Institute of Research and Publication Indonesia
IJEERE: Indonesian Journal of Electrical Engineering and Renewable Energy
Journal Homepage: https://journal.irpi.or.id/index.php/ijeere
Vol. 3 Iss. 2 December 2023, pp: 96-100
P-ISSN: 2797-1155 E-ISSN: 2797-0868

The Implementation of Data Mining Techniques for Predicting Student Study Period Using the C4.5 Algorithm

Penerapan Teknik Data Mining Terhadap Prediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5

Anugrah Rizki Putra[1], Lidya Septiani[2], Welberkat Angga[3]

[1,2,3] Program Studi Sistem Informasi Fakultas Teknik Informatika Universitas Bina Sarana Informatika

Corresponden E-Mail: [1]anugrahrizkiputra670@gmail.com, [2]lidyaseptiani233@gmail.com, [3]welberkat18plp@gmail.com

Makalah: Diterima 9 November 2023; Diperbaiki 13 November 2023; Disetujui 17 December 2023
Corresponding Author: Anugrah Rizki Putra

Abstract

Higher education is a key element in human resource development, and in the information age, data generated by universities is becoming increasingly abundant, including data related to students' study duration. The duration of a student's study is a crucial indicator in evaluating the efficiency and effectiveness of higher education systems. This research presents the application of the C4.5 algorithm in the analysis of student study duration data using the RapidMiner software. The research findings indicate that the IPS1 attribute (Grade Point Average for Semester 1) is a determining factor in whether a student will graduate on time or be delayed. In the data analysis, if the IPS1 value exceeds 2.950, the student is considered "Graduated," while if it is less than or equal to 2.950, they are considered "Delayed." These results provide valuable insights for decision-makers in the field of higher education, demonstrating the potential of leveraging information technology and data mining to enhance the efficiency of the education system.

Keywords: *C4.5 Algorithm, Data Mining, RapidMiner*

Abstrak

Pendidikan tinggi adalah elemen kunci dalam pengembangan sumber daya manusia, dan dalam era informasi, data yang dihasilkan oleh perguruan tinggi semakin melimpah, termasuk data mengenai masa studi mahasiswa. Masa studi mahasiswa menjadi indikator penting dalam mengevaluasi efisiensi dan efektivitas sistem pendidikan tinggi. Penelitian ini memaparkan penerapan algoritma C4.5 pada analisis data masa studi mahasiswa dengan menggunakan perangkat lunak RapidMiner. Hasil penelitian ini menunjukkan bahwa atribut IPS1 (Indeks Prestasi Semester 1) merupakan faktor penentu dalam menentukan apakah seorang mahasiswa akan lulus atau terlambat. Dalam analisis data, jika nilai IPS1 melebihi 2.950, mahasiswa dianggap "Lulus," sedangkan jika kurang dari atau sama dengan 2.950, mereka dianggap "Terlambat." Hasil ini memberikan wawasan yang berharga bagi pengambil kebijakan di dunia pendidikan tinggi, menunjukkan potensi dalam memanfaatkan teknologi informasi dan data mining untuk meningkatkan efisiensi sistem pendidikan.

Kata kunci: Algoritma C4.5, Data Mining, RapidMiner

1. Introduction

Higher education is one of the most critical aspects of developing human resources. In this information era, data generated by universities is becoming increasingly abundant, including data related to students' study periods. The duration of students' study is a key indicator for evaluating the efficiency and effectiveness of the education system.

In efforts to enhance the efficiency and quality of education, information technology has played a crucial role, particularly in data analysis. Data mining techniques have become a relevant and beneficial approach to deal with the growing volume of data. Data mining algorithms, such as the C4.5 algorithm, can be used to analyze student study period data.

The use of the C4.5 algorithm in analyzing student study period data can be facilitated by the use of software like RapidMiner. RapidMiner is a data analysis platform that provides various tools for efficiently managing, processing, and analyzing data. In the context of applying data mining techniques for predicting students' study periods, RapidMiner can be an invaluable tool. Through RapidMiner, data on students, their age, GPA from the first to the eighth semester, can be imported and analyzed. The C4.5 algorithm can be applied to this data to identify patterns related to the duration of students' study. Moreover, RapidMiner also enables users to create prediction models that can provide more accurate estimates of students' study periods.

Therefore, this research is expected to provide valuable insights for decision-makers in higher education and those interested in data analysis to improve the education system by harnessing advanced technologies like RapidMiner.

2. Materials and Method

Data Mining

According to Mardi (2017), data mining is the process of discovering significant patterns and information in selected data through the utilization of various techniques, methods, or algorithms. The range of techniques and methods used in data mining is diverse, and the appropriate selection of methods or algorithms is tailored to the objectives and the entire Knowledge Discovery in Database (KDD) process.

C4.5 Algorithm

The C4.5 algorithm is classified under decision tree algorithms, representing an extension of the ID3 (Iterative Digital Calculator 3) algorithm developed by J. Ross Quinlan. This algorithm involves inputs in the form of training samples, training labels, and attributes. The C4.5 algorithm, as an enhancement of ID3, possesses additional capabilities to address issues such as missing data, continuous data, and pruning. When constructing a decision tree in the C4.5 algorithm, the attribute that will become the root of the tree is selected based on the highest gain value among the available attributes (Azwanti & Elisa, 2020). In calculating the gain value, the formula described below is used.

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

Information:
S: set of cases
A : attribute
n : total partition S
pi: proportion of Si to S

Meanwhile, the calculation of the entropy value can be seen from the following equation 2.

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

Information:
S: set of cases
A : attribute
n : total partition attribute A
|Si| : total cases in the i-th partition
|S| : total cases in S

Rapid Miner

RapidMiner is an open-source software that provides solutions for analyzing various fields, including data mining, text mining, and predictive analysis. RapidMiner utilizes various descriptive and predictive methods to provide information to its users, which can assist them in making informed decisions. RapidMiner is written in the Java programming language, allowing it to be used on multiple operating systems (Manullang et al., 2021).

3. Results and Discussion

The experiment's results

Test data refers to data that has been prepared for testing purposes. The results from this test data are used for classification using specific variables and attributes. This data is then divided into two groups: training data with 157 data entries and test data with 100 data entries. The next step involves using the C4.5 Algorithm to predict students' study periods.

In this study, the C4.5 Algorithm is used to estimate students' study periods, and a decision tree is built based on the calculation of Entropy and Gain. Once the tree is formed, the next step is to extract rules according to the decision tree. This study will discuss the steps of manual calculation and analysis and will use the Rapid Miner tool to support the analysis. The results of the calculations using the C4.5 Algorithm can be found in Table 1.

Table 1. Entropy and Gain Values

Nilai Entropy dan Gain

Simpul	Atribut	Nilai	Jumlah Kasus	Lulus	Terlambat	Entropy	Gain
Akar	Total		100	93	7	0,045479	
	IPS1						0,061589
		<3.00	16	14	2	0,07332	
		>3.00	84	84	0	0	
	IPS2						0,0865017
		<3.00	7	5	2	0,135467	
		>3.00	93	87	6	0,042454	
	IPS3						0,1259841
		<3.00	7	5	2	0,135467	
		>3.00	93	93	0	0	
	IPS4						0,0675651
		<3.00	7	6	1	0,081529	
		>3.00	93	92	1	0,008878	
	IPS5						0,1002125
		<3.00	5	4	1	0,105487	
		>3.00	95	95	0	0	
	IPS6						0,0783879
		<3.00	5	4	1	0,105487	
		>3.00	95	92	3	0,022973	
	IPS7						-0,021969
		<3.00	0	0	0	0	
		>3.00	100	97	3	0,021969	
	IPS8						-0,021969
		<3.00	0	0	0	0	
		>3.00	100	97	3	0,021969	

Table 1 explains that student data or test data has been analyzed using the C4.5 algorithm, and the results are grouped based on attributes that show the highest information gain, which, in this case, is the students' major. Therefore, it can be concluded that the attribute providing the highest information gain and used to build the decision tree is IPS1. IPS1 is a determining factor in identifying the graduation status or delay in this study.

Testing Data Using RapidMiner

After the data has undergone analysis and classification using the C4.5 Algorithm, the next step is to validate the manually calculated analysis results. To perform this step, the RapidMiner application is used, as shown in the following image. The RapidMiner application is used for testing the classification of student majors.

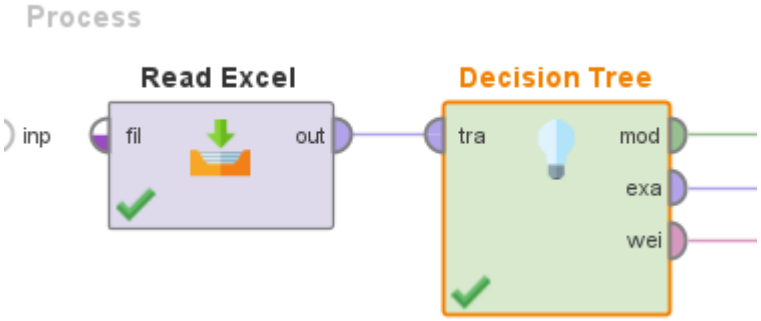

Figure 1. Decision tree relationship

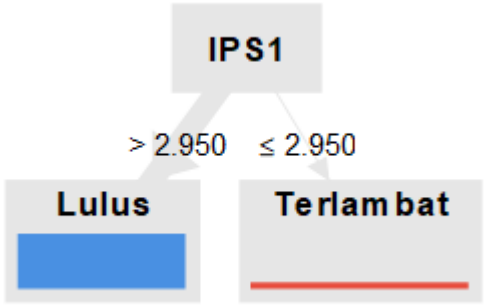

Figure 2. Display of decision tree results

Discussion

The statement is a rule or criterion related to the IPS1 (Semester 1 Grade Point Average) of an individual in the context of determining their status, whether they graduate or are delayed. There are two conditions explained in this statement. First, if someone's IPS1 is greater than 2.950, they will be declared "Graduated." In this situation, 93 individuals have achieved this score and are declared as graduates, with none being delayed. Second, if someone's IPS1 is less than or equal to 2.950, their status is "Delayed." In this condition, there are no individuals who have graduated, but there are 7 individuals who are declared as delayed. Therefore, it can be concluded that IPS1 is a determining factor in identifying the graduation or delay status in this study, and this statement provides a clear overview of these criteria.

4. Conclusion

This research reveals that the use of the C4.5 algorithm in analyzing students' academic duration data, with the assistance of the RapidMiner software, has successfully identified that IPS1 (Semester 1 Grade Point Average) is a determining factor in whether a student graduates or experiences a delay. These findings provide valuable insights for decision-makers in higher education, demonstrate the potential of utilizing information technology and data mining to enhance the efficiency of the education system, and stimulate interest in leveraging advanced technology for educational system improvement.

5. Acknowledgments

The author would like to express gratitude to all parties who have contributed to this research. Thanks to everyone for the assistance and resources provided. for their participation and support in this research.

References

Azwanti, N., & Elisa, E. (2020). Analisa Kepuasan Konsumen Menggunakan Algoritma C4.5. *Prosiding Seminar Nasional Ilmu Sosial Dan Teknologi*, *3*, 126–131.

Manullang, N., Sembiring, R. W., Gunawan, I., Parlina, I., & Irawan, I. (2021). Implementasi Teknik Data Mining untuk Prediksi Peminatan Jurusan Siswa Menggunakan Algoritma C4.5. *Jurnal Ilmu Komputer Dan Teknologi*, *2*(2), 1–5. https://doi.org/10.35960/ikomti.v2i2.700

Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. *Edik Informatika*, *2*(2), 213–219. https://doi.org/10.22202/ei.2016.v2i2.1465