# 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:

1. Load the data and explore them, showing size, structure and histograms of numeric data; show the histogram of the frequencies of the *target* labels, contained in the **"language"** column ..... **1pt**
2. Comment the exploration of step 1 pointing out if there are imbalanced distributions, outliers, missing values, features that seem not to be relevant ..... **1pt**
3. Drop the rows with NaN values, if any, show the shape of the dataset after this cleaning ..... **1pt**
4. *Model1:* tune the hyper–parameters of a classifier with Cross Validation, optimize for recall without considering the frequencies of class labels **4pt**
5. produce a classification report for Model1 ..... **1pt**
6. display the confusion matrix for Model1, normalised with respect to true values ..... **2pt**
7. *Model2:* tune the hyper–parameters of another classifier with Cross Validation, optimize for recall without considering the frequencies of class labels ..... **3pt**
8. produce a classification report for Model2 ..... **1pt**
9. display the confusion matrix for Model2, normalised with respect to true values ..... **1pt**
10. comment the comparison between the results of the two models .. **1pt**

*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