Vol. 10 No. 1 March 2022
PIKSEL
Penelitian Ilmu Komputer
Sistem Embedded and Logic

Penelitian Ilmu Komputer, Sistem Embedded and Logic
p-ISSN: 2303-3304, e-ISSN: 2620-3553
Vol. 10 (1): 67 – 78 (March 2022)
https://doi.org/10.33558/piksel.v10i1.4418

Decision Tree-Based Weather Prediction

Achmad Noeman [1,*], Dwipa Handayani [1], Abrar Hiswara [1]

* Corespondence Author: e-mail: achmad.noeman@dsn.ubharajaya.ac.id

[1] Informatics; Universitas Bhayangkara Jakarta Raya; Jl. Perjuangan No.81, RT.003/RW.002, Marga Mulya, Kec. Bekasi Utara; e-mail: achmad.noeman@dsn.ubharajaya.ac.id, dwipa.handayani@dsn.ubharajaya.ac.id, abrar.hiswara@dsn.ubharajaya.ac.id.

Submitted : **24/01/2022**
Revised : **08/02/2022**
Accepted : **07/03/2022**
Published : **26/03/2022**

Abstract

Weather is formed from a group of weather elements and can be changed fast. For example, in the morning, afternoon, or evening, the weather can be different for each place and every hour. Weather is the condition of the air that occurs in a narrow place and lasts for a short time. Weather conditions in a place can be determined by many factors, such as air pressure, humidity, wind, sunlight, and so on. Therefore, by looking at these factors it can be estimated the weather that will occur the next day. Fishermen and farmers are fields of work that are closely related to weather forecasting, accurate and fast weather predictions are needed by these fields to carry out various activities. The amount of rainfall that occurs cannot be determined exactly but can be predicted or estimated. The application of determining weather information is needed, hence, the prediction can be utilized optimally by the community. The design of a system that will classify automatically can be developed by applying machine learning methods, one of them is used in this study, i.e., C.4.5 Algorithm using weather data that a reference in determining the weather conditions whether not rainy, rainy, light rain, or heavy rain.

Keywords: *C4.5 algorithm, data mining, machine learning, rain prediction, weather datasets*

1. Introduction

Weather is the state of the air at a certain time and in a certain narrow area and in a short time. Of course, the weather is dynamic because it can change daily. One way to determine weather changes is checking the temperature. Some of the factors that influence weather changes include how long the sun shines on the earth. The intensity of solar radiation in the hemisphere varies greatly depending on the latitude and altitude of a place. If a place is flatter, the heat received is certainly greater

When there are many clouds in the atmosphere, the heat received by the earth will be smaller because it is absorbed by the clouds. Another factor that determines the temperature on earth is the condition of the plants on the surface.

Another way to determine weather changes is the angle of sunshine. This angle is formed by the sun's rays on the earth's surface.

The amount of rainfall that occurs cannot be determined with certainty, but can be predicted. By using historical data on the amount of rainfall in the past, it can be predicted how much rainfall will occur in the future. Weather is the condition of the air that occurs in a narrow place and lasts for a short time. It can be determined by many factors, such as air pressure, humidity, wind, sunlight, and so on. Therefore, by looking at these factors, the weather can be estimated for the next days. Fishermen and farmers are fields of work that are closely related to weather forecasting, accurate and fast weather predictions are needed by these fields to carry out various activities. In this study, the C4.5 classification algorithm is used for weather predictions.

2. Research Method

The official information about weather in Indonesia is from the Meteorology, Climatology, and Geophysics Bureau (abbreviated BMKG), an Indonesian Non-Departmental Government Institution which has the task of carrying out government duties in the fields of meteorology, climatology, and geophysics.

Temperature is one of the variables of climate change. Based on the Big Indonesian Dictionary, temperature is defined as a quantitative measure of temperature, hot and cold, measured with a thermometer (Dynes Rizky Navianti, I Gusti Ngurah Ray, Farida Agustini W, 2012). Temperature is the state of hot or cold air. The highest air temperature on earth is in the tropics and the colder it gets to the poles. When viewed from the plains, the lowest plains tend to have high temperatures and the higher the plains the temperature tends to decrease (Setio et al., 2020)(Desmonda et al., 2018). The instrument for measuring temperature is a thermometer. There are two types of thermometers, namely the maximum thermometer and the minimum thermometer. Also read: Differences in Seasons, Climate, and Weather Temperature measurements are usually expressed on the Celsius (C), Reamur (R) and Fahrenheit (F) scales (Rofiq et al., 2020). Microscopically, temperature shows the energy possessed by an object. Each atom in an object each moves, either in the form of displacement or movement in place in the form of vibration. Air temperature plays a very important role in the evaporation of water as well as the ability to hold water in the air and chemical processes in the

air. The higher the air temperature, the higher the rate of evaporation of water, the higher the water vapor retained in the air and the faster the chemical reaction. The lower the air temperature, the ability to hold water vapor also decreases. This causes the air to become saturated with water vapor. When the air reaches its maximum water vapor limit, condensation will start, and the rain begins to fall.

The C4.5 algorithm is one of the algorithms used to classify or group datasets. The basis of the C4.5 algorithm is the formation of a decision tree. The branches of the decision tree are a classification question while the leaves are the classes or groups. Because the purpose of the C4.5 algorithm is to classify, the result of processing the dataset is in the form of grouping the data into certain classes.

The first thing to do to form a decision tree is to determine which attribute/variable is the root of the decision tree. The way to determine the root variable is to use entropy, gain, split info, and gain ratio.

Entropy

Entropy is a parameter to measure the level of diversity (heterogeneity) of the data set. If the value of entropy is getting bigger, then the level of diversity of a data set is getting bigger. The formula for calculating entropy is as follows

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

Where:

m = number of classification classes

pi = the number of sample proportions (probability) for class i

While the formula for the entropy of each variable is

$$Entropy_A(S) = \sum_{v} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

Where :

A = variable

V = possible values for the variable A

|S| = number of samples for the value of v

Entropy(S_v) = Entropy for samples that have a value of *v*

In this calculation, the formula used to find the value of impurity uses the entropy formula. The gain value consists of the attributes of MINIMUM TEMPERATURE, MAX TEMPERATURE, AVERAGE TEMPERATURE, HUMIDITY , and LONG EXPOSURE (LP). The following is an example of finding the root node in calculating the Minimum Temperature attribute. In the dataset there are a total of 1790 records in the form of 1110 No Rain (TH), 225 Mild (R), 140 Medium (S), 139 Heavy (L), and 176 Very Heavy (SL). Therefore, the process of calculating Entropy(S) is:

Entropy(S) = - (1110/1790)$\log_2$(1110/1790)

- (225/1790)$\log_2$(225/1790) - (140/1790)$\log_2$(140/1790)

- (139/1790)$\log_2$(139/1790) - (176/1790)$\log_2$(176/1790) =1.70645

After the entropy has been obtained, then look for the highest gain value for each attribute. In this example it is calculated only on the Minimum Temperature attribute. Here are the steps for calculating the gain for the minimum temperature.

a. Data Sort and output transition search 20.1 203 20.4 20.5 TH | S

b. The gain calculation for each transition point found is at the point: < 20.45 [354,10,6,5,4] and > 20.45 [756,215,134,134,172]

e[354,10,6,5,4] = - (354/379) is the gain value at point 20.5 Gain value this still has to be compared with the values of all points on the MINIMUM TEMPERATURE attribute. The largest gain value from the gain options at the existing split points which will represent the MINIMUM TEMPERATURE attribute gain value. This calculation is performed on all existing attributes. And the largest gain value of all attributes will be the root node (Node at level 1). To search for nodes at a level below the root node, use the same method. With a note that the value of impurity uses the value of the parent node or the node at the level above it.

Gain

Gain is a measure of the effectiveness of a variable in classifying data. The gain of a variable is the difference between the total entropy value and the entropy of the variable. Gain can be calculated using equation 3:

$$Gain(A) = Entropy(S) - Entropy_A(S) \quad (3)$$

In C4.5 algorithm, the gain value is used to determine which variable is the node of a decision tree. A variable that has the highest gain will be used as a node in the decision tree.

Split Info

Split info is used as the divisor of Gain(A) which will produce Gain Ratio for example:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} . \log_2 \left(\frac{|D_j|}{|D|} \right) \quad (4)$$

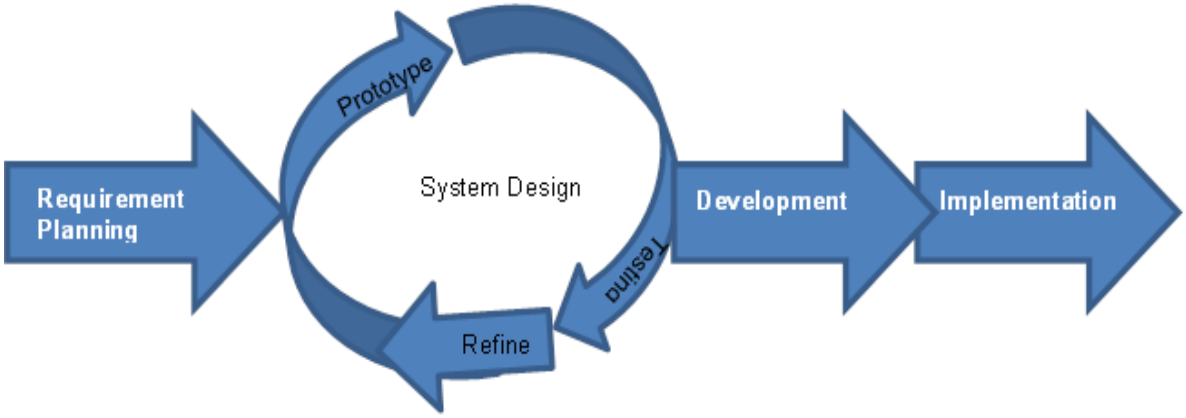
Gain Ratio

Gain Ratio is another measure used to solve problems on attributes that have very varied values. The highest Gain Ratio is chosen as the test attribute for the node using the following relation:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \quad (5)$$

Rapid Application Development (RAD)

Rapid Application Development a software development process that emphasizes a short development cycle. Another definition states that the RAD software development method is a method that uses an object-oriented approach to system development which includes device and software development (Pandey et al., 2013). The stages of RAD are shown in the figure 1.



Source: (Pandey et al., 2013)

Figure 1. Stages of RAD Source from the author

RAD is an object-oriented approach to the development of a system that aims to shorten the time normally required in the design and implementation of information systems.

System Testing

System testing is a complete and integrated testing of software programs. Software or what is often known as software is just a unit element of a larger computer-based system. Usually, software is associated with other software and hardware. Manual testing steps are: 1). Analyze Requirements, 2). Make a test plan, 3). Create test cases, 4). Execution of test cases, 5). Fix from system.

Software testing can be divided into two, namely Black Box Testing and White Box Testing.

a. Black Box Testing

   Black Box Testing is a software testing method used to test software without knowing the internal structure of the code or program. In this test, the tester is aware of what the program must do but has no knowledge of how to do it.

   Blackbox testing focuses on those program units that meet the requirements stated in the specification. Blackbox testing run or execute in a unit or module, then observe whether the results are in accordance with the desired business process.

b. White Box Testing

   White Box Testing is a software testing method in which the internal structure is known to test who will test the software. This test requires internal knowledge of system and programming capabilities.

Data Mining

Data mining is the process of collecting important information on a large amount of data. The data collection can be done through a statistical calculation process, mathematics, or the use of Artificial Intelligence (AI) technology.

Another term for data mining itself can mean data mining in the form of a tool to perform analysis with information filtering techniques more accurately (Rady & Anwar, 2019)(Wang et al., 2020)(Asroni et al., 2018)(Hendrian, 2018). This technique is usually done to find certain patterns that still have relevance to the goals or instructions of the user (user).

Functions of the Application of Data Mining: (1) Association is the process of identifying the relationship (relationship) of each event or events that have occurred

at a certain time; (2) Classification serves to conclude several definitions of characteristics in a group or groups, e.g., determining weather predictions such as sunny, light rain, heavy rain; (3) Clustering is the process of identifying weather predictions that have special characteristics; (4) Descriptive is a function for the purpose of understanding more deeply about the data, so that you can observe any changes in behavior in the information; (5) Forecasting is a technique that is carried out to obtain an overview of the value of a data in the future according to the collection of information with a large amount of information. For example, of weather prediction; (6) Predictive is a function that is used to describe a process in determining a certain pattern in a data. This pattern is used by various variables in the data; (7) Sequencing is the process of identifying each different relationship at a certain period.

Data mining aims to collect information or data with a large size in predicting the weather (Sularno & Anggraini, 2017).

Weather research data from BMKG in 2020

Research stage to be carried out is to identify the existence of weather predictions using the C4.5 Algorithm on the weather dataset taken from the BMKG. This study aims to predict the weather from collecting data to analyzing it, designing, and making menus and sub menus that will be displayed on the information system. The weather data set is data that will be processed in predicting the weather as shown in table 1.

Table 1. Weather Data

Date	Minimum Of Temperature	Maximum of Temeperature	Average of Temperature	Humidity	Rainfall	Exposure Time
10/1/2020	26	33	29	71	0	8
10/2/2020	26	32	29	73	0	6
10/3/2020	26	32	29	72	0	5
10/4/2020	26	33	29	75	0	7
10/5/2020	26	34	30	67	0	7
10/6/2020	26	34	29	71	0	8
10/7/2020	26	35	30	63	0	8
10/8/2020	26	35	30	60	0	7
10/9/2020	25	35	30	58	0	8
10/10/2020	26	33	30	67	0	8

Date	Minimum Of Temperature	Maximum of Temeperature	Average of Temperature	Humidity	Rainfall	Exposure Time
10/11/2020	25	33	29	67	0	8
10/12/2020	25	33	29	68	0	8
10/13/2020	25	34	30	70	0	8
10/14/2020	25	32	29	71	0	8

Source: Research Result

3. Result and Analysis

Data from BMKG can be processed using equation 1-5. Root node value should be calculated before implementation in an web-base application.

Calculation of Root Node Value

The following is the node calculation of Root Node Value as shown in table 2.

Table 2. Calculation of Root Node

Date	Minimum Temperature	Maximum Temperature	Average Temperature	Humidity	Rainfall	Long Exposure Time	Rain
10/5/2020	27	33	30	68	0	8	Yes
10/10/2020	26	34	29	68	0	7	No
10/11/2020	27	34	30	68	0	7	No

Source: Research Result

Root Node: 0.251629167

The following is the node calculation as shown in table 3.

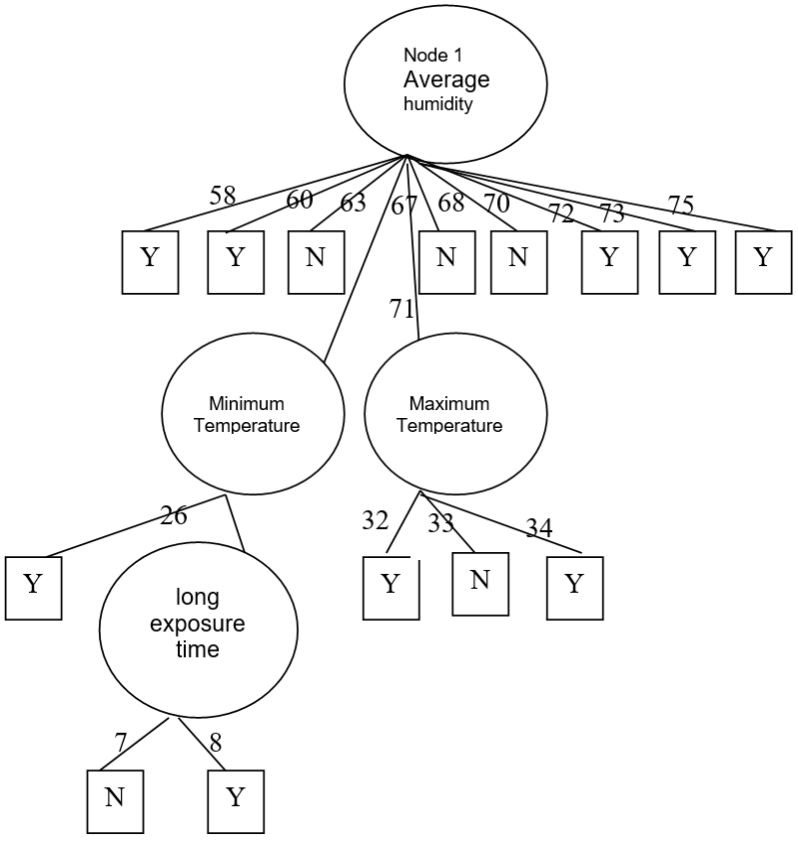
Table 3. Node Calculation

		number of cases	No	Yes	Entropy	Information Gain
TOTAL		3	2	1	0.918295834	
Minimum Temperature						0.251629167
	26	1	1	0	0	
	27	2	1	1	1	
Maximum Temperature						0.251629167
	33	2	1	1	1	
	34	1	1	0	0	
Average Temperature						0.251629167
	29	1	1	0	0	

		number of cases	No	Yes	Entropy	Information Gain
	30	2	1	1	1	
humidity						0
	0	3	2	1	0.918295834	
rainfall						0.251629167
	7	1	1	0	0	
	8	2	1	1	1	

Source: Research Result

From the data above, it can be used as a decision tree as shown in Figure 2.



Source: Research Result

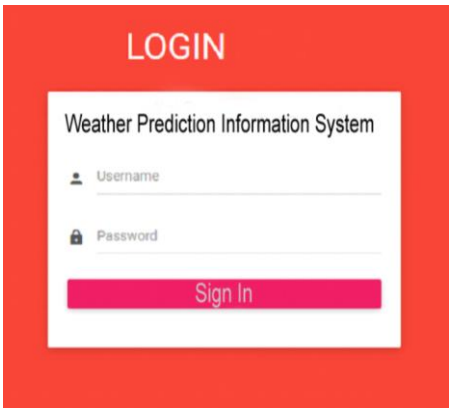
Figure 2. Decision Tree Weather Prediction

Implementation

At this stage identify the design objectives and the information that can be conveyed through the application to be designed. The purpose of this design is to make it easier for users to get information. After knowing the purpose of this design, the authors identify the needs that must be prepared for the design of this system and identify what information will be displayed on this system. User displaying data

that is processed by the admin with the help of a system that is presented to the user in information form. Admin processing the data that has been obtained by the admin by uploading data, downloading data and search data.

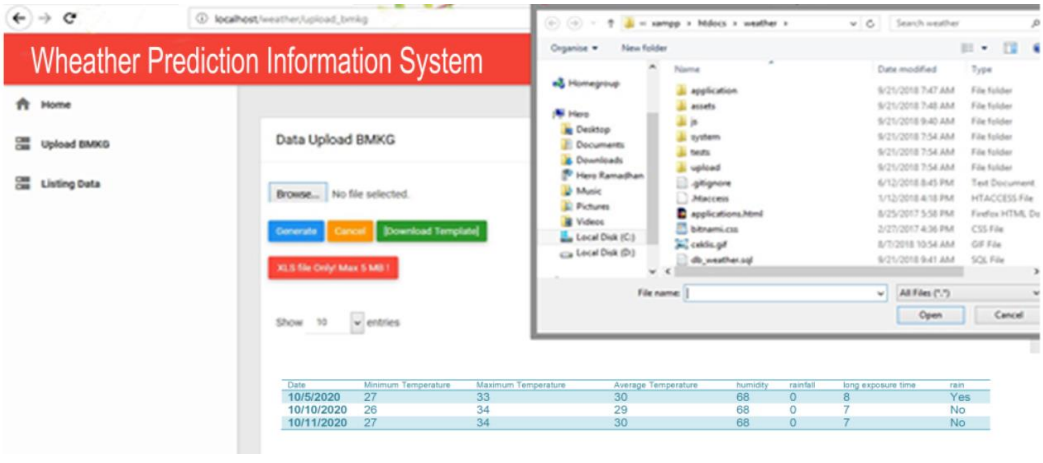
Interface Design



Source: Research Result

Figure 3. Login page to Weather Prediction Information System

Upload implementation page contains uploaded BMKG data menu before running weather prediction information system (Figure 4). The rain prediction will be shown including other parameters, e.g., date, minimum temperature, maximum temperature, average temperature, humidity, rainfall, and long exposure time.



Source: Research Result

Figure 4. Proposed Web-based Application

4. Conclusion

Based on the results of the research on the application of the C4.5 Classification Algorithm for Weather Prediction Based on the Dataset, it can be

concluded as follows: (1) The Weather Prediction Information System can visualize the processed data presented in a form that is more easily understood by public; (2) Application of the C45 Algorithm can be implemented in an application website for processing weather datasets.

Author Contributions

Achmad Noe'man, Dwipa Handayani, Abrar Hiswara conceived models and designed the experiments; Achmad Noe'man, Dwipa Handayani, Abrar Hiswara conceived the optimisation algorithms; Achmad Noe'man, Dwipa Handayani, Abrar Hiswara analysed the result.

Conflicts of Interest

The author declare no conflict of interest.

References

Asroni, A., Masajeng Respati, B., & Riyadi, S. (2018). Penerapan Algoritma C4.5 untuk Klasifikasi Jenis Pekerjaan Alumni di Universitas Muhammadiyah Yogyakarta. *Semesta Teknika*, *21*(2), 158–165. https://doi.org/10.18196/st.212222

Desmonda, D., Tursina, T., & Irwansyah, M. A. (2018). Prediksi Besaran Curah Hujan Menggunakan Metode Fuzzy Time Series. *Jurnal Sistem Dan Teknologi Informasi (JUSTIN)*, *6*(4), 141. https://doi.org/10.26418/justin.v6i4.27036

Dynes Rizky Navianti, I Gusti Ngurah Ray, Farida Agustini W. (2012). Penerapan Fuzzy Inference System Pada Prediksi Curah Hujan di Surabaya Utara. *Jurnal Sains Dan Seni ITS*, *I*(1), I.

Hendrian, S. (2018). Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan. *Faktor Exacta*, *11*(3), 266–274. https://doi.org/10.30998/faktorexacta.v11i3.2777

Pandey, V., Bairwa, A., & Bhattacharya, S. (2013). Application of the Pareto principle in Rapid Application Development Model. *International Journal of Engineering and Technology*, *5*(3), 2649–2654.

Rady, E. H. A., & Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked*, *15*(April), 100178. https://doi.org/10.1016/j.imu.2019.100178

Rofiq, H., Pelangi, K. C., & Lasena, Y. (2020). Penerapan Data Mining Untuk Menentukan Potensi Hujan Harian Dengan Menggunakan Algoritma Naive Bayes. *Jurnal Manajemen Informatika Dan Sistem Informasi*, *3*(1), 8–15. http://mahasiswa.dinus.ac.id/docs/skripsi/jurnal/19417.pdf

Setio, P. B. N., Saputro, D. R. S., & Bowo Winarno. (2020). Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5. *PRISMA, Prosiding Seminar Nasional Matematika*, *3*, 64–71.

Sularno, S., & Anggraini, P. (2017). PENERAPAN ALGORITMA C4.5 UNTUK KLASIFIKASI TINGKAT KEGANASAN HAMA PADA TANAMAN PADI (Studi Kasus : Dinas Pertanian Kabupaten Kerinci). *Jurnal Sains Dan Informatika*, *3*(2), 161. https://doi.org/10.22216/jsi.v3i2.2779

Wang, S., Cao, J., & Yu, P. (2020). Deep Learning for Spatio-Temporal Data Mining: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, *14*(8), 1–1. https://doi.org/10.1109/tkde.2020.3025580