Experiment :

AIM : Write a program to implement K-Means clustering algorithm.

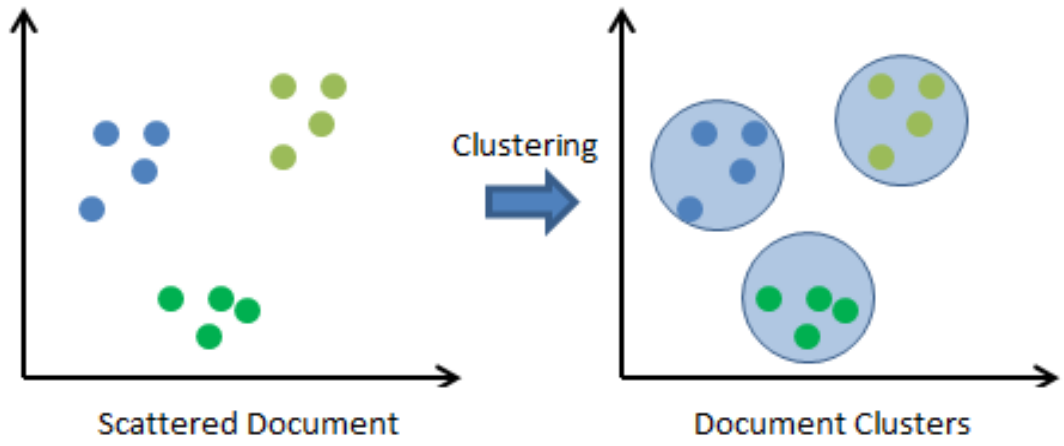
THEORY :

Clustering is a type of unsupervised learning method . An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses.

Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

For example : The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture. There may be N number of clusters.



Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present. There are no criteria for a good clustering. It depends on the user, what is the criteria they may use which satisfy their need.

For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).

This algorithm must make some assumptions which constitute the similarity of points and each assumption make different and equally valid clusters.

Types of Clustering methods:

1. **Hierarchical Based Methods :** The clusters formed in this method forms a tree type structure based on the hierarchy. New clusters are formed using the previously formed one.
   Examples: CURE (Clustering Using Representatives).

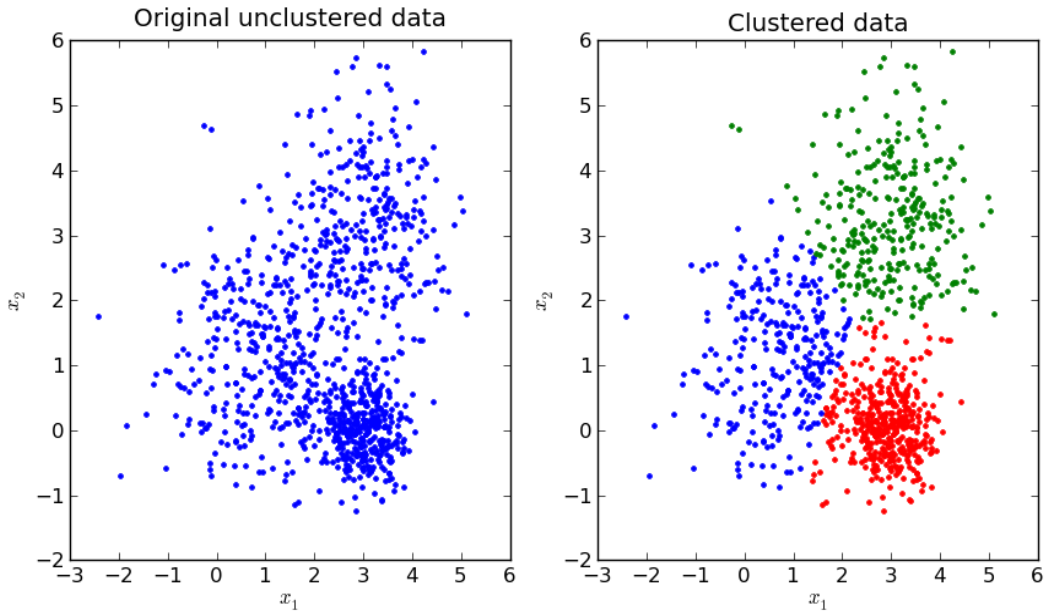
2. **Partitioning Methods :** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter.
   Examples: K-means.

3. **Density-Based Methods :** These methods consider the clusters as the dense region having some similarity and different from the lower dense region of the space. These methods have good accuracy and ability to merge two clusters.
   Examples: OPTICS (Ordering Points to Identify Clustering Structure) etc.

4. **Grid-based Methods :** In this method the data space are formulated into a finite number of cells that form a grid-like structure. All the clustering operation done on these grids are fast and independent of the number of data objects.
   Examples: wave cluster, CLIQUE (CLustering In Quest) etc.

K Means Clustering Algorithm

It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partition n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster .



Working of K means algorithm :

1. INITIALIZE: Firstly, we initialize k points, called means, randomly.
2. ASSIGNMENT: We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
3. UPDATE: We repeat the process for a given number of iterations and at the end, we have our clusters.

   The points mentioned above are called means, because they hold the mean values of the items categorized in it. To initialize these means, we have a lot of options. An intuitive method is to initialize the means at random items in the data set. Another method is to initialize the means at random values between the boundaries of the data set.

ALGORITHM :

1: Accept the number of clusters to group data into and the dataset to cluster as input values

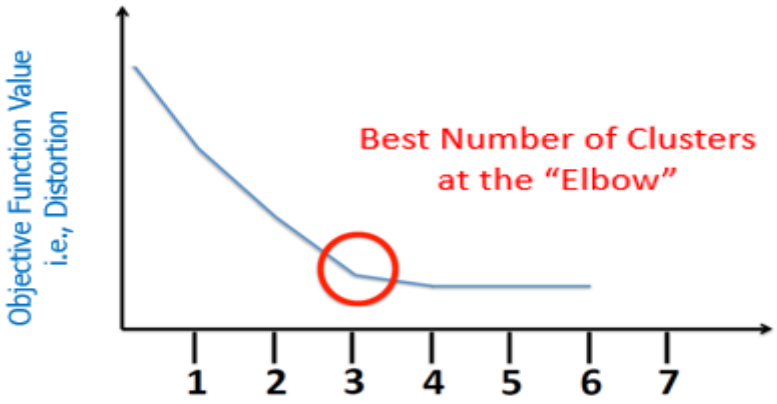
2: Initialize the first K clusters
- Take first k instances or
- Take Random sampling of k elements

3: Calculate the arithmetic means of each cluster formed in the dataset.

4: K-means assigns each record in the dataset to only one of the initial clusters
- Each record is assigned to the nearest cluster using a measure of distance (e.g Euclidean distance).

5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.

ELBOW METHOD FOR K MEANS ALGORITHM

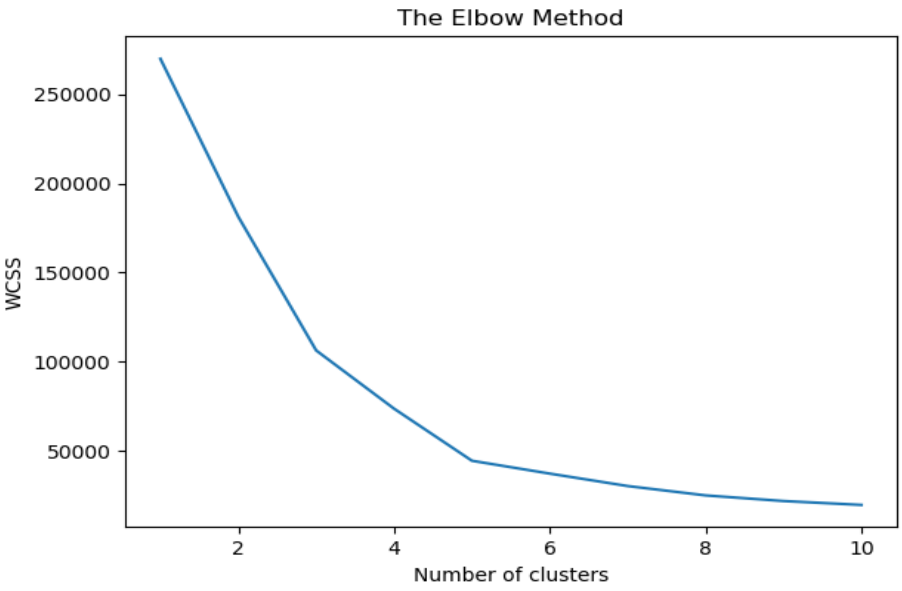


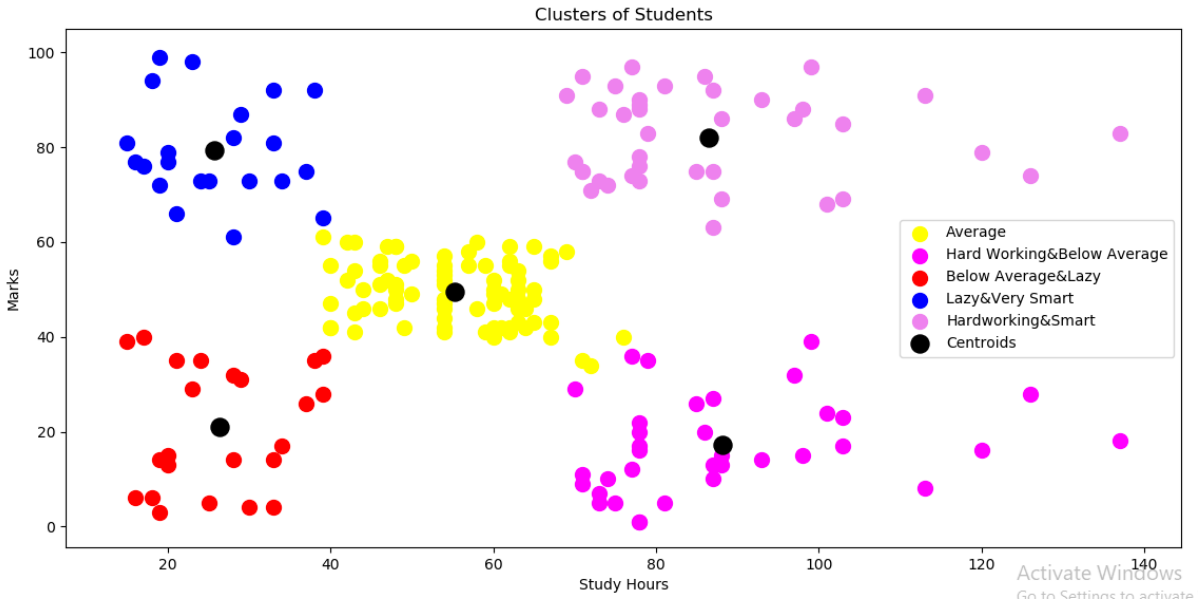
CODE :

```
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# Importing the dataset
dataset = pd.read_csv('student.csv')
X = dataset.iloc[:, [3, 4]].values


# Using the elbow method to find the optimal number of clusters
from sklearn.cluster import KMeans
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 100)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
plt.plot(range(1, 11), wcss)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()

# Fitting K-Means to the dataset
kmeans = KMeans(n_clusters = 5, init = 'k-means++', random_state = 100)
y_kmeans = kmeans.fit_predict(X)

# Visualising the clusters
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s = 100, c = 'yellow', label =
'Average')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s = 100, c = 'magenta', label = 'Hard
Working&Below Average')
plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s = 100, c = 'red', label = 'Below
Average&Lazy')
plt.scatter(X[y_kmeans == 3, 0], X[y_kmeans == 3, 1], s = 100, c = 'blue', label =
'Lazy&Very Smart')
plt.scatter(X[y_kmeans == 4, 0], X[y_kmeans == 4, 1], s = 100, c = 'violet', label =
'Hardworking&Smart')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s = 150, c = 'black',
label = 'Centroids')
plt.title('Clusters of Students')
plt.xlabel('Study Hours')
plt.ylabel('Marks')
plt.legend()
plt.show()
```

OUTPUT :

StudentID	Gender	Age	Study Hours	Marks
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14


The Elbow Method
WCSS
250000
200000
150000
100000
50000
2
4
6
8
10
Number of clusters


Clusters of Students
Marks
100
80
60
40
20
0
20
40
60
80
100
120
140
Study Hours
Average
Hard Working&Below Average
Below Average&Lazy
Lazy&Very Smart
Hardworking&Smart
Centroids
Activate Windows

DISCUSSION :

Clustering is basically a type of unsupervised learning method. It is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

FINDING AND LEARNING :

From this experiment, we learn about implementing and understanding the concept of K-means clustering algorithm by applying it using sklearn python library on a student dataset using the attributes ‘Hours Studied’ and ‘Marks Scored’ in an exam which made 5 clusters of data points classifying students in 5 clusters and thus types.