# 

A picture containing text

Description automatically generated

**CS 521 Information Structures With Python**

**Spring 2022**

**Wine Quality Analysis Report**

**Xiangxian Song, Weiyao Xu, Longyuan Tang**

May 4th, 2022

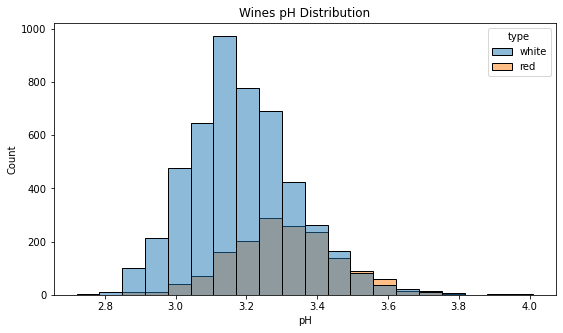
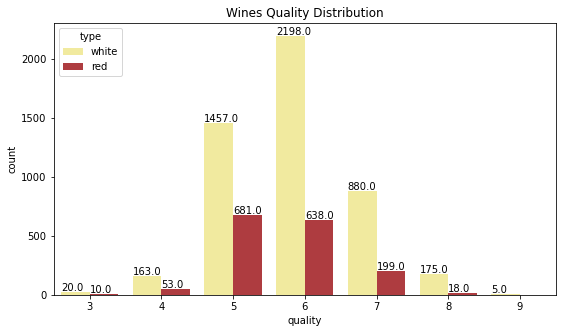
1. Background and Introduction

(A description of the problem you tried to solve)

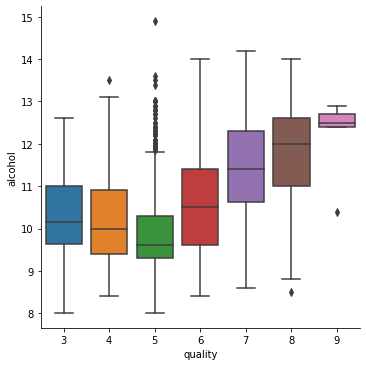
(The files you included and what they do)

1. Analysis Process

( Methods and Techniques you used along with "brief" explanations of how they should work)

First of all, we need to have a general understanding of the data. After knowing the attribute information of the original data, we combined the red wine and white wine and did some visualization. We use matplotlib.pyplot and seaborn python libraries to create plots of different dimensions, such as heatmaps, boxplots, scatterplots, etc. Through the graphs, we got a preliminary understanding of some basic characteristics of wine, the commonalities and differences between two wines, as well as the distribution of different indicators and their relationship with quality.

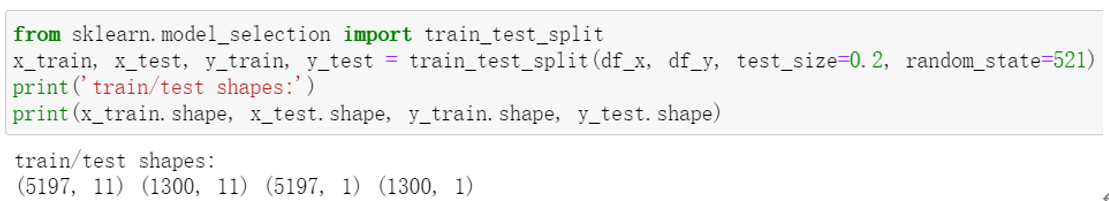
The top left plot shows the quality distribution of red and white wine, we can clearly see that the quliaty of wine is mainly concentrated in 5 and 6. Among them, the quality range of White wine is more extensive, ranging from 3 to 9, mainly concentrated in 6, and the total number of wine is 1.5 times that of quality 5; Red wine’s distribution of quality 5 and quality 6 is relatively similar, but significantly more than other quality rate. And through the top right plot which shows the pH distribution of red wine and white wine, we can find that the pH of the two wines is similar, and the white wine is generally more acidic. Most white wines have a pH in the 3.0-3.2 range, while most red wines have a pH in the 3.2-3.4 range. Also, in the range of 3.8-4.0 is almost red wine.

To go further, after general checking the potential relationships between different data attributes through heatmap. we also explored the relationship between individual index and quality. For example, the boxplot on the right shows the relationship between wine quality and alcohol content. we can notice that the average alcohol content of quality 3-5 decreases, while the average alcohol content of quality 5-9 gradually increases. In the part of quality 5 and up, the higher the quality, the higher the alcohol content. Among them, the alcohol content of wine with quality 5 has a wider distribution range, and the alcohol content of wine with quality 9 is more concentrated than other qualities.

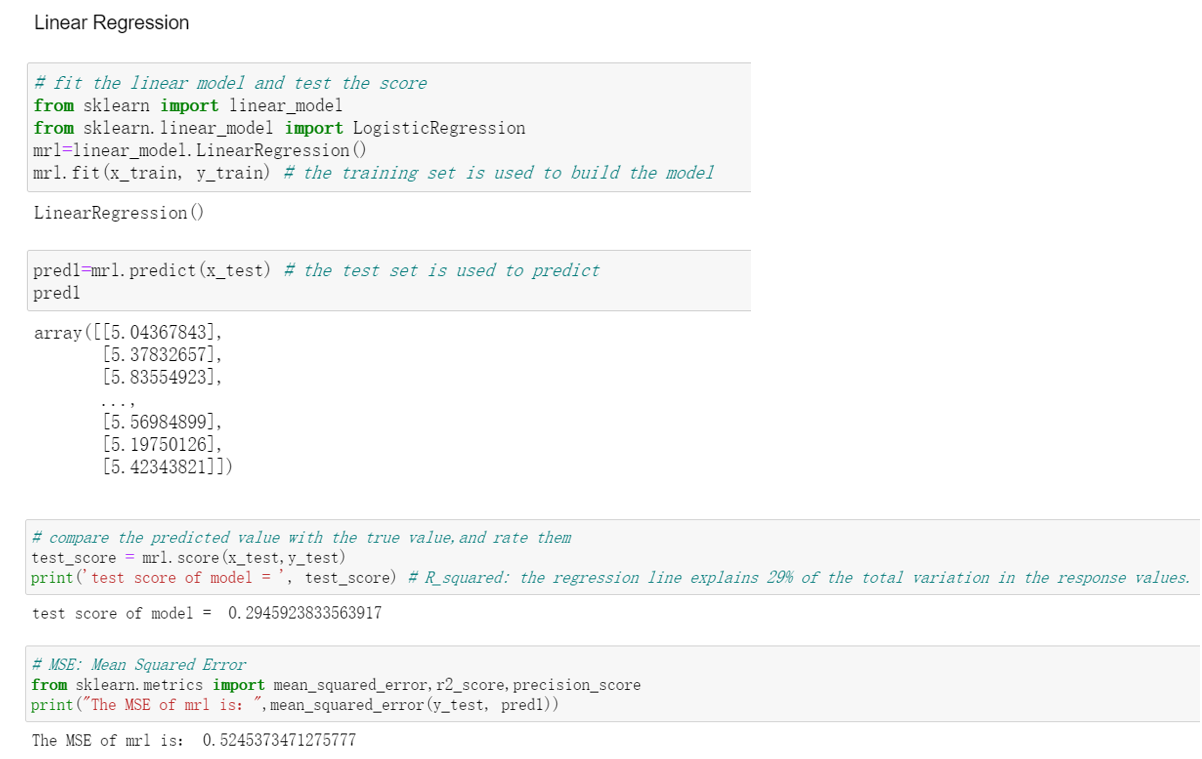
In the next part, we performed regression analysis.

1. **Regression**

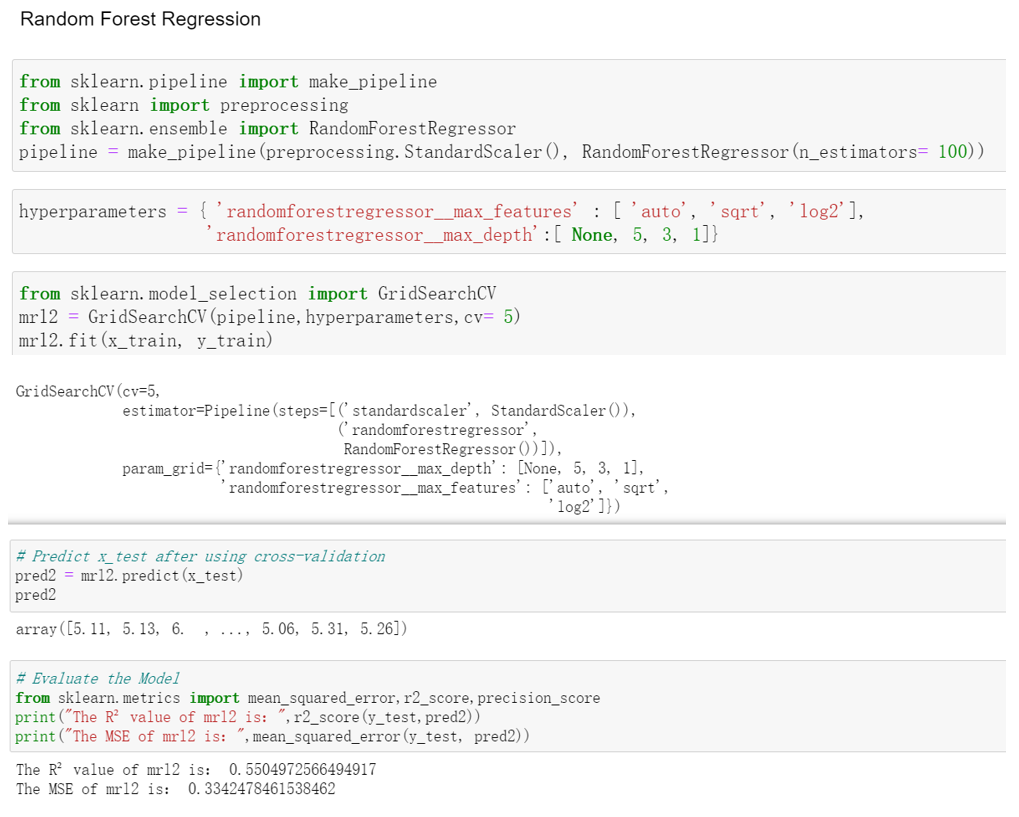
****Before modeling, we first need to check whether the data contain NA values. If there are a lot of NA values in the data, it means that we will lose much helpful information in the process of data mining modeling. Also, data containing NA values can confuse the modeling process, resulting in unreliable output. Such uncertainties will also be more significant, and the law implied in the model will be more challenging to grasp. So, we first use isnull(). sum() function and heatmap to detect NA values. Fortunately, our data is very clean, with no missing values.



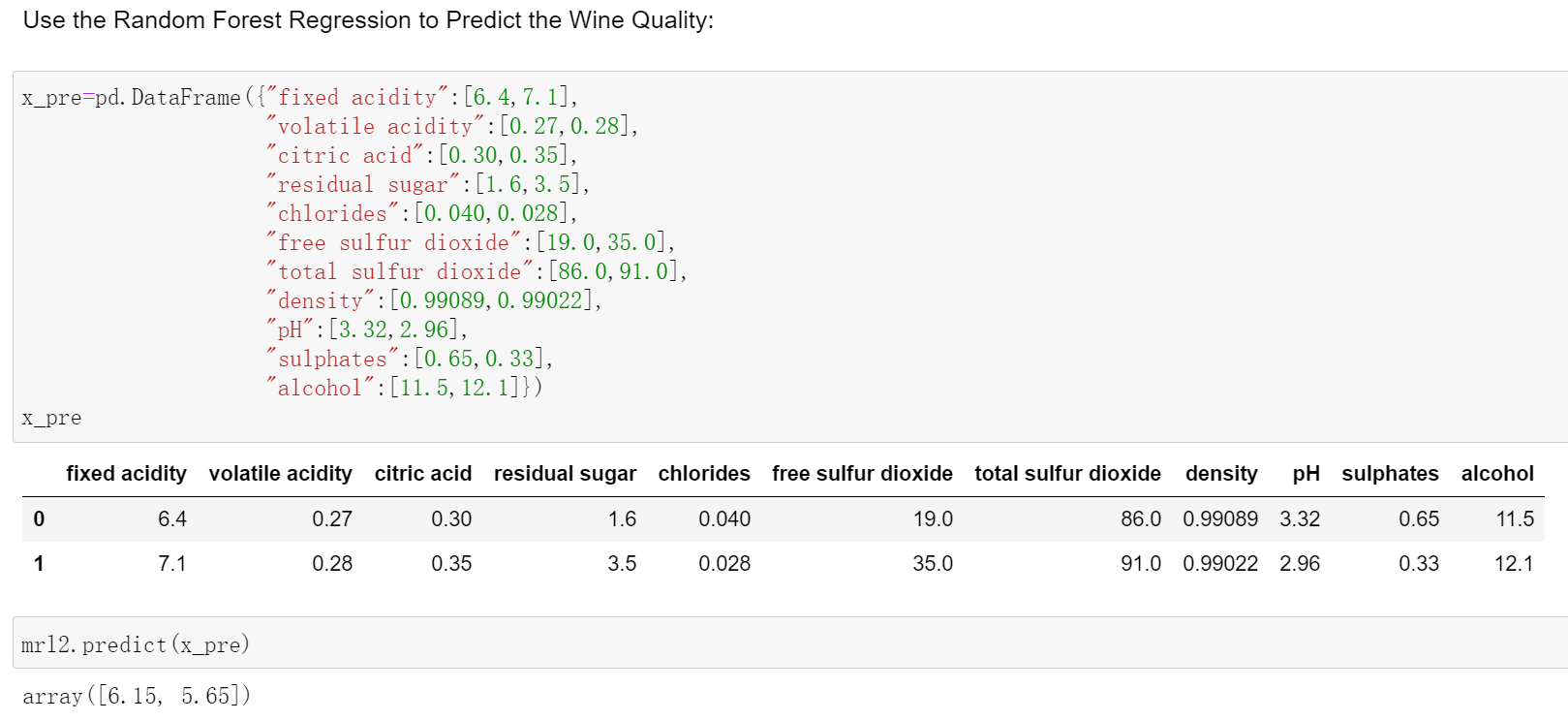
Before using machine learning algorithms, usually, we need to divide the dataset into a training set and a test set. When assigning training and test sets, the smaller the data in the test set, the less accurate the estimation of the model's generalization error will be. Therefore, we need to make trade-offs when dividing the dataset. Moreover, we should ensure that the divided data should be random; otherwise, the model will be useless. So based on the size of the entire dataset, we set the division ratio of training set data and test set data to 8:2. The method we use is the train\_test\_split method provided by sklearn. We use 80% of the data as the training set to fit the parameters in the model and then put the trained model on the remaining 20% of the data for testing to obtain the performance indicators of the model.



To further explore the relationship between wine quality and other variables, we will perform a Multiple Regression Analysis. Set wine quality as the dependent variable and other variables as independent variables. To achieve our goal, we use the linear\_model.LinearRegression() function to build a regression model. Then we use the test set to test the model to get the performance indicators of the model. We choose R^2 and MSE to judge the quality of the model. R^2 is called the coefficient of determination. It is used to determine the degree of fit. Its value range is [0,1]. If the result is 0, the model fits poorly; if the result is 1, the model is error-free. The larger the R-Squared, the better the model fitting effect. However, our R^2 is only 0.29, which is relatively small. That shows that our regression model does not fit well. Another indicator, MSE, is called Mean Squared Error. It uses the difference between the predicted and actual values to judge the pros and cons of the model, where the smaller the difference, the better the model. Moreover, the size of MSE for our model is 0.52, which means our model is not accurate enough.

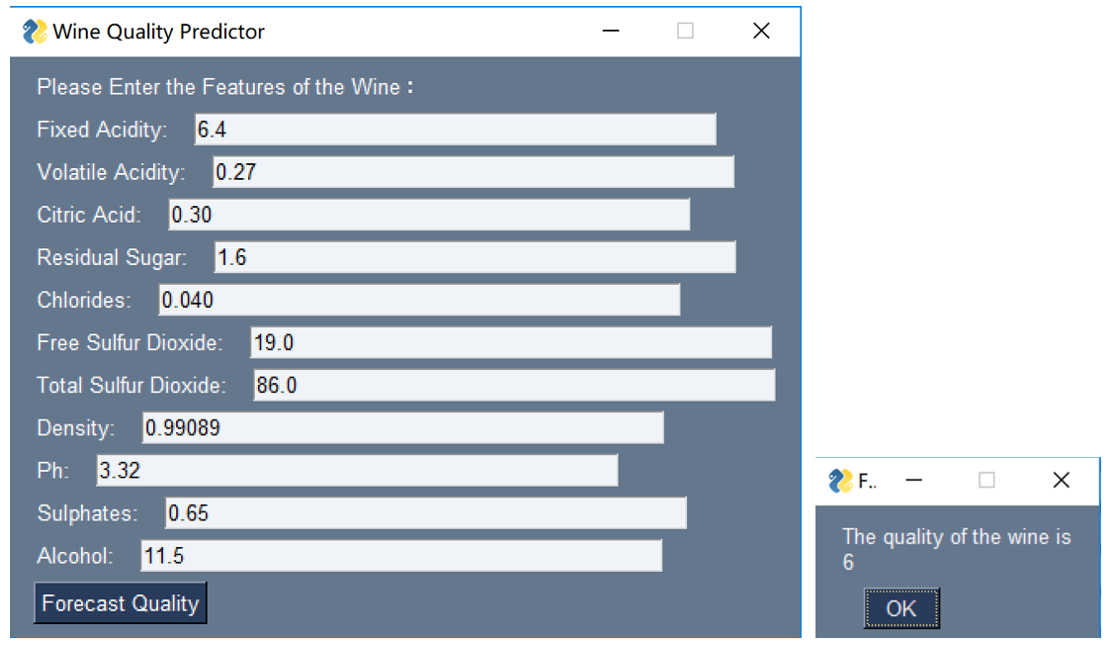


To improve performance, we reflect on the reasons why the model may not fit well and be inaccurate. One of the significant reasons is that the precision of the imported wine data is not the same, so we need to normalize it first. Normalization assumes that all features are centered around zero and have roughly the same variance. To achieve this, we need a modeling pipeline. We first transform the data using StandardScaler() and then use the RandomForestRegressor() function to fit the model. At the same time, we need to use hyperparameters. It is a parameter whose value is set before starting the learning process, not parameter data obtained through training. Typically, hyperparameters need to be optimized. It selects a set of optimal hyperparameters for the learning machine to improve the performance and effectiveness of learning. After that, we employed cross-validation. It is a process of reliably estimating the performance of a method that builds a model by training and evaluating our model multiple times using the same method. Finally, we still use R^2 and MSE to judge the quality of our model. We can see that our R^2 increased from 29% to 55%. It means that the fit of our model has become higher. Moreover, the MSE dropped from 52% to 33%. It means that the smaller the average distance of points in the test dataset from the model, the more accurate our model will be.

****

Due to the improved fit and accuracy of the Random Forest Regression model, we use it to predict the wine quality. We fabricated two wines to verify the feasibility of the model. After that, we created a data frame to store values for the two wines for each one of the predictor variables in our model. And the quality of the two wines was 6.15 and 5.65, respectively.

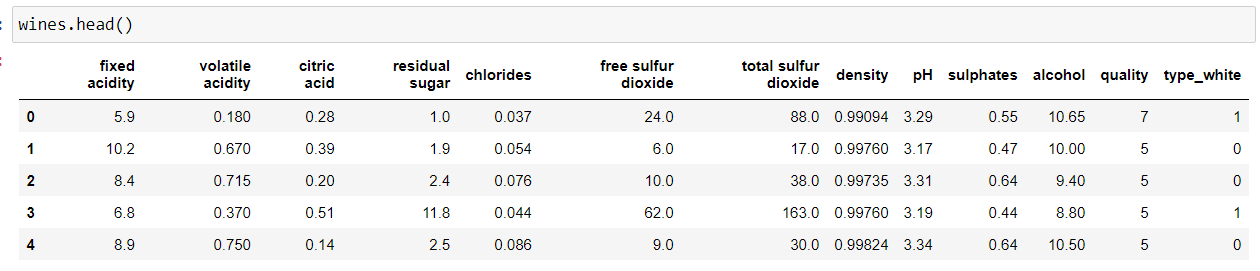
User Interface



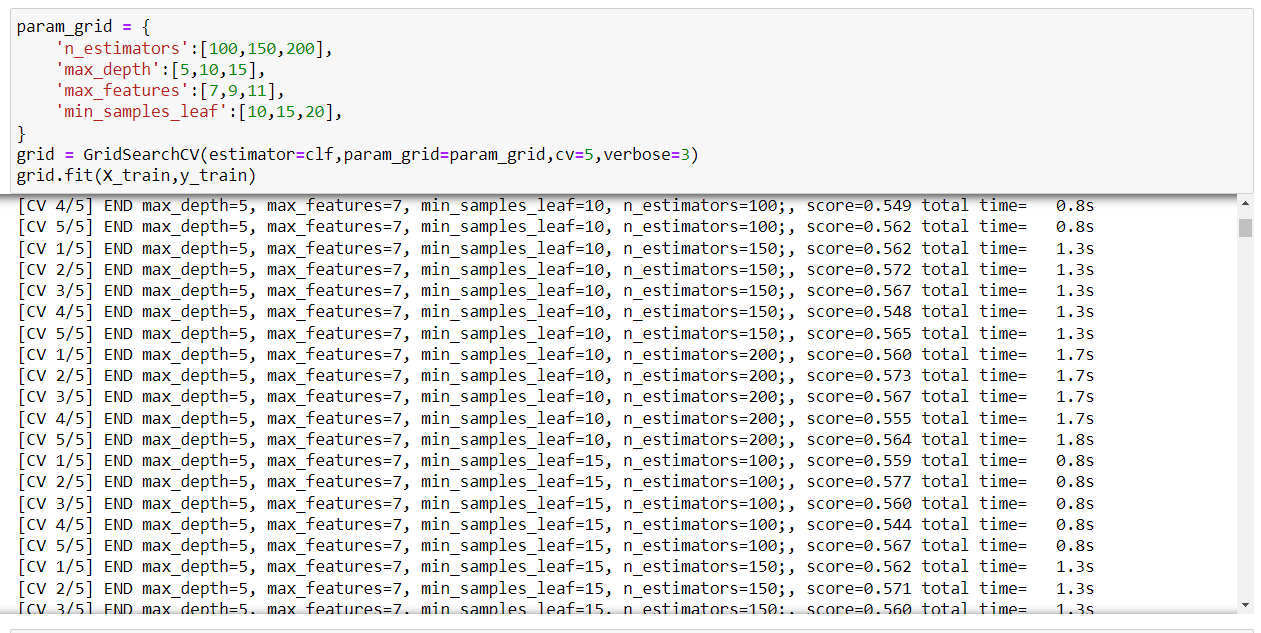
In order to allow the user to correlate with the prediction results using the input, we designed a concise user interface with the help of the PySimpleGUI package. We built a popup window that allows people to enter various attributes of the wine and then click on "Forecast Quality" to predict the quality of the wine. This result was achieved using the Random Forest Regression model created earlier.

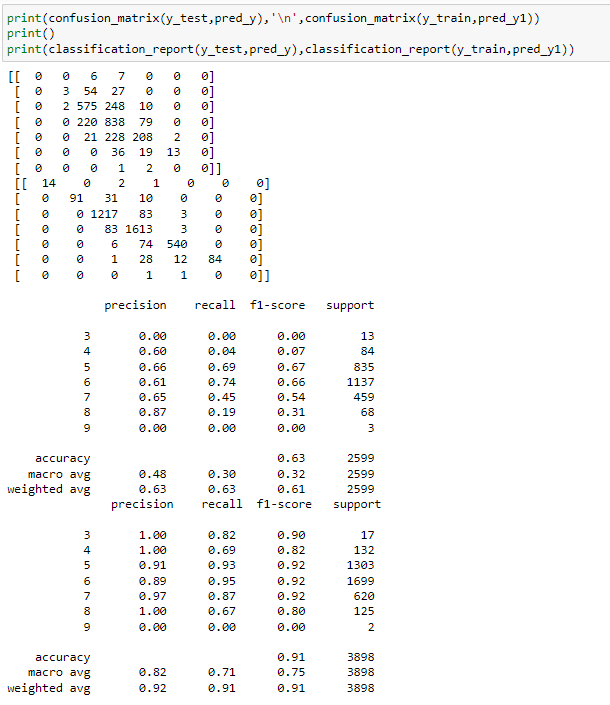
1. Classification

Since the column that we should predict is from 1 to 10. Instead of using a regression model, our team decided to use a classification model and pick which one is better. There are two models we used, random forest classification model and support vector machine model. First of all, since the type\_wine is a categorical variable, the computer cannot recognize string type. We need a dummified type\_wine column to 0 and 1 where white wine is 0 and red wine is 1. We don’t need to scale the data as we do in a regression session, because a classification model can not build a linear model which needs everything in the same standard deviation.

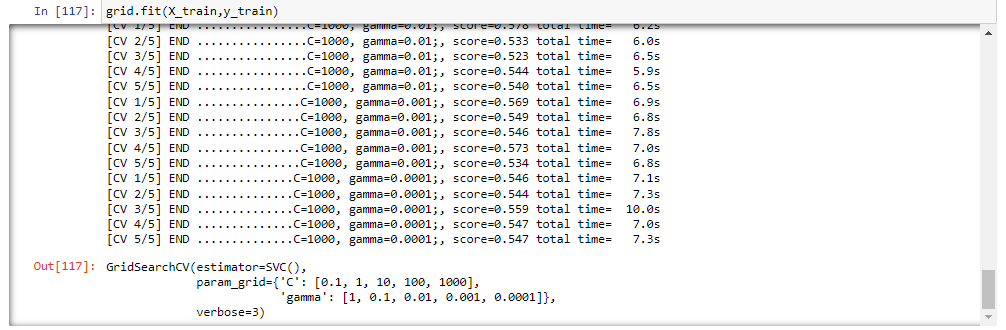


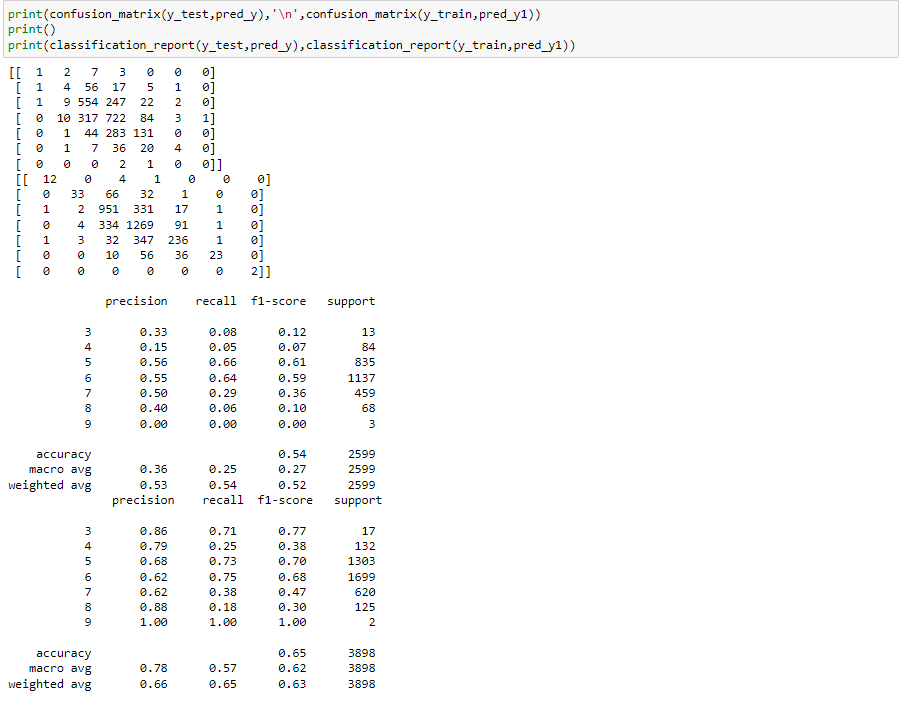
Supervised experiment is necessary for classification models, because we need to use a grid search CV function to list all the parameters and select the best one from it, we need to use supervised experiment to know whether our data is overfitting to the training dataset or not. However, an overfitting model is not a bad model for random forest, if the predicted result is good, it doesn’t matter if the model is overfitting. After listing all the parameters, the best pattern is max\_depth = 11, max\_features=9, n\_estimators=150. The accuracy of the model is 63%. While splitting the test and training data, we decide to use 0.45 because the data while wine quality less than 3 or greater than 8 are few. If defining only 30% data as train data, it cannot train any data marked as quality of 1, which will influence the result of the model..





The accuracy is pretty good since it is a real dataset, it is hard to achieve an accuracy of 0.70 or more. In order to make sure the model is the best classification model for this dataset, the other model, the support vector machine, is applied to the same dataset. It also applied a grid search CV function to find the best pattern by changing C or gamma value. As a result, the best model has an accuracy of 0.54 and the best pattern is when C is 0.1 and gamma is 0.001.





From those two models we infer that the random forest model is better for classification in this particular dataset. However, the flaw is obvious, the confusion matrix shows there is still no data where wine quality is equal to 1 in the test group. Thus we don’t know the accuracy when using this model to estimate the low quality wines.

1. Conclusion and Suggestion

By comparing the results from different models, our team concluded that a random forest regression model is better for the entire dataset. Firstly, a classification model has uncertainty due lack of data in low quality wines. A regression model can predict value through some linear or nonlinear relationship which has a better prediction on those lacking data.