**Final Project Report**

|  |  |  |
| --- | --- | --- |
| **Selected Topic** | **Visual Analytics on Red Wine Quality** | |
| **Group Membership** | **Student ID** | **Name** |
| **Member 1** | **17427827** | **LIN JIEQIN**  **(Main Contributor)** |
| **Member 2** | **17437156** | **SU YING** |
| **Member 3** | **17415381** | **ZHENG YAWEN** |
| **Member 4** | **17420229** | **QIN XIN** |

**Part 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

**Part 2: 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 an "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 conclude 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.

**Part 3: Tools and Task Processes (Designed System)**

This study involves Python language and Matplotlib, Scikit-Learn three main tools to analyze, process data and visualize our results.

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.

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Our Project involves many other library of Python, such as Graphviz to visualize the decision tree model, Panda or Numpy to process the data.

Main process:

The first step is to observe probability density distribution of each attribute in different quality of red wine and white wine. In the probability density distribution curve, we can examine the peak value and analyze the effect of different quality under the specific attribute.

We made a mean distribution histogram 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.

Next, scatterplots Matrix is 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 correlation matrix to examine the correlation of every two attributes, 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 logistics regression.

In this report, we use four computing models to examine the accuracy of making prediction by given variables and dataset.

* Decision Tree(DT): DT is a decision support tool by utilizing tree graph or decision model to analyzing the possible of outcomes, including random event results, resource costs, and utility.
* Logistic Regression: Logistic regression model is used to analyze the relationship between independent and dependent variables.
* K-Nearest Neighbors: KNN uses the vector space model to classify, in the same category of cases with high similarity. It is able to evaluate the classification of the unknown class cases by calculating the similarity to the known class cases.
* Artificial Neural Networks(ANNs)：Ann is used to evaluate the given dataset for judging the quality of given datasets.

**Part 4: Data Description and Pre-Processing**

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.

**Part 5: Results**

* 1. **The relationship between quality and variables**

The first step is that using python to create 22 probability density distribution charts for each attribute in different quality in red wine and white wine datasets, which indicates how much influence between each attribute and quality.

The diffusion of an attribute in different quality density distribution charts shows that the attribute is good for distinguishing different quality. On the contrary, if the distribution of probability density is denser, more overlapping, this means that the attribute has little effect on quality. Taking figure 5.1 and figure 5.2 for examples. Abscissa represents different quality, and ordinate represents probability density. The previous one is a probability density distribution charts for alcohol in red wine dataset in different quality. Figure 5.2 is a probability density distribution charts for residual sugar in the dataset of red wine in different quality. Comparing two figures, obviously, we can know the distribution of second one is overlapping than first, and the absolute value of correlation is lower. It tells us that alcohol has a greater impact on the quality of red wine.

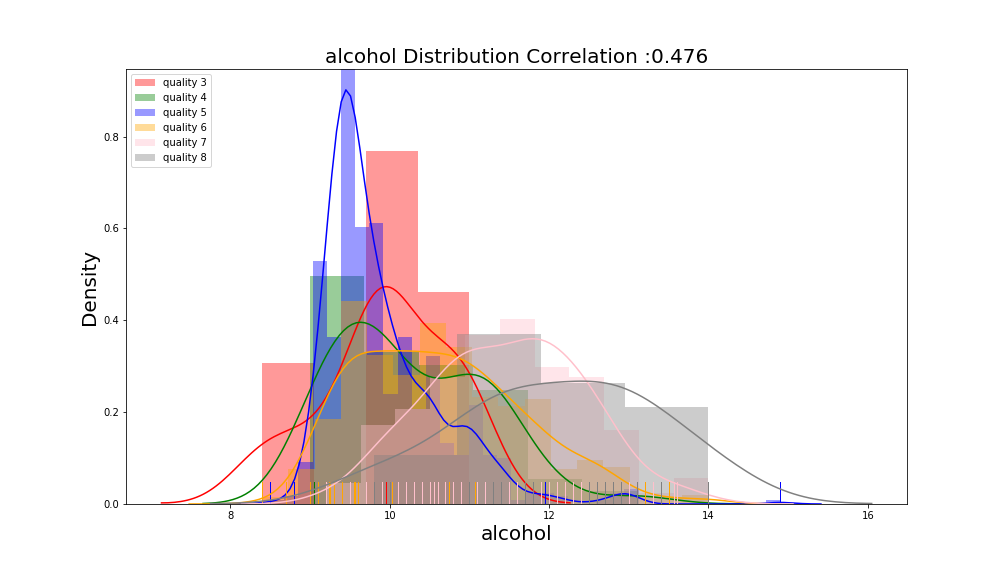


Figure 5.1

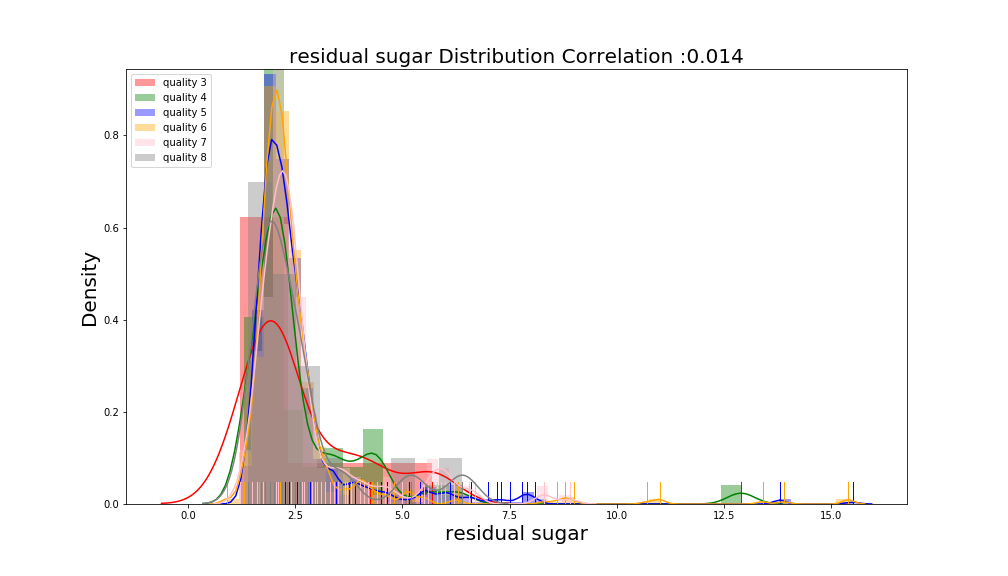


Figure 5.2

In order to better understand the relation between each attribute and quality, we calculate average value of each attribute in different quality and analyze the specific effect of each attribute on quality, whether positive or negative. We encode data of relation by using bar char generated by Python matplotlib. Choosing two attributes for explaining, the chlorides and alcohol, figure 5.3 and figure 5.4. The horizontal coordinates represent the quality, and the vertical coordinates represent the mean value of each attribute in different quality. In figure 5.3, it shows that higher quality with lower chlorides. However, the white wine has higher quality with higher alcohol. This shows that alcohol has a positive effect on white wine, while chlorides has a negative effect on wine.

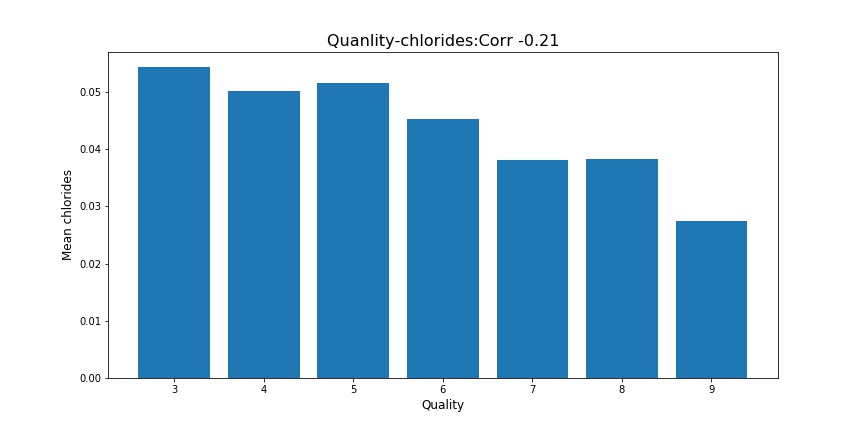


Figure 5.3

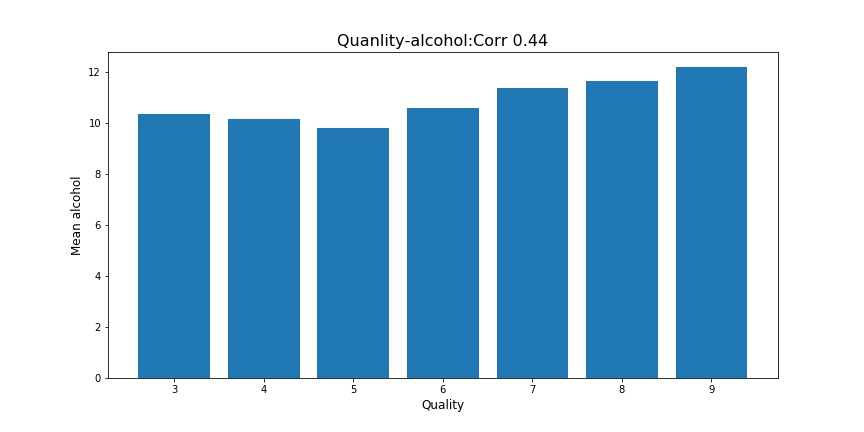
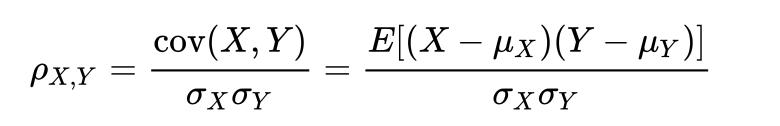


Figure 5.4

After that, because probability density distribution charts and mean value of each attribute in different quality unable directly comparing the relationship between different attributes and the quality in two wines, creating the bar chart which only explains the value of relationship between variables and quality (figure 5.5, figure 5.6). The horizontal coordinate represents the absolute value of the correlation coefficient, and the ordinate is a different variable. From figure 5.5 and figure 5.5, we are able to intuitively see that alcohol has a great impact on the quality of red wine (0.476) and white wine (0.436). Whereas, the influence of residual sugar can be ignored in red wine, the number of free sulfur cannot determine the quality of white wine.

In addition, the top three variables which are play important parts on the quality of red wine are alcohol, volatile acidity, sulphates. On the contrary, anyone in alcohol, density or chlorides can change the quality of white wine.

The algorithm is shown as following [2]



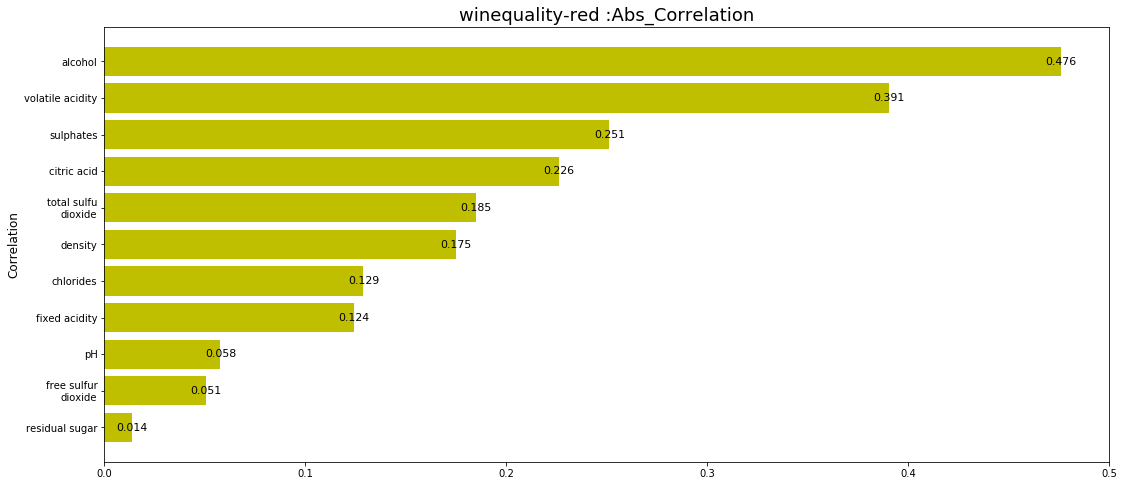


Figure 5.5

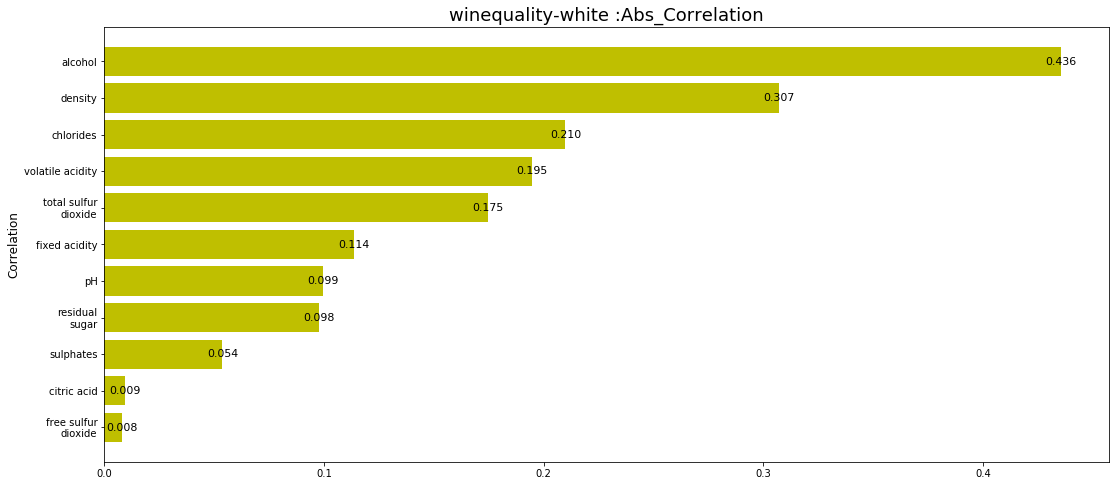
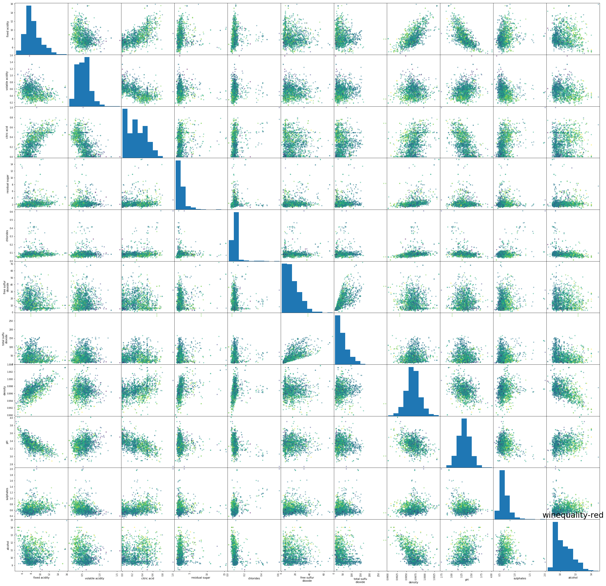
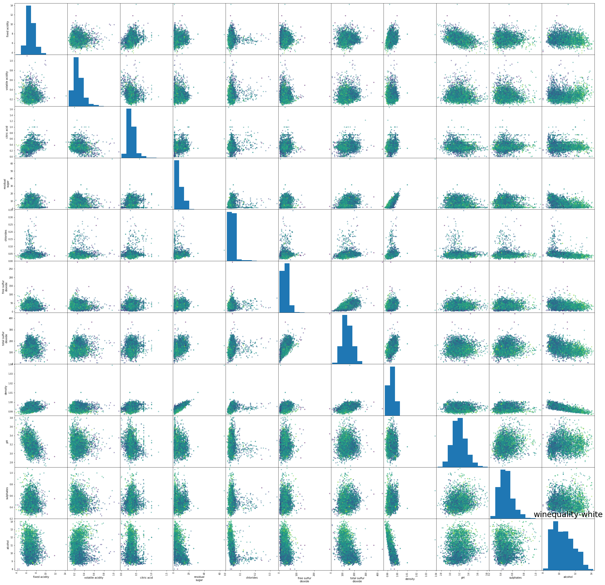


Figure 5.6

**5.2 The relationship between different variables**

Firstly, the correlation analysis between variables was 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.

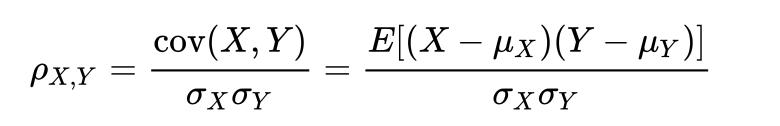
The Scatter diagram of correlation between variables:

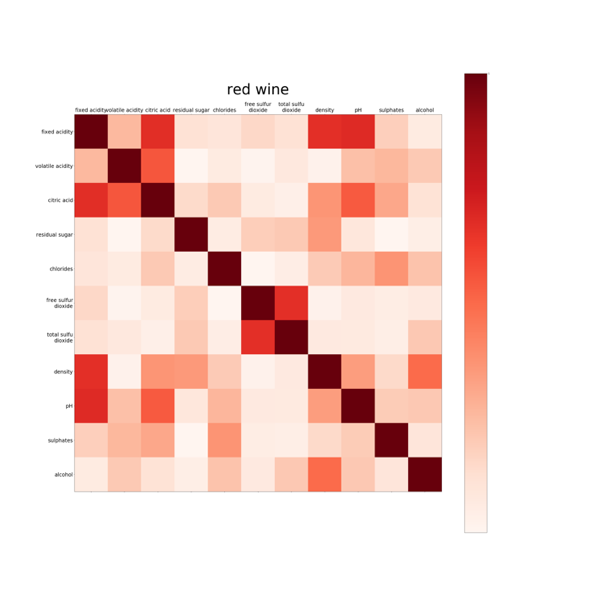
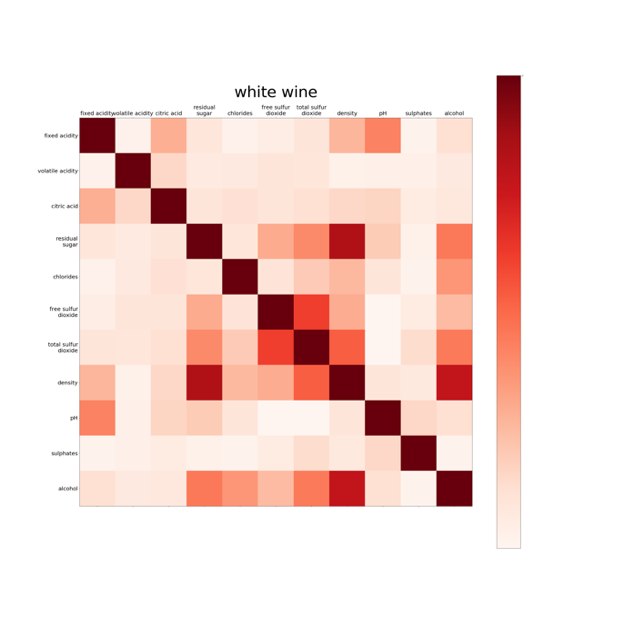
 

Red Wine White Wine

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 (0.67), PH (0.68) and density (0.67). And the correlation between free sulfur and total sulfur dioxide is strong (0.67). For white wine, we can see that the correlation between free sulfur dioxide and total sulfur dioxide (0.62) is stronger. Also, the correlation between residual sugar and density or density and alcohol are also closely related (0.84).

In order to illustrate the correlation between variables more accurately, we also use excel and Pandas DataFrame 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 and white wine DataFrame figure of attribute correlation:

DataFrame Lookup Tables:

For Red Wine:



For White Wine:



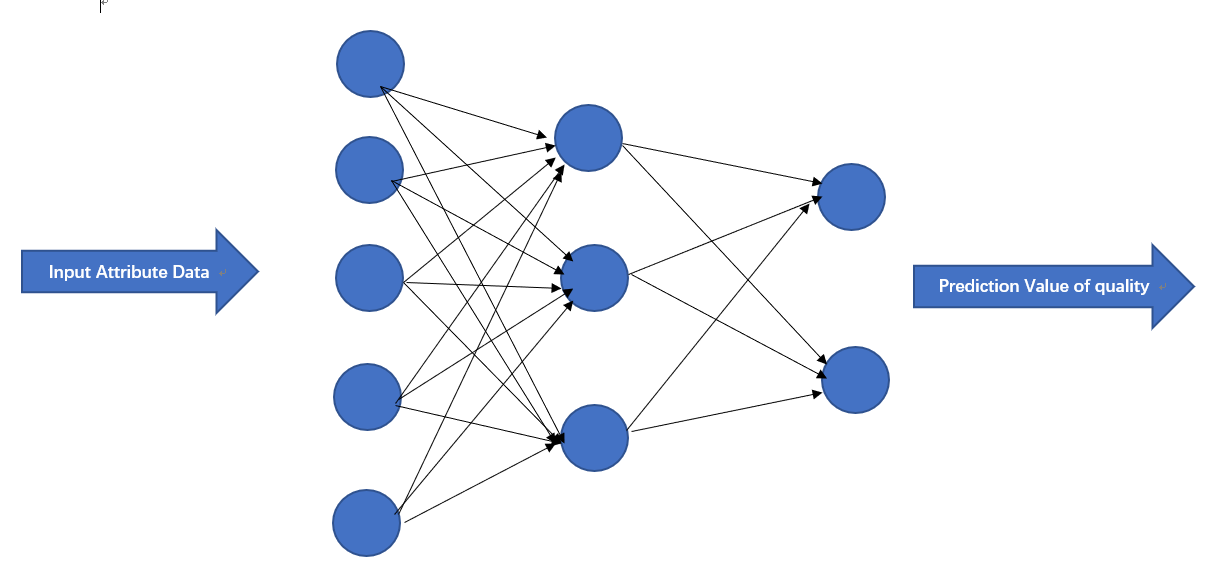
**5.3 Prediction**

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. **Four models:**
2. Neural Network:
   1. Design of Neural hidden layers:

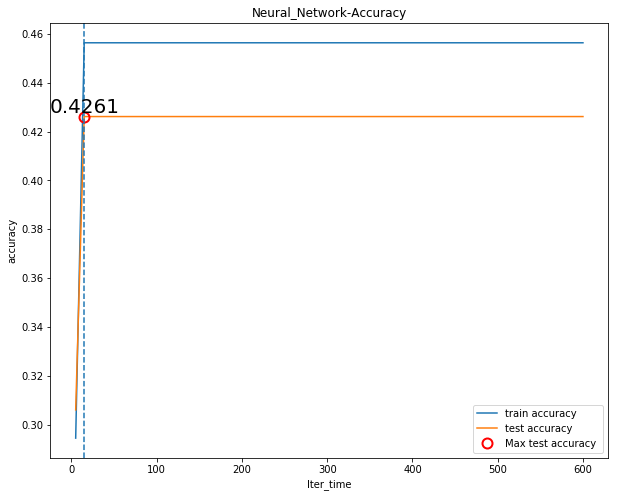
Every nodes use sigmoid function as activation function for Neural Network.



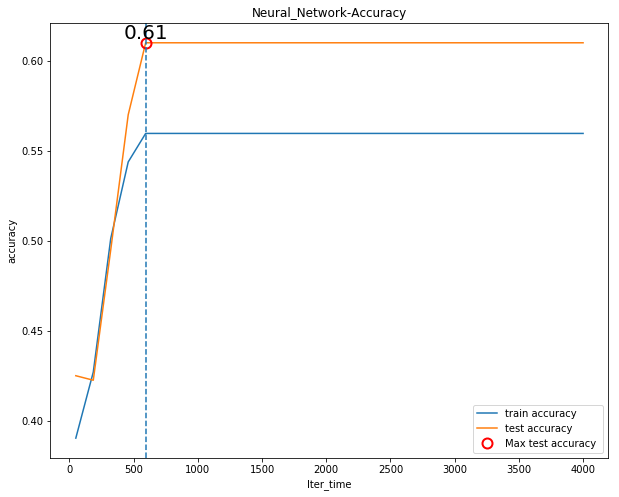
* 1. Train Accuracy, Test Accuracy relationship with iteration time:

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

* + - White Wine:



* + - Red Wine:

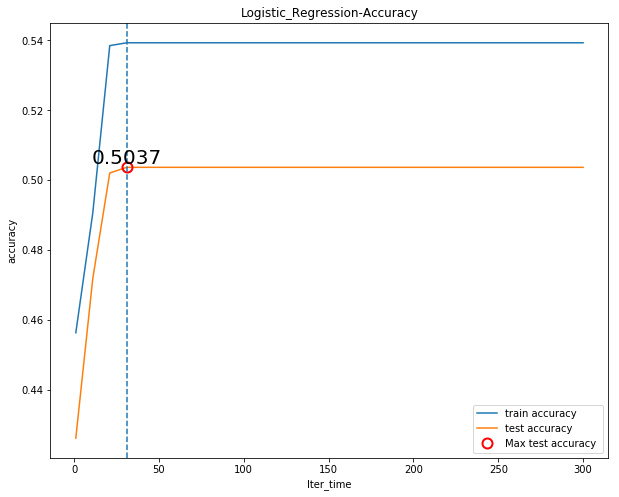


1. Logistic Regression:

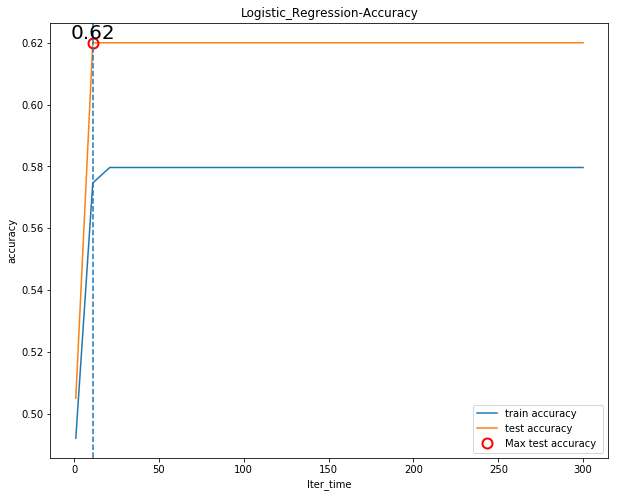
We visualize the accuracy of result, whose iteration time ranges from 0 to 300.

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.

* + - White Wine:



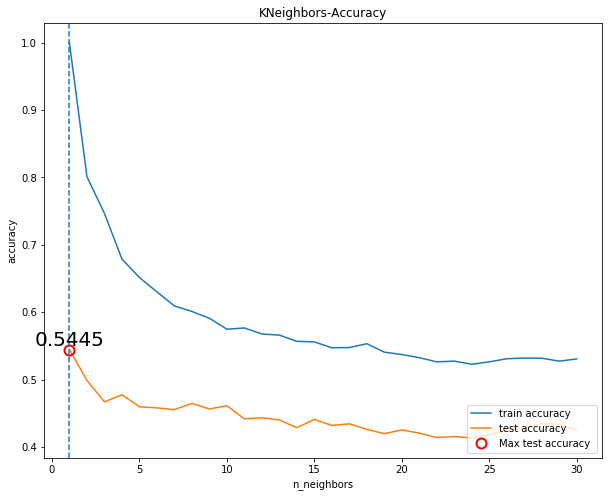
* + - Red Wine:



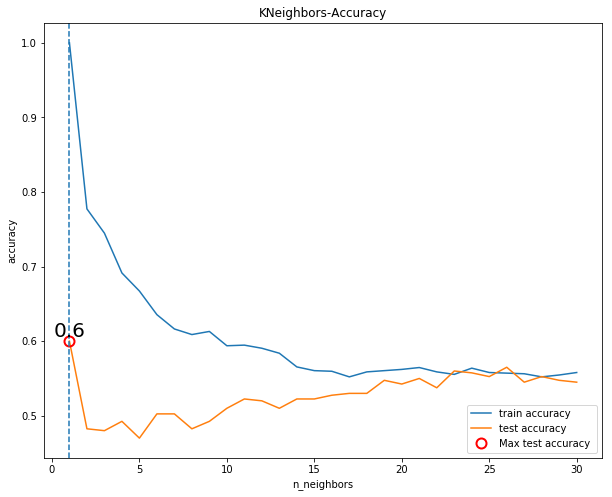
1. K-Nearest Neighbor classifier:

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 augment.

* + - White Wine:



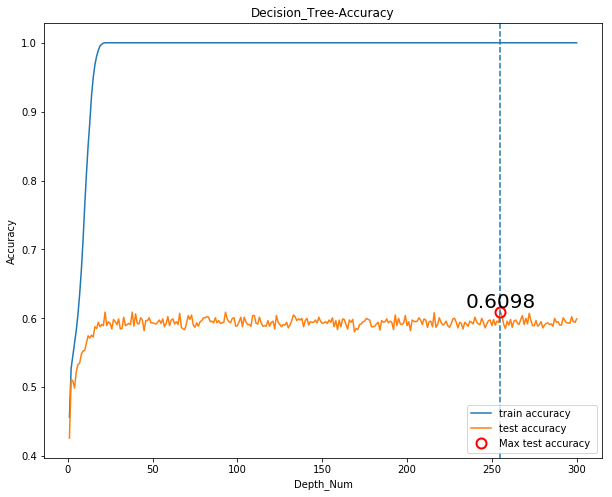
* + - Red Wine:



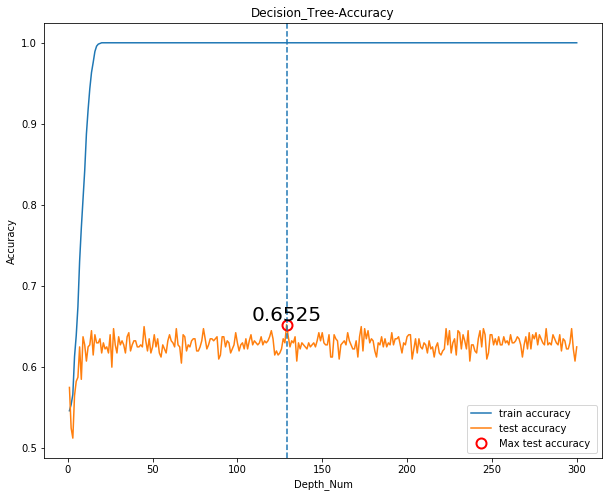
1. Decision Tree Classifier:

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.

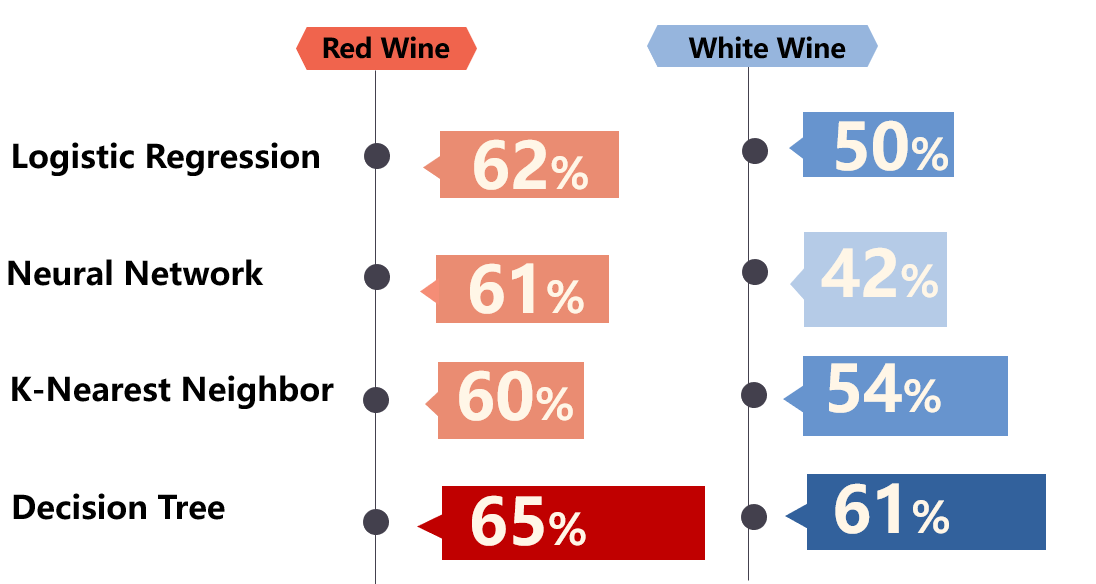
* + - White Wine:



* + - Red Wine:



1. **Prediction results:**

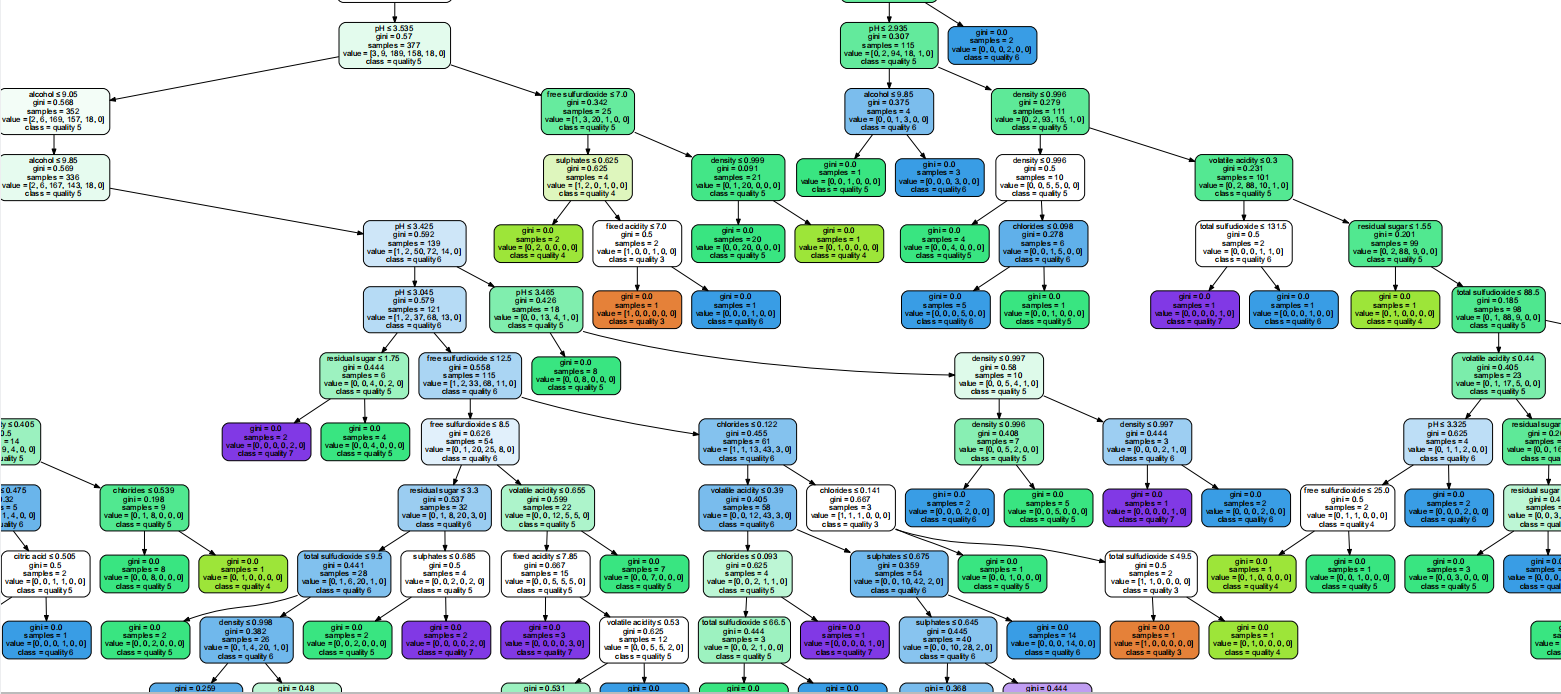
****

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.

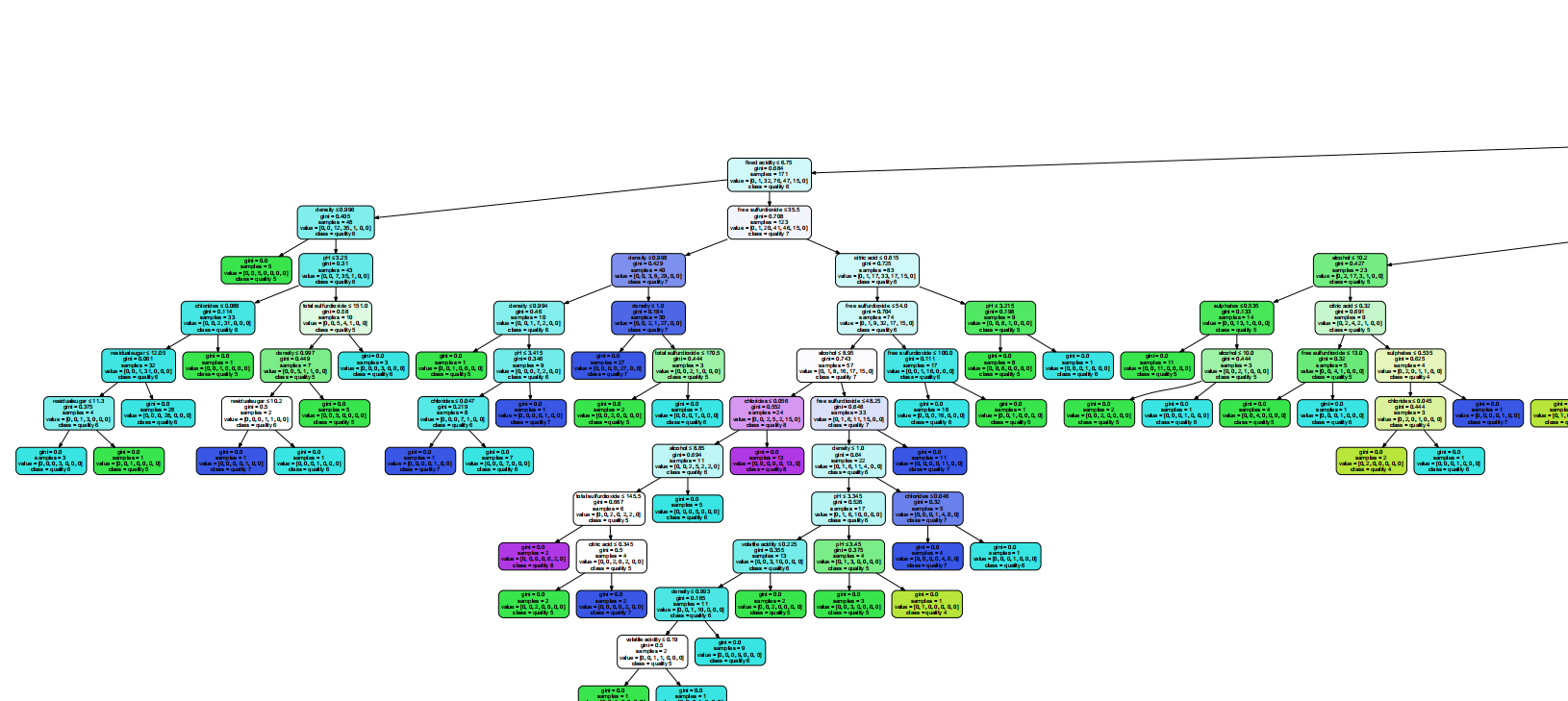
So, the decision tree model is the best model to predict the quality of white wine and red wine data. We use Graphviz - Graph Visualization Software to visualize the decision tree models.

Uncompleted of tree structures are shown below:

For Red Wine:



For White Wine:

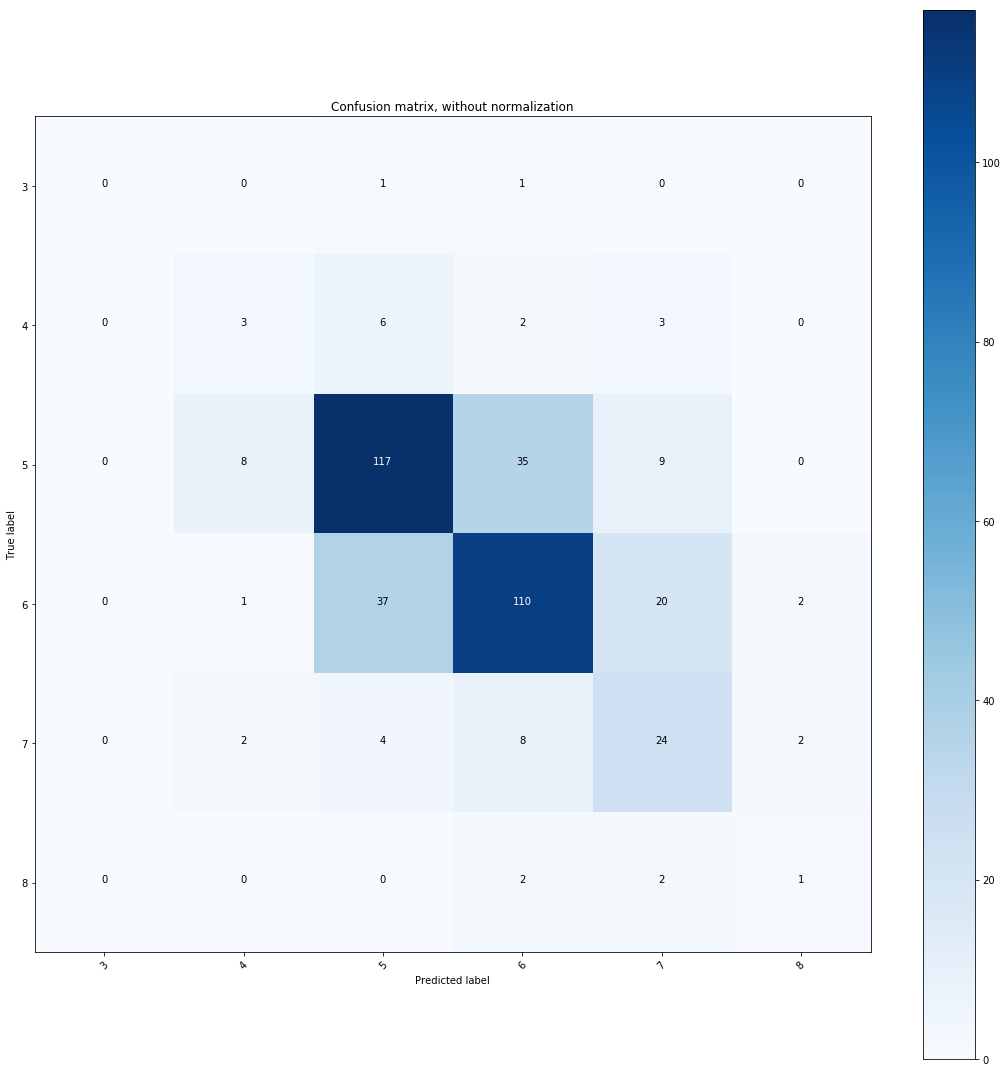
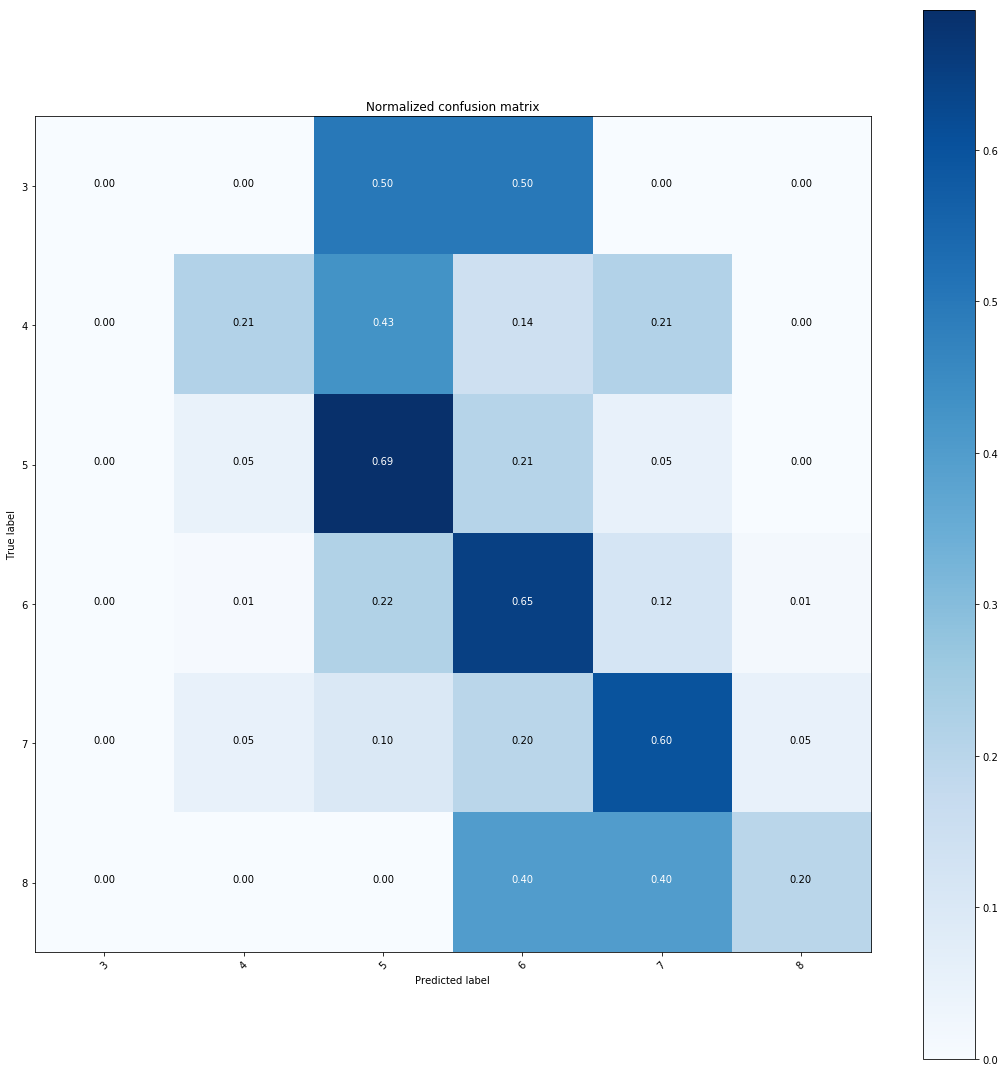


1. **Multi-Classes Evaluation and conclusion:**
2. **Multi-Classes Evaluation**

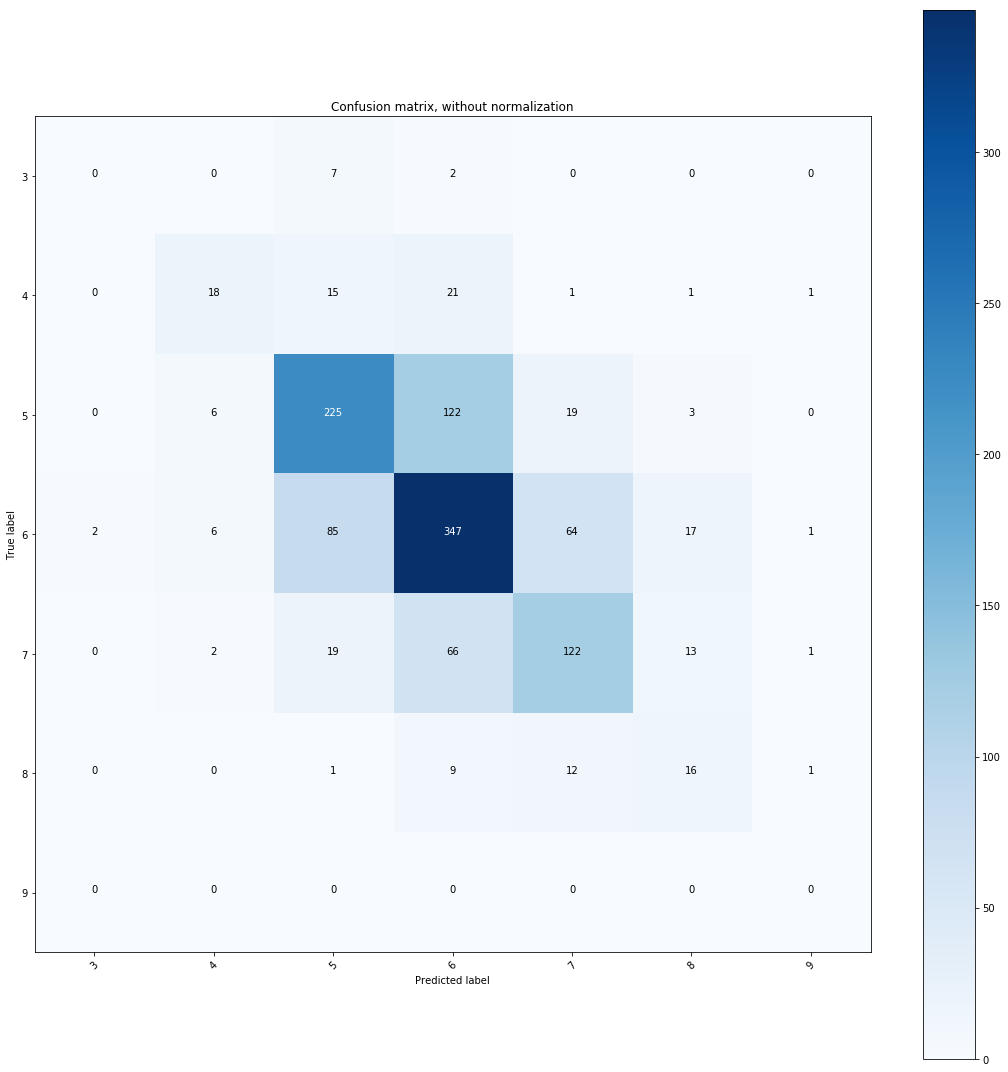
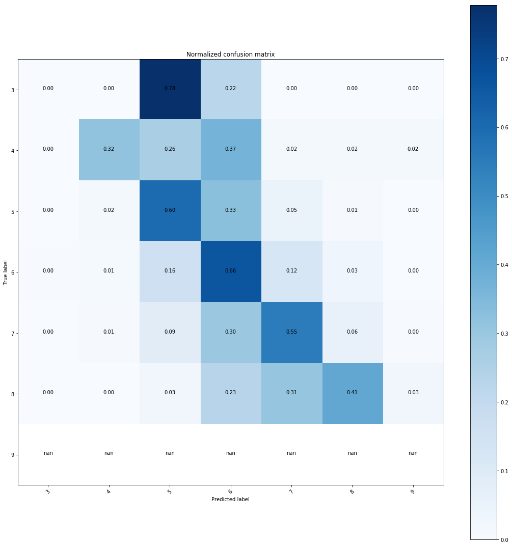
The decision tree is best-performance model, so we only generate Multi-Class Confusion Matrix of it to evaluate its effectiveness.

Multi-Class Confusion Matrix is an all-round method to evaluate the prediction of multi-Class. Because of imbalanced classes, we normalized the numbers of confusion matrix. The numbers and depth of color in the diagonal represent correct predictions.

**For red wine:**

**** 

**For white wine:**

**** 

1. **Conclusion and result**

From the Confusion Matrix color bar, we infer that the data is with imbalanced classes. The majority of quality is quality 5,6 for red wine, and quality 5,6,7 for white wine. The training data with target labels quality 3, 4, 7,8 for red wine, and quality 3, 4, 8,9 is insufficient for models to train. The decision tree is good for predicting the quality 5,6 for red wine, quality 5,6,7 but it is poor to predict other quality. To great extent, because of imbalanced classes and insufficient data, the prediction accuracy is limited by 65% for red wine and 61% for white wine.

In addition, some attributes are correlated to others. For example, in the red wine, fixed acidity is correlated with citric acid, so the acidity information may redundant for training and to some extent it is useless for prediction.

In the white wine, residual sugar has a very strong correlation to density with 0.84. The correlation between two variables values will cause that the number of variables cannot be fully used, and the highly correlated variable can be deleted because of redundant information. In this situation, actually this reduce the number of variables for prediction.

What’s more, the variables are less correlated with target label quality, the maximum one is only alcohol with 0.476 for red wine and 0.436 for white wine. This means that a few attributes cannot be used to predict the wine quality.

In summary, Because of imbalanced classes, highly correlation with variables, and low correlation with variables and quality, the predict model cannot achieve better performance in accuracy. So, our conclusion is that these given variables is not enough for making prediction and evaluation on the quality of the red/white wine.

To tackle with this problem, we can add some other variables highly correlated with quality, such as rank for wine taster, productive year. Meanwhile, we also need to input more data with low quality and high quality to make the classes balanced.

**Part 6: Future Work and Discussion**

* 1. **Try more prediction models:**

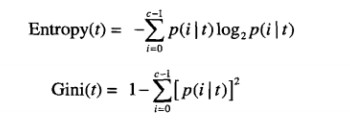
We know that the prediction accuracy is not high enough and there some adjusted ways to improve the prediction performance.

In out project, we use only four models to fit these data. It is highly possible that other machine learning models can do a better job to predict the wine quality, such Vector Support Machine, Radom Forest Models etc.

* 1. **Adjust parameters of models:**

A machine learning model has many relevant parameters to adjust its prediction algorithm. These adjusted parameters may help models to achieve a better performance.

Taking Decision Tree as an example, the default splitting criterion from scikit-learn is CART and it split subset of data by reducing its Gini.

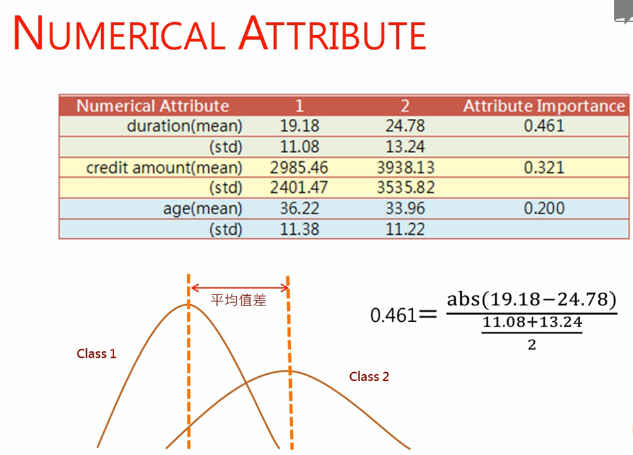


So, we can change splitting algorithm, such as ID3 and C4.5. It is possible that after change of algorithm the accuracy will be higher. Therefore, in the future, we can change parameters of each machine learning models to overserve these models’ performance.

* 1. **Try other way to calculate attribute influence for wine quality.**

We simply use Pearson correlation coefficient to measure the influence of attributes for wine quality and every two attributes.

**For example:**

****

**Influence = Abs (Mean Difference)/ (Mean of variety)**

In the different quality, the mean of each attribute is a good indication to discriminate the quality but it may be influenced by attributes’ variety because if the overlapping area of variety is too large, this attribute is difficult to be used to classify the quality.

* 1. **Possible multiple, nonlinear correlation：**

We just use Pearson correlation coefficient to calculate relationship between every two variables. In many cases, it is possible that multiple correlation may exist. For example, the sum of Attribute and Attribute B is positive correlative to Attribute C. In this case, our project does not find out this kind of relationship and it is an area of improvement for our project. In addition, Pearson correlation coefficient is good to find the linear correlation but bad for nonlinear correlation. Our project does not include complex technique to find out nonlinear correlation.

* 1. **Change target table to be real number：**

In our prediction models, we have an assumption that the value of quality is discrete, categories of different quality. We input quality attribute as multiclass labels. If we change data type to be real number and use regression models, the evaluation index will be different. In this way, we can change evaluation method to appraise our prediction outcomes, for example, calculating Mean of Square Error, or Mean Absolute of Error. Although, the classifier has low prediction accuracy, but the regression may have better performance and have low prediction errors in MSE or MAE

**Part 7: Author Contributions**

|  |  |
| --- | --- |
| Group Membership | Contribution |
| LIN JIEQIN  (Main Contributor) | Responsible for analytics visualization, prediction modes building and evaluation, and provide technical support and optimization for groupmates; Writing report and correcting some errors. |
| SU YING | Analyzing the relationship between quality and variables by writing python code;  Writing report with introducing four models, analyzing result and integrating report; |
| ZHENG YAWEN | Analyzing the relationship between different variables by writing python code;  Writing report with relating works, analyzing result. |
| QIN XIN | Writing report with motivation, what tool we used;  Data Description and Pre-Processing. |

**References**

[1] *Kamiński, B.; Jakubczyk, M.; Szufel, P. (2017)*. "A framework for sensitivity analysis of decision trees". Central European Journal of Operations Research. doi:10.1007/s10100-017-0479-6

[2] *real statistics using excel. (n.d.)*. Retrieved May 1, 2018, from: <http://www.real-statistics.com/correlation/basic-concepts-correlation/>

[3] *CART决策树的sklearn实现及其GraphViz可视化*. Retrieved May 2, 2018, from: <https://blog.csdn.net/chai_zheng/article/details/78226556>

[4] *3.3. Model evaluation: quantifying the quality of predictions*. Retrieved May 2, 2018, from: <http://scikit-learn.org/stable/modules/model_evaluation.html>