RED WINE QUALITY PREDICTION

SUPERVISED LEARNING

CAPSTONE

OVERVIEW OF THE STUDY

- For the current capstone project, Red Wine Quality Prediction is considered with various **Supervised Machine Learning** algorithms
- We have used machine learning to determine which **physiochemical properties** make a wine **'good' !**
- The **wine data** used in this study comes from the north-west region, named **Minho, of Portugal**, and we have used only the data w.r.t. red wine
- This dataset is also available from the **UCI machine learning repository**,

https://archive.ics.uci.edu/ml/datasets/wine+quality

- **Relevant publication:** P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

AGENDA

1. DATA ANALYSIS AND FEATURE EVALUATION

2. MACHINE LEARNING MODELS

OVERVIEW OF THE DATA

- The dataset of red wine contains **1599 rows** and **12 columns.** The following are the variables in the dataset

- Input variables - **"features" :** They are based on physicochemical tests namely **fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol**

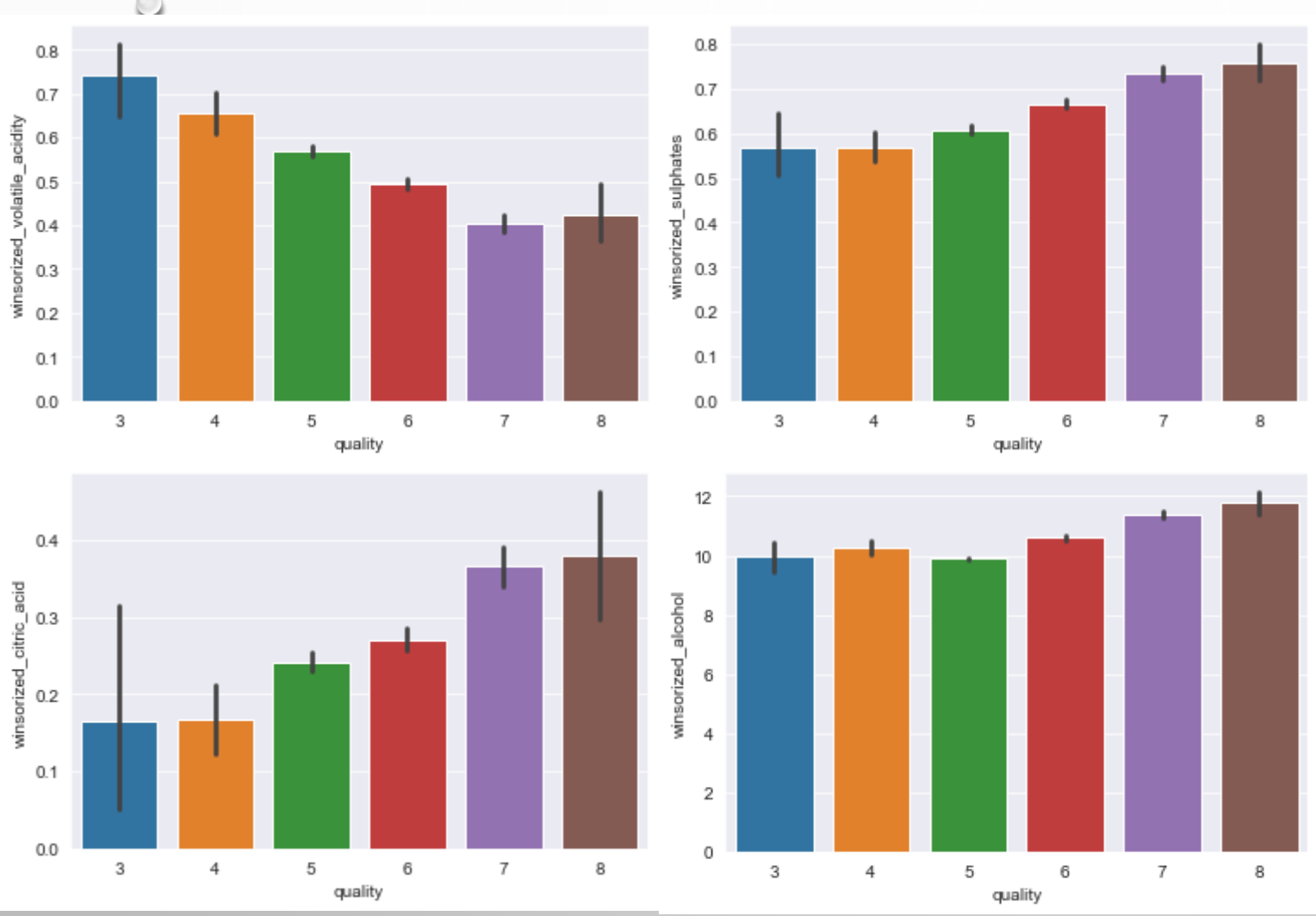
- Output variables - **"target" :** It is based on sensory data namely **quality** (score between 0 and 10)

- A basic exploratory data analysis reveals that there are **no missing values in the dataset**

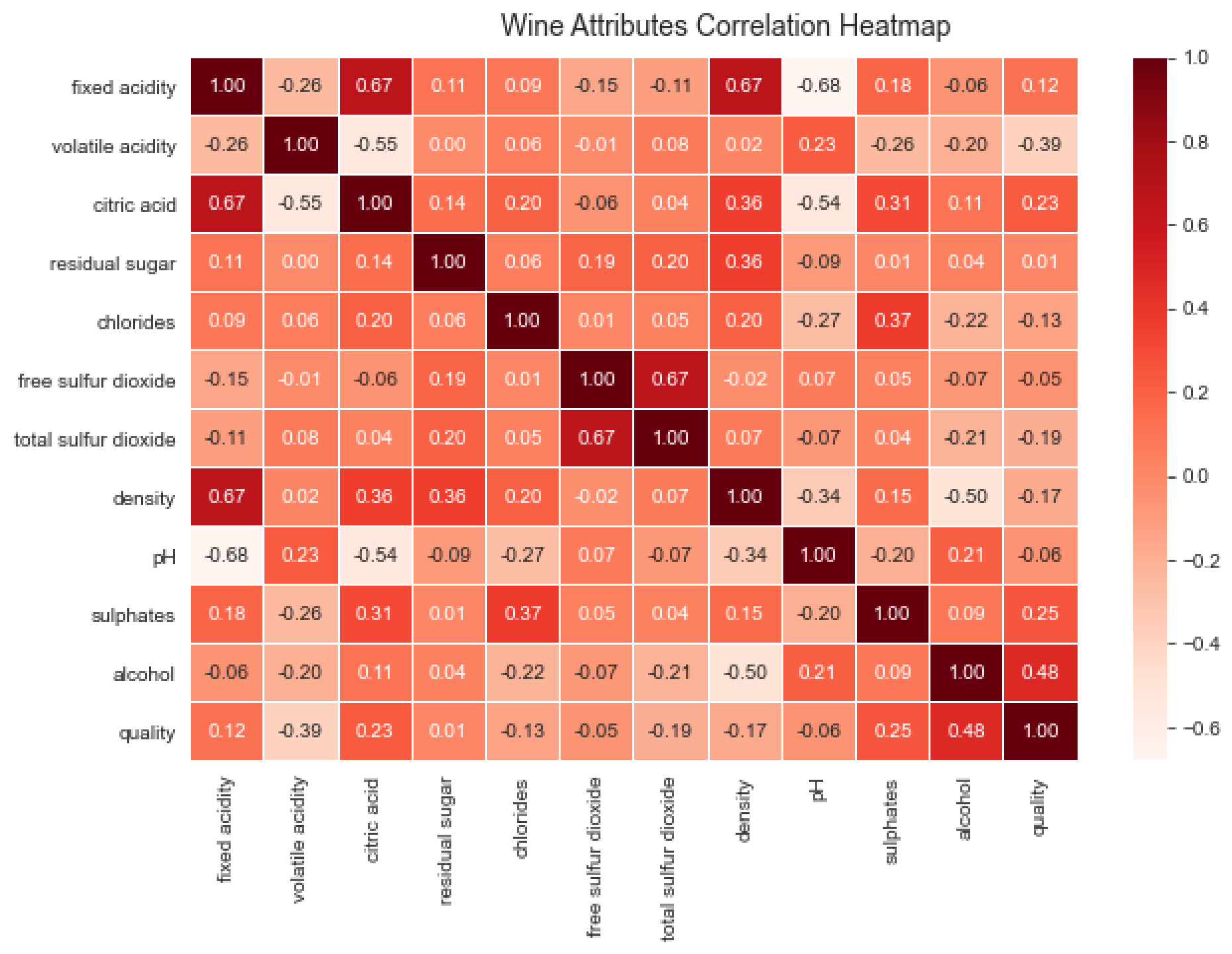
- All the variable columns were analyzed for **outliers** with **Tukey's method**

- The data was subjected to **winsorization** to eliminate outliers

BIVARIATE ANALYSIS



- Quality decreases with volatile acidity
- Quality increases with citric acid
- Quality increases with sulphates
- Quality increases with alcohol content



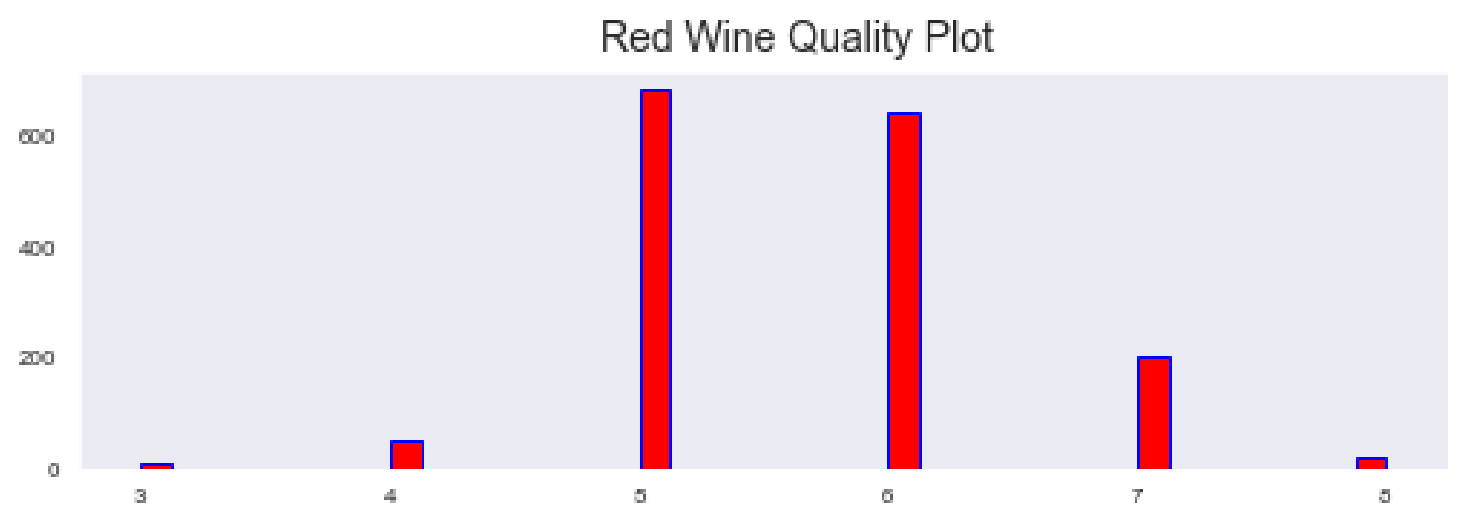
Variable Correlation Matrix

1. **Alcohol** level has **strongest** positive correlation (0.48) with **quality**

2. **Negative correlation** between **'pH'** and the **'fixed acidity'** of the wine (Wine is mostly acidic with pH 3-4)

TARGET "REVIEW" AND LABEL ENCODING

- Quality is represented by scores ranging from 0 to 10
- 0 is **"worst"** and 10 is **"best" in the data**
- Create a new **target** column **"Review"**
- **Label Encoding in Review:** Quality < = 5 as "0", Quality > 5 as "1"



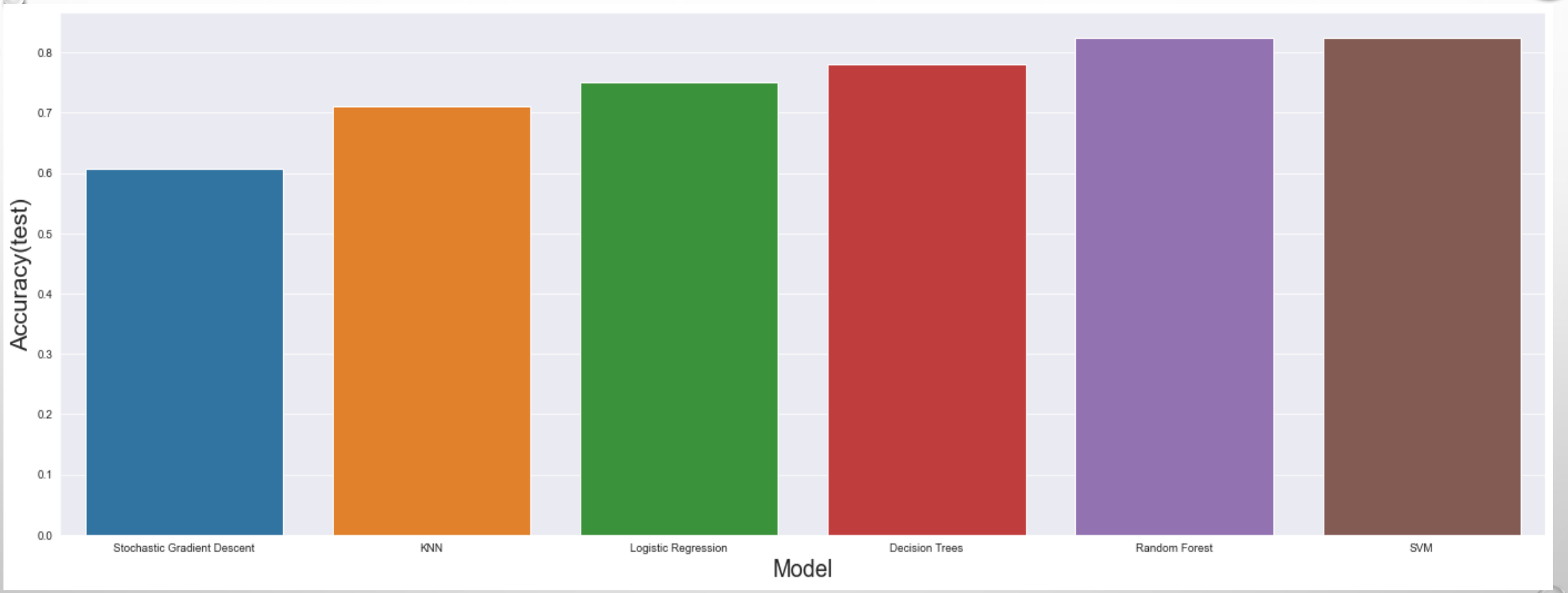
Quality	Bin Count
3	10
4	53
5	681
6	638
7	199

MACHINE LEARNING MODELS

The following are the different machine leaning models that we used to get the prediction

- A. LOGISTIC REGRESSION
- B. DECISION TREES
- C. RANDOM FORESTS
- D. SVM
- E. STOCHASTIC GRADIENT DESCENT
- F. KNN

VISUALIZING MODEL PERFORMANCE



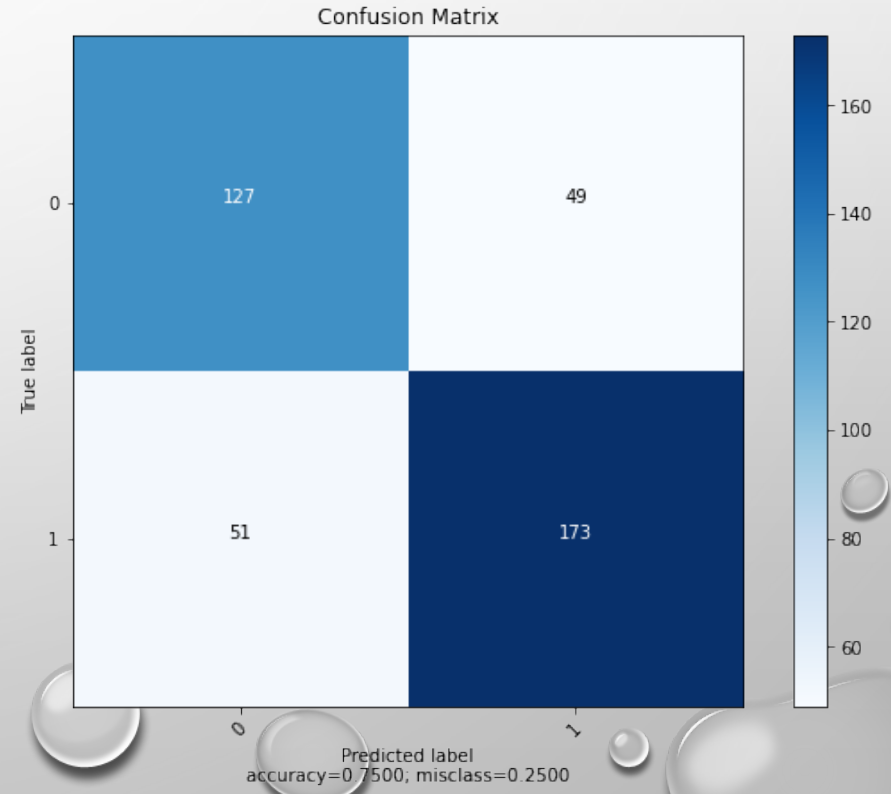
THE TOP THREE MACHINE LEARNING MODELS FOR THE CURRENT DATASET ARE

1. **SVM**
2. **RANDOM FOREST**
3. **DECISION TREES**

1. SUPPORT VECTOR MACHINES

	precision	recall	f1-core	support
0	0.8	0.8	0.8	176
1	0.84	0.85	0.84	224
accuracy			0.82	400
macro avg	0.82	0.82	0.82	400
weighted avg	0.82	0.82	0.82	400

Accuracy = 82.5 %



Parameters:

Kernel = 'poly'

Degree = 3

Gamma = 'auto'

2. RANDOM FOREST

	precision	recall	f1-core	support
0	0.78	0.84	0.81	176
1	0.87	0.82	0.84	224
accuracy			0.83	400
macro avg	0.83	0.83	0.83	400
weighted avg	0.83	0.83	0.83	400

Accuracy = 82.5 %

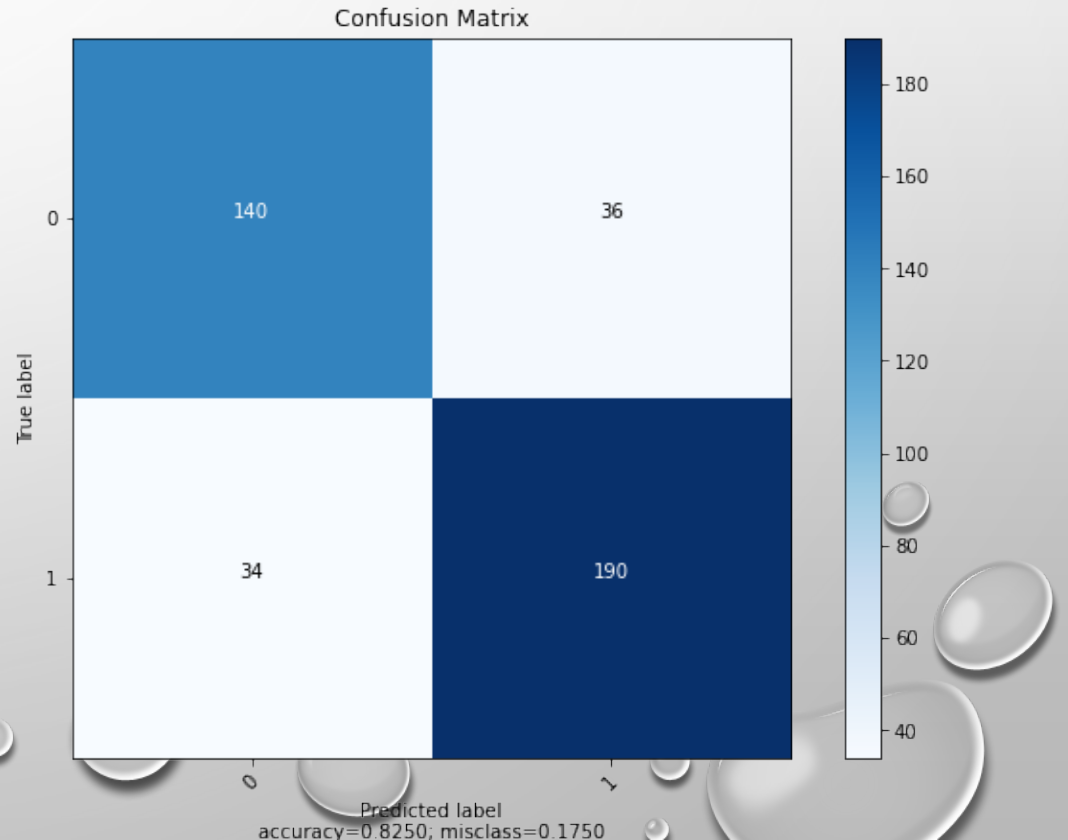
Parameters:

criterion = 'entropy'

n_estimators = 100

max_features = 'sqrt'

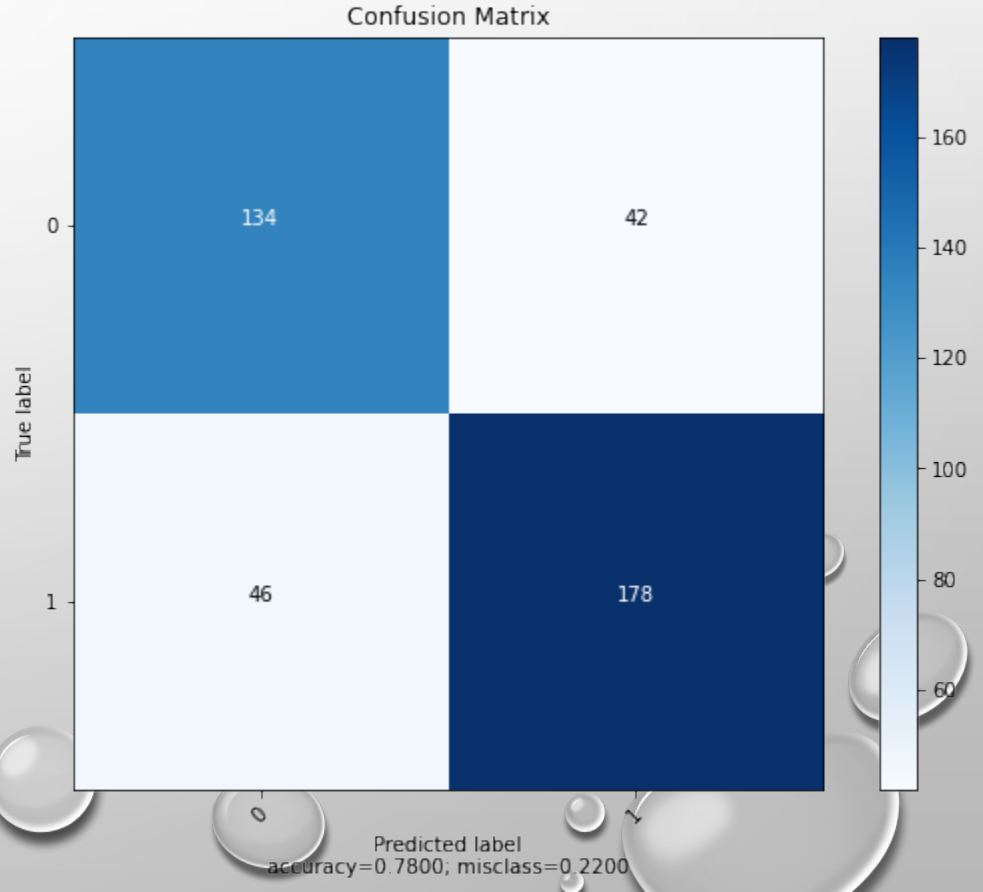
min_samples_leaf = 1



3. DECISION TREES

	precision	recall	f1-core	support
0	0.76	0.72	0.74	176
1	0.79	0.82	0.8	224
accuracy			0.78	400
macro avg	0.77	0.77	0.77	400
weighted avg	0.77	0.78	0.77	400

Accuracy = 78%



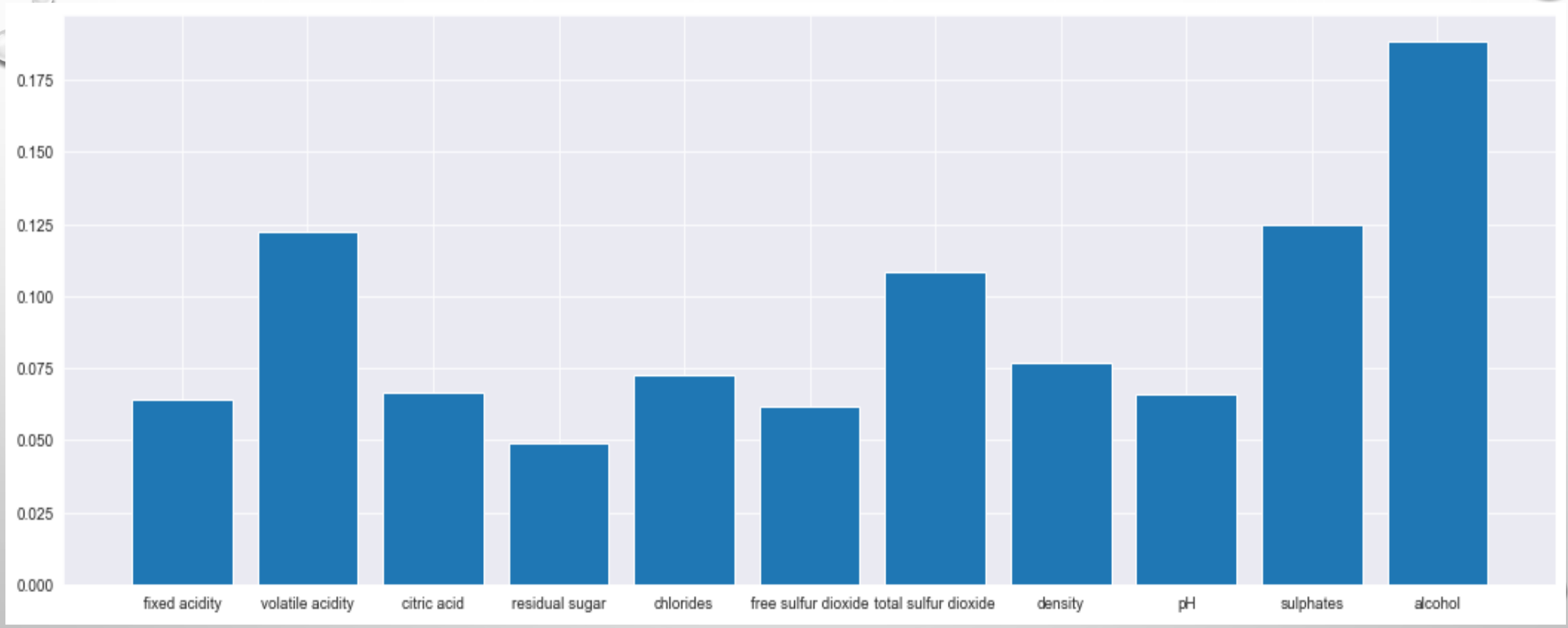
Parameters:

criterion = 'entropy'

max_depth = None,

min_samples_split = 3

VARIABLE IMPORTANCE PLOT



It indicates that alcohol is the most important variable

CONCLUSIONS

- The red wine data was **analyzed for quality**
- It was found that **quality of the wine** is mostly **correlated to alcohol**
- Data was split into training and test set. Machine learning models were trained in training set and tested for accuracy on the test set.
- **SVM model** does the best with an accuracy of **82.5%**
- Machine learning also conclude that **"alcohol"** is the most important variable that **determines the quality** of wine
- **Future Scope:** ML models accuracy is not high. The **low prevalence of quality levels 3, 4 and 8** and the large distribution overlapping area stratified by quality is a reason.
- We **do not have information** about **the composition of grape varieties in each wine, the mix of experts that evaluated wine quality, or the production year.**
- **Lack of information** about how the dataset was created may **impact the prediction of quality** using the physicochemical properties as predictors.

THANK YOU!