# Formal Proposal:
# Analyzing Airline Delays Based on Weather Data and Airport Metrics

SENG 474: Data Mining

University of Victoria

February 13, 2023

Group Members:
Nathaniel Roberts,
Chenuri Gunaratne,
Dylan Smith,
Harry Zhao,
Freya Engman,
Katelyn Dhami

## 1. Problem Statement

Weather-related flight delays and cancellations often occur at the last minute and because of this can be difficult to plan around. It could be helpful to predict these cancellations and delays ahead of time so that travelers, and airports, have advance warning and can arrange alternate plans if needed. Additionally, it will be interesting to see if basic weather data, such as precipitation and snowfall, and airport metrics, such as location and airline, are enough to accurately predict departure status.

After several years of restricted air travel, flyers are eager to travel in 2023. However, air travel is increasingly challenging and unpredictable for jetsetters. Cancellations are at their highest rate since 2014 and airlines report a rate of 20.1% in delayed domestic flights in 2022 within the continental US, which is an increase from 2021's rate of 11.4% [5]. While the economic climate and staff shortages can impact the reliability of flights, we want to examine the effects of weather and airline details to measure correlations between these features and cancellations or delays. With the rising cost of flights due to inflation, it is critical that buyers have the best possible information to make their flight decisions.

## 2. Goals

Our goal with this project is to analyse models that can predict flight delays and cancellations with at least 85% accuracy. Based on our research, we are expecting the decision tree to be the most accurate model [2]. We will use two supervised learning models, a decision tree and a neural network to generate a predictor for cancellations and delays, and one unsupervised model using a clustering algorithm to examine correlations between flight conditions and their outcomes. It will be interesting to see if certain locations, times of the year, or airlines result in more cancellations and delays. We hope the results can help inform travelers to make the best decisions on when and how to fly.

We will use clustering techniques to find similar groups within the data to better inform our conclusions about the problem. This can be useful to shed light on important features of the data which can affect the outcome of flight delays and cancellations that can be enhanced with deeper analytics. For example, the highest reported number of delays by airline are Allegiant Air, Frontier Airlines, and JetBlue with delay rates of over 30% [5]. Our goal is to determine the similarity in features, such as airline, between flights that are and are not delayed. The selection of an optimal clustering approach is subject to the intrinsic characteristics of the data. For instance, K-clustering may be advantageous in scenarios where the clusters exhibit a spherical topology, yet its efficacy may wane when applied to elliptical clusters. Conversely, hierarchical clustering can offer higher precision, regardless of the shapes of the clusters. However, this method is often constrained by its computational demands, which can pose challenges for clustering larger datasets. In spite of these limitations of clustering, it is a valuable tool for analyzing flight delay data and discovering underlying patterns and trends because it can efficiently and effectively identify groups of similar flights based on shared characteristics, providing insights into the factors contributing to flight delays.

Through the use of neural networks, we hope to find reasonable evidence that points to certain relations within the data set. The purpose of this is naturally to draw further information from the data used, specifically in terms of our overarching goal of predicting flight cancellations themselves. Ideally, we'll be able to find consistent patterns behind underlying relations between flights cancelled/delayed and instances such as poor weather conditions or particular times of the year. As previously mentioned in our initial goal statement, we will of course be aiming for the minimum accuracy of 85%.

## 3. Plan

We plan to use public flight and weather data to predict the chance of a flight being cancelled or delayed. However, we may narrow or increase the scope of the project based on difficulty. For example, the project could be simplified to a binary classifier that predicts whether a flight is or isn't cancelled, or it could be made more complex by attempting to predict the exact time a flight departs. We intend to use a dataset sourced from Kaggle, examining US continental flight data between 2009-2019 [1]. We will choose years to examine with the highest rate of cancellations and delays similar to current conditions. Good candidates for analysis are 2019 with a rate of delays of 18.9%, 2017 with a rate of delays of 19.1%, and 2015 with a rate of 19.2% [5]. The analysis will be broken into three parts. First the data needs to be prepared for each type of model. Second, the models will need to be tuned to the data to ensure the best performance of the model. Third, we will analyze the results to form conclusions about the reasons behind cancellations and delays.

To complete analysis based on a clustering model, the data will need to be preprocessed so that the dataset examples can be properly segmented. We will convert any categorical features into numerical features and normalize the data so that each feature lies on the same scale. To ensure the best results possible, multiple clustering techniques will be tested, including k-means clustering, hierarchical clustering, and density-based clustering. One method will be chosen as the most appropriate model for the dataset, on which further analysis will be performed. To see patterns in the data we will examine if natural groupings occur based on the length of delay observed across flights. Based on the results of cluster modeling, we may be able to better inform how to structure the supervised learning models included in this project.

Neural networks possess similar data preprocessing concerns as the clustering model described above. Different methods of preprocessing will be attempted, such as ensuring the maximum and minimum values of a feature lie between 0 and 1, or shifting the values of each feature such that their standard deviation is 1 and the mean is 0. Like clustering, we will experiment with multiple approaches to ensure the best results. The current starting point will be a multilayer perceptron with a single hidden layer, but this will change in the future as the results and problems from the initial strategy are observed.

Aiming at the flight delays and cancellation problems within the continental US, we propose to construct a weather-related flight delays prediction model based on classification regression decision tree algorithm and random forest model. With the training results of Logistic regression algorithm, decision tree algorithm and the fitting results, we may conclude that this method is able to deal with the data with massive features and reduce the possibility of over-fitting.

## 4. Task Breakdown

As previously stated, we have an exciting plan in place to tackle the challenging task of analysing flight delay data using three different machine learning algorithms. Our team will work in pairs to train and develop a decision tree model, a neural network model, and clustering models. Each pair will be responsible for conducting in-depth analyses of the models they develop and draw valuable insights into the reasons behind flight cancellations and delays. With this approach, we aim to provide reliable predictions and a deeper understanding of the factors that contribute to flight disruptions, potentially paving the way for more effective strategies to reduce delays and cancellations.

The division of this workload, and other individual tasks are as follows. Harry Zhao and Katelyn Dhami will conduct the work to train a decision tree and form predictions based on that decision tree. Nathaniel Roberts and Dylan Smith will be responsible for training and analysing the neural network model. Freya Engman and Chenuri Gunaratne will use clustering methods to find patterns in the data.

| **Date** | **Group Members** | **Description** |
|---|---|---|
| Feb 13 | All | Formal Proposal |

| | | |
|---|---|---|
| Mar 1 | Harry Zhao<br>Katelyn Dhami | Complete model research and data preparation for decision tree. |
| | Nathaniel Roberts<br>Dylan Smith | Complete model research and data preparation for neural network. |
| Mar 1 | Freya Engman<br>Chenuri Gunaratne | Complete model research and data preparation for cluster model. |
| Mar 10 | Harry Zhao<br>Katelyn Dhami | Complete decision tree parameter tuning and training |
| | Nathaniel Roberts<br>Dylan Smith | Complete neural network parameter tuning and training |
| Mar 10 | Freya Engman<br>Chenuri Gunaratne | Complete cluster model parameter tuning and training |
| Mar 13 | All | Progress Report |
| Mar 27 | Harry Zhao<br>Katelyn Dhami | Complete decision tree analysis |
| | Nathaniel Roberts<br>Dylan Smith | Complete neural network analysis |
| Mar 27 | Freya Engman<br>Chenuri Gunaratne | Complete cluster model analysis |
| Mar 31 – Apr 5 | All | Final Presentation |
| Apr 10 | All | Final Report |

## 5. References

[1] 'Airline Delay and Cancellation Data, 2009-2018'. Kaggle https://www.kaggle.com/datasets /yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018 (Accessed Feb 3, 2023)

[2] Tang, Y. (2022). Airline flight delay prediction using machine learning models. *Association for Computing Machinery*. https://doi.org/10.1145/3497701.3497725

[3] Wadkins, J. (2019). 2019 Airline delays w/weather and airport detail. Kaggle https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations/code?select=full_data_flightdelay.csv (Accessed Feb 3, 2023)

[4] Herbas, J. (2020). Using Machine Learning to Predict Flight Delays. Analytics Vidhya https://medium.com/analytics-vidhya/using-machine-learning-to-predict-flight-delays-e8a50b0bb64c

[5] Adams, Kurt. (2022) 2022 Has Brought More Air Travel Delays and Cancellations. LendngTree. 2022 Has Brought More Air Travel Delays and Cancellations — And Nearly Double the Risk of Having a Bag Mishandled - ValuePenguin Accessed Feb 12, 2023.