QUALITY OF WINE

EXCUTIVE SUMMARY:

Wine is a good beverage which is made from the fermented grapes. A glass of good wine gives good health. Spain is the major producer of wine. But there was drastic drop in the sales of wine in 2003 in Spain. So, to increase the exports of wine the data mining solutions are provided, and these solutions helps Spain to improve their quality of wine and helps them to find major sales of whether it is red or the white wine. The main motive of this project is to find the best wine based on the chemical components of the wine. Sales can be increased by improving the Quality of wine. The can checked using the Linear Regression.

The major components that are used in the wine are chlorides and the sulfuric acids and alcohol content in the wine. White wine contains high number of chlorides. I have used the different modeling techniques to find the best quality of wine in both the red and the white wine using the Logistic regression, J48 tree, Naïve Bayes. But out of the three algorithms or the modeling techniques Logistic Regression provides best results because 98% of the instances are correctly classified. I have used the linear regression techniques to calculate the correlation coefficient and the quality of the wine depending on the maximum content of the alcohol and the chlorides in both red and the white wine. The quality of the wine is calculated on the scale of 1 to 10. When the wine measured above 6 it is known as the good wine.

Visual Analytics gives better understanding. So, I have used the Tableau to show the quality of the wine based on the alcohol content in the wine. Red wine has more quality as compared to the white wine. In every chemical component the red wine shows best and as well as in the sales the red wine as more sales. So, this helps the Spain to improve their productivity and the sales as per the niche market standards in both the red and the white wine productivity.

BUSINESS UNDERSTANDING:

A sip of wine gives good flavor to food and is enjoyed widely all over the world in all the happy movements. Spain is one of the top ten countries who produce their majority of the income through the production of wine. Spain is more popular to produce white wine. The Exports of **Vinho Verde wine** which produces both the white and the red wine have decreased to 3.17% in the year 2005 but the end of the year 2007 the exports were increased to 36%. During the fermentation of wine in Vinho Verde wine has faced many problems due to the impact on the quality of wine which in turn reduced the productivity.

This analysis is done for predicting the preferences of the human wine taste. This can be done through the analytical tests using a large data set. We need to focus on the ingredients that are used in the wine to find the concentration of the wine. Depending on the concentration and the quality of the wine the best wine is preferred. Depending on the niche market we can predict which wine as the more sales whether the red wine or the white wine.

Depending on the different chemical constituents like the alcohol constituents, nitric acid we can find the best quality of wine among the best cultivators of wine in Italy.

DATA PREPARATION:

Source of data: http://www.sciencedirect.com/science/article/pii/S0167923609001377

http://archive.ics.uci.edu/ml/datasets/Wine+Quality

This data set is mostly related to the red and the white wine variants that are used in SPAIN especially in the “Vinho Verde” wine. From the different variants the quality of wine can be determined, and also which wine has major sales and better quality. There are totally 12 attributes in the data set and around 4000 instances.

The attributes that are there in the data set are
Fixed acidity
Volatile acidity

QUALITY OF WINE

Citric acid
Residual sugar
Chlorides
Free sulfur dioxide
Total sulfur dioxide
Density
PH
Sulphates
Alcohol
R/W
The quality of the wine is tested on the scale of 0 to 10.

- Now days the interest in wine has increased as a result the manufacturing companies take important step to improve the wine taste by using the latest technologies to improve production of the industry and increase the selling.
- The fixed acids generally produced by the metabolism of carbohydrates, fats and proteins which is incomplete to produce wine. The maximum quantity of fixed acids in red wine is 15.9g per tartaric acid quantity in dm3 and maximum quantity of fixed acids in white wine is 14.2g per tartaric acid in dm3.
- The quantity of residual sugar content is more in white wine as compared to the red wine. Where the residual sugar content in red wine up to maximum limit is 15.5g/dm3 whereas the white wine has a maximum residual sugar content of 65.8g/dm3.
- The PH content of white of Red wine is 4 in range of 0 to 14 and white wine has maximum PH of 3.8 in the scale of 0 to 14.
- The maximum alcohol content in volume of one liter quantity of red wine is 14.9 and the maximum content of alcohol in white wine in volume of one liter is 14.2.

DATA UNDERSTANDING:

Source: http://www3.dsi.uminho.pt/pcortez/wine/

This Data was obtained from the paper that got published in “Modeling the wine preferences from the Physicochemical Properties”. This Data had all the chemical components that are used in the red and white wine and the alcohol content in the red and white wine.

Limitations:

- In the data I could not find the exact components ratio that is used for the Red and white wine. What is differentiation of the chemical components used in both the types of wine?
- I could not find the information about the competitor’s brands and the selling price of both red and white wine which made the production of “Vinho Verde” lower as compared to other brands in Spain.
- The data did not give the information about the type of the grape used for white wine and red wine and the chemical constituents. “Alvarinho” is the type of white grape which produces the best wine out if this grapes and especially white wine is prepared. “Espadeiro” is the Black Grapes which produces the best red wine.
- These kinds of Chemical compositions and the variety of grapes are not produced in the data to know why the sales of “Vinho Verde” as reduced.

Cleaning of the data:

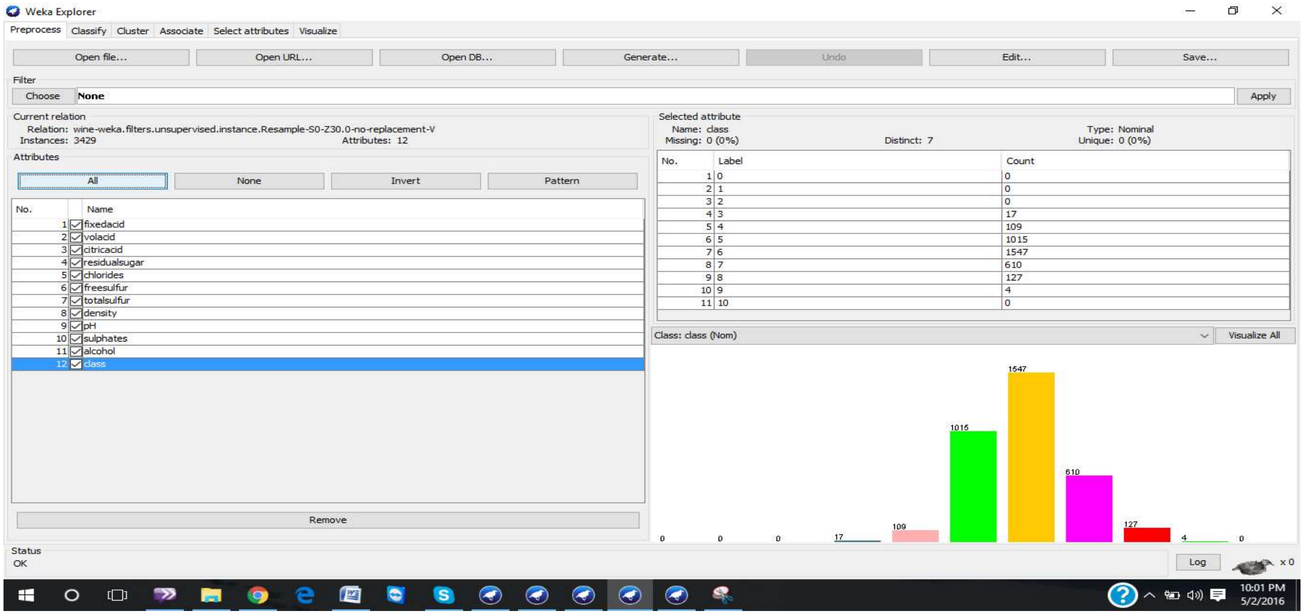
I am using CSV file to load in the WEKA tool. But the CSV file I used had many special characters like the double quotes and inverted commas. So, I have cleaned every data and placed in the form of separate columns so that WEKA can accept it easily. And again, I have even prepared a normal Excel sheet to demonstrate the chemical characters in the wine using the Tableau graphical analytics instead of algorithms.
All the columns which are in the CSV file are very useful. So, it is not required to remove any kinds of columns. So, I have just used the data but just tried to remove duplicates.

QUALITY OF WINE

MODELLING:

Attributes:

Total numbers of attributes are 12 and below is the screen shot for the attributes that are in the given data. Using the below attributes, we can even find the rank of the attributes depending on the type of wine whether it is the red or the white wine.



Current relation
Relation: wine-weka.filters.unsupervised.instance.Resample-S0-Z30.0-no-replacement-V
Instances: 3429 Attributes: 12

Attributes

No.		Name
1	☑	fixedacid
2	☑	volacid
3	☑	citricacid
4	☑	residualsugar
5	☑	chlorides
6	☑	freesulfur
7	☑	totalsulfur
8	☑	density
9	☑	pH
10	☑	sulphates
11	☑	alcohol
12	☑	class

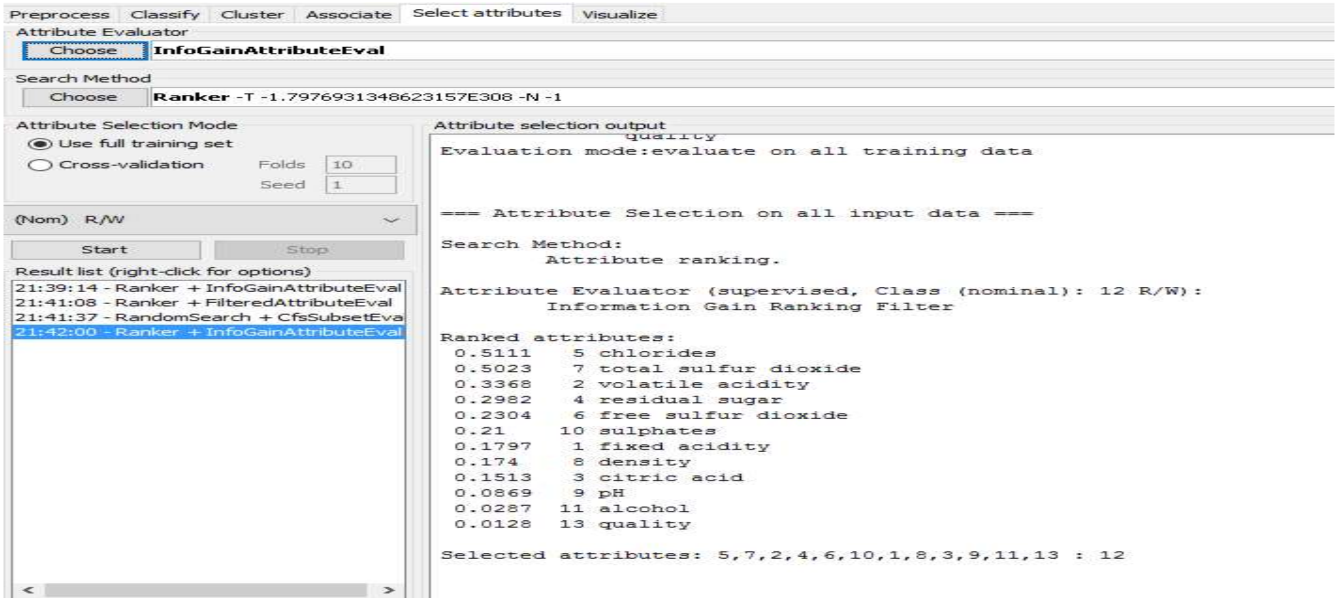
Selected attribute
Name: class Type: Nominal
Missing: 0 (0%) Distinct: 7 Unique: 0 (0%)

No.	Label	Count
1	0	0
2	1	0
3	2	0
4	3	17
5	4	109
6	5	1015
7	6	1547
8	7	610
9	8	127
10	9	4
11	10	0

Class: class (Nom)

Status
OK

Info Gain for the attributes



Attribute Evaluator
InfoGainAttributeEval

Search Method
Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode
● Use full training set
○ Cross-validation Folds 10 Seed 1

(Nom) R/W

Result list (right-click for options)
21:39:14 - Ranker + InfoGainAttributeEval
21:41:08 - Ranker + FilteredAttributeEval
21:41:37 - RandomSearch + CfsSubsetEva
21:42:00 - Ranker + InfoGainAttributeEval

Attribute selection output

```
              quality
Evaluation mode:evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 12 R/W):
        Information Gain Ranking Filter

Ranked attributes:
 0.5111    5 chlorides
 0.5023    7 total sulfur dioxide
 0.3368    2 volatile acidity
 0.2982    4 residual sugar
 0.2304    6 free sulfur dioxide
 0.21     10 sulphates
 0.1797    1 fixed acidity
 0.174     8 density
 0.1513    3 citric acid
 0.0869    9 pH
 0.0287   11 alcohol
 0.0128   13 quality

Selected attributes: 5,7,2,4,6,10,1,8,3,9,11,13 : 12
```

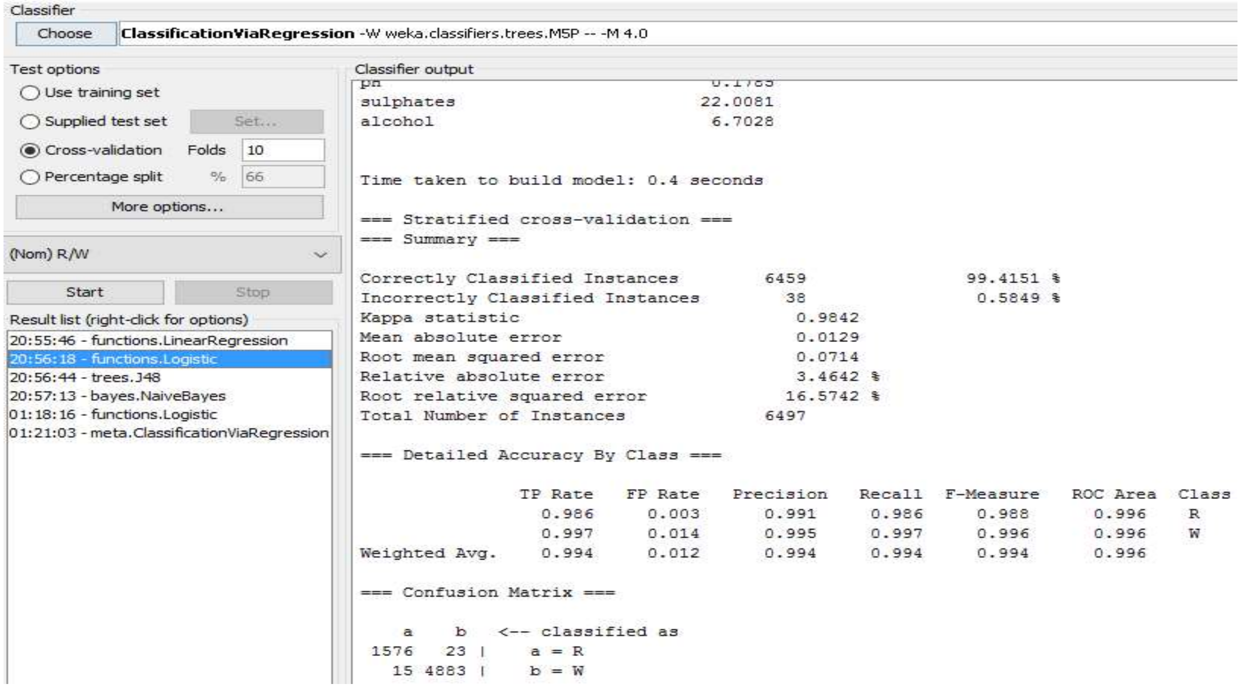
This helps to find the ranking of the attributes. So, the major chemical constituent in the wine is the Chlorides. The ranking of attributes is found using the Info Gain Attribute Evaluation

QUALITY OF WINE

I am using different modeling techniques to find out the most appropriate techniques for finding the best quality of the wine. I used Tableau to give a clear visual effect to find the best quality of wine in both the red and white wine as per the chemical components.

Logistic Regression

The Logistic Regression is used to find the true positives and the false positives of the Red wine and the white wine. This modeling technique helps us to find the confusion matrix.



Classifier: **ClassificationViaRegression** -W weka.classifiers.trees.M5P -- -M 4.0

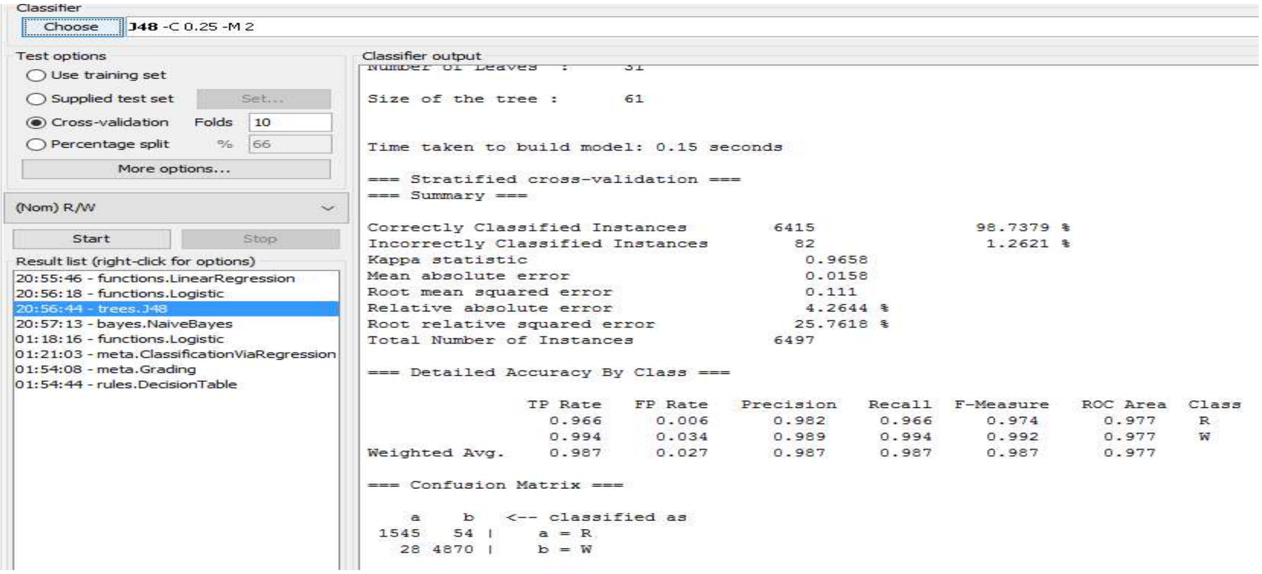
Test options:
- ○ Use training set
- ○ Supplied test set
- ● Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) R/W

Result list (right-click for options)
- 20:55:46 - functions.LinearRegression
- 20:56:18 - functions.Logistic
- 20:56:44 - trees.J48
- 20:57:13 - bayes.NaiveBayes
- 01:18:16 - functions.Logistic
- 01:21:03 - meta.ClassificationViaRegression

```
ph                          0.1785
sulphates                  22.0081
alcohol                     6.7028


Time taken to build model: 0.4 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6459               99.4151 %
Incorrectly Classified Instances        38                0.5849 %
Kappa statistic                          0.9842
Mean absolute error                      0.0129
Root mean squared error                  0.0714
Relative absolute error                  3.4642 %
Root relative squared error             16.5742 %
Total Number of Instances             6497

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.986     0.003      0.991     0.986     0.988      0.996    R
                 0.997     0.014      0.995     0.997     0.996      0.996    W
Weighted Avg.    0.994     0.012      0.994     0.994     0.994      0.996

=== Confusion Matrix ===

    a    b   <-- classified as
 1576   23 |    a = R
   15 4883 |    b = W
```

J48 Tree:



Classifier: **J48** -C 0.25 -M 2

Test options:
- ○ Use training set
- ○ Supplied test set
- ● Cross-validation Folds 10
- ○ Percentage split % 66

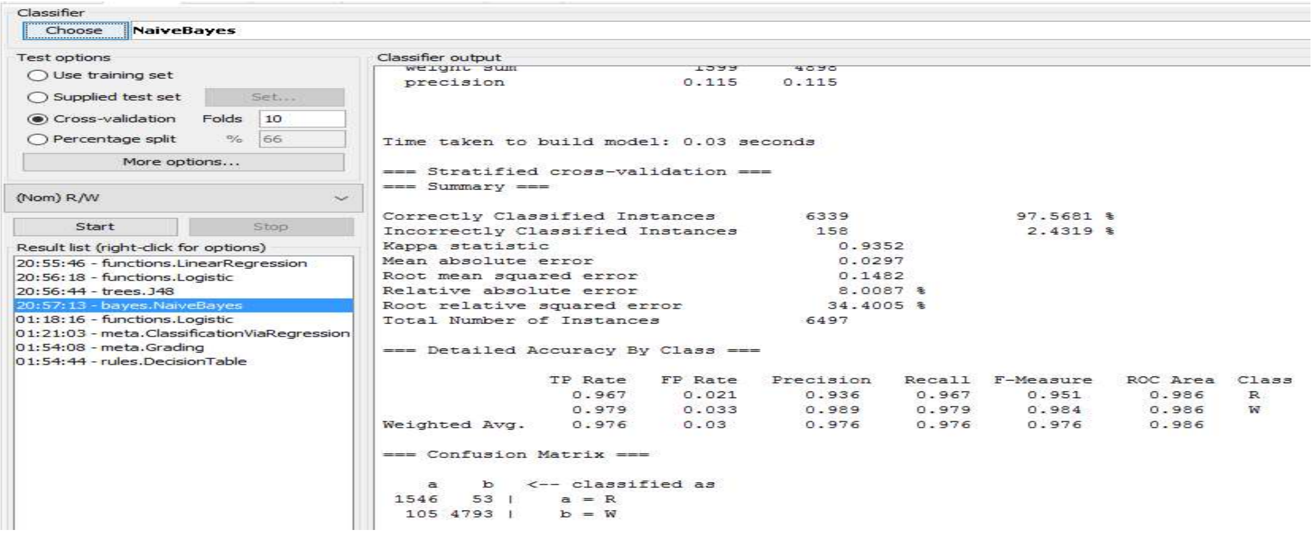
(Nom) R/W

Result list (right-click for options)
- 20:55:46 - functions.LinearRegression
- 20:56:18 - functions.Logistic
- 20:56:44 - trees.J48
- 20:57:13 - bayes.NaiveBayes
- 01:18:16 - functions.Logistic
- 01:21:03 - meta.ClassificationViaRegression
- 01:54:08 - meta.Grading
- 01:54:44 - rules.DecisionTable

```
Number of Leaves  :     31

Size of the tree :      61


Time taken to build model: 0.15 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6415               98.7379 %
Incorrectly Classified Instances        82                1.2621 %
Kappa statistic                          0.9658
Mean absolute error                      0.0158
Root mean squared error                  0.111
Relative absolute error                  4.2644 %
Root relative squared error             25.7618 %
Total Number of Instances             6497

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.966     0.006      0.982     0.966     0.974      0.977    R
                 0.994     0.034      0.989     0.994     0.992      0.977    W
Weighted Avg.    0.987     0.027      0.987     0.987     0.987      0.977

=== Confusion Matrix ===

    a    b   <-- classified as
 1545   54 |    a = R
   28 4870 |    b = W
```

QUALITY OF WINE

<u>Naïve Bayes:</u>



Classifier
Choose NaiveBayes

Test options
○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66
More options...

(Nom) R/W

Start Stop

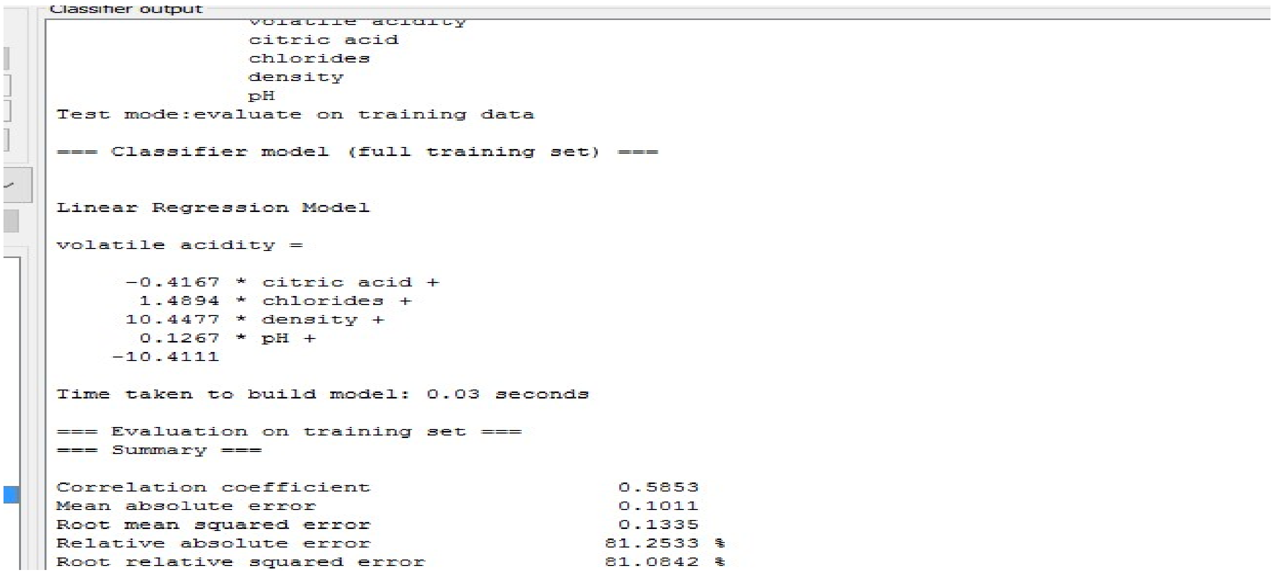
Result list (right-click for options)
20:55:46 - functions.LinearRegression
20:56:18 - functions.Logistic
20:56:44 - trees.J48
20:57:13 - bayes.NaiveBayes
01:18:16 - functions.Logistic
01:21:03 - meta.ClassificationViaRegression
01:54:08 - meta.Grading
01:54:44 - rules.DecisionTable

Classifier output

```
  weight sum                  1599     4898
  precision                  0.115    0.115


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6339               97.5681 %
Incorrectly Classified Instances       158                2.4319 %
Kappa statistic                          0.9352
Mean absolute error                      0.0297
Root mean squared error                  0.1482
Relative absolute error                  8.0087 %
Root relative squared error             34.4005 %
Total Number of Instances             6497

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.967     0.021      0.936     0.967     0.951      0.986    R
                 0.979     0.033      0.989     0.979     0.984      0.986    W
Weighted Avg.    0.976     0.03       0.976     0.976     0.976      0.986

=== Confusion Matrix ===

    a    b   <-- classified as
 1546   53 |    a = R
  105 4793 |    b = W
```

Linear Regression:

The quality of both red and the white wine is found by using the linear regression techniques.



Classifier output

```
                volatile acidity
                citric acid
                chlorides
                density
                pH
Test mode:evaluate on training data

=== Classifier model (full training set) ===


Linear Regression Model

volatile acidity =

     -0.4167 * citric acid +
      1.4894 * chlorides +
     10.4477 * density +
      0.1267 * pH +
    -10.4111

Time taken to build model: 0.03 seconds

=== Evaluation on training set ===
=== Summary ===

Correlation coefficient                  0.5853
Mean absolute error                      0.1011
Root mean squared error                  0.1335
Relative absolute error                 81.2533 %
Root relative squared error             81.0842 %
```

Quality = -0.4167*X1 + 1.4894*X2 + 10.4477*X3 + 0.1267 *X4 + 10.4111

Where the attributes I have selected are

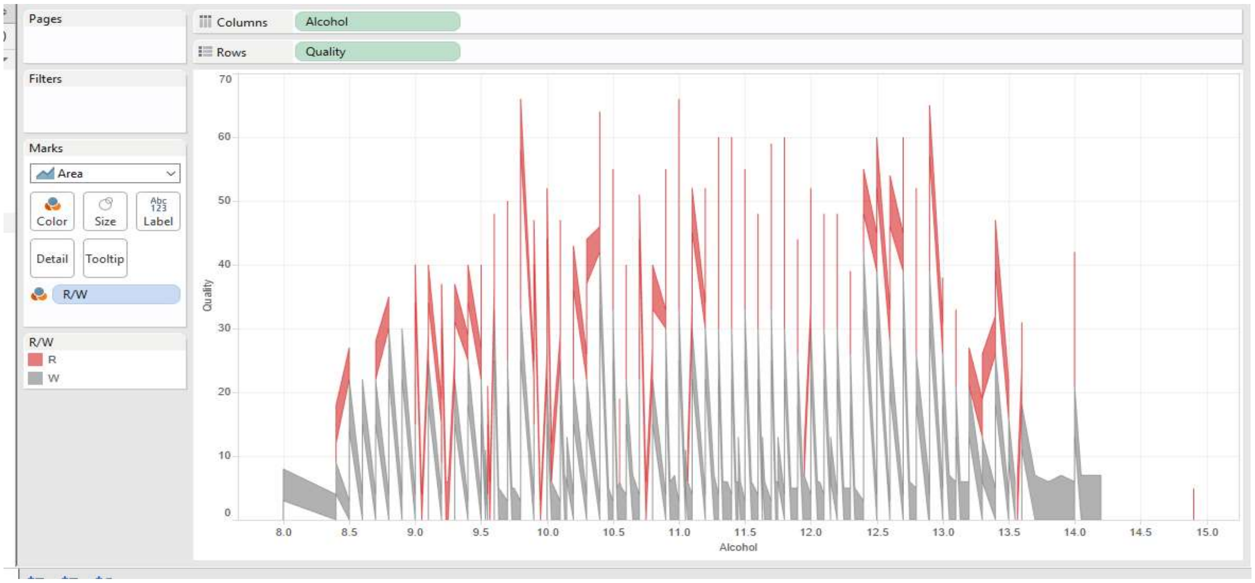
X1= Citric acid	X3= Density
X2 = Chlorides	X4= PH

Quality = -0.69+0.908+ 10.854+ 0.504 -10.411
Quality = 1.165

QUALITY OF WINE

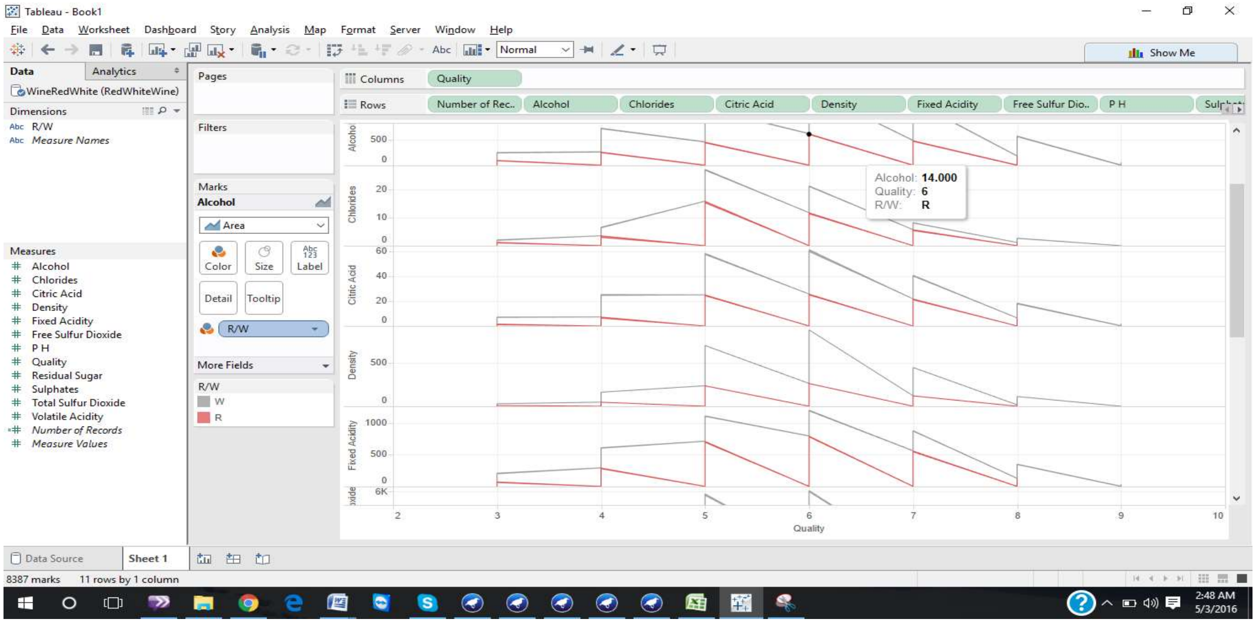
The quality of the wine when using the maximum content of the above chemical residues is greater than 1 (Quality > 1) that means the quality of the wine is best as per the above Linear Regression.
The correlation coefficient is 0.5853 which means it is greater than 0.5 so the model is correctly classified. I have considered the above attributes because the alcohol content and the quality of the wine is majorly dependent on the chemical constituents like the citric acid, density, chlorides and the PH. The chlorides play a major role in the production of the alcohol and during the fermentation of the grapes the chlorides are formed.

Tableau Analysis:



This is the Tableau analysis which I have considered to find between the quality and the alcohol content of the wine. The red wine has the best quality as per the visual analytics and even the machine learning techniques such as the WEKA also produced the same analysis as the chemical components. Since the Red wine is the best quality the production of the wine should be increased in Spain.
The below graph is the separate chemical components to the quality of the wine. The chemical contents are more in the white wine as per the below visual graphics in the tableau. The best quality of the wine is 8 as per the scale of 0-10. This is present in the red wine. Even the white wine as the quality of 9 on the scale but the chlorides are more in content in the white wine which in turn reduces the quality of the wine.

QUALITY OF WINE



Columns: Quality

Rows: Number of Rec.. Alcohol Chlorides Citric Acid Density Fixed Acidity Free Sulfur Dio.. P H Sulph..

Alcohol: 14.000
Quality: 6
R/W: R

I am using three different modeling techniques. But out of the three modeling techniques Logistic Regression gives an accurate result and the number of instances that are correctly classified are more as compared to the other two techniques. I have selected all the three types of modeling techniques to compare which techniques produces the best and accurate results in find the best quality of the wine on both red and the white wine.

The Linear regression modeling technique is used to calculate the quality of wine by selecting the different attributes depending in the ranking which we calculated from the Info Gain for the attributes. The quality of the wine is ranked in the scale of the 1 to 10. These regression techniques help to calculate manually depending on the chemical constituents.
I have used the Tableau visualization to give visual effects about the quality of the wine and the alcohol content in the wine. This helps to the business to understand which wine is rich is quality and the alcohol concentration. As per the analysis the red wine has the good quality and even the content of alcohol is as required by human body.

EVALUATION:

Total Number of Attributes	12	
Total Number of Instances	6497	
Logistic Regression	99.41% (6459 Instances correctly classified)	Time taken 0.4
Naïve Bayes	97.56% (6339 In stances correctly classified)	Time taken 0.15
J48 Tree	98.73% (6415 instances correctly classified	Time taken 0.03

As per the above table analysis and the modeling techniques analysis Logistic Regression provided best results. But, slight little drawback is the time constraint
Apart from that the true positives and the confusion matrix show better results in logistic regression as compared to the J48 and Naïve Bayes.

Logistic Regression:

Detailed accuracy: 98.6% are correctly modeled as Red wine and 99.7% of White wine is modeled and totally 99.4% is the weighted average that is correctly modeled.
The false positive rate is 0.3% is incorrectly modeled as Red wine and 1.4% is incorrectly modeled ad white wine and the totally weighted average that is incorrectly modeled is 1.2%.
Confusion Matrix:
1576 are modeled as Red wine 23 is modeled as White wine where 99.6% is correctly modeled s red wine. 15 are modeled as white wine 4883 are modeled as White wine where 99.6% is correctly modeled as the white wine

Linear regression:

Linear Regression is used to calculate the quality of the wine on the scale of 1-10. As per the analysis the quality of the wine can be calculated by the chemical components used in the wine. As per the given data the best quality of the wine is the White wine with the scale of 9 on the quality evaluation scale. The best quality of the red wine is 8 as the chemical constituents. This will be helpful for the business to estimate the quality of the wine and if so the quality is reducing then the sales may also reduce. So increase the sales the quality of the wine plays a major role. This Linear regression modeling helps to find the quality of the wine by selecting different chemical components. The major chemical components that I have considered chlorides, the alcohol content, density of wine and the PH of the wine. Since the wine is acidic in nature the PH of the wine are 4. The correlation coefficient is 0.587 which means the modeling is almost accurate for the company's recommendation.

Tableau:

The above modeling techniques help to understand the performance by calculating. But visual effects can give a better understanding about the drawbacks in the product and can also be circulated in required form. And as per the Tableau analysis the Red wine is the best quality as compared to the White wine. So, both the WEKA and the Tableau produces the same results, but the representation form is different.
All the models help the business needs but as per the results the **Logistic Regression** can best meet the business needs and for the visual analysis Tableau also provides best results.

DEPLOYMENT:

In any food products or the Wine, the quality and the components play a major role. So, these modeling techniques help the business to produce better quality of the wine or any food products. But presently I am working on the exports of **Vinho Verde wine** in Spain. So, these modeling techniques can be deployed by the Spain wine industry to export the best quality of Red wine or the White wine. As we see Spin is best exporter of Red wine.

These modeling techniques are very useful for the organization because there are no drawbacks found in the modeling techniques and the accuracy and the correct classification of the data in this modeling technique is 99.6 % which means that it is a good modeling technique and can be recommended to the organization.

References

Below are the references I have considered to complete the project.

http://www.sciencedirect.com/science/article/pii/S0167923609001377
http://archive.ics.uci.edu/ml/datasets/Wine+Quality
https://pdfs.semanticscholar.org/e6cf/6dc995ec0a0efc42c038d54d680a4f48529a.pdf