

*Machine Learning Bootcamp - 2020*

**Team No. 4**

**Final Abstract**

**Submitted By-**

**Miguel Mendoza - Business Analyst**

**Celine Cherian - Developer**

**Pratik Ghatake – Data Engineer**

**Fatmir Likrama – Data Scientist**

**Abstract**

The air transportation is increasing tremendously nowadays. With high globalization, ease of flight journeys, rapidity of transport, it has become one of the most popular modes of transportation worldwide. This has caused lots of problems; because of growing air traffic and many more reasons, airways are becoming congested, leading to flight delays. The delay not just wastes the time of multiple people traveling but also causes environmental damage by rise in fuelwaste, causing increased unnecessary emissions. According to the US Department of Transportation, 18.87% of all flights in the US were delayed in 2019. These can cause billions of dollars in losses to businesses and consumers alike.

This motivates us to formulate this business problem and find probable solutions for it by learning from historical data. For this project, our focus would be predicting the flight delays with the use of informative attributes. To carry out the predictive analysis, the model will encompass various statistical techniques from Supervised Machine learning specifically, nonlinear algorithms, along with K-Means clustering. These techniques will study from historical data, specifically 2019, and will try to make future predictions. This detection will help airline companies in making well assessed decisions for designing their airways journeys and schedules, along with helping consumers by warning them about possible delays. The developed model can be used in real world scenarios with the purpose of improvisation of the airways experience and can save a lot of wastage, financial and environmental.

**Data Preparation and Analysis**

Dataset Resource: <https://www.kaggle.com/divyansh22/flight-delay-prediction>

Data Cleaning: The csv file we will be applying our classification algorithm is based on flight delays: it gives us info about each flight, along with whether it was delayed or not. This csv file has 21 columns: [Day of Month, Day of Week, Unique Carrier Code, Carrier Airline ID, Carrier, Tail Number, Flight Number, Origin Airport ID, Origin Airport Sequence ID, Origin Airport, Destination Airport ID, Destination Airport Sequence ID, Destination Airport, Departure Time, Departure Delay Indicator, Departure Time Block, Arrival Time, Arrival Delay, Cancelled Flight, Diverted Flight, Distance between airports]. Delay, cancelled, and diverted indicators are given with a 1 for yes or 0 for no.

This spreadsheet has 607,346 rows; one for each unique flight in the US. This allows us to simply drop any rows with missing values, after which we end up with 599,268. After this, we drop any column that has repeating information: for example, drop Destination Airport since we already have Destination Airport ID. After doing this, we end up with 15 columns instead of 21. We will be trying to predict whether a flight’s departure will be delayed based on the other columns in our spreadsheet; therefore, the columns indicating if a flight was cancelled, diverted, or delayed for arrival will also be removed, leaving us with 12 columns.

**Hypothesis**

After data preparation, Exploratory data analysis will be carried upon the dataset in order to find patterns and develop data understanding before decoding any algorithm. This may involve answering common questions such as how many flights were delayed, which airports are busiest, calculating delays mathematically etc.

Feature selection techniques will be applied to identify most informative attributes for the prediction. After which, the nonlinear algorithms will be applied. We will be implementing random forest classifier, decision tree. As far as decision tree are concerned gradient boosting algorithms will be used as an ensemble method. If we find any weak classifiers in the dataset, we would try to combine those weak classifiers into one and using Ada Boost classifier. Lastly, neural network by using keras will be used to learn about delays.

After using our non-linear algorithms, we will also implement k-means and k-modes clustering on our data and assess which technique yielded better results. We will be using two clusters: one for delayed and one for not delayed.