Group 12 Project Report

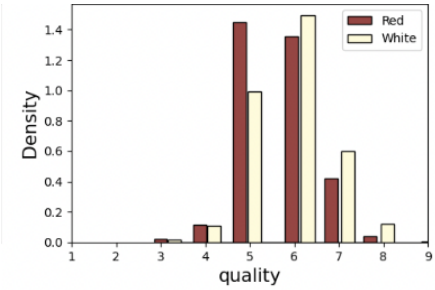
Noah Eisenberg, Matteo Magcalas, Gavin Fancher, Ethan Ferger, Michael Gentleman

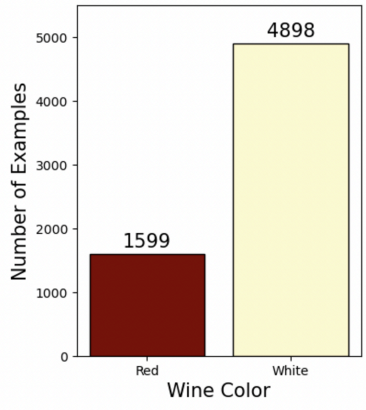
Introduction

We chose to investigate how to distinguish between red and white wine when given different measurements of the contents of the respective wines. Our investigative questions include: What features are most indicative of wine quality? Are those features different for white and red wine respectively? What features differentiate red wine from white wine? While attempting various models, we found that k-NN and logistic regression were the most effective classifiers between red and white wine while random forest regression was best for selecting indicative features.

Our Data[1]

Our dataset consisted of two sub-datasets: **Red Wine** and **White Wine**.





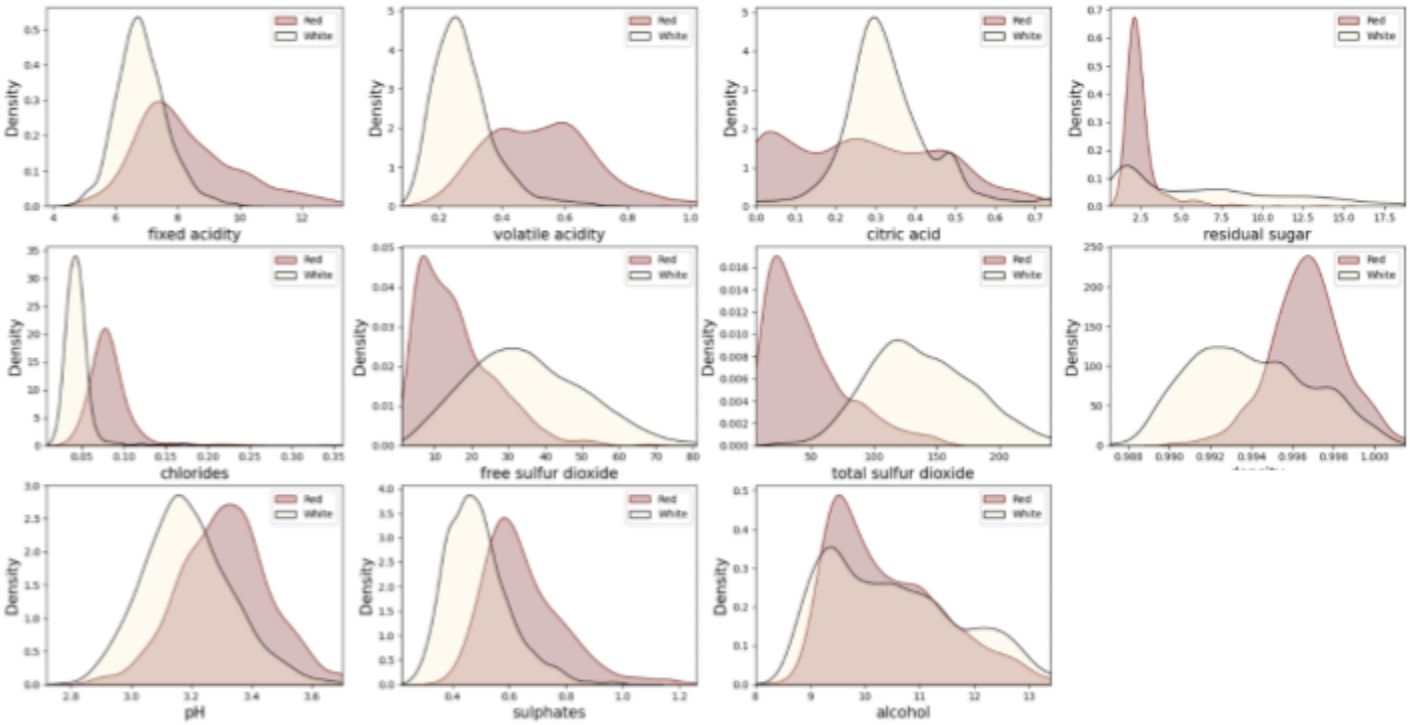
Features

- Fixed acidity (g(tartaric acid)/L) (FA)
- Volatile acidity (g(acetic acid)/L) (VA)
- Citric acid (g/L)
- Residual sugar (g/L) (RS)
- Chlorides (g(sodium chloride)/L)
- Free sulfur dioxide (mg/L) (FSD)
- Total Sulfur Dioxide (mg/L) (TSD)
- Density (g/mL)
- pH
- Sulphates (g(potassium sulphate)/L)
- Alcohol (% vol)

Target Label: Quality (1-10), Type (Red, White)

1. https://www.kaggle.com/datasets/xuzihe2010/wine-quality-red?select=winequality-white.csv

Distribution of Features



Wine Classification: White or Red?

Linear SVM

- Multi-dimensional linear SVM model, c=1
- Linear model was chosen after running grid search to determine model type and hyperparameters
- Model had a 97% accuracy predicting Red v. White wine when run on unseen test data
- Used permutation importance on this model to determine which variables were the most important in the success of the model
    - *Fixed Acidity* and *chlorides*
- Ran a two-variable linear SVM with c=1, on these features alone, seeing a 93% accuracy on seen test data

Decision Tree

- Used decision tree classifier with criterion of entropy and no max depth
- 98% accuracy with a depth of 12, first node was *chlorides*

- *Total sulfur dioxide* levels are regulated by wine type
- When removed regulated feature, depth: 14 and accuracy: 97%
- Second level node changed from *total sulfur dioxide* to *volatile acid* and *residual sugar*

kNN Classification

- Used a kNN model (k = 5) to predict both quality and wine type from the chemical features in the data.
- Explored whether there were any relationships between certain chemical properties and these target variables.

Wine Quality:

- Performance hindered by class imbalance and the complexity of the task.
- Performed best for quality labels 5 and 6 (most frequent classes), achieving recall scores above 50%.
- More extreme labels were more difficult to predict accurately.
    - Label 9 had no correct predictions.
- Overall accuracy: 54%

Wine Type:

- Excelled in predicting wine type, with or without regulated features
- Distinguished between red and white wines effectively (accuracy of 99% for both types).
    - Suggests that other features (i.e. density, volatile acidity, residual sugar) played critical roles in success.
- Precision, recall, and F1-scores were consistently high across both wines.

Logistic Regression

Wine Type

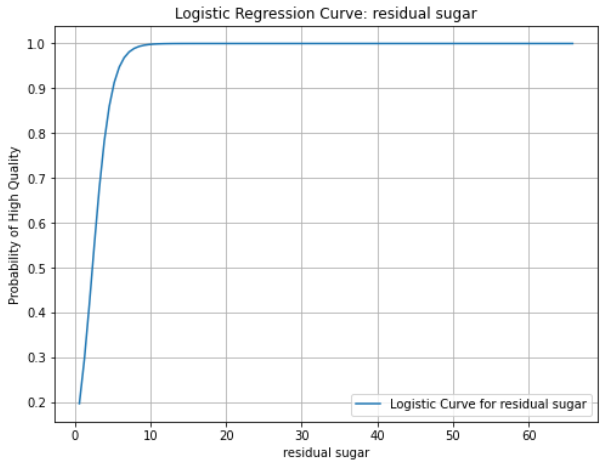
- With and without regulated features, accuracy was 99%

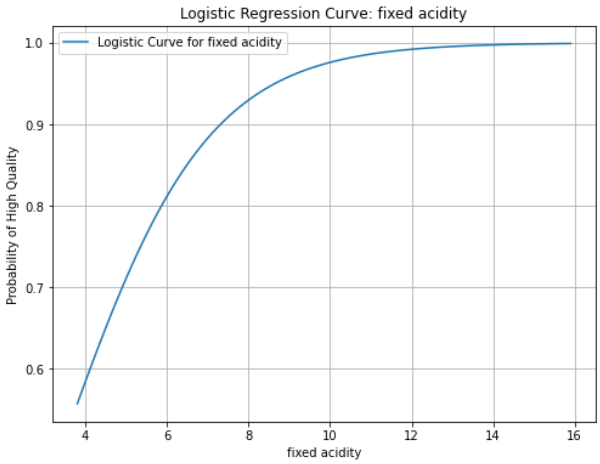
Distinguishing Features

- *Residual sugar* was the best way to distinguish a white wine
- *Density* and *alcohol* were best to distinguish a red wine

Wine Quality

- Seen below, *residual sugar* and *fixed acidity* significantly contribute to predicting quality (>6)





Wine Quality (Regression):

Quality vs Features

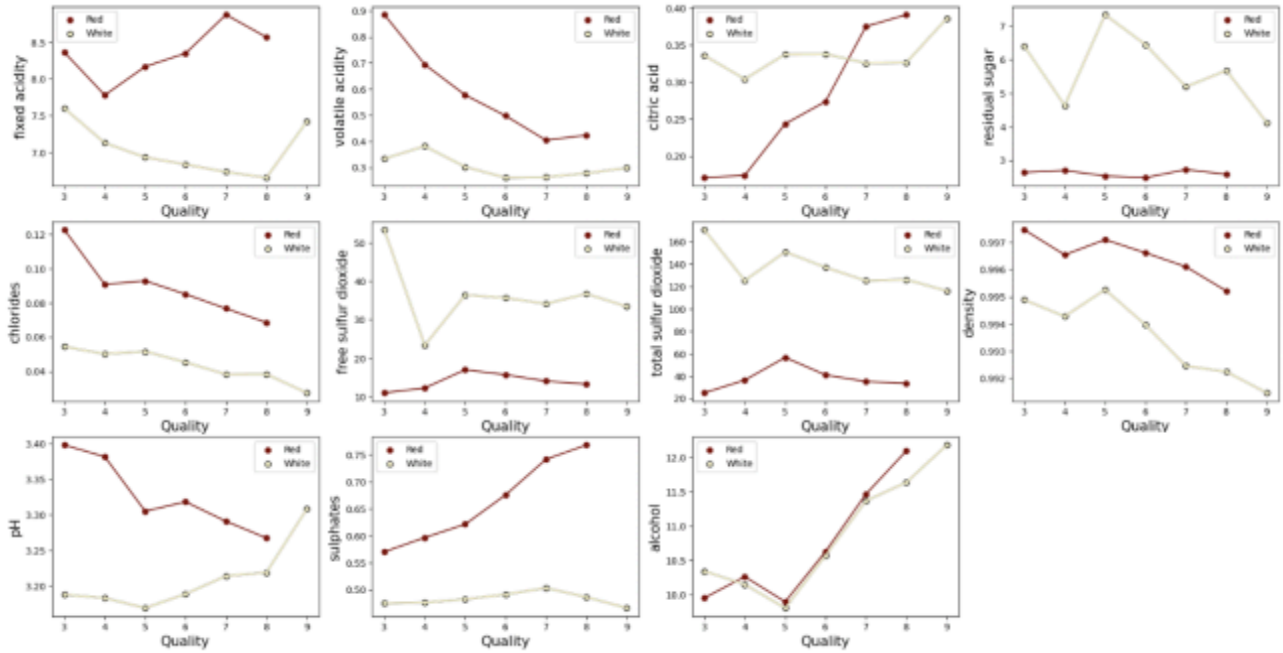
Positive trends:

- Alcohol (most notable for both wine colors)
- Sulphates and citric acid (red wine)

Negative trends:

- Chlorides
- Density

Model Training/Testing



- Training: 80%, Testing: 20%
- Hyperparameter Tuning: Grid Search (5-Fold Cross Validation)

Model	Hyperparameters	Metrics
Linear Regression	N/A	color rmse r2 red 0.620057 0.328389 white 0.812309 0.251348
Lasso Regression	Alpha: 0.02 Features Used • Red: FA/VA, Sulphates, Alcohol • White: FA, RS Alcohol	color rmse r2 red 0.628015 0.311041 white 0.848582 0.182994
Random Forest	Red: Max_Depth: 20 / N_Estimators: 200 White: Max_Depth: 20 / N_Estimators: 175	color rmse r2 red 0.559744 0.452689 white 0.691112 0.458080
kNN	Distance: Manhattan / Neighbors: 7 Distance: Manhattan / Neighbors: 7	color rmse r2 red 0.697847 0.149305 white 0.867347 0.146460

Best Model: Random Forest Regression

Red Wine:

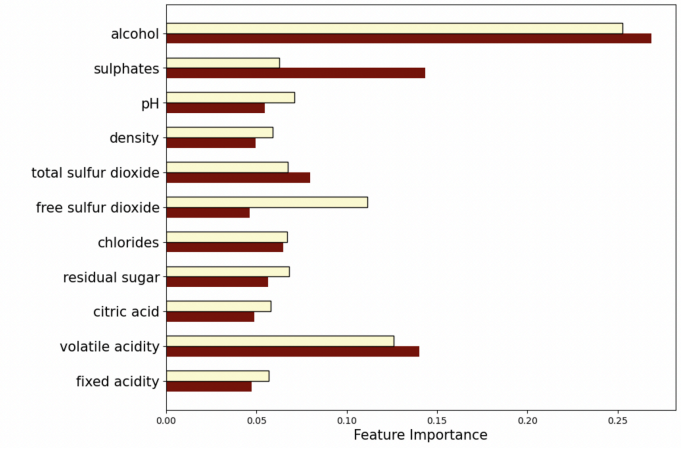
Notable features: **Alcohol, Sulphates, VA**.

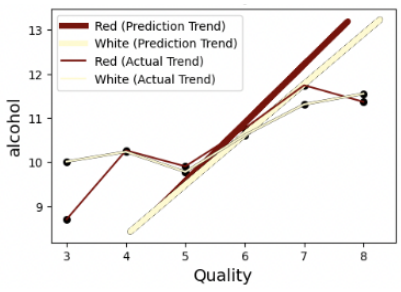
- Alcohol and sulphates have a positive correlation with quality, and VA has a negative correlation. These trends were similar to the anticipated results.

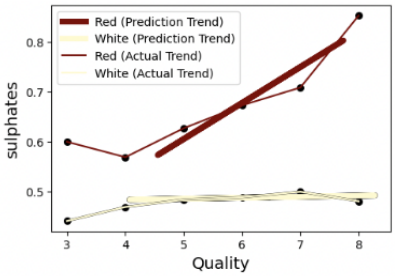
White Wine:

Notable features:**Alcohol, VA**, **FSD**.

- Alcohol and FSD have a positive correlation with quality, and VA has a negative correlation. The model exhibited more extreme correlation on VA and FSD than expected.

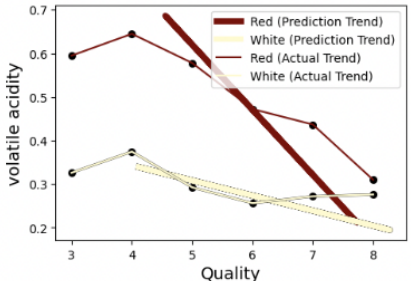


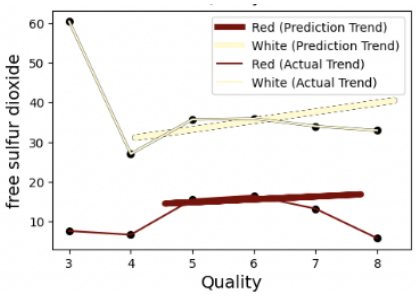


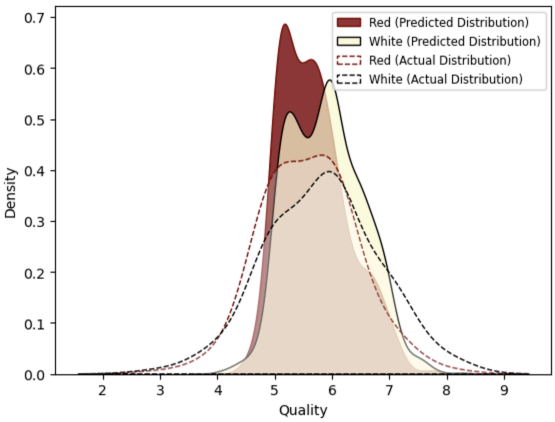


Note: Failure of Model to Predict Outliers

The distribution failed to predict outlier outcomes of quality < 4 , quality > 8.







Conclusion

We found we can best predict the classification of wines using a k-NN and logistic regression model. The quality of the wines is best determined using random forest regression. Future areas of improvement: Correctly classify a rosé, which is a mix of white and red wine.

Member	Proposal	Coding	Presentation	Report
Michael	1	1	1	1
Noah	1	1	1	1
Matteo	1	1	1	1
Gavin	1	1	1	1
Ethan	1	1	1	1