**Section A**

1. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
2. Density-Based Spatial Clustering of Applications with Noise is a very popular unsupervised learning algorithm uses for data clustering. “It is a density-based clustering non-parametric algorithm”[[1]](#endnote-1). In other words, given a set of points, DBSCAN groups points that are very close together based on the density. Any point alone, or outside the density is consider an outlier. DBSCAN is affected by the presence of noise and outliner in the data.
3. Compute the neighborhood of all data points

Group the data points that have at least some specific number of points in their cluster.

**4a.** Numerical and categorial data

**4b.** It requires at least two observations

**4c.** DBSCAN does not require the number of clusters in data a priori.

**5a.** DBSCAN does not quite well work when dealing with clusters of varying densities, but it works well for finding clusters with different sizes and shapes[[2]](#endnote-2).

**5b.**

**6a**. We need to import DBSCAN from sklearn. cluster

**6b.** We call fit () and fit-predict () on a DBSCAN object.

**7a.** To evaluate we need to calculate the silhouette score. “The silhouette score is calculated utilizing the mean intra- cluster distance between points, AND the mean nearest-cluster distance.[[3]](#endnote-3)”

**7b.** The silhouette score varies from -1 to 1. The score of -1 indicates a worse performance, and silhouette score of 1 signifies a best performance. Silhouette score of 0 indicates that there is an overlapping among the clusters

**7c.** We import silhouette score form sklearn. Metrics. Metris.silhouette\_score()

718 789 1007

Section B

**8.** This dataset is available at this link.

<https://archive.ics.uci.edu/ml/machine-learning-databases/00292/>

According to the description given in the official [UCI machine learning repository of this dataset](https://archive.ics.uci.edu/ml/datasets/wholesale+customers), information about the features of the dataset is as follows:

Fresh: annual spending on fresh products

Milk: annual spending on milk products

Grocery: annual spending on grocery products.

Frozen: annual spending on frozen products.

Detergent Paper: annual spending on detergents and paper products.

Delicatessen: annual spending on and delicatessen products.

Channel: customers Channel - Horeca (Hotel/Restaurant/ coffee) or Retail channel (REGION)

**9**. Researching DBSCAN algorithm helped us to understand the differences between K-Means and DBSCAN**.** They both are clustering algorithms, where DBSCAN does not require to specify the number of clusters in the data a priori, as to opposed to K-Means. DBSCAN detects outlier very easy.

**10**.

1. <https://en.wikipedia.org/wiki/DBSCAN> [↑](#endnote-ref-1)
2. <https://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818> [↑](#endnote-ref-2)
3. <https://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818> [↑](#endnote-ref-3)