*A project report on*

**PREDICTION IN FLIGHT DELAYS**

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

Computer Science and Engineering

*for*

*CSE3050 – Data Visualization*



*By*

# AAYUSHI (20BCE1791)

# 

# *Presented to*

# Dr. Joshan Athanesious J

# 

# SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

April, 2022

## ABSTRACT

**Flight delays are gradually getting worse, which causes airline businesses to have more trouble making money and to lose customers. Supervised machine learning models were used to forecast flight delays in order to fix this issue. For the prediction, information from flights leaving from JFK airport over the course of a year was gathered in a data collection. To finish the binary classification of flight delays, seven algorithms—Logistic Regression, K-Nearest Neighbor, Gaussian Naive Bayes, Decision Tree, Support Vector Machine, Random Forest, and Gradient Boosted Tree—were trained and evaluated. By contrasting the results of four measurements—accuracy, precision, memory, and f1-score—algorithms were evaluated. In order to correct for the imbalance in the chosen data collection, these measures were weighted. According to the comparative study, the KNN algorithm performs the worst with a f1-score of 0.8039, and the Decision Tree algorithm performs the best with an accuracy of 0.9777. In general, tree-based ensemble classifiers work better than other base classifiers.**

*Keywords—Flight delays, weather, Linear Models, Machine Learning, Random Forest*

Table of Contents

1. INTRODUCTION………………………………………………………………..5
   1. OBJECTIVE …………………………………………………………………6
   2. METHODS USED……………………………………………………………6
   3. [PROJECT STATEMENT](#_heading=h.2et92p0)…………………………………………………….6
   4. SCOPE………………………………………………………………………...6
   5. OUTCOME OF THE PROJECT……………………………………………...6
2. BACKGROUND………………………………………………………………….7
   1. INTRODUCTION…………………………………………………………….7
   2. LITERATURE SURVEY……………………………………………………..7
3. MATERIALS AND METHODS USED…………………………………………16

3.1. ABOUT THE MODELS…………………………………………………….16

3.2. DATASET…………………………………………………………………...18

1. PROPOSED WORKS……………………………………………………………20
   1. NOVELTY……………………………………………………………………20
2. RESULTS AND DISCUSSION………………………………………………….22
3. CONCLUSION…………………………………………………………………...23
4. FUTURE WORKS………………………………………………………………..24
   1. REFERENCES………………………………………………………………….24

Appendix……………………………………………………………………………..26

LIST OF FIGURES

4. PERFORMANCE EVALUATION

Fig.1. Flights dataset………………………………………………………...18

Fig.2. Airlines dataset……………………………………………………….18

Fig.3. Airports dataset……………………………………………………….19

Fig. 4. Number of flights per month…………………………………………26

Fig 5. Number of flights per month in every category………………………27

Fig 6. Total number of flights per day of a week…………………………….28

Fig 7. Number of flights per year…………………………………………….29

Fig 8. Number of flights per day name per week…………………………….30

# 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 is an upscale affair for the airlines, making issues in scheduling and operations for the end users therefore inflicting unhealthy name and client discontent and dissatisfaction. 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.  According to the estimation by the Total Delay Impact Study, the total cost of air transportation delay to air travelers and the airline industry in 2007 was $32.9 billion in the US, resulting in a $4 billion reduction in GDP. Therefore, predicting flight delays can improve airline operations and passenger satisfaction, which will result in a positive impact on the economy. Major problems which cause delay in flights can be delay propagation, delay caused on the departure point or the root of the flight, and cancellation of flights. This problem that causes delay affects Airline, Airport and the enroute airspace which are independent entities which work 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 Representation.

### PROJECT STATEMENT

Our 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. We have used R studio to thoroughly analyses the delay behavior and generate favorable outcomes from them.

* 1. OBJECTIVE

The objective of this project is to design a system that would predict the delays in flights using some machine learning algorithms like Random Forest algorithm and Decision Tree classifier in R language and Python.

* 1. SCOPE OF THE PROJECT

Since R programming is a low-cost measure with server capabilities that have been

proved the scope of our project is we are able to provide self-hosted low cost and low maintenance servers that can be used by the masses. Python is an advanced machine language which contains various predefined libraries ultimately making our work easier and quicker.

.

**BACKGROUND**

We all know that, as airline company customers, we frequently do not learn of flight delays prior to departure, and neither the airline company's ground employees nor the airline itself receives the airline delay prognosis supported by various circumstances. Delay propagation and delay at the departure point are two major issues that might cause delays in flights. Machine learning, probabilistic models, statistical analysis, and network representations are some of the techniques that can be utilized to create a system that forecasts flight delays.

**LITERATURE SURVEY**

This is a collection of articles and research papers on how to predict flight delays because of natural occurrences or operational shortcomings to inflict lesser client dissatisfaction and find approaches to make it better.

*[1]* The research used machine learning to transform the problem into a prediction of the likelihood of aircraft delays. The prediction was made using a binary classification method of supervised machine learning. Four metrics were used to assess the performance of the algorithms, while seven methods were utilized to anticipate delays. The evaluation metrics were weighted to remove the dominating effect of non-delayed flights over delayed flights due to the data set's imbalance. To assess each model's performance, the values of their four metrics were compared after classifiers had been applied to the delay prediction.

*[2]* The individual who wrote this research investigated flight delays, which are now a major topic for air travel around the globe due to the resulting financial losses that the aviation sector is experiencing. Over 20% of US aircraft had delays in 2018, according to data from the US Bureau of Transportation Statistics (BTS). This had a significant economic impact of 41 billion US dollars. These delays are inconvenient for both the airlines and the customers. As a result, journey times lengthen, resulting in higher accommodation and food costs as well as increased stress among travellers. The airlines suffer from increased expenses related to their crews, aircraft repositioning, fuel use while attempting to shorten flight durations, and many other factors that damage their reputation and frequently cause a decline in consumer demand.

*[3]* This study's author looked at flight delays, which are currently a hot topic for air travel worldwide because of the ensuing financial losses the aviation industry is suffering. According to data from the US Bureau of Transportation Statistics, over 20% of US planes experienced delays in 2018. (BTS). A major 41 billion US dollar economic impact resulted from this. Both the airlines and the customers find these delays to be inconvenient. As a result, travel durations grow longer, increasing the expense of lodging and meals as well as the level of stress experienced by passengers. The airlines are burdened by rising costs for their employees, aircraft repositioning, fuel use when seeking to cut travel times, and a host of other issues that harm their reputation and frequently result in a reduction in customer demand.

*[4]* Flight delays are a result of air traffic jams brought on by the expansion of the aviation industries. Flight delays have negative environmental effects in addition to negative economic effects. Controlling air traffic is getting harder and harder. There are many reasons for impede in flights, some of them are due to security issues, mechanical problems, due to weather conditions, airport congestion, etc. we are proposing machine learning algorithms like XG Boost regressed, Linear regression Techniques. Airlines delays make enormous loss for business field as well as in budget loss for a country. The goal of this research project is to predict flight delays because they are a major source of lost revenue for many nations and because this mode of transportation is the fastest and most comfortable. By identifying and reducing flight delays, machine-learning algorithms can significantly reduce them and save significant amounts of money.

*[5]* Light delay prediction is essential to building a more effective airline industry. The complexity of the air transportation system, the variety of forecast techniques, and the abundance of flight data made it difficult to construct precise prediction models for flight delays. This study provides a full assessment of the methodologies utilized to create flight delay prediction models in this setting. Airlines delays cause significant economic harm to a nation's budget as well as to the corporate world. Flight delays are bad for travelers, airports, and airlines. We suggest using machine learning techniques like linear regression. The goal of this research project is to predict flight delays because they are a major source of lost revenue for many nations and because this mode of transportation is the fastest and most comfortable. By identifying and reducing flight delays, machine-learning algorithms can significantly reduce them and save significant amounts of money. Flight delays might always be inconvenient, especially if they lasted so long that it might even be possible to miss the next flight. But if there was a way to foretell if there would be a delay or, even better, how long it may be, then customers might plan ahead and reschedule subsequent flights in advance.

*[6]* Flight delays are bad for travelers, airports, and airlines. All participants in commercial aviation must consider their prediction while making decisions. Additionally, the complexity of the air transportation system, the variety of forecast approaches, and the abundance of flight data made it difficult to construct precise prediction models for flight delays. This study provides an in-depth analysis of the methods used to create aircraft delay prediction models from a data science perspective. According to scope, information, and computational approaches, we suggest a taxonomy and provide a summary of the initiatives utilized to address the flight delay prediction problem, paying special emphasis to a greater use of machine learning techniques. Additionally, we provide a history of major works that shows the connections between issues with flight delay prediction and the developments in research to solve them.

*[7]* In this dissertation, many sorts of variables are used to create machine-learning models that forecast airline on-time performance. In one of the chapters, a methodology is described for enhancing model performance and minority class prediction when using an unbalanced dataset. The other chapter creates, under the guidance of a decision-maker at an airline firm, a machine learning framework to categories whether the flight is on-time or delayed. The proposed framework produces better results than those in the literature when the methodology introduced in the second chapter is used. For domestic flights, another chapter incorporates data from the Federal Aviation Administration covering all US airports from January 2015 to December 2017. There it examines the effects of previously unconsidered weather variables, aircraft variables, airport variables, and scheduling variables on predicting delays. A decision-maker for an airline firm can also get the answers to three questions using the model created in the chapter, including whether a flight is delayed, why it is delayed, and how much longer than expected.

*[8]* Flight delays are a result of air traffic jams brought on by the expansion of the aviation industries. Flight delays have negative environmental effects in addition to negative economic effects. Controlling air traffic is getting harder and harder. There are various reasons for hinder in flights, some of them are related to security issues, mechanical problems, due to weather conditions, airport congestion, etc. Airlines delays make enormous loss for business field as well as in budget loss for a country. Machine learning methods like Random Forest, Decision Tree, MLP Classifier, Naive Bayes, KNN, Gradient Boosting Classifier, Voting Classifier, SVM, Logistic Regression, Ridge Regression, and Neural Network Techniques are what we are recommending. The goal of this research project is to forecast airline delays, which are the biggest economic drag on many nations and the fastest and most comfortable mode of transportation. By applying machine learning algorithms to identify and eliminate flight delays, significant amounts of cash can be saved.

*[9]* In airports, flight delays frequently lead to frustration if they are not effectively managed. The advancement of machine learning models is encouraging scientists and researchers to use current research problems. A robust decision-making process is required to mobilize the air traffic with the least amount of delay due to an increase in customer satisfaction with the air transportation system. 19% of domestic flights in the US arrive at their destination with an average delay of 15 minutes. Furthermore, the availability of precise prediction models is constrained by the complexity of the aviation system. Because delays are stochastic, this study looks into the qualitative prediction of airline delays in order to make the required adjustments and improve customer service. The source for developing prediction models is the collection of operational data during departure and arrival at airports and historical weather data. A decision tree model and a logistic regression model are contrasted in order to determine the delay's status and performance, respectively. The effectiveness of the decision tree algorithm compared to logistic regression is empirically assessed in the suggested study. Based on the results of this simulation, which took into account the time of day, the weather, and other factors, there should be a minimum amount of delays at major airports.

*[10]* Unmanaged flight delays in airports cause frustration. Machine learning methods encourage scientists to use current research problems. Increased consumer satisfaction with the air transportation system requires a sophisticated decision-making mechanism to mobilize air traffic with minimal delay. 19% of US domestic flights arrive 15 minutes late. Aviation's complexity limits the availability of exact prediction models. Because airline delays are stochastic, this study looks into qualitative prediction to improve customer experience. Airport operating data and historical weather data are used to construct prediction models. The delay's status and performance are determined using a decision tree model and a logistic regression model. The recommended study compares decision tree algorithm to logistic regression. Based on this simulation, which considered time of day, weather, and other factors, major airports should have little delays.

*[11]* Air traffic congestion caused by aviation industry growth delay flights. Flight delays have economic and environmental impacts. Air traffic control is difficult. Security issues, technical problems, weather, airport congestion, etc. might cause aircraft delays. Airline delays hurt companies and a country's finances. Random Forest, Decision Tree, MLP Classifier, Naive Bayes, KNN, Gradient Boosting Classifier, Voting Classifier, SVM, Logistic Regression, and Neural Network Techniques are recommended. This research study aims to predict aero plane delays, the biggest economic drain on many nations' fastest and most convenient means of transportation. Machine learning systems can identify and eliminate aircraft delays, saving money.

*[12]* In this dissertation, machine-learning algorithms are utilized to predict airline on-time performance. In one chapter, an approach for improving model performance and minority class prediction with unbalanced data is described. The other chapter offers a machine learning framework to determine if an aircraft is on-time or delayed. Using the second chapter's technique, the proposed framework outperforms the literature. Another part covers domestic flights from January 2015 to December 2017 using FAA data. Weather, aircraft, airport, and scheduling variables are examined for their effects on predicting delays. Using the chapter's model, an airline decision-maker can determine if a flight is delayed, why it's delayed, and how long it's been delayed.

*[13]* Delays hurt passengers, airports, and airlines. Commercial aviation participants must consider predictions while making decisions. Complexity of the air transportation system, variety of forecast methodologies, and amount of flight data made it impossible to estimate flight delays precisely. This study analyses aircraft delay prediction models using data science. We present a taxonomy and provide a review of projects used to address the flight delay prediction problem, with a focus on machine learning. We also present a history of key works that highlights flight delay prediction concerns and research to overcome them

*[14]* Airline industry efficiency depends on light delay prediction. Complexity of the air transportation system, variety of forecast approaches, and amount of flight data made it impossible to estimate flight delays precisely. This study evaluates the flight delay prediction models used in this situation. Airline delays hurt national budgets and corporations. Delays hurt passengers, airports, and airlines. We recommend linear regression machine learning. This study endeavor aims to predict aircraft delays because they cost many nations money and are the fastest and most comfortable means of transportation. Machine-learning algorithms can cut aircraft delays and save money. Flight delays can be inconvenient, especially if they last long enough to cause missed connections. If customers could predict if there will be a delay or how lengthy it may be, they might postpone future flights.

*[15]* Air traffic congestion caused by aviation industry growth delay flights. Flight delays have economic and environmental impacts. Air traffic control is difficult. Security issues, mechanical problems, weather, airport congestion, etc. can impede flights. We propose Boost regressed, Linear regression techniques. Airline delays hurt companies and a country's finances. This study endeavor aims to predict aircraft delays because they cost many nations money and are the fastest and most comfortable means of transportation. Machine-learning algorithms can cut aircraft delays and save money.

*[16]* The author of this paper examined flight delays, which are today a major issue in the aviation sector due to the resulting financial losses. U.S. Bureau of Transportation Statistics statistics shows that in 2018, over 20% of domestic flights in the United States were delayed. (BTS). This had a significant economic impact of $41 billion in the United States. These delays are inconvenient for both the airlines and the passengers. This causes trips to take longer, which raises costs for things like lodging and food as well as passenger tension. Rising labor expenses, aircraft repositioning, fuel use to reduce journey times, and other problems plague the airline industry, damaging its brand and often leading to lower client demand.

*[17]* The author of this study looked at the problem of flight delays, which has recently become a hot topic in the aviation industry due to the financial losses it has caused. The US Bureau of Transportation Statistics reports that in 2018, over 20% of US flights experienced delays (BTS). The 41 billion dollars in economic impact is substantial. These holdups are frustrating for everyone involved: the airlines, the passengers, and the passengers' schedules. This causes delays, which lengthens the duration of trips and raises prices for things like lodging and meals while also increasing traveler stress. Crew costs, aircraft repositioning, fuel consumption, and other things all add up to higher costs for airlines as they try to reduce travel times and, in turn, customer demand.

*[18]* The problem was transformed into a forecast of the possibility of airplane delays by means of machine learning, which was used in the research. A form of supervised machine learning known as binary classification was utilized in order to make the prediction. There were four criteria that were used to evaluate how well the algorithms performed, and seven different approaches were utilized to predict when delays would occur. Because of the imbalance in the data set, the evaluation metrics were weighted so as to exclude the preponderant influence of non-delayed flights in comparison to delayed flights. Following the application of classifiers to the delay prediction, the values of each model's four metrics were compared in order to determine how well each model performed.

*[19]* To ensure on-time arrivals for customers, airline schedules are based on the meticulous planning of resources, airports, aircraft, crews, etc. When airport capacity falls below demand, the timetable cannot be maintained and arrivals are late. When initial flight delays occur, crews and aircraft for subsequent flights are frequently unavailable, resulting in additional delays. This analysis introduces the concept of a delay multiplier to quantify this propagation and assesses the proposed metric using data from American Airlines. The delay multiplier appears to provide a very straightforward measurement of how a delay affects an airline's schedule, thereby contributing to airline and FAA choices and minimizing the system-wide impact of initial delays. There is a need for additional research to comprehend how the delay multiplier varies based on the airport of arrival and how airline connection models vary.

*[20]* This article compares ways to anticipating air traffic network delays. Three models are considered: A recently created aggregate model of delay network dynamics, the Markov Jump Linear System (MJLS), classical machine learning approaches like CART, and three possible Artificial Neural Network (ANN) topologies. We show that prediction performance depends on model/algorithm choice and prediction kind (for example, classification vs. regression). We also explore how choosing the correct predictor variables might increase algorithm performance. The models are evaluated using NAS data. The ANN is a decent method for classifying whether delays on the 100 most-delayed lines will surpass 60 minutes two hours in the future. The MJLS model predicts real delay levels on different links with a mean error of 4.7 min for a 2-hour horizon. MJLS predicts outbound delays at 30 major airports with a 6.8-minute mean inaccuracy for a 2-hour forecast horizon. Comparing temporal parameters and present delays' spatial distribution to anticipate future delays. The MJLS model, developed to represent aggregate air traffic dynamics, uses these parameters to anticipate the future spatial distribution of delays. So, model simplicity and prediction accuracy trade off.

*[21]* About 20% of airline flights arrive more than 15 minutes late, costing billions of dollars annually. In this situation, airlines and travelers must predict flight delays. This effort aims to implement a weather-based flight delay prediction. The anticipated arrival delay takes flight information (origin airport, destination airport, scheduled departure and arrival time) and weather conditions into account. Parallel algorithms deployed as MapReduce applications on a Cloud platform examined airline flight and weather observation datasets. Predicting delays above a threshold is accurate. With a 15min delay threshold, we get 74.2% accuracy and 71.8% recall for delayed flights; with a 60min barrier, we score 85.8% accuracy and 86.9% recall. Experiments show prediction scalability when data preparation and mining are done as MapReduce programmed in the cloud.

*[22]* In this article, we investigate the possibility of delays spreading over aircraft networks that are used by passenger airlines. The purpose of this study is to gain a deeper comprehension of the connection that may be established between the scheduling of aircraft and crew members and the operational performance of such schedules. When transportation companies, in particular, are deciding how to plan these limited and expensive resources, a significant portion of their attention is directed into determining how to maximize resource usage. The resulting plans, on the other hand, frequently contain very little wiggle room, which limits the capacity of the network to absorb disruption. As a result, initial flight delays may cause following flights to be delayed as well. A crucial step in the process of developing tools for the construction of more resilient airline plans is coming to terms with the relationship that exists between network plans and the propagation of delay. In this study, we analyze this link by using flight data given by two large U.S. carriers, one of which is a "low-fare" carrier and the other is a classic hub-and-spoke carrier.

*[23]* We construct a model in this work that can estimate flight departure delay distributions, which is a requirement for models that predict air traffic congestion. We identify and investigate the primary causes that contribute to airplane departure delays, and we create a predictive model for these delays using a strategic perspective. This model takes advantage of nonparametric approaches for analyzing daily and seasonal changes. In addition to that, the model makes use of a mixed distribution in order to estimate the errors that are still left over. We construct a global optimization version of the expectation–maximization algorithm by adopting ideas from genetic algorithms in order to circumvent issues that arise due to the presence of local optimum solutions in the mixed distribution. The model exhibits a level of goodness of fit that is acceptable, robustness with regard to the parameters that were selected for the model, and excellent predictive capabilities. In order to train and test our model, we made use of flight data from United Airlines and Denver International Airport during the years 2000 and 2001.

*[24]* Deep learning has improved image identification, speech recognition, machine translation, etc. Inspired by the paradigm's success, deep learning algorithms have been used to large data analytics challenges like traffic flow prediction. Deep learning techniques haven't been used to analyze air traffic data. This study analyses deep learning algorithms' air traffic delay prediction effectiveness. By merging numerous deep learning models, an accurate and resilient prediction model was created to analyze air traffic delay trends. RNN models sequential data quite accurately. Long Short-Term Memory RNN models daily departure and arrival flight delays at an airport. Deeper RNNs improve accuracy. This paper discusses four deep RNN architectures. The proposed prediction model's accuracy was assessed, examined, and compared to existing methods. It's the most accurate way.

*[25]* This research examines a number of different cost components that are generated by flight delays. Some of these cost components include the cost to airlines, the cost to passengers, the cost of lost demand, and the indirect impact that delay has on the economy of the United States. This study takes into account a wider range of relevant expenses than do traditional estimates of cost-of-delay, and it uses a variety of cutting-edge methods to determine how best to evaluate the magnitudes of these costs. It is important to take note of the passenger delay cost estimates since they take into account the fact that flight cancellations and missed connections can result in significant passenger delays that are not reflected in standard flight delay data.

**3. MATERIALS USED**

### 3.1. MODELS

Algorithm/ technique used

## Random Forest :

## Random Forest is a classifier that uses multiple decision trees on different subsets of the provided dataset and averages the results to increase the dataset's predictive accuracy. Instead of depending on a single decision tree, the random forest uses predictions from each tree and predicts the result based on the votes of the majority of predictions.

## Higher accuracy and overfitting are prevented by the larger number of trees in the woodland.

## Popular machine learning algorithm Random Forest is a part of the guided learning methodology. It can be applied to ML issues involving both classification and regression. It is founded on the idea of ensemble learning, which is a method of combining various classifiers to address complex issues and enhance model performance.

## Decision Tree Classifier:

## A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favored for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each child node for the classification result.

## The Decision Node and Leaf Node are the two elements in a decision tree. While Leaf nodes are the results of decisions and do not have any additional branches, Decision nodes are used to make decisions and have numerous branches.

## The given dataset's features are used to conduct the test or make the decisions.

## It is a graphical representation for obtaining all feasible answers to a decision or issue based on predetermined conditions.

## Equations

The highest value for each measure is shown in the table with a bold font. The lowest value for each measure is labeled with an underscore.

It can be seen that during the prediction, the one with the best performance among the seven algorithms is the Decision Tree model. For example, the accuracy value for the Decision Tree is 0.9778. This value is significantly higher than that of Gradient Boosted Tree, with the second-greatest value of 0.9334 accuracy. Similar patterns of noticeable differences for performance scores of Decision Tree can also be seen in the other three measures. Besides the Decision Tree, the two tree-based ensemble classifiers Random Forest and Gradient Boosted Tree, are also better performed than others. The values of measures of these two algorithms are relatively similar. The difference between their performance scores and the other four algorithms is also significant.

The lowest scores for the four measures occur in two of the seven models, and they are both base classifiers. KNN model has the lowest accuracy and recall, while Gaussian Naïve Bayes has the lowest precision and f1-score. Meanwhile, the precision value of KNN is particularly low, which is only 0.7501. KNN also has the second-smallest value of accuracy and recall, while the second-smallest F1-Score belongs to the Logistic Regression model instead of the Gaussian Naïve Bayes model. Therefore, it is possible to conclude that the worst-performing algorithm among the seven selected models when predicting the given data set is the KNN model.

## Statistical Analysis

## Statistical model requires the use of correlation analysis, parametric and non-parametric tests, multivariate analysis and econometric models.

## Probabilistic Model

## Probabilistic model requires analysis tools that estimates the probability of an event based on the historic data. The estimated outcome is given in form of a distribution function of the probability. The factor of randomness always makes an impact on the decision or the outcome produced by the probabilistic model.

## Machine Learning

### Supervised Machine learning could be a task where the dataset input and also the output are recognized, then many algorithms are used to analyze this data to map new examples. Here in this case is that the prediction of delay in flight

### 3.2. **DATASET**

The implementation of my proposed solution has been tested on 3 datasets:

1. flights.csv /Kaggle/input/flight-delays/flights.csv

Dimensions: 25 columns and 10,951 rows

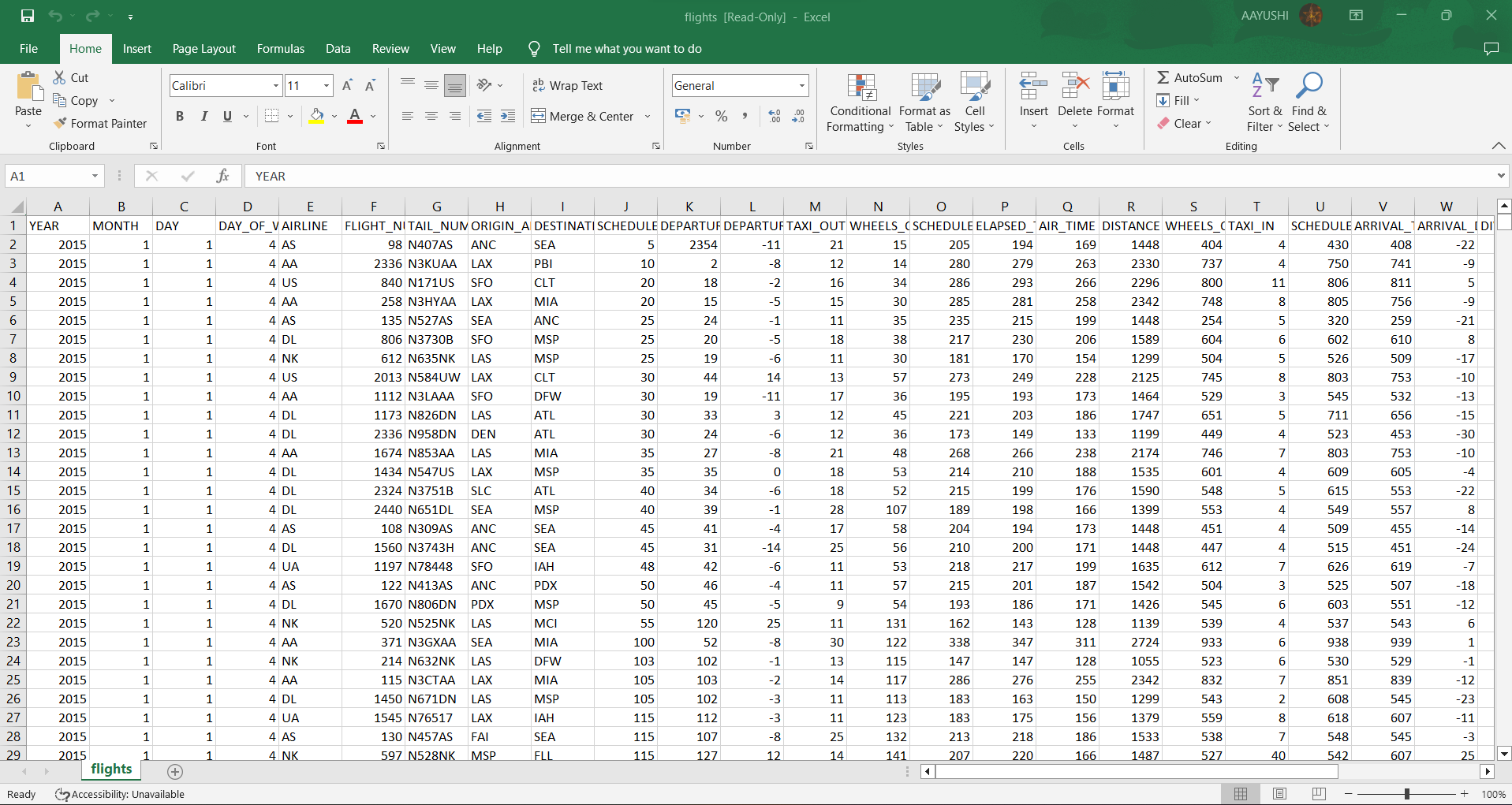


Fig.1.1 flights dataset

2. airlines.csv /Kaggle/input/flight-delays/airlines.csv

Dimensions: 2 columns, 15 rows

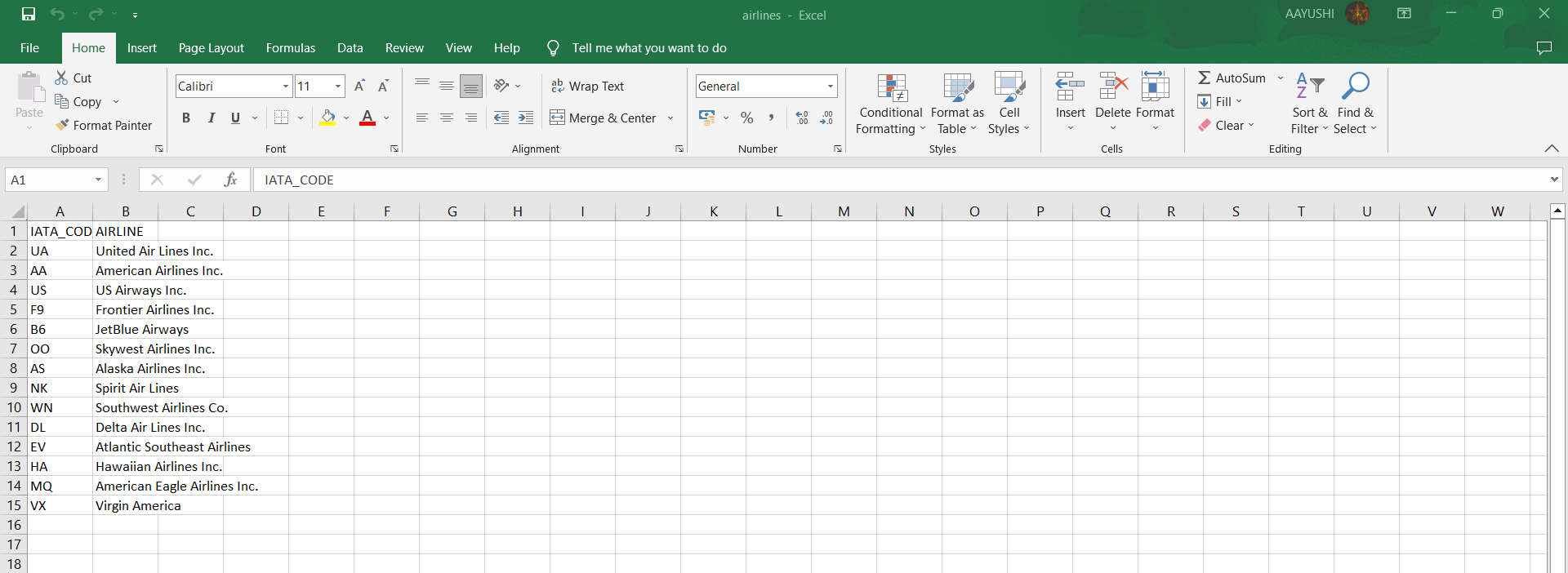
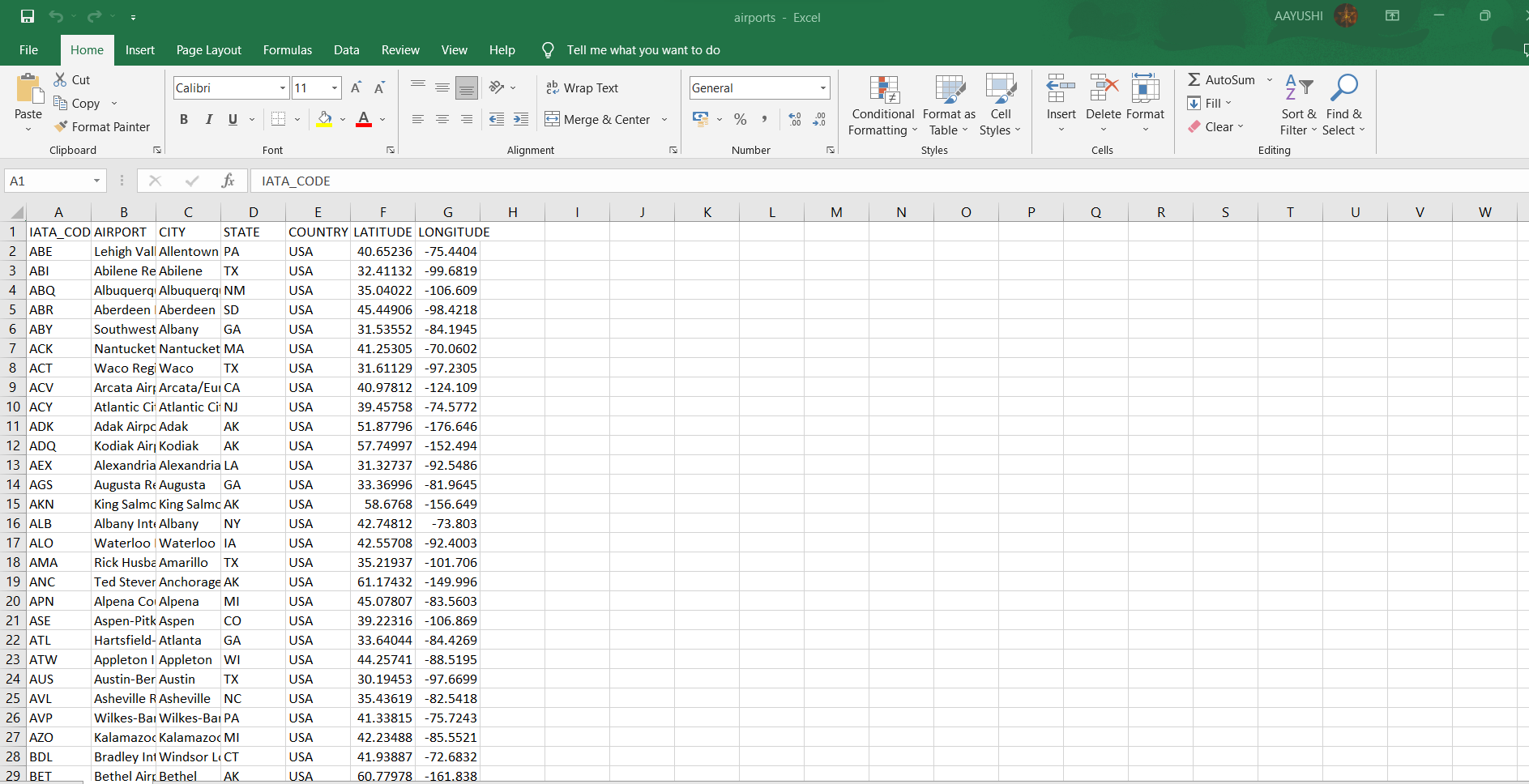


Fig.2. Airlines dataset

3. airports. csv /Kaggle/input/flight-delays/airports.csv

Dimensions: 7 columns, 323 rows



### Fig. 3. Airports dataset

# PROPOSED WORKS

# 4.1. NOVELTY

# The use of Random Forest algorithm for predicting flight delays offers several novel advantages compared to other methods. Here are some of the key novelties:

# Ensemble of decision trees: Random Forest is an ensemble learning technique that combines multiple decision trees to make predictions. Each tree in the forest is trained on a random subset of the data and a random subset of features, making it more robust to overfitting and capable of capturing complex interactions in the data.

# Feature importance: Random Forest provides a measure of feature importance, which allows for identifying the most influential features in predicting flight delays. This information can be used for feature engineering, selecting relevant features, and gaining insights into the factors that contribute to flight delays.

# Handling missing data: Random Forest can handle missing data effectively. It can make accurate predictions even when some of the data points have missing values. This is particularly useful in flight delay prediction where missing data can be common due to various reasons such as incomplete flight records or sensor failures.

# Non-linear relationships: Random Forest is capable of capturing non-linear relationships in the data. Flight delays are often influenced by various complex and non-linear factors such as weather conditions, air traffic congestion, and airline operations. Random Forest can capture such non-linear patterns in the data, making it suitable for predicting flight delays accurately.

# Robustness to outliers: Random Forest is robust to outliers in the data. Outliers in flight delay prediction can be caused by unusual events such as extreme weather conditions or operational disruptions. Random Forest can handle such outliers effectively without being overly influenced by them, leading to more robust predictions.

# Scalability: Random Forest is scalable and can handle large datasets with high-dimensional features. Flight delay prediction often involves a large amount of data, including information such as historical flight records, weather data, and airport operations. Random Forest can handle such large datasets efficiently, making it suitable for real-world applications.

# Interpretability: While Random Forest is an ensemble method, it can still provide some interpretability. Feature importance measures can help understand the relative importance of different features in predicting flight delays. Additionally, decision trees in the Random Forest can be visualized, allowing for a better understanding of the decision-making process and model outputs.

# Overall, the use of Random Forest for flight delay prediction offers several novel advantages, including ensemble of decision trees, handling missing data, capturing non-linear relationships, robustness to outliers, scalability, and interpretability. These advantages make Random Forest a powerful and effective approach for predicting flight delays in the aviation industry. So, it can be considered as a novel approach in the field of flight delay prediction. However, it is important to note that the performance of the Random Forest model depends on the quality and quantity of data available, the specific implementation, and the problem at hand. Proper validation and evaluation should be performed before deploying any model in real-world applications. Always consult with domain experts and follow best practices for model development and evaluation. Additionally, other machine learning algorithms, such as deep learning, can also be used in conjunction with Random Forest to further improve prediction accuracy in flight delay prediction tasks. Further research and experimentation may be required to continuously improve the accuracy and effectiveness of flight delay prediction models using Random Forest or other machine learning techniques. It's always a good practice to stay updated with the latest research and advancements in the field of machine learning for aviation-related applications.

# RESULTS AND DISCUSSION

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 passengers 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 passenger and the airlines which in turn is liked to another set of airline and passengers 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 monetary loss and for the higher and smooth operation. Classification or regression 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. Also, the scope of this project is very much confined to flight and weather data of United States, but we can include more countries like China, India, and Russia. Expanding the scope of this project, we can also add the flight data from international flights and not just restrict our self to the domestic flights.

1. **CONCLUSION**

The results obtained from using two machine learning algorithms: SVM Classifier and Random Forest combined with decision tree classifier can be observed from the given table: -

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **SVM Algorithm** | **Random Forest and Decision Tree Classifier Algorithm** |
| Precision | 0.68 | 0.32 |
| Recall | 0.34 | 0.21 |
| Accuracy | 60% | 47% |
| Weather forecast precision | 0.85 | 0.60 |
| Visibility | 72% | 53% |
| Delay prediction | 60.2% | 42% |

The accuracy score of 60% is already encouraging, meaning that even this first classification model guesses right 60% of the time, better than a coin flip. We know that the positive case of a delay of more than 15 minutes is less frequent than no delay (around 2800 delayed instances to more than 7000 for no delay), so we may prefer to look at the F1 measure (47% here), which is calculated by using the harmonic mean of precision and recall.

1. **FUTURE WORKS**

Further supportive studies are needed to relate all issues, scope and methods in order to obtain the most accurate results possible. Weather conditions are the main reason for flight delays, but unprecedented events such as large-scale natural and man-made disasters can cause significant flight delays. The demand for air travel is currently increasing significantly. Therefore, flight delay analysis has become a popular research area. Different researchers conducted the research using different machine learning and data mining techniques. They were interested in various aspects such as the location of airport facilities, weather conditions and airport capacity. Machine learning allows researchers to process large amounts of flight data to store and process.

This study, conducted by Khaksar and Sheikholeslami, used Bayesian modeling, decision trees, and cluster classification to estimate the occurrence and severity of delays in the US and Iranian aviation networks [1]. They found that visibility, wind and departure times, the age of Iran's fleet and aircraft type are the main parameters that affect U.S. airline networks.

Esmaeilzadeh and Mokhtari Mousavi used support vector machines (SVMs) to investigate causes and patterns of flight delays at his three major airports in New York [2]. Several explanatory variables were tested to determine their association with flight delays, airport operations, and flow management. Their probabilities of causing delays were calculated and compared to better understand the causes of departure delays.

* 1. REFERENCES

1. Tang, Y. (2021, October). *Airline Flight Delay Prediction Using Machine Learning Model*. Airline flight delay prediction using machine learning models. Retrieved November 30, 2022, from<https://dl.acm.org/doi/fullHtml/10.1145/3497701.3497725>
2. Javier, H. (2020, October). Using Machine Learning to Predict Flight Delays. Retrieved November 30, 2022, from<https://medium.com/analytics-vidhya/using-machine-learning-to-predict-flight-delays-e8a50b0bb64c>
3. Sternberg Alice, Soares Jorge, Carvalho Diego & Ogasawara Eduardo. (2017) A Review on Flight Delay Prediction, from<https://www.researchgate.net/publication/315382748_A_Review_on_Flight_Delay_Prediction>
4. Shaik Subhani & Teja, KP. (2019). Flight Delay Prediction Using Machine Learning Algorithm XGBoost. Journal of Advanced Research in Dynamical and Control Systems. 11.
5. Vishrut , R., Viran, R., Satyam, S., & Adityanath, M. (2021, June). *Flight delay prediction - IJIRT* . Flight Delay Prediction. Retrieved November 30, 2022, from<https://ijirt.org/master/publishedpaper/IJIRT151753_PAPER.pdf>
6. Sternberg, A., Soares, J., Carvalho, D., & Ogasawara, E. (2021, April 4). *A review on flight delay prediction*. arXiv.org. Retrieved November 30, 2022, from<https://arxiv.org/abs/1703.06118v3>
7. Alok, D. (2020, December). *Airline delay prediction using machine learning algorithms*. wichita. Retrieved November 30, 2022, from<https://soar.wichita.edu/bitstream/handle/10057/19743/d20039s_Dand_Sequester.pdf>
8. Mohammed, H. K. A. (2022, May). *Flight Delay Prediction Using Machine Learning* . Flight Delay Prediction Using Machine Learning . Retrieved November 30, 2022, from<https://jespublication.com/upload/2022-V13I5032.pdf>
9. V. Natarajan, S. Meenakshisundaram, G. Balasubramanian and S. Sinha, "A Novel Approach: Airline Delay Prediction Using Machine Learning," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 2018, pp. 1081-1086, doi: 10.1109/CSCI46756.2018.00210.
10. Yogita, B. (2022, March). *Flight delay prediction system* . Flight Delay Prediction System, from<https://www.ijert.org/research/flight-delay-prediction-system-IJERTV9IS030148.pdf>
11. Britto, Rodrigo & Dresner, Martin & Voltes, Augusto. (2012). The impact of flight delays on passenger demand and societal welfare. Transportation Research Part E-logistics and Transportation Review - TRANSP RES PT E-LOGIST TRANSP. 48. 10.1016/j.tre.2011.10.009.
12. Dand, A. (2020, December 1). *Airline delay prediction using machine learning algorithms*. Shocker Open Access Repository Home. Retrieved December 4, 2022, from https://soar.wichita.edu/handle/10057/19743
13. Carvalho, Diego & Ogasawara, Soares, Jorge Eduardo. (2017). Flight Delay Prediction.
14. Sternberg, Alice & Soares, Jorge & Carvalho, Diego & Ogasawara, Eduardo. (2017). A Review on Flight Delay Prediction.
15. Abramson, N. A. (2020). *Future flight: A review of the small aircraft transportation system concept* . Air Transportation Challenges . Retrieved December 4, 2022, from https://nap.nationalacademies.org/read/10319/chapter/5
16. *Reporting of causes of flight delays and cancellations* (2022). Understanding the Reporting of Causes of Flight Delays and Cancellations | Bureau of Transportation Statistics
17. *Understanding the reporting of causes of flight delays and cancellations* (2022, January 10). Understanding the Reporting of Causes of Flight Delays and Cancellations | Bureau of Transportation Statistics, from https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations
18. Zoutendijk, Mike & Mitici, Mihaela. (2021). Probabilistic Flight Delay Predictions Using Machine Learning and Applications to the Flight-to-Gate Assignment Problem. Aerospace. 8. 152. 10.3390/aerospace8060152.
19. Beatty, R., Hsu, R., Lee Berry, & James Rome. (2016, August). *AIAA Aerospace Research Central*. Aerospace Research Central. Retrieved December 3, 2022, from<https://arc.aiaa.org/doi/abs/10.2514/atcq.7.4.259?cookieSet=1>
20. 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
21. Calabria, L. B. U. of, Belcastro, L., Calabria, U. of, Calabria, F. M. U. of, Marozzo, F., Calabria, D. T. U. of, Talia, D., Calabria, P. T. U. of, Trunfio, P., Research, M., & Metrics, O. M. V. A. (2017, January 1). *Using scalable data mining for predicting flight delays*. ACM Transactions on Intelligent Systems and Technology. Retrieved December 3, 2022, from<https://dl.acm.org/doi/abs/10.1145/2888402>
22. Yihan , G., Shervin AhmadBeygi, & Amy Cohn. (2007, April). *Analysis of the potential for delay propagation in passenger ... - mit*. University of Michigan. Retrieved December 3, 2022, from<http://web.mit.edu/sis07/www/cohn.pdf>
23. Yufeng Tu, Michael O Ball & Wolfgang S Jank (2008) Estimating Flight Departure Delay Distributions—A Statistical Approach With Long-Term Trend and Short-Term Pattern, Journal of the American Statistical Association, 103:481, 112-125, DOI: [10.1198/016214507000000257](https://doi.org/10.1198/016214507000000257)
24. Kim, Y. J., Dimitri Mavris, Simon Briceno, & Sun Choi. (2016). *A deep learning approach to flight delay prediction* . IEEEXPLORE. Retrieved December 3, 2022, from https://ieeexplore.ieee.org/document/7778092
25. Michael , B., & Cynthia Barnhart. (2010). *Total Delay Impact Study*. NEXTOR. Retrieved December 3, 2022, from https://isr.umd.edu/NEXTOR/pubs/TDI\_Report\_Final\_11\_03\_10.pdf

**Appendix**

*Dataset Analysis plots:*

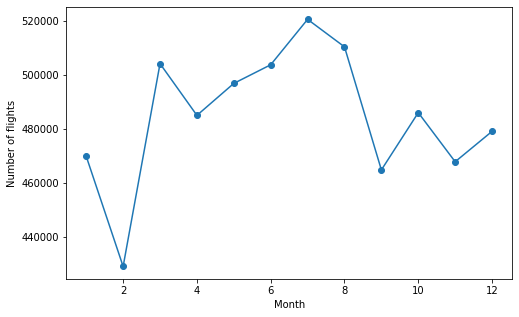


Fig. 4. Number of flights per month

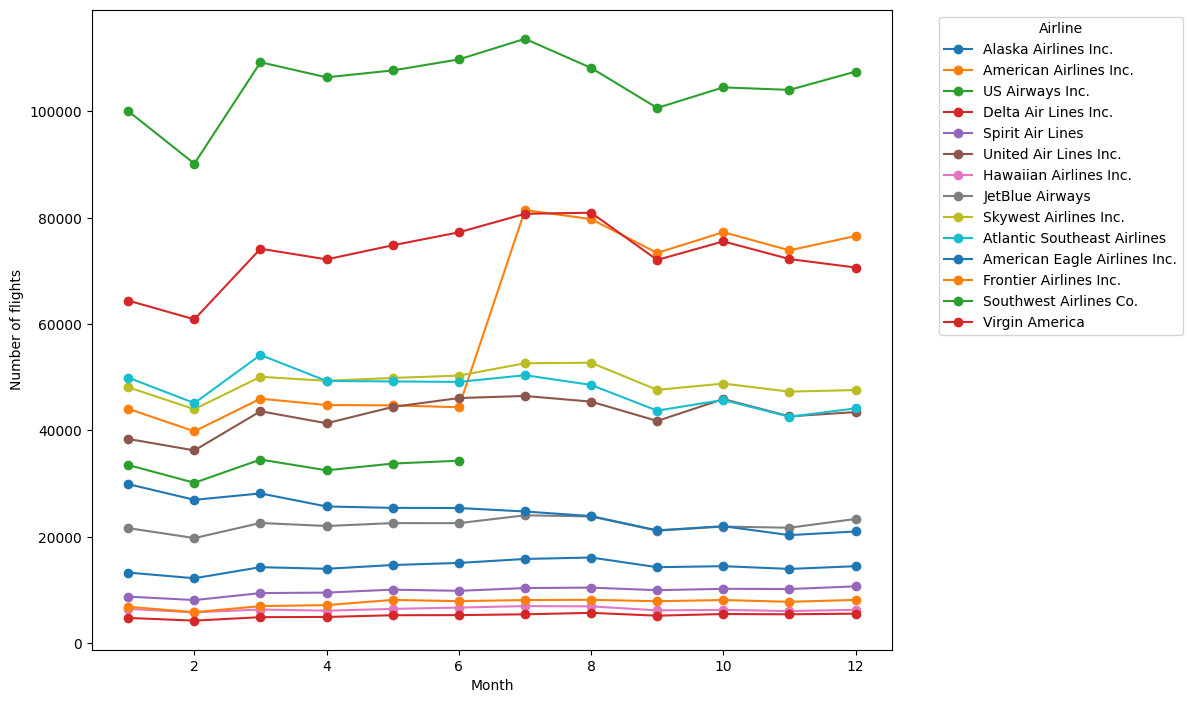


Fig 5. Number of flights per month in every category

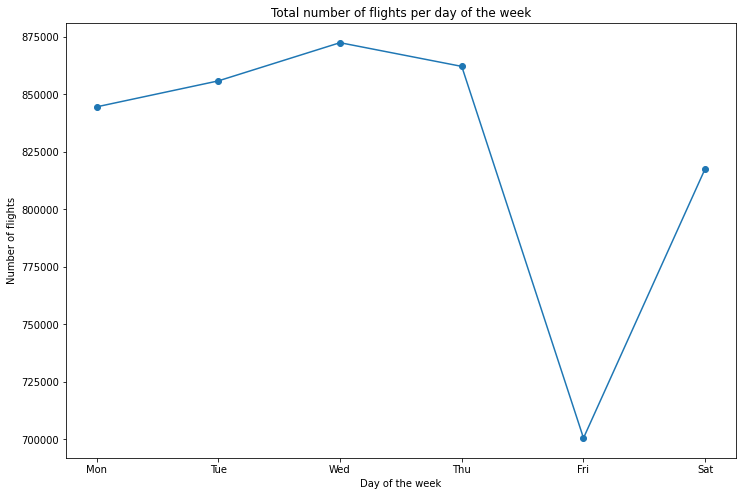


Fig 6. Total number of flights per day of a week

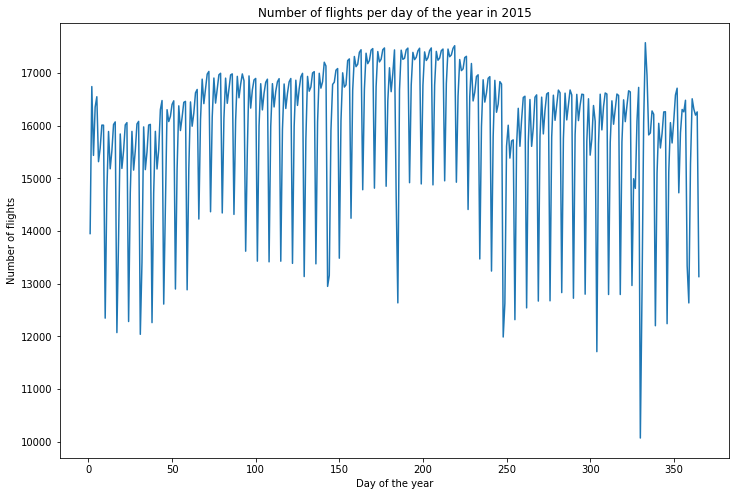


Fig 7. Number of flights per year

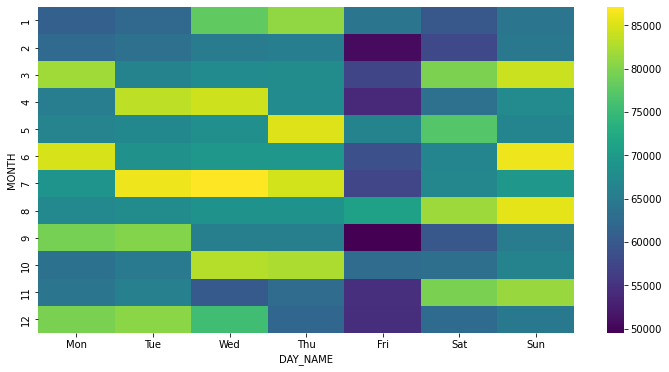


Fig 8. Number of flights per day name per week