Prediction of Teachers' Lateness Factors Coming to School Using C4.5, Random Tree, Random Forest Algorithm

Windu Gata, Grand Grand
Computer Science
Post Graduate STMIK Nusa Mandiri
Jakarta, Indonesia
windugata@nusamandiri.ac.id; grandnoowen@gmail.com

Yuyun Elizabeth Patras, Rais Hidayat
Education Management
Pakuan University
Bogor, Indonesia
yuyunpatras64@gmail.com; rais72rais@gmail.com

Rhini Fatmasari
Education Management
Universitas Terbuka
Jakarta, Indonesia
riens@ecampus.ut.ac.id

Siswanto Tohari
Faculty of IT
Budi Luhur University
Jakarta, Indonesia
siswantobl@gmail.com

Baharuddin Baharuddin
Faculty of Islamic Religion
Universitas Muhammadiyah Makassar
Makassar, Indonesia
afinyeyen@yahoo.com

Nia Kusuma Wardhani
Faculty of Economic
Mercubuana University
Jakarta, Indonesia
nia.kusuma@mercubuana.ac.id

***Abstract*—Lateness arrives at work can be experienced by anyone, including teachers. Teachers who are late arriving at school have shown examples of bad behavior for students. It takes a study to determine the factors that cause a teacher to arrive late to school. Data Mining is selected to process the data that has been available. Processing uses 3 classification algorithms which are decision tree (C4.5, Random Tree, and Random Forest) algorithms. All three algorithms will be tested for known performance, where the best algorithm is determined by accuracy and AUC. The results of the research were obtained that Random Forest with pruning and pre-pruning is the best for accuracy value with 74.63% and also AUC value with 0.743. The teacher's delay in this study is often done by teachers who have a vehicle compared to those who do not have a vehicle.**

***Keywords*—data mining; C4.5; random tree; random forest; accuracy; AUC**

I. INTRODUCTION

Lateness is connected with late [1]. Lateness refers to an assumption that all activities that can't be done prematurely or precisely at a given time. From various examples of lateness that can be encountered, lateness arrives at work is one of them. Lateness arrives at work can be experienced by anyone, including teachers. Lateness arrives at the workplace can be categorized as an undisciplined form of attitude and responsible for work, agency and workplace organization, and against others. It is a duty for everyone who works to arrive before time or on time at work.

Teachers who are late arriving at school have shown examples of bad behavior for students. Teacher's lateness will be used by the students when they arrive late for school. A teacher should not come late to school, even teachers are expected to arrive before the deadline is set. It is also late in relation to discipline, by showing discipline it can be concluded that the teacher is a professional teacher. Through disciplined action, the teacher can serve as an example for the disciples. Exemplary teachers can be seen in the attitude shown teachers in everyday life, both inside and outside school [2, 3].

There are several factors that may affect a teacher coming late to school, including: the distance of the house, the vehicle used, and so forth. Various statements are given regarding the most influencing factors on teacher's lateness coming to school. It takes a study to determine the factors that cause a teacher to arrive late to school.

Based on search that has been conducted on related research, it is known that there is still a lack of research on the prediction of teacher's lateness factors, especially in the field of Data Mining.

II. Related Works

Related research used as reference is a study that also uses Data Mining by classification method [4]. Using two classification algorithms, namely C4.5 and Naive Bayes. These two algorithms are compared to find a better algorithm between the two. Measurement metrics used as evaluators are: time to build models (second), correctly classify, incorrectly classify, accuracy, precision for yes, precision for no, AUC (Area under Curve). This study uses techniques from previous studies to extract actionable knowledge. It works post processing technique to mine actionable knowledge from decision tree.

Next research by Wajhillah, comparing the standard C4.5 algorithm with C4.5 that has been optimized with PSO (Particle Swarm Optimization) [5]. This research uses famous Data Mining methodology, CRISP-DM. Metric measurement used as performance evaluator for the algorithms are accuracy and AUC value. Defiyanti comparing ID3 with its successor, C4.5. This study tried to prove whether C4.5 can be better than its predecessor algorithm, ID3. Metric measurement used as performance evaluator for the algorithms are precision, recall, and accuracy value [6].

Uses 5 decision tree algorithms: J48, LMT, Random Forest, Random Tree, and Decision Stump. This study tried to get the best decision tree algorithm used [7]. Metric measurement used as performance evaluator for the algorithms are TP rate, FP rate, precision, recall, f-measure, ROC area. Uses 2 classification algorithms: C4.5 and Naïve Bayes [8]. To compare its two algorithms, measurement metric used are time to build model (second), accuracy, TP rate, FP rate, Kappa, Precision, Recall, ROC area. This study not only uses the research dataset that has been formed, but also uses 2 additional datasets: breast cancer and Irish. In this study also compared 2 validation models, they are: percentage split and 10-fold cross validation. The results to be obtained is which validation model is better.

Uses 7 classification algorithms: AD Tree, C4.5, J48 graft, LAD Tree, NB Tree, Random Tree, Random Forest, and REP Tree [9]. To get best algorithm, 2 feature selections are used: CONS (Consistency Subset) and CFS (Correlation Feature). These 2 features selection used to remove redundant or irrelevant features from the data to increase classification accuracy and decrease computational costs. It is the process of choosing a subset of original features that optimally reduces the feature space to evaluation criterion. Measurement metric used are time to build model (second), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), Kappa Statistic, Recall, Precision, F-Measure, False Alarm Rate, Accuracy, Error rate.

Uses dataset from news website. Process used are stop word removal, stemming, tokenization and ultimately generated the frequency matrix [10]. Stemming is used as many times when news is printed, for a same there can be many variants depending on the tense used or whether it is singular or plural. Such words when processed for stemming, generates a unique word. Stop words needs to be removed as they do not contribute much in the decision-making process. Frequency matrix thus generated can be processed for generating a model and the model so generated was used in further decision process. 3 decision trees algorithm chosen in this study: Random Tree, Random Forest, and LAD Tree.

As performance evaluator, 6 measurement metrics chosen: TP rate, FP rate, Precision, Recall, F-Measure, and ROC area. All algorithms tested in all news article selected. And the last study from uses 4 classification algorithms, they are: C4.5, Naïve Bayes, Neural Network, Logistic Regression [11]. As performance evaluator, 6 measurement metrics chosen: Accuracy and AUC.

III. Literature Survey

A. *Data Mining*

Data Mining is the process of extracting interesting information from large amount of data so as to obtain useful patterns and knowledge. Information that is extracted can't be trivial, implicit, unknown, and potentially useful [12,13]. This process should be automatic or more often semi-automatic [13]. The patterns found can be called structural, because they have a form that can be examined, considered, and used for information in the future. In other words, helps explain the data [13].

The beginning of the presence of Data mining, because it requires a discipline to process data that amounts to very large. People can call this large amount of data as an abundance of data. The abundance of untreated data leads to the emergence of "data grave" statements. Decision makers have not been able to use the data because it is still a raw data [13]. Berry Divide the main task of Data Mining into 6, namely: classification, estimation, prediction, association rule, clustering, and description / profiling [14].

Data Mining Process by Anggarwal consists of 3 step blocks [15]:

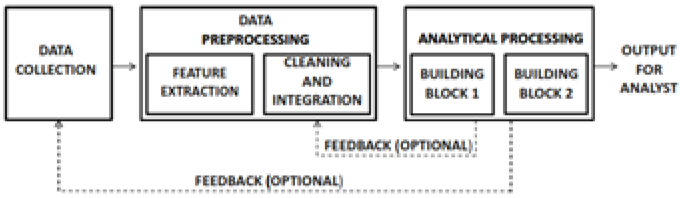


Fig. 1. Data mining process.

B. *Classification*

Classification is a type of data analysis that can help people predict class labels from samples to be classified [16]. In predicting class label used classification model [12]. Class labels can be "yes" and "no", "late" and "not late", "A", "B" and "C". Classification is included in supervised learning, because the data has class labels [12]. Various classification techniques have been proposed in several areas, such as: machine learning, expert systems and statistics [12,16].

Typically, model / classification algorithms are trained first in historical data sets (i.e., training sets) with class labels already known. Then, a trained classifier is applied to predict the new sample class label [16]. Classification has been applied to several things, such as: cheating detection, target marketing, performance prediction, manufacturing, and medical diagnosis [12]. Some evaluation and selection measurement models can be used to test the performance of classification algorithms, consists of: confusion matrix, accuracy, ROC curve, precision, recall, f-measure, fβ.

C. C4.5

C4.5 is a series of systematically arranged questions so that each question asks attributes and branches based on attribute values. In the leaf placed predicted class variables [17]. C4.5 is the successor of ID3 introduced by Quinlan [12]. C4.5 is one of decision tree algorithm [15]. C4.5 becomes a benchmark that is often compared to newer supervised learning algorithms [12]. C4.5 is included as an algorithm that handles classification problems in machine learning and data mining [17].

C4.5 adopts a greedy approach, where decision trees are constructed by recursive separation from top to bottom and the top most attribute is the most influential attribute of the attributes under it. C4.5 uses the pessimistic pruning method, where the decision to prune part of the tree is determined based on the estimated error rate [12]. The C4.5 algorithm begins with a root node that is subdivided into another part of the node resulting from testing the attribute variable that meets the test value. If the test result is a node that can still be tested again, it is called a branch and if it can’t be tested again or is the end result, it is called a leaf known as a label / class. This process will end if all branch nodes tested end in the leaf [17].

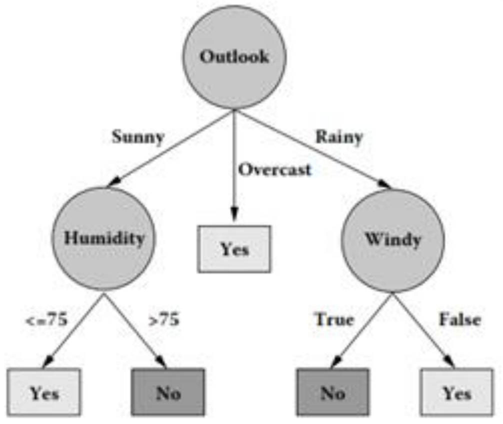


Fig. 2. Decision tree inducted by C4.5.

D. Random Tree

Random Tree is a supervised learning algorithm developed by Breiman [10,18]. Random Tree is a combination of two ideas of classification algorithms, a single tree decision model and Random Forest [19]. Random Tree uses an alternative approach: trees are roughly offset by just splitting the median of some attributes. The approximate procedure for median calculations was recently described on. This procedure requires only two linear scans as long as the data approaches the median.

The random rod model uses this result for split selection and thus induces a fairly balanced tree in which a global setting for the ridge value works on all leaves, making it simpler for optimization procedures. In addition, to prevent extrapolation of extreme extremes, each leaf (or hypercube) records the local minimum and maximum values for the target. Predictions from the local model are then compared with this threshold and are limited if necessary. This simple procedure proves to be very effective, since a single extreme value can have a major effect on the size of such an average squared error, even after multiple predictions of the model tree ensemble.

Finally, because the trees are semi-random and therefore certainly not optimal in isolation, the average number of appropriate trees is critical for good predictive performance. At least 30 trees should always be counted, and counting more (and sometimes more) more trees improves their performance. Of course, because the random nature of the process of adding more trees to the ensemble will never degrade performance significantly, but for most group methods, improvement is reduced.

E. Random Forest

Random Forest was first introduced by Leo Breiman through his paper in 2001. Random Forest is a combination of tree predictors, so that each tree depends on randomly selected random vectors and with the same distribution for all trees in the forest [10]. Random Forest cultivates many classification trees. To classify a new object from the input vector, enter the input vector into each tree in the forest. Each tree gives a classification called "voice" for the class. Forests choose the classification that has the most votes (over all trees in the forest) [20].

Random Forest builds a CART tree classification group using a bagging mechanism. By using bagging, each tree node chooses only a small part of the feature for separation, allowing the algorithm to create a classifier for high-dimensional data very quickly. This somewhat contradictory strategy works very well compared to the latest method in classification and regression. In addition, Random Forest runs efficiently on large data sets with many features and fast execution speed. Random Forest generates additional facilities, especially important variables with numerical values [10].

Stages of tree formation are divided into 3 [20]:

- If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
- If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
- Each tree is grown to the largest extent possible. There is no pruning.

IV. Research Metodology

A. Research Type

The type of research on this thesis is empirical and applied research. Empirical, because this research is a data-based research, namely: teacher data, weather, teacher absenteeism. Experiments were conducted to prove whether or not the hypothesis has been given by the researcher in the early stages, for example: house spacing is the most influential factor on teacher delay coming to school. Applied, because this study aims to find solutions to practical problems that will be faced by society, organizations and agencies. In this study, the agency is Kalam Kudus Christian School III Jakarta.

B. Research Approach

The research approach that underlies this thesis research is an experimental approach. The experimental approach is part of a quantitative approach [21]. Experimental approach was chosen because the research of this thesis aims to find out the attributes that are the determinants of teacher delay coming to school and to know the relationship between attributes through the decision tree.

C. Data Collection Method

The data used in this study consists of 2, namely primary and secondary. For primary data collection, conducted through the provision of questions written on a form to obtain teacher data. While collecting secondary data obtained through 2 sources, namely BMKG official website for weather data and fingerprint sensors for teacher attendance data [22]. Data duration time used for 2.5 months.

D. Research Instruments

This research uses 3 research instruments, consists of:

1) Teacher's data form: Used to collect teacher data. The data written on the next form will be entered into the dataset for processing.

2) RapidMiner Studio Educational versi 8.1: Used to process teacher data, weather, and teacher attendance are incorporated into a dataset.

3) PHP: Hypertext Preprocessor: Used to build the GUI (Graphical User Interface) from the results of this study.

E. Research Method

The method used is the classification of Data Mining. As a classifier selected 3 decision tree algorithms, namely: C4.5, Random Tree, and Random Forest. At the evaluation stage a number of classification metrics will be used. The best performing algorithms will be selected for their use as a basis for GUI creation.

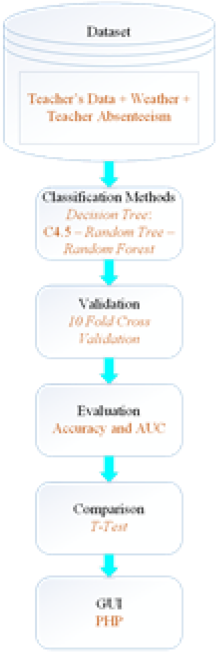


Fig. 3. Research framework.

V. Experimental Results

A. Evaluation

The experiments were performed using hardware and software specifications: Intel Core i7-4710HQ 2.5 GHz CPU, 8GB RAM, Microsoft Windows 8.1 Enterprise 64 bit. To process the data, used application RapidMiner Studio Educational version 8.1 which is a Data Mining application.

The performance test of the algorithm uses some standard classification metrics: sensitivity, specificity, PPV, NPV, accuracy, AUC, precision, recall, f-measure, fβ. Testing is done on algorithm which use parameter pruning and pre-pruning and which do not use it. In previous studies [10], there has never been a comparison of decision tree algorithms with pruning and pre-pruning parameters and those not using them.

TABLE I. Algorithm Test Performance Result

	Accuracy	**AUC**
C4.5 - PP	**71.85%**	**0.705**
C4.5	**67.81%**	**0.587**
Random Tree - PP	**68.69%**	**0.594**
Random Tree	**67.54%**	**0.621**
Random Forest - PP	**74.63%**	**0.743**
Random Forest	**67.49%**	**0.686**

PP = Pruning and Pre-Pruning

In Table 1, the best performance algorithm for AUC and accuracy is Random Forest with pruning and pre-pruning parameters. The value of AUC obtained is 0.743 while the accuracy value is 74.63%. C4.5 with pruning and pre-pruning parameters and Random Tree also obtained good classification results for AUC values.

In Table 1, almost all algorithms that use pruning and pre-pruning parameters have improved performance. For AUC values, C4.5 and Random Forest algorithms with pruning and pre-pruning parameters are fair classification. While Random Tree and Random Forest without pruning and pre-pruning parameters are poor classification. C4.5 without pruning and pre-pruning parameters and Random Tree with pruning and pre-pruning parameters are fail classification.

In this study also used the Parametric Method, T-test to test the differences of all algorithms used. Where the value of alpha ≥ 0.05 is the limitation of an algorithm is said to be better than other algorithms. In previous studies, no one has ever used a different test method between algorithms [4-11]. Still limited to classification metrics.

TABLE II. SIGNIFICANCE RESULT

	0.718 +/- 0.033	**0.678 +/- 0.057**	**0.687 +/- 0.085**	**0.675 +/- 0.102**	**0.746 +/- 0.050**	**0.675 +/- 0.053**
0.718 +/- 0.033		0.069	0.289	0.220	0.160	0.040
0.678 +/- 0.057			0.788	0.941	0.011	0.897
0.687 +/- 0.085				0.786	0.073	0.708
0.675 +/- 0.102					0.064	0.990
0.746 +/- 0.050						0.006
0.675 +/- 0.053						

In Table II, there are 3 significance differences were founded, they are:

- C4.5 with pre-pruning and pruning parameters and Random Forest without pre-pruning and pruning parameters.
- C4.5 without pre-pruning and pruning parameters and Random Forest with pre-pruning and pruning parameters.
- Random Forest with pre-pruning and pruning parameters and Random Forest without pre-pruning and pruning parameters.

In addition to determining the best performing algorithm, this study also has a goal to find the factors that most influence the delay in coming to school teachers. Determination of the most influential attribute based on Gain Ratio value.

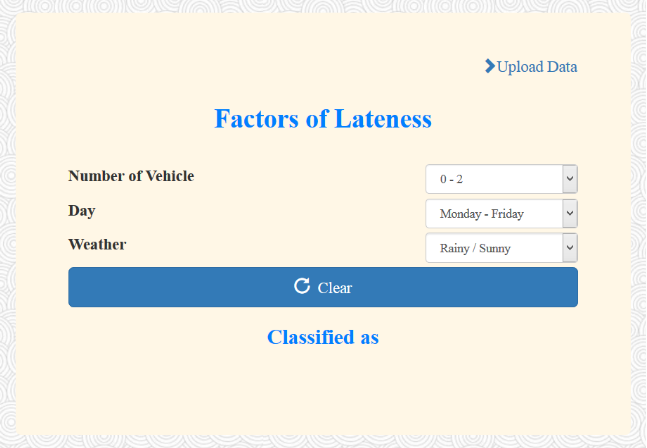
TABLE III. GAIN RATIO VALUE

Attribute	Gain Ratio
Sex	0.00081486
Age	0.00041352
Level	0.02226304
Position	0.01389266
Distance	0.06741947
Num_vehicle	0.07563392
Day	0.00269039
Weather	0.000015647

Based on Table III, the attribute of num_vehicle attribute is the most influential because it obtains the highest Gain Ratio value. While the attribute with the lowest Gain Ratio value is the weather which means only has a small effect.

B. *Deployment*

After evaluation phase, we get 1 best performance algorithm that is Random Forest with parameter pruning and pre-pruning. The resulting pattern of this algorithm is used to make the GUI as one solution to solve the problems that occur in the object of research. GUI is a web application built with PHP programming language.

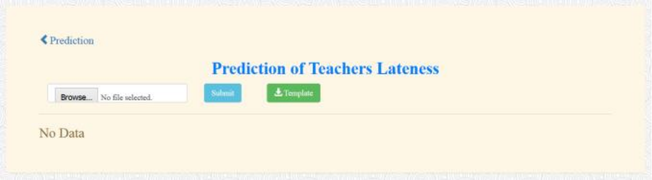


Factors of Lateness

Number of Vehicle: 0 - 2
Day: Monday - Friday
Weather: Rainy / Sunny

Classified as

Fig. 4. GUI (1).



Prediction of Teachers Lateness

No file selected.

No Data

Fig. 5. GUI (2).

Figure 4 and 5 are a view of a GUI that has been created. Figure 4 is the view to make predictions, where the user only needs to select the attribute and will get the result of classification. Figure 5 is a view to upload an excel file so that it can know how much the accuracy of the GUI.

VI. CONCLUSION

Based on the result of experiment that has been obtained, it can be concluded that the attribute that has the most influence on teachers' delay factor is num_vehicle. The use of pruning and pre-pruning successfully improved the performance of Algorithm C4.5 and Random Forest. Random Forest with pruning and pre-pruning obtains the AUC value and the highest accuracy value. And based on the results of different test using Parametric T-Test method, 3 significant differences were founded, they are: C4.5 with pre-pruning and pruning parameters and Random Forest without pre-pruning and pruning parameters, C4.5 without pre-pruning and pruning parameters and Random Forest with pre-pruning and pruning parameters, and Random Forest with pre-pruning and pruning parameters and Random Forest without pre-pruning and pruning parameters. For further research, the research can be developed using feature selection or PSO (Particle Swarm Optimization) as well as increasing the number of attributes with the attributes of long journey and the state of the road. Data duration can also be added, for example: 6 months - 1 year. in this research showed teacher delay was caused by long distance while the middle one on Monday, Friday, and Friday was confirmed late. The teacher's delay in this study is often done by teachers who have a vehicle compared to those who do not have a vehicle.

ACKNOWLEDGMENT

Thanks to Kalam Kudus III Christian School Jakarta who has given me permission to do research and give permission to use various research instruments at school. And also, to the teachers who have given permission for the data can be used for research.

REFERENCES

[1] Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan dan Kebudayaan Republik Indonesia, Keterlambatan [Online] Retrivied from https://kbbi.kemdikbud.go.id/entri/keterlambatan, accessed on 25 November 2017.

[2] Haryati. L, Upaya meningkatkan disiplin guru dalam kehadiran mengajar di kelas melalui penerapan "Reward and Punishment". MEDIA DIDAKTIKA, vol. 2(2), pp. 191–200, 2016.

[3] Sariana, "Upaya meningkatkan disiplin guru dalam kehadiran mengajar di kelas melalui waskat kepala sekolah pada smp negeri 4 rimba melintang kabupaten rokan hilir". Perspektif Pendidikan Dan Keguruan, vol. VIII(1), pp. 12–17, 2017.

[4] Karim. M., and Rahman. R.M, "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing". Journal of Software Engineering and Applications, (6), pp. 196–206, 2013.

[5] Wajhillah. R, "Optimasi algoritma klasifikasi c4.5 berbasis particle swarm optimization untuk prediksi penyakit jantung". SWABUMI, vol. I(1), pp. 26–36, 2014.

[6] Defiyanti. S, and Pardede. D.L.C, "Perbandingan kinerja algoritma id3 dan c4.5 dalam klasifikasi spam-mail". ReCALL, 2008.

[7] Sewaiwar. P, and Verma. K.K, "Comparative Study of Various Decision Tree Classification Algorithm Using WEKA". International Journal of Emerging Research in Management & Technology, vol. 4(10), pp. 87–91, 2015.

[8] Georgina. O, Alhasan. J, and Abdullahi. M.B, "Classification of Crime Data for Crime Control Using C4.5 and Naïve Bayes Techniques". International Journal of Mathematical Analysis And Optimization: Theory And Applications, pp. 139–153, 2017.

[9] Thaseen. S, and Kumar. C.A, "An analysis of supervised tree based classifiers for intrusion detection system". Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering(PRIME), pp. 294–299, 2013.

[10] Kalmegh. S.R, "Comparative Analysis of WEKA Data Mining Algorithm RandomForest, RandomTree and LADTree for Classification of Indigenous News Data". International Journal of Emerging Technology and Advanced Engineering, vol. 5(1), pp. 507–517, 2015.

[11] Rizal, "Komparasi Algoritma Klasifikasi Data Mining Untuk Memprediksi Penyakit Tuberculosis (TB)", 2013.

[12] Han. J, Kamber. M, and Pei. J, Data Mining: Concepts and Techniques (3rd ed.). (San Francisco: Morgan Kaufmann), 2012.

[13] Witten. I.H, Frank. E, and Hall. M.A, Data Mining: Practical machine learning tools and techniques (3rd ed.). (Burlington: Morgan Kaufmann), 2011.

[14] Berry. M.J.A, and Linoff. G.S, Data mining techniques: for marketing, sales, and customer relationship management (2nd ed.). (Indiana: Wiley Publishing), 2004.

[15] Anggarwal. C.C, Data Mining: The Textbook. (Switzerland: Springer), 2015.

[16] Yu. L, Chen. G, Koronios. A, Zhu. S, and Guo. X, Application and Comparison of Classification Techniques in Controlling Credit Risk. (Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications), pp. 111-145, 2007.

[17] Wu. X, and Kumar. V, The Top Ten Algorithms in Data Mining. (Boca Raton: CRC Press), 2009.

[18] Shajahaan. S.S, Shanthi. S, and Manochitra. V, "Application of Data Mining Techniques to Model Breast Cancer Data". International Journal of Emerging Technology and Advanced Engineering, vol. 3(11), pp. 1–8, 2013.

[19] Pfahringer. B, "Random model trees: an effective and scalable regression method", 2010.

[20] Random Forest (tm), RandomForests. [Online] Retrivied from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm, accessed on 31 Desember 2017.

[21] Kothari. C.R, Research Methodology: Methods & Techniques (2nd ed.). (New Delhi: New Age International Publishers), 2004.

[22] Badan Meteorologi Krimatologi dan Geofisika, Data Online Pusat Database BMKG. [Online] Retrivied from http://dataonline.bmkg.go.id/data_iklim, accessed on 30 November 2017.