

RAINFALL PREDICTION USING DATA MINING TECHNIQUES

^[1]Riko Herwanto, ^[2]Rosyana Fitria Purnomo, ^[3]Sriyanto

Department of Informatics Engineering

Faculty of Computer Science

Informatics and Business Institute Darmajaya

riko.1721211022@mail.darmajaya.ac.id rosyana.1721211024@mail.darmajaya.ac.id sriyanto@darmajaya.ac.id

ABSTRACT

Rainfall is an important factor in agrarian countries such as Indonesia. Rainfall prediction has become one of the most challenging technological challenges and challenges in the world. And also the most significant and difficult task for researchers in recent years. In Data Mining, the classification algorithm is primarily used to predict rainfall, temperature, various methods of available rainfall estimation that will be used to determine the cultivation time for a particular crop, a particular crop varieties. The reliability of this prediction depends on accuracy in choosing correlated variables. If existing historical databases fail to record the most correlated variables, then the reliability of these data-driven forecast approaches is questionable. In this paper, an attempt has been made to develop a methodological framework that leverages the power of a predefined data mining analysis (decision tree). The decision-based rainfall prediction model developed maps climate variables, namely; a) temperature, b) humidity, and c) wind speed over the observed rainfall database.

This paper uses data mining techniques such as Clustering Technique, Decision Tree and classification for rainfall prediction.

Keywords: *Rainfall, Rainy Season, Data Mining, Classification, Decision Tree, Bayesian Technique.*

1. Introduction

Weather is an air condition somewhere at a relatively short time, expressed by parameter values such as wind speed, temperature, pressure, rainfall, and other atmospheric phenomena as the main component [1]. Weather is an important thing that will never be separated from human life. The course of human activity can be affected by weather conditions, such as agriculture, plantation fields. For example a good time decision to plant crops. The magnitude of climate-induced influences encourages the development of weather-determining systems that determine weather conditions. The supported approach using available technology is the application of the current weather determination system [2]

The determination of weather is the process of applying science and technology to determine the state of the earth's atmosphere. The weather determination process uses a method to produce weather determination output [3]. The use of good methods, has recently become an activity that is often done by weather researchers or the atmosphere. To get better and more accurate results, researchers are encouraged to continue researching weather-related atmospheres and develop methods to meet the need for information on weather or atmospheric conditions [1].

Naive Bayes Classifier (NBC) is a supervised, reliable, easy, effective, efficient, and efficient method of dealing with data noise such as attributes that are lacking or irrelevant. Large datasets either that attribute discrete or continuous variables can also be handled by Naive bayes clasifier [4] [5]. The use of this algorithm is expected to facilitate the process of determining the weather with good accuracy.

There are various functions used in Data Mining in processing data in general, such as Description, Clustering, Classification, Association Rules, Estimation. And each Function has its own algorithm. But in this writing the author uses the Decision Tree algorithm entered in the Function Classification.

The result of the paper is to find out whether to compare the Naive Bayes algorithm method, Decision Tree and kNN Classification can be used to help predict rain.

2. Research Methods

This research uses several methods that aim to solve the problems that exist in this research. The method that is proposed in this research is as follows:

2.1 Preliminary Data Processing (Preparation Data)

In this research, there are several stages to be done on the data already obtained in the data, all data or records and attributes are not used all, data or records and attributes must go through several stages of initial processing of data (preparation data). So get quality data, some techniques are done as follows[6] :

a. Data Cleaning

Good quality data are the basic keys to produce quality data, data nois data that still outliers or errors, data incomplete data missing attribute values, and inconsistent data inconsistent data in the charging attributes.

Stages in data cleaning:

1. Eliminate or identify outliers.
2. Completing values that are not complete or missing (missing value), naive bayes algorithm has its own keunggullan that can handle the data is incomplete or missing (missing value).
3. Inconsistent data is fixed.
Break down redundancy caused by data interrogation.

b. Data intergration and transformasion

The next step is integration techniques used to analyze correlation data, redundant attributes and duplicate data, and transformation is used to improve the accuracy and efficiency of the algorithm. Excess naive bayes algorithm is able to process data that is nominal, continuous, and ordinal. Therefore the value of each attribute contained in the dataset does not have to be transformed.

c. Data reduction

Data rection is the process of reducing the dataset by reducing the number of attributes or records that are not needed for less but still be informative. Obtain representations in reduced data volumes but still obtain similar or similar analytical results and data descriptions that are part of data reduction, an important part of numerical data.

2.2. Naive Bayes Method of Classification and Algorithm

Classification can be defined in detail as an activity of learning or training on the function f maps the vector x into one of several available class y labels. The activity will give the result of a model which is then stored as model [7]. Models created in later learning can be used to determine the label of the class. During the process of learning in making the model, required a learning algorithm, among others, namely: KNN, Naive Bayes, Decission Tree.

Naive Bayes Classifier (NBC) also referred to as Bayesian Classification is a method of classifying statistics useful for the process of determining the probability of a membership of a class. Bayes's theorem

underlies the Naive Bayes Classifier that has similar classification capabilities to the Decision Tree and Neural Network. NBC also efficiently, effectively, and reliably handles data noise such as irrelevant attributes. NBC can also overcome large datasets with both variable and continuous attributes [4] [5].

2.3. Decision Tree

Decision Tree is a very powerful method of classification and prediction famous. The decision tree method turns a very large fact into a decision tree that represents the rule. Rules can be easily understood with natural language. In addition, rules can also be expressed in the form of database languages such as Structured Query Language (SQL) to search for records in certain categories. The decision tree is also useful for exploring data, finding the hidden relationship between a number of potential input variables with a target variable. Because the decision tree combines data exploration and Decision Tree modeling is used for cases where the output is discrete.

A decision tree is a structure that can be used to divide large datasets into smaller record sets by applying a set of decision rules. With each set of divisions, members of the result set become similar to others [Berry & Linoff,2004].

3. Results And Discussion

3.1. Data Processing

In conducting this experiment required weather data as much as 2192 data ie weather data from January 2011 to December of 2016 used as a dataset. The dataset is obtained from the location S 05° 23 '45.0852 "E 105° 22' 43.1616" Before the classification process is done, data or records and attributes must go through several stages of initial processing of data (preparation data).

In incomplete or missing data leads to unfairness or even error, therefore to obtain quality data, several stages of initial data processing (preparation data) are done as follows:

a. Data Cleaning

Incomplete data deletion (incomplete data), noise data, and missing attribute values are as big as records.

b. Data Reduction

Removes noise data and breaks down data redundancies by deleting less important attributes and values. Removing the Date (Date) attribute will not provide the required information in mining.

The following is the result of the initial data processing (preparation data) ie Data Cleaning and Data Reduction. 2192 records obtained from 11 to 9 attributes attribute i.e. 5 numeric attributes and

kategorial attributes 3 and 1 labels the target. Numeric attribute is the attribute that the domain be integer or real value, i.e. the Model Year, Min-Max Temperature-Temperature, Rain, Light Intensity. Kategorial attribute is an attribute that its domain is a set, the value of himpunannya is the Month, Weather Morning, Noon, and Dry Weather and Rainy which is the label target.

The next step that is collecting a whole dataset data into weather testing. Based on the overall existing weather dataset i.e. 2192 data, comparison with the composition of 100% for testing data, composition of the comparison based on research [12]. So data testing owned as much as 2192 records

3.2. Implementation Result

3.2.1. Naive Bayes Methods

Naive bayes method, obtained as a result of the matching of data between testing results of the proposed method with the use of the weather data 2192 tested and actual weather conditions i.e. with 91.56% accuracy level.

The accuracy of the stated that the value of the ratio of the amount of data the weather that classified in class correctly (true positive) and the amount of data that is classified in the class that otherwise (true negative) with all the weather data is classified.

The result of the test data matching is then inserted into the table of confusion matrix. Based on table 1 the value of accuracy can be calculated using the following calculation::

Table 1. Confusion Matriks Naïve Bayes

	True Dry	True Rainy	Class Precision
Pred.Dry	1856	0	100%
Pred. Rainy	185	151	44.94%
Clas Recall	90.94%	100%	

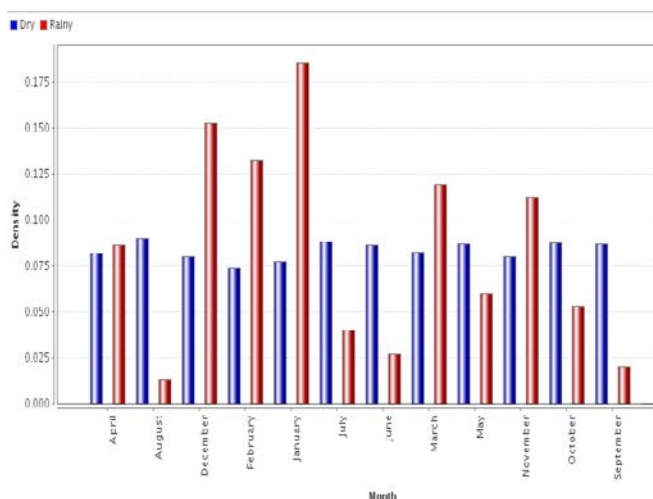


Figure 1. Graph View From Naive Bayes Methods

Based on the graph above by using the method of naïve bayes, we can know the density of the precipitation Dry and Rainy months based on data obtained from the testing that has been tested.

3.2.2 K-NN Methods

K-NN method, obtained as a result of the matching of data between testing results of the proposed method with the use of the weather data 2192 tested and actual weather conditions i.e. with 97.77% accuracy level. The accuracy of the stated that the value of the ratio of the amount of data the weather that classified in class correctly (true positive) and the amount of data that is classified in the class that otherwise (true negative) with all the weather data is classified.

The result of the test data matching is then inserted into the table of confusion matrix. Based on Table 2 and the values of accuracy can be calculated using the following calculation:

Table 2. Confusion Matriks k-NN

	True Dry	True Rainy	Class Precision
Pred.Dry	2016	24	98.82%
Pred. Rainy	25	127	83.55%
Clas Recall	98.78%	84.11%	

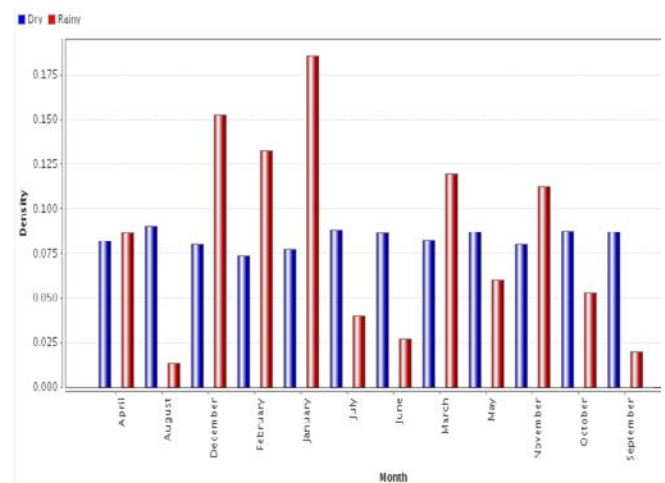


Figure 2. Graph View k-NN Methods

Based on the above graph using the kNN method, we can know the density of Dry and Rainy rainfall by month based on the tested data testing.

3.2.3. Decision Tree Method

Method of Decision Tree, got result of match of data testing between result of proposed method by using 2192 tested weather data and actual weather condition

The results of matching the test data are then fed into the confusion matrix table. Based on Table 3 then the accuracy value can be calculated using the following calculation:

Table 3. *ConfusionMatriks Decision Tree*

	True Dry	True Rainy	Class Precision
Pred.Dry	2020	20	99.02%
Pred. Rainy	21	131	86.18%
Clas Recall	98.97%	84.11%	

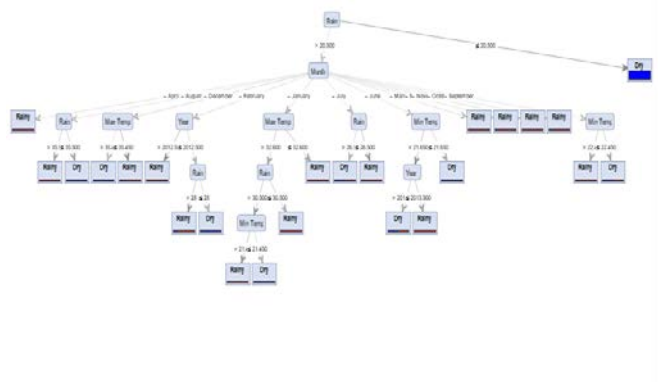


Figure. 3. View Decision Tree Method

Based on the figure above, by using Ddecision Tree method, we can know the density of Rain and Rainy rainfall with based on selected attribute with label that has been determined that is Dry and Rainy obtained from tested data testing.

4. Conclusion

Based on the result of daily weather research experiments using naive bayes method has been done, so it can be concluded some things as follows:

From the test results show that the classification of daily weather determination by using the method naive bayes get an accuracy value of 91.56% .k-NN Method, accuracy value 97.77% and Decision Tree with accuracy 98.13%

From the results of research and testing has been done, then the method proposed in this research is the daily weather classification (Rain and Not Rain) using Decision Tree method can be used to determine the daily weather (Rain and Not Rain) for the next years.

In this paper, Naive Bayes, k-NN and decision tree building process were implemented; both are the most common data mining techniques tried to highlight the method that the stored data about past measures can be used for the future ones. Here, j48 (decision tree algorithm) was tried to create decision-trees & rules for the classification of parameters of weather such as minimum temperature, maximum temperature, precipitation and wind-speed per months and years. Experimental trends about sufficient data over-time was analyzed and the significant deviations were identified that showing the change in climate patterns. Future work can include expanded database with other important weather parameters and include using this weather information in agriculture sector reform with cutting edge technologies.

Bibliographies

- [1] H. Maharani, "Konstruksi Bayesian Network Dengan Algoritma K2 Pada Kasus Prediksi Cuaca", 2015.
- [2] R. Chaniago, T. H. Liong dan K. R. R. Wardani, "Prediksi Cuaca Menggunakan Metode Case Based Reasoning dan Adaptive Neuro Fuzzy Inference System", Jurnal Informatika, vol. 12, 2014.
- [3] [Y. Yang and G. I. Webb, "On Why Discretization Works for Naive-Bayes Classifier," 2003.
- [4] D. c. Mahendra dan A. W. Kurniawan, "Klasifikasi Data Debitur Untuk Menentukan Kelayakan Kredit Dengan Menggunakan Metode Naive Bayes," 2015.
- [5] H. Witten, E. Frank dan M. A. Hall, "Data Mining Practical Machine Learning Tools and Techniques, Burlington: Morgan Kaufmann", 2011.
- [6] J. Han, M. Kamber dan J. Pei, Data Mining Concepts and Techniques, Waltham: Morgan Kaufmann, 2012.
- [7] S. Mujiasih , "Pemanfatan Data Mining Untuk Prakiraan Cuaca," 2011.
- [8] M. G. Salman , B. Kanigoro dan Y. Heryadi , "Weather Forecasting using Deep Learning Techniques," 2015.
- [9] Daniel T Larose. "Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Inc." 2005
- [10] Xindong Wu, "The Top Ten Algorithms in Data Mining. CRC Press, Inc." .2009
- [11] Kusriani, Emha Taufiq Luthfi, "Algoritma Data Mining. Yogyakarta. Andi", 2009.

- [12] Ian H. Witten, Eibe Frank, Mark A. Hall, “Data Mining Practical Machine Learning Tools and Techniques” Morgan Kaufmann, 2011.
- [13] Han, Jiawei and Micheline Kamber,. “Data Mining Concepts and Techniques”, Morgan Kaufmann, California, 2001.
- [14] Remco R. Bouckaert, Eibe Frank, Mark A. Hall, WEKA “Manual for Version 3-6-2s”. University of Waikato, Hamilton, New Zealand, 2010
- [15] Ronald Mak, “The Java Programmer's Guide to Numerical Computing”. Prentice Hall PTR, 2002.
- [16] Mark Matthews, Jim Cole, Joseph D. Gradecki, “MySQL and Java Developer's Guide”. Wiley Publishing, Inc, 2010.
- [17] Perdita Stevens, Rob Pooley, “Using UML Software engineering with objects and components”. Addison-Wesley, Inc, 2006.
- [18] Adeyemo, A. “Soft Computing Techniques for Weather and Climate Change Studies”. African Journal of Computing & ICT (Volume 6), Pages 14, 2013.
- [19] Agboola A.H., Gabriel A. J., & Aliyu E.O., Alese B.K. “Development of a Fuzzy Logic Based Rainfall Prediction Model”. International Journal of Engineering and Technology, Volume 3, Pages 9, 2013.
- [20] Badhiye S. S., Wakode B. V., & Chatur P. N. “Analysis of Temperature and Humidity Data for Future value prediction”. International Journal of Computer Science and Information Technologies, Volume 3, Pages 3, 2012.
- [21] Juraj Bartoka, Ondrej Habalab, Peter Bednarc, Martin Gazaka, & Ladislav Hluchýb. “Data Mining and Integration for Predicting Significant Meteorological Phenomena”. Elsevier, Pages 10, 2012.
- [22] Jyotisma Goswami, & Alok Choudhury. “Dynamic Modeling Technique for Weather Prediction”. International Journal of Computer Science & Engineering Technology, Volume 5, Pages 8, 2014.
- [23] K. Somvanshi, & Et Al. Modeling And Prediction Of Rainfall Using Artificial Neural Network And Arima Techniques”. J. Ind. Geophys. Union, Vol. 10 (No. 2), Pp. 141-151, 2006.
- [24] M. Viswambari, & Dr. R. Anbu Selvi. “Data Mining Techniques to Predict Weather: A Survey”. International Journal of Innovative Science, Engineering & Technology, Volume 1 (Issue 4), Pages 3, 2014.
- [25] Meghali A, Kalyankar, & Prof. S. J. Alaspurkar. “Data Mining Technique to Analyse the Metrological Data”. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3 (Issue 2), Pages 5, 2013.
- [26] Pabreja, & Kavita. “Clustering technique to interpret Numerical Weather Prediction output products for forecast of Cloudburst”. International Journal of Computer Science and Information Technologies, Volume 3, Pages 4, 2012.
- [27] R. Nagalakshmi, M. Usha, & RM. A. N. Ramanathan. “Application of Data Mining Techniques in Maximum Temperature Forecasting: A Comprehensive Literature Review”. International Journal of Advance Research in Computer Science and Management Studies, Pages 9, 2013.
- [28] S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, & K. Menagias. “A Hybrid Data Mining Technique for Estimating Mean Daily Temperature Values”, 2006.
- [29] S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, & K. Menagias. “Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values”. International Journal of Mathematical, Physical and Engineering Sciences, Volume 1, Pages 5, 2007.
- [30] S. S. Mesakar, & M. S. Chaudhari. “Review Paper On Data Clustering Of Categorical Data”. International Journal of Engineering Research & Technology, Vol. 1, pp. 1-3, 2012.
- [31] Sarah N. Kohail, & Alaa M. El-Halees. “Implementation of Data Mining Techniques for Meteorological Data Analysis”. International Journal of Information and Communication Technology Research, Volume 1, Pages 5, 2011.
- [32] Valmik B Nikam¹, & B. Meshram². “Modeling Rainfall Prediction Using Data Mining Method”. Fifth International Conference on Computational Intelligence, Modelling and Simulation, 2013.
- [33] Zahoor Jan¹, M. Abrar², Shariq Bashir³, & Anwar M. Mirza⁴. “Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique”. Springer, Pages 13, 2008.
- [34] Folorunsho Olaiya, & Adesesan Barnabas Adeyemo. “Application of Data Mining Techniques in Weather Prediction and Climate Change Studies”. Information Engineering and Electronic Business, Pages 9, 2012.
- [35] Quinlan, J. R. “C4.5: Programs for Machine Learning”. Morgan Kaufmann Publishers, 1993.
- [36] M. Kuhn, & K. Johnson. “Applied Predictive Modeling”. Springer, 2013.

- [37] Azevedo, J.M., Almeida, R. & Almeida, P., "Using Data Mining with Time Series Data in Short-Term Stocks Prediction: A Literature Review, *International Journal of Intelligence Science*", 2, pp. 76-180, 2012.

- [38] Ramsundram N, Sathya S, Karthikeyan S "Comparison of Decision Tree Based Rainfall Prediction Model with Data Driven Model Considering Climatic Variables" , Jan 2016

- [39] Cios KJ, Pedrycz W, Swiniarski RW, Kurgan IA "Data-Mining: A knowledge discovery approach. Springer. New york", 2010.

- [40] Rajesh Kumar, "Decision tree for the weather forecasting", *International Journal of Computer Applications (0975-8887)*, vol. 2, August 2013.