ISTE 780 Project Proposal
**Forecasting Flight Delays: A Data-Driven Approach**

Chelsi Santiago
Brandon Habschied
Geordy Aponte

## Problem Description

The problem utilizing the Flight Status dataset will be to determine whether or not any particular upcoming flight will have a delayed arrival. There is a lot of value in being able to predict if a flight will have a delayed arrival. The definition of a delayed arrival in the terms of our data set will be if the plane arrives at its intended destination 15 minutes or later than the scheduled arrival time. We will be using data for all flights from 2018 through 2022 when trying to create a system to solve this problem.

## Data and Methodologies Being Tested

The Flight Status dataset was originally found on Kaggle and can be accessed here: https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022. This dataset captures flight data for five years, from 2018 to 2022, and has 61 features ranging from airline, flight date, departure time, and arrival delay. While a combination of features will be considered throughout this project, the primary focus will be on the delayed flight status. Principal component analysis will be used in order to reduce the dimensionality of the dataset.

In order to address the problem stated above, a variety of algorithms and machine learning methods will be used. Our system will utilize supervised machine learning to develop a classification system that will help us determine if a flight and its passengers will be delayed at their destination by 15 minutes or more. To successfully predict the delayed status of a flight, logistic regression, support vector machines, multilayer perceptron neural networks, traditional random forests and random forests with gradient boosting are considered candidate models to be tested and fitted to this dataset. These methods are good at handling large datasets that have a high dimension. The final results of this project will be compared to any prior work published on Kaggle using this dataset.

## Potential Real World Applications of Solution

Airlines and air traffic control would greatly benefit from a predictive model that can anticipate airline arrival delays and cancellations. Such a model would enable airlines to optimize their operations by proactively adjusting schedules, allocating resources efficiently, and making informed decisions about crew assignments. This would result in fewer delays and cancellations, reduced operational costs, and increased customer satisfaction. Air traffic control

can use this data to better manage airspace and runway capacity, reducing congestion and improving overall safety. By minimizing disruptions and enhancing operational efficiency, the aviation industry can save costs, maintain better on-time performance, and provide a more reliable and seamless travel experience for passengers.

To capture more market share, airlines can leverage this data-driven prediction model to gain a competitive edge. By consistently providing a more reliable and punctual service, airlines can attract and retain more passengers. Travelers are more likely to choose carriers with a reputation for on-time arrivals and fewer cancellations, as this reliability directly impacts their travel experiences. Airlines that can effectively utilize this predictive model to improve their overall service quality can differentiate themselves in a highly competitive market, ultimately expanding their customer base and market share. Additionally, proactive communication with passengers about potential delays or cancellations can enhance the airline's brand image and build loyalty among travelers, further contributing to market share growth.