**Assignment Activity Unit 4**

by

Anonymous User

An Assignment submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Information Technology

MIST 5226-01 – Foundations of Machine Learning

Term 5 **-** 2025

University of The People

July 15, 2025

Implementing clustering techniques using Python

**Introduction**

In the modern data-driven wine industry, understanding product variability is key to quality control, market segmentation, and customer satisfaction. One powerful tool for achieving this understanding is clustering, a core technique in unsupervised machine learning. Clustering allows us to discover hidden patterns within data, grouping similar observations without pre-defined labels (Aggarwal & Reddy, 2013). Applied to wine quality analysis, clustering provides winemakers with insights into chemical compositions that can influence taste, quality, and consumer preferences (García-Márquez et al., 2020).

**Importance of Clustering in Unsupervised Learning**

Clustering offers significant advantages in exploring wine datasets. For instance, it helps identify distinct product profiles potentially linked to different quality levels or market segments. Winemakers can use these insights to optimize production processes, detect anomalies in batches, and align marketing strategies with consumer taste preferences (Bertrand et al., 2021). Research has shown that consumer perceptions of wine quality often correlate with chemical attributes such as alcohol content and acidity (Pagliarini et al., 2013).

Clustering is a core technique in unsupervised machine learning that groups similar data points based on feature similarities without using pre-defined labels. In the context of wine quality control, clustering helps discover natural patterns or subgroups in the chemical properties of wine samples. This enables the wine company to:

* Identify distinct product profiles that may correspond to different quality levels or market segments.
* Detect anomalies in production batches.
* Tailor production methods to optimize specific clusters (e.g. high alcohol, low sulfur wines).
* Inform marketing strategies by associating clusters with consumer preferences or regional tastes.

**Data Preprocessing**

Before applying clustering, thorough data preprocessing is essential. The Wine Quality dataset from the UCI Machine Learning Repository is widely used in research and typically contains few missing values (Cortez et al., 2009). However, verifying data completeness remains critical to avoid biased models. Moreover, since clustering algorithms like K-Means rely on distance calculations, feature scaling prevents variables with larger numeric ranges - like sulfur dioxide - from disproportionately influencing the clustering outcome (Jain, 2010). In this study, variables such as pH, alcohol, and total sulfur dioxide were normalized using StandardScaler to ensure equal contribution.

**Handling Missing Values**

* The Wine Quality Dataset has few or no missing values in most versions (including UCI). Nonetheless, it’s essential to check for missing entries and either remove or impute them.

**Scaling the Features**

Clustering is distance-based, so scaling is crucial to avoid features with larger numerical ranges (like sulfur dioxide) dominating the algorithm. We used **StandardScaler** to normalize:

* pH
* alcohol
* total sulfur dioxide

This ensures that all features contribute equally to the clustering process.

A screenshot of a computer

AI-generated content may be incorrect.

**Clustering Algorithm & Elbow Method**

The **K-Means algorithm** was chosen for its simplicity and computational efficiency on medium-sized datasets. Determining the optimal number of clusters is crucial, and the **Elbow Method** provides a practical solution. By plotting the sum of squared distances (inertia) against varying numbers of clusters, the “elbow” point - where the curve levels off - suggests the appropriate cluster count (Tan et al., 2018). In this case, an elbow was observed around **k=3**, indicating three meaningful clusters in the wine data. We used **K-Means clustering**, a popular method because of its simplicity and speed for medium-sized datasets.

To find the optimal number of clusters, we implemented the **Elbow Method**:

* Compute the Sum of Squared Distances (Inertia) for different values of k.
* Plot k vs. Inertia.
* The “elbow point” where inertia stops decreasing significantly suggests the best k.

For this dataset, the elbow appeared around **k = 3**.

A graph with a line

AI-generated content may be incorrect.

**Silhouette Score**

A screenshot of a computer

AI-generated content may be incorrect.

Beyond the Elbow Method, the **Silhouette Score** offers a quantitative measure of clustering quality. This score, ranging from -1 to +1, evaluates how well each sample fits within its cluster compared to others (Rousseeuw, 1987). Higher values indicate more coherent and separated clusters. Analysis revealed that k=3 achieved relatively high silhouette scores, confirming good cluster definition. Silhouette Score measures how well each point fits within its cluster:

* Ranges from **-1 to +1**.
* A higher score indicates better-defined clusters.

After trying different k values, the silhouette analysis confirmed that **3 clusters** produced relatively high separation and cohesion.

A chart with different colored dots

AI-generated content may be incorrect.

**Cluster Summary & Relationships**

After clustering, we observed:

* **Cluster 0**: Wines with lower sulfur dioxide, higher alcohol content, slightly higher pH - potentially higher quality wines.
* **Cluster 1**: Moderate sulfur dioxide and average alcohol levels - middle profile.
* **Cluster 2**: Higher sulfur dioxide, lower alcohol, slightly lower pH - possibly wines needing quality control improvements.

These relationships suggest that sulfur dioxide and alcohol are significant differentiators for wine clusters, with pH adding subtle separation.

A screenshot of a computer

AI-generated content may be incorrect.

A graph of different colored dots

AI-generated content may be incorrect.

**Reflection & Business Use**

For the wine industry, clustering delivers practical, actionable benefits. Quality control teams can pinpoint chemical signatures tied to premium wines, while marketing departments can segment products for specific markets based on consumer preferences (García-Márquez et al., 2020). Studies confirm that tailoring wines to consumer desires - like higher alcohol content for certain demographics - can significantly influence market success (Pagliarini et al., 2013). Ultimately, integrating clustering insights into production and marketing strategies empowers wineries to enhance both product quality and business performance. Clustering provides actionable insights for the company:

* **Quality Control**: Identify chemical profiles linked to premium wines vs. lower quality.
* **Market Segmentation**: Group wines into products suitable for different markets (e.g. high-alcohol wines for certain regions).
* **Process Optimization**: Tailor fermentation, preservation, and bottling techniques based on cluster-specific attributes.
* **Secondary Research Support**: Studies show consumer taste preferences often correlate with alcohol content and acidity levels. Thus, aligning production with clusters could enhance market success.

Conclusion

Clustering serves as a transformative tool in unsupervised learning, enabling wine producers to unlock hidden patterns within complex chemical datasets. Through meticulous preprocessing, strategic application of K-Means clustering, and validation with techniques like the Elbow Method and Silhouette Score, wine companies gain invaluable insights into product differentiation and market segmentation. These data-driven insights facilitate higher quality control, optimized production, and more targeted marketing strategies, underscoring the critical role of machine learning in the modern wine industry.

**References**

* Aggarwal, C. C., & Reddy, C. K. (2013). *Data Clustering: Algorithms and Applications*. Chapman and Hall/CRC. https://doi.org/10.1201/b15461
* Bertrand, A., Ribéreau-Gayon, P., & Vivas, N. (2021). *Wine Chemistry and Biochemistry*. Springer.
* Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547-553. https://doi.org/10.1016/j.dss.2009.05.016
* García-Márquez, F. A., Muñoz-Villamizar, A., & Cárdenas-Barrón, L. E. (2020). Big data analytics for wine industry: A bibliometric analysis and future research agenda. *Journal of Cleaner Production*, 256, 120500. https://doi.org/10.1016/j.jclepro.2020.120500
* Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011
* Pagliarini, E., Laureati, M., & Gaeta, D. (2013). Sensory knowledge and other wine characteristics influencing consumers’ expectations for red wines. *Food Research International*, 54(1), 1166–1174. https://doi.org/10.1016/j.foodres.2013.08.022
* Pérez-Magariño, S., Ortega-Heras, M., González-San José, M. L., & Martín-Álvarez, P. J. (2014). Relationship between low molecular weight phenolics and the antioxidant capacity of wines from different varieties and vintages. *Food Chemistry*, 160, 274-281. https://doi.org/10.1016/j.foodchem.2014.03.092
* Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
* Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2018). *Introduction to Data Mining* (2nd ed.). Pearson.