Explainable AI

TD2 :Linear regression and trees

In this class we will use data from a cardiology study to try to asses the impact of cholesterol on the probability of having a heart attack.

- 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://archive.ics.uci.edu/dataset/45/heart+disease
- This TD will be graded, please submit it one week after the end of the second TD (3hours)

Question 1.

Download and extract the data. Using a notebook load your data through pandas and start exploring the dataset. Get your data from the "Import in python" button from the download page

- Describe the different variables and find their definition in the documentation
- For each variable make an educated guess on the effect that you expect that variable to have
- Check with a quick statistic if the guess was correct and state if your guess was correct
- Clean the variables removing or imputing NAs

Question 2.
Make a few plots to get a better understanding of the data.

Question 3.
We are going to do a quick logistic regression with two classes, target 0 from one side (healthy) and target 1 to 4 on the other side (sick)

- Do one hot encoding to categorical variables, and select their base.
- with scikit learn, use the logsitic regressor to determine the probability.
- What are the performance of your model? which metrics can you use?
- use .coef_ and .intercept_ to access the parameters of the model. What can you say about them?
- How the interpretation of these parameters is different from the interpretation of the coefficient of a linear regression?
- Which are the most important variables? Does it make sense? at the parameter for cholesterol. How does it compare to the statistic computed on the previous question?

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 probability of sickness.
- 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?
- Which model yields better results, the linear regression or the decision tree?
- explain how the decision tree weights are computed

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?