eJAICT
ISSN 2355 1771
JOURNAL OF APPLIED INFORMATION, COMMUNICATION AND TECHNOLOGY
Published by:
Research Center of Technology and Innovation
Academic Research and Community Services
Swiss German University

Building Data Mining Decision Tree Model for Predicting Employee Performance

Ira Mellisa

Department of Information Technology, Swiss German University, Tangerang 15143, Indonesia

Article Information

Received: 15 June 2019
Accepted: 20 September 2019
Published: 25 October 2019
DOI: 10.33555/ejaict.v6i2.79

Corresponding Author:

Ira Mellisa
Email: ira.mellisa@student.sgu.ac.id

ISSN 2355-1771

ABSTRACT

Human resource is one of the functions of a company that is considered as an asset. Therefore, the theory of performance qualification was adopted by the company in order to get an overview of employee performance. Furthermore, the company needs an effective method to predict the performance not only for the employees but also for the new applicants. The goals of this research are to get a decision tree model of the employee performance. By learning employee data, the performance of the new applicants could be predicted. The study would provide the characteristic of new applicants who will give better performance than other applicants. The data from a company in Indonesia will have been used for this research. The data mining technique will be applied to the data of operators (such as admins, clerks, cashiers, machine operators, and security officers). The data mining technique was used is decision tree. The decision tree technique was commonly used for a supervised learning data. The decision tree technique also has advantages compared others, because of its ability to produce information that is easy to understand. The result of this research shown the high dependency of employee performance with employment type (work contract). It also means that employees are encouraged to provide good performance to the company if those employees have become permanent employees. This research also showed that there is no relationship between employee performances with gender or position grade.

Keywords: *Data Mining, Prediction, Decision Tree, Employe Performance*

1. Introduction

Human resources were considered as an asset to the company. The data of human resources also considered as an issue by top management. Top management realized that human resources have an important role in the growth of the company. Information Technology (IT) has been applied by some medium-to-large scale companies in order to simplify and standardize the human resources operational processes. The simplification and standardization of human resources processes are handled by the Human Resources Information System.

Human resources management consists of several fields or activities, one of it is recruitment. Recruitment aims to get prospective applicants by doing a series of selection processes. The applicant will be selected by the company based on the suitability of the candidate profile with the criteria of a position. It also became a concern of the company during the recruitment process.

The theory of performance qualification needs to be applied by the company to determine the employee performance criteria. The performance qualification is a method for classifying employee performance based on the employee achievement in a certain period. However, the employee performance criteria of one position could be different with others. Therefore, the employee performance should be grouped by the level of position (such as operator, junior management, middle management, and top management). The approach of this analysis is based on the similarity of employee performance criteria for each position level.

Each company has different perspectives and definitions of employee performance. This research used data from a company in Indonesia that adopt the employee performance concept with five levels or categories. Those five levels are below par, par, quality, outstanding, and exceptional. Quality, outstanding, and exceptional were identified as good performance. While below par and par are shown bad performance.

In general, data mining was categorized into four models, namely *clustering, association, classification* and *prediction* (Chein and Chen, 2006). The model used as a technique to verify and validate the data of human resources. The characteristic of human resources data might be different with others because most of human resources data is qualitative or non-numerical data. The data mining techniques are usually used in the recruitment process (personnel selection). The data mining techniques are used in order to choose and find the right candidates for the company. The idea of this concept is predicting candidates performance by using the past experience knowledge from the employee database.

According to previous studies by Jantan and Puteh (2009), classification is the most suitable model of data mining for predicting employee performance. The classification technique has been used and proposed in terms of build the machine learning process (Jantan and Puteh, 2009). The classification technique is known as supervised learning, which means the class or target for a study is already known or determined. The class or target for this study is the performance result with two categories, good and bad performance.

The classification technique was chosen as data mining model for this research because of its ability to predict based on the selected class labels. The classification techniques are also aligned with the purpose of this research, which is to find the pattern of good and bad performance of employees. There are very few studies related to the prediction of employee performance using this approach. From the literature study, there are some attributes that commonly used as research materials for predicting employee performance such as age, gender, marital status, education degree, university type and length of services. Besides that, the attributes could be adjusted based on research objectives. For example, is the data of university type. This attribute is needed to predict employee performance. However, the data of university type is not necessarily needed for this research. Because this research is focused on the performance of employees in operator level. The operator level does not require a high education. It was aligned with the requirement of operator positions. Therefore, the university type will not be used as an attribute for this research.

There are many algorithms that are used for classification in data mining, such as decision tree, neural network, association rule mining, k-nearest neighbor, and many more. Among those techniques, the decision tree is the most common technique used in the previous study (Kotalwar and Chavan, 2014). The decision tree also has benefits in order to help the decision-making process because of its ability to visualize the decision into a tree structure. This approach will make management easier to understand the decision. Moreover, the attributes and dependencies of each attribute are also easily understood by studying the result of decision tree structure.

In order to produce an accurate decision tree, the data mining software is needed in this research. There is some software that can be used, one of them is WEKA. WEKA is an open source software for data mining. Data mining software is used to measure the accuracy of each experiment.

The decision tree technique can be built using some algorithm such as ID3, C4.5 (or J4.8), and Naïve Bayes. In the previous study by (Al-Radaideh and Nagi, 2012), the most accurate algorithm of the decision tree for predicting performance was identified by comparing the accuracy percentages of each algorithm. However, that step is not necessary if the accuracy percentage of one algorithm is considered high or relevant.

2. Methodology

2.1 Data Mining Concept

Data mining concept can be applied in order to get the characteristics of the good employee performance. In general, data mining can be described as a process of extracting data from database in order to obtain a pattern of knowledge. The resulting pattern itself must be explicit and easy to understand. To achieve that goal, there are several steps that need to be done. Data mining is part of the pattern search process.

Data mining consists of set of activities and techniques that can be used to extract knowledge from data. According to (Han and Kamber, 2012), following are the activities to extract knowledge from data:

a. Data cleaning
This stage is to eliminate data with extreme values that could potentially disturbing the research. Although the company has adopted a human resources information system, but that does not mean there is no data problem. Data problems can be triggered by several things, such as mistakes in data entry or errors in the system (bugs). Those data problems will generate invalid data or commonly called noise data. For example, there are some mistakes in data entry for the birth date of employees. Those data will disrupt the employee age analysis. Because it can produce a large range or spans of employee age. Therefore noise data will be deleted and not used in research.

b. Data integration
This stage is to combine data from multiple sources or databases. The data used as research material can be sourced from one or more databases. Data compilation and comparison will be performed at this stage. For example, the format of employee numbers in one database may be different to other databases. Employee numbers need to be equated first, so the data from each databases can connected.

c. Data selection
This stage is to select the relevant data for research. At this stage, the attributes that used for research material will be selected. The selection of attributes is based on research objectives. The references for attributes can be sourced from previous research or based on the needs of the company. The attributes that commonly used for predicting employee performance are age, gender, length of services, education level, and employment type (Chein and Chen, 2006).

d. Data transformation
This stage is to convert data into an appropriate format. Data can be easily analyzed if it is quantitative data. However, most of data in human resources information system is qualitative data. Therefore, the qualitative data needs to categorized and transformed into quantitative data. In the human resources information system, there is no classification based on the age of employees. Then, the age of employees need to be categorized. The categorization used in this research is showed in Section 3 (Result).

e. Data mining
The purpose of this stage is to apply algorithm to extract the pattern. Basically, extracting data is divided into two functionalities, namely descriptive and predictive. The data mining method chosen depends on the research objectives and characteristics of the research material. Some examples of the data mining methods are *clustering*, *regression*, and *classification*. Clustering is the method of data mining for grouping and aggregating data according to the similarities. Regression is the method to identify the relationship and estimate the value of the target. Classification is the method of data mining to assign attributes to categories or classes. This research used the classification method to identify employee performance according two classes, namely good and bad. Because this research used data from Company XYZ, which was categorizing employee performance into five categories. The highest and the second highest level of employee performance will be considered as good, while the others will be considered as bad.

f. Pattern evaluation:

This stage is to recognize and study the results of the pattern. If the method used is not able to produce the expected pattern, then there should be a review of the research material or algorithm.

g. Pattern presentation
This stage is to visualize the pattern. The result of the research will be concluded to be a statement. Presentation can be a table or structure that contain research results along with the level of accuracy for those methods.

2.2. Data Mining Classification Technique

Data mining classification method has several algorithms and techniques that can be used in research. One of the methods is the decision tree. The decision tree is a classification method that uses a representation of a tree structure that contains alternatives for solving problems. The decision tree also shows the factors that influence the outcome of the decision, along with its estimation.

This research used only one technique of data mining, which is decision tree. Other data mining techniques are not included in this study because the research will take longer research time. Due to the limited time of research, this research focuses only on decision tree techniques that are better known by many people.

The decision tree methods are often used in research that related to employee performance predictions. The decision tree considered as a method that capable to produce the pattern of employee performance (good or bad). The decision tree produces an illustration of trees and roots of each attribute and classes that used in the research. For example, in order to analyze the class of good performance, what attributes are the main factors or determinants of good performance.

2.3. Data Mining Tools

This research used IBM SPSS Statistics 24 for analysis, because of its ability to provide the decision tree algorithm and visualizing the tree of the chart. The software also shows the percentage of every node in the decision tree. It is also including the accuracy percentage of the research experiment. SPSS was known as one of computing software for mathematics. However, SPSS also have some algorithm for supervised learning in data mining. Supervised learning means learning data from labeled training data (Mohri, Rostamizadeh, and Talwakar, 2012). It also means the correct results or targets are known.

2.4. Research Material and Sampling Method

This study used research material from the HR system in XYZ Company. XYZ Company is a private company in Indonesia with more than 10,000 employees. For this research purposes, the data was selected is the data for employees in operator level.

The number of records of this data is 3.126 rows. Those data will be used as a dataset in research material. Afterwards, the dataset will be divided into two parts, viz. training set and test set. The training set contains 80% of the data from the original dataset, while the test set contains 20% of data from the original dataset. This method was required for supervised learning to build and evaluate the model of decision tree algorithm.

3. Result

3.1 Attributes of Research

There are nine attributes or variables for this research as shown in Table 1. Table 1 shown that employee performance result is the dependent variable, while the rest of attributes is an independent variable. The selection of attributes used is based on previous research and XYZ Company's needs. The attributes used are similar to those conducted by (Karthick & Yousuf, 2016). However, there are some attributes that are adjusted to XYZ Company's needs. For example the absence of attributes for *education specialization* and *education degree*. It because the target for this research are employees at the operator level, where the level of education of them is not up to university level.

Table 1. Attributes of Research.

No	Attributes	Description	Dependency	Type
1	Gender	Gender of employees. Classified into two categories: Male and Female	Independent	Categorical
2	Marital_Status	Marital status of employees. Classified into three categories: Single, Married, and Divorced	Independent	Categorical
3	Position_Grade	Job grade of operator levels. Classified into three categories: A, B, and C.	Independent	Categorical
4	AgeGroup	The age range of employees. Classified into five categories: Under 21, 21 to 30, 31 to 40, 41 to 50, and Above 50.	Independent	Categorical
5	Employment_Type	Type of contract. Classified into two categories: Contract and Permanent.	Independent	Categorical
6	Operator_Type	Type of operator. Classified into two categories: Direct and Indirect. This attribute is to identify whether the operator is direct or indirect to the production process.	Independent	Categorical
7	Highest_Education	Educational level. Classified into four categories: Elementary, Junior High, Senior High, and Diploma.	Independent	Categorical
8	Year_of_Service	Service range of employees. Classified into five categories: Under 5 years, 5 to 10 years, 10 to 15 years, 15 to 20 years, and more than 20 years	Independent	Categorical
9	Number_of_Children	The number represented the total of children that employee has.	Independent	Numerical
10	Employee_Performance_Result	Classified into two categories: bad and good.	Dependent	Categorical

3.2. Decision Tree

Those attributes were analyzed by using decision tree (CRT/CART) algorithm in SPSS. Figure 1 shows the result of the training set (80% data from the original dataset). The experiment used 2.522 of 3.126 rows data. The result for training set sample shown that there are eighteen nodes in decision tree chart. It was also described that not all attributes give a contribution to building the decision tree model. The first node that determined the Employee_Performance_Result is Employement_Type.

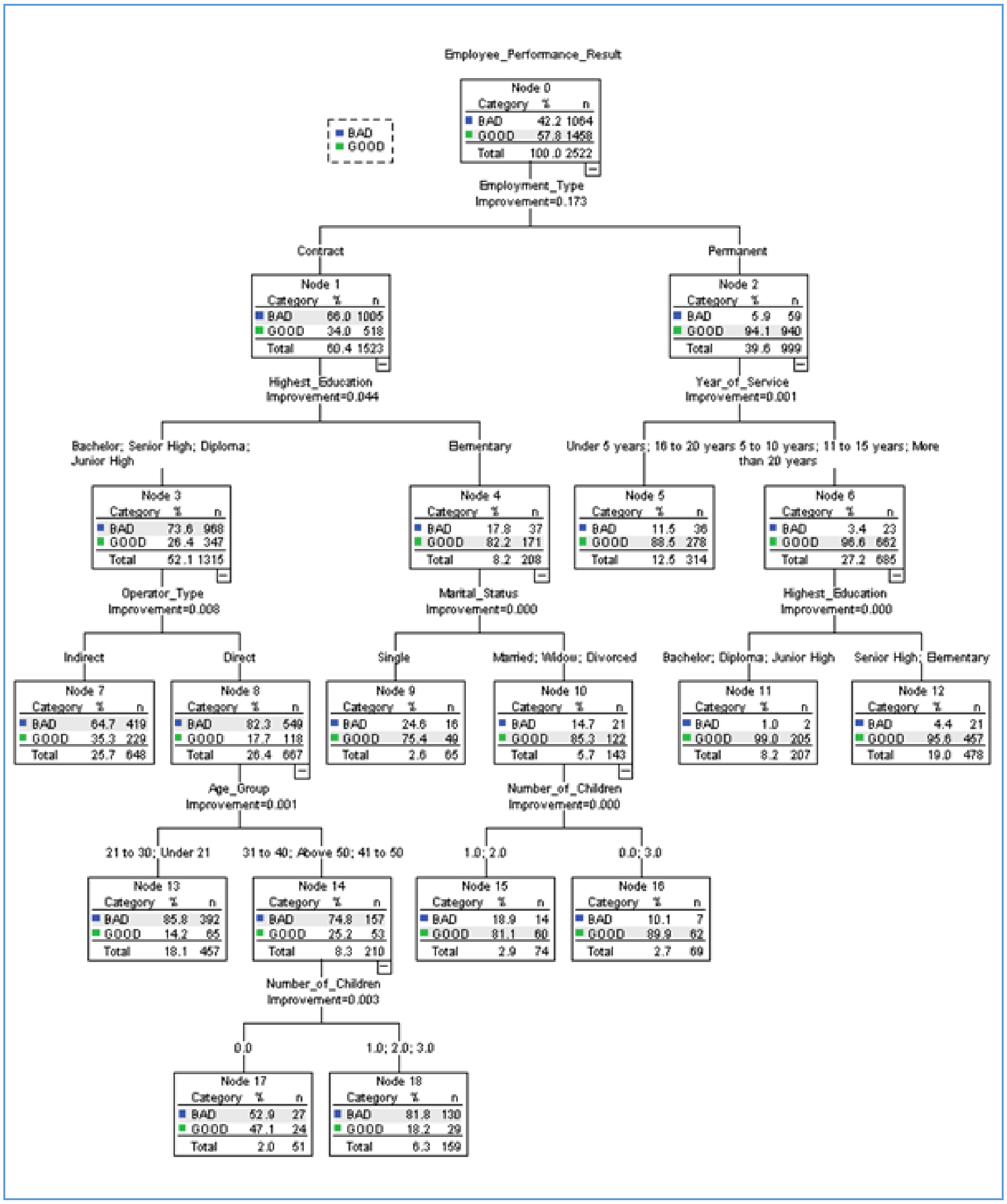


Figure 1. Decision tree model for training set.

While the result for test set sample was shown in Figure 2. This test set used 603 of 3.126 rows data. The decision tree chart shows the similar result compared to Figure 1. Figure 2 also has eighteen nodes and the same contains as Figure 1. The only difference is the amount of the data and the percentage of the data.

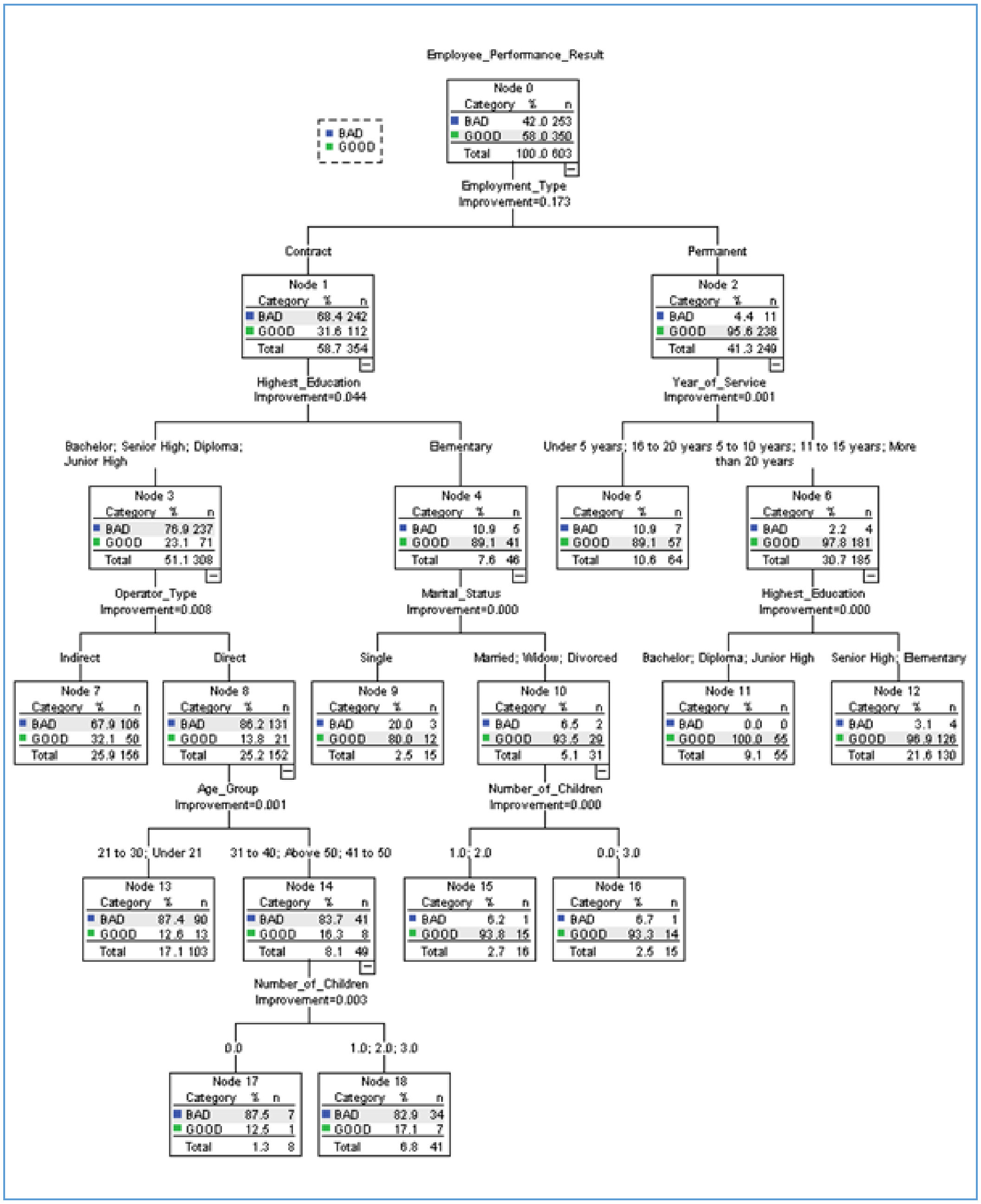


Figure 2. Decision tree model for test set.

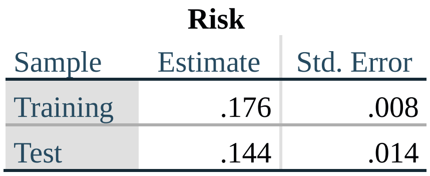
The result for each node in *training set* and *test set* decision tree are shown the same result. Table 2 shown the nodes in *training set* and *test set*.

Table 2. The result for training set nodes and test set nodes.

No	Nodes No	Training Set Nodes	Test Set Nodes
1	Node 0	Employee_Performance_Result	Employee_Performance_Result
2	Node 1	Contract	Contract
3	Node 2	Permanent	Permanent
4	Node 3	Bachelor, Senior High, Diploma, Junior High	Bachelor, Senior High, Diploma, Junior High
5	Node 4	Elementary	Elementary
6	Node 5	Under 5 years, 15 to 20 years	Under 5 years, 15 to 20 years
7	Node 6	5 to 10 years, 11 to 15 years, More than 20 years	5 to 10 years, 11 to 15 years, More than 20 years
8	Node 7	Indirect	Indirect
9	Node 8	Direct	Direct
10	Node 9	Single	Single
11	Node 10	Married, Widow, Divorced	Married, Widow, Divorced
12	Node 11	Bachelor, Diploma, Junior High	Bachelor, Diploma, Junior High
13	Node 12	Senior High, Elementary	Senior High, Elementary
14	Node 13	21 to 30, Under 21	21 to 30, Under 21
15	Node 14	31 to 40, Above 50, 41 to 50	31 to 40, Above 50, 41 to 50
16	Node 15	1, 2	1, 2
17	Node 16	0, 3	0, 3
18	Node 17	0	0
19	Node 18	1, 2, 3	1, 2, 3

The SPSS software also generated the estimates and standard of errors. The risk or the standard deviation of its sampling distribution was reflected in Table 3. The standard deviation of errors in the training set is lower than in test set.

Table 3. The comparison of standard deviation for training set and test set.



Risk

Sample	Estimate	Std. Error
Training	.176	.008
Test	.144	.014

The experiment also calculated the percentage of correct prediction. The prediction was compared between the accuracy of prediction based on experimental data in training set and test set. The overall accuracy percentage of the training set is 82.4%. This result is quite similar to the experiment in the test set, which is 85.6%. The result of classification technique for this study was shown in Table 4.

Table 4. The result of correct prediction in training set and test set.

Classification

		Predicted		
				Percent
Sample	Observed	BAD	GOOD	Correct
Training	BAD	968	96	91.0%
	GOOD	347	1111	76.2%
	Overall Percentage	52.1%	47.9%	82.4%
Test	BAD	237	16	93.7%
	GOOD	71	279	79.7%
	Overall Percentage	51.1%	48.9%	85.6%

Table 4 also showed that the correct prediction for predicting bad is higher than the correct prediction for good. And the result for training set is lower than test set. The class for bad and good performance is given by the HR information system based on the Performance Management System in XYZ Company. XYZ Company has been implemented Performance Management System for several years to identify the achievement of every employee. The good performance indicated employee with achievement A to C, while bad performance is employees with achievement D to E.

4. Discussion

The study has found that there are several attributes might have a great impact on employee performance. The most effective attribute is the employment type. The result showed that permanent employees have high possibility to give good performances. As seen in Figure 1, the possibility of the permanent employees with good performance is 96.6%. It can be assumed that employees are willing to work harder and giving the good performance if those employees became permanent employees because the result has shown the permanent employees are given the higher possibility to produce the good performance. However, employment type is not the only factor that determines the prediction of employee performance. Highest education became another factor to determine employee performance for the employees who work in the contract.

The study is also showing that not all dependent attributes or variables are used in order to build the decision tree. The gender attribute is not included in decision tree graph, which means gender does not have a significant relation to employee performance. It is probably caused by the distribution of gender data is very varied, so it cannot be analyzed.

In Table 4, the result of percent correct for predicting bad performance is higher than the prediction of good performance. This result was given by analyzing training set and test set. There are some possibilities for this result. The data diversity had the greatest impact on the experimental result. In training set, there are 2.522 rows of data which contains 1.054 rows of bad performance and 1.458 rows of good performance. The data is plenty balanced to produce a high percentage of data accuracy.

The experiment result is given a high accuracy for predicting bad performance. The accuracy for predicting good performance is not as high the result for bad performance. However, the result of this research still valid, because the standard deviation of error is less than 0.10 (10%) and the overall percentage of correct prediction is more than 80%. The result of this research might be different with other researchers. It depends on the varieties and contains.

The number of nodes in contract data is more than the number of nodes in permanent data. This result shows the factors that determined the performance of contract employees is more than permanent employees. Those factors can be a reference to showing the characteristics of employees who perform both good or bad.

The training set and test set were giving the same result of the most impacted attribute for predicting employee performance because the training set and test set were chosen randomly from the original dataset. And it could be that the distribution of both datasets is quite smooth.

5. Conclusion

Based on the result of this research, the data mining can be applied for predicting employee performance. The data mining technique which suitable for this research is classification. The reason is that the target or goal of the research has been defined. From some classification techniques in data mining, the decision tree technique was chosen because of its ability to visualize the probability and accuracy of the experiment into a graph.

References

Al-Radaideh, Q.A & Nagi, E.A. (2012) " Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance", International Journal of Advanced Computer Science and Applications (IJACSA) 3.

Chein, C. & Chen, L. (2006) "Data mining to improve personnel selection and enhance human capital: A case study in high technology industry", Expert Systems with Applications, In Press.

Han, J., Kamber, M. & Pei, J. (2012). *Data mining: concepts and techniques*. Amsterdam Boston: Elsevier/Morgan Kaufmann.

Jantan, H et al. (2010) "Human Talent Prediction in HRM using C4.5 Classification Algorithm", International Journal of Computer Science and Engineering (IJCSE) 2.

Jantan, H. & Hamdan, A.R (2010) "Human Talent Forecasting using Data Mining Classification Techniques", International Journal of Technology Diffusion edition October – December 2010.

Jantan, H. & Puteh, M. (2009) "Applying Data Mining Classification Techniques for Employees Performance Prediction", International Journal of Computer Science and Engineering (IJCSE).

Karthick, J., & Yousuf, M. (2016). Predicting Human Productivity for an Organization. *Indian Journal of Science and Technology, 9*(48).

Kotalwar, R. & Chavan, R (2014) "Data Mining : Evaluating Performance of Employee's using Classification Algorithm Based on Decision Tree", IRACST - Engineering Science and Technology : An International Journal (ESTIJ) 4.

Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2012). *Foundations of machine learning*. Cambridge, MA: MIT Press.