# Flight Delay Prediction

## Execution steps:

1. Upload the following dataset to the google drive:

   https://www.kaggle.com/datasets/giovamata/airlinedelaycauses

2. Open the following Colab link and mount the google drive. Then, provide the dataset location in the cell where spark reads the csv data.

   https://colab.research.google.com/drive/1W-6uqRekqods8xcOpH8_OO_68t7SoS7q?usp=sharing

3. This program stores the processed data back in the drive. Note the folder/file path.
4. Now, Open the following to run the Linear Regression Model by replacing the read csv path with the processed data folder path.

   https://colab.research.google.com/drive/1aSrda8tZLgr1p8S5CZkyN2gHpuSReBYe?usp=sharing

5. Next, Open the following to run the Random Forest Model by replacing the read csv path with the processed data folder path.

   https://colab.research.google.com/drive/19YFiUIzoHomoKhfheKIxanqJX4fnWuhk?usp=sharing

6. Then, Open the following to run the Gradient Boosted Tree Model by replacing the read csv path with the processed data folder path.

   https://colab.research.google.com/drive/1wD21hiwZUSAfN-Stzk7rYG6CvXmlr7ai?usp=sharing

7. Finally, you can also execute the following code to evaluate Gradient Boosted Tree Model for various system metrics.

   https://colab.research.google.com/drive/14XqcBWC61ArfnK6AfNS-eydqS77_pVCC?usp=sharing

8. Test data will be run automatically for each model and graph is displayed at end.

Note:  We used multiple colab notebooks as it would be easy to evaluate each model individually.