CUSTOMER SEGMENTATION

# **SUBJECT:** Customer Segmentation using K-means Clustering algorithm.

The most fundamental steps required to implement K-means Clustering in the field of Customer Segmentation are as follows -

## STEP 1:- Understanding K-means Clustering Algorithm.

**K-means Clustering** is an **unsupervised machine learning algorithm** that divides **n** observations into **k** groups or clusters based on similar characteristics or attributes. It is an **exclusive clustering technique**, hence a particular observation or data point can be assigned to only one cluster.

## STEP 2:- Gathering and Cleaning Data.

**Dataset** relevant to the customer segmentation is collected by a method called **web scraping**. This dataset may include **demographic (gender, age, income)** and **geographic (location)** information about the **customers** as well as the **products** they have bought.

Now this data set is cleaned to remove any missing or inconsistent data.

## STEP 3:- Deciding the Attributes on which Clustering should be done.

It is assumed that a dataset from a supermarket has the following attributes:

## Customer ID

Labelled as 1, 2, 3, 4, …

## Gender

Possible values-

* + Male
  + Female

## Area

Possible values-

* + Town
  + City

## Location

Possible values-

* + 1
  + 2
  + 3
  + 4
  + 5
  + 6

Where 1,2,3 belongs to Town and 4,5,6 belongs to City.

## Product Type

Possible values-

* + Grocery
  + Clothing

## Product ID

Possible values-

* + 1
  + 2
  + 3
  + 4
  + 5
  + 6

Where 1,2,3 belongs to Grocery and 4,5,6 belongs to Clothing.

It is also assumed that a single customer can buy only one type of product (either grocery or clothing). In case a customer shops from both the categories that will be treated as separate data points, as our target is to check the demand.

Now it is required to analyze which products are bought by the customers more in both Town areas and City areas to increase the sales.

So **Product ID** is chosen to be plotted against **Location**.

## STEP 4:- Deciding the number of clusters ‘K’.

Based on the aforementioned dataset the customers are divided into **4 clusters. Hence k = 4.** These clusters are as follows:

* Customers from Town area buying Grocery items
* Customers from Town area buying Clothing items
* Customers from City area buying Grocery items
* Customers from City area buying Clothing items

## STEP 5:- Implementing the K-means Clustering on the Dataset (Coding).

**Python**, a high level language, has several libraries like **Pandas** and distributions like **Anaconda**. This programming tool is dedicated for **Data Science and Machine Learning Applications**, which can be used to implement the code of the k-means clustering algorithm conveniently and efficiently.

This is an **iterative** process which is continued until the **within cluster sum of squares is minimised** . After varying the number of clusters k within an estimated limit **Elbow Method** can be used to obtain the required number centroids and thus a **scatter plot** of the data can be done clearly showing the different clusters.

If the Hartigan-Wong K-means algorithm (1979) is considered, then the total within-cluster variation can be formulated as the sum of squared distances between the items and the corresponding centroid:

W(Ck)=∑xi∈Ck(xi−μk)^2(6)

where:

1. x is a data point belonging to the cluster C.
2. μk is the mean value of the points assigned to the cluster Ck.

Thus the total within-cluster variation can be defined as follows:

total\_wss=k∑k=1W(Ck)=k∑k=1∑xi∈Ck(xi−μk)2(7)

The *total within-cluster sum of square* measures the compactness of the clustering and our target will be to minimise this .

## STEP 6:- Drawing Inferences from the Clusters.

It is assumed that the cluster of **Customers from Town area buying Grocery** items is very **sparse** as compared to the other three clusters. Hence it can be concluded that the customers from town buy clothing items more than grocery items from that supermarket, whereas customers from city area buy both the types of item.

Based on this inference the supermarket can target one or more clusters to increase their sales. The buying preferences of a person based on his/her location, can also be predicted.