**Introduction to Machine Learning 2024: project first part**

The objective of the first part of the project is to develop a basic Machine Learning pipeline, applying all the techniques that we introduced during the first part of the course. You will be provided with a simple dataset and you will have to perform a series of tasks on the said dataset. This document serves as a guide to understanding the different steps of the project, what to do, and what to deliver.

**Data**

You will be working on a partition of the FolkTables dataset, fully contained in the ‘*pums.csv’* file. You will find 15 variables and 1 boolean target variable. This is a modified portion of the original dataset, covering three states for a period of 1 year. The full documentation and repository can be found here: <https://github.com/socialfoundations/folktables>. Further documentation to understand the semantics and values of variables can be found on the official site of the American Community Survey Public Use Microdata Sample: <https://www.census.gov/programs-surveys/acs/microdata/documentation.html>. In particular pay attention to the data dictionary info: <https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2022.pdf>.

**Objectives**

complete the first part of the project you will need to solve the following tasks:

1. **Data exploration:** explore the data, specifically understand variable distributions, data types, and correlations. Sensibly manage null values.
2. **Clustering:** apply unsupervised learning techniques to the variables, i.e., removing the target variable. Estimate parameters, compare methods with sound evaluation, and report results.
3. **Classification:** apply supervised learning techniques, targeting the Boolean variable. Estimate parameters, compare methods with sound evaluation, and report results.

All the tasks must be solved using the Python programming language. You can provide notebook files and/or .py files with functions and code. *Everything* has to be properly commented on and documented. Prepare a small, *3 to 5 pages* report detailing the most interesting choices and/or findings. Implementation details can be commented on just in the code. More general considerations and explanations, with plots, must be documented in the report. *Do not* exceed the indicated page limit for your report. Try to be concise and to the point.

**Organization**  
Please, sign in the group form at: <https://docs.google.com/spreadsheets/d/189sqJOsWEIBTXOqpXk8Dds4dkPc9WCk4YY-gqmS0hw4/edit?usp=sharing>. Here, we’ve reported the names and contacts of students who provided this information at the beginning of the course. Organize in groups of 2 people at most. If you need to create a group larger than 2, contact the teachers *first.*

Deliver your project into a folder with *the same name as your group* in: <https://drive.google.com/drive/folders/11fRcxR0zf0PzWBN7KCgCnR3ZrUZMObQ0?usp=sharing>.

**Deadline is: 31st of May.**