Report

For this project, we conducted data analyses on a wine quality dataset from UC Irvine's machine learning repository; the data was collected in a 2009 study on wine from Minho, Portugal. The dataset comprises 12 variables: fixed acidity (a wines natural acids), volatile acidity (gaseous acids in wine), citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol (abv), quality, and color (red or white).

We had two questions of interest: (1) is there a separable relationship between red and white wines based on the variables given and (2) can we accurately predict the quality. For the first question, we utilized decision trees and support vector machines (SVMs) to classify wines by color based on chemical composition variables. To approach question two, we compared four different models –SVM, Linear Regression, K Nearest Neighbors, and Random Forest. – for predicting wine quality rating, using GridSearch to find the most accurate model.

The dataset, originally from separate tables for white and red wines, was merged into a single dataset of 6497 rows and 13 columns. Although the data was clean, we used one hot encoding to convert wine color into a numerical category.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	color
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	red
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5	red
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5	red
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6	red
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	red
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6492	6.2	0.21	0.29	1.6	0.039	24.0	92.0	0.99114	3.27	0.50	11.2	6	white
6493	6.6	0.32	0.36	8.0	0.047	57.0	168.0	0.99490	3.15	0.46	9.6	5	white
6494	6.5	0.24	0.19	1.2	0.041	30.0	111.0	0.99254	2.99	0.46	9.4	6	white
6495	5.5	0.29	0.30	1.1	0.022	20.0	110.0	0.98869	3.34	0.38	12.8	7	white
6496	6.0	0.21	0.38	0.8	0.020	22.0	98.0	0.98941	3.26	0.32	11.8	6	white

6497 rows × 13 columns

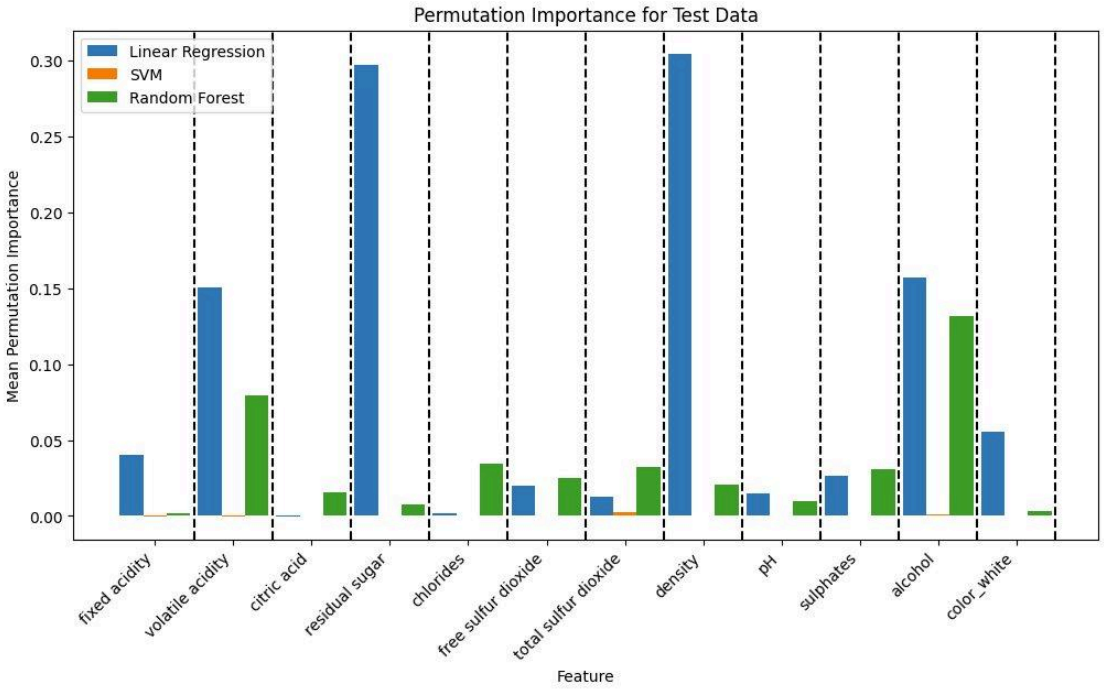
To address question one regarding color separability, we used a decision tree to classify wine color based on their chemical makeup. We utilized a train test split of the data to evaluate the prediction accuracy on unseen data. Next we tested a range of max depths–1, 2, 3, 5, 7, 9, 11– to determine the most accurate tree. We found that a max depth of 5 scored the highest with a score of .988.

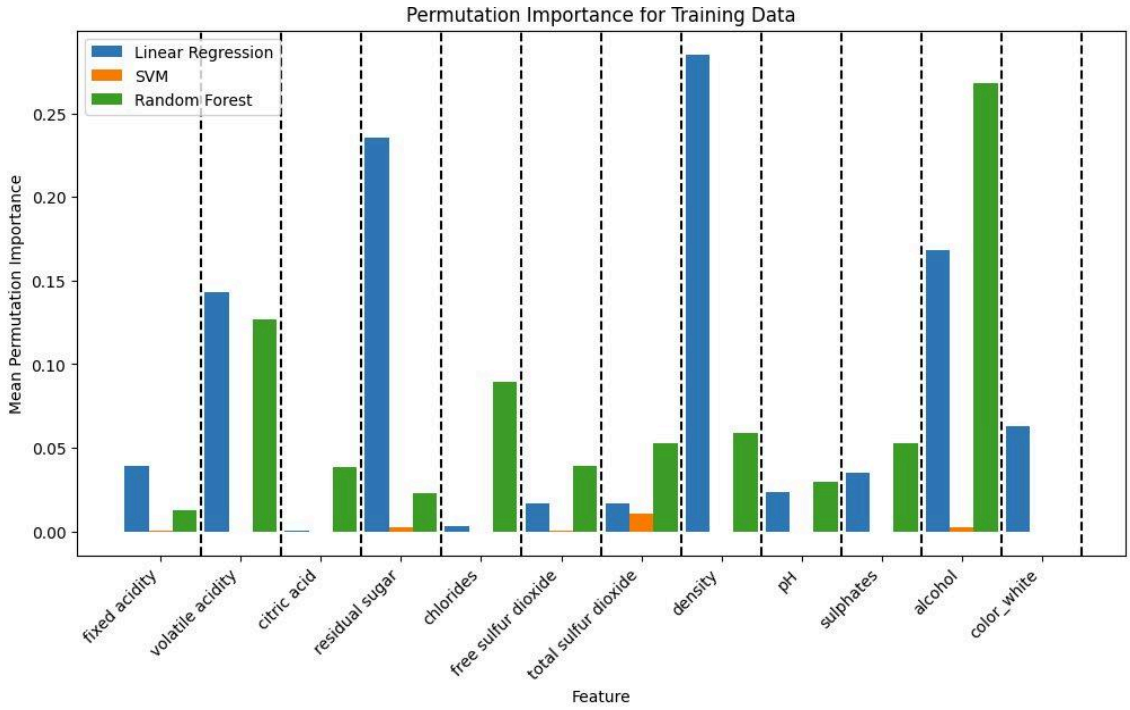
Max Depth	Accuracy on Test Data	Accuracy on Validation Data
1	.929	.917
2	.957	.971
3	.974	.977
5	.985	.988
7	.988	.982
9	.986	.986
11	.986	.986

We wanted to compare the decision tree approach against an alternative classification model, SVM. We fit the classifiers on the training data and evaluated them on the test data to test the models on unseen data. We used a grid search on linear and rbf models with C values (0.1, 1, and 100) and found that a linear model with C = 100 gave us the best results with an accuracy of 0.991.

Kernel Type	C value	Accuracy on Test Data
Linear	.01	.958
Linear	1	.989
Linear	100	.991
Non-linear	.01	.925
Non-linear	1	.938
Non-linear	100	.957

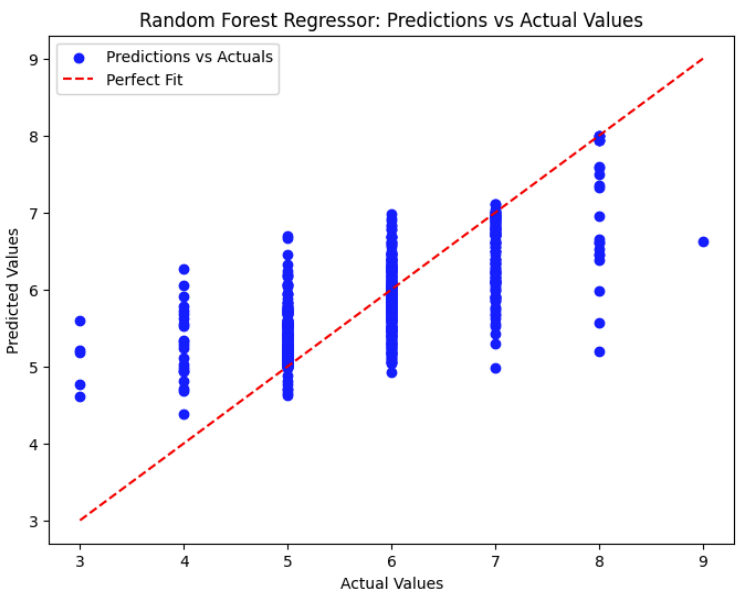
In attempting to predict wine quality, we used permutation feature importance to identify relevant features. However, due to the discrete nature of quality scores (3-9), outliers outside this range posed challenges for our models. Despite efforts to reduce the number of features, accuracy remained low, indicating that wine quality is more subjective than solely determined by chemical properties.

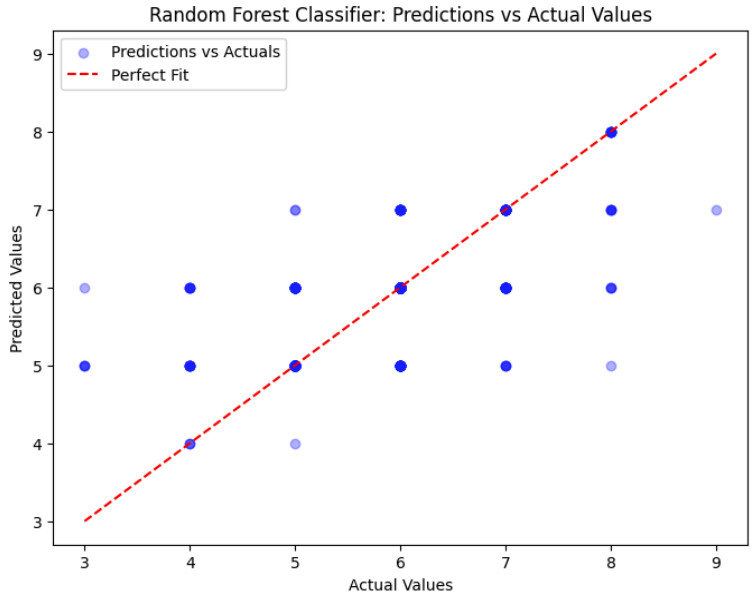




After splitting the data into a training and test set, we performed two GridSearches, one for regressors and classifiers, to find the best model and hyperparameters. For Regressors, we trained an SVM, linear model, KNearestNeighbors, and RandomForest. For classifiers, we trained an SVM, KNeighbors, and a RandomForest. While RandomForest with 200 estimators

yielded the best results for both GridSearches, the most accurate model is a Random Forest Classifier with 200 estimators. We believe that the quality ratings consisting of whole numbers ranging from 3 to 9 affected the RandomForest classifier's higher accuracy scores in comparison to the RandomForest regression model. For the regressor, the accuracy score is .455 while the accuracy score for the classifier is .671. We then plotted each of the models' predictions against the actual values, as shown below. Despite the classifier's high accuracy score, the graphs still suggest that quality is more subjective than determined by chemical properties.





The key takeaway from this analysis shows that wine can effectively be classified in terms of color using the given variables, but we cannot accurately predict the wine's quality when considering these same variables. Our scores show that to predict color, a decision tree classifier scores well on both training data and test data which suggests that the features in the data are distinct between red and white wine. We found that sulfur dioxide and chlorides are most important in distinguishing color, both of which come from grapes during fermentation. As for wine quality we were not able to find a model that accurately predicted the quality on the test data. This led us to the conclusion that quality is subjective to the critic, and not based on the chemical properties of the wine. However, this leaves potential for further exploration of features of wines, and whether there are specific factors that do directly influence the accuracy. There could be potential to use machine learning alongside critics to standardize the quality rating system based on factors of the wines.

Member	Proposal	Coding	Presentation	Report
David Blatz	1	1	1	1
Kylie Scowcroft	1	1	1	1
Matthew Dorn	1	1	1	1
Mya Shanahan	1	1	1	1
Henry Olenec	1	1	1	1