INDIAN AIRLINE PASSENGERS SATISFACTION DATA ANALYSIS DURING COVID-19

Ravikanth Misra
Department of Computer Science and Engineering
Dhanekula Institute of Engineering and Technology
Vijayawada, India
ravikanthmishra@gmail.com

Phani Kishore Rompicharla
Department of Computer Science and Engineering
Dhanekula Institute of Engineering and Technology
Vijayawada, India
phani.rompicharla@gmail.com

Navya Bade
Department *of Computer Science and Engineering*
Dhanekula Institute of Engineering and Technology
Vijayawada, India
navyabade7@gmail.com

Ritikaa Tiwari
Department of Computer Science and Engineering
Dhanekula Institute of Engineering and Technology
Vijayawada, India
ritikaatiwari7@gmail.com

Chandini Mohammad
Department of Computer Science and Engineering
Dhanekula Institute of Engineering and Technology
Vijayawada, India
chandini.simmu@gmail.com

Jai Surya Balasa
Department of Computer Science and Engineering
Dhanekula Institute of Engineering and *Technology*
Vijayawada, India
jaisurya@diet.ac.in

***Abstract*—Customer satisfaction is the key to success of any business organization. Hence, it is necessary to analyze and improve the factors that have high impact on customer satisfaction as it affects growth and reputation of the business organization. This paper is aimed at Indian airline passengers satisfaction analysis during COVID-19. The analysis is carried out by random forest and C4.5 classification algorithms of machine learning. Later, the results of both the classification algorithms are compared to identify the algorithm with higher accuracy. The analysis predicts that both the algorithms have better accuracy. The comparison of both reports that random forest has better accuracy than C4.5 algorithm. Besides, we have also implemented feature importance score to identify the attributes that have high impact on passengers satisfaction and thus we can guide the airline company that if they can improve these attributes, they can achieve high customer satisfaction and improve their business. With regard to service attributes, we conclude that (1) distancing in seat assigning, (2) face mask provided by airlines, (3) hand sanitizer provided by airlines, and (4) passenger temperature screening are the top 4 crucial services that have high impact on passengers satisfaction.**

Keywords—passenger, satisfaction, airline, random forest, C4.5, feature importance

I. INTRODUCTION

Passenger satisfaction is the greatest concern of airline business [8]. The airline companies should understand how well the passengers are satisfied with their services and also identify the services that if they can improve them can achieve high customer satisfaction rate which leads to the growth of the business and also gain reputation in the industry. It's important for an airline company to remain an excellent experience to the passengers every time they travel through their airlines. There are several factors that need to be considered while analyzing the passengers satisfaction. In the researches carried earlier, they taken the attributes into account like inflight wi-fi service, seat comfort, baggage handling, inflight entertainment, boarding, departure and arrival time convenient etc. With the outburst of COVID-19, it is ideal to consider service attributes like hand sanitizer provided by airlines, distancing in seat assigning, passenger temperature screening, face mask provided by airlines, pandemic prevention policy compliance while analyzing the passengers satisfaction. In this paper we considered the above attributes along with the service attributes which were used in previous researches for analyzing customer satisfaction during COVID-19. This paper is aimed at analyzing the Indian airline passengers satisfaction data during COVID-19 using random forest and C4.5 classification algorithms in machine learning. A comparison is also made to identify the algorithm with higher accuracy [1][13]. Besides this comparison, the concept of feature importance score is implemented to identify the service attributes that have greater impact on passengers satisfaction which the airline company if can improve achieve success, gain profit and good reputation in the industry.

II. LITERATURE SURVEY

There are several studies on prediction of customer satisfaction using machine learning algorithms [6][14]. Comparative analysis between decision tree algorithms is also made for classification of airline customer satisfaction [1]. Machine learning algorithms like logistic regressions and neural

networks are used to predict customer behavior in Credit card churn forecasting and wireless telecommunication industry [9][11][12]. Airline businesses around the world have been destroyed by Covid-19 as most international air travel has been banned [2]. Some service attributes are more important than others for passengers' overall satisfaction [3][4]. One research provided a tool for measuring air passenger satisfaction and for identifying the critical service aspects available in the terminal in order to offer services characterized by a high level of quality [5]. In one of the papers, an investigation is performed on the linkages between customer service, customer satisfaction, and firm performance in the US airline industry [10]. During the literature review, we have identified research papers that motivated us to develop a machine learning model for analyzing Indian airline passengers satisfaction.

III. METHODOLOGY

The methodology of this paper used KDD process [7] which involves data cleaning, data integration, data reduction, data transformation and evaluation. A comparison is made between the most popular decision tree data mining algorithms namely random forest and C4.5. This paper implemented the concept of feature importance score to identify service attributes that have higher impact on passenger satisfaction.

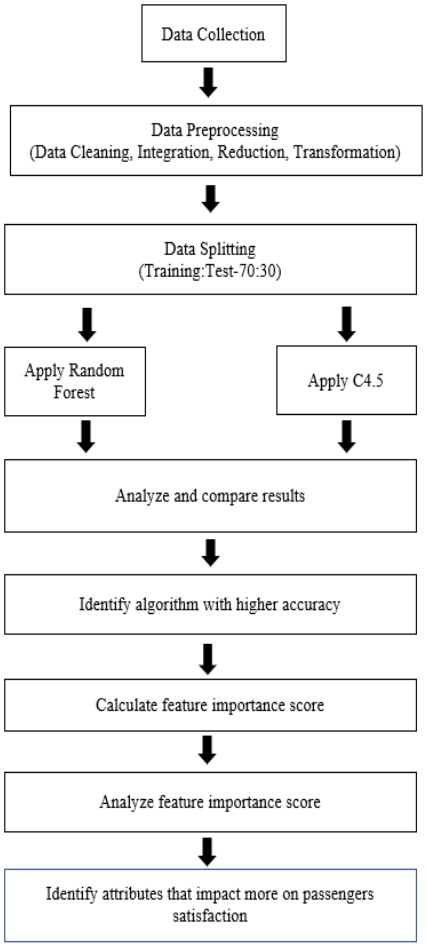

Fig. 1. Methodology

A. *Data Collection*

The dataset of Indian airline passengers satisfaction data analysis during COVID-19 was collected through google form. The airline passengers satisfaction attribute names are obtained from Kaggle and 5 more service attributes that are likely to show impact on passengers satisfaction during COVID-19 are identified for generating a google form. In total, the dataset had 24 attributes. The TABLE I. shows the various attributes that are taken into consideration.

B. *Data Preprocessing*

The data collected through the google form is in the raw format. The data may consist of several missing and inappropriate values. Data preprocessing is the process that transforms the raw data into useful format. In order to convert the data collected through google form suitable for analyzing using the random forest and C4.5 algorithms data preprocessing is performed. The preprocessing of data includes data cleaning, data integration, data reduction and data transformation.

C. *Data Splitting*

The partitioning of data into training and test data is referred as data splitting. The data collected through google form is partitioned into training and test data in the ratio of 70:30 where 70% of collected data is used for training the model and 30% of data is used for testing the model and finding the accuracy.

TABLE I. ATTRIBUTES OF INDIAN AIRLINE PASSENGERS SATISFACTION DATASET

Field	Type	Value
Age	Nominal	8..95
Flight Distance	Nominal	42..5674
Inflight wi-fi service	Real	1..5
Departure/Arrival time convenient	Real	1..5
Ease of Online booking	Real	1..5
Gate location	Real	1..5
Food and drink	Real	1..5
Online boarding	Real	1..5
Seat comfort	Real	1..5
Inflight entertainment	Real	1..5
On-board service	Real	1..5
Leg room service	Real	1..5
Baggage handling	Real	1..5
Check-in service	Real	1..5
Inflight service	Real	1..5
Cleanliness	Real	1..5
Departure Delay in Minutes	Real	1..5
Arrival Delay in Minutes	Real	1..5
Distancing in seat assigning	Real	1..5
Passenger temperature screening	Real	1..5
Face mask provided by airlines	Real	1..5
Hand sanitizer provided by airlines	Real	1..5
Pandemic prevention policy compliance	Real	1..5
satisfaction	Nominal	Satisfied; neutral or dissatisfied

D. Random Forest

Random forest algorithm builds number of decision trees based on different subsets of given data. The structure of decision tree is similar to a flow chart in which the leaf nodes represent the class label and the internal nodes represent the test on a feature. The prediction in random forest algorithm is based on average of all the decision trees to improve accuracy. The greater the number of trees in the random forest, the higher the accuracy and avoids overfitting. The advantage of random forest algorithm over other classification algorithms is that it takes less training time when compared to other algorithms. Random forest algorithm is best suitable for classification even the size of the data set is large.

E. C4.5 Algorithm

C4.5 algorithm is one of the decision tree classifiers. It is an extension of ID3 algorithm. The training data that is provided to the C4.5 algorithm is used for building the decision tree. The improvements that have been made by C4.5 algorithm over ID3 algorithm are, capable of handling both discrete and continuous attributes and handling attributes with varying costs.

F. Feature Importance Score

The importance of each feature for a model can be represented by calculating the feature importance score for each feature in the dataset that builds the model. The effect of a specific feature on the model can be identified based on the value of its feature importance score. The higher the score of a feature, greater the effect of the feature on the model. It represents how important the feature is in predicting the target class. In this research the feature importance score of all the attributes is represented in the form of a bar graph. Thus, we can identify the service attributes that shows higher impact on customer satisfaction.

IV. Results and Discussion

The effectiveness of a machine learning model can be evaluated based on the evaluation metrics like accuracy, confusion matrix, recall, precision, f1 score etc. The performance of a classification algorithm can be determined by one of the classification metrics known as confusion matrix. In confusion matrix there are N number of rows and N number of columns where N represents the count of target classes in a given data set. The dataset used in this research consists of 2 target classes namely satisfied, neutral or dissatisfied. Hence the confusion matrix is of size 2x2. Recall is the measure of the model's ability to determine the relevant cases. The precision of a model represents the ratio of correctly identified positives to the total number of positives identified by the model. The f1 score is obtained by calculating the weighted average of recall and precision. The most frequently used evaluation metric is accuracy. Accuracy is the ratio of number of correct predictions the model had made to the total number of predictions.

A. Random Forest Results

Random forest algorithm applied on Indian airline passengers satisfaction data during COVID-19 containing 24 attributes by splitting the dataset in the ratio of 70:30 as training and testing data produced the results in TABLE II. The confusion matrix of random forest algorithm is represented in Fig. 2.

TABLE II. Random Forest Results

	precision	recall	f1-score	support
neutral or dissatisfied	1.00	0.93	0.97	60
satisfied	0.94	1.00	0.97	58
Accuracy			0.97	118

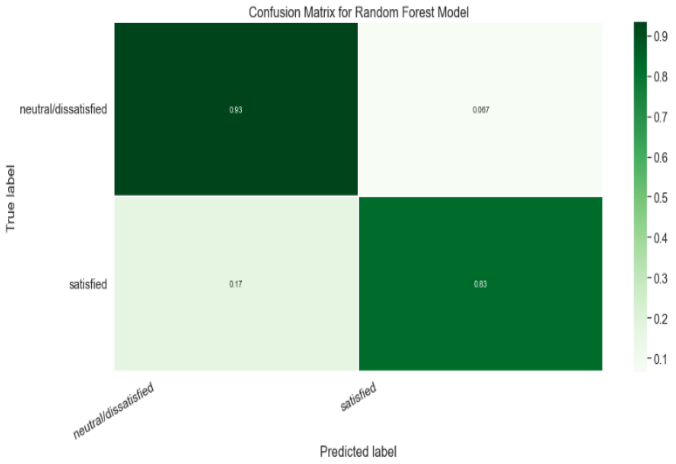

Fig. 2. Confusion matrix for random forest

B. C4.5 Algorithm Results

C4.5 algorithm applied on Indian airline passengers satisfaction data during COVID-19 containing 24 attributes by splitting the dataset in the ratio of 70:30 as training and testing data produced the results in TABLE III. The confusion matrix of C4.5 algorithm is represented in Fig. 3.

TABLE III. C4.5 Algorithm Results

	precision	recall	f1-score	support
neutral or dissatisfied	0.85	0.93	0.89	60
satisfied	0.92	0.83	0.87	58
Accuracy			0.88	118

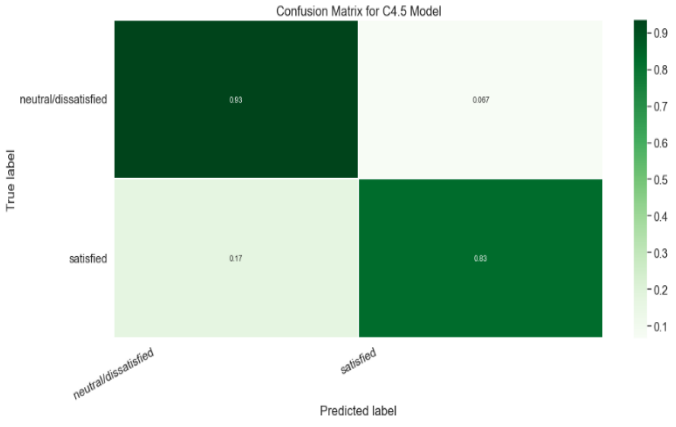

Fig. 3. Confusion matrix for C4.5 Model

C. Comparing Accuracy

After calculating the accuracy, precision, recall, f1-score, support and confusion matrix of random forest and C4.5 algorithms a comparison of accuracy between random forest and C4.5 algorithm is made to identify the algorithm with higher accuracy. The TABLE IV below shows the results of comparison of accuracy between random forest and C4.5 algorithms. After analyzing the results, it is observed that random forest algorithm records an accuracy of 97 and C4.5 records an accuracy of 88.

TABLE IV. COMPARING ACCURACY OF RANDOM FOREST AND C4.5

Algorithm	Accuracy
Random Forest	97
C4.5	88

D. Feature Importance Score Results

Feature Importance score of each feature in the Indian 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. 4.

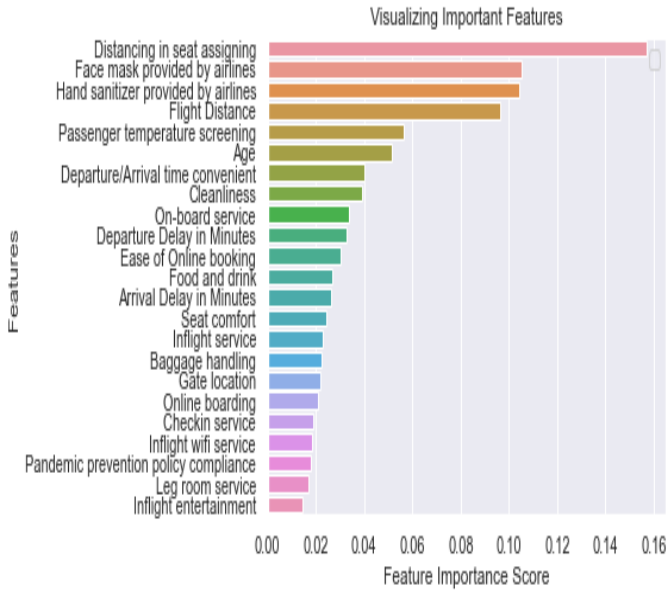

Fig. 4. Feature Importance Score

V. CONCLUSION

After analyzing the results of random forest and C4.5 algorithms on Indian airline passengers satisfaction data with training and testing split ratio as 70:30, it is observed that random forest algorithm records an accuracy of 97 and C4.5 algorithm records an accuracy of 88. It is observed that the same data set results in different accuracy when classified with different classification algorithms. The analysis of feature importance score graph demonstrates that distancing in seat assigning, face mask provided by airlines, hand sanitizer provided by airlines, and passenger temperature screening are the top 4 crucial services attributes that are having high impact on passengers satisfaction. Thus, if the airline companies can improve those services can attract more passengers and improve their business as well as grow the reputation of the company. This paper used the decision tree algorithms namely random forest and C4.5 for analyzing the airline passengers satisfaction. The analysis of passengers satisfaction can also be done with other classification algorithms like KNN, Naïve Bayes etc., with varying split ratios and a comparison can be made among them. In this paper we considered all the attributes in the dataset for the construction of the model. Instead of considering all the features in model construction feature selection techniques can be applied to identify the better features for constructing a model so that a much better accuracy can be drawn. In order to identify the service attributes that have high impact on passenger satisfaction this paper implemented feature importance score, instead correlation between every pair of features can be calculated and represented graphically so that the features which are having high correlation with target class can be identified as the high impact services on passengers satisfaction.

REFERENCES

[1] Comparative analysis of decision tree algorithms: Random forest and C4.5 for airlines customer satisfaction classification: W Baswardono et al 2019 J. Phys.: Conf. Ser. 1402 066055. Journal of Physics: Conference Series.

[2] Predicting Airline Passenger Satisfaction with Classification Algorithms B. Herawan Hayadi 1,* , Jin-Mook Kim 2 , Khodijah Hulliyah 3 , Husni Teja Sukmana 4. International Journal of Informatics and Information System Vol. 4, No. 1, March 2021, pp. 82-94.

[3] Determinants of Customer Satisfaction at the San Francisco International Airport: Ashok K. Singh1 *, Myongjee Yoo2 and Rohan J. Dalpatadu3. Journal of Tourism & Hospitality.

[4] Wang X, Hong M, Berger PD. Customer-satisfaction analysis at San Francisco international airport. International Journal of Management Studies. 2015;2(1):1-12.

[5] Eboli L, Mazzulla G. An ordinal logistic regression model for analysing airport passenger satisfaction. EuroMed Journal of Business. 2009;4(1):40-57.

[6] Research on Bank Customer Satisfaction Classification on Random Forest Algorithm Li Yang*. International Journal of Current Advanced Research, ISSN: O: 2319-6475, ISSN: P: 2319-6505, Volume 7; Issue 6(E).

[7] Linear and Non-Linear Clustering Algorithms for Data Mining Applications Neelakantappa M*. International Journal of Current Advanced Research, ISSN: O: 2319-6475, ISSN: P: 2319-6505,Volume 7; Issue 2(G).

[8] A machine learning approach to analyze customer satisfaction from airline tweets Sachin Kumar* and Mikhail Zymbler. Journal of Big Data (2019) 6:62 https://doi.org/10.1186/s40537-019-0224-1.

[9] Nie, Guangli, et al. "Credit card churn forecasting by logistic regression and decision tree." Expert Systems with Applications 38.12 (2011): 15273-15285.

[10] Steven, Adams B., Yan Dong, and Martin Dresner. "Linkages between customer service, customer satisfaction and performance in the airline industry: Investigation of non-linearities and moderating effects." Transportation Research Part E: Logistics and Transportation Review 48.4 (2012): 743-754.

[11] Mozer, Michael C., et al. "Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry." Neural Networks, IEEE Transactions on 11.3 (2000): 690-696.

[12] Hadden, John, et al. "Computer assisted customer churn management: State-of-the-art and future trends." Computers & Operations Research 34.10 (2007): 2902-2917.

[13] A Comparative Study of Classification Algorithms for Spam Email Data Analysis Aman Kumar Sharma, Suruchi Sahni. International Journal on Computer Science and Engineering, ISSN : 0975-3397, Vol. 3 No. 5.

[14] Prediction of Customer Satisfaction Using Naive Bayes, MultiClass Classifier, K-Star and IBK Sanjiban Sekhar Roy, Deeksha Kaul, Reetika Roy, Cornel Barna, Suhasini Mehta, Anusha Misra. International Workshop Soft Computing Applications, DOI:10.1007/978-3-319-62524-9_12.