JOURNAL OF DATA SCIENCE | Vol.2024:58
eISSN:2805-5160

Comparative Accuracy of Data Mining Models for Predicting Extreme Weather Events in West Sumatra

Desindra Deddy Kurniawan[1], Tri Basuki Kurniawan[1 *]

[1]Magister of Information Technology, Universitas Bina Darma, Palembang, Indonesia

***Email:** tribasukikurniawan@binadarma.ac.id

Abstract

Weather factors have a vital role in human activities, especially extreme weather phenomena. Extreme weather can result in potential hydrometeorological disasters that cause loss of life and property. Climate change is also contributing to the higher frequency of extreme weather events. For this reason, research related to predicting extreme weather, especially very heavy rain, is needed to anticipate its impact. Research related to the prediction of extreme weather events is currently still being carried out using various models. By utilizing aerial observation data from Radiosonde (RASON) and daily rainfall data at the Minangkabau Meteorological Station Padang Pariaman, West Sumatra, extreme weather prediction modeling was carried out with the criteria of heavy rain events having rainfall intensity above 50 mm/day or 50 mm/day. 24 hours. From the data mining prediction model that has been carried out using the Support Vector Machine (SVM) Model, in this case, the Support Vector Regression (SVR), the Mean Squared Error (MSE) value is 502.88, and the R2 (Coefficient of Determination) score is 0.09. For the Artificial Neural Network (ANN) model, the Mean Squared Error (MSE) value was 590.03, and the R2 (Coefficient of Determination) score was -0.73 with an accuracy value of only 0.11 and a loss model value of 590. Meanwhile, for the data mining classification model using the Decision Tree Model, the value obtained The model accuracy was 0.47, and the Naïve Bayes (NB) model obtained a model accuracy value of 0.34. From the results of this comparison, it was found that the prediction model using the Decision Tree Model was more accurate in predicting extreme rain events in the West Sumatra region.

Keywords

Extreme Weather, Data Mining, Model Predicting

Introduction

Human life is greatly influenced by weather factors. Therefore, accurate weather information is a need that must be fulfilled to support human activities (Rifqi & Aldisa, 2024). Weather information related to natural disasters, such as the potential for extreme weather, is also very important in the context of disaster mitigation to avoid losses that occur, both loss of life and property. Extreme weather phenomena such as very heavy rain, strong winds, tornadoes, and lightning at this time also often cause hydrometeorological disasters. Extreme weather is an abnormal, unusual weather

Submission: 23 October 2024; **Acceptance:** 25 November 2024

event that can result in loss, especially the safety of life and property (BMKG Head Regulation Kep.009 of 2010) (Zulfiani & Fauzi, 2023). Extreme weather is a natural phenomenon that is quite difficult to predict. Modeling extreme weather is a challenge for modeling experts in Indonesia and the world (Hikmah et al., 2023).

Problems faced in weather forecasting include unstable changing atmospheric conditions, measurement errors, data that is too large, and an incomplete understanding of the performance of the resulting weather forecasts (Intan et al., n.d.). The role of extreme weather modeling, such as prediction or forecasting of the possibility of the same event in several future periods, is certainly very important in efforts to minimize risks to human survival and as evaluation and study material for the government in decision-making (Hikmah et al., 2023). Research and development of weather prediction models is also continuing to be carried out to find prediction models that are truly suitable for the Indonesian Maritime Continent (BMI) region, which is a tropical country located on the equator with very dynamic weather conditions (Maharani & Rejeki, 2021).

One way to predict the weather is to use data mining, which uses patterns contained in the rainfall database to classify the rainfall that occurs. The methods that will be used are classified into classification functions with several algorithms (Arif et al., 2022). Data mining is a suitable method, where the process is to implement mathematics and statistical techniques in artificial intelligence to machine learning against various databases, then processed and extracted, then identified knowledge and utilized data that has the potential to produce information (Mulyati et al., 2020). Data mining can be broken down based on the tasks and functions that can be performed, including prediction, classification, clustering, description, estimation, and association (Agung et al., 2023).

Research using data mining methods to predict weather has been carried out previously by Mujiasih (Mujiasih, 2011) using the Association rule model, Decision Tree (C 4.5), where the accuracy of the Decision Tree is higher with a value of 68.5%. The data used for this research is synoptic observation data or surface observations from 9 BMKG Maritime Meteorological Stations. The Artificial Neural Network (ANN) model has also been used to research rainfall prediction models in Padang City with performance based on a fairly good MSE value of 0.03 (Ali et al., 2021) then the Regression Support Vector Machine (SVM) model has also been used to predict the location of hot spots. (hotspot) forest fires in the South Sumatra region with an RSME value of 2.1 and an R2 value of 0.83 (Yandi et al., 2021). The Support Vector Machine method is also used in research to predict the weather in the Medan City area with an accuracy of 54.55% (Rifqi & Aldisa, 2024). Then, Kirana et al. (Kirana et al., 2024) also researched data mining models with the Naïve Bayes algorithm to predict the weather with an accuracy value of 84.26%.

In this research, the data mining method used is by comparing the Forecast method with Classification to create an extreme weather forecast model, namely predicting very heavy rain events. The prediction data mining method with the algorithm used is a Support Vector Machine (SVM) and Artificial neural network (ANN) (Pratama et al., 2022). Meanwhile, for classification data mining, the algorithms used are Decision Tree and Naïve Bayes (Siregar et al., 2020). The data used is data from the Minangkabau Padang Pariaman BMKG Meteorological Station, namely Upper-Air observation data using Radiosonde (RASON) and F-Klim 71 observation data. These four data mining models can carry out good predictions and training based on existing input

information. So, the four models were implemented to determine the level of accuracy of weather prediction results based on measured physical atmospheric parameter data. By knowing the level of accuracy of each model, you can choose the best model that can be applied in operational activities related to predicting extreme weather, such as very heavy rain that occurs in the West Sumatra region.

Methodology

2.1 Data Collection

The data collection method is implemented by collecting and processing upper air observation data (RASON) and rainfall observations for 1 (one) year from January 1, 2023 to December 31, 2023. There are 2 x REASON upper air observation data in 1 (one) day, namely data at 00UTC (07.00 WIB) and data at 12UTC (19.00 WIB). The elements taken from RASON data are Temperature (T), Humidity (RH), Direction (DDD), and Wind Speed (FF) in the significant layers 850mb (5000 ft), 700mb (10,000 ft) and 500mb (18,000ft). RASON data and rainfall data are compiled to create a dataset by determining the features/attributes and class/label. The dataset created has 24 attributes and one label, as in Table 1. For numeric label data, the amount of rainfall is changed to nominal by making the category None. Rain, Light Rain, Moderate Rain, Heavy Rain, Very Heavy Rain, and Extreme Rain as in Table 2.

Table 1. Dataset Attributes and Labels

Column	Name	Data	Description
1	T850mb 00Z	Temperature in layer 850mb at 00 UTC	Attribute
2	T700mb 00Z	Temperature in layer 700mb at 00 UTC	Attribute
3	T500mb 00Z	Temperature in layer 500mb at 00 UTC	Attribute
4	RH850mb 00Z	Moisture in layer 850mb at 00 UTC	Attribute
5	RH700mb 00Z	Moisture in layer 700mb at 00 UTC	Attribute
6	RH 500mb 00Z	Moisture in layer 500mb at 00 UTC	Attribute
7	ddd850mb 00Z	Wind direction in layer 850mb at 00 UTC	Attribute
8	ff850mb 00Z	Wind speed in layer 850mb at 00 UTC	Attribute
9	ddd700mb 00Z	Wind direction in layer 700mb at 00 UTC	Attribute
10	ff700mb 00Z	Wind speed in layer 700mb at 00 UTC	Attribute
11	ddd5000mb 00Z	Wind direction in layer 500mb at 00 UTC	Attribute
12	ff500mb 00Z	Wind speed in layer 500mb at 00 UTC	Attribute
13	T850mb 12Z	Temperature in layer 850mb at 12 UTC	Attribute
14	T700mb 12Z	Temperature in layer 700mb at 12 UTC	Attribute
15	T500mb 12Z	Temperature in layer 500mb at 12 UTC	Attribute
16	RH850mb 12Z	Moisture in layer 850mb at 12 UTC	Attribute
17	RH700mb 12Z	Moisture in layer 700mb at 12 UTC	Attribute

18	RH 500mb 12Z	Moisture in layer 500mb at 12 UTC	Attribute
19	ddd850mb 12Z	Wind direction in layer 850mb at 12 UTC	Attribute
20	ff850mb 12Z	Wind speed in layer 850mb at 12 UTC	Attribute
21	ddd700mb 12Z	Wind direction in layer 700mb at 12 UTC	Attribute
22	ff700mb 12Z	Wind speed in layer 700mb at 00 UTC	Attribute
23	ddd5000mb 12Z	Wind direction in layer 500mb at 00 UTC	Attribute
24	ff500mb 12Z	Wind speed in layer 500mb at 00 UTC	Attribute
25	RAIN	Total Rainfall in 24 hours	Label

Table 2. Rainfall Criteria for Label/Class

No	Intensitas Curah Hujan	Kriteria Curah Hujan
1	TTU atau > 0.5 mm	Tidak Ada Hujan
2	0.5 mm – 20 mm / Hari	Hujan Ringan
3	20 mm – 50 mm / Hari	Hujan Sedang
4	50 mm – 100 mm / Hari	Hujang Lebat
5	100 mm – 150 mm / Hari	Hujan Sangat Lebat
6	>150 mm / Hari	Hujan Ekstrim

2.2 Creating a Data Mining Model

The following is a flow plan for the stages of creating a data mining model that will be implemented, namely:

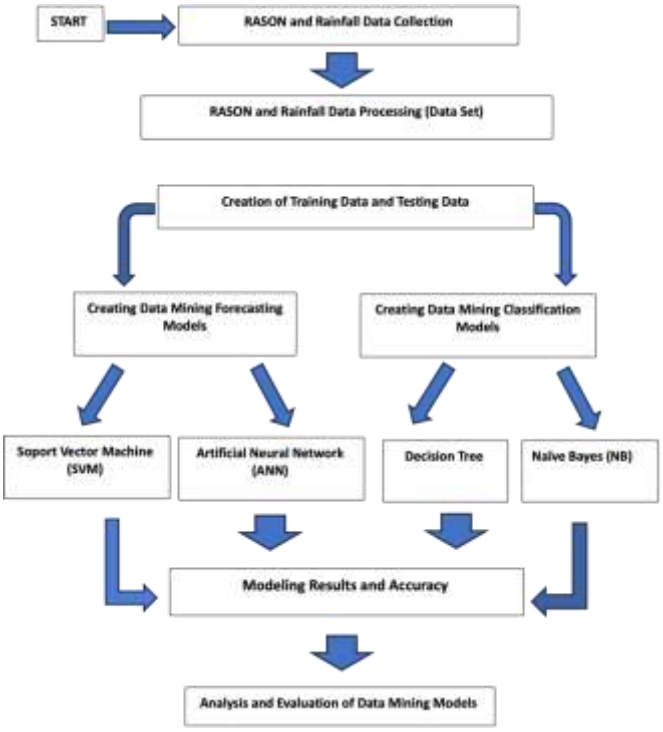

Figure 1. Research Process Flow

The data mining methods used are data mining prediction models and data mining classification models. For data mining prediction models using the Support Vector Machine (SVM) and Artificial Neural Network (ANN) methods. And for the data mining classification model using the Decision Tree and Naïve Bayes methods. The Support Vector Machine (SVM) method used is the Support Vector Regression (SVR) method because Support Vector Regression (SVR) is an application of Support Vector Machine (SVM), which is used in regression cases. The Support Vector Regression (SVR) method can be used on time series data, data that is not normally distributed, and data that is not linear. Data mining model creation is carried out using the Python programming language in the Visual Studio Code (VSC) application. The steps taken in creating a Python algorithm in VSC are as follows:

1. Import the required libraries
2. Read and prepare the dataset
3. Data preprocessing
4. Building a model
5. Training the model
6. Model evaluation
7. Visualization of Model Evaluation Results

2.3 Model Evaluation Stages

At this stage, the author will evaluate the data mining model, where there are 4 (four) data mining models used, namely the Support Vector Regression (SVR), Artificial Neural Network (ANN), Decision Tree, and Naïve Bayes models. The evaluation carried out was to look at the Mean

Squared Error (MSE) R2 (Coefficient of Determination) and model accuracy. So you can know the advantages and disadvantages of the 4 (four) data mining models.

Results and Discussion

3.1. Dataset Creation and Preprocessing

The dataset initially consisted of 365 rows and 25 columns, but before processing, a preprocessing process was carried out by deleting nine rows of incomplete data (missing values) due to damage to the Rason observation tool. So, the dataset that can be processed is 356 lines with 24 features/attributes and one target/class. After the data was cleaned from noise, the data, which was initially in Excel format, was then converted into CSV format with the aim of smaller data, supporting many libraries in Python, easy to read, simplicity of format, and good performance. Then, carry out the process of splitting the dataset with a split ratio of 70:30. Which means 70% of the data is used for training data, and 30% is used as testing data. So, 249 rows of data are used as training data, and the remaining 107 rows of data are used as testing data.

3.2. Support Vector Regression (SVR) Model

From the results of modeling using Python, the value for the Support Vector Regression (SVR) model is a Mean Squared Error (MSE) of 502.88 and an R2 (Coefficient of Determination) score of 0.09. Meanwhile, for visualization, the plotting of model prediction results with actual values can be seen in Figure 2.

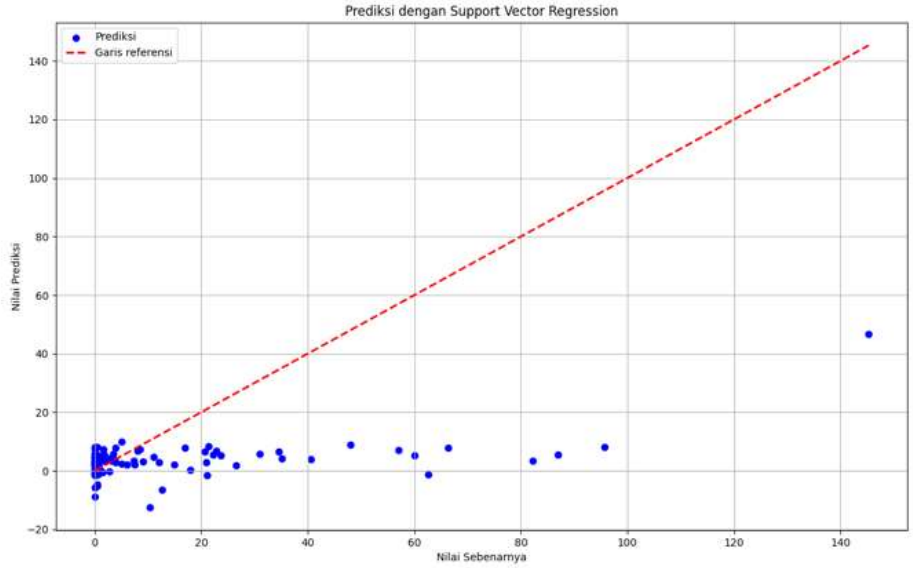

Figure 2. Plotting of SVR Model Prediction Results with Actual Values

3.3. Artificial Neural Network Model (ANN)

The results of modeling using the Artificial Neural Network (ANN) method obtained a Mean Squared Error (MSE) value of 590.03 and an R2 (Coefficient of Determination) score of -0.73. Plotting the results of the Loss Model and Accuracy Model can be seen in Figure 3, while visualization of the comparison between prediction results and accuracy can be seen in Figure 4.

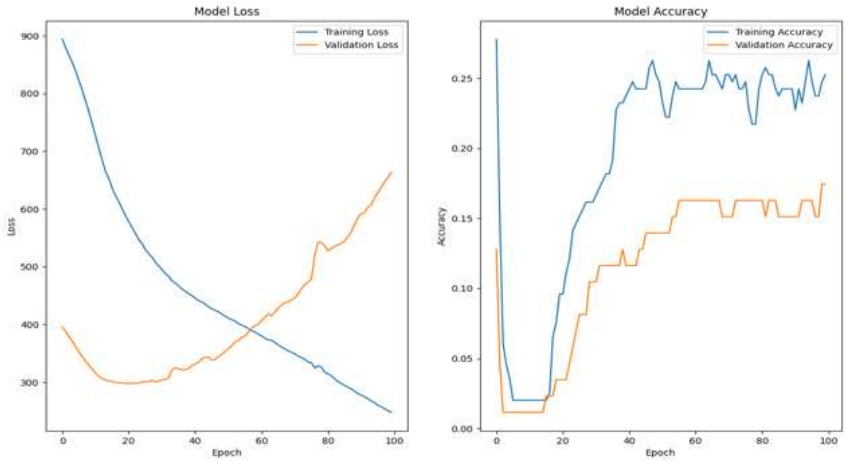

Figure 3. Graph of ANN Model Loss and Accuracy Values

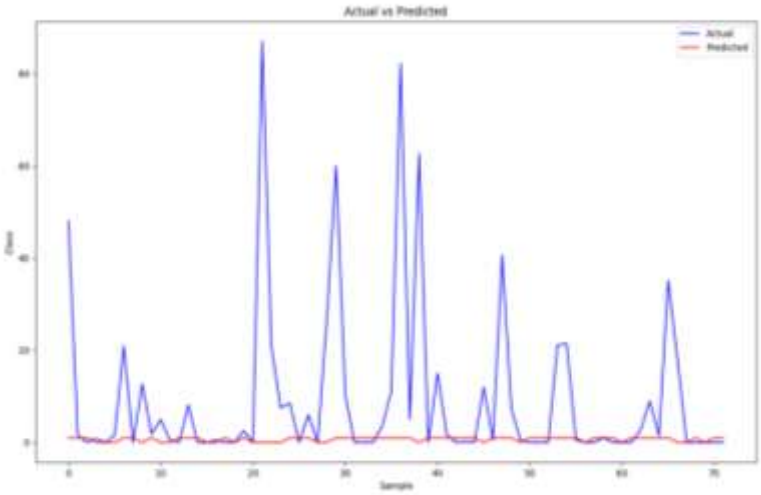

Figure 4. Plotting Predicted Values with Actual Values from the ANN Model

3.4. Decision Tree Model

The Decision Tree model created is a data mining model that is classification in nature. In making this model, the label data, which was previously in the form of numeric, was then changed to Nominal with the categories No Rain, Light Rain, Moderate Rain, Heavy Rain, Very Heavy Rain, and Extreme Rain. Based on the Head of BMKG Regulation Number 009 of 2010, the threshold for the Extreme Rain category is the amount of rainfall above 50mm/day, so the Heavy Rain, Very Heavy Rain, and Extreme Rain categories were changed to just Extreme Rain, this is to suit the research objectives. Namely data mining modeling for predicting extreme weather events. The modeling results using the Decision Tree method obtained an accuracy value of 0.47. To visualize the decision tree from Decision Tree modeling, it can be seen in Figure 5.

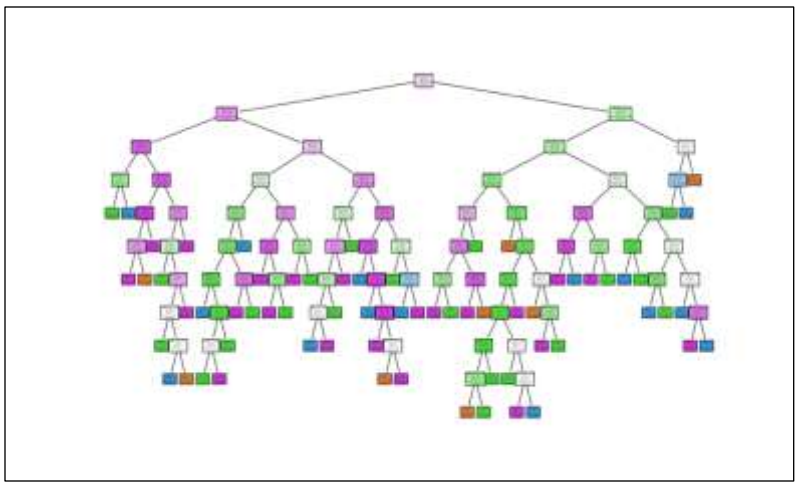

Figure 5. Decision Tree Model

3.5. Naïve Bayes Model

The Naïve Bayes model is a type of classification data mining modeling, so it is used as a comparison with the Decision Tree classification model. The dataset used in this modeling is the same as the Decision Tree model, namely, the labels/targets are in nominal format with four categories: Heavy Rain, Very Heavy Rain, and Extreme Rain. The results of modeling with Naïve Bayes the confusion matrix graph from the modeling results are shown in Table 3 and Figure 6.

Table 3. Naïve Bayes Modeling Accuracy Results

	precision	recall	f1-score	support
Ekstrim	0.14	0.38	0.21	8
Ringan	0.27	0.10	0.15	39
Sedang	0.08	0.14	0.11	14
Tidak Ada Hujan	0.57	0.59	0.58	46
accuracy			**0.34**	107
macro avg	0.27	0.30	0.26	107
weighted avg	0.37	0.34	0.33	107

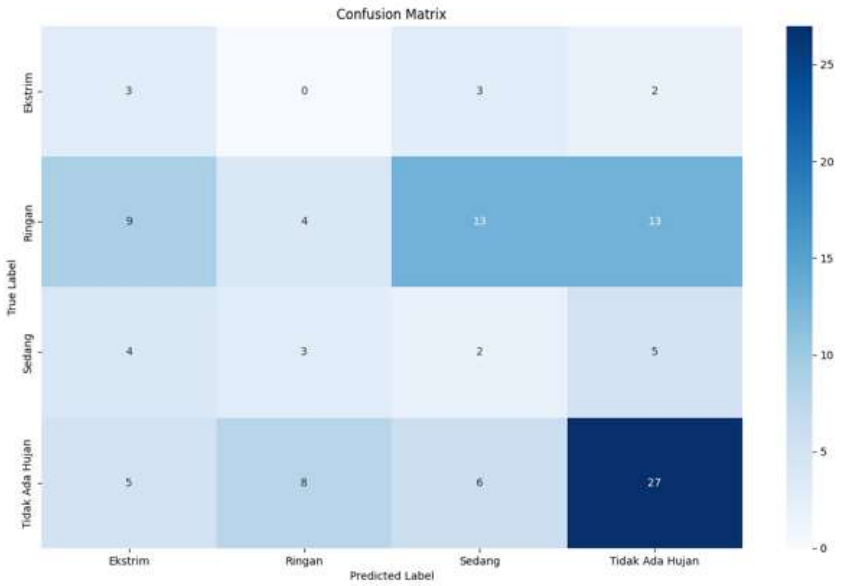

Figure 6. Confusion Matrix Model Naïve Bayes

4.6 Analysis of Data Mining Modeling Results

Modeling for extreme weather predictions, especially rainfall predictions in the West Sumatra region using data mining methods, this type of prediction has been carried out using Support Vector Regression (SVR) and Artificial Neural Network (ANN) models. Where the SVR model has a Mean Squared Error (MSE) value of 502.88 and an R2 (Coefficient of Determination) score of 0.09. Meanwhile, for Artificial Neural Network (ANN) modeling, the Mean Squared Error (MSE) score was 590.03, and the R2 (Coefficient of Determination) score was -0.73 with an accuracy value of only 0.11 and a model loss value of 590. These two data mining models turned out to have low accuracy, The SVR model of the R2 (Coefficient of Determination) value has poor performance and cannot explain data variability.

Likewise, with ANN modeling, the value of each epoch has a high loss value, so this shows poor model performance. Even the R2 (Coefficient of Determination) value is negative, meaning the model performance is very bad, even lower than the average value. This is because the two models cannot read patterns in the dataset in the form of time series. Each parameter or attribute in the data set does not have a pattern following the time series.

This may also be because upper air observations (RASON) are carried out 2 (two) times a day with a period of 12 hours, namely at 07.00 LT (00 UTC) and 19.00 LT (12.00 UTC), meaning they are carried out in the morning and evening. So when rain occurs in the afternoon or evening, it is caused by morning parameters, while rain that occurs at night or early in the morning is caused by weather parameters at night. Unless it rains from morning to morning for 24 hours, something like this has never happened. So, the parameters at 00UTC and 12 UTC will often be different in describing real weather conditions. And this is what causes the model's accuracy to decrease in making rainy weather predictions.

For data mining modeling, this type of classification has been carried out using Decision Trees and Naïve Bayes models. From the research results, the Decision Tree model has an accuracy level of 0.47, while the Naïve Bayes model has an accuracy of 0.34. This classification data mining method turns out to have slightly better accuracy than the prediction data mining method. This is because the classification data mining model can read parameter patterns or attributes from the dataset by grouping or classifying weather parameters such as Temperature (T), Humidity (RH), Direction (D), and Wind Speed (ff) in the 850mb and 700mb layers. And 500mb. So, these parameters or attributes can be used as rainfall prediction parameters using Decision Tree and Naïve Bayes even though the accuracy is still below 50%.

Conclusion

The modeling results of the 4 (four) data mining methods used, both prediction and classification, showed that prediction models such as Support Vector Regression (SVR) and Artificial Neural Network (ANN) had very low or even poor accuracy. The accuracy value of Support Vector Regression (SVR) and Artificial Neural Network (ANN) is very low, even the R2 (Coefficient of Determination) of Artificial Neural Network (ANN) is negative. Meanwhile, the classification data mining models used, such as Decision Tree and Naïve Bayes, have higher accuracy than prediction methods (forecasting).

The Artificial Neural Network (ANN) model has the lowest accuracy and is worse than Support Vector Regression (SVR), so these two models cannot be used to make rainfall predictions in the West Sumatra region. The Decision Tree and Naïve Bayes models also have low accuracy but are still better when compared to Support Vector Regression (SVR) and Artificial Neural Network (ANN). So, the Decision Tree and Naïve Bayes models can be used as rainfall prediction models, but the Decision Tree model is more appropriate for making rainfall predictions in the West Sumatra region because it has the highest accuracy of the 4 (four) data mining models.

References

Agung, A. S., Fauzi, A. A., Risal, A. A. N., & Adiba, F. (2023). Implementasi Teknik Data Mining terhadap Klasifikasi Data Prediksi Curah Hujan BMKG Di Sulawesi Selatan. *Jurnal Tekno Insentif, 17*(1), Article 1. https://doi.org/10.36787/jti.v17i1.955

Ali, Z. I., Nur, I. M., & Fauzi, F. (2021). *ARTIFICIAL NEURAL NETWORK UNTUK MEMPREDIKSI CURAH HUJAN DI KOTA PADANG DENGAN METODE BACKPROPAGATION DAN ADALINE*. http://repository.unimus.ac.id/4559/11/Jurnal%20Zakia%20Intan%20Ali.pdf

Arif, A. A., Firdaus, M., Rahmaddeni, & Maruhawa, Y. (2022). Perbandingan Metode Data Mining untuk Prediksi Curah Hujan dengan Algoritma C4.5, Naïve Bayes, dan KNN: Comparison of Data Mining Methods for Predition of Rainfall with C4.5, Naïve Bayes, and KNN Algorithm. *SENTIMAS: Seminar Nasional Penelitian Dan Pengabdian Masyarakat*, 187–197. https://journal.irpi.or.id/index.php/sentimas/article/download/308/129/2243

Hikmah, H., Asrirawan, A., Apriyanto, A., & Nilawati, N. (2023). Peramalan Data Cuaca Ekstrim Indonesia Menggunakan Model ARIMA dan Recurrent Neural Network. *Jambura Journal of Mathematics*, *5*(1), Article 1. https://doi.org/10.34312/jjom.v5i1.17496

Intan, Indo, et al. "Performance Analysis of Weather Forecasting Using Machine Learning Algorithms (Analisis Performansi Prakiraan Cuaca Menggunakan Algoritma Machine Learning)." *Pekommas*, vol. 6, no. 2, 31 Oct. 2021, pp. 1-8, doi:10.30818/jpkm.2021.2060221.

Kirana, A. N., Nurhakim, B., Permana, S. E., Prihartono, W., & Dwilestari, G. (2024). IMPLEMENTASI ALGORITMA NAIVE BAYES UNTUK MEMPREDIKSI CUACA MENGGUNAKAN RAPIDMINER. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(2), Article 2. https://doi.org/10.36040/jati.v8i2.8967

Maharani, S., & Rejeki, H. A. (2021). PENGARUH PROPAGASI MADDEN JULIAN OSCILLATION (MJO) DI BENUA MARITIM INDONESIA (BMI) TERHADAP SIKLUS DIURNAL DINAMIKA ATMOSFER DAN CURAH HUJAN DI PROVINSI LAMPUNG TAHUN 2018. *Jurnal Sains & Teknologi Modifikasi Cuaca*, *22*(2), Article 2. https://doi.org/10.29122/jstmc.v22i2.4528

Mujiasih, S. (2011). PEMANFATAN DATA MINING UNTUK PRAKIRAAN CUACA. *Jurnal Meteorologi Dan Geofisika*, *12*(2). https://doi.org/10.31172/jmg.v12i2.100

Mulyati, S., Husein, S. M., & Ramdhan, R. (2020). RANCANG BANGUN APLIKASI DATA MINING PREDIKSI KELULUSAN UJIAN NASIONAL MENGGUNAKAN ALGORITMA (KNN) K-NEAREST NEIGHBOR DENGAN METODE EUCLIDEAN DISTANCE PADA SMPN 2 PAGEDANGAN. *JIKA (Jurnal Informatika)*, *4*(1), Article 1. https://doi.org/10.31000/jika.v4i1.2288

Pratama, A. R. I., Latipah, S. A., & Sari, B. N. (2022). OPTIMASI KLASIFIKASI CURAH HUJAN MENGGUNAKAN SUPPORT VECTOR MACHINE (SVM) DAN RECURSIVE FEATURE ELIMINATION (RFE). *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, *7*(2), Article 2. https://doi.org/10.29100/jipi.v7i2.2675

Rifqi, M. N., & Aldisa, R. T. (2024). Penerapan Metode Support Vector Machine Dalam Memprediksi Prediksi Cuaca. *Journal of Computer System and Informatics (JoSYC)*, *5*(2), Article 2. https://doi.org/10.47065/josyc.v5i2.4961

Siregar, A. M., Faisal, S., Cahyana, Y., & Priyatna, B. (2020). Perbandingan Algoritme Klasifikasi Untuk Prediksi Cuaca. *Jurnal Accounting Information System (AIMS)*, *3*(1), Article 1. https://doi.org/10.32627/aims.v3i1.92

Jepri Yandi, Kurniawan, T. B., Edi Surya Negara, & Akbar, M. (2021). Prediksi Lokasi Titik Panas Kebaran Hutan menggunakan Model Regresion SVM (Support Vector Machine) pada Data Kebakaran Hutan Daops Manggala Agni Oki Provinsi Sumatera Selatan Tahun 2019. *InfoTekJar (Jurnal Nasional Informatika Dan Teknologi Jaringan)*, *6*(1), 10–15. https://doi.org/10.30743/infotekjar.v6i1.4101

Zulfiani, A., & Fauzi, C. (2023). Penerapan Algorimta Backpropagation Untuk Prakiraan Cuaca Harian Dibandingkan Dengan Support Vector Machine dan Logistic Regression. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *7*(3), Article 3. https://doi.org/10.30865/mib.v7i3.6173