http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 10, No.1, Desember 2021 ISSN : 2302-9706

Implementation Of The Naive Bayes Method To Classification Good Air Quality

Zalika Riswan Dini[1], Dedi Purwanto[2], Nova Mayasari[3]
Faculty Sains and Technology, Program Study Computer ,
Universitas Pembangunan Pancabudi Medan
Email: zalikariswandini04@gmail.com

Abstract

Article Info

Received : 30 November 2021
Revised : 19 December 2021
Accepted : 28 December 2021

Naïve Bayesisa- method- of doing statistical classification based on the probability value of the members in the class on a data. This research was conducted by finding several stages of the process carried out, namely the process of sharing training data and test data. Then the-classification-process-is-carried-out by-finding the-accuracy-value-of-the-Naïve-Bayes method. The success of increasing the-accuracy-of-the-dataset-using-the-weighting-of-the Naïve Bayes algorithm which greatly affects all attributes, so that it greatly affects the-accuracy-of-the-data. -In-the-data classification-process-using the Naïve Bayes algorithm that uses the Air Quality dataset, the accuracy results are 39.97%. The author hopes that there will be development of the Naïve Bayes method by comparing it to other weighting methods in order to get various kinds of analysis results so as to get a higher accuracy value and use more datasets.

Keywords: Accuracy, Classification, Naïve Bayes, Air Quality

1. Introduction

According to research (Siagian, 2020) Naïve Bayes is a method of classifying statistics based on the probability value of members in a class on a data. In the research made, the accuracy of the water quality dataset and Haberman's reached 88.57% and 78.95%, respectively.

In research (Niloy, 2018) conducted research on data mining with 2 techniques whose results have differences in method errors. However, it has a large latetive in the ratio. Naïve Bayes has a high level of accuracy compared to the tree method because in classifying the measurement data accuracy is needed.

Meanwhile, according to (Yuliana, 2017) Nave Bayes is a method that is often used in classifying data that uses a lot of data. The researcher uses DSS (Decision Support System) in conducting the experiment because it does not require weights in calculating but uses pre-existing probability values. The researcher uses the Naïve Bayes method because the classification opportunities are very simple and very easy to understand, the calculations are very efficient and the accuracy is higher in the data classification process.

Naive Bayes is a probabilistic classification model that facilitates machine learning by performing computations on data sets aimed at predicting intra-class probabilities with strong independent assumptions. Naive Bayes is one of the classification models. Nave Bayes has lighter probabilities in machine learning by performing computations on data sets aimed at predicting probabilities in classes that assume strong independence.

Naïve Bayes is a machine learning tool for solving probability problems. Nave Bayes is more complex and slows down cycle spread when no training sample increases because when the number of training samples increases Nave Bayes requires more storage. Naïve Bayes was originally used in the classification process on binary numbers and there are formulations of multi-class cases that are

not easy and include existing problems. In Naïve Bayes there are 2 ways to classify multiclasses; the first is to consider all data in one optimization formulation, and the second is to simplify a multi-class problem into several binary problems. Both solutions are initial approximations, because classification on binary numbers is very easy to use and also some powerful Algorithms like Naïve Bayes are inherently binary.

2. Methods

Research methodology is a process that has a structure in carrying out activities in conducting a research. The flow in research that uses the Naïve Bayes method in classifying data can see on charts is below:

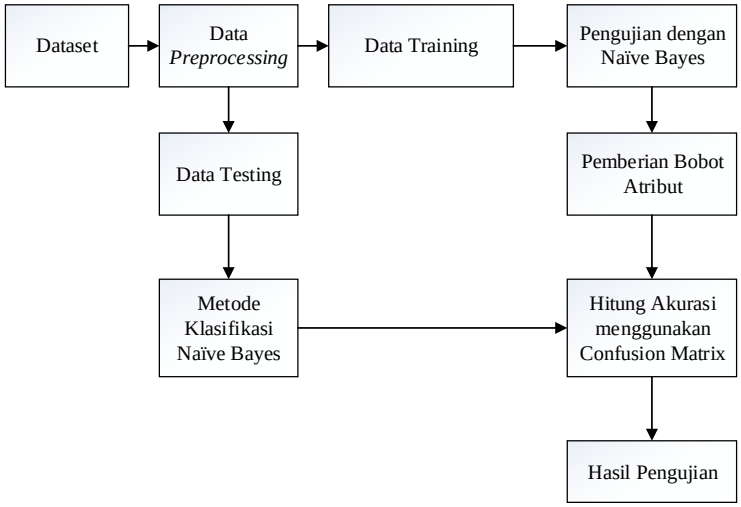


Figure 1. Design Of Research

In the picture above, it can be seen that the stages of the research process are as follows:

The dataset selection process is carried out, the authors use two data sets, namely the Air Quality Dataset. Then the data is processed by classifying the Naïve Bayes method. The next step is to analyze the classification results obtained by testing the classification model using the Confusion Matrix tabulation. This study uses air quality data that can be seen in uci machibe learning. Here is the dataset:

3. Research Results and Discussion

3.1 Research Result

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 second 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 carried out standardized data normalization so that the intervals become more proportional, which can be seen below:

$$z = \frac{x-\mu}{\sigma}$$

z score which has a standard value, x as a dataset used in the analysis process for improving classification performance, is the average value (mean) then is the standard deviation of each data variable. Then the process of Z score is is 0 and is number 1.

In short, the observation data will be reduced by the mean of each variable and divided by the standard deviation. In this process, the first several processes are missing values that replace numeric

values into mean values of similar attributes and columns. Then the second that is process deletion of duplicate data (double). The recorded data used has 13 kinds of attributes, 3 classes and 9357 instances.

Then the third process is data normalization of standardized data so that the interval becomes more proportional and the next step is cleaning the data by removing duplicate (same) data which returns to the number of the initial dataset.

In this process, it is actually not a problem if it is lost in a small amount for the entire data, but the percentage that is lost is small, such as only 1% of the entire dataset. If the missing data reaches a large amount, it is necessary to re test the data whether the data is feasible for further processing or not. The calculation of missing/value can be seen below:

$$\overline{x} = \frac{\sum x}{n} = \frac{x_1+x_2+x_3+\cdots x_n}{n}$$

$\bar{x}$ = Average

x = Data

n = the amount of data

$$\overline{x} = \frac{\sum x}{n} = \frac{x_1+x_2+x_3+\cdots x_n}{n}$$

$$\bar{x} = \frac{\sum x}{n} = \frac{2.6+2+2.2+2.2+1.6+1.2+1.2+1+0.9+0.6+(-200)}{\cdots+2.2} \quad 9357$$

$\bar{x}$ = 34.208

The following are the results of data sharing (Train Test Splitting) using the Python programming language:

```
Rekapitulasi Data Latih
----------------------------------------------------------------
[[-0.86514866 -0.58890849 -0.84343716 ...  1.93954639  0.50924955
  0.19726939]
[-0.86514866 -1.00352517 -0.68235048 ...  0.62871699 -1.5159393
  0.49072555]
[ 0.36857938 -0.1742918  -0.26620988 ... -0.93224234  1.23253128
 -1.27001143]
...
[ 0.80401281  0.52594971  0.90166857 ...  0.96255376  0.07528051
  1.07763788]
[ 2.61831875  2.27194664  1.73394976 ...  0.57398559  0.79856224
 -0.6830991 ]
[-1.01029313 -0.00844513 -0.33332933 ... -0.92704162 -0.50334488
 -1.56346759]]
```

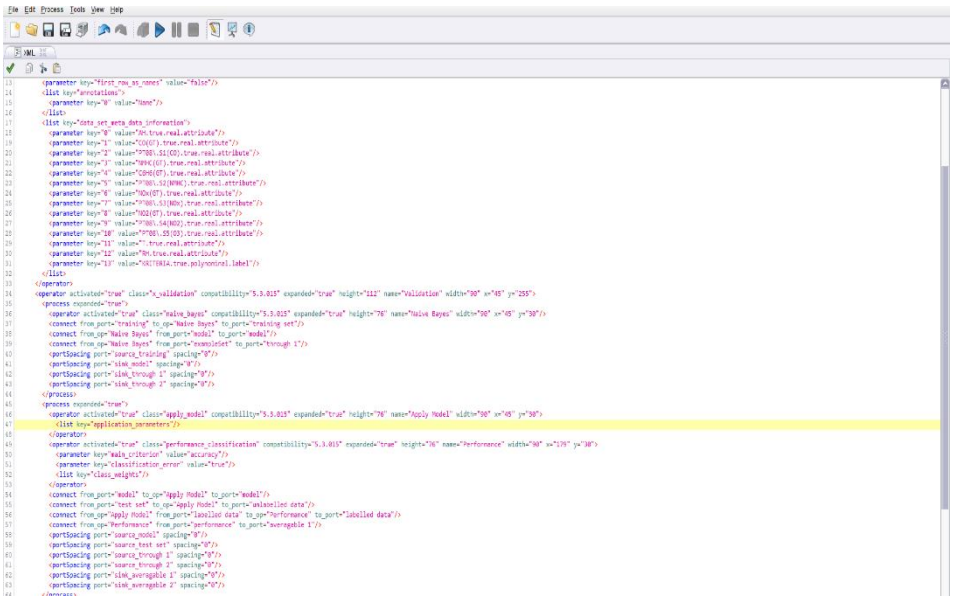
Figures 2. Air Quality Training Data

In the picture above, it is training data from the amount of data of 9326 and the class of 9357 data obtained from 70% of the data that became training data. This training data is used to process data classification in the Naïve Bayes method.

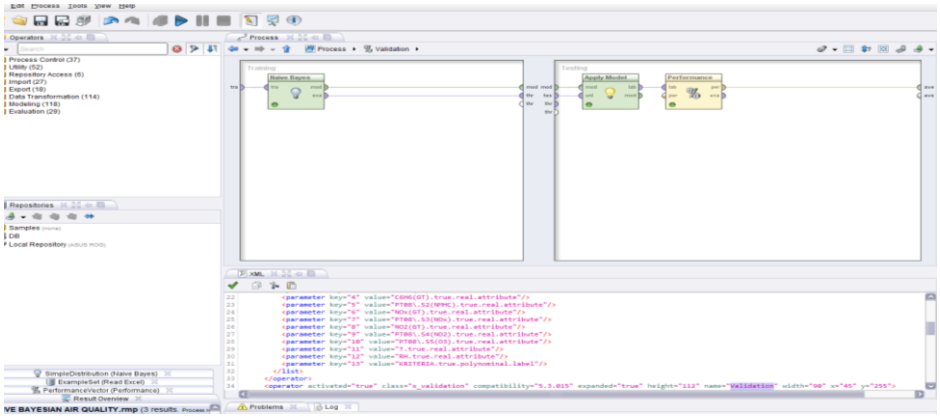
```
Rekapitulasi Data Uji
----------------------------------------------------------------------
[[-0.50228747  0.21268377 -0.19909043 ...  0.2391582   1.52184397
  -0.6830991 ]
 [ 3.12632441  2.34565627  2.6333504  ... -0.09616449  1.08787493
  -0.97655527]
 [-0.71146627  0.08369191 -0.1051232  ...  1.33007139 -1.37128295
   1.07763788]
 ...
 [ 1.02172952  0.88989102  1.04933136 ... -0.92704162  0.07528051
   1.6645502 ]
 [ 0.29600714 -0.10518902 -0.06485153 ... -1.39733541 -0.06937584
   1.6645502 ]
 [ 0.22974554  1.20776382  0.9956358  ...  1.35706561 -0.35868853
   0.19726939]]
```

Figures 3. Test Data

Another experiment was used in other applications using Rapidminer 5.3.0. The results of the accuracy of the application used are as follows:



Figures 4 Source Codes Rapidminer



Figures 5. Results of Data Accuracy using Rapidminer

3.2 Discussion

The results of the research that have been carried out have found that the processes that have been tested and carried out by the author, as well as the process of sharing training data and test data. So then carry out the classification process by finding the accuracy value of the Naïve Bayes method. The results of the Naïve Bayes accuracy on the Air Quality dataset are as follows:

Tables 2. Naïve Bayes . 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%	

The calculations in the confusion matrix are as follows:

a. Accuracy = $\frac{3383+345}{3383+345+5503+95} = \frac{3728}{9326}$ = 0.39974* 100% = **39.97%**

b. Classification_*error*= $\frac{5503+95}{3383+345+5503+95} = \frac{5598}{9326}$= 0.60025*100% = **60.02%**

4. Conclusion

The success of increasing the accuracy of the dataset using the weighting of the Naïve Bayes algorithm which greatly affects all attributes, so that it greatly affects the accuracy of the data. In the process of classifying data using the Naïve Bayes algorithm using the Air Quality dataset, the accuracy result is 39.97%. The author hopes that there will be development of the Naïve Bayes method by comparing it to other weighting methods in order to get various kinds of analysis results so as to get a higher accuracy value and use more datasets.

Reference

1. Adityawarman, Dimas Zebua, O., & Hakim, L. (20116). Design and Build a Current Measuring Tool Using Atmega32 Microcontroller Based Current Transformer. Electrician, *8*(2), 45–56.
2. Andi Aulia Rahman. (2019). *Design And Construction Of Air Quality Monitoring System Using Web-Based Wireless Signal Network (WSN)*. 63.
3. Deptt, E., & Jabalpur, J. E. C. (2013). *CONTROL OF STARTING CURRENT IN THREE PHASE Sharda Patwa*. *01*, 27–32.
4. Jadot, F., Malrait, F., Moreno-Valenzuela, J., & Sepulchre, R. (2009). Adaptive regulation of vector-controlled induction motors. *IEEE Transactions on Control Systems Technology*, *17*(3), 646–657. https://doi.org/10.1109/TCST.2008.2003434
5. Lubis, S. A., Hariyanto, E., Perangin-angin, M. I., Saputra, S., Niska, D. Y., Wahyuni, S., Nasution, D., & Iqbal, M. (2015). *APPLICATION HYBRID ECO CAMPUS VEHICLE BASED ON SOLAR POWER*. *3*(2).
6. Marliani, N. (2014). Utilization of Household Waste (Inorganic Waste) as a Form of Implementation. *Formatif*, *4*(2), 124–132.
7. solly Aryza. (2017). A Novelty Design Of Minimization Of Electrical Losses In A Vector Controlled Induction Machine Drive. *Scopus*, *1*, 20155.