http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 9, No. 2,Juni 2021 ISSN : 2302-9706

IMPLEMENTATION OF DECISION TREE FOR STUDENT DATA PROCESSING

Lestari Ramadewi[1], Volvo Sihombing[2] Masrizal[3]
[1,2,3]Informatics Management
[1,2,3]Labuhanbatu University, Rantauprapat, Indonesia

lestariramadewi05@gmail.com,volvolumbantoruan@gmail.com, masrizal120405@gmail.com

Article Info
Received 02 March 2021
Revised 15 April 2021
Accepted 01 June 2021

Abstract

SD Negeri 106184 Sekip is one of the public elementary schools in Lubuk Pakam, Deli Serdang Regency. The graduation rate is one of the main problems in an elementary school. The higher the graduation rate of students at an educational institution, the more it will raise the popularity of the school. Likewise, if there is a decrease in the graduation rate, the school will experience a decline in popularity, thus allowing a decrease in applicants / prospective students to enter the school in the following years.

This of course raises concerns on the school side. Data mining is intended to provide a solution so that the school knows the most dominant factors affecting the graduation rate. For this reason, the authors are interested in raising this problem into a thesis research entitled "Application of Data Mining for student data processing using the Decision Tree Study Case SDN 106184 Sekip method".

The purpose of this study was to determine which factors were most dominant in influencing a student's passing rate, so that in the future, it is hoped that the school will be able to increase student graduation rates in the following years.

Keywords : *Data Mining*, Decision tree, Pass rate

1. Introduction

Utilization of existing data in the information system to support action-taking activities, it is not enough to rely solely on operational data, a data analysis is needed to explore the potential of existing information. Decision makers try to take advantage of existing data warehouses to dig up useful information to help retrieve the necessary data, this encourages the emergence of new branches of knowledge to solve the problem of extracting information or patterns that are important or extracting from large amounts of data, called data mining. The use of data mining techniques is expected to provide knowledge that was previously hidden in the data warehouse so that it becomes valuable information

Decision Tree Methodis a method for determining the main factor based on the comparison between one factor to another. In determining a graduation level, of course there are various considerations that need to be considered. The factors supporting the passing rates are compared with one another. To facilitate comparison, these factors are grouped into several categories by making the percentage of each factor first. From each of these categories, the most supporting factors are determined [1], [2].

SD Negeri No.106184 Sekip is an educational institution in the city of Lubuk Pakam, which still uses traditional methods of data presentation and processing, allowing frequent errors and slow processing of student data. Recently, SD Negeri No.106184 Sekip has decreasedstudent graduation rate, especially in 2017, so that it raises several new problems that must be discussed and reviewed. It is hoped that by using data mining, namely by utilizing previous data and comparing daily scores, semester

scores, UN scores, US scores and assessments of student behavior are expected to determine which factors are the most dominant support in a student's graduation rate.

2. Literature Review

2.1 Data

Data is a collection of informationobtained from an observation, can be in the form of numbers, symbols or characteristics. According to the Webster New World Dictionary, [3] Pergertian is things known or assumed, which means that data is something that is known or considered. It is known that what has happened is a fact (evidence). Data can provide an overview of a situation or problem. Data can also be defined as a collection of information or values obtained from the observation (observation) of an object. Good data is data that can be trusted to be true (reliable), timely and covers a broad scope or can provide a comprehensive picture of a problem which is relevant data. Opinion data is an object in line with what was stated [4], [5] that "Data is a value that presents a description of an object or event (event)". Thus a conclusion can be drawn that data is an object, event, or fact that is documented by having a structured codification to describe an object, event, or fact.

2.2 Data Mining

Data mining or often referred to as knowledge discovery in database (KDD) is an activity that includes the collection, use of historical data to find regularities, patterns or relationships in large data. This data mining expenditure can be used to help make decisions in the future. The development of KDD causes the use of pattern recognition to decrease because it has become part of data mining[3]

The terms data mining and knowledge discovery in database (KDD) is often used take turns to explain the excavation process hidden information in a database big. In fact, both terms have a different concept, but it has interrelation with each other, which is where the stages throughout the Knowledge Discovery in Database (KDD) is data mining

KDD process sizes are as follows:[3]

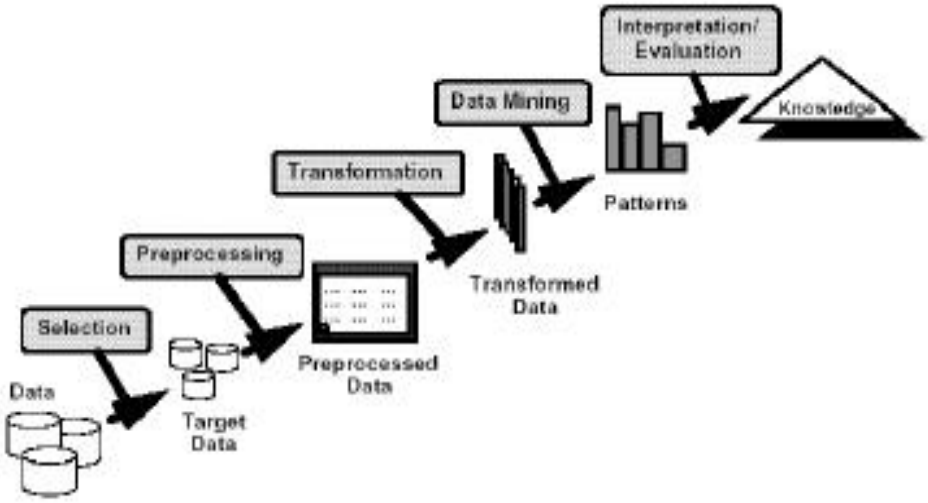

Figure 1. Knowledge Discovery stage on Data Mining (KDD)

2.3 Decision Tree

*Decision tree*is a classification method that uses a tree structure representation where each node represents an attribute, the branches represent the value of the attribute, and the leaves represent the class. The node at the top of the decision tree is known as the root. The decision tree is the most popular classification method used. Apart from being relatively fast in development, the results of the model that were built were easy to understand.
In the decision tree there are 3 types of nodes, namely:
1. Root Node, is the top node, at this node there is no input and may have no output or have more than one output.
2. Internal Node, which is a branching node, this node has only one input and has at least two outputs.
3. Leaf node or terminal node, is the end node, at this node there is only one input and no output.[7]

2.4 Algorithm C4.5

The C4.5 algorithm is a method for making a decision tree based on the training data that has been provided. Some of the developments carried out at C45 are, among others, able to overcome missing values, can overcome continuous data, and pruning [9]. Here is the basic algorithm of C4.5:
1. Build a decision tree from the training set
2. Pruning to simplify the tree.
3. Changing the resulting tree in several rules. The number of rules is equal to the number of possible paths that can be built from the root to the leaf node. [8], [10]To choose an attribute as the root, it is based on the highest gain value of the existing attributes. To calculate the gain, a formula is used as stated in formula 1 [11]:

$$\text{Gain (S, A)} = \text{Entropy (S)} - \sum_{i-}^{n} \frac{S_i}{S} \text{x Entropy ()} \quad \text{..... (1)} S_i$$

With:
S: Set of cases
A: Attribute
n: The number of partitions attribute A
[Si]: the number of cases on the ith partition
[S]: number of cases S

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

$$\text{Entropy (S)} = \text{..... ..... (2)} \sum_{i-1}^{a} -pi * log_2 pi$$

With:
S: Set of Cases
a: Features
I: The number of partitions S
Pi: The proportion of Si to S

3. Results and Discussion

The following is the data that will be used as samples for analysis and also for testing the application of data mining with the decision tree method, the data taken is student data for the last 3 years, as shown in the following table.

Table 1. Data on Student Values in 2017/2018

No.	Student's name	Daily tests	Semester Deuteronomy	UAS	US value	UN value	Behavior	Ket?

1	Adrian Nikholas	9.3	9.0	9.2	8.6	8.7	S.BAIK	GRADUATED
2	Agil Sambora	7.6	7.1	7.2	7.1	7.7	GOOD	Q. PASS
3	Alvi Nabawi	9.0	8.6	8.1	8.1	7.9	GOOD	GRADUATED
4	Dita Yolanda	5.5	6.0	6.1	6.2	6.8	K.BAIK	Q. PASS
5	Generous Dicky	9.2	8.7	9.0	8.4	8.5	GOOD	GRADUATED
6	Gio Ramadan	9.1	8.7	8.7	8.5	8.3	GOOD	GRADUATED
7	Fehby	5.4	5.4	6.5	6.2	7.1	K.BAIK	Q. PASS
8	Dita Yolanda	5.9	6.6	6.1	6.5	7.0	GOOD	GRADUATED
9	Ibrah Ramadhan	9.1	9.1	9.3	8.5	8.5	GOOD	GRADUATED
10	Jefri Kurniawan	9.3	9.3	9.0	8.8	8.6	GOOD	GRADUATED
11	Lidia Izati	7.6	7.4	7.6	7.3	7.9	K.BAIK	GRADUATED
12	M.Firza Rizky	7.3	7.6	7.8	7.2	7.6	GOOD	GRADUATED
13	Nailah Zahiyyah	9.4	9.4	9.6	8.7	8.8	S.BAIK	GRADUATED
14	Nayla Sahara	8.4	8.5	8.7	8.0	8.1	GOOD	GRADUATED
15	M.Ridho	6.7	6.5	6.5	6.5	7.0	GOOD	Q. PASS
16	M. Abid	6.3	7.5	8.0	7.1	7.7	GOOD	GRADUATED
17	Raya Paramitha	8.0	8.0	8.0	7.4	7.8	GOOD	GRADUATED
18	Rian Kurniawan	7.3	7.6	8.2	7.3	7.8	K.BAIK	GRADUATED
19	Putri Kusma	5.2	5.5	5.4	6.0	6.5	S.BAIK	GRADUATED
20	Natasya Maylani	5.8	5.4	5.8	5.9	6.7	GOOD	GRADUATED
21	Fadhlan Arifin	7.2	6.9	7.8	6.9	7.0	NOT GOOD	Q. PASS
22	Hadija	6.8	7.3	7.5	7.2	7.6	GOOD	GRADUATED
23	Meutia	8.2	9.2	8.8	8.4	8.5	S.BAIK	GRADUATED
24	Nur Alif	6.0	6.6	6.3	6.4	6.1	GOOD	GRADUATED
25	Rara Mustika	5.5	6.3	7.4	6.5	7.0	GOOD	Q. PASS

26	Sartika Putri	8.2	7.0	7.4	7.1	7.4	GOOD	Q. PASS
27	Ruri Aguslina	9.1	9.0	8.9	8.6	8.6	GOOD	GRADUATED
28	Shella Gustira	6.0	6.3	6.0	6.3	6.8	GOOD	GRADUATED
29	Sandy Pratama	9.0	8.5	8.4	8.1	8.5	GOOD	Q. PASS
30	Holy Pratiwi	6.6	6.4	6.5	6.7	7.2	GOOD	GRADUATED
31	Tamara Aulia	8.3	8.5	8.4	8.0	8.2	K.BAIK	Q. PASS
32	Sindy Aulia	7.0	6.9	6.9	7.0	7.4	S.BAIK	GRADUATED

In general, the C.45 algorithm for building a decision tree is as follows.
1. Selection of Attributes as the root
2. Create a branch for each value
3. Divide cases into branches
3. Creating a Decision Tree
5. Repeating the process for each branch until all cases on the branch have the same class.
From student data, information can be obtained as shown in the following table.

Table 2. Number of Students Based on Daily Test Values

Daily Test Score	*Total students*
<7.00	60
= 7.00	3
> 7.00	39
Total	102

Calculating all Entropy:

$$\text{Entropy (S)} = \sum_{i-1}^{a} -pi * log_2 pi$$

Entropy (Total) = (- * log2 ()) + (- * log2 ())$\frac{70}{102}\frac{70}{102}\frac{52}{102}\frac{52}{102}$

= 0.43

Entropy at the Daily Value can be calculated as follows:

Entropy (daily value, <7.00) =(- * log2 ()) + (- * log2 ())$\frac{45}{60}\frac{45}{60}\frac{15}{60}\frac{15}{60}$

= 0.13

Entropy (Daily Value, = 7.00) = (- * log2 ()) + (- * log2 ())$\frac{2}{3}\frac{2}{3}\frac{1}{3}\frac{1}{3}$

= 0.6

Entropy (Daily Value,> 7.00) =(- * log2 ()) + (- * log2 ())$\frac{22}{39}\frac{22}{39}\frac{17}{39}\frac{17}{39}$

= 0.26

Meanwhile, the value of the gain at the daily value is calculated by the following equation:

Gain (Total Daily Value) = 0.43 - ((x0.13) + (x 0.6) + (x 0.26))$\frac{60}{102}\frac{3}{102}\frac{39}{102}$

= 0.26

From the results of the table it can be seen that the attribute with the highest gain is behavior which is equal to 0.29. thus, the Behavior becomes the root node
From these results, a universal Decision Tree can be described as shown in the following figure:

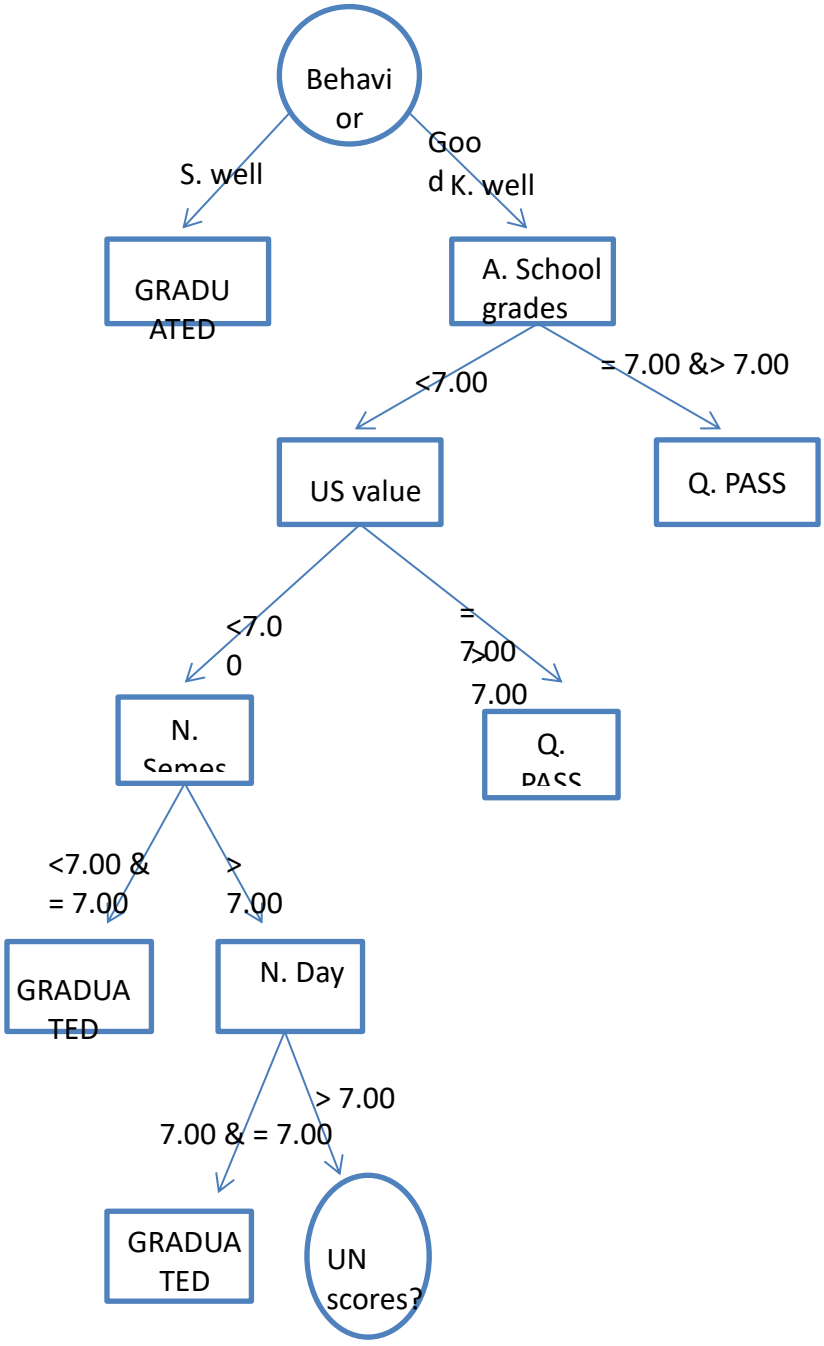

Figure 2.Node 5 Calculation Result Decision Tree

3.1. Counting Number of Students Based on National Examination Value.

Determination of the internal node for the UN value> 7.00 by counting the number of cases for Pass decisions, the number of decisions T. Passed and Entropy from all cases and cases divided by the UN Value attribute which can be the root node of the attribute value. as well as the calculation of Gain for each attribute.

Entropy value at Daily value> 7.00 can be calculated as follows:

Entropy (UN value, <7.00) =(- * log2 ()) + (- * log2 ())$\frac{12}{12}\frac{12}{12}\frac{0}{12}\frac{0}{12}$
= 0

Entropy (UN value, = 7.00) = (- * log2 ()) + (- * log2 ())$\frac{6}{6}\frac{6}{6}\frac{0}{6}\frac{0}{6}$
= 0

Entropy (UN value,> 7.00) = (- * log2 ()) + (- * log2 ())$\frac{0}{11}\frac{0}{11}\frac{11}{11}\frac{11}{11}$
= 0

Meanwhile, the Gain Value at the UN Value is calculated by the following equation:

Gain (Total UN Value) = 0.35 - ((x0) + (x 0) + (x 0))$\frac{12}{29}\frac{6}{29}\frac{11}{29}$
= 0.35

Table 3. Node Calculation 6

Node 6	Attribute	Knot	Jlh students	Student Status L	TL	ENTROPY	GAIN
		> 7.00	29	12	17	0.35	
	UN value						0.35
		<7.00	12	12	0	0	
		= 7.00	6	6	0	0	
		> 7.00	11	0	11	0	

From the table above it can be seen that the attribute UN Value with a Gain of 0.35 and all Entropy has been calculated, thus the UN Value is the last calculation.

From these results, the final Decision Tree can be described as in the following figure:

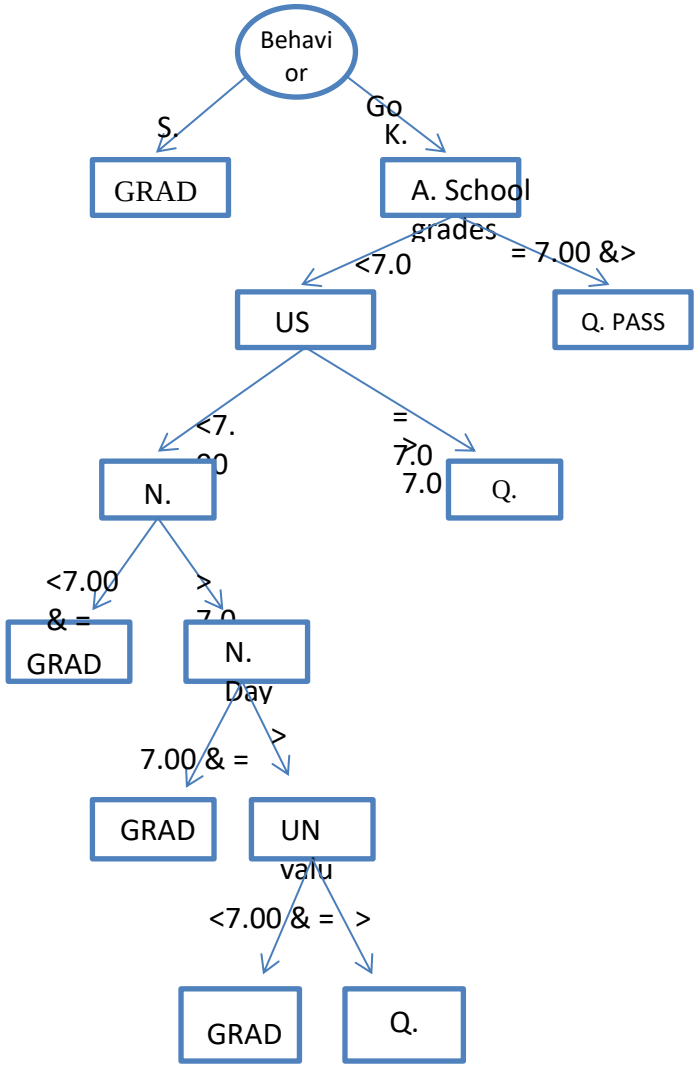

Figure 3.Node Calculation Result Decision Tree 6

In the picture, it can be seen that the first attribute is behavior which consists of assessments with very good, good and bad behavior, very good behavior has been determined to pass but bad and good behavior is developed back into school final grades, the final school grades consist of three the assessments were,> 7.00, = 7.00 and <7.00. The final school grade that has a value> 7.00 and = 7.00 has been determined not to pass but a value <7.00 will be developed back into US grades. The US grades consist of assessments with US grades <7.00, = 7.00 and <7.00, US grades = 7.00 and> 7.00 have been determined not to pass, but US grades <7.00 are developed in semester scores. The semester scores consist of semester scores with semester scores <7.00, = 7.00 and> 7.00, semester scores <7.00, = 7. 00 has been determined to pass, but a semester grade with a grade of> 7.00 is developed into the daily score. The daily scores consist of values <7.00, = 7.00 and> 7.00, the daily scores <7.00, = 7.00 have been determined to pass but the daily scores> 7.00 are developed on the UN scores. The UN scores consist of values <7.00, = 7.00 and> 7.00, the UN scores <7.00, = 7.00 have been determined to pass and the UN scores> 7.00 have been determined not to pass.

Based on the picture above, a list of rules is taken from the decision tree, namely:

1. If student behavior = Very good
   Then students Pass
2. If behavior = good and K. good

Final school grades = 7.00 and> 7.00
Then T. Pass students

3. If behavior = good and K. good
   Final school grades = <7.00
   US value = 7.00 and> 7.00
   Then T. Pass students
4. If behavior = good and K. good
   Final school grades = <7.00
   US value = <7.00
   NTSemester = <7.00 and = 7.00
   Then Students Pass
5. If behavior = good and K. good
   Final school grades = <7.00
   US value = <7.00
   NTSemester => 7.00
   Daily Value = 7.00 and <7.00
   Then Students Pass
6. If behavior = good and K. good
   Final school grades = <7.00
   US value = <7.00
   NTSemester => 7.00
   Daily Value => 7.00
   UN value => 7.00
   Then T. Students Pass
7. If behavior = good and K. good
   Final school grades = <7.00
   US value = <7.00
   NTSemester => 7.00
   Daily Value => 7.00
   UN value = 7.00 and <7.00
   Then Students Pass

Based on the list of rules above, conclusions can be drawn, namely:

1. Behavior is very influential in determining student grades and the results of the grades to be obtained by a student, so the school must further improve discipline and regulations within the school for the future.
2. The daily score has a big impact on the value of the National Examination, because with a good daily score, it will certainly reflect the readiness of a student to take the National Exam, it is hoped that the school will make extra extracurricular or additional tutoring.

4. Conclusion

In closing the discussion in the research conducted, the authors draw conclusions as well as provide suggestions to readers and those who want to re-develop the application of data mining for student data processing using the Decision Tree method.

The conclusion that the authors obtained is that the application of data mining for processing student data using the Decision Tree method is a process to produce new knowledge in the form of comparisons between the factors that affect student data, especially data on graduation rates. The results of data mining using the Decision Tree method are a sequence of activities that support each other in the student assessment process so that it is easier to understand by looking at the stages of the decision tree image.

Reference

[1] E. Elisa, "Analysis and Application of the C4.5 Algorithm in Data Mining to Identify Factors Causing PT.Arupadhatu Adisesanti Construction Accidents," *J. Online Inform.*, vol. 2, no. 1, p. 36, 2017, doi: 10.15575 / join.v2i1.71.

[2] S. Normasari, S. Kumadji, and A. Kusumawati, "The Influence of Service Quality on Customer Satisfaction, Company Image and Customer Loyalty," *J. Adm. Business*, 2013.

[3] M. Fahmi and FA Sianturi, "ANALYSIS OF APRIORIC ALGORITHM ON CONSUMER ORDERING AT CAFÉ THE L. COFFE COFFEE,"*SAINTEK (Journal of Science and Technology.*, vol. 1, no. 1, pp. 52–57, 2019.

[4] A. Mukminin and D. Riana, "Comparison of C4 Algorithms. 5, Naïve Bayes And Neural Network For Soil Classification,"*J. Inform.*, 2017.

[5] FA Sianturi, "Decision Tree Analysis in Student Data Processing," *MEANS (Media Inf. Anal. And Sist.*, vol. 3, no. 2, pp. 166–172, 2018, [Online]. Available: http://ejournal.ust.ac.id/index.php/Jurnal_Means/.

[6] J. Simarmata *et al.*, "Multimedia of number recognition for early childhood using image object," Int. J. Eng. Technol., Vol. 7, no. 3.2 Special Issue 2, pp. 796–798, 2018, doi: 10.14419 / ijet.v7i3.2.18760.

[7] YI Kurniawan, "Comparison of Naive Bayes Algorithm and C.45 in Data Classification Mining," J. Technol. Inf. And Computer Science., 2018, doi: 10.25126 / jtiik.201854803.

[8] FA Sianturi, B. Sinaga, PM Hasugian, T. Informatics, and S. Utara, "Fuzzy Multiple Attribute Decisison Macking Using Oreste Method to Determine Promotion Location," *J. Inform. Pelita Nusant.*, vol. 3, no. 1, pp. 63–68, 2018, [Online]. Available: http://e-jurnal.pelitanusantara.ac.id/index.php/JIPN/article/view/289.

[9] DS Kusumo, MA Bijaksana, and D. Darmantoro, "DATA MINING WITH APRIORIC ALGORITHM IN ORACLE RDBMS," *TECHNOLOGY - J. Researcher. and Developer. Telecomun. Control, Computer, Electr. and Electrons.*, 2016, doi: 10.25124 / tektrika.v8i1.215.

[10] EP Cynthia and E. Ismanto, "The C.45 Algorithm Decision Tree Method in Classifying Data on Sales of Fast Food Store Business," *Jurasik (Jurnal Ris. Sist. Inf. And Tek. Inform.*, vol. 3, no. July, p. 1, 2018, doi: 10.30645 / jurasik.v3i0.60.

[11] IH Witten, E. Frank, MA Hall, and CJ Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. 2016.