Zili Bu G36963059

Ziwei Li G40721135

Midway Project Report

1. Introduction
   1. Background Introduction

Wine is one of the most ancient drinks of humankind and the earliest evidence of wine was around 6000 B.C in Georgia[[1]](#footnote-1). Winemaking involves a lot of chemistry and there is a complex mixture of different chemical components in wine. Typically, there are natural molecules present in wine like acids, sugars, sulfur dioxides etc. and it is important to balance these chemical components to produce good quality wine[[2]](#footnote-2).

* 1. Data Set Introduction

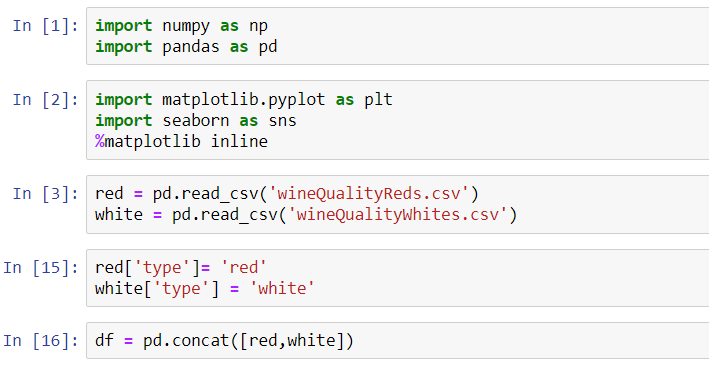
This data set is related to red and white variants of the Portuguese "Vinho Verde" wine, and it contains only physicochemical (inputs) and sensory (the output) variables are available due to privacy and logistic issues (Wine Quality, Kaggle)[[3]](#footnote-3). For example, there is no data about grape types, wine brand, wine selling price, etc. (Wine Quality, Kaggle) Here are the introductions to the key variables[[4]](#footnote-4):

* Fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily). These acids like tartaric acids (“maintain the chemical stability of wine and its color”[[5]](#footnote-5)), malic acids (naturally exists in grapes and is responsible for the tart taste of wine. The process of the malolactic fermentation will reduce the tartness by transforming malic acids to lactic acids)[[6]](#footnote-6).
* Volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste. Excessive amount of acetic acid is a wine fault, and the process of “acetification” (exposing to oxygen) is the key for wine degradation into vinegar[[7]](#footnote-7).
* Citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines. Meanwhile, it also helps to remove excess iron and copper from the wine[[8]](#footnote-8).
* Residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet. The level of residual sugars can classify the wine into different “sweetness”[[9]](#footnote-9).

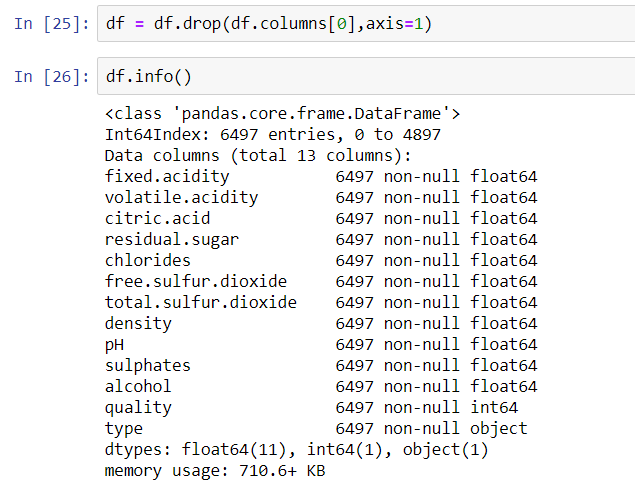
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Dry | Medium Dry | Medium | Sweet |
| Sweetness | up to 4 g/l | up to 12 g/l | up to 45 g/l | more than 45 g/l |

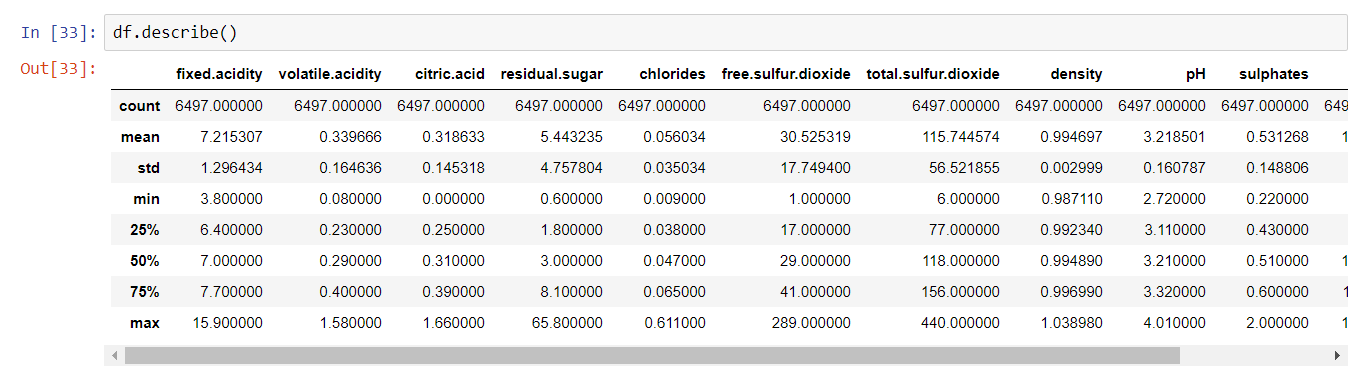
* Chlorides: the amount of salt in the wine
* Free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
* Total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine.
* Density: the density of water is close to that of water depending on the percent alcohol and sugar content.
* pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale.
* Sulphates: a wine additive which can contribute to sulfur dioxide gas (SO2) levels, which acts as an antimicrobial and antioxidant.
* Alcohol: the percent alcohol content of the wine
* Quality: the output variable, score from 0 to 10.
  1. Problems of Interest
     1. We would like to see the relationship between the quality and the volume of different chemical components the dataset provided. For example, we can find the relationship between quality of wine and acidity level, residual sugars and sulfur dioxides.
     2. Some variables might not be relevant to describe the quality of wine so that we can use PCA or feature selection methods to sift the best predictors, and produce a simpler model.
     3. One of the key interests is the classification of the wine for this dataset. We plan to create a new variable “review” and distribute the score into three levels: bad (0-3), medium (4-7), good (8-10). After that, we can use different algorithms to perform classification and evaluate the outcomes.
  2. Concerns about the analysis
     1. The dataset only contains the chemical contents but no information about “brand type”, “grape type”, “selling price” etc., it might be hard for the audiences to understand (need more information about these chemicals).
     2. Although the number of predictors is relatively not low, they are basically the same type. This could limit our approaches to perform different analyses to predict quality.
     3. The relationship between the quality of wine and the volume of acids might not be linear, we might need to think more when coming to interpretation.

1. Data Set Analysis
   1. Import Our Data

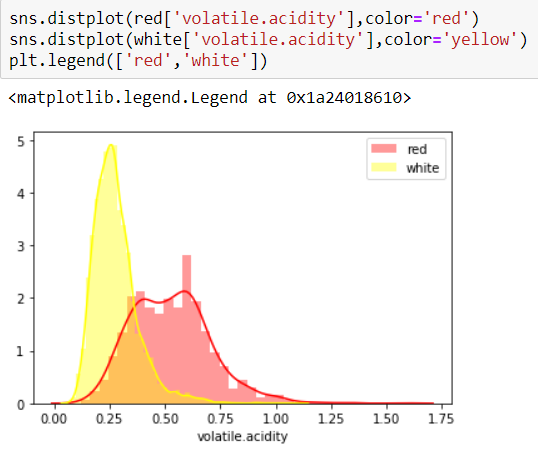
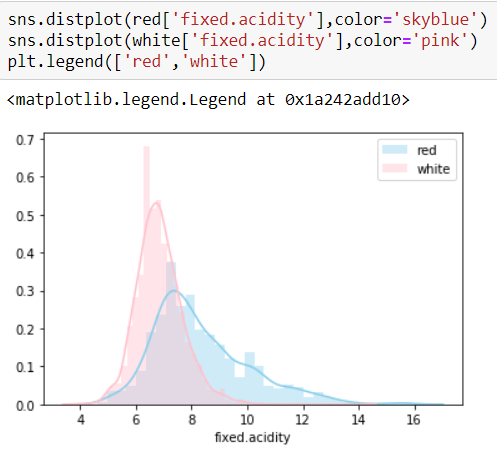


* 1. Exploratory Data Analysis
     1. Overall review of each variables

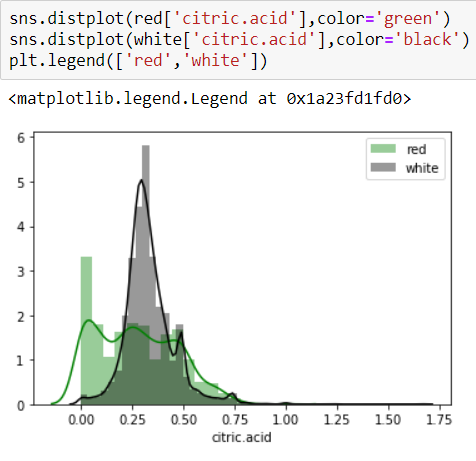
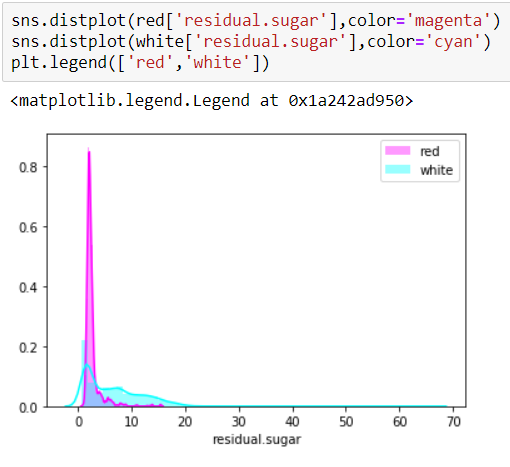




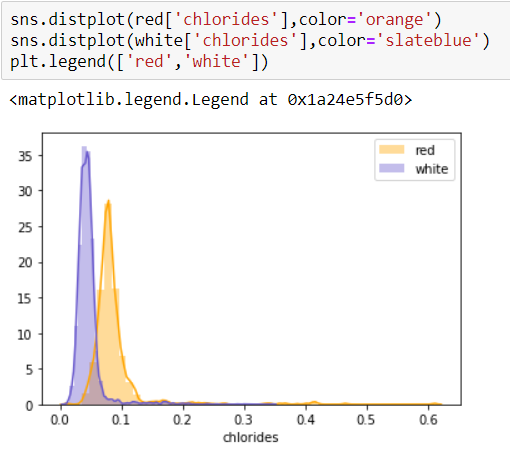
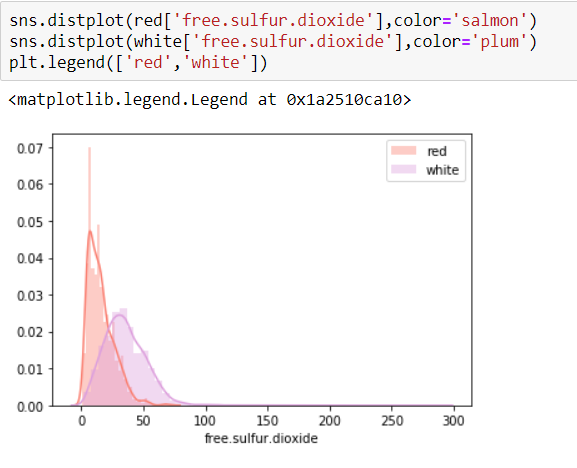
* + 1. Explore the distribution of two types of wine in terms of each variables
       - Fixed Acid: average fixed acid of white wine is less than the average of red.
       - Volatile Acid: Average volatile acid of white wine is also less, and it is more concentrated around mean than the red wine.



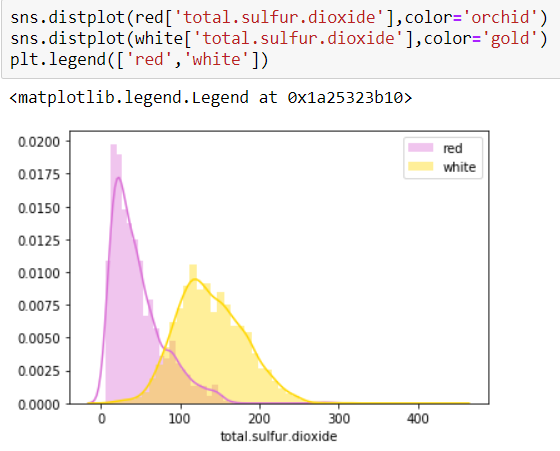
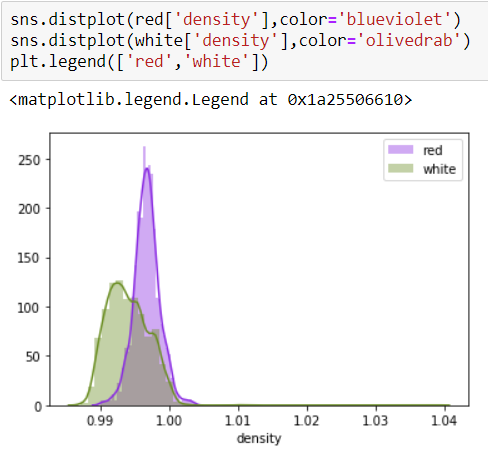
* + - * Citric Acid: White wine has a more concentrated distribution of citric acid than red wine,
      * Residual Sugar: Residual sugar for red wine is highly crowded at 2-3g, but white wine is more spread out. However, both of them have low level of residual sugars.



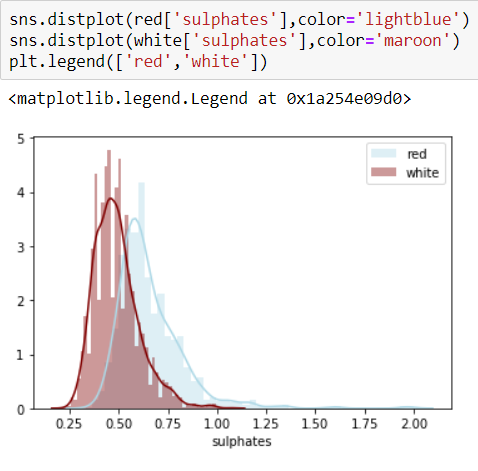
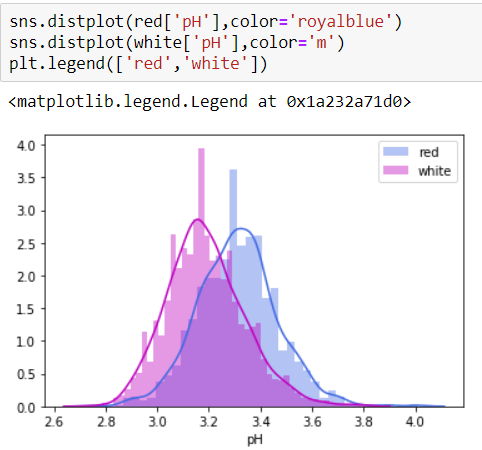
* + - * Chlorides: Red wine has lower average value than white wine, but they are very similar and concentrated in the same way.
      * Free Sulfur Dioxide: the value of red wine is lower on average and more concentrated, but white wine is more spread out. However, most of them are lower than 100.



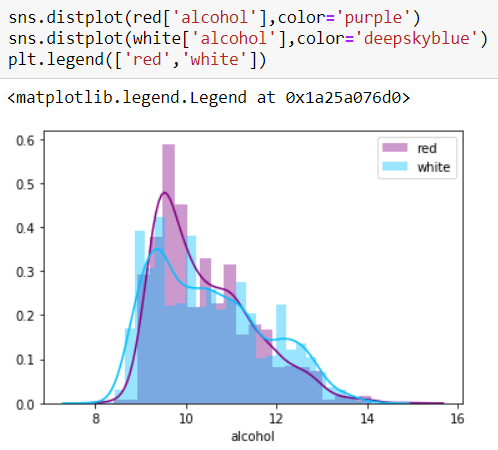
* + - * Total Sulfur Dioxide: the value of white wine and red wine are clearly different in this variable. Red wine has much lower total sulfur dioxide than white wine, and it is more concentrated at lower level as white wine’s value is more or less spreading out.
      * Density: Red wine density is more concentrated at mean than white wine.



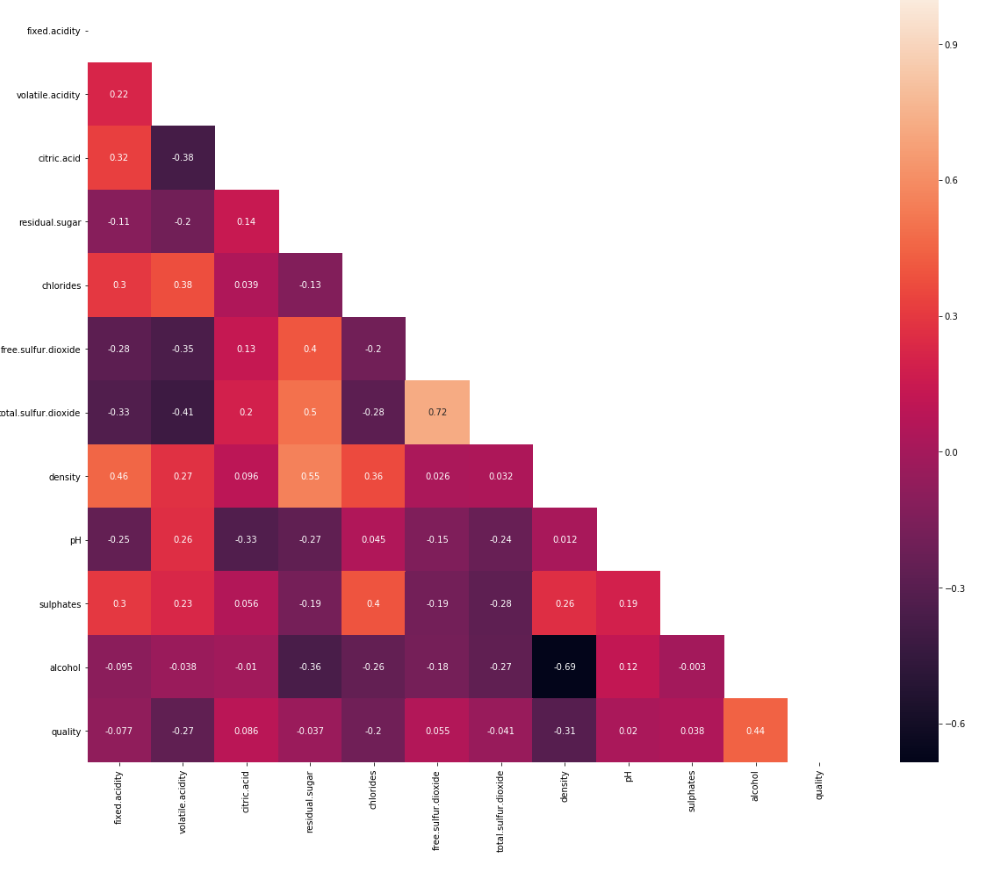
* + - * PH: Ph value is very similar of both wine in terms of the shape, and there is not much difference in the mean or variance from the graph.
      * Sulphates: This variable is still very similar between two types of wine.



* + - * Alcohol: The shape of distribution is similar between two types of wines, however, red wine has relatively higher alcohol value than white.



* + - * Summary：There are some obvious differences among each variables of different types of wine. We assume that the classification of different type of wine would be feasible and the outcome should be relatively well-performed.
    1. Correlation between each variable



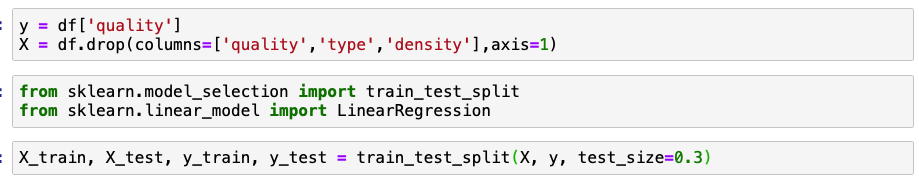
1. Regression

In this part, we are going to use linear regression to explore the relationship between quality (dependent variable) and the rest of the variables (independent variable).

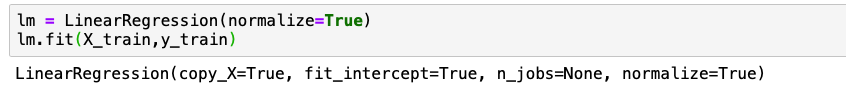
Before doing that, we need to manually choose what variables we might include. As we can see from the columns, 'total.sulfur.dioxide' might include 'free.sulfur.dioxide', if we choose both variables into our analysis, we might face a multicollinearity issue. Therefore, in this part, we are going to use 'total.sulfur.dioxide' other than 'free.sulfur.dioxide'.

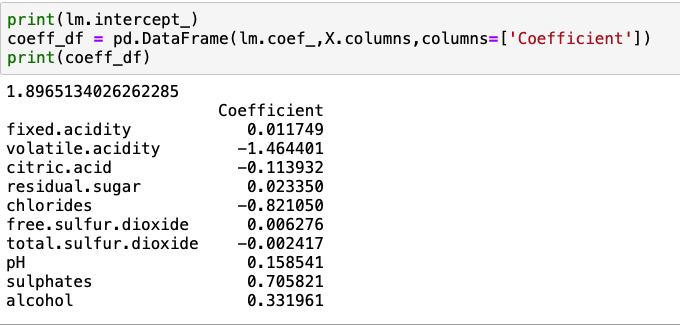
* 1. Regression Preprocessing

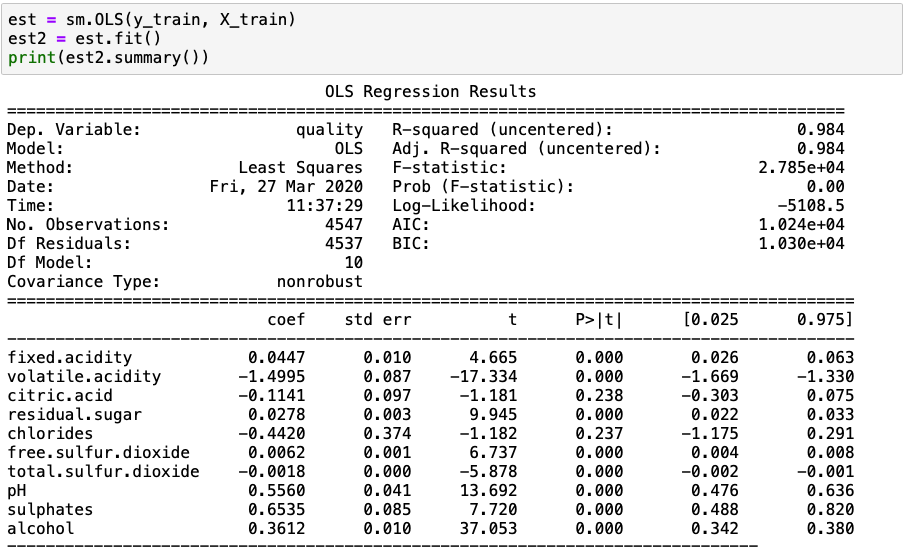
We are going to designate the dependent variables and the independent variable. And split the data into training and test set using test\_size = 0.3.



* 1. Linear Regression

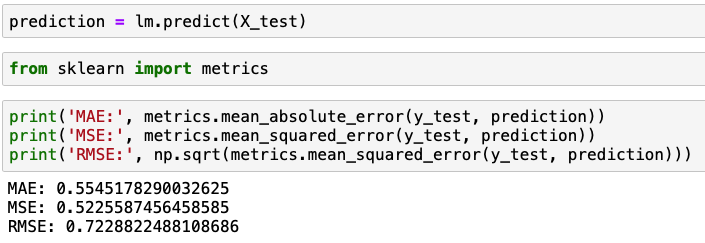
First, we will fit the data using the training set and check the intercept, coefficients and each variable's p-value.





We might notice that there might be some variable's p-value is bigger than 0.05, which means they are not significant. But right now, we will still include these variables to make some predictions.

After getting the intercept and coefficients, we can use this equation to make the prediction on the test set and calculate the test error.

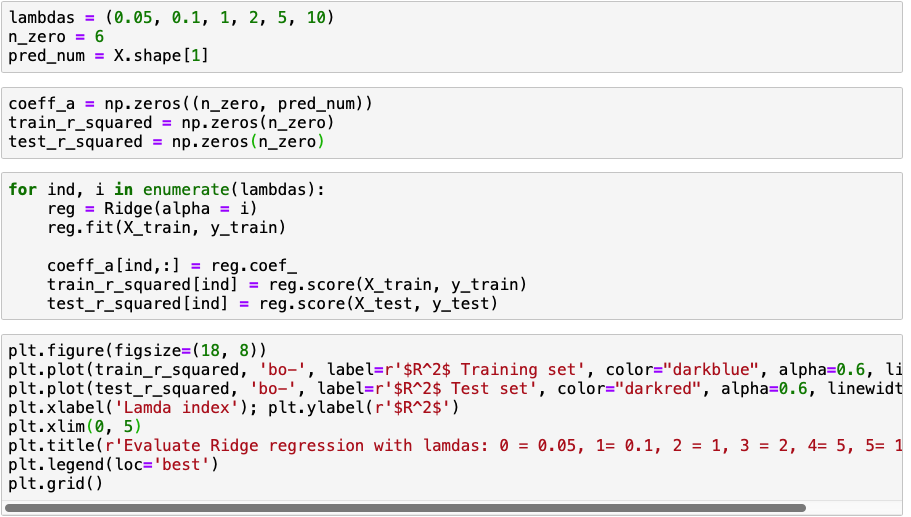


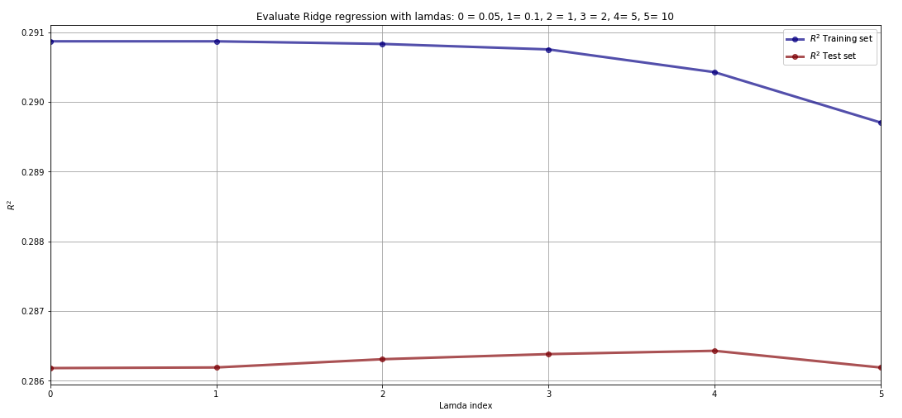
We can see the RMSE of linear regression is approximately 0.7229. It is slightly high, but since we do not have other models yet, it is hard to decide whether this model is good or not.

In order to modify the linear model, we will use lasso and ridge regression to check if the model's performance can improve.

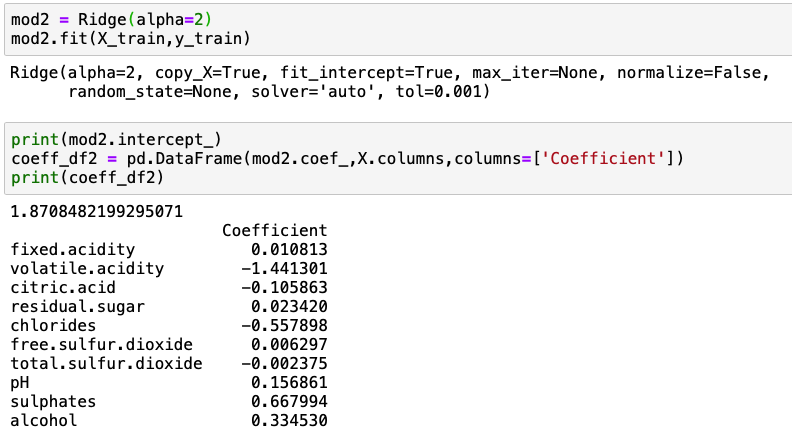
* 1. Ridge Regression

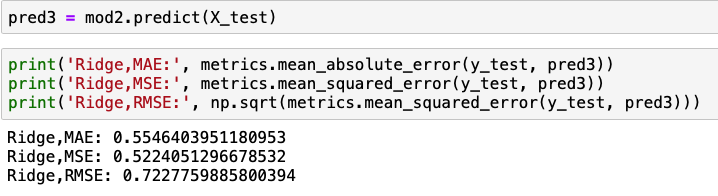
According to the concept of ridge regression, first, we need to choose the best lambda we want to use. Here, we are going to include 6 values and using the plot to see which lambda can give us the highest R-squared.





We can see from the plot that when lambda is 1, we will get the highest R-squared. Therefore, we will use the lambda=1 to fit the ridge regression model.



And then we can use the results we obtain to make the prediction on the test set can calculate the test error.

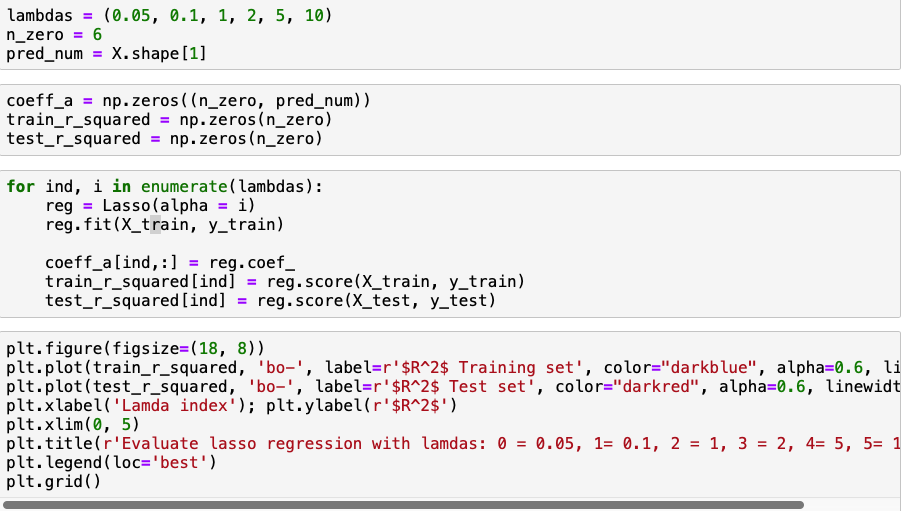
According to the RMSE, we can see that ridge regression performs slightly better than linear regression, but the results are almost the same.

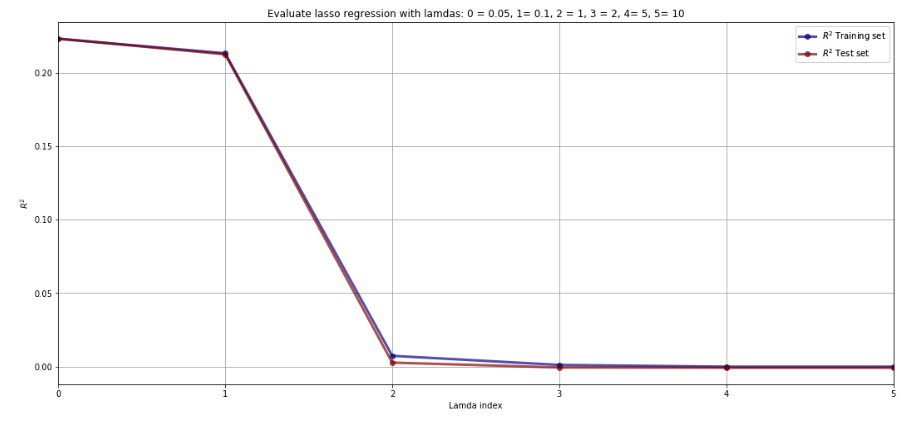
Finally, we will use the lasso regression to fit the model.

* 1. LASSO Regression

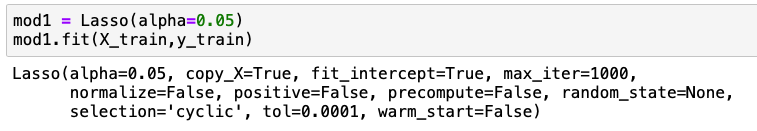
Same as ridge regression, we need to decide what is the best lambda that we are going to use to fit our model.

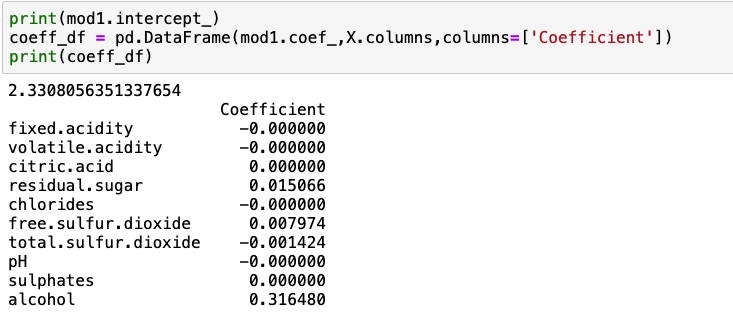
The steps are almost the same, we will still use the 6 values we set and choose the biggest R-squared from the plot.





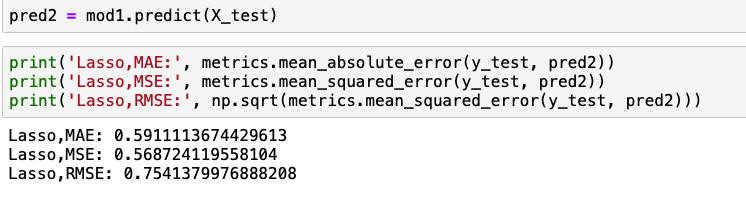
When lambda is the smallest, we can get the highest R-squared. That is lambda = 0.05. Therefore, we will use the value to fit the model.





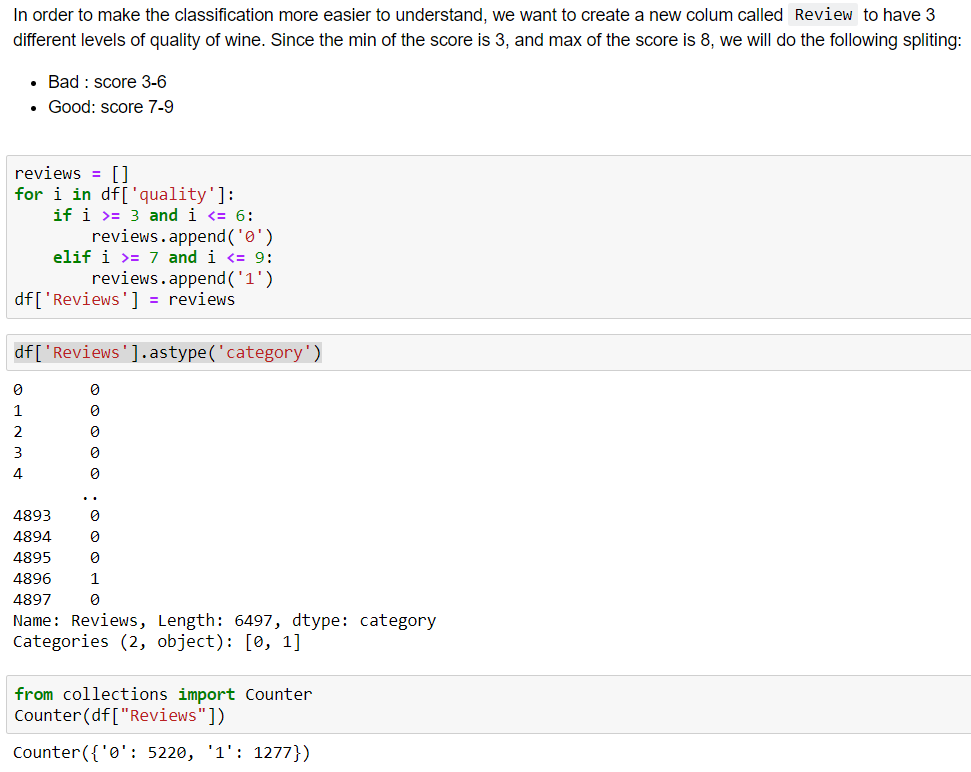
The difference between ridge and lasso regression is that lasso can push some variable's coefficient exactly to 0, while the ridge cannot, as we can see from the results we obtained.

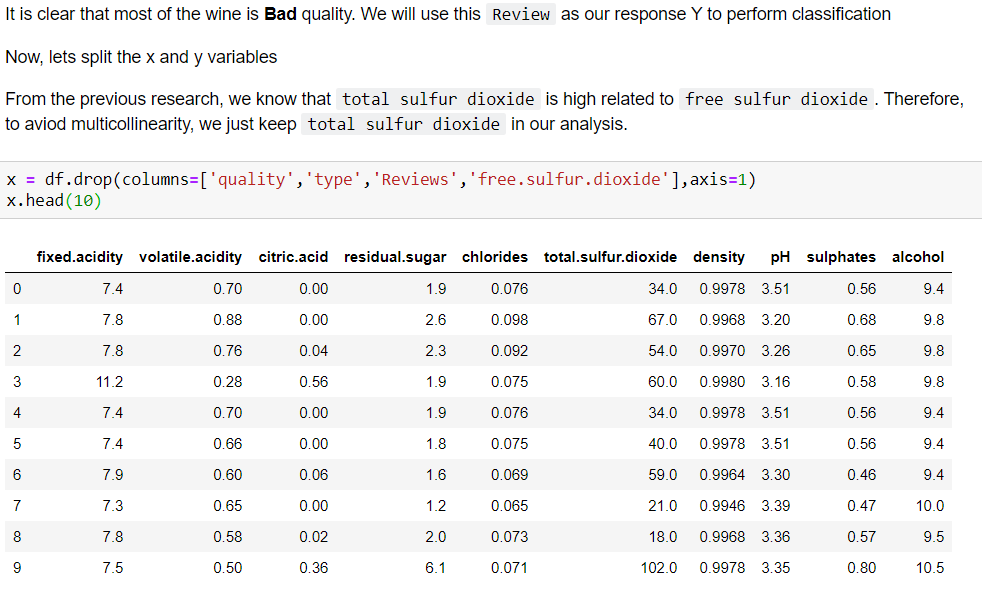
Then we will use this result to make prediction on test set and calculate the test error.

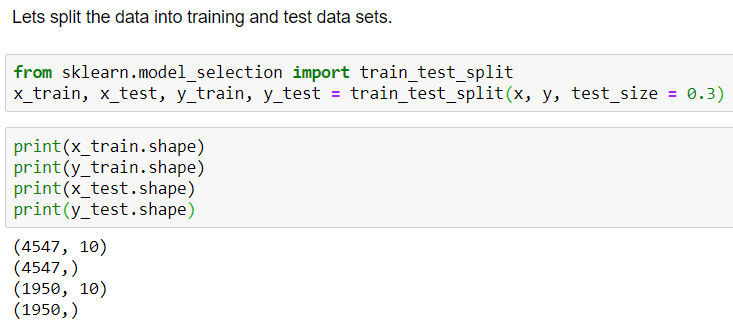


We can see the RMSE of lasso regression is 0.75, which is the biggest among these three methods. The performance is not good.

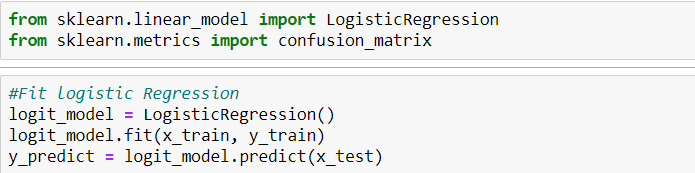
1. Classification

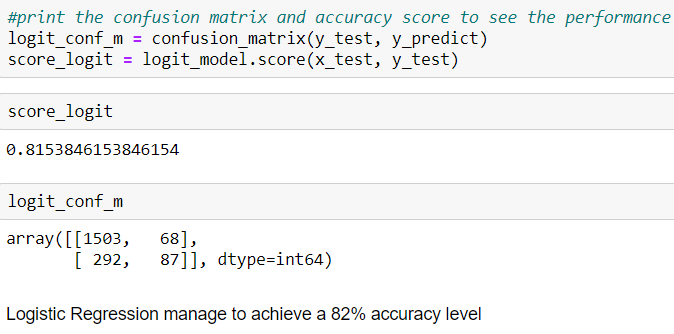




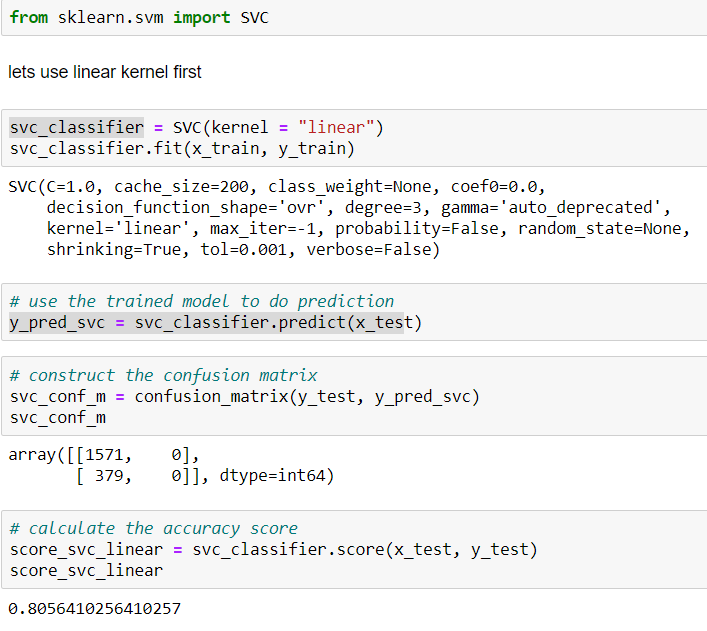


* 1. Logistic Regression for Review Category

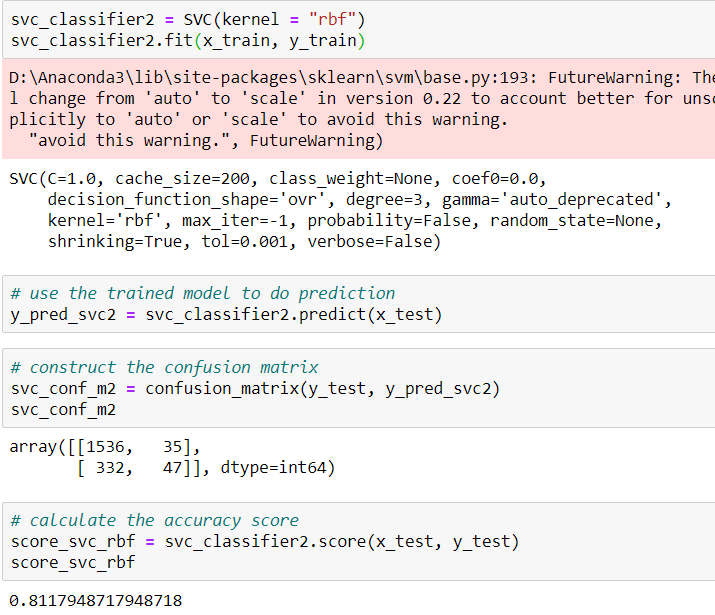




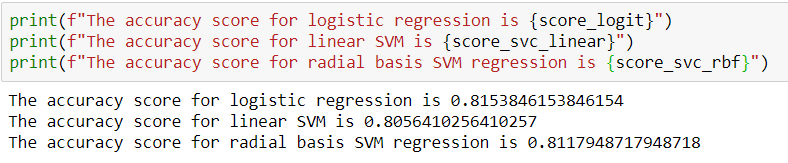
* 1. SVM for Review Category
     1. Linear Kernel



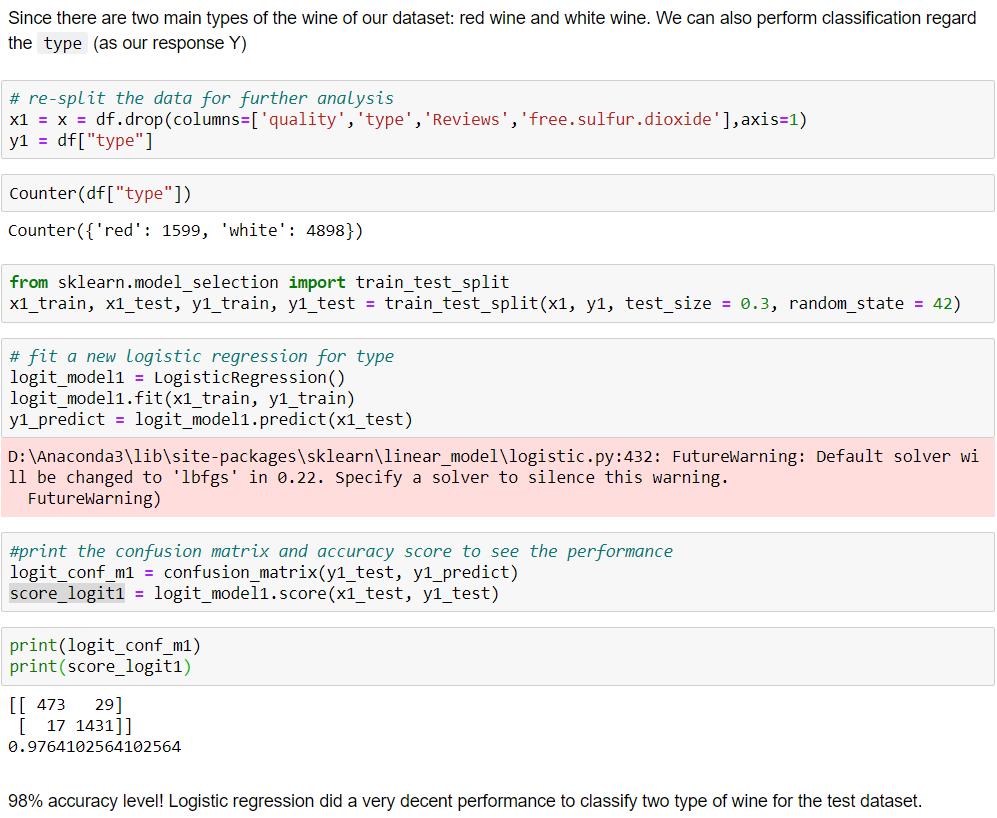
* + 1. Radial Basis Kernel



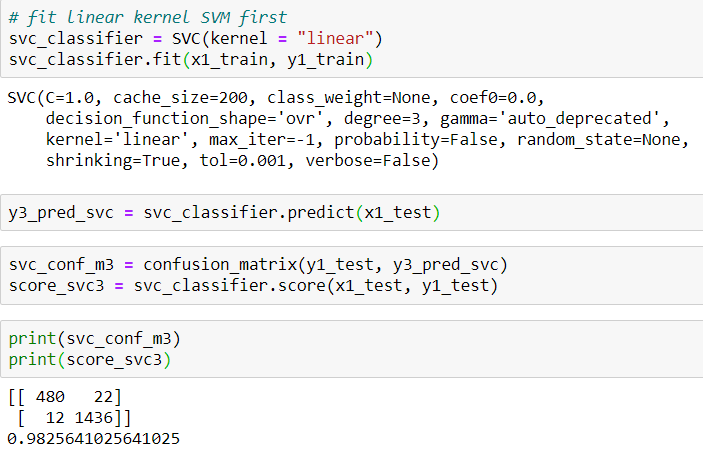
* 1. Decision Tree for Review Category
  2. Random Forest for Review Category
  3. Summary of Classification for Review Category



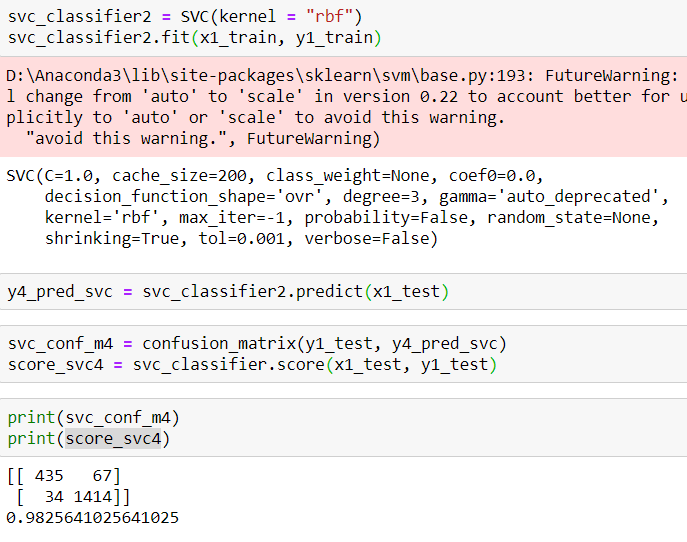
* 1. Logistic Regression for Wine Type



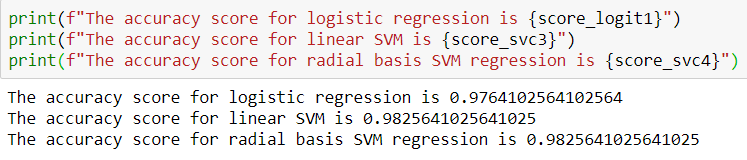
* 1. SVM for Wine Type
     1. Linear Kernel



* + 1. Radial Basis Kernel



* 1. Decision Tree for Wine Type
  2. Random Forest for Wine Type
  3. Summary of Classification for Wine Type



1. Wine, Wikipedia. URL: <https://en.wikipedia.org/wiki/Wine> [↑](#footnote-ref-1)
2. Wine Chemistry, Wikipedia. URL: <https://en.wikipedia.org/wiki/Wine_chemistry> [↑](#footnote-ref-2)
3. Wine Quality, Kaggle.com. URL:<https://www.kaggle.com/danielpanizzo/wine-quality> [↑](#footnote-ref-3)
4. Wine Quality, Kaggle.com. URL:<https://www.kaggle.com/danielpanizzo/wine-quality> [↑](#footnote-ref-4)
5. Acids in Wine, Wikipedia. URL: <https://en.wikipedia.org/wiki/Acids_in_wine> [↑](#footnote-ref-5)
6. A Detailed Explanation of Acids Used in Wine Making, Homebrewit.com <https://www.homebrewit.com/a-detailed-explanation-of-acids-used-in-wine-making> [↑](#footnote-ref-6)
7. Acids in Wine, Wikipedia. URL: <https://en.wikipedia.org/wiki/Acids_in_wine> [↑](#footnote-ref-7)
8. Acids in Wine, Wikipedia. URL: <https://en.wikipedia.org/wiki/Acids_in_wine> [↑](#footnote-ref-8)
9. Sweetness in Wine, Wikipedia. URL: <https://en.wikipedia.org/wiki/Sweetness_of_wine> [↑](#footnote-ref-9)