Wine Quality Prediction Team 2

Wine Quality + Machine Learning

Table of Contents:

Overview

Objective:
We are a new wine startup based in the Adelaide Hills, this is our research project into our competitors wines, to analyse what is the most influential characteristic of a wine to produce the highest rated product.

We will also use a Machine Learning Model to evaluate and predict “good wines” and optimise the model to assist with our own production methods and adequately assess wine quality before pushing our wine to market.

Real World Applications of Our Model:

1. Knowing the quality of a new wine would allow us to predict the demand for that wine and price the wine accordingly.

2. Help us predict the shelf life of a new untested wine. Higher quality wine we hope will sell faster.

3. Knowing which characteristic is the most influential would allow us use wine making techniques that maximise these characteristics.

4. Save money by skipping a whole trial and error phase. Focusing in on the characteristics identified for a highly rated product.

Process

Find the Data Set

We researched and found the data set on-line
https://archive.ics.uci.edu/dataset/186/wine+quality

Data Analysis & Model Assessment

We consolidated the data into a SQL database. And ran various data analysis and pre-processing. We then validated xx models prior to selecting two models to optimise.

Optimise the Model

In this final phase, we optimised two models to produce our results and conclusion.

Find the Data Set

01

Two datasets are included, related to red and white wine samples, from the north of Portugal.

The goal is to model wine quality based on physicochemical test results. (see [Cortez et al., 2009], http://www3.dsi.uminho.pt/pcortez/wine/).

In this step we created a SQLite DB using the CSV files for both red & white wine and create a new value “Type” for Red or White wine test results.

The Data

Red Wine

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

White Wine

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

Data Engineering

Data Loading and Labeling

- Two datasets were utilized, red and white, with various chemical properties
- Added a ‘type’ column to each dataset, marking them as ‘Red’ or ‘White’ for analysis later.

Database Creation and Storage

- Created an SQL database to store our datasets. Separated the database to red and white wine named `red_wine_quality.db`and `white_wine_quality.db`respectively.
- Why SQLite?

Data Verification

- SQL queries to retrieve the stored data to verify for integrity and completeness of the data.

Step 01 | Summary

The data was merged the CSV files into a SQLite DB, adding a wine “type” for red or white wine, allowing for access and creation of Dataframes as necessary.

We included all of the data sets values and did not exclude or drop any outliers.

The two items influencing red wine quality was Alcohol & Sulphates.

The two items influencing white wine quality was Alcohol & Density.

Conduct Analysis

02

What makes a good Quality wine?

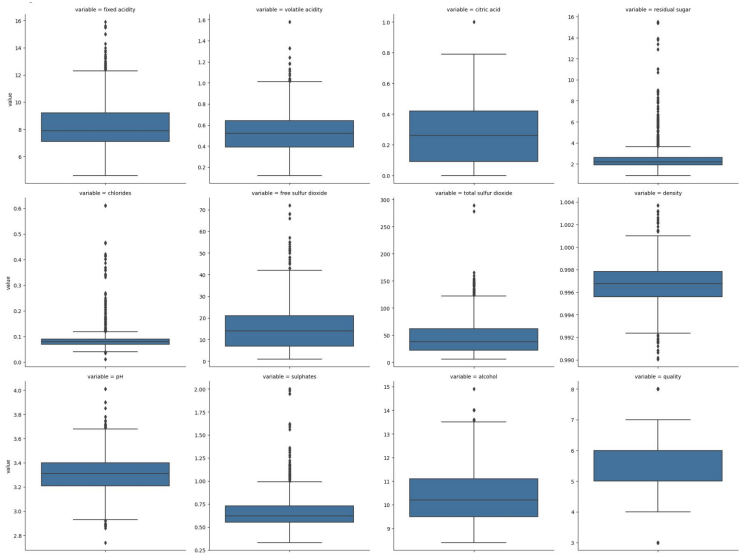
In this phase we undertook some pre-processing and data analysis. This included box plot, heat mapping, a PCA and feature importance assessments.

We then split the data to based on quality rating of >7, to a ‘good quality’ or not.

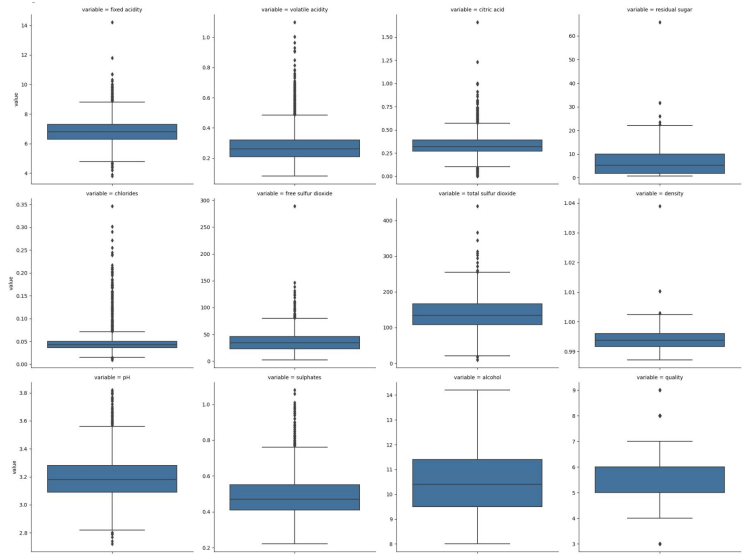
We the undertook to standardise the data sets using the `from sklearn.preprocessing import StandardScaler`.

Box Plot

In this image we can see the box plot of the data sampling variable features.
Each box plot Y-axis has been set to show the scaled values.



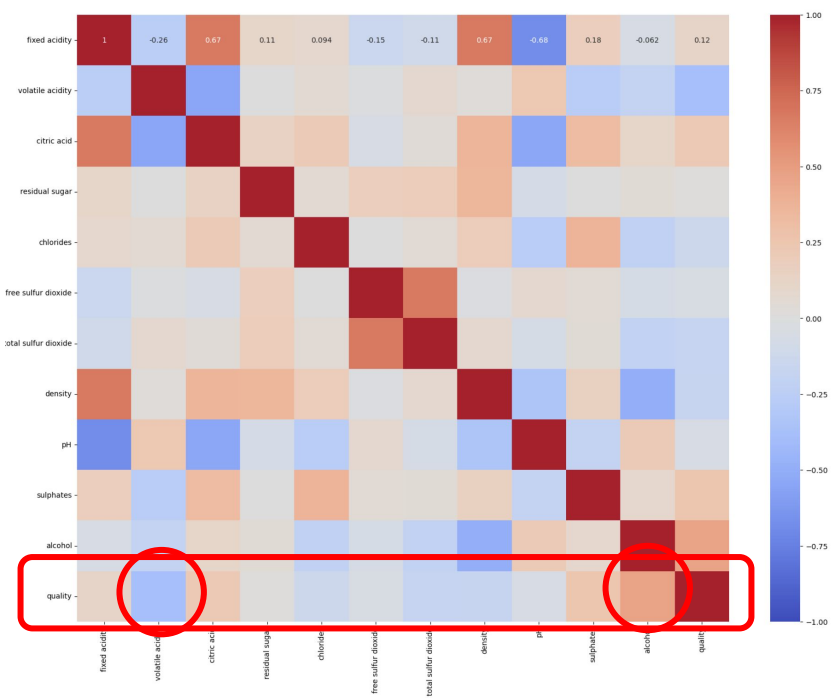
Red Wine



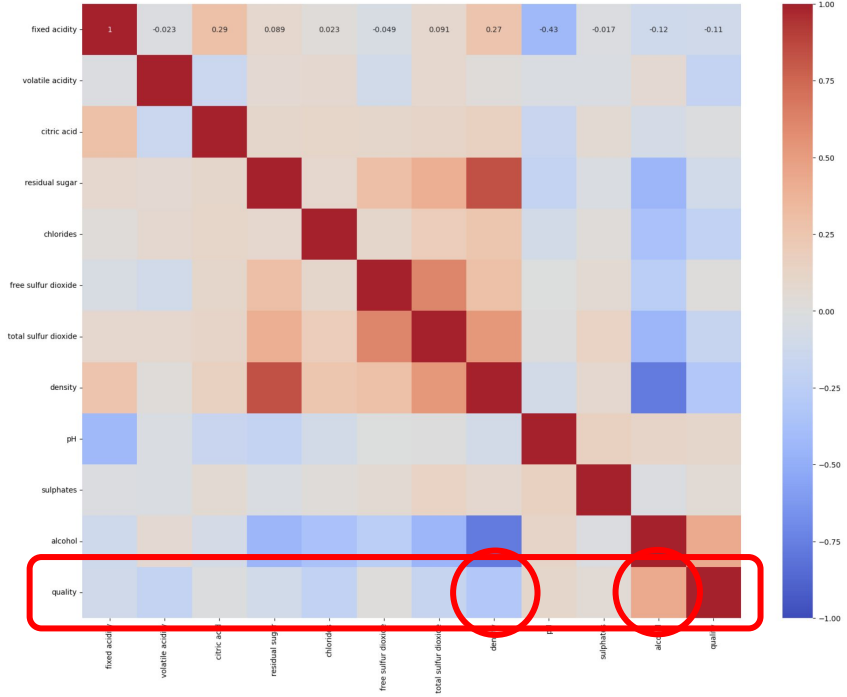
White Wine

Heat Map

In this image we can see the feature correlation in the below heat map. Interestingly, alcohol content was a high positive factor in both red and white wine quality.



Red Wine



White Wine

Step 02 Analysis | Summary

The data was assessed and visualised in a Box Plot and a Heat Map this showed the range of data and the correlations to the wine Quality rating.

We included all of the data sets values and did not exclude or drop any outliers. As this is important for the wine testing points.

The two items influencing red wine quality was Alcohol & Volatile Acidity (Which is the a measure of the low molecular weight (or steam distillable) fatty acids and is generally perceived as the odour of vinegar.).

The two items influencing white wine quality was Alcohol & Density.

Model Selection

02

We tested 15 different binary classification Machine Learning models to find the most accurate model. This included;

Logistic Regression	**Linear Discriminant Analysis**	**Support Vector Machine**	
DecisionTreeClassifier	**RandomForestClassifier**	**Gradient Boosting Classifier**	**AdaBoostClassifier**
Bagging Classifier	**K-Nearest Neighbors**	**Gaussian Naive Bayes**	**Quadratic Discriminant Analysis**
Multilayer Perceptron	**RidgeClassifier**	**Extra Trees Classifier**	**Isolation Forest**

ML Summary | Red

	Model	Precision	Recall	F1-Score	Support	Accuracy
2	Logistic Regression	0.5818181818181818	0.28193832599118945	0.3798219584569733	227.0	0.786734693877551
3	Linear Discriminant Analysis	0.5666666666666667	0.29955947136563876	0.3919308357348703	227.0	0.7846938775510204
4	Support Vector Machine	0.7549019607843137	0.3392070484581498	0.46808510638297873	227.0	0.8214285714285714
5	Decision Tree Classifier	0.606425702811245	0.6651982378854625	0.634453781512605	227.0	0.8224489795918367
6	Random Forest Classifier	0.8529411764705882	0.6387665198237885	0.7304785894206548	227.0	0.8908163265306123
7	Gradient Boosting Classifier	0.704225352112676	0.44052863436123346	0.5420054200542005	227.0	0.8275510204081633
8	AdaBoost	0.6163522012578616	0.43171806167400884	0.5077720207253885	227.0	0.8061224489795918
9	Bagging Classifier	0.8058823529411765	0.6035242290748899	0.6901763224181361	227.0	0.8744897959183674
10	K-Nearest Neighbors	0.6853932584269663	0.5374449339207048	0.6024691358024692	227.0	0.8357142857142857
11	Gaussian Naive Bayes	0.4262734584450402	0.7004405286343612	0.53	227.0	0.7122448979591837
12	Quadratic Discriminant Analysis	0.4595375722543353	0.7004405286343612	0.5549738219895288	227.0	0.7397959183673469
13	Multilayer Perceptron	0.7181208053691275	0.4713656387665198	0.5691489361702127	227.0	0.8346938775510204
14	Ridge Classifier	0.6865671641791045	0.2026431718061674	0.3129251700680272	227.0	0.7938775510204081
15	ExtraTrees Classifier	0.8323699421965318	0.6343612334801763	0.7200000000000001	227.0	0.8857142857142857
16	Isolation Forest	0.23777777777777778	0.9427312775330396	0.3797692990239574	227.0	0.21836734693877552

The best model for Red wine was Random Forest Classifier with >89% accuracy.

Precision >85%

Recall >63%

F1 Score > 73%

ML Summary | White

	Model	Precision	Recall	F1-Score	Support	Accuracy
2	Logistic Regression	0.5818181818181818	0.28193832599118945	0.3798219584569733	227.0	0.786734693877551
3	Linear Discriminant Analysis	0.5666666666666667	0.29955947136563876	0.3919308357348703	227.0	0.7846938775510204
4	Support Vector Machine	0.7549019607843137	0.3392070484581498	0.46808510638297873	227.0	0.8214285714285714
5	Decision Tree Classifier	0.606425702811245	0.6651982378854625	0.634453781512605	227.0	0.8224489795918367
6	Random Forest Classifier	0.8529411764705882	0.6387665198237885	0.7304785894206548	227.0	0.8908163265306123
7	Gradient Boosting Classifier	0.704225352112676	0.44052863436123346	0.5420054200542005	227.0	0.8275510204081633
8	AdaBoost	0.6163522012578616	0.43171806167400884	0.5077720207253885	227.0	0.8061224489795918
9	Bagging Classifier	0.8058823529411765	0.6035242290748899	0.6901763224181361	227.0	0.8744897959183674
10	K-Nearest Neighbors	0.6853932584269663	0.5374449339207048	0.6024691358024692	227.0	0.8357142857142857
11	Gaussian Naive Bayes	0.4262734584450402	0.7004405286343612	0.53	227.0	0.7122448979591837
12	Quadratic Discriminant Analysis	0.4595375722543353	0.7004405286343612	0.5549738219895288	227.0	0.7397959183673469
13	Multilayer Perceptron	0.7181208053691275	0.4713656387665198	0.5691489361702127	227.0	0.8346938775510204
14	Ridge Classifier	0.6865671641791045	0.2026431718061674	0.3129251700680272	227.0	0.7938775510204081
15	ExtraTrees Classifier	0.8323699421965318	0.6343612334801763	0.7200000000000001	227.0	0.8857142857142857
16	Isolation Forest	0.23777777777777778	0.9427312775330396	0.3797692990239574	227.0	0.21836734693877552

The best model for White wine was Random Forest Classifier with >89% accuracy.

Precision >85%

Recall >63%

F1 Score > 73%

Step 02 Summary

The Random Forest Classifier consistently showed the highest accuracy and precision in predicting whether a wine is of good quality.

- Precision prevents False Positives → False Positive means poor quality wine is classified as good quality.
- Recall ensures we don’t miss an opportunity for a good wine.

In our binary classification task (good quality vs. not good quality), precision ensures that most of the wines predicted to be of good quality actually are, while recall ensures that we are identifying most of the good quality wines.

We chose RandomForestClassifier over Bagging or extra trees because, the Additional random feature selection used in Random Forest, decorrelates the features by selecting a random subset of features and determines the best split from this subset.

This process reduces overfitting and ensure greater accuracy.

Optimise the Models

03

Based on our project objective to find the best ML model, we further optimised two models with the best results from the Data Analysis step.

Random Forest Classifier - Chosen Model - Maximum Accuracy of 91%

Neural Network - Discarded Model - Maximum Accuracy of 83%

Data Implications:
Our work shows that the default models were well robust and in our optimisationation were unable significantly to improve the results.

Random Forest Classifier | Red & White

The model optimisation was modified as follows;
1. The number of Estimators was increased from the default (100) to 300, this had limited benefit.
2. The maximum Features was changed from default (sqrt) to Auto, this had limited benefit.
3. GridsearchCV was used to optimise over multiple parameters which yielded a minor benefit from 0.9 to 0.91 accuracy for Red wine and

In summary, we were not able to significantly improve the Random Forest Classifier Model for Red Wine

Random Forest Classifier | Results

Dataset	Model	Precision	Recall	F1-Score	Accuracy
Red Wine Initial	Random Forest	0.8916	0.9000	0.8925	0.9000
Red Wine Optimized	Random Forest	0.9027	0.9094	0.9020	0.9094
White Wine Initial	Random Forest	0.8882	0.8908	0.8850	0.8908
White Wine Optimized	Random Forest	0.8821	0.8857	0.8800	0.8857

Conclusion

Based on various analysis, we can confirm that wine quality can be predicted prior to its production, focus on Alcohol Content. Which perfectly make sense because, not only is it about the feelings after drinking the wine - it can influence the taste, texture and quality.

Cheers !

Thank you.