Decision Tree C 4.5 Algorithm for Classification of Poor Family Scholarship Recipients

Y Kustiyahningsih¹, B K Khotimah¹, D R Anamisa¹, M Yusuf¹, T Rahayu¹, J Purnama²
¹Departement of Informatics Engineering, University of Trunojoyo Madura, Bangkalan – Madura, Indonesia
²Departement of Engineering, university of 17 Agustus 1945 Surabaya, Indonesia

¹ykustiyahningsih@trunojoyo.ac.id

Abstract The poor family scholarship program is a scholarship program by Lamongan District Education Office. This scholarship is given to underprivileged students for study costs at university. The many indicators of scholarship determination result in difficulties education office in making objective decisions. The purpose is to determine the classification of scholarship recipients using the decision tree C4.5 algorithm. This algorithm has the advantage of speed informing and reading models, high accuracy, flexibility, and processing discrete and continuous data. The contribution of this research is to classify poor family scholarship participants with the C 4.5 algorithm without pruning, pre pruning, and post pruning methods. This study uses 460 data and for each attribute, it is calculated using the gain ratio formula to produce trees and decision rules. Based on the scenario that has been done the accuracy value generated after pruning is better than without pruning. The classification process using scenario post pruning technique is used as a reference for the classification of new registrants because it has the highest accuracy of 85.55%.

Keywords: C 4.5 Algorithm, Classification, Poor family, Scholarship Program.

1. Introduction

Education is the main pillar in the progress of a nation, a country is said to be advanced if the country's education develops rapidly and adequately. Education has become one of the important elements to support the working world, especially higher education. However, the cost of education is relatively not cheap, an obstacle for people who have financial limitations. According to Law No. 20 of 2003 which states that "every citizen has the right to have the opportunity to improve lifelong education"[1]. This law states that Indonesian citizens have the right to improve education because education can increase their potential and ways of thinking. With this regulation, the Lamongan District Education Office made policy for students who graduated from high school and vocational school who wanted to continue to state or private universities but were financially constrained[2]. The Lamongan Education Office provides scholarships or assistance for them to continue their education to college.

From 2006 to 2019, as many as 3,890 underprivileged students in Lamongan Regency, East Java, were helped by their tuition fees through a scholarship scheme by the local government [3]. The main
requirement for scholarship is to have a Lamongan regency family card and a certificate of being unable (economic difficulties) and report grades while they are in high school or vocational high school[2]. The scholarships provided by the Lamongan District Education Office each use a manual system. From the beginning until now, there is no system to determine students who are entitled to get a scholarship based on the specified requirements. The amount of aid and labor funds is limited, so it is feared that there will be errors in selection and not on target. This study using data mining methods to assist the education office in accelerating the selection process of scholarship recipients for poor families. The data mining method used is C 4.5. This method is a decision tree algorithm for decision making by classifying scholarship recipients based on attributes or indicators used[4]. Based on previous research, C 4.5 method has been widely applied to predict student graduation rates [5]. customer loyalty predictions [6]. Decision Support System Applications Using C 4.5. Algorithms, For Senior High School majors [7], to classify project tender winners [8], the attribute selection used in previous studies is to use information gain and Correlation-based Feature Selection (CFS) and C 4.5 are the most stable algorithms and more accurate values high[9][10][11]. Comparison of the performance of ID3 and C 4.5 decision trees in the classification of motorcycle creditworthiness, this study shows that the C 4.5 algorithm has a degree of accuracy [12]. Therefore in this study using C 4.5 algorithm in classification. The purpose of this study can assist the department in determining the receipt of scholarships quickly, based on the rules generated by the algorithm and used as decision support. C4.5 algorithm can make predictions by providing an ideal level of accuracy for receiving scholarships. The classification process in this study is divided into three namely without pruning, pre pruning, and post pruning.

2. Research Methods
2.1. Determine System Flowchart

The flowchart is a picture of the flow of the system that is done as a whole or separately in a particular process and explains the procedures that exist in the system.

![Flowchart](image)

**Figure 1.** Flowchart system for receiving a Poor Family scholarship

2.2. Formation of Decision Trees with C4.5 Algorithm[13][14].

a. Training data input.

b. Calculate the gain ratio, split info, and entropy of each of the existing training data attributes.

\[
GainRatio(S, A) = \frac{Gain(S, A)}{split \text{ info}(S, A)}
\]  

(1)

With \( S = Data \text{ Training} \), \( A = Attributes \), \( Gain \ (S, A) = Information \ gain \ on \ attribute \ A \), Split Info \ (S, A) = Split information on attribute A

\[
Entropy_S = - \left( \frac{p_x}{S} \log \frac{p_x}{S} + \frac{p_y}{S} \log \frac{p_y}{S} \right)
\]  

(2)

\[
SplitInfo(S, A) = - \sum_{i=1}^{n} \frac{S_i}{S} \log \frac{S_i}{S}
\]  

(3)
With S = training data, A = Attribute, Si = Number of Training for attribute i

c. Make the root node of attribute selection which has the largest gain ratio.

d. Calculate the gain ratio, split info and entropy of each attribute by removing previously selected attributes.

e. Create an internal node from attribute selection that has the largest gain ratio. Check whether all attributes have been formed in the tree.

f. If not, then repeat the d and e processes, if so then continue to the next process.

g. Trim the trees to eliminate unnecessary branches. Then decision rules are generated following a tree that was formed before.

2.3. Determine Pre Pruning Process

Pre pruning is pruning done since the beginning of tree formation by stopping the construction of a subtree earlier, namely by deciding not to further partition the training data. How it works pre pruning is to first calculate the value of gain ratio to find out the value of parent and child. After the parent and child are known then the error value is calculated, if the child error value is a smaller parent then the parent forms a subtree again, but conversely, if the child error value is greater than the parent then the pruning is done and subtree formation stops. The following equation 3 is used for the pre pruning approach [12][15].

\[ e = \frac{r + \frac{z^2}{2n} + \frac{z^2}{n} - \frac{z^2}{4n^2}}{1 - \frac{z^2}{n}} \]  

where the value of r is the value of the error rate comparison, n is the total sample while determining the value of z using a normal distribution table by determining confidence. In general, calculating a confidence interval with a 95% probability of the true value[16]. because the actual value is 95%, the z value is 1.645 by looking at the normal distribution table.

2.4. Determine the Post Pruning Process

The post pruning approach uses the Reduced Error Pruning method [17][18]. Reduced Error Pruning is a post pruning algorithm. This algorithm divides data into two, namely training and testing data. Training data is the data used to form a decision tree, while testing data is used to calculate the error rate in the tree after pruning. How it works is to trim the internal node that starts from the bottom of the internal node to the top. Pruning is done by replacing attributes with leaf nodes that have a dominant class appearing. After that the test data is processed using the pruning rule, then the error value is calculated. Test data is also processed with the initial rule, the rule that is formed before the tree is pruned, and then the error rate is calculated. If the error rate resulting from pruning trees is smaller, then pruning is done[19][20].

3. Results And Discussion

3.1. Data Sharing

The amount of data used in this study is 460 data, consisting of 350 data having the class "YES" and 110 data having the class "NO". The data is divided into 2 namely training data and testing data. Training data is used to form decision trees, testing data is used to test trees that have been formed to calculate error rates, recall, precision, test accuracy and pruning process of trees. The Scholarship Recipients Dataset is in Table 1 and the Poor Family Scholarship Attribute Data is in Table 2.

3.2. Tree Decision Formation Process C4.5

As a first step to calculating the search for the value of the gain ratio, split info and entropy can be seen in Table 3. Furthermore, the C4.5 algorithm for building a decision tree as follows: select attribute as the root, create a branch for each value, divide cases into branches, and repeat the process for each branch until all cases in the branch have the same class.

**Table 1. Scholarship Recipients Dataset**

| No. | Name | Gender | Card | Income | The number of dependents | Profession | Electricity cost | Class |
|-----|------|--------|------|--------|--------------------------|------------|-----------------|-------|
| 1   |      |        |      |        |                          |            |                 |       |
Table 2. The Poor Family Scholarship Attribute Data

| No | Attribute          | Attribute Value | Information                        |
|----|-------------------|-----------------|------------------------------------|
| 1  | Card              | Very Good       | More than 4 card                   |
|    |                   | Good            | More than 3 card                   |
|    |                   | Enough          | 2 card                            |
|    |                   | Less            | 1 card                            |
|    |                   | Very Less       | No card                           |
| 2  | Income            | Good            | < 500.000                         |
|    |                   | Enough          | 500.000 – 1000.000                |
|    |                   | Less            | 1,000.000-2,000.000               |
|    |                   | Very Less       | > 2,000.000                       |
| 3  | The number of family dependents | Good          | More than four-person              |
|    |                   | Enough          | More than three-person             |
|    |                   | Less            | Two-person                        |
|    |                   | Very Less       | one person                        |
| 4  | Profession        | Very Good       | unemployment                       |
|    |                   | Good            | Farmers, fishermen, pedicab drivers|
|    |                   | Enough          | Bricklayer, Motorcycle taxi driver, Woodworker|
|    |                   | Less            | entrepreneur, Trader               |
|    |                   | Very Less       | Civil servants, Polri, TNI, Businessmen|
| 5  | Electricity cost  | Good            | < 50.000                          |
|    |                   | Enough          | 50.000 -149.000                   |
|    |                   | Less            | 150.000 – 250.000                 |
|    |                   | Very Less       | > 250.000                        |

3.3. Decision Tree Formation

The decision tree is formed from the highest gain ratio value from the data in Table 1 and the algorithm in the research method. The highest gain ratio value is the card attribute, so the root/parent is occupied by the card attribute and the card attribute class is very good, good, sufficient, less and very less. The rules formed are as follows. If the card is very good, then the registrant is accepted, but if the card is lacking, then the system will look for the next leaf by repeating until it reaches the same class. The next calculation is done with the card attribute = less and card = very less. because the number of cases of YES and NO values does not occupy a greater value between the two, it is necessary to calculate to form a branch below, the card attribute = less and card = very less, and so on as in Figure 2.

![Figure 2. Results of Rule decision tree C 4.5](image-url)
Testing on this system uses 3 data partitions. Each data partition has several training data, data testing, and pruning tests. Each data partition is carried out 3 tree formation processes. First, tree formation without pruning techniques. Second, tree formation pre pruning technique, Third, tree formation post pruning technique. The scenario in this study is to compare the results obtained by using a 70:30 data comparison partition consisting of 308 training data, 132 data testing data, and 20 data pruning tests. Rule Decisions tree formation are divided into 3, such as:

1. Decision rules obtained from tree formation without pruning techniques, the rule is show in Figure 3. Rule decision tree without pruning.

![Figure 3. Rule decision tree without pruning](image)

2. Decision rules obtained from tree formation with pre pruning techniques, the rule is show in Figure 4. Rule decision tree with pre pruning.

![Figure 4. A rule decision tree with pre pruning](image)

3. Decision rules obtained from tree formation with post pruning techniques, the rule is shown in Figure 5. Rule decision tree with post pruning.

![Figure 5. A rule decision tree with post pruning](image)

The result of testing and rule decision tree above, Precision, Recall, Error Rate, and accuracy of each decision rule without pruning, pre pruning, and post pruning can be shown in Table 3. Comparison of accuracy results without pruning, pre pruning, and post pruning and Figure 6. Results of a decision tree comparison. Analysis of the results Comparison between accuracy without pruning, pre- pruning and post- pruning, the highest accuracy produced is 85.50% with the lowest difficulty level of 14.5%. The best accuracy is obtained from the pruning process that is post pruning. Based on this, post pruning becomes the best decision rule for determining the acceptance of poor family scholarships for the Lamongan community.

**Table 3. Comparison of accuracy results without pruning, pre pruning and post pruning**

| Testing   | Without Pruning | Prepruning | Post pruning |
|-----------|-----------------|------------|--------------|
| Precision | 81.55%          | 83.00%     | 83.17%       |
| Recall    | 97.67%          | 96.51%     | 97.67%       |
4. Conclusion

Based on the evaluation conducted it can be seen that the process of tree formation using pruning techniques has an average accuracy value greater than without using the pruning process. The attribute that is the root of the decision tree in this study is the card, as well as the main determining parameters that are determined for scholarship recipients from the Department of Education. According to the results of the training data that using the C4.5 algorithm in the scholarship recipient dataset, it produces good rules to be used as classifications for the classification of new scholarship recipient data. The rule used for classification in new registrants is to use partition rules after pruning because it has a good level of accuracy than other partitions of 85.55%.

References

[1] The Law Number 20 of 2003 concerning the National Education System, 2003.
[2] Lamongan Regent Regulation No. 51 of 2011 concerning Guidelines for Implementing Scholarship Assistance for Students from Poor Families in Lamongan District, 2011.
[3] https://faktualnews.co/2019/02/26/pemkab-lamongan-help-lecture-3-890-underprivileged students/26092/: Saturday, 11 April 2020.
[4] Rizky H, Pambudi, Budi D S, Indriati, "Application of C4.5 Algorithm to Predict High School Graduation Grades Based on External Factors", Journal of Information Technology and Computer Science Development e-ISSN: 2548-964X Vol. 2, No. 7: 2637-2643, 2018.
[5] Kamagi, D.H. and Hansun, S,"Implementation of Data Mining with C4.5 Algorithm to predict Student Graduation Rates", UMN, Tangerang, 2014.
[6] Santoso T B, "Analysis and Leading Method C4.5 for Predicting Customer Loyalty", Satya Negara Indonesia University. Jakarta, 2013.
[7] Kustiyahningsih Y, Rahmanita, "Application of Decision Support Systems Using C4 Algorithm. 5. For majors in high school". Simantec Journal, 5 (2), pp. 101-108, 2016.
[8] Duarte G, Jose A, Mudjihartono, Paulus, Ernawati, "Analysis of the Effectiveness of the C4.5 Algorithm in Determining Project Tender Winner Participants", Yogyakarta: Master of Informatics Engineering, Atma Jaya University Yogyakarta, 2012.
[9] Bhumika G, Aditya R, Akshay J, Arpit A, Naresh D, "Analysis of Various Decision Tree Algorithms for Classification in Data Mining", International Journal of Computer Applications 163 (8), 2017.
[10] Muhammad R, Deni M, "Application of Classification Algorithm C 4.5 for Predicting Asset Maintenance", International Journal of Recent Contributions from Engineering, Science & IT (iJES) 7 (4), 2019.
[11] Muhammad S E and Helmie A W,"Prediction of Diabetes Using Algorithms ID3 with the Selection of the Best Attributes (Diabetes Prediction using ID3 Algorithm with Best Attribute Selection) " ,JUITA VI 1: 29-35, 2018.
[12] Nugroho M.W. Comparison of ID3 and C4.5 decision tree performance in the classification of motorcycle credit worthiness (Bussan Auto Finance case study) [Thesis]. Trunojoyo
[13] Pandeeswary P, Janaki M, "Performance Analysis of Big data Classification Techniques on Diabetes Prediction", International Journal of Innovative Technology and Exploring Engineering (IJITEE) 8 (10), 2019.

[14] Badr, Abdelkarim, Mohammed,"A comparative study of decision tree ID3 and C4.5 (IJACSA)”, International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, 2014.

[15] Venkatesh Naganathan, ‘Comparative Analysis of Big Data, Big Data Analytics: Challenges and Trends’, (JET), Vol: 5, Issue: 05, May-2018.

[16] Wahyuni S, "Implementation of Data Mining to Analyze Drug Cases Using C4.5 Decision Tree", Journal of Physics: Conf. Series 970 doi :10.1088/1742-6596/970/1/012030, 2018.

[17] David B A, Ari W, "C4.5 Decision Tree Implementation in Sistem Informasi Zakat (Sizakat) to Automatically Determining the Amount of Zakat Received By Mustahik", Journal of information systems, Vol 10 No 1, 2014.

[18] Saleh H, "Enterprise bankruptcy prediction using c4.5 algorithm based on forward selection", jurnal ilmiah ILKOM Vol 9, No 2, 2017.

[19] Elisa E, "Analysis and Application of C4.5 Algorithm in Data Mining to Identify Factors Causing PT.Arupadhatu Adisesanti's Construction Work Accidents", Journal of Informatic online Vol. 2 No. 1, 2018.

[20] Olson D. L, Delen D & Meng Y, “Comparative analysis of data mining methods for bankruptcy prediction”. Decision Support Systems, 52, 464–473, 2012.