

**Table of Contents**

[1.1 Project # 2](#_heading=)

[1.2 Project Description 2](#_heading=h.30j0zll)

[1.3 Date Submitted 2](#_heading=h.1fob9te)

[1.4 Project Priority 2](#_heading=h.3znysh7)

[1.5 Step 1. Project Deliverables 2](#_heading=h.2et92p0)

[1.6 Step 2. List of Project Tasks 2](#_heading=h.tyjcwt)

[1.7 Step 3. Out of Scope 2](#_heading=h.3dy6vkm)

[1.8 Step 4. Project Assumptions 2](#_heading=h.1t3h5sf)

[1.9 Step 5. Project Constraints 3](#_heading=h.17dp8vu)

[1.10 Step 6. Updated Estimates 3](#_heading=h.3rdcrjn)

[1.11 Step7. Approvals 3](#_heading=h.26in1rg)

[2 Introduction 4](#_heading=h.lnxbz9)

[3 Analysis of the dataset and Trained Model 4](#_heading=h.35nkun2)

[3.1 Exploratory Analysis and Visualization 4](#_heading=h.1ksv4uv)

[3.2 Baseline Model 7](#_heading=h.44sinio)

[4 Model Selection 8](#_heading=h.2jxsxqh)

[4.1 Wine Quality Prediction 8](#_heading=h.7e3dxqnztw6p)

[4.2 Model Performance Evaluation 9](#_heading=)

[4.3 Appendix 11](#_heading=h.1s0l00shfz36)

| Project # | Project Description | Date Submitted | Project Priority |
| --- | --- | --- | --- |
| 1 | Jiazu Wines is a local grocery store that proudly sources its wine from Mengjia Farms in the DC metropolitan region. They would like to predict the quality of their wines before releasing their latest red wine selection. We are committed to building a machine learning application that their managers can use as a reference for deciding the quality of red wine to increase sales. | Oct 20th, 2023 | Priority 01 |

## Step 1. Project Deliverables

Please list *all project deliverables* listed in the Project Charter and, if necessary, elaborate on them. *Do not list dates*. Add more rows as necessary.

| **Deliverable ID#** | **Description** |
| --- | --- |
| 1 | Project Charter |
| 2 | Ensemble trained model with preliminary results of the training and testing. |
| 3 | Flask and Heroku application. (customer will choose whether to deploy it locally or on the cloud) |
| 4 | Final report of the project in addition to the cloud (heroku) deployed link. |

## Step 2. List of Project Tasks

Please list ***all project tasks*** to be completed, based on the “Deliverables” specified in the Project Charter. *Do not list dates*. Add more rows as necessary. Optional: you may substitute a work breakdown structure (WBS) or mind-map in lieu of Step 2. Please attach WBS or mind-map to document.

| **Task ID#** | **Task to be completed** | **Delivery Date** | **For Deliverable #** |
| --- | --- | --- | --- |
| 1 | Submit Project Charter | 09/16/2023 | 1 |
| 2 | Ensemble trained model with preliminary results of the training and testing. | 10/20/2023 | 2 |
| 3 | Flask and Heroku application. (customer will choose whether to deploy it locally or on the cloud) | 11/02/2023 | 3 |
| 4 | Final report of the project in addition to the cloud (heroku) deployed link. | 11/15/2023\* | 4 |

## Step 3. Out of Scope

| This project **will NOT accomplish or include** the following: | This project will not include any API analysis, non-generative methods would be also considered out of scope as well as predicting the quality using features that are not present in the dataset such as customer preferences, brand names, etc. |
| --- | --- |

## Step 4. Project Assumptions

Please list any project factors that will be considered to be true, real, or certain. Assumptions generally involve a certain degree of risk.

| **#** | **Assumption** |
| --- | --- |
| 1 | On the Kaggle site, it is mentioned that the quality of the wine ranges from 0-10, however, in the dataset, the range is between 3-8. |
| 2 | Another assumption would be whether the quality of red wine in the dataset represents the industry standards of red wine. |
| 3 | Was the dataset collected in a short/long period of time? As wine quality tends to be associated with its age, the longer the dataset was collected, the more better our models would be in predicting higher/lower quality of wine. |

## Step 5. Project Constraints

| Project Start Date | 08/25/2023 |
| --- | --- |
| Launch/Go-Live Date | 09/13/2023 |
| Project End Date | 11/16/2023 |
| List any hard deadline(s) | 09/16/2023  10/20/2023  11/02/2023  11/15/2023\* |
| List other dates/descriptions of key milestones | None |
| Budget constraints Enter information about project budget limitations | N/A |
| Quality or performance constraints Enter any other requirements for the functionality, performance, or quality of the project | Software should load in a reasonable time and the system must provide 99.9% uptime. |
| Equipment/personnel Constraints Enter any constraints regarding equipment or people that will impact the project | N/A |
| Regulatory constraints  Enter any legal, policy or other regulatory constraints | N/A |

## Step 6. Updated Estimates

| Estimate T&C hours required to complete project | N/A | If charge-back project, list total estimated T&C cost | N/A |
| --- | --- | --- | --- |

## Step7. Approvals

| Required For Project Class… | Role of Approver | Submitted for Approval on: | Approval Received on: |
| --- | --- | --- | --- |
| All classes | 1. Client + Client Supervisor | Mengjia Wei | 8/25/2023 |
| All classes | 2. T&C Supervising Manager | Jiazu Zhang | 8/25/2023 |
| Class 3 + 4 only | 4. VP for Technology & Communication | Ahmed Khair | 8/25/2023 |
| Class 3 + 4 only | 5. Project Review Board | Dr. Nakul Padalkar | 8/25/2023 |

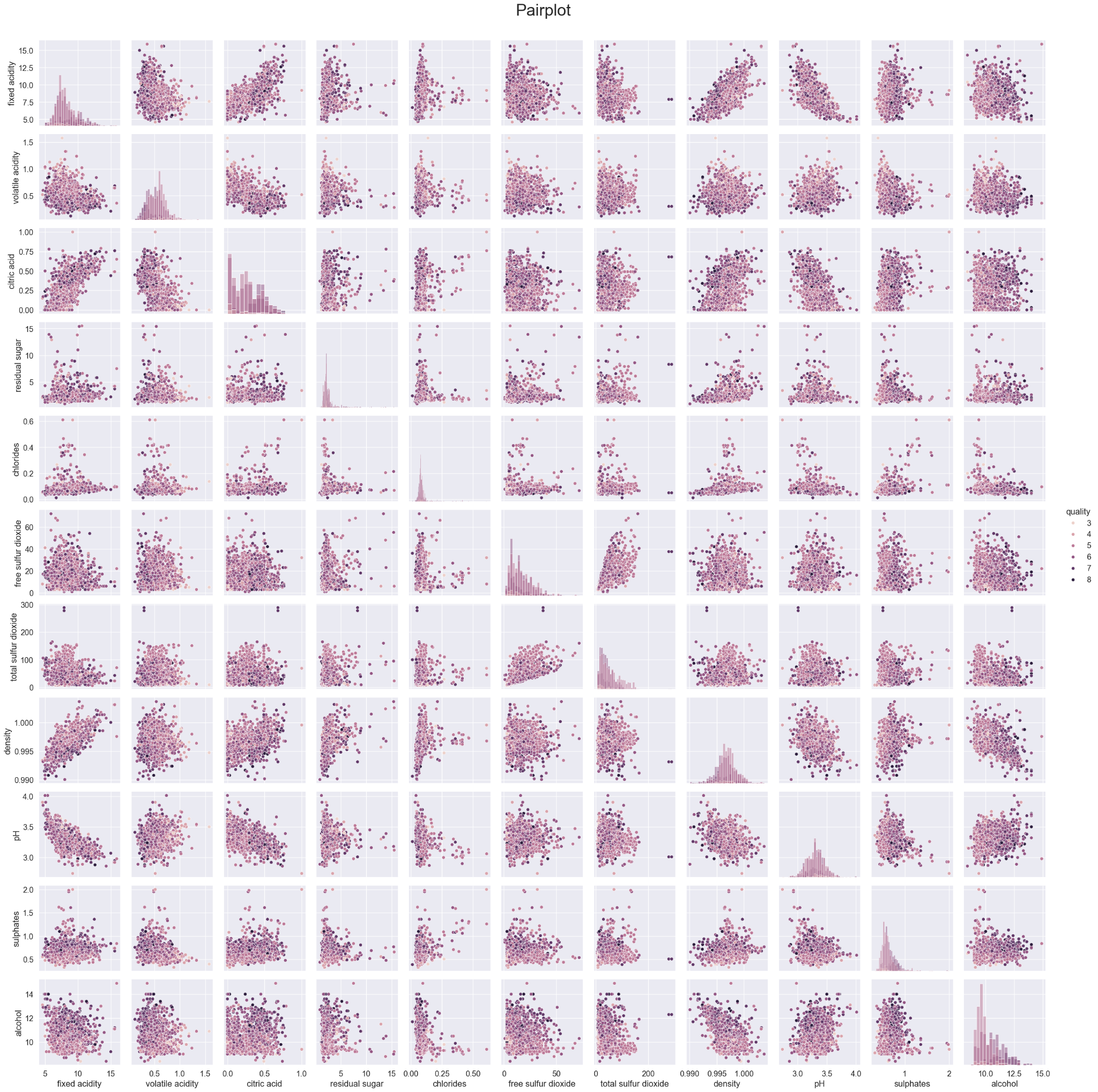
# Introduction

Jiazu Wines is a premium local grocery in the Washington-DC metropolitan region. Mengjia Wei, President and CEO of this grocery, requested to build a machine learning application to predict the quality of red wine prices. Since the outcome will be used as a reference for deciding the quality of wine, we have decided to train our machine learning model on a dataset specific to red wine. The dataset contains details of red wine sold in the Washington-DC metropolitan area. It contains information about the fixed acidity of red wine; the volatile which is the amount of acetic acid in wine; citric acid which is found in small quantities; residual sugar is the amount of sugar after the fermentation stops; chlorides; free sulfur dioxide; the density of water is close to that of water (density); pH to describe how acidic or alkaline a wine is from a scale of 0 to 14 (0 being very acidic and 14 being very alkaline; sulphates; quality of the wine (from 0 to 10).

# Analysis of the Dataset and Trained Model

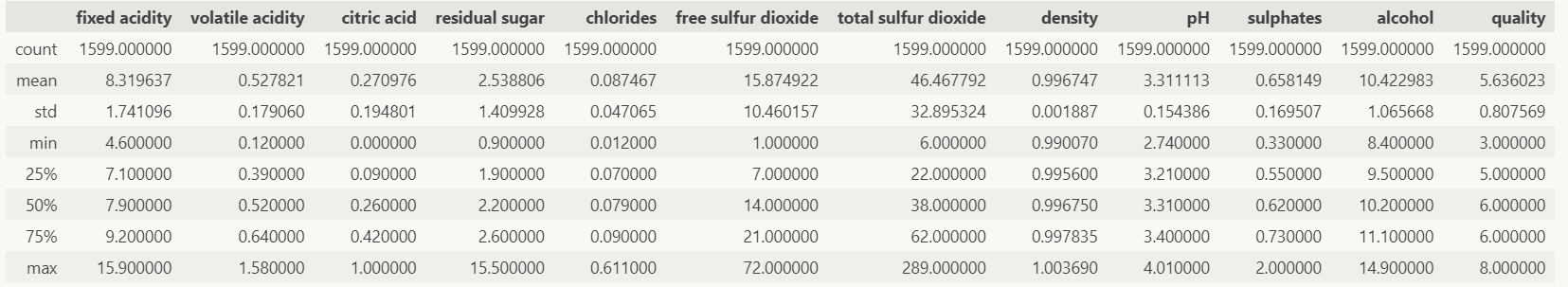
## Exploratory Analysis and Visualization

We start with a pair plot of all variables to examine the distributions of wine features and the general trend of quality as the change of wine feature values. The graph shows that some features have strong positive correlations such as “PH” & “Alcohol” and “Citric Acid” & “Fixed Acidity”. Some features have strong negative correlation such as “PH” & “Fixed Acidity” and “Alcohol” & “Density”. It also shows correlations between wine quality and features. For example, high-quality wines are more frequently distributed around wines with low “volatile acidity”, high “citric acid”, high “Sulphates” and high “Alcohol”, while the other features don’t show significant patterns related to wine quality. Next, we will understand the correlation strengths and directions between wine quality and wine features through a correlation analysis.

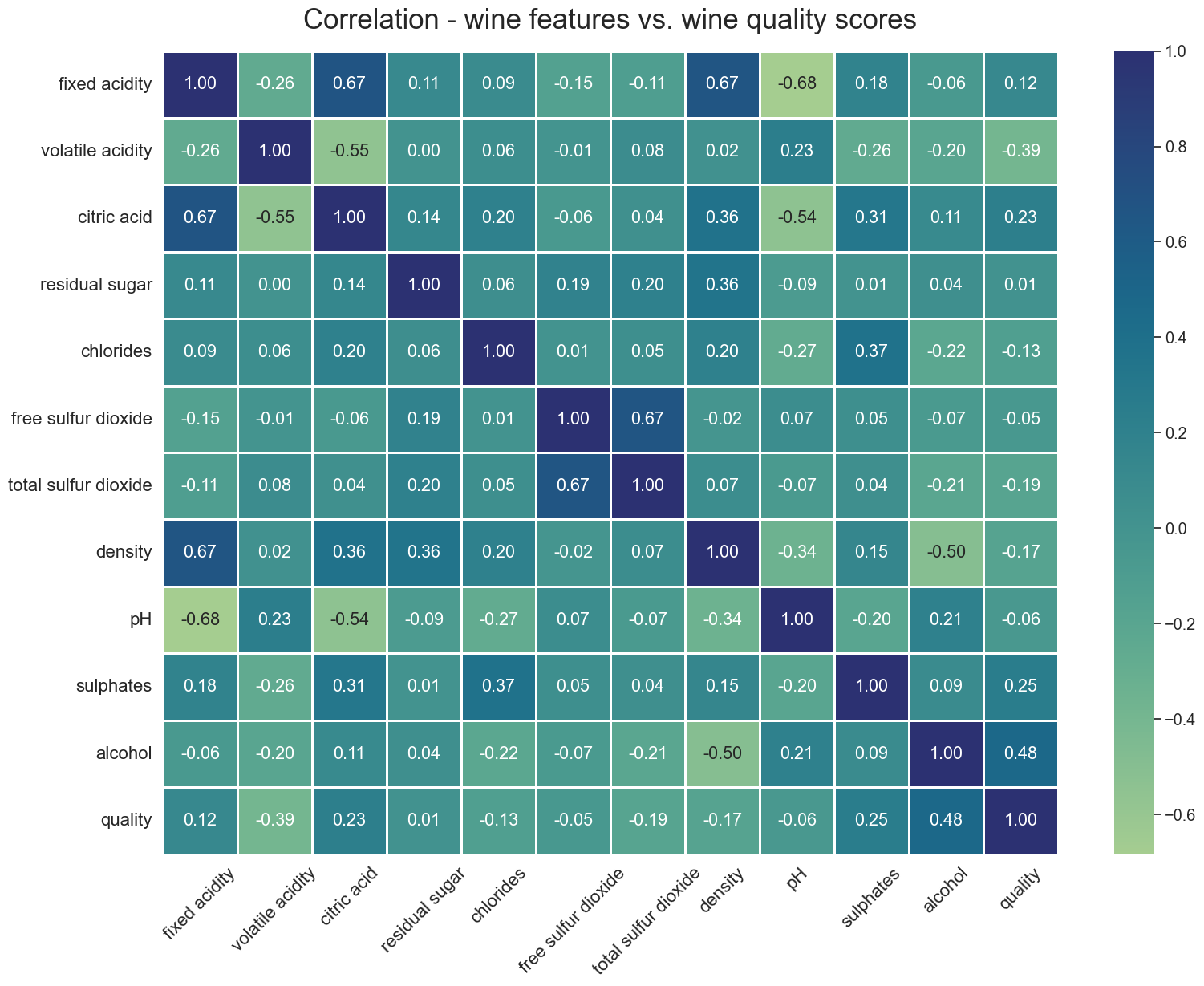
****

*Figure1 Pairplot of all variables*

Figure1 shows that the high quality wines are more frequently distributed around wines with low “volatile acidity”, high “citric acid”, high “Sulphates” and high “Alcohol”.



*Table 1 Summary of the Variables*

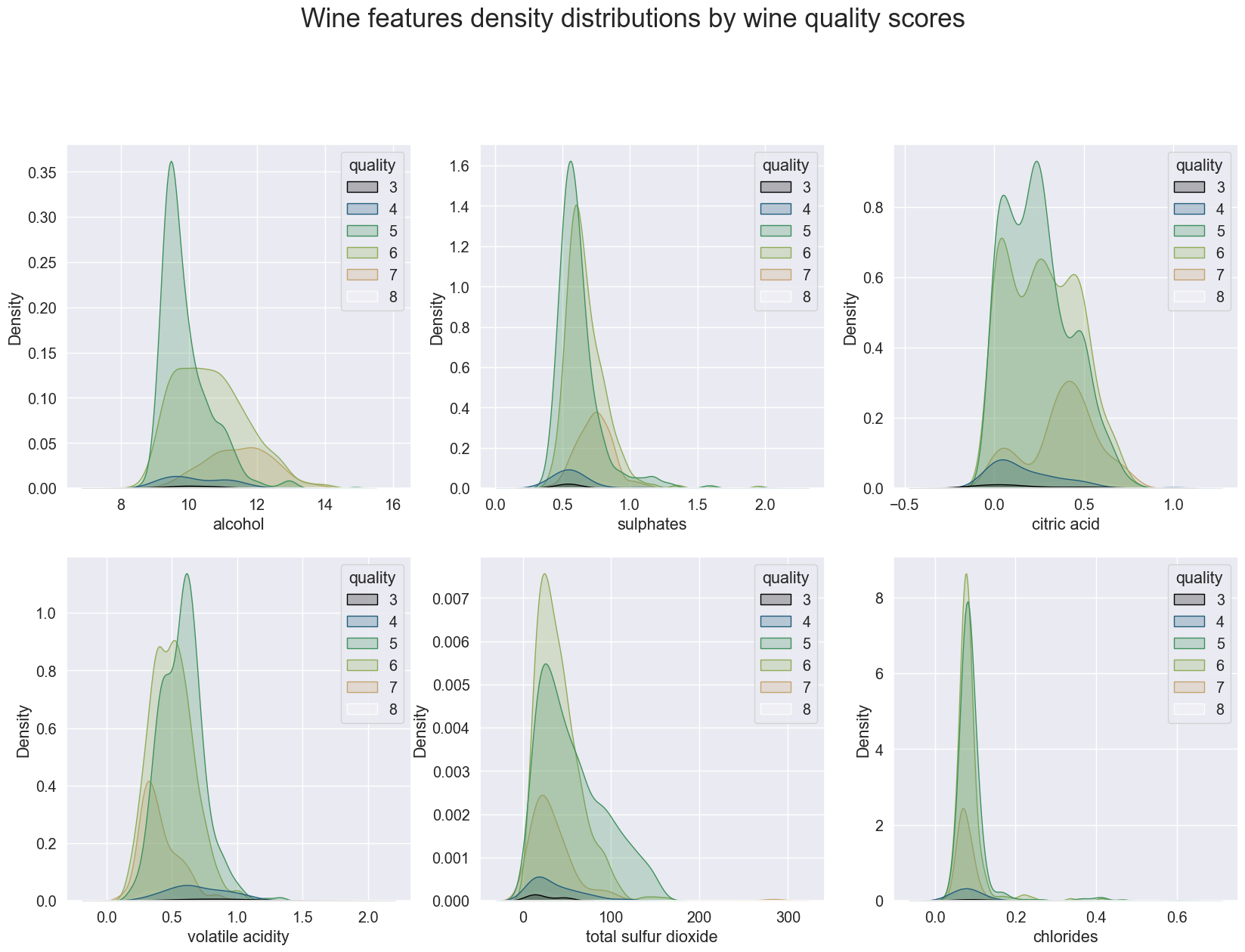
**

*Figure2 Correlation heatmap*

Figure2 shows that wine quality has a positive correlation with levels of “alcohol” (0.48), “sulphates” (0.25), and “citric acid” in the wine.

The correlation heat map shows that wine quality is significantly positively correlated with “alcohol” (0.48), “sulphates” (0.25), and “citric acid” (0.23). Wine quality shows negative correlations with “volatile acidity” (-0.39). This means higher levels of alcohol, sulphates, and citric acid increases wine quality, while high wine quality comes with low volatile acidity. “Total sulfur dioxide” and “chlorides” also show a weaker negative correlation with wine quality.

Now that we find features that are most related to the wine quality, the third part of EDA is to further understand the feature value distributions of different quality wines. We draw six KDE plots (figure3) for each of the six features that are highly correlated with wine quality discussed above, and grouped by wine quality (3-8). The KDE plots show that a). Wines with a quality of 5 are densely distributed around an alcohol level of 9-10, while wines with a quality of 7 are densely distributed around an alcohol level of 11-13. b). Wines with a quality of 4 are densely distributed around a citric acid level of 0.0, while wines with a quality of 7 are densely distributed around a citric acid level of 0.5. c). Wines with a quality of 5 are densely distributed around a volatile acidity of 0.7, while wines with a quality of 7 are densely distributed around a volatile acidity of 0.3. The other three features also show different distributions on different wine qualities but with lower variances. The KDE plots and correlation analysis provide us with some insight into the most important wine features in terms of wine quality prediction.



*Figure3 KED plots of wine features distributions*

*Figure 3 shows that wines with high qualities have alcohol, sulphates, and citric acid levels that are densely distributed around lower values, while they have volatile acidity, total sulfur dioxide and chlorides that are more densely distributed around higher values.*

## Baseline Model

Since this is the logistic regression dataset, we have decided to analyze the model using multi-class classification. We will also assume this to be the baseline model and improve the performance by using multiple ensemble techniques. For simplicity, all the model code is provided with an attached jupyter notebook, not in the report. Additionally, a baseline model was created which has an accuracy score of 60%. In our dataset, we have 6 unique qualities (3,4,5,6,7,8), and as a result, we have 6 sets of coefficients which are shown in table 2.

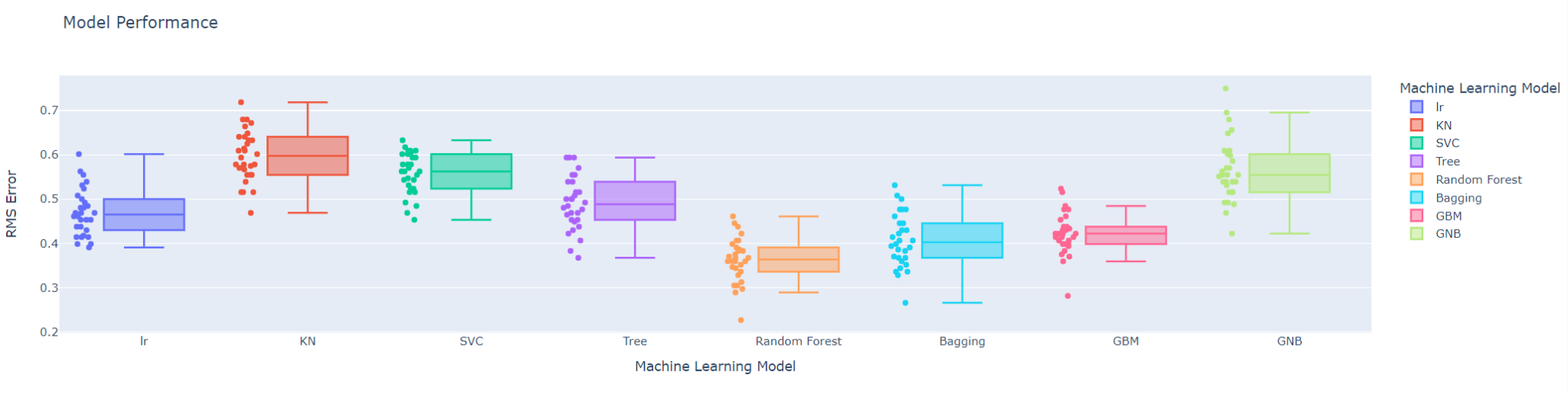
*Table 2 Logistic Regression Coefficients for each wine quality*

A table with numbers and symbols

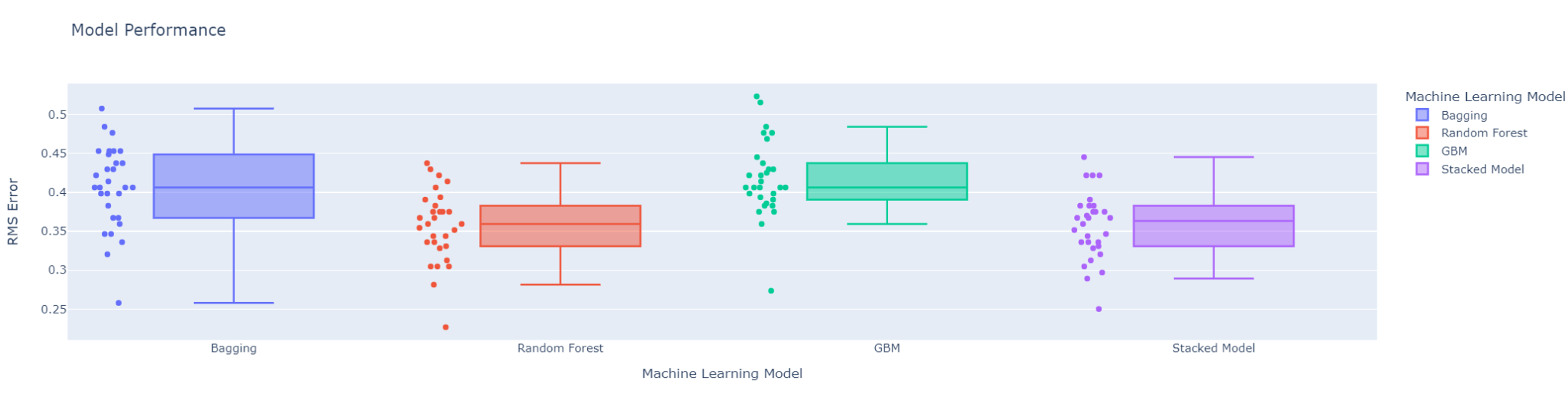
Description automatically generated

# **Model** Selection

Obtaining a baseline provided us insight into the baseline error score, which helps us with model selection. Following the preliminary regression analysis, we have decided to evaluate, Logistic Regression, KNeighbors Classifier, SVM, Decision Tree Classifier, Random Forest Classifier, Bagging Classifier, Gradient Boosting Classifier, and GaussianNB as candidate models. Figure 4 shows the model results, and we can see that Random forests, Bagging, and Gradient boosting perform the best with the smallest mean RMSE. We will select them as level 0 models for the stacking regression. We will use LogisticRegression as the level 1 combiner to aggregate the results of the level 0 models.



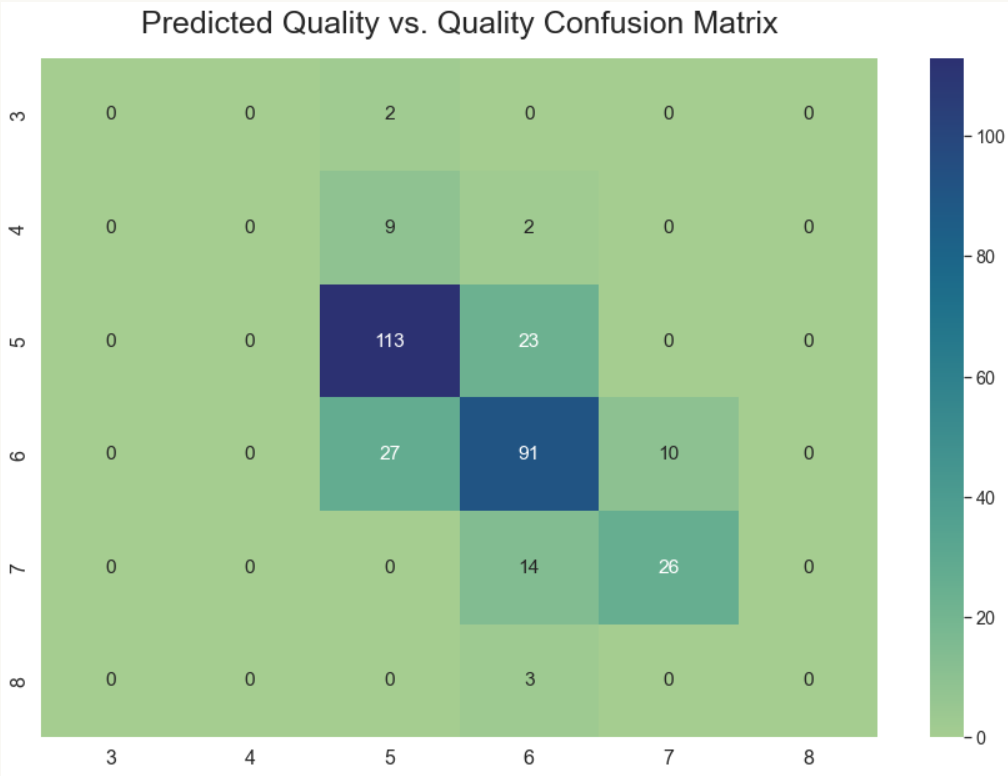
*Figure 4 Machine Learning Model Result*

Figure 5 shows the performance of the standalone and stacked models. The performance of the stacked model has very similar performance to the Random Forest model. We will use the stacked model to perform wine quality prediction in the next steps.

*Figure 5 Top three performing models and stacked model performance*

## Wine Quality Prediction

We use a stacked model to predict wine quality based on all the features in the dataset. The model accuracy score is 71.9%. The confusion matrix (figure 6) shows that 83% of 5-score quality wines, 77% of 6-score quality wines, and 65% of 7-score quality wines are successfully predicted. 3,4,8-score quality wines are not well predicted.



*Figure 6 Confusion matrix for predicted quality vs. true quality*

## Model Performance Evaluation

To visualize the model performance, we drew two plots for 1,599 wines in the datasets and colored them by true quality and predicted quality separately. As shown in Figure 7, the predicted quality overlaps with true quality for most of the wines with quality scores 5-7. Annotations point out instances where the predicted quality is different from true quality, especially that extremely low (3) and extremely high (8) quality wines are barely output by the model.

# 

*Figure 7 True vs Predicted wine quality*

## Appendix

Optimization terminated successfully.

Current function value: 0.916486

Iterations 11

MNLogit Regression Results

==============================================================================

Dep. Variable: y No. Observations: 1279

Model: MNLogit Df Residuals: 1224

Method: MLE Df Model: 50

Date: Thu, 19 Oct 2023 Pseudo R-squ.: 0.2228

Time: 18:11:51 Log-Likelihood: -1172.2

converged: True LL-Null: -1508.2

Covariance Type: nonrobust LLR p-value: 8.845e-110

========================================================================================

y=1 coef std err z P>|z| [0.025 0.975]

----------------------------------------------------------------------------------------

fixed acidity -0.6063 0.573 -1.057 0.290 -1.730 0.518

volatile acidity -3.5227 2.483 -1.419 0.156 -8.390 1.344

citric acid 0.4499 4.963 0.091 0.928 -9.278 10.177

residual sugar -0.1716 0.257 -0.667 0.504 -0.675 0.332

chlorides -8.6083 9.485 -0.908 0.364 -27.200 9.983

free sulfur dioxide -0.0419 0.098 -0.426 0.670 -0.234 0.151

total sulfur dioxide 0.0288 0.044 0.659 0.510 -0.057 0.114

density 13.2386 18.790 0.705 0.481 -23.588 50.066

pH -4.9867 5.467 -0.912 0.362 -15.701 5.728

sulphates 4.6253 5.835 0.793 0.428 -6.811 16.061

alcohol 1.1035 0.821 1.344 0.179 -0.506 2.713

----------------------------------------------------------------------------------------

y=2 coef std err z P>|z| [0.025 0.975]

----------------------------------------------------------------------------------------

fixed acidity -0.3946 0.549 -0.718 0.472 -1.471 0.682

volatile acidity -6.6577 2.425 -2.745 0.006 -11.411 -1.904

citric acid -1.8224 4.822 -0.378 0.705 -11.274 7.629

residual sugar -0.3419 0.236 -1.448 0.148 -0.805 0.121

chlorides -10.2583 8.924 -1.150 0.250 -27.749 7.232

free sulfur dioxide -0.0177 0.094 -0.189 0.850 -0.201 0.166

total sulfur dioxide 0.0506 0.043 1.188 0.235 -0.033 0.134

density 25.0952 18.022 1.392 0.164 -10.228 60.418

pH -6.8828 5.263 -1.308 0.191 -17.199 3.433

sulphates 4.6012 5.673 0.811 0.417 -6.519 15.721

alcohol 0.8371 0.799 1.048 0.295 -0.729 2.403

----------------------------------------------------------------------------------------

y=3 coef std err z P>|z| [0.025 0.975]

----------------------------------------------------------------------------------------

fixed acidity -0.3320 0.551 -0.603 0.547 -1.412 0.748

volatile acidity -9.3240 2.461 -3.789 0.000 -14.147 -4.501

citric acid -2.8858 4.839 -0.596 0.551 -12.369 6.598

residual sugar -0.2833 0.237 -1.196 0.232 -0.747 0.181

chlorides -13.0933 9.001 -1.455 0.146 -30.735 4.548

free sulfur dioxide 0.0007 0.094 0.007 0.994 -0.183 0.185

total sulfur dioxide 0.0365 0.043 0.855 0.392 -0.047 0.120

density 18.4771 18.090 1.021 0.307 -16.978 53.933

pH -7.2574 5.280 -1.375 0.169 -17.605 3.090

sulphates 6.7460 5.680 1.188 0.235 -4.387 17.879

alcohol 1.6298 0.800 2.037 0.042 0.062 3.198

----------------------------------------------------------------------------------------

y=4 coef std err z P>|z| [0.025 0.975]

----------------------------------------------------------------------------------------

fixed acidity -0.2550 0.557 -0.458 0.647 -1.346 0.836

volatile acidity -11.2469 2.590 -4.342 0.000 -16.324 -6.170

citric acid -2.5203 4.909 -0.513 0.608 -12.143 7.102

residual sugar -0.1651 0.244 -0.676 0.499 -0.643 0.313

chlorides -19.9251 9.649 -2.065 0.039 -38.837 -1.013

free sulfur dioxide 0.0068 0.095 0.071 0.943 -0.179 0.192

total sulfur dioxide 0.0260 0.043 0.606 0.544 -0.058 0.110

density 10.2849 18.412 0.559 0.576 -25.803 46.373

pH -8.0759 5.351 -1.509 0.131 -18.563 2.412

sulphates 9.0549 5.705 1.587 0.112 -2.126 20.236

alcohol 2.4191 0.805 3.003 0.003 0.840 3.998

----------------------------------------------------------------------------------------

y=5 coef std err z P>|z| [0.025 0.975]

----------------------------------------------------------------------------------------

fixed acidity -0.6887 0.622 -1.107 0.268 -1.908 0.531

volatile acidity -5.7394 3.474 -1.652 0.099 -12.548 1.070

citric acid 1.2150 5.475 0.222 0.824 -9.515 11.945

residual sugar -0.2081 0.363 -0.574 0.566 -0.919 0.503

chlorides -55.8193 20.883 -2.673 0.008 -96.750 -14.889

free sulfur dioxide 0.0174 0.102 0.170 0.865 -0.183 0.218

total sulfur dioxide 0.0063 0.046 0.138 0.890 -0.084 0.096

density 19.0944 21.388 0.893 0.372 -22.825 61.014

pH -13.7971 6.101 -2.261 0.024 -25.755 -1.839

sulphates 11.9895 5.890 2.036 0.042 0.446 23.533

alcohol 3.1567 0.855 3.693 0.000 1.481 4.832

========================================================================================