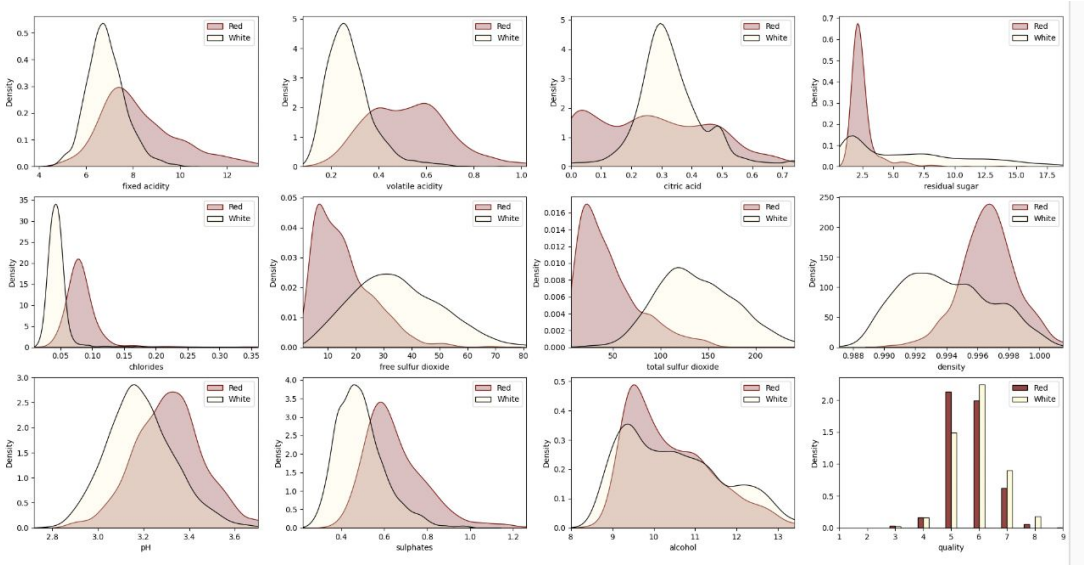
STAT 451 Project: Wine

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

Exploratory Data Analysis: By Feature



Key Observations:

Acidity levels

- **White:** more consistent, lesser overall
- **Red:** More overall, less consistent

Sulfur Dioxide/Density

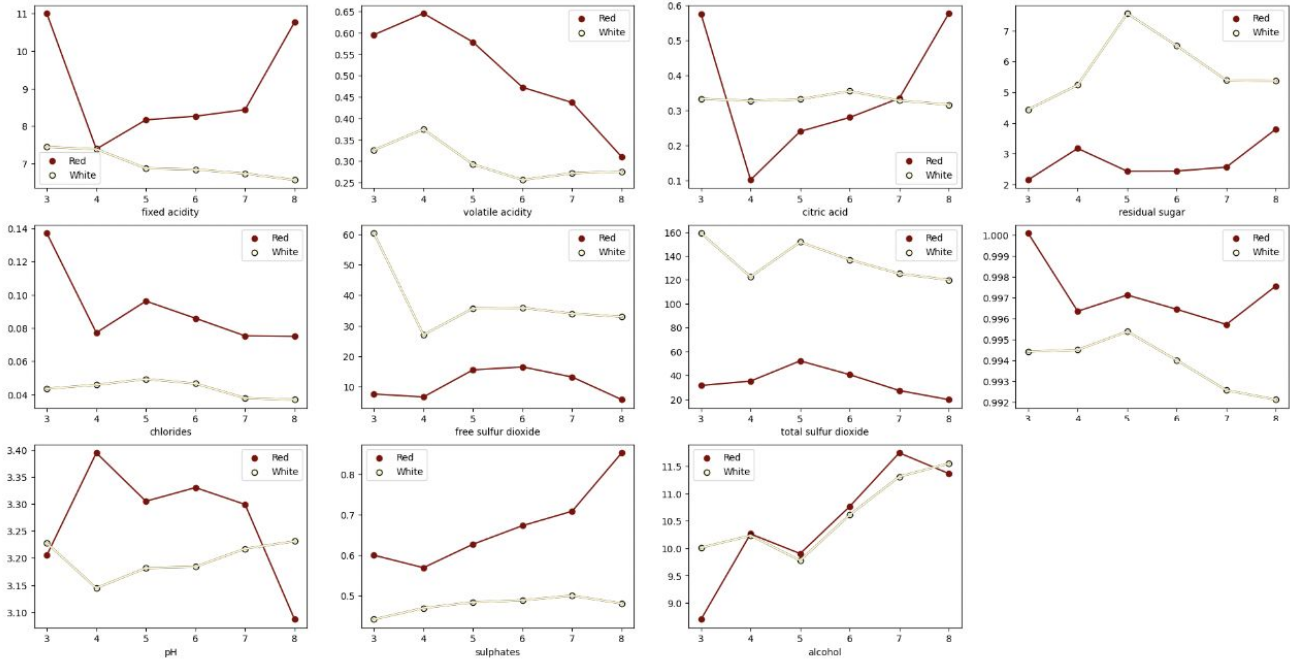
- **White:** Less consistent, more overall
- **Red:** More consistent, lesser overall

Quality

- Similar Distributions
- White slightly higher quality on average

Exploratory Data Analysis: Feature vs Quality

Note: Each point represents the mean of each feature at each level of quality



Key Observations:

Alcohol

- Most notable positive correlation among both colors

Many features show similar trends across both colors

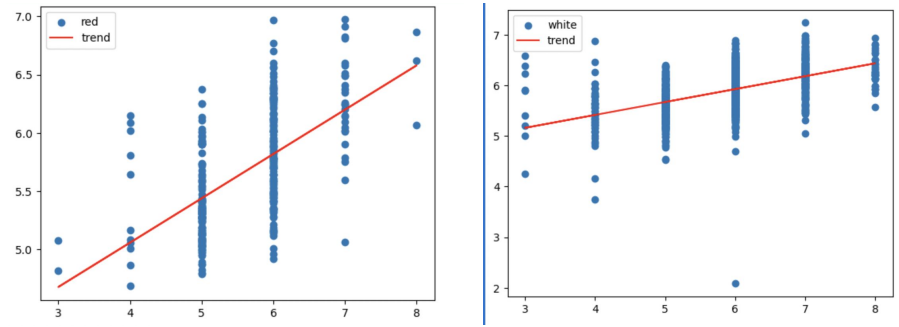
- At varying quantities
- Similar Distributions

Some features show deviation from trend at outliers

- Quality levels of 3 and 8 are higher/lower than rest of trend

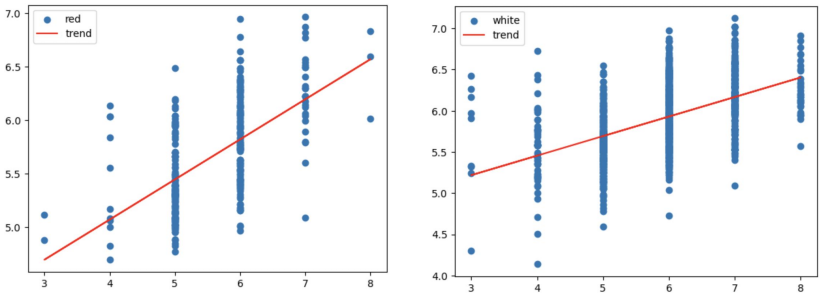
Regression Models: Linear Regression and Lasso

Linear Regression



color	rmse	r2
red	0.620057	0.328389
white	0.812309	0.251348

Lasso Regression



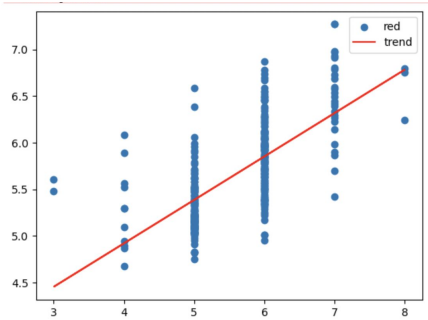
	color	rmse	r2
0	red	0.619194	0.330259
1	white	0.814208	0.247843

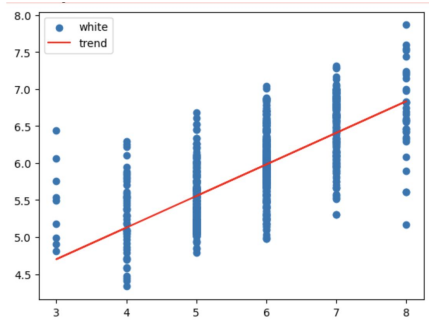
Note: Feature Selection (Coefficient < 0.01)

- Red: Free sulfur dioxide, total sulfur dioxide, density
- White: Free sulfur dioxide, total sulfur dioxide, density, chlorides, citric acid

Regression Models: Random Forest and kNN

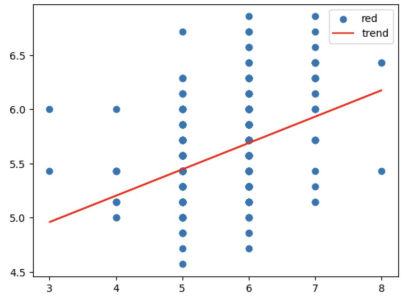
Random Forest

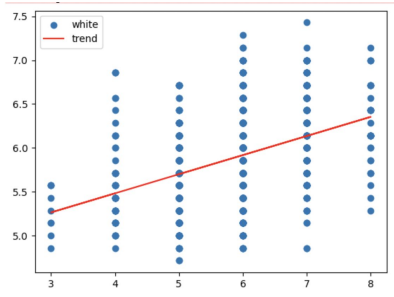




color	rmse	r2
red	0.559744	0.452689
white	0.691112	0.458080

kNN

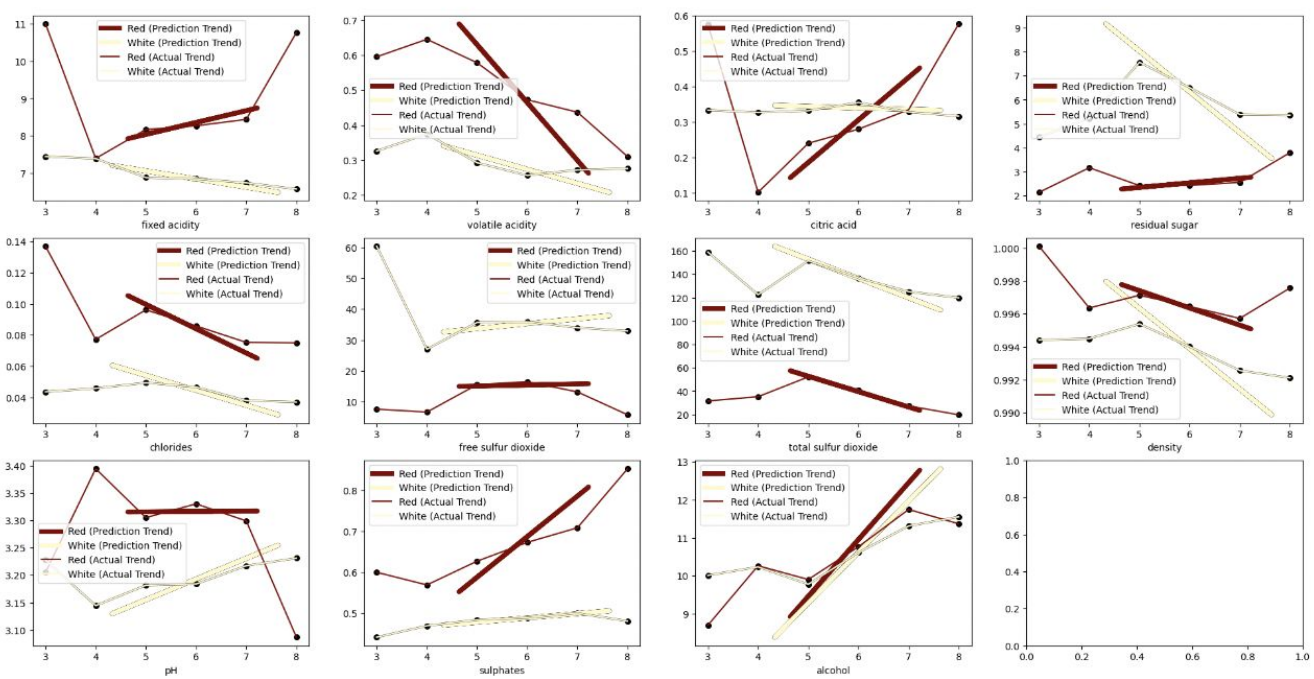




color	rmse	r2
red	0.697847	0.149305
white	0.867347	0.146460

Best Model: Random Forest Regression

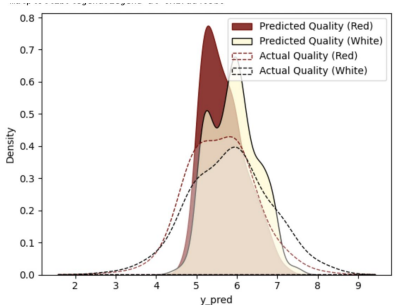
Note: Each point represents the mean of each feature at each level of quality



color	rmse	r2
red	0.559744	0.452689
white	0.691112	0.458080

Key Observations:

Model typically mimics overall trends when quality is between 5-7



Model fails to predict outlier values of <4 and >8

Decision Tree Classification

classification report for wine type with all features

	precision	recall	f1-score	support
red	0.95	0.98	0.96	626
white	0.99	0.98	0.99	1973
accuracy			0.98	2599
macro avg	0.97	0.98	0.98	2599
weighted avg	0.98	0.98	0.98	2599

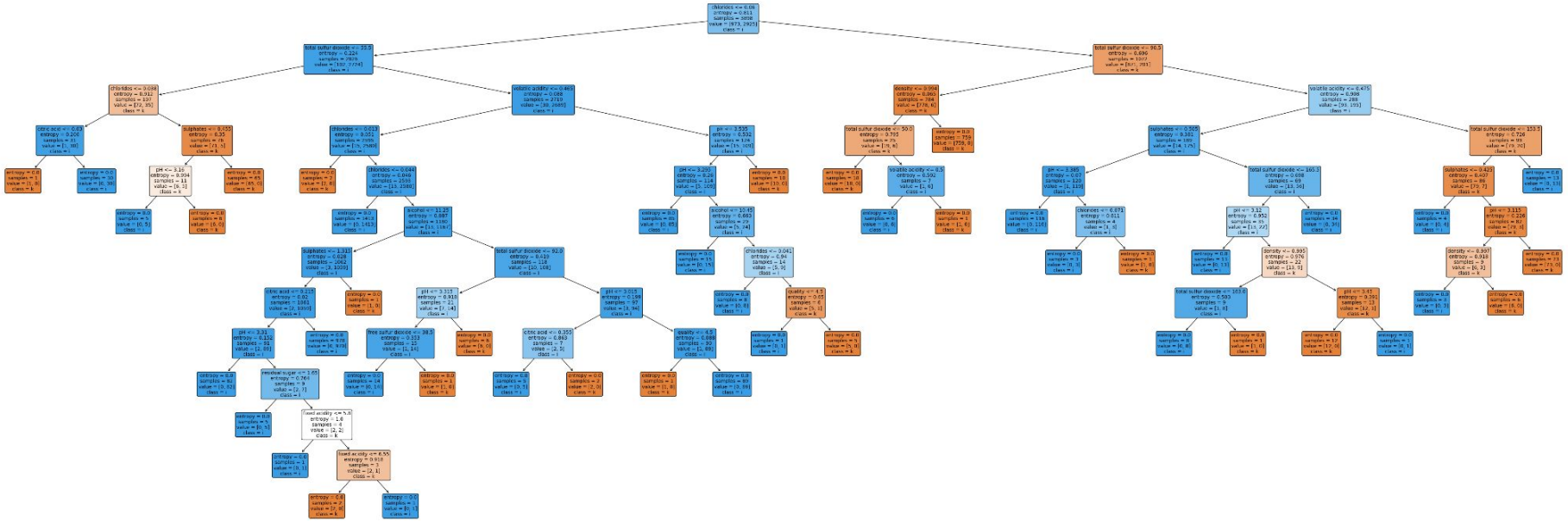
classification report for wine type with only non-regulated features

	precision	recall	f1-score	support
red	0.95	0.95	0.95	626
white	0.98	0.98	0.98	1973
accuracy			0.97	2599
macro avg	0.97	0.97	0.97	2599
weighted avg	0.98	0.97	0.98	2599

- 75.4% of data is on white wine
- Used criterion of entropy and no max depth
- Regulations for sulfur dioxide content in different wine types
- Model performs slightly worse with regulated features removed

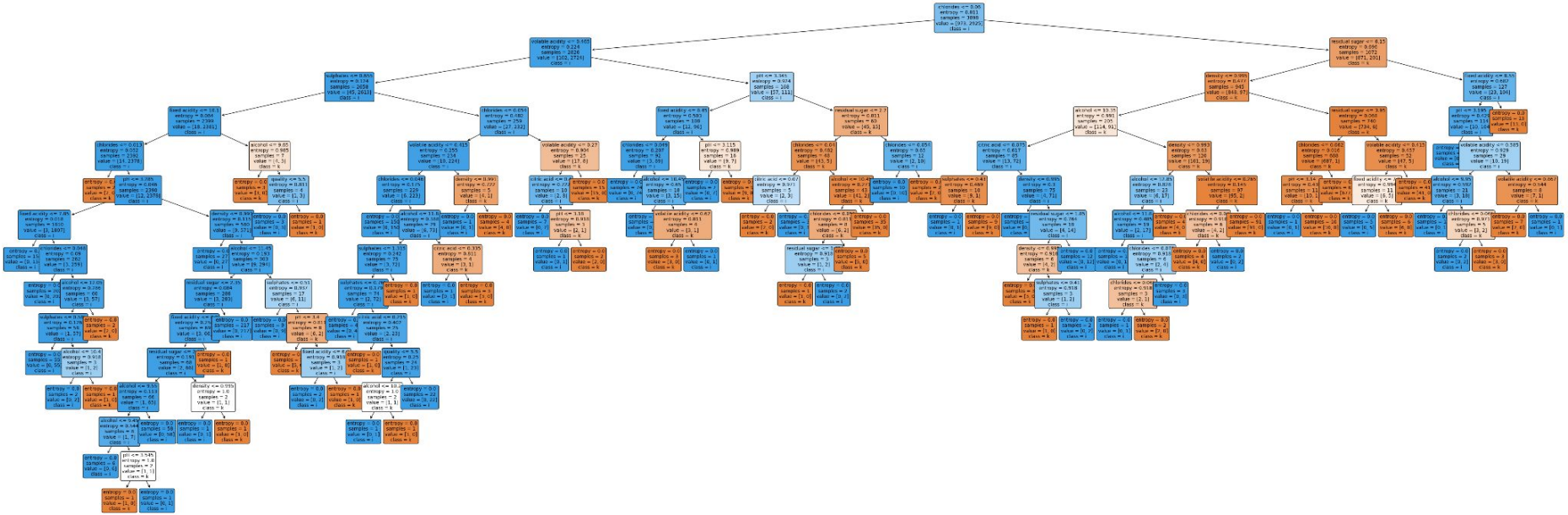
Decision Tree Classification

All features, depth of 12



Decision Tree Classification

Only unregulated features, depth of 14



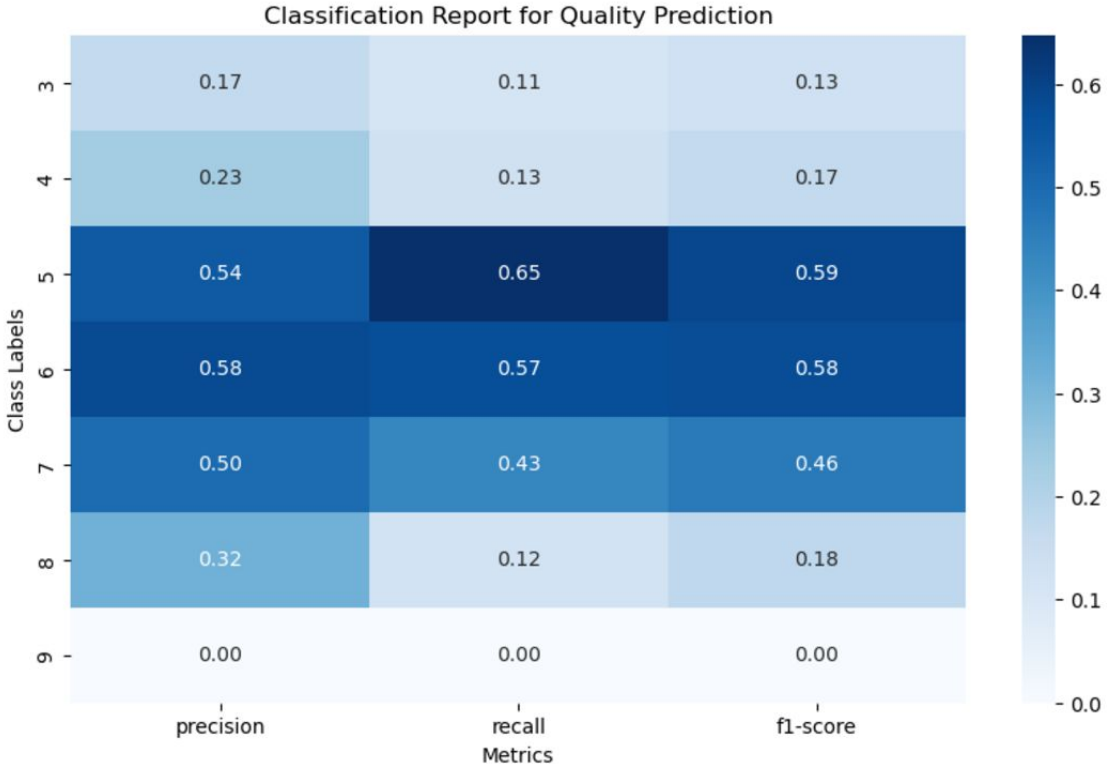
kNN Classification (Quality Prediction)

Classification Report for Quality Prediction:

	precision	recall	f1-score	support
3	0.17	0.11	0.13	9
4	0.23	0.13	0.17	69
5	0.54	0.65	0.59	613
6	0.58	0.57	0.58	894
7	0.50	0.43	0.46	315
8	0.32	0.12	0.18	49
9	0.00	0.00	0.00	1
accuracy			0.54	1950
macro avg	0.33	0.29	0.30	1950
weighted avg	0.53	0.54	0.54	1950

- Most frequent quality labels: 5 and 6 (model also performs well with these labels, with recalls >50% for both)
- Performs poorly for rare quality labels (3, 4, 8, 9)
- Overall model accuracy: 54%

kNN Classification (Quality Prediction)



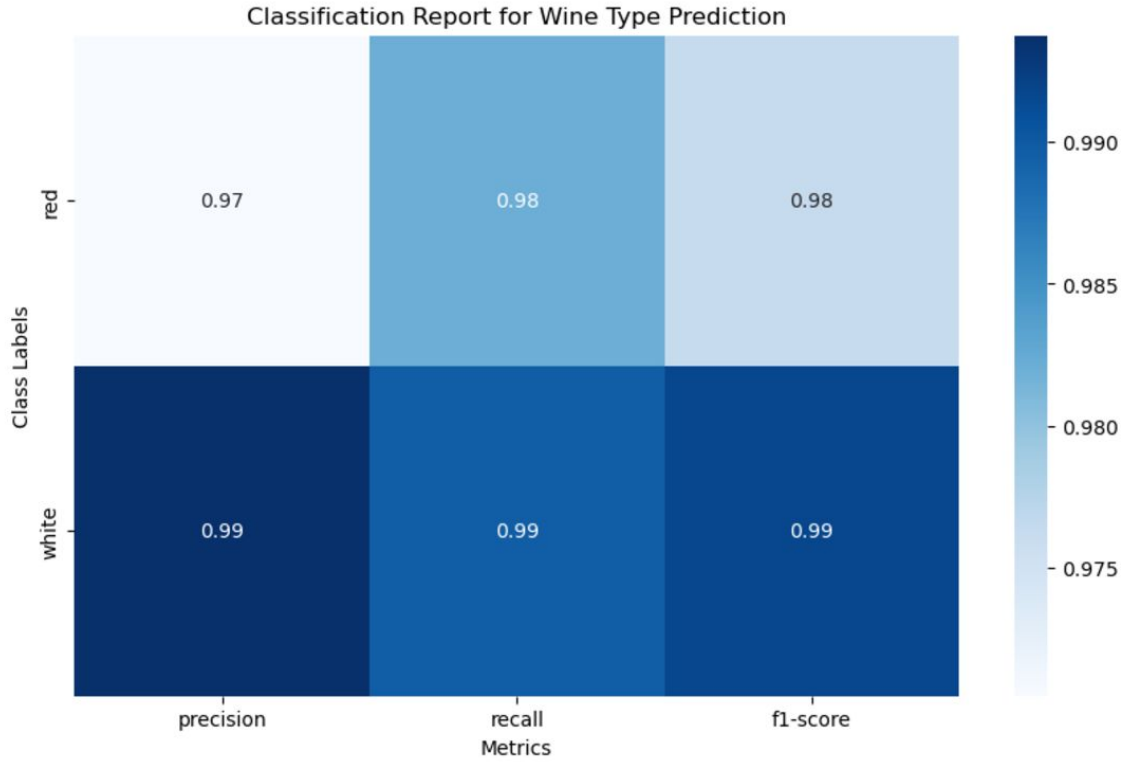
kNN Classification (Wine Type Prediction)

Classification Report for Wine Type Prediction:

	precision	recall	f1-score	support
red	0.97	0.98	0.98	502
white	0.99	0.99	0.99	1448
accuracy			0.99	1950
macro avg	0.98	0.99	0.98	1950
weighted avg	0.99	0.99	0.99	1950

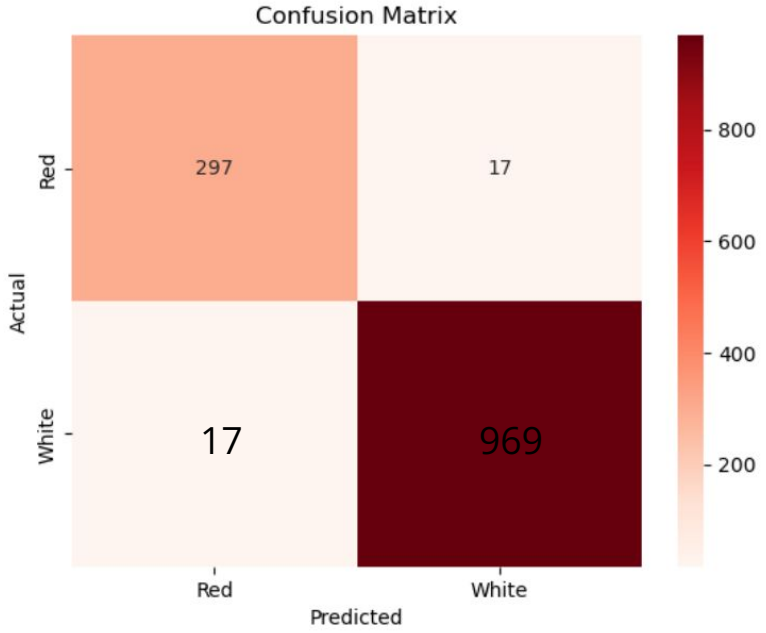
- Model accuracy for both wine types: ~99%
- Model easily distinguishes between red and white wines, due to features like density, volatile acidity, and/or residual sugar that likely play a role

kNN Classification (Wine Type Prediction)



SVM Classification

- Linear SVM with C=1 in order to help avoid overfitting over many features
- Model accuracy at 97%, slightly better at predicting red wines
- Total of 34 incorrect predictions from the model on 1300 test data points

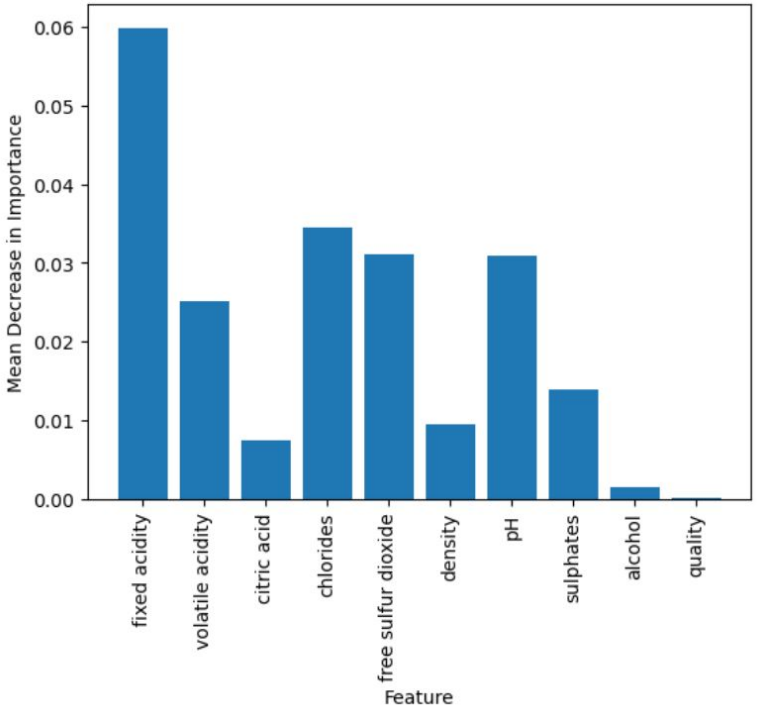


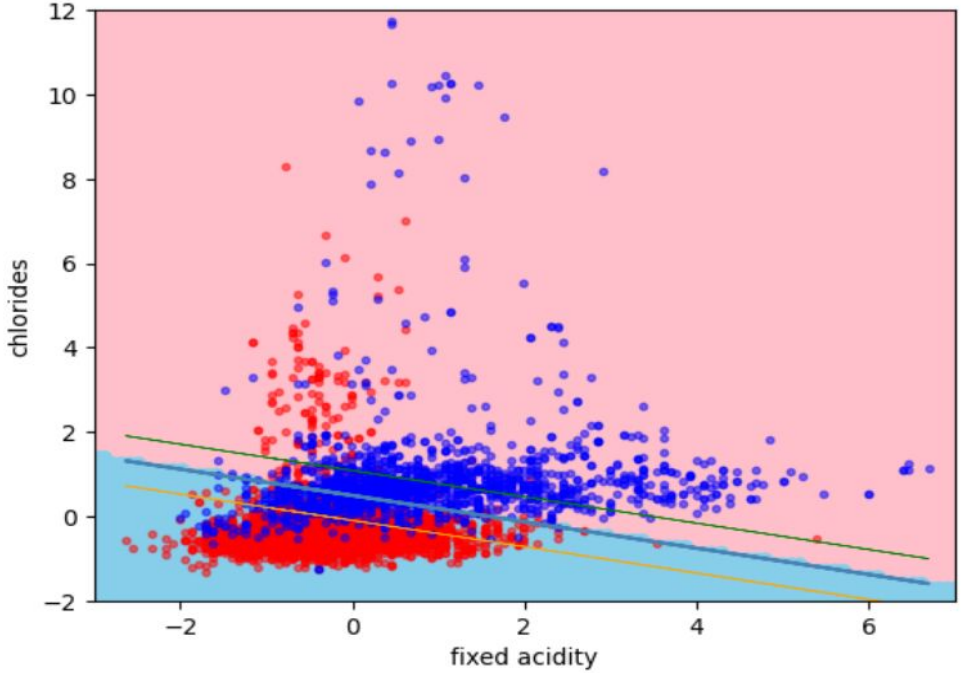
Classification Report:

	precision	recall	f1-score	support
0	0.95	0.95	0.95	314
1	0.98	0.98	0.98	986
accuracy			0.97	1300
macro avg	0.96	0.96	0.96	1300
weighted avg	0.97	0.97	0.97	1300

SVM Classification

- Determined ‘Fixed Acidity’ and ‘Chlorides’ were best predictors through permutation importance
- Linear SVM model accuracy of 93%





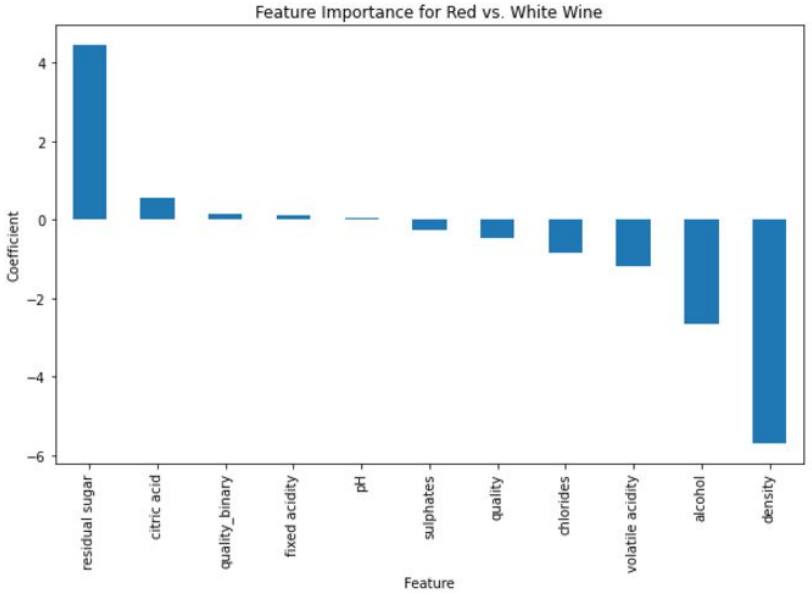
Logistic Regression

Classification Report for Wine Type Prediction with all features

	precision	recall	f1-score	support
red	0.98	1.00	0.99	320
white	1.00	0.99	1.00	980
accuracy			1.00	1300
macro avg	0.99	1.00	0.99	1300
weighted avg	1.00	1.00	1.00	1300

Classification Report for Wine Type Prediction with non-regulated features

	precision	recall	f1-score	support
red	0.95	0.99	0.97	320
white	1.00	0.98	0.99	980
accuracy			0.99	1300
macro avg	0.98	0.99	0.98	1300
weighted avg	0.99	0.99	0.99	1300



- Accuracy for the model is very high and is able to distinguish between the white and red wines
- The positive coefficients for feature importance indicate those features are associated with predicting white wine while negative is red wine.

Logistic Regression

Some Interesting logistic regression curves

- Residual sugar and fixed acidity significantly contribute to predicting quality

- As pH increases the probability of higher quality wine also increases but only slightly