



Comparison of Naive Bayes Method and Support Vector Machine in Classifying Diabetes Mellitus Disease

Indah Kusuma Sari, Rizky Putra Wijaya

Sekolah Tinggi Teknologi Jayapura, Papua, Indonesia

Article Info

Keywords:

Data Mining,
Diabetes Mellitus,
Naïve Bayes,
Support Vector Machine

ABSTRACT

Diabetes mellitus is a chronic disease that occurs due to excessively high blood glucose levels resulting in the absence of insulin. In the period of data at the Siti Khadijah Islamic Hospital in Palembang, which is influenced by the number of patients undergoing health checks such as diabetes mellitus, it affects the classification of data that will complicate the hospital. So by utilizing data mining, classification to determine patients who have undergone examinations including diabetes sufferers or not. With these problems, the author conducted a comparative analysis of two algorithms, namely the naïve Bayes algorithm and the support vector machine algorithm for the classification of diabetes by using the WEKA tool with the Cross Validation and Confusion Matrix options tools with the highest accuracy results, namely the support vector machine algorithm with a polynomial kernel, the results of which are 96.2704% and an error rate of 3.7296%, it can be concluded that the most accurate algorithm in the classification of diabetes is the support vector machine algorithm with a polynomial kernel.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Indah Kusuma Sari
Sekolah Tinggi Teknologi Jayapura
Email: indahkusuma@sttjaya.ac.id

1. INTRODUCTION

Diabetes mellitus, better known in Indonesia as "diabetes mellitus," is a disease that causes uncontrolled blood sugar levels, leading to a lack of insulin, which prevents the body from functioning normally. Siti Khadijah Hospital in Palembang, a private hospital accredited type B, is located on Jalan Demang Lebar Daun Lorok Pakjo, Ilir Barat I District, Palembang City, South Sumatra. Over time, the large number of patients undergoing medical examinations has resulted in a significant accumulation of data at the hospital. These examination results can potentially diagnose patients with diabetes mellitus, making it difficult for the hospital to classify the data. Therefore, techniques that can support the solution to the above problem are needed, such as applying data mining for classification using algorithms such as the Naive Bayes algorithm and Support Vector Machine.

Previously, there were a number of previous studies conducted by several researchers such as a comparison of the Naive Bayes and Support Vector Machine algorithms, including a study conducted by (Riyanto, 2018) to classify the number of online readers by comparing the Naive Bayes and SVM methods,

which produced an accurate level of accuracy, namely the SVM algorithm with a value of 63.39%. Furthermore, a study conducted by (Arifin & Sasongko, 2018) compared the Naive Bayes and SVM algorithms for the classification of high school interest paths, the accuracy results of which showed that the SVM method was more accurate than the Naive Bayes algorithm with an accuracy value of 97.01%. There was also a study conducted by (Widyawati & Sutanto, 2019) which compared the Naive Bayes and SVM algorithms in the classification of Indonesian language spam SMS, the accuracy results of which showed that the Naive Bayes algorithm was superior in terms of 94% recall and 95% precision. So the question will arise in this study "Which algorithm has a more accurate level of accuracy in classifying diabetes mellitus between Naive Bayes and Support-Vector-Machine?".

This study used 613 records from 2017-2019. The purpose of this study was to determine which algorithm has the highest accuracy for classifying diabetes mellitus, between the Naive-Bayes algorithm and the Support Vector Machine. The benefits of this study include understanding the concept of data mining and providing useful references for future research on similar topics.

2. RESEARCH METHOD

2.1 Data Mining

According to research by Saleh & Utama (2016), data mining is a process of automatically extracting knowledge or analyzing data to find information that employs one or more machine learning techniques. Broadly speaking, data mining is a process or method for automatically searching data to obtain information that functions as a storage location for large data (big data). The method for classifying diabetes mellitus uses a classification method with the characteristics and attributes of the data that have been defined. The algorithms used in the classification process are the naïve Bayes algorithm and the support vector machine.

2.2 Diabetes Mellitus

According to research from (Nugroho et al., 2017) Diabetes Mellitus (DM) or in Indonesia better known as diabetes is a disorder or disease caused by increased blood sugar due to decreased insulin secretion or an imbalance in the body's use of insulin, resulting in uncontrolled glucose or blood sugar levels. Type 2 diabetes is a hyperglycemic disease or metabolic disorder characterized by increased blood sugar levels due to decreased insulin secretion.

2.3 Naïve Bayes Algorithm

In research (Andriyanto & Santoso, 2017) Naïve Bayes Algorithm is a method that utilizes probability and statistics to solve classification problems. This method performs classification by calculating the value of the probability $P(x|y)$ by knowing the probability of class X. Determination of the class in classification is done by selecting the maximum value of $P(x|y)$ based on probability. The advantage of classification is that it requires a relatively small amount of training data to estimate the parameters needed for classification.

2.4 Support Vector Machine Algorithm

According to research (Sukmawati et al., 2013), Support Vector Machine (SVM) is a method or algorithm for classification and prediction. The working principle of this method is to find the most optimal separating space from a dataset into different classes. In everyday life, we are often faced with non-linear problems or data that cannot be truly separated linearly, namely a condition where there is no line or plane that can be created to separate the data classes.

2.5 K-Fold Cross Validation

According to research (Arifin & Sasongko, 2018) that K-folds cross validation is a validation technique for testing that functions to assess the performance of an algorithm process and is used to reduce time in the classification process while maintaining accuracy values by dividing data randomly and grouping data as many as K-Folds. The function of K-folds cross validation is to prevent overlapping in the testing data.

1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10

Figure 1. Illustration of K-Fold Cross Validation

2.6 Confusion matrix

Rahman et al. (2017) argue that a confusion matrix is a method for providing information on the results of a classification system, which is useful for analyzing how well a classifier recognizes tuples from different classes. For example, for a two-class confusion matrix, the positive and negative classes will be mentioned. True positives refer to positive classes that are correctly marked by the classifier, while true negatives are negative classes that are correctly marked by the classifier. False positives are negative classes that are incorrectly marked. Furthermore, false negatives are positive classes that are incorrectly labeled.

Table 1. *Confusion matrix*

Correct classification	Classification	
	Positive (+)	Negative (-)
Positively true	<i>True Positive (TP)</i>	<i>False Negative (FN)</i>
Negative true	<i>False Positive (FP)</i>	<i>True Negative (TN)</i>

Description: 1) True Positive (TP), Total of positive data classified as positive by the system. 2) False Positive (FP), Total of negative data classified as positive by the system. 3) False Negative (FN), Total of positive data classified as positive by the system. 4) True Negative (TN), Total of negative data classified as negative by the system. Description: 1) True Positive (TP), Total of positive data classified as positive by the system. 2) False Positive (FP), Total of negative data classified as positive by the system. 3) False Negative (FN), Total of positive data classified as positive by the system. 4) True Negative (TN), Total of negative data classified as negative by the system.

2.7 Data Collection Methods

The data comes from medical records at Siti Khadijah Hospital, Palembang, totaling 613 records with 9 attributes over a period of almost three years. For more details on the dataset used, see the image below: data on diabetes patients in PDF format.

No.	Tgl masuk	Tgl keluar	Kadar glukosa puasa(mg/dl)	Kadar glukosa setelah 2 jam(mg/dl)	Tekanan darah (mm/hg)	Berat Badan(Kg)	Umur(Th)	Jenis kelamin(L/P)
1	03-Jun	07-Apr	300	262	120/80	77	52	P
2	05-Apr	09-Apr	271	217	120/80	68	50	P
3	29-Apr	02-May	251	201	120/70	54	58	L
4	01-Jan	04-Jan	300	248	120/70	60	61	P
5	21-Mar	21-Mar	216	178	120/70	47	45	P
6	02-Feb	10-Apr	218	180	120/80	61	63	P
7	30-Jun	30-Jun	226	150	120/70	55	46	L
8	23-Mar	26-Mar	265	215	120/80	59	56	L
9	17-Mar	20-Mar	290	245	120/70	62	65	P
10	26-Feb	26-Feb	159	106	120/80	55	59	L
11	03-Jun	07-Jun	340	284	130/70	58	49	L
12	05-Apr	10-Apr	290	221	120/80	55	63	P
13	17-Apr	21-Apr	232	212	120/80	64	54	L
14	01-Jan	04-Jan	310	254	130/70	70	62	L
15	21-Mar	25-Mar	255	198	120/70	40	50	P

Figure 2. Dataset of diabetes mellitus patients

2.8 Data Analysis Techniques

From the data collection described previously, not all data obtained can be processed for classification. Prior data analysis is necessary because some attributes are unnecessary and must go through various stages of data analysis to obtain quality data. There are several stages in data processing aimed at obtaining quality data. The techniques used for the data analysis process are as follows:

- 1) Data Selection. In the data selection stage, data is selected from the data cleaning stage. Because not all of the obtained dataset is used, several attributes containing information not included in the data processing requirements, such as entry date and exit date, will be removed from the dataset. The results of the data selection will be used in the next analysis stage.
- 2) Data Cleaning. The data cleaning stage corrects data by removing missing values. Data is considered missing if an attribute in the data is missing or empty (null), as well as incomplete data. The cleaned data is then cleaned using the Replace Missing Values process to fill in the average value of a particular attribute in each empty area that refers to the attribute.
- 3) Data Transformation. This stage is the process of transforming or normalizing data into a format that can be managed by the system. By normalizing using the WEKA tool and changing the initial data format according to the category because association analysis can only accept categorical data input, transformation on continuous data columns is carried out by a discretization process followed by a binarization process.

3. RESULTS AND DISCUSSION

In the testing phase, the diabetes dataset was divided using 10-fold cross-validation. The aim was to determine accuracy and analyze the naive Bayes and support vector machine algorithms based on accuracy and error results. The algorithms were then applied to the polynomial and RBF kernels to achieve the best accuracy level.

3.1 Naïve Bayes Classification Analysis

The classification analysis stage process for the Naïve Bayes algorithm with the following Confusion matrix results:

Table 2. Naïve Bayes confusion matrix results			
Class	Classification		Total
	Positive	Negative	
Positive	323	24	347
Negative	10	72	82
Total	333	96	429

Based on table 2 above, the results of the comparison on the positive status variable were obtained with a classification value of 323 cases, in the class that entered the negative status but in the classification entered the positive status as many as 10 cases. For the classification of negative status, the classification was correct as many as 72 cases, however, there were 24 cases for which the system misclassified. The total data for the positive status class was 347, while for the negative status there were 82. The number of classification data for positive status was 333 while for negative status there were 39. To determine the level of accuracy

and error in the classification of diabetes mellitus disease on training data with the naive Bayes algorithm, namely 92 and 7 percent.

3.2 Classification Analysis of Support Vector Machine Kernel Polynomial

Classification analysis for the support vector machine algorithm with a polynomial kernel with the following confusion matrix results:

Table 3. Confusion matrix of SVM with Polynomial kernel

Class	Classification		Total
	Positive	Negative	
Positive	345	2	347
Negative	14	68	82
Total	359	70	429

Based on table 3, the results of the analysis of the accuracy of the support vector machine algorithm with a polynomial kernel with positive classification results of 345 cases, for positive classification but not included in the negative class with a total of 14 cases, then negative classification but included in the positive class there are only 2 cases, then the case of correct negative classification is 68 cases. So the accuracy and error values are 96 and 3 percent.

3.3 Classification Analysis of Support Vector Machine Kernel RBF

Classification analysis for the support vector machine algorithm with the RBF kernel with the following confusion matrix results:

Table 4. Confusion matrix of SVM with RBF kernel

Class	Classification		Total
	Positive	Negative	
Positive	347	0	347
Negative	82	0	82
Total	429	0	429

Table 4 above shows the classification results using the support vector machine algorithm with the RBF kernel. For the positive status, 347 cases were correctly classified, 82 cases were classified as negative, 0 cases were incorrectly classified, and 0 cases were correctly classified as negative. The accuracy and error values are 80 and 9 percent, respectively.

3.4 Comparison of Naïve Bayes and Support Vector Machine classification results

A comparison of the results of these two algorithms was used to determine the method with the highest level of accuracy. The accuracy results obtained from both algorithms are shown in the following table:

Table 5. Results of Algorithm Comparison

	<i>Naïve Bayes</i>	<i>SVM Kernel Polynomial</i>	<i>SVM Kernel RBF</i>
Accuracy	92.0746%	96.2704%	80.8857%

For the comparison results between the naive Bayes algorithm and the support vector machine, it can be concluded by looking at table 5 that the value obtained from the results of the 10-fold cross validation test with accuracy results is the SVM algorithm with the Polynomial Kernel which has the highest accuracy value of 96.2704%.

4. CONCLUSION

Based on the discussion that has been described previously, the conclusion of the study on the comparison of the naïve bayes method and support vector machine for the classification of diabetes mellitus at RSI Siti Khadijah Palembang is as follows: By using training data and testing data, it is then measured to determine the level of accuracy by evaluating the confusion matrix against several best experiments. For the SVM algorithm using two kernels, namely the polynomial kernel and RBF, the accuracy value of the SVM algorithm with the Polynomial kernel has the highest accuracy value with a value of 96.2704%, it can be said to be more accurate when compared to the naïve bayes algorithm with an accuracy level of 92.0746%. So it

can be concluded that the most appropriate algorithm for the classification of diabetes mellitus at RSI Siti Khadijah Palembang is the support vector machine algorithm. From the study, it can be concluded why there are factors that cause one of the results to be different, namely the naïve bayes algorithm which is superior. This is because researchers use the stopword removal stage, which means there is a possibility of changing the content of the context which basically will have a different meaning or even in the opposite direction. Based on the analysis, it can also be concluded that manual calculations do not guarantee accuracy in data classification.

REFERENCES

- Andriyanto, I., & Santoso, E. (2017). Pemodelan Sistem Pakar Untuk Menentukan Penyakit Diabetes Mellitus Menggunakan Metode Naive Bayes Studi Kasus : Puskesmas Poncokusumo Malang. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, 2(2), 880–887.
- Arifin, O., & Sasongko, T. B. (2018). ANALISA PERBANDINGAN TINGKAT PERFORMANSI METODE SUPPORT VECTOR MACHINE DAN NAÏVE BAYES CLASSIFIER. 2016, 67–72.
- Dalam, C., Penyakit, P., Di, M., Sakit, R., & Siwi, K. (n.d.). PERBANDINGAN KINERJA METODE NAIVE BAYES DAN. 1–8.
- Nugroho, R. A., Tarno, & Prahutama, A. (2017). Klasifikasi Pasien Diabetes Mellitus Menggunakan Metode Smooth Support Vector Machine (Ssvm). *Gaussian*, 6, 439–448.
- Rahman, M. F., Alamsah, D., Darmawidjaja, M. I., & Nurma, I. (2017). Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN). *Jurnal Informatika*, 11(1), 36.<https://doi.org/10.26555/jifo.v11i1.a5452>
- Riyanto, U. (2018). ANALISIS PERBANDINGAN ALGORITMA NAIVE BAYES DAN SUPPORT VECTOR MACHINE DALAM MENGLASIFIKASIKAN JUMLAH PEMBACA. 62–72.
- Saleh, A., & Utama, U. P. (2016). Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. *June*.
- Sukmawati, Lawi, A., & Thamrin, S. A. (2013). Klasifikasi Penyakit Diabetes Melitus Tipe II Menggunakan Metode Support Vector Machine.
- Widyawati, & Sutanto. (2019). Perbandingan Algoritma Naïve Bayes Dan Support Vector Machine (Svm). 3(2), 178–194.