Exam for Machine Learning Python Lab

Consider `exam_superv.csv`, explore the data, exclude the column texttty and find the best clustering scheme and hyper parameters which can reproduce the column y. Then apply the clustering scheme found in the previous step to `exam_unsuperv.csv` and plot the data with the assigned labels
The solution must be produced as a Python Notebook, assuming that the dataset is in the same folder as the notebook.

The notebook must include appropriate comments and must operate as follows:

1. Load the `dataset1.csv` and explore the data, showing size, structure and histograms of numeric data; show the histogram of the frequencies of the class labels, contained in the "y" column **4pt**

2. drop the column "y" and find the best clustering scheme and hyper–parameters able to reproduce the y column (hint: before clustering you can consider dropping columns with little correlation to "y"; perhaps you should consider more than one estimator for clustering) **10pt**

3. show the difference between the original "y" column and the labels generated by the clustering, it can be expressed as "accuracy", produce also the confusion matrix **4pt**

4. apply the same transformations, to `dataset2.csv`, then apply the best clustering scheme and hyper parameters and plot the data with the assigned labels **6pt**

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 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

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

- The notebook name must be `youremailusername.ipynb` in lowercase letters (underscore instead of dot inside the email username can also be accepted
  E.G. if your email is `mario.rossi45@studio.unibo.it`, the notebook filename will be `mario.rossi45.ipynb` (`mario_rossi45.ipynb` can also be accepted)
- 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. If the notebook is developed using *Google Colab*, the code must be able to work also out of the Google Colab environment without any change.
- 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

Cooperative work will be heavily sanctioned

The candidate can freely access any kind of materials.