CLUSTERING SIMILAR FACES USING K-MEANS AND DBSCAN

IBM Machine Learning Professional Certificate

Course 05: Unsupervised Machine Learning

By Moustafa Abada

Content:

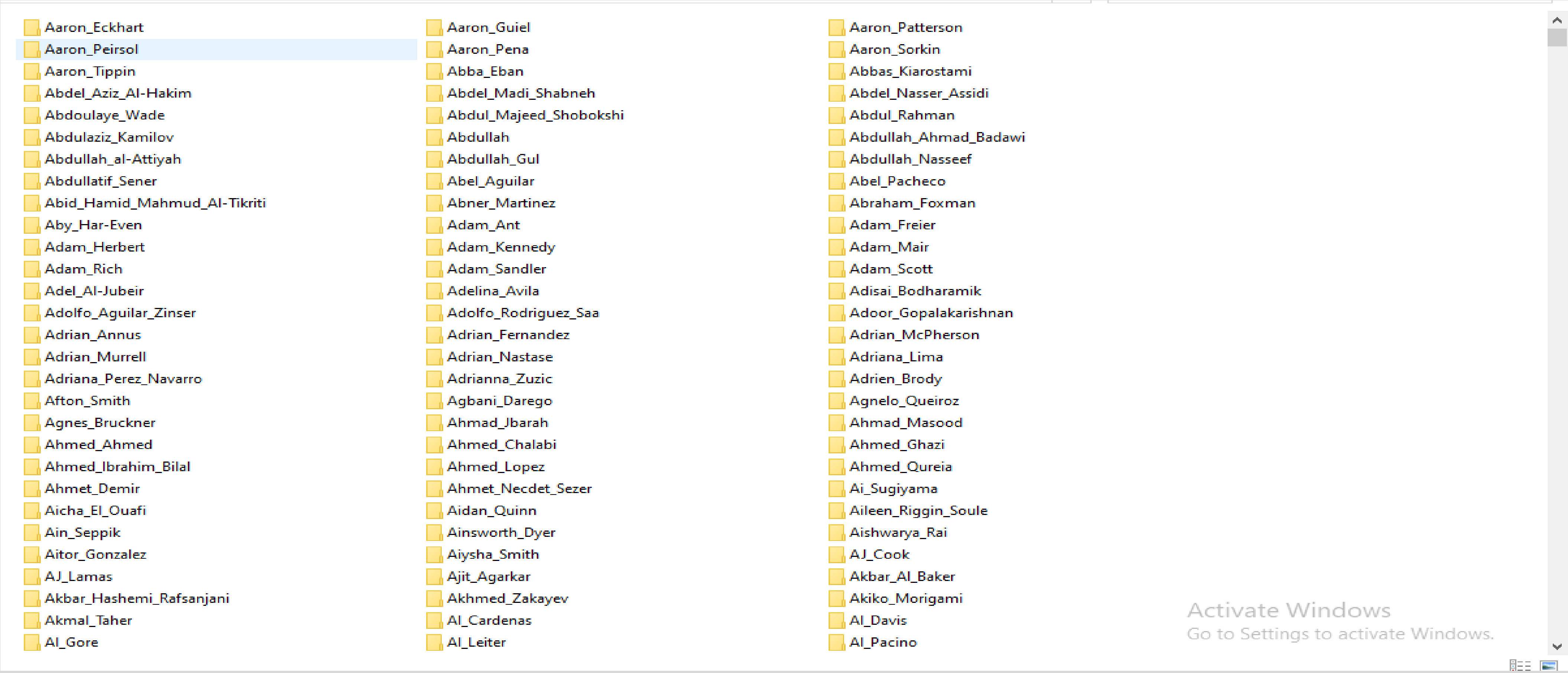
- Dataset Description
- Main objectives of the analysis.
- Applying various machine learning models.
- Model analysis and findings.
- Models flaws and advanced steps.

DATASET DESCRIPTION

Dataset Description

- Labeled Faces in the Wild (LFW), a database of face photographs designed for studying the problem of unconstrained face recognition. The data set contains more than 13,000 images of faces collected from the web. Each face has been labeled with the name of the person pictured. 1680 of the people pictured have two or more distinct photos in the data set.
- You can download the dataset from here: http://vis-www.cs.umass.edu/lfw/

Sample of the Dataset



Aaron_Eckhart
Aaron_Peirsol
Aaron_Tippin
Abdel_Aziz_Al-Hakim
Abdoulaye_Wade
Abdulaziz_Kamilov
Abdullah_al-Attiyah
Abdullatif_Sener
Abid_Hamid_Mahmud_Al-Tikriti
Aby_Har-Even
Adam_Herbert
Adam_Rich
Adel_Al-Jubeir
Adolfo_Aguilar_Zinser
Adrian_Annus
Adrian_Murrell
Adriana_Perez_Navarro
Afton_Smith
Agnes_Bruckner
Ahmed_Ahmed
Ahmed_Ibrahim_Bilal
Ahmet_Demir
Aicha_El_Ouafi
Ain_Seppik
Aitor_Gonzalez
AJ_Lamas
Akbar_Hashemi_Rafsanjani
Akmal_Taher
Al_Gore

Aaron_Guiel
Aaron_Pena
Abba_Eban
Abdel_Madi_Shabneh
Abdul_Majeed_Shobokshi
Abdullah
Abdullah_Gul
Abel_Aguilar
Abner_Martinez
Adam_Ant
Adam_Kennedy
Adam_Sandler
Adelina_Avila
Adolfo_Rodriguez_Saa
Adrian_Fernandez
Adrian_Nastase
Adrianna_Zuzic
Agbani_Darego
Ahmad_Jbarah
Ahmed_Chalabi
Ahmed_Lopez
Ahmet_Necdet_Sezer
Aidan_Quinn
Ainsworth_Dyer
Aiysha_Smith
Ajit_Agarkar
Akhmed_Zakayev
Al_Cardenas
Al_Leiter

Aaron_Patterson
Aaron_Sorkin
Abbas_Kiarostami
Abdel_Nasser_Assidi
Abdul_Rahman
Abdullah_Ahmad_Badawi
Abdullah_Nasseef
Abel_Pacheco
Abraham_Foxman
Adam_Freier
Adam_Mair
Adam_Scott
Adisai_Bodharamik
Adoor_Gopalakarishnan
Adrian_McPherson
Adriana_Lima
Adrien_Brody
Agnelo_Queiroz
Ahmad_Masood
Ahmed_Ghazi
Ahmed_Qureia
Ai_Sugiyama
Aileen_Riggin_Soule
Aishwarya_Rai
AJ_Cook
Akbar_Al_Baker
Akiko_Morigami
Al_Davis
Al_Pacino

DATA ANALYSIS

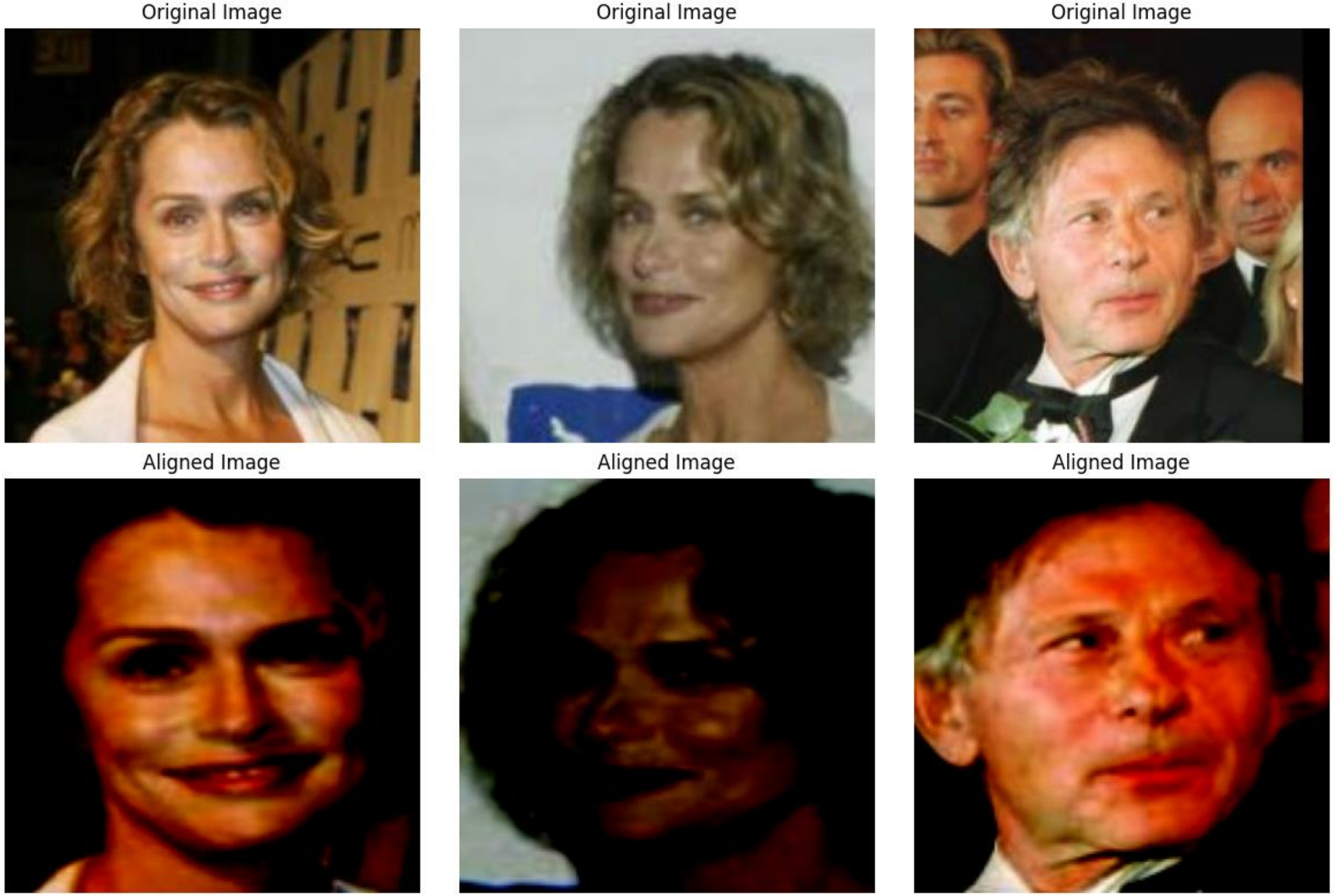
Main objectives of the analysis:

- This analysis has the following goals:

1. Load and Align Faces.
2. Find the Embeddings of Face.
3. Use Euclidian Distance to Find the Closest Embeddings and Cluster Similar Face Together.

DATA ANALYSIS:

Original Faces Vs Aligned Faces



Original Image

Original Image

Original Image

Aligned Image

Aligned Image

Aligned Image

Main Functions Implemented:

1. **load_and_align_data:** Initialize the MTCNN face detector and align images using specified parameters.
2. **generate_embeddings:** Load images, then generate embeddings using a pretrained InceptionResnetV1 model.
3. **calculate_distance_matrix:** Calculate the Euclidean distance matrix between embeddings.
4. **sort_faces:** Sort faces based on their distances to create a sorted list.
5. **Cluster Faces:** Cluster Facing using K-means and DBSCAN Clustering Algorithms.
6. **Visualizing The Embeddings:** Use PCA to reduce the dimensionality of face embeddings to 2D and visualizes the embeddings with cluster labels using a scatter plot.

APPLYING VARIOUS MACHINE LEARNING MODELS

Unsupervised Machine Learning Models Used:

K-means:

- K-means is a clustering algorithm that aims to partition data points, in this case face embeddings, into a specified number of clusters based on minimizing the within-cluster sum of squares. It assigns each embedding to the nearest centroid, which effectively groups similar faces together based on their embeddings. By specifying the number of clusters (n_clusters=8), the algorithm identifies distinct groups of similar faces.

DBSCAN:

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that groups together data points, or face embeddings, that are close to each other and separates regions of higher density from regions of lower density. It identifies dense regions of embeddings and considers them as clusters, effectively grouping similar faces together. The eps parameter (0.55 in this analysis) defines the maximum distance between two embeddings for them to be considered part of the same neighborhood, while the min_samples parameter (4 in this analysis) specifies the minimum number of embeddings required to form a dense region. Unlike K-means, DBSCAN does not require specifying the number of clusters in advance and can automatically detect the number of clusters based on the density of the embeddings.

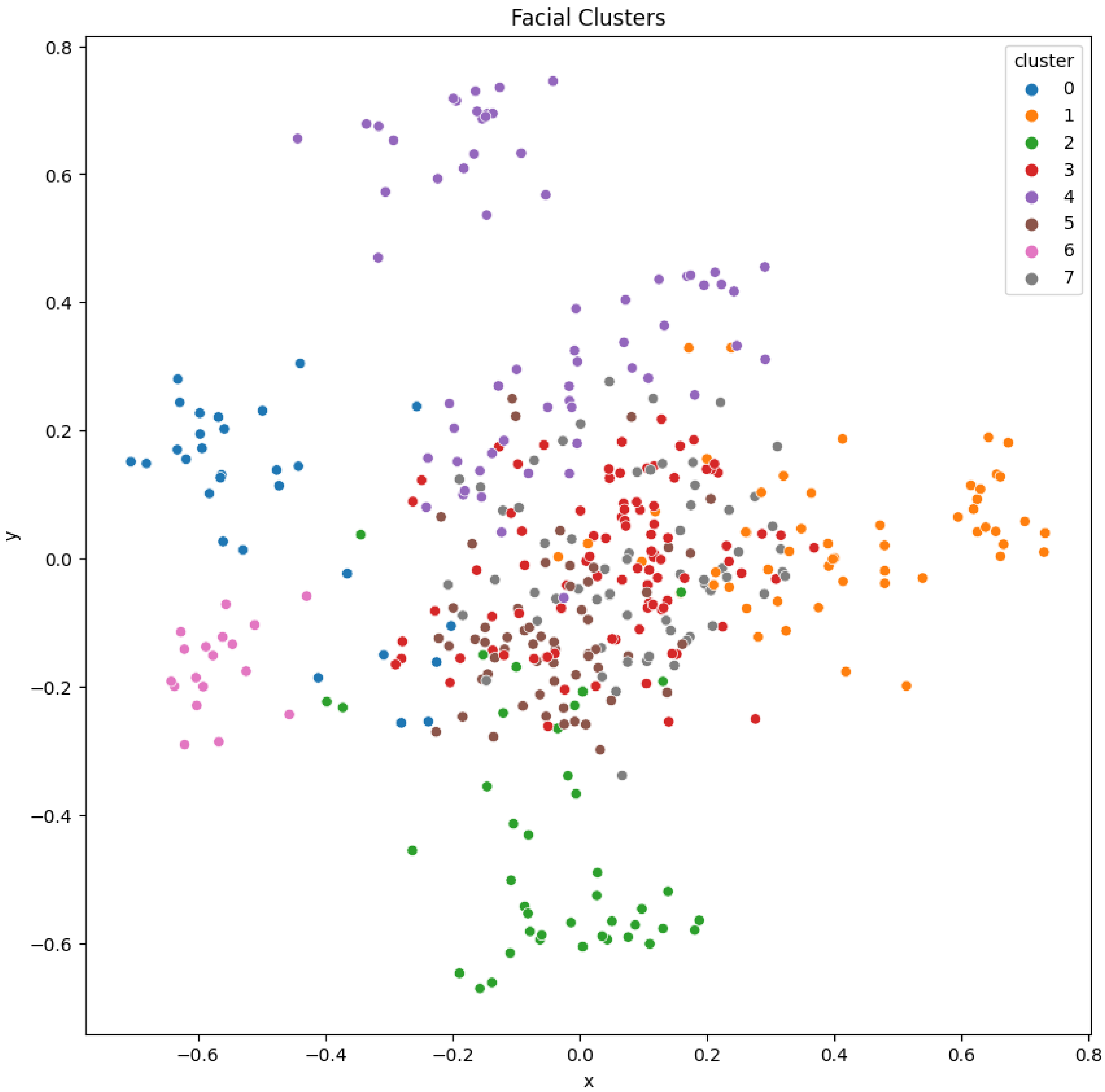
PCA

- It’s used to reduce the dimensionality of face embeddings to 2D and visualizes the embeddings with cluster labels using a scatter plot.

MACHINE LEARNING ANALYSIS AND FINDINGS

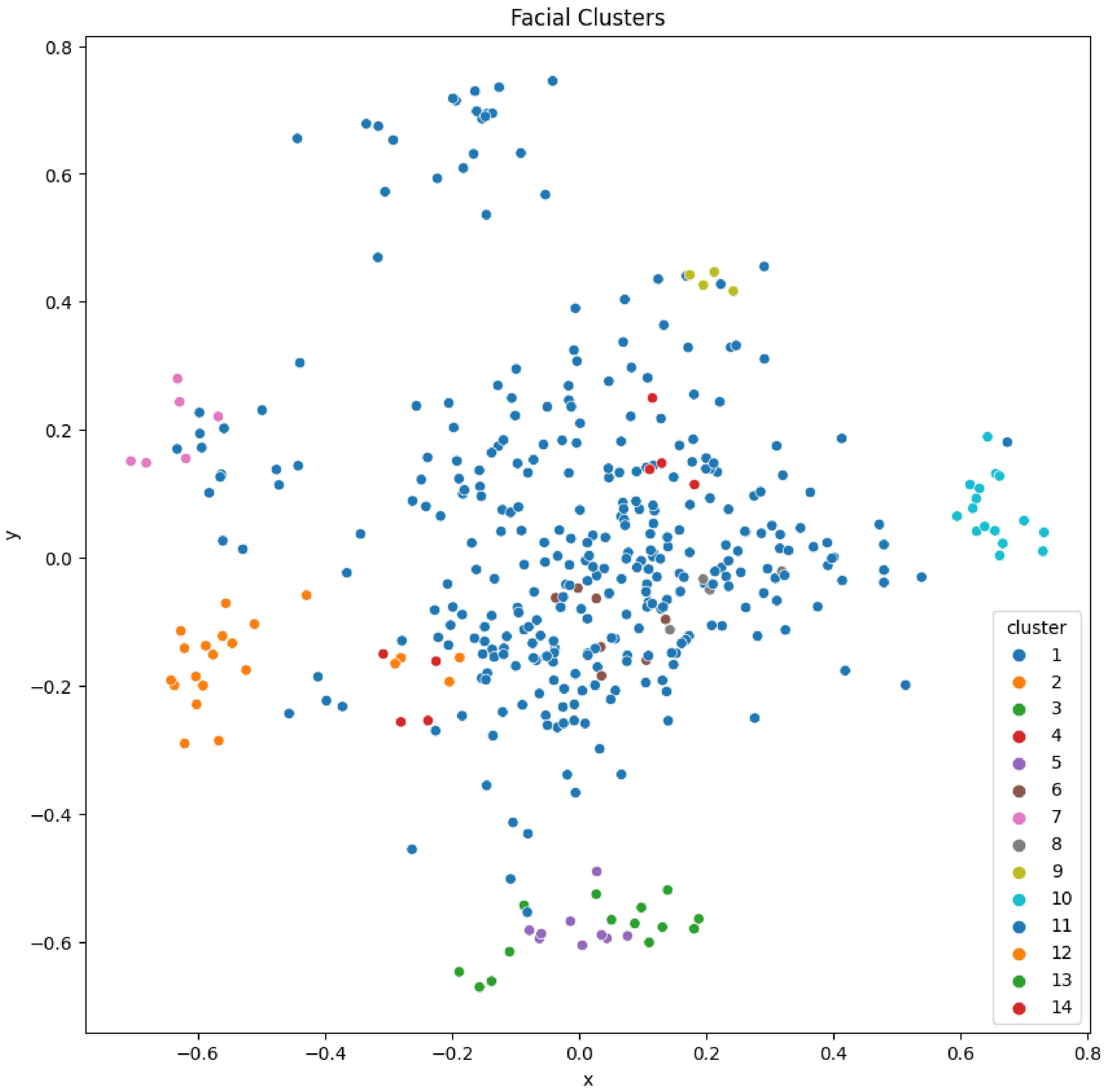
MODELS ANALYSIS AND FINDINGS:

- PCA has successfully reduced the dimensionality of face embeddings to 2D and helped visualizing the embeddings with cluster labels using a scatter plot.
- K-means has demonstrated effective clustering of similar embeddings by assigning each embedding to the nearest centroid, enabling the grouping of similar faces based on their embeddings.
- DBSCAN has shown median success in grouping face embeddings that are close to each other and separating regions of different densities. DBSCAN struggled to differentiate between individuals with slightly similar embeddings, leading to a large portion of the dataset being clustered together as a single group. Nonetheless, DBSCAN tends to successfully identify individuals with highly distinctive embeddings, assigning them to unique clusters.



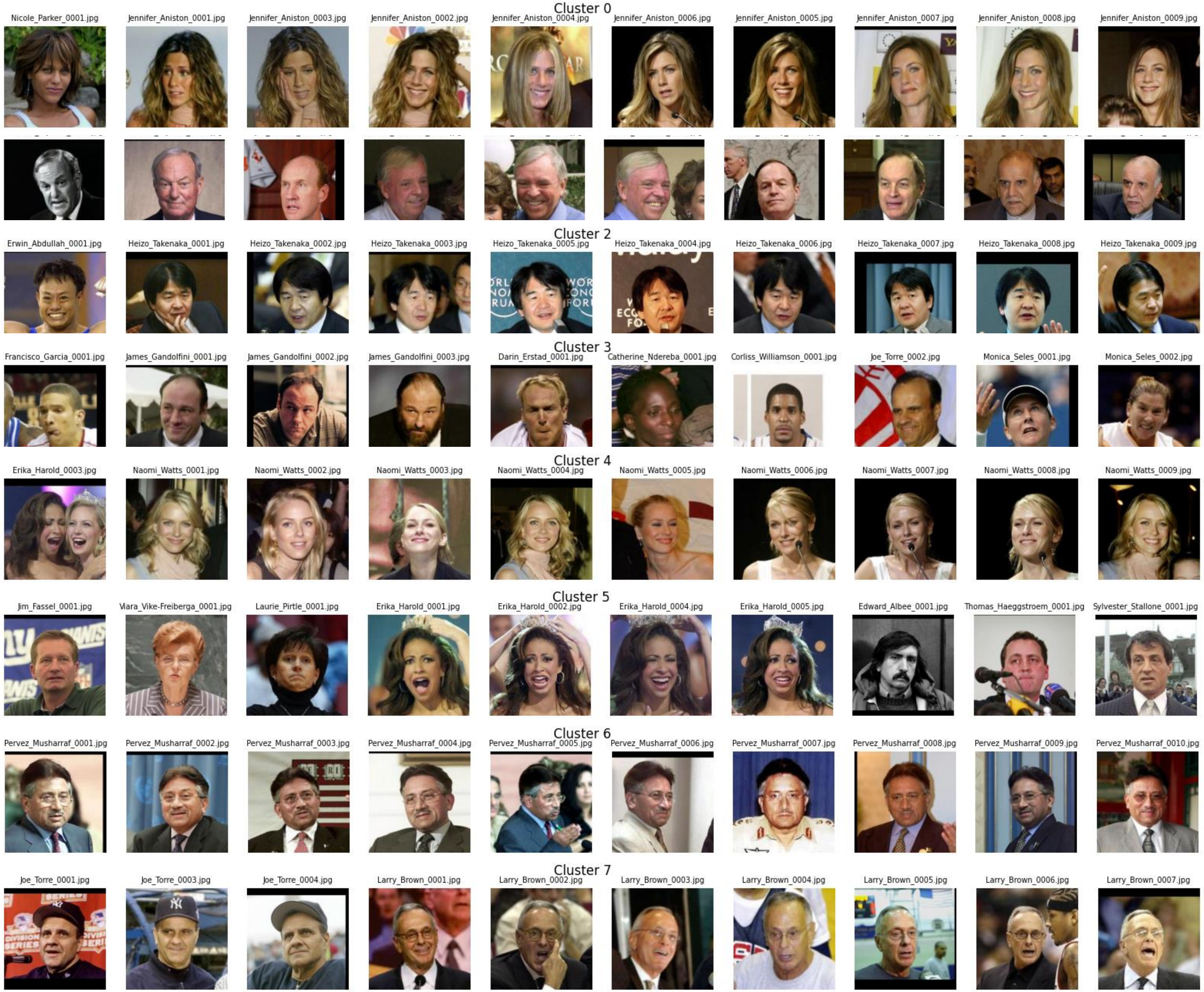
MODEL ANALYSIS AND FINDINGS:

2D embeddings visualization using PCA for K-means 8 Clusters:

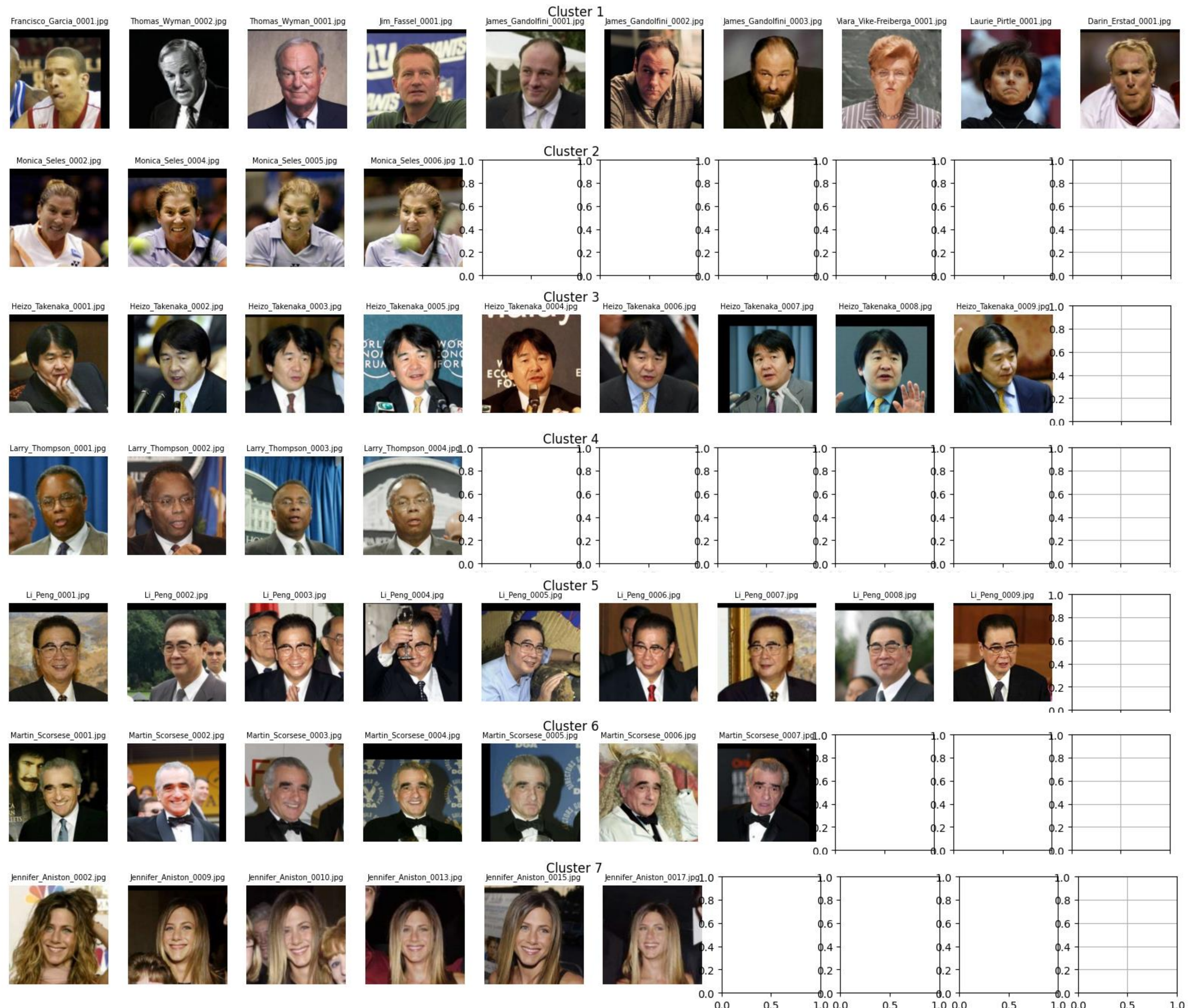


MODEL ANALYSIS AND FINDINGS:

2D embeddings visualization using PCA for DBSCAN Clustering Algorithm:



THE 8 CLUSTERS MADE BY K-MEAN.



SOME CLUSTERS MADE BY DBSCAN.

MODELS FLAWS AND ADVANCED STEPS

Model Flaws and Advanced Steps:

- One potential flaw in the clustering process is that it relies solely on the embeddings to determine similarity. While embeddings capture certain facial features, they may not capture all aspects that humans use to assess similarity, such as hairstyle, facial expressions, or accessories. This could result in misclassifications or grouping of dissimilar faces together.
- Another potential flaw is inaccurate face alignment, which can be addressed by using the original aligned dataset or exploring different alignment techniques for improved accuracy.
- Another potential flaw is the sensitivity of clustering algorithms to the choice of hyperparameters. The number of clusters for K-means and the epsilon and min_samples values for DBSCAN need to be carefully selected. Suboptimal choices can lead to overfitting or underfitting, impacting the quality of the clustering results.

Advanced Steps To be Made:

- To mitigate the limitations of relying solely on embeddings, it is beneficial to incorporate additional features, such as facial landmarks or other facial descriptors, in the clustering process. This can provide a more comprehensive representation of facial similarity.
- Exploring alternative clustering algorithms beyond K-means and DBSCAN can be beneficial. Algorithms like spectral clustering, agglomerative clustering, or hierarchical clustering may offer different perspectives on grouping similar faces together and could potentially yield improved results.
- Consider incorporating a validation or evaluation step to assess the quality of the clustering results. This can involve using external criteria, such as labeled data or human evaluation, to measure the accuracy and effectiveness of the clustering algorithm. This step can help identify potential shortcomings and refine the clustering process.

THANK YOU!

IBM Machine Learning Professional Certificate

Course 05: Unsupervised Machine Learning

By Moustafa Abada