Used of the C.45

by Supratman Zakir

Submission date: 06-May-2023 06:21AM (UTC+0700)
Submission ID: 2085511215
File name: C45_Algorithm_2021_J._Phys.__Conf._Ser._1779_012009.pdf (823.49K)
Word count: 3123
Character count: 16723

Journal of Physics: Conference Series

Use of the C4.5 Algorithm in Determining Scholarship Recipients

Use of the C4.5 Algorithm in Determining Scholarship Recipients

Rina Novita[1], Supratman Zakir[2], Agus Nur Khomarudin[3], Efmi Maiyana[4], Hamimah Hasyim[5]

[1,2,3]Institut Agama Islam Negeri (IAIN) Bukittinggi, Bukittinggi, Indonesia
[4] Akademi Manajemen Informatika & Komputer (AMIK) Boekittinggi, Bukittinggi, Indonesia
[5] Faculty of Education, University Teknologi Mara (UiTM), Malaysia

* rinanovita@iainbukittinggi.ac.id

Abstract. The importance of education is increasingly felt by the community. Various attempts were made by parents to continue education for their children. This is in line with the government's efforts to equalize education. Currently, there are quite a lot of scholarships provided by the government in developing education. One of these scholarships is the Bidik Misi Scholarship. This scholarship is in great demand by prospective students because the period for giving is 4 years and is given every semester. Every year the number of applicants for this scholarship continues to grow. The process of selecting scholarship recipients is getting more complicated due to a large number of candidates for scholarship recipients. The purpose of this research is to develop a system that can assist in selecting prospective scholarship recipients. The method used is Research and Development (RnD) with 4D (Define, Design, Develop, and Disseminate). The system development takes advantage of data mining in classifying the pile of scholarship applicant data using the C4.5 algorithm. The result of the product test shows that the value of the test of vitality is 0.91 which means valid, the practicality test is 0.89 which means that it is practical and the value of the effectiveness test is 0.93 which means that it is effective. These results are shown by the research results obtained from 7 rules with an accuracy rate of 72.5%.

Keyword: C4.5 Algorithm, Scholarship, Research and Development, Data Mining

1. Introduction

Public awareness of the importance of education is increasingly being felt. Economic reasons are not a barrier to not continuing education. Especially now that there are so many scholarships offered by the government. One of the scholarships that are very popular with the community is the Bidikmisi scholarship. Bidikmisi scholarships are a form of assistance from the Indonesian government in the field of education. Assistance in the form of tuition fees provided to prospective students who have good academic achievements but have economic limitations. This scholarship is given to prospective students until graduation on time or 4 years for S1 and 3 years for D3 programs [1]. This scholarship has been started since 2010. IAIN Bukittinggi as one of the state universities of Islamic religion in the city of Bukittinggi also provides opportunities for prospective students to study well through this Bidikmisi scholarship program.

Every year the number of applicants for the Bidikmisi scholarship at IAIN Bukittinggi is increasing. Prospective students are allowed to register either by admitting new students through the SPAN, UM-PTKIN, and UM-Mandiri pathways. The number of Bidikmisi scholarship enthusiasts that continues to increase has resulted in data accumulation. Stored data can be reprocessed to obtain important

information using Data Mining technology. Application of data mining technology [2]. Data Mining is a process in the Knowledge Database in Discovery (KDD). KDD is an organized process to identify valid, new, useful and understandable patterns contained in a large and complex data set [3]. KDD consists of several stages, namely: [4]

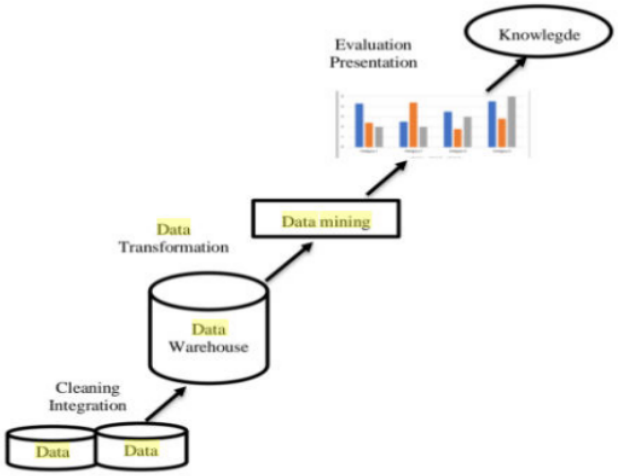

Figure 1. Stages of KDD Process

The stages in KDD begin with data cleaning. The next stage is data integration, which is to integrate data against inconsistent data and correct errors in the data. Next, perform data transformation, data mining processes, evaluate patterns, and finally present knowledge. Several techniques can be applied to extracting important information in data mining. One of them is a classification technique by applying the C4.5 algorithm.

The results of processing data mining classification techniques using the C4.5 Algorithm will form a decision tree [2]. In order to form a decision tree, the first step in the c4.5 algorithm process is to determine the attributes that will serve as the root. The root selection is based on the calculation of the highest gain valu. After the roots are formed, then proceed to make a branch for each value. The process will be repeated until all cases on the branch have the same class.

Researches that have been carried out related to the application of the C4.5 algorithm were also carried out by Mahindra Suryaning Praja and Erna Zuni Astuti regarding the scholarship recommendations at SMA N 1 Mlonggo. In research, scholarship recommendations in high school require data on majors, grade, grade, parental stages, and attribute number of siblings. The attribute number of values above the average becomes the root of the formed decision tree [5]. Nanda Dimas Prayoga also conducted research on the application of the C4.5 Algorithm in predicting graduation on time in college. Some of the factors that cause students to not graduate on time are low GPA, incomplete SKS, bad ethics, and the absence of student achievement. The results of these studies help students to graduate on time and assist academics in making academic rules for students to graduate on time [6]. David Hartanto Kamagi and Seng Hansum have also implemented the C.45 algorithm to predict the pass rate of students. David and Seng Hansum used GPA, Achievement, Ethics, and SKS data as attributes. From this study, it can be seen that the attributes of SKS and GPA are very influential in determining student graduation[7].

Other research on the implementation of the c4.5 algorithm to determine scholarship recipients was also carried out by Rismayana in 2016. Rismayana conducted research in determining recipients of Academic Achievement Improvement (PPA) scholarships given to outstanding students. There are 4 predictor attributes used in the research, namely GPA, semester, parent's income, and the number of parental dependents. Of the 4 attributes, the GPA attribute is a determining factor in predicting a decision where the GPA is Cum Laude and is Very Good with a large number of parental dependents. Meanwhile, the classification of students who have a Good GPA and a Very Good GPA with the number of dependents of a parent with a value of Enough is not rejected in the PPA scholarship[8].

The application of Data Mining using another C4.5 algorithm was also carried out by Erlin Elsa in identifying the factors causing the construction work accident of PT. Arupadhatu Adisesanti. The results of Erlin Elisa's research found several factors causing work accidents including the workplace environment, safety, and worker signs, and work methods. Worker factors and work methods are the roots of the decision tree formed [9]. Also, the application of the C4.5 algorithm for the classification of students with the potential to drop out has also been carried out by Asmaul Hasanah Nasrullah in the scientific journal Volume 10 of 2018. Each year there are approximately students dropping out. The attributes used for the classification of potential dropout students are gender, age, religion, regional origin, class, semester I social studies scores, semester II social studies scores, semester III social studies scores, and semester IV social studies scores. From the results of research conducted by Hasanah, it is known that the IPS semester IV scores become the root of the decision tree that is formed. There are 17 rules that can be used as a class/group in determining students who have the potential to drop out before being accepted as students at the tertiary institution[10].

The application of the C4.5 Algorithm in this study was to determine the recipients of the Bidikmisi scholarship at the IAIN Bukittinggi college. So far, the process of determining scholarship recipients has been carried out by following the rules and weightings set by the academic division. The amount of data on the incoming bidikmisi scholarship candidates will be inputted one by one and obtained by the manager. This certainly requires a relatively long time and effort. The results of this research are expected to be a solution in determining bisikmisi scholarship recipients at IAIN Bukittinggi by following the pattern of rules formed from the decision tree. From this research, the classification of students recommended receiving Bidikmisi scholarships will also be obtained.

2. Methodology

The research methodology used is the ADDIE version of Research and Development (RnD) (Analyze, Design, Develop, Evaluate, Implement). The 4D stages can be seen in Figure 2 below:

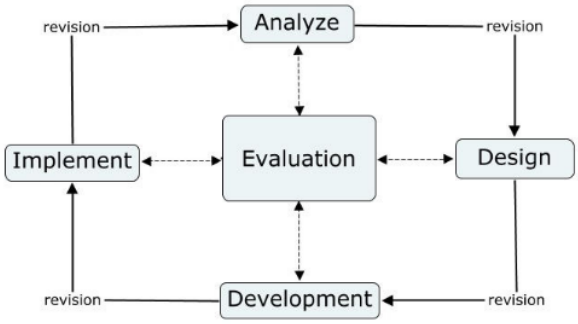

Figure 2. The Flow of Research Methods

Analyze; determine the substance of the problem in selecting prospective scholarship recipients. Some of the problems found were the effects of several aspects that caused these problems to occur, such as the number of candidate recipient files, non-standard selection methods and the time required for selection was quite long[11]. Design; make a general design system that will be developed with the help of a system development tool, namely the Unified Modeling Languag(UML). Develop; The data that has been collected is then processed with several stages contained in the Knowledge Discovery in Database (KDD) as follows:

a. Data Cleaning, so that the data obtained is avoided from incomplete, and inconsistent, the data cleaning process is carried out first.
b. Data Selection, in this stage, data selection is carried out on unused attributes in the data processing process. Attributes that do not contribute to the research objectives are not used in this study.
c. In data transformation, knowledge is extracted from real data. Data is changed or combined into a suitable format that can then be processed in data mining.
d. Data mining, data mining is carried out to obtain new knowledge. The data mining technique used in this study is a classification technique using the C4.5 algorithm.

The next stage is to analyze the results of data processing. The data that has been processed using classification techniques using the C4.5 algorithm is then analyzed. After analyzing the results of data processing, testing is then carried out. Tests were carried out using WEKA open-source data mining software. The final stage in this research is drawing concluding on the results of the research and providing recommendations with the aim that weaknesses can be eliminated and future solutions can be implemented according to the expected objectives.
Implement; Perform testing or use directly on products that have been designed. Evaluate; Evaluating and maintaining the products that have been made.

3. Result and Discussion

Application of Data Mining in this study aims to obtain a classification of prospective new students who are eligible for Bidikmisi scholarships. There are 4 attributes used in this study. The first attribute is the economic condition or income of the parents in one month. The second attribute used in this study is the average national exam results. The next attribute used in this study is the average report card value and ranking each semester by seeing an increase or a decrease.

The sample data used in this study were 40 data. Data processing begins with the Data Cleaning process for invalid, incomplete, and inconsistent data. Furthermore, the Data Selection process is carried out by selecting which data will be used in this study. The selection result data which is nominal value is then transformed according to the criteria set by the academic party. The results of the Data Transformation process are shown in table 1 below:

Table 1. Attributes of Bidikmisi Scholarship Recipient Eligibility Classification

NO.	ATTRIBUTE	ATTRIBUTES VALUE
1.	Economic / Income Conditions	Big Intermediate Small
2.	Average national exam results	High Moderate Small

3	Average report card scores and semester rankings	Ride Down
4	Home Status	Owned - paid off Owned - in installments Family house/boarding Contract

The next stage is the Data Mining process in the classification of Bidikmisi scholarship recipients using the C4.5 algorithm. The steps for solving the C4.5 algorithm are in the following flowchart Figure 3 [12]:

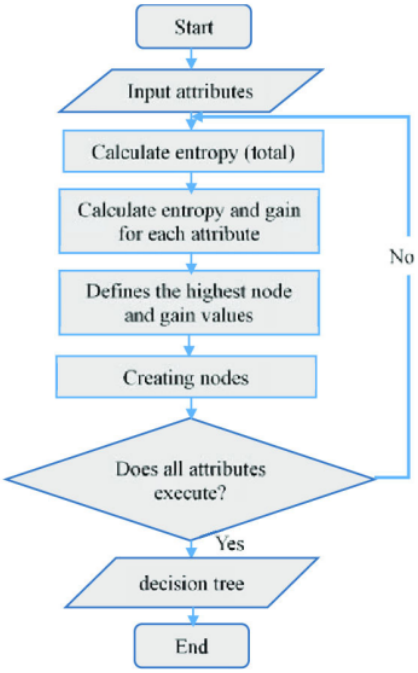

Figure 3. Flowchart C4.5 Algorithm

Based on Figure 2, it can be seen that the stages in the c4.5 algorithm are as follows [13]:
1. Prepare data that will be used as training data. The training data is based on past data or historical data obtained previously. Data will be grouped into certain classes.
2. Calculate the value of each attribute that will serve as the root. The selection is based on the attribute that gets the highest gain value from the existing attributes. Before calculating the gain value of the attribute, first, calculate the entropy value. To calculate the entropy value, the formula is used:

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

Information :
S: case set
N: the number of partitions S
Pi: the proportion of Si to S

After the entropy value is obtained, the next step is to calculate the gain value. The calculation of the gain value uses the following formula:

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

Information:
S: case set
A: attribute
N: the number of partitions attributes A
[Si]: number of cases on the ith partition
[S]: number of cases in S

3. Divide the case of steps 2 and 3 until all records are partitioned
4. The process will stop when:
   a. The same value is obtained for all records in node N;
   b. There are no attributes present on the record to be partitioned again;
   c. There are no more records in the empty branch.

The results of data mining processing in determining scholarships using the C4.5 algorithm in this study are shown in Figure 4 below:

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          29               72.5    %
Incorrectly Classified Instances        11               27.5    %
Kappa statistic                          0.4359
Mean absolute error                      0.268
Root mean squared error                  0.4237
Relative absolute error                 53.8762 %
Root relative squared error             84.7828 %
Total Number of Instances               40

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,818    0,389    0,720      0,818    0,766      0,441    0,835     0,840     YA
                 0,611    0,182    0,733      0,611    0,667      0,441    0,835     0,840     TIDAK
Weighted Avg.    0,725    0,296    0,726      0,725    0,721      0,441    0,835     0,840

=== Confusion Matrix ===

  a  b   <-- classified as
 18  4 |  a = YA
  7 11 |  b = TIDAK
```

Figure 4. Results of the C4.5 Algorithm Processing in the Classification of Bidikmisi

The decision tree resulting from the calculation of the c4.5 algorithm in the classification of Bidikmisi scholarship recipients is shown in Figure 5 below:

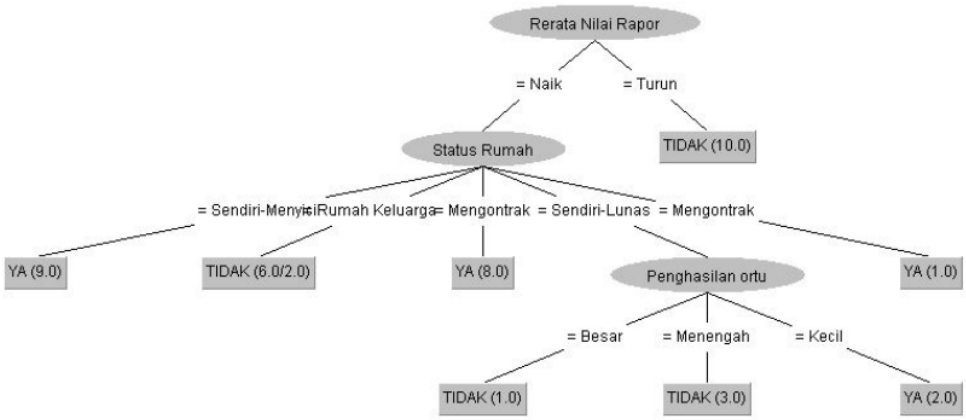

Figure 5. Decision Tree C4.5 Algorithm for Bidikmisi Scholarship Recipients

Based on the decision tree in Figure 5, the following rules are obtained:
The Decision for the results of being eligible for a scholarship:

a. IF Average Report Card Value = Increase and House Status = Own-Installment Then Recommendation = **YES**
b. IF Average Report Card Value = Up and House Status = Contracting Then Recommendation = **YES**
c. IF Average Report Card Value = Increased and House Status = Self-Paid and Parents' Income = Small Then Recommendation = **YES**

The decision for the result is NOT entitled to receive:

a. IF Average Report Card Value = Decrease Then Recommendation = **NO**
b. IF Average Report Card Value = Increased and Status of House = Family House / Ride Then Recommendation = **NO**
c. IF Average Report Card Value = Increased and House Status = Self-Paid and Parents' Income = Large Then Recommendation = **NO**
d. IF Average Report Card Value = Increased and House Status = Self-Paid and Parents' Income = Intermediate Then Recommendation = **NO**

4. Conclusion

Based on the results of testing conducted on the data of prospective students who are entitled to receive Bidik Misi scholarships at IAIN Bukittinggi, it can be concluded that the classification technique using the C4.5 algorithm is quite good. The resulting accuracy rate is 72.5%. The attribute that is the main factor in granting scholarship eligibility is the average report card score. The average grade of the semester report cards that went up received recommendations for eligibility to receive Bidik Misi scholarships and was supported by home status and parents' income.

References

[1] Wikipedia, "Beasiswa Bidikmisi." https://id.wikipedia.org/wiki/Beasiswa_Bidikmisi.
[2] K. & E. T. Luthfi, *Algoritma Data Mining*, Edisi I. Yogyakarta: CV Andi Offset, 2009.
[3] L. R. Oded Maimon, Ed., *Data Mining and Knowledge Discovery Handbook*, 2nd ed. USA: Springer Science+Business Media, 2010.
[4] E. Buulolo, *Data Mining Untuk Perguruan Tinggi*, Edisi I. Yogyakarta: Deepublish Publiser, 2020.

[5] E. Z. A. Mahindra Suryaning Praja, "PENERAPAN DATA MINING UNTUK REKOMENDASI BEASISWA PADA SMA N 1 MLONGGO," *Skripsi, Univ. Dian Sastro, Semarang*, 2016.
[6] N. D. Prayoga, "Penerapan Algoritma C.45 Dalam Memprediksi Kelulusan Tepat Waktu Pada Perguruan Tinggi (Studi Kasus : Stmik Royal Kisaran)," 2018, doi: 10.31227/osf.io/unqt4.
[7] D. H. Kamagi and S. Hansun, "Implementasi Data Mining dengan Algoritma C4 . 5 untuk Memprediksi Tingkat Kelulusan Mahasiswa," vol. VI, no. 1, pp. 15–20, 2014.
[8] Rismayanti, "Implementasi Algoritma C4.5 Untuk Menentukan Penerima Beasiswa Di Stt Harapan Medan," *Media Infotama*, vol. 12, no. 2, pp. 116–120, 2016.
[9] E. Elisa, "Analisa dan Penerapan Algoritma C4 . 5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT . Arupadhatu Adisesanti," vol. 2, no. 1, pp. 36–41, 2017.
[10] A. H. Nasrullah, "Penerapan Metode C4.5 untuk Klasifikasi Mahasiswa Berpotensi Drop Out," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 244–250, 2018, doi: 10.33096/ilkom.v10i2.300.244-250.
[11] Supratman, T. Arianto, and E. Maiyana, "Development of Local Web-Based Learning (LWBL) as Low-Cost Digital Learning Efforts," *J. Phys. Conf. Ser.*, vol. 1471, no. 1, 2020, doi: 10.1088/1742-6596/1471/1/012003.
[12] N. Anwar, A. Pranolo, and R. Kurnaiwan, "Grouping the community health center patients based on the disease characteristics using C4.5 decision tree," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 403, no. 1, 2018, doi: 10.1088/1757-899X/403/1/012084.
[13] A. Andriani, "Sistem Prediksi Penyakit Diabetes Berbasis Decision Tree," *Bianglala Inform.*, vol. 1, no. 1, pp. 1–10, 2013.

Used of the C.45

ORIGINALITY REPORT

18%	17%	18%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	risbang.unuja.ac.id Internet Source	5%
2	elar.urfu.ru Internet Source	2%
3	jurnal.polgan.ac.id Internet Source	2%
4	infor.seaninstitute.org Internet Source	1%
5	"Prediction using C4.5 Method and RFM Method for Selling Furniture", International Journal of Engineering and Advanced Technology, 2019 Publication	1%
6	Sepsa Nur Rahman, Suparmi, Annisak Izzaty Jamhur, Yesri Elva, Surmayanti, Eva Rianti. "Comparison of the Effectiveness of C.45 Algorithm with Naive Bayes Algorithm in Determining Scholarship Recipients", 2021 International Conference on Computer Science and Engineering (IC2SE), 2021 Publication	1%

7	digilib.esaunggul.ac.id Internet Source	1 %
8	M. Lintang, N. Pandiangan, D. Hyronimus. "Use of the C4.5 Algorithm to Analyze Student Interest in Continuing to College", SHS Web of Conferences, 2022 Publication	1 %
9	mail.ijair.id Internet Source	1 %
10	jurnal.polines.ac.id Internet Source	1 %
11	repository.pnj.ac.id Internet Source	1 %
12	Fitriana Harahap, Evri Ekadiansyah, Erwin Ginting, Nidia Enjelita Saragih, Robiatul Adawiyah, Ermayanti Astuti. "Factors of the Decrease in Student's Interest in Learning During the Covid-19 Pandemic Using the C4.5 Method", 2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS), 2021 Publication	1 %
13	litapdimas.kemenag.go.id Internet Source	1 %

Exclude quotes	On	Exclude matches	< 20 words
Exclude bibliography	On		