Wine Quality Prediction

How accurately can we predict red wine quality with their physicochemical properties?

Group: Sophia Mei, Vicky Ge, Leah Lee, Peter Lin

Link to GitHub repository: https://github.com/Sophiameiman/Wine-Quality-Prediction/tree/main

The best wine!

Why Wine?

- Diverse wine types presents a challenge for consumers
- For amateur, there is no clear criteria for selecting high-quality wine
- With ML, our goal would be ESTABLISHING CRITERIA FOR THE BEST WINE

Red Wine Quality Dataset

From UCI Machine Learning Repository

Wine Quality

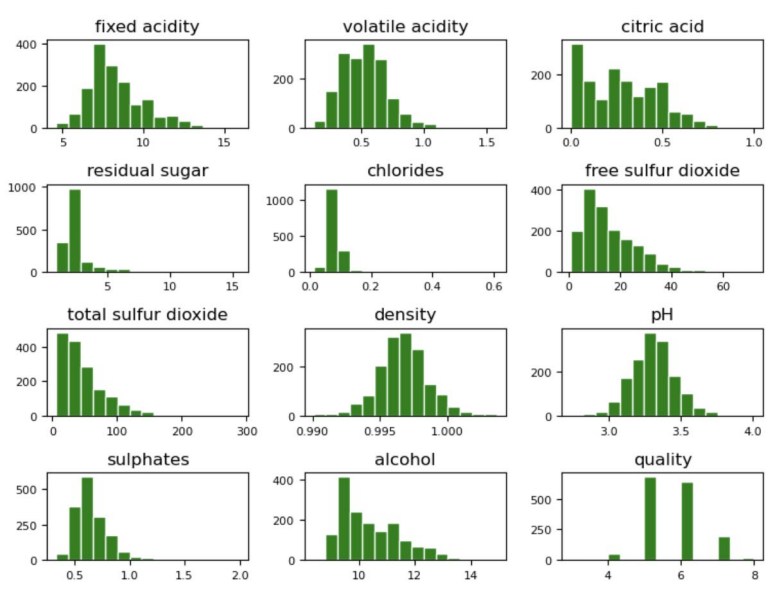
Donated on 10/6/2009

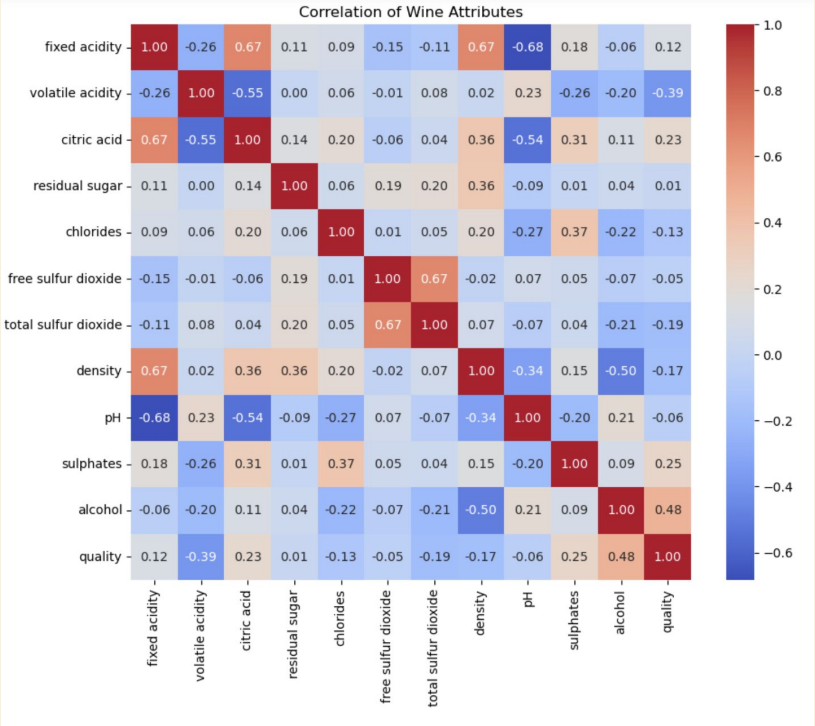
Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests (see [Cortez et al., 2009], http://www3.dsi.uminho.pt/pcortez/wine/).

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Business	Classification, Regression
Feature Type	**# Instances**	**# Features**
Real	4898	11

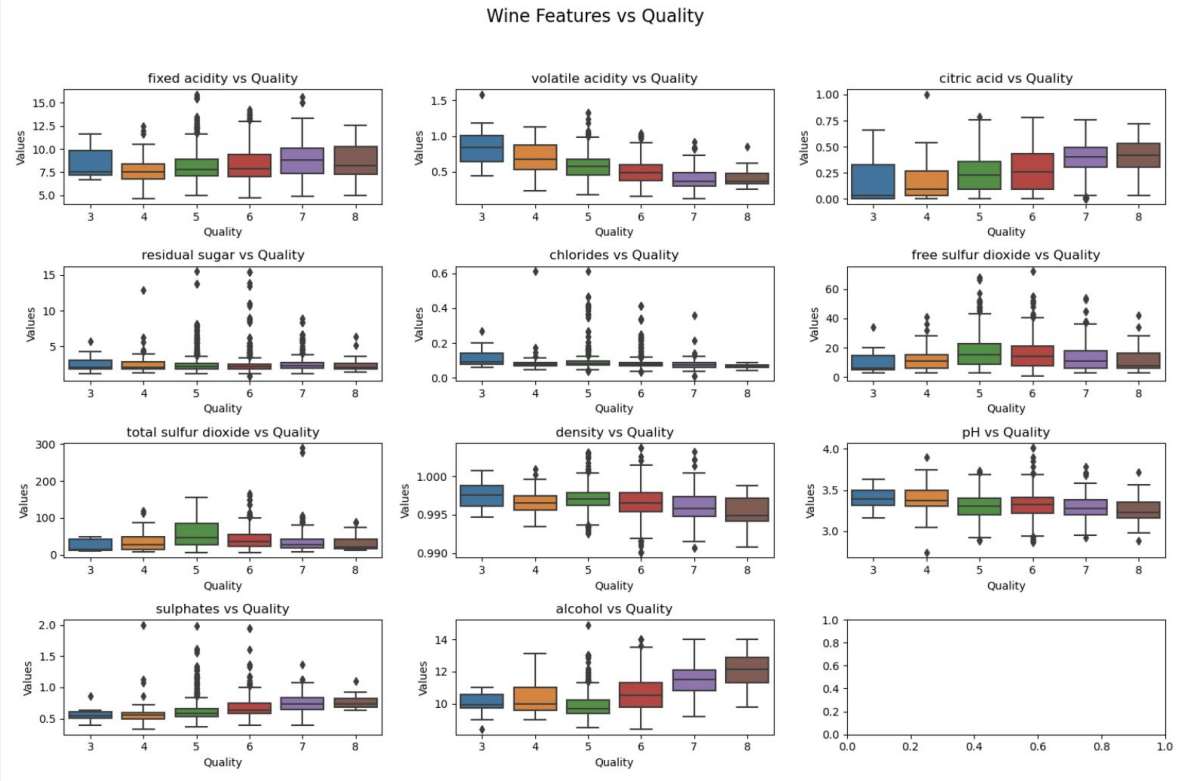
fixed acidity	continuous	**total sulfur dioxide**	continuous
volatile acidity	continuous	**density**	continuous
citric acid	continuous	**pH**	continuous
residual sugar	continuous	**sulphates**	continuous
chlorides	continuous	**alcohol**	continuous
free sulfur dioxide	continuous	**quality** Output Variable	categorical

EDA

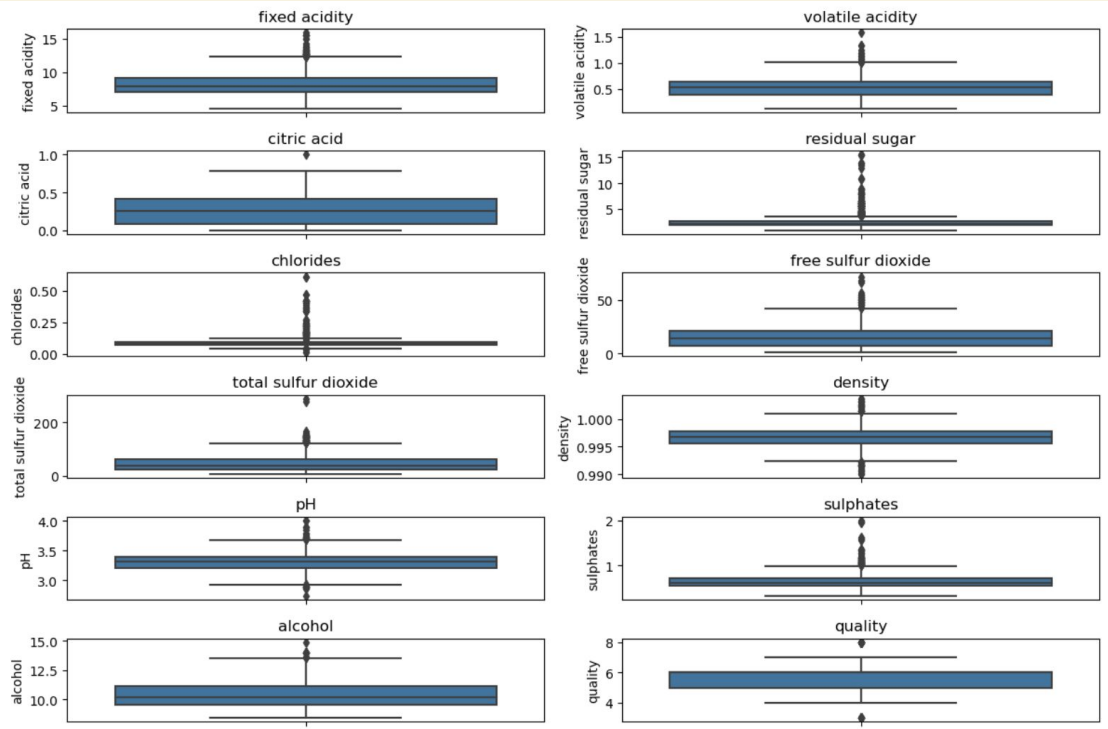

fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality


Correlation of Wine Attributes
fixed acidity 1.00 -0.26 0.67 0.11 0.09 -0.15 -0.11 0.67 -0.68 0.18 -0.06 0.12
volatile acidity -0.26 1.00 -0.55 0.00 0.06 -0.01 0.08 0.02 0.23 -0.26 -0.20 -0.39
citric acid 0.67 -0.55 1.00 0.14 0.20 -0.06 0.04 0.36 -0.54 0.31 0.11 0.23
residual sugar 0.11 0.00 0.14 1.00 0.06 0.19 0.20 0.36 -0.09 0.01 0.04 0.01
chlorides 0.09 0.06 0.20 0.06 1.00 0.01 0.05 0.20 -0.27 0.37 -0.22 -0.13
free sulfur dioxide -0.15 -0.01 -0.06 0.19 0.01 1.00 0.67 -0.02 0.07 0.05 -0.07 -0.05
total sulfur dioxide -0.11 0.08 0.04 0.20 0.05 0.67 1.00 0.07 -0.07 0.04 -0.21 -0.19
density 0.67 0.02 0.36 0.36 0.20 -0.02 0.07 1.00 -0.34 0.15 -0.50 -0.17
pH -0.68 0.23 -0.54 -0.09 -0.27 0.07 -0.07 -0.34 1.00 -0.20 0.21 -0.06
sulphates 0.18 -0.26 0.31 0.01 0.37 0.05 0.04 0.15 -0.20 1.00 0.09 0.25
alcohol -0.06 -0.20 0.11 0.04 -0.22 -0.07 -0.21 -0.50 0.21 0.09 1.00 0.48
quality 0.12 -0.39 0.23 0.01 -0.13 -0.05 -0.19 -0.17 -0.06 0.25 0.48 1.00
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

EDA-Subplots



Data Preprocessing



- Used box plots to visualize interquartile range
- Train-test split (80/20 split)
- Data cleaning by removing the outliers

Model 1: Logistics Regression + Best Subset Selection

- **Best Subset Selection:** exhaustive search over all possible combinations of features based on cross-validation accuracy
  - Stratified K-Fold cross-validation (K=4) to maintain the balance of classes across different subsets
  - Best subset of 9 predictors: `fixed acidity`, `volatile acidity`, `citric acid`, `residual sugar`, `chlorides`, `free sulfur dioxide`, `total sulfur dioxide`, `sulphates`, `alcohol`
- **Predictive Model:** Logistic Regression pipeline with StandardScaler and iterate up to 10,000 times

```
Testing Data Metrics:
Confusion Matrix:
[[  0   0   2   0   0   0]
 [  0   0   6   4   1   0]
 [  0   0 104  30   1   0]
 [  0   0  39  88  15   0]
 [  0   0   2  15  10   0]
 [  0   0   0   1   2   0]]

Classification Report:
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         2
           4       0.00      0.00      0.00        11
           5       0.68      0.77      0.72       135
           6       0.64      0.62      0.63       142
           7       0.34      0.37      0.36        27
           8       0.00      0.00      0.00         3

    accuracy                           0.63       320
   macro avg       0.28      0.29      0.28       320
weighted avg       0.60      0.63      0.61       320
```

Model 2: Random Forest Classifier

- Utilizing Randomized SearchCV: Fitting 5 folds cross validation.

- **Best Parameters:**
  'n_estimators': 600,
  'min_samples_split': 5,
  'min_samples_leaf': 1

  Overall Accuracy: 73%

Confusion Matrix:

```
[[  0   0   0   2   0   0]
 [  0   0   6   5   0   0]
 [  0   0 115  18   2   0]
 [  0   0  30 105   7   0]
 [  0   0   1   9  15   2]
 [  0   0   0   1   2   0]]
```

Classification Report:

	precision	recall	f1-score	support
3	0.00	0.00	0.00	2
4	0.00	0.00	0.00	11
5	0.76	0.85	0.80	135
6	0.75	0.74	0.74	142
7	0.58	0.56	0.57	27
8	0.00	0.00	0.00	3
accuracy			0.73	320
macro avg	0.35	0.36	0.35	320
weighted avg	0.70	0.73	0.72	320

Model 3: KNN

- **Best parameters** (n_neighbors) is 1
  - Best Cross-validation Score: 0.61–a moderate predictive performance
- **Overall accuracy** is 62%
- A high number of misclassifications between adjacent quality classes like 5 and 6
- This model has better performance in distinguishing some of the **middle classes** (5, 6, and 7)
  - The poor performance on minority classes (3 and 8), likely due to class imbalance

```
Best parameters: {'n_neighbors': 1}
Best cross-validation score: 0.61
Confusion Matrix:
[[ 0  0  0  1  0  0]
 [ 2  1  1  6  0  0]
 [ 1  4 87 35  3  0]
 [ 0  1 31 86 13  1]
 [ 0  0  2 14 25  1]
 [ 0  0  0  2  3  0]]

Classification Report:
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.17      0.10      0.12        10
           5       0.72      0.67      0.69       130
           6       0.60      0.65      0.62       132
           7       0.57      0.60      0.58        42
           8       0.00      0.00      0.00         5

    accuracy                           0.62       320
   macro avg       0.34      0.34      0.34       320
weighted avg       0.62      0.62      0.62       320
```

Model Comparison

Model	Accuracy	Avg. Precision	Avg. Recall	Avg. F1 Score
LR + Best Subset	0.63	0.60	0.63	0.61
Random Forests	0.73	0.70	0.73	0.72
KNN	0.62	0.62	0.62	0.62

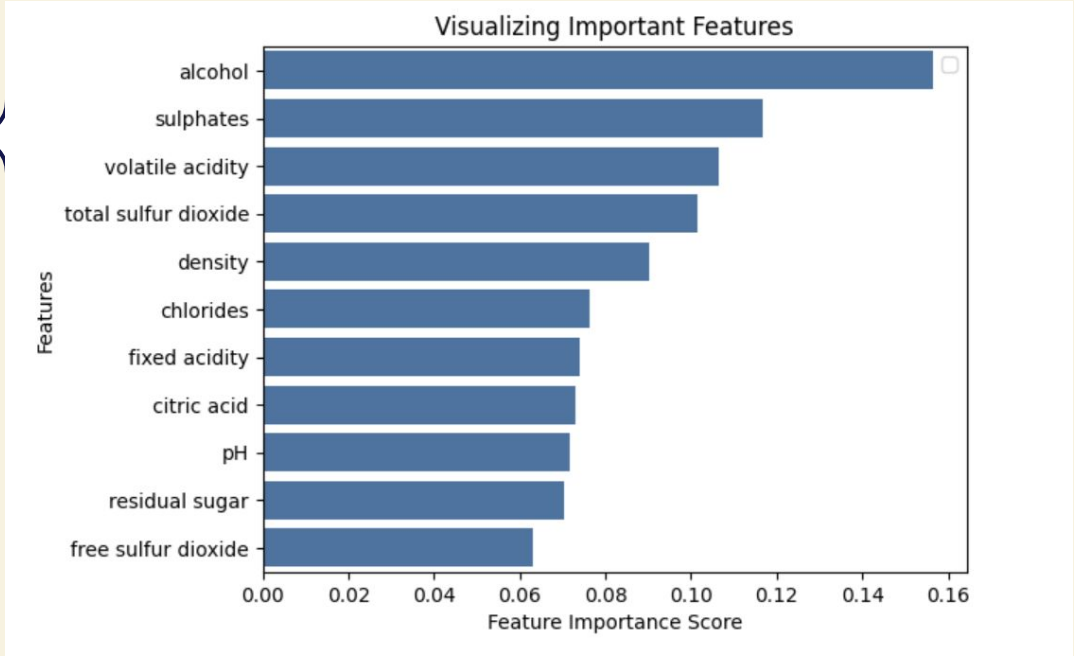
Random Forests Model:

- Highest accuracy of **0.73**, indicating the most reliable predictions
- Highest precision of **0.70**, meaning when it predicts a wine to be of a certain quality, it is correct 70% of the time
- Highest recall of **0.73**, identifying nearly three-quarters of all high-quality wines
- Highest F1 score of **0.72**

Strengths of RF model:

- Handle feature Interactions well
- Robust to overfitting
- Capture non-linear relationship

Conclusion




```
alcohol                 0.156508
sulphates               0.116797
volatile acidity        0.106611
total sulfur dioxide    0.101647
density                 0.090210
chlorides               0.076197
fixed acidity           0.074036
citric acid             0.072896
pH                      0.071600
residual sugar          0.070467
free sulfur dioxide     0.063031
dtype: float64
```

THANKS!

Any questions?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**