**Data Analytics Laboratory**

**Task 4**

**Clustering using K-Means Algorithm**

**Introduction**

* Cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as a clustering.
* Clustering can be used to discover clusters or “subclasses” in handwritten character recognition systems.
* Clustering is known as unsupervised learning because the class label information is not present. For this reason, clustering is a form of learning by observation, rather than learning by examples.
* Suppose a data set, *D*, contains *n* objects in Euclidean space. Partitioning methods distribute the objects in D into k clusters, *C1,...,Ck*, that is, *Ci ⊂D* and *Ci∩Cj=∅* for *(1 ≤ i, j ≤ k).*
* A picture containing diagram

  Description automatically generated

**Prerequisites**

List the steps in performing k-means clustering.

How do you decide the number of clusters needed for the given dataset?

How do you find the distance between two points in Euclidean space? List some alternative methods.

**Clustering-Exercise Problem.**

For the given practise problem dataset ([*Mall\_Customer.csv*](https://github.com/ArunkumarGoge/DataAnalyticsLab/blob/master/Exercise%204/Mall_Customers.csv)) consider all the attribute and perform k-means clustering.

1. Use appropriate data pre-processing techniques on the dataset. List the pre-processing techniques used for each attribute.
2. How many number of clusters can be formed for the dataset.
3. What is the WCSS value for 2 and 3 clusters.
4. How many datapoint are in each clusters.

**Results**

The program is implemented in python and the output is observed.

**Faculty Signature**