

Practical – 7

Name: Shivam Tawari

Roll no: A-58

Aim: Solving clustering problems using Scikitlearn

Theory:

Classification:

Clustering:

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 than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

K-Means Clustering Algorithm:

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if $K=2$, there will be two clusters, and for $K=3$, there will be three clusters, and so on.

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

DBSCAN Clustering:

Density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in data mining and machine learning.

Code & Output:

 Skills Practical 7.ipynb ☆
File Edit View Insert Runtime Tools Help

Comment Share Settings User

+ Code + Text

RAM Disk

Editing ^

▼ Practical 7

Name: Shivam Tawari

Roll no: A-58

✓ [15] In

```
import numpy as np
import pandas as pd
import plotly.graph_objects as go
from utils import set_default, plot_clusters3d
from sklearn.cluster import KMeans, DBSCAN
```

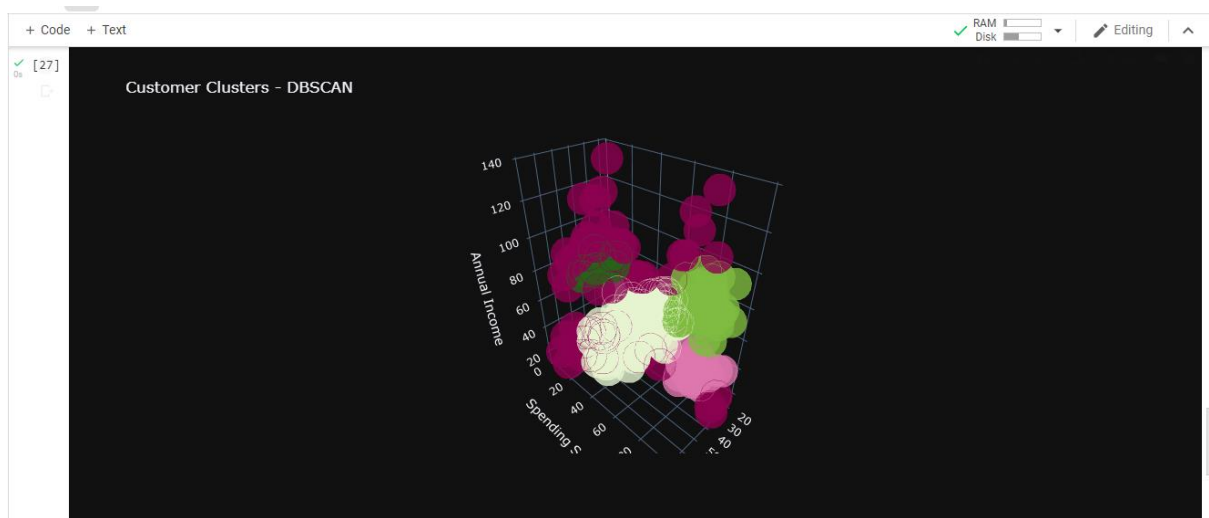
✓ [16] In

```
set_default()
```

✓ [17] In

```
df = pd.read_csv('Mall_Customers.csv')
df.head()
```

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81



Conclusion: Hence, successfully performed solving clustering problem using Scikit-learn.