K means Clustering – Introduction

> K-Means Clustering is an Unsupervised Machine Learning algorithm, which groups the unlabeled dataset into different clusters.

K-means Clustering

> Unsupervised Machine Learning is the process of teaching a computer to use unlabeled, unclassified data and enabling the algorithm to operate on that data without supervision.

> Without any previous data training, the machine’s job in this case is to organize unsorted data according to parallels, patterns, and variations.

> K means clustering, assigns data points to one of the K clusters depending on their distance from the center of the clusters.

> It starts by randomly assigning the clusters centroid in the space. Then each data point assign to one of the cluster based on its distance from centroid of the cluster.

> After assigning each point to one of the cluster, new cluster centroids are assigned. This process runs iteratively until it finds good cluster.

> In the analysis we assume that number of cluster is given in advanced and we have to put points in one of the group.

Objective of k-means clustering

> The goal of clustering is to divide the population or set of data points into a number of groups so that the data points within each group are more comparable to one another and different from the data points within the other groups.

> It is essentially a grouping of things based on how similar and different they are to one another.

Important points

In [ ]:

```
1) Select the value of k i.e k=8 deafult
2) Select random value of k clusters centroids
3) Calculte the distance between centroids and all datapoints using eucledian distance formula
4) Assign each datapoints to closet clusters / centroid
5) Update the value of centroid by using mean of clusters datapoints
6) Keep iterating untill and unless there will be no change in centroid value or no movement in datapoints from clusters to clust
To know number of clusters we used Elbow method and which used concept of WCSS (Withing clusters sum of squared distance)
>>. to know how well are clusters we used silhoutte score >> range -1 to +1
```