Comparative analysis of decision tree algorithms: Random forest and C4.5 for airlines customer satisfaction classification

W Baswardono1,*, D Kurniadi1, A Mulyani1 and D M Arifin2

1 Department of Informatics, Sekolah Tinggi Teknologi Garut, Jalan Mayor Syamsu 1, Garut 44151, Indonesia
2 Department of Industrial Engineering, Sekolah Tinggi Teknologi Garut, Jalan Mayor Syamsu 1, Garut 44151, Indonesia

*wiyo.b@sttgarut.ac.id

Abstract. This article aims to a comparative analysis of decision tree algorithms between random forest and C4.5 for airlines customer satisfaction classification. The comparative study predicts both algorithms have better accuracy, precision, recall AUC (area under the curve) for analyzing data set of customer satisfaction on airlines, which are useful for later if have some same kind set of data set and problem. In this particular comparative analysis, first, need to select the dataset and transform so it can be used for data mining technique classification after choosing the algorithm to analyze the data set. The analyzing of the dataset it will go through validation, testing, and also result for each algorithm used. Then will compare the result from each algorithm, to determine which algorithm are best to use in this particular dataset or problem for customer satisfaction for airlines. The results of the comparative analysis are the best alternative algorithm choice for use in airline customer satisfaction classifications. For this comparison, the Random forest algorithm has a better result than the C4.5 algorithms.

1. Introduction

In the business world, especially business people, compete with each other to provide the best service for customers to create the best customer satisfaction. One of the biggest business services is airlines; the authors of the data get these data airlines from the Kaggle.com site, which contains a dataset for customer satisfaction along with the parameters.

The data will be tried out to find the level of accuracy of customer satisfaction. Classification is a technique for data collection after that to process the data, data mining science will be used [1–3]. Data Mining is a process for extracting to get important information that is implicit and previously unknown from data. To extract it, the data set must be selected as well as some data preprocessing techniques that can be used to produce quality data [4,5]. Data cleaning can be applied to eliminate noise and inconsistent data that is transformed first as part of the process after carrying out the process, it will be continued by selecting a certain method to process the existing data set.

It is chosen to use two kinds of algorithms consisting of Random Forest and C4.5. After the two algorithms are processed, it is expected that the results of the accuracy will be obtained from the data set, from the two results will be compared to find out which algorithm is more accurate and good according to other parameters as well. In the comparison in the validation and the test, the split test data will be shared for testing and training, moreover, in the condition of ten-fold cross-validation. Then after
the data will be compared according to their respective conditions. Then the results of these comparisons are the results of this experiment, the Random Forest and C4.5 comparison analysis for customer classification.

2. Method
The methodology used for the comparison of Random Forest and C4.5 analysis uses the KDD process stages consisting of data selection, data preprocessing/cleaning, data transformation, data mining, and Interpretation/Evaluation [6]. Random Forest and C4.5 being compared, which are the most popular decision algorithm in data mining for decision tree and also both can be used in the large data set.

2.1. Stages of research methods

In the research, the method starts from data collection (data is collected to be made into data sets later) after that data selection (making data sets/selection of data sets) if the data has been collected, then the data transformation is done when needed. After the data set is ready, then an algorithm is chosen, which will be used for later rule models for research that will be done using data mining using the chosen algorithm. To determine the results of the data, do validation and testing to determine the Accuracy level, precision, recall, error classification, and AUC from the results of the classification prediction. Finally, the analysis of the test results is done to ensure that the test results are by the discussion. The analysis is done by calculating the results of validation and testing (the level of accuracy, precision, recall, error classification and AUC (Area under Curve), which will then be compared between the predefined classification algorithms.

2.2. Decision tree
Decision tree learning is a learning method that is very popular and widely used practically [7]. This method is a method that seeks to find functions that have a discrete value and are resistant to data that has errors (data noise) and can learn disjunctive expressions.

2.3. Random forest
Random Forest is one method in the Decision Tree. Decision Tree or Decision Tree is a flow diagram shaped like a tree that has a root node that is used to collect data, an inner node that is in the root node that contains questions about the data and a leaf node that is used to solve problems and make a decision. Decision tree classifies a sample of data whose class is not yet known into existing classes. Using the
decision tree to avoid overfitting on a data set when achieving maximum accuracy. Also, the Random Forest Algorithm is an algorithm that is suitable for use for large data classifications [8].

2.4. **C4.5 Algorithm**

Is a group of Decision Tree algorithms. This algorithm has input in the form of training samples and samples. Training samples in the form of sample data will be used to build a tree that has been tested properly. While samples are data fields which we will use as parameters in classifying data. C4.5 algorithms is a development from ID3. The development that exists in the C4.5 algorithm is that it can overcome overcoming continuous data, missing value, and training data [6].

3. **Results and discussion**

Discussion and results of comparative analysis research for customer satisfaction airlines classification

3.1. **Analyze data**

The data set of customer satisfaction airlines was obtained from Kaggle.com (US Airline Passenger Satisfaction) and had 129881 records and has 23 attributes. After pre-processing the data transform, column determination, filtering data that has no value, and unbalanced data checking. Then the distribution of variables and data classes is like the table below:

| Variable   | Field           | Type  | Value                      |
|------------|-----------------|-------|----------------------------|
| Y          | Satisfaction    | Nominal | Satisfied; dissatisfied   |
| X1         | Gender          | Nominal | Male; Female               |
| X2         | Customer_Type   | Nominal | Loyal; Disloyal           |
| X3         | Age             | Nominal | 7 .. 85                   |
| X4         | Travel_Type     | Nominal | Personal; Business        |
| X5         | Class           | Nominal | Eco; Eco Plus; Business   |
| X6         | Distance        | Nominal | 31 .. 4983                |
| X7         | Wi_fi           | Real   | 1 .. 5                    |
| X8         | Service         | Real   | 1 .. 5                    |
| X9         | Depart/arrive_time | Real | 1 .. 5                   |
| X10        | Convinient      | Real   | 1 .. 5                    |
| X11        | Booking         | Real   | 1 .. 5                    |
| X12        | Gate_location   | Real   | 1 .. 5                    |
| X13        | Beverages       | Real   | 1 .. 5                    |
| X14        | Boarding        | Real   | 1 .. 5                    |
| X15        | Seat_comfort    | Real   | 1 .. 5                    |
| X16        | Entertainment   | Real   | 1 .. 5                    |
| X17        | Baggage         | Real   | 1 .. 5                    |
| X18        | Check_in        | Real   | 1 .. 5                    |
| X19        | Inflight        | Real   | 1 .. 5                    |
| X20        | Cleanliness     | Real   | 1 .. 5                    |

The data used for internal analysis is 129881 records from the original data and 21 attributes after processing the data transform.

3.2. **Discussion**

Classification methods and algorithm will be evaluated on Accuracy, precision, recall, error classification, and AUC (Area Under Curve). The experiment to find out the comparison between the two algorithms the application used to help this experiment is Rapid Miner software. After the data is processed (data that has been processed with the algorithm specified) and produces a model. After that,
the model is carried out using Ten fold Cross-validation. Ten-fold Cross-Validation is the nested operator. The Operator has two processes, one for training and another for testing. The model that already train will be followed by testing [9–11].

The confusion matrix is one method that can be used to measure the performance of a classification method; one of them is to search for Accuracy, Error Classification, Precision, and Recall. For the confusion matrix, it is common to test for the accuracy model [5,12,13].

3.3. Analysis result comparison testing between random forest and C4.5

3.3.1. Random forest. Testing data with 129881 data and 21 attributes with random forest using the gain ratio criteria and voting strategy with confidence vote and assisted by Ten-Fold Cross Validation 80:20 and 90:10 produce like the data below:

| Split Data | Accuracy | Class Error | Precision | Recall | AUC   |
|------------|----------|-------------|-----------|--------|-------|
| 70:30      | 93.30    | 6.70        | 92.81     | 95.56  | 0.983 |
| 80:20      | 93.31    | 6.69        | 92.83     | 95.54  | 0.983 |
| 90:10      | 93.32    | 6.68        | 92.82     | 95.59  | 0.984 |

From the comparison of the data above which has been tested with a rapid miner, for the discussion here compared this algorithm will only discuss split data 90:10 because as seen in the table the data above has the best accuracy among the three data. Here Random Forest produces 100 Decision Tree as a comparison, from that all found the highest accuracy level.

For Accuracy, there are 45892 data which are predicted to be satisfied and in fact, satisfied; there are 63191 data predicted at dissatisfied and in fact dissatisfied, there are 4893 data which are predicted to be satisfied and in fact dissatisfied, there are 2916 data which are predicted dissatisfied and in fact satisfied. Then the value of accuracy is 93.32%.

3.3.2. C4.5. Testing data with 129881 data and 21 attributes with C4.5 using the gain ratio and 0.1 confidence criteria as well as helping Cross Validation Ten-Fold using fold 10 and stratified sampling with split data comparisons between 70:30, 80:20 and 90:10 produces like the data below:

| Split Data | Accuracy | Class Error | Precision | Recall | AUC   |
|------------|----------|-------------|-----------|--------|-------|
| 70:30      | 92.21    | 7.79        | 92.21     | 94.22  | 0.968 |
| 80:20      | 92.22    | 7.78        | 91.90     | 94.63  | 0.968 |
| 90:10      | 92.55    | 7.45        | 92.48     | 94.53  | 0.971 |

From the comparison of the data above which has been tested with the rapid miner, for the discussion here compared this algorithm will only discuss split data 90:10 because as seen in the table the data above has the best accuracy among the three data.

For Accuracy, there are 45696 data that are predicted to be satisfied and in fact satisfied; there are 62489 data predicted at dissatisfied and in fact dissatisfied, there are 5089 data which are predicted to be satisfied and in fact dissatisfied, there are 3618 data which are predicted dissatisfied and in fact satisfied. Then the value of accuracy is 92.55%.

3.4. Comparative result
After calculating the accuracy, calcifying error, precision, recall, and AUC, with the airline's dataset and rules specified above by comparing Random Forest and C4.5. Seen in the table below:
Table 4. Comparative algorithms random forest and C4.5 on split 90:10.

| Algorithm     | Accuracy | Class Error | Precision | Recall | AUC  |
|---------------|----------|-------------|-----------|--------|------|
| Random Forest | 93.32    | 6.68        | 92.82     | 95.59  | 0.984|
| C4.5          | 93.31    | 6.69        | 92.83     | 95.54  | 0.971|

In the analysis of each comparison of these algorithms, it turns out that high Accuracy data is data that uses data split 90:10. Then the comparison data table above uses that data. As seen for the accuracy of the comparison of data seen, Random Forest has Accuracy of 93.32% while C4.5 has 93.31%, the highest accuracy in this comparison is Random Forest.

Comparing ROC (Receiving operating characteristic) comparison can be done on rapid miner simultaneously based on two algorithms that are done, then generate curve ROC data as below.

![Figure 2. Graphic compare ROC.](image)

As seen in the curve between Random Forest and C4.5. Graphically looks almost the same because the calculation between these algorithms is almost the same method as generating the decision tree as a result by comparing it, of course, random forest produces more decision tree patterns so that it can be more accurate than C4.5.

4. Conclusion

Based on the results of comparative experiments for the Random Forest and C4.5 Tree on customer satisfaction airlines classifications can be taken as with the same dataset between Random Forest and C4.5 with split data between 70:30, 80:20 and 90:10. Split data 90:10 in the process, also with Ten Cross Validation, has the highest accuracy. With Random Forest, the data has an Accuracy of 93.32% while C4.5 has 92.55%. Random Forest has the highest accuracy.

With the same dataset between Random Forest and C4.5 with data split 9:10 for AUC (area under a curve) the highest value is in the random forest. This data is taken from the mean values of AUC (pessimistic) and AUC (optimistic) which are not displayed here. So what can be seen in the data if need a high AUC result then Random Forest is the right one to use.

Based on the results of a comparison experiment for the Random Forest and C4.5 on customer satisfaction airlines classifications recommendation for further works can be data sets can be tested again with different methods for different purposes. Such as clustering or classification but with different algorithms or different models and also different parameters are expected to produce different results for comparison. Also, optimization Data sets can be done, for example, with modified techniques such as example, the data set is selected using genetic algorithms or other algorithms before being compared with a certain technique.

Acknowledgment

Authors wishing to acknowledge Sekolah Tinggi Teknologi Garut that supports and funds this research publication.
References

[1] Lacic E, Kowald D and Lex E 2016 High Enough?: Explaining and Predicting Traveler Satisfaction Using Airline Reviews Proceedings of the 27th ACM Conference on Hypertext and Social Media - HT ’16 249–254

[2] Wan Y and Gao Q 2015 An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (IEEE) 1318–1325

[3] Kurniadi D, Sasmoko, Warnars H L H S and Gaol F L 2017 Software size measurement of student information terminal with use case point 2017 IEEE International Conference on Cybernetics and Computational Intelligence 164–169

[4] Hand D J 2013 Data Mining Based in part on the article “Data mining” by David Hand, which appeared in the Encyclopedia of Environmetrics, Encyclopedia of Environmetrics (Chichester, UK: John Wiley & Sons, Ltd)

[5] Kurniadi D, Abdurachman E, Warnars H L H S and Suparta W 2018 The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm IOP Conf. Ser. Mater. Sci. Eng. 434 012039

[6] Cynthia E P and Ismanto E 2018 Metode Decision Tree Algoritma C.45 Dalam Mengklasifikasi Data Penjualan Bisnis Gerai Makanan Cepat Saji Jurnal Ris. Sist. Inf. dan Tek. Inform. 31

[7] Harrison P A, Dunford R, Barton D N, Kelemen E, Martin-López B, Norton L, Termansen M, Hendriks K, Gómez-Baggethun E, Czúcz B, Garcia-Llorente M, Howard D, Jacobs S, Karlsen M, Kopperoinen L, Madsen A, Rusch G, van Eupen M, Verweij P, Smith R, Tuomasjukka D and Zulian G 2018 Selecting methods for ecosystem service assessment: A decision tree approach Ecosyst. Serv. 29 481–498

[8] Díaz-Uriarte R and Alvarez de Andrés S 2006 Gene selection and classification of microarray data using random forest BMC Bioinformatics 7 3

[9] Naik A and Samant L 2016 Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime Procedia Comput. Sci. 85 662–628

[10] Septiana Y, Kurniadi D, Mulyani A and Baswardono W 2018 Design of decision support system for blood analysis MATEC Web Conf. 197 03018

[11] Septiana Y 2018 Design of prototype decision support system for flood detection based on ultrasonic sensor MATEC Web Conf. 197 03017

[12] Kurniadi D, Mulyani A, Septiana Y and Aulawi H 2018 Estimated software measurement base on use case for online admission system IOP Conf. Ser. Mater. Sci. Eng. 434 012062

[13] Kunwar V, Chandel K, Sabitha A S and Bansal A 2016 Chronic Kidney Disease analysis using data mining classification techniques 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence) (IEEE) 300–305