**Flight Delay Prediction**

**Abstract**

The goal of this project is to use the 'DecisionTreeClassifier' model from sklearn library to predict the flight delay in January month for US flights, and to know what destination or aircraft that causes more delays.

**Design**

The project is done for learning purpose and could be enhanced with more data to predict more information, to prevent future delays that could be avoid. The data is provided by (kaggle.com) collected from the Bureau of Transportation Statistics, This data is open-sourced, its unbalanced data since most of flights are on time, so we need more features to be more accurate.

**Data**

Dataset contains 583986 flights information for January 2019 and 607030 flights for January 2020, with 20 features for each, 16 are categorical. And 4 features are numrical like departure time, arrival time, departure time block and distance. Other features could be grouped into one category, for example 'Carrier code' and 'Airline ID. \*\* data included in the data folder as CSV file.

**Algorithms**

Feature Engineering

1. Decide the features that will affect the delay.
2. Give weights to the categorical features.
3. Converting the categorical features to numerical values

Models

Logistic regression and Decision tree classifier

**Tools**

* Numpy and Pandas for data manipulation
* Scikit-learn for modeling
* Pandas\_profiling for plotting
* Seaborn for plotting

**Communication**

Slides will be uploaded to my github