# Report on Olivetti Faces Dataset Analysis

## 1. Retrieval and Loading of the Olivetti Faces Dataset

The Olivetti faces dataset was successfully retrieved and loaded using the fetch\_olivetti\_faces() function. This dataset contains images of faces, with a total of 400 samples and each image having 4096 features (pixel values).

Number of Samples: 400

Number of Features: 4096

## 2. Stratified Sampling for Train, Validation, and Test Sets

The dataset was split into training, validation, and test sets using a stratified sampling method to maintain the same distribution of images per person in each set. A 70-15-15 split ratio was chosen for the following reasons:

- Training Set (70%): Provides a robust base of data for training the model, allowing it to learn the features effectively.

- Validation Set (15%): Used for tuning hyperparameters and preventing overfitting. This ensures that the model's performance is assessed on unseen data.

- Test Set (15%): Evaluates the final performance of the model after training and validation, providing an unbiased measure of accuracy.

The sizes of the sets are as follows:

Training Set Size: 280

Validation Set Size: 60

Test Set Size: 60

## 3. K-Fold Cross Validation and Classifier Training

A classifier was trained using k-fold cross-validation (with k=5) to predict the identity of the individuals in the images. A logistic regression model was implemented within a pipeline that included standard scaling. The average metrics obtained from the cross-validation process were:

Average Accuracy: 0.9536

Average Precision: 0.9450

Average Recall: 0.9550

Average F1 Score: 0.9440

After training, the classifier was evaluated on the validation set, yielding the following classification report:

Class Precision Recall F1-Score Support  
0 1.00 1.00 1.00 2  
1 1.00 1.00 1.00 1  
2 1.00 1.00 1.00 1  
3 1.00 1.00 1.00 1  
4 1.00 0.50 0.67 2  
...  
39 0.67 1.00 0.80 2  
  
Overall accuracy on the validation set was reported as 95%.

## 4. Dimensionality Reduction using K-Means

K-Means clustering was used to reduce the dimensionality of the dataset. Before clustering, PCA was applied for initial dimensionality reduction, selecting 100 components. The optimal number of clusters was determined using the silhouette score method. The results indicated that for n\_clusters = 14, the silhouette score achieved was 0.0952, suggesting this as an appropriate choice for clustering.

## 5. Training Classifier on K-Means Reduced Set

Using the feature set obtained from K-Means clustering, a logistic regression classifier was trained similarly to the previous step (3). This allowed for assessing how dimensionality reduction impacts classification performance.

## 6. Clustering with DBSCAN

DBSCAN was applied to the Olivetti Faces dataset for further clustering. Prior to clustering, images were preprocessed and transformed into feature vectors using PCA. The parameters for DBSCAN were set with eps = 0.5 and min\_samples = 5.

Rationale for Similarity Measure: DBSCAN utilizes a density-based approach, which is particularly suitable for facial image data as it can identify clusters of varying shapes and sizes, accommodating noise. This is critical when dealing with facial images that may have different expressions, lighting conditions, or partial occlusions.

## Conclusion

The analysis provided valuable insights into the Olivetti Faces dataset, successfully implementing various techniques for classification and clustering. The findings indicate high accuracy in predicting individual identities and demonstrate effective dimensionality reduction and clustering strategies suitable for facial image data.