International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 6, December 2018, pp. 4763~4771
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i6.pp4763-4771

Comparative Study of Classification Method on Customer Candidate Data to Predict its Potential Risk

Mujiono Sadikin, Fahri Alfiandi
Faculty of Computer Science, Universitas Mercu Buana, Indonesia

Article Info

Article history:

Received Jan 8, 2018
Revised Jul 19, 2018
Accepted Jul 29, 2018

Keyword:

C4.5 algorithm
Data mining
Leasing
Naive bayes algorithm

ABSTRACT

Leasing vehicles are a company engaged in the field of vehicle loans. Purchase by way of credit becomes a mainstay because it can attract potential customers to generate more profit. But if there is a mistake in approving a customer candidate, the risk of stalled credit payments can happen. To minimize the risk, it can be applied the certain data mining technique to predict the future behavior of the customers. In this study, it is explored in some data mining techniques such as C4.5 and Naive Bayes for this purpose. The customer attributes used in this study are: salary, age, marital status, other installments and worthiness. The experiments are performed by using the Weka software. Based on evaluation criteria, i.e. accuracy, C4.5 algorithm outperforms compared to Naive Bayes. The percentage split experiment scenarios provide the precision value of 89.16% and the accuracy value of 83.33% wheres the cross validation experiment scenarios give the higher accuracy values of all used k-fold. The C4.5 experiment results also confirm that the most influential instant data attribute in this research is the salary.

Corresponding Author:

Fahri Alfiandi
Faculty of Computer Science,
Universitas Mercu Buana,
Meruya Selatan No. 1, Kembangan, Jakarta Barat 11650, Indonesia.
Email: 41514010101@student.mercubuana.ac.id

1. INTRODUCTION

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions [1]–[3]. The process is performed by extracting or recognizing the important pattern from the data contained in the database. In the data mining there are many techniques to do it, among of them are C4.5, Naive Bayes algorithm, Apriori, K-NN and many others.

Bank credit risk assessment is widely used at banks around the world. Some of bank risks include: credit risk, the risk that the loan won't be return back on time or at all; liquidity risk, the risk that too many deposits will be withdrawn too quickly, leaving the bank all; liquidity risk, the risk that too many deposits will be withdrawn too quickly, leaving the bank short on immediate cash; and interest rate risk, the risk that the interest rates priced on bank loans will be too low to earn the bank adequate money [4]. As credit risk evaluation is very crucial, a variety of techniques are used for risk level calculation. In addition, credit risk is one of the main functions of the banking community. Banks classify clients according to their profiles. While classifying, the financial background of the customers and subjective factors related to them are evaluated [5]. To facilitate the company in processing the large data, then the system would be needed to produce a decision on potential customer's risk. One of them is using data mining techniques, the much so that the data can be used optimally. By exploiting these data, it is expected to assist in addressing the

customer candidates whom are predicted will have payment problems in the future to assist in determining the prospective customer credit more as well.

In the study published in the Journal entitled "C4.5 Algorithm to Predict the Impact of the Earthquake" it is describe about the earthquake that cannot be predicted when it would happen, but we can predict the expected impact of the quake based on seismic data that never happened before. One of the methods used to dig or to search for information on old data is data mining algorithm C4.5. The output of the algorithm C4.5 in predicting the impact of the quake is divided into three parts. Namely, there are no impact/minor damage, severe damage, and the damage and tsunami. By predicting the implications of the earthquake, it is expected to minimize the quack impact. This study uses the C4.5 algorithm to predict the effects of earthquakes while the attributes that are used are the epicenter, distance from the beach, depth, scale, duration, and effect. The results of the study show the pattern to predict is based on the effects of earthquakes. If the scale is low, it does not cause any effect. If the scale is medium and in short duration, then there is no effect. If the scale is medium and in long duration, then it will cause the broken. If the scale is height and in a certain distance from the coast or it is happening on the land, it will cause the broken too. If the scale is height and its distance from the coast is very far, then it will cause broken and tsunami. If the scale is height and its distance from the coast is far and the epicenter in the sea, it will cause broken and tsunami [6].

The other study that utilizes the C4.5 is also presented in [7]. The study describes about rainfall, soil data and climate dataset that are used to predict the crop production. These types of datasets are preprocessed to remove the unwanted and null data in the dataset. The feature extraction method is used to extract a subset of new features from the datasets through functional mapping to maintain the information. In feature selection, genetic algorithm is used to select optimal features. The genetic algorithm provides the opportunity to discover the optimum solution. The enhanced ANFIS classifier then is used. The ANFIS classifier is the improvement of C4.5 classifier in hidden layer to generate the rules to predict the yield. By enhancing the C4.5, the experimental results of proposed work show better accuracy of 92.50 % than existing classifier. The comparative study of decision tree variants performance of information mining in the forest burned area is conducted by Putri et al as published in [8]. The study conducted comparative analysis of three decision tree variants ie. CART, C5.0, and C4.5 algorithm. Of these three decision techniques, the C5.0 algorithm is the most suitable for spatial data of the forest burned area. The algorithm is outperform shown by its accuracy is 99.79%.

In [9] authors show their study in using Naive Bayes classifier to predict the patient's hypertension disease. The hypertension disease is a significant health problem, and patients may not be able to recognize this disease for years. But in the other side, it's still difficult to answer complex queries such as "Given patient records, predict the probability of patients getting hypertension". Most of the time, clinical decisions are often made based on doctors intuition and experience rather than on the knowledge rich data hidden in the database. In this study, the Naive Bayes algorithm is employed to make a model with predictive capabilities. It provides new ways that of exploring and understanding knowledge. Attributes used in this research are as follows sex, chest pain, exam, age, systolic BP, diastolic BP, cholesterol, fasting blood sugar, thalach, old peak, the risk of hypertension. The Naive Bayes experiments in the study give performances as: the recall is 83.70%, the precision is 83.60% and the accuracy is 83.67%. Another interesting of naïve Bayes application for classification purpose is presented in [10]. In the study author present the result of the Zakah receiver classification experiment that utilizes the naïve Bayes classifier. According the experiment results, the classifier provides good accuracy i.e. 85 %. One of the application of naïve Bayes classifiers in media social mining domain is discussed in [11]. The study explored the application of Multinomial Naïve Bayes classifier technique to mine the sentiment opinion pattern of GSM based on customer's twitter account. By using 1665 features of the dataset, the technique provides the accuracy results of 73.15 %.

In this work we perform an experimental study of Naive Bayes and C4.5 algorithm that applied to the company leasing customer data history. The purpose of the data is to evaluate the performance of both algorithms in assisting the company leasing to make the decision regarding the approval of customers candidate who apply the leasing. The such study is critical to local Indonesia context since the financial technology is currently growing quickly while the information technology, especially the software/ application, the environment is still in the initial phase. According to the author's knowledge, there is a very limited publication related the application of Artificial Intelligent or Machine Learning to this domain for Indonesia cases.

2. MATERIAL AND METHOD

2.1. Classification

Classification is one of the Data Mining techniques that is mainly used to analyze a given dataset and takes each instance of it and assigns this instance to a particular class such that classification error will be least. It is used to extract models that accurately define important data classes within the given dataset. Classification is a two step process. During the first step the model is created by applying a classification algorithm for training data set, then in the second step the extracted model is tested against a predefined test dataset to measure the model trained performance and accuracy. So classification is the process to assign a class label from dataset whose class label is unknown [9].

2.2. C4.5 Algorithm

C4.5 algorithm is an algorithm used to construct a decision tree [12], a classification and prediction methods are extremely powerful and famous. Decision tree method changes the very large fact into a decision tree that represents the rule. The decision tree is also useful to explore the data in finding the relationship between input variables and a certain output/target variable. In general, C4.5 algorithm to construct a decision tree is described as follows:

a. Select an attribute as root.
b. Create a branch for each value.
c. For the case of the branches.
d. Repeat the process for each branch until all cases the branches have the same class.

To select an attribute as roots, is based on the highest gain value from the existing attributes. To calculate the gain used formula as follows:

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

Information:
S: The sets of cases
A: Attribute
n: The number of partitions attribute A
$|S_i|$: Number of cases in the i partitions
|S|: Number of cases on S
Meanwhile, the calculation of entropy value follows:

$$Entropy = \sum_{i=1}^{n} - pi * log_2\ pi$$

Information:
S: The sets of cases
A: Feature
n: The number of partitions S
pi: The proportion of S_i againts S

2.3. Naive Bayes Algorithm

Naive Bayes algorithm studies the events of the database record by calculating the variables which are analyzed with other variables [13]. The result of this process is we can predict something such as whether or a person coming from certain groups based on variables attached to it. Additionally, Naive Bayes can also analyze the variables that most influence in the form of probabilities. Naive Bayes is a simple probability-based prediction techniques based on the application of Bayes theorem to assume strong independence. The steps below are Naive Bayes stages process:

a. Counting the number of classes / labels
b. Counting the number of cases per class
c. Multiply all class variables
d. Compare results per class

The formula of Naive Bayes Algorithm is as follows:

$$P\ (C\ |\ X\ ) = \frac{P(x\ |\ c)\ P(c)}{P(X)}$$

Information:
x : Data with unknown class
c : Hypothesis of data is a specific class
P (c | x) : The probability of a hypothesis based on the conditions
P (c) : The probability of a hypothesis
P (x | c) : Probability based on hypothetical conditions
P (x) : Probability c

2.4. Weka Tools

Weka is a collection of machine learning algorithms for data mining tasks. Weka stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, New Zealand. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization [14]. The workflow of Weka would be follows in Figure 1.

Figure 1. Weka Flow

2.5. Data Set

The data source used in this research is collected from one of the leasing companies located in the area of Cikupa-Tangerang, Banten Province. The total amount of data collected are 560 record data, each instant contains 5 attributes, namely: age, marital status, salary, other installments and worthiness as presented as Table 1. Worthiness attribute is the target variable/label. Some samples of data instant are described in Table 2.

Table 1. Data Set Attribute

No.	Attribute	Attribute Value
1	Age(Years)	23, 40, 50 so on
2	Salary(Rupiah)	1 Milion, 4 Milion so on
3	Other Installments	Yes, No
4	Marital Status	Married, Single
5	Worthiness	Worth It, Not Worth It

Table 2. Example of Data Set Attribute Value

No.	Age	Salary	Other Installments	Marital Status	Worthiness
1	21	4.400.000 IDR	No	Married	Not Worth It
2	23	10.600.000 IDR	Yes	Married	Not Worth It
3	43	14.000.000 IDR	Yes	Married	Worth It
4	54	13.000.000 IDR	No	Married	Worth It
5	25	4.700.000 IDR	Yes	Single	Not Worth It

Two of four attributes, age and salary, can contain values in wide range, so this condition will make suffer in its computation. To deal with this problem we apply the categorization mechanism to both of attribute values as presented in Table 3. Table 4 shows data example.

Table 3. Data Set Attribute Categorization

No.	Attribute	Attribute Value	Attribute Categorization
1	Age(Years)	23, 40, 50 so on	- Age < 45 : Young - Age > 45 : Old
2	Salary(Rupiah)	1 million, 4 million so on	- < 5 million: Low - 5 – 10 million: Middle - > 10 million: High
3	Other Installments	Yes, No	Yes, No
4	Marital Status	Married, Single	Married, Single
5	Worthiness	Worth It, Not Worth It	Worth It, Not Worth It

Table 4. Example of Data Set Categorization

No	Age	Salary	Other Installments	Marital Status	Worthiness
1	Young	Low	No	Married	Not Worth It
2	Young	High	Yes	Married	Not Worth It
3	Young	High	Yes	Married	Worth It
4	Old	High	No	Married	Worth It
5	Young	Low	Yes	Single	Not Worth It

2.6. Experiment Scenario

The main parts of experiment scenario consist of two steps. The first step is to obtain the best model from each algorithm and the second is to compete the both best models obtained. The detail of experiment stages and scenario is illustrated as the Figure 2.

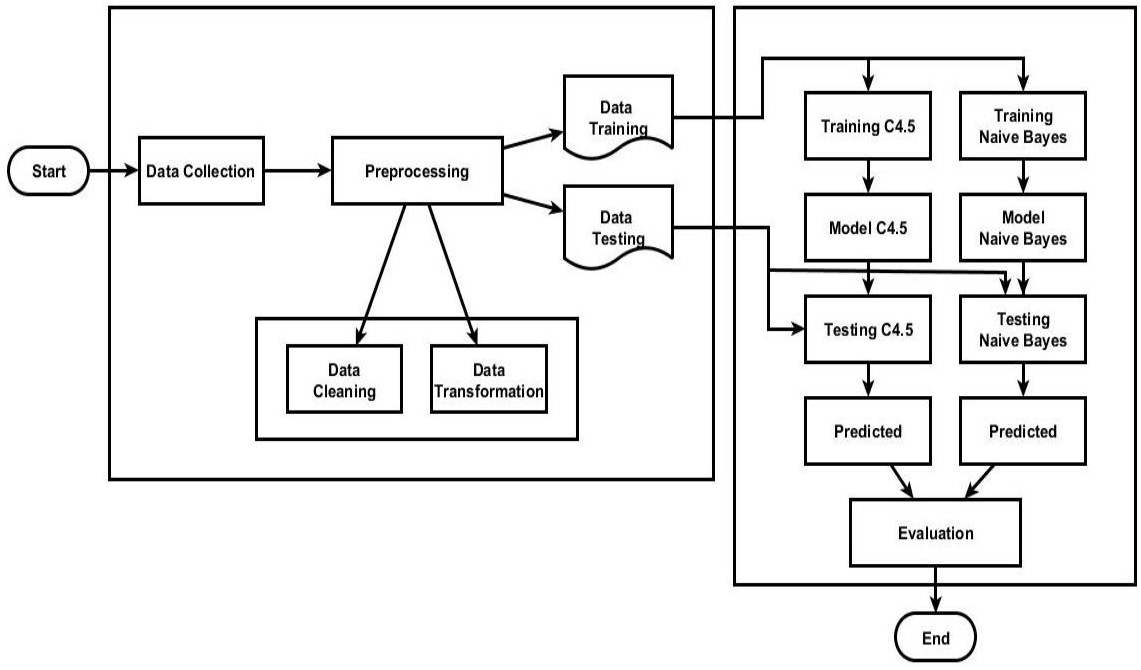

Figure 2.The Experiment Scenario

The data collected is not ready yet to be processed by the algorithm since there are too many biases or ambiguous contained on it, so it needs to perform the data preprocessing operation. In this step we perform data cleaning by ignoring the uncompleted data. The next step of data preprocessing is a data transformation that transforms the data format to format that compatible with Weka tools. Data splitting is then applied to the data to divide the data into two parts: training data and testing data. In this case, we use 80% parts of the data for data training, and the rest as data testing. The same training data is then used to train both of algorithm to provide the models which will be tested with the same data testing. For both algorithms used, we perform twenty experiment runs to get the best model of each algorithm. Both of the best models are then competed to evaluate their performance and to get the best model among of C4.5 and Naive Bayes.

2.7. Data Preprocessing

Data preprocessing is required to improve the quality of the data by removing the unwanted data from the original data [15]. Preprocessing data is important since the raw data contains missing values, noisy, and inconsistent data it will result in data not qualified. In this study, we do data preprocessing as follows:

a. Data Cleaning

Data cleaning is to do data cleaning of the noise found in the form of missing values, inconsistent data, and redundant data. All the above attributes will then be selected to obtain attributes that contain relevant values, not missing values, and not redundant, where the three requirements are the prerequisites that must be done in data mining so that will be obtained a clean dataset for use in the data mining stage . In this dataset found 1 missing value, the technique that will be done for 1 missing value record is to delete it record.

b. Data Transformation

The data transformation stage is at this stage the data is converted into the appropriate form for processing in data mining. In this study the data will be processed from Microsoft excel will be converted into a CSV file (Comma Separated Values) which can be used for data processing on Weka tools.

2.8. Evaluation

To evaluate the performance of both algorithms, we use the common criteria in data mining i.e. precision, recall, and accuracy. The calculation of those parameters is performed by to provide a confusion matrix. A confusion matrix contains information about actual and predicted class provided by a classification system [16]. All correct classifications that lie along the diagonal from the north-west corner to the south-east corner also is called True Positives (TP) and True Negatives (TN) while other cells are stated as the False Positives (FP) and False Negatives (FN)[17]. In this study, the likely cases are considered as the positive case, while the unlikely and probable cases are the negative cases. The definitions of these parameters are presented as follows:

a. True positives (TP) are correctly classified yes cases.
b. False positives (FP) are incorrectly classified no cases.
c. True negatives (TN) are correctly classified no cases.
d. False negatives (FN) are incorrectly classified yes cases.

The true positive/negative and false positive/negative values recorded from the confusion matrix, then can be used to evaluate the performance of the prediction model. A description of the definition and expressions of the metrics is presented as follows[18]:

a. Recall is an average per-class effectiveness of a classifier to identify class labels.

$$Recall = \frac{TP}{TP + FN}$$

b. Precision is the ability of a classifier to determine the positive labels by using one versus all approach.

$$Precision = \frac{TP}{TP + FP}$$

c. Accuracy is the sum of the ratios of correct classifications to the number of total classifications by using a one versus all approach.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

3. RESULTS AND DISCUSSION

This section presents the experimental results and analysis of this study which utilize two clasifiers, C4.5 and Naive Bayes. Three experiments scenarios based on percentage data splitting are performed to each algorithm. The first experiment uses 60% of training data and 40% of the data testing, the second experiment uses 70% of training data and 30% of the data testing, and the third experiment uses 80% training data and 20 % data testing. The experiment which provides the highest performance values for each method is used as a model to find the best method by re-testing on provided data testing. The Table 5 presents the average performance parameter values of each experiment scenario of C4.5 on model testing stages, while Table 6 shows the results of Naive Bayes. Based on the achieved value of accuracy criteria, the first experiment

scenario is the best for both of the algorithm. In the first scenario, the C4.5 accuracy is 82.59 %, whereas the Naive Bayes accuracy is 80.35 %.

Table 5. C4.5 Algorithm Test Performance

Experiment	Accuracy	Precision	Recall
1	82.59%	86.77%	82.03%
2	80.37%	85.10%	80.80%
3	80.37%	87.50%	80%

Table 6. Naive Bayes Algorithm Test Performance

Experiment	Accuracy	Precision	Recall
1	80.35%	80.16%	82.90%
2	77.38%	78.72%	80.43%
3	77.68%	81.25%	81.25%

The next stage of the experiment is to compare the best model provided from each experiment scenario which are run for both algorithms. These two models then are applied to the data testing that has been provided to get which of algorithm that is suitable for the study case. The results of this comparison stage are presented as Table 7. Table 7 shows that the C4.5 algorithm is superior compared to the Naive Bayes algorithm with its accuracy is 83.33%, while the Naive Bayes algorithm achieved is 80.67%.

Table 7. Comparison C4.5 Best Model and Naive Bayes Best Model on Testing Stage

Criteria	C4.5 Algorithm	Naive Bayes Algorithm
Accuracy	83.33%	80.67%
Precision	89.16%	80.72%
Recall	82.22%	83.75%

To validate the result above, we perform the next experiment based on the cross validation evaluation scenario. Three different *k*-folds are used in the scenario i.e. 5-fold, 10-fold, 20-fold and each these *k-fold* is applied to both C4.5 and Naïve Bayes as well. The results are presented as Table 8 and Table 9. Table 8 presents C4.5 performance, whereas Table 9 presents Naïve Bayes performance. The cross validation experiment confirms that, in this case, C4.5 achieves better performance compared to Naïve Bayes. Of all k-folds applied C4.5 presents better accuracy than Naïve Bayes. The other information presented by the results is their different performance pattern. C4.5 gives a better accuracy performance for the less k-fold, whereas Naïve Bayes better accuracy performances are provided by the bigger k-fold.

Table 8. C4.5 Cross Validation Scenario Performance

	Precision	Recall	Accuracy
5-fold	80.48%	83.07%	81.58%
10-fold	80.73%	83.07%	81.56%
20-fold	81.17%	83.06%	81.50%

Table 9. Naive Bayes C4.5 Cross Validation Scenario Performance

	Precision	Recall	Accuracy
5-fold	76.47%	84.86%	80.39%
10-fold	76.73%	84.87%	80.41%
20-fold	77.25%	84.85%	80.41%

The superiority of C4.5 compared to Naive Bayes can be understood since all of the input variable are independence each other, so C4.5 is more suitable to this characteristic of data. On the other side, the nature of the Naive Bayes algorithm is based on the conditional probability of input variables, so in this case the advantages of Naive Bayes is less use. Another implication shown by the results is that the customer leasing application tends to fall into recommender application rather than classification.

4. CONCLUSION AND FUTURE STUDY

In this study, C4.5 Algorithm and Naive Bayes Algorithm were implemented on a customer credit dataset to predict the potential risk in the future. Based on two types of experiments scenario results, C4.5 algorithm achieves better performance. The results study presents that the recommender system as the characteristics of C4.5 is more suitable than Naive Bayes which work based on conditional probability of the input variables. Whereas, on C4.5 algorithm salary attribute is the most influential attribute shown by the its significant value of entropy gain compared to other input variables. The dominant influence of the salary attribute is also presented in every experiment scenario where the attribute is always selected as the root node of the tree. In the future study, we will explore some opportunities to apply the others technique in this domain. We also will investigate the other real applications which still open to exploit such as: customer care, sales recommender, and micro finance which is growing quickly.

REFERENCES

[1] Karim M, Rahman RM. Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. J Softw Eng Appl 2013; 06: 196–206.
[2] Dimitoglou G, Dimitoglou G, Adams JA, et al. Comparison of the C4 . 5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Comparison of the C4 . 5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability.
[3] Arifin MF, Fitrianah D. Penerapan Algoritma Klasifikasi C4.5 Dalam Rekomendasi Penerimaan Mitra Penjualan Studi Kasus : PT Atria Artha Persada. InComTech 2018; 8: 87–102.
[4] Jafar Hamid A, Ahmed TM. Developing Prediction Model of Loan Risk in Banks Using Data Mining. Mach Learn Appl An Int J 2016; 3: 1–9.
[5] Krichene A. Using a naive Bayesian classifier methodology for loan risk assessment. J Econ Financ Adm Sci 2017; 22: 3–24.
[6] Buulolo E, Silalahi N, Fadlina, et al. C4.5 Algorithm To Predict the Impact of the Earthquake. Int J Eng Res Technol 2017; 6: 10–15.
[7] Poongodi S, Babu MR. Prediction of Crop Production using Improved C4 . 5 with ANFIS Classifier. 10.
[8] Thariqa P, Sitanggang IS, Syaufina L. Comparative Analysis of Spatial Decision Tree Algorithms for Burned Area of Peatland in Rokan Hilir Riau. Telkomnika (Telecommunication Comput Electron Control 2016; 14: 684–691.
[9] Nikam SS. A Comparative Study of Classification Techniques in Data Mining Algorithms. Orient J Comput Sci Technol 2015; 8: 13–19.
[10] Basri Hasanuddin Z, Syarif S. Zakah Management System using Approach Classification. Telkomnika (Telecommunication Comput Electron Control 2017; 15: 1852–1857.
[11] Susanti AR, Djatna T, Kusuma WA. Twitter's Sentiment Analysis on Gsm Services using Multinomial Naïve Bayes. TELKOMNIKA (Telecommunication Comput Electron Control 2017; 15: 1354.
[12] Larose DT. DISCOVERING KNOWLEDGE IN DATA An Introduction to Data Mining. John Wiley & Sons, Inc., 2015.
[13] Patil, T. R., Sherekar MS. No Title. Perform Anal Naive Bayes J48 Classif Algorithm Data Classif; 6.
[14] waikato. Weka 3: Data Mining Software in Java.
[15] Ȧ SR, Sonika Ȧ. Effectiveness of Data Preprocessing for Data Mining. 2014; 4: 3480–3483.
[16] Santra a. K, Christy CJ. Genetic Algorithm and Confusion Matrix for Document Clustering. Int J Comput Sci 2012; 9: 322–328.
[17] Sadikin M, Fanany MI, Basaruddin T. A New Data Representation Based on Training Data Characteristics to Extract Drug Name Entity in Medical Text. 2016.
[18] Mehdiyev N, Enke D, Fettke P, et al. Evaluating Forecasting Methods by Considering Different Accuracy Measures. Procedia Comput Sci 2016; 95: 264–271.

BIOGRAPHIES OF AUTHORS

Mujiono Sadikin is faculty member of Faculty of Computer Science Universitas Mercu Buana Jakarta. He held doctoral degree from Univiersitas Indonesia, Jakarta 2017. His research area is in Data Mining, Machine Learning, and IT Governance as well. Some of his experiences are: As team leader in IT Governance an Procedure preparation of Directorate Land & Transportations Ministery of Transportation, Team leader of IT Audit and Assessment Universitas Mercu Buana, and some more. Since 2012 he leads the University of Mercu Buana IT Directorate as the Director.

Fahri Alfiandi is a student in Faculty of Computer Science, Universitas Mercu Buana, Indonesia. He was born in Jakarta on December 16th, 1995. He is interested in data mining, algorithm analysis and programming.