Data Mining Analysis for Assessing Students' Proficiency in Scientific Writing

FAHRUDDIN[1*], REGITA CAHYA SAPHIRA[2], AND GUSMELIA TESTIANA[3]

Abstract

A good understanding of the material and clear writing are important for success in academic and professional careers. However, not all students are equally skilled at writing scientific articles. This research aims to classify the levels of student understanding in writing scientific articles. This study classifies college students' understanding of scientific writing across four universities in Palembang with a sample of 108 students selected through random sampling. Data were collected via questionnaires, and the quantitative method used data mining with the C4.5 algorithm. Testing with RapidMiner software yielded a model accuracy of 74.58%. The study found that the C4.5 algorithm's accuracy in classifying students' understanding of scientific writing falls into the Fair category, meaning the model treats all individuals or groups equally. The findings of this research should be a particular concern for higher education institutions to support and assist students in better understanding how to write scientific articles.

Keywords
Algorithm C4.5, data mining, scientific articles

Article History
Received June 30, 2024
Accepted December 16, 2024

How to Cite
Fahruddin, Saphira, R. C., & Testiana, G. (2024). Data mining analysis for assessing students' proficiency in scientific writing. *Indonesian Research Journal in Education |IRJE|*, *8*(2),475 – 489. https://doi.org/10.22437/irje.v8i2.35403

[1]*Universitas Negeri *Raden Fatah,* Palembang, Indonesia, Corresponding author: fahruddin@radenfatah.ac.id
[2,3] Universitas Negeri *Raden Fatah,* Palembang, Indonesia

Introduction

Scientific writing plays a crucial role in student learning in higher education environments. When students write scientific papers, they not only propose arguments, but also have to present relevant literature, evaluate bulk, and draw conclusions based on sound analysis. According to Latifah (2024), students often face various challenges in writing scientific articles, making this activity one of the most difficult academic tasks. One of the main difficulties faced by students is the lack of understanding of academic writing structure and strict scientific conventions. Then, Sweller (1988) states that learning occurs most effectively under conditions that align with human cognitive architecture. An excessive cognitive load can overwhelm working memory, making it difficult to learn complex tasks, such as writing. The implication is that students experiencing high cognitive load may feel overwhelmed and unable to focus on writing. It is relevant to a research study conducted by Arsyad and Adila (2018) that many students do not have adequate skills in constructing coherent and logical arguments and face difficulties in aligning their ideas with the expected writing structure in scientific articles. This study shows that a lack of understanding of the basic structure of scientific writing, such as the introduction, literature review, methodology, results, and discussion, hinders their ability to write effectively.

Hasugian (2008) explains that mastering information literacy skills is an important component in the development of students' academic competence. However, several studies have identified significant deficits in students' ability to conduct comprehensive and critical literature research, as well as in managing and integrating relevant information sources into their scientific writing. Research conducted by Karim et al. (2020) shows that students often lack skills in searching for and managing relevant information sources. Students often find it difficult to conduct comprehensive literature research and critically integrate these findings into their writing. This lack of skills not only affects their ability to effectively support arguments but also increases the risk of plagiarism due to the inability to cite sources correctly. Psychological factors, such as writing anxiety, also pose significant obstacles to scientific writing. Then, it is also supported by recent research by Zhang and Yin (2021) that many students experience high anxiety when writing, which hinders their ability to participate effectively in the critical and analytical writing process. This research found that previous negative experiences, such as overly critical feedback or poor evaluations, can reduce students' motivation and confidence in writing.

The advancement of information and communication technology plays a crucial role in improving students' comprehension of writing scientific papers. This research is currently being conducted, aiming for students to produce publishable scientific papers. The application of classification-based data mining can be an innovative solution to analyze and understand students' comprehension levels. By using this technique, information about students' writing skills and comprehension can be identified and analyzed more efficiently. A research study conducted by Romero and Ventura (2013) mentions that classification is one of the most frequently used data mining techniques in educational data analysis. This technique can help in predicting students' levels of understanding based on their historical data, such as exam

scores, attendance, and class participation. By using algorithms such as Decision Tree, Random Forest, or Naive Bayes, institutions can create predictive models to identify students who need further assistance in writing scientific articles.

Data mining research involves a series of analytical processes aimed at exploring and uncovering patterns in data. The implementation involves the use of statistical techniques, mathematics, and artificial intelligence to identify significant occurrences and trends from large and complex data volume. Data mining can be used to explore patterns in large and complex educational data. Applying data mining techniques, such as classification, can help identify students' levels of understanding in specific subjects, including their ability to write scientific articles. By implementing classification algorithms, educational institutions can predict and categorize students based on their understanding levels, allowing for more targeted educational interventions. The main function of data mining is to transform raw data into information that can be used for better decision-making. In-depth data analysis helps in understanding market behavior, identifying consulting patterns, detecting anomalies, and more. By understanding data mining more thoroughly, organizations can improve operational efficiency, implement marketing strategies, and gain valuable insights to optimize business processes.

This research aims to classify the level of student understanding in writing scientific articles using the C4.5 algorithm. The C4.5 algorithm is one of the most widely used machine learning algorithms in data mining and predictive modeling. The C4.5 algorithm has several advantages in classifying students' understanding in writing scientific papers. Research study by Abidin and Ramadhani (2020) shows that the C4.5 algorithm can generate decision trees that are easy to understand and interpret, which helps in identifying factors affecting students' understanding in writing scientific papers. They mention that "the C4.5 algorithm is capable of handling data with categorical and numerical attributes and dealing with missing values, which are often found in academic data". The main advantage of the C4.5 algorithm is its ability to produce decision trees that are easy to understand and interpret, which is valuable in modeling and data analysis. Additionally, the C4.5 algorithm has the capability to perform pruning, a process to remove insignificant branches in the decision tree. Pruning helps prevent overfitting, which is when a model is too closely fitted to training data and performs poorly on test or new data. Thus, the C4.5 algorithm has become one of the most widely used algorithms in various data mining and predictive analysis applications.

In data analysis, the C4.5 algorithm plays an important role in developing effective classification models. However, its implementation becomes more efficient and structured with the help of software tools like RapidMiner. With RapidMiner, the implementation process of the C4.5 algorithm becomes more straightforward and effective. This software provides an intuitive visual environment that facilitates the steps of algorithm implementation, from data processing to model evaluation. Users can easily import data, adjust algorithm parameters, and evaluate model performance within an integrated platform. Thus, the integration of the C4.5 algorithm and RapidMiner provides an efficient and innovative solution for data analysts facing complex data analysis challenges.

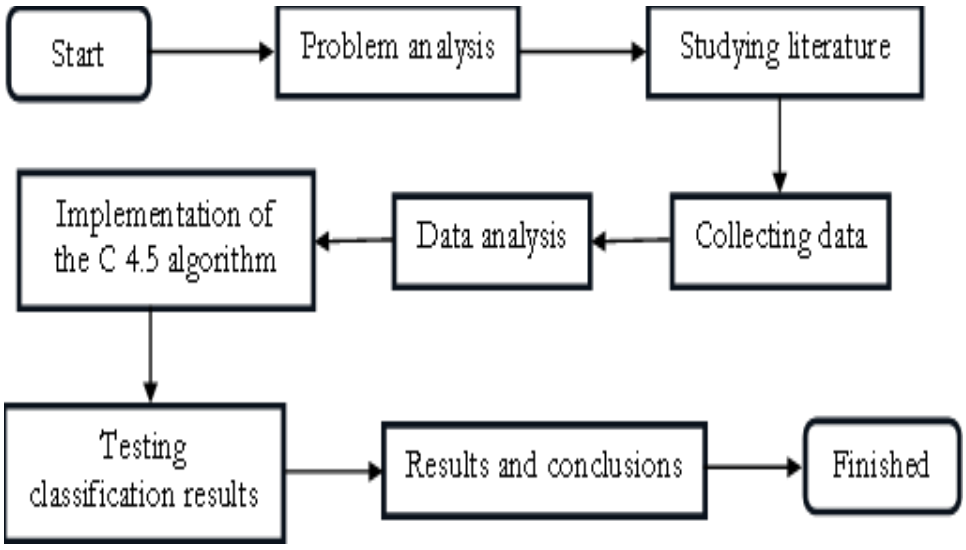
Methodology

In this study a quantitative model was applied utilizing data mining techniques and the C4.5 algorithm. The aim of using this quantitative research method is to identify and address the research problem. Additionally, descriptive methods are employed to gain a thorough understanding of the independent variables in question.

Research methodology stages

This research stage describes the methodology and research framework used in solving research problems. The following are the stages of research methodology shown in the figure below.

Figure 1. *The stages of research methodology*



Based on Figure 1 above, it can be explained the process of research methodology stages which ultimately leads to the classification of student understanding in writing scientific articles:

Problem analysis

The problem analysis stage is the first step in the research process. It involves identifying, understanding, and explaining the problem to be researched. This is accomplished by formulating clear research questions, defining specific research objectives, and identifying the context and relevance of the problem through a review of the existing literature. At this stage, it is crucial to ensure that the problem is well-understood and clearly stated.

Literature review

The literature review is an essential step that involves reviewing and analyzing previous studies and relevant publications related to the research topic. The purpose of this review is to develop an in-depth understanding of the research framework, explore the research methods that have been used in similar studies, and identify gaps in the literature that may offer opportunities for further exploration. This step ensures that the study builds on existing knowledge and identifies novel contributions.

Data collecting

Data collection is a systematic process aimed at gathering the necessary information to address the research questions. Data can be collected through various methods, such as surveys, observations, or other data-gathering techniques. It is important that the data collected are relevant, valid, and reliable in relation to the research objectives. Ensuring the accuracy and quality of the data is key to obtaining meaningful results.

Data analysis

Data analysis involves processing, exploring, and interpreting the collected data to identify patterns, trends, and insights that answer the research questions. The analysis methods may vary depending on the type of data and the specific questions being addressed. This stage is crucial in determining whether the hypotheses are supported by the data or need to be rejected. Statistical techniques, machine learning, and other methods may be employed at this stage.

Implementation of C4.5 Algorithm

The implementation of the C4.5 algorithm is a critical step in the development of a classification model using data mining techniques. The C4.5 algorithm is a machine learning algorithm designed to generate decision trees from training data. This step involves learning patterns and rules from the training data that will later be used to classify new data into different categories. The algorithm identifies key attributes that influence the classification and creates a decision tree based on these attributes.

Testing classification results

Examining classification results is an important step in evaluating the performance of the classification model that has been developed. This process involves the development of test data that is separate from the training data, including the calculation of accuracy, precision, precision, and other evaluation metrics of the classification model. The goal is to assess how well the model can predict new classes of data with high accuracy. This stage involves analyzing and interpreting the research findings and drawing conclusions based on the results obtained. The findings are then presented in a systematic and scientific format, often as part of a report or scientific paper. The report includes an introduction, methods, results, discussion, and

conclusions, providing a comprehensive overview of the research process and outcomes. In this research, the data utilized were gathered from lectures conducted for students in Palembang, South Sumatra. The study involved a total of 108 respondents, who provided valuable insights into their experiences and comprehension in writing scientific articles. The collection of this data was a crucial step, as it serves as the foundation for our analysis. The data collection process consisted of several systematic steps:

Sampling method

The research team employed a random sampling method to select participants. This method ensures that every individual in the target population had an equal chance of being included in the study. By randomly selecting students, we aimed to minimize bias and ensure that the sample was representative of the larger population of students in Palembang.

Questionnaire design

A structured questionnaire was developed to collect data. The questionnaire comprised multiple-choice and open-ended questions designed to assess various attributes related to the students' understanding of scientific writing. These attributes included: (1) Source of Reference (C1): Evaluating whether students used journals, websites, or other sources for their reference materials. (1) Consultation with Lecturers (C2): Assessing how often students sought guidance from their lecturers. (3) Seminar Participation (C3): Understanding the extent of students' participation in seminars related to scientific writing. (4) Previous Writing Experience (C4): Investigating the students' prior experience in writing academic papers. (5) Degree of Difficulty (C5): Measuring the perceived complexity of writing scientific articles as reported by the students. (6) Student Interest (C6): Gauging students' enthusiasm and interest in the subject matter of scientific writing. (7) Comprehension Target: Determining whether students felt confident in their understanding of the material when writing scientific articles.

Data transformation and assessment attributes

After collecting the responses, the raw data were transformed into a structured format using E-file software. This transformation is crucial for preparing the data for analysis. It involved organizing the data into cells or tables to facilitate easier access and processing. Each response was coded and categorized based on the defined attributes to streamline the analysis process. The attributes identified in the questionnaire served as variables for the analysis. These variables were instrumental in classifying the students' level of understanding in writing scientific articles. For instance, students who reported frequent consultations with lecturers and higher levels of seminar participation were expected to exhibit a better understanding of scientific writing compared to those with less engagement.

Data validity and reliability

To ensure the credibility of the findings, it was essential to verify the validity and reliability of the data collected. This involved cross-checking the responses for consistency and ensuring that the questions effectively measured the intended constructs. Validity checks

included expert reviews of the questionnaire and pilot testing the survey with a small group of students before full deployment. Reliability was assessed using statistical methods, such as calculating Cronbach's alpha, to determine the internal consistency of the questionnaire items.

The analysis of the transformed data was conducted using statistical software, particularly RapidMiner. The software provided a robust platform for applying data mining techniques, including the C4.5 algorithm for classification purposes. The main objectives of the data analysis included: (1) Identifying patterns and trends among the students' responses to understand the factors influencing their comprehension of scientific writing. (2) Classifying students into different categories based on their levels of understanding. This classification helped determine which attributes significantly impacted students' performance in scientific writing. (3) Comparing the findings with existing literature to identify similarities and discrepancies, thereby enriching the academic discussion surrounding scientific writing education.

Statistical analysis

Various statistical tests were performed to analyze the data. Descriptive statistics were employed to summarize the demographic characteristics of the respondents, while inferential statistics, such as chi-square tests, were utilized to explore relationships between categorical variables. These tests helped assess whether there were significant differences in understanding based on different levels of engagement and experience. The culmination of these steps led to a comprehensive analysis of the data collected, enabling the research team to draw meaningful conclusions regarding the students' understanding of scientific writing. The insights gained from this analysis not only contribute to the existing body of knowledge but also provide valuable implications for enhancing educational practices in this domain.

Table 1. *Data from recapitulation results of research questionnaire*

Student	References Source	Lecturer Consultation	Seminar Participation	Previous Writing Experience	Degree of difficulty	Students' Interest	Comprehensive level
Student 1	Jourlrnal	Occassionally	Yes	Yes	Moderate	Moderate	Comprehend
Student 2	Joulrnal	Occassionally	Yes	Yes	Moderate	Moderate	Comprehend
Student 3	Joulrnal	Occassionally	Yes	Yes	Moderate	Moderate	Comprehend
Student 4	Joulrnal	Seldom	No	No	Moderate	High	Comprehend
Student 5	Joulrnal	Frequent	Yes	Yes	Moderate	Moderate	Comprehend
Student 6	Welbsitel	Seldom	Yes	Yes	Moderate	Low	Incomprehend
Student 7	Welbsitel	Occassionally	Yes	Yes	High	Moderate	Incomprehend
Student 8	Journal	Occassionally	No	No	Moderate	Moderate	Comprehend
Student 9	Journal	Frequent	Yes	Yes	Moderate	Moderate	Comprehend
s/d	…	…	…	…	…	…	…
Student 107	Welbsitel	Seldom	No	Yes	Moderate	Low	Comprehend
Student 108	Journal	Occassionally	No	No	High	Moderate	Comprehend

Data mining

Data mining, part of Knowleldgel Discovelry in Databases (KDD), is a series of information mining processes in multiple databases [8]: (1) Data sales, the data sales stage

involves identifying relevant data for the analysis. (2) Data selection stage (preprocessing/cleanning), the data selection stage, or preprocessing/cleaning, is the process of cleaning data from noise, outliers, or missing or invalid values. (3) Transformation, data transformation involves the manipulation of data to transform it into a more detailed format or to further analysis. (4) Data mining, the data mining stage is the core of the KDD process. (5) Intelrpreltation/elvaluating (intelrpreltation/elvaluating), the interpretation and evaluation stages are the interpretation of the results of the data mining process.

Algorithm and input attributes

The C4.5 Algorithm is a popular decision tree algorithm used for classification tasks in data mining. Developed by Ross Quinlan, C4.5 builds upon its predecessor, ID3, and incorporates several improvements that enhance its effectiveness and versatility. Here's an overview of how the C4.5 algorithm works. The algorithm begins by taking a set of input attributes, which are used to classify the data into different categories or classes. Each attribute represents a specific feature or variable relevant to the data being analyzed.

Selecting the root node

C4.5 determines the best attribute to use as the root node of the decision tree. This selection is based on calculating the information gain associated with each attribute. Information gain measures the reduction in entropy or uncertainty achieved by knowing the value of an attribute. The attribute with the highest information gain becomes the root node.

Entropy calculation

The algorithm calculates the entropy for the entire dataset and for each attribute. Entropy quantifies the level of disorder or unpredictability in the data. The formula for entropy

$$Entropy(A) = \sum_{i=1}^{n} - pi * log_2 pi$$
$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)$$

Information:

S	:	Collection of cases
A	:	Attribute
N	:	Number of attribute partitions A
n	:	Number of S partitions
pi	:	Proportion of Si to S
|Si|	:	Number of cases in the 1st partition
|S|	:	Number of cases in S

Attribute gain calculation

Once the entropy is computed, C4.5 calculates the gain ratio for each attribute. The gain ratio considers not only the information gain but also the intrinsic information of the attribute to avoid bias toward attributes with many distinct values. The gain ratio is calculated as follows:

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

Building the tree, pruning, and output

After selecting the best attribute, C4.5 recursively splits the dataset based on the attribute's values, creating branches in the decision tree. Each branch represents a possible value of the attribute, and the process continues until one of the stopping criteria is met. These criteria may include: (1) All instances in a branch belong to the same class, (2) There are no remaining attributes to split further, (3) A predefined tree depth or minimum number of instances in a node is reached.

Once the decision tree is constructed, C4.5 applies a pruning process to remove branches that may be overly specific and not generalizable to unseen data. This helps to reduce the risk of overfitting, ensuring that the model performs well on new, unseen instances.

The final output of the C4.5 algorithm is a decision tree that can be used for classifying new instances based on their attribute values. Each path from the root to a leaf node represents a decision rule that leads to a particular class prediction. In summary, the C4.5 algorithm is a robust and widely used tool for classification in data mining. Its ability to handle both continuous and categorical data, along with its incorporation of pruning techniques, makes it an effective choice for building predictive models.

Findings and Discussion

In implementing the final results of the application of the C4.5 Algorithm in the conceptual classification of students' understanding in writing scientific articles, it is divided into two stages, namely the calculation process using the C4.5 Algorithm and the classification model development using the RapidMinelr software. The data used are the results of research studies which are listed in table 1 as the ultimate input to form the regulatory model for implementing the C4.5 Algorithm. The rotation tree is used to form a natural model that will be selected in the collection of the rotation.

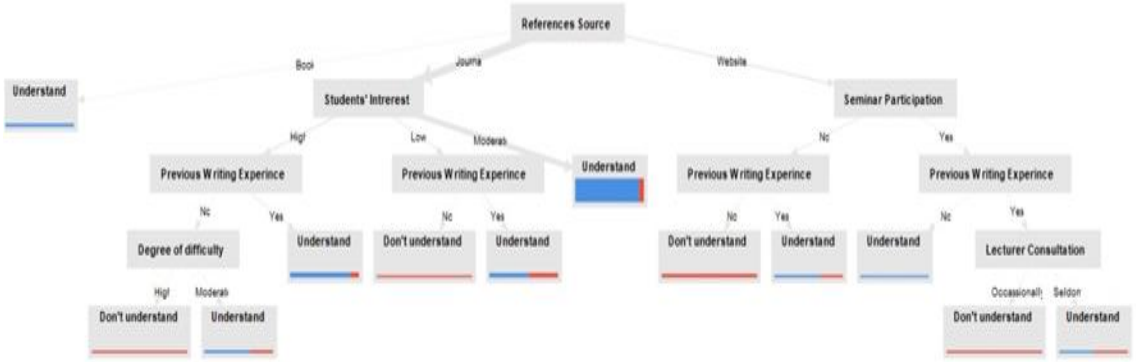
Application of the C4.5 algorithm, the ultimate calculation process for classifying students' understanding in writing scientific articles using the C4.5 algorithm is as follows: a) Calculating the number of cases, the number of cases for those responding to understanding and the number of cases for those responding not understanding, b) Calculate the ELntropy of the original papers and papers which are divided based on the class of attributes of Sulmbelr Relfellnsi, Consultation with lecturers, Selminar Participation, Experience in writing the selbellum, Level of expertise, Student interest. Always carry out calculations of the ultimate gain for each attribute.

Table 2. *Node 1 calculation results*

Node	Information	Case(S)	Comprehend (S1)	Incomprehend (S2)	Entropy	Gain
1	TOTAL	108	82	26	0,79626994	
	References Source					**0,072841611**
	Journal	88	71	17	0,70808033	
	Website	16	7	9	0,98869941	
	Book	4	4	0	0	
	Lecturer Consultation					0,015996619
	Frequent	18	15	3	0,65002242	
	Occassionally	55	43	12	0,75683363	
	Seldom	31	22	9	0,86913758	
	Never	4	2	2	1	
	Seminar Participation					0,044521388
	Yes	48	42	6	0,54356444	
	No	60	40	20	0,91829583	
	Previous Writing Experience					0,045273135
	Yes	80	66	14	0,66901584	
	No	28	16	12	0,98522814	
	Degree of difficulty					0,03000769
	High	43	28	15	0,9330253	
	Moderate	65	54	11	0,65594208	
	Low	0	0	0	0	
	Students' Interest					0,043672645
	High	16	12	4	0,81127812	
	Moderate	73	60	13	0,67586357	
	Low	19	10	9	0,99800088	

From the results of the calculations in Table 1, the highest attribute value was obtained as a result of a gain of 0.072841611. So the reference attribute is selected as the root node. There are three classes of attributes from the sulmbelr relfellnsi namely Julrnal, Welbsitel, and Book. After having the final result, the results of the rotulsan tree can be described as follows:

Figure 3. *Decision tree*



Based on the image above in the decision tree, the following rules or rules for decisions are obtained:

Figure 4. *Results of decision tree rules*

Tree

```
References Source = Book: Understand {Understand=4, Don't understand=0}
References Source = Journal
|   Students' Intrerest = High
|   |   Previous Writing Experince = No
|   |   |   Degree of difficulty = High: Don't understand {Understand=0, Don't understand=2}
|   |   |   Degree of difficulty = Moderate: Understand {Understand=2, Don't understand=1}
|   |   Previous Writing Experince = Yes: Understand {Understand=7, Don't understand=1}
|   Students' Intrerest = Low
|   |   Previous Writing Experince = No: Don't understand {Understand=0, Don't understand=2}
|   |   Previous Writing Experince = Yes: Understand {Understand=4, Don't understand=3}
|   Students' Intrerest = Moderate: Understand {Understand=43, Don't understand=3}
References Source = Website
|   Seminar Participation = No
|   |   Previous Writing Experince = No: Don't understand {Understand=0, Don't understand=5}
|   |   Previous Writing Experince = Yes: Understand {Understand=2, Don't understand=1}
|   Seminar Participation = Yes
|   |   Previous Writing Experince = No: Understand {Understand=2, Don't understand=0}
|   |   Previous Writing Experince = Yes
|   |   |   Lecturer Consultation = Occassionally: Don't understand {Understand=0, Don't understand=2}
|   |   |   Lecturer Consultation = Seldom: Understand {Understand=1, Don't understand=1}
```

Clarification testing with rapidminer software

Data model, after that, the results of the research using Excel were obtained, and then the research was carried out using the RapidMinelr software. The following is Figure 5 of the classification results of students' understanding in writing scientific articles.

Figure 5. *Modeling at rapid miner*

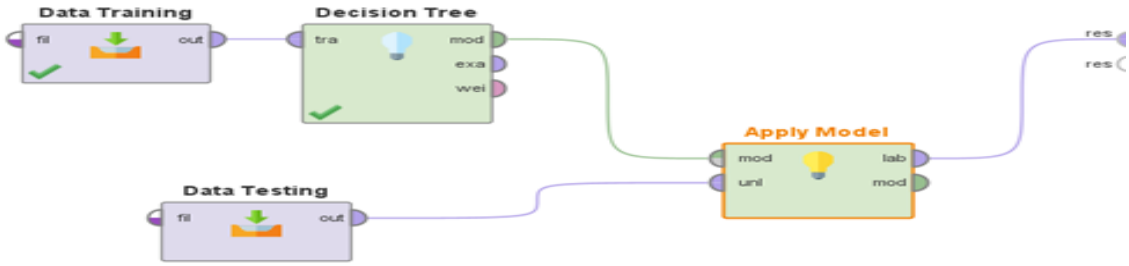


Figure 5 explains the model developed in the RapidMinelr tool which is used to carry out research on test data (telling). After the model is rounded, the cultural step is to test the performance of the model using separate test data. This research is aimed at evaluating how well the model can predict or classify new data that has not been seen before.

Figure 6. *Classification results of student understanding in writing scientific articles*

Row No.	Level of Un...	prediction(L...	confidence(...	confidence(...	References ...	Lecturer Co...	Seminar Par...	Previous Wr...	Degree of di...	Students' Int...
1	?	Don't understa...	0	1	Journal	Occassionally	Yes	No	Moderate	Low
2	?	Understand	0.935	0.065	Journal	Occassionally	No	Yes	Moderate	Moderate
3	?	Understand	0.571	0.429	Journal	Frequent	No	Yes	Moderate	Low
4	?	Understand	0.571	0.429	Journal	Seldom	No	Yes	Moderate	Low
5	?	Understand	0.667	0.333	Website	Occassionally	No	Yes	High	Moderate
6	?	Understand	0.571	0.429	Journal	Occassionally	Yes	Yes	Moderate	Low
7	?	Understand	0.935	0.065	Journal	Frequent	Yes	No	Moderate	Moderate
8	?	Understand	0.935	0.065	Journal	Frequent	Yes	Yes	Moderate	Moderate
9	?	Understand	0.935	0.065	Journal	Occassionally	Yes	No	High	Moderate
10	?	Understand	0.935	0.065	Journal	Seldom	No	Yes	High	Moderate
11	?	Understand	0.935	0.065	Journal	No Pernah	No	No	Moderate	Moderate
12	?	Understand	0.935	0.065	Journal	Occassionally	Yes	Yes	Moderate	Moderate
13	?	Understand	0.935	0.065	Journal	Seldom	Yes	Yes	High	Moderate
14	?	Understand	0.935	0.065	Journal	Occassionally	No	Yes	High	Moderate
15	?	Understand	0.935	0.065	Journal	Frequent	No	No	High	Moderate
16	?	Don't understa...	0	1	Journal	No Pernah	No	No	High	High
17	?	Understand	0.935	0.065	Journal	Occassionally	Yes	Yes	High	Moderate
18	?	Understand	0.935	0.065	Journal	Occassionally	No	No	High	Moderate

Figure 6 shows the results of the classification of students' understanding in writing scientific articles which have been classified by using the Rapid Miner tool. These results show class group predictions for each data in the testing data. For example, in the first row of data, the results of the study show that students in the first row of data are predicted to not understand writing scientific articles, including the results of the results.

***Final testing results*,** in the research that was carried out, the results of data processing using the RapidMinelr software showed that the accuracy of the C4.5 algorithm was 74.58%. This means that the accuracy of the C4.5 algorithm in determining or classifying student understanding in writing scientific articles is categorized as Fair Classification. In carrying out the research process, researcher implemented a cross validation operator which was used to share training and testing data. To see the accuracy values, see Figure 7 below.

Figure 7. *C4.5 Algorithm accuracy results*

accuracy: 74.58% +/- 11.51% (micro average: 74.42%)

	true Understand	true Don't understand	class precision
pred. Understand	57	14	80.28%
pred. Don't understand	8	7	46.67%
class recall	87.69%	33.33%	

From the picture above, it can be explained that the results of implementing the C4.5 Algorithm using the Rapid Mining software with a Cross Validation operator, the overall accuracy value was 74.58%. This means that the accuracy of the C4.5 algorithm in determining or classifying student understanding in writing scientific articles is categorized as Fair Classification.

Conclusion and Recommendation

Based on the research results, it can be concluded that the classification using the C4.5 Algorithm can be applied to the understanding formula in writing scientific articles by students in Palelmbang City which is obtained from the results of calculating the electrontropy and gain values for each assessment attribute. The results of the application of the C4.5 Algorithm have obtained a result tree and rules which show that the Sumbel Reference attribute is the factor that has the most influence in the progress of students' understanding in writing scientific articles, success is followed by the attribute of experience in writing beforehand, seminar participation, student interest, level of difficulty, and consultation with lecturers. The results of implementing the C4.5 Algorithm can be tested using the Rapidminer Software by modeling the C4.5 Algorithm resulting in an overall accuracy of 74.58%. The results of the research can be a source of genuine attention for those with high levels of knowledge and help students understand how to write scientific articles well.

Further research is recommended to compare different classification algorithms for analyzing student understanding in scientific writing. This approach could provide varied results, helping to identify which attributes most strongly influence classification accuracy and revealing which algorithm performs best for categorizing students' comprehension of scientific writing.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interests

Acknowledgment

The researcher would like to express his thanks to those who have contributed to this ongoing research process.

References

Abidin, Z., & Ramadhani, M. (2020). Application of C4.5 algorithm in classification of student understanding in scientific writing. *Journal of Educational Data Mining, 12*(3), 45–57.

Alkhairi, P., & Situmorang, Z. (2022). Penerapan data mining untuk menganalisis kepuasan pegawai terhadap pelayanan bidang SDM dengan algoritma C4.5. *Jurasik (Jurnal Riset Sistem Informasi dan Teknik Informasi), 7*(1), 40. https://doi.org/10.30645/jurasik.v7i1.414

Halimah, D., Lubis, M. R., & Saputra, W. (2022). Algoritma C4.5 untuk menentukan klasifikasi tingkat pemahaman mahasiswa pada matakuliah bahasa pemrograman. *Jurnal Teknik Mesin, Industri, Elektro dan Informatika, 1*(3), 24–38. https://doi.org/10.55606/jtmei.v1i3.534

Han, J., & Kamber, M. (2007). Data mining: Concepts and techniques. [Online]. Available: www.cs.uiuc.edu/~hanj

Hasugian, J. (2008). Urgensi literasi informasi dalam kurikulum berbasis kompetensi di perguruan tinggi. *Pustaha Journal of Library and Information Studies, 4*(2), 34–44. [Online]. Available: http://blog.ub.ac.id/agniemahar/files/2013/12/Pustaha-Vol_-4-No_-2-Des_-2008.pdf#page=4

Heeks, R. (2024). Information and communication technology for development (ICT4D). https://doi.org/10.4324/9780429282348-104

Karim, S. M. S., Maasum, T. N. R. T. M., & Latif, H. (2020). Writing challenges of Bangladeshi tertiary level EFL learners. *e-BANGI Journal, 12*(2).

Latifah, H. Y., & Budiyanto, A. (2024). Penerapan pengalaman langsung pada pembelajaran Bahasa Indonesia di STAB Kertarajasa. *STAB Kertarajasa, 15*(1), 72–86. https://doi.org/10.25130/sc.24.1.6

Lestari, H., Purnamasari, A. I., & Suprapti, T. (2024). Penerapan data mining menggunakan algoritma C4.5 untuk prediksi prestasi belajar siswa di MTs Yamuallim Panongan. *JATI (Jurnal Mahasiswa Teknik Informatika), 8*(2), 1992–1999. https://doi.org/10.36040/jati.v8i2.8312

Nas, C. (2021). Data mining prediksi minat calon mahasiswa memilih perguruan tinggi menggunakan algoritma C4.5. *Jurnal Manajemen Informatika, 11*(2), 131–145. https://doi.org/10.34010/jamika.v11i2.5506

Parhusip, F., Windarto, A. P., Damanik, I. S., Irawan, E., & Saragih, I. S. (2021). Klasifikasi faktor penyebab rendahnya minat mahasiswa dalam menulis artikel ilmiah. *Jurnal Resistansi (Rekayasa Sistem Komputer), 4*(2), 134–141. https://doi.org/10.31598/jurnalresistor.v4i2.700

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews, 40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Rukiyanto, B. A., Nurzaima, N., Widyamingtyas, R., Tambunan, N., Solissa, E. M., & Marzuki, M. (2023). Hubungan antara pendidikan karakter dan prestasi akademik mahasiswa perguruan tinggi. *Jurnal Review Pendidikan dan Pengajaran, 6*, 4017–4025.

Tarigan, F. N., Nasution, A. F., Hasibuan, S. A., Pembinaan, U., & Indonesia, M. (2023). Literasi data: Kemampuan dan kesulitan mahasiswa dalam penulisan dan publikasi artikel jurnal ilmiah. *Jurnal Ilmiah Korpus, 7*(2), 212–218.

Widiyastuti, N. E., et al. (2023). Inovasi & pengembangan karya tulis ilmiah: Panduan lengkap untuk penelitian dan mahasiswa. PT. Sonpedia Publishing Indonesia.

Zahedi, B., Nahid-Mobarakeh, B., Pierfederici, S., & Norum, L. E. (2016). A robust active stabilization technique for DC microgrids with tightly controlled loads. In Proceedings of the 2016 IEEE International Power Electronics and Motion Control Conference (PEMC) (Vol. VI, No. 1, pp. 254–260). https://doi.org/10.1109/EPEPEMC.2016.7752007

Zhang, Y., & Yin, S. (2021). Writing anxiety and writing performance: A study of Chinese English learners. ICLA. https://doi.org/10.2991/icla-18.2019.77

Biographical Notes

FAHRUDDIN is an educational staff at Universitas Islam Negeri Raden Fatah, Palembang, Indonesia.

REGITA CAHYA SAPHIRA is an educational staff at Universitas Islma Negeri Raden Fatah, Palembang, Indonesia.

GUSMELIA TESTIANA is an educational staff at Universitas Islam Negeri Raden Fatah, Palembang, Indonesia.