**PCA, KMeans, and Agglomerative Clustering on Wine Dataset**

Mahamadou Mahi Diallo

Northcentral University

TIM-8555

**Dr. Mohammad Yavarimanesh**

**November 2, 2024**

**PCA, KMeans, and Agglomerative Clustering on Wine Dataset**

**Introduction**

Using a dataset containing chemical properties of various wines from different cultivars grown in a single region in Italy, this project aims to uncover inherent groupings in the data through unsupervised learning. Since the data lacks specific identifiers, we focus on using the chemical attributes to discover patterns and similarities across the wines. The analysis begins with exploratory data analysis and data preprocessing to ensure that the dataset is well-prepared for clustering.

Principal Component Analysis (PCA) is then applied to reduce the dataset’s dimensionality while preserving significant variance, making it suitable for clustering. Both Agglomerative Clustering and K-Means Clustering techniques are employed to group similar wines together. The resulting clusters provide a deeper understanding of the relationships among the wines based on their chemical profiles.

**Exploratory Data Analysis**

The exploratory data analysis (EDA) phase provided an in-depth look at the distribution and relationships of the chemical attributes in the wine dataset. Initial inspection of the dataset showed that the features, such as alcohol content, malic acid concentration, and phenolic composition, varied widely across samples, reflecting the diversity of the wine types. Univariate analysis through histograms and box plots revealed that some features were skewed, while others had a normal distribution. These insights informed the need for standardization before applying PCA.

Bivariate analysis through scatter plots and correlation heatmaps showed significant correlations among certain chemical properties (VanderPlas, J. (2016))., such as between alcohol content and total phenols, which indicates potential redundancy that PCA could address by reducing dimensionality (Wold, S., Esbensen, K., & Geladi). Overall, EDA helped to identify underlying patterns, correlations, and potential outliers, setting a solid foundation for dimensionality reduction and clustering.

**PCA**

Principal Component Analysis (PCA) was applied to the dataset to reduce its dimensionality while retaining the maximum possible variance. This technique is particularly useful for high-dimensional data, as it transforms the original features into a set of new, uncorrelated variables called principal components. After scaling the features to ensure equal contribution, the dataset was analyzed for its explained variance to determine the optimal number of components to retain.

Based on the explained variance ratio and the cumulative variance graph, five principal components were chosen, capturing a significant portion of the dataset’s variability. By reducing the number of dimensions, PCA helps simplify the data for clustering and visualization without sacrificing much of the information. The transformed dataset, consisting of these five principal components, was then used for clustering analysis.

**KMeans Clustering**

K-Means Clustering was employed to group similar observations within the dataset based on the principal components obtained from PCA. This algorithm partitions data points into a specified number of clusters by minimizing the variance within each cluster (Grus, J. (2019)), effectively grouping similar samples together.

After experimenting with different values for the number of clusters, the optimal choice was determined using methods such as the Elbow Method, which visually identifies the point where adding more clusters results in diminishing returns in variance reduction. For this analysis, three clusters were chosen, capturing distinct groupings within the dataset. Each wine sample was then assigned a cluster label, allowing us to explore the characteristics of each group and identify patterns within the clusters. This approach provides a deeper understanding of the dataset’s structure, enabling us to distinguish between different types of wines based on their chemical properties.

**Agglomerative Clustering**

Agglomerative Clustering was utilized to group the wine samples based on the five principal components obtained from PCA. This hierarchical clustering method starts with each observation as an individual cluster and iteratively merges the closest pairs of clusters based on a defined distance metric.

By choosing an appropriate number of clusters, determined from previous analyses such as the Elbow Method, we set the algorithm to partition the dataset into three distinct clusters. The resulting cluster assignments were integrated into the PCA DataFrame, enabling further analysis of the clusters' characteristics. This approach allowed us to explore the inherent groupings within the data and provided insights into the relationships among the different types of wine based on their chemical properties, facilitating a more nuanced understanding of the dataset's structure.

**Results**

The clustering analysis revealed three distinct groups of wines, each characterized by unique chemical properties and potential flavor profiles.

**Cluster 0** consists of wines that are lower in alcohol, moderate in flavonoids, and low in proline, with high OD280 values and a wide range of hues. These wines are generally milder, exhibiting a lighter body and lower viscosity. In contrast, **Cluster 1** features wines with moderate alcohol levels, low flavonoids, and low proline content. These wines possess darker hues and higher color intensity, suggesting a bolder profile, yet they are less bitter due to lower flavonoid content. Lastly, **Cluster 2** includes wines with the highest alcohol content, most flavonoids, and elevated proline levels. This cluster is characterized by high OD280 values and darker hues, indicating a full-bodied, strong selection with a rich and potentially bitter flavor profile.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster** | **Alcohol** | **Flavonoids** | **Proline** | **OD280** | **Hue** | **Color Intensity** | **Description** |
| 0 | Lower | Moderate | Low | High | Wide range | Low | Milder wines, lighter body, lower viscosity |
| 1 | Moderate | Low | Low | Low | Lighter | High | Stronger wines, darker varietal, less bitter |
| 2 | Higher | Most | High | High | Darker | Moderate | Full-bodied wines, strong flavor, high bitterness |

**Conclusion**

This project utilized unsupervised learning techniques, specifically Principal Component Analysis (PCA) and clustering methods, to analyze the chemical properties of various wines. By applying PCA, we effectively reduced dimensionality while preserving significant variance, enabling clearer visualization and analysis of the data. The Elbow Method helped determine the optimal number of clusters for K-Means clustering, revealing distinct groupings among the wine samples. Additionally, Agglomerative Clustering reinforced these findings, providing insights into the relationships between the clusters. Overall, this analysis highlighted meaningful patterns in the dataset, demonstrating the effectiveness of combining PCA with clustering techniques to enhance our understanding of wine properties.

**References**

Wold, S., Esbensen, K., & Geladi, P. (1987). "Principal component analysis." *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), 37-52

VanderPlas, J. (2016). "Python Data Science Handbook: Essential Tools for Working with Data." O'Reilly Media.

Grus, J. (2019). *Data Science from Scratch: First Principles with Python*. O'Reilly Media