CS 5805: Machine Learning I

Final Term Project Proposal

Student: Krishna Mattaparthi
Date: September 28, 2025
Dataset: US Flights 2023 with Meteo & Aircraft Data

Dataset Overview

This project will utilize the "US Flights 2023 with Meteo & Aircraft Data" dataset, a comprehensive, real-world collection of over 6.7 million US domestic flights from Kaggle. With direct applications in the aviation industry for operational planning and customer experience improvement, this multivariate dataset is ideal for analysis. It contains a rich mix of numerical variables (e.g., `ARRIVAL_DELAY`, `DISTANCE`) and categorical variables (e.g., `AIRLINE`, `ORIGIN_AIRPORT`).

Regression Analysis

The objective of this task is to develop a model that accurately predicts the arrival delay of a flight in minutes.

- **Dependent Variable:** `ARRIVAL_DELAY`. This continuous numerical feature is the primary target.
- **Independent Variables:** A comprehensive set of predictors will be used, including `DEPARTURE_DELAY`, `DISTANCE`, `SCHEDULED_TIME`, `AIRLINE_NAME`, `ORIGIN_STATE`, and `DAY_OF_WEEK`.
- **Methodology:** Data preparation and model selection are critical. Categorical variables such as `AIRLINE_NAME` and `ORIGIN_STATE` will be one-hot encoded to ensure compatibility with regression algorithms. Models ranging from a Linear Regression baseline to more complex ensembles like Random Forest or Gradient Boosting will be explored to maximize predictive accuracy.

Clustering and Classification

This task aims to classify airports into distinct performance tiers, providing a high-level "quality report" of the national aviation network. This will be a multi-class classification problem, and model performance will be assessed using accuracy, a confusion matrix, and other methods.

- **Dependent Variable:** The engineered `airport_performance_tier` (e.g., 'Tier 1: High Performance', 'Tier 2: Average', 'Tier 3: Low Performance') for now.
- **Independent Variables:** The aggregated airport statistics used for regression will also serve as the features for this model.
- **Methodology:** A two-stage, cluster-then-classify approach will be employed. First, an airport-level dataset will be engineered by aggregating flight data to calculate summary statistics for each unique airport (e.g., `average_arrival_delay`, `delay_variability`, `total_flights`, `cancellation_rate`). A K-Means clustering algorithm will be applied to this dataset to identify natural groupings of airports, creating the target variable: `airport_performance_tier` for now.

Association Rule Mining

The objective of this analysis is to uncover specific, human-readable conditional rules that describe the co-occurrence of flight characteristics and delay events. The accuracy and relevance of these rules will be measured by their support, confidence, and lift.

- **Methodology:** Each flight will be treated as a transaction. Numerical variables like `ARRIVAL_DELAY` have been discretized into categorical "items" (e.g., `delay_level='Low <5min'`). These, along with existing categorical features like `AIRLINE_NAME` and `ORIGIN_STATE`, will form the item sets for the appropriate algorithm.
- **Justification:** This technique provides a unique level of granular insight not available from other models. It can reveal high-confidence rules such as **`{AIRLINE=Spirit, DAY_OF_WEEK=Sunday, ORIGIN_STATE=FL} => {delay_level=High}`**. Such rules are valuable for operational planning, as they identify precise, multi-factor scenarios that frequently lead to disruptions.

Conclusion

This proposal outlines a comprehensive, three-pronged analytical approach to the 2023 US Flights dataset. By combining predictive regression, unsupervised clustering and classification, and pattern-based association rule mining, this project will move beyond a single analysis to provide a holistic understanding of flight delays. The planned methodologies will not only predict the magnitude of delays but also classify airport operational performance and uncover the specific, multi-factor conditions that lead to disruptions. The combined insights from these three analyses will provide a robust and actionable understanding of the dynamics of the US domestic flight network.