Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2024. Vol. 3. No. 3; e-ISSN: 2808-4519

Performance Analysis of the Support Vector Machine Algorithm in Predicting Rain Potential in DKI Jakarta

Aina Latifa Riyana Putri

Program Studi S1 Sains Data, Institut Teknologi Telkom Purwokerto
ainalatifariyanaputri@ittelkom-pwt.ac.id

Abstract

Indonesia is a country with a tropical climate which is located on the equator which tends to get sunlight throughout the year. Not only gives beauty but also saves disasters that can be dangerous such as floods, which usually occur due to high rainfall. The impact is also large on facilities with damage to buildings and health problems. It is very important to prepare it so that the possibility of damage or loss can be minimized. This research will apply the Support Vector Machine (SVM) Algorithm by selecting the distribution ratio of training and test data as well as the best kernel function to predict the potential for rain using daily climate data from the Meteorology, Climatology and Geophysics Agency (BMKG) in DKI Jakarta with the help of Rstudio software. The performance evaluated using the confusion matrix method produces the highest accuracy value of 89% is the SVM model with a training data distribution ratio of 90% and the Linear kernel as the chosen model for predicting rain potential.

Keywords: *Support Vector Machine, Classification, Prediction, Confusion Matrix*

1. Introduction

Tropical climate is a type of climate that covers the earth between the equator in the range 23.5° N to 23.5° S. One of the countries with a tropical climate is Indonesia, which is located on the equator, which means it tends to get sunlight all year round. Usually tropical climate areas have higher temperatures than other climates. According to meteorology, tropical climates have an average temperature above 18° C. This is what makes areas passing through tropical climates tend to only have two seasons, namely the rainy and dry seasons. The advantage of areas with tropical climates is that they have a lot of biodiversity and natural resources that are worth preserving. Because almost all types of living creatures, both plants and animals, can live under a balance between sunlight and rain, for example Raflessia Arnoldi, edelweiss, pitcher plant, one-horned rhinoceros, and others.

Not only does it provide beauty but also saves from dangerous disasters such as floods, which usually occur due to high rainfall. Several areas in Indonesia are often prone to flooding, especially the DKI Jakarta area which is almost certain that when the rainy season arrives it will be affected by flooding [1]. The impact is also large on facilities with damage to buildings and infrastructure, health problems, and water and environmental pollution. According to BPBD in 2020, the total number of flood evacuation posts in DKI Jakarta reached 269 posts with a total of 31,232 evacuees and 19 people who died due to the flood. Therefore, it is very important to prepare it so that the possibility of damage or loss can also be minimized. Efforts to predict potential rain can be used as an early warning of potential flooding [2]. Because the government and the public receive information quickly and accurately about the potential for flooding that will occur, they can prevent losses due to the impact of flooding.

Several studies related to weather or rain predictions have been carried out, such as [3] predicting rain based on meteorological data from the Australian Meteorological Bureau by creating Logistic Regression and Decision Tree models. It was obtained that the model using Logistic Regression had the best AUC value of 0.871. On [4] build a weather forecast prediction model based on data from the Kaggle website with the Neural Network, Naïve Baiyes, Random Forest, and K-Nearest Neighbor algorithms. Obtained with the Random Forest algorithm achieving the best accuracy of 89%. In [5] analyzed weather data using 2188 data samples with the Random Forest algorithm. In [6] predicts the potential for rain in Ternate City using the Naïve Bayes algorithm. Obtained accuracy of 76%. In [7] identified rainfall in the state of Andhra Pradesh using the SVM (Support Vector Machine), Random Forest, K-Nearest Neighbor, Decision Tree algorithm. It was found that the efficiency of the SVM algorithm was better compared to other algorithms. The five studies are classification analyzes using various algorithms.

Classification analysis is carried out with the aim of determining the group or collective membership of an individual. The Support Vector Machine (SVM) algorithm is a classification technique that includes kernel concepts such as Linear, Polynomial, Sigmoid kernels, etc. in high-dimensional space. The goal of this algorithm is to divide the dataset into classes to find the best hyperplane by measuring the margin and finding the maximum point of distance between classes. Hyperplane is a function that can be used to separate classes. Algorithm SVM is widely implemented in various problems in the real world [8] such as the medical field for example [9] detecting heart disease

classification using the SVM algorithm provides an AUC value of 0.868. Up to [10] to estimate battery life in electric vehicles using an algorithm SVM with an MSE value of 0.01505.

The uniqueness of the SVM algorithm when compared to other algorithms is that most other algorithms perform separation based on the average pattern of existing data. Meanwhile, the SVM algorithm performs classification based on data points that are similar but in different classes. Apart from that, this algorithm can classify data without looking at the level of significance of each independent variable to be studied [11].

Therefore, by looking at the impact of flooding, this research intends to apply the Support Vector Machine (SVM) algorithm to predict the potential for rain using daily climate data in DKI Jakarta from the Meteorology, Climatology and Geophysics Agency (BMKG) with the help of Rstudio software. Accuracy in predictions is also an important factor in research so that the model obtained can be directly applied to a problem. The application of the SVM algorithm can be done by optimizing based on the selection of kernel type and the size of the data splitting ratio. The aim of this research is to determine optimal accuracy results based on the confusion matrix of each experiment, the ratio of training data and test data as well as the type of kernel function in the SVM algorithm as a model for predicting potential rain in DKI Jakarta. The research stages consist of several stages, namely data collection, data preprocessing, dividing training data and test data, determining the best kernel function and optimal parameters with training data, applying the best kernel function and optimal parameters to the test data.

2. Research Methods

This research method includes data sources and data analysis methods.

2.1. Data Source

The data that will be used is secondary data from the Online Data Center Database of the Meteorology, Climatology and Geophysics Agency (BMKG), especially at the Tanjung Priok Maritime Meteorological Station, North Jakarta from 01 January 2019 to 10 November 2022 with parameters that can be seen in Table 1. January 2019 to 10 November 2022 with 10 parameters which can be seen in Table 1.

Table 1: Variable Description

No	Variable	Description
1	ddd_x	The wind direction at maximum speed (°)
2	ddd_car	Maximum wind speed (m/s)
3	RR	Average wind speed (m/s)
4	ff_x	Average Humidity (%)
5	ff_avg	Length of sunshine (hours)
6	RH_avg	Maximum temperature (°C)
7	ss	Minimum temperature (°C)
8	Tx	Average temperature (°C)
9	Tn	Most wind direction (°)
10	Tavg	Rainfall (mm)

Based on Table 1, it is known that the daily climate data collection used uses 10 parameter variables.

2.2. Data Analysis Method

The following are the experimental stages carried out in the research

1. Carrying out data collection
   The data used will be downloaded and saved as a file with the extension .xlsx.
2. Perform data pre-processing
   At this stage, the first step is to handle missing values. Missing values are a problem because almost all statistical methods assume all information for all variables entered in the analysis is complete. So these errors must be cleaned and corrected. There are two ways to handle missing values, namely removing data and data imputation. Removing data can be done by deleting all observation data that contains one or more missing values (listwise deletion) [12] or if there is a variable that has a lot of missing values, and this variable is not a variable that significantly influences the response, then this variable can be removed (dropping variables). Meanwhile, for data imputation, missing values can be filled in by guessing directly or using a correlation-based estimator. There are various approaches to imputation such as mean imputation [3], regression imputation [14], and others. Second, variable modification, where the variables used in this research consist of the target variable (Y) and predictor variables (X). The variable Yused is the category of rainfall in the area, which consists of 2 categories, namely:
   $Y = (0)$; No rain occurs if the variable "RR" has a value equal to 0.
   $Y = (1)$; Rain occurs if the variable "RR" has a value of ≥ 0. Meanwhile, the predictor variable (X), can be seen in Table 2 below.

Table 2: Predictor Variable

No	Variable	Data Type
1	ddd_x	Numeric
2	ddd_car	Categorial 1= C ; 2=E; 3=N; 4=NE; 5=NW; 6=S; 7=SE; 8=SW; 9=W
3	ff_x	Numeric
4	ff_avg	Numeric

5	RH_avg	Numeric
6	ss	Numeric
7	Tx	Numeric
8	Tn	Numeric
9	Tavg	Numeric

Based on Table 2, 9 parameter variables have been determined for predictor variables to determine rain potential.

3. Carry out classification with the SVM algorithm

At this stage, the first step is to form test data and training data. In this study, a ratio splitting data experiment was determined to obtain the greatest accuracy, which can be seen in Table 3.

Table 3: Ratio Splitting Data

No	Ratio	Frequency of Train Data	Frequency of Test Data
1	90:10	787	87
2	80:20	699	175
3	70:30	612	262
4	60:40	525	349
5	50:50	436	438

Table 3 shows the amount of training data and test data for each data splitting ratio experiment before building the SVM model. Second, standardize data attributes. The process of standardizing data on numerical variables for training data and test data is carried out so that the data has the same range of values (scale). To form standardization values, use the following formula.

$$X_{stand} = \frac{x-\mu}{\sigma} \quad (1)$$

Where is the standardization value (X_{stand}) obtained by dividing the difference between the data value (x)and the distribution mean (μ) with the distribution standard deviation value (σ).

Third, form a Support vector Machine model. The Support Vector Machine algorithm was first introduced by Vapnik in 1992. In simple terms, this algorithm is a Supervised Machine Learning method which is used as a classification algorithm for solving classification problems that separate 2 classes. An illustration of the SVM algorithm can be seen in Figure 1.

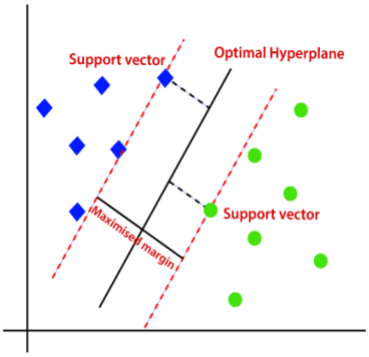

Fig. 1: Ilustration of Algorithm Support Vector Machine

The concept, according to Figure 1, is that each data is plotted as a point in n-dimensional space (where n is the number of features it has) with the value of each feature being a certain coordinate value. Then it is classified by looking for a hyperplane that differentiates the two classes well. The aim of the Support Vector Machine algorithm is to find the best hyperplane in n-dimensional space by maximizing the distance between the hyperplane and the closest pattern from each class. This closest pattern is called the support vector. This distance can function to clearly classify data points or as a separator of two classes in the input space.

One of the advantages of the Support Vector Machine algorithm is that it is a technique in soft computing [15] where you can choose other options in the model parameters. In this research, kernel selection experiments were also carried out in Table 4 based on experiments on each ratio splitting data in Table 3 to obtain the greatest accuracy in the Support Vector Machine model.

Table 3: Kernel Support Vector Machine Algorithm

No	Kernel	Definition of Function
1	Linear	$K(x_1, x_2) = x_1.x_2$
2	Polinomial (Cost 0.5 &1.0)	$K(x_1, x_2) = (x_1.x_2 + r)^d$
3	Radial (Cost 0.5 &1.0)	$K(x_1, x_2) = e^{(-\gamma\|x_1 - x_2\|^2)}$
4	Sigmoid (Cost 0.5 &1.0)	$K(x_1, x_2) = \tanh(\gamma(x_1.x_2) + c)$
5	Linear	$K(x_1, x_2) = x_1.x_2$

Table 4 shows the types of kernels that will be compared in this study along with the formula where x_1 and x_2 are two different observations in the dataset, r are polynomial coefficients, d are a set of polynomial degrees, γ are squared distance scales, c are sigmoid coefficients, and cost is a parameter that provides contributions to hyperplanes in n-dimensions. The kernel type and data splitting ratio work as optimization of the performance of the Support Vector Machine algorithm [16] to maximize margins and avoid classification errors in each sample in the training data.

4. Conduct model comparisons
   At this stage, a comparison of the evaluation of each Support Vector Machine Algorithm classification model using a Confusion Matrix is displayed with the calculation of accuracy values on training data and test data based on experiments with various data and kernel splitting ratios.
5. Draw a conclusion
   At this stage, conclusions are drawn by determining which Support Vector Machine Algorithm model is the best to choose as a model for determining predictions of rain in the city of Jakarta based on the accuracy values in the confusion matrix. Furthermore, the best model can also obtain precision, recall and F1-Score values along with their interpretation.

3. Results and Discussion

In this research, the Support Vector Machine (SVM) algorithm was used to carry out testing and analysis as a model for determining the potential for rain in DKI Jakarta. The testing process is carried out by comparing parameters, namely kernel type and data splitting ratio.

3.1. Parameter Comparison Testing

The kernel types and data splitting ratio values that will be compared in this research have been explained in the previous section. The results of the comparison of kernel types and data splitting ratio values for training data and test data are shown in graphical form in Figure 2 and Figure 3.

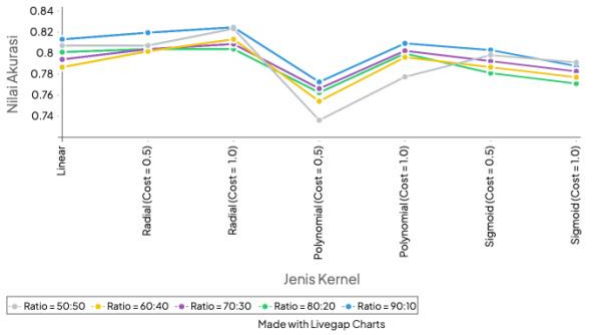

Fig. 2: Results on Training Data

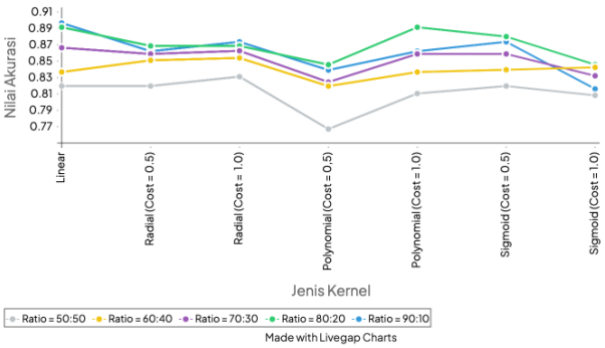

Fig. 3: Results on Testing Data

Figure 2 and Figure 3 show the accuracy results of 35 Support Vector Machine models on training data and test data in graphical form. Based on these two figures, it can be seen that the choice of kernel type and data splitting ratio has an influence on the SVM model that is formed.

In Figure 2, the Radial kernel type (Cost = 1.0) with a data splitting ratio of 90:10 has the highest accuracy for testing on training data of 0.8247. Meanwhile, the Polynomial kernel type (Cost = 0.5) with a data splitting ratio of 50:50 has the lowest accuracy for testing on training data of 0.7362. In the figure, determining a data splitting ratio of 90:10 is considered to have the best accuracy for each model with different kernel types, with the exception of the Sigmoid kernel (Cost = 1.0). Giving the Cost value also turns out to have an effect on the accuracy of the SVM model in testing using training data. In the Sigmoid and Polynomial kernel types, giving a high Cost value has an impact on increasing the accuracy of the SVM model. This is different from the Sigmoid kernel type, where Figure 2 shows a decrease in accuracy as the Cost parameter value increases.

In Figure 3, the Linear kernel type with a data splitting ratio of 90:10 has the highest accuracy for testing on test data of 0.8966. Meanwhile, the Polynomial kernel type (Cost = 0.5) with a data splitting ratio of 50:50 has the lowest accuracy for testing on test data of 0.7671. In the

figure, determining a data splitting ratio of 80:20 is considered to have the best accuracy for each model with different kernel types, with the exception of Linear and Radial kernels (Cost = 1.0). Giving the Cost value affects the accuracy of the SVM model in testing using test data. In the Sigmoid and Polynomial kernel types, giving a high Cost value has an impact on increasing the accuracy of the SVM model. This is different from the Sigmoid kernel type, where Figure 3 shows a decrease in accuracy with increasing Cost parameter values, except for determining the data splitting ratio of 60:40.

Because accuracy on training data is not an indicator of model performance for future data, the highest accuracy value of 89% obtained for test data is the SVM model with a 90:10 Linear kernel ratio as the selected model for predicting potential rain. Comparison of the accuracy of training data and test data needs to be done to ensure that the model is not overfitted. When the accuracy of the training data significantly outperforms the accuracy of the test data, there is a high probability of overfitting [17].

3.2. Performance of Selected Models

Apart from accuracy, the performance of a classification model can also be seen from the precision, recall and F1-Score values. It can be seen in the image that table is obtained confusion matrix in the SVM model with a data splitting ratio of 90:10 Linear kernel.
The value of precision, recall and F1-Score obtained from the classification model above are shown in Table 5.

Table 5: Best Model Performance

	Precision	Recall	F1-Score
Nilai	0,8913	0,9111	0,9011

Based on Table 5, the precision value obtained is 0.8913 which means that there are 89% correct predictions of potential rain from all days that are predicted to have the potential to rain. So it can be seen that the precision value is also high in this classification. The Recall value obtained is 0.9111 which means that there are 91% predicted to have the potential for rain compared to all days that are predicted to have the potential to rain. So it can be seen that the recall value is also high in this classification. In case in above, the F1 - Score "potential for rain" is 90%. Therefore, the classification using the C5.0 Algorithm is considered very good.

3.3. Prediction of Potential Rain

To test the model, predictions of potential rain were made using test data. In the test data, the target variable (class) will be made unknown and the value of this variable will be predicted using the SVM model. The results of predicting potential rain using test data with the selected SVM algorithm model can be seen in Figure 4.

```
## Confusion Matrix and Statistics
##
##    datauji.prediksi
##       0  1
##    0 37  5
##    1  4 41
```

Fig. 4: Results of Rain Prediction

Figure 4 shows the prediction results of the best SVM model regarding the potential for rain in DKI Jakarta. Predictions were made using 87 test data because the selected SVM model had a data splitting ratio of 90:10 with a Linear kernel type. In this figure, it can be seen that of the 87 data, 37 data were predicted correctly when it was raining and 41 data were predicted correctly when it was not raining. By seeing that the accuracy is quite large and the classification error in Figure 4 is quite small, this model can be used to determine results related to potential rain.

3.4. Previous Research

A comparison of the accuracy regarding the formation of a prediction model for potential rain in DKI Jakarta using the Support Vector Machine algorithm based on previous studies can be seen in Table 6.

Table 6: Comparison of Accuracy Analysis Against Previous Research

Reference	Data	Kernel Type & Ratio	Accuracy
[18]	Curah hujan bulanan di India tahun 1951-2000	- 80:20	6.97%
[19]	Curah hujan harian di India tahun 1979-2013	Linear 70:30	53.9%
[20]	Curah hujan bulanan stasiun Shivajinagar di India tahun 2000-2018	Polynomial 80:20	82.1%

In Table 6 it can be seen that the accuracy comparison that this research proposes, comparing the selection of the combination of kernel type and data splitting ratio to obtain optimal model accuracy, provides the highest accuracy compared to previous rain prediction research. By looking at the data used in several previous studies, this research also shows that SVM has good accuracy results for solving problems on small datasets [21].

4. Conclusion

Based on research conducted using Rstudio on daily climate data from 01 January 2019 to 10 November 2022 with 10 parameters from the Online Data Database Center of the Meteorology, Climatology and Geophysics Agency (BMKG) especially at the Tanjung Priok Maritime Meteorological Station which was tested with the Support Vector algorithm Machine by experimenting with various data splitting ratios and kernels, performance was obtained which was evaluated using the accuracy value of 89%, namely the SVM model with a 90:10 Linear kernel ratio as the selected model for predicting potential rain. Also obtained was a precision value of 89 %, recall of 91%, and F1-Score of 90%. The method in this research shows that the resulting classification is very good, so the model can be used for predictions using this pattern to determine results related to potential rain.

References

[1] Chalid, A., & Prasetya, B. (2020, February). Utilization of a pond in East Jakarta for a sustainable urban drainage system model. In IOP conference series: Earth and environmental science (Vol. 437, No. 1, p. 012018). IOP Publishing.
[2] Rani, D. S., Jayalakshmi, G. N., & Baligar, V. P. (2020, March). Low cost IoT based flood monitoring system using machine learning and neural networks: flood alerting and rainfall prediction. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 261-267). IEEE.
[3] Deng, F. (2020, November). Research on the applicability of weather forecast model—based on logistic regression and decision tree. In Journal of Physics: Conference Series (Vol. 1678, No. 1, p. 012110). IOP Publishing.
[4] Kareem, F. Q., Abdulazeez, A. M., & Hasan, D. A. (2021). Predicting Weather Forecasting State Based on Data Mining Classification Algorithms. Asian J. Res. Comput. Sci. AJRCOS, 9(3), 13-24.
[5] Primajaya, A., & Sari, B. N. (2018). Random forest algorithm for prediction of precipitation. Indonesian Journal of Artificial Intelligence and Data Mining, 1(1), 27-31.
[6] Ali, A., Khairan, A., Tempola, F., & Fuad, A. (2021). Application Of Naïve Bayes to Predict the Potential of Rain in Ternate City. In E3S Web of Conferences (Vol. 328, p. 04011). EDP Sciences.
[7] Tharwat, A. (2019). Parameter investigation of support vector machine classifier with kernel functions. Knowledge and Information Systems, 61(3), 1269-1302.
[8] Ghosh, S., Dasgupta, A., & Swetapadma, A. (2019, February). A study on support vector machine based linear and non-linear pattern classification. In 2019 International Conference on Intelligent Sustainable Systems (ICISS) (pp. 24-28). IEEE.
[9] Harimoorthy, K., & Thangavelu, M. (2021). Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. Journal of Ambient Intelligence and Humanized Computing, 12, 3715-3723.
[10] Chandran, V., Patil, C. K., Karthick, A., Ganeshaperumal, D., Rahim, R., & Ghosh, A. (2021). State of charge estimation of lithium-ion battery for electric vehicles using machine learning algorithms. World Electric Vehicle Journal, 12(1), 38.
[11] Putra, A. P., Debataraja, N. N., & Kusnandar, D. (2020). TINGKAT AKURASI KLASIFIKASI JARAK KELAHIRAN DI KAMPUNG KELUARGA BERENCANA (KB) DENGAN METODE SUPPORT VECTOR MACHINE (SVM). Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya, 9(3).
[12] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. Journal of Big Data, 8(1), 1-37.
[13] Khan, S. I., & Hoque, A. S. M. L. (2020). SICE: an improved missing data imputation technique. Journal of big Data, 7(1), 1-21.
[14] Chalid, A., & Prasetya, B. (2020, February). Utilization of a pond in East Jakarta for a sustainable urban drainage system model. In IOP conference series: Earth and environmental science (Vol. 437, No. 1, p. 012018). IOP Publishing.
[15] Dou, J., Yunus, A. P., Bui, D. T., Merghadi, A., Sahana, M., Zhu, Z., ... & Pham, B. T. (2020). Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan. Landslides, 17, 641-658.
[16] Tao, Z., Huiling, L., Wenwen, W., & Xia, Y. (2019). GA-SVM based feature selection and parameter optimization in hospitalization expense modeling. Applied soft computing, 75, 323-332.
[17] Rahaman, M. M., Li, C., Yao, Y., Kulwa, F., Rahman, M. A., Wang, Q., ... & Zhao, X. (2020). Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches. Journal of X-ray Science and Technology, 28(5), 821-839.
[18] Harita, U., Kumar, V. U., Sudarsa, D., Krishna, G. R., Basha, C. Z., & Kumar, B. S. S. (2020, November). A fundamental study on suicides and rainfall datasets using basic machine learning algorithms. In 2020 4th international conference on electronics, communication and aerospace technology (iceca) (pp. 1239-1243). IEEE.
[19] Shah, U., Garg, S., Sisodiya, N., Dube, N., & Sharma, S. (2018, December). Rainfall prediction: Accuracy enhancement using machine learning and forecasting techniques. In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 776-782). IEEE
[20] Hudnurkar, S., & Rayavarapu, N. (2022). Binary classification of rainfall time-series using machine learning algorithms. International Journal of Electrical and Computer Engineering, 12(2), 1945-1954.
[21] Pradana, M. R., Hardinata, R. S., & Syahputra, Z. (2022). ANALISA ALGORITMA SUPPORT VECTOR MACHINE PADA DATA BUNGA IRIS. Jurnal Darma Agung, 30(1), 477-487.