

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Abstract

Diabetes mellitus is a chronic disease whose prevalence continues to increase and has the potential to cause serious complications if not detected early. Therefore, a classification method is needed capable of assisting the diagnosis process quickly and accurately. This study aims to analyze and compare the performance of the K-Nearest Neighbor (KNN) and Naïve Bayes algorithms in the classification of diabetes. The dataset used was obtained from Kaggle, totaling 768 data points with 8 attributes: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and the Outcome label. The dataset is divided into 80% training data and 20% testing data. Model evaluation was carried out using accuracy, precision, recall, and F1-score metrics. The results showed that the KNN method produced the best accuracy of 75% at a value of k = 3, while the Naïve Bayes method produced an accuracy of 77.92%. Based on these evaluation results, the Naïve Bayes method has better performance compared to KNN in classifying diabetes in the dataset used. This research is expected to be a reference in the development of clinical decision support systems for the early diagnosis of diabetes.

1. Introduction

Diabetes mellitus is a non-communicable disease that has become a serious global concern, including in Indonesia. This disease occurs due to disturbances in the production or function of insulin, which results in high levels of glucose in the blood. Diabetes is defined as a state of hyperglycemia both in a fasting state and after eating. Chronic hyperglycemia in diabetes mellitus (DM) is associated with end-organ damage, dysfunction, and failure of organs and tissues including the retina, kidneys, nerves, heart, and blood vessels [1]. If not treated optimally, this condition can develop into serious complications, such as coronary heart disease, stroke, chronic kidney failure, retinopathy, and premature death [2]. Therefore, prompt and accurate early detection is the main key to preventing the deterioration of the patient's condition.

Advancements in information technology, particularly in the field of Artificial Intelligence, open up significant opportunities in supporting the medical diagnosis process [3]. One branch of artificial intelligence that is widely applied is machine learning, which is capable of building prediction models based on patients' historical data [4]. This approach allows the system to support clinical decision-making more quickly, efficiently, and evidence-based. In the context of diabetes disease classification, various machine learning algorithms have been developed and evaluated to improve prediction accuracy. Among the various available algorithms, K-Nearest Neighbor (KNN) and Naïve Bayes are two of the most frequently used methods in medical data classification.

KNN works on a distance-based principle, namely determining the class of a data point based on the majority of its nearest neighbors [5]. Although relatively simple and intuitive, this algorithm is quite sensitive to the selection of the parameter *k* value and the scale of the data features. On the other hand, Naïve Bayes is a probabilistic algorithm based on Bayes' Theorem with the assumption of independence

©2026. The Author(s). This is an open acces article under cc-by-sa

among features [6]. This method is known to have high computational efficiency and relatively stable performance, especially on numerical data. Several previous studies have shown that the performance of these two algorithms can vary, depending on the dataset characteristics and the evaluation methods used. Nevertheless, comparative studies that comprehensively evaluate both algorithms using various classification metrics are still limited. This indicates a research gap that needs to be filled to provide a more objective guide in selecting diabetes classification methods.

Various studies have been conducted previously to test the capability of machine learning in predicting diabetes. Prasetya and Sujatmiko (2022) designed an application comparing KNN and Naïve Bayes, which proved the need for further comparison regarding the specifications of medical dataset [7]. A similar study by Arrohman and Fatah (2024) focused specifically on the use of the KNN algorithm on the Pima Indian Diabetes dataset, demonstrating KNN's effectiveness in recognizing clinical data patterns based on race and heredity [8]. On the other hand, Anisa and Jumanto (2022) highlighted the advantages of Naïve Bayes, which is considered highly efficient and does not require long computational time when faced with varied numerical attributes [9].

Another comparative study was also conducted by Delvika et al. (2022), comparing KNN and Naïve Bayes to measure the specific risk of diabetes in pregnant women, where dataset characteristics are crucial in determining final accuracy [10]. Kurniawan (2025) even expanded this comparison by including algorithms such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and KNN, providing insight that KNN has a dynamic nature whose performance is highly dependent on the value of parameter *k* [11]. Irsyad et al. (2025) strengthened this argument in their research comparing KNN with SVM, emphasizing the importance of data preprocessing before measuring feature distances [12].

Development based on intelligent systems was also investigated in the Journal of Computer Science (2026), where KNN classification was directly integrated into a web-based system for real-time detection [13]. Another study in the Journal of Artificial Intelligence and Software Engineering (2025) explicitly compared these two algorithms on a more heterogeneous diabetes dataset [14]. Furthermore, research related to class imbalance in diabetes datasets (2024) highlighted the vulnerability of models if the distribution of positive and negative classes is unbalanced, demanding evaluation through a confusion matrix [15]. This is supported by a study in the SMATIKA Journal (2024), which outlines the clinical risks when algorithms fail to predict precisely [16].

Based on the description above, this study aims to analyze and compare the performance of the K-Nearest Neighbor and Naïve Bayes algorithms in classifying diabetes disease using a dataset sourced from Kaggle. The evaluation is carried out comprehensively using four main metrics: accuracy, precision, recall, and F1-score, so it is hoped that the results of this study can provide a real contribution to the development of a more reliable and effective diabetes prediction system.

2. Research Methodology

This study utilizes a quantitative experimental approach by applying machine learning methods for the classification of diabetes. The methods used are K-Nearest Neighbor (KNN) and Naïve Bayes, whose performances are then compared based on several evaluation metrics.

2.1 Research Dataset

The dataset used in this study was obtained from the Kaggle platform, namely a diabetes dataset consisting of 768 data points with 8 numerical attributes and 1 class label (Outcome). The attributes used include Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age. Clinically, these attributes represent the following:

1. Pregnancies: Number of times pregnant. Relates to the risk of gestational diabetes.
2. Glucose: Plasma glucose concentration over 2 hours in an oral glucose tolerance test.
3. BloodPressure: Diastolic blood pressure (mm Hg).
4. SkinThickness: Triceps skinfold thickness (mm).
5. Insulin: 2-Hour serum insulin (mu U/ml).
6. BMI: Body Mass Index (weight in kg/(height in meters)^2).
7. DiabetesPedigreeFunction: A function that scores the likelihood of diabetes based on family history.

MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (1) 2026, p. 01-09
E-ISSN : XXXX-XXXX
©2026. The Author(s). This is an open acces article under cc-by-sa

8. Age: Age of the patient (years).

The Outcome label has two classes: a value of 0 for non-diabetic patients and a value of 1 for diabetic patients. This dataset was selected because it is frequently used in diabetes classification research and contains attributes relevant to medical analysis.

2.2 Research Stages

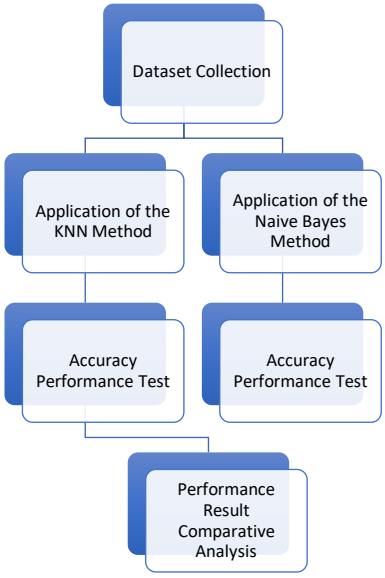

Fig. 1. Research Method Flowchart

In the analysis of diabetes research, there are several stages. The explanation of the research methods for diabetes disease is as follows:

- Dataset Collection: The initial stage in the diabetes analysis research is to prepare the dataset. In this study, the dataset was obtained from kaggle.com, totaling 768 data points. There are 8 classes used in the diabetes analysis, namely Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Outcome.
- Application of K-Nearest Neighbor and Naive Bayes Methods: The next stage is implementing the prepared dataset. For the Outcome value, there are 2 numbers, namely 0 and 1, where a value of 0 means not suffering from diabetes and 1 is stated as suffering from diabetes. The dataset is then processed to determine whether it falls into the diabetes category or not.
- Method Accuracy Performance Testing: After the processing stage of the KNN and Naive Bayes methods, it will enter the process of calculating the accuracy results for diabetes.
- Comparative Analysis of Performance Results: The final stage of the research is the analysis of the accuracy results of the KNN and Naive Bayes methods. It will then be visible which method performs better for use.

2.3 K-Nearest Neighbor (KNN) Method

K-Nearest Neighbor is a distance-based classification algorithm that determines the class of a data point based on the majority class of its *k* nearest neighbors. In this study, the distance between data is calculated using Euclidean Distance. The mathematical formula for calculating the Euclidean Distance between data points *X* and *Y* with *n* attributes is:

$$d(X,Y) = \sqrt{\sum_{i-1}^{n}(x_i - y_i)^2}$$

MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (1) 2026, p. 01-09
E-ISSN : XXXX-XXXX
©2026. The Author(s). This is an open acces article under cc-by-sa

Where d is the proximity distance, x_i is the attribute from the testing data, and y_i is the attribute from the training data. The value of parameter k was tested at several values to obtain the best results. Based on the test results, the value of k = 3 produced the best performance with the highest accuracy value.

2.4 Naïve Bayes Method

Naïve Bayes is a probabilistic classification algorithm based on Bayes' Theorem with the assumption that every attribute is independent of the other attributes. The general formula for Bayes' Theorem is:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

In this study, Naïve Bayes uses the Gaussian Naïve Bayes approach, because all dataset attributes are numerical data. The model calculates the probability of each class based on the normal distribution of each attribute, then determines the class with the highest probability. The Gaussian probability density function is defined as:

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

2.5 Model Evaluation Scenario

To evaluate the performance of the K-Nearest Neighbor (KNN) and Naïve Bayes methods, this study uses a Confusion Matrix. The confusion matrix provides a comprehensive overview of the distribution of the model's prediction results, which includes four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Based on these four components, the evaluation metrics are calculated using the formulas:

1. Accuracy :

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

2. Precision:

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

3. Recall:

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

4. F1-Score:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3. Results and Discussion

3.1 Application of K-Nearest Neighbor and Naive Bayes Methods

In the KNN Method process, 20% of the data was used for testing and 80% of the data for training from the total diabetes dataset; the number of neighbors (k) used was 5 data points. Meanwhile, in the Naïve Bayes Method process, numpy classification was used and performed manually based on normal

MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (1) 2026, p. 01-09
E-ISSN : XXXX-XXXX
©2026. The Author(s). This is an open acces article under cc-by-sa

distribution, using 5 testing datasets. The results of applying the KNN and Naïve Bayes Methods are in the following tables:

Table 1. KNN Data

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Table 2. Naïve Bayes Data

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	Predicted_Class
575	1	119	44	47	63	35.5	0.280	25	0	0
206	8	196	76	29	280	37.5	0.605	57	1	1
127	1	118	58	36	94	33.3	0.261	23	0	0
550	1	116	70	28	0	27.4	0.204	21	0	0
118	4	97	60	23	0	28.2	0.443	22	0	0

Table 1 illustrates a representative subset of five data points processed during the K-Nearest Neighbor (KNN) algorithm's iteration phase. This table highlights eight crucial clinical attributes, including Glucose levels, Blood Pressure, BMI, and Age, which are utilized by the algorithm to calculate the Euclidean distance between data points. In this specific nearest-neighbor sample, the target variable is represented by the 'Outcome' column, where a value of '1' indicates a positive diagnosis for diabetes and '0' indicates a negative diagnosis. As stated in the initial evaluation results, the specific samples designated as K1, K3, and K5 (which correspond to the rows displaying an Outcome of '1') are correctly identified by the model as patients suffering from diabetes, demonstrating how the algorithm isolates positive cases based on feature proximity.

Meanwhile, Table 2 presents the classification results generated by the Gaussian Naïve Bayes model, utilizing a randomly selected testing subset (as indicated by the unique index numbers like 575, 206, and 127). Unlike the KNN table, this dataset explicitly features a 'Predicted_Class' column alongside the actual 'Outcome' column, allowing for a direct and transparent comparison between the real historical data and the model's probabilistic predictions. For instance, in row index 206, the model successfully identifies the patient as a diabetes sufferer, matching the actual Outcome '1' with a Predicted_Class '1'. Furthermore, it accurately classifies the remaining four samples as non-diabetic (Outcome '0' and Predicted_Class '0'). This precise alignment in the presented sample underscores the Naïve Bayes model's effectiveness in calculating class probabilities based on the normal distribution of the clinical features.

Both tables serve as empirical evidence of the rigorous evaluation and iteration stages implemented in this study to secure the most optimal datasets for model testing. By displaying these specific subsets, the methodology transparently shows how raw patient data—ranging from the number of pregnancies to the Diabetes Pedigree Function—is computationally transformed into binary medical classifications. The juxtaposition of these two tables not only clarifies the internal mechanics of both the distance-based KNN and the probabilistic Naïve Bayes algorithms, but it also provides a concrete visual representation of the data processing steps that occur right before the final confusion matrix and overall accuracy metrics are computed.

MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (1) 2026, p. 01-09
E-ISSN : XXXX-XXXX
©2026. The Author(s). This is an open acces article under cc-by-sa

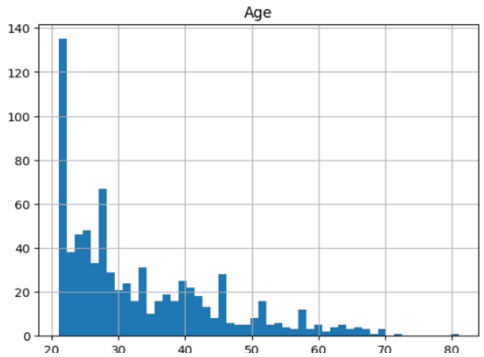

Fig. 2. Age Class of KNN Model

Image 2 presents a histogram illustrating the age distribution within the dataset utilized for training and evaluating the K-Nearest Neighbor (KNN) model. As visually depicted in the graph, the distribution is prominently right-skewed, indicating a significant concentration of individuals in the younger age brackets, particularly between 20 and 30 years old, where the highest frequency peaks sharply. Conversely, the frequency of recorded patient cases steadily declines as the age progresses towards 80 years. This specific demographic concentration is a crucial characteristic of the underlying data, as the distance-based mechanics of the KNN algorithm will inherently rely on this continuous age distribution when calculating the Euclidean proximity between clinical data points to classify potential diabetes outcomes.

3.2 Klasifikasi Model Evaluation Using Confusion Matrix and Classification Metrics

The evaluation of the classification model's performance was carried out using a confusion matrix as the primary tool to measure the ability of each algorithm to classify diabetes data accurately. The confusion matrix provides a comprehensive overview of the distribution of model predictions, including four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These four components form the basis for calculating the evaluation metrics used in this study, which include accuracy, precision, recall, and F1-score.

The use of these four metrics is considered important because the accuracy value alone is not sufficient to represent the quality of the model comprehensively, especially in medical datasets that potentially have an unbalanced class distribution (class imbalance). In the context of diabetes classification, an error in predicting a patient who is actually positive for diabetes as negative (False Negative) has more serious clinical consequences than other types of errors. Therefore, the recall value becomes a highly critical metric, as it reflects the model's ability to detect all positive cases correctly.

Meanwhile, precision measures the level of accuracy of the positive predictions generated by the model, while the F1-score is the harmonic mean between precision and recall, providing a balanced assessment of overall model performance. The evaluation results from both algorithms, namely K-Nearest Neighbor (KNN) and Naïve Bayes, are presented based on these four metrics to obtain an objective and comprehensive comparison. This analysis is expected to provide a more representative picture of the strengths and weaknesses of each method in the context of predicting diabetes.

3.3 KNN Method Evaluation Results

In the KNN method, several k values were tested. Based on the test results, the highest accuracy value was obtained at k = 3 with an accuracy of 75%. The accuracy value for each parameter k test is as follows:

Table 3. KNN Evaluation Results Table

Value of k	Accuracy
1	68,8%
2	70,7%
3	75,3%
4	75,3%

MIoSA : Machine Intelligence for Societal Advancement
Volume 1, (1) 2026, p. 01-09
E-ISSN : XXXX-XXXX
©2026. The Author(s). This is an open acces article under cc-by-sa

These results indicate that increasing the value of *k* can improve model performance up to a certain point. However, using a *k* value that is too large can lead to a decrease in accuracy because the model becomes too generalized in determining classes. Furthermore, based on the training and testing results, the KNN method obtained a train score of 81.7% and a test score of 72.7%, indicating a performance difference between the training and testing data. This indicates that the KNN model has a slight tendency for overfitting, which is good performance on training data but decreases on testing data.

3.4 Naïve Bayes Method Evaluation Results

The Naïve Bayes method was tested using a probabilistic approach with a Gaussian distribution. The evaluation results of the Naïve Bayes method produced an accuracy of 77.92%. In addition to accuracy, this method also produced precision, recall, and F1-score values based on the obtained confusion matrix. The use of the confusion matrix in the Naïve Bayes method helps explain the model's performance in detecting both diabetic and non-diabetic classes. In general, Naïve Bayes showed better performance compared to KNN on this dataset, because this method is able to perform classification more stably even though the dataset has varied numerical attributes.

3.5 Performance Comparison of KNN and Naïve Bayes Methods

Perbandingan hasil performa antara metode KNN dan Naïve Bayes ditunjukkan pada Tabel berikut:

Table 4. Accuracy Comparison of KNN and Naïve Bayes

Method	Accuracy
KNN (k=3)	75%
Naïve Bayes	77,92%

Based on the table, the Naïve Bayes method has a higher accuracy value compared to the KNN method. This difference in performance can be caused by the characteristics of the respective algorithms. KNN is a distance-based method, making it highly sensitive to data scaling, outliers, and the selection of parameter *k*. If the data is not normalized or has large variations between attributes, the distance results can be biased and affect the classification outcome. Conversely, Naïve Bayes uses a probabilistic approach that calculates class probabilities based on data distribution, making it generally more stable. This method is also faster in computation because it does not need to calculate distances against all training data as in KNN.

Thus, the Naïve Bayes method can be considered more effective in classifying diabetes on the Kaggle dataset used in this study. Although Naïve Bayes produces higher accuracy, this method has a weakness, namely the assumption of independence among features. In medical data, several attributes such as glucose levels, BMI, and age tend to have correlations. Nevertheless, the study results show that this assumption is still quite effective in providing good predictions. This indicates that Naïve Bayes is suitable for use as a baseline method in diabetes classification.

3.6 Analysis of Feature Impact and Data Scale

A critical factor contributing to the performance gap between K-Nearest Neighbor and Naïve Bayes in this study is the nature of the dataset's features. The Kaggle diabetes dataset contains attributes with vastly different numerical scales. For instance, the 'DiabetesPedigreeFunction' attribute generally has values less than 2.0, while the 'Insulin' attribute can reach values over 800. Because KNN calculates the Euclidean distance uniformly across all features, attributes with larger scales (like Insulin or Glucose) mathematically dominate the distance calculation, potentially overshadowing the importance of smaller-scaled features unless rigorous feature scaling (normalization or standardization) is applied.

In contrast, the Gaussian Naïve Bayes algorithm handles these varying scales inherently well. By calculating the mean and variance for each feature independently within each class, the algorithm normalizes the influence of each feature probabilistically. This characteristic explains why Naïve Bayes

©2026. The Author(s). This is an open acces article under cc-by-sa

maintains a more robust and slightly higher accuracy (77.92%) without requiring complex hyperparameter tuning or strict data transformation processes, making it highly adaptable to raw medical data.

3.7 Computational Complexity and Efficiency Analysis

Beyond accuracy, evaluating the computational efficiency of both algorithms is essential for implementing a real-time clinical decision support system. KNN is categorized as a "lazy learner." It does not build an explicit generalized model during the training phase; instead, it stores the entire training dataset. Consequently, the computational complexity during the testing phase is $O(n \times d)$, where n is the number of training samples and d is the number of dimensions (features). When evaluating a new patient's data, the system must compute the distance to all 768 data points, which consumes higher memory and processing time. Conversely, Naïve Bayes is an "eager learner." It processes the training data to calculate the prior probabilities and probability density functions, requiring a computational complexity of $O(n \times d)$ during training. However, during the testing phase, the complexity drops significantly to $O(c \times d)$, where c is the number of classes. Because the testing phase only involves looking up and multiplying the pre-calculated probabilities, Naïve Bayes is computationally lighter and faster. This makes Naïve Bayes more scalable and suitable for web-based or mobile-based healthcare applications with limited processing power.

3.8 Clinical Implications of the Classification Models

From a clinical perspective, the application of these machine learning models serves as a primary screening tool rather than an absolute diagnostic mechanism. An accuracy of 77.92% achieved by Naïve Bayes demonstrates a strong baseline for identifying potential diabetic patients based on non-invasive or routinely collected data (such as BMI, age, and blood pressure).

In medical diagnostics, the cost of a False Negative (failing to identify a diabetic patient) is significantly higher than a False Positive (diagnosing a healthy patient with diabetes), as the former delays critical interventions. While both models demonstrate reliable accuracy, integrating this predictive model into a hospital information system can help medical practitioners prioritize high-risk patients for further laboratory testing, such as HbA1c tests. The probabilistic nature of Naïve Bayes also provides an advantage here, as it can output the "probability percentage" of a patient having diabetes, giving doctors a confidence score rather than just a binary "Yes/No" classification.

4. Conclusion

This study has conducted a comparative analysis of two machine learning algorithms, namely K-Nearest Neighbor (KNN) and Naïve Bayes, in classifying diabetes disease using a dataset sourced from Kaggle. Based on the testing and evaluation conducted using accuracy, precision, recall, and F1-score metrics, several main findings were obtained as follows. The KNN algorithm produced the highest accuracy value of 75% at a parameter configuration of k = 3, while the Naïve Bayes algorithm achieved an accuracy of 77.92%. Overall, the evaluation results indicate that Naïve Bayes has superior classification performance compared to KNN on the dataset used in this study.

This superiority is reflected not only in the higher accuracy value but also in the consistency of performance across other evaluation metrics, indicating the model's ability to handle data distribution more effectively. The probabilistic approach in Naïve Bayes proved to be more robust to variances in the scale of health attribute values (such as glucose levels and blood pressure) compared to the metric distance calculations in KNN. Therefore, the Naïve Bayes algorithm can be recommended as a more effective classification method for the purpose of predicting diabetes, especially on dataset characteristics similar to those used in this study.

Nevertheless, it should be underlined that these results are contextual and cannot be generalized absolutely, considering that algorithm performance can vary depending on data characteristics, preprocessing techniques, and the validation methods applied. For future research, it is recommended to explore other more complex classification algorithms, such as Random Forest, Support Vector Machine (SVM), or Gradient Boosting, as well as the application of class balancing techniques like SMOTE to improve prediction performance significantly. In addition, testing using larger and more diverse datasets also needs to be carried out to obtain more representative results that can be generalized more broadly.

©2026. The Author(s). This is an open acces article under cc-by-sa

References

[1] U. Alam, "General aspects of diabetes mellitus," vol. 126, 2014. https://doi.org/10.5005/jp/books/12220_38

[2] F. K. R. Noor, "ASUHAN KEPERAWATAN PASIEN YANG MENGALAMI DIABETES MELITUS TIPE 2 DENGAN KETIDAKSTABILAN KADAR GLUKOSA DARAH DI RSUD PASAR REBO," 2024.

[3] P. Kecerdasan Buatan Dan Dampaknya Pada Dunia Teknologi, I. Zaenuddin, and A. Bani Riyan, "28 Creative Commons Attribution 4.0 International License," 2024.

[4] D. Kusuma Ningrum and A. Maytsa Ismawardi, "EFEKTIVITAS ALGORITMA KECERDASAN BUATAN DALAM IMPLEMENTASI KESEHATAN MENTAL : SYSTEMATIC LITERATURE REVIEW," 2025. https://doi.org/10.36040/jati.v9i1.12457

[5] F. Malik Namus Akbar, "Metode KNN (K-Nearest Neighbor) untuk Menentukan Kualitas Air," vol. 18, no. 1.

[6] A. Sirojul Munir *et al.*, "Perbandingan Akurasi Algoritma Naive Bayes dan Algoritma Decision Tree dalam Pengklasifikasian Penyakit Kanker Payudara".

[7] W. Dwi Prasetya and B. Sujatmiko, "Rancang Bangun Aplikasi dengan Perbandingan Metode K-Nearest Neighbor (KNN) dan Naive Bayes dalam Klasifikasi Penderita Penyakit Diabetes".

[8] J. Kecerdasan Buatan, K. dan Teknologi Informasi, L. Octa Sofyan Firmandala, Z. Fatah, R. Artikel, and K. Kunci Data Mining, "Implementasi Data Mining Klasifikasi Kelulusan Mahasiswa di Perguruan Tinggi Menggunakan K-Nearest Neighbors," *Tahun*, vol. 5, no. 2, 2024, [Online]. Available: https://ejournal.unuja.ac.id/index.php/core. https://doi.org/10.33650/coreai.v5i2.9729

[9] D. Nurul Anisa, "KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN ALGORITMA NAIVE BAYES," *Dinamika Informatika*, vol. 14, no. 1, 2022. https://doi.org/10.35315/informatika.v14i1.9135

[10] J. Homepage *et al.*, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Comparison of Classification Between Naive Bayes and K-Nearest Neighbor on Diabetes Risk in Pregnant Women Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes Pada Ibu Hamil," vol. 2, pp. 68–75, 2022. https://doi.org/10.57152/malcom.v2i2.432

[11] M. Fadli Kurniawan and D. Ayu Megawaty, "Comparison of Logistic Regression, Random Forest, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) Algorithms in Diabetes Prediction," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC. https://doi.org/10.30871/jaic.v9i5.9815

[12] H. Hatta Irsyad, M. I. Syafwan, and D. Ramadhani, "Journal of System & Technology ANALISIS PERBANDINGAN KINERJA ALGORITMA K-NEAREST NEIGHBORS DAN SUPPORT VECTOR MACHINE UNTUK KLASIFIKASI PENYAKIT DIABETES," vol. 1, no. 2, 2025, [Online]. Available: https://doi.org/XX.XXXXX/systec.X.X.X-XX

[13] Muhammad Randy Fachrezi, Hafiz Aryanda, Alwi Syahputra, and Risma Riansyah, "Sistem Klasifikasi Diabetes Mellitus Menggunakan Algoritma K-Nearest Neighbor (KNN) Berbasis Web," *Jurnal Ilmu Komputer dan Teknik Informatika*, vol. 2, no. 1, pp. 119–130, Jan. 2026, doi: https://doi.org/10.64803/juikti.v2i1.116

[14] E. B. Susanto, A. N. Anzila, and B. Ismanto, "Comparison Of The Effectiveness Of K-Nearest Neighbor (KNN) And Naive Bayes Algorithms In Identifying Diabetes Patients," *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 5, no. 1, p. 22, Mar. 2025, doi: https://doi.org/10.30811/jaise.v5i1.6275

[15] M. R. Hunafa and A. Hermawan, "KLIK: Kajian Ilmiah Informatika dan Komputer Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Imbalace Class Dataset Penyakit Diabetes," *Media Online*, vol. 4, no. 3, pp. 1551–1561, 2023, doi: 10.30865/klik.v4i3.1486.

[16] R. A. Safitri and R. Hidayati, "Komparasi Metode K-Nearest Neighbor dan NaÃ¯ve Bayes untuk Mengklasifikasi Resiko Diabetes Di Posbindu Desa Bulupitu," *SMATIKA JURNAL*, vol. 14, no. 02, pp. 297–303, Dec. 2024, doi: https://doi.org/10.32664/smatika.v14i02.1350