**Final Project Report**

**Red Wine Quality Analysis**

**University of North Texas**

**Info 5082: Seminar in Research and Research Methodology**

**Dr. Junhua Ding**

**Manasa Athota**

**11374793**

**April 26, 2021**

**RED WINE QUALITY ANALYSIS**

**Introduction and Statement of the Problem:**

The Quality of the red wine is measured on basis of few metrics such as the PH, Critic acid, Acidity Levels. These data set includes the various red wines of the “Vinho Verde” wine of Portuguese. We look at a series of findings on a variety of red wine types, including chemical properties and taster rankings. As casual drinking becomes more common, the wine industry has experienced a recent growth spurt. The price of wine is determined by a very subjective definition of wine appreciation by wine tasters, whose opinions which differ significantly. Physicochemical studies, which are laboratory-based and consider factors like acidity, pH level, the presence of sugar, and other chemical properties, are another important component in wine certification and quality evaluation. For the wine industry, it would be interesting if human taste quality could be linked to chemical properties of wine, allowing for certification and quality assurance. In this Report I have performed various methods and see which method does a good decision.

**Review of Literature:**

In [3,] Y. Subba Reddy et al. proposed a user-centric similarity system in which product similarity is determined by user tastes. The aim of creating such a scheme is to assist and guide wine consumers in making smarter choices and winemakers in producing higher-quality wines. It provides a critical analysis of current research developments in the field of wine and can assess the quality of wine depending on these consumer interest categories. The findings of the experiment will help product shoppers make smarter decisions.

Data mining technology for wine research was invented by Dimitrija Angelkovet al. [4]. Data mining is a method of collecting valuable and creative information from data that is both interactive and interactive. Data collection is a step in the scientific process. They need a large base of measurement in order to accurately determine the quality of wine. The key benefit of this low-cost prototype is that it can be used by small winemakers for grape fermentation control and tracking. The proposed system was tested in a winery in the Tikves area, and it met the initial wine production stage expectations. It provides a manufacturing market for the manufacture of high-quality wine.

The aim of the analysis was to find an outlier or exception in the sample wine collection in order to detect wine adulteration. In comparison to red wine, white wine is more susceptible to physio-chemical modifications. The consistency of red wine is influenced by less factors than that of white wine. The thesis also discovered that two characteristics, alcohol, and volatile acidity, have a significant impact on wine production. The key goal of this research was to use physicochemical data to forecast wine quality. One recommendation is for a higher degree of distinction between white and red wine production lines, and another is for winemakers to insist on ensuring a suitable alcohol content. The study concludes that analysis would provide opportunities for corrective action to be taken in order to improve quality.

**Objectives of the Study:**

The main objective of the project is to predict the Quality ranking from the chemical properties of the wines. Generally, most of the research figured out that which attribute contributes major for finding the red wine quality. Which predictive model is best fit to describe the quality by categorizing the data into the good and bad wine. Which model best fit for classification can be decided.

Techniques used in this Project for predicting is:

1. Logistic Regression
2. KNeighborsClassifiers
3. SVM
4. Decision Tree classifiers
5. Random Forest Classification

**Data Collection:**

This data has been collected from the Kaggle and the UCI machine learning repository.

Data set Link: <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>

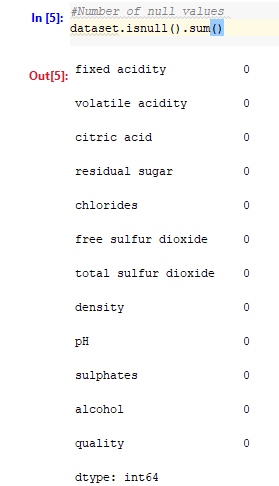
<https://archive.ics.uci.edu/ml/datasets/wine+quality>

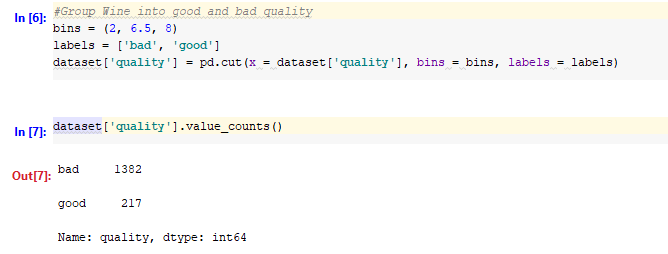
The Red wine data consists of 12 columns each variable indicate the properties of the wine. Each attribute has its own importance. The below table shows the description about the attributes.

|  |  |
| --- | --- |
| Fixed Acidity | Most acids involved with wine or fixed or non-volatile. |
| Volatile Acidity | The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste. |
| Citric Acid | Found in small quantities, citric acid can add 'freshness' and flavour to wines. |
| Residual sugar | The amount of sugar remaining after fermentation stops, it is rare to find wines with less than 1 gram/lite. |
| Chlorides | The amount of salt in the wine. |
| Free sulphur dioxide | The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfited ion. |
| Total Sulphur dioxide | Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 |
| Density | The density of water is close to that of water depending on the percent alcohol and sugar content. |
| PH | Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14. |
| Sulphates | A wine additive which can contribute to sulphur dioxide gas (S02) levels, which acts as an antimicrobial. |
| Alcohol | The percent alcohol content of the wine |
| Quality | Output variable (based on sensory data, score between 0 and 10) |

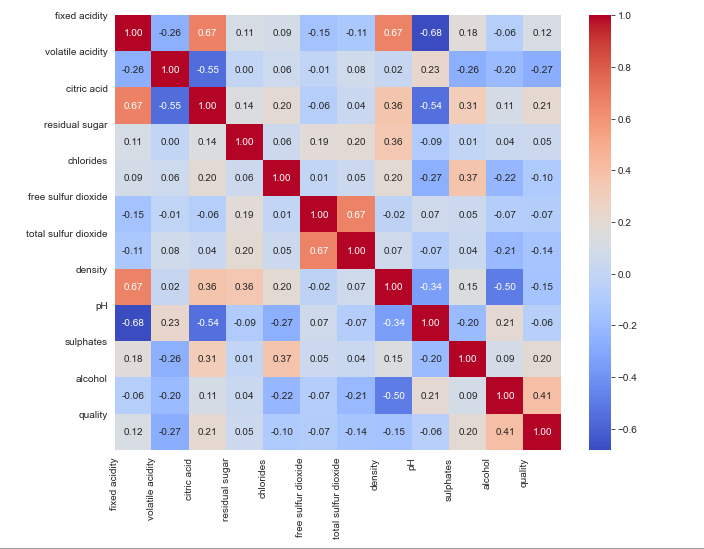
**Exploratory data analysis (EDA) and Hypotheses for the Study:**

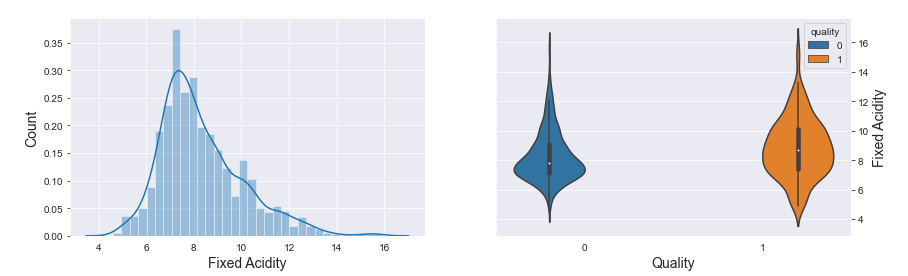
The dataset has been collected from the original sources and there is no Missing Value and Null Values in this EDA process I represented the correlation value between the attributes. Grouped the data into the good and bad based on the attribute Quality which is a scaled between 0-8. The quality scaled 6.5 and above is categorized into the good quality wine and below 6.5 is categorized into the bad quality wine. Total 1382 are in bad quality and 217 are good quality.



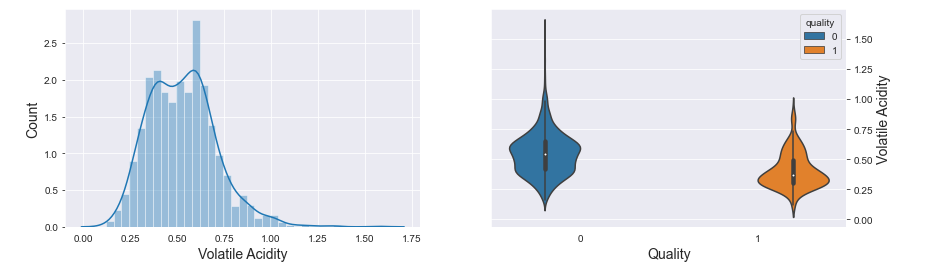


Correlation between the variables:

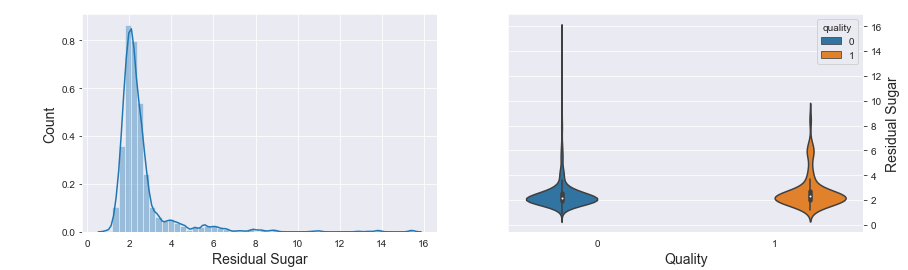
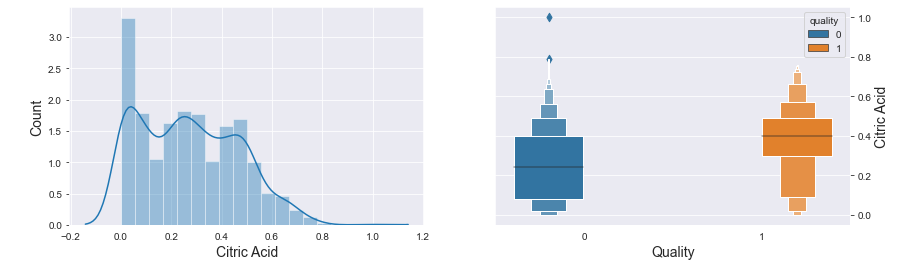


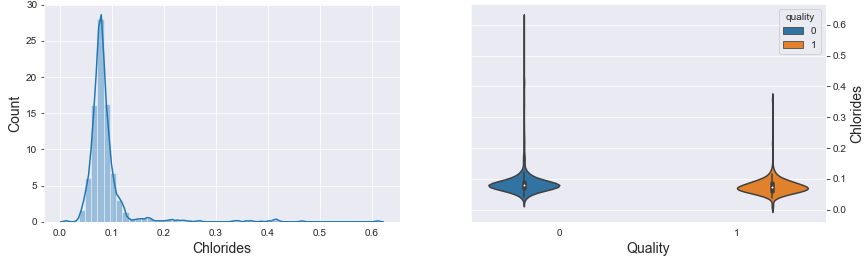
Each and every variable is shown with respective to the quality how the specific property varies and what value of it can be considered as a good quality. The Fixed Acidity of a wine can be said as a bad quality when the Acidity level is between 6 and 8.

The below graph histogram shows most of the data lies In which range of the volatile Acidity and from the whole data set 20 different wines are with the volatile Acidity of the around 0.60 and the wine with range of 0.40 and 0.70 volatile acidity is considered as a bad quality.

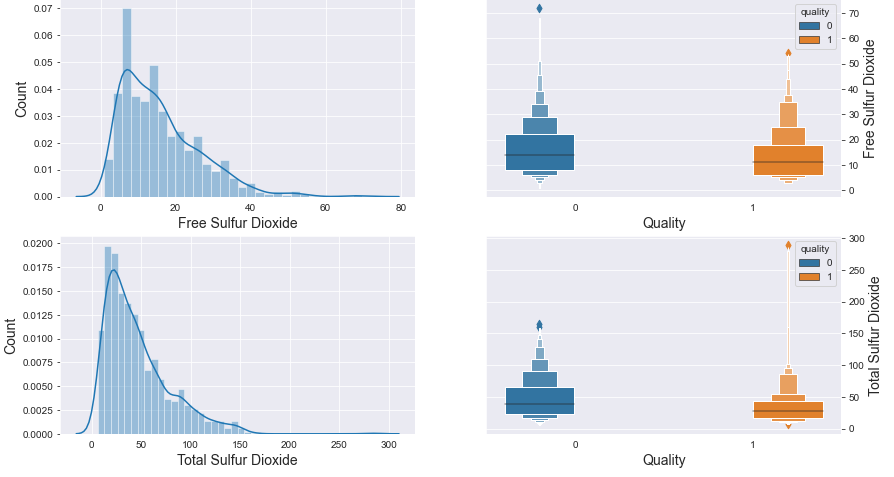


The data set is having more 18 different wines with citric acid with 0.0 and the citric acid with range of 0.1 to 0.4 is considered as a bad quality. We cannot decide on basis of residual sugar as they just vary in small quantity.

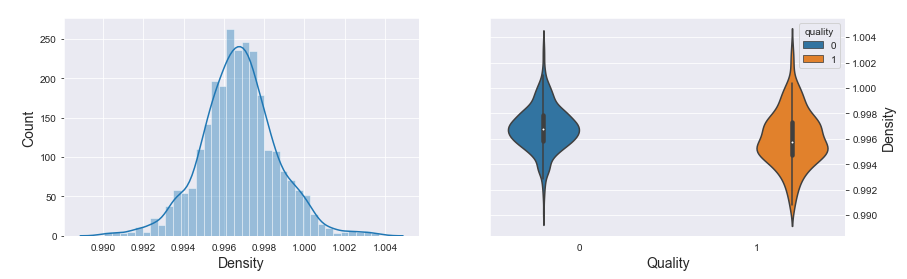




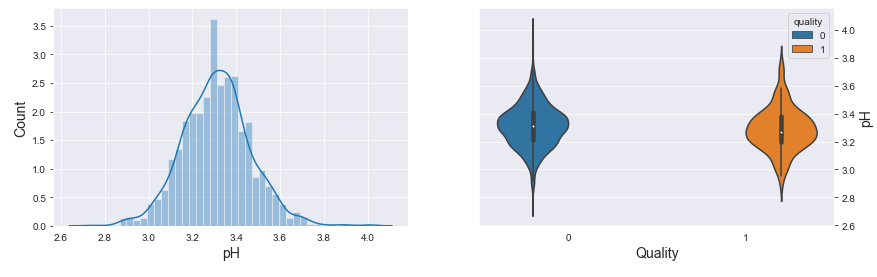
Most of the wines are maintaining the chlorides in the range of 0.0 to 0.1 we cannot decide a wine is good or bad on basis of chloride as both the good quality and bad quality for chlorides does not vary. With the free sulfur dioxide values, we can say a bad quality is between the range of 9-21 and in the data set most of the wine is having the free sulfur dioxide with the value of 5. The total sulfur dioxide with range of 25 to 49 is said to be a good quality.

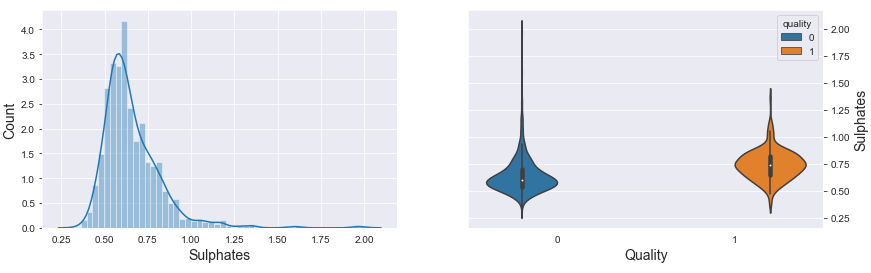


The density of the wine with the range of 0.994 to 0.998 as a good quality and bad quality is in the range like good the value between the range of 0.995 to 0.997 as the bad quality.

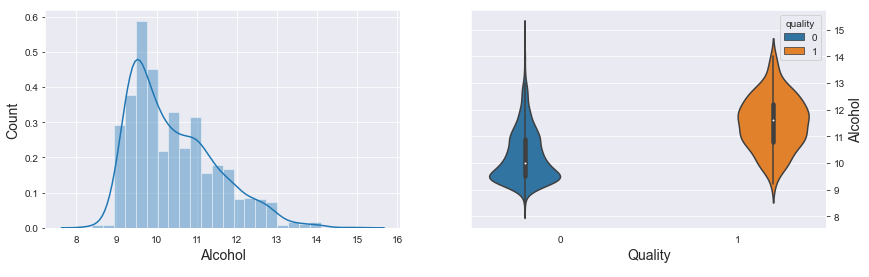


The PH of the wine in the dataset is mostly with the 3.3, Most of the good quality wine is between the range of 3.2 to 3.4.



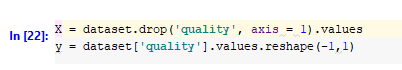
The sulphates of the range of 0.60 to 0.80 as a good Quality and the bad 0.55 to 0. 65 vales as a bad qulity wine and in this dataset most of the sulphates has a sulphates of 0.60 . 

The Alchol quantity in the wine can define the wine is good or bad the range of 10 to 13 is said to be a good quality and most of the wines in this dataset is having the alchol with the value of 9.5 .

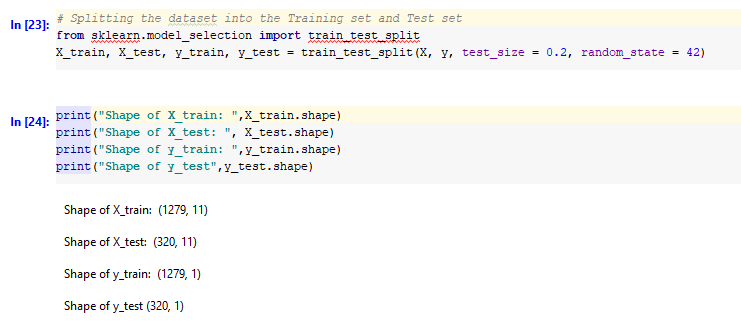


**Data Analytics:**

The Dataset includes an attribute Quality which is the main factor we wanted to find out based on other factors. But if the attribute Quality is already present in the dataset through which the models can predict easy. So, dropping the attribute from the table and storing the Quality attribute in another table.

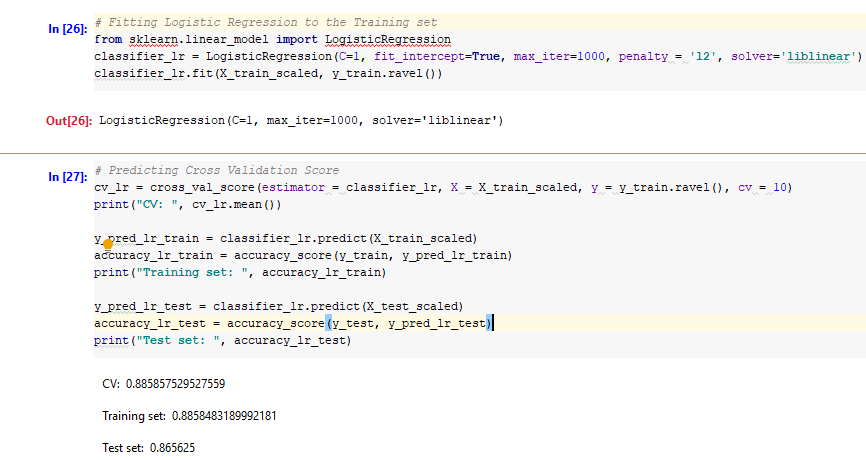


To predict the analysis the data set has been split into the training and testing for both the tables which contains whole data set except the Quality and another table which consists only the Quality variable. Which is been used for classifying the data and predicting the analysis.

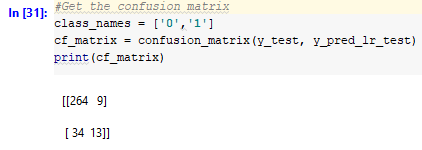


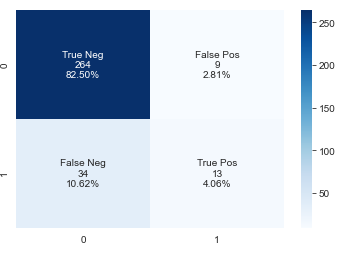
**Data Visualization and Results Report:**

1. **Logistic Regression:**  The logistic regression is performed on the nominal variable i.e., the quality and the other independent variables. It shows the accuracy for cross validation is 88.5%, training set is 88.5% and testing set is 86.5%. Here the logistic regression accuracy is above the 86%.

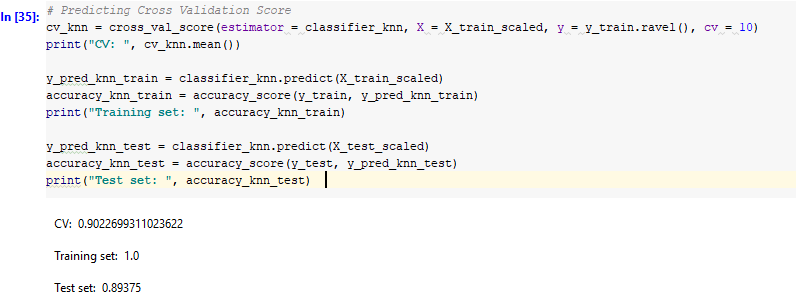


The confusion matrix for the logistic regression is shown below it shows that 82.5% chances for predicting the bad red wine quality.

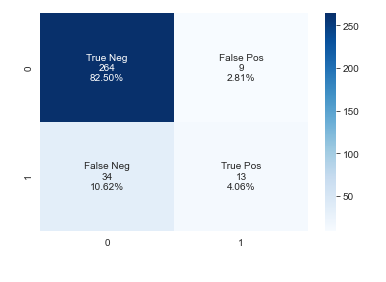




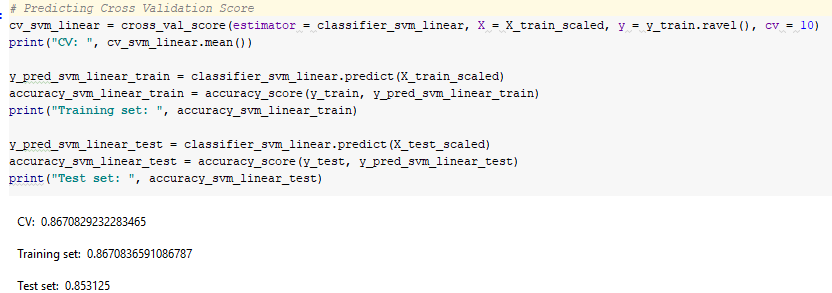
**2.KNeighnhoursClassifiers:** In this classification considering the n-neighbours as 32. The accuracy of the KNN with cross validation is 90.2% and for the training and testing datasets are 100% and 89%. Comparing with the logistic regression the KNN Classifiers have a very good accuracy.



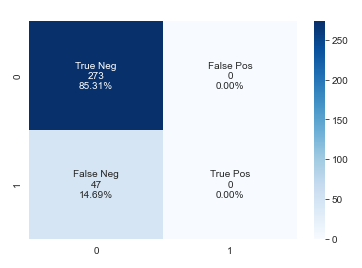
The confusion matrix of the KNN shows similar results as logistic regression it showed 82.5% for prediction of the bad wine quality.



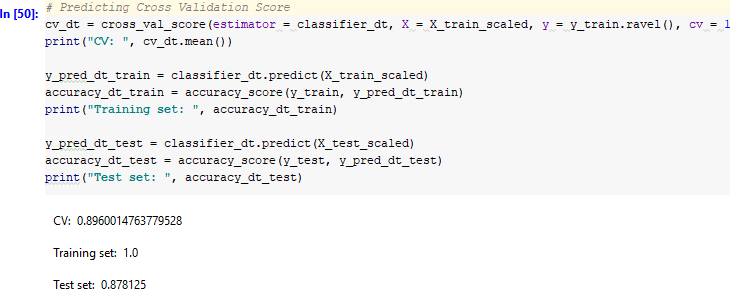
**3.Support vector Classifiers (SVC):** The SVC is the machine learning concept used to predict the best fit hyperplane which divides the data into the categories. In this model for the red wine data set the accuracy is less compared with the logistic regression and the KNN classifiers. SVC came up with the accuracy of 86% approximately in the cross validation and training sets and the accuracy for the testing dataset portion showed is 85%.



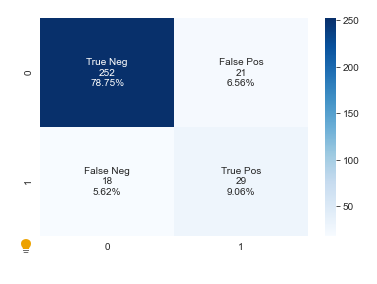
The confusion matrix shows it have a high percentage rate to show the bad wine data set prediction rate is high with 85.3%. But the True Positive nature of the dataset is shown as 0% which is unlike to choose this model as the best as it shows 0% for both the True Positive and False Positive.

****

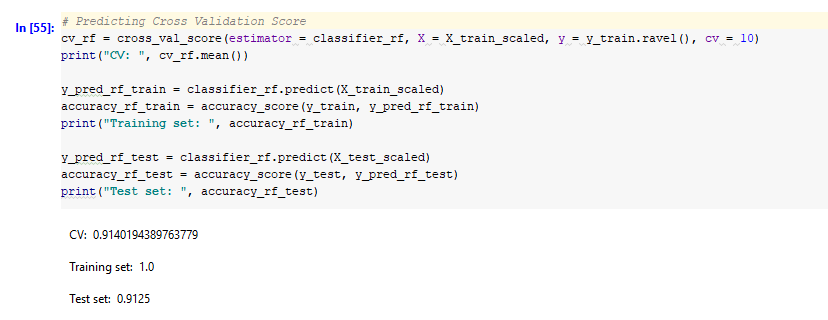
**4.Decision Tree Classifiers:**  The Decision Tree classifiers has been set with the random state of 33 and the leaf nodes of the 400. The accuracy shown with this model 89% for cross validation, for training set 100% and for testing set came up with 87%.



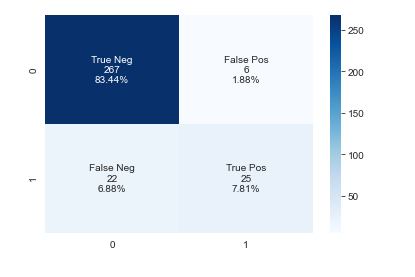
The confusion matrix shows the true negative percentage as 78.5%, False Positive as 6.5%, False Negative as 5.6% and the true positive Percentage as 9.0%.



**5.Random Forest Classifiers**: The random forest classifiers operates by constructing a multiple decision tree at the time and for prediction of individual trees. The data set is given for this classification with estimators of 800 and random state of 33. The accuracy percentage of cross validation and the testing set is shown as 91%, the accuracy for training is shown as 100% in this model.

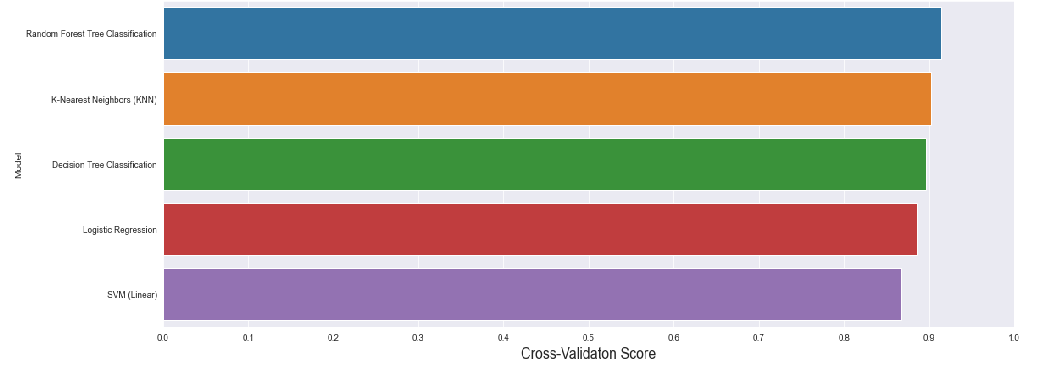


The confusion matrix of the Random forest classifiers is shown as true negative with 83.4% , false positive of 1.8%, false negative with 6.8%, True Positive 7.8%.

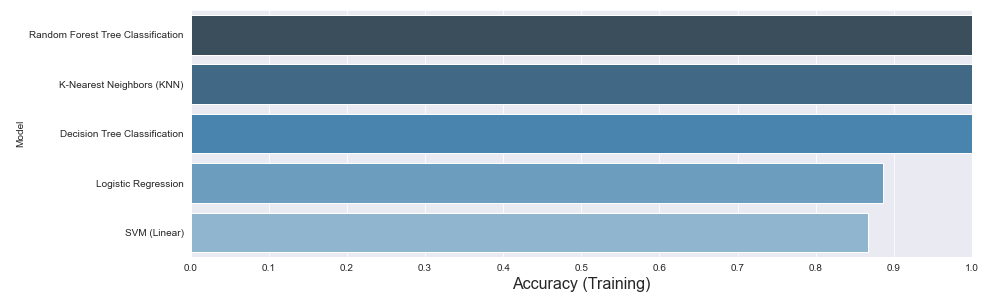


**Conclusion:**

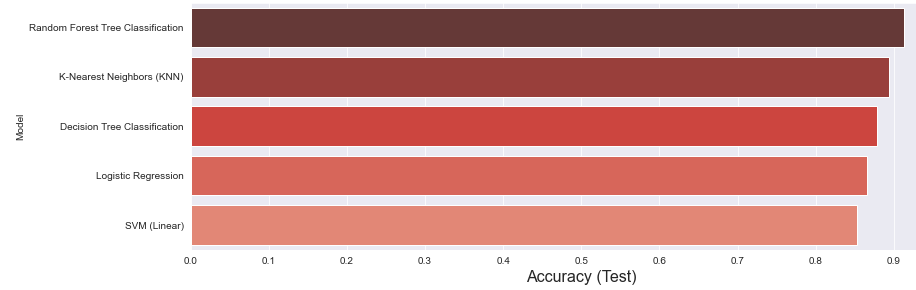
Comparing all the models on the basis of the cross-validation results of the models from the below picture we can see Random Forest Tree classifiers. KNN and decision tree have a similar result and close to random forest tree classifiers. Random Forest Tree classifier gives the best result when we consider the Cross validation.



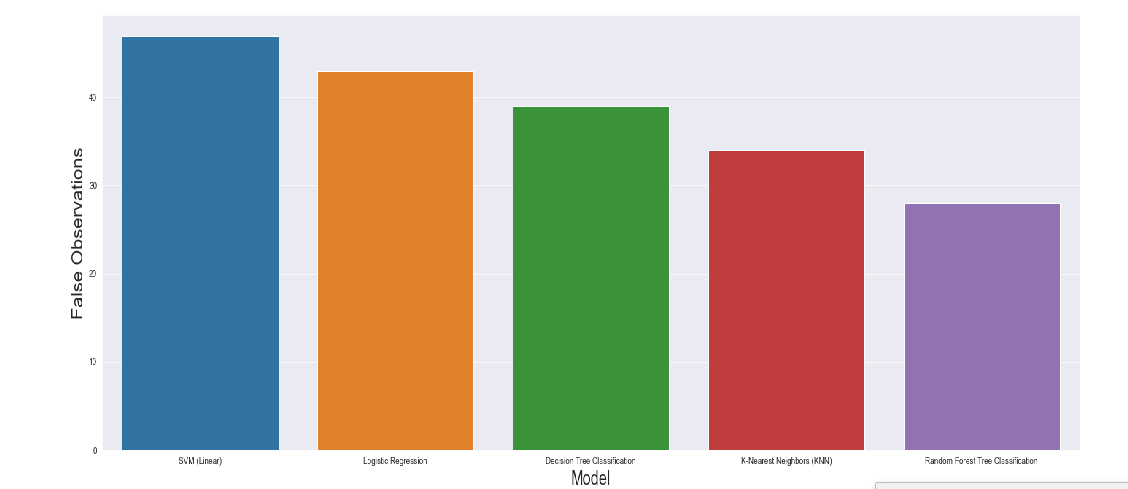
On basis of accuracy, we cannot say one model as the best as tree models showed the same results. Random Forest tree classifiers, KNN, Decision Tree Classification showed 100% as the accuracy results.



Accuracy of the testing data set shows random forest tree classification as the best model as the Random Forest Tree classification with accuracy of 90% approximately.



Form the above three comparing we can say Random Forest Tree Classification suits the best for predicting the quality of the dataset. But we need to consider the prediction of the false observation can be done best through which model. which can be calculated by summing up false positive and false negative values from the confusion matrix. Comparing all the models SVM model is best for predicting the False observation.



**Bibliography:**

Aich, S., Al-Absi, A. A., Lee Hui, K., & Sain, M. (2019). Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques. *2019 21st International Conference on Advanced Communication Technology (ICACT)*. <https://doi.org/10.23919/icact.2019.8702017>

Kumar, S., Agrawal, K., & Mandan, N. (2020). Red Wine Quality Prediction Using Machine Learning Techniques. *2020 International Conference on Computer Communication and Informatics (ICCCI)*. <https://doi.org/10.1109/iccci48352.2020.9104095>

Pawar, D., Mahajan, A., & Bhoithe, S. (2019). Wine Quality Prediction using Machine Learning Algorithms. *International Journal of Computer Applications Technology and Research*, *8*(9), 385–388. <https://doi.org/10.7753/ijcatr0809.1010>

Wollan, D. (2010). Membrane and other techniques for the management of wine composition. *Managing Wine Quality*, 133–163. <https://doi.org/10.1533/9781845699987.1.133>

Ye, C., Li, K., & Jia, G.-zhu. (2020). A new red wine prediction framework using machine learning. *Journal of Physics: Conference Series*, *1684*, 012067. <https://doi.org/10.1088/1742-6596/1684/1/012067>