Journal of Autonomous Intelligence (2024) Volume 7 Issue 2
doi: 10.32629/jai.v7i2.888

ORIGINAL RESEARCH ARTICLE

Improvement of support vector machine for predicting diabetes mellitus with machine learning approach

Christine Dewi[1,], Jernius Zendrato[1], Henoch Juli Christanto[2,*]

[1] *Department of Information Technology, Satya Wacana Christian University, Salatiga 50711, Indonesia*

[2] *Department of Information System, Atma Jaya Catholic University of Indonesia, Jakarta 12930, Indonesia*

*** Corresponding author:** Henoch Juli Christanto, henoch.christanto@atmajaya.ac.id

ABSTRACT

The prevalence of diabetes is currently increasing worldwide, including in Indonesia, due to the increasing levels of stress and lack of physical activity that led to obesity and related complications such as hypertension. However, only about 25% of diabetes patients are aware of their condition. Therefore, this study aims to find an algorithm that can help predict with better accuracy using the diabetes mellitus dataset obtained from Kaggle. To obtain information about the accuracy level of diabetes diagnosis, the data will be processed using two methods, namely support vector machine and naive bayes. To obtain the most accurate results, we optimize each variant and parameter of every algorithm used. The best method in this study was produced by the support vector machine method with a radial basis function (RBF) kernel, which achieved an accuracy level of 98.25%, superior to the naive bayes method which obtained the highest accuracy of only 77.25%. Additionally, this study also applied the proposed method using the diabetes mellitus dataset from LAB01 DAT263x taken from the Kaggle website. The results of the experiment indicate that the suggested model outperforms other methods in terms of performance, with a tendency for high accuracy generated in every experiment for all datasets.

Keywords: support vector machine; naive bayes; diabetes mellitus

ARTICLE INFO

Received: 4 July 2023
Accepted: 19 September 2023
Available online: 13 December 2023

1. Introduction

Diabetes mellitus is a common name for a heterogeneous metabolic disorder, characterized by chronic hyperglycemia. Its cause is a disturbance in insulin secretion or insulin effect, or both[1]. Millions of individuals worldwide are affected by diabetes mellitus (DM), which is a chronic non-communicable disease. The prevalence of diabetes has risen across various socioeconomic groups due to heightened stress levels and a decline in physical activity. This ultimately results in obesity and associated complications, including hypertension and type II diabetes[2].

According to the international diabetes federation (IDF), the prevalence of diabetes among adults aged 20 to 79 has significantly risen since its initial publication in 2000. The estimated number of diabetes cases increased from approximately 151 million individuals (constituting 4.6% of the global population at that time) to 537 million in 2021 (representing 10.5% of the population). This signifies a more than threefold increase over the mentioned period[3]. Individuals with diabetes are at risk of experiencing various complications that can lead to disability or death[3].

According to the findings of the 2018 basic health research (Riskesdas), the prevalence of diabetes in Indonesia among individuals aged 15 years and above, as diagnosed by doctors, was recorded at 2%. If compared to the prevalence of diabetes in the population aged 15 years and above in the 2013 Riskesdas results, there has been an increase of 1.5%. From the results of blood sugar level tests, it can be seen that the prevalence of diabetes mellitus increased from 6.9% in 2013 to 8.5% in 2018. This data indicates that only about 25% of people with diabetes are aware that they have the condition[4]. Therefore, it is necessary to have an algorithm that can help predict whether someone has diabetes or not. To make accurate predictions about the likelihood of disease occurrence, the implementation of data mining algorithms is necessary. Data mining is a discipline that focuses on analyzing data to obtain additional information that is broader than the currently available information through the use of keywords or existing information[5].

This study will also compare data mining methods in predicting whether someone has diabetes or not. The data mining methods used are classification methods using the support vector machine and naïve bayes algorithms. In this research, we will present several main contributions. First, we will test the accuracy of the support vector machine and naive bayes methods in processing the diabetes mellitus dataset. Second, we will compare the accuracy level with previous research using the support vector machine and naive bayes methods. Third, we will adjust the most optimal parameters for each method used, such as gamma, C, var_smoothing, Alpha parameters. Fourth, we will compare the accuracy values of the Support Vector Machine and Naive bayes methods from the conducted experiments.

This research consists of three main parts. The materials and methods section will include related work and the methodology that we plan to apply in this study. The results and discussion section will describe the settings and results of the experiments that we have conducted. Finally, in the conclusion section, we will provide conclusions and suggestions for further research.

2. Related work

In recent years, researchers have been developing various recommendation algorithms based on provided text data. Alghamdi et al.[6] investigated the relative performance of different machine learning methods, developed an ensemble-based predictive model, and utilized the Synthetic Minority Oversampling Technique (SMOTE) approach to predict the occurrence of diabetes using cardiorespiratory fitness medical records. The research by Poonia et al.[7] is titled, “Intelligent diagnostic prediction and classification models for detection of kidney disease”. The study employs predictive analysis based on machine learning to detect kidney disease at an early stage. This research provides a feature-based predictive model for kidney disease detection. Various machine learning algorithms are utilized, including k-nearest neighbors (KNN), artificial neural networks (ANN), support vector machine (SVM), naive bayes (NB), and others. The use of recursive feature elimination (RFE) and the chi-square feature selection technique is necessary to construct and analyze various predictive models on a dataset comprising healthy and kidney disease patients. In another study, Zou et al.[8] titled, “Predicting diabetes mellitus using machine learning”, decision trees, random forests, and neural networks were used to predict diabetes mellitus. The dataset used consisted of hospital physical examination data from Luzhou, China, with 14 attributes. A five-fold cross-validation technique was used to test the models. Similarly, with Vigneswari et al.[9] applied a machine learning classifier to predict patient diseases and evaluated the performance of a decision tree classifier in predicting diabetes mellitus (DM). This analysis was based on accuracy and true positive rate (TPR). The research titled, “Predicting the risk of incident type two diabetes mellitus in Chinese elderly using machine learning techniques” developed by Liu et al.[10] aims to construct an effective predictive model based on machine learning (ML) for the risk of type two diabetes mellitus (T2DM) among the elderly population in China. This study utilizes health examination data from adults aged over 65 in Wuhan, China, spanning from 2018 to 2020. Four ML algorithms are employed in this research: Logistic regression (LR), decision tree (DT), random forest (RF), and extreme gradient boosting (XGBoost). Performance evaluation of the models will be conducted based on the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and accuracy. Maulidah et al.[11] conducted a study titled, "Prediction of diabetes mellitus using support vector machine and naive bayes methods". The study was developed through processing secondary health database data using the support vector machine and naive bayes methods to determine the accuracy of diabetes diagnosis. Based on several conducted studies, they can serve as a reference for developing improved accuracy in predicting the diagnosis of diabetes. The research conducted by Faruque et al.[12] is focused on, "Predicting diabetes mellitus and analysing risk-factors correlation". This study aims to explore various risk factors such as kidney complications, blood pressure, hearing impairments, and skin complications related to this disease using machine learning techniques and decision-making. It employs four popular machine learning algorithms, namely support vector machine (SVM), naive bayes (NB), k-nearest neighbor (KNN), and C4.5 decision tree (DT). The data used consists of diagnostic medical data from 200 diabetic patients at the Medical Center Chittagong, Bangladesh, comprising 16 attributes. Mushtaq et al.[13] conducted a study on, "Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques". The dataset used was obtained from an online repository. In the initial stage of this research, logistic regression, Support Vector Machine, k-nearest neighbors, gradient boosting, naive bayes, and random forests algorithms were applied to assess the prediction efficiency based on patient preconditions. Subsequently, a voting algorithm was employed with the three best-performing algorithms.

3. Proposed work

3.1. Data collection

The information gathering process will involve using the diabetes mellitus database obtained from Kaggle, which originates from a hospital in Frankfurt, Germany[14], and also the diabetes mellitus dataset from LAB01 DAT263x taken from the Kaggle website[15]. The dataset comprises 2000 and 15,000 records, respectively, with various medical predictor variables or attributes such as pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome. Subsequently, the data will be processed using Python tools.

3.2. Data preprocessing

Preprocessing plays a vital role in machine learning as it is an essential step. In other words, this step aims to convert raw data into a more understandable and usable format. Data sets often contain errors or imperfections, so this step can help address these issues and facilitate data processing[16]. To ensure the validity of the data, this study will also perform several preprocessing actions on irrelevant or unused data.

Data Separation is part of the preprocessing stage. In this step, the dataset will be split into two sections: the training dataset and the testing dataset. The training dataset will be utilized to construct the model and will still include labeled data. Meanwhile, the testing dataset is used to validate the model to assess the accuracy of the algorithm. In the testing dataset, label data will be removed and separated as the actual target value[17]. In this research, the data was divided into 80% for the training dataset and 20% for the testing dataset.

Next in the preprocessing process, there is a normalization stage which is a good method for reducing data differences and increasing efficiency. In cases of very large data, normalization methods must have simple rules and be able to run quickly[18]. Fortunately, in this study, data normalization as one of the preprocessing steps is not necessary because the dataset used is already appropriate and does not require further adjustments.

3.3. Applied data analytics methods

3.3.1. Support vector machine

The SVM method has a strong ability to build classifications[19]. An overview can be found in some study[20–22] and is also considered an extension of maximum margin classification[23]. Support vector machine (SVM) is an algorithm that leverages examples to acquire knowledge on how to assign labels to objects[24]. SVM is a highly effective discriminative classification method formally characterized by an optimal hyperplane. The optimal hyperplane results in classifications for new examples and the dataset that supports the hyperplane is referred to as the support vector[25]. The hyperplane is adjusted to ensure it has the maximum distance from the closest data points of each class. The closest data points are referred to as support vectors. This applies to the training dataset with Equation (1)[19].

$$(x_1, y_1), \ldots, (x_n, y_n), x_i \in R^d \text{ end } y_i \in (-1, +1) \tag{1}$$

It can be seen that x_i represents the feature vector representation, while y_i corresponds to the class label, which can be either positive or negative, of the i training compound. Therefore, Equation (2) can be employed to define the optimal hyperplane.

$$wx^T + b = 0 \tag{2}$$

The weight vector is represented as w, where x denotes the input feature vector. The bias is denoted as b. Both the weight vector, w, and the bias, b, must satisfy the following inequality for all elements in the training set as per Equation (3).

$$wx_i^T + b \geq +1 \; if \; y_i = 1; \; wx_i^T + b \leq -1 \text{ if } y_i = -1 \tag{3}$$

The purpose of training an SVM model is to determine the values of w and b that allow the hyperplane to effectively separate the data with a maximum margin, which is defined by Equation (3).

$$\frac{1}{\|w\|^2} \tag{4}$$

Support vector is a term used for x_i vectors that satisfy the equation $|y_i| \left(wx_i^T + b\right) = 1$, as shown in **Figure 1**.

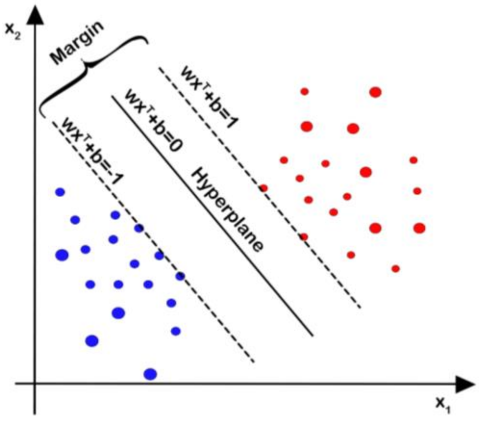


Figure 1. A linear SVM model is used to classify two classes, namely red and blue.

Moreover, one alternative technique currently used is the support vector machine (SVM) nonlinear, which uses kernel functions to classify a set of data with the aim of finding the optimal hyperplane in a high-dimensional feature space[26] such as in (**Figure 2**). Kernel is a method used to solve nonlinear problems by utilizing linear classification and involving the transformation of data that cannot be linearly separated[25]. By using $K(x_n, x_i)$, the original data undergoes transformation into a higher-dimensional space. In this process, a transformation function is applied to the dot product of $\emptyset(x)$, as depicted in Equation (5)[27].

$$K(x_n, x_i) = \emptyset(x_n)\emptyset(x_i) \tag{5}$$

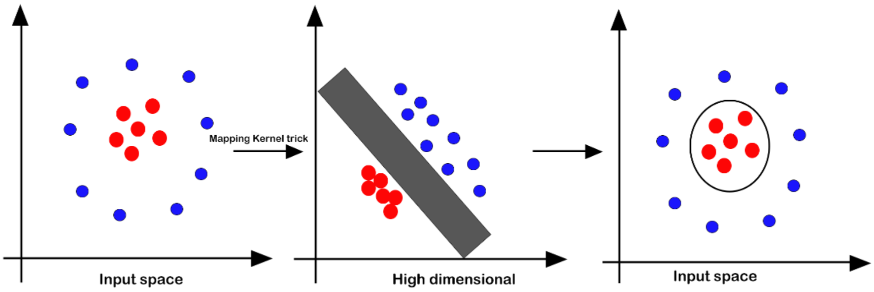


Figure 2. A kernel function is used to transform and separate data that cannot be separated by a linear SVM.

In SVM algorithm, there are several common types of kernel functions such as linear, radial basis function (RBF), sigmoid, and polynomial listed in **Table 1**. Each kernel function has specific parameters that need to be optimized to achieve the best performance[27].

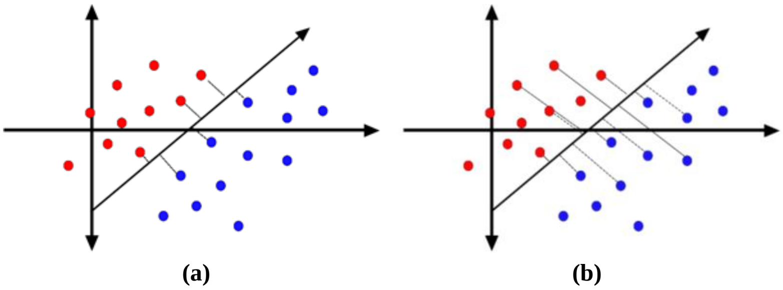
Table 1. Four common kernels.

No	Kernel function	Formula
1	Linear	$K(x_n, x_i) = (x_n, x_i)$
2	RBF	$K(x_n, x_i) = \exp(-\gamma\|x_n - x_i\|^2 + C)$
3	Sigmoid	$K(x_n, x_i) = \tanh(\gamma(x_n, x_i) + r)$
4	Polynomial	$K(x_n, x_i) = (\gamma(x_n, x_i) + r)^d$

C: cost; γ: gamma; r: coefficient; d: degree.

3.3.2. Regularization parameter SVM

Regularization parameter C is used to determine the magnitude of penalty for errors. This impacts the balance between the smoothness of the decision boundary and the capability to accurately classify the training data[28]. If the value of C is high, the training data will be accurately classified based on the hyperplane; conversely, if the value of C is low, the optimization will seek a higher margin that separates the hyperplane[25].



(a) **(b)**

Figure 3. (a) high gamma; **(b)** low gamma.

One of the factors that affects the performance of SVM classification is gamma, which is included in the sample subspace with complex changes[29]. A high value of gamma (**Figure 3a**) gives more weight to data points close to the decision boundary. Conversely, a low value of gamma (**Figure 3b**) takes into account data points far from the decision boundary in the computation of the decision boundary[25].

3.3.3. Naïve bayes

Naive bayes classifier is a technique in text mining that is useful for handling problems in opinion mining by categorizing into two categories, positive and negative opinions. Therefore, naive bayes classifier is effective as a method for classifying text[30].

The naive bayes classifier is an approach that utilizes Bayes' theorem, which is a simple probability-based classification method that assumes that each attribute in the data is independent[31]. In classifying data using naive bayes, it is represented by a set of attributes, “$x_1, x_2, \dots, x_n$”. The Equation (6) can be used to express the probability model of each class k.

$$P(y_k|x_1, x_2, \dots, x_n) \tag{6}$$

Further, n represents the count of attributes, while k represents the number of classes present in the set of class y data. In the probability perspective, classification is described as a Bayes rule which can be written according to Equation (7):

$$P(y_k|x_i) = \frac{P(y_k).P(x_i|y_k)}{P(x_i)} \tag{7}$$

In the following formula, $P(y_k|x_i)$ Denotes the likelihood of event y_k happening in the presence of event x_i, $P(x_i|y_k)$ represents the probability of the event x_n occurring when y_k occurs, $P(y_k)$ represents the probability of the event y_k, and $P(x_i)$ represents the probability of the event x_i. To find the highest probability value for each class that can be selected as the optimal class, Equation (8) can be used:

$$\arg\max_{y_k \in y} = \frac{P(y_k).P(x_i|y_k)}{P(x_i)} \tag{8}$$

Equation (9) is derived by assuming a constant value for $P(x_i)$ across all classes.

$$\arg\max_{y_k \in y} = P(y_k).P(x_i|y_k) \tag{9}$$

3.4. Model evaluation

There are numerous metrics to evaluate text processing and information retrieval systems. The performance of systems that classify documents into categories can be measured using various measures, such as precision, recall, and macro-average[32]. The definitions of precision and recall can be found in **Table 2**.

Table 2. Definitions of FN, FP, TN, and TP.

	Negative (N)	**Positive (P)**
False (F)	FN predicted result: N actual result: P	FP predicted result: P actual result: N
True (T)	TN predicted result: N actual result: N	TP predicted result: P actual result: P

TP stands for the number of relevant documents classified by humans and classifiers, FN is the number of documents deemed relevant by humans but not classified as relevant by the classifier, FP is the number of documents deemed irrelevant by humans but classified as relevant by the classifier, and TN is the number of documents deemed irrelevant by humans and the classifier[32].

Precision pertains to the proportion of correct positive predictions relative to the overall positive predictions[33]. Precision is determined by utilizing the provided Equation (10) to calculate its value:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

Recall, which is also referred to as the true positive rate or sensitivity, refers to the proportion of correct (positive) predictions out of the total actual positive samples. The Equation (11) can be used to calculate the recall.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

As a result, The F1 score, derived from the harmonic mean of precision and recall, can be computed using the specific Equation (12).

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \quad (12)$$

4. Result analysis

At this phase, the results of the conducted experiments will be explained. This research tests and evaluates the performance of an algorithm for determining diabetes mellitus patients using a dataset from a hospital in Frankfurt, Germany. The dataset used contains 2000 data points, with various medical predictor variables or attributes such as pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome. The description of the datasets used in these experiments can be seen in **Table 3**. Further, **Tables 4** and **5** provide an illustration of Dataset 1 that was utilized as an example. **Figure 4** shows the number of data points for diabetes and non-diabetes cases (Dataset 1).

Table 3. Dataset descriptions.

NO	Name	Dataset	Number of data	Feature
1	Dataset 1	Diabetes mellitus dataset from a hospital in Frankfurt, Germany[14].	2000	8
2	Dataset 2	Diabetes mellitus dataset from LAB01 DAT263x[15].	15,000	8

Table 4. Dataset of diabetes mellitus from a hospital in Frankfurt, Germany (Dataset 1).

Pregnancy	Glucose	Blood pressure	Skin thickness	Insulin	BMI	Diabetes pedigree function	Age	Outcome
2	138	62	35	0	33.6	0.127	47	2
0	84	82	31	125	38.2	0.233	23	0
0	145	0	0	0	44.2	0.630	31	0
0	135	68	42	250	42.3	0.365	24	0
1	139	62	41	480	40.7	0.536	21	1

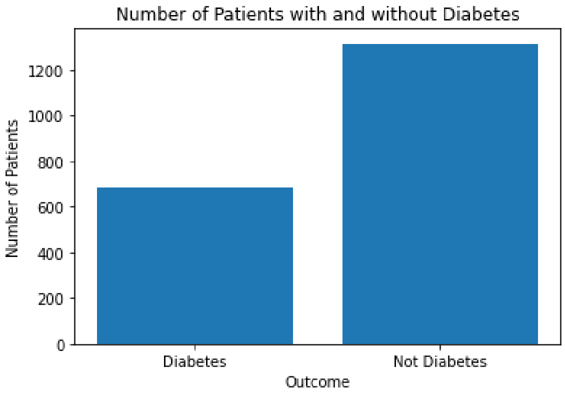


Figure 4. The number of data points for diabetes and non-diabetes cases (Dataset 1).

In the initial stage of the experiment, a study was conducted to test the accuracy of methods such as Support Vector Machine and Naive Bayes in processing the diabetes mellitus dataset obtained from a hospital in Frankfurt, Germany. The results of the research in **Table 5** show the accuracy test results of the Support vector machine method. Experiments were conducted using several types of kernels in support vector machine, such as linear kernel, polynomial (poly) kernel, and radial basis function (RBF) kernel. From the experimental results obtained from the testing dataset, the highest accuracy of 83.50% was achieved when using the radial basis function (RBF) kernel.

Meanwhile, the accuracy test results of the naive bayes method with Dataset 1 presented in **Table 6** were also conducted with several variations, such as naive bayes gaussian, naive bayes multinomial, and naive bayes Bernoulli. From the experiment results, it was found that the highest accuracy of 77.00% was obtained when using the naive bayes gaussian variation. To achieve optimal accuracy levels, adjustments of the most optimal parameters are required for each method. For support vector machine, the parameters to be adjusted are *C* and gamma for radial basis function (RBF) kernel, while for Naive Bayes method, the parameters to be adjusted are var_smoothing for naive bayes gaussian variation and alpha parameter for naive bayes multinomial.

Table 5. Results of support vector machine method experimentation with Dataset 1.

Support vector machine	Accuracy (random_state = 0)
Kernel = Linear	77.75%
Kernel = poly	77.00%
Kernel = RBF	78.50%
Kernel = RBF, gamma = 0.001	83.50%

Table 6. Results of naive bayes method experimentation with Dataset 1.

Naive bayes	Accuracy (random_state = 0)
naive_bayes = GaussianNB ()	75.75%
naive_bayes = GaussianNB (var_smoothing = 0.001)	77.00%
naive_bayes = BernoulliNB ()	66.25%
naive_bayes = MultinomialNB ()	60.25%

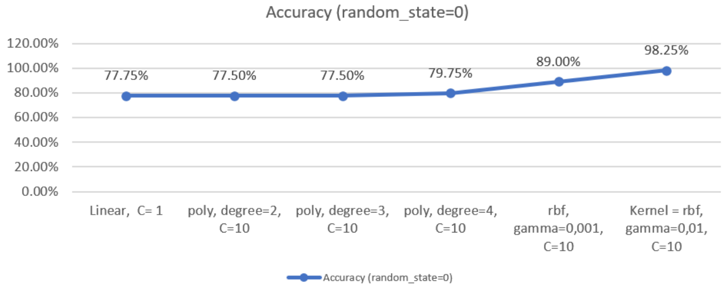


Figure 5. The results of experiments using support vector machine method with several kernels.

In **Figure 5**, an experiment was conducted using the support vector machine method with regularization parameter *C* and parameters that affect training examples such as degree and gamma. Initially, the experiment was conducted using a linear kernel with *C* = 1, which resulted in an accuracy of 77.75%. However, after trying to use a polynomial kernel with different degree values, it was found that degree = 3 and *C* = 10 resulted in the highest accuracy of 79.75%. Next, the experiment was conducted using the radial basis function (RBF) kernel and various gamma parameters. As a result, it was found that gamma = 0.01 and *C* = 10 resulted in the highest accuracy of 98.25%. **Table 7** shows that the support vector machine method with RBF kernel and gamma = 0.01 resulted in the highest accuracy.

Table 7. The classification report results of support vector machine with *C* = 10 and gamma = 0.01.

Items	Precision	Recall	F1-score	Support
0	0.98	1.00	0.99	272
1	0.99	0.95	0.97	128
Accuracy	-	-	0.98	400
Macro avg	0.99	0.97	0.98	400
Weighted avg	0.98	0.98	0.98	400

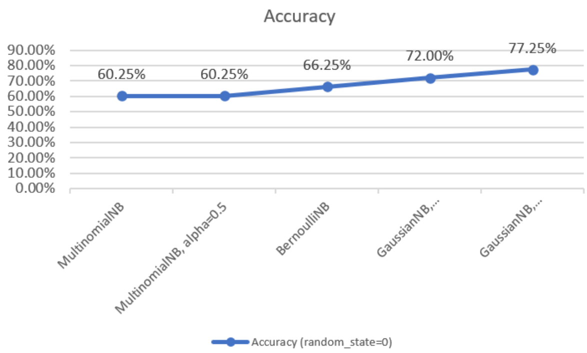


Figure 6. The results of the naive bayes method experiment with Dataset 1 and several variations.

In **Figure 6**, an experiment was conducted using various parameters in the naive bayes classification method. The experiment results using multinomial naive bayes with various values of alpha parameter showed an accuracy of 60.25%. Furthermore, an experiment was conducted using Bernoulli Naive Bayes variation, which yielded an accuracy of 66.25%. Then, an experiment was conducted using gaussian naive bayes with several values of var_smoothing parameter, and the highest accuracy found was 77.25%. **Table 8** shows that the naive bayes classification method with gaussian variation and var_smoothing parameter value yields the highest accuracy of 77.25%.

Table 8. Classification report results of naive_bayes with GaussianNB (var_smoothing = 0.01).

Items	**Precision**	**Recall**	**F1-score**	**Support**
0	0.79	0.91	0.84	272
1	0.71	0.48	0.58	128
Accuracy	-	-	0.77	400
Macro avg	0.75	0.70	0.71	400
Weighted avg	0.76	0.77	0.76	400

Therefore, the conclusion drawn is that the use of the support vector machine method with an accuracy rate of 98.25% is superior to the application of the naive bayes method. This makes the support vector machine with the radial basis function (RBF) kernel the best method in this research. Furthermore, this research will involve comparing the accuracy rates with the previous study using the support vector machine and naive bayes methods.

Table 9. Comparison of experimental results with the previous study.

Algorithm	**Dataset**	**Accuracy results**
Support vector machine[11]	Dataset 1	78.04%
Naive bayes[11]	Dataset 1	76.98%
Proposes method		
Support vector machine with *C* = 10 and gamma = 0.01.	Dataset 1	98.25%
Naive bayes with Gaussian NB (var_smoothing = 0.01).	Dataset 1	77.25%

The previous research conducted by Maulidah et al. can be seen in **Table 9**. By using the support vector machine (SVM) and naive bayes methods, they achieved the highest accuracy of 78.04% using support vector machine[11]. In this study, we employ support vector machine with *C* = 10 and gamma = 0.01 and naive bayes with Gaussian NB (var_smoothing = 0.01). However, our proposed method successfully achieved the highest accuracy of 98.25% using SVM. This success was achieved through our research, which focused on parameter optimization, aimed at attaining a higher level of accuracy. Careful parameter selection has made this study more optimal and accurate compared to previous research. The next step is to perform experiments on several different datasets. These experiments are intended to ensure that the model developed using support vector machine with radial basis function (RBF) can achieve high levels of accuracy even when tested on different sets of data.

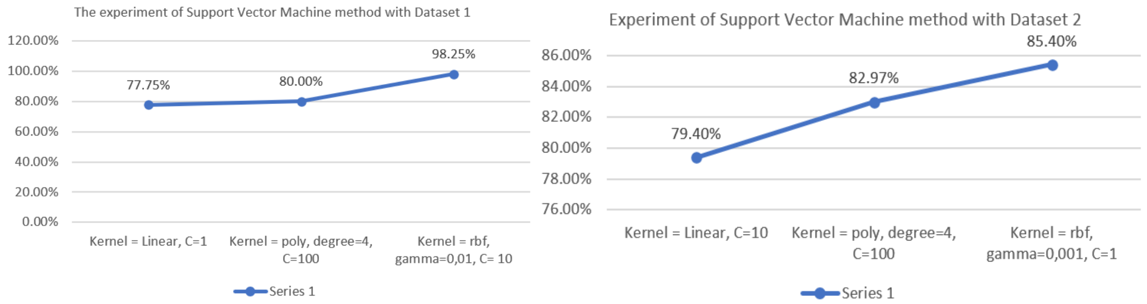
Table 10. The result of the support vector machine method experiments with the Dataset 2.

Support vector machine	**Accuracy (random state = 0)**
Kernel = Linear, C = 10	79.40%
Kernel = Poly, degree = 4, C = 100	82.97%
Kernel = RBF, gamma = 0.001, C = 1	85.40%

Table 11. The results of the naive bayes method experiment with Dataset 2.

Naive bayes	**Accuracy (random state = 0)**
Naive bayes = Multinomial NB ()	62.30%
Naive bayes = Bernoulli NB ()	66.73%
Naive bayes = Gaussian NB (var_smoothing = 0.00001)	79.53%

The results of the experiments using the support vector machine method on the diabetes mellitus dataset from LAB01 DAT263x are recorded in **Table 10**, while the results using the naive bayes method are recorded in **Table 11**. From the results obtained from the testing dataset, it can be concluded that the support vector machine method with the radial basis function (RBF) kernel is still the best compared to other methods. This method is able to provide higher accuracy compared to others, which is 85.40%. In this study, higher accuracy values generally indicate better performance, while lower accuracy values indicate less satisfactory performance. If the highest accuracy percentage from multiple experimental models is found, then that experiment is the best classifier[34]. The research findings show a trend of increasing accuracy values in each experiment performed for all datasets. Based on the evaluation results presented in **Figure 7**, it can be concluded that the proposed model using support vector machine with radial basis function (RBF) kernel has better performance compared to other methods on each dataset. The Support Vector Machine method with radial basis function (RBF) kernel can produce the best accuracy in the classification performed. Further, we upload our experiment in GitHub link https://github.com/jerniusZendrato/penelitian.git.



(a) **(b)**

Figure 7. (a) Dataset 1; **(b)** Dataset 2.

5. Conclusions

This research aims to improve the accuracy level of the previous study by exploring various methods and parameters. In the experiments conducted, it was found that the support vector machine (SVM) method performed better compared to naive bayes. To further improve the accuracy of both methods, the next experiments were conducted by varying parameters such as regularization parameter C, training parameters such as degree and gamma on SVM, and various parameters in naive bayes classification method such as alpha and var_smoothing. From the experiment results in **Figures 5–7**, it was found that SVM method with radial basis function (RBF) kernel could achieve higher accuracy. Compared to other kernel functions, the use of SVM with RBF kernel could increase the accuracy from 77.75% to 98.25% and outperform the highest accuracy in naive bayes method which was 77.25%. This indicates that adjusting the method used is crucial to optimize the resulting accuracy. The experiment results also prove the feasibility and accuracy of the proposed algorithm. With the high accuracy achieved by SVM with the RBF kernel, there is significant potential for its future application in medical pattern recognition systems. This can be utilized to support doctors in diagnosing diseases, not limited to diabetes but also including diseases like cancer and heart disease, with greater precision and speed, thereby enhancing patient care. In future work, it is necessary to consider comparing different methods and updating the dataset to improve the quality of the research results. Besides, we will implement the SHAP to describe the importance feature.

Author contributions

Conceptualization, CD, and JZ; methodology, JZ; software, HJC; validation, HJC, CD and JZ; formal analysis, CD; investigation, CD; resources, JZ; data curation, JZ; writing—original draft preparation, CD and JZ; writing—review and editing, CD, JZ, and HJC; visualization, JZ; supervision, CD and HJC; project administration, HJC; funding acquisition, CD. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Petersmann A, Müller-Wieland D, Müller UA, et al. Definition, classification and diagnosis of diabetes mellitus. *Experimental and Clinical Endocrinology and Diabetes* 2019; 127(S01): S1–S7. doi: 10.1055/a-1018-9078
2. John JE, John NA. Imminent risk of COVID-19 in diabetes mellitus and undiagnosed diabetes mellitus patients. *Pan African Medical Journal* 2020; 36. doi: 10.11604/pamj.2020.36.158.24011
3. Federation D. *IDF Diabetes Atlas Tenth Edition 2021*. International Diabetes Federation; 2021.
4. Kemenkes RI. Information data center ministry of health 2020 diabetes mellitus (Indonesian). *Kementrian Kesehatan RI* 2020; 15(2).
5. Tiwari AK, Ramakrishna G, Sharma KL, Kashyap SK. Academic performance prediction algorithm based on fuzzy data mining. *IAES International Journal of Artificial Intelligence* 2019; 8(1): 26. doi: 10.11591/ijai.v8.i1.pp26-32
6. Alghamdi M, Al-Mallah M, Keteyian S, et al. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The henry ford exercise testing (FIT) project. *PLoS One* 2017; 12(7): e0179805. doi: 10.1371/journal.pone.0179805
7. Poonia RC, Gupta MK, Abunadi I, et al. Intelligent diagnostic prediction and classification models for detection of kidney disease. *Healthcare* 2022; 10(2): 371. doi: 10.3390/healthcare10020371
8. Zou Q, Qu K, Luo Y, et al. Predicting diabetes mellitus with machine learning techniques. *Front Genet* 2018; 9. doi: 10.3389/fgene.2018.00515
9. Vigneswari D, Kumar NK, Ganesh Raj V, et al. Machine learning tree classifiers in predicting diabetes mellitus. In: Proceedings of the 5th International Conference on Advanced Computing and Communication Systems (ICACCS 2019); 15–16 March 2019; Coimbatore, India.
10. Liu Q, Zhang M, He Y, et al. Predicting the risk of incident type 2 diabetes mellitus in Chinese elderly using machine learning techniques. *Journal of Personalized Medicine* 2022; 12(6): 905. doi: 10.3390/jpm12060905
11. Maulidah N, Supriyadi R, Utami DY, et al. Prediction of diabetes mellitus using support vector machine and naive bayes methods (Indonesian). *Indonesian Journal on Software Engineering (IJSE)* 2021; 7(1): 63–68. doi: 10.31294/ijse.v7i1.10279
12. Faruque MF, Asaduzzaman A, Hossain SMM, et al. Predicting diabetes mellitus and analysing risk-factors correlation. *EAI Endorsed Trans Pervasive Health Technol* 2020; 5(20): 164173. doi: 10.4108/eai.13-7-2018.164173
13. Mushtaq Z, Ramzan MF, Ali S, et al. Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques. *Mobile Information Systems* 2022; 2022: 1–16. doi: 10.1155/2022/6521532
14. Diabetes. Available online: https://www.kaggle.com/datasets/johndasilva/diabetes (accessed on 20 April 2023).

15. Diabetes from DAT263x Lab01. Available online: https://www.kaggle.com/datasets/fmendes/diabetes-from-dat263x-lab01 (accessed on 20 April 2023).
16. Ghorbani R, Ghousi R. Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access* 2020; 8: 67899–67911. doi: 10.1109/access.2020.2986809
17. Anggoro DA, Supriyanti W. Improving accuracy by applying Z-score normalization in linear regression and polynomial regression model for real estate data. *International Journal of Emerging Trends in Engineering Research* 2019; 7(11). doi: 10.30534/ijeter/2019/247112019
18. Li W, Liu Z. A method of SVM with normalization in intrusion detection. *Procedia Environmental Sciences* 2011; 11: 256–262. doi: 10.1016/j.proenv.2011.12.040
19. Huang S, Cai N, Pacheco PP, et al. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics and Proteomics* 2018; 15(1): 41–51. doi: 10.21873/cgp.20063
20. Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press; 2000.
21. Schölkopf B, Smola AJ. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press; 2001.
22. González C, Mira-McWilliams J, Juárez I. Important variable assessment and electricity price forecasting based on regression tree models: Classification and regression trees, bagging and random forests. *IET Generation, Transmission and Distribution* 2015; 9(11): 1120–1128. doi: 10.1049/iet-gtd.2014.0655
23. Golpour P, Ghayour-Mobarhan M, Saki A, et al. Comparison of support vector machine, naïve bayes and logistic regression for assessing the necessity for coronary angiography. *International Journal of Environmental Research and Public Health* 2020; 17(18): 6449. doi: 10.3390/ijerph17186449
24. Noble WS. What is a support vector machine? *Nature Biotechnology* 2006; 24(12): 1565–1567. doi: 10.1038/nbt1206-1565
25. Battineni G, Chintalapudi N, Amenta F. Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). *Informatics in Medicine Unlocked* 2019; 16: 100200. doi: 10.1016/j.imu.2019.100200
26. Cheng H, Tan PN, Jin R. Efficient algorithm for localized support vector machine. *IEEE Transactions on Knowledge and Data Engineering* 2010; 22(4): 537–549. doi: 10.1109/tkde.2009.116
27. Achirul Nanda M, Boro Seminar K, Nandika D, Maddu A. A comparison study of kernel functions in the support vector machine and its application for termite detection. *Information* 2018; 9(1): 5. doi: 10.3390/info9010005
28. Kamble M, Shrivastava P, Jain M. Digitized spiral drawing classification for Parkinson's disease diagnosis. *Measurement: Sensors* 2021; 16: 100047. doi: 10.1016/j.measen.2021.100047
29. Wu Y, Lu Y. An intelligent machine vision system for detecting surface defects on packing boxes based on support vector machine. *Measurement and Control* 2019; 52(7–8): 1102–1110. doi: 10.1177/0020294019858175
30. Sunarya POA, Refianti R, Benny A, Octaviani W. Comparison of accuracy between convolutional neural networks and naïve bayes classifiers in sentiment analysis on twitter. *International Journal of Advanced Computer Science and Applications* 2019;10(5): 77–86. doi: 10.14569/ijacsa.2019.0100511
31. Malani R, Putra ABW, Rifani M. Implementation of the naive bayes classifier method for potential network port selection. *International Journal of Computer Network and Information Security* 2020; 12(2): 32–40. doi: 10.5815/ijcnis.2020.02.04
32. Rezaeian N, Novikova G. Persian text classification using naive bayes algorithms and support vector machine algorithm. *Indonesian Journal of Electrical Engineering and Informatics* 2020; 8(1): 178–188. doi: 10.11591/ijeei.v8i1.1696
33. Guo J, Wan B, Wu H, et al. A virtual reality and online learning immersion experience evaluation model based on SVM and wearable recordings. *Electronics* 2022; 11(9): 1429. doi: 10.3390/electronics11091429
34. Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data* 2020; 7(1): 1–26. doi: 10.1186/s40537-020-00327-4