Red Wine Quality Prediction

In this blog-post, I will go through the whole process of creating a machine learning model on the Red Wine Quality Prediction. The dataset is related to red and white variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.)

This dataset can be viewed as classification task. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

Importing the Libraries

# linear algebra

import numpy as np

# data processing

import pandas as pd

# data visualization

import seaborn as sns

%matplotlib inline

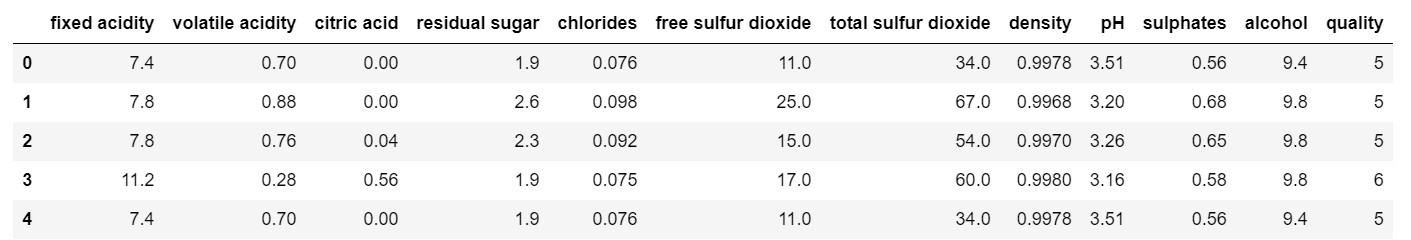
from matplotlib import pyplot as plt

from matplotlib import style

# Getting the Data

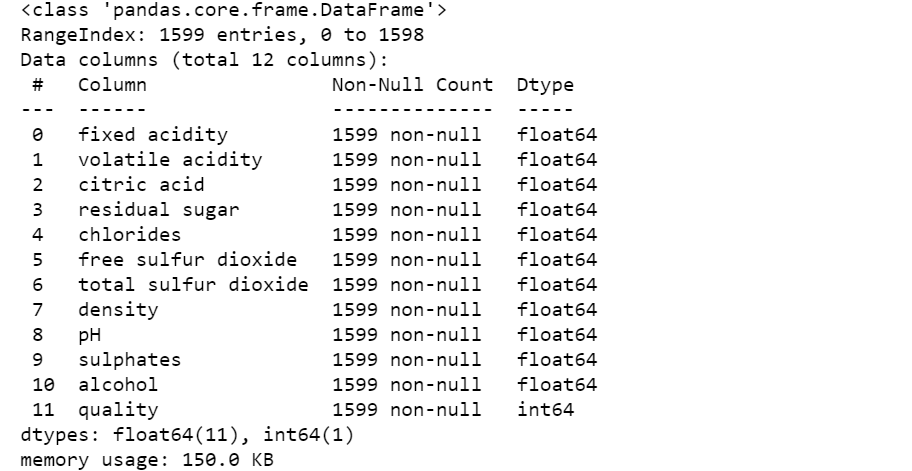
data = pd.read\_csv('https://raw.githubusercontent.com/dsrscientist/DSData/master/winequality-red.csv')

data.head()

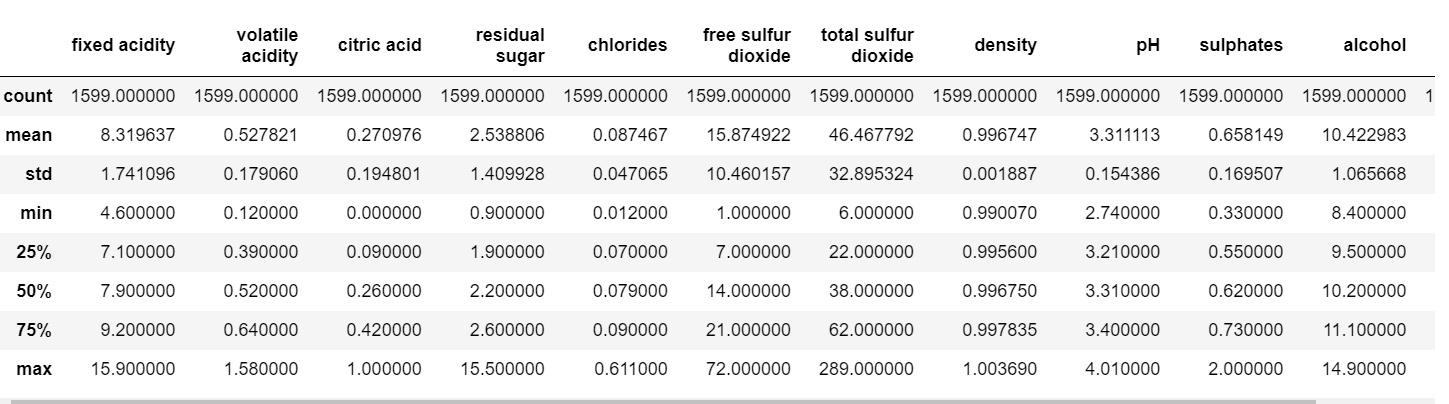


# Data Exploration/Analysis

Data.info()

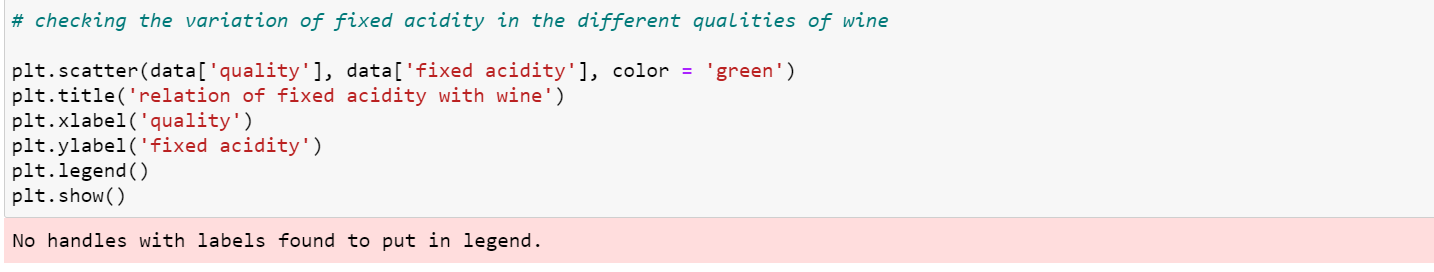


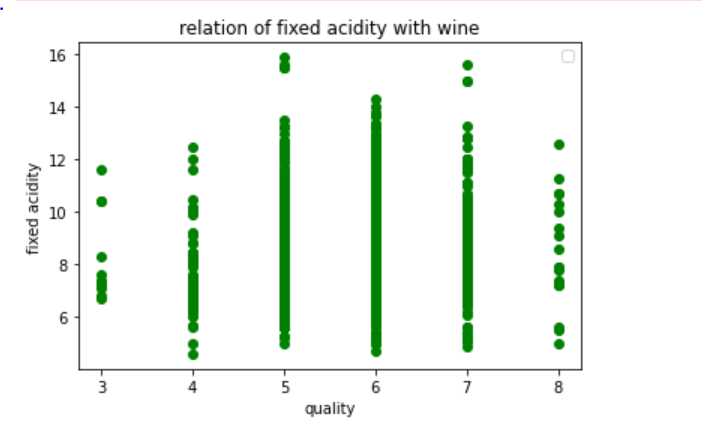
Data.describe()

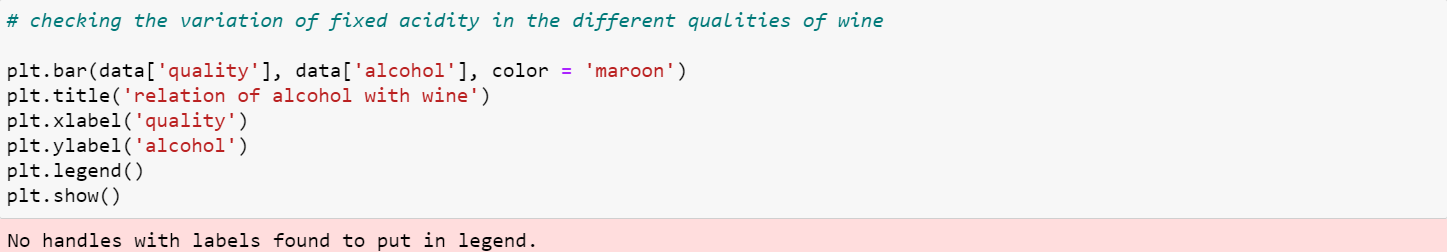


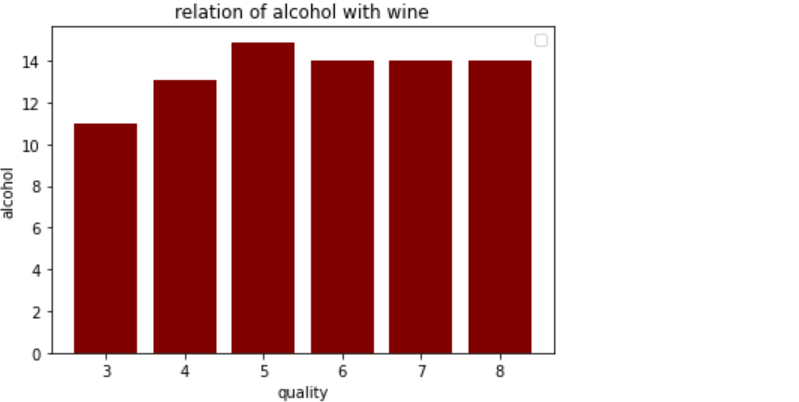
DATA VISUALIZATION

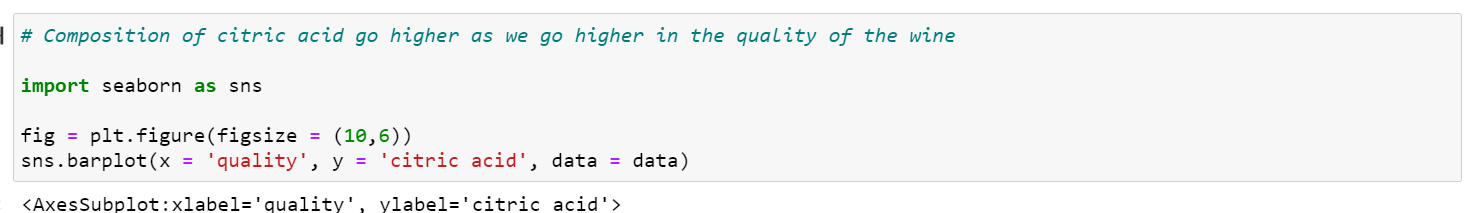
Data[‘Quality].value\_counts()

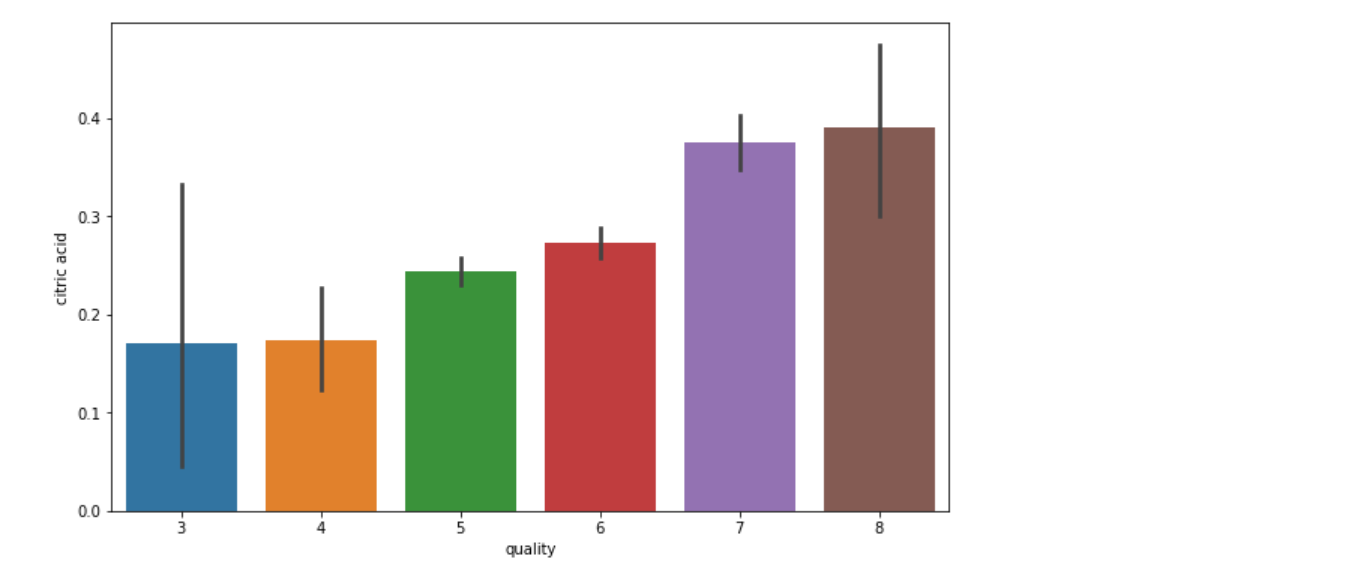


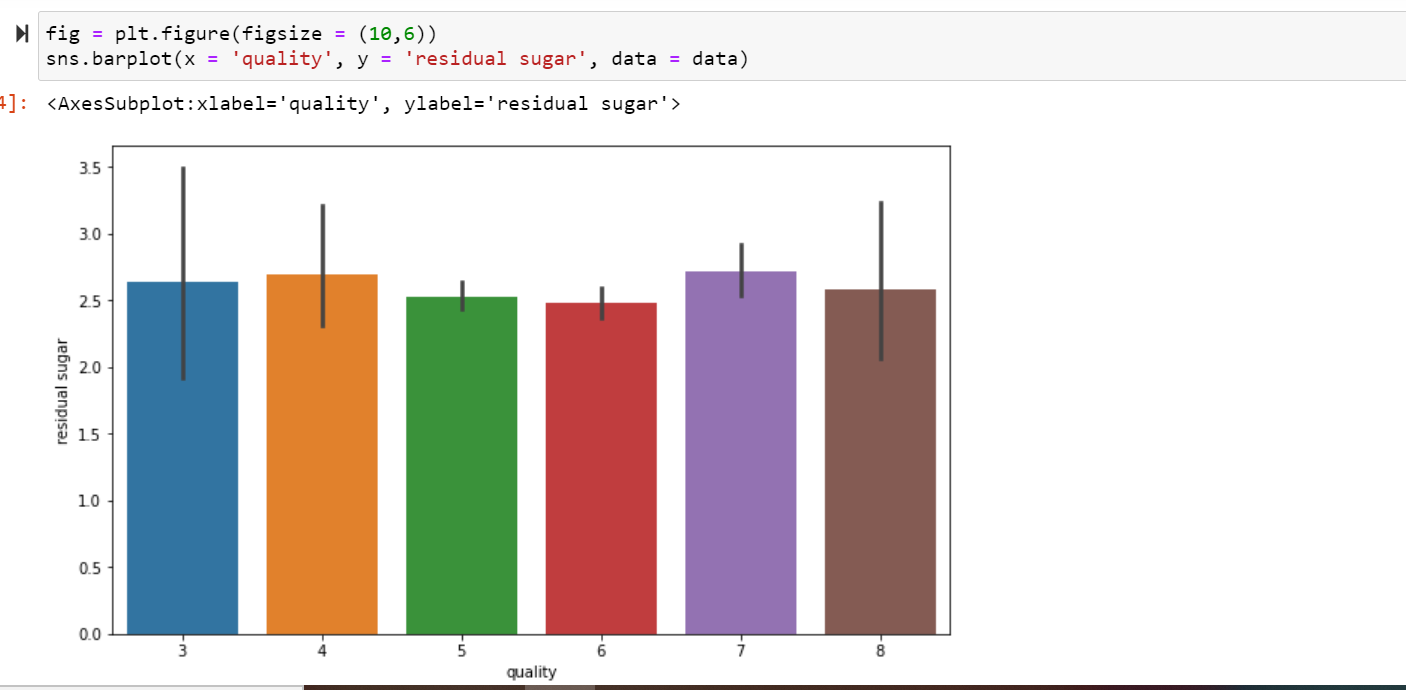


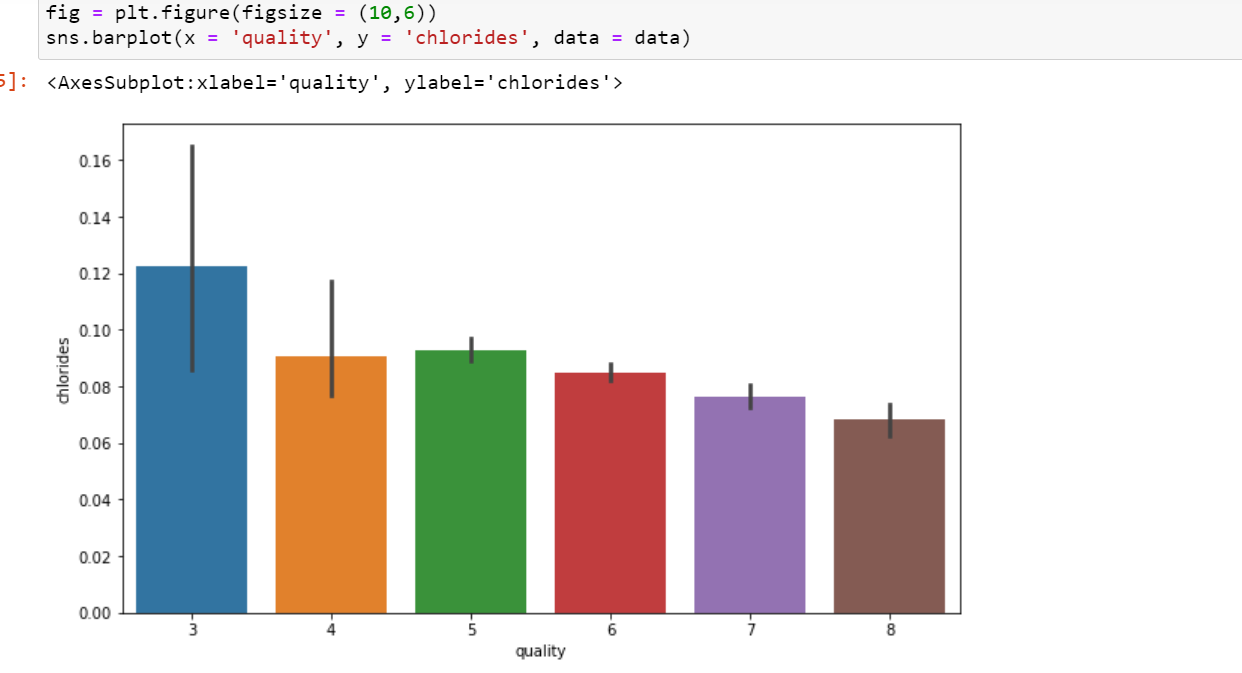


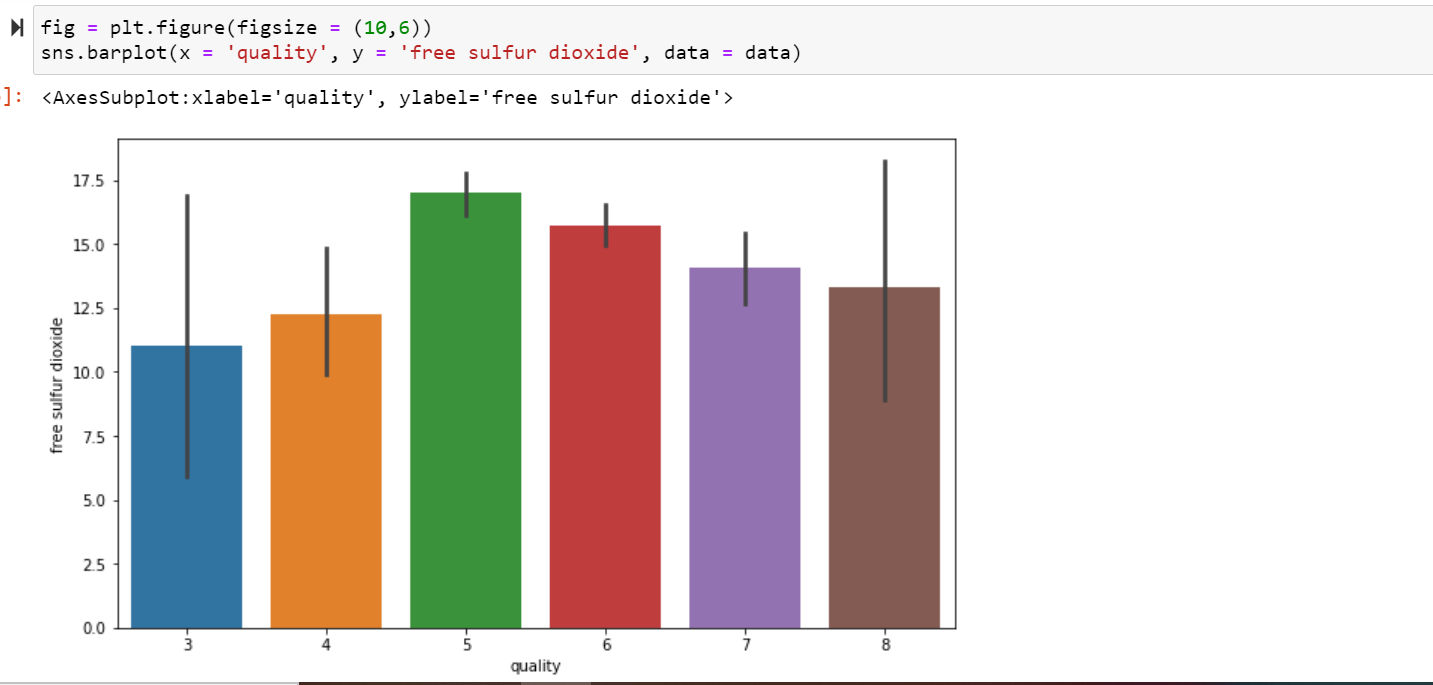


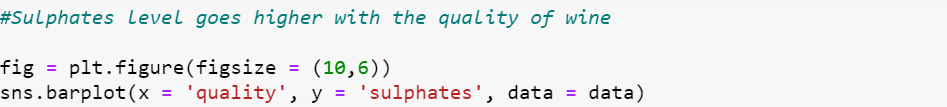


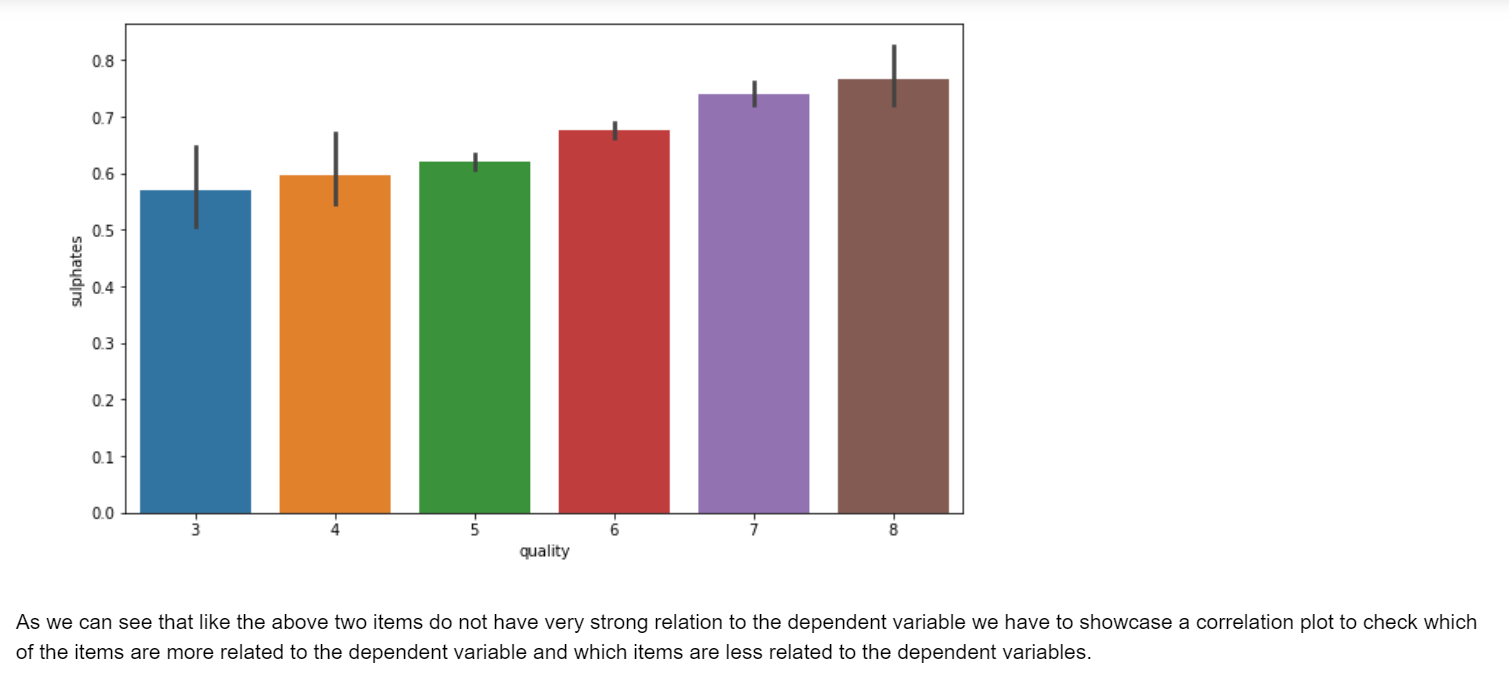


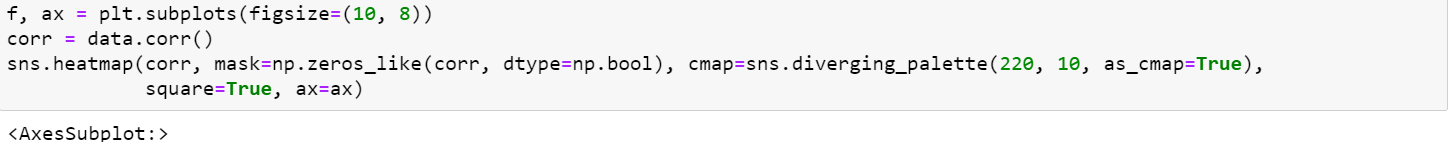


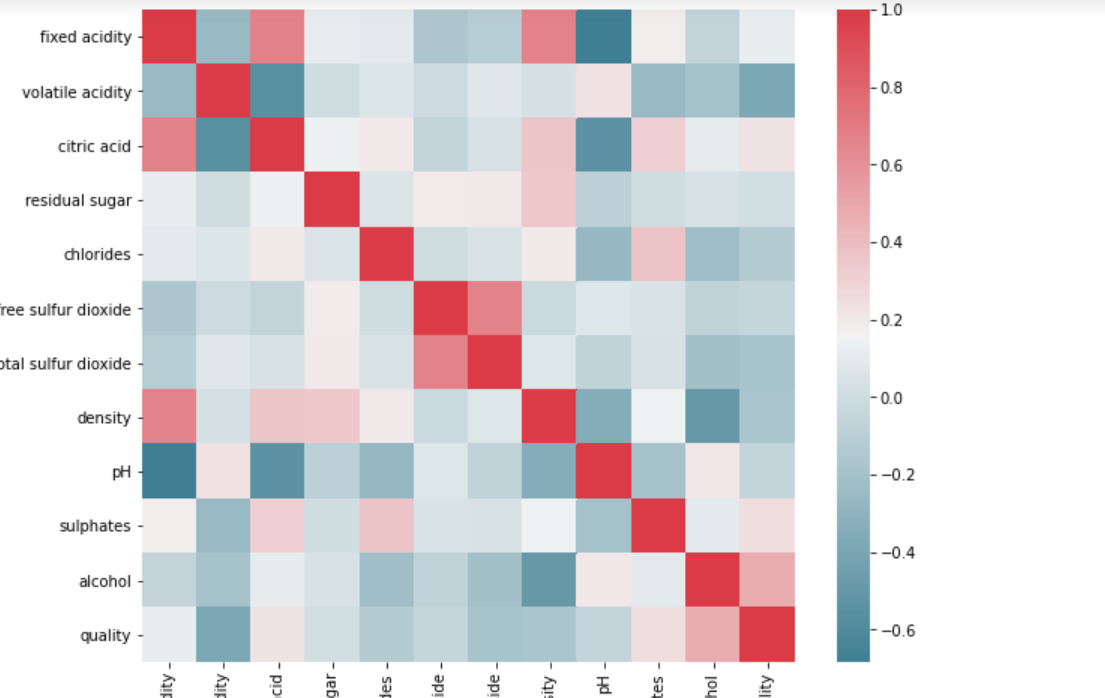


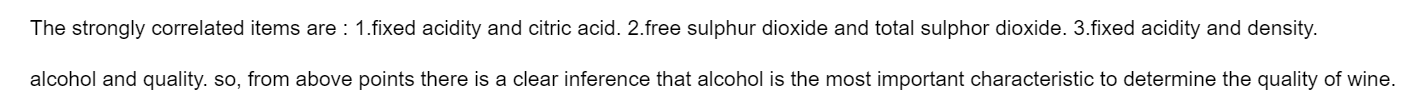


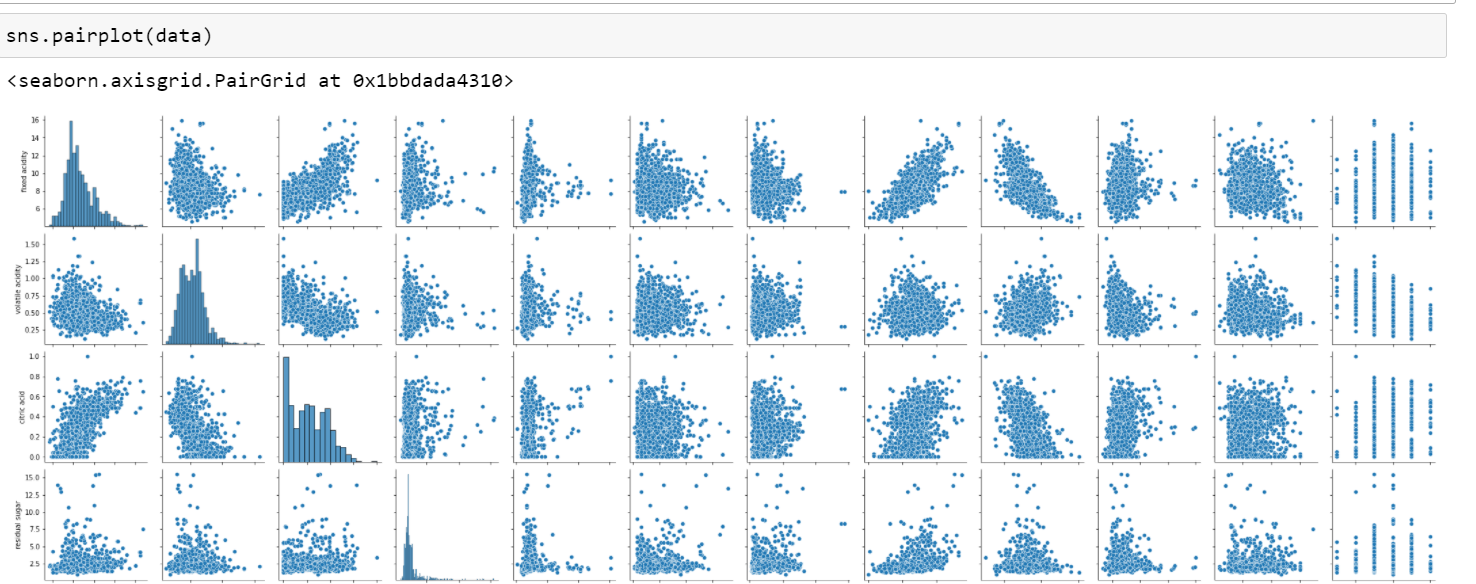


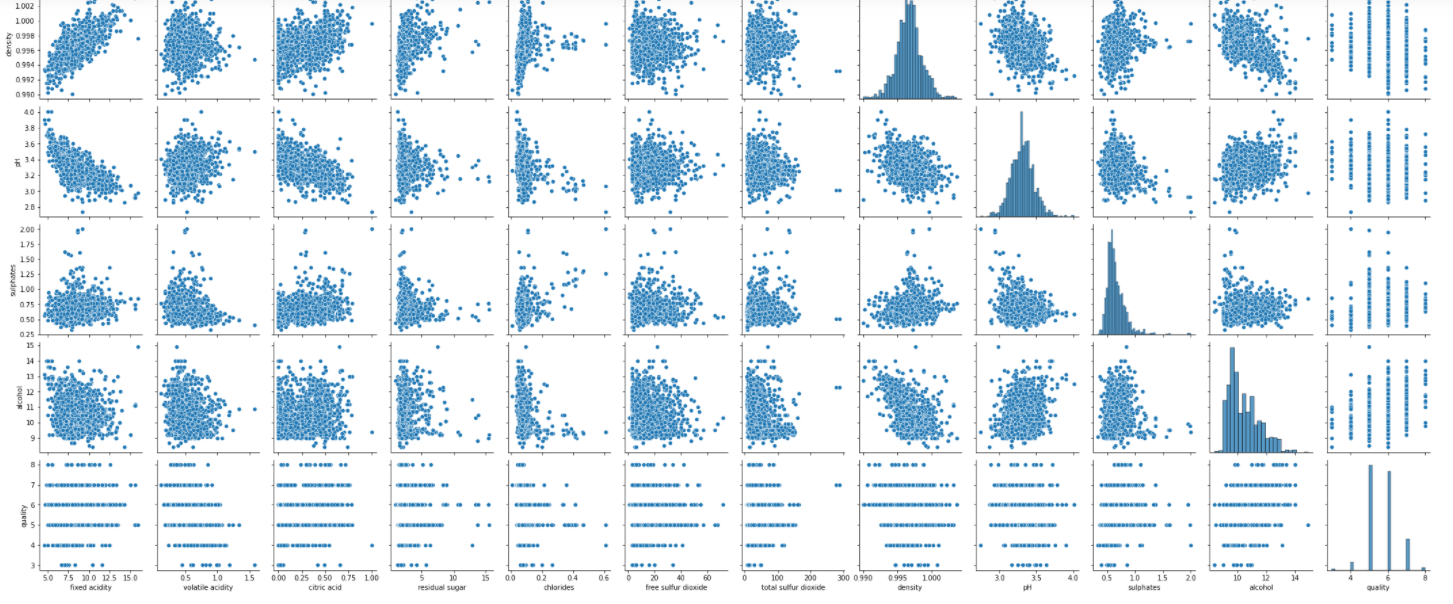


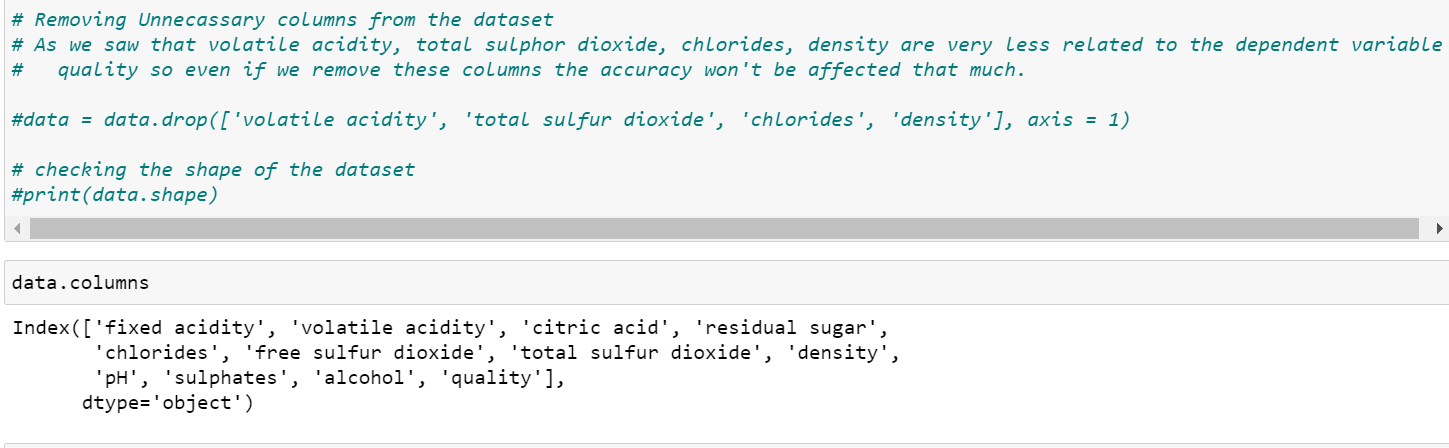


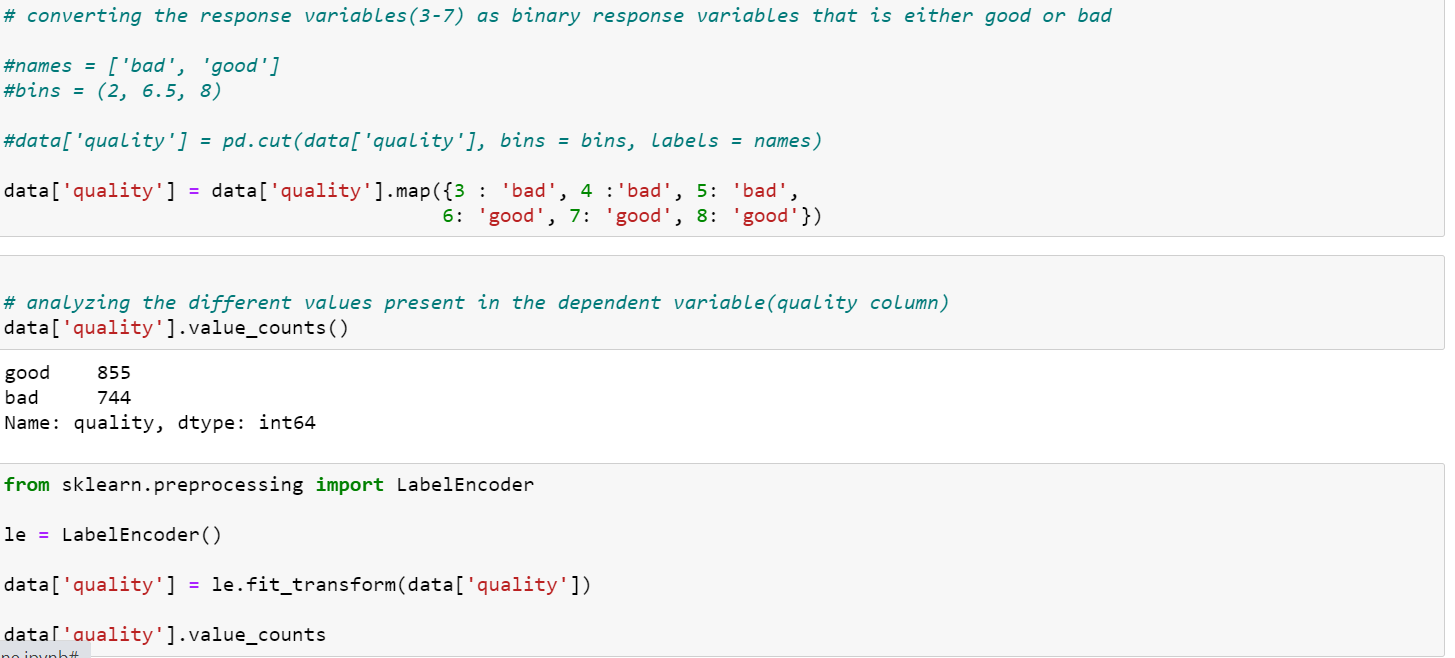


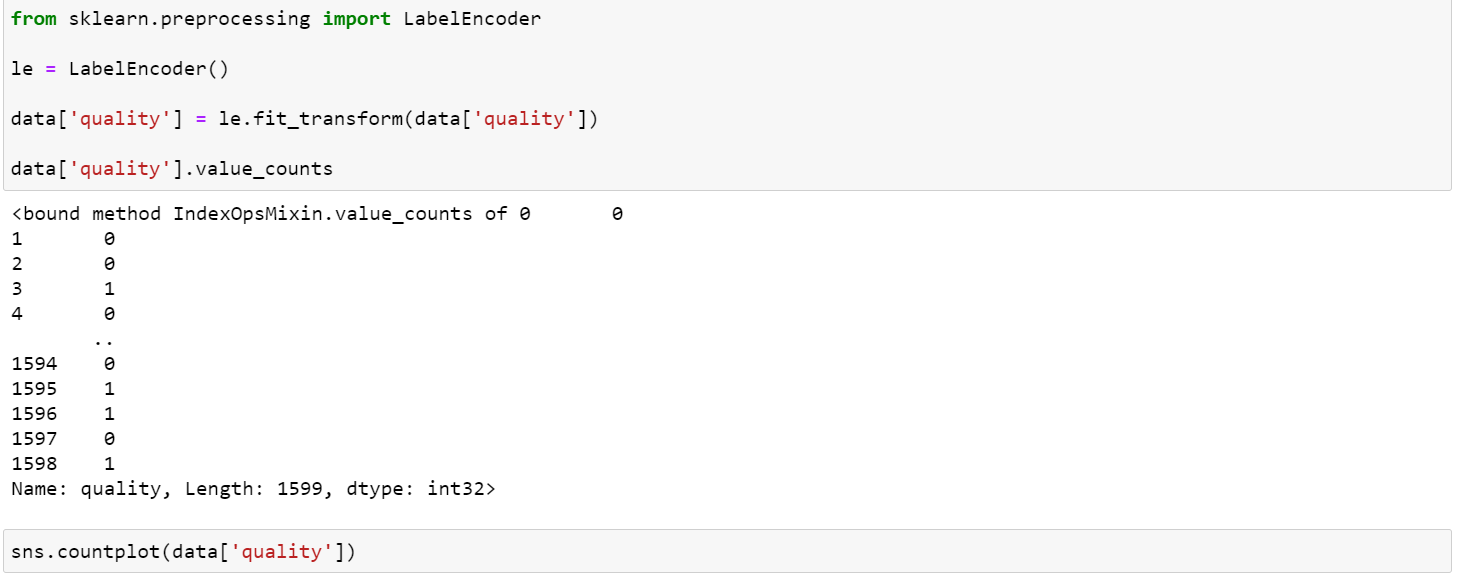


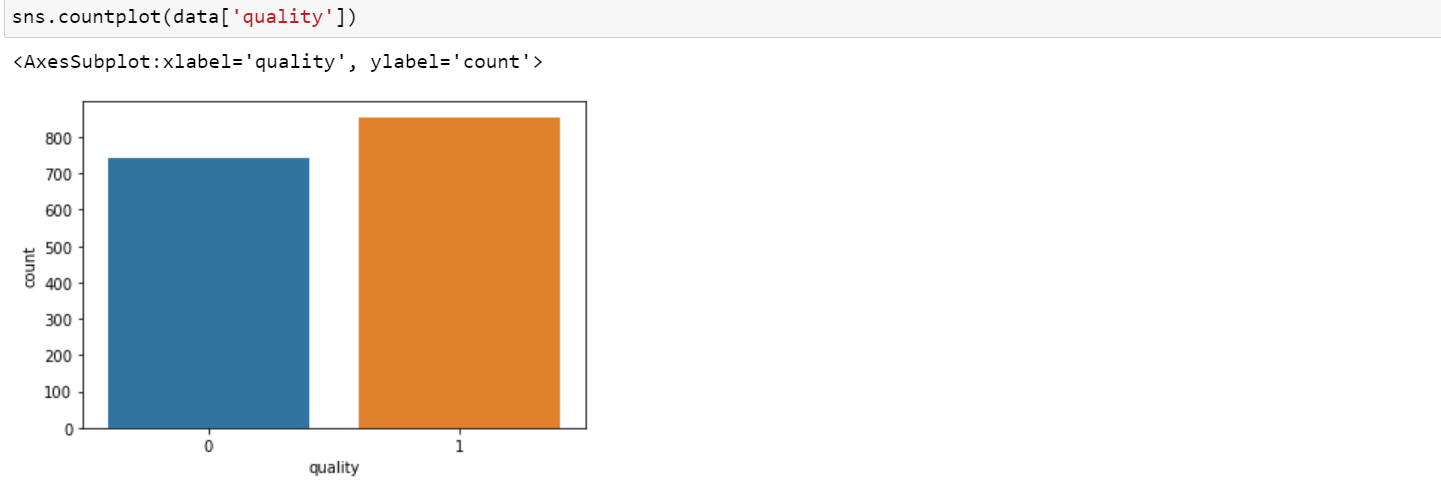








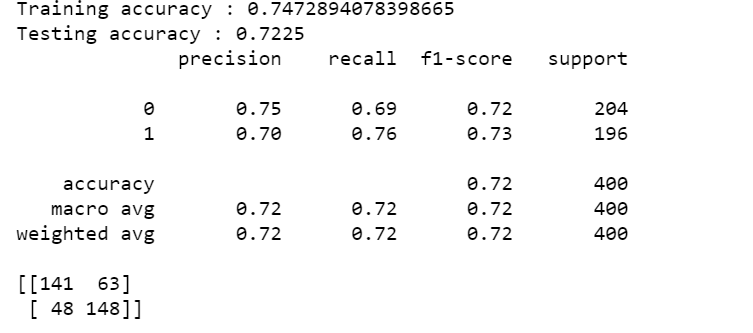




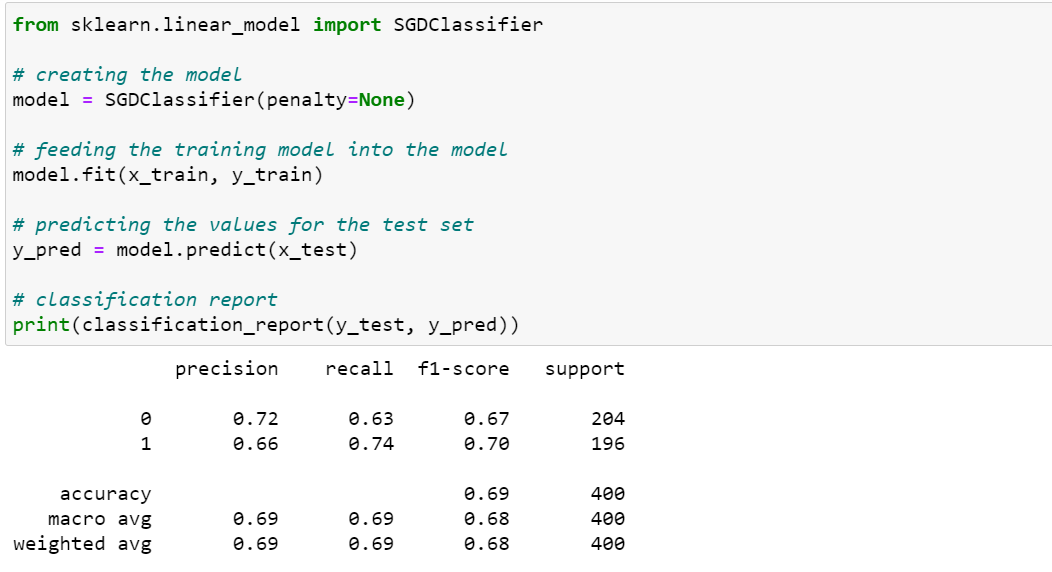


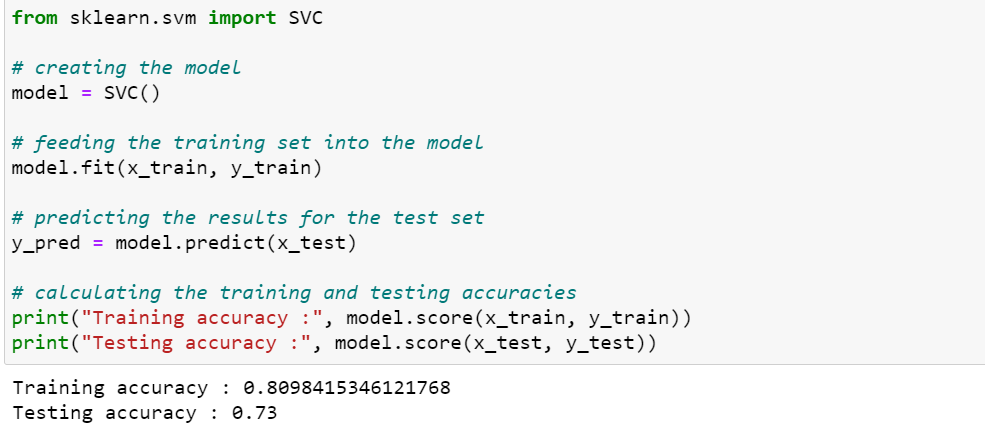


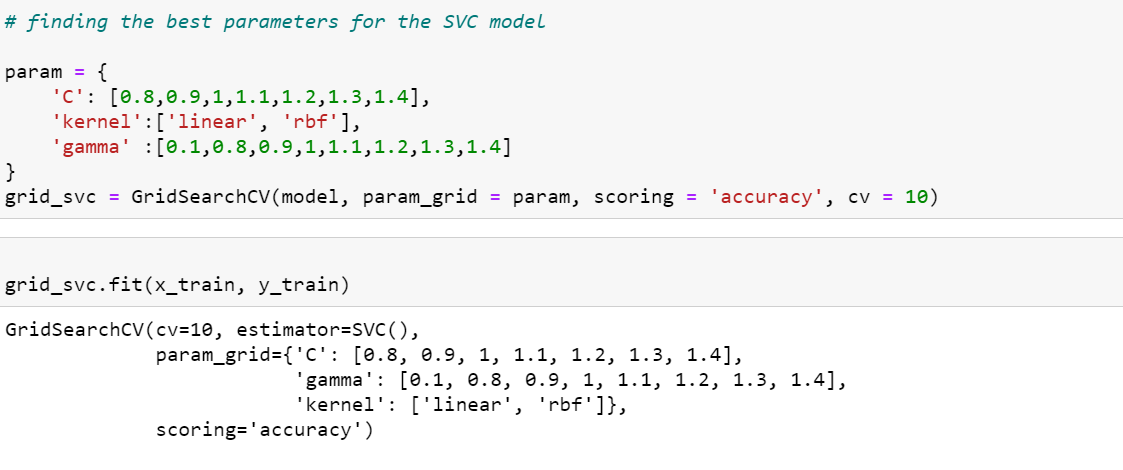




**Stochastic Gradient Descent (SGD):**

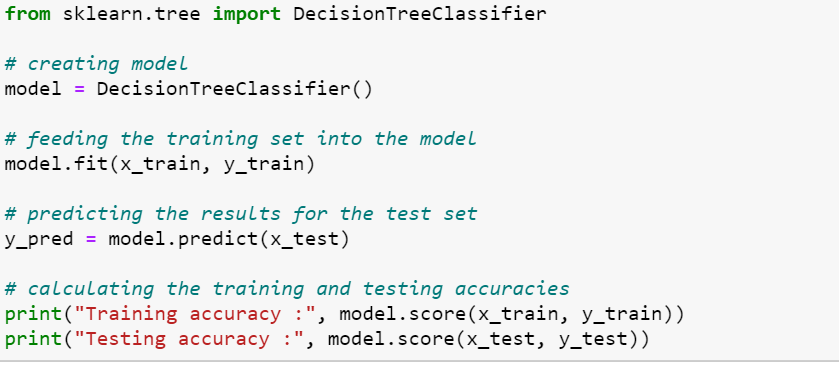


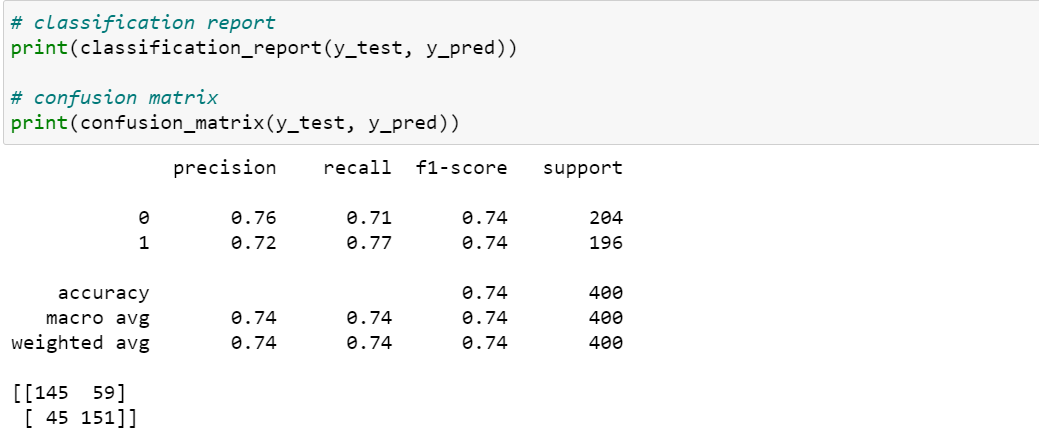




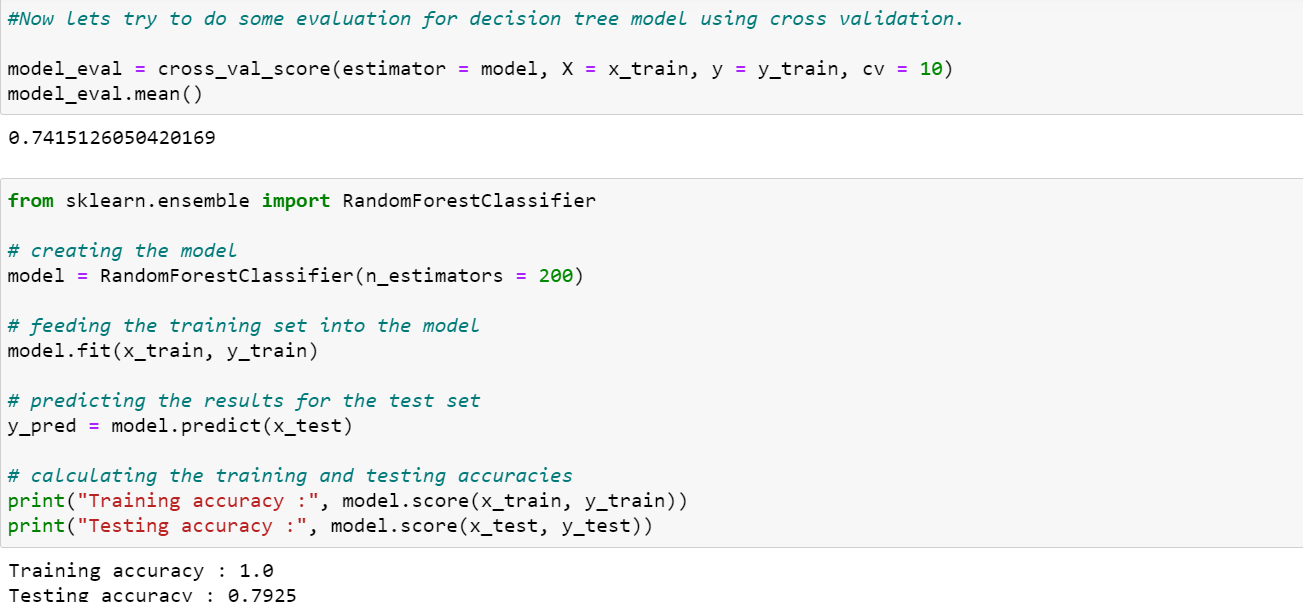


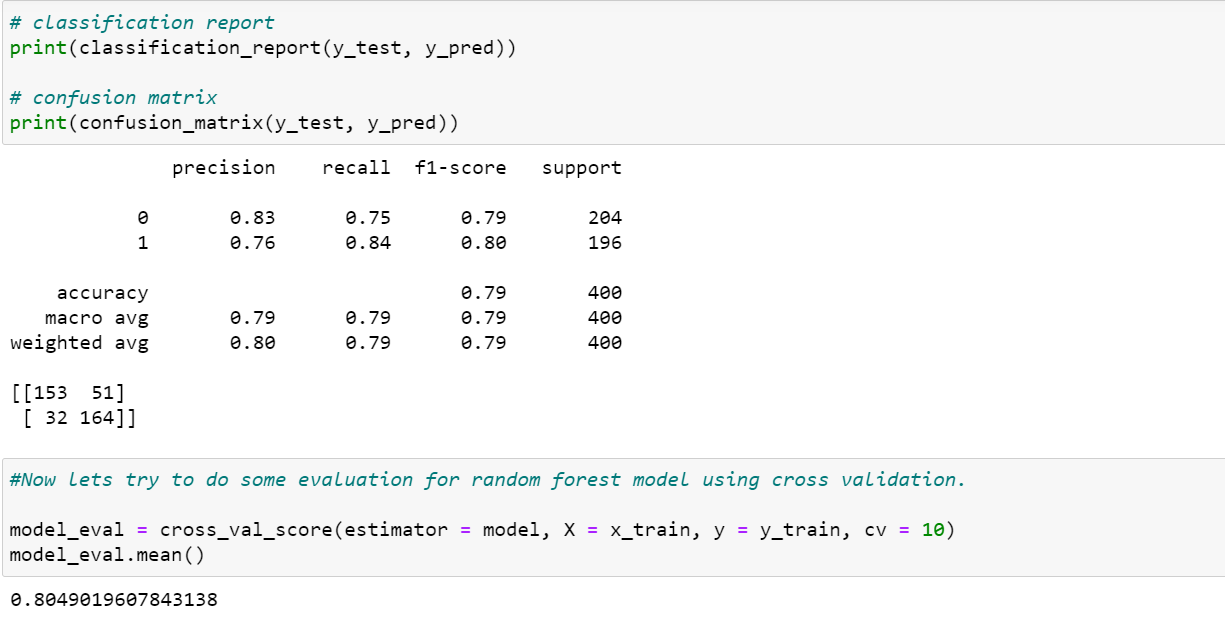
**Decision Tree**





**Random Forest:**





CONCLUSION

Based on the bar plots plotted we come to a conclusion that not all input features are essential and affect the data, for example from the bar plot against quality and residual sugar we see that as the quality increases residual sugar is moderate and does not have change drastically. So this feature is not so essential as compared to others like alcohol and citric acid, so we can drop this feature while feature selection. For classifying the wine quality, we have implemented

1) Logistic Regression

2) Stochastic gradient descent

3) Support Vector Classifier

4) Random Forest

We were able to achieve maximum accuracy using random forest of 88%. Stochastic gradient descent giving an accuracy of 81% .SVC has an accuracy of 85% and logistic regression of 86%.