ENSEMBLE
Journal of Applied Mathematics, Statistics, and Data Science
eISSN: xxxx-xxxx
Vol. X, No. X, XXXXXX
Page: 50-56

Classification of Diabetes Disease Using the Naive Bayes Algorithm

Nurul Mutmainnah[*1], Khaerunnisa[2], Nurhana[3], Difha[4], Annisa Risky Aulia[5], Yuwanda Purnamasari Pasrun[6*], Mardiawati[7]

[1,2,3,4,5,6,7] Department of Information System, Universitas Sembilanbelas November, Kolaka 93517, Indonesia

E-mail Correspondence Author: *yuwandapurnamasari@gmail.com

: xxxxxx

ARTICLE INFO

Article history

Received : Feb 18, 2026
Revised : Apr 20, 2026
Accepted : May 3, 2026

Keywords
Classification
Data Mining
Diabetes
Naïve Bayes
RapidMiner

Abstract

Diabetes is one of the leading causes of death worldwide, thus requiring an accurate early detection method to minimize the risk of complications. The urgency of this study is based on the need for a computational approach capable of identifying diabetes risk more quickly and objectively. The Naïve Bayes method was selected due to its simple yet effective classification capability in processing both numerical and categorical data. The dataset used was obtained from the Kaggle platform, consisting of 995 records, and was processed through several stages, including preprocessing, splitting the data into 80% training and 20% testing, modeling using the Naïve Bayes algorithm, and evaluation through RapidMiner and manual calculations. The results show that the model achieved an accuracy of 89.95% and an AUC value of 0.968, indicating excellent predictive ability in distinguishing diabetic and non-diabetic cases. Although this study is limited to only two attributes, the findings demonstrate that Naïve Bayes can serve as the basis for a fast and efficient early diabetes detection system. Overall, this research highlights the importance of applying data mining techniques to support the early diagnosis of diabetes.

1. INTRODUCTION

Diabetes is a chronic disease caused by insufficient production of insulin, a hormone produced by the pancreas to regulate glucose levels, in the human body. Glucose is the main energy source for body cells. Insufficient insulin secretion causes blood glucose levels to rise above normal limits [1]. High blood glucose levels can damage body organs and lead to organ and tissue failure [2]. As the number of diabetes cases continues to rise worldwide, data from the World Health Organization (WHO) shows that diabetes ranks seventh among the leading causes of death globally, accounting for approximately 1.5 million deaths each year [3].

One approach that can help predict diabetes is data mining using classification principles. Classification is an algorithmic process for pattern discovery that builds a classification model based on predictor and target class variables [4]. Applying classification to diabetes diagnosis offers several significant advantages. Through this approach, the identification of diabetes risk can be carried out more quickly and objectively, since decisions are based on data patterns rather than subjective judgment alone.

Within classification, one of the most commonly used methods for disease diagnosis is the naïve Bayes method. Naïve Bayes is a classification method based on probability and

statistics, and it is capable of processing multiple variables simultaneously while uncovering relational patterns that may be difficult to detect through manual analysis [5]. A study conducted by Muhammad found that naïve Bayes offers several advantages, including high accuracy achieved through a simple yet fast algorithm [6].

A previous study [7] reported that diabetes classification using the Naïve Bayes method achieved an accuracy of 90.20%. In addition, another study comparing the KNN and Naïve Bayes methods showed that Naïve Bayes achieved a higher accuracy of 95% in predicting diabetes, compared to only 93% for KNN [8]. Other findings indicate that a Naïve Bayes-based model can classify diabetes risk with an accuracy of 87.88% and an error rate of 12.12%, demonstrating that the method is sufficiently reliable for early diabetes detection, particularly when applied to large sample datasets [9].

Based on these findings, the researchers were motivated to apply the naïve Bayes method to detect diabetes, as the algorithm has proven effective for classification tasks. Previous studies suggest that naïve Bayes performs well and is reliable for diabetes prediction. The purpose of this study is to evaluate how effectively the naïve Bayes method predicts diabetes and to improve prediction accuracy by refining the model through feature selection and probability estimation techniques.

2. METHOD

2.1. Problem Identification

Diabetes is a fairly dangerous disease that can lead to severe complications and even death. It needs to be identified early so that it can be treated as soon as possible. Therefore, an accurate classification approach is required to recognize diabetes. The Naïve Bayes algorithm is the classification algorithm used in this study. Naïve Bayes builds models quickly, has strong predictive capability, and offers a practical way to explore and understand data.

2.2. Data Collection

Data collection was carried out using an open dataset available online through the Kaggle platform. The dataset used is a diabetes dataset uploaded by Himansu Nakrani, consisting of 995 records with attributes suitable for the classification process. The attributes used in this study include glucose, blood pressure, and the diabetes class label, which consists of two categories: 0 (No) and 1 (Yes). This dataset was selected because it fits the requirements for implementing the Naïve Bayes method.

Table 1. Research Dataset

No	Glucose	Bloodpressure	Diabetes
1	40	85	0
2	45	92	0
3	45	63	1
4	40	80	0
5	45	73	1
6	40	82	0
7	30	85	0
8	65	63	1
9	45	65	1
10	35	82	0
11	45	73	1
12	50	90	0
13	40	68	1
14	35	93	0
15	50	80	1
16	40	70	1
17	40	73	1
18	40	67	1
19	40	75	1
1	40	85	0

2.3. Data Preprocessing

Data preprocessing was carried out to ensure that the dataset was clean and ready for building the classification model. The first step involved checking for missing or invalid values, which were then corrected or removed so as not to affect the probability calculation process. Data transformation was then performed by converting the class label from numeric form (0 and 1) into categorical form ("No" and "Yes") to make the results easier to interprets.

2.4. Modeling

This stage describes the implementation of the selected algorithm, naïve Bayes, in the process of classifying diabetes data. Manual calculations were performed using Microsoft Excel, while model testing was carried out using RapidMiner software. For testing purposes, sample data were randomly drawn from the available dataset. The probability of each class was calculated to determine the final prediction, with the class having the highest probability assigned as the classification result**.**

2.5 Evaluation

The evaluation stage was carried out to measure the performance of the naïve Bayes model in classifying diabetes data. Evaluation was performed using test data to assess the accuracy of the model's predictions. In addition, the manual calculation results were compared with the RapidMiner testing results to ensure consistency and correctness of the calculation process.

3. RESULTS AND DISCUSSION

3.1. Data Preprocessing

At this stage, the initial dataset was transformed to convert several numeric attribute values into nominal form. This transformation was applied to the diabetes attribute, which serves as the class label. The label values, originally numeric (0 and 1) as shown in the initial dataset in Table 1, were converted into the categorical form "No" and "Yes." This transformation was carried out to simplify the interpretation of the classification results. Meanwhile, the other attributes were kept in their original form since no changes were required. Table 2 presents the results of this label transformation.

Table 2. Preprocessed Dataset

No	Glucose	Bloodpressure	Diabetes
1	40	85	No
2	45	92	No
3	45	63	Yes
4	40	80	No
5	45	73	Yes
6	40	82	No
7	30	85	No
8	65	63	Yes
9	45	65	Yes
10	35	82	No
11	45	73	Yes
12	50	90	No
13	40	68	Yes
14	35	93	No
15	50	80	Yes
16	40	70	Yes
17	40	73	Yes
18	40	67	Yes
19	40	75	Yes

3.2. Modeling

This study applies classification using the selected algorithm, naïve Bayes, to determine whether a sample is classified as having diabetes or not, based on the class probabilities, the probability of each class-wise event, and the resulting accuracy. Before modeling, the initial dataset was divided into training data and testing data using the Split Train Test operator, which splits the data into training and testing sets. The data split ratio used was 80% for training data and 20% for testing data. An 80:20 split ratio can provide the best model performance [10]. Accordingly, from the dataset of 995 records, this study used 796 training records and 199 testing records. The modeling steps using the naïve Bayes algorithm are as follows:

a. Calculating Class Probability

Of the data used, there are two classes: Yes and No. The Yes class (diabetic patients) contains 100 records, while the No class (non-diabetic patients) contains 99 records. The probability of each class was calculated by dividing the number of records in each class by the total number of records. Table 3 presents the resulting class probabilities:

Table 3. Class Probability Results

Manual	Class Probability
Yes	0.495
No	0.505

b. Calculating Attribute Probability Based on Class

Once the class probabilities were determined, the next step was to calculate the probability of each attribute for each class. This was done by counting the frequency of each attribute value within a class and dividing it by the total number of records in that class. This calculation determines how likely a given attribute value is to appear in the Yes or No class, which is then used in the final naïve Bayes calculation. These probability values were obtained from processing the dataset generated during the naïve Bayes model training. Table 4 presents the classification results for each class.

Table 4. Attribute Probability Results by Class

Glucose	Bloodpressure	P(X|H) – Yes	P(X|H) – No
25	73	0.169	0.831
30	73	1.000	0.000
35	83	0.169	0.831
40	92	0.994	0.006
45	92	0.996	0.004
50	83	0.037	0.963
55	57	1.000	0.000
60	68	1.000	0.000

c. Calculating the Probability of Each Class-wise Event

The probability values obtained can be used to predict or determine the label of the testing data using the naïve Bayes formula, shown in Equation (1):

$$P(x_i|y) = \left[1/\sqrt{(2\pi\sigma^2 y)}\right] \cdot exp(-(x_i - \mu y)^2/2\sigma^2 y) \quad (1)$$

Table 5. Example of Testing Data

Glucose	Bloodpressure	Diabetes
40	92	?

Using the example of testing data shown in Table 5, a manual calculation was performed using Microsoft Excel. The results of this calculation are presented in Table 6 below.

Table 6. Manual Calculation Results

Manual Calculation	Value	Result
P(Glucose = 40 | Result = Yes)		0.02558994
P(Bloodpressure = 92 | Result = Yes)		2.789E-08
P(Glucose = 40 | Result = No)		0.0800
P(Bloodpressure = 92 | Result = No)		2.2619E-05
P(Yes, Data) = P(Yes) × P(40|Yes) × P(92|Yes)	0.4444 × 0.02558994 × 2.789E-08	1.0048777E-06
P(Yes | Data) = P(Yes|Data) / [P(Yes|Data) + P(No|Data)]		3.1551812E-04
P(No, Data) = P(No) × P(40|No) × P(92|No)	0.5556 × 0.0800 × 2.2619E-05	1.0051948E-06
P(No | Data) = P(No|Data) / [P(No|Data) + P(Yes|Data)]		0.9996845
Maximum value		0.9996845

Based on the manual calculation using the naïve Bayes formula, the class with the highest probability value was then determined. The calculation results for the testing data example in Table 5 show that the No class has the highest probability value, 0.9996845, indicating that the testing data in Table 5 is not classified as having diabetes.

Data testing using RapidMiner was performed to obtain accurate information to support decision-making. RapidMiner is a data mining engine that can be used either as standalone software or integrated into other products for data analysis [9]. The testing process began by importing the dataset into RapidMiner using the Read CSV or Read Excel operator, ensuring that all attributes and data were read correctly. Next, a sample of data—for example, 10 records—was selected and then divided into training and testing data using the Split Data operator, with 9 records used to train the model and 1 record used for testing.

The classification model was built using the naïve Bayes operator and trained on the selected training data. Once trained, the model was applied to the testing data using the Apply Model operator to obtain the prediction results. These results could then be evaluated directly using the Performance operator or saved for further analysis. This structured and systematic testing procedure produced output consistent with the Excel-based calculations. One of the evaluation methods used was the ROC (Receiver Operating Characteristic) curve, which illustrates the model's ability to distinguish between the positive and negative classes, as shown in Figure 1.

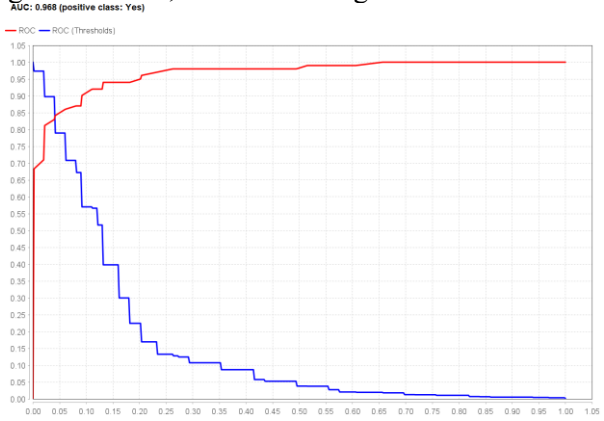

Figure 1. ROC Curve

3.3. Evaluation

The evaluation stage aims to determine the accuracy performance of the model that has been built. Evaluation was carried out in RapidMiner by adding a confusion matrix function, using an 80:20 data split, i.e., 80% training data and 20% testing data. Table 7 below presents the evaluation results.

Table 7. Naive Bayes Algorithm Accuracy Evaluation

Predicted Class	True No	True Yes	Precision
Pred. No	87	8	91.58%
Pred. Yes	12	92	88.46%
Recall	87.88%	92.00%	
Total	99	100	
Accuracy	89.95%		

The evaluation results show that the accuracy of the naïve Bayes algorithm is 89.95%. This indicates that the naïve Bayes algorithm falls into the "good" classification category for determining or classifying diabetes.

4. CONCLUSION

Based on all the research stages that have been carried out, it can be concluded that the naïve Bayes algorithm can be effectively applied in the classification process to detect whether a person is indicated to have diabetes. Testing using an 80:20 training-to-testing data split shows that the model successfully classified 104 records as having diabetes and 95 records as not having diabetes, with an accuracy rate of 89.95%, which falls into the "good" classification category. This result indicates that naïve Bayes has good predictive ability on the dataset used, particularly in processing both numerical and categorical attributes.

Nevertheless, the model still has limitations, as this study used only two attributes as predictor variables, which limits the information available for classification. The small number of attributes may reduce the model's ability to capture a broader range of characteristics among diabetes patients, so the model's performance could potentially improve if more diverse and relevant attributes were added.

For future research, it is recommended that the number of attributes used in the classification process be increased so that the model can learn more information about the characteristics of diabetes patients. Adding attributes such as family history, body mass index, insulin levels, age, and lifestyle patterns could provide a more comprehensive picture for the model's predictions. In addition, feature selection techniques could be considered to identify the most influential attributes, thereby optimizing the classification results. Future research could also compare the performance of the naïve Bayes algorithm with other classification methods, such as Decision Tree, Random Forest, or Support Vector Machine, to determine which method provides the best accuracy and reliability for diabetes detection. With these developments, the model's ability to detect diabetes is expected to improve, resulting in a more accurate and reliable system.

Funding Information
There is no funding in this article.

REFERENCES

[1] Y. B. Widodo, S. A. Anggraeini, and T. Sutabri, "Perancangan Sistem Pakar Diagnosis Penyakit Diabetes Berbasis Web Menggunakan Algoritma Naive Bayes," vol. 7, no. 1, pp. 112–123, 2021.
[2] A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," vol. IV, no. September, pp. 15–21, 2020.
[3] F. Handayani, R. Firdaus, A. Wahyudi, H. Fu, and R. M. Taufiq, "Kombinasi Algoritma Gaussian Naïve Bayes Dan Adaboost Untuk Meningkatkan Akurasi Dalam Klasifikasi Penyakit Diabetes," vol. 15, no. 2, pp. 238–244, 2025.
[4] D. N. Anisa et al., "Klasifikasi Penyakit Diabetes Melitus Anak Nonketotik Berbasis Naive Bayes," vol. 14, no. 1, pp. 33–42, 2022.

[5] H. Sitorus, V. Yasin, and A. B. Yulianto, "Perancangan sistem pakar diagnosis penyakit diabetes berbasis web menggunakan algoritma naive bayes," Jurnal Sains dan Teknologi Widyaloka, vol. 1, pp. 135–144, 2022.
[6] M. Ari, M. T. Informatika, and U. A. Yogayakarta, "Perancangan Sistem Klasifikasi Penyakit Jantung Mengunakan Naïve Bayes," vol. 6, no. 1, 2019.
[7] A. Fauzi and A. H. Yunial, "Optimasi Algoritma Klasifikasi Naive Bayes, Decision Tree, K-Nearest Neighbor, dan Random Forest menggunakan Algoritma Particle Swarm Optimization pada Diabetes Dataset," vol. 8, no. 3, pp. 470–481, 2022.
[8] W. D. Prasetya and B. Sujatmiko, "Rancang Bangun Aplikasi dengan Perbandingan Metode K-Nearest Neighbor (KNN) dan Naive Bayes dalam Klasifikasi Penderita Penyakit Diabetes," vol. 22, pp. 515–525.
[9] A. Davinka, S. Depari, and C. C. Kirana, "Prediksi Risiko Diabetes dengan Metode Naive Bayes: Identifikasi Faktor Risiko Utama dan Evaluasi Akurasi Model," vol. 9, no. 4, pp. 6372–6377, 2025.
[10] A. F. Riany, G. Testiana, S. S. Informasi, and K. Palembang, "Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes," vol. 9, pp. 42–54, 2023.