Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2022. Vol. 1. No. 2

# Application Of The C4.5 Algorithm To Determining Student's Level Of Understanding

**Adeita A. Ndraha[1]*, Jaya Tata Hardinata[2], Yuegilion Pranayama Purba[3]**

*[1,2,3]STIKOM Tunas Bangsa Pematangsiantar, NorthSumatra, Indonesia*
*[3]AMIK Tunas Bangsa Pematangsiantar, NorthSumatra, Indonesia*
**adendraha99@gmail.com*

**Abstract**

This research was conducted to find the rules of the model in measuring the level of students' understanding of the subject. During this pandemic, the learning process is carried out online, so it is difficult to measure students' ability to master the material. This measurement needs to be done so that the evaluation process can be carried out so that the ability of students in one group to achieve the target level of understanding. Currently, evaluation activities have never been carried out because they do not have a model so that evaluation can only be done by giving quizzes and exercises. This problem can be solved by using data mining algorithm C4.5. Attributes used as parameters for assessing student understanding of lessons such as Teaching Method (C1), Learning Media (C2), Communication (C3), Experience (C4), Teaching Materials/Modules/Assignments (C5), Learning Duration (C6). The six attributes are used to find the relationship between each other that influence each other to get the highest root so that a decision tree will be obtained that produces the rules of the relationship between each attribute in determining student understanding of the subject. This rule or rule will be used as the basis for making an information system so that it can be applied to end users, namely schools.

***Keywords***: *Data Mining; Algorithm C4.5; Rule, Level of Understanding; Student*

## 1. Introduction

Along with the times, many changes in technology and information. The role of computers is very helpful for human work so that it is faster to recognize various aspects, the development of technology all computers perform data processing which is one of the important things in information technology[1],[2], one method in data processing is classification which is a way of grouping data according to the characteristics or characteristics of the data. Data mining can classify students' abilities in understanding lessons[3],[4],[5], one of the data mining that can be used to classify students' understanding when learning online is the C4.5 algorithm[6],[7]. The main basis for many students who do not understand learning is the interest of the students themselves, Other factors that also affect students' understanding are the way an educator teaches, the learning media used, experience, communication, teaching materials/modules/assignments, and the duration of learning[8]. Based on some of the factors above, the school wants to know how the level of students' understanding during the online teaching and learning process.

From these problems the method applied is the C4.5 algorithm[9],[10], which is widely known and used for data classification that has numeric and categorical attributes, The data collection method used is filling out the questionnaire, the content of the questionnaire is in the form of factors that affect the level of students' understanding. Research that has utilized the advantages of the classification method with the C4.5 algorithm to solve problems, including in the case of student difficulty factors in programming languages[11]. The results of the study stated that the interest factor was the first node that influenced students' understanding of programming languages. Furthermore, research on the case of students' level of understanding of the subject with the accuracy level of the C4.5 algorithm classification model is 87.10%.

## 2. Research methodology

This research was conducted based on the problem formulation described in the previous chapter, namely to classify students' understanding of the lesson. The purpose of writing this thesis is to classify the concept of the extent to which students understand the lesson, and can help a teacher in improving teaching methods, as well as providing motivation for students. In completing this thesis the author uses quantitative research which demands more on the use of numbers. Where numerical computing is an approach to mathematical problems using several numerical methods. This research work activity diagram is a description in the form of work that will be carried out on the classification analysis of students' understanding levels when learning online. The workflow carried out by the author in this study is presented in Figure 1 in the activity diagram.

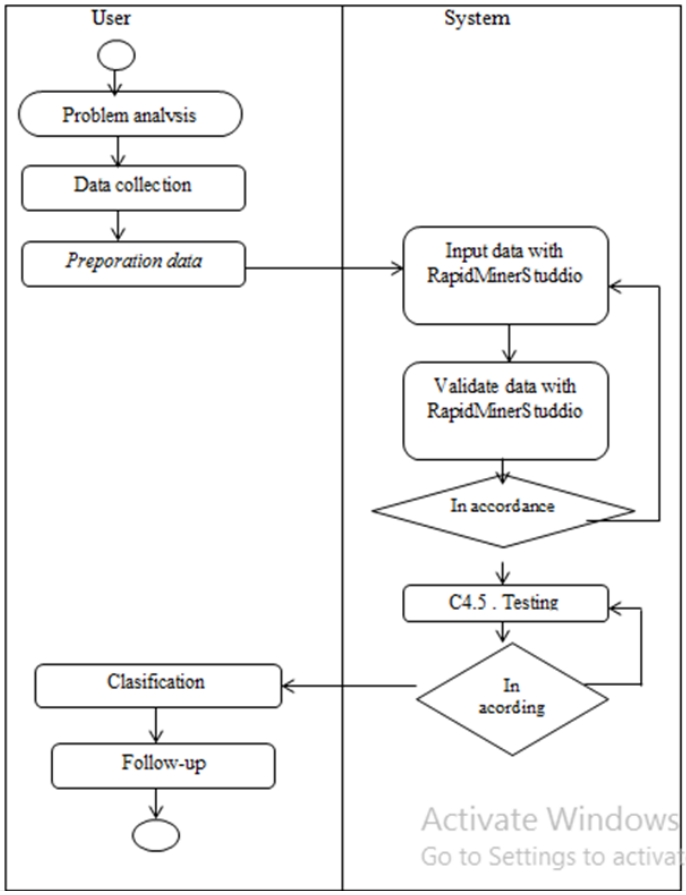


**Figure 1.** Work Activity Diagram

The method used in this study is the C4.5 algorithm. In this modeling the C4.5 algorithm will identify the sample from the data set, then calculate the entrophy (S) of all attributes, after the entrophy (S) is found then calculate the highest gain of all attributes, then get the attribute that will be used as the root / node. Create a branch for each value, divide the cases in the branch, repeat the Gain calculation until all data are included in the same class. The selected attribute is no longer included in the calculation, the decision tree formation process stops when there are no partitioned attributes and all tuples in node N have the same class. The steps taken to determine the decision tree are preparing training data and testing data taken from previous history, calculating the entropy and gain for each attribute, and the highest gain value being the root of the tree. Before calculating the gain value, you must first calculate the entropy value, The formula for entropy is:

$$\text{Entropy(S)} = \sum_{i=1}^{n} -\text{pi} * log2pi \quad (1)$$

After calculating the entropy value, calculate the gain value, using the formula:

$$\text{Gain (S, A)} = \text{Entropy (S)} - \sum_{i=1}^{n} \left(\frac{Si}{S}\right) * \quad Entropy\ (S) \quad (2)$$

Information :
S : Case Collection
A : Attributes
n : Number of attribute partitions A
|Si| : Number of cases partition to i
Pi : Proportion of Si to S
n : Number of partitions to S
|S| : Number of cases in S

Repeat step 2 until all records are partitioned and the tree partitioning process will stop when all records in node N get the same class, no attributes in the records are partitioned again and no records in the branch are empty.

## 3. Results And Discussion

In implementing the final result of the C4.5 algorithm, it is carried out in two stages, namely manual calculations and adjusting the final results of manual calculations with RapidMiner software. The following are the calculation steps in the C4.5 algorithm in solving the classification case of students who understand the lesson that will be divided into two labels understand and do not understand. The first

process calculates the entropy of all cases divided by way of teaching (C1), learning media (C2), communication (C3), teacher experience (C4), teaching materials /modules / assignments (C5), duration of learning (C6). After that, the calculation of gain is done for each attribute.

$$Entropy\ (\text{Total}) = \left(-\frac{51}{120} * \log_2\left(\frac{51}{120}\right)\right) + \left(-\frac{69}{120} * \log_2\left(\frac{69}{120}\right)\right) = 0{,}98371$$

The results can be seen in the following table:

**Tabel 1.** Entropy Total Kasus

| Total Cases | Total (Understood) | Total (Don't Understand) | ***Entropy*** |
|---|---|---|---|
| 120 | 51 | 69 | 0,98371 |

Then calculate the entropy of all cases which are divided based on the attributes of teaching method, learning media, communication, teacher experience, teaching materials/modules/assignments, and learning duration.
The calculation results can be seen in table2 below:

**Table 2.** calculation results of node 1

| Atribut | | amount | understanding | Do not understanding | Entropy | Gain |
|---|---|---|---|---|---|---|
| Total | | 120 | 51 | 69 | 0,98371 | |
| how to teach | | | | | | 0,13124 |
| | Not enough | 25 | 4 | 21 | 0,63431 | |
| | enough | 36 | 10 | 26 | 0,85241 | |
| | good | 42 | 25 | 17 | 0,97367 | |
| | Very good | 17 | 12 | 5 | 0,87398 | |
| learning Media | | | | | | 0,51436 |
| | Not enough | 9 | 0 | 9 | 0 | |
| | enough | 70 | 12 | 58 | 0,66096 | |
| | Good | 25 | 23 | 2 | 0,40218 | |
| | Very good | 16 | 16 | 0 | 0 | |
| experience | | | | | | 0,19452 |
| | Not enough | 6 | 1 | 5 | 0,65002 | |
| | enough | 45 | 8 | 37 | 0,67519 | |
| | Good | 61 | 34 | 27 | 0,99048 | |
| | Very good | 8 | 8 | 0 | 0 | |
| Comunication | | | | | | 0,26449 |
| | Not enough | 6 | 3 | 3 | 1,00000 | |
| | enough | 50 | 8 | 42 | 0,63431 | |
| | good | 39 | 17 | 22 | 0,98811 | |
| | Very good | 25 | 23 | 2 | 0,40218 | |
| teaching materials | | | | | | 0,35639 |
| | Not enough | 7 | 0 | 7 | 0 | |
| | Enough | 51 | 7 | 44 | 0,57700 | |
| | Good | 36 | 20 | 16 | 0,99108 | |
| | Very good | 26 | 24 | 2 | 0,39124 | |
| Duration | | | | | | 0,49924 |
| | Not enough | 32 | 2 | 30 | 0,33729 | |
| | Enough | 23 | 0 | 23 | 0 | |
| | Good | 54 | 38 | 16 | 0,87672 | |
| | Very good | 11 | 11 | 0 | 0 | |

From the table above, we can see that the attribute of learning media has the highest Gain, namely 0.51436, then the learning media becomes Nodeakar, learning media has 4 values, namely less, enough, good, and very good. The less and very good scores have classified the cases into one, namely understand and don't understand decisions, while for sufficient and good scores, further calculations are needed because they still have results between understand and don't understand. The calculation is carried out on the next node and the final result of the calculation can be seen in table 3 below:

**Table 3.** Calculation Results for Nodes 2

| Learning media – good=materials learning – enough = how to teach– very good | | amount | understanding | Do not understanding | Entropy | Gain |
|---|---|---|---|---|---|---|
| Total | | 2 | 1 | 1 | 1,00000 | |
| | | | | | | 0,00000 |
| | Not enough | 0 | 0 | 0 | 0 | |
| Experience | enough | 0 | 0 | 0 | 0 | |
| | good | 2 | 1 | 1 | 1,00000 | |
| | Very good | 0 | 0 | 0 | 0 | |
| | | | | | | 1,00000 |
| | Not enough | 1 | 1 | 0 | 0 | |
| Comunication | enough | 0 | 0 | 0 | 0 | |
| | good | 0 | 0 | 0 | 0 | |
| | Very good | 1 | 0 | 1 | 0 | |
| | | | | | | 0,00000 |
| | Not enough | 0 | 0 | 0 | 0 | |
| Duration | enough | 0 | 0 | 0 | 0 | |
| | good | 2 | 1 | 1 | 1,00000 | |
| | Very good | 0 | 0 | 0 | 0 | |

From the results of the calculation in table 3 there is the highest gain value, namely communication of 1.00000 which consists of four sub-attributes less, sufficient, good, and very good. In the table above, we can see that the sub-attribute of the communication pad does not have an entropy value, then the calculation has been completed at node 2. The rules resulting from the above calculations are as follows:

1. If the learning media is lacking, the results do not understand.
2. If the learning media is very good then the result is understood
3. If the learning media is sufficient and the duration is lacking and the communication is good, the result is not understood.
4. If the learning media is sufficient and the duration is insufficient and the communication is sufficient then the result is not understood.
5. If the learning media is sufficient and the duration is less and the communication is lacking, the results are not understood.
6. If the learning media is sufficient and the duration is less and the communication is very good and the teaching method is good, the result is not understood, the result is not understood.
7. If the learning media is sufficient and the duration is less and the communication is very good and the teaching method is sufficient, the result is not understood, the result is not understood.
8. If the learning media is sufficient and the duration is sufficient then the results are not understood.
9. If the learning media is sufficient and the duration is very good, the results are understandable
10. If the learning media is sufficient and the duration is good and the communication is lacking, the result is not understood.
11. If the learning media is sufficient and the duration is good and the communication is very good then the result is understandable.
12. If the learning media is sufficient and the duration is good and the communication is good and the teaching materials are very good then the result is understandable.
13. If the learning media is sufficient and the duration is good and the communication is good and the teaching materials are good and the teaching method is good then the result is not understood.
14. If the learning media is sufficient and the duration is good and the communication is good and the teaching materials are good and the teaching method is sufficient then the result is understandable.
15. If the learning media is good and the teaching materials are very good, the results are understandable.
16. If the learning media is good and the teaching materials are good, the results are understandable.
17. If the learning media is good and the teaching materials are sufficient and the teaching method is good, the result is understood.
18. If the learning media is good and the teaching materials are sufficient and the teaching method is lacking, the result is not understood.
19. If the learning media is good and the teaching materials are sufficient and the teaching method is very good and the communication is lacking, the result is understanding.
20. If the learning media is good and the teaching materials are sufficient and the teaching method is very good and the communication is very good then the result is not understood.

Experiments using Rapidminer can be seen in the following image:

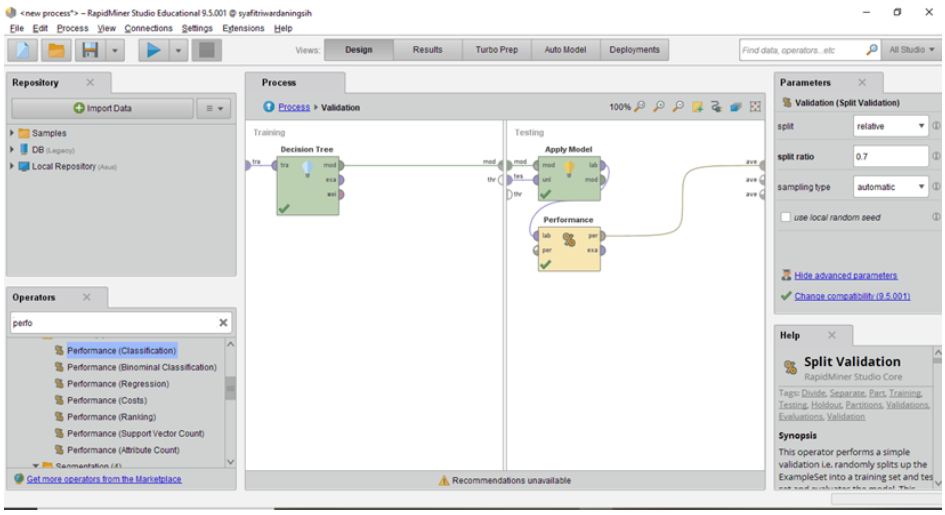


**Figure 2.** Input Operators Perfomance

The next step is to connect the ports from the decision tree operators, apply model operators, and performance operators, then click the run icon on the toolbar to display the results.

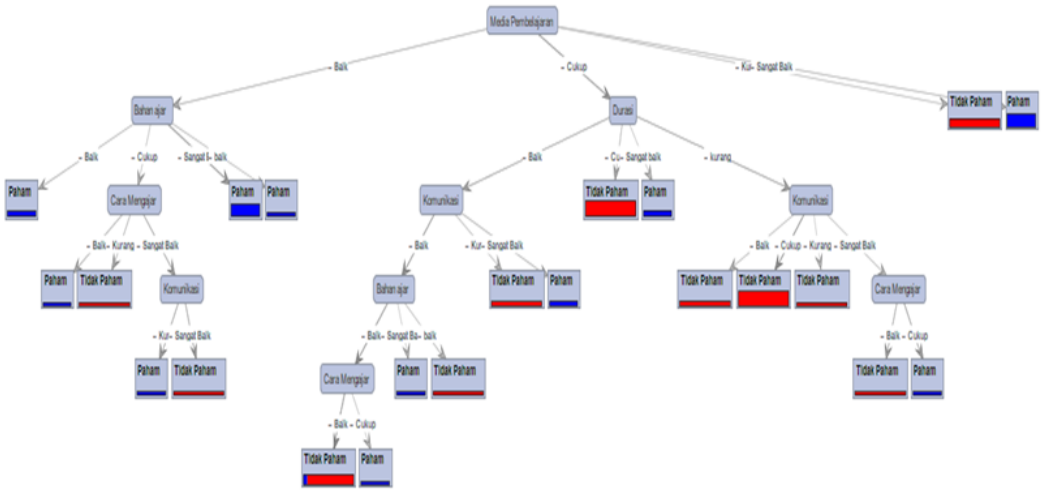


**Figure 3.** Decision Tree results

After calculating and testing the data with the C4.5 algorithm, the final decision pattern is obtained.

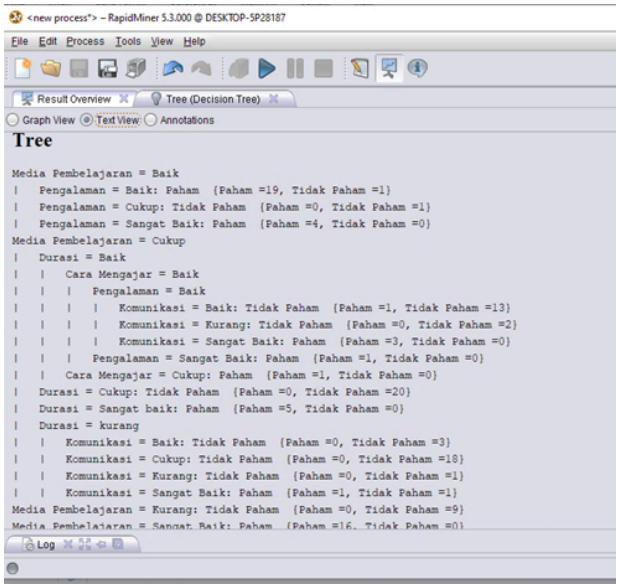


**Figure 4.** Rule Decision Tree on Rapidminer

From manual calculations on node 1, it was found that the learning media factor was the dominant factor that influenced the level of students' understanding when learning online and had also been tested with the repidminer application on the decision tree results and decision tree rule results in the learning media results. So from manual calculations and using the rapidminer application that has been compared, the same results are obtained, namely learning media.

## Acknowledgement

Acknowledgments to the supervisors and examiners who are lecturers at AMIK and STIKOM Tunas Bangsa so that this research can be arranged as one of the requirements for completing Bachelor's education (S1) at STIKOM Tunas Bangsa. I hope this research can be a reference for other research related to the methods and algorithms used. I hope for constructive suggestions for the readers for the perfection of this research in the future.

## Conclusion

The problem of determining the level of students' understanding when learning online was successfully solved by data mining techniques, namely the C4.5 algorithm which produced 20 rules. The understanding value of the recall system accuracy is 73.33% and the understanding does not understand is 95.24% with the understanding precision value is 91.67% and does not understand is 83.33%. With the application of data mining algorithm C4.5 is expected to be able to provide a solution in determining the level of understanding of students.

## References

[1] W. Yadiati and Meiryani, "The role of information technology in E-Commerce," *Int. J. Sci. Technol. Res.*, vol. 8, no. 1, pp. 173–176, 2019.

[2] W. Cascio and R. Montealegre, "How Technology Is Changing Work and Organizations," *Annu. Rev. Organ. Psychol. Organ. Behav.*, vol. 3, pp. 349–375, Mar. 2016, doi: 10.1146/annurev-orgpsych-041015-062352.

[3] M. A. Saleh, S. Palaniappan, N. Ali, and A. Abdalla, "Education is An Overview of Data Mining and The Ability to Predict the Performance of Students," vol. 15, no. 1, pp. 19–28, 2021.

[4] M. Pechenizkiy, T. Calders, E. Vasilyeva, and P. De Bra, *Mining the Student Assessment Data: Lessons Drawn from a Small Scale Case Study.* 2008.

[5] S. Pisani, D. Fioriti, M. P. Conte, F. Chiarini, L. Seganti, and A. M. Degener, "Involvement of herpes simplex virus type 2 in modulation of gene expression of human papillomavirus type 18," *Int. J. Immunopathol. Pharmacol.*, vol. 15, no. 1, pp. 59–63, 2002, doi: 10.1177/039463200201500108.

[6] S. W. Siahaan, K. D. R. Sianipar, P. P. P. A. N. . F. Ilmi R.H Zer, and D. Hartama, "Application of C4.5 Algorithm in Improving English Skills in Students," *J. Inform. Univ. Pamulang*, vol. 5, no. 3, p. 261, 2020, doi: 10.32493/informatika.v5i3.5268.

[7] S. Sucipto, K. Kusrini, and E. Luthfi, *Classification method of multi-class on C4.5 algorithm for fish diseases*. 2016.

[8] Rohmatillah, "A Study On Students' Difficulties In Learning Vocabulary (Bachellors' Degree)," *J. Raden Intan Lampung*, vol. 3, no. 1, pp. 69–86, 2014, [Online]. Available: http://repository.unej.ac.id/handle/123456789/18942.

[9] A. A. Aprilia Lestari, "Increasing Accuracy of C4 . 5 Algorithm Using Information Gain Ratio and Adaboost for Classification of Chronic Kidney Disease," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 32–38, 2020.

[10] A. A. Septiantina and E. Sugiharti, "Optimization of C4 . 5 Algorithm Using K-Means Algorithm and Particle Swarm Optimization Feature Selection on Breast Cancer Diagnosis," vol. 2, no. April, pp. 51–60, 2020.

[11] A. B. U. Nájera and J. de la Calleja Mora, "Brief review of educational applications using data mining and machine learning," *Rev. Electron. Investig. Educ.*, vol. 19, no. 4, pp. 84–96, 2017, doi: 10.24320/redie.2017.19.4.1305.