

Comprehensive Analysis of Wine Quality Prediction and Clustering

G.H. AMILA HETTIARACHCHI
STATISTICAL LEARNING – COURSE COORDINATOR: ILIAS THOMAS
JUNE 02, 2024

Introduction

Wine quality prediction is a significant area of interest in both the wine industry and the field of data science. Machine learning algorithms provide an opportunity to explore the intricate relationship between various attributes of wine and its perceived quality. In this research, the application of Decision Trees, Random Forests, and Support Vector Machines (SVM) to predict wine quality is investigated. Furthermore, the effects of parameter optimization and pruning techniques on model performance are explored. Additionally, clustering techniques such as K-means and hierarchical clustering are employed to uncover natural groupings within the wine dataset. Through this research, a deeper understanding of wine quality prediction is aimed for, along with insights into potential improvements in model accuracy and clustering performance.

Research Question

1. How can wine quality be predicted using Decision Tree, Random Forest, and SVM models?
2. What are the effects of cross-validation on model performance?
3. What are the effects of pruning on the performance of the Decision Tree model?
4. What are the results of kernel selection and parameter optimization in SVM models?
5. How can the performance of Random Forest models be further improved?
6. Which model demonstrates the best performance and accuracy?
7. What are the optimal numbers of clusters when using K-means and hierarchical clustering on the wine dataset?
8. Can the clusters obtained using K-means with $k=2$ be validated against the wine type labels (red and white)?
9. Do K-means and hierarchical clustering provide the same clustering results?

Methodology

Data Preprocessing

The given datasets "winequality-red.csv" and "winequality-white.csv" were utilized to create a combined dataset. A new column called "wine_type" was introduced to distinguish between red and white wines, with red wine encoded as 1 and white wine encoded as 0 for the prediction problem. After combining the datasets, the total dataset consisted of 6497 observations and 13 variables. Table I describes the data dictionary including features, data types, and descriptions about the data.

Table I Data Description

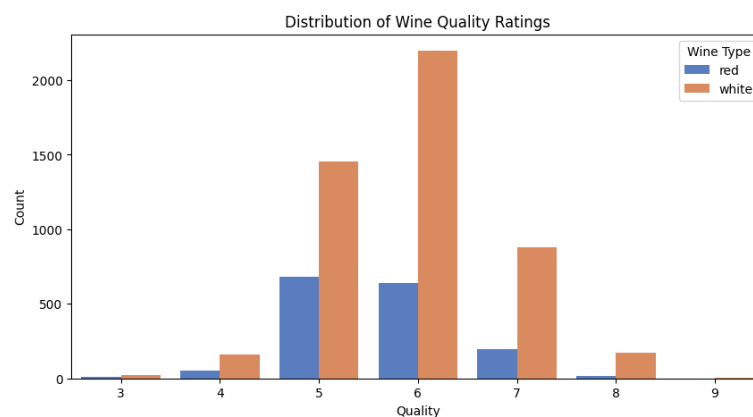
Feature	Data Type	Description
fixed acidity	float64	The amount of fixed acids (such as tartaric acid) in wine.
volatile acidity	float64	The amount of volatile acids (such as acetic acid) in wine.
citric acid	float64	The amount of citric acid in wine.
residual sugar	float64	The amount of residual sugar in wine after fermentation.
chlorides	float64	The amount of chlorides (salt) in wine.
free sulfur dioxide	float64	The amount of free SO ₂ in wine.
total sulfur dioxide	float64	The total amount of SO ₂ in wine.
density	float64	The density of wine.

pH	float64	The pH level of wine.
sulphates	float64	The amount of sulfates in wine.
alcohol	float64	The alcohol content in wine.
quality	int64	The quality rating of wine (typically on a scale from 0 to 10).
wine_type	String	The type of wine (e.g., red or white).

The dataset was cleaned to handle missing values, with none detected. All variables were converted to numeric values to facilitate linear regression analysis.

Figure 1.0 illustrates the distribution of white and red wine qualities. The distinct distributions indicate that white wine generally exhibits better quality compared to red wine. This observation is highlighted by the superior quality of white wine evident in the distribution.

Figure 1.0 Wine Quality Rating Distribution



Model Development

To predict wine quality, the dataset was randomized before being split into 80% training and 20% testing subsets. Each model (Decision Tree Regression, SVM Regression, and Random Forest Regression) was trained on the training subset to learn the relationship between attributes and wine quality using default parameters.

Model Optimization

10-fold cross-validation was used to boost model accuracy and optimize parameters, with 10% of the data reserved for validation. Parameter adjustments and pruning techniques were applied for improved performance.

1. Decision Tree Regression: Evaluation was conducted before and after pruning, adjusting the `ccp_alpha` parameter for Cost-Complexity Pruning to control tree complexity.
2. SVM Regression: Performance analysis involved testing different kernel functions (e.g., linear, RBF) and optimized parameters (e.g., `C` and `gamma`) to find the best configuration.
3. Random Forest Regression: Evaluation included testing various configurations of `n_estimators`, representing the number of trees, to optimize predictive accuracy.

Additionally, the bagging method with bootstrapping was employed to enhance model robustness and performance.

Performance Comparison

The performance of each regression model was evaluated using several metrics:

1. Root Mean Squared Error (RMSE): This metric measures the square root of the average squared difference between observed and predicted values, providing a measure of the model's accuracy.
2. R-squared (R^2): This metric indicates the proportion of variance in the dependent variable that is explained by the independent variables, reflecting the model's explanatory power.

K-means Clustering

For the K-means Clustering analysis, the goal was to determine the optimal number of clusters in the wine dataset. This was achieved through the utilization of the elbow method and silhouette analysis. The K-means algorithm was executed for various values of k , representing the number of clusters, and the within-cluster sum of squares was plotted. The objective was to identify the "elbow point," where the addition of more clusters led to diminishing returns in reducing the within-cluster sum of squares, indicating an optimal number of clusters. Furthermore, when $k=2$, the resulting clusters were analyzed to evaluate their alignment with the red and white wine labels. This validation process involved comparing the cluster assignments with the known wine type labels to determine the effectiveness of the clustering algorithm in segregating the data based on wine types.

Hierarchical clustering

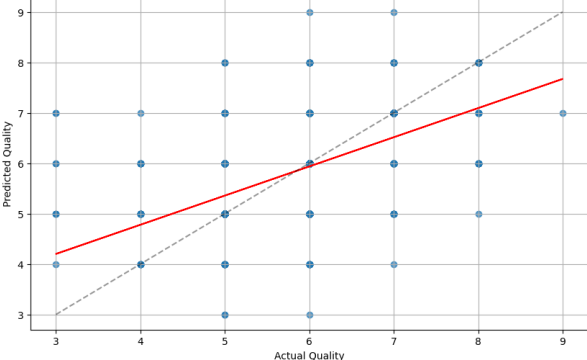
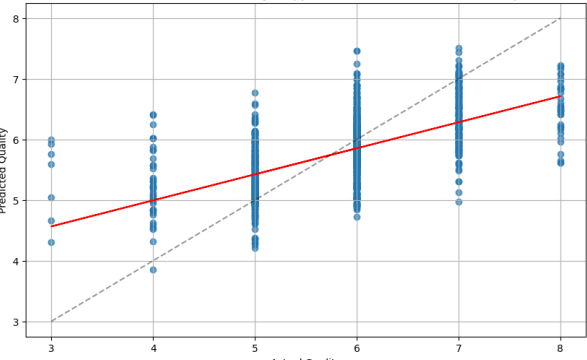
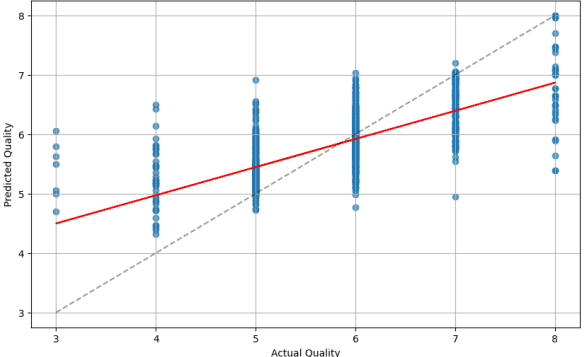
Hierarchical clustering was employed to provide an alternative perspective on the data's structure. Dendrograms, graphical representations illustrating the hierarchical clustering process, were utilized to visualize the clustering hierarchy and determine the optimal number of clusters. Utilizing various distance metrics such as Euclidean, Minkowski, and Pearson Correlation Distance, along with linkage methods including Single, Complete, Average, and Centroid, agglomerative clustering, a hierarchical clustering technique, was applied to the dataset. Dendrograms were then analyzed to identify the most suitable cluster separation. Subsequently, the outcomes of hierarchical clustering were compared with those of K-means clustering to ensure consistency in the clustering results, facilitating a comprehensive understanding of the inherent groupings within the data.

Data Analysis

The analysis and results directly address the outlined research questions, detailing performance metrics of Decision Tree, Random Forest, and SVM models. The impact of pruning on Decision Tree models is examined, as well as kernel selection and parameter optimization in SVM models. Parameter optimization effects on Random Forest models are evaluated. Additionally, optimal clustering techniques, like K-means and hierarchical clustering, are explored to validate clustering results. Recommendations are provided for enhancing model accuracy and clustering performance.

1. **How can wine quality be predicted using Decision Tree, Random Forest, and SVM models?** Each model effectively predicts wine quality, leveraging unique strengths (see Table II).

Table II Model evaluation Train test Splits

Model Parameters		Model Evaluation		Decision tree with train and test split
Criterion	squared_error	Test RMSE	0.817	
Splitter	best	R ²	0.096	
Train Split (%)	80	Test Split (%)	20	
The Root Mean Squared Error (RMSE) of approximately 0.817 indicates the average prediction error, while the R-squared (R ²) value of around 0.096 suggests limited explanatory power of the model. Further refinement or exploration of alternative models may be necessary to enhance predictive accuracy.				
Model parameters		Model evaluation		SVM with train and test split
kernel	rbf	RMSE	0.817	
C	2	R ²	0.096	
Gamma	scale			
Train Split (%)	80	Test Split (%)	20	
The Model yielded a RMSE of 0.679, with an R-squared value of 0.371. These results were obtained with the RBF kernel, a C value of 2.0, and Gamma set to 'scale'. Overall, the SVM model demonstrated reasonable performance in predicting wine quality.				
Model parameters		Model evaluation		Random forest with train and test split
Criterion	squared_error	RMSE	0.603	
n_estimators	550	R ²	0.505	
max_features	sqrt			
Train Split (%)	80	Test Split (%)	20	
The Random Forest model attained an RMSE of 0.606 and an R ² value of 0.500 with n_estimators set to 100. Increasing the n_estimators to 550 improved the R ² slightly to 0.504 and reduced the RMSE to 0.603. These findings indicate that the Random Forest model demonstrated moderate performance in predicting wine quality.				

2. **What are the effects of cross-validation on model performance?**

10-fold cross-validation improved model reliability, highlighting overfitting in some cases (see Table III)

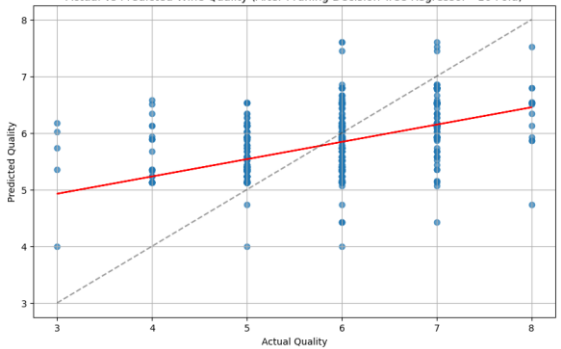
Table III Cross-validation on model performance

Model Parameters		Model Evaluation		Decision tree with 10-fold cross validation
Criterion	squared_error	Tra. RMSE	0.847	
Splitter	best	Test RMSE	0.968	
Train Split	10-fold	R ²	-0.257	
		Unseen (%)	10	
<p>After 10-fold cross-validation, the Decision Tree model showed increased RMSE (0.968) and decreased R-squared (-0.257), indicating higher prediction error and reduced explanatory power. This decline suggests potential overfitting or insufficient generalization, underscoring the need for further optimization or alternative modeling approaches.</p>				
Model parameters		Model evaluation		SVM with 10-fold cross validation
kernel	rbf	Tr. RMSE	0.674	
C	1	Test RMSE	0.692	
Gamma	scale	R ²	0.358	
Train Split	10-fold	Unseen (%)	10	
<p>After 10-fold cross-validation, the SVM model exhibited a training Mean RMSE of 0.674 and a test RMSE of 0.692. The R-squared value on the test data was approximately 0.358. These metrics suggest that the model performed moderately well in explaining the variance in the wine quality data, with a relatively low level of prediction error.</p>				
Model parameters		Model evaluation		Random forest with 10-fold cross validation
Criterion	squared_error	Tr. RMSE	0.668	
n_estimators	550	Test RMSE	0.699	
max_features	sqrt	R ²	0.446	
Train Split	10-fold	Unseen (%)	10	
<p>After 10-fold cross-validation, the Random Forest model demonstrated a train RMSE of 0.592, a test RMSE of 0.642, and an R-squared value of 0.446. These results indicate that the model performed well, achieving a good balance between prediction accuracy and the ability to explain the variance in the wine quality data.</p>				

3. What are the effects of pruning on the performance of the Decision Tree model?

Pruning reduced RMSE and improved R-squared, enhancing model generalization (see Table IV).

Table IV After Pruning the Decision tree with 10-fold cross validation.

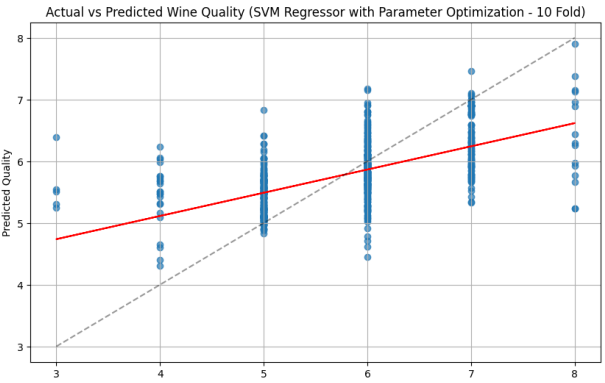
Model Parameters		Model Evaluation		Decision tree with 10-fold after pruning 
Criterion	squared_error	Tra. RMSE	0.735	
Splitter	best	Test RMSE	0.774	
Train Split	10-fold	R ²	0.197	
		Unseen (%)	10	
After 10-fold cross-validation, the Decision Tree After pruning the Decision Tree model, the training RMSE was 0.735, and the test RMSE was 0.774. The test R-squared improved to 0.197, indicating better generalization and reduced overfitting.				

4. What are the results of kernel selection and parameter optimization in SVM models?

Optimizing kernel parameters improved SVM performance, with better accuracy metrics (see Table V).

- C: Balances training error and model complexity.
- Gamma: Sets the influence range of training examples.
- RBF kernel: Maps inputs to a higher-dimensional space for non-linear data.

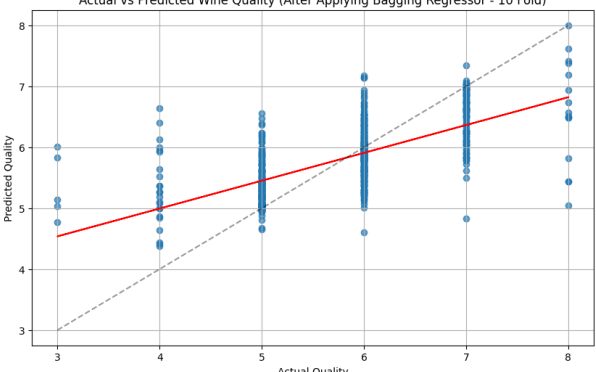
Table V SVM after parameter optimization K = 10

Model parameters		Model evaluation		SVM after parameter optimization K = 10 
Kernel	rbf	Tr. RMSE	0.649	
C	1	Test RMSE	0.693	
Gamma	1	R ²	0.356	
Train Split	10-fold	Unseen (%)	10	
After parameter optimization in SVM, the best parameters found were {'C': 1, 'gamma': 1, 'kernel': 'rbf'}. This resulted in improved model performance, with a train mean RMSE of 0.649 and a test RMSE of 0.693. Additionally, the R-squared value increased to 0.356, indicating better predictive accuracy. ----- Considered Parameters and tested values----- C: [0.1, 1, 10, 100], Gamma: [0.01, 0.1, 1, 10] Kernel: [rbf]				

5. How can the performance of Random Forest models be further improved?

Fine-tuning `n_estimators` and using bagging with bootstrapping enhanced Random Forest performance (see Table VI). This did not improve the model's performance as expected.

Table VI Bagging approach to the Random Forest with bootstrapping

Model parameters		Model evaluation		<h3>Bagging approach to the Random Forest with bootstrapping</h3> 
n_estimators	550	Tr. RMSE	0.593	
Bootstrap	enabled	Test RMSE	0.641	
		R ²	0.449	
Train Split	10-fold	Unseen (%)	10	
<p>After applying bagging with bootstrapping to the Random Forest model, the train mean RMSE decreased to 0.593, and the test RMSE decreased to 0.641. Additionally, the R-squared value increased to 0.449, but this is not over run the best model in predictive accuracy compared to previous results.</p>				

6. Which model demonstrates the best performance and accuracy?

Random Forest showed the best performance with the lowest RMSE and highest R-squared.

7. How many clusters are optimal based on the results?

The optimal number of clusters was determined for K-means and hierarchical clustering using methods like the elbow method and silhouette analysis. Based on the Elbow plot analysis, described in Figure 3.0 the optimal number of clusters is determined. The Elbow plot indicates that three clusters could be suitable for K-means clustering. However, given that the dataset includes the actual wine type values, K equal to 2 is considered. As per the assignment instructions, it is decided to proceed with K equal to 2.

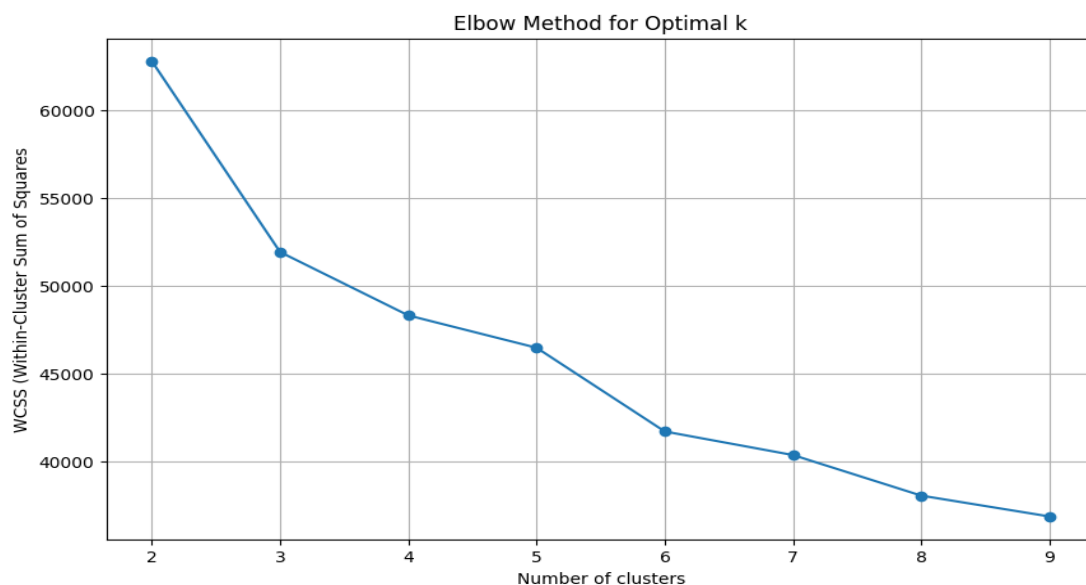


Figure 3.0 Elbo plot for the K means Clustering.

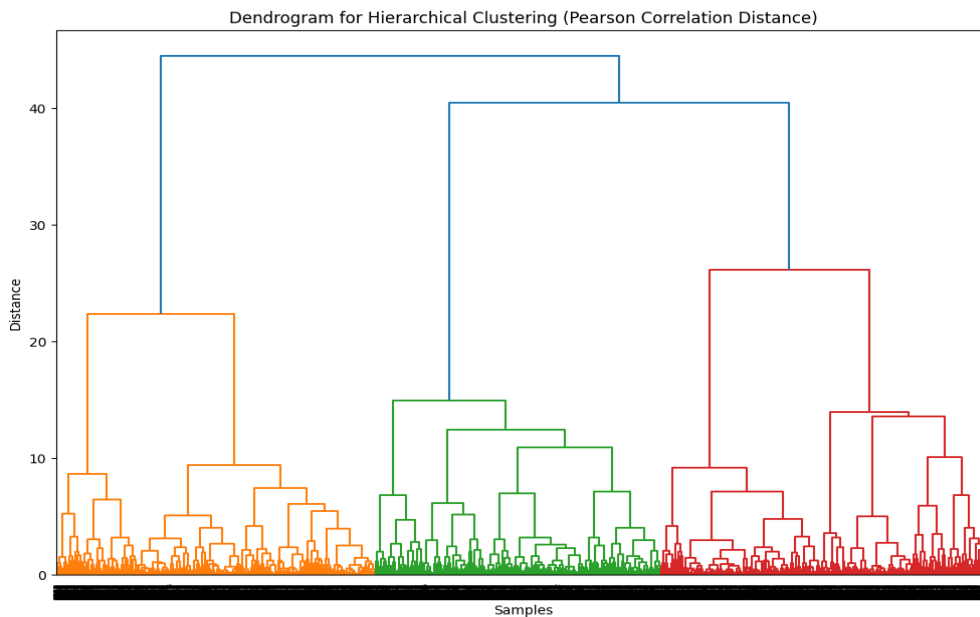


Figure 4.0 Elbo plot for the Hierarchical Clustering.

Complete linkage method used in the dendrogram. In the clusters could be interpreted as two at a higher level. However, upon further visualization, a potential cluster of three can be observed. Therefore, cluster three could be considered the most suitable option. However, for the purpose of this assignment, it is decided to proceed with cluster 2.

8. Do K-means and hierarchical clustering provide the same clustering results?

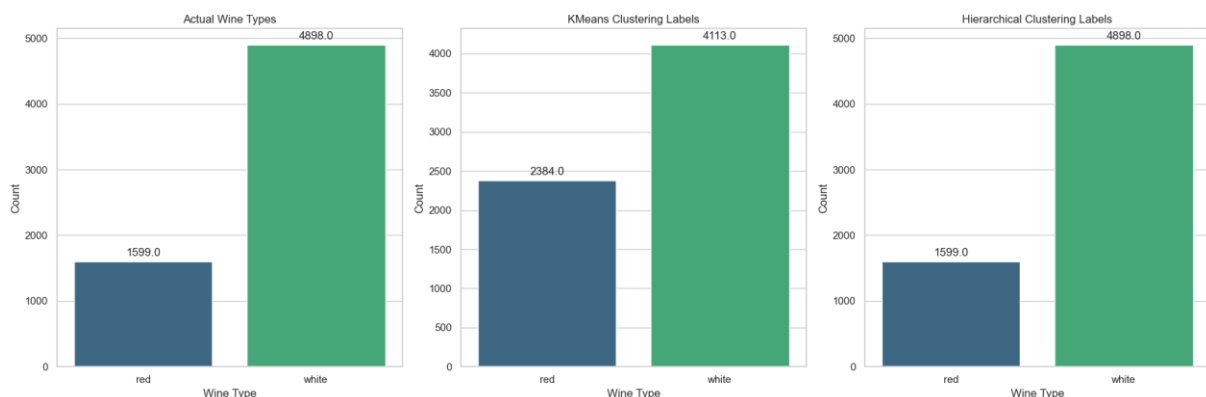


Figure 5.0 Comparison of Wine Type counts between the actual data and the results obtained from K-means and Hierarchical Clustering

As depicted in Figure 5.0, the actual wine types consist of 4898 white wine entries and 1599 red wine entries. In K-means clustering, there are 4113 white and 2384 red wine entries, while in hierarchical clustering, the counts remain the same as the actual data, with 4898 white and 1599 red wine entries. Both clustering methods provide good predictions, but hierarchical clustering achieves 100% accuracy in replicating the actual wine types.

9. Does k-means and hierarchical clustering provide the same results?

Both KMeans and Hierarchical Clustering demonstrate high performance, as shown by their confusion matrices. In KMeans, out of 5,497 instances, only 28 were misclassified, resulting in an accuracy of

99.6%. Similarly, Hierarchical Clustering achieved an accuracy of 99.5%, with only 30 misclassifications out of 5,497 instances. Both models exhibit excellent precision, recall, and F1-score, indicating robust predictive capability with minimal errors. (See Table VII)

Table VII: Accuracy of KMeans and Hierarchical Clustering from Confusion Matrix

Confusion Matrix		
KMeans	[[1586 13] [15 4883]]	Accuracy: 0.996
		Precision: 0.997
		Recall: 0.997
		F1-score: 0.997
Hierarchical Clustering	[[1587 12] [18 4880]]	Accuracy: 0.995
		Precision: 0.998
		Recall: 0.996
		F1-score: 0.997

Conclusion

The exploration of wine quality prediction using machine learning techniques has provided valuable insights into the predictive capabilities of various models and the underlying structure of the wine dataset. Among the models tested, Random Forests generally showed superior performance, particularly after parameter optimization.

Cross-validation played a vital role in improving model performance, underscoring the importance of robust validation techniques in achieving efficient and reliable predictions. This comprehensive analysis highlights the significance of parameter tuning and validation methods in optimizing model efficiency.

Clustering analysis using K-means and hierarchical methods offered a deeper understanding of natural groupings within the data, aligning well with known wine types. The visualization of clusters further identified previously unknown patterns in the dataset.

Future research could focus on integrating additional features, exploring other machine learning algorithms, and applying advanced ensemble methods to further enhance prediction accuracy and robustness.