Available *online* at : http://bit.ly/InfoTekJar

InfoTekJar : Jurnal Nasional Informatika dan Teknologi Jaringan

ISSN (Print) 2540-7597 | ISSN (Online) 2540-7600

Comparative Analysis of Accuracy in Identifying Types of Glass

Novriadi Antonius Siagian

University of Sumatera Utara, Medan, Indonesia.

KEYWORDS

Type of Glass, K-NN, C4.5, Naïve Bayes, Accuracy, Error, Confusion Matrix

CORRESPONDENCE

E-mail: novriadi.antonius95@gmail.com

A B S T R A C T

To see whether the proposed research model is able to improve the performance of the classification of the Glass Type Identification data using the K-Nearest Neighbor (K-NN) method then the results will be compared with the C4.5 method and the Naïve Bayes method, a performance analysis of the methods will be carried out. The results are based on the results of the Confusion Matrix tabulation (two-class prediction. In this study, only three preprocessing processes were carried out. The first process is handling missing value. The missing value for attributes with numeric values is replaced by the mean (mean) value of the attributes in the same column. Meanwhile, the missing values for attributes with nominal values are replaced by the most likely values for the attributes in the same column. Then the second process is the handling of duplicated data. The data recorded were 214 data, the number of attributes was 9 attributes and the number of classes was 6 classes.The results of this study show that the highest accuracy value is in the C4.5 method with an accuracy of 73.45% with a value of K = 2 and an error rate of 26.55%, while the method with low accuracy is the KNN method. with an accuracy value of 61.95% and an error rate of 38.05%. Naïve Bayes has an accuracy of 63.33% and an error rate of 36.67. Therefore C4.5 is more effective than the two methods.

INTRODUCTION

The C4.5 algorithm is one of the Decision Tree methods in the classification process using the information entropy concept. The C4.5 algorithm uses the split criteria from ID3, the Gain Ratio is a modification of the method. The ID3 algorithm uses Information Gain (IG) for the split attribute criteria, while the C4.5 algorithm with Gain Ratio (GR), where the root value comes from high gain. The step of the C4.5 algorithm process is by calculating the Entropy value. With each attribute, the Gain Ratio value is calculated, then the attribute that has a high Gain Ratio value will be selected as the root and the low one will become the branch, then recalculate the Gain Ratio value of each attribute by not using the selected attribute as the root of the process. Previously, the next process was carried out to produce a Gain value of 0 on the remaining attributes.

Naïve Bayes is a probability classification model that is easier in machine learning by performing calculations from a dataset that aims to predict probability in a class with the assumption of strong dependability.

Whereas the KNN (K-Nearest Neighbor) K method used in each class has a large effect on the K value. If k is less than the classification that is useful for data is not fulfilled, if the value of k is large it can more easily cause existing outliers. in the neighborhood k which is close to the classroom center.

METHOD

To see whether the proposed research model is able to improve the performance of the classification of the Glass Type Identification data using the K-Nearest Neighbor (K-NN) method then the results will be compared with the C4.5 method and the Naïve Bayes method, a performance analysis of the methods will be carried out. based on the results of the Confusion Matrix tab (two-class prediction.

In this study, only three preprocessing processes were carried out. The first process is handling missing value. The missing value for attributes with numeric values is replaced by the mean (mean) value of the attributes in the same column. Meanwhile, the missing values for attributes with nominal values are replaced by the most likely values for the attributes in the same column. Then the second process is the handling of duplicated data. The data recorded were 214 data, the number of attributes was 9 attributes and the number of classes was 6 classes.

The next process is data normalization carried out by standardizing the data so that the interval or range of data becomes more proportional using the Z-Score method as follows:

$z = (x-\mu) / \sigma$

z: standard score, x: observed data, μ: mean per variable and σ: standard deviation per variable. The result of the Z-score is data with mean = 0 and standard deviation = 1.

Simply put, the Z-scoring process is: each observed data on a variable minus the mean of the variable and divided by the

https://doi.org/10.30743/infotekjar.v5i2.3655

standard deviation (in other words, each row per column minus the column mean, divided by the standard deviation of the same column).

In the process of forming the K-NN, C4.5, and Naïve Bayes classification models the results of preprocessing data, namely cleaning data from the *Kaggle* dataset, obtained as many as 214 observational data then divided into 90% data as training data and 10% data as test data.

RESULTS AND DISCUSSION

In this study, using the Glass Type Identification dataset from Kaggle.com, after carrying out the preprocessing process, the results of the test data will then be tested using the KNN, C4.5, and Naïve Bayes methods based on the Confusion Matrix. The data used are as follows:

Table 1. Data for testing

No.	RI	Na	Mg	Al	…	Type
1	**0.344632**	**0.288878**	**1.254.284**	**-0.70471**	**…**	**1**
2	**0.336591**	**0.595002**	**0.637808**	**-0.18044**	**…**	**1**
3	**0.333209**	**0.154183**	**0.603175**	**0.18252**	**…**	**1**
4	**0.336709**	**-0.23766**	**0.700148**	**-0.32159**	**…**	**1**
5	**0.336142**	**-0.16419**	**0.651662**	**-0.42241**	**…**	**1**
6	**0.332689**	**-0.75194**	**0.644735**	**0.343835**	**…**	**1**
7	**0.336165**	**-0.12745**	**0.637808**	**-0.62405**	**…**	**1**
8	**0.336473**	**-0.31113**	**0.644735**	**-0.80553**	**…**	**1**
9	**0.340304**	**0.778676**	**0.623955**	**-0.16027**	**…**	**1**
…	**…**	**…**	**…**	**…**	**…**	**…**
214	**0.335408**	**101.133**	**-18.558**	**1.271.394**	**…**	**7**

Information Attribute:
RI : refractive index
Na : Sodium
Mg : Magnesium
Al : Aluminum
Si : Silicon
K : Potassium
Ca : Calcium
Ba : Barium
Fe : Iron
Type of glass: (class attribute)
1 : building windows float processed
2 : building windows non float processed
3 : vehicle windows float processed
4 : vehicle windows non float processed (none in this database)
5 : containers
6 : tableware
7 : headlamps

After obtaining preprocessing, the classification model is then tested using the test data dataset on the Identification of Glass Types. The test was carried out with the K-Nearest Neighbors classification model. When the K = 2 value, there were 214 instances. In order for the expected accuracy results to be more accurate, data partitioning is carried out using the K-Fold Cross Validation method. K-fold is one of the popular Cross Validation methods by folding K as much data and repeating (iterating) the experiment as much as K as well. Then, to see the error ratio of each K value, an iteration is carried out to calculate the error-rate to produce the optimal K value when testing the K-Nearest Neighbors classification model. The following is the output of the test:

Table 2. Confusion Matrix K-NN

Actual	Predicted Class 1	Predicted Class 2	Predicted Class 3	Predicted Class 5	Predicted Class 6	Predicted Class 7
Actual Class 1	32	16	7	0	1	1
Actual Class 2	5	21	2	2	2	2
Actual Class 3	0	0	0	0	0	0
Actual Class 5	0	2	0	4	1	0
Actual Class 6	0	0	0	0	1	0
Actual Class 7	0	1	0	1	0	12

$$\text{Accuracy} = \frac{32+21+0+4+1+12}{32+21+0+4+1+12+16+7+1+1+5+2+2+2+2+1+1+1} = \frac{70}{113} = 0.61946*100\% = 61.95\%$$

The level of closeness between class predictions and actual class or the number of correct class predictions from the KNN classification model is 61.95%. While the results of the Classification Error are as follows:

$$\text{Classification\_Error} = \frac{16+7+1+1+5+2+2+2+2+1+1+1}{32+21+0+4+1+12+16+7+1+1+5+2+2+2+2+1+1+1} = \frac{43}{113} = 0.3805 * 100\% = 38.05\%$$

Then testing the C4.5 classification model. The following is the output of the test:

Table 3. Confusion Matrix C4.5

Actual	Predicted Class 1	Predicted Class 2	Predicted Class 3	Predicted Class 5	Predicted Class 6	Predicted Class 7
Actual Class 1	29	8	5	0	0	0
Actual Class 2	4	29	2	0	0	0
Actual Class 3	0	0	0	0	0	0
Actual Class 5	0	6	0	8	0	1
Actual Class 6	0	2	0	0	5	0
Actual Class 7	0	1	0	1	0	12

$$\text{Accuracy} = \frac{29+29+0+8+5+12}{29+29+0+8+5+12+8+5+4+2+6+1+2+1+1} = \frac{83}{113} = 0.7345*100\% = 73.45\%$$

The level of closeness between class predictions and actual class or the number of correct class predictions from the KNN classification model is 73.45%. While the results of the Classification Error are as follows:

Classification Error= $\frac{8+5+4+2+6+1+2+1+1}{29+29+0+8+5+12+8+5+4+2+6+1+2+1+1} = \frac{29}{113} = 0.2566 * 100\% = 25.66\%$

Then testing the Naïve Bayes classification model. The following is the output of the test:

Table 4. Confusion Matrix Naïve Bayes

Accuracy=

Actual	Predicted Class 1	Predicted Class 2	Predicted Class 3	Predicted Class 5	Predicted Class 6	Predicted Class 7
Actual Class 1	55	31	13	0	1	1
Actual Class 2	9	40	4	3	0	1
Actual Class 3	5	0	0	0	0	0
Actual Class 5	0	2	0	9	0	3
Actual Class 6	0	2	0	0	7	0
Actual Class 7	0	1	0	1	1	24

$\frac{55+40++9+7+24}{55+40+9+7+24+31+13+1+1+9+4+3+1+5+2+3+2+1+1+1} = \frac{235}{213} = 0.7345*100\% = 73.45\%$

The level of closeness between class predictions and actual class or the number of correct class predictions from the KNN classification model is 73.45%. While the results of the Classification Error are as follows:

Classification_Error=

$\frac{31+13+1+1+9+4+3+1+5+2+3+2+1+1+1}{55+40+9+7+24+31+13+1+1+9+4+3+1+5+2+3+2+1+1+1} = \frac{78}{213} = 0.3661 * 100\% = 36.61\%$

Table 5. Accuracy Comparison Results

RESULT		
METHOD	ACCURACY	CLASSIFICATION ERROR
KNN	61.95%	38.05%
C4.5	73.45%	26.55%
NAÏVE BAYES	63.33%	36.67%

The comparison chart can be seen in the following image:

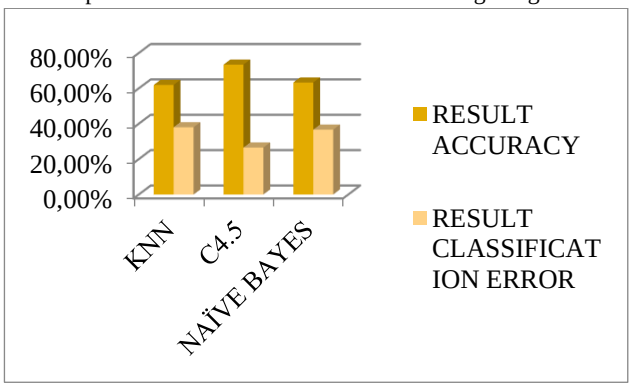

Figure 1. Comparison Result Graph

CONCLUSIONS

The results of this study indicate that the highest accuracy value is found in the C4.5 method with an accuracy of 73.45% with a value of K = 2 and an error rate of 26.55%, while the method with low accuracy is the KNN method with an accuracy value of 61.95% and an error rate of 38.05%. Naïve Bayes has an accuracy of 63.33% and an error rate of 36.67. Therefore, C4.5 is more effective than the two methods.

ACKNOWLEDGMENT

The writer hopes that the next researcher can add another method in the comparison, or the next researcher can modify the method by combining it with other attribute features in order to get higher accuracy from the research that I did. Researchers also hope that the dataset can also be analyzed using other methods to make it more developed in computer science.

REFERENCES

[1] Arifin, Toni. 2015. *Implementasi Metode K-Nearest Neighbor untukKlasifikasi Citra Sel PAP Smear menggunakan Analisis Tekstur Nukleus*. Jurnal Informatika. Volume. II. Pp. 1-4. ISSN : 2355-6579.

[2] Dai Qin-yun,. Zang Chun-Ping., Wu Hao. 2016. *Research of Decision tree Classification Algorithm in Data Mining*. Dept. of Electric and Electronic Engineering, Shijiazhuang Vocational and Technology Institute. China

[3] Danades, A., Pratama, D., Anggraini, D., Anggriani, D. 2016. Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status. *International*

https://doi.org/10.30743/infotekjar.v5i2.3655

Conference on System Engineering and Technology, pp. 137-141.

[4] Fadillah, Annisa Pulungan. 2018. AnalisisKinerja Bray Curtis Distance dan Canberra Distance Pada Algoritma K-Nearest Neighbor. Tesis. UniversitasSumatera Utara.

[5] Han, J., Kamber, M. & Pei, J. 2012.*Data Mining: Concepts and Techniques*. 3rd Edition. Morgan Kaufmann Publishers: San Francisco.

[6] NH Niloy, MAI Navid. *Naïve Bayesian Classifier and Classification Trees for the Predictive Accuracy of Probability of Default Credit Card Clients*. Department of Science, Ruhea College, Rangpur, Bangladesh

[7] Pattekari, S. A., Parveen, A., 2012. *Prediction System for Heart Disease Using Naive Bayes, International Journal of Advanced Computer and Mathematical Sciences*, ISSN 2230-9624, Vol. 3, No 3, Hal 290-294

[8] Raviya. Kaushik H & Gajjar, Biren. 2013. *Performance Evaluation of Different Data Mining Classification Algoritma Using WEKA. Indian Journal of Research*. Volume. 2. Issue.1. ISSN: 2250-1991.