Machine Learning module – Python Lab – Exam 19/06/2020

Find clusters for the included dataset.
The solution must be produced as a Python Notebook.
The notebook must include appropriate comments and must operate as follows:

1. load the data into a dataframe **df**, show its size and head, eliminate the rows containing null values and show the number of remaining rows (2pt)
2. produce a pairplot of the numeric columns of **df** and comment relevant situations (2pt)
3. Produce a box plot of the numeric columns of **df** and comment relevant situations (2pt)
4. Produce the correlation matrix of the data and eliminate the redundant attributes, if it is adequate (4pt)
    - For example, if attributes **a** and **b** have high correlation (e.g. absolute value higher than 0.95) one of the two can be eliminated
    - Refer to this https://stackoverflow.com/questions/29432629/plot-correlation-matrix-using-pandas for the generation of the correlation matrix
5. Split the reduced data: store the first column in a vector **keys** and the others in a matrix **X** (2pt)
6. Find the best clustering scheme for the data (possibly reduced after step 4) with a method of your choice, plot global silhouette index for an appropriate range of hyperparameter(s) and show the chosen hyperparameter(s) (4pt)
7. Fit the clustering scheme to **y**, then produce the silhouette plot using the function plot_silhouette contained in the attached file (4pt)
8. Perform a logarithmic transformation of the data (4pt)
    - This means simply to apply the *log* function of *numpy*
    - If one of the columns has zero or negative values, avoid its transformation
9. repeat points 6 and 7 above and comment the comparison with the result of point 6 (2pt)

Quality of the code (6pt):

- Include appropriate comments with reference to the numbered requirements
- Useless cells, pieces of code and non-required output will be penalized
    - Remove the code you use for testing and inspecting the variables during the development
- Naming style of variables must be uniform and in English
- Bad indentation and messy code will be penalized

Additional directions, the assignments not compliant with the rules below will not be considered

1. The notebook name must be ***lastname.ipynb*** in lowercase letters
2. The first cell must contain the student last name
3. The solution must directly access the data in the same folder of the notebook

Cooperative work will be **heavily sanctioned**
The candidate can freely access the manuals available on line
The candidate can freely access the teaching materials available in the course website, including the available examples of python notebooks.
The notebook must be zip-compressed and then uploaded in EoL