Naive bayes methods for rainfall prediction classification in Banyuwangi

A U Azmi, A F Hadi, D Anggraeni and A Riski
Master of Mathematics Department, Faculty of Mathematics and Natural Science, University of Jember, Jember 68121, Indonesia.

E-mail: aanaulul@gmail.com, afhadi@unej.ac.id, dian_a_fmipa@unej.ac.id, riski.fmipa@unej.ac.id

Abstract. Banyuwangi is the largest district in East Java with an area of 5,782.50 km². It has a long coastline of about 175.8 km which stretches along the southern eastern boundary of Banyuwangi Regency, and there are 10 islands. The BMKG estimates that the dry season in the Banyuwangi area is due to the appearance of the beach having hot weather and rarely rains. Banyuwangi also predicts that the dry season is due to the slight influence of cloud growth. Rainfall is a factor of the rainy season which has a big influence on life such as aviation, plantations and agriculture. Agriculture and plantations in Banyuwangi are mostly located in remote areas. Remote areas are likely to lack weather and climate data information. Climate elements of a region cannot be ignored, especially rainfall. Based on data from BMKG (Meteorology, Climatology and Geophysics), the weather data used needs to be classified. Rainfall classification can be categorized into three, namely, light, normal and heavy. There are quite a lot of classification methods, there are several new methods that are quite good such as Naive Bayes (NB). Naive Bayes Classifier (NBC) is an algorithm in data mining techniques that is used to determine the probability of a member of a group. Large and irrelevant datasets can be solved using the Naive Bayes Classifier (NBC) method. The rainfall data used is known first, observed then identified to form a training dataset. Determining the accuracy of rainfall with the Naive Bayes Classifier (NBC) can use several parameters that have a physical relationship between the atmosphere and rainfall. The parameters used to determine rainfall are humidity, rainfall and precipitation. From this study, from 49 data testing, 47 data were predicted correctly with an accuracy of 96%.

1. Introduction
Banyuwangi is the largest district in East Java with an area of 5782.50 km². It has a long coastline of about 175.8 km which stretches along the southern eastern boundary of Banyuwangi Regency, and there are 10 islands. The Banyuwangi BMKG estimates that the dry season in the Banyuwangi area is due to the appearance of the beach having hot weather and rarely rains. Banyuwangi also predicts dry season due to a slight cloud growth. The emergence of the potential seeds of tropical storms and tropical storms that hit the northwestern region of Australia and low air pressure in northwest Australia has an impact on increasing wind speed, sea waves and also reducing cloud growth, so that sunlight is not blocked by clouds. This is what causes Banyuwangi to be dry, but its air is quite cold, because the heat absorbed by the earth from morning to evening will be released back into the atmosphere at night with more intensity. Therefore Banyuwangi has a wet dry season.

Wet dry means that rain can still occur during the dry season. Rainfall in the wet dry season is classified as light rain and normal rain. Rainfall is sediment or water sediment in liquid or solid form that comes from the atmosphere. Rain characteristics of an area need to be known to determine water availability and the possibility of problems and disasters related to water resources [10]. Rainfall is one of the factors in the rainy season which has a big impact on life, such as aviation, plantations and agriculture. Rainfall can hamper agricultural and plantation products. Farmers need information about rainfall to find alternatives in case of rainfall problems such as drought and excess water. Agriculture
and plantations in Banyuwangi are mostly located in remote areas. Remote areas tend to lack information on weather and climate data. The climatic elements of an area cannot be ignored, especially rainfall. Based on BMKG data (Meteorology, Climatology, and Geophysics), the weather data used need to be classified. Classification is processing to find a model or function that describes and characterizes a concept or class of data, for a specific purpose. As done by Arsyad [1] who analyzed the characteristics and classification of rainfall in the Maros Karst Region and Sasmito [13] which determined the climate in Ponorogo with the Schmidt Ferguson and Oldeman classification. Classification has several methods, including the Naïve Bayes Classifier, Support Vector Machine, Artificial Neural Network, Classification Tree, K-Nearest Neighbor, Discriminant Analysis, and others. Marthin [5] compares the rainfall classification using the Support Vector Machine and Naïve Bayes Classifier methods.

The Naïve Bayes Classifier is suitable for forecasting rainfall because it does not require fulfilling assumptions like other classification methods such as discontinuity analysis. Related research was previously conducted by [3] entitled "Classification of Rainfall in Bali Province Based on the Naïve Bayesian Method". The results of Gunadi's research were the classification in the category of light and heavy rainfall, the system was able to classify well. The possibility that the yield is still low for the normal rain category lies in the data, there should be no significant difference between the three classes. It is necessary to retest with more data. Based on the above points, the authors are interested in conducting research on the classification of rainfall predictions in Banyuwangi using the Naive Bayes method. The data used is the rainfall data in Banyuwangi in the last one year which was obtained from BMKG.

2. Material and Methods

2.1. Study Region

Study Area Research location is in Banyuwangi Regency. This district is located at the easternmost tip of the island of Java, in the Horseshoe area, and is bordered by Situbondo Regency in the north, the Bali Strait in the east, the Indian Ocean in the south and Jember Regency and Bondowoso Regency in the west. Banyuwangi Regency is the largest district in East Java as well as the largest in Java, with an area of 5,782.50 km². Geographically, Banyuwangi Regency is located at the coordinates 7°45’15" – 8°43’2" LS and 113°38’10" East Longitude. Banyuwangi district is quite diverse, from lowlands to mountains. The border area with Bondowoso Regency, there is a series of Ijen Plateau with the peaks of Mount Raung (3,344 m) and Mount Merapi (2,799 m). Banyuwangi Has a long coastline of about 175.8 km that stretches along the southern eastern boundary of Banyuwangi Regency. The temperature in the lowland areas ranges from 20°C–34°C, while the highland areas have temperatures less than 19°C. The humidity level in Banyuwangi Regency varies between 73–84%. Based on the Koppen climate classification, almost the entire area of Banyuwangi Regency is included in the category of wet and dry tropical climate with two seasons, namely the dry season and the rainy season. The dry season in Banyuwangi Regency takes place from May to October with the peak of the dry season being August. Meanwhile, the rainy season in the Banyuwangi region takes place from November to April with the wettest month being January, where the monthly rainfall is more than 230 mm per month. The annual rainfall in the Banyuwangi region ranges from 1,000–1,600 mm per year with the number of rainy days varying from 80–140 rainy days per year.

2.2 Data Description

The data used in this research is secondary data obtained from the BMKG Banyuwangi Meteorological Station, the data taken is daily climate data and can be accessed on the website www.bmkg.go.id. The variables used in the study consisted of the dependent variable and the independent variable. Dependent variable (Y). In this study, the depositional status is divided into two categories, namely light rain, normal rain, and heavy rain. Independent variable (X). Used as many as four variables, namely rainfall, average humidity, and average wind speed. The stages of data analysis using the NBC method are as follows, Collect data. The data collection is then formed a dataset that
Descriptive analysis. This process is carried out on each variable. Because the data used are numerical data, the solution is using the Gaussian density. This process is carried out on each observation by calculating the average value and standard deviation. After the mean and standard deviation are found, the likelihood value is calculated. Next, calculate the posterior probability of each category (class) on the testing data. Then the next stage of normality of the posterior probability value for determining the category (class) that has the highest posterior probability value will become the category (class) of the data testing. After the NBC process is carried out, the results of the confusion matrix table will be obtained. The table is used to find the accuracy value for the NBC method.

2.3. Data analysis method

2.3.1. Naive Bayes Classifier (NBC) Algorithm.

Naive Bayes Classifier which was first proposed by Revered Thomas Bayes. The use of the Naive Bayes Classifier has been introduced since 1702-1761. Naive Bayes Classification or also known as Bayesian Classification is a statistical classification method based on the Bayes theorem which can be used to predict the probability of class membership [4]. Meanwhile, Kononenko and Langley concluded that the Naive Bayes Classifier is a possible data class label or it can be assumed to be a labeled class attribute [6]. Bayesian Classification is proven to have high accuracy and speed when applied to large databases [7]. Naïve Bayes Classifier is an algorithm in data mining techniques that applies the Bayes theory in classification [11]. NBC is reliable in handling large datasets and can handle irrelevant data. The symbol for X is an input vector containing data and Y is the class label. The equation of the Bayes theorem is as follows:

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

(1)

Where :

\( X \) = Data with unknown class

\( Y \) = Data hypothesis X is a specific class

\( P(Y|X) \) = probability of hypothesis H based on condition X (posterior probability)

\( P(Y) \) = Hypothesis probability H (prior probability)

To explain the Naïve Bayes method, it should be noted that the classification process requires a number of clues to determine what class is suitable for the sample being analyzed [12]. Therefore, the Naïve Bayes method above is adjusted as follows:

\[ P(Y|X_1, ..., X_n) = \frac{P(Y)P(X_1, ..., X_n|Y)}{P(X_1, ..., X_n)} \]  

(2)

Where the variable \( Y \) represents the class, while the variables \( X_1, ..., X_n \) represent the characteristics of the instructions needed to perform classification. Then the formula explains that the chance of the entry of a certain characteristic sample in the \( Y \) (Posterior) class is the chance of the \( Y \) class (before the entry of the sample, it is often called the prior), multiplied by the probability of the occurrence of the sample characteristics in the \( Y \) class (also called likelihood), divided by with the probability of the emergence of sample characteristics globally (also called evidence). Therefore, the formula above can also be written simply as follows:

\[ \text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}} \]  

(3)
The Evidence value is always fixed for each class in one sample. The value of the posterior will later be compared with the values of the other posterior classes to determine which class a sample will be classified into. Further elaboration of the Bayes formula is carried out by describing \((Y_j|X_1,...,X_n)\) using the following multiplication rules:

\[
P(Y_j|X_1,...,X_n) = P(Y_j)P(X_1,...,X_n|Y_j) \\
= P(Y_j)P(X_1|Y_j)P(X_2,...,X_n|Y_j,X_1) \\
= P(Y_j)P(X_1|Y_j)P(X_2|Y_j,X_1)P(X_3,...,X_n|Y_j,X_1,X_2) \\
= P(Y_j)P(X_1|Y_j)P(X_2|Y_j,X_1)P(X_3|Y_j,X_1,X_2)P(X_4,...,X_n|Y_j,X_1,X_2,X_3) \\
= P(Y_j)P(X_1|Y_j)P(X_2|Y_j,X_1)P(X_3|Y_j,X_1,X_2)P(X_4|Y_j,X_1,X_2,X_3) ... P\left(X_n|Y_j,X_1,X_2,X_3,...X_{n-1}\right)
\]

(4)

It can be seen that the results of this translation cause the complexity of the factors that affect the probability value, which is almost impossible to analyze one by one. As a result, these calculations become difficult to carry out. This is where the assumption of very high independence (naive) is used, that each point \((X_1,X_2,...,X_n)\) is independent of one another. With this assumption, one thing in common is as follows:

\[
(X_i|Y_j) = \frac{P(X_i|Y_j)}{p(X_j)} = \frac{P(X_i|Y_j)p(X_j)}{p(Y_j)} = P(X_i)
\]

(5)

For \(i \neq j\), so that

\[
P(X_i|Y,Y_j) = P(X_i|Y_j)
\]

(6)

From the above equation it can be concluded that the assumption of naive independence makes the terms of the opportunity simple, so that the calculation becomes possible. Furthermore, the translation of \(P(Y_j|X_1,...,X_n)\) can be simplified to:

\[
P(Y_j|X_1,...,X_n) = P(Y_j)P(X_1|Y_j)P(X_2|Y_j,X_1)P(X_3|Y_j,X_1,X_2) ... P\left(X_n|Y_j,X_1,X_2,X_3,...X_{n-1}\right) \\
= (P(Y_j)\prod_{i=1}^{n} P\left(X_i|Y_j\right))
\]

(7)

So that the classification results are a class that produces a maximum probability value or can be expressed in the following equation:

\[Y_{MAP} = \arg \max_{Y_j \in Y} \left( P(Y_j)\prod_{i=1}^{n} P\left(X_i|Y_j\right)\right)\]

(8)

Information:

\[
P(Y_j|X_1,...,X_n) :\text{ Posterior Probability} \\
P(X|Y) :\text{ Likelihood} \\
P(Y) :\text{ Prior Probability} \\
Y_{MAP} :\text{ Class with Maximum A Posterior Probability}
\]

The equation above is a model of the Naïve Bayes theorem which will then be used in the classification process. For classification with quantitative or continuous data the Gauss Density formula is used:
\begin{equation}
P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi} \sigma_{ij}} \exp \left( -\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2} \right)
\end{equation}

Where:
- \( P \) = Opportunity
- \( X_i \) = Attribute to \( i \)
- \( x_i \) = Attribute value to \( i \)
- \( Y \) = Class sought
- \( y_j \) = Y sub class you are looking for
- \( \mu_{ij} \) = The sample mean of the training data that belongs to \( y_j \)

### 2.3.2. Confusion Matrix.

According to [9] the confusion matrix is a table that records the classification results. Generally, the measurement of classification performance is carried out with a confusion matrix. The confusion matrix is a table that records the results of the classification work. The following is the result of confusion matrix [8].

|       | Prediktion |
|-------|------------|
| Actual|            |
|       | C1         | C2         |
| C1    | TP         | FN         |
| C2    | FP         | TN         |

**Information:**
- \( \text{TN} \) = The number of correct predictions is negative (True Negative)
- \( \text{FN} \) = Number of false predictions is positive (False Negative)
- \( \text{FP} \) = Number of wrong predictions is negative (False Positive)
- \( \text{TP} \) = The number of correct predictions is positive (True Positive)

Actual is a classification of rain status which has been previously classified. Prediction is the result of the classification of the status variable produced by the program / software. From the formation of the configuration matrix, several other values can be calculated that can be used as classification performance values [2]. These values are as follows:

a. **Accuracy** is the proportion of correct classification prediction. The accuracy formula is:

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}
\]

b. **Error Rate** is the proportion of the classification making prediction errors, with the calculation of the following equation:

\[
\text{Error Rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}
\]
3. Results and Discussion

3.1. Dataset
The first stage in the data mining process is to understand the application domain to find out and explore initial knowledge and what the target users are. From the available data, humidity and rainfall are as data with an unknown class or as X, while precipitation is used as data hypothesis X which is a specific class or as Y. Local rainfall data used in this case starts from January 2020 to Last August, there were 245 data. The data obtained are daily data.

| DATE       | HUMIDITY | WIND VELOCITY | RAINFALL | PRECIPITATION |
|------------|----------|---------------|----------|---------------|
| 01-01-2020 | 89       | 0             | 0        | 0             |
| 02-01-2020 | 84       | 1             | 19       | 1             |
| 03-01-2020 | 87       | 2             | 25       | 2             |
| 04-01-2020 | 83       | 1             | 4        | 0             |
| 05-01-2020 | 73       | 2             | 1        | 0             |

3.2. Prediction Classification
At this stage, I predict the data to classify or classify the data into three parts, namely light rain, normal rain, and heavy rain. The prediction of this class is obtained from processing humidity and rainfall.

3.3. Data Training and Data Testing
The training process is conducting the data training process on the model (naïve Bayes). While the testing process is to test data that will produce a graph or pattern. Daily rainfall data, divided into training data and testing data, 80% for training data, namely 196 data and 20% for testing data, namely 49.

![Figure 1. Plot of Training Data Result.](image)
The local rainfall data plot is divided into three types of rain, namely light rain, normal rain and heavy rain. For light rain it is symbolized in red, and for normal rain it is symbolized by green, while for heavy rain it is symbolized in blue. The higher the rainfall and humidity, the heavier the rain will be, and vice versa, if the rainfall is low and the humidity is low, then the rain will be light. The yellow color in the plot indicates that the rainfall is high, and results in heavy rain.

The results of this study can be seen from the confusion matrix. The above confirmation matrix states that 41 days of light rain, 5 days of moderate rain. And 1 day it rained heavily. From this figure, it can be explained that 41, 5, and 1, which are in the diagonal matrix, are correct predictions. While others predict wrong. So that the predicted data is correct as much as 47 days / data from 49 days / data. Meanwhile, 2 days / other data is predicted to be less precise, for 1 data it is predicted that light rain is supposed to be normal rain, while 1 data which is supposed to be heavy rain becomes normal
rain. From the picture above, the data accuracy is wrong or the error rate is 4%. Meanwhile, the accuracy for correct data reaches 96%.

4. Conclusion
From the results of the tests that have been done, it can be concluded that, the test results show that the classification of the rainfall category is light, normal, and heavy with a fairly high level of accuracy, namely 95.91836734693877%. Classification using the Naive Bayes Classifier (NBC) method in the categories of light, normal and heavy rainfall, the system is able to classify well. The possibility of low yields in the normal rain category lies in the data, there should be no significant difference between the three classes. It is necessary to retest with more data.

Acknowledgement
We would like to thank LP2M of University of Jember. This paper was part of the research that funded by the internal grant funds with contract number 3566/UN25.3.1/LT/2020. We also thanks to all member of Data Science Research Group at Department of Mathematics, University of Jember.

References
[1] Arsya M, Sulistiawati and Vistarani A T 2015 Analisis Karakteristik dan Klasifikasi Curah Hujan di Kawasan Karst Maros Proceeding Internasional Seminar On mathematics, Science, and Computer Education 57–62
[2] Faisal M and Nugrahadi D 2019 Belajar Data Science Klasifikasi dengan Bahasa Pemrograman R Banjarbaru Kalimantan Selatan Scripta Cendekia
[3] Gunadi G A and Dew A A k 2018 Klasifikasi Curah Hujan di Provinsi Bali Berdasarkan Metode Naïve Bayesian Wahana Maematika dan Sains : Jurnal Matematika, Sains, dan Pembelajaranannya 12
[4] Hand David J and YU Keming 2001 Idiot’s Bayes: Not So Stupid after All? International Statistical Review 69 (3) 385–398
[5] Laia M L and Setyawan Y 2020 Perbandingan Hasil Klasifikasi Curah Hujan Menggunakan Metode SVM dan NBC Jurnal Statistika Industri dan Komputasi 5 51–61
[6] Langley and Sage S 1994 Induction of Selective Bayesian Classifier Proceeding of The Tenth Conference on Uncertainty in Artificial Intelligence Morgan Kaufmann US 399–406
[7] Lewis D 1998 Naïve Bayes at forty: The independence assumption in information retrieval Proceedings of the Tenth European Conference on Machine Learning 4–15
[8] Novandya A and Oktria I 2017 Penerapan Algoritma Klasifikasi Data Mining C4.5 Pada Dataset Cuaca Wilayah Bekasi Jurnal Format 98–106
[9] Prasetyo 2012 Data Mining Konsep dan Aplikasi Menggunakan Matlab Yogyakarta Andi Offset
[10] Prawirowardoyo S 1996 Meteorologi Bandung : ITB
[11] Ridwan, M, Suyono H and Sarosa M 2013 Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier Jurnal EECCIS 59–64
[12] Saleh A 2015 Implementasi Metode Klasifikasi Naive Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga Jurnal Citec 207–217
[13] Sasminto R A, Alexander T J and Bambang R W 2015 Analisis Spasial Penentuan Iklim Menurut Klasifikasi Schmidt-Ferguson dan Oldeman di Kabupaten Ponorogo Jurnal Sumberdaya Alam dan Lingkungan Malang: Universitas Brawijaya