Application of Decision Tree of Classification English Premier League

Syaiful Bahri[1*], Maya Silvi Lydia[1], Rahmat Widia Sembiring[2]
[1]*Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, Medan-Indonesia*
[2]*Politeknik Negeri Medan, Medan-Indonesia*

Keyword: Football, Decision Tree, C4.5 Algorithm

Abstract: Football today has become a hobby for most people. For people who are very hobby will always be waiting for the results of the final standings of the league's favorite club and hopes to become champion. Results standings each season can be analyzed based on the attributes of the table standings. With a set of results of the standings, each season can be classified how big the club is able to become a champion by most dominant. Penelitian attributes using classification method C4.5 decision tree algorithm using software RapidMiner 8.1. The dataset used is club Manchester United's season from 1992 through 2016, the number of goals that are created, the number of goals conceded, the number of points and the status of champions or not at the end of the season. The results of the study with the algorithm C4. 5 show that the club will be able to become a champion at season's end goal number above 65 and points earned over 89 points. If the number of points at the end of the season under 65, then the club has never been champions at the end of the season.

1 INTRODUCTION

Football cannot be avoided by technology. Football nowadays always grow with the support of technology. The development of technology in predicting the fitness of players until the end of the match result is always using technology. The result of the match will be stored in the league table as a determinant champion late in the season for that position number 1. In the case of prediction with a track record of a club can be classified results of the standings every year. The classification results predictions or conclusions can be drawn in various ways, one of which is to classify what are the chances that a club can win in the league with the final results of the standings in the league a few years ago.

Decision tree whose methods are C.45 will make the classification of the results of English premier league standings from 1992 to 2016. The processed data will classify the number of goals per season, the number of goals conceded per season, the endpoints of the season and the status of champions or not at the end season will be tested with a sample club Manchester United. With the existence of the classification, it can be concluded in a number of attributes which most determine the chances of the club is able to champions in every season. To simplify the classification of data, this study uses RapidMiner 8.1 applications.

2 METHODS

2.1 Data Mining

Data mining is the process of discovering interesting knowledge, such as associations, patterns, changes, significant structures, and anomalies, from large amounts of the data stored in databases or data warehouses or other information repositories (Han, 2006). Some techniques that are often mentioned in the literature of DM include clustering, classification, association rule mining, neural networks and genetic algorithms (Lindawati, 2008). Stages of Data Mining by (Budanis 2014), Data Mining has phases - phases include:

1. Cleaning the data (data cleaning) Data cleansing is the process of eliminating noise and inconsistent data or data irrelevant.

Bahri, S., Lydia, M. and Sembiring, R.
Application of Decision Tree of Classification English Premier League.
DOI: 10.5220/0010044704520456
In *Proceedings of the 3rd International Conference of Computer, Environment, Agriculture, Social Science, Health Science, Engineering and Technology (ICEST 2018)*, pages 452-456
ISBN: 978-989-758-496-1

2. Integration of data (data integration) Data integration is the combination of data from different databases into the new database.

3. Selection of Data (Data Selection) Data contained in the database is often not all used, therefore only the appropriate data to be analyzed to be retrieved from the database.

4. The data transformation (Data Transformation Services) Data amended or merged into a format suitable for processing in Data Mining.

5. The process of mining, It is a major process when the method is applied to find valuable and hidden knowledge from data.

6. Evaluation of the pattern (pattern evaluation) To identify interesting patterns into knowledge-based alerts.

7. Presentation of knowledge (knowledge presentation) A visualization and presentation of knowledge about the methods used to obtain the knowledge acquired.

2.1.1 Classification

Classification is a data mining technique that maps the data into predefined groups or classes. It is a supervised learning method labeled roommates training requires the data to generate rules for classifying the data into predetermined test groups or classes (Dunham, 2003). The method of classification refers to the formation of groups of data by applying known algorithms to the data warehouse under examination. This method is useful for business processes that require categorical information such as marketing or sales. It can use various algorithms such as nearest neighbor, decision tree learning, and others.Decision Trees are also used to explore the data, find hidden relationships among a number of candidates for the input variables with a target variable.The decision tree combines data exploration and modeling, so it is great as a first step in the modeling process even when used as the final models of several other techniques.

2.2 Decision Tree

The decision tree is a prediction model technique that can be used for classification and prediction of tasks. The Decision Tree uses the technique of "divide and conquers" to divide the problem-finding space into a set of problems (Dunham, 2003). The process on the decision tree is to change the shape of the data table into a model tree. The model tree will generate rule and simplified (Basuki & Syarif, 2003).

The advantages of the decision tree method are:

1. The area of decision making that was previously complex and very global, can be changed to be more simple and specific.

2. Elimination calculations are not necessary because when using decision tree method the samples tested was based criteria or a particular class.

3. Flexible to choose features from the different internal nodes, feature selected will distinguish criteria other than the criteria in the same node. The flexibility of this decision tree method increases the quality of the resulting decisions than when using the method of calculating the phase of a more conventional

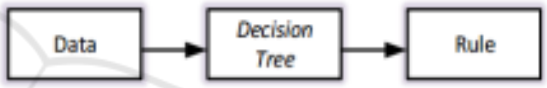

Figure 1. Decision Tree Concept

2.3 C4.5 Algorithm

There are several steps in making a decision tree algorithm C4.5, Larose, namely:

1. Prepare the training data. The training data are typically taken from historical data that never happened before or referred to the past data and is already classified in a particular class.

2. Calculate the root of the tree. The roots will be taken of the attributes to be elected, by calculating the value of the gain of each attribute, the highest gain value which will be the first roots. Before calculating the gain of attribute values, first, calculate the value of entropy. To calculate the value of entropy used the formula:

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

by:

S: The set Case

A: Features

n: number of partitions S

pi: The proportion of Si to S

4. Calculate the value of Gain using the equation.

$$Gain(S, A) = entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} Entropy\ (Si) \quad \text{..... (2)}$$

Information:
S = The set of cases
A = Features
n = number of partitions attribute A
| Si | = The proportion of Si to S
| S | = Number of cases in S

4. Repeat step 2 and step 3 until all the records partitioned.
5. The process decision tree partition will stop when:
   a. All records in the node N gets the same class.
   b. No attribute in the record is partitioned again. There are no records in the empty branch.

2.4 RapidMiner 8.1

RapidMiner is a neighborhood machine learning data mining, Text mining and predictive analytics (Han, 2006). Work performed by text mining RapidMiner is ranged with text analysis, extract patterns from large data sets and combine with statistical methods, artificial intelligence, and database. The purpose of the analysis of this text is to obtain the highest quality information from text that is processed.

RapidMiner provides procedures for data mining and machine learning, it includes ETL (extraction, transformation, loading), the data preprocessing, visualization, modeling, and evaluation. The data mining process is composed of operators nestable, described by XML, and is made with a GUI. The presentation is written in the Java programming language.

2.5 Data Preprocessing

Data cleaning is applied to add the attribute is missing or empty, and change the data inconsistent.

1. Data Transformation In this process, data is transformed into a form suitable for data mining processes.
2. Data Reduction Data reduction is done by eliminating unnecessary attributes so that the size of the database to be small, and only include the required attributes in the data mining process, as it will be more efficient on smaller data. Here is a table of data sets to be used.

Table 1. The dataset of results Preprocessing

No	Seasons	Goal Forced	Goal Allowed	Points	Champions
1	2016-17	54	29	69	No
2	2015-16	49	35	66	No
3	2014-15	62	37	70	No
4	2013-14	64	43	64	No
5	2012-13	86	43	89	Yes
6	2011-12	89	33	89	No
7	2010–11	78	37	80	Yes
8	2009–10	86	28	85	No
9	2008–09	68	24	90	Yes
10	2007–08	80	22	87	Yes
11	2006–07	83	27	89	Yes
12	2005–06	72	34	83	No
13	2004–05	58	26	77	No
14	2003–04	64	35	75	No
15	2002–03	74	34	83	Yes
16	2001–02	87	45	77	No
17	2000–01	79	31	80	Yes
18	1999–00	97	45	91	Yes
19	1998–99	80	37	79	Yes
20	1997–98	73	26	77	No
21	1996–97	76	44	75	Yes
22	1995–96	73	35	82	Yes
23	1994–95	77	28	88	No
24	1993–94	80	38	92	Yes
25	1992–93	67	31	84	Yes

Data preprocessing is important in the process of data mining. From these data still to be done preprocessing so that the data is processed in accordance with that required in the process.

3 RESULT AND DISCUSSIONS

3.1 Determining the Value Attribute

In the first set of sample data node elected, by calculating the gain value of each attribute to determine the selected node, use the value of the greatest gain information. By using the equation -P (+) log2P (+) - P (-) log2P (-).

At this stage of the study will be performed using software RapidMiner 8.1 with decision tree method. Data were processed using M.excel 2010, so in RapidMiner 8.1 application using excel read operator function to read the file to be processed. Then enter operator Decision Tree to create processed data generates decision trees.

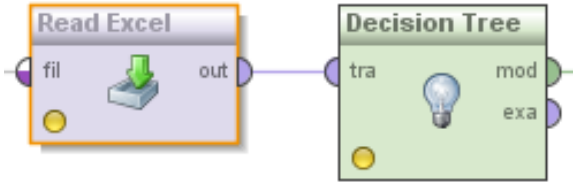

Figure 2. First Process RapidMiner

3.2 Determining Parameters

The criterion used was gain_ratio based Opera- obtained against intrinsic information which serves to

reduce the bias towards Multi-value attributes to take into account the number and size of branches when selecting attributes. On the choice of maximal depth indicates the maximum number of long branching from treetop branches.

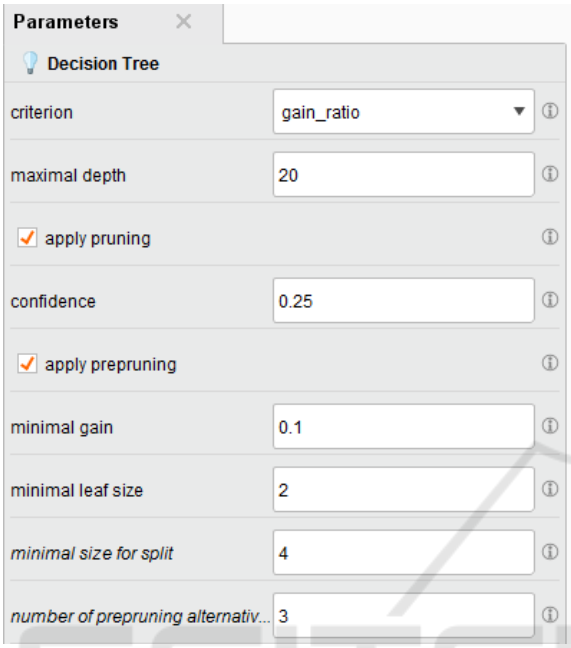

Figure 3. Setting Parameters

3.3 Example Set

The example set is a data set that has been uploaded to the application RapidMiner 8.1. Here is the data set that is ready to be processed by the application RapidMiner 8.1.

ExampleSet (25 examples, 1 special attribute, 5 regular attributes) Filter (

Row No.	Champions	No	Musim	Goal Forced	Goal Allowed	Points
4	No	4	2013-14	64	43	64
5	Yes	5	2012-13	86	43	89
6	No	6	2011-12	89	33	89
7	Yes	7	2010-11	78	37	80
8	No	8	2009-10	86	28	85
9	Yes	9	2008-09	68	24	90
10	Yes	10	2007-08	80	22	87
11	Yes	11	2006-07	83	27	89
12	No	12	2005-06	72	34	83
13	No	13	2004-05	58	26	77
14	No	14	2003-04	64	35	75
15	Yes	15	2002-03	74	34	83
16	No	16	2001-02	87	45	77
17	Yes	17	2000-01	79	31	80
18	Yes	18	1999-00	97	45	91
19	Yes	19	1998-99	80	37	79
20	No	20	1997-98	73	26	77

Figure 4. Example Set

3.4 Description

The description describes the results of the classification of the data sets in though. Results of the description can be seen in the following figure.

Tree

```
Goal Forced > 65.500
|   Points > 89.500: Yes {No=0, Yes=3}
|   Points ≤ 89.500
|   |   Goal Forced > 86.500: No {No=2, Yes=0}
|   |   Goal Forced ≤ 86.500: Yes {No=4, Yes=10}
Goal Forced ≤ 65.500: No {No=6, Yes=0}
```

Figure 5. Example set read excel

3.5 Graph View

Graph view shows the results of classification by branching can be produced conclusions. Here is the result of a decision tree.

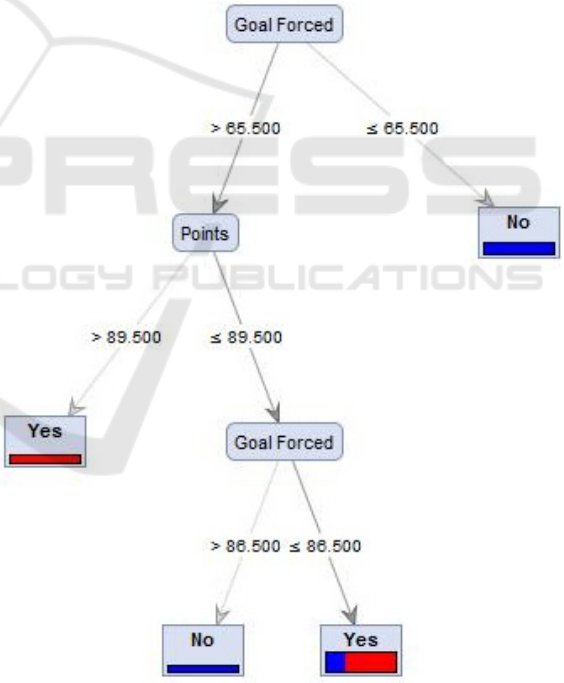

Figure 6. Decision Tree Results

4 CONCLUSIONS

1. Decision tree method can be used to classify the results of the final standings of the league to see what are the chances of a club can win the final table results of the previous season's standings.

2. The results of the data that has been processed to show that the club Manchester United will be champions if the acquisition point each season 89 or> 89. If the pointy end of the season under 65 it is ascertained that the club will not be champion.

3. It turned out that the club was also able to get the title 7 times apart from the number of points above 89, but through a number of points below 89 and averaged 86 goals per season below.

REFERENCES

Basuki, Ahmad dan Syarif, Iwan, 2003. "*Decision Tree*". Surabaya: Politeknik Electronika Negeri

Dunham, Margareth H., 2003. "*Data Mining Introductory and Advanced Topics*". New Jersey:Prentice Hall.

Han, J. and Kamber, M., (2006) Data Mining: Concepts and Techniques, Elsevier.

Tjahyono, A. dan Anggara, A. M., 2010, *Sistem Pendukung Keputusan Penerimaan Pegawai Baru pada PT. Kanasritex Semarang*, *Techno.com, Vol. 9 No.3*

Kursini, Luthfi. E. T. 2009. Algoritma Data Mining.PT Andi Offset.

Lindawati (2008), Data Mining Dengan Teknik Clustering Dalam Pengklasifikasian Data Mahasiswa Studi Kasus Prediksi Lama Studi Mahasiswa Universitas Bina Nusantara, Seminar Nasional Informatika (semnasIF 2008), ISSN :1979-2328