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: 101-105
P-ISSN: 2797-1155 E-ISSN: 2797-0868

Application of Data Mining Techniques for Predicting Students' Selection of Science or Social Studies Major Using the C4.5 Algorithm

Penerapan Teknik Data Mining terhadap Prediksi Pemilihan Jurusan IPA/IPS Siswa Menggunakan Algoritma C4.5

Jean Rahmatika Jhody[1], Muhammad Ramadhan Tanjung[2], Snowhart Novolulu[3]

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

Corresponden E-Mail: [1]jeanrahmaa13@gmail.com, [2]19210348@bsi.ac.id, [3]gameeonnly@gmail.com

Makalah: Diterima 08 November 2023; Diperbaiki 13 November 2023; Disetujui 17 December 2023
Corresponding Author: Jean Rahmatika Jhody

Abstract

The selection of science (IPA) or social studies (IPS) majors is a crucial decision in a student's educational journey with significant implications for their academic and career development. This research discusses the implementation of data mining techniques using the C4.5 algorithm to predict the choice of IPA or IPS majors by students. The research results indicate that the most influential attribute in students' major selection is their National Examination (UN) scores. By employing the C4.5 algorithm, students with specific UN scores are categorized into either the IPA or IPS group. In data testing using RapidMiner, the analysis confirms the effectiveness of the C4.5 algorithm in predicting students' major choices. With the assistance of data mining technology, the data analysis process becomes more efficient, providing valuable insights for educational decision-making. This research makes a significant contribution to the development of more effective and efficient educational recommendation systems.

Keyword: Data mining, Prediction, RapidMiner, C4.5 Algorithm

Abstrak

Pemilihan jurusan IPA atau IPS adalah keputusan penting dalam perjalanan pendidikan siswa yang berdampak signifikan pada perkembangan akademik dan karir mereka. Penelitian ini membahas penerapan teknik data mining mempergunakan algoritma C4.5 dalam memperkirakan pemilihan jurusan IPA atau IPS oleh siswa. Hasil penelitian menunjukkan bahwa atribut yang paling memengaruhi pemilihan jurusan siswa adalah nilai UN. Dengan menggunakan algoritma C4.5, siswa dengan nilai UN tertentu ditempatkan dalam kelompok IPA atau IPS. Dalam pengujian data menggunakan RapidMiner, hasil analisis mengkonfirmasi bahwa algoritma C4.5 efektif dalam memprediksi pemilihan jurusan siswa. Dengan bantuan teknologi data mining, proses analisis data menjadi lebih efisien dan memberikan wawasan berharga dalam pengambilan keputusan pendidikan. Penelitian ini memberikan kontribusi penting dalam pengembangan sistem rekomendasi pendidikan yang lebih baik serta efisien.

Kata kunci: Data mining, Prediksi, RapidMiner, Algoritma C4.5

1. Introduction

The Choice of Science (IPA) or Social Studies (IPS) major is one of the critical decisions in a student's educational journey, which can significantly impact their academic development and future career. In the digital era and with the advancement of information technology, data mining has become an effective tool for analyzing student data and aiding the decision-making process. One of the algorithms proven to be effective in applying data mining techniques to predict students' major selection is the C4.5 algorithm.

In this study, the author discusses the application of data mining methods using the C4.5 algorithm to facilitate the prediction of whether students will choose a major in science (IPA) or social studies (IPS). The author will explain the basic concept of data mining, the C4.5 algorithm, and how they are used to analyze student data

DOI: https://doi.org/10.57152/ijeere.v3i1

and provide accurate recommendations. Furthermore, the researcher will explore the benefits of this study in improving our understanding of the factors influencing students' major selection decisions and the potential development of better recommendation systems in the field of education.

2. Materials and Method

Data Mining

According to Hermawati, as cited in Sikumbang (2018), data mining is a process that involves the use of one or more machine learning methods that automatically analyze and extract knowledge. Data mining is an iterative and interactive process aimed at obtaining new models that have significant value and can be understood in the context of massive databases. The goal of data mining is to search for relevant trends or patterns in these large databases to facilitate decision-making in future situations. These patterns are identified through specialized tools that provide meaningful data analysis and insights, which can serve as a basis for further research.

C4.5 Algorithm

The C4.5 algorithm is a method used in constructing decision trees. A decision tree is a well-known and highly effective classification and prediction technique. The decision tree method transforms complex information into decision trees that represent rules. These rules can be easily understood in natural language and can be expressed in a database language, such as Structured Query Language (SQL), to search for data in specific categories. The C4.5 algorithm follows these steps to build a decision tree: (1) Selection of attributes to be nodes; (3) Division of data cases into these branches; (4) Repeating this process for each branch until all cases in that branch have the same class. After preparing the training data, the next step is to determine the root of the decision tree. The selection of the root is done by calculating the gain value for each attribute, and the attribute with the highest gain becomes the root of the tree (Eska, 2016). Before calculating gain, it is necessary to calculate the entropy value first. In calculating gain, the formula explained in the following formula is used.

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

Keterangan:
S : himpunan kasus
A : atribut
n : jumlah partisi S
pi : proporsi dari Si terhadap S

Next, determine the entropy value by using the following formula:

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

Information:
S: set of cases
A : attribute
n : number of attribute A partitions
|Si| : number of cases in the ith partition
|S| : number of cases in S

RapidMiner

RapidMiner is a software developed by Dr. Markus Hofmann from the Institute of Technology Blanchardstown and Ralf Klinkenberg from rapid-i.com. This software features a graphical user interface (GUI) that simplifies its application. RapidMiner is open-source software created using the Java programming language and is licensed under the GNU Public License. It can be operated on various operating systems. Using RapidMiner does not require specific programming skills because all the necessary facilities are provided.

RapidMiner is specifically designed for data analysis and data mining. The software offers a complete range of models and algorithms, such as Bayesian Models, Tree Induction, Neural Network, and more. RapidMiner also provides various methods, including classification, clustering, association, and others. If users cannot find a specific model or algorithm in RapidMiner, they can add additional modules because this software is open-source, allowing anyone to contribute to its development (Siska et al., 2020).

3. Results and Discussion

Experiment Results

The test data refers to the data prepared in advance for testing purposes, which will be used in the classification process using specific variables and attributes. This data is then divided into two groups, consisting of training data with 164 data entries and testing data with 100 data entries. The next step involves the application of the C4.5 Algorithm to predict students' choice of the science (IPA) or social science (IPS) stream.

In this study, where the C4.5 Algorithm is used to predict students' choice of the science (IPA) or social science (IPS) stream, a decision tree is constructed based on the calculation of Entropy and Gain. Once the decision tree is formed, the next step is to extract rules according to the branches of the decision tree. In this study, the authors will explain the steps of manual calculation and analysis and will also use the RapidMiner tool to support the analysis. The results of the calculations using the C4.5 Algorithm can be found in Table 1.

Table 1. Results of the C4.5 Algorithm Calculation

Entropy and Gain Value

Simpul	Atribut	Nilai	Jumlah Kasus	IPA	IPS	Entropy	Gain
Akar	Total		100	62	38	0,160654	
	Usia						0,044664098
		Usia 16	33	15	18	0,184364	
		Usia 17	34	31	3	0,055184	
		Usia 18	33	16	17	0,182113	
	Nilai UN						0,006425887
		70-75	31	16	15	0,178764	
		80-85	35	16	19	0,18421	
		90-95	34	22	12	0,154257	
	Nilai IPA						-0,109272639
		65-70	6	0	6	0	
		75-80	45	32	13	0,136447	
		85-90	36	26	10	0,132977	
		95	12	12	0	0	
	Nilai IPS					0	0,011228177
		70-75	31	16	15	0,178764	
		80-85	44	28	16	0,156866	
		90-95	25	14	11	0,172391	

Table 1 illustrates the results of analyzing student data or testing data using the C4.5 algorithm, and the result groups are created based on the attribute that produces the highest information gain, which in this situation is the students' major. Therefore, it can be concluded that the attribute that provides the highest information gain and is used as the basis for building the decision tree is the students' major. From this analysis, it can be inferred that the major preferred by students is science (IPA) with a National Exam score between 80-85.

Testing Data Using Rapid Miner

After the data has undergone analysis and has been grouped using the C4.5 Algorithm, the next step is to confirm the results of the manual calculation analysis. To perform this step, the RapidMiner application is used, as seen in the illustration in Figure 1 below. The RapidMiner application is used in the testing process of classifying students' majors.

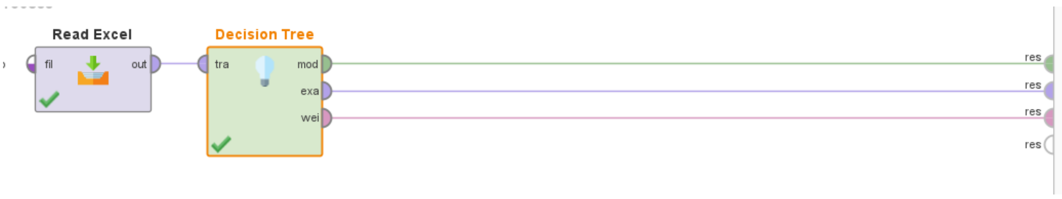

Figure 1. Decision tree relationship

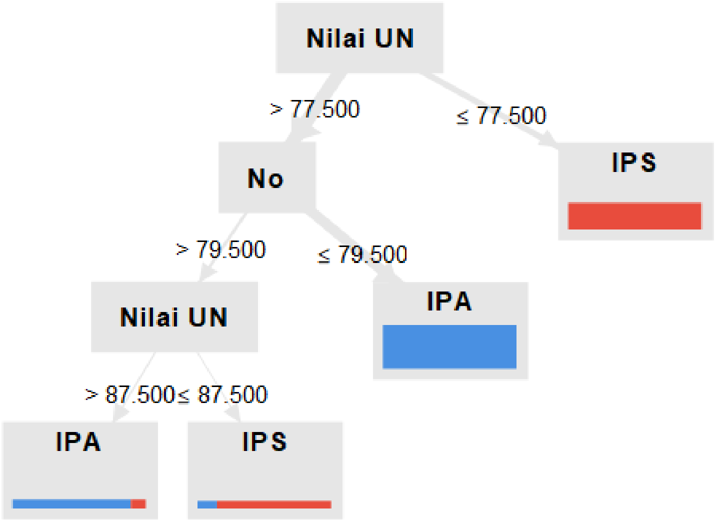

Figure 2. Display of decision tree results

Discussion

In the given sequence of conditions, the separation of students is based on their National Exam (UN) scores. First, students with UN scores greater than 77.500 are examined in the following conditions. Next, if someone's UN score is greater than 79.500, the next rule is checked. If the UN score is also greater than 87.500, students are classified into the science (IPA) group. In this situation, there is information about the number of students entering the IPA and IPS groups, which is IPA=8 and IPS=1. However, if their UN score is less than or equal to 87.500, students are classified into the IPS group, with details of the number of students as IPA=1 and IPS=6. If someone's UN score is less than 79.500, they are classified into the IPA group, and the information states that 53 students enter the IPA group, and 0 students enter the IPS group in this condition. Finally, if someone's UN score is less than or equal to 77.500, they are classified into the IPS group, and the information indicates that no students enter the IPA group, while there are 31 students entering the IPS group in this condition. So, this system categorizes students based on their UN scores, directing them to the IPA or IPS group according to specific score ranges.

4. Conclusion

The choice between science (IPA) or social studies (IPS) majors is a crucial decision in a student's education, and it can significantly impact their academic development and future career. In the digital era with the advancement of information technology, data mining has proven to be an effective tool for analyzing student data and aiding in decision-making processes. In this research, the C4.5 algorithm was used to predict students' major choices based on their data.

Data mining aims to discover new patterns or models with significant value within large databases, which can later be used for recommendations. The analysis results indicate that the attribute that most influences students' major choices is their National Exam (UN) score. Students with specific UN scores will be placed into either the IPA or IPS groups.

This study provides a deeper understanding of the factors affecting students' major choices and the potential for developing better recommendation systems in the field of education. With the assistance of the C4.5 algorithm and software like RapidMiner, the data analysis process becomes more efficient and can provide valuable insights in the field of education.

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

Eska, J. (2016). Penerapan Data Mining Untuk Prekdiksi Penjualan Wallpaper Menggunakan Algoritma C4.5 STMIK Royal Ksiaran. *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi)*, *2*, 9–13.

Sikumbang, E. D. (2018). Penerapan Data Mining Dengan Algoritma Apriori. *Jurnal Teknik Komputer AMIK BSI (JTK)*, *9986*(September), 1–4.

Siska, H., Aji, S., & Eko, S. (2020). Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu). *Jurnal Media Infotama*, *11*(2), 130–138.