Application Of Naïve Bayes to Predict the Potential of Rain in Ternate City

Alwin Ali*, Amal Khairan, Firman Tempola and Achmad Fuad

Departmen Informatic, Faculty Engineering Khairun University, 97728 Ternate, Indonesia

Abstract. The amount of rainfall that occurs cannot be determined with certainty, but it can be predicted or estimated. In predicting the potential for rain, data mining techniques can be used by classifying data using the naive Bayes method. Naïve Bayes algorithm is a classification method using probability and statistical methods. The purpose of this study is how to implement the naive Bayes method to predict the potential for rain in Ternate City, and be able to calculate the accuracy of the Naive Bayes method from system created. The highest calculation results with new data with a total of 400 training data and 30 test data, obtained 30 correct data with 100% precision, 100% recall and 100% accuracy and the lowest calculation results with new data with a total of 500 training data and 50 test data, obtained 38 correct data and 12 incorrect data with a percentage of precision 61.29%, recall 100% and accuracy 76%.

Keywords: Naïve Bayes, Statistical Methods, Data Mining

1 Introduction

North Maluku Province is geographically located at 0°-2° North Latitude and 126°-128° East Longitude. Ternate City and also generally coastal areas in North Maluku Province have a tropical climate type that is influenced by the marine climate which is usually heterogeneous according to general indications of a tropical climate. This area is known for its two seasons, namely north-west and east-south which are often interspersed with two transition periods each year.

Weather is the condition of the air in a place in a relatively short time, which is expressed by parameter values such as wind speed, temperature, pressure, rainfall, and other atmospheric phenomena as the main components. Weather is an important thing that will never be separated from human life. The course of human activities can be influenced by weather conditions [1].

Rain is a form of precipitation of water vapor from clouds in the atmosphere. The amount of rainfall that occurs is not can be determined with certainty, but can be predicted or estimated. By using historical data on the amount of rainfall some time ago, it can be predicted how much rainfall will occur in the future [2].

One approach used in predicting the potential for rain is to utilize the concept of data mining. Data mining is used to analyze and find patterns in weather changes and to model the relationship between weather variables. From the resulting model, calculations can be made to predict the condition or value of the weather variable you want to know, for example the level of

rainfall. One of the data *mining* methods commonly used to model the relationship between variables is *Naïve Bayes* [3].

Naïve Bayes is a simple probability classifier that calculates a set of probabilities by adding up the frequencies and combinations of values from a given dataset. The algorithm uses Bayes theorem and assumes all independent attributes are not interdependent by the value of the class variable [4]. The application of naive bayes has been carried out by [5] on volcanic activity data. This system will also be tested for algorithm performance by applying accuracy, precision, and recall tests.

2 Methodology

2.1 Data Mining

Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from large databases. The term data mining has the essence as a discipline whose main goal is to find, explore, or mine knowledge from the data or information that we have. Data mining, often also referred to as Knowledge Discovery in Database (KDD). Knowledge Discovery in Database is an activity that includes collecting, using historical data to find regularities, patterns or relationships in large data sets [5].

* Corresponding author : alwinalisaja@gmail.com

2.2 Naive Bayes Algorithm

The Naïve Bayes algorithm is one of the algorithms contained in the classification technique. Naïve Bayes is a classification with probability and statistical methods proposed by British scientist Thomas Bayes, which predicts future opportunities based on previous experience, so it is known as Bayes' theorem. The theorem is combined with Naïve which assumes conditions between independent attributes. The Naïve Bayes classification assumes that a trait is present or not certain of a class has nothing to do with the characteristics of the class [6].

The advantage of using this method is that it only requires a small amount of training data to determine the parameter estimates needed in the classification process. Because it is assumed to be an independent variable, only the variance of a variable in a class is needed to determine the classification, not the entire covariance matrix. Naïve Bayes Stages [7].

Calculating the *Mean* and *Standard Deviation* (Numeric Data). Finding the mean can be seen in equation 1. While to find the standard deviation used equation 2.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (1)$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (2)$$

1. Calculating Probability (Discrete Data) can be seen in equation 3.

$$P(E) = X/N \quad (3)$$
2. Calculating gauss density can be seen in equation 4.

$$f(x_i) = \frac{1}{\sqrt{2\pi}s} e^{-\frac{(x-\bar{x})^2}{2s^2}} \quad (4)$$

3. Calculating Likelihood can be seen in the equation 5.

$$L(x) = f(x_1) \times f(x_2) \times \dots \times f(x_n) \quad (5)$$

4. Calculating Likelihood Probability is seen in equation 6.

$$P(L_i) = \frac{L_i}{\sum_{i=1}^n L_i} \quad (6)$$

5. Laplace Smoothing

In a large dataset, random selection of training data will lead to the possibility of zero values in the probability model. These zero values will cause the Naïve Bayes Classifier to be unable to classify an input data. Therefore, we need a smoothing method that can avoid zero values in the probability model. Laplacian Smoothing is a smoothing method commonly used in the Naïve Bayes Classifier. Laplacian Smoothing is commonly known as add one smoothing, because in its calculation, each variable in each parameter is added by 1 [8]. The following is the Laplace Smoothing equation (7).

$$P(X_k|C) = \frac{P(X_k|C)+1}{P(C)+|V|} \quad (7)$$

2.3 Confusion Matrix

Confusion matrix is a method that is usually used to perform accuracy calculations on the concept of data mining. The confusion matrix is depicted by a table that states the number of test data that is correctly classified and the number of test data that is incorrectly classified [9].

Prediction accuracy needs to be done to see the percentage of accuracy of the Naive Bayes method system in predicting patterns. The accuracy of the model predictions Naive Bayes measured precision, recall, accuracy [10].

Precision is the level of accuracy between the information requested by the user and the answer given by the system. To find the value of precision used equation 8.

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

Recall is the success rate of the system in retrieving information. to find the recall value used equation 9.

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

Accuracy is defined as the degree of closeness between the value of proximity to the actual. to get the accuracy value, that is, all correct data is divided by the number of test data, as shown in Equation 10.

$$Accuracy = \frac{TP}{TP + FP + FN + TN} \quad (10)$$

2.4 Methode Prototype

The prototyping method begins with gathering requirements, involving system developers and users to determine the objectives, functions and operational requirements of the system [11]. The following is an illustration of the prototyping model which can be seen in Figure 1.

Fig. 1. Model Prototype

3 RESULT AND DISCUSSION

3.1 Dataset

The dataset in this study is data collected from the City of Ternate (BMKG). The criteria used in the study are wind speed, temperature, humidity, air pressure,

weather phenomena and brightness. The data used are rainy weather data, the amount of data used is 797 data where data is No Rain 616, Light Rain 172, Moderate Rain 4 and Heavy Rain 5. Data obtained from the Meteorology, Climatology and Geophysics Agency (BMKG) of Ternate City. Output specifications displayed are Light Rain, Moderate Rain, Heavy Rain. The dataset results obtained are as shown in Table 1.

Table 1. Rainfall dataset

Data ke	Wind velocity °C	Temperature m/s	humidity RH	Air pressure hPa	Weather phenomenon	brightness	Status CH
1	5	27.8	86	1012.9	cloudy	cloudy	Not raining
2	4	28.4	79	1013.5	cloudy	cloudy	Not raining
3	6	29.6	77	1013.1	cloudy	Sunny cloudy	Not raining
4	5	30.6	71	1012.3	cloudy	Sunny cloudy	Not raining
5	9	31	74	1011.2	cloudy	Sunny cloudy	Not raining
6	10	31.4	74	1010.1	cloudy	Sunny cloudy	Not raining
7	12	31	74	1009.8	cloudy	Sunny cloudy	Not raining
8	12	30.5	75	1010.1	cloudy	Sunny cloudy	Not raining
9	8	29.4	80	1010.1	cloudy	Sunny cloudy	Not raining
10	6	28.8	82	1011.1	cloudy	Sunny cloudy	Not raining
-	-	-	-	-	-	-	-
500	3	27.6	89	1011.7	rain	cloudy	Light rain
-	-	-	-	-	-	-	-
797	11	30.6	67	1011.7	cloudy	Sunny cloudy	Not raining

3.2 Flowchart Naïve bayes

Flowchart for Naïve Bayes calculation can be seen in Figure 2.

Fig. 2 Flowchart naïve bayes

Based on Figure 2, there are several steps that will be carried out in the Naïve Bayes method:

1. Input the data to be calculated.
2. If the inputted data contains numeric data, find the mean and standard deviation. If the inputted data contains discrete data, calculate it using the probability formula.
3. After getting the mean and standard deviation, calculate the data using gauss density.

4. Calculating the likelihood of the calculation of the gauss density and probability.
5. Calculate likelihood probability.
6. choose a class as the result of prediction or classification based on the final result of the calculation seen from the largest likelihood value.

3.3 Implemented Naïve bayes

At this stage the rainfall data is calculated using the Naïve Bayes method. Naïve Bayes calculation steps. The training data that will be used to test are 797 training data, as follows:

3.3.1 The first data testing

The test on this first test data has the data as shown in Table 2. while in Table 3 is the result of prediction of weather forecast data based on the data in Table 2.

Tabel 2. The First data testing

No	Kriteria	Data
1	Wind velocity	11
2	temparature	27
3	humadity	90
4	Air pressure	1011.5
5	Weaahther phenomenon	Cloudy
6	brightness	Cloudy

Table. 3 The results of the first data test

Categori	Probability Likelihood	Prediction
No raining	4.75524316E-7	Not raining
Light rain	3.21884042E-7	
Moderate rain	0	
Heavy rain	0	

3.3.2 The Second data testing

Table 4 is the data for which the prediction process will be carried out. as for the prediction results can be seen in Table 5.

Table 4. The second data testing

No	Kriteria	Data
1	Wind velocity	10
2	temparature	29
3	humadity	90
4	Air pressure	1010
5	Weahtther phenomenon	Cloudy
6	brightness	Cloudy

Table 5. Result of the second data testing

Categori	Probability Likelihood	Classify
Not raining	7.30093923E-7	Not rain
Light rain	1.47905146E-7	
Moderate rain	0	
Heavy rain	0	

3.3.3 The third data testing

Table 6 is the third data tested to predict the weather in the city of Ternate. the prediction results are shown in table 7.

Table 6. The third data testing

No	Criteria	Data
1	Wind velocity	8
2	Temaparature	27.5
3	humadity	86
4	Air pressure	1012.2
5	Weahtther phenomenon	Cloudy
6	brightness	Cloudy

Table. 7. Result of the third data testing

Categori	Probability Likelihood	classify
Not raining	2.734539339E-6	Not raining
Light rain	1.173954139E-6	
Moderate rain	0	
Heavy rain	0	

3.4 Result of accuracy system

In this study, several test scenarios were carried out to see the performance of the Naive Bayes algorithm on weather data in the city of Ternate. in measuring algorithm performance including accuracy, precision and recall tests. test scenarios three times by applying different amounts of training data. as for the test results are shown in Figure 3, Figure 4 and Figure 5.

Fig. 3. Result testing scenario 1.

Fig. 4. Result testing scenario 2.

Fig. 3. Result testing scenario 3.

Based on the results of tests carried out by the system using the naive Bayes method, the results with the highest accuracy were obtained, namely in the 3rd experiment with 400 training data and 30 test data, it managed to get an accuracy of 100% with precision and recall values also 100% respectively. while the lowest accuracy is in the 1st experiment with 500 training data and 50 test data, managed to get an accuracy of 76%. the amount of training data and testing data cannot affect the level of accuracy, this is because in taking training data or testing data it is done randomly. Otherwise, unbalanced labels are applied.

4 CONCLUSION

How to implement the Naïve Bayes method, namely, determine how many data sets are then perform calculations from the dataset, namely calculating the mean and standard deviation for numeric data after the data is calculated then calculate the probability for discrete data then calculate for new data, after that calculate likelihood then calculate the likelihood probability after that the results and predictions are shown. The results obtained that the Naïve Bayes method was successfully implemented to determine the potential for rain in Ternate City with the amount of data obtained as many as 797 data where the data is No Rain 616, Light Rain 172, Moderate Rain 4 and Heavy Rain 5. The criteria used in the study are: Wind speed, temperature, humidity, air pressure, weather phenomena and brightness. From the system accuracy test, 3 tests were carried out, where the highest result was 100% with training data used as much as 300 and testing data as much as 300 while, the lowest accuracy was 90% where the training data used is 600 and the testing data is 40. Testing of the system by using a white box indicates that the implementation of the system is successfully created.

References

- [1] A. Subhan, and A. Z. Fanani, A. 'Penerapan Data Mining Untuk Menentukan Potensi Hujan Harian Dengan Menggunakan Algoritma Naive Bayes', *Jurnal Informatika*, **6**, 1., pp. 1–6. (2016)
- [2] C. Oktaviani., & Afdal. *Prediksi Curah Hujan Bulanan Menggunakan Jaringan Syaraf Tiruan Dengan Beberapa Fungsi Pelatihan Backpropagation*. **2**, 4, pp. 228–237. (2013)
- [3] H. Maharani, 'Konstruksi Bayesian Network Dengan Algoritma K2 Pada Kasus Prediksi Cuaca', *Jurnal Telematika*, **2**, 1, pp. 1–4. (2015)
- [4] A. Saleh., & U. P. Utama. *Implementasi Metode Klasifikasi Naive Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga*. *Citec Journal*, **2**, 3, 207–217. (2015)
- [5] F. Tempola., M. Muhammad dan A. Khairan. 2018. Naive Bayes Classifier for Prediction of Volcanic Status in Indonesia. Proc. - 2018 5th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2018, pp. 365–369. Indonesia (2018)
- [6] F. Tempola and A. Mubarak (2013) 'Optimization Naive Bayes using Particle Swarm Optimization in Volcanic Activities', *Journal of Physics: Conference Series* **1569**, 2, pp. 1–9. (2020)
- [6] Bustami 'Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi', *Jurnal Informatika*, **8**, 1, pp. 884–898. (2014)
- [7] S. Kusumadewi. 'Klasifikasi Status Gizi Menggunakan Naive Bayesian Classification', *Jurnal CommIT*, **3**, 1, pp. 6–11. (2009)
- [8] A. F. Cahyanti. (2015). Penentuan Model Terbaik pada Metode Naive Bayes Classifier dalam Menentukan Status Gizi Balita dengan Mempertimbangkan Independensi Parameter. **4**, 1, 28–35. (2015)
- [9] M. F. Rahman., M. I. Darmawidjadja, and D. Alamsah., (2017) 'Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN)', *Jurnal Informatika*, **11**, 1, pp. 36–45. (2017)
- [10] F. Tempola "Case Based Reasoning For Determining The Feasibility Of Scholarship Grantees Using Case Adaptation," *2018 5th Int. Conf. Inf. Technol. Comput. Electr. Eng.*, pp. 372–376, (2018).
- [11] D. Purnomo, 'Model Prototyping Pada Pengembangan Sistem Informasi', *Jurnal Informatika Merdeka Pasuruan*, **2**, 2, pp. 54–61. (2017)