Indonesian Journal of Informatics Education

Research Paper ISSN: 2549-0389

Predicted Student Study Period with C4.5 Data Mining Algorithm

Agus Supriyanto
Informatics and Computer Education Program,
Faculty of Teacher Training and Education,
Sebelas Maret University
agusjimok@student.uns.ac.id

Dwi Maryono, S.Si., M.Kom.
Informatics and Computer Education Program,
Faculty of Teacher Training and Education,
Sebelas Maret University
dwimarus@yahoo.com

Febri Liantoni, S.ST., M.Kom.
Informatics and Computer Education Program,
Faculty of Teacher Training and Education,
Sebelas Maret University
febri.liantoni@gmail.com

Abstract:

Data of alumni from 2012 to 2015 found that the average percentage of students graduating on time was 22%. The comparison between the number of students who graduate on time and new students who enter each year is not comparable, therefore a study is needed to find out the factors that affect student graduation and to predict the graduation period of the student through data mining research using the C4.5 algorithm. The data tested was student alumni data from 2012 to 2015. The instruments studied include study period, academic year, GPA, corner focus, gender, the intensity of work during college, type of thesis, intensity of campus internal organization, the intensity of external organization of campus, UKT group, scholarship status, pre-college education, hobby intensity, the intensity of gameplay, academic competition participation status, non-academic competition participation status, and availability of facilities and infrastructure. The best test results using percentage-split 75% obtained 83.33% accuracy as well as the rules contained in the decision tree.

Keywords: C4.5 algorithm, Data mining, Predicted study period

1. Introduction

Data of Alumni in 2012 had a total of 52 students with students who graduated on time as many as 24 students resulting in a percentage of 46%. For 2013, there were 68 students with 9 students graduating on time, resulting in a percentage of 13%. Meanwhile, in 2014, there were 43 students with students who graduated on time as well as 5 students, resulting in a percentage of 11%. And for 2015, there were 57 students with 12 students graduating on time, resulting in a percentage of 21%. Based on the above data, the average percentage of students who graduate on time is 22%.

The comparison between the number of students who graduate on time and the new students who enter each year is not comparable, therefore a study is needed to find out the factors that affect student graduation and to the prediction of the graduation period of the student. This prediction study requires a certain amount of data to be processed to predict the study period of students whether to graduate on time or not on time. With the fact that factors affect the sooner or later students graduate, academics can create a policy to optimize students to graduate on time.

Data mining is a process to obtain useful information from a large set of data (Kasus et al., 2015). The term data mining is also often referred to as knowledge covered. In data mining, it has various techniques, namely Association Rule Mining, Clustering, and Classification techniques. To be able to predict a thing used data mining classification techniques. Classification techniques have various methods, including the C4.5 method, Naive Bayes, kNN, and others. In this study, the C4.5 method was applied to predict the study period of students. The C4.5 algorithm is a classification algorithm that uses the decision tree (Astuti, 2017). The decision tree is a large interpretation of data into a simpler form of using the decision tree structure. The process in the decision tree is to turn the table data into a tree model, and then become rules.

(Mustafa et al., 2018) in his research obtained the most dominant factors in determining the classification of student academic performance, namely Compulsive Achievement Index (GPA), Semester Achievement Index (IPS) semester 1st and 4th, and student gender. This research applies data mining method C4.5 to determine the predicted study period of students based on gender, the origin of high school, and social sciences semesters 1st to 6th.

(Agustina & Wijanarto, 2016) in his research for the classification of recipients of drinking water installation grants in PDAM Kendal district compared the algorithms ID3 and C4.5. from his research produced an accuracy value for algorithm ID3 of 98.91%. The C4.5 algorithm obtained an accuracy of 99.14% with the amount of data processed as much as 1473 data processed as much as 9 times from 10% to 90% data.

This research was conducted to predict the level of the study period, C4.5 prediction model, and accuracy of the C4.5 prediction model in predicting the study period of students.

2. Research Method

According to Kasus et al., (2015) formulates that data mining is a process to obtain useful information from a large set of data. The term data mining is also often referred to as knowledge discovery. The C4.5 algorithm is a classification algorithm that applies decision tree techniques (Dhika, 2015). C4.5 algorithm is a development of the ID3 algorithm. Algorithm ID3 or Iterative Dichotomiser 3 is one of the algorithms used to produce decision trees. The ID3 algorithm uses the concept of information entropy. While the decision tree is a way to classify data based on certain conditions. With the decision tree, it can cluster large datasets into smaller data based on a set of decision rules. Algortitma C4.5 is often used because it can process numbers and discrete data, can handle the value of missing data attributes, resulting in certain rules that can be easily understood.

The research was conducted with the data used is graduation data of alumni from 2012 to 2015. The data will then be processed using the C4.5 data mining method, resulting in a decision tree containing rules to predict the student's study period. The steps that will be taken in this study include data collection, data processing, method testing, and evaluation and validation of results.

The data collection techniques used in this study were to use questionnaires. The questionnaire that will be used for data collection in this study will contain closed questions and questionnaires filled out by respondents online on a google form. In filling out the questionnaire, respondents will be given questions that already have several answers available.

The initial stage in the data analysis process is to determine the attributes that will be used as parameters in the data classification stage. The parameters that have been created will be used to form the decision tree. In the tree, the decision will later contain rules that divide the still homogeneous population into heterogeneous.

From the relationship of the student's study period with the academic data of the student as well as other data obtained from the student's daily life will be taken some attributes which of each attribute affect the student's graduation performance. The attributes to be processed include the relationship of the study period with the year of the class, with the GPA, with a focus on the direction, with gender, with the intensity of working during college, with the type of thesis, with the intensity of the internal organization of the campus, with the intensity of the external organization of the campus, with the ukt group, with the status of the scholarship recipient, with education before college, with the intensity of the hobby, with the intensity of playing the game, with the status of having entered the academic competition, with the status of ever participating in non-academic competitions and with the availability of facilities and infrastructure.

3. Result and Analysis

The data that has been collected through google form with the link https://bit.ly/FormDataAlumniPTIK a total of 73 respondents. The data collected include full name data, year of generation, gender, NIM, pre-college education, corner focus, type of thesis, study period, graduate GPA, scholarship recipients, UKT group, the intensity of side work during college, the intensity of internal campus organizing during college, the intensity of organization/activities of the off-campus community during college, the intensity of gameplay, intensity of hobbies, once participated in academic competitions during college, and availability of facilities and infrastructure during college and final assignment preparation.

The application used to process the collected data is to use the weka application with Weka version 3.8.4. Weka is software that is used to process various machine learning algorithms to process data mining. After entering the collected data file into the Weka application, next is to remove the attributes that will not be processed such as full name and NIM because it is unique.

Before the data enters the test stage, it is necessary to preprocessing to improve the data it, among others by deleting double data and correcting the value of data instruments that do not comply with the provisions. The identification results found 6 data that are double data and subsequently deleted. As well as improvement in the value of GPA instruments in some data that previously used commas was changed to a dot-like 3.51 to 3.51. The next stage is the classification process using the C4.5 method found in the classifier in the tree directory and then using the J48 classification. The data will be processed using the percentage-split test option to get the highest accuracy score and the best decision tree.

Using percentage-split, the tested split percentage value is 10% to 90% percentage. Percentage Split (Fixed or Holdout) is a re-sampling method that leaves out random N% of the original data. The test results are shown in Table 1 as follows:

Table 1. Percentage-split test accuracy

Percentage-split	Accuracy (%)
10	65,15
20	70,68
25	65,45
30	64,70
33	63,26
40	50,00
50	72,22
60	58,62

66	68,00
70	63,63
75	83,33
80	73,33
90	71,42

So if depicted in a graphic form produce figure 1 as follows:

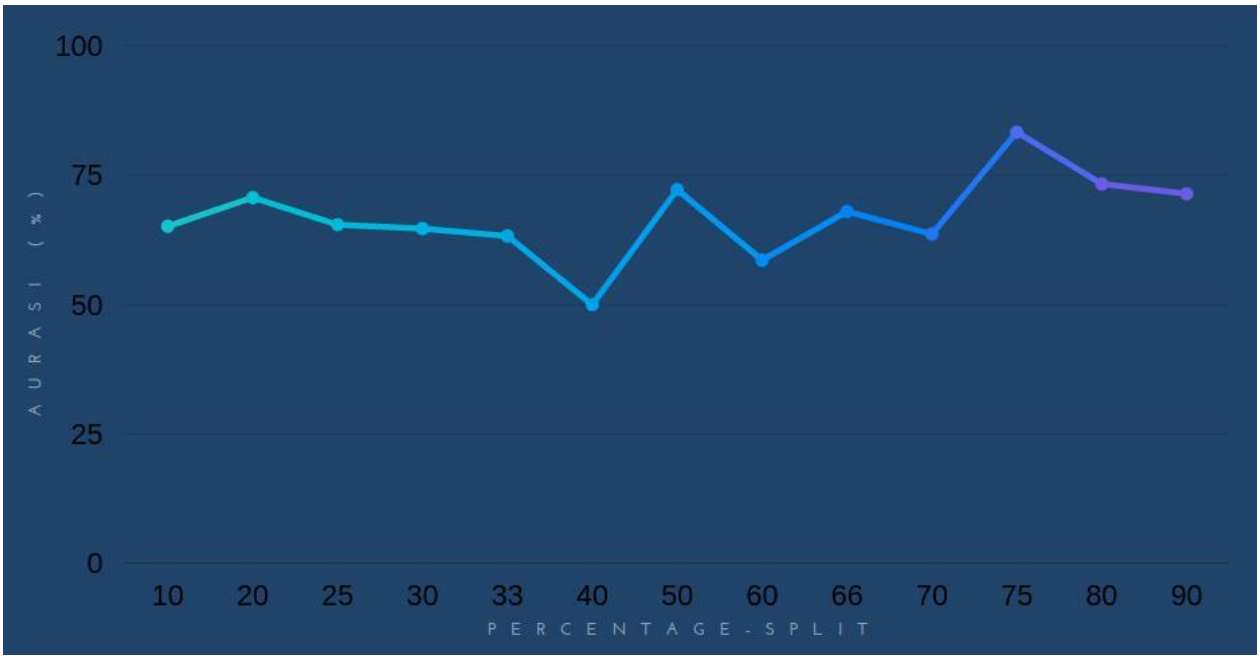

Figure 1. Percentage-split test results graph

From the results of the percentage-split test 75% obtained pruned tree that is loaded in figure 2 as follows:

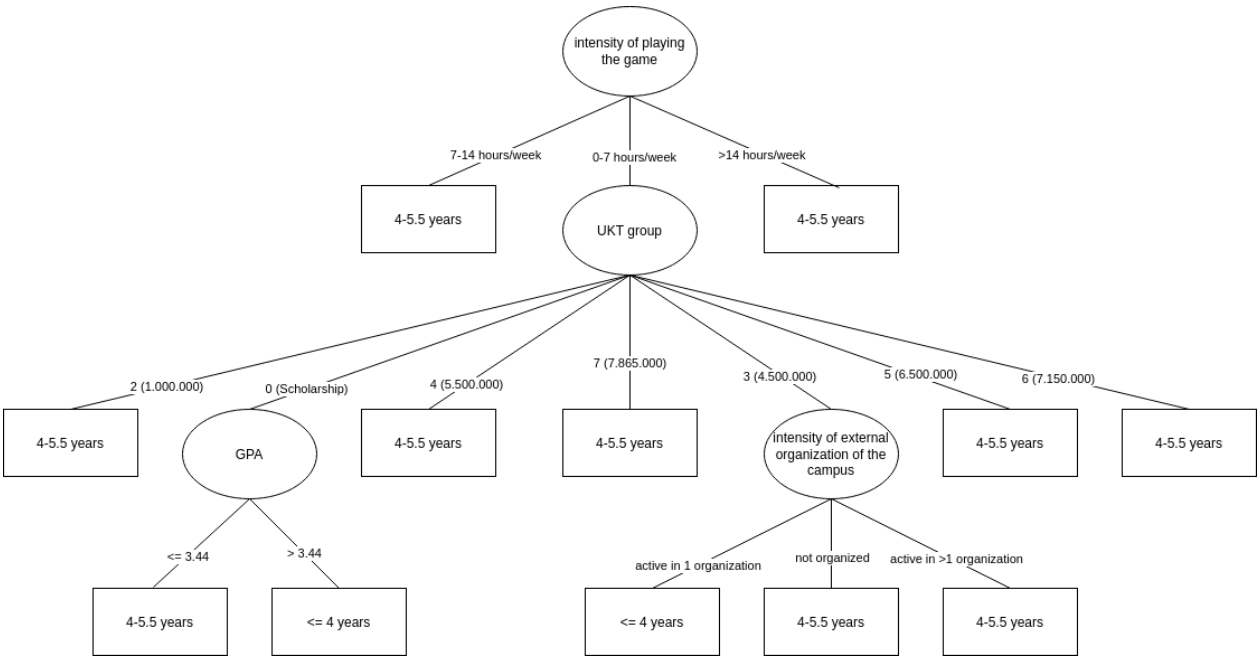

Figure 2. 75% percentage-split decision tree

From the result of processing obtained rules from the pruned tree as follows:

1. If the intensity of playing a game is >14 hours/week then graduated 4-5.5 years
2. If the intensity of playing a game is 7-14 hours/week then graduated 4-5.5 years
3. If the intensity of playing games 0-7 hours/week and group UKT 2 (1,000,000) then graduated 4-5.5 years
4. If the intensity of playing games 0-7 hours/week and group UKT 4 (5,500,000) then graduated 4-5.5 years
5. If the intensity of playing game 0-7 hours/week and group UKT 7 (7,865,000) then graduated 4-5.5 years
6. If the intensity of playing games 0-7 hours/week and class UKT 5 (6,500,000) then graduated 4-5.5 years
7. If the intensity of playing games 0-7 hours/week and class UKT 6 (7,150,000) then graduated 4-5.5 years
8. If the intensity of playing the game is 0-7 hours/week, the UKT 0 group (scholarships) and GPA <= 3.44 then graduated 4-5.5 years
9. If the intensity of playing the game is 0-7 hours/week, the UKT 0 group (scholarships) and GPA > 3.44 then graduated <= 4 years
10. If the intensity of playing games is 0-7 hours /week, the UKT 3 class (4,500,000) and the intensity of external organizations of active campuses in 1 organization then graduated <=4 years
11. If the intensity of playing games 0-7 hours/week, class UKT 3 (4,500,000) and inactive external organizations of the campus then graduated 4-5.5 years
12. If the intensity of playing games is 0-7 hours/week, the UKT 3 class (4,500,000) and the intensity of active campus external organizations in >1 organization then graduated 4-5.5 years

Based on the rules that have been obtained, the instrument of the intensity of playing games becomes the first factor with which it can be concluded that students will graduate on time or not, namely if the intensity of playing > games 7 hours/week then the student will graduate 4-5.5 years. If the intensity of playing the game is < 7 hours/week, the next instrument that will be seen is the UKT class as the second instrument, if it has a group of UKT 2, 4, 5, 6, and 7 then the student will graduate 4-5,5 years. If the intensity

of playing the game < 7 hours/week and the group UKTnya 0 then the third instrument to be reviewed is the GPA. For students with a GPA of < = 3.44 then the study period is 4-5.5 years, while for a GPA of > 3.44 then the study period is < = 4 years, especially for students with the group OF UKT 3, then the third instrument will be reviewed the intensity of external organizations of the campus, if active in 1 organization then it can be concluded the study period < = 4 years, while if a not active organization or active in the > 1 organization then the study period is 4-5.5 years.

The confusion matrix resulting from the test is contained in figure 3 as follows:

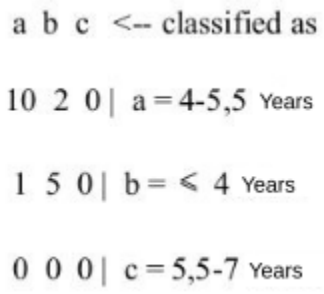

Figure 3. 75% percentage-split confusion matrix

To calculate accuracy using a formula (TP + TN) / (TP+FP+FN+TN), resulting (10+5) / (10+2+1+5) = 83.33%. To calculate Precission using the formula = (TP) / (TP+FP), so that it returns (10) / (10+2) = 83.33%. And to calculate the Recall using the formula = (TP) / (TP + FN), resulting (10) / (10 + 1) = 90.90%.

4. Conclusions and Suggestions

The conclusions that can be inferred from this study are that using the C4.5 algorithm got an accuracy score of 83,33%, a precision percentage of 90,9%, and a recall percentage of 83,3% to predict the study period of students using the decision tree that has been successfully made. The instrument that becomes the seed or the first instrument to be considered is the intensity instrument of playing the game. The second instrument being considered is the value of the UKT. The third instrument that is considered if the UKT class is 0 is the value of the instrument GPA value. The third instrument considered if the UKT group 3 is the value of the intensity instrument of the external organization of the campus.

The Study Program can optimize and improve the curriculum so that students who graduate on time can improve and can use this research to improve the quality of students in lectures and optimize the study period of students. For students to be able to introspect themselves to what factors affect the performance of their study period, such as by reducing the intensity of playing games and optimizing other factors.

Suggestions for further research may improve this research to produce similar research with a better degree of accuracy to predict the study period of students such as by using other algorithms, developing researched instruments and/or multiplying the number of research respondents. This research can be used as a reference for similar research to be done using data that has different characteristics elsewhere. The instruments contained in the decision tree can be evaluated in optimizing the student study period by students and courses.

References

Agustina, D. melina, & Wijanarto. (2016). Analisis Perbandingan Algoritma ID3 Dan C4 . 5 Untuk Klasifikasi Penerima Hibah Pemasangan Air Minum pada PDAM Kabupaten Kendal. *Journal of Applied Intelligent System*, *1*(3), 234–244.

Astuti, I. P. (2017). *Prediksi Ketepatan Waktu Kelulusan Dengan Algoritma Data Mining C4 . 5*. *2*(2).

Dhika, H. (2015). *Kajian Komparasi Penerapan Algoritma C4 . 5 , Naïve Bayes , dan Neural Network dalam Pemilihan Mitra Kerja Penyedia Jasa Transportasi : Studi Kasus CV . Viradi Global Pratama*. 197–202.

Kasus, S., Dehasen, U., Haryati, S., Sudarsono, A., & Suryana, E. (2015). *IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA C4 . 5*. *11*(2), 130–138.

Mustafa, M. S., Ramadhan, M. R., & Thenata, A. P. (2018). Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Creative Information Technology Journal*, *4*(2), 151. https://doi.org/10.24076/citec.2017v4i2.106