Salma Chaaban – 301216551

COMP257 - Assignment 2 – Analysis Report

1. The Olivetti dataset, consisting of 400 instances of 64x64 images of 40 different faces, was loaded using the fetch\_livetti\_faces function from sklearn. Then it was split into features (X) and targets (y).
2. Stratified shuffle split was used to split the data into training, validation, and testing sets to ensure that each set contains the same number of images per person. The ratio for splitting is 80% for training, 10% for validation, and 10% for testing. It was also tried with a 70/15/15 ratio. It is common practice for training sets to be between 70%-90%, validation between 5%-25%, and testing between 5%-25% (Purdue University, 2024).
3. An SVM classifier with k-fold cross-validation (k=5) was trained on the training set and evaluated on the validation set. The average cross-validation accuracy score was 0.95 and the validation accuracy score = 1, which raises the concern of overfitting. For the 70/15/15 ratio, the results were a little different. The average cross-validation accuracy score was also 0.95, but the validation accuracy was 0.98.
4. K-Means was used for clustering the images. The range for the number of images was 2 to 150, it was changed a few times with smaller ranges but was always choosing the highest value possible, therefore it was set to a high range of 150. According to the silhouette score, the best number of clusters was 136, which is high. Since there are 40 different faces, each with 10 images, making up 400 in total, at first, it was assumed that there would be 40 clusters (each with 10 images of a person); however, this was proven not to be true. The images are from different angles which might have made them look like different people. For the 70/15/15 split, the best number of clusters based on the silhouette score was 90.
5. After reducing the dimensionality with K-Means and finding the best number of clusters, the classifier was trained again. The validation accuracy on the reduced data was 0.9, and 0.85 on the 70/15/15 split.
6. The data was scaled then DBSCAN was applied with epsilon 0.5 (default), minimum samples of 5, and the Euclidean distance metric, then the estimated number of clusters by DBSCAN was displayed which was 0, which was also the same as the 70/15/15 split. However, with values eps=0.5, min\_samples=5, metric='cosine’ the estimated number of clusters was 6, a few other epsilon and minimum sample values were tested, with the Euclidean distance metric the values were always 0, but with the cosine, they were greater than 0, and the best epsilon value, that yielded the greatest number of clusters, was 0.4 with 16 estimated number of clusters and 96 noise points (which is high). The value of estimated number of clusters was greater with lower minimum sample values; however, 5 minimum samples were kept as the dataset is moderately large, and a value less than 5 raises the concern of over-segmenting. Overall, cosine similarity is a more suitable algorithm for facial recognition.

**References**

Purdue University. (2024). *Cross Validation: training, validation and testing splits :: The*

*Examples Book*. The Examples Book. <https://the-examples-book.com/starter-guides/data-science/data-modeling/resampling-methods/cross-validation/train-valid-test>