# U.S. FLIGHT DELAY & CANCELLATION PREDICTION

**PROJECT PROPOSAL**

**ISYE 6740: COMPUTATIONAL DATA ANALYTICS**

Maynard Miranda (ID# 903652021), Alfonzo Miguel Sabado (ID# 903656898)

## Problem Statement

Flight delays and cancellations is a significant problem for both travelers and airline companies. For travelers, it can cause inconvenience and missed events. For airline companies, it may cause profit loss and reputation damage. Predicting flight delays and cancellation is challenging due to various situations or events that can happen before the flight such as weather problems, aircraft downtime/mechanical problems, queuing issues, system problems, crew scheduling issues, and air traffic congestion. The goal of this project is to help address this problem, and improve an airline's ability to cancel flights in advance in order to improve:

(1) Customer experience. Canceling flights 4 hours before the flight, allows customers to avoid traveling to the airport only to find their flights canceled or moved. It also enables customers to rebook flights earlier, and find better routes.

(2) Planning of route recovery. By being more intentional on the flights/routes that are canceled, airlines can better plan how to get back to 100% operational capacity, with the least amount of delays.

The team will evaluate different models such as logistic regression, decision trees, neural networks, k-nearest neighbors, and support vector machines. Hyperparameter experimentation will be performed, and various feature engineering techniques discussed in the class will be explored to find the optimal feature combination that should be used in the model.

## Dataset

The data to be used for this project is the Airline On-Time Performance dataset which is created and maintained by the Bureau of Transportation Statistics. The data contains United States domestic flight information from 1987 to 2023, and has multiple features such as: airline carrier, airport departure and destination, arrival time, delay times, cancellation information, etc. In addition to flight data, the team will also be using airports local weather data from National Centers for Environmental Information (NCEI). The team will use 10 years of pre-pandemic data (2009 to 2019) for training and testing our models, and then use post-pandemic data (2022) to validate our models.

Per the team's initial review of the data, it was noted that there are missing values, outliers, data transformation, and data type problems that need to be fixed before creating the classification models. The team will use the appropriate data cleaning and preprocessing techniques to handle these identified issues to ensure optimal and accurate results.

## Methodology

The study will use both supervised and unsupervised machine learning techniques to create models that will predict flight delay and cancellation.

- *Pre-Processing:* The team will connect to the Bureau of Transportation Statistics and NCEI APIs to pull the relevant dataset. Various data pre-processing and feature engineering techniques will be used to clean and prepare the dataset for analysis.
- *Model:* Various classification models such as logistic regression, k-nearest neighbor, support vector machine, neural network, decision trees, and naive bayes will be developed to predict flight delay and cancellation. Experimentation on the model's hyperparameter and features will be performed to find the optimal result.
- *Evaluation:* Each developed model will then be evaluated using cross-validation technique and various classification measures such as accuracy, precision, f1-score, recall, AUC-ROC, true positive rate, false positive rate, true negative rate, and false negative rate to compare how well each model is performing.

## Evaluation and Final Results

The team will evaluate the performance of the model using cross-validation and evaluate the results using the different classification metrics such as accuracy, precision, recall, f1-score, and AUC-ROC, true positive rate, false positive rate, true negative rate and false negative rate. The team will also provide recommendations for future model improvement and future studies.

In conclusion, predicting flight delays and cancellations will significantly help airline companies and air travelers. The study will use the flight information dataset created and maintained by the Bureau of Transportation Statistics to create various supervised and unsupervised classification models and determine which model is the best based on different criterias.