Machine Learning

- Gradient Descent Algorithm
- Linear Regression
- Non-Linear Regression
- Logistic Regression
- Decision Trees
  - Regression Trees
  - Classification Trees
  - Model complexity and ensemble models
- Clustering Algorithms
  - K-Means
  - Hierarchical clustering
  - DB-Scan
  - Mean Shift
  - GMM
- Support Vector Machine

Deep Learning

- MLP
- CNN

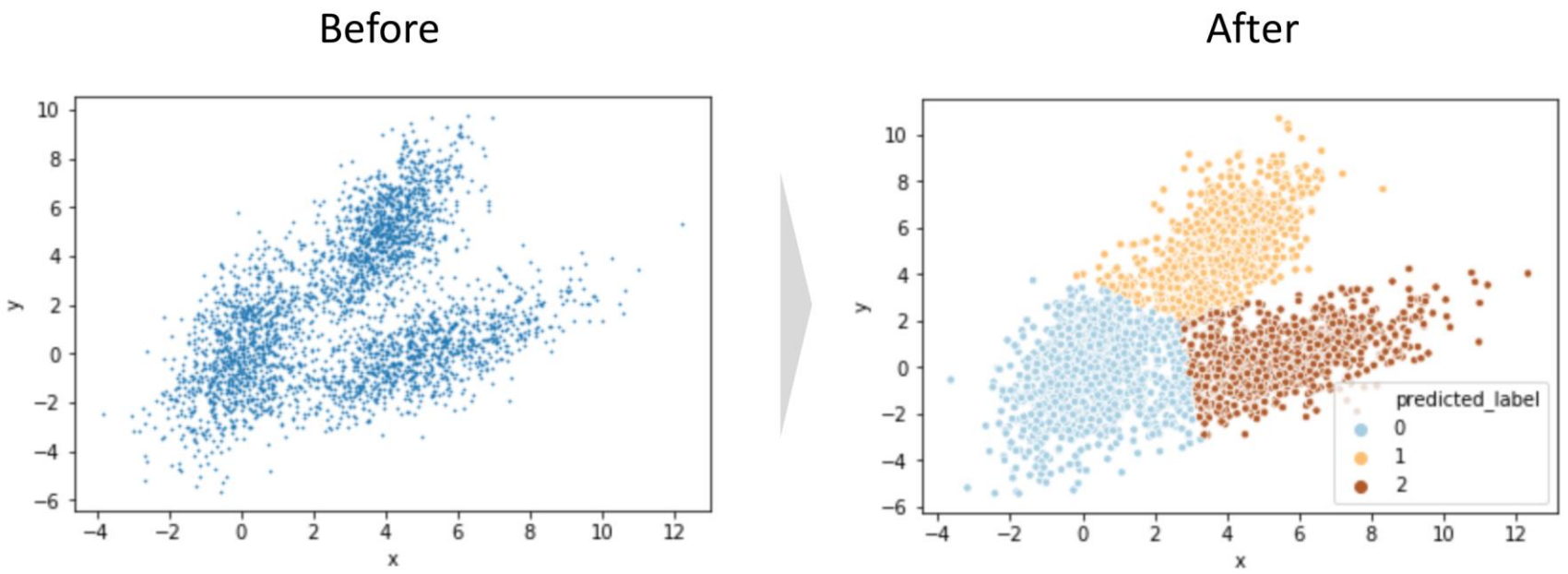
Datasets

- Breast Cancer Wisconsin
- MIMIC-III
- Framingham Heart Study
- Alzheimer’s Disease Neuroimaging Initiative
- Drug discovery
- Microbiome

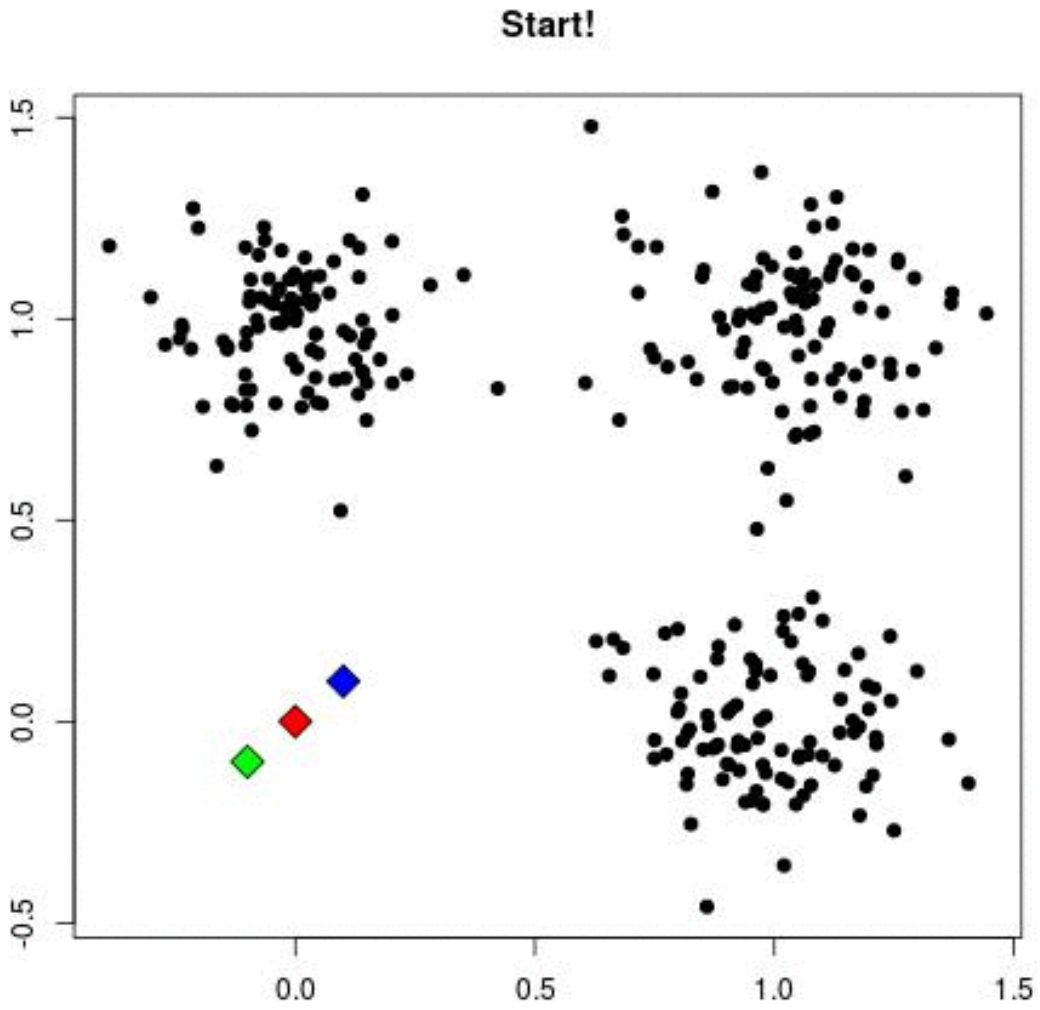
C7 - Clustering

Clustering

- ML technique, like Regression and Classification
- Unsupervised technique
- **Clustering** is the task of dividing the data points into a number of groups such that the two data points in the same group have same features whereas two data points each in different groups contain dissimilar features. It is basically the grouping of data points based on similarity and dissimilarity.

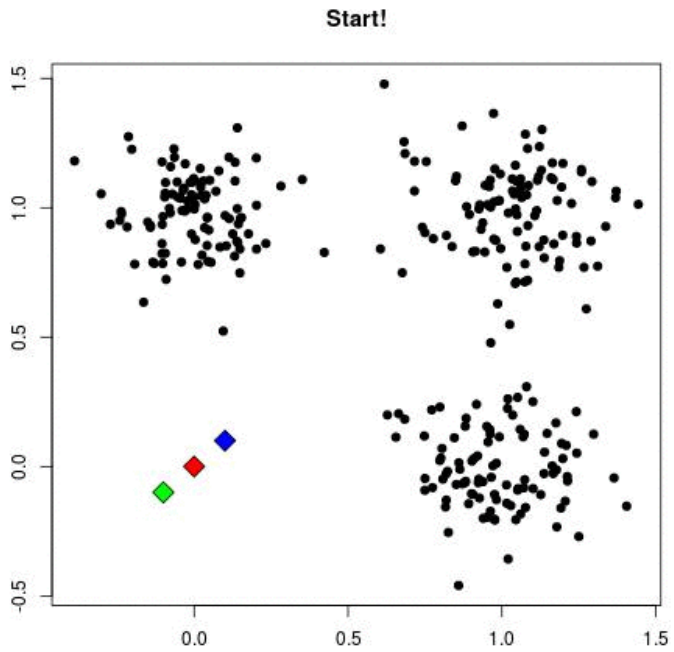


KMeans



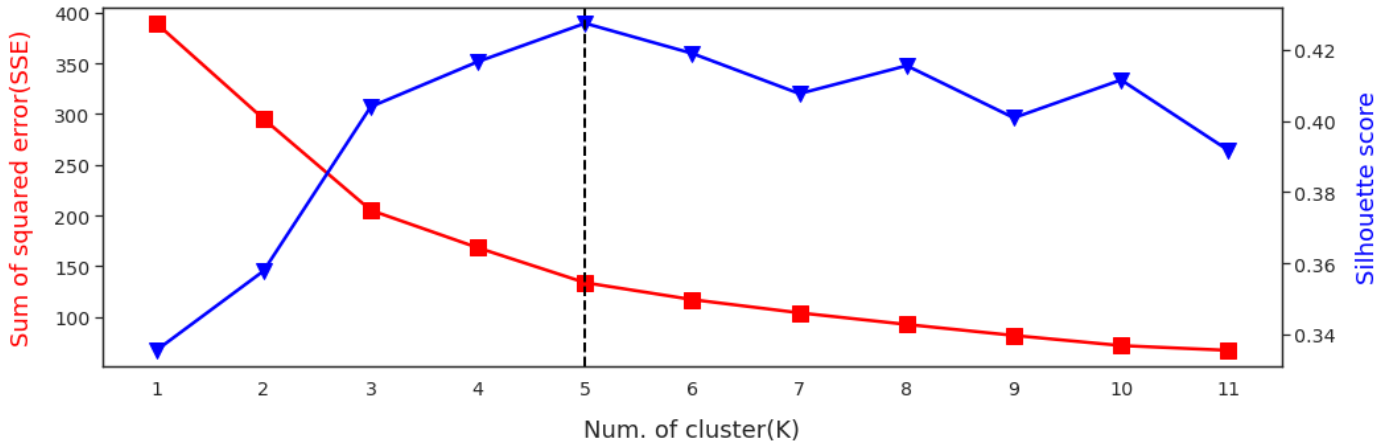
K-Means

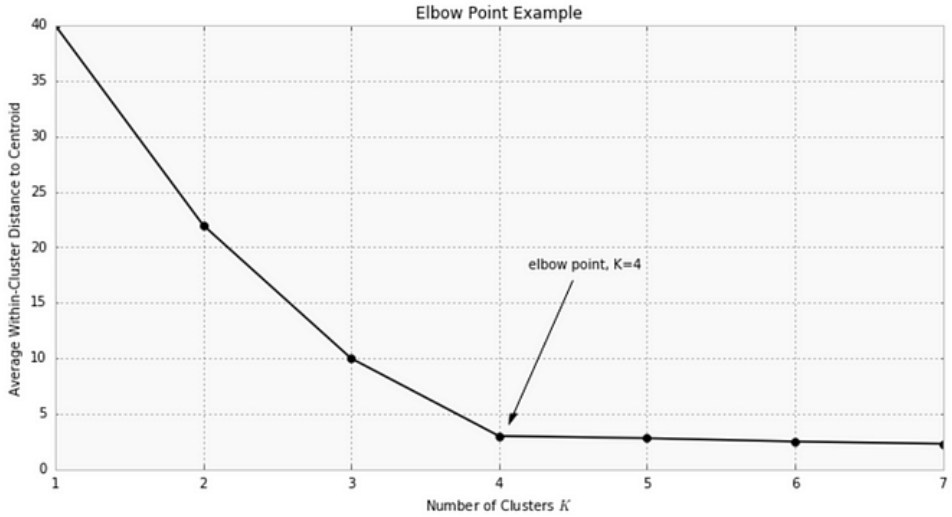
1. Choose the number of clusters(K) and obtain the data points
2. Place the centroids c_1, c_2, ... c_k randomly
3. Repeat steps 4 and 5 until convergence or until the end of a fixed number of iterations
4. for each data point x_i:
    - find the nearest centroid(c_1, c_2 .. c_k)
    - assign the point to that cluster
5. for each cluster j = 1..k
    - new centroid = mean of all points assigned to that cluster
6. End



KMeans

- Best K? Elbow method
- Centroid initialization? K-means++





WCSS (Within-cluster sums of squares)

$$\sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \left\| \mathbf{x}_j - \boldsymbol{\mu}_i \right\|^2$$

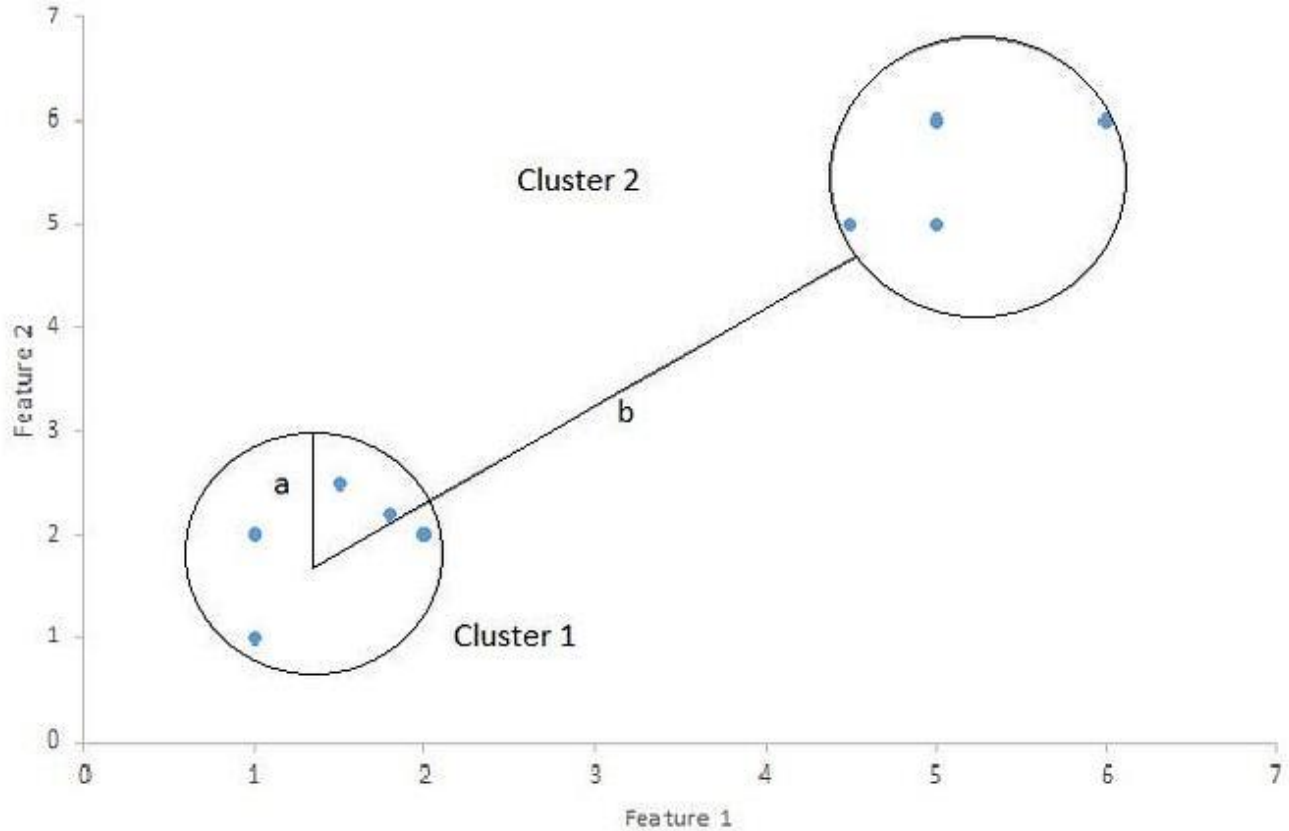
donde $\boldsymbol{\mu}_i$ es la media de puntos en S_i.

Silhouette score

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)}$$

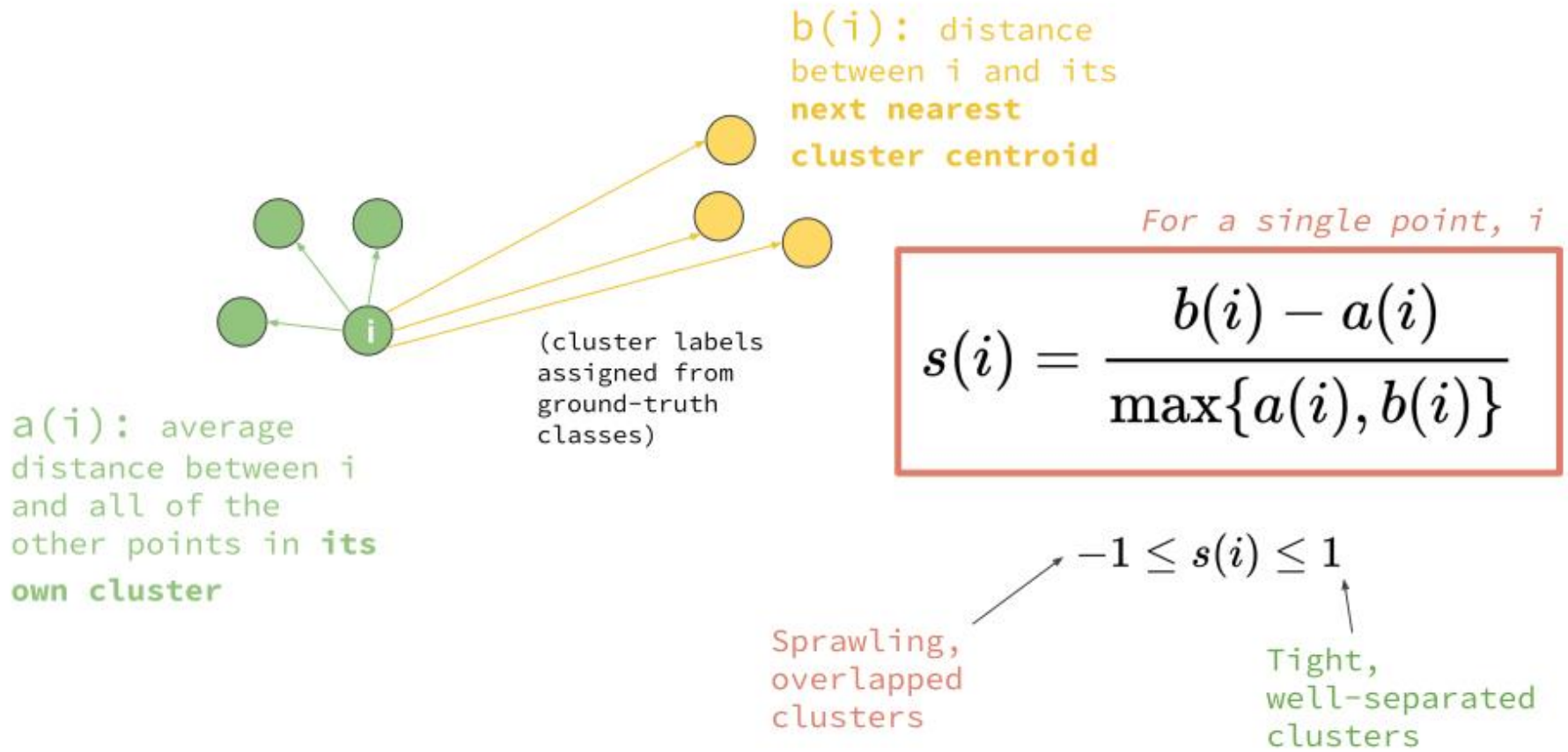
$$b_i = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

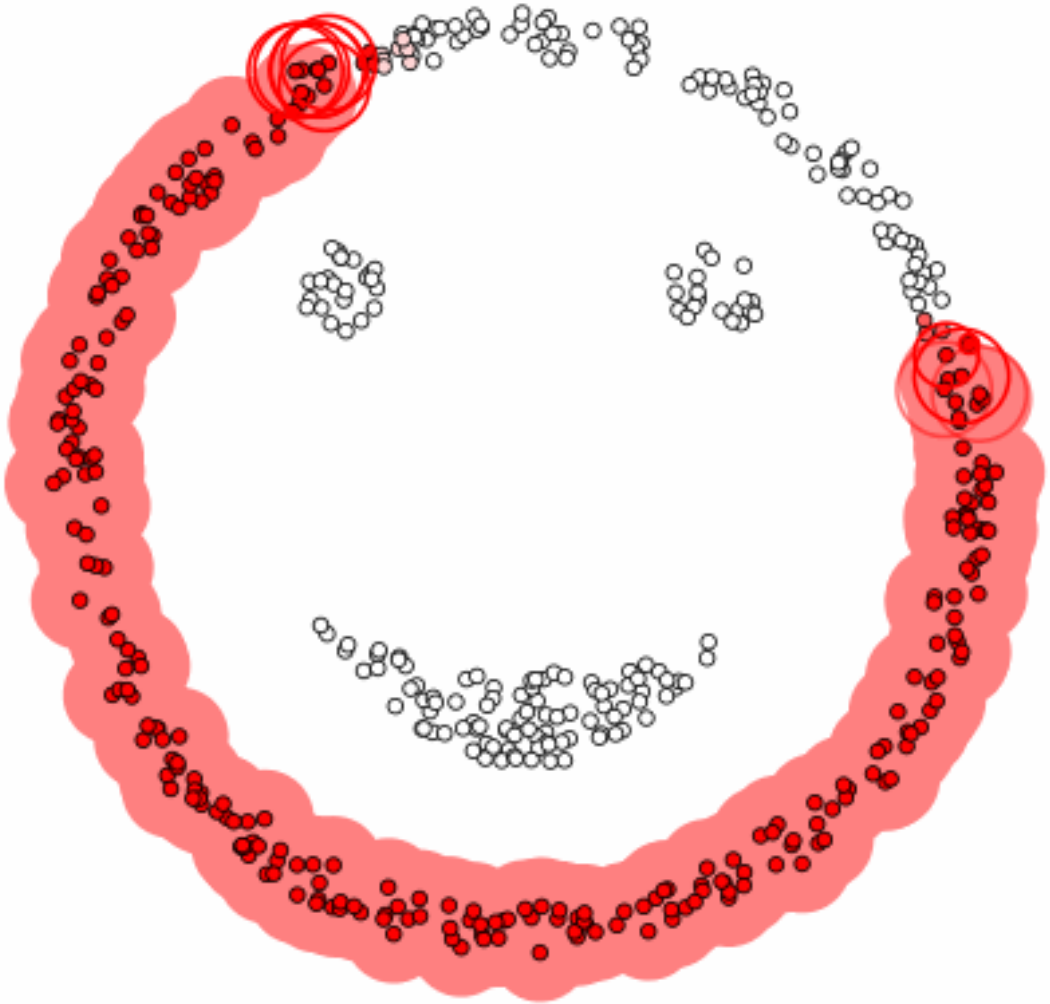


Silhouette score

1: Means clusters are well apart from each other and clearly distinguished.
0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.
-1: Means clusters are assigned in the wrong way.

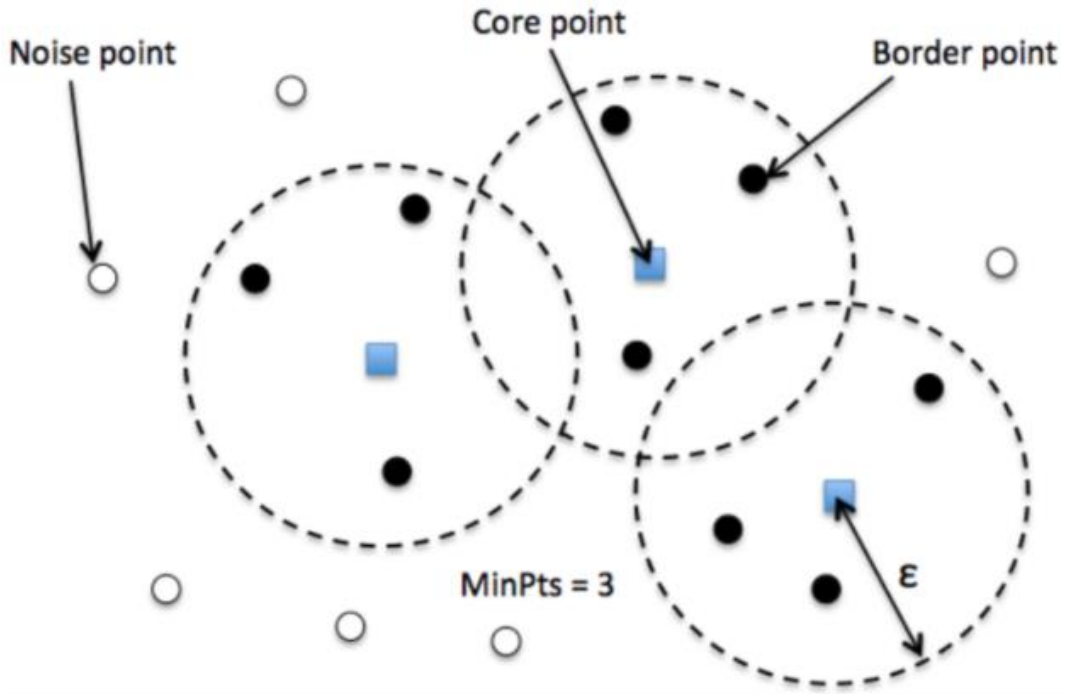


DBScan



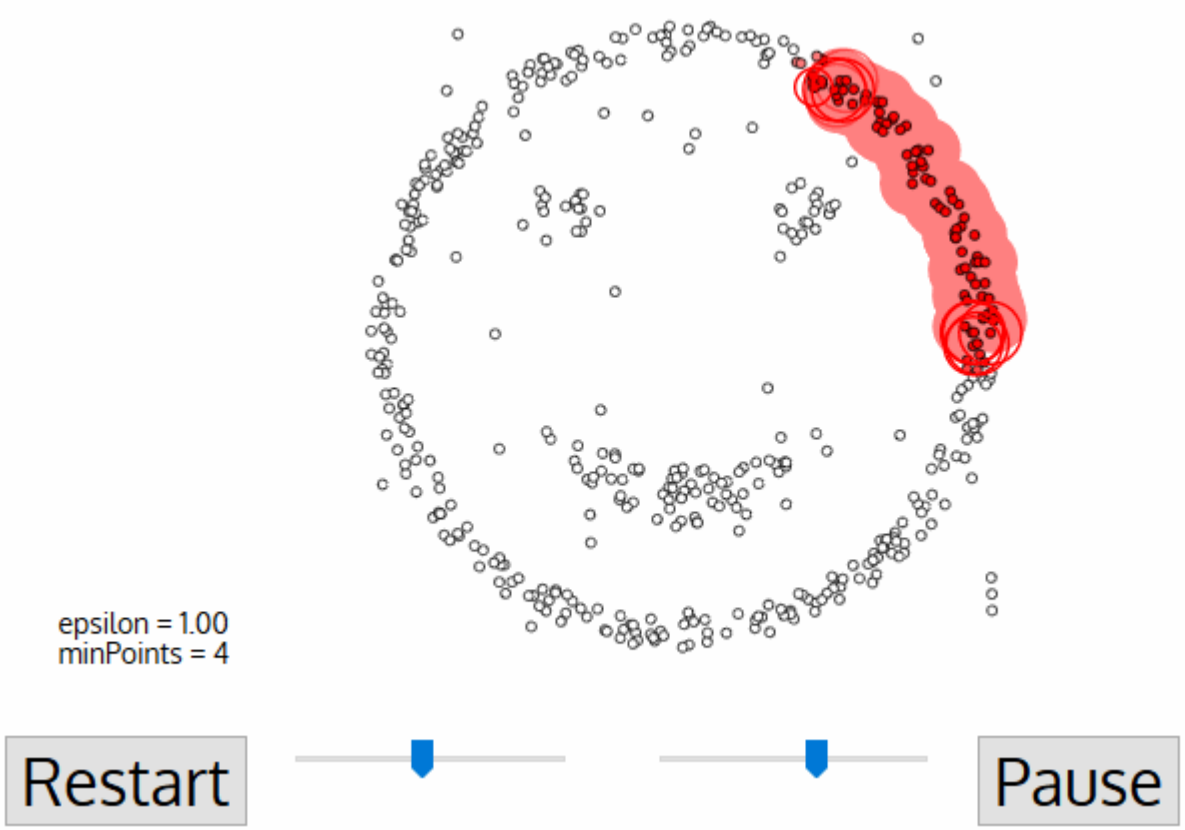
DBScan

- Identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density.
- 2 hyperparameter:
  - **minPts:** The minimum number of points (a threshold) clustered together for a region to be considered dense.
  - **eps (ε):** A distance measure that will be used to locate the points in the neighborhood of any point.



DBScan

1. The algorithm proceeds by arbitrarily picking up a point in the dataset (until all points have been visited).
2. If there are at least ‘minPoint’ points within a radius of ‘ε’ to the point then we consider all these points to be part of the same cluster.
3. The clusters are then expanded by recursively repeating the neighborhood calculation for each neighboring point



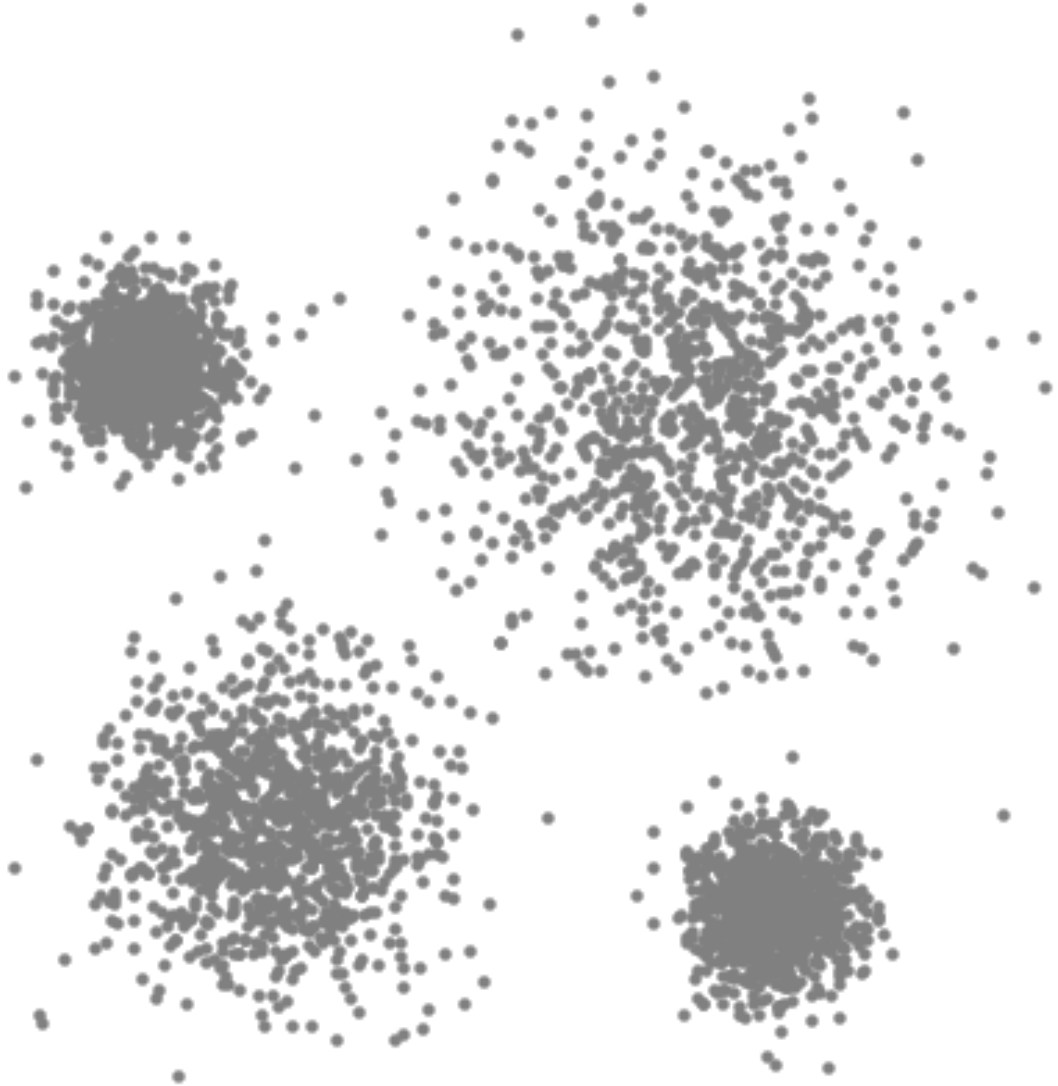
DBScan

ALGORITHM 1: Pseudocode of Original Sequential DBSCAN Algorithm

```
Input: DB: Database
Input: ε: Radius
Input: minPts: Density threshold
Input: dist: Distance function
Data: label: Point labels, initially undefined
1  foreach point p in database DB do                          // Iterate over every point
2      if label(p) ≠ undefined then continue                  // Skip processed points
3      Neighbors N ← RangeQuery(DB, dist, p, ε)               // Find initial neighbors
4      if |N| < minPts then                                   // Non-core points are noise
5          label(p) ← Noise
6          continue
7      c ← next cluster label                                 // Start a new cluster
8      label(p) ← c
9      Seed set S ← N \ {p}                                   // Expand neighborhood
10     foreach q in S do
11         if label(q) = Noise then label(q) ← c
12         if label(q) ≠ undefined then continue
13         Neighbors N ← RangeQuery(DB, dist, q, ε)
14         label(q) ← c
15         if |N| < minPts then continue                      // Core-point check
16         S ← S ∪ N
```

S.No.	K-means Clustering	DBScan Clustering
1.	Clusters formed are more or less spherical or convex in shape and must have same feature size.	Clusters formed are arbitrary in shape and may not have same feature size.
2.	K-means clustering is sensitive to the number of clusters specified.	Number of clusters need not be specified.
3.	K-means Clustering is more efficient for large datasets.	DBSCan Clustering can not efficiently handle high dimensional datasets.
4.	K-means Clustering does not work well with outliers and noisy datasets.	DBScan clustering efficiently handles outliers and noisy datasets.
5.	In the domain of anomaly detection, this algorithm causes problems as anomalous points will be assigned to the same cluster as “normal” data points.	DBScan algorithm, on the other hand, locates regions of high density that are separated from one another by regions of low density.
6.	It requires one parameter : Number of clusters (K)	It requires two parameters : Radius(R) and Minimum Points(M) R determines a chosen radius such that if it includes enough points within it, it is a dense area. M determines the minimum number of data points required in a neighborhood to be defined as a cluster.
7.	Varying densities of the data points doesn’t affect K-means clustering algorithm.	DBScan clustering does not work very well for sparse datasets or for data points with varying density.

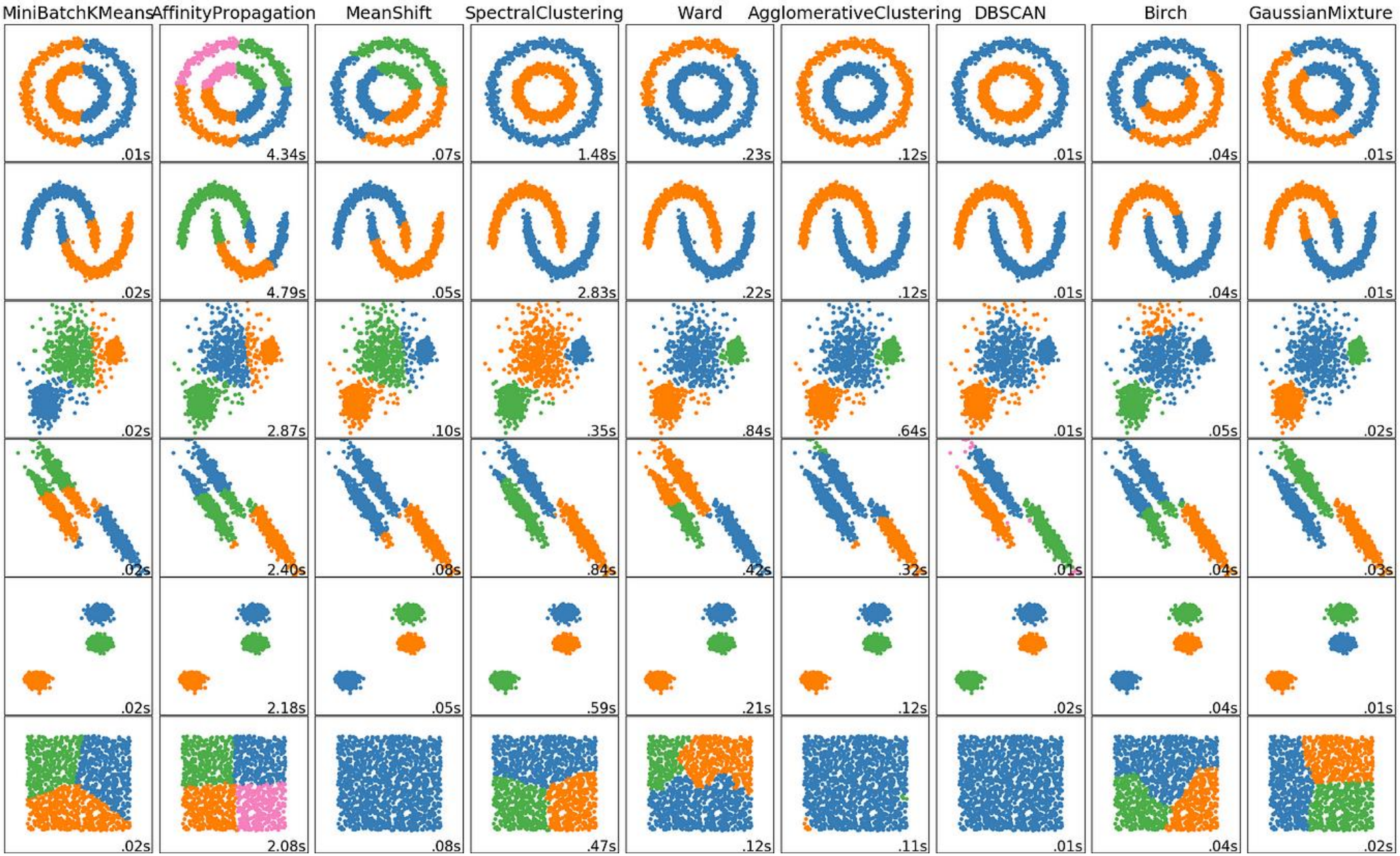
MeanShift



DBSCAN

k-means

my humble ML course



Bibliography

https://github.com/jeremy-jmc/IA-P002
https://drive.google.com/drive/u/0/folders/1X3AjOQ2G7NZO3U_0KGYzPXex49vx3Iy-
https://towardsdatascience.com/k-means-clustering-introduction-to-machine-learning-algorithms-c96bf0d5d57a
https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html
https://www.kaggle.com/code/alirezahanifi/customer-segmentation-k-means-dbscan-meanshift
https://www.geeksforgeeks.org/difference-between-k-means-and-dbscan-clustering/
https://soroushhashemifar.medium.com/kmeans-vs-dbscan-d9d5f9dbee8b

https://es.wikipedia.org/wiki/Silhouette_(clustering)
https://tushar-joshi-89.medium.com/silhouette-score-a9f7d8d78f29
https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

https://www.kaggle.com/code/alirezahanifi/customer-segmentation-k-means-dbscan-meanshift

https://medium.com/@ainsupriyofficial/unsupervised-learning-clustering-algorithms-fad2d86cce6a
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68