# **CAPSTONE Project Approach – Berkley Haas College of Engineering March 2024 Cohort**

Table of Contents

[CAPSTONE Project Approach – Berkley Haas College of Engineering March 2024 Cohort 1](#_Toc175040591)

[Problem Statement 1](#_Toc175040592)

[Research Question: 1](#_Toc175040593)

[Expected Data Sources: 1](#_Toc175040594)

[Techniques Expected to be used for analysis: 2](#_Toc175040595)

[1. Data Preparation 2](#_Toc175040596)

[2. Exploratory Data Analysis 2](#_Toc175040597)

[3. Feature Engineering 2](#_Toc175040598)

[4. Iterative Modeling and Evaluation 3](#_Toc175040599)

[Expected Results 3](#_Toc175040600)

[Why is this question important? 3](#_Toc175040601)

[**1**. Enhanced Operational Efficiency 3](#_Toc175040602)

[2. Improved Customer Experience 4](#_Toc175040603)

[3. Data-Driven Decision Making 4](#_Toc175040604)

[4. Operational Resilience 4](#_Toc175040605)

[5. Financial Impact 4](#_Toc175040606)

[6. Safety and Compliance 4](#_Toc175040607)

[7. Competitive Advantage 5](#_Toc175040608)

[8. Route Optimization 5](#_Toc175040609)

[Summary 5](#_Toc175040610)

## Problem Statement

Predict Flight Delays based on, a) Flight Status, b) Weather, c) Air Traffic, d) Aircraft specifics, e) Ground and Passenger handling info

### Research Question:

How can a multi-class classification model be developed to predict flight delays by assessing both departure and arrival delays?

### Expected Data Sources:

Kaggle Dataset from [here](https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations/data), that is comprised of multiple csv's listed below.

1. Air Carrier Summary
2. Aircraft Inventory
3. Air Carrier employee support (Ground Crew, Flight Attendants)
4. Flight Status with Air Carrier info for 2019-2020
5. Airport Weather
6. Airport and Carrier look-up codes

The analysis will involve creating a combined dataset from multiple CSV files, resulting in approximately 26-35 features. The dataset will be appropriately sized to ensure it remains manageable for effective model training.

## Techniques Expected to be used for analysis:

### 1. Data Preparation

The data preparation process will involve cleaning and merging multiple CSV files to create a unified dataset with approximately 26-35 features. This step will include handling missing values, correcting inconsistencies, and ensuring data integrity before merging the files. The resulting dataset will be appropriately sized and structured to maintain manageability for effective model training.

### 2. Exploratory Data Analysis

Clustering algorithms, such as K-Means or DBSCAN, will be utilized to discover any natural groupings within the data. Investigate to see if these clusters reveal hidden patterns to help in segmenting flights by delay patterns or distance. Also, investigate these clusters for anomaly detection, identifying outliers. These insights will be used toward feature engineering.

### 3. Feature Engineering

Explore creating new features from existing data:

* Delay Categories: Classify delays into categories like No Delay, Moderate Delay, and Severe Delay for both departure and arrival times.
* Time-Based Features: Extract day of the week and part of the day from departure and arrival times.
* Weather Impact: Create features for weather conditions at airports, such as visibility, temperature, and overall severity.
* Route-Based Features**:** Include normalized or categorized flight distance and duration, and flag busy routes.
* Using Clustering Results:
* Cluster Membership: Add cluster labels as new categorical features to indicate patterns in delays or weather.
* Anomaly Scores: Create a feature representing the distance of each flight from its cluster center to identify potential outliers
* Interaction Features:
* Departure-Arrival Interaction: Combine departure and arrival delays to capture their relationship.
* Weather and Delay Interaction: Create interaction terms between weather severity and departure delays.
* Aggregation Features:
* Historical Averages: Average delays for routes, days, or time blocks based on historical data.
* Rolling Averages: moving averages for delay times or weather conditions to capture trends.
* Dimensionality Reduction:
* PCA: Use Principal Component Analysis to reduce dimensionality if this results in many features to retain most of the variance.
* Handling Categorical Features:
* One-Hot Encoding: Convert categorical features into one-hot encoded vectors
* Target Encoding as needed

### 4. Iterative Modeling and Evaluation

* Classification algorithms with iterative hyper-parameter tuning: Decision Trees, Random Forest, Gradient Boost, Logistic Regression - multinomial, SVM
* Evaluation metrics: Need to define the metrics to be used, possibly one/more of these: Accuracy, Precision, Recall, F1 score, PR AUC,  ROC AUC
* Feature Importance: Start with permutation feature importance to measure the contribution of each feature. Based on importance, rank the feature that contributes the most to the model and re-evaluate the model with subset of these features to cross check the selection.
* Dashboards: As time permits to present results

## Expected Results

This multi-class classification model assessing both departure and arrival delays is expected categorize flights into distinct classes based on relationships between departure and arrival times. This will enable a granular understanding of how delays in departure affect arrival times and overall travel experiences.

* + **On-time Departure and Arrival:** Flights that depart and arrive within their scheduled times.
  + **Delayed Departure, On-time Arrival:** Flights that experience delays during departure but still arrive on time or within a minimal delay window (e.g., less than 15 minutes).
  + **Delayed Departure and Arrival:** Flights that experience delays both in departure and arrival, with arrival delays exceeding a specified threshold (e.g., more than 15 minutes).

## Why is this question important?

Understanding and predicting flight delays through a multiclass classification model that assesses both departure and arrival delays is important for several reasons:

### **1**. Enhanced Operational Efficiency

* + **Resource Allocation:** By classifying flights based on delay categories, airlines can better allocate resources, such as ground crew and maintenance staff, to address potential issues proactively.
  + **Scheduling:** Accurate predictions help in optimizing flight schedules and turnaround times, reducing the risk of cascading delays and improving overall operational efficiency.

### 2. Improved Customer Experience

* + **Better Communication:** Knowing the likely classification of delays allows airlines to provide more accurate and timely information to passengers, improving customer satisfaction and trust.
  + **Compensation and Support:** Understanding delay patterns helps airlines plan better compensation strategies and offer appropriate support to passengers, such as meal vouchers or rebooking options.

### 3. Data-Driven Decision Making

* + **Strategic Planning:** Insights from delay classifications can inform strategic decisions, such as route adjustments, flight frequency changes, and investment in infrastructure improvements.
  + **Performance Monitoring:** Tracking the frequency and severity of different delay types allows airlines to monitor performance and implement corrective actions to minimize delays.

### 4. Operational Resilience

* + **Contingency Planning:** Identifying patterns in delays helps airlines develop contingency plans for various scenarios, such as adverse weather conditions or high traffic volumes.
  + **Crisis Management:** Accurate delay predictions can improve the airline’s ability to manage crises, such as unexpected disruptions, by providing a clearer understanding of potential impacts on operations.

### 5. Financial Impact

* + **Cost Management:** By reducing the incidence and impact of delays, airlines can lower costs associated with operational disruptions, compensation claims, and customer dissatisfaction.
  + **Revenue Optimization:** Minimizing delays helps in maximizing revenue opportunities, as timely operations lead to higher customer satisfaction and better utilization of aircraft and staff.

### 6. Safety and Compliance

* + **Regulatory Compliance:** Airlines need to comply with various regulations related to on-time performance and passenger treatment. Accurate delay predictions help ensure compliance with these regulations.
  + **Safety Considerations:** Understanding delay patterns helps in assessing and mitigating safety risks associated with tight turnaround times and high passenger volumes.

### 7. Competitive Advantage

* + **Market Positioning:** Airlines that can consistently provide on-time performance and manage delays effectively have a competitive edge in the market. This can be a key differentiator in customer choice and loyalty.

### 8. Route Optimization

* + **Efficient Routing:** Accurate delay predictions and classifications help airlines optimize flight routes based on historical delay patterns and current conditions, leading to more efficient flight planning.
  + **Dynamic Adjustments:** Airlines can adjust routes dynamically to avoid known delay-prone areas or times, reducing overall delay risks and improving operational efficiency.
  + **Network Planning:** Understanding delay patterns helps in designing more robust flight networks that minimize disruptions and ensure smoother connections between flights.

## Summary

By creating this multi-class classification model that captures the interplay between departure and arrival delays, airlines will gain a comprehensive understanding of flight punctuality. This granular view supports effective planning, enhances customer experience, improves operational efficiency, and contributes to better financial, safety, and competitive outcomes. It will also facilitate route optimization, enabling airlines to design more efficient flight plans and improve overall network performance.