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 decide which pair of columns is most interesting for a 2d scatterplot, ad produce the scatterplot (5pt)
2. find the best clustering scheme for X with a method of your choice, plot ssd and global silhouhette index for an appropriate range of parameters and show the chosen hyperparameter(s) (5pt)
    1. consider carefully the number of clusters, simple optimisation of the silhouette will not be enough, consider also the elbow plot and decide visually the best number of clusters
3. fit the clustering scheme to y_km, then produce the confusion matrix comparing y and y_km with sklearn.metrics.confusion_matrix, the resulting confusion matrix must be "sorted" using the function max_diag provided below, producing the final confusion matrix cm_km (5pt)
4. in a comment explain why function max_diag is useful (2pt)
5. compute the accuracy a_km of y_km versus y as the ratio the sum of the main diagonal of cm_km and the number of samples in X (2pt)
6. rescale X using sklearn.preprocessing.MinMaxScaler, producing the scaled dataset X_mms (3pt)
7. repeat point 3 and 5 above, fitting X_mms to y_km_mms and producing the confusion matrix cm_km_mms reordered with max_diag and the accuracy a_km_mms as above (3pt)

Quality of the code (6pt):

1. The python cells must be preceded by appropriate comments
2. Useless cells and pieces of code will be penalised
3. Naming style of variables must be uniform and in English
4. Bad indentation and messy code will be penalised

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

1. The notebook name must be ***machineNumber_lastname_firstname.ipynb,*** the number must have three digits, with leading zeroes, if necessary
    1. for example, if I am sitting on the machine lab2, my notebook will be 002_sartori_claudio.ipynb
2. The first cell must contain the machine number, the last name and first name of the student.
3. The solution must directly access the data in the same folder of the notebook

Cooperative work will be **heavily sanctioned.**

The candidate can freely access the manuals available online in:

scikit-learn.org
docs.scipy.org
pandas.pydata.org
matplotlib.org
seaborn.pydata.org

The candidate can freely access the teaching materials available in the course website, including the available examples of python notebooks.
The notebook must be uploaded in both original and pdf form, as two separate files.