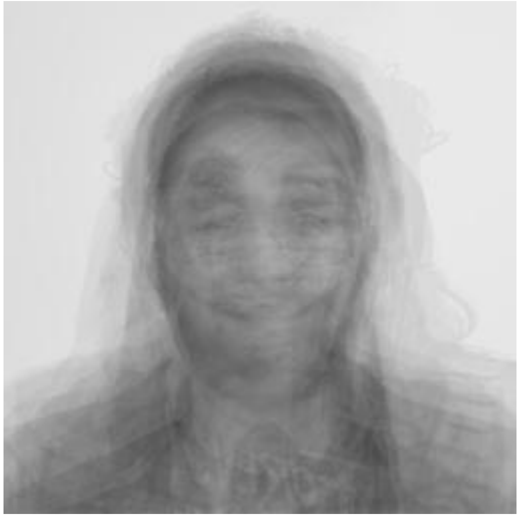
Lab I – Dimensionality Reduction

1

A rectangular matrix was created, and its Rank, Trace, and Determinant were calculated. Since it is not a square matrix, it is not invertible. The transpose of the matrix was computed, and the operations AA’ and A’A were performed to then calculate the eigenvalues and eigenvectors. It was observed that the number of eigenvalues for each matrix depends on the Rank, and they share the same count of non-zero and equal values since they are the squares of the singular values of A

2

A picture of the face was taken and resized to 256 x 256 pixels in grayscale. The photos of the classmates from the course were loaded, and the average of the matrices forming the images was calculated and displayed as the result. Finally, the Euclidean distance between the image of my face and the calculated average with the other images was computed.



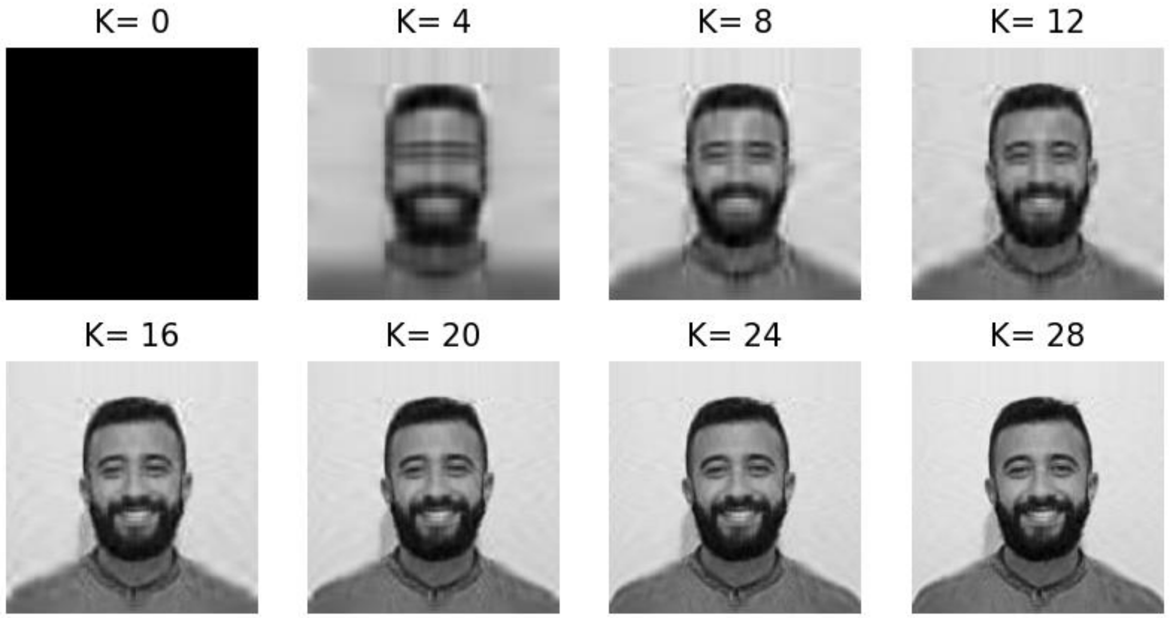
The distance between my face and the average is: 38338.10557395866

3

A package named "unsupervised_package" was created, containing modules for SVD and PCA developed using Python and NumPy.

4

The SVD module was used for reconstructing images based on singular values. It was observed that in this case, starting from k = 12, it was already possible to identify the image.



5

A logistic regression model was trained using the MNIST dataset to classify images between 0 and 8. The performance of this model is reflected in an accuracy of 99%.

6

The dimensionality reduction modules were used in the previous exercise, and it was identified that it affected the model's accuracy, resulting in a 95% accuracy using PCA and SVD.

7

The same process as the previous point was carried out, but in this case, the integrated modules of Scikit-Learn were used. It was identified that using these modules produces similar results to point 6, allowing us to conclude that there are no differences in using the modules built with NumPy and those from the Scikit-Learn library.

8

Data Standardization

• Scale Equality: Standardization could contribute to ensuring that all variables have comparable variance, potentially avoiding those with larger magnitudes dominating PCA. This suggests that it would be crucial to ensure that all features contribute equitably to the total variability and that the analysis is based on the actual structure of the data rather than the scale of the variables.

• Reduction of Outlier Impact: Standardization could reduce the influence of outliers by measuring all variables in standard deviation units. This might enhance the robustness of PCA by minimizing the impact of extreme observations on the estimation of covariance and principal components.

9

UMAP (Uniform Manifold Approximation and Projection):

UMAP is based on the principle of preserving proximity. It uses graph theory and differential geometry to model local relationships between points in a high-dimensional space. The algorithm seeks to find a representation in a low-dimensional space that preserves these local relationships. UMAP utilizes cost functions that minimize the difference between local neighborhood distributions in high and low-dimensional spaces. Additionally, it incorporates stochastic optimization techniques to enhance efficiency and scalability.

10

LDA (Latent Dirichlet Allocation):

LDA is based on probabilistic inference and statistical distribution theory. It employs the Dirichlet distribution to model variability in the mixture of topics in documents and the distribution of words across topics. The model assumes that each document is a mixture of various topics, and each word in the document is generated from one of those topics. Inference is performed using methods like Gibbs sampling or variational inference to estimate latent topic distributions in the document corpus.

11

An HTML and JavaScript code has been developed to create an interactive web interface that enables users to draw on a canvas of 28x28 pixels. The drawn pixels are converted into an array and sent via an HTTP POST request to a server on port 8000. Subsequently, this array is compared with the trained MNIST model to make a prediction