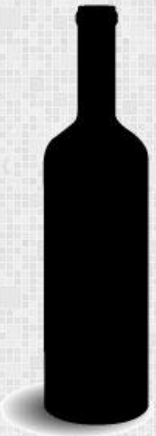
Red Wine Quality Prediction



Team 6:
Priya Ramakrishnan, Chirag Madhukar, Parth Parsana, Noor Zia,Pratheek Praveen Kumar.

Agenda

1. Introduction
2. Project Purpose
3. Data Dictionary
4. Analysis Process
5. Models & Prediction
6. Correlation Plots
7. Conclusion & Takeaways

Introduction

Facts:

- Quality Certification
- Physicochemical properties

Problem statement:

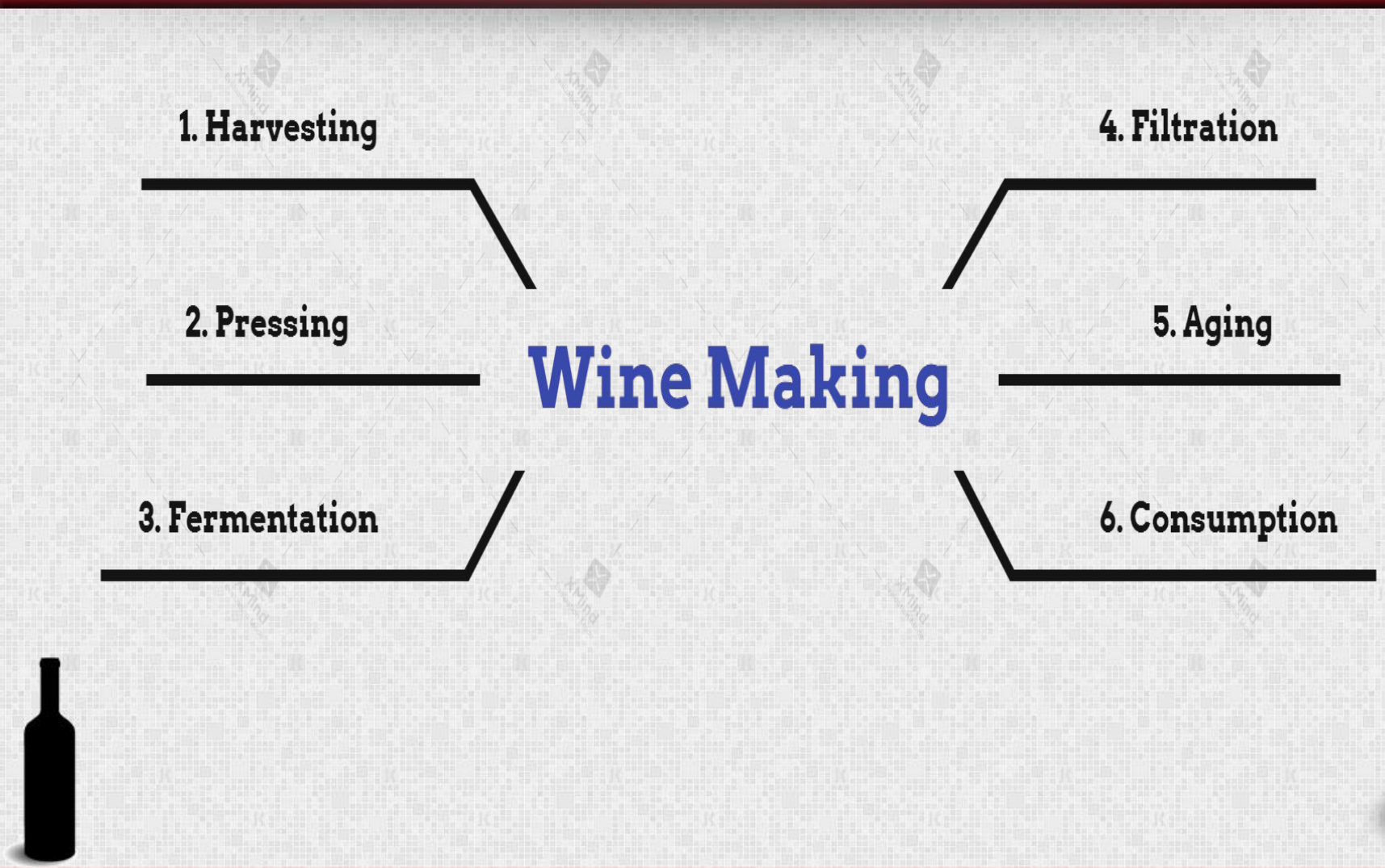
- Biased opinion
- Expensive

Value:

Controlled certification and assessment process reducing human intervention.

“I’m a wine enthusiast. The more wine I drink, the more enthusiastic I get.”

Wine Making Procedure

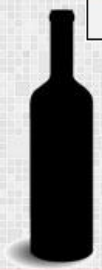


Chemical Process

- Fermentation -> alcohol & residual sugars
- Grape -> acids, chlorides
- Preservation -> sulphates , free sulphur dioxide

Good tasting wines should have:

Alcohol	Volatile Acidity	Fixed Acidity	Sulphates	Citric acid	Residual sugars
High	Low	High	Low	Low	low
pH	Chlorides	Free sulphur dioxide	Density	Total sulphur dioxide	
low	Low	Low	high	low	



Project Purpose

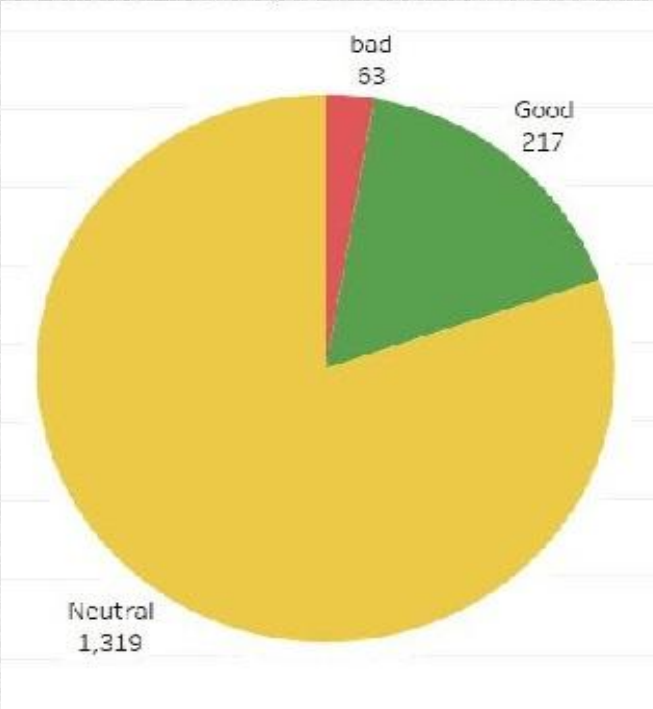
- Determine the best red wine quality indicators.
- Quality of wine = Taste Preferences
- Multi Class Classification Problem

Data Dictionary

- Fixed acidity - Unit: Density in float
- Volatile acidity - Unit: Density in float
- Citric acid - Unit: Density in float
- Residual sugar - Unit: Grams in float
- Chlorides - Unit: Density in float
- Free sulphur dioxide - Unit: Density in float
- Total sulphur dioxide - Unit: Density in float
- Density - Unit: Density in float
- pH - Unit: Integer, Scale: 1-14
- Sulphates - Unit: Density in Float
- Alcohol - Unit: Percentage by volume
- Quality - Class - Scale 1- 10

Analysis Process

- Dataset from: UCI Data Repository
- Sample size: 1599
- Predictor Variable: Quality(1-10 scale)
- Pre-Processing
  - Classified the data in 3 labels
    - Quality > 6 - Good
    - 6 > Quality > 4 - Okay
    - Quality < 4 - Bad
- Smote Analysis to balance data
- Train: 70% ; Test: 30%
- 4 models analyzed



Models Used

- Multinomial Regression
- XGBoost
- Random Forest
- Support Vector Machine (SVM)

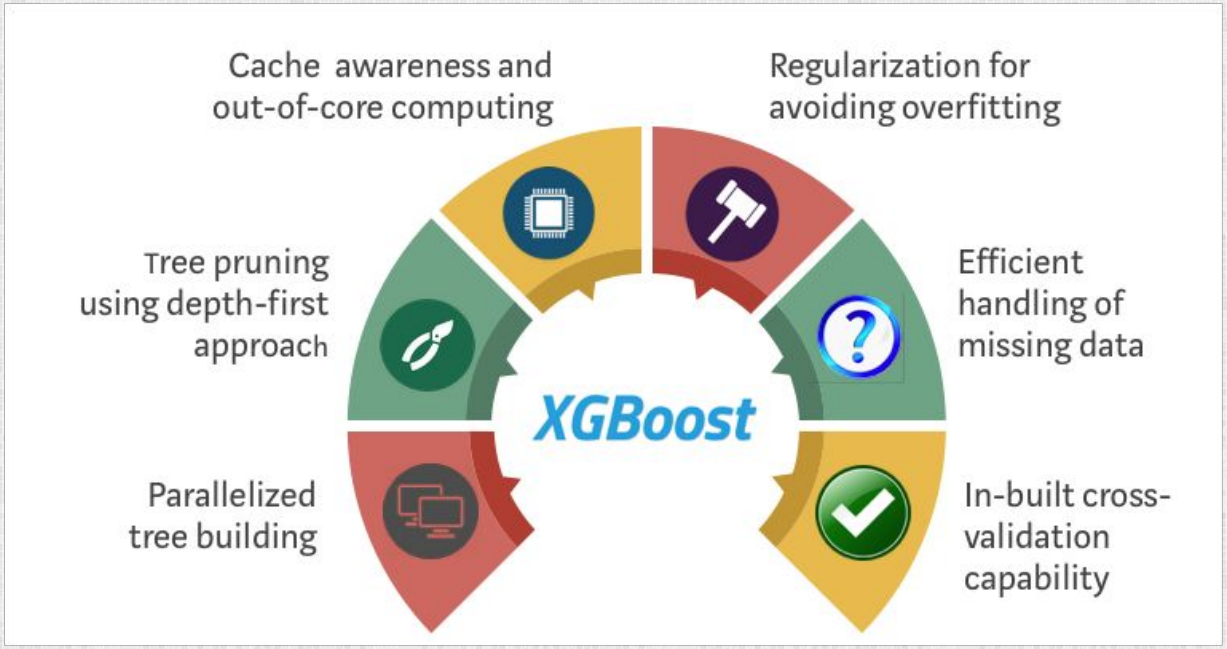
Multinomial Regression

```
> head(pvalues)
     (Intercept) fixed.acidity volatile.acidity citric.acid
Good           0  7.405287e-02     3.361148e-18  0.03002649
Ok             0  4.366036e-15     9.326824e-20  0.70608666
     residual.sugar  chlorides free.sulfur.dioxide
Good   3.983799e-06 0.03989669        3.999020e-07
Ok     4.091226e-16 0.71735877        4.848941e-04
     total.sulfur.dioxide density           pH    sulphates
Good          0.001380443       0 1.712158e-10 4.551088e-11
Ok            0.123183015       0 4.432288e-22 3.792674e-01
          alcohol
Good 9.881660e-39
Ok   1.035369e-13
```

XGBoost

Preprocessing:

- Conversion to xgb.DMatrix format
- Hyperparameters chosen:
  - nrounds = 100
  - objective = "multi:softmax"



XGBoost Confusion Matrix

Accuracy : 0.7531

Confusion Matrix :-

	Reference		
Prediction	bad	good	ok
bad	9	0	36
good	0	50	58
ok	9	15	301

XGBoost Parameters

	Class: bad	Class: good	Class: ok
Sensitivity	0.50	0.77	0.76
Specificity	0.92	0.86	0.71
Balanced Accuracy	0.71	0.81	0.74
Precision	0.04	0.18	0.85
F1 Score	0.08	0.29	0.80

XGBoost Features

Sr No.	Feature	Gain	Cover	Frequency
1	alcohol	0.24718990	0.2576002	0.1901004
2	sulphates	0.24216656	0.1995195	0.1744978
3	volatile acidity	0.23474276	0.1638620	0.1608680
4	pH	0.11091620	0.1515180	0.1807747
5	density	0.08595417	0.1036645	0.1576399
6	citric acid	0.07903041	0.1238358	0.1361191

Random Forest

Hyperparameters:

Variables: 4

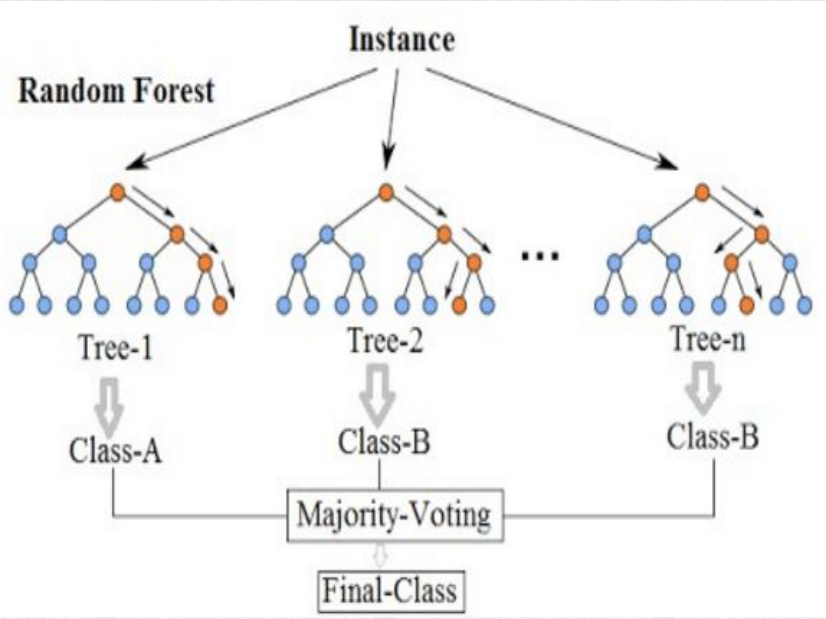
Terminal nodes: 6

ntree:100

Confusion Matrix:

	Reference		
Prediction	bad	good	ok
bad	9	0	38
good	0	52	57
ok	9	13	300

Accuracy: 75.52%



Random Forest

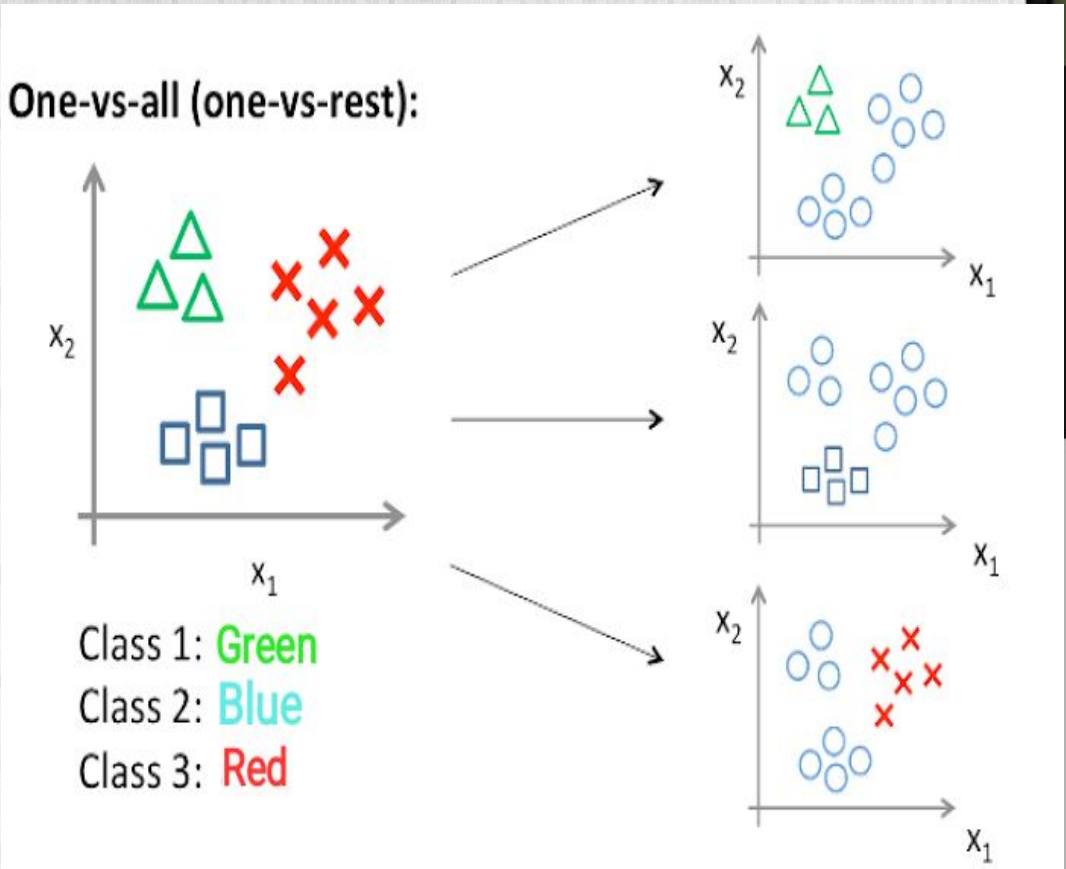
	Class: bad	Class: good	Class: ok
Sensitivity	0.50	0.80	0.76
Specificity	0.92	0.86	0.73
Balanced Accuracy	0.71	0.83	0.75
Precision	0.19	0.48	0.93
F1 Score	0.28	0.60	0.84

Support Vector Machine

Confusion Matrix :-

	Reference		
Prediction	bad	good	ok
bad	11	1	71
good	1	53	67
ok	6	11	257

Accuracy : 0.6715

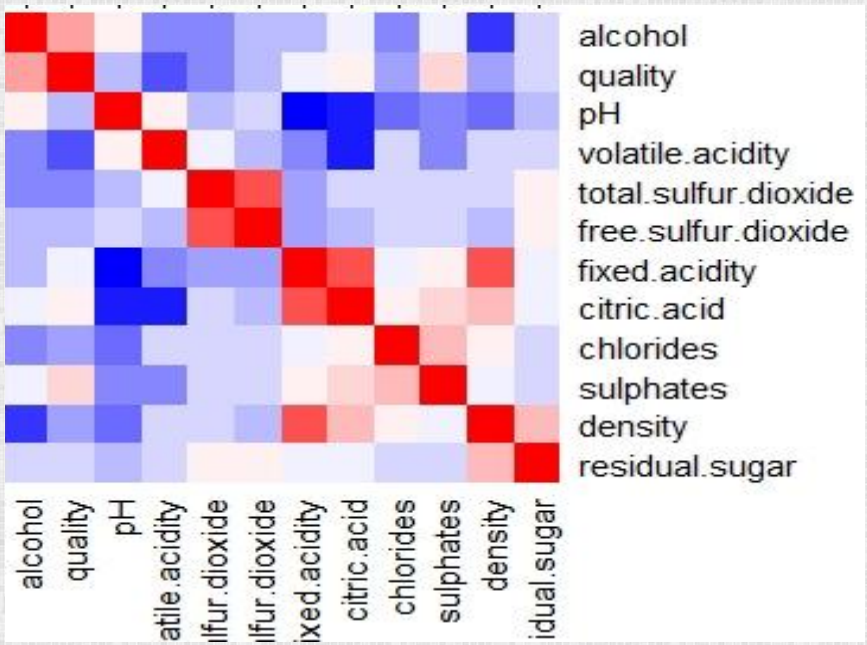


Support Vector Parameters

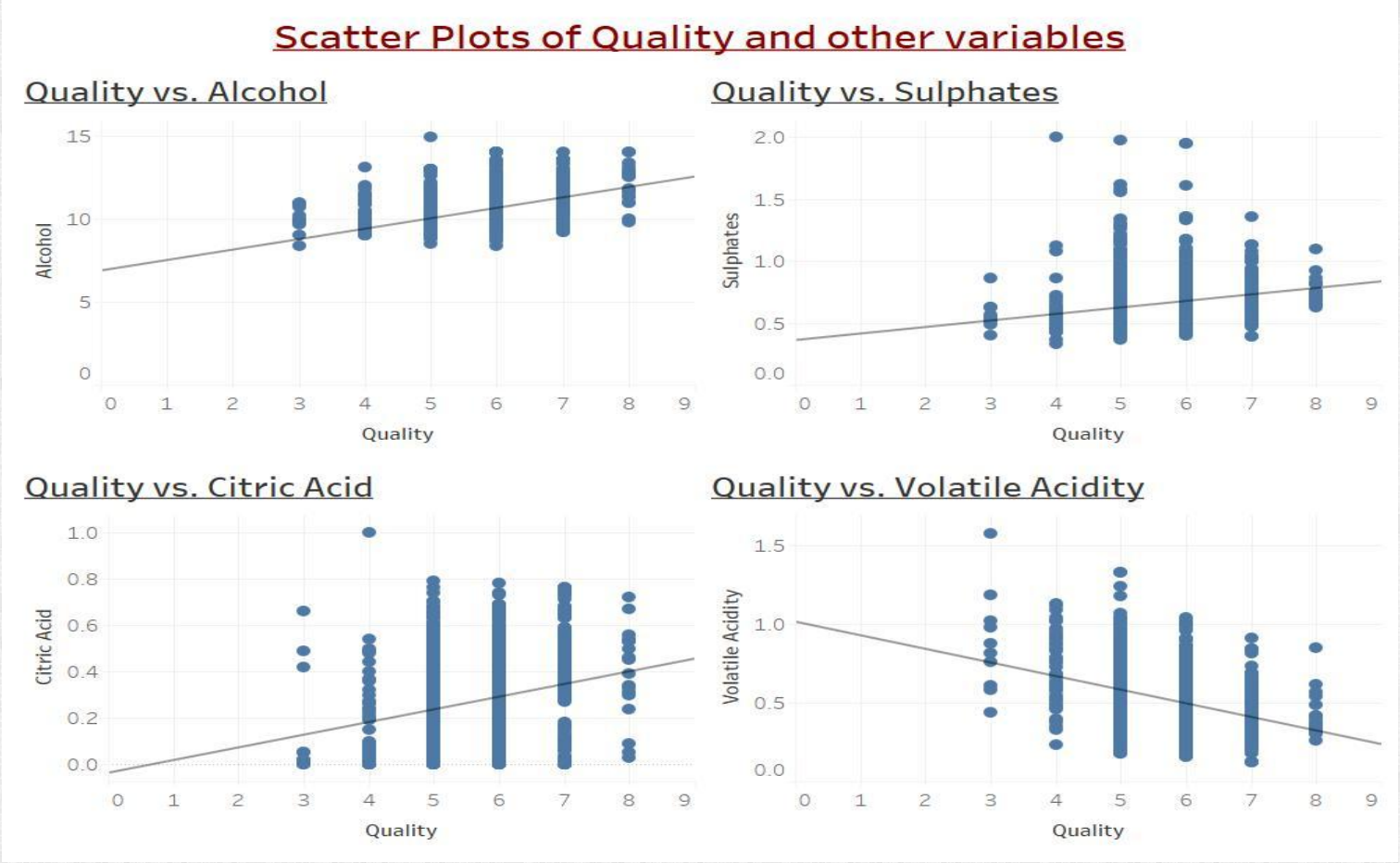
	Class: bad	Class: good	Class: ok
Sensitivity	0.61	0.82	0.65
Specificity	0.84	0.84	0.79
Balanced Accuracy	0.73	0.83	0.72
Precision	0.13	0.44	0.94
F1 Score	0.21	0.57	0.77

Correlation Matrix

From the following matrix, we observe that following variables are the most correlated:
-‘alcohol’: positive correlation
-‘sulphates’: positive correlation
-‘citric acid’: positive correlation
-‘volatile acidity’: negative correlation.



Quality Correlation Plot



Conclusion

Model	Accuracy	Takeaways
Random Forest	**75.52%**	**Best model overall, better precision, sensitivity and specificity**
Support Vector Machine	**67.15%**	**Worse than RF in every metric**
XGBoost	**75.31%**	**Excellent accuracy, better precision and sensitivity for some classes**

1. Most significant variables to assess the quality of wine: Alcohol, Volatile Acidity, pH, citric acid and density

2. From our findings, a good quality wine should have high alcohol (increases density), low volatile acidity, low pH (3 to 4), and some presence of citric acid.