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 is a samples from kaggle red and white wine types.
- .The dataset has 6497 rows and 13 columns.
- The data has no null values.

Data Preparation and processing

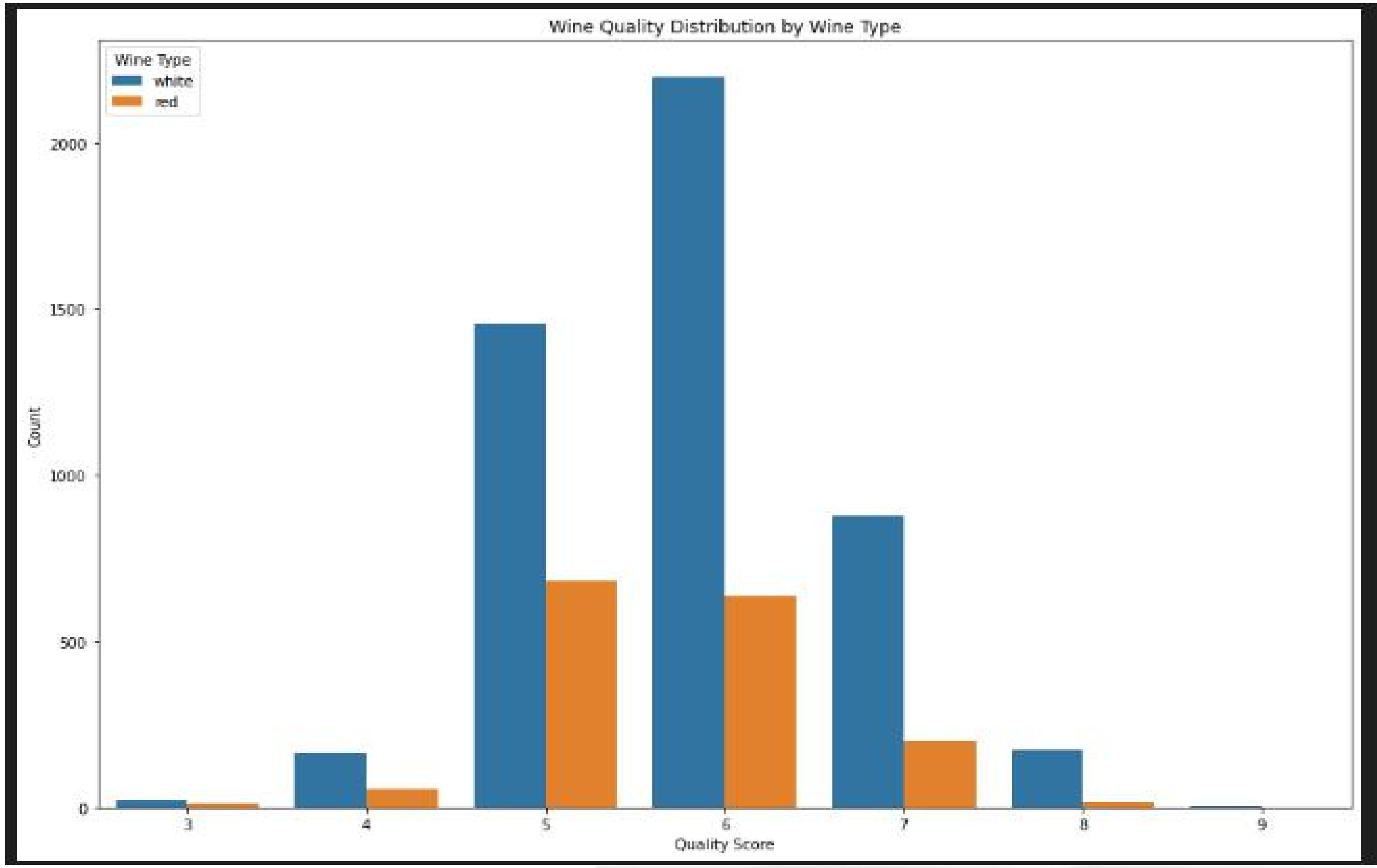
- Checked the data Description,shape,info.
- Checked for missing values.
- encode type to binary and column quality label binary where quality.

EDA Analysis

- The output shows that most wines are rated 5 and 6 shownning that the medium quality wines ast average dominate.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

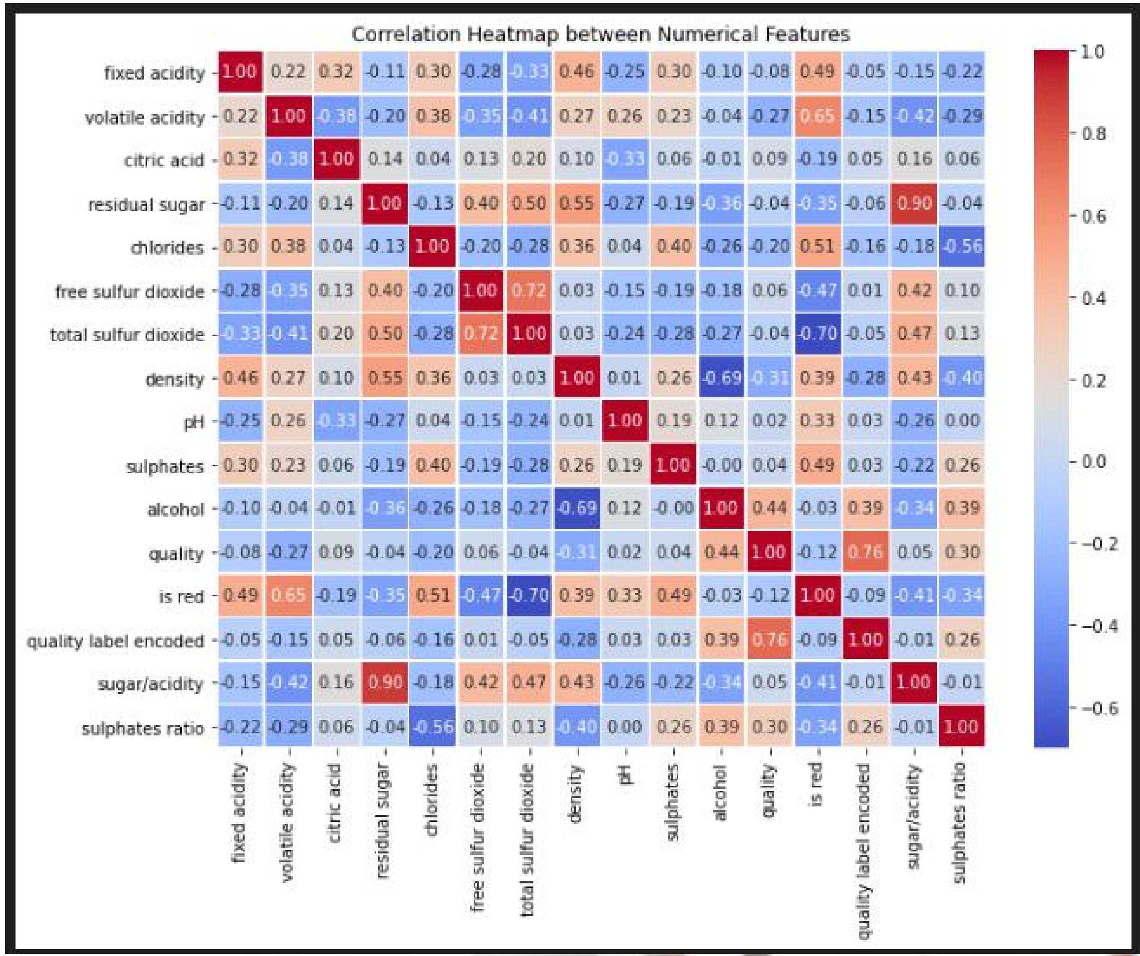


Correlation coeffficients range from -1 to 1 where:

- +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.

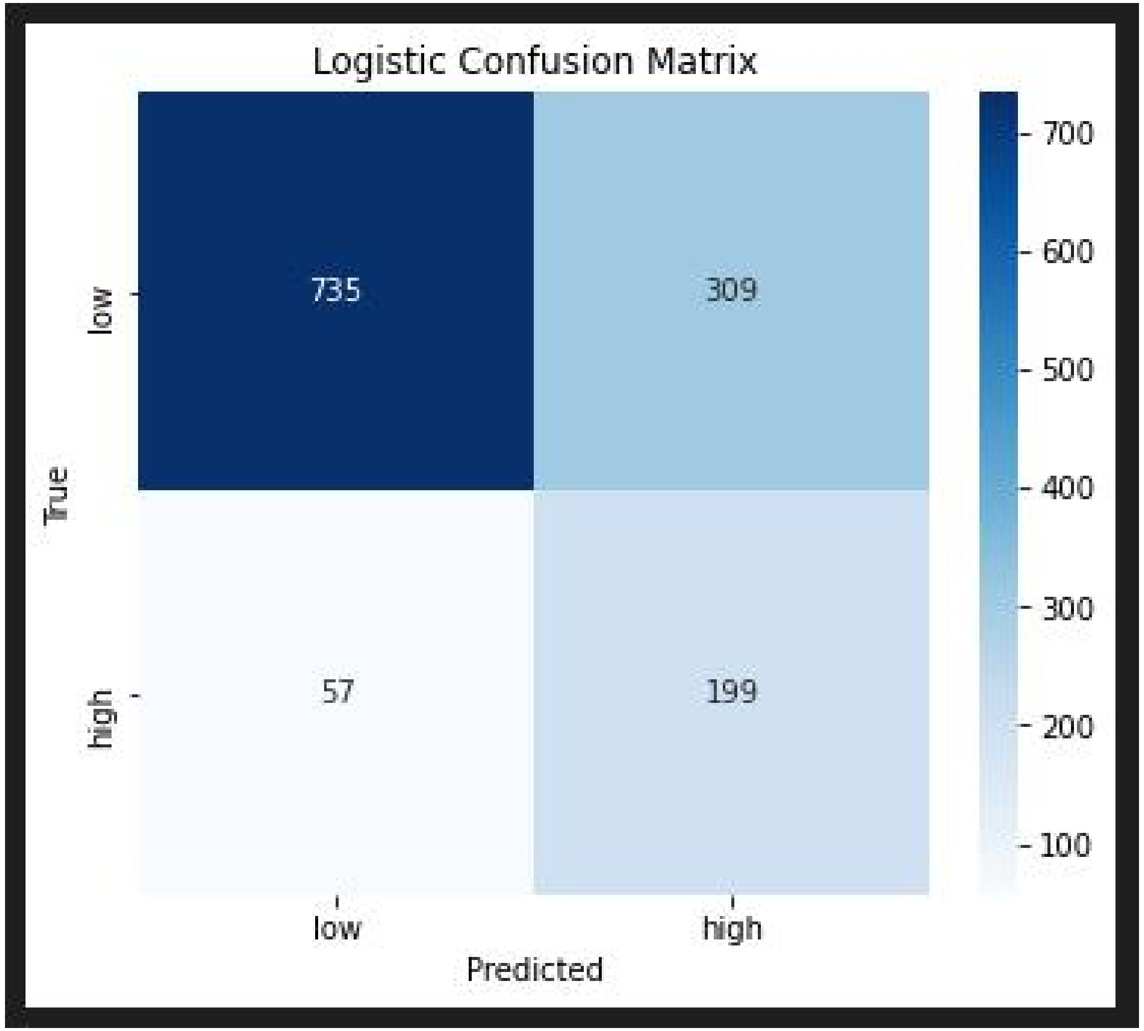
From the matrix above we conclude the features with the highest correlation to qualitya are alcohol with 0.44(positive correlation) and volatile acidity` with 0.27 (negative correlation)



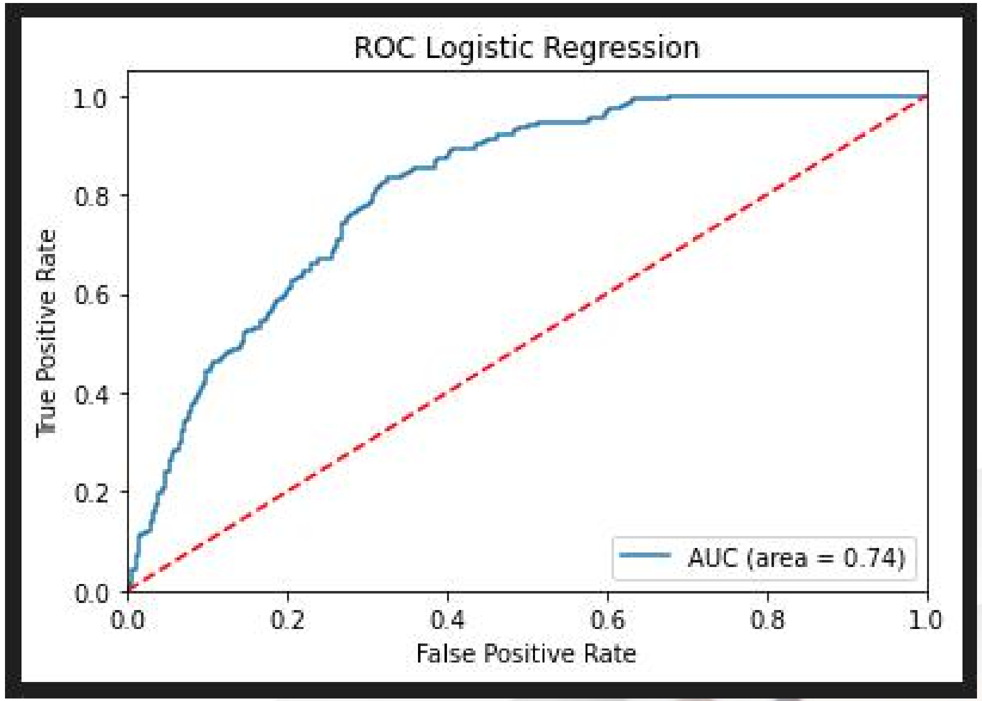
Correlation Heatmap between Numerical Features

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	is red	quality label encoded	sugar/acidity	sulphates ratio
fixed acidity	1.00	0.22	0.32	-0.11	0.30	-0.28	-0.33	0.46	-0.25	0.30	-0.10	-0.08	0.49	-0.05	-0.15	-0.22
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	-0.15	-0.42	-0.29
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	0.05	0.16	0.06
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	-0.06	0.90	-0.04
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	-0.16	-0.18	-0.56
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	0.01	0.42	0.10
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	-0.70	-0.05	0.47	0.13
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	-0.28	0.43	-0.40
pH	-0.25	0.26	-0.33	-0.27	0.04	-0.15	-0.24	0.01	1.00	0.19	0.12	0.02	0.33	0.03	-0.26	0.00
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	0.03	-0.22	0.26
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	0.39	-0.34	0.39
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	0.76	0.05	0.30
is red	0.49	0.65	-0.19	-0.35	0.51	-0.47	-0.70	0.39	0.33	0.49	-0.03	-0.12	1.00	-0.09	-0.41	-0.34
quality label encoded	-0.05	-0.15	0.05	-0.06	-0.16	0.01	-0.05	-0.28	0.03	0.03	0.39	0.76	-0.09	1.00	-0.01	0.26
sugar/acidity	-0.15	-0.42	0.16	0.90	-0.18	0.42	0.47	0.43	-0.26	-0.22	-0.34	0.05	-0.41	-0.01	1.00	-0.01
sulphates ratio	-0.22	-0.29	0.06	-0.04	-0.56	0.10	0.13	-0.40	0.00	0.26	0.39	0.30	-0.34	0.26	-0.01	1.00

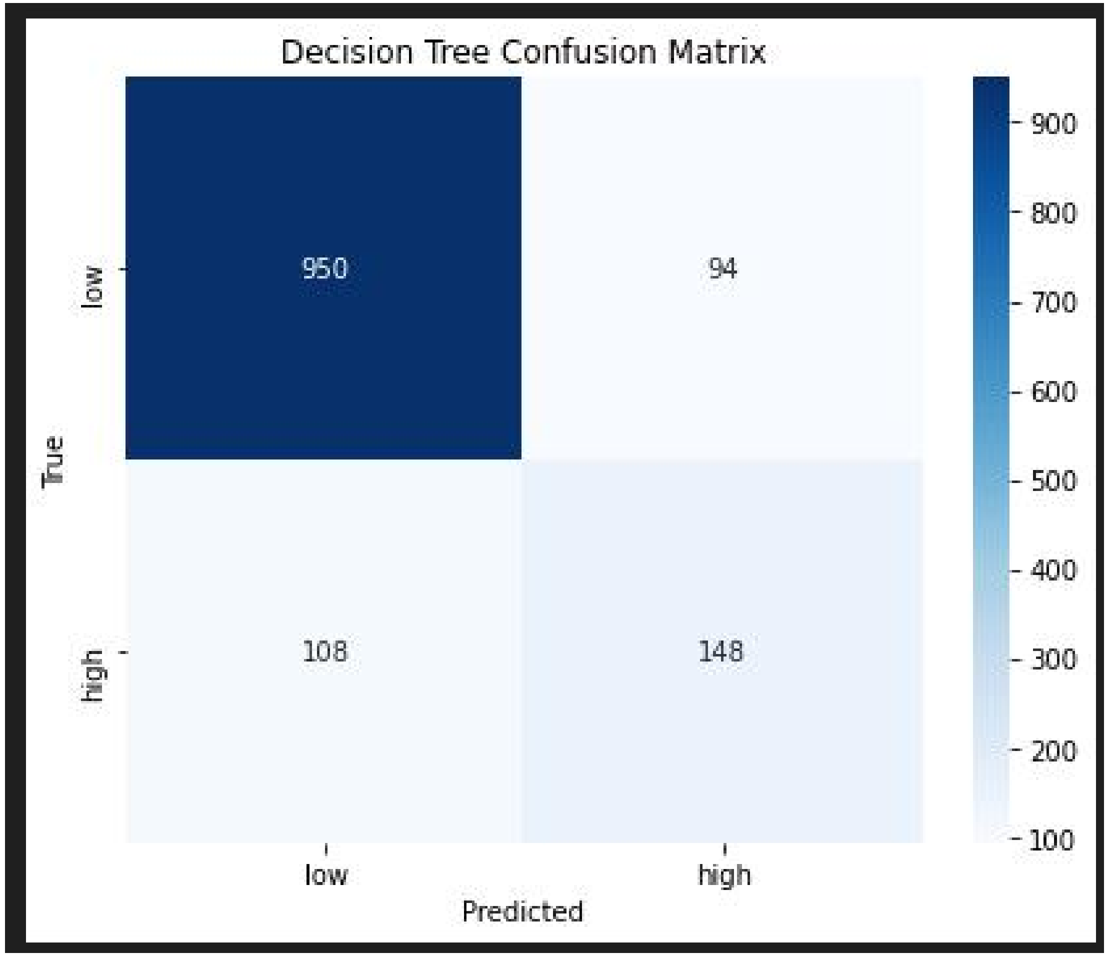
Logistic Regression



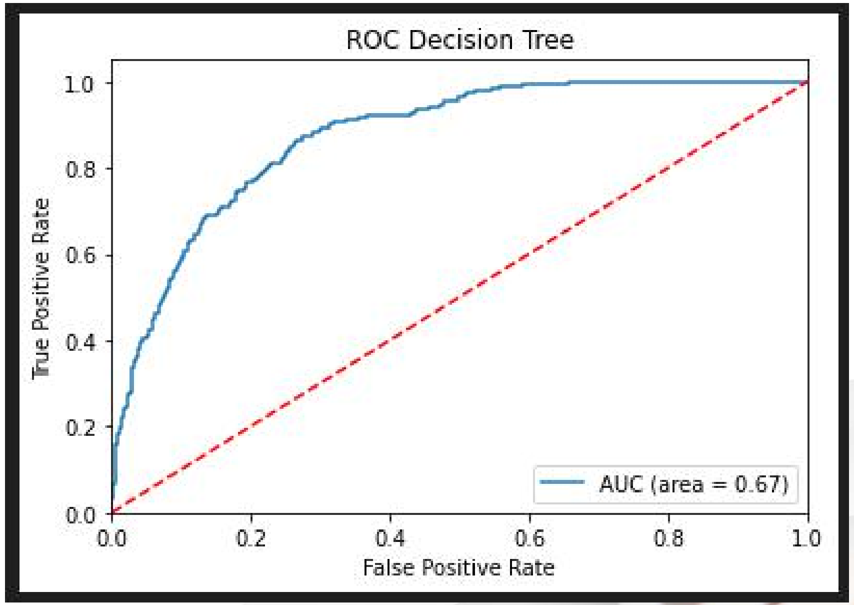
- - 735 samples were correctly classified as low.
- - 199 samples were correctly classified as high.
- - 309 low samples were wrongly predicted as high.
- - 57 high samples were wrongly predicted as low.



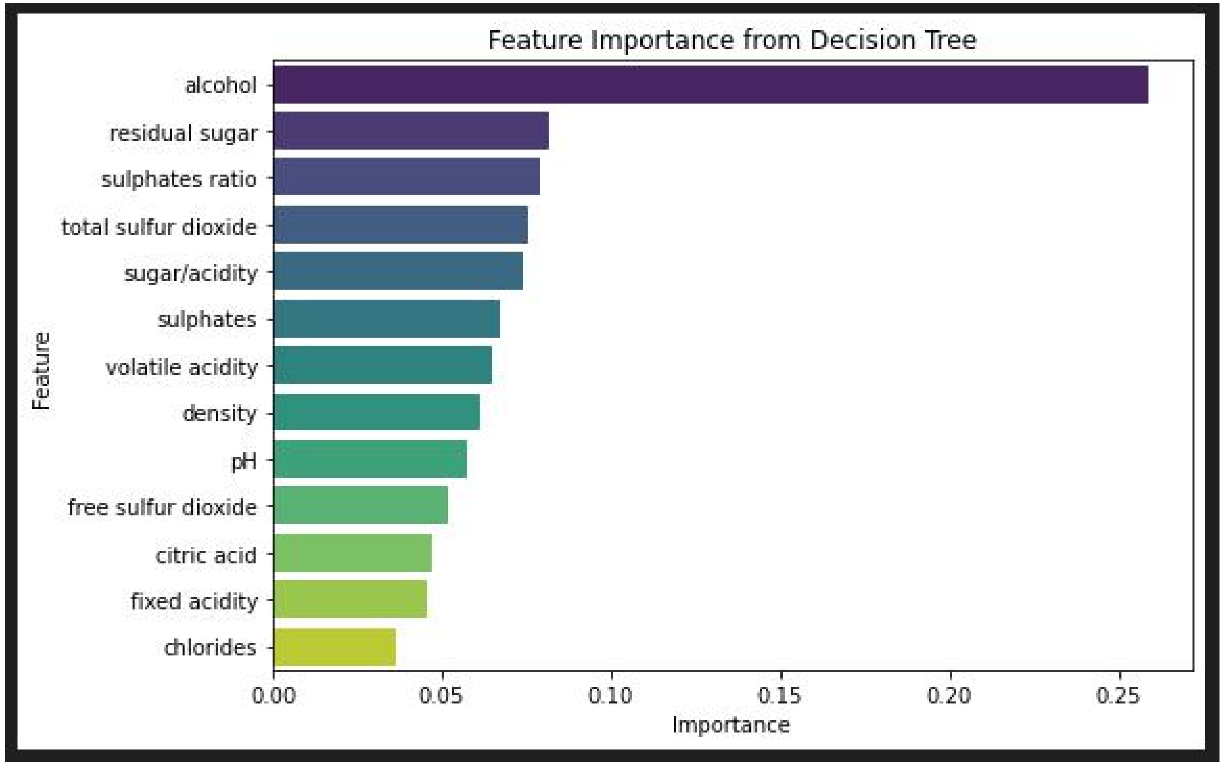
Decision Tree

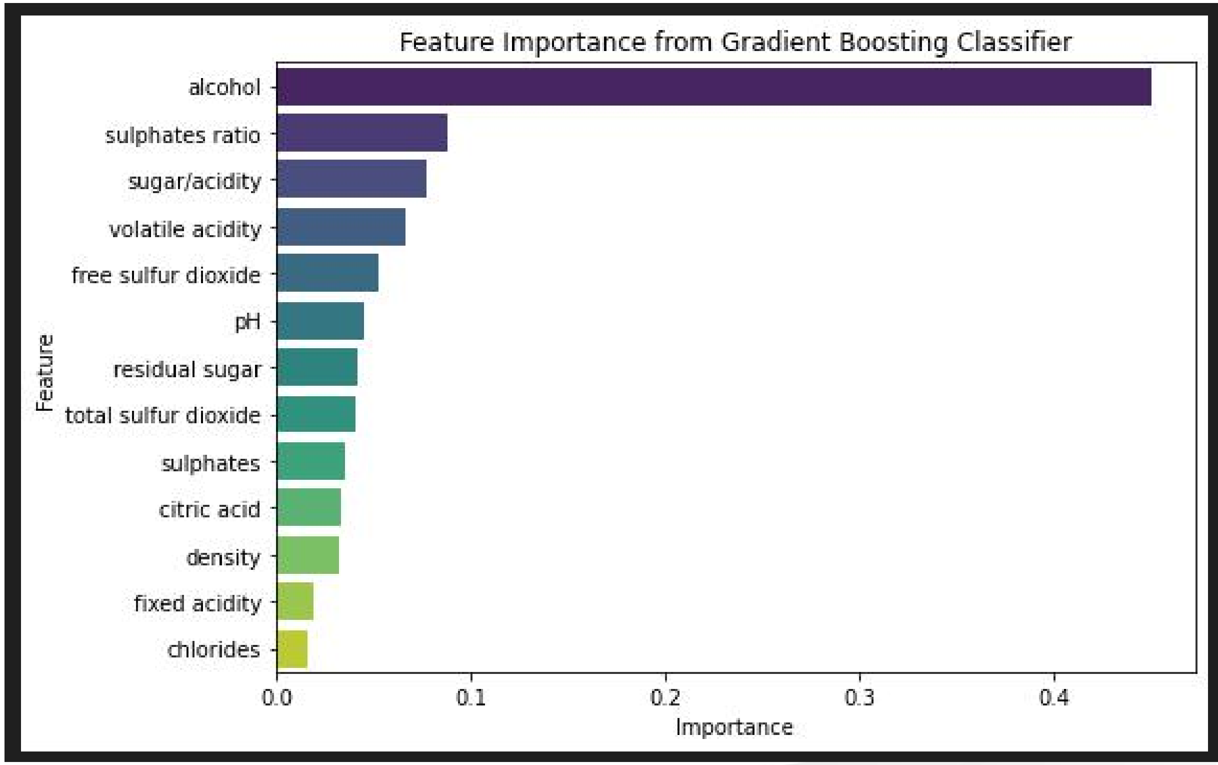


- 950 samples were correctly classified as low.
- 148 samples were correctly classified as high.
- 108 low samples were wrongly predicted as high.
- 94 high samples were wrongly predicted as low.



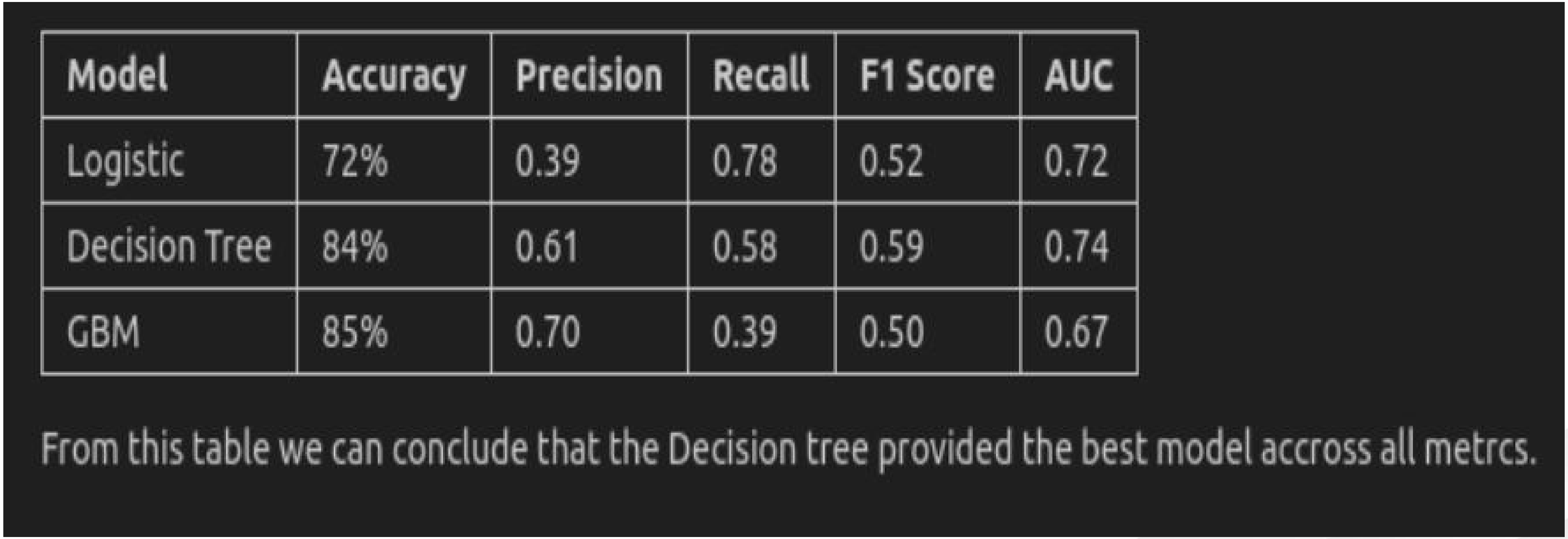
Feature Importance





The two images above indicate to us that `alcohol` is the feature tha 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 lables by adding a medium class.
- 3. Reclass or reweight to handle the class imbalance orrectly

Limitation

- 1. Dataset Imbalance on Fewer higher qality 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

?