Unsupervised Learning - Final Project

Isaac Benadiba 339694705, Yuval Ramot 208115840

April 21, 2025

Abstract

This project explores the application of unsupervised learning techniques to uncover latent structure in wine data based solely on chemical properties. Without relying on any labels or external annotations, we aim to identify natural groupings of wines, detect anomalies, and interpret the underlying geometry of the dataset.

We begin by standardizing the data and applying Principal Component Analysis (PCA) to reduce its dimensionality. Clustering is then performed using K-Means, DBSCAN, and Gaussian Mixture Models (GMM), each offering a different view of the data's organization: centroid-based, density-based, and probabilistic. We determine the optimal number of clusters using the elbow method and validate clustering performance through the silhouette score. A t-distributed Stochastic Neighbor Embedding (t-SNE) visualization is used to illustrate the separation between clusters in a nonlinear projection.

Overall, the analysis demonstrates that unsupervised learning can reveal meaningful wine archetypes and detect unusual samples without any prior labeling. The combination of visual, statistical, and algorithmic tools provides a robust pipeline for exploring hidden structure in complex datasets. The code is available at Git Repository.

1 Introduction

Unsupervised learning is a core area of machine learning that focuses on discovering hidden patterns, structures, or groupings within unlabeled data. Unlike supervised learning, which relies on labeled examples, unsupervised methods aim to find meaningful organization in data without any prior annotations. This makes them especially valuable in exploratory data analysis, where no ground truth is available.

In this project, we investigate whether unsupervised learning algorithms can reveal latent structure in a real-world dataset of wines. Specifically, we aim to identify natural groupings — or "archetypes" — of wine based solely on their chemical properties, without access to quality labels or human-crafted categories. We explore whether the chemical composition of wine is sufficient to distinguish styles or profiles, and whether different algorithms converge on similar clusterings.

To do this, we apply three fundamental techniques in unsupervised learning: clustering [2, 3, 4], dimensionality reduction [1, 5], and anomaly detection. **Clustering** involves partitioning data points into groups such that members of the same group are more similar to each other than to those in other groups. **Dimensionality reduction** transforms high-dimensional data into a lower-dimensional representation, preserving the most important structure while enabling visualization and reducing computational complexity. **Anomaly detection** seeks to identify samples that deviate significantly from the norm — these outliers may represent rare or unusual patterns in the data.

We apply these techniques to a dataset of wine samples described by 11 chemical attributes. After standardizing the features, we perform Principal Component Analysis (PCA) for dimensionality reduction, followed by clustering using K-Means, DBSCAN, and Gaussian Mixture Models (GMM). We evaluate clustering quality using the silhouette score and validate our findings with statistical tests including ANOVA and paired t-tests. For anomaly detection, we compare results from DBSCAN and GMM-based likelihood methods. Finally, we use t-SNE for non-linear visualization of the clustering structure, providing intuitive insights into the discovered groupings.

This analysis demonstrates how unsupervised learning can uncover latent structure in real-world data, even in the absence of labeled outcomes. The project emphasizes both interpretability and statistical rigor, combining visual and quantitative methods to support every claim.

2 Methods

Data Description

We used the Wine Quality dataset from Kaggle[1], containing 11 numerical chemical features. The quality label was removed to ensure a fully unsupervised setting.

Data Preprocessing

All features were standardized using Scikit-learn's `StandardScaler`, ensuring zero mean and unit variance. The scaled data was used throughout the analysis.

Dimensionality Reduction

We applied Principal Component Analysis (PCA) [1] to reduce dimensionality for visualization and clustering. The first two principal components were retained. PCA was also used to initialize t-SNE [5] for non-linear projection.

Clustering Algorithms

We applied three clustering algorithms:

- **K-Means** [2]: using the elbow method to determine $K = 3$.
- **DBSCAN** [3]: parameters chosen using a k-distance graph ($\epsilon = 0.17$, $\texttt{min\_samples} = 5$).

[1] https://www.kaggle.com/datasets/taweilo/wine-quality-dataset-balanced-classification

- **Gaussian Mixture Models (GMM)** [4]: with three components to match K-Means.

Anomaly Detection

Outliers were identified using:

- DBSCAN label = -1 (noise points).
- GMM log-likelihood threshold: $\mu - 3\sigma$.

Statistical Evaluation

We computed silhouette scores [6] across algorithms and used one-way ANOVA [7] and paired t-tests to compare clustering performance.

3 Results

Dimensionality Reduction with PCA

The first principal component captured approximately 60% of the variance, and the first three components explained about 80%. Projecting the samples onto the first two components revealed two visible high-density regions (Figure 1, 2).

K-Means Clustering

K-Means with $K = 3$ produced three well-separated clusters (Figure 4), supported by a silhouette score of **0.5119**. This indicates good internal cohesion and separation.

DBSCAN Clustering and Outlier Detection

DBSCAN identified 8 clusters and 155 outliers (Figure 6). While its silhouette score was lower (**0.1854**), it successfully detected complex, non-convex structures and low-density outliers.

GMM Clustering

GMM with three components generated clusters similar to K-Means (Figure 7). It achieved a silhouette score of **0.5140**, slightly higher than K-Means, while providing soft assignments and modeling elliptical shapes.

Cluster Interpretation

Feature	Cluster 0	Cluster 1	Cluster 2
Alcohol	-0.677	0.388	0.243
Residual Sugar	-1.360	0.559	0.644
Fixed Acidity	-1.070	0.274	0.625
pH	0.349	0.770	-0.815
Sulphates	-1.199	0.445	0.602

Table 1: Mean z-score of selected features per cluster (K-Means, $K = 3$).

To characterize each cluster, we calculated the average values of key chemical features. Cluster 0 has lower alcohol and sugar, Cluster 1 shows higher alcohol and sulphates, and Cluster 2 is more acidic with lower pH.

Anomaly Detection

DBSCAN flagged 155 samples (0.74%) as outliers. GMM identified 352 anomalies (1.68%) using a log-likelihood cutoff (Figure 9, 8). The two methods captured different types of rare samples, with partial overlap. Most GMM anomalies lie near the fringes of clusters in PCA space.

While our method used threshold rules to identify outliers, future analysis could include statistical tests to confirm whether these unusual samples are truly different from the rest of the data.

Statistical Comparison of Clustering Quality

Silhouette scores [6] across algorithms are shown in Figure 10. A one-way ANOVA [7] yielded $p = 1.29 \times 10^{-36}$, indicating significant differences. Paired t-tests showed a slight but significant difference between K-Means and GMM ($p = 0.0162$), and a strong difference between K-Means and DBSCAN ($p = 2.3 \times 10^{-14}$).

Comparison	p-value	Interpretation
K-Means vs GMM	0.0162	Slightly significant
K-Means vs DBSCAN	2.3×10^{-14}	Highly significant

Table 2: Paired t-test results comparing silhouette scores across clustering algorithms.

Visualization

t-SNE [5] projected the wine samples into two dimensions (Figure 11). The resulting plot shows three well-defined regions consistent with K-Means clustering, highlighting the separation in a non-linear embedding.

Figures

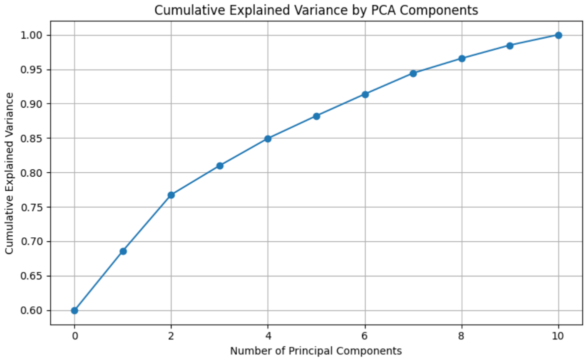

Figure 1: Cumulative explained variance by number of principal components.

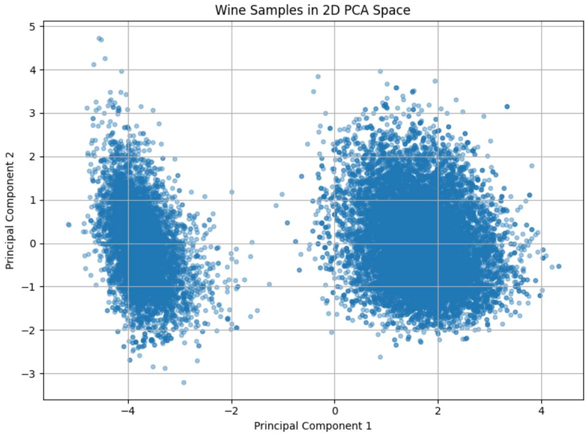

Figure 2: 2D PCA scatter plot of wine samples in the first two components.

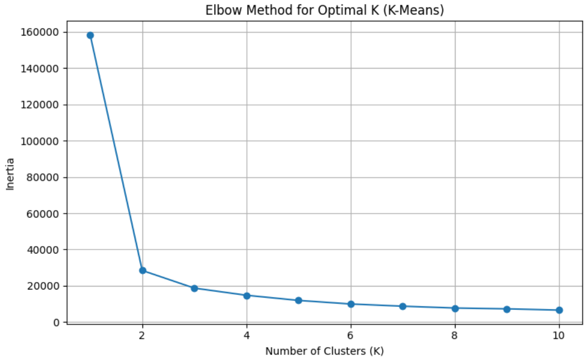

Figure 3: Elbow method for determining optimal K in K-Means ($K = 3$).

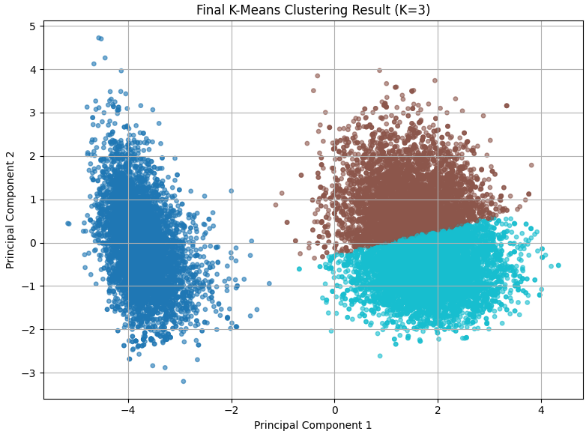

Figure 4: K-Means clustering result with $K = 3$ in PCA space.

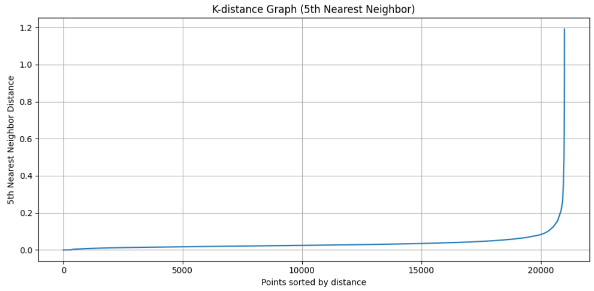

Figure 5: K-distance graph for selecting ϵ in DBSCAN.

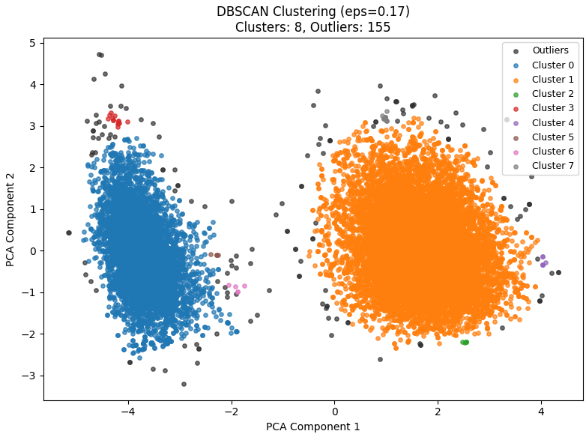

Figure 6: DBSCAN result with 155 outliers shown in gray.

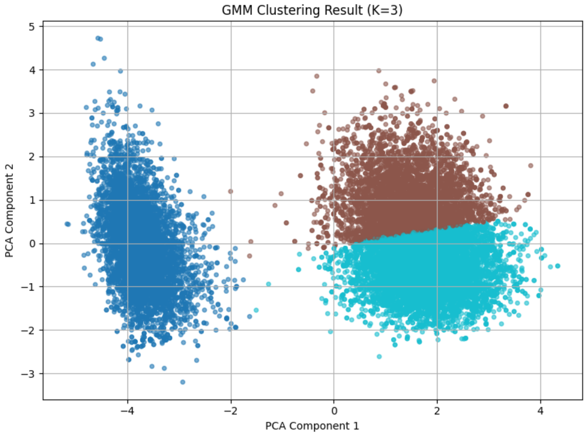

Figure 7: GMM clustering with $K = 3$ in PCA space (soft assignments).

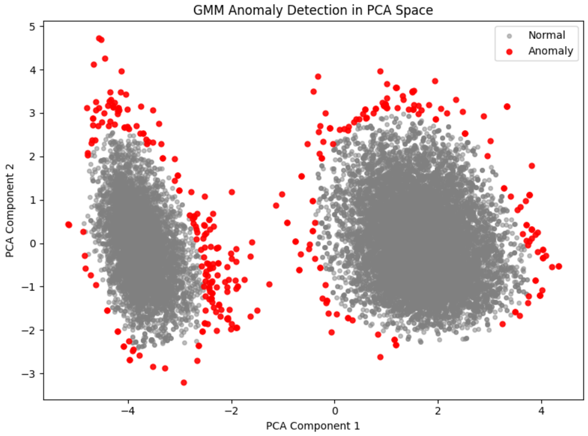

Figure 8: GMM anomaly detection: outliers in red (low log-likelihood).

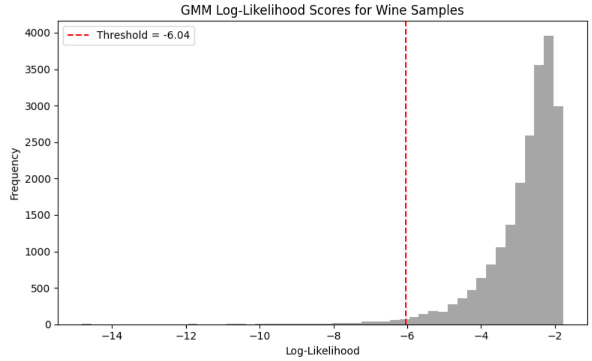

Figure 9: Histogram of GMM log-likelihood scores. Red line shows anomaly threshold.

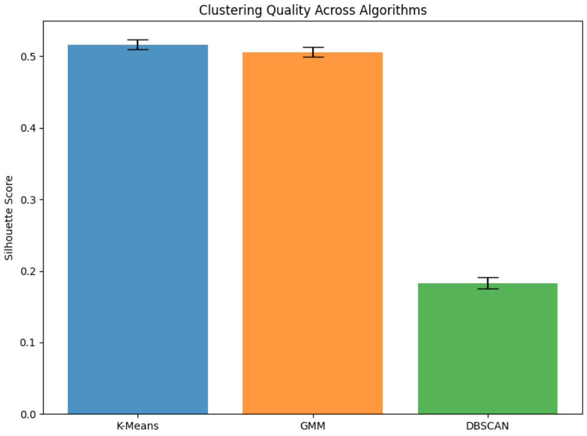

Figure 10: Silhouette scores across algorithms with standard deviation.

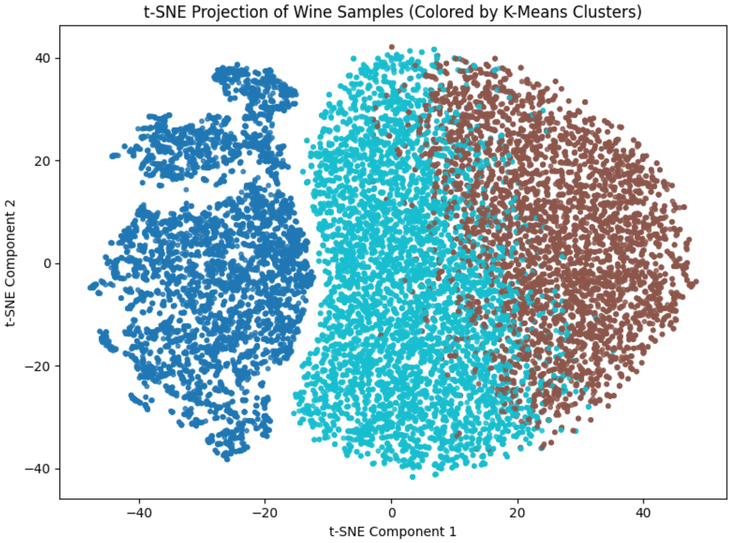

Figure 11: t-SNE projection of wine samples colored by K-Means cluster labels.

4 Discussion

In this project, we explored the latent structure of a wine dataset using unsupervised learning. By applying dimensionality reduction, clustering, and anomaly detection methods, we uncovered meaningful subgroups based solely on chemical composition—without using quality labels. Each claim was supported by visual or statistical evidence, forming a robust analysis pipeline.

Beyond technical performance, our results raise broader questions: Can wines be grouped into distinct chemical profiles that reflect different wine styles? Do some samples deviate enough to suggest alternative production methods or origins? Can such groupings help winemakers better understand their products?

K-Means and GMM revealed three compact, well-separated clusters with similar silhouette scores (≈ 0.51), suggesting consistent structure. DBSCAN detected more complex shapes and 155 outliers but with lower cohesion (score 0.18). Each method offered complementary perspectives—centroid-based, density-based, and probabilistic.

Analyzing the chemical profiles of the clusters revealed distinctive traits: Cluster 0 had low alcohol and sugar, Cluster 1 had higher alcohol and sulphates, and Cluster 2 had high acidity and low pH. These suggest that Cluster 0 may represent lighter wines, Cluster 1 stronger or preserved wines (e.g., reds), and Cluster 2 sharper or longer-lasting profiles. While speculative, these patterns hint at meaningful styles emerging naturally from chemistry alone.

Anomalies flagged by DBSCAN and GMM may indicate experimental or flawed batches. For example, low-alcohol/high-acidity wines could signal fermentation issues or niche production. Identifying such outliers can assist in quality control or innovation.

Visual tools like t-SNE proved invaluable in revealing non-linear separation, enhancing interpretability. Future work could integrate external labels (e.g., grape type, origin) to validate or enrich the discovered structure.

Overall, unsupervised learning not only exposed hidden patterns but suggested how they may translate into real-world wine segmentation—helping producers, marketers, and researchers better understand complex, unlabeled data.

References

[1] I. T. Jolliffe, *Principal Component Analysis.* Springer Series in Statistics, 2002.

[2] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations", in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[5] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE", Journal of Machine Learning Research, vol. 9, 2008.

[6] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics, 1987.

[7] R. A. Fisher, *Statistical Methods for Research Workers.* Oliver and Boyd, 1925.