Capstone Report- White Wine Quality Analysis

Gina Cavill

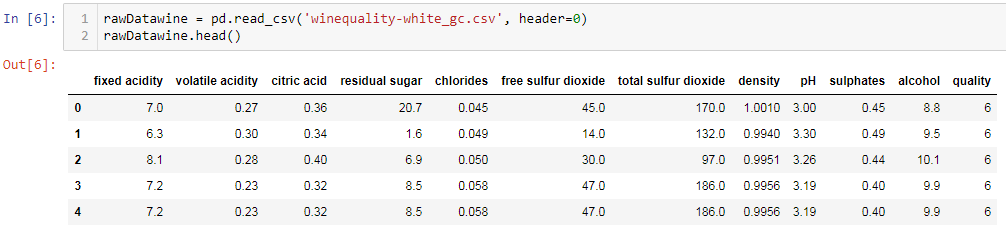
A research data set from the UCI research website was selected (See footnote) related to wine quality. “The inputs include objective tests (e.g. PH values) and the output is based on sensory data

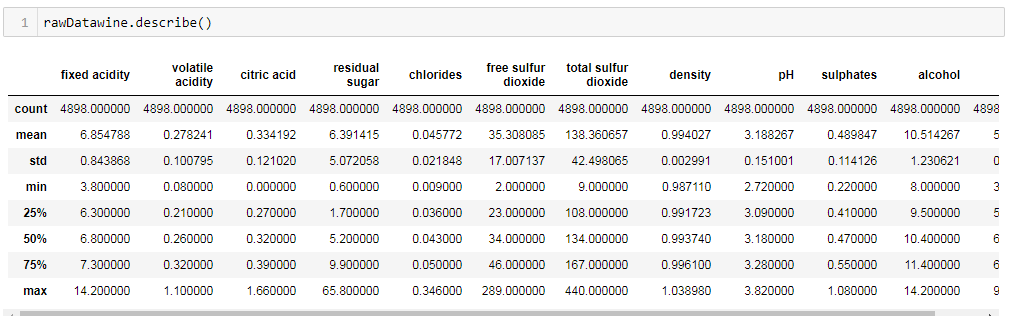
(median of at least 3 evaluations made by wine experts). Each expert graded the wine quality

between 0 (very bad) and 10 (very excellent). The goal of this analysis is to determine if any of the objective tests have any correlation to the quality grade of the white wine dataset. Which modeling approach would be deemed most effective, regression or classification? Which inputs if any had any impact on the output variable “quality” score? What is the confidence level of the best model predictions to be used for determining the best quality wines?

**Data preparation and cleaning analysis**

Below is information regarding the white wine dataset input and output variables. The data was prepared well with no non-null attributes and all input datatypes were float numbers with the output variable being an integer datatype. No major preparation or cleaning was needed prior to performing the analysis.

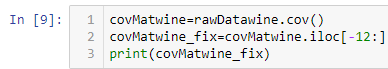


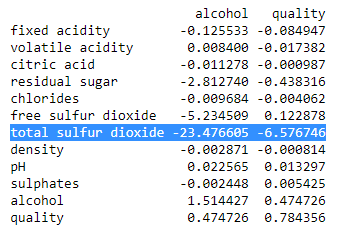


**Analyzing correlation of attributes of the dataset**

Several techniques were utilized to measure the correlation or impact of different attributes with the output variable “quality”.

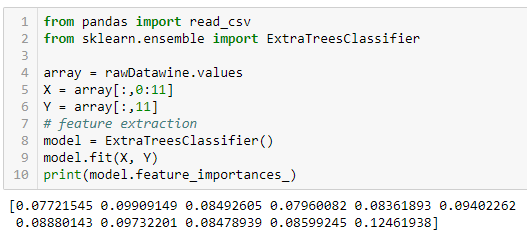
First, the covariance function was run against the dataset to measure influence of each attribute has against the y variable. See below results.





Based on the covariance function results, it appears the total sulfur dioxide test attribute has by far the largest relationship with the quality variable in an inverse way.

The second approach to analyze correlation was using the ExtraTreesClassifier model to determine feature selection importance. See below results.

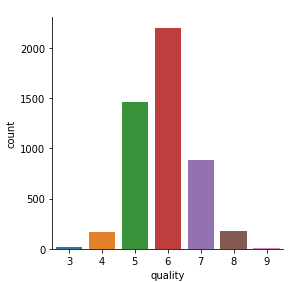


Based on the feature importance model run, the last attribute “alcohol” had the highest score(0.12461938), with all the other variables being within a similar range of each other in terms of score importance.

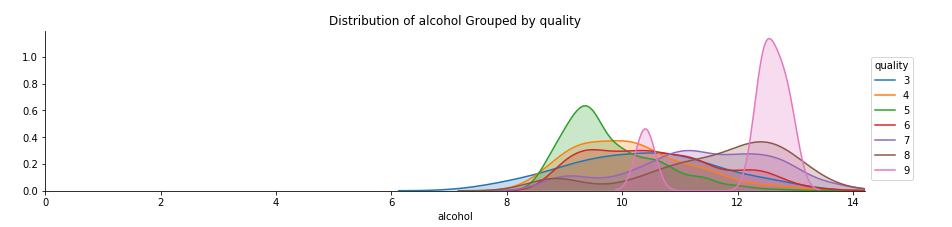
**EDA (Exploratory Data Analysis)**

Along with running models to determine feature importance, several functions were run to visualize the dataset and to see if any of them assisted in understanding the attribute relationships better through EDA.

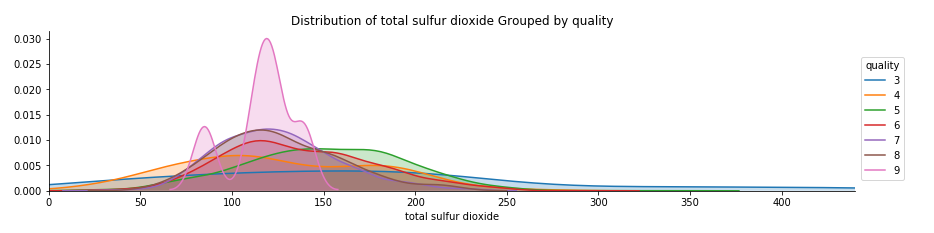
Y Variable “quality” scores- most wine quality scores were in the 5-6 range with the 0-10 possible range.



Alcohol was determined by feature selection analysis to be an important attribute. Below is a graph showing the alcohol levels as relation to quality. The higher quality wine had alcohol levels between the 12 and 13 levels. The purple spike displays the highest quality.

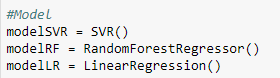


Total sulfur dioxide was the other attribute which showed a high level of correlation with the quality score. Below is a graph showing the relationship between total sulfur dioxide test numbers and the corresponding quality score. The higher quality scores were in between 100 and 125, with the quality dropping quickly once the numbers go above the 125 test result number.



**Regression model evaluation**

The white wine was evaluated by two approaches regression and classification modeling. Below regression models were selected to determine effectiveness of prediction first. All features were utilized in the below models, none were removed for the following regression results.



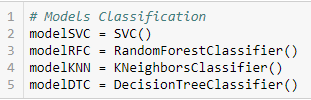
***SVR model results*** ***RF Regressor results Linear Regression results***

The most effective regression model tested was the Random Forest Regressor model with an R Squared of 0.432. Cross validation and fitting functions were utilized during prediction analysis. Please refer to Jupyter file for more details. The regression model resulted in not having the most confident of prediction models. Classification model analysis was required to explore a different approach.

**Classfication model evaluation**

Below are the classification models which were selected to determine effectiveness of this approach with the white wine dataset. All features were utilized in the below models, none were removed for the following classification results.



***SVC Results RF Classifier KNeighbors DecisionTree***

**Accuracy:** 56.215% 65.801% 46.506% 58.194%

Cross validation and fitting functions were used on the above models and based on the accuracy scores RF Classifier was the strongest prediction classification model.

**Conclusion**

The following questions were asked at the beginning of the analysis. The analysis assisted with the following answers.

**Which modeling approach would be deemed most effective, regression or classification?**

Based on the R squared results of the regression models and the accuracy scores of the classification models, the best predictive model to utilized would be the RandomForestClassifier() with the highest accuracy score of 65.801%.

**Which inputs if any had any impact on the output variable “quality” score?**

Based on the covariance function and ExtraTreesClassifier model, two input variable were determined to affect the output variable the most (Alcohol and Total Sulfur Dioxide).

**What is the confidence level of the best model predictions to be used for determining the best quality wines?**

The RandomForestClassifier model with the white wine dataset was the best model, but the accuracy score of 65% is a mediocre score and preferably the prediction model would have a higher accuracy rate. However, the models are only as good as the dataset provided and the results give a reasonable confidence of prediction of what the quality score would be based on the input variable test results.

Footnote: (White Wine dataset from Sources

Created by: Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009)