Student Name: Artem Kamov

Student ID: 301220613

2. The split ration is 80% for the training test and 20% for the testing and validation set. The split ratio for the testing and validation set is 0.5. This ration was chosen because it provides the greater amount of data for the training set compared to other rations, such as 70-75%, which is useful in the given case, as there are limited amount of data in the dataset (400 pictures). So, it improves the performance of the classifiers we are going to train.

Training set size: 320

Validation set size: 40

Test set size: 40

3. Random Forest classifier was chosen to predict which person is represented in each picture. It is quite effective because the non-parametric nature of Random Forest provides flexibility and adaptability allowing the model to capture the complex, non-linear, and high-dimensional nature of the face data without being restricted by predefined assumptions about the data structure.

Random Forest Cross-validation scores: [0.859375 0.921875 0.9375 0.875 0.796875]

Mean cross-validation score: 0.878125

precision recall f1-score support

0 0.50 1.00 0.67 1

1 1.00 1.00 1.00 1

2 0.50 1.00 0.67 1

3 1.00 1.00 1.00 1

4 0.50 1.00 0.67 1

5 1.00 1.00 1.00 1

6 1.00 1.00 1.00 1

7 1.00 1.00 1.00 1

8 1.00 1.00 1.00 1

9 0.00 0.00 0.00 1

10 1.00 1.00 1.00 1

11 1.00 1.00 1.00 1

12 1.00 1.00 1.00 1

13 1.00 1.00 1.00 1

14 0.50 1.00 0.67 1

15 0.00 0.00 0.00 1

16 1.00 1.00 1.00 1

17 1.00 1.00 1.00 1

18 1.00 1.00 1.00 1

19 1.00 1.00 1.00 1

20 1.00 1.00 1.00 1

21 1.00 1.00 1.00 1

22 1.00 1.00 1.00 1

23 0.50 1.00 0.67 1

24 1.00 1.00 1.00 1

25 0.00 0.00 0.00 1

26 1.00 1.00 1.00 1

27 1.00 1.00 1.00 1

28 1.00 1.00 1.00 1

29 1.00 1.00 1.00 1

30 1.00 1.00 1.00 1

31 0.00 0.00 0.00 1

32 1.00 1.00 1.00 1

33 1.00 1.00 1.00 1

34 1.00 1.00 1.00 1

35 0.00 0.00 0.00 1

36 1.00 1.00 1.00 1

37 1.00 1.00 1.00 1

38 1.00 1.00 1.00 1

39 1.00 1.00 1.00 1

accuracy 0.88 40

macro avg 0.81 0.88 0.83 40

weighted avg 0.81 0.88 0.83 40

The mean score cross-validation score of about 0.88 is quite high indicating good performance of the Random Forest classifier on the Olivetti faces dataset.

The test set accuracy is 88%, and precision/recall/f1-scores indicate that most classes were well predicted, though some outliers, for instance class 9, 15, 25, 31 and 35 had no true positives, leading to zero scores.

4. Euclidean distance:



The dendrogram for the Euclidean distance provides a quite balanced hierarchical structure of the data points based on their geometric distance in Euclidean space. It shows multiple branches with almost similar lengths indicating that clusters were formed at relatively similar distance levels. The points appear to be grouped in a well-distributed manner, with distinct clusters formed at varying levels of the hierarchy. Besides, some clusters merge at lower heights, showing that the data points within those clusters are more similar, while others merge at higher levels, indicating less similarity. Overall, the dendrogram demonstrates that Euclidean distance effectively captured the data structure, providing distinct and balanced clusters.

#### Manhattan Distance:



The Manhattan distance dendrogram displays a different structure due to the nature of the distance metric which calculates the absolute differences along axes. The overall structure of the dendrogram has branches that extend to greater heights showing that the Manhattan metric tends to group points more loosely creating clusters that are more spread out. There are fewer balanced clusters compared to the Euclidean dendrogram. The branches are more irregular, with some clusters merging early and others much later which indicates that some groups of points are much more distant from one another in this metric. The variance in branch heights suggests that some clusters are more compact, while others contain data points that are only loosely related which is a hallmark of how Manhattan distance measures the dissimilarity between points. To sum up, the dendrogram for Manhattan distance reflects more distant and less compact clusters.

Cosine Similarity:



The dendrogram shows shorter branches compared to the Euclidean and Manhattan versions. This means that cosine similarity groups points based on how similar their directions are, with values between 0 (completely different directions) and 1 (perfectly aligned vectors). Additionally, the clusters form early at lower levels indicating that the data points are more tightly grouped by their angular similarity. While the dendrogram forms clusters more quickly (at lower levels), there are still distinct groups. The dendrogram is more compact meaning that data points are closely related when using cosine similarity. This dendrogram demonstrates that cosine similarity forms more tightly knit groups than either Euclidean or Manhattan distance.

5. Since it was required to use the silhouette score approach to choose the number of clusters for each method the range of clusters was set to 150 to find the number of clusters with the highest silhouette score. While, it is expected to find about 40 clusters, since there are images of 40 individuals in the dataset, the highest silhouette score might be for higher numbers of clusters.

The best scores:

Euclidean: 0.3 with 123 clusters

Manhattan: 0.28 with 131 clusters

Cosine Similarity: 0.27 with 97 clusters

As we can see, the highest silhouette scores are for cluster numbers of 123, 131 and 97 clusters for each method respectively. Although none of the silhouette scores are particularly high, Euclidean distance clustering offers the highest silhouette score in this case.

6. Euclidean distance for clustering was chosen for training the final classifier because of several reasons. Firstly, the clusters formed using Euclidean distance are well-separated and less likely to overfit to outliers, which is important for enhancing the dataset's performance when training a classifier. Meanwhile, cosine similarity can work well in high-dimensional spaces where direction is more important than magnitude, in this case, Euclidean distance likely captures both the direction and magnitude of the features well. This ensures that clustering augments the dataset in a way that complements both PCA (which is based on Euclidean geometry) and Random Forest. Manhattan Distance is not the best choice either, since it forms more irregular clusters and may not perform as well as Euclidean distance in this scenario, where the features likely do not have a strong grid-like or discrete structure.

Mean cross-validation score (with clusters): 0.884375

precision recall f1-score support

0 0.00 0.00 0.00 1

1 0.50 1.00 0.67 1

2 1.00 1.00 1.00 1

3 1.00 1.00 1.00 1

4 1.00 1.00 1.00 1

5 1.00 1.00 1.00 1

6 1.00 1.00 1.00 1

7 1.00 1.00 1.00 1

8 1.00 1.00 1.00 1

9 0.00 0.00 0.00 1

10 1.00 1.00 1.00 1

11 1.00 1.00 1.00 1

12 1.00 1.00 1.00 1

13 1.00 1.00 1.00 1

14 1.00 1.00 1.00 1

15 0.00 0.00 0.00 1

16 1.00 1.00 1.00 1

17 1.00 1.00 1.00 1

18 1.00 1.00 1.00 1

19 1.00 1.00 1.00 1

20 1.00 1.00 1.00 1

21 1.00 1.00 1.00 1

22 1.00 1.00 1.00 1

23 0.50 1.00 0.67 1

24 1.00 1.00 1.00 1

25 1.00 1.00 1.00 1

26 1.00 1.00 1.00 1

27 1.00 1.00 1.00 1

28 1.00 1.00 1.00 1

29 1.00 1.00 1.00 1

30 1.00 1.00 1.00 1

31 1.00 1.00 1.00 1

32 1.00 1.00 1.00 1

33 1.00 1.00 1.00 1

34 1.00 1.00 1.00 1

35 0.00 0.00 0.00 1

36 1.00 1.00 1.00 1

37 1.00 1.00 1.00 1

38 1.00 1.00 1.00 1

39 0.33 1.00 0.50 1

accuracy 0.90 40

macro avg 0.86 0.90 0.87 40

weighted avg 0.86 0.90 0.87 40

In conclusion, the mean cross-validation score has improved to about 0.884 providing a slight model performance improvement. The test set accuracy of 90% shows a 2% improvement compared to when clustering was not employed. The use of clustering labels has enhanced the F1-scores for several classes, particularly class 24, which saw a notable increase. However, challenges remain, as classes like 9, 15, and 31 continue to have zero scores, pointing to areas that require further attention.

To sum up, the Euclidean distance-based clustering provided the best structure for the data with a silhouette score of 0.3, but it did not significantly enhance classification performance across all classes. Clustering still shows the potential to improve the model’s performance. However, because the Random Forest classifier already performed well, the new cluster labels just slightly improved the classifier results.