Data mining project

Alicja Czajkowska

magda leszczynska

Paulina Luty

Anna Skowron

HDSDA

SEP2

Contents

[**Problem Statement** 2](#_Toc511629353)

[**Data set description** 2](#_Toc511629354)

[**Tools, Data pre-processing and formatting** 2](#_Toc511629355)

[**Models used** 2](#_Toc511629356)

[**Construction of data mining model** 2](#_Toc511629357)

[**Decision** **Tree** 3](#_Toc511629358)

[**Clustering** 3](#_Toc511629359)

[**Conclusions** 3](#_Toc511629360)

[**References** 3](#_Toc511629361)

# **Problem Statement**

Sdfsdf

# **Data set description**

We chose “Wine quality - white” as a data set for our project. 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). 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 | - | amount of |
| 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

Volatile acidity relates to wine spoilage

Sulfites or sulfur dioxide is a fruit preservative commonly used in wine production because of its anti-oxidative and anti-microbial properties. It is also used as a cleaning agent for barrels and winery facilities. Free sulfur dioxide (not associated with wine molecules) is the protection against microbes and oxidation. Bound sulfur dioxide (associated with wine molecules) has already done its work and cannot be useful any longer in this context ([www.winobrothers.com](http://www.winobrothers.com)).

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

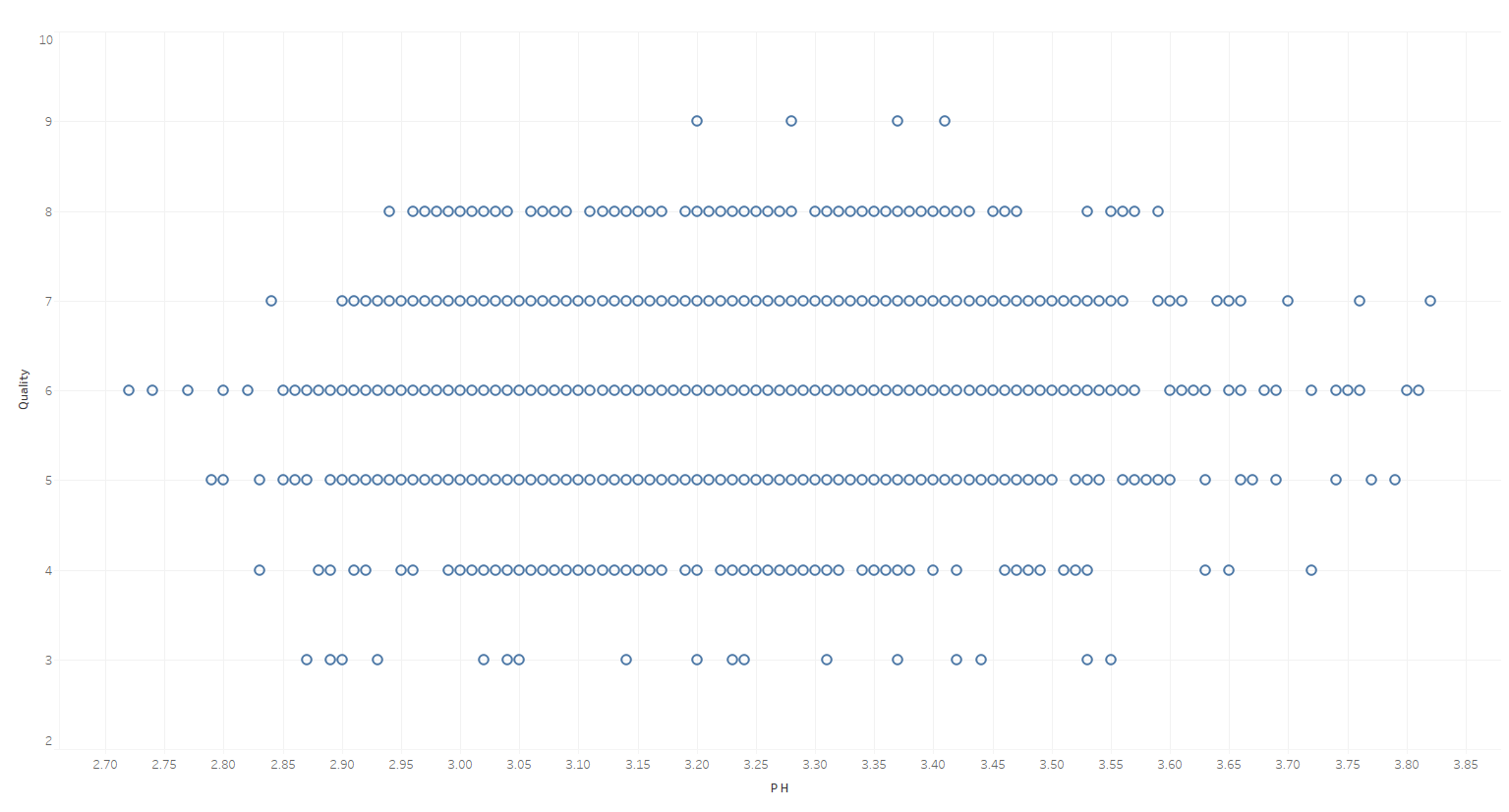
# **Tools, Data pre-processing and formatting**

We will use Tableau, Excel, RapidMiner, R and Python in our project.

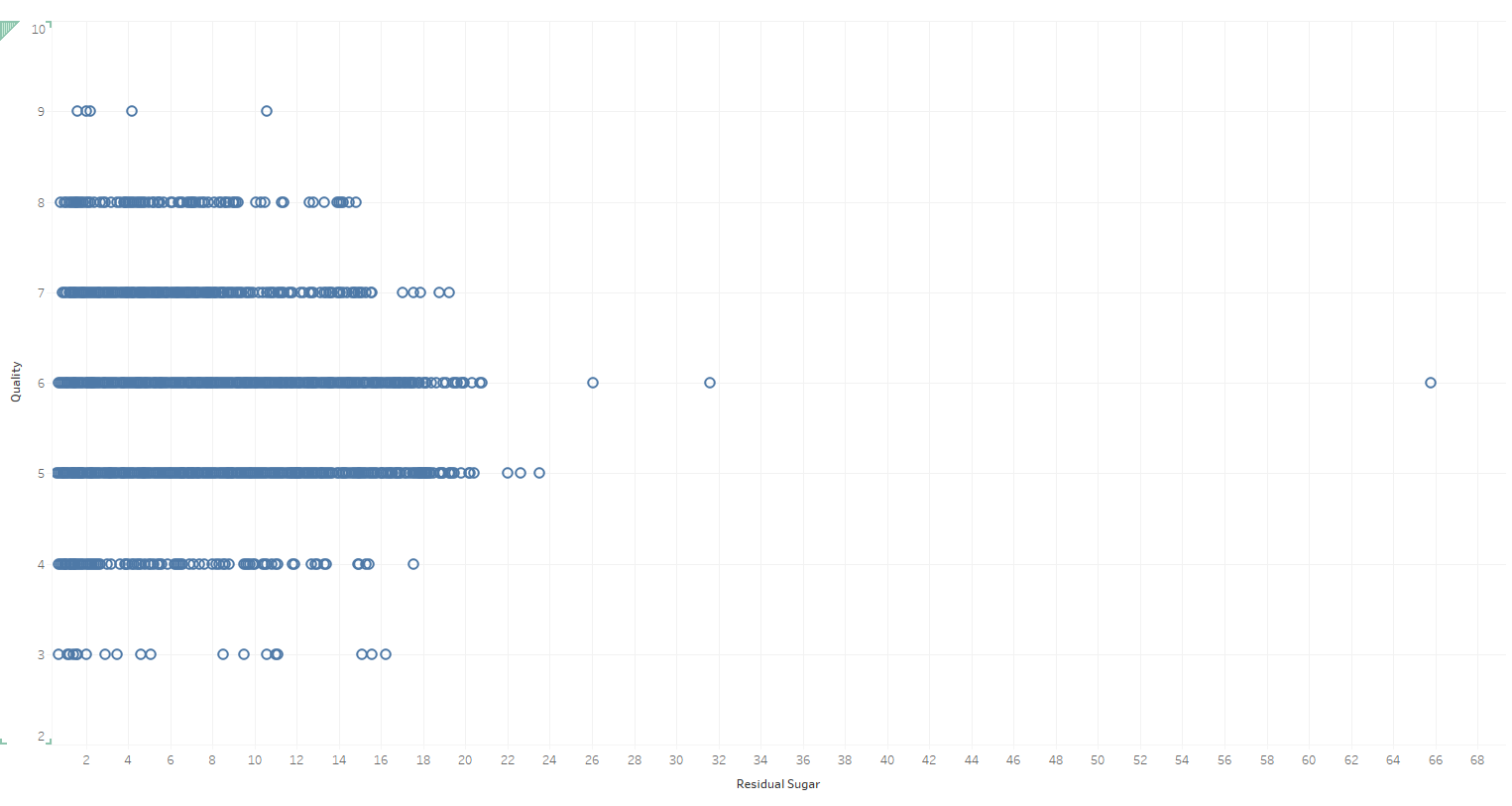
First, we checked the data set for NANs and duplications. There were no NAN values in our dataset. All duplicates have been carefully checked removed if occur.

We run some visualisation to see any outliers and correlation in this same time.

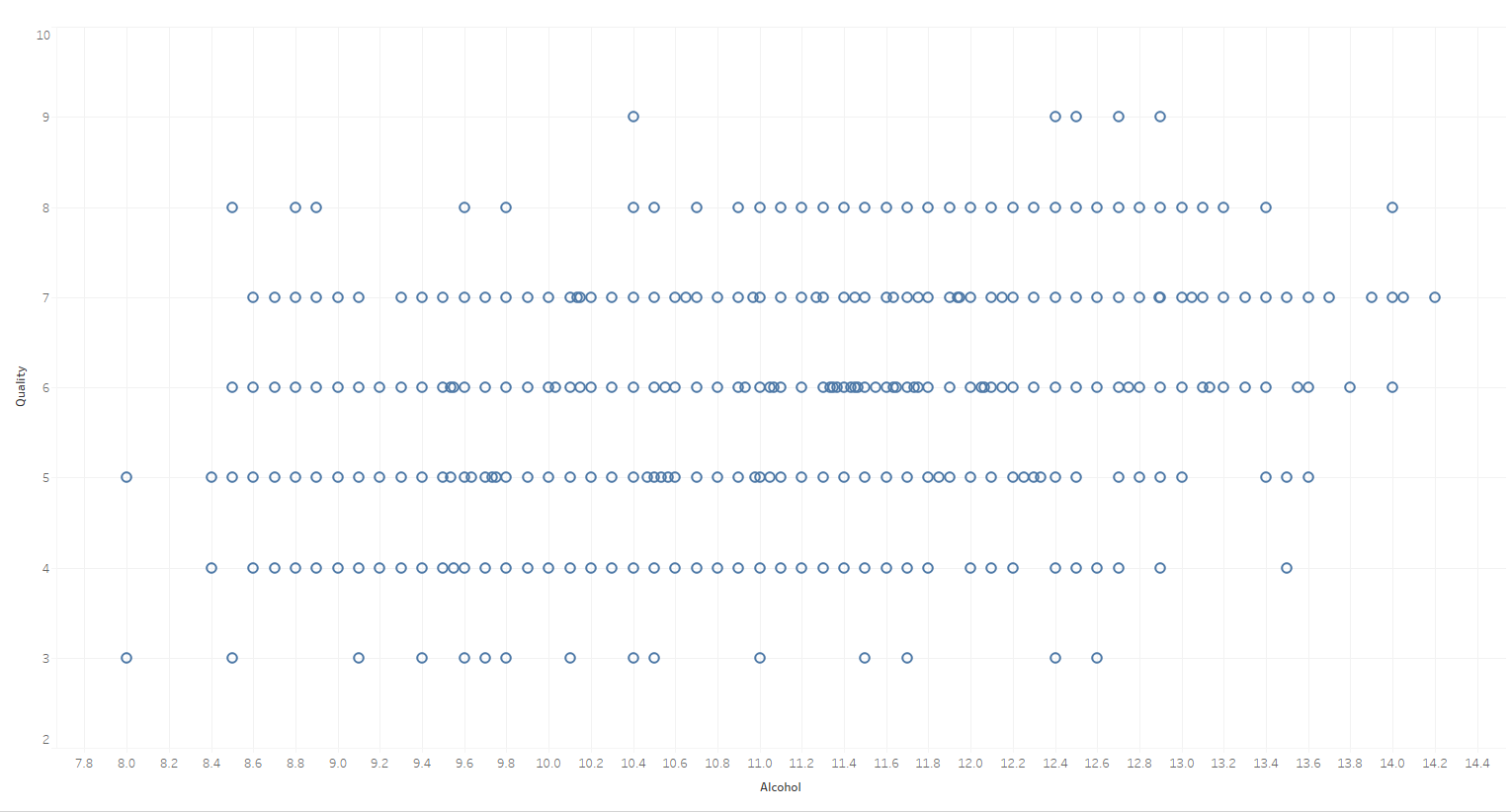
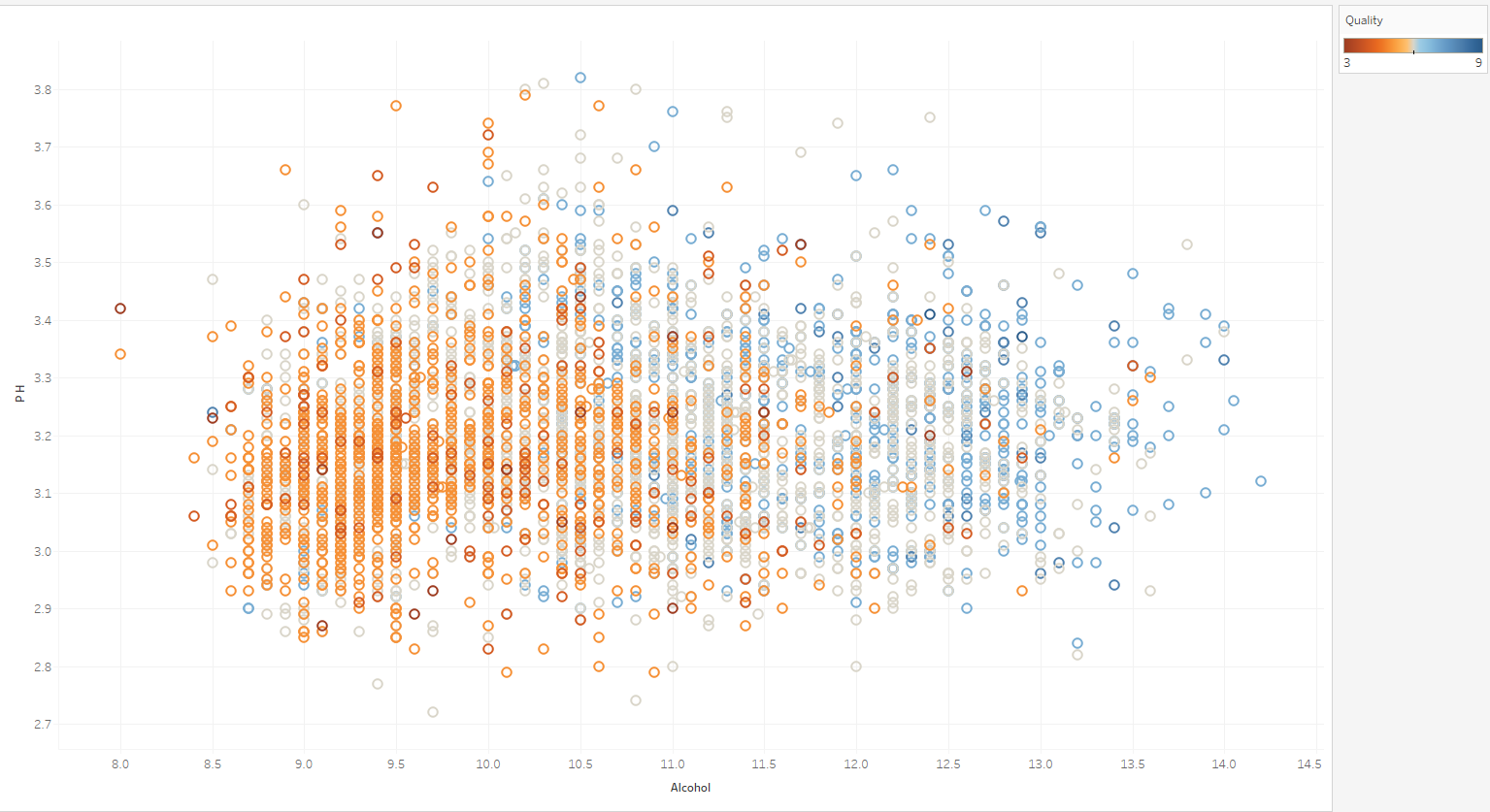
1. PH to Quality scatter graph

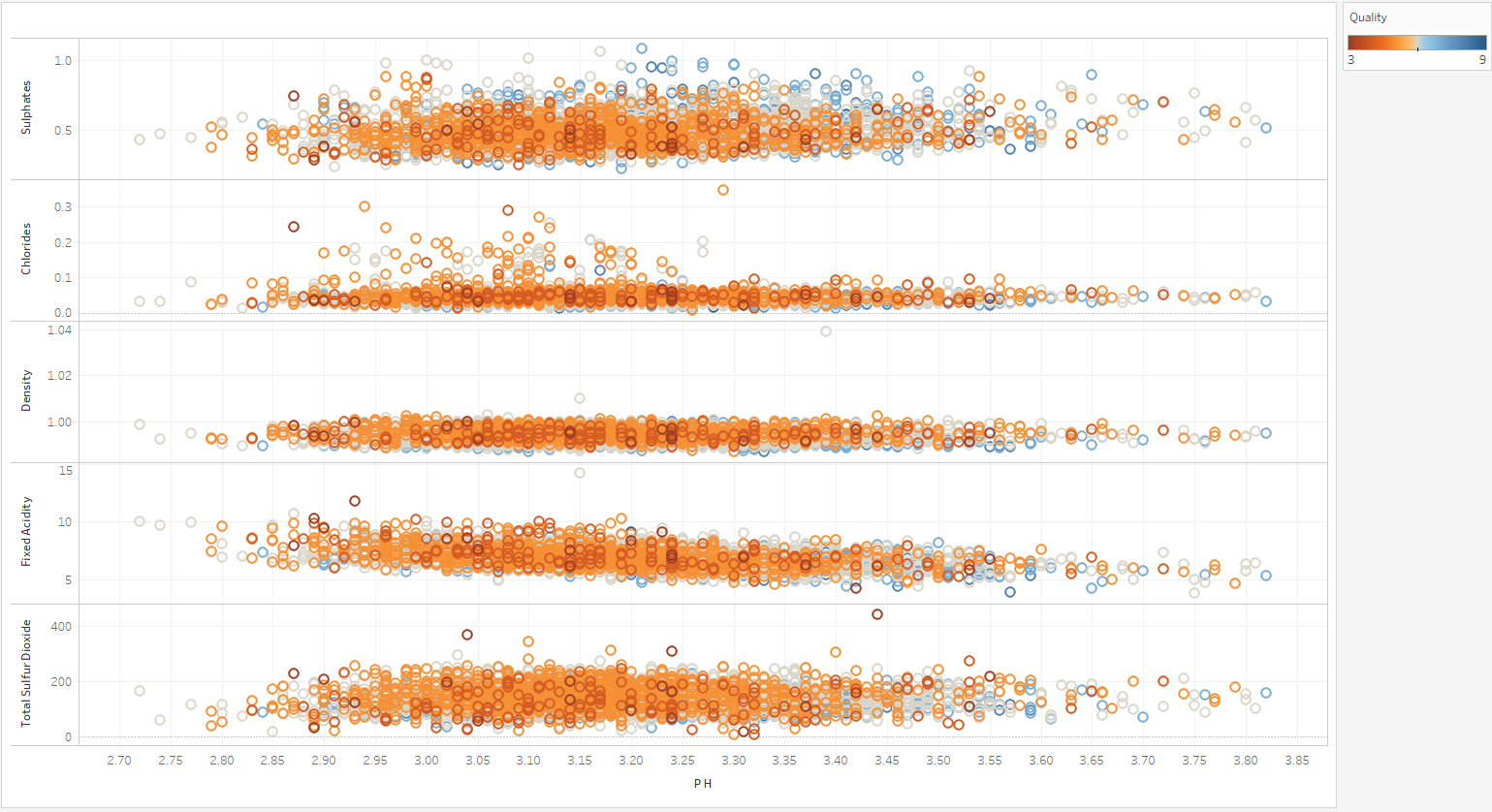


1. Sugar to Quality Scatter graph



On this graph we can see very clearly outlier. It is a sugar at level of 65.8. It is very high level of sugar. But after small research about level of sugar in white wines. It is possible, ther eis many kinds of grapes and sweet white wines like for example Hungarian Tokaj – with that kind of level of sweeteners. It seems that it’s not mistake. We decided to leave that outlier in our model.

1. Alcohol to Quality Scatter plot
2. Alcohol to PH and Quality in Colour Scatter plot
3. Other variables to PH and Quality in Coulor scatter plot.



Low – below - 9%

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).

# **Classification rules:**

After studying materials about wine production[[2]](#footnote-2). We find some rules about wines.

**Target attributes:**

After first selection of attributes. We excluded:

- density as it is actually alcohol content.

Our target variables are:

* Cost range – polynominal – categorical variables.

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

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

Low – Others

* Sugar/Ph ratio – level of sugar/ level of Ph
* Alcohol content

High – above - 11%

Medium – 9-11%

**Explanation:**

1. Sugar/PH ratio. It is common knowledge that wine cannot be too sweet or too acidity in taste. The biggest ratio then the less probable that the wine will be in good quality.
2. Price range – basically there is some preservatives used in wines to make wine last longer. If we have more preservatives then it will be cheaper and can last longer. Our preservatives are:

# Sulphates

* Chlorides

If less preservatives than it will be more expensive wine due to short shelf life.

Are we sure?

1. Alcohol Content – below 9% it is a wine with low level of alcohol and between 9 – 11% it will be medium, above it is high alcohol content

**Classification Rules:**

* If (Sugar/ Ph ratio <= 3.5 ) and (Price range = Medium or Expensive) and (Alcohol content = medium or High) THAN Quality = HIGH (8-9) (0.6/0.0)
* If (Sugar/ Ph ratio <4.5) or (Price range = Medium or Expensive) and (Alcohol content = medium or High) THAN Quality = MEDIUM(6 -7) (0.3/0.0)
* Other THAN Low Quality (5-1) (0.5/0.0)

# **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.

# **Construction of data mining model**

xxxx

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

## 

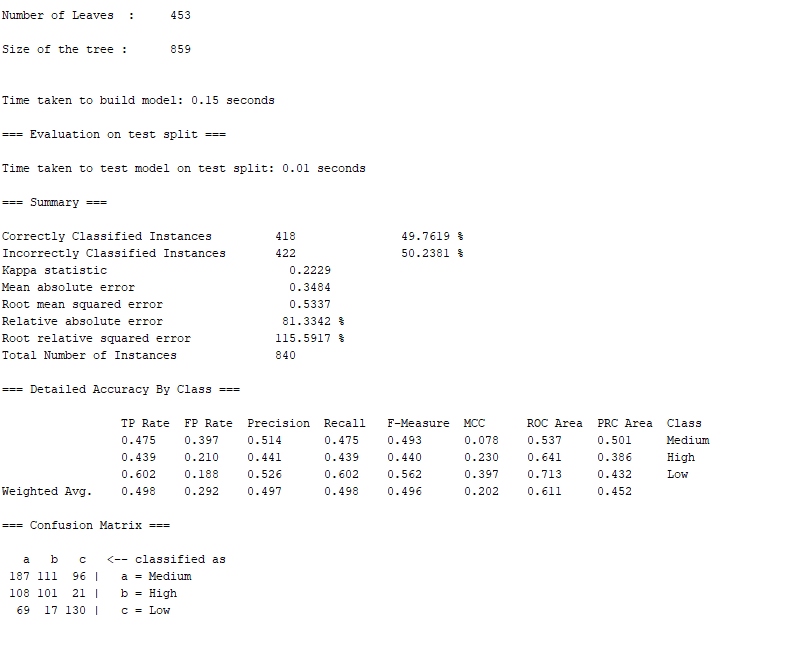
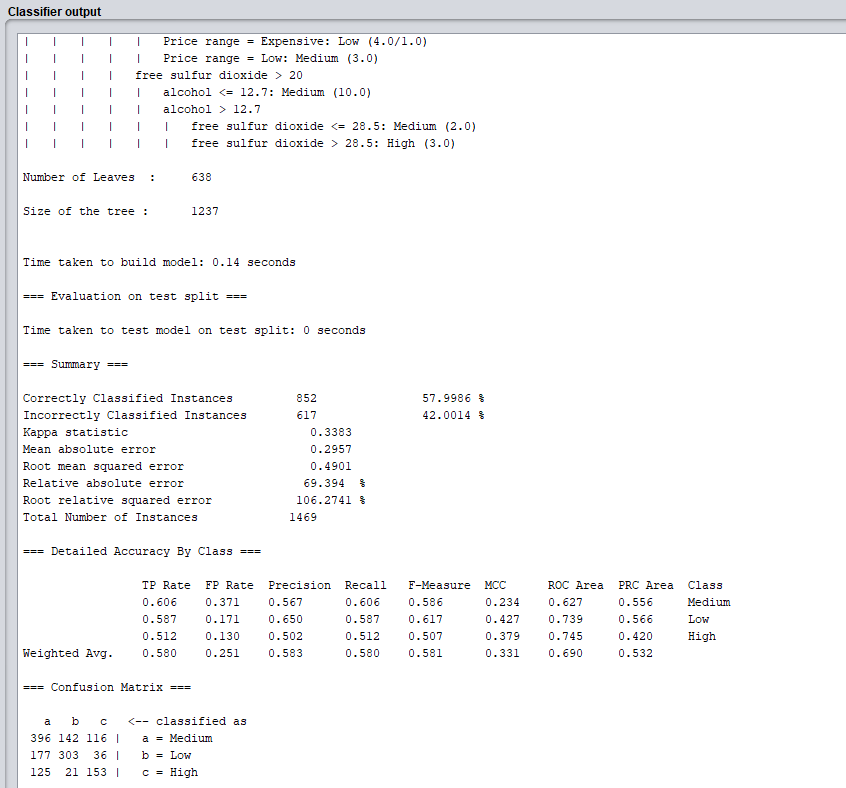


Fig. X.1 Decision Tree Model J-48 evaluation output (Weka) – target Quality



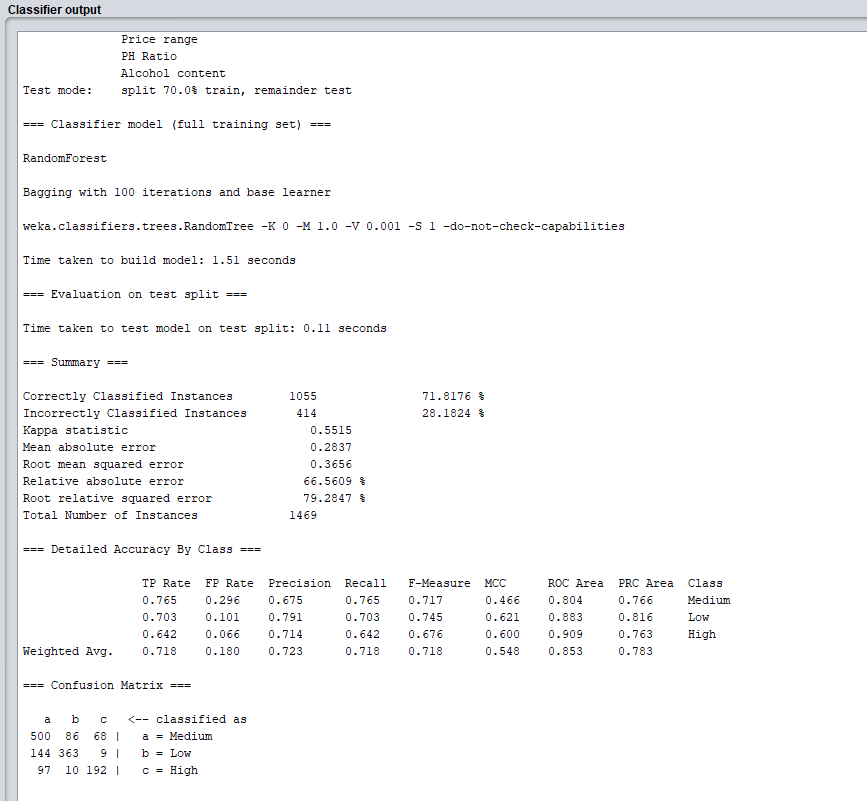
Output with dataset with all 4898 observations – target variable Quality (3 bins: 3-5 Low, 6 Medium, 7-9 High)

## 

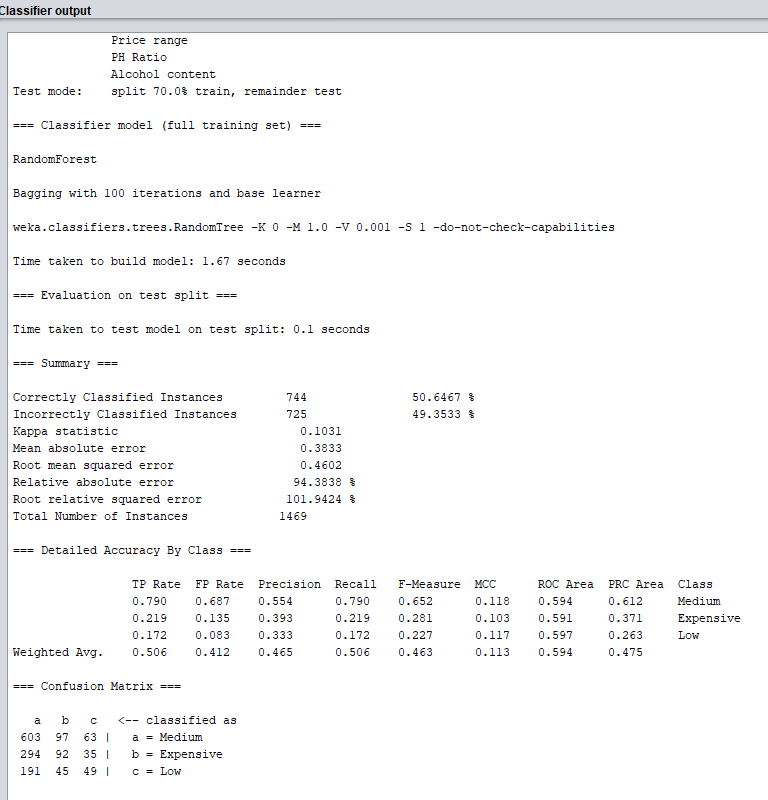
Decision Tree J-48, target variable Price Range (removed Sulphates)

## 

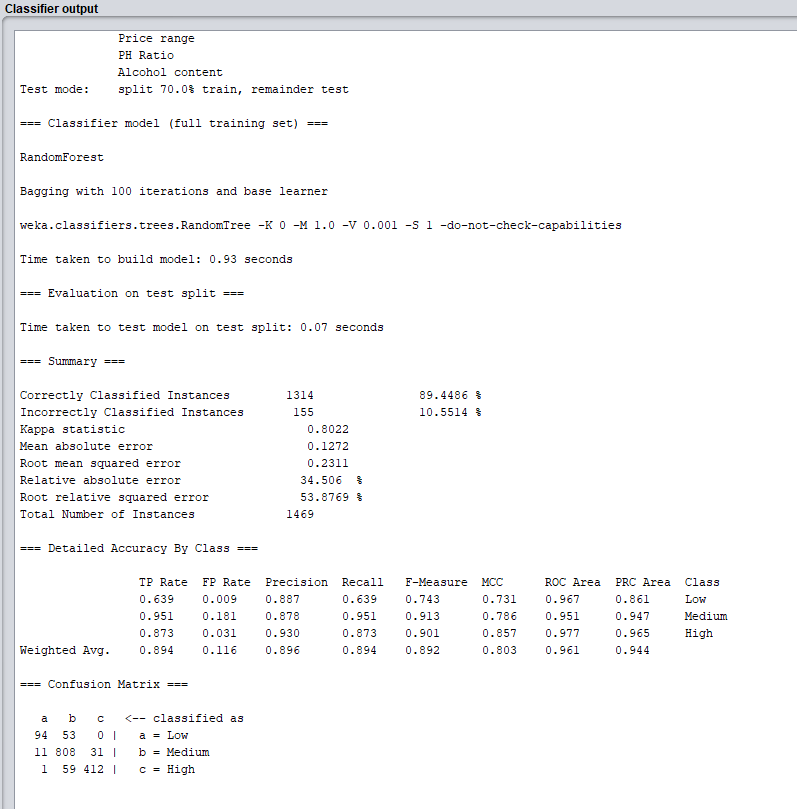
Decision Tree J-48, target variable Price Range (removed Chlorides)



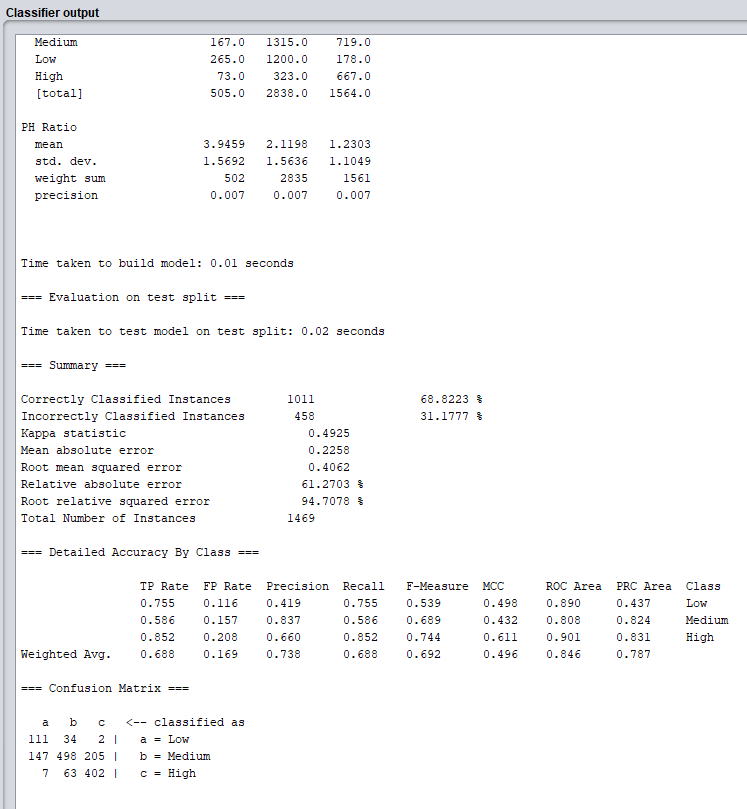
Random Forest model – target Quality (3 bins)



Random Forest model – target Price Range



Random Forest – target Alcohol content (alcohol removed)



Naïve Bayes – target Alcohol content (alcohol attribute removed)

## **Clustering**

Xxx

# **Conclusions**

# **References**

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from[[3]](#footnote-3) physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

<https://www.wardsci.com/www.wardsci.com/images/Chemistry_of_Wine.pdf>

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. https://www.wardsci.com/www.wardsci.com/images/Chemistry\_of\_Wine.pdf [↑](#footnote-ref-2)
3. [↑](#footnote-ref-3)