National University of Computer
and Emerging Sciences

Department: Data Science

Name of Assignment: ASSIGNMENT # 3

Name: Muhammad Talal

Section: “DS-N”

Subject: Data Mining

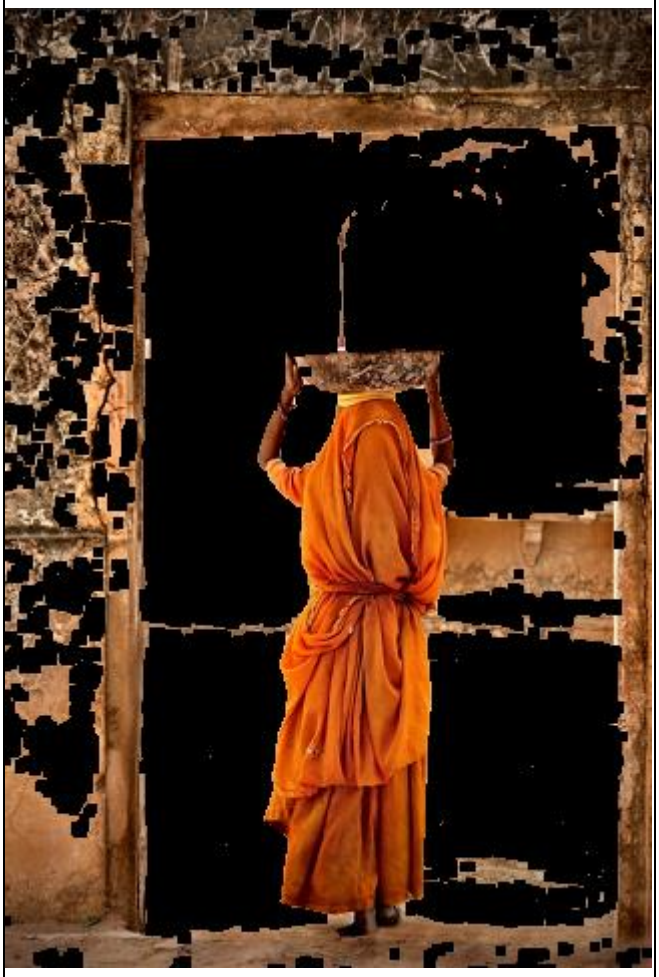
Date of Submission: 30/03/2024

Question:1

Using Custom K-Mean:

Total time: 4.1s

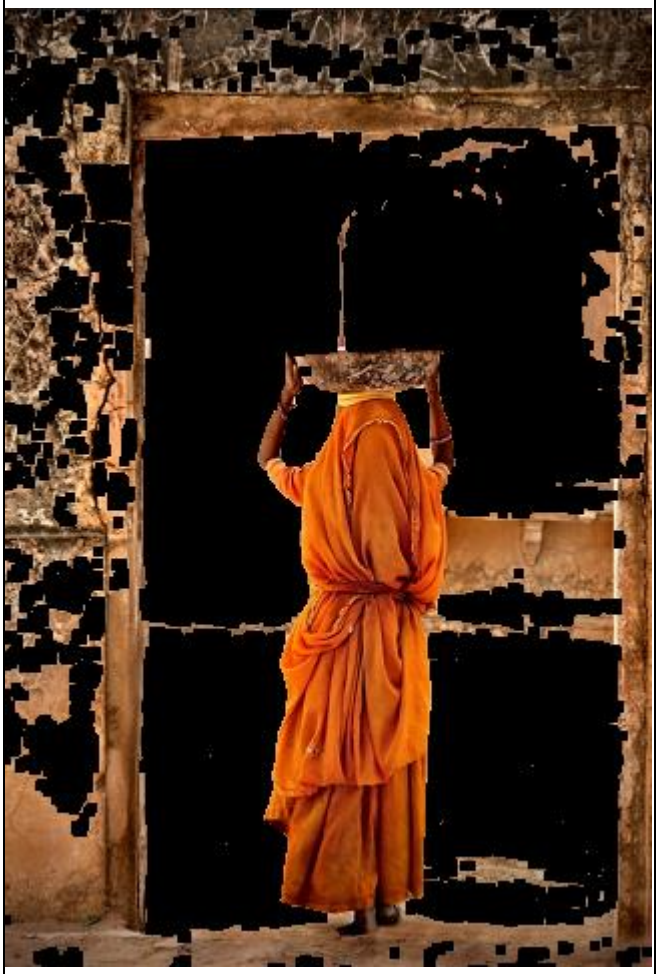
Clusters=8

Original Image:	**Converted Image:**
	

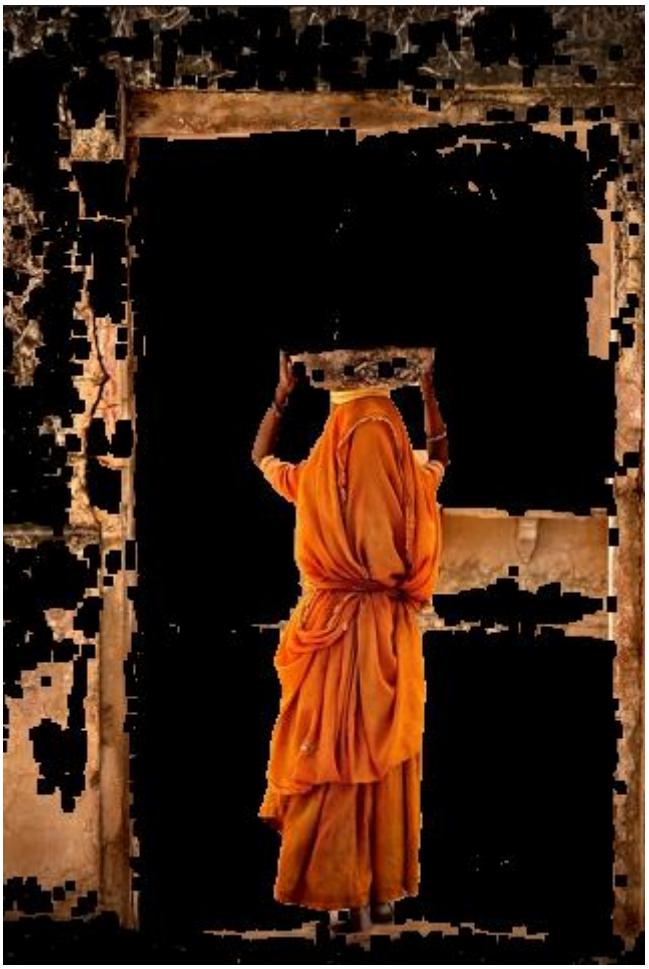
Using Sk Learn K-Mean:

Total time: 2.6s

Clusters=8

Original Image:	**Converted Image:**
	

Using K=3 with custom k-mean:



Question : 2

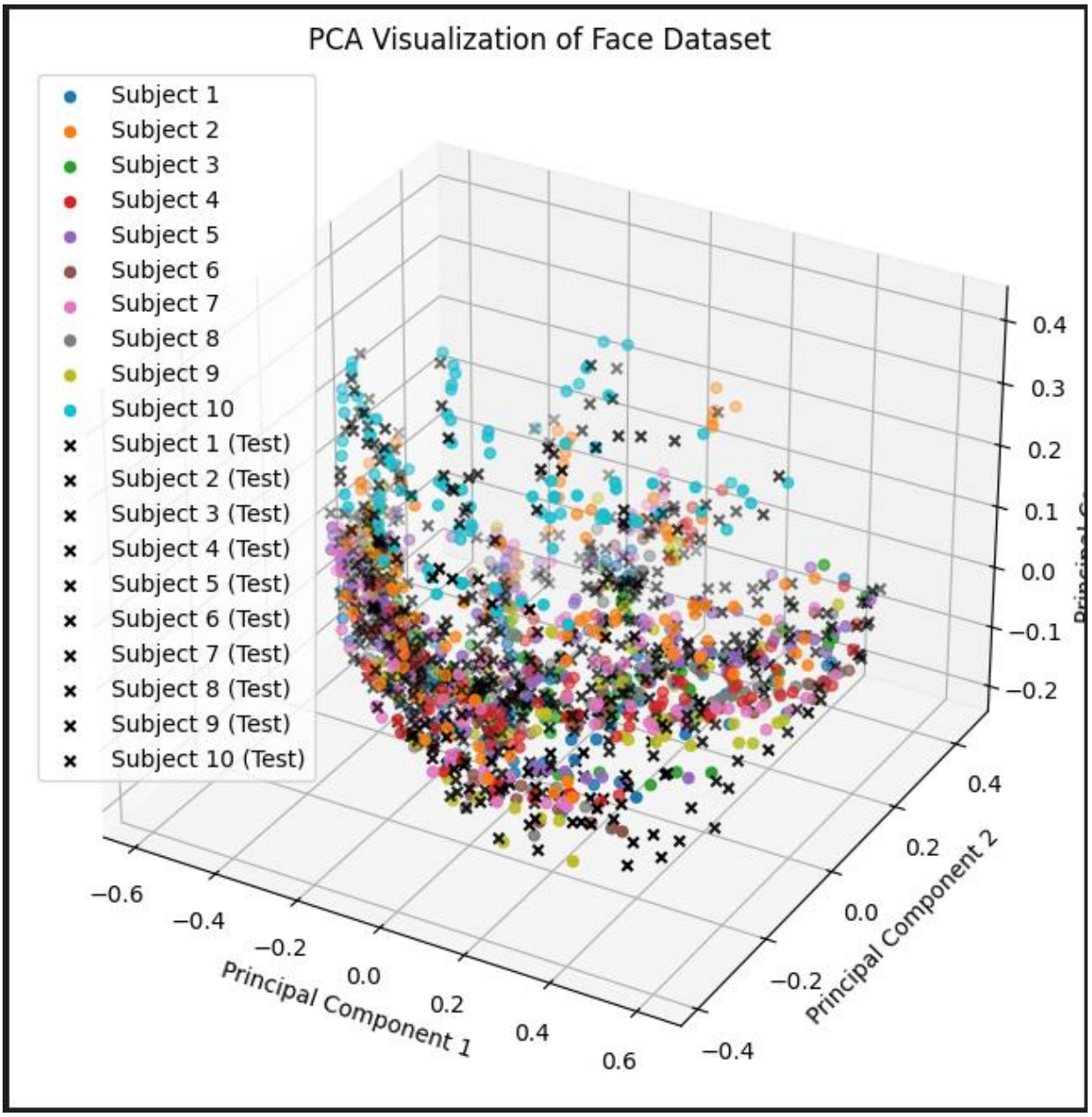
Accuracy:

Using Custom Knn and training data 150 , testing data 20:	Using Custom Knn and training data 150 , testing data 20:
✓ 1m 30.8s [(2, 'euclidean', 0.99), (2, 'cosine', 0.99), (5, 'euclidean', 0.98), (5, 'cosine', 0.98), (7, 'euclidean', 0.96), (7, 'cosine', 0.96), (11, 'euclidean', 0.955), (11, 'cosine', 0.955)]	[7] ✓ 4m 9.7s [(2, 'euclidean', 0.9514285714285714), (2, 'cosine', 0.9514285714285714), (5, 'euclidean', 0.9342857142857143), (5, 'cosine', 0.9342857142857143), (7, 'euclidean', 0.9242857142857143), (7, 'cosine', 0.9242857142857143), (11, 'euclidean', 0.9114285714285715), (11, 'cosine', 0.9114285714285715)]

Using Built in SVM and Gaussian:

```
SVM Accuracy: 1.0
GaussianNB Accuracy: 0.85
```

PCA Analysis:



Explanation:

1. Data Clustering: The plot shows that the data points (representing images) are clustered around their respective subjects. Each subject's data points are depicted in a unique color and symbol, differentiating between training and testing datasets (solid dots for training and crosses for testing).
2. Dimensionality Reduction: Since the original data has 1024 features (each image is 32x32 pixels), PCA has been used to reduce these features to the three most important principal components, which are the axes of the 3D plot.
3. Variance: The spread of data points along each principal component axis represents the variance captured by that component. PCA aims to capture the maximum variance in the least number of components. Typically, the first three components capture the most variance.
4. Overlap between Classes: There's some overlap between subjects in the center of the plot, indicating that some images of different subjects are similar.