Comparative Analysis of Naïve Bayes Classifier, Support Vector Machine and Decision Tree in Rainfall Classification Using Confusion Matrix

Elvira Vidya Berliana, Mardhani Riasetiawan*
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia

***Abstract*—The climate in Indonesia is sometimes unstable to this day. This unstable climate change will cause difficulties in predicting rainfall conditions. With unstable climate change, an algorithm is needed that helps the public predict rainfall conditions using rainfall, temperature and humidity parameters. The research process uses daily climate data from the Indonesia Climatology Agency with time span 2018 – 2023. The classification system using the Naïve Bayes Classifier (NBC) algorithm is less able to capture complexity and complex feature interactions with an accuracy of 97%-98%, Support Vector Machine (SVM) has an accuracy of 92%-94% and fewer prediction errors than NBC and Decision Tree which experienced overfitting especially when testing sets with 50% data with an accuracy of 99%-100%. Even though the Decision Tree shows the best performance, there is still a risk of overfitting so, SVM is a stable choice in this research.**

Keywords*—Naïve Bayes Classifier (NBC); Support Vector Machine (SVM); decision tree; confusion matrix; classification; rainfall; temperature; humidity

I. INTRODUCTION

The climate in Indonesia is sometimes unstable now. This unstable climate change will cause difficulties when predicting rainfall conditions. According to the Indonesian Meteorology, Climatology and Geophysics Agency (BMKG), the average air temperature in July in Indonesia for the period 1981 – 2010 was 26.39°C. In 2020, the average air temperature in February was 27.22°C so the anomalous increase in average air temperature was 0.83°C [1]. Rainfall predictions are very important because good rainfall predictions will avoid many disasters and accidents. Unpredictable rainfall can cause crop damage, major floods and droughts, ultimately exploiting animal, plant and human life [2].

In agriculture, accurate rainfall predictions will help farmers to plan their agricultural activities. Until now, farmers still carry out planting activities based on their intuition [3]. Machine learning and deep learning algorithms have emerged as powerful tools for analyzing vast amounts of data from various sources, including satellite imagery and atmospheric conditions, to enhance rainfall prediction accuracy [4]. The agriculture sector benefits from more accurate rainfall forecast, as they help mitigate the effects of abnormal precipitation on crop cultivation and influence decisions regarding planting, harvesting, and agricultural inputs [5].

Machine learning tries to process observational data and then gets results, namely weather patterns, which in turn can help analyze rainfall which often changes so that it can make more accurate rainfall predictions [6]. Several studies on machine learning for optimal classification have been carried out. Research by Fallo (2021) [7] utilized the linear kernel SVM method, NBC, and Ordinal Logistic Regression with SVM accuracy results of 67.99%, NBC with accuracy results of 69.63%, and Ordinal Logistic Regression with accuracy results of 69.63 %. Azmi et al. (2021) [8] achieved 96% accuracy using NBC for rainfall classification in Banyuwangi, Indonesia. Husain H., Dawoodi, and Patil (2023) [9] found SVM to be the best model for rainfall prediction in North Maharashtra, India, with 93% accuracy. Sivanantham et al. (2023) [10] compared multiple classification algorithms for rainfall prediction in Indian States, including Logistic Regression, Decision Tree, Random Forest, and SVM. These studies demonstrate the potential of machine learning techniques in improving rainfall classification and prediction accuracy.

The Climatology Station Indonesian Meteorology, Climatology and Geophysics Agency has historical weather data but the data is still small in quantity and there is data that is less accurate because data takers still depend on historical data so there are still errors and incomplete data, and this causes rainfall predictions to be less accurate. Therefore, the researcher is interested in developing previous research related to rainfall classification using the Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), Decision Tree algorithm.

This paper aims to provide a comprehensive analysis of three machine learning algorithms, Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), and Decision Tree, in the context of rainfall classification. The study will evaluate the performance of these algorithms using confusion matrix to determine and evaluate their accuracy and reliability.

II. RESEARCH METHODS

A. Problem Analysis

The system design for this research aims to predict rainfall for agricultural activities using the Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), and Decision Tree methods and data taken from the DI Yogyakarta Climatology Station (dataonline.bmkg.go.id) to create and train a model to make accurate predictions. The research process used daily climate data from Indonesia Climatology Agency (BMKG) with time span 2018 – 2023 by using three parameters: rainfall, temperature and humidity.

B. Research Step

In conducting this research, several stages are required to achieve the research objectives. Starting with determining environmental system requirements such as research methods, data used, and measurement methods. After knowing the system analysis, you can obtain weather parameter data, then form a Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), Decision Tree model architecture. The desired model parameters are high accuracy and produce model performance measurement parameters in accordance with Fig. 1.

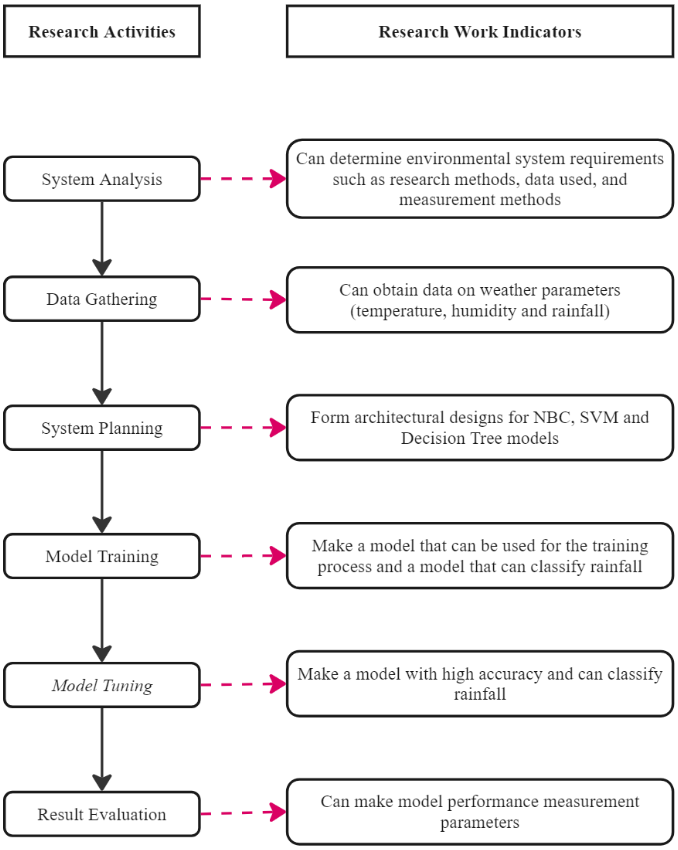


Fig. 1. Research steps flowchart.

C. System Workflow

System workflow describes the whole system. There are several stages so that the system can work well. The first stage that must be carried out is data loading, this process is the stage for adding the dataset to the algorithm system. Then, the data preprocessing stage includes classifying and labelling the dataset. The next process is Exploratory Data Analysis (EDA) and Data Preparation, this stage is the stage to understand the nature of the dataset. Then, the most crucial stage is data classification using the Naïve Bayes Classifier, Support Vector Machine and Decision Tree. Then, model evaluated using the confusion matrix and the final stage is testing the results of the three algorithms. All stages are visualized as a flow diagram as in Fig. 2.

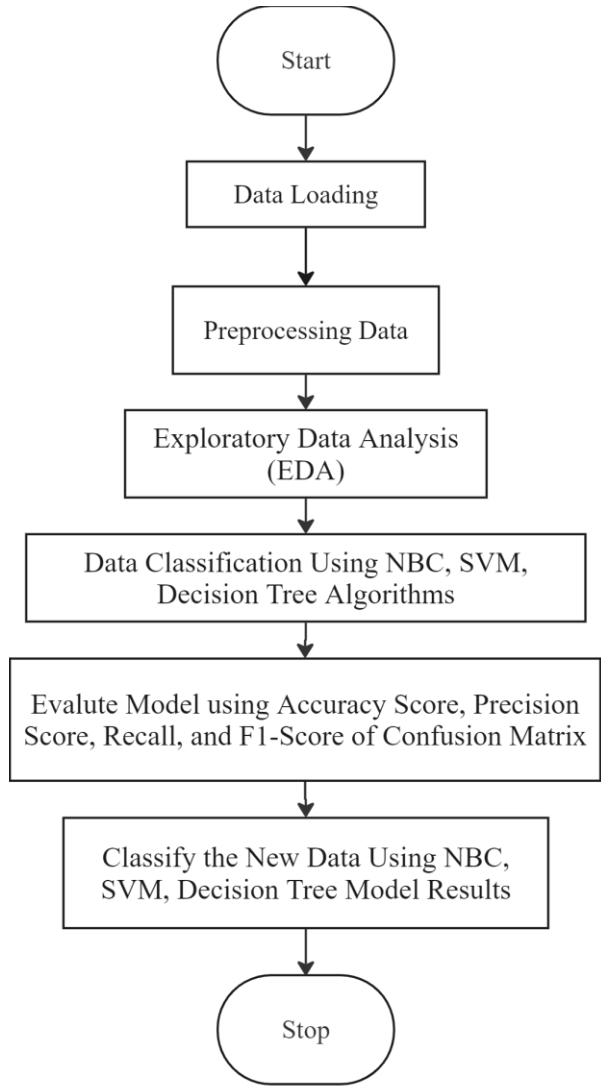


Fig. 2. System workflow.

D. Preprocessing Data

After the data loading process, the next stage is data preprocessing with labelling the frequency of rainfall into different labels based on specified ranges. The labelling method checks the frequency of rainfall in increasing ranges and assigns a corresponding label based on where the frequency falls. This structured approach helps in categorizing rainfall data efficiently and those processes are visualized as in Fig 3.

E. Data Description

The data used is open-source data from the BMKG page using rainfall, humidity and temperature parameters with a time span from 2018 – 2023. And the following is the classification of rainfall according to Indonesian Meteorology, Climatology and Geophysics Agency (BMKG):

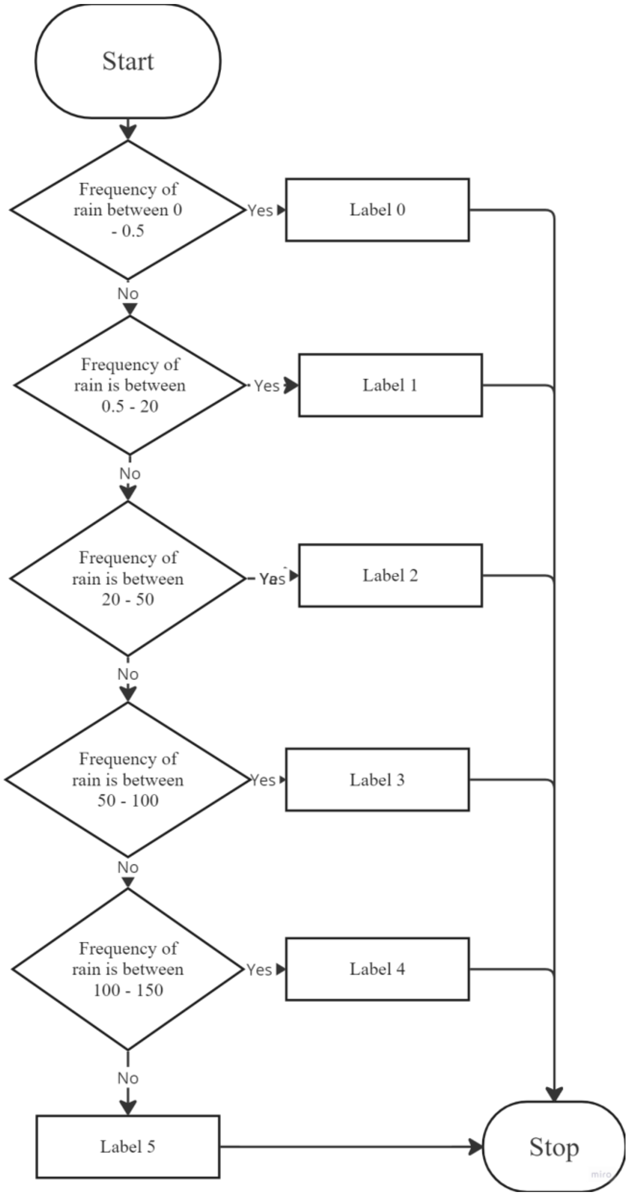


Fig. 3. Pre-processing data.

TABLE I. RAINFALL CLASSIFICATION ACCORDING BMKG

No	Rainfall Intensity	Classification
1.	0 mm/day	Cloudy
2.	0.5 – 20 mm/day	Light Rain
3.	20 – 50 mm/day	Moderate Rain
4.	50 – 100 mm/day	Heavy Rain
5.	100 – 150 mm/day	Very Heavy Rain
6.	>150 mm/day	Extreme Rain

F. *Confusion Matrix*

The confusion matrix is widely used in machine learning for evaluating classification model performance [11]. It compares predicted and actual class labels, typically in a tabular format [12]. While traditionally used for binary classification, confusion matrix can be extended to multiclass problems [13]. The term "confusion" in the matrix name refers to how the model may confuse or mislabel classes [11]. In credit scoring, the confusion matrix is an essential measure of model accuracy, with 16 possible variants, of which only eight are considered reasonable (Zeng, 2019) [14]. The matrix's entries typically include true negatives, false positives, false negatives, and true positives (Piegorsch, 2020) [15]. Confusion matrix calculated the F1-Score, a widely used measure of classification accuracy, combines precision and recall using a harmonic mean (Yadavendra & Chand, 2020) [16].

G. *Naïve Bayes Classifier*

Gaussian Naïve Bayes (GNB) Classifier is a popular machine learning technique that assumes conditional independence between features. While efficient, this assumption can be limiting (Ali Haghpanah Jahromi & Taheri, 2017) [17]. Gaussian Naïve Bayes (GNB) algorithm has been successfully applied to rainfall classification and prediction in various studies. It has shown high accuracy in classifying daily and monthly rainfall patterns (Indra Kusuma Yoga et al., 2022) [18] and in predicting drought and flood risks based on multiple atmospheric variables (Oluwatobi Aiyelokun et al., 2020) [19]. Research has also shown that the Naïve Bayes Classifier can effectively classify rainfall based on air temperature and wind speed (Meilani Nisa Abdilla et al., 2024) [20].

H. *Support Vector Machine*

Support Vector Machines are widely used for classification tasks, with linear SVM being particularly effective for linearly separable classes (Murty & Raghava, 2016) [21]. The choice of kernel function is crucial for SVM performance, with linear kernels often outperforming others in certain applications (Keumala Intan, 2019) [22]. Support Vector Machine (SVM) have been developed for rainfall classification and prediction tasks with varying kernel functions. Linear kernel SVM have shown effectiveness in modeling pore-water pressure responses to rainfall, achieving high accuracy while offering computational efficiency (K. Yusof et al., 2017) [23]. In rainfall classification, linear and polynomial kernels demonstrated superior performance (78.38% accuracy) compared to Gaussian kernels when using a 90:10 training-testing split (Novia Pratiwi & Yudi Setyawan, 2021) [24]. SVM with linear and RBF kernels have been used to classify rainfall as heavy or light based on temperature and humidity data (S. Sunori et al., 2021) [25]. for rainfall prediction, a comparative study of different kernels revealed that the linear kernel produced the lowest average mean square error (15.04%) on test data, outperforming polynomial, RBF, and sigmoid kernels (J. Mohanty & M. Mohapatra, 2018) [26].

I. *Decision Tree*

Decision trees are popular classification algorithms in data mining, utilizing a divide-and-conquer strategy to create a flowchart-like structure (Dai et al., 2016) [27]. The tree consists of internal nodes representing attribute tests, branches denoting test outcomes, and leaf nodes holding class labels (Sharma & Kumar, 2016) [28]. Classification involves traversing tree from root to leaf, with the leaf node indicating the final classification (Dai et al., 2016) [27]. Decision trees are widely applicable in various fields, including rainfall classification (see Table I). Decision trees outperform their effectiveness in binary classification of rainfall events (Manoj Chhetri & Lily Gurung, 2023) [29] and long-term rainfall prediction (B. Revathi et al., 2021) [30]. Their popularity stems from their ability to handle large, complex datasets and extract useful knowledge from incomplete or noisy data (Sharma & Kumar, 2016) [28].

J. System Analysis Methods

This paper uses NBC, SVM and Decision Tree classification analysis methods. After carrying out the classification process, the next thing is to evaluate using the Confusion Matrix. After getting the table, this table will be used to find the accuracy value, recall value, precision value, and F1-Score for three models. The next stage is evaluating the model by using learning curve of each model. The final in the data analysis method is to check whether each model is performing well, underfitting, or overfitting as shown in Fig. 4.

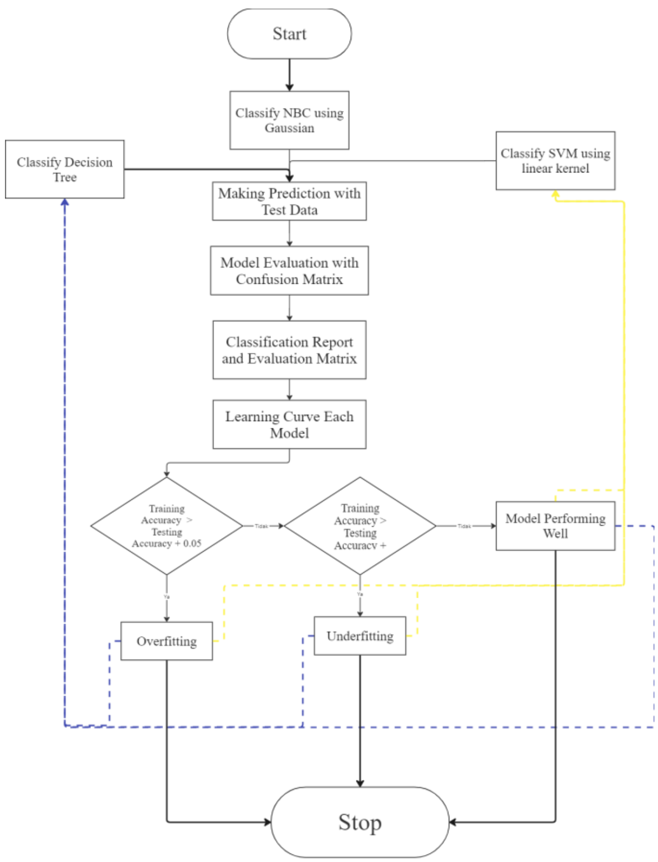


Fig. 4. NBC, SVM and decision tree model process.

K. Classification System Results

System testing was carried out using new data consisting of three parameters: rainfall, temperature and humidity. And the output of this test is a classification of rainfall based on the new parameters that have been entered.

III. Results and Discussion

The system built is a system to classify the rainfall in Yogyakarta, Indonesia with applied three methods in machine learning, they are Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), and Decision Tree. While the programming language used in this paper is Python.

A. Naïve Bayes System Evaluation Results

In the 50% test set with 98% accuracy, in class 0 most of the data (666 out of 677) was correctly classified as class 0 with 11 data incorrectly classified as class 1, in class 1 classified 307 data correctly and 7 data incorrectly classified as class 2, in class 3 most of the data (24 of 25) were classified correctly and 1 data was misclassified as class 4, and the data in class 4 were all classified correctly as in Fig. 5.

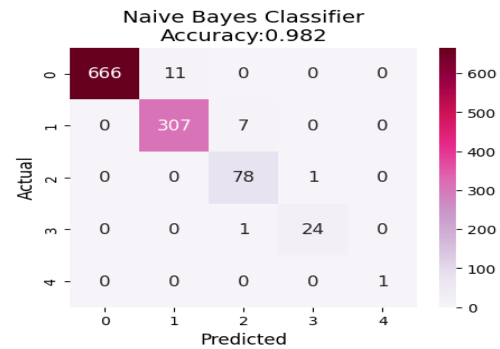


Fig. 5. Heatmap of NBC confusion matrix result on 50% test set.

In the 20% test set with 98% accuracy, class 0 in this test set most of the data (256 out of 259) is correctly classified as class 0 and 3 data is incorrectly classified as class 1, in class 0 124 data are correctly classified as class 1 and 3 incorrect data was classified as class 2, all data (44 data) in class 2 were classified correctly, and in class 4 some of the data was classified correctly and 1 data was incorrectly classified as class 2 as in Fig. 6.

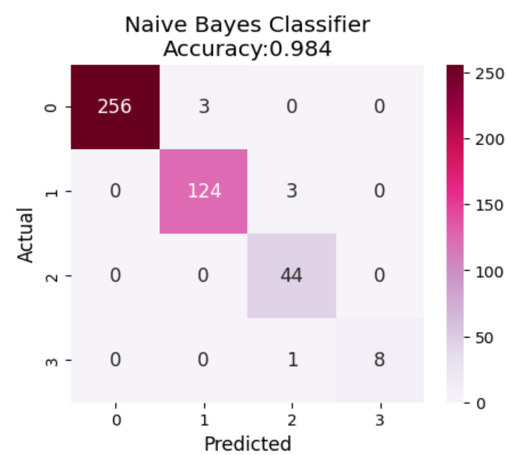


Fig. 6. Heatmap of NBC confusion matrix result on 20% test set.

In the 10% test set with 97% accuracy, class 0 in this test set most of the data (113 out of 114) is classified correctly and 1 data is incorrectly classified as class 1, in class 1 71 data are classified correctly but 4 data are incorrectly classified as class 2, all data (25 data) in class 2 and all data (6 data) in class 3 are classified correctly as in Fig. 7.

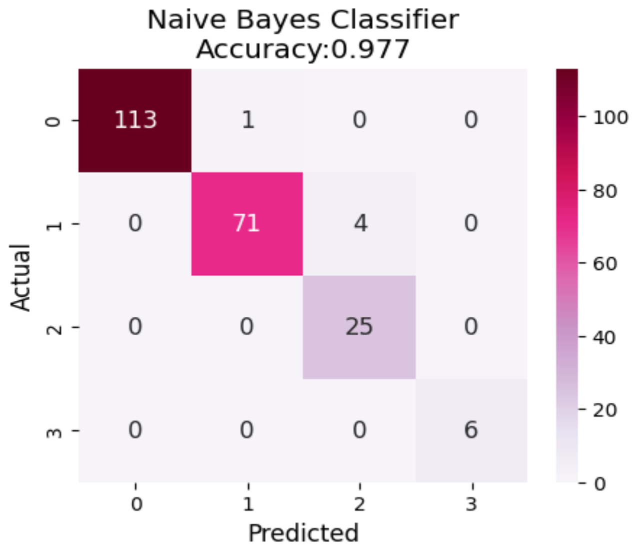
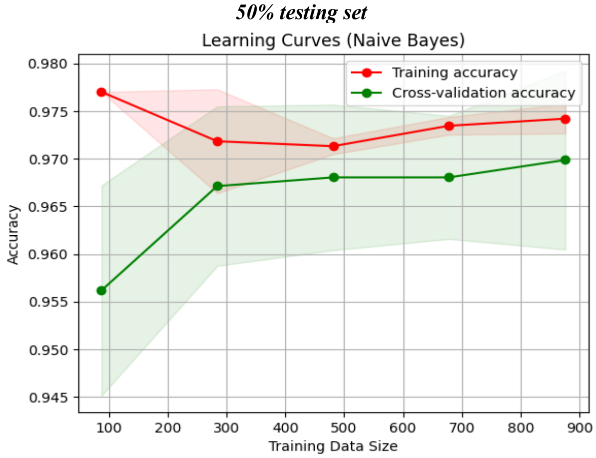


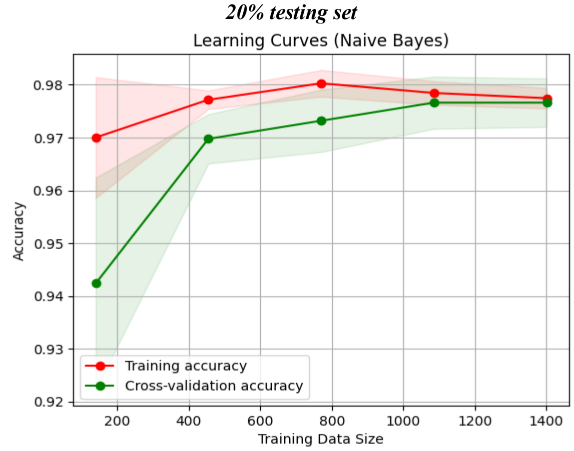
Fig. 7. Heatmap of NBC confusion matrix result on 10% test set.

Referring to the NBC heatmap results above, this algorithm has a fairly good level of accuracy, but there are still prediction errors in class 0 and class 1 and these prediction errors tend to occur in adjacent classes, indicating that NBC still faces challenges in distinguishing classes with the same characteristics.

50% testing set



20% testing set



10% testing set

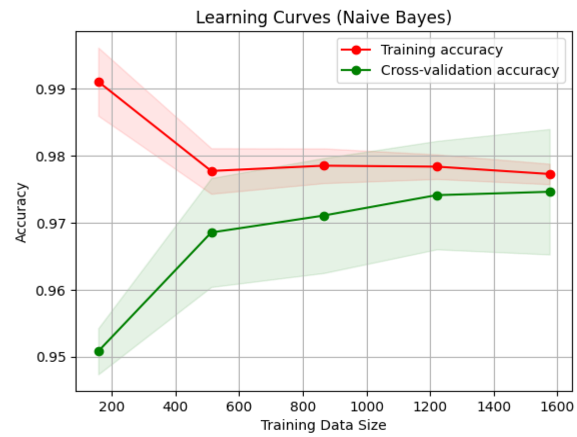


Fig. 8. The learning curve of Naïve Bayes classifier model.

As shown in Fig. 8, the cross-validation curve of 20% test set has increased too quickly indicating that the model has difficulty capturing the complexity of the data so 600 – 800 data is enough to achieve optimal NBC performance.

B. Support Vector Machine System Evaluation Results

In the 50% test set with 94% accuracy, all data (677 out of 677) in class 0 are predicted correctly. In class 1, most of the data (260 out of 314) was predicted correctly but there were 54 data that were incorrectly predicted as class 0. In class 2, 72 data were correctly predicted as class 2 and 7 data were incorrectly predicted as class 1. In class 3, 24 data were predicted correctly while 1 data was incorrectly predicted as class 2. In class 4, all data were predicted correctly as class 4 as in Fig. 9.

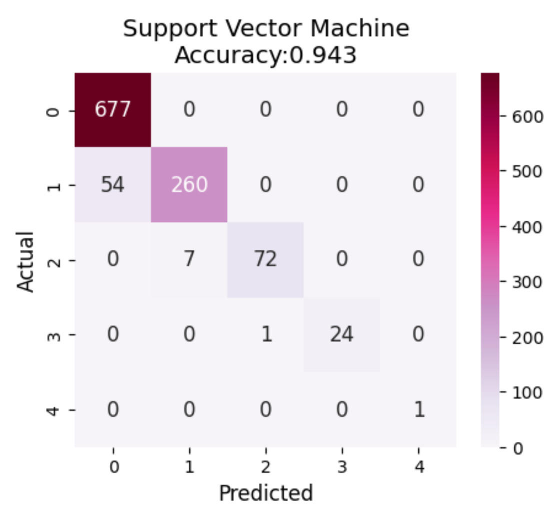


Fig. 9. Heatmap of SVM confusion matrix result on 50% test set.

In the 20% test set with 94% accuracy, all data (259 out of 259) in class 0 are predicted correctly. Predictions in class 1 were 109 out of 127 data predicted correctly and 18 data incorrectly predicted as class 0. In class 2, 40 out of 44 data were predicted correctly and 4 data were incorrectly predicted as class 1. Meanwhile in class 3, 8 data was predicted correctly and 1 data that was incorrectly predicted as class 2 as in Fig. 10.

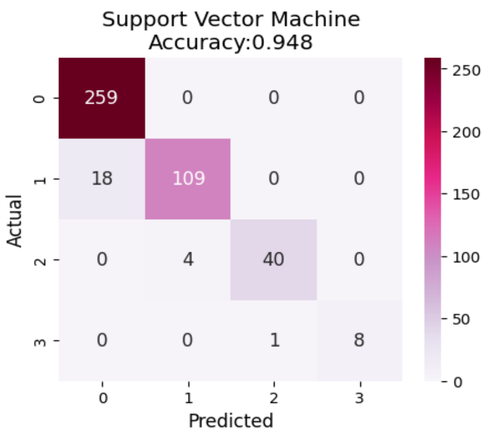


Fig. 10. Heatmap of SVM confusion matrix result on 20% test set.

In the 10% test se with 96% accuracy, all data (114 out of 114) in class 0 are predicted correctly. In class 1, some data (68 out of 75) were predicted correctly, and 7 data were incorrectly predicted as class 0. In class 2, 24 data were predicted correctly, and 1 data was incorrectly predicted as class 1. Meanwhile, in class 3 all data (6 of 6) were predicted correctly as in Fig. 11.

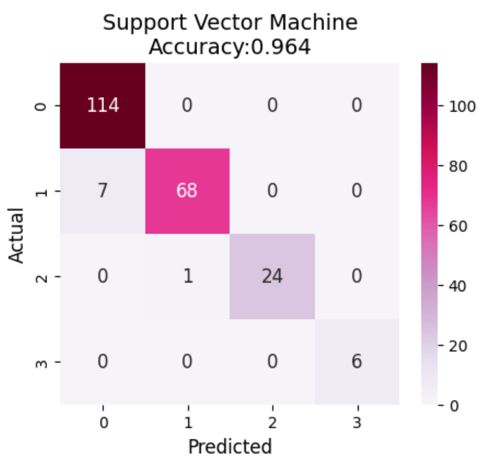
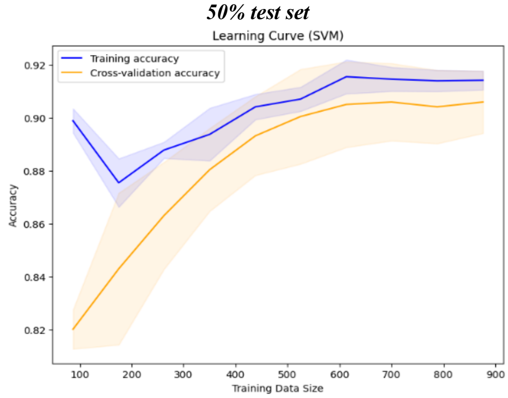


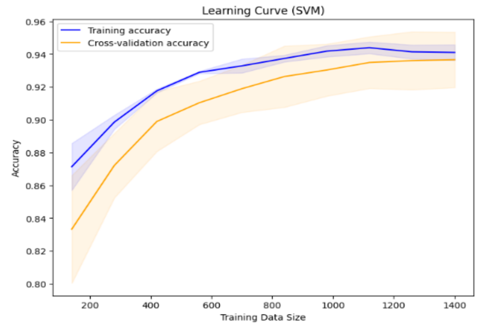
Fig. 11. Heatmap of SVM confusion matrix result on 10% test set.

All SVM models in the three test sets have an accuracy above 94%, but there are still errors in class 1 and class 2 classification, however, this SVM model is very good at classifying data from class 0 very accurately in all test sets.

50% test set



20% test set



10% test set

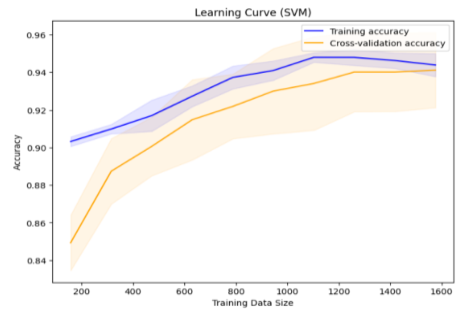


Fig. 12. The learning curve of support vector machine model.

In the SVM learning curve, the three test sets show that the model has good abilities in learning data as seen from the training accuracy and cross-validation accuracy which increase and decrease the distance from each other as shown in Fig. 12.

C. Decision Tree Evaluation Results

In the 10% test set, all data (677 out of 677) in class 0 are predicted correctly. In class 1, 313 of 314 data were predicted correctly and 1 data was incorrectly predicted as class 5. In class 2, class 3, and class 4, all data were predicted accurately as in Fig. 13.

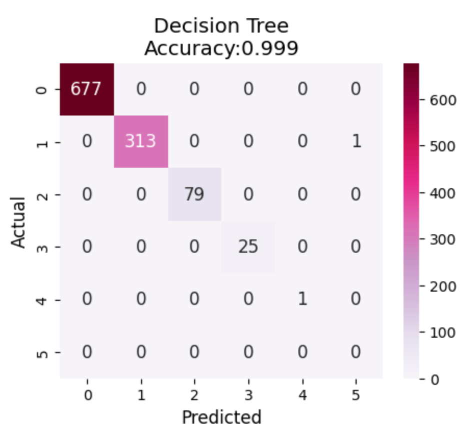


Fig. 13. Heatmap of decision tree confusion matrix result on 50% test set.

In the 20% test set, all data from all classes (class 0 with 259 data, class 1 with 127 data, class 2 with 44 data, and class 3 with 9 data) were predicted correctly and accurately as in Fig. 14.

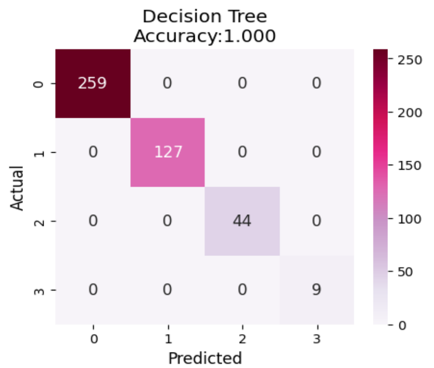


Fig. 14. Heatmap of decision tree confusion matrix result on 20% test set.

Likewise, in the 10% test set, all data from all classes (class 0 with 114 data, class 1 with 75 data, class 2 with 25 data, and class 3 with 6 data) were predicted correctly as in Fig. 15.

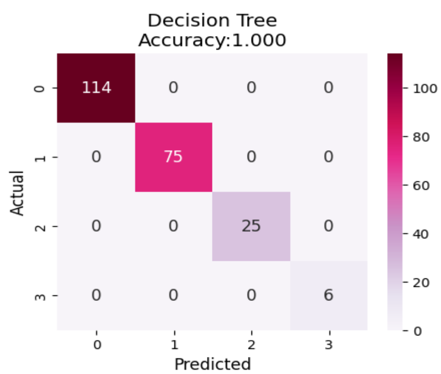
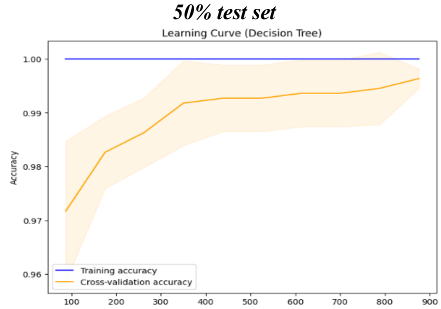


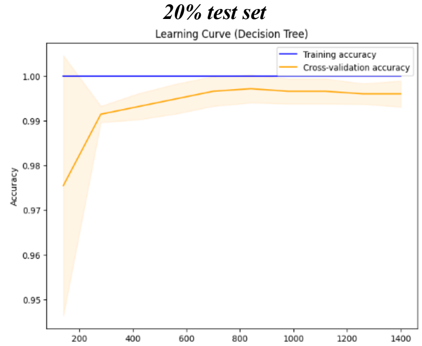
Fig. 15. Heatmap of decision tree confusion matrix result on 10% test set.

The Decision Tree model has 100% accuracy and the lowest prediction error among the three algorithms.

50% test set



20% test set



10% test set

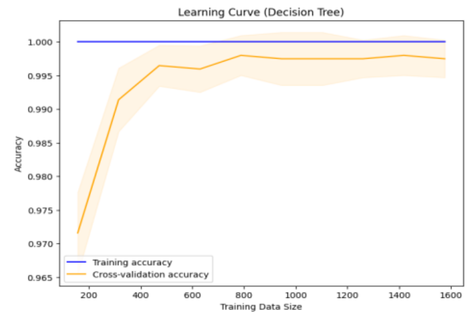


Fig. 16. Learning curve of decision tree model.

There is overfitting of the three test sets, especially in the 50% test set, it is proven that there is quite a large gap between testing accuracy and cross-validation accuracy as shown in Fig. 16.

IV. Conclusion and Suggestion

A. Conclusion

Based on research on rainfall classification using the Naïve Bayes Classifier (NBC), Support Vector Machine (SVM) and Decision Tree algorithms. The conclusion that can be drawn is that NBC accuracy is 98%, SVM accuracy is 94%, Decision Tree accuracy is 100%. NBC has quite high accuracy but still experiences difficulty in distinguishing classes, especially class 0 and class 1, SVM has good performance in classifying class data, even though there are errors in class 1 and class 2, and even though Decision Tree has perfect accuracy, however This happens due to overfitting, especially on a large test set (50% test set), so Support Vector Machine (SVM) becomes a stable choice among the three models.

B. Suggestion

Suggestions given from research that have been carried out for further research are improving the three models, such as reducing the dataset to 600-800 for Naïve Bayes Classifier (NBC), exploring other than linear kernels in Support Vector Machine (SVM), and using pruning techniques in Decision Tree.

References

[1] Sudirman, H. “Average air temperature anomalies in 2023.” [Online]. Available: https://www.bmkg.go.id/berita/?p=anomali-suhu-udara-rata-rata-tahun-2023&lang=ID&tag=anomali-suhu-udara. 2024

[2] Chaudhary, H., Mishra, U., Gupta, A., & Singh, A. (2022). “Comparative analysis of rainfall prediction using machine learning and deep learning techniques.” *In 2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)* (pp. 1-6). IEEE.

[3] Dhenanta, R. P., & Kholifah, I. B. (2022). “Prediction of monthly rainfall for Trenggalek district in 2022 and 2023 using the ARIMA method.” *In National Seminar on Official Statistics (Vol. 2022, No. 1, pp. 1135-1144).*

[4] Kothai, G., Sushmitha, Y., Saranya, K. L., Sri, P. N. R., & Amulya, P. (2023). Rainfall Prediction Using Deep Learning and Machine Learning Techniques. In 2023 *International Conference on Advances in Computing, Communication and Applied Informatics* (ACCAI) (pp. 1-7). IEEE.

[5] Bhatnagar, P. (2023). Machine Learning-based Rainfall Prediction on Indian Agriculture Land. In 2023 *IEEE International Conference on ICT in Business Industry & Government* (ICTBIG) (pp. 1-5). IEEE.

[6] Supeno, H. (2019). "14th machine learning meeting." In Indonesia: Universitas Pasundan.

[7] Fallo, S. I., (2021). "Support Vector Machine, Naïve Bayes Classifier, and Ordinal Logistic regression in weather prediction." *In 2021 Thesis Universitas Gadjah Mada*.

[8] Azmi, A. U., Hadi, A. F., Anggraeni, D., & Riski, A. (2021). Naive bayes methods for rainfall prediction classification in Banyuwangi. In *Journal of Physics: Conference Series* (Vol. 1872, No. 1, p. 012028). IOP Publishing.

[9] Dawoodi, H. H., & Patil, M. P. (2023). Rainfall prediction for north maharashtra, india using advanced machine learning models. *Indian Journal of Science and Technology*, 16(13), 956-966.

[10] Sivanantham, S., Kumar, P. H., Vardhan, S. N., Kumar, S. C., Kumar, T. A., & Pradeep, T. (2023). Rainfall Prediction using Machine Learning Techniques–A Comparative Approach. In *2023 Third International Conference on Artificial Intelligence and Smart Energy* (ICAIS) (pp. 1472-1480). IEEE.

[11] Amin, F., & Mahmoud, M. (2022). Confusion matrix in binary classification problems: A step-by-step tutorial. *Journal of Engineering* Research, 6(5), 0-0.

[12] Görtler, J., Hohman, F., Moritz, D., Wongsuphasawat, K., Ren, D., Nair, R., ... & Patel, K. (2022). Neo: Generalizing confusion matrix visualization to hierarchical and multi-output labels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

[13] Haghighi, S., Jasemi, M., Hessabi, S., & Zolanvari, A. (2018). PyCM: Multiclass confusion matrix library in Python. *Journal of Open Source Software*, 3(25), 729.

[14] Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. Communications in Statistics-Theory and Methods, 49(9), 2080-2093.

[15] Piegorsch, W.W. 2020. Confusion Matrix. in Wiley. StatsRef: Statistics Reference Online.

[16] Yadavendra, & Chand, S. (2020). A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method. Machine Vision and Applications, 31(6), 46.

[17] Jahromi, A. H., & Taheri, M. (2017, October). A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. In 2017 Artificial intelligence and signal processing conference (AISP) (pp. 209-212). IEEE.

[18] Yoga, I. K., Prasetyowati, S. S., & Sibaroni, Y. (2022). Prediction And Mapping Rainfall Classification Using Naive Bayes And Simple Kriging. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 7(4), 1244-1253.

[19] Aiyelokun, O., Ogunsanwo, G., Ojelabi, A., & Agbede, O. (2021). Gaussian Naïve Bayes classification algorithm for drought and flood risk reduction. Intelligent Data Analytics for Decision-Support Systems in Hazard Mitigation: Theory and Practice of Hazard Mitigation, 49-62.

[20] Abdilah, M. N., Ruhiat, Y., & Guntara, Y. (2024). Rainfall Classification Analysis Using Naïve Bayes Classifier Based on Air And Wind Temperatures in Serang City. Spektra: Jurnal Fisika dan Aplikasinya, 9(1), 39-48.

[21] Murty, M. N., Raghava, R., Murty, M. N., & Raghava, R. (2016). Kernel-based SVM. Support vector machines and perceptrons: Learning, optimization, classification, and application to social networks, 57-67.

[22] Intan, P. K. (2019). Comparison of kernel function on support vector machine in classification of childbirth. Jurnal Matematika MANTIK, 5(2), 90-99.

[23] Yusof, K. W., Babangida, N. M., Mustafa, M. R., & Isa, M. H. (2017). Linear kernel support vector machines for modeling pore-water pressure responses. Journal of Engineering Science and Technology, 12(8), 2202-2212.

[24] Pratiwi, N., & Setyawan, Y. (2021). Analisis Akurasi Dari Perbedaan Fungsi Kernel Dan Cost Pada Support Vector Machine Studi Kasus Klasifikasi Curah Hujan Di Jakarta. Journal of Fundamental Mathematics and Applications (JFMA), 4(2), 203-212.

[25] Sunori, S. K., Singh, D. K., Mittal, A., Maurya, S., Mamodiya, U., & Juneja, P. K. (2021, November). Rainfall classification using support vector machine. In 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 433-437). IEEE.

[26] Mohanty, J. R., & Mohapatra, M. R. (2018). Rainfall prediction using support vector machine (SVM). IOSR J. Comput. Eng, 20, 6-13.

[27] Dai, Q. Y., Zhang, C. P., & Wu, H. (2016). Research of decision tree classification algorithm in data mining. International journal of database theory and application, 9(5), 1-8.

[28] Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research (IJSR), 5(4), 2094-2097.

[29] Chhetri, M., & Gurung, L. (2023). Rainfall Prediction using Decision Tree: A Case Study of CST, Phuntsholing. Zorig Melong-A Technical Journal of Science, Engineering and Technology, 8(1).

[30] Revathi, B., & Usharani, C. (2021). Rainfall prediction using machine learning classification algorithms. Int. J. Creat. Res. Thoughts (IJCRT), 9(1).