# 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*, 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:

1. Load the file `data.csv`, explore the data showing size and do some data exploration, the dataset is unsupervised ..... **1 pt**
2. Do the appropriate pre-processing in order to use the sklearn algorithms on this dataset; the values are qualitative and must be considered as nominal ..... **2 pt**
3. In this dataset the elbow method will show an almost "vanishing" elbow for inertia, and the silhouette is totally non-effective. As external background knowledge, we are told that here a requirement for a good clustering scheme is to have clusters with *balanced cluster sizes*, e.g. a scheme with cluster sizes (333, 667) is less acceptable than one with (333, 333, 334). In order to obtain this, we to compute, for each clustering scheme a *Balanced Partition Index* (BPI) computed from the cluster sizes with the function defined in the attached file. BPI values range from 0=worst index to 1=best index. For varying `n_clusters` fit KMeans and compute the inertia, the silhouette index, and the above mentioned BPI. ..... **4 pt**
4. Make two plots, one with inertia and silhouette, another with inertia and BPI, then decide the best number of clusters and refit KMeans using that value. ..... **2 pt**
5. Repeat the experiment using a different clustering algorithm of your choice, trying to generate a number of clusters near the one chosen in step 4 and show the results with the best hyperparameter values. **4 pt**
6. Comment the results of the two experiments. ..... **3 pt**

*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