

| **Title: Implementation of k-means clustering algorithm** |
| --- |

**Objective:** To implement k-means clustering algorithm.

**Expected Outcome of Experiment:**

CO3: Analyse unsupervised learning methods.

**Books/ Journals/ Websites referred:**

<https://www.geeksforgeeks.org/k-means-clustering-introduction/>

**Theory of k-means clustering algorithm**

K-means clustering is a popular unsupervised machine learning algorithm used for grouping similar data points together. Here's a breakdown of the theory behind it:

**Objective:**

The goal of k-means clustering is to partition a dataset into a predefined number of groups (k) such that the points within a group are similar to each other and dissimilar to points in other groups. This similarity is typically measured by distance, often the squared Euclidean distance. The algorithm aims to minimize the within-cluster sum of squares (WCSS), which is the total squared distance between each data point and its assigned cluster's centroid (mean).

**Process:**

K-means is an iterative process that follows these general steps:

1. **Initialization:**
   * Define the number of clusters (k). This is a crucial step as it dictates the outcome.
   * Choose k initial centroids, which represent the center points of each cluster. There are various methods for centroid initialization, with k-means++ being a popular strategy that helps spread out the centroids for better convergence.
2. **Assignment:**
   * For each data point, calculate the distance to all centroids.
   * Assign the data point to the cluster with the nearest centroid.
3. **Recalculate Centroids:**
   * Once all data points are assigned, recompute the centroid (mean) of each cluster based on the data points currently assigned to it.
4. **Repeat:**
   * Iterate through steps 2 and 3 until a stopping criterion is met. This criterion is often when the centroids no longer move significantly between iterations, indicating convergence.

**Key Points:**

* K-means works best with numerical data and spherical clusters.
* It's sensitive to the initial placement of centroids, which can lead to local minima and affect the final clustering.
* The number of clusters (k) needs to be predetermined, and choosing the right k can be challenging. There are methods to evaluate the quality of clustering for different k values, but it often involves domain knowledge and experimentation..

**Dataset details used in the Experiment** Sample dataset of points was used.

**Explanation of API/Tool used for implementation** - Coding done in pure python without any APIs

**Source Code** Attached in separate file.

**Results/Output** Attached in separate file.

**Conclusion:** We have implemented K menas clustering algorithm in Python. We have done so without using any external libraries. By implementing everything from scratch, we have understood how K means algorithm works. K-means clustering is a simple and efficient algorithm with a wide range of applications in various fields, including customer segmentation, image compression, and anomaly detection