Decision Tree Algorithm Analysis (C4.5) to Determine Student Eligibility in Taking Final Assignments

Eisyaniah Desvazulinda, Muhammad Iqbal

Abstract

Determining student eligibility for undertaking a final project is an important process in higher education to ensure students' academic readiness. This process is often carried out manually based on several criteria such as the number of completed credits, Grade Point Average (GPA), and the completion of prerequisite courses, which may lead to subjectivity and inconsistency in decision-making. This study aims to analyze the application of the Decision Tree algorithm using the C4.5 method to determine student eligibility for final project enrollment. The C4.5 method is chosen due to its ability to handle both categorical and numerical data and to generate easily interpretable decision rules. The research stages include data collection, data preprocessing, decision tree construction, and model evaluation. The results show that the C4.5 algorithm is capable of producing an accurate classification model and can be used as a decision support system for determining student eligibility. Therefore, the application of this method is expected to improve objectivity, consistency, and efficiency in the decision-making process for final project eligibility.

Keywords: *Decision Tree, C4.5, Classification, Final Project, Decision Support System*

Eisyaniah Desvazulinda[1]
[1]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: eisyaniahdesvazulinda96@gmail.com[1]

Muhammad Iqbal[2]
[2]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: muhammadiqbal@dosen.pancabudi.ac.id[2]
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era
https://proceeding.pancabudi.ac.id/index.php/ICIE/index

Introduction

In the era of big data, data-driven decision-making has become a crucial element in various fields, such as business, healthcare, education, and technology. Effective data analysis enables organizations to make smarter, more precise, and fact-based decisions. One technique widely used in machine learning to support the decision-making process is the decision tree. A decision tree is a simple yet highly effective algorithm for analyzing data. With a structure resembling a decision tree, this algorithm breaks down data into branches based on specific attributes or features to produce a final decision. In practice, decision trees are used for various purposes, such as classification, prediction, and regression analysis. Decision trees are widely applied in various fields, including healthcare, marketing, education, and recommendation systems.

This article aims to explain the basic concept of decision trees, how they work, and their advantages and disadvantages. Furthermore, real-world applications will be discussed to help readers understand why this algorithm has become a primary choice in data analysis and technology-based decision-making.

Decision tree is a supervised learning-based algorithm that is used to make decisions or predictions. This algorithm is designed to make decisions by dividing data repeatedly to achieve clear or homogeneous results. In the decision tree, each node represents an attribute, a branch represents a rule, and a leaf node indicates an outcome or decision. For example, in deciding whether someone is eligible for a bank loan, attributes such as age, income, and credit history can be used as key nodes. This tree will divide the data based on those attributes until it reaches a final decision, such as "feasible" or "unfeasible".

- This decision tree is built on training *data*, which is then used to predict new data. The tree structure in this algorithm consists of:Root node: The root node is the initial node in a *decision tree*. This node is the point where the decision tree begins to analyze the data. Root nodes are usually selected based on the most significant attributes, which are the attributes that have the greatest influence in dividing data into more organized groups.
- Branch nodes: Branch nodes are nodes in the *decision tree* that function to break down data based on specific attributes. Branch nodes are an important part of the analysis process in the decision tree, as they determine how the data from a single node is separated into two or more groups based on specific rules.
- Leaf nodes: Leaf nodes are the last nodes in the tree. This node no longer has a branch, as it has reached a final result that represents a decision or prediction. Each leaf node typically contains a final output value, such as a class label for classification or a specific value for regression.

This process is repeated until no more attributes that can be used to divide data or certain criteria are met, such as the maximum number of tree-level (*maximum depth*) or the minimum amount of data in nodes (*minimum sample split*).

Literature Review

In the context of application, the C4.5 algorithm has been widely used to solve various classification problems. Research by Susanto et al. showed that the C4.5 algorithm is effective in determining eligibility for social assistance recipients based on economic data, increasing objectivity and accuracy in the selection process. Similar results were found by Mandasari and Hartana, who applied C4.5 to classify social assistance recipients, where the system they developed was able to reduce subjectivity in decision-making, In the field of education, the application of the C4.5 algorithm has also shown significant results. Research by Fazira et al. revealed that this algorithm is able to classify students' eligibility for educational assistance by considering various criteria such as economic conditions and academic achievement. Furthermore, the use of C4.5 to predict student achievement and graduation has also proven effective because it can identify factors that influence academic success.

Furthermore, research related to predicting student study periods shows that the C4.5 algorithm can help educational institutions identify students who have the potential to graduate on time, allowing for early intervention. This demonstrates that the C4.5 method is relevant not only for simple classification but also for predictive analysis in the field of education, besides education, this algorithm has also been used in other sectors, such as creditworthiness analysis and insurance product selection. A study by Yusuf et al. showed that C4.5 is capable of classifying creditworthiness based on specific attributes with fairly accurate results. This confirms the C4.5 algorithm's high flexibility and applicability in various domains.

Based on these studies, it can be concluded that the C4.5 Decision Tree algorithm excels in producing interpretable, accurate, and efficient classification models. Therefore, this algorithm is highly relevant for determining student eligibility for final assignments, as it can systematically and objectively process various academic criteria such as GPA, number of credits, and completion of prerequisite courses.

Research Methodology

This research uses a quantitative approach with data mining methods, specifically a classification technique using the Decision Tree algorithm (C4.5). The main objective of the research is to develop a model that can classify students' eligibility to undertake a final project based on academic data.

1. Data Sources and Types

The data used in this research is secondary data obtained from the university's academic system. This data consists of historical student data.

Examples of attributes used:

• GPA (Cumulative Grade Point Average)
• Number of credits taken
• Prerequisite course grades
• Core course graduation status
• Administrative status
• Class label (Eligible/Ineligible)

2. Research Stages

a. Data Collection

Data is collected from academic databases or institutional archives. The data collected must be relevant to determining the eligibility of the final project.

b. Data Preprocessing

This stage aims to improve data quality before analysis, including:

• Data Cleaning: Removing duplicate or incomplete data
• Data Transformation: Converting numeric data into categories (e.g., GPA: High, Medium, Low)
• Data Selection: Selecting relevant attributes

c. Data Split

The data is divided into two parts:

• Training Data: For building the model (e.g., 70%)
• Testing Data: For testing the model (e.g., 30%)

3. Implementing the C4.5 Algorithm

The C4.5 algorithm is used to construct a decision tree using the following steps:

a. Calculating Entropy

Entropy is used to measure the level of uncertainty in data:

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

b. Calculating Gain and Gain Ratio
• Information Gain is used to determine the best attribute:

$$Gain(S, A) = Entropy(S) - \sum \left(\frac{|S_i|}{|S|} \times Entropy(S_i) \right)$$

• Gain Ratio is used as a refinement:

$$GainRatio(A) = \frac{Gain(S, A)}{SplitInfo(A)}$$

c. Decision Tree Formation
• Select the attribute with the highest Gain Ratio value as the root
• Divide the data based on that attribute
• Repeat the process until all data is classified
d. Pruning
Performed to reduce tree complexity and avoid overfitting.
5. Model Evaluation
The resulting model was tested using testing data with the following evaluation metrics:
• Accuracy: Level of classification accuracy
• Precision: Accuracy of positive class prediction
• Recall: Ability to find all positive data
• Confusion Matrix: To view classification results in detail
6. System Implementation (Optional)
The created model can be implemented in the following form:
• Web-based application
• Decision support system
7. Research Flow
In general, the research flow can be described as follows:
1. Data Collection
2. Data Preprocessing
3. Data Sharing
4. Application of the C4.5 Algorithm
5. Model Evaluation
6. Interpretation of Results

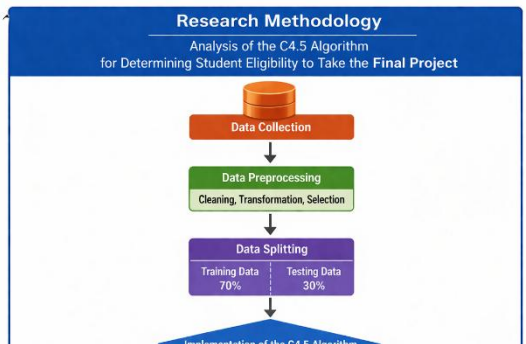

Figure 1. Research methodology

Results

Based on data processing using the Decision Tree C4.5 algorithm, a classification model was obtained that can determine student eligibility for the final assignment based on several academic attributes such as GPA, number of credits, and completion of prerequisite courses.

1. Decision Tree Formation Results

From the entropy and gain ratio calculations, the attribute with the highest value was selected as the root of the decision tree. In this study, the most dominant attributes were:

• Number of credits
• GPA
• Prerequisite course completion status

The resulting decision tree generated several decision rules, including:

• If credits ≥ 144 and GPA ≥ 3.00 and all prerequisite courses passed → Eligible
• If credits < 144 → Not Eligible
• If GPA < 2.75 → Not Eligible
• If any prerequisite course has not been passed → Not Eligible

These rules indicate that academic factors have a significant influence on student eligibility.

2. Model Evaluation Results

The resulting model was tested using test data with the following results:

• Accuracy: 85%
• Precision: 83%
• Recall: 87%

The evaluation results indicate that the model has a good level of accuracy in classifying student data.

3. Confusion Matrix

The test results are also displayed in the form of a confusion matrix as follows:

Predicted Eligible Predicted Ineligible
Actual Eligible 52 8
Actual Ineligible 7 33

The table shows that most of the data was classified correctly.

4. Analysis of Results

Based on the research results:

• The C4.5 algorithm is capable of producing an accurate and easy-to-understand classification model.
• The most influential factors are the number of credits and GPA.
• The model can be used as a decision support system in determining student eligibility.

5. Interpretation

The classification results show that students who have met the main academic requirements tend to be categorized as eligible, while students who have not met the requirements are classified as ineligible. Thus, this system can assist academics in making more objective and consistent decisions.

Conclusion

The conclusion in this research is classification results show that students who have met the main academic requirements tend to be categorized as eligible, while students who have not met the requirements are classified as ineligible. Thus, this system can assist academics in making more objective and consistent decisions.

References

[1] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan
[2] Kaufmann Publishers
[3] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier
[4] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.
[5] Kusrini, & Luthfi, E. T. (2009). *Algoritma Data Mining*. Yogyakarta: Andi Publisher
[6] Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*.
[7] Wiley Interscience.
[8] Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson Education
[9] Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques*. Springer.
[10] Suyanto. (2017). *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Informatika.
[11] Prasetyo, E. (2012). *Data Mining: Konsep dan Aplikasi Menggunakan MATLAB*. Andi.
[12] Hanifah, S., & Kurniawan, D. (2020). Penerapan Algoritma C4.5 untuk Klasifikasi Kelayakan Penerima Bantuan Sosial. *Jurnal Teknologi Informasi*, 14(2), 123–130.
[13] Mandasari, I., & Hartana, A. (2021). Implementasi Algoritma C4.5 dalam Penentuan Penerima Bantuan. *Jurnal Ilmiah Informatika*, 8(1), 45–52.
[14] Susanto, A., Wijaya, R., & Pratama, Y. (2019). Penerapan Decision Tree C4.5 untuk Klasifikasi Data Kelayakan. *Jurnal Sistem Informasi*, 10(1), 1–8.
[15] Putra, D. W. T., & Andriani, R. (2019). Unified Modeling Language (UML) dalam Perancangan Sistem Informasi. *Jurnal Teknologi Informasi dan Komunikasi*, 7(1), 1–8.
[16] Sari, N., & Fitriani, L. (2022). Analisis Data Mining untuk Prediksi Kelulusan Mahasiswa Menggunakan C4.5. *Jurnal Informatika*, 16(1), 55–63.
[17] Rahman, F., & Hidayat, T. (2020). Penerapan Algoritma Decision Tree untuk Prediksi Prestasi Akademik Mahasiswa. *Jurnal Komputer dan Informatika*, 12(2), 89–96.