DAMA61 - 3rd assignment

Exercise 1

1 - 4

The code was constructed as dictated by the exercise. There was a long runtime noticed and the attribute n_jobs=-1 as well as the preprocessor module of the sklearn was used to optimize the performance. After the code run, the following results of the scores were produced:

Decision Tree Score: 0.7932
Random Forest Score: 0.9329
AdaBoost Score: 0.7
Linear SVC Score: 0.8996
Logistic Regression Score: 0.9216
Stacking Classifier Score: 0.9316

5)

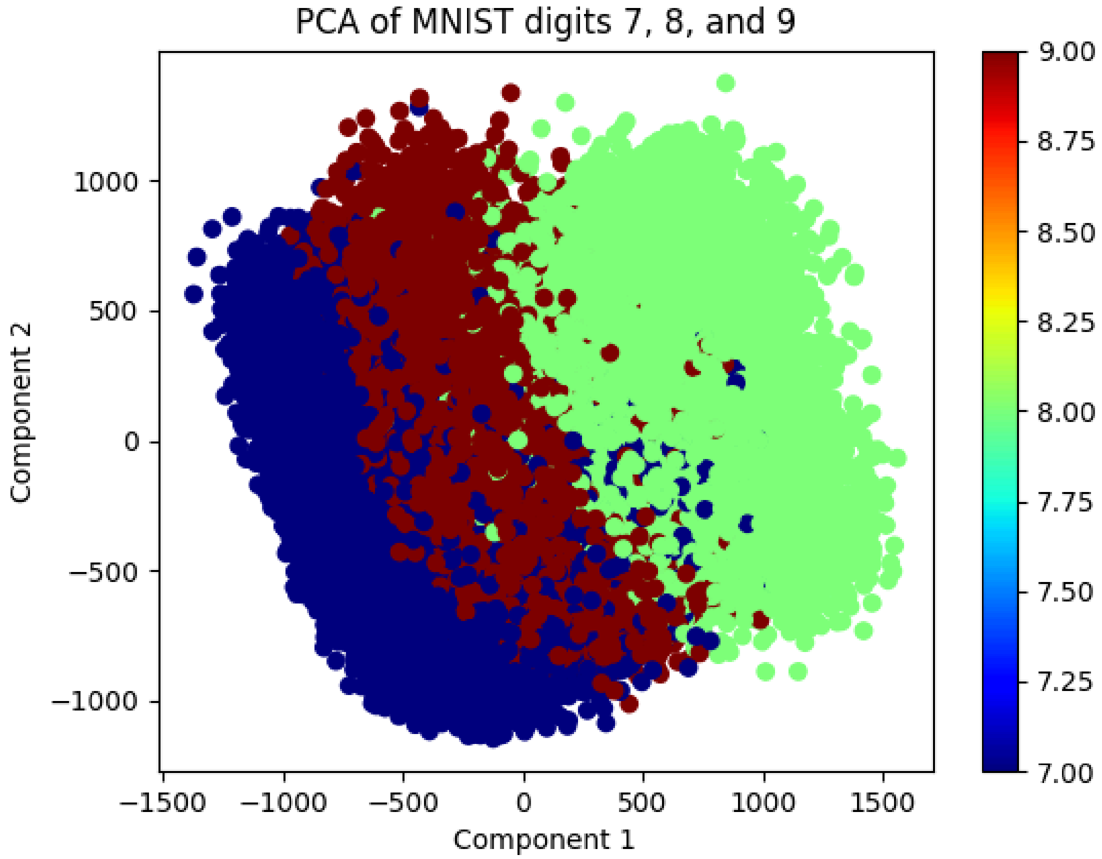
The general conclusion is that the Stacking Classifier showed a performance improvement compared to individual classifiers, with a score of 0.9316. This suggests that combining diverse classifiers using ensemble techniques can yield enhanced predictive capabilities. The Random Forest final estimator in the stacking ensemble played a crucial role in achieving this improved performance. The long runtime could impact the practical applicability of the model or the machine that is running on, especially in scenarios where quick predictions are essential. Consideration should be given to balancing computational efficiency with model performance, and further exploration of alternative models or optimization techniques may be needed.

Exercise 2

1 - 2

The code was constructed as dictated by the exercise.

3)



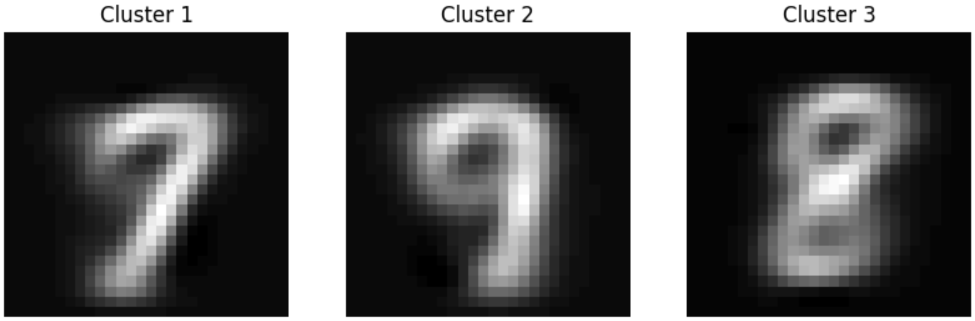
4 - 5

The code produced the following results:
Number of clusters: 2, Silhouette Score: 0.3727215528488159
Number of clusters: 3, Silhouette Score: 0.4312571585178375
Number of clusters: 4, Silhouette Score: 0.3685321509838104
Number of clusters: 5, Silhouette Score: 0.3844033479690552

```
Number of clusters: 6, Silhouette Score: 0.38583704829216003
Number of clusters: 7, Silhouette Score: 0.38510382175445557
Number of clusters: 8, Silhouette Score: 0.3736627995967865
Number of clusters: 9, Silhouette Score: 0.3698442578315735
Number of clusters: 10, Silhouette Score: 0.362096905708313
Best number of clusters: 3
```

As dictated by the score, the best number of clusters is 3 which agrees with the number of digits in the data, so this number will be used for the next question.

6


Cluster 1
Cluster 2
Cluster 3