Prediction Competition Result of Indonesian Football Club with C.45 Algorithm

Gde Brahupadhya Subiksa[1*], Made Dinda Pradnya Pramita[2], and Komang Oka Saputra[3]
[1,2]Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University
[3]Department of Electrical and Computer Engineering, Udayana University
*Email: brahupadhya@gmail.com

***Abstract* — Indonesia is a country that is very fond of soccer sport, according to Gojek-Traveloka League data there are an average of 13,423 spectators watching every game. With a sizeable audience and a fairly regular schedule of matches by clubs and leagues in Indonesia. It makes predictions interesting because it will provide an overview and direction of support for the competing club. In some studies relating to predictions and decisions recommending the use of excessive C4.5 methods or algorithms, such as the C4.5 algorithm can produce decision trees that are easy to interpret, have an acceptable level of accuracy, efficient in dealing with discrete type attributes and can handle attribute of discrete and numeric type. So based on the explanation, this research will discuss about the prediction of Indonesian football club match using C4.5 method based on game history that has been done before. The purpose of this research is to know how far C4.5 method can be applied in predicting soccer match. The results of this study indicate that the C4.5 algorithm can be implemented in predictions of the results of the game proved with a fairly good accuracy of 63.04%.**

Keywords — C4.5 Algorithm, Data Mining, Indonesia, Prediction, Soccer.

I. INTRODUCTION

Indonesia is a country that has the biggest soccer fans in the world. According to research conducted by Nielsen Sport, 77% of Indonesians have an interest in soccer, Indonesia is second only to Nigeria with 83%. According to the data of League 1 Indonesia in 2017 there are 18 major clubs that competed in the league. Some clubs have had fans base or very big fanatic fans.

Indonesian football began in 1914 when Indonesia was still colonized by the Dutch East Indies government. Inter-city competition in Java is only in champion by two teams or in the dominance of two teams only, namely Batavia City, Soerabaja City. In 2017 the rise of Indonesian soccer carried out the Gojek-Traveloka league according to the league's data contained an average of 13,423 spectators who witnessed directly at stadiums in Indonesia [1].

Based on the data history of 2017 matches from various clubs in Indonesia, there are approximately 324 domestic or national matches and 15 international matches. Total matches held as many as 339 With a very large number of matches and solid make predictions between supporters and mass media often occur.

Predictions are predictions of events that may occur in the future or have not occurred at this time. Predictions about sports matches, especially football in Indonesia, became very interesting discussed, especially shortly before the big club game was held. Prediction becomes interesting because it will provide an overview and direction of support for the competing club. Until now there is no special method shown by the tv station or newspaper to provide a mathematical prediction with methods that can be said to be close to the truth.

In a study conducted by Muhamad Ardiansyah Sembiring in 2016 on the application of C4.5 algorithm decision tree method to predict student learning outcomes based on academic history, using the aid tools rapid miner so as to produce a decision tree. Decision tree can also be a prediction or prediction of events that have not happened based on history data that has been previously owned. The results of the study mentioned that the C4.5 method was able to generate rules to predict the achievement of learning outcomes based on previous academic history [2].

Similar research has also been conducted by Muhamad Arif Rahman in 2015 regarding Algorithm C4.5 to determine the scholarship recipients. Based on the results of this study found that the C4.5 algorithm can be applied to determine the scholarship recipients, the researchers used 40 sample data by applying the C4.5 method to produce as many as 18 students who are not eligible for scholarship because the GPA is below 3.00 and 14 eligible students become a scholarship recipient in terms of GPA, employment and tenure based on previously owned data [3].

Based on some research methods C4.5 has many advantages, such method can produce decision tree that is easy interpreted, has acceptable level of accuracy, efficient in handling discrete type attribute and can handle attribute of type of discrete and numeric [4].

Based on the above explanation, the researcher wishes to conduct research related to prediction of Indonesian soccer club match using C4.5 method based on game history which has been done before. The purpose of this research is to know how far C4.5 method can be applied in predicting soccer match.

II. STUDY LITERATURE

A. *Data Mining*

Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases [4]. Data mining is also mentioned as a series of processes to explore the added value of a data collection of knowledge that has not been known manually, is also interpreted as an automated analysis of large or complex data in order to find important patterns or trends that are usually not realized its existence The term data mining has the essence as a discipline whose sole purpose is to discover, dig, or mine the knowledge of the data or information we have. Data mining, often referred to as Knowledge Discovery in Database (KDD). KDD is an activity that includes collecting, using data, historically to find regularities, patterns or relationships in large data sets [5]. The stages of data mining are as follows:

1. *Data Cleaning*

In general, the data obtained, either from the database of a company or the results of experiments, has the stuff that is not perfect such as missing data, invalid data or also just a typo. In addition, there are also data attributes that are not relevant to the data mining hypothesis we have. Irrelevant data is also better discarded because of its existence can reduce the quality or accuracy of the results of data mining later. Garbage in garbage out (only garbage that will be generated when garbage is included) is a term often used to describe this stage. Data cleaning will also affect the performance of data mining systems as the data handled will decrease in number and complexity.

2. *Data Integration*

Not infrequently the data needed for data mining not only comes from one database but also comes from several databases or text files. Data integration is performed on attributes that identify unique entities such as name attributes, product types, customer numbers etc. Data integration needs to be done carefully because errors in data integration can produce deviant and even misleading results. For example, if the integration of data by product type turns out to combine products from different categories, then there will be correlation between products that do not actually exist. In the integration of this data also needs to be done transformation and data cleaning because often the data from two different databases are not the same way of writing or even existing data in one database was not in the other database.

3. *Data Transformation*

Some data mining techniques require special data formats before they can be applied. For example some standard techniques such as association and cluster analysis can only accept categorical data input. Hence data in the form of continuous numeric numbers needs to be divided into several intervals. This process is often called binning. Here also conducted data selection required by data mining techniques used. This data transformation and selection also determines the quality of the data mining results as there are some characteristics of certain data mining techniques that depend on this stage.

4. *Application of data mining techniques*

Application of data mining techniques itself is only one part of the data mining process. There are several data mining techniques that are commonly used. We will discuss more about the techniques in the next section. It should be noted that there are times when common data mining techniques available on the market are insufficient to implement data mining in a particular field or for certain data.

5. *Evaluation of found patterns*

In this stage, the results of data mining techniques in the form of typical patterns and predictive models are evaluated to assess whether the existing hypothesis is indeed achieved. If the result is not in accordance with the hypothesis there are several alternatives that can be taken such as: making it a feedback to improve the process of data mining, try other data mining techniques more appropriate, or accept this result as an unexpected result that may be useful.

6. *Presentation patterns are found to generate action*

The last stage of data mining process is how to formulate a decision or action from the analysis result obtained. Sometimes this should involve people who do not understand data mining. Therefore the presentation of data mining results in the form of knowledge that can be understood by everyone is a necessary step in the process of data mining. In this presentation, visualization can also help communicate mini data results

B. *C4.5 Method*

Method or Algorithm C4.5 is an algorithm used to form decision tree (Decision Tree). Decision tree is a well-known method of classification and prediction. Decision tree is useful for exploiting the data, finding the hidden relationship between a number of candidate input variables with a target variable. Many algorithms can be used in decision tree formation, among others: ID3, CART, and C4.5. The C4.5 algorithm is the development of the ID3 algorithm. The process of the decision tree is to transform the data form (table) into a tree model, change the tree model to rule, and simplify rule [6].

Decision tree is a prediction model technique that can be used for classification and prediction of tasks. The decision tree uses the "divide and conquer" technique to divide the problem-finding space into a problem set [7]. The process on the decision tree is to change the shape of the data table into a model tree. Model tree will generate rule and simplified. The concept of Decision Tree can be illustrated in Figure 1.

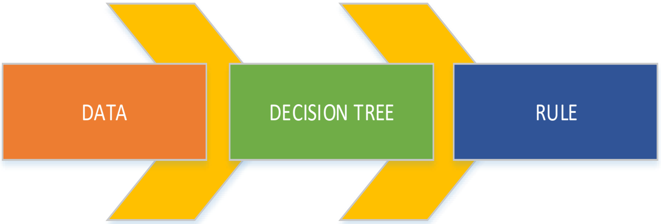


Figure 1 Concept of Decision Tree

The data in the decision tree is usually expressed in tabular form with attributes and records. The attribute states a parameter created as a criterion in the formation of a tree. In Table 1 is an example of attribute and record data in predicting the results of Indonesian football club matches.

Table 1 Here are the Attributes Involved in the Prediction of Indonesian Football Club Results.

LOCATION	FIGHT	TYPES OF RIVAL	PREVIOUS RESULTS	FINAL RESULT
T	N	N	M	K
T	N	N	M	S
T	N	N	M	S
K	N	N	M	M
T	IN	IN	M	S
K	N	N	M	M
K	IN	IN	M	M
K	N	N	M	M
ATTRIBUTE				***CLASS***

Information :
T = Away.
K = Cage
N = National
IN = International
M = Win
S = Series
K = Lose

Decision tree is one of classification technique to object or record. This technique consists of a set of decision nodes, and is connected by a branch, moving down from the root node until it ends in the leaf node. The basic concept of the decision tree in Figure 2.

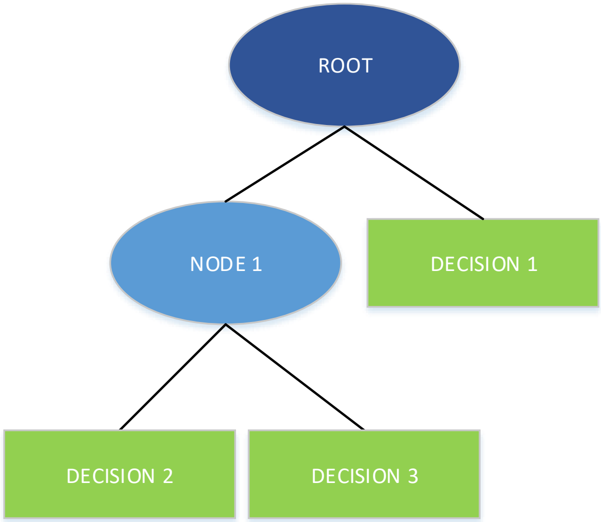


Figure 2 Basic Concepts of Decision Tree

III. Methodology

A. Data Collection

Primary data from this research is taken from the results of the match that has been implemented by several football clubs in Indonesia. Data that have not been processed and adjusted to the attributes or criteria used can be seen in Table 2. The data obtained from the website scoreboard.com and sofascore.com which collects all the results of matches ever implemented.

Table 2. Examples of Pre-arranged Data History Previously based on scoreboard.com data

PRESIDENTIAL CUP			
19.01.2018	**Bali United**	Pusamania Borneo	**3 : 2**
CHAMPION AFC FIGHT			
16.01.2018	**Bali United (Ina)**	Tampines (Sin)	**3 : 1**
1 INDONESIA FIGHT			
12.11.2017	**Bali United**	Gresik United	**3 : 0**
06.11.2017	PSM Macassar	**Bali United**	**0 : 1**
30.10.2017	**Bali United**	Sriwijaya FC	**3 : 2**
25.10.2017	**PS Barito P**	Bali United	**1 : 1**
20.10.2017	**Bali United**	PS TNI	**2 : 1**
16.10.2017	**Persiba**	Bali United	**3 : 2**
08.10.2017	**Bali United**	Arema FC	**6 : 1**
etc			

The data in Table 2 is processed and adjusted to the criteria or attributes that will be used in this research are:

1. Location of the Game about the location of the game, whether the football club as the host which means Cage or as a visiting team which means Away.
2. Types of matches or competitions attended by the football club, there are two types of matches: domestic (national) and international matches followed by Asian and international team.
3. Type of Opponent, about the opposite football club, whether the local club domestic (national) or an international club originating from another country.
4. The results of the previous game from the football club, this will concern the team's confidence, the results of the previous game will be an effect on the club's mental.

The data that have been processed according to the criteria in the study of predictive analysis of the results of Indonesian football matches using the C4.5 method can be seen in Table 3.

Table 3. Game of Indonesian Football Football Game by Prediction Attribute

LOCATION	FIGHT	TYPES OF RIVAL	PREVIOUS RESULTS	FINAL RESULT
T	N	N	M	K
T	N	N	M	S

T	N	N	M	S
K	N	N	M	M
T	IN	IN	M	S
K	N	N	M	M
K	IN	IN	M	M
Etc				

B. *Application of Algorithm C4.5*

To apply the C4.5 algorithm in a case, two main stages are needed: root search process and branch formation resulting in decision tree [8]. The process stages of C4.5 algorithm implementation can be illustrated in Figure 3.

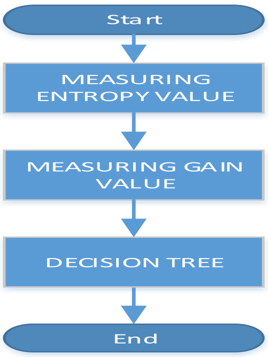


Figure 3. Flowchart Implementation Algorithm C4.5

Root search process is done first is to calculate the entropy value of each category. Then do the gain calculation to know the value of root. The highest gain value in each category will be the root value. Once the root value is specified, the category with the highest gain value (root) will be the basis for determining the establishment of Node (branch). The category with the highest gain value will be the next Node.

In the data mining known process training data, the algorithm C4.5 The initial process of training (learning) is root search. Data entered in this process comes from the weighting result. It then calculates the information gain (IG) of each attribute where the attribute that has the highest information gain (IG) value is set as root. The result of the root search process is one attribute as root. Then calculate the gain value of each attribute. The result of the process obtained value of information gain (IG) of each attribute.

IV. PEMBAHASAN

A. *Proses Pencarian Root, Node dan Leaf*

The first process is the determination of root in the decision tree, it becomes important because root is the initial decision maker that will affect the prediction in the end, so it takes the calculation to find the highest Gain value based on the history data of football club matches in Indonesia that have been processed. Before finding the value of Gain it takes the Entropy value of each attribute, the formula in calculating Entropy as follows:

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

After getting the entropy value, then proceed by calculating the gain with the following formula:

$$Gain\ (S, A) = Entropy(S) - \sum_{ve\ values(A)} \frac{|Sv|}{|S|}\ Entropy\ (Sv)$$

The results of these calculations can be seen in Table 4.

Tabel 4. Perhitungan Entropy dan Gain dalam Menentukan Root

ATTRIBUTE	SUM CASES	W	L	D	ENT	GAIN
SUM	155	63	53	39	1,55819469	
MATCH LOCATION						1,46
OPPONENT	79	16	41	22	1,47128546	
CAGE	76	47	12	17	1,33250522	
FIGHT						0,16
NATIONAL	147	61	50	36	1,552835017	
INTERNATIONAL	8	2	3	3	1,561278124	
RIVAL						0,103
NATIONAL	150	62	51	37	1,554127769	
INTERNATIONAL	5	1	2	2	1,521928095	
PREVIOUS GAME						1,83
WIN	63	25	22	16	1,561344997	
LOST	54	25	18	11	1,510280047	
SERIES	38	13	13	12	1,583954285	

Based on the calculated value of Gain in table 4 the highest score is found in the attributes of previous matches that were implemented by soccer clubs in Indonesia. So the root is the attribute of the previous game. As for the other attributes will be nodes, in searching for nodes done by the same process just remove the attributes that have become the root or node. Root prediction action can be shown in Figure 4 which is equipped with twig result is win or series or lose.

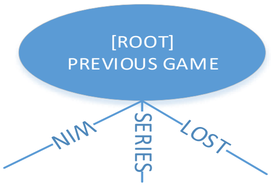
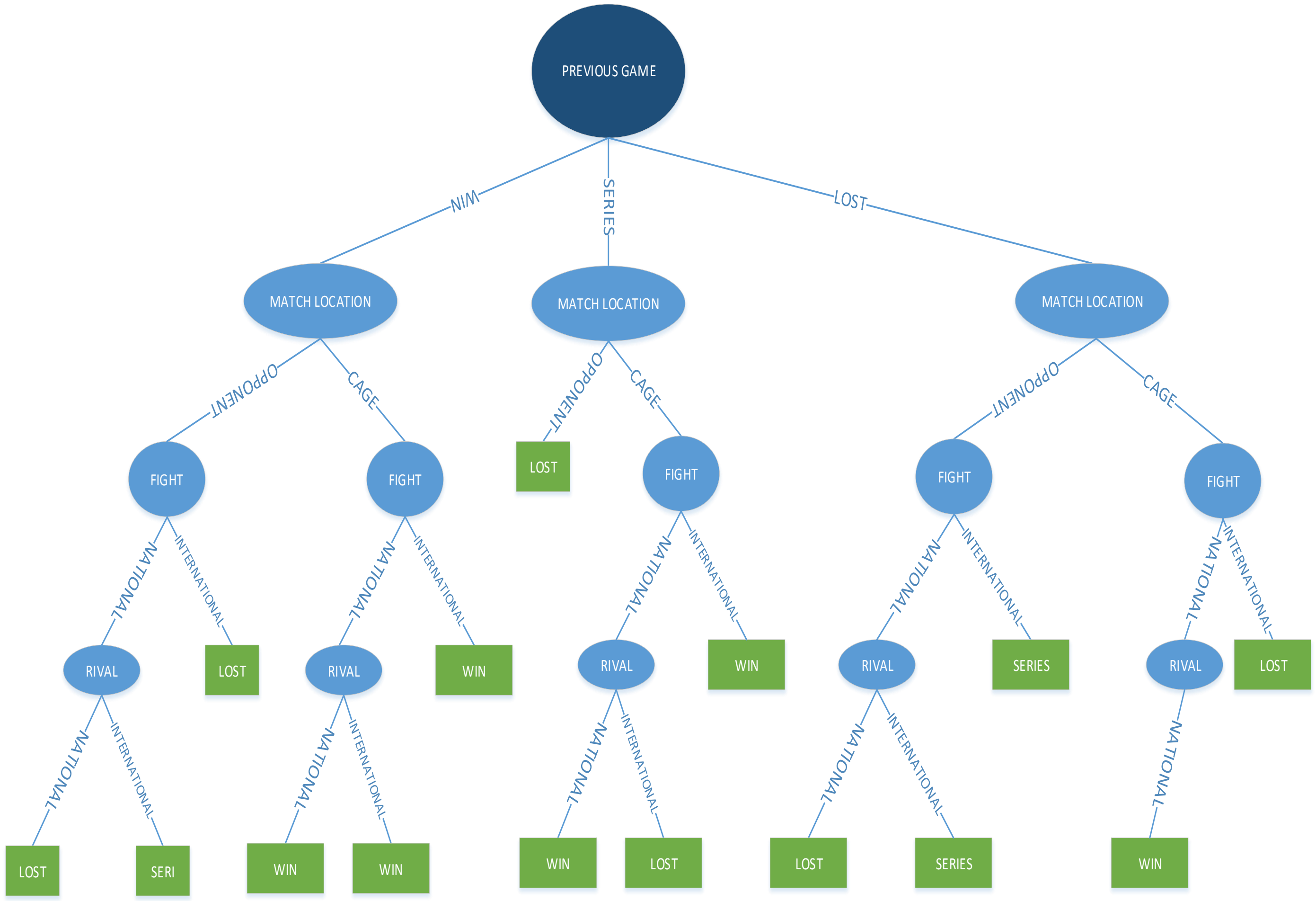


Figure 4. Previous Game Root

Next determine the specify node and leaf based on the remaining attributes, in the same way as specifying the previous root, up to the highest gain value which will arrive at leaf as the value or prediction. The decision tree is shown in figure 5.



Picture 5. Prediction Tree Based on Game History of Indonesian Football Club

B. *Testing Accuracy*

There are 339 data matches Indonesian football club ever held. Accuracy testing utilizes 184 match data of various clubs that have been implemented with the aim of calculating the accuracy of the prediction tree obtained through C4.5 algorithm. The formula used in measuring accuracy are:

$$accuracy = \frac{\text{total data valid}}{\text{total data testing}}\ x\ 100\%$$

$$accuracy = \frac{116}{184}\ x\ 100\%$$

$$\boldsymbol{accuracy = 63.04}\%$$

From the above calculation, the accuracy of the prediction tree is 63.04% which means that the decision tree has a good prediction of the matches held by soccer clubs in Indonesia. For more details can be seen in the picture below:

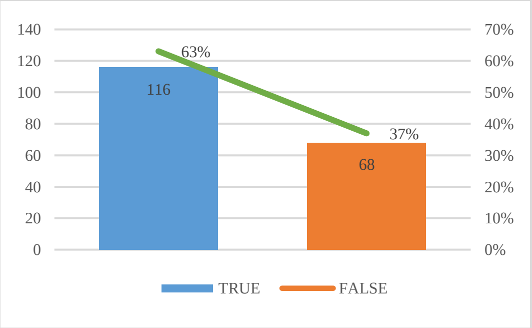


Figure 6. Accuracy of valid and invalid amount of data

If the match data is divided into two based on the type of match that is domestic or national game and international match then it can be depicted in the graph of Figure 7 and Figure 8.

In Figure 7 shows the prediction accuracy of the national game reached 63%, while in the international game the predictive accuracy rate reached 67%.

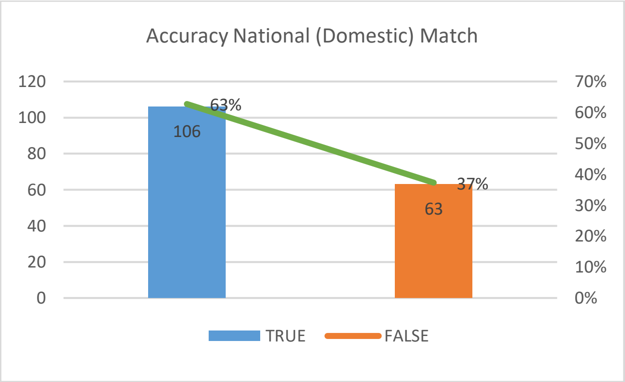


Figure 7. National Match Accuracy

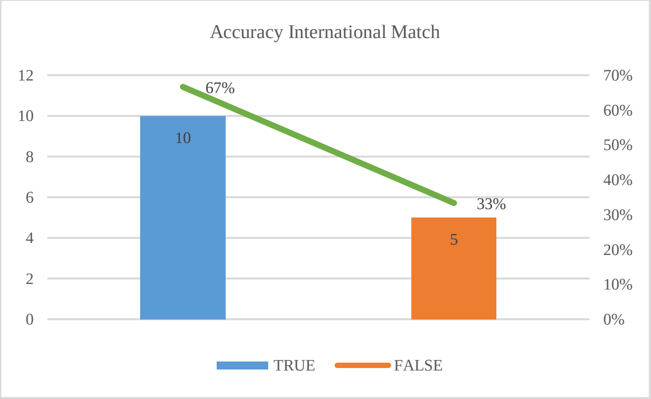


Figure 8. International Match Accuracy

Based on the above picture shows a higher degree of accuracy found in matches made at home with international match type match.

V. CONCLUSION

Based on the result of this research, decision tree with C4.5 method can be used as prediction tree with good enough. These results can be proved with an accuracy of 63.04%. Prediction tree making with C4.5 algorithm can be applied in predicting soccer club matches in Indonesia. Some of the most influential attributes to the victory of a football club in Indonesia based on the location of the Cage or Away match and the results of the previous game. Both attributes are very influential on the results of game predictions.

Suggestions for further research is the application of the decision tree into a prediction tree with different methods or data mining algorithms so as to obtain a better level of accuracy. And need more research about the attributes used as a reference prediction matches.

REFERENCE

[1] [1] Indonesian League, "Indonesian League Supporters," Official League Website Indonesia, https://liga-indonesia.id/pojok-suporter (accessed March 31, 2018)

[2] [2] K. Editor et al., "Application of the Decission Tree Algorithm Method C4.5 To Predict Student Results Based on Academic History," Jurteksi, vol. 3, no. 1, pp. 60-65, 2016.

[3] [3] M. A. Rahman, P. P. S. Iain, R. Intan, and B. Lampung, "C4.5 Algorithm for Determining Scholarship Students (Case Study: Pps Iain Raden Intan Bandar Lampung)," J. TIM Darmajaya, vol. 1, no. 2, pp. 118-128, 2015.

[4] [4] T. B. Santoso, "Analysis and Application of Method C4.5 For Customer Loyalty Prediction," CEUR Workshop Proc., Vol. 1542, no. 1, pp. 33-36, 2015.

[5] [5] D. H. Kamagi and S. Hansun, "Implementation of Data Mining with C4 Algorithm. 5 to Predict the Graduation Level of Students, "Ultim. Vol. VI, No. 1 | June 2014, vol. VI, no. 1, pp. 15-20, 2014.

[6] [6] R. H. Pambudi, B. D. Setiawan, and Indriat, "Application of C4 Algorithm. 5 In Programs To Predict High School Student Performance, "J. Pengemb. Teknol. Inf. and Computational Science., vol. 2, no. 7, pp. 2637-2643, 2017.

[7] [7] A. Putra, "The predictive solution of students dropping out on the faculty of information system of computer science faculty of university bina darma," J. SIMETRIS, vol. 8, no. 1, pp. 177-184, 2017.

[8] [8] G. A. Aryanata, P. Suta, A. Dharma, and Y. S. Sudarmojo, "Prediction of DOTA 2 Match Result by Using Analytical Hierarchy Process Method," Int. J. Eng. Emerg. Technol., Vol. 2, no. 1, pp. 22-26, 2017.

[9] [9] I. Good, A. Peling, I. N. Arnawan, I. P. A. Arthawan, and I. G. N. Janardana, "Implementation of Data Mining to Predict Period of Study Students Using Naive Bayes Algorithm," Int. J. Eng. Emerg. Technol., Vol. 2, no. 1, pp. 53-57, 2017.

[10] Score Board, "Match Data Bali United, Persija, Persib etc.," Official Score Board Website, http://www.scoreboard.com/id/tim/ (accessed April 25, 2018)