Performance of Naive Bayes method with data weighting

Afdhaluzzikri[1)*], Herman Mawengkang[2)], Opim Salim Sitompul[3)]
[1,2,3)]Universitas Sumatera Utara, Medan, Indonesia
[1)]90afdhal@gmail.com, [2)]mawengkang@usu.ac.id, [3)]opim@usu.ac.id

Submitted : July 4, 2022 | **Accepted** : July 5, 2022 | **Published** : July 6, 2022

Abstract: Classification using naive bayes algorithm for air quality dataset has an accuracy rate of 39.97%. This result is considered not good and by using all existing data attributes. By doing pre-processing, namely feature selection using the gain ratio algorithm, the accuracy of the Naive Bayes algorithm increases to 61.76%. This proves that the gain ratio algorithm can improve the performance of the naive bayes algorithm for air quality dataset classification. Classification using naive bayes algorithm for air quality dataset. While the Water Quality dataset has an accuracy rate of 93.18%. These results are considered good and by using all existing data attributes. By doing pre-processing, namely feature selection using the gain ratio algorithm, the accuracy of the Naive Bayes algorithm increases to 95.73%. This proves that the gain ratio algorithm can improve the performance of the naive bayes algorithm for air quality dataset classification. Classification using Naive Bayes algorithm for Water Quality dataset. Based on the tests that have been carried out on all data, it can be seen that the Weight nave Bayes classification model can provide better accuracy values because there is a change in the weighting of the attribute values in the dataset used. The value of the weighted Gain ratio is used to calculate the probability in Naïve Bayes, which is a parameter to see the relationship between each attribute in the data, and is used as the basis for the weighting of each attribute of the dataset. The higher the Gain ratio of an attribute, the greater the relationship to the data class. So that the accuracy value increases than the accuracy value generated by the Naïve Bayes classification model. The increase in accuracy in the Naïve Bayes classification model is due to the amount of weight accuracy from the attribute selection in the Gain ratio.

Keywords: Naïve Bayes, Gain Ratio, Air Quality, Water Quality, Accuracy.

INTRODUCTION

Naive Bayes is a probability classification model that is easier in machine learning by performing calculations from datasets that aim to predict probabilities in a class with the assumption of strong independence. Classification is a directed learning step. Classification functions as an estimate of the class of objects whose parts are unknown(Raviya & Gajjar, 2013). Decision Tree, K-Nearest Neighbor, Naïve Bayes, Neural Network and Support Vector Machines are classification methods that are very often used (Sahu, Nagwani, Verma, & Shirke, 2015).

In research conducted by (Mao, Zhao, & Sun, 2017) local attribute weighting uses the K-Nearest Neighbors Classifier (KNN) method. In his research, he produced the closest K number from the data used and calculated the probability value for each attribute which would later be given a weighting for each attribute and then the data was classified using Naïve Bayes. The data used is the travel history of bus routes and weather condition data from August to December in 2014. The results of the study achieved an accuracy rate of 89% using the Naïve Bayes Local Attribute Weighted KNN method.

One of several methods used in the process of classifying hidden data and getting information from a lot of data. The Naïve Bayes classification method in this study is combined with a Decision Support System because this method does not need to use weights in the data calculation process, but the data used has a high probability.

there has been. The reason for choosing Naïve Bayes to carry out classification in his research is because the classification opportunities are easier, the computations are very effective, the accuracy is good, and they have the ability to classify based on simple mathematical categories (Yuliana & Erlangga, 2017).

Based on research from (Niloy & Navid, 2018), conducting research on two data mining techniques, the results of his research show difference in error rates between the two methods. However, there are also relatively

*name of corresponding author

large differences in the area ratio. Naïve Bayes performs a more precise classification than the Classification Tree. Classification is needed in measuring data accuracy.

In a study conducted (Raymond & Jason, 2011) to compare Naïve Bayes with the performance of the Weighting feature method that has been tested using the dataset that has been provided. The research was conducted using traditional methods and using feature weighting where the comprehensive experimental results validate the effectiveness of the Feature Weighting method. The results show that the performance of Naïve Bayes improves after using the feature weighting method. Naïve Bayes is a statistical classification method by predicting the probability in a class with the assumption of a strong independent. So, the researcher proposes the attribute weighting of the data sample before the classification stage is carried out. It aims to increase the accuracy value of the conventional Naïve Bayesian method by applying attribute reduction by assigning weights to the attributes of the dataset in order to obtain higher accuracy results from the classification process carried out.

LITERATURE REVIEW

Some researchers mention classification with the term forecast in conducting class categorization and processing data classified into several parts or creating a pattern or model by processing existing training data in order to create a value for classes on each attribute in the dataset and using data groups. new. Usually, the classification is used in processing bank users in doing credit, sales targets, diagnosis of illness in hospitals, and analysis of several problems that require decision results.

Naïve Bayes Classifier

In this classification process requires classes, data to predict, as well as data to train, and test data. This process is carried out by parsing the existing set of classes and using a pattern that aims to classify data tuples whose class has not been detected. Then the predictor aims at determining patterns in data classification, Decision Tree, or mathematical formulas. So, the purpose of the training data is to get the results of grouping data with existing classes and predictors. While the test data is generated from the grouping of the classification model that has been processed to obtain new data.

Naïve Bayes' probabilistic classification with simple calculations using a set of probabilities with the number of frequencies and the combination of values from the dataset. The Naïve Bayes algorithm is assumed to have each attribute that is independent or not dependent on the data contained in each class variable (Patil & Sherekar, 2013). Naïve Bayes has another meaning, namely classifying using probability and statistical methods that was once put forward by a British scientist named Thomas Bayes by predicting future opportunities based on previous experiences (Bustami, 2014).

Cross-Validation

In the cross-validation approach, each record is used the same number of times for training and once for testing. This method can be simulated as follows; suppose we divide the data into two subsets of the same size. First, take one subset for the training data then one for the testing data. Next, perform the exchange of functions from the subset that was previously the training set to the test set and vice versa. This modeling uses two-fold cross-validation. The number of errors is obtained from the sum of the errors for the two processes. In this illustration, each record is used once for training and once for testing. In general, the k-fold cross-validation method applies data division into k parts of the same size.

In certain cases, the k-fold cross-validation method uses the equation k = N, N is the size of the data set. This method is called leave-one-out, each test set has one record. The advantage of this method is that it can use a lot of data for training. Test sets are mutually exclusive and effectively cover the entire data set. The disadvantage of this method is that the computation will be performed N times (Tan, Steinbach, & Kumar, 2013)

Stratified 10-fold cross-validation can be an option to get accurate validation results in standard evaluation methods. The way the 10-fold cross-validation works is by repeating the test ten times and the average value obtained from ten tests becomes the measurement result. The benefit of this method is that it avoids overlapping of the testing data. Test sets are mutually exclusive and effectively cover the entire data set. The disadvantage of this method is that computation will be repeated N times (Gorunescu, 2011).

Accuracy with unbalanced data will be dominated by accuracy in minority class data, so the AUC (Area Under the ROC Curve) matrix is used, F-Measurement, G-Mean, overall accuracy, and accuracy for the minority class (Zhang & Wang, 2011). The results of the F-Measure assessment, the average integer multiple of two numbers tends to be close to a number smaller than two. So, a high F-Measure value can be used as a reference for both recall (sensitivity) and high precision. True Positive rate (TP rate) and Positive Predictive Value (PP value) are considered important for positive class performance. PP value is defined as the precision that

*name of corresponding author

represents the relevant object presentation defined for retrieval. TP Rate in retrieval is defined as a recall to show the object taken is relevant. The combined precision measures of recall are called harmonics.
In the test data that has been processed so that prediction results are obtained, then the process is carried out to obtain good prediction results then the next step is carried out by measuring the performance of the method used. The measurements made are very important. Because the performance measurement is done by comparing the prediction results of the classification algorithm with the target value on the test data variable from the actual data (Wang, 2014).

METHOD

Research methodology is structured concepts about the rules, activities, and procedures used by researchers in telling how the research process is carried out. The research method owned by the researcher is about the steps and processes experienced during conducting research. The research design on the Naïve Bayes algorithm is as follows:

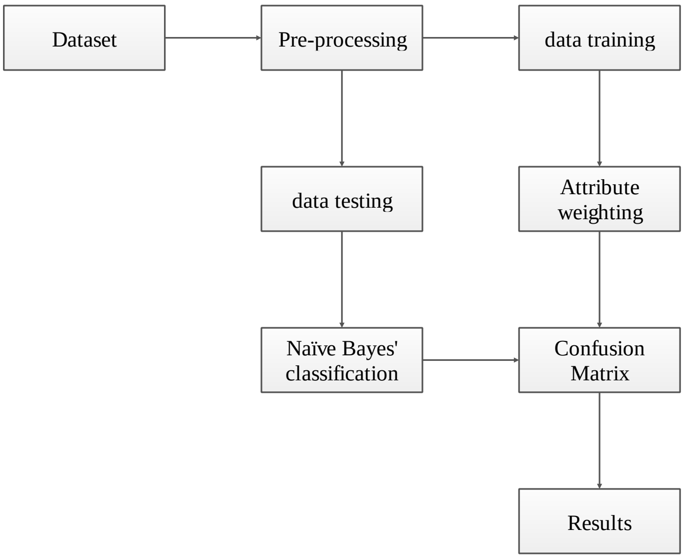

Fig 1. Model Design

Data Preprocessing

In this study, a preprocessing process is carried out, namely the process of handling missing values, which in the process is to process attributes that have numeric values replaced into mean values with attributes that have similar columns. Then the nominal ones are replaced with the most possible values in the same column for each attribute. Then the two data processing is handled with data that contains duplicates. The attributes used have 13 kinds of attributes, 3 classes and 9357 instances. Then the third process is the use of the Z-Score method which has performed standardized data normalization so that the interval becomes more proportional.

Naïve Bayesian Classification Model

The process of forming a Nave Bayesian Classification Model The next step is to form a classification model using the Nave Bayesian method. Bayes' rule is that the outcome (C, target) can be estimated based on the number of test samples (X, attribute) being observed. There are several important things from Bayes' rules, namely: 1) A prior probability (C, target) or P(C) is the probability of a hypothesis before the evidence is observed. 2) A final probability (C, target) or P(C|X) is the probability of a hypothesis after the evidence has been observed. Please note that the classification process requires a number of clues to determine what class is appropriate for the sample data being analyzed.

Confusion Matrix

In the classification that uses Naïve Bayes with test data and training data, calculations are carried out in carrying out test data which will then be displayed in the confusion table. matrix (Witten, 2005). The classification generated by the test data has different classes, namely positive and negative.

RESULT

In the weighting scheme of each attribute of the air quality dataset obtained from UCI Machine Learning. This study uses the Gain ratio as a parameter. To simplify calculations, the Rapid Miner® version 5.3 application is used. There is a highest and lowest value for each attribute. After that, the Naïve Bayes classification model was

*name of corresponding author

carried out and used the Gain ratio approach to the dataset used. Testing the results of the accuracy value based on the Confusion matrix.

In this study, two preprocessing processes were carried out, namely attributes that had numeric values (numbers) replaced with average values (means) of attributes in the same column. Cleaning is a way to get rid of duplicate data so that the original Air Quality dataset is 9357 instances to 9326 instances and the original Air Quality dataset is 120 instances to 117 instances.

Table 1 Naïve Bayes Air Quality Accuracy Results

	true 1	true 0	class precision
pred. 1	3383	5503	38.07%
pred. 0	95	345	78.41%
class recall	97.27%	5.90%	

Table 2 Nave Bayes Gain Ratio Air Quality Accuracy

	True 1.0	True 0.0	Class Precision
Pred. 1.0	136	224	37.78%
Pred. 0.0	3342	5624	62.73%
Class Recall	3.91%	96.17%	

The results of the research that have been carried out have found that the stages of the process carried out in this study, after the weighting process for each attribute is obtained using a Gain ratio, then the next step is to classify using the Naïve Bayes method, each attribute is given a weight which is generally called Weight Naïve Bayes (WNB). In this study, this method has been able to reduce the influence of irrelevant attributes on the data class so that it can affect the accuracy value.

In order to explain the accuracy value obtained from the Naïve Bayes classification model and the Weight Naïve Bayes classification model (gain ratio) to the air quality dataset, it is seen from the results of the measurement of the accuracy of predicting each classification model which can be described as in the following table:

Table 3 Result

NAÏVE BAYES		NAÏVE BAYES+GAIN RATIO	
Water quality	Air Quality	Water quality	Air Quality
93.18%	39.97%	95.73%	61.76%

Based on the tests that have been carried out on all data, it can be seen that the Weight nave Bayes classification model can provide better accuracy values because there is a change in the weighting of attribute values in the dataset used. The value of the weighted Gain ratio is used to calculate the probability in Naïve Bayes as a parameter to see the relationship between each attribute in the data, and is used as the basis for the weighting of each attribute of the dataset. The higher the Gain ratio of an attribute, the greater the relationship to the data class. So that the accuracy value increases than the accuracy value generated by the Naïve Bayes classification model. The increasing accuracy of the Naïve Bayes classification model is due to the number of weights from the attribute selection to the Gain ratio accuracy.

DISCUSSIONS

In the process of classifying data with the Naïve Bayes algorithm using the Air Quality dataset, the accuracy result is 39.97%. The result of this accuracy is a poor level of accuracy by using all the attributes in the Air Quality dataset. So that the researchers carried out the feature selection process using the Gain Ratio method so as to get better accuracy results, which was 61.76%. The Gain Ratio method has improved the accuracy performance of the Naïve Bayes algorithm in the process of classifying Air Quality datasets. There is a 21.79% increase in accuracy obtained after selecting features using the gain ratio.

In the process of classifying the data using the Naïve Bayes algorithm that uses the Water Quality dataset, the accuracy is 93.18%. The result of this accuracy is a good level of accuracy by using all the attributes in the Water Quality dataset. So that the researchers carried out the feature selection process using the Gain Ratio method so as to get an accuracy of 95.73%. The Gain Ratio method has improved the accuracy performance of the Naïve Bayes algorithm in the water quality dataset classification process. There is a 2.55% increase in accuracy obtained after selecting features using the gain ratio.

CONCLUSION

The success of increasing the accuracy of the dataset by using the weighting of the Naïve Bayes algorithm is very influential on all attributes, so it greatly affects the accuracy of the resulting data. The Gain Ratio method is a technique used to calculate probability in Naïve Bayes, as a parameter to see the relationship between each attribute in the data, and is used as a basis for weighting each attribute of the dataset. The higher the Gain ratio of an attribute, the greater the relationship to the data class. In the Naïve Bayes algorithm using the Air Quality dataset, the accuracy results are 39.97% and the Air Quality dataset 93.18%. The results of this accuracy are not good accuracy levels using all existing attributes. While the application of the Gain Ratio method got an increase in accuracy results of 61.76% on the air quality dataset, then the accuracy results also increased accuracy to 95.73% on the water quality dataset. These results prove that the application of the Gain Ratio method is very influential on accuracy. Application to different datasets will be considered in the future while other algorithms will be considered.

REFERENCES

Tan, P.-N., Steinbach, M., & Kumar, V. (2013). *Introduction to Data Mining.* Boston: Pearson Addison-Wesley.

Bustami, B. (2014). PENERAPAN ALGORITMA NAIVE BAYES UNTUK MENGKLASIFIKASI DATA NASABAH ASURANSI. *Jurnal Informatika (JIFO)*, 884-898.

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques.* Heidelberg: Springer Berlin.

Mao, X., Zhao, G., & Sun, R. (2017). Naive Bayesian algorithm classification model with local attribute weighted based on KNN. *IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC).* IEEE.

Niloy, N., & Navid, M. (2018). Naïve Bayesian Classifier and Classification Trees for the Predictive Accuracy of Probability of Default Credit Card Clients. *American Journal of Data Mining and Knowledge Discovery*, 1-12.

Patil, T. R., & Sherekar, S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications*, 256-261.

Raviya, K. H., & Gajjar, B. (2013). Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA. *PARIPEX - INDIAN JOURNAL OF RESEARCH*, 19-21.

Raymond, C., & Jason, O. (2011). *General Chemistry: The Essential Concepts Sixth Edition.* New York: McGraw-Hill .

Sahu, M., Nagwani, N. K., Verma, S., & Shirke, S. (2015). Performance Evaluation of Different Classifier for Eye State Prediction Using EEG Signal. *International Journal of Knowledge Engineering (IJKE)*, 141.

Wang, Q. (2014). A Hybrid Sampling SVM Approach to Imbalanced Data Classification. *Abstract and Applied Analysis*, 1-7.

Witten, I. E. (2005). *Data Mining Practical Machine Learning Tools and Techniques. 2nd Edition.* San Francisco: Morgan Kaufmann Publishers.

Yuliana , Y., & Erlangga, E. (2017). Analysis Of Data Mining Methods Naive Bayes Classifier (NBC). *International Conference on Engineering and Technology Development (ICETD)* (pp. 246-260). Bandar Lampung: Information System, Computer Science Faculty, Bandar Lampung University.

Zhang, H., & Wang, Z. (2011). A normal distribution-based over-sampling approach to imbalanced data classification. *ADMA'11: Proceedings of the 7th international conference on Advanced Data Mining and Applications* (pp. 83–96). Beijing China: ADMA.