Machine Learning Project Proposal

by Dorian Buijse and Konrad Krawczyk

1. Data & Usage:

The dataset to be analysed consists of data about all domestic flights in the United States for the whole year 2015, for all airports and airlines. The data was collected and published by the DOT's Bureau of Transportation Statistics. The raw data file, with appropriate documentation, can be found on: https://www.kaggle.com/usdot/flight-delays.

The whole set has over 5,400,000 examples. By default, every example has 31 features, although not all information is complete for all examples. It is nevertheless an extremely large and comprehensive dataset, which allows for very accurate statistical inferences. For the sake of better performance, we shrunk the dataset by a hundred times, leaving 54,000 examples - still large enough, but far more manageable.

The primary use case of the algorithm will be: **predicting a potential delay, on a given day, for a given airport and airline.**

2. Features and Target:

We found several features that could be useful for data analysis and prediction algorithm:

Feature (Key)	Data type	Significance
Day and month (DAY, MONTH)	Numeric	Depending on the time of the year, different factors (weather, holidays, other events) could affect the traffic across all airports
Day of the week (DAY_OF_WEEK)	Numeric (counted from Monday onwards)	Trends can also unfold for particular days of the week (over the weekend e.g.)
Airline identifier (AIRLINE)	String (two-letter identifier)	Using the prediction algorithm, a user can compare airlines' punctuality which fosters informed consumer decision-making
Departure airport (ORIGIN_AIRPORT)	String (IATA code)	Creating a prediction model like this would potentially help in identifying airports that are poorly organized; also in estimating real travel time and minimizing the risk of missed connections
Destination time (SCHEDULED_DEPARTURE)	Numeric (format: HHMM)	At different times of the day traffic blocks are more likely to happen (for example during the night)

Target: total delay on departure (DEPARTURE_DELAY)	Boolean (true if delay in minutes > 5, else: false)	The value that will be predicted based on the above features

3. Algorithms (tentative):

- Decision trees
- Support Vector Machine
- Neural Network

First, the data will be split into training and test sets (test size: 0.3). The above algorithms will be trained using sklearn packages, and optimized (respectively) for number of splits, C-value and number of layers, using a cross-validation set. Confusion matrices and precision/recall/f1 tables for the test data will be generated, and based on accuracy and f1 values, the most accurate predictor of whether a flight is delayed or not will be picked.