**Airline delays and cancellations data : Predicting Flight Delays**

**Background:**

Flight delay has become a very important subject for air transportation all over the world because of the associated financial losses that the aviation industry experiences. These delays not only cause inconveniences to the passengers but also to the airlines. Due to the increased travel time, passengers are forced to spend more money on food and lodging, ultimately resulting in stress. Airlines face extra costs due to crew repositioning, fuel consumption while trying to reduce elapsed times, and a variety of other factors that could potentially reduce their reputation and lower passenger demand. These delays can be attributed to a variety of factors, including air congestion, poor weather conditions, mechanical problems, difficulties boarding passengers, and simply the airline's inability to handle the demand.

**Proposal:**

The U.S Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT’s monthly Air Travel Consumer Report. The particular dataset has been made available on Kaggle, which was the source of reference. The entire dataset consists of data ranging from the year 2009 till 2018. For the purpose of analysis, we’ll be picking up the dataset for 2015 only. The total observations in the dataset exceed 5 million rows and 35 different variables. To tackle this for computational reasons, we plan on randomly sampling the dataset to reduce the size. The final number of observations is still to be determined. Apart from Flight details, we have 2 additional files for airlines and airport details that we’ll use for mapping into the main dataset.

Some SMART questions that we will answer through our analysis are:

1. Is there a delay\_label imbalance in the dataset and how to go around it? Being a classification problem, what is the best metric to evaluate our model?
2. Which feature is highly correlated with delay\_label?
3. Which classification methodology produces the best results to predict flight delay?
4. With what accuracy / precision / recall can we predict using the best model whether a flight is going to be delayed before it is announced on the departure board?
5. Which classification methodology produces the best results to predict flight cancellations?
6. With what accuracy / precision / recall can we predict using the best model whether a flight is going to be canceled?

By using Statistical Modeling and Machine learning methods, we will try to predict whether a Flight is going to be delayed or canceled, based on features that are available prior to this information being displayed on the departure boards.

**Repository & Dataset Source**

Dataset Source: Click [HERE](https://www.kaggle.com/datasets/usdot/flight-delays) to view the source of our dataset.

Github Repo: Follow this [LINK](https://github.com/KumarAditya98/DATS6101_Group5_Project2) to view our GitHub repository.