Main objective

This document is a summary of a data analysis using python sklearn library. One of the difficulty during clustering might be that there is no straight metrics or methods that allow us to validate our model. The elbow method works fine for the k-means, though graphical. But what about the others clustering algorithms ? How can we evaluate which algorithm is better over another ? Which hyperparameters are the best ? We will try to address these question in this small study by 2 ways :

- using silhouette score and calinski harabasz score from sklearn to evaluate best parameters
- PCA analysis is famous to reduce dimensionality of a dataset to smaller set of features and speed up performance. Another common use of PCA is for data visualization in 2D or 3D, which is what we are going to explore here. We will also use t-SNE which is another technique of dimensionality reduction. After this, we hope to be able to pick up the best algorithm

Raw data, data preparation

We will use the wine dataset, it can also be obtained from :
https://archive.ics.uci.edu/ml/datasets/wine+quality
We know there are 2 type of wines, red and white. This seems like a clustering problem. The dataset contains a total of 6497 entries. It is slightly unbalanced with 4898 entries white wine and 1599 red wine. We will remove the feature “quality” and “color” as if they were not known. The dataset contains afterwards 11 numeric features that we will scale using standardscaler.

Tested models

We evaluate 4 models : k-means, agglomerative clustering, dbscan and mean shift. We cannot use gridsearch as it is unsupervised learning, so no usual scoring as for classification or regression problems. To address this, we implement a custom search function that will evaluate each model (see paramsearch.py) given a set of parameter using silhouette score and calinski harabasz score.

Results

For each clustering method, the best parameters showing the highest silhouette and calinski harabasz score were :
Kmeans : {'n_clusters': 2}
agglomerative clustering: {'n_clusters': 2, 'linkage': 'ward'}
dbscan: {'eps': 5, 'min_samples': 300, 'metric': 'manhattan'}
mean shift: {'bandwidth': 10}
The mean shift output 1 cluster so we will put it on the side for the rest of the discussions given the poor performance. The 3 others provide 2 clusters with a repartition close to red and white, not bad. Normally we do not know this, so we will try to visualize our results and maybe decide which model is the best.

Visualization and best model

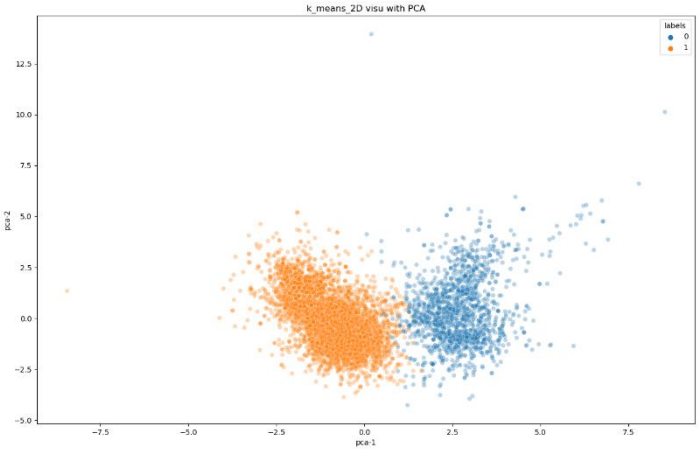
Let’s plot the data and our clustering results using PCA first and t-SNE on the next page.
We plot the following figures:

- page 2, PCA decomposition in 2D (2 components)
- page 2, PCA decomposition in 3D (3 components)
- page 3, t-SNE reduction 2D (2 components)

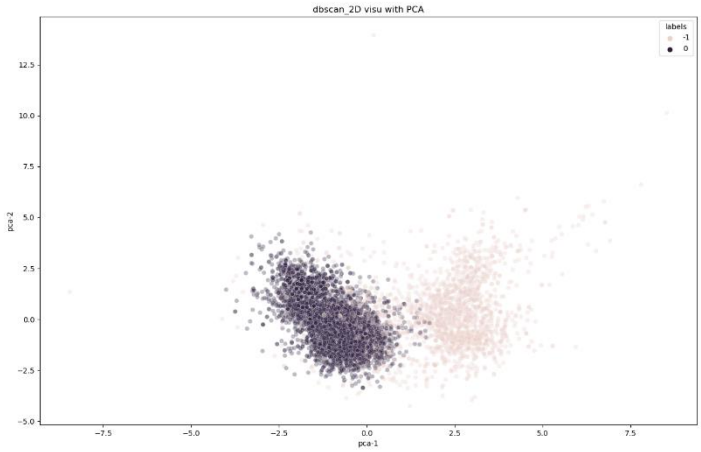
Conclusion

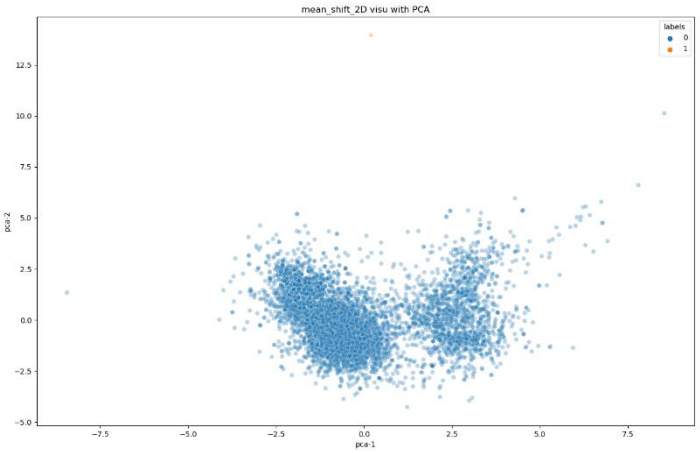
From the figures, we see the low performance of mean shift with no cluster. dbscan is less clear as per 3D plot (clouds intermix with each other) so we might exclude this model. The 2 best models seem to be k-means and agglomerative clustering. If we look at the 3D plots, the number of 2 clusters is not obvious. However, from 2D PCA plots, there are clearly 2 main clouds though the separation is not obvious. Finally, when we look at t-SNE, there is a clear separation between 2 data clouds which match with our previous findings. Here the problem was relatively simple. Maybe this approach can provide additional insight when dealing with more complex datasets.

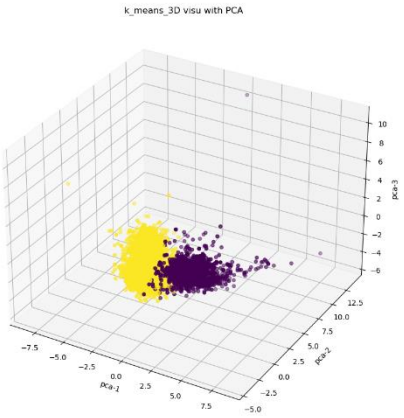
Author: NL

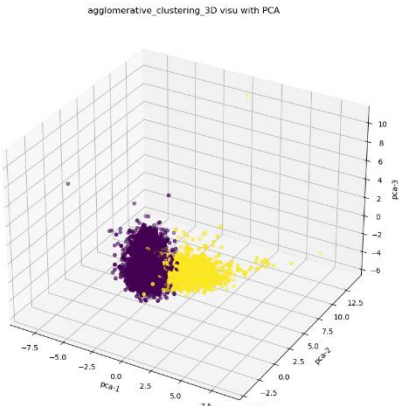

k_means_2D visu with PCA
labels
0
1
pca-1
pca-2

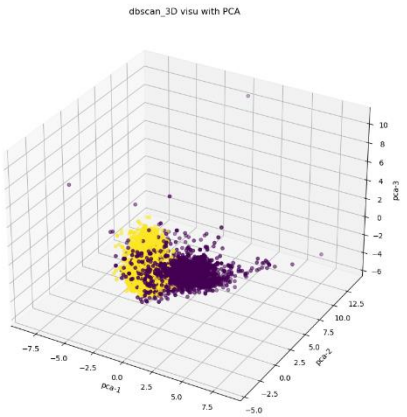

agglomerative_clustering_2D visu with PCA
labels
0
1
pca-1
pca-2

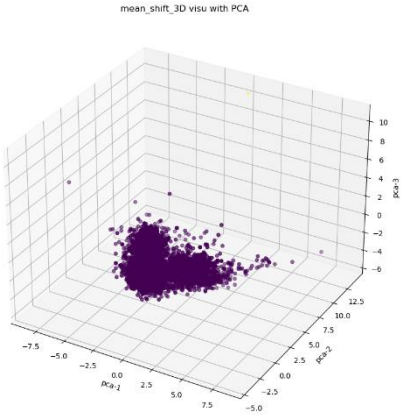

dbscan_2D visu with PCA
labels
-1
0
pca-1
pca-2


mean_shift_2D visu with PCA
labels
0
1
pca-1
pca-2

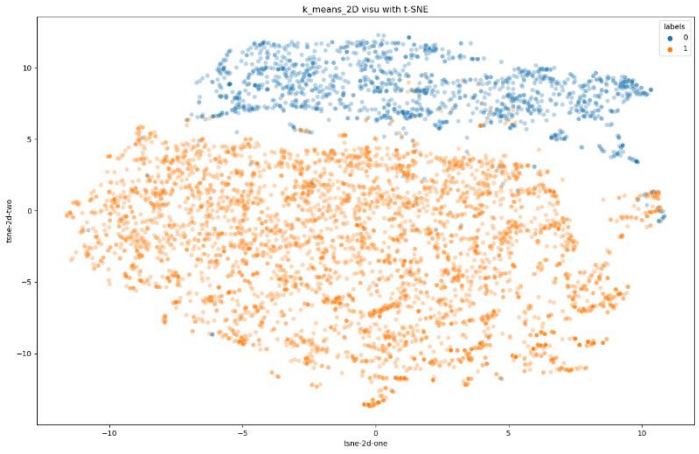

k_means_3D visu with PCA
pca-1
pca-2
pca-3

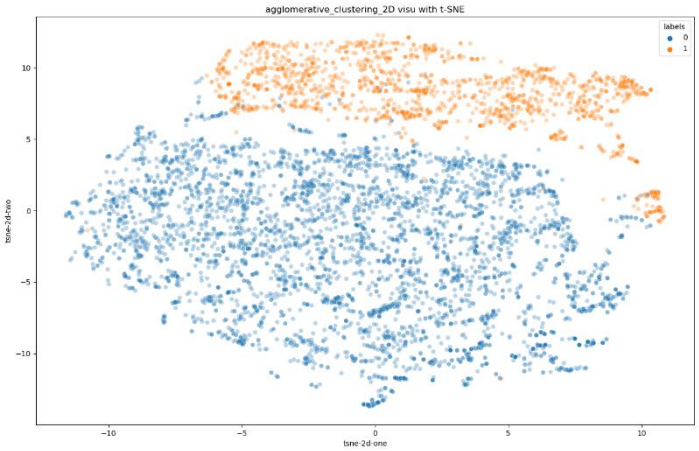

agglomerative_clustering_3D visu with PCA
pca-1
pca-2
pca-3

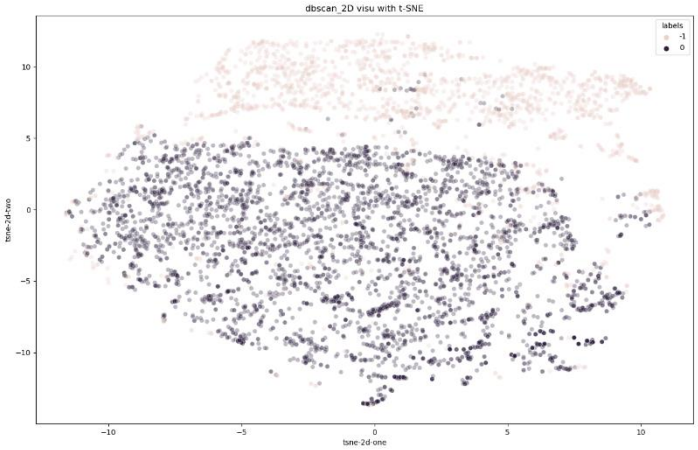

dbscan_3D visu with PCA
pca-1
pca-2
pca-3

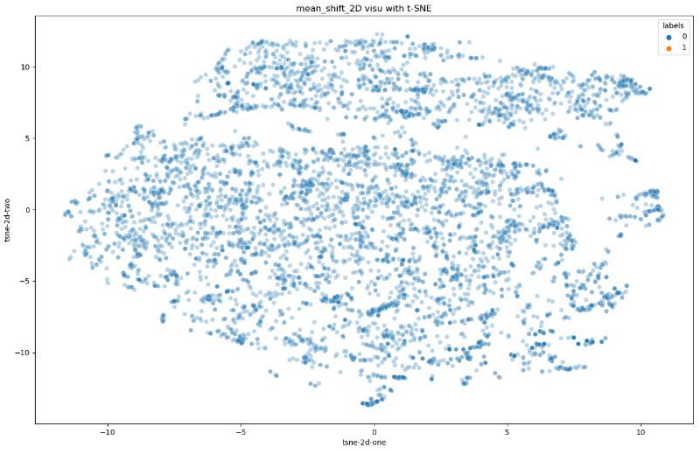

mean_shift_3D visu with PCA
pca-1
pca-2
pca-3

Author: NL


k_means_2D visu with t-SNE
labels
0
1
tsne-2d-two
tsne-2d-one


agglomerative_clustering_2D visu with t-SNE
labels
0
1
tsne-2d-two
tsne-2d-one


dbscan_2D visu with t-SNE
labels
-1
0
tsne-2d-two
tsne-2d-one


mean_shift_2D visu with t-SNE
labels
0
1
tsne-2d-two
tsne-2d-one

Author: NL