**FLIGHT DELAY PREDICTION FOR AVIATION INDUTRY USING MACHINE**

**LEARNING**

**SUBMITTED BY**

L.JEEVA (20445ER003)

S. PRAVEENA (20445ER005)

P. SANGEETHA (20445ER006)

S. SNEHA (20445ER008)

P. SAHAYARANI (19445ER027)

**TEAM ID: NM2023TMID21296**

**UNDER THE GUIDENCE OF**

MRS.EZILARASI B.E., MCA., M.PHIL., M.TECH.,

PROFFESSOR

**DEPARTMENT OF COMPUTER APPLICATION**

**GOVERNMENT ARTS AND SCIENCE COLLEGE FOR WOMEN**

**KODAIKANAL-624101**

**APRIL-2023**

TABLE OFCONTENT

**INTRODUCTION**

Overview

A brief description about your project

Purpose

The use of this project. What can be achieved using this.

**PROBLEM DEFINING & DESIGN THINGING**

Empathy Map

Paste the empathy map screenshot

Brainstrom & Ideation map

Paste the brainstorm & ideation map screenshot

**BASIC CONCEPT**

Flight data

Weather data

**DATA COLLACTION**

**MODELING AND ANALYSIS**

Random Forest Classifier

**ADVANTAGES & DISADVANTAGES**

List of advantages and disadvantages of the proposed solution

**APPLICATIONS**

The areas where this solution can be applied

**RESULT**

Final findings (Output) of the project along with screenshots.

**CONCLUSION**

Conclusion summarizing the entire work and findings.

**FUTURE SCOPE**

Enhancements that can be made in the future.

**REFERENCES**

**APPENDIX**

Source code

Attach the code for the solution build

**Introduction**

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 endusers 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 the most reason 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 delay[2].Airline

corporations had to pay the passengers compensations of over Rs. twenty five crore

for Varied inconveniences throughtout 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.

Research Motivation

Average aircraft delay is regularly referred to as an indication of airport capacity.

Flight delay is a prevailing problem in this world. It's very tough to explain the

reason for a delay. A few factors responsible for the flight delays like runway

construction to excessive traffic are rare, but bad weather seems to be a common

cause. Some flights are delayed because of the reactionary delays, due to the late

arrival of the previous flight. It hurts airports, airlines, and affects a company's

marketing strategies as companies rely on customer loyalty to support their

frequent flying programs.

Problem statements:

My case study was about LaGuardia Airport in New York, Logan International

Airport in Boston, San Francisco International Airport in San Francisco, and

O’Hare International Airport in Chicago, which are four major airports in the

United States of America. But we focused the idea and research on LaGuardia

International Airport. Compared with the data produced by all airports in USA, the

data which we gathered was very limited, but it gave us a great direction on how

weather plays a part in flight delays. In this project, the goal is to use exploratory

analysis and to build machine learning models to predict airline departure and

arrival delays.

Report structure:

This master project report is organized into nine chapters. The preface of the

project, research motivation, and problem statement form chapter 1. Chapter 2

describes the basic concepts of flight and weather data. Chapter 3 focuses on

structures of the project. Chapter 4 and 5 explain the data collection and data

exploration part of the flight data, while the chapter 6 focuses on predictive

modelling implemented on the flight data. Chapter 7 focuses on predictive

modelling implemented on the weather data. Chapter 8 starts with the introduction

of the Twitter data and some tweets exploration that helped me in the course of

building the project. It focuses on predictive modeling of Twitter data using

Random Forest and Support VectorMachine. concludes the paper and finally

chapter 10 talks about the future scope of the project. 1.4 Related Work The main

concern of the researchers and analysts is to predict the reasons for flight delays

and for that they have put in their efforts on collecting data about flight and the

weather. Mohamed et al. [2] have studied the pattern of arrival delay for non-stop

domestic flights at the Orlando International Airport. They focused primarily on

the cyclic variations that happen in the air travel demand and the weather at that

particular airport. In Shervin et al.’s work [3], their motive of research is to

propose an approach that improves the operational performance without hampering

or effecting the planned cost. Adrian et al. [4] have created a data mining model

which enables the flight delays by observing the weather conditions. They have

used WEKA and R to build their models by selecting different classifiers and

choosing the one with the best results. They have used different machine learning

techniques like Naïve Bayes and Linear Discriminant Analysis classifier Choi et al.

[5] have focused on overcoming the effects of the data imbalancing caused during

data training. They have used techniques like Decision Trees, AdaBoost, and K-

Nearest Neighbors for predicting individual flight delays. A binary classification

was performed by the model to predict the scheduled flight delay. Schaefer et al.

[6] have made Detailed Policy Assessment Tool (DPAT) that is used to stimulate

the minor changes in the flight delay caused by the weather changes. Bing Liu [7]

has done a sentiment analysis and opinion mining that analyzes people’s opinions,

sentiments, and studies their behavior. The output of the research is a feature-based

opinion summary which is also known as sentiment classification. Using

techniques such as Natural Language Processing, Naïve Bayes, and Support Vector

Machine, researchers built algorithms for analysis that helped them in extracting

features in the model. Most of them focused on predicting overall flight delays.

Our research concentrated mainly on predicting flight delays for a particular airport

over a specific period of time. First, we used a regression model to examine the

significance of each feature and then, a feature selection approach to examine the

impact of feature combination. These two techniques determined the features to

retain in the model. Instead of using the whole set, we sampled 5,000 records at a

time to run through different machine learning models. The machine learning

models implemented here were Random Forest classifier and Support Vector

Machine (SVM) classifier. Further, we applied an approach 5 called One-Hot-

Encoder to create a variant of the model for evaluating potential prediction

performance.

OVERVIEW:

Flights delay 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 Neighbour

guassian naïve bayes, decision tree, support vector machine ,random forest, and

gradient boosted tree) were trained and tested to complete the binary classification

of flight delay . the evaluation of algorithms as fulfilled by comparing the values of

four measures : accuracy, precision, recall, and f1- score .these measues wee

weigthed to adjust the imbalance of the selected data set. The comparative analysis

showed that the Decision Tree algorithm has the best performance with an

acuuracy 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.

PURPOSE:

Predicting flight delays can improve airline operations and passenger

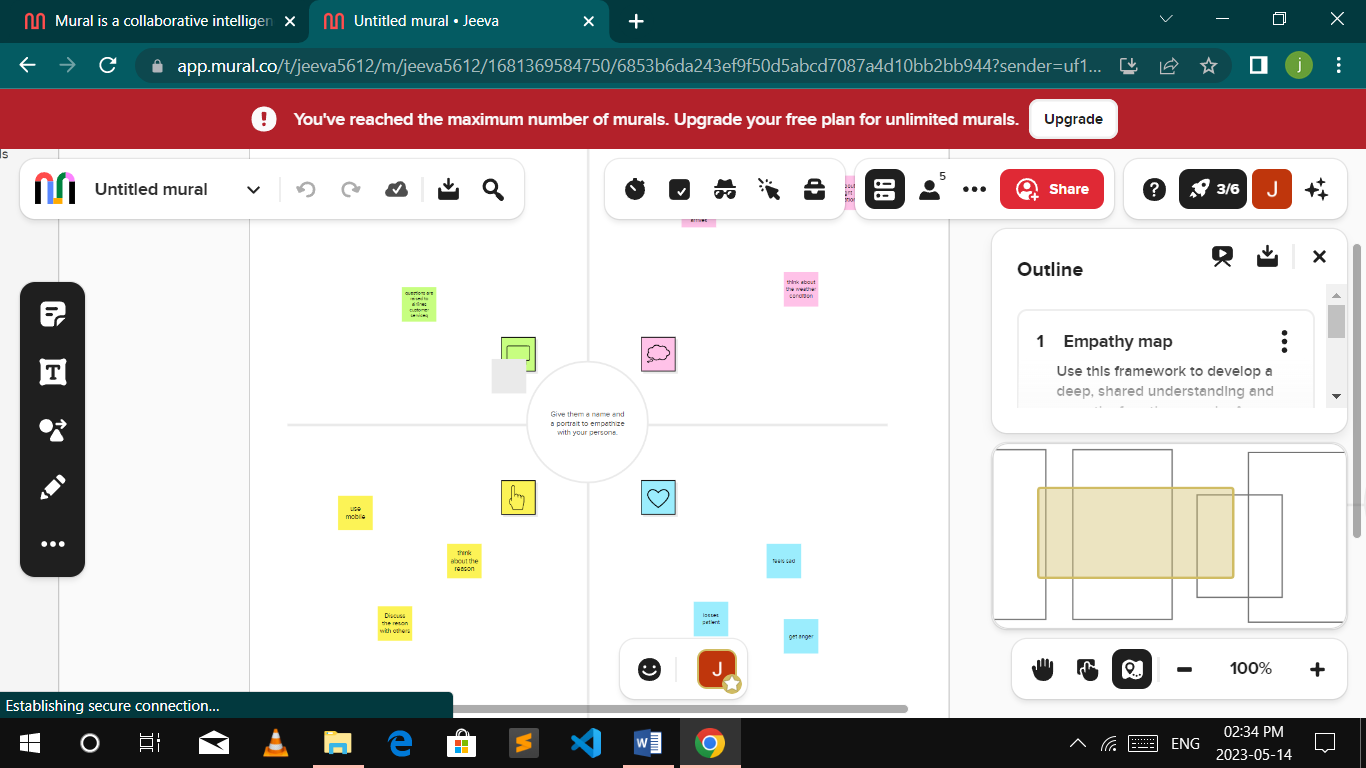
satistifaction, which will result in a positive import on the economy.

In this study, the main goal is to compare the performance of learning

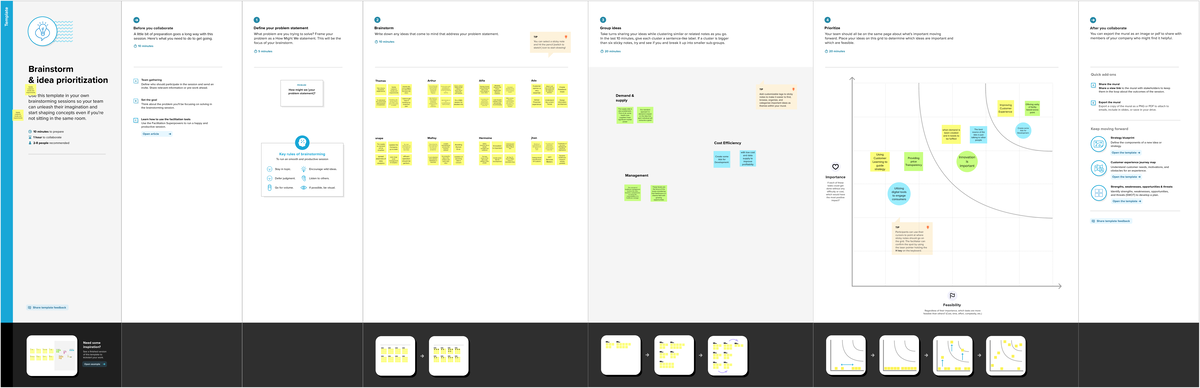
classification algorithm when predicting flight delays.

**Problem Definition & Design Thinking**

Empathy Map



Brainstroming & Ideation mapping



BASIC CONCEPT

We will be looking into some of the variables that play a vital role in determining

the flight delay. Each variable has a different weight that is used while training the

model. Prior knowledge of these variables should be present to understand the

dataset.

Flight Data

The flight data is from the year 2007 and 2008, which is taken from Bureau of

Transportation Statics. Every flight has many variables, which give detailed

information about the specific flight.

1. Flight Number

2. Carrier {American, United}

3. Destination {Airport Code: SFO}

4. Origin {Airport Code: LAX}

5. Date {MM/DD/YY}

6. Day of Week {Mon, Tue, Wed, Thu, Fri, Sat, Sun}

7. Scheduled Departure Time {HH:MM AM/PM}

8. Actual Arrival Time {HH:MM AM/PM}

9. Actual Departure Time {HH:MM AM/PM}

10. Minutes Late {+Late/-Early} 11. Scheduled Arrival Time {HH:MM AM/PM}

The 2008 on-time performance data contains 7 million records. The average

number of daily flights is 19,178, with 24 data elements included in the flight

database.

Weather Data

The Aviation System Performance Metrics (ASPM) [9] data from the FAA

operations performance data website contains airport specific weather, flight

demand, and airport capacity information. The data used in this project is from

2008. It provides hourly weather and operational data for 77 US airports. A list of

data fields is presented below:

1. Airport ID

2. Year

3. Adjusted Month

4. Adjusted Day

5. Adjusted Hour

6. Time Zone

7. Visibility

8. Dry Bulb Farenheit

9. Dry Bulb Celsius

10. Dew Point Farenheit

11. Dew Point Celsius

12. Relative Humidity

13. Wind Speed

14. Altimeter

DATA COLLECTION

Once the project undertaking is completely comprehended, our subsequent stage is

to gather the information that is required for future model building. The

information accumulation was an issue as data was not situated at a single source.

The data was kept in unique information design. To achieve the end goal, it

requires a clear understanding of the correct location of the data. As we can see in

the US Bureau of Transportation Statistics gives detailed information on every

single household flight, which incorporates their booking and take off

circumstances and real takeoff, origin, destination, date, and carrier. We

consolidated a portion of the information properties with Local Climatological

Data from National Oceanic and Atmospheric Administration (NOAA) to shape a

join data set. Since the datasets for every year are very massive, we decrease our

concentration to one-year, i.e., 2008, which as of now contains 1 million records

for the most significant airplane terminals. In this venture, we have taken 2007 as

our preparation set and 2008 as our test set. Handling speed is a noteworthy

thought since the machine learning methodology that functions admirably on

smaller datasets cause issues with the Jupyter Notebook establishments on our

PCs.

DATA COLLECTIONS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Number of operations** | **% of Total operations** | **Delayed minutes** | **% of Total Delayed minutes** |
| **On time** | 5,473,439 | 73.42% | N/A | N/A |
| **Air carrier delay** | 520,597 | 6.98% | 28,827,070 | 28.55% |
| **Weather delay** | 72,307 | 0.97% | 5,745,298 | 5.69% |
| **National Aviation System Delay** | 598,258 | 8.02% | 28,209,475 | 27.94% |
| **Security Delay** | 4,939 | 0.07% | 176,946 | 0.18% |
| **Aircraft Arriving Late** | 607,928 | 8.15% | 38,006,943 | 37.64% |
| **Cancelled** | 160,809 | 2.16% | N/A | N/A |
| **Diverted** | 17,182 | 0.23% | N/A | N/A |
| **Total Operations** | 7,455,458 | 100.00% | 100,965,732 | 100.00% |

MODELING AND ANALYSIS

Random Forest

Random Forest Classifier: Random Forest is a classifier algorithm that includes a

number of decision trees on several subsets of the given dataset and considers the

average to enhance the predictive precision of that dataset. Rather than relying on

one decision tree, the random forest takes the prediction from each tree and based

on the bulk votes of predictions, and it predicts the final output.



Decision Tree Classifier:

Decision Tree is a Supervised Machine Learning Algorithm that utilizes a set of

rules to build decisions, comparably to how humans make decisions. One way to

think of a Machine Learning classification algorithm is that it is constructed to

make decisions.

**ADVANTAGES:**

• The Random Forest approach outperforms the SVM model in terms of

performance.

• Some way or another the SVM model is exceptionally tedious and doesn't

be guaranteed to create improved results.

• In the end, 91% of non-delayed flights are correctly predicted by our model.

**DISADVANTAGES:**

• The SVM model takes a lot of time, is inaccurate,

• The worst part is cleaning and formatting the data.

•Finding an accuracy of flight delay is less

•It does not have required parameters for finding delay.

**APPLICATIONS:**

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

There are no federal laws requiring airlines to provide passengers with money or other compensation when their flights are delayed. Each airline has its own policies about what it will do for delayed passengers. If your flight is experiencing a long delay, ask airline staff if they will pay for meals or a hotel room.

A flight delay is when an airline flight takes off and/or lands later than its scheduled time. The United States Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes later than its scheduled time.

**FUTURE SCOPE:**

In order to obtain the most precise result, additional supporting research is

required to correlate the entire problem, scope, and approach. Even though the

most common cause of a flight delay is bad weather, major natural or man-made

disasters and other unprecedented events can also delay flights significantly.

**RESULT:**

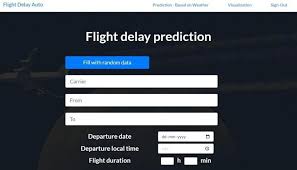
After comparing multiple algorithms we have found that random forest was the

best of them and by using random forest algorithm we have implemented a user

interface which firstly consists of login page and by signing up or logging in we

can enter the page where entering different attribute details gives us the delay

prediction in minutes.



**CONCLUSION:**

The methodology for developing a system to anticipate flight delays is

presented in this project. The paper provides specifics about the various

methodologies that are utilized or can be utilized to determine the cause of flight

delays. Flight delays are currently the topic of conversation due to the significant

financial and environmental costs they impose on airlines and passengers. Flight

delays drive up costs because they have a significant impact on operations. They

may result in higher costs for airlines and their customers. Since the outcome is

directly related to the passenger and the airlines, which in turn is related to another

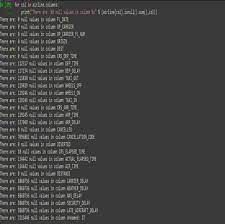
group of airlines and passengers, it is very important to have a real time delay for

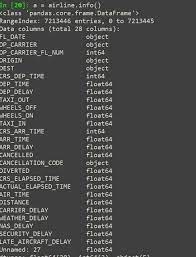
all of the players in the air transportation system. As a result, it is necessary to

develop a method for anticipating flight delays in order to reduce financial losses

and ensure efficient operation.

SOURCE CODE:





REFERENCES:

1. “Airlines Delay” data fromwww.kaggle.com/datasets.

2. U.S. Airlines Delay Analysis- <http://www.milantomin.com>

3. https://developers.amadeus.com/flight-delaypredictionmachine.

4. developers.amadeus.com/flight-delay-prediction machine learning.

5. [www.icao.int/annual-report-2018/Pages/theworld-of-airtransport-in-2018](http://www.icao.int/annual-report-2018/Pages/theworld-of-airtransport-in-2018)

6. machine learning random forest classifier from <https://www.javapoint.com>

7. <https://www.airhelp.com>

8. towardsdatascience.com.