# Final Project

## Data Science pipeline on-premises and on-the-cloud

Due on Sunday, 1 November (13:00)

### Objective

In this assignment, you will be provided with a real-world dataset, and you are required to implement the whole pipeline of building the data science pipeline on-premises and on-the-cloud. This includes understanding the business problem, preparing data, exploring the data, performing feature engineering, building, and deploying models.

### Introducing the business scenario

You work for a travel booking website that wants to improve the customer experience for flights that were delayed. The company wants to create a service to let the customers know the likelihood of the flight being delayed based on the weather conditions before they book the flight to or from the busiest airports in US.

You are tasked for solving parts of this problem by using machine learning (ML) to identify how likely the flight will be delayed based on the available weather data. You have been given access to the dataset of the on-time performance of the domestic flights that are operated by large air carriers. You can use this data to train a ML model to *predict if the flight will be delayed for the busiest airports*.

### About the dataset

The provided dataset contains scheduled and actual departure and arrival times reported by certified US air carriers that account for at least 1 percent of domestic scheduled passenger revenues. The data was collected by the Office of Airline Information, Bureau of Transportation Statistics (BTS). The dataset contains date, time, origin, destination, airline, distance, and delay status of flights for flights between 2014 and 2018.

The data are in 60 compressed files, where each file contains a CSV for the flight details in a month for the five years (from 2014 - 2018). The data can be downloaded from this link: [compressed data].

**Features of the dataset**

Dataset(s) used in this assignment were compiled by the Office of Airline Information, Bureau of Transportation Statistics (BTS), Airline On-Time Performance Data, available with the following link: [dataset attributes].

---

## Tasks

The tasks of this assignment are divided in two parts as follows:

---

**Part A –Data Science on-premises (60 marks)**

In this part you are expected to:

- Understand the dataset and describe the business problem;
- Document an exploratory data analysis and whenever possible draw conclusions about the analysis;
- Employ popular graphical modules (matplotlib, seaborn or tableau) to answer questions;
- Implement machine learning techniques to predict whether the flights will be delayed or not.

**You are given a Jupyter notebook named "onpremises.ipynb", which contains a starter code and instructions to go ahead with this part. You are required to answer the questions in this notebook and upload it as a response to this part.**

---

**Part B –Data Science on-cloud** **(40 marks)**

In this part you are expected to:

- Use your skills in performing the machine learning pipeline using the Amazon SageMaker.
- Compare the results of implementing the ML pipeline on premises versus on the cloud.

**You are given a Jupyter notebook named "oncloud.ipynb", which contains a starter code and instructions to go ahead with this part. You are required to answer the questions in this notebook and upload it as a response to this part.**

---

## Deliverables

You are required to submit a compressed (e.g. ZIP) file to the Canvas website of the unit with the following files:

1- A Python Jupyter Notebook with the code for parts A
2- A Python Jupyter Notebook with the code for parts B
3- [Optional] A PDF document with your reflection on the unit highlighting what you liked and what you didn't.

---

Good Luck 😊

Ibrahim Radwan