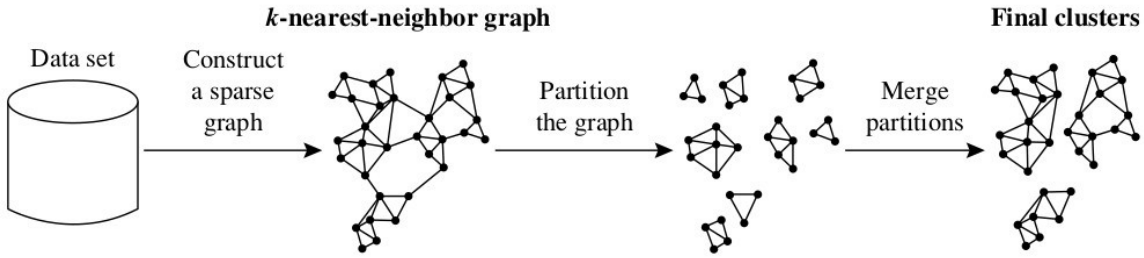
HW1 REPORT

CANBERK ARICI
171044062

1- Implementation is in 171044062_canberk_arici.py

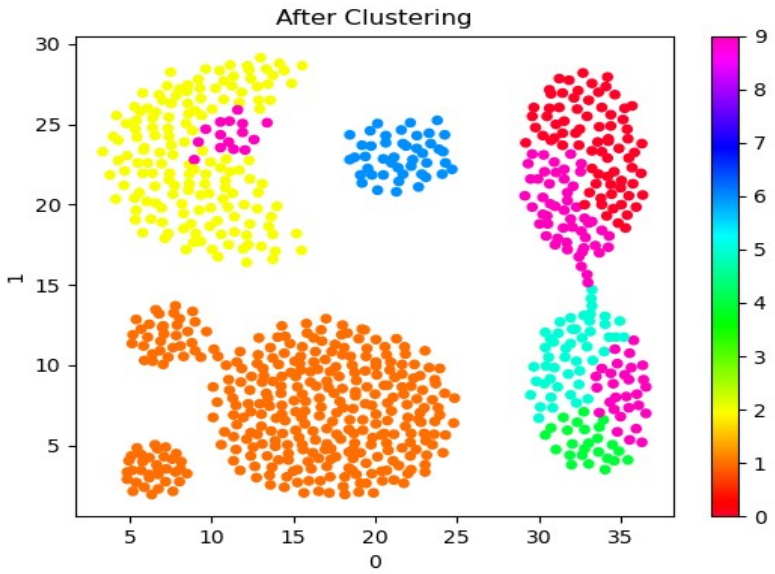


I implemented Chameleon Clustering Algorithm as explained in the book.

Chameleon is a hierarchical clustering algorithm that uses dynamic modeling to determine the similarity between pairs of clusters. In Chameleon, cluster similarity is assessed based on how well connected objects are within a cluster and the proximity of clusters. That is, two clusters are merged if their interconnectivity is high and they are close together. Thus, Chameleon does not depend on a static, user-supplied model and can automatically adapt to the internal characteristics of the clusters being merged. The merge process facilitates the discovery of natural and homogeneous clusters and applies to all data types as long as a similarity function can be specified.

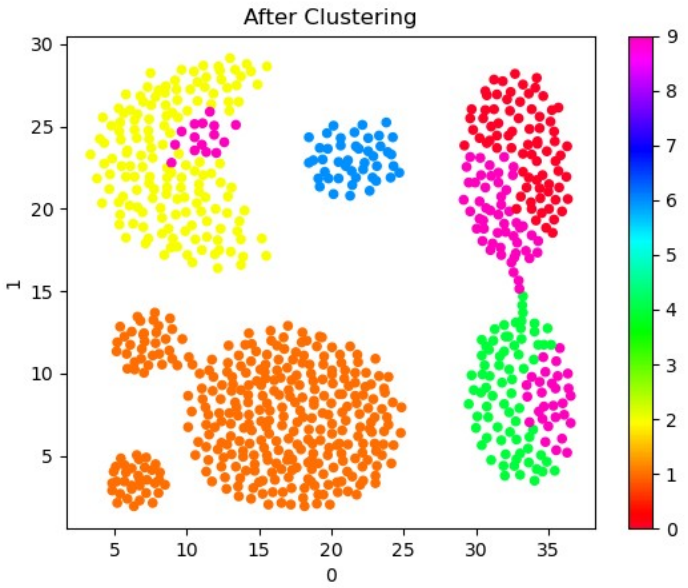
2-

Extracted clusters for knn value = 20 and number of clusters = 7



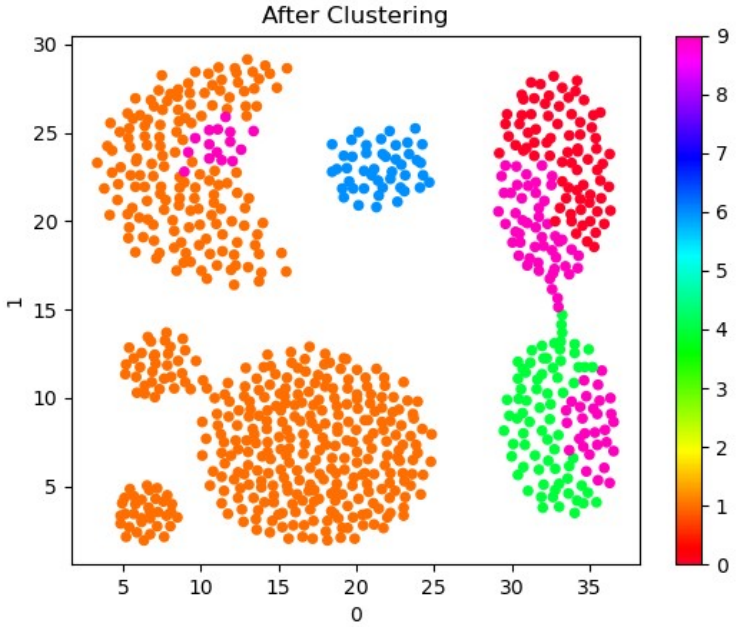
Colors of clusters: **Yellow, Blue, Pink, Orange, Red, Turquoise, Green**

Extracted clusters for knn value = 20 and number of clusters = 6



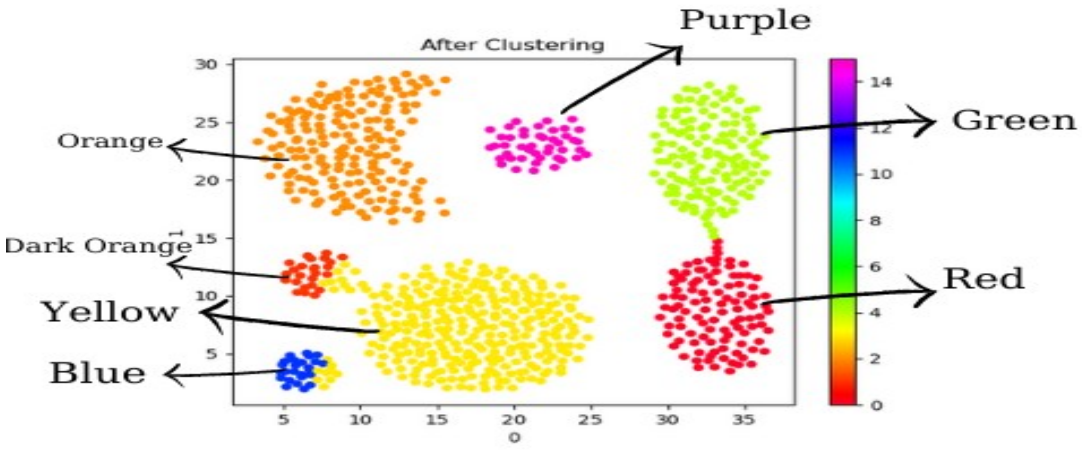
Colors of clusters: **Yellow, Blue, Pink, Orange, Red, Green**

Extracted clusters for knn value = 20 and number of clusters = 5



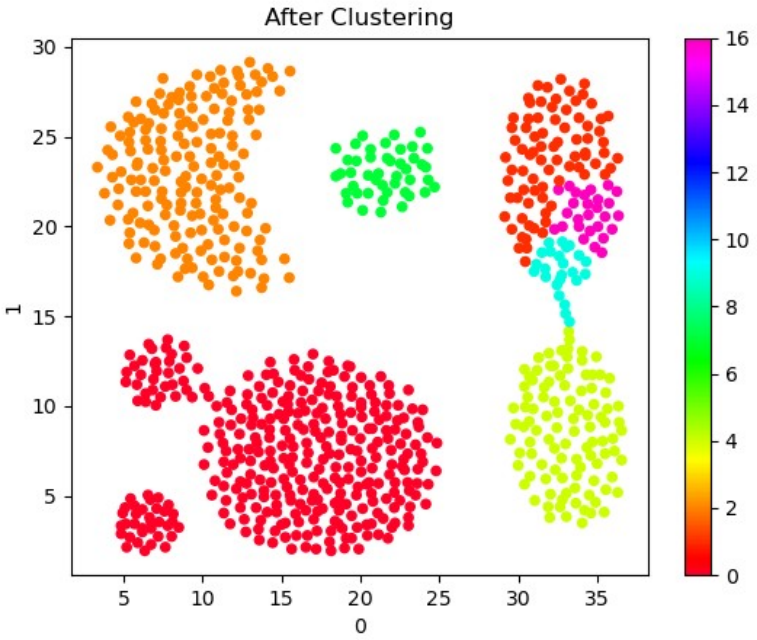
Colors of clusters: **Blue, Pink, Orange, Red, Green**

Extracted clusters for knn value = 10 and number of clusters = 7



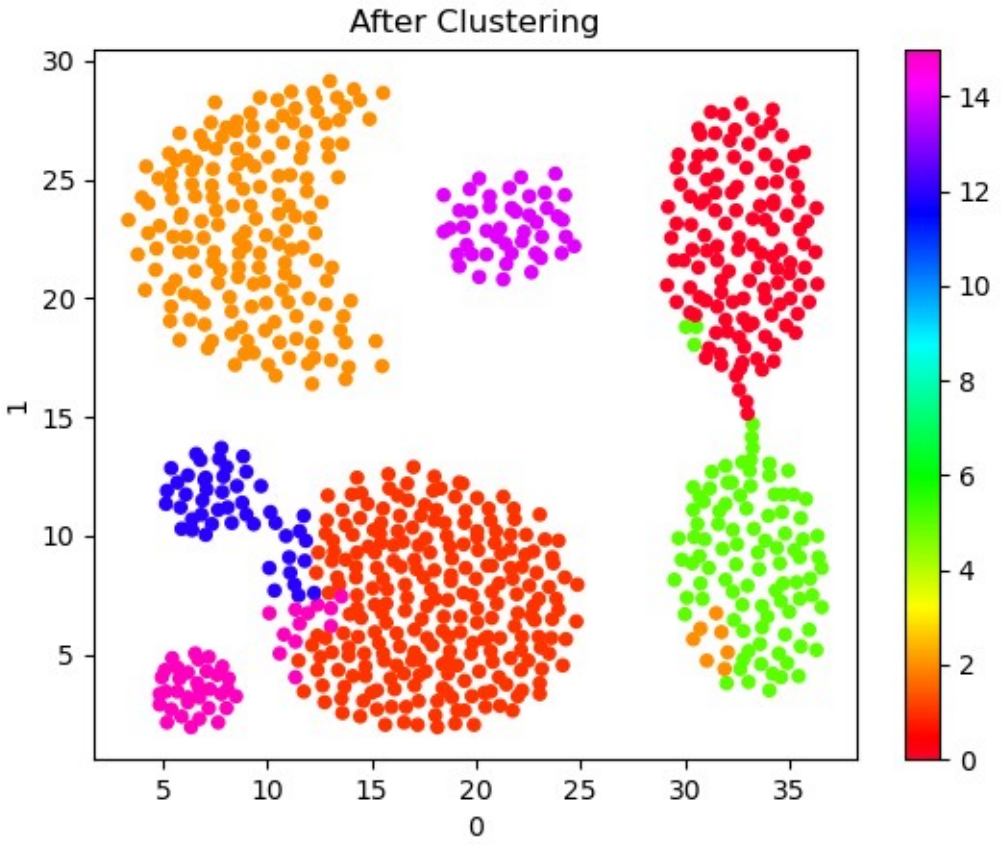
Colors of clusters: **Yellow, Blue, Dark Orange, Orange, Red, Purple, Green**

Extracted clusters for knn value = 15 and number of clusters = 7



Colors of clusters: **Yellow, Blue, Pink, Orange, Red, Turquoise, Green**

Extracted clusters for knn value = 30 and number of clusters = 7



Colors of clusters: Light Orange, Blue, Dark Orange(connected to light purple and blue), Light Purple, Dark Purple(between light orange and red), Red, Yellow

3- Knn value is for computing k-nearest neighbor graph. When knn value gets bigger, size of clusters gets bigger as I observed from figures. Algorithm is very sensitive to the value of k for spherical shapes of clusters therefore it is very important to choose an optimal value of k. If value of number of clusters increases then number of clusters increases.

4-

Advantages:

- Chameleon can discover arbitrary clusters of any shape and size.
- Chameleon uses relative inter-connectivity and relative closeness to select the most similar pairs of clusters therefore it has a very good clustering quality.
- Chameleon can find clusters of different shapes, densities, and sizes in two-dimensional space.
- Works on data of all attributes

Disadvantages:

- Handling large databases by using Chameleon can be inefficient. BIRCH can handle noise but it cannot handle very large amount of data properly as Chameleon cannot.
- Chameleon does not handle noise like DBSCAN.

5-

Time complexity of Chameleon depends on the time which it requires to make a KNN Graph. Chameleon's worst-case time complexity is O(n(log2 n + m)), m is the number of clusters after completion of the first phase of the algorithm. For high dimensional datasets, the amount of time required to find the k-nearest neighbors of a data item is O(n), leading to an overall complexity of $O(n^2)$. BIRCH has O(n) time complexity. Chameleon has one of the worst time-complexity among other clustering algorithms.

Method	Algorithm	Scalability	Cluster shape	Outliers sensitivity	Time complexity
Hierarchical	BIRCH	Active	Spherical	No	O(n)
	CURE	Active	Arbitrary	No	$O(n^2 \log n)$
	ROCK	Moderate	Arbitrary	No	$O(n^2 \log n)$
	Chameleon	Active	Arbitrary	No	$O(n^2)$
Partitioning	K-means	Moderate	Spherical	Yes	O(n k d)
	K-medoids	Passive	Spherical	No	$O(k(n-k)^2)$
Density-based	DBSCAN	Moderate	Arbitrary	No	O(n log n)
	OPTICS	Moderate	Arbitrary	No	O(n log n)
Grid-based	STING	Active	Arbitrary	No	O(n)
	CLIQUE	Active	Spherical	Yes	$O(n+d^2)$