IST707

Qingqing Hu

Yafei Lou

Project proposal presentation guidelines

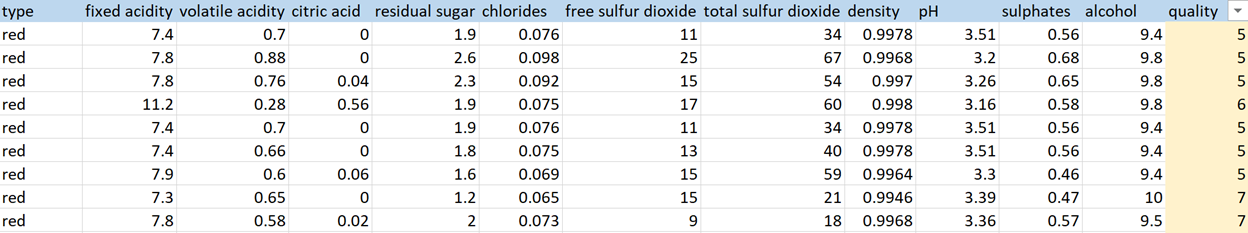
Since traditionally the standards of performance that wine experts use to judge wine quality include the following: sweetness, acidity, tannin, and alcohol, people usually taste the wine to see if it is balanced when nothing sticks out, like harsh tannin or too much sweetness. Therefore, wine quality is easier to detect than define, because the quality is defined subjectively and influenced by extrinsic factors. Nonetheless, novices, unlike experts, cannot easily determine what great characteristics a wine should possess but still pay for a high price blindly. What is not certain is if this is a complex example of the blind leading the blind. However, if we know what influences and signifies wine quality, we will be in a better position to make good purchases. Then, our appreciation for wines will deepen once we are familiar with wine quality levels and how wines vary in taste from region to region.

In this study, the model will predict wine’s quality score between 0 and 10, given the attributes such as acidity, residual sugar, chlorides, density, pH and alcohol. As a result, we could apply the model to impartially evaluate the wine quality using the scientific method.

The data mining task will focus on classification mainly. By doing the classification task,we can map wines that contain certain characteristics into higher or lower score ranges. In addition, we will also use a clustering algorithm to determine what factors influence the score the most and possibly find the pattern among the data points.

In this project, we will use a data set downloaded from the UCI machine learning library. This data set is about the red variety of Portuguese "Vinho Verde" wine. Due to privacy and logic issues, only physicochemical (input) and sensory (output) variables are available (for example, no data on grape type, wine brand, wine selling price, etc.). The dataset consists of 1599 rows and 12 Input variables (based on physicochemical tests) and one Output variable (based on sensory data). Input variables include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol. quality (score between 0 and 10) is the Output variable.

Here is an example of the data, as shown below:



Based on this data set, we will use two models, classification and clustering, to analyze the factors that mostly affect the quality of wine. First, we divided the wine into two classes, Good (7~8) with 217 rows and Bad (3~6) with 1382 rows, we set the quality column (score range 3~8) as the target variable. In the classification task, we will analyze and compare the mean distribution of the different attributes of the two grades to find the factors that have the most influence on the quality of wine. Taking pH as an example, we plot the pH distribution for both good and bad levels. Based on this, see if the two grades have a significantly different distribution in pH, that is, whether the pH value significantly affects the quality of the wine grade classification. For other variables, we used the same method for comparison and analysis. In the association rule model, we will take the two levels of wine quality as RHS, and take all the remaining 10 attributes as the analysis factors, LHS. Through the comparison of support, confidence and lift, we can find the key factors to classify the wine grade.

Finally, we will construct a Confusion Matrix to calculate Accuracy, Precision, Sensitivity, Specificity and Score. According to the calculated value, the quality of the model is tested and evaluated.