Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 723−731
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1978

Prediction Map of Rainfall Classification Using Random Forest and Inverse Distance Weighted (IDW)

Ibnu Muzakky M Noor*, Sri Suryani Prasetyowati, Yuliant Sibaroni

Informatika, S1 Informatika, Telkom University, Bandung, Indonesia
Email: [1,*]ibnumuzakky@student.telkomuniversity.ac.id, [2]srisuryani@telkomuniversity.ac.id, [3]yuliant@telkomuniversity.ac.id
Correspondence Author Email: ibnumuzakky@student.telkomuniversity.ac.id
Submitted:**26/07/2022**; Accepted:**23/08/2022**; Published: **30/09/2022**

Abstract−The amount of rainfall that occurs can affect natural disasters and even food production to economic activities. the factor of the area where the rain occurs is one of the main parameters for how the change occurs. So, it is necessary to have a rainfall prediction approach that aims to find out when and what type of rain will occur. Spatial classification and interpolation are two methods used to make predictions. Random Forest is a classification method that can be used to predict rainfall. and Inverse Distance Weighted is one of the stochastic interpolation techniques to calculate the estimated rainfall from the data points of rainfall that occur so that the distribution can be visualized. In the implementation of random forest, the model that is built on a daily basis gets the best level of accuracy in the 5D model sub model C with an accuracy of 0.8238 while the monthly model gets the best level of accuracy in the sub-model B 4M 0.9362. and the results of predictions and mapping using IDW show that daily predictions from June 1-4 2022 show that Most of Java Island will experience light rain, June 5-7 2022 most of Java Island will experience sunny cloudy days. And for monthly predictions, August and June 2022 show the distribution of monthly rainfall with predictions that most of Java is cloudy, while May, July, October, September have light rainfall in most of Java.

Keywords: Rainfall; Inverse Distance Weighted; Random Forest; Classification;

1. INTRODUCTION

Rainfall is very significant information since it can have an impact on many elements of human life, either directly or indirectly. Floods and landslides are two examples of natural disasters that can result from rainfall. Or the impact on urban life, including the impact on the sewer system, the impact of traffic, etc. [1]. The geographical location of the rainfall, such as locations with a tropical climate, notably in Indonesia, which tends to have a diverse climate, as well as the shape and direction of the islands, can all have an impact on this [2, 15]. Tropical climate regions typically have hot temperatures, with temperature serving as the primary criterion. But in reality, climate change does not primarily depend on temperature. Instead, rainfall—particularly during the rainy season—plays a significant role in climate change [3]. Changes in rainfall volume can have an impact on economic activity, food production, and even natural disasters.

To overcome this, there needs to be an approach to rainfall prediction that aims to find out when and what kind of rainfall will occur. The Random Forest machine learning method is a method that can be used to predict rainfall. Suhaila Zainudin, et al in his research showed the results of a comparison of five classification techniques (Naïve Bayes, Decision Tree, Support Vector Machine, Neural Network and Random Forest) which were carried out for prediction of rainfall in Malaysia, where the results were Decision Tree and Random Forest performed well for prediction. rainfall due to their ability to train less data and predict larger amounts of data with higher F-measures, and Random Forest correctly identified 1043 events with small training data (10%) of 1581 events [4]. While Jie Dou, et al in his research compared the performance of two machine learning models Decision Tree and Random Forest to model large landslide events triggered by rainfall. The results show that the Decision Tree and Random Forest models produce an almost accurate vulnerability map (AUC N 0.9). Random Forest (AUC = 0.956) has a much higher overall efficiency than Decision Tree (AUC = 0.928) [5].

However, the research by [4] and [5] only focuses on the predictive results of the machine learning model built. has examined the prediction results using the scattering pattern by predicting the distribution of rainfall in the central part of Taiwan using the Inverse Distance Weighting (IDW) method from rainfall data between 1981 and 2010 taken from 46 rainfall monitoring stations. where 12 rainfall monitoring stations belonging to the Taichung Irrigation Association (TIA) were used for cross-validation. The influence of the radius value, and the control-a parameter were selected to calculate the RMSE value in order to obtain optimal rainfall interpolation data. From the results of the study, the IDW parameter that has optimal values for rainfall interpolation data is in a radius of up to 10–30 km in most cases. And from the results of its application, the IDW method has better predictive accuracy in the dry season (October to April) than the flood season (May–September).

From the research above and several other studies on rainfall predictions, they still have not focused on the distribution of the prediction results that have been made. Like other studies by [7] which examined the rainfall prediction system with the random forest model explained that the random forest model was suitable for use in predicting rainfall. Evaluation of the model used to use training data, data testing and evaluation of the model using k-fold cross validation. And the results of the study indicate that the implementation of random forest is more suitable to use all data for data training and data testing when viewed from the evaluation of accuracy, precision, recall, f-measurement, kappa statistics, MAE, RMSE and ROC Area. The accuracy of random forest using the 10-fold cross validation technique is 71.09% while the whole data technique is 99.45%. Regarding a different study by [8] that

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 723−731
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1978

utilized the distribution pattern from the predictions of the Support Vector Machine (SVM) and Random Forest (RF) methods on air temperature. Through data scanning techniques, the accuracy of the SVM and RF approaches in determining long-term air temperature was determined in this study. SVM and RF models were contrasted using data from 30 Iranian radio stations. The k-fold test approach, which is a superior method than the traditional train/test method to judge the model's accuracy, was then used to construct the SVM and RF models. In comparison to the SVM model, which has equivalent values of 0.142, 1.855, and 0.945, the RF model has a better average scatter index, average absolute error, and average variance account of 0.111, 1.456, and 0.968, respectively. The results, however, demonstrate that the geostatistical-based kriging method can spatially replicate air temperature.

A similar study in applying the distribution pattern by [9] examined the development of a prediction map for the spread of Dengue Hemorrhagic Fever (DHF) in 2016 to 2018 based on data from 2010 to 2015. The prediction method used was GSTAR combined with Inverse Distance Weighting (DHF). IDW) and Kriging (Ordinary Kriging and Universal Kriging) to analyze the distribution pattern of the DHF area distribution. From this combination, a map of the distribution of Dengue Hemorrhagic Fever (DHF) was obtained with comparison results to the accuracy of the resulting disease distribution pattern based on the RMSE value of each combination. Then from using this combination to predict the pattern of the spread of DHF in the Bandung area, it was revealed that the IDW and Kriging methods were not significantly different or the estimation results produced by the two methods were not significantly different, but the Ordinary Kriging and Universal Kriging methods gave different results.

Research by [10] also applies a location assessment method using Inverse Distance Weighting (IDW) in the Laterite Nickel application (a case study in Block R, Konawe Regency, Southeast Sulawesi) researching on mapping the lateral distribution of limonite ore and estimating nickel resources, using IDW interpolation method in estimating Ni content and thickness of mineralized zone. The evaluation of the method used is to find the smallest RMSE value of the power parameters used, namely 1, 2, 3, 4 and 5. The results show that using the power = 1 parameter produces the best RMSE value for estimating Ni content and the thickness of the limonite ore zone. And from these results, mapping of ore distribution shows that additional potential for nickel resources is still open to the south and northwest of the study area.

Based on the problems and research studies that have been described and taking into account the results of various research results and their advantages and disadvantages, a rainfall classification prediction map system was made using the Random Forest method to determine the identification of the intensity of rainfall in a particular area by identifying the results. of the resulting method evaluation. And add a method of spatial analysis, namely Inverse Distance Weighted (IDW) to describe the distribution pattern of rainfall based on various existing climatic characteristics.

2. RESEARCH METHODOLOGY

2.1 Research Framework

In this study, the system that will be built is the development of a rainfall prediction map by combining the Random Forest machine learning method and the Inverse Distance Weighting (IDW) spatial analysis method. The following Figure 1 is a flowchart of the system design that was built:

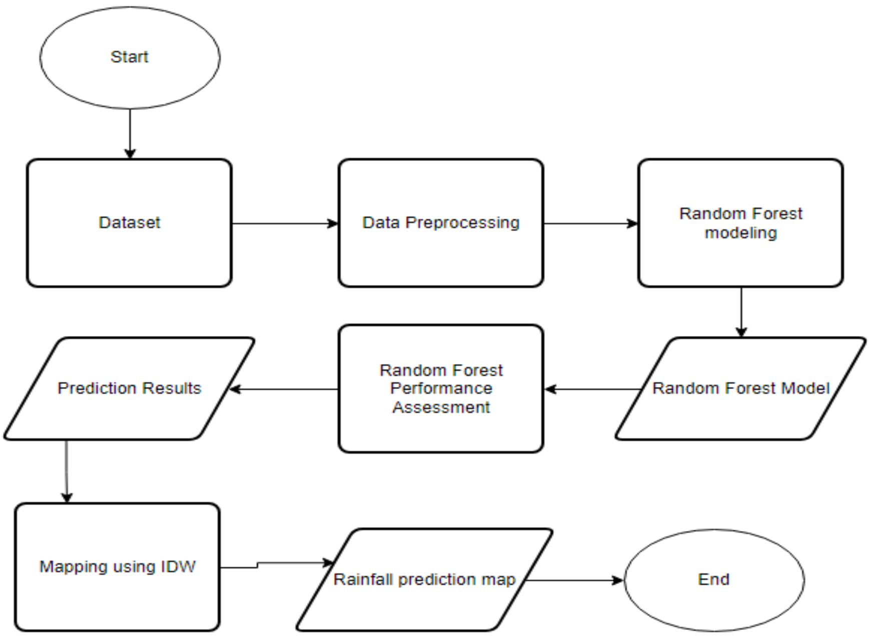

Figure 1. System Flowchart

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 723−731
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1978

2.2 Dataset

The data used were taken from several climatological stations on the island of Java with attribute data including minimum temperature (ºC), maximum temperature (ºC), average temperature (ºC), duration of sunshine (hours), rainfall (mm), average humidity (%), maximum wind speed (m/s), wind direction at maximum speed (°), average wind speed (m/s). obtained from the Meteorology, Climatology and Geophysics Agency (BMKG) from January 1, 2010 to March 31, 2022 with daily data details. At this stage, the research data used is data with a total of 120771 data records and starting from January 1, 2010 to March 31, 2022 with detailed data per day. The following is the initial dataset used in this study.

Table 1. Dataset

Date	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	Location
01-01-2010	26	32.2	28.8	75	0	3.4	5	250	4	Tanjung Priok Maritime Meteorological Station
…	…	…	…	…	…	…	…	…	…	…
31-03-2022	22.6	33.5	26.7	77	0	7.2	5	320	1	Bogor Climatology Station

2.3 Data Preprocessing

Preparing the raw data for the next procedure is a task performed during the preprocessing stage. Preprocessing can be done by eliminating unstructured (unsupervised) input or by converting it to a simpler data format for processing by users or machines [16]. At this stage, the data obtained will then be processed to check whether the data quality is good or not for use in the next stage. Each data will be checked such as checking for missing values, wrong data, noise data and others. And at this stage too, labeling value data is needed to find out the label/class of each data row so that it can be categorized from any class.

2.3.1 Handling Missing Values

In this research, the linear interpolation method is very useful in dealing with missing values in the data, this method is very suitable for the system being built. The reason for using this method is because there are not many data that have missing values and for each row of data it is important to use because the data has time data so that when there are rows that have missing values, they cannot be deleted. Therefore, it is necessary to fill in the sample data using the linear interpolation method. The following is an illustration after and after handling missing values.

Table 2. Data Before Handling Missing Values

Column	Date	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	Location
Non-Null Count	120711	113113	113135	114245	113720	100840	110421	115989	115206	116027	120771
Dtype	Object	Float64	Float64	Float64	Float64	Float64	Float64	Float64	Float64	Float64	Object

Table 3. Data After Handling Missing Values

Column	Date	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	Location
Non-Null Count	120711	120711	120711	120711	120711	120711	120711	120711	120711	120711	120771
Dtype	Object	Float64	Float64	Float64	Float64	Float64	Float64	Float64	Float64	Float64	Object

2.3.2 Data Labeling

In the classification system, data labeling is needed to identify the overall value of the data. In this study, data labeling is used to identify the RR data column with data labeling rules that refer to the category of rainfall data from the BMKG. The following are the categories of rainfall that are implemented in this system.

Table 4. Data Category Result

Category	Class	Range (RR)	Total data
Cloudy	0	$Ch \geq 0$	62773
Light Rain	1	$0.5 \leq Ch \leq 20$	43645
Moderate Rain	2	$21 \leq Ch \leq 50$	11012
Heavy Rain	3	$51 \leq Ch \leq 100$	2938
Very Heavy Rain	4	$101 \leq Ch \leq 150$	346
Extreamly Rain	5	$Ch > 150$	57

Based on the table above, the data after labeling shows that the cloudy category is the highest with a total of 62733, the light rain category is 43654, 11012 for moderate rain, 2938 for heavy rain, 346 very heavy rain, and 57 extreme rains.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 723−731
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1978

2.4 Random Forest

Random Forest is an ensemble learning algorithm that can handle the use of large-scale classification and high-dimensional regression problems [8]. The Random Forest method is similar to the decision tree method, which is a combination of predictor trees that depend on the values of random vectors collected independently and with the same distribution for all trees which aims to see the behavior of each tree [12]. The main principle of the Random Forest method is to build a tree ensemble based on the fact that a number of weak lessons can come together to form strong learnings. Random Forest is a collection of individual trees. As a result, every tree is a weak learner. This model ends up being a powerful lesson when trees are merged into a Random Forest [4, 11].

Random forest is a predictor type consisting of M random regression trees. The projected value at query point x is represented as m_n (x; Θ_j, D_n) for the jth tree in the family, where _1,...,Θ_m is an independent random variable, distributed the same as the generic random variable and independent of D_n. In practice, variables are used to re-sample the training set before individual tree growth and to select the next splitting direction. The j-th tree estimate is expressed mathematically as follows [13].

$$m_n(x;\Theta_j,\mathcal{D}_n) = \sum_{i\,\epsilon\,\mathcal{D}_n^*(\Theta_j)} \frac{X_i A_n(x;\Theta_j,\mathcal{D}_n)^{Y_i}}{N_n(x;\Theta_j,\mathcal{D}_n)} \tag{1}$$

Where D_n^* (Θ_j) is the set of data points selected before the tree is constructed, A_n (x;Θ_j,D_n) is the cell containing x, and N_n (x;Θ_j,D_n) is the number of (pre-selected) points. which belongs to A_n (x;Θ_j,D_n).

2.4.1 Random Forest Performance Measurement

Because the model to be built only focuses on the classification stage. Then the performance assessment that will be used is the assessment of Accuracy (ACC), Precision, Recall, and F-measure. Accuracy is obtained based on the percentage of pixels accurately categorized as erosion and non-landslide [5].

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Precision evaluates a True Positive (TP) entity, which is an accurately categorized entity, against a False Positive (FP) entity, which is a misclassified entity. And for recall is to compare True Positives with False Negatives, which are entities that are not categorized at all. The following is a formula that can be used:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{|TP| + |FN|} \tag{4}$$

F-measure is used to determine the average value of precision and recall. This value needs to be known because it is a solution when you want to know whether one algorithm is better than another when it only has two values, namely precision and recall [4]. The following is a formula that can be used:

$$F - measure = 2 \times \frac{Precison\ \times Recall}{Precision + Recall} \tag{5}$$

2.5 Inversi Distance Weighted (IDW)

Inverse Distance Weighted (IDW) is a deterministic interpolation method that is used to fulfill certain statistical assumptions which assumes that the attribute value of the no-sampled point is the weighted average of the known values in the surrounding area [6, 14]. IDW is divided into two groups, namely global interpolation and local interpolation. Global interpolation is used to calculate predictions using all data sets. Meanwhile, local interpolation is calculating predictions using data with a smaller area within a larger study area [9].

Prediction results can be influenced by several determining factors, namely the power factor and the surrounding area (neighboring radius). The commonly used power parameter values are 1, 2, 3, 4 and 5. And the IDW equation used in determining the weighting is as follows [10]:

$$w_i = \frac{1/di^p}{\sum_{i=1}^{n} 1/di^p} \tag{6}$$

And to find out the value of the prediction point, you can use the following equation:

$$\hat{Z}_0 = \sum_{i=1}^{n} w_i . Z_i \tag{7}$$

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 723−731
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1978

In the two equations above, Z _0 means the predicted point value, w_i is the weighting factor of point i, the distance between point i and the predicted point is interpreted as di, and p is the power value or exponential factor.

3. RESULT AND DISCUSSION

In this section, the data used in the system is rainfall data taken from the BMKG station with a large number of data as many as 120771 data records and starting from January 1, 2010 to March 31, 2022 with detailed data per day. Then the data will be checked for quality before entering the modeling and prediction stage which includes handling missing values, and labeling value data. In this study, because the data is time series data, the handling of missing values in the system built uses a missing value filling system with linear interpolation techniques. The data that has been processed will enter the modeling stage using a random forest model to see the level of accuracy of each model built and then the best model will be used in the future data prediction stage.

3.1 Random Forest Modeling

In Random Forest data modeling, the data used is taken based on the input month and the output month. The input month is the data used as test data and train data, while the output data is used as the target class data. the data to be used in the modeling must be calculated on the average of the entire data with the conditions of calculating the average monthly when the model is made based on the month and calculating the average per day when the condition of the model is made on a daily basis.

Train data and test data are taken based on the separation set by the year parameter with data conditions less than the year parameter including as train data and conditions more than the year parameter as test data. The following is an overview of the conditions for separating train data and test data.

$$input\ data = \begin{cases} train, & year\ input < year \\ test, & year\ input \geq year \end{cases} \quad (8)$$

The separation of train data and test data with the input year parameter is carried out because the data used in this system is time series data so that each data record is very influential on time so that data separation can only be done by setting the data separation taking into account the time in the data and in the above formula using input year as data separator parameter.

Then after the separation of test data and train data is done, the next scheme builds several models to see the comparison of the accuracy of each model. In this study, 5 models were built for each type of input data based on the total data divided into 4 data sections. The data distribution was taken based on the index data which later on each section of the data would be used as random forest models. and the last 1 part is a model built with total data without data sharing. The following is an illustration of the data sharing scheme used for Random Forest models.

Table 5. Random Forest Modeling Scenario

Input Month/Day Data	Model Type	Data sharing based on index	Accuracy
2	A	0:958	0.85
	B	958:1916	0.74
	C	1916:2874	0.88
	D	2874:3834	0.80
	Gab	0:3834	0.93

The table 5 is a model schema based on data sharing by index with a total sample of 3834 data rows divided into 5 models with 4 data separations for 4 models A, B, C, D and 1 section using the total data in Gab models. After the model is built, it will look at the best model for the best daily model and the best monthly model for each model built and later the model that gets the best accuracy will be used as a model for future predictions.

3.2 Feature Combination Selection

After determining what model to build, then the next step is to try a combination of features from existing features. This feature combination is carried out to find the best combination of features which will later produce an accuracy value so that later it can be a comparison between the accuracy of the sub-models. The number of feature combinations is obtained from the feature extraction function which extracts all combinations of each feature in the sub-model that has been built. The following is one of the combinations of features obtained from the sub-model with a combination of 24 feature combinations. Feature extraction continues to be carried out until it reaches the maximum limit of all possible combinations of existing features, and later from each feature combination will produce accuracy so that the combination with the best accuracy will be used as a feature to be used as predictive feature data.

3.3 Random Forest Model Performance Measurement

At this stage, the performance assessment of the model that has been built is useful to see the level of accuracy of the model which will later be used as a predictive model. The performance assessment used is an assessment of Accuracy

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 723−731
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1978

(ACC), Precision, Recall, and F-measure. However, in this study, the Precision, Recall, and F-measure values are considered when the condition is that there are several models that have the same accuracy value so that when the conditions are not the same for accuracy, this study only looks at the accuracy value.

As explained in sub-chapter 3.1, there are 5 model scenarios, including 4 models based on data sharing and 1 model without data sharing or a combined model. Then for the data separation scheme with the year parameter used, namely 2020 with the same separation conditions as in the formula in chapter 3.3, for data with a time less than 2020 as train data and data more than 2020 as test data. Then there are 2 types of models built, namely the daily model and the monthly model. For the monthly model, there are 6 models built, including 2M, 3M, 4M, 5M, 6M, 7M, then for the daily model there are 5 models, including 3D, 4D, 5D, 6D, 7D. and each model has sub-models A, B, C, D, and a combination. And the following are the results of the models that have been built and the best accuracy obtained from each model.

Table 6. Best Accuracy Results on Monthly Model

	2M	3M	4M	5M	6M	7M
A	0,8557	0,8391	0,8657	0,8673	0,8722	0,8592
B	0,8928	0,9109	0,9362	0,9115	0,9065	0,9315
C	0,9310	0,8951	0,9003	0,9350	0,9047	0,9239
D	0,8552	0,9259	0,9209	0,8982	0,8939	0,9118
Gab	0,8757	0,8906	0,8854	0,8835	0,8989	0,8922

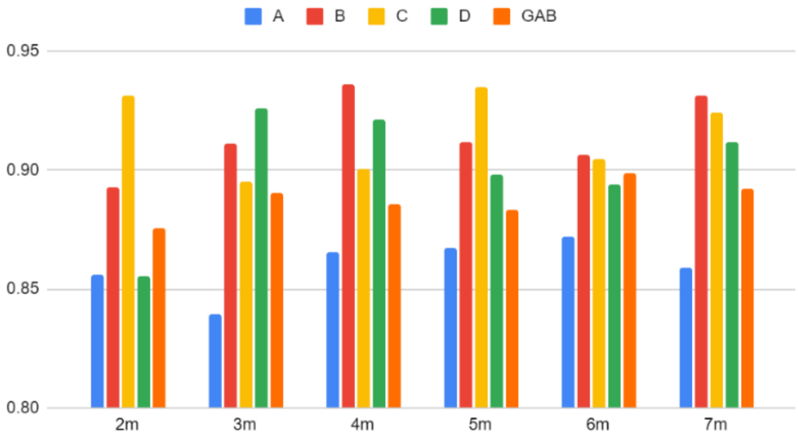

Figure 2. Best Accuracy on Histogram for Monthly Model

Based on the Table 6 and Figure 2 above, it can be seen that the monthly modeling shows an increasing trend of accuracy in the B, C, D sub models in each model when compared to the accuracy in the A and Gab models. then it can be seen that the 4M model in sub model B shows the best level of accuracy of 0.9362.

Table 7. Best Accuracy Results on Daily Models

	3D	4D	5D	6D	7D
A	0.8001	0.8024	0.7973	0.7988	0.7972
B	0.7713	0.7752	0.7757	0.7669	0.7714
C	0.8221	0.8163	0.8238	0.773	0.819
D	0.7627	0.7649	0.7668	0.7957	0.7664
GAB	0.784	0.7897	0.7911	0.7893	0.7868

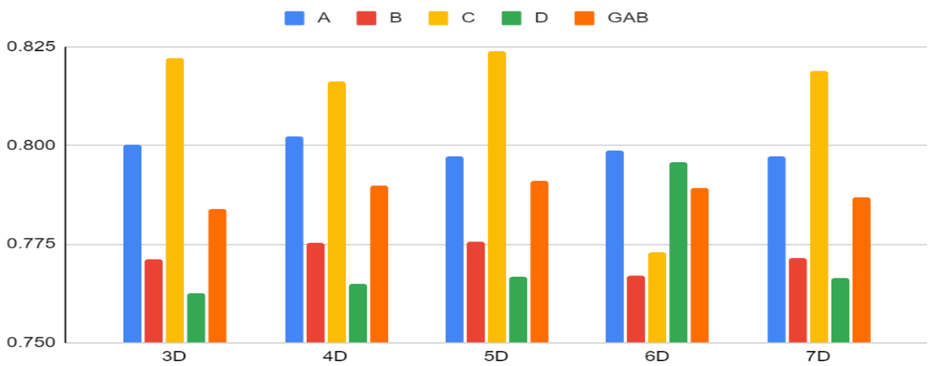

Figure 3. Best Accuracy on Histogram for Daily Model

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 723−731
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1978

While the daily model shown in table 7 and figure 3, the accuracy trend obtained experienced a significant increase in sub-model C in each model and for all models looks the same for all sub-models. of all the models obtained that the best level of accuracy in the 5D model sub model C with an accuracy of 0.8238. Then after getting the best accuracy from the daily and monthly models. Then the next step is to use these models to be used in predicting rainfall. For the monthly model, rainfall predictions are made for the next 6 months, while for the daily model predictions are made for the next 7 days.

3.4 Mapping using IDW

In this study, the mapping of the prediction results was carried out after using the Random Forest model and then the prediction data was processed using ArcMap software to see the distribution of the rainfall prediction results using IDW on the island of Java with a total of 26 stations. The shapefile file was first created to find out the initial map as an observation of the distribution of rainfall. Then determine the x-axis, y-axis and z-axis to see the distribution of the prediction results on the map of the island of Java. For the x and y axes, fill in with the coordinates for each station, totaling 26. And followed by the z axis, which is the predicted data that will be seen the distribution based on the location of the station.

After determining the x, y, and z axes. then interpolated the points from each existing station using the IDW method by inputting the z-axis point feature filled with the predicted data column. Then in the rainfall distribution class, the class distribution is obtained from the classification or grouping results sourced from the IDW default class on the arcmap. The following is the result of mapping from predictions made based on predictions for the next 6 months starting from May 2022 to October 2022. And mapping predictions for the next 7 days starting from May 1, 2022 to May 7, 2022.

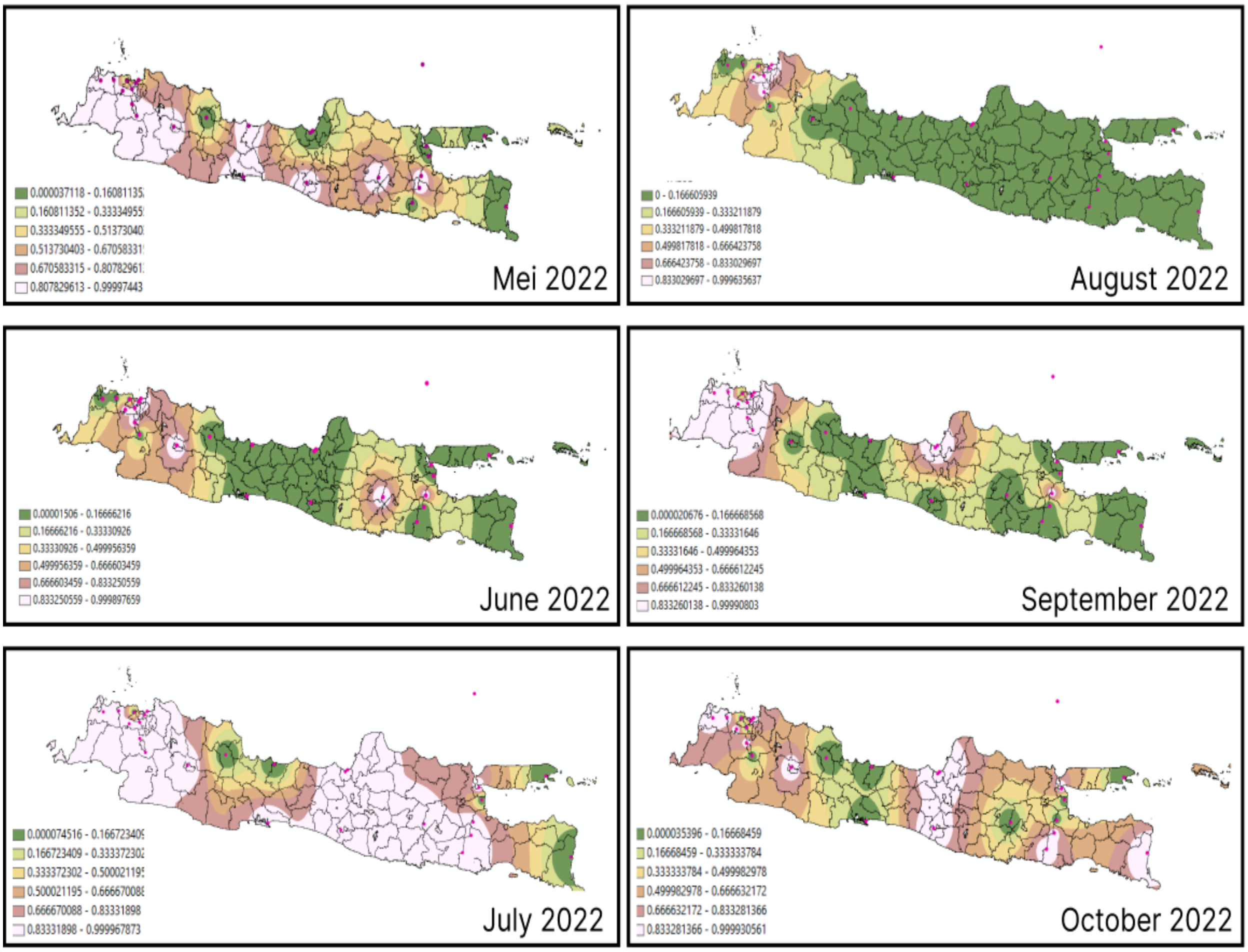

Figure 4. Monthly Rainfall Prediction Map

Figure 4 shows that May has a predominance of light rain which indicates that Banten and most areas of Jakarta, West Java experience light rain which is shown in pink. Then in June it showed the dominance of sunny cloudy as indicated by the dominance of the green color found in Central Java and most of West Java. in July and October showed a predominance of light rain over most of the island of Java. The difference, in August and September shows that most of the island of Java has a predominance of sunny and cloudy which is indicated by green and yellow colors.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 723−731
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1978

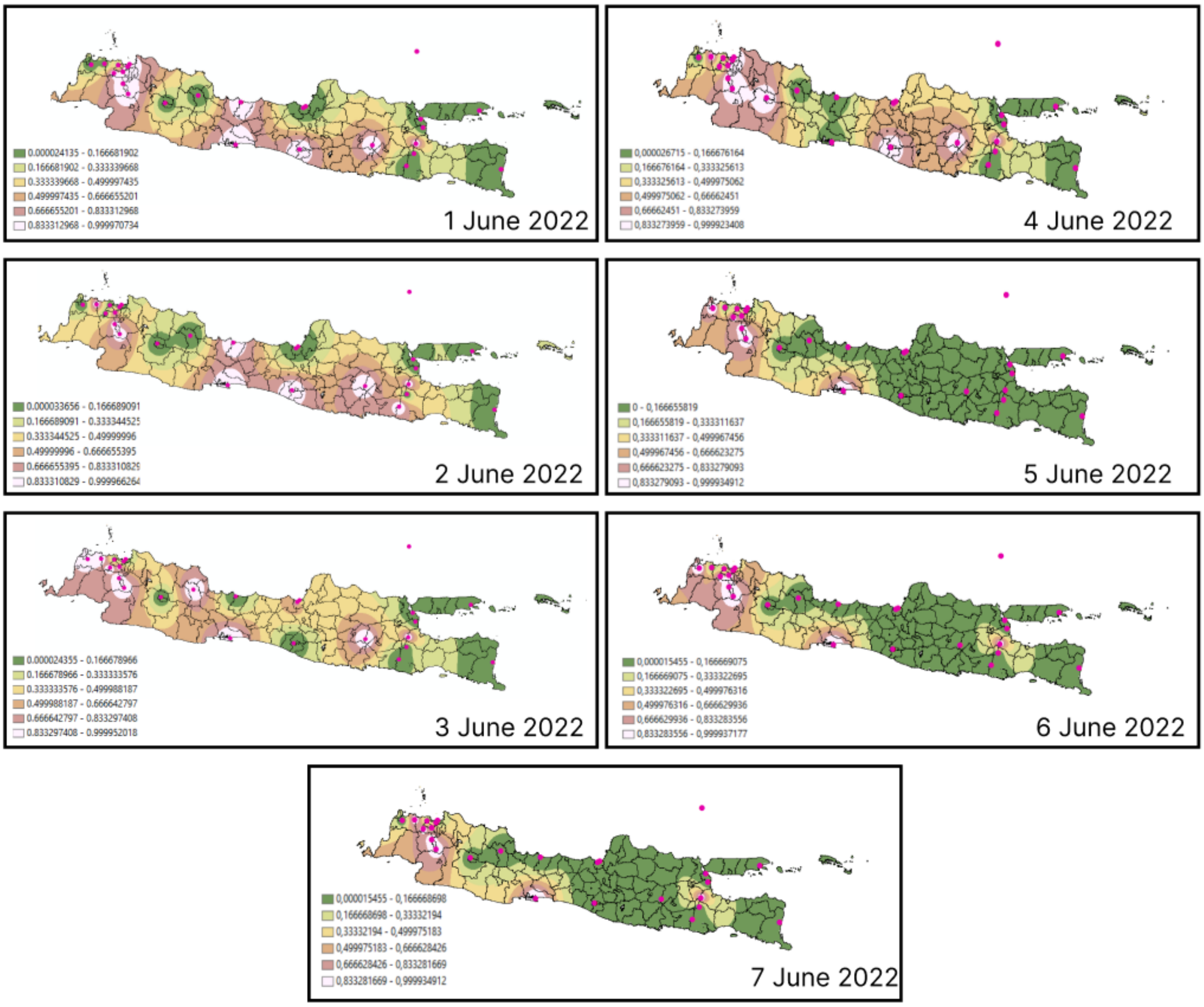

Figure 5. Daily Rainfall Prediction Map

While the daily prediction mapping can be seen in Figure 5. From June 1 to June 4, it shows the dominance of pink color which indicates that most areas of Java Island experience light rain and there is a small part which is indicated by green and semi-yellow clouds. for June 5 to June 7 shows dominance in most areas of Java Island with sunny cloudy weather.

4. CONCLUSION

Based on the research described previously, the Random Forest Model works well in modeling the data and predicting the data. The daily model gets the best level of accuracy on the sub model C 5D model with an accuracy of 0.8238 while the monthly model gets the 4M sub model B model with the best accuracy rate of 0.9362. the model with the best accuracy from each type of model is influenced by experiments carried out such as feature selection by trying all the available combinations of features, then performing a data separation scheme so that it can create many models and submodels so that the accuracy data is more and more. Then after getting the best model accuracy values are daily and monthly, then predictions are made and see the distribution of predictive data using IDW. There are two types of predictions made, namely daily and monthly based on the best model that has been built. For daily predictions, this study predicts 7 days from June 1 to June 7, 2022. Then for monthly predictions, this study predicts the next 6 months from May 2022 to October 2022. And the prediction results show that daily predictions from June 1-4 2022 show that Most of Java Island will experience light rain, June 5-7 2022 most of Java Island will experience sunny cloudy days. And for monthly predictions, August and June 2022 show the distribution of monthly rainfall with predictions that most of Java is cloudy, while May, July, October, September has light rainfall in most of Java.

REFERENCES

[1] K. Gao, T. Liu, B. Hu, M. Hao, Y. Zhang, and X. Ning, "Establishment of Economic Forecasting Model of High-Tech Industry Based on Genetic Optimization Neural Network," *Intell. Neurosci.*, vol. 2022, Jan. 2022, doi: 10.1155/2022/2128370.
[2] Y. Swarinoto and S. Sugiyono, "PEMANFAATAN SUHU UDARA DAN KELEMBAPAN UDARA DALAM

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 723−731
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1978

PERSAMAAN REGRESI UNTUK SIMULASI PREDIKSI TOTAL HUJAN BULANAN DI BANDAR LAMPUNG," *J. Meteorol. dan Geofis.*, vol. 12, 2011, doi: 10.31172/jmg.v12i3.109.

[3] F. Altarazi *et al.*, "Analysis and Implementation of Thermal Heat Exchanger Tube Performance with Helically Pierced Twisted Tape Inserts Using ANFIS Model," *Math. Probl. Eng.*, vol. 2021, p. 1734909, 2021, doi: 10.1155/2021/1734909.

[4] S. Zainudin, D. Jasim, and A. Abu Bakar, "Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, p. 1148, 2016, doi: 10.18517/ijaseit.6.6.1487.

[5] J. Dou *et al.*, "Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan," *Sci. Total Environ.*, 2019, doi: 10.1016/j.scitotenv.2019.01.221.

[6] F.-W. Chen and C.-W. Liu, "Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan," *Paddy Water Environ.*, vol. 10, 2012, doi: 10.1007/s10333-012-0319-1.

[7] A. Primajaya and B. Sari, "Random Forest Algorithm for Prediction of Precipitation," *Indones. J. Artif. Intell. Data Min.*, vol. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.

[8] S. Mohsenzadeh Karimi, O. Kisi, M. Porrajabali, F. Rouhani-Nia, and J. Shiri, "Evaluation of the support vector machine, random forest and geo-statistical methodologies for predicting long-term air temperature," *ISH J. Hydraul. Eng.*, vol. 26, pp. 1–11, 2018, doi: 10.1080/09715010.2018.1495583.

[9] S. Prasetiyowati and Y. Sibaroni, "Prediction of DHF disease spreading patterns using inverse distances weighted (IDW), ordinary and universal kriging," *J. Phys. Conf. Ser.*, vol. 971, p. 12010, 2018, doi: 10.1088/1742-6596/971/1/012010.

[10] H. Purnomo, "APLIKASI METODE INTERPOLASI INVERSE DISTANCE WEIGHTING DALAM PENAKSIRAN SUMBERDAYA LATERIT NIKEL (Studi kasus di Blok R, Kabupaten Konawe-Sulawesi Tenggara)," *Angkasa J. Ilm. Bid. Teknol.*, vol. 10, no. 1, pp. 49–60, 2018.

[11] A. Salam, S. Prasetiyowati, and Y. Sibaroni, "Prediction Vulnerability Level of Dengue Fever Using KNN and Random Forest," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, pp. 531–536, 2020, doi: 10.29207/resti.v4i3.1926.

[12] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[13] G. Biau and E. Scornet, "A Random Forest Guided Tour," *TEST*, vol. 25, 2015, doi: 10.1007/s11749-016-0481-7.

[14] A. Faudzan, S. Suryani, and T. Budiawati, "Perbandingan Metode Inverse Distance Weighted (IDW) dengan Metode Ordinary Kriging untuk Estimasi Sebaran Polusi Udara di Bandung," *E-Prosiding Eng.*, vol. 2, no. 2, pp. 6726–6734, 2015

[15] I. Jhonson Arizona Saragih *et al.*, "Prediksi Curah Hujan Bulanan Di Deli Serdang Menggunakan Persamaan Regresi Dengan Prediktor Data Suhu Dan Kelembapan Udara," *J. Meteorol. Klimatologi dan Geofis.*, vol. 7, no. 2, pp. 6–14, 2020, [Online]. Available: http://bmkgsoft.database.bmkg.go.id

[16] A. Mustofa and R. Novita, "Klasifikasi Sentimen Masyarakat Terhadap Pemberlakuan Pembatasan Kegiatan Masyarakat Menggunakan Text Mining Pada Twitter," vol. 4, no. 1, pp. 200–208, 2022, doi: 10.47065/bits.v4i1.1628.