

COMPARATIVE PERFORMANCE ANALYSIS OF SUPPORT VECTOR MACHINE AND RANDOM FOREST ON DIABETES PATIENT DATA FROM HOSPITALS IN THE UNITED STATES

¹Darmstater Albertus Albert D., ²Hironimus Leong

^{1,2}Program Studi Teknik Informatika Fakultas Ilmu Komputer,
Universitas Katolik Soegijapranata
²marlon.leong@unika.ac.id

ABSTRACT

The era of technological advancement at this time has begun to help a lot in many job sections, especially in the medical section. Especially in the development of Machine Learning which has a significant impact, the model built can help to predict the patient's disease from the symptoms and tests performed. Therefore, high accuracy and a short time are required for the machine-learning model to be built. The author build a model using the Random Forest algorithm and SVM algorithm, then compare these two models. What is compared between these two models is the computation time required by each algorithm and the level of accuracy, precision, recall, and F1-Score with stepwise data usage. The result to be achieved is that one of the algorithms produces stable and maximum results with the existing data. Among the eight experiments, SVM showed better performance in experiments 1, 3, 4, and 6, while random forest showed better performance in experiments 1, 2, 3, 4, and 6. The sixth experiment yielded the highest accuracy on both the minimum and maximum datasets. Here, SVM achieved 99.992 curacy in 142.0238 seconds and Random Forest achieved 99.982 curacy in 8.045849 seconds. Random Forest was 133.977951 seconds faster, but had a slightly lower accuracy of 0.01%.

Keyword: *Random Forest, Support Vector Machine, Machine Learning, Diabetes Disease*

BACKGROUND

According to Sudianto[1], the random forest algorithm is a better algorithm than Support Vector Machine for classifying Twitter data, and. In contrast with C-45, Naive Bayes, Random Forest research, did not use any datasets that could guarantee the consistency of the datasets used for the analysis. Azhari [2] where also compared the accuracy of several algorithm algorithms. In his research, he concluded that SVM was the more accurate algorithm than his other three algorithms. He utilized a small amount of data for the experiment. The authors conclude that the performance issues of the two algorithms are not maximal, as different numbers of datasets used for classification lead to different conclusions.

Based on the different results in his two cited journals, the researchers plan to conduct an intensive study to compare the performance of Random Forest and his SVM algorithm. This study uses datasets from the same source, increasing the amount of data used. This study, therefore, aims to investigate and compare the accuracy of the two algorithms as the number and diversity of datasets used increase.

By comparing the computation time and accuracy of the two algorithms, researchers can determine which algorithm processes data more effectively and produces more accurate predictive models. The study also modifies the dataset used to improve the accuracy rate. It is important to note that the performance of both algorithms can be tested on both small and large datasets, so the results of this study can provide more comprehensive information.

LITERATURE STUDY

Saputro [3] conducted research to predict the closing price of gold using the Support Vector Machine algorithm method to compare variable A (open, high, low, and close) with variable B (open, high, low, close, and factory news), which improves performance by maximizing parameters. By using data obtained from PT Rifan Financindo Futures, the author processes it with preprocessing techniques to compare variable A (open, high, low, and close) with variable B (open, high, low, close, and factory news), which improves performance by maximizing parameters. By using data obtained from PT Rifan Financindo Berjangka, the author processes it first with preprocessing techniques. The research conducted by the author with the amount of gold price data input shows that the use of the support vector machine algorithm by maximizing the parameter value for predicting the closing price of gold can get a pretty good value. When testing 10% of the data, the dataset using open, high, low, and close variables with the SVM algorithm and optimization of kernel type and parameter C (cost) resulted in an RMSE of 4.695. However, the dataset, which also included the Factory News variable using the same SVM algorithm and parameter optimization, gave an RMSE of 4.620. Therefore, it can be concluded that adding the factory news variable to the dataset gives a better RMSE improvement.

Dwiasnati and Devianto [4] conducted the same research using the SVM algorithm where they analyzed flood-prone areas through a data mining approach to find out which areas were included in flood-prone areas in the Bandung Regency area. And they tried to model flood-prone areas and facilitate the delivery of information to the surrounding community in Bandung Regency, which is included in the flood and non-flood zones. And based on the results of the research they did, they concluded that the accuracy level produced by the algorithm they used, namely SVM, was 85.71 n AUC, giving 0.841, while the accuracy level produced by the PSM-based SVM algorithm was 97.62 n AUC, giving 1,000. They have not used Weight Selection other than feature Weight Selection by correlation in their research.

Supriyanto et al. [5] conducted joint research on applying data mining techniques using Linear Regression and Random Forest algorithms in predicting palm oil prices, which will provide useful information for commodity export decision-making. They used a quantitative approach, where they used three scenarios of data sharing: 90:10, 80:20, and 70:30. They evaluated the two algorithms used and determined the best algorithm to use. In the 90:10 data division scenario, the best algorithm they concluded was Random Forest with an RMSE of 25,106, in the second scenario with data division 80:20, the best algorithm they concluded was Linear Regression with an RMSE of 31,174, in the third scenario with data division 70:30 Linear regression has the best results, with

an RMSE of 30,227. And of the three scenarios, the Linear Regression algorithm gets the best performance.

Inversely proportional to the conclusions of Fachid and Triayudi [6] conducted research to find out how much data was affected, died, and recovered from COVID-19 and how to analyze linear regression and random forest regression calculations. The data used was obtained through the website kawalkovid19.co.id. The data set they use is from the Indonesian COVID-19 Monitoring Agency, in this case, the data selection starts from January 1, 2021, to December 19, 2021. The results of their research provide an RMSE value of 3031.127 MAPE 47.66 and an accuracy of 94% in the linear regression algorithm, while the random forest algorithm provides an RMSE value of 1886.555 MAPE 14.85 and a resulting accuracy of 97%. From this, they can conclude that random forest regression is better to use than linear regression in this study in contrast to the research from Supriyanto et al. [5].

However, the accuracy of the random forest algorithm was proven and they concluded that a good algorithm for classification data characteristics is the Random Forest algorithm using shuffle sampling (gain ratio), which achieved an accuracy of 98.96%. Ismanto and Novalia [7], looked for the best classification algorithm in data mining for commodity data classification and tested the performance of various algorithms that are often used, making it easier for the Riau Provincial Government to get information on its superior commodities. In this study, they have not been able to analyze in detail the effect of the data parameters used, so the results of accuracy, recall, and precision have not been affected, and have not added comparisons with several other algorithms.

But there is also research using five years of stock trading data for the period September 30, 2014–September 30, 2019, that can be obtained from the Yahoo Finance website, Utomo et al.[8] researched and developed a prototype that supports stock prediction. They used the SVM algorithm as an algorithm that helps predict stock prices. The F-Score method requires much less processing than the maximum possible processing to find the best features, especially after the number of features exceeds 5. The F-Score method also provides a 70% F-Score and 71% accuracy. These results are only 3% worse than the best feature choice. And the authors concluded that assigning values of 0 and 1 to the zero division parameter did not change their prediction results.

In the naked eye, beef and pork are difficult to distinguish, causing consumers to be deceived when buying beef due to human visual limitations. Purnomo et al.[9] researched this matter so that Muslims can distinguish pork and beef because meat traders cheat by mixing pork and beef. They took 450 images of pork, mixed meat, and beef obtained from primary and secondary data as the basis of this research. After this research, they concluded that the number of trees affects the classification results of meat images using the random forest method. Increasing the number of trees can also increase the resulting accuracy. The optimal number of trees in this study is 280, using 10-fold data division to achieve an accuracy of 78.22%.

Purnomo et al.[9] researched distinguishing pork and beef to address the issue of meat traders mixing pork and beef, while Rindiyanı et al.[10] examined the best sales platform for Omah Branded and the accuracy of classifying its product sales data using the random forest method. Omah Branded owners face difficulties in determining the best sales platform for Omah Branded (Instagram or Shopee platform) and difficulties in determining the main inventory that matches the interests of Omah Branded customers. Then Rindiyanı et al. [10] examined this because it directly affects the income turnover of Omah Branded. They took data from sales events for all products sold in a period of 20 weeks or 5 months (December 2021– April 2022). Omah Branded sales data can be classified using the random forest method because the data stored in the training dataset does not match the data in the test dataset. However, their research only has two platforms, namely Instagram and Shoope, as research variables, which is a shortcoming in their research. Then it is known that the accuracy value using Random Forest classification of Omah Branded product sales data based on the results of the confusion matrix calculation provides an accuracy of 92%.

Studies are comparing the performance of several classification algorithms namely C4.5, Random Forest, SVM, and Naive Bayes. This work was supported by Azhari et al.[2] carried out. The four algorithms were compared based on his 200 data on the outcomes of participants of the Yogyakarta International Scout Camp 2020 (JISC2020) held in the Yogyakarta Special Region in 2020. And the C4.5 algorithm gave an accuracy of 86.67%. The random forest algorithm achieved 83.33% accuracy. The SVM algorithm achieved 95% accuracy. Naive Bayes' algorithm achieved an accuracy of 86.67%. The algorithm has the highest algorithmic accuracy and the Random Forest algorithm has the lowest. However, the amount of data used to demonstrate the performance of the two algorithms is still small, and more needed.

Instead to the research conducted by Sudioanto et al. [1]. Sudioanto conducted research using 1000 data taken from crawling results on Twitter social media, using the keyword "Rachel" with a period from October 1, 2021, to December 20, 2021. With this data, Sudioanto tried to compare classification methods to public sentiment regarding the case of Rachel Venny's escape from quarantine using the Random Forest and SVM methods. And get the conclusion that the Random Forest Algorithm gets more optimal results in conducting sentiment analysis. The Random Forest algorithm gets a precision, recall, f1-score, and accuracy value of 94%. While the SVM algorithm gets an accuracy value of 93%, the average precision value is 93%, recall is 94% and f1-score is 93%.

From the 10 journals that the author has reviewed, the author concludes that the two algorithms need to be compared using the same dataset and using stepwise data usage techniques. Where the author can analyze the performance of the model at each level of data increase and obtain the scalability of both algorithm model performances.

RESEARCH METHODOLOGY

Find Literature

The author has collected 10 references on the performance of the two algorithms and used them as research material. The selected documents are compared with other documents with respect to the conclusions of one document. Some literature supports each other, while others reach very different conclusions. Intended to provide the authors with a basis for comparative studies and the issues of this study and for the research methodology used can be seen in Figure 1.

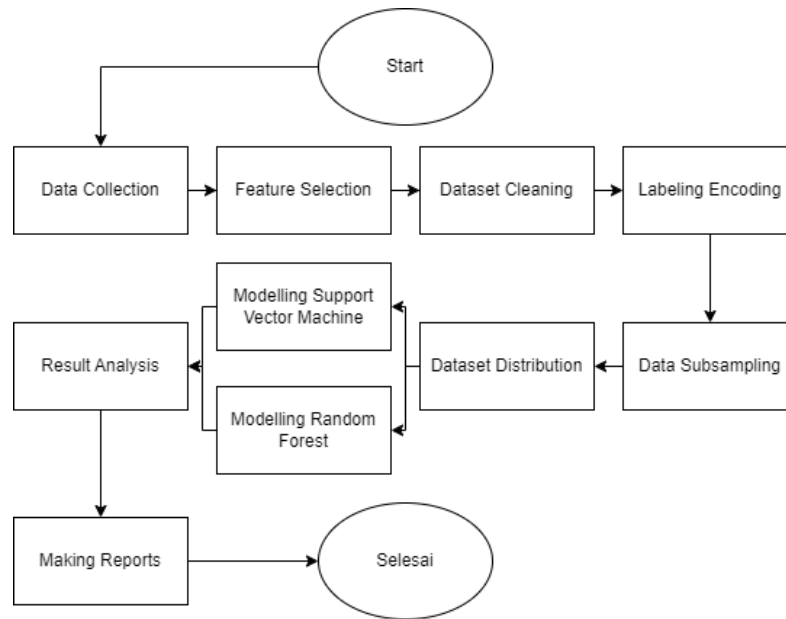


Figure 1. Research Method

Dataset Collection

encounte	patient_nb	race	gender	age	weight	ad	dis	tin	pa	medical_specialty	lal	pr	me	ou	en	inj	diag_1	diag_2	diag_3	dis	max	gl	A1C	res	met	forr	repaj	nateg	chlorig	limepi	acet
2278392	8222157	Caucasian	Female	[0-10]	?	6	25	1	1	Pediatrics-Endocrinology	41	0	1	0	0	0	250.83	?	?	1	None	None	No	No	No	No	No	No	No	No	No
149190	55629189	Caucasian	Female	[10-20]	?	1	1	7	3	?	59	0	18	0	0	0	276	250.01	255	9	None	None	No	No	No	No	No	No	No	No	No
64410	86047875	AfricanAmerican	Female	[20-30]	?	1	1	7	2	?	11	5	13	2	0	1	648	250	V27	6	None	None	No	No	No	No	No	No	No	No	No
500364	82442376	Caucasian	Male	[30-40]	?	1	1	7	2	?	44	1	16	0	0	0	8	250.43	403	7	None	None	No	No	No	No	No	No	No	No	No
16680	42519267	Caucasian	Male	[40-50]	?	1	1	7	1	?	51	0	8	0	0	0	197	157	250	5	None	None	No	No	No	No	No	No	No	No	No
35754	82637451	Caucasian	Male	[50-60]	?	2	1	2	3	?	31	6	16	0	0	0	414	411	250	9	None	None	No	No	No	No	No	No	No	No	No
55842	84259809	Caucasian	Male	[60-70]	?	3	1	2	4	?	70	1	21	0	0	0	414	411	V45	7	None	None	Steady	No	No	No	No	No	Steady	No	No
63768	114882984	Caucasian	Male	[70-80]	?	1	1	7	5	?	73	0	12	0	0	0	428	492	250	8	None	None	No	No	No	No	No	No	No	No	No
12522	48330783	Caucasian	Female	[80-90]	?	2	1	4	13	?	68	2	28	0	0	0	398	427	38	8	None	None	No	No	No	No	No	No	No	No	No
15738	63555939	Caucasian	Female	[90-100]	?	3	3	4	12	?	33	3	18	0	0	0	434	198	486	8	None	None	No	No	No	No	No	No	No	No	No
28236	89869032	AfricanAmerican	Female	[40-50]	?	1	1	7	9	?	47	2	17	0	0	0	250.7	403	996	9	None	None	No	No	No	No	No	No	No	No	No
36900	77391171	AfricanAmerican	Male	[60-70]	?	2	1	4	7	?	62	0	11	0	0	0	157	288	197	7	None	None	No	No	No	No	No	No	No	No	No
40926	85504905	Caucasian	Female	[40-50]	?	1	3	7	7	?	60	0	15	0	1	0	428	250.43	250.6	8	None	None	Steady	Up	No	No	No	No	No	No	No
42570	77586282	Caucasian	Male	[80-90]	?	1	6	7	10	?	55	1	31	0	0	0	428	411	427	8	None	None	No	No	No	No	No	No	No	No	No
62256	49726791	AfricanAmerican	Female	[60-70]	?	3	1	2	1	?	49	5	2	0	0	0	518	998	627	8	None	None	No	No	No	No	No	No	No	No	No
73578	86328819	AfricanAmerican	Male	[60-70]	?	1	3	7	12	?	75	5	13	0	0	0	999	507	996	9	None	None	No	No	No	No	No	No	No	No	No
77076	92519352	AfricanAmerican	Male	[50-60]	?	1	1	7	4	?	45	4	17	0	0	0	410	411	414	8	None	None	No	No	No	No	No	No	No	No	No
84222	108662661	Caucasian	Female	[50-60]	?	1	1	7	3	?	29	0	11	0	0	0	682	174	250	3	None	None	No	No	No	No	No	No	No	No	No
89682	107389323	AfricanAmerican	Male	[70-80]	?	1	1	7	5	?	35	5	23	0	0	0	402	425	416	9	None	None	No	No	No	No	No	No	No	No	No
148530	69422211	?	Male	[70-80]	?	3	6	2	6	?	42	2	23	0	0	0	737	427	714	8	None	None	No	No	No	No	No	No	No	No	No
150006	22864131	?	Female	[50-60]	?	2	1	4	2	?	66	1	19	0	0	0	410	427	428	7	None	None	No	No	No	No	No	No	No	No	No
150048	21239181	?	Male	[60-70]	?	2	1	4	2	?	36	2	11	0	0	0	572	456	427	6	None	None	Steady	No	No	No	No	Steady	No	No	
182796	63000108	AfricanAmerican	Female	[70-80]	?	2	1	4	2	?	47	0	12	0	0	0	410	401	582	8	None	None	No	No	No	No	No	No	No	No	No
183930	107400762	Caucasian	Female	[80-90]	?	2	6	1	11	?	42	2	19	0	0	0	V57	715	V43	8	None	None	No	No	No	No	No	No	No	No	
216156	62718876	AfricanAmerican	Female	[70-80]	?	3	1	2	3	?	19	4	18	0	0	0	189	496	427	6	None	None	No	No	No	No	No	No	No	No	No
221634	21861756	Other	Female	[50-60]	?	1	1	7	1	?	33	0	7	0	0	0	786	401	250	3	None	None	Steady	No	No	No	No	No	No	No	

Figure 2. Dataset

The data for this study was obtained from Kaggle, and the dataset used was data from diabetes patients in 130 hospitals in the US from 1999-2008 which can be accessed through https://www.kaggle.com/datasets/brandao/diabetes?select=diabetic_data.csv. Dataset on the research conducted by Strack B. et al [11] contains data on patients who are diagnosed with positive and negative diabetes with several tests that have been carried out by the patient. The authors chose this dataset because this dataset can be used by both algorithms.

Design System

Feature Selection

The author performs feature selection, in which only the most relevant features are selected. Initially, there were 50 features available in the dataset, but the author narrowed it down to 37. The eliminated features are those whose contents are the same as other data, empty, and instead to the research conducted by Strack B. et al [11]. The features/columns used are the main diagnostic values in the final dataset. In the analysis, groups representing less than 3.5% of meetings were grouped into the “other” category.

Dataset Cleaning

In this study, we use the mean substitution method to replace missing values according to Lin and Tsai [12]. This step begins by replacing the "?" value with NaN, and then converting the column's data type to numeric. Then, as you iterate through the columns of the DataFrame, each NaN value is replaced with the mean (average) value of the column. This is so that complete or more complete data can be used for further analysis in the hope of producing more accurate estimates. Although this method is commonly used, it is important to note that this decision can affect the distribution of the data and the overall analysis results.

Labeling Encoding

Start by separating the data columns into two groups, `int_column` (columns with int data type) and `object_column` (columns with object data type). This is important because labeling will only be done on columns with object data type. Perform label encoding on the columns in `object_column`. Label encoding is a technique that replaces values in categorical columns with corresponding numeric values. In the author's dataset have a "diabetesMed" column with the values "Yes" and "No" label encoding will convert them to 0 and 1. This makes categorical data into numeric data.

After completed the label encoding on `object_column`, merge `int_column` (which is still in the form of numeric data) with the columns that have label encoded in `object_column`. This merge results in a complete dataset, where all columns are already numeric data ready to be used in machine learning. This process is important to prepare data before training the machine learning model because most machine learning algorithms require numeric input. By doing so, convert the categorical data into a format that the algorithm can use. This is one of the key steps in the machine learning model development cycle.

Data Subsampling















 data_1000.csv	 me	75 KB	⋮
 data_2500.csv	 me	187 KB	⋮
 data_5000.csv	 me	373 KB	⋮
 data_10000.csv	 me	745 KB	⋮
 data_20000.csv	 me	1.5 MB	⋮
 data_50000.csv	 me	3.6 MB	⋮
 data_100000.csv	 me	7.3 MB	⋮

Figure 3. Splitted Data

At this stage, the author doesn't take all the data provided but only takes some data. After doing dataset cleaning, the author only took gradually from 1000, 2500, 5000, 10000, 20000, 50000, and 100000 rows down from 101,766 rows available in the dataset. Stepwise data usage techniques are used to avoid overfitting which occurs when the model is too complex and too specific to the training data, resulting in degraded performance when applied to new data and ensuring efficient use of available data, obtaining better model performance, and speeding up the model development process.

Dataset Distribution

The division of datasets between training datasets and testing datasets follows an 70:30 ratio, where 70% of the data is allocated for training purposes, while the remaining 30% is reserved for testing and evaluation. The purpose of splitting the data is to train the model on the most data (70%) and test its performance on different data (30%) to measure its accuracy in real-world situations and it has been proven by Nguyen et al. [13] and Supriyanto et al. [5] in their research that the 70:30 ratio is more suitable for machine learning modeling.

Model Training

The division of datasets between training datasets and testing datasets follows an 70:30 ratio, where 70% of the data is allocated for training purposes, while the remaining 30% is reserved for testing and evaluation. The purpose of splitting the data is to train the model on the most data (70%) and test its performance on different data (30%) to measure its accuracy in real-world situations and it has been proven by Nguyen et al. [13] and Supriyanto et al. [5] in their research that the 70:30 ratio is more suitable for machine learning modeling.

Support Vector Machine (SVM)

In this study, the Support Vector Machine (SVM) classification algorithm was used. In simple terms, SVM can be likened to an attempt to find the best hyperplane that functions as a separator of two classes. This is in accordance with the opinion of Kasim and Sudarsono [13]. SVM works by finding the hyperplane that has the largest margin. Margin is the distance between the hyperplane and the closest data from each class. Where the purpose of SVM is to find a hyperplane that can separate two classes of data as well as possible, with the largest margin. The parameters used by researchers are default parameters.

Random Forest

This research uses the Random Forest Classification Algorithm. Decision making uses decision trees and random forests. By applying the random forest algorithm, it can classify or classify patients diagnosed with diabetes mellitus [10]. The parameters used by researchers are default parameters. The first way random forest works is bootstrapping where from the original dataset, a number of equally large subsets of data (usually with replacement) are randomly generated. This is referred to as "bootstrap samples". Then, Decision Tree Construction is useful for each bootstrap sample, a decision tree is constructed. At each node, instead of considering all features, only a random subset of features is selected for the split data. This helps reduce overfitting and improve generalizability. Finally, to predict the class of new data, the data is passed through each decision tree. Each tree produces a class prediction. The final prediction is the one most selected by the individual trees.

Model Optimization and Performance Improvement

In this study, researchers optimized the model and tried to improve the performance of both models to get the model performance at the highest point on a particular dataset unit.

a. Normalization

First, the author used a normalization technique on the dataset. Specifically, the author used the MinMax scaler, which is used to change the value of a variable so that the value of the variable is within a uniform range. This ensures that variables with a larger range of values are not prioritized over variables with a smaller range of values when performing analysis or modeling. This alone can improve the performance of some machine learning models, especially those that are highly sensitive to variable scaling. Normalization is a technique that ensures that all data in a database has the same range. This is very important when your data is unstructured and contains very different values. MinMaxScaler normalization is useful for high-dimensional data. EEG signal values are expressed in microvolts and vary widely from channel to channel. This variation causes problems when training the model. MinMaxScaler is a type of normalization that can scale all EEG signal values to values between 0 and 1. Equation (1) and Equation (1) (2) Specify the normalization method for MinMaxScaler.

$$X_{std} = \frac{(X - X_{min})}{(X_{max} - X_{min})} \quad (1)$$

$$X_{scaled} = X_{std} * (X_{max} - X_{min}) + X_{min} \quad (2)$$

10987 In Eq. (1) and Eq. (2), min and max values are the minimum and maximum voltage values for the channel X under consideration. A channel will be made up of EEG voltage readings at 256Hz sampling rate. Eq. (1) and Eq. (2) provide the normalized values for the particular channel. The values are fit and transformed for all of the dataset and then used for training and testing according to Deepa and Ramesh [14]

b. Feature Selection

This research uses Chi-Square method is one of the most useful machines learning tools. Chi-Square equation is:

$$x^2(t, c) = \frac{N(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \quad (3)$$

If A contains t and belongs to class c, the variant frequency, B contains t and is not a child of c, then C represents the frequency of documents that do not contain t and do not belong to class c, and N is the amount of documents bravely of Mahmood MR [15].

c. Best Parameter

In this study, the authors use grid search to find the optimal parameters seperti . Grid search is inherently a comprehensive search based on a defined subset of hyperparameter space. Hyperparameters are specified in terms of minimum value (lower bound), maximum value (upper bound), and number of steps. The performance of each combination is evaluated using several performance metrics.

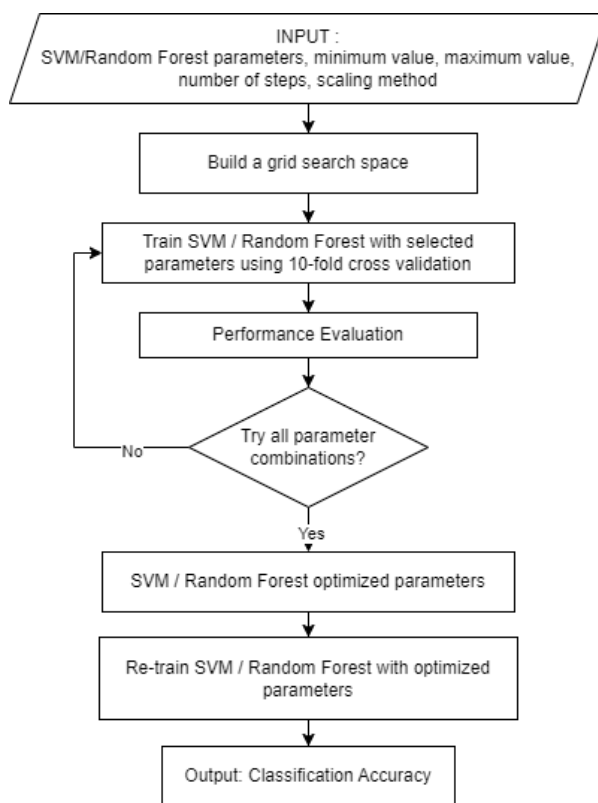


Figure 4. SVM and Random Forest parameter using GRID search

In Figure 4, the SVM and Random Forest parameters are optimized using GRID search. Grid search uses cross-validation (CV) techniques as a performance metric to optimize SVM and Random Forest parameters. The goal is to identify suitable hyperparameter combinations so that the classifier can accurately predict unknown data. The subset is used as test data and evaluated against the remaining k-1 training subsets. Then, calculate the CV error based on this splitting error for the SVM and random forest classifier using different values of other parameters. Different combinations of hyperparameter values are

input, and the one with the highest cross-validation accuracy (or lowest CV error) is selected and used to train the SVM and random forest on the entire dataset.

Coding

In this research, the author uses the Python programming language with the sci-kit-learn, numpy, pandas, matplotlib, and seaborn libraries. Storage of data to be processed in CSV(Comma-Separated Values) files.

Result Analysis

The author records the results of the time, accuracy, precision, recall, or F1-score of the random forest algorithm and the SVM algorithm using stratified data from the author's modified dataset. Then the final results of this experiment are compared, and conclusions are drawn on the algorithm that has a fast computation time and has superior accuracy, precision, recall, or F1-score. After producing the classification, the test results are evaluated from the confusion matrix and measured using accuracy, precision, recall, and F1-score. According to Rindiyani et al. [10].

Accuracy

Accuracy The accuracy value shows the prediction accuracy of the model built in both the positive class (yes) and the negative class (no).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (4)$$

Description:

TP: Number of positive data correctly classified by the system.

TN: Number of negative data correctly classified by the system.

FP : The number of positive data classified incorrectly by the system.

FN: The number of negative data classified incorrectly by the system.

Recall

Recall (Sensitivity) The ratio of true positive predictions compared to the total number of true positive data.

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (5)$$

Precision

Precision The ratio of true positive predictions compared to the total number of true positive data.

$$Precision = \frac{TP + TN}{TP + FP} \times 100\% \quad (6)$$

F1-Score

F1-score is a classification performance measure that combines precision and recall. Precision is the percentage of positive data that is actually predicted to be positive, while recall is the percentage of positive data that is actually detected.

$$F1 = 2 \times \frac{Precision + Recall}{Precision \times Recall} \quad (7)$$

Making Reports

After doing all that, the author compiles a report based on the research that has been done. This report includes the background of the issues raised, problem formulation, literature review, research methodology, implementation and results obtained, and conclusions that can be drawn.

IMPLEMENTATION

So the experiments that the author did in this study by 1 by 1 adding several attributes such as Feature Selection, Best Parameter, and also Normalization. And there are also several combinations that were tested. The following are some of the important experiments in this report and the rest of the experiments can be found in the appendix. The author only writes down the resume of other test trials.

Resume

From the experiments authors will briefly show what happened in this experiment. And this is the description of each experiments: Experiment 1 both algorithms were trained with default parameters without adding any attributes. Experiment 2 is where both algorithms are given data that has been normalised with the minmax scaler and still with the default parameters. Experiment 3 is an experiment with the addition of Feature Selection, where both algorithms are trained with data whose features have been selected with the best features using the Select K Best variant of Chi squared. Experiment 4 is the algorithm with the Best Parameter, which uses grid search in finding the best parameter for each algorithm. Then in Experiment 5 we combine data normalisation and Feature Selection before being trained by the models of both algorithms. Experiment 6 is Normalisation of data and then used to train the model with both algorithms with the Best Parameter of both algorithms in the dataset that will be used. Experiment 7 is Feature Selection and Best Parameter. The last experiment is a combination of all the experiments, namely by normalising the data, then selecting the top 5 best features from the dataset, then used to train the models of both algorithms using the best parameters of each algorithm.

Table 1. Summary of the 8 experiments conducted (SVM - Accuracy)

Data	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8
1.000	0.775	0.999	0.774	0.987	0.957	0.975	0.927	0.957
2.500	0.7688	0.9972	0.9112	0.996	0.9532	0.9948	0.9276	0.9532
5.000	0.7742	1	0.917	0.9982	0.9564	0.9952	0.928	0.9564
10.000	0.8234	0.9998	0.924	0.9987	0.9572	0.9975	0.9268	0.9572
20.000	0.96615	0.99995	0.92485	0.99955	0.95685	0.9992	0.92555	0.95685
50.000	0.9871	0.99998	0.9268	0.99976	0.95674	0.99984	0.92702	0.95674
100.000	0.99162	0.99999	0.92776	0.99985	0.957	0.99992	0.92785	0.957

Table 2. Summary of the 8 experiments conducted (Random Forest - Accuracy)

Data	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8
1.000	0.993	0.989	0.92	0.994	0.957	0.993	0.921	0.957
2.500	0.9948	0.9948	0.9264	0.9956	0.9532	0.996	0.9272	0.9532
5.000	0.9964	0.9962	0.9268	0.9978	0.9564	0.9976	0.9274	0.9564
10.000	0.9973	0.9977	0.9257	0.9984	0.9572	0.9987	0.9266	0.9572
20.000	0.9989	0.99915	0.9253	0.9995	0.95685	0.9996	0.92525	0.95685
50.000	0.99954	0.99948	0.92692	0.99976	0.95674	0.99968	0.92698	0.95674
100.000	0.99973	0.99972	0.92762	0.99981	0.957	0.99982	0.92768	0.957

In Table 1 it can be seen that the accuracy of the SVM model which experienced a significant increase was in experiment 1 from 77.5% to 99.162% on the highest data. Then in experiment 4 it can be seen that from the 1000 data used it can touch 98.7% and experience an insignificant increase or arguably stable until it touches 99.985% and does not experience a decrease in accuracy at all. Whereas in Table 2 the accuracy of the Random Forest model is good when in experiment 1 with an accuracy level that continues to increase from 1000 data with 99.3% accuracy continues to experience an insignificant increase to 100000 data with 99.973% accuracy level. In experiments 4 and 6 the model with the Random Forest algorithm has a higher level of accuracy and continues to increase as the amount of data used increases. In both algorithms there are experiments that produce exactly the same evaluation value, namely in experiment 5 and experiment 8, the difference is in computation time only.

Table 3. Summary of the 8 experiments conducted (SVM – Precision)

Data	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8
1.000	0.775	0.998718	0.774774	0.992315	1	1	0.998519	0.957
2.500	0.7688	0.999479	0.984123	0.998437	1	1	1	0.9532

5.000	0.7742	1	0.992646	1	1	1	1	0.9564
10.000	0.81566	1	0.997589	0.999871	1	1	0.999858	0.9572
20.000	0.98468	1	0.99843	0.999871	1	1	0.999929	0.95685
50.000	0.99924	1	0.999514	0.999897	1	1	0.999972	0.95674
100.000	0.99976	0.999987	0.999728	0.999948	1	1	0.999986	0.957

Tabel 4. Summary of the 8 experiments conducted (Random Forest - Precision)

Data	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8
1.000	1	0.997452	0.987266	1	1	1	0.990407	0.957
2.500	1	1	0.992657	1	1	1	0.993226	0.9532
5.000	1	1	0.998287	1	1	1	0.999141	0.9564
10.000	1	1	0.997475	1	1	1	0.999009	0.9572
20.000	1	1	0.999361	1	1	1	0.999363	0.95685
50.000	1	1	0.9998	1	1	1	0.999885	0.95674
100.000	1	1	0.999613	1	1	1	0.999742	0.957

In tables 3 and 4 we can see a summary of the precision score results of the two algorithms tested in this study. In experiments 1, 4, 5, and 6 Random Forest obtained 100% in each data increment, meaning that the trained model did not vary significantly when tested on different datasets / different situations. Experiments 7 and 8 show the ability of the Random Forest algorithm model where the addition of data that occurs here makes the model continue to reduce its false positive classification so that the performance obtained becomes better. Then in experiments 2, 5, and 6, it can be seen that the SVM model in table 4.17 obtained a precision value that was stable / did not vary even though in experiment 2 it had the lowest value in the experiment which was 99.8718%. And in experiments 1, 3, and 8 show that there is an increase in the precision value of the model with the SVM algorithm.

Tabel 5. Summary of the 8 experiments conducted (SVM - Recall)

Data	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8
1.000	1	1	0.99871	0.990968	0.944516	0.967742	0.907097	0.957
2.500	1	0.996875	0.89906	0.99636	0.939123	0.993235	0.905824	0.9532
5.000	1	1	0.899507	0.997676	0.943683	0.9938	0.906998	0.9564
10.000	0.99833	0.999743	0.904364	0.998455	0.94491	0.996782	0.905909	0.9572
20.000	0.97137	0.999935	0.904285	0.999548	0.944225	0.998966	0.903833	0.95685
50.000	0.98407	0.999974	0.905785	0.999793	0.944061	0.999793	0.905656	0.95674
100.000	0.98935	1	0.906409	0.999857	0.944144	0.999896	0.906292	0.957

Tabel 6. Summary of the 8 experiments conducted (Random Forest - Recall)

Data	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8
1.000	0.990968	0.988387	0.908387	0.992258	0.944516	0.990968	0.907097	0.957
2.500	0.993235	0.993235	0.911028	0.994276	0.939123	0.994797	0.911549	0.9532
5.000	0.99535	0.995092	0.906998	0.997159	0.943683	0.996901	0.906998	0.9564
10.000	0.996525	0.99704	0.906681	0.997941	0.94491	0.998327	0.906424	0.9572
20.000	0.998578	0.998901	0.904027	0.999354	0.944225	0.999483	0.903962	0.95685
50.000	0.999405	0.999328	0.905682	0.99969	0.944061	0.999586	0.905682	0.95674
100.000	0.999649	0.999636	0.906331	0.999753	0.944144	0.999766	0.906292	0.957

In tables 5 and 6 we can see a summary of the recall score results of the two algorithms tested in this study. In experiments 4 and 6 SVM models has improved in reducing false negative detections, but in other experiments experienced an increase in some subdata and also experienced a decrease in some subdata not consistently increasing or decreasing. Lalu pada tabke 4.20 rangkuman Random Forest, eksperimen 1, 2, 4, dan 6 mengalami kenaikan skor dan tidak ada penurunan. Random Forest lebih bisa meningkatkan pendeteksian false negatif nya di 4 eksperimen dibandingkan dengan SVM.

Tabel 7. Summary of the 8 experiments conducted (SVM – F1 - Score)

Data	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8
1.000	0.87324	0.999357	0.872603	0.991611	0.971308	0.98329	0.950455	0.957
2.500	0.86929	0.998173	0.939621	0.997395	0.968597	0.996602	0.95057	0.9532
5.000	0.87273	1	0.943753	0.998835	0.97102	0.996885	0.951217	0.9564
10.000	0.89779	0.999871	0.948675	0.999162	0.971672	0.998386	0.950558	0.9572
20.000	0.97798	0.999968	0.949022	0.999709	0.97131	0.999483	0.949448	0.95685
50.000	0.99159	0.999987	0.950343	0.999845	0.971225	0.999897	0.950478	0.95674
100.000	0.99453	0.999994	0.950782	0.999903	0.971269	0.999948	0.950834	0.957

Tabel 8. Summary of the 8 experiments conducted (Random Forest – F1 - Score)

Data	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8
1.000	0.995452	0.992863	0.946048	0.996104	0.971308	0.995444	0.946631	0.957
2.500	0.996602	0.996602	0.950067	0.997126	0.968597	0.997388	0.95061	0.9532
5.000	0.997668	0.997538	0.950445	0.998576	0.97102	0.998446	0.950831	0.9564
10.000	0.998258	0.998517	0.949893	0.998969	0.971672	0.999162	0.950455	0.9572

20.000	0.999289	0.99945	0.949297	0.999677	0.97131	0.999741	0.949261	0.95685
50.000	0.999702	0.999664	0.950415	0.999845	0.971225	0.999793	0.950454	0.95674
100.000	0.999825	0.999818	0.950687	0.999877	0.971269	0.999883	0.950724	0.957

In the tables above, tables 7 And 8 contain a summary of F1-Score or the balance of the ability of the algorithm model to detect false negatives and false positives, or the balance between precision and recall values. In experiments 3, 4, 6, and 7 the SVM Algorithm model obtained an ever-increasing F1-Score which means that the performance shown is good because the model is increasingly learning the data given. Then in experiments 1, 2, 3, 4, 6, and 7 the Random Forest Algorithm model also increased to produce good performance in detecting false negatives and false positives.

Tabel 9. Summary of the 8 experiments conducted (SVM – Time Computation)

Data	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8
1.000	0.1206962109	0.037697	0.089477	0.210883	0.039998	0.056578	0.252385	0.04042
2.500	0.566228199	0.097362	0.407193	0.570106	0.147101	0.239392	0.55315	0.11767
5.000	3.235619211	0.218984	1.113969	1.387589	0.609599	1.072012	1.492679	0.36703
10.000	5.920486021	0.51281	2.970849	5.39408	1.319793	1.715585	5.177019	0.82240
20.000	21.76576767	1.860737	18.82178	5.97956	3.854241	4.802966	23.11707	3.39460
50.000	116.2835956	4.616325	45.93975	18.27541	27.10332	15.96853	105.043	16.6240
100.000	453.3295238	12.58395	296.1649	84.43455	68.29472	142.0238	338.0198	63.7107

Tabel 10. Summary of the 8 experiments conducted (Random Forest – Time Computation)

Data	Experiment 1	Experiment 2	Experiment 3	Experiment 4	Experiment 5	Experiment 6	Experiment 7	Experiment 8
1.000	0.624436	0.209953	0.970077	0.310097	0.181798	0.768703	0.261332	0.241522789
2.500	0.467948	0.28313	1.011642	0.58169	0.202509	0.88805	0.35273	0.2717500687
5.000	0.631278	0.390907	1.153611	0.969641	0.287118	0.859536	0.458257	0.4629778862
10.000	1.166907	0.634728	0.755974	1.333934	0.366638	2.189115	1.264488	0.9641777515
20.000	1.544448	1.575055	1.003495	2.388723	0.529711	2.750263	1.107552	0.7826076031
50.000	3.389108	3.154572	5.027922	6.911544	1.52787	6.254779	2.810311	1.960704565
100.000	7.31551	6.665683	3.126494	11.33091	2.079292	8.045849	5.409761	3.009775162

Tables 9 and 10 contain a summary of the computation time required by both algorithms in processing data and modelling. In experiment 1 which is as much as 446.014 seconds on the amount of data 100000 data and the least in experiment 2 which is 5.918267 seconds with 100000 data used. And the SVM algorithm model reached the maximum speed in the 2nd experiment with

a speed of 12.58395 seconds to process 100,000 data, and reached a long time in experiment 1 when processing 100,000 data reaching 453.3295238 seconds. While in Random Forest there is not too significant a difference between experiments conducted on this algorithm, the fastest speed is in the 5th experiment on 100,000 data and the slowest speed is in the 4th experiment with a time of 11.33091 seconds on 100,000 data. Between the two algorithms, there is a difference where SVM is faster at processing data with an amount that is not massive compared to Random Forest, this can be seen when the SVM model processes 10,000 data. The model slows down because it receives a lot of data, in contrast to Random Forest, this algorithm slows down as data is added but the slowdown is very insignificant, for example in experiment 4, Random Forest increases computation time from 0.310097 to 11.33091 at the highest. If in SVM in Experiment 1 from 1000 data with a speed of 0.1206962109 seconds where the time is faster than Random Forest, but began to weaken to 453.3295238 seconds at 100,000 data.

CONCLUSION

In this study, the authors compared the SVM algorithm and the Random Forest algorithm in diabetes classification. So from the 8 experiments that have been carried out, the performance of the svm algorithm model that performs well in experiment 1, experiment 3, experiment 4 and experiment 6. And the performance of the random forest model that performs well is in experiment 1, experiment 2, experiment 3, experiment 4, and experiment 6. The highest accuracy on the smallest data (1,000 data) and the largest data (100,000 data) is in the 6th experiment. The highest accuracy on the largest data is the SVM algorithm model with an accuracy of 99.992% with a time speed of 142.0238 seconds, while Random Forest gets an accuracy of 99.982% with a time speed of 8.045849 seconds. Random Forest algorithm model is faster 133.977951 seconds than the svm algorithm model with a lower accuracy of 0.01%, but the best accuracy on the SVM algorithm model but slower time. The highest accuracy on the smallest data is the Random Forest algorithm model with 99.4% accuracy with a time speed of 0.310097 seconds while SVM gets 98.7% accuracy with a time speed of 0.210883 seconds. SVM algorithm model is faster 0.099214 seconds than the Random Forest algorithm model with a lower accuracy of 0.7%, but the best accuracy on the Random Forest algorithm model but slower time.

This research can be further developed by replacing the dataset with a simpler one or with another topic, because the dataset used in this research has a very large and varied number of features or columns. And using other algorithms that can survive in large amounts of data such as Random Forest.

REFERENCES

- [1] Sudianto, Wahyuningtias P, Utami HW, et al. COMPARISON OF RANDOM FOREST AND SUPPORT VECTOR MACHINE METHODS ON TWITTER SENTIMENT ANALYSIS (CASE STUDY: INTERNET SELEBGRAM RACHEL VENNYA ESCAPE FROM QUARANTINE). *Jurnal Teknik Informatika (Jutif)* 2022; 3: 141–145.

- [2] Azhari M, Situmorang Z, Rosnelly R. Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA* 2021; 5: 640–651.
- [3] Saputro ND. Penerapan Algoritma Support Vector Machine untuk Prediksi Harga Emas. *Jurnal Informatika Upgris*; 1. Epub ahead of print 2015. DOI: 10.26877/jiu.v1i1.
- [4] Dwiasnati S, Devianto Y. Optimasi Prediksi Bencana Banjir menggunakan Algoritma SVM untuk penentuan Daerah Rawan Bencana Banjir. *Prosiding SISFOTEK* 2021; 5: 202–207.
- [5] Supriyanto Y, Ilhamsyah M, Enri U. Prediksi Harga Minyak Kelapa Sawit Menggunakan Linear Regression Dan Random Forest. *Jurnal Ilmiah Wahana Pendidikan* 2022; 8: 178–185.
- [6] Fachid S, Triayudi A. Perbandingan Algoritma Regresi Linier dan Regresi Random Forest Dalam Memprediksi Kasus Positif Covid-19. *JURNAL MEDIA INFORMATIKA BUDIDARMA* 2022; 6: 68–73.
- [7] Ismanto E, Novalia M. Komparasi Kinerja Algoritma C4.5, Random Forest, dan Gradient Boosting untuk Klasifikasi Komoditas. *TechnoCom* 2021; 20: 400–410.
- [8] Utomo VG, Wakhidah N, Putri AN. PREDIKSI HARGA SAHAM DENGAN SVM (SUPPORT VECTOR MACHINE) DAN PEMILIHAN FITUR F-SCORE. *Jurnal Informatika Upgris*; 6. Epub ahead of print 7 July 2020. DOI: 10.26877/jiu.v6i1.5306.
- [9] Purnomo TY, Yanto F, Insani F, et al. Penerapan Algoritma Random Forest pada Klasifikasi Daging. *Jurnal Intra Tech* 2022; 6: 21–34.
- [10] Rindiyani R, Primadewi A, Maimunah M, et al. Klasifikasi Penjualan berdasarkan Platform pada UMKM Omah Branded Menggunakan Random Forest. *JURIKOM (Jurnal Riset Komputer)* 2022; 9: 1520–1529.
- [11] Strack B, DeShazo JP, Gennings C, et al. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International* 2014; 2014: 1–11.
- [12] Lin W-C, Tsai C-F. Missing value imputation: a review and analysis of the literature (2006–2017). *Artif Intell Rev* 2020; 53: 1487–1509.
- [13] Nguyen QH, Ly H-B, Ho LS, et al. Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. *Mathematical Problems in Engineering* 2021; 2021: e4832864.
- [14] Deepa B, Ramesh K. Epileptic seizure detection using deep learning through min max scaler normalization. *ijhs* 2022; 10981–10996.
- [15] Mahmood MR. Two Feature Selection Methods Comparison Chi-square and Relief-F for Facial Expression Recognition. *J Phys: Conf Ser* 2021; 1804: 012056.