Pattern Recognition and Machine Learning

Lab - 6 Assignment Report

Aryan Himmatlal Prajapati (B21EE012)

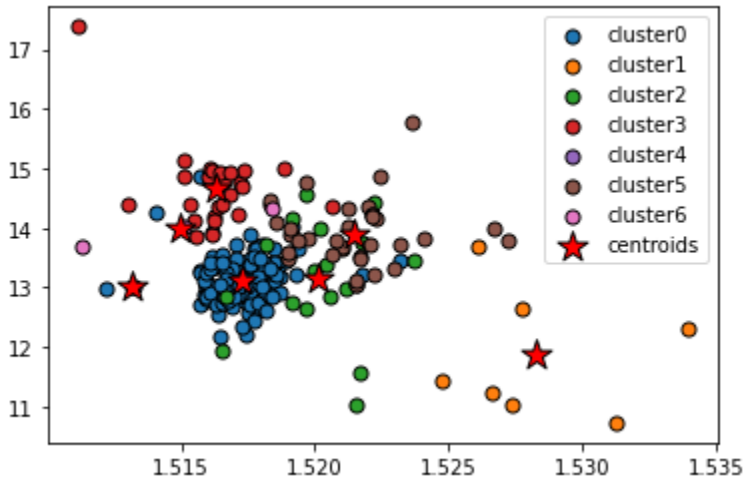
Question 1.

Pre-Processed Glass Classification Dataset

Part 1

Builded a k-means clustering algorithm and implemented it using the value of k = 7. Visualized this part by showing the clusters along with the centroids.

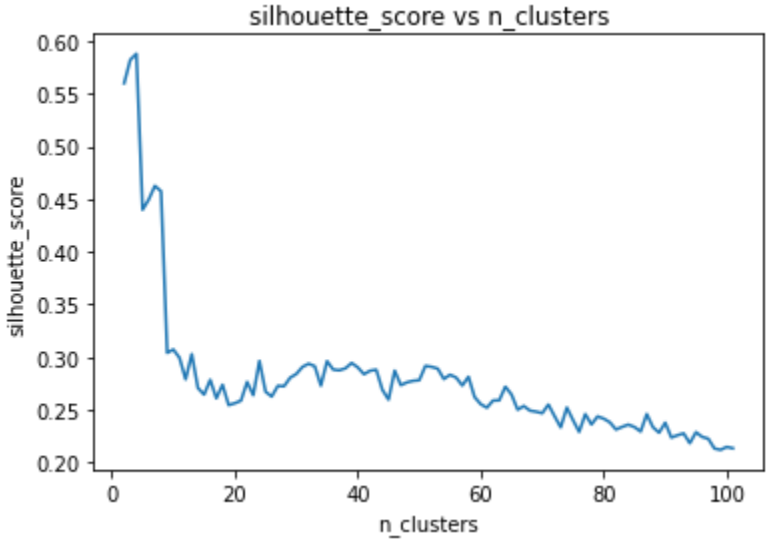
Visualization of Clusters with Centroids:



Part 2

Used different values of k and found the Silhouette score. Plotted Silhouette score vs K.

Plot of Silhouette score vs K(n_clusters):



`silhouette_score is maximum, when n_clusters = 4.0`

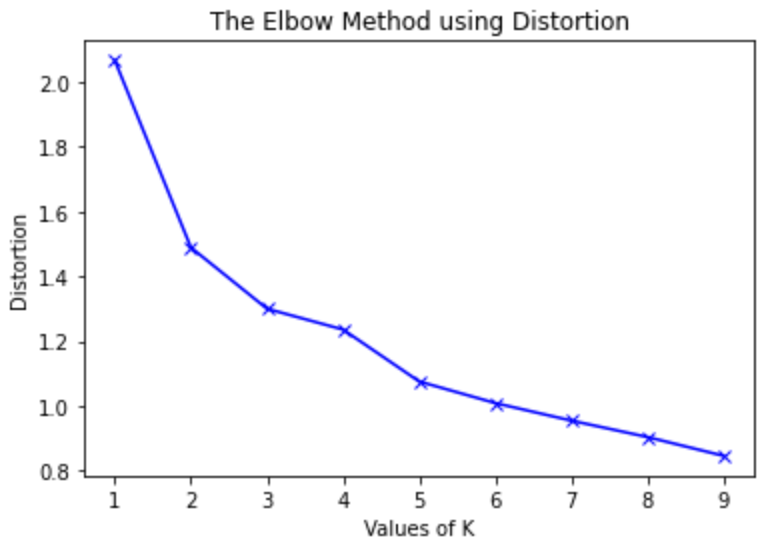
Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.
1: Means clusters are well apart from each other and clearly distinguished.
0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.
-1: Means clusters are assigned in the wrong way.

Source:
https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

From silhouette score for different values of k, we can conclude that optimal value for k is 4 as silhouette score is maximum for it.

Part 3

Used Elbow Method(using distortion) to Find Optimal K value.



To determine the optimal number of clusters, we have to select the value of k at the “elbow” i.e. the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is 4.

Part 4

Applied bagging with the KNN classifier as the base model. Showed results with different values of K(=1,2,3).

Results -

For K = 1:

KNN Accuracy Score: `0.7692307692307693`

Bagging Accuracy Score: `0.7384615384615385`

For K = 2:

KNN Accuracy Score: `0.6615384615384615`

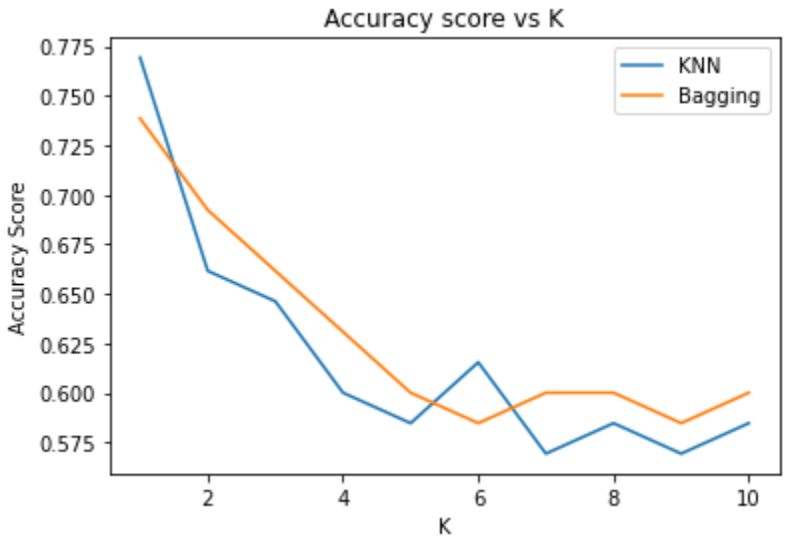
Bagging Accuracy Score: `0.6923076923076923`

For K = 3:

KNN Accuracy Score: `0.6461538461538462`

Bagging Accuracy Score: `0.6615384615384615`

Plot of Accuracy Score vs K:



In General, The reason for the improvement in accuracy with bagging is due to the reduction of variance. By creating multiple samples of the dataset and training separate models on each sample, we are effectively reducing the variance of the model. This reduction in variance leads to a decrease in overfitting and an improvement in the model's generalization performance.

In terms of bias, bagging does not typically have a significant effect. Since the base model is a KNN classifier, which has low bias, we do not expect bagging to have a significant impact on bias.

In summary, we have applied bagging with the KNN classifier as the base model and observed an improvement in accuracy. The improvement in accuracy is due to the reduction of variance achieved by bagging.

Question 2.

Part A & B

Implemented K-Means Clustering Algorithm from Scratch.

Which Has Following Functionalities:

i) Class which will be able to store the cluster centers.
ii) Take a value of k from users to give k clusters.
iii) Able to take initial cluster center points from the user as its initialization.
iv) Stop iterating when it converges (cluster centers are not changing anymore) or,a maximum iteration (given as max_iter by user) is reached.

Part C

Trained the k-means model on Olivetti data with k = 40. Reported the number of points in each cluster.

```
Number of point in cluster 0 is: 4
Number of point in cluster 1 is: 11
Number of point in cluster 2 is: 5
Number of point in cluster 3 is: 4
Number of point in cluster 4 is: 5
Number of point in cluster 5 is: 1
Number of point in cluster 6 is: 4
Number of point in cluster 7 is: 16
Number of point in cluster 8 is: 1
Number of point in cluster 9 is: 10
Number of point in cluster 10 is: 14
Number of point in cluster 11 is: 12
Number of point in cluster 12 is: 25
Number of point in cluster 13 is: 7
Number of point in cluster 14 is: 12
Number of point in cluster 15 is: 15
Number of point in cluster 16 is: 8
Number of point in cluster 17 is: 19
Number of point in cluster 18 is: 7
Number of point in cluster 19 is: 4
Number of point in cluster 20 is: 11
Number of point in cluster 21 is: 10
Number of point in cluster 22 is: 6
Number of point in cluster 23 is: 17
```

```
Number of point in cluster 24 is: 10
Number of point in cluster 25 is: 27
Number of point in cluster 26 is: 4
Number of point in cluster 27 is: 10
Number of point in cluster 28 is: 3
Number of point in cluster 29 is: 30
Number of point in cluster 30 is: 9
Number of point in cluster 31 is: 7
Number of point in cluster 32 is: 3
Number of point in cluster 33 is: 13
Number of point in cluster 34 is: 3
Number of point in cluster 35 is: 16
Number of point in cluster 36 is: 5
Number of point in cluster 37 is: 14
Number of point in cluster 38 is: 6
Number of point in cluster 39 is: 12
```

Part D

Visualized the cluster centers of each cluster as 2-d images of all clusters.

Part E

Visualized 10 images corresponding to each cluster**.**
Note: Every Cluster Doesn't necessarily have 10 images.
Please Check .ipynb Q.2 Part E for Image Visualization.

Part F

Train another k-means model. reported the number of points in each cluster and visualized the cluster centers.

Number of point in cluster 0 is: 6
Number of point in cluster 1 is: 4
Number of point in cluster 2 is: 5
Number of point in cluster 3 is: 14
Number of point in cluster 4 is: 17
Number of point in cluster 5 is: 2
Number of point in cluster 6 is: 10
Number of point in cluster 7 is: 4
Number of point in cluster 8 is: 3
Number of point in cluster 9 is: 11
Number of point in cluster 10 is: 3
Number of point in cluster 11 is: 10
Number of point in cluster 12 is: 19
Number of point in cluster 13 is: 10
Number of point in cluster 14 is: 9
Number of point in cluster 15 is: 4
Number of point in cluster 16 is: 5
Number of point in cluster 17 is: 16
Number of point in cluster 18 is: 7
Number of point in cluster 19 is: 10
Number of point in cluster 20 is: 4
Number of point in cluster 21 is: 2
Number of point in cluster 22 is: 5
Number of point in cluster 23 is: 18
Number of point in cluster 24 is: 15
Number of point in cluster 25 is: 16
Number of point in cluster 26 is: 21
Number of point in cluster 27 is: 3
Number of point in cluster 28 is: 5
Number of point in cluster 29 is: 4
Number of point in cluster 30 is: 22
Number of point in cluster 31 is: 21
Number of point in cluster 32 is: 10
Number of point in cluster 33 is: 4

Number of point in cluster 34 is: 9
Number of point in cluster 35 is: 21
Number of point in cluster 36 is: 22
Number of point in cluster 37 is: 8
Number of point in cluster 38 is: 10
Number of point in cluster 39 is: 11

Visualization of the cluster centers of each cluster as 2-d images of all clusters:

Part G

Visualized 10 images corresponding to each cluster**.**
Note: Every Cluster Doesn't necessarily have 10 images.
Please Check .ipynb Q.2 Part G for Image Visualization.

Question 3.

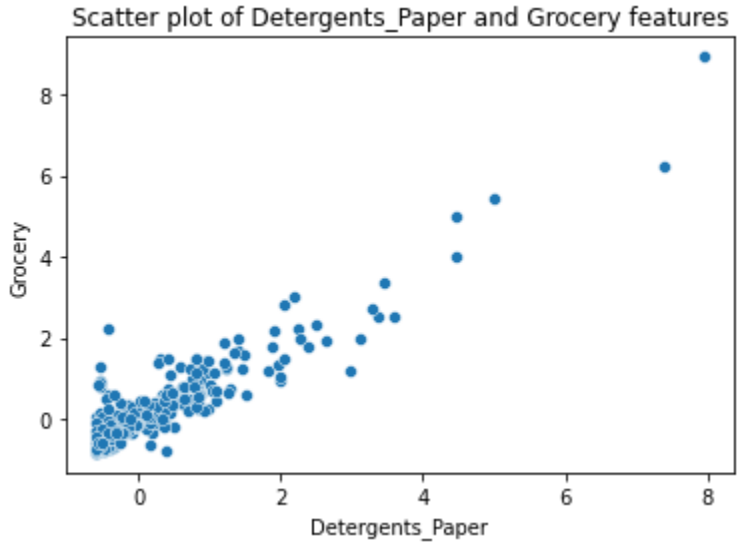
Part 1

Checked out the dataset & preprocessed the data so that the scale of each variable will be the same.

Part 2

Found out the covariance between the pair of features with which we can best visualize the outliers. Also, visualized the same set of features.

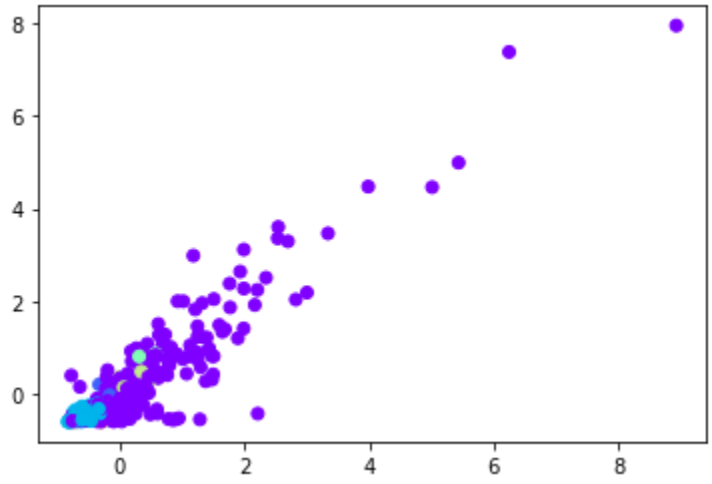
Visualization:



Part 3

Applied DBSCAN to cluster the data points and visualized the same.

Visualization:




```
Labels : [-1  0  1  2  3  4  5  6  7]
```

Part 4

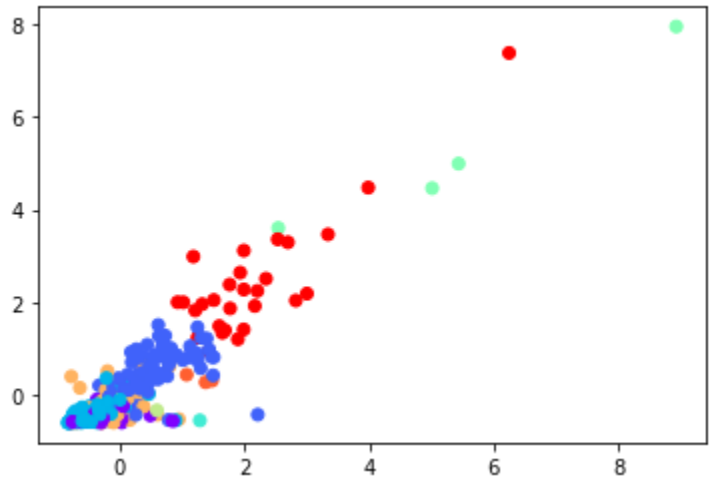
Applied K Means Clustering on the same dataset and visualized the same.

Comparing the visualizations of K-means clustering and DBSCAN, we can observe the following:

- K-means clustering assumes that the clusters have a spherical shape and a similar size, while DBSCAN can handle clusters of different shapes and sizes.
- K-means clustering assigns each data point to exactly one cluster, while DBSCAN can assign points to multiple clusters if they are located near the border of two or more clusters.
- K-means clustering can be sensitive to the initial placement of the centroids, while DBSCAN is more robust to initialization.
- K-means clustering requires specifying the number of clusters beforehand, while DBSCAN doesn't.

In conclusion, the choice of clustering algorithm depends on the characteristics of the dataset and the specific goals of the analysis. K-means clustering may be a good choice for datasets with distinct, spherical clusters, while DBSCAN may be more suitable for datasets with complex, non-linear structures.

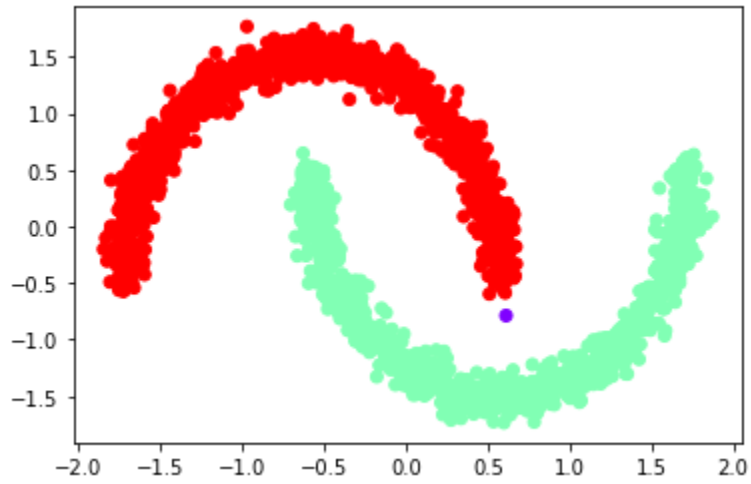
Visualization:



Part 5

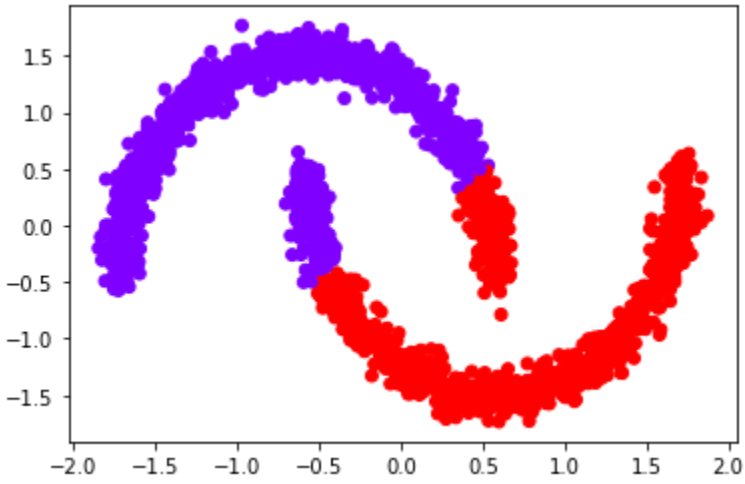
Used the make_moons function of sklearn to create a dataset of 2000 points. Applied DBSCAN and K Means Clustering to cluster them and finally compared the plots.

DBSCAN:




```
Labels : [-1  0  1]
```

K Means Clustering:



On this dataset, DBSCAN is performing better than K Means Clustering because it is able to discriminate between two classes better and also DBSCAN is able to detect outliers in Dataset.

THANK YOU