**Lab Assignment #2: K-Means & DBSCAN Clustering**

The objective of this assignment is to explore clustering techniques on the Olivetti faces dataset using both the K-Means and DBSCAN algorithms.

**1. Data Preparation**

The Olivetti faces dataset, provided by the fetch\_olivetti\_faces function from Scikit-Learn, contains grayscale images of 40 distinct individuals. Scikit-Learn conveniently offers both flattened 1D and 2D versions of these image features, as well as their corresponding labels.

A screenshot of a computer program

Description automatically generated

For enhanced visualization and analysis, we packed these features and labels into a Pandas DataFrame. The dataset consists of 400 samples, where each sample is a vector of 4097 dimensions (4096 pixels + 1 label).

A screenshot of a computer

Description automatically generated

To understand the intrinsic patterns and similarities, we plotted 40 sample images.

A collage of different faces

Description automatically generated

**2. Data Splitting**

Given the limited number of samples, it's crucial to ensure maximum utility from each data point. We allocated 70% of the data for training, and equally divided the remainder between validation and testing (15% each).

A screenshot of a computer program

Description automatically generated

**3. Initial SVM Classification**

Our initial classification using a Support Vector Machine (SVM) yielded promising results. The 5-fold cross-validation on the training data reported an average accuracy of approximately 88%. This is commendable, especially given the dataset's complexity and high dimensionality.

The validation set achieved slightly better accuracy, while the test set performance aligned closely with our cross-validation results.

A screenshot of a computer

Description automatically generated

**4. K-Means Clustering for Dimensionality Reduction**

We subjected the training data to K-Means clustering across a range of potential cluster sizes, evaluating the silhouette score for each configuration.

A screenshot of a computer program

Description automatically generated

The silhouette analysis revealed the highest score for a configuration with 79 clusters.

A screen shot of a graph

Description automatically generated

However, reducing our data's dimensionality to these 79 clusters and reapplying our SVM classifier led to decreased accuracy. This suggests that the centroids of the 79 clusters might not capture the dataset's underlying variance adequately.

A screenshot of a computer

Description automatically generated

**5. DBSCAN Clustering**

Turning our attention to DBSCAN, we focused on optimizing its two primary hyperparameters: eps and min\_samples. Our goal was to approximate the natural division of the dataset, targeting 40 clusters, corresponding to the 40 distinct individuals.

A screenshot of a computer program

Description automatically generated

Unfortunately, the high dimensionality combined with the scarcity of samples made accurate clustering challenging. DBSCAN often misgrouped different individuals into the same cluster.

A collage of a person's face

Description automatically generated

However, in certain configurations, DBSCAN showed potential, delivering reasonably accurate clusters.

A collage of a person's face

Description automatically generated

**Conclusion**

This assignment showcased the challenges of clustering high-dimensional data with a limited number of samples. While SVM delivered decent classification accuracy on the original data, dimensionality reduction via K-Means and clustering via DBSCAN were less straightforward. Fine-tuning and further experimentation, possibly with other techniques or algorithms, could lead to better results in future endeavors.