**COMP 257 SEC 003**

**Sirada Thoungvitayasutee 301292632**

**Assignment 2**

Github: <https://github.com/venussirada/COMP257/blob/main/Assignment2/COMP257_Sirada_Assignment2.ipynb>

**What I did?**

Answer For this assignment, I started by loading the Olivetti faces dataset and visualizing sample images to understand the dataset. Then, I split the training, validation, and testing sets using 80-10-10 ratio and StratifiedShuffleSplit to ensure equal class distribution in each set. Given the dataset’s small size (400 images), the 80-10-10 ratio ensured a sufficient amount of training data.

After that, I trained the data using Logistic Regression model using 5-fold cross-validation, achieving validation accuracy of 95%. I also experimented with SVC, however, its validation accuracy of SVC was 92.5%, little lower than Logistic Regression. Then I applied K-Means and used the silhouette score to find the optimal number of clusters, which was K = 123.

With the reduced dataset, I retrained the classifier and applied DBSCAN for clustering. Initially, using K-Means for dimensionality reduction, DBSCAN showed several noise points, so I used PCA to reduce the data to 2D for better visualization and clustering.

**Lesson Learn:**

* To find the optimal number of clusters, it is important to explore wide range of K values. A small range might miss the true optimal value.
* To visualize the results of DBSCAN, the dataset should be reduced to 2 dimensions, such as using PCA, to make the clusters and points easier to interpret.
* Using a reduced dataset with 123 dimensions can lead to lower accuracy (from 95% to 90%), this might be because some information is lost during the reduction process.

**Output**

1. Dataset

A collage of different faces

Description automatically generated

1. Split ratio

Answer I split the dataset into 80% training, 10% validation and 10% testing. As the dataset contains only 400 images, it is important to have enough data for training.

A number with numbers on it

Description automatically generated with medium confidence

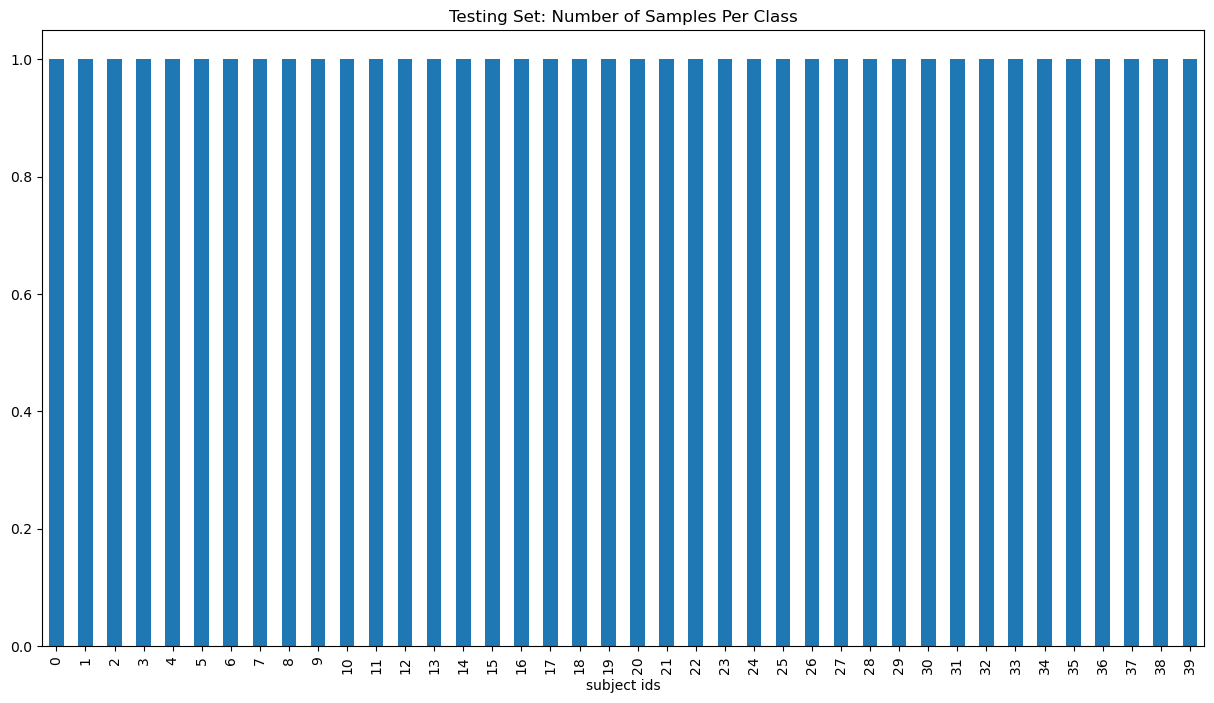
1. Split images per each class

A graph of blue and white lines

Description automatically generated

A graph of blue and white vertical lines

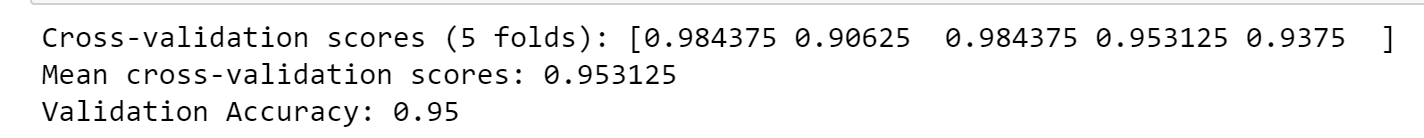
Description automatically generated

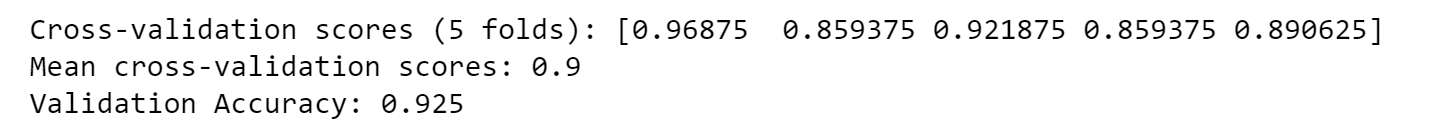


1. Validation accuracy score from using k-fold cross validation

I experimented with both Logistic Regression and SVC, but Logistic Regression achieved higher accuracy, therefore I selected Logistic Regression for training (95%).

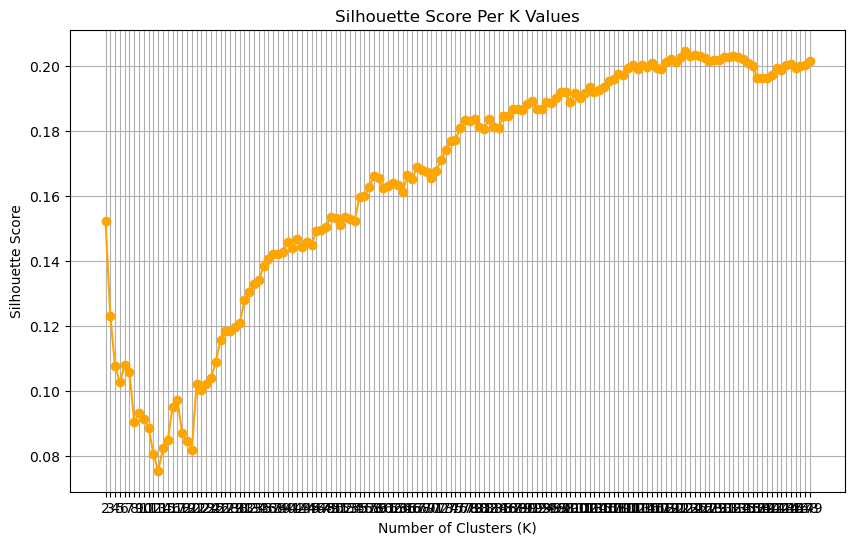
Logistic Regression:



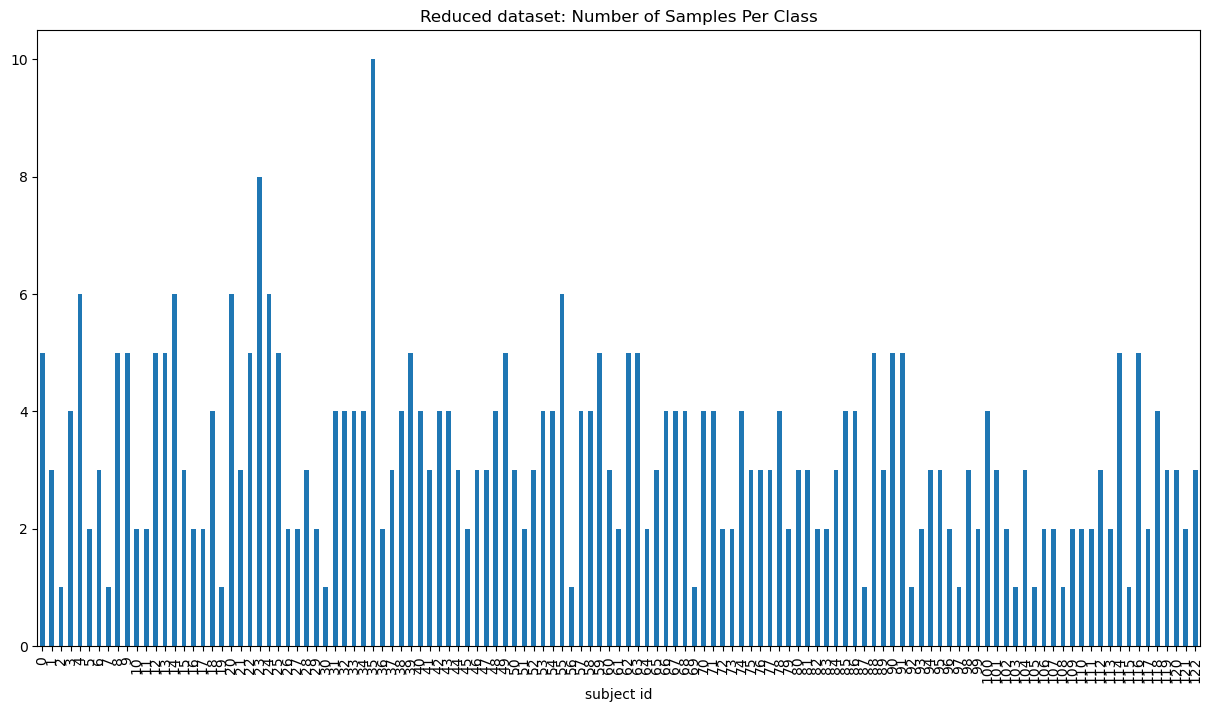
SVC: (For reference) 

1. Optimal K

Best K (optimal K) is 123



1. Reduced dataset has 123 classes

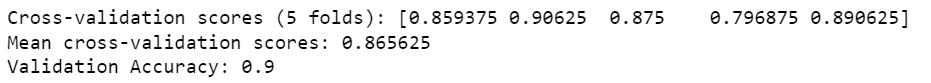


1. Split the dataset from reduced dataset

A white background with black text

Description automatically generated

1. Validation accuracy score for reduced dataset



1. DBSCAN result

