Predicting Wine Quality Using Ridge Regression and Random Forest Approaches

Saajan Grewal
Department of Atmospheric and Oceanic Science, University of California Los Angeles
AOS C111: Introduction to Machine Learning for Physical Sciences
Dr. Alexander Lozinski

1. Introduction

Wine quality assessment is an important aspect of the wine industry, impacting production processes and consumer satisfaction. Usually, this assessment has been performed by expert tasters through sensory evaluations, which can be subjective and inconsistent due to individual preferences. With the growth of data availability and machine learning, this provides an opportunity to develop a more objective and efficient model that predicts wine quality based on measurable physicochemical properties.

This project aims to use machine learning techniques to predict the quality of red wines by analyzing their physicochemical attributes, such as acidity, sugar content, and alcohol levels. Using the "Wine Quality" dataset from the UCI Machine Learning Repository, which includes quality ratings on a scale from 0 to 10, the study uses Ridge Regression for continuous quality score prediction and Random Forest Classification for categorizing wines into quality levels Low, Medium, and High. By identifying key factors that influence wine quality and improving prediction accuracy, the project aims to gain insights into the wine industry and support more consistent quality evaluations.

2. Data

For this project, I used the "Wine Quality" dataset available from the UCI Machine Learning Repository. The dataset contains various physicochemical properties of red wine samples, along with their quality ratings. It includes 1,599 instances of red wine from the Vinho Verde region of Portugal. Each wine sample is characterized by 11 physicochemical attributes, and a quality score rated by wine experts on a scale from 0 to 10.

The features in the dataset are as follows:

1. **Fixed Acidity**: Concentration of nonvolatile acids (mainly tartaric and malic) in g/dm^3.
2. **Volatile Acidity**: Concentration of volatile acids (mainly acetic acid) in g/dm^3.
3. **Citric Acid**: Amount of citric acid in g/dm^3, which can add freshness and flavor to wines.
4. **Residual Sugar**: Amount of sugar remaining after fermentation stops, measured in g/dm^3.
5. **Chlorides**: Concentration of salts, particularly sodium chloride, in g/dm^3.
6. **Free Sulfur Dioxide**: Amount of SO_2 not bound to other molecules, in mg/dm^3; acts as an antimicrobial and antioxidant.
7. **Total Sulfur Dioxide**: Total amount of SO_2, both free and bound forms, in mg/dm^3.
8. **Density**: Density of the wine in g/cm^3; closely related to alcohol and sugar content.
9. **pH**: Acidity or basicity of the wine on a scale from 0 (very acidic) to 14 (very basic).

10. **Sulphates**: Concentration of potassium sulphate in g/dm³, which acts as an additive contributing to sulfur dioxide levels.
11. **Alcohol**: Alcohol content of the wine in % volume.

The target variable is:

Quality: wine quality score between 0 and 10, assigned by expert wine tasters.

To prepare the dataset for modeling, I first imported it into a Pandas DataFrame using Python's pandas library. Initial data analysis involved checking for missing values, outliers, and understanding the distribution of each feature. The dataset did not contain any missing values, which allowed it to proceed without the need for imputation.

I performed exploratory data analysis to gain insights into the data:

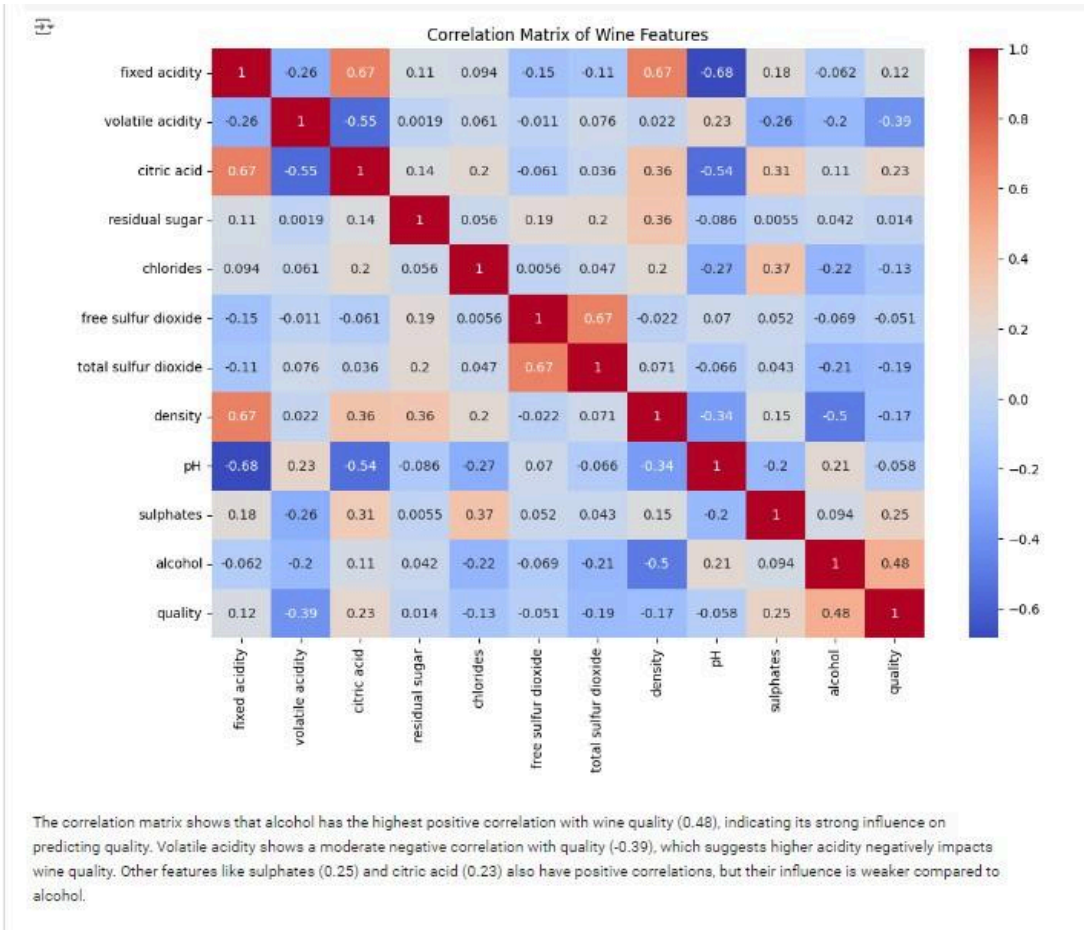
- **Statistical Summary**: Calculated mean, median, standard deviation, and range for each feature.

	fixed acidity	volatile acidity	citric acid	residual sugar	\
count	1599.000000	1599.000000	1599.000000	1599.000000	
mean	8.319637	0.527821	0.270976	2.538806	
std	1.741096	0.179060	0.194801	1.409928	
min	4.600000	0.120000	0.000000	0.900000	
25%	7.100000	0.390000	0.090000	1.900000	
50%	7.900000	0.520000	0.260000	2.200000	
75%	9.200000	0.640000	0.420000	2.600000	
max	15.900000	1.580000	1.000000	15.500000	

	chlorides	free sulfur dioxide	total sulfur dioxide	density	\
count	1599.000000	1599.000000	1599.000000	1599.000000	
mean	0.087467	15.874922	46.467792	0.996747	
std	0.047065	10.460157	32.895324	0.001887	
min	0.012000	1.000000	6.000000	0.990070	
25%	0.070000	7.000000	22.000000	0.995600	
50%	0.079000	14.000000	38.000000	0.996750	
75%	0.090000	21.000000	62.000000	0.997835	
max	0.611000	72.000000	289.000000	1.003690	

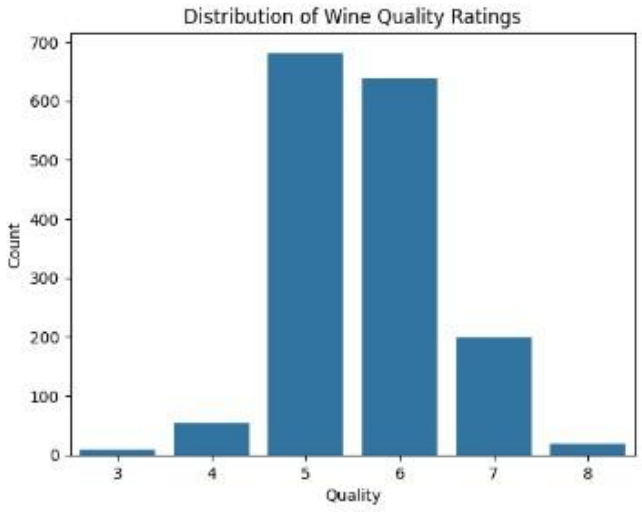
	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000
mean	3.311113	0.658149	10.422983	5.636023
std	0.154386	0.169507	1.065668	0.807569
min	2.740000	0.330000	8.400000	3.000000
25%	3.210000	0.550000	9.500000	5.000000
50%	3.310000	0.620000	10.200000	6.000000
75%	3.400000	0.730000	11.100000	6.000000
max	4.010000	2.000000	14.900000	8.000000

- **Correlation Matrix**: Analyzed the Pearson correlation coefficients between features and the target variable to identify any strong relationships.



The correlation matrix shows that alcohol has the highest positive correlation with wine quality (0.48), indicating its strong influence on predicting quality. Volatile acidity shows a moderate negative correlation with quality (-0.39), which suggests higher acidity negatively impacts wine quality. Other features like sulphates (0.25) and citric acid (0.23) also have positive correlations, but their influence is weaker compared to alcohol.

- **Distribution of Quality Scores**: Observed that the quality ratings are imbalanced, with most wines rated between 5 and 6 on the quality scale.



Given the imbalance in the quality ratings, I decided to process the target variable for the classification task, by mapping them into 3 separate categories.

- **Low Quality**: Scores of 3 and 4
- **Medium Quality**: Scores of 5 and 6
- **High Quality**: Scores of 7 and 8

This mapping helped in creating a more balanced target variable for the classification models.

Additionally, feature scaling was performed using StandardScaler to standardize the features, ensuring that each feature contributes equally to the model training. For feature selection, I employed the SelectKBest method with the ANOVA F-value as the scoring function to identify the most significant features influencing wine quality.

3. Modeling

To predict wine quality based on physicochemical attributes, I initially used Ridge Regression to predict the continuous quality scores. Seeing the limitations in handling imbalanced and nonlinear data, I then implemented a Random Forest Classifier. To further improve performance, I modified the Random Forest approach by mapping quality scores into categorical labels and applying feature selection.

Ridge Regression was chosen as the initial modeling approach for predicting the continuous wine quality scores. This linear regression technique includes a regularization term (L2 penalty) that helps mitigate multicollinearity among features by shrinking their coefficients. The correlated features like fixed acidity, citric acid, and density in our dataset made Ridge Regression a potential choice.

I prepared the data by separating the features from the target variable, which is the quality score. The dataset was split into training and testing sets using an 80/20 split to evaluate the model's generalization capability. To ensure that all features contributed equally to the model, I then standardized them using StandardScaler.

The Ridge Regression model was trained on the scaled training data. After evaluation on the test set, the model achieved a Root Mean Squared Error (RMSE) of 0.6245 and a Mean Absolute Error (MAE) of 0.5036. The R^2 score was 0.4032, indicating that approximately 40% of the variance in wine quality could be explained by the model. While this performance is moderate, it suggests that the relationship between the physicochemical properties and wine quality is not strictly linear and may benefit from more complex modeling techniques.

Considering the limitations of linear models in capturing nonlinear relationships, I then decided to use a Random Forest Classifier. Random Forest is an ensemble learning method that

constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. It is effective in handling nonlinear data and reducing overfitting through averaging.

Using the same feature set and data splitting strategy, I trained the Random Forest Classifier on the standardized features. However, there were challenges due to the imbalanced distribution of the quality scores. Most wines were rated between 5 and 6, with very few instances of extreme quality scores. This imbalance led to poor predictive performance for minority classes, as reflected in the confusion matrix and classification report. The overall accuracy of the model was 66%, and it struggled to correctly classify wines of Low or High quality.

To address the class imbalance and improve model performance, I implemented several modifications:

1. Quality Mapping: mapped the original quality scores into three categories:
    - Low Quality: Scores of 3 and 4
    - Medium Quality: Scores of 5 and 6
    - High Quality: Scores of 7 and 8
2. This mapping helped balance the class distribution and simplified the classification task.
3. Label Encoding: The categorical quality labels were encoded into numerical values using LabelEncoder to make them suitable for the machine learning algorithm.
4. Feature Selection: Applied SelectKBest with the ANOVA F-value as the scoring function to select the top six features most correlated with the target variable. The selected features were:
    - Fixed Acidity
    - Volatile Acidity
    - Citric Acid
    - Total Sulfur Dioxide
    - Sulphates
    - Alcohol
5. Model Training: With the reduced feature set, the Random Forest Classifier was retrained. The data was again split into training and testing sets with stratification to maintain class distribution, and features were standardized.

The modified Random Forest Classifier showed significant improvement. The overall accuracy increased to 88%, and the model performed well in predicting the Medium quality category, which constituted the majority of the dataset. The confusion matrix indicated better classification results, especially for the High quality wines. However, the model still struggled with the Low quality category due to the limited number of samples.

4. Results

The Ridge Regression model was used to predict the continuous quality scores of the wines. After training the model on the standardized training data, the performance was evaluated:

- Root Mean Squared Error (RMSE): 0.6245
- Mean Absolute Error (MAE): 0.5036
- Coefficient of Determination (R^2 Score): 0.4032

The RMSE of 0.6245 indicates that, on average, the predicted quality scores deviate from the actual scores by approximately 0.62 units on the quality scale. The MAE of 0.5036 suggests that the average absolute difference between the predicted and actual quality scores is about 0.50 units. An R^2 score of 0.4032 means that the model explains around 40.32% of the variance in the wine quality scores.

Interpretation:

The moderate R^2 score suggests that while the model captures some of the relationships between the physicochemical properties and wine quality, a significant portion of the variability remains unexplained. This indicates that the linear assumptions of Ridge Regression may not fully capture the complex, potentially nonlinear relationships in the data.

The Random Forest Classifier was initially applied without modifications to predict the categorical quality scores ranging from 3 to 8. The model's performance was evaluated using a classification report and accuracy score:

Classification Report:

	precision	recall	f1-score	support
3	0.00	0.00	0.00	1
4	0.00	0.00	0.00	10
5	0.71	0.74	0.72	130
6	0.63	0.70	0.66	132
7	0.64	0.55	0.59	42
8	0.00	0.00	0.00	5
accuracy			0.66	320
macro avg	0.33	0.33	0.33	320

weighted avg 0.63 0.66 0.64 320

- Accuracy Score: 0.66

Interpretation:

The overall accuracy of 66% indicates that the model correctly classified approximately two thirds of the test instances. However, the classification report reveals that the model performs poorly on wines with quality scores at the extremes (3, 4, and 8), with precision, recall, and f1-scores of 0.00. This poor performance is due to the imbalanced distribution of quality scores in the dataset, where the majority of wines are rated as 5 or 6. The model performs moderately well on the majority classes (quality scores 5 and 6) but fails to generalize to minority classes.

To address the issues observed with the initial Random Forest model, I implemented two key modifications:

1. Quality Mapping: The original quality scores were mapped into three categories to balance the class distribution:
    - Low Quality: Scores of 3 and 4
    - Medium Quality: Scores of 5 and 6
    - High Quality: Scores of 7 and 8
2. Feature Selection: selected the top six features most correlated with the target variable using the SelectKBest method with ANOVA F-value as the scoring function. The selected features are:
    - Fixed Acidity
    - Volatile Acidity
    - Citric Acid
    - Total Sulfur Dioxide
    - Sulphates
    - Alcohol

Model Evaluation:

The modified Random Forest Classifier was trained on the selected features and evaluated on the test set.

Confusion Matrix:

```
[[ 26   0  17]
 [  0   0  13]
 [  8   0 256]]
```

Classification Report:

	precision	recall	f1-score	support
High	0.76	0.60	0.68	43
Low	0.00	0.00	0.00	13
Medium	0.90	0.97	0.93	264
accuracy			0.88	320
macro avg	0.55	0.52	0.54	320
weighted avg	0.84	0.88	0.86	320

Accuracy Score: 0.88

Interpretation:

The overall accuracy improved significantly to 88%, showing that the modifications improved the model's performance. The model achieved high precision and recall for the Medium quality class, which comprises the majority of the dataset. Specifically:

- Medium Quality Wines:
    - Precision: 0.90
    - Recall: 0.97
    - F1-Score: 0.93
- The model performs well in identifying Medium quality wines, with a high true positive rate.
- High Quality Wines:
    - Precision: 0.76
    - Recall: 0.60
    - F1-Score: 0.68
- For High quality wines, the model shows moderate performance. While the precision is acceptable, the recall indicates that 60% of High quality wines were correctly identified.
- Low Quality Wines:
    - Precision: 0.00
    - Recall: 0.00

    - F1-Score: 0.00
- The model fails to correctly classify Low quality wines. This is likely due to the small number of samples (13 instances) in this category, making it challenging for the model to learn distinguishing patterns.

Confusion Matrix Analysis:

- The model misclassified some High quality wines as Medium and all Low quality wines as Medium.
- There were no instances where Medium quality wines were misclassified as Low or High.

Feature Importance:

Analyzing the feature importances from the Random Forest model revealed that the most significant predictors of wine quality are:

1. Alcohol
2. Volatile Acidity
3. Sulphates

- Ridge Regression provided a moderate prediction capability with an R^2 score of 0.4032. While it had some variance in the data, the linear nature of the model limited its performance on the complex relationships within the dataset.
- Initial Random Forest Classification struggled with class imbalance, which resulted in poor performance on minority classes and an overall accuracy of 66%.
- Modified Random Forest Classification with quality mapping and feature selection significantly improved the model's performance, achieving an accuracy of 88%. The model effectively classified Medium quality wines but still had challenges with Low quality wines due to limited data.

5. Discussion

The performance of Ridge Regression can be attributed to its linear nature. Wine quality is influenced by complex, nonlinear interactions among various physicochemical properties. Factors such as acidity, sugar content, and alcohol levels may interact in ways that are not correctly captured by a linear model. Additionally, the presence of multicollinearity among features, although mitigated by Ridge Regression's regularization, may still affect the model's ability to assign appropriate weights to each feature.

Challenges with Initial Random Forest Classification

The initial Random Forest Classifier achieved an overall accuracy of 66%. While this shows a slight improvement over Ridge Regression in terms of classification, the detailed analysis shows significant shortcomings:

- Poor Performance on Minority Classes: The model failed to correctly classify wines with quality scores of 3, 4, and 8, as evidenced by precision and recall scores of 0.00 for these classes.
- Imbalanced Dataset: The majority of the wines were rated as 5 or 6, leading the model to be biased toward these classes. This imbalance caused the model to underperform on classes with fewer instances.

Impact of Class Imbalance

Class imbalance is a common issue in classification tasks and can significantly degrade model performance on minority classes. The model tends to prioritize accuracy on the majority class, which in this case are the Medium quality wines, at the expense of minority classes. This results in poor recall for the underrepresented classes, making the model less useful for distinguishing wines of Low or High quality.

Improvements with Quality Mapping and Feature Selection

By mapping the quality scores into three broader categories (Low, Medium, and High) this aimed to address the class imbalance. Additionally, selecting the top six features most correlated with the target variable focused the model on the most relevant predictors.

Analysis of Modifications

- Quality Mapping Effectiveness: Grouping quality scores into broader categories helped balance the class distribution, allowing the model to learn more general patterns associated with each category. This approach reduced the bias toward the majority class and improved overall accuracy.
- Feature Selection Benefits: By selecting the most influential features Alcohol, Volatile Acidity, Citric Acid, Total Sulfur Dioxide, Sulphates, and Fixed Acidity, this reduces noise in the data and focused the model on variables with the strongest relationships to wine quality. This likely contributed to the improved performance, as the model could more effectively discern patterns relevant to quality prediction.

Remaining Limitations

- Underrepresentation of Low Quality Wines: The Low quality class remained underrepresented, with only 13 instances in the test set. The lack of sufficient data for this

class hindered the model's ability to learn distinguishing features, resulting in poor performance.
- Potential Overfitting to Majority Class: While the model performed exceptionally well on the Medium quality class, there is a risk of overfitting to this class due to its dominance in the dataset. This could limit the model's generalizability to new, unseen data with a different class distribution.

Feature Importance and Domain Alignment

The analysis of feature importance from the Random Forest model revealed that Alcohol was the most significant predictor of wine quality, followed by Volatile Acidity and Sulphates. This aligns with enological principles:

- Alcohol: Higher alcohol content is often associated with better quality wines, as it can enhance flavor and mouthfeel.
- Volatile Acidity: Lower levels of volatile acidity are preferable, as high levels can lead to undesirable vinegar like flavors.
- Sulphates: Appropriate levels of sulphates contribute to the preservation and antioxidant properties of wine.

Limitations and Future Work

Limitations

- Data Imbalance: Despite quality mapping, the Low quality class remained underrepresented, limiting the model's ability to accurately predict this category.
- Generalizability: The models were trained and tested on a specific dataset of Portuguese red wines. The applicability of the findings to other wine types or regions may be limited.
- Model Complexity vs. Interpretability: While Random Forest models offer strong predictive performance, they can be less interpretable than linear models. Understanding the decision making process within the ensemble can be challenging.

Future Work

- Addressing Class Imbalance: Implementing techniques such as Synthetic Minority Over sampling Technique (SMOTE) or adjusting class weights could help improve the model's performance on minority classes.
- Hyperparameter Tuning: Exploring different hyperparameters for the Random Forest Classifier, such as the number of trees, depth, and split criteria, may allow further improvements.

- Alternative Models: Testing other machine learning algorithms, such as Gradient Boosting Machines or Support Vector Machines, could provide additional insights and potentially better performance.
- Feature Engineering: Creating new features or transforming existing ones based on domain knowledge might enhance the model's ability to capture complex relationships.
- Cross Validation: Using cross validation techniques would provide a more robust evaluation of the models performance.

6. Conclusion

From this study, the following conclusions can be made:

- **Random Forest Classification with Quality Mapping and Feature Selection performed the best**, achieving an accuracy score of 88%. This approach effectively addressed class imbalance and focused on the most significant features influencing wine quality.
- **The three most important features identified were alcohol, volatile acidity, and sulfates**. These features were consistently found to have the greatest impact on the model's predictive ability within the Random Forest framework.
- **Ridge Regression demonstrated moderate performance**, with an R^2 score of 0.4032. While it captured some of the relationships between physicochemical properties and wine quality, it was limited by its linear assumptions and could not fully model the complex, nonlinear interactions present in the data.
- **The initial Random Forest Classifier without modifications struggled due to class imbalance**, resulting in poor predictive performance for minority classes and an overall accuracy of 66%.
- **Addressing class imbalance through quality mapping and focusing on key features significantly improved model performance**, highlighting the importance of data preprocessing and feature selection in machine learning tasks involving imbalanced datasets