P-ISSN: 2715-2448 | E-ISSN: 2715-7199
Vol.6 No.1 January 2025
Buana Information Technology and Computer Sciences (BIT and CS)

Implementation of the C4.5 Algorithm in Predicting the Interest of Prospective Students in Choosing Higher Education

Bambang Triraharjo[1]*, Prilian Ayu Minarni[2], Baskoro[3]
[1,3]Universitas Muhammadiyah Pringsewu/Fakultas Ilmu Komputer, Pringsewu,Lampung, Indonesia
[2]Universitas Muhammadiyah Pringsewu/Fakultas Kesehatan, Pringsewu,Lampung, Indonesia
E-mail: bambangtriraharjo@umpri.ac.id, prilianayuminarni@umpri.ac.id, baskoro@umpri.ac.id

Received: 2024-07-02 | Revised: 2024-10-30 | Accepted: 2025-01-31

Abstract

Lampung Province has many diverse private universities and offers a variety of majors. Due to the large number of private universities that exist, competition between universities to attract prospective new students is very tight. So, to be able to compete with these universities, the campus needs to predict the interests of prospective new students by knowing what factors motivate prospective new students to choose a university. The aim of this research is to predict prospective students' interest in choosing a university. In this research, the data processed are the results of a survey conducted on prospective new students in the Information Systems and Technology Undergraduate Study Program. Data from prospective students will be processed using a Data Mining process with the Classification Method. Furthermore, using the C4.5 algorithm in the Classification method, literacy was obtained up to node 4 with 3 assessment data which became a factor in determining prospective students' interest in the study program. Furthermore, the results of data processing using the classification method with the C4.5 algorithm were tested using Rapid Software. Miner, where an accuracy rate of 100% is obtained. The final result of using Data Mining with the C4.5 algorithm is that it is able to predict the interests of prospective new students based on assessment factors in choosing a university.

Keywords: *Prediction, Classification, Data Mining, C4.5 Algorithm, Decision Tree*

I. Introduction

The development of a university, one of which is seen based on the number of students obtained every year. In universities, especially private, the acquisition of students every year is the main factor in the development of the university. Because, students are the main resource for private universities to get funds in carrying out operations in universities. The more students, the greater the income for the university, so that it is easy for the university to run campus operations without limited funds. Currently, the Information Systems and Technology Study Program, University of Muhammadiyah Pringsewu is a new program majoring in Computer Science. In 2021-2022, the Information Systems and Technology Study Program will only start opening registration. So currently the Study Program is trying its best to get prospective new students to be able to continue their education at the Information Systems and Technology Study Program, University of Muhammadiyah Pringsewu.

In choosing a university, a student will usually look for information about the university they are going to [8]. Apart from that, there are also factors that affect prospective students in choosing a university, including factors that influence parents, friends, relatives, scholarships, job opportunities and others. With these factors, universities can predict what factors are the main drivers for prospective students to choose universities, so that universities can make the best decisions to be able to recruit as many prospective students as possible [3]. Interest can be interpreted as a high tendency and passion or a great desire for something. The purpose of this study is to find out the interest of prospective new

students in a university and how to apply the Data Mining process to predict the request of prospective students based on the assessment factors they use [6].

Data Mining can also be interpreted as the extraction of new information extracted from chunks of big data that helps in decision-making [15]. Data mining is part of Knowledge Discovery in Database (KDD) which consists of several stages such as data selection, data pre-processing, transformation, data mining and evaluation of results [11]. By using data mining, valuable information in the data set can be mined. By using data mining, data on the interests of prospective new students can be processed with an algorithm [10]. In data mining there are algorithms that are able to analyze data. In this study, the algorithm used is the C4.5 algorithm. The C4.5 algorithm is a classification and prediction method used to form a Decission Tree based on training data [9]. A decission tree is a flowchart structure that has a Tree, where each internal node indicates an attribute test, each branch represents the test result and the leaf node represents the class or class distribution [7].

In the previous study, the Theory of Planned Behavior approach was used in predicting student interest[5]. This approach measures students' interest in entrepreneurship by prioritizing 3 determinants of desire to be entrepreneurial, namely Attitudes Towards Behavior, Subjective Norms, and Behavior Control. The result of this approach is a conclusion where attitudes towards behavior do not affect students' interest while 2 other factors do [6]. Another algorithm used in previous research to predict the interest of vocational school students to enter college is the Naive Bayes algorithm. The data used is graduate data in 2018 and 2019 where there are 158 data. Furthermore, the data was processed using the Naive Bayes algorithm with 9 attributes, namely expertise, report card scores, UN scores, parental work, and others. The data was tested using the RapidMinner application [2]. The results of the processor were obtained from the prediction of parental income attributes, report card scores, and student desires, which affected students' interest in continuing their education to higher education with an accuracy rate of 92.96% [12].

In previous research, the C4.5 algorithm has been used to predict the factors that cause students to repeat a course. In the study, variables were determined in the form of Number of Semesters, GPA, Grades, Economic Condition and Status. Furthermore, data processing is carried out with the C4.5 algorithm as many as 141 data records as training data. After data processing using the C4.5 algorithm, tests were then carried out using the WEKA application. The result of this study is information in the form of rules that can predict students who repeat a course [1]. Furthermore, the C4.5 algorithm is also used for the classification of student success rate predictions at AMIK Tunas Bangsa. In this study, several attribute variables were determined such as Gender, Attendance, Lecture Session, Average Score and School Origin. Furthermore, data processing was carried out using the C4.5 algorithm. after data processing using the C4.5 algorithm, the results were tested using Rapid Minner. The results of the processing of the RapidMinner application found that the results of decisions obtained based on manual calculations and using Rapid Minner resulted in an accuracy of 92% against predictions [14]. From the previous research that has been explained, the use of the Theory of Planned Behavior approach has not been optimal in helping decision-makers. This is because the prediction accuracy level is not accurate, while the Naive Bayes algorithm compared to the C4.5 algorithm has a low accuracy level, and also does not describe the shape of a decision with a Decession Tree. With the use of the C4.5 algorithm, the data will be tested with a high degree of accuracy to predict the interest of prospective new students in choosing a university so that the university can make the best decision in attracting prospective new students.

II. Methods

In this study, the stages in the process of getting the best decision using the C4.5 algorithm based on the data that have been obtained are explained. The stages carried out can be seen in Figure 1.

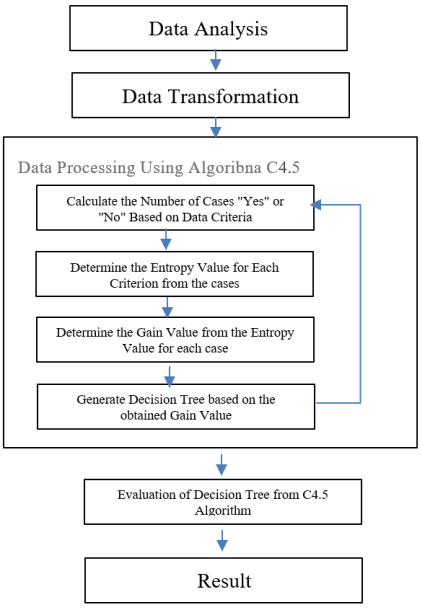

Figure 1. Research Stages

A. Data Needs Analysis

The data needed in this study is data in the form of survey results data on students and criteria data that are factors in determining prospective students interested in studying at the campus. There are 5 criteria and subcriteria that are the determining factors that can be seen in Table 1.

Table 1. Criteria for Determining Factors of Interest in Prospective Students Choosing a Campus

It	Criterion	Subkriteria		
1	Tuition	Cheap	Affordable	Expensive
2	Facilities	Excellent	Adequate	Enough
3	Accreditation	Excellent	Good	Enough
4	Service	Excellent	Good	Enough
5	Location	Near	Affordable	Far

The reason for choosing the criteria for Accreditation, Services and Facilities is seen in terms of student needs for higher education. In terms of tuition fees, because many families in Lampung have a lower middle monthly income and in terms of location, many students come from outside the Lampung area. Based on the criteria of the above determining factors, a survey was conducted on several students based on the criteria of these determining factors so that data on the interest survey of students who chose and did not choose to study at the campus were obtained. Furthermore, the data from the survey results was carried out Data Transformation for processing.

B. Data Transformation

Data Transformation is a change in data by changing the initial attribute value into an attribute value that is in accordance with the needs of the data in its processing [4]. Based on the results of the survey of prospective students' interest in choosing a campus against the criteria, then Data Transformation was carried out. For testing using the C4.5 algorithm, 25 sample data is used as a training dataset. The results of the survey Data Transformation can be seen in Table 2:

Table 2. Data Transformation Based on Survey Data

MHS	Accreditation	Tuition	Facilities	Service	Location	Interest
MHS 1	Good	Affordable	Excellent	Good	Affordable	***Yes***
MHS 2	Enough	Cheap	Enough	Enough	Affordable	***Yes***
MHS 3	Enough	Affordable	Enough	Enough	Far	***It***

MHS 4	Excellent	Cheap	Excellent	Excellent	Affordable	***Yes***
MHS 5	Good	Cheap	Adequate	Good	Far	***Yes***
MHS 6	Excellent	Affordable	Adequate	Good	Affordable	***Yes***
MHS 7	Enough	Affordable	Enough	Enough	Far	***It***
MHS 8	Good	Affordable	Enough	Good	Near	***Yes***
MHS 9	Enough	Expensive	Enough	Enough	Affordable	***It***
MHS 10	Enough	Affordable	Enough	Enough	Far	***It***
MHS 11	Good	Cheap	Excellent	Excellent	Near	***Yes***
MHS 12	Enough	Affordable	Excellent	Enough	Far	***Yes***
MHS 13	Enough	Affordable	Adequate	Good	Affordable	***Yes***
MHS 14	**Enough**	**Expensive**	**Adequate**	**Good**	**Near**	***It***

The data contained in the Data Transformation is sample data that will be used for testing using the C4.5 Algorithm.

C. C4.5 Algorithm Process

Based on the results of the Data Transformation that has been obtained, the C4.5 algorithm process is then carried out. The C4.5 algorithm has stages or steps in the problem-solving process so that it reaches the stage of results. The stages in the C4.5 Algorithm process can be seen as follows [13]:

1. Calculate the number of "Yes" and "No" cases based on Transformation data.
2. Determine the Entropy value of each criterion on a case-by-case basis. To calculate the Entropy value can use the formula:

$$\text{Entropy}(S) = \sum n_{i=1} - pi * log_{2pi} \quad (1)$$

Where:
S = case set
n = number of partitions *S*
pi = proportion of *Si* to *S*

3. Determine the highest Gain value based on the Entropy score that has been obtained. To calculate the Gain value you can use the formula:

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum n_{i=1} |Si/S| * Entropy(Si) \quad (2)$$

   a. Where:
   b. *S* = case set
   c. *A* = features
   d. n= number of partitions Attribute A

|Si| = the proportion of *Si* to *S*
|S|= number of cases in *S*

4. Generate a Decission Tree To generate the first node, go through processes 1 to 4 until there are no records in the empty Decission Tree branch.

III. Results and Discussions

A. Data Processing

Based on the algorithm process described above, then testing is carried out using Data Mining software, namely RapidMiner. For testing using the RapidMiner application, 592 data testing data was used. Testing is carried out by importing test data in excel format into Rapid Miner software as shown in Figure 2.

Row No.	Minat	C MHS	Akreditasi	Uang Kuliah	Fasilitas	Pelayanan	Lokasi
1	Yes	C MHS 1	Baik	Terjangkau	Sangat Baik	Baik	Terjangkau
2	Yes	C MHS 2	Cukup	Murah	Cukup	Cukup	Terjangkau
3	No	C MHS 3	Cukup	Terjangkau	Cukup	Cukup	Jauh
4	Yes	C MHS 4	Sangat Baik	Murah	Sangat Baik	Sangat Baik	Terjangkau
5	Yes	C MHS 5	Baik	Murah	Memadai	Baik	Jauh
6	Yes	C MHS 6	Sangat Baik	Terjangkau	Memadai	Baik	Terjangkau
7	No	C MHS 7	Cukup	Terjangkau	Cukup	Cukup	Jauh
8	Yes	C MHS 8	Baik	Terjangkau	Cukup	Baik	Dekat
9	No	C MHS 9	Cukup	Mahal	Cukup	Cukup	Terjangkau
10	No	C MHS 10	Cukup	Terjangkau	Cukup	Cukup	Jauh
11	Yes	C MHS 11	Baik	Murah	Sangat Baik	Sangat Baik	Dekat
12	Yes	C MHS 12	Cukup	Terjangkau	Sangat Baik	Cukup	Jauh
13	Yes	C MHS 13	Cukup	Terjangkau	Memadai	Baik	Terjangkau
14	No	C MHS 14	Cukup	Mahal	Memadai	Baik	Dekat
15	Yes	C MHS 15	Baik	Terjangkau	Memadai	Cukup	Terjangkau
16	No	C MHS 16	Baik	Mahal	Cukup	Baik	Terjangkau
17	Yes	C MHS 17	Baik	Terjangkau	Sangat Baik	Baik	Dekat
18	Yes	C MHS 18	Cukup	Murah	Cukup	Cukup	Terjangkau
19	No	C MHS 19	Cukup	Terjangkau	Cukup	Cukup	Jauh
20	No	C MHS 20	Cukup	Mahal	Cukup	Cukup	Terjangkau
21	Yes	C MHS 21	Sangat Baik	Murah	Sangat Baik	Baik	Terjangkau
22	Yes	C MHS 22	Baik	Terjangkau	Cukup	Sangat Baik	Jauh
23	Yes	C MHS 23	Baik	Terjangkau	Cukup	Cukup	Dekat
24	Yes	C MHS 24	Sangat Baik	Murah	Sangat Baik	Sangat Baik	Terjangkau

ExampleSet (592 examples,1 special attribute,6 regular attributes)

Figure 2. Dataset Snippets

Decision Tree Based Algorithm on the imported data, drag and drop the learning dataset table into the process view and set the decession tree operator which can be seen in Figure 3:

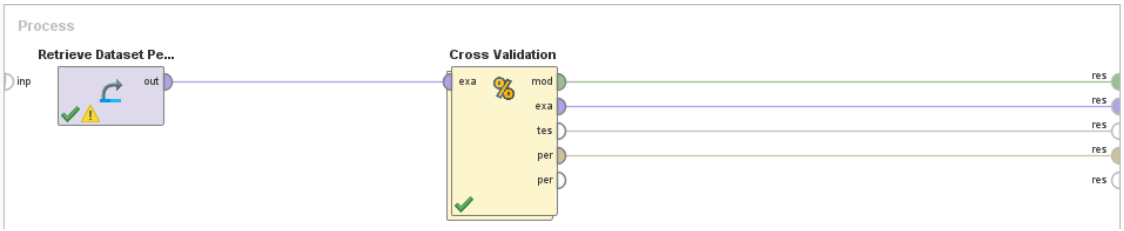

Figure 3. Process View in RapidMiner Software

Based on the process view that has been made, the acquisition of Entropy and Gain values from Node 4, then to obtain the Gain value of the entire category is 0, then for the Service and Location categories are not a factor in determining student interest in choosing a campus. The shape of the final Decession Tree can be seen in Figure 4:

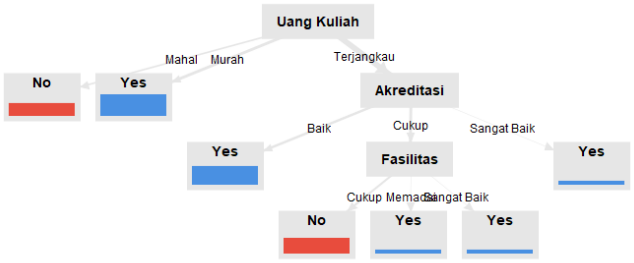

Figure 4. Decession Tree Test Results

From the results of the test using Rapidminer using the Decession Tree algorithm, it is explained that the criteria of Facility - Simply have a "No" decision for students' interest. So the Decession Tree above is used as the final decision in the prediction of determining the request for prospective students in choosing a university.

B. Evaluate Results

The results of the Decession Tree from the C4.5 algorithm are then evaluated on the results of the Decession Tree based on the Transformation data. The evaluation aims to see if there are errors in the results of the Decession Tree obtained and whether it is necessary to re-test. It is then translated into the form of decisions. Based on the translation of the decision, it can be a tool for a decision-maker to predict what criteria factors support the interest of prospective students in choosing the desired university. So that the prediction can be taken by the campus to take appropriate actions in dealing with this problem. The form of the decision generated through the Rapid Miner software can be seen in Figure 5:

Tree

```
Uang Kuliah = Mahal: No {Yes=0, No=95}
Uang Kuliah = Murah: Yes {Yes=165, No=0}
Uang Kuliah = Terjangkau
|   Akreditasi = Baik: Yes {Yes=142, No=0}
|   Akreditasi = Cukup
|   |   Fasilitas = Cukup: No {Yes=0, No=118}
|   |   Fasilitas = Memadai: Yes {Yes=24, No=0}
|   |   Fasilitas = Sangat Baik: Yes {Yes=24, No=0}
|   Akreditasi = Sangat Baik: Yes {Yes=24, No=0}
```

Figure 5. Test Results

Based on the tests that have been carried out, it can be concluded that the main factors that affect the interest of prospective new students in choosing a university using the C4.5 algorithm are as follows:

1. If the tuition fee is "Cheap", then the interest of prospective students is "YES" to choose a university and if the tuition fee is "Expensive", then the interest of prospective students is "NO" to choose a university.
2. If the tuition fee is "Affordable" and the campus accreditation is "Very Good" or "Good", then the interest of prospective students is "YES" to choose a college.
3. If the tuition fee is "Affordable" and the campus accreditation is "Sufficient" and the campus facilities are "Very Good" or "Adequate", then the interest of prospective students is "YES" to choose a university, but if the tuition fee is "Affordable" and the campus accreditation is "Sufficient" and the campus facilities are "Adequate", then the interest of prospective students is "NO" to choose a university.

IV. Conclusions

Based on the results of the research obtained, it is concluded that the data mining process using the C4.5 Algorithm has been able to produce a decision to predict the interest of prospective new students in choosing universities based on the criteria factors. Based on the results of data processing using the C4.5 Algorithm with 25 sample data, the criteria that are the main factors for the interest of prospective new students in choosing a university are the Tuition, Accreditation and Facilities criteria, where Cheap Tuition, Excellent or Good Accreditation and Very Good or Adequate Facilities attract the interest of prospective new students in choosing a university. Meanwhile, Expensive Tuition, Sufficient Accreditation and Sufficient Facilities do not attract the interest of prospective new students in choosing a university. Based on testing using Rapid Miner software with 592 data, the results of the Decession Tree and the decision between manual testing are the same, with an accuracy level of 100%. So that the use of the C4.5 algorithm has produced an assessment factor that can predict the interest of prospective students, and the university can predict the interest of prospective new students in the future.

References

[1] Alwarthan, S. A., Aslam, N., & Khan, I. U. (2022). Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review. In *Applied Computational Intelligence and Soft Computing* (Vol. 2022). Hindawi Limited.

[2] Berrar, D. (2019). Bayes' Theorem and Naive Bayes Classifier. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 403–412). Academic Press.

[3] Feng, L. (2021). Research on Higher Education Evaluation and Decision-Making Based on Data Mining. *Scientific Programming*, *2021*. https://doi.org/10.1155/2021/6195067

[4] Galit Shmueli, P. C. B. I. Y. N. R. P. K. C. L. Jr. (2018). *DATA MINING FOR BUSINESS ANALYTICS*.

[5] Gerhana, Y. A., Fallah, I., Zulfikar, W. B., Maylawati, D. S., & Ramdhani, M. A. (2019). Comparison of naive Bayes classifier and C4.5 algorithms in predicting student study period. *Journal of Physics: Conference Series*, *1280*(2).

[6] Hu, J., & Li, H. (2021). Composition and Optimization of Higher Education Management System Based on Data Mining Technology. *Scientific Programming*, *2021*.

[7] Masters, T. (2018). Data Mining Algorithms in C++. In *Data Mining Algorithms in C++*. Apress. https://doi.org/10.1007/978-1-4842-3315-3

[8] Naga, J. F., & Tinam-Isan, M. A. C. (2024). EXPLORING THE INFLUENCE OF PERSONALITY TRAITS ON STUDENTS' INFORMATION SECURITY RISK-TAKING BEHAVIORS: A BFI ASSESSMENT. *Procedia Computer Science*, *234*, 527–536. https://doi.org/10.1016/j.procs.2024.03.036

[9] Ngoc, P. V., Ngoc, C. V. T., Ngoc, T. V. T., & Duy, D. N. (2019). A C4.5 algorithm for english emotional classification. *Evolving Systems*, *10*(3), 425–451.

[10] Nurmalitasari, Awang Long, Z., & Faizuddin Mohd Noor, M. (2023). Factors Influencing Dropout Students in Higher Education. *Education Research International*, *2023*.

[11] Parteek Bhatia. (2019). *Data Mining and Data Warehousing*.

[12] Wang, C. (2021). Analysis of Students' Behavior in English Online Education Based on Data Mining. *Mobile Information Systems*, *2021*. https://doi.org/10.1155/2021/1856690

[13] Wang, J. (2022). Application of C4.5 Decision Tree Algorithm for Evaluating the College Music Education. *Mobile Information Systems*, *2022*. https://doi.org/10.1155/2022/7442352

[14] Yang, X., & Ge, J. (2022). Predicting Student Learning Effectiveness in Higher Education Based on Big Data Analysis. *Mobile Information Systems*, *2022*. https://doi.org/10.1155/2022/8409780

[15] Ye, H., & Li, C. (2022). Engineering Education Understanding Expert Decision System Research and Application. *Computational Intelligence and Neuroscience*, *2022*.