**CLUSTER ANALYSIS BY K-MEANS**

Clustering exploits similarities between the data to be analyzed, similarities that can be of various nature but which are essentially a distance between the dataset points.

K-means is one of the most widespread and best performing clustering algorithms.

The k means is based on the distance of the values ​​from its centroid.

The centroid represents the center of gravity of the cluster itself.

You must first decide the number of classes for which to divide the dataset. The number of classes is defined with k (hence the name of the k-means method).

The term "means" means the use of centroids or the mean values.

To find the number of clusters you can use two different methods: The Elbow method and the Silhouette method.

1) **Elbow Method**: This approach plots the WCSS (Within Clusters Summed Squares) values ​​and selects the point where the parameter value falls more than the previous value.

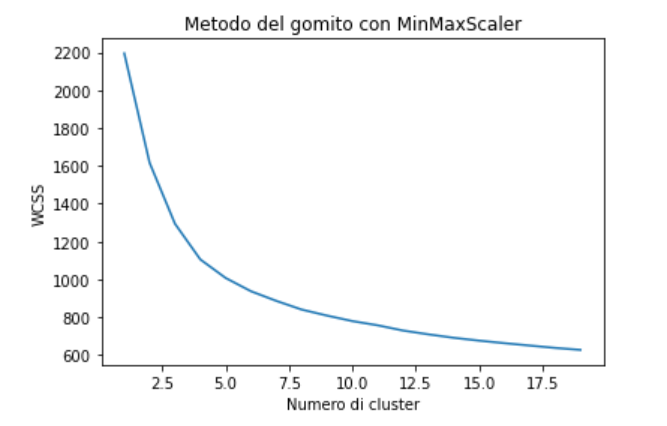
2) **Silhouette method:** in this procedure the silhouette coefficient is plotted and the maximum value is selected.

**ELBOW METHOD**

This method uses the WCSS parameter, which represents the square distance of each point from the centroid of the cluster to which it belongs. The centroid is defined as the geometric point where most of the data in the cluster is concentrated.

The WCSS parameter expires as the number of clusters increases. In this way we obtain a curve whose slope is variable.

The optimal number of clusters is defined when the decrease in the WCSS parameter is no longer significant, that is, the slope of the curve varies visibly.



From the graph it can be seen that there is a significant difference when the number of clusters is 3. After this number, the difference between the clusters is attenuated or not significant. So in this case, the optimal number of clusters is 3.

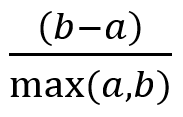
**Silhouette Method**

The Silhouette method uses a unit of measurement whose range varies from -1 to 1 [-1; 1]. This unit of measurement allows us to understand the distance of each point in a cluster to the points in neighboring clusters.

With a Silhouette coefficient close to +1, it is indicated that the sample under examination is located far from the neighboring clusters (and therefore belongs to a well-defined cluster). A value of 0 indicates that the sample is located or is very close to the decision boundary between two neighboring clusters. A negative value (-1) instead indicates that those samples may have been assigned to the wrong cluster (and therefore possible outliers).

Ideally, the distance of a sample from the points of the clusters to which it belongs should be less than the distance of the sample from the points of the other clusters.

With "a" we indicate the distance of the sample with respect to the points of its own cluster, and with "b" instead we indicate the distance of the sample with respect to the points of the most distant clusters.

To calculate the Silhouette coefficient we use the following form: .

