Exam for Machine Learning Python Lab

Consider the file provided with the assignment and execute the analysis described below according to the best practices of Machine Learning. You are allowed to use *only the computers of the lab, use the operating system Ubuntu*, you are not allowed to use any other device, email or any other messaging tool. You can use *only the websites accessible through the computers of the lab, as listed in the following page*.
Cooperative work will be heavily sanctioned

The notebook must operate as follows:

Group A

1. Load the data file and explore the data, showing size, data descriptions, data distributions with boxplot, pairplots ..... **1 pt**
2. Comment the exploration of step 1 pointing out if there are imbalanced distributions, outliers, missing values, non–numeric fields with number of distinct values similar to the number of records ..... **2 pt**
3. Drop the columns that are not relevant for the clustering operation, if any, and explain why you do that.
   Deal with missing values, if any ..... **4 pt**
4. find the best clustering scheme with KMeans, require not less than 3 clusters, show the hyperparameters, show the silhouette plots of clusters, show the distribution of the resulting cluster labels (e.g. histogram or pie plot) ..... **4 pt**
5. find the best clustering scheme with Agglomerative Clustering or DBSCAN (your choice), require not less than 3 clusters, show the hyperparameters, show the silhouette plots of clusters, show the distribution of the resulting cluster labels (e.g. histogram or pie plot) ..... **3 pt**
6. Comment the results ..... **2 pt**

   Total points for tasks 16

Quality of the code **4pt**

- Include appropriate comments with reference to the numbered requirements
- Useless cells, pieces of code and non-required output will be penalised
- 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 penalised
- Non generalised solution, such as three sequential statements with the same kind of operation instead of a loop, will be penalised

Total grade:20
Additional directions, the assignments not compliant with the rules below will not be considered:

- The notebook name must be `yourworkplace_youremailusername.ipynb` in lowercase letters
  E.G. if your worplace is `lab9_35` and your email is `mario.rossi45@studio.unibo.it`, the notebook filename will be `lab9_35_mario.rossi45.ipynb`
- The solution must directly access the data in the same folder of the notebook, the name of the file must be the same as the file provided.
- Upload the notebook only to `http://eol.unibo.it` in the activity specified by the teacher, any other way of submitting the notebook will be ignored

Allowed websites

- https://numpy.org
- https://scipy.org
- https://pandas.pydata.org
- https://matplotlib.org
- https://seaborn.pydata.org
- https://scikit-learn.org/stable