CPS803 Assignment 4: Dry Bean Clustering Analysis

Alden Shin-Culhane 501036054

Background

For this assignment, I selected the Dry Bean dataset from the UCI Machine Learning Repository [1]. The dataset consists of images of 13,611 grains of seven registered types of dry beans, such as Seker, Barbunya, and Bombay. Each bean's physical characteristics were extracted into 16 numerical features. The goal of this analysis was to apply clustering techniques to group these beans based on their shared physical properties, independent of their labeled class.

This dataset is particularly interesting because it highlights the practical applications of machine learning in agriculture. Clustering similar beans based on their features could help improve sorting and packaging processes for farmers or suppliers, ensuring uniform quality in production. The dataset was accessed programmatically using the ucimlrepo Python package, and pre-processing steps were undertaken to prepare the data for clustering.

Methods

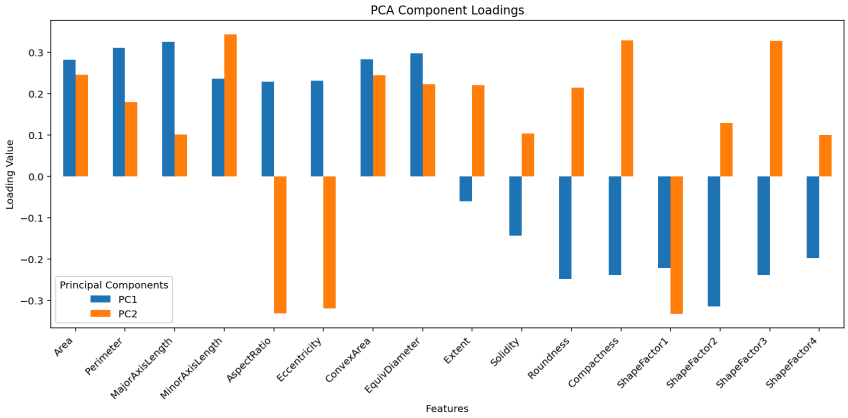
I began the analysis by pre-processing the dataset. This involved extracting the numeric features and excluding the labeled class column, ensuring that clustering was performed without the influence of predefined labels. To address differences in feature scales, I standardized the dataset using the StandardScaler from Scikit-Learn. This step ensured that all features contributed equally to the clustering process, preventing features with larger ranges, such as Area, from dominating the results. The raw data, along with these numeric features, was saved in a file named dry_bean_original.csv.

To facilitate visualization and reduce dimensionality, I applied Principal Component Analysis (PCA). This transformation reduced the dataset to two principal components while retaining most of the variance. A PCA Loadings Plot was created to interpret the contributions of individual features to each principal component. Principal Component 1 (PC1) was influenced heavily by size-related features such as Area, Perimeter, and Major Axis Length, while Principal Component 2 (PC2) captured shape-related features like Aspect Ratio and Eccentricity. This step provided insights into the underlying structure of the data and its key characteristics.

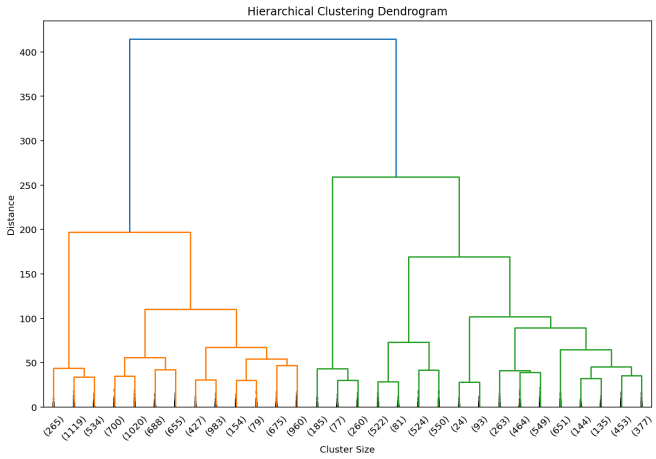
The clustering was performed using hierarchical clustering with Ward’s linkage, which minimizes variance within clusters. A dendrogram was generated to visualize the hierarchical relationships between data points and determine the optimal number of clusters. Various numbers of clusters were tested, based on the dendrogram and visualizations I determined that four clusters would provide the best balance between compactness and separation. This decision was further supported by the PCA Scatter Plot, which showed clear separability in two-dimensional space.

To further analyze the clusters, I computed summary statistics, including the mean, median, and mode for all features within each cluster. These results were saved in three separate CSV files: hierarchical_cluster_mean_stats.csv, hierarchical_cluster_median_stats.csv, and hierarchical_cluster_mode_stats.csv. These statistics helped to identify distinguishing features across clusters and provided numerical support for the clustering outcomes.

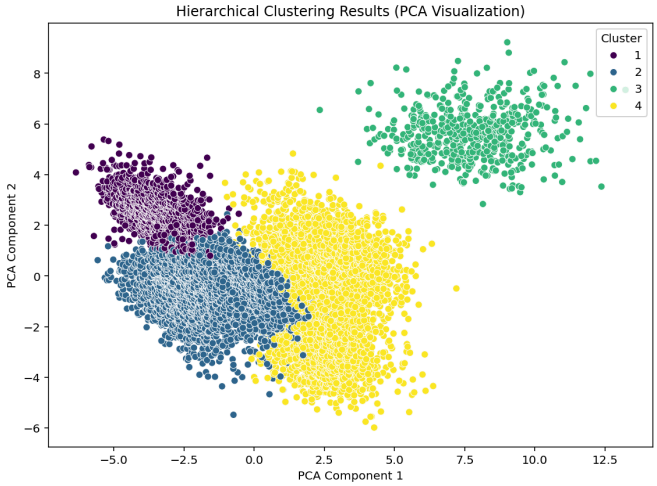
Several visualizations were created to validate and interpret the results. The PCA Loadings Plot illustrated the contribution of each feature to the two principal components, helping to explain the structure of the PCA-transformed data.



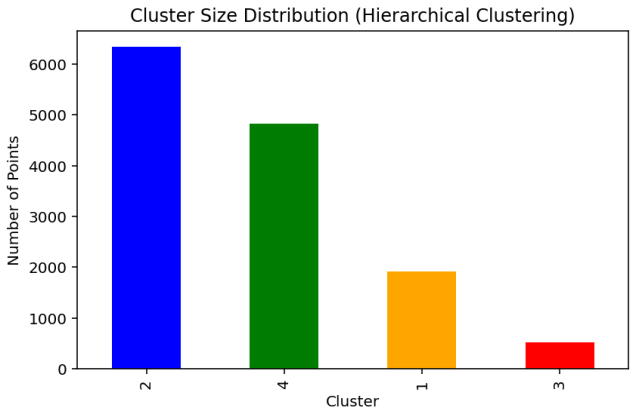
The dendrogram confirmed the number of clusters by revealing a significant increase in linkage distance below four clusters.



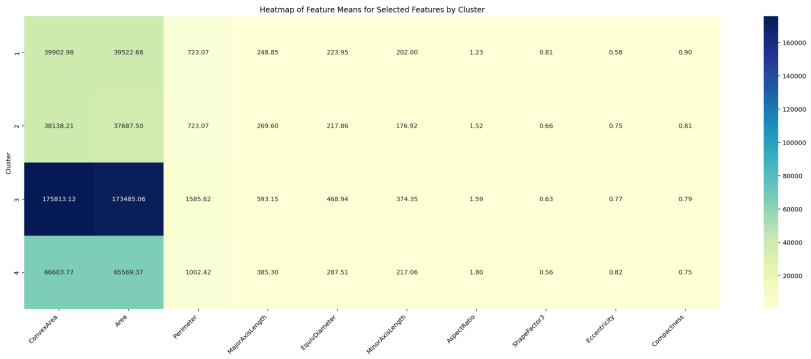
A PCA Scatter Plot was used to visualize the separability of clusters in two-dimensional space, showing distinct groupings with minimal overlap.



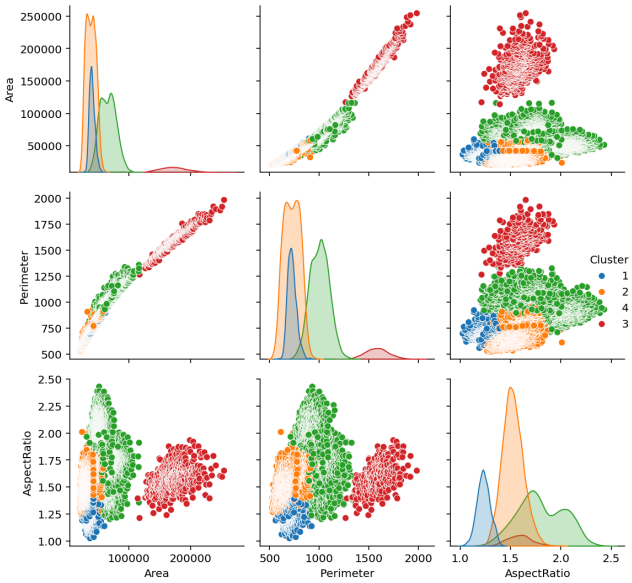
The Cluster Size Distribution Plot revealed the relative sizes of the clusters, where the largest cluster contained approximately 6,600 beans and the smallest contained about 1,200. This distribution aligns with expectations, as certain bean types are more prevalent than others.



A Heatmap of Feature Means provided a visual representation of how the top 10 features with the highest variance differed across clusters.



Finally, a Pairplot was generated to explore the relationships between key features, such as Area, Perimeter, and Aspect Ratio, and how they varied across clusters.



These manipulations, including standardization, PCA, and hierarchical clustering, were critical for achieving meaningful clustering results. Standardization ensured equal feature contributions, while PCA reduced noise and highlighted the dataset's most important dimensions. Hierarchical clustering with Ward's linkage effectively grouped the beans based on their physical attributes, and the visualizations provided robust evidence of the clustering's validity.

Results

After setting the number of clusters to four, the results demonstrated moderately well-defined and distinct groupings. The dendrogram provided a clear visual representation of hierarchical relationships in the data, and it showed a significant jump in linkage distance when reducing the number of clusters below four. This observation reinforced the choice of four clusters as optimal. The PCA Loadings Plot clarified the contributions of each feature to the principal components, with size-related features dominating PC1 and shape-related features influencing PC2. This provided insight into which characteristics were most influential in the clustering process.

The PCA Scatter Plot illustrated how well-separated the clusters were, providing strong visual evidence of distinct groupings. The clear separation observed in the plot suggests that the clusters have meaningful differences based on the principal components, emphasizing the role of size- and shape-related features in distinguishing the clusters. The Cluster Size Distribution Plot showed that the largest cluster contained over 6,600 points while the smallest contained approximately 1,200. This imbalance reflects the natural distribution of bean types in the dataset.

The size distribution aligns with expectations, as it suggests that some bean types are more common than others in the dataset, which is typical in agricultural datasets.

The Heatmap of Feature Means highlighted key differences in feature values across clusters, such as Convex Area and Major Axis Length. This helped to understand the unique characteristics of each cluster and further validate that the clusters represent distinct groupings of beans. For example, features like Convex Area and Major Axis Length had notably different means across clusters, suggesting that these features were significant in differentiating the beans.

Lastly, the Pairplot provided a more detailed look at the relationships between selected features across clusters. It showcased how features like Area, Perimeter, and Aspect Ratio varied systematically among the four groups, offering deeper insights into the relationships between these physical characteristics. The CSV files produced during this analysis (hierarchical_clusters.csv, hierarchical_cluster_mean_stats.csv, hierarchical_cluster_median_stats.csv, and hierarchical_cluster_mode_stats.csv) serve as interpretable records of the clustering process. They summarize the characteristics of each cluster and provide further insights into the nature of the groups formed.

Conclusions

The results of this analysis demonstrate that hierarchical clustering is effective for grouping dry beans based on their physical characteristics. The selection of four clusters was supported by visual evidence from the dendrogram and the PCA scatter plot. These clusters reflect meaningful groupings of the beans, providing insights into their physical attributes that could potentially be applied in agricultural quality control processes, such as sorting and packaging.

The PCA transformation facilitated effective visualization and interpretation of the data, highlighting features that contributed most to cluster separability, such as Area, Perimeter, and Aspect Ratio. This dimensionality reduction was critical in understanding the underlying structure of the dataset and effectively visualizing the results. The combination of hierarchical clustering and PCA provided robust evidence of distinct bean groupings, and the insights derived from this analysis could have practical applications in optimizing agricultural processes, ensuring uniformity and quality in production.

References

Dry Bean [Dataset]. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C50S4B.