ML Project

Problem: Model a two stage predictive machine learning engine that forecasts the on-time performance of flight from the given datasets.

Module 1 (DATA PREPROCESSING): The weather folder consists of the weather data at 15 different airports in the USA between 2016 and 2017. The flight data of all the flights that flew inside the USA is provided. Merge these two data appropriately such that each record of the flight should have the corresponding weather data available.

Preprocess the data only for the airports mentioned below for the years 2016 and 2017 only.

Airport : codes

ATL	CLT	DEN	DFW	EWR
IAH	JFK	LAS	LAX	MCO
MIA	ORD	PHX	SEA	SFO

Recommended Weather features to consider:

WindSpeedKmph	WindDirDegree	WeatherCode	precipMM
Visibilty	Pressure	Cloudcover	DewPointF
WindGustKmph	tempF	WindChillF	Humidity
date	time	airport	

Recommended Flight features to consider:

FlightDate	Quarter	Year	Month
DayofMonth	DepTime	DepDel15	CRSDepTime
DepDelayMinutes	OriginAirportID	DestAirportID	ArrTime
CRSArrTime	ArrDel15	ArrDelayMinutes	

Using Google Colab/Jupyter notebook is highly encouraged as it makes it easier for us to help you out.
To install all the packages you will be requiring to do ML, visit [here](#).

Link to flight data: [here](#)
Link to weather data: [here](#)
Workbooks for numpy, pandas and json: [here](#)
Lookups: [Dask](#) ,[Numpy](#), [Pandas](#), [datetime](#), [os](#), [json](#)

Deadline: 20 days

_____

Module 2 (CLASSIFICATION): Model a classifier that classifies whether a given flight will be delayed or not using the features.

Lookup : [Logistic Regression](#), [Decision trees](#), [SVM](#), [Extra Trees Classifier](#), [XGBoost](#)

Deadline : 15 days

_____

Module 3 (REGRESSION): Off those flights that were classified to be delayed, train a regression engine which predicts the Arrival delay period (in minutes).

Lookup : [Linear Regression](#), [SVR](#), [Extra Trees Regressor](#), [XGBoost Regressor](#).

Deadline : 15 days

_____

Please contact the TAs if any clarification is required.