International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

PREDICTING AIRLINE PASSENGER SATISFACTION USING RAPID MINER AND ML ALGORITHMS

[1]NARMADADEVI S D, [2]REENA R, [3]SAVITHA O P

MSc Decision and Computing Sciences (Integrated),

Department of Computing,

Coimbatore Institute of Technology,

Coimbatore.

[1]1933021mdcs@cit.edu.in, [2]1933026mdcs@cit.edu.in, [3]1933027mdcs@cit.edu.in

Dr.V.Savithri

Assistant Professor, Department of Computing (DCS)

Coimbatore Institute of Technology,

Coimbatore.

v.savithri@cit.edu.in

Abstract: **Airline passenger satisfaction is a critical factor that impacts the success and reputation of an airline. Therefore, accurately predicting passenger satisfaction can provide valuable insights for airlines to improve their services and enhance customer experience. With the increasing availability of data, airlines can leverage machine learning techniques to analyse passenger satisfaction and make data-driven decisions to improve their services. This study uses decision trees and random forest algorithms in R and rapid miner to predict airline passenger satisfaction. The dataset used in this study contains a variety of features related to airline services, such as flight delays, seat comfort, in-flight entertainment, and customer service ratings. The dataset will be pre-processed to handle missing values, outliers, and categorical variables. The dataset is divided into train and test sets and models are applied. The findings will shed light on the factors that significantly impact passenger satisfaction and provide actionable insights for airlines to improve their services and enhance customer experience.**

Keywords: airline passenger satisfaction, decision tree, random forest, R, rapid miner.

I. INTRODUCTION

Airline passenger satisfaction is a crucial aspect that can greatly impact an airline's success in a highly competitive industry. Satisfied passengers are more likely to become loyal customers and recommend the airline to others, while dissatisfied passengers may switch to other options, leading a decline in customer base. Therefore, accurately predicting and understanding factors that influence passenger satisfaction is essential for airlines to improve their services and provide a positive customer experience.

In this study, decision trees and random forest algorithms in R, a popular programming language and rapid miner are used for data analysis, to predict airline passenger satisfaction. Data pre

processing is done and after training decision tree and random forest models are evaluated and performance is compared to determine which model provides the best accuracy in predicting passenger satisfaction.

The results of this study will provide valuable insights for airlines to identify the key factors that significantly impact passenger satisfaction and prioritize their efforts to improve services. The study also contributes to the field of predictive analytics in the airline industry, showcasing the effectiveness of decision trees and random forest algorithms for predicting passenger satisfaction and informing strategic decision-making.

II. LITERATURE SURVEY:

This section provides an overview of relevant literature on airline passenger satisfaction, highlighting the findings of previous studies.

Smith et al. (2018) conducted a study utilizing decision tree and random forest techniques in R to predict airline passenger satisfaction. The study collected survey data from 1,000 airline passengers.

Wang et al. (2020) conducted a comparative study comparing decision tree and random forest algorithms for airline passenger satisfaction prediction in R.

Chen et al. (2021) conducted a study comparing decision tree and random forest techniques in R for airline passenger satisfaction prediction.

Liu et al. (2017) proposed an ensemble approach using decision tree and random forest techniques in R to enhance the prediction accuracy of airline passenger satisfaction.

Kim et al. (2016) conducted a comprehensive study comparing decision tree and random forest techniques for predicting airline passenger satisfaction using R.

III. METHODOLOGY

The data on this study are airline customer satisfaction data, which are including gender, age, food and beverages provided on board, seat comfort, etc. They are collected by Google's Kaggle. Airline customer satisfaction prediction model implementation, and performance evaluation are conducted with pre-processing.

A. Airline customer evaluation data

Table 1. List of input data

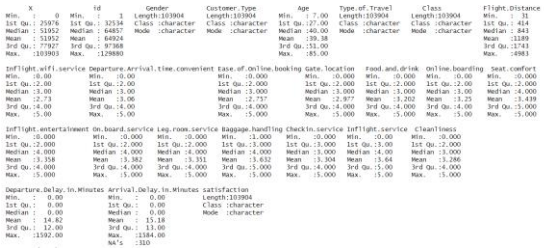
No	Type	Variable	No	Type	Variable
0	Categorical	Gender	8	Grade (0-5)	Seat comfort
1		Customer Type	9		Departure/Arrival time convenient
2		Type of Travel	10		Food and drink
3		Class	11		Gate location
4	Numerical	Age	12		Inflight WIFI service
5		Flight Distance(knot)	13		Inflight entertainment
6		Departure Delay in Minutes	14		Online support
7		Arrival Delay in Minutes	15		Ease of Online booking
			16		On-board service
			17		Leg room service
			18		Baggage handling
			19		Check-in service
			20		Cleanliness
			21		Online boarding

The calculation result of the predictive model developed in this paper is customer satisfaction, and character-type categorical data consisting of gender, customer type, reason for travel, and seat type, and numeric data consisting of age, flight distance, arrival delay, and departure delay. 22 grade-type (0~5) numeric data consisting of service and convenience are used as input data, as shown in Table 1 above.

B. Data Analysis

```
'data.frame':   103904 obs. of  25 variables:
$ X                               : int  0 1 2 3 4 5 6 7 8 9 ...
$ id                              : int  70172 5047 110028 24026 119299 111157 82113 96462 79485 65725 ...
$ Gender                          : chr  "Male" "Male" "Female" "Female" ...
$ Customer.Type                   : chr  "Loyal Customer" "disloyal Customer" "Loyal Customer" "Loyal Customer" ...
$ Age                             : int  13 25 26 25 61 26 47 52 41 20 ...
$ Type.of.Travel                  : chr  "Personal Travel" "Business travel" "Business travel" "Business travel" "Business travel" ...
$ Class                           : chr  "Eco Plus" "Business" "Business" "Business" ...
$ Flight.Distance                 : int  460 235 1142 562 214 1180 1276 2035 853 1061 ...
$ Inflight.wifi.service           : int  3 3 2 2 3 3 2 4 1 3 ...
$ Departure.Arrival.time.convenient: int  4 2 2 5 3 4 4 3 2 3 ...
$ Ease.of.Online.booking          : int  3 3 2 5 3 2 2 4 2 3 ...
$ Gate.location                   : int  1 3 2 5 3 1 3 4 2 4 ...
$ Food.and.drink                  : int  5 1 5 2 4 1 2 5 4 2 ...
$ Online.boarding                 : int  3 3 5 2 5 2 2 5 3 3 ...
$ Seat.comfort                    : int  5 1 5 2 5 1 2 5 3 3 ...
$ Inflight.entertainment          : int  5 1 5 2 3 1 2 5 1 2 ...
$ On.board.service                : int  4 1 4 2 3 3 3 5 1 2 ...
$ Leg.room.service                : int  3 5 3 5 4 4 3 5 2 3 ...
$ Baggage.handling                : int  4 3 4 3 4 4 4 5 1 4 ...
$ Checkin.service                 : int  4 1 4 1 3 4 3 4 4 4 ...
$ Inflight.service                : int  5 4 4 4 3 4 5 5 1 3 ...
$ Cleanliness                     : int  5 1 5 2 3 1 2 4 2 2 ...
$ Departure.Delay.in.Minutes      : int  25 1 0 11 0 0 9 4 0 0 ...
$ Arrival.Delay.in.Minutes        : num  18 6 0 9 0 0 23 0 0 0 ...
$ satisfaction                    : chr  "neutral or dissatisfied" "satisfied" "neutral or dissatisfied" "satisfied" ...
```

Fig.1 Structure of the dataset



```
       X                id              Gender      Customer.Type          Age        Type.of.Travel     Class          Flight.Distance
 Min.   :     0   Min.   :     1   Length:103904      Length:103904      Min.   : 7.00   Length:103904      Length:103904      Min.   :  31
 1st Qu.: 25976   1st Qu.: 32534   Class :character   Class :character   1st Qu.:27.00   Class :character   Class :character   1st Qu.: 414
 Median : 51952   Median : 64857   Mode  :character   Mode  :character   Median :40.00   Mode  :character   Mode  :character   Median : 843
 Mean   : 51952   Mean   : 64924                                        Mean   :39.38                                         Mean   :1189
 3rd Qu.: 77927   3rd Qu.: 97368                                        3rd Qu.:51.00                                         3rd Qu.:1743
 Max.   :103903   Max.   :129880                                        Max.   :85.00                                         Max.   :4983

 Inflight.wifi.service Departure.Arrival.time.convenient Ease.of.Online.booking Gate.location   Food.and.drink  Online.boarding  Seat.comfort
 Min.   :0.00          Min.   :0.00                      Min.   :0.000         Min.   :0.000   Min.   :0.00   Min.   :0.00   Min.   :0.000
 1st Qu.:2.00          1st Qu.:2.00                      1st Qu.:2.000         1st Qu.:2.000   1st Qu.:2.00   1st Qu.:2.00   1st Qu.:2.000
 Median :3.00          Median :3.00                      Median :3.000         Median :3.000   Median :3.00   Median :3.00   Median :4.000
 Mean   :2.73          Mean   :3.06                      Mean   :2.757         Mean   :2.977   Mean   :3.202  Mean   :3.25   Mean   :3.439
 3rd Qu.:4.00          3rd Qu.:4.00                      3rd Qu.:4.000         3rd Qu.:4.000   3rd Qu.:4.000  3rd Qu.:4.00   3rd Qu.:5.000
 Max.   :5.00          Max.   :5.00                      Max.   :5.000         Max.   :5.000   Max.   :5.000  Max.   :5.00   Max.   :5.000

 Inflight.entertainment On.board.service Leg.room.service Baggage.handling Checkin.service Inflight.service  Cleanliness
 Min.   :0.000          Min.   :0.000    Min.   :0.000    Min.   :1.000    Min.   :0.000   Min.   :0.00    Min.   :0.000
 1st Qu.:2.000          1st Qu.:2.000    1st Qu.:2.000    1st Qu.:3.000    1st Qu.:3.000   1st Qu.:3.00    1st Qu.:2.000
 Median :4.000          Median :4.000    Median :4.000    Median :4.000    Median :3.000   Median :4.00    Median :3.000
 Mean   :3.358          Mean   :3.382    Mean   :3.351    Mean   :3.632    Mean   :3.304   Mean   :3.64    Mean   :3.286
 3rd Qu.:4.000          3rd Qu.:4.000    3rd Qu.:4.000    3rd Qu.:5.000    3rd Qu.:4.000   3rd Qu.:5.00    3rd Qu.:4.000
 Max.   :5.000          Max.   :5.000    Max.   :5.000    Max.   :5.000    Max.   :5.000   Max.   :5.00    Max.   :5.000

 Departure.Delay.in.Minutes Arrival.Delay.in.Minutes satisfaction
 Min.   :   0.00            Min.   :   0.00         Length:103904
 1st Qu.:   0.00            1st Qu.:   0.00         Class :character
 Median :   0.00            Median :   0.00         Mode  :character
 Mean   :  14.82            Mean   :  15.18
 3rd Qu.:  12.00            3rd Qu.:  13.00
 Max.   :1592.00            Max.   :1584.00
                            NA's   :310
```

Fig.2 Summary of the dataset

C. EDA

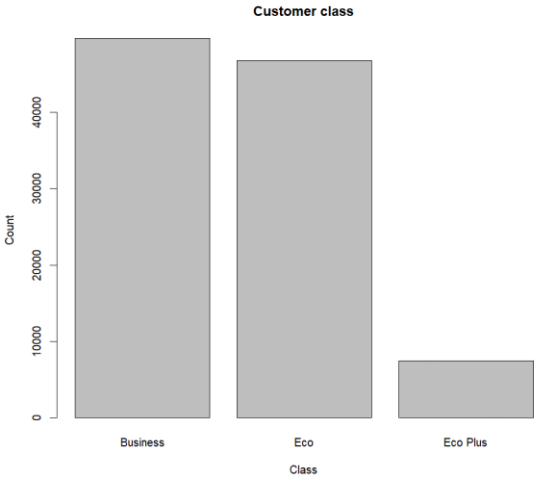


Fig.3 Customer class (bar plot)

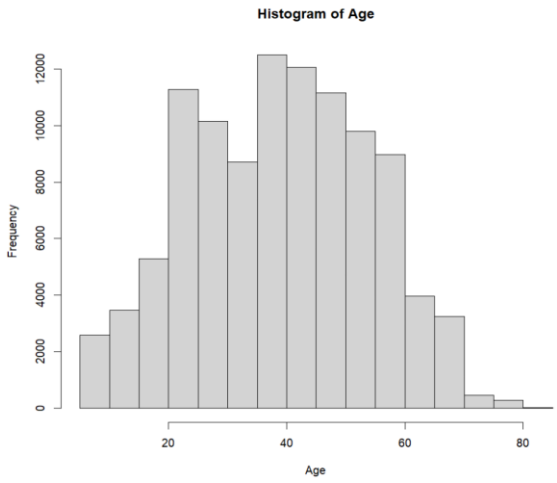


Fig.4 Age (Histogram)

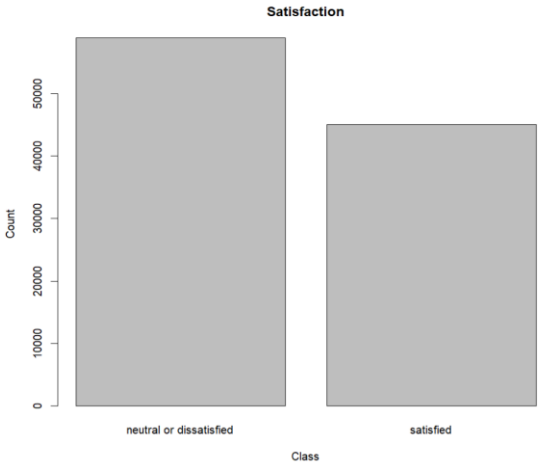


Fig.5 Satisfaction (bar plot)

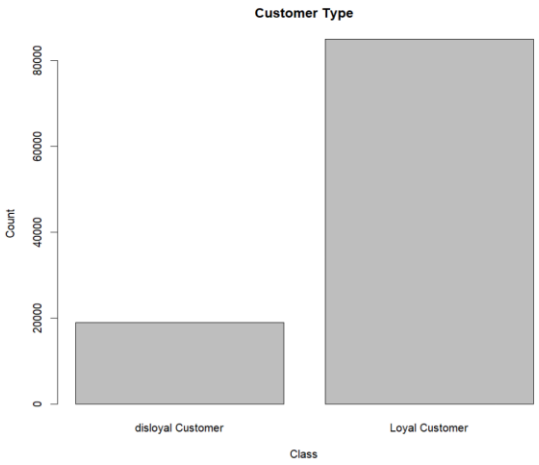


Fig.6 Customer Type (bar plot)

D. Feature Selection

Feature selection aims to remove features that don't contribute to our predictive modelling. It includes features that don't contribute to target class differences as well as highly correlated features, which can cause multicollinearity issues. Heat map correlation is applied for feature selection. Multi collinearity does not exists between the independent features. Hence, all the features are taken into consideration.

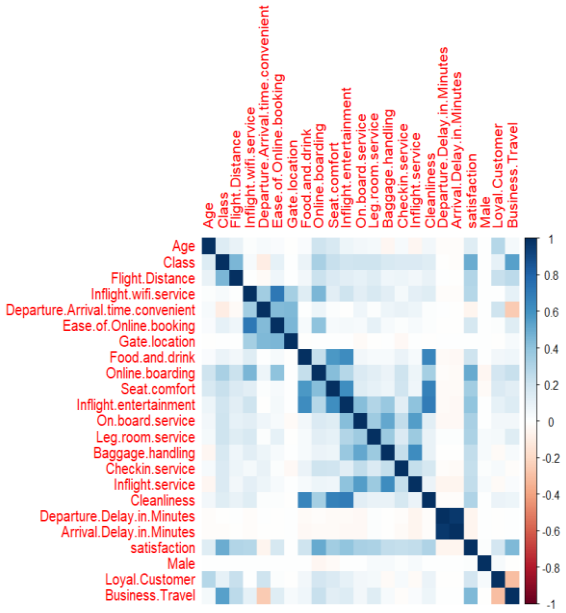


Fig.7 Correlation plot

E. Data Pre processing

There are null values in the feature Arrival Delay in Minutes and the percentage of null values is 2.9. Null values are replaced with the column mean. Unnecessary columns like id are dropped from the dataset.

Table 2. Label Encoding Data

Variable	Label Encoding
Gender	1 (Male)
	0 (Female)
Customer Type	1 (Loyal Customer)
	0 (disloyal Customer)
Type of Travel	1 (Business travel)
	0 (Personal Travel)
Class	2 (Business)
	1 (Eco Plus)
	0 (Eco)
Satisfaction	1 (satisfied)
	0 (dissatisfied)

Label Encoding is performed to convert four character-type characteristics (gender, customer type, travel purpose, boarding seat) into real-type categorical data, as shown in Table 2 above.

F. Classification algorithm

1. Decision Tree

A decision tree is a type of supervised machine learning algorithm used for classification or regression tasks. It is a tree-like structure where each internal node represents a decision based on the values of predictor variables, and each leaf node represents a predicted outcome. The decision tree algorithm recursively splits the data into subsets based on the values of predictor variables, aiming to create subsets with homogeneous outcomes. The resulting decision tree can be used for making predictions on new data by following the decision path from the root to a leaf node based on the values of predictor variables.

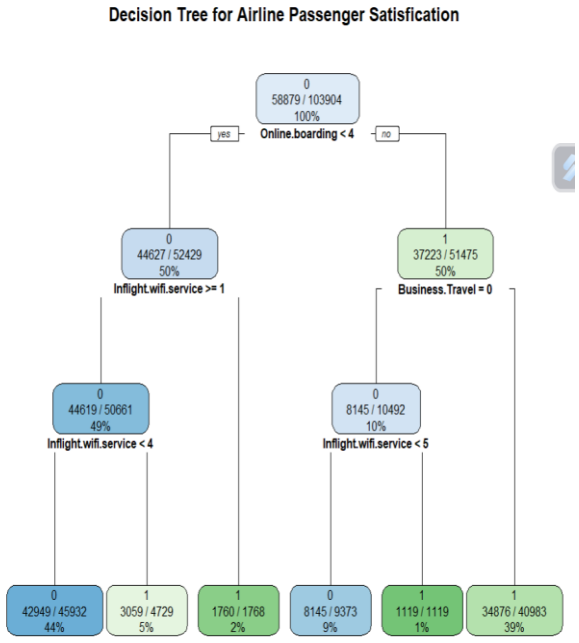


Fig.8 Decision tree model

2. Random Forest

Random Forest is a classifier that composes an ensemble through bagging (Bootstrap Aggregation) of multiple decision trees, and two parameters (number of trees, number of randomly selected variables) must be set. In this paper, the number of trees is set to 500, and the designated variable is used as the input data variable, as shown in Figure 9 below.

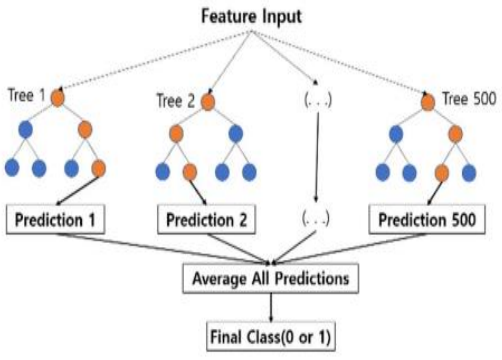


Fig.9 Random Forest model

IV. RESULTS AND MODEL EVALUATION

Feature Importance score of each feature in the airline passengers satisfaction data set is calculated so that the features that are having greater impact on the target class which is passenger satisfaction can be identified. To visualize the feature importance scores of each attribute clearly a bar graph is plotted by taking Feature Importance Score of each feature on X- axis and the corresponding Feature names on Y axis and the results produced are represented in Fig. 11.

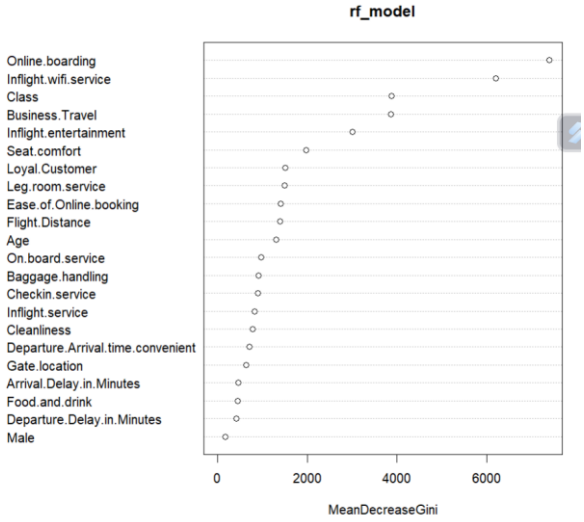


Fig.10 Feature Importance

Table 3. Accuracy of each prediction model

Models	Accuracy	Precision	Recall	F1 score
Decision Tree	88.3	92.3	86.4	89.3
Random Forest	96.3	95.6	97.9	96.8

The accuracy calculated through each customer satisfaction prediction experiment is compared and analysed, and is shown in Table 3 above.

(a)

```
                Reference
Prediction      0     1
         0  12599  1045
         1   1974 10358
```

(b)

```
                Reference
Prediction      0     1
         0  14278   648
         1    295 10755
```

Figure 11. Confusion Matrix of Decision Tree (a), Random Forest (b)

The confusion matrix is a visualization tool used to see at a glance the performance of supervised learning in the field of artificial intelligence. If true is said to be true, it is true positive (TP), if true is said to be false is negative, false (NP), if it is said to be true is false positive (FP), and finally false is said to be false, it is expressed as a true negative (TN). In this study, the highest accuracy is calculated from the random forest model among the two classification algorithms, and the airline customer satisfaction prediction model. Confusion matrix is shown in Figure 6 above.

As a result of confirming the Confusion Matrix, it is calculated accuracy in the two classification algorithms, which are the Random Forest, and Decision Tree prediction model. It is shown as high performance in order, which is consistent with the ranking of accuracy.

IV. RAPID MINER

RapidMiner is a data science platform that provides a range of tools and operators for predictive modelling, including data preparation, modelling, validation, and deployment. It allows users to build and deploy predictive models without writing any code, using a visual workflow interface.

1. Passenger dataset is loaded into RapidMiner and separate the features and the target variable into different columns.

2. Pre process the dataset to handle missing values, outliers, and other issues RapidMiner's data preparation operators to perform tasks such as imputation, normalization, and feature selection.

3. Split the dataset into training and testing sets using Rapid Miner's data splitting operators. Evaluate the performance on unseen data.

4. Select a machine learning algorithm to use for passenger satisfaction prediction. RapidMiner provides a range of classification algorithms, including decision trees, random forests, and support vector machines. Compare the performance of different algorithms using RapidMiner's validation operators.

5. Train the model on the training set using RapidMiner's modelling operators. This will involve specifying the algorithm and its parameters, as well as the features and target variable to use. Here decision tree classifier and random forest classifier are used.

6. Evaluate the performance on the testing set using RapidMiner's performance operators. Metrics such as accuracy, precision, recall, and F1 score, which you can use to assess the quality of model.

Overall, RapidMiner provides a range of tools and operators for performing passenger satisfaction prediction using machine learning. By following these steps and experimenting with different algorithms and parameters, a robust and accurate model can be created for predicting airline passenger satisfaction.

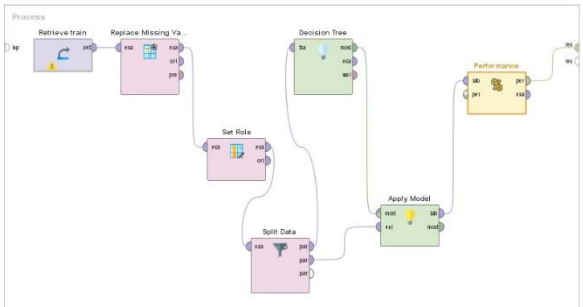


Figure 12. Rapid miner

accuracy: 91.31%

	true neutral or dissatisfied	true satisfied	class precision
pred. neutral or dissatisfied	16742	1786	90.36%
pred. satisfied	922	11722	92.71%
class recall	94.78%	86.78%	

Figure 13. Output

V. CONCLUSION

For airlines, a highly precise classification model is built to help recognise critical bottlenecks and improve passenger satisfaction. Airlines concentrate on optimizing the In-Flight Wi-Fi Service experience based on many simulations. Airlines may, for example, create improved tools to make accessing in-flight wi-fi easier, or lower the cost of accessing in-flight wi-fi so that more economy class travellers would take advantage of the service In this paper, the customer satisfaction prediction model is proposed using customer evaluation data for airlines, and a model is constructed from selected data. As a result of the experiment, the accuracy of 95.7% is calculated in the random forest model with the highest accuracy and performance. And if an airline establishes service management and marketing strategies by further analysing customer evaluation data and improving the airline customer satisfaction prediction model, it will be useful to help customers re-use services and to gain a relative advantage in the fiercely competitive market.

References

[1] Kyungdoo NAM, Thomas SCHAEFER, “Forecasting international airline passenger traffic using neural networks”, The Logistics and Transportation Review, 1995, 31.3: 239-252.

[2] Insil Park, “Influence of Customer Satisfaction and Reuse Intention on Service Quality of Airline Outsourcing: Focusing on National Airlines”, Tourism Management Research, Vol. 13, No. 39, pp.27-60, 2009.

[3] Czepiel, J. A., Rosenberg, L. J., Akerele, “Perspectives on consumer satisfaction”, AMA Conference Proceedings, pp.119-123, 1997.

[4] Cronin, J, Joseph, Jr. and Steven A, Taylor, "SERVPERF Versus SERVQUAL: Reconciling Performance -Based and Perceptions-Minus- Expectations Measurement of Service Quality", Journal of Marketing, 1994, p.127.DOI: https://doi.org/10.1177/002224299405800110

[5] Seongsuk Ahn, “The Influence of Service Integrity on Customer Satisfaction, Word of Mouth and Reuse Intention: Focused on Airline Service”, Korean Society for Aviation Management, Vol. 16, No. 1, pp.91-106, 2018.

[6] Pan-ho Choi, “A Study on the Influence of Selecting Attributes of Airline on Customer Satisfaction and Loyalty”, Korea Data Analysis Society, Vol. 21, No. 1, pp.305-317, 2019.

[7] Hyeon Mi Yoo, “A Study of the Nonlinear Relationship between Customer Satisfaction and Repurchase Intension”, Journal of Channel and Retailing, 2017, 22(3): 19-38.