**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 consumption and gas emissions. Hence, these factors indicate how necessary and relevant it has become to predict the delays no matter the wide-range of airline meshes. To carry out the predictive analysis, which encompasses a range of statistical techniques from supervised machine learning and, data mining, that studies current and historical data to make predictions or just analyze about the future delays, with help of Regression Analysis using regularization technique in Python 3.

**1 INTRODUCTION**

* 1. Over View
  2. Purpose

**2 LITERATURE SURVEY**

2.1 Existing Problem

2.2 Proposed Solution

**3 THEORITICAL ANALYSIS**

3.1 Block Diagram

3.2 Hardware/Software designing

**4 EXPERIMENTAL INVESTIGATIONS**

**5 FLOW CHART**

**6 RESULT**

**7 ADVANTAGES AND DISADVANTAGES**

**8 APPLICATIONS**

**9 CONCLUSION**

**10 FUTURE SCOPE**

**11 BIBILOGRAPHY**

**12 APPENDIX**

**CHAPTER-1 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.

**1.1 OVER VIEW**

MACHINE LEARNING consists of

Supervised Learning and Unsupervised Learning

***A. Supervised Machine Learning:***

It is a machine learning task where the 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. Regression Analysis Methods*

The main focus of regression analysis is to model and

determine the expected value of a dependent variable *y* in

terms of the value of one or more independent variables (x)*.*

• ***Linear Regression***

Linear Regression is used to model and establish a

relationship between dependent and independent set of

variables by fitting the best line possible. The best fit line hence formed as the result of prediction carried out is known

as our regression line and is represented by a linear equation

1. *:y = b0+b1x1* (1)

In case of logistic regression, which is very much

compared to linear regression, the outcome (dependent

variable) has only limited number of discrete possible

values. Whereas, linear regression analysis is the first bestsuited

method because it results in any one among the range

of an infinite number of possible values.

• ***Polynomial Regression***

In practice, rather than performing a simple linear

regression, we can improve the model doing a fit with a

polynomial of order N, because, in many situations, such a

linear regression model may not hold true, or even if it does,

the accuracy is decreased. Doing so, it is necessary to define

the degree N which is optimal to represent the data. Hence,

here it is where polynomial regression analysis becomes the

next best-suited method for the prediction analysis.

Represented by equation (2):

*y=b0+b1x1+b2x1*

*2………..+bnx1*

*n* (2)

• ***Multiple Linear Regression***

If set of variables have a linear relationship with the

dependent variable, then the regression is known as multiple

linear regression. A multiple regression is represented by the

following equation(3):

*y=b0+b1x1+b2x2…..+bnxn*(3)

In all three equations above,(1),(2) and(3), *b0,b1,…bn* are

the coefficients of the equation whose values we need to

determine in any model; the *x1,.. xn* are the dependent

variables involved; and *y* is the independent variable here.

Multiple Linear Regressions is an even wider class of

regression that combines linear and nonlinear regressions

with multiple explanatory variables. In this case, because of

the broad range of prediction possibilities it offers, using

multiple regression in some of the models, which attempts to

explain dependent variable using more than one independent

**1.2 PURPOSE**

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.

**CHAPTER 2 - LITERATURE SURVEY**

**2.1 EXISTING PROBLEM**

OVER the last twenty years, air travel has been increasingly preferred among travellers, mainly because of its speed and in some cases comfort. This has led to phenomenal growth in the air traffic and on the ground. 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. Moreover, the economic impact of flight delays for domestic flights in the US is estimated to be more than $19 Billion per year to the airlines and over $41 Billion per year to the national economy. In response to growing concerns of fuel emissions and their negative impact on health, there is an active research in the aviation industry for finding techniques to predict flight delays accurately in order to optimize flight operations and minimize delays.

**2.2 PROPOSED SOLUTION**

Using machine learning model, we can predict the 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 for various figures of merit.Finally, it will be integrated to web based application.

**CHAPTER-3 THEORITICAL ANALYSIS**

**3.1 BLOCK DIAGRAM**

A close up of a map

Description automatically generated

**3.2 HARDWARE / SOFTWARE SOLUTION**

**Strategy :** Using Application Of Machine Learning Algorithms To Predict Flight Delays

**Dataset Creation :** Data Collection

**Dataset Preprocessing :**

* Importing general Dataset from Library
* Configuring Dataset to Generator class

**Training and Testing Dataset :** By Applying Dataset Generator Class Functionality

**Model Building :**

* Import Model Building Libraries
* Initialising the Model
* Loading Preprocessing Data
* Adding Machine Learning Algorithms
* Configuring Learning Process
* Train and Test the Model
* Optimize and save the Model

**Application Building :**

* Create HTML file
* Build Python Code

**CHAPTER -5 FLOW CHART**

A screenshot of a cell phone

Description automatically generated

**CHAPTER – 6 RESULT**

Based on the input given by the user,the model predicts risk and also displays whether the flight is delayed or not.

**CHAPTER -7 ADVANTAGES AND DISADVANTAGES**

**Advantages :**

* Perfect prediction of Risk along with the reason for Delay
* Very Accurate Results
* Extremely Easy Interface
* Straight Forward Results

**Disadvantages :**

* User Should have the idea on inputs

**CHAPTER – 8 APPLICATION**

***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 model.

***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.

**CHAPTER – 9 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.

**CHAPTER – 10 FUTURE SCOPE**

The model gives very appropriate prediction accuracy, extra variables can be considered to increase a predictive model. For example, Weather information can be extracted and used to better enhance a predictive model for flight delay. The future scope of this study entails a range of techniques that can be used to analyze the data. Principal element analysis or transformation can be accomplished to uncover hidden relations between variables. In addition, considering the statistics is not exactly linear, artificial neural networks or Support vector machines can be used to analyze the impact of a number of variables on flight delay.

**CHAPTER -11 BIBILOGRAPHY AND APPENDIX**

**Model Building**

* Dataset
* Notebook

**Application Building**

* HTML file
* CSS file
* Flask