CS 584- Data Mining and Applications
Kaustubh Prashant Karanjkar
Identifier: **Kaustubh**
#G01314326

HW3: Image Clustering

On Miner:

Rank: 62

Accuracy: 61%

Aim:

- Test/Implement k-means and k-means++ on a sample IRIS dataset
- Test/Implement k-means and k-means++ on a sample IRIS dataset
- Check accuracy using VScore and Silhouette coefficient

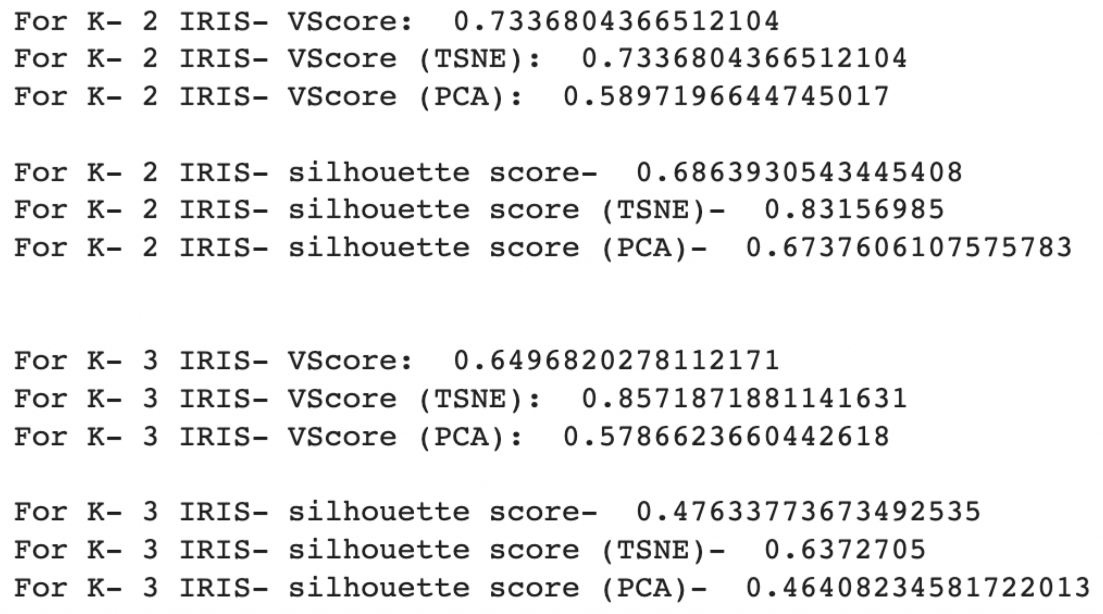
Approach:

- Initially, I implemented and tested my k means algorithm on the IRIS data set.
- To create a k means algorithm, the first thing which we must do is to find, the centroids. So, I have used two ways to find centroid.
    1. Randomly generating initial centroids.
    2. By using the k-means++ method, which basically is works by taking a single centroid and then generating the other centroids based on the probability distribution.
- Iris dataset contains 4 features, I have considered all the features and tested the algorithm, moreover, I have tried a couple of dimensionality reduction methods named PCA and TSNE and reduced the features to 2 from 4.

  Randomly generated centroids:

- Below are the results that I got after applying these reduction techniques, I have tested it for the multiple centroids, where k is the number of clusters or centroids.

- After comparing those results, I could see that I got the highest score in VScore and Silhouette coefficient when I applied the TSNE reduction technique, the score was 85.71% and 63.72% respectively, for k=3.



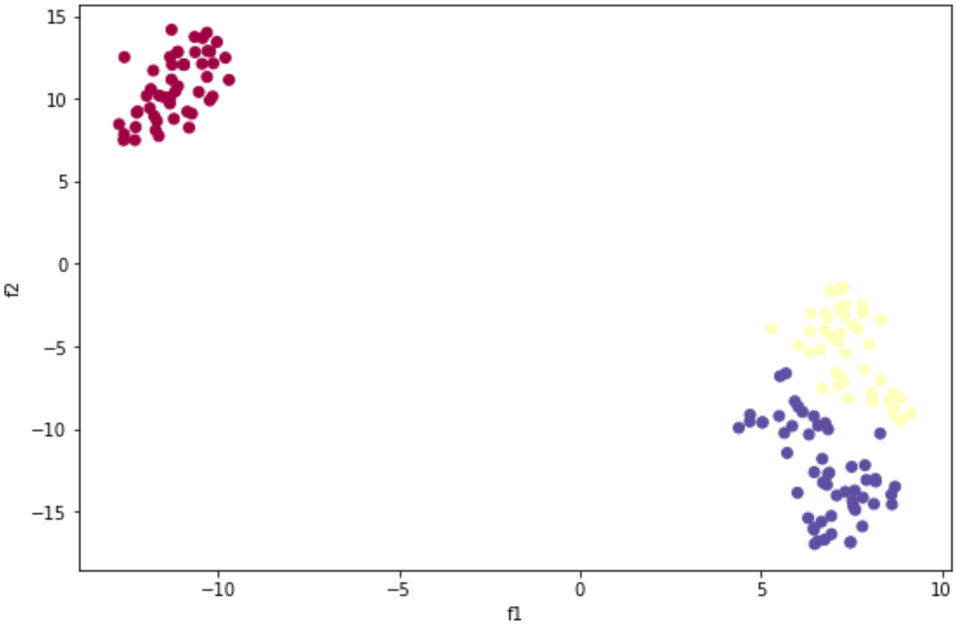
```
For K- 2 IRIS- VScore:  0.7336804366512104
For K- 2 IRIS- VScore (TSNE):  0.7336804366512104
For K- 2 IRIS- VScore (PCA):  0.5897196644745017

For K- 2 IRIS- silhouette score-  0.6863930543445408
For K- 2 IRIS- silhouette score (TSNE)-  0.83156985
For K- 2 IRIS- silhouette score (PCA)-  0.6737606107575783


For K- 3 IRIS- VScore:  0.6496820278112171
For K- 3 IRIS- VScore (TSNE):  0.8571871881141631
For K- 3 IRIS- VScore (PCA):  0.5786623660442618

For K- 3 IRIS- silhouette score-  0.47633773673492535
For K- 3 IRIS- silhouette score (TSNE)-  0.6372705
For K- 3 IRIS- silhouette score (PCA)-  0.46408234581722013
```

- Below is the cluster diagram, which I have plotted when I used TSNE reduction technique. Below are the 3 clusters.

- 

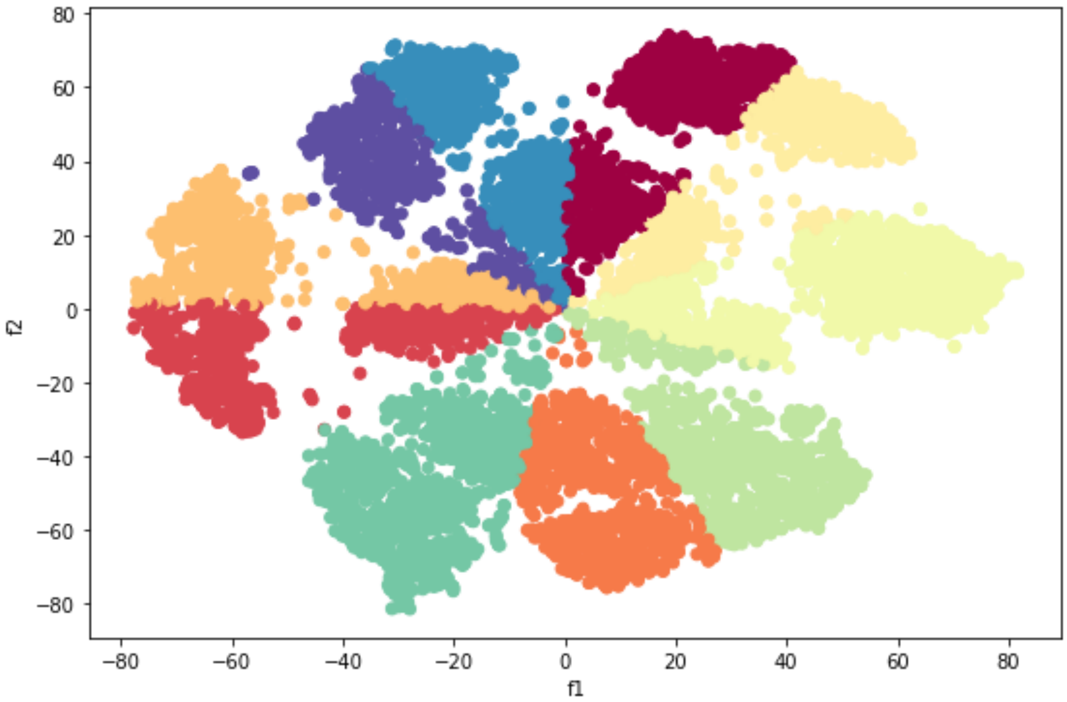
K-Means++ Implementation on IRIS dataset:

- As per my observation, applying TSNE reduction on the dataset gave more accuracy when tested using VScore and Silhouette coefficient, so applied this reduction technique on the kmeans++ algorithm and got a score of 89.9% on VScore and 48.65% on the Silhouette coefficient.

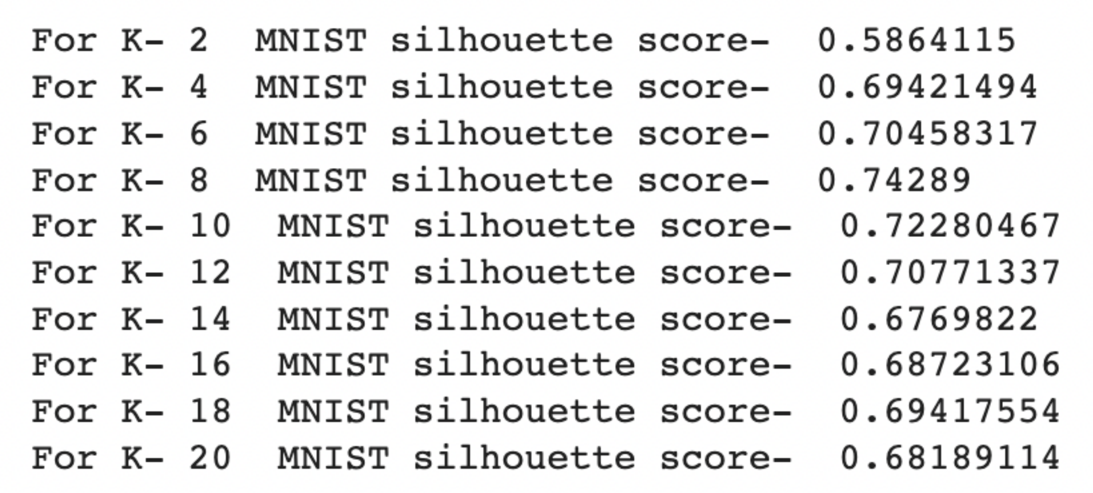
Implementing the K-Means++ algorithm on MNIST dataset:

- In this dataset, the images were scanned and scaled into 28x28 pixels. For every digit, each pixel can be represented as an integer in the range [0, 255] where 0 corresponds to the pixel being completely white, and 255 corresponds to the pixel being completely black. This gives us a 28x28 matrix of integers for each digit. We can then flatten each matrix into a 1x784 vector. Which means there are 784 features.

- It is not feasible to apply the k-means algorithm to 784 features, so we must apply a dimensionality reduction technique to reduce the features.
- As per my observation, the TSNE reduction technique performed better as compared to PCA. With PCA dimensionality reduction, I got a silhouette score of 68.51% on the other hand I got 73.24% with the TSNE reduction technique.
- Additionally, I have used normalization on the dataset to rescale the real-valued numeric attributes.
- After doing the pre-processing and then using k means++ on the pre-processed dataset, I got a silhouette score of 73.24% when cluster and centroid are 10. I have used the parameter "*metric = 'cosine'* "
- I got a score of 61% on the Miner.
- Below are the clusters which I have plotted of this result.



- Then I tried the k means algorithm for different numbers of centroids and clusters, from 2 to 20, and below are the results of the Silhouette score I got:



```
For K- 2  MNIST silhouette score-  0.5864115
For K- 4  MNIST silhouette score-  0.69421494
For K- 6  MNIST silhouette score-  0.70458317
For K- 8  MNIST silhouette score-  0.74289
For K- 10  MNIST silhouette score-  0.72280467
For K- 12  MNIST silhouette score-  0.70771337
For K- 14  MNIST silhouette score-  0.6769822
For K- 16  MNIST silhouette score-  0.68723106
For K- 18  MNIST silhouette score-  0.69417554
For K- 20  MNIST silhouette score-  0.68189114
```