Use of the C4.5 Algorithm to Analyze Student Interest in Continuing to College

M Lintang1, N Pandiangan2, D Hyronimus3

123 Department of Computer Education, Universitas Musamus, Merauke, Indonesia

Abstract: Higher education is one of the goals of several options when a student completes education in high school. However, not all students wish to continue their education in college. There are many reasons behind why students do not continue their education in college. Several factors that influence in this research include self-motivation, scholarships provided, parental support, ease of getting a job and academic competition. This study took a sample of 25 student data that will be used to find the most influential factors in continuing education in higher education. This study uses a data mining classification approach using the c4.5 method in making a decision tree. The results that can be concluded, a student has a high desire to continue his education in college because it will make it easier for them to get a job and the factor in getting an education scholarship.

Keywords: C4.5 Algorithm, Student interest, College,

1 Introduction

The quality of human resources is seen as one of the key factors in the era of free trade. The higher competition and demands in the world of work also require quality human resources with all the competencies they have, able to develop themselves and together build the nation [1]. One way to improve the quality of human resources is through education [2].

Higher education is organized to prepare students to become members of the community who have academic and/or professional abilities who can apply, develop and/or create science, technology and/or art. Therefore, universities as educational units that organize higher education play a very important role in creating quality human resources, so that rapid global changes can be responded to by existing education products.

Given the difficulty of getting a job in the midst of competition in the wider community, it is very important for higher education to have sufficient abilities and skills as a provision to become a workforce [3]. The quality of a job when attending higher education is not only concerned with theory but also practice. In addition, it will be easier to get a job with a more and more decent position. In fact, many students do not want to continue their education to college. There are various threatening reasons. However, in this study we have conducted research in high school on factors that generally have an influence on students who continue their education in college and students who do not continue their education in college. Among them are parental support, self-motivation, scholarships, ease of getting a job and academic competition.

To solve this problem into deeper results, then, we use the concept of data mining. Data mining is able to detect factors that occur in the educational environment like analyze the factors influencing active commuting to school [4]. Next, we use the c4.5 algorithm to classify in extracting data and information. The c4.5 algorithm is able to provide optimal results and is able to provide a fairly high level of accuracy [5]. The c4.5 algorithm is able to predict well on training data and provide increased accuracy when using larger training data [6]. This study is expected to find the most influential and dominant factor for high school students that causes them to have an interest in continuing their education to college using the data set that we have collected at their school.

2 Research Methods

Decision tree is able to solve difficult problems in ideological education, where the research puts forward a decision tree optimization algorithm for ideological education for decision making using classification techniques on cloud services. The experimental results in this study are to verify the effectiveness of the decision tree algorithm. The paper is able to carry out an empirical analysis for the ideological education of students based on the level of orientation [7]. The decision tree method is one of the most popular classification methods. This method is included in the

* Corresponding author: lintang@unmus.ac.id

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).
concept of machine learning. Decision tree has also been successfully implemented into an intelligent m-learning system, so that it can provide an effective learning system where this method plays a role in improving students’ learning abilities, so that it also has an impact on improving students' academic performance. [8].

Decision trees in the world of education can also be used to evaluate the form of scholarships in universities. The provision of this scholarship is very important because it is aimed at students who work hard to achieve achievements both at home and abroad. The use of this method is able to evaluate fair and efficient scholarships [9]. The classification method in the machine learning concept can predict student performance in the learning process [10]. This classification is associated based on the talent of students in the academic field with certain skills to improve their achievement. This analysis helps educational institutions to avoid failure in educating students. This study predicts student performance in a course which is also related to student and teacher data. The c4.5 method can classify and predict the solution of the problem accurately [11].

The method suggested in this paper to predict the most influential factors on students' interest in continuing their education to college, uses a combination of Data Mining and Decision Tree Learning. There are four main stages in this method: Data Collection, Classification, Predictive Modeling and Evaluation [12].

2.1 Decision Tree

Decision tree is a well-known predictive model and can be applied to a number of different fields [13]. In general, decision trees are built using an algorithmic approach so that they can be identified for data collection in various conditions [14]. This method is one of the most widely used in terms of supervised learning. This method is a non-parametric supervised learning method, which is used for classification assignments and regression assignments. Decision rules are generally in the form of 'if-then-else' statements. The deeper the decision tree, the more complex the rules [15].

The decision tree has a usable node where we can choose a suitable attribute to ask a question and its leaves represent the actual class label. Decision tree is used for non-linear decision making. Decision trees classify by sorting them down to the root and to each of its leaves, which leaves give the appropriate classification. Each node in the tree serves as a test case representing some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. Decision trees can be easily converted into classification rules. Decision tree used in data mining [16], there are two types: a) Classification tree, which is used to predict. b) regression, which is used to predict and generate real processed values findings.

2.2 C4.5 Algorithm

The C4.5 approach is a form of development of the ID3 algorithm proposed by Ross Quinlan [17]. The C4.5 algorithm is used as the basis for selecting attributes to carry out the testing process in the classification concept. The steps using the C4.5 algorithm to calculate student interest predictions are as follows.

Step 1: Calculate the initial entropy for sample distribution.

\[ Gain(S, A) = \text{Entropy}(S) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j) \]  

where:

- \( S \) : Case Collection
- \( A \) : Feature
- \( f \) : Number of partitions \( A \)
- \(|D_j|\) : Proportion of \( D_j \) to \( S \)
- \(|D|\) : Number of cases in \( S \)

Step 2 : Calculate entropy for any test property.

\[ \text{Entropy}(S) = \sum_{i=1}^{c} -P_i \times \log 2P_i \]

where:

- \( S \) : Case Collection
- \( A \) : Feature
- \( f \) : Number of partitions \( A \)
- \( P_i \) : Proportion of \( P_i \) to \( S \)

1) Prepare training data. Training data is usually taken from historical data has happened before or is called past data and has been grouped in certain classes;
2) Counting the roots of the tree. The root will be taken from the attribute to be selected, with how to calculate the gain value of each attribute, the highest gain value which will be the first root. Before calculating the gain value of the attribute, calculate first the entropy value;
3) Calculate the Gain value;
4) Repeat step 2 and step 3 until all records are partitioned;
5) The decision tree partitioning process will stop when: (a) All records in node \( N \) have the same class. (b) There are no attributes in the partitioned record anymore. There are no records in the empty branch.

3 Results And Discussion

3.1 Case Study

3.1.1 Data Selection and Data Preparation

Variables that affect students' interest in continuing their studies to college will then be processed using data mining techniques. From the existing data, attributes or labels will then be given to determine the final decision. The variables used in the implementation of the decision tree are academic competition, parental support, self-motivation, ease of getting a job and get scholarship.
3.1.2 Data Preprocessing

Data preprocessing is applied to add the contents of missing or inappropriate data attributes, as well as changing data inconsistencies. The steps needed for data preprocessing include, (1) the data is transformed into a suitable form for carrying out the data mining process, (2) performing data reduction to eliminate unnecessary attributes. Table 1 shows the data that has been repaired and is ready to be computed.

<table>
<thead>
<tr>
<th>Academic Competition</th>
<th>Parental Support</th>
<th>Self-motivation</th>
<th>Ease of Getting Job</th>
<th>Scholarship</th>
<th>decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Yes</td>
<td>Motivated</td>
<td>Easy</td>
<td>Yes</td>
<td>Continue</td>
</tr>
<tr>
<td>Have Competition</td>
<td>Yes</td>
<td>Motivated</td>
<td>difficult</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Have Competition</td>
<td>No</td>
<td>Motivated</td>
<td>difficult</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Easy</td>
<td>Yes</td>
<td>Continue</td>
</tr>
<tr>
<td>Have Competition</td>
<td>Yes</td>
<td>Motivated</td>
<td>difficult</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Have Competition</td>
<td>Yes</td>
<td>Motivated</td>
<td>Easy</td>
<td>Yes</td>
<td>Continue</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>difficult</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Have Competition</td>
<td>Yes</td>
<td>Motivated</td>
<td>difficult</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Have Competition</td>
<td>Yes</td>
<td>Motivated</td>
<td>Easy</td>
<td>Yes</td>
<td>Continue</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Easy</td>
<td>Yes</td>
<td>Continue</td>
</tr>
<tr>
<td>Have Competition</td>
<td>No</td>
<td>Motivated</td>
<td>Easy</td>
<td>Yes</td>
<td>Continue</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>difficult</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Have Competition</td>
<td>Yes</td>
<td>Motivated</td>
<td>difficult</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Have Competition</td>
<td>Yes</td>
<td>Motivated</td>
<td>Easy</td>
<td>Yes</td>
<td>Continue</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Easy</td>
<td>Yes</td>
<td>Continue</td>
</tr>
<tr>
<td>Have Competition</td>
<td>Yes</td>
<td>Motivated</td>
<td>difficult</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Have Competition</td>
<td>Yes</td>
<td>Motivated</td>
<td>difficult</td>
<td>Yes</td>
<td>Continue</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Easy</td>
<td>Yes</td>
<td>Continue</td>
</tr>
<tr>
<td>Have Competition</td>
<td>No</td>
<td>Motivated</td>
<td>Easy</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>difficult</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Have Competition</td>
<td>Yes</td>
<td>Motivated</td>
<td>Easy</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Motivated</td>
<td>Easy</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 1. Dataset of factors that influence students to continue their studies to college

<table>
<thead>
<tr>
<th>Node</th>
<th>Information</th>
<th>Amount Case</th>
<th>Continue</th>
<th>No</th>
<th>Entropy</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Total</td>
<td>25</td>
<td>16</td>
<td>9</td>
<td>0.942</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Academic Competition</td>
<td>Have Competition</td>
<td>15</td>
<td>10</td>
<td>5</td>
<td>0.918</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>0.970</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parental Support</td>
<td>Yes</td>
<td>17</td>
<td>12</td>
<td>5</td>
<td>0.873</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>4.583</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self-Motivation</td>
<td>Motivated</td>
<td>17</td>
<td>10</td>
<td>7</td>
<td>0.977</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>0.954</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ease of Getting Job</td>
<td>Easy</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>0.413</td>
</tr>
<tr>
<td></td>
<td>Difficult</td>
<td>13</td>
<td>5</td>
<td>8</td>
<td>0.961</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scholarship</td>
<td>Yes</td>
<td>13</td>
<td>12</td>
<td>1</td>
<td>0.391</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>12</td>
<td>4</td>
<td>8</td>
<td>0.918</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Dataset of factors that influence students to continue their studies to college
3.1.3 **Turning Data Into Trees**

In converting the data into a tree, the data is first expressed in the form of a table with attributes and records. Attribute states a parameter that is created as a criterion in the formation of the tree.

3.2 **Result**

3.2.1 **Determining the Entropy Value and Gain Value**

In the sample data, the first thing to do is to calculate the entropy values in all tables to find the first node:

\[
\text{Entropy Total} = ((-9/25) \times \log_2(9/25) + ((-16/25) \times \log_2(16/25)) = 0.942 \]  

(3)

Then calculate the entropy of all cases which are divided based on the attributes of academic competition, parental support, self-motivation, ease of getting a job and scholarship. After getting the entropy value, the next step is to calculate the gain value for each attribute. The gain value for each attribute is:

- Academic competition, \( \frac{0.942-(15/25) \times 0.918+(10/25) \times 0.970}{(15/25) \times 0.918+(10/25) \times 0.970} = 0.780 \).
- Parental Support, \( \frac{0.942-(17/25) \times 0.873+(8/25) \times 1}{(17/25) \times 0.873+(8/25) \times 1} = 0.668 \).
- Self-Motivation, \( \frac{0.942-(17/25) \times 0.977+(8/25) \times 0.954}{(17/25) \times 0.977+(8/25) \times 0.954} = 0.583 \).
- Ease of Getting Job, \( \frac{0.942-(12/25) \times 0.413+(13/25) \times 0.961}{(12/25) \times 0.413+(13/25) \times 0.961} = 1.243 \).
- Scholarship, \( \frac{0.942-(13/25) \times 0.391+(12/25) \times 0.918}{(13/25) \times 0.391+(12/25) \times 0.918} = 1.18 \).

From the results of the calculation of the search for the 1st node, the ease of getting job is obtained as the attribute with the largest gain value. Then the ease of getting job will be used as the root node in the next search. This calculation continues until all the attributes in all branches have the same class by selecting the attribute based on the highest "gain" value of the existing attribute. Where the calculation results for node 1 are as shown in Table 2. After all the calculation processes to determine branching have been completed, a decision tree is generated as shown in Figure 1 below.

![Fig.1 Decision tree to analyze the most influential student factors for continuing education.](image)

3.2.2 **Turning Trees Into Rules**

Rule 1: IF Ease getting job = difficult ^ Scholarship = Yes THEN decision = Continue to higher education.

Rule 2: IF Ease getting job = difficult ^ Scholarship = No THEN decision = No.

Rule 3: IF Ease getting job = easy THEN decision = Continue to higher education.

3.2.3 **Description**

The results of the classification of the processed data sets are as follows:

- If Ease of Getting Job = Easy, then Continue to Higher Education \{Continue to higher education = 11, No = 1\}.
- If Ease of Getting Job = Difficult, then Scholarship = No \{Continue to higher education = 1, No = 6\}.
- If Ease of Getting Job = Difficult, then Scholarship = Yes: Continue to higher education \{Continue to higher education = 4, No = 2\}.

4 **Conclusions**

The decision tree is a system developed to help find and make decisions for a problem and by taking into account various factors that exist within the scope of the problem. With a decision tree, humans can easily identify and see the relationship between the factors that affect a problem and can find the best solution by taking into account these factors. The results of the study using the decision tree concept with the c4.5 method show that from a total of 25 datasets it can be concluded that the attributes of easy work and scholarship attributes have a great influence on a student to continue to college level.

4 Acknowledgment

We would like to thank the Principal of SMA Muhammadiyah Merauke who has facilitated us so that we are able to carry out this research well. We also thank the leadership of Musamus University who has supported us in completing this research.

4 References


