Group 4
Huan He | Qijia Zheng | Kelsey Strachan

Modeling Wine Quality Distribution

Section I: Abstract

To predict the quality for white and red wines, we applied Generalized Linear Models (GLM), Generalized Additive Models (GAM), Gaussian Mixture Models (GMM) and Random Forest modeling to identify the best-fitting approach. Both GLM and GMM were found to be ineffective in this context. Evaluation results indicated that both the GAM with a Gamma distribution and the Random Forest classifier demonstrated the strongest predictive power. The GAM with Gamma distribution yielded the lowest Akaike Information Criterion (AIC) scores (red wine: 2526.03; white wine: 8654.79) and the lowest misclassification rates (red wine: ~39.0%; white wine: ~46.0%). The Random Forest classifier achieved notable accuracy levels for both varieties (red wine: 65.00%; white wine: 71.12%). Our analysis highlighted the significant role of the `alcohol` feature in predicting wine quality for both white and red wines, followed by `density`, `volatile acidity`, `sulphates`, and `total sulfur dioxide`. Our process also identified challenges related to outlier sensitivity and complex interactions between features.

Section II: Introduction

The primary objective of our analysis is to explore the characteristics of white and red wines through the following key questions:

- Can we develop a model to predict the distribution of wine quality based on its physicochemical properties?
- Are there significant differences in the physicochemical properties between red and white wines, and do these differences impact their overall quality?
- Which physicochemical features are most strongly correlated with wine quality, and how do they influence consumer preferences and expert evaluations?

Based on those research questions, we will investigate the following scientific hypotheses:

1. Physicochemical Properties Predict Wine Quality
    - Statement: The physicochemical properties of wines, such as acidity, alcohol content, and sulfur dioxide levels, can predict the distribution of wine quality on a scale from 0 (very poor) to 10 (excellent).
    - Rationale: Wine quality is influenced by a complex interplay of various physicochemical properties. Previous studies have shown that factors like acidity, alcohol content, and sulfur dioxide levels play crucial roles in determining the sensory attributes of wine. By

analyzing these properties, we aim to build predictive models that can estimate the quality distribution for new wine samples.

2. Differences in Physicochemical Properties Between Red and White Wines Influence Quality
    - Statement: There are significant differences in the physicochemical properties between red and white wines.
    - Rationale: Red and white wines are produced using different processes, which can lead to distinct chemical compositions. These differences can affect the sensory attributes and, consequently, the perceived quality of the wines. By comparing the physicochemical properties of red and white wines, we aim to understand how production methods and chemical composition impact wine quality.

3. Certain Physicochemical Features are More Strongly Correlated with Wine Quality
    - Statement: Specific physicochemical features are more strongly correlated with wine quality than others and contribute significantly to the perceived quality of wine.
    - Rationale: Not all physicochemical properties may have an equal impact on wine quality. Some features might have a stronger influence on the sensory attributes that determine quality. By identifying the most influential features, we can provide insights into which properties are critical for improving wine quality.

Section III: Discussion and Analysis

To address the scientific hypotheses, we employ a combination of statistical analysis, machine learning, and data visualization techniques.

Shapiro-Wilk test

To assess the normality of the physicochemical properties in both white and red wine datasets, we employed the Shapiro-Wilk test, a widely used statistical test for normality. This test evaluates whether a dataset is likely to have come from a normally distributed population. For each feature in the datasets, we calculated the test statistic and the corresponding p-value. A small p-value (typically < 0.05) indicates strong evidence against the null hypothesis of normality, suggesting that the data are not normally distributed.The Shapiro-Wilk test results (Table 1) indicate that the physicochemical properties of both white and red wines do not follow a normal distribution. This finding is crucial for selecting appropriate statistical methods and models that do not assume normality.

Table 1. Shapiro-Wilk test Results

Attribute	Red Wine		White Wine	
	Statistics	p-value	Statistics	p-value
fixed acidity	0.942	0.000	0.977	0.000
volatile acidity	0.974	0.000	0.905	0.000
citric acidity	0.955	0.000	0.922	0.000
residual sugar	0.566	0.000	0.885	0.000
chlorides	0.484	0.000	0.591	0.000
free sulfur dioxide	0.902	0.000	0.942	0.000
total sulfur dioxide	0.873	0.000	0.989	0.000
density	0.991	0.000	0.955	0.000
pH	0.993	0.000	0.988	0.000
sulphates	0.833	0.000	0.952	0.000
alcohol	0.929	0.000	0.955	0.000

Mann-Whitney U test

Since both red and white wine features are not normally distributed, we performed Mann-Whitney U test to investigate whether there are significant differences in the physicochemical properties between red and white wines. For the test result, a smaller test statistic indicates a larger difference between the groups and a small p-value (typically < 0.05) suggests that the difference between the groups is statistically significant. According to Table 2, except for alcohol content (p-value > 0.05), all other physicochemical properties have extremely small p-values, indicating that there are highly significant differences in most of the physicochemical properties between red and white wines. Therefore, the hypothesis that there are significant differences in the physicochemical properties between red and white wines is strongly supported by the Mann-Whitney U test results.

Table 2. Mann-Whitney U test Result

Attribute	Median (Red)	Median (White)	Test Statistic	p-value
fixed acidity	7.90	6.80	6138507.0	1.438930e-255
volatile acidity	0.52	0.26	7059623.5	0.00
citric acidity	0.26	0.32	3070088.5	1.312558e-38
residual sugar	2.20	5.20	2569687.0	5.634073e-95
chlorides	0.08	0.04	7407015.5	0.00
free sulfur dioxide	14.00	34.00	1186396.5	0.00
total sulfur dioxide	38.00	134.00	366639.5	0.00
density	0.997	0.994	6059284.5	1.453091e-237
pH	3.31	3.18	5681839.5	5.472258e-162
sulphates	0.62	0.47	6509961.0	0.00
alcohol	10.20	10.40	3829043.5	0.182

Model Selection

To predict the distribution of wine quality based on its physicochemical properties, we applied various models, specifically focusing on Generalized Linear Models (GLM), Generalized Additive Models (GAM) and Gaussian Mixture Models (GMM). Based on correlation analysis performed in exploratory data analysis (EDA), we remove `residual sugar` when building a predictive model for white wine since `density` and `residual sugar` have high correlation (~0.84). The models were evaluated based on two key metrics: the Akaike Information Criterion (AIC) and the misclassification rate. AIC provides a measure of the relative quality of a statistical model for a given set of data, with a lower AIC indicating a better fit. The misclassification rate, on the other hand, quantifies the proportion of instances incorrectly classified by the model, with a lower rate indicating higher accuracy.

Table 3. Model Comparison

Method	Red Wine	
	AIC	Misclassification Rate
GLM (Poisson)	4706.13	0.43
GLM (Negative Binomial)	7207.13	0.42
GAM (Gamma)	2526.03	0.39
GAM (Gaussian)	3098.28	0.39
GAM (InverseGaussian)	3181.59	0.39
GAM (Tweedie)	83971.44	0.39
GMM	7481.6989	0.69
Method	White Wine	
	AIC	Misclassification Rate
GLM (Poisson)	14610.79	0.497
GLM (Negative Binomial)	22339.93	0.493
GAM (Gamma)	8654.79	0.46
GAM (Gaussian)	10774.67	0.47
GAM (InverseGaussian)	10984.10	0.47
GAM (Tweedie)	203860.75	0.47
GMM	23306.5083	0.70

The results, as summarized in Table 3, show a clear distinction in performance between the GLM, GAM and GMM approaches for both red and white wines. Based on the AIC and misclassification rate, the GAM with the Gamma family consistently outperformed other models for both red and white wines, indicating that it provides the best balance between model complexity and goodness of fit. The Gamma family's lower AIC values and lower misclassification rates suggest that it captures the relationship between the predictors and the response variable more effectively than the other distributions considered.

Generalized Linear Models (GLMs)

Poisson Regression Results:
Generalized Linear Model Regression Results

Dep. Variable:	quality	No. Observations:	3918
Model:	GLM	Df Residuals:	3907
Model Family:	Poisson	Df Model:	10
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-7294.4
Date:	Fri, 21 Mar 2025	Deviance:	384.68
Time:	17:05:49	Pearson chi2:	383.
No. Iterations:	4	Pseudo R-squ. (CS):	0.03565
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-6.2637	3.836	-1.633	0.102	-13.781	1.254
fixed acidity	-0.0149	0.010	-1.558	0.119	-0.034	0.004
volatile acidity	-0.3578	0.071	-5.038	0.000	-0.497	-0.219
citric acid	-0.0190	0.060	-0.316	0.752	-0.137	0.099
chlorides	-0.2094	0.343	-0.610	0.542	-0.882	0.463
free sulfur dioxide	0.0011	0.001	2.158	0.031	0.000	0.002
total sulfur dioxide	-0.0001	0.000	-0.491	0.623	-0.001	0.000
density	7.5354	3.832	1.966	0.049	0.025	15.046
pH	-0.0053	0.050	-0.105	0.916	-0.104	0.093
sulphates	0.0648	0.059	1.104	0.269	-0.050	0.180
alcohol	0.0685	0.009	7.485	0.000	0.051	0.086

Poisson Regression Results:
Generalized Linear Model Regression Results

Dep. Variable:	quality	No. Observations:	1279
Model:	GLM	Df Residuals:	1267
Model Family:	Poisson	Df Model:	11
Link Function:	Log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2341.1
Date:	Fri, 21 Mar 2025	Deviance:	97.339
Time:	21:06:29	Pearson chi2:	95.7
No. Iterations:	4	Pseudo R-squ. (CS):	0.03911
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.3258	15.166	0.153	0.878	-27.398	32.050
fixed acidity	0.0033	0.018	0.182	0.855	-0.033	0.039
volatile acidity	-0.1835	0.091	-2.027	0.043	-0.361	-0.006
citric acid	-0.0285	0.109	-0.262	0.793	-0.241	0.184
residual sugar	0.0008	0.011	0.076	0.940	-0.020	0.022
chlorides	-0.3196	0.299	-1.069	0.285	-0.906	0.266
free sulfur dioxide	0.0011	0.002	0.653	0.514	-0.002	0.004
total sulfur dioxide	-0.0007	0.001	-1.266	0.206	-0.002	0.000
density	-0.8613	15.477	-0.056	0.956	-31.195	29.472
pH	-0.0707	0.138	-0.512	0.609	-0.341	0.200
sulphates	0.1464	0.080	1.837	0.066	-0.010	0.303
alcohol	0.0492	0.019	2.595	0.009	0.012	0.086

Figure 1. Generalized Linear Regression (Poisson) of White Wine (Top) and Red Wine (Bottom)

Poisson regression was employed within the framework of Generalized Linear Models (GLM), as we considered it appropriate for modeling wine quality ratings. Alongside Poisson regression, Negative Binomial was also fitted for comparison (Figure 1).

The GLM regression results indicated that several predictors of white wine quality were statistically significant, including `volatile acidity`, `free sulfur dioxide`, `density`, and `alcohol`. Among these, `alcohol` exhibited the most substantial positive effect on white wine quality suggesting that higher alcohol content is associated with improved ratings. In contrast, `volatile acidity` displayed the strongest negative effect, implying that higher levels of `volatile acidity` are detrimental to white wine quality. The remaining variables included in the model, however, demonstrated negligible effects on the quality ratings and were statistically insignificant in predicting quality. These findings suggest that the relationship between wine quality and certain physicochemical properties may be more complex or influenced by factors not captured by variables in the current model.

Red wine GLM results suggest that `volatile acidity` and `alcohol` are the most influential predictors of `quality`, with `alcohol` positively related to `quality` and `volatile acidity` negatively related. The overall model explains only a small portion of the variation in red wine `quality`, so other factors are influencing and not being captured by the model.

Regarding model performance, the white wine misclassification rates were found to be 49.7% for the Poisson model and 49.3% for the Negative Binomial model. Red wine rates were found to be 43.0% for Poisson and 42.0% for Negative Binomial. Both models showed similar predictive accuracy, with very comparable misclassification rates. However, when considering model fit, the Poisson model yielded lower AIC scores (red: 4706.13; white: 14610.79). This is in contrast to the Negative Binomial model's AIC values (red: 7207.13; white: 22339.93). Given the substantially lower AIC of the Poisson model, it was considered a better fit to the data. Despite this, the high misclassification rate (~49.7%) indicated suboptimal performance for both models. Consequently, this prompted the exploration of further models to improve prediction accuracy and further refine the analysis.

Generalized Additive Models (GAMs) & Splines

The GAM utilizing a Gamma distribution with an inverse power link function offered insightful findings regarding the key predictors of wine quality. The analysis revealed that `fixed acidity`, `volatile acidity`, `density` and `pH` are statistically significant factors influencing white wine quality. Among these, `volatile acidity` and `density` exhibited the most substantial effects on the predicted quality (Figure 2).

The coefficient for white wine `fixed acidity` is 0.0098 with a highly significant p-value of less than 0.0001, suggesting a positive relationship between `fixed acidity` and wine quality. This indicates that higher levels of `fixed acidity` tend to result in higher quality white wine, assuming other factors are constant. Similarly, `volatile acidity` showed a coefficient of 0.0495, with a p-value indicating statistical

significance, suggesting that this `volatile acidity` also has a positive effect on white wine quality, even though it is commonly expected to negatively impact quality. The surprising positive effect of `volatile acidity` warrants further exploration, as it contrasts with typical understanding of wine chemistry.

The `density` variable exhibited a coefficient is 0.0272, with a low p-value, indicating a significant positive relationship with white wine quality. This suggests that white wines with higher `density` tend to be rated higher in quality. Additionally, `pH` was found to have a coefficient of 0.0404, with a low p-value, demonstrating that as pH increases, white wine quality is positively influenced. This aligns with the general understanding that wines with higher pH tend to exhibit more desirable characteristics. Alternatively, certain variables such as `citric acid`, `chlorides`, `free sulfur dioxide` and `total sulfur dioxide` were found to be statistically insignificant in predicting white wine quality. These variables either have weak effects or no discernible impact in the context of this model, as indicated by their high p-values.

Similarly, the red wine model identified several significant predictors of wine quality, including `volatile acidity` (coefficient: 0.0398; p-value: <0.001), `chlorides` (coefficient: 0.0293; p-value: 0.030), `density` (coefficient: 0.0723; p-value: <0.001) and `pH` (coefficient: 0.0443; p-value: <0.001). Features `fixed acidity`, `citric acid`, `free sulfur dioxide`, and `sulphates` showed negligible effects on the model. The red wine model explained ~48% of the variability in wine quality (Figure 3). Overall, both models provided a nuanced understanding of the factors affecting wine quality.

Figure 4 presents a visual comparison of the true and predicted wine quality distributions using the GAM with a Gamma family for both white and red wines. The density plots illustrate the distribution of true quality (blue) and predicted quality (red) across different quality scores, ranging from 3 to 9. For white wine (left panel), the predicted quality distribution closely aligns with the true quality distribution, particularly around the peaks at quality scores of 6 and 7. This indicates that the model effectively captures the central tendency of the data. However, there is a noticeable discrepancy at higher quality scores (8 and 9), where the predicted distribution shows lower density compared to the true distribution. This suggests that the model may underestimate the probability of high-quality white wines.

In contrast, the red wine analysis (right panel) reveals a slightly more accurate prediction, as evidenced by the misclassification rate of 0.39. The predicted quality distribution peaks at a slightly lower quality score compared to the true distribution, particularly around the main peak at quality score 5. This shift in the peak suggests that the model tends to predict red wines to be of slightly lower quality than they actually are.

Overall, while the GAM model demonstrates a reasonable ability to predict wine quality for both white and red wines, there are areas for improvement, particularly in accurately predicting high-quality wines and reducing the spread of predicted quality scores.

Generalized Linear Model Regression Results

Dep. Variable:	quality	No. Observations:	3918
Model:	GLMGam	Df Residuals:	3877.00
Model Family:	Gamma	Df Model:	40.00
Link Function:	InversePower	Scale:	0.015418
Method:	PIRLS	Log-Likelihood:	-4286.4
Date:	Thu, 20 Mar 2025	Deviance:	60.587
Time:	18:33:38	Pearson chi2:	59.8
No. Iterations:	5	Pseudo R-squ. (CS):	0.3989
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	0.025	0.975]
const	0.0361	0.008	4.804	0.000	0.021	0.051
fixed acidity	0.0098	0.002	5.475	0.000	0.006	0.013
volatile acidity	0.0495	0.007	6.679	0.000	0.035	0.064
citric acid	0.0038	0.006	0.635	0.526	-0.008	0.016
chlorides	0.0078	0.006	1.301	0.193	-0.004	0.020
free sulfur dioxide	-4.966e-05	7e-05	-0.710	0.478	-0.000	8.75e-05
total sulfur dioxide	-1.792e-05	3.77e-05	-0.476	0.634	-9.17e-05	5.59e-05
density	0.0272	0.007	3.769	0.000	0.013	0.041
pH	0.0404	0.005	8.054	0.000	0.031	0.050
sulphates	-0.0003	0.005	-0.060	0.952	-0.011	0.010
alcohol	-0.0016	0.001	-1.262	0.207	-0.004	0.001
fixed acidity_s0	0.0181	0.011	1.577	0.115	-0.004	0.041
fixed acidity_s1	-0.0196	0.005	-3.672	0.000	-0.030	-0.009
fixed acidity_s2	-0.0281	0.013	-2.217	0.027	-0.053	-0.003
fixed acidity_s3	0.0130	0.012	1.102	0.270	-0.010	0.036
volatile acidity_s0	0.0001	0.005	0.023	0.982	-0.010	0.010
volatile acidity_s1	0.0195	0.004	4.498	0.000	0.011	0.028
volatile acidity_s2	-0.0399	0.014	-2.799	0.005	-0.068	-0.012
volatile acidity_s3	0.0670	0.015	4.373	0.000	0.037	0.097
citric acid_s0	-0.0135	0.005	-2.829	0.005	-0.023	-0.004
citric acid_s1	-0.0347	0.007	-4.940	0.000	-0.049	-0.021
citric acid_s2	0.0345	0.014	2.404	0.016	0.006	0.063
citric acid_s3	-0.0082	0.012	-0.706	0.480	-0.031	0.015
chlorides_s0	-0.0085	0.007	-1.294	0.196	-0.021	0.004
chlorides_s1	0.0291	0.007	4.175	0.000	0.015	0.043
chlorides_s2	-0.0587	0.017	-3.430	0.001	-0.092	-0.025
chlorides_s3	0.0530	0.021	2.582	0.010	0.013	0.093
free sulfur dioxide_s0	-0.0343	0.004	-8.188	0.000	-0.042	-0.026
free sulfur dioxide_s1	-0.0335	0.005	-7.227	0.000	-0.043	-0.024
free sulfur dioxide_s2	-0.0467	0.015	-3.206	0.001	-0.075	-0.018
free sulfur dioxide_s3	0.0495	0.011	4.494	0.000	0.028	0.071
total sulfur dioxide_s0	-0.0021	0.009	-0.218	0.828	-0.021	0.017
total sulfur dioxide_s1	-0.0130	0.004	-3.303	0.001	-0.021	-0.005
total sulfur dioxide_s2	0.0149	0.010	1.510	0.131	-0.004	0.034
total sulfur dioxide_s3	-0.0068	0.009	-0.760	0.447	-0.024	0.011
density_s0	-0.0150	0.007	-2.238	0.025	-0.028	-0.002
density_s1	-0.0129	0.010	-1.261	0.207	-0.033	0.007
density_s2	-0.0921	0.041	-2.261	0.024	-0.172	-0.012
density_s3	-0.0918	0.024	-3.788	0.000	-0.139	-0.044
pH_s0	-0.0024	0.011	-0.229	0.819	-0.023	0.018
pH_s1	-0.0014	0.006	-0.230	0.818	-0.014	0.011
pH_s2	-0.0435	0.010	-4.387	0.000	-0.063	-0.024
pH_s3	-0.0164	0.009	-1.805	0.071	-0.034	0.001
sulphates_s0	0.0057	0.007	0.808	0.419	-0.008	0.020
sulphates_s1	-0.0021	0.004	-0.520	0.603	-0.010	0.006
sulphates_s2	0.0021	0.008	0.268	0.789	-0.013	0.017
sulphates_s3	-0.0108	0.006	-1.708	0.088	-0.023	0.002
alcohol_s0	0.0142	0.008	1.753	0.080	-0.002	0.030
alcohol_s1	0.0031	0.004	0.780	0.435	-0.005	0.011
alcohol_s2	-0.0253	0.006	-3.934	0.000	-0.038	-0.013
alcohol_s3	-0.0300	0.006	-4.897	0.000	-0.042	-0.018

Figure 2. Generalized Additive Model (Gamma) of White Wine Features

Generalized Linear Model Regression Results

Dep. Variable:	quality	No. Observations:	1279
Model:	GLMGam	Df Residuals:	1234.00
Model Family:	Gamma	Df Model:	44.00
Link Function:	InversePower	Scale:	0.012664
Method:	PIRLS	Log-Likelihood:	-1218.0
Date:	Fri, 21 Mar 2025	Deviance:	16.206
Time:	21:18:22	Pearson chi2:	15.6
No. Iterations:	5	Pseudo R-squ. (CS):	0.4802
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	0.025	0.975]
const	0.0719	0.016	4.556	0.000	0.041	0.103
fixed acidity	-0.0010	0.001	-0.703	0.482	-0.004	0.002
volatile acidity	0.0398	0.008	4.892	0.000	0.024	0.056
citric acid	-0.0106	0.008	-1.356	0.175	-0.026	0.005
residual sugar	-0.0022	0.001	-1.908	0.056	-0.004	5.95e-05
chlorides	0.0293	0.014	2.167	0.030	0.003	0.056
free sulfur dioxide	-0.0002	0.000	-1.494	0.135	-0.001	7.34e-05
total sulfur dioxide	8.524e-05	4.46e-05	1.913	0.056	-2.1e-06	0.000
density	0.0723	0.016	4.594	0.000	0.041	0.103
pH	0.0443	0.009	4.824	0.000	0.026	0.062
sulphates	-0.0099	0.007	-1.496	0.135	-0.023	0.003
alcohol	-0.0013	0.003	-0.503	0.615	-0.006	0.004
fixed acidity_s0	-0.0169	0.012	-1.465	0.143	-0.040	0.006
fixed acidity_s1	-0.0195	0.008	-2.454	0.014	-0.035	-0.004
fixed acidity_s2	-0.0160	0.011	-1.416	0.157	-0.038	0.006
fixed acidity_s3	-0.0071	0.009	-0.756	0.450	-0.025	0.011
volatile acidity_s0	-0.0106	0.008	-1.275	0.202	-0.027	0.006
volatile acidity_s1	-0.0013	0.007	-0.182	0.855	-0.015	0.013
volatile acidity_s2	-0.0414	0.017	-2.408	0.016	-0.075	-0.008
volatile acidity_s3	0.0543	0.017	3.239	0.001	0.021	0.087
citric acid_s0	-0.0011	0.004	-0.273	0.785	-0.009	0.007
citric acid_s1	0.0142	0.006	2.448	0.014	0.003	0.026
citric acid_s2	0.0153	0.016	0.969	0.332	-0.016	0.046
citric acid_s3	-0.0280	0.018	-1.547	0.122	-0.063	0.007
residual sugar_s0	-0.0244	0.011	-2.207	0.027	-0.046	-0.003
residual sugar_s1	0.0008	0.010	0.080	0.936	-0.019	0.020
residual sugar_s2	-0.0223	0.018	-1.263	0.206	-0.057	0.012
residual sugar_s3	0.0114	0.011	1.021	0.307	-0.010	0.033
chlorides_s0	0.0029	0.014	0.204	0.838	-0.025	0.031
chlorides_s1	0.0134	0.013	1.002	0.316	-0.013	0.040
chlorides_s2	-0.0184	0.022	-0.838	0.402	-0.061	0.025
chlorides_s3	0.0553	0.020	2.716	0.007	0.015	0.095
free sulfur dioxide_s0	-0.0047	0.007	-0.648	0.517	-0.019	0.010
free sulfur dioxide_s1	-0.0087	0.007	-1.323	0.186	-0.022	0.004
free sulfur dioxide_s2	0.0121	0.011	1.073	0.283	-0.010	0.034
free sulfur dioxide_s3	-0.0061	0.008	-0.790	0.429	-0.021	0.009
total sulfur dioxide_s0	0.0029	0.005	0.552	0.581	-0.007	0.013
total sulfur dioxide_s1	-0.0039	0.010	-0.402	0.687	-0.023	0.015
total sulfur dioxide_s2	0.0265	0.015	1.718	0.086	-0.004	0.057
total sulfur dioxide_s3	-0.0188	0.010	-1.909	0.056	-0.038	0.001
density_s0	0.0353	0.015	2.311	0.021	0.005	0.065
density_s1	0.0281	0.013	2.107	0.035	0.002	0.054
density_s2	0.0431	0.017	2.518	0.012	0.010	0.077
density_s3	0.0260	0.019	1.358	0.174	-0.012	0.064
pH_s0	-0.0277	0.024	-1.161	0.246	-0.074	0.019
pH_s1	-0.0311	0.014	-2.245	0.025	-0.058	-0.004
pH_s2	-0.0501	0.020	-2.564	0.010	-0.088	-0.012
pH_s3	-0.0601	0.020	-3.019	0.003	-0.099	-0.021
sulphates_s0	-0.0045	0.009	-0.522	0.602	-0.021	0.012
sulphates_s1	-0.0605	0.008	-7.512	0.000	-0.076	-0.045
sulphates_s2	0.0248	0.016	1.592	0.111	-0.006	0.055
sulphates_s3	-0.0167	0.010	-1.591	0.112	-0.037	0.004
alcohol_s0	-0.0456	0.015	-3.123	0.002	-0.074	-0.017
alcohol_s1	-0.0231	0.008	-2.813	0.005	-0.039	-0.007
alcohol_s2	-0.0981	0.016	-6.031	0.000	-0.130	-0.066
alcohol_s3	-0.0046	0.014	-0.318	0.750	-0.033	0.024

Figure 3. Generalized Additive Model (Gamma) of Red Wine Features

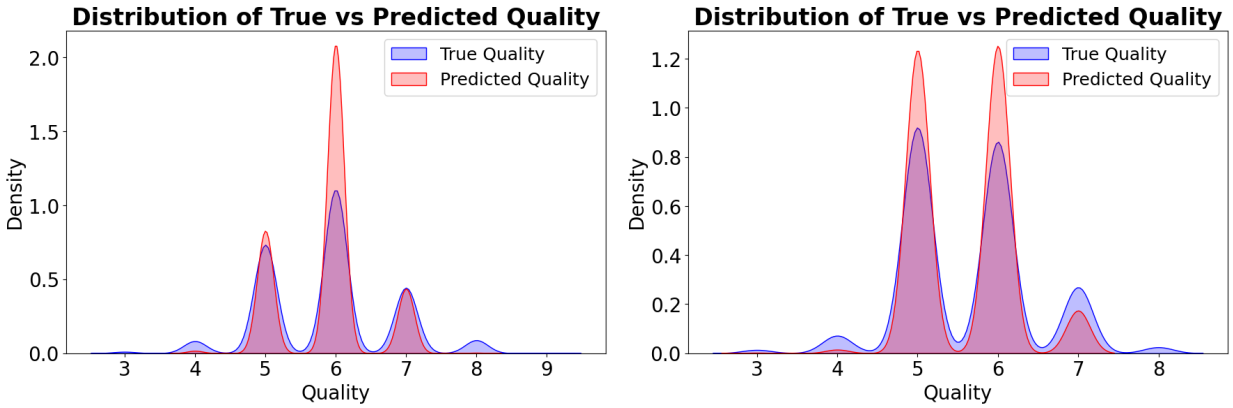

Figure 4. Generalized Additive Model (Gamma) Distribution of True vs. Predicted
(Left: White Wine; Right: Red Wine)

Predictive Modeling (Random Forest)

The Random Forest model not only provides a robust method for evaluating feature importance, allowing us to identify which physicochemical properties have the most significant impact on wine quality, but also serves as a powerful predictive model. The feature importance scores derived from the trained models indicate the relative importance of each feature in predicting wine quality, measured by the decrease in impurity (e.g., Gini impurity) across all splits that include the feature.

The Random Forest classifier was employed to predict wine quality for both white and red wine datasets. For white wine, the Random Forest classifier achieved an accuracy of 71.12%, indicating that the model has a reasonable ability to predict wine quality for white wines. For red wine, the Random Forest classifier achieved an accuracy of 65.00%. The Random Forest model performed better on white wines compared to red wines, with a higher accuracy and lower misclassification rate for white wines. This difference could be attributed to variations in the physicochemical properties and their relationships with wine quality between the two types of wines.

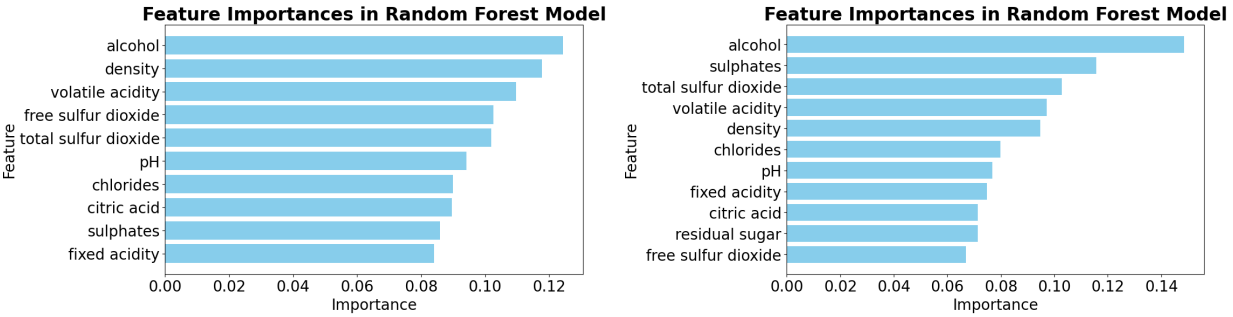

Figure 5. Feature Importances in Random Forest Model (Left: White Wine; Right: Red Wine)

The feature importance analysis reveals distinct patterns in the physicochemical properties that significantly influence wine quality for white and red wines (Figure 5). For white wines, alcohol content emerges as the most important feature, contributing 12.44% to the model's performance, followed closely by density (11.77%) and volatile acidity (10.96%). These features highlight the critical role of alcohol and density in determining the perceived quality of white wines. In contrast, for red wines, alcohol content is even more dominant, contributing 14.86% to the model's performance, while sulphates (11.56%) and total sulfur dioxide (10.27%) play significant roles. This suggests that sulphates and sulfur dioxide levels are particularly influential in red wines, likely due to their impact on the wine's sensory attributes and preservation. The differences in feature importance underscore the unique characteristics and production processes of white and red wines, indicating that optimizing these specific physicochemical properties can enhance wine quality.

Section IV: Conclusion

Based on our analysis, we can draw several conclusions regarding the outlined hypotheses and provide recommendations for future work.

Hypothesis Conclusions

1. Physicochemical Properties and Wine Quality
   Our findings support the hypothesis that physicochemical properties such as alcohol content, density, and volatile acidity significantly predict wine quality. The importance of these features varies between white and red wines, indicating that different properties influence the prediction model of each wine type.

2. Differences in Physicochemical Properties
   The analysis confirms significant differences in physicochemical properties between red and white wines. These differences are crucial for understanding how production methods and chemical composition impact wine quality

3. Correlation with Wine Quality
   Specific physicochemical features are more strongly correlated with wine quality than others. For white wines, alcohol content, density, and volatile acidity are most influential, while for red wines, alcohol content, sulphates, and total sulfur dioxide are key predictors.

Important Variables

The most influential feature for both white and red wines is `alcohol`, highlighting its critical role in determining wine quality. For white wine, `density` and `volatile acidity` are important, suggesting their impact on sensory attributes. For red wine, `sulphates` and `total sulfur dioxide` are significant, likely due to their effects on sensory attributes and preservation.

Most Appropriate Model

For predicting the distribution of wine quality, the GAM with the Gamma family is the most appropriate model, offering the best balance between model complexity and goodness of fit. If the focus is on classification without needing the distribution, the Random Forest model is the best choice, providing high accuracy and feature importance insights.

Statistical Assumptions and Model Performance

The non-normal distribution of both white and red wine features violates the assumptions of many traditional statistical models, potentially affecting model performance. This highlights the importance of using models that do not rely on normality assumptions, such as GAM and Random Forest.

Difficulties Faced when Modelling

When fitting Gaussian Mixture Models (GMM) to our data, we encountered a significant challenge: the misclassification rate reached 60-70%. This is largely related to the curse of dimensionality. With 11 features in our dataset, the sparsity of data in high-dimensional space makes it difficult for GMM to accurately estimate the parameters of Gaussian distributions, leading to degraded model performance. Additionally, complex relationships between features (e.g., nonlinear dependencies) in high-dimensional spaces may not be well-captured by simple Gaussian assumptions, further reducing the model's predictive power.

Shortcomings of the Current Model

While the GAM with the Gamma family is the most appropriate model to predict the distribution of wine quality for both white and red wine, the performance of both GLM and GAM models can be sensitive to outliers, which might skew the results if not properly handled. Moreover, the current models may not fully capture the complex interactions between different physicochemical properties that could influence

wine quality. The models might not generalize well to datasets from different regions or vintages due to variations in wine production practices and climate conditions.

In order to further improve the current models, future work would be focus on:

- Model Refinement: Further refine the GAM and Random Forest models by tuning hyperparameters and exploring more sophisticated feature engineering techniques.
- Feature Interaction Analysis: Investigate the interactions between different physicochemical properties to better understand their combined effects on wine quality.
- Outlier Treatment: Implement outlier detection and treatment methods to improve the robustness of the models.

Interesting Findings

- The significant differences in feature importance between white and red wines underscore the unique characteristics of each wine type.
- The GAM with the Gamma family consistently outperformed other models, indicating its effectiveness in capturing the relationship between predictors and wine quality.

In conclusion, our analysis has provided valuable insights into the complex relationship between physicochemical properties and wine quality. The findings underscore the importance of `alcohol` across both white and red wines, while also highlighting the unique influences of other properties such as `density`, `volatile acidity`, `sulphates`, and `total sulfur dioxide`. The GAM with the Gamma family emerged as the most suitable model for predicting wine quality distributions, while Random Forest offered a robust alternative for classification tasks. Despite the promising results, our models faced challenges, particularly with respect to outlier sensitivity and the capture of complex feature interactions. Future work should focus on refining these models, investigating feature interactions, and addressing outlier treatment to enhance predictive accuracy and model robustness. The interesting findings from this study not only advance our understanding of wine quality prediction but also provide actionable insights for winemakers aiming to optimize production processes and enhance the sensory attributes of their wines.

References

1. P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Wine quality," *UCI Machine Learning Repository*, 2009. [Online]. Available: https://doi.org/10.24432/C56S3T.
2. P. Cortez, F. Almeida, and T. Matos, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547-553, 2009.
3. A. Luque, M. Mazzoleni, F. Zamora-Polo, A. Ferramosca, J. R. Lama, and F. Previdi, "Determining the importance of physicochemical properties in the perceived quality of wines,"

Dept. de Ingeniería del Diseño, Escuela Politécnica Superior, Universidad de Sevilla, Seville, Spain, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10287348

4. Krishnamurthy, Akshay. "High-dimensional clustering with sparse gaussian mixture models." Unpublished paper (2011): 191-192.