LONTAR KOMPUTER VOL. 9, NO. 2, AUGUST 2018 **p-ISSN 2088-1541**
DOI : 10.24843/LKJITI.2018.v09.i02.p05 **e-ISSN 2541-5832**
Accredited B by RISTEKDIKTI Decree No. 51/E/KPT/2017

Implementation of Equal-Width Interval Discretization in Naive Bayes Method for Increasing Accuracy of Students' Majors Prediction

Alfa Saleh[a1], Fina Nasari[a2]

[a]Faculty of Computer Science and Engineering, Potensi Utama University
Jl. K.L Yos Sudarso KM 6.5 Tanjung Mulia Medan, Indonesia
[1]alfa@potensi-utama.ac.id
[2]fina@potensi-utama.ac.id

Abstract

The Selection of majors for students is a positive step that is done to focus students in accordance with their potential, it is considered important because with the majors, students are expected to develop academic ability according to the field of interest. In previous research, Naive Bayes method has been tested to classify the student's department based on the criteria that support the case study on Private Madrasah Aliyah PAB 6 Helvetia students and the accuracy of the test from 100 student data is 90%. in this study, the researcher developed a previously used method by applying an equal-width interval discretization that would transform numerical or continuous criteria into a categorical criteria with a predetermined k value, different k values would be tested to find the best accuracy value. from the 120-student data that have been tested, it is proved that the result of the classification of the application of equal-width interval discretization on the Naive Bayes method with the value of k = 8 is better and increased the accuracy value 91.7% to 93.3%.

Keywords: *Data Mining, Naive Bayes, Equal-Width Interval Discretization, Students' Majors*

1. INTRODUCTION

The role of education is very important in supporting the development of technology that almost has penetrated into all areas. It also affects the determination of majors for high school / equivalent students, where the determination of the student's department is a process to focus students in a particular area of the interested field, this is done so that each student can learn more in the subjects that are in accordance with the concentration which has been specified for the student. The problem is the ongoing system of private school Madrasah Aliyah PAB 2 Helvetia Medan, the place where researchers conduct research is not entirely effective because students are given a questionnaire to determine which majors they are interested in regardless of other criteria that may have a stake in determining eligibility students in terms of choosing majors. Through the process of determining the majors for students is an important step in preparing students to concentrate on the field that students are interested in when it should continue to the next education level. In the previous research, researchers also have done the process of mining to dig information about the determination of student majors using Naive Bayes method, the results of the research were tested 100 student data based on several criteria include the average score of natural science subjects, the average value of science social, classroom teacher recommendation and the questionnaire value filled by the students concerned. from the 100 data tested using the Naive Bayes method, it is obtained the accuracy value of determining student majors by 90% with an error of 10% [1] . The Naive Bayes method was chosen because it was widely implemented in various fields of science, as in the Xingxing Zhou research (2016), the Naive Bayes method was used to classify images to improve the accuracy of brain diagnosis using NMR imagery, where 94.5% sensitivity classification was obtained, 91.70% and the overall accuracy of 92.60 [2]. Naive Bayes is one of the top ten (10) data mining algorithms for simplicity and efficiency, as evidenced by the performance of Naive Bayes in classifying text [3], [4]. In addition, Naive Bayes is widely recognized as a simple and effective probabilistic classification method [5]–[7], and its performance is proportional to or higher than the decision tree [8] and artificial neural networks [9].

However, researchers wanted to expand their previous research by applying Unsupervised Discretization [10] to improve the performance of the Naive Bayes method so that the percentage of predicted accuracy results could increase compared to the previous one. Where Unsupervised Discretization techniques in transforming numerical criteria / attributes are excellent [11].

2. Research Methods

2.1. Naïve Bayes

Naive Bayes is a model-based classification method and offers competitive classification performance compared with other data-driven classification methods [12]–[15], such as neural network, support vector machine (SVM), logistic regression, and k-nearest neighbors. The naive Bayes applies the Bayes' theorem with the "naive" assumption that any pair of features is independent for a given class. The classification decision is made based upon the maximum-a-posteriori (MAP) rule. Usually, three distribution models, including Bernoulli model, multinomial model and Poisson model, have commonly been incorporated into the Bayesian framework and have resulted in classifiers of Bernoulli naive Bayes (BNB), multinomial naive Bayes (MNB) and Poisson naive Bayes (PNB), respectively[4]. The formula of Bayes's theorem is [16]:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Where variable X represents Data with unknown class, H represents The data hypothesis is a specific class, P (H|X) represents The probability of hypothesis H is based on condition X (posterior probability), P (H) represents Hypothesis probability H (prior probability), while P (X|H) represents The probability of X is based on the conditions in hypothesis H and P (X) represents Probability X. Therefore, the method of Naive Bayes above is adjusted as follows:

$$P(C|F1 \dots Fn) = \frac{P(C)P(F1 \dots Fn|C)}{P(F1 \dots Fn)} \quad (2)$$

Where Variable C represents the class, while the F1 ... Fn represents the characteristics of the user for the classification process. Therefore, the above formula can also be written simply as follows:

$$Posterior = \frac{prior\ x\ likelihood}{evidence} \quad (3)$$

2.2. Unsupervised Discretization

Discretization is the process of converting a continuous attribute value into a limited number of intervals and associated with each interval with a discrete numerical value. Discretization process is carried out before the learning process [17]. Among the methods of Unsupervised Discretization, there are several simple methods. (Equal-width Interval Discretization and equal-frequency Interval Discretization) and more sophisticated, based on clustering analysis, such as k-means discretization. The Continuous range is divided into subranges by user-specified width or Frequency[18]. But in this study, researchers used Equal-width interval Discretization technique, which is the simplest discretization method that divides the observed range of values in each feature / attribute. The process involves sorting the observed values of the continuous feature / attribute and finding the minimum (Vmin) and maximum (Vmax) values. The interval can be calculated by dividing the observed range of values for the variables into k of the same size using the following formula [18].

$$Interval = \frac{V_{max} - V_{min}}{k} \quad (4)$$

$$Boundaries = V_{min} + (i\ x\ Interval) \quad (5)$$

Then the limits can be constructed for i = 1 ... k-1 using the above equation. This type of discretization does not depend on multi-relational data structures. However, this discretization method is sensitive to outliers that can drastically reduce the range. The limitations of this method are given by the uneven distribution of data points: some intervals may contain more data points than others.

2.3. Research Stages

In the Naïve Bayes method, the constant (categorical) String data is distinguished from continuous numerical data, this difference will be seen when determining the probability value of each criterion whether it is a criterion with a string data value or a criterion with a numeric data value. The stages of applying the method of Naive Bayes in this study can be seen in Figure 1 below.

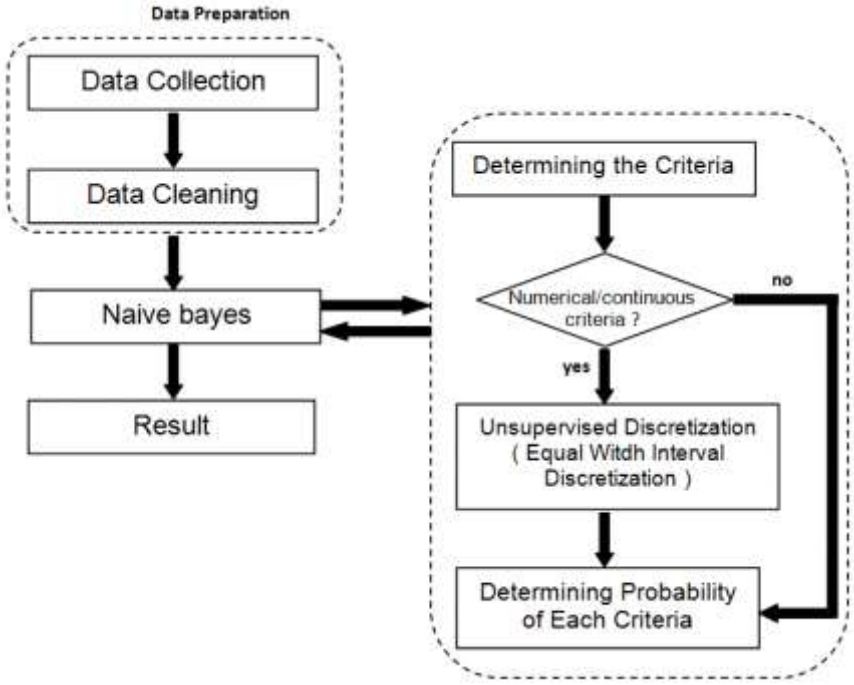

Figure 1. Research Stages of Equal-Width Interval Discretization on Naive Bayes

2.3.1. Data Collection

The data that will be used as training data is the academic data of the students as respondents, where the sample of student data is taken as much as 120 data, they consist of The students' academic data such as the score of Mathematics, Physics, Chemistry, Biology, Economics, Geography, History and Sociology ,the questionnaire that is filled by students and recommendation from the homeroom.

2.3.2. Data Cleaning

In the process of data cleaning, the data that eventually used in this research is the exact value of subjects, non-exact subjects, a recommendation from the homeroom, and questionnaires filled by students.

2.3.3. Determining the Criteria

The criteria that used based on data that has been collected is as in table 1 below:

Table 1. Criteria

NO	Criterion	Type of Criterion	Value
1	The average score of exact subjects	Numerical/Continuous	0 - 100
2	The average score of non-exact subjects	Numerical/Continuous	0 - 100
3	Recommendation	Categorical	Science, Social Studies
4	Questionnaire	Categorical	Science, Social Studies

There are four (4) criteria used in this research, namely the average score of exact subjects, the average value of non-exact subjects, recommendation and lift. Two (2) of them are numerical / continuous criteria and two (2) categorical criteria. To improve the accuracy of the Naive Bayes method, discretization is performed using unsupervised discretization techniques on numerical / continuous criteria, the goal is to transform numerical/continuous criteria into categorical criteria using formulas 4 and 5. The following table 2 discriminates numerical criteria / continuous.

Table 2. The results of Discretization with k=8

Numerical/Continuous Criteria	
The average score of exact subjects	The average score of non-exact subjects
<71.9125	<71.875
71.9125 – 73.825	71.875 – 73.75
73.825 – 75.7375	73.75 – 75.625
75.7375 – 77.65	75.625 – 77.5
77.65 – 79.5625	77.5 – 79.375
79.5625 – 81.475	79.375 – 81.25
81.475 – 83.3875	81.25 – 83.125
83.3875>	83.125>

In table 2 above, you can see the results of the discretization process using the Unsupervised Discretization technique. Where the criteria / attributes of The average values of exact and non-exact subjects with numerical or continuous type are transformed into categorical criteria with 8 categories. The first category is the average value of exact sciences that are below 71.9125, the second category is the average value of exact subjects which are between 71.9125-73.825, the third category is the average value of exact subjects which are between 73.825- 75.7375, the fourth category is the average value of exact subjects that are between 75.7375-77.65, the fifth category is the average value of exact subjects that are between 77.65-79.5625, the sixth category is the average value of exact subjects that are between 79.5625-81.475, the seventh category is the average value of exact subjects which are between 81.475-83.3875, and the eighth category is the average value of exact sciences that are above 83.3875.

Furthermore, the results of the discretization of the criteria for the average value of non-exact subjects are also divided into 8 categories, where the first category is the average value of non-exact subjects under 71,875, the second category is the average value of non-exact subjects - acts that are between 71,875-73,75, the third category is the average value of non-exact subjects that are between 73.75-75.625, the fourth category is the average value of non-exact subjects that are between 75.625-77.5, the fifth category is the average value of non-exact subjects that are between 77.5-79.375, the sixth category is the average value of non-exact subjects that are between 79.375-81.25, the seventh category is the average value of non-exact subjects between 81.25-83.125, and the eighth category are the average values of non-exact subjects above 83.125.

2.3.4. The Probability of Each Criterion

Several criteria have been set as a reference in classifying students' majors using Unsupervised Discretization techniques on the Naive Bayes method. The next step, determining the probability value of each criterion, for example, the probability value of the average scores of the exact scores of subjects to be shown is the probability value with the value k = 8.

Here the value of probability criteria of the average value of the exact sciences can be seen in table 3.

Table 3. The Probability of The average score of exact subjects with k=8

The Average Score of Exact Subject	Probability	
	Science	Social Studies
<71.9125	0.067	0.283
71.9125 – 73.825	0.05	0.133
73.825 – 75.7375	0.2	0.2
75.7375 – 77.65	0.017	0.05
77.65 – 79.5625	0.033	0.033
79.5625 – 81.475	0.217	0.15
81.475 – 83.3875	0.133	0.1
83.3875>	0.283	0.050

from table 3 above, there were 60 students placed in the science studies major and 60 students were placed in the social studies major . Based on these data, there were 4 students with the average value of exact subjects below 71.9125 placed in the science studies major and the probability value of 0.067, 3 student with an average value of exact subjects between 71.9125-73.825 placed in the science studies major and the probability value of 0.05 , 12 students with the average value of exact subjects between 73.825-75.7375 are placed in the science studies major and the probability value is 0.2, 1 student with an average value of exact subjects between 75.7375-77.65 is placed in the science studies major and the probability value is 0.017, 2 students with the average value of exact subjects between 77.65-79.5625 are placed in the science studies major and the probability value is 0.033, 13 students with the the average value of exact subjects between 79.5625-81.475 are placed in the science studies major and the probability value is 0.217, 8 students with the average value of exact subjects between 81,475-83.3875 is placed in the science studies major and the probability value is 0.133, 17 students with the average value of exact subjects above 83.3875 are placed in the science studies major and the probability value is 0.283. Meanwhile, there were 17 students with the average value of exact subjects below 71.9125 placed at the social studies major and the probability value was 0.283, 8 students with the average value of exact subjects between 71.9125-73.825 were placed in the social studies major and the probability value was 0.133, 12 students with the average value of exact subjects between 73.825-75.7375 were placed in the social studies major and the probability value was 0.2, 3 students with the average value of exact subjects between 75.7375-77.65 were placed in the social studies major and the probability value was 0.05, 2 students the average value of exact subjects between 77.65-79.5625 are placed in the social studies major and the probability value is 0.033, 9 students with the average value of exact subjects between 79.5625-81.475 are placed in the social studies major and the probability value is 0.15, 6 students with the average value of exact subjects is between 81,475-8 3.3875 is placed at the social studies major and the probability value is 0.1, 3 students with an average value of exact subjects above 83.3875 are placed at the social studies major and the probability value is 0.05.

The probability value of the average score of non-exact subjects with a value of k = 8, be shown in table 4 as follows.

Table 4. The Probability of The average score of non-exact subjects with k=8

The average score of non-exact subjects	Probability	
	Science	Social Studies
<71.875	0.3	0.05
71.875 – 73.75	0.167	0.1
73.75 – 75.625	0.15	0.25
75.625 – 77.5	0.033	0.017
77.5 – 79.375	0	0.067
79.375 – 81.25	0.167	0.183
81.25 – 83.125	0.133	0.167
83.125>	0.05	0.167

from table 4 above, there were 60 students placed in the science studies major and 60 students were placed in the social studies major. Based on these data, there were 18 students with the average value of non-exact subjects below 71.9125 placed in the science studies major and the probability value of 0.3, 10 student with an average value of non-exact subjects between 71.9125-73.825 placed in the science studies major and the probability value of 0.167, 9 students with the average value of non-exact subjects between 73.825-75.7375 are placed in the science studies major and the probability value is 0.15, 2 student with an average value of non-exact subjects between 75.7375-77.65 is placed in the science studies major and the probability value is 0.033, there is no student with the average value of non-exact subjects between 77.65-79.5625 are placed in the science studies major and the probability value is 0, 10 students with the the average value of non-exact subjects between 79.5625-81.475 are placed in the science studies major and the probability value is 0.167, 8 students with the average value of non-exact subjects between 81,475-83.3875 is placed in the science studies major and the probability value is 0.133, 3 students with the average value of non-exact subjects above 83.3875 are placed in the science studies major and the probability value is 0.05. Meanwhile, there were 3 students with the average value of non-exact subjects below 71.9125 placed at the social studies major and the probability value was 0.05, 6 students with the average value of non-exact subjects between 71.9125-73.825 were placed in the social studies major and the probability value was 0.1 , 15 students with the average value of non-exact subjects between 73.825-75.7375 were placed in the social studies major and the probability value was 0.25, 1 students with the average value of non-exact subjects between 75.7375-77.65 were placed in the social studies major and the probability value was 0.033, 4 students the average value of non-exact subjects between 77.65-79.5625 are placed in the social studies major and the probability value is 0.067, 11 students with the average value of non-exact subjects between 79.5625-81.475 are placed in the social studies major and the probability value is 0.183, 10 students with the average value of non-exact subjects is between 81,475-8 3.3875 is placed at the social studies major and the probability value is 0.167, 10 students with an average value of non-exact subjects above 83.3875 are placed at the social studies major and the probability value is 0.167.

The probability value for the recommendation criteria can be seen in table 5.

Table 5. The Probability of the recommendation criteria with k=8

Recommendation	Probability	
	Science	Social Studies
Science	0.967	0.15
Social Studies	0.033	0.85

The number of students used was 120 students who had been recommended by the previous homeroom teacher, there were 60 students were placed in the science studies major and 60 students were placed in the social studies major. Based on these data there were 59 students who were recommended to enter the science studies major and placed in the science studies major, while there was 1 student who was recommended to enter the social studies major but was placed in the science studies major. Furthermore, there were 9 students who were recommended to enter the science studies major but were placed at the social studies major while there were 51 students who were recommended to enter the social studies major and placed at the social studies major. Thus, the probability of students who are recommended to enter the science studies major and be placed in the science studies major is 0.967 while the probability of students who are recommended to enter the social studies major but is placed at the science studies major is 0.033. While the probability of students who were recommended to enter the science studies major but placed in the social studies major was 0.15. then, the probability of students being recommended to enter the social studies major and placed in the social studies major was 0.85. The probability value for the questionnaire criteria can be seen in table 6.

The probability value for the Questionnaire criteria can be seen in table 6.

Table 6. The Probability of the Questionnaire criteria with k=8

Questionnaire	Probability	
	Science	Social Studies
Science	0.833	0.15
Social Studies	0.167	0.85

The number of students used was 120 students who had been given questionnaires, it was recorded as many as 60 students were placed in the science studies majors and 60 more students were placed in the social studies major. Based on these data there were 50 students who chose the science studies major and were placed in the science studies majors, while there were 10 students who chose the social studies major but were placed in the science studies major. Then there were 9 students who chose the science studies major but were placed in the social studies majors while there were 51 students who chose the social studies major and were placed in the social studies major. Thus the probability of students who choose the science studies major can be calculated and placed at the science studies major of 0.833, the probability of students who choose the social studies major but placed in the science studies majors is 0.167. Whereas, the probability of students who choose the science studies major but placed at the social studies major is 0.15 while the probability of students who choose the social studies major and placed at the social studies major is 0.85.

3. Result and Discussion

To see the consistency of the use of equal-width interval discretization in the Naive Bayes method, it was tested for some data, The following test of the implementation of unsupervised discretization on The Naive Bayes method by using sample 60 data can be seen in table 7.

Table 7. Testing Results with 60 data

Amount of 'K' value	Weighted Average				
	TP Rate	FP Rate	Precision	Recall	F-Measure
4	0.917	0.082	0.917	0.917	0.917
6	0.917	0.082	0.917	0.917	0.917
8	0.933	0.067	0.933	0.933	0.933
10	0.967	0.033	0.967	0.967	0.967

From the test results using 60 sample data, the application of equal-width interval discretization technique on the Naive Bayes method with the value of k = 4 successfully classify the data with

the accuracy of 91.7%, while for the value k = 6, obtained a level of accuracy of 91.7%, then for value k = 8, the obtained accuracy of 93.3% and for the value k = 10, the accuracy rate obtained is 0.967%. meanwhile, testing is also done with 90 data, the test result can be seen in table 8 below.

Table 8. Testing Results with 90 data

Amount of 'K' value	Weighted Average				
	TP Rate	FP Rate	Precision	Recall	F-Measure
4	0.9	0.1	0.9	0.9	0.9
6	0.922	0.078	0.922	0.922	0.922
8	0.933	0.067	0.933	0.933	0.933
10	0.889	0.111	0.889	0.889	0.889

From the test result using 90 sample data, the application of equal-width interval discretization technique on Naive Bayes method with k = 4 value succeeded in classifying the data with 90% accuracy, while for k = 6, the accuracy level was 92.5%, then the value k = 8, the accuracy of 93.3% and k = 10, the accuracy of 9.25%. meanwhile, testing is also done with 120 data, the test result can be seen in table 9 below.

Table 9. Testing Results with 120 data

Amount of 'K' value	Weighted Average				
	TP Rate	FP Rate	Precision	Recall	F-Measure
4	0.9	0.1	0.9	0.9	0.9
6	0.925	0.075	0.925	0.925	0.925
8	0.933	0.067	0.933	0.933	0.933
10	0.925	0.075	0.925	0.925	0.925

The test result using 120 sample data, the application of equal-width interval discretization technique on Naive Bayes method with value k = 4 succeeded in classifying data with 90% accuracy, while for k = 6, the accuracy level was 92.2%, then for the value k = 8,
the accuracy of 93.3% and k = 10, the accuracy of 88.9%.

The graph of the test results with some previous data can be seen in Figure 2 below:

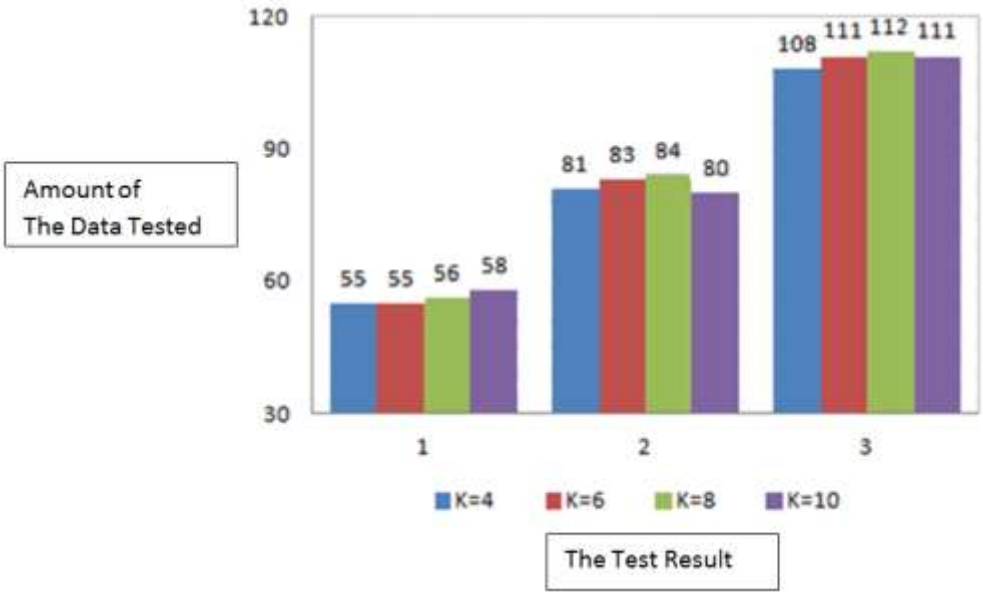

Figure 2. The test results of Unsupervised Discretization Implementation on the Naive Bayes method

from the figure 2 above can be seen the results of testing the application of equal-width interval discretization on the Naive Bayes method in predicting the suitability of students' majors. In the test with 60 sample data, the accuracy value of k = 10 was the best result with 58 successfully classified data correctly. Furthermore, in the test with 90 sample data, the best classification result is owned by the value of k = 8 with 84 data successfully classified correctly, and the last test with 120 sample data, got the best result at value k = 8 where there are 112 data successfully classified with correct.

4. Conclusion

The conclusion that can be summarized in this study is the application of Unsupervised Discretization on the Naive Bayes method has quite an impact on the test results, where the criteria used for this test are: data on the average value of exact courses, data on the average value of non-exact courses, recommendation data and student questionnaire data. And the application of Unsupervised Discretization especially equal-width discretization to Naive Bayes method in predicting the suitability of the student majors increased from the result of accuracy in the previous study by 90% to 93.3%.

5. Acknowledgments

Researchers would like to thank the Ministry of Research and Technology Higher Education Republic of Indonesia (KEMENRISTEKDIKTI) which has helped this research morally and financially.

References

[1] A. Saleh, “KLASIFIKASI METODE NAIVE BAYES DALAM DATA MINING UNTUK MENENTUKAN KONSENTRASI SISWA (STUDI KASUS DI MAS PAB 2 MEDAN),” in *Konferensi Nasional Pengembangan Teknologi Informasi dan Komunikasi (KeTIK) 2014*, 2014, pp. 200–207.

[2] X. Zhou, S. Wang, W. Xu, G. Ji, P. Phillips, P. Sun, and Y. Zhang, “Detection of Pathological Brain in MRI Scanning Based on Wavelet-Entropy and Naive Bayes Classifier,” Springer, Cham, 2015, pp. 201–209.

[3] L. Jiang, C. Li, S. Wang, and L. Zhang, “Deep feature weighting for naive Bayes and its application to text classification,” *Engineering Application of Artificial Inteligence*, vol. 52, pp. 26–39, Jun. 2016.

[4] B. Tang, S. Kay, and H. He, “Toward Optimal Feature Selection in Naive Bayes for Text Categorization,” Feb. 2016.

[5] A. M. P. and D. S. R., “A sequential naïve Bayes classifier for DNA barcodes,” *Stat. Appl. Genet. Mol. Biol.*, vol. 13, no. 4, pp. 1–12, 2014.

[6] J. Wu, S. Pan, X. Zhu, Z. Cai, P. Zhang, and C. Zhang, “Self-adaptive attribute weighting for Naive Bayes classification,” *Expert Systems With Application*, vol. 42, no. 3, pp. 1487–1502, Feb. 2015.

[7] N. Mohamad, N. Jusoh, Z. Htike, and S. Win, “Bacteria identification from microscopic morphology using naive bayes,” *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, vol. 4, no. 2, pp. 1–9, 2014.

[8] Y. Zhang, S. Wang, P. Phillips, and G. Ji, “Binary PSO with mutation operator for feature selection using decision tree applied to spam detection,” *Knowledge-Based Systems*, vol. 64, pp. 22–31, Jul. 2014.

[9] S. Kotsiantis, “Integrating Global and Local Application of Naive Bayes Classifier.,” *International Arab Journal of Information Technology*, vol. 11, no. 3, pp. 300–307, 2014.

[10] S. Palaniappan and T. Kim Hong, “Discretization of continuous valued dimensions in OLAP data cubes,” *International Journal of Computer Science and Network Security*, vol. 8, no. 11, pp. 116–126, 2008.

[11] I. Kareem and M. Duaimi, “Improved accuracy for decision tree algorithm based on unsupervised discretization,” *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 6, pp. 176–183, 2014.

[12] G. Forman, “An Extensive Empirical Study of Feature Selection Metrics for Text Classification,” *The Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1289–1305,

2003.
[13] Y. Yang and J. Pedersen, “A comparative study on feature selection in text categorization,” in *14th International Conference on Machine Learning*, 1997, pp. 412–420.
[14] A. Genkin, D. D. Lewis, and D. Madigan, “Large-Scale Bayesian Logistic Regression for Text Categorization,” *Technometrics*, vol. 49, no. 3, pp. 291–304, Aug. 2007.
[15] B. Tang and H. He, “ENN: Extended Nearest Neighbor Method for Pattern Recognition [Research Frontier],” *IEEE Computational Intelligence Magazine*, vol. 10, no. 3, pp. 52–60, Aug. 2015.
[16] A. Saleh, “Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga,” *Creat. Inf. Technol. J.*, vol. 2, no. 3, pp. 207–217, 2015.
[17] A. Al-Ibrahim, “Discretization of Continuous Attributes in Supervised Learning algorithms,” *Res. Bull. Jordan ACM*, vol. 2, no. 4, pp. 158–166, 2011.
[18] D. Joiţa, “Unsupervised static discretization methods in data mining,” Titu Maiorescu University, 2010.