# Report on the Olivetti Faces Dataset Analysis

## Introduction

This report outlines the analysis conducted on the Olivetti Faces dataset, focusing on data preparation, model training, clustering using various distance metrics, and evaluation using silhouette scores and cross-validation. The objectives include classifying individuals based on facial images and clustering the dataset using different distance measures.

## Data Retrieval and Loading

The Olivetti Faces dataset was retrieved and successfully loaded into the environment. This dataset consists of grayscale images of faces, with each image representing one of 40 different individuals.

## Data Splitting

The dataset was split into training, validation, and test sets using stratified sampling to maintain equal representation of images per person across each set. The initial split allocated 80% of the data for training (including a validation subset) and 20% for testing. Subsequently, the training set was further divided into actual training and validation sets using a stratified shuffle split, ensuring balanced representation in each subset.

## Classifier Training and Evaluation

A Random Forest Classifier was trained using k-fold cross-validation on the training set. This method enabled a comprehensive assessment of the classifier's performance on unseen data. The evaluation was performed on the validation set, producing a mean cross-validation score of approximately 87.08%, indicating a solid model performance.

## Clustering Using Different Distance Measures

Clustering was conducted using Agglomerative Hierarchical Clustering (AHC) with three different distance metrics:

1. Euclidean Distance: The dataset was clustered, and the results were visualized. The silhouette score for this clustering method was computed, revealing a best score of approximately 0.15 with 2 clusters.

2. Minkowski Distance: Using a specified parameter for the Minkowski distance, clustering was performed again. The silhouette score obtained was approximately 0.09 with 2 clusters, demonstrating a lower clustering quality compared to the Euclidean method.

3. Cosine Similarity: The clustering was conducted using cosine similarity transformed into a distance matrix. The silhouette score was around 0.14, indicating clustering performance that was moderately better than the Minkowski distance but less than the Euclidean distance.

## Discrepancies Between Distance Measures

The clustering results varied across the three distance metrics. Euclidean distance provided the best silhouette score, suggesting better-defined clusters. Minkowski distance yielded lower scores, implying that the clustering structure was less distinct. Cosine similarity showed moderate performance, indicating that the nature of the data (images) affected clustering outcomes differently depending on the distance measure used.

## Silhouette Score Analysis

Silhouette scores were computed for each distance measure to determine the optimal number of clusters. The results indicated:

- For Euclidean Distance, the best silhouette score occurred with 2 clusters.

- For Minkowski Distance, a similar outcome was observed with 2 clusters.

- Cosine Similarity yielded its best silhouette score with 5 clusters, reflecting a more complex clustering structure.

## Final Classifier Training Using Clustering Labels

The cluster labels obtained from the best performing distance metric (Euclidean distance) were used as new target labels for a Random Forest Classifier. k-fold cross-validation was performed on this adjusted dataset, yielding a mean cross-validation score of approximately 90.25%. This indicates that the clustering labels were informative enough to improve classification performance.

## Conclusion

The analysis of the Olivetti Faces dataset demonstrated effective data handling through stratified sampling and robust classification performance using k-fold cross-validation. Clustering results varied significantly with different distance measures, highlighting the impact of the chosen metric on the clustering outcome. The use of silhouette scores provided a quantitative approach to determine the optimal number of clusters, ultimately enhancing the classification accuracy through the integration of clustering labels. Future work may explore additional distance metrics or more sophisticated clustering algorithms to further improve the results.