Analysis of Wine Quality Based on Data Visualization

1. **Motivation**

Traditional wine identification uses organoleptic tasting to judge its quality. This requires that the taster is a trained wine tasting expert. However, the organ-tasting results are easily affected by various factors. Therefore, data mining methods are increasingly used in wine quality identification. The article examines the key variables that affect the quality of red/white wine, the correlation between variables and the consistency of the influence variables on red/white wine. At the same time, we will make a judgment to determine whether it is enough to predict the quality of red/white wine

1. **Related Works**

2.1 Classic techniques to appraise the quality of red/white wine

In order to reflect the different quality of the wine, the traditional method is often to detect the senses, physical and chemical components. Generally speaking, for most traditional consumers, they will judge the quality of the wine by simple looking at the origin, year, grape variety and level of the wine label. Spain is the world's largest grape growing country, which the place with sufficient sunshine and natural conditions, it can provide favorable conditions for the growth of grapes. But because of the vast territory, the qualities of wine vary greatly. Thus, the customers will also look the wine international wine competition awards as a reference. In these later, the final looking of the bottle design and construction will help the observation and inference wine quality. Secondly, the wine inspector will also look at the appearance, color, concentration, tone, clarification, the consistency of the bubbles, the type of aroma, the degree of between concentration and harmony, and the aftertaste of the taste. The quality of a wine is assessed by the balance between appearance, aroma and taste. Finally, the relevant testing institutions will detect some physical and chemical indicators and analyze the composition of the wine, which will detect the flavor and stability of different wine quality.

2.2 State-of-the-art techniques appraise the quality of red/white wine

The French chemistry professor invented the SINF and NMR methods, which can be used to accurately test the authenticity of alcohol. The SINF method is based on the isotope technique to detect the purity of the wine and it can also be used to accurately identify whether or not the wine is added with sugar, spices or other harmful chemical compounds. This technique is widely used in France, Germany and Italy. Then, the NMR method uses a device to scan a wine's molecular structure to determine the composition of the wine. The second, in 2008, French scientists invented a way to use particle accelerators to identify wine. And they can use x-rays to pinpoint the date and place of production. The last one, the scientists have invented a "electronic tongue" at the university of south Australia. It can be to electrochemical analysis of vintage, quality, measurement of component content and then the information will put in the computer system to make contrast. We can concluded that the type and quality of alcohol.

However, these methods are based on existing wines that are tested for quality assessment. If we want to predict the future quality of wine before the output of wine, it is very important to data mining and analysis on factors that affect the quality of wine.

1. **Designed System**

This study combines Python language and Matplotlib to analyze and process data. Python is a widely used high-level programming language, which emphasizes the readability of the code. Compared with the C language, python allows developers to express ideas with less code. Matplotlib is a visual interface for the Python programming language and its numerical extension package NumPy, which provides the reader with visual images.

Then we made a mean distribution histogram and the normal distribution curve to illustrate the relationship between each attribute and quality. The histogram is an intuitive manifestation of the simplification of large amounts of data, which has a better visual effect. In the normal distribution curve, we can examine the peak value and analyze the effect of different quality under the specific attribute. Next, scatter plots are used to illustrate the relevance of each attribute. Scatter plots determine the correlation between two attributes by the degree of density of points and the color. Due to too many data samples, the scatter chart does not look intuitive. We use linear regression to examine the correlation of all attributes by cross test, which can be achieved quickly by Seaborn visualization. The redder the square area, the higher the correlation between the representative attributes. Through the regional colors, we can intuitively experience. Finally, we forecast the quality of red wine and white wine by establishing four models: decision tree, neural network, K-mean algorithm and linear regression.

1. **Data Description**

Firstly, it is a priority to import data, explore and preprocess. The data source we used was obtained from UC Irvine Machine Learning Repository. A total of two dataset were red wine and white wine respectively. The data contains 11 variables, which are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol respectively. A total of six levels of quality indicators, the minimum index of 3, the maximum index of 8. When importing data, it is found that the Excel table is segmented by semicolons, so when importing, we need to remove the semicolons before we can read the data normally.

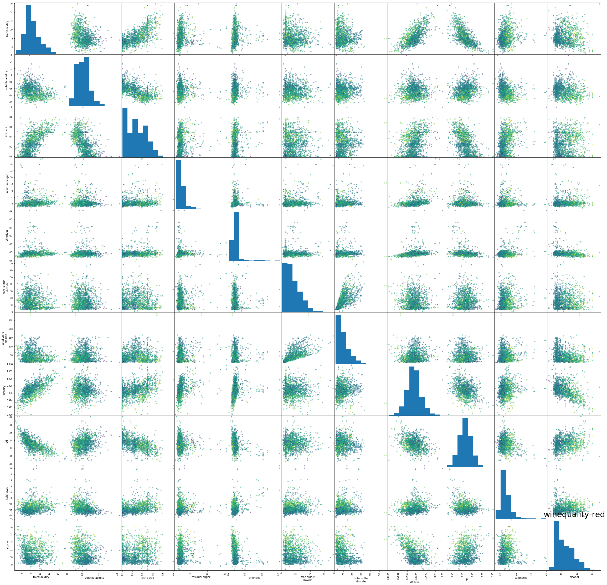
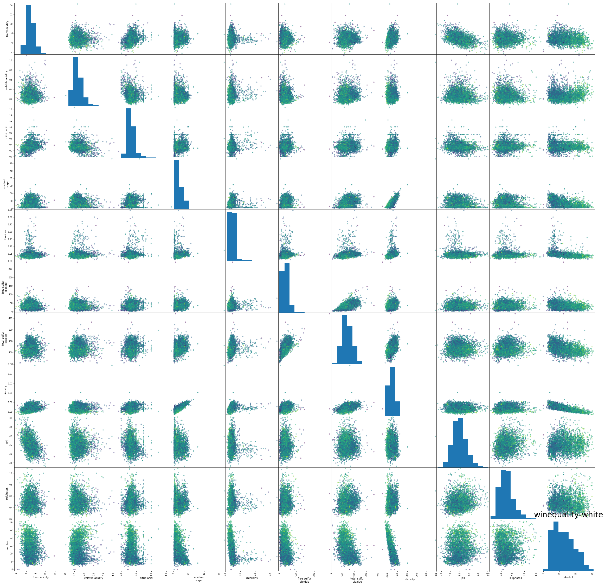
1. **Results**

**第一题：Which variable(s) is (are) the most influential one(s) to the quality of the red/white wine? （苏）**

**第二题：Are there any correlations between different variables? （郑）**

Firstly, the correlation analysis between variables is carried out. The following figure shows the scatter plot between variables. The scatter plot matrix can simultaneously draw the scatter diagram of each variable, so as to quickly find the main correlation between multiple variables. Through a scatter diagram we can preliminary find that it is difficult to observe from the vision of a specific related degree between every two attributes. Probably, it can be seen that for red wine, the correlation between the density, PH and fixe acid is higher. It has a clear trend of positive correlation or negative correlation between them. But for white wine, the distribution of scatter plots is not so easy to observe the correlation.

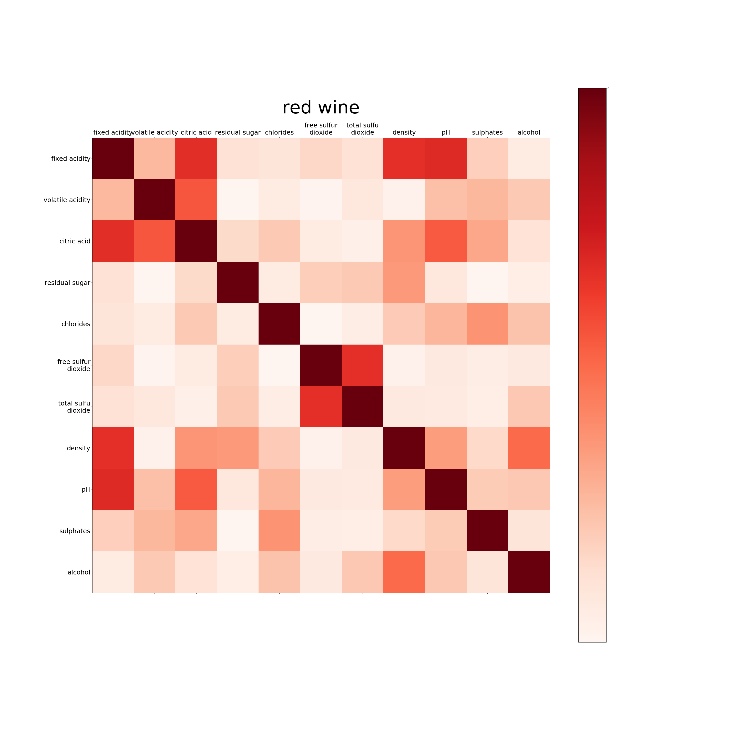
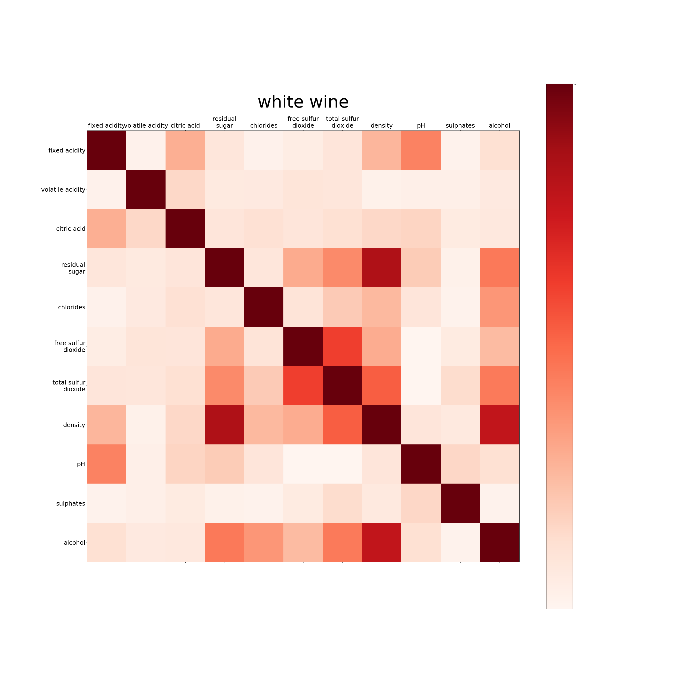
The Scatter diagram of correlation between variables:

Red Wine White Wine

Because of the limitation of scatter diagram, we also chose the visualization of data analysis of heat map. Heat map can simply and intuitively show the difference of data by color difference and brightness. The graph below is the correlation between the two attributes. If the correlation stronger, the corresponding color will more red.

Heat map of attribute correlation：

Red Wine White Wine

From the figure, we can observe that for red wine, fixed acid is strongly correlated with citric acid, PH and density. And the correlation between free sulfur and total sulfur dioxide is strong. For white wine, we can see that the correlation between free sulfur dioxide and total sulfur dioxide is stronger. Also, the correlation between residual sugar and density or density and alcohol are also closely related.

In order to illustrate the correlation between variables more accurately, we also use excel to process and analyze the basic data. The correlation degree between two attributes is obtained through the specific numerical value. And the higher the value, the stronger the correlation. For example, the following picture is the red wine excel figure of attribute correlation:

Red Wine:



White Wine:



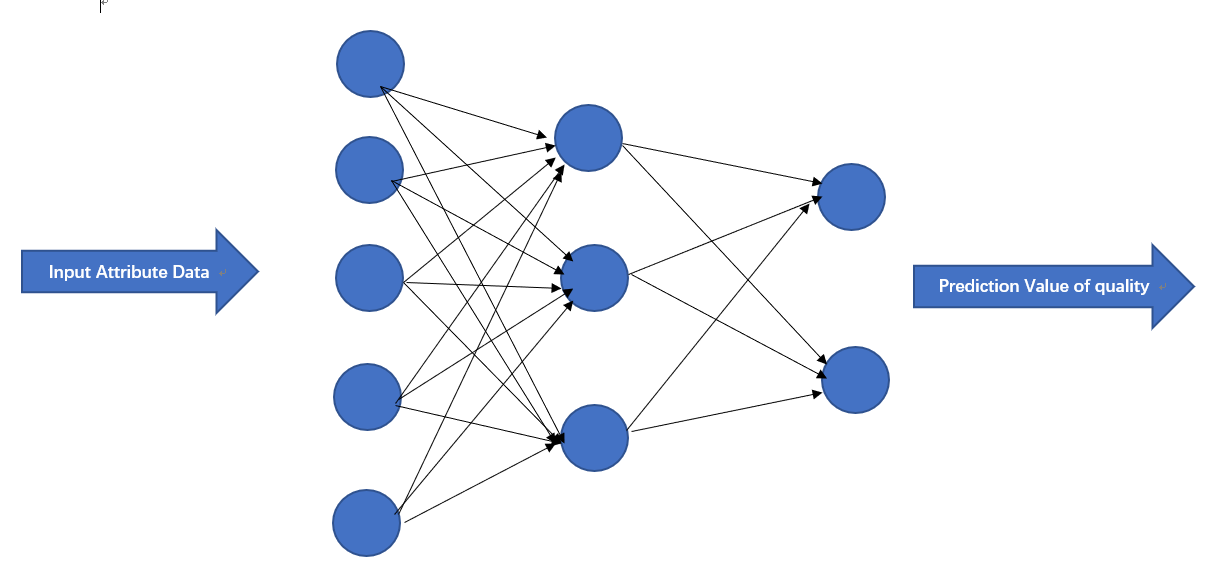
**第三题：Are the influential variables of the red wine the same as those of the white wine?¬（苏苏，同第一题）**

**第四题：Are these given variables enough for making prediction/evaluation on the quality of the red/white wine?**

In order to fully use data to predict the quality of wine and exclude other factors that may influence the conclusion of whether these given variables enough for making prediction, we use four machine learning models, including K-Nearest Neighbor classifier, Neural Network, Logistics Regression and Decision Tree Classifier. And each single model has its own different parameters to achieve the best prediction performance.

Firstly, we split the data into two sets, 75% for training data and 25 % for testing data. We use training data to train these four machine learning models. We compare these four models result and find out the best-performance models with optimized parameters.

1. Neural Network:
   1. Design of Neural hidden layers:

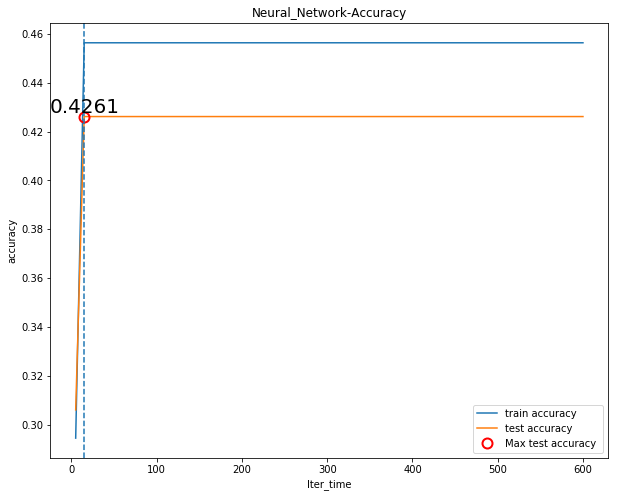


* 1. Train Accuracy, Test Accuracy relationship with iteration time:

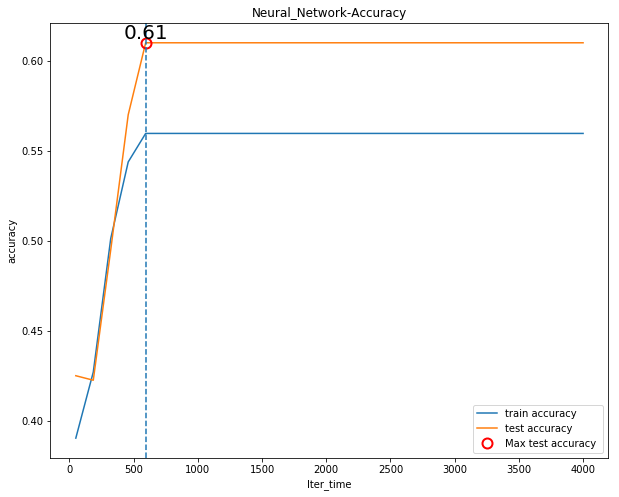
We visualize the accuracy of result, whose iteration time ranges from 0 to 600.

We visualize the accuracy of Neural Network, whose number of iteration from 1 to 600. And find out the highest accuracy and corresponding iteration time.

* + 1. White Wine:



* + 1. Red Wine:

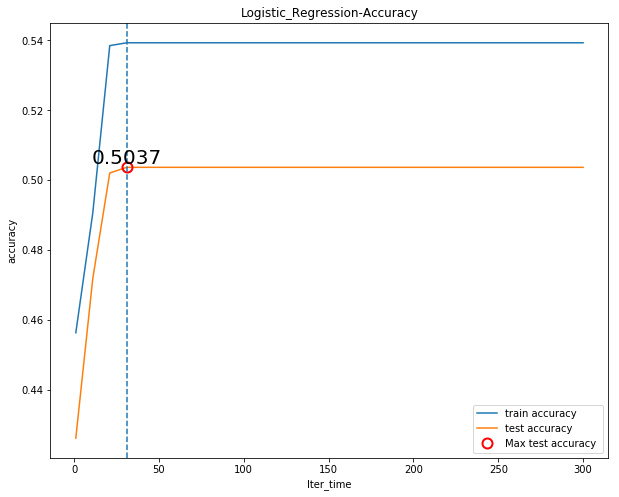


1. Logistic Regression:

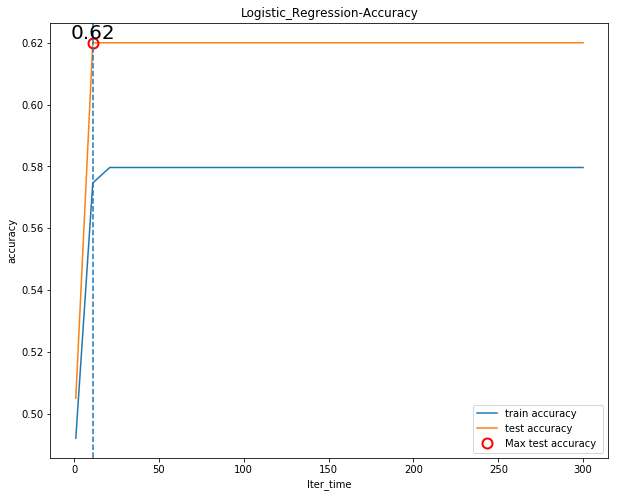
We visualize the accuracy of result, whose iteration time ranges from 0 to 300.

We visualize the accuracy of Logistic Regression, whose number of iteration from 1 to 300. And find out the highest accuracy and corresponding iteration time.

* + 1. White Wine:



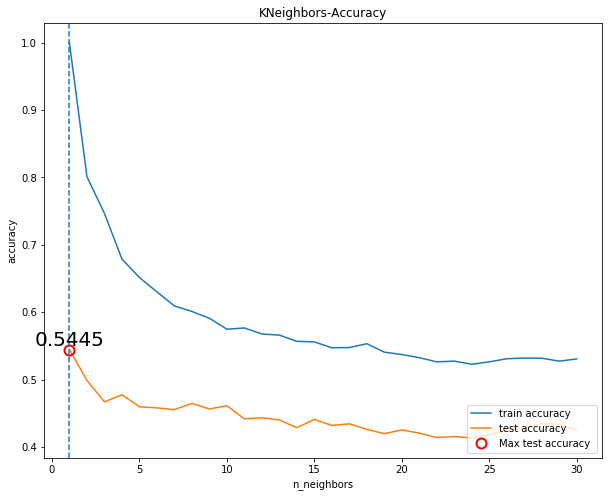
* + 1. Red Wine:



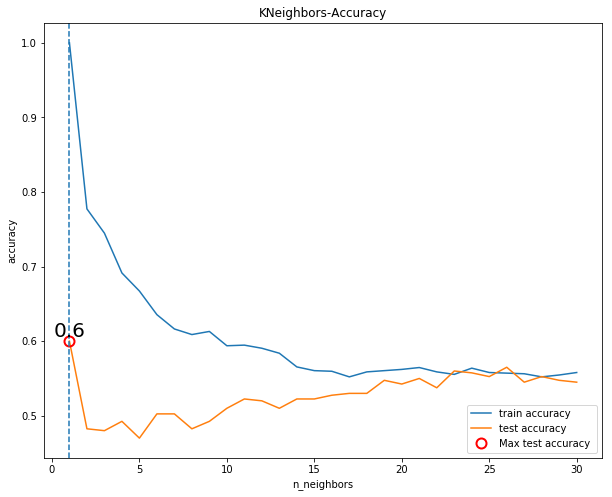
1. K-Nearest Neighbor classifier:

We visualize the accuracy of K-Nearest Neighbor classifier, whose number of neighbor ranges from 1 to 30. And find out the highest accuracy and corresponding n\_neighbors augument.

* + 1. White Wine:



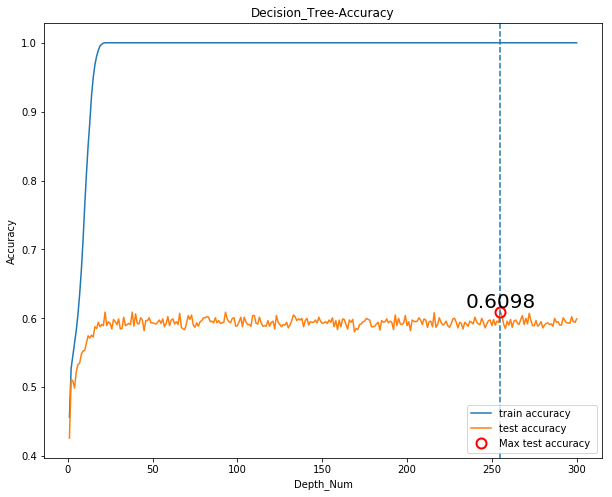
* + 1. Red Wine:



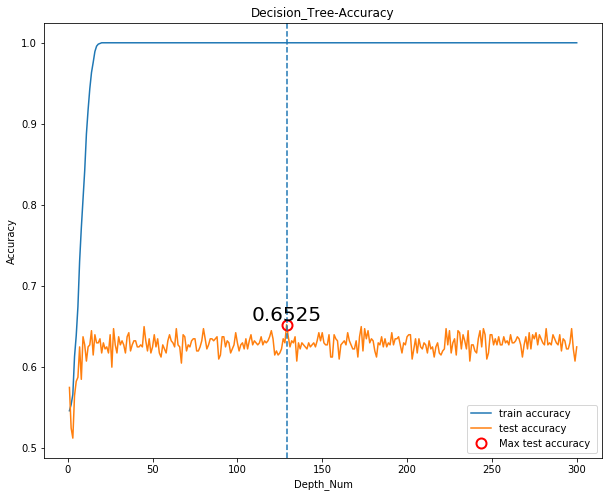
1. Decision Tree Classifier:

We visualize the accuracy of Decision Tree Classifier using ‘CART’ algorithm , whose number of depth ranges from 1 to 300. And find out the highest accuracy and corresponding Depth\_Num augment.

* + 1. White Wine:



* + 1. Red Wine:



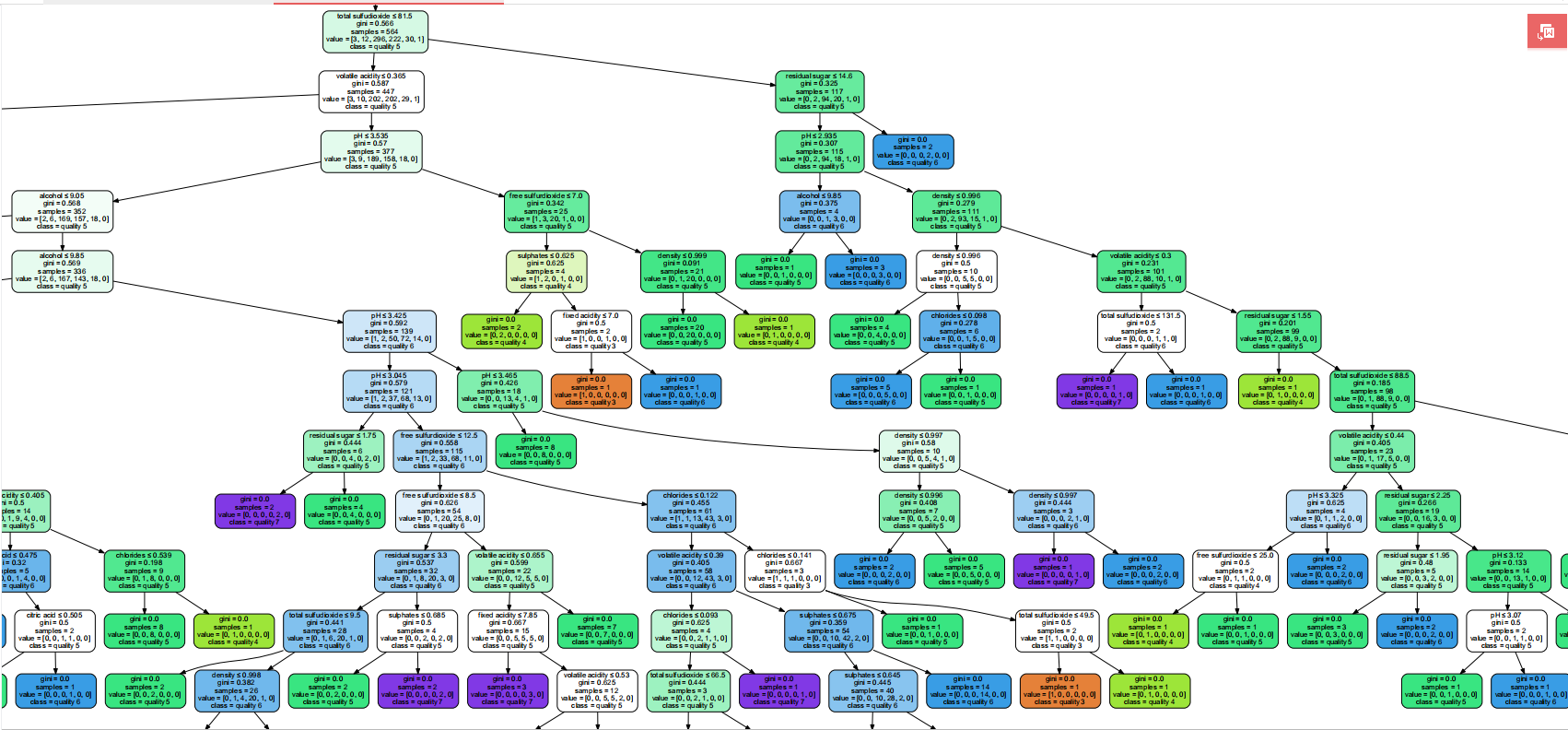
Conclusion and result:

Even we use four different machine learning models, and test many corresponding parameter of models, but the test accuracy is not high enough. The best-performance prediction model is Decision Tree Classifier and it achieves the highest accuracy 0.625 for Red Wine with Depth\_Num parameter around 130, and 0.6098 for White Wine with Depth\_Num parameter around 250.

So, the decision tree model is the best model to predict the quality of white wine and red wine data. We use Graphviz - Graph Visualization Software to visualize the decision tree models.\

Uncompleted of tree structures are shown below:

For Red Wine:



For White Wine:

