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

Application of C4.5 Algorithm with Calculating Entropy-Based on Classification Method for Determining Prediction Graduate Students

Elnandi Nur, Dr. Riza Adrianti Supono., MMSI
Department of Computer Science, Gunadarma University, Jakarta, Indonesia, 16424

Abstract- *The high level of competition in the education world make each individual or university continues to develop talent and ability competency. But the success rate of students is lowered, when students are not able to graduate on time. The time of student graduation becomes a benchmark for student and university success. Ways that can be done to answer these challenges is make data analysis using data mining techniques. University can do this prediction using data mining techniques. This research uses the C4.5 algorithm to construct a decision tree. Conducted a research of 115 active students in one of the state universities of Jakarta, the results of this research is, there were 15 students who are predicted not be able to graduate on time.*

Keywords- *Datamining, Prediction, Algorithm C4.5, Calculating Entrop, Classification Method.*

I. INTRODUCTION

Planned education, directed and sustained can help students to develop their skills optimally, both cognitive, affective, and psychomotic aspects. Therefore, the ability of students is important to be a benchmark fot the success of students themselves.

A. Data Mining

Data mining allows organizations to use existing report management capabilities to find and understand hidden patterns (hidden patterns) in a large database. Then these patterns are built into the model and the data used to predict an event. With this understanding, institution can allocate available resources and staff more effectively, for example by the information from data mining obtained more accurate estimate of how many students will take a certain course so that the allocation of its resources can be more effective.

B. Development

The rapid development of data mining is inseparable from the development of information technology which causes to the accumulation of data in a very large number or data explosion. But the rapid growth of data accumulation it has created the conditions in which many of us have a lot of data but have little information.

C. Definition of Data Mining

A simple definition of data mining is the extraction of information or interesting information or patterns from existing data in large databases. But the database is not the only field of science that affect data mining, many more fields of science that enriched data mining such as: statistics, neural networks (neural network), machine learning, information science, mathematics, visualization and much more.

In general, the C4.5 algorithm to construct a decision tree is as follows.

1. Select an attribute as root
2. Create a branch for each value
3. For cases in branch
4. Repeat the process for each branch until all cases the branches have the same class.

To select an attribute as roots, is based on the highest gain nial of existing attributes.

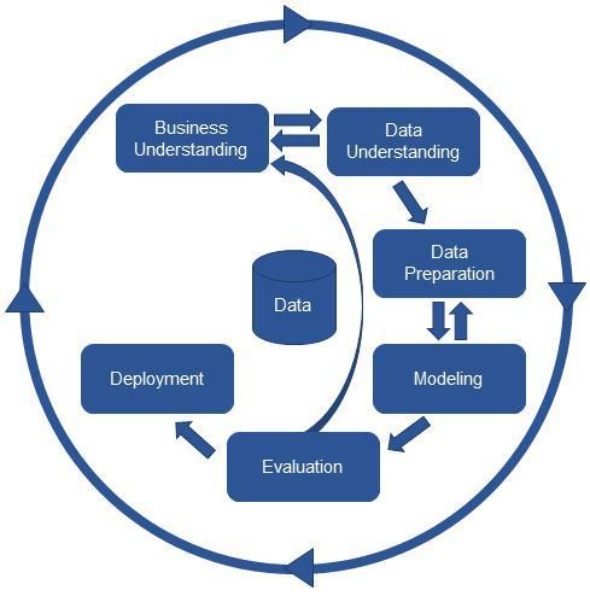
II. RELATED WORKS

The research approach used is a quantitative approach, which is by considering that quantitative research is a method for testing certain theory by examining the relationship between variables. According to Cresswell (1994) Quantitative research is an investigation into social issues based on testing the theory with variables that can be analyzed by numbers and statistics. The methods of quantitative research generally involves the collecting, analysis, and interpretation of data, as well as writing the results of research. Quantitative research is research that examines theory objectively. These variables can be measured in an instrument, so some data can be analyzed using statistical procedures.

The methodology used in developing data mining is CRISP-DM (Cross Industry Standard Process for Data Mining). CRISP-DM is a method established by the European Commission in 1996 which applies standards in the data mining process. According Kusrini and Rogan (2009) is the standard data mining process as strategy general troubleshooting of business or for research that has a life cycle is divided into six phases:

Elnandi Nur, and Dr. Riza Adrianti Supono., MMSI "Application of C4.5 Algorithm with Calculating Entropy-Based on Classification Method for Determining Prediction Graduate Students," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 2, pp. 218-221, 2020.

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



III. THE GRADUATE STUDENTS DATA SET AND PREPROCESSING

This study uses secondary data obtained from the student database owned by the University of Indonesia, namely SIAK NG. The collected data is data from FIA UI student undergraduate courses (S1) class of 2016-2018. The data consists of 20 attributes and one attribute outcome predictor. The attributes that became parameters in this study are:

Attribute	Status of Usage Details	
NIM	√	ID
Gender	√	value Model
Student Status	√	value Model
Marital Status	√	value Model
Age	√	value Model
IPS1	√	value Model
IPS2	√	value Model
IPS3	√	value Model
IPS4	√	value Model
IPS5	√	value Model
IPS6	√	value Model
IPS7	√	value Model
IPS8	√	value Model
SKS1	√	No.
SKS2	√	No.
SKS3	√	No.
SKS4	√	No.
SKS5	√	No.
SKS6	√	No.
SKS7	√	No.
SKS8	√	No.
Total Credits	√	No.
GPA	√	value Model
Pass On Time	√	Target label

The table above describes the attributes that will be used in research, the indicator yes (√) indicates that the attribute in question will be used in the study, whereas no indicator (×) indicates this attributes will be eliminated at the data preparation stage.

Data with the attributes that have been selected are then converted to facilitate the process of mining, because the data will be processed with the help of data mining tools. Here are the attributes that the conversion process will do.

a. Age

The age attributes has an integer data type value with many value will be converted into two categories: value "≥ 20" and "<20".

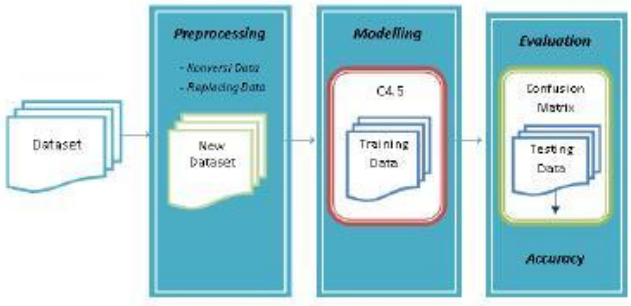
b. IPS

Attributes IPS1, IPS2, IPS3, IPS4 previously it contained student achievement index in each semester in float data type with many grades and was converted into 3 categories value "0 - 1:30", "1:31 - 2.60" and "2.61 - 4:00".

c. SKS

Attributes SKS1, SKS2, SKS3 AND SKS4 previously contained semester credit units are taken each semester by the students in the integer data types with many value then be converted into two categories: value "≥20" and "<20".

The method that will be used in this research is the algorithm C4.5. To measure accuracy in this research will use a framework RapidMiner Studio 6.0.003:



The purpose of this study is to analyze the predictions of the accuracy of students graduation by applying data mining classification techniques with the decision tree algorithm C4.5.

Researchers take a few steps in building the algorithm C4.5 decision tree as follows:

1. Calculate the amount of data based on member attributes results with certain conditions. In the first process requirements are still empty.
2. Select Attributes Node
3. Create a branch for each - each member of the Node
4. Checkwether there is zero entropy from Node member. If so, determine the leaves formed. If all entropy value for Node is zero, then the process stops.
5. If there are member node that have an entropy value greater than zero, repeat the process again from the beginning with the node as a condition until all the members of Node are zero.

Node is an attribute that has the highest gain value from the existing attributes.

Elnandi Nur, and Dr. Riza Adrianti Supono., MMSI "Application of C4.5 Algorithm with Calculating Entropy-Based on Classification Method for Determining Prediction Graduate Students," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 2, pp. 218-221, 2020.

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

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

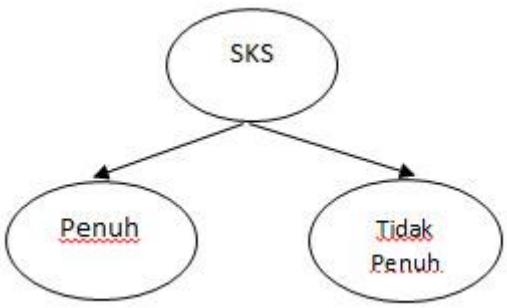
Information:
S = set case
A = attribute
n = number of partitions attribute A
|Si| = proportion of Si to S

IV. Application of Data Mining Techniques to Graduate Students Dataset: Result And Discussion

The analysis in this study requires data on GPA and IP values from semesters 1 to 8 that researchers get from faculty. There are 115 samples analyzed. Attachment 1 is a sample of raw data that will be extracted in according with the method of classification algorithm C4.5 After getting the sample data, then the process of calculating the amount of data, entropy, and the gain. These results are in the table below,

Node 1	Jumlah (S)	Tepat Waktu	Tidak Tepat Waktu	Entropy	Gain
Total	115	100	15	0,558629373	
Jenis Kelamin					-0,021301216
Laki-laki	36	30	6	0,650022422	
Perempuan	79	69	10	0,547990008	
Status Mahasiswa					0,002709729
Mahasiswa	111	97	14	0,546717537	
Bekerja	4	3	1	0,811278124	
Umur					0,020952155
<20	111	98	13	0,521016938	
>20	4	2	2	1	
IPK					0,036788247
Tinggi	112	99	13	0,517961871	
Sedang	1	0	1	0	
Rendah	2	1	1	1	
Totak SKS					0,088894067
Penuh	105	96	9	0,422000517	
Tidak Penuh	10	4	6	0,970950594	

In the table above shows that the highest gain value is the total credits, compared to gender attribute, student status, age, and GPA. Then SKS becomes root. As in the picture below,



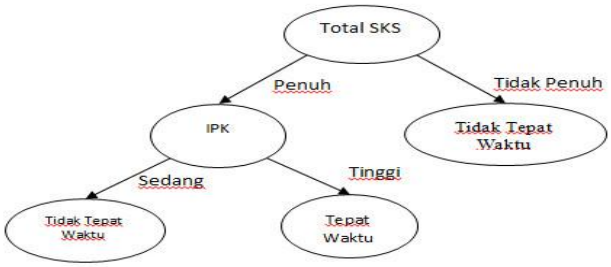
The picture above shows that the SKS is a root that have two members, which is full and not full, as well as its feasibility, there are two decisions, which is Pass or Do not Pass. Full entrophy value and not full have a value, so the result from the tree above is still questionable for the eligibility, whether passed or not passed.

Node SKS Penuh	Jumlah (S)	Tepat Waktu	Tidak Tepat Waktu	Entropy	Gain
Total SKS					
Penuh	105	96	9	0,422000517	
IPK					0,0344830824
Tinggi	104	96	8	0,391243564	
Sedang	1	0	1	0	
Rendah	0	0	0	0	
Umur					0,0104679424
<20	102	94	8	0,396627773	
>20	3	2	1	0,918295834	
Status Mahasiswa					0,00692715667
Mahasiswa	101	93	8	0,399382082	
Bekerja	4	3	1	0,811278124	
Jenis Kelamin					0,00102618022
Laki-laki	29	27	2	0,362051252	
Perempuan	76	69	7	0,443458145	

Node SKS Tidak Penuh	Jumlah (S)	Tepat Waktu	Tidak Tepat Waktu	Entropy	Gain
Total SKS					
Tidak Penuh	10	4	6	0,970950594	
IPK					-0,029049406
Tinggi	8	4	4	1	
Sedang	2	1	1	1	
Rendah	0	0	0	0	
Umur					0,574520171
<20	9	4	5	0,99107606	
>20	1	0	1	0	
Status Mahasiswa					0,089659695
Mahasiwa	10	3	7	0,881290899	
Bekerja	0	0	0	0	
Jenis Kelamin					0,005802149
Laki-laki	7	3	4	0,985228136	
Perempuan	3	1	2	0,918295834	

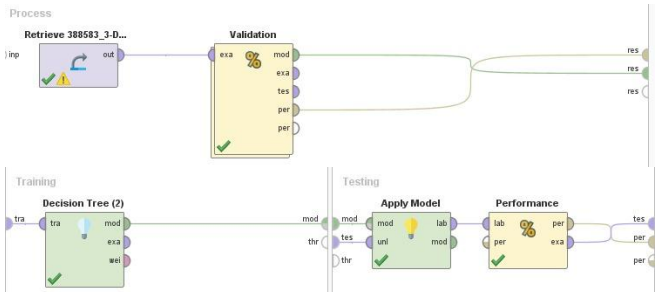
Explanation of the above table is that the table next node is SKS sold and not sold. Then repeat calculating the entropy and gain value. Then there exists among the attributes contained in the table, the highest and influence are the IPK, compared to other attribute such as achievement and ethics. After getting the next decision attribute, the decision tree is made again which starts from if SKS is full and the GPA is high, it will be declared PASSED, the if the SKS is full and the GPA is medium, it will be declared PASSED, and if the SKS is full while the GPA is low, it will be declared NOT PASSED. Consider the following decision tree. If the SKS IS not full then it does not pass,

Elnandi Nur, and Dr. Riza Adrianti Supono., MMSI "Application of C4.5 Algorithm with Calculating Entropy-Based on Classification Method for Determining Prediction Graduate Students," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 2, pp. 218-221, 2020.



The decision tree only until the GPA, because the value between the pass and do not pass is 0, then the feasibility of the decision is immediately obtained. While other variables such as student status, gender, and age are not include affecting student graduation on time.

The next step is to test the sample data in the form of tables that are in excel through the tools RapidMiner 5 starts from the connection process between the sample data base, operator and validation as shown below:



A. Result

Based on the conclusions of the result of RapidMiner above, students who get GPA more than 3.535 in semesters 2, the number of students who graduating on time by 57 students, if in this semester below or equal to 3,535 can pass if the value of IPS 1 above 3,255. GPA semesters 6 above 3,838 who graduated on time were 3 students, if in this semester below or equal to 3.610 the students who did not graduate on time were students. The GPA in semester is below or equal to 3.520, then there are 39 students who graduate on time and 1 student who not graduate on time. The GPA in semester 1 is below 3,255 students who not graduate on time are students. Total credits if below or equal to 137, students who graduate on time were 1 student and those who did not graduate on time were 8 students.

The most influential variable on the time of graduation of students is GPA and Semester Credit Units (SKS) are taken by students.

Based on the predictions of the faculty and university can carry out some several policies for students who are predicted not pass.

V. Conclusion and Future Work

The conclusion of this research using the algorithm C4.5 shows that:

1. Data Mining System on C4.5 algorithm can be used to predict the timely of graduation in university.
2. The factors that most influence student graduation are SKS and GPA.
3. The test results on the RapidMiner 5 tools produce accuracy with C4.5 algorithm performance level of 86.82% +/- 8.94% (micro average: 86.96%).
4. A total of 115 active students under study can be predicted the students who graduate on time are 100 students and 15 students predicted not to graduate on time.

Researchers provide suggestions based on the result of the analysis on the application C4.5 algorithm.

1. Students who do not yet have a full credits SKS can be tolerance and guidance in order to graduate on time.
2. Students who are predicted to not be able to graduate on time given direction and motivation to be able to improve their, so that the target of graduation on time can be achieved.

References

[1] Triyanto, E., Anitah, S., & Suryani, N. (2013). Peran Kepemimpinan Kepala Sekolah dalam Pemanfaatan Media Pembelajaran Sebagai Upaya Peningkatan Kualitas Proses Pembelajaran. Teknologi Pendidikan, 1(2), 226-238..

[2] A. Sri Padmini, " Analysis of Students Graduation Time With CHAID method," the e-Journal of Mathematics, vol. 1, no. 1, pp. 89-93, 2012.H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.

[3] IH Witten and E. Frank, Data Mining. 2005.

[4] NM Huda, "Data Mining Applications to Displays Information on Student Graduation rates," 2010.C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[5] Dawn, Ardi S. " Study on The Application Naie Bayes and C4.5 in Predicting product offerings at XYZ Bank Tbk." Budi Luhur Univerisity: Jakarta, 2014M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.

[6] Tai, M. Mohammed Abu and El-halees, Alaa M., "Education Mining Data to Improve Students' Performance: A Case Study". International Journal of Information and Communication Technology Research, Vol.2, No.2, 140-146, 2012.

[7] Turban, et al. "Decision Support System and Intelligent System". Prentice-Hall of India: New Delhi, 7th ed, 2007.

[8] Han, et al. "Data Mining Concepts and Techniques". Springer: Heidelberg, Vol.12, 2011.