Rain Prediction Using Rule-Based Machine Learning Approach

Muchamad Taufiq Anwar[1], Saptono Nugrohadi[2], Vita Tantriyati[3], Vikky Aprelia Windarni[4]

[1]Faculty of Information Technology, Universitas Stikubank, Jl. Tri Lomba Juang No 1 Semarang 50241, Central Java, Indonesia

[2]Faculty of Engineering and Informatics, Universitas PGRI Semarang, Jl. Sidodadi-Timur No.24 Semarang, Central Java 50232, Indonesia

[3]Faculty of Information Technology, Universitas Kristen Satya Wacana, Jl. Dr. O. Notohamidjodjo, Salatiga 50715, Central Java, Indonesia

[4]Faculty of Computer Science, Universitas Amikom Yogyakarta, Jl. Ring Road Utara Yogyakarta, Special Region of Yogyakarta 55283, Indonesia

taufiq@edu.unisbank.ac.id

Abstract. Rain prediction is an important topic that continues to gain attention throughout the world. The rain has a big impact on various aspects of human life both socially and economically, for example in agriculture, health, transportation, etc. Rain also affects natural disasters such as landslides and floods. The various impact of rain on human life prompts us to build a model to understand and predict rain to provide early warning in various fields/needs such as agriculture, transportation, etc. This research aims to build a rain prediction model using a rule-based Machine Learning approach by utilizing historical meteorological data. The experiment using the J48 method resulted in up to 77.8% accuracy in the training model and gave accurate prediction results of 86% when tested against actual weather data in 2020.

Keywords: rain prediction, machine learning, J48, data mining

1. Introduction

Rain prediction is an important topic that continues to gain attention throughout the world. The rain has a big impact on various aspects of human life both socially and economically, for example in agriculture, health, transportation, etc. Rain also affects natural disasters such as landslides and floods. So much the impact of rain on human life, then we need a model to understand and predict predictions to provide early warning in various fields/needs such as agriculture, transportation, etc. Modeling can be made based on historical weather data that has been recorded by meteorological stations that are scattered in various locations in Indonesia. This data has been provided by the Climatology, Meteorology, and Geophysics Agency (BMKG) to be accessed by the public for various purposes including research purposes. It is known that Machine Learning / Data Mining can be used for weather prediction and forecasting[1][2]. This study aims to build a rain prediction model using a data mining approach by

utilizing historical meteorological data.

2. Methods

2.1. Research on Weather Predictions

Several studies on weather/rain prediction have been conducted. Some studies use a statistical approach while others use a data mining approach. Research on weather/rain prediction with a data mining / statistical approach is summarized in Table 1. In weather timeseries research, there are statistical approaches such as ARIMA, Exponential Smoothing[3], etc and Data Mining / Machine Learning such as Artificial Neural Networks, etc. [4]. Some studies combine the elements of weather prediction to be associated with certain phenomena such as Dengue Fever [5], agriculture [6], dan foods[7].

Table 1. Research on weather/rain prediction

Reference	Variables	Method
[8]	Temperature	Fuzzy
[9]	Barometric pressure, temperature, dew point, humidity, wind speed	Fuzzy
[2]	Temperature, rainfall, humidity, exposure time, duration of fog, evaporation, wind, atmospheric pressure, number of clouds	Decision trees, bagging, random forests, and boosting
[10]	Minimum temperature, maximum temperature, rainfall	Multiple Linear Regression
[11]	Temperature, air pressure, relative humidity, vapor pressure, wind speed	Bayesian
[12]	Maximum humidity, average humidity, rainfall	Naïve Bayes
[13]	Temperature, wind speed, wind direction, humidity, atmospheric pressure, rainfall	Multiple Linear Regression
[14]	Maximum temperature, minimum temperature, evaporation, wind speed, cloud cover	J48, ANN, dan Naïve Bayes

2.2. Decision Tree

The Decision Tree (DT) model has a top-down hierarchical structure that describes the rules for dividing large data sets into small groups given a specific target variable. There are three distinct algorithms for categorical target variables in the DT model, i.e Entropy Reduction, Gini, and Chi-square tests. Previously, research on weather forecasting and climate change found that models produced using DT have small errors compared to other techniques in predicting data mining with large historical data [15], [16].

The DT model is one of the most powerful and useful for predictions that explore large and complex data. The mechanism in the DT model is transparent and is produced easily to understand the model for researchers. Besides, the DT model can convert raw data into information in a simple way, by complying with a set of rules that can be read by humans. The resulting structure represents a decision or rule for the classification of datasets. These rules are made to make groups as homogeneous as possible in terms of response variables. At each step, the input variable is used to divide the observations into groups. If the specified input values are identified to have a strong relationship with the response values, then all of these values are grouped in the same branch in the decision tree.

Trees can fit better as the grouping of observation data split into smaller groups (i.e 'branches'). In this situation, the DT model will remember data patterns rather than generalize them. The pruning algorithm in the DT model helps overcome the problem of overfitting by pruning trees using certain algorithms, namely, CART, CHAID, and C4.5. CART and CHAID both use the Gini Index and Chi-squared test, respectively, to classify records in the target variable. This research uses the C4.5 / J48 method which uses the measure of Entropy and information gain.

2.3. C4.5 / J48 Algorithm

C4.5 is the successor of the Iterative Dichotomiser 3 (ID3) algorithm developed by the same author, Ross Quinlan, in 1993 [17]. This has several improvements over the original ID3 such as the ability to handle continuous and discrete attributes and the ability to prune trees after it is created. C4.5 works by creating trees based on entropy and information gain to select which attributes are useful in classifying the data. Entropy is a measure of the heterogeneity of data, while information-gain is a measure of how much information is obtained by comparing entropy before and after separating the dataset based on certain attributes. Formulas for entropy and information-gain are shown in (1) and (2) respectively. Pruning in C4.5 is based on the confidence factor. Pruning is useful for minimizing model overfitting and reducing tree size but in lower model accuracy costs. The well-known implementation of C4.5 is the J48 function which is written in Java is provided within the Waikato Environment for Knowledge Analysis (WEKA) software [18]. The pseudocode for the C4.5 algorithm is shown in Algorithm 1 [19]. J48 will produce a tree by which the rules could be easily read by humans. This J48 method has also been used to find rules for forest fire cases in Indonesia[20]. Research [21] also showed that a decision tree is very suitable for rain prediction.

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \tag{1}$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

Algorithm 1: C4.5

Input: an attribute-valued dataset D
1: Tree = {}
2: **if** D is ”pure” OR other stopping criteria met **then**
3: terminate
4: **end if**
5: **for all** attribute a ∈ D **do**
6: Compute information-theoretic criteria if we split *a*
7: **end for**
8: a_best=Best attribute according to above-computed criteria
9: Tree = Create a decision node that tests a_best in the root
10: D_v = Induced sub-datasets from D based on a_best
11: **for all** D_v **do**
12: Tree v = C4.5(D_v)
13: Attach Tree v to the corresponding branch of Tree
14: **end for**
15: **return** Tree

The research methodology is shown in Figure 1. Daily historical weather data was obtained from the BMKG website for the Tanjung Mas meteorological station, in Semarang City, Indonesia. The original data consisted of 12 attributes, but for this study, only 8 attributes were used, as shown in Table 1. The attribute of wind direction was not used since the numerical scale was not appropriate for this study. One additional attribute is added, i.e the class which shows whether it rained or not on each particular day. The class is obtained by evaluating the RR (rainfall) attribute column, if RR> 0 then class = ‘rain’; otherwise, class = ‘norain’. Data cleaning is done to remove entries with missing values. Data is then

stored in CSV format and then converted to the ARFF file format to be able being processed using the WEKA software. Experiments were carried out using the J48 function under the classification tab. The attributes of the meteorological data are shown in Table 2.

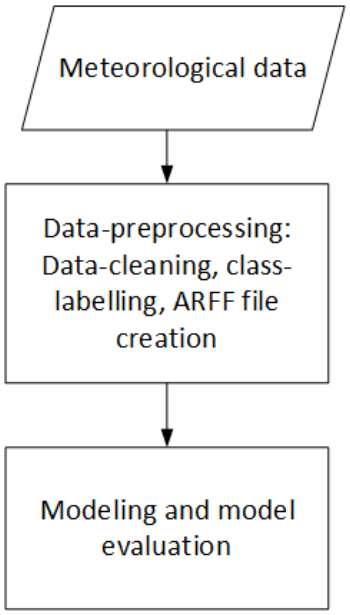

Figure 1. Research Methodology

Table 2. Attributes of the meteorological data

Attribute	Data type	Description
Tn	Numeric	Minimum temperature
Tx	Numeric	Maximum temperature
Tavg	Numeric	Average temperature
RH_avg	Numeric	Average Humidity (%)
RR	Numeric	Rainfall (mm)
ss	Numeric	Sun exposure time (hours)
ff_x	Numeric	Maximum wind speed (m/s)
ff_avg	Numeric	Average wind speed (m/s)

3. Results and Discussion

Meteorological data gathered from the year 2013 to 2019 with a total of 2536 rows of data were used in this experiment. Evaluation of model accuracy is done by using the 10-fold cross-validation for the training data and tested against actual weather data in 2020. The training model gives an accuracy of 77.8% whereas the results of experiments against 2020 data gave an accuracy of 86%. The lower accuracy of the trained model might be caused by the overfitting of the model or that there is a huge variation in the large amount of training data being used to build the model. These findings also agree with another research which showed that decision trees and k-mean clustering are best-suited data mining techniques for weather data, with the increase in the size of the training set, the accuracy is first increased but then decreased after a certain limit [21]. The model also showed that the factors that predict rain the most are the average humidity (RH_avg), followed by the minimum temperature (Tn). The hight accuracy achieved by the J48 method is in line with other research which stated that the Decision Tree model is better as compared to the other predictive models [14]. The resulted tree which also shows the

rules is shown in Figure 2. The model accuracy on various minumum number of cases per leaf is shown in Table 3.

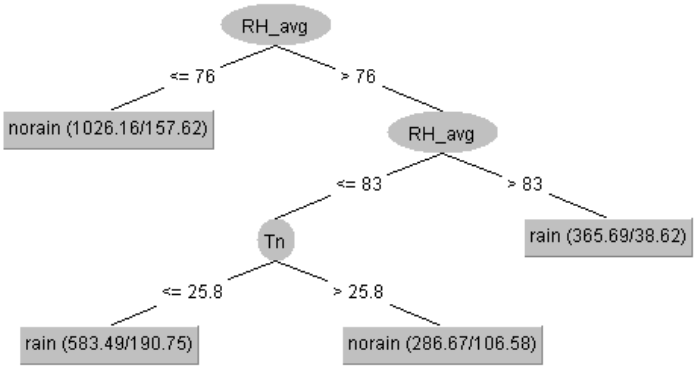

Figure 2. The (simplified) decision tree produced by J48 (with minumum case of 10 per leaf)

Table 3. The model accuracy on various number of minimum case per leaf

Minimum cases per leaf	Model accuracy (%)
2 (default)	76.0
5	76.1
10	77.4
20	77.7
100	77.8

4. Conclusion

A rain prediction model is very useful for human activities. This research attempted to build a rain prediction model by using a rule-based machine learning approach applied to historical meteorological data. The decision tree model produced by the J48 algorithm could give an accuracy up to 77.8% from the training data and give an accuracy of 86% when tested against actual weather data in 2020. The result showed that rainfall is mainly affected by the average humidity and by minimum temperature for a particular day of observation. This result gave us a better understanding of the phenomenon of rain and the model could be used for several purposes such as in agriculture, transportation, etc.

References

[1] M. R. Mahmood, R. K. Patra, R. Raja, and G. R. Sinha, "A novel approach for weather prediction using forecasting analysis and data mining techniques," in *Innovations in Electronics and Communication Engineering*, Springer, 2019, pp. 479–489.

[2] C. Choi, J. Kim, J. Kim, D. Kim, Y. Bae, and H. S. Kim, "Development of heavy rain damage prediction model using machine learning based on big data," *Adv. Meteorol.*, vol. 2018, 2018.

[3] K. D. Hartomo, S. Y. J. Prasetyo, M. T. Anwar, and H. D. Purnomo, "Rainfall Prediction Model Using Exponential Smoothing Seasonal Planting Index (ESSPI) For Determination of Crop Planting Pattern," in *Computational Intelligence in the Internet of Things*, IGI Global, 2019, pp. 234–255.

[4] N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay, "A comprehensive survey of data mining techniques on time series data for rainfall prediction," *J. ICT Res. Appl.*, vol. 11, no. 2, pp. 168–184, 2017.

[5] N. Agarwal, S. R. Koti, S. Saran, and A. S. Kumar, "Data mining techniques for predicting dengue outbreak in geospatial domain using weather parameters for New Delhi, India," *Curr. Sci.*, vol. 114, no. 11, pp. 2281–2291, 2018.

[6] P. S. Tayde, B. K. Patil, and R. A. Auti, "Applying Data Mining Technique to Predict Annual Yield of Major Crops," *Int. J.*, vol. 2, no. 2, 2017.
[7] U. K. Dey, A. H. Masud, and M. N. Uddin, "Rice yield prediction model using data mining," in *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2017, pp. 321–326.
[8] K. Kar, N. Thakur, and P. Sanghvi, "Prediction of Rainfall Using Fuzzy Dataset," 2019.
[9] N. Z. M. Safar, A. A. Ramli, H. Mahdin, D. Ndzi, and K. M. N. K. Khalif, "Rain prediction using fuzzy rule based system in North-West Malaysia," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1572–1581, 2019.
[10] E. Sreehari and S. Srivastava, "Prediction of Climate Variable using Multiple Linear Regression," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 2018, pp. 1–4.
[11] V. B. Nikam and B. B. Meshram, "Modeling rainfall prediction using data mining method: A Bayesian approach," in *2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation*, 2013, pp. 132–136.
[12] S. Navadia, P. Yadav, J. Thomas, and S. Shaikh, "Weather prediction: a novel approach for measuring and analyzing weather data," in *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, 2017, pp. 414–417.
[13] N. Anusha, M. S. Chaithanya, and G. J. Reddy, "Weather Prediction Using Multi Linear Regression Algorithm," in *IOP Conference Series: Materials Science and Engineering*, 2019, vol. 590, no. 1, p. 12034.
[14] N. W. Zamani and S. S. M. Khairi, "A comparative study on data mining techniques for rainfall prediction in Subang," in *AIP Conference Proceedings*, 2018, vol. 2013, no. 1, p. 20042.
[15] A. Joshi, B. Kamble, V. Joshi, K. Kajale, and N. Dhange, "Weather forecasting and climate changing using data mining application," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 3, pp. 19–21, 2015.
[16] F. Sheikh, S. Karthick, D. Malathi, J. S. Sudarsan, and C. Arun, "Analysis of data mining techniques for weather prediction," *Indian J. Sci. Technol.*, vol. 9, no. 38, 2016.
[17] J. Quinlan, *C4. 5: programs for machine learning*. 2014.
[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.
[19] P. Nevlud, M. Bures, L. Kapicak, and J. Zdralek, "Anomaly-based network intrusion detection methods," *Adv. Electr. Electron. Eng.*, vol. 11, no. 6, pp. 468–474, 2013.
[20] M. T. Anwar, H. D. Pumomo, S. Y. J. Prasetyo, and K. D. Hartomo, "Decision Tree Learning Approach To Wildfire Modeling on Peat and Non-Peat Land in Riau Province," in *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2018, pp. 409–415.
[21] R. S. Kumar and C. Ramesh, "A study on prediction of rainfall using datamining technique," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, 2016, vol. 3, pp. 1–9.