### DEVELOPING A FLIGHT DELAY PREDICTION MODEL USING MACHINE LEARNING

### Submitted by

**TEAM ID: PNT2022TMID52569**

**1905016 HANITAA SRI VR**

**2005203 GOMATHI T**

**1905002 AISHWARYA V**

**1905043 SADASVI M**

**in partial fulfillment for the award of the degree of**

**BACHELOR OF ENGINEERING**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

****

**COIMBATORE INSTITUTE OF TECHNOLOGY**

**(Government Aided Autonomous Institution Affiliated to Anna University)**

**COIMBATORE-641 014**

**ANNA UNIVERSITY - CHENNAI 600 025**

**NOVEMBER 2022**

**1. INTRODUCTION**

**1.1 Project overview**

One of the key business issues that airlines face is that the vital prices that are related to flights being delayed because of natural occurrences and operational shortcomings that is an upscale affair for the airlines, making issues in scheduling and operations for the end users therefore inflicting unhealthy name and client discontent. As we all know that we have a tendency to not get the flight delay before departure as customers of the Airline Company neither the airline company’s ground staff gets the airline delay prediction supported varied conditions. However, we all know that one in all the most reasons for delay in flights is that the weather. This motivates us to use the live weather knowledge in conjunction with different metrics to calculate the delay on the wing before departure. Indian state of affairs, in 2017, in line with the reports by the directorate General of Civil Aviation (DGCA), between January and April, close to 5.12 hundred thousand domestic passengers in India faced issues because of airline corporations denying boarding, moreover as flight cancellations and delays [2]. Airline corporations had to pay the passengers compensations of over Rs. twenty five core for varied inconveniences throughout the first four months of this year. Hence, the prediction analysis retrieved from this project can contribute within the form of a prototype in helping to identify operational variables that contribute to delays in any country scenario[2] The main issues associated with flight delay prediction are known and arranged in taxonomy. It includes the problem that causes the flight delay, the range of institution it affects, and ways that of handling flight delay prediction downside. It considers flight domain options, like problem and scope. Major problem which causes delay in flights can be delay propagation, delay caused on the departure point or the root of the flight, and cancellation of flights. These problems cannot be eliminated forever, but a delay prediction tool will allow the operator and the administrators to take the concerned actions for smooth operation. This problem that is causes delay affects Airline, Airport and the enroute airspace which are independent entities which works in synchronization and hence delay in flight causes issues in all the sectors. Various methods that can be used to develop a system which predicts the delay in flights can be Machine Learning, Probabilistic models, Statistical analysis or Network Representations.

**1.2 Purpose**

Flight Planning is one of the challenges in industrial world which faces many uncertain conditions. One such condition is delay occurrence, which stems from various factors and imposes considerable costs on airlines, operators, and travelers. Delays in departure can occur due to bad weather conditions, seasonal and holiday demands, airline policies, technical issue such as problems in airport facilities, luggage handling and mechanical apparatus, and accumulation of delays from preceding flights. Here in flight delay prediction system based on the weather parameters which can result in delays. The system considers the temperature, humidity, rain in mm, visibility and month number as important parameters for prediction of delay.

**2. LITERATURE SURVEY**

**Flight Delay Prediction Based on Aviation Big Data and Machine Learning**

Accurate flight delay prediction is fundamental to establish the more efficient airline business. Recent studies have been focused on applying machine learning methods to predict the flight delay. Most of the previous prediction methods are conducted in a single route or airport. This paper explores a broader scope of factors which may potentially influence the flight delay, and compares several machine learning-based models in designed generalized flight delay prediction tasks. To build a dataset for the proposed scheme, automatic dependent surveillance broadcast (ADS-B) messages are received, pre-processed, and integrated with other information such as weather condition, flight schedule, and airport information. The designed prediction tasks contain different classification tasks and a regression task. Experimental results show that long short-term memory (LSTM) is capable of handling the obtained aviation sequence data, but overfitting problem occurs in our limited dataset. Compared with the previous schemes, the proposed random forest-based model can obtain higher prediction accuracy (90.2% for the binary classification) and can overcome the overfitting problem.

**Machine Learning Model - based Prediction of Flight Delay**

Prior prediction of flight arrival delays is necessary for both travellers and airlines because delays in flights not only trigger huge economic loss but also airlines end up losing their reputation that was built for several years and passengers lose their valuable time. Our paper aims at predicting the arrival delay of a scheduled individual flight at the destination airport by utilizing available data. The predictive model presented in this work is to foresee airline arrival delays by employing supervised machine learning algorithms. US domestic flight data along with the weather data from July 2019 to December 2019 were acquired and are used while training the predictive model. XG Boost and linear regression algorithms were applied to develop the predictive model that aims at predicting flight delays. The performance of each algorithm was analyzed. Flight data along with the weather data was given to the model. Using this data, binary classification was carried out by the XG Boost trained model to predict whether there would be any arrival delay or not, and then linear regression model predicts the delay time of the flight. the estimation of delay time in minutes using machine learning algorithms namely Decision Tree Algorithm (XGBoost) and Linear regression. Data set of both flights and weather will be taken to compare with the given inputs and validate them by applying classification and Regression concepts of Machine Learning. Also having done feature extraction, handling missing values using appropriate methods, sampling to handle imbalanced data and also tuning the hyperparameters with which better accuracy was able to be achieved.

**A Deep Learning Approach to Flight Delay Prediction**

Deep learning has achieved significant improvement in various machine learning tasks including image recognition, speech recognition, machine translation and etc. Inspired by the huge success of the paradigm, there have been lots of tries to apply deep learning algorithms to data analytics problems with big data including traffic flow prediction. However, there has been no attempt to apply the deep learning algorithms to the analysis of air traffic data. This paper investigates the effectiveness of the deep learning models in the air traffic delay prediction tasks. By combining multiple models based on the deep learning paradigm, an accurate and robust prediction model has been built which enables an elaborate analysis of the patterns in air traffic delays. In particular, Recurrent Neural Networks (RNN) has shown its great accuracy in modeling sequential data. Day - to-day sequences of the departure and arrival flight delays of an individual airport have been modeled by the Long Short Term Memory RNN architecture. It has been shown that the accuracy of RNN improves with deeper architectures. In this study, four different ways of building deep RNN architecture are also discussed. Finally, the accuracy of the proposed prediction model was measured, analyzed and compared with previous prediction methods. It shows best accuracy compared with all other methods.

**Prediction of Weather-induced Airline Delays Based on Machine Learning Algorithms**

The primary goal of the model proposed in this paper is to predict airline delays caused by inclement weather conditions using data mining and supervised machine learning algorithms. US domestic flight data and the weather data from 2005 to 2015 were extracted and used to train the model. To overcome the effects of imbalanced training data, sampling techniques are applied. Decision trees, random forest, the AdaBoost and the k Nearest-Neighbors were implemented to build models which can predict delays of individual flights. Then, each of the algorithms’ prediction accuracy and the receiver operating characteristic (ROC) curve were compared. In the prediction step, flight schedule and weather forecast were gathered and fed into the model. Using those data, the trained model performed a binary classification to predicted whether a scheduled flight will be delayed or on-time.

The model was built on historical weather and traffic data of individual OD pair by utilizing machine learning algorithms. Supervised machine learning algorithms implemented in this study includes random forest, AdaBoost, k Nearest-Neighbors and Decision Trees. Because the data was imbalanced, the combination of SMOTE and random under sampling were applied. The model’s prediction performance on the validation set and the test set was analyzed. There are still possible approaches that can improve the model in the future. If the costs of false positive and false negative are taken into account, preferred performance of classifiers could be clearly determined. Then it could be a solid foundation for a decision support tool for predicting aircraft arrival. Also a thorough consideration of uncertainty in forecast would enhance the model’s predictive performance.

**2.1 Existing problem**

As discussed, considering the standard taxonomy of the flight delay and its problems, one will contemplate the scope of prediction to be one in every of these factors or combination of those factors[3]. The models developed during this system may be applied to predict the incidence of flight delay at airports. Such prognosticative capabilities would facilitate traffic managers and airline dispatchers to organize mitigation methods for reducing traffic disruptions. This issue can be reduced by developing the flight delay prediction tool which can be developed using following methods.

**2.2 References**

[1] Kuhn, Nathalie and Navaneeth Jamadagni. “Application of Machine Learning Algorithms to Predict Flight Arrival Delays.” (2017).

[2] N, Prabakaran & Kannadasan, Rajendran. (2018). Airline Delay Predictions using Supervised Machine Learning. International Journal of Pure and Applied Mathematics. 119.

[3] A Review on Flight Delay Prediction Alice Sternberg, Jorge Soares, Diego Carvalho, Eduardo Ogasawara \_ CEFET/RJ Rio de Janeiro, Brazil November 6, 2017 International Journal of Engineering Research & Technology (IJERT) [http://www.ijert.org](http://www.ijert.org/) ISSN: 2278-0181 IJERTV9IS030148 (This work is licensed under a Creative Commons Attribution 4.0 International License.) Published by : [www.ijert.org](http://www.ijert.org/) Vol. 9 Issue 03, March-2020 91

[4] Gopalakrishnan, Karthik and Hamsa Balakrishnan. “A Comparative Analysis of Models for Predicting Delays in Air Traffic Networks.” Air Traffic Management Research and Development Seminar, June 2017, Seattle, Washington, USA, ATM Seminar, June 2017 © 2017 ATM Seminar

[5] Rebollo, Juan Jose and Balakrishnan, Hamsa. “Characterization and Prediction of Air Traffic Delays.” Transportation Research Part C: Emerging Technologies 44 (July 2014): 231–241 © 2014 Elsevier LtdA model for acurracy prediction using geoRSS using naive bayes

<https://doi.org/10.24200/sci.2017.20020>

<https://doi.org/10.1177/0361198120930014>

**2.3 Problem Statement definition**

In this project, the goal is to use exploratory analysis and to build **machine learning** models to **predict airline** departure and arrival **delays**.

**OBJECTIVES:**

* Flight delays are gradually increasing and bring more financial difficulties and customer dissatisfaction to airline companies.
* To resolve this situation, supervised machine learning models were implemented to predict flight delays.
* The data set that records information of flights departing from JFK airport during one year was used for the prediction.
* The comparative analysis showed that the Decision Tree algorithm has the best performance with an accuracy of 0.9777, and the KNN algorithm has the worst performance with an f1-score of 0.8039.
* Tree-based ensemble classifiers generally have better performance over other base classifiers.

**3. PROPOSED SOLUTION**:

**3.1 Proposed Solution**

As discussed, weather condition plays an important role in proper and timely functioning of flights. We propose a flight delay prediction system which focuses mainly on predicting delay of a flight based on the weather situation. To make the system more scalable it is necessary to choose an algorithm which considers all the parameters to be independent. Supervised learning as the name indicates a presence of supervisor as teacher. Essentially supervised learning could be a learning that within which we tend to teach or train the machine exploitation data which is well tagged which means some data is already labeled with correct answer. After that, machine is given new set of examples(data) so supervised learning algorithmic rule analyses the coaching knowledge(set of training examples) and produces an correct outcome from tagged data Using supervised machine learning approach, the labeled data gives it authenticity. Naïve bayed model is one of the algorithm which is proven to be efficient for real time prediction as well as the fact that it considers every attribute to be independent from each other makes it an apt algorithm for the concerned project The proposed system takes the city of departure as its input in a textbox field as shown in fig 5 and 6. It then returns the predicted weather data using an API (Application Program Interface) and passes the data into the algorithm. The attributes considered for calculations and taken by the API are as follows weather, temperature, humidity, Rain in mm, Visibility and Month number. As discussed that supervised machine learning is based on having a set of correct labeled data form which the algorithm bases its prediction. We use a CSV file for storing that data as a flat file format is easier to edit , update and retrieve it for calculations.

**3.4 Problem Solution fit**

### Flight delays are gradually increasing and bring more financial difficulties and customer dissatisfaction to airline companies. To resolve this situation, supervised machine learning models were implemented to predict flight delays. The data set that records information of flights departing from JFK airport during one year was used for the prediction. Seven algorithms (Logistic Regression, K-Nearest Neighbor, Gaussian Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest, and Gradient Boosted Tree) were trained and tested to complete the binary classification of flight delays. The evaluation of algorithms was fulfilled by comparing the values of four measures: accuracy, precision, recall, and f1-score. These measures were weighted to adjust the imbalance of the selected data set. The comparative analysis showed that the Decision Tree algorithm has the best performance with an accuracy of 0.9777, and the KNN algorithm has the worst performance with an f1-score of 0.8039. Tree-based ensemble classifiers generally have better performance over other base classifiers.Machine learning is the designation of algorithms that enable the computer to analyze the data, obtain potential patterns, and then use them to predict. Learning algorithms can give insight into the relative difficulty of learning in different environments . Machine learning algorithms are divided into several categories, and the two most common types are supervised learning and unsupervised learning. Algorithms of supervised learning generated a function that translates inputs to desired outputs. The primary forms of supervised learning algorithms include regression and classification. Unsupervised learning models a collection of inputs in the absence of labeled examples.

### Minor adjustments were made to the data set, including changing the variable “DEW\_POINT” from object datatype to integer datatype since it has numerical values and deleting rows containing null values. There were only two out of 28820 rows with missing values, which means that the deletion will have little effect on the overall data set distribution. Some of the selected algorithms can only deal with numerical data. As presented in the previous section, the data set contains both numerical variables and categorical variables. These variables need to be converted to numerical variables to avoid that these algorithms are failing to work when facing categorical data. Therefore, the technique of integer encoding, converting category labels to unique integer numbers, was used for training and new data [[7](https://dl.acm.org/doi/fullHtml/10.1145/3497701.3497725#bib7)]. Also, categorical variables such as “TAIL\_NUM”, which have little influence on predicting flight delays, were dropped.

In this study, a supervised machine learning approach was applied. The data set has a target variable, and the goal is often to let the computer learn a created classification system [[**8**](https://dl.acm.org/doi/fullHtml/10.1145/3497701.3497725#bib8)]. The main objective of this study is to predict flight delays based on labels data. Therefore, a supervised learning classification algorithm was selected as the appropriate one. The prediction of flight delays was considered a binary classification problem that uses given data to predict whether a flight delay will take place or not. After consultations with experts and previous works from the airline domain, the criteria was made that if the variable “DEP\_DELAY” (minute difference between scheduled departure time and actual departure time) is greater than 15, the flight is considered as delayed. Else it is not delayed. As a result, an additional binary variable “IS\_DELAY” was created with the value 1 when the flight is delayed and 0 when not delayed.Based on the variable “IS\_DELAY”, it can be seen that the data set consists of 3873 delayed flights and 24945 non-delayed flights, showing an imbalanced distribution since the majority of flights were not delayed. 10-fold cross-validation was used to resolve this problem. It created the training set and testing set. Each algorithm was run with the default parameters in the scikit-learn python package on the testing set, and the same training set was used for all algorithms.

**4. REQUIREMENT ANALYSIS**

**4.1 Functional requirements**

1) Supervised and Unsupervised learning concepts.

2) Classification, regression and clustering.

3) Flask.

4) Decision tree.

5) Gini index.

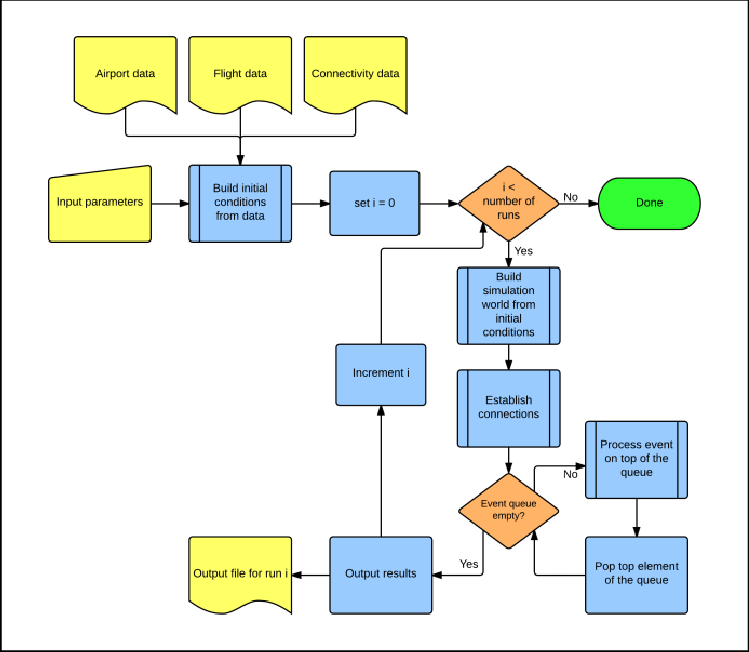
**4.2 Non- Functional requirements**

1) Accuracy predicting the delay timings.

2) Availability.

**5. PROJECT DESIGN**

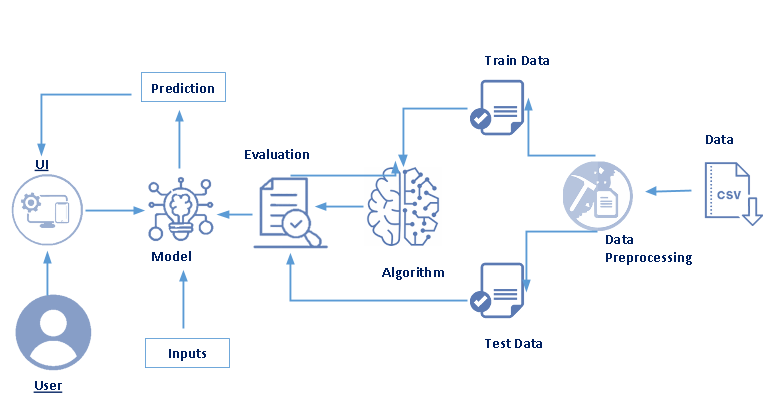
**5.1 Data flow diagrams**



**5.2 Solution & Technical Architecture**

**Solution**: Over the last twenty years, air travel has been increasingly preferred among travelers, mainly because of its speed and in some cases comfort. This has led to phenomenal growth in air traffic and on the ground. An increase in air traffic growth has also resulted in massive levels of aircraft delays on the ground and in the air. These delays are responsible for large economic and environmental losses. The main objective of the model is to predict flight delays accurately in order to optimize flight operations and minimize delays. Using a machine learning model, we can predict flight arrival delays. The input to our algorithm is rows of feature vector like departure date, departure delay, distance between the two airports, scheduled arrival time etc. We then use decision tree classifier to predict if the flight arrival will be delayed or not. A flight is considered to be delayed when difference between scheduled and actual arrival times is greater than 15 minutes. Furthermore, we compare decision tree classifier with logistic regression and a simple neural network for various figures of merit.

**Technical Architecture**



**6. PROJECT PLANNING & SCHEDULING**

**6.1 Sprint Planning & Estimation**

|  |  |
| --- | --- |
| **Problem statement planning** | **Timing** |
| [DATA COLLECTION](https://careereducation.smartinternz.com/Student/guided_project_workspace/48724#collapse6) | 05-09-2022 - 07-09-2022 |
| [DATA PRE-PROCESSING & MODEL BUILDING](https://careereducation.smartinternz.com/Student/guided_project_workspace/48724#collapse7) | 07-09-2022 - 12-09-2022 |
| T[RAIN THE MODEL ON IBM](https://careereducation.smartinternz.com/Student/guided_project_workspace/48724#collapse9) | 12-09-2022 - 14-09-2022 |
| [IDEATION PHASE](https://careereducation.smartinternz.com/Student/guided_project_workspace/48724#collapse10)-1 & 2 | 14-09-2022 - 26-09-2022 |
| [APPLICATION BUILDING](https://careereducation.smartinternz.com/Student/guided_project_workspace/48724#collapse8) | 26-09-2022 - 03-10-2022 |
| [PROJECT PLANNING PHASE](https://careereducation.smartinternz.com/Student/guided_project_workspace/48724#collapse13) | 10-10-2022 - 17-10-2022 |
| [PROJECT DEVELOPMENT PHASE](https://careereducation.smartinternz.com/Student/guided_project_workspace/48724#collapse14) | 17-10-2022 - 26-10-2022 |

**6.2 Sprint Delivery Schedule**

|  |  |  |
| --- | --- | --- |
| **Day** | **Date** | **Timeslot** |
| Day-1 | 5-09-2022 | 1:00pm - 4:00pm |
| Day-2 | 7-09-2022 | 6:00pm - 9:00pm |
| Day-3 | 12-09-2022 | 1:00pm - 4:00pm |
| Day-4 | 14-09-2022 | 6:00pm - 9:00pm |
| Day-5 | 19-09-2022 | 1:00pm - 4:00pm |
| Day-6 | 21-09-2022 | 6:00pm - 9:00pm |
| Day-7 | 26-09-2022 | 1:00pm - 4:00pm |
| **Day** | **Date** | **Timeslot** |
| Day-8 | 28-09-2022 | 6:00pm - 9:00pm |
| Day-9 | 3-10-2022 | 1:00pm - 4:00pm |
| Day-10 | 10-10-2022 | 1:00pm - 4:00pm |
| Day-11 | 12-10-2022 | 6:00pm - 9:00pm |
| Day-12 | 17-10-2022 | 1:00pm - 4:00pm |
| Day-13 | 19-10-2022 | 6:00pm - 9:00pm |
| Day-14 | 26-10-2022 | 6:00pm - 9:00pm |

**7. CODING & SOLUTIONING (Explain the features added in the project along with code)**

**7.1 Feature 1(phase 1)**

### Solution Architecture

**MODULES:**

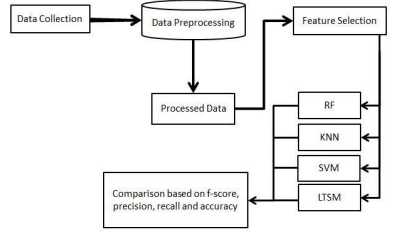
• Load dataset: We will upload our dataset into application.

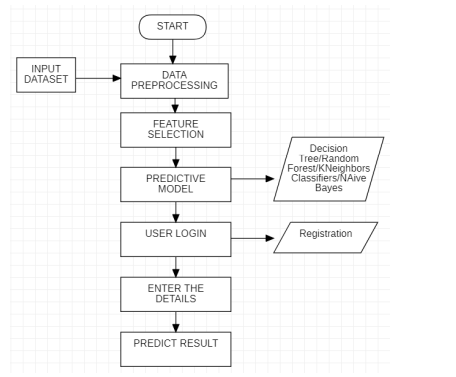
• Data Preprocessing: The quality of the data should be checked before applying our algorithms.

• Feature Extraction: Transforming raw data into numerical features that can be processed while preserving the information in the original data set.

• Generate models: After extracting features from the dataset we will generate our algorithms on that dataset.

• Accuracy Graph: We will plot the accuracies comparison graph between all the algorithms.



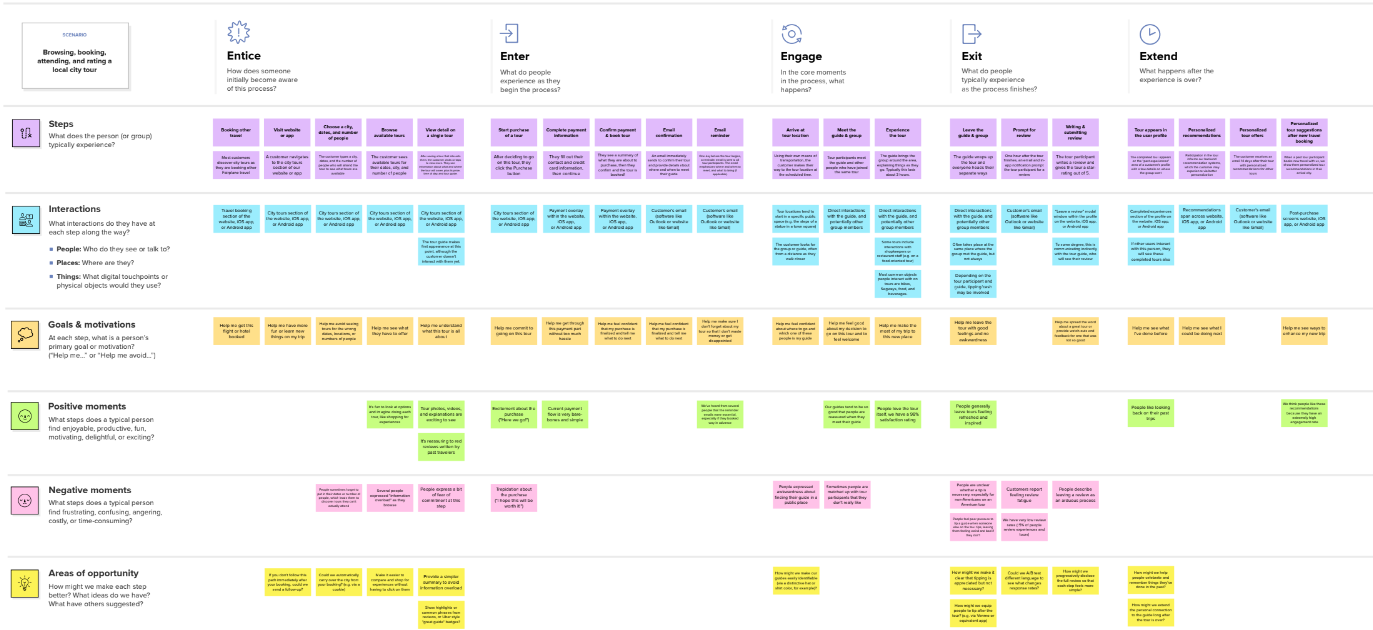


The above figure shows how the machine learning model was built. Deploying a machine learning model includes the following steps namely data collection, input Dataset, data preprocessing ,feature selection, predictive models such as Random Forest Model, Naïve Bayes, Decision Tree and KNeighbors classifier Input Dataset: Input dataset is given to the system to find the desired predicted output. Data preprocessing: Data preprocessing is the process in which the collected data is processed and cleaning the data and suitable for machine learning model so as to make execution efficiently. Feature Selection: Feature Selection is the process where we select the suitable features required for the model and avoiding unwanted data. Predictive Model is that to select the efficient model required and it will predict the result by analysing the existing data. Here after the user registration and then by logging into the system and we have to enter the required essentials. After entering the details in the required fields we can see the result of the flight delay.

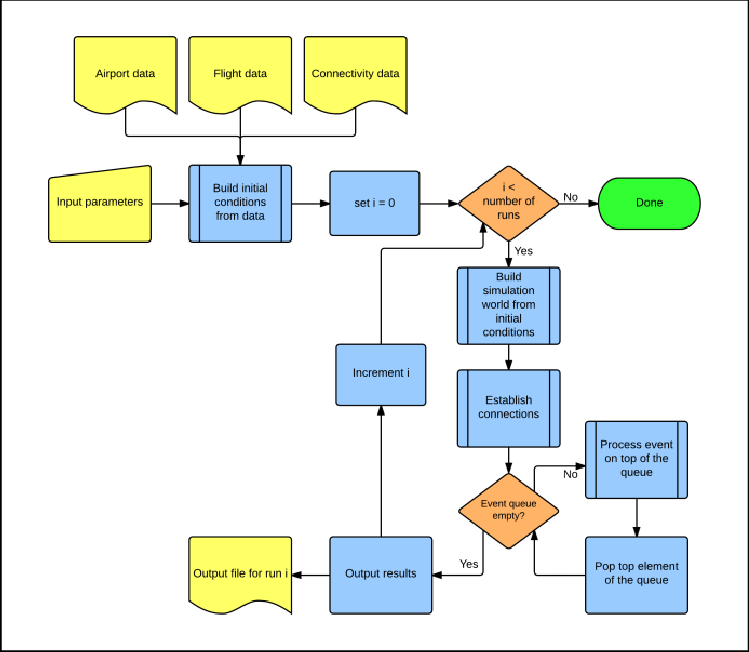
**7.2 Feature 2(phase-2)**

**Customer Journey**

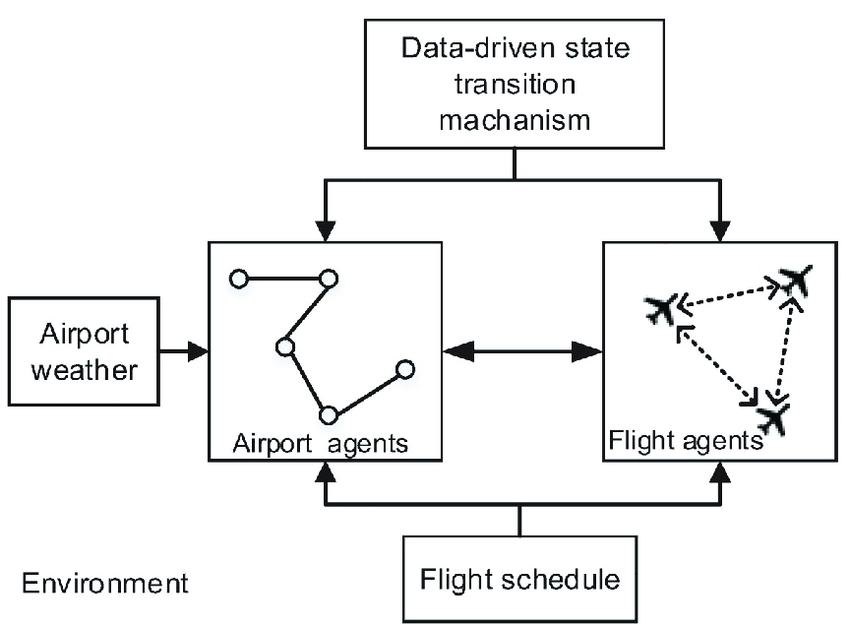
Since ﬂight delays cause economic consequences to passengers and airlines, recognizing them through prediction may improve marketing decisions. Due to that, several forecast models have been built over the last twenty years. These models have sought to understand how delays propagate through the network of ﬂights or airports, to predict root delay in the system or to comprehend the cancellation process. Beyond these three points of view for treating the ﬂight delay prediction problem, models could also diﬀer by their scope of application, data issues and methods.

****

**Data Flow Diagram & User Stories**



### Technology Architecture



Prior work has used simulation to investigate air traffic delay. However, these studies have notemphatically focused on the real-time parameters that significantly affect flight delay. Hence, itis difficult to apply these studies to flight delay prediction in the tactical phase. To fill this gap,we proposed a method combining simulation and data mining aiming to predict flight delaysin the entire air traffic system in the future. In particular, an agent-based model was con-structed, and several crucial time-varying parameters affected agent states obtained by datamining methods.

**8. TESTING**

**8.1 User Acceptance Testing**

If one flight is delayed, but still able to depart and arrive within its currently assigned ATFM slots, delay needs to be propagated to the next leg in the aircraft’s rotation and the passenger/crew connections (if any).

If a flight is delayed (because of a primary or a reactionary delay) and lost its slots, the simulation tries to find a new suitable pair of slots (first through re-scheduling, then through slot swapping), which also may cause delay to be propagated. If these processes fail, the flight and all the successive legs in the same aircraft’s rotation are cancelled.

If a flight has no delay, no measure is necessary, i.e. the flight will depart and land as scheduled.

**9. RESULTS**

**9.1 Performance Metrics**

Our proposed model had able to provide a better accuracy and have differentiated into training and testing model.

x\_test.shape

(67356, 8)

x\_train.shape

(269420, 8)

y\_test.shape

(269420, 1)

y\_train.shape

(67356, 1)

**10. ADVANTAGES & DISADVANTAGES**

**Advantage**

Therefore, predicting flight delays can **improve airline operations and passenger satisfaction**, which will result in a positive impact on the economy. In this study, the main goal is to compare the performance of machine learning classification algorithms when predicting flight delays.

**Disadvantage**

Ithas too much negative economic effects on passengers, agencies and airport [[7](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR7),[8](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR8),[9](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR9),[10](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR10),[11](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR11)]. Furthermore, delay can damage the environment through fuel consumption increment and also leads to emission of pollutant gases [[1](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR1), [12](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR12),[13](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR13),[14](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR14),[15](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR15),[16](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR16)]. In addition, the delay affects the trade, because goods’ transport is highly dependent to customer trust, which can increase or decrease the ticket sales, so that on time flight leads to customer confidence [[17](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR17), [18](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR18)]. So that, flight prediction can cause a skillful decision and operation for agencies and airports, and also a good passenger information system can relatively satisfy the customer [[19](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z#ref-CR19)].

**11. CONCLUSION**

This paper presented the need to develop a system to predict the delay in flights along with its methodology. The paper gives details about the range of different methodology that is used or can be used to find out the delay in flights. As flight delay cost a lot to the airlines as well as passangers in financial and environmental terms, flight delay is a the talk of the hour. Flight delay causes surging of prices by costing a lot on operational purpose They may increase prices to customers and operational prices to airlines. As the outcome is directly associated with the passanger and the airlines which inturn is liked to another set of airline and passangers it is very crucial to get real time delay for each player within the air transport system. hence there is a requirement to develop a system to predict the delay in flights to scale back monetory loss and for the higher and smooth operation. Classification or reggrerssion ways are often accustomed determine the delay which includes Feed forward network, Neural Network, Random Forrest, decision tress, Naïve Bayes Classification Tree, Regression Tree, etc. As seen from the articles and papers these methodologies offer virtually identical accuracy however we want an algorithmic rule that is good with real world prediction and analysis and thus: naïve Bayes.Except being smart with real time prediction algorithmic rule that considers or assumes independence among predictors that makes the system scalable as other independent attribute may be superimposed up to the algorithmic rule for computation of the delay. the expected delay can thus facilitate the ground employees for creating correct and smooth operation plans and therefore the data if sent to the passengers will profit the airlines also because the passengers

**12. FUTURE SCOPE**

Further supportive study is required to correlate all the problem, scope and method for getting most accurate result. Although weather conditions are the major reasons for flight delay, other unprecedented events such as major calamities, natural or man-made can cause major delay in flight.

**13. APPENDIX**

**Source Code:**

**import numpy as np**

**import pandas as pd**

import seaborn as sns

import pickle

%matplotlib inline

from sklearn.preprocessing import LabelEncoder

from sklearn.preprocessing import OneHotEncoder

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy\_score

import sklearn.metrics as metrics

dataset = pd.read\_csv("flightdata.csv")

dataset.info()

dataset.describe()

dataset.isnull().sum()

dataset['dest'].unique()

import matplotlib.pyplot as plt

plt.figure(1,figsize=(100,257))

plt.show()

def draw\_barplot(\_style, \_x,\_y,\_dataset,\_suptitle,\_xLabels,\_yLabels):

sns.set(style=\_style)

g = sns.catplot(x=\_x, y=\_y, data=\_dataset,aspect=2,

height=5, kind="bar", palette="muted")

plt.subplots\_adjust(top=0.9)

g.fig.suptitle(\_suptitle)

g.set\_xlabels(\_xLabels)

g.set\_ylabels(\_yLabels)

def plot\_CountPlot(X,Dataframe,Title):

fig = plt.figure(figsize=(15,5))

ax = sns.countplot(x=X, data=Dataframe ,palette='pastel' ,edgecolor=sns.color\_palette("dark", 10))

ax.set\_title(Title)

ax.legend(loc='upper right')

for t in ax.patches:

if (np.isnan(float(t.get\_height()))):

ax.annotate(0, (t.get\_x(), 0))

else:

ax.annotate(str(format(int(t.get\_height()), ',d')), (t.get\_x(), t.get\_height()\*1.01))

plt.show();

dt=dataset.groupby(by='origin').count().reset\_index().sort\_values(by='year', ascending=False)

dt['flightsCount']=dt.year

#Call Function to draw Bar Graph

draw\_barplot(\_style='whitegrid',\_x='origin',\_y='flightsCount',\_xLabels='Airport',

\_yLabels='Number of Flights', \_dataset=dt,

\_suptitle="Number of Flights from different Airports"

)

sns.heatmap(dataset.corr())

plot\_CountPlot('origin',dataset,'Number of Flights from different Airports')

dataset = dataset[["year","month","day","dep\_time","sched\_dep\_time","dep\_delay","arr\_time","sched\_arr\_time","arr\_delay","flight","air\_time","distance","hour","minute"]]

dataset.isnull().sum()

dataset= dataset.fillna({'arr\_delay' : 1})

dataset= dataset.fillna({'dep\_delay' : 0})

dataset.iloc[160:173]

dataset.head(5)

import numpy as np

import sklearn.metrics as metrics

import matplotlib.pyplot as plt

dataset= pd.get\_dummies(dataset, columns = ['flight','distance'])

dataset.head()

x= dataset.iloc[:, 0:8].values

y= dataset.iloc[:, 8:9].values

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_test,y\_train = train\_test\_split(x,y,test\_size = 0.2, random\_state =0)

x\_test.shape

x\_train.shape

y\_test.shape

y\_trainshape

**GitHub & Project Demo Link**:

https://github.com/IBM-EPBL/IBM-Project-48721-1660811989