Predicting Quality of Red Wine

Christina Cha
DH 150 | July 27th, 2022

Agenda:

1. Introduction
2. Logistic Regression
3. Random Forest
4. k Nearest Neighbors
5. Prior Work
6. Conclusion

I. Introduction

Predicting Quality of Red Wine

Currently, wine evaluation relies heavily on human specialists to conduct tasting of the wines.

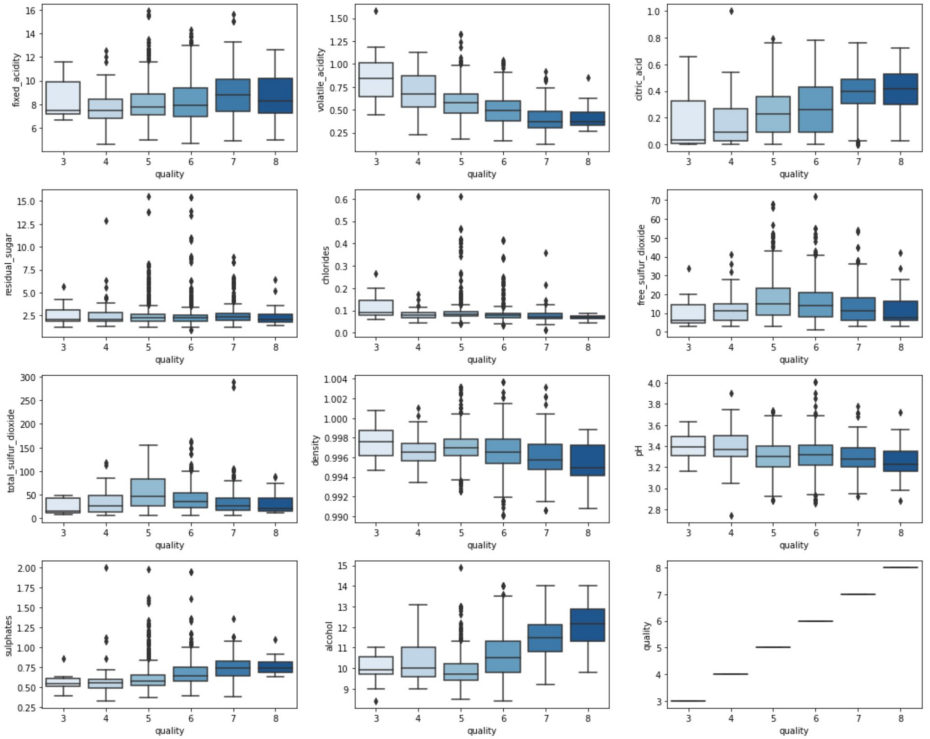
Red Wine Quality Dataset UCI Machine Learning Repository

- 12 variables
  - ***11 features:*** fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol
  - ***1 target:*** quality
- 1599 observations

Understanding the Data

Based on the box plots, we can conclude that…

1. High quality wines have higher levels of alcohol, sulphates, and citric acid.
2. Low quality wines have high volatile acidity, density, and pH
3. Attributes such as residual sugar, total sulfur dioxide, free sulfur dioxide, and chlorides have no effect with the quality of wine



Machine Learning Methods

3 Different Approaches

Logistic Regression, Random Forest, k Nearest Neighbors

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
...	...	...	...	...	...	...	...	...	...	...	...	...
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

Arrays

x = features
y = target (quality)

Machine Learning Methods

3 Different Approaches

Logistic Regression, Random Forest, k Nearest Neighbors

Classifying the Quality of Wine

Divided the quality of the wine into two categories: low quality and high quality.

Quality scores 3 to 6 = 0 (low quality wine)
Quality scores between 7 and 8 = 1 (high quality wine)

II. Logistic Regression

Logistic Regression

Results: The logistic model accurately predicted 86.5% of the red wines to be low or high quality. In this case, **89% of the wines** are predicted to be low quality while **53% of the wines** are predicted to be high quality.

Classification Report:

		precision	recall	f1-score	support
Low Quality Wine	0	0.89	0.96	0.92	413
High Quality Wine	1	0.53	0.25	0.34	67
accuracy				0.86	480
macro avg		0.71	0.61	0.63	480
weighted avg		0.84	0.86	0.84	480

Confusion Metrics:

```
[[398  15]
 [ 50  17]]
```

III. Random Forest

Random Forest

Results: The random forest model accurately predicted 89.8% of the red wines to be low or high quality. In this case, **93% of the wines** are predicted to be low quality while **67% of the wines** are predicted to be high quality.

Low Quality Wine

High Quality Wine

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.96	0.94	413
1	0.67	0.54	0.60	67
accuracy			0.90	480
macro avg	0.80	0.75	0.77	480
weighted avg	0.89	0.90	0.89	480

Confusion Metrics:

```
[[395  18]
 [ 31  36]]
```

IV. k Nearest Neighbor

k Nearest Neighbor

Results: The k nearest neighbors model accurately predicted 84.6% of the red wines to be low or high quality. In this case, **87% of the wines** are predicted to be low quality while **32% of the wines** are predicted to be high quality.

Classification Report:

		precision	recall	f1-score	support
Low Quality Wine	0	0.87	0.97	0.92	413
High Quality Wine	1	0.32	0.09	0.14	67
accuracy				0.85	480
macro avg		0.59	0.53	0.53	480
weighted avg		0.79	0.85	0.81	480

Confusion Metrics:

```
[[400  13]
 [ 61   6]]
```

V. Comparison to Prior Work

Research Paper 1:

Methods: naive payes algorithm , support vector machine, and random forest

Results: The results demonstrate that **Support Vector Machine outperformed** the other models achieving an accuracy of **67.25%** for prediction of red wine quality, followed by **Random Forest** with an accuracy of **65.83%** and Naive Bayes with an accuracy of 55.91%.

TABLE I. Performance measures of training set of red wine dataset using Naïve Bayes

	Training set					
	Wine Quality 3	Wine Quality 4	Wine Quality 5	Wine Quality 6	Wine Quality 7	Wine Quality 8
Precision	0.3333333	0.5	0.9939577	0.9813953	0.556044	0.3333333
Recall	0.5	0.4545455	0.9648094	0.9906103	0.9775281	0.6666667
Specificity	0.00175901	0.9953638	0.9982159	0.4761905	0.5094608	0.991684
F-measure	0.4	0.4761905	0.9792611	0.9859813	0.337037	0.4444444
Accuracy(%)	0.9591586					
Error (%)	0.4408414					

TABLE II. Performance measures of testing set of red wine dataset using Naïve Bayes algorithm.

	Testing set					
	Wine Quality 3	Wine Quality 4	Wine Quality 5	Wine Quality 6	Wine Quality 7	Wine Quality 8
Precision	0	0.3333333	1	0.9767442	0.5939939	0
Recall	0	0.25	0.672028	0.9883353	0.96875	0.5
Specificity	0.00220264	0.9953704	1	0.4304636	0.9192708	0.9900498
F-measure	0	0.2857143	0.9858156	0.9824561	0	0
Accuracy(%)	0.558552					
Error (%)	0.441048					

TABLE III. Performance measures of training set of red wine dataset using Support Vector Machine algorithm.

	Training set					
	Wine Quality 3	Wine Quality 4	Wine Quality 5	Wine Quality 6	Wine Quality 7	Wine Quality 8
Precision	0	0.5	0.9446547	0.5781361	0.8923077	0
Recall	0	0.5	0.9846547	1	1	0
Specificity	0	0.9990834	0.9946572	0.6796117	0.9429695	0.9859719
F-measure	0	0.5	0.9736438	0.5889764	0	0
Accuracy(%)	0.6725821					
Error(%)	0.3274179					

TABLE IV. Performance measures of testing set of red wine dataset using Support Vector Machine algorithm.

	Testing set					
	Wine Quality 3	Wine Quality 4	Wine Quality 5	Wine Quality 6	Wine Quality 7	Wine Quality 8
Precision	0	0	0.9763314	0.9602632	0.7692308	0
Recall	0	0	1	1	1	0
Specificity	0	1	0.9537571	0.785124	0.9766355	0.9859155
F-measure	0	0	0.9914531	0.9800392	0	0
Accuracy(%)	0.6864407					
Error(%)	0.3135593					

TABLE VI. Performance measures of training set of red wine dataset using Random Forest algorithm.

	Training set					
	Wine Quality 3	Wine Quality 4	Wine Quality 5	Wine Quality 6	Wine Quality 7	Wine Quality 8
Precision	0	0.3333333	0.986911	0.9830508	0.9367089	0
Recall	0	0.25	0.9573545	1	1	0
Specificity	0	0.9981618	0.9955397	0.6219136	0.9267327	0.99002
F-measure	0	0.2857143	0.9804941	0.991453	0	0
Accuracy (%)	0.6583851					
Error (%)	0.3416149					

TABLE VI. Performance measures of testing set of red wine dataset using Random Forest algorithm.

	Testing set					
	Wine Quality 3	Wine Quality 4	Wine Quality 5	Wine Quality 6	Wine Quality 7	Wine Quality 8
Precision	0	0	0.5748428	0.977778	0.88	0
Recall	0	0	0.9872611	1	1	0
Specificity	0	1	0.5914163	0.6920152	0.9461358	0.985782
F-measure	0	0	0.5677329	0.988764	0	0
Accuracy(%)	0.654661					
Error(%)	0.3453339					

MY RESULTS

	Model	Accuracy_score
0	Logistic Regression	0.864583
1	Random Forest	0.897917
2	KNeighbours	0.845833

Research Paper 2:

Methods: naive payes algorithm, support vector machine, prism algorithm, and random decision forest

Results: The results demonstrate that random decision forest outperformed the other models achieving the **lowest RMSE** of 0.6430 for white wine and **0.6322 for red wine**, the best model to predict wine quality prediction.

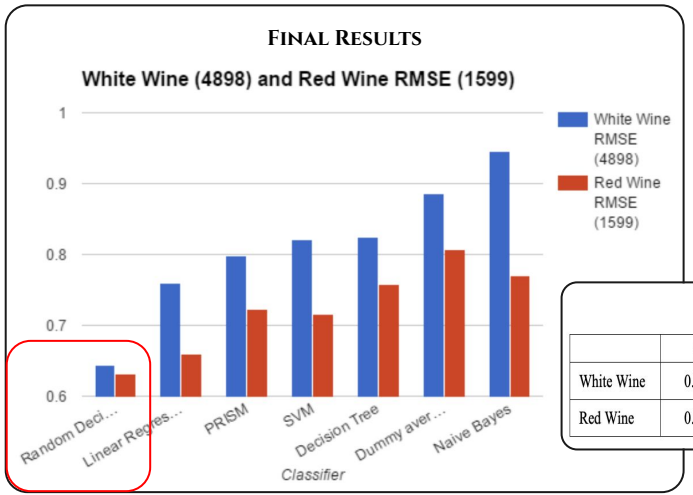


Figure 10 – Initial Results

	NB	*Decision Trees*	*SVM*	*PRISM*
White Wine	0.9455	0.8249	0.8375	0.3193
Red Wine	0.7706	0.7577	0.7730	0.3369

MY RESULTS

	Model	Accuracy_score
0	Logistic Regression	0.864583
1	Random Forest	0.897917
2	KNeighbours	0.845833

VI. Conclusion

Conclusion

Results: Random Forest performed the best out of the three methods.

1. Random Forest
2. Logistic Regression
3. kNeighbours

Top three features that impact the quality of red wine (using logistics regression):

1. Alcohol
2. Sulphates
3. Fixed Acidity