Olivetti Faces – Clustering Analysis

Yifan Wang, Zeyu Wang

Abstract

In this study, we address the challenge of clustering human faces using the Olivetti faces dataset. We applied various clustering algorithms, including k-means, Agglomerative Clustering, DBSCAN, and GMM, following dimensionality reduction through PCA. Initially using silhouette score for evaluation, we shifted to a customized V-Measure combining penalized homogeneity and completeness due to the former's limitations. Our results identified Agglomerative Clustering with 59 clusters as the most effective method, offering a balance between cluster quantity and quality, as evidenced by our novel evaluation metric.

Introduction

In the realm of machine learning, the clustering of high-dimensional data remains a significant challenge, particularly in the context of image recognition. Human faces, with their intricate features and subtle variations, present a unique case for clustering algorithms.

This report focuses on the Olivetti faces dataset, a collection of facial images with varying expressions, lighting conditions, and accessories (like glasses or beards). The dataset's diversity and high dimensionality (64x64 pixel images) make it an ideal candidate for exploring advanced clustering techniques.

The primary objective of this study is to identify an effective method for clustering these facial images. Given the high-dimensional nature of the data, a straightforward application of clustering algorithms could lead to suboptimal results. Therefore, we first employ Principal Component Analysis (PCA) for dimensionality reduction, a crucial step in managing the computational complexity and enhancing the performance of the clustering algorithms.

Our exploration begins with the application of several well-known clustering algorithms: k-means, Agglomerative Clustering, DBSCAN, and Gaussian Mixture Models (GMM). Each algorithm has its strengths and limitations, which we assess in the context of our specific dataset.

Initially, we evaluate the clustering results using the silhouette score. However, this metric's tendency to favor a higher number of clusters led to many sparsely populated or single-member clusters, which are not practical for our purposes. This challenge prompted a shift in our evaluation strategy to focus on homogeneity, leading to the development of a customized v-measure. This new metric combines penalized homogeneity and completeness, offering a more balanced approach to evaluating the effectiveness of the clustering.

Our comprehensive analysis concludes with the identification of Agglomerative Clustering, configured to form 59 clusters, as the most effective method. This approach not only achieved the highest score on our customized v-measure but also produced visually coherent and meaningful clusters.

Through this study, we aim to contribute to the broader understanding of clustering high-dimensional data, particularly in the field of image recognition, and demonstrate the importance of tailored evaluation metrics in achieving meaningful clustering results.

Background

This project builds upon established concepts and algorithms within the field of machine learning, particularly in unsupervised learning and clustering. Understanding these foundational elements is crucial for comprehending the methods and approaches used in this study.

- Principal Component Analysis (PCA): A technique for reducing the dimensionality of large datasets, increasing interpretability while minimizing information loss.
- K-Means: Divides data into K clusters, each represented by the mean of its points.
- Agglomerative Clustering: A hierarchical clustering method that merges data points or clusters based on their similarity.
- DBSCAN: Clusters points based on density, effectively separating high-density groups from low-density outliers.
- Gaussian Mixture Models (GMM): Models the data as a mixture of multiple Gaussian distributions.
- Silhouette Score: Assesses the separation distance between the resulting clusters.
- Homogeneity, Completeness, V-Measure: Evaluate the quality of clustering based on how well the clusters contain only members of a single class and how completely all members of a single class are in a cluster.

Related Work

In the realm of clustering algorithms and facial data analysis, a substantial body of work exists that provided the foundation and inspiration for our project. Generally, these works focus on applying traditional clustering techniques, such as K-Means, Agglomerative Clustering, and DBSCAN, to various datasets. Notably, many studies have employed these methods on facial datasets, albeit often on lower-dimensional data compared to the Olivetti faces dataset. These studies typically utilize standard metrics like Silhouette scores to evaluate clustering performance.

What sets our work apart is the focused application and adaptation of these traditional clustering techniques to a high-dimensional dataset like the Olivetti faces. Our project diverged from the conventional path by challenging the effectiveness of the Silhouette score in high-dimensional spaces and introducing a novel metric - the Penalized V-Measure. This measure was specifically designed to address the over-segmentation issue prevalent in complex datasets when using standard evaluation metrics.

Other potential methods that could be applied to this problem include deep learning-based clustering techniques, such as autoencoders for dimensionality reduction followed by clustering. While these methods are promising, especially in their ability to handle high-dimensional data, they were not used in our project due to their complexity, the requirement for larger computational resources, and the need for a more profound understanding of neural network architectures. Additionally, these methods often require a larger dataset to perform effectively, which contrasts with the relatively small size of the Olivetti faces dataset.

Our approach, therefore, was more aligned with traditional clustering methods, adapted with a novel evaluation metric to suit the specific challenges posed by the Olivetti dataset. This adaptation not only allowed for a more nuanced understanding of the dataset's clustering structure but also provided a means to critically assess and improve upon the standard practices in clustering high-dimensional data.

Empirical results

- Data Acquisition, Preprocessing and Exploration

In the initial phase of the project, we acquired the Olivetti faces dataset, which includes 400 images of 40 distinct subjects. The dataset was reshaped into a 3D array (400x64x64) to facilitate visualization. We then created a custom function to display the first image of each unique individual, highlighting the dataset's diversity in facial features and expressions. This visualization was crucial for understanding the dataset's structure and composition before proceeding with further analysis.

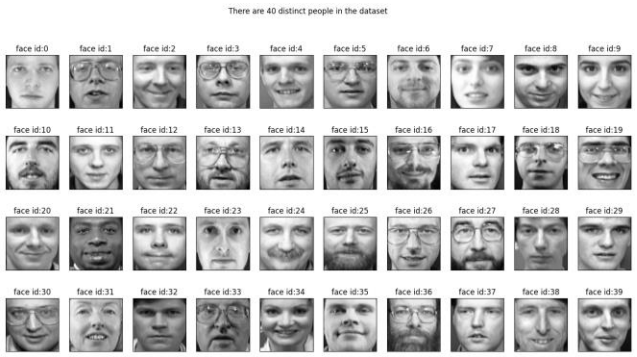

Figure 1: 40 Distinct People

- Dimensionality Reduction with PCA

We applied PCA to the Olivetti faces dataset, configuring it to retain 99% of the variance. This process reduced the data to 260 principal components. A plot of the cumulative explained variance against these components illustrated the plot effective balance between dimensionality reduction and information preservation.

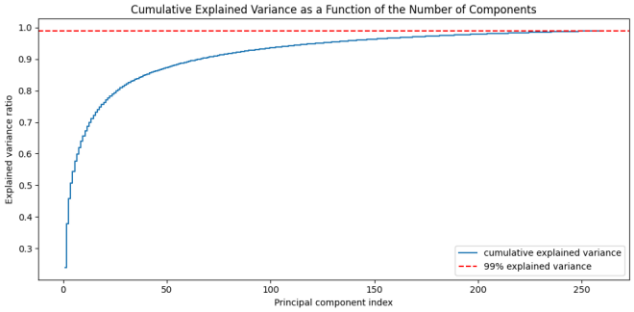

Figure 2: PCA

- K-Means

K-Means was applied to the PCA-reduced data from the Olivetti faces dataset, exploring a range of cluster numbers from 5 to 200. Each potential cluster count was tested, with the models initialized with multiple centroid seeds for robustness. We evaluated the clustering effectiveness using the silhouette score for each model, identifying the optimal number of clusters based on the highest score. Inertia was also considered for model selection, but the absence of a distinct 'elbow' in its plot led us to rely on the silhouette score for finalizing the best model.

Then we utilized the best K-Means model, identified through silhouette score, to examine the clustering results of the Olivetti faces dataset. A custom function was implemented to visually display the faces in each cluster. This visual assessment was key in evaluating the effectiveness of the clustering, allowing us to observe the grouping characteristics and assess the coherence of the clusters formed

by the best model.

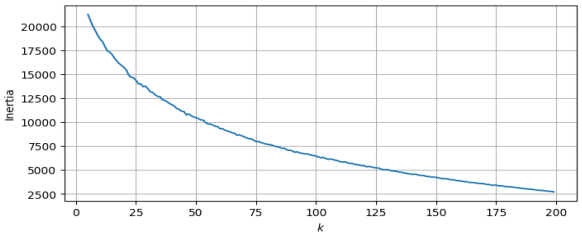

Figure 3: K-Means Clustering Inertias

To evaluate the effectiveness of our clustering, we utilized silhouette scores, a metric that assesses how similar an object is to its own cluster compared to other clusters. The best clustering performance, as indicated by the highest silhouette score, was achieved with 140 clusters. However, this optimal number appeared counterintuitive, as it significantly exceeded the expected 40 distinct individuals in the dataset.

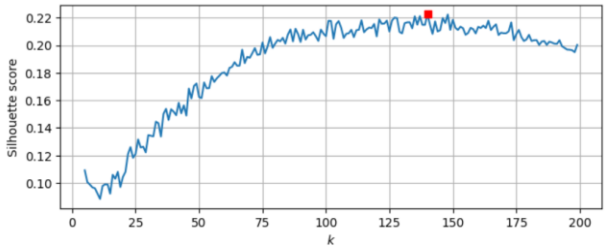

Figure 4: K-Means Clustering with Silhouette Score

Additionally, we examined the inertia of the K-Means clusters, which reflects the sum of squared distances of samples to their nearest cluster center. While lower inertia values are generally desirable, indicating tighter clusters, our analysis did not reveal a clear elbow in the inertia plot. This absence of a distinct turning point in the inertia graph further complicated the determination of an ideal number of clusters.

Figure 5: K-Means Clustering Result $k = 140$

• Agglomerative Clustering with Silhouette Scores

Following the K-Means analysis, we explored Agglomerative Clustering, a method well-suited for hierarchical data structures. This algorithm was applied under various configurations, adjusting the linkage criteria (ward, complete, average, single) and affinity metrics (Euclidean, Manhattan, cosine). Similar to our approach with K-Means, we computed silhouette scores for each cluster count to measure the clustering quality.

The results from these configurations presented a nuanced picture. The 'Euclidean-ward' linkage emerged as the most effective, achieving the best silhouette score at 132 clusters. This score, albeit slightly higher than that of K-Means, still reflected moderate clustering quality. The alternative configurations, notably those with *'Manhattan-single'* and *'cosine-single'* linkages, registered significantly lower scores, indicating less distinct clustering.

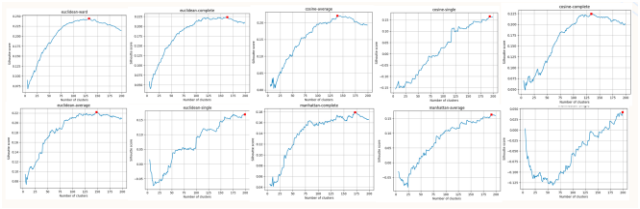

Figure 6: Agglomerative Clustering with Different Configurations

Table 1: Results with Different Configurations

Configuration	Best K	Best Score
Euclidean-ward	132	0.244383
Euclidean-complete	163	0.223546
Euclidean-average	148	0.222101
Euclidean-single	199	0.168866
Manhattan-complete	172	0.179192
Manhattan-average	191	0.163170
Manhattan-single	199	0.042458
cosine-complete	134	0.224513
cosine-average	139	0.220701
cosine-single	193	0.165933

• DBSCAN Clustering

In our pursuit to explore different clustering methodologies, we employed the Density-Based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN is particularly known for its effectiveness in identifying outliers and handling clusters of arbitrary shapes. However, its performance is highly sensitive to the choice of *eps* (the maximum distance between two samples for one to be considered as in the neighborhood of the other) and *min_samples* parameters.

Our experiments with DBSCAN revealed a significant challenge: the majority of data points were identified as noise. This outcome is attributable to the high-dimensional space of the Olivetti faces dataset, where data points are sparsely distributed. As we varied *eps* from 1.0 to 10.0, the algorithm mostly labeled data points as noise, with the number of clusters and noise points fluctuating significantly across different *eps* values. The algorithm was unable to form meaningful clusters, as indicated by the high number of noise points even at larger *eps* values.

• Gaussian Mixture Model (GMM)

We then explored clustering using Gaussian Mixture Models (GMM). GMM is a probabilistic model that assumes the data points are generated from a mixture of several Gaussian distributions with unknown parameters. Unlike K-means, GMM can infer clusters with different shapes and densities, making it a versatile choice for complex datasets.

Using silhouette scores to assess the clustering quality, we applied GMM across the same range of cluster numbers as K-means and Agglomerative Clustering. The best silhouette score was observed at 138 clusters, with a score of approximately 0.218.

Like previous methods, the optimal number of clusters suggested by the silhouette score was significantly higher than the expected 40 distinct individuals, and the score itself indicated moderate clustering quality.

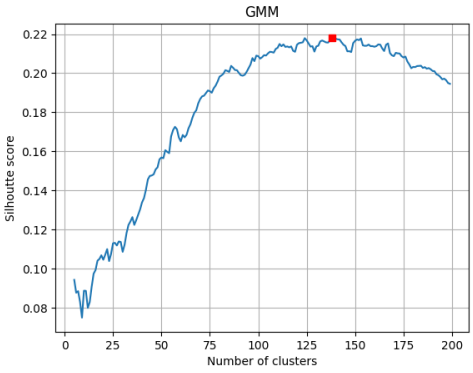

Figure 7: GMM with Silhouette Score

• Summary of Clustering Algorithms Applied to the Olivetti faces dataset with Silhouette Score

In our comprehensive analysis of clustering algorithms applied to the Olivetti faces dataset, we observed distinct behaviors and outcomes across the methods. K-Means and Agglomerative Clustering, while straightforward in their application, tended to over-segment the dataset, as evidenced by the high optimal cluster counts suggested by silhouette scores. DBSCAN, on the other hand, struggled with the high-dimensional nature of the dataset, resulting in a majority of data points being labeled as noise. Gaussian Mixture Models (GMM) showed a somewhat better ability to form clusters but still suggested a higher than expected number of clusters. None of the methods, with the metrics used, could effectively balance between capturing the true number of distinct individuals and maintaining high clustering quality, highlighting the challenges of clustering in complex, high-dimensional spaces like facial datasets.

• Refined Clustering Analysis

In our continued exploration of the Olivetti faces dataset, we employed more sophisticated evaluation metrics such as Homogeneity, Penalized Homogeneity, and Completeness, aiming to overcome the limitations observed with silhouette scores. This shift in approach provided us with a deeper and more nuanced understanding of the clustering dynamics within this complex dataset.

Our analysis using Homogeneity, which assesses whether each cluster is composed of members from a single class, revealed a highly homogeneous clustering at $k = 199$. However, this result, close to a perfect homogeneity score, suggested an overfitting scenario where clusters were predominantly representing individual or very few faces. This outcome, although statistically sound, was practically impractical, as it essentially mirrored the number of images in the dataset.

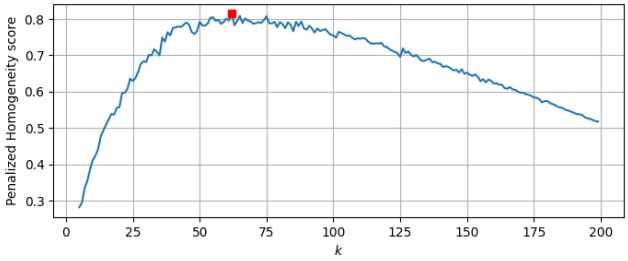

Figure 8: K-Means with Homogeneity

To counter the over-segmentation tendency observed in our clustering analysis, we introduced a novel metric known as Penalized Homogeneity. This approach effectively combines the standard homogeneity measure with a penalty term for cluster counts exceeding a predetermined threshold. Specifically, the penalty was calculated using the formula:

$$penalty = (k - max\_clusters) * penalty\_weight$$

where k is the number of clusters, $max\_clusters$ is the predefined threshold, and $penalty\_weight$ is a factor that determines the severity of the penalty for exceeding the threshold. This strategic adjustment aimed to strike a balance between clustering quality and the number of clusters, thus preventing the creation of an excessive number of clusters.

Applying Penalized Homogeneity to our dataset, we obtained an optimal cluster count at $k = 62$, achieving a score of approximately 0.814. This result was significant in its implications. It not only indicated a statistically robust clustering arrangement but also ensured practical relevance by avoiding the overfitting issue typically associated with pure homogeneity. The optimal cluster count, markedly lower than the initial estimates, pointed towards a more realistic

and meaningful clustering solution that aligns more closely with the intrinsic structure of the Olivetti faces dataset.

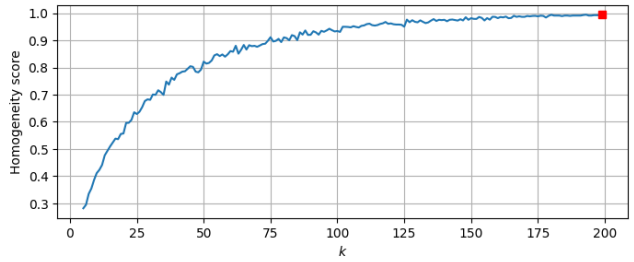

Figure 9: K-Means with Penalized Homogeneity

Further reinforcing our findings, the Completeness score, which evaluates if all members of a given class are assigned to the same cluster, also indicated $k = 62$ as the optimal clustering solution, achieving a score close to 0.808. The consistency observed between Penalized Homogeneity and Completeness in determining the optimal number of clusters lent credibility to our revised approach.

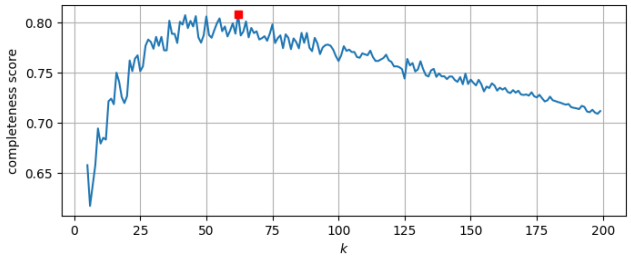

Figure 10: K-Means with Completeness

Building upon our previous clustering analysis, we next explored the V-Measure, a metric that combines the elements of both homogeneity and completeness to provide a balanced view of clustering quality. The standard V-Measure is particularly useful as it does not disproportionately favor larger cluster numbers, unlike pure homogeneity. In our application of V-Measure to the K-Means clustering results, we found the highest score at $k = 62$, with a V-Measure score of approximately 0.857. This result indicated a fair balance between homogeneity and completeness in the clustering outcome.

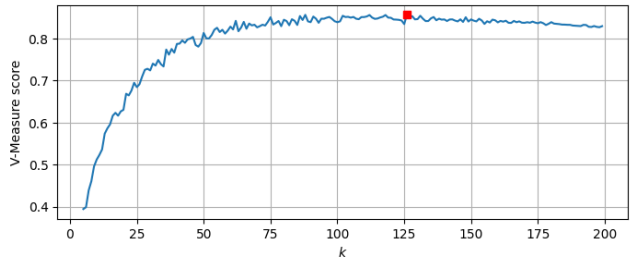

Figure 11: K-Means with V-Measure

However, recognizing the potential for overfitting with standard homogeneity, as observed in our previous analyses, we opted to incorporate our Penalized Homogeneity measure into the V-Measure calculation. This customized approach intended to refine the V-Measure by penalizing cluster counts exceeding a certain threshold, thus ensuring a more realistic clustering solution. The penalized V-Measure was computed by adjusting the homogeneity component using the formula:

$$penalty = (k - max\_clusters) * penalty\_weight$$

, and then combining it with the standard completeness measure.

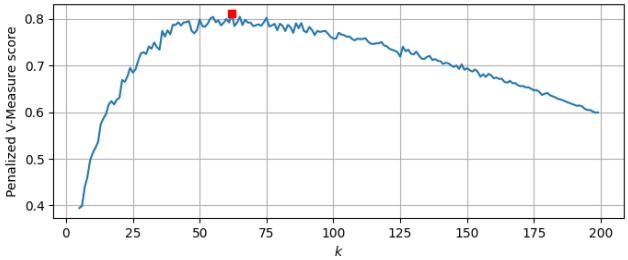

Figure 12: K-Means with Penalized V-Measure

The application of this Customized V-Measure to our clustering results yielded an optimal cluster count at $k = 62$, with a score of approximately 0.811. This outcome was not only in line with our previous findings using Penalized Homogeneity and Completeness but also validated our approach of penalizing excessive clustering. The penalized V-Measure effectively captured the essence of meaningful clustering by ensuring that the clusters were neither too broad nor overly granular.

• Evaluating Clustering Algorithms with Customized Penalized V-Measure

To validate the effectiveness of our customized penalized V-Measure, we applied it to Gaussian Mixture Models (GMM) and the highest-scoring configuration of Agglomerative Clustering (Euclidean-ward). This was an essential step in verifying whether our new metric would provide consistent and meaningful clustering results across different algorithms.

In the case of GMM, we recalculated the clustering scores across a range of clusters from 5 to 200 using our penalized V-Measure. This metric, which adjusts homogeneity to account for over-clustering, revealed that the best clustering performance for GMM occurred at $k = 55$, with a score of approximately 0.789. This outcome was significant as it demonstrated a reasonably good clustering quality, suggesting that GMM, when evaluated with this new metric, could produce meaningful clusters without over-segmentation.

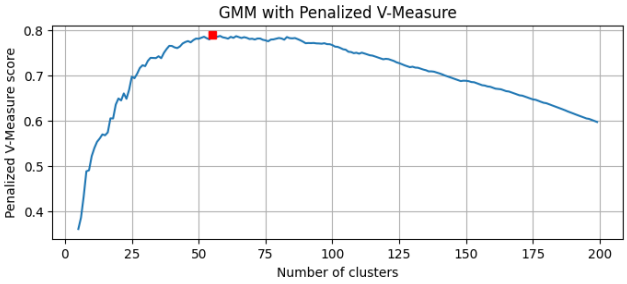

Figure 13: GMM with Penalized V-Measure

Similarly, we reassessed Agglomerative Clustering using the Euclidean-ward linkage and our penalized V-Measure. This method had previously shown promising results, and with the application of the penalized V-Measure, it reached its optimal performance at $k = 59$, achieving a score of around 0.835. This score was not only higher than that obtained with GMM but also indicative of a robust clustering performance, balancing between homogeneity and completeness while avoiding excessive cluster counts.

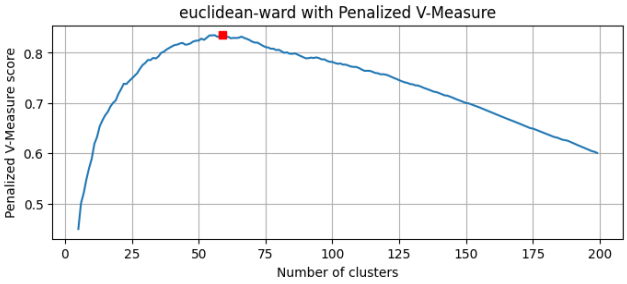

Figure 14: Agglomerative Clustering (Euclidean-ward) with Penalized V-Measure

These findings confirmed the utility of the penalized V-Measure as a more effective and realistic metric for evaluating clustering quality in complex datasets like the Olivetti faces. Unlike standard metrics, which often led to over-segmentation, the penalized V-Measure provided a more balanced perspective, guiding us towards clustering solutions that were statistically sound and practically meaningful. The consistency in the optimal cluster counts across different algorithms further underscored the robustness and adaptability of our approach.

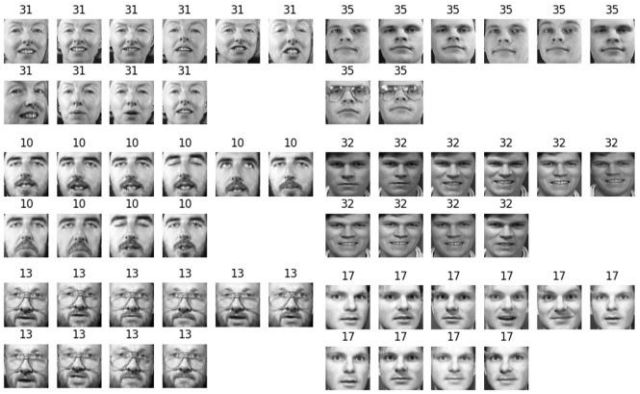

Figure 15: Agglomerative Clustering (Euclidean-ward) Result when $k = 59$

Conclusions/future directions

The conclusion of our project on clustering human faces using the Olivetti faces dataset highlighted significant insights, particularly in dealing with high-dimensional data. We learned that conventional metrics like the silhouette score can be misleading in such contexts, sometimes leading to the selection of suboptimal models. This experience emphasized the importance of choosing evaluation measures that are tailored to the specific structure and complexities of the data.

With more time, we would enhance the project by integrating advanced clustering, deep learning techniques like autoencoders for dimensionality reduction, and a semi-supervised approach. Autoencoders could reveal complex data patterns beyond PCA's scope, while partial data labeling would refine clustering accuracy and context. Testing the models on a related dataset would also assess their robustness and generalization. This comprehensive strategy aims to improve clustering results and establish a more versatile framework for similar analyses.

For future DS 5230 students taking on similar projects, two main pieces of advice are offered. Starting the project early is recommended. Understanding the mechanics of the algorithm helps a lot in doing these projects.

GitHub Link

https://github.com/cclgdxdw/Olivetti-Faces