Explainable AI

First level of explicabity: statistics

In this class we will use data from the new york housing market in order to understand what are the drivers of it's price

- 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, matplotlib ans sklearn
- you can find the necessary data here: https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market
- This TD will be graded, please submit it one week after the end of the second TD

Question 1.

Download and extract the data. Using a notebook load your data through pandas and start exploring the dataset.

- How many apartments are there?
- What are the characteristics of the most expensive apartment?
- Transform the size of the apartment to m2. Which are more expensive per square meters, big apartments or small apartments?
- what is the distribution of the prices for the main brokers? Can you find luxury brokers?
- what is the most expensive area? the cheapest?
- what area has the most variance in prices? and in price per square meter?

Question 2.

Make a few plots to illustrate the previous questions and deepen your knowlege of the dataset

Question 3.

We are going to do a quick linear regression.

- Select the variables that are the most likely to have an effect
- Do one hot encoding to categorical variables, and select their base.
- with scikit learn, use the linear regressor to predict the price of the apartment.
- use .coef_ and .intercept_ to access the parameters of the model. What can you say about them?
- Transform the price with the log, fit your new model and interpret the parameters.

Question 4.

Now lets work on a different model. We will use a decision tree.
Can you describe how a decision tree works?

- Use the scikitlearn library and fit a decision tree to predict the price of the apartment.
- Plot the decision tree. Is there anything surprising?
- Split your data on train and test sets and try to find the best hyper-parameters to fit the model. Is the tree different?
- Wich model yields better results, the linear regression or the decision tree?

Question 5.

We are going to use a random forest algorithm now.

- Use the scikitlearn library and fit a random forest on a training set.
- Which model yields better results? (don't forget to tune the hyperparameters)
- look at the variable weights. Which variables are more important? How can you interpret that?