Classification Analysis Using C4.5 Algorithm To Predict The Level of Graduation of Nurul Falah Pekanbaru High School Students

Fana Wiza[1], *Bayu Febriadi*[1]

[1]*Department of Information System, Faculty of Computer Science, Lancang Kuning University*
JYos Sudarso Km 8 Pekanbaru, Telp. (+628117532015)
[1]*fana@unilak.ac.id*, [2]*bayufebriadi@gmail.com*

Abstract

School as one of the processes for implementing formal education is required to carry out the learning process optimally to produce quality students. Regarding the research process carried out to predict the graduation rate of SMA Nurul Falah students by using the decision tree method. The data used in this study are student data using the criteria for student names, majors, average report cards from semester one (I), two (II), three (III), four (IV), five (V), and the average value of the National Standard School Examination (USBN). The data is then managed using Rapidminer 5.3 software to make it easier to predict student graduation rates. The application of data mining is used to predict the graduation rate by using the decision tree method and C4.5 algorithm as a supporter as well as to find out information on the graduation rate of Nurul Falah High School students. This study aims to predict student graduation rates in order to get useful information and the school can make policies in the coming year.

Keywords: *Decision Tree, data mining and C4.5 Algorithm*

1. Introduction

Education is an effort to develop highly capable human resources (HR) and also beneficial for the development that is being carried out in our country. This is in accordance with the objectives of national education, namely to educate the nation's life and develop Indonesian people as a whole, devoted to the Almighty God and virtuous character, possessing knowledge and skills, physical and spiritual health, a solid and independent personality and social and national responsibility [1][2].

School as one of the processes for implementing formal education is required to carry out the learning process optimally to produce quality students. In the 2016/2017 school year, SMA Nurul Falah Pekanbaru has 103 students of class XII (twelve), calculated from two majors namely Science and Social Sciences. In each class there are approximately 30 (thirty) students. So from that every year the school must pass these students, so the school always worries about the level of graduation of their students. For now the National Standardized School Examination (USBN) is the last stage in determining the graduation and failure of students while the Computer Based National Examination (UNBK) does not play a role in determining student graduation again, only as a formality[3-7].

Failure of students in facing examinations can hinder the student's graduation process, so graduation that is uncertain will cause anxiety for students of Nurul Falah High School Pekanbaru and teachers. For the school, especially the teachers, the student graduation rate is a top priority because it involves school accreditation. For that, the right strategy is needed to boost the graduation rate of Nurul Falah High School students. One of them is by predicting students who pass and who do not pass. Thus, teachers can concentrate more on improving the teaching and learning process especially for students who are predicted not to graduate.

Copyright ⓒ 2019

2. Rudimentary

In this study predicting the graduation rate of students at Pekanbaru Nurul Falah High School using data mining applications with the decision tree classification method, namely C4.5 algorithm and also using the Rapidminer 5.3 tool so that later it can help the school in predicting the graduation rate of students..

3. Research Methodology

Pekanbaru Nurul Falah High School. In this case the data mining method used is Classification. Where the process classifies the existing student data set. The classification model used is the Decision Tree approach. The algorithm used as the decision tree forming algorithm is the C4.5 algorithm. In this study the author uses existing data mining applications, namely Rapidminer 5.3. The stages that will be carried out in this study are as follows:

a. Describe the Problem
   Describe the problem or provide an explanation of the problems to be studied, structured or systematically to solve problems and to make a better decision.
b. Analysis of the problem
   Problem analysis is a step to be able to understand the problem that has been determined in the scope or limitations. By analyzing the predetermined problem, it is expected that the problem can be well understood and the solution can be obtained to the maximum and with a suitable method.
c. Studying Literature
   After the problem is analyzed, the literature is studied that deals with the problem. Then the literature studied was selected to determine which literature would be used in the study. Literature sources are obtained from libraries, journals, articles that discuss data mining, C.45 algorithms, Decision Tree and other supporting concepts in completing this research.
d. Data Collection
   In conducting this research, collecting data and information at this stage is done to find out about the system under study. From the data and information collected will be obtained data for research supporters.
e. Selection of techniques used
   This stage aims to determine the techniques used in system design. The method used is the classification of Decision Tree by changing large facts into decision trees that present the rules. Rules can be easily understood with natural language. Decision trees are also useful for exploring data, finding hidden relationships between a number of potential input variables with target variables. The algorithm used is C.45

4. Results and Discussion

Conducting a Pre-Process At this stage the stages of data mining will be explained. Steps taken in the pre-process phase of this data are data cleaning (data cleaning), data integration (data integration), data selection (data selection), data transformation (data transformation).

a. Data Cleaning

At this stage, part of the 2016/2017 student year data cleaning will be carried out with 103 students consisting of two majors, namely Science, Social Sciences and USBN. The number of students majoring in Natural Sciences were 46 students, while the number of IPS students was 57 students. Data previously taken at the research site is not necessarily all the data is complete and directly used in the process of data mining, therefore incomplete data is not included in this study. Complete data selected several attributes that will be processed as needed. To facilitate the cleaning of data, IPA data and IPS data are separated because the data was previously integrated. The following is the original data of students majoring in Natural Sciences and Social Sciences before the data cleaning stage is shown in the table below.

In the data of the original students majoring in Science there are several attributes such as student identity, study code, the value of each per subject there are grades of Indonesian, English, Mathematics, Physics, Chemistry, Biology, Average scores per semester and Remarks. In the identity attribute there are student exam numbers and date of birth that may not be used during subsequent analysis. For data that is processed into data mining attributes, not all of them are used and should be deleted. Because the conditions for determining graduation in a research location are seen from the average student report card and the USBN value. In the data of the original IPS students there are also several attributes such as Student Identity, Code of study, the value of each subject, namely Indonesian, English, Mathematics, Economics, Sociology, Geography, Average scores per semester and Remarks. The 2016/2017 academic year data is used as a data set, then determines the attributes that will be used as data sets. The attributes used are the names of students, majors, average semester scores one (I) to five (V), USBN average values and information (Class). The following is the student data table for each department after data cleaning analysis.

Table 1. Data on the Natural Sciences Department in 2016/2017 after the Cleaning Stage

No	Name	Dept	Rt2 Smt 1	Rt2 Smt 2	Rt2 Smt 3	Rt2 Smt 4	Rt2 Smt 5	Average USBN	Class
1	Angel Bertha Panggabean	IPA	76.92	79.17	85.00	87.17	90.17	86.33	Pass
2	Anita Laila	IPA	78.29	78.33	83.33	85.17	88.83	85.00	Pass
3	Annisa	IPA	75.54	75.00	78.67	77.83	83.00	82.67	Failed
4	Apriandi Said Tanjung	IPA	76.92	76.67	79.83	80.50	82.17	81.83	Failed
5	Ayu Wulandari	IPA	79.13	75.83	82.33	82.33	86.67	83.17	Pass
...	...	...	...	...	...	...	...	...	....
46	Yogi Indracaya	IPA	78.29	79.23	81.67	81.33	86.50	84.50	Pass

Table 4. Data on Social Sciences Department in 2016/2017 after the Cleaning Stage

No	Name	Dept	Rt2 Smt 1	Rt2 Smt 2	Rt2 Smt 3	Rt2 Smt 4	Rt2 Smt 5	Average USBN	Class
1	Ahsani Taqwim	IPS	78.29	77.75	79.33	78.67	82.50	82.17	Failed
2	Ansar	IPS	74.17	78.29	79.83	80.00	83.67	82.50	Failed
3	Bagus Kurniawan	IPS	79.67	79.13	83.83	83.83	83.17	84.50	Pass
4	Carissa Amanda	IPS	78.29	83.25	79.17	80.00	81.67	82.50	Failed
5	Dinda Kusuma Dewi	IPS	79.67	81.88	79.50	83.00	85.17	84.67	Pass
...	...	...	...	...	...	...	...	...	...
57	Yulia Novriamirda Warni	IPS	79.13	84.63	82.00	82.83	83.50	84.00	Pass

In the tables 1 and 2 this is the result of a data cleaning process where there are only a few attributes, namely the name of the student, department, average semester one (I) to semester five (V), USBN average value and description (Class).

b. Data Integration

Not infrequently the data obtained for data mining not only comes from a database but also comes from several databases or text files. the combined table is between the majors of science and social studies, and the value of USBN then the two tables are integrated into one table, the results of integration in table 3 below are shown.

Table 3. Integration of Science and Social Sciences Departments

No	Name	Dept	Rt2 Smt 1	Rt2 Smt 2	Rt2 Smt 3	Rt2 Smt 4	Rt2 Smt 5	Average USBN	Class
1	Angel Bertha Panggabean	IPA	76.92	79.17	85.00	87.17	90.17	86.33	Pass
2	Anita Laila	IPA	78.29	78.33	83.33	85.17	88.83	85.00	Pass
3	Annisa	IPA	75.54	75.00	78.67	77.83	83.00	82.67	Failed
4	Apriandi Said Tanjung	IPA	76.92	76.67	79.83	80.50	82.17	81.83	Failed
5	Ayu Wulandari	IPA	79.13	75.83	82.33	82.33	86.67	83.17	Pass
...	...		...	...	...	...	...	...	...
103	Yulia Novriamirda Warni	IPS	79.13	84.63	82.00	82.83	83.50	84.00	Pass

c. Data Selection (Data Selection)

In the data selection phase the number of attributes is also determined for the analysis process. Previously there were name and department attributes and therefore the name and department attributes were not used in the mining process.

d. Data Transformation (Data Transformation)

This stage is the stage where the data is converted into a format or form that is suitable for data mining. Because the data is in accordance with the desired format, which is stored in the excel data format (. Xls), it is not necessary to transform the data. In the next stage, the grouping of values is based on the data selected for the analysis process. The following are data that have been initialized into the letter form in table 4 and table 5 below.

Table 4. Semester Value Category

Semester	**Classification**
90-100	A
80-89	B
70-79	C
60-69	D
0-59	E

Table 4. Semester Value Category

Average USBN	**Classification**
90-100	A
80-89	B
70-79	C
60-69	D
0-59	E

The pre-process results consist of six attributes and one class (label), namely the first semester (I), two (II), three (III), four (IV), five (V), USBN averages and classes. (label). In table 5.9 below is a display of some pre-process data that has been done. The final data format is obtained based on attribute values that have been grouped or classified, for example the average semester one (I) value is "76.92" then after being classified as "C", the average semester two (II) is "80.00" then after classified into "B", and so on.

3.1 Process (Classification Analysis with C4.5 Algorithm)

Calculation of Gain and Entropy To find the Gain value, use the formula:

$$Gain\ (S,A) = \ Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} \times Entropy\ (Si) \qquad (1)$$

Information:

S	: Case Set
A	: Attribute
N	: Number of partition attributes A
| Si |	: Number of cases on the i-partition
| S |	: Number of cases in S

Whereas to see Entropy can be seen in the formula:

$$Entropy\ (S) = \sum_{i=1}^{n} - Pi \times \log_2 Pi \qquad (2)$$

Information:

S	: Set of cases
A	: Features
n	: Number of partitions s
Pi	: Proportion of Si to S

By using the two equations above, it is found that Entropy and Gain are used as roots in making decision trees.

Entropy(Total) $=(-\frac{10}{20} * \log_2(\frac{10}{20}))+(-\frac{10}{20} * \log_2(\frac{10}{20}))= 1$

Entropy(Total) =1

Entropy **SMT 1**

Entropy (Total **,** B) $= (-\frac{2}{3} * \log_2(\frac{2}{3}))+(-\frac{1}{3} * \log_2(\frac{1}{3}))= 0{,}9182958$

Entropy (Total , C) $= (-\frac{8}{17} * \log_2(\frac{8}{17}))+(-\frac{9}{17} * \log_2(\frac{9}{17}))= 0{,}997502546$

Entropy **SMT 2**

Entropy (Total **,**B) $= (-\frac{4}{7} * \log_2(\frac{4}{7}))+(-\frac{3}{7} * \log_2(\frac{3}{7})) = 0{,}985228136$

Entropy (Total **,**C) $= (-\frac{6}{13} * \log_2(\frac{6}{13}))+(-\frac{7}{13} * \log_2(\frac{7}{13})) = 0{,}995727452$

Entropy **SMT 3**

Entropy (Total **,** B) $= (-\frac{8}{10} * \log_2(\frac{8}{10}))+(-\frac{2}{10} * \log_2(\frac{2}{10})) = 0{,}721928095$

Entropy (Total **,** C) $= (-\frac{2}{10} * \log_2(\frac{2}{10}))+(-\frac{8}{10} * \log_2(\frac{8}{10})) = 0{,}721928095$

Entropy **SMT 4**

Entropy (Total, B) $= (-\frac{10}{16} * \log_2(\frac{10}{16}))+(-\frac{6}{16} * \log_2(\frac{6}{16})) = 0{,}954434003$

Entropy (Total **,**C) $= (-\frac{0}{4} * \log_2(\frac{0}{4}))+(-\frac{4}{4} * \log_2(\frac{4}{4})) = 0$

Entropy **SMT 5**

Entropy (Total, A) $= (-\frac{1}{1} * \log_2(\frac{1}{1}))+(-\frac{0}{1} * \log_2(\frac{0}{1})) = 0$

Entropy (Total **,**B) $= (-\frac{9}{19} * \log_2(\frac{9}{19}))+(-\frac{10}{19} * \log_2(\frac{10}{19})) = 0{,}998000884$

Entropy **USBN**

Entropy (Total, B) $= (-\frac{10}{20} * \log_2(\frac{10}{20}))+(-\frac{10}{20} * \log_2(\frac{10}{20})) = 1$

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

Gain(Total, SMT 1) = 1 - $\left(\frac{3}{20} * 0{,}9182958\right) + \left(\frac{17}{20} * \ 0{,}995727452\right) = 0{,}0143785$

Gain(Total, SMT 2) = 1- $\left(\frac{7}{20} * 0{,}985228136\right) + \left(\frac{13}{20} * 0{,}995727452\right)$ =0,0079473

Gain(Total, SMT 3) =1– $\left(\frac{10}{20} * \ 0{,}721928095\right) + \left(\frac{10}{20} * 0{,}721928095\right)$ =0,278072

Gain(Total, SMT 4) = 1 – $\left(\frac{16}{20} * \ 0{,}954434003\right) + \left(\frac{4}{20} * 0\ \right)$ =0,236453

Gain(Total, SMT 5) $= 1 - \left(\frac{1}{20} * 0\right) + \left(\frac{19}{20} * 0{,}998000884\right) = 0{,}051899$

Gain(Total, USBN) $= 1 - \left(\frac{20}{20} * 1\right) = 0$

From it can be seen that the attribute with the highest Gain is Semester 3 which is 0.278072, for the values of B and C the same so it needs to be done further calculations. The decision tree formed until this stage is shown in the image below.

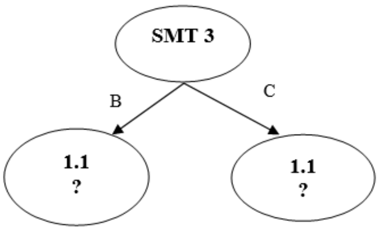

Figure 1. Tree of Decision on Node Calculation Results 1

Calculating the number of cases, the number of cases for passing decisions, the number of cases for decisions not passed, and the entropy of all cases and cases divided by semester 1, 2, 4, 5 and USBN attributes which can be root nodes of attribute values are B and C semester 3. After that, do the Gain calculation for each attribute. The total entropy line is as follows.

Entropy (Total **,** SMT 3 B) $= (-\frac{8}{10} * \log_2(\frac{8}{10})) + (-\frac{2}{10} * \log_2(\frac{2}{10})) =$ 0.72192805

***Entropy* SMT 1**

Entropy (SMT 3 B**,** B) $= (-\frac{2}{3} * \log_2(\frac{2}{3})) + (-\frac{1}{3} * \log_2(\frac{1}{3})) = 0{,}918295834$

Entropy (SMT 3 B, C) $= (-\frac{6}{7} * \log_2\left(\frac{6}{7}\right)) + (-\frac{1}{7} * \log_2(\frac{1}{7})) = 0{,}591672779$

***Entropy* SMT 2**

Entropy (SMT 3 B**,** B) $= (-\frac{2}{4} * \log_2(\frac{2}{4})) + (-\frac{2}{4} * \log_2(\frac{2}{4})) = 1$

Entropy (SMT 3 B**,** C) $= (-\frac{6}{6} * \log_2(\frac{6}{6})) + (-\frac{0}{6} * \log_2(\frac{0}{6})) = 0$

***Entropy* SMT 4**

Entropy (SMT 3 B, B) $= (-\frac{8}{9} * \log_2(\frac{8}{9})) + (-\frac{1}{9} * \log_2(\frac{1}{9})) = 0{,}503258335$

Entropy (SMT 3 B**,** C) $= (-\frac{0}{1} * \log_2(\frac{0}{1})) + (-\frac{1}{1} * \log_2(\frac{1}{1})) = 0$

***Entropy* SMT 5**

Entropy (SMT 3 B, A) $= (-\frac{1}{1} * \log_2(\frac{1}{1})) + (-\frac{0}{1} * \log_2(\frac{0}{1})) = 0$

Entropy (SMT 3 B**,** B) $= (-\frac{7}{9} * \log_2(\frac{7}{9})) + (-\frac{2}{9} * \log_2(\frac{2}{9})) = 0{,}764204507$

***Entropy* USBN**

Entropy (SMT 3 B, B) $= (-\frac{8}{10} * \log_2(\frac{8}{10})) + (-\frac{2}{10} * \log_2(\frac{2}{10})) = 0{,}721928095$

Gain(SMT 3 B, SMT 1) = *Gain(S, A)* = *Entropy(S)*-$\sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Sn)$

$= 0.72192805 - (\left(\frac{3}{10} * 0{,}918295834\right) + \left(\frac{7}{10} * 0{,}591672779\right)$

$= 0{,}0812619$

Gain (SMT 3 B, SMT 2) = *Gain(S, A)* = *Entropy(S)*-$\sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Sn)$

$= 0.72192805 - (\left(\frac{4}{10} * 1\right) + \left(\frac{6}{10} * 0\right)$

$= 0{,}3219281$

Gain (SMT 3 B, SMT 4) $= Gain(S, A) = Entropy(S)\text{-}\sum_{i=1}^{n} \frac{|si|}{|s|} * Entropy(Sn)$
$= 0.72192805 \text{-}(\left(\frac{9}{10} * 0{,}503258335\right) + \left(\frac{1}{10} * 0\right))$
$= 0{,}268996$

Gain (SMT 3 B, SMT 5) $= Gain(S, A) = Entropy(S)\text{-}\sum_{i=1}^{n} \frac{|si|}{|s|} * Entropy(Sn)$
$= 0.72192805 \text{-}(\left(\frac{1}{10} * 0\right) + \left(\frac{9}{10} * 0{,}764204507\right))$
$= 0{,}034144$

Gain (SMT 3 B, USBN) $= Gain(S, A) = Entropy(S)\text{-}\sum_{i=1}^{n} \frac{|si|}{|s|} * Entropy(Sn)$
$= 0.72192805 \text{-}(\left(\frac{10}{10} * 0{,}721928095\right)$
$= 0$

From the results it can be seen that the attribute with the highest Gain is semester 2, which is equal to 0.3219281. There are two attribute values from semester 2, namely B and C. From the two attribute values, the value of attribute C already classifies the case, namely the decision to pass, but for the value of attribute B it still needs to be calculated again. thus semester 2 can be a branch node of attribute value of semester 3 B. then the attribute value is calculated from semester 3 C. The total entropy line is calculated by the following equation.

Entropy (Total **,** SMT 3 C) $= (-\frac{2}{10} * \log_2(\frac{2}{10}))+ (-\frac{8}{10} * \log_2(\frac{8}{10}))= 0.72192805$

***Entropy* SMT 1**

Entropy (SMT 3 C**,** B) $= (-\frac{0}{0} * \log_2(\frac{0}{0})) + (-\frac{0}{0} * \log_2(\frac{0}{0}))= 0$

Entropy (SMT 3 C, C) $= (-\frac{2}{10} * \log_2\left(\frac{2}{10}\right)) + (-\frac{8}{10} * \log_2(\frac{8}{10}))= 0.72192805$

***Entropy* SMT 2**

Entropy (SMT 3 C**,** B) $= (-\frac{2}{3} * \log_2(\frac{2}{3})) + (-\frac{1}{3} * \log_2(\frac{1}{3}))= 0{,}918296$

Entropy (SMT 3 C, C) $= (-\frac{0}{7} * \log_2\left(\frac{0}{7}\right)) + (-\frac{7}{7} * \log_2(\frac{7}{7}))= 0$

***Entropy* SMT 4**

Entropy (SMT 3 C**,** B) $= (-\frac{2}{7} * \log_2(\frac{2}{7})) + (-\frac{5}{7} * \log_2(\frac{5}{7}))= 0{,}863121$

Entropy (SMT 3 C, C) $= (-\frac{0}{3} * \log_2\left(\frac{0}{3}\right)) + (-\frac{3}{3} * \log_2(\frac{3}{3}))= 0$

***Entropy* SMT 5**

Entropy (SMT 3 C**,** A) $= (-\frac{0}{0} * \log_2(\frac{0}{0})) + (-\frac{0}{0} * \log_2(\frac{0}{0}))= 0$

Entropy (SMT 3 C, B) $= (-\frac{2}{10} * \log_2\left(\frac{2}{10}\right)) + (-\frac{8}{10} * \log_2(\frac{8}{10}))= 0{,}721928095$

***Entropy* USBN**

Entropy (SMT 3 C**,** B) $= (-\frac{2}{10} * \log_2(\frac{2}{10})) + (-\frac{8}{10} * \log_2(\frac{8}{10}))= 0{,}721928095$

Gain(SMT 3 C, SMT 1) $= Gain(S, A) = Entropy(S)\text{-}\sum_{i=1}^{n} \frac{|si|}{|s|} * Entropy(Sn)$
$= 0.72192805\text{-}(\left(\frac{0}{10} * 0\right) + \left(\frac{10}{10} * 0.72192805\right) = 0{,}108289$

Gain(SMT 3 C, SMT 2) $= Gain(S, A) = Entropy(S)\text{-}\sum_{i=1}^{n} \frac{|si|}{|s|} * Entropy(Sn)$
$= 0.72192805\text{-}(\left(\frac{3}{10} * 0{,}918296\right) + \left(\frac{7}{10} * 0\right) = 0{,}446439$

Gain(SMT 3 C, SMT 4) $= Gain(S, A) = Entropy(S)\text{-}\sum_{i=1}^{n} \frac{|si|}{|s|} * Entropy(Sn)$

$= 0.72192805\text{-}(\left(\frac{7}{10} * 0{,}863121\right) + \left(\frac{3}{10} * 0\right) = 0{,}117744$

Gain(SMT 3 C, SMT 5) $= Gain(S, A) = Entropy(S)\text{-}\sum_{i=1}^{n} \frac{|si|}{|s|} * Entropy(Sn)$

$= 0.72192805\text{-}(\left(\frac{0}{10} * 0\right) + \left(\frac{10}{10} * 0{,}721928095\right) = 0{,}108289$

Gain (SMT 3 B, USBN) $= Gain(S, A) = Entropy(S)\text{-}\sum_{i=1}^{n} \frac{|si|}{|s|} * Entropy(Sn)$

$= 0.72192805\ \text{-}(\left(\frac{10}{10} * 0{,}721928095\right) = 0$

Decision trees that are formed up to the stage :

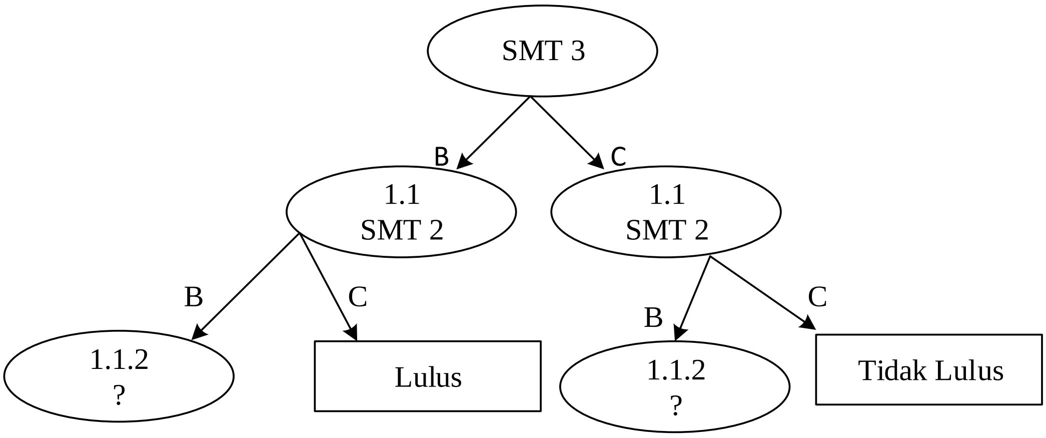

Figure 2. Decision Result Tree of Node 1.1 Decision

Calculating the number of cases, the number of cases for the graduation decision, the number of cases for the decision not to pass, and the entropy of all cases and cases divided by semester 1, 4, 5, and USBN attributes. Which can be a branch node of attribute value Semester 2 B of Semester 3 B root node and branch node of attribute value Semester 2 B of Semester 3 C. root node.

Entropy (Total **,** SMT 2 B) $= (-\frac{2}{4} * \log_2(\frac{2}{4}))+ (-\frac{2}{4} * \log_2(\frac{2}{4}))= 1$

***Entropy* SMT 1**

Entropy (SMT 2 B**,** B) $= (-\frac{1}{2} * \log_2(\frac{1}{2})) + (-\frac{1}{2} * \log_2(\frac{1}{2})) = 1$

Entropy (SMT 2 B, C) $= (-\frac{1}{2} * \log_2\left(\frac{1}{2}\right)) + (-\frac{1}{2} * \log_2(\frac{1}{2})) = 1$

***Entropy* SMT 4**

Entropy (SMT 2 B**,** B) $= (-\frac{2}{3} * \log_2(\frac{2}{3})) + (-\frac{1}{3} * \log_2(\frac{1}{3})) = 0{,}918295834$

Entropy (SMT 2 B, C) $= (-\frac{0}{1} * \log_2\left(\frac{0}{1}\right)) + (-\frac{1}{1} * \log_2(\frac{1}{1})) = 0$

***Entropy* SMT 5**

Entropy (SMT 2 B**,** A) $= (-\frac{0}{0} * \log_2(\frac{0}{0})) + (-\frac{0}{0} * \log_2(\frac{0}{0})) = 0$

Entropy (SMT 2 B, B) $= (-\frac{2}{4} * \log_2\left(\frac{2}{4}\right)) + (-\frac{2}{4} * \log_2(\frac{2}{4})) = 1$

***Entropy* USBN**

Entropy (SMT 2 B**,** B) $= (-\frac{2}{4} * \log_2(\frac{2}{4})) + (-\frac{2}{4} * \log_2(\frac{2}{4})) = 1$

Gain(SMT 2 B, SMT 1) = *Gain(S, A) = Entropy(S)*-$\sum_{i=1}^{n}\frac{|si|}{|s|} * Entropy(Sn)$

$$=1-((\frac{2}{4} * 1) + (\frac{2}{4} * 1) = 0$$

Gain(SMT 2 B, SMT 4) = *Gain(S, A) = Entropy(S)*-$\sum_{i=1}^{n}\frac{|si|}{|s|} * Entropy(Sn)$

$$=1-((\frac{3}{4} * 0{,}918295834) + (\frac{1}{4} * 0) = 0{,}31128$$

Gain(SMT 2 B, SMT 5) = *Gain(S, A) = Entropy(S)*-$\sum_{i=1}^{n}\frac{|si|}{|s|} * Entropy(Sn)$

$$= 1-((\frac{0}{4} * 0) + (\frac{4}{4} * 1) = 0$$

Gain(SMT 2 B, USBN) = *Gain(S, A) = Entropy(S)*-$\sum_{i=1}^{n}\frac{|si|}{|s|} * Entropy(Sn)$

$$=1-((\frac{4}{4} * 1) = 0$$

From the results it can be seen that the highest gain attribute is semester 4 which is equal to 0.31128. There are two attribute values from semester 4, namely B and C. From the two attribute values the value of attribute C already classifies the case, that is the decision "Not Passed", but for the value of attribute B it still needs to be calculated again. Thus 4th semester can be a branch node of the 2nd semester attribute value B. Next below is the overall decision tree formation.

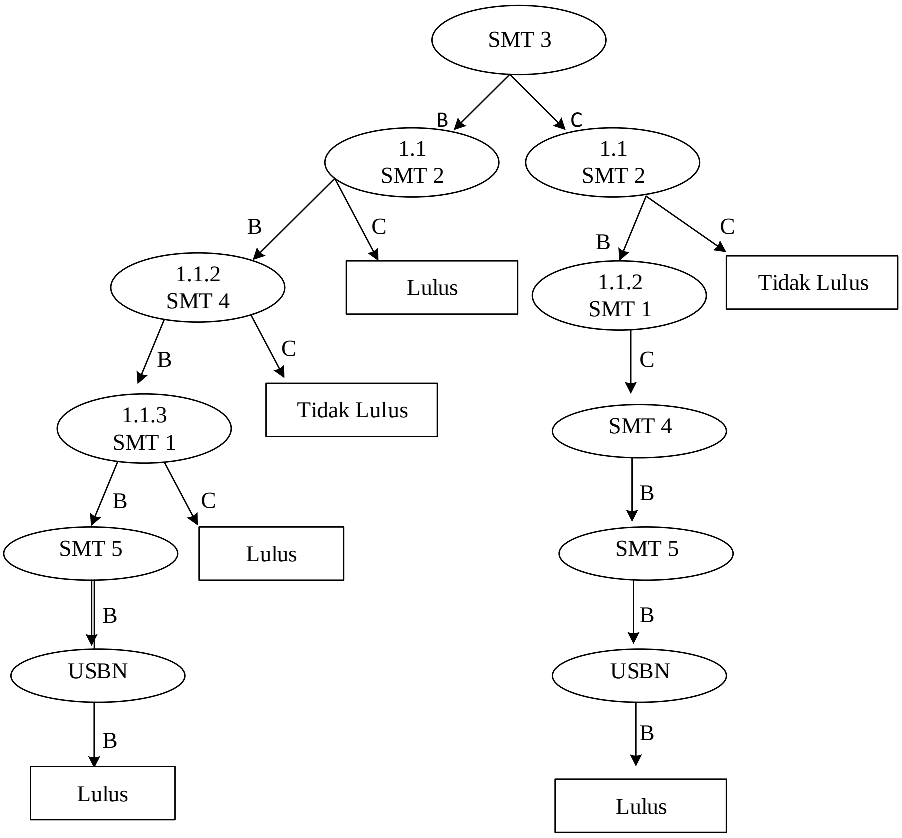

Figure 3. Decision Tree of C4.5 Algorithm Calculation Results

The following is the Rule formed in the C4.5 Algorithm decision tree as follows.

1. If SMT 3 = B and SMT 2 = C then "Pass" (pass = 6, not pass = 0).
2. If SMT 3 = B and SMT 2 = B and SMT 4 = C then "No Pass" (pass = 0, not pass = 1).

3. If SMT 3 = B and SMT 2 = B and SMT 4 = B and SMT 1 = C then "Pass" (pass = 1, not pass = 0).
4. If SMT 3 = B and SMT 2 = B and SMT 4 = B and SMT 1 = B and SMT 5 = B and USBN = B then (pass = 1, not pass = 1).
5. If SMT 3 = C and SMT 2 = C then "Not Passed" (graduated = 0, did not pass = 7).
6. If SMT 3 = C and SMT 2 = B and SMT 1 = C and SMT 4 = B and SMT 5 = B and USBN = B then (pass = 2, not pass = 1).

5. Conclusion

Through this research it is shown that:

a. Data mining processing using the classification method and C4.5 algorithm can predict the graduation rate of students with two categories, namely graduating and not graduating, and the most influential attribute in the prediction results is the third semester.
b. Analysis of student data to predict the graduation of these students has been successfully made using the classification method and the C4.5 algorithm.
c. Thus things that can be used as input based on the results of this study are better to add a tree form visually about the highest Gain value attribute until the calculation process is complete so that it can be known what attributes occupy the root, branch, leaf positions..

References

[1] Asril, E., & Wiza, F. (2015). Penerapan Data Mining Untuk Menggali Informasi Tersembunyi Dalam Big Data Nilai Mata Kuliah, 129–134.

[2] Asril, E., Wiza, F., & Yunefri, Y. (2015). Analisis Data Lulusan dengan Data Mining untuk Mendukung Strategi Promosi Universitas Lancang Kuning, *x*(x), 24–32.

[3] Indrawan, G. (2016). Penerapan Metode Decision Tree (Data Mining) Untuk Memprediksi Tingkat Kelulusan Siswa Smpn1, 35–44.

[4] Irfan, M. (2015). Issn 1979-8911 analisa pola asosiasi jalur masuk terhadap kelulusan mahasiswa dengan menggunakan metode, *IX*(2).

[5] Kamagi, D. H., & Hansun, S. (2014). Implementasi Data Mining dengan Algoritma C4 . 5 untuk Memprediksi Tingkat Kelulusan Mahasiswa. *ULTIMATICS, Vol. VI, No. 1 | Juni 2014*, *VI*(1), 15–20.

[6] Kartini, D. (2016). Rancang Bangun Aplikasi K-Means untuk Klasifikasi Kelulusan Siswa Sekolah Kepolisian Negara Daerah Kalimantan Selatan. *ProTekInfo*, *3*(1), 14–21.

[7] Meilani, B. D., & Susanti, N. (2014). Aplikasi Data Mining Untuk Menghasilkan Pola, *21*(2), 1–6.

Authors

1st Author
Fana Wiza
Department of Information System
Faculty of Computer Science,
Lancang Kuning University

2nd Author
Bayu Febriadi
Department of Information System
Faculty of Computer Science,
Lancang Kuning University