P-ISSN: 2087-1244
E-ISSN: 2476-907X

The Application of C4.5 Algorithm for Selecting Scholarship Recipients

Fristi Riandari[1]* and Sarjon Defit[2]

[1]Computer Engineering, STMIK Pelita Nusantara
Jln. Iskandar Muda, Sumatera Utara 20152, Indonesia
[2]Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang
Jln. Raya Lubuk Begalung, Sumatera Barat 25145, Indonesia
[1]fristy.rianda@ymail.com; [2]sarjond@yahoo.co.uk

Received: 26th April 2021/ **Revised:** 3rd July 2021/ **Accepted:** 5th July 2021

How to Cite: Riandari, F., & Defit, S. (2022). The Application of C4.5 Algorithm for Selecting Scholarship Recipients. *ComTech: Computer, Mathematics and Engineering Applications, 13*(1), 11−21. https://doi.org/10.21512/comtech.v13i1.7307

***Abstract* -** The scholarship program is one of the promotional techniques used by many universities, and the right scholarship award will certainly be an attraction for many people. STMIK Pelita Nusantara is one of the universities that organizes a scholarship program. In the current difficult economic conditions, the scholarship program is the target of many prospective students who want to continue their education in higher education. However, the absence of tools to process large amounts of data make determining scholarship recipients less effective and time-consuming. This situation is seen by the fact that some students are still unable to maintain the scholarships they receive. In the research, a classification model was proposed using the C4.5 algorithm approach by utilizing past data to facilitate the decision making of the scholarship program. This classification process produced a decision tree that could be used as a decision-making tool. Scholarships were awarded based on several criteria: academic potential, vocational potential, parents' income, number of dependents, and employment status. Based on the data processing results of students who apply for scholarships in 2020 with predetermined criteria, the highest root is obtained. It consists of node 1 for academic potential, node 1.1 for vocational potential, and node 1.2 for parental income. The resulting decision tree model is expected to help to make decisions quickly and on target.

Keywords: C4.5 algorithm, scholarship program, data mining, decision tree, data classification

I. INTRODUCTION

Current technological developments have been used as a tool in various fields, such as education, to enable teachers to quickly, precisely, and accurately process student data. Hence, the purpose of a job can be achieved effectively and efficiently (Hidayad, Defit, & Sumijan, 2020). Data processing by involving the technology in question will certainly be easier when associated with the right data processing model. One of the ways to process large amounts of data is by data mining techniques.

Data mining is an implementation model applied to look for patterns based on previous data to extract knowledge from large amounts of data (Guntur, Santony, & Yuhandri, 2018). The purpose of data mining is usually predictive (Dardzinska & Zdrodowska, 2020). According to Daryl Pregibon, data mining is a mixture of statistics, artificial intelligence, and database research which is still developing (Sulastri & Gufroni, 2017). Various techniques are available in data mining for knowledge extraction, including prediction, description, classification, estimation, association, grouping and classification (Ariawan, 2019; Afrianto, Suseno, & Warsito, 2020).

According to Azmi and Dahria (2013), data mining is an iterative process that requires human interaction to find a new pattern or model that can be generalized for the future and useful to carry out an action. Data mining has the concept of capturing and storing data, converting raw data into information and information into knowledge (Condrobimo, Sano, & Nindito, 2016).

*Corresponding Author

Data mining is also called Knowledge Discovery in Database (KDD) or pattern recognition (Hidayad et al., 2020). Data mining includes collecting and using historical sources and data to find regularities, patterns, or relationships in large datasets (Santoso, Hariyadi, & Prayitno, 2016). The process includes understanding the application field, making target data determined from the raw data contained in the database, and preprocessing and cleaning data (Virgo, Defit, & Yunus, 2020). The main goal of KDD is to extract high-level knowledge from low-level information (Putra & Defit, 2019). The KDD process generally consists of the following steps: data selection, data transformation, exploration like extraction of knowledge from data, and interpretation of results (Dardzinska & Zdrodowska, 2020). Data mining has also been implemented to predict students' study periods. The test results show that the error rate in predicting students' study period is only 5% (Haryati, Sudarsono, & Suryana, 2015).

Next, classification is finding a model or function that differentiates concepts or data classes. It aims to predict the class of objects whose class labels are unknown based on training data analysis (data objects with known class) (Afrianto et al., 2020). It is also the most commonly applied data extraction technique to predict categorical attribute values (discrete or nominal). It uses a set of previously classified examples to develop a model to classify entire population records with a decision tree or neural network-based classification algorithm. The process involves two stages, learning and classification. At the learning stage, the classification algorithm analyzes the training data. At the classification stage, test data is used to estimate the accuracy of the classification rules. The rules can be applied to the new data tuples if the accuracy is acceptable. The classifier training algorithm uses pre-classified examples to determine the set of parameters required for true discrimination and encodes these parameters into the model called classifier (Bedregal-Alpaca, Cornejo-Aparicio, Zárate-Valderrama, & Yanque-Churo, 2020).

Moreover, a decision tree is a data mining method used for classification. Decision tree classification is a simple classification technique that is widely used. Previous researchers have developed various decision tree algorithms over several periods by improving the performance and ability to handle various data types. Examples are the Chi-squared Automatic Interaction Detector (CHAID), Classification and Regression Tree (CHART), Iterative Dichotomiser 3 (ID3), C4.5 algorithm, C5.0 algorithm, Hunt's algorithm, and Ordinal Class Classifier (OCC) (Effendy & Purbandini, 2018).

In the research, the C4.5 algorithm is used. C4.5 algorithm is a classification technique using entropy and profit information as a separator in a decision tree (Florence & Savithri, 2013). The C4.5 algorithm constructs a decision tree from training data in the form of cases or records (tuples) in the database (Riandari & Simangunsong, 2019). The C4.5 algorithm is also used to build a decision tree. There is a study comparing the C4.5 algorithm and the CART algorithm in the student grade classification. It explains that the C4.5 algorithm has a higher accuracy value of 85,61%, while the CART algorithm has 84,95% (Rahmayuni, 2014). Moreover, the C4.5 algorithm generates a decision tree, which provides input in the form of a classification sample. The application of the C4.5 algorithm functions to produce a level of data accuracy as a dataset containing large amounts of data (Fiandra, Defit, & Yuhandri, 2017).

Scholarships are one of the leading programs offered by many universities. In the current difficult economic conditions, scholarship programs target many prospective students who want to pursue higher education. However, there are still some difficulties in determining the eligible prospective students due to the many applicants and the variables assessed in its implementation. Besides that, there are no tools that determine the selection. It takes a long time, and the possibility of inaccurate selection results is quite high. Based on the previous explanation of the data, data mining can be used to extract student data based on the characteristics of the selection results for scholarship recipients. The classification algorithm used is a decision tree with the C4.5 algorithm approach. Then, the classification results in the form of a decision tree that can be used as a tool in making decisions in the process of receiving scholarships quickly and staying on target. In this way, it is expected to help to make decisions quickly and on target.

II. METHODS

In tree formation with the C4.5 algorithm, there are several stages. Training data is usually taken from historical data that has occurred previously and grouped into certain classes. Second, it determines the roots of the tree. The root will be taken from the selected attribute by calculating the acquisition value of each attribute. Then, the highest value will be the first root (Dhika & Destiawati, 2015).

Analyzing the C4.5 algorithm is a stage after the problem to be analyzed is found. Then, the existing data will be processed. So, the C4.5 Algorithm design will be carried out after all existing data are processed, and all required data are complete. Data processing is carried out in accordance with the KDD stages (Rahmayuni, 2014).

First, it is selection. The object of the research is students who apply for scholarships in 2020. The research is carried out at STMIK Pelita Nusantara Medan. Then, the data collection uses observation and interviews with implementers. The data obtained are qualitative, containing information on each variable determined by the college in receiving scholarships, such as the value of academic potential, potential vocational test, parents' income, number of dependents, and employment status. The number of new students who apply for the scholarship that year is 150 people.

Second, there is preprocessing or cleaning.

After the data from the selection results are obtained (the data of prospective scholarship recipients in 2020, amounting to 150), the selection data proceeds to the data cleaning stage to remove inconsistent/noise and with the same value data. It can be said that this stage discards the data of prospective scholarship recipients with the same score as the other potential recipients in each criterion. Different patterns will be searched at this stage, if the same pattern is found, only one representative pattern will be left, and the rest will be cleaned. So, the final result of the cleaning stage gets 16 different patterns from 150 participant data in the scholarship acceptance process. Hence, the final result of this stage obtains 16 people from the previous data, amounting to 150 people.

Third, it is transformation. The preprocessed qualitative data will be grouped and transformed into an appropriate assessment form to be processed into data mining. In the research, the data are converted into quantitative form. This process makes it easier to define during testing.

Fourth, in data mining, the data of 16 students will be processed in the C4.5 algorithm data classification. It is carried out by making a decision tree to identify the conditions for objectively giving scholarships by looking at the value of each attribute of the new applicants for the scholarship (academic potential, potential vocational test, parents' income, total dependents, and employment status). It is based on the highest gain value of the existing attributes to choose an attribute as the root. Equation (1) is used to calculate gain. It shows S as a case set, A as features, n as the number of partitions S, and pi and the proportion of Si to S.

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

Meanwhile, Equation (2) calculates the entropy value in the entropy (total) formula. It shows as the number of partitions attribute A, | Si | as a number of cases on the i-th partition, and | S | as a number of cases in S.

$$\text{Entropy (total)} = -\sum_{i=1}^{n} - Pi\ x\ Log2\ Pi \tag{2}$$

In making a decision tree, it must count the number of cases, the number of cases for the decision of "Accepted" (S1), the number of cases for the decision of "Rejected" (S2), and cases divided based on the attributes of academic potential, vocational potential, parents' income, number of dependents, and employment status. Then, the gain will be calculated for each attribute. In making a decision tree, there are several stages. It determines the attribute as the root and calculates the value of the attribute gain information. It is based on the highest gain value of the existing attributes to select the attribute as the root. An entropy value is needed to determine the highest gain.

Fifth, the purpose of interpretation or evaluation is to objectively obtain the results of the decision analysis of students who receive scholarships. It is based on the attributes of academic potential, vocational potential, parents' income, number of dependents, and employment status. The data will be analyzed, and the method will be implemented to get the desired results.

Table 1 Data in the Test

No	Academic Potential	Vocational Potential	Parents' Income	Total Dependents	Employment Status	Decision
1	Good	Enough	Low	Low	High	Accepted
2	Enough	Good	High	High	Low	Accepted
3	Enough	Enough	High	Low	Low	Rejected
4	Enough	Enough	Low	High	Low	Accepted
5	Enough	Good	High	Low	High	Accepted
6	Good	Low	High	High	High	Accepted
7	Enough	Low	Low	Low	Low	Rejected
8	Enough	Good	High	Low	Low	Accepted
9	Low	Enough	Low	High	Low	Rejected
10	Low	Good	High	Low	High	Rejected
11	Enough	Low	High	Low	Low	Rejected
12	Low	Good	Low	Low	High	Rejected
13	Enough	Enough	Low	Low	High	Accepted
14	Low	Good	Low	High	Low	Rejected
15	Enough	Good	Low	High	High	Accepted
16	Enough	Enough	High	High	Low	Rejected

III. RESULTS AND DISCUSSIONS

Based on the test data in Table 1, the attribute as the root is determined. Then, the value of the attribute acquisition information is calculated. It is based on the highest gain value of the existing attribute to determine the attribute as root. In determining the highest gain value, the entropy value is needed. Then, to find the entropy of each case, the total number of sub-criteria values is calculated. The sub-criteria (see Table 2) in finding entropy is transformed into the following form.

Table 2 Description of Sub-Criteria

Description	
Good	G
High	H
Enough	E
Low	L

The calculation of the entropy value for each attribute uses Equation (3). Entropy (total) calculates the total value of the decision. The "Accepted" (S1) is 8, and "Rejected" is 8. Hence, the total number of cases is 16.

$$Entropy\ (total) = -\textstyle\sum_{i=1}^{n} -\mathrm{P}\,i \times \log_2 \mathrm{P}i$$

$$Entropy(total) = \left(-\frac{8}{16} * \log_2\left(\frac{8}{16}\right)\right) + \left(-\frac{8}{16} * \log_2\left(\frac{8}{16}\right)\right) = 1 \quad (3)$$

As seen in Table 1, the value of academic potential has a good score of 2 cases in the attributes of academic potential. Then, the rejected value has a good score of 0 cases. With a total of 2 cases, Equations (4), (5), and (6) calculate the entropy of each case. The same way is done for the other attributes.

$$Entropy(G) = \left(-\frac{2}{2} * \log_2\left(\frac{2}{2}\right)\right) + \left(-\frac{0}{2} * \log_2\left(\frac{0}{2}\right)\right) = 0 \quad (4)$$

$$Entropy(E) = \left(-\frac{6}{10} * \log_2\left(\frac{6}{10}\right)\right) + \left(-\frac{4}{10} * \log_2\left(\frac{4}{10}\right)\right) = 0{,}97095 \quad (5)$$

$$Entropy(L) = \left(-\frac{0}{4} * \log_2\left(\frac{0}{4}\right)\right) + \left(-\frac{4}{4} * \log_2\left(\frac{4}{4}\right)\right) = 0 \quad (6)$$

Attributes of vocational potential:

$$Entropy(G) = \left(-\frac{4}{7} * \log_2\left(\frac{4}{7}\right)\right) + \left(-\frac{3}{7} * \log_2\left(\frac{3}{7}\right)\right) = 0{,}98523 \quad (7)$$

$$Entropy(E) = \left(-\frac{3}{6} * \log_2\left(\frac{3}{6}\right)\right) + \left(-\frac{3}{6} * \log_2\left(\frac{3}{6}\right)\right) = 1 \quad (8)$$

$$Entropy(L) = \left(-\frac{1}{3} * \log_2\left(\frac{1}{3}\right)\right) + \left(-\frac{2}{3} * \log_2\left(\frac{2}{3}\right)\right) = 0{,}9183 \quad (9)$$

Attributes of parents' income:

$$Entropy(L) = \left(-\frac{4}{8} * \log_2\left(\frac{4}{8}\right)\right) + \left(-\frac{4}{8} * \log_2\left(\frac{4}{8}\right)\right) = 1 \quad (10)$$

$$Entropy(H) = \left(-\frac{4}{8} * \log_2\left(\frac{4}{8}\right)\right) + \left(-\frac{4}{8} * \log_2\left(\frac{4}{8}\right)\right) = 1 \quad (11)$$

Attributes of number of dependents:

$$Entropy(H) = \left(-\frac{4}{7} * \log_2\left(\frac{4}{7}\right)\right) + \left(-\frac{3}{7} * \log_2\left(\frac{3}{7}\right)\right) = 0{,}98523 \quad (12)$$

$$Entropy(L) = \left(-\frac{4}{9} * \log_2\left(\frac{4}{9}\right)\right) + \left(-\frac{5}{9} * \log_2\left(\frac{5}{9}\right)\right) = 0{,}99108 \quad (13)$$

Attributes of employment status:

$$Entropy(L) = \left(-\frac{3}{9} * \log_2\left(\frac{3}{9}\right)\right) + \left(-\frac{6}{9} * \log_2\left(\frac{6}{9}\right)\right) = 0{,}9183 \quad (14)$$

$$Entropy(H) = \left(-\frac{5}{7} * \log_2\left(\frac{5}{7}\right)\right) + \left(-\frac{2}{7} * \log_2\left(\frac{2}{7}\right)\right) = 0{,}86312 \quad (15)$$

In searching for the gain value for each attribute, several equations are used.

Gain (Total, Academic Potential)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\text{Academic Potential}_i|}{|Total|} * Entropy(\text{Academic Potential}_i)$$

$$= 1 - \left(\left(\frac{2}{16} * 0\right) + \left(\frac{10}{16} * 0{,}97095\right) + \left(\frac{4}{16} * 0\right)\right) = 0{,}39316 \quad (16)$$

Gain (Total, Vocational Potential)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\text{Vocational Potential}_i|}{|Total|} * Entropy(Vocational\ Potential_i)$$

$$= 1 - \left(\left(\frac{7}{16} * 0{,}98523\right) + \left(\frac{6}{16} * 1\right) + \left(\frac{3}{16} * 0{,}9183\right)\right) = 0{,}02178 \quad (17)$$

Gain (Total, Parents' Income)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Parents'\ Income_i|}{|Total|}$$

$$* Entropy(Parents'\ Income_i)$$

$$= 1 - \left(\left(\frac{8}{16} * 1\right) + \left(\frac{8}{16} * 1\right)\right) = 0 \qquad (18)$$

Gain (Total, Number of Dependents)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{(Number\ of\ Dependents_i)}{|Total|}$$

$$* Entropy\ (Number\ of\ Dependents_i)$$

$$= 1 - \left(\left(\frac{7}{16} * 0{,}98523\right) + \left(\frac{9}{16} * 0{,}99108\right)\right) = 0{,}01148 \qquad (19)$$

Gain (Total, Employment Status)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Working\ Status_i|}{|Total|}$$

$$* Entropy\ (Working\ Status_i)$$

$$= 1 - \left(\left(\frac{9}{16} * 0{,}9183\right) + \left(\frac{7}{16} * 0{,}86312\right)\right) = 0{,}10584 \qquad (20)$$

When all the entropy and gain values for each attribute have been calculated, the calculation results are recorded in Table 3. The calculations in Table 1 show that the attribute with the highest gain value is academic potential, with a gain value of 0,39316. So, this attribute is used as the root method with the others. The attribute with a lower value can be said to be "Rejected". However, attributes with enough value still have to be recalculated. Node 1 of the decision tree can be seen in Figure 1.

Furthermore, a solution is carried out to calculate Node 1.1 as the root. It is done in the same way as calculating the entropy value of the remaining attributes, such as vocational potential, parents' income, number of dependents, and employment status. After entropy is calculated, the gain for each attribute is measured. Entropy (value, academic potential, enough) is calculated with the following equations.

Table 3 Node 1: Calculation Results to Determine the Main Root

Node	Cases (S)		Accepted (S1)	Rejected (S2)	Entropy	Gain
1	Total	16	8	8	1	
	Academic Potential					0,39316
	Good	2	2	0	0	
	Enough	10	6	4	0,97095	
	Less	4	0	4	0	
	Vocational Potential					0,02178
	Good	7	4	3	0,98523	
	Enough	6	3	3	1	
	Less	3	1	2	0,9183	
	Parents' Income					0
	Less	8	4	4	1	
	High	8	4	4	1	
	Number of Dependents					0,01148
	High	7	4	3	0,98523	
	Less	9	4	5	0,99108	
	Employment Status					0,10584
	Less	9	3	6	0,9183	
	High	7	5	2	0,86312	

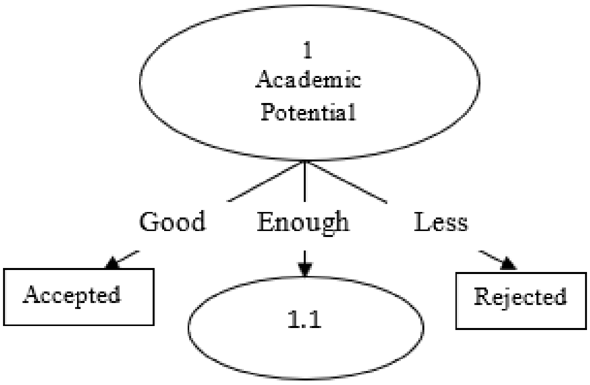


Figure 1 Decision Tree of Node 1

$$Entropy\ (total) = -\sum_{i=1}^{n} -\mathrm{P}i \times \log_2 \mathrm{P}i$$
$$= \left(-\frac{6}{10} * log_2\left(\frac{6}{10}\right)\right) + \left(-\frac{4}{10} * log_2\left(\frac{4}{10}\right)\right)$$
$$= 0{,}97095 \qquad (21)$$

Attributes of vocational potential

$$Entropy(G) = \left(-\frac{4}{4} * log_2\left(\frac{4}{4}\right)\right) + \left(-\frac{0}{4} * log_2\left(\frac{0}{4}\right)\right) = 0 \quad (22)$$

$$Entropy(E) = \left(-\frac{2}{4} * log_2\left(\frac{2}{4}\right)\right) + \left(-\frac{2}{4} * log_2\left(\frac{2}{4}\right)\right) = 0 \quad (23)$$

$$Entropy(L) = \left(-\frac{0}{2} * log_2\left(\frac{0}{2}\right)\right) + \left(-\frac{2}{2} * log_2\left(\frac{2}{2}\right)\right) = 0 \quad (24)$$

Attributes of parents' income

$$Entropy\ (L) = \left(-\frac{3}{4} * log_2\left(\frac{3}{4}\right)\right) + \left(-\frac{1}{4} * log_2\left(\frac{1}{4}\right)\right)$$
$$= 0{,}81128 \qquad (25)$$

$$Entropy\ (H) = \left(-\frac{3}{6} * log_2\left(\frac{3}{6}\right)\right) + \left(-\frac{3}{6} * log_2\left(\frac{3}{6}\right)\right) = 0 \quad (26)$$

Attributes of number of dependents

$$Entropy(H) = \left(-\frac{3}{4} * log_2\left(\frac{3}{4}\right)\right) + \left(-\frac{1}{4} * log_2\left(\frac{1}{4}\right)\right)$$
$$= 0{,}81128 \qquad (27)$$

$$Entropy(L) = \left(-\frac{3}{6} * log_2\left(\frac{3}{6}\right)\right) + \left(-\frac{3}{6} * log_2\left(\frac{3}{6}\right)\right) = 1 \quad (28)$$

Attributes of employment status

$$Entropy(L) = \left(-\frac{3}{7} * log_2\left(\frac{3}{7}\right)\right) + \left(-\frac{4}{7} * log_2\left(\frac{4}{7}\right)\right)$$
$$= 0{,}98523 \qquad (29)$$

$$Entropy(H) = \left(-\frac{3}{3} * log_2\left(\frac{3}{3}\right)\right) + \left(-\frac{0}{3} * log_2\left(\frac{0}{3}\right)\right) = 0 \quad (30)$$

Find the values of each attribute:

Gain (Total, Vocational Potential)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Vocational\ Potential_i|}{|Total|}$$
$$* Entropy\ (Vocational\ Potential_i)$$
$$= 0{,}97095 - \left(\left(\frac{4}{10} * 0\right) + \left(\frac{4}{10} * 1\right) + \left(\frac{42}{10} * 0\right)\right)$$
$$= 0{,}57095 \qquad (31)$$

Gain (Total, Parents' Income)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Parents'\ Income_i|}{|Total|}$$
$$* Entropy(Parents'\ Income\ _i)$$
$$= 0{,}97095 - \left(\left(\frac{4}{10} * 0{,}81128\right) + \left(\frac{6}{10} * 1\right)\right)$$
$$= 0{,}04644 \qquad (32)$$

Gain (Total, Number of Dependents)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Number\ of\ Dependents\ _i|}{|Total|}$$
$$* Entropy\ (Number\ of\ Dependents_i)$$
$$= 0{,}97095 - \left(\left(\frac{4}{10} * 0{,}81128\right) + \left(\frac{6}{10} * 1\right)\right)$$
$$= 0{,}04644 \qquad (33)$$

Gain (Total, Employment Status)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Employment\ Status_i|}{|Total|}$$
$$* Entropy(Employment\ Status_i)$$
$$= 0{,}97095 - \left(\left(\frac{7}{10} * 0{,}98523\right) + \left(\frac{3}{10} * 0\right)\right)$$
$$= 0{,}28129 \qquad (34)$$

When all entropy values and gain values have been calculated, the calculation results are put in Table 4. It can be seen that the highest gain attribute is the vocational potential, with a value of 0,57095. Thus, it can be interpreted that the vocational potential can become the next root node so that a decision tree is formed in Figure 2.

Next, the research calculates Node 1.2 as the root. It calculates the entropy and gain values in the same way, using the entropy value of the remaining attributes of parents' income, total dependence, and employment status. After calculating entropy, the gain is measured for each attribute. Entropy (Vocational potential, C) has the following equations.

Table 4 Node 1.1: Calculation Results to Determine the Branch from the Main Root

Node	Cases(S)		Accepted (S1)	Rejected (S2)	Entropy	Gain
1.1	Academic Potential: Enough	10	6	4	0,97095	
	Vocational Potential:					0,57095
	Good	4	4	0	0	
	Enough	4	2	2	1	
	Less	2	0	2	0	
	Parents' Income:					0,04644
	Low	4	3	1	0,81128	
	High	6	3	3	1	
	Number of Dependents:					0,04644
	High	4	3	1	0,81128	
	Low	6	3	3	1	
	Employment Status:					0,28129
	Low	7	3	4	0,98523	
	High	3	3	0	0	

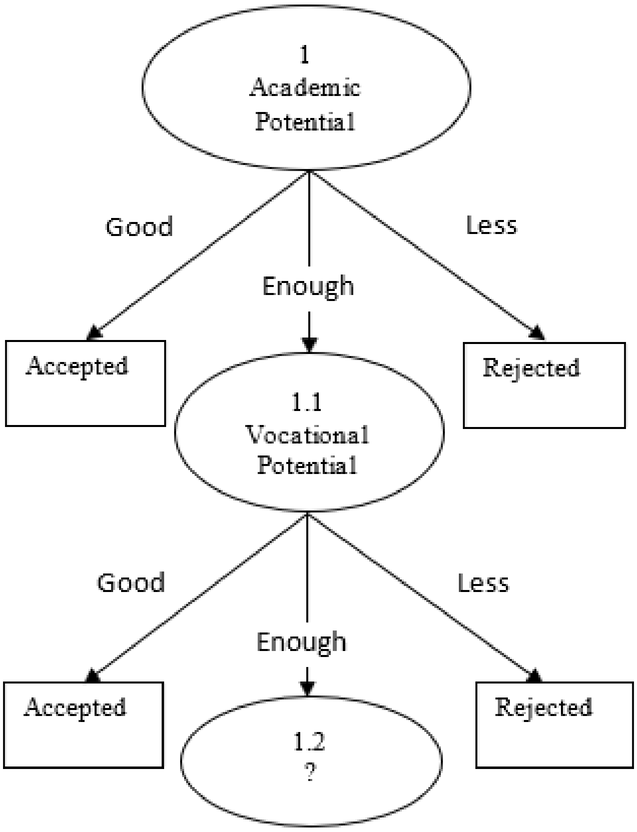


Figure 2 Decision Tree of Node 1.1

$$Entropy\ (total) = -\sum_{i=1}^{n} -\mathrm{P}i \times \log_2 \mathrm{P}i$$

$$= \left(-\frac{2}{4} * \log_2\left(\frac{2}{4}\right)\right) + \left(-\frac{2}{4} * \log_2\left(\frac{2}{4}\right)\right) = 1 \quad (35)$$

Attributes of parents' income

$$Entropy(R) = \left(-\frac{2}{2} * log_2\left(\frac{2}{2}\right)\right) + \left(-\frac{0}{2} * log_2\left(\frac{0}{2}\right)\right) = 0 \quad (36)$$

$$Entropy(T) = \left(-\frac{0}{2} * log_2\left(\frac{0}{2}\right)\right) + \left(-\frac{2}{2} * log_2\left(\frac{2}{2}\right)\right) = 0 \quad (37)$$

Attributes of number of dependents

$$Entropy(H) = \left(-\frac{1}{2} * log_2\left(\frac{1}{2}\right)\right) + \left(-\frac{1}{2} * log_2\left(\frac{1}{2}\right)\right) = 1 \quad (38)$$

$$Entropy(L) = \left(-\frac{1}{2} * log_2\left(\frac{1}{2}\right)\right) + \left(-\frac{1}{2} * log_2\left(\frac{1}{2}\right)\right) = 1 \quad (39)$$

Attributes of employment status

$$Entropy(L) = \left(-\frac{1}{3} * log_2\left(\frac{1}{3}\right)\right) + \left(-\frac{2}{3} * log_2\left(\frac{2}{3}\right)\right) = 0{,}9183 \quad (40)$$

$$Entropy(H) = \left(-\frac{1}{1} * log_2\left(\frac{1}{1}\right)\right) + \left(-\frac{0}{1} * log_2\left(\frac{0}{1}\right)\right) = 0 \quad (41)$$

Calculate the other values for each attribute:

Gain (Total, Parents' Income)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Parents'\ Income_i|}{|Total|}$$

$$* Entropy(Parents'\ Income_i)$$

$$= 1 - \left(\left(\frac{2}{4} * 0\right) + \left(\frac{2}{4} * 0\right)\right) = 1 \quad (42)$$

Gain (Total, Number of Dependents)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|Number\ of\ Dependents_i|}{|Total|}$$

$$* Entropy(Number\ of\ Dependents_i)$$

$$= 1 - \left(\left(\frac{2}{4} * 1\right) + \left(\frac{2}{4} * 1\right)\right) = 0 \quad (43)$$

Gain (Total, Employment Status)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\mathrm{Employment\ Status}_i|}{|Total|}$$

$$** Entropy(\mathrm{Employment\ Status}_i)$$

$$= 1 - \left(\left(\frac{3}{4} * 0{,}9183\right) + \left(\frac{1}{4} * 0\right)\right) = 0{,}31128 \quad (44)$$

After the entropy and gain values are calculated, the results of these calculations are put in Table 5. It can be seen that the highest gain attribute is the parents' income with a value of 1. So, it can be the next root node. The value of the T attribute is high, and the value of the R attribute is low. The decision tree formed can be seen in Figure 3.

The rules obtained based on the decision tree formed are as follows: IF Academic Potential = Good THEN Decision = Accepted, IF Academic Potential = Sufficient AND Vocational Potential = Good THEN Decision = Accepted, IF Academic Potential = Sufficient AND Vocational Potential = Sufficient AND Parents' Income = Low THEN Decision = Acceptable, IF Academic Potential = Enough AND Vocational Potential = Enough AND Parents' Income = High THEN Decision = Rejected, IF Academic Potential = Sufficient AND Vocational Potential = Low THEN Decision = Rejected, and IF Academic Potential = Low THEN Decision = Rejected.

Table 5 Node 1.2: Calculation Results to Determine the Next Branch

Node	**Cases (S)**		**Accepted (S1)**	**Rejected (S2)**	**Entropy**	**Gain**
1.2	Vocational Potential Enough	4	2	2	1	
	Parents' Income					1
	Less	2	2	0	0	
	High	2	0	2	0	
	Number of Dependents					0
	High	2	1	1	1	
	Less	2	1	1	1	
	Working Status					0,31128
	Less	3	1	2	0,9183	
	High	1	1	0	0	

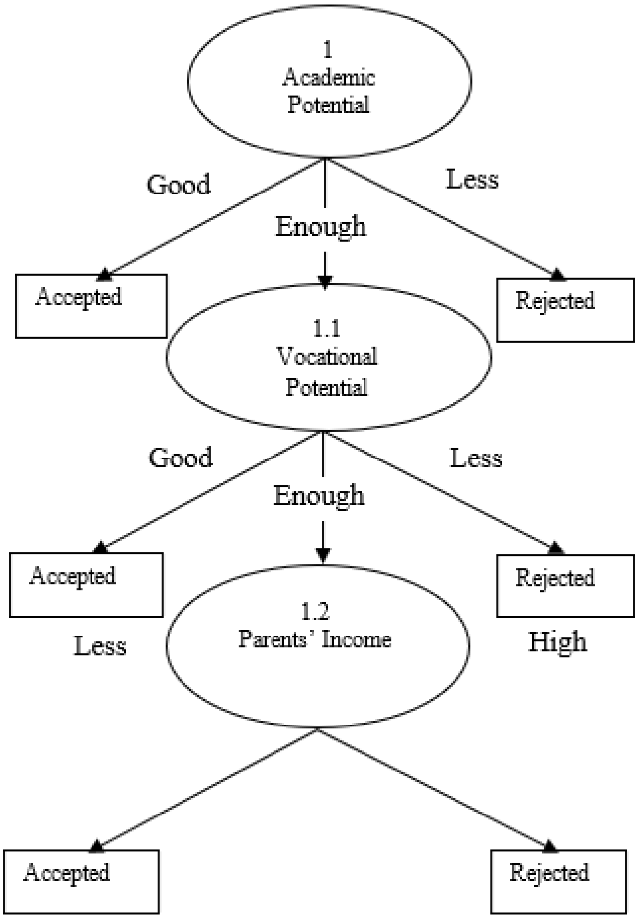


Figure 3 Decision Tree of Node 1.2

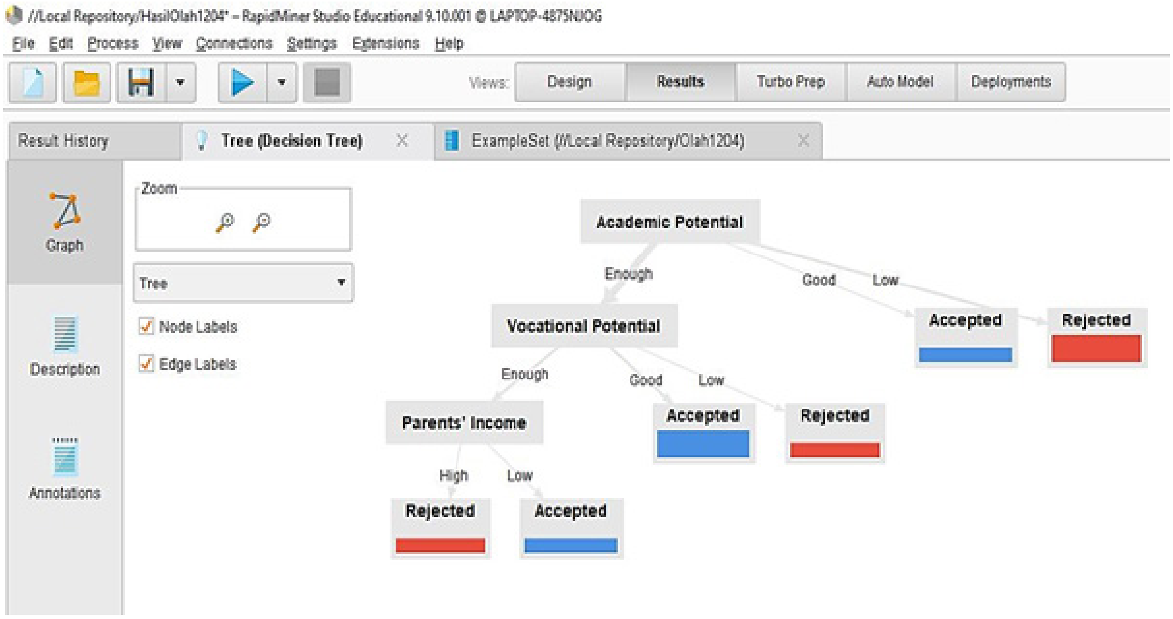
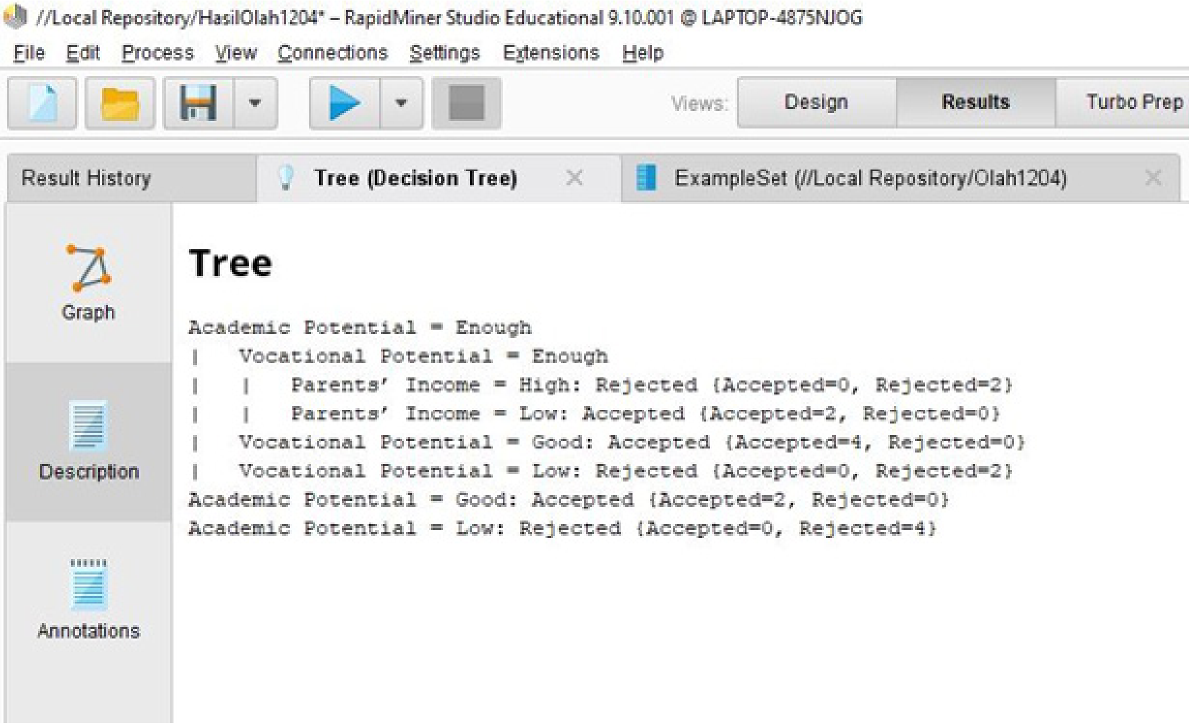

//Local Repository/HasilOlah1204* – RapidMiner Studio Educational 9.10.001 @ LAPTOP-4875NJOG
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Result History
Tree (Decision Tree)
ExampleSet (//Local Repository/Olah1204)
Graph
Description
Annotations
Zoom
Tree
☑ Node Labels
☑ Edge Labels
Academic Potential
Enough
Good
Low
Vocational Potential
Accepted
Rejected
Enough
Good
Low
Parents' Income
Accepted
Rejected
High
Low
Rejected
Accepted

Figure 4 Testing with Rapid Miner



//Local Repository/HasilOlah1204* – RapidMiner Studio Educational 9.10.001 @ LAPTOP-4875NJOG
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep
Result History
Tree (Decision Tree)
ExampleSet (//Local Repository/Olah1204)
Graph
Description
Annotations
Tree

```
Academic Potential = Enough
|   Vocational Potential = Enough
|   |   Parents' Income = High: Rejected {Accepted=0, Rejected=2}
|   |   Parents' Income = Low: Accepted {Accepted=2, Rejected=0}
|   Vocational Potential = Good: Accepted {Accepted=4, Rejected=0}
|   Vocational Potential = Low: Rejected {Accepted=0, Rejected=2}
Academic Potential = Good: Accepted {Accepted=2, Rejected=0}
Academic Potential = Low: Rejected {Accepted=0, Rejected=4}
```

Figure 5 Rule Formed From Rapid Miner Test Results

After the rules are obtained from the C4.5 algorithm classification process, further testing is carried out using one of the data mining applications, Rapid Miner. A decision tree and rules are obtained from the test results with the Rapid Miner application, which can be seen in Figure 4. The branch formed in a graph is from the same dataset as Table 1. The roots formed from the application test show the same shape as the manual calculation performed with the C4.5 algorithm in Figure 3. Then, in Figure 5, the rules are formed in a description.

IV. CONCLUSIONS

From the discussion results, it is concluded that the decision tree with the C4.5 algorithm can be used to classify the attributes used in analyzing prospective scholarship recipients. It can be a tool in making decisions about scholarship recipients and shorten the decision-making time. So, it can analyze prospective students who entitle to a scholarship with the most influential attributes, namely the academic potential, vocational potential, and parents' income. There are three influential variables from five variables used in selecting prospective scholarship recipients. The utilization of these three variables is based on branch in accordance with Figure 3, which is formed and translated into rules. It can shorten the timeline used in the selection because it has known the rules in the assessment. Hence, the assessment can be started on the criteria with the main priority or the highest root and only carried out on the influencing criteria. Then, the results obtained can be more efficient and on target.

It is hoped that future studies add variables related to the expected socioeconomic status, such as the parents' occupation, electricity bills, and homeownership status, to expand the research results. Hence, the scholarship recipient can be the right person regarding academic and socioeconomic status. At the same time, increasing the number of variables will allow the algorithm to work with larger data sets. In addition, future research can use other approaches in classifying scholarship patterns to determine the performance of each algorithm used, so universities can use decision-making tools that best suit their needs.

REFERENCES

Afrianto, E., Suseno, J. E., & Warsito, B. (2020). Decision tree method with C4.5 algorithm for students classification who is entitled to receive Indonesian Smart Card (KIP). In *IOP Conference Series: Materials Science and Engineering*. IOP Publishing. https://doi.org/10.1088/1757-899X/879/1/012072

Ariawan, P. A. (2019). Optimasi pengelompokan data pada metode K-means dengan analisis outlier. *Jurnal Nasional Teknologi & Sistem Informasi*, *5*(2), 88–95. https://doi.org/10.25077/teknosi.v5i2.2019.88-95

Azmi, Z., & Dahria, M. (2013). Decision tree berbasis algoritma untuk pengambilan keputusan. *Jurnal SAINTIKOM*, *12*(3), 157–164.

Bedregal-Alpaca, N., Cornejo-Aparicio, V., Zárate-Valderrama, J., & Yanque-Churo, P. (2020). Classification models for determining types of academic risk and predicting dropout in university students. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *11*(1), 266–272. https://doi.org/10.14569/ijacsa.2020.0110133

Condrobimo, A. R., Sano, A. V. D., & Nindito, H. (2016). The application of K-means algorithm for LQ45 index on Indonesia Stock Exchange. *ComTech: Computer, Mathematics and Engineering Applications*, *7*(2), 151–159. https://doi.org/10.21512/comtech.v7i2.2256

Dardzinska, A., & Zdrodowska, M. (2020). Classification algorithms in the material science and engineering data mining techniques. In *IOP Conference Series: Materials Science and Engineering*. IOP Publishing. https://doi.org/10.1088/1757-899X/770/1/012096

Dhika, H., & Destiawati, F. (2015). Application of data mining algorithm to recipient of motorcycle installment. *ComTech: Computer, Mathematics and Engineering Applications*, *6*(4), 569–579. https://doi.org/10.21512/comtech.v6i4.2192

Effendy, F., & Purbandini. (2018). Klasifikasi rumah tangga miskin menggunakan ordinal class classifier. *Jurnal Nasional Teknologi & Sistem Informasi*, *4*(1), 30–36. https://doi.org/10.25077/teknosi.v4i1.2018.30-36

Fiandra, Y. A., Defit, S., & Yuhandri. (2017). Penerapan algoritma C4.5 untuk klasifikasi data rekam medis berdasarkan International Classification Diseases (ICD-10). *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *1*(2), 82–89. https://doi.org/10.29207/resti.v1i2.48

Florence, A. M., & Savithri, R. (2013). Talent knowledge acquisition using C4.5 classification algorithm. *International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)*, *4*(4), 406–410.

Guntur, M., Santony, J., & Yuhandri. (2018). Prediksi harga emas dengan menggunakan metode Naïve Bayes dalam investasi untuk meminimalisasi resiko. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *2*(1), 354–360. https://doi.org/10.29207/resti.v2i1.276

Haryati, S., Sudarsono, A., & Suryana, E. (2015). Implementasi data mining untuk memprediksi masa studi mahasiswa menggunakan algoritma C4.5 (Studi kasus: Universitas Dehasen Bengkulu). *Jurnal Media Infotama, 11*(2), 130–138.

Hidayad, A., Defit, S., & Sumijan, S. (2020). Penerapan algoritma K-means clustering untuk melihat hubungan kegiatan Tahfiz dengan hasil belajar (Studi kasus Madrasah Aliyah Negeri 1 Bukiktinggi). *Jurnal Sistim Informasi dan Teknologi*, *2*(2), 41–47. https://doi.org/10.37034/jsisfotek.v2i2.34

Putra, R. A., & Defit, S. (2019). Data mining menggunakan rough set dalam menganalisa modal upah produksi pada industri seragam sekolah. *Jurnal Sistim Informasi dan Teknologi*, *1*(4), 72–78. https://doi.org/10.35134/jsisfotek.v1i4.18

Rahmayuni, I. (2014). Perbandingan performansi algoritma C4.5 dan Cart dalam klasifiksi data nilai mahasiswa Prodi Teknik Komputer Politeknik Negeri Padang. *Teknoif*, *2*(1), 40–46.

Riandari, F., & Simangunsong, A. (2019). *Penerapan algoritma C4.5 untuk mengukur tingkat kepuasan mahasiswa*. CV. Rudang Mayang.

Santoso, H., Hariyadi, I. P., & Prayitno. (2016). Data mining analisa pola pembelian produk dengan menggunakan metode algoritma Apriori. *Semnasteknomedia Online, 4(1),* 19–24.

Sulastri, H., & Gufroni, A. I. (2017). Penerapan data mining dalam pengelompokan penderita thalassaemia. *Jurnal Nasional Teknologi & Sistem Informasi*, *3*(2), 299–305. https://doi.org/10.25077/teknosi.v3i2.2017.299-305

Virgo, I., Defit, S., & Yunus, Y. (2020). Klasterisasi tingkat kehadiran dosen menggunakan algoritma K-means clustering (Studi kasus Institut Agama Islam Batusangkar). *Jurnal Sistim Informasi dan Teknologi, 2*(1), 23–28. https://doi.org/10.37034/jsisfotek.v2i1.22