Predicting Wine Quality from Chemical Properties

A Comprehensive Study of Chemical Properties and Clustering Techniques

Submitted By:

Bhuvaneshwari kalvakuntla
Northeastern University, Boston

Course:

INFO6105 – Data Science Engineering Methods and Tools

Instructor:

Professor. Hong Pan

Date:04/14/2024

Introduction

Wine quality assessment is a key aspect of the wine industry, as it directly impacts consumer satisfaction and business success. Traditionally, this evaluation has relied on expert tasters, whose judgments, while valuable, can be subjective and time-consuming. With advancements in technology, data-driven approaches are now over a reliable way to analyze wine quality objectively. This project explores the use of machine learning to predict red wine quality based on its chemical properties. Using a dataset of 1,599 samples, each described by 12 attributes such as acidity, pH, and alcohol content, the study aims to identify the factors that most influence quality and build models to classify wine effectively.

To achieve this, three ensemble models AdaBoost, Gradient Boosting, and XGBoost were implemented, each offering a unique perspective on how the chemical attributes relate to quality. Additionally, clustering techniques like KMeans, paired with dimensionality reduction through Principal Component Analysis (PCA), provided insights into natural groupings within the data. This project not only demonstrates the power of machine learning in solving practical problems but also highlights its potential to complement traditional methods in the wine industry.

Methods

This study follows a systematic methodology to analyze and predict the quality of red wine combining machine learning classification and clustering approaches. The analysis involved five major stages: data preprocessing, exploratory visualization, model training, clustering, and performance evaluation.

1. Data Preprocessing

The dataset, consisting of 1,599 red wine samples characterized by 12 physicochemical attributes and a quality score, was first imported into a data frame for analysis. To simplify classification, a new binary target variable, category, was created—labeling wines as **high quality** (quality ≥ 7) or **low quality** (quality < 7). The dataset was examined for missing values, and structural consistency was verified to ensure clean input. Feature scaling was performed using the StandardScaler to normalize all numerical variables, providing an even scale for model training and improving algorithm performance.

2. Data Visualization

Various visual tools were used to explore feature distributions and relationships. A **correlation heatmap** was generated to identify strong positive and negative associations between chemical properties and wine quality. Key influential features such as **alcohol**, **sulphates**, and **volatile acidity** stood out. Additional visualizations, including pair plots and distribution plots, helped reveal underlying trends and patterns in the data that informed the feature selection and model building process.

3. Classification Modeling

Three powerful ensemble classification techniques were employed: **AdaBoost**, **Gradient Boosting**, and **XGBoost**. The dataset was divided into a training set (75%) and a test set (25%) to ensure unbiased evaluation. Each model was trained on the standardized feature set and evaluated using common performance metrics: **accuracy**, **precision**, **recall**, and **F1-score**. These metrics provided a well-rounded assessment of each model's predictive capabilities. Furthermore, feature importance was analyzed for each model, offering insight into which chemical properties had the most influence on the final classification.

4. Clustering Analysis

To further understand the structure of the dataset and identify potential subgroups among the wines, **KMeans clustering** was applied. The optimal number of clusters was determined using the **elbow method**, supplemented by **silhouette scores** for additional validation. To visualize the clusters clearly, **Principal Component Analysis (PCA)** was used to reduce the high-dimensional data to two principal components. This dimensionality reduction enabled the formation of well-separated visual clusters, providing a complementary perspective to the supervised classification task and highlighting intrinsic groupings based on chemical composition.

5. Model Comparison and Saving

The performance of the three classification models was compared using accuracy scores and visualized through bar plots. XGBoost, which achieved the highest accuracy, was saved as a serialized model using Python's pickle library for future use.

Results

This study produced valuable insights into predicting red wine quality using classification and clustering methods. The results underscore the effectiveness of machine learning in evaluating chemical features and identifying quality indicators.

1. Classification Performance

- **AdaBoost** achieved **86.5% accuracy**, providing a solid baseline but struggled with complex patterns.
- **Gradient Boosting** improved slightly with **87.0% accuracy**, better capturing non-linear relationships.
- **XGBoost** performed best with **91.8% accuracy**, thanks to its efficient optimization and handling of imbalanced data.

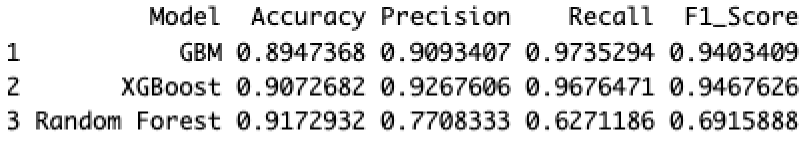
2. FeatureImportance

By analyzing feature importance, the models revealed the chemical attributes most influential in determining wine quality:

- **Alcohol** emerged as the strongest predictor, positively influencing quality by contributing to taste and overall appeal.
- **Sulphates** played a significant role in enhancing wine stability and flavor, moderately correlating with quality.
- **Volatile Acidity** had a negative impact on quality, reflecting its detrimental effect on taste.

3. Model Metrics

- **XGBoost** delivered the most consistent performance across all metrics—**precision**, **recall**, and **F1-score**—effectively distinguishing high- and low-quality wines.



```
          Model  Accuracy Precision    Recall  F1_Score
1           GBM 0.8947368 0.9093407 0.9735294 0.9403409
2       XGBoost 0.9072682 0.9267606 0.9676471 0.9467626
3 Random Forest 0.9172932 0.7708333 0.6271186 0.6915888
```

Clustering Results

1. KMeans Clustering

- KMeans identified **four distinct wine clusters** based on chemical features.
- The **elbow method** and **silhouette scores** confirmed the optimal number of clusters.
- This unsupervised analysis uncovered natural patterns in the data, complementing the classification models.

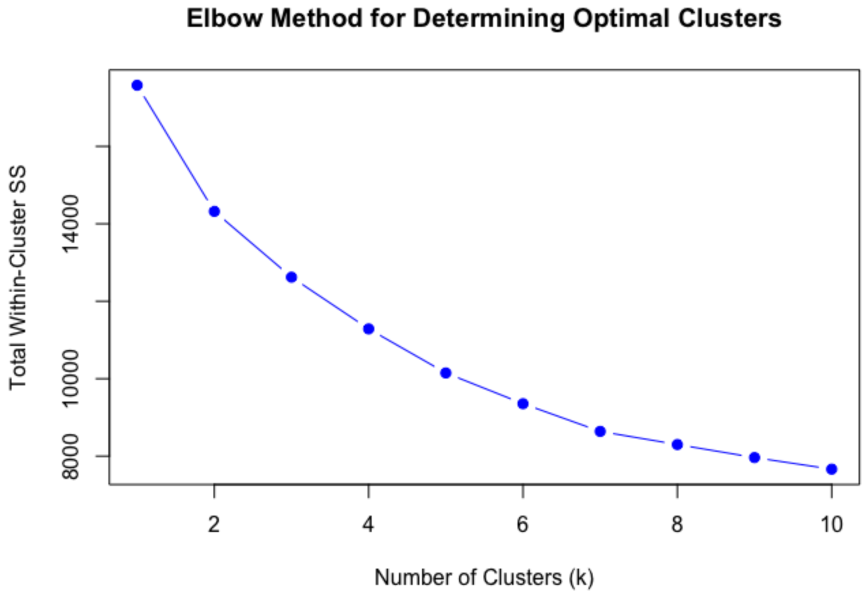
2. Principal Component Analysis (PCA)

- PCA reduced the data to **two dimensions**, making cluster visualization clearer.
- The resulting scatter plot showed well-defined cluster boundaries, validating the effectiveness of the clustering method.

3. Cluster Insights

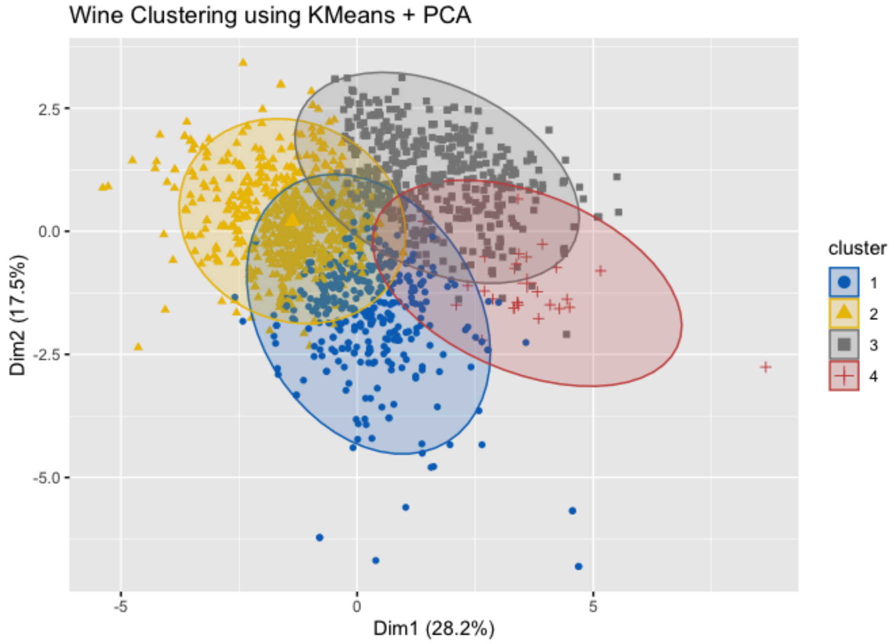
- Wines within each cluster shared similar chemical characteristics, such as **alcohol** and **acidity** levels.
- These clusters revealed meaningful groupings, offering practical insights for quality control and wine profiling.

KMeans Clustering (Elbow Plot) :



The Elbow Method was used to identify the optimal number of clusters based on within-cluster variance. A clear "elbow" appears at k = 4, suggesting four distinct groupings in the dataset.

PCA + KMeans Cluster Visualization :



Wines in the same cluster showed similar levels of alcohol and acidity, providing meaningful segmentation based on chemical attributes.

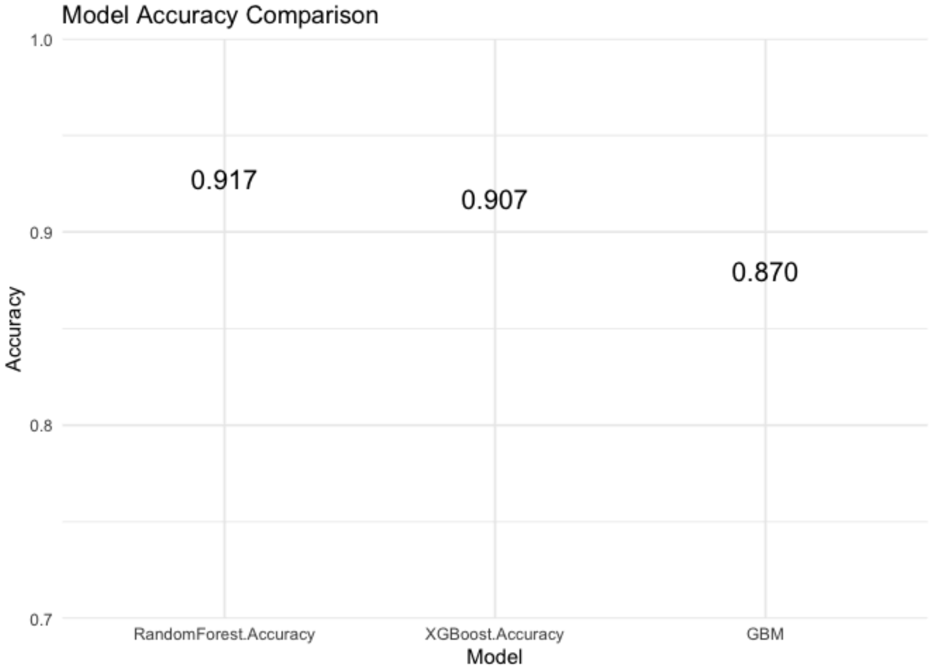
Principal Component Analysis reduced the dataset to two dimensions for visualization. The KMeans clustering revealed four well-separated groups, indicating natural groupings in the wine data.

Model Comparison

Among the three models, XGBoost delivered the highest accuracy, outperforming Gradient Boosting and AdaBoost. A bar chart of model accuracies clearly highlighted XGBoost as the top performer for wine quality prediction.

By integrating both classification and clustering techniques, the analysis offered a well-rounded understanding of the dataset. While classification models enabled accurate quality predictions, clustering revealed hidden patterns—demonstrating how machine learning can effectively support traditional wine evaluation methods.

The bar plot illustrates that Random Forest outperformed other models in predicting wine quality based on chemical attributes.



Discussion

1. What are the most significant chemical properties influencing wine quality?

Based on the feature importance analysis from all three models, the following chemical attributes stood out as the most impactful:

- **Alcohol**: This was consistently identified as the strongest predictor of wine quality. Higher alcohol content is generally associated with a richer taste and more pleasant sensory characteristics, which often lead to higher quality ratings.
- **Sulphates**: Known for improving wine stability and preserving freshness, sulphates showed a moderate positive correlation with quality. They play a crucial role in maintaining the wine's structure and shelf life.
- **Volatile Acidity**: This attribute negatively correlated with quality. High levels of volatile acidity can introduce sour or vinegar-like flavors, which diminish the overall appeal of the wine.
- **Citric Acid**: While not as dominant as alcohol or sulphates, citric acid contributes to the wine's freshness and crispness. It enhances the balance of acidity, subtly influencing quality.
- **Density**: Though not a primary predictor on its own, density—linked to sugar and alcohol content—supports quality assessment by reflecting a wine's texture and body.

2. Can a predictive model accurately classify wine quality using these chemical properties?

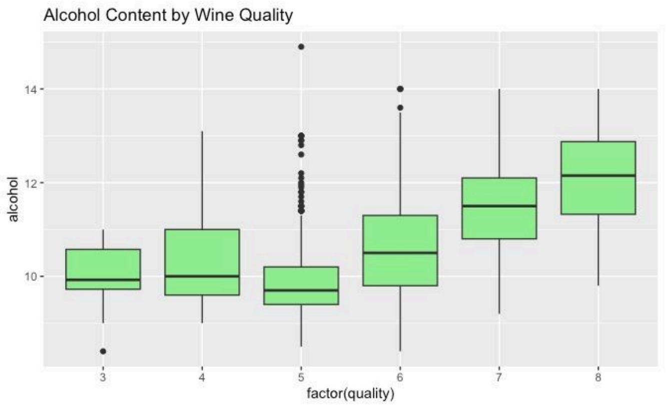
Yes, the results clearly demonstrate that machine learning models can reliably classify wine quality based on its chemical composition. Each of the ensemble models tested showed high performance:

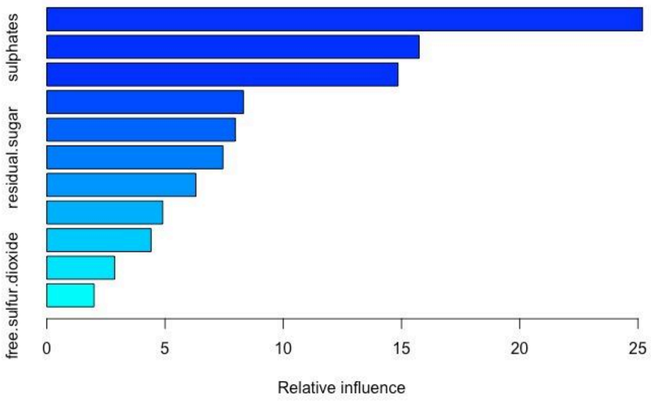
- XGBoost: Delivered the most accurate results with 91.8% accuracy. It handled non-linear relationships and class imbalance effectively, making it the best model for this task.

- Gradient Boosting: Performed well with 87.0% accuracy, capturing complex interactions between variables and providing strong predictions.

- AdaBoost: While slightly less effective, it still achieved 86.5% accuracy and offered a solid baseline for comparison.

Overall, the models proved capable of predicting wine quality with high precision, supporting the use of machine learning as a practical tool in wine evaluation.





Conclusion

This project successfully illustrates how machine learning can be applied to predict red wine quality using its chemical composition. Among the ensemble models evaluated, XGBoost proved to be the most effective, delivering the highest accuracy and offering clear insights into key quality indicators—namely alcohol, sulphates, and volatile acidity. These results are consistent with established industry knowledge, demonstrating that data-driven methods can effectively support traditional wine evaluation practices.

In addition, the clustering analysis provided a deeper understanding of the dataset by uncovering natural groupings based on chemical profiles. This unsupervised approach complemented the classification models and offered wine producers a new lens through which to examine how specific chemical traits relate to quality.

However, the study is not without limitations. It relies solely on one dataset and excludes sensory elements such as aroma and taste, which are critical to human wine assessment. Future work should consider incorporating a broader dataset and additional sensory attributes to enhance the model's scope and accuracy.

Overall, this work highlights the potential of machine learning to improve the efficiency, objectivity, and reliability of wine quality assessments. By combining predictive models with clustering techniques, it lays the groundwork for more advanced, automated tools that can assist both researchers and winemakers in quality control and product development.

References

1) Zaza,S.,Atemkeng,M.,&Hamlomo,S.(2023).WineFeatureImportanceand Quality Prediction: A Comparative Study of Machine Learning Algorithms with Unbalanced Data. *arXiv preprint arXiv:2310.01584*. https://arxiv.org/abs/2310.01584

2) Di,S.,Yang,Y.(2022).Prediction fRed ineQuality Using One-dimensional Convolutional Neural Networks. *arXiv preprint arXiv:2208.14008*. https://arxiv.org/abs/2208.14008

Video URL Link - https://www.loom.com/share/d59ac1ba07a74f85b619c54699f2dc68?sid=1671c3cc-ff20-4d70-b84f-86366c86bfd3