International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Comparison of Student Graduation Classification Analysis Based on Study Length Using Naïve-Bayes and C4.5 Algorithms

Dwi Ismiyana Putri[1], Novita Sulistyowati[2]
[1]Department of Computer Science, Gunadarma University, Jakarta, Indonesia, 16424
[2]Department of Computer Science, Gunadarma University, Jakarta, Indonesia, 16424

Abstract**—** *Data mining is the process of finding patterns and knowledge from large number of data. An important part of data mining is data classification. Classification is used to classify data based on the nature of the data that each class has recognized. There are various techniques used to classify data, two of which are C4.5 and Naive Bayes. C4.5 is an algorithm used to form a decision tree while Naive Bayes is a classification method using the Bayes theorem. Based on some researchers, the C4.5 and Naive Bayes methods have good performance so the system was made with the aim to compare the performance of C4.5 and Naive Bayes. The data that will be used in this study are academic students of one of the state universities in Indonesia. Based on the results of research using WEKA 3.9.3 tools it is known that the performance of the C4.5 algorithm is better than the performance of the Naive Bayes algorithm. This can be seen from the value of accuracy, recall, precision produced by C4.5 is greater than that of Naive Bayes..*

Keywords**—** *Data Mining, Classification, Students, Naïve Bayes, Decision Tree, C4.5, WEKA 3.9.3.*

I. INTRODUCTION

Right now, education in Indonesia is developing rapidly. The more development of universities and schools is a sign of the development of education. The number of applicants who want to enter universities is increasing every year, causing a competition between both state and private universities. So this condition become motivation for each universities to continue to improve the quality of its education.

Data in the academic field is becoming an important and interesting thing for every educational[1]. The high level of student success and the low level of student failure is a reflection of the quality of a college. Graduation rate is considered as one of institutional[2]. Paying attention to the number of graduates of a tertiary institution becomes an importan thing. Besides being in a very competitive environment, right now every university is also trying to continuously improve management to improve quality of the grade.

Badan Akreditasi Nasional Perguruan Tinggi (BAN PT) tasked with assessing the quality of a college[7]. The quality of institution is able to influence the accreditation of its institution by watching to one of the accreditation standards according to BAN PT. And if a student has exceeded 8 (eight) semesters, then the student is relatively slow in completing his studies. To determine the level of graduation of students in one school year, a classification can be done based on student data at the level or the first school year.

Data mining is an analysis of a data set review to find unexpected relationships and be able to summarize data in a different way than before, which can be understood and beneficial for the data owner[3]. One of the most important part of data mining is data classification. Classification is used to classify data based on the nature of the data that each class has recognized. There are various techniques used to classify data, two of which are the C4.5 algorithm and the Naïve Bayes algorithm. From some of the studies above, the Naïve classification method as an alternative method in evaluating academics. The Naïve Bayes classification method was chosen because the Naïve Bayes method is a simple statistical probability method but produces accurate results. However, the most accurate algorithm is still unknown in determining the graduation of students based on the length of their studies.

The rest of the paper is structured into 4 sections. In section 2, a review of the related work is presented. Section 3 contains the data mining process implemented in this study, which includes a representation of the collected dataset, an exploration and visualization of the data, and finally the implementation of the data mining tasks and the final results. In section 4, insights about future work are included. Finally, section 5 contains the outcomes of this study.

II. RELATED WORK

Study Makhtar M[4] analyze the potential use of one data mining technique called the Naïve Bayesian algorithm to improve the quality of student performance at Sijil Pelajaran Malaysia. The purpose of this study is to test the Naïve Bayes algorithm which is one of the classification methods in data mining, to identify hidden information between subjects that affect student performance at Sijil Pelajaran Malaysia (SPM). For future research, there will be a process of redefining a technique with more attributes and data to get more accurate results that can be useful for teachers to improve the outcomes of student learning for semester exams at Sijil Pelajaran Malaysia.

Next Dangi A[5] suggested from the results of his research analysis that the use of the Naïve Bayes algorithm can be used as a media for classifying small data sets. However, the percentage that is classified can still be increased by using other classifiers such as Support Vector Machine (SVM), K-

Dwi Ismiyana Putri and Novita Sulistyowati, "Comparison of Student Graduation Classification Analysis Based on Study Length Using Naïve-Bayes and C4.5 Algorithms," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 1, pp. 233-238, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Nearest Neighbor (K-NN), and others with shorter time. For future research he suggested implementing the same procedure with performance comparisons conducted using other classification techniques and larger data sets.

Then the research conducted by Haviluddin[6] concluded about the analysis using the Naïve Bayes Classifier (NBC) technique to achieve a student academic performance model that has been implemented at the CSIT Faculty, Mulawarman University. This study confirms that the NBC algorithm has better accuracy in evaluating student academic performance. In other words, the NBC algorithm can be used as an alternative model in student academic evaluation performance. Therefore, other machine learning algorithms to get better accuracy performance are the future work.

III. DATA MINING PROCESS

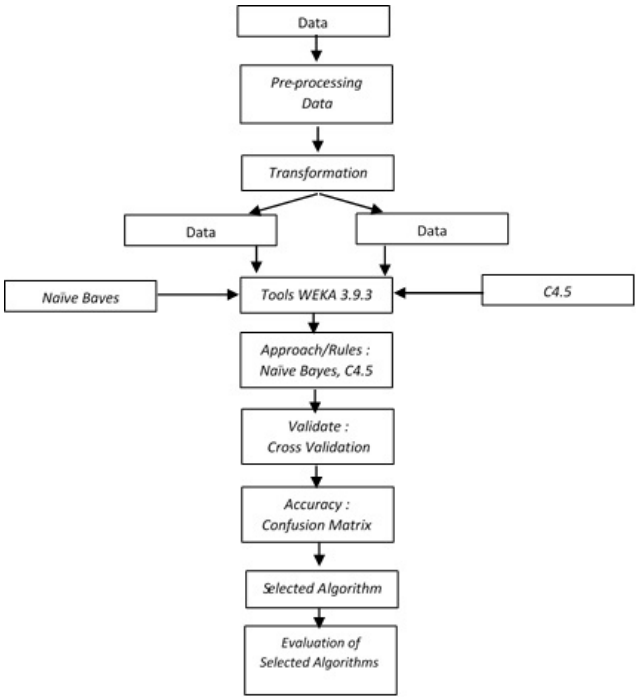

Fig. 1. Problem-solving framework

In this study, the type of research taken is a comparative experiment. This comparative experimental research is based on a problem-solving framework that can be seen in Fig.1[8].

A. *Pre-processing Data*

Before implementing data mining classification techniques, at this stage the attributes used such as student ID, gender, study program, class, GPA, total credits, and graduation status must first be converted from two stages, namely data extraction and data transformation into the form which can be used in WEKA data mining tools to identify patterns with techniques data mining classification.

B. *Extraction and Data Transformation*

The data sources of the academic attributes of students used came from the following two tables:

a. Gender attribute is obtained from the jk column in the mhs table.
b. The prodi attribute is obtained from the prodi column in the mhs table.
c. Angkatan attribute is obtained from the angkatan column in the mhs table.
d. The IPK attribute is obtained from the ipk column in the perkuliahanMhs table.
e. The SKS Total attribute is obtained from the TotalSKS column in the mhs table.
f. The status lulus attribute is obtained from the reduction of the StatusKeluar column and the angkatan column in the mhs table.

id	jk	prodi	smstr_mul	status_kelu	ipk	sks_tot	smester_lul
ID2	L	Sistem Informasi S1	20131	1	3.19	149	20172
ID3	L	Sistem Informasi S1	20131	1	3.18	149	20172
ID4	L	Sistem Informasi S1	20131	1	3.46	148	20172
ID5	L	Sistem Informasi S1	20131	1	3.41	152	20162
ID6	L	Sistem Informasi S1	20131	1	3.21	147	20171
ID7	L	Sistem Informasi S1	20131	1	3.2	147	20162
ID8	L	Sistem Informasi S1	20131	1	3.25	149	20171
ID9	L	Sistem Informasi S1	20131	1	3.35	150	20172
ID10	P	Sistem Informasi S1	20131	1	2.94	155	20172
ID12	P	Sistem Informasi S1	20131	1	3.34	148	20162
ID13	P	Sistem Informasi S1	20131	1	3.49	152	20171
ID14	P	Sistem Informasi S1	20131	1	3.27	153	20172
ID15	P	Sistem Informasi S1	20131	1	3.29	150	20171
ID17	P	Sistem Informasi S1	20131	1	3.24	151	20172
ID19	P	Sistem Informasi S1	20131	1	3.05	149	20172
ID20	P	Sistem Informasi S1	20131	1	3.47	150	20162
ID21	P	Sistem Informasi S1	20131	1	3.14	147	20162
ID22	P	Sistem Informasi S1	20131	1	3.39	150	20172
ID23	P	Sistem Informasi S1	20131	1	3.59	155	20162
ID24	P	Sistem Informasi S1	20131	1	3.3	147	20162
ID25	P	Sistem Informasi S1	20131	1	3.42	149	20162

Fig. 2. Examples of Student Academic Data Before the Data Pre-processing Stage

From the total of the academic data tables for bachelor degree students from the Faculty of Computer Science Department of Information Systems, Computer Systems, and Informatics Engineering from 2008 to 2015 amounted to 4,306 student data. But the data used in this study will only use as many as 1,676 student data. Because in the mhs table there is an StatusKeluar column as shown in Fig. 2 with NULL information or other than status 1 already graduate, data cleansing will be performed.

Based on the attributes needed to obtain a dataset whose results will be used in classification techniques, the academic attributes of students can be described by following table:

TABLE I. Attributes Data Type and Description

Attribute	Data Type	Description
Jenis Kelamin	Text	Student Gender / L (Male), P (Female).
Prodi	Text	The majors of the Faculty of Computer Science are Information Systems (SI), Computer Systems (SK), Informatics Engineering (TI).
Angkatan	*Numeric*	Year of student entry / 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015.
IPK	*Numeric*	Cumulative Performance Index (GPA) is a measure of a student's ability up to a certain time.
Total SKS	*Numeric*	Total semester credit units taken by students.
Status Kelulusan	Text	Calculated study status starts when registered as a student until graduation / TEPAT (ON TIME), TERLAMBAT (LATE)

Based on the data extraction carried out in the akademik database in accordance with the existing Status Kelulusan, the

Dwi Ismiyana Putri and Novita Sulistyowati, "Comparison of Student Graduation Classification Analysis Based on Study Length Using Naïve-Bayes and C4.5 Algorithms," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 1, pp. 233-238, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

number of training sets and testing sets obtained can be seen in TABLE II.

TABLE II. Number of Data Records Used

Status Kelulusan	Student Academic (*Train Set*)	Student Academic (*Test Set*)
TEPAT (ON TME)	490	123
TERLAMBAT (LATE)	850	213

After the data extraction stage, there will be a transformation into a form that can be used in WEKA data mining tools to identify patterns with data mining classification techniques such as in Fig. 3 below:

jenis_kelamin	prodi	angkatan	ipk	sks_total	status_lulus
L	SI	2013	3.19	149	TERLAMBAT
L	SI	2013	3.18	149	TERLAMBAT
L	SI	2013	3.46	148	TERLAMBAT
L	SI	2013	3.41	152	TEPAT
L	SI	2013	3.21	147	TEPAT
L	SI	2013	3.2	147	TEPAT
L	SI	2013	3.25	149	TEPAT
L	SI	2013	3.35	150	TERLAMBAT
P	SI	2013	2.94	155	TERLAMBAT
P	SI	2013	3.34	148	TEPAT
P	SI	2013	3.49	152	TEPAT
P	SI	2013	3.27	153	TERLAMBAT
P	SI	2013	3.29	150	TEPAT
P	SI	2013	3.24	151	TERLAMBAT
P	SI	2013	3.05	149	TERLAMBAT
P	SI	2013	3.47	150	TEPAT
P	SI	2013	3.14	147	TEPAT
P	SI	2013	3.39	150	TERLAMBAT
P	SI	2013	3.59	155	TEPAT
P	SI	2013	3.3	147	TEPAT
P	SI	2013	3.42	149	TEPAT

Fig. 3. Transforming Student Academic Data after the Data Pre-processing Stage

C. Classification Using Algoritma Naïve Bayes

In research using the classification method using the Naïve Bayes Algorithm, researchers used 1,341 training sets and 335 testing sets to view academic data with the status kelulusan result attributes.

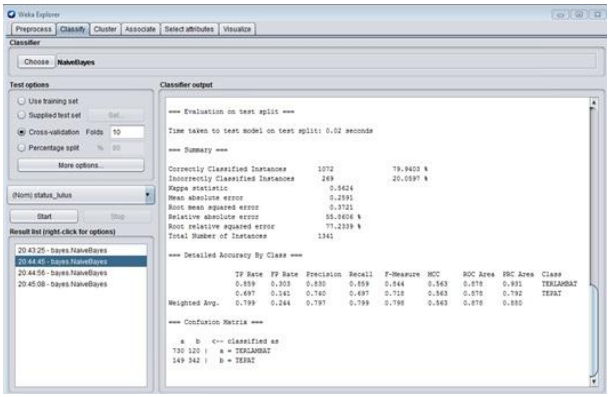

Fig. 4. Display of Naïve Bayes Training Data Classification Results

The results of classification techniques using the Naïve Bayes Algorithm for the training set have an accuracy value of 79.9%, then the value of precision which explains the level of accuracy between the information requested by the user with the answer requested by the system by 79.7%, the recall value describes the success rate of the system in finding back information by 79.9%, and the F-Measure Recall which is the mean harmonic weight of recall and precision obtained a value of 79.8%.

TABLE III. Confusion Matrix of Academic Data Training on the Naïve Bayes Algorithm

	TERLAMBAT (LATE)	TEPAT (ON TIME)
TERLAMBAT (LATE)	730	120
TEPAT (ON TIME)	149	342

Naïve Bayes algorithm in training data produces an accuracy value of 79.9% on the TERLAMBAT (LATE) test set predictions with 120 prediction errors in the TERLAMBAT (LATE) data, and 149 errors in predicting the TEPAT (ON TIME) data.

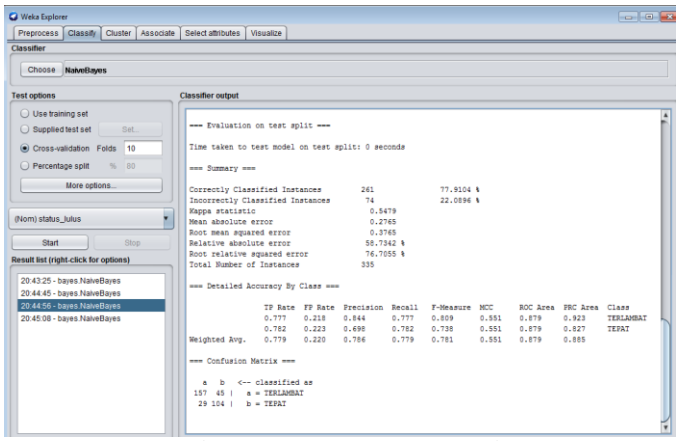

Fig. 5. Display of Naïve Bayes Testing Data Classification Results

Classification techniques using the Naïve Bayes Algorithm for testing sets have an accuracy value of 77.9%, then a precision value of 78.6%, a recall value describes the success rate of the system in finding information back at 77.9%, and F-Measure Recall which is the harmonic mean weight of recall and precision obtained a value of 78.1%. And for confusion matrix training data can be seen in TABLE IV below:

TABLE IV. Confusion Matrix of Academic Data Testing on the Naïve Bayes Algorithm

	TERLAMBAT (LATE)	TEPAT (ON TIME)
TERLAMBAT (LATE)	157	45
TEPAT (ON TIME)	29	104

From the results of the trial training set and akademik data set testing using the Naïve Bayes Algorithm, it can be concluded from the TABLE V below:

TABLE V. Classification Results for Student Graduation in the Naïve Bayes Algorithm

Algorithm	Training Data Accuracy	Data Testing Accuracy	Precision Data Testing	Recall Data Testing	F-Masure Recall Data Testing
Naïve Bayes	79.9%	77.9%	78.6%	77.9%	78.1%

Dwi Ismiyana Putri and Novita Sulistyowati, "Comparison of Student Graduation Classification Analysis Based on Study Length Using Naïve-Bayes and C4.5 Algorithms," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 1, pp. 233-238, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

As can be seen in the TABLE V the results of classification techniques using the Naïve Bayes algorithm have an accuracy of testing data of 77.9%, which means the level of closeness between the predicted value and the actual value is close to 100%, it can be concluded that testing with the composition of training data and data percentage testing used is quite valid, with a precision data testing value of 78.6%, a recall test value of 77.9%, and F-Measure Recall which is the harmonic mean weight of recall and precision testing obtained a value of 78.1%.

D. Classification Using Algoritma C4.5

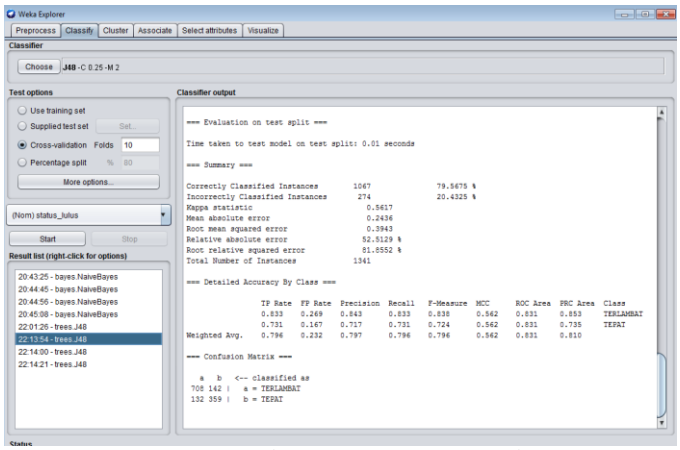

Fig. 6. Display Results of Naïve Training Data Classification C4.5

As can be seen in Figure the results of classification techniques using C4.5 Algorithm for the training set have an accuracy value of 79.6%, then the value of precision which explains the level of accuracy between the information requested by the user with the answer requested by the system by 79.7%, recall value explains the success rate of the system in finding back information was 79.6%, and the F-Measure Recall which is the harmonic mean weight of recall and precision obtained a value of 79.6%.

TABLE VI. Confusion Matrix of Academic Data Training on the C4.5 Algorithm

	TERLAMBAT (LATE)	TEPAT (ON TIME)
TERLAMBAT (LATE)	708	142
TEPAT (ON TIME)	132	359

In TABLE VI explains that the C4.5 Algorithm in training data produces an accuracy value of 79.7% in the TERLAMBAT test set predictions with 142 prediction errors in the TERLAMBAT (LATE) data, and 132 errors in predicting the TEPAT (ON TIME) data.

While the results of classification techniques using the C4.5 algorithm for testing set has an accuracy value of 80.9%, then the value of precision which explains the level of accuracy between the information requested by the user with the answer requested by the system by 80.8%, the recall value describes the level of success the system in finding information back was 80.9%, and the F-Measure Recall which is the harmonic mean weight of recall and precision obtained a value of 80.8%.

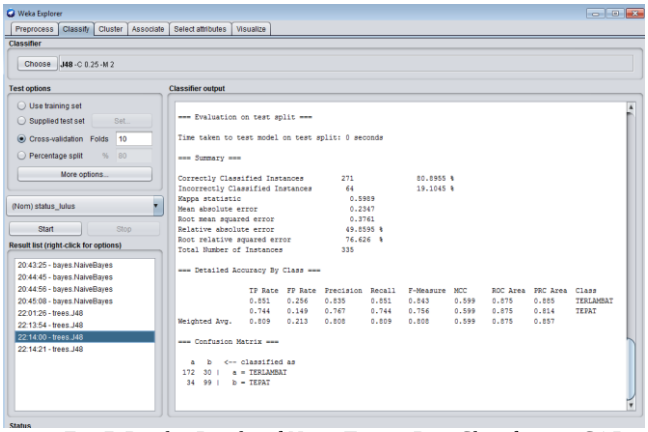

Fig. 7. Display Results of Naïve Testing Data Classification C4.5

In Confusion Matrix testing the data produces an accuracy value of 80.9%, in the TERLAMBAT (LATE) test set predictions with 30 prediction errors in the TERLAMBAT (LATE) data, and 34 errors in predicting the TEPAT (ON TIME) data.

TABLE VII. Confusion Matrix of Academic Data Testing on the C4.5 Algorithm

	TERLAMBAT (LATE)	TEPAT (ON TIME)
TERLAMBAT (LATE)	172	30
TEPAT (ON TIME)	34	99

From the results of the trial training set and academic data set testing using C4.5 Algorithm, it can be concluded from TABLE VIII below:

TABLE VIII. Classification Results for Student Graduation in the C4.5 Algorithm

Algorithm	Training Data Accuracy	Data Testing Accuracy	Precision Data Testing	Recall Data Testing	F-Masure Recall Data Testing
C4.5	79.6%	80.9%	80.8%	80.9%	80.8%

Classification using C4.5 algorithm has an accuracy value of testing data of 80.9%, which means the level of closeness between the predicted value and the actual value is close to 100%, then it can be concluded that the test with the composition of the training data and testing data used is quite valid, with a value of precision data testing is 80.8%, recall testing is 80.9%, and F-Measure Recall which is the harmonic mean weight of recall and precision testing is 80.8%.

E. Analysis and Summary

The Naïve Bayes algorithm has a higher AUC value of 0.879 than the C4.5 algorithm with a value of 0.853. As for accuracy, C4.5 algorithm is superior with a percentage of 80.25% compared to Naïve Bayes which only has a percentage of 78.9%. However, both the Naïve Bayes algorithm and the C4.5 algorithm are included in a good classification method with a range of AUC values of 0.80 - 0.90 = good classification.

Dwi Ismiyana Putri and Novita Sulistyowati, "Comparison of Student Graduation Classification Analysis Based on Study Length Using Naïve-Bayes and C4.5 Algorithms," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 1, pp. 233-238, 2020.

TABLE IX. Comparative Results of Accuracy Confusion Matrix and AUC Value in Student Graduation Testing Data

Algorithm	*Confusion Matrix*		AUC		Comparison of Accuracy Values	Comparison of AUC values
	Training	*Testing Data*	*Training Data*	*Testing Data*		
Naïve Bayes	79.9%	77.9%	0.878	0.879	78.9%	**0.879**
C4.5	79.6%	80.9%	0.831	0.875	**80.25%**	0.853

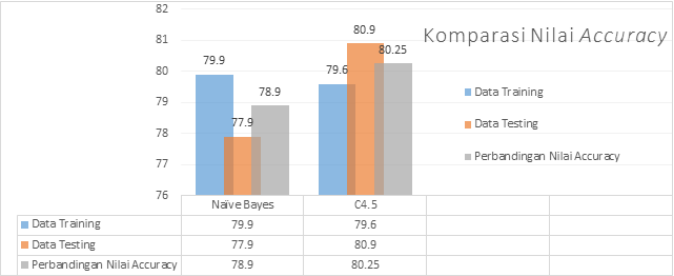

Fig. 8. Comparison of Accuracy Values in the Naïve Bayes Algorithm and C4.5 Algorithm

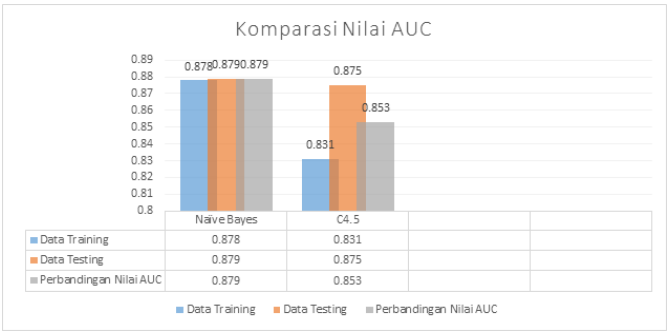

Fig. 9. Comparison of AUC values in the Naïve Bayes algorithm and the C4.5 algorithm

Regarding the accuracy value obtained in the classification technique carried out, then in accordance with the classification of accuracy values generated in the application of classification techniques is to use the C4.5 algorithm method in Klasifikasi Kelulusan Mahasiswa Berdasarkan Masa Studi into the very good classification category.

IV. CONCLUSION

The results of classification techniques with the C4.5 algorithm in this study produce a very good prediction accuracy percentage in the process of identifying the graduation punctuality based on the study period has an average prediction accuracy value of 80.25% and AUC value of 0.853.

Therefore the classification technique with the C4.5 algorithm is very suitable to be used to assess the accuracy of the accuracy of the student graduation punctuation using attribute data that has been stored in the akademik database, but does not rule out the possibility of future research along with the more advanced development methods can be used - other data mining methods from two algorithms that have been tested in this study and the results can be used as a comparison of the data mining methods used today. Because in this study there are still results of student graduation that are not appropriate based from the attributes used in data processing. Henceforth, it is expected that research will be able to obtain better predictive accuracy.

And from the evaluation of the classification model obtained, thus the university can determine the pattern of the algorithm that has been tested.

V. FUTURE WORK

As the times develops, the future is predicted to emerge the latest patterns of new algorithms with data mining classification techniques. With the advent of the new algorithm, it is expected to be implemented in various data mining tools and developed into an application.

Further research is expected research can use a larger number of datasets to be processed and produce a better classification model and to be able to compare with other algorithms as well as combined with clustering models that should be tried to test the accuracy of training data and testing data to measure suitability in the classification of student graduation based on the length of study.

ACKNOWLEDGMENT

I would like to express my very great appreciation to Dr. Novita Sulistyowati, S.Kom., M.M. for his valuable efforts in teaching me data mining and exploration module in my Masters study, as well as, for her valuable and constructive suggestions and overall help in conducting this research.

As well as, I would like to thank and express my very great appreciation to Riza Adrianti Supono, S.Kom, MMSI for her great support and inspiration during the whole time of conducting this research, as well as, her willingness to allocate her time so generously.

Last but not least, I would like to thank my husband, my family, and my friends for continuously supporting and inspiring me during my study and research.

REFERENCES

[1] N. Putpuek, K. Atchariyachanvanich, N. Rojanaprasert, and T. Thamrongthanyawong, "Comparative Study of Prediction Models for Final GPA Score: A Case Study of Rajabhat Rajanagarindra University," IEEE ICIS 2018, 978-1-5386-5892-5/18, pp. 92-97, 2018.

[2] Qudri M. N. and Kalyankar N. V, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques," Global Journal of Computer Science and Technology, pp. 2-4, 2010..

[3] Larose, D. T, "Discovering Knowledge in Data: An Introduction to Data Mining," John Wiley & Sons, Inc, Hoboken, New Jersey, 2005.

[4] M. Makhtar, H. Nawang, and S.N.W Shamsuddin, "Analysis on Students Performance Using Naïve Bayes Classifier," Journal of Theoretical and Applied Information Technology, Vol.95. No.16, pp. 3993 – 4000, 2017.

[5] A. Dangi and S. Srivastava, "Educational data Classification using Selective Naïve Bayes for Quota categorization," IEEE International Conference on MOOC, Innovation and Technology in Education (MITE), 978-1-4799-6876-3/14, pp. 118-121, 2014.

[6] Haviluddin, N. Dengen, E. Budiman, M. Wati, and U. Hairah, "Student Academic Evaluation using Naïve Bayes Classifier Algorithm," IEEE, The 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT), 978-1-5386-8050-6/18, pp. 104-107, 2018.

[7] Akreditasi Perguruan Tinggi, Badan Akreditasi Nasional Perguruan Tinggi, Naskah Akademik Iapt 3.0, Jakarta, 2019.

[8] Rahmawati, E, "Analisa Komparasi Algoritma Naive Bayes Dan C4.5 Untuk Prediksi Penyakit Liver," Jurnal Techno Nusa Mandiri : Vol. XII No. 2, pp 27-37, 2015.

Dwi Ismiyana Putri and Novita Sulistyowati, "Comparison of Student Graduation Classification Analysis Based on Study Length Using Naïve-Bayes and C4.5 Algorithms," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 1, pp. 233-238, 2020.