Green AI Project II: Getting some data analysis done

In this class we will begin to perform statistical analysis on the collected data. This step is of crucial importance in all ML projects as it helps us understand the data and the phenomenons we are working with.

- We will be using python for the data analysis (R is accepted as well if needed)
- The main libraries we are going to use are pandas and matplotlib
- If your type of data is not tabular (ex: text, images..) the teacher will directly guide you on the steps to take

Question 1.

Finish your data collection if it's not done. you can search Arxiv for topics that are related to your project.

Question 2.

Using Jupyter notebook load your data through pandas. Is the data the same as in the documentation? are the column names correct and understandable for your team? Make sure you understand the nature of the data you're working with.

Question 3.

Lets clean a bit your data. Make sure the following is correct:

- All the columns are assigned their correct type?
- Are missing values correctly identified?
- Aberrant values are identified?
- If you have categorical variables, i there any with too many modes?

Question 4.

Let's start analysing! Identify the main variables:

- What is the target variable that you will use to predict?
- Are there variables that ”just make sense” that they would impact our target? why?
- Compute the correlation matrix. are there variables that are very correlated/anti correlated ?

Question 5.

Lets make a few plots:

- plot the distribution of all the main variables
- Make the pairwise scatter plots. Is there any that seems interesting?
- If you have temporal data, don't hesitate to plot the relevant time series instead of the scatter plots.

Question 6.

Your own project might have specific plots or metrics that are relevant to compute. Work with the teacher to identify them and plot them