

Wine Quality prediction

This is a model that predicts wine quality depending on the features.



Business Problem

A wine manufacturing company wants to automate its quality control process to help in reducing manual errors and increase consistency in wine grading in terms of quality. They aim to develop a predictive model that classifies wines into 'high quality' or 'low quality' sections based on provided features and chemical properties to optimize production, marketing and pricing strategies.



Data Understanding

- The data used are 2 samples from kaggle one with red and white wine types and the other with red only.
- .The dataset has 6497 rows and 13 columns.



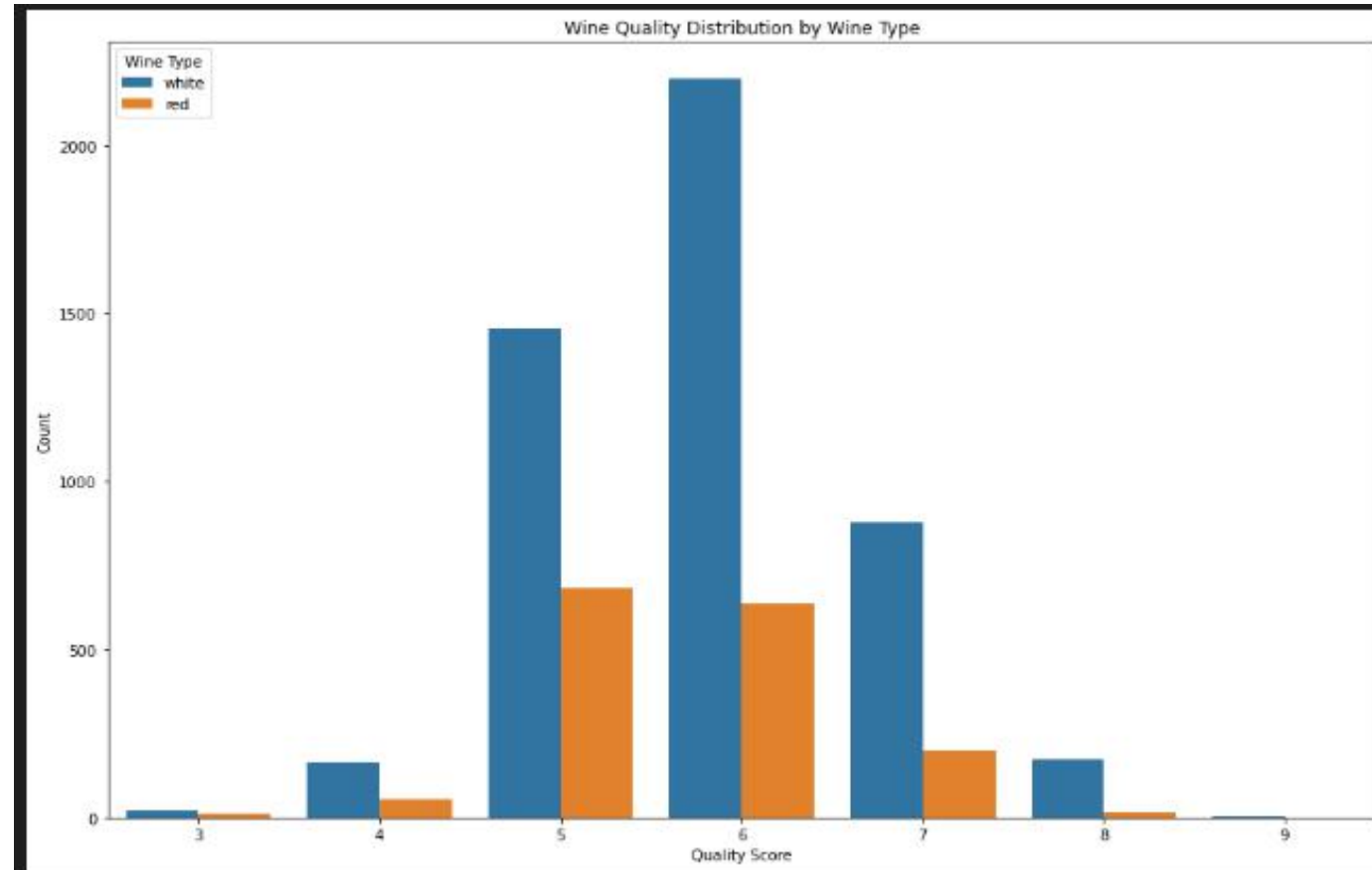
Data Preparation and processing

- Checked the data Description,shape,info.
- Checked for missing values.
- encode type to binary and column quality label binary where quality.



EDA Analysis

- The output shows that most wines are rated 5 and 6 showing that the medium quality wines are average dominated. This shows the dataset is imbalanced as most wines fall under the moderate category.
- Very few wines in low quality and high quality wines
- The Data is imbalanced so we will consider alternative solutions like Resampling and class weights

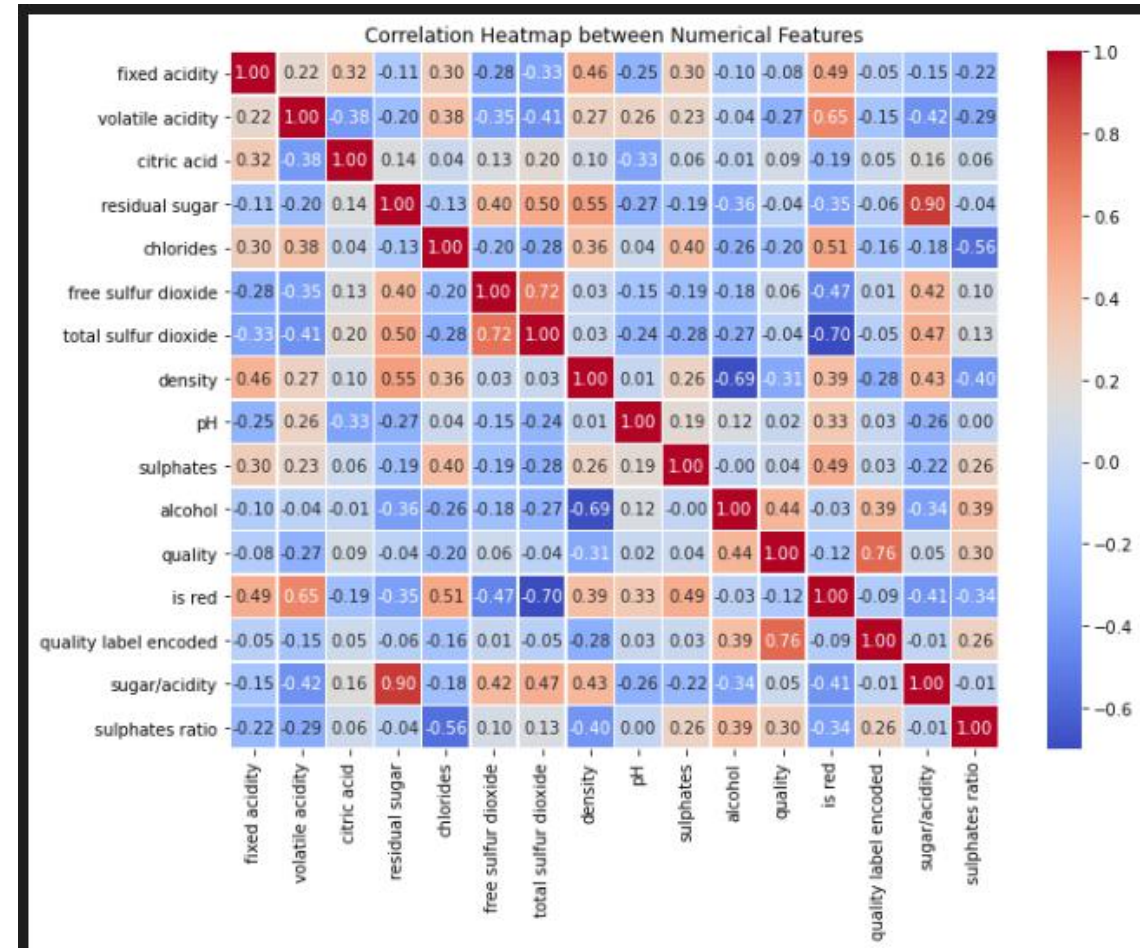


Correlation coefficients range from -1 to 1 where:

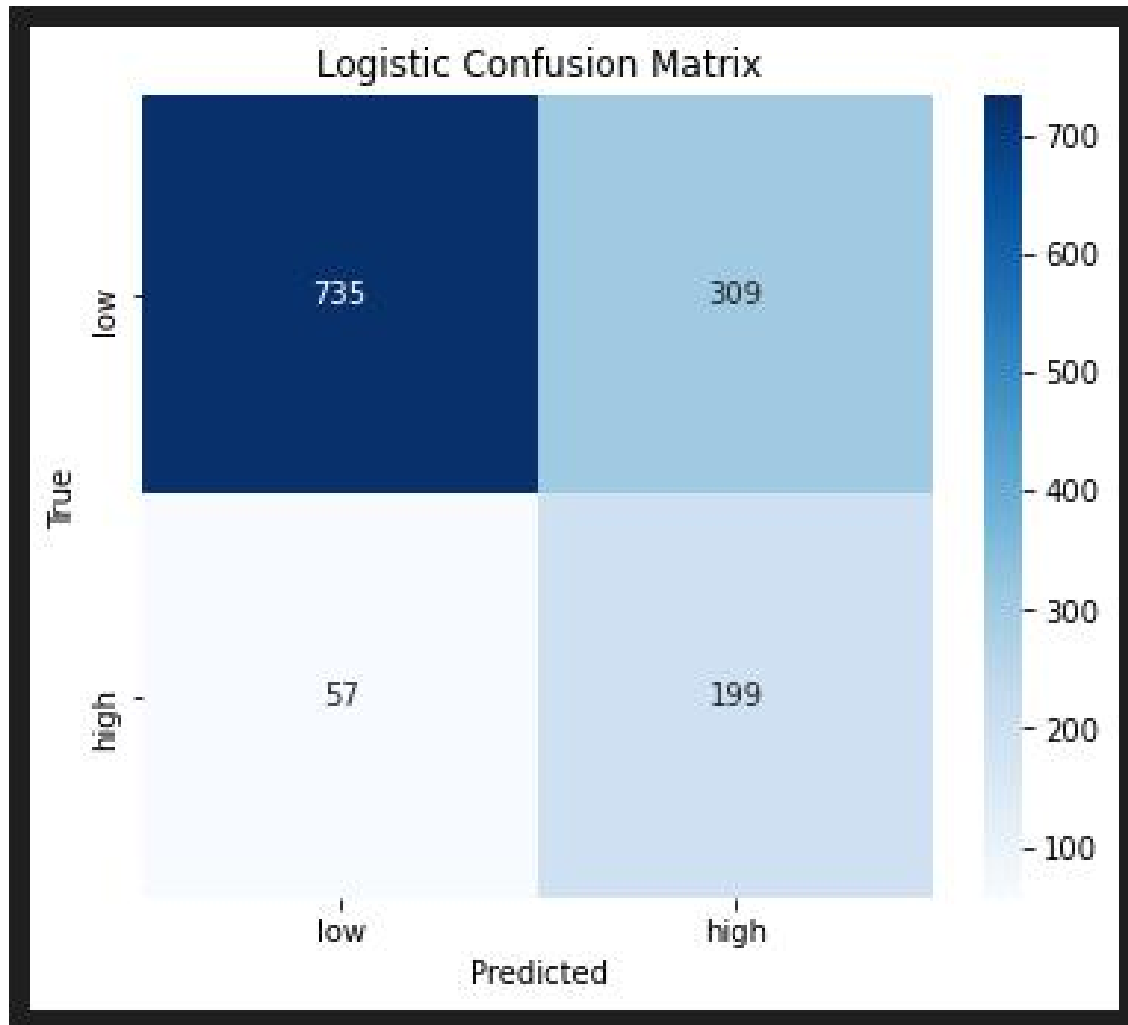
- +1 strong positive correlation(an increase in one leads to an increase in the other)
- -1 strong negative correlation(as one increases the other decreases)
- 0 There is no correlation

Our main concern is how features relate to quality.

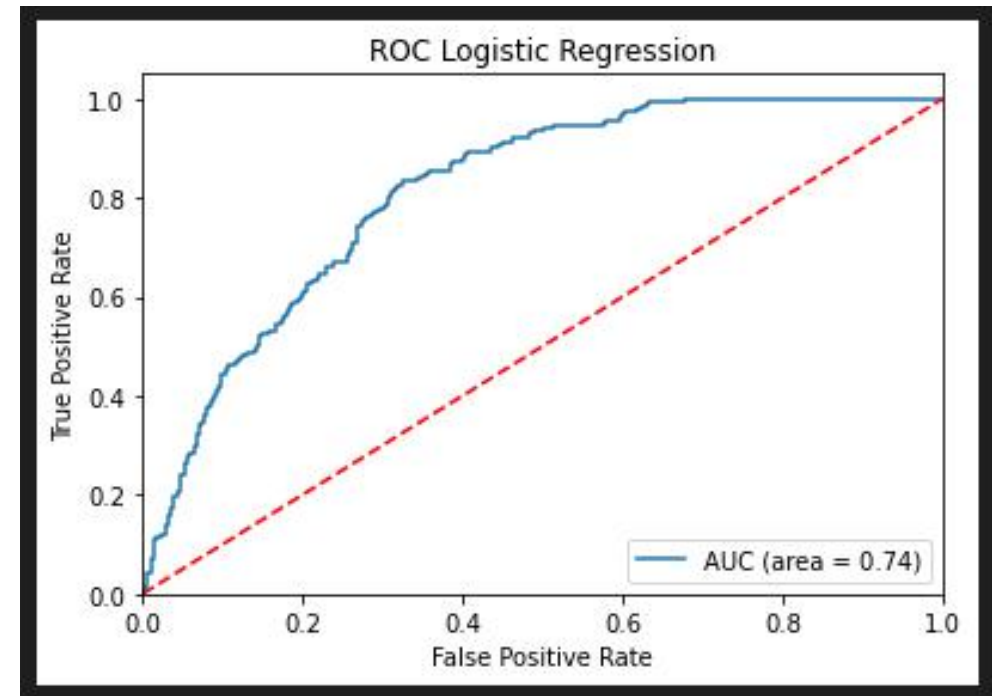
From the matrix above we conclude the features with the highest correlation to quality are alcohol with 0.44(positive correlation) and volatile acidity` with 0.27 (negative correlation)



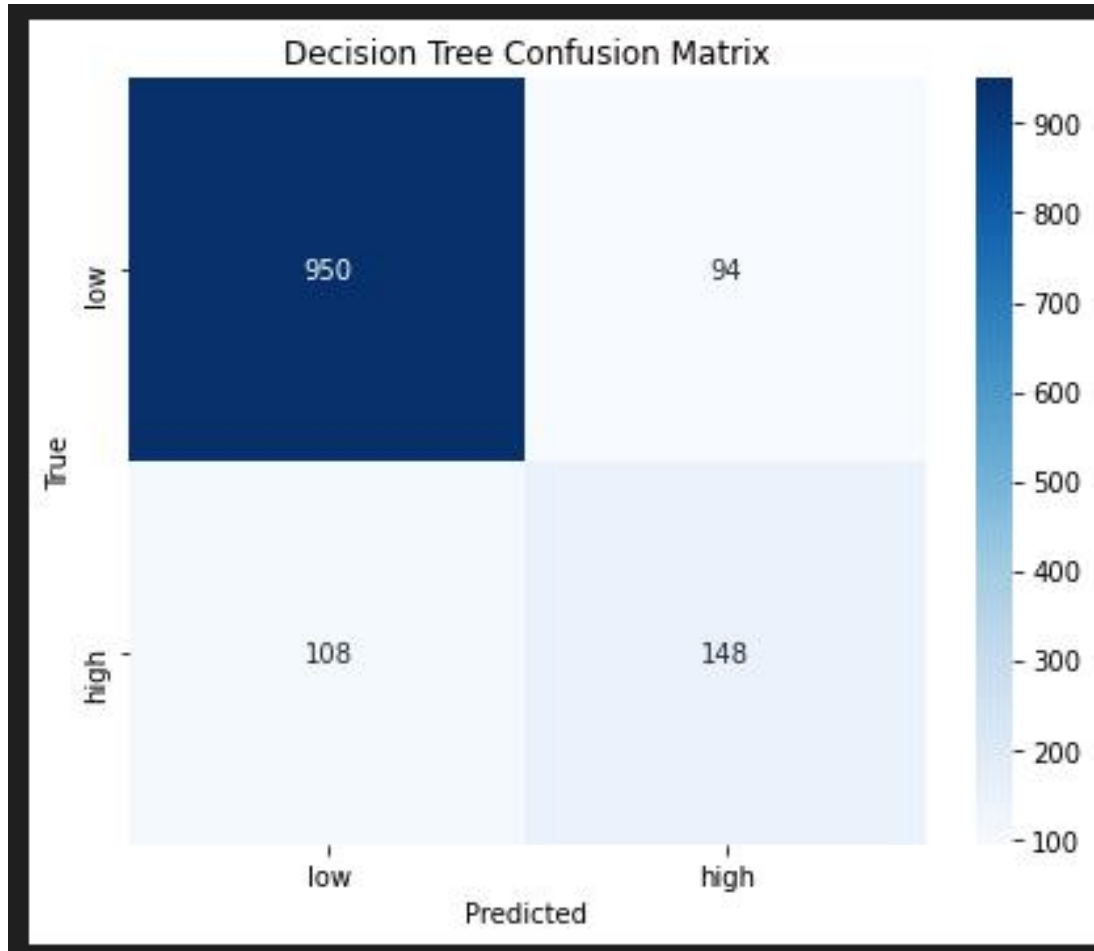
Logistic Regression



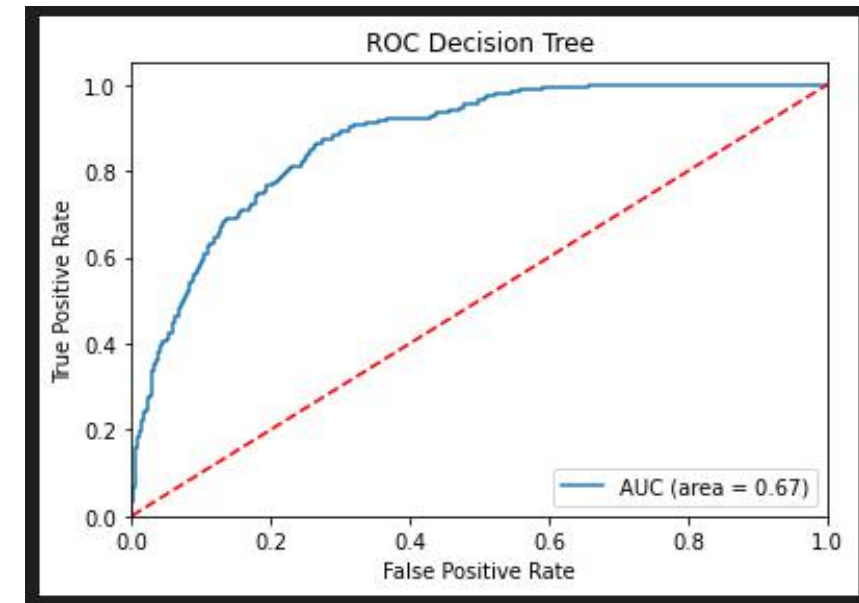
- - 735 samples were correctly classified as low.
- - 199 samples were correctly classified as high.
- - 309 low samples were wrongly predicted as high.
- - 57 high samples were wrongly predicted as low.



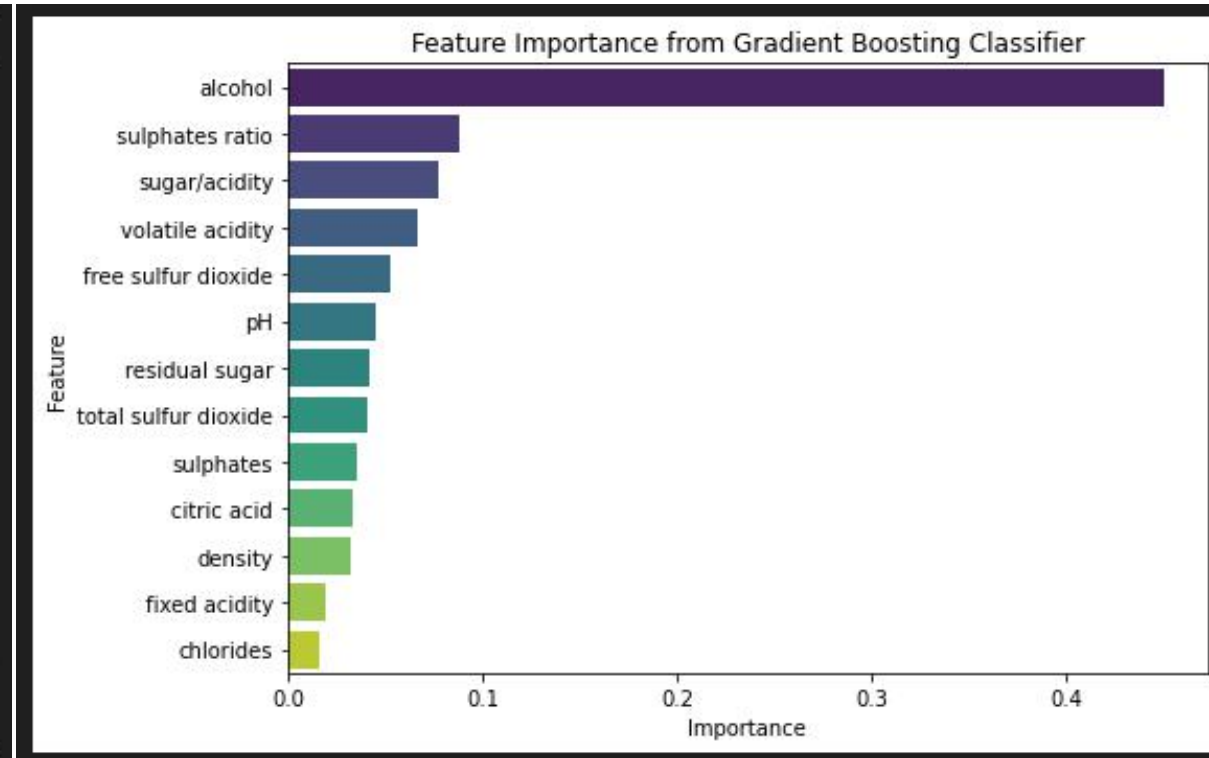
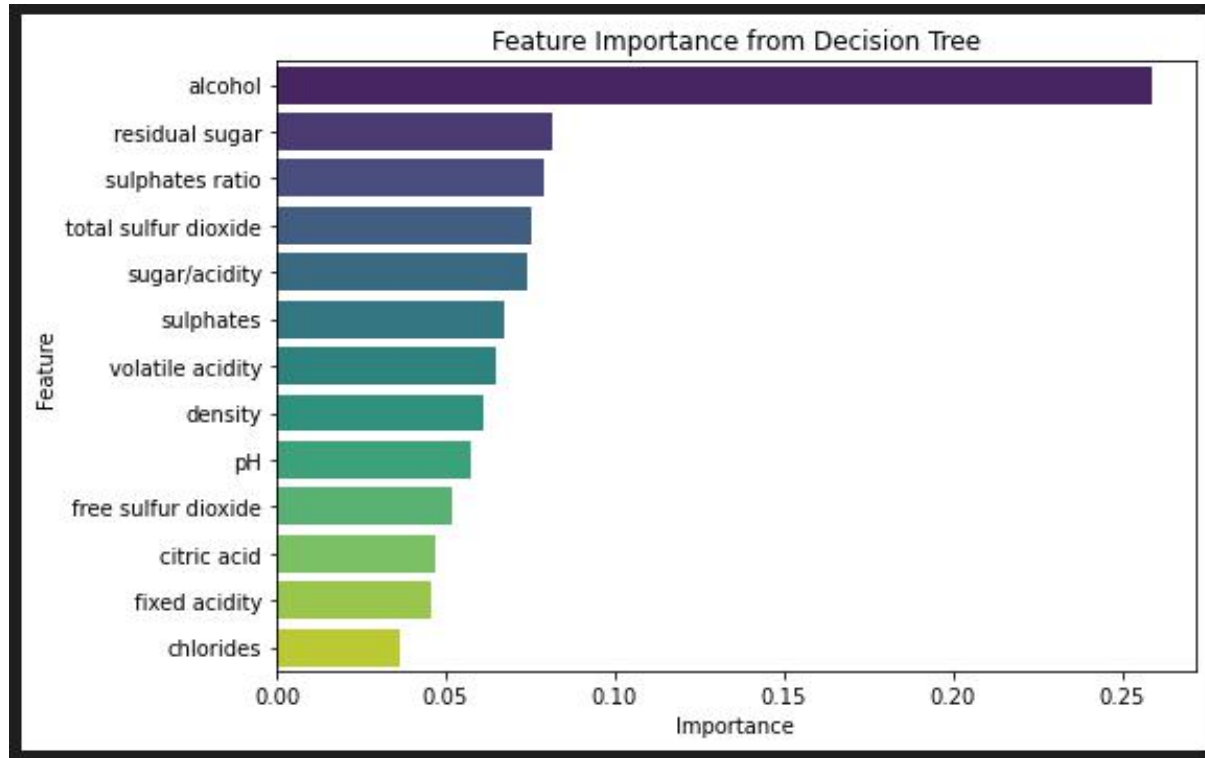
Decision Tree



- 950 samples were correctly classified as low.
- 148 samples were correctly classified as high.
- 108 low samples were wrongly predicted as high.
- 94 high samples were wrongly predicted as low.



Feature Importance



The two images above indicate to us that `alcohol` is the feature that affects the data quality most



Model Comparison

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic	72%	0.39	0.78	0.52	0.72
Decision Tree	84%	0.61	0.58	0.59	0.74
GBM	85%	0.70	0.39	0.50	0.67

From this table we can conclude that the Decision tree provided the best model accross all metrcs.



Recommendation

- 1. Perform external validations using other datasets
- 2. One can enhance quality labels by adding a medium class.
- 3. Reclass or reweight to handle the class imbalance correctly



Limitation

- 1. Dataset Imbalance on Fewer higher quality wines and more low quality wines create a class imbalance.
- 2. No external test data was used to test the model



Conclusion

- 1. Alcohol and volatile acidity play a crucial role in predicting wine quality. This shows that they play a great role in the quality of the data.
- 2. Decision tree provide the most balanced classification.
- 3. Wine type should be included as it has moderate effect on quality.



Questions

Any Questions

