Predicting Rainfall Intensity using Naïve Bayes and Information Gain Methods (Case Study: Sleman Regency)

I G W Sena¹, J W Dillak², P Leunupun³ and A J Santoso⁴

¹,²,³,⁴Universitas Atma Jaya Yogyakarta, Jl. Babarsari No.44, Janti, Caturtunggal, Depok, Sleman, Daerah Istimewa Yogyakarta 55281
E-mail: albjoko@staff.uajy.ac.id

Abstract. Climate change, which has an impact on environmental problems in tropical countries, is still a severe problem, and efforts to prevent and manage it are continuously pursued. Indonesia, as a tropical country with topographical conditions and strategic geographical position, causes Indonesia to have different weather and climate patterns. Climatologically there are significant differences between the rainy season and the dry season. Both these seasons can bring blessings but also disasters if not appropriately handled, flooding in the rainy season and drought in the dry season. High rainfall can cause floods and landslides, whereas if using excess water in the rainy season can be a solution for water shortages in the dry season. The purpose of this study is to predict the rainfall intensity with the Naïve Bayes method and what parameters are considered the most dominant causes of heavy rainfall using the information gain method. The source of the data in this study came from BMKG data, which was during the rainy season between October to March from 2016 - 2019 in the Sleman Regency. The results of this study showed that the Naïve Bayes method could be used to predict rainfall intensity in Sleman Regency. Also, the most influential parameter on rainfall intensity is the average temperature with an information gain value of 0.047811028.

1. Introduction

Various conditions arising from climate change are global warming, which affects work productivity and impacts on economic decline [1], increased global emissions, greenhouse effect, warmer sea, rising sea level. These conditions cause damage to coral reefs and coastal settlements, high strength of tropical cyclones and rainstorms, groundwater that cannot be consumed because it becomes saltier [2].

As a tropical country with a rainy and dry season, Indonesia also feels the impact of climate change that is happening. Rising land-sea temperature gradients, affecting rainfall patterns that also impact human life, drought, floods, health, water and food supplies, changes in ecosystems, and infrastructure [3]. One way to overcome the disaster is through crisis planning and control, by predicting Iranian rainfall applying rainwater harvest management that has been proven to control the drought crisis and save water needs [4].

Rainfall intensity prediction can be useful information for planning flood and drought risk management, also could be useful for preparing outdoor activities. The purpose of this study is to determine the prediction of rainfall intensity and determine which parameters are the most dominant causes of rainfall intensity using Naïve Bayes method and Information Gain in Sleman Regency, Daerah Istimewa Yogyakarta (DIY) Province, Indonesia. Researchers used the Naïve Bayes method because, this method is considered more accurate to be used in terms of comparison and prediction [5].
2. Literature Review

Data Mining (DM) is one of the fields in computer science that demands the extraction of hidden information from various data storage, data marts, and repositories [6]. DM allows researchers to find beneficial information from available data. This discovery is significant in business because it can be used as a key for making decisions [7]. There are several methods to do DM, such as decision tree, neural networks, Naïve Bayes, K-nearest neighbor, etc. This research will use the Naïve Bayes method to predict the student's performance.

The Naïve Bayes method has a simple algorithm but can produce better results than other methods, mainly when used at high input dimensions [8]. Research by T. Devasia et al. [5] using the Naive Bayes method to predict student performance in Amrita Vishwa Vidyapeetham University, Mysuru. In this study, there were 700 responses with 19 attributes. The results of this study indicate that, in terms of accuracy and comparison, the Naive Bayes method algorithm is more accurate compared to other methods such as Neural Network, Decision Tree, etc.

Another research to [9] predicted rainfall intensity in Malang City using Naïve Bayes and Laplace Estimator method. The result of this research proved that Naïve Bayes and Laplace Estimator method could be used to predict rainfall time series for seven days with a percentage accuracy of 97.74%, 97.74% for precision, 100% for sensitivity and error rate of 2.26%.

Conceptually, the studies above are not much different from this research, which is to make predictions using Naïve Bayes. However, this study focuses on the prediction of rainfall intensity in Sleman Regency by using four parameters, namely average temperature, average humidity, duration of sun exposure, and average wind speed. This study will also determine which parameters most influence the intensity of rainfall using Information Gain, which is not done in the studies above.

3. Material and Methods

This section will explain some of the stages of researchers in the process of analyzing data. This stage is divided into data collection, data pre-processing, data analysis.

![Research stages in analyzing data](image)

**Figure 1. Research stages in analyzing data**

3.1. Data collection

Data sources in this study were obtained from the Indonesian Meteorology, Climatology and Geophysics Agency, or Badan Meteorologi Klimatologi dan Geofisika (BMKG) Indonesia, which was downloaded on the official BMKG website (http://dataonline.bmkg.go.id/). Data that have been downloaded are rainfall data, as well as data for each parameter used to predict rainfall, include average temperature, average humidity, duration of sun exposure, and average wind speed in Sleman Regency, Province of DIY, Indonesia. Data collected are only the data during the rainy season in Indonesia, namely in the range of October - March with a total of 577 data, starting from October
(2016-2018), November (2016-2018), December (2016-2018), January (2017-2019), February (2017-2019), and March (2017-2019).

3.2. Data pre-processing
After the data is collected, the next step is data pre-processing. Pre-processing data is a necessary step for preparing the dataset before applying classification techniques [7]. At this step, all the data is inputted in Microsoft Excel, then checking whether there is an error or invalid data to be used. The unused data is the rainfall data with a value of 0, as well as data with a value of 8888 (unmeasured data) and 9999 (no data or no measurement). From a total of 577 data, after passing the pre-processing step, there were 329 valid data and 248 invalid data.

At this step, classification and weighting of class values are also carried out for each parameter. The researcher classified each parameter into three classes (Table 1). Weighting class values are based on the minimum and maximum values of each parameter, based on the data that has been pre-processed before.

| Parameter Class | Class Value |
|-----------------|-------------|
| Average temperature (°C) | 25 |
| Low | ≤ 25 |
| Medium | > 25 - ≤ 26.5 |
| High | > 26.5 |
| Average humidity (%) | 84 |
| Fairy Damp | ≤ 84 |
| Damp | > 84 - ≤ 91 |
| Very Damp | > 91 |
| Sun exposure duration (hour) | 3.3 |
| Briefly | < 3.3 |
| Medium | ≥ 3.3 - < 6.6 |
| Long | ≥ 6.6 |
| Average wind speed (m/s) | 2 |
| Slow | < 1 |
| Medium | ≥ 1 - < 2 |
| Fast | ≥ 2 |

While for the classification of rainfall, researchers are based on the standard categories of rainfall from BMKG, namely, (1) low categories ranging from 1-5 mm/hour; or 5-20 mm/day, (2) medium categories ranging from 5-10 mm/hour; or 20-50 mm/day, (3) heavy categories ranging from 10-20 mm/hour; or 50-100 mm/day, and (4) very heavy categories > 20 mm/hour; or > 100 mm/day. However, in this study, we divide rainfall into only three categories or classes, because based on the data that we have got, there is no value above 100 mm/day.

| Class | Class Value (mm/day) |
|-------|---------------------|
| Low   | ≤ 20                |
| Medium| > 20 - ≤ 50         |
| High  | > 50                |
3.3. Data Analysis

Researchers use the Naïve Bayes method to predict rainfall, and Information Gain to determine the most influential parameters in predicting rainfall.

3.3.1. Naïve Bayes

The method invented in the 18th century by Thomas Bayes used the Bayes theorem. The theorem is used based on conditional probabilities based on the past for future probabilities [10]. The Naïve Bayes equation is shown in formula (1).

\[ P(H|X) = \frac{P(X|H)P(H)}{P(X)} \]  

(1)

with, \( P(H|X) \) = the probability that the hypothesis H is true for condition X; \( P(X|H) \) = the probability that proof X is true for hypothesis H; \( P(H) \) = the previous probability; \( P(X) \) = the observed probability; \( X \) = proof (data sample); and \( H \) = hypothesis.

3.3.2. Information Gain

To find the effectiveness of each parameter, researchers used information gain as a method in this analysis to determine which parameters are most influential for high rainfall [10]. Information gain for parameter A is shown in formula (2).

\[ Gain(S,A) = \text{Entropy}(S) - \sum_{\text{values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \]  

(2)

with, \( A \) = parameter; \( V \) = the declared value for parameter A; \( \text{Values}(A) \) = a set of values of A; \( |S_v| \) = the number of samples for value v; \( |S| \) = the sum of all data samples; and \( \text{Entropy} \) = entropy for samples that have a value of v.

For entropy itself has its own calculations. Entropy is formulated in equation (3).

\[ \text{Entropy}(S) = \sum_i c - p_i \log_2 p_i \]  

(3)

with, \( c \) = the number of values in the target parameter; and \( p_i \) = the ratio between the number of samples in class i.

4. Result

4.1. Predicting Rainfall Intensity with Naïve Bayes

There are two types of data used as test material, such as data-training and data-testing. Data-training is the data result in the pre-processing step with 329 data. Meanwhile, data-testing is a combination of class of each parameter, that is not contained in the data-training. This research used five parameter class combinations for data-testing. The result of data-training testing can be seen in Table 3.
Table 3. Data-training testing to predict rainfall intensity

| No | Average temperature | Average humidity | Sun exposure duration | Average wind speed | Low      | Medium     | High      |
|----|---------------------|------------------|-----------------------|-------------------|----------|------------|-----------|
| 1  | High                | Very Damp        | Long                  | Slow              | 0.00151  | 0.00018    | 0.00009   |
| 2  | Low                 | Damp             | Long                  | Slow              | 0.00156  | 0.00073    | 0.00123   |
| 3  | Medium              | Fairy Damp       | Briefly               | Slow              | 0.01139  | 0.00196    | 0.00030   |
| 4  | High                | Very Damp        | Briefly               | Slow              | 0.00582  | 0.00095    | 0.00010   |
| 5  | Low                 | Damp             | Briefly               | Medium            | 0.00816  | 0.00590    | 0.00267   |

Prediction result in Table 3 shows that, (1) the highest entropy score of rainfall intensity is in low class which is mean that for condition number 1 the rainfall intensity is low; (2) the highest entropy score of rainfall intensity is in low class which is mean that for condition number 2 the rainfall intensity is low; (3) the highest entropy score of rainfall intensity is in low class which is mean that for condition number 3 the rainfall intensity is low; (4) the highest entropy score of rainfall intensity is in low class which is mean that for condition number 4 the rainfall intensity is low; and (5) the highest entropy score of rainfall intensity is in low class which is mean that for condition number 5 the rainfall intensity is low.

4.2. Effectiveness of parameters with Information Gain

In this section, the results of the calculation of the most influential parameters will be displayed using the information gain method. The first step is to search for the Entropy (S) value and look for the entropy value of each parameter. The next step is to find the gain value of each parameter. The following results of the calculation of each gain value will be displayed using a pie chart in Figure 2.

![Figure 2. Gain of each parameter](image)

Based on the formula in Chapter 3 (Material and Method), the value of Entropy(S) is 1.06393. Then, this value is used to calculate information gain. Based on calculation using Information Gain, the most influential parameter in predicting rainfall is the average temperature with the amount of gain value is
0.047811028. The second one is sun exposure duration with 0.026614401, followed by average humidity and average wind speed with 0.01290302 and 0.012079021.

5. Conclusion

Based on the results of this study, it can be concluded that the Naïve Bayes method can be used to predict rainfall intensity in Sleman Regency. However, keep in mind, the standard class value for each parameter is different in each region. Also, based on the four parameters analyzed using Information Gain, it can be concluded that the most influential parameter on rainfall intensity is the average temperature with an entropy value of 0.047811028.

6. References

[1] T. Kjellstrom, “Impact of Climate Conditions on Occupational Health and Related Economic Losses: A New Feature of Global and Urban Health in the Context of Climate Change,” Asia-Pacific J. Public Heal., vol. 28, pp. 28S-37S, 2014.
[2] T. Weir, L. Dovey, and D. Orcherton, “Social and cultural issues raised by climate change in Pacific Island countries: an overview,” Reg. Environ. Chang., vol. 17, no. 4, pp. 1017–1028, 2017.
[3] R. Chadwick, P. Good, G. Martin, and D. P. Rowell, “Large rainfall changes consistently projected over substantial areas of tropical land,” Nat. Clim. Chang., vol. 6, no. 2, pp. 177–181, 2016.
[4] M. Ravanshadnia, M. Ghanbari, and M. H. Fardani, “Planning Crisis Management of Water Resources in Tehran in the Next Twenty Five Years,” J. Appl. Environ. Biol. Sci., vol. 5, no. 10, pp. 50–59, 2015.
[5] T. Devasia, T. P. Vinushree, and V. Hegde, “Prediction of students performance using Educational Data Mining,” Proc. 2016 Int. Conf. Data Min. Adv. Comput. SAPIENCE 2016, pp. 91–95, 2016.
[6] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, “Educational data mining and analysis of students’ academic performance using WEKA,” Indones. J. Electr. Eng. Comput. Sci., vol. 9, no. 2, pp. 447–459, 2018.
[7] A. Mueen, B. Zafar, and U. Manzoor, “Modeling and Predicting Students’ Academic Performance Using Data Mining Techniques,” Int. J. Mod. Educ. Comput. Sci., vol. 8, no. 11, pp. 36–42, 2016.
[8] C. Anuradha and T. Velmurugan, “Feature Selection Techniques To Analyse Student Academic Performance Using Naïve Bayes Classifier,” 3rd Int. Conf. Small Mediu. Bus., no. February, pp. 345–350, 2016.
[9] M. Muthmainnah, M. Ashar, I. M. Wirawan, and T. Widiyaningtyas, “Time Series Forecast for Rainfall Intensity in Malang City with Naive Bayes Methodology,” 3rd Int. Conf. Sustain. Inf. Eng. Technol. SIET 2018 - Proc., pp. 137–141, 2019.
[10] Suyatno, DATA MINING UNTUK KLASIFIKASI DAN KLASTERISASI DATA. Penerbit INFORMATIKA, 2017.