Comparison of Classification Data Mining C4.5 and Naïve Bayes Algorithms of EDM Dataset

Joseph Teguh Santoso [1], Ni Luh Wiwik Sri Rahayu Ginantra [2], Muhammad Arifin [3], R Riinawati [4], Dadang Sudrajat [5], Robbi Rahim [6]

[1]*Universitas STEKOM, Semarang, Indonesia*
[2]*STMIK STIKOM Indonesia, Denpasar, Indonesia*
[3]*Universitas Muria Kudus, Kudus, Indonesia*
[4]*Universitas Islam Negeri Antasari, Banjarmasin, Indonesia*
[5]*STMIK IKMI Cirebon, Cirebon, Indonesia*
[6]*Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia*

Abstract **- The purpose of this research is to choose the best method by comparing two classification methods of data mining C4.5 and Naïve Bayes on Educational Data Mining, in which the data used is student graduation data consisting of 79 records. Both methods are tested for validation with 10-ford X Validation and perform a T-Test difference test to produce a table that contains the best method ranking. Different results were obtained for each method. Based on the results of these two methods, it is very influential on the dataset and the value of the area under curve in the Naïve Bayes method is better than the C4.5 method in various datasets. Comparison of the method with the 10-Ford X Validation test and the T-Test difference test is that the Naïve Bayes method is better than C4.5 with an average accuracy value of 73.41% and an under-curve area of 0.664.**

Keywords: **Comparison, data mining, Classification, C4.5, Naive Bayes, Performance, EDM.**

DOI: 10.18421/TEM104-34
https://doi.org/10.18421/TEM104-34

Corresponding author: Robbi Rahim,
Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia.
Email: usurobbi85@zoho.com

Received: 04 July 2021.
Revised: 26 October 2021.
Accepted: 30 October 2021.
Published: 26 November 2021.

© 2021 Joseph Teguh Santoso et al; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at www.temjournal.com

1. Introduction

Making comparisons with several data mining classification methods to find out the best results is one way to get the best classification and prediction results [1], [2], [3], [4]. Some of the most widely used classification methods in classifying and predicting are the Decision Tree method (C4.5), Naive Bayes and KNN [5], [6], [7]. There is related research that has been done in determining the best classification method with various existing problems. The results obtained are also different for each problem [8]. Research [9] is done on the prediction of lung cancer survival. This paper proposes two classification techniques, namely the C4.5 algorithm and the Naive Bayes algorithm. The aim of the project is to verify the effectiveness of the predictions of the two techniques on actual historical data. The results show that the C4.5 algorithm is slightly better than Naive Bayes.

Furthermore, research by [10] on evaluating the classification accuracy of the C4.5 and Naïve Bayes methods using the Ljubljana Breast Cancer dataset. The paper proposes an evaluation of the accuracy of two classification methods (C4.5 and Naive Bayes). The results of the classification methods C4.5 and Naïve Bayes in terms of accuracy for various cross-validation folds show the same accuracy results in the Ljubljana Breast Cancer dataset.

Furthermore, research by Suseno [11] on the classification of people who receive zakat (*mustahik*). This paper proposes a comparison of the two classification methods in the case of people receiving zakat. By using split validation, the results of the study concluded that the use of the C4.5 method was better than the Naive Bayes method. This is evidenced by the level of accuracy ranging from 75% - 100% and execution time of 0 seconds.

Further research by Pujianto [12] on Diabetes Patients with HbA1c Measurement. This paper proposes two comparisons of classification methods C4.5 and Naive Bayes with HbA1c measurement in seeing the performance of the two methods. By involving a combination of preprocessing methods, namely Synthetic Minority Over-Sampling Technique (SMOTE) and the Wrapper feature selection method, with both classification techniques. The result of the research states that the C4.5 method produces the best performance in classifying diabetic patients with an accuracy value of 82.74%, a precision value of 87.1%, and a recall value of 82.7%.

Based on related research, the results obtained from the best classification method differ according to the cases used. The purpose of this study was to evaluate the performance of the C4.5 and Naïve Bayes classification methods by performing a validation test with 10-Ford X Validation and performing a T-Test differential test [13]. The case raised is Educational Data Mining (EDM) [14] on the student graduation dataset in the research conducted [15], [16]. By changing the dataset as training and testing data, it is hoped that it can evaluate the selection of the best classification method.

2. Methodology

2.1. Dataset

The dataset used is the graduation data of students majoring in informatics engineering at University XYZ which consists of 79 data from students who have graduated from various generations. The attributes used are regional origin, type of school, entrance, predicate cumulative graduation (IPK), predicate of graduation in the first semester (IP1), predicate of second semester graduation (IP2), predicate of third semester graduation (IP3), predicate of fourth semester graduation (IP4), predicate of fifth semester graduation (IP5), boarding school and information. The following is the student graduation dataset as shown in Table 1 below:

Table 1. Research data

Origin	School	Entrance	IPK	IP1	IP2	IP3	IP4	IP5	Boarding school	Information
East Java	MAN	1	C	B	B	B	B	B	No	Incorrect
East Java	MAS	2	B	C	C	B	B	C	No	Incorrect
East Java	SMAN	1	B	B	B	C	C	B	No	Incorrect
Madura	SMAN	2	B	A	B	B	B	B	No	Incorrect
East Java	MAN	1	B	C	B	C	B	B	Yes	Incorrect
West Java	SMKS	2	B	B	B	B	B	B	No	Incorrect
East Java	MAS	1	B	C	B	C	B	B	Yes	Incorrect
East Java	MAN	1	B	C	B	C	C	B	No	Incorrect
East Java	MAS	1	B	B	C	B	B	C	Yes	Incorrect
East Java	MAS	1	B	B	B	B	B	B	Yes	Incorrect
East Java	MAN	4	B	B	B	B	B	B	No	Incorrect
East Java	SMAN	3	B	B	B	B	B	B	Yes	Incorrect
East Java	SMAN	1	B	B	B	B	B	B	No	Incorrect
East Java	SMAN	1	C	B	C	B	B	C	No	Incorrect
East Java	SMAN	1	B	B	C	B	B	B	No	Incorrect
Sumatra	MAN	1	B	B	B	C	B	C	Yes	Incorrect
Sumatra	MAS	1	B	C	C	B	B	B	Yes	Incorrect
East Java	SMAN	1	B	B	B	C	C	B	No	Incorrect
East Java	SMAN	1	B	B	C	B	B	C	No	Incorrect
Sumatra	SMAN	1	B	B	B	C	C	B	No	Incorrect
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…
East Java	SMAN	1	B	B	C	C	B	B	Yes	Incorrect
East Java	SMAS	3	B	B	B	B	B	A	No	Incorrect
Madura	SMAS	1	B	C	C	B	B	B	Yes	Incorrect

source: [15]

In Table 1 it can be explained that the dataset will be used to compare the best classification method by performing a validation test with 10-Ford X Validation and performing a T-Test difference test. Training and testing data is divided into three datasets, namely dataset 1 (25 records), dataset 2 (50 records) and dataset 3 (79 records). The analysis process was carried out using the Rapid Miner software. The following is an explanation of the attributes of the type of school and the predicate of graduation using numbers and letters based on the research dataset.

Table 2. Predicate of Graduation

No	Achievement Index Value	Letter Value	Predicate
1	3,51 – 4,00	A	With Praise (Cumlaude)
2	3,00 – 3,50	B	Very satisfy
3	2,51 – 2,99	C	Satisfactory
4	2,00 – 2,50	D	Enough

Table 3. Entrance

No	Entrance	Score
1	*Mandiri Ujian Tulis*	1
2	*SNMPTN Ujian Tulis*	2
3	*Mandiri Prestasi*	3
4	*SNMPTN Undangan*	4
5	*SPMB - PTAIN*	5

2.2. Classification Algorithms

The proposed classification algorithm aims to achieve a balance between the classification methods used by comparing the performance of these models [6]. The methods used are the decision tree (C4.5) and the traditional statistical classifier (Naïve Bayes) [17].

2.3. Model Validation

The validation model [18] used is cross validation 10-fold stratified, which means dividing the training dataset into 10 equal parts and then do the learning process 10 times and use the rest of the dataset to perform the test. Several tests mention the use of this validation model stratification slightly increased yield [11].

2.4. Model Evaluation

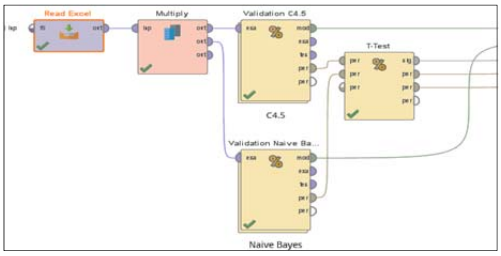
Apply the area under the curve (AUC) for accuracy indicator is to increase increasing convergence across experiments. The following is guidance Table for classifying accuracy using AUC as shown in Table 4 [6].

Table 4. AUC value

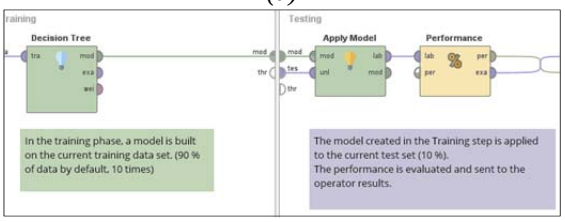
AUC	Meaning
0.90 - 1.00	Excellent Classification
0.80 - 0.90	Good Classification
0.70 - 0.80	Fair Classification
0.60 - 0.70	Poor Classification
< 0.60	Failure

3. Results and Discussion

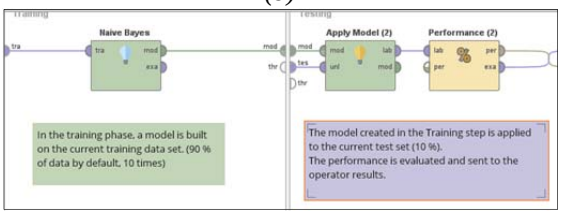
This stage of the analysis process uses the assistance of Rapid Miner software in comparing the best classification method on the student graduation dataset which is divided into three datasets. The following is a design model for the comparison of the C4.5 method with Naïve Bayes using the Rapid Miner software as in Figure 1 below:



(a)



(b)



(c)

Figure 1. Classification Method Comparison Design Model (a)(b)(c)

In Figure 1 (a), the input dataset uses three different datasets (25, 50 and 79) with excel format (.xls). The design uses the multiply operator which functions as a bridge in comparing to the classification methods used at once (C4.5 and Naïve Bayes). After that each method uses the Cross Validation operator which uses the 10-Ford X Validation test on training and testing data (b) (c). Then the further determination is used testing using statistical tests, namely by using the T-Test to compare two methods alternately. Following are the results of a comparative analysis of the C4.5 and Naïve Bayes methods for various datasets (25, 50 and 79).

3.1. Results of Comparative Analysis of Methods with Dataset 3 (79 records)

a) C4.5 (Decision tree)

The following are the results of the analysis of the C4.5 method using RapidMiner software for validation tests with 10-Ford X Validation and T-Test differences as shown in Figures 2 and 3 below:

accuracy: 87.32% +/- 8.35% (micro average: 87.34%)

	true Incorrect	true On time	class precision
pred. Incorrect	67	7	90.54%
pred. On time	3	2	40.00%
class recall	95.71%	22.22%	

Figure 2. C4.5 accuracy results (79 records)

In Figure 2, it is explained that the accuracy value obtained is 87.32%. Here is the AUC value for measuring discriminatory performance with estimate the probability of output from the sample randomly selected from a positive or negative population.

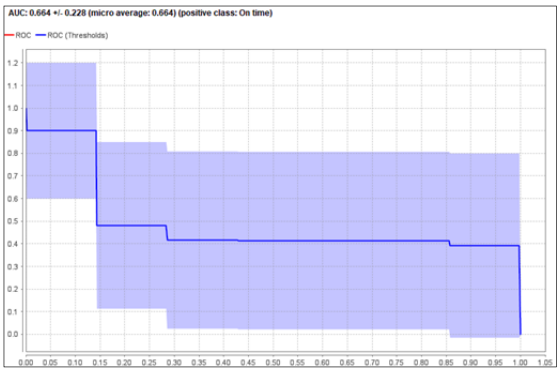


Figure 3. Result of AUC (Area Under the ROC Curve) at C4.5 (79 records)

In Figure 3, the best AUC value is 0.664 and is included in the category of "Poor Classification".

b) Naive Bayes

The following are the results of the analysis of the Naïve Bayes method using RapidMiner software for validation tests with 10-Ford X Validation and T-Test differences as shown in Figures 4 and 5 below:

accuracy: 83.57% +/- 14.45% (micro average: 83.54%)

	true Incorrect	true On time	class precision
pred. Incorrect	63	6	91.30%
pred. On time	7	3	30.00%
class recall	90.00%	33.33%	

Figure 4. Naïve Bayes accuracy results (79 records)

In Figure 4, it is explained that the accuracy value obtained is 83.57%. Here is the AUC value for measuring discriminatory performance with estimate the probability of output from the sample randomly selected from a positive or negative population.

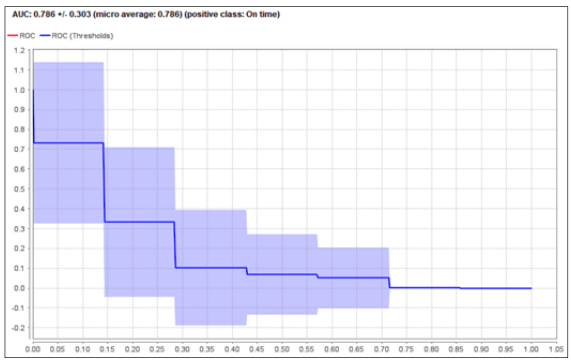


Figure 5. Result of AUC (Area Under the ROC Curve) at Naïve Bayes (79 records)

In Figure 5, the best AUC value is 0.786 and is included in the category of "Fair Classification".

A	B	C
	0.873 +/- 0.084	0.836 +/- 0.145
0.873 +/- 0.084		0.487
0.836 +/- 0.145		

Figure 6. T-Test Statistics Test (79 record)

From the t-test above, that results comparison between C4.5 and Naive Bayes methods there is no significant difference (H0).

Table 5. Comparison results of all tests (79 records)

Dataset	C4.5		Naïve Bayes	
	Accuracy	AUC value	Accuracy	AUC value
79	87.32%	0.664	83.57%	0.786

Based on the Table above, it can be seen that the C4.5 algorithm has an accuracy value the highest was 87.32% and Naïve Bayes 83.57%. Meanwhile, the ROC curve (AUC) test shows that Naïve Bayes achieved the best AUC value of 0.786 compared to C4.5, which is 0.664.

3.2. Results of Comparative Analysis of Methods with Dataset 2 (50 records)

a) C4.5 (Decision tree)

The following are the results of the analysis of the C4.5 method using RapidMiner software for validation tests with 10-Ford X Validation and T-Test differences as shown in Figures 7 and 8 below:

accuracy: 76.00% +/- 12.65% (micro average: 76.00%)

	true Incorrect	true On time	class precision
pred. Incorrect	37	8	82.22%
pred. On time	4	1	20.00%
class recall	90.24%	11.11%	

Figure 7. C4.5 accuracy results (50 records)

In Figure 7, it is explained that the accuracy value obtained is 76%. Here is the AUC value for measuring discriminatory performance with estimate the probability of output from the sample randomly selected from a positive or negative population.

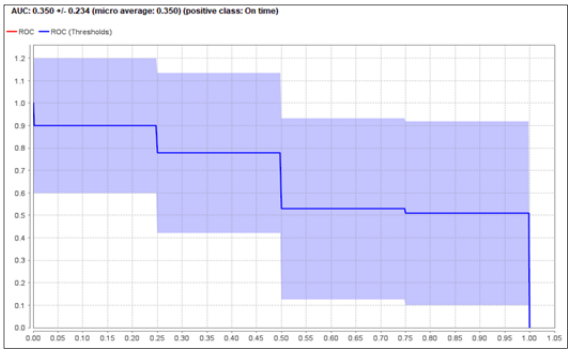


Figure 8. Result of AUC (Area Under the ROC Curve) at C4.5 (50 records)

In Figure 3, the best AUC value is 0.350 and is included in the category of "Failure ".

b) Naive Bayes

The following are the results of the analysis of the Naïve Bayes method using RapidMiner software for validation tests with 10-Ford X Validation and T-Test differences as shown in Figures 9 and 10 below:

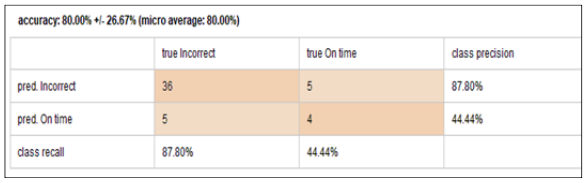


Figure 9. Naïve Bayes accuracy results (50 records)

In Figure 4, it is explained that the accuracy value obtained is 80%. Here is the AUC value for measuring discriminatory performance with estimate the probability of output from the sample randomly selected from a positive or negative population.

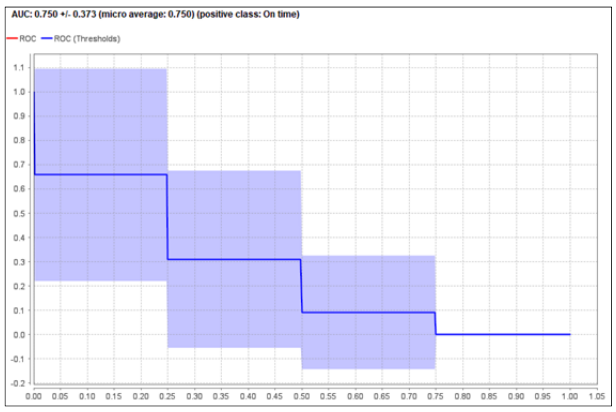


Figure 10. Result of AUC (Area Under the ROC Curve) at Naïve Bayes (50 records)

In Figure 10, the best AUC value is 0.750 and is included in the category of "Fair Classification".

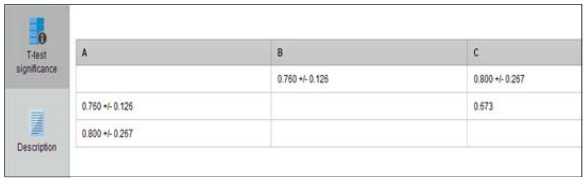


Figure 11. T-Test Statistics Test (50 record)

From the t-test above, that results comparison between C4.5 and Naive Bayes methods there is no significant difference (H0).

Table 6. Comparison results of all tests (50 records)

Dataset	C4.5		Naïve Bayes	
	Accuracy	AUC value	Accuracy	AUC value
50	76%	0.35	80%	0.75

Based on the Table above, it can be seen that the Naive Bayes algorithm has an accuracy value the highest is 80% and C4.5 is 76%. Meanwhile, the ROC curve (AUC) test shows that Naïve Bayes achieved the best AUC value, namely 0.75. Meanwhile, Method C4.5 is in the Failed category because it is <0.60.

3.3. Results of Comparative Analysis of Methods with Dataset 1 (25 records)

a) C4.5 (Decision tree)

The following are the results of the analysis of the C4.5 method using RapidMiner software for validation tests with 10-Ford X Validation and T-Test differences as shown in Figures 12 and 13 below:

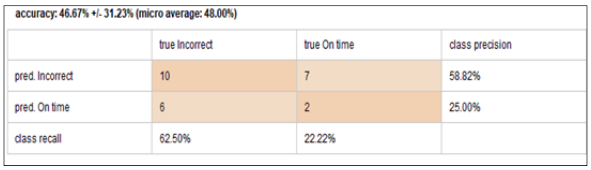


Figure 12. C4.5 accuracy results (25 records)

In Figure 12, it is explained that the accuracy value obtained is 46.67%. Here is the AUC value for measuring discriminatory performance with estimate the probability of output from the sample randomly selected from a positive or negative population.

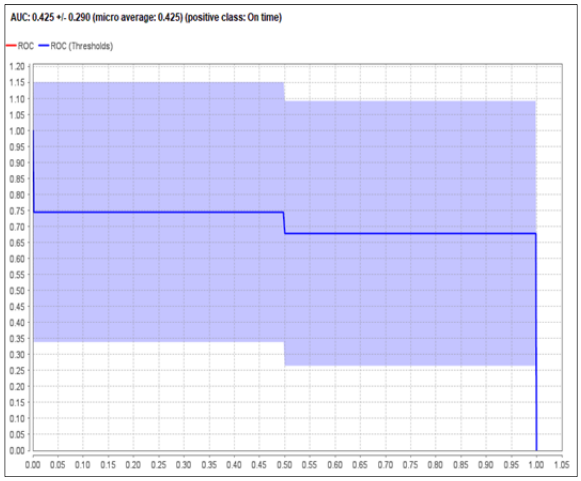


Figure 13. Result of AUC (Area Under the ROC Curve) at C4.5 (25 records)

In Figure 13, the best AUC value is 0.425 and is included in the category of "Failure ".

b) Naive Bayes

The following are the results of the analysis of the Naïve Bayes method using RapidMiner software for validation tests with 10-Ford X Validation and T-Test differences as shown in Figures 14 and 15 below:

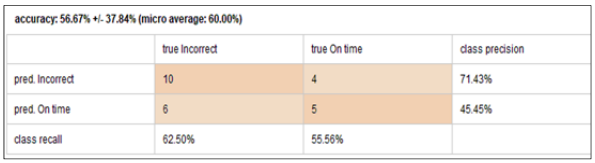


Figure 14. Naïve Bayes accuracy results (25 records)

In Figure 14, it is explained that the accuracy value obtained is 56%. Here is the AUC value for measuring discriminatory performance with estimate the probability of output from the sample randomly selected from a positive or negative population.

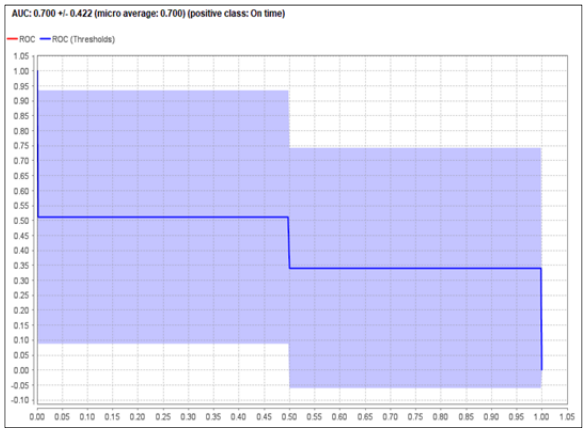


Figure 15. Result of AUC (Area Under the ROC Curve) at Naïve Bayes (25 records)

In Figure 10, the best AUC value is 0.7 and is included in the category of "Fair Classification".

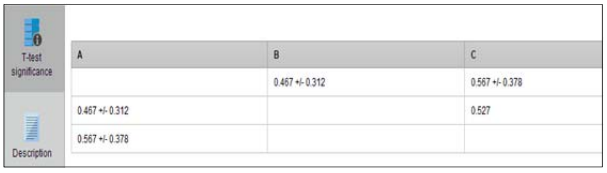


Figure 16. T-Test Statistics Test (25 record)

From the t-test above, that results comparison between C4.5 and Naive Bayes methods there is no significant difference (H0).

Table 7. Comparison results of all tests (25 records)

Dataset	C4.5		Naïve Bayes	
	Accuracy	AUC value	Accuracy	AUC value
25	46.67%	0.425	56.67%	0.7

Based on the Table above, it can be seen that the Naive Bayes algorithm has an accuracy value the highest is 56.67% and C4.5 is 46.67%. Meanwhile, the ROC curve (AUC) test shows that Naïve Bayes achieved the best AUC value, namely 0.7. Meanwhile, Method C4.5 is in the Failed category because it is <0.60.

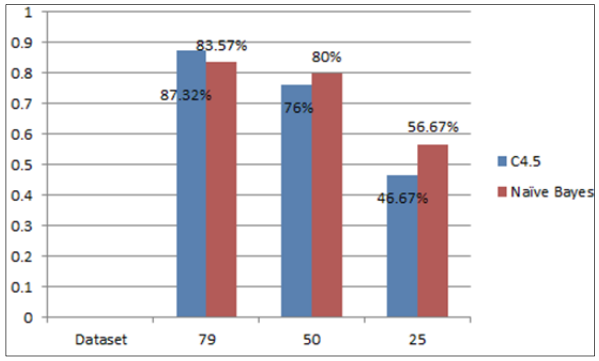
3.4. Discussion

Experiments were carried out on a laptop based on an Intel Core i5, Processor with 8 GB RAM and an operating system Windows 8 is used. Applications are used to carry out research using the Rapid Miner software. In Table 8, you can see the accuracy and AUC values of each method based on the AUC evaluation model.

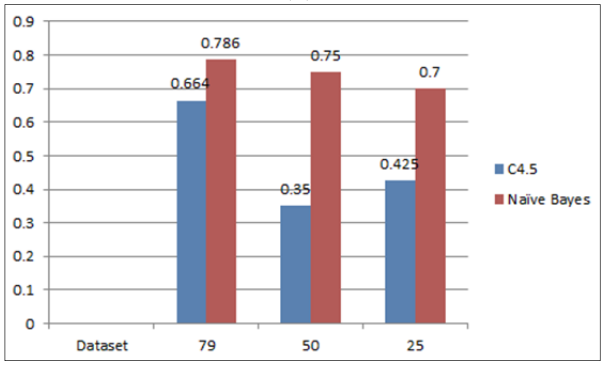
Table 8. Comparison results of all tests

Dataset	C4.5 Accuracy	Naïve Bayes Accuracy	C4.5 AUC value	Naïve Bayes AUC value
79	87.32%	83.57%	0.664	0.786
50	76%	80%	0.35	0.75
25	46.67%	56.67%	0.425	0.7
Average	70.00%	73.41%	0.48	0.75

Based on Table 8, it can be concluded that in the T-Test, the C4.5 and Naïve Bayes methods showed non-dominant results. Based on the three datasets used, the Naïve Bayes method shows better results than the C4.5 method. In 3 experiments (3 datasets), all methods showed a decrease in accuracy for each data test. In the C4.5 method, the highest accuracy is 87.32% on the 79 record dataset. While the Naïve Bayes method, the highest accuracy is 83.57% and 80% on the 79 and 50 datasets. Meanwhile, the AUC value that achieves the maximum results is the Naïve Bayes method with values of 0.789, 0.75 and 0.7. While the AUC value that reaches the maximum result is 0.664. Based on these results, the Naïve Bayes Method is the best classification method on the EDM dataset with an average accuracy value of 73.41% and an area of under curve (AUC) of 0.664. The following is a comparison chart of the classification method based on the accuracy and AUC values for each dataset.



(a)



(b)

Figure 17. The results of the comparison graph of all tests (a)(b)

4. Conclusion

Based on the comparison of data mining classification algorithms, namely C.45 and Naïve Bayes, for the graduation dataset of informatics engineering students at University XYZ which consists of 79 records that have passed from various generations can be done. The results of accuracy and AUC are very influential on the dataset used. The larger the dataset used, the better the accuracy value results. Evidenced by a small dataset, the accuracy value for both methods is greatly reduced. Meanwhile for AUC, the Naïve Bayes method is better than the C.45 method.

References

[1]. Afiyah, A. N., Dengen, N., & Taruk, M. (2019, October). Comparison Performance of C4. 5, Naïve Bayes and K-Nearest Neighbor in Determination Drug Rehabilitation. In *2019 5th International Conference on Science in Information Technology (ICSITech)* (pp. 112-117). IEEE.

[2]. Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications*, *4*(2), 13-19. Doi: 10.14569/SpecialIssue.2014.040203.

[3]. Arumugam, P., & Christy, V. (2018). Analysis of clustering and classification methods for actionable knowledge. *Materials Today: Proceedings*, *5*(1), 1839-1845. Doi: 10.1016/j.matpr.2017.11.283

[4]. Efendi, Y., Imardi, S., Muzawi, R., Syaifullah, M., 2021. Application of RFID internet of things for school empowerment towards smart school. Jurnal Pengabdian Dan Pemberdayaan Masyarakat Indonesia 1(2): 48–58.

[5]. Suyadi, S., Setyanto, A., & Al Fattah, H. (2017). Analisis perbandingan algoritma Decision tree (C4. 5) dan K-naïve bayes untuk mengklasifikasi penerimaan mahasiswa baru tingkat universitas. *Indonesian Journal of Applied Informatics*, *2*(1), 59-68. Doi: 10.20961/ijai.v2i1.13258.

[6]. Wahono, R. S., Herman, N. S., & Ahmad, S. (2014). A comparison framework of classification models for software defect prediction. *Advanced Science Letters*, *20*(10-11), 1945-1950. Doi: 10.1166/asl.2014.5640

[7]. Yousefi, M., Hooshyar, D., Remezani, A., Sahari, K. S. M., Khaksar, W., & Alnaimi, F. B. I. (2015). Short-term wind speed forecasting by an adaptive network-based fuzzy inference system (ANFIS): an attempt towards an ensemble forecasting method. *International Journal of Advances in Intelligent Informatics*, *1*(3), 140-149. Doi: 10.26555/ijain.v1i3.45.

[8]. Conejero, J. M., Preciado, J. C., Fernández-García, A. J., Prieto, A. E., & Rodríguez-Echeverría, R. (2021). Towards the use of Data Engineering, Advanced Visualization techniques and Association Rules to support knowledge discovery for public policies. *Expert Systems with Applications*, *170*, 114509. Doi: 10.1016/j.eswa.2020.114509.

[9]. Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability. *arXiv preprint arXiv:1206.1121*.

[10]. Sivakumari, S., Praveena Priyadarsini, R., & Amudha, P. (2009). Accuracy evaluation of C4. 5 and Naive Bayes classifiers using attribute ranking method. *International journal of computational intelligence systems*, *2*(1), 60-68.

[11]. H. Suseno, A. Wanhari, and S. Masruroh, "Comparison of C4.5 and Naïve Bayes Algorithm for Mustahik Classification," Proc. 2nd Int. Colloq. Interdiscip. Islam. Stud. Conjunction with 3rd Int. Conf. Quran Hadith Stud., 2020, doi: 10.4108/eai.7-11-2019.2294560.

[12]. Pujianto, U., Setiawan, A. L., Rosyid, H. A., & Salah, A. M. M. (2019). Comparison of naïve bayes algorithm and decision tree C4. 5 for hospital readmission diabetes patients using hba1c measurement. *Knowledge Engineering and Data Science*, *2*(2), 58-71. Doi: 10.17977/um018v2i22019p58-71.

[13]. Triyanto, Y., Sepriani, Y., Mustamu, N. E., Siregar, R. A., & Rambe, B. H. (2021, June). Implementation of PROMETHEE Method for Potential Suitability of Land Oil Palm Plant. In *Journal of Physics: Conference Series* (Vol. 1933, No. 1, p. 012060). IOP Publishing.

[14]. Shin, D., & Shim, J. (2021). A Systematic Review on Data Mining for Mathematics and Science Education. *International Journal of Science & Mathematics Education, 19*(4). Doi: 10.1007/s10763-020-10085-7.

[15]. Mashlahah, S. (2013). *Prediksi kelulusan mahasiswa menggunakan metode decision tree dengan penerapan algoritma C4. 5* (Doctoral dissertation, Universitas Islam Negeri Maulana Malik Ibrahim).

[16]. Suryani, L., Lestari, B., Al_barazanchi, I., Ahmad, I., & Singh, M. K. S. (2021). Al-qur'an education park at el-hasan terban mosque, gondokusuman, yogyakarta city. *Jurnal Pengabdian dan Pemberdayaan Masyarakat Indonesia*, *1*(1), 1.

[17]. Xing, W., Li, C., Chen, G., Huang, X., Chao, J., Massicotte, J., & Xie, C. (2020). Automatic Assessment of Students' Engineering Design Performance Using a Bayesian Network Model. *Journal of Educational Computing Research*. Doi: 10.1177/0735633120960422.

[18]. Susanto, E., Novitasari, Y., Rahman, W., & Amane, A. P. O. (2019, December). Designing Software to Introduce the Musical Instruments. In *Journal of Physics: Conference Series* (Vol. 1364, No. 1, p. 012031). IOP Publishing.