LP2

Assignment 2.

Date of Completion
27.8.2020

Date of Submission:-
30.9.2020

Title:- Visualize the clusters using suitable tool

Problem Statement :- Consider a suitable dataset. For clustering of data instances in different groups, apply different clustering techniques. Visualize the clusters using suitable tool

Learning Objective :- Understand clustering and different algorithms used for clustering data

Learning Outcomes :- Students will be able to understand different clustering methods and implement them.

~~Theor.~~ Software / Hardware requirements :- Python, numpy, sklearn (packages)

Theory:-

Clustering Algorithms:-

1. Clustering is a machine Learning technique that involves the grouping of points.

2. It is used to analyse patterns and grouping.
3. It is a unsupervised learning algorithm.

Types of ~~etc~~ clustering algorithms:-

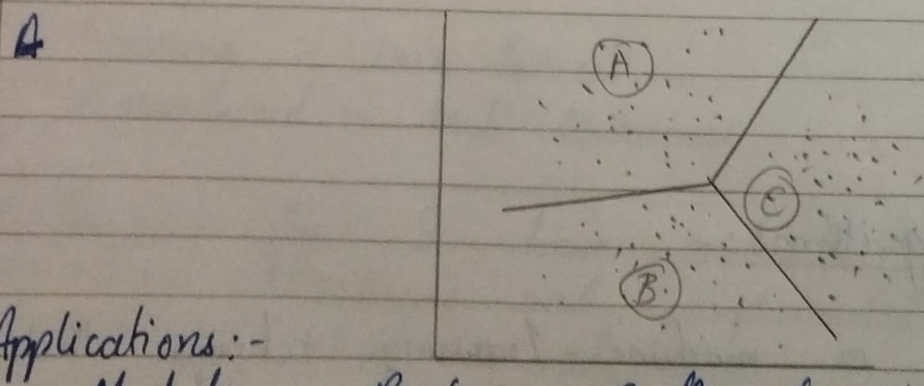
1) K Means clustering
2) Hierarchical clustering
3) Mean shift clustering
4) Fuzzy C clustering
5) spectral Clustering etc.

Clustering Methods:-

1. Density Based
2. Hierarchical Based
3. Partitioning Based
4. Grid-based

K-Means algorithm:-

1. Simplest unsupervised clustering algorithm.
2. It partions n observation into k ~~the~~ clusters where each observation belongs to clusters nearest mean serving as a prototype cluster.



Applications:-

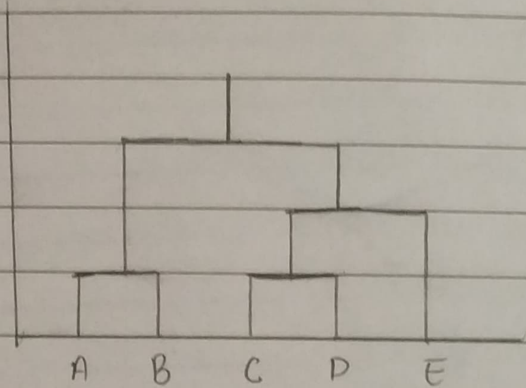
Marketing, Biology, Earthquake studies etc.

Hierarchical clustering:-

1. It builds a hierarchy of clusters.
2) 2 types :-
   a) Agglometric :- "Bottom up" approach
   b) Divisive :- "Top down" approach.
3) Results are usually presented in a dendogram.
4) Slow.

Linkage Criteria:-

To compute the distance between two similar clusters many linkage criteria have developed.



Dataset Used:-

K-means :- iris dataset
Hierarchical :- Mall Customers.

Conclusion:- Thus I have understood different clustering algorithms and implemented K-means & hierarchical clustering algorithm.

CODE:

Heriarical Clustering:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

dataset = pd.read_csv('Mall_Customers.csv')
X = dataset.iloc[:, [3, 4]].values

#using dendogram for finding optimal number of clusterings
import scipy.cluster.hierarchy as sch
dendrogram = sch.dendrogram(sch.linkage(X, method='ward'))
plt.title('Dendrogram')
plt.xlabel('Customers')
plt.ylabel('Eucladian Distances')
plt.show()

#training the cluster Using HC
from sklearn.cluster import AgglomerativeClustering
hc = AgglomerativeClustering(n_clusters = 3, affinity='euclidean', linkage='ward')
y_hc = hc.fit_predict(X)

plt.scatter(X[y_hc == 0, 0], X[y_hc == 0, 1], s = 100, c = 'red', label = 'Cluster 1')
plt.scatter(X[y_hc == 1, 0], X[y_hc == 1, 1], s = 100, c = 'blue', label = 'Cluster 2')
plt.scatter(X[y_hc == 2, 0], X[y_hc == 2, 1], s = 100, c = 'green', label = 'Cluster 3')
plt.title('Clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```

KMeans Clustering:

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

#Dataset
iris = datasets.load_iris()

# Importing the dataset
x = iris.data[:, :2]
y = iris.data[:, :2]
```

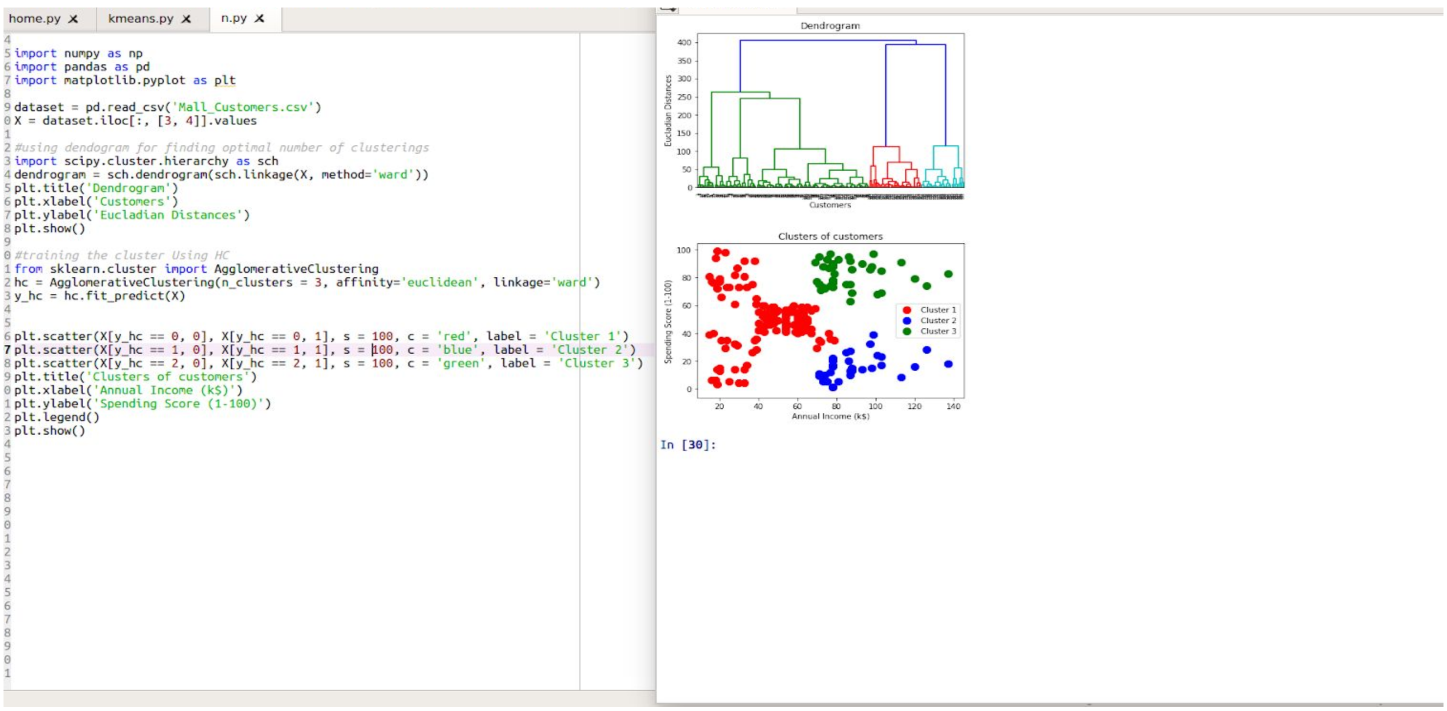
```
# Using the elbow method to find the optimal number of clusters
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=4)
y_kmeans = kmeans.fit_predict(x)

# print(y_kmeans)
kmeans.cluster_centers_

# Fitting K-Means to the dataset
plt.scatter(x[:,0], x[:,1], c=y_kmeans, cmap='gist_rainbow')
plt.xlabel('Sepal Length')
plt.ylabel('Sepal Width')
```

OUTPUT:

HIERARCHICAL CLUSTERING



K-MEANS CLUSTERING: