Analysis of Students Graduation Target Based on Academic Data Record Using C4.5 Algorithm Case Study: Information Systems Students of Telkom University

[1]Dela Youlina Putri, [2]Rachmadita Andreswari, [3]Muhammad Azani Hasibuan
[1,2,3]Study Program of Information System, School of Industrial Engineering, Telkom University
Jl. Telekomunikasi No 01, Terusan Buah Batu, Bandung, West Java 40257 INDONESIA
[1]delayoulina.putri@gmail.com, [2]andreswari@gmail.com, [3]muhammad.azani@gmail.com

***Abstract*- Study program of Information Systems is one of the existing study programs at Telkom University which has produced many graduates until 2017. However, not all graduates produced successfully completed the study period during four years of normal study period in which may cause the decrease of study programs quality and affect the assessment of study program if there is an audit or evaluation so it can affect the achievement level of the study program. To solve the problem can be by making a prediction model of student graduation that can be obtained from data classification process using decision tree with algorithm C4.5 and implement it to the academic data record of existing student so that got two group of student, that is student which predicted pass on time and student predicted to pass late. From the results of the classification of student data can be done an analysis of what factors that can affect the graduation of students who are predicted to pass on time and plan appropriate strategies for groups of students who may not pass on time. The data classification process is done with the help of open source based tools using RapidMiner application. The result of the classification is a prediction model that has an accuracy value of 82.24% and states that the most influential factor in predicting students' graduation is GPA in the second year. The result of the student's graduation classification is expected to be used as the reference base to support the academic planner in making the right decision to the student groups generated so that all students can graduate on time.**

Keywords- student passing prediction, data classification, decision tree, C4.5 algorithm, RapidMiner

I. INTRODUCTION

Telkom University is one of the private universities in Indonesia that provides education services to its students in order to create an integrity and competence graduates with national and international competitiveness. Telkom University offers 31 study programs sheltered by seven faculties, one of them is Information System (IS) study program in the School of Industrial Engineering. This study program emphasizes on the use of information and communication technology (ICT) to solve business or organizational problems and able to provide an evaluation of the strategic value of ICT utilization in the achievement of business or organizational goals, so that IS graduates will have competencies in the field of technology and business [1]. Every year, IS always accepts new students with a big capacity that come from different regions all around Indonesia. The students who are accepted in this study program have to complete four levels science of information systems based on a predetermined curriculum that are IS fundamental, IS core, IS depth, and IS breadth during the four years of normal study. Until 2016, this study program has produced approximately 800 graduates since its establishment in 2008.

Based on the number of graduates, it is known that not all students complete their studies in a timely manner within four years of normal study period. There are several factors that might affect the students to complete their studies. As shown in Fig. 1, the percentage of students who can graduate on time over the past 4-years has decreased. This can certainly affect the assessment of the study program during the accreditation and achievement of the university so that the study program needs to carry out continuous monitoring and measurement on the performance of its students in order to complete its study on time and to support the achievement of the study program's targets.

To do so, it is required to create a prediction model that can be obtained from the process of seeking information on existing data.

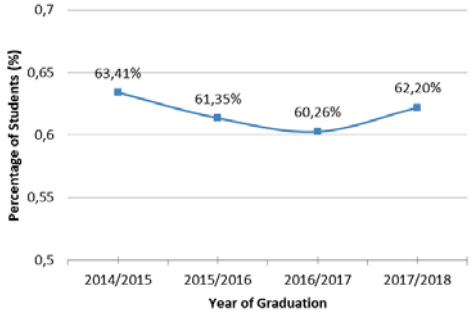

Fig. 1. Chart of Students Graduation on Time

As the number of students grow every year, the data that stored in the database of student information system which in Telkom University is called *i-Gracias* will increase as well. The process of seeking information on those data can be done using some existing data mining techniques which in this research using C4.5 algorithm.

Previous research conducted by Adhatrao et al [5] uses the classification algorithm of ID3 and C4.5 algorithm to perform student performance analysis in order to provide a better perspective for student academic performance in the future. It can help educational institutions to identify outstanding students and also improve students who may still get lower grades. Therefore, the study conducted an analysis based on several parameters such as incoming test scores, gender, percentage values on physics, chemistry and mathematics during class XII, as well as admission type which then resulted in a web-based system that can used by decision makers at these educational institutions to predict future student performance based on academic data records to be able to support appropriate decision-making.

Yadav & Pal [6] compared performance of three decision tree algorithms for data classification of 90 student data of Veer Bahadur Singh Purvanchal University, India which proves that classification with C4.5 algorithm is more accurate than ID3 algorithm or CART algorithm. It is supported by the level of accuracy of C4.5 algorithm is 67.78% which is bigger than second of other algorithm that is each 62.22%.

Another study conducted by Guleria et al [7] has the main objective of collecting knowledge of student performance and to identify students who need special attention which can help in predicting student outcomes. It also helps educational institutions in identifying student performances whose attendance numbers are lacking and shows poor performance in the learning sessions. This research used decision tree classifier that is C4.5 algorithm to 120 student dataset using five attribute of student data such as student class performance, attendance, task, laboratory work and student learning performance. This study also performs entropy calculations of each attribute taken in the Educational Data Set and attributes with the highest information gain used as the root node. The result of this research is an analysis which states that poor performance of students in learning session is the main factor of student failure in final examination.

Based on some previous research, it can be concluded that this research uses decision tree with algorithm C4.5 to classify students who can pass on time based on the record of academic data.

II. THEORY

A. *Knowledge Discovery in Database*

Knowledge Discovery in Database (KDD) is a process of extracting information on large amounts of data. In general, the whole KDD process can be explained as follows [12]:

1. Data Selection
   Data selection from a set of operational data needs to be done before the stage of extracting information in KDD begins. Selected data will be used for data mining process, stored in a file, separate from the operational database.
2. Preprocessing
   The preprocessing / cleaning process includes removing data duplication, checking inconsistent data, and correcting data errors, such as typographical errors.
3. Transformation
   Coding is a transformation process in the data that has been selected, so the data is appropriate for the process of data mining. The coding process in KDD is a creative process and depends on the type or pattern of information to be searched in the database.
4. Data Mining
   Data mining is the process of finding patterns or interesting information in selected data by using a particular technique or method. Techniques, methods, or algorithms in data mining vary widely. The choice of the appropriate method or algorithm depends heavily on the purpose and process of KDD as a whole.
5. Interpretation / Evaluation
   The pattern of information generated from the data mining process needs to be displayed in a form that is easily understood by interested parties. This stage is part of the KDD process called interpretation. This stage includes examining whether the pattern or information found is against the previous fact or hypothesis.

B. *Data Classification*

Classification is a form of data analysis that extracts a model that describes important data classes. Classification is a process for finding models (or functions) that describe and differentiate between data classes or data concepts. The classification of data consists of two steps of the learning process, which in this step will build the classification model, and the second step is the classification phase where the model has been built will be used to predict the class label for the new data given [2].

According to Gorunescu [3], the classification process is based on four basic components as follows:

1. *Class* is a dependent variable in the form of categorical variables of the model that represents a label on the object after classification. Examples are customer loyalty, earthquake type, and so on.
2. *Predictor* is an independent variable indicated by the characteristics (attributes) of data to be classified and based on the classification that has been made. Examples are the frequency of purchase, the marital status of the customer, the direction and speed of the wind, the seasons and the location of the earthquake.

3. *Training set* is a set of data that contains values from the above two components and is used to classify the appropriate classes based on existing predictors. Examples are customer groups in supermarkets (generated from internal polls), databases on storms and databases on earthquake research.
4. *Testing set* is a new set of data that will be classified with pre-made models and classification accuracy can be evaluated.

C. Decision Tree

Decision tree is used to predict the membership of an object against a different class by considering the value corresponding to its attributes (predictor variable). The purpose of the decision tree is to divide the data set into groups whose variables are homogeneous in order to make predictions [4].

Decision tree is a tree-shaped flowchart that can be characterized as follows: each internal node in addition to the leaf-node represents testing based on a particular attribute, whereas each branch node represents the test result and each leaf node represents the class label.

D. C4.5 Algorithm

Algorithm C4.5 is one of the algorithm used to perform the process of data classification by using decision tree technique. Algorithm C4.5 is the development of ID3 algorithm which is also an algorithm to build a decision tree. The C4.5 algorithm recursively visits each decision node, chooses optimal branching, until no more branches are possible [13].

The steps in building the decision tree using the C4.5 algorithm are as follows [8]:

1. Prepare training dataset. Training datasets are usually obtained from pre-existing data history and have been grouped into specific classes.
2. Specifies the root attribute of the decision tree. The root attribute is determined by calculating the gain value of each attribute and selecting the attribute with the highest gain value as the first root attribute. To calculate the gain can use the formula in the following equation:

$$Gain(S,A)=Entropy(S)-\sum_{i=1}^{n}\frac{|S_i|}{|S|}*Entropy(S_i) \quad (1)$$

Which is, S : set of cases
A : attribute
n : number of attribute partition A
$|S_i|$: number of cases on the *i*-th partition
$|S|$: number of cases in S

And to calculate the value of entropy can use the formula in the following equation:

$$Entropy(S)=\sum_{i=1}^{n}-pi*\log_2 p_i \quad (2)$$

Which is, S : set of cases
n : number of partitions S
p_i : proportion of S_i to S

The default splitting criterion used by the C4.5 algorithm is the gain ratio of the following formula:

$$GainRation(S,A)=\frac{Gain(S,A)}{SplitInformation(S,A)} \quad (3)$$

Which is, S : set of cases
A : attribute
Gain (S, A) : information gain on attribute A
SplitInformation (S, A) : split information on attribute A

Split information states entropy or potential information with the following formula:

$$SplitInfo(S,A)=-\sum_{i=1}^{n}\frac{S_i}{S}\log_2\frac{S_i}{S} \quad (4)$$

Which is, S : set of cases
A : attribute
S_i = number of samples for attribute *i*

3. Create a branch for each attribut
4. Divide the case in the branch
5. Repeat the process for each branch until all the cases on the branch have the same class
6. The decision tree partition process will stop when:
   a. All records in node N get the same class.
   b. No attributes or variables in the record are partitioned again.
   c. No record in the empty branch.

III. METHODOLOGY

A. Study Case Analysis

In this study, we used academic data records of Information System students from 2009 to 2012. Based on the data obtained the result shows that this study program has a total of 486 graduates in year 2009 - 2012. There is only 80.25% of graduates who can complete their studies on time. Therefore, it is necessary to have a prediction model of graduation which can be used to predict the graduation on time of students who are studying in study program of Information System.

B. Data Preparation

In this study, we used data records of students from 2009 to 2012 which is not the data of students who are "drop-out" or "resign" that obtained from the SISFO of Telkom University. The data used as input data is student background data and student academic data with a total of 554 records. The raw data will be divided into two parts as follows:

1. *Training set* : data are used in the system training process and consists of input data pair and target data. From the total data obtained, 70% - 80% part will be used as training data.
2. *Testing set* : data are used to test the ability of the system and also consists of input data pair and target data. Data testing used amounted to 20% -30% of the data obtained.

TABLE I
LIST OF ATTRIBUTES

Attribute	Values	Description
SID	Unique values	Student identification number
SHS	1 – 6	Senior High School of Students which is 1 : Public High School 2 : Private High School 3 : Public Vocational High School 4 : Private Vocational High School 5 : Public School of Madrasah Aliyah 6 : Private School of Madrasah Aliyah
PJ	1 – 7	Parents Job which is 1 : Civil Servants 2 : Private Employees 3 : Entrepreneur 4 : State-owned Enterprise Employees 5 : Army/Police 6 : Unemployed 7 : Others
PI	Low, High	Parents income which is Low ≤ 2.5 millions High > 2.5 millions
RC	Yes, No	Repetition of course which is Yes : students have retaken course No : students never retake the course
GPA	A, AB, B, BC, C, D	GPA score for the 4th semester which is A = 3.51 – 4.00 AB = 3.01 – 3.50 B = 2.51 – 3.00 BC = 2.01 – 2.50 C = 1.01 – 2.00 D < 1.00
TAK	Less, Enough	Points of Student Activity during college which is Less < 60 Enough >= 60
GradStat	Ontime, Late	Graduation status of students which is Ontime = 4-years of study Late > 4-years of study

There are many factors that affect students on-time graduation as in [2] [9] [10] [11] [14]. For this research, we selected factors that might affects students on-time graduation based on data records that we have collected. There were eight attributes that we used as predictors that shown at Table I above.

C. Data Preprocessing

The raw data obtained is divided into three different CSV files. Those three raw data that have been obtained are then processed to become input data through the process of preprocessing data. Preprocessing data is done using Pentaho Data Integration (PDI) application which consists of data transformation and data cleaning. The transformation stage is performed to obtain data in accordance with the required input data format. Furthermore, for data cleaning, we had to normalize the data by correcting and doing some adjustment to data that is not appropriate so that no missing value is found again.

D. Data Processing

Data processing was performed by implementing C4.5 algorithm to the training set by calculating the gain ratio for all attributes that have been determined, then select the attribute with the highest gain ratio to be the root node. After that, repeat the gain ratio calculation process and form a node that contains the attribute until all data has been included in the same class. The step of data processing is shown in Fig 2.

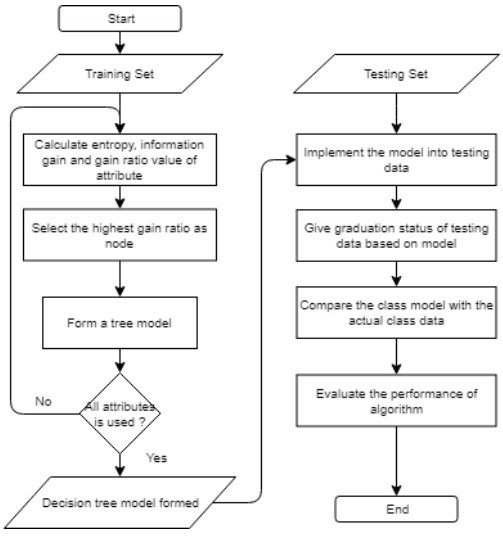

Fig. 2. Flowchart of Data Processing

E. Implementation

This step is done by implementation of C4.5 algorithm using RapidMiner tool as shown in Fig. 3. The input data is a CSV file that contains of student id, six attributes and class label, that have been splitted as training set and testing set with ratio 8:2. We are using decision tree operators with gain ration as the criteria to create the model and apply the model that has been created into testing set. Performance operator is used to evaluate the performance of the algorithm.

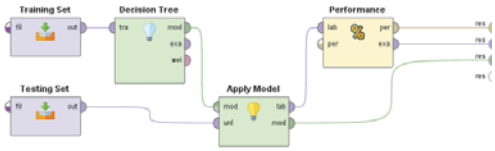

Fig. 3. Implementation in RapidMiner

IV. Result and Analysis

We used cross-validation for evaluating the performance of C4.5 algorithm and the result shows that the accuracy is 82.24% with the class precision and class recall show in Table II below.

TABLE II
Performance Evaluation Result

	True Ontime	**True Late**	**Class Precision**
Prediction Ontime	81	12	87.10%
Prediction Late	7	7	50.00%
Class Recall	92.05%	36.84%	

The results of the classification process of the graduation prediction with the decision tree method are shown in Fig. 4. Based on the decision tree result, it can be seen that the attribute that has the highest influence to determine the classification of the student's graduate on time is GPA score in the 4th semester. This is indicated by the GPA attribute occupying as the root node. If the GPA score are A or B, it can be predicted that the students can graduate on time while if the GPA score is AB then should pay attention to other attributes such as repetition of course, parents income, TAK points and the origin of the senior high school of students. Students with GPA score AB are predicted to graduate on time if they never retake the course. Then if GPA score are BC and C, the students have the highest probability to be predicted to graduate late, so the academic planners should give more attention to those students.

V. Conclusions and Future Work

Based on the results of research conducted it can be concluded that has been obtained classification model of students graduation on time or late from 554 records of Information Systems students in Telkom University with the accuracy of 82.24%. The attribute that has the highest influence on the classification result is the student's GPA score in the fourth semester.

Interpretation of research results indicate that the attribute that needs to be used as a consideration to obtain timely graduation is the GPA score in the fourth semester is more than 2.51. Therefore, the results of this study can be used as the reference for academic planners to increase the target on-time student graduation by making academic strategies such as hold short semester for students who get a GPA score below 2.51 in the fourth semester to help them improve their score.

In the future, we aim to develop this model to website and define another factor that might affect the students graduation on time so the model would be more accurate. Then the project will be connected to university system and tested in a real-time environment.

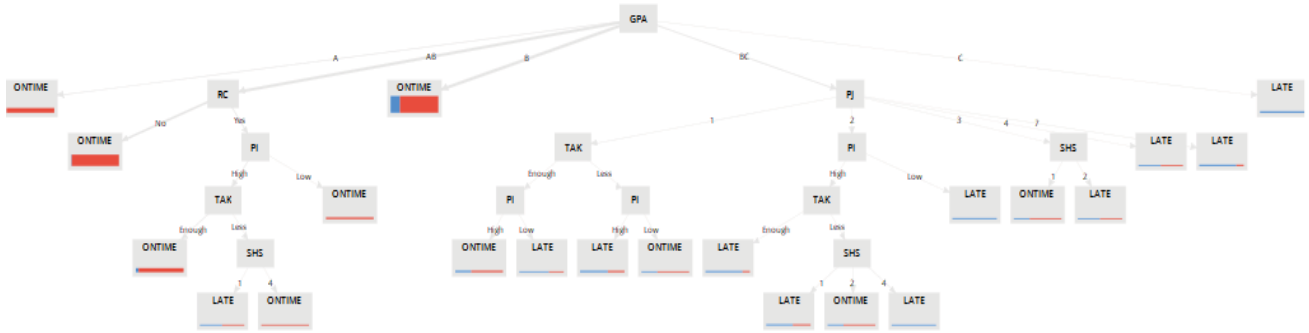

Fig. 4. Decision Tree of Students Graduation on Time

ACKNOWLEDGMENT

This work cannot complete without the great support from all faculty staffs in Faculty of Industrial Engineering. Thanks to our mentor for the guidance and evaluation of this work and always give a very useful suggestions for better result. Lastly, we would like to express our gratitude to our parents for their support and encouragement that keeps us motivated to finish this work. If there is any mistake in this work, we would like to improve it in the future.

REFERENCES

[1] About Information System of Telkom University, 2018. [Online] Available: http://bis.telkomuniversity.ac.id/web/about-us/

[2] J. Han, M. Kamber, and J. Pei, *Data Mining : Concepts and Techniques*, 3rd ed., USA: Elsevier Inc, 2012.

[3] F. Gorunescu, *Data Mining : Concepts, Models and Techniques*, vol. 12, Verlag Berlin Heidelberg: Springer, 2011.

[4] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree ID3 and C4.5," *International Journal of Advanced Computer Science and Applications (IJACSA),* 2014, pp.13-19.

[5] K. Adhatrao, A. Gaykar, A. Dhawan, R. Jha, and V. Honrao, "Predicting students' performance using ID3 and C4.5 classification algorithms," *International Journal of Data Mining & Knowledge Management Process (IJDKP),* 2013, pp.39-52.

[6] S. Yadav, & S. Pal, "Data mining: a prediction for performance improvement of engineering students using classification," *World of Computer Science and Information Technology Journal (WCSIT)*, 2012, pp.51-56.

[7] P. Guleria, N. Thakur, and M. Sood, "Predicting student performance using decision tree classifiers and information gain," *International Conference on Parallel, Distributed and Grid Computing*, 2014, pp.126-129.

[8] Kusrini, and E.T. Luthfi, *Algoritma Data Mining,* Yogyakarta: Penerbit Andi, 2009.

[9] I. Tahyudin, E. Utami, A. Amborowati, "Comparing clasification algorithm of data mining to predict the graduation students on time," *Information Systems International Conference (ISICO),* 2013, pp.379-384.

[10] M. Dragičević, M.P. Bach, V. Šimičević, "Improving university operations with data mining: predicting student performance," *International Journal of Economics and Management Engineering,* 2014, vol. 4, pp.1101-1106.

[11] M. Goga, S. Kuyoro, N. Goga, "A recommender for improving the student academic performance," *International Conference Edu World,* 2014, pp.1481-1488.

[12] F. Nasari, and S. Darma, "Penerapan k-means clustering pada data penerimaan mahasiswa baru (studi kasus : universitas potensi utama)," *Seminar Nasional Teknologi Informasi dan Multimedia,* 2015, pp.73-78.

[13] I. Rahmayuni, "Perbandingan performansi algoritma c4.5 dan cart dalam klasifiksi data nilai mahasiswa prodi teknik komputer politeknik negeri padang," *Jurnal TEKNOIF,* 2014, pp.40-46.

[14] I.A Ganiyu, "Data Mining: A Prediction for Academic Performance Improvement of Science Students using Classification," *International Journal of Information and Communication Technology Research*, 2016, vol. 6.