Rainfall Prediction with Support Vector Machines: A Case Study in Tanjungpinang City, Indonesia

Nurul Hayaty[1,*], *Hendra* Kurniawan[1], *Muhamad Radzi* Rathomi[1], *Ferdi* Chahyadi[1] and *Martaleli* Bettiza[1]

[1]Informatics, Faculty of Engineering and Maritime Technology, 29111 Tanjungpinang, Indonesia

Abstract. Rainfall forecasting is becoming more challenging due to extreme climate change. Especially for the archipelago which has a unique geography compared to the mainland. The aim of this study is to test the performance of the support vector machine in predicting rainfall in Tanjungpinang, Kepulauan Riau, Indonesia. The variables used to predict are temperature, humidity, wind speed, and rainfall. The results obtained is a precision value of 82% for rain, with a ROC curve evaluation score of 0.74. These results show that the model built has a fairly good ability to separate between positive and negative results in predicting rainfall. *

1 Introduction

One of the island regions in Indonesia is the Kepulauan Riau Province. Kepulauan Riau Province consists of 96% ocean and 4% land which is near the equator and has more than 2,408 islands [1]. Archipelagic regions close to the equator are vulnerable to monsoons and tropical storms. Tropical storms can bring heavy rain, strong winds, and high ocean waves, causing serious damage.

Rainfall forecasting is particularly important in island areas for several reasons. Firstly, islands often have limited water resources and rely heavily on rainfall for freshwater supply. Accurate rainfall forecasts are crucial for proper water management and ensuring sufficient water supply for both human consumption and agriculture. Additionally, islands are more vulnerable to the impacts of extreme weather events, such as tropical storms and hurricanes. Rainfall forecasting in these areas plays a critical role in enabling early warning systems and facilitating evacuation plans to minimize potential loss of life and infrastructure damage. Furthermore, the accuracy of rainfall forecasting in island areas is vital for industries such as tourism and outdoor event planning. and important in transportation that links the islands, as it helps to anticipate any disruptions or safety concerns that may arise due to heavy rainfall or storms. Rainfall forecasts are especially challenging in island areas due to the complex terrain and unique atmospheric conditions.

SVM has been shown to be superior to other models in predicting extreme rainfall and has been successfully applied in various areas such as groundwater level prediction, flood

* Corresponding author: nurul.hayaty@umrah.ac.id

stage prediction, and river flow forecasting [2]. SVM has also been employed in hydrological modeling for streamflow forecasting, rainfall-runoff modeling, and forecasting of reservoir

inflows. In island areas, accurate rainfall forecasting is particularly important due to the potential impact on water resources, agriculture, and infrastructure. However, forecasting rainfall in island areas can be challenging due to unique geographical characteristics that can influence weather patterns, such as the influence of ocean currents and mountain ranges. Therefore, developing a rainfall forecasting model using Support Vector Machine specifically designed for island areas can greatly improve the accuracy and reliability of rainfall predictions. By incorporating data from local meteorological stations, the SVM model can effectively capture the complex interactions that contribute to rainfall variability in these regions [3][4]. This can help local authorities and communities make informed decisions regarding water management, agricultural planning, disaster preparedness, and infrastructure development. Overall, the utilization of SVM in rainfall forecasting for island areas can provide significant benefits in terms of improving the accuracy and reliability of rainfall predictions, enabling better resource management and decision-making in these vulnerable environments. In conclusion, the use of Support Vector Machine in rainfall forecasting for island areas has shown promise in improving the accuracy and reliability of rainfall predictions in these regions, which is crucial for effective water resource management, agricultural planning, infrastructure development, and disaster preparedness.

2 Literature Review

Support Vector Machines (SVMs) are a class of supervised machine learning algorithms that have gained significant attention and popularity in various fields since their introduction in the 1990s. SVMs are primarily used for classification and regression tasks, and they have proven to be effective in a wide range of applications due to their ability to handle high-dimensional data and nonlinear relationships. This literature review provides an overview of key concepts, developments, and applications of Support Vector Machines in the years leading up to 2023.

Support Vector Machine (SVM) is a powerful machine learning technique used for a wide range of applications, including weather prediction and rainfall forecasting. Support Vector Machines have remained a prominent machine learning technique, showcasing versatility and effectiveness across diverse applications [5]. The reviewed journal sources highlight SVM's role in addressing complex challenges, from big data classification [6] to healthcare [7] and renewable energy forecasting [8][9]. As SVM continues to evolve, its adaptability and robustness make it a valuable tool in the data scientist's arsenal, with ongoing research aiming to further enhance its capabilities.

Support Vector Machine has proven to be a valuable tool in weather prediction and rainfall forecasting [10][11] The reviewed journal sources illustrate its diverse applications, from short-term weather forecasting to climate change impact assessment. Combining SVM with other techniques [12] and integrating satellite and climate data further enhances its capabilities in addressing weather-related challenges. Researchers continue to explore innovative approaches to leverage SVM for improved accuracy and early detection of extreme weather events.

3 Methodology

The steps used to achieve the objectives in this study are:

1. Preparation
2. Data collection

3. Data preprocessing
4. Data training

3.1 Preparation

This study uses tools to assist in processing data such as Google Colab, Panda, Numpy, Keras, Matplotlib, and Tensorflow. Google colab is a cloud-based computing platform provided by Google. It is specifically designed for the development and training of machine learning and deep learning models, as well as for data analysing using a Python-based environment. Panda is a library used for data analysis and powerful functions for working with structured data. Numpy is a library that provides multidimensional array support and various mathematical functions. Keras is a deep learning library that provides a high-level interface for building, training, and evaluating neural networks models. Matplotlib is a library for data visualization in the form of graphs and plots. Tensorflow is a library that provides various tools and frameworks for developing, training, and evaluating machine learning models.

3.2 Data Collection

The data used in this study were taken at the Raja Haji Fi Sabilillah meteorological and geophysical station for Tanjungpinang City, Riau Islands, Indonesia. The data collected are data from January 2021 to January 2023, totaling 758 data. The attributes used are temperature (°C), humidity (%), wind speed (m/s), and rainfall (in) which play a vital role in weather. There is a small amount of empty data but this is accommodated by the large data size. The data will be divided into 530 training data and 228 testing data.

Table 1. Input parameter

Date	Temp (°C)	Hum (%)	Wind Speed (m/s)	Rainfall (in)
01-01-2021	25.9	91	5	5.2
02-01-2021	23.5	96	2	282.6
03-01-2021	24.8	95	2	201.4
04-01-2021	25	94	2	3.7
05-01-2021	26.7	87	3	1.95
06-01-2021	26.3	87	3	0.2
07-01-2021	24.9	92	3	4
08-01-2021	25.2	94	2	57.6
09-01-2021	25.1	95	4	28.9

3.3 Data Preprocessing

A crucial first step is data preprocessing. Data preprocessing helps to clean the data from errors, missing values, and outliers. It improves the quality of the data, which in turn can lead to better models. Data preprocessing makes it possible to transform data into a form that is more suitable for machine learning algorithms. This includes normalization, standardization,

categorical variable coding, and more. In addition, it can help in the selection of the most relevant features for building models. This can reduce data dimensionality and improve computational efficiency.

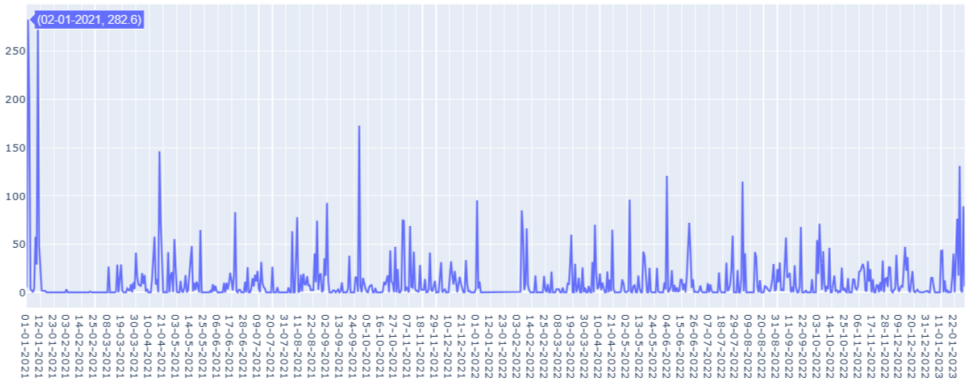

Fig.1. Rainfall in inch

3.3 Training Data

In using the SVM technique, it tries to find the optimal classifier to separate two different classes. It tries to find the best dividing function (hyperplane) among an unconstrained number of functions to divide two types of objects. A good hyperplane is one that lies in the middle between two sets of sobjects from two classes. The definition of the equation on the separating hyperplane can be written in (1)[7].

$$W \cdot X + b = 0 \tag{1}$$

W is the weight of a vector, i.e. $W = \{w_1, w_2, w_3, \dots, w_n\}$; where n is the number of attributes and b is a scalar value or often called bias. The data will be processed to see if the SVM model built is compatible and meets the convergent aspects of a method/algorithm. The learning model is built using the SVC or Support Vector Classifier model. SVC is a form of SVM used for classification problems. It works by finding a hyperplane (line or surface) that separates two classes of data in such a way that the margin around the hyperplane is maximized. SVC tries to find the hyperplane that best separates the data and has a larger margin.

4 Results

After analyzing the model that was built, the results can be seen in the table 2. The table shows the highest result with a precision value for rain prediction of 0.82 or 82%. Recall on the model prediction of non-rain data is 0.71 and rain is 0.77. F1 and Support values are the number of testing data used for prediction. In the rainfall dataset, the data for "no rain" is 82 and for "rain" is 146. We use confusion matrix to evaluate the performance of the prediction model. It contains the actual value and the predicted value. Researchers use this value to calculate the accuracy of the model. The accuracy value of the model is 0.75 or 75%, visualization with a heat map can be seen in the figure 2.

Table 2. Results

Report	Precision	Recall	F1-Score	Support
No Rain	0.64	0.71	0.67	82
Rain	**0.82**	0.77	0.80	146
Accuracy			0.75	228
Macro AVG	0.73	0.74	0.75	228
Weighted AVF	0.76	0.75	0.75	228

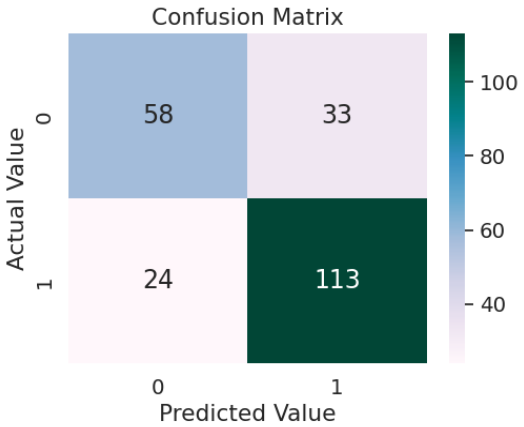

Fig. 2. Confusion matrix of prediction model

Based on this, it can be predicted that there will be 58 days without rain and 113 days with rain. The model also has a prediction error of 33 non-rainy days and 24 rainy days in one year. The 75% accuracy shows that the model performs well for testing the dataset.

Evaluation of classification model performance using the ROC Curve. The ROC Curve, or Receiver Operating Characteristic Curve, is a visual tool used in the evaluation of classification model performance, especially in the context of machine learning and statistics. The ROC Curve illustrates the relationship between two important metrics: the True Positive rate (Sensitivity) and the False Positive rate (1-Specificity) at various threshold values used to classify the data. Model evaluation conducted using the ROC curve shows an area value of 0.74. This indicates that the model has adequate performance in distinguishing between positive and negative categories.

So, in the context of the ROC curve, a value of 0.74 indicates that the model has a fairly good ability to separate the different classes, but not quite close to perfection (a value of 1). This value can be considered a positive indication in most cases. ROC curve prediction model can be seen in figure 3.

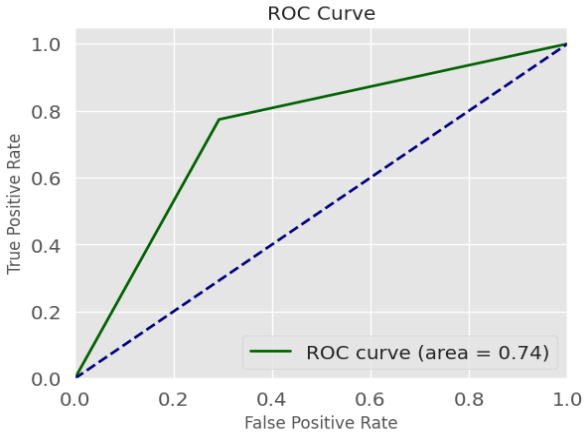

Figure 3. ROC curve of prediction model

5 Conclusion

The Support vector machine prediction model has been successfully built with a result of 0.82 or 82% for the precision value in the rain category. ROC curve was used to evaluate the model. The result shows 0.74 which indicates a fairly good performance in distinguishing between 2 classes.

References

1. Pemprov Kepri. Tentang Kepri. Retrieved October 19, 2023, from https://kepriprov.go.id/laman/tentang-kepri
2. Pham, Q. B., Yang, T. C., Kuo, C. M., Tseng, H. W., & Yu, P. S. Combing Random Forest and Least Square Support Vector Regression for Improving Extreme Rainfall Downscaling. Water, **11**, 3 (2019). https://doi.org/10.3390/W11030451.
3. Chen, S. T. Mining Informative Hydrologic Data by Using Support Vector Machines and Elucidating Mined Data according to Information Entropy. Entropy, **17**, 3, 1023–1041 (2015). https://doi.org/10.3390/E17031023
4. Li, G., Sun, Y., He, Y., Li, X., & Tu, Q. Short-Term Power Generation Energy Forecasting Model for Small Hydropower Stations Using GA-SVM. *Mathematical Problems in Engineering* (2014). https://doi.org/10.1155/2014/381387.
5. Bochenek, B., & Ustrnul, Z. Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives. Atmosphere, **13**, 2 (2022). https://doi.org/10.3390/atmos13020180.
6. Mohammad, R. M. A. An Enhanced Multiclass Support Vector Machine Model and its Application to Classifying File Systems Affected by a Digital Crime. Journal of King Saud University - Computer and Information Sciences, **34**, 2, 179–190 (2022). https://doi.org/10.1016/J.JKSUCI.2019.10.010.
7. Shi, B., Ye, H., Heidari, A. A., Zheng, L., Hu, Z., Chen, H., Turabieh, H., Mafarja, M., & Wu, P. Analysis of COVID-19 severity from the perspective of coagulation index using evolutionary machine learning with enhanced brain storm optimization. Journal of King Saud University - Computer and Information Sciences, **34**, 8, 4874–4887 (2022). https://doi.org/10.1016/J.JKSUCI.2021.09.019
8. Zendehboudi, A., Baseer, M. A., & Saidur, R. Application of support vector machine models for forecasting solar and wind energy resources: A review. Journal

of Cleaner Production, **199**, 272–285 (2018). https://doi.org/10.1016/J.JCLEPRO.2018.07.164.

9. Xue, J., Cai, D., & Zhou, G. Application of support vector machines in photovoltaic power prediction. *Proceedings - 2022 14th International Conference on Intelligent Human-Machine Systems and Cybernetics* (IHMSC 2022), 56–59. https://doi.org/10.1109/IHMSC55436.2022.00022.
10. Yin, G., Yoshikane, T., Yamamoto, K., Kubota, T., & Yoshimura, K. A support vector machine-based method for improving real-time hourly precipitation forecast in Japan. Journal of Hydrology, **612**, 128125 (2022). https://doi.org/10.1016/j.jhydrol.2022.128125.
11. Hussein, E., Ghaziasgar, M., & Thron, C. Regional rainfall prediction using support vector machine classification of large-scale precipitation maps. *Proceedings of 2020 23rd International Conference on Information Fusion* (FUSION 2020). https://doi.org/10.23919/FUSION45008.2020.9190285.
12. Zhu, Y., Zhao, Y., Zhang, J., Geng, N., & Huang, D. Spring onion seed demand forecasting using a hybrid Holt-Winters and support vector machine model. PLoS ONE, **14**, 7 (2019). https://doi.org/10.1371/journal.pone.0219889.