

Wine Dataset Analysis

Author: Kanchana Weerasinghe
Statistical Learning (AMI22T)
Home Exercise 2
June 02, 2024

Abstract: *"This study investigates wine quality prediction utilizing decision trees, random forests, and support vector machines (SVM) through statistical learning techniques. Two datasets comprising red and white wine attributes are merged and analyzed. Decision trees are pruned to enhance model performance, SVM models undergo parameter optimization, and random forests are tuned for optimal accuracy. Additionally, clustering analysis using k-means and hierarchical techniques is performed to identify optimal clusters. Results show improved model performance post-pruning, moderate predictive accuracy of SVM models, and medium predictive accuracy of random forest models. Clustering analysis identifies optimal clusters primarily based on wine type"*

1. Introduction

This assignment involves analyzing and modeling wine quality using two datasets: "winequality-red.csv" and "winequality-white.csv". Initially, the datasets will be merged, and labels ("red" or "white") will be assigned to the observations.

The first task includes building predictive models using decision trees, SVM, and random forests to estimate wine quality based on attributes. Decision trees will be pruned, SVM models will explore various kernels and parameters, and random forest parameters will be fine-tuned for optimal performance. Cross-validation will be conducted, and models will be evaluated on a test dataset.

In the second task, k-means and hierarchical clustering techniques will determine the optimal number of clusters in the combined dataset. Clusters, particularly $k=2$, will be validated against assigned labels to align with wine types. The consistency between k-means and hierarchical clustering results will be compared.

1.1 Following research questions will be answered in this assignment:

1. How can the quality of wine be predicted using decision trees, random forests, and SVM models based on the attributes described in the datasets?
2. What are the results of cross-validation and performance evaluation of the models on the testing dataset?
3. What are the outcomes when the decision trees are pruned?
4. What are the effects of different kernels and parameter optimizations in the SVM model on prediction accuracy?
5. How does parameter optimization in the random forest model impact its performance?
6. How many clusters are optimal based on the results of k-means and hierarchical clustering?
7. How can k-means and hierarchical clustering be applied to perform clustering on the full wine dataset?
8. If $k=2$ is used, can the clusters be validated based on the labels assigned to the wine types (red and white)?
9. Do k-means and hierarchical clustering provide the same clustering results?

1.2 Dataset

The wine dataset consists of two distinct datasets, "winequality-red.csv" and "winequality-white.csv,"

No of rows: 6497

<i>Attribute Name</i>	<i>Description</i>	<i>Type</i>
fixed acidity	The amount of fixed acids (tartaric, citric, etc.) in g/dm ³ .	Numeric
volatile acidity	The amount of acetic acid in g/dm ³ , which can lead to an unpleasant, vinegar taste.	Numeric
citric acid	The amount of citric acid in g/dm ³ , which can add freshness and flavor to wines.	Numeric
residual sugar	The amount of sugar remaining after fermentation in g/dm ³ .	Numeric
chlorides	The amount of salt in the wine in g/dm ³ .	Numeric
free sulfur dioxide	The amount of free SO ₂ in mg/L, which prevents microbial growth and oxidation.	Numeric
total sulfur dioxide	The total amount of SO ₂ in mg/L, including both free and bound forms.	Numeric
density	The density of the wine in g/cm ³ , which is influenced by the sugar content.	Numeric
pH	The pH level of the wine, indicating its acidity or basicity.	Numeric

sulphates	The amount of potassium sulphate in g/dm ³ , which acts as a wine preservative.	Numeric
alcohol	The alcohol content in % vol.	Numeric
quality	The quality rating of the wine (0 to 10).	Numeric
type	The type of wine, either "red" or "white".	Categorical

Table 1: Dataset

2. Methodology

2.1 Data Mining Methods:

2.1.1 Data Preparation

- Import the datasets "winequality-red.csv" and "winequality-white.csv".
- Assign labels to the observations as "red" for red wines and "white" for white wines.
- Merge the two datasets into one comprehensive dataset.
- Kept the 10% as unseen and 90% for model building.
- Perform data cleaning and preprocessing, including handling missing values and normalizing features.

2.1.2 Model Building

- Decision Trees: Build a decision tree model using the combined dataset. Train the model to predict wine quality based on the available attributes.
- Random Forests: Construct a random forest model with multiple trees to improve prediction accuracy.
- Support Vector Machines (SVM): Develop an SVM model to classify wine quality. Experiment with different kernels (linear, polynomial, RBF) and perform parameter optimization.

2.1.3 Model Evaluation

- Cross-Validation: Perform k-fold cross-validation (e.g., k=10) to assess the robustness and generalizability of each model.
- Performance Metrics: Evaluate the models using performance metrics such as RMSE, STD RMSE, R-Squared.

2.1.4 Model Optimization

- Decision Trees: Prune the decision trees to avoid overfitting and improve generalization. Compare the performance before and after pruning.
- SVM: Explore the impact of different kernel functions and optimize the SVM parameters (e.g., C, gamma) using grid search or random search techniques.
- Random Forests: Optimize the random forest model by tuning the number of trees, maximum depth, and other hyperparameters.

2.1.5 Clustering Analysis

- K-Means Clustering: Apply k-means clustering to the full dataset. Determine the optimal number of clusters using the Elbow Method, Silhouette Score, or other relevant metrics.
- Hierarchical Clustering: Perform hierarchical clustering using methods like Ward's linkage. Compare the dendrogram to identify the optimal number of clusters.

2.1.6 Cluster Validation

- For k=2, validate the clusters based on the labels assigned to the wine types (red and white). Assess how well the clusters align with these labels.
- Compare the clustering results from k-means and hierarchical clustering to evaluate consistency and differences.

2.2 Exploratory data analysis:

2.2.1. Combined the Datasets:

Merged the "winequality-red.csv" and "winequality-white.csv" datasets into a single comprehensive dataset, enabling a unified analysis of both red and white wines. Combined dataset was divided as unseen(10%) and training data (90 %)

No of rows: 6497	X_train samples: 5847	X_unseen samples: 650	y_train samples: 5847	y_unseen samples: 650
------------------	-----------------------	-----------------------	-----------------------	-----------------------

2.2.2. Cleaning and Preprocessing:

Addressed missing values by imputing or removing them as necessary.

Normalized the features to ensure all attributes are on a comparable scale, facilitating more accurate model training and analysis in the clustering phase

2.2.3. Frequency Distribution of Wine Type:

Analyzed the frequency distribution of the wine types, determining the proportion of red and white wine samples within the dataset.

Quality	Red Wine	White Wine	Quality	Red Wine	White Wine
3	10	20	6	638	2198
4	53	163	7	199	880
5	681	1457	8	18	175
			9	0	5

Table 2: Wine Quality distribution

2.2.3. Feature Encoding:

Applied feature encoding to transform the categorical wine type attribute into numerical values: red wine was encoded as 0, and white wine was encoded as 1, preparing the data for machine learning models.

2.2.4. Correlation

Correlation Matrix													
fixed acidity	1.00	0.22	0.32	-0.11	0.30	0.28	0.41	0.46	-0.25	0.30	-0.10	-0.08	0.49
volatile acidity	-0.22	1.00	0.38	-0.20	0.38	-0.35	-0.41	0.27	0.26	0.23	-0.04	-0.27	-0.65
citric acid	0.32	0.38	1.00	0.14	0.04	0.13	0.20	0.10	-0.33	0.06	-0.01	0.09	0.19
residual sugar	-0.11	-0.20	0.14	1.00	0.13	0.40	0.50	0.55	-0.27	-0.19	-0.36	-0.04	0.35
chlorides	0.30	0.38	0.04	-0.13	1.00	-0.20	-0.28	0.36	0.04	0.40	-0.26	-0.20	-0.51
free sulfur dioxide	-0.28	-0.35	0.13	0.40	-0.20	1.00	0.72	0.03	-0.15	-0.19	-0.18	0.06	0.47
total sulfur dioxide	-0.33	-0.41	0.20	0.50	-0.28	0.72	1.00	0.03	-0.24	-0.28	-0.27	-0.04	0.76
density	-0.46	0.27	0.10	0.55	0.36	0.03	0.03	1.00	0.01	0.26	-0.69	-0.31	-0.39
pH	-0.25	0.26	-0.33	-0.27	0.04	-0.15	-0.24	0.03	1.00	0.19	0.12	0.02	-0.33
sulphates	0.30	0.23	0.06	-0.19	0.40	-0.19	-0.28	0.26	0.19	1.00	-0.00	0.04	-0.49
alcohol	-0.10	-0.04	-0.01	-0.36	-0.26	-0.18	-0.27	-0.69	0.12	-0.00	1.00	0.44	0.03
quality	-0.08	-0.27	0.09	-0.04	-0.20	0.06	-0.04	-0.31	0.02	0.04	0.44	1.00	0.12
wine_type	-0.49	-0.65	0.19	0.35	-0.51	0.47	0.76	-0.39	-0.33	-0.40	0.03	0.12	1.00

Attribute Pair	Correlation Coefficient	Description
Density and Fixed Acidity	0.67	Strong positive correlation; as fixed acidity increases, density tends to increase.
Alcohol and Density	-0.50	Negative correlation; as alcohol content increases, density tends to decrease.
Alcohol and Volatile Acidity	-0.38	Negative correlation; higher alcohol content is associated with lower volatile acidity.
Citric Acid and Fixed Acidity	0.67	Moderate positive correlation; when fixed acidity increases, citric acid tends to increase.
Total Sulfur Dioxide and Free Sulfur Dioxide	0.72	Moderate positive relationship; as one increases, the other tends to increase too.

Table 3: Correlation

3. Results and Analysis

3.1 How can the quality of wine be predicted using decision trees, models based on the attributes described in the datasets?

What are the results of cross-validation and performance evaluation of the models on the testing dataset?

Mode Type: Decision Tree Regression Model (**Without K-Fold Validation**) | Target Variable: quality | Predicators: All other variables including Wine type | random_state: 456

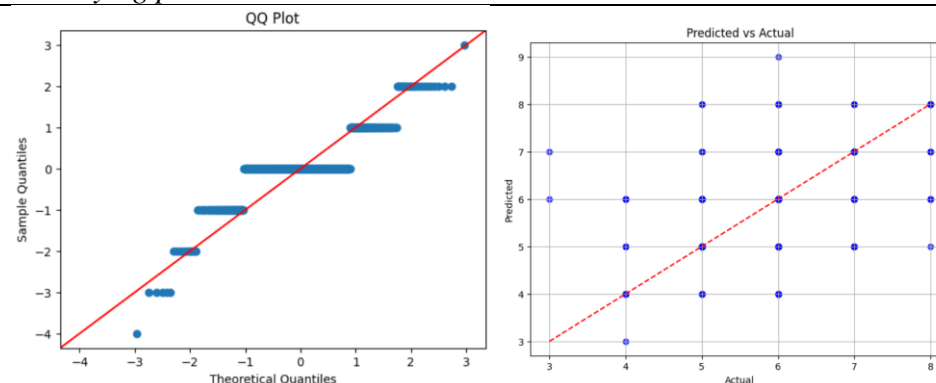
Test RMSE	0.7795
-----------	--------

STD of RMSE	0
-------------	---

R-Squared	0.1455
-----------	--------

Evaluation: *

The decision tree regression without cross validation model demonstrates a Root Mean Squared Error (RMSE) of approximately 0.7795, indicating an average error magnitude of about 0.7795 units in the model's predictions of the quality variable. The R-Squared value is low at 0.1455, suggesting that the model explains only about 14.55% of the variance in the quality scores. These metrics indicate that the model has limited predictive accuracy and may require improvements or alternative approaches to better capture the underlying patterns in the data



Mode Type: Decision Tree Regression Model (**With K-Fold Validation**) | Target Variable: quality | Predicators: All other variables including Wine type | random_state: 456

Training RMSE	0.83
---------------	------

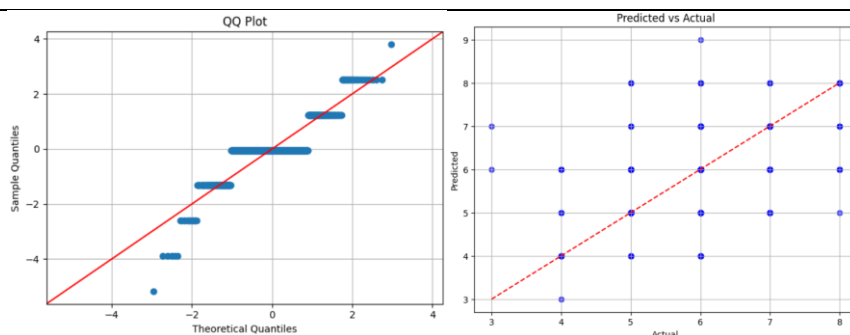
Training STD RMSE	0.024
-------------------	-------

Test RMSE	0.779
-----------	-------

R-Squared	0.14
-----------	------

Evaluation: *

The results indicate that the model has an RMSE of 0.83 for the training data, with a slight variation (standard deviation of RMSE) of 0.024, showing consistency in model performance across different training sets. For the test data, the RMSE is slightly lower at 0.779, suggesting the model performs a bit better on unseen data. However, the R-squared value of 0.14 is low, indicating that the model only explains about 14% of the variance in the target variable, reflecting limited predictive accuracy.



Mode Type: Decision Tree Regression **Pruned Model (With K-Fold Validation)** | Target Variable: quality | Predicators: All other variables including Wine type | random_state: 456 | Best ccp_alpha: 0.00131 | K value:10

Training RMSE	0.735
---------------	-------

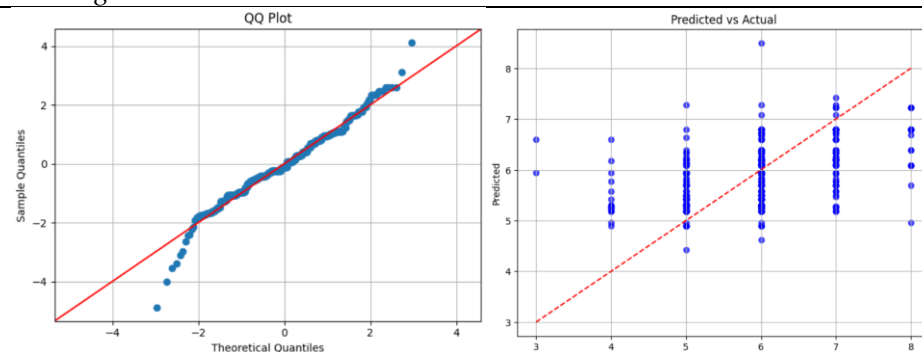
Training STD RMSE	0.025
-------------------	-------

Test RMSE	0.736
-----------	-------

R-Squared	0.238
-----------	-------

Evaluation: **

The pruned Decision Tree Regression model displays a training RMSE of 0.735, with a consistent standard deviation of 0.025 across different training sets. On the test data, the RMSE is comparable at 0.736, indicating similar performance on unseen data. Despite this, the R-squared value of 0.238 shows an improved explanatory power compared to the unpruned model, explaining around 23.8% of the variance in the target variable.



Mode Type: Decision Tree Regression Model (**Bagging**) | Target Variable: quality | Predicators: All other variables including Wine type | random_state: 456 | n_estimators=200 | bootstrap=True | K Value:10

Training RMSE	0.604
---------------	-------

Training STD RMSE	0.018
-------------------	-------

Test RMSE	0.594
-----------	-------

R-Squared	0.502
-----------	-------

Evaluation: ***

The results indicate that the pruned Decision Tree Regression model achieves a training RMSE of 0.604, with a stable standard deviation of 0.018 across various training sets. On the test data, the RMSE slightly improves to 0.594, suggesting consistent performance on unseen data. Notably, the model's R-squared value of 0.502 signifies a substantial improvement in explanatory power compared to the unpruned model, elucidating approximately 50.2% of the variance in the target variable.

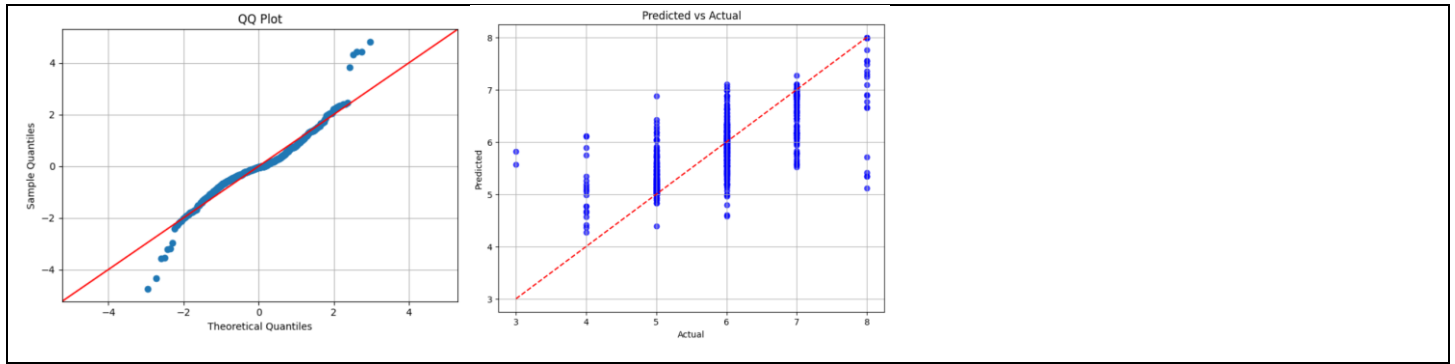


Table 4: Decision Tree Results

3.2 What are the outcomes when the decision trees are pruned?

When comparing the unpruned and pruned Decision Tree Regression models, both exhibit similar performance in terms of test RMSE, with the pruned model slightly outperforming the unpruned model. However, the pruned model demonstrates a notably higher R-squared value, indicating a significant improvement in explanatory power compared to the unpruned model. This suggests that the pruned model better captures the underlying patterns in the data and provides a more accurate representation of the target variable.

3.3 What are the effects of different kernels and parameter optimizations in the SVM model on prediction accuracy?

Mode Type: SVM Model (**Without K-Fold Validation**) | Target Variable: quality | Predicators: All other variables including Wine type | random_state: 456 | kernel: 'rbf'

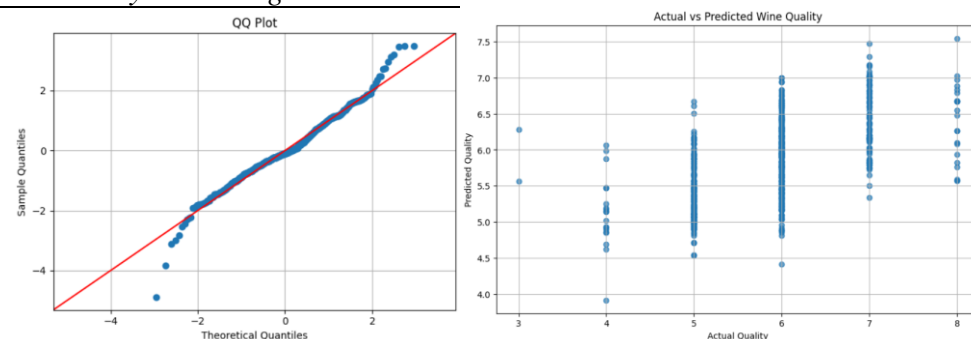
Test RMSE	0.682
-----------	-------

STD of RMSE	0
-------------	---

R-Squared	0.345
-----------	-------

Evaluation: *

The results indicate that the model has a test RMSE of 0.682, with a standard deviation of RMSE at 0, showing no variation across different test sets. The R-squared value of 0.345 indicates that the model explains about 34.5% of the variance in the target variable, reflecting moderate predictive accuracy. This suggests that while the model has consistent performance on the test data, its ability to explain the variability in the target variable is limited.



Mode Type: SVM Model (**With K-Fold Validation**) | Target Variable: quality | Predicators: All other variables including Wine type | random_state: 456 | K:10 | kernel: 'rbf'

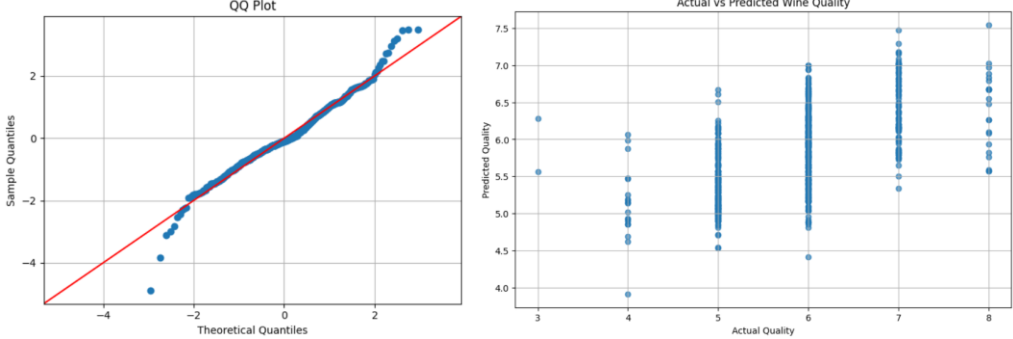
Training RMSE	0.675
Training STD RMSE	0.020
Test RMSE	0.682
Test R-squared:	0.345
Best value for the kernel: 'rbf' was C:1 and Gamma:1	
Training RMSE	0.659
Training STD RMSE	0.032
Test RMSE	0.632
Test R-squared:	0.436
<p>Evaluation:</p> <p>The results indicate that the pruned Decision Tree Regression model achieves a training RMSE of 0.675, with a stable standard deviation of 0.020 across various training sets. On the test data, the RMSE is comparable at 0.682, indicating consistent performance on unseen data. Additionally, the model's R-squared value of 0.345 signifies an improved explanatory power compared to the unpruned model, explaining approximately 34.5% of the variance in the target variable.</p>	
	
Kernels and parameter Optimization	
kernel: 'rbf' was C:10 and Gamma:10	Tr RMSE: 0.740 Std RMSE: 0.032 Test RMSE: 0.676 R-Squred: 0.357
kernel: 'rbf' was C:10 and Gamma:20	Tr RMSE: 0.747 Std RMSE: 0.032 Test RMSE: 0.682 R-Squred: 0.344
kernel: 'linear' was C:10 and Gamma:20	Tr RMSE: 0.735 Std RMSE: 0.027 Test RMSE: 0.742 R-Squred:0.225

Table 4: SVM Results Summary

How does parameter optimization in the random forest model impact its performance?

When adjusting the kernel, **C**, and **gamma** parameters in SVM regression models, changes in the model's complexity, regularization strength, and data representation occur. These adjustments influence the model's ability to capture patterns in the data, potentially altering both the RMSE (Root Mean Squared Error) and R-squared (R^2) values

3.4 How can the quality of wine be predicted using random forest, models based on the attributes described in the datasets?

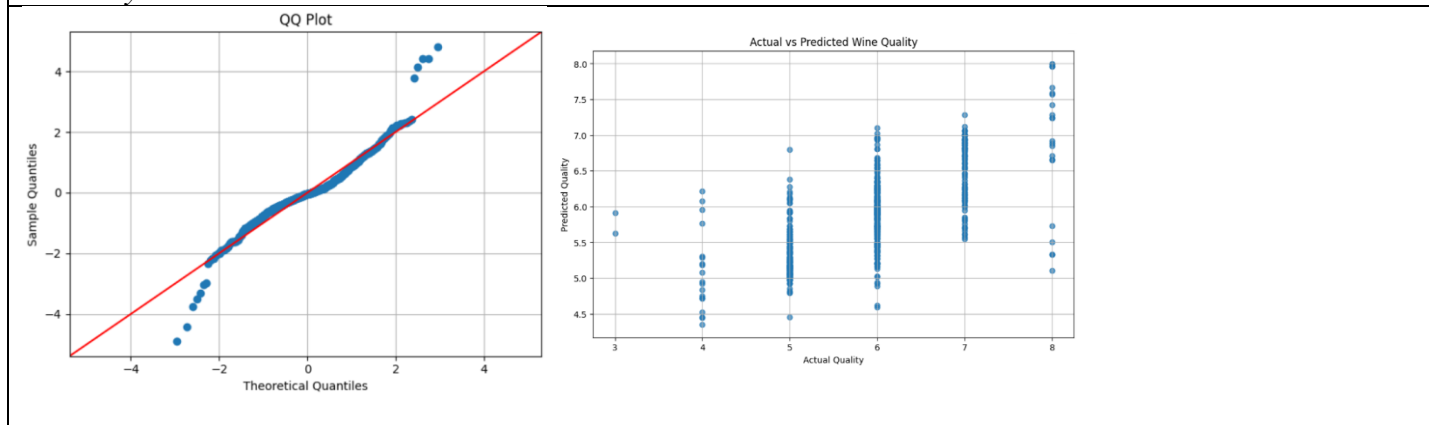
What are the results of cross-validation and performance evaluation of the models on the testing dataset?

Mode Type: Random Forest Model (Without K-Fold Validation) Target Variable: quality Predicators: All other variables including Wine type random_state: 456 n_estimators=500	
Training RMSE	0.218
Training R-Squred	0.937

Test RMSE	0.597
STD of RMSE	0
R-Squared	0.497

Evaluation: **

The results indicate that the Random Forest Regression model achieves a training RMSE of 0.218, with a very high R-squared value of 0.937, showing that the model explains 93.7% of the variance in the training data. On the test data, the RMSE is 0.597, and the standard deviation of RMSE is 0, suggesting consistent performance across different test sets. The test R-squared value of 0.497 indicates that the model explains approximately 49.7% of the variance in the target variable on unseen data, reflecting medium predictive accuracy.

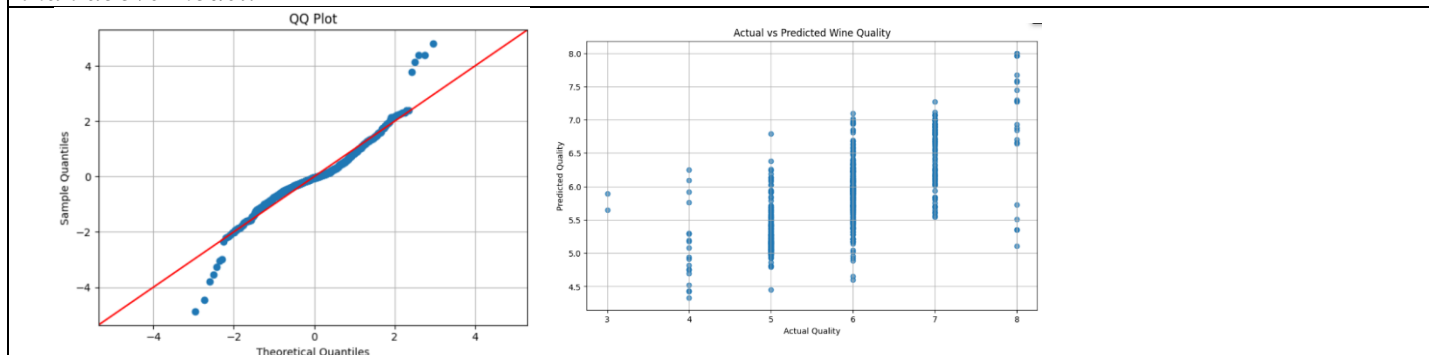


Mode Type: Random Forest Model (**With K-Fold Validation**) | Target Variable: quality | Predicators: All other variables including Wine type | random_state: 456 | n_estimators=500 | K:10

Training RMSE	0.602
Training STD RMSE	0.017
Test RMSE	0.596
R-Squared	0.499

Evaluation: ***

The results indicate that the model achieves a training RMSE of 0.602, with a consistent standard deviation of 0.017 across different training sets. On the test data, the RMSE is similar at 0.597, suggesting stable performance on unseen data. Moreover, the R-squared value of 0.499 indicates that the model explains approximately 49.9% of the variance in the target variable, demonstrating little more predictive accuracy than above model.



Mode Type: Random Forest Model (**Bagging**) | Target Variable: quality | Predicators: All other variables including Wine type | random_state: 456 | n_estimators=500 | bootstrap=True | K Value:10

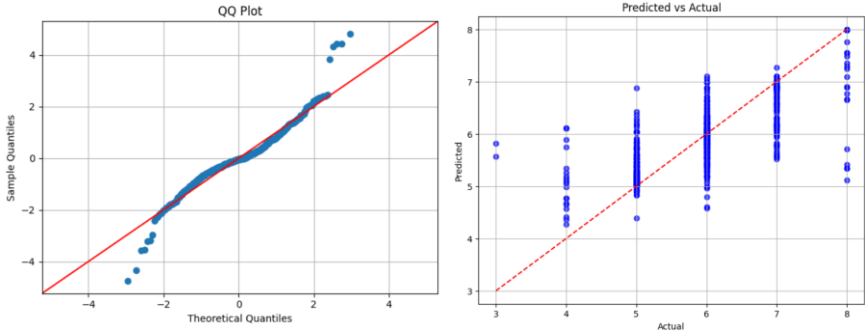
Test RMSE	0.612
R-Squared	0.471
<p>Evaluation: *</p> <p>The results indicate that the pruned Decision Tree Regression model achieves a training RMSE of 0.604, with a stable standard deviation of 0.018 across various training sets. On the test data, the RMSE slightly improves to 0.594, suggesting consistent performance on unseen data. Notably, the model's R-squared value of 0.502 signifies a substantial improvement in explanatory power compared to the unpruned model, elucidating approximately 50.2% of the variance in the target variable.</p>	
	

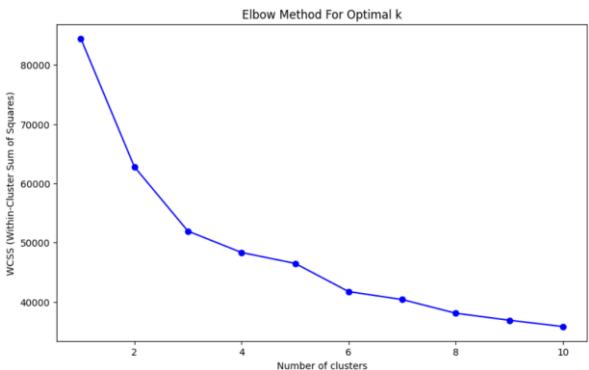
Table 5: Random Forest Results Summary

Among all the models, the Decision Tree Bagging model and the Random Forest K-Fold validated models are outperforming the others.

3.5 How many clusters are optimal based on the results of k-means and hierarchical clustering for all features?

How can the clusters be labeled

How can k-means and hierarchical clustering be applied to perform clustering on the full wine dataset?

With all the features	
<p>Standardize the inputs (Z-Score)</p> <p>K-value range – 1-11</p> <p>Elbow method for optimal K</p>	
According to the above graph best K values will be 2 or 3	

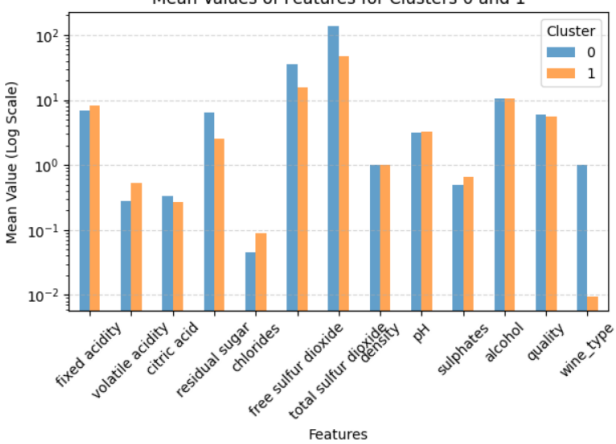
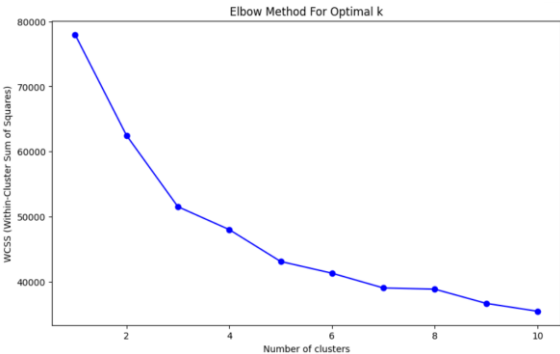
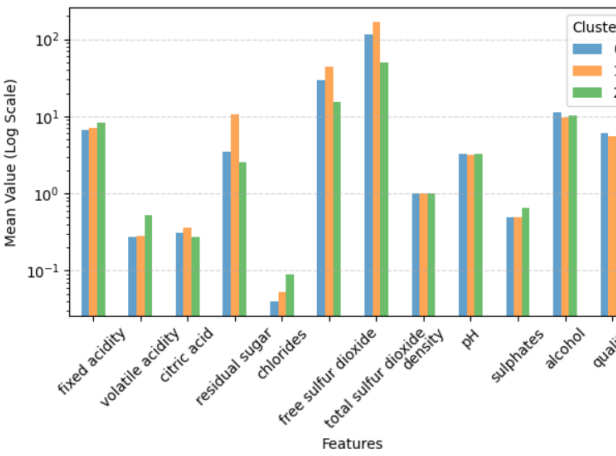
	<p>Cluster 0 Red Wine – 13 White Wine – 4883</p> <p>Cluster 1 Red Wine – 1586 White Wine – 15</p> <p>Based on the summary above, it is clear that the two clusters have been identified primarily based on wine type rather than any other features. Therefore, Cluster 0 can be named the Red Wine cluster, and Cluster 1 can be labeled the White Wine cluster</p>
After dropping the wine type	
<p>Standardize the inputs (Z-Score) K-value range – 1-11 Elbow method for optimal K</p>	
According to the above graph best K values will be 3	
	<p>According to the graph on the left, it is evident that the clusters have been identified based on <i>citric acid</i>, <i>residual sugar</i>, <i>chlorides</i>, and <i>free sulfur dioxide</i>. Therefore, we can label the clusters as follows:</p> <ul style="list-style-type: none"> • Low content of the above chemicals • Medium content of the above chemicals • High content of the above chemicals

Table 6: Clustering Results Summary

3.8 Do k-means and hierarchical clustering provide the same clustering results?

Yes, that as per the following results it has the similarity of 99.27%

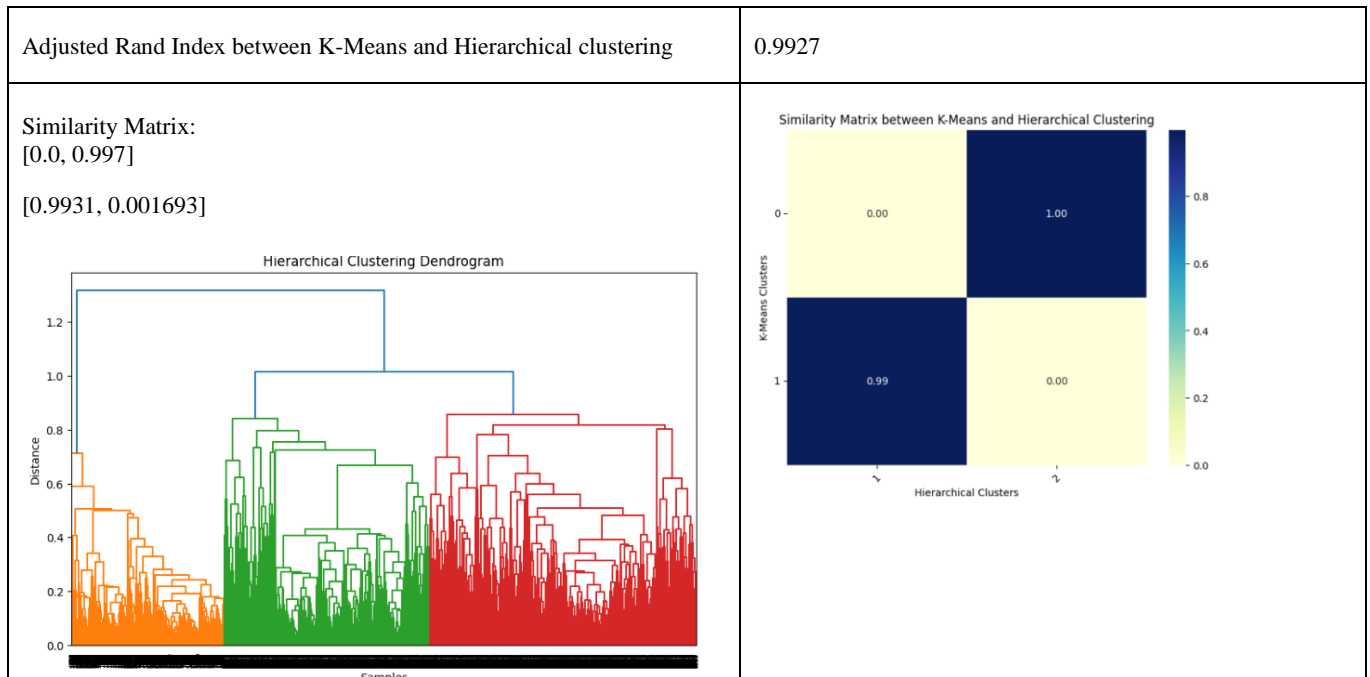


Table 7: K mean and Hierarchical clustering comparison Summary

4. Conclusion

In the study conducted, nine models were analyzed using Decision Trees, SVM, and Random Forest algorithms with various parameter optimization techniques. It was found that the Decision Tree Bagging model (RMSE-0.59 | R-Squared -0.502) and the Random Forest K-Fold validated models (RMSE-0.602 | R-Squared- 0.49) performed better than the others. The accuracy of the models was heavily influenced by the data predominantly scattering around Wine quality 5 and 6. Additionally, a high similarity of 99.27% was noted between the K-means and hierarchical clustering at K=2, demonstrating significant consistency in clustering outcomes