IDC4251C Module 9 Project: Logistic Regression with PowerBI

In this assignment we will explore the use of the Python programming language to manipulate data which we can use for exploratory data analysis and visualization in Power BI.

The original analysis is described in the document "Predicting Titanic Survivability Using Logistic Regression".

Python is a popular programming language used in many industries, commonly used for data analytics and machine learning. Many libraries are available which can simplify complex tasks, including libraries for data manipulation and training machine learning applications.

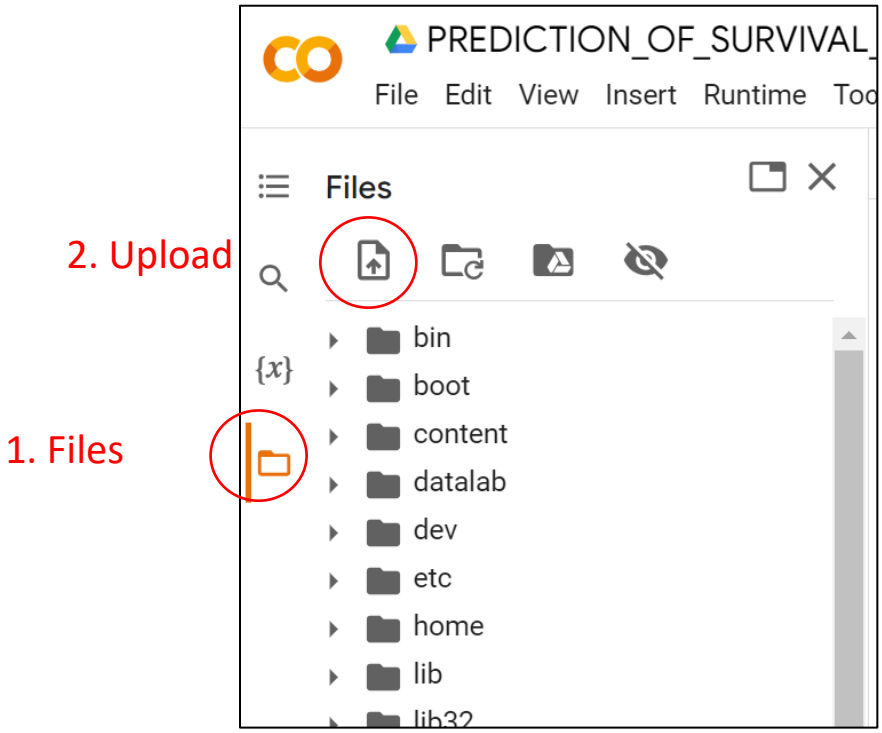
You do not need to be familiar with Python to complete this program, we will use the Jupyter Notebook application which will lead you step by step through the process of training a model using logistic regression.

You will need a Google account to access Google CoLab which will allow you to set up the Jupyter notebook.

The GitHub classroom repo contains the document mentioned above, a training data file (train.csv), a test data file (test.csv), the Jupyter notebook (.ipynb) file (created by the analysis author), and an image by the analysis author containing sample visualizations you can use as a reference for your own analysis.

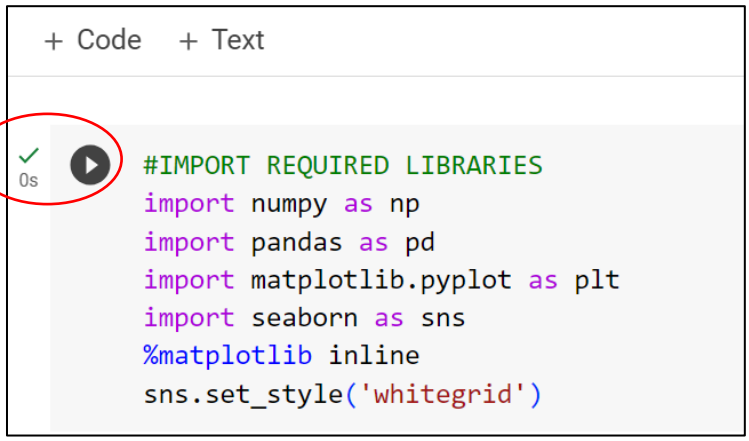
When you first connect to Google CoLab you will be shown a dialog to initiate an upload of the Jupyter Notebook file from your repo; select Upload and then use the Choose File to find your .ipnyb file:

After loading the notebook, Upload the training and test data to the notebook by selecting the Files icon on the top left of the page and then selecting the Upload icon:



Use the scrollbar on the right to scroll to the bottom of the files pane to verify the files have been uploaded.

After uploading the data files, follow the notebook steps by executing each code cell in the order in which they appear. Press the "Play" arrow to execute a cell, when it has successfully executed a green checkmark will appear. Here is the first cell in the notebook after executing, the Play arrow and checkmark are highlighted:



After executing the last cell in the notebook, a new .csv file will be generated which contains the calculated survival data from the original test dataset. Download this file and use it in combination with the original test data file to create a relationship to display your data.

You do not need to reproduce the visuals from the reference image, this assignment will allow you to use your own creativity to produce an informative report. Be sure to include a title on the report as well as informative labels for the visuals.

Submit your Power BI workbook and results.csv file to your GitHub classroom repo.