Written Report

#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

The split ratio I used was 60% training, 20% validation set, and 20% test set. The reason I choose this was it generalizes the model better when a bigger part of the data is used for training. The remaining part is split evenly into two to allow fine-tuning parameters, and test the models performance after the training is complete. This lets there be enough data for training while still saving data for tuning and testing

# 4 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.

The Euclidian distance works well in this data since K-Means naturally works well when the clusters are spherical and equally sized in feature space I already normalized the features which are pixel intensities ranging from 0 to 1.The distance metric isn’t going to be skewed by features that have larger scales. This aligns with how K-Means calculates cluster centroids by averaging features

#6 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

The similarity measure used for clustering is based on the Euclidean distance between normalized pixel intensity feature vectors. This gets the proximity of images in high-dimensional space, letting us have more accurate grouping of similar faces. The density-based nature of DBSCAN lets us identify clusters with different shapes and sizes. This approach helps group similar images while minimizing the influence of noise and outliers in the dataset.