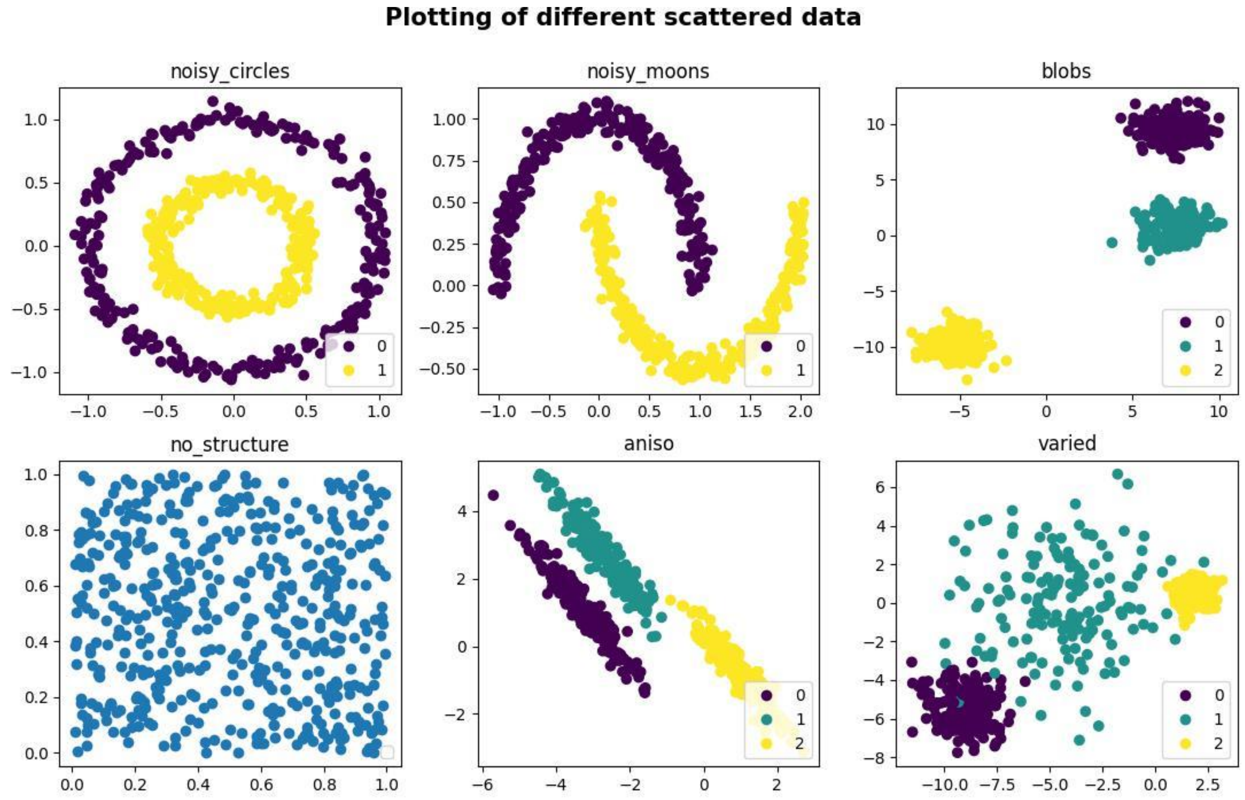
WORKSHOP II SOLUTION - POINT 6

APPLYING K-MEANS, K-MEDOIDS, DBSCAN AND SPECTRAL CLUSTERING FROM SCIKIT-LEARN

a) What can you say about the different datasets?

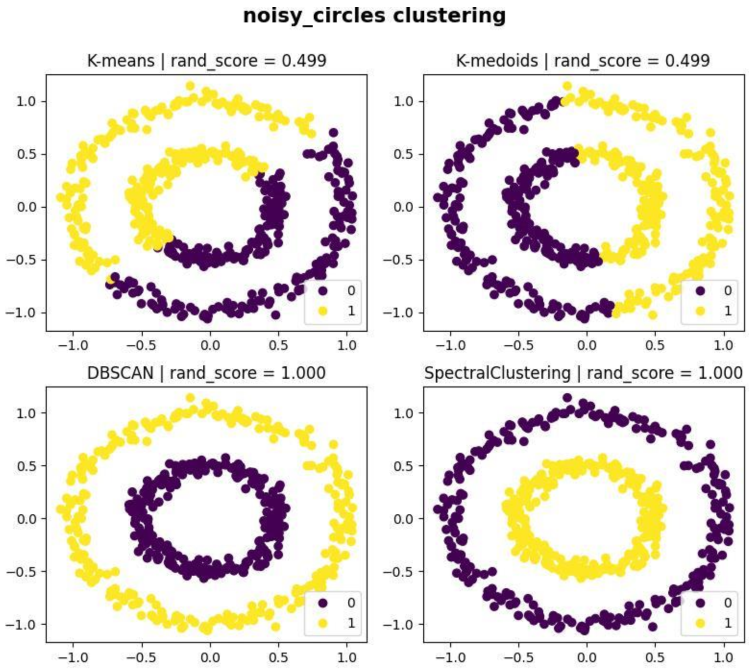


About the scatterplots of the different datasets, it can be said that:

- The **noisy_circles** and **noisy_moons** data sets contain two clusters in which the points are not closer to each other and are not compact towards the center of the cluster. The **noisy_circles** dataset contains 2 concentric circle-shaped clusters and the **noisy_moons** dataset contains 2 half-moon-shaped clusters.
- The **blobs** dataset contains 3 compact clusters of similar density and with some noise.
- The **no_structure** dataset contains sparse data where no cluster or pattern can be identified in the data.
- The **aniso** dataset contains 3 elliptical clusters in which not all points are close to each other.
- The **varied** data set contains 3 clusters of different density, one of them very sparse.

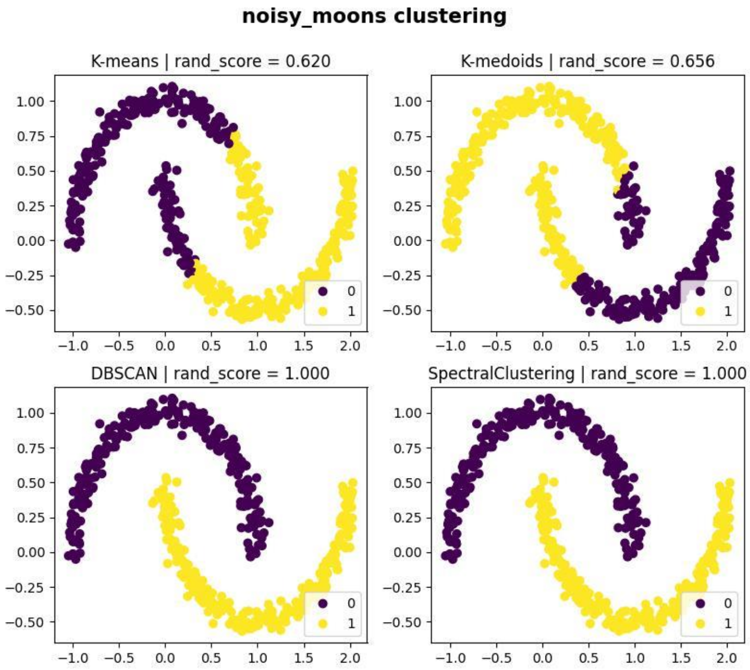
b) Compare the results of each algorithm with respect to each dataset:

NOISY_CIRCLES



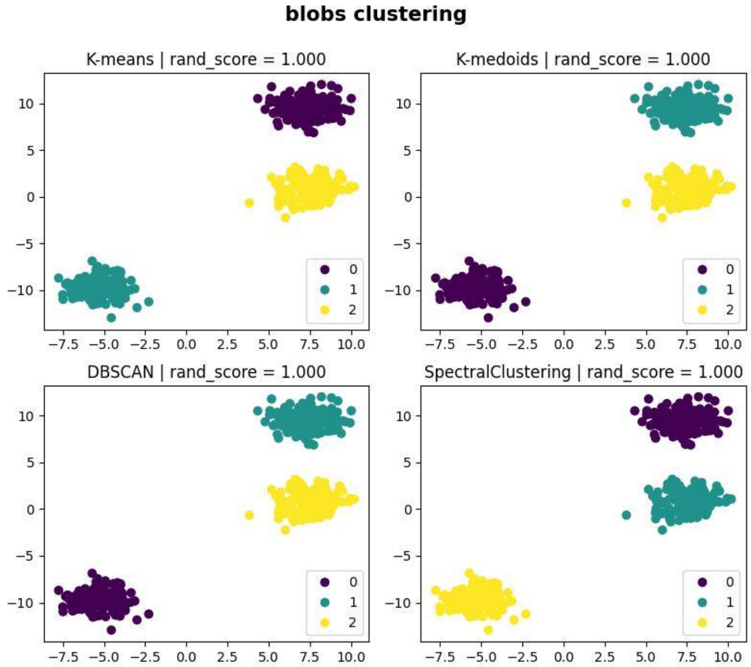
K-means and k_medoids try in vain to cluster the data around two distinct centers located in space. DBSCAN and SpectralClustering succeed in clustering the data into two concentric circles, as indicated by the labels.

NOISY_MOONS



The clustering results are similar to those of the Noisy_circles dataset. K-means and k_medoids fail and DBSCAN and SpectralClustering succeed.

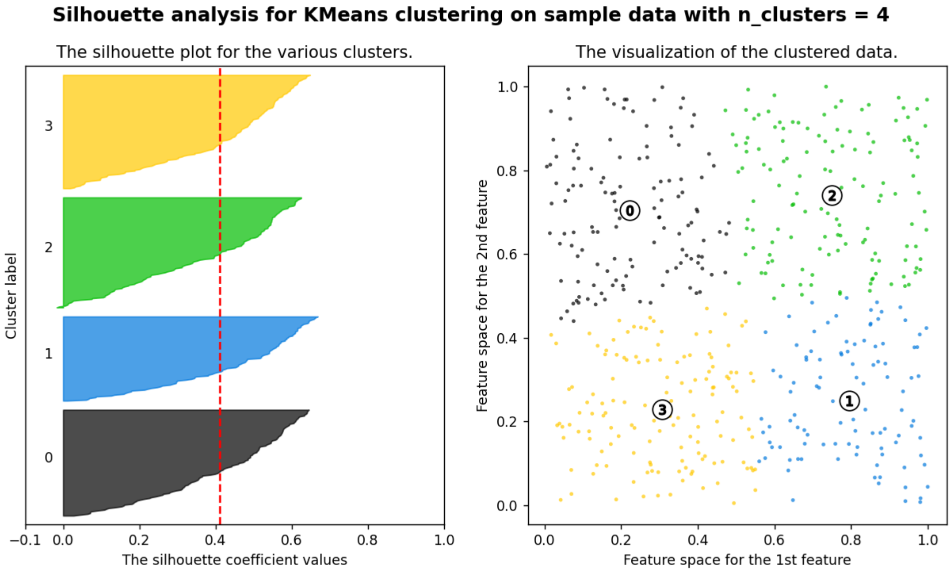
BLOBS

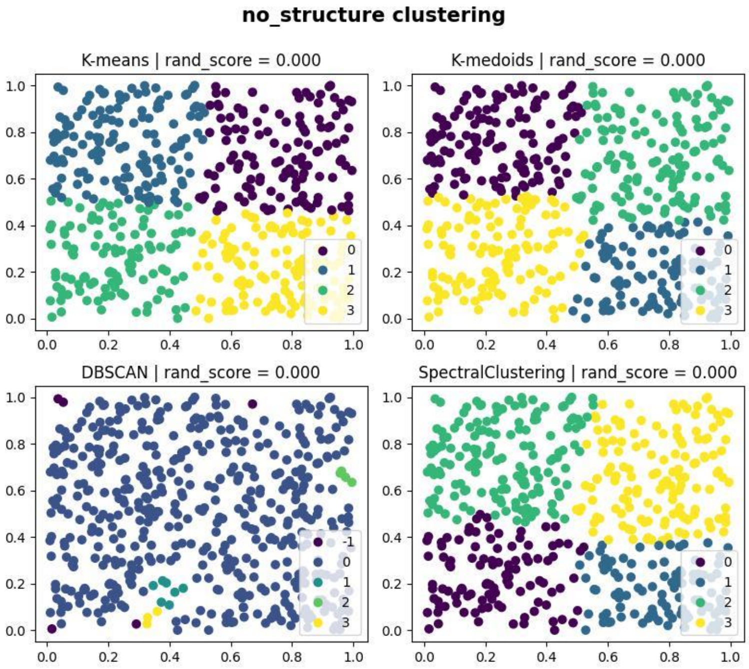


For this dataset, all clustering algorithms succeed in correctly grouping the data into 3 clusters after adjusting the parameters of each algorithm.

NO_STRUCTURED

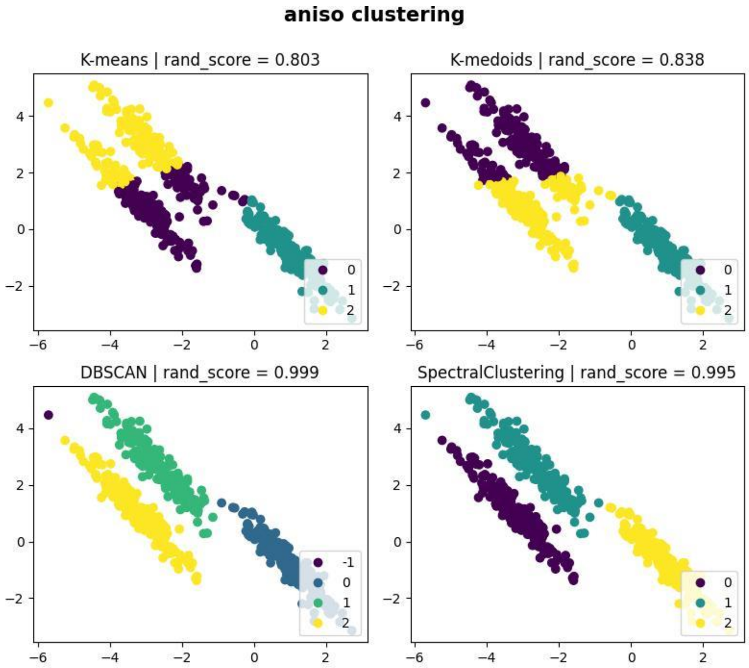
Before applying the clustering algorithms, the Silhouette analysis is performed to determine the number of clusters to set in the parameters of the algorithms. In this analysis it is defined that for n_clusters = 4 the best average silhouette_score is obtained.





Using this parameter for the k_means, k-medoids and SpectralClustering algorithms, which allow setting this value, it is observed that k-means and SpectralClustering obtain similar results, while k-medoids groups the data with different positions of the cluster centers. On the other hand, DBSCAN does not find a way to cluster the data, after varying the parameters it is observed that, if min_samples is not set, the algorithm starts to build small clusters among a large cluster, or by setting min_samples it groups the data in a single cluster, regardless of the distance eps.

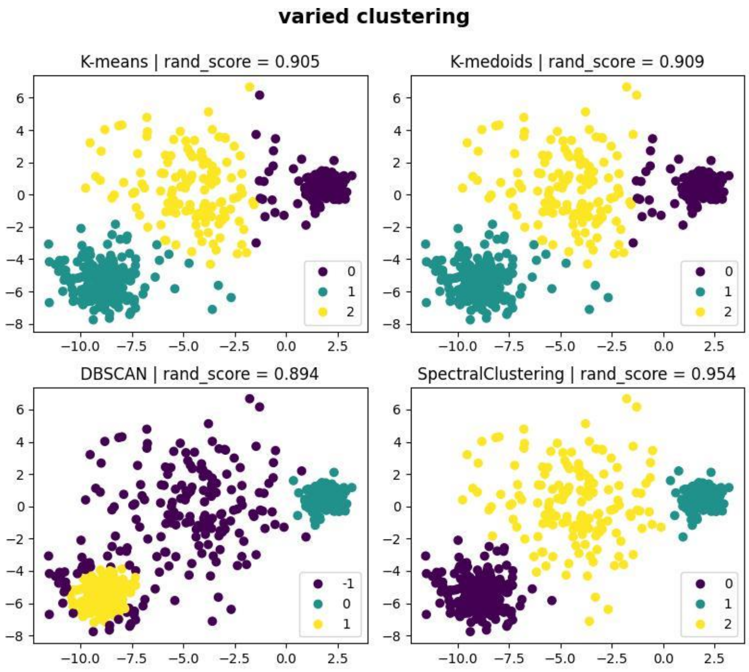
ANISO



K-means and k_medoids get similar results by correctly clustering the outermost cluster but fail on the other two clusters because they are too close together and the points are not compact towards the center of each cluster.

DBSCAN and SpectralClusterng do a great job clustering the data, but they are not perfect. There is a point too far away from the cluster and neither algorithm manages to cluster it correctly, DBSCAN creates a new cluster with just that point while SpectralClustering rarely assigns that point to the wrong cluster which even appears to be the farthest away.

VARIED



K-means and k_medoids correctly assign cluster centers but cannot correctly cluster points from sparse clusters that are closer to other clusters.

In general, DBSCAN and SpectralClusterng group the data correctly, except for some points belonging to the sparse cluster.