CPS803 Assignment 4 Report- Wine Dataset (UCI)

(1) Background

The Wine dataset, sourced from the UCI Machine Learning Repository, is a well-known dataset often used for classification and clustering tasks in machine learning. It contains 178 samples of wines derived from three different cultivars (wine-producing grape varieties). Each sample is characterized by 13 chemical attributes, including alcohol content, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavonoids, nonflavonoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines, and proline [2].

The dataset is interesting because it provides a real-world example of clustering, where groups of wine samples can be divided based on chemical properties. These clusters can reflect potential patterns like flavor profiles, production methods, or quality categorizations. For example, clustering could help a winemaker group wines into premium and standard qualities or help a retailer organize wines into intuitive categories for marketing purposes.

Clustering is an unsupervised learning method used to group data points based on their similarities. Unlike classification, clustering doesn't rely on labeled outputs. Instead, it aims to discover hidden structures within the data. In this report, two clustering techniques were used: **K-Means Clustering**, which partitions the data into clusters based on centroids, and **Hierarchical Clustering**, which organizes data into a nested structure based on proximity. These methods were applied to understand the grouping of wines and to evaluate how well the clusters align with natural patterns in the dataset.

(2) Methods

Data Preprocessing

The Wine dataset was loaded as a CSV file with predefined column headers corresponding to its 13 features and target class. Preprocessing was necessary to ensure the data was clean and ready for clustering:

1. **Feature Selection:** The Class attribute, which denotes the cultivar each wine belongs to, was excluded since clustering is an unsupervised task that does not use class labels. This ensured the algorithm relied only on the numerical features to find patterns.
2. **Standardization:** Since the features have varying ranges (e.g., alcohol content ranges from 11 to 15, while OD280/OD315 ranges from 1 to 4) [2], they were standardized to ensure equal contribution. Standardization transformed each feature to have a mean of 0 and a standard deviation of 1. Without this, clustering algorithms like K-Means would bias toward features with larger magnitudes.

Clustering Algorithms

1. **K-Means Clustering:**
    - **Algorithm Overview:** K-Means is an iterative algorithm that partitions data into k clusters. The algorithm randomly initializes k centroids and assigns each data point to the cluster with the nearest centroid. The centroids are updated iteratively by taking the mean of the points in each cluster, and this process continues until the clusters stabilize [1][4].
    - **Parameter Selection:**
        - The number of clusters (k) was set to 3, reflecting the number of cultivars.
        - Random state was initialized for centroids, to ensure reproduction occurs.
    - **Function used:**
        - The algorithm minimizes the Sum of Squared Errors (SSE) [4]:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

where x is a data point in cluster C_i and m_i is the centroid (mean) for cluster C_i. SSE improves in each iteration of K-means until it reaches a local or global minimal

2. **Hierarchical Clustering:**
    - **Algorithm:** Hierarchical clustering builds a tree-like structure called a dendrogram (Figure 1.0) [4]. An agglomerative approach was used, where at each step, the two closest clusters are merged until a single cluster remains. This approach provides flexibility, as the desired number of clusters can be obtained by cutting the dendrogram at the correct level.

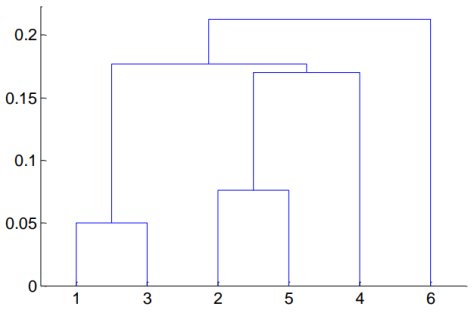

Figure 1.0: Example of a dendrogram

- **Distance Metric:** Euclidean distance was used to measure the similarity between points, ensuring consistency with K-Means [4].

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

where p and q are two points in Euclidean n-space, q_i and p_i are Euclidean vectors, starting from the origin of the space (initial point), and n is the n-space

Evaluation Strategy [4]

1. Sum of Squared Errors (SSE) ($|C_i|$ *is the size of cluster i)*:
   - Cluster Cohesion: Measures how closely related are objects in a cluster using SSE

     $$SSE = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

   - A lower SSE indicates that points within a cluster are closer to their centroid, meaning higher cohesion.
2. Sum of Squared Between-Cluster Distances (SSB) ($|C_i|$ *is the size of cluster i):*
   - Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters using SSB

     $$SSB = \sum_{i} |C_i|(m - m_i)^2$$

   - Higher SSB values indicate greater separation between clusters, which is a desirable property for distinct grouping.

The quality of the clusters was then assessed using the **Silhouette Score** [3]. Silhouette Score combines cohesion and separation into one metric. The difference of the three can be seen in the chart below:

Metric	What it measures	Goal	Used for
SSE	Compactness of clusters (intra-cluster distances)	Minimize for better clusters	Cluster compactness evaluation
SSB	Separation between clusters (inter-cluster distances)	Maximize for better clusters	Cluster separation evaluation
Silhouette Score	Balance between cohesion and separation for each point	Closer to 1 is better	Overall cluster validity

- Silhouette Score = (b - a) / max(a, b),

  where a is the average distance between each point within a cluster (mean intra-cluster distance), and b is the average distance between all clusters (mean nearest-cluster distance)

- A higher score (closer to 1) indicates well-separated clusters, while a lower score (near 0) suggests overlapping clusters [6].

(3) Results

Clustering Quality

1. **K-Means Clustering:**
    - The Silhouette Score was **0.285**, indicating moderate clustering quality. This value reflects some degree of overlap between the clusters, but the separation is significant enough to identify distinct groups.
    - Analysis of cluster centers revealed distinct profiles for each cluster, with different values for key chemical attributes like flavonoids, magnesium, and total phenols. This suggests that K-Means successfully grouped wines with similar chemical compositions.
2. **Hierarchical Clustering:**
    - The Silhouette Score was **0.277**, slightly lower than K-Means. While the dendrogram showed distinct groupings, the proximity of some clusters led to overlaps, reducing the overall separation quality.
    - Hierarchical clustering produced results similar to K-Means, supporting the strength of the patterns in the dataset.

Cluster Visualization

To visualize the clusters in a 2D space, the 13-dimensional data was projected onto two principal components using PCA. Principal Component Analysis is a tool used to simplify data by reducing its number of dimensions while still keeping the most important information [5].

- **K-Means Clustering:**
    - Points in the scatter plot were color-coded to indicate cluster membership.
    - While the clusters were generally well-separated, some overlap was observed in the boundaries, particularly between Clusters 2 and 3.
- **Hierarchical Clustering:**
    - The dendrogram indicated nested clusters, with some clusters merging late in the process. The scatter plot confirmed the moderate separation of clusters.

Insights from K-Means Cluster Centers

The cluster centers provide an average representation of the chemical properties for each cluster. Key insights:

- Cluster 1 had higher levels of flavonoids and total phenols, indicating wines with potentially richer flavor profiles.
- Cluster 2 had higher levels of proanthocyanins and lower magnesium content, possibly reflecting differences in grape type or production techniques.
- Cluster 3 demonstrated balanced chemical properties, suggesting a middle ground between the other two clusters.

(4) Conclusions

The clustering analysis of the Wine dataset provided insights into the natural groupings of wines based on chemical properties. K-Means clustering demonstrated slightly better performance than Hierarchical Clustering, as reflected in the higher Silhouette Score (0.285 vs. 0.277). Both methods identified three distinct clusters that broadly align with the three known wine cultivars. The results suggest that clustering techniques have practical applications in the wine industry, such as segmenting wines for quality control, flavor profiling, or market differentiation. By analyzing chemical attributes, producers and retailers can identify distinct groups and tailor their strategies accordingly. However, the moderate Silhouette Scores highlight limitations in cluster separation. This could be caused from overlapping chemical profiles between cultivars or the complexity of the dataset, which could possibly be improved using more advanced techniques. In conclusion, the current analysis provides valuable insights and lays a strong foundation for further exploration, with the potential to uncover even more meaningful patterns.

References

[1] *2.3. clustering*. scikit. (n.d.-a). https://scikit-learn.org/1.5/modules/clustering.html

[2] Aeberhard, S. & Forina, M. (1992). Wine [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5PC7J.

[3] Gültekin, H. (2023, September 7). *What is silhouette score?*. Medium. https://medium.com/@hazallgultekin/what-is-silhouette-score-f428fb39bf9a

[4] Lugez, E. (2024, November). *Machine learning: Unsupervised Learning Clustering*. *Unit6*.

[5] *Principal Component Analysis (PCA) explained*. Built In. (n.d.). https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[6] *Silhouette_score*. scikit. (n.d.-b). https://scikit-learn.org/dev/modules/generated/sklearn.metrics.silhouette_score.html