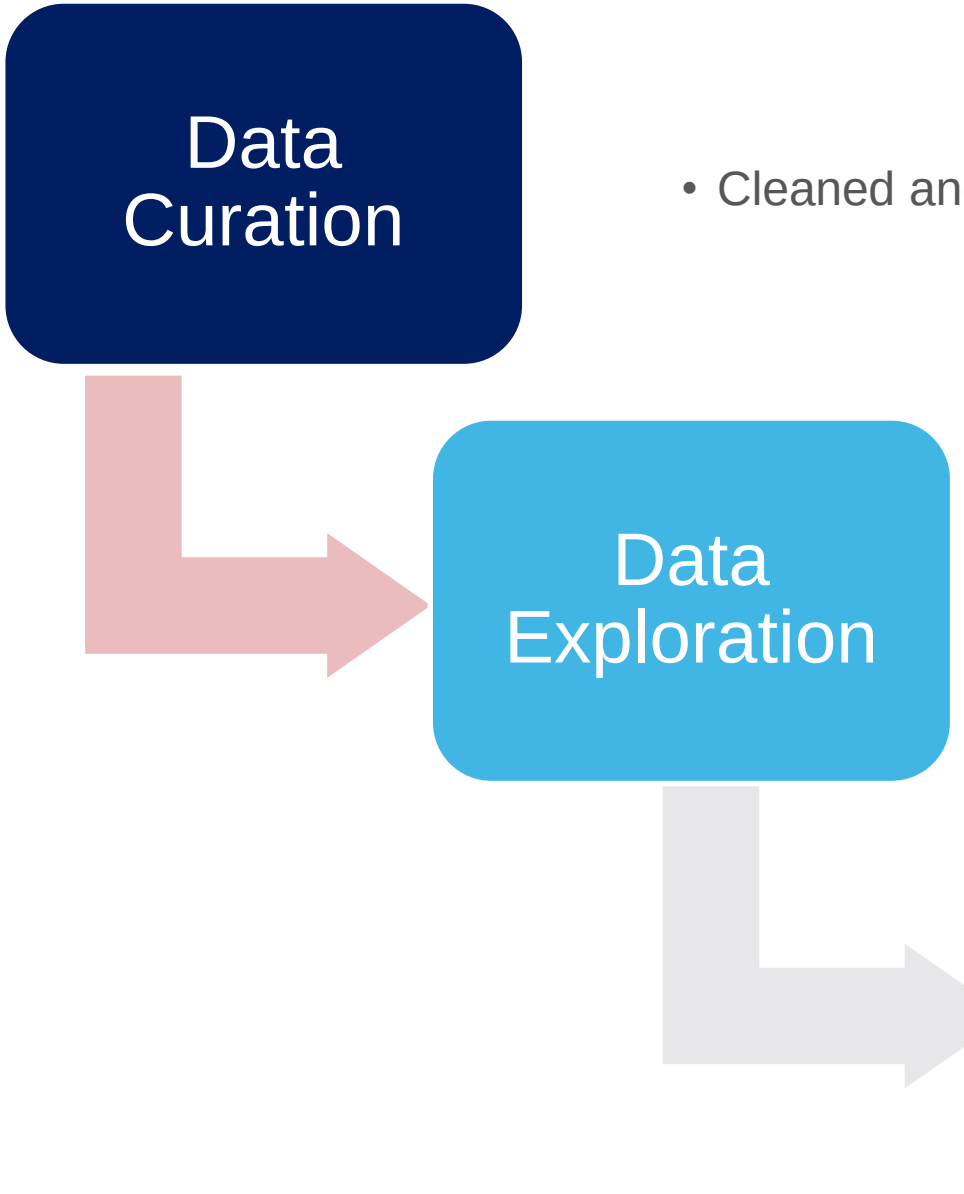
IDS 575 - Machine Learning

Customer Segmentation

Group 18:

Sangeeta Baitalik - 653896036
Suruchi Jain - 659781101
Shruti Chanda - 670617428
Jin rong Qiu - 673250167

UIC UNIVERSITY OF ILLINOIS CHICAGO



- Cleaned and analysed the data

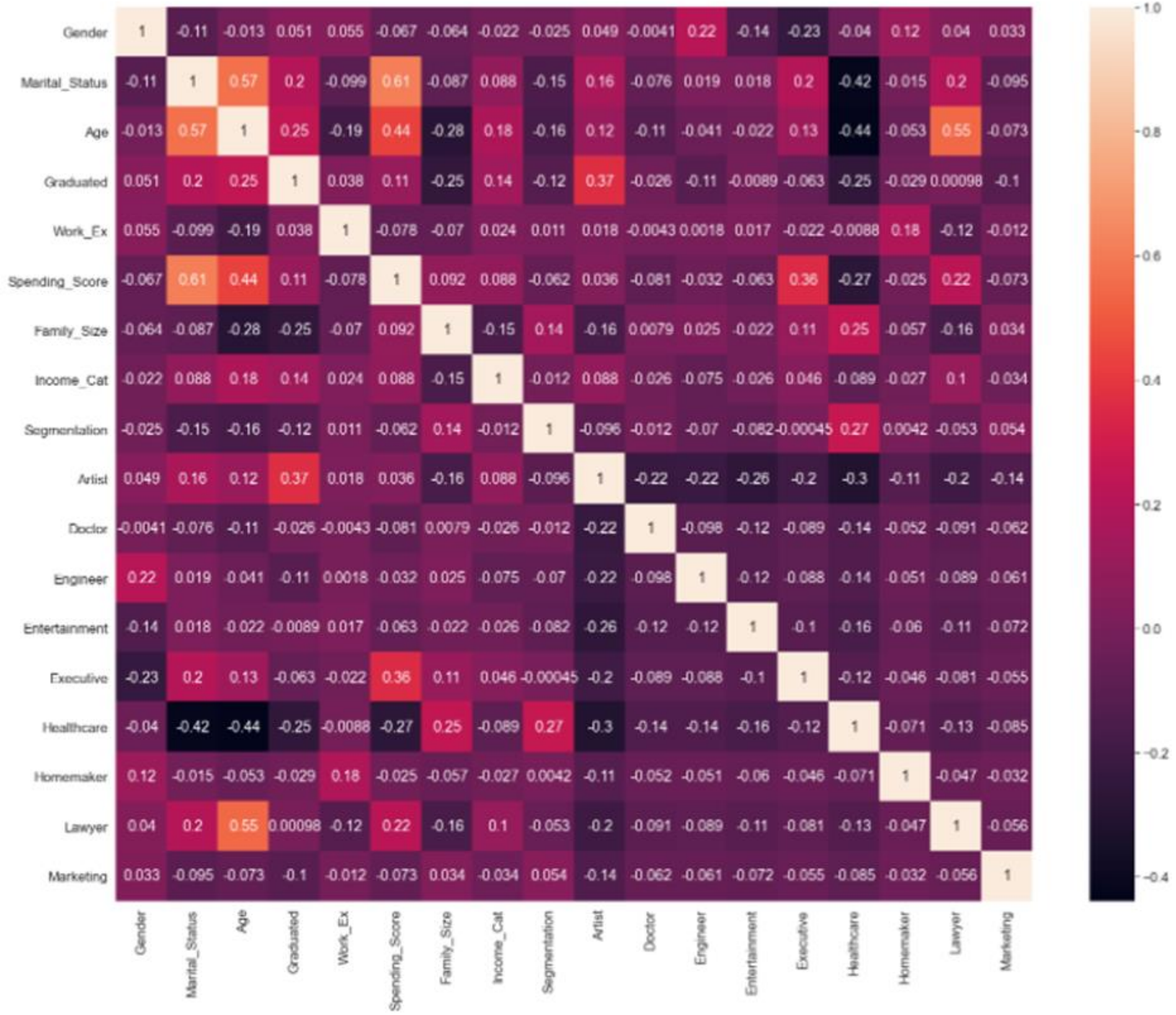
- Studied the features of the dataset like age, gender, etc.

- Kmeans
- DBSCAN
- KNN

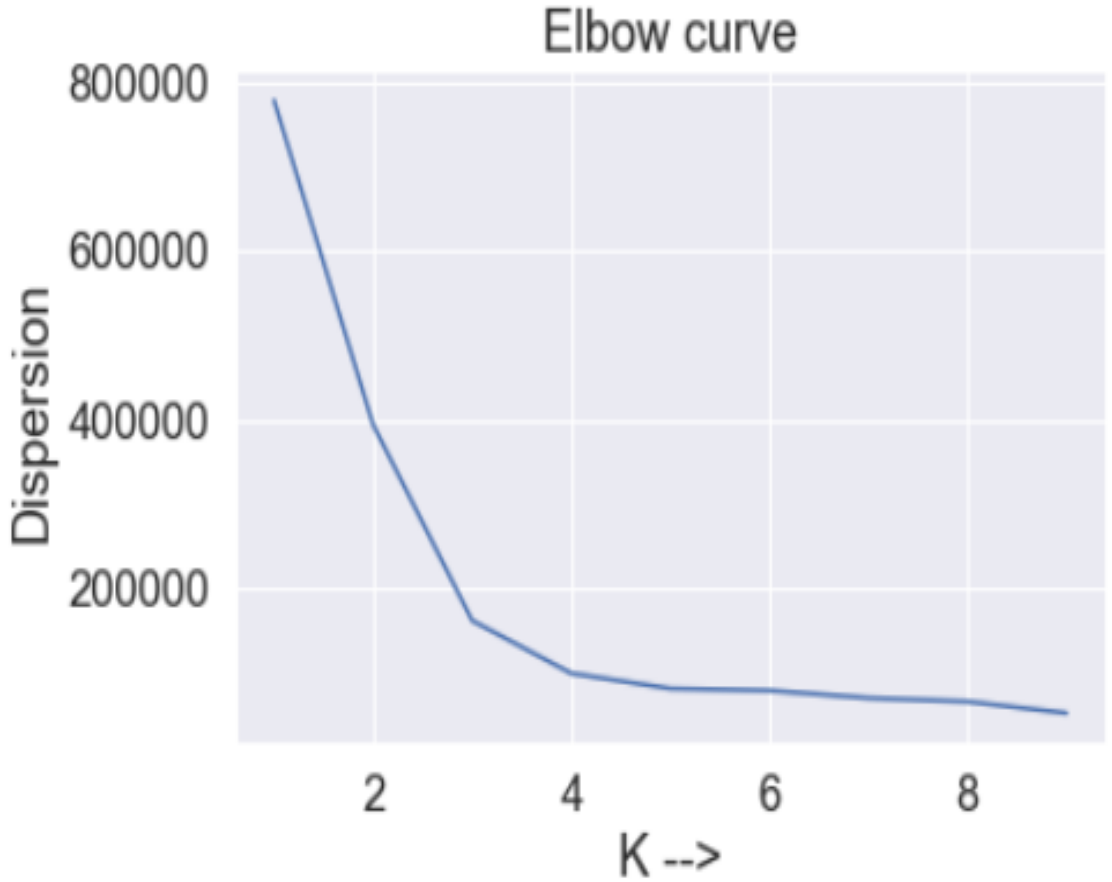
Data Exploration

Correlation Matrix

- The heatmap shows correlation of each variable present in the dataset
- Highest correlation to be seen amongst Age, Gender, work experience and income category
- Work Experience of the population is high for 0-2 years while lowest for 12 years



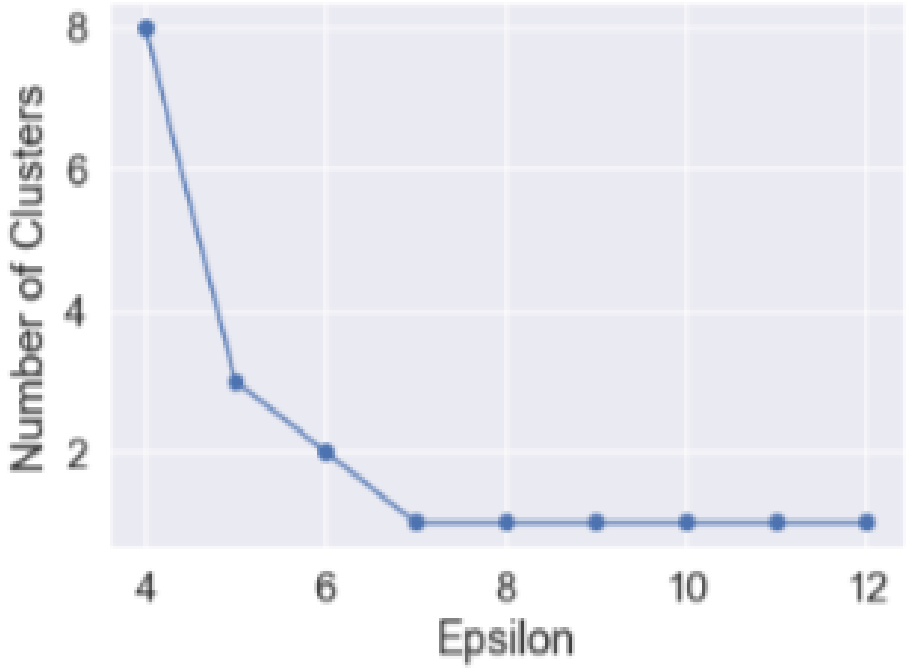
K-Means Algorithm

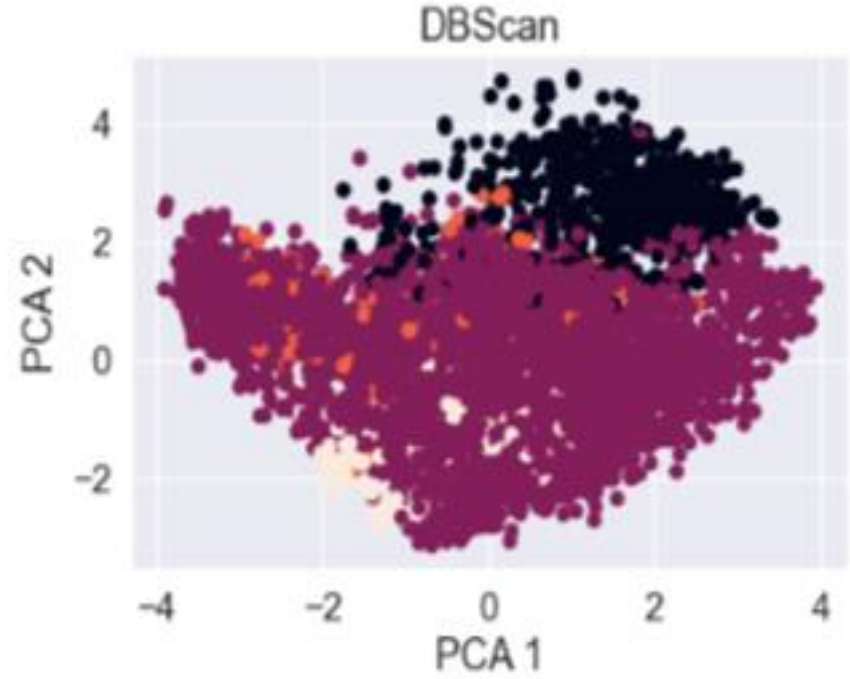


Since we are using random test and training splitting method the elbow curve and silhouette score for every rum varies

- Here, we have implemented the k-means unsupervised learning algorithm and it assigns every data point to the closest cluster, while maintaining the centroids as small as possible.
- Elbow Curve Plot **takes the costs** and the number of trials corresponding to different ‘k’.
- Silhouette Score Plot **takes the scores** and the number of trials and corresponding to different ‘k’ with the **accuracy of 44.8%**

DBSCAN – Unsupervised Algorithm





- DBSCAN is not useful with customer segmentation. The matching score between customer segmentation and the DBScan cluster is only 0.47%.

DBSCAN - Unsupervised Algorithm

Accuracy with eps = 4.35

```
In [313]: 1 adjusted_rand_score(data['Segmentation'], dbscan.labels_)
Out[313]: 0.004704905550590988
```

Accuracy with eps = 4

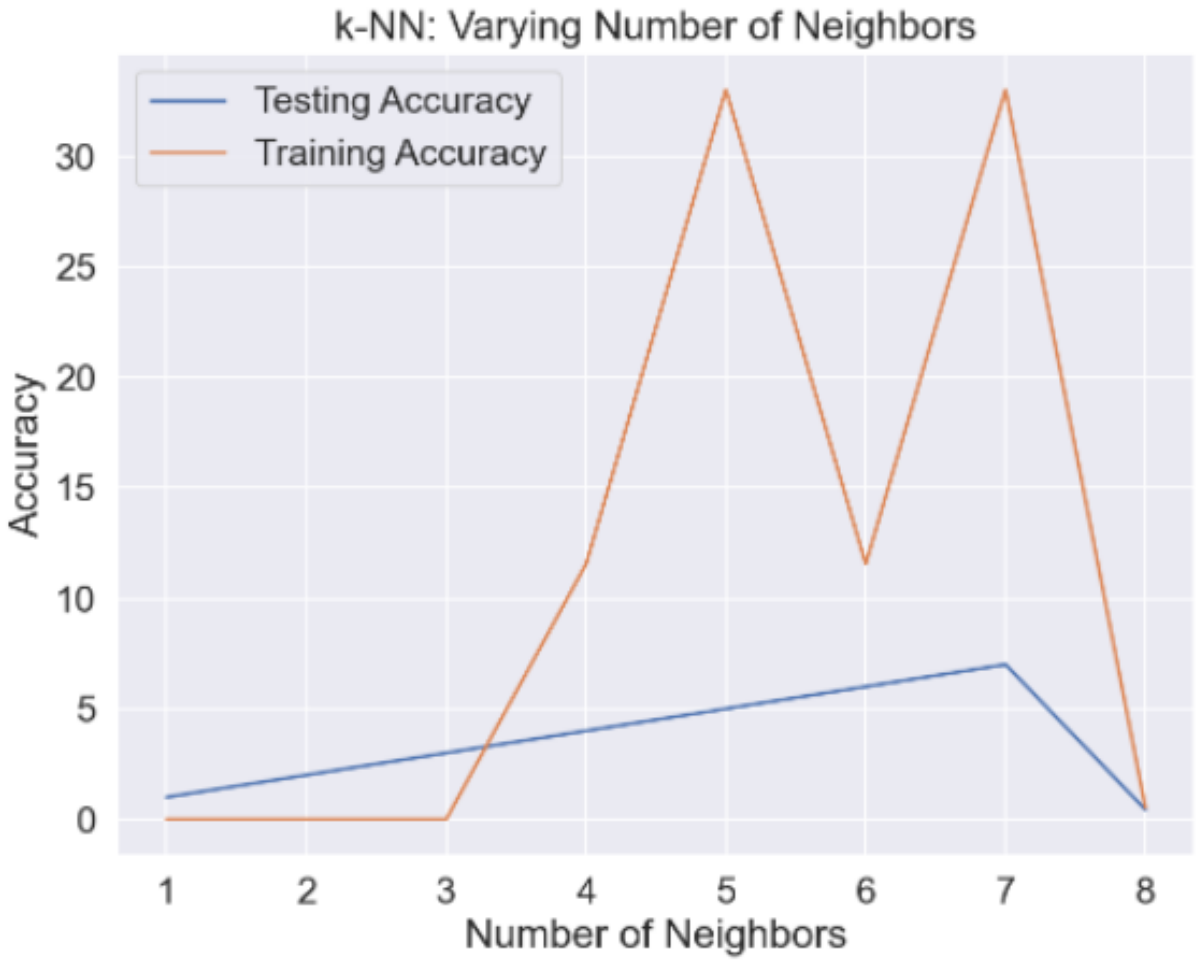
```
In [380]: 1 adjusted_rand_score(data['Segmentation'], dbscan.labels_)
Out[380]: -0.011269595374259244
```

Accuracy with eps = 5

```
In [375]: 1 adjusted_rand_score(data['Segmentation'], dbscan.labels_)
Out[375]: -0.00024565633673403174
```

- At eps = 4, model returns 8 clusters with the accuracy of -0.011%.
- At eps = 5, model returns only 3 clusters with the accuracy of -0.000245%.
- So, a higher or lower eps value would give us a very poor results.

KNN - Supervised Algorithm

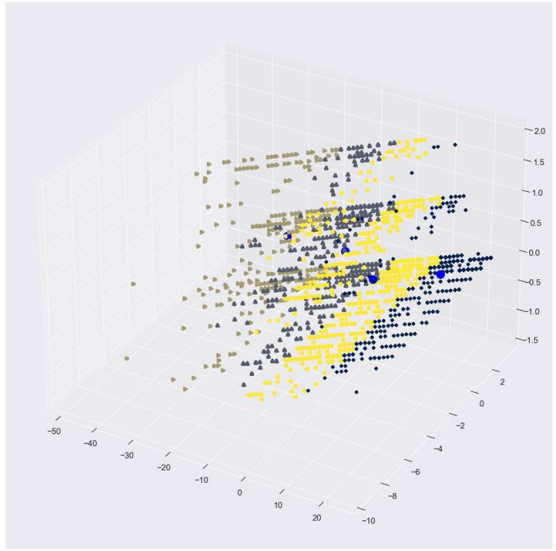


- We have implemented KNN supervised algorithm to predict the future subscription tiers based in their age, graduated, work experience, and spending score and put them in a correct segment with the closest proximity.

- From the graph, we can conclude that testing data is underfitting with the **accuracy of 48.36**

- KNN's main disadvantage of becoming significantly slower as the volume of data increases makes it an impractical choice in environments where predictions need to be made rapidly.

Insights

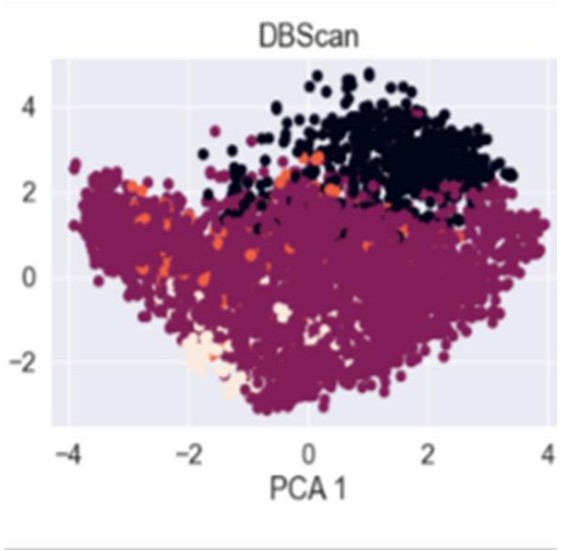


K- Means

Accuracy – 44.8%

Unsupervised Algorithm

Cluster based

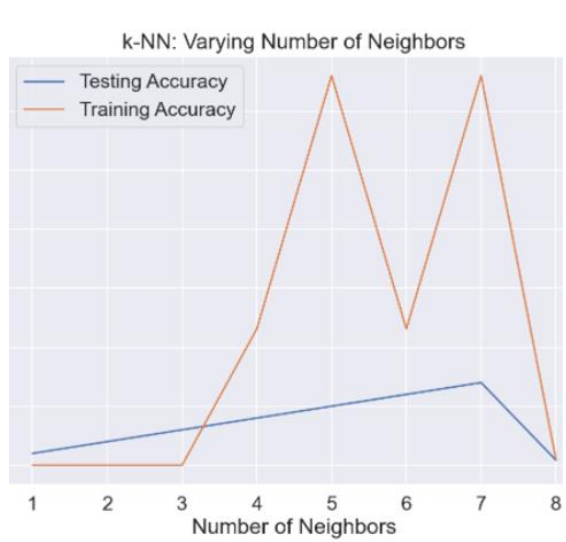


DBSCAN

Accuracy – 4% to 5%

Unsupervised Algorithm

Radial Based Clustering



KNN

Accuracy – 48.36%

Supervised Algorithm

Neighbours Based

Thank You!