JES
Journal of Engineering Sciences

Vol 13, Issue 05, MAY/2022

ISSN NO:0377-9254

# Flight Delay Prediction Using Machine Learning

Mohammed Ayaz Hussain Khan, BE, Department of CSE, ayazhussaink1@gmail.com

Mohammed Farhan Uddin, BE, Department of CSE, Mohdfarhan2001624@gmail.com

Mohammed Abdul Wajid Sarshaar, BE, Department of CSE, wajidsarshaar07@gmail.com

Dr. Jameel Hashmi, Ph.D, Associate Professor, Department of CSE, drjhashmiqa@gmail.com

**ABSTRACT: Growth in aviation industries has resulted in air-traffic jamming causing flight delays. Flight delays not only have economic impact but also injurious environmental properties. Air-traffic supervision is becoming increasingly challenging. Airlines delays make immense loss for business field as well as in budget loss for a country, there are so many reasons for impede in flights some of them are, some of them are due to security issues, mechanical problems, due to weather conditions, Airport congestion etc. we are proposing machine learning algorithms like Random Forest,Decision Tree,MLP Classifier, Naive Bayes, KNN,Gradient Boosting Classifier, Voting Classifier, SVM,Logistic Regression,Ridge Regression and Neural Network Techniques. The aim of this research work is to predict Flight Delay, Which is highest economy producing field for many countries and among many transportation this one is fastest and comfort, so to identify and reduce flight delays, can dramatically reduce the flight delays to saves huge amount of turnovers, using machine-learning algorithms.**

**Keywords-** *Random Forest, Decision Tree, MLP Classifier, Naive Bayes, KNN, Gradient Boosting Classifier, Voting Classifier, SVM, Logistic Regression, Ridge Regression and Neural Network Techniques*

## 1. INTRODUCTION

Air transportation system is one of the crucial modes of modern versatility. With increasing congestion in airtraffic and passenger-traffic, it is important to maintain persistence and resilience. Availability of land and resources contribute to the infrastructure of airports. The norms of improving technology and procedure are to maintain safety, efficiency, capacity, etc., Therefore, the National Airspace System (NAS) focuses on minimizing the environmental effects as a result of improvisation. With the current technology in hand, passengers can visualize their flight path, altitude, heading and other related parameters during their journey. However, air-traffic authorities continuously try to depreciate the delay in departure and arrival of flights. Though their efforts were in phase, the outcome is undesirable as the delays are in terms of hours sometimes causing chaos. Some important parameters that cause delay include weather, maintenance, security, and carrier. Corporate travel and tourism are the two major contributors to flight transportation system which is expected to be doubled by 2030. As a result of this increase, the airtraffic is also expected to increase in the same multiple. To minimize the air-traffic congestion new airports can be constructed. But, the complexity still grows exponentially. Hence, the only possible way of minimizing the delay is to improvise the existing airports. Considering the limited availability of land resources, the latter is more of a logical solution. Delay basically represents the period by which the aircraft is late or cancelled. Commercial aviation is likely to

be affected if there is a delay in their mobility. This delay results in the dissatisfaction of trusted customers and sometimes even marketing strategies. With a view of understanding the flight system, scientists and researchers stored the vast amount of data recorded over the entire course of a flight journey.

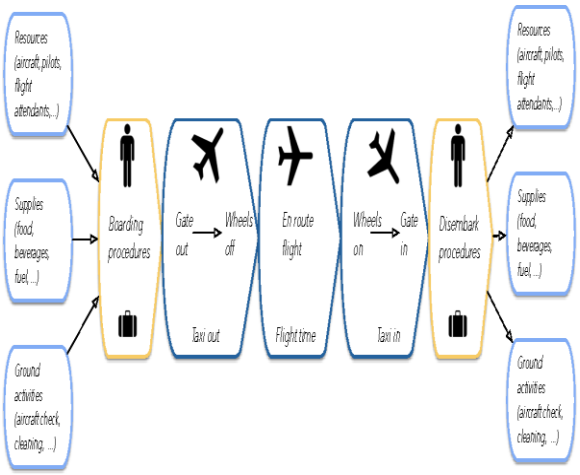



**Fig.1: A typical operation of a commercial flight**

As population increases tremendously and time is everything for many billionaire. Here the importance of Flights were raised, but due to high cost and some continuous delay of flight made less eyes on flights in 1960's, but due to government help many companies have been started manufacturing flights with less cost and more comfort and many Airports, this made control of airlines traffic. Airlines Economy play a predominant role in countries economy, so there is huge losses had occurred, we all know recent technology of Machine learning is one of the way to determine the flight delays. Mining techniques for instances applied to airlines topics rise rapidly due to their high concert in predicting outcomes, reducing costs of cancellation, promoting excellent airline transportation, improves customers counting and making real time choice to save people's time, money and completing their work smoothly.

## 2. LITERATURE REVIEW

### *2.1 A machine learning approach for prediction of on-time performance of flights:*

One of the major business problems that airlines face is the significant costs that are associated with flights being delayed due to natural occurrences and operational shortcomings, which is an expensive affair for the airlines, creating problems in scheduling and operations for the end-users thus causing bad reputation and customer dissatisfaction. In our paper, a two-stage predictive model was developed employing supervised machine learning algorithms for the prediction of flight ontime performance. The first stage of the model performs binary classification to predict the occurrence of flight delays and the second stage does regression to predict the value of the delay in minutes. The dataset used for evaluating the model was obtained from historical data which contains flight schedules and weather data for 5 years. It was observed that, in the classification stage, Gradient Boosting Classifier performed the best and in the regression stage, Extra-Trees Regressor performed the best. The performance of the other algorithms is also extensively documented in the paper. Furthermore, a real-time Decision Support Tool was built using the model which utilizes features that are readily available before the departure of an airplane and can inform passengers and airlines about flight delays in advance, helping them reduce possible monetary losses.

### *2.2 Analysis of the potential for delay propagation in passenger airline networks:*

The paper analyzes the potential for delays to propagate in passenger airline networks. The aim is to better understand the relationship between the scheduling of aircraft and crew, and the operational performance of such schedules. In particular, when carriers decide how to schedule costly resources, the focus is primarily on achieving high levels of utilization. The resulting plans, however, often have little slack, limiting the schedule's ability to absorb disruption; instead, initial flight delays may propagate to delay subsequent flights as well. Understanding the relationship between planned schedules and delay propagation is a requisite precursor to developing tools for building more robust airline plans. This relationship is investigated using the flight data provided by two major US carriers, one traditional hub-and-spoke and one low-fare carrier operating a predominantly point-to-point network.

### *2.3 Estimation of arrival flight delay and delay propagation in a busy hub-airport:*

In recent years, flight delay problem blocks the development of the civil aviation industry all over the world. And delay propagation always is a main factor that impacts the flight's delay. All kinds of delays often happen in nearly-saturated or overloaded airports. This paper we take one busy hub-airport as the main research object to estimate the arrival delay in this airport, and to discuss the influence of propagation within and from this airport. First, a delay propagation model is described qualitatively in mathematics after sorting and analyzing the relationships between all flights, especially focused on the frequently type, named aircraft correlation. Second, an arrival delay model is established based on Bayesian network. By training the model, the arrival delay in this airport can be estimated. Third, after clarifying the arrival status of one airport, the impact from propagation of arrival delays within and from this busy airport is discussed, especially between the flights belonging to one same air company. All the data used in our experiments is come from real records, for the industry secret, the name of the airport and the air company is hidden.

### *2.4 Flight delay prediction system using weighted multiple linear regression:*

Airline delays caused by bad weather, traffic control problems and mechanical repairs are difficult to predict. If your flight is canceled, most airlines will rebook you on the earliest flight possible to your destination, at no additional charge. Unfortunately for airline travelers, however, many of these flights do not leave on-time. The issue of delay is paramount for any airlines. Therefore we intend to aid the airlines by predicting the delays by using certain data patterns from the previous information. This system explores what factors influence the occurrence of flight delays along with the intensity of the delays. Our method is based on archived data at major airports in current flight information systems. Classification in this scenario is hindered by the large number of attributes, which might occlude the dominant patterns of flight delays. The results of data analysis will suggest that flight delays follow certain patterns that distinguish them from on-time flights. Our system also provides current weather details along with the weather delay probability. We have achieved much better accuracy in predicting delays. We may also discover that fairly good predictions can be made on the basis on a few attribute.

## 2. EXISTING SYSTEM

An accurate estimation of flight delay is critical for airlines because the results can be applied to increase customer satisfaction and

incomes of airline agencies. There have been many researches on modeling and predicting flight delays, where most of them have been trying to predict the delay through extracting important characteristics and most related features. However, most of the proposed methods are not accurate enough because of massive volume data, dependencies and extreme number of parameters.

Disadvantages:

- Finding an accuracy of flight delay is less.
- It does not have required parameters for finding flight delay.

## 3. PROPOSED SYSTEM

This research work is to predict Flight Delay, Which is highest economy producing field for many countries and among many transportation this one is fastest and comfort, so to identify and reduce flight delays, can dramatically reduce the flight delays to saves huge amount of turnovers, using machine-learning algorithms. The results of this simulation indicate the potential delays in major airports including the time, day, weather, etc., and hence the volume of delay shall be minimum based on the constructed model.

Advantages:

- Due to the stochastic nature of delays, this research investigates the qualitative prediction of airline delays to implement necessary changes and provide better customer experience.

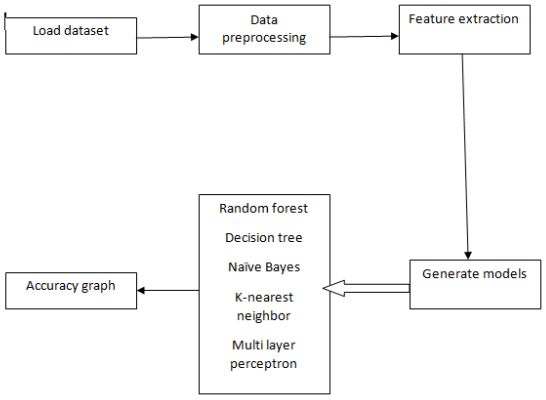



**Fig.2: System architecture**

**MODULES:**

- **Load dataset:** We will upload our dataset into application.
- **Data Preprocessing:** The quality of the data should be checked before applying our algorithms.
- **Feature Extraction:** Transforming raw data into numerical features that can be processed while preserving the information in the original data set.
- **Generate models:** After extracting features from the dataset we will generate our algorithms on that dataset.
- **Accuracy Graph:** We will plot the accuracies comparison graph between all the algorithms.

**ALGORITHMS:**

1. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

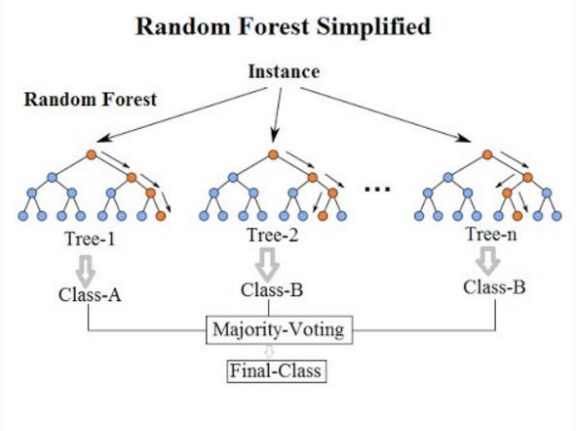



**Fig.3: Random forest architecture**

2. Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

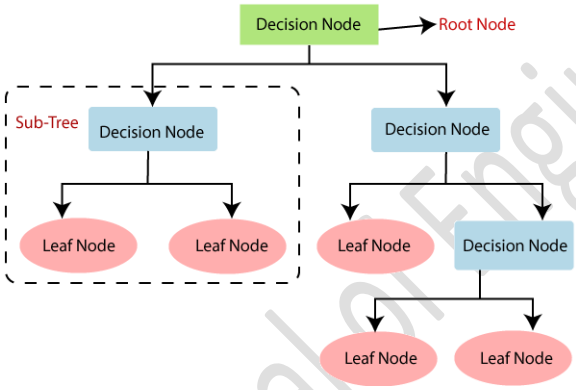



**Fig.4: Decision tree architecture**

3. Multi layer perceptron (MLPs) are suitable for classification prediction problems where inputs are assigned a class or label. They are also suitable for regression prediction problems where a real-valued quantity is predicted given a set of inputs.

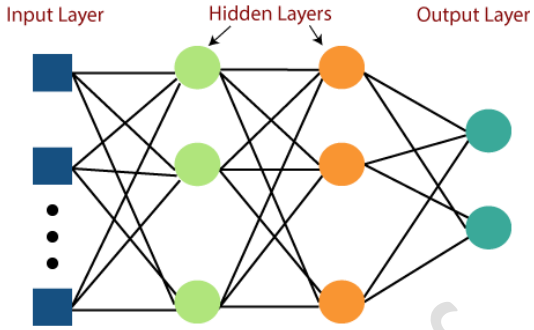



**Fig.5: Multilayer perceptron architecture**

4. Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.

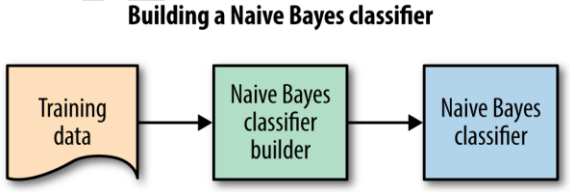



**Fig.6: Naïve bayes architecture**

5. The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems.

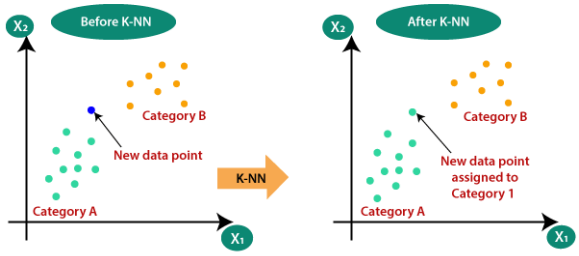



**Fig.7: KNN architecture**

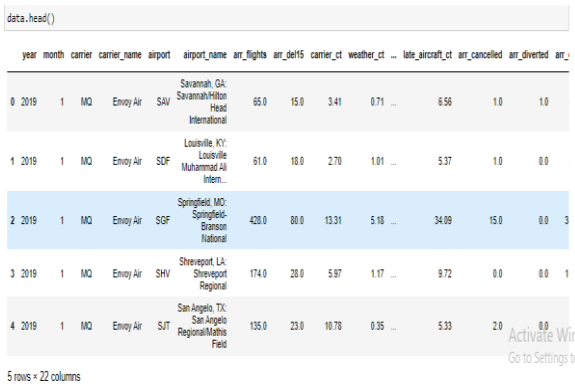
## 4. DATASET DESCRIPTION

To train and test models, we used a publicly available kaggle dataset for United States domestic air traffic. The original source of

our dataset is the on-line Bureau and Transportation Statistics database [4].The data set is for the year 2015 and consists of well over 3 Million examples with 19 features categorized as follows:

*3.1 Loading Dataset*

Step 1:- Open the python API module and select the CSV files and browser the data set to a Num_Py array and use it for machine learning. Here, we used read_csv method of pandas to import the dataset. Before improving the dataset, check the present working directory and select the directory where the data is available.



data.head()

| | year | month | carrier | carrier_name | airport | airport_name | arr_flights | arr_del15 | carrier_ct | weather_ct | ... | late_aircraft_ct | arr_cancelled | arr_diverted | arr_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2019 | 1 | MQ | Envoy Air | SAV | Savannah, GA: Savannah/Hilton Head International | 65.0 | 15.0 | 3.41 | 0.71 | ... | 6.56 | 1.0 | 1.0 | |
| 1 | 2019 | 1 | MQ | Envoy Air | SDF | Louisville, KY: Louisville Muhammad Ali Intern... | 61.0 | 18.0 | 2.70 | 1.01 | ... | 5.37 | 1.0 | 0.0 | |
| 2 | 2019 | 1 | MQ | Envoy Air | SGF | Springfield, MO: Springfield-Branson National | 428.0 | 80.0 | 13.31 | 5.18 | ... | 34.09 | 15.0 | 0.0 | 3 |
| 3 | 2019 | 1 | MQ | Envoy Air | SHV | Shreveport, LA: Shreveport Regional | 174.0 | 28.0 | 5.97 | 1.17 | ... | 9.72 | 0.0 | 0.0 | 1 |
| 4 | 2019 | 1 | MQ | Envoy Air | SJT | San Angelo, TX: San Angelo Regional/Mathis Field | 135.0 | 23.0 | 10.78 | 0.35 | ... | 5.33 | 2.0 | 0.0 | |

5 rows × 22 columns

**Fig.3: Dataset description**

*3.2 Data Frame*

A Data frame is a 2-dimensional data structure, i.e., data is associated with a tabular fashion in rows and columns. Features of Data Frame:

1. Potentially columns are of different types

2. Size – Mutable.

3. Labeled axes (rows and columns)

4. Can Perform Arithmetic operations on rows and columns.

*3.3 Dealing with Missing Values*

Datasets may contain the missing values, often encoded as blanks, NaNs or other placeholders. Such datasets however are incompatible with scikit-learn estimators which assume that all values in an array are numerical, and that all have and hold meaning. A basic strategy to use incomplete datasets is to discard entire rows and/or columns containing missing values. However, this comes at the price of losing data which may be valuable (even though incomplete). A better strategy is to impute the missing values, i.e., to infer them from the known part of the data. Here we use Imputer class of sklearn module and the methods used are transform and fit_transform.

*3.4 Splitting the Dataset*

A machine learning algorithm works in two stages-the testing and training stage. The training dataset (also called training set, learning set, or AI training data) is the initial dataset used to train an algorithm to understand how to apply technologies such as neural networks, to learn and produce complex results. It includes both input data and the corresponding expected output. The purpose of the training dataset is to provide your algorithm with “ground truth” data. The test dataset, however, is used to assess how well your algorithm was trained with the training dataset. You can’t simply reuse the training dataset in the testing stage because the algorithm will already “know” the expected output, which defeats the purpose of testing the algorithm. Here’s a cool flowchart that shows the training process and the different functions of training data and test data.

JES
Journal of Engineering Sciences

Vol 13, Issue 05, MAY/2022

ISSN NO:0377-9254

## 5. EXPERIMENTAL RESULTS

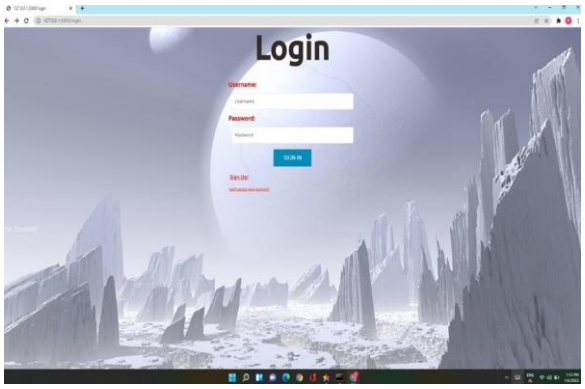



**Fig.4: Login page**

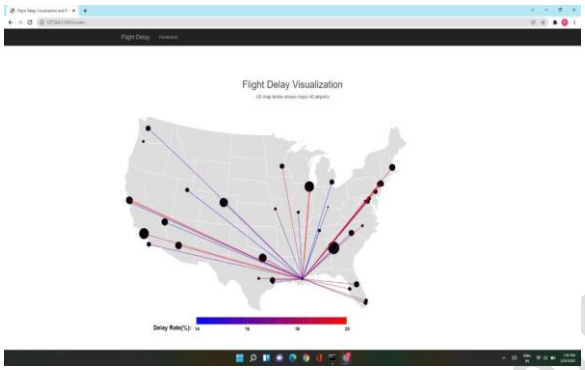



**Fig.5: Flight delay visualization**

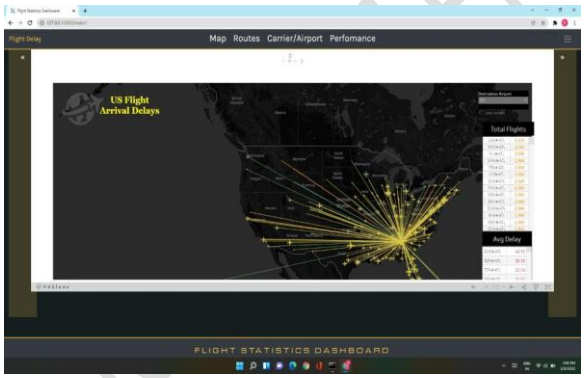



**Fig.6: Statistics**


**Fig.7 Graph**

**Table.1: Comparison table**

| S.No | Algorithm | Accuracy (%) |
|---|---|---|
| 1 | Random Forest | 74% |
| 2 | Decision Tree | 75% |
| 3 | Naive Bayes | 80% |
| 4 | Multi Layer Perceptron | 82% |
| 5 | K-Nearest Neighbor | 76% |

## 6. CONCLUSION

Predicting flight delays is on interesting research topic and required many attentions these years. Majority of research have tried to develop and expand their models in order to increase the precision and accuracy of predicting flight delays. Since the issue of flights being on-time is very important, flight delay prediction models must have high precision and accuracy.Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and regression can provide promising tools for inference in the cancer domain. Regardless of the type of prediction task at hand; regression or classification. It has become the state-of-the-art machine learning algorithm to deal with structured data. Compare to all algorithms MLP algorithm gives high accuracy that is 82%.

## 7. FUTURE SCOPE

Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and regression can provide promising tools for inference in the cancer domain. The XGBoost is used in the analysis of this paper because XGBoost is one of the most popular machine learning algorithms these days. Regardless of the type of prediction task at hand; regression or classification. It has become the state-of-the-art machine learning algorithm to deal with structured data.

## REFERENCES

1. Rebollo JJ, Balakrishnan H. Characterization and prediction of air traffic delays. Transportation Res Part C Emerg Technol. 2014;44:231–41.

2. Thiagarajan B, et al. A machine learning approach for prediction of on-time performance of flights. In 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC). New York: IEEE. 2017.

3. Reynolds-Feighan AJ, Button KJ. An assessment of the capacity and congestion levels at European airports. J Air Transp Manag. 1999;5(3):113–34.

4. Hunter G, Boisvert B, Ramamoorthy K. Advanced national airspace traffic flow management simulation experiments and vlidation. In 2007 Winter Simulation Conference. New York: IEEE. 2007.

5. AhmadBeygi S, et al. Analysis of the potential for delay propagation in passenger airline networks. J Air Transp Manag. 2008;14(5):221–36.

6. Liu YJ, Cao WD, Ma S. Estimation of arrival flight delay and delay propagation in a busy hub-airport. In 2008 Fourth International Conference on Natural Computation. New York: IEEE. 2008.

7. Tu Y, Ball MO, Jank WS. Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. J Am Stat Assoc. 2008;103(481):112–25

8. Oza S, et al. Flight delay prediction system using weighted multiple linear regression. Int J Eng Comp Sci. 2015;4(05):11765.

9. Evans JE, Allan S, Robinson M. Quantifying delay reduction benefits for aviation convective weather decision support systems. In Proceedings of the 11th Conference on Aviation, Range, and Aerospace Meteorology, Hyannis. 2004.

10. Hsiao C-Y, Hansen M. Air transportation network flows: equilibrium model. Transp Res Rec. 2005;1915(1):12–9.

11. Britto R, Dresner M, Voltes A. The impact of flight delays on passenger demand and societal welfare. Transp Res Part E Logist Transp Rev. 2012;48(2):460–9.

12. Pejovic T, et al. A tentative analysis of the impacts of an airport closure. J Air Transp Manag. 2009a;15(5):241–8.