Data Mining Project: Wine Quality

By: Bright Ekeigwe
CIS 675

Abstract

In the original form of the dataset, two datasets were created, using red and white wine samples. Here, these two datasets have been combined into one dataset. 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)

Dataset: Wine Quality

Attributes: 11 numeric, 1 nominal

Class: Numeric

Instances: 6497

Missing Values: None

Abstract

The two datasets are related to red and white wine variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling, etc.)

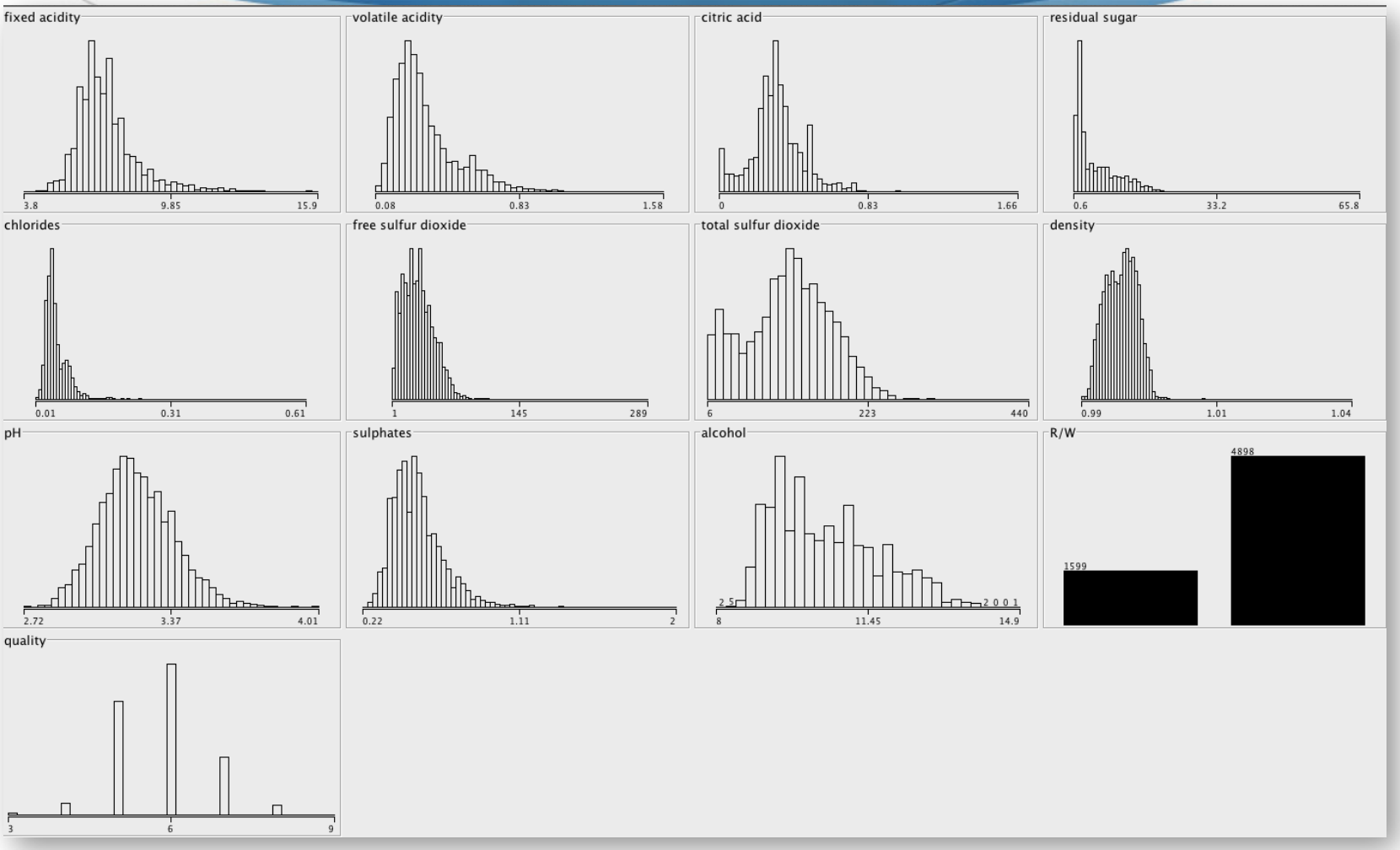
Attributes

1) Fixed Acidity, numeric
2) Volatile Acidity, numeric
3) Citric Acid, numeric
4) Residual Sugar, numeric
5) Chlorides, numeric
6) Free Sulfur, numeric
7) Total Sulfur, numeric
8) Density, numeric
9) pH, numeric
10) Sulphates, numeric
11) Alcohol, numeric
12) R/W, nominal – R (Red), W (White)

Class:

- Quality (Score between 0 and 10)

Visualize Data



Red Wine

Attribute (units)	Red wine		
	Min	Max	Mean
Fixed acidity (g(tartaric acid)/dm^3)	4.6	15.9	8.3
Volatile acidity (g(acetic acid)/dm^3)	0.1	1.6	0.5
Citric acid (g/dm^3)	0.0	1.0	0.3
Residual sugar (g/dm^3)	0.9	15.5	2.5
Chlorides (g(sodium chloride)/dm^3)	0.01	0.61	0.08
Free sulfur dioxide (mg/dm^3)	1	72	14
Total sulfur dioxide (mg/dm^3)	6	289	46
Density (g/cm^3)	0.990	1.004	0.996
pH	2.7	4.0	3.3
Sulphates (g(potassium sulphate)/dm^3)	0.3	2.0	0.7
Alcohol (vol.%)	8.4	14.9	10.4

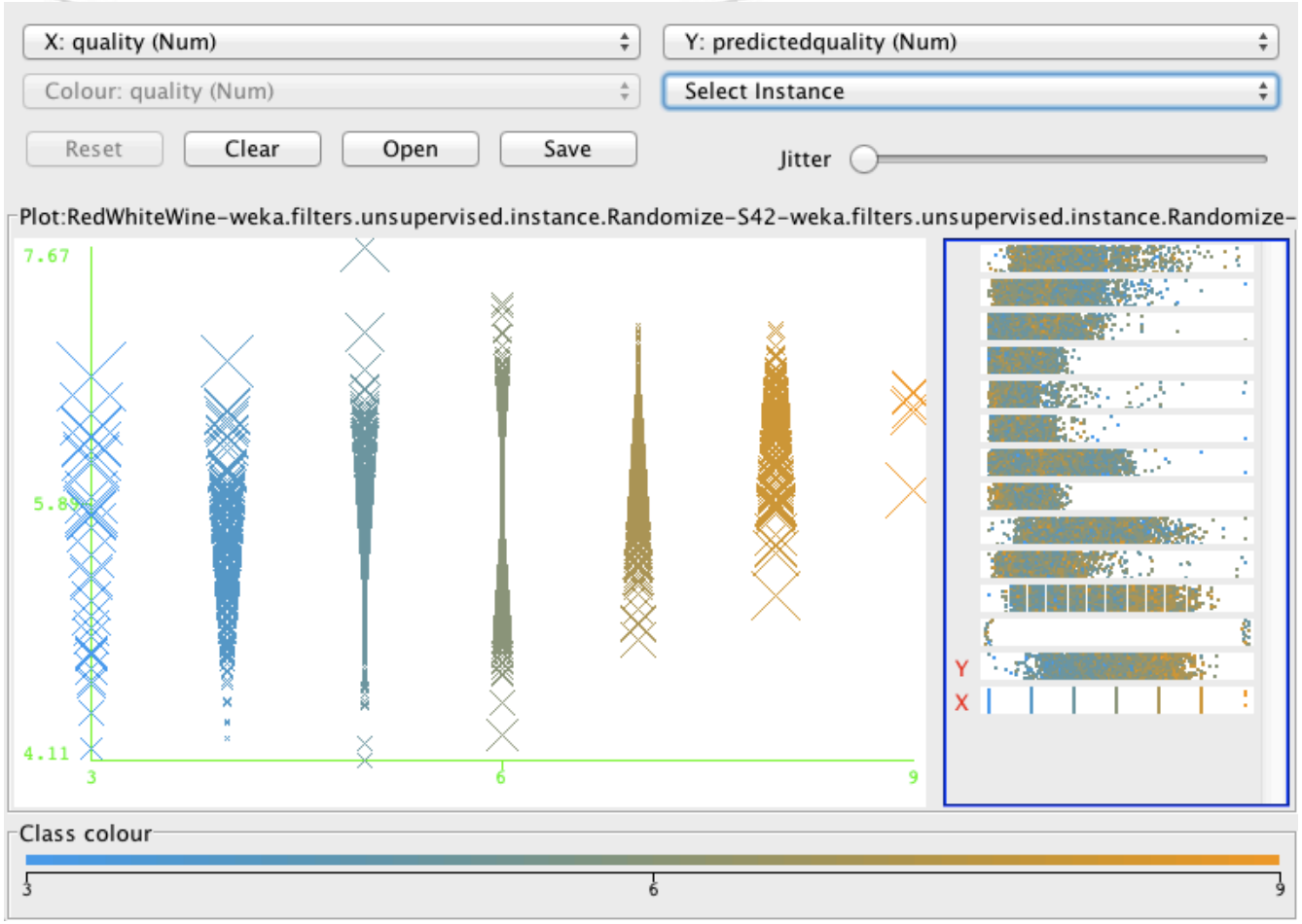
White Wine

Attribute (units)	White wine		
	Min	Max	Mean
Fixed acidity (g(tartaric acid)/dm^3)	3.8	14.2	6.9
Volatile acidity (g(acetic acid)/dm^3)	0.1	1.1	0.3
Citric acid (g/dm^3)	0.0	1.7	0.3
Residual sugar (g/dm^3)	0.6	65.8	6.4
Chlorides (g(sodium chloride)/dm^3)	0.01	0.35	0.05
Free sulfur dioxide (mg/dm^3)	2	289	35
Total sulfur dioxide (mg/dm^3)	9	440	138
Density (g/cm^3)	0.987	1.039	0.994
pH	2.7	3.8	3.1
Sulphates (g(potassium sulphate)/dm^3)	0.2	1.1	0.5
Alcohol (vol.%)	8.0	14.2	10.4

Linear Regression

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


Linear Regression Model

quality =

      0.0823 * fixed acidity +
     -1.4714 * volatile acidity +
      0.0625 * residual sugar +
     -0.8011 * chlorides +
      0.0049 * free sulfur dioxide +
     -0.0014 * total sulfur dioxide +
   -104.4192 * density +
      0.5038 * pH +
      0.7185 * sulphates +
      0.221  * alcohol +
     -0.3651 * R/W=W +
    105.2617

Time taken to build model: 0.52 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.5412
Mean absolute error                      0.57
Root mean squared error                  0.7343
Relative absolute error                 83.1287 %
Root relative squared error             84.0756 %
Total Number of Instances             6497
```

Linear Regression Plot



Decision Tree

```
Test mode:10-fold cross-validation

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

Decision Stump

Classifications

alcohol <= 10.649999999999999 : 5.528453181583031
alcohol > 10.649999999999999 : 6.244393766628658
alcohol is missing : 5.818377712790519


Time taken to build model: 0.8 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3963
Mean absolute error                      0.6605
Root mean squared error                  0.8017
Relative absolute error                 96.3354 %
Root relative squared error             91.7996 %
Total Number of Instances             6497
```

Decision Tree

```
Number of Leaves  :     31

Size of the tree :      61


Time taken to build model: 2.85 seconds

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

Correctly Classified Instances        6471               99.5998 %
Incorrectly Classified Instances        26                0.4002 %
Kappa statistic                          0.9892
Mean absolute error                      0.0076
Root mean squared error                  0.0617
Relative absolute error                  2.0491 %
Root relative squared error             14.3154 %
Total Number of Instances             6497

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.986     0.001       0.997     0.986     0.992      0.995    R
                 0.999     0.014       0.996     0.999     0.997      0.995    W
Weighted Avg.    0.996     0.011       0.996     0.996     0.996      0.995

=== Confusion Matrix ===

    a    b   <-- classified as
 1577   22 |    a = R
    4 4894 |    b = W
```

Decision Tree