WHAT IS K-MEANS ALGORITHM?

K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems. It follows a simple procedure of classifying a given data set into a number of clusters, defined by the letter "k," which is fixed beforehand. The clusters are then positioned as points and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process starts over using the new adjustments until a desired result is reached.

K-means clustering has uses in search engines, market segmentation, statistics and even astronomy.

HOW TO IMPLEMENT K-MEANS ALGORITHM

1. Start with K centroids by putting them at random place.

2. Compute distance of all points from centroid and cluster them accordingly

3. Adjust centroids so them become center of gravity for given cluster

4. Re-cluster all points based on their distance with centroid, again

5. Adjust centroids, again

6. Repeat step 4 and step 5 until data points stop changing clusters

HOW TO DETERMINE OPTIMAL NUMBER OF CLUSTER (K) FOR DATASET?

A method called ellow is used to find it.

In this method, you will try to compute sum of squared error in many case k

=

=

Draw a plot with y axis is SSE and x axis is k, and the elbow in that plot is the best number of cluster.