**Flight Delay Predictions using Supervised Machine Learning**

***Adhithyaa Vishal .T and Vishaal.S, Prince Shri Venkateshwara Padmavathy Engineering College, Ponmar.***

Abstract­ **—**The primary goal of this project is to predict airline

delays caused by various factors. Flight delays lead to negative

impacts, mainly economical for commuters, airline industries

and airport authorities. Furthermore, in the domain of sustainability,it can even cause environmental harm by the rise in fuel consumptionand gas emissions. The growth of the aviation sector has made flightdelays more common across the world.They cause inconvenience tothe travelers and incur monetary losses to the airlines.We analyzed the various factors responsible for flight delays and applied machine learning models such as Random Forest, XGBoost, KNN, Decision

Tree to predictwhether a given flight would be delayed or not.Also with certain features we can predict how far the delay is going to be using some regression techniques like Random Forest Regression and Decision Tree Regression.

I. INTRODUCTION

During the most defining period of human history, where

computing has moved from mainframes to PCs to cloud, and now to artificial intelligence. A fundamental sub-area of artificial intelligence has come into notice, called as

Machine Learning, which enables computers to get into a

mode of self-learning without being explicitly programmed.

With the concept of machine learning, we have been able to

apply complex mathematical computations to big data

iteratively and automatically, that too with efficient speed,

this phenomenon has been encompassing momentum over

the last several years. On the other hand, data mining

involves data discovery and sorting it among large data sets

available to identify the required patterns and establish

relationships with the aim of solving problems through data

analysis. Simply combining, machine learning and data

mining use the same type of approach and set of algorithms,

except the kind of data pre-processing and end prediction

varies. BY combining these two core areas to predict and

present the most accurate results possible.

*A. Supervised Machine Learning*

It is a machine learning task where dataset inputs and

outputs are clearly recognized and already given, then

several type of algorithms are trained using labeled

examples. A supervised learning algorithm contains an entire dataset, which is further divided into training and test data; the algorithm examines the training dataset and produces an inferred function, which is then used for mapping new examples. In case of the aviation industry, commercialized aviation is a type of transportation system that is complexly distributed. It tends to deal with several important resources, demand fluctuations, and various other kinds of stages. Stages are bound to take place at terminal boundaries, runways, airports, and distinguished airspaces that may be susceptible to different kind of delays or cancellations. Summing up, some set of examples include weather conditions, ground delays, air traffic control and several other constraints and unforeseen circumstances that lead to delays and cancellations in the entire aviation industry. Hence, this becomes an optimal scenario which will allow us to implement a supervised machine learning algorithm to precisely determine and predict the class labels for unrevealed instances. Supervised Learning algorithm here will model relationships and dependencies between the aimed prediction output and the input features, such that I’ll be predicting the output values for new data based on the relationships which are learned from the previous data set. Supervised Learning problems can be further categorized into following problems.

• ***Classification*** – It is a type problem in which the output

variable is an entire category itself, such as “Win” or

“Lose”, the entire input data is classified into the category

variables; it is generally used largely for recommendation

problems.

*•* ***Regression*** *–* It is a type of problem is which the output

variable is a real value, such as few raw data values related

to something. This is the problem type massively used for

prediction analysis, and hence will be used in this project.

*B. Classification Methods*

**Random forest Classification**

It consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.

**Decision Tree Classification**

It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and

 leaf nodes.

**K Nearest Neighbours Classification**

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN is used in statistical estimation and pattern recognition.

**XGBoost Classification**

Gradient boosting is a method where the new models are created that computes the error in the previous model and then leftover is added to make the final prediction. It uses a gradient descent algorithm that is the reason it is called a **“Gradient Boosting Algorithm”**

C. Regression Methods

**Random Tree Regression**

**It uses ensemble learning** method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

**Decision Tree Regression**

A Decision Tree imposes a series of questions to the data, each question narrowing possible values, until the model is trained well to make predictions. It can make a prediction by running through the entire tree, asking true/false questions, until it reaches a leaf node. The final prediction is given by the average of the value of the dependent variable in that leaf node.

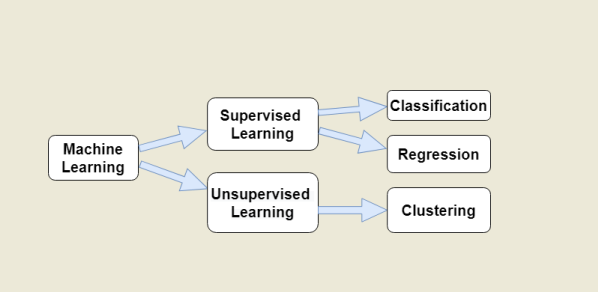


Fig.1: Overview and classification of Machine Learning

II. RELATED WORKS

Flight Delays has become a common and complex

phenomenon, it occurs due to the problems at the originairport,at the destination-airport, any ground reasons or a combination of these entire factors can also give rise to

delays. Delays are also being regarded as caused due to specific airlines. Even if it is complex, it is still measurable

with decent accuracy And with respect to the schedule and

on-time performance of airlines, their generally exists some

pattern of flight delay (Wu, 2005)[4]. The results obtained

from this project, Airline Delay Predictions using Supervised Machine Learning, it can help to better understand the phenomenon and up to a very large extent.

In 2013, it was estimated that approx. 36% of flights were

delayed by more than five minutes in Europe, 32% of flights

delayed by more than 15 minutes in the US, and 16% of

flights were cancelled or sobered delays greater than 30-40

minutes in Brazil[1]. Hence, it indicates how important this

indicator is and how it acts no matter how wide the scale of

airline meshes exists.

Furthermore, coming to the Indian scenario, in 2017,

according to the reports by the Directorate General of Civil

Aviation (DGCA), between January and April, close to 5.12

lakh domestic passengers in India faced issues due to airline

companies denying boarding, as well as flight cancellations

and delays [2]. Airline companies had to pay the passengers

compensations of over Rs. 25 crore for various

inconveniences during the first four months of this year.

Hence, the prediction analysis retrieved from this project can

contribute in the form of a prototype in helping to identify

operational variables that contribute to delays in any country

scenario.

(Allan et al., 2001)[3]analysed delays at NYC Airports

from September `96 through August`00, with the aim of

finding out some major causes of delay occurred during the

first year of an Integrated Terminal Weather System (or

ITWS) use and delays occurred with ITWS in operation that

were “avoidable” if in case weather conditions would have

been improved. The methodology used in the study has

considered some major causes of delays (for example,

convective weather inside and outside the terminal area, and

high winds), and these causes were generally neglected in

previous studies of capacity constrained airports such as

Newark International Airport (EWR). The research

concluded that the usual methods of assessing delays only in

terms of Instrument Meteorological Conditions (IMC)

,Visual Meteorological Conditions (VMC) and the

respective airport capacities is way more simplified than

required for determining the type of air traffic management

investments that in the best ways reduces the possible

“avoidable” delays.

(Hansen and Hsiao, 2005)[5] analysed the rise in flight

delay in the United States domestic system by estimating an

econometric model of average daily delay that combines the

effects of arrival queuing, terminal weather conditions,

seasonal effects, and secular effects (such as a half year).

The results suggested that even after controlling these

factors altogether, the delays decreased gradually from 2000

through mid-2003, but the trend reversed drastically

thereafter.

(Rosen, 2002)[6] measured the rate of change in flight

timings that resulted due to infrastructure-constant changes

in passenger demand. Results indicated that as the ratio of

demand to fix infrastructure increased, the delays increased

proportionately, which resulted in proper decrease in

average flight times by approx. 7 minutes after the rapid

decrease in the fall’01. The flight time differences between

the airlines in the data sample were small, though the United

Airlines had lesser average flight times in the winter quarter

than America West, which is considered even smaller

airline.

Over the past couple of years, various analytical models and simulation methods have been used to analyze flight delay, including deterministic queuing models, neural

networks, econometric models etc. Although it is evident

that the analysis on delays carried is either on macroscopic

or microscopic data over a period of couple of days and this

has happened because of the huge data of flights every day.

Hence, the predictions led to less accurate results or relapse

in the trend among the results. So here, obtaining the airline

on-time performance data set from the U.S. DOT Bureau of

Transportation Statistics (BTS) website, and the linear and

polynomial regression models to be used along with

regularization technique in machine learning is far better to

identify the delay pattern. In this project, studies on airport

delay and individual airlines delay behavior analysis are

carried out, using linear regression model, polynomial

regression models, and regularization. The performances of

the models are tested using various metrics, e.g., CV

Method, MSE/RMSE Scores, etc. This project will be able

to complete several objectives like the statistical description

of airlines, temporal variability of delays, the relation of

delays with the origin airports, estimating geographically the

flights from each airport, etc., along with the main prediction analysis.

III. REGRESSION ANALYSIS MODELLING

*A. Overview of the Dataset*

The dataset has been taken from a reliable online available

government agency website that provides the air traffic

delay statistics in the United States. The U.S. Department of

Transportation's (DOT) Bureau of Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers. BTS compiles daily data for the benefit of the customers or for any data analysts.

TABLE I

DESCRIPTION OF THE ATTRIBUTES INVOLVED IN THE DATASET

|  |  |
| --- | --- |
| Attributes | Descriptions of Attributes |
| YEAR, MONTH, DAY,  DAY\_OF\_WEEK  AIRLINES | dates of the flight  It is the IATA Code to  identify unique airlines |
|  |  |
| ORIGIN\_AIRPORT and  DESTINATION\_AIRPORT  SCHEDULED\_DEPARTUR  E and  SCHEDULED\_ARRIVAL | Code attributed by IATA to  identify the airports  scheduled times of take-off  and landing |
|  |  |
| DEPARTURE\_TIME and  ARRIVAL\_TIME  DEPARTURE\_DELAY and  ARRIVAL\_DELAY  DISTANCE | real times at which take-off and landing took place  difference (in minutes)  between planned and real times |
|  |  |

|  |  |  |
| --- | --- | --- |
|  | IATA\_CODE | AIRLINE |
| 0 | UA | United Airlines Inc. |
| 1 | AA | American Airlines Inc. |
| 2 | US | US Airways Inc. |
| 3 | F9 | Frontier Airlines Inc. |
| 4 | B6 | JetBlue Airways Inc. |
| 5 | OO | Skywest Airlines Inc. |
| 6 | AS | Alaskha Airlines Inc. |
| 7 | NK | Spirit Airlines |
| 8 | WN | Southwest Airlines Co |
| 9 | DL | Delta Airlines Inc. |
| 10 | EV | Altantic Southeast Airlines. |
| 11 | HA | Hawaiian Airlines Inc. |
| 12 | MQ | American Eagle Airlines Inc. |
| 13 | VX | Virgin America |

Fig.2: All the airlines in the dataset associated with particular IATA carrier codes.

*B. Data Exploration and Visualization*

Data cleaning is the critical initial step in evaluating the

dataset for final analysis. With the enormous amount of data

available, databases are prone to have noisy, missing and

inconsistent data. The data in this project is obtained from

BTS source, which has varying kinds of 31 variables

involved, and may not be compatible with the format in

which we require the data to use in Python. Data Cleaning

helps in removing noisy data, and removing inconsistencies.

**Data Visualization**

It is a way to graphically represent the data especially when the data is numeral.It allows us to recognize new patterns and relations with the data.Here in this project we use bar chart , pie chart and scatter plot for identifying trends and clusters.A Heatmap is visually represented to find the correlationship between different attributes.

# 

Fig.4:Dropping unused attributes and values

The dataset contains small percentage of missing values for certain columns and these values are dropped as they make up a very small portion of the dataset.

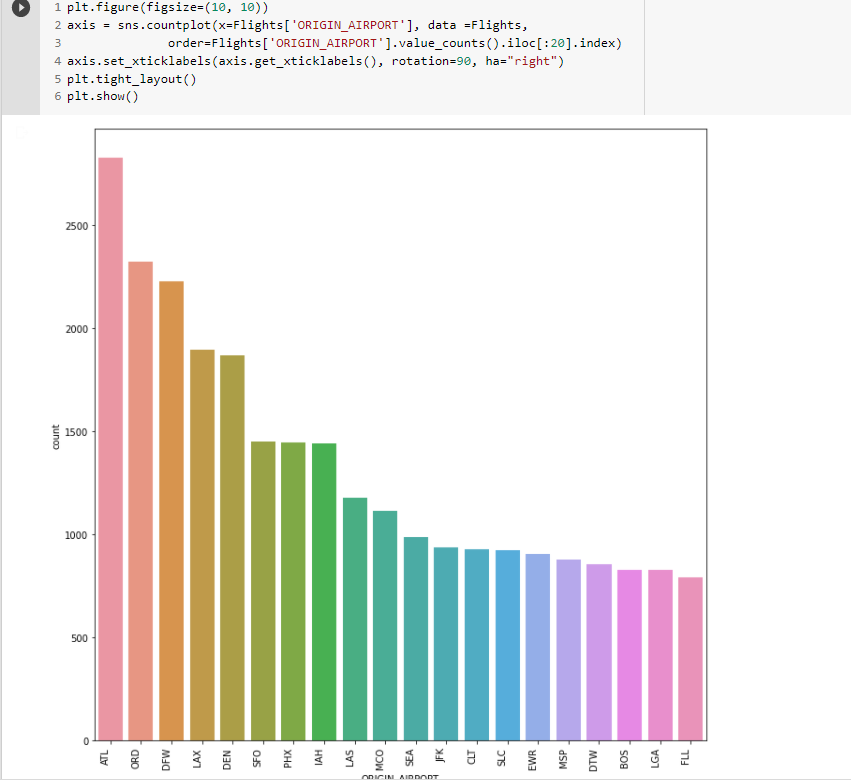
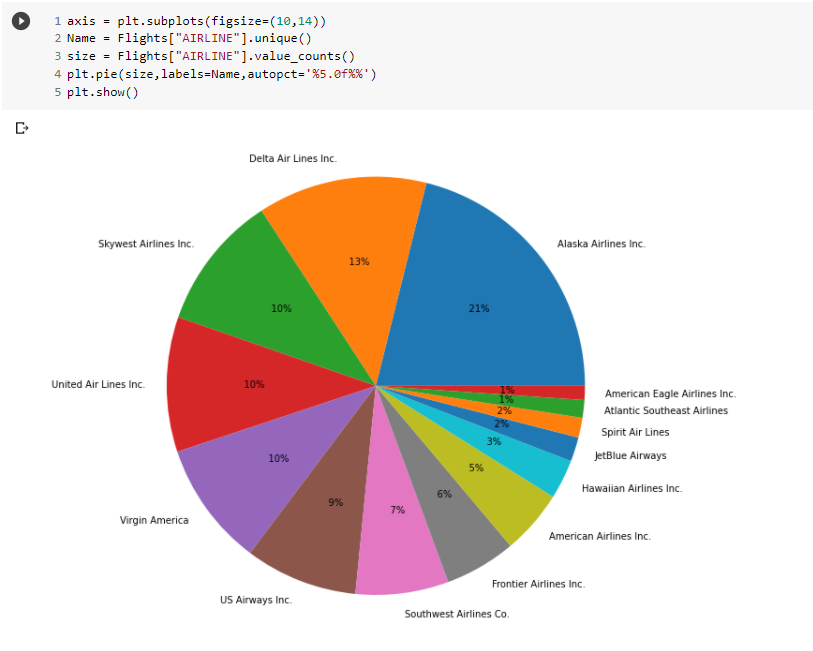


Fig.3:A plot between Origin City and the Number of flights

 Fig.5:Pie chart with % of flights per company

The above diagram depicts the percentage of flights travelled by each airlines.

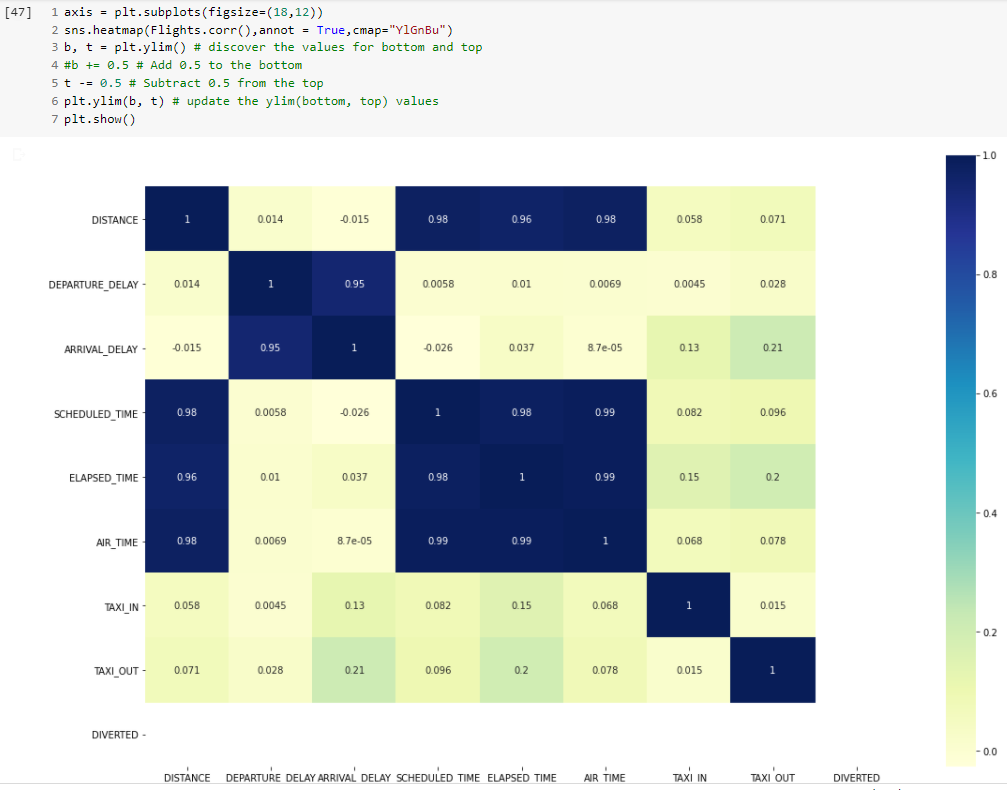


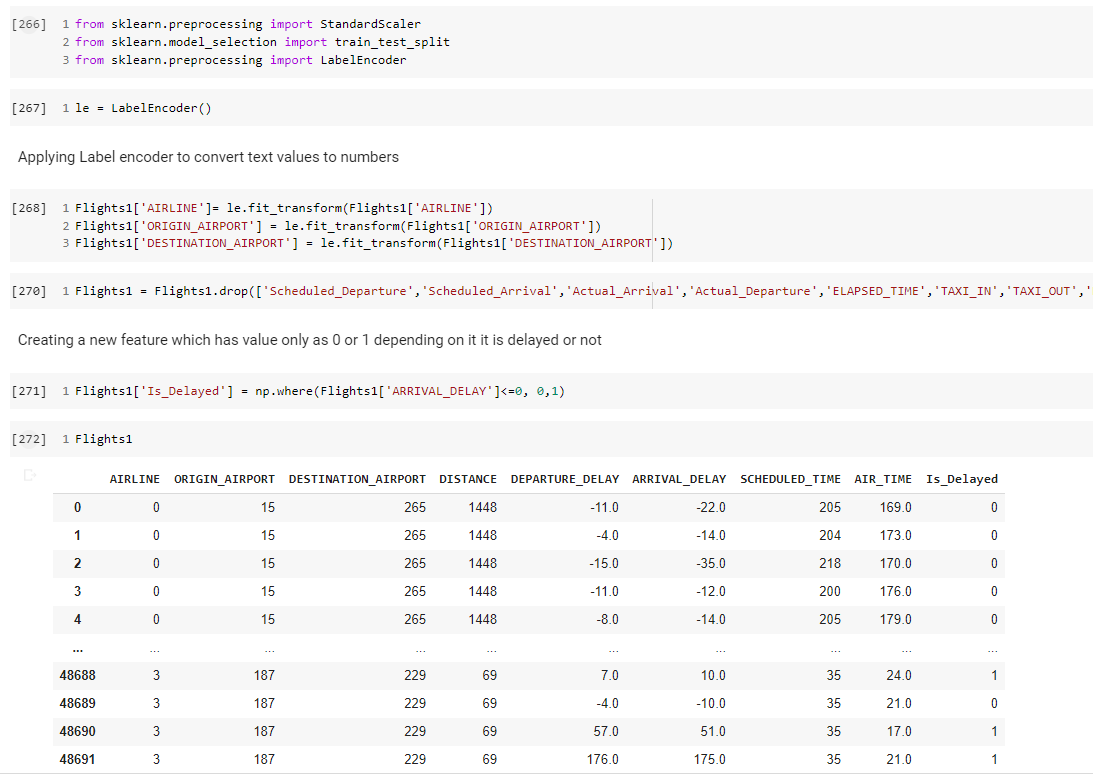
Fig.6:Correlation between variables using Heatmap

*C. Feature Selection*

Not all the columns are not really needed for the prediction of delays, so the unneeded features are dropped and few features necessary were kept for the prediction purposes.Some of the features are in the form of a string these were converted to number values using Label encoder and were assigned number beginning with zero, this is done so that the dataset is more machine learning friendly as models tend to not perform well with strings as features.To Predict the flight delay an extra feature is to be added in the data called ‘Is\_Delayed’ which is actually a binary value indicating

0 = Flight Not delayed

1 = Flight Delayed



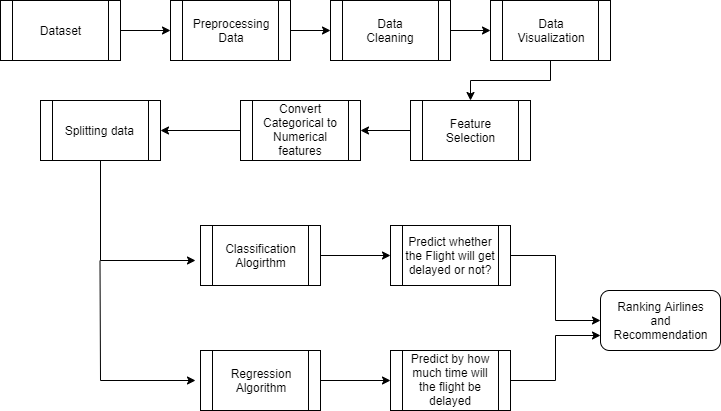


Fig.7: System Architecture Diagram

IV. PERFORMANCE METRIC

***Cross Validation Technique and K-Fold Technique***

Cross Validation is a very important technique for

assessing the performance of machine learning models. It

enables us in knowing how a machine learning model would

generalize to an independent data set. The model dataset is divided into three sets: Training, test,

and validation. The entire set is divided into K-folds or

subsets, which is basically applying the K-fold technique,

one of the ways of Cross Validation. Then, the K-1 folds are

sent for training and the learning is done on it, then the

model’s generalization is checked on the test set, which

contains just the remaining one fold; and this process goes

on till the last fold.

***MSE***

The Mean Squared Error (MSE) is a measure of how

close a fitted line is to the real data points. For every data

point on the line, we take the distance vertically from the

real point to the corresponding Y value on the curve fitted

(which is the error), and square the value. The next step is to

carry out the summation of all the squared error values

corresponding to all the data points, and, in the case of a

linear fit, the value we get is divided by the total number of

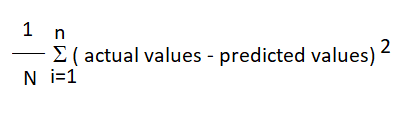
observations minus 2. The squaring is to avoid negative

values cancelling the positive values. The quality of the

model is assessed by the Mean Squared Error score we get,

the smaller the value, the closer the fit is to the real data and

the accurate the machine learning model.

**MSE FORMULA**

Where n=Total no. of attributes or points taken into account.

RMSE

Root Mean Squared Error (RMSE) is another quality that

we calculate to measure the accuracy of a model. It is equal

to the square root of the mean square error. It is considered

as one of the most easily interpreted statistics, as it has the

same units as the quantity plotted on the ordinate, which is

the y-axis.

RMSE = sqrt( sum( (predicted\_i - actual\_i)^2 ) / total predictions)

V. CONCLUSION

This project and the analysis retrieved are useful not only

for passengers point of view, but for every decision maker in

the aviation industry. Apart from the financial losses

incurred by the industry, flight delay also portray a negative

reputation of the airlines, and decreases their reliability. It

causes various sustainability issues, for example, increase in

fuel consumption and gas emissions. The analysis carried

here not only predicts delays based on the previous available

data, but also give statistical description of airlines, their

rankings based on their on-time performance, and delays

with respect to time, showing the peak hours of delay. This

project can be used as a prototype by any aviation authority

for their benefit, in the Indian Scenario too, it can work as an

efficient model or a proper prototype to study delay analysis,

based on the real dataset provided. This project has

encompassed and showed the importance of Regression

Analysis in Machine Learning, Data Mining concepts for

efficient data cleaning, Cross Validation technique and

Regularization in ML for making proper models and its

predictive analysis.

REFERENCES

[1] ANAC. AgˆenciaNacional de Aviac¸ ˜ao Civil. Technical report,

http://www.anac.gov.br/ , 2017.

[2] Indian Economic Times https://economictimes.indiatimes.com/

[3] MIT, Lexington, Massachusetts, Allan, S.S., S.G. Gaddy, and J.E.

Evans, (2001) Delay Causality and Reduction at the New York City

Airports Using Terminal Weather Information.

[4] Wu, C. (2005), Inherent delays and operational reliability of airline

schedules, Journal of Air Transport Management Volume 11, Issue

4(273-282)

[5] Hansen, M., and C. Y. Hsiao (2005), Going South? An Econometric

Analysis of US Airline Flight Delays from 2000 to 2004, Presented at

the 84th Annual Meeting of the Transportation Research Board

(TRB), Washington D.C.’05.

[6] Rosen, A. (2002), Flight Delays on US Airlines: The Impact of

Congestion Externalities in Hub and Spoke Networks, Department of

Economics, Stanford University

[7] Programming in Python 3: A Complete Introduction to the Python

Language , By Mark Summerfield.

[8] Prabakaran, N., and R. JagadeeshKannan. "Sustainable life-span of

WSN nodes using participatory devices in pervasive environment."

Microsystem Technologies 23.3 (2017): 651-657.