Data Mining Application on Weather Prediction Using Classification Tree, Naïve Bayes and K-Nearest Neighbor Algorithm With Model Testing of Supervised Learning Probabilistic Brier Score, Confusion Matrix and ROC

Ratih Prasetya[1], Anggraeni Ridwan[2]

[1] *Education and Training Centre, The Agency for Meteorology Climatology and Geophysics, Jakarta, 10720, Indonesia*
[2] *Computer Science Department, University of Gunadarma, Depok, 16424, Indonesia*

Abstract— *One of data mining techniques is Classification, used to predict relationships between data on a dataset. The prediction performed by classifying data into several different classes considering certain factor. Classification is a performance of Supervised Learning application where the training data already has a label when entered as input data. Classification is an approach of empirical techniques that can be utilized for short-term weather prediction. The most widely used algorithms in Classification Techniques are Classification Tree, Naïve Bayes and K-Nearest Neighbors. In this study, the author used these three algorithms to predict rain with validation parameters of Brier Score, Confusion Matrix and ROC curves. The input data is synoptic data of Kemayoran Meteorological Station, Jakarta (96745) for 10 years (2006 - 2015) consists of 3528 datasets and 8 attributes. Based on a series of data processing, selection and model testing shows that the Naïve Bayes Algorithm has the best accuracy rate of 77.1% with the category of fair classification so it is quite potential to be used in the operational. The dominant weather attributes in rain formation are moisture (RHavg), minimum temperature (Tmin), maximum temperature (Tmax), average temperature (Tavg) and wind direction (ddd).*

Index Terms—Data Mining, Classification Tree, Naïve Bayes, K-Nearest Neighbour

1. Introduction

Weather prediction is a challenge in meteorology that has been a major subject of meteorological research. Research on weather prediction has been done by various methods which each method has deficiencies and advantages. Approach in weather prediction can be done by empirical or dynamic method. Short-term weather predictions have been using dynamic methods which are an analytical approach based on the principles of fluid dynamics, while empirical methods performed with statistical and mathematical approaches are more widely used for long-term weather predictions. Both approaches have their own flaws and advantages. The use of empirical methods in BMKG for short-term weather prediction are not much done yet. Related to this, the researchers are interested to examine more about how the use of empirical methods, especially data mining techniques for short-term weather prediction.

Researchers in the field of meteorology have studied how to get an accurate prediction methods with data mining techniques have comparing the methods of classification in data mining for weather prediction. The weather parameters discussed in this study are dew point, humidity, wind speed, pressure, mean sea level, wind speed and rainfall using kNN, Naïve Bayes, Multiple Regression and ID3 algorithm with good result in predicting weather [1].

2. Literature Review

The research in [2] using Artificial Neural Network (ANN) and Decision Tree (DT) algorithms to analyze meteorological data (wind speed, evaporation, radiation, minimum temperature, maximum temperature and rainfall) for 10 years (2000 - 2009) period at the Ibadan Meteorology Flight Station, Nigeria, the results show that this technique is good for weather prediction, the C5 Decision Tree is used to generate decision trees and rules to classify weather parameters.

Based on the research, it is concluded that the application of data mining technique for weather prediction by analyzing weather parameter can be done and have a good accuracy. Furthermore, the aim of data mining testing in this study using classification methods on weather prediction is to know the accuracy performance of each algorithm and which classification algorithm has the best performance for weather prediction.

Weather prediction is closely related to probability theory so that the determination of the algorithm in this study is adapted to the theory. Comparative algorithm is done by comparing the three algorithms to know which algorithm is best used for weather predictions.

The selection of algorithms in this study is based on the results of previous studies. Bayes decision theorem is a fundamental statistical approach to recognize a pattern, Naïve Bayes Classifier is a probabilistic learning method based on the Bayesian theorem [3]. Climate and groundwater prediction using K-Nearest Neighbor approach showed good performance (Mucherino et al., 2009), the use of k-NN also showed an accurate result of 93.44% [4]. The Decision Tree algorithm produces good 88.2% accuracy when applied with weather data because it can generate a good classification [5]. The three algorithms are data mining techniques that commonly used in weather prediction, it shows effectiveness and have a strong theoretical basis in hourly rainfall prediction computation model [6].

There are two methods in weather prediction, empirical and dynamical methods [7]:

a. Empirical Approach

This approach relies on research on past data to forecast future circumstances and look for relationships between attributes. The most widely used methods in the empirical approach to weather prediction are regression, decision tree, artificial neural network, fuzzy logic and other data processing methods.

b. Dynamical Approach

In dynamical approach, it is expected that the results will close to the actual state by physics modeling to predict future conditions.

The most predictable weather element is rain. Rain is a fall of hydrometeor water particles that have a diameter of 0.5 mm or more and reaches the ground.

The rain forms in the cloud after meets several conditions such as temperature, humidity, rising air currents (vortices) and environmental air conditions also the availability of this sufficient condensation.

Short-term weather predictions incorporate analogy / subjective methods with different weather models. Initially the method used is very conventional, i.e. by studying the weather conditions before and compare with the latest weather conditions to see what the future weather trends. Along with the development of science, especially mathematics and technology, a new era of numerical weather prediction concepts were introduced, Lewis Fry Richardson first introduced the concept in 1922.

This subjective method of short-term weather prediction emphasized the present and ongoing weather characteristics, observation and monitoring the synoptic-scale atmospheric conditions using numerical weather modeling as well as remote sensing such as radar and satellites. From the results of observation and monitoring combined with physical and meteorological knowledge we can identify the possibilities of unstable weather area that has chance of rain.

This dynamical methods are quite good when used during the rainy and dry seasons due to similar weather patterns from day to day during these seasons, whereas in the transitional season is not good because during the transition season, local and convective influence is greater than the synoptic scale.

Data mining is activity of finding interesting patterns from large amounts of data. The data can be stored in database, data warehouse or other information storage. Data mining deals with other fields of science such as database systems, data warehousing, statistics, machine learning, information retrieval and high-level computing. In addition, data mining is supported by other sciences such as neural network, pattern recognition, spatial data analysis, image database and signal processing.

Supervised learning is a machine-learning technique of creating a function of training data. The training data consists of the input value (usually in vector) and the expected output for the input. The task of the supervised learning machine is to predict the value of the function for all possible input values after experiencing some training data.

Classification and prediction are two forms of data analysis that can be used to extract models from data containing classes or to predict future data trends. Classification predicts data in the form of categories, whereas predictions model the functions of a continuous value. A common approach used in classification problems is training sets containing records that have known class labels to be available. The training set is used to construct a classification model, which is then applied to the test set, which contains records with unknown class labels.

2.1 *Classification Tree*

Classification and Regression Tree (CART) is one of the methods or algorithms of decision tree techniques. CART is a nonparametric statistic method that can describe the relationship between response variables (dependent variables) with one or more predictor variables (independent variables).

The formation of the classification tree consists of several stages requiring learning sample L. The first stage is the selection of the divider. Each sorting depends only on the value derived from an independent variable. For the continuous independent variable Xj with the sample space of size n and there are n different sample observation values, then there will be n - 1 different sorting.

While for Xj is the nominal category variable with L, it will get sorting as much as $2^{L-1} - 1$. But if variable Xj is ordinal category it will get L - 1 sorting possible. The most commonly used sorting method is the Gini index with the following functions [8]:

$$i(t) = \sum_{i \neq j} p(i|t)p(j|t) \quad \ldots \ (1)$$

with i (t) is the heterogeneous function of the gini index, p (i|t) is the proportion of class i in the t symbol, and p (j|t) is the proportion of class j at node t.

The development of the tree is done by looking for all possible dividers at vertex t1 to find the s * s that gives the highest degrees of heterogeneity,

$$\Delta i(s^*, t_1) = \max_{s \in S} \Delta i(s, t_1) \quad \ldots .(2)$$

With $\phi(s, t)$ is *goodness of split criteria*, P_Li (t_L) is the proportion of observations from node t to right node.

The second stage is the determination of the terminal node. The t knot can be used as a terminal node if there is no significant drop of heterogeneity at the sorting, there is only one observation (n = 1) in each child node or a minimum limit of n and a limitation of the maximum level or depth of the tree.

The third stage is labeling each terminal node based on the rules of the most number of class members, namely:

$$p(j_0|t) = \max_j \ p(j|t) = \max_j \frac{N_j(t)}{N(t)} \quad \ldots .(3)$$

with $p(jo|t)$ is j class proportion at node t, $Nj(t)$ is the number of j class observation at node t, and $N(t)$ is observation number at node t. Class label at terminal node *t* and jo that gives the highest error guessing error value of the *t* knot.

The process of forming a classification tree stops when there is only one observation in each child node or a minimum n limit, all observations in each child's node are identical and there is a limit to the maximum number of tree levels.

2.2 *Naïve Bayes*

Naïve Bayes is one of the most effective and efficient inductive learning algorithms for machine learning and data mining. The performance of Naïve Bayes is competitive in the classification process although using the assumption of attribute independence (no attribute linkage). The assumption of the independence of these attributes on the data is rare, but although the assumption of attribute independence is breached the performance of Naïve Bayes classification is quite high, as evidenced by the various empirical studies [9].

The Naïve Bayes prediction is based on Bayes's theorem with the following formula for classification (Prasetyo, 2012):

$$P(Y|X) = \frac{P(Y) \prod_{i=1}^{q} P(Xi\,|Y)}{P\ (X)} \quad \ldots\ldots..(4)$$

While Naïve Bayes with continuous features have a formula:

$$\mathrm{P(X|\ Y)} = \frac{1}{\sqrt{2\pi}\ \sigma}\, exp^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \ldots\ldots\ (5)$$

Remarks:

$P(Y \mid X)$ = Data probability with X vector on Y class

$P(Y)$ = Initial probability Y Class

$\prod_{i=1}^{q} P(Xi\,|Y)$ = Independent probability of Y class from all features from X vector

μ = mean value of attribute with attribute with continue features

σ = standard deviation

2.3 *K – Nearest Neighbour*

The kNN method is a Machine Learning algorithm which is considered as a simple method to be applied in data analysis with many dimensions [11]. Although this method is simple, this method has advantages compared to other methods, which can generalize a relatively small set of training data (Rokach., 2010).

The k-nearest neighbor (k-NN) method is a method for classifying objects based on learning data that is the closest distance to the object. Learning data is projected into multi-dimensional space, where each dimension represents the features of the data. This space is divided into sections based on the classification of learning data. A point in this space is marked as class c if class c is the most common classification in the nearest k of the point. Near or far neighbors are usually calculated based on Euclidian distances whose equations are as follows [13]:

Untuk $P = (p_1, p_2, \ldots, p_n)$ dan $Q = (q_1, q_2, \ldots, q_n)$, maka

$$Jarak = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2}$$

$$Jarak = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \qquad ..(6)$$

2.4 Classification Performance Testing

Cross Validation is a general method used to evaluate classifier performance. Cross Validation is a simple form of statistical technique. The standard fold number to predict the error rate of the data is to use 10-fold cross validation [14].

In general, the use of prediction systems depends on several predictive quality attributes. Predictive quality attributes for the probabilistic category are sharpness, resolution, discriminant, bias, reliability *(calibration)*, *accuracy* and *skill*. The attribute that will be discussed in this research is accuracy.

Accuracy is a measurement of the compatibility between predictions and observations. The appropriate measurement for this probability forecast is the Brier Score.

Accuracy analysis in the dichotomy prediction is prediction in the form of two categories of rain or cannot be done to determine the level of accuracy of the classification algorithm used. The accuracy value is obtained from a contingency matrix that is a square matrix called the "error matrix" or "confusion matrix" and the ROC (receiver operating characteristic) curve. Confusion matrix table is a table created to link the classification results with the data obtained for accuracy testing.

3. Research Methods

This type of research is quantitative research (quantitative research). A 10 years (2006-2015) as many as 3528 datasets obtained from the Database Subdivision, Deputy IV, BMKG. Surface observational data (synoptic) is observed every hour, the synoptic data in the form of a password is then translated and input into Ms Excel. Weather elements observed were Temperature, Pressure, Visibility, weather conditions, wind direction, wind speed, dew point, cloud type, number of clouds, solar radiation, duration of sun exposure and others.

The location of the study is Jakarta Kemayoran Meteorological Station (96745) which is located at Jalan Angkasa I No.2, Kemayoran Jakarta Pusat. The thinking chart as a guidance in performing this research as shown in Figure 1.

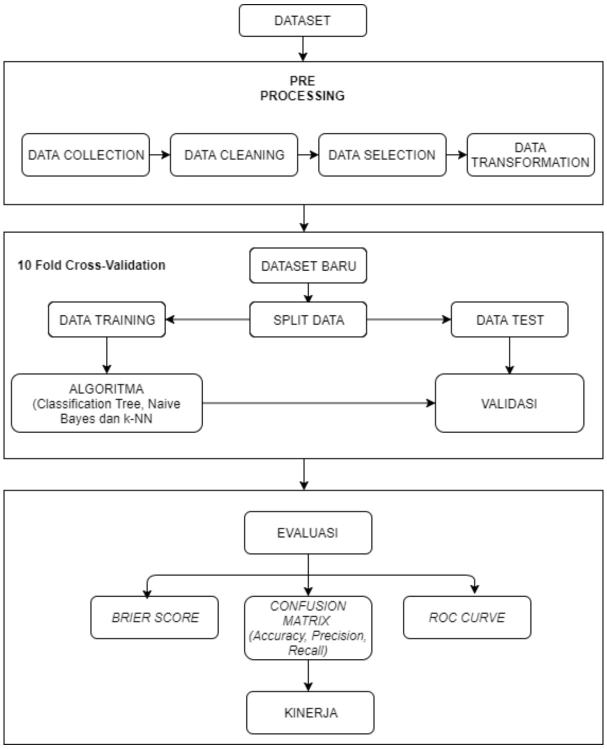


Figure 1. Thinking Chart

4. Results and Analysis

Overall the process of making a model using the Classification Tree, Naïve Bayes and k-Nearest Neighbor algorithms looks like in Figure 2.

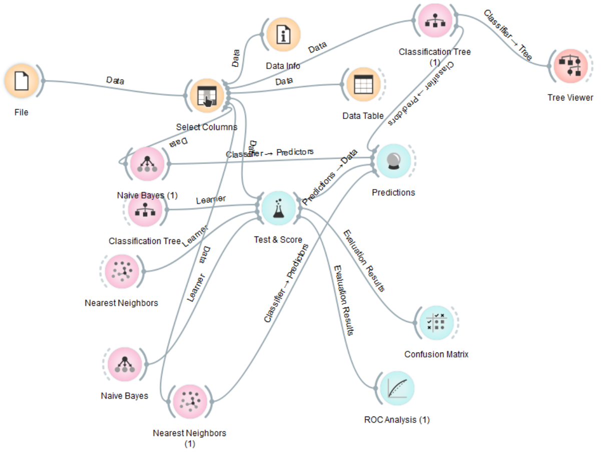


Figure 2. The Classification Process in Data Mining Software (Orange Ver. 3.23.1)

Surface weather (synoptic) observation data is processed using 3 classification algorithms, namely Classification Tree, Naïve Bayes and k-NN. The output model of each method is tested with some input data to see the reliability of the model. Furthermore, the test results are compared to get the highest accuracy value so that the algorithm can be determined which is the best.

In the process of testing the data is divided into 2 namely training data (training) and test data (test data). 70% training data is used as a mining process and obtains probability values while 30% test data is used to test the probability values that have been formed.

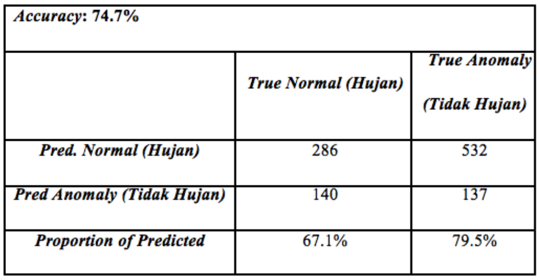
4.1 Brier Score, Confusion Matrix and ROC Curve

In general, the Brier Score value on the three algorithms has a tendency of an average error value of 0.09% obtained from the total Brier Score value on the three algorithms divided by the number of algorithms used. In the Classification Tree algorithm with an error value of 0.11, Naïve Bayes with an error value of 0.10 and k-NN with an error value of 0.07. From table 4.2 it can be concluded that the prediction of rain with probabilistic categories through classification has a maximum confidence level of 93%, namely the lowest maximum Brier Score indicates a value of 0.07 (Table 4.2 column k-NN). To find out the percentage of confidence from rain predictions with the probabilistic category from the data in table 4.2, all values are simply multiplied by 100%, so the results will represent a percentage of probabilistic prediction trust from the supervised learning model with the Classification method.

The highest prediction of rain chance prediction performance is in the k-NN algorithm with a confidence level of 93% (the result of a 100% -7% reduction), in the Naïve Bayes algorithm the confidence level reaches 90% while in the Classification Tree algorithm the confidence value reaches 89% .

Confusion matrix test is performed to obtain the value of precision, recall and accuracy of the test results. The test results are to measure the accuracy and Area Under Curve (AUC) of the determination by the 10-fold Cross Validation method. Following are the test results of each algorithm:

Table 1 Confusion Matrix table for Classification Tree

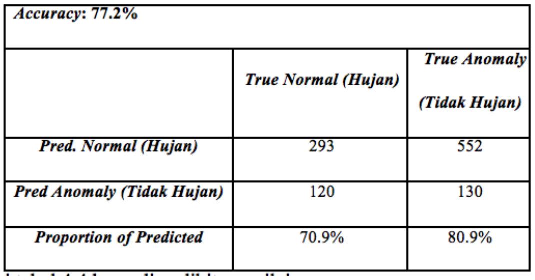


Accuracy: 74.7%		
	True Normal (Hujan)	*True Anomaly* *(Tidak Hujan)*
Pred. Normal (Hujan)	286	532
Pred Anomaly (Tidak Hujan)	140	137
Proportion of Predicted	67.1%	79.5%

Precision = 0.74; *Recall* = 0.74; *Accuracy* = 0.74 dan *AUC* = 0.73

Based on the results in Table 1, it can be seen that the level of accuracy using the Classification Tree algorithm is 74% with the number of true predictions is 818 datasets from the total amount of data tested that is 1095 datasets.

Table 2 Confusion Matrix results for Naïve Bayes

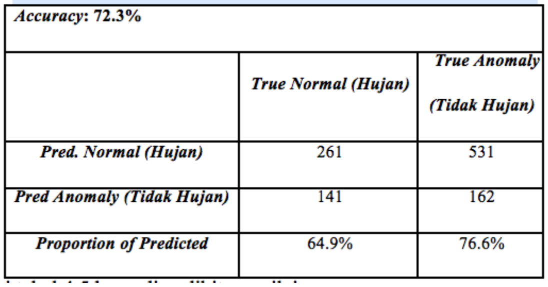


Accuracy: 77.2%		
	True Normal (Hujan)	*True Anomaly* *(Tidak Hujan)*
Pred. Normal (Hujan)	293	552
Pred Anomaly (Tidak Hujan)	120	130
Proportion of Predicted	70.9%	80.9%

Precision = 0.77; *Recall* = 0.77; *Accuracy* = 0.77 dan *AUC* = 0.75

Based on the results in Table 2, it can be seen that the level of accuracy using the Naïve Bayes algorithm is 77% with the number of true predictions is 845 datasets out of the total amount of data tested, namely 1095 datasets.

Table 3 *Confusion Matrix* for *k-Nearest Neighbour*



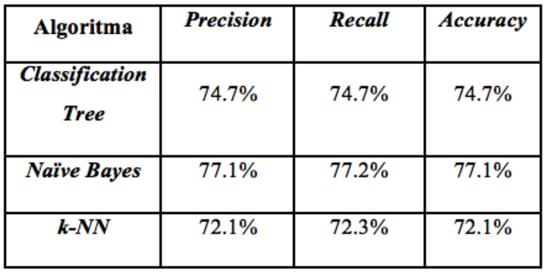
Accuracy: 72.3%		
	True Normal (Hujan)	*True Anomaly* *(Tidak Hujan)*
Pred. Normal (Hujan)	261	531
Pred Anomaly (Tidak Hujan)	141	162
Proportion of Predicted	64.9%	76.6%

Precision = 0.72; *Recall* = 0.72; *Accuracy* = 0.72 dan *AUC* = 0.70

Based on the results in Table 3, it can be seen that the level of accuracy using the Naïve Bayes algorithm is 72% with the number of correct predictions is 792 datasets out of the total amount of data tested namely 1095 datasets.

The results of the performance accuracy values for the three algorithms can be seen in table 4.6 with the 10-fold cross validation sampling method.

Table 4 Table of test results for accuracy values



Algoritma	***Precision***	***Recall***	***Accuracy***
Classification Tree	74.7%	74.7%	74.7%
Naïve Bayes	77.1%	77.2%	77.1%
k-NN	72.1%	72.3%	72.1%

From the test results it can be seen that the values of precision, recall, accuracy and AUC for each experiment. The highest precision value of all tests is in the Naïve Bayes algorithm which is 77.1%, while the lowest precision value is in the k-NN algorithm which is 72.1%. The highest Recall value of all tests is in the Naïve Bayes algorithm which is 77.2%, while the lowest Recall value is in the k-NN algorithm which is 72.3%.

Classifier accuracy measures or classification accuracy measurement values are the percentage of the number of data records correctly classified by an algorithm after testing the classification results. Accuracy can also be defined as the level of closeness between the predicted value and the actual value. The highest accuracy value is in the Naïve Bayes algorithm which is 77% where the value is the maximum value for accuracy which means that the algorithm is very good at predicting weather, followed by the Classification Tree algorithm at 75% measurement results and the k-NN algorithm at measurement results 72 % The three algorithms have the equation which shows fair classification which means it is good enough to be used for weather prediction.

4.2 ROC Curve

The ROC curve shows accuracy and compares classification visually. ROC expresses confusion matrix. ROC is a two-dimensional graph with false positives as horizontal lines and true positives as vertical lines (Vercellis, 2009).

To find out the discriminant strength of the three algorithms, one way that can be used is through the ROC curve. The results of the ROC curve analysis can be seen in Figure 3.

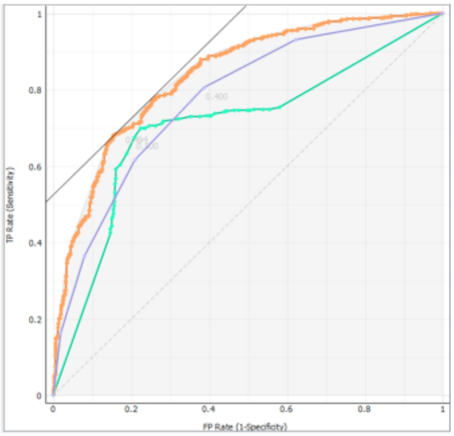


Figure 3. ROC values for Classification Tree, Naïve Bayes and k-NN

The closer the ROC curve is to the Y line (0.1), the better the model predicts the weather. Based on Figure 3 it is known that the curve has a shape that tends towards the Y line where the whole curve is above the dashed line. To make sure the analysis can be seen from the AUC table. The value of the area under the curve will usually be at 0.5 and 1. If the value of the area under the curve approaches 1, the model is more accurate. The highest AUC value is in the Naïve Bayes algorithm with a value of 75.7, Classification Tree with a value of 75.7 and k-NN with the lowest value of 70.4. The area under the ROC curve for the three algorithms is shown in table 5 below.

Table 5 Test results with AUC values

	Classification Tree	*Naïve Bayes*	*K-NN*
AUC	73.4	75.7	70.4

The results obtained for the AUC value of the Classification Tree, Naïve Bayes and k-NN algorithms can be seen in Table 4.5 with a fair classification level which means that the accuracy of the model is good enough so that the three algorithms can be used for weather prediction.

Probabilistic Forecast is a weather prediction information product that is made based on the chance event format. Probabilistic forecast information is based on the ensemble distribution forecast [16]. In this study the chance prediction is obtained by counting the number of events in a particular category from a rain prediction distribution from the run of three classification algorithms which are then divided by the number of datasets (1095 datasets), then multiplied by 100% to get the percentage value of the rain prediction probability.

The prediction parameter used in the processing of probabilistic forecast is rainfall, the representation of the results of the run of three classification algorithms is the chance of rain events. The results of testing of the three algorithms can be seen in Figure 4.

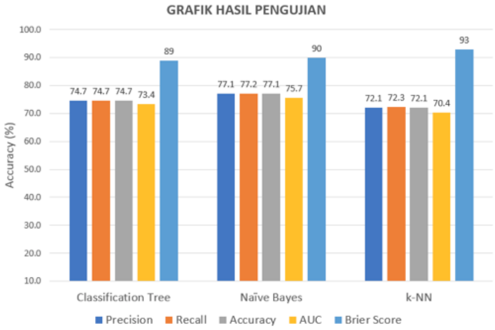


Figure 4. Graph of test results

Based on Figure 4 it can be seen that the Naïve Bayes algorithm has the highest prediction accuracy based on the Confusion Matrix test parameters and the ROC curve, namely precision 77.1%, recall 77.2%, accuracy 77.1%, AUC 75.7%. While the test value with the Brier Score parameter shows a higher predictive accuracy value than the other test parameters namely Classification Tree 89%, Naïve Bayes 90% and k-NN 93%.

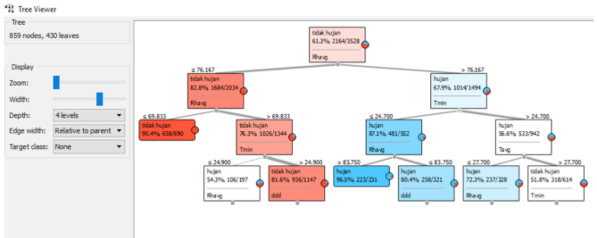
In general, it can be seen that the Naïve Bayes algorithm has the potential to be used operationally with an average accuracy rate of 75.42%. This can be caused because in this study the dataset used is quite large, amounting to 3528 and in accordance with the characteristics of Naïve Bayes which is suitable to be applied to large datasets (> 1500 datasets)[17].

The accuracy value in the Brier Score test has a greater value than the other test parameters with an average of

90%. This is because the Brier Score test parameters are very in accordance with the definition of probabilistic predictions with the possibilities displayed in the form of numbers from 0 to 1 and are suitable for binary forecasting [18] . This test parameter has been widely used by researchers from the World Meteorological Organization (WMO) for rain prediction. Basic measurements based on probabilistic forecast skills for binary events are the Brier Score and the ROC [19].

The Classification Tree parameter is the Induce Binary Tree which functions to build a binary tree by breaking it into two child nodes. Min Number of Instances in Leaves is where the algorithm will not produce nodes other than the specified number. Do Not Split Smaller Than Subsets functions to prevent the algorithm from breaking the node with a smaller number of instances specified. Limit the Maximal Tree Depth serves to limit the number of node depths to the specified number of node depths. Stop When Majority Reaches (%) stops the node after a certain threshold has been reached.

Problems in classification can be processed by submitting a number of attributes from the test record. Each time a node is obtained, a number of attributes are processed again until a conclusion is reached about the class label of the record. The series of processes can be associated in the form of a decision tree which is a hierarchical structure consisting of nodes and nodes. The decision tree for processing the Classification Tree algorithm can be seen in Figure 5.


Tree Viewer

Tree

859 nodes, 430 leaves

Display

Zoom:

Width:

Depth: 4 levels

Edge width: Relative to parent

Target class: None

Figure 5. Classification Tree for weather classification (rain or not)

The results of data processing in the Classification Tree produced 859 nodes and 430 leaves with a depth of 5 levels. The root node, which is the average Receiver Humidity attribute (RHavg), is used to separate rain or no rain weather conditions, from the total dataset of 3528 there are 2164 datasets or 61.3% chance of non-rainy weather. Furthermore, from the root node is broken down into 2 internal nodes namely RHavg attribute and minimum temperature (Tmin). If the RH value is. 76.167%, then the weather conditions have a chance of not raining and if the value of RH> 76.167% then the weather conditions have the opportunity to rain. From the internal node Tmin is broken down again into 2 internal nodes namely RHavg and average temperature (Tavg), if Tmin ≤ 24.70 C then the weather conditions have a chance of rain and if Tmin> 24.70 C then the weather conditions have little chance of rain.

In the last process, the Internal node RH produces a leaf node of rain when the RH value is> 83.75%, the chance of rain is quite large namely 96.5% and if the RH value is ≤ 69.83% then the chance of not raining is large enough that is 95.4%.

The Classification Tree succeeded in classifying which parameters were most influential on rainfall prediction, based on their successive levels, namely RHavg, Tmin, RH, ddd (wind direction), LPM (duration of solar radiation) and Tmax. In addition, from the classification tree formed, it is also known the chance of rain intensity that will occur, this can be seen from the stronger colors of each node.

5. Conclusion

Based on research on short-term weather prediction using the classification algorithm on rain prediction based on probabilistic supervised learning with Brier Score, Confusion Matrix and Receiver Operating Characteristic test parameters, the following conclusions can be drawn:

1. Based on the results of the three test parameters namely Brier Score, Confusion Matrix and ROC curve, the three algorithms can be applied to weather data with a fairly good category, namely fair classification.

2. Comparison of classification algorithms namely Classification Tree, Naïve Bayes and k-NN test results show that Naïve Bayes has the best predictive probability for short-term weather, namely values of precision 77.1%, recall 77.2%, accuracy 77.1%, area values under the AUC 75.7 curve and 90% Brier Score. Thus, the Naïve Bayes algorithm is quite potential to be used operationally.

3. Knowledge interpretation that can be applied for short-term weather prediction, especially for the Classification Tree algorithm is that the algorithm successfully classifies 1095 dataset test data into 287 nodes and 144 leaves with the most significant weather parameters to the formation of rain is humidity (RHavg), temperature minimum (Tmin), maximum temperature (Tmax), average temperature (Tavg) and wind direction (ddd).

6. Recommendation

This research should be continued by using other test parameters such as RMSE (Root Mean Squared Error) or by increasing the number of test parameters. Brier Score test parameters should be applied with several different dataset ranges.

The empirical method with this data mining model should be tested not only at one point in the weather station but also uses observational data distribution of several weather stations (spatial) so that it can better represent certain areas.

References

[1] G. Gaikwad and V. B. Nikam, "Different Rainfall Prediction Models And General Data Mining Rainfall Prediction Model," *Int. J. Eng. Res. Technol.*, vol. 2, no. 7, pp. 115–123, 2013.

[2] D. Chauhan and J. Thakur, "Data Mining Techniques for Weather Prediction: A Review," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 2, no. 8, pp. 2184–2189, 2014.

[3] L. Lan and S. Vucetic, "Improving accuracy of microarray classification by a simple multi-task feature selection filter," *Int. J. data Min.*, vol. 5, no. 2, pp. 189–208, 2011.

[4] V. Kunwar, K. Chandel, A. S. Sabitha, and A. Bansal, "Chronic Kidney Disease Analysis Using Data Mining Classification," pp. 300–305, 2016.

[5] S. D. E.Manjula, "Analysis of Data Mining Techniques for Agriculture Data," *Int. J. Comput. Sci. Eng. Commun.*, vol. 4, no. 2, pp. 1311–1313, 2016.

[6] B. N. Lakshmi, "A Comparative Study of Classification Algorithms for Risk Prediction in Pregnancy," pp. 0–5, 2015.

[7] S. S. Bhatkande and R. G. Hubballi, "Weather Prediction Based on Decision Tree Algorithm Using Data Mining Techniques," vol. 5, no. 5, pp. 483–487, 2016.

[8] D. R. Tobergte and S. Curtis, "Metode Classification," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2013.

[9] M. A. Shadiq, "Keoptimalan Naïve Bayes Dalam Klasifikasi," no. 1, p. 31, 2009.

[10] "No Title," 2013.

[11] K. Alkhatib, H. Najadat, I. Hmeidi, and M. Shatnawi, "Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm," *Ijbhtnet.Com*, vol. 3, no. 3, pp. 32–44, 2013.

[12] D. Prediction, U. Classification, M. Naive, and D. Tree, "Prediksi Keputusan Menggunakan Metode Klasifikasi Naïve Bayes , One-R , Dan Decision Tree," pp. 1–10, 2016.

[13] N. C. Barde and M. Patole, "Classification and Forecasting of Weather using ANN, k-NN and Naïve Bayes Algorithms," *Int. J. Sci. Res.*, vol. 5, no. 2, pp. 1740–1742, 2016.

[14] I. H. Witten, E. Frank, and M. a Hall, *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. 2011.

[15] M. N. Sholikhin and Y. Rahayu, "Analisis Delay Penerbangan Akibat Cuaca di Bandara Ahmad Yani Semarang dengan Algoritma C4 . 5," vol. 5, pp. 1–10, 2013.

[16] T. M. Hamill, "Brier Skill Scores, ROCs, and Economic Value Diagrams Can Overestimate Forecast Skill," *Mon. Weather Rev.*, vol. 1, no. 303, pp. 1–29, 2005.

[17] S. D. Jadhav and H. P. Channe, "Comparative Study of K-NN , Naive Bayes and Decision Tree Classification Techniques," vol. 5, no. 1, pp. 2014–2017, 2016.

[18] J. Lo, "Help from Weather Forecasters From Verification to Validation The Brier Score," pp. 1–6, 2013.

[19] T. Memorandum and T. N. Palmer, "Predicting uncertainty in forecasts of weather," no. 294, 2003.