GENERATION JOURNAL

Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

Classification of Diabetes Mellitus (DM) Using the Naïve Bayes Method with Chi-Square Variable Selection

Farhan Arizal Ginanjar[1], Ambar Winarni[2], Nur'aini Muhassanah[3]
[123] Universitas Nahdlatul Ulama Purwokerto, Jalan Sultan Agung No. 42, Purwokerto, Indonesia
Email: farhanarizalginanjar@gmail.com[1], ambarwinaarni@gmail.com[2], nuraini.muhassanah8790@gmail.com3
*corresponding author : farhanarizalginanjar@gmail.com

Article Info

Received	2025-12-18
Revised	2026-01-31
Accepted	2026-02-01

Keywords *: Diabetes Mellitus; Classification;Naïve Bayes; Chi-Square*

Abstract

Diabetes mellitus (DM) is a chronic disease that can cause serious complications, making early detection essential. Technological advances enable the use of data mining techniques, particularly the Naïve Bayes classification method, to support early diabetes detection. Although Chi-Square variable selection is known to improve Naïve Bayes accuracy, studies examining the impact of different significance levels remain limited. Therefore, this study applies the Naïve Bayes method with and without Chi-Square variable selection at three significance levels (α = 0.05, α = 0.01, and α = 0.001) to evaluate their effects on classification performance and identify the optimal significance level. The results show that Naïve Bayes without variable selection achieved an accuracy of 87.50%, precision of 93.01%, and recall of 86.21%. After applying Chi-Square selection, performance improved across all significance levels. At α = 0.05, the accuracy reached 87.88%, with precision of 93.06% and recall of 86.85%. At α = 0.01, accuracy increased to 88.46%, precision to 94.25%, and recall to 86.53%. The best performance was obtained at α = 0.001, achieving an accuracy of 88.65%, precision of 94.19%, and recall of 86.86%. These findings indicate that Chi-Square variable selection effectively enhances the performance of the Naïve Bayes algorithm for diabetes classification.

INTRODUCTION

Diabetes mellitus (DM) is a chronic disease characterized by blood sugar levels exceeding normal limits, which can lead to serious complications if left untreated [1][2]. According to the International Diabetes Federation (IDF), there were approximately 589 million people with diabetes worldwide in 2024, and this number is expected to rise to 853 million by 2050. In the same year, over 3.4 million deaths were attributed to diabetes complications, and 43% of individuals with diabetes remain undiagnosed[3]. This situation underscores the urgency of early detection as a preventive measure and to reduce the risk of complications and mortality. Diabetes is generally accompanied by symptoms such as polyuria, polydipsia, polyphagia, weight loss, fatigue, blurred vision, and wounds that are difficult to heal[4][5]. These symptoms are important indicators in the process of symptom-based diabetes diagnosis, which can be used to support the early detection system for diabetes..

With the advancement of technology, the early detection of diabetes can be aided by data-based methods. One widely used approach is data mining. Data mining is the process of extracting and analysing relevant information from large databases using statistical, mathematical, machine learning, and artificial intelligence approaches [6]. Data mining uses various approaches to obtain information,

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

including description, estimation, clustering, association, prediction, and classification[7]. Among these various approaches, classification is one of the methods that is often used in various fields, including health, because it can group objects into certain groups, for example, positive or negative for a disease, based on the available data information. Thus, the use of classification methods can assist in the early detection of diabetes.

Classification is the process of categorizing objects based on similar characteristics and developing models and functions that can define, distinguishing, and predicting the class of previously unknown objects [8][9]. There are several classification methods, namely Naïve Bayes, K-Nearest Neighbor (KNN), Decision Tree (C4.5, C5.0), Artificial Neural Network (ANN), Random Forest, and Support Vector Machine (SVM) [10]. This study applies the Naïve Bayes method, which is used for various reasons, including calculation speed, algorithm simplicity, and its ability to achieve high accuracy[11]. Naïve Bayes classification is a statistical method for identifying membership of a class. The principle of this classification was introduced by Thomas Bayes, a British scientist, which is to estimate future probabilities based on previous events, which is widely known as Bayes' theorem [12]. Several previous studies have shown that the Naive Bayes method can produce good classification model accuracy. In a study by [13] using the Naïve Bayes method to classify diabetes in women, an accuracy of 78.50% was obtained. Another study by [14] evaluated the prediction of early-stage diabetes risk using the Naïve Bayes method with cross-validation, achieving an accuracy level of 87.88%. Although the Naïve Bayes method has been proven to have advantages in classifying diabetes, further research is needed to evaluate its effectiveness in various dataset conditions, including those that have undergone variable selection. This stage is important for identifying influential variables in the dataset, so that the classification model focuses on variables that have a significant influence.

Variable selection is the process of identifying and selecting a subset of influential independent variables from a large data set that will be used to form a classification model [15]. Variable selection aims to reduce irrelevant variables, speed up the model training process, and improve classifier performance [16]. However, finding the optimal variables is not an easy task because it has the potential to cause high bias or variance [17]. Therefore, an appropriate variable selection strategy that is relevant to the data is needed. Chi-Square and Mutual Information are two widely used filter-based variable selection approaches [18]. The Chi-Square method is a variable selection technique used to assess the strength of the relationship between a variable and a class[19]. According to [20] , his research proved that the application of Chi-Square variable selection in Naïve Bayes increased accuracy by 2.38%. Similar research by [21] showed an accuracy increase of 6.78%.

Previous studies have confirmed that the use of the Naive Bayes method with Chi-Square variable selection can improve the accuracy performance of classification models. Based on this description, the problem in this study is that there has been no specific study analysing the effect of differences in the significance level of Chi-Square variable selection on the accuracy performance of the Naive Bayes method in the classification of diabetes. Based on this problem, this study classified diabetes using the Naive Bayes method, both with and without Chi-Square variable selection, at three levels of significance, namely $\alpha = 0.05$; $\alpha = 0.01$; and $\alpha = 0.001$. This study aims to compare the accuracy performance of Naive Bayes without variable selection and Naive Bayes with Chi-Square variable selection at various levels of significance so that it can be determined to what extent variable selection affects the improvement in the accuracy of the diabetes classification model.

RESEARCH METHODS

Overall, this study was conducted to optimise the performance of the Naive Bayes classification method in classifying diabetes patients by comparing the effectiveness of classification models with and without Chi-Square variable selection at three different levels of significance.

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

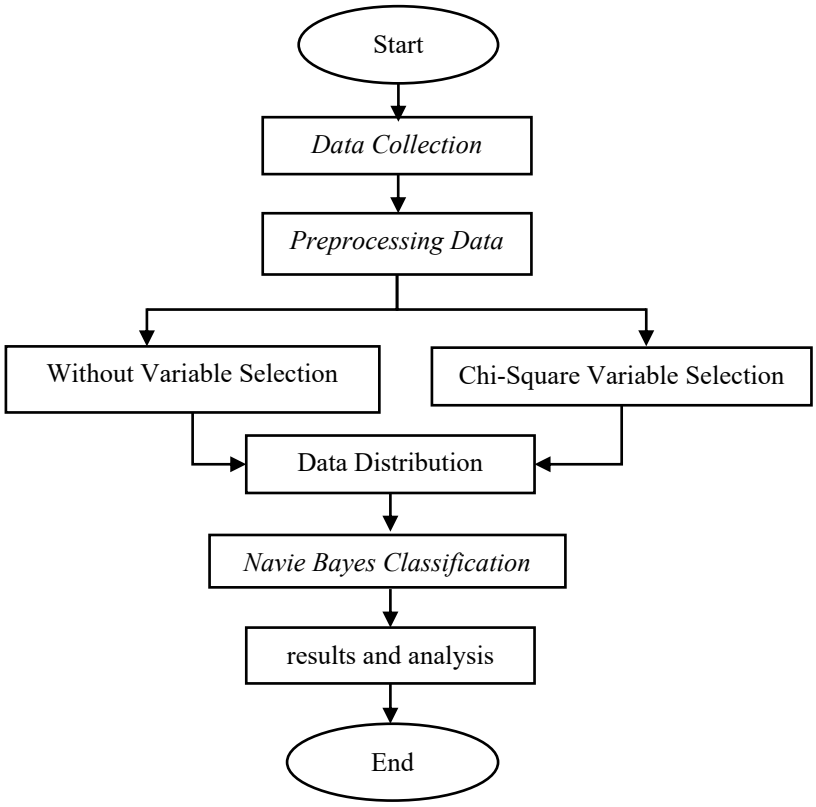

Figure 1. Research methodology

Data Collection

The data used in this study is secondary data obtained from Kaggle entitled Early-Stage Diabetes Risk Prediction Dataset. The dataset is based on questionnaire results obtained directly from patients at Sylhet Diabetes Hospital in Sylhet, Bangladesh, and has been approved by medical professionals. This study uses 520 data points with 16 variables, consisting of 15 independent variables and 1 dependent variable.

Data Preprocessing

Data preprocessing is an important process that aims to process data so that it is more structured and can be used in the analysis process [22]. The preprocessing stage used in this study encodes labels by transforming categorical data into numerical form.

Variable Selection

Variable selection is the process of identifying and selecting a subset of variables to be used in a machine learning model. The variable selection process is used to optimise performance by eliminating variables that have no effect or contribute little to the final classification model results[23].

Chi-Square

Chi-Square is a nonparametric statistical test method used to assess whether two variables show a statistically significant relationship [24]. The higher the Chi-Square value, the stronger the relationship between the variable and the target class [25].

The following is the Chi-Square test equation [26]:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad \text{.....(1)}$$

where

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

X^2 = Chi-square test distribution
O_i = i^{th} observed value
E_i = i^{th} expected value

The Chi-Square test procedure is as follows:

1. Establish hypotheses H_0 and H_1
   H_0 : There is no significant relationship between the two variables.
   H_1 : There is a significant relationship between the two variables.
2. Calculate the expected frequency (E_i)

$$E_i \; for \; each \; cell = \frac{(Total \; Row) \times (Total \; Columns)}{Total \; Overall} \quad \text{.....(2)}$$

3. Calculate the Chi-Square test distribution using equation (1)
4. Determine the significance level. In this study, three significance levels are used, namely $a = 0{,}05$, $a = 0{,}01$ and $a = 0{,}001$
5. Determining the X^2 table value

$$d.f = (Number \; of \; row - 1) \times (Number \; of \; columns - 1) \text{.....(3)}$$

6. Formulating the test criteria
   If the calculated X^2 value $\leq$ the X^2 table value, then the H_0 hypothesis is accepted
   If the calculated X^2 value > the X^2 table value, then the H_0 hypothesis is rejected
7. Comparing the calculated X^2 and X^2 table
8. Making a decision on whether there is a relationship between variables

Data Distribution

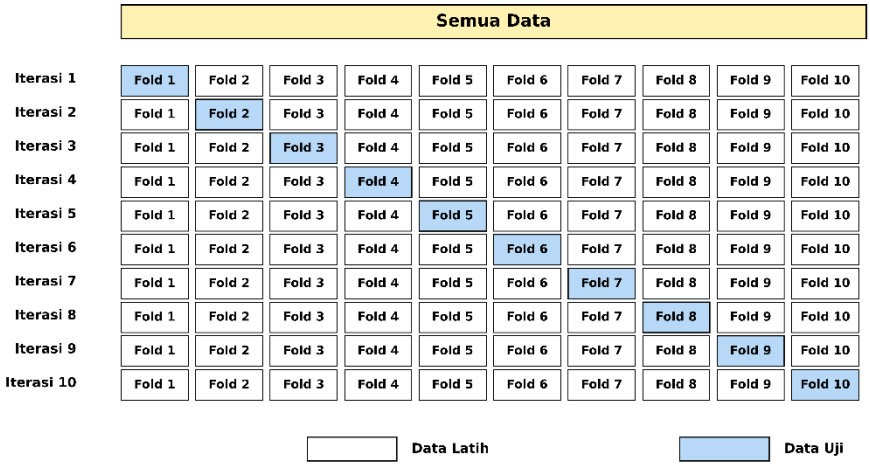

Figure 2. K-Fold Cross Validation Scheme k=10

Data splitting is a method of separating data into two groups, namely training data and test data, but unbalanced data splitting can result in bias. K-Fold Cross Validation is an approach that can be used to optimise the performance of classification models by splitting the dataset into k folds, with (k-1) used for training data and the rest as test data. The analysis stage is carried out using classification methods for k iterations. The final stage of this process is to calculate the average results of all iterations to obtain a more stable model evaluation[27][28]. The K-Fold Cross Validation approach with k=10 can produce accuracy with relatively low variance and bias [29]. Therefore, in this study, the K-Fold Cross Validation method with k=10 was applied to divide the data in the classification of diabetes.

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

Naïve Bayes Classification Method

Naïve Bayes classification is a statistical classification method used to predict the probability of data belonging to a particular class[30]. This method is based on the concept of probabilistic theory and statistics introduced by a British scientist named Thomas Bayes, who estimated future probabilities based on past events, which later became widely known as Bayes' theorem [31]. Naïve Bayes classification can be applied in the following stages [32]:

1. Reading training data
2. Calculate the prior probability for class Y_i using the following equation:

$$P(Y_i) = \frac{Y_i}{Y} \quad \text{.....(4)}$$

dengan

$P(Y_i)$ = prior probability of class Y_i occuring
Y_i = classification result of the training data
Y = total number of training data

3. Calculate the probability of each independent variable relative to its dependent variable using the following conditional probability equation:

$$P(Y|X) = \frac{P(X \cap Y)}{P(X)} \quad \text{.....(5)}$$

where

$P(Y|X)$ = probability of Y occurring given that X has already occurred
$P(X \cap Y)$ = initial probability of X given that Y has already occurred simultaneously
$P(X)$ = probability of condition X occurring

4. Calculate the total posterior probability value for each class using the following equation:

$$P(Y_i|X) = \prod_{k=1}^{n} P(x_k|Y_i) = P(x_1|Y_i) \times P(x_2|Y_i) \times \ldots \times P(x_n|Y_i) \quad (6)$$

5. Calculate the product of the prior probability and the total probability of the independent variable for each class.
6. The final classification result is determined based on the highest probability value between the two classes.

In an effort to clarify the process flow of diabetes classification using the Naïve Bayes method with and without Chi-Square variable selection, the following pseudocode briefly and systematically describes the main stages of the research.

```
Input: Diabetes dataset (520 data points, 16 variables)
Output: Accuracy, Precision, Recall values
Begin
 1. Read the diabetes dataset
 2. perform preprocessing data
    a. Encoding Gender → (1 (Male), 2 (Female))
    b. Encoding symptoms → (1 (Yes), 0 (No))
    c. Encoding Class → (1 (Positive), 0 (Negative))
 3. Determine the research scenario
        a. Without variable selection
        b. With Chi-Square variable selection (a=0.05, a=0.01, and a=0.001)
 4. If using Chi-Square, then
    a. Calculate the Chi-Square value of each variable against Class
    b. Compare the calculate X² with the table X²
    c. Select variables with X² calculate > X² Table
    d. create a dataset of variable selection results
        Else
     a. Use all variables
 5. Divide the dataset randomly using K-Fold Cross Validation (k = 10)
 6. For each fold i (i = 1 to 10) do the following
      a. Determine the training data and test data
      b. Calculate the prior probability of each class
```

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

```
      c. Calculate the conditional probability of each variable for each class
      d. Calculate the posterior probability of each class
      e. Calculate the product of the prior probability and the total posterior
         probability for each class
      f. Determine the classification result based on the highest probability
      g. Save the prediction results
    7. Calculate the confusion matrix (TP, TN, FP, FN)
    8. Calculate the evaluation values
       a. Accuracy
       b. Precision
       c. Recall
   9. Calculate the average evaluation results from all folds
End
```

Model Evaluation

Model evaluation in this research stage aims to measure the performance of the model that has been built by referring to the accuracy, precision, and recall values using a confusion matrix. Model evaluation with a confusion matrix is used to measure the effectiveness of the classification model's performance after the data mining process. This evaluation produces information that serves to measure the performance of the classification model results with the actual classification [33].

Table 1. *Confusion Matrix*

Confusion matrix		**Actual**	
		Positive	Negative
Prediction	Positive	*TP*	*FP*
	Negative	*FN*	*TN*

Explanation:

- *TP (True Positive)* the number of actual positive class data predicted as positive values.
- *TN (True Negative)* is the number of actual negative class data predicted as negative values.
- *FN (False Negative)* is the number of actual positive class data predicted as negative values.
- *FP (False Positive)* is the number of actual negative class data predicted as positive values.

The confusion matrix contains three methods for evaluating classification models on test data [34]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad \text{.....(7)}$$

Accuracy serves as the result of model performance assessment during the classification process.

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad \text{..... (8)}$$

Precision serves to compare the number of True Positive (TP) data with the total number of data predicted as positive

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad \text{..... (9)}$$

Recall serves to compare the number of True Positive (TP) data with the total number of data that are actually positive

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

RESULT AND DISCUSSION

Data Collection

This study utilised secondary data from the kaggle.com platform entitled 'Early Stage Diabetes Risk Prediction Dataset'. This dataset was collected from surveys conducted directly with patients at Sylhet Diabetes Hospital in Sylhet, Bangladesh, and was approved by authorised medical personnel. This study examined 520 observational data sets, covering 16 variables, including 15 independent variables and one dependent variable.

Table 2. Description of Diabetes Variables

Variable	Description	Category	Description
X_1	*Gender*	Male Female	gender
X_2	*Polyuria*	Yes No	frequent urination with abnormal frequency
X_3	*Polydipsia*	Yes No	hyperphysiological thirst
X_4	*Sudden weight loss*	Yes No	significant weight loss
X_5	*Weakness*	Yes No	decreased physical endurance
X_6	*Polyphagia*	Yes No	excessive increase in appetite
X_7	*Genital thrush*	Yes No	fungal and bacterial infection
X_8	*Visual blurring*	Yes No	blurred vision
X_9	*Itching*	Yes No	Persistent itching
X_{10}	*Irritability*	Yes No	emotional instability
X_{11}	*Delayed healing*	Yes No	wounds that are difficult to heal
X_{12}	*Partial paresis*	Yes No	weakness or limpness in the body
X_{13}	*Muscle stiffness*	Yes No	difficulty moving or stiff muscles
X_{14}	*Alopecia*	Yes No	hair loss
X_{15}	*Obesity*	Yes No	obesity
Y	*Class*	Positive Negative	classified as positive or negative diabetes

Preprocessing Data

The preprocessing stage is carried out to prepare the dataset so that it can be processed by the Naïve Bayes method. This process includes label encoding to convert categorical data into numerical data. The Gender variable is labelled with a value of 1 for 'Male' and 2 for 'Female', while all symptom variables

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

are labelled 1 for "Yes" and 0 for 'No'. The Class variable as the output label is given a value of 1 for the category 'Positive' and 0 for the category 'Negative'.

Table 3. Result of Preprocessing Label Encoding

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	Y
1	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1
2	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1
3	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0	1
4	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1
5	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
520	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Chi-Square Variable Selection Results

The variable selection method used for this study was Chi-Square. Variable selection is a method for selecting independent variables that show a significant relationship with the dependent variable in order to improve the performance of the classification model. Figure 3 shows the results of the Chi-Square method calculation.

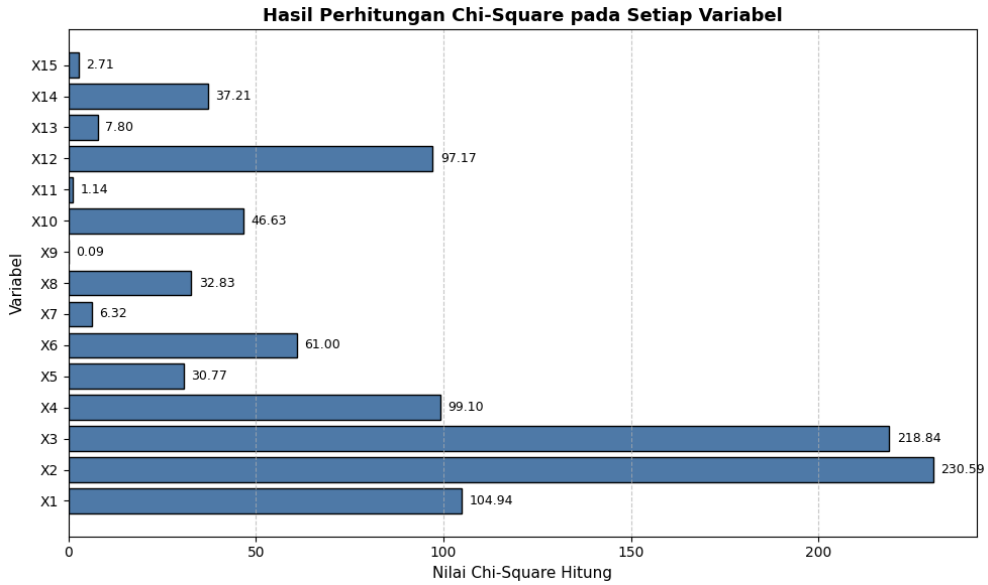

Figure 3. Chi-Square Calculation Results

The figure above shows the calculated Chi-Square value in the diabetes dataset, which illustrates the level of influence of the independent variables on the dependent variable. A high calculated Chi-Square value indicates a significant influence of the symptom variables on diabetes status. In the variable selection stage, testing was carried out using three levels of significance, namely a=0.05, a=0.01 and a=0.001. In the Chi-Square test steps described above, variables that are significantly influential are retained, and variables that are not significantly influential are not included in the classification process. Based on the results of the Chi-Square test on the diabetes dataset, several variables meet the significance requirements.

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

Table 4. Results of Chi-Square Test Variable Selection

Significant Variables Chi-Square Test $a = 0,05$	**Significant Variables Chi-Square Test $a = 0,01$**	**Significant Variables Chi-Square Test$a = 0,001$**
Gender, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Irritability, Partial paresis, Muscle stiffness, Alopecia	*Gender, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Visual blurring, Irritability, Partial paresis, Muscle stiffness, Alopecia*	*Gender, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Visual blurring, Irritability, Partial paresis, Alopecia*

Data Division

In this study, the data set was divided randomly using the K-Fold Cross Validation method with k=10, where each fold consisted of 52 observations, with a total of 468 observations in the training data and 52 observations in the test data.

Naïve Bayes Classification Results

The results of the Naïve Bayes classification method for diabetes using the K-Fold Cross Validation scheme at k= 10 obtained an average accuracy value of 87.50%, precision of 93.01% and recall of 86.21%.

Table 5. Naïve Bayes Classification Results

Fold	***Accuracy***	***Precision***	***Recall***
1	86,54%	92,86%	83,87%
2	94,23%	96,15%	92,59%
3	86,54%	90,63%	87,88%
4	92,31%	92,59%	92,59%
5	90,38%	97,06%	89,19%
6	82,69%	88,57%	86,11%
7	78,85%	86,21%	78,13%
8	82,69%	95,45%	72,41%
9	94,23%	100,00%	91,43%
10	86,54%	90,63%	87,88%
Average	**87,50%**	**93,01%**	**86,21%**

Results of Naïve Bayes Classification with Chi-Square

The results of the Naïve Bayes classification method with Chi-Square at three significance levels using the K-Fold Cross Validation technique at k = 10 obtained the following accuracy, recall, and precision classification models:

Table 6. Results of Naïve Bayes Classification with Chi-Square

Model	**Fold**	***Accuracy***	***Precision***	***Recall***
Results of Naïve Bayes Classification with Chi-Square $a = 0,05$	1	90,38%	93,33%	90,32%
	2	94,23%	96,15%	92,59%
	3	86,54%	90,63%	87,88%
	4	92,31%	92,59%	92,59%
	5	90,38%	97,06%	89,19%
	6	82,69%	88,57%	86,11%
	7	78,85%	86,21%	78,13%
	8	82,69%	95,45%	72,41%
	9	94,23%	100,00%	91,43%
	10	86,54%	90,63%	87,88%
	Average	**87,88%**	**93,06%**	**86,85%**

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

Model	Fold	*Accuracy*	*Precision*	*Recall*
Results of Naïve Bayes Classification with Chi-Square $a = 0{,}01$	1	88,46%	93,10%	87,10%
	2	92,31%	96,00%	88,89%
	3	86,54%	93,33%	84,85%
	4	96,15%	96,30%	96,30%
	5	92,31%	100,00%	89,19%
	6	84,62%	91,18%	86,11%
	7	78,85%	86,21%	78,13%
	8	82,69%	95,45%	72,41%
	9	94,23%	100,00%	91,43%
	10	88,46%	90,91%	90,91%
	Average	**88,46%**	**94,25%**	**86,53%**
Results of Naïve Bayes Classification with Chi-Square $a = 0{,}001$	1	86,54%	92,86%	83,87%
	2	94,23%	96,15%	92,59%
	3	90,38%	93,75%	90,91%
	4	92,31%	92,59%	92,59%
	5	92,31%	100,00%	89,19%
	6	84,62%	91,18%	86,11%
	7	78,85%	86,21%	78,13%
	8	84,62%	95,65%	75,86%
	9	94,23%	100,00%	91,43%
	10	88,46%	93,55%	87,88%
	Average	**88,65%**	**94,19%**	**86,86%**

Comparison of Classification Models

A comparison of the accuracy performance of the Naïve Bayes classification model without variable selection and the Naïve Bayes classification model using Chi-Square variable selection at three levels of significance was used to determine the extent to which variable selection affects the improvement in the accuracy of the diabetes classification model. Based on the evaluation results, a comparison of the accuracy, recall and precision values of the classification model is shown in Table 7 below:

Table 7. Comparison of Accuracy, Recall, and Precision of Classification Models

Classification Model	*Accuracy*	*Precision*	*Recall*
Naïve Bayes	87,50%	93,01%	86,21%
Naïve Bayes + Chi-Square a = 0,05	87,88%	93,06%	86,85%
Naïve Bayes + Chi-Square a = 0,01	88,46%	94,25%	86,53%
Naïve Bayes + Chi-Square a = 0,001	88,65%	94,19%	86,86%

The results of the experiment show that the best performance value can be measured based on the accuracy level. The application of Chi-Square variable selection has been proven to improve the accuracy of the Naïve Bayes classification model. At a significance level of a=0.05, the accuracy reached 87.88%, an increase of 0.38% from the model without variable selection. At a significance level of a=0.01, accuracy reached 88.46%, an increase of 0.96%, and at a=0.001, accuracy reached 88.65%, an increase of 1.15% compared to the model without variable selection. Based on these results, the best accuracy value was obtained in the Naïve Bayes classification model with Chi-Square when using a significance level of a=0.001, which was 88.65%. This finding indicates that a smaller a value results in a more rigorous variable selection process that is more relevant to the dependent variable. These results support the principle of the Chi-Square test, namely that a higher Chi-Square value indicates that the variable has strong discriminatory power in distinguishing between different class labels [35].

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

CONCLUSION

Based on the results of research conducted on the classification of diabetes mellitus using the Naïve Bayes method with the application of Chi-Square variable selection, several conclusions were obtained as follows:

1. The Naïve Bayes method without variable selection showed good classification capabilities with an accuracy value of 87.50%, precision of 93.01%, and recall of 86.21%. These results indicate that the Naïve Bayes method is capable of identifying patterns in diabetes symptom data with a prediction accuracy of 87.50%.
2. The application of Chi-Square variable selection has been proven to have a positive effect on improving model performance. At a=0.05, the model accuracy increased to 87.88%; at a=0.01, the model accuracy increased to 88.46%; and at a=0.001, it reached 88.65%. Thus, there was an increase in accuracy of 0.38% to 1.15% compared to the model without variable selection. This indicates that the variable selection process helps the model focus on the most relevant variables, resulting in more accurate classification.
3. These findings show that a more rigorous variable selection process and the use of a smaller significance value improve the model's ability to recognise significant data patterns. Overall, the combination of Naïve Bayes classification with Chi-Square variable selection proved to be effective in the early detection of diabetes because it was able to provide more accurate results.

SUGGESTION

Several suggestions and recommendations for further research development related to the application of the Naïve Bayes classification method with Chi-Square variable selection are as follows:

1. Further research is recommended to combine the Chi-Square method with other variable selection techniques, such as Mutual Information, Forward Selection, Backward Elimination, and Recursive Feature Elimination (RFE).
2. The variable selection method that has been developed is recommended to be applied to other classification methods, such as Artificial Neural Network (ANN), Random Forest, K-Nearest Neighbour (KNN), Decision Tree (C4.5, C5.0), and Support Vector Machine (SVM). This application aims to broaden the scope of research while comparing the performance results of each algorithm, thereby obtaining the most optimal classification model.

REFERENCES

[1] I. Roifah, "Analisis Hubungan Lama Menderita Diabetes Mellitus Dengan Kualitas Hidup Penderita Diabetes Mellitus," *Jurnal Ilmu Kesehatan*, vol. 4, no. 2, 2016, doi: https://doi.org/10.32831/jik.v4i2.84.

[2] K. Yudianto, H. Rizmadewi, and I. Maryati, "KUALITAS HIDUP PENDERITA DIABETES MELLITUS DI RUMAH SAKIT UMUM DAERAH CIANJUR," 2008.

[3] International Diabetes Federation, "IDF Diabetes Atlas 11th Edition," 2025.

[4] Lestari, Zulkarnain, and S. Aisyah Sijid, "Diabetes Melitus: Review Etiologi, Patofisiologi, Gejala, Penyebab, Cara Pemeriksaan, Cara Pengobatan dan Cara Pencegahan," *Jurusan Biologi, Fakultas Sains dan Teknologi*, p. 237, Nov. 2021, [Online]. Available: http://journal.uin-alauddin.ac.id/index.php/psb

[5] O. R. Simatupang, M. Kristina, S. Nauli, and H. Sibolga, "PENYULUHAN TENTANG DIABETES MELITUS PADA LANSIA PENDERITA DM," *JPM Jurnal Pengabdian Mandiri*, vol. 2, no. 3, 2023, [Online]. Available: http://bajangjournal.com/index.php/JPM

[6] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Jurnal Edik informatika*, vol. 2, pp. 213–219, 2017, doi: https://doi.org/10.22202/ei.2016.v2i2.1465.

Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

[7] T. Novika, P. Poningsih, H. Okprana, A. P. Windarto, and H. Siahaan, "Penerapan Data Mining Klasifikasi Tingkat Pemahaman Siswa Pada Pelajaran Matematika," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 1, pp. 9–17, Jan. 2021, doi: 10.30865/mib.v5i1.2498.

[8] E. K. Putri and T. Setiadi, "PENERAPAN TEXT MINING PADA SISTEM KLASIFIKASI EMAIL SPAM MENGGUNAKAN NAIVE BAYES," *Jurnal Sarjana Teknik Informatika*, vol. 2, pp. 73–83, 2014, doi: https://doi.org/10.12928/jstie.v2i3.2877.

[9] D. Septhya *et al.*, "Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, pp. 15–19, Apr. 2023, doi: https://doi.org/10.57152/malcom.v3i1.591.

[10] M. Arifin, F. Helmi, and R. Bagus Hikmawansyah, "ANALISIS METODE DAN ALGORITMA DALAM SISTEM PENDUKUNG KEPUTUSAN UNTUK MEMPREDIKSI KELULUSAN," *Jurnal Advance Research Informatika*, vol. 3, no. 1, p. 73, 2024, doi: https://doi.org/10.24929/jars.v3i1.4045.

[11] H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, and I. Cholissodin, "OPTIMASI NAÏVE BAYES CLASSIFIER DENGAN MENGGUNAKAN PARTICLE SWARM OPTIMIZATION PADA DATA IRIS," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 4, no. 3, pp. 180–184, Sep. 2017, doi: 10.25126/jtiik.201743251.

[12] I. Riswanto and R. H. Laluma, "KLASIFIKASI KELAYAKAN PINJAMAN PADA KOPERASI KARYAWAN MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER BERBASIS WEB," *Infotronik : Jurnal Teknologi Informasi dan Elektronika*, vol. 5, no. 1, pp. 11–16, Jun. 2020, doi: 10.32897/infotronik.2020.5.1.2.

[13] A. V. Agustin and A. Voutama, "IMPLEMENTASI DATA MINING KLASIFIKASI PENYAKIT DIABETES PADA PEREMPUAN MENGGUNAKAN NAÏVE BAYES," 2023. doi: https://doi.org/10.36040/jati.v7i2.6808.

[14] M. Danny and A. Muhidin, "Analisis Prediksi Resiko Diabetes Tahap Awal Menggunakan Algoritma Naive Bayes," *Jurnal Teknologi Informatika dan Komputer MH. Thamrin*, vol. 9, no. 2, pp. 1443–1459, Sep. 2023, doi: 10.37012/jtik.v9i2.2017.

[15] R. Aziz, C. K. Verma, and N. Srivastava, "Dimension reduction methods for microarray data: a review," *AIMS Bioeng.*, vol. 4, no. 1, pp. 179–197, 2017, doi: 10.3934/bioeng.2017.1.179.

[16] G. Kicska and A. Kiss, "Comparing swarm intelligence algorithms for dimension reduction in machine learning," *Big Data and Cognitive Computing*, vol. 5, no. 3, Sep. 2021, doi: 10.3390/bdcc5030036.

[17] D. H. Jeong, B. K. Jeong, N. Leslie, C. Kamhoua, and S.-Y. Ji, "Designing a supervised feature selection technique for mixed attribute data analysis," *Machine Learning with Applications*, vol. 10, p. 100431, Dec. 2022, doi: 10.1016/j.mlwa.2022.100431.

[18] H. Chauhan, K. Modi, and S. Shrivastava, "Development of a classifier with analysis of feature selection methods for COVID-19 diagnosis," *World Journal of Engineering*, vol. 19, no. 1, pp. 49–57, Feb. 2021, doi: 10.1108/WJE-10-2020-0537.

[19] D. R. Anamisa, F. A. Mufarroha, and A. Jauhari, "Visitor Decision System in Selection of Tourist Sites Based on Hybrid of Chi-Square And K-NN Methods," *Elinvo (Electronics, Informatics, and Vocational Education)*, vol. 8, no. 2, pp. 248–254, Jan. 2024, doi: 10.21831/elinvo.v8i2.55702.

[20] D. Ryanto Fernandes, N. Jacky Pratama Hasan, and N. Wijaya, "Optimasi Akurasi Sentimen Komentar Xiaomi SU7 di YouTube Menggunakan Naive Bayes dan Chi-Square," vol. 2, no. 1.

[21] R. Yunita Kisworini and M. Akbar Setiawan, "Peningkatan Performa Naivee Bayes Dengan Seleksi Atribut Menggunakan Chi Square Untuk Klasifikasi Loyalitas Pelanggan GRAB," *Journal of Informatics, Information System, Software Engineering and Applications*, vol. 2, no. 2, pp. 69–075, 2020, doi: 10.20895/INISTA.V2I2.

Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.27878

[22] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.

[23] D. Kurnia, M. Itqan Mazdadi, D. Kartini, R. Adi Nugroho, and F. Abadi, "Seleksi Fitur dengan Particle Swarm Optimization pada Klasifikasi Penyakit Parkinson Menggunakan XGBoost," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 5, pp. 1083–1094, Oct. 2023, doi: 10.25126/jtiik.2023107252.

[24] Y. Kamila, A. Sa'idah, A. S. Akbar, F. A. N. Azzen, A. Y. B. Rohim, and N. Chamidah, "Analisis Hubungan Antara Jalur Masuk Universitas dengan Predikat Kelulusan Mahasiswa," *Zeta - Math Journal*, vol. 8, no. 1, pp. 23–29, May 2023, doi: 10.31102/zeta.2023.8.1.23-29.

[25] N. Adliani Awalia, R. Nur Shofa, and S. Yuliyanti, "Perbandingan Algoritma Pendekatan Supervised Learning Menggunakan Seleksi Fitur Chi-Square untuk Klasifikasi Status Kesehatan Jemaah Haji," *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, vol. 13, pp. 166–2, Jan. 2025, doi: 10.26418/justin.v13i1.86639.

[26] I. Cahya Negara and A. Prabowo, "PENGGUNAAN UJI CHI–SQUARE UNTUK MENGETAHUI PENGARUH TINGKAT PENDIDIKAN DAN UMUR TERHADAP PENGETAHUAN PENASUN MENGENAI HIV–AIDS DI PROVINSI DKI JAKARTA," Purwokerto, Sep. 2018.

[27] J. Homepage, A. Putri, and B. Purnama, "Classification of Scholarship Eligibility Using Naïve Bayes with Attribute Optimization Based on K-Means Clustering," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, pp. 1450–1462, 2025, doi: 10.57152/malcom.v5i4.2312.

[28] Moch. A. Aprihartha and I. Idham, "Optimization of Classification Algorithms Performance with k-Fold Cross Validation," *EIGEN MATHEMATICS JOURNAL*, vol. 7, no. 2, pp. 61–66, Sep. 2024, doi: 10.29303/emj.v7i2.212.

[29] E. Etriyanti, D. Syamsuar, and Y. Novaria Kunang, "Implementasi Data Mining Menggunakan Algoritme Naive Bayes Classifier dan C4.5 untuk Memprediksi Kelulusan Mahasiswa," *Telematika*, vol. 13, no. 1, pp. 56–67, Feb. 2020, doi: 10.35671/telematika.v13i1.881.

[30] I. D. Ratih, S. M. Retnaningsih, and V. M. Dewi, "Klasifikasi Kualitas Tanah Menggunakan Metode Naive Bayes Classifier," *Jurnal Aplikasi Matematika dan Statistik)*, vol. 1, pp. 11–20, 2022, doi: 10.53625/jams.v1i1.4227.

[31] Fadlisyah and S. Eliyanda, "PENGELOMPOKAN SISWA PENYANDANG DISABILITAS BERDASARKAN TINGKAT TUNAGRAHITA MENGGUNAKAN METODE NAÏVE BAYES," vol. 2, Aug. 2021, doi: 10.29103/tts.v2i1.3703.

[32] L. U. Khasanah, Y. N. Nasution, F. Deny, and T. Amijaya, "Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma Naïve Bayes Classifier," *Jurnal Ilmiah Matematika*, vol. 1, no. 1, pp. 41–50, 2022, doi: https://doi.org/10.30872/basis.v1i1.918.

[33] P. L. Romadloni, B. A. Kusuma, and W. M. Baihaqi, "KOMPARASI METODE PEMBELAJARAN MESIN UNTUK IMPLEMENTASI PENGAMBILAN KEPUTUSAN DALAM MENENTUKAN PROMOSI JABATAN KARYAWAN," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 6, pp. 622–628, 2022, doi: https://doi.org/10.36040/jati.v6i2.5238.

[34] A. B. Alpiansah and Y. Ramdhani, "Optimasi Fitur dengan Forward Selection pada Estimasi Tingkat Obesitas menggunakan Random Forest," 2023. doi: https://doi.org/10.32520/stmsi.v12i3.3125.

[35] C. Shi, J. Gao, J. Yu, L. Zhao, and F. Jia, "A novel similarity-constrained feature selection method for epilepsy detection via EEG signals," *Journal of King Saud University - Computer and Information Sciences*, vol. 37, no. 6, pp. 1–24, Aug. 2025, doi: 10.1007/s44443-025-00152-w.