Application of Data mining to Prediction of Timeliness Graduation of Students (A Case Study)

1st Chandra Wirawan, 2nd Eva Khudzaeva, 3rd Tuhfatul Habibah Hasibuan, 4th Karjono, 5th Yeni Hilmi Khairani Lubis
[1,3,4,5] *STIMK Islam International*
[2]*UIN Syarif Hidayatullah Jakarta*
Jakarta, Indonesia
chandrawirawan50@gmail.com, eva.khudzaeva@uinjkt.ac.id, tuhfatulhabibah95@gmail.com,
aryo.y2k@gmail.com, yeni.hilmi@gmail.com

***Abstract*— Timely graduation is one indicator of student success in obtaining a bachelor's degree. Timely graduation is one indicator of the assessment of the quality of higher education, because of one of the assessments of accreditation of the National Accreditation Board (BAN-PT). So that if the graduation and student levels are not balanced, it will affect the accreditation assessment of the Study Program and the University. According to University accreditation data at data.uinjkt.ac.id, there are insignificant data between students who enter and the student graduation rate on time. The number of graduates in the last 7 years S-1 students who graduate on time is less than 50%. In the long term, this can lead to a buildup of students at Syarif Hidayatullah State Islamic University in Jakarta. In this paper, we compare the best data mining methods that will be used for predictions and identify graduation rates on time. 3 data mining technique methods used are decision tree using C.4.5 algorithm, Naïve Bayes and KNN. The data tested is student data for 2012-2014. The result obtained is C.4.5 algorithm decision tree model because it has the highest accuracy value compared to KNN and Naïve Bayes. So the results of this study are expected to provide a strategic plan for Syarif Hidayatullah UIN Jakarta, in controlling and monitoring students graduating on time to maintain the quality of education at the Syarif Hidayatullah UIN Jakarta.**

Keywords— Decision tree, algorithm C.4.5, Naïve Bayes, KNN, Prediction of Timeliness Graduation of Students, Data mining, Comparison Data mining

I. INTRODUCTION

Timely graduation is one indicator of student success in obtaining a bachelor's degree. In practice, students cannot always complete undergraduate education in four years.

Accreditation of tertiary institutions as a provider of higher education provides a method of evaluation of higher education institutions in measuring, determining the quality and feasibility of their programs. Timely graduation is one indicator of the assessment of the quality of higher education because one of the accreditation assessments of the National Higher Education Accreditation Board (BAN-PT) contained in standard 3 concerning students and graduates is the profile of students graduating on time. So that if the graduation and student levels are not balanced, it will affect the accreditation assessment of the Study Program and the University.

UIN Syarif Hidayatullah Jakarta was established with the Republic of Indonesia's Presidential Decree Number 031 of 2002. According to University accreditation data at data.uinjkt.ac.id, there are insignificant data between students who enter and the student graduation rate on time. The number of students who graduate on time, which is by the provisions of BAN-PT accreditation is much smaller than the number of students who enter the Syarif Hidayatullah UIN Jakarta.

The number of graduates in the last 7 years S-1 students who graduate on time does not reach 50%. In the long term, this can lead to a buildup of students at the Syarif Hidayatullah UIN in Jakarta and of course it will result in a decrease in the value of accreditation for the Study Program and the University.

One technique for predicting can be done using data mining, data mining method can be used to provide knowledge previously hidden in the data warehouse so that it becomes valuable information, data mining is used to predict student travel time. [1]

Prediction of the graduation rate on time has been done by Salmu and Solichin [2], the method used by Salmu and Solicin uses the Naïve Bayes method with an accuracy of 80.72%. while this paper uses a comparison of 3 methods of data mining techniques to predict on-time graduation rates using the decision tree method using the C.4.5 algorithm and then Naïve Bayes and KNN. From these comparisons it is expected to obtain the best method for predicting graduation rates on time.

II. LITERATURE REVIEW

Data mining is the activity of finding interesting patterns of large amounts of data, data stored in a database, data warehouse, or other information storage [3]. Some researchers use a comparison of data mining classification techniques, Janwata and Tsai (2013), this study builds graduate employment models using classification, by comparing several data mining approaches, such as Bayesian methods and Tree methods. The Bayesian Method includes 5 algorithms, including AODE, BayesNet, HNB, Naive Bayes, WAODE. The Tree method includes 5 algorithms, including BFTree, NBTree, REPTree, ID3, C4.5. and it turns out that from the comparison, the C.4.5 algorithm is superior inaccuracy, which is equal to 98.71%, while for the other algorithms it is below the accuracy value of algorithm C.4.5. [1]. According to Kaur et al (2015) in Classification and prediction based data mining, the slow learners predict to predict in the education sector. identifying students who are slow among other students, displaying them with predictions of data mining models using classification-based algorithms, data compared with using Naïve Bayes, SMO, J48, and REPTree, and after testing the accuracy value C.4.5 is also higher than others, that is equal to 69.73%. [4]. F. Marbouti et al (2014). use a comparison of data mining methods to create predictive models early prediction of at-risk students in a course, by comparing 7 data mining methods, and naive bayes and ensembles have the best value of the 7 methods tested [5]. According A.M . Shahiri et al, Data mining can be used to identify the most important attributes in student data.

so that it can improve student achievement and success more effectively [6].

III. METHODOLOGY

A. Sample Selection Methods

According to statistical law in determining the number of samples, the greater the number of samples, the more the population and the use of large numbers of samples are highly recommended, taking into account the various limitations of the researcher, so researchers try to take the minimum sample with statistical requirements and rules but are met as recommended Issac and Michael.

B. Data Collection Methods

Data collection is the most important part of a study. The availability of data is crucial in the process of data processing and subsequent analysis, therefore in data collection must pay attention that the data obtained must be accurate and can be scientifically accountable

A Analysis Techniques

The analysis technique used in the development of data mining techniques in this study is using the CRISP-DM method (Cross Standard Industries for Data Mining). There are 6 phases in this method, the stages are as follows: [7]

1) Business Understanding

According to university accreditation data at data.uinjkt.ac.id, there are insignificant data between students entering and leaving. The number of graduates in the last 7 years S-1 students who graduate on time does not reach 50%. In the long term, this can lead to a buildup of students at Syarif Hidayatullah State Islamic University in Jakarta. So from that, this study will predict the graduation rate on time for students at Syarif Hidayatullah UIN Jakarta

2) Data Understanding

To predict student graduation on time, researchers used academic data for 2012-2014 students as many as 754 records, the data was obtained from the Data Center for Information and Data (PUSTIPANDA) Syarif Hidayatullah State Islamic University Jakarta, the data used consisted of 9 Predictive Attributes and 1 result attribute, so the total attributes used are 10 parameter attributes. Consisting of type_gender, program_studi, type_School, majors_SLTA, region, IPSmt 1-4, and graduation. Presented in table 1. Data set variable

TABLE I. DATA SET VARIABLE

No	Attribute	Value	Type
1	Gender	Male	Binominal
		Female	
2	Program Studi	religion	Binominal
		general	
3	Type_school	MAS	Polynomial
		MAN	
		SMAN	
		SMAS	
		SMK	
4	Major_SLTA	Religion	Polynomial
		IPA	
		IPS	
		other	
5	Region	Java	Binominal
		other	
6	IPSmt 1	Numerik	Polynomial
7	IPSmt 2	Numerik	Polynomial
8	IPSmt 3	Numerik	Polynomial
9	IPSmt 4	Numerik	Polynomial
10	graduation	On-time	Binominal
		Not on time	

3) Data Preparation

Initial data processing needs to be done to prepare correct and valid data before processing. Academic data that has been obtained will be made preprocessing data.

4) Modeling Phase

At this stage, it is also called the learning stage because at this stage the processing of training data which is classified by the model then produces some rules. In paper using a comparison of 3 methods, namely the C.45 algorithm Decision tree, Naïve Bayes Classification and KNN, using RapidMiner 7.3 tools to process the data.

a) Decision Tree is a modeling method based on partitioning. In each step, it partitions the data based on one variable until all data in each node have only one category label or all variables have been used [8].

b) Naive Bayes Classifier (NBC) is a simple probabilistic classifier that calculates a conditional probability distribution over the output of a function based on applying Bayes theorem with the (naive) assumption of independence between the predictive variables[9].

c) K-Nearest Neighbor (KNN) is a non-parametric classifier. Unlike the methods described above, it does not train a model with parameters. KNN classifies an object (e.g., a student) by a majority vote of its K neighbors [10]

5) Evaluation Phase

This stage is used to test the model that aims to get the most accurate model into Rapid Miner 7.3 frameworks. Evaluation and validation in this study using the confusion matrix method, this method is used to measure precision, Recall, and Accuracy.

6) Deployment Phase

At this stage the researchers implemented a data mining model using a comparison of 3 methods namely the C.45 Decision tree algorithm, Naïve Bayes Classification and KNN. to predict the graduation rate on time for UIN Syarif Hidayatullah Jakarta students at LPM UIN Syarif Hidyatullah Jakarta

IV. RESULT

A. Grouping and Data Analysis

The data used in this study consisted of training data and testing data totaling 754 student data. Before processing using the RapidMiner data application, validation is done by deleting incomplete or empty data (null), the data that has been validated is then divided into two parts data training and testing data with a ratio of 70% data transfer and 30% data testing using Stratified random sampling, data used in this study has 10 attributes 9 attributes as predictors and 1 attribute results

B. Test Results with Tools

The stages of the process that will be carried out in testing the data set begin with the process of entering the verified data set into the RapidMiner program, there are stages of the process as follows:

1) Collecting datasets for training: is made in Excel format (.xlsx) then the Excel data is imported into the Rapid Miner application, following the dataset that has been imported into the Rapid Miner program

2) Determine the sample by dividing the dataset into two parts, namely data training by 70% and data testing by 30% using Split Data operators as shown in Figure 1.

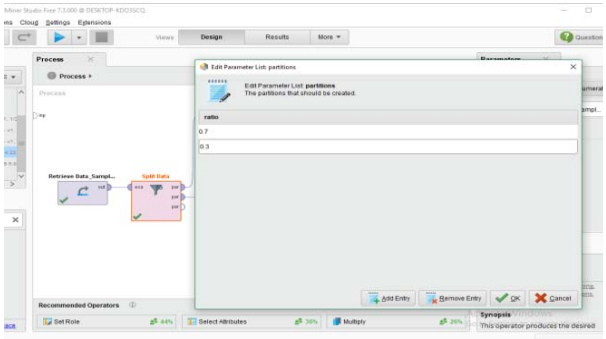

Fig. 1. Application of Split Data Operators in Data Datasets

3) Implement the selected algorithm operator, using the Decision Tree C.4.5 algorithm, naïve bayes and KNN. On the training data, then proceed with the stage of applying the model to the testing data using the Apply Model operator. Design can be seen as shown in Figure 2

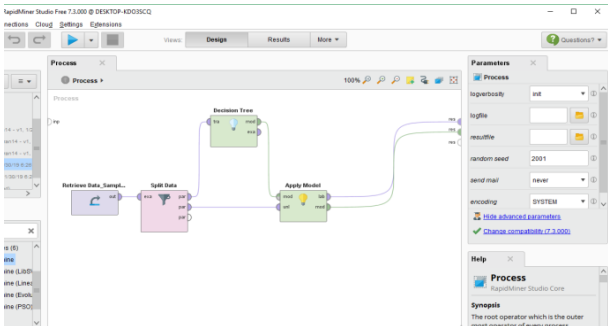

Fig. 2. Application of Decision Tree C.4.5 Algorithm Operators

4) The data set during this works is tested and analyze with 3 classification methods those are Decision Tree, Naïve Bayes and KNN The last stage is testing the model using the confusion matrix method, at this stage the Performance operator is used to evaluate the model obtained at the present stage in the Rapid Miner program, namely in the form of accuracy, precision, recall,

a) Accuracy: is the amount of comparison of correct data with the total amount of data. The highest accuracy of the comparison of data mining methods is the decision tree of 89.82%, that is, predictions regarding graduation rates on time show accurate results. The accuracy data is presented in Table 2 below

TABLE II. ACCURACY OF CLASSIFIER COMPARISON

Data Mining Technique	Accuracy
Decision Tree	89.82%
Naïve Bayes	85.40%
KNN	84.07%

b) Precision: is used to measure the proportion of the positive data class that has been correctly predicted from the overall positive class prediction results, The highest of precision from the comparison of 3 data mining methods is decision tree of 52.63%. The result of precision comparison are shown in the below Table 3.

TABLE III. PRECICION OF CLASSIFIER COMPARISON

Data Mining Technique	Precision
Decision Tree	52.63%
Naïve Bayes	39.20%
KNN	25%

c) The recall: is used to show the percentage of positive data classes that have been successfully predicted correctly from the overall positive class data, The smallest percentage of recall from the comparison of 3 data mining methods is decision tree of 41.67% . all the comparison result of Recall are provided in Table 4.

TABLE IV. RECALL OF CLASSIFIER COMPARISON

Data Mining Technique	Recall
Decision Tree	41.67%
Naïve Bayes	67%
KNN	66.67%

Based on an accuracy value of 89.82%, that means, Prediction of Timeliness Graduation of Students show accurate results of 89.82%, it can be seen from small prediction errors, according to this model, the correct time prediction is 193 and has a prediction error of 14, while the prediction is not on the correct time of 10 and which has a prediction error not as much as 9.

V. CONCLUSION

The purpose of this paper is to compare the best prediction method from the 3 methods tested to determine the prediction passed on time. Of the 3 methods, the decision tree method

using C.4.5 algorithm is the method that has the highest accuracy value compared to the other two methods, which is 89.82%, while the accuracy of the Naïve Bayes method is 85.4% and the accuracy of the KNN method is 84.07%. In this paper, the data set used to predict graduation rates is 754 students consisting of 9 prediction attributes and 1 outcome attribute, so the total attributes used are 10 parameter attributes. Consists of type_gender, program_studi, types_School, majors_SLTA, region, IPSmt 1-4, and graduation.

REFERENCES

[1] Jantawan, B, "The Application of Data Mining to Build Classification Model for Predicting Graduate Employment," 11(10) 2013.

[2] Supardi, S., Solichin, A, "Prediction of Timeliness Graduation of Students Using Naïve Bayes: A Case Study at Islamic State University Syarif Hidayatullah Jakarta," Prosiding Seminar Disiplin Ilmu, Universitas Budi Luhur, 2017. ISSN : 2087-0930

[3] Han, J., Kamber, M., & Pei, J, *Data Mining – Concepts & Techniques*, 2012, https://doi.org/10.1016/B978-0-12-381479-1.00001-0

[4] Kaur, P., Singh, M., & Singh, G, "Classification and prediction based data mining algorithms to predict slow learners in the education sector," Procedia - Procedia Computer Science, 57, 500–508, 2015 https://doi.org/10.1016/j.procs.2015.07.372

[5] Marbouti, F., Diefes_Dux, H.A., Madhavan, K, "Models for early prediction of at-risk students in a course using standards-based grading," Computer & Educations, 103, 2016, 1-15

[6] Shahiri, A, M., Husain, W., Rashid, N, A., "A Review on Predicting Student's Performance using Data Mining Techniques," Procedia Computer Science, 72, 2015, 414-422

[7] Larose, D. T. "Discovering Knowledge In Data An Introduction to Data Mining. Automotive Industries AI (Vol. 190)," New Jersey: John Wiley & Sons, Inc., Hoboken, New Jersey, 2005, https://doi.org/10.1016/j.cll.2007.10.008

[8] Hand, D. J., Mannila, H., & Smyth, P, Principles of data mining. New York: MIT press. 2001

[9] Russell, S., & Norvig, P, Artificial intelligence: A modern approach. Englewood Cliffs, NJ: Prentice-Hall, 1995

[10] Friedman, J. H., Bentley, J. L., & Finkel, R. A, An algorithm for finding best matches in logarithmic expected time. ACM Transactions on Mathematical Software (TOMS), 3(3), 209e226, 1997

.