Máster en Big Data. Tecnología y Analítica Avanzada (MBD).

Fundamentos Matemáticos del Análisis de Datos (FMAD). 2022-2023.

Final Course Project: Teamwork

- The goal of the Final Course Project is to test all the Data Analysis skills and Python tools that we have met in the course. In order to that you should write a report, in the form of a Jupyter notebook, describing the analysis of a data set. See below for more details about the choice of a data set.
- Keep it relevant! Using Jupyter it is easy to make reports with extension equivalent to 100 or more printed pages. Make sure that the information you provide in your report is useful, do not include uncommented figures, long data tables, etc.

Main Goal of the Project

- The main goal of this project is to give you the opportunity to showcase the Exploratory Data Analysis (EDA) and Python skills that you have been practicing in this course. That is, you should select a *rich enough* data set so that you can use it to (at least) do the following:
  - Describe and visualize all common types of variables (qualitative, discrete, continuous).
  - In particular make sure that you include a rich set of graphs, covering all the standard types of graphs (bar plots, histograms, density curves, bar plots, scatter plots to name the most relevant). Bot remember: do not include a graph without including a comment about the information it provides!
  - Make some inference about the variables in the data set: get some confidence intervals, test hypothesis, etc.
  - Analyze the possible relations between variables in you data. The kind of analysis we look for is exploratory or based in elementary linear or logistic models. There is no need for more complicated modeling, you will have plenty of that in Machine Learning.
  - Some amount of *Data Wrangling* (dealing with missing data, outliers, untidy data sets, etc.) is highly encouraged and will be taken into account for grading, but keep in mind that it is easy to get lost in this! Keep it under control and talk to us when in doubt.

Suggested Data Sets for the Project

- **Kaggle Data Sets**. Kaggle is one of the most popular websites in the Data Science and Machine Learning communities. From the main page use the *DataSets* link on top and use the filters in the search field to narrow down the search. We suggest some popular data sets but feel free to explore.
- The **UC Irvine Machine Learning Repository** contains a smaller collection of data sets, but they are very popular, widely used and in general good quality.
- **Open Data Portals from Institutions and Organizations** Such as the European Union, Spanish Government or Parliament, Regional or Local Institutions. Also many NGOs, newspapers, etc. Many of these resources contain a huge collection of data sets and the main problem is usually filtering out an interesting one. If you run into trouble or if you have doubts about whether a particular data set is adequate for this project please let us know.
- Remember in any case that you must clearly identify and cite the source of your data.

Submission and Deadlines

- The project must be submitted as a zip or similarly compressed file containing a Jupyter notebook (or set of notebooks) and all of the auxiliary files that your code requires (data sets must be included or linked for download).
- In any case there must be a main Jupyter notebook called *Final_Project_Main* which contains a section with a description of the project. If you decide to keep all your code in a file then this will be your only notebook, but if you use more than a notebook then the role they play and the order in which they should be read must be clearly described in the main notebook. Similarly if you have any other kind of code files outside notebooks.
- If you use any Python. module besides those that we have seen in class then you must clearly state that in the first cells of the main notebook. Do not include setup commands for these modules in your notebooks.
- **Deadlines:**
  - We should receive an email with your **project proposal** (choice of dataset and a basic sketch of what you aim to do) **no later than Friday 23rd September**.
  - The final version of the project should be submitted **no later than Friday 11th November**