**Assignment 2: K-Means & DBSCAN**

**Written report**

1. Retrieve and load the Olivetti faces dataset
2. Split the training set, a validation set, and a test set using stratified sampling to ensure that there are the same number of images per person in each set. Provide your rationale for the split ratio.
3. Using k-fold cross validation, train a classifier to predict which person is represented in each picture, and evaluate it on the validation set.

**OUTPUT:**  
Cross-validation scores: [0.91666667 0.875 0.9375 0.89583333 0.9375]

Validation score: 0.95

Analysis: The cross-validation scores we obtained represent the accuracy of the SVM classifier on different folds of the training data. These scores are relatively consistent, ranging from 0.875 to 0.9375, indicating that the model performs consistently across various subsets of the training data.

In contrast, the validation score of 0.95 represents the accuracy of the SVM classifier on the validation set, which was not used during training. This high score suggests that the model has effectively learned the patterns in the training data and can generalize well to new data.

Comparing the mean cross-validation score of 0.91 with the validation score of 0.95, we observe that the validation score is slightly higher than the average cross-validation score. This comparison indicates that the model remains robust and performs well on unseen data.

1. Use K-Means to reduce the dimensionality of the set. Provide your rationale for the similarity measure used to perform the clustering. Use the silhouette score approach to choose the number of clusters.
2. Use the set from step (4) to train a classifier as in step (3).

**OUTPUT:**  
Cross-validation scores on reduced data: [0.8125 0.8125 0.83333333 0.79166667 0.83333333]

Validation score on reduced data: 0.875

Analysis: The cross-validation scores obtained on the reduced data from K-Means are fairly consistent, indicating that the model performs consistently across different subsets of the reduced training data. However, when we compare the mean cross-validation score of 0.81 from this step with the mean cross-validation score of 0.91 from step 3, we observe a significant difference. The mean cross-validation score on the reduced data is lower than that on the original data, suggesting a slight drop in performance after reducing the data dimensions.

Similarly, the validation score on the reduced data (0.87) is also lower than the validation score on the original data (0.95). This further indicates a slight decrease in performance following dimensionality reduction. Despite this slight decrease in model performance, dimensionality reduction might still be beneficial if it significantly reduces computational cost or improves model interpretability.

1. Apply DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to the Olivetti Faces dataset for clustering. Preprocess the images and convert them into feature vectors, then use DBSCAN to group similar images together based on their density. Provide your rationale for the similarity measure used to perform the clustering, considering the nature of facial image data.

A graph with blue dots and a line

Description automatically generated

**OUTPUT:**  
For eps=40:

- Number of clusters: 0

- Number of noise points: 240

------------------------------

For eps=45:

- Number of clusters: 2

- Number of noise points: 225

------------------------------

For eps=50:

- Number of clusters: 5

- Number of noise points: 165

------------------------------

For eps=55:

- Number of clusters: 8

- Number of noise points: 106

Analysis: According to the DBSCAN results, the algorithm is unable to find clusters in the dataset if the parameter eps is set below 40. Without using the elbow technique, we might not be aware of this behavior and could struggle to find clusters using minimal eps values. By running the elbow method along with DBSCAN, we discovered that an epsilon ranges from 40 to 60 is optimal for identifying clusters, as suggested by the plot.

When we set eps to 55, the results showed that DBSCAN identified 8 clusters and 106 noise points.

This exercise allowed us to find the optimal number of clusters in the Olivetti faces dataset based on density. This density-based clustering algorithm is excellent for identifying clusters of varying shapes and sizes and detecting noise (outliers). It does not require specifying the number of clusters beforehand and is particularly useful for grouping data points that are similar based on density. However, it might not directly correspond to the number of distinct categories or labels in the data, such as different individuals in the Olivetti faces dataset.