JIT Journal of Innovation and Technology
e-ISSN: 2721-8562; p-ISSN: 2721-8570
Vol.4, No.2, October 2023. Page: 62 – 66

Weather Forecast for Bandar Lampung City Using Random Forest and C4.5

Rahma Ferika Shaumi[1*] Sri Ratna Sulistiyanti[1], F.X. Arinto Setyawan[1], Helmy Fitriawan[1], and Sri Purwiyanti[1]

[1] Department of Electrical Engineering, Faculty of Engineering, Universitas Lampung, Jl. Prof. Soemnantri BrojonegoroNo. 1, Bandar Lampung 35143, Indonesia

*Corresponding Author: rahmaferika@gmail.com

Article history
Received: 04.04.2023
Revised: 18.08.2023
Accepted: 01.09.2023

DOI:10.31629/jit.v4i1.5625

Abstract

Weather forecasts are important information for various agencies and the wider community. Weather forecasts are usually used to benefit various sectors such as transportation, tourism, plantations and others. This study aims to create a new model regarding weather forecasting using the random forest and C4.5 algorithms using the WEKA application. The dataset uses data from the Panjang Maritime Meteorological Station with 365 days of data and six attributes: rain intensity, average temperature, humidity, rainfall, sunshine duration and average wind speed. The results obtained from this study between the random forest algorithm and C4.5, namely cross-validation trials fold 5, 10 and 15 random forests, have better results than C4.5 by using the MAE and RMSE evaluation values, then in testing with a percentage split 25% on the evaluation of the MAE value is better at C4.5. Still, the random forest has better results for all experiments evaluating the RMSE value and two evaluations of the MAE value.

Keywords: weather forecast, random forest, C4.5, WEKA.

1. Introduction

Weather forecasts are usually also called weather forecasts, which are air conditions such as an area for a short period. This weather condition can change easily at certain times with a relatively small area coverage [1]. This situation can be due to the influence of wind, rainfall, temperature, air pressure, humidity etc.

Weather forecasting is very important information because it can be used to determine a decision regarding individual, group or agency activities (study of the BMKG). Several fields, such as the transportation sector, are closely related to weather forecasting information. Some modes of transportation use navigation, so if the weather is bad, it will interfere with navigation, resulting in delayed trips. Weather forecasting also plays a very important role in the tourism, plantation, and telecommunications sectors and can play an important role in preventing natural disasters such as floods.

The weather forecasting process is carried out officially by the Meteorology, Climatology and Geophysics Agency (BMKG), a government agency. BMKG will go through several processes to obtain weather forecasts with several pieces of equipment, which are then processed so that they can finally inform the public or agencies [2].

Technology development is very rapid, so many innovations have emerged, including weather forecasting using several algorithms. Previous studies using random forests are better than some models, such as K-Nearest Neighbors and Least Medium Square Regression networks. Radial Basic Function and Multilayer Perceptron [3]. Then the

DOI:10.31629/jit.v4i1.5625

random forest also gets better results than the Artificial Neural Network and Support Vector Machine [4]. Another study used C4.5, which got better results than Naïve Bayes [5]. In the previous study, C4.5 also obtained high accuracy, namely 81.94% (and a hike). C4.5 also gets the highest accuracy from the comparison using the KNN and Naïve Bayes algorithms [6].

This study hypothesizes that the two random forest and C4.5 algorithms get a small error value and compare the two algorithms. Thus, one of the benefits of this research is to obtain alternative weather forecasts for the city of Bandar Lampung.

1.1. Random Forest

The random forest recursively divides the data set by selecting, at each node, the variable and threshold that maximizes the dissimilarity metric (e.g., information gain) until the termination criterion is met (e.g., the Number of samples of the data set falls below a specified number, etc.) [8].
To make a prediction:

$$f(X) = \frac{1}{K}\sum\sum_{k=1}^{K} T_k(x) \qquad (1)$$

Where:
K = Number of trees in the forest
F = Number of input variables randomly chosen at each split, respectively
T = variable output
X = New predicted Number

1.2. Algorithms C4.5

C4.5 is one of the algorithms used to form decision trees based on training data (by themselves). C.45 is a predictive model for a decision using a hierarchical or tree structure. Every tree has branches; the branch represents the attributes that must be met to go to the next branch until it ends in a leaf (no more branches) [6].

Algorithm C4.5 is one of the algorithms used to form a decision tree based on training data. This algorithm is a very powerful and well-known classification and prediction method. Input training samples and samples where the training samples are in the form of sample data used to build a tree that has been tested for its validity, while the samples are data fields that will be used as parameters in conducting data classification [7].

Determine the root of the tree by calculating the entropy:

$$Entropy\ (S) = \sum_{i=1}^{N} -pi \log 2\ pi \qquad (2)$$

Where:
S = case set
N = Number of partitions S
Pi = Proportion of Si to S

Then determine the gain :

$$Gain\ (S, A) = entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{S} * entropy\ (Si) \qquad (3)$$

Where:
S = case set
A = Features
N = Number of partitions attribute A
|Si| = Proportion of Si to S
|S| = Number of cases in S

2. Materials and Methods

2.1. Datasets

The dataset from the Panjang Maritime Meteorological Station has 365 days of data with six attributes: rain intensity, average temperature, average humidity, rainfall, sunshine duration and average wind speed. The dataset can be seen in Figure 1.

Tanggal	Intensitas Hujan	Temperatur Rata-Rata	Kelembaban Rata-Rata	Curah Hujan	Lamanya Penyinaran Matahari	Kecepatan Angin Rata-Rata
11-01-2022	Hujan Sedang	27	86	33	2	2
12-01-2022	Hujan Ringan	26,8	83	5,3	4,5	2
13-01-2022	Hujan Ringan	27	83	11,1	4,5	4
14-01-2022	Hujan Ringan	27,9	81	24,2		2
15-01-2022	Hujan Ringan	27,6	77	11,6	8	2
16-01-2022	Hujan Ringan	28,6	78	4,8	5	2
17-01-2022	Hujan Lebat	28,1	80	72,5	7,2	2
18-01-2022	Hujan Ringan	27,9	78	3,5	6	3
19-01-2022	Berawan	27,1	80	0,2	1,5	2
20-01-2022	Hujan Ringan			8,2	4,8	2
21-01-2022	Hujan Ringan	27,1	82	11,1	2	2

Figure 1. Dataset

2.2. Pre Processing Data

The dataset is used in CSV format later. The next step is to process the missing data with the Replace Missing Value filter. In Figure 2, the rain intensity attribute, which previously had missing data, was 38%, then after using the missing data filter, it became 0%.

DOI:10.31629/jit.v4i1.5625

Selected attribute

Name: Intensitas Hujan | Type: Nominal
Missing: 0 (0%) | Distinct: 5 | Unique: 1 (0%)

No.	Label	Count	Weight
1	Hujan Ringan	286	286
2	Berawan	54	54
3	Hujan Sedang	19	19
4	Hujan Lebat	5	5
5	hujan sangat lebat	1	1

Figure 2. Replace Missing Value

Then the six attributes are processed to evaluate attribute values by measuring the information gained about the class. This stage can be done with ranking attributes. Figure 3 is the result of evaluating attribute values from 6 attributes to 4 attributes.

```
Attribute Evaluator (supervised, Class (nominal): 2 Intensitas Hujan):
       Information Gain Ranking Filter

Ranked attributes:
 1.0137  1 Tanggal
 0.6371  5 Curah Hujan
 0.3545  6 Lamanya Penyinaran Matahari
 0.2786  3 Temperatur Rata-Rata
 0       7 Kecepatan Angin Rata-rata
 0       4 Kelembapan Rata-Rata
```

Figure 3. Evaluating attribute Values

3. Results and Discussion

3.1. Training Set with Random Forest

Figure 4 shows the prediction results on the training set with a random forest that takes 0.05 s to build the model.

```
Time taken to build model: 0.05 seconds

=== Predictions on training set ===

    inst#     actual  predicted error prediction
        1 1:'Hujan Ringan' 1:Hujan Ringan       0.933
        2 1:'Hujan Ringan' 1:Hujan Ringan       0.935
        3  2:Berawan  2:Berawan       0.671
        4 1:'Hujan Ringan' 1:Hujan Ringan       0.924
        5  2:Berawan  2:Berawan       0.696
        6 1:'Hujan Ringan' 1:Hujan Ringan       0.946
        7 3:'Hujan Sedang' 3:Hujan Sedang       0.667
        8  2:Berawan  2:Berawan       0.726
        9 1:'Hujan Ringan' 1:Hujan Ringan       0.908
       10 1:'Hujan Ringan' 1:Hujan Ringan       0.918
```

Figure 4. Training set with random forest

Figure 5 shows the evaluation of the training set, which takes 0.01 s to build the training data. Correctly Classified Instances of 100%, MAE Value of 0.05 and RMSE of 0.0966.

```
      364 1:'Hujan Ringan' 1:Hujan Ringan       0.939
      365  2:Berawan  2:Berawan       0.694

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances         365               100      %
Incorrectly Classified Instances         0                 0      %
Mean absolute error                      0.05
Root mean squared error                  0.0966
Total Number of Instances              365
```

Figure 5. Evaluation on training set with random forest

3.2. Training Set with C4.5

Figure 6 shows the prediction results on the training set with C4.5 by requiring 0 s to build the model.

```
Time taken to build model: 0 seconds

=== Predictions on training set ===

    inst#     actual  predicted error prediction
        1 1:'Hujan Ringan' 1:Hujan Ringan       0.784
        2 1:'Hujan Ringan' 1:Hujan Ringan       0.784
        3  2:Berawan 1:Hujan Ringan   +   0.784
        4 1:'Hujan Ringan' 1:Hujan Ringan       0.784
        5  2:Berawan 1:Hujan Ringan   +   0.784
        6 1:'Hujan Ringan' 1:Hujan Ringan       0.784
        7 3:'Hujan Sedang' 1:Hujan Ringan   +   0.784
        8  2:Berawan 1:Hujan Ringan   +   0.784
        9 1:'Hujan Ringan' 1:Hujan Ringan       0.784
```

Figure 6. Training set with C4.5

Figure 7 shows the evaluation of the training set, which takes 0.02 s to build the training data. Correctly Classified Instances of 78.3562%, MAE Value of 0.1445 and RMSE of 0.2688.

```
      364 1:'Hujan Ringan' 1:Hujan Ringan       0.784
      365  2:Berawan 1:Hujan Ringan   +   0.784

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances         286               78.3562 %
Incorrectly Classified Instances        79               21.6438 %
Mean absolute error                      0.1445
Root mean squared error                  0.2688
Total Number of Instances              365
```

Figure 7. Evaluation on training set C4.5

DOI:10.31629/jit.v4i1.5625

3.3. Testing with Cross-Validation and Percentage Split

Tabel 1. Results of cross-validation and percentage split testing

	CCI	MAE	RSME
Cross Validation Fold 5 - RF	78,3562%	0,1366	0,2579
Cross Validation Fold 5 – C4.5	78,3562%	0,1445	0,2688
Cross Validation Fold 10 – RF	78,3562%	0,136	0,2571
Cross Validation Fold 10 – C4.5	78,3562%	0,145	0,2688
Cross Validation Fold 15 – RF	78,3562%	0,1363	0,2573
Cross Validation Fold 15 – C4.5	78,3562%	0,1445	0,2689
Percentage Split 25% - RF	77,7372%	0,143	0,2717
Percentage Split 25% - C4.5	77,7372%	0,1407	0,2728
Percentage Split 66% - RF	75%	0,145	0,2821
Percentage Split 66% - C4.5	75%	0,15	0,2887
Percentage Split 90% - RF	69,4444%	0,1532	0,2981
Percentage Split 90% - C4.5	69,4444%	0,1665	0,3113

Based on Table 1, the results of testing between random forest and C4.5 with cross-validation folds 5 for random forest have the same CCI value as C4.5, then for MAE and RMSE random forest values have a better evaluation value with a difference of 0, 0079 and 0.0109. Cross-validation folds 10 for random forest has an MAE value, and RMSE random forest has a better evaluation value with a difference of 0.0085 and 0.0117. Cross-validation folds 15 for random forest have a better MAE and RMSE evaluation value than C4.5, with a difference of 0.082 and 0.0116. The three tests have the same CCI value of 78.3562%. Based on the 25% percentage split test between random forest and C4.5, they have the same value, namely 77.7372%. The MAE value is better at C4.5 with a difference of 0.0023, and the RMSE value is better at random forest with a difference of 0.0011. Then testing with a percentage split of 66% has the same CCI value of 75% with a better random forest evaluation value of MAE and RMSE with a difference of 0.005 and 0.0063. Tests with a percentage of 90% have a CCI value of 69.4444% with a better MAE and RMSE evaluation score with a difference in values of 0.0133 and 0.0132. The smaller the MAE value (closer to 0) [9]. The more accurate the prediction results; the smaller the RMSE value (closer to 0), the more accurate the prediction [10].

4. Conclusion

The dataset used in this study was formed from information on the Panjang Maritime Meteorological Station for the Bandar Lampung City area of 365 data with six attributes. Of the two algorithms used, it has been found that the random forest has better prediction accuracy based on the MAE and RMSE evaluation values in the cross-validation fold tests 5, 10, and 15. The percentage split test is 66% and 90% for the MAE evaluation values, and the RMSE evaluation for the three percentage split tests has better results. At the same time, C4.5 has a better score only on the 25% percentage split test on the MAE evaluation value.

References

[1] Kertasapoetra, A.G. (2010). Teknologi Konservasi Tanah dan Air. *Rineka Cipta*.

[2] Diani, F., Permana, H., Sarah N, P. (2012). Kajian Sistem Informasi Prakiraan Cuaca BMKG Pada BMKG Bandung.*Seminar Nasional Aplikasi Informasi 2012(SNATI 2012)*, B-16-B21. ISSN:1907-5022.

[3] Naing, W.Y.N., and Htike, Z.Z. (2015). Forecasting of Monthly Temperature Variations using Random Forest. *Journal of Arp*, 10(21), 10109-10112.

[4] Lahouar, A., Slama, J.B.H. (2015). Day-Ahead Load Forecast using Random Forest and Expert Input Selection. *Journal of Elsevier*, 103, 1040-1051.

[5] Bamaruckmani, P., Kausalya, R. (2019). Efficient Analysis of Weather Prediction Using C4.5 Decision Tree and Naïve Bayes Algorithm. *International Journal of Research in Advent Technology*, 7(5S), 253-260.

DOI:10.31629/jit.v4i1.5625

[6] Findawati, Y, et al. (2019). Comparative Analysis of Naïve Bayes, K Nearest Neighbor and C4.5 Method in Weather Forecast. *Journal of Physics: Conference Series*, 1402, 1-6.

[7] Pramudito, D.K. (2022). Data Mining Implementation on Java North Coast Weather Forecast Dataset Using C4.5 Algorithm. *Jurnal Teknologi Pelita Bangsa*, 13(3), 139-148.

[8] Loken, E.D., et al. (2019). Postprocessing Next-Day Ensemble Probabilistic Precipitation Forecast Using Random Forest. *American Meteorological Society, 34*, 2017-2033.

[9] Suryanto, A. A., Muqtadir, A. Penerapan Metode Mean Absolute Error Dalam Algoritma Regresi Linear Untuk Prediksi Produksi Padi. *Saintekbu: Jurnal Sains dan Teknologi, 11(1), 78-83.*

[10] Azmi, U., Hadi, Z. N., Soraya, S. (2020). Ardl Method: Foreasting Data Jumlah Hari Terjadinya Hujan di NTB. *Jurnal Varian, 3(2), 73-82.*