Journal of Applied Engineering and Technological Science
Vol 3(2) 2022 : 208-213

IMPLEMENTATION OF THE C4.5 ALGORITHM IN DESCRIBING THE TRENDS OF THE HUMAN CONSCIOUSNESS AND UNCONSCIOUSNESS

Ronaldo Syahputra[1*], Yeviki Maisyah Putra[2]
Information Systems Study Program, Putra Indonesia University YPTK Padang, Indonesia[1,2]
ronaldo_syahputra@upiyptk.ac.id,

Received : 26 May 2022, Revised: 28 June 2022, Accepted : 30 June 2022
*Corresponding Author

ABSTRACT
The human mind has two properties that have different and conflicting functions. The two characteristics of the mind are the conscious mind and the subconscious mind. This study uses the C4.5 data mining algorithm to describe or see the tendency of the conscious and subconscious or the level of suggestion. Suggestibility is the most important thing in hypnotherapy. Hypnotherapy is a therapy performed under hypnosis. Hypnosis is communication with the human subconscious. The C4.5 algorithm for Data Mining is used to form a decision tree. This research will produce a decision tree that can explain the suggestive level of a series of tests that have been carried out. Testing is done with RapidMiner software to get a decision tree. The test consists of a series of tests consisting of four types of tests, where test 1 is to measuring right brain dominance , test 2 is to measure the speed of receiving instructions, test 3 is to measure a person's creativity, and test 4 is to know person's level of understanding, reasoning and imagination. The results of manual calculations were carried out in this study later with the results obtained from the results of testing with the RapidMiner software.
Keywords:*Data Mining, C4.5, Suggestibility, Decision Trees, RapidMiner.*

1. Introduction

Data Mining is the process of discovering a new set of patterns from a large set of data. (Florence & Savithri, 2013)(Hssina et al., 2014). Actually, Data Mining is a step in knowledge discovery in databases (KDD). Knowledge discovery is a process consisting of data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation(Hosseini & Sardo, 2021; Dias, et al., 2021; Yuliansyah, et al., 2021).

To apply the concept of Data Mining in classifying data and forming decision trees, one way that can be done is to apply the C4.5 algorithm. The C4.5 algorithm, also known as the decision tree algorithm, is a very powerful and well-known classification and prediction method (Dai & Ji, 2014)(Nursela, 2010). The C4.5 algorithm has been widely used in various scientific studies. Dwi Ayu Nursela (2010) in her research used the C4.5 algorithm to classify the degree of malignancy of breast cancer.

Zeidi, et al(2022) in their research also used the C4.5 algorithm for data classification of Indian Pima diabetes. The C4.5 algorithm is used to build a decision tree (decision making). A decision tree is a structure that can be used to divide a large data set into smaller record sets by applying a set of decision rules (Mambang & Marleny, 2015). In their research, Mambang and Finki (2015) used the C4.5 algorithm for data classification which resulted in a decision tree in predicting prospective new students. The data classification process that produces this decision tree makes it easy for humans to interpret the data set that represents the rules in the decision tree. Rules can be easily understood in natural language. The decision tree will provide useful knowledge by finding hidden relationships between a number of potential input variables and target variables(Sinaga, et al., 2021; Tempola, et al., 2022; Abdullah, et al., 2021).

The human mind has two characteristics that have different and contradictory functions. The two characteristics of the mind are the conscious mind and the subconscious mind. Generally known as the conscious and subconscious. The subconscious plays an important role in hypnotherapy. Hypnotherapy is a therapy performed under hypnosis. Hypnotherapy is basically the art of communication, that hypnosis is a state of relaxation of the mind accompanied by relaxation of the body. According to the Big Indonesian Dictionary, hypnosis is a state of hypnosis; related to hypnosis. Meanwhile, hypnosis itself is a "sleep-like state due to suggestions,

which at the initial stage the person is under the influence of the person giving the suggestion, but at the next stage becomes completely unconscious"(Sanyal, et al., 2022).

2. Research Methods

Research is a process of searching for something systematically in a relatively long time by using the scientific method based on applicable procedures and regulations. Research activities require a methodology that contains a framework of thought. The framework of thought is a description of the steps that will be carried out so that the research can run systematically and the expected goals can be achieved. This framework is the steps that will be taken in solving the problems that will be discussed. The framework of this research can be described in Figure 2.1 below :

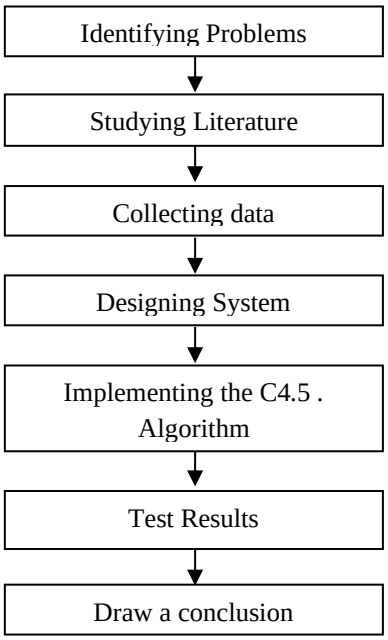

Fig. 1. Research Methods.

On the other hand, Data Mining is a process that using ic statistics, math, intelligence artificial, and machine learning to find out useful information. Data mining is an important step in the process of finding knowledge by exploring the added value so far unknown manually (Lakshmi & Raghunandhan, 2011) Model proposed in this study is the algorithm Decision C4.5(Wang, 2022).

At the data-learning stage, the C4.5 algorithm constructs a decision tree from the training data, which are cases or records (tuples) in the database. Each case contains the values of the attributes for a class. Each attribute can contain discrete or continuous (numeric) data. C4.5 also handles cases where there is no value for one or more attributes. However, class attributes are only of a discrete type and cannot be empty(Wang & Gao, 2021; Mijwil & Abttan, 2021).

The three working principles of the C4.5 algorithm at the data-learning stage are:

1. Decision tree creation.

The objective of the decision tree algorithm is to construct a tree data structure (called a decision tree) that can be used to predict the class of a new case or record that does not yet have a class. C4.5 constructs a decision tree with a divide and conquer strategy. At first, only root nodes are created by applying the divide and conquer algorithm. This algorithm chooses the best case solution by calculating and comparing the gain ratio, then on the nodes formed at the next level, the divide and conquer algorithm will be applied again. And so on until the leaves are formed. The C4.5 algorithm can produce a decision tree, where the square symbol represents the node and the ellipse represents the leaf(Permana, et al., 2021).

2. Decision tree pruning and evaluation (optional).

Because the constructed tree can be large and not easy to read, C4.5 can simplify the tree by pruning it based on the confidence level value. In addition to reducing tree size, pruning also aims to reduce the prediction error rate in new cases (records).

3. Generating rules from a decision tree (optional).

The rules in the form of if-then are derived from the decision tree by tracing from the root to the leaf. Each node and its branching conditions will be given in the if, while the value for the leaf will be written in then. After all the rules are created, the rules will be simplified (merged or generalized). In general, the C4.5 algorithm for building a decision tree is as follows (Saikhu et al., 2011) :

a. Select attribute as root
b. Create a branch for each value
c. Split cases in branches
d. Repeat the process for each branch until all cases on the branch have the same class.

Select an attribute as the root, it is based on the highest gain value of the existing attributes. To calculate the gain, the formula as stated in the formula is used:

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

Where:
S:Case set
A:Attribute
n:Number of partitions A
|*Si*|:Number of cases on partition i
|*S*|:Number of cases in S

While the calculation of the entropy value can be seen in the following formula 2:

$$Entropy(S) = \sum_{i=1}^{n} -pi * log2\ pi \quad (2)$$

Where:
S: Case set
n: number of partitions S
pi: Proportion of Si to S

There are 4 psychological test tools used in this study. Each test has a different purpose. The test also has a value that will be used for the calculation of the C4.5 algorithm. The objectives of some of these psychological tests are as follows:

1. Test 1

This test is used to measure the ability to catch and the speed of receiving instructions or commands.

2. Test 2

This test is the Torrance test. In psychology, this test is used to measure a person's creativity.

3. Test 3

This test is a Spatial psychological test that is used to determine a person's level of understanding, reasoning, and imagination

4. Test 4

This test is a test used to measure right brain dominance.

3. Results and Discussions

After calculating Node 1, Node 2, Node 3, then proceed with the calculation of Node 4. The calculation results can be seen in the following table:

Table 1 – Node 4 calculation.

nodes			amount	high	currently	entropy	gain
4	1,2,4-high test		5	3	2	0.970950594	
	test 3						0.970950594
		high	3	3	0	0	
		low	2	0	2	0	

The results of the above calculations are calculated by the following equation:

$$Entrophy(Total) = \sum_{i=s}^{n} -pi * \log_2 pi$$

$$Entrophy\ (Total) = \left(-\frac{3}{5} * \log_2\left(\frac{3}{5}\right)\right) + \left(-\frac{2}{5} * \log_2\left(\frac{2}{5}\right)\right)$$

$$= 0.970950594$$

Entropy test value 3

$$Entrophy\ (Test\ 3 - high) = \left(-\frac{3}{3} * \log_2\left(\frac{3}{3}\right)\right) + \left(-\frac{0}{3} * \log_2\left(\frac{0}{3}\right)\right)$$

$$= 0$$

$$Entrophy\ (Test\ 3 - low) = \left(-\frac{0}{2} * \log_2\left(\frac{0}{2}\right)\right) + \left(-\frac{2}{2} * \log_2\left(\frac{2}{2}\right)\right)$$

$$= 0$$

After obtaining the entropy value for each criterion, the next step is to calculate the gain value using the following equation:

Score*gain* Test 3

$$Gain(Test\ 4, test\ 3) = Entrophy(Test\ 4) - \sum_{i=n}^{n} \frac{|Test\ 3_i|}{|Test\ 4|} * Entrophy\ (Test\ 3_i)$$

$$Gain(Test\ 4, test\ 3) = 0{,}970950594 - \left(\left(\frac{3}{5} * 0\right) + \left(\frac{2}{5} * 0{,}970959594\right)\right)$$

$$= 0,,970950594$$

Based on table 1, it can be seen that the last criterion is test 3 with a gain value of 0.970950594. There are two attribute values from test 3, namely HIGH and MEDIUM. Of the two attribute values, the HIGH attribute value has classified the case into 1, i.e. the decision is HIGH, so there is no need for further calculations, while the LOW attribute value has also classified the case into 1, i.e. the decision is MEDIUM, so there is also no need for further calculations. carry on.

From these results, a decision tree can be drawn as follows:

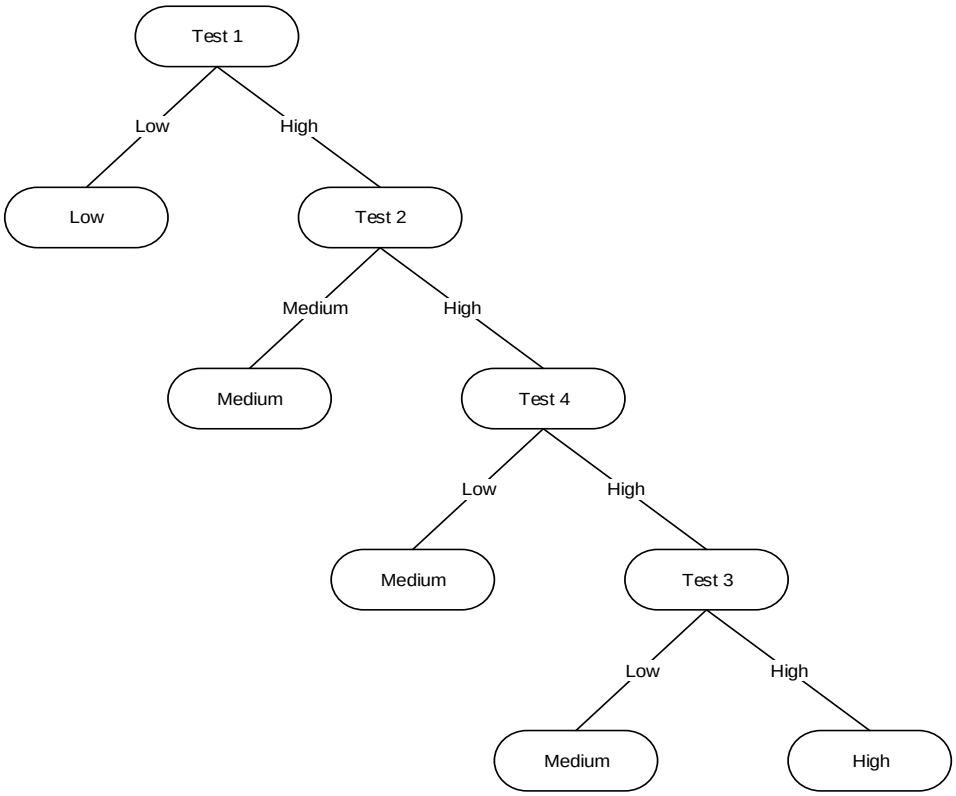

Fig. 2 Decision Tree Calculation Results Node 4

Figure 2 above is an image of the last decision tree that was formed because there are no more criteria that can be used as the next branch, and all the attributes in the last branch, namely test criteria 4 have classified cases into 1.

1. IF test 1 = low THEN decision = low
2. IF test 1 = high AND test 2 = low THEN decision = moderate
3. IF test 1 = high AND test 2 = high AND test 3 = low THEN decision = moderate
4. IF test 1 = high AND test 2 high AND test 3 high AND test 4 high THEN decision = high
5. IF test = 1 high AND test 2 = high AND test 3 = high and test 4 = low THEN decision = moderate

4. Conclusion

Based on the rule above, it can be concluded that respondents who get high results on test 1, test 2, test 4, and 3 have a high subconscious tendency. Meanwhile, respondents who get a low test result 1 have a low subconscious tendency. Then respondents with high results on test 1, test 2, and test 4, and low test 3 have a moderate subconscious tendency. Then respondents with high test results 1 and low test 2 have a moderate subconscious tendency. Then respondents with high results on test 1 and test 2, and low test 4 have a moderate subconscious tendency.

References

Abdullah, A. Z., Winarno, B., & Saputro, D. R. S. (2021, February). The decision tree classification with C4. 5 and C5. 0 algorithm based on R to detect case fatality rate of dengue hemorrhagic fever in Indonesia. In *Journal of Physics: Conference Series (Vol. 1776, No. 1, p. 012040)*. IOP Publishing.

Dai, W., & Ji, W. (2014). A mapreduce implementation of C4.5 decision tree algorithm. *International Journal of Database Theory and Application, 7*(1), 49–60. https://doi.org/10.14257/ijdta.2014.7.1.05

Dias, J. L., Sott, M. K., Ferrão, C. C., Furtado, J. C., & Moraes, J. A. R. (2021). Data mining and knowledge discovery in databases for urban solid waste management: A scientific literature review. *Waste Management & Research*, *39*(11), 1331-1340.

Florence, A. M., & Savithri, R. (2013). Talent knowledge acquisition using C4. 5 classification algorithm. *International Journal Of Emerging Technologies in Computational and Applied Sciences (IJETCAS)*.

Hosseini, S., & Sardo, S. R. (2021). Data mining tools-a case study for network intrusion detection. *Multimedia Tools and Applications*, *80*(4), 4999-5019.

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications*, *4*(2), 13-19.

Lakshmi, B. N., & Raghunandhan, G. H. (2011, February). A conceptual overview of data mining. In *2011 National Conference on Innovations in Emerging Technology (pp. 27-32)*. IEEE.

Mambang, M., & Marleny, F. D. (2015). Prediksi Calon Mahasiswa Baru Mengunakan Metode Klasifikasi Decision Tree. *CSRID (Computer Science Research and Its Development Journal)*, *7*(1), 48-56.

Mijwil, M. M., & Abttan, R. A. (2021). Utilizing the genetic algorithm to pruning the C4. 5 decision tree algorithm. *Asian Journal of Applied Sciences*, *9*(1).

Nursela, D. A. (2014). *Penerapan Algoritma C4. 5 untuk Klasifikasi Tingkat Keganasan Kanker Payudara*. Skripsi Teknik Informatika S, 1.

Permana, B. A. C., Ahmad, R., Bahtiar, H., Sudianto, A., & Gunawan, I. (2021, April). Classification of diabetes disease using decision tree algorithm (C4. 5). In *Journal of Physics: Conference Series (Vol. 1869, No. 1, p. 012082)*. IOP Publishing.

Sinaga, T. H., Wanto, A., Gunawan, I., Sumarno, S., & Nasution, Z. M. (2021). Implementation of Data Mining Using C4. 5 Algorithm on Customer Satisfaction in Tirta Lihou PDAM. *Journal of Computer Networks, Architecture and High Performance Computing*, *3*(1), 9-20.

Tempola, F., Muhammad, M., Maswara, A. K., & Rosihan, R. (2022). Rule Formation Application based on C4. 5 Algorithm for Household Electricity Usage Prediction. *Trends in Sciences*, *1*9(3), 2167-2167.

Wang, J. (2022). Application of C4. 5 Decision Tree Algorithm for Evaluating the College Music Education. *Mobile Information Systems*, 2022.

Wang, H. B., & Gao, Y. J. (2021). Research on C4. 5 algorithm improvement strategy based on MapReduce. *Procedia Computer Science*, *183*, 160-165.

Yuliansyah, H., Imaniati, R. A. P., Wirasto, A., & Wibowo, M. (2021). Predicting Students Graduate on Time Using C4. 5 Algorithm. *Journal of Information Systems Engineering and Business Intelligence*, *7*(1), 67-73.