Machine Learning module – Python Lab – Exam 15/01/2021

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 and separate in X all the columns but the last one, in y the last column, then produce a pairplot of X and comment what you see (4pt)
2. find the best clustering scheme for X with a method of your choice, plot the silhouhette index for an appropriate range of parameters and show the chosen hyperparameter(s) (4pt)
    1. consider carefully the number of clusters, simple optimisation of the silhouette will not be enough, decide visually the best number of clusters
3. fit the clustering scheme store the cluster labels in y_km and output the silhouette score (2pt)
4. use the labels in the last column of the input file as the "gold standard" for the clustering and compare y_km and y; for an effective comparison, each label in y_km must be remapped to the best label in y; compute and apply this re-mapping (5pt)
    1. hint for each subset of the data with x in y_km find the most frequent label in y
5. produce the confusion matrix comparing y and y_km with sklearn.metrics.confusion_matrix, (2pt)
6. consider possible pre-processing actions, repeat the fitting and evaluate as before the result of the new fitting (8pt)

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 ***emailusername.ipynb*** in lowercase letters
    a. E.G. if your email is mario.rossi45@studio.unibo.it the notebook filename will be mario.rossi45.ipynb
2. The first cell must contain the student first name, last name and email
3. The solution must directly access the data in the same folder of the notebook
4. Upload the notebook only to Virtuale

Cooperative work will be **heavily sanctioned**
The candidate can freely access any kind of materials