Name: Anooshka Bajaj

1 a.

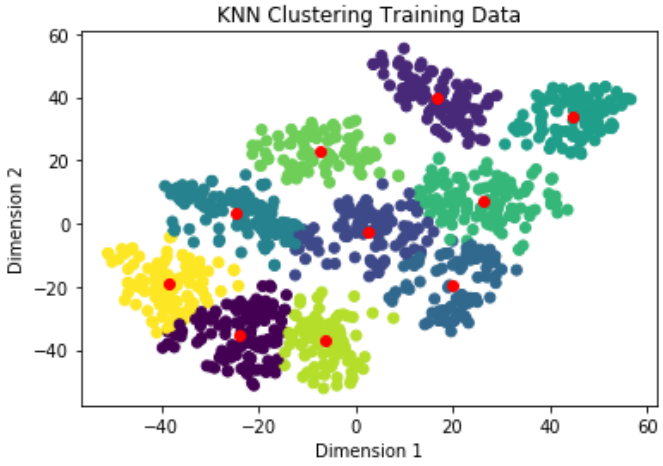


Figure 1 K-means (K=10) clustering on the mnist tsne training data

Inferences:

1. The K-Means algorithm has performed reasonably well because all the 10 clusters are distinctly visible.
2. The boundaries of clusters obtained in K-means algorithm are circular.

b. The purity score after training examples are assigned to the clusters is 0.694

c.

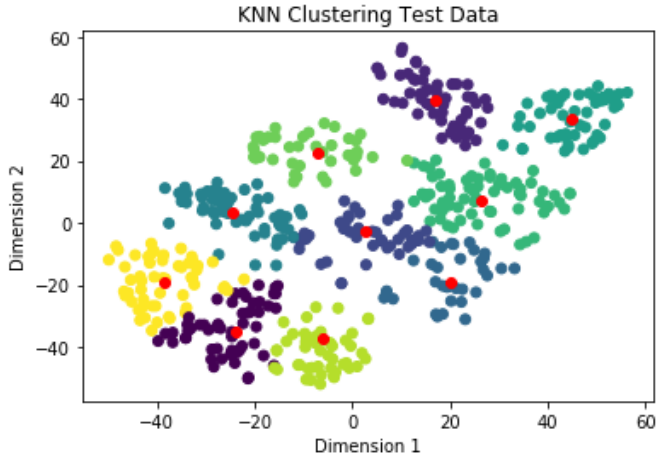


Figure 2 K-means (K=10) clustering on the mnist tsne test data

Inferences:

1. There is not much difference in the distribution of the test and train data except for the number of data points.

d. The purity score after test examples are assigned to the clusters is 0.684

Inferences:

1. The train purity score is higher than the test purity score. This is because the model was formed with respect to the training data and the test data was assigned those clusters.
2. This clustering approach doesn't form clusters of arbitrary shape and the value of number of clusters k has to be provided as an input to the algorithm.

2 a.

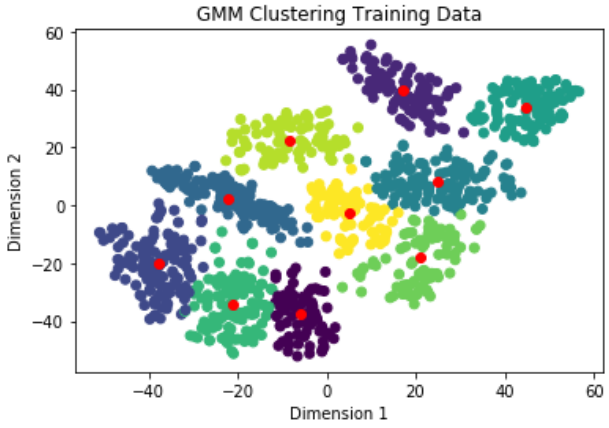


Figure 3 GMM clustering on the mnist tsne training data

Inferences:

1. The GMM algorithm has performed reasonably well.
2. The boundaries of clusters obtained in GMM algorithm are elliptical.
3. There is not much observable difference between clusters formed using K-means in 1.a and 2.a.

b. The purity score after training examples are assigned to the clusters is 0.708

c.

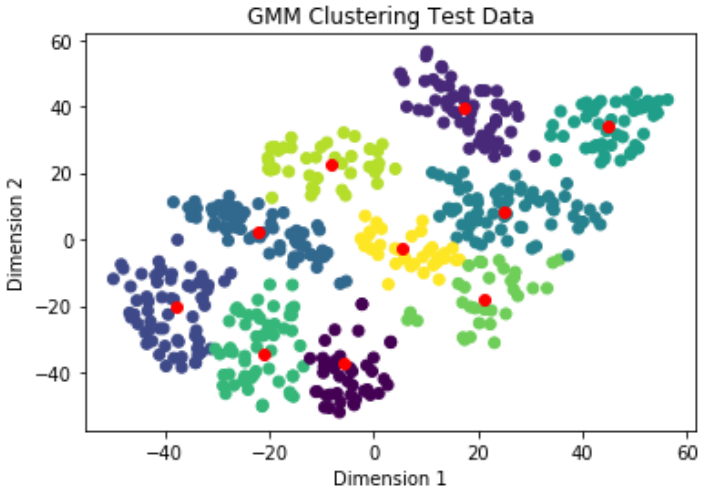


Figure 4 GMM clustering on the mnist tsne test data

Inferences:

1. There is not much difference in the distribution of the test and train data except for the number of data points.

d. The purity score after test examples are assigned to the clusters is 0.7

Inferences:

1. The train purity score is higher than the test purity score. This is because the model was formed with respect to the training data and the test data was assigned those clusters.
2. It is very time consuming for a data set with higher dimensions.

3 a.

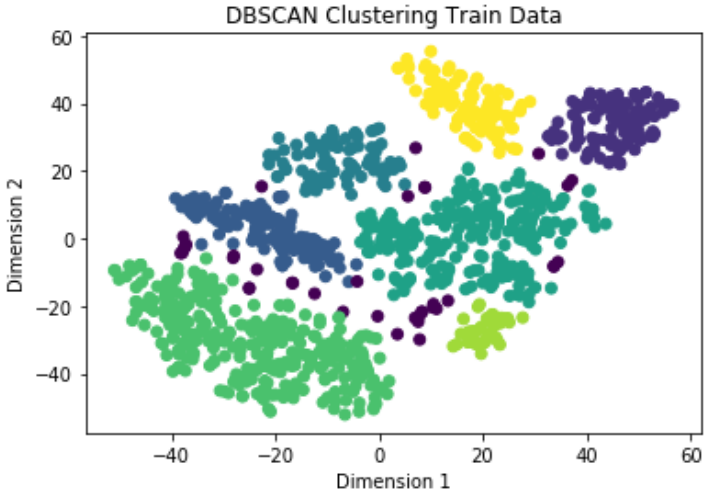


Figure 5 DBSCAN clustering on the mnist tsne training data

Inferences:

1. DBSCAN does not assume any shape or boundary hence it can discover arbitrarily shaped clusters. It can find clusters completely surrounded by other clusters also.
2. Observable difference between clusters formed using K-means in 1.a, GMM in 2.a and DBSCAN in 3.a is between the number of clusters formed. There are 8 clusters while using DBSCAN but while using K-Means or GMM we found 10 clusters. This is because DBSCAN finds the clusters solely based on the data distribution and we do not have to manually define the number of clusters.

b. The purity score after training examples are assigned to the clusters is 0.585

c.

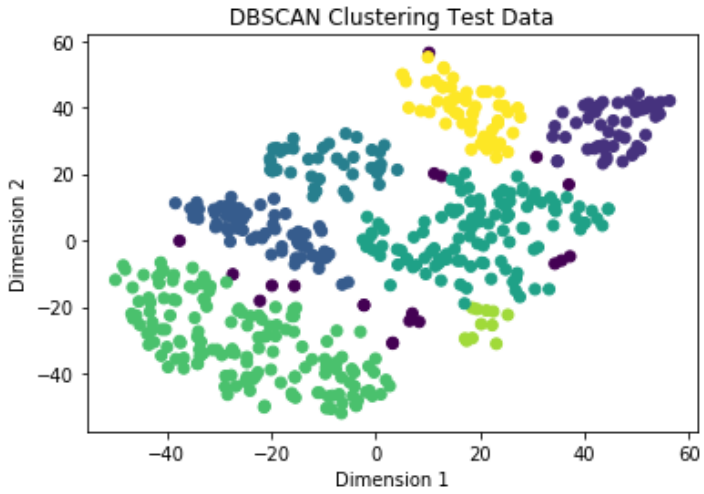


Figure 6 DBSCAN clustering on the mnist tsne test data

Inferences:

1. There is not much difference in the distribution of the test and train data except for less density in test data case.

d. The purity score after test examples are assigned to the clusters is 0.584

Inferences:

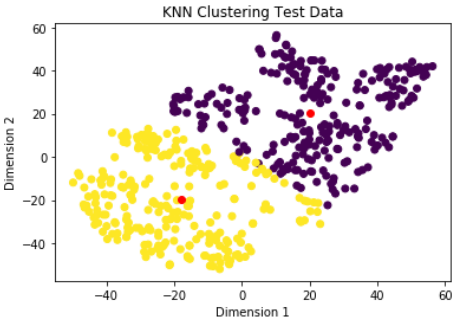
1. The train purity score is higher than the test purity score. This is because the model was formed with respect to the training data and the test data was assigned those clusters.
2. This clustering approach fails when the data is completely dense and there are no regions of low density which can act as cluster boundaries. The value of epsilon and minpoints have to be carefully chosen by us.

A.

1. K-means

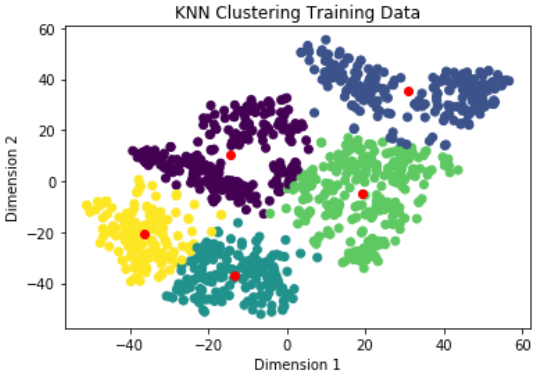
(i) k=2

Purity Score: 0.2

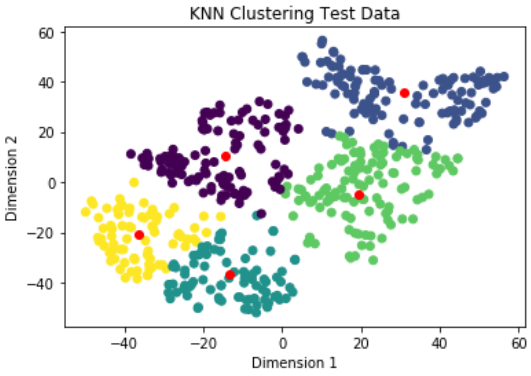


Purity Score: 0.2

(ii) k=5

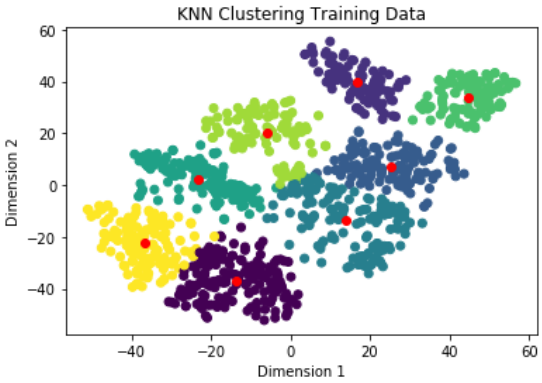


Purity Score: 0.392

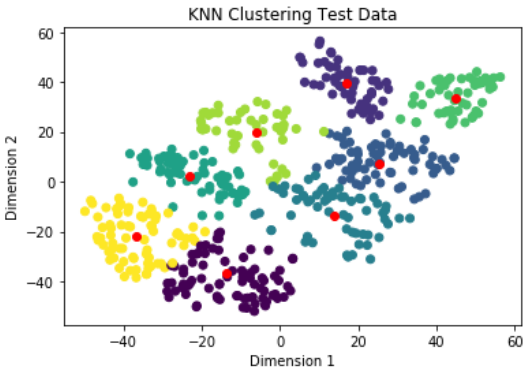


Purity Score: 0.402

(iii) k=8

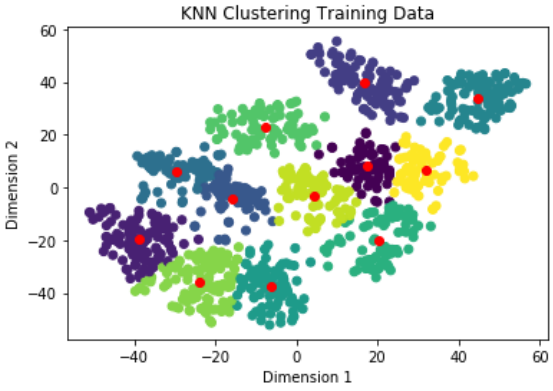


Purity Score: 0.63

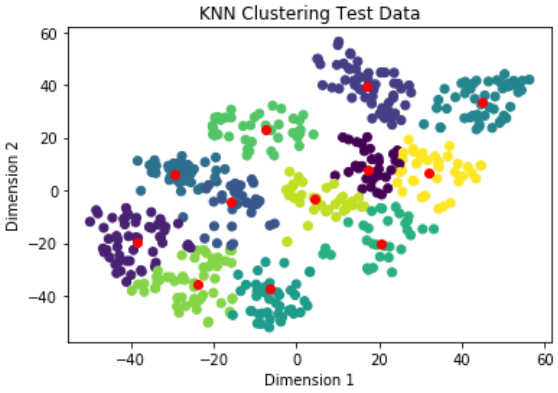


Purity Score: 0.624

(iv) k=12

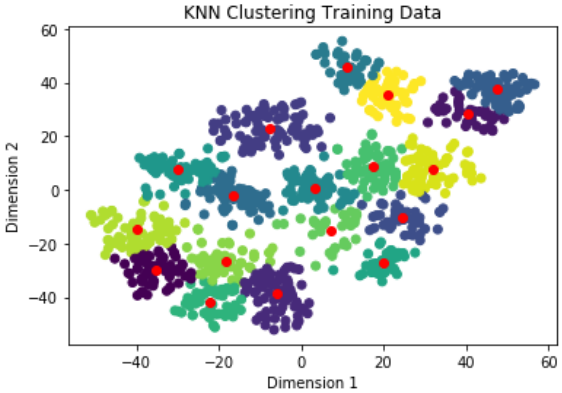


Purity Score: 0.626

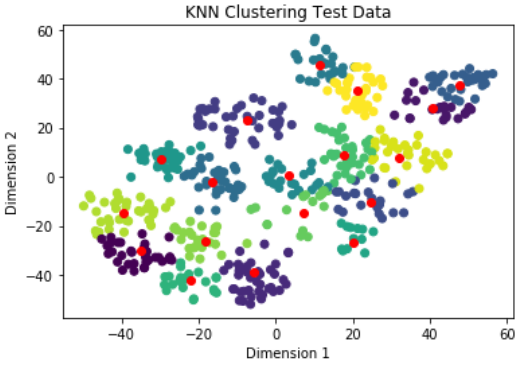


Purity Score: 0.624

(v) k=18

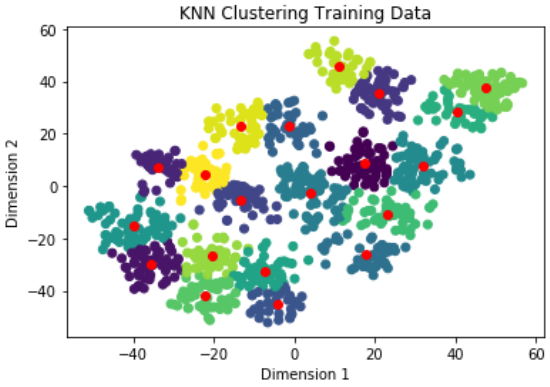


Purity Score: 0.484

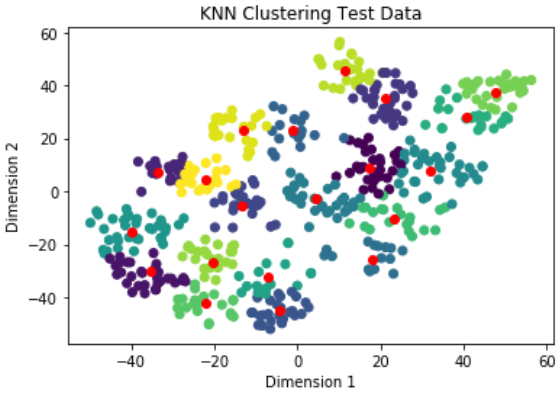


Purity Score: 0.468

(vi) k=20



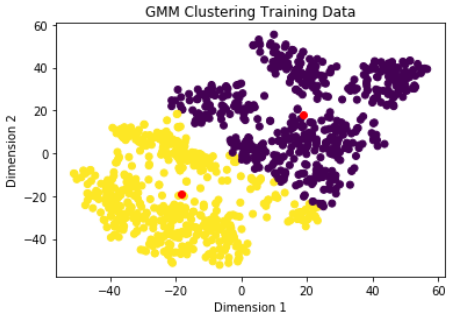
Purity Score: 0.434



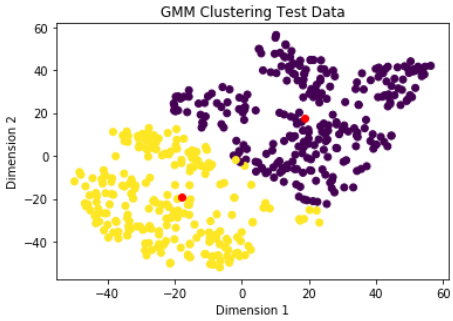
Purity Score: 0.426

2. GMM

(i) k=2

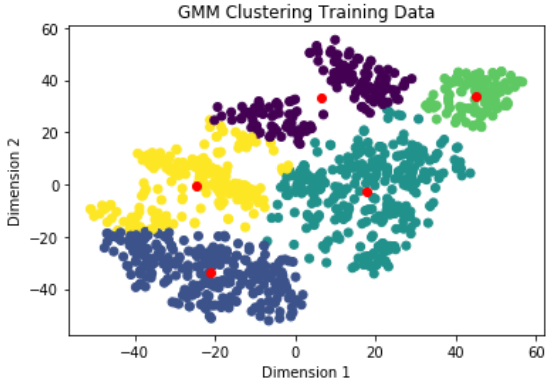


Purity Score: 0.2

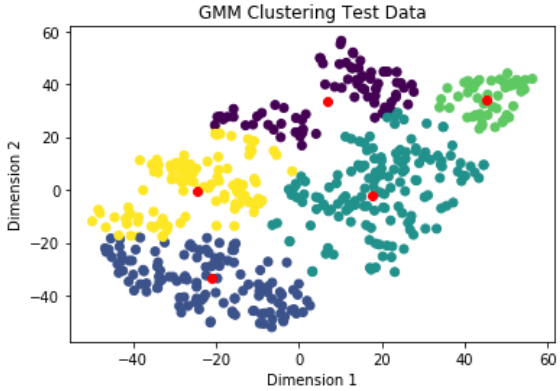


Purity Score: 0.2

(ii) k=5

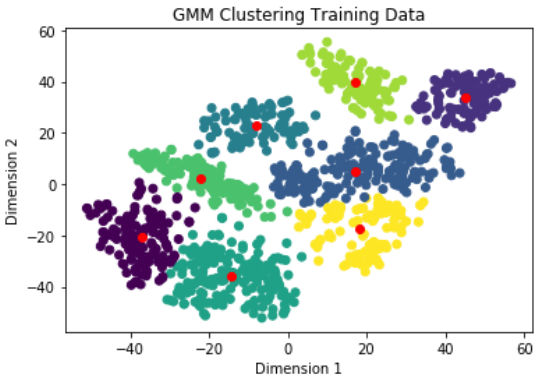


Purity Score: 0.47

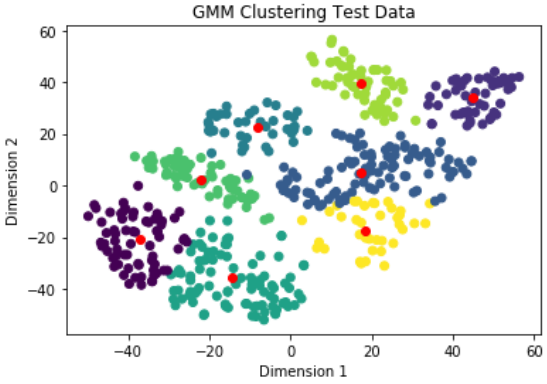


Purity Score: 0.464

(iii) k=8

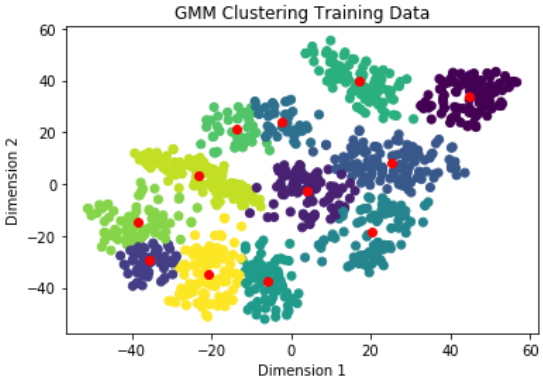


Purity Score: 0.629

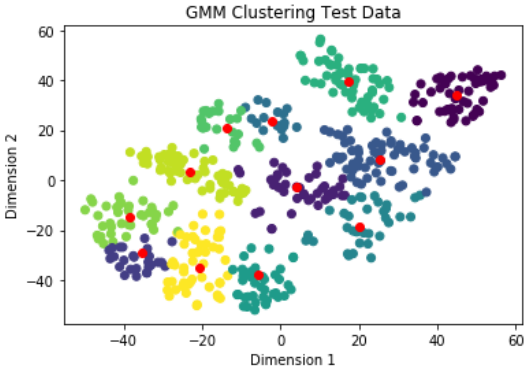


Purity Score: 0.628

(iv) k=12

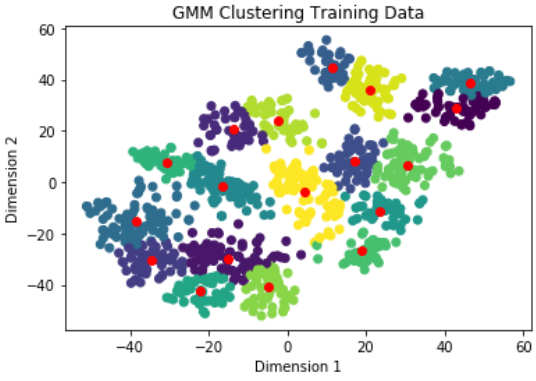


Purity Score: 0.656

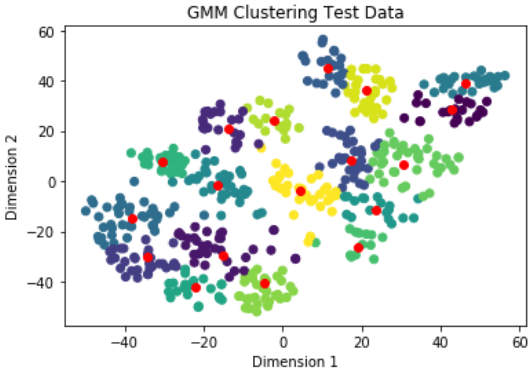


Purity Score: 0.662

(v) k=18

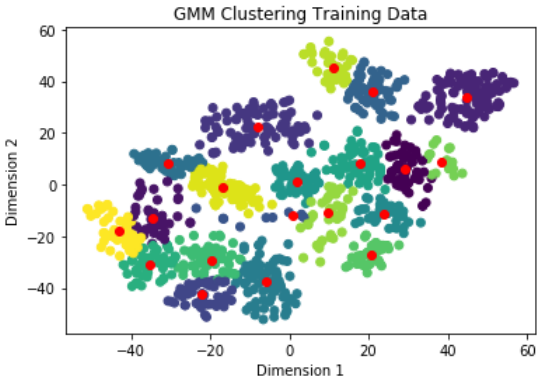


Purity Score: 0.448

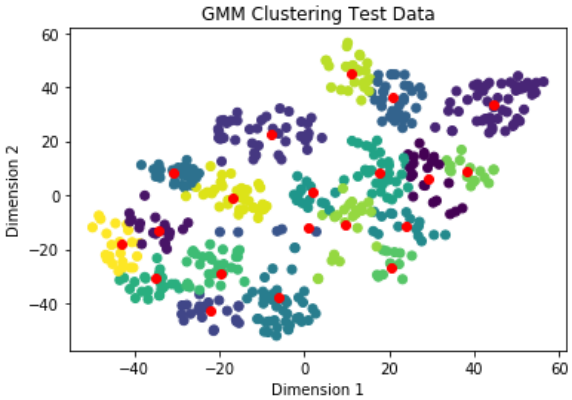


Purity Score: 0.446

(vi) k=20

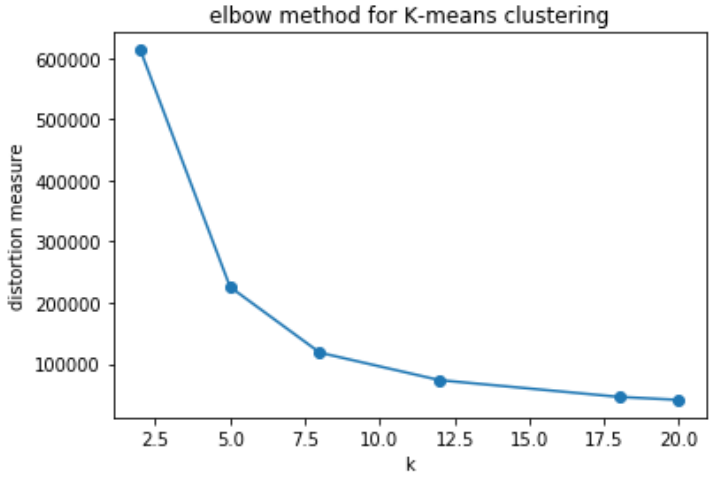


Purity Score: 0.52

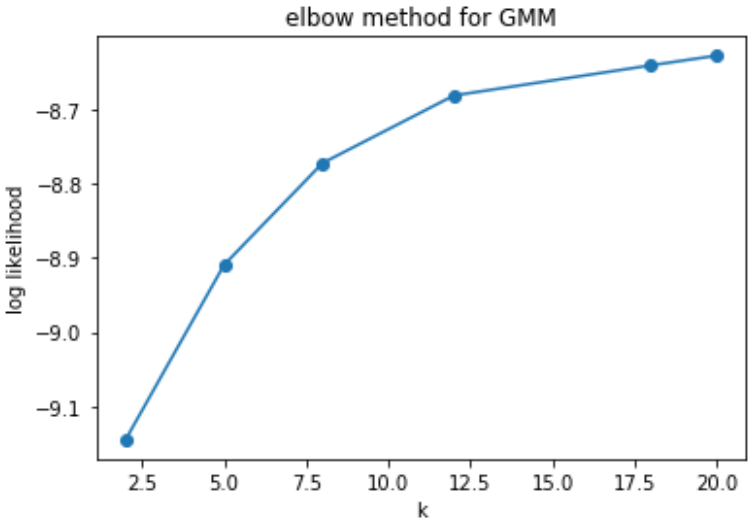


Purity Score: 0.494

Optimal number of clusters:



Optimum number of clusters using elbow method for K-means clustering = 8

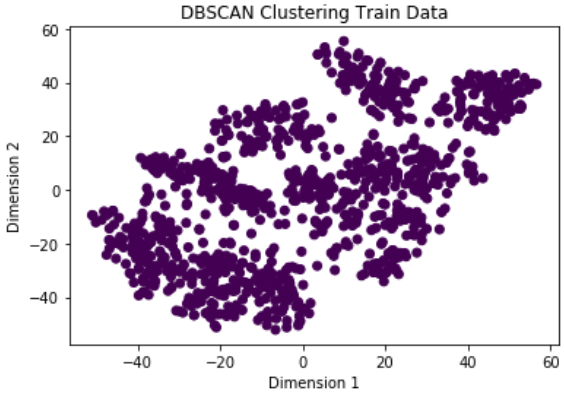


Optimum number of clusters using elbow method for GMM clustering = 8

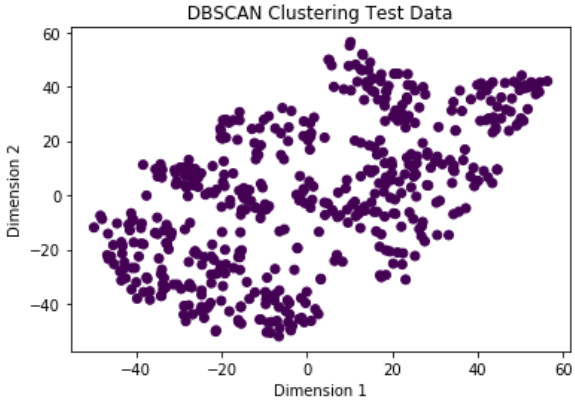
B.

(i) min_samples = 10

1. **Eps = 1**

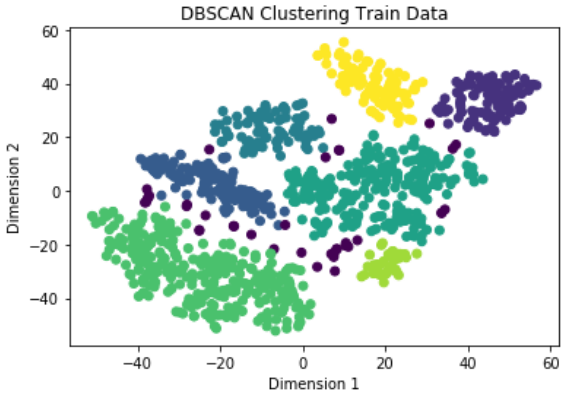


Purity Score: 0.1

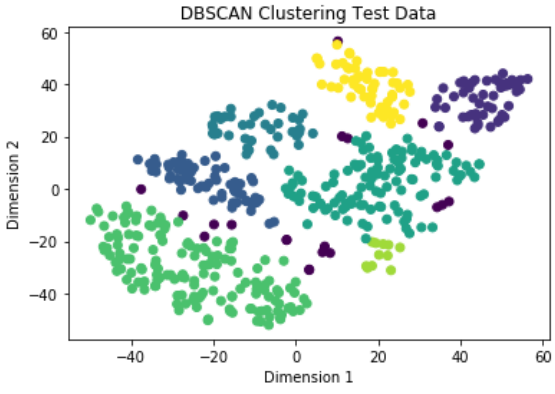


Purity Score: 0.1

2. **Eps = 5**

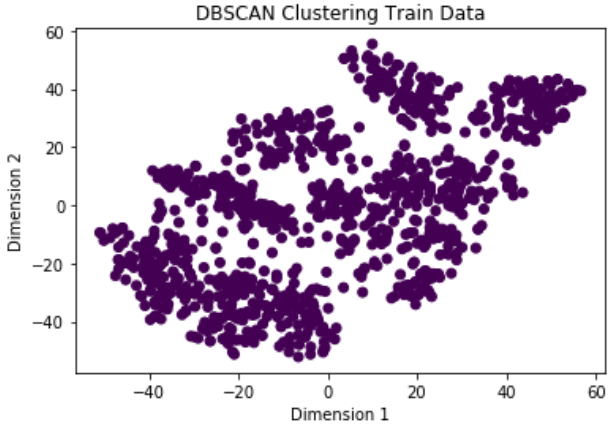


Purity Score: 0.585

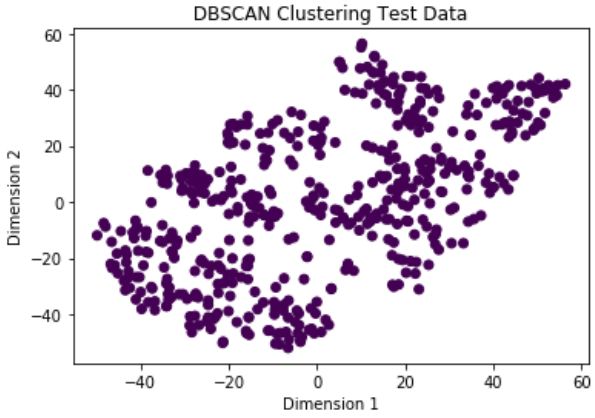


Purity Score: 0.584

3. **Eps = 10**



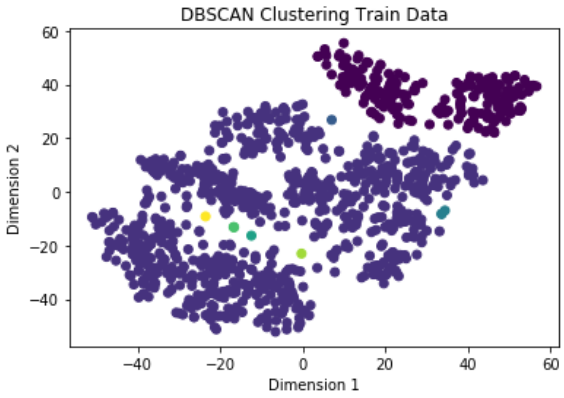
Purity Score: 0.1



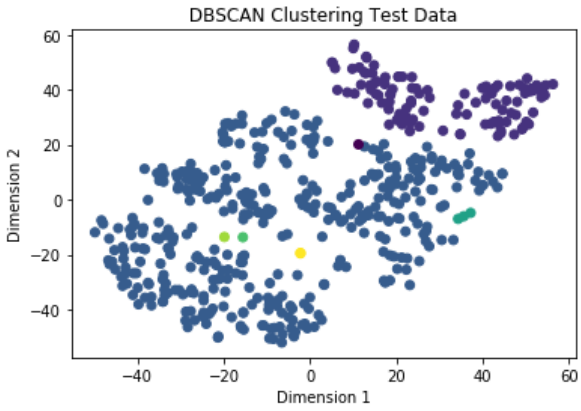
Purity Score: 0.1

(ii) eps = 5

1. **min_samples = 1**

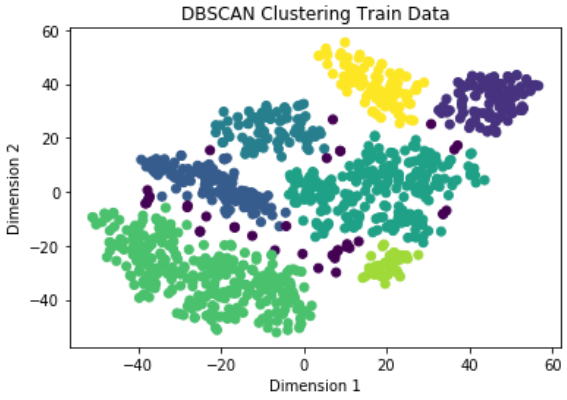


Purity Score: 0.208

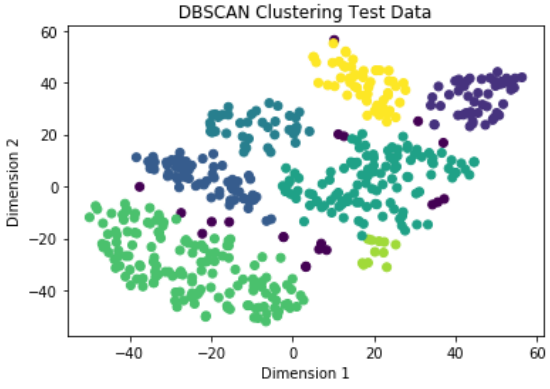


Purity Score: 0.212

2. **min_samples = 10**

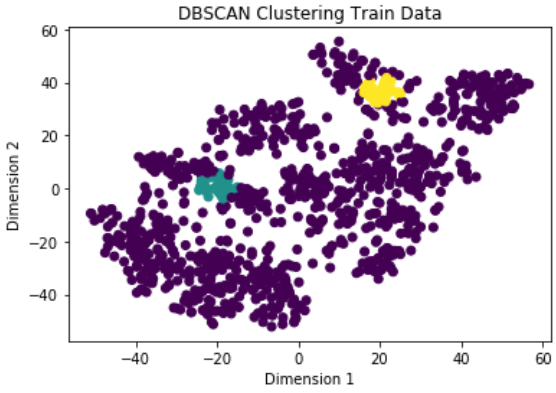


Purity Score: 0.585

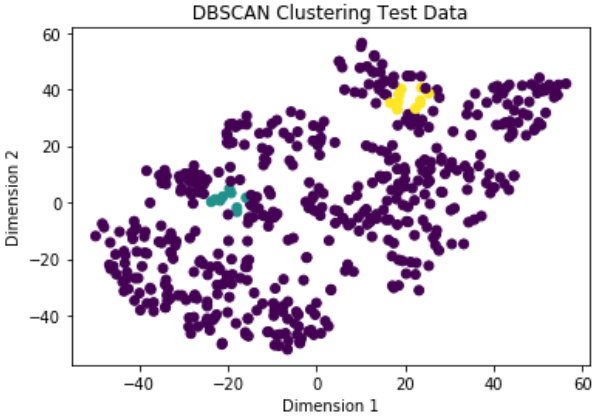


Purity Score: 0.584

3. **min_samples = 30**

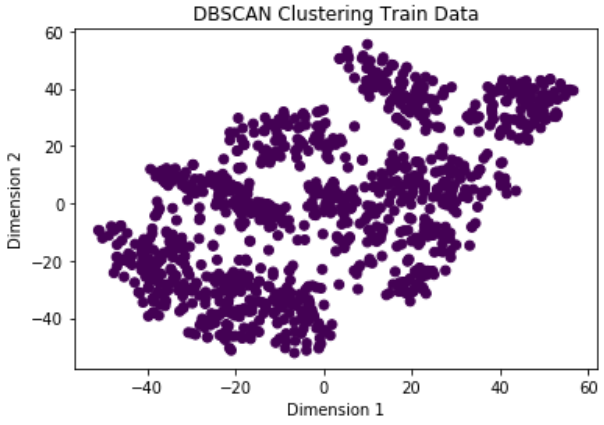


Purity Score: 0.158

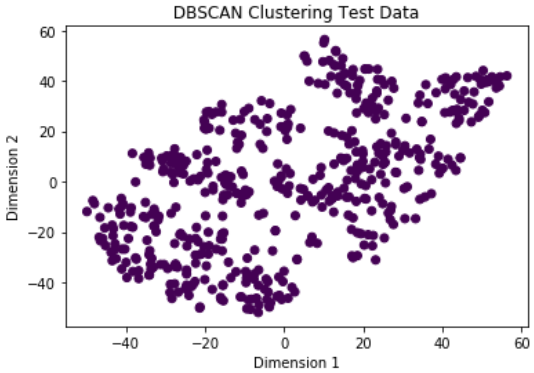


Purity Score: 0.14

4. **min_samples = 50**



Purity Score: 0.1



Purity Score: 0.1