WINE QUALITY PREDICTION

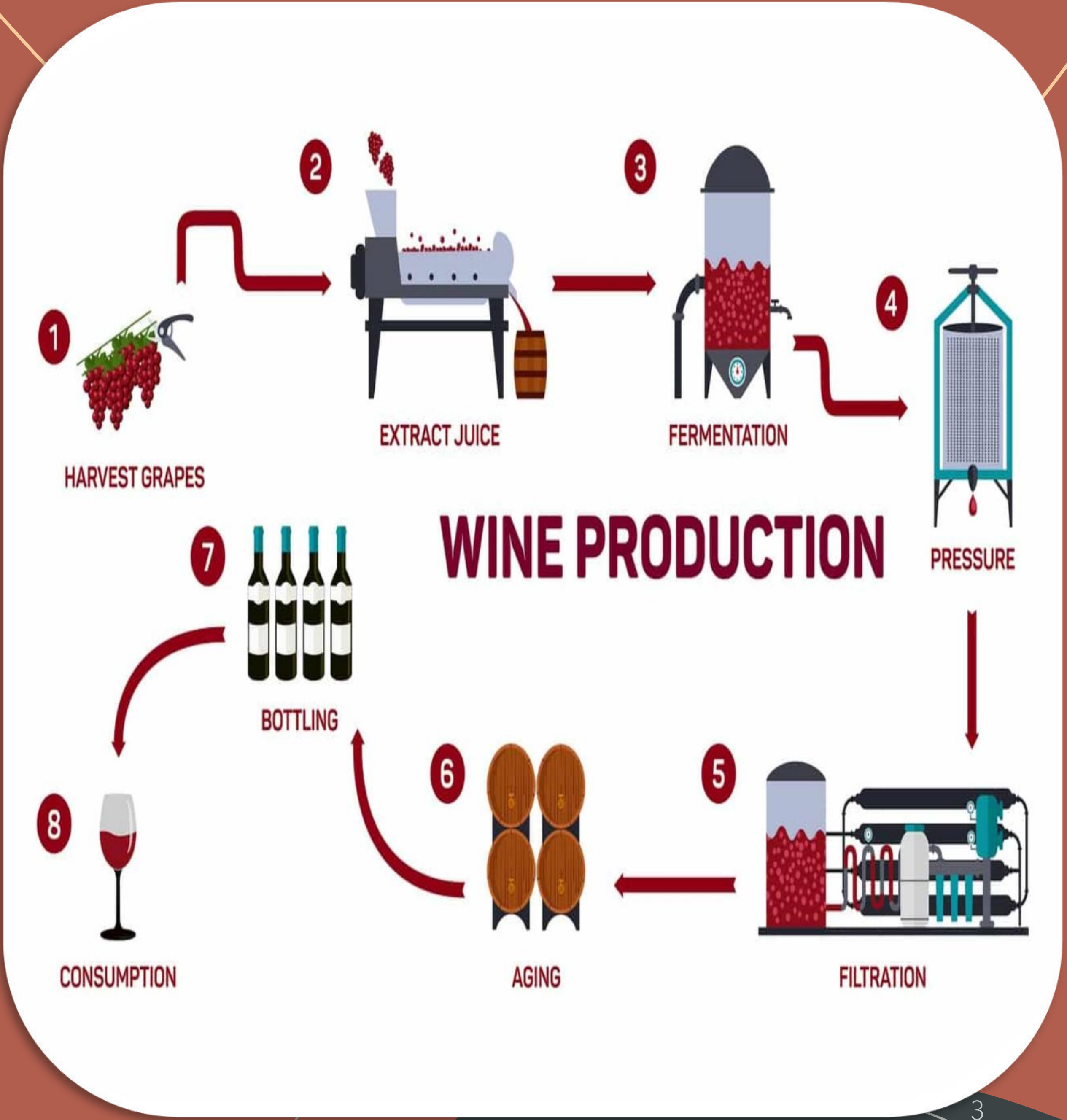
Francis Osei

Agenda

1. Introduction
2. Data Preparation
3. Machine Learning Models
4. Results and Discussion
5. Summary

Introduction

- Wine is an alcoholic drink typically made from fermented grapes.
- It is very difficult to assess the quality of wine just by reading the label.
- Many quality wines are made to mature over a long period of time before finally reaching their best.



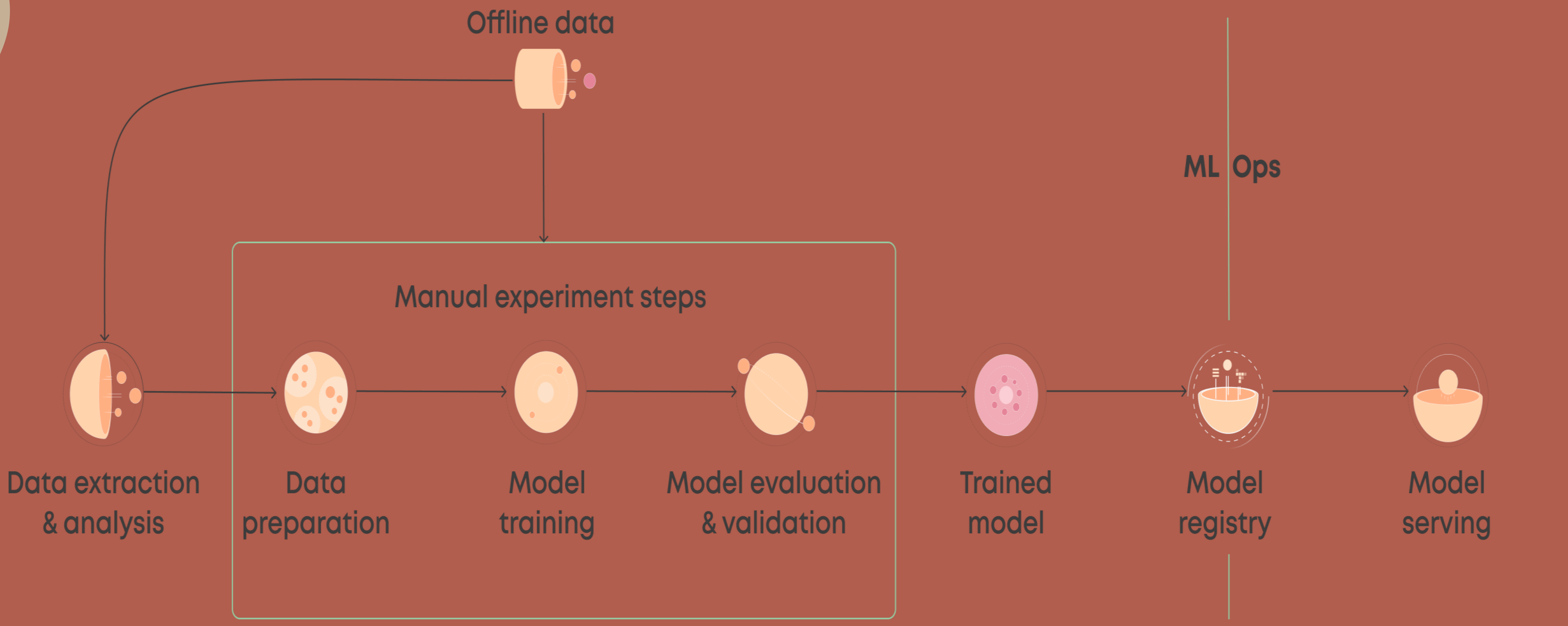
Problem Statement

- Price is only an indication of quality when similar wines are being compared.
- Determining wine quality is a complex task for companies due to various factors involved in assessing and evaluating wine.

Objective

- To discover the quality of a wine (red and white variants of the Portuguese "Vinho Verde" wine) considering only its chemical properties.
- To identify the qualities of an excellent wine that are most indicative.

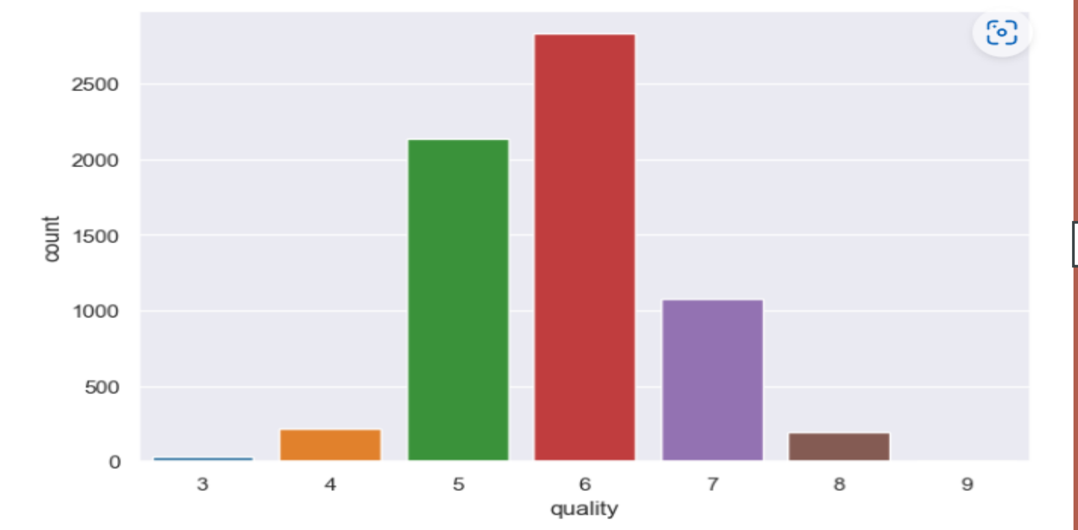
Machine Learning Pipeline



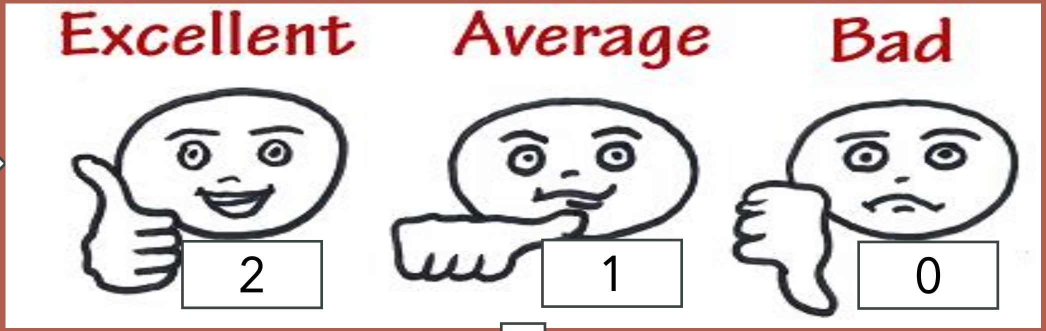
Data Description

Variables	Description
Type	Red/White wine
Fixed acidity	The acids that naturally occur in the grapes used to ferment the wine and carry over into the wine.
Volatile acidity	Acids that evaporate at low temperatures
Citric acid	Citric acid is used as an acid supplement which boosts the acidity of the wine.
Residual sugar	The amount of sugar remaining after fermentation stops.
Chlorides	The amount of chloride salts (sodium chloride) present in the wine
Free sulfur dioxide	The free form of SO exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion
Total sulfur dioxide	The amount of free and bound forms of S02
Density	The density of wine juice depending on the percent alcohol and sugar content
pH	A measure of the acidity of wine
Sulphates	Amount of potassium sulphate as a wine additive which can contribute to sulfur dioxide gas (S02) levels
Alcohol	How much alcohol is contained in each volume of wine
Quality	score between 0 (very bad) and 10 (very excellent) by wine experts

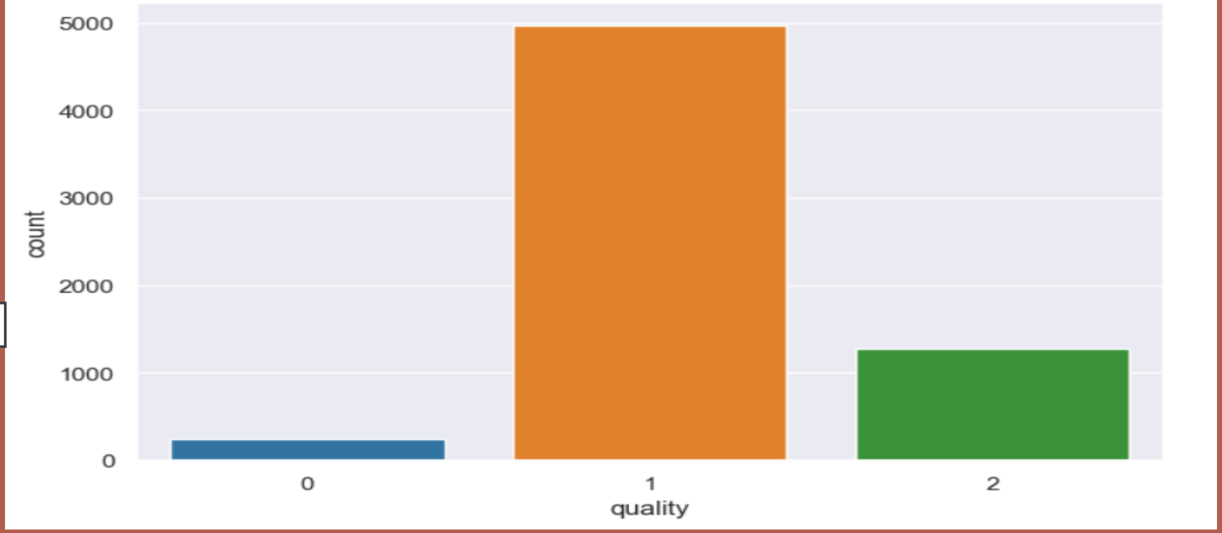
Data Preparation (Data Visualization)



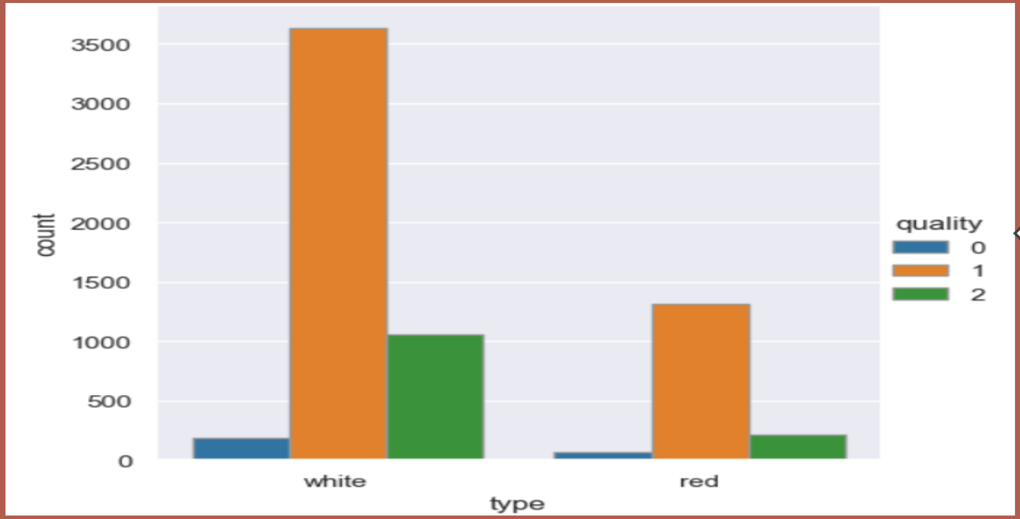
(A)



(B)

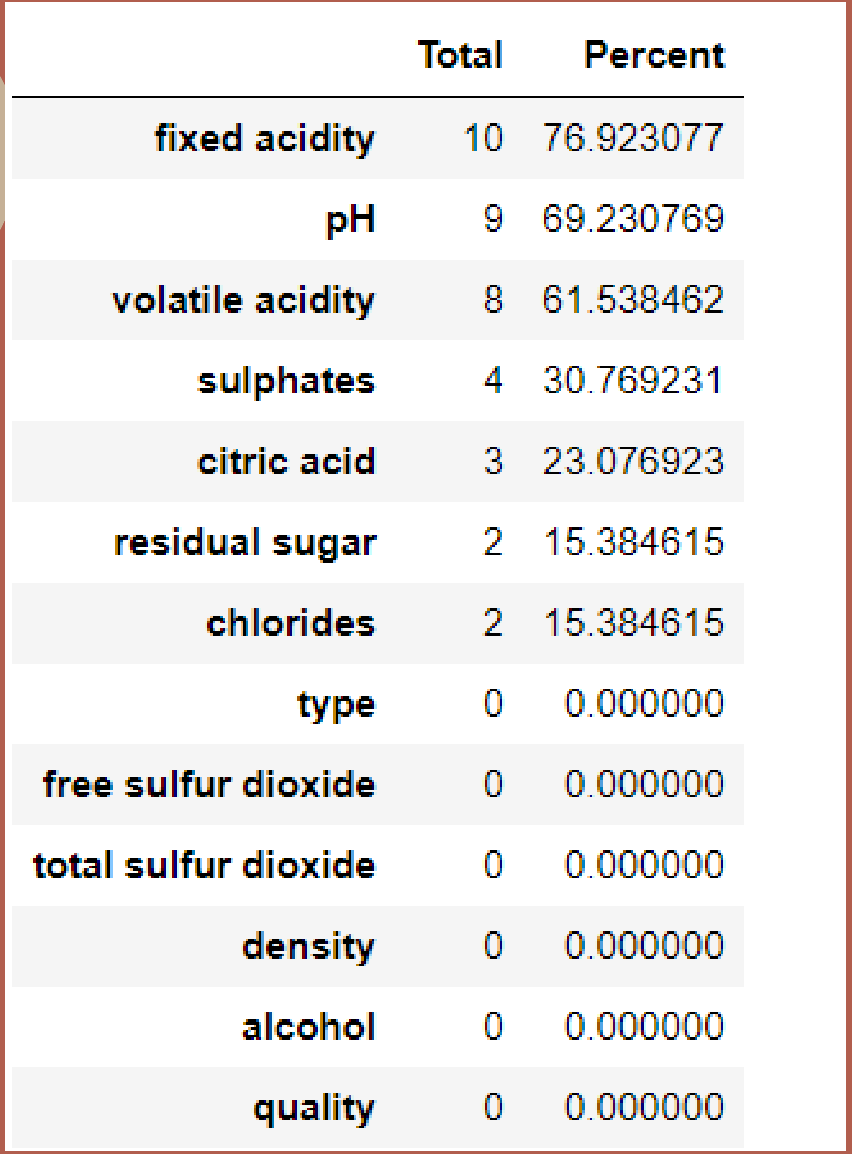


(C)



(D)

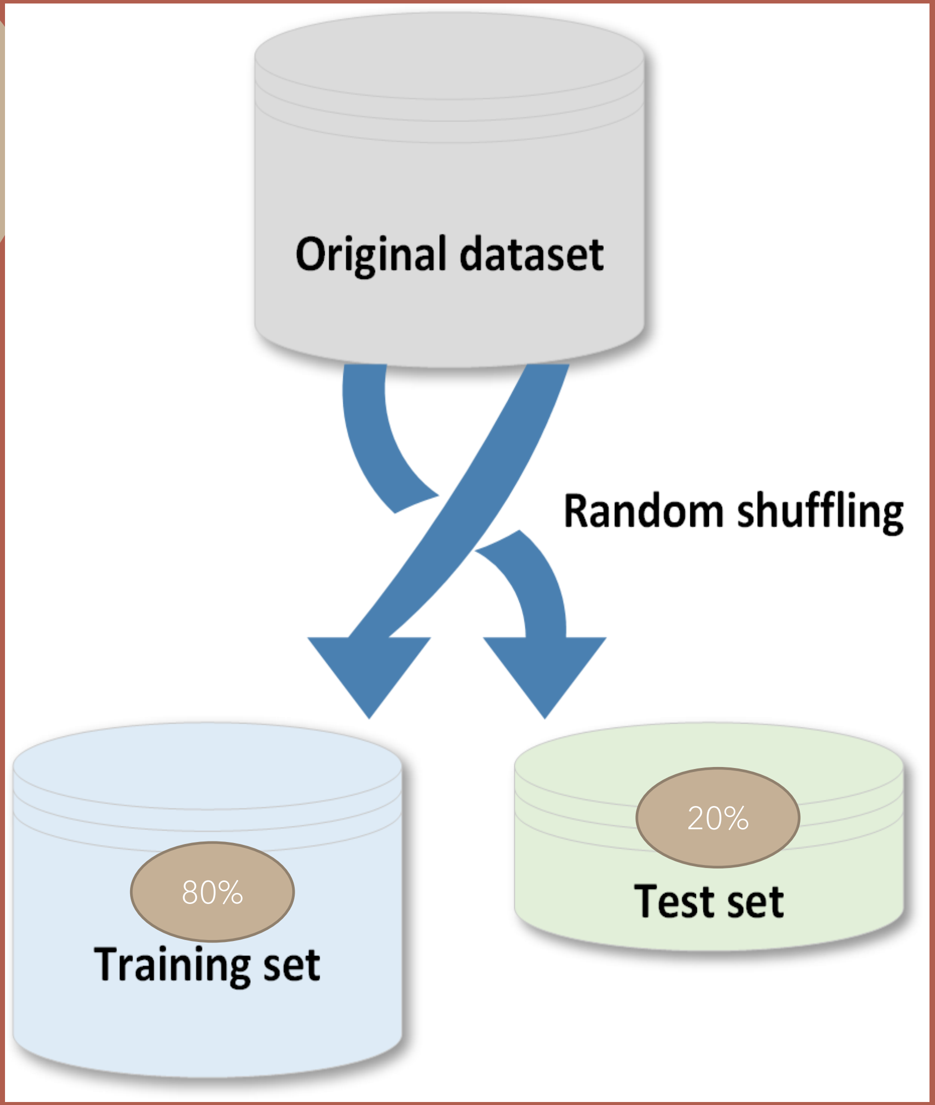
Data Preparation (Missing Values)



	Total	Percent
fixed acidity	10	76.923077
pH	9	69.230769
volatile acidity	8	61.538462
sulphates	4	30.769231
citric acid	3	23.076923
residual sugar	2	15.384615
chlorides	2	15.384615
type	0	0.000000
free sulfur dioxide	0	0.000000
total sulfur dioxide	0	0.000000
density	0	0.000000
alcohol	0	0.000000
quality	0	0.000000

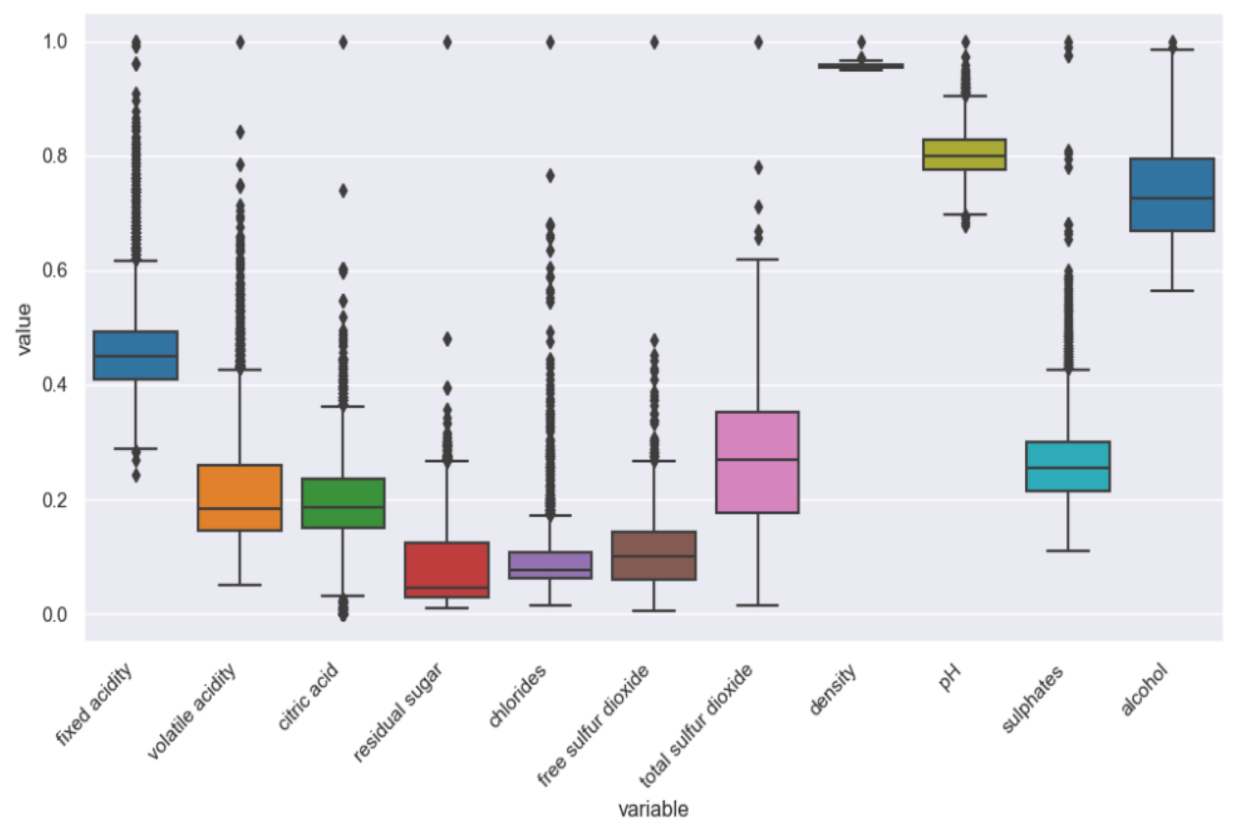
- Missing data can significantly affect the conclusions that can be drawn from the data.
- 0.005% of information is missing

Data Preparation (Data Splitting)

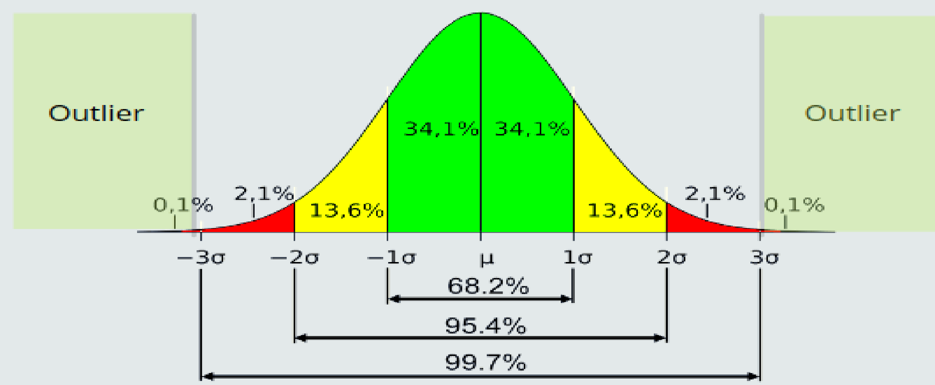


- Data splitting is an essential step in machine learning and data analysis.
- Training datasets comprise samples used to fit models under construction.
- A test dataset is a separate sample to provide an unbiased final evaluation of a model fit.

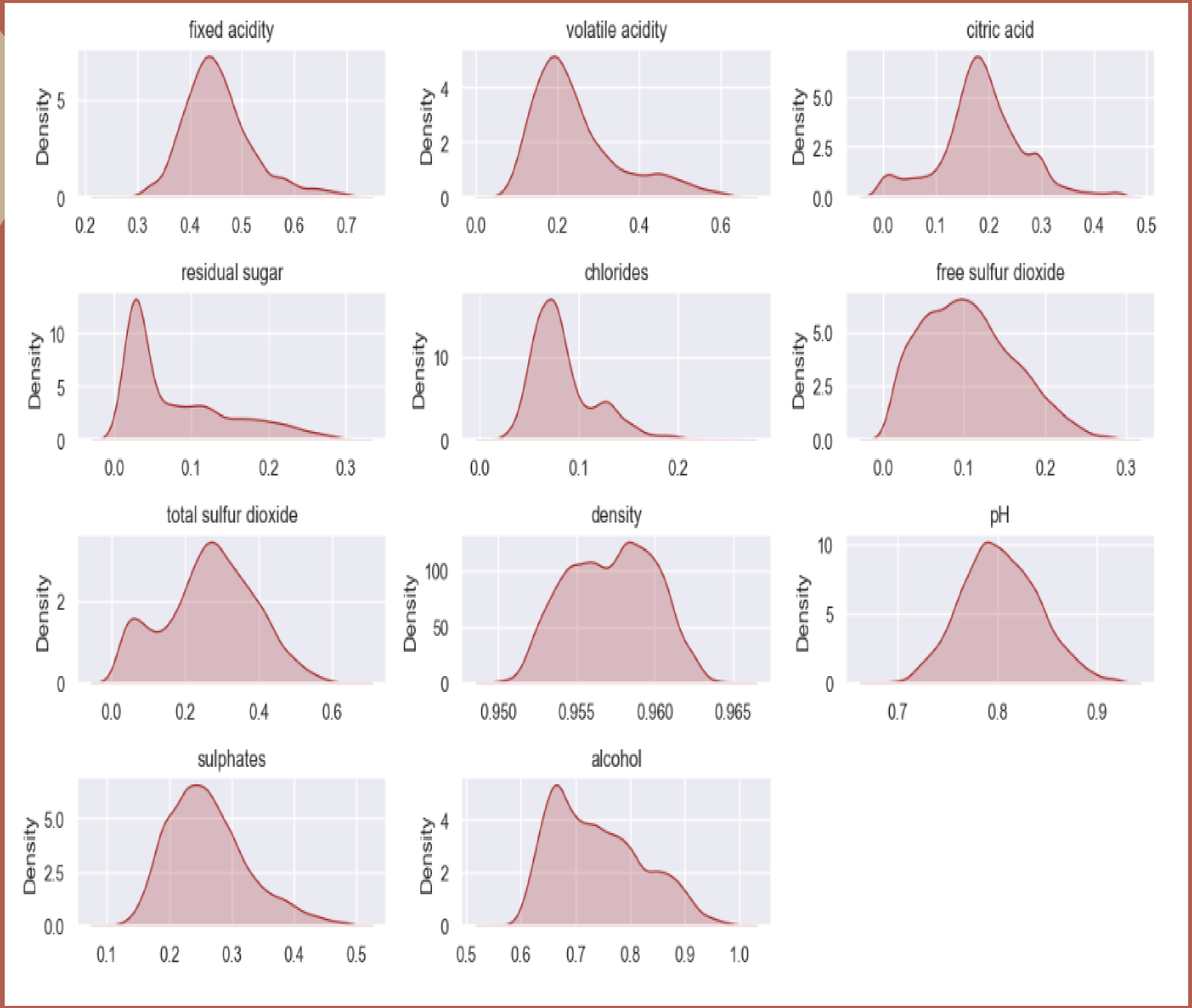
Data Preparation (Outliers)



- Outliers in the data may causes problems during model fitting
- Outliers may inflate the error metrics which give higher weights to large errors

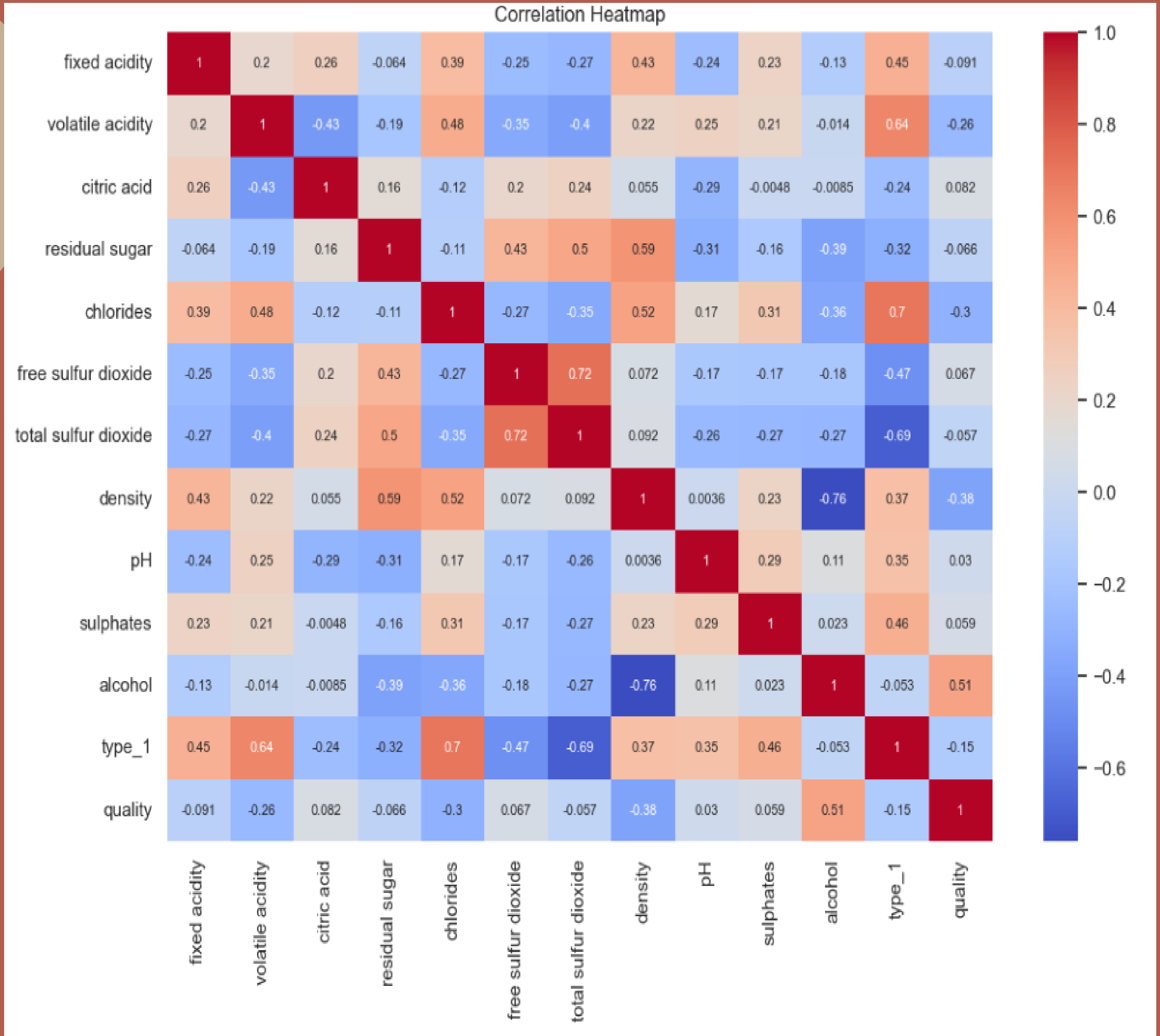


Data Preparation (Distribution)



- Different machine learning models have different assumptions about the distribution of the data.
- The distribution can provide insights into the normality or non-normality of the data.

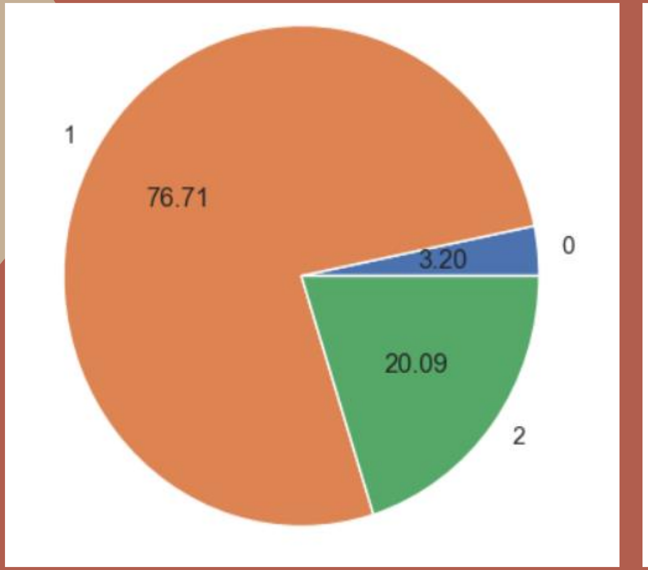
Data Preparation (Multicollinearity)

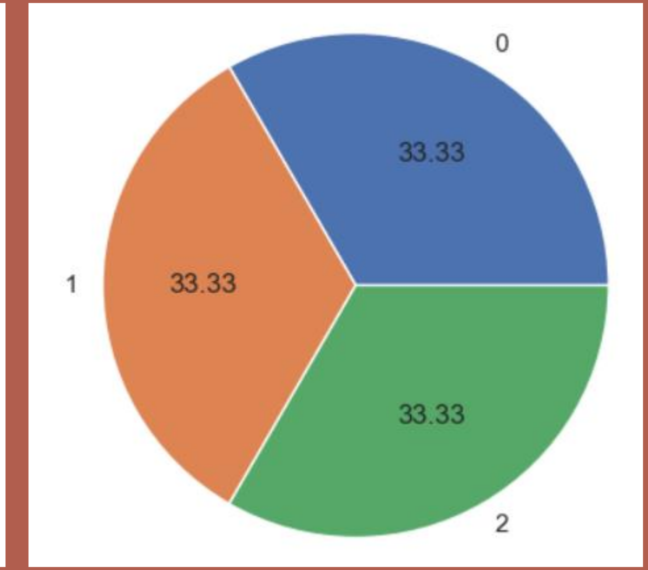


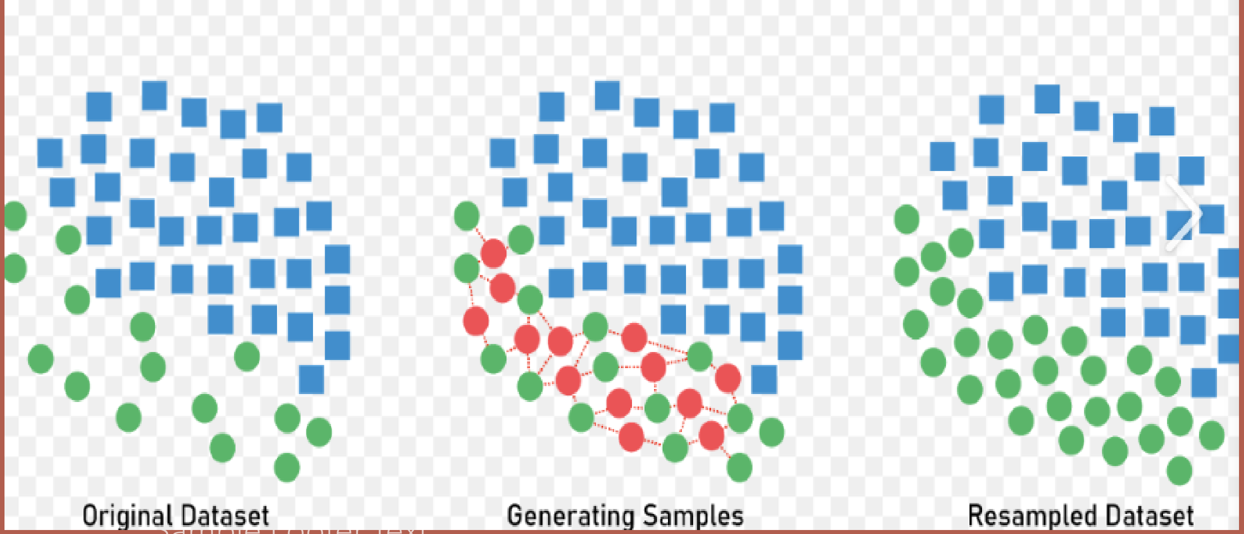
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	type_1	quality
fixed acidity	1	0.2	0.26	-0.064	0.39	-0.25	-0.27	0.43	-0.24	0.23	-0.13	0.45	-0.091
volatile acidity	0.2	1	-0.43	-0.19	0.48	-0.35	-0.4	0.22	0.25	0.21	-0.014	0.64	-0.26
citric acid	0.26	-0.43	1	0.16	-0.12	0.2	0.24	0.055	-0.29	-0.0048	-0.0085	-0.24	0.082
residual sugar	-0.064	-0.19	0.16	1	-0.11	0.43	0.5	0.59	-0.31	-0.16	-0.39	-0.32	-0.066
chlorides	0.39	0.48	-0.12	-0.11	1	-0.27	-0.35	0.52	0.17	0.31	-0.36	0.7	-0.3
free sulfur dioxide	-0.25	-0.35	0.2	0.43	-0.27	1	0.72	0.072	-0.17	-0.17	-0.18	-0.47	0.067
total sulfur dioxide	-0.27	-0.4	0.24	0.5	-0.35	0.72	1	0.092	-0.26	-0.27	-0.27	-0.69	-0.057
density	0.43	0.22	0.055	0.59	0.52	0.072	0.092	1	0.0036	0.23	-0.76	0.37	-0.38
pH	-0.24	0.25	-0.29	-0.31	0.17	-0.17	-0.26	0.0036	1	0.29	0.11	0.35	0.03
sulphates	0.23	0.21	-0.0048	-0.16	0.31	-0.17	-0.27	0.23	0.29	1	0.023	0.46	0.059
alcohol	-0.13	-0.014	-0.0085	-0.39	-0.36	-0.18	-0.27	-0.76	0.11	0.023	1	-0.053	0.51
type_1	0.45	0.64	-0.24	-0.32	0.7	-0.47	-0.69	0.37	0.35	0.46	-0.053	1	-0.15
quality	-0.091	-0.26	0.082	-0.066	-0.3	0.067	-0.057	-0.38	0.03	0.059	0.51	-0.15	1

- Correlation is statistical technique which determines how one variables changes in relation with the other variable.
- Multicollinearity will severely affect performance outside of the original data sample.

Data Preparation (Imbalance Dataset)

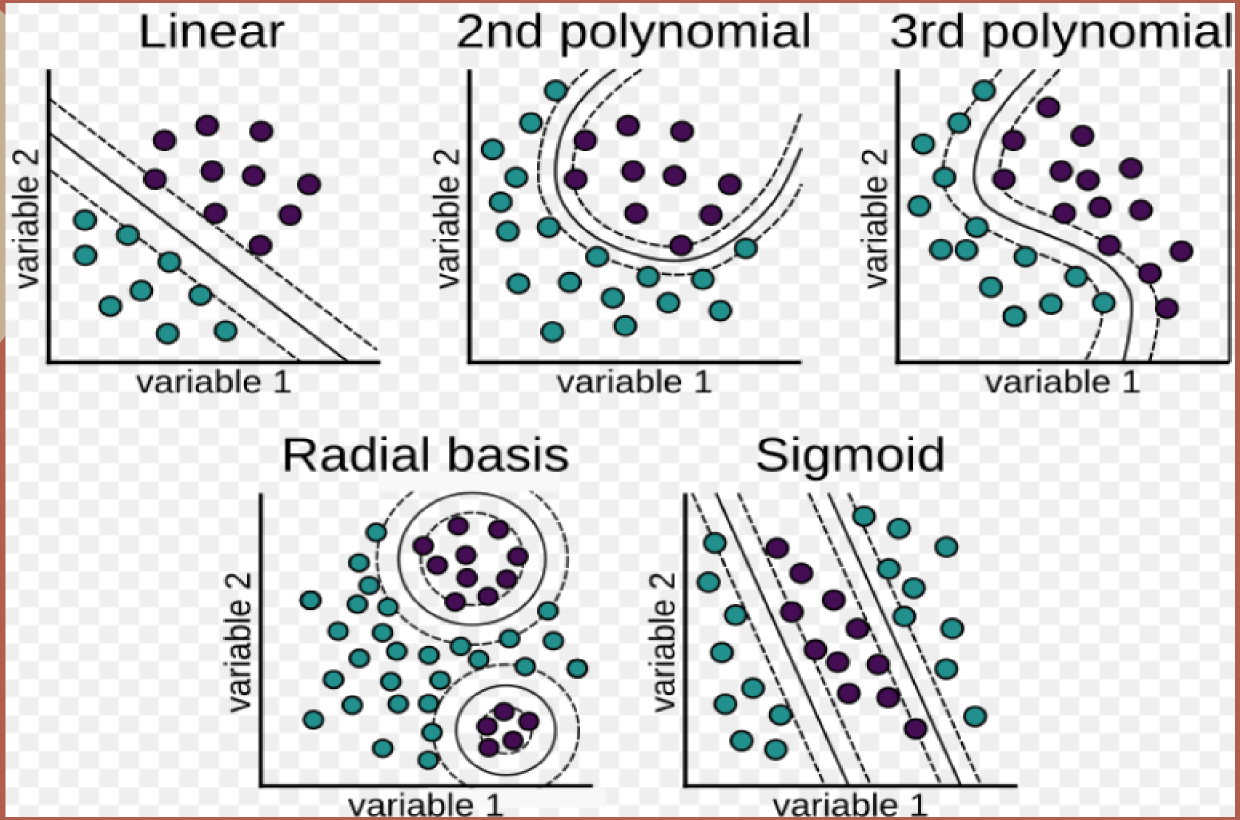






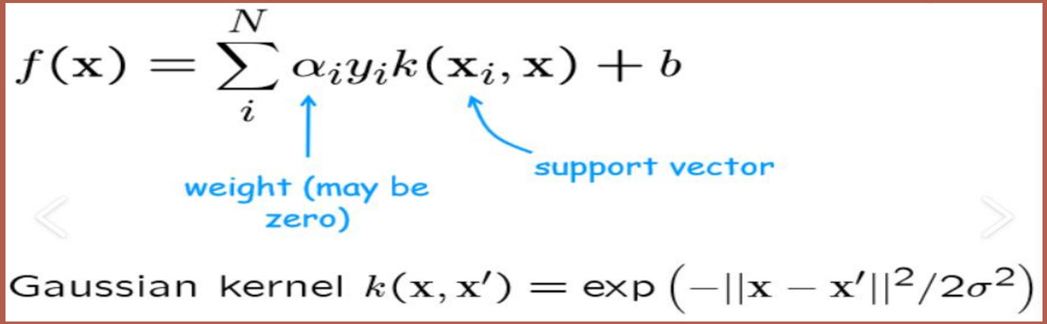
- Imbalance dataset arises when one set of classes dominate over another set of classes.
- it causes the machine learning model to be more biased towards majority class.
- SMOTE techniques generates the virtual training records by linear interpolation for the minority class.

Model training (Support Vector Machine)



Source: The Support Vector Machine: Basic Concept | by Aminah Mardiyyah Rufai | The Startup | Medium

- Support vector machine results from enlarging the feature space in a specific way using kernels
- The Radial Basis Function depends only on the distance between the input and some fixed point



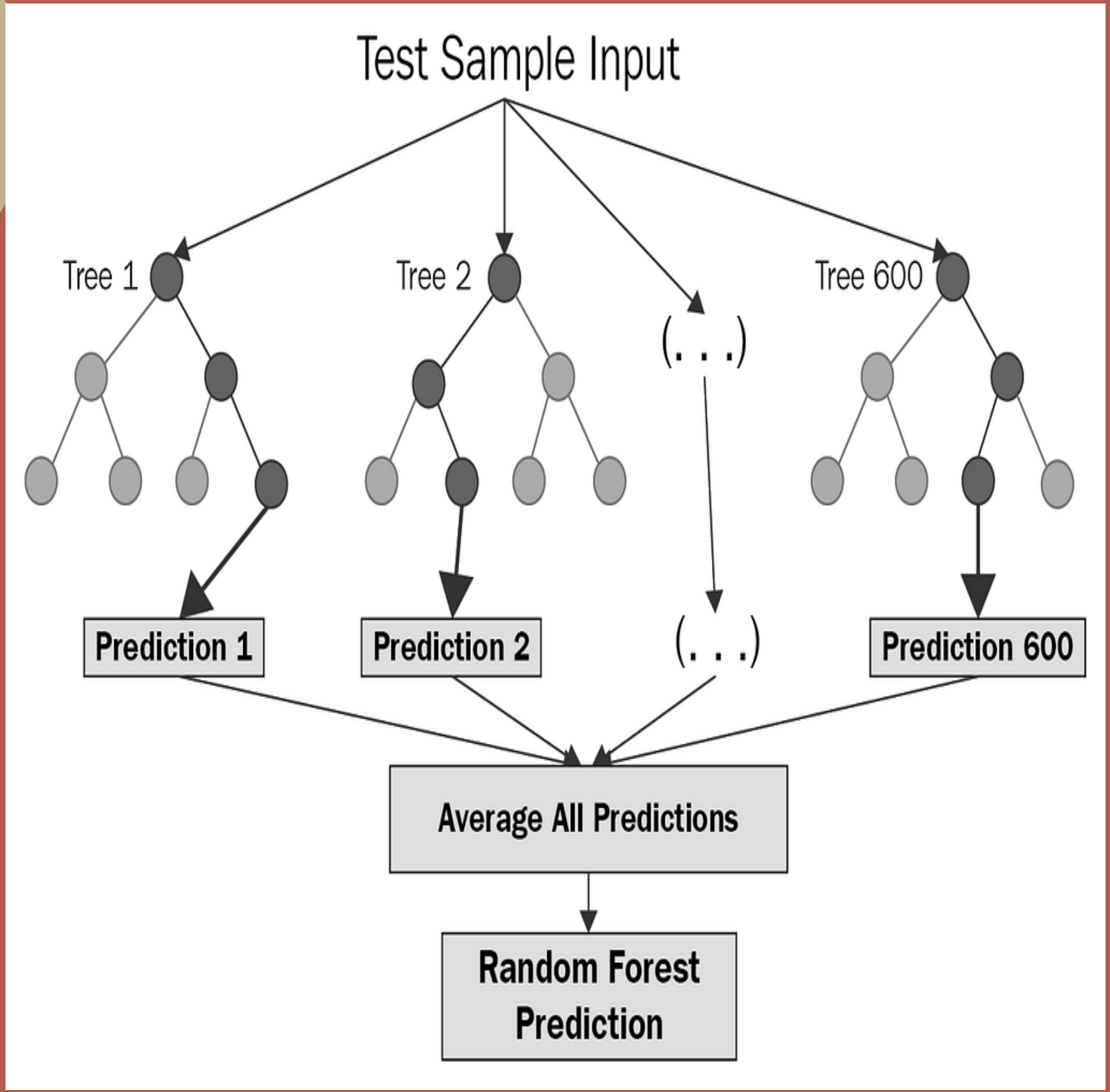
$$f(\mathbf{x}) = \sum_{i}^{N} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b$$

weight (may be zero)

support vector

Gaussian kernel $k(\mathbf{x}, \mathbf{x}') = \exp\left(-||\mathbf{x} - \mathbf{x}'||^2/2\sigma^2\right)$

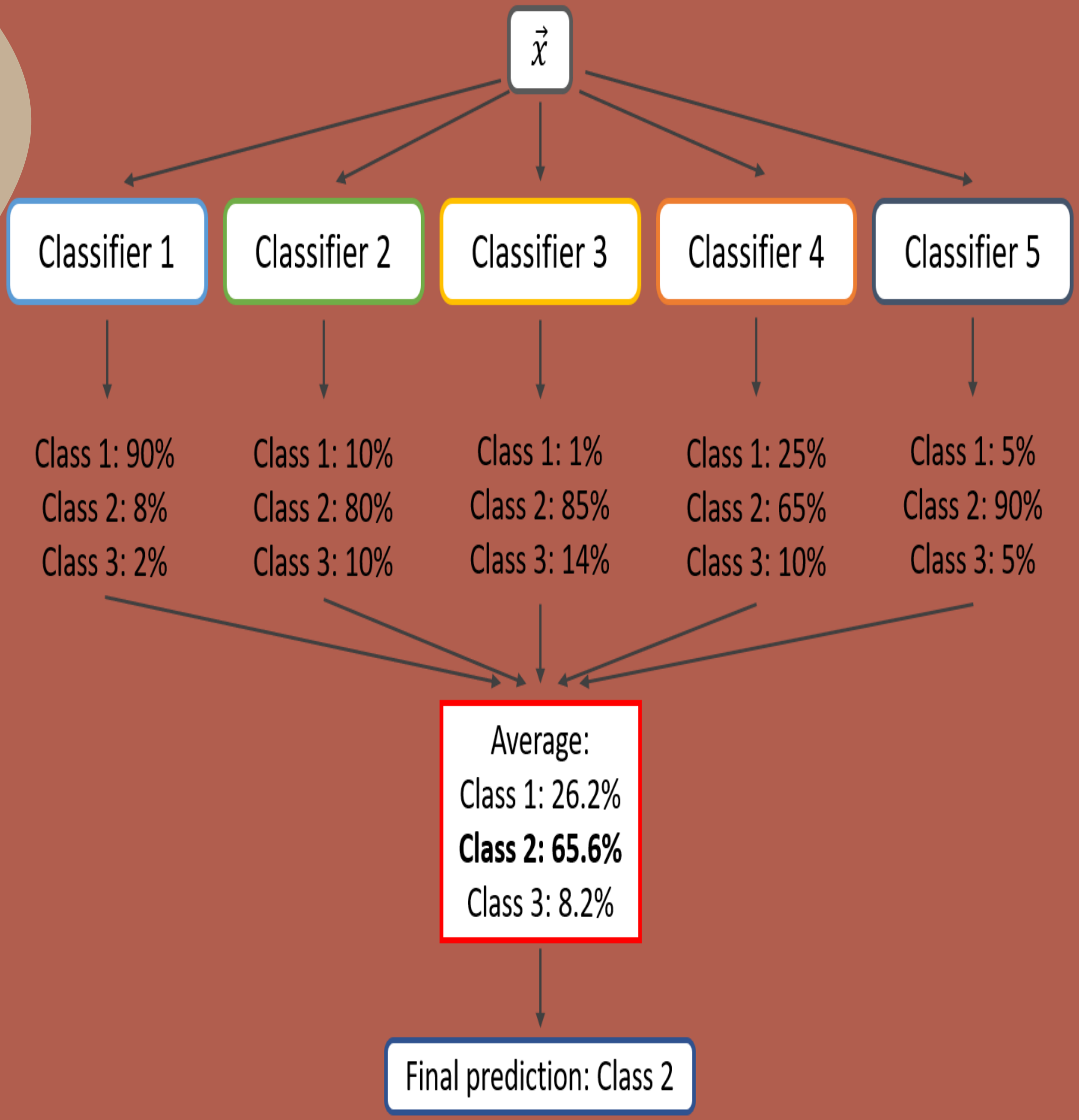
Model training (Random Forest)



Source: Databricks

- Random forest ensemble is an ensemble of decision trees and a natural extension of bagging.
- Random forest handle large datasets with high dimensionality.
- It provides feature importance measures, allowing you to understand the relative importance of different features in the model.

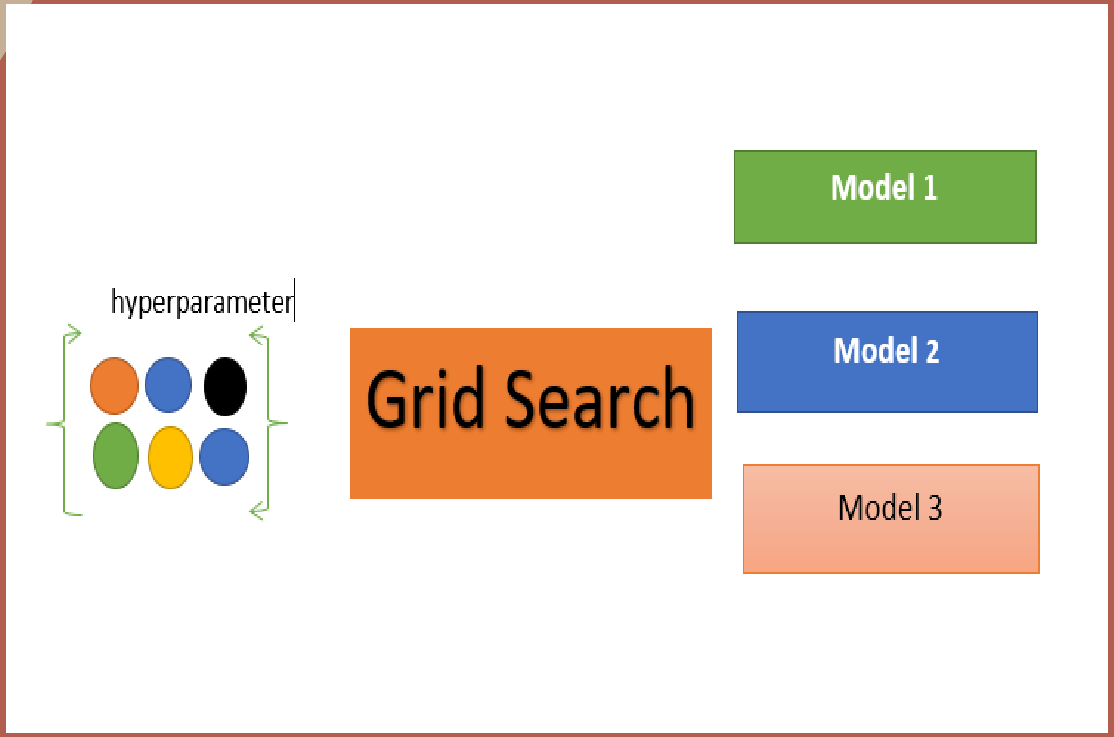
Model training (Voting Classifier)



Source: https://iq.opengenus.org/voting-classifier/

- The Voting Classifier is an ensemble learning method that combines multiple individual classifiers to make predictions by voting.
- The predicted class probabilities from each classifier are averaged, and the class with the highest average probability is selected as the final prediction.

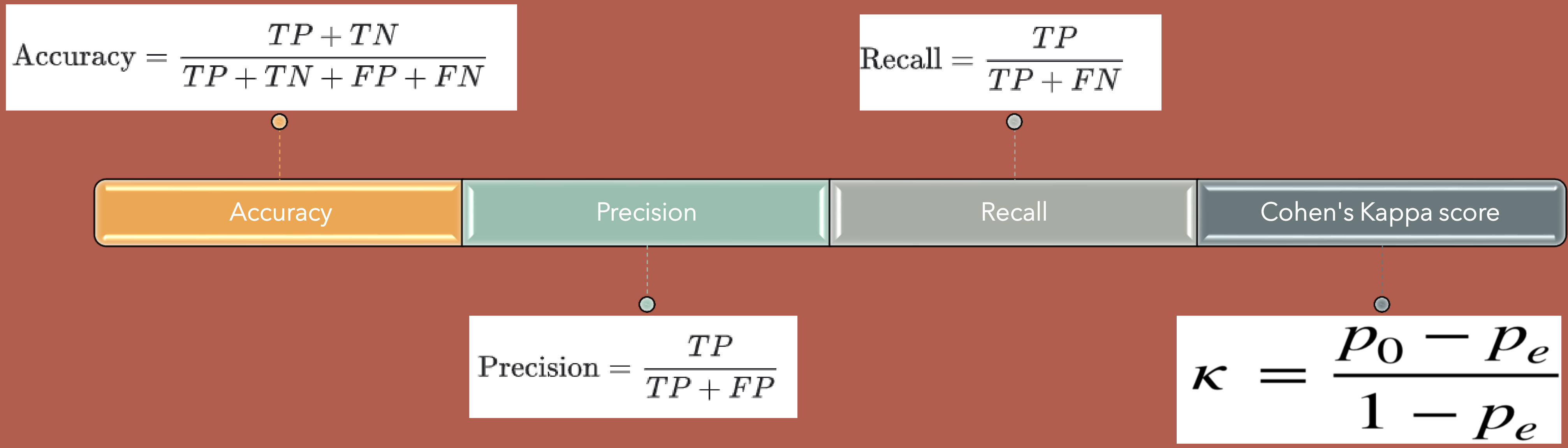
Model training (Hyperparameter Tuning)



- Hyperparameter tuning is an essential step in machine learning to optimize the performance of your models.
- Gridsearch tries all possible combinations of hyperparameters and evaluates each combination using cross-validation or a validation set.

Model Evaluation (Evaluation Metrics)

- Model evaluation is an essential step in machine learning to assess the performance and effectiveness of trained models.



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Accuracy

Precision

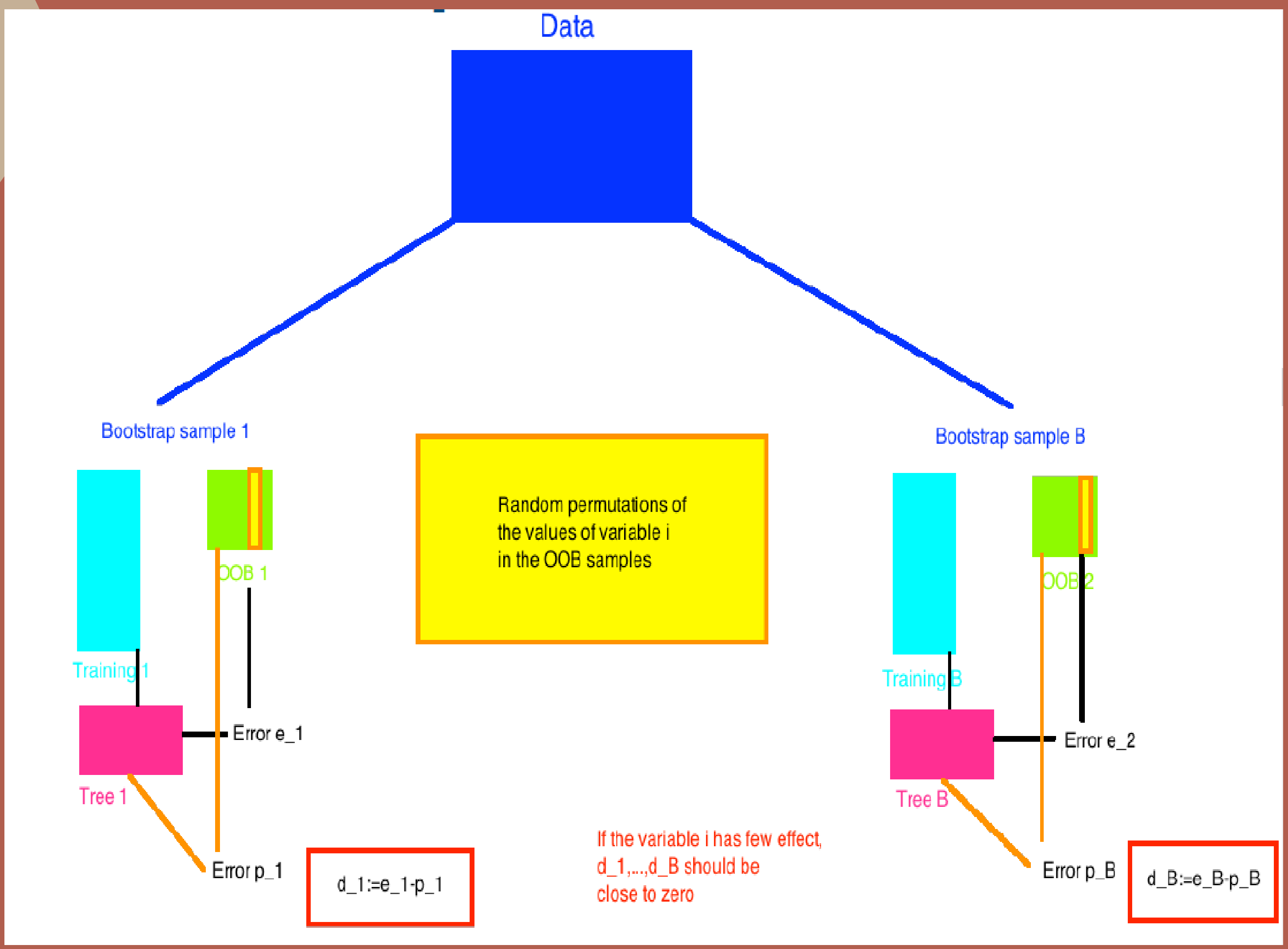
Recall

Cohen's Kappa score

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

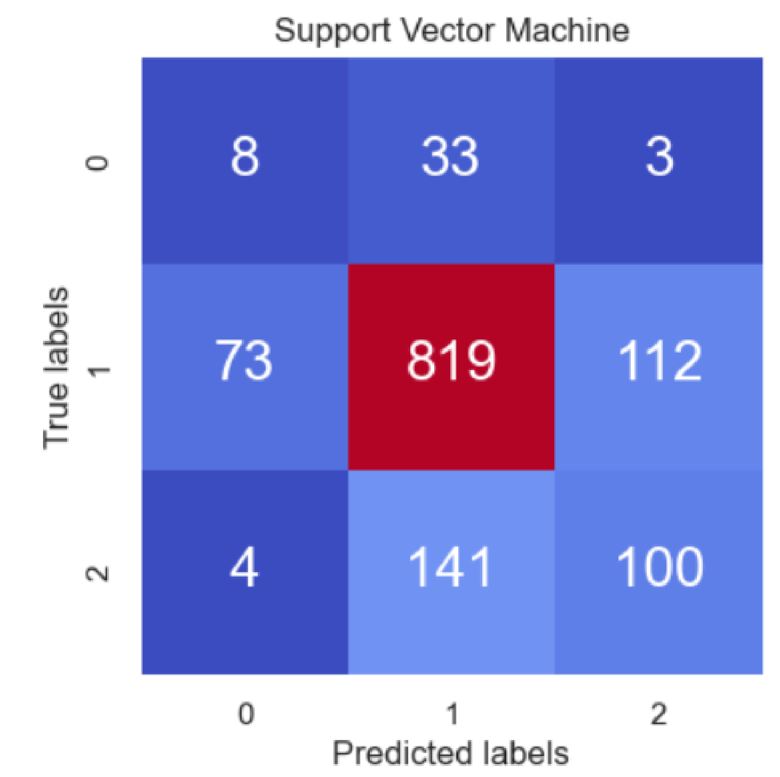
Model Evaluation (Variable Importance)



- Feature importance is a technique used in machine learning to determine the relevance or contribution of each feature in a predictive model.

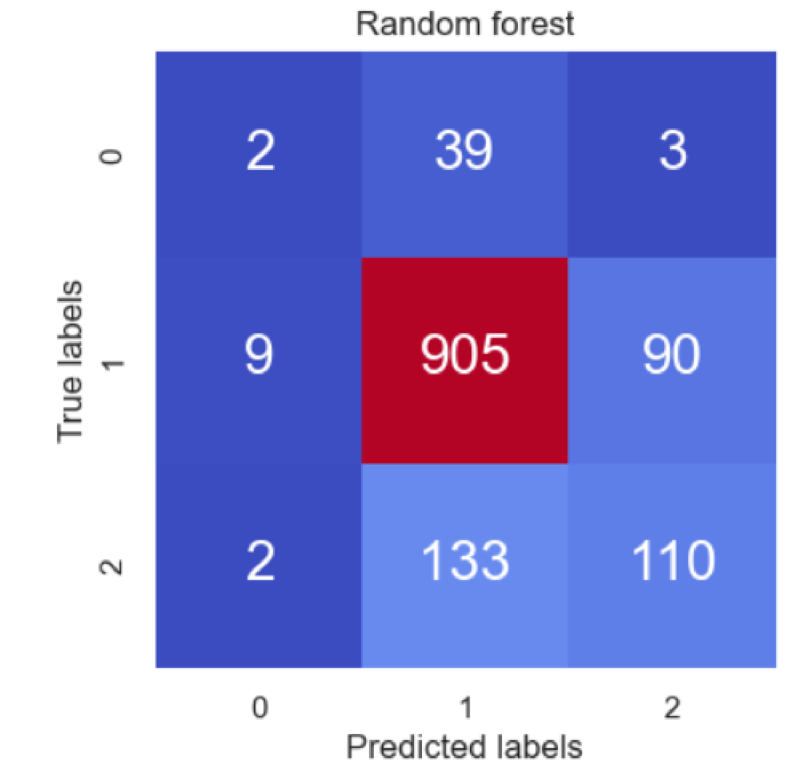
Result and Discussion (1)

- Voting classifiers can identity bad wines compared to the other models
- ML models are very good in identifying average wines



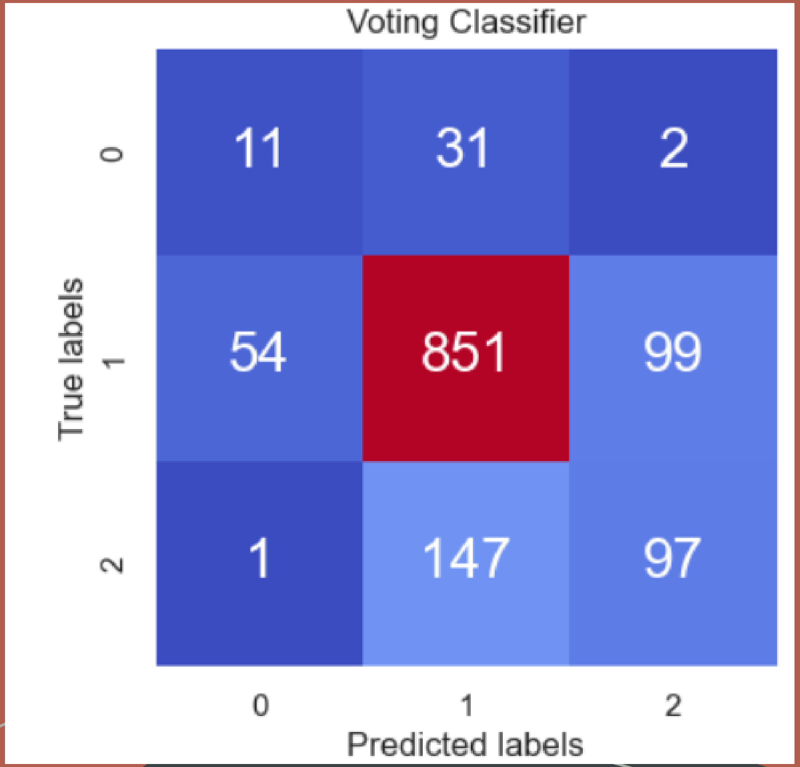
Support Vector Machine

True labels \ Predicted labels	0	1	2
0	8	33	3
1	73	819	112
2	4	141	100



Random forest

True labels \ Predicted labels	0	1	2
0	2	39	3
1	9	905	90
2	2	133	110



Voting Classifier

True labels \ Predicted labels	0	1	2
0	11	31	2
1	54	851	99
2	1	147	97

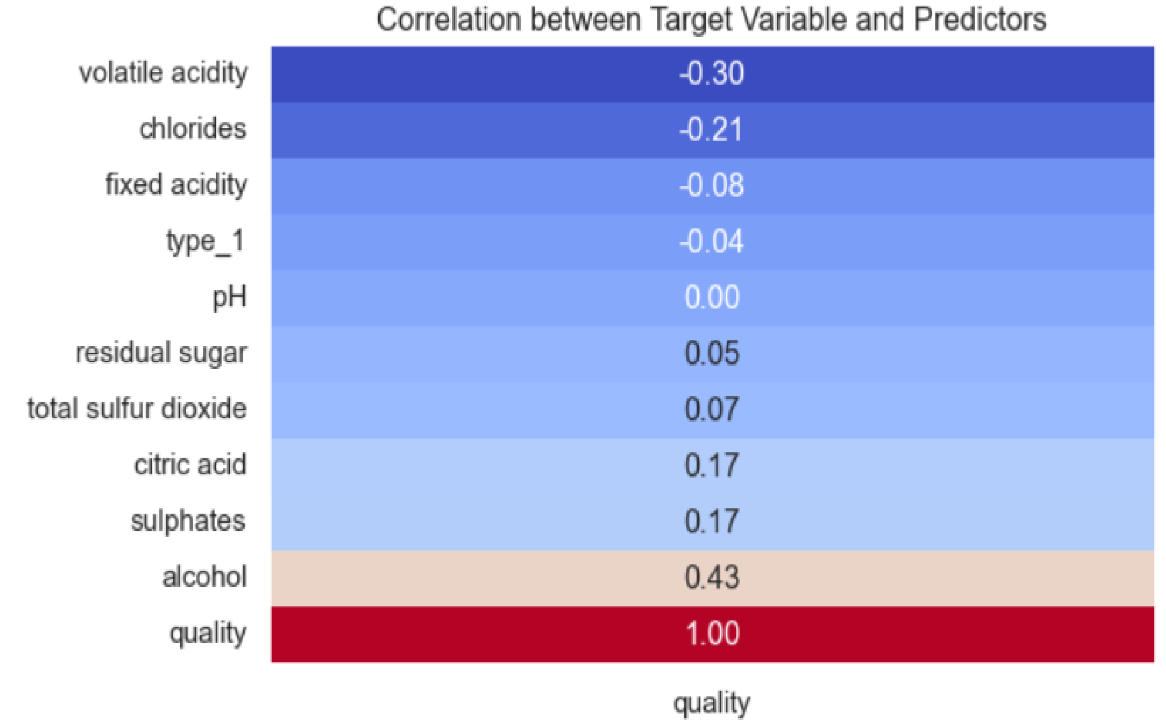
Result and Discussion (2)

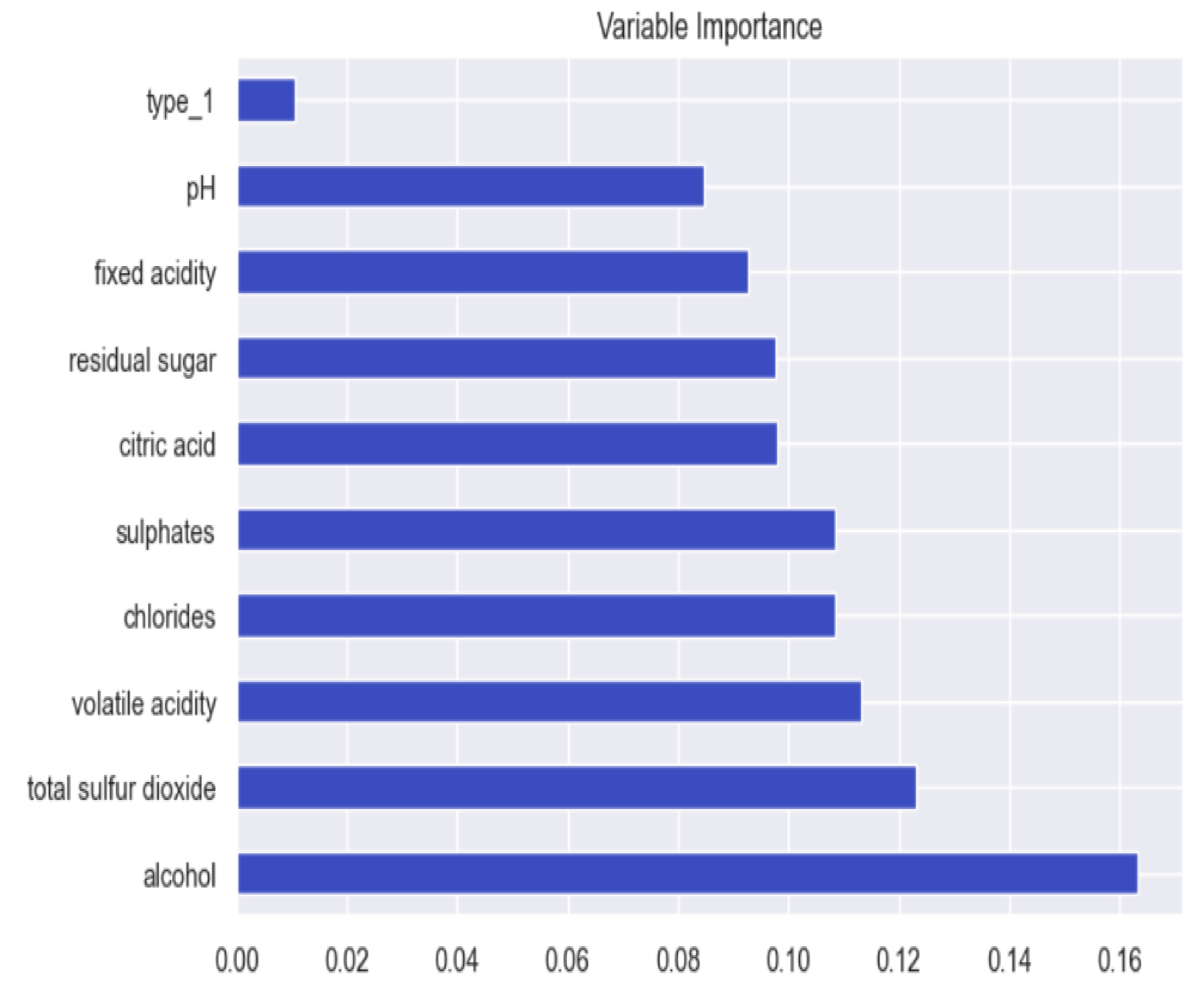
- Classification Report

Models	Accuracy	Precision (weighted average)	Recall (weighted average)	Kohen's Kappa Score
Support Vector Machine	0.72	0.73	0.72	0.23
Random Forest	0.78	0.75	0.78	0.31
Voting Classifier	0.74	0.74	0.74	0.27

Result and Discussion (3)

- Feature Importance





Summary

- ML models are very good in identifying average wines, but it's terrible in finding poor wines.
- In Business perspective, the voting classify will be preferred over the Random forest
- More features are needed to improve the model performance

References

- Olivier Lopez (2022), Machine Learning and Insurance : Workshop Actuarial Science
- Martin Gubri & Salah Ghamizi (2022), Applied Machine Learning Lecture Note
- Gareth James & Daniela Witten & Trevor Hastie & Robert Tisbshirani (2013), An Introduction to Statistical Learning

Thank You

https://github.com/Franosei/Wine-Quality