Data mining project

Alicja Czajkowska

magda leszczynska

Paulina Luty

Anna Skowron

HDSDA

SEP2

Contents

[**Problem Statement** 2](#_Toc512181717)

[**Data set description and pre-processing** 2](#_Toc512181718)

[**Tools used and data overview** 4](#_Toc512181719)

[**Construction of data mining model** 7](#_Toc512181720)

[**Models used** 7](#_Toc512181721)

[**Naïve Bayes** 7](#_Toc512181722)

[**Decision** **Tree** 9](#_Toc512181723)

[**Clustering** 9](#_Toc512181724)

[**Conclusions** 9](#_Toc512181725)

[**References** 9](#_Toc512181726)

# **Problem Statement**

Sdfsdfsdfsfffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa

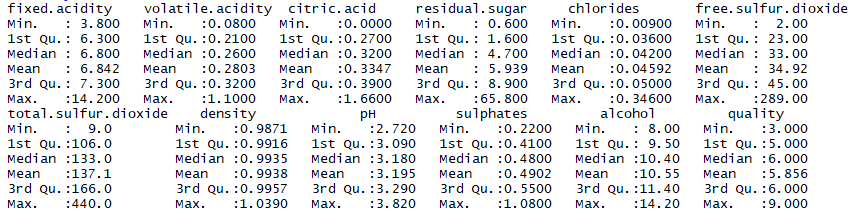
# **Data set description and pre-processing**

We chose “Wine quality - white” as a data set for our project. Data set contains a samples wine testing of Portuguese *vinho verde*. Data set is available to download as a .csv file on <http://www3.dsi.uminho.pt/pcortez/wine/> [Cortez et al., 2009][[1]](#footnote-1). Original data set consists of 4899 rows and the following 12 columns:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| fixed acidity | - | real number | - | measure of tartaric acid [g/dm3] |
| volatile acidity | - | real number | - | amount of acetic acid [g/dm3] |
| citric acid | - | real number | - | amount of citric acid [g/dm3] |
| residual sugar | - | real number | - | amount of residual sugar [g/dm3] |
| Chlorides | - | real number | - | amount of sodium chloride [g/dm3] |
| free sulfur dioxide | - | real number | - | measure of sulfur dioxide [mg/dm3] |
| total sulfur dioxide | - | real number | - | measure of sulfur dioxide [mg/dm3] |
| Density | - | real number | - | [g/cm3] |
| pH | - | real number | - | potential of hydrogen of wine [mole] |
| sulphates | - | real number | - | potassium sulphate [g/dm3] |
| Alcohol | - | real number | - | [vol.%] |
| Quality | - | ordinal | - | number from range 0 -10 |

First 11 attributes are the input and 12th - “quality” is an output attribute. There is no missing attributes. All are numeric. There was 898 duplicates in the data set which we removed and that left us with 4000 of rows of data.

We ran some descriptive statistics in R to see how the data is structured and to get a better understanding of it.



Acids (fixed, citric and volatile) – important components of wine, occurring in grapes and during the winemaking process. Acids control level of pH and are also antibacterial.

Residual sugar – natural grape sugar used in winemaking that is left after the fermentation process. It determines sweetness of the wine. Our wines have from 0.600 to 65.800 mg/dm3.

Chlorides – group of preservatives used for different purposes in wine making.

Sulphates (including total and free sulphur oxide – preservative and also a cleaning product. Although it is not a desirable ingredient, it is not possible to make a wine without it as some amount is produced during the winemaking process and some amount is required because of their antioxidant and antibacterial properties. It is a common allergen. In tested wines the amount of total sulphur dioxide has a big range – it varies from 9 to 440 mg/dm3.

Density - represents the concentration of dissolved sugar, in weight percent (wt%).

pH - potential of hydrogen value measures acidity in wine. It ranges from 2.720 to 3.820 in our data set. The lower pH value is the higher acidity level is observed.

Alcohol – percentage of alcohol content is an important factor for potential customer, in our dataset alcohol level is from 8 to 14.20 % with a mean of 10.55%.

Quality – the only original output in our database and only attribute that is based on sensory assessor’s’ rating (each wine was tested by a minimum of three assessors using blind tests and quality final quality score is a median of these tests) [Cortez et al., 2009][[2]](#footnote-2). Scale available to assessors was 0 (very bad) to 10 (excellent) but in the data set quality varies from 3 to 9

For the purpose of this project we created three more output attributes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sugar /PH ratio | - | real number | - | ratio between sugar and pH |
| level of preservatives | - | polynomial | - | low, medium, high |
| alcohol content | - | polynomial | - | low, medium, high |

We used the following classification rules to establish our output attributes:

For sugar /pH ratio:

Sugar/PH ratio

For level of preservatives

Low– If sulphates <=0.45 and chlorides <= 0.045

Medium – If sulphates <=0.6 and chlorides <= 0.06

High – Others

For alcohol content:

High – above - 11%

Medium – 9-11%

Low – 9% and less

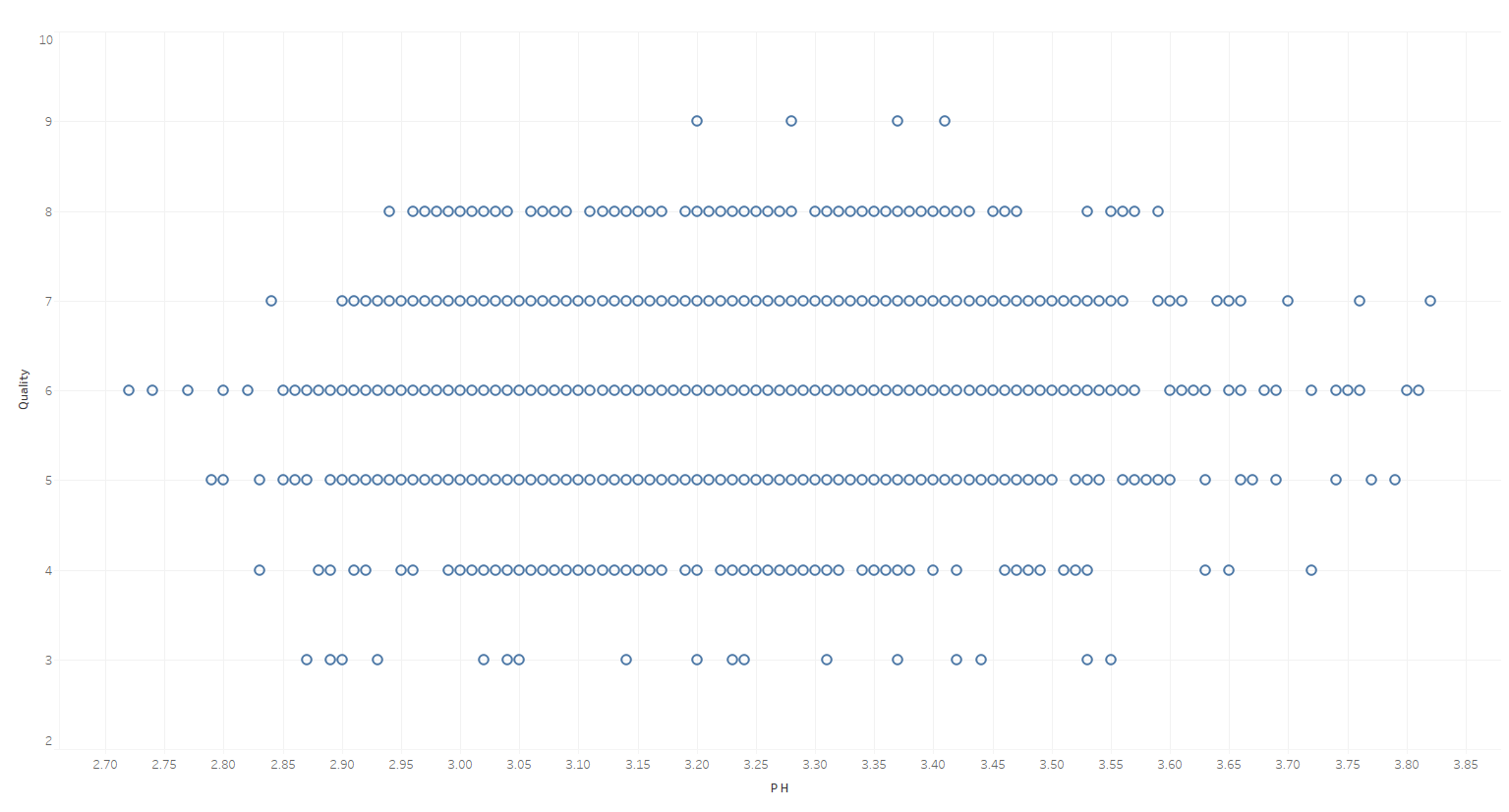
We decided to use sugar /pH ratio as the wine should not be either too sweet or too sour. We expect to find some correlation between the ratio and quality.

There are two main preservatives in wine: sulphates and chlorides. We want to test whether the level of preservatives has any correlation to quality. Sulfur dioxide (SO2) is widely use in winemaking mainly because of its anti-oxidative and anti-microbial properties in wine. It is also used for cleaning purposes at the wineries[[3]](#footnote-3). We suspect that too much of preservatives will have negative impact on quality.

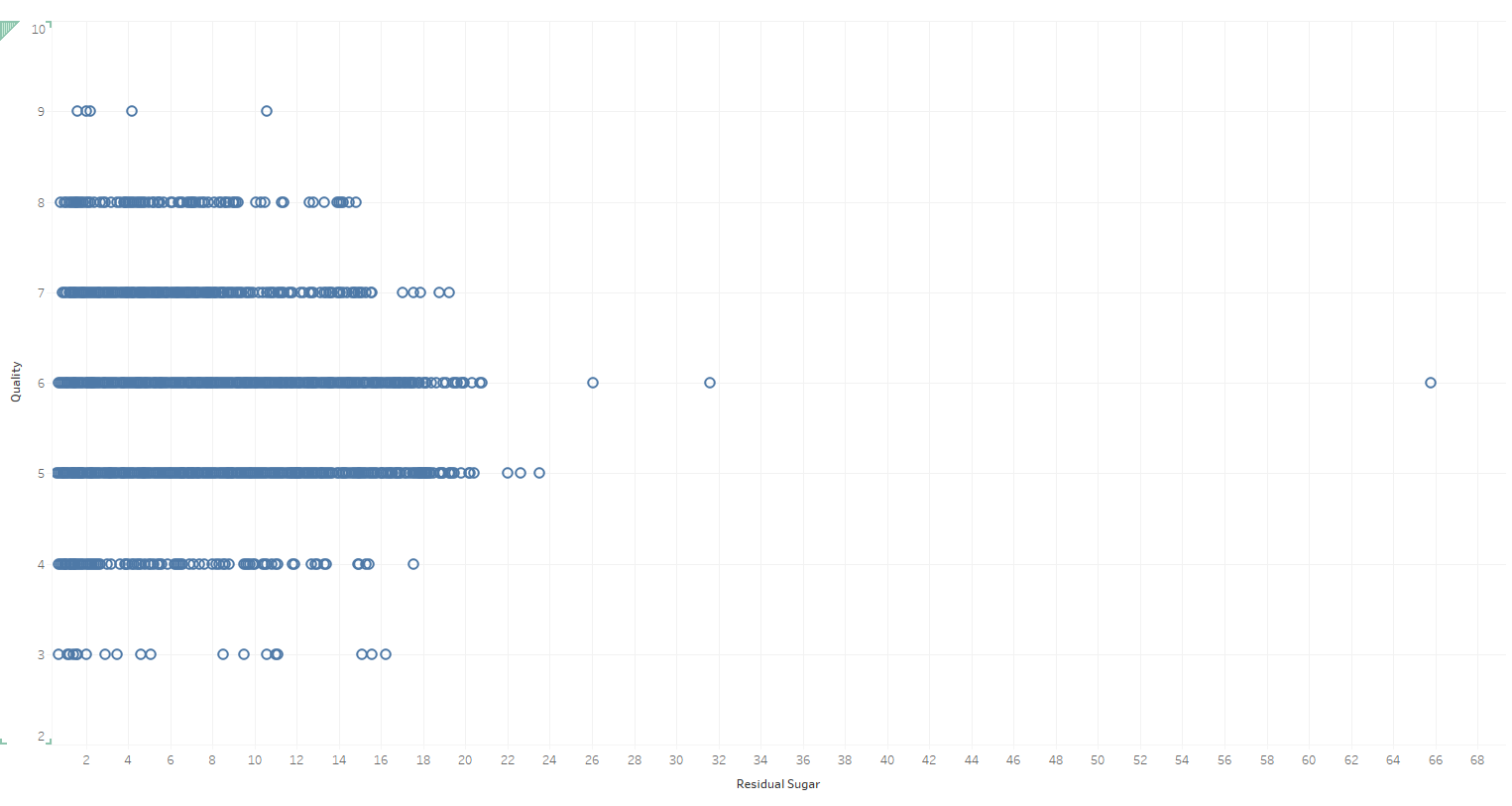
For the alcohol content, we wanted to group wines into three categories based on their alcohol level.

# **Tools used and data overview**

We used Tableau, Excel, Weka, RapidMiner, R and Python in our project.

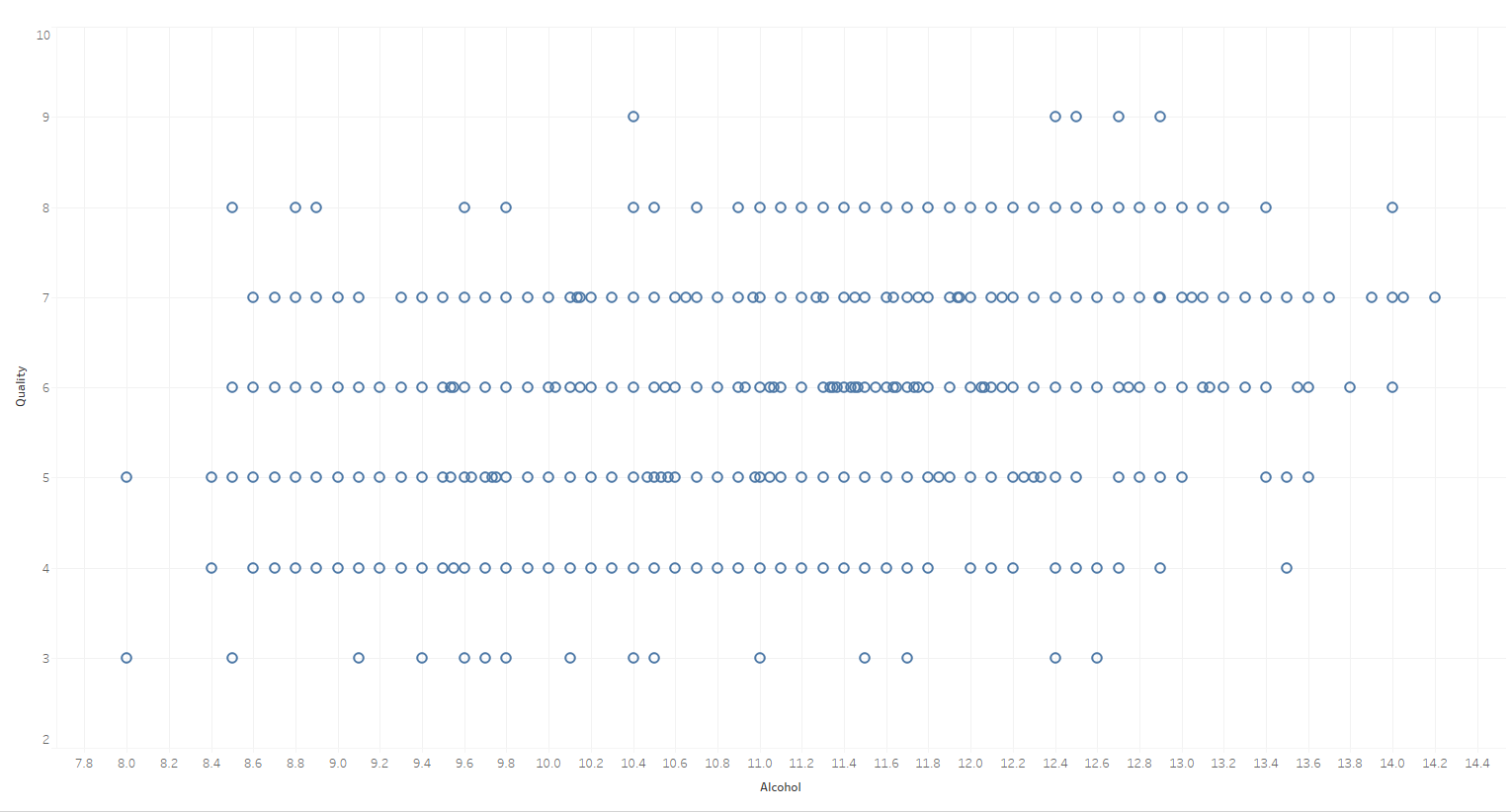
First, we have ran some visualisations to see how data is distributed and checked for outliers and possible correlations.

*Graph 1. PH to quality scatter plot*

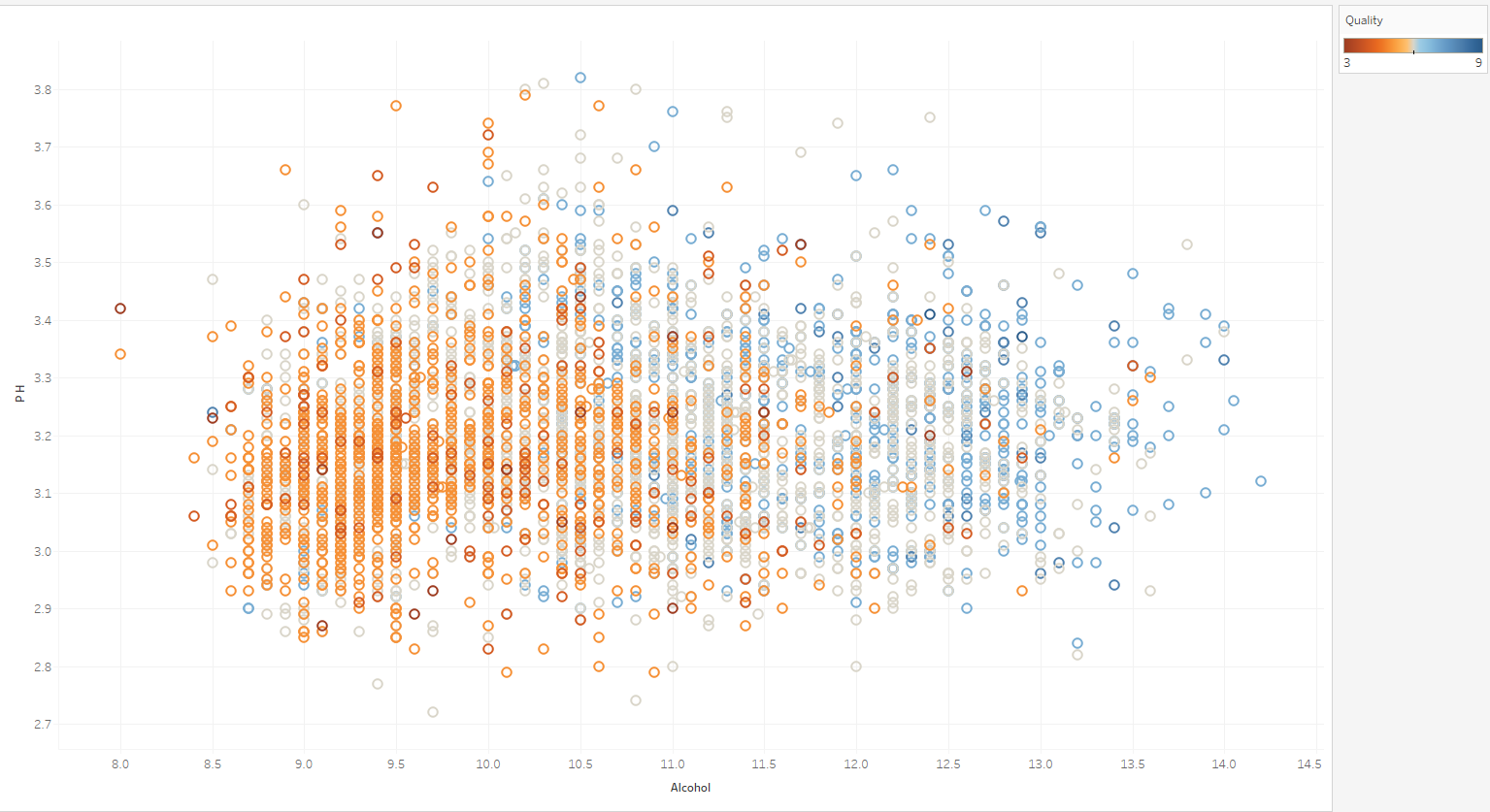


*Graph 2. Sugar to quality scatter plot*

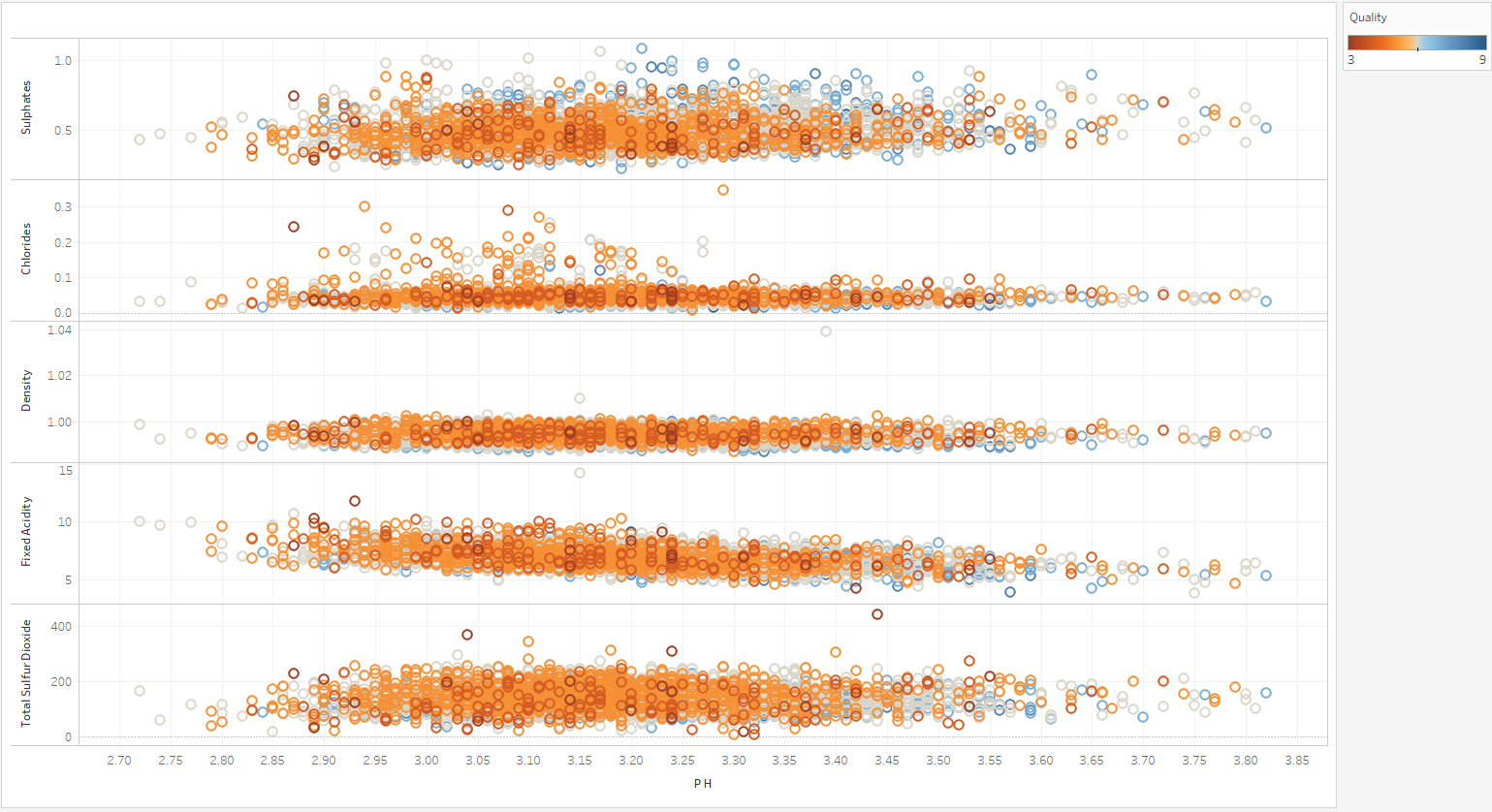
There is an outlier on this graph - a sugar at level of 65.8. We decided to leave that outlier in our model as the amount of sugar in white wines can vary from 0 to 220 grams per litre[[4]](#footnote-4). It is possible that one of the tested wines had that amount of sugar.



*Graph 3. Alcohol to Quality Scatter plot*



*Graph 4. Alcohol to pH with quality as a factor scatter plot*



*Graph 5. Sulfur dioxide, fixed acidity, density, chlorides and sulphates to pH scatter plots*

From the scatter plots we learn that any deviation from the mean it’s taking points from wine quality (excluding alcohol content – if less alcohol in wine, then it is less probable that it will good quality wine).

# **Construction of data mining model**

# **Models used**

We will use decision trees, clustering and association models.

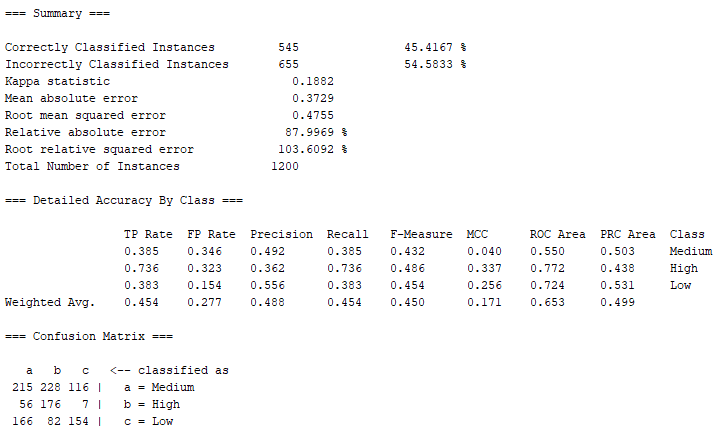
Decision tree is a classification method and is also a predictive method. Classification methods use existing data to create a model that will allow to classify new data. In our project we will use decision tree to predict xxxxxx.

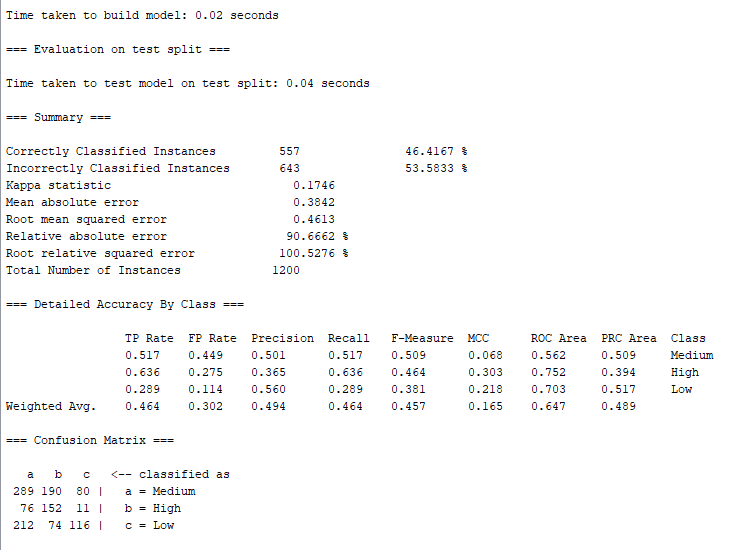
Clustering is a descriptive technique that finds groups of observations (clusters) that share similar characteristics in a data set.

## **Naïve Bayes**

Naïve Bayesmodel was run in Weka on using the dataset with 4,000 instances (split 70 /30 - 70.0% of data is for a training, remaining 30% is for testing purposes) and 12 attributes (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, Quality and Alcohol content).

The accuracy of that model was only 45.4% (Fig 1.1).

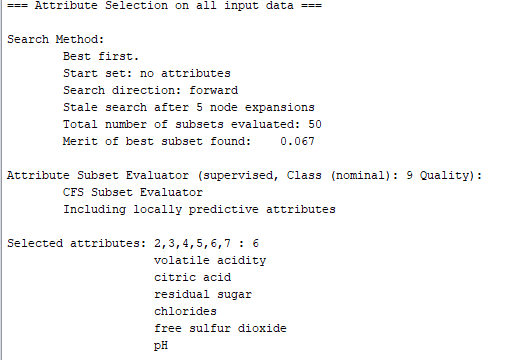




*Figure 1.1 Naïve Bayes classifiers’ output comparison*

The same model was run after removing total sulfur dioxide attribute (correlated to free sulfur dioxide) and slight improvement in accuracy level was observed – 46.4%.

In the next step we used the correlation-based feature selection (CFS) algorithm in Weka to further eliminate the correlated redundant features from the dataset (Fig. 1.2).

**

*Figure 1.2. Attributes selected by applying the CFS algorithm in Weka*

However, after executing the Naïve Bayes model in Weka using only attributes selected by CFS algorithm (volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, pH and Quality) we did not observe any major improvement in accuracy. The accuracy for this model was 45.5%.

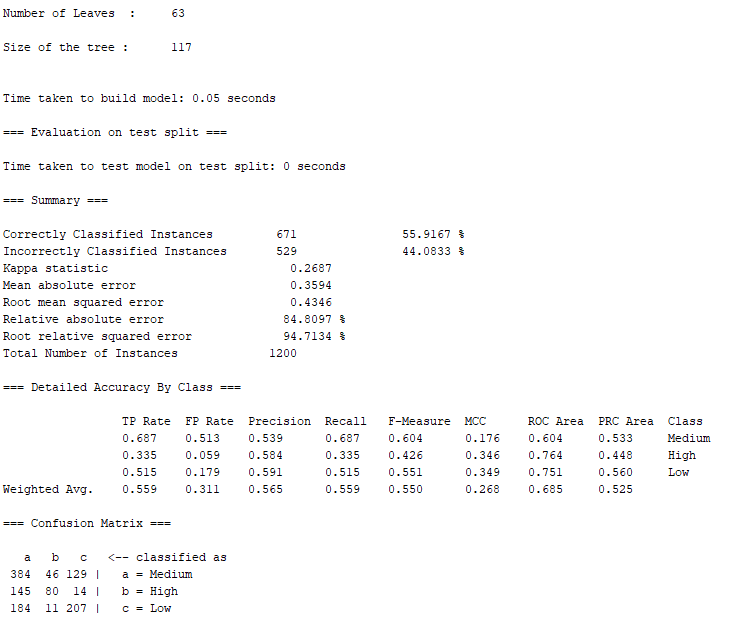
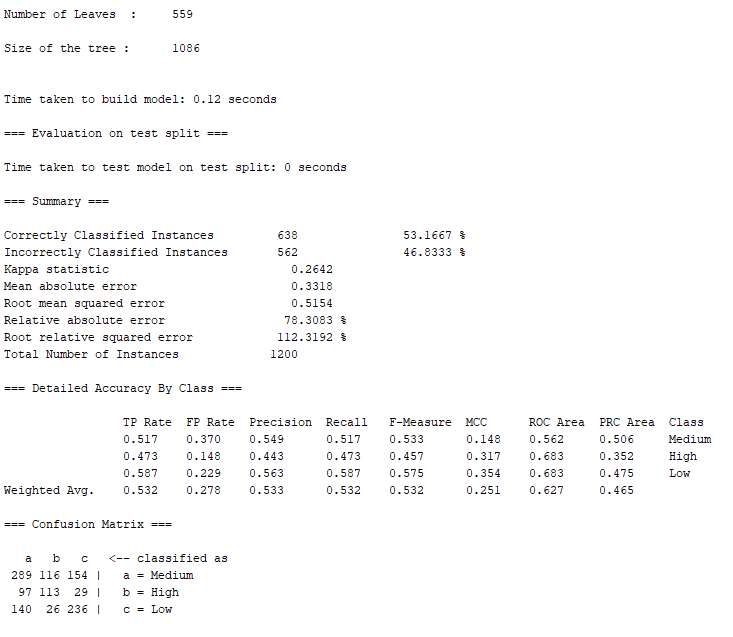
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Class** | |  |  |
| **Attribute** | | **High** | **Medium** | **Low** |
|  | | 0.21 | 0.45 | 0.34 |
| **fixed acidity** | |  |  |  |
| mean | | 6.6968 | 6.8267 | 6.9602 |
| std. dev. | | 0.7846 | 0.8485 | 0.9121 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.1552 | 0.1552 | 0.1552 |
| **volatile acidity** | |  |  |  |
| mean | | 0.2689 | 0.2615 | 0.3121 |
| std. dev. | | 0.0946 | 0.0895 | 0.1167 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.0082 | 0.0082 | 0.0082 |
| **citric acid** | |  |  |  |
| mean | | 0.3306 | 0.3394 | 0.3329 |
| std. dev. | | 0.0828 | 0.1206 | 0.1436 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.0193 | 0.0193 | 0.0193 |
| **residual sugar** | |  |  |  |
| mean | | 4.6206 | 6.0033 | 6.6776 |
| std. dev. | | 3.7365 | 4.9864 | 5.1673 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.211 | 0.211 | 0.211 |
| **chlorides** | |  |  |  |
| mean | | 0.0374 | 0.0452 | 0.0522 |
| std. dev. | | 0.0108 | 0.0209 | 0.0288 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.0021 | 0.0021 | 0.0021 |
| **free sulfur dioxide** | | | | |
| mean | | 34.1524 | 35.313 | 34.7731 |
| std. dev. | | 14.2492 | 15.7009 | 20.5611 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 2.1908 | 2.1908 | 2.1908 |
| **pH** | |  |  |  |
| mean | | 3.2288 | 3.1951 | 3.1728 |
| std. dev. | | 0.1536 | 0.1512 | 0.147 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.0108 | 0.0108 | 0.0108 |
| **sulphates** | |  |  |  |
| mean | | 0.5 | 0.4921 | 0.4816 |
| std. dev. | | 0.1345 | 0.1118 | 0.1008 |
| weight sum | | 834 | 1806 | 1360 |
| precision | | 0.011 | 0.011 | 0.011 |
| **Alcohol content** | |  |  |  |
| High | | 300 | 564 | 473 |
| Medium | | 467 | 1008 | 777 |
| Low | | 70 | 237 | 113 |
| [total] | | 837 | 1809 | 1363 |

*Table 1.1 Naïve Bayes classifier model*

It would appear that the quality is decreasing with higher level of fixed acidity, residual sugar and chlorides and it is improved with higher pH level. It appears that higher quality wines have lower alcohol level.

## **Decision** **Tree**

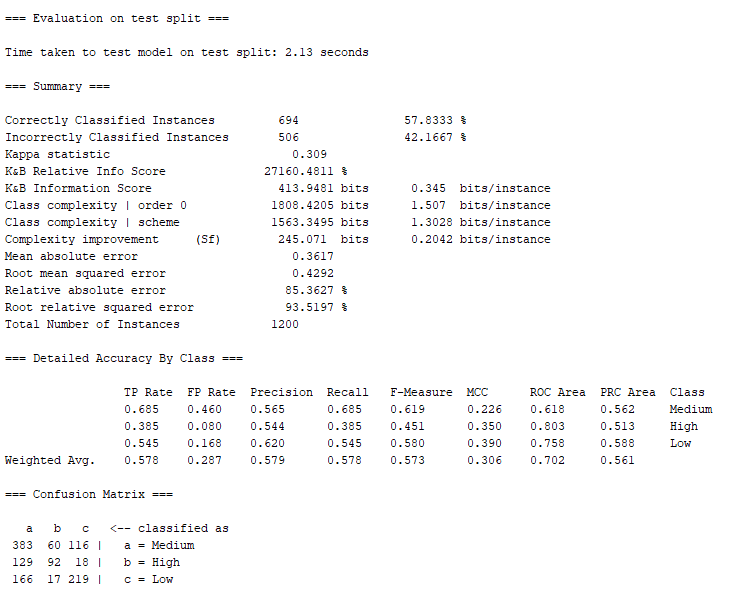
Decision Tree J48 model was executed in Weka using the dataset with 4,000 instances (split 70.0% train, remainder test) and 12 attributes (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, Quality and Alcohol content). The first Decision Tree model was executed with default settings in Weka (batch size 100, confidence factor 0.25, minimum number of instances per leaf 2). The accuracy of that model was 53.2%. The model was simplified by increasing the minimum number of instances per leaf to 28 and it improved its accuracy to 55.9%.



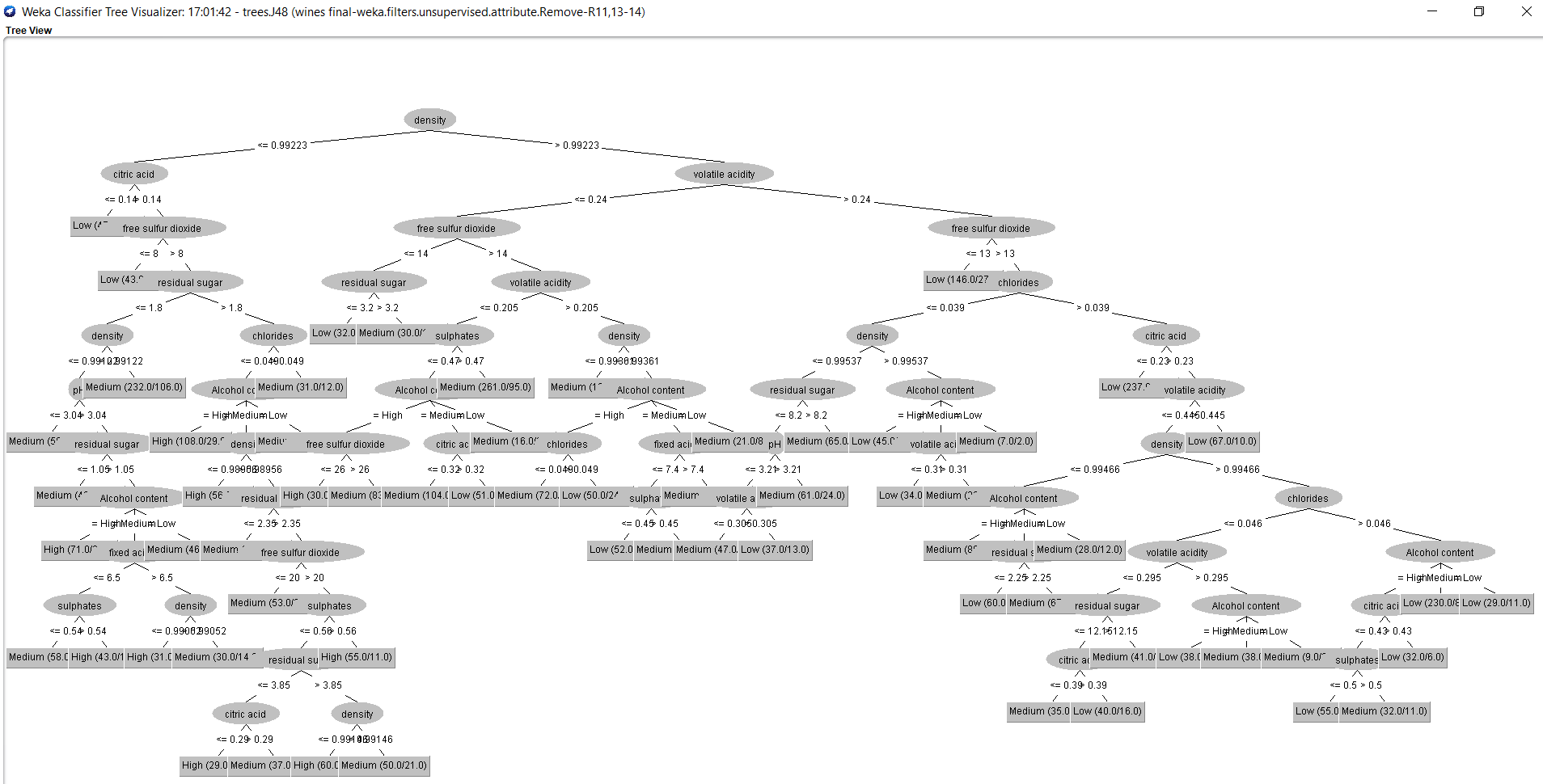
*Figure 1.3. Decision Tree J48 models’ comparison*

We also used Support Vector Machine model in Weka – SMO to predict the quality attribute. The accuracy of that model was 57.8% (Fig. 1.5).

Support vector machine algorithm transforms training data into a higher dimension, where it searches for “decision boundary” (linear optimal separating hyperplane) that separates the data from one class from another. To find this hyperplane SVM model uses support vectors (“essential” training tuples) and margins (deﬁned by the support vectors).



*Figure 1.4. SMO model Weka – evaluation output*



*Figure 1.5. Decision Tree J48 model in Weka*

The best accuracy level 67.4% (Fig. 1.6) was achieved using the Random Forest model using the dataset with 4,000 instances (split 70.0% train, remainder test) and 11 attributes (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, density, pH, sulphates, Quality and Alcohol content).

This is not surprising as Random Forest is an ensemble method where the decision is based on the outcome of a number (forest) of decision tree classifiers. The individual decision trees are generated using a random selection of attributes at each node to determine the split. During classiﬁcation, each tree votes and the most popular class is returned.

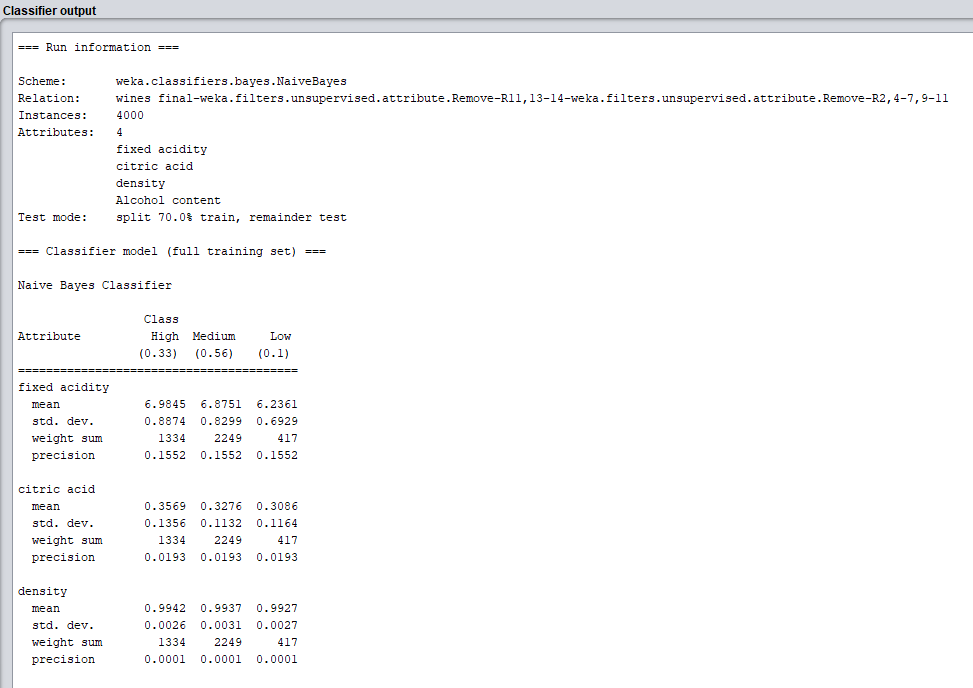
## 

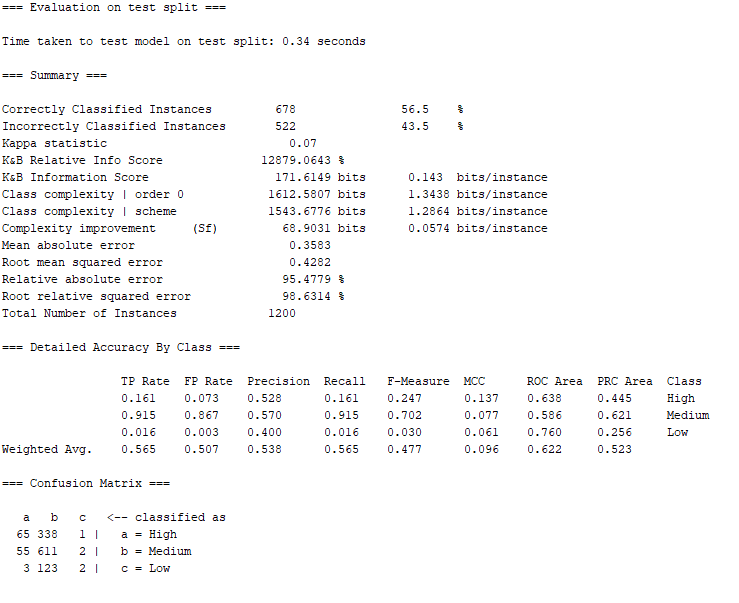
*Figure 1.6. Random Forest model in Weka*

**Classification Models – alcohol content**

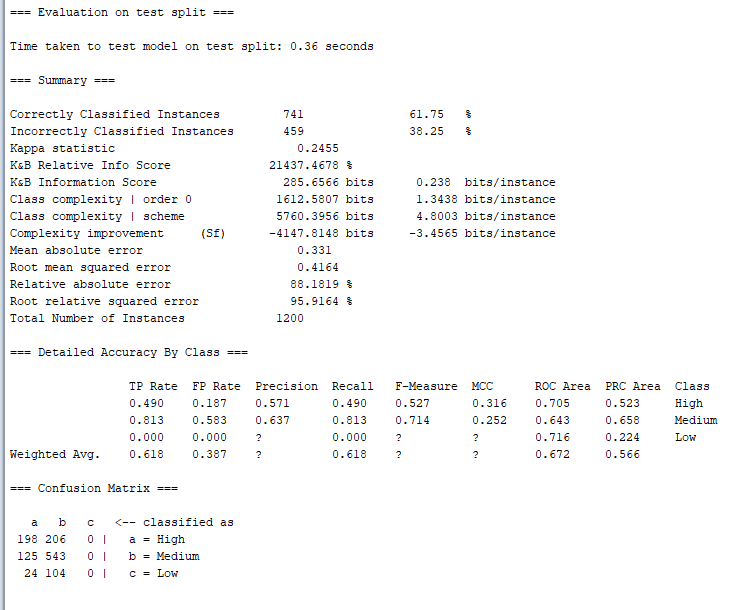
We also used **Naïve Bayes and Decision Tree** models to review alcohol content. The original alcohol values were discretised: values lower than 9% were classed as low, 9 to 11% as medium and above 11% as high. After running the correlation-based feature selection (CFS) algorithm in Weka to we eliminated all attributes but fixed acidity, citric acid and density to classify alcohol content. The accuracy of the model was 56.5% (Fig. 1.7). An interesting insight learnt from this model is that alcohol content seems to be higher in wines with higher fixed acidity levels and amount of citric acid and higher density.

Decision Tree (pruned J48) model was also executed using 12 attributes (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, Quality and Alcohol content) with minimum number of instances per leaf 28. The accuracy of that model was 61.7% (Fig. 1.8)

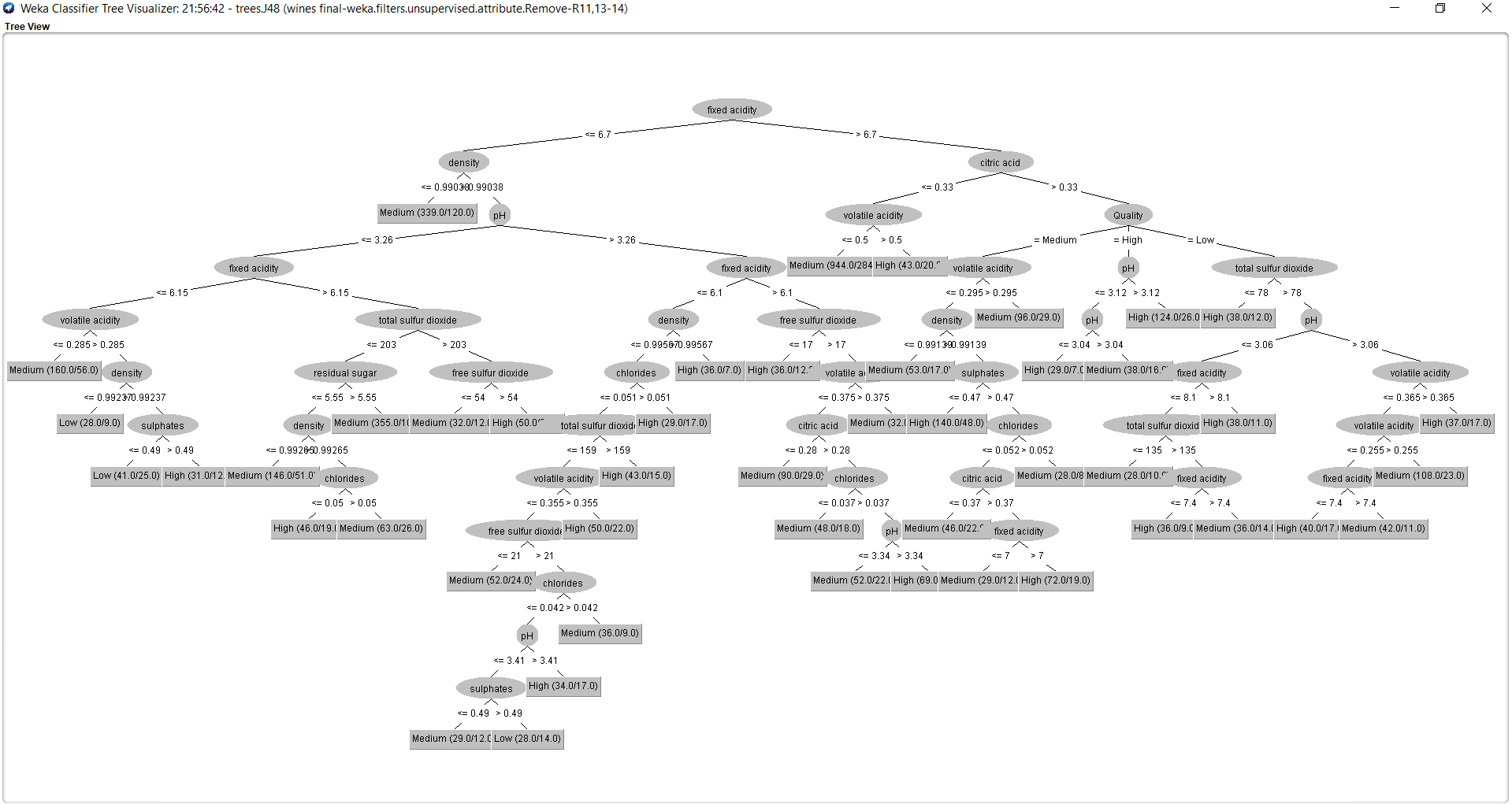




*Figure 1.7. Naïve Bayes model – alcohol content*



*Figure 1.8. Decision Tree J48 – alcohol content prediction – evaluation output Weka*



*Figure 1.9. Decision Tree J48 model – alcohol content prediction (Weka)*

# **Clustering**

# **Conclusions**

# **References**

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

2. Sources

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

**Bibliography**

1. Cortez P., Cerderira A., Almeida F., Matos T. and Reis J. (2009) ’Modeling wine preferences by data mining from physicochemical properties’. Decision Support Systems, 47 (2009): pp. 547 – 533.
2. <https://www.wardsci.com/www.wardsci.com/images/Chemistry_of_Wine.pdf> [Accessed 20 April 2018]
3. <http://waterhouse.ucdavis.edu/whats-in-wine> [Accessed 20 April 2018]
4. <https://winobrothers.com/2011/10/11/sulfur-dioxide-so2-in-wine/> [Accessed 20 April 2018]

1. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

   Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems>, Elsevier, 47(4):547-553. ISSN: 0167-9236. Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016 [Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf [bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib [↑](#footnote-ref-1)
2. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, Page 6 [↑](#footnote-ref-2)
3. <https://winobrothers.com/2011/10/11/sulfur-dioxide-so2-in-wine/> [↑](#footnote-ref-3)
4. http://winefolly.com/review/sugar-in-wine-chart/ [↑](#footnote-ref-4)