Jurnal Mantik, 6 (2) (2022) 2228-2235

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

Student Graduation Value Analysis Based On External Factors With C4.5 Algorithm

Fristi Riandari[1], Hengki Tamando Sihotang[2], Rohit Gautama[3], Sethu Ramen[4]

[1,2,3,4] STMIK Pelita Nusantara, Jl. Iskandar Muda No. 1 Medan, Sumatera Utara, 20154

E-mail: fristy.rianda@ymail.com

ARTICLE INFO

Article history:
Received: Jul 15, 2022
Revised: Jul 30, 2022
Accepted: Aug 08, 2022

Keywords:
Data Mining;
Decision Tree;
C4.5 Algorithm.

ABSTRACT

Data mining is the process of extracting data into information that has not previously been conveyed, with the right techniques the data mining process will provide optimal results. Data Mining is divided into several methods. Data classification is a process of finding the same properties in a set of objects in a database and classifying them into different classes according to the defined classification model. The purpose of classification is to find a model from the training set that distinguishes attributes into the appropriate category or class, the model is then used to classify attributes whose class has not been previously known. The classification technique is divided into several techniques, one of which is the Decision Tree. One of the existing approaches in the classification technique is the C4.5 algorithm. The C4.5 algorithm is an approach in data mining classification techniques that can predict students' final grades. The variables used in analyzing the passing grades will be classified based on their attributes. The C4.5 algorithm with the decision tree method can provide predictive rule information to describe the process associated with analyzing student passing grades. The characteristics of the classified data can be obtained clearly, both in the form of a decision tree structure and rules so that in the testing phase the RapidMiner software can help predict student passing grades. With the formation of rules that can become new information that can be used as a tool in analyzing student passing grades.

1. Introduction

Data mining is an iterative process and requires human interaction in the process to find new patterns or models that can be generalized for the future, and are useful if used to perform an action [1]. There are several methods that can be used in data mining, namely Association Rule, Apriori, Roughset, K-Means, C4.5 Algorithm and others. Data mining, often also called knowledge discovery in database (KDD), is an activity that includes collecting, using historical and data to find regularities, patterns or relationships in large data sets [2].

The C4.5 algorithm is one of the case-solving solutions that is often used to make decision trees in solving problems in classification techniques that have characteristics, namely the process of determining the entropy value and gain value. A study explains that the C4.5 algorithm is a classification technique using entropy and information gain as a separator in the decision tree [3]. A study comparing the C4.5 algorithm and the CART algorithm in classifying student scores explains that the C4.5 algorithm has a higher accuracy value of 85.61% while the CART algorithm has an accuracy value of 84.95% [4]. In other studies, data mining is implemented to measure student study period and obtained test results which explain that the error rate in measuring student study period is only 5% [5].

STMIK Pelita Nusantara is an educational foundation that always follows the development of knowledge and technology that can assist in making policies that support success, one of which is in the student learning process. The success of student learning which has been identical to the value obtained can be the cause of the student's enthusiasm for learning to decline so that he does not graduate on time due to having to repeat

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

the failed grades. The success of student learning does not only come from internal factors but can also be caused by external factors such as type of class, age, study time, failure in the previous class, school support, family support, additional guidance, motivation for college, after-school activities, absenteeism, grades, student problems. This of course can be overcome by making an approach in the form of predicting student graduation scores based on external factors so that for students who are predicted to have failed grades, universities can make policies that can overcome or minimize this situation.

Based on the problems faced, an approach will be carried out with a prediction technique using data mining with the C4.5 algorithm to analyze student graduation scores. The sample used in this research is active students who are sitting in semester VI. This study aims to produce a rule in the form of a decision tree so that information about the passing grade will be obtained based on external factors and is expected to be a tool for educational institutions to be able to make early policies regarding this problem.

2. Method

2.1 Knowledge Discovery in Database (KDD)

Apart from the understanding that has been explained in the background, data mining is also called Knowledge Discovery in Database (KDD) is defined as the extraction of implicit and unknown potential information from a set of data. The Knowlegde Discovery in Database process involves the results of the data mining process (the process of extracting the tendency of a data pattern), then converting the results accurately into easy information [10].

2.2 Data Mining

Data mining is the process of extracting data into information that has not previously been conveyed, with the right techniques the data mining process will provide optimal results [12].

2.3 Classification

Data classification is a process of finding the same properties in a set of objects in a database and classifying them into different classes according to a defined classification model. The purpose of classification is to find a model from the training set that distinguishes attributes into the appropriate category or class, the model is then used to classify attributes whose class has not been previously known. The classification technique is divided into several techniques, one of which is a tree [13]. There are also those who explain that classification is the process of finding a set of models that describe and differentiate data classes. The purpose of classification is that the resulting model can be used to predict the class of data that does not have a class label. If given a set of data consisting of several features and classes, then classification is to find a model of that class as a function of other features [14].

2.4 C4.5 Algorithm

The C4.5 algorithm or decision tree is a commonly used method for classifying data mining. As previously explained, classification is a technique of finding a collection of patterns or functions that describe and separate data classes from one another to declare the object to be in a certain category by looking at the behavior and attributes of the defined group. This method is popular because it is able to classify as well as show the relationship between attributes. Many algorithms can be used to build a decision tree, one of which is the C4.5 algorithm. The C4.5 algorithm can handle both numeric and discrete data. The C4.5 algorithm uses the gain ratio. Before calculating the gain ratio, it is necessary to calculate the information value in bits from a collection of objects, using the concept of entropy [15].

3 Result and Discussion

3.1 Problem analysis

Analysis of student graduation scores is an activity that needs to be carried out by universities to be able to find out the graduation scores of students as early as possible and if it is taken into account that there are values that can interfere with the student administration process, it can be quickly addressed by making supportive policies as a form of service to students. student. In order for these goals to work efficiently and appropriately, suggestions are needed for an approach that can process past data, one of which is data mining. Data mining will be used as a data processing tool that can be used as a reference in decision making.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

3.2 Data Mining Analysis in Analysis of Student Graduation Scores

The study used qualitative data from 100 students. After carrying out the stages in the KDD, the final data that will be used as test data is obtained, namely as many as 16 student data with different patterns. From this, it is obtained variables that become indicators in analyzing student graduation scores based on external factors such as the Social Environment, Family Environment, Organizational Activeness and Work Status. This study will provide information on external factors that affect the value obtained by students, which is the input in this study is the condition experienced by students based on external factors that have been determined and the achievement of the value obtained and the resulting output is external factors that affect the value student.

*Output*or the results in this study are divided into two categories, namely Passed and Failed. Based on the resulting output, data mining with classification techniques is used because in this technique there are categorical variables that will produce information:

a. *Data Selection*namely combining data from several sources.
b. *Data Cleaning*i.e. to get rid of inconsistent data and noise.
c. *Data Transformation*namely transforming data into a form suitable for mining.
d. *Data Mining*i.e. essential process in which intelligent methods are used to extract data patterns, then Pattern evolution to identify really interesting patterns which represent knowledge based on some interesting actions.
e. *Graphical User Interface (GUI)*that is for End Users.

3.3 Application of C4.5 . Method

The design of the system that will be used in determining the external factors that affect the achievement of student scores. The initial data that will be divided by class is in the form of numeric or non-numeric to facilitate further analysis. After all the data to be processed is divided by class, then the classification process will be carried out by forming a decision tree as the output.

The decision-making process to analyze student scores based on external factors is as follows:

a. Social environment
b. Family environment
c. Campus environment
d. Economic Status
e. Employment status

The categories that will be decided are Passed and Failed.

3.4 Process Algorithm C4.5

As explained above, the number of samples used is 16 and the variables used as external factors are 5 variables, so the final data format will be obtained as table 1.

TABLE 1
Test Data

No	Social environment	Family environment	Campus environment	Economic Status	Status Work	Decision
1	Well	Enough	Low	Not enough	Tall	Graduated
2	Enough	Well	Well	Well	Low	Graduated
3	Enough	Enough	Well	Not enough	Low	Not pass
4	Enough	Enough	Low	Well	Low	Graduated
5	Enough	Well	Well	Not enough	Tall	Graduated
6	Well	Low	Well	Well	Tall	Graduated
7	Enough	Low	Low	Not enough	Low	Not pass
8	Enough	Well	Well	Not enough	Low	Graduated
9	Low	Enough	Low	Well	Low	Not pass
10	Low	Well	Well	Not enough	Tall	Not pass
11	Enough	Low	Well	Not enough	Low	Not pass
12	Low	Well	Low	Not enough	Tall	Not pass
13	Enough	Enough	Low	Not enough	Tall	Graduated
14	Low	Well	Low	Well	Low	Not pass
15	Enough	Well	Low	Well	Tall	Graduated
16	Enough	Enough	Well	Well	Low	Not pass

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

From the attributes that have been grouped or classified, the final data format is obtained in table 5.1 above, for example data on the Social Environment whose assessment is divided into 3, namely Good, Enough and Low as well as other attributes.

3.5 C4.5 . Data Classification Algorithm

From the student test score data in table 5.1 above, the C4.5 algorithm data classification will be carried out by making a decision tree. The cases listed in table 5.1 will make a decision tree to identify external factors that affect the achievement of student passing grades. To select an attribute as the root, it is based on the highest gain value of the existing attributes.

In making a decision tree, what must be done is to count the number of cases, the number of cases for the decision "Pass" (S1), the number of cases for the decision "Not Passed" (S2) and cases that are divided based on the attributes of the Social Environment, Family Environment, Campus Environment, Economic Status, Employment Status, then the calculation of the gain for each attribute is carried out. The steps for making a decision tree are:

Specifies the attribute as the root and calculates the value of the attribute gain information.

To select an attribute as the root, it is based on the highest gain value of the existing attributes. It takes the Entropy value to determine the highest gain.

Calculating the Entropy Value of each attribute:

***Entropy*(Total)** with the following formula:

Entropy(total) = log2 Pi− $\sum_{i=1}^{n} -\mathrm{P}\,i \times$ (1)

$$Entropy(total) = \left(-\frac{8}{16} * log_2\left(\frac{8}{16}\right)\right) + \left(-\frac{8}{16} * log_2\left(\frac{8}{16}\right)\right) = 1$$

Entropy(total) is to calculate the total score of Passed (8) and Failed (8) decisions, while 16 is the total number of cases.

1. Social Environment Attributes

   To calculate the entropy of the Social Environment, it can be seen from table 5.1, the value of the Passed Social Environment which has a value ofboth cases amounted to 2 cases and the Disqualified Social Environment which had a good score was 0 cases with a total of 2 cases, here are the entropy of each case:

$$Entropy(B) = \left(-\frac{2}{2} * log_2\left(\frac{2}{2}\right)\right) + \left(-\frac{0}{2} * log_2\left(\frac{0}{2}\right)\right) = 0$$

$$Entropy(C) = \left(-\frac{6}{10} * log_2\left(\frac{6}{10}\right)\right) + \left(-\frac{4}{10} * log_2\left(\frac{4}{10}\right)\right) = 0{,}97095$$

$$Entropy(R) = \left(-\frac{0}{4} * log_2\left(\frac{0}{4}\right)\right) + \left(-\frac{4}{4} * log_2\left(\frac{4}{4}\right)\right) = 0$$

   For the case of other variables do the same.

2. Family Environment Attributes

$$Entropy(B) = \left(-\frac{4}{7} * log_2\left(\frac{4}{7}\right)\right) + \left(-\frac{3}{7} * log_2\left(\frac{3}{7}\right)\right) = 0.98523$$

$$Entropy(C) = \left(-\frac{3}{6} * log_2\left(\frac{3}{6}\right)\right) + \left(-\frac{3}{6} * log_2\left(\frac{3}{6}\right)\right) = 1$$

$$Entropy(R) = \left(-\frac{1}{3} * log_2\left(\frac{1}{3}\right)\right) + \left(-\frac{2}{3} * log_2\left(\frac{2}{3}\right)\right) = 0.9183$$

3. Campus Environment Attributes

$$Entropy(K) = \left(-\frac{4}{8} * log_2\left(\frac{4}{8}\right)\right) + \left(-\frac{4}{8} * log_2\left(\frac{4}{8}\right)\right) = 1$$

$$Entropy(B) = \left(-\frac{4}{8} * log_2\left(\frac{4}{8}\right)\right) + \left(-\frac{4}{8} * log_2\left(\frac{4}{8}\right)\right) = 1$$

4. Economic Status Attribute

$$Entropy(T) = \left(-\frac{4}{7} * log_2\left(\frac{4}{7}\right)\right) + \left(-\frac{3}{7} * log_2\left(\frac{3}{7}\right)\right) = 0.98523$$

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

$$Entropy(R) = \left(-\frac{4}{9} * log_2\left(\frac{4}{9}\right)\right) + \left(-\frac{5}{9} * log_2\left(\frac{5}{9}\right)\right) = 0.99108$$

5. Work Status Attribute

$$Entropy(R) = \left(-\frac{3}{9} * log_2\left(\frac{3}{9}\right)\right) + \left(-\frac{6}{9} * log_2\left(\frac{6}{9}\right)\right) = 0.9183$$

$$Entropy(T) = \left(-\frac{5}{7} * log_2\left(\frac{5}{7}\right)\right) + \left(-\frac{2}{7} * log_2\left(\frac{2}{7}\right)\right) = 0.86312$$

Calculate the gain value of each attribute:

1. *Gain*(Total, Social Environment)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\text{Social Environment}_i|}{|Total|} * Entropy(\text{Social Environment}_i)$$

$$= 1 - \left(\left(\frac{2}{16} * 0\right) + \left(\frac{10}{16} * 0.97095\right) + \left(\frac{4}{16} * 0\right)\right) = 0.39316$$

2. *Gain*(Total, Family Environment)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\text{Family Environment}_i|}{|Total|} * Entropy(\text{Family Environment}_i)$$

$$= 1 - \left(\left(\frac{7}{16} * 0.98523\right) + \left(\frac{6}{16} * 1\right) + \left(\frac{3}{16} * 0.9183\right)\right) = 0.02178$$

3. *Gain*(Total, Campus Environment)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\text{Campus Environment}_i|}{|Total|} * Entropy(\text{Campus Environment}_i)$$

$$= 1 - \left(\left(\frac{8}{16} * 1\right) + \left(\frac{8}{16} * 1\right)\right) = 0$$

4. *Gain*(Total, Economic Status)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\text{Economic Status}_i|}{|Total|} * Entropy(\text{Economic Status}_i)$$

$$= 1 - \left(\left(\frac{7}{16} * 0.98523\right) + \left(\frac{9}{16} * 0.99108\right)\right) = 0.01148$$

5. *Gain*(Total, Working Status)

$$= Entropy(S) - \sum_{1=1}^{n} \frac{|\text{Working Status}_i|}{|Total|} * Entropy(\text{Working Status}_i)$$

$$= 1 - \left(\left(\frac{9}{16} * 0.9183\right) + \left(\frac{7}{16} * 0.86312\right)\right) = 0.10584$$

After all the entropy and gain values of each attribute are calculated, then the results of these calculations are entered into table 2.

TABLE 2
Node 1 . Calculation

Node	Number of Cases(S)		Graduated (S1)	Not Passed (S2)	Entropy	Gain
1	Total	16	8	8	1	
	Social environment					0.39316
	Well	2	2	0	0	
	Enough	10	6	4	0.97095	
	Low	4	0	4	0	
	Family environment					0.02178
	Well	7	4	3	0.98523	

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Enough	6	3	3	1	
Low	3	1	2	0.9183	
Campus environment					0
Low	8	4	4	1	
Tall	8	4	4	1	
Economic Status					0.01148
Tall	7	4	3	0.98523	
Low	9	4	5	0.99108	
Employment status					0.10584
Low	9	3	6	0.9183	
Tall	7	5	2	0.86312	

From the calculations in the table5.2 it can be seen that the attribute with the highest gain is the Social Environment with a value of 0.39316, then the attribute is used as the root node, where the low attribute value can be said to be Disqualified. However, the Sufficient attribute value still needs to be calculated again, the decision tree from the results of node 1 can be seen in Figure 1.

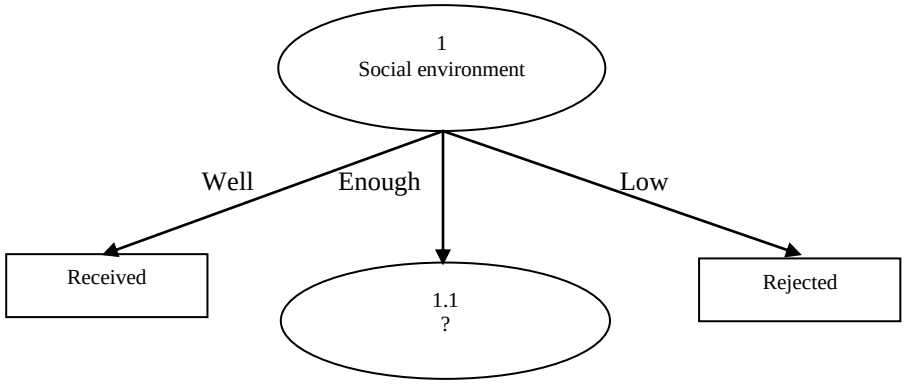

Figure 1. Decision Tree Calculation Results Node 1

Next, it will be solved to calculate Node 1.1 as the root, the same way as above by calculating the entropy value of the remaining attributes, namely Family Environment, Campus Environment, Social Status, and Work Status after entropy is calculated, then calculates the gain for each attribute. . And the final result is obtained as follows:

Table 3.
Node Calculation 1.2

Node			Accepted(S1)	Rejected(S2)	Entropy	Gain
1.2	Vocational Test Enough	4	2	2	1	
	Parent's Income					1
	Low	2	2	0	0	
	Tall	2	0	2	0	
	Dependent Parents					0
	Tall	2	1	1	1	
	Low	2	1	1	1	
	Employment status					0.31128
	Low	3	1	2	0.9183	
	Tall	1	1	0	0	

The decision tree formed can be seen in Figure 2.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

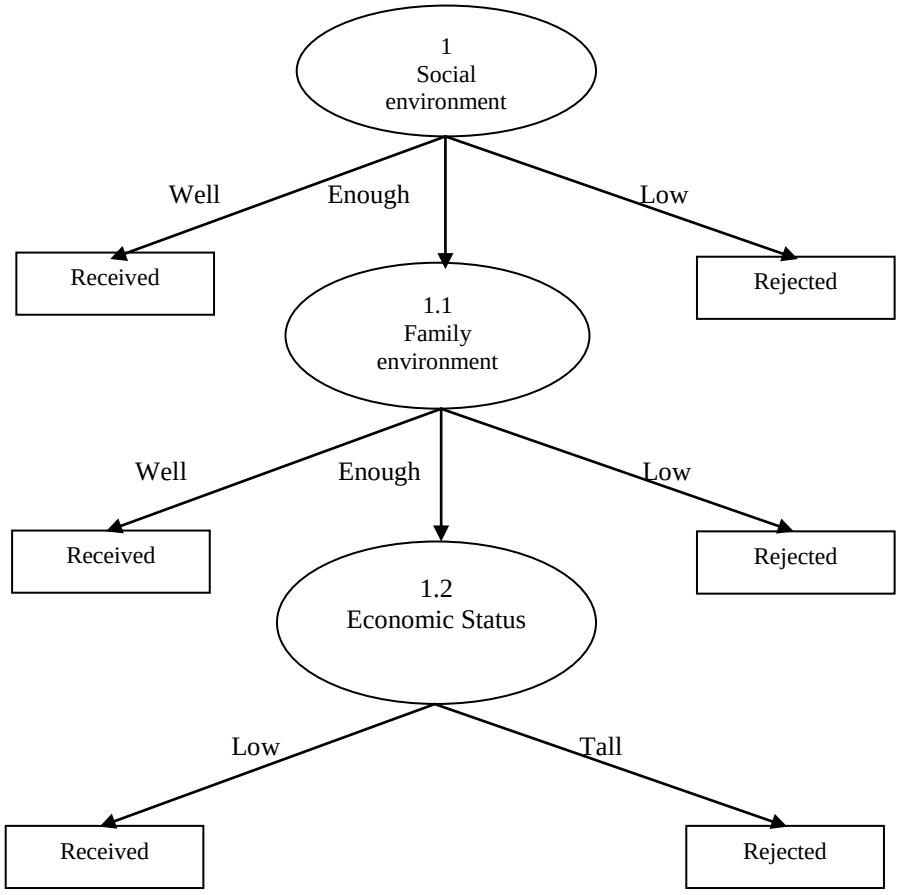

Figure 2. Decision Tree Result of Node Calculation 1.2

The rules or rules formed based on the last decision tree as shown in Figure 2 above are as follows:

a. IF Social Environment = Good THEN Decision = Accepted
b. IF Social Environment = Enough AND Family Environment = Good THEN Decision = Accepted
c. IF Social Environment = Enough AND Family Environment = Enough AND Economic Status = Low THEN Decision = Accepted
d. IF Social Environment = Enough AND Family Environment = Enough AND Economic Status = High THEN Decision = Rejected
e. IF Social Environment = Enough AND Family Environment = Low THEN Decision = Rejected
f. IF Social Environment = Low THEN Decision =Rejected

4 Conclusion

Based on the implementation of the research stages that have been carried out, the conclusion of this study is that the C4.5 algorithm can be applied as a tool in analyzing student graduation scores based on external factors, by applying the C4.5 algorithm in analyzing student graduation scores based on external factors, it can help universities in making policies early if it is calculated that there are student graduation scores that do not meet the specified requirements, the C4.5 algorithm can be a solution in solving problems in analyzing student graduation scores based on external factors.

Reference

[1] Z. Azmi and M. Dahria, "Decision Tree Berbasis Algoritma Untuk," *Saintikom*, vol. 12, pp. 157–164, 2013.
[2] H. Santoso, I. P. Hariyadi, and Prayitno, "Data Mining Analisa Pola Pembelian Produk Dengan Menggunakan

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Metode Algoritma Apriori," *Tek. Inform. ISSN 2302-3805*, no. 1, pp. 19–24, 2016, [Online]. Available: http://ojs.amikom.ac.id/index.php/semnasteknomedia/article/download/1267/1200

[3] A. M. Florence.T; and Ms.Savithri.R, "International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS) TALENT KNOWLEDGE ACQUISITION USING C4 . 5 CLASSIFICATION ALGORITHM," *Int. J. Emerg. Technol. Comput. Appl. Sci.*, vol. 4, no. 4, pp. 406–410, 2013.

[4] I. Rahmayuni, "Perbandingan performansi algoritma c4.5 dan cart dalam klasifiksi data nilai mahasiswa prodi teknik komputer politeknik negeri padang," *Teknoif*, vol. 2, no. 1, pp. 40–46, 2014, doi: 10.1016/j.jnc.2008.09.001.

[5] E. S. Siska Haryati, Aji Sudarsono, "Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu)," *J. Media Infotama*, vol. 11, no. 2, pp. 130–138, 2015.

[6] R. P. S. Putri and I. Waspada, "Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 4, no. 1, p. 1, 2018, doi: 10.23917/khif.v4i1.5975.

[7] S. N. Hermawanti, Asriyanik, and A. A. Sunarto, "Implementasi Algoritma C4.5 untuk Prediksi Kelulusan Tepat Waktu (Studi Kasus : Program Studi Teknik Informatika)," *J. Ilm. SANTIKA*, vol. 9, no. 1, pp. 853–864, 2019, [Online]. Available: http://jurnalummi.agungprasetyo.net/index.php/santika/article/download/552/253

[8] I. Iskandar, L. Hiryanto, and J. Hendryli, "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Decision Tree C4.5 dengan Teknik Pruning," *J. Ilmu Komput. dan Sist. Inf.*, vol. 6, no. 1, p. 64, 2018, [Online]. Available: https://journal.untar.ac.id/index.php/jiksi/article/view/2599

[9] S. W. Siahaan, K. D. R. Sianipar, P. P. P. A. N. . F. I. R.H Zer, and D. Hartama, "Penerapan Algoritma C4.5 dalam Menentukan Faktor yang Dapat Meningkatkan Kemampuan Bahasa Inggris pada Mahasiswa," *J. Eksplora Inform.*, vol. 10, no. 1, pp. 59–67, 2020, doi: 10.30864/eksplora.v10i1.396.

[10] K. Tampubolon, H. Saragih, B. Reza, K. Epicentrum, A. Asosiasi, and A. Apriori, "Implementasi Data Mining Algoritma Apriori Pada Sistem Persediaan Alat-Alat Kesehatan," pp. 93–106, 2013.

[11] F. Nasari and S. Darma, "Penerapan K-Means Clustering Pada Data Penerimaan Mahasiswa Baru," *Semin. Nas. Teknol. Inf. dan Multimed. 2015*, pp. 73–78, 2015.

[12] D. W. T. Putra, "Algoritma C4.5 untuk Menentukan Tingkat Kelayakan Motor Bekas yang Akan Dijual," *J. TEKNOIF*, vol. 4, no. 1, pp. 16–22, 2016.

[13] S. Lorena, W. Zarman, and I. Hamidah, "Analisis Dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Memprediksi Masa Studi Mahasiswa Berdasarkan Data Nilai Akademik," *Prosiding Seminar Nasional Aplikasi Sains dan Teknologi (SNAST)*, no. November. pp. 263–272, 2014.

[14] A. S. Sukardi and C. Supriyanto, "Klasifikasi Spam Email Menggunakan Algoritma C4.5 Dengan Seleksi Fitur," *J. Teknol. Inf.*, vol. 10, no. 1, pp. 19–30, 2014, [Online]. Available: http://research.pps.dinus.ac.id/lib/jurnal/Vol 10.1 019-030.pdf

[15] W. Supriyanti, Kusrini, and A. Amborowati, "Perbandingan Kinerja Algoritma c4.5 Dan Naive Bayes Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa," *J. Inf. Politek. Indonusa*, vol. 1, no. 3, pp. 61–67, 2016.