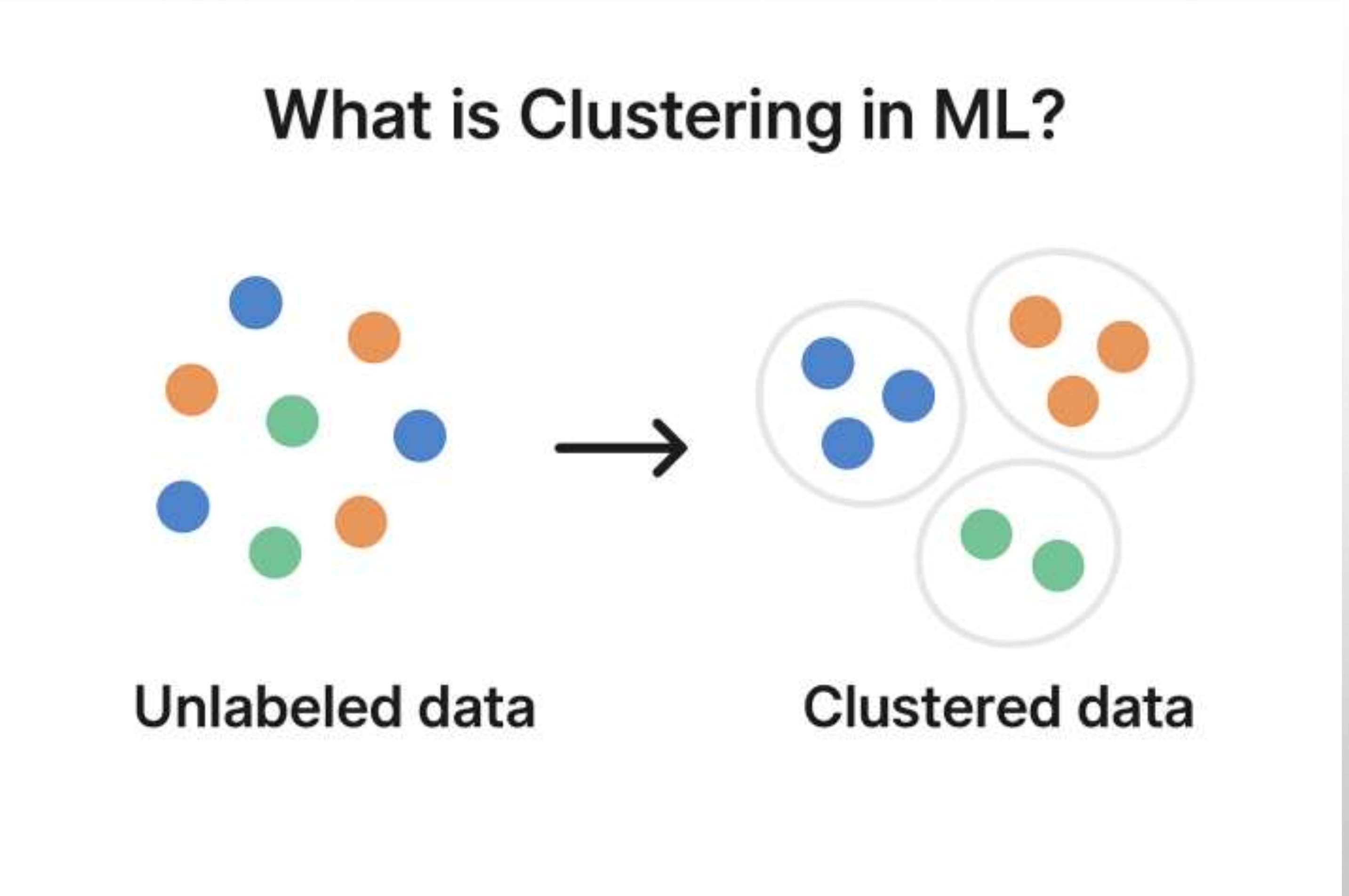
CLUSTERING ALGORITHMS IN MACHINE LEARNING

OVERVIEW OF CLUSTERING: KEY METHODS, VISUALIZATIONS, AND MATHEMATICAL FORMULAS

PRESENTED BY: MANI MARAN . R

WHAT IS CLUSTERING IN ML?

- *CLUSTERING IS AN **UNSUPERVISED LEARNING TECHNIQUE.***
- *IT GROUPS DATA POINTS INTO CLUSTERS SO THAT:*
  - *POINTS IN THE **SAME CLUSTER** ARE MORE SIMILAR TO EACH OTHER.*
  - *POINTS IN **DIFFERENT CLUSTERS** ARE MORE DISSIMILAR.*
- *USED WHEN WE **DON'T HAVE LABELS** IN THE DATASET.*
- *APPLICATIONS: CUSTOMER SEGMENTATION, ANOMALY DETECTION, IMAGE COMPRESSION, DOCUMENT GROUPING, ETC.*



What is Clustering in ML?

K-MEANS CLUSTERING

- GROUPS DATA INTO K CLUSTERS BY MINIMIZING DISTANCE TO CLUSTER CENTROIDS. WORKS WELL FOR SPHERICAL AND EVENLY SIZED CLUSTERS.

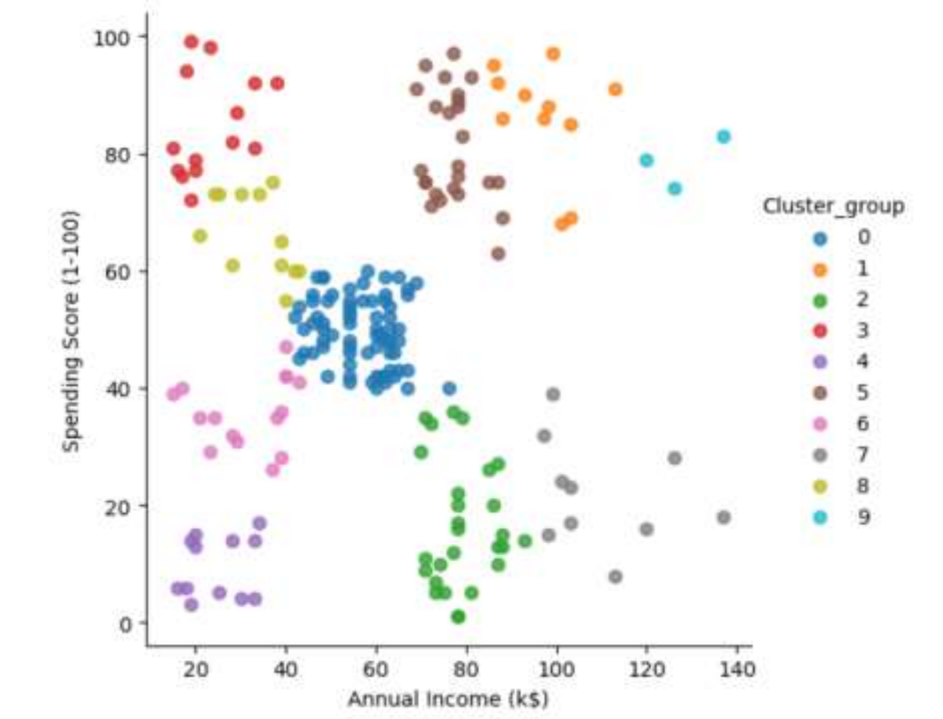
```
FROM SKLEARN.CLUSTER IMPORT KMEANS

KMEANS = KMEANS(N_CLUSTERS = I, INIT = 'K-MEANS++’,
                              RANDOM_STATE = 42)

Y_KMEANS = KMEANS.FIT_PREDICT(X)
```

- ADVANTAGES:
- - SIMPLE, FAST, SCALABLE FOR LARGE DATASETS.
- DISADVANTAGES:
- - MUST PREDEFINE K, SENSITIVE TO OUTLIERS AND INITIALIZATION.

VISUALIZATION OF K-MEANS CLUSTERING



AGGLOMERATIVE CLUSTERING

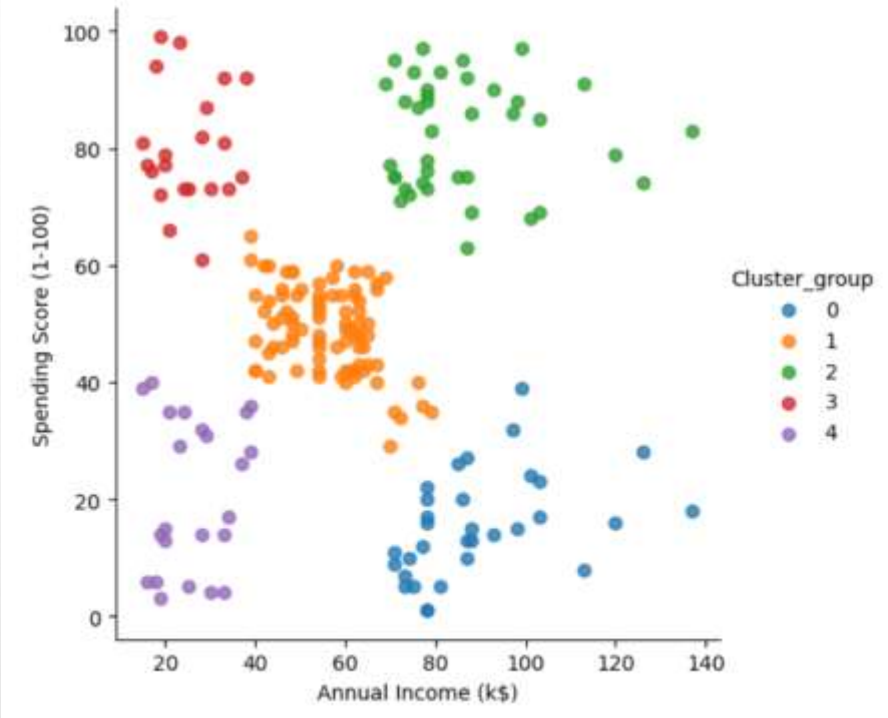
- A HIERARCHICAL, BOTTOM-UP CLUSTERING METHOD THAT MERGES CLUSTERS STEP BY STEP. PRODUCES A DENDROGRAM FOR VISUALIZATION.

```
FROM SKLEARN.CLUSTER IMPORT
AGGLOMERATIVECLUSTERINGCLUSMODEL =
AGGLOMERATIVECLUSTERING(N_CLUSTERS = 5)

LABEL = CLUSMODEL.FIT_PREDICT(X)
```

- ADVANTAGES:
- - NO NEED TO PREDEFINE NUMBER OF CLUSTERS, CAPTURES HIERARCHY.
- DISADVANTAGES:
- - COMPUTATIONALLY EXPENSIVE ($O(N^2)$), SENSITIVE TO NOISE.

VISUALIZATION OF AGGLOMERATIVE CLUSTERING



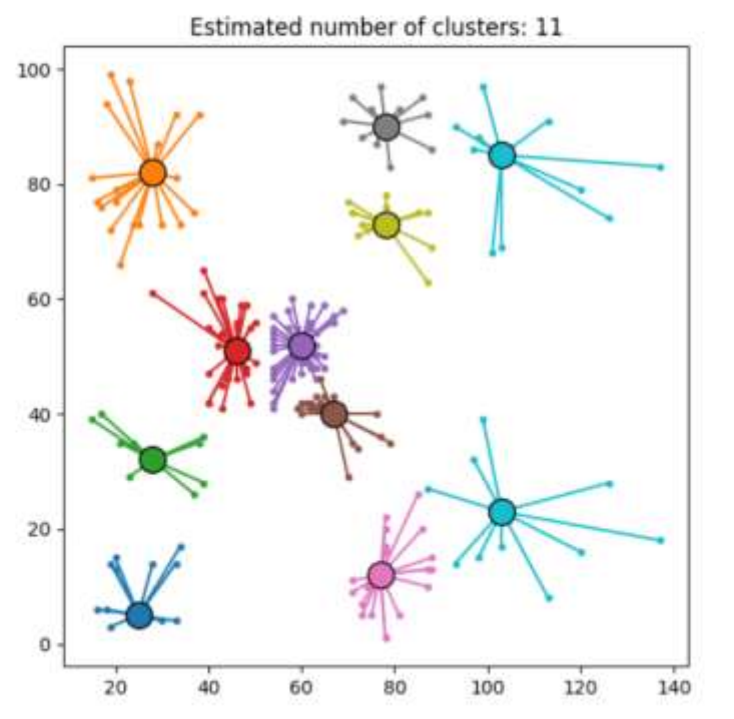
AFFINITY PROPAGATION CLUSTERING

- IDENTIFIES EXEMPLAR POINTS AS CLUSTER CENTERS USING MESSAGE-PASSING BETWEEN DATA POINTS. DOES NOT REQUIRE K.

```
FROM SKLEARN.CLUSTER IMPORT AFFINITYPROPAGATIONAFF =
AFFINITYPROPAGATION(RANDOM_STATE=5)
Y_AFF=AFF.FIT_PREDICT(X)
```

- ADVANTAGES:
- - AUTOMATICALLY FINDS NUMBER OF CLUSTERS, FLEXIBLE.
- DISADVANTAGES:
- - HIGH MEMORY AND CPU COST, MAY FORM MANY SMALL CLUSTERS.

VISUALIZATION OF AFFINITY PROPAGATION



MEAN-SHIFT CLUSTERING

- FINDS CLUSTERS BY SHIFTING POINTS TOWARDS REGIONS OF HIGH DATA DENSITY. WORKS WELL FOR NON-SPHERICAL CLUSTERS.

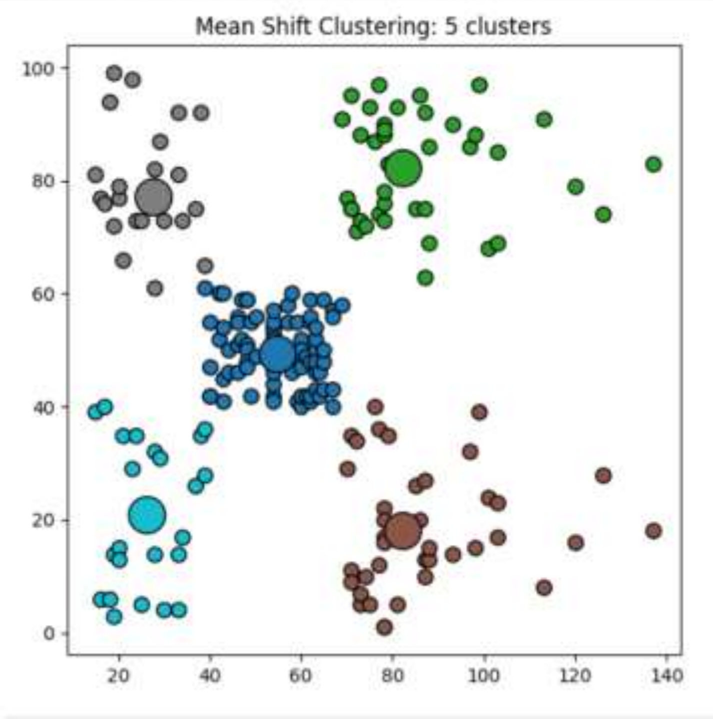
```
FROM SKLEARN.CLUSTER IMPORT MEANSHIFT

MS = MEANSHIFT(BANDWIDTH=25).FIT(X)

Y_MS = MS.FIT_PREDICT(X)
```

- ADVANTAGES:
- - NO NEED FOR NUMBER OF CLUSTERS, DETECTS ARBITRARY SHAPES.
- DISADVANTAGES:
- - COMPUTATIONALLY EXPENSIVE, SENSITIVE TO BANDWIDTH CHOICE.

VISUALIZATION OF MEAN SHIFT



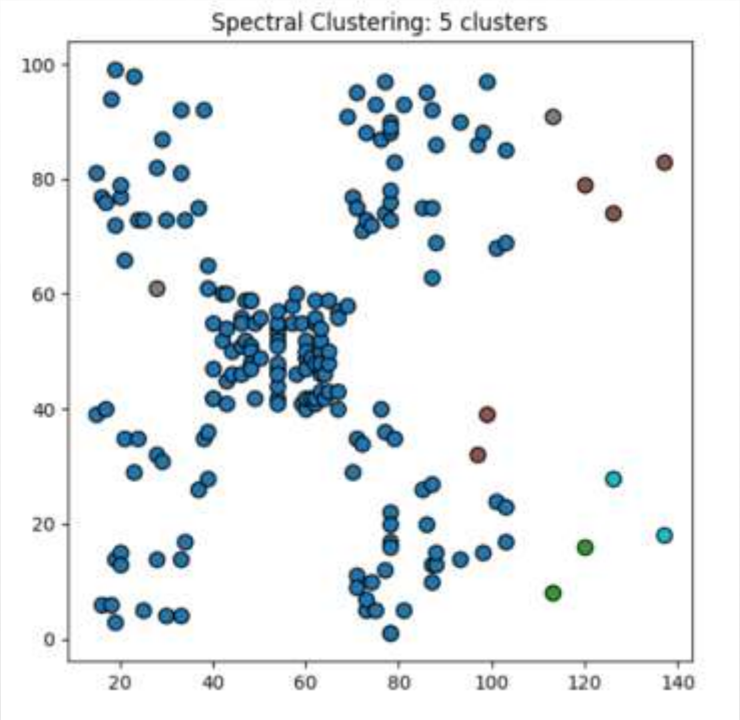
SPECTRAL CLUSTERING

- USES GRAPH LAPLACIAN AND EIGENVALUES TO TRANSFORM DATA BEFORE CLUSTERING. CAPTURES COMPLEX STRUCTURES.

```
FROM SKLEARN.CLUSTER IMPORT SPECTRALCLUSTERING
SC = SPECTRALCLUSTERING(N_CLUSTERS = 5, ASSIGN_LABELS = 'DISCRETIZE', EIGEN_SOLVER =  'ARPACK', RANDOM_STATE = 0)
Y_SC = SC.FIT_PREDICT(X)
```

- ADVANTAGES:
- - WORKS FOR NON-LINEAR CLUSTER BOUNDARIES, FLEXIBLE.
- DISADVANTAGES:
- - NEEDS PREDEFINED K, COMPUTATIONALLY HEAVY FOR LARGE DATA.

VISUALIZATION OF SPECTRAL CLUSTERING



DBSCAN DENSITY-BASED CLUSTERING

- GROUPS DENSE REGIONS TOGETHER AND LABELS SPARSE POINTS AS NOISE. SUITABLE FOR IRREGULARLY SHAPED CLUSTERS.

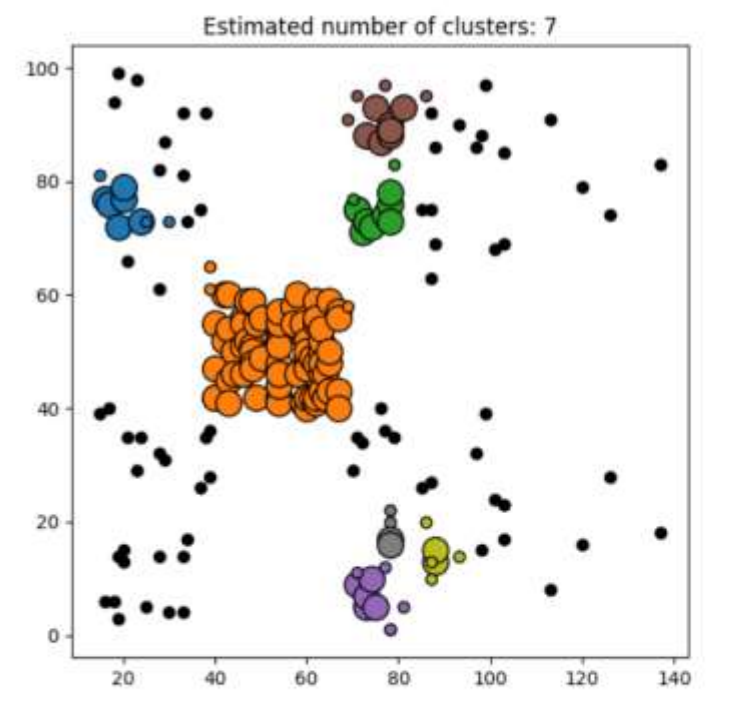
```
FROM SKLEARN.CLUSTER IMPORT DBSCAN

DB = DBSCAN(EPS=6, MIN_SAMPLES=5)

Y_DB = DB.FIT_PREDICT(X)
```

- ADVANTAGES:
- - NO NEED TO SPECIFY CLUSTERS, HANDLES NOISE, ARBITRARY SHAPES.
- DISADVANTAGES:
- - STRUGGLES WITH VARYING DENSITY, NEEDS CAREFUL PARAMETER TUNING.

VISUALIZATION OF DBSCAN



OPTICS – ORDERING POINTS TO IDENTIFY THE CLUSTERING STRUCTURE

- EXTENSION OF DBSCAN THAT ORDERS POINTS BY REACHABILITY TO DETECT CLUSTERS AT VARYING DENSITIES.

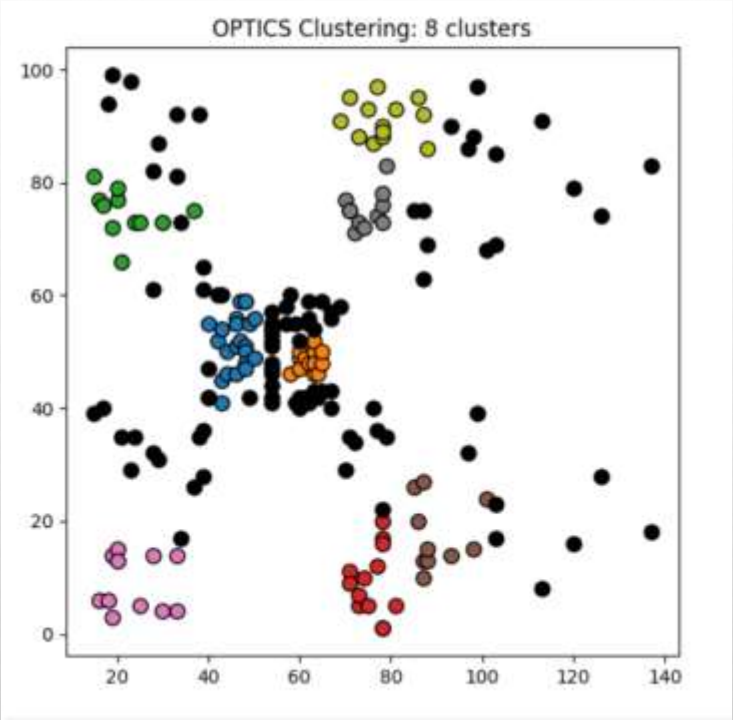
```
FROM SKLEARN.CLUSTER IMPORT OPTICS

OP = OPTICS(MIN_SAMPLES=5, XI = 0.05,
MIN_CLUSTER_SIZE =            0.05).FIT(X)

Y_OP = OP.FIT_PREDICT(X)
```

- ADVANTAGES:
- - HANDLES VARIABLE DENSITY CLUSTERS, NO FIXED K.
- DISADVANTAGES:
- - MORE COMPLEX AND COMPUTATIONALLY INTENSIVE, HARDER TO INTERPRET.

VISUALIZATION OF OPTICS



BIRCH - (BALANCED ITERATIVE REDUCING AND CLUSTERING USING HIERARCHIES)

- USES A CF TREE TO INCREMENTALLY CLUSTER VERY LARGE DATASETS EFFICIENTLY. SUMMARIZES DATA FOR CLUSTERING.

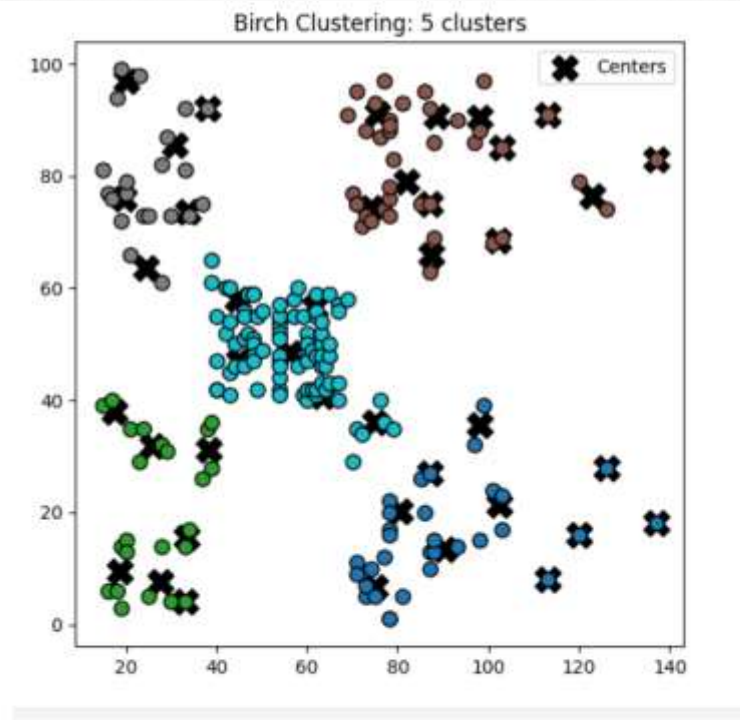
```
FROM SKLEARN.CLUSTER IMPORT BIRCH

BRC = BIRCH(THRESHOLD = 5.0, N_CLUSTERS = 5)

Y_BRC = BRC.FIT_PREDICT(X)
```

- ADVANTAGES:
- - SCALES WELL TO MASSIVE DATA, MEMORY EFFICIENT.
- DISADVANTAGES:
- - BEST FOR SPHERICAL CLUSTERS, SENSITIVE TO THRESHOLD PARAMETERS.

VISUALIZATION OF BIRCH


Birch Clustering: 5 clusters
Centers