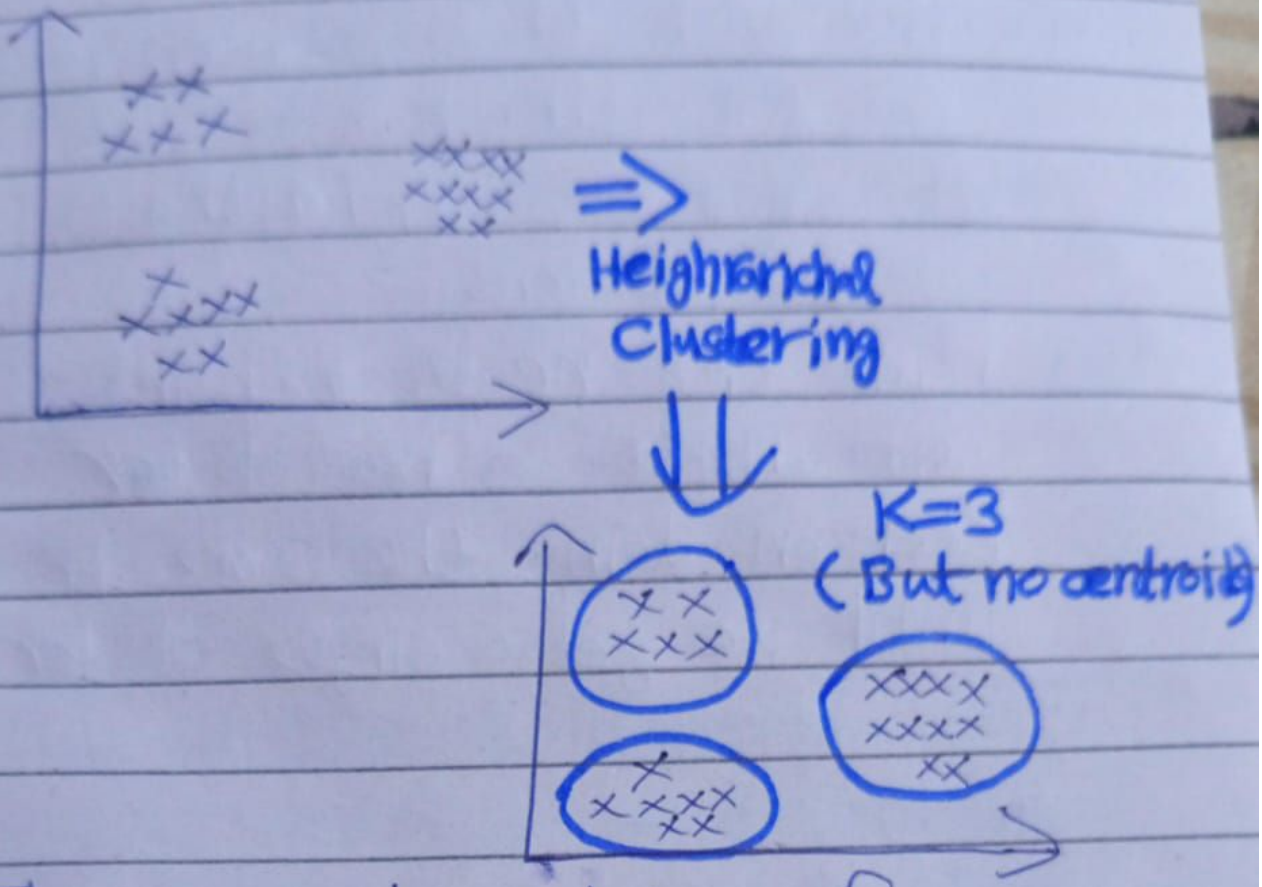
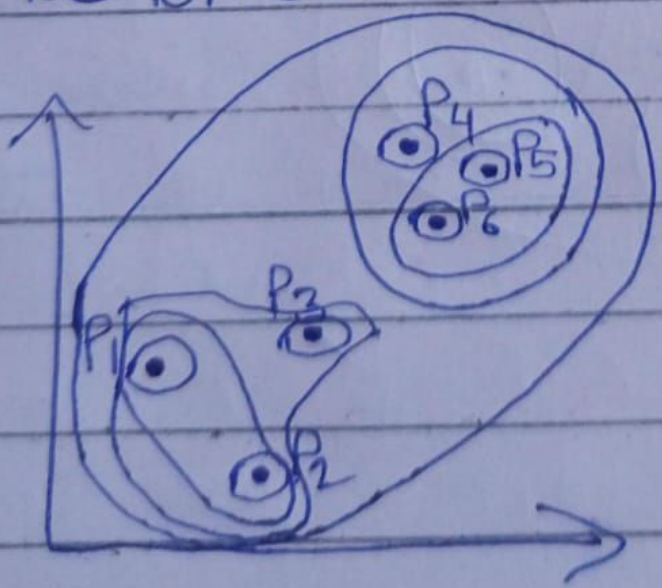
⇒ Heirarchical Clustering:



There are two types of Heirarchical clustering:

1. Algomerative
2. Devisive

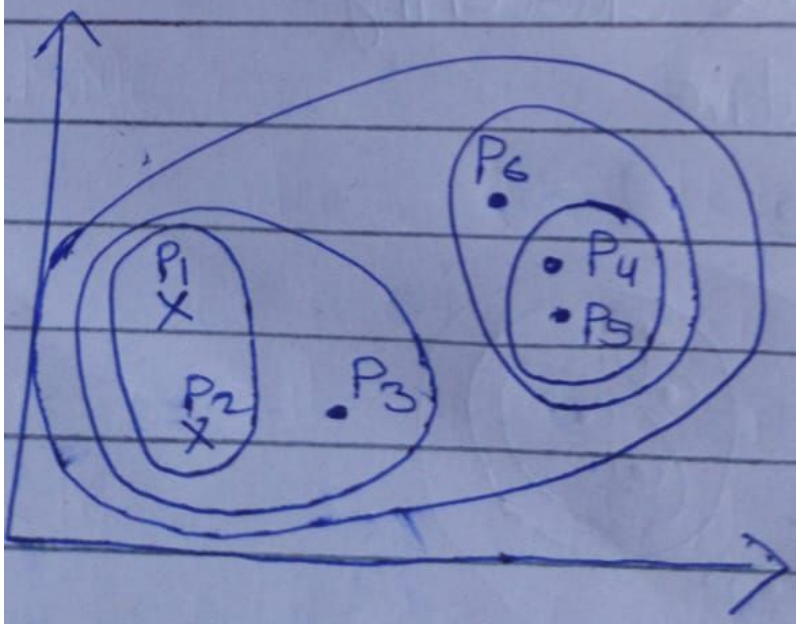


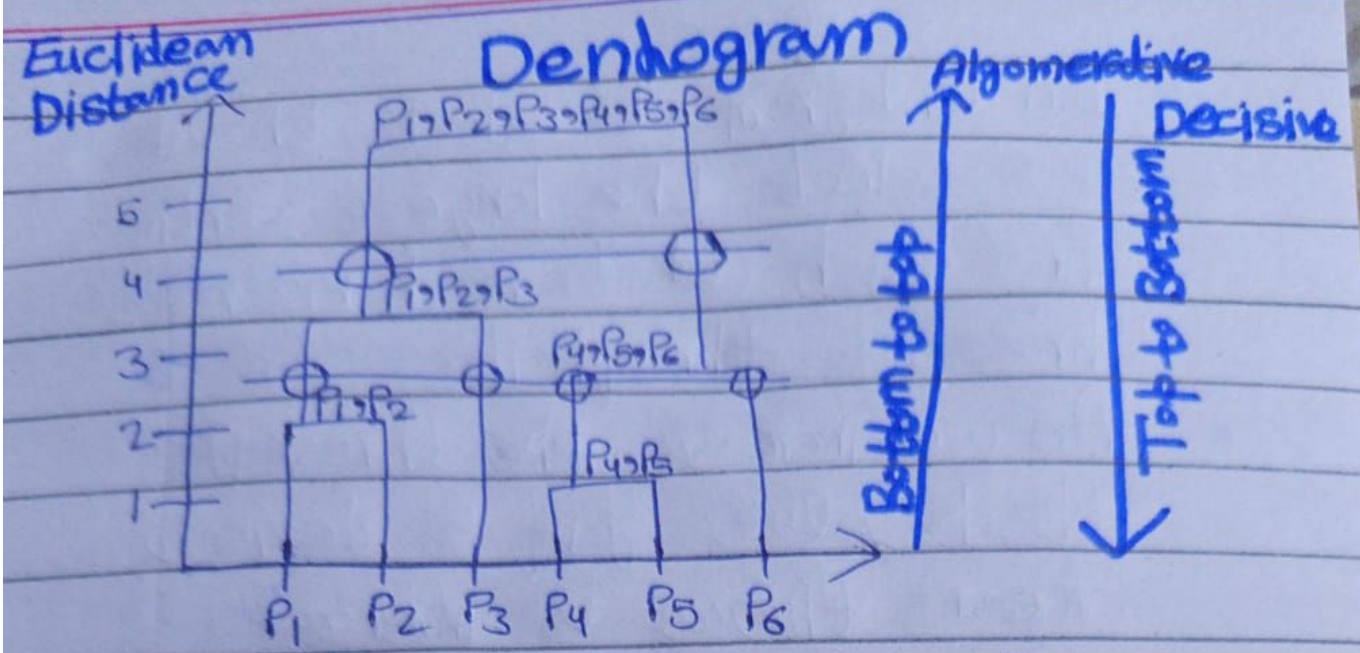
Steps:

1. For each datapoint initially we will consider it as a seperate cluster i.e there are 6 clusters
2. Find the nearest point and create a new cluster
3. Keep on doing the same process until we get a single cluster

This approach is called algomerative approach

Dendogram:





- We will find nearest points using Euclidean or Manhatann Distance
- How we will find out how many clusters are there i.e what is the value of K?

⇒ To find number of clusters or value of K we set a threshold value of Eudidean distance i.e if my threshold=4 K=2 as horizontal line passing thrugh 2 point and if my threshold=2.5 K=4.

⇒ In dendogram, to select threshold we select the longest vertical line such that no horizontal line pass through it.

- As we increase the threshold the K value going to decrease means there is inverse relationship b/w threshold and K-value.

⇒ KMeans vs Heirarchical Clustering:

1. Dataset Size
   - → Huge → KMeans
   - → Small → Heirarchical

2. KMeans is only for numerical data where [redacted] Heirarchical can be used for non-numeric even, as it can use cosine similarity. KMeans use centroid and to find number of clusters use elbow method which can be

difficult where Heirarchical clustering use Dendogram

⇒ DBSCAN Clustering

Intuition in Repo

- Pros and Cons of DBSCAN Clustering:

Advantages:

1. DBSCAN doesn't require one to specify the number of clusters in data
2. DBSCAN can find arbitrarily-shaped clusters. It can even find a cluster completely surrounded by a different cluster
3. DBSCAN has a notion of noise and robust to outliers
4. DBSCAN is designed for use with databases that can accelerate region queries e.g using R* tree

Disadvantages:

1. DBSCAN is not entirely deterministic: border points that are reachable from more than one cluster can be part of either cluster
2. The quality of DBSCAN depends on distance measure i.e Euclidean Distance so the quality may vary.
3. DBScan cannot cluster dataset with large difference in densities
4. If data and scale are not well understood, choosing threshold (radius) Ecan be difficult To overcome this we standardize data.