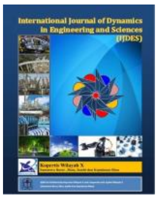
INTERNATIONAL JOURNAL of DYNAMICS
in ENGINEERING and SCIENCES (IJDES)
KOPERTIS WILAYAH X
Vol. 7, No. 2, Nov 2022
e-ISSN: 2502-0692



IMPLEMENTATION OF C4.5 ALGORITHM IN DATA MINING FOR PREDICTING SCHOLARSHIP RECIPIENT STUDENTS IN UNIVERSITY

Halifia Hendri[1], Harkamsyah Andrianof[2,*] Aggy Pramana Gusman[3]
[1, 2, 3] Sistem Komputer, Universitas Putra Indonesia YPTK Padang

10.22216/jod.v17i2
* Corresponding Email: harkamsyah.andrianof@upiyptk.ac.id

Submitted
10June 2022

Accepted :
15 Sept 2022

Published :
28 Sept 2022

Abstract

Every student who is enrolled in school wishes to receive a scholarship, especially those who come from low-income homes. The existence of a scholarship can relieve the financial load on parents who are paying for their child's college education and make the lecture process easier. However, the majority of students are unsure of how to apply for the award. Students, understandably, want to know what characteristics and conditions have the most impact on their ability to obtain a scholarship. The goal of this research is to see if the C4.5 approach can be used to forecast the number of students who will receive university scholarships. In research, the Rapid Miner software version 9.7.002 is utilized to process data. According to the findings of this study, 9 students (9%) will obtain a scholarship and 91 students (91%) will not win a scholarship out of 100 students

Keywords: Data Mining, C.45 Method, Students, Scholarship Recipients

INTRODUCTION

A student will undoubtedly require everything that will aid in the educational process, including equipment, transportation, communication and internet, food, fees, housing, books, and so on. The expense of schooling is one of the most crucial supporting factors. The cost of education is a collection of costs that students pay to the educational institution in which they are enrolled. Aside from that, costs are required to purchase equipment, transportation, food, and shelter, among other things [1], [2]. The tuition fees paid to educational institutions are the most crucial of the numerous items that must be financed to pursue education. Students' tuition fees are determined by the educational institution they attend. Some have a high price, while others have a low fee. The amount of the charge has no bearing on the quality of the education, nor does it depend on whether the school is public or private. The cost of education at the highest level, meaning college, has the highest average cost of all levels of education [3]–[5].

Universities with students are given names after them. Of course, the student must pay several fees (money) to the college where he or she is enrolled. Students reared in well-to-do homes or whose parents have a high income will find it simple to pay for their education, whereas students raised in low-income families or whose parents have low salaries will find it difficult to pay for their education. There are a variety of options available to students who are unable to pay their tuition fees, including scholarships. Scholarships are available from a variety of government, commercial, and public institutions, although the number is restricted.

All students will not be able to take advantage of all offered scholarships [6]–[8]. As a result, educational institutions always make a selection or selection among a large number of students who apply for the available scholarships.

The process of dredging, collecting, or mining essential information from a very big data set is referred to as data mining. To apply artificial intelligence technologies, this data mining procedure mostly employs statistical methodologies and mathematics. Educational Institutions use a variety of approaches to identify or select individuals who will get scholarships. Some people do it by hand, while others use computers. Many computerized methods for predicting scholarship recipients can be applied [9], [10]. The C4.5 approach is the one that the researcher employs. The C4.5 approach is a data mining methodology that can be used to reveal forecasts or predictions based on data.

After this study is done, the goal will be to determine ways to anticipate or predict which students will obtain scholarships utilizing the C4.5 approach for computer science students. Universitas Putra Indonesia YPTK Padang is a big and well-known private institution in Indonesia, particularly in the province of West Sumatra. The Faculty of Computer Science is one of the computer science faculties of this university (FILKOM). In this study, students who attended the computer science department at Putra Indonesia University YPTK Padang in 2019 were used as source data.

Scholarships for class 1 and 2 winners, Bidik Misi scholarships, BBM scholarships, PPA scholarships, and scholarships from major Indonesian banks are all available in the Faculty of Computer Science, Putra Indonesia University, Padang. All available scholarships are distributed evenly among the students who qualify. As a result, the process of selecting students who are eligible for scholarships is carried out. Predictive action employing data mining approaches, such as the C4.5 method, is required so that future students can learn what are the determining elements that decide whether or not they will win a scholarship.

Previous research that has been conducted and is relevant to this research is by Nurul Azwanti in 2018 who came from Putra University Batam, Riau Islands. Nurul Azwanti, a researcher from Putra University Batam in the Riau Islands [11], completed previous research that is related to this study in 2018. The results gained in this study are based on two tests, both the manual method and the WEKA software. It can be concluded that the test results are extremely good because the rules obtained are practically identical. The difference is in the value attribute that is entered into WEKA, but it has no bearing on the decision's outcome. WEKA uses 141 records, while the manual count uses 34. The study's flaw is that it only analyzes three sorts of data categories: GPA attributes, economic situations, and demographics.

The second research ever conducted was by Tukino in 2019 from Putra University, Batam, Riau Islands [12]. The outcomes of this investigation are six rules with a performance level of 92.60 percent +/- 1.36 percent on the C.45 algorithm. The discovered principles can be utilized as a basis for corporate managers to predict the achievement of profit targets, allowing them to anticipate by taking suitable business activities in the pursuit of profit. The study's main flaw is the short number of data lines analyzed, which totaled only 12 in all.

RESEARCH METHODS

The process of formulating or determining what critical and major problems that occur in the field that must be solved in this research is known as problem formulation [13], [14]. Literature study is the process of gathering or reviewing information from prior research and library books to solve a predetermined problem [15], [16]. The process of gathering data in the field for problem-solving is known as data collection. The data for this study came from the Vice Dean III of the Faculty of Computer Science at Putra Indonesia University YPTK Padang (WD III). The total number of lines of testing data that have been obtained is 100, which equals 100 students. The data collected by 100 students were used as data to make predictions using the C4.5 method. The data collected consists of 6 data attributes which can be seen in Table 1 below:

Table 1. Data Attribute

No	Data Attribute	Data Type
1	BP Number	Text
2	Name	Text
3	Parents Income (PO)	Integer
4	Commulative Index (GPA)	Integer
5	Parents Status (SO)	Text
6	Scholarship Receipt (Yes/No)	Binominal

RapidMiner version 9.7.002 was used to process the data, and the C4.5 method was used for data analysis. The steps of the C4.5 approach for data analysis are as follows:

a. Determine the data attributes that are used as root nodes or predictions in the decision tree and count the number of YES and NO values in each data row.

b. Determine the branch from the root (root) for each value after determining the root of the decision tree by calculating the Gain value.

Gain Formula is [17], [18]:

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

Note: $Gain(S, A) =$ Total Gain value with the attribute, $Entropy(S) =$ Total Entropy Value, $Entropy(S_i) =$ Entropy Value of each attribute, $n =$ number of *cluster*

Entropi formula is [19], [20]:

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2\, pi \qquad (2)$$

Note: $Entropy(S) =$ Total Entropy Value, $pi =$ proportion of S_i for S

c. Split cases on each existing branch form

d. Repeat steps 2 to step 3 on each branch until each branch case gets the same class.

The C4.5 approach is used to process and generate data, which is then examined and concluded. How much data is required to determine which students are likely to receive scholarships and which students are unlikely to receive scholarships using the C4.5 approach among all the data that has been processed? Rapid Miner software version 9.7.002 was used to process the data. Data analysis results and conclusions are implemented in the field, which is highly beneficial and assists an institution, particularly educational institutions, in forecasting whether or not their students will receive scholarships.

RESULTS AND DISCUSSION

A. Result

As shown in 2 below, the researcher collected 100 data lines or 100 scholarship recipients who were not recipients of the previous year's scholarship. Table 2 shows the first data that was collected, which consisted of 6 attributes and 100 data records:

Table 2. Prelimanary Data

No	BP Number	Name	Parents Income (PO) (Rp./Mounth)	Commulative Index (GPA)	Parents Status (SO)	Scholarship Receipt
1	17..027	Williem Kusnedi	2.000.000	3,33	Have	Yes
2	18..007	Wira Wahyuni	800.000	2,86	Haven't	Yes
3	18..020	Raesa Islamiati	1.000.000	3,87	Haven't	No
4	18..022	Yogi Hensyah	1.500.000	3,42	Have	Yes
5	18..029	Wirda Jihadita	1.500.000	3,74	Have	Yes
….	…	…	…	…	…	….
100	17..444	Admel Brina	2.000.000	3,25	Have	Yes

The C4.5 approach is then used to forecast which students will receive the scholarship the following year based on the initial data. The steps of the C4.5 algorithm, as well as the outcomes of the calculations, are as follows:

Step 1: Count the number of YES and NO values in each data row for the data attribute that will be utilized as a root node or decision tree prediction. Table 2 shows the data that has been labeled YES and NO, as well as the addition of criteria to be eligible for a scholarship.

Table 3. Prediction of Data Mining

No	PO	Group	GPA	Group	SO	MB
1	3	Low	3.33	Low	Have	Yes
1	3	Low	3.33	Low	Have	Yes
2	3	Low	3.27	Low	Haven't	Yes
3	1	High	3.35	Low	Have	No
4	3	Low	3.4	Low	Have	Yes
5	3	Low	3.35	Low	Have	Yes
….	….	….	….	….	….	….

100	3	Low	3.1	Low	Have	Yes

Step 2: After identifying the root of the decision tree by computing the Gain value, determine the branch from the root (root) for each value. The Gain value in one of the data rows is calculated as follows:

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

$$Gain(PO,A) = 0.905 - (0.77 * 0.806) = 0.2844 \qquad Gain(IPK,A) = 0.905 - (0.94 * 0.925) = 0.0355$$

$$Gain(Status,A) = 0.905 - (0.78 * 0.977) = 0.1429$$

Step 3: For each existing branch, divide the cases.

Step 4: For each branch, repeat steps 2–3 until all cases in the branch have the same class.

Because the C4.5 method's calculation is performed several times, it necessitates the use of data mining software to speed up and simplify calculations. Rapid Miner, version 9.7.002, was utilized as the application. The results of prediction calculations using the C4.5 algorithm are shown in table 4 below.

Table 4. Prediction of Data

NO		True (YES)	True (No)	Class Precission
1	pred. YES	5	4	55.56%
2	pred. NO	0	1	100.00%
3	class recall	100.00%	20.00%	
4	pred. YES	5	4	55.56%

Accuracy: 80.00%

The projected data above can be shown as a graph (plot view) so that the prediction results can be easily seen, as shown in Figure 2 below:

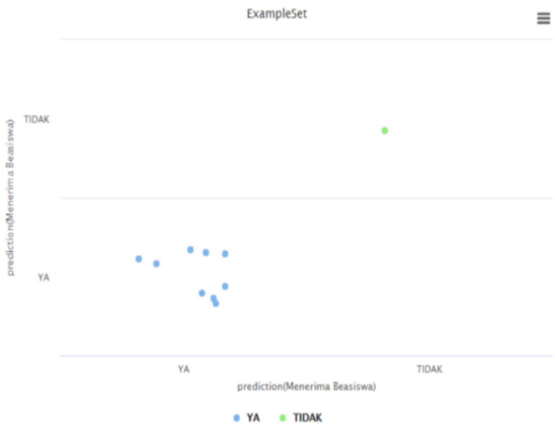


Figure 2. Graphical form (plot view) of predicted data

In order to process data in the Rapid Miner application, you'll need a block design that outlines the data processing sequence. The design of the data processing block using the Rapid Miner application can be seen in Figure 3 below to generate data as mentioned in table 3 and also figure 3:

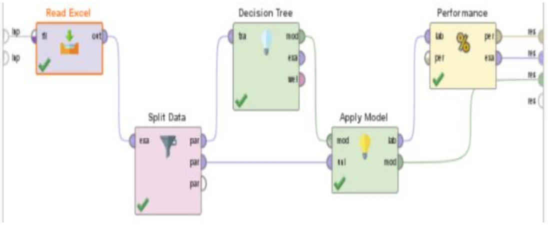


Figure 3. Design block Method C4.5 RapidMiner application

The outcomes of the table 3 predictions can take the shape of statistics or data conclusions. The data is presented in the form of a description decision tree text view in the following.

Tree

```
PO = Low
|   STATUS = HAVE
|   |   GPA = LOW: YES {YES=2,
NO=0}
|   |   GPA = HIGH: YES {YES=41,
NO=15}
|   STATUS = TIDAK: YA {YA=11,
TIDAK=0}
PO = VERY LOW: YES {YES=9, NO=0}
PO = HIGH: NO {YES=0, NO=12}
```

Figure 4 shows the form of the decision tree based on the prediction results:

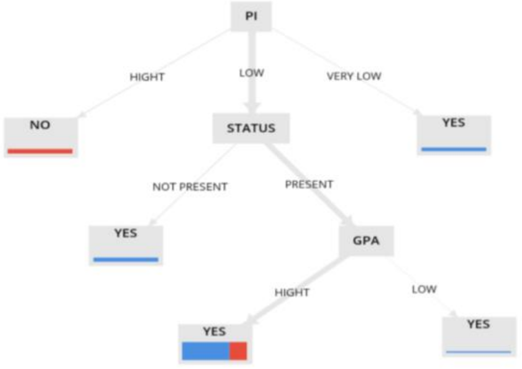


Figure 4. Decision Tree Predicted Results

B. Discussion

According to the data in table 4, the results of processing the prediction data using the C4.5 approach utilizing the Rapid Miner program reveal that out of a total of 100 student data processed, 9 students with a value of YES received scholarships and 91 people with a value of NO did not. Figure 2 shows a plot view of the data, which shows points in the form of scholarship recipient forecasts. Students who are projected to obtain YES scholarships are represented by blue plots, while students who are predicted to not receive the scholarships are represented by green plots. The number of scholarship recipients is represented by the number of points.

The Design Block Method C4.5 for the RapidMiner application is shown in Figure 3. The RapidMiner application requires input data in the first block, after which the input data is split / divided into two blocks, the decision tree block and the apply model block, which are then combined back into the final block, the performance block, to determine the percentage of data accuracy. In Figure 4, the decision tree shows that the first step in anticipating the process flow is to look at the parental income (PO), which might be low, extremely low, or large. Check to see if the student's parental status is still active if the parent's income is low.

CONCLUSION

To make the resulting decision tree easier to understand and evaluate, the data should be converted to polynomial data or given a name in advance based on the data group (cluster). Furthermore, it was shown that 9 students (9%) were expected to win scholarships and 91 students (91%) did not receive scholarships out of 100 students whose data was examined. Based on the conclusions obtained, several things are recommended for future research, namely: We recommend using more attributes than we have done so that the results of clustering and predictions are better. We recommend that the data attribute for clustering with the K-Means method is in the form of numerical data and then the results are used as polynomial data so that they can be used for predictions of the C4.5 method.

BIBLIOGRAPHY

[1] K. Vaughan *et al.*, “Immunization costs, from evidence to policy: Findings from a nationally representative costing study and policy translation effort in Tanzania,” *Vaccine*, vol. 38, no. 48, pp. 7659–7667, 2020, doi: 10.1016/j.vaccine.2020.10.004.

[2] J. T. Kraiss, B. Wijnen, R. W. Kupka, E. T. Bohlmeijer, and J. Lokkerbol, “Economic evaluations of non-pharmacological interventions and cost-of-illness studies in bipolar disorder: A systematic review,” *J. Affect. Disord.*, vol. 276, no. May, pp. 388–401, 2020, doi: 10.1016/j.jad.2020.06.064.

[3] M. Mongioi, J. Bodzio, K. Eck, and A. Levine, “Nutrition Education and College Students’ Nutrition-Related Knowledge, Attitudes and Behaviors,” *J. Acad. Nutr. Diet.*, vol. 121, no. 9, p. A57, 2021, doi: 10.1016/j.jand.2021.06.165.

[4] A. G. Forster, H. G. van de Werfhorst, and T. Leopold, “Who benefits most from college? Dimensions of selection and heterogeneous returns to higher education in the United States and the Netherlands,” *Res. Soc. Stratif. Mobil.*, vol. 73, no. February 2020, p. 100607, 2021, doi: 10.1016/j.rssm.2021.100607.

[5] J. Lee, N. Kim, and M. Su, “Immigrant and international college students’ learning gaps: Improving academic and sociocultural readiness for career and graduate/professional education,” *Int. J. Educ. Res. Open*, vol. 2–2, no. January, p. 100047, 2021, doi: 10.1016/j.ijedro.2021.100047.

[6] N. A. Matrose, K. Obikese, Z. A. Belay, and O. J. Caleb, "Equality of opportunity and human capital accumulation: Motivational effect of a nationwide scholarship in Colombia," *Sci. Total Environ.*, p. 135907, 2019, doi: 10.1016/j.jdeveco.2021.102754.

[7] T. Naidu, "Says who? Northern ventriloquism, or epistemic disobedience in global health scholarship," *Lancet Glob. Heal.*, vol. 9, no. 9, pp. e1332–e1335, 2021, doi: 10.1016/S2214-109X(21)00198-4.

[8] R. Kleinpell, B. B. Kennedy, M. Piano, and L. D. Norman, "Advancing Clinical Scholarship among Non-tenure Track Faculty: A Faculty Scholarship Program," *J. Prof. Nurs.*, 2021, doi: 10.1016/j.profnurs.2021.08.008.

[9] V. Mittal and S. Sridhar, "Customer based execution and strategy: Enhancing the relevance & utilization of B2B scholarship in the C-suite," *Ind. Mark. Manag.*, vol. 88, no. May, pp. 396–409, 2020, doi: 10.1016/j.indmarman.2020.05.036.

[10] K. Schmiedeknecht, M. Perera, E. Geoffroy, E. Schell, S. Rankin, and J. Jere, "Predictors of workforce retention among malawian nurse graduates from the GAIA nursing scholarship program: A mixed methods study," *Ann. Glob. Heal.*, vol. 81, no. 1, p. 51, 2015, doi: 10.1016/j.aogh.2015.02.626.

[11] N. Azwanti, "C4.5 Algorithm For Predicting Students Who Repeat Courses (Case Study In Amik Labuhan Batu)," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 9, no. 1, pp. 11–22, 2018, doi: 10.24176/simet.v9i1.1627.

[12] T. Tukino, "Application of the C4.5 Algorithm to Predict Profits at PT SMOE Indonesia," *J. Sist. Inf. Bisnis*, vol. 9, no. 1, p. 39, 2019, doi: 10.21456/vol9iss1pp39-46.

[13] H. Hendri, H. Awal, and Mardiosn, "Indonesian Journal of Computer Science (IJCS)," *STMIK Indones. Padang*, vol. 8, no. 2, p. 121, 2019.

[14] H. Hendri, H. Awal, and Mardison, "Solar-Cell Implementation for Supporting Tourist Facilities and Tourism Promotion Media," *J. Phys. Conf. Ser.*, vol. 1783, no. 1, p. 012058, 2021, doi: 10.1088/1742-6596/1783/1/012058.

[15] I. Zasada *et al.*, "A conceptual model to integrate the regional context in landscape policy, management and contribution to rural development: Literature review and European case study evidence," *Geoforum*, vol. 82, no. March, pp. 1–12, 2017, doi: 10.1016/j.geoforum.2017.03.012.

[16] J. Fulford *et al.*, "The neural correlates of visual imagery vividness – An fMRI study and literature review," *Cortex*, vol. 105, pp. 26–40, 2018, doi: 10.1016/j.cortex.2017.09.014.

[17] H. Wang and Y. Gao, "ScienceDirect ScienceDirect Research on algorithm improvement improvement strategy strategy based based on on MapReduce MapReduce," *Procedia Comput. Sci.*, vol. 183, pp. 160–165, 2021, doi: 10.1016/j.procs.2021.02.045.

[18] A. Joshuva, R. S. Kumar, S. Sivakumar, G. Deenadayalan, and R. Vishnuvardhan, "An insight on VMD for diagnosing wind turbine blade faults using C4 . 5 as feature selection and discriminating through multilayer perceptron," *Alexandria Eng. J.*, vol. 59, no. 5, pp. 3863–3879, 2020, doi: 10.1016/j.aej.2020.06.041.

[19] C. L. H. Chen, P. L. M. Ku, and L. O. C. Chuang, "Smart Dynamic Resource Allocation Model for Patient-Driven Mobile Medical Information System Using C4 . 5 Algorithm," *J. Electron. Sci. Technol.*, vol. 17, no. 3, pp. 231–241, 2019, doi: 10.11989/JEST.1674-862X.71018117.

[20] X. Wang, C. Zhou, X. Wang, C. Zhou, and X. Xu, "ScienceDirect ScienceDirect Application of C4 . 5 decision tree for scholarship evaluations Application of C4 . 5 decision tree for scholarship evaluations," vol. 00, no. 2018, 2019, doi: 10.1016/j.procs.2019.04.027.