Predicting Wine Quality From Wine Chemistry

Modeling Red and White Wine

Robert V. Moel

May 1, 2022

Abstract

Wine is loved by people from around the world. Being able to use wine chemistry to determine wine quality would speed selection and permit for a tasty repast.

1 Data Source, EDA, and Preprocessing

The source of data used for this evaluation comes from https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/ This data table consists of the following

Name of Parameter	Value Type
Fixed acidity	Real
Volatile acidity	Real
Citric Acid	Real
Residual sugar	Real
Chlorides	Real
Free sulfur dioxide	Real
Total sulfur dioxide	Real
Density	Real
pH	Real
Sulphates	Real
Alcohol	Real
Quality (score between 0 and 10)	Integer

The data is presented in two tables: one for red (1599 rows) and one for white (4898 rows) with 12 different attributes - 11 explanatory attributes and 1 predictive attribute called "quality." Each wine's quality is ranked from 1 through 10 with 10 representing a wine with the highest quality based on a human taster. For the analysis, an additional column, color, was added to each table where, 1 means a red wine and 0 means a white wine. The tables were then joined together to create a single table to analyze. Both tables were very complete and no row contained missing data.

Descriptive statistics where run on the data with the following results:

	Count	Mean	Std	Min	25%		75%	Max
Fixed acidity	6497.0	7.215307	1.296434	3.80000	6.40000	7.00000	7.70000	15.90000
Volatile acidity	6497.0	0.339666	0.164636	0.08000	0.23000	0.29000	0.40000	1.58000
Citric acid	6497.0	0.318633	0.145318	0.00000	0.25000	0.31000	0.39000	1.66000
Residual sugar	6497.0	5.443235	4.757804	0.60000	1.80000	3.00000	8.10000	65.80000
Chlorides	6497.0	0.056034	0.035034	0.00900	0.03800	0.04700	0.06500	0.61100
Free sulfur dioxide	6497.0	30.525319	17.749400	1.00000	17.00000	29.00000	41.00000	289.00000
Total sulfur dioxide	6497.0	115.744574	56.521855	6.00000	77.00000	118.00000	156.00000	440.00000
Density	6497.0	0.994697	0.002999	0.98711	0.99234	0.99489	0.99699	1.03898
pH	6497.0	3.218501	0.160787	2.72000	3.11000	3.21000	3.32000	4.01000
Sulfates	6497.0	0.531268	0.148806	0.22000	0.43000	0.51000	0.60000	2.00000
Alcohol	6497.0	10.491801	1.192712	8.00000	9.50000	10.30000	11.30000	14.90000
Quality	6497.0	5.818378	0.873255	3.00000	5.00000	6.00000	6.00000	9.00000
Wine_color	6497.0	0.753886	0.430779	0.00000	1.000	1.00000	1.00000	1.0000000

We can see from these density plots that the data are not particularly normal. Let's look at two versions of the correlation heat map.

Some explanatory variables are not independent. As expected, measures of acidity are related

Let's look at a pair-plots diagram to get a better look at some of these relationships:

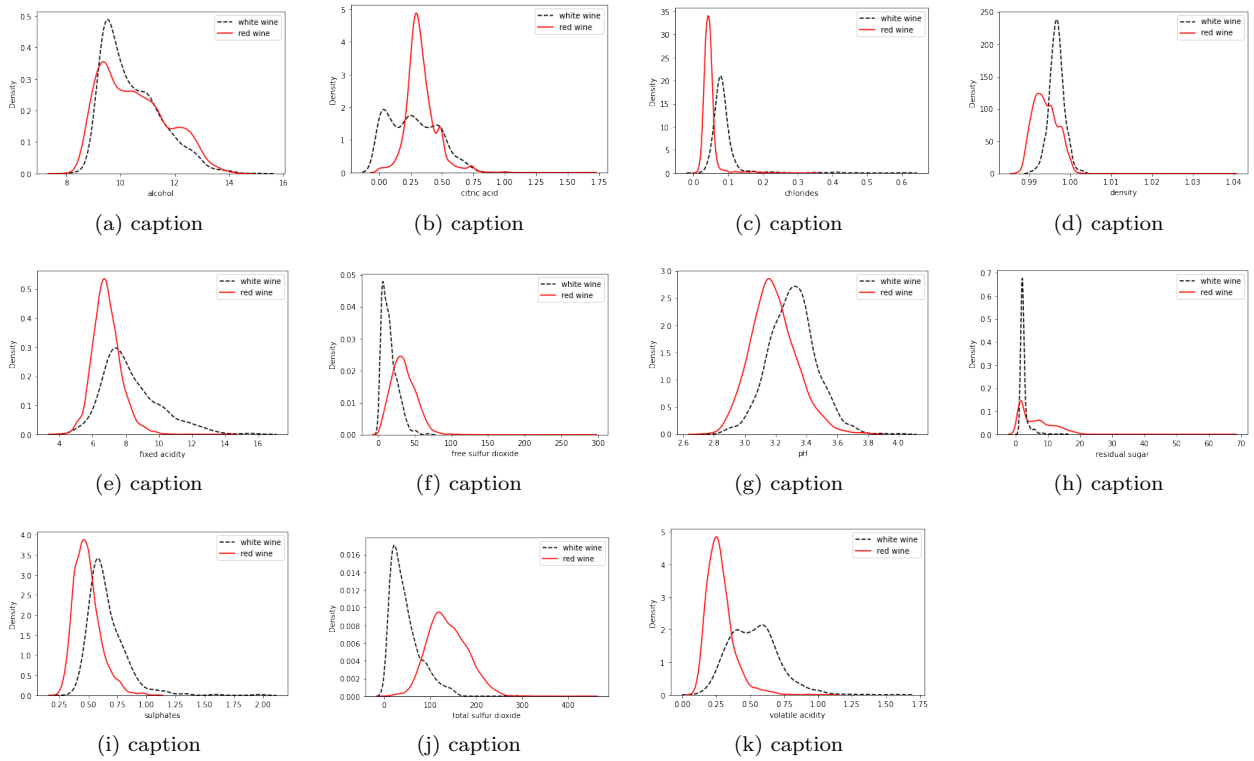
The data are now ready to be used in modeling.

2 Modeling Approaches

Several modeling approaches will be tried to determine if chemistry can be used to find wine quality for both read and white whites. The methods to be considered will be ordered logistic regression with ordered probit kernel, logistic regression with logit kernel, ordered logit kernel, a plain random forest regressor, and a tuned random forest regressor. First, the data will be scaled, and a training and test sample will be drawn. This will be used to find appropriate models.

2.1 Choosing Test and Training

To ensure we get a representative sample of each wine score and each wine color, the training and test data are created using a stratification variable created from wine_color and from each of the quality scores. Twenty-percent of the data were set aside as test. The dummy stratifying variable was then removed. The training data were then scaled using StandardScaler and this was applied to the test data.



(a) caption (b) caption (c) caption (d) caption

(e) caption (f) caption (g) caption (h) caption

(i) caption (j) caption (k) caption

Figure 1: Density Plots of Red and White Wines

2.2 Order Logistic Model

The first ordered logistic model included all explanatory variables. The model output is included below:

From this first run, we can see that citric acid and chlorides are not significant and can be dropped.

This model appears to have all variables as significant. Let's use this model against the test data and see how well it does in predicting quality. An output of the multi-label confusion matrix is as follows: From this table we can see that labels at

Label	Precision	Recall	Specificity	Accuracy	F1
3	1.0000	0.9915	0.0000	0.9915	0.9957
4	1.0000	0.9700	1.0000	0.9700	0.9848
5	0.8126	0.8045	0.5901	0.7377	0.8085
6	0.4743	0.6772	0.5233	0.5831	0.5579
7	0.9474	0.8544	0.4184	0.8215	0.8985
8	0.9992	0.9723	0.0000	0.9715	0.9856
9	1.0000	0.9969	0.0000	0.9969	0.9984

either end of the quality spectrum do well but labels in the center have difficulty determining average wines. Overall, on the test y variables the model correctly predicts the quality 53.6% of the time.

2.3 Ordered Probit Model

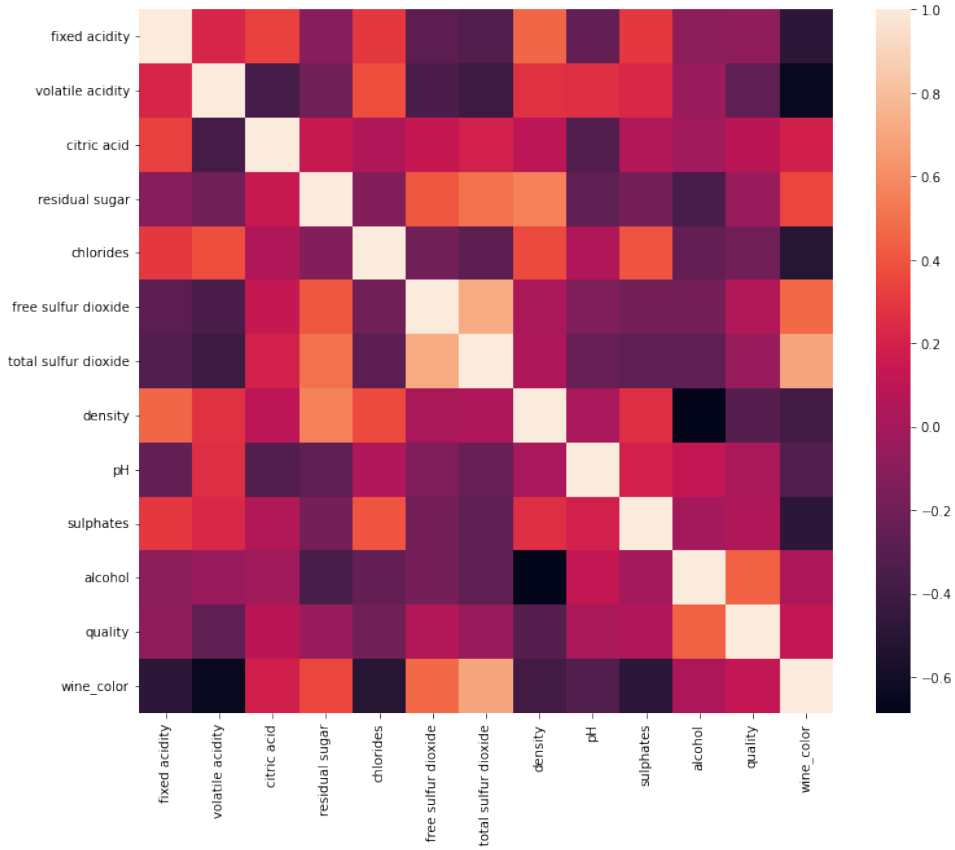
Let's use a probit kernel and see if this performs better using the same subset of variables as the second logit model.

From this we can see the model is valid and all explanatory variables are significant. Let's look at the results of the multi-label confusion matrix and the overall accuracy:

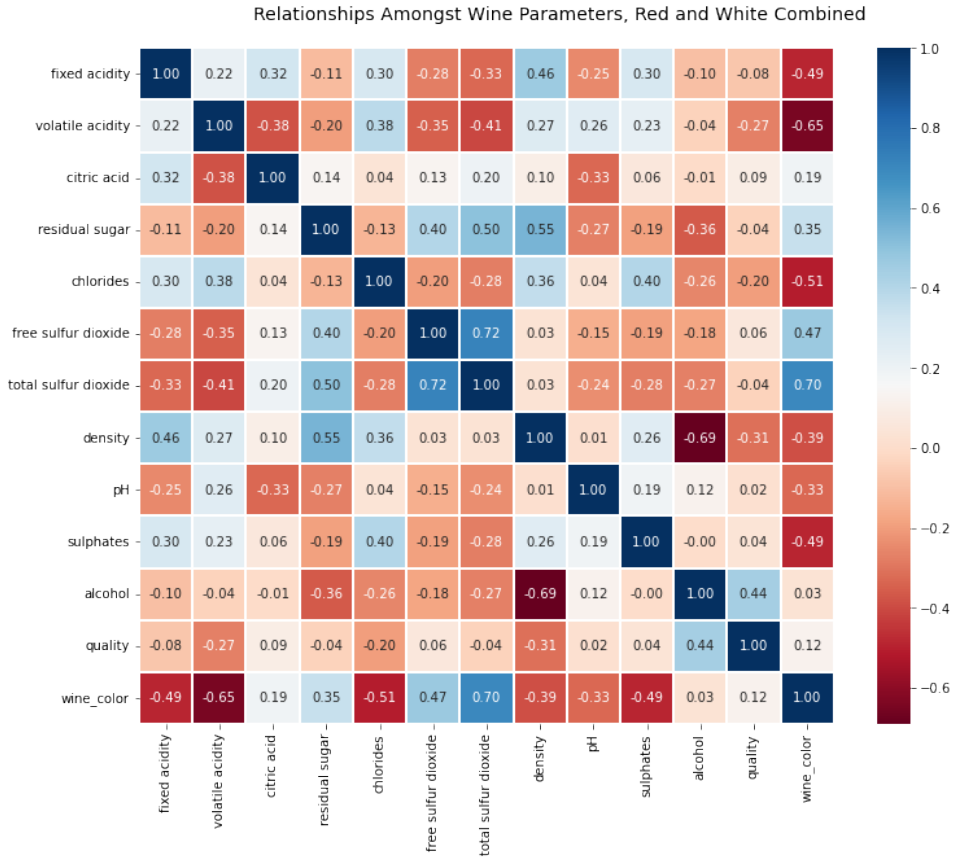
The overall ability to correctly predict quality for this model was 52.5% not much better. Again, we see weakness in its ability to predict in the middle quality score region.

2.4 Random Forest Regressor

Let's look at an untuned random forest regressor and we use the default values and max_depth of 2. Running this regressor and checking its ability to predict quality it does so about 51.6% of the time which, isn't much better than a standard logistic regression.

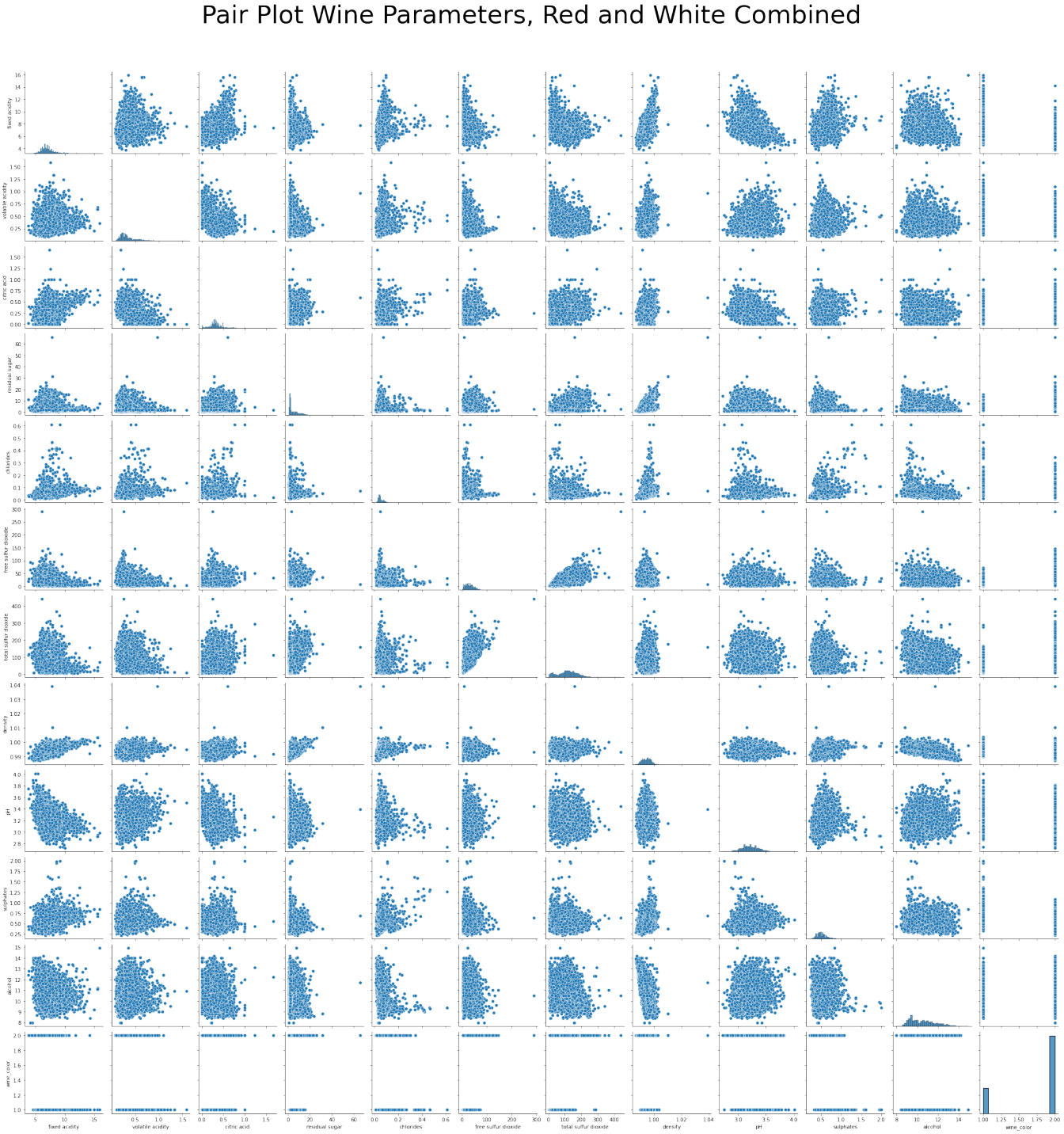


(a) Heat Map



(b) Correlation

Figure 2: Heat Maps



(a) caption

Figure 3: Pair Plots

OrderedModel Results

Dep. Variable:	quality	**Log-Likelihood:**	-5620.5
Model:	OrderedModel	**AIC:**	1.128e+04
Method:	Maximum Likelihood	**BIC:**	1.140e+04
Date:	Sun, 01 May 2022		
Time:	12:07:40		
No. Observations:	5197		
Df Residuals:	5179		
Df Model:	18		

	coef	std err	z	P>|z|	[0.025	0.975]
fixed acidity	0.2941	0.065	4.557	0.000	0.168	0.421
volatile acidity	-0.6709	0.042	-16.017	0.000	-0.753	-0.589
citric acid	-0.0029	0.034	-0.085	0.932	-0.070	0.064
residual sugar	0.8289	0.090	9.195	0.000	0.652	1.006
chlorides	-0.0673	0.035	-1.904	0.057	-0.137	0.002
free sulfur dioxide	0.2665	0.042	6.399	0.000	0.185	0.348
total sulfur dioxide	-0.2427	0.054	-4.480	0.000	-0.349	-0.137
density	-0.8935	0.142	-6.275	0.000	-1.173	-0.614
pH	0.2383	0.045	5.308	0.000	0.150	0.326
sulphates	0.2661	0.034	7.906	0.000	0.200	0.332
alcohol	0.6983	0.071	9.889	0.000	0.560	0.837
wine_color	-0.4289	0.076	-5.658	0.000	-0.577	-0.280
3/4	-6.3900	0.233	-27.405	0.000	-6.847	-5.933
4/5	0.8883	0.091	9.781	0.000	0.710	1.066
5/6	1.1694	0.025	47.157	0.000	1.121	1.218
6/7	0.9562	0.019	49.141	0.000	0.918	0.994
7/8	0.8491	0.035	24.489	0.000	0.781	0.917
8/9	1.6386	0.194	8.451	0.000	1.259	2.019

(a) caption

Figure 4: First Logit Model

OrderedModel Results

Dep. Variable:	quality	**Log-Likelihood:**	-5634.8
Model:	OrderedModel	**AIC:**	1.130e+04
Method:	Maximum Likelihood	**BIC:**	1.140e+04
Date:	Sun, 01 May 2022		
Time:	12:07:53		
No. Observations:	5197		
Df Residuals:	5182		
Df Model:	15		

	coef	std err	z	P>|z|	[0.025	0.975]
volatile acidity	-0.7067	0.039	-18.171	0.000	-0.783	-0.631
residual sugar	0.5320	0.058	9.209	0.000	0.419	0.645
free sulfur dioxide	0.2620	0.042	6.305	0.000	0.181	0.343
total sulfur dioxide	-0.2593	0.054	-4.836	0.000	-0.364	-0.154
density	-0.3858	0.078	-4.968	0.000	-0.538	-0.234
pH	0.0962	0.030	3.243	0.001	0.038	0.154
sulphates	0.2254	0.032	6.939	0.000	0.162	0.289
alcohol	0.9324	0.048	19.501	0.000	0.839	1.026
wine_color	-0.3219	0.070	-4.597	0.000	-0.459	-0.185
3/4	-6.3863	0.233	-27.376	0.000	-6.844	-5.929
4/5	0.8892	0.091	9.796	0.000	0.711	1.067
5/6	1.1695	0.025	47.085	0.000	1.121	1.218
6/7	0.9501	0.019	48.942	0.000	0.912	0.988
7/8	0.8473	0.035	24.398	0.000	0.779	0.915
8/9	1.6397	0.194	8.446	0.000	1.259	2.020

(a) caption

Figure 5: Second Logit Model

OrderedModel Results

Dep. Variable:	quality	**Log-Likelihood:**	-5663.7
Model:	OrderedModel	**AIC:**	1.137e+04
Method:	Maximum Likelihood	**BIC:**	1.149e+04
Date:	Sun, 01 May 2022		
Time:	12:08:01		
No. Observations:	5197		
Df Residuals:	5178		
Df Model:	19		

	coef	std err	z	P>|z|	[0.025	0.975]
fixed acidity	0.1558	0.034	4.600	0.000	0.089	0.222
volatile acidity	-0.3877	0.023	-16.870	0.000	-0.433	-0.343
citric acid	-0.0082	0.020	-0.417	0.676	-0.046	0.030
residual sugar	0.4460	0.048	9.372	0.000	0.353	0.539
chlorides	-0.0419	0.020	-2.125	0.034	-0.081	-0.003
free sulfur dioxide	0.1282	0.023	5.672	0.000	0.084	0.172
total sulfur dioxide	-0.1254	0.030	-4.124	0.000	-0.185	-0.066
density	-0.4748	0.071	-6.719	0.000	-0.613	-0.336
pH	0.1293	0.024	5.298	0.000	0.081	0.177
sulphates	0.1448	0.019	7.645	0.000	0.108	0.182
alcohol	2.3090	0.557	4.147	0.000	1.218	3.400
wine_color	-0.2402	0.040	-5.945	0.000	-0.319	-0.161
logalco	-1.9277	0.557	-3.463	0.001	-3.019	-0.837
3/4	-3.1326	0.083	-37.951	0.000	-3.294	-2.971
4/5	-0.0248	0.080	-0.309	0.757	-0.182	0.133
5/6	0.5463	0.022	24.786	0.000	0.503	0.590
6/7	0.4067	0.018	22.522	0.000	0.371	0.442
7/8	0.2102	0.032	6.620	0.000	0.148	0.272
8/9	0.6525	0.155	4.220	0.000	0.350	0.956

(a) caption

Figure 6: First Probit Model

Label	Precision	Recall	Specificity	Accuracy	F1
3	1.0000	0.9915	0.0000	0.9915	0.9957
4	1.0000	0.9692	0.0000	0.9692	0.9844
5	0.8036	0.8000	0.5756	0.7292	0.8018
6	0.4674	0.6634	0.5152	0.5731	0.5484
7	0.9456	0.8527	0.3980	0.8185	0.8968
8	1.0000	0.9723	0.0000	0.9723	0.9860
9	1.0000	0.9969	0.0000	0.9969	0.9984

2.5 Tuned Random Forest

In this next model we tune the following parameters: n_estimators, max_features, max_depth, min_samples_split, and min_samples_leaf. We set bootstrap to True. We used 5-fold cross-validation.

The best parameters are:

Hyper-parameter	Value
Bootstrap	True
Max Depth	43
Max Features	Auto
Min Samples Leaf	1
Min Samples Split	2
N Estimators	134

The overall accuracy of this model is 64.5% or 10-percentage points higher than logistic regression

A visual of the tree is produced below:

The confusion matrix produced the following results for each of the quality levels:

Label	Precision	Recall	Specificity	Accuracy	F1
3	1.0000	0.9915	0.0000	0.9915	0.9957
4	1.0000	0.9700	1.0000	0.9700	0.9848
5	0.8488	0.8714	0.6934	0.8115	0.8600
6	0.6630	0.7772	0.6453	0.7077	0.7156
7	0.9327	0.8979	0.5805	0.8554	0.9150
8	1.0000	0.9746	1.0000	0.9746	0.9871
9	1.0000	0.9969	0.0000	0.9969	0.9984

What's clear here is that the random tree performs much better in the mid-range quality. This improved the model's performance significantly.

3 Further Research

While 65% is a great improvement over plain logistic regression models, better predictive power should be achievable with additional samples and not just Portuguese wine. Because of the time it takes to tune this model, only a sample of hyper-parameters were tuned if more time and a more powerful CPU were available, more parameters could be tuned. The wines considered were Portuguese. Additional wine nationalities should be included. Additionally, only red or white were represented. Ideally, rose should be added as a category.

4 Top Three Recommendations on Using this Model

First, this model can be used to identify great versus good wine brands from the thousands that are out there by just knowing the chemistry of the wine. While this won't put great sommeliers out of work, it would help the average consumer find a great bottle of wine. Second, this model can be used in the wine production and quality control process as a wine is produced and to help wines conform to chemistry that will make them taste better. Finally, this model can be used to improve wines by adjusting their chemistry to wines that appear to taste better. Because the model is tuned to both white and red wines, the specific tuning can be created

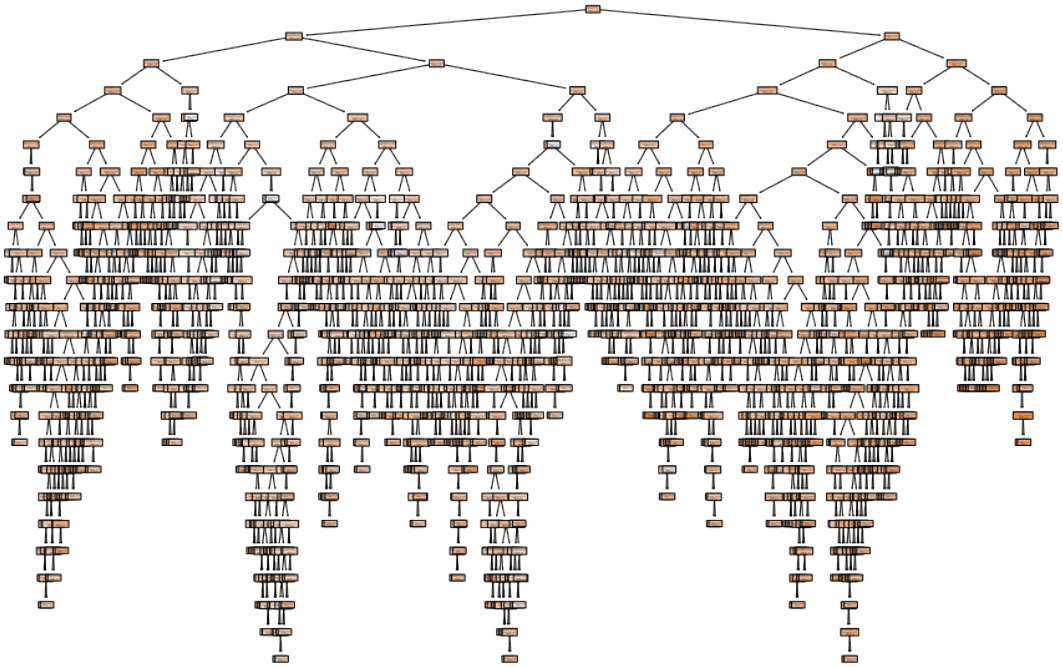


Figure 7: Tuned Tree Regressor Model