Explainable AI

First level of explicabity: statistics

In this class we will gather INSEE Data, and start analyzing it in rder to have a bigger understanding on the underlying mechanisms that have an effect on salary

- 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
- you can find the necessary data here: https://www.insee.fr/fr/statistiques/7651654

Question 1.

Download and extract the data. There should be two files, the data itself and a documentation file.

Question 2.

Using Jupyter notebook load your data through pandas. Are you able to understand the variables? With the help of the documentation choose 5-10 variables that you think that the salary might be related to. (ex: gender, type of contract...)

Question 3.
Lets clean a bit the data. With the documentation file, replace the codes by their values. For salary, replace it with the lower part of the bracket.

Question 4.

Let's start analysing! Keep in mind that the salary is the main variable we want to understand.

- Compute the correlation matrix for numerical variables. Are there variables that are very correlated/anti correlated ? Do they make sense?
- Look at the average salary grouped by the categorical variables. Any variable look like it has a big effect? Does it look logical?

Question 5.
Lets make a few plots to illustrate what we have:

- Make the pairwise scatter plots between the salary and other numerical variables. Is there any that seems interesting?
- Plot the salary box-plot for each category of your categorical variables. Are there shocking things?

Question 6.
Let's create a linear model.

- Import the linear regressor from the sklearn library.
- Split your data in train/test
- Train your model
- Evaluate the r2 score and the RSME (Root squared mean error). What can you conclude?
- Look at the parameters of your regression. which ones are the most important? How do you interpret them?

Question 7.
Let's improve the model. Usually taking the log of the salary yields better results. Let's try it. How do the interpretation of the parameters change?