Data Mining and Machine Learning

Abstract— **The Airline Industry is one of the major sectors that generate a large amount of the nation's revenue. It consists of many unpredictable circumstances and problems that can cause a very huge loss of economy for a nation. Problems like flight delays due to various factors not only affect the airlines, airport but also affect the passengers. To retain their valuable customers, airlines usually get the customer's feedback and based on that they try to improve their services. So, classifying the customers on the level of their satisfaction as well as predicting airline delays is equally important for the airline industry. In this project, the customer's satisfaction is classified using KNN classification and the Decision Tree algorithms. Also, the Logistic Regression Model is applied to airline past data to predict airline delays. Using the data on various airlines with their prices during a period we can predict the future prices of flight tickets with the past data. This project uses Random Forest Regression and Support vector machine algorithms for this prediction. Flight fare prediction can be a hard guess, the particular flight price today might be different tomorrow. The main motive of customers is to buy the airline ticket at the minimum price whereas the airline department is trying to obtain maximum profit. Thus, this prediction will help both the customers as well as the airline department.**

I. INTRODUCTION

Machine learning is mainly a field of Artificial Intelligence, which is a core component in digitalization solutions that has gained a great deal of interest in the modern arena. The Airline Industry is one of the major sectors that generate a large amount of the nation's revenue but also consists of many problems that may result in a huge loss of economy for a nation. One of the critical modern life challenges of airline industries is a flight delay. In 2007, this affected approximately $33 billion as a direct or indirect cost to customers, airlines, and other parts of the industry[15]. Flight delays are unavoidable and have many negative economic impacts on passengers as well as the industry. To retain their valuable customers, airlines usually get the customer's feedback and based on that they try to improve their services. So, classifying the customers on the level of their satisfaction is equally important for the airline industry. There are currently 39 airlines operating in India including scheduled, regional, chartered, and cargo airlines that make this market a highly competitive one with rapid growth. Using the data for these airlines one can predict the airfares using the machine learning models.

In this project, an attempt has been made in using numerous machine learning algorithms that are most widely used and applied to different datasets related to the airline industry. This project compares the different machine learning models in the R language and looks at the advantages of supervised machine learning algorithms in terms of accuracy, measuring error rate, complexity, and risk of overfitting measures. The main objective of this project is to give an overall comparison with the state of art machine learning techniques.

The first dataset consists of airline on-time statistics along with causes for delay in flights for the United States. Ordinal logistic regression is applied to this dataset to predict flight delays.

The data in this second dataset is provided by an airline organization with the given name Invistico airlines. The machine learning techniques for classification like K-Nearest Neighbours (KNN) and Decision tree are applied to this dataset to predict the satisfaction of airline passengers.

The third dataset is about the different airlines in India and their details along with the prices. The price prediction is made using the machine learning algorithms for regression like Support Vector Machines(SVM) and Random Forest.

The following sequence of sections is used in this project: Section 1 offers an overview of the different predictions performed for the airline industry and the methodologies used. Section 2 explains the appropriate analysis that has been carried out using research papers. Section 3 is a thorough description of the data mining methodologies. Section 4 shows the results and talks about the various evaluation methods seen. The last section Section 5 finally provides a conclusion for the project.

II. RELATED WORK

A general distinction is made in paper[1] between different machine learning algorithms such as Artificial neural networks, Decision trees, KNN algorithm, Logistic regression, Random Forest, and help vector machines. The authors of this paper grouped the areas according to the machine learning models where they could be implemented and produce good results.

In the research paper [2], several machine learning algorithm was applied to the airline data which consists of cabin class passenger data, cabin class supplied data, the distance of flights, etc. Amongst all the applied algorithm SMO classification model outperformed with an accuracy of approximately 86%. The approach used in this paper was tried implemented on this project, but the results were not

as per expectations as they were classifying properly with the predictors applied in this research paper.

To predict flight delays statistically is given in this paper. One of the popular machine learning algorithms Gradient Boosted Decision Tree is used to analyze air traffic delay prediction tasks. In this paper, the data taken from the U.S Department of Transportation on Passenger Flight on-time Performance is used for modeling and has predicted the delays. Gradient Boosted Decision Tree model has used six important attributes to predict flight delays. This model has provided the highest coefficient of Determination in arrival as well as departure delays. The same prediction can be done with a more straight forward approach by applying logistic regression as shown in the current project[3].

In this paper [4], the author proposes a framework to evaluate customer satisfaction based on their comments provided on social media platforms. With the help of the analysis in this paper, potential customers can pick up the most suitable flight according to their demands. Using the constructed dictionaries author has done the sentiment score calculation of Twitter comments with different perspectives. The comments that are calculated by scores are classified into three different categories: positive sentiment, neutral sentiment, and negative sentiment by using the method of sentiment classification. The analysis of the result predicted in this paper can provide great and valuable insight for different companies resulting in improvement in their services and quality, so that they can be more active and competitive in the market[4]. The implementation tried in this paper is exceptionally good and has taken a more correct approach in evaluating customer satisfaction. It can be future work for this ongoing project to evaluate customer satisfaction using sentiment analysis.

In this paper [5], the author has explained the need for a buyer to know the price trends for flight tickets. A model that predicts the price trends without using the official information from the airlines is being proposed in this paper. According to the findings of the proposed model, the trends can be predicted also the model can tell the airfare changes up to the departure dates with the help of the public airfare data that is available online. Evaluating different conventional machine learning algorithms like Decision Tree, MLP, AdaBoost, Gradient boosting, Random Forest, KNN, and SVMs the model predicts the airfares. The final interpretation regarding predicting the airfares with different machine algorithms is built by stacking two separate models Random Forest and MLP[5]. Stacking algorithms is a more complex approach in predicting the target variables. A more simple approach towards predicting is applying the conventional machine learning algorithms and comparing the results with the non-conventional ones.

In this research paper [6], the author is examining the influence of airline service quality on passenger satisfaction and loyalty. According to this paper, pre-flight, in-flight, and post-flight experience matters the most to a passenger. The analysis of this paper is outstanding but it could be more reliable to use machine learning models in predicting the customer's satisfaction by passing different parameters related to airlines.

In this research paper [7], prediction of passenger traffic for a certain airport is carried out. It is predicted as a two-step process. The first is, prediction of aircraft departures from the particular airport, and the second is, predicting the number of departures for that current route. Multiple regression is applied along with a Time series analysis model i.e. Holts winter additive method. Applying other conventional machine learning algorithms like Random Forest for classification may perform better than Time series analysis.

The efficiency of several machine learning algorithms for various datasets was contrasted in Paper[8]. Authors also found that the use of predictive methods as a benchmark is not perfect. The conclusion of this paper blends into the ongoing research paper, as in the application of traditional methodologies such as linear regression to the imputation of missing values on such dynamic data.

The models for data mining are explained sequentially in the paper[13]. The authors compared KDD, CRISP-DM, and SEMMA and offered a brief understanding of them. KDD is a reliable model for the data mining process and is adopted by the majority of researchers in this field. KDD is the most effective approach to be applied to the current project.

In this paper [10], prediction of passenger count is done for a particular flight for 355 days which are open in the reservations system. Box-Jenkins and artificial neural networks are used for the creation of the model. Box-Jenkins method outperformed the ANN. This research could be a future scope of this project in predicting the count of the passenger through the sales window.

In this paper,[11] the author explains different data mining techniques to lessen the risk of investments in the stock market. The data mining learning models like the Random Forest model and support vector machines are used to obtain reliable prediction of the stock prices based on the historical data. However, SVMS can better result in

financial forecasting when the data is time-series data. The author explains in detail the implementation of all three models along with the equations. The analysis is done on the 30 companies. The analysis using the optimum parameters that are influential in prediction is missing I this paper and could be taken as a further part to improve better.

In this research paper [12], It explores the usefulness of deep-learning models in air traffic delay prediction tasks. It shows that by applying deep RNN (Recurring neural networks), a good prediction model for delay states of a flight can be created which is much reliable. It can be used to create a model that can show important patterns in flight delay data.

III. METHODOLOGY

The Knowledge Discovery Databases (KDD) model is an iterative and interactive model, which involves many steps where many decisions are made by the user. It consists of five stages of data mining with nine steps in all. It is the process of pulling out the hidden knowledge from the databases[3]. Below are the 5 different stages of the KDD process explained in detail for each dataset.

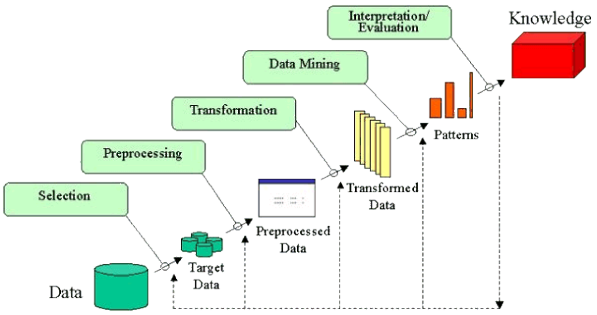


Fig. 1: Knowledge Discovery Databases Process Model

1. *Data Selection*

The first step of KDD involves appropriate data selection for further performing the data mining techniques on it. In this stage understanding and developing the domain knowledge for the project is important. This will then help to do data selection properly as the basic knowledge of the domain will help to easily understand the dataset columns and finalize a target dataset. The target dataset is selected from the available data sources and then is used for further steps.

i. Airline Delay Dataset:

The tracking of domestic flights for large air carriers for their on-time arrivals is done by the U.S. Department of Transportation (DOT) Bureau of Transportation Statistics (BTS). The BTS maintains this large data and publishes it after every 30 days. The need to lessen the flight delays and improve the percentage of on-time arrivals of flights is the need of the hour to save the huge economic loss of the nation. This dataset consists of 30 variables describing the flight details like the actual and scheduled arrival time, departure time, if any delays occurred, which type of delays, etc. and have in all 1936758 observations shown in Fig 2.

```
> str(flightdelay)
'data.frame':   1936758 obs. of  30 variables:
 $ X                 : int  0 1 2 4 5 6 10 11 15 16 ...
 $ Year              : int  2008 2008 2008 2008 2008 2008 2008 2008 2008 2008 ...
 $ Month             : int  1 1 1 1 1 1 1 1 1 1 ...
 $ DayofMonth        : int  3 3 3 3 3 3 3 3 3 3 ...
 $ DayOfWeek         : int  4 4 4 4 4 4 4 4 4 4 ...
 $ DepTime           : num  2003 754 628 1829 1940 ...
 $ CRSDepTime        : int  1955 735 620 1755 1915 1830 700 1510 1020 1425 ...
 $ ArrTime           : num  2211 1002 804 1959 2121 ...
 $ CRSArrTime        : int  2225 1000 750 1925 2110 1940 915 1725 1010 1625 ...
 $ UniqueCarrier     : chr  "WN" "WN" "WN" "WN" "WN" ...
 $ FlightNum         : int  335 3231 448 3920 378 509 100 1333 2272 675 ...
 $ TailNum           : chr  "N712SW" "N772SW" "N428WN" "N464WN" ...
 $ ActualElapsedTime: num  128 128 96 90 101 240 130 121 52 228 ...
 $ CRSElapsedTime    : num  150 145 90 90 115 250 135 135 50 240 ...
 $ AirTime           : num  116 113 76 77 87 230 106 107 37 213 ...
 $ ArrDelay          : num  -14 2 14 34 11 57 1 80 11 15 ...
 $ DepDelay          : num  8 19 8 34 25 67 6 94 9 27 ...
 $ Origin            : chr  "IAD" "IAD" "IND" "IND" ...
 $ Dest              : chr  "TPA" "TPA" "BWI" "BWI" ...
 $ Distance          : int  810 810 515 515 688 1591 828 828 162 1489 ...
 $ TaxiIn            : num  4 5 3 3 4 3 5 6 6 7 ...
 $ TaxiOut           : num  8 10 17 10 10 7 19 8 9 8 ...
 $ Cancelled         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CancellationCode  : chr  "N" "N" "N" "N" ...
 $ Diverted          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CarrierDelay      : num  NA NA NA 2 NA 10 NA 8 NA 3 ...
 $ WeatherDelay      : num  NA NA NA 0 NA 0 NA 0 NA 0 ...
 $ NASDelay          : num  NA NA NA 0 NA 0 NA 0 NA 0 ...
 $ SecurityDelay     : num  NA NA NA 0 NA 0 NA 0 NA 0 ...
 $ LateAircraftDelay: num  NA NA NA 32 NA 47 NA 72 NA 12 ...
```

Fig 2: Structure of the Flight Delay Dataset.

ii. Airline Passenger Dataset:

The feedback of airline passengers is very important for airlines to improve their services and retain their customers. Also, the airlines must know the parameters on which they need to focus on satisfying their passengers. This dataset helps to survey the airline organization which is named Invistico airlines and predict whether the future passenger would be satisfied or not with the service provided by the airlines. This dataset contains 23 variables in all along with the target variable satisfaction and 129880 observations as shown in the below figure.

```
> str(airlinedata)
'data.frame':   129880 obs. of  23 variables:
 $ satisfaction                    : Factor w/ 2 levels "dissatisfied",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Gender                          : Factor w/ 2 levels "Female","Male": 1 2 1 1 1 2 1 2 1 2 ...
 $ Customer.Type                   : Factor w/ 2 levels "disloyal customer",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Age                             : int  65 47 15 60 70 30 66 10 56 22 ...
 $ Type.of.Travel                  : Factor w/ 2 levels "Business travel",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Class                           : Factor w/ 3 levels "Business","Eco",..: 2 1 2 2 2 2 2 2 1 2 ...
 $ Flight.Distance                 : int  265 2464 2138 623 354 1894 227 1812 73 1556 ...
 $ Seat.comfort                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Departure.Arrival.time.convenient: int  0 0 0 0 0 0 0 0 0 0 ...
 $ Food.and.drink                  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Gate.location                   : int  2 3 3 3 3 3 3 3 3 3 ...
 $ Inflight.wifi.service           : int  2 0 2 3 4 2 2 2 5 2 ...
 $ Inflight.entertainment          : int  4 2 0 4 3 0 5 0 3 0 ...
 $ Online.support                  : int  2 2 2 3 4 2 5 2 5 2 ...
 $ Ease.of.online.booking          : int  3 3 2 1 2 2 5 2 4 2 ...
 $ On.board.service                : int  3 4 3 1 2 5 5 3 4 2 ...
 $ Leg.room.service                : int  0 4 3 0 0 4 0 3 0 4 ...
 $ Baggage.handling                : int  3 4 4 1 2 5 5 4 1 5 ...
 $ Checkin.service                 : int  5 2 4 4 4 5 5 5 5 3 ...
 $ Cleanliness                     : int  3 3 4 1 2 4 5 4 4 4 ...
 $ Online.boarding                 : int  2 2 2 3 5 2 3 2 4 2 ...
 $ Departure.Delay.in.Minutes      : int  0 310 0 0 0 0 17 0 0 30 ...
 $ Arrival.Delay.in.Minutes        : int  0 305 0 0 0 0 15 0 0 26 ...
>
```

Fig 3: Structure of the Airline Passenger Dataset.

iii. Flight Fare Dataset:

Knowing the airline ticket price well in advance before booking a flight is what every customer wants. As the customer is looking for a minimum flight fare and the airline industry that is selling the ticket is trying to get the maximum amount of the ticket, balanced airfare prediction is the need of the hour. This dataset contains data for various airlines in India which helps to predict the future

price of the flight based on the airline's different attributes in the past. This dataset contains 10683 observations and 11 attributes for the different airlines.

```
> str(airfare)
tibble [10,683 x 11] (S3: tbl_df/tbl/data.frame)
 $ Airline        : chr [1:10683] "IndiGo" "Air India" "Jet Airways" "IndiGo" ...
 $ Date_of_Journey: chr [1:10683] "24/03/2019" "1/05/2019" "9/06/2019" "12/05/2019" ...
 $ Source         : chr [1:10683] "Banglore" "Kolkata" "Delhi" "Kolkata" ...
 $ Destination    : chr [1:10683] "New Delhi" "Banglore" "Cochin" "Banglore" ...
 $ Route          : chr [1:10683] "BLR <U+2192> DEL" "CCU <U+2192> IXR <U+2192> BBI <U+21
BOM <U+2192> COK" "CCU <U+2192> NAG <U+2192> BLR" ...
 $ Dep_Time       : chr [1:10683] "22:20" "05:50" "09:25" "18:05" ...
 $ Arrival_Time   : chr [1:10683] "01:10 22 Mar" "13:15" "04:25 10 Jun" "23:30" ...
 $ Duration       : chr [1:10683] "2h 50m" "7h 25m" "19h" "5h 25m" ...
 $ Total_Stops    : chr [1:10683] "non-stop" "2 stops" "2 stops" "1 stop" ...
 $ Additional_Info: chr [1:10683] "No info" "No info" "No info" "No info" ...
 $ Price          : num [1:10683] 3897 7662 13882 6218 13302 ...
>
```

Fig 4: Structure of the Flight Fare Dataset.

The above three selected datasets are selected from the data source named Kaggle considered as our target datasets for performing machine learning algorithms and are sent further to the next stage for preprocessing.

2. *Data Cleaning and Preprocessing*

This is the next stage of KDD in which various strategies are applied to target data to clean and pre-process the data before applying any model. This step removes the noisy and inconsistent data from the target dataset. Below are the preprocessing and cleaning steps applied to the three datasets.

i. Airline Delay Dataset:

a. Removing missing data

In this dataset 11 columns comprised of missing data as seen in the below figure which had to be either removed or replaced from the dataset.

```
> #Data Cleaning
> #Checking for missing values
> sapply(flightdelay,function(x) sum(is.na(x)))
             X           Year          Month     DayofMonth      DayOfWeek        DepTime
             0              0              0              0              0              0
   CRSDepTime        ArrTime     CRSArrTime  UniqueCarrier      FlightNum        TailNum
             0           7110              0              0              0              5
ActualElapsedTime CRSElapsedTime        AirTime       ArrDelay       DepDelay         Origin
          8387            198           8387           8387              0              0
          Dest       Distance         TaxiIn        TaxiOut      Cancelled CancellationCode
             0              0           7110            455              0              0
     Diverted   CarrierDelay   WeatherDelay       NASDelay  SecurityDelay LateAircraftDelay
             0         689270         689270         689270         689270         689270
>
```

Fig 5: Missing Data checking in Airline Delay Dataset

So, starting with the LateAircaftDelay column, 689270 missing values need to be removed. After imputing the missing values with the mean values of the column as well as through the linear model the results were not improved significantly and as this dataset was way too large with around 19 lakh observations, removing the observations that have missing values wasn't a problem here. After removing the data we can see that only one column named TailNum was left with missing values of about 2 observations which were then removed from the dataset using the same command. This step was repeated until we have all the columns with 0 null values as seen in the below figure.

```
> #Removing missing values in column LateAircraftDelay
> flightdelay<-flightdelay[!is.na(flightdelay$LateAircraftDelay), ]
> #Again checking for missing values
> sapply(flightdelay,function(x) sum(is.na(x)))
             X           Year          Month     DayofMonth      DayOfWeek        DepTime
             0              0              0              0              0              0
   CRSDepTime        ArrTime     CRSArrTime  UniqueCarrier      FlightNum        TailNum
             0              0              0              0              0              2
ActualElapsedTime CRSElapsedTime        AirTime       ArrDelay       DepDelay         Origin
             0              0              0              0              0              0
          Dest       Distance         TaxiIn        TaxiOut      Cancelled CancellationCode
             0              0              0              0              0              0
     Diverted   CarrierDelay   WeatherDelay       NASDelay  SecurityDelay LateAircraftDelay
             0              0              0              0              0              0
> #Removing missing values in TailNum column
> flightdelay<-flightdelay[!is.na(flightdelay$TailNum), ]
> #Again checking for missing values until all the columns have 0 missing data
> sapply(flightdelay,function(x) sum(is.na(x)))
             X           Year          Month     DayofMonth      DayOfWeek        DepTime
             0              0              0              0              0              0
   CRSDepTime        ArrTime     CRSArrTime  UniqueCarrier      FlightNum        TailNum
             0              0              0              0              0              0
ActualElapsedTime CRSElapsedTime        AirTime       ArrDelay       DepDelay         Origin
             0              0              0              0              0              0
          Dest       Distance         TaxiIn        TaxiOut      Cancelled CancellationCode
             0              0              0              0              0              0
     Diverted   CarrierDelay   WeatherDelay       NASDelay  SecurityDelay LateAircraftDelay
             0              0              0              0              0              0
```

Fig 6: Removing missing values from Airline Delay Dataset

b. Removing Outliers

In this dataset, there were multiple outliers present in the target variable ArrDelay. These needed to be removed to get better fit the machine learning model. The below figures explains the difference. Fig 7. shows the boxplot for the ArrDelay variable which had multiple outliers and Fig 8. Shows the boxplot after removing the outliers from the variable.

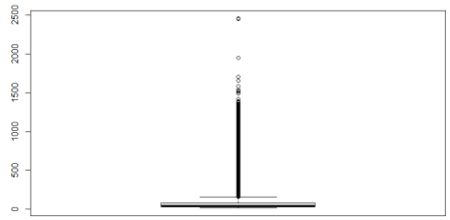


Fig 7: Boxplot for ArrDelay Column before removing outliers

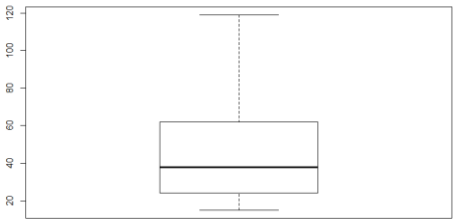


Fig 8: Boxplot for ArrDelay column after removing outliers

The other variables were also checked for outliers by plotting against boxplot and the ones having significant outliers were cleaned by removing the outliers from the variables. Variables like CarrierDelay, NASDelay, WeatherDElay, Security Delay, TaxiIn, TaxiOut, AirTime had to be cleaned by removing outliers. This way it might help in improving the probability of the model to fit better.

ii. Airline Passenger Dataset:

a. Removing Missing data

In this dataset, there are missing values only in one column named Arrival.Delay.in.Minutes. The column

consists of 393 missing values in airline delay(shown in Fig.9) which can be imputed using the Departure.Delay.in.Minutes column by applying the linear model between them. This will give the missing data values in Arrival.Delay.in.Minutes column with respect to the Departure.Delay.in.Minutes.

```
> #checking for missing values
> sapply(airlinedata,function(x) sum(is.na(x)))
                satisfaction                      Gender               Customer.Type
                           0                           0                           0
                         Age              Type.of.Travel                       Class
                           0                           0                           0
           Flight.Distance              Seat.comfort Departure.Arrival.time.convenient
                           0                           0                           0
            Food.and.drink             Gate.location          Inflight.wifi.service
                           0                           0                           0
     Inflight.entertainment            Online.support        Ease.of.Online.booking
                           0                           0                           0
          On.board.service          Leg.room.service             Baggage.handling
                           0                           0                           0
           Checkin.service               Cleanliness             Online.boarding
                           0                           0                           0
Departure.Delay.in.Minutes   Arrival.Delay.in.Minutes
                           0                         393
```

Fig 9: Checking for missing values in Airline Passenger dataset

After applying the linear model for predicting Arrival.Delay.in.Minutes by the predictor Departure.Delay.in.Minutes there are zero missing values left (as shown in fig.10) in the dataset and the arrival delay column consists of the appropriate delay values that will help in our prediction model further.

```
> sapply(ImputedArrivalDelayLM,function(x) sum(is.na(x)))
                        ID                satisfaction                      Gender
                         0                           0                           0
               Customer.Type                         Age              Type.of.Travel
                         0                           0                           0
                       Class           Flight.Distance                Seat.comfort
                         0                           0                           0
Departure.Arrival.time.convenient          Food.and.drink               Gate.location
                         0                           0                           0
          Inflight.wifi.service   Inflight.entertainment              Online.support
                         0                           0                           0
        Ease.of.Online.booking        On.board.service            Leg.room.service
                         0                           0                           0
             Baggage.handling         Checkin.service                 Cleanliness
                         0                           0                           0
             Online.boarding Departure.Delay.in.Minutes   Arrival.Delay.in.Minutes
                         0                           0                           0
```

Fig 10: Checking for missing values in Airline Passenger dataset after imputation of the linear model

b. Removing Outliers

In this dataset, there was a column named Flight Distance where outliers were present. These needed to be removed to get better fit the machine learning model. The below figures explains the difference. Fig 11. shows the boxplot for the FlightDistance variable which had multiple outliers and Fig 12. Shows the boxplot after removing the outliers from the variable.

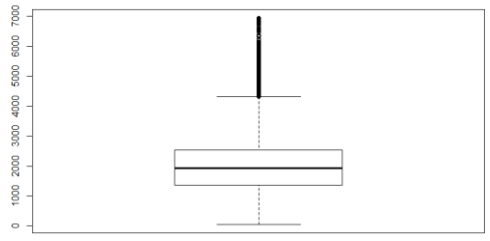


Fig 11: Boxplot for Flight Distance Column before removing outliers

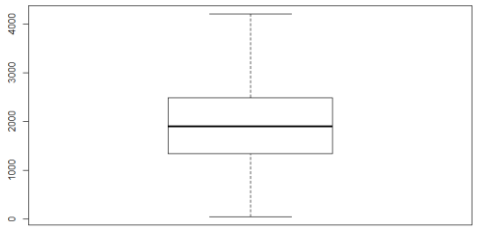


Fig 12: Boxplot for Flight Distance Column before removing outliers

iii. Flight Fare Dataset:

a. Removing Missing values

This dataset consists of only 2 variables with missing values that too in only one observation each so this can be removed from the dataset without any worries. The below fig 13. shows the before and after version of the dataset for missing values.

```
> sapply(airfare,function(x) sum(is.na(x)))
      Airline Date_of_Journey          Source     Destination           Route        Dep_Time
            0               0               0               0               1               0
 Arrival_Time        Duration     Total_Stops Additional_Info           Price
            0               0               1               0               0
> #removing the observation with missing value in Route variable
> airfare<-airfare[!is.na(airfare$Route), ]
> #checking again for missing values in airfare dataset
> sapply(airfare,function(x) sum(is.na(x)))
      Airline Date_of_Journey          Source     Destination           Route        Dep_Time
            0               0               0               0               0               0
 Arrival_Time        Duration     Total_Stops Additional_Info           Price
            0               0               0               0               0
```

Fig 13: Removing Missing values in Flight Fare dataset

b. Removing Outliers

In this dataset, a column named Price which is the target variable consists of some outliers. These needed to be removed to get a better fit of the machine learning model. The below figures explains the difference. Fig 14. shows the boxplot for the Price variable which had some outliers and Fig 15. Shows the boxplot after removing the outliers from the variable.

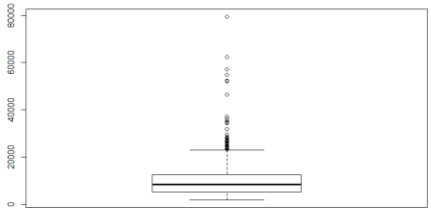


Fig 14: Boxplot for Price Column before removing outliers

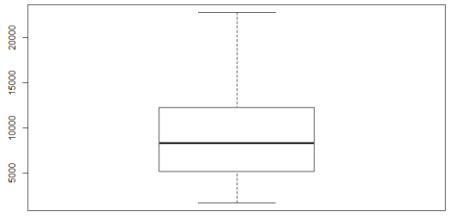


Fig 15: Boxplot for Price Column after removing outliers

So, here this stage was completed with all the cleaning and preprocessing of data for three datasets. Further, to stage 3 where data transformation is done on the clean data.

3. *Data Transformation*

This is the third stage of the KDD process in which data is transformed from one form to another using different strategies to suitably be used in implementing the data mining techniques. The data transformation for the considered datasets is taken place in the following manner.

i. Airline Delay Dataset:

a. Dimensionality Reduction

Before the transformation of variables is performed removing the irrelevant columns from the dataset like X, Cancelled, and CancellationCode hence resulting in changes in the structure of the dataset.

```
> flightdelay<-flightdelay[-c(1,23,24)]
> str(flightdelay)
'data.frame':   714306 obs. of  27 variables:
 $ Year             : int  2008 2008 2008 2008 2008 2008 2008 2008 2008 2008 ...
 $ Month            : int  1 1 1 1 1 1 1 1 1 1 ...
 $ DayofMonth       : int  3 3 3 3 3 3 3 3 3 3 ...
 $ DayOfWeek        : int  4 4 4 4 4 4 4 4 4 4 ...
 $ DepTime          : num  1829 1644 1452 1323 1416 ...
 $ CRSDepTime       : int  1755 1510 1425 1255 1325 1625 1945 1650 1230 1435 ...
 $ ArrTime          : num  1959 1845 1640 1526 1512 ...
 $ CRSArrTime       : int  1925 1725 1625 1510 1435 1735 2230 1815 1530 1745 ...
 $ UniqueCarrier    : chr  "WN" "WN" "WN" "WN" ...
 $ FlightNum        : int  3920 1333 675 4 54 623 362 422 1056 3244 ...
 $ TailNum          : chr  "N464WN" "N334SW" "N286WN" "N674AA" ...
 $ ActualElapsedTime: num  90 121 228 123 56 57 147 135 153 136 ...
 $ CRSElapsedTime   : num  90 135 240 135 70 70 165 145 180 130 ...
 $ AirTime          : num  77 107 213 110 49 47 134 118 143 121 ...
 $ ArrDelay         : num  34 80 15 16 37 19 64 72 29 21 ...
 $ DepDelay         : num  34 94 27 28 51 32 82 82 56 15 ...
 $ Origin           : chr  "IND" "IND" "IND" "IND" ...
 $ Dest             : chr  "BWI" "MCO" "PHX" "TPA" ...
 $ Distance         : int  515 828 1489 838 220 220 972 765 1052 888 ...
 $ TaxiIn           : num  3 6 7 4 2 5 6 6 5 7 ...
 $ TaxiOut          : num  10 8 8 9 5 5 7 11 5 8 ...
 $ Diverted         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CarrierDelay     : num  2 8 3 0 12 7 5 3 0 0 ...
 $ WeatherDelay     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ NASDelay         : num  0 0 0 0 0 0 0 0 0 6 ...
 $ SecurityDelay    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ LateAircraftDelay: num  32 72 12 16 25 12 59 69 29 15 ...
```

Fig 16: Removing unwanted columns from airline delay dataset

ii. Airline Passenger Dataset

a. Dimensionality Reduction

The columns that are not useful in performing the analysis are removed in order to reduce the dimension of the data, which is a step of dimensionality reduction. The variables namely Id, Gender, Customer type, Type of Travel, Departure Delay in minutes, Arrival Delay in minutes are not useful in the analysis and hence removed thus resulting in dimensionality reduction.

b. Ensuring an equal proportion of the target variable

In order to get proper results from the model, the target variable must be in equal proportions so that the output of the model is not biased.

iii. Flight Fare Dataset

a. Conversion of datatypes

In this dataset, a lot of data transformation was required to get the data in the required format. The variable Date_of_Journey was transformed from char to date datatype. Variables Dep_Time and Arrival_Time are transformed in the time format. The duration column was converted in minutes such that it can be easily used in the model prediction. The below figure shows a glimpse of variable transformation.

```
> glimpse(airfare)
Rows: 9,558
Columns: 11
$ Airline          <chr> "IndiGo", "Air India", "IndiGo", "IndiGo",
$ Date_of_Journey  <date> 2019-03-24, 2019-05-01, 2019-05-12, 2019-0
$ Source           <chr> "Banglore", "Kolkata", "Kolkata", "Banglore
$ Destination      <chr> "New Delhi", "Banglore", "Banglore", "New D
$ Route            <chr> "BLR → DEL", "CCU → IXR → BBI → BLR", "CCU
$ Dep_Time         <time> 22:20:00, 05:50:00, 18:05:00, 16:50:00, 09
$ Arrival_Time     <time> 01:10:00, 13:15:00, 23:30:00, 21:35:00, 11
$ Duration         <dbl> 170, 445, 325, 285, 145, 930, 1265, 1530, 4
$ Total_Stops      <chr> "non-stop", "2 stops", "1 stop", "1 stop",
$ Additional_Info  <chr> "No info", "No info", "No info", "No info",
$ Price            <dbl> 3897, 7662, 6218, 13302, 3873, 11087, 2227(
>
```

Fig 17: Glimpse of data transformation to Flight Fare Dataset

4. *Data Mining*

In the fourth stage of the KDD process, proper data mining methods are chosen based on the research questions defined above. The different types of data mining tasks are clustering, classification, regression, etc. which have different approaches in applying the algorithms. After choosing a suitable method of data mining, data mining algorithms falling into these tasks are then selected based on different patterns of the data to perform the predictions. Then, finally, the selected algorithm is implemented and the results are obtained for a particular data mining model.

These three steps of data mining are performed on the below three datasets.

i. Airline Delay Dataset

This dataset was used to predict the delays in flight. The data mining task like classification is used to classify the result of predictions into different classes. The passengers as well as the airlines usually want to know the probability of flight delays and to know the factors which might result in the delaying of flights. Using logistic regression for finding these probabilities is easily interpretable as the sign of weights for the features selected tells whether a flight will be less likely or more likely delayed depending on the value of that feature.

A. Implementing Logistic Regression

- The delay in flights had many approaches to be predicted. Either predicting if the flight was delayed or not in two classes, predicting by how much time the flight will be delayed or predicting the delay into multiple classes like less delayed, more delayed or highly delayed so that the passengers are notified prior which will help them to plan their schedule accordingly.
- The third approach of classifying the delay into three ordered classes was chosen to perform logistic regression. This required the delay variable to be divided into three categories so that it could be used as the target variable.
- The logistic regression with target variable consisting of multiple factors that are ordinal that is they are in some level or order, then it is termed as ordinal logistic regression.
- The ordinal logistic regression is then applied to this dataset.
- The independent variables are chosen such that they have a high probability of predicting the price delay.

ii. Airline Passenger Dataset

Data mining is the process of getting patterns from the data. In this dataset, the satisfaction of airline passengers is predicted. The passenger needs to be classified into a satisfied or dissatisfied class. Thus, the classification machine learning method is needed here to predict the appropriate class. The most commonly used classification algorithms namely K-Nearest Neighbours and Decision Tree are selected here for performing data mining on this dataset.

A. Implementing the K-NN Algorithm :

- Preprocessed and Transformed Data in the previous stage is used for the application of K-NN Classification to the dataset.
- In order to provide the factor variable like the Class variable to the model, it was first converted to ordered factors and then to numeric data.
- The complete dataset is divided into training and test data using a random sampling of a dataset. 85% of the complete dataset is used for training the model and 15% is used for testing the model.
- The predictors and target variables are stored separately in different data structures for both training data as well as test data.
- The training and testing data are tried to be divided such that the model is trained equally for both the classes while training the data.
- Once we have the training and test data ready, the algorithm is implemented by sending the parameters to the model like training data, test data, class into which the model is to be classified(here it is target variable, satisfied and dissatisfied), and the k-value.
- K-value in the K-NN model is usually determined as the square root of the number of observations considered in the model. In this case, the observations are too huge, and hence square root won't be an option. So, different k-values were tried to fit the model and the best value for k was chosen which was k=7.
- The results of the model were then interpreted and evaluated that are explained in the next section.

B. Implementing Decision Tree

- A similar process of providing the decision tree model the preprocessed and transformed is applied here.
- The data is again divided into training and testing data for modeling purposes. The model predicts future data based on past data. This can be well performed by using the approach of training the model on training data first and then predicting the target values for test data.
- The training data is 80% of the complete dataset and test data is the remaining part.
- The target variable in both training, as well as testing, is checked if both have an equal distribution of classes in the target variable so that the model is well trained for both the classes.
- Rpart library is used for modeling decision tree with satisfaction as the dependent variable and all other variables except for the target variable are given as predictors to the model.
- Predict function is used to predict the model on test data.
- Finally, the results are interpreted and are explained n the next section.

iii. Flight Fare Dataset

- In this dataset, the flight prices were to be predicted and so the regression methods for data mining were selected. Since the regression models require a continuous target variable, the price of the airlines given in the dataset is our target variable and is predicted using two different regression models. The Random Forest and Support vector machines are two commonly used regression models and hence are selected in doing this price prediction.

  The steps followed in implementing these algorithms were as follows:

A. Implementing the Random Forest Algorithm :

- The Random Forest model requires all the numeric variables as predictors. So, transforming the variables Destination, Source, airline, and Total stops are first converted to factors and then to numeric variables.
- Creating features and target variables is one of the important steps in this model. The features are selected namely Airline, date of journey, destination, source, Departure time and arrival time, Duration, and Total_stops. The target variable Price is selected and assigned to a variable Y.
- Data is then split into training and test dataset using a stratified sampling method with training data 75 % of the complete dataset.
- The model is then implemented using the training data and testing data variables created in the previous step.

The other parameters like maxnodes and ntree are also best chosen in order to obtain the best results from the model.

- The price prediction for test data is done on the model obtained from training data using the predict() function.

B. Implementing Support Vector Machines Regression

- Implementing this model requires all the predictor variables either in factors or numeric values. They should be in a particular range to avoid the larger range variable dominance
- The similar steps are repeated for dividing the data into train and test using a random sampling method.
- Apply the SVM model to the training dataset with predictors same as that of the Random forest model and target variable Price that are passed to the sm() function.
- Predict the test results with the help of the predicted function.
- Plot the SVM model with actual data and predicted data that will help to interpret the results and will be explained in the next section.

IV. EVALUATION

After the implementation of all five models above, the interpretation and evaluation of these models is an important stage of the KDD process. The Prediction Models evaluation is explained below for all the five Models.

i. Airline Delay Dataset

a. Logistic Regression for airline delay prediction

In this model, the random sampling method is used for partitioning the dataset into training and testing.

The AIC value is obtained by checking the model summary. The lesser the AIC value better is the model.

```
polr(formula = data_train$ArrDelay_div ~ CarrierDelay + WeatherDelay +
    LateAircraftDelay + SecurityDelay + TaxiIn + TaxiOut, data = data_train,
    Hess = TRUE)

Coefficients:
                 Value Std. Error t value
CarrierDelay   0.10386  0.0002471  420.40
WeatherDelay   0.10950  0.0004317  253.63
LateAircraftDelay 0.10983  0.0002428  452.36
SecurityDelay  0.09444  0.0017290   54.63
TaxiIn         0.09185  0.0010889   84.35
TaxiOut        0.05482  0.0004582  119.62

Intercepts:
                            Value    Std. Error t value
less delay|moderately delay   3.3612   0.0122   274.9144
moderately delay|high delay   7.4607   0.0167   446.1484

Residual Deviance: 761650.34
AIC: 761666.34
> coefs
     CarrierDelay      WeatherDelay LateAircraftDelay     SecurityDelay           TaxiIn          TaxiOut
       0.10386420        0.10949816        0.10982867        0.09444492        0.09185173        0.05481570
>
```

Fig 17: summary for the ordinal logistic regression model

The confusion matrix shows the correctly classified 71% of the total observations. The coefficients of the predictor variables determine the log of odds of the variable to predict the target variable. The coefficients seen by the summary of the model say that weather delay has the highest probability of predicting the airline delay.

```
> table(data_test$ArrDelay_div, predictdelay)
                 predictdelay
                  less delay moderately delay high delay
less delay             50284             23618          0
moderately delay       15464             99965       3667
high delay              7072              4653      49457
```

Fig 18: confusion matrix for the ordinal logistic regression model

ii. Airline Passenger Dataset

a. KNN Classification for predicting airline passenger satisfaction

In this model, the random sampling method is used to partition the dataset into train and test. Various evaluation parameters are used to interpret the result obtained from the model.

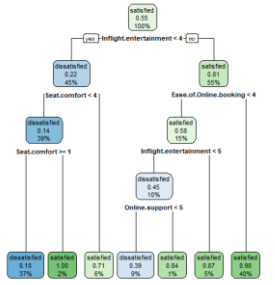
One of the common measures for evaluating the performance of the model is using the CrossTable function in the gmodels package. A cross tabular structure is created to indicate the result between two vectors. The below figure shows the CrossTable for the knn model with the initial value of k. The accuracy for the model can be calculated using the formula as (TP+TN)/(TP+FP+TN+FN) Accuracy of the knn model with k value 7 is 77% which is quite good. The Kappa statistic is 0.54 which is moderate.

```
Total Observations in Table:  19043

| knn_result
test_variable | dissatisfied |    satisfied |    Row Total |
--------------|--------------|--------------|--------------|
 dissatisfied |         6514 |         2135 |         8649 |
              |        0.753 |        0.247 |        0.454 |
              |        0.749 |        0.206 |              |
              |        0.342 |        0.112 |              |
--------------|--------------|--------------|--------------|
    satisfied |         2178 |         8216 |        10394 |
              |        0.210 |        0.790 |        0.546 |
              |        0.251 |        0.794 |              |
              |        0.114 |        0.431 |              |
--------------|--------------|--------------|--------------|
 Column Total |         8692 |        10351 |        19043 |
              |        0.456 |        0.544 |              |
--------------|--------------|--------------|--------------|

Confusion Matrix and Statistics

               test_variable
knn_result      dissatisfied satisfied
  dissatisfied          6514      2178
  satisfied             2135      8216

               Accuracy : 0.7735
                 95% CI : (0.7675, 0.7794)
    No Information Rate : 0.5458
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5434

 Mcnemar's Test P-Value : 0.5225

            Sensitivity : 0.7532
            Specificity : 0.7905
         Pos Pred Value : 0.7494
         Neg Pred Value : 0.7937
             Prevalence : 0.4542
         Detection Rate : 0.3421
   Detection Prevalence : 0.4564
      Balanced Accuracy : 0.7718

       'Positive' Class : dissatisfied
```

Fig 19: Confusion Matrix and Cross Table for KNN Classification model

b. Decision Tree Model for predicting airline passenger satisfaction

Random sampling is done in this model for splitting the data. The decision tree considers the best affecting predictors to give a decision for the classification of the dependent variable. The model can be run by selecting appropriate predictors to get better accuracy. The Decision Tree model is evaluated using the confusion matrix table which shows an accuracy of 86.5%




```
> confusionMatrix(table(data_test$satisfaction,predict_unseen))
Confusion Matrix and Statistics

               predict_unseen
                dissatisfied satisfied
  dissatisfied          12390      2048
  satisfied              2234     15065

               Accuracy : 0.8651
                 95% CI : (0.8613, 0.8688)
    No Information Rate : 0.5392
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7282

 Mcnemar's Test P-Value : 0.004696

            Sensitivity : 0.8472
            Specificity : 0.8803
         Pos Pred Value : 0.8582
         Neg Pred Value : 0.8709
             Prevalence : 0.4608
         Detection Rate : 0.3904
   Detection Prevalence : 0.4549
      Balanced Accuracy : 0.8638

       'Positive' Class : dissatisfied
```

Fig 20: Decision Tree and confusion matrix with statistics

Around 14% of the passengers were incorrectly classified as satisfied and around 12% were incorrectly classified as dissatisfied. These values are very less compared to the correct classification. Hence, we can say the model was well fit to answer the research question of predicting airline passenger satisfaction compared to the K-NN classification model.

```
> table_mat
              predict_unseen
               dissatisfied satisfied
  dissatisfied        12390      2048
  satisfied            2234     15065
> accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
> print(paste('Accuracy for test', accuracy_Test))
[1] "Accuracy for test 0.865078614865929"
> |
```

Fig 21: Confusion matrix for Decision Tree

iii. Flight Fare Dataset

a. Random Forest Regression Model for predicting price prediction of airlines

In this Model, the Stratified data sampling method is used. Stratified sampling retains the distribution of the target variable. The function named createDataPartition is used for this type of sampling which is a part of the "caret" package in R.

Below evaluation parameters are checked for Random Forest Regression Model.

Initially, on training the model the values for R^2 was 69% and RMSE values around 2338 were noted which can be taken as were considerate values as 69% accurately the model was able to predict the price. Checking the actual value vs. predicted value, we can say that the values are close enough as seen in the below figure. This accuracy of the model can be considered to be good. But it can be improved by tuning the model with different ntree and max nodes values.

actual_price	predicted_price
<dbl>	<dbl>
6218	6676.
11087	11560.
5830	7004.
10262	11109.
12898	11109.
4423	4755.

Fig 22: actual vs predicted values for airfare

The graph below explains that as the number of trees increases the error rate is decreasing. But a large number of trees in the model might result in overfitting. So, the trees must be a moderate value, that is not too less and not too large to avoid overfitting as well as underfitting of values.

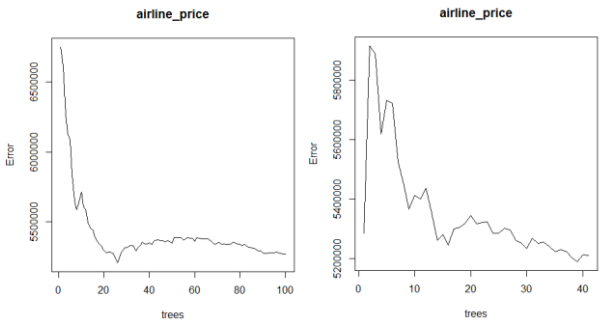


Fig 23: Error plot with a number of trees 100 and 41

The minimum number of trees with the lowest RMSE can be obtained from the function seen below.

```
> which.min(airline_price$mse)
[1] 26
> # RMSE of this optimal random forest
> sqrt(airline_price$mse[which.min(airline_price$mse)])
[1] 2281.321
```

Fig 24: Tree with lowest RMSE

As the 26th tree is having the lowest RMSE value of 2281. We can select the number of trees in this range of values. Hence, applying random forest again with the max nodes of 20 and number of trees equal to 41. This lessens the RMSE value than the initial model as seen in the below figure.

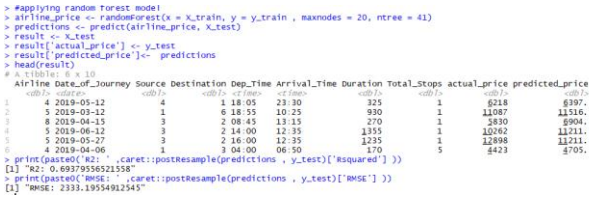


Fig 25: predicted results for the model

Thus, the price prediction of airlines with Random forest is done with around 70% accuracy.

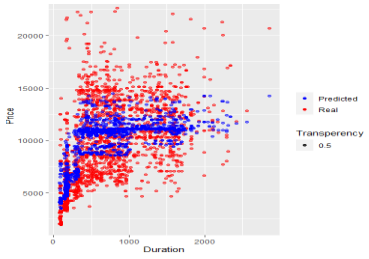


Fig 26:Plot for Actual vs predicted values

b. Support Vector Machine-Regression for predicting price prediction of airlines

The goal of SVM Regression is the same as that of Classification that is to extract the maximum margin, which is nothing but to minimize the error. The complete dataset is divided into training and testing data using random sampling in a 3:1 ratio.

The evaluation parameters used for SVM regression are RMSE values for the initial model was noted which was about 2280. The actual values vs the predicted values are plotted against each other in different colors such that we

can determine the margin of error. In the below first plot significant outliers can be seen after plotting the initial SVM.

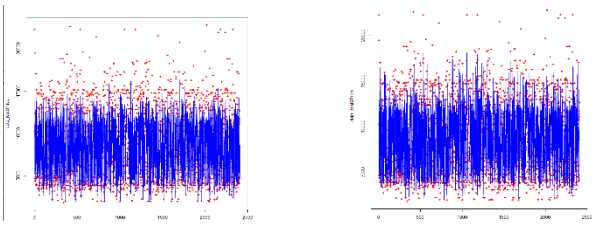


Fig 27:Plot SVM Regression model using different cost values

Tuning of SVR is done to minimize the error rate and best fit the model with various parameters like maximum allowable error rate and cost parameter. Optimum value parameters are found using the tuned model.

The tune function evaluates the performance of each combination of the MAE called epsilon and cost parameter. It evaluates 50*11 i.e. 550 models. Hence we can find the best model with the lowest MSE by performing the tuning of the model.

The best model for SVR is chosen with a cost parameter equal to 45 and radial kernel with type eps_regression. This model has reduced the RMSE value to 20% reduced errors.

V. Conclusions and Future Work

This paper on airline data study explains and shows a comparative analysis of the results of 5 machine learning models using the KDD process of data mining. These models are successfully applied to the three different datasets and have provided adequate results concerning the data given to the model. The Logistic Regression Model, KNN classification model, Decision Tree, and Random Forest Regression and Support Vector Machine Regression Model gave considerate results on the performed datasets. As the models were applied on different datasets direct comparison between the accuracies is difficult. Overall Decision Tree model performed the best for fitting airline passenger data to predict their satisfaction with an accuracy of 86.5%. SVR (Support Vector Machine Regression) was polished by evaluating the performance of 550 models and the model with the least RMSE value was chosen as the final model for predicting future prices for airlines. Among Random forest and SVM, SVM predicted the airline prices with more accuracy as the RMSE value for the SVM model was less compared to the Random Forest Model. The evaluation of models was done using R2, RMSE, AIC, CrossTables, and Confusion matrix.

Comparing the Decision Tree and KNN classification both the models were well fitted and can be tuned more for better predictions in future work. Stacking machine learning algorithms can be applied in delay prediction of flights to create a better prediction model. Feature engineering can be performed on Flight Fare Dataset to better analyze the dynamic pricing of flights.

REFERENCES

[1] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 1310-1315.

[2] Bahadir, Cuneyt, and Adem Karahoca. “Airline Revenue Management via Data Mining.” Global Journal of Information Technology: Emerging Technologies 7.3 128–148. Web.

[3] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta and S. Barman, "A statistical approach to predicting flight delay using gradient boosted decision tree," 2017 International Conference on Computational Intelligence in Data Science(ICCIDS), Chennai, 2017, pp. 1-5, doi: 10.1109/ICCIDS.2017.8272656.

[4] L. Zhang, Y. Sun and T. Luo, "A framework for evaluating customer satisfaction," 2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), Chengdu, 2016, pp. 448-453, doi: 10.1109/SKIMA.2016.7916264.

[5] V. H. Vu, Q. T. Minh and P. H. Phung, "An airfare prediction model for developing markets," 2018 International Conference on Information Networking (ICOIN), Chiang Mai, 2018, pp. 765-770, doi: 10.1109/ICOIN.2018.8343221.

[6] Namukasa, Juliet. (2013). The influence of airline service quality on passenger satisfaction and loyalty: The case of Uganda airline industry. The TQM Journal. 25. 10.1108/TQM-11-2012-0092.

[7] S. Srinidhi, "Development of an airline traffic forecasting model on international sectors," 2009 IEEE International Conference on Automation Science and Engineering, Bangalore, 2009, pp. 322-327, DOI: 10.1109/COASE.2009.5234138.

[8] V. Tsoukas, K. Kolomvatsos, V. Chioktour and A. Kakarountas, "A Comparative Assessment of Machine Learning Algorithms for Events Detection," 2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Piraeus, Greece, 2019, pp. 1-4, doi: 10.1109/SEEDA-CECNSM.2019.8908366.

[9] Shafique, Umair & Qaiser, Haseeb. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research. 12. 2351-8014.

[10] S. M. T. F. Ghomi and K. Forghani, "Airline passenger forecasting using neural networks and Box-Jenkins," 2016 12th International Conference on Industrial Engineering (ICIE), Tehran, 2016, pp. 10-13, doi: 10.1109/INDUSENG.2016.7519342.

[11] S. S. Maini and K. Govinda, "Stock market prediction using data mining techniques," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, 2017, pp. 654-661, doi: 10.1109/ISS1.2017.8389253.

[12] Y. J. Kim, S. Choi, S. Briceno and D. Mavris, "A deep learning approach to flight delay prediction," 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, 2016, pp. 1-6, doi: 10.1109/DASC.2016.7778092.