2021 International Conference on Artificial Intelligence and Big Data Analytics (ICAIBDA) | 978-1-6654-0890-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIBDA53487.2021.9689701

Comparative Study of J48 Decision Tree Classification Algorithm, Random Tree, and Random Forest on In-Vehicle Coupon Recommendation Data

Dicky Rahma Hermawan[ab1], Mohamad Fahrio Ghanial Fatihah[ab2], Linda Kurniawati[bc3], Afrida Helen [ab4]
[a]*Department of Computer Science, Universitas Padjadjaran*
[b]*Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran*
[c]*Department of Business Administration, Universitas Padjadjaran*
Sumedang, Indonesia
e-mail: [1]dicky19002@mail.unpad.ac.id, [2]mohamad19001@mail.unpad.ac.id, [3]linda.kurniawati@unpad.ac.id, [4]helen@unpad.ac.id

Abstract— **Coupons are one of the media used to increase sales and invite customers to repurchase products. A study to investigate the effectiveness of the distribution of coupons, especially coupons for restaurants and bar, can be carried out by collecting data through an in-vehicle survey. The data can then be analyzed using classification techniques in data mining. This paper presents a classification on the problem of in-vehicle coupon recommendation to determine the decision of coupon acceptance through the J48, Random Tree, and Random Forest decision tree classification algorithm. The dataset used consists of 23 attributes including the class Y attribute which indicates the receipt of coupons by customers. The performance of the three algorithms is evaluated to determine the best classification algorithm by looking at accuracy, time to build the model, and other variables that appear in the class classification experiment. The results reveal that the Random Tree classification algorithm takes the least amount of time (0.28 seconds) and has the lowest accuracy (67.38%). The J48 algorithm is more accurate than the Random Tree algorithm (72.79%) but takes significantly longer time (0.36 seconds). The Random Forest technique has the best accuracy (77.0%), but the time it takes for model creation is substantially longer than the Random Tree and J48 algorithms (10.89 seconds).**

Keywords—classification, data mining, J48, random tree, random forest

I. Introduction

Coupons are one of the media used to increase sales. Coupons are distributed with the intention of inviting potential customers to make more purchases. The coupon distribution target can be in the form of regular customers or non-customers. When a customer makes a trial purchase on a product being sold, it is also expected that the product will continue to be purchased by the customer. Coupons can also be distributed to specific target customers, for example the best customers with pre-set preferences. This is considered to have built a lot of customer loyalty [1]. Several studies have attempted to investigate how the attitude of potential customers when a coupon is given to them as an antecedent variable and whether the coupon will be used. To collect customer data, especially in different scenarios, the questionnaire method is widely used. Through the data collected, an analysis was carried out to investigate the correlation of the response variations from customers to the effectiveness of coupon distribution. In case of analyzing data and concluding it, data mining is an effective tool.

Data mining is a method that aims to find and examine a structured pattern in the data to gain an understanding. The basic principles in data mining include analyzing data through various directions, categorizing, and finally making conclusions from visible patterns. The pattern can be broken down into information to predictions. In the case of machine learning and data mining, the knowledge structures and structural descriptions obtained in a data are as important as their ability to sample new data. In this case, people use data mining not only to make predictions, but to gain new knowledge that emerges from a dataset [2].

Data mining can also be referred to as *knowledge discovery in database* (KDD) which follows the following steps: data cleaning, data integration, data selection, data transformation, data mining, pattern evolution, knowledge evolution, and data reduction. There are also many techniques in data mining, including classification, clustering, data pre-processing, pattern recognition, association, and visualization. Classification is a process in data mining to describe and distinguish data classes and concepts through searching for classification models [3]. The purpose of the classification is to use a model that has been successfully searched to determine the class of unknown data labels. Currently, there are several classifier algorithms available. The classifier that is commonly used and has good performance is a decision tree [4]. Decision trees are expressive enough to model many data partitions that cannot be achieved in classifiers such as logistic regression and Support Vector Machines (SVM) which only rely on one decision scope. Decision trees are flexible in processing data with a combination of real types and categorical features and attributes that have missing data.

Classifiers such as J48, Logistic Model Tree (LMT), Random Tree, Simple Cart, Random Forest, and Reduced Error Pruning (REP) Tree are part of a decision tree and are used for the purpose of classifying datasets. In the decision tree, the results of the analysis are in the form of sequential rules that lead to a certain class or value and form a tree structure. A positive classification is obtained if there is a rule path to a positive leaf [5]. This can be used to derive conclusions from datasets that have classes with dichotomous labels such as yes and no.

In this study, we use the in-vehicle coupon recommendation dataset as a data source in conducting comparative analysis on decision tree-based classification

978-1-6654-0890-5/21/$31.00 ©2021 IEEE

algorithms, namely J48, Random Tree, and Random Forest. We are comparing these three data mining classifiers to attain the most accurate decision tree classifiers for this type of case.

II. Research Methods

A. J48 Algorithm

J48 is a form of classifier that uses the C4.5 algorithm and is part of the classification method in data mining. The C4.5 method is a well-known and commonly used technique for categorizing data with numerical and categorical properties. The results of the classification procedure within the sort of rules are frequently utilized to estimate the value of the new record's discrete type property. The C4.5 algorithm is a progression of the ID3 algorithm. The development is carried out in terms of the ability to overcome missing data, the ability to handle continuous data, and the ability to prune. The C45 algorithm has an advantage over the ID3 algorithm. The advantage is in the way it gathers data. Because C4.5 employs the gain ratio as a metric of attribute selection, C4.5 is superior [6].

The J48 method works on the principle of dividing data into ranges based on attribute values for items in the training data set. Missing values, which are values for elements that can be predicted based on what is known about attribute values in other rows, are ignored by the J48 method [7].

J48 uses a greedy technique, in which decision trees are built by recursively separating attributes from top to bottom, with the topmost attribute being the most influential attribute of the attributes beneath it. J48 employs the pessimistic pruning approach, which determines whether to prune a portion of the tree depending on the expected error rate. The J48 algorithm starts with a root node, which is then subdivided into another section of the node as a result of evaluating the attribute variable to see if it meets the test value. If the test result is a node that can be tested again, it is referred to as a branch; if it cannot be tested again or is the final result, it is referred to as a leaf, also known as a label or class [10].

The stages of the J48 algorithm in the construction of a decision tree are as follows:

1. Make attribute the root attribute.

2. For each value, make a branch.

3. Separate the cases into branches.

4. Repeat the process for each branch until all cases on that branch have the same class.

The largest gain value of the existing attributes is used to choose an attribute as the root. The following equation is used to compute the gain [8]:

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

Where:

S = set of instances

A = attribute

n = number of partitions of A

|Si| = number of cases on partition i

|S| = number of cases in S

The basic formula for the entropy is as follows:

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2 pi \quad \dots\dots\dots\dots\dots (2)$$

Where:

S = case set

A = feature

n = number of partition S

pi = proportion of Si to S

B. Random Forest Algorithm

Random Forest is a classifier made up of a series of tree-structured classifiers {h (x, k) k=1, 2, ….}, where {Θk} is a uniformly distributed random vector and each tree awards one vote unit to the most popular class at input x. Random Forest is made up of a huge number of individual decision trees that work together as a set, as the name implies. In Random Forest, each tree generates a class prediction, with the most votes being the model prediction [8].

The key of Random Forest is that each model has a low correlation. An uncorrelated model can generate ensemble forecasts that are more accurate than any individual prediction, just as low-correlated assets (such as stocks and bonds) combine to create a portfolio that is larger than the sum of its parts. The trees will shield one other from each other's errors if they don't continually make the same mistakes. Therefore, an uncorrelated model can produce ensemble predictions that are more accurate than any single prediction. While some trees will be incorrect, many others will be correct. As a result, the trees might move in the same direction as a group. Random Forest constructs each tree using two methods: bagging and feature randomness, to generate a forest of uncorrelated trees [9].

Because decision trees are highly dependent on the data they are trained on, even little modifications to the training set might result in drastically different tree architectures. Random Forest takes use of this by allowing each tree in the data set to be randomly sampled via replacement, resulting in a new tree, this process known as bagging (bootstrap aggregation). When splitting a node in a normal decision tree, it considers all candidate features and chooses the one that creates the maximum separation between the observations at the left and right nodes. Each tree in the Random Forest, on the other hand, can only choose from a random subset of features. This causes greater variety among the model's trees, which leads to a reduced correlation between them and increased diversification, this process known as feature randomness.

Trees are not only trained on diverse data sets (bagging) in the Random Forest algorithm, but they also use distinct features to make judgments (feature randomness). As a result, an uncorrelated tree emerges, which both supports and protects each other from errors.

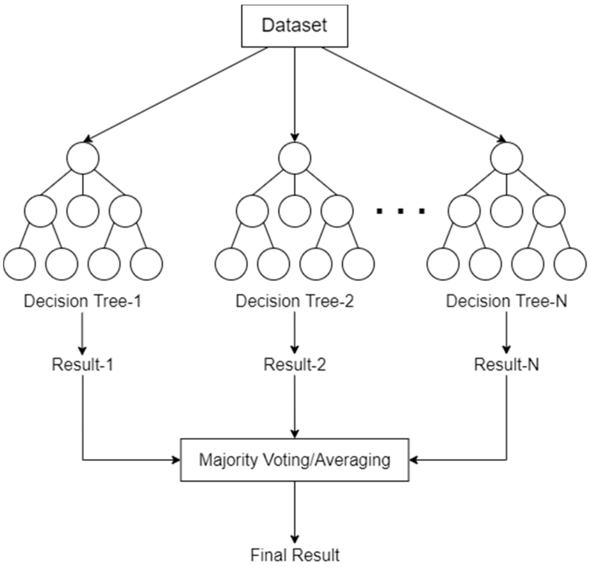

Fig. 1. Diagram of Random Forest algorithm

C. *Random Tree Algorithm*

Random Tree is a supervised classifier that produces many unique learners. The Random Tree algorithm is a type of ensemble learning algorithm. It builds a decision tree using the principle of bagging to generate a random set of data. Each node in a traditional tree is split using the best split across all variables.

The Random Tree algorithm can deal with both classification and regression problems. Leo Breiman and Adele Cutler invented this algorithm. A forest is an ensemble of prediction trees known as a random tree. The following is how the classification works: Random Tree takes an input feature vector, classifies it with each tree in the forest, and then assigns the class label with the most "votes" to the input feature vector. The Random Tree answer in regression is the average of all the responses from all the trees in the forest.

Random Tree is a machine learning algorithm that combines two algorithms: a single model tree and the Random Forest concept. A model tree is a decision tree that has a linear model for each leaf that is optimized for the local subspace specified by that leaf. Random Forests have been demonstrated to considerably increase the performance of single decision trees, using two randomization methods used to generate tree diversity. As in bagging, the training data is first sampled using a substitute for every single tree. Second, instead of computing the best split for each node all the time, when building a tree, only a random subset of all attributes is considered for every node, and thus the best alternative for that subset is calculated. Random Tree takes a different strategy, splitting the median of several attributes to roughly offset the trees. Recently, the approximate approach for calculating medians was described. As long as the data is close to the median, this process only requires two linear scans [10][11].

III. METHODOLOGY

The flowchart to find the model in this study using data mining for each classifier is as follows.

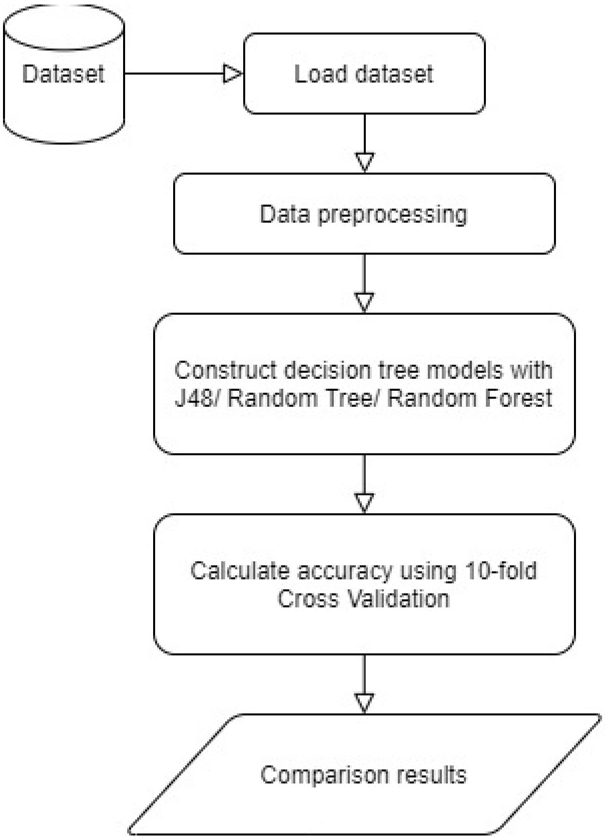
Dataset
Load dataset
Data preprocessing
Construct decision tree models with J48/ Random Tree/ Random Forest
Calculate accuracy using 10-fold Cross Validation
Comparison results

Fig. 2. Data mining process flowchart

The method proposed for classification of in-vehicle coupon recommendation dataset is implemented using a computer with AMD Radeon R2 with 2GB of memory, 4 GB DDR3-SDRAM memory, and AMD E2 1.5GHz processor.

A. *Dataset Preparation*

The dataset used in this study was sourced from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation) which was collected through a survey on Amazon Mechanical Turk, a crowdsourcing website that owned by Amazon. There are 12,684 rows of data and 26 attributes in the dataset. The problem of this classification experiment is to see whether potential customers will accept the coupons offered under certain circumstances, such as the place of validity of the coupons, demographic attributes, and certain contextual attributes. For a positive answer which means that the potential customer will 'redeem the coupon immediately' or 'exchange it later but before the expiration date' is labeled with 'Y = 1'. For negative answers which mean the potential customer is 'not interested in receiving or exchanging coupons' is labeled with 'Y = 0'. The list of each attribute and their meaning are as follows.

- destination: Driving destination
- passenger: Passengers in the vehicle
- weather: Weather conditions when driving
- temperature: Temperature (°F) when driving
- time: Time when driving
- coupon: Coupon exchange type

- expiration: Does the coupon expire in one day or in two hours?
- gender: Gender of the driver
- age: Age of the driver
- maritalStatus: Marital status of the driver
- hasChildren: Does the rider have children?
- education: Education status of the driver
- occupation: Job role of the driver
- income: Yearly income of the driver
- car: Type of the vehicle used when driving
- Bar: How many times does the driver go to the bar every month?
- CoffeeHouse: How many times does the driver go to the coffeehouse each month?
- CarryAway: How many times does the driver order take away food each month?
- RestaurantLessThan20: How many times have the driver go to a restaurant that averaged less than $20 per person per month?
- Restaurant20To50: How many times have the driver go to a restaurant that averaged less than $20 - $50 per person per month?
- toCouponGEQ5min: Does it take more than 5 minutes to get to the restaurant/bar to redeem the coupon?
- toCouponGEQ15min: Does it take more than 15 minutes to get to the restaurant/bar to redeem the coupon?
- toCouponGEQ25min: Does it take more than 25 minutes to get to the restaurant/bar to redeem the coupon?
- directionsame: Is the restaurant/ bar in the same direction as the current destination?
- direction_opp: Is the restaurant/ bar in a different direction from its current destination?
- Y: Will the coupon be accepted and exchanged by the driver?

B. *Data Preprocessing*

First, data preprocessing on the dataset is carried out to improve the quality of the data. Things that are done at this stage include:

- Remove redundant attributes such as an attribute containing only one unique value (*toCouponGEQ5min*) and an attribute whose value is complementary to another attribute (*directionsame* with *direction_opp*).
- Removing attribute that have a very large amount of empty data (> 90%) (*car*).
- There are several types of attributes that are recognized as strings with several different data values < 10. These attributes are then converted into nominal type.
- There are several attributes that have an empty data row (missing). Therefore, the attributes with empty data are filled with the value that has the highest number of occurrences.

The remaining attributes are 23 and have nominal type. This data is ready to be used for the classification stage.

C. *Training the Model and Testing Using 5-fold, 7-fold, and 10-fold Cross Validation*

The preprocessed data are then loaded to train data mining algorithms, namely J48, Random Tree, and Random Forest for classification purposes. Parameters for each algorithm are set to default values.

After training, the classification model is then evaluated using stratified k-fold cross validation. This technique works by partitioning the dataset into k parts. Each part is used as test data and the rest is used as training data until k repetitions are completed. This time, we chose values of k=5, k=7, and k=10. In the field of applied machine learning, k=10 is a popular choice. 10-fold cross validation is carried out because it is suitable to avoid biased results and can produce good classification results. Hyperparameters tuning is possible only with original training set. This allows keeping test set as an unseen dataset for selecting the final model. Through this technique, all the correct or incorrect classification results from each iteration will be recorded and then the classification accuracy will be generated. The best classifier is known by comparing the accuracy values of each algorithm.

D. *Search for the Most Important Attributes (Feature Selection)*

After evaluating the model, feature selection will be carried out to find the attributes that are most important or have the most impact on the classification model. Feature selection was done by calculating the information gain. The information gain (also known as entropy) for each attribute of the output variable was calculated. The values for the entries range from 0 (no information) to 1 (a lot of information) (maximum information). Information gain value is higher for attributes that contribute more information, while it is lower for attributes that do not contribute much information.

IV. Results and Discussion

Test results with 5-fold validation can be seen in Table I below.

TABLE I. Table for classification performance, accuracy, and training time for creating respective models using 5-fold cross validation

Classifier	Correct	Incorrect	Accuracy	Time (seconds)
J48	9,174	3,510	72.33%	0.35
Random Forest	9,665	3,019	76.20%	3.18
Random Tree	8,475	4,209	66.82%	0.05

TABLE II. Confusion Matrix Table for The Entire Decision Tree using 5-Fold Cross Validation

Classifier	a	b	Parametric Variable
J48	3,302	2,172	0
	1,338	5,872	1

Classifier	a	b	Parametric Variable
Random Forest	3,647	1,827	0
	1,192	6,018	1
Random Tree	3,438	2,036	0
	2,173	5,037	1

TABLE III. TABLE FOR CLASSIFICATION PERFORMANCE, ACCURACY, AND TRAINING TIME FOR CREATING RESPECTIVE MODELS USING 7-FOLD CROSS VALIDATION

Classifier	Correct	Incorrect	Accuracy	Time (seconds)
J48	9,211	3,473	72.62%	0.1
Random Forest	9,718	2,966	76.62%	2.69
Random Tree	8,369	4,315	65.98%	0.04

TABLE IV. CONFUSION MATRIX TABLE FOR THE ENTIRE DECISION TREE USING 7-FOLD CROSS VALIDATION

Classifier	a	b	Parametric Variable
J48	3,301	2,173	0
	1,300	5,910	1
Random Forest	3,655	1,819	0
	1,147	6,063	1
Random Tree	3,409	2,065	0
	2,250	4,960	1

Test results with 7-fold validation can be seen in Table III below. Test results with 10-fold validation can be seen in Table V. Table VI shows the performance of decision tree in this study.

TABLE V. TABLE FOR CLASSIFICATION PERFORMANCE, ACCURACY, AND TRAINING TIME FOR CREATING RESPECTIVE MODELS USING 10-FOLD CROSS VALIDATION

Classifier	Correct	Incorrect	Accuracy	Time (seconds)
J48	9,233	3,451	72.79%	0.29
Random Forest	9,778	2,906	77.09%	3.1
Random Tree	8,547	4,137	67.38%	0.14

TABLE VI. CONFUSION MATRIX TABLE FOR THE ENTIRE DECISION TREE USING 10-FOLD CROSS VALIDATION

Classifier	a	b	Parametric Variable
J48	3,311	2,163	0
	1,288	5,922	1
Random Forest	3,704	1,770	0
	1,136	6,074	1
Random Tree	3,488	1,986	0
	2,151	5,059	1

TABLE VII. TABLE OF ATTRIBUTES RANKING

Rank	Value	Attributes
1	0.050799	coupon
2	0.015898	CoffeeHouse
3	0.012838	passanger
4	0.01237	destination
5	0.01218	expiration
6	0.01002	time
7	0.007814	weather
8	0.007665	toCoupon_GEQ25min
9	0.006874	occupation
10	0.004818	toCoupon_GEQ15min
11	0.004511	Bar
12	0.003854	Restaurant20To50
13	0.003591	age
14	0.003186	temperature
15	0.002821	income
16	0.002779	maritalStatus
17	0.002477	CarryAway
18	0.00239	education
19	0.001496	has_children
20	0.001395	gender
21	0.001327	RestaurantLessThan20
22	0.000153	direction_same

From Table 7, it can be seen that the test using 10-fold cross-validation has better accuracy than 5-fold and 7-fold cross-validation. Table 5 shows all the results after testing using stratified 10-fold cross validation. It is known that Random Forest has the highest accuracy (77.09%), and Random Tree has the lowest accuracy (67.38%). However, Random Tree has the fastest time for model training (0.14 seconds) while Random Forest has the slowest (3.1 seconds). Based on all the results obtained, it is found that Random Forest has the highest accuracy results, but the training time is far longer. From the attribute ranking table, it shows that coupon is the most influential attribute, with an information gain value of 0.050799.

V. CONCLUSION

This study shows how to classify coupon recommendation data in vehicles using three data mining decision tree classification algorithms: J48, Random Forest, and Random Tree. The classification was done using stratified 10-fold cross-validation and based on the given data set. The results reveal that the Random Tree classification algorithm takes the least amount of time (0.14 seconds) and has the lowest accuracy of the three algorithms (67.38%). The J48 algorithm is more accurate than the Random Tree algorithm (72.79%), but the time it takes is significantly longer (0.29 seconds). The Random Forest technique has the best accuracy (77.09%), but the time it takes to develop the classification model is substantially longer than the Random Tree and J48 algorithms (3.1 seconds). Thus, if less time is required, the Random Tree method can be used, and if great accuracy is required, the Random Forest approach can be employed. From the results, it can also be concluded that coupon is the most influential attribute with an information gain value of 0.050799 so that the type of coupon is the most considered by customers in choosing whether they will accept the coupon or not.

ACKNOWLEDGMENT

The Author thanks to the Research Center for Artificial Intelligence and Big Data Universitas Padjadjaran Indonesia that supported the study. This work also supported by Conference Grant funded by Directorate of Research & Community Service, Universitas Padjadjaran Contract No: 2574/UN6.3.1/TU.00/2021.

REFERENCES

[1] S. Barat and L. Ye, "Effects of Coupons on Consumer Purchase Behavior: A Meta-Analysis," Journal of Marketing Development and Competitiveness, vol. 6, no. 5, pp. 131-145, 2012.

[2] S. R. Kalmegh, "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, no. 1, 2015.

[3] I. N. Yulita, M. I. Fanany, and A. M. Arymurthy, "Combining deep belief networks and bidirectional long short-term memory: Case study: Sleep stage classification", 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 1-6), 2017.

[4] S. Kiranmai and A. J. Laxmi, "Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy," Protection and Control of Modern Power Systems, vol. 3, no. 29, pp. 1-12, 2018.

[5] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, K. E and P. MacNeille, "A bayesian framework for learning rule sets for interpretable classification," Journal of Machine Learning Research, vol. 18, pp. 1-37, 2017.

[6] U. Bashir and M. Chachoo, "Performance Evaluation of J48 and Bayes Algorithms for Intrusion Detection System," International Journal of Network Security & Its Applications (IJNSA), vol. 9, no. 4, 2017.

[7] N. Saravanan and V. Gayathri, "Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48)," International Journal of Computational Intelligence and Informatics, vol. 7, no. 4, 2018.

[8] M. Schounlau and R. Y. Zou, "The random forest algorithm for statistical learning," The Stata Journal: Promoting communications on statistics and Stata, vol. 20, no. 1, 2020.

[9] I. M. Wildani,and I. N. Yulita, "Classifying botnet attack on internet of things device using random forest", IOP Conference Series: Earth and Environmental Science (Vol. 248, No. 1, p. 012002), 2019

[10] W. Gata, G. Y. E. Patras, R. Hidayat, R. Fatmasari, S. Tohari, B. and N. K. Wardhani, "Prediction of Teachers' Lateness Factors Coming to School Using C4.5, Random Tree, Random Forest Algorithm," in 2nd International Conference on Research of Educational Administration and Management (ICREAM 2018), Bandung, 2019.

[11] A. K. Mishra and B. K. Ratha, "Study of Random Tree and Random Forest data mining algoritms for microarray data analysis," International Journal on Advanced Electrical and Computer Engineering (IJAECE), vol. 3, no. 4, 2016.