Wine Quality Prediction using classification

Under the supervision of:
Submitted by:

Dr. Devesh Kumar Srivastava
Professor (IT Department)

Santosh Maturi
(169108083)

Information Technology
MANIPAL UNIVERSITY JAIPUR
JAIPUR – 303007
RAJASTHAN, INDIA
2019

Abstract:

Wine Quality Red dataset is included, related to red vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests and predict

Introduction:

Wine is an alcoholic drink made from fermented grapes. Yeast consumes the sugar in the grapes and converts it to ethanol, carbon dioxide, and heat. Different varieties of grapes and strains of yeasts produce different styles of wine. These variations result from the complex interactions between the biochemical development of the grape, the reactions involved in fermentation, the terroir, and the production process. Many countries enact legal appellations intended to define styles and qualities of wine. These typically restrict the geographical origin and permitted varieties of grapes, as well as other aspects of wine production. Wines not made from grapes include rice wine and fruit wines such as plum, cherry, pomegranate, currant and elderberry.

Attribute Information:

Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

Objective:

-Data Preparation:diving the quality score into 2 different categories good and bad
And then encoded into 0 and 1
-Create visualizations to depict how residual sugar
-density and alcohol affect the quality of the wine
-Univariate and bivariate analysis
-Other variable observations
-Faulty wines

Methodology:

First of all i have applied the Multiple linear regression since the prediction score is too less then i have applied the logistic regression and later on i have applied support vector machine and k nearest neighbours and then Random forest classifier
And naive bayes

Multiple Linear Regression:

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable *x* is associated with a value of the dependent variable *y*. The population regression line for *p* explanatory variables x_1, x_2, ... , x_p is defined to be $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$. This line describes how the mean response μ_y changes with the explanatory variables. The observed values for *y* vary about their means μ_y and are assumed to have the same standard deviation σ. The fitted values b_0, b_1, ..., b_p estimate the parameters β_0, β_1, ..., β_p of the population regression line.

Logistic Regression:

Logistic regression predicts categorical outcomes (binomial / multinomial values of y), whereas linear Regression is good for predicting continuous-valued outcomes.The predictions of Logistic Regression (henceforth, LogR in this article) are in the form of probabilities of an event occurring, ie the probability of y=1, given certain values of input variables x. Thus, the results of LogR range between 0-1.

$$\frac{1}{1+ e^{-x}} = \frac{e^{x}}{1+ e^{x}}$$

K Nearest Neighbours:

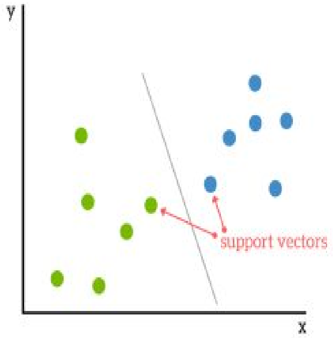
After selecting the value of *k*, you can make predictions based on the *KNN* examples. For regression, *KNN* predictions is the average of the *k*-nearest neighbors outcome.

$$y = \frac{1}{K}\sum_{i=1}^{K} y_i$$

where y_i is the *i*th case of the examples sample and *y* is the prediction (outcome) of the query point. In contrast to regression, in classification problems, *KNN* predictions are based on a voting scheme in which the winner is used to label the query.

Support Vector Machine:

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVMs are more commonly used in classification problems.



Support vectors are the data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a data set

Naive Bayes Classifier:

Naive Bayes classifiers can handle an arbitrary number of independent variables whether continuous or categorical. Given a set of variables, X = {x1,x2,x...,xd}, we want to construct the posterior probability for the event Cj among a set of possible outcomes C = {c1,c2,c...,cd}. In a more familiar language, X is the predictors and C is the set of categorical levels present in the dependent variable. Using Bayes' rule:

$$p\left(C_j \mid x_1, x_2, \ldots, x_d\right) \propto p\left(x_1, x_2, \ldots, x_d \mid C_j\right) p\left(C_j\right)$$

where p(Cj | x1,x2,x...,xd) is the posterior probability of class membership, i.e., the probability that X belongs to Cj. Since Naive Bayes assumes that the conditional probabilities of the independent variables are statistically independent we can decompose the likelihood to a product of terms:

$$p\left(X \mid C_j\right) \propto \prod_{k=1}^{d} p\left(x_k \mid C_j\right)$$

and rewrite the posterior as:

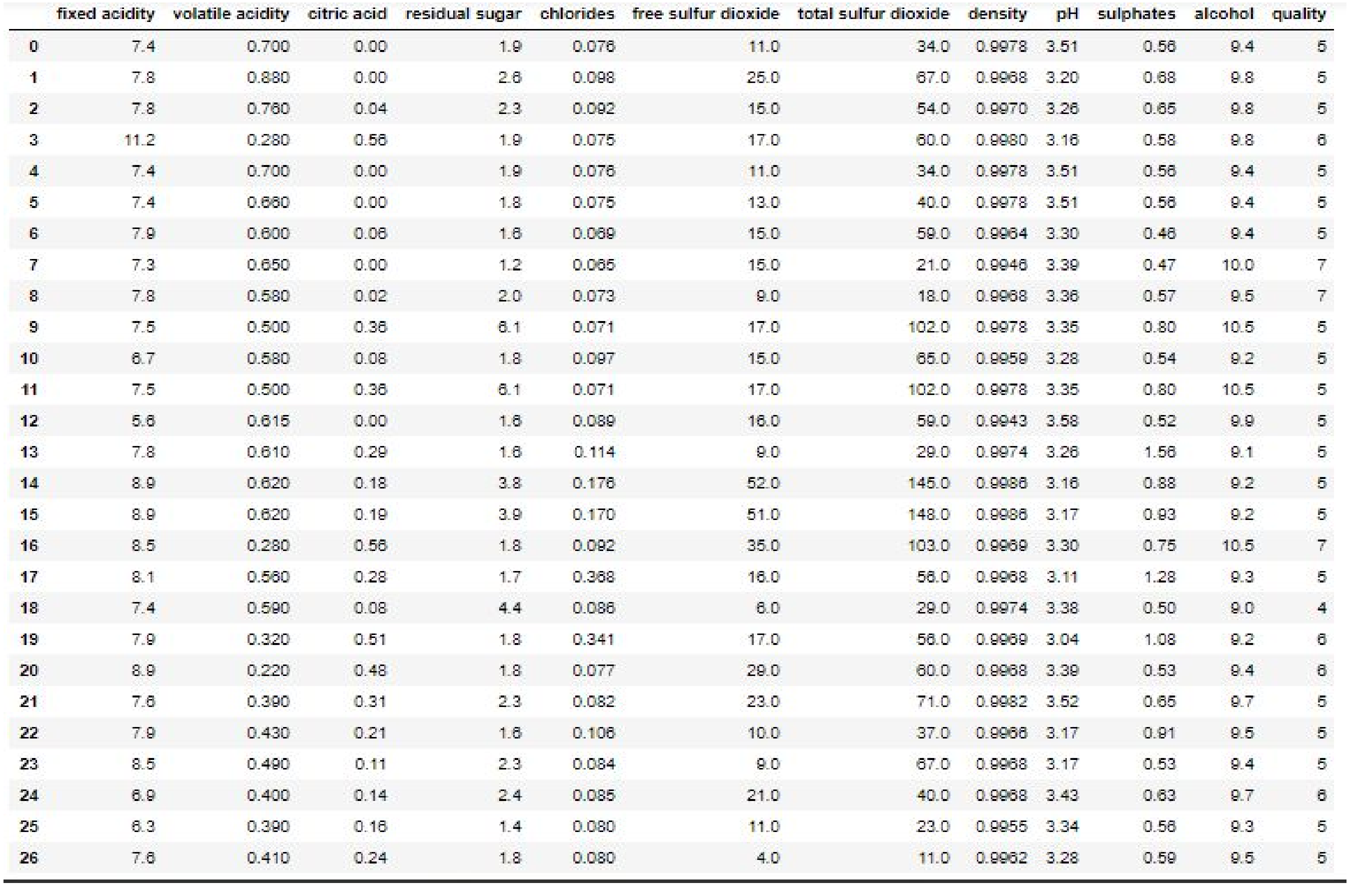
$$p\left(C_j \mid X\right) \propto p\left(C_j\right) \prod_{k=1}^{d} p\left(x_k \mid C_j\right)$$

Proposed Work:

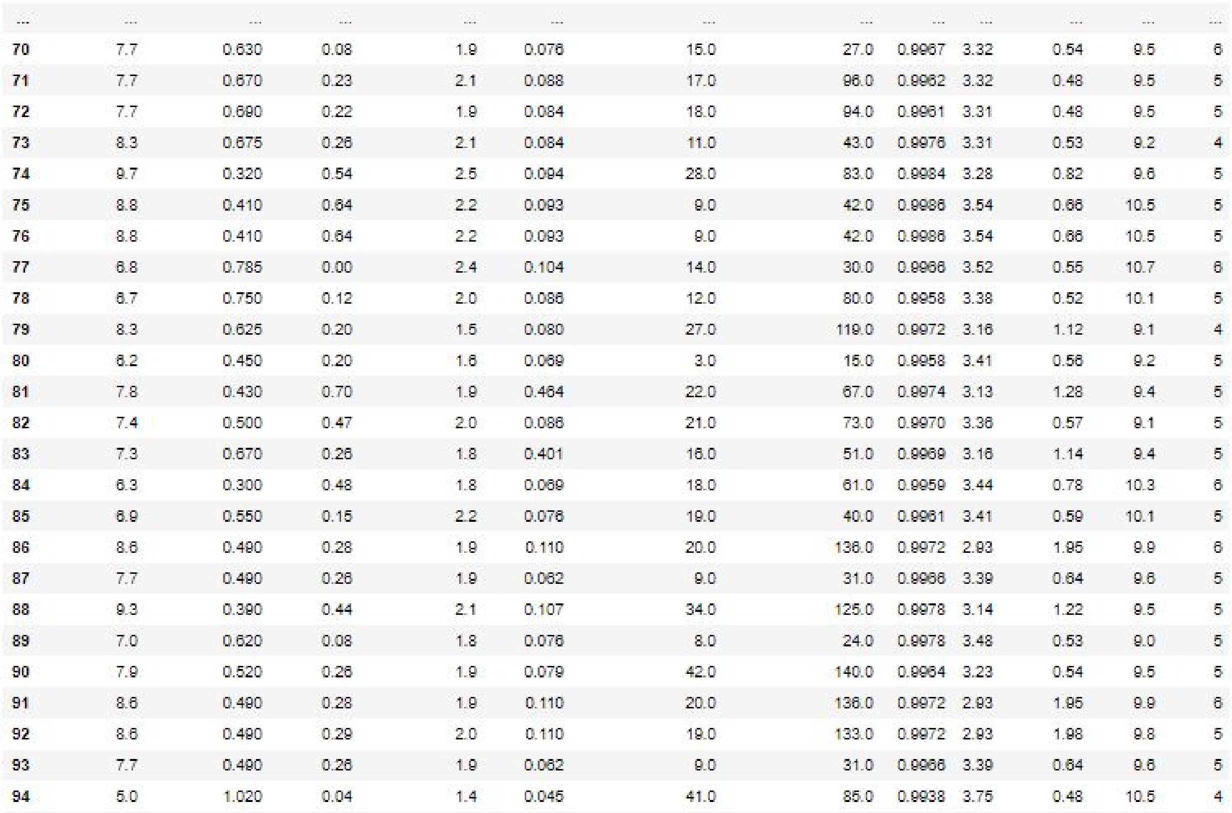
In this work, machine learning techniques are used to determine dependency of wine quality on other variables and in wine quality predictions. This section gives insights of proposed methodology. First Wine dataset is preprocessed as explained in previous section. Further, linear regression is applied to determine dependency of Wine quality on other 11 independent variables (predictors). Then after, important predictors are selected according to dependency of wine quality on independent variables. At last, Wine quality is predicted with the help of support vector machine and Random Forest considering all predictors and selected predictors.

Results:

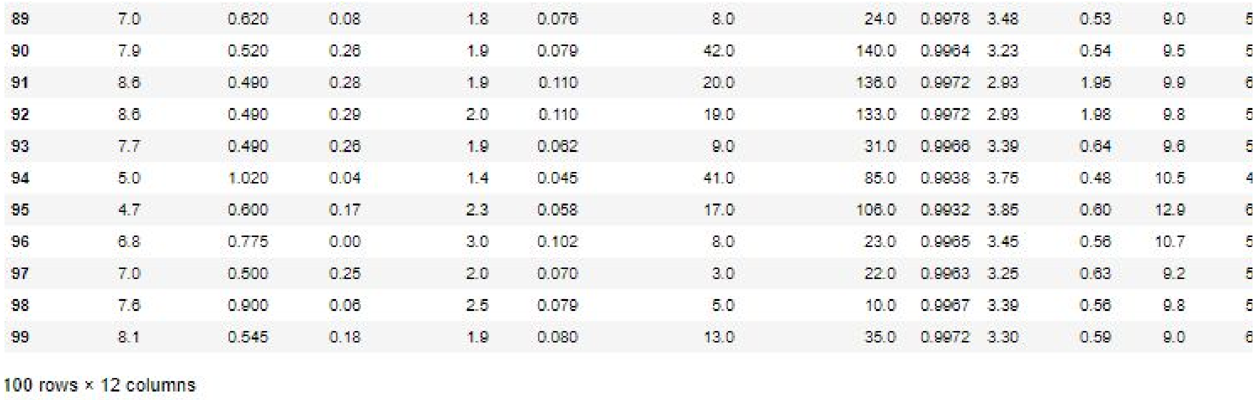
DataSet:



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
5	7.4	0.660	0.00	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	5
6	7.9	0.600	0.06	1.6	0.069	15.0	59.0	0.9964	3.30	0.46	9.4	5
7	7.3	0.650	0.00	1.2	0.065	15.0	21.0	0.9946	3.39	0.47	10.0	7
8	7.8	0.580	0.02	2.0	0.073	9.0	18.0	0.9968	3.36	0.57	9.5	7
9	7.5	0.500	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.80	10.5	5
10	6.7	0.580	0.08	1.8	0.097	15.0	65.0	0.9959	3.28	0.54	9.2	5
11	7.5	0.500	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.80	10.5	5
12	5.6	0.615	0.00	1.6	0.089	16.0	59.0	0.9943	3.58	0.52	9.9	5
13	7.8	0.610	0.29	1.6	0.114	9.0	29.0	0.9974	3.26	1.56	9.1	5
14	8.9	0.620	0.18	3.8	0.176	52.0	145.0	0.9986	3.16	0.88	9.2	5
15	8.9	0.620	0.19	3.9	0.170	51.0	148.0	0.9986	3.17	0.93	9.2	5
16	8.5	0.280	0.56	1.8	0.092	35.0	103.0	0.9969	3.30	0.75	10.5	7
17	8.1	0.560	0.28	1.7	0.368	16.0	56.0	0.9968	3.11	1.28	9.3	5
18	7.4	0.590	0.08	4.4	0.086	6.0	29.0	0.9974	3.38	0.50	9.0	4
19	7.9	0.320	0.51	1.8	0.341	17.0	56.0	0.9969	3.04	1.08	9.2	6
20	8.9	0.220	0.48	1.8	0.077	29.0	60.0	0.9968	3.39	0.53	9.4	6
21	7.6	0.390	0.31	2.3	0.082	23.0	71.0	0.9982	3.52	0.65	9.7	5
22	7.9	0.430	0.21	1.6	0.106	10.0	37.0	0.9966	3.17	0.91	9.5	5
23	8.5	0.490	0.11	2.3	0.084	9.0	67.0	0.9968	3.17	0.53	9.4	5
24	6.9	0.400	0.14	2.4	0.085	21.0	40.0	0.9968	3.43	0.63	9.7	6
25	6.3	0.390	0.16	1.4	0.080	11.0	23.0	0.9955	3.34	0.56	9.3	5
26	7.6	0.410	0.24	1.8	0.080	4.0	11.0	0.9962	3.28	0.59	9.5	5



...	...	...	...	...	...	...	...	...	...	...	...	...
70	7.7	0.630	0.08	1.9	0.076	15.0	27.0	0.9967	3.32	0.54	9.5	6
71	7.7	0.670	0.23	2.1	0.088	17.0	96.0	0.9962	3.32	0.48	9.5	5
72	7.7	0.690	0.22	1.9	0.084	18.0	94.0	0.9961	3.31	0.48	9.5	5
73	8.3	0.675	0.26	2.1	0.084	11.0	43.0	0.9976	3.31	0.53	9.2	4
74	9.7	0.320	0.54	2.5	0.094	28.0	83.0	0.9984	3.28	0.82	9.6	5
75	8.8	0.410	0.64	2.2	0.093	9.0	42.0	0.9986	3.54	0.66	10.5	5
76	8.8	0.410	0.64	2.2	0.093	9.0	42.0	0.9986	3.54	0.66	10.5	5
77	6.8	0.785	0.00	2.4	0.104	14.0	30.0	0.9966	3.52	0.55	10.7	6
78	6.7	0.750	0.12	2.0	0.086	12.0	80.0	0.9958	3.38	0.52	10.1	5
79	8.3	0.625	0.20	1.5	0.080	27.0	119.0	0.9972	3.16	1.12	9.1	4
80	6.2	0.450	0.20	1.6	0.069	3.0	15.0	0.9958	3.41	0.56	9.2	5
81	7.8	0.430	0.70	1.9	0.464	22.0	67.0	0.9974	3.13	1.28	9.4	5
82	7.4	0.500	0.47	2.0	0.086	21.0	73.0	0.9970	3.36	0.57	9.1	5
83	7.3	0.670	0.26	1.8	0.401	16.0	51.0	0.9969	3.16	1.14	9.4	5
84	6.3	0.300	0.48	1.8	0.069	18.0	61.0	0.9959	3.44	0.78	10.3	6
85	6.9	0.550	0.15	2.2	0.076	19.0	40.0	0.9961	3.41	0.59	10.1	5
86	8.6	0.490	0.28	1.9	0.110	20.0	136.0	0.9972	2.93	1.95	9.9	6
87	7.7	0.490	0.26	1.9	0.062	9.0	31.0	0.9966	3.39	0.64	9.6	5
88	9.3	0.390	0.44	2.1	0.107	34.0	125.0	0.9978	3.14	1.22	9.5	5
89	7.0	0.620	0.08	1.8	0.076	8.0	24.0	0.9978	3.48	0.53	9.0	5
90	7.9	0.520	0.26	1.9	0.079	42.0	140.0	0.9964	3.23	0.54	9.5	5
91	8.6	0.490	0.28	1.9	0.110	20.0	136.0	0.9972	2.93	1.95	9.9	6
92	8.6	0.490	0.29	2.0	0.110	19.0	133.0	0.9972	2.93	1.98	9.8	5
93	7.7	0.490	0.26	1.9	0.062	9.0	31.0	0.9966	3.39	0.64	9.6	5
94	5.0	1.020	0.04	1.4	0.045	41.0	85.0	0.9938	3.75	0.48	10.5	4



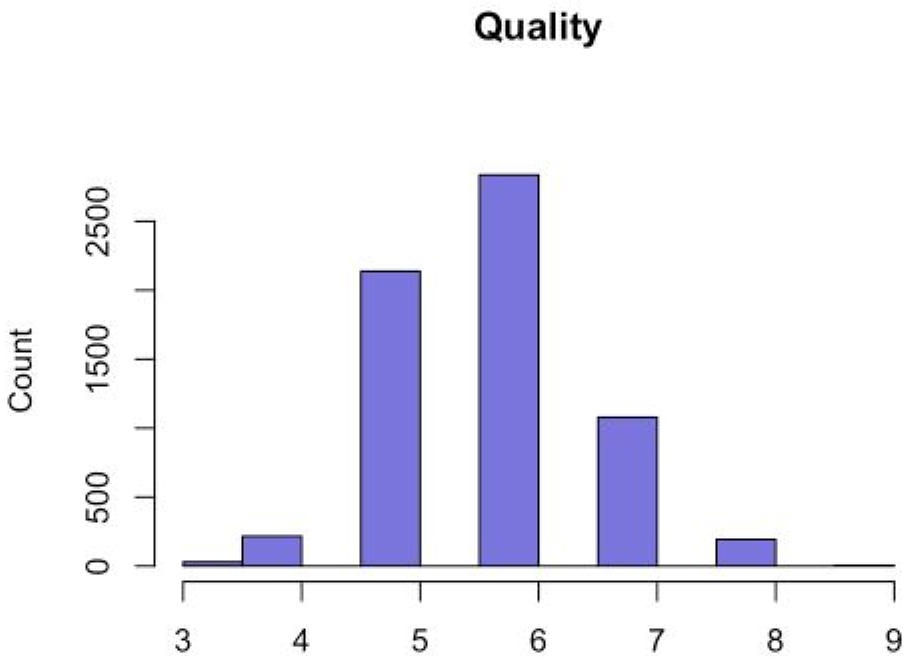
89	7.0	0.620	0.08	1.8	0.076	8.0	24.0	0.9978	3.48	0.53	9.0	5
90	7.9	0.520	0.26	1.9	0.079	42.0	140.0	0.9964	3.23	0.54	9.5	5
91	8.6	0.490	0.28	1.9	0.110	20.0	136.0	0.9972	2.93	1.95	9.9	6
92	8.6	0.490	0.29	2.0	0.110	19.0	133.0	0.9972	2.93	1.98	9.8	5
93	7.7	0.490	0.26	1.9	0.062	9.0	31.0	0.9966	3.39	0.64	9.6	5
94	5.0	1.020	0.04	1.4	0.045	41.0	85.0	0.9938	3.75	0.48	10.5	4
95	4.7	0.600	0.17	2.3	0.058	17.0	106.0	0.9932	3.85	0.60	12.9	6
96	6.8	0.775	0.00	3.0	0.102	8.0	23.0	0.9965	3.45	0.56	10.7	5
97	7.0	0.500	0.25	2.0	0.070	3.0	22.0	0.9963	3.25	0.63	9.2	5
98	7.6	0.900	0.06	2.5	0.079	5.0	10.0	0.9967	3.39	0.56	9.8	5
99	8.1	0.545	0.18	1.9	0.080	13.0	35.0	0.9972	3.30	0.59	9.0	6

100 rows × 12 columns

Quality:

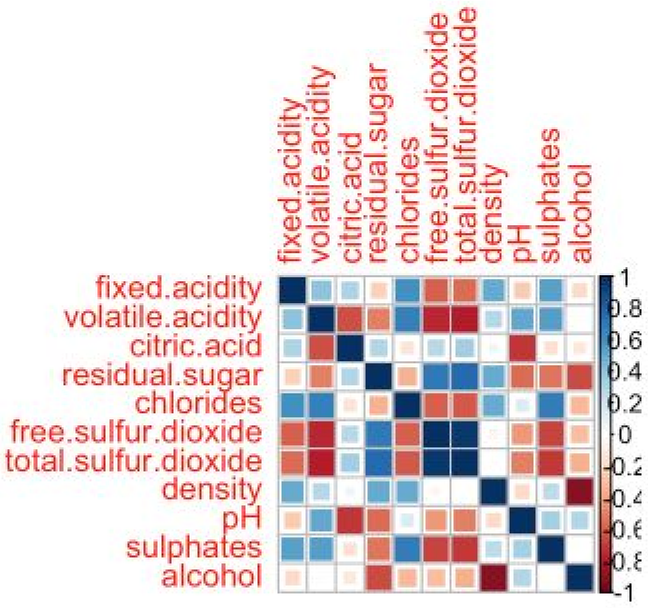
displays the first 10 observations of the combined data frame of red and white wine files.There are 1599 observations of red wine each frame with 13 variables giving us a 6497 by 13 dataframe.

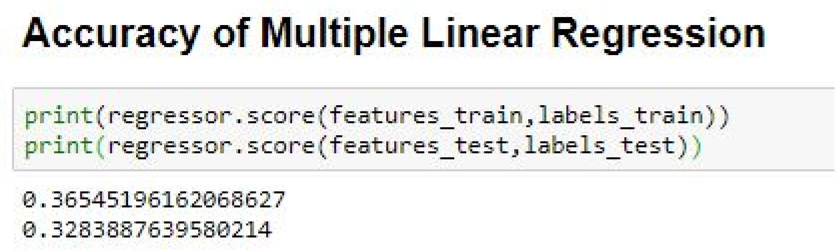
First, I will look at the variable quality. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). Below, I see that the bulk of the wine quality is at a quality of 5, 6 and 7. There is no observations below a quality of 3 and none above 9.



Correlation:

I look at the correlation between the continuous variables. A common concern in data analysis is multicollinearity, where one predictor variables is highly correlated with another variables. The problem with multicollinearity is that it makes parameter estimation unstable as well as difficult to understand the effect that the predictor has on the response



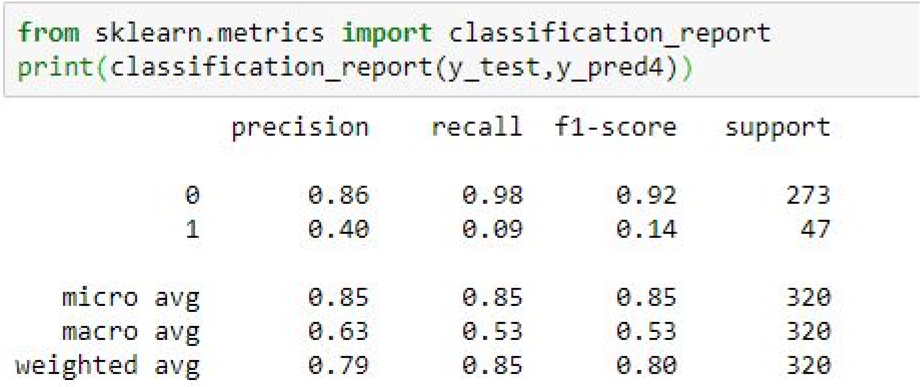


Accuracy of Multiple Linear Regression

```
print(regressor.score(features_train,labels_train))
print(regressor.score(features_test,labels_test))
```

```
0.36545196162068627
0.3283887639580214
```

Accuracy of Logistic Regression



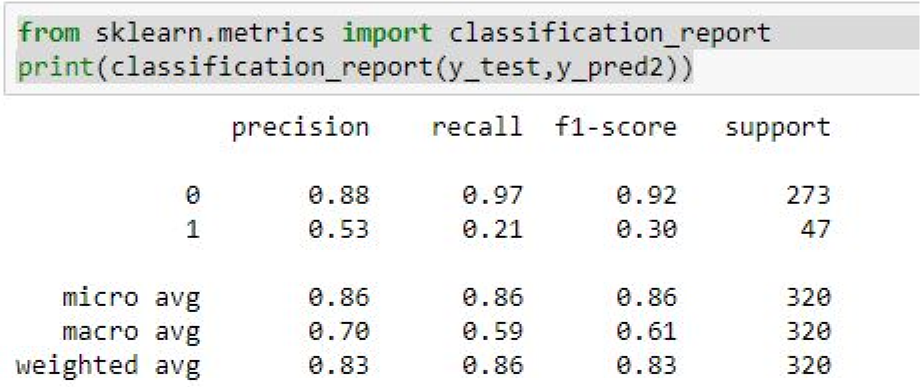
```
from sklearn.metrics import classification_report
print(classification_report(y_test,y_pred4))
```

```
              precision    recall  f1-score   support

           0       0.86      0.98      0.92       273
           1       0.40      0.09      0.14        47

   micro avg       0.85      0.85      0.85       320
   macro avg       0.63      0.53      0.53       320
weighted avg       0.79      0.85      0.80       320
```

Accuracy of K nearest neighbours:



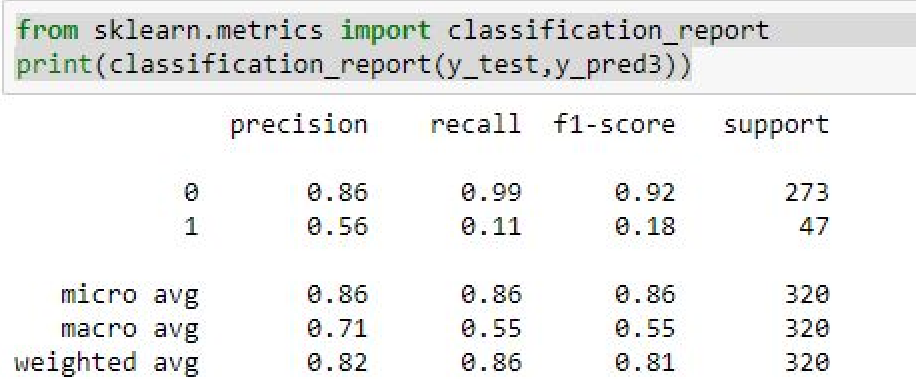
```
from sklearn.metrics import classification_report
print(classification_report(y_test,y_pred2))
```

```
              precision    recall  f1-score   support

           0       0.88      0.97      0.92       273
           1       0.53      0.21      0.30        47

   micro avg       0.86      0.86      0.86       320
   macro avg       0.70      0.59      0.61       320
weighted avg       0.83      0.86      0.83       320
```

Results for Support vector machine:



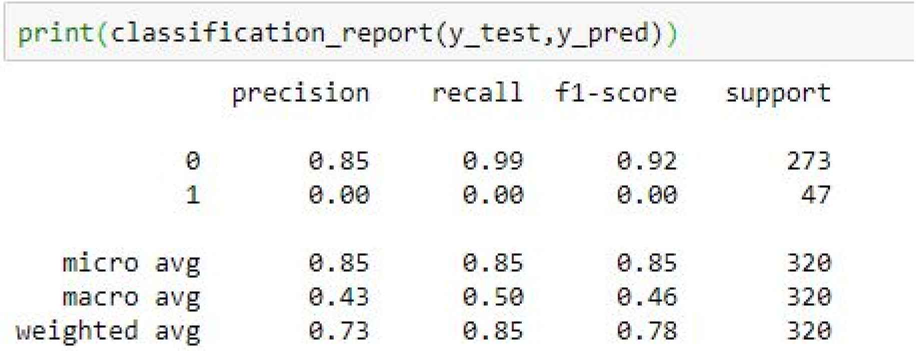
```
from sklearn.metrics import classification_report
print(classification_report(y_test,y_pred3))
```

```
              precision    recall  f1-score   support

           0       0.86      0.99      0.92       273
           1       0.56      0.11      0.18        47

   micro avg       0.86      0.86      0.86       320
   macro avg       0.71      0.55      0.55       320
weighted avg       0.82      0.86      0.81       320
```

Results for naive bayes:



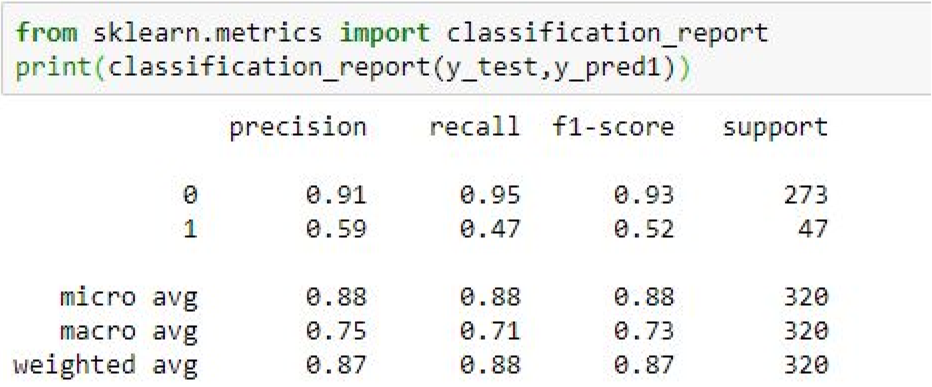
```
print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

           0       0.85      0.99      0.92       273
           1       0.00      0.00      0.00        47

   micro avg       0.85      0.85      0.85       320
   macro avg       0.43      0.50      0.46       320
weighted avg       0.73      0.85      0.78       320
```

Results for Random Forest:



```
from sklearn.metrics import classification_report
print(classification_report(y_test,y_pred1))
```

```
              precision    recall  f1-score   support

           0       0.91      0.95      0.93       273
           1       0.59      0.47      0.52        47

   micro avg       0.88      0.88      0.88       320
   macro avg       0.75      0.71      0.73       320
weighted avg       0.87      0.88      0.87       320
```

And finally among all these the best prediction that has occured is through the Random forest classifier as 87%

Conclusion:

At the end of the story I can say that Wine quality is a very complex study.Good wine is more than perfect combination of different chemical components. Future improvement can be made if more data can be collected on both low-quality and high-quality wine.If the data set has more records on both the low end and high end, the quality of analysis can be improved. We can be more certain about whether there is a significant correlation between a chemical component and the wine quality.

References:

[1] Ebeler S. (1999) “Flavor Chemistry — Thirty Years of Progress: chapter Linking flavour chemistry to sensory analysis of wine”. Kluwer Academic Publishers, 409–422.

[2] Legin, Rudnitskaya, Luvova, Vlasov, Natale and D'Amico. (2003) “Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception”. Analytica Chimica Acta 484 (1): 33–34.

[3] Sun, Danzer and Thiel. (1997) “Classification of wine samples by means of artificial neural networks and discrimination analytical methods”. Fresenius Journal of Analytical Chemistry 359 (2) 143–149.

[4] Vlassides, Ferrier and Block. (2001) “Using historical data for bioprocess optimization: modeling wine characteristics using classification and archived process information”. Biotechnology and Bioengineering 73 (1) 55-68.

[5] Moreno, Gonzalez-Weller, Gutierrez, Marino, Camean, Gonzalez and Hardisson. (2007) “Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic”. Talanta 72 263–268