Data mining project

Alicja Czajkowska

magda leszczynska

Paulina Luty

Anna Skowron

HDSDA

SEP2

Contents

[**Abstract** 2](#_Toc512463170)

[**Problem Statement** 2](#_Toc512463171)

[**Data set description and pre-processing** 2](#_Toc512463172)

[**Tools used and data overview** 5](#_Toc512463173)

[**Target attributes analysis of white wines** 8](#_Toc512463174)

[**Models used** 13](#_Toc512463175)

[**Naïve Bayes** 13](#_Toc512463176)

[**Decision** **Tree** 16](#_Toc512463177)

[**Unsupervised learning – cluster analysis** 22](#_Toc512463178)

[**Conclusions** 28](#_Toc512463179)

# **Abstract**

White Wines are very specific goods. The quality depends from many factors. Barrels, kind of grapes, location etc. but also from ingredients and some of them do not depends from kind of grapes but from recipe and the activities carried over production process. The level of preservatives, level of alcohol and ph ratio and many other depends is the manufacturer of wines will get desired certification and will not loose on quality of wine.

The Quality management is very important in many industries. Data mining is a nowadays a necessary tool in this field.

We used the classification technique – clustering to find the similarities between attributes and decision tree, generalised linear regression, Random forest, Gradient Boosted Tree to see the dependency of attributes from the target attributes and predict the target variables from our dataset of white wines variables.

Keywords: wine quality, data mining, classification, clustering, k-means, x-means, decision tree, gradient boosted tree, random forest, generalised linear regression

# **Problem Statement**

The level of preservatives in wines starts to be a big problem but also necessary ingredient.

How much level of preservatives influences the other attributes of wine? Is that wine still a balanced and healthy for us? How about level of alcohol in wine?

How about answering that question just because of study of already made wines? Very expensive process of developing a new quality wines can be set aside. There is literature about manufacturing wines and this expertise is necessary but we would like to show just small scope for improvement.

And hopefully this project will help little bit more understand the dependencies between attributes of wines.

Keywords: wine quality, data mining, classification, clustering, k-means, x-means, decision tree, gradient boosted tree, random forest, generalised linear regression

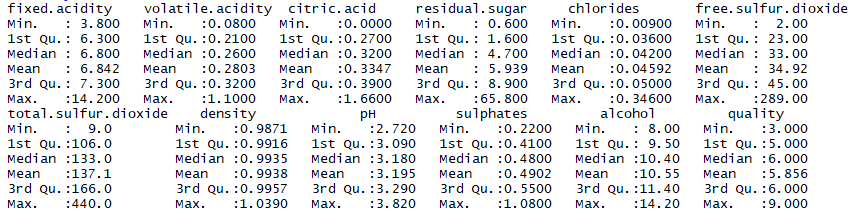
# **Data set description and pre-processing**

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

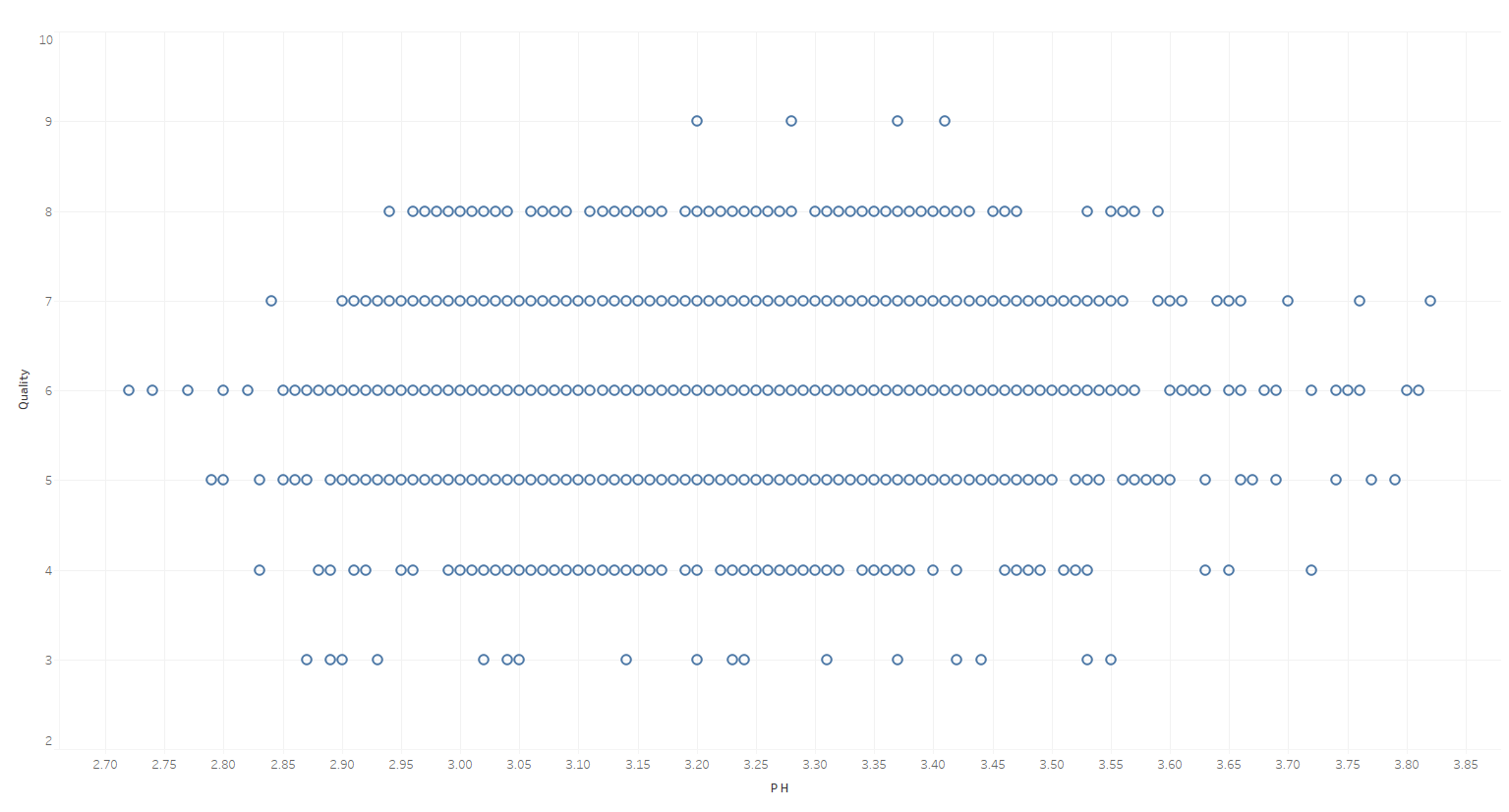
We decided to use sugar /pH ratio as the wine should not be either too sweet or too sour. We expect to find some correlation between the ratio and quality.

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.

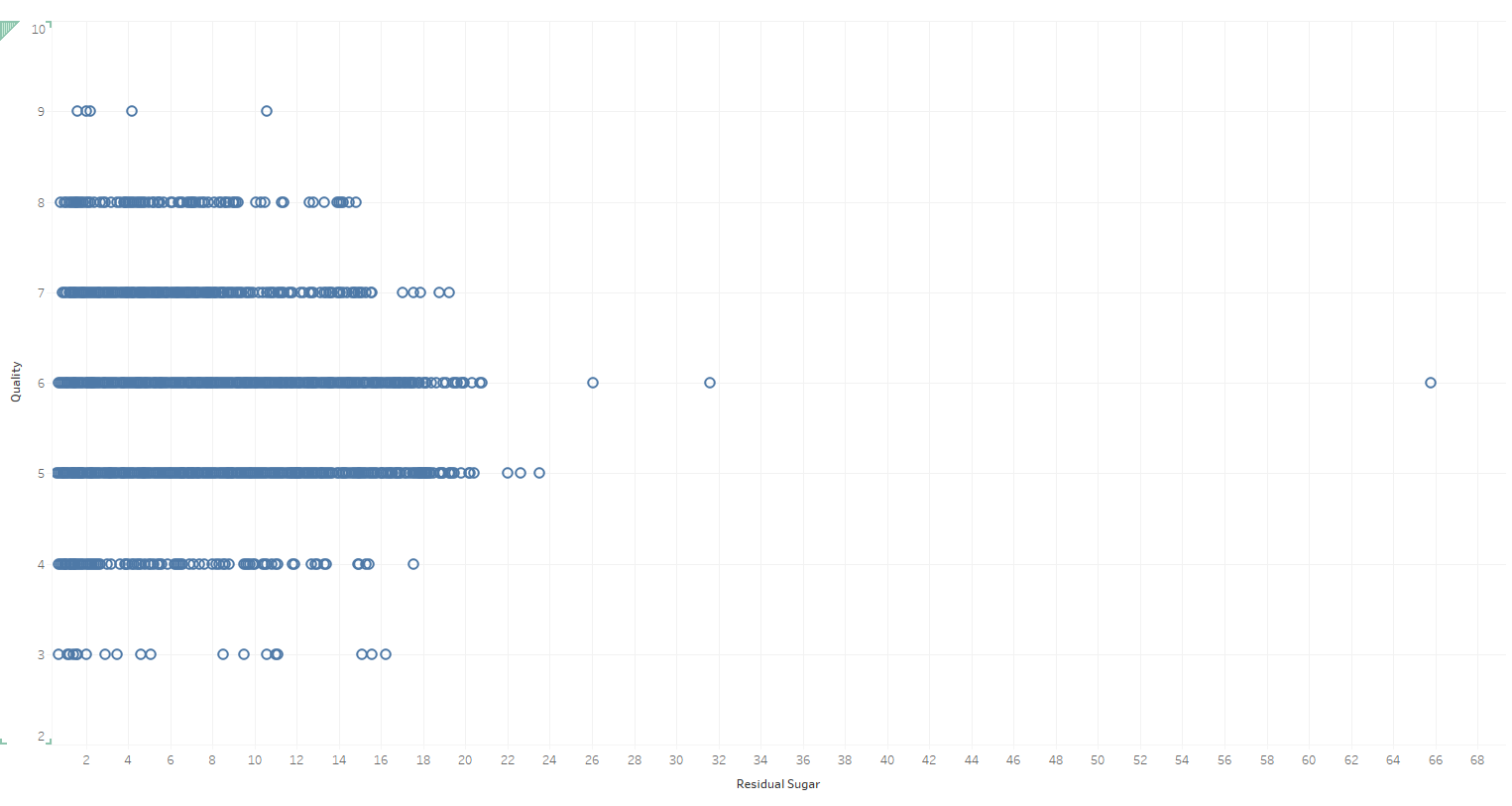
For the alcohol content, we wanted to group wines into three categories based on their alcohol level.

# **Tools used and data overview**

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

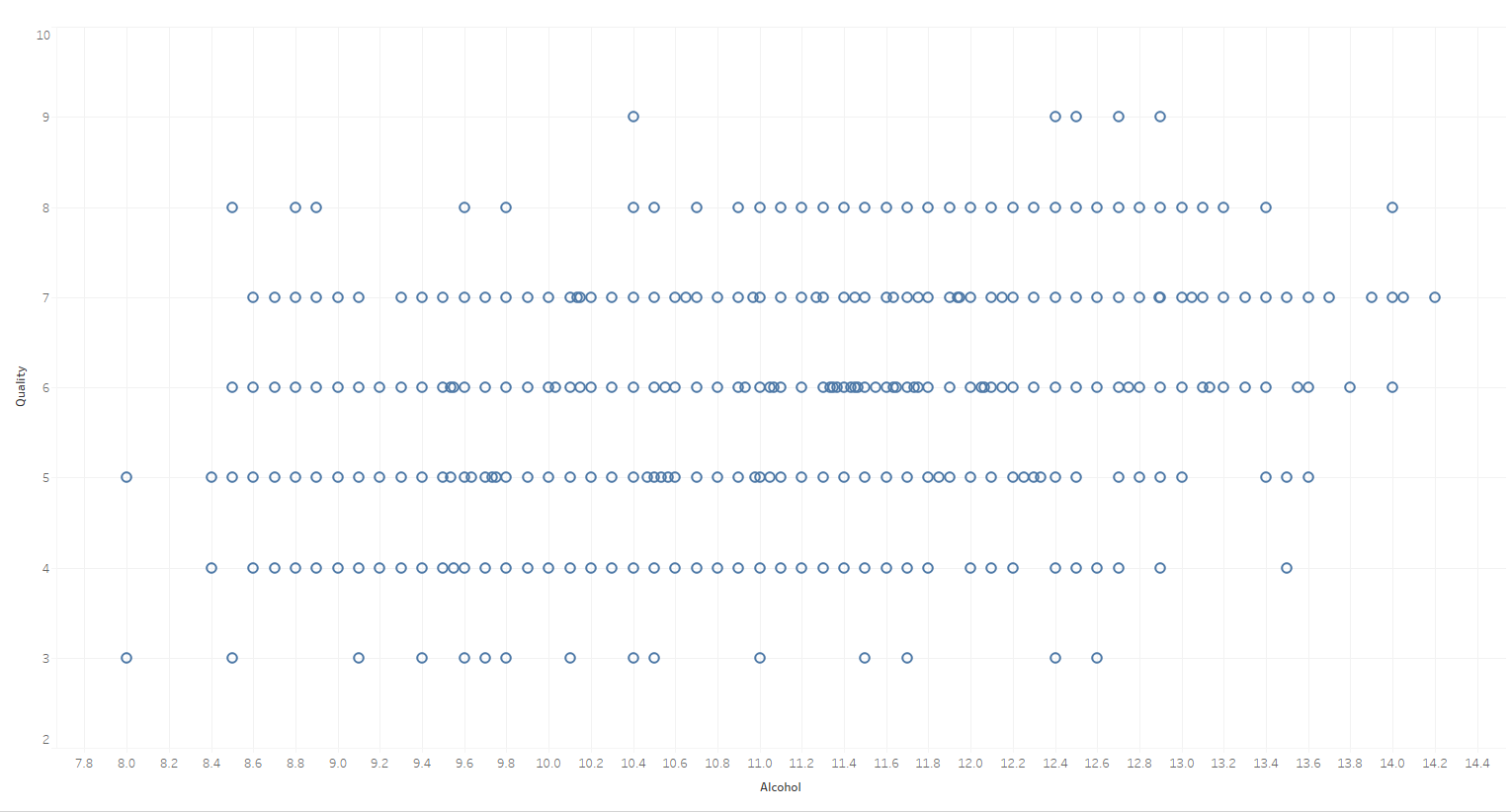
First, we have ran some visualisations to see how data is distributed and checked for outliers and possible correlations.

*Graph 1. PH to quality scatter plot*

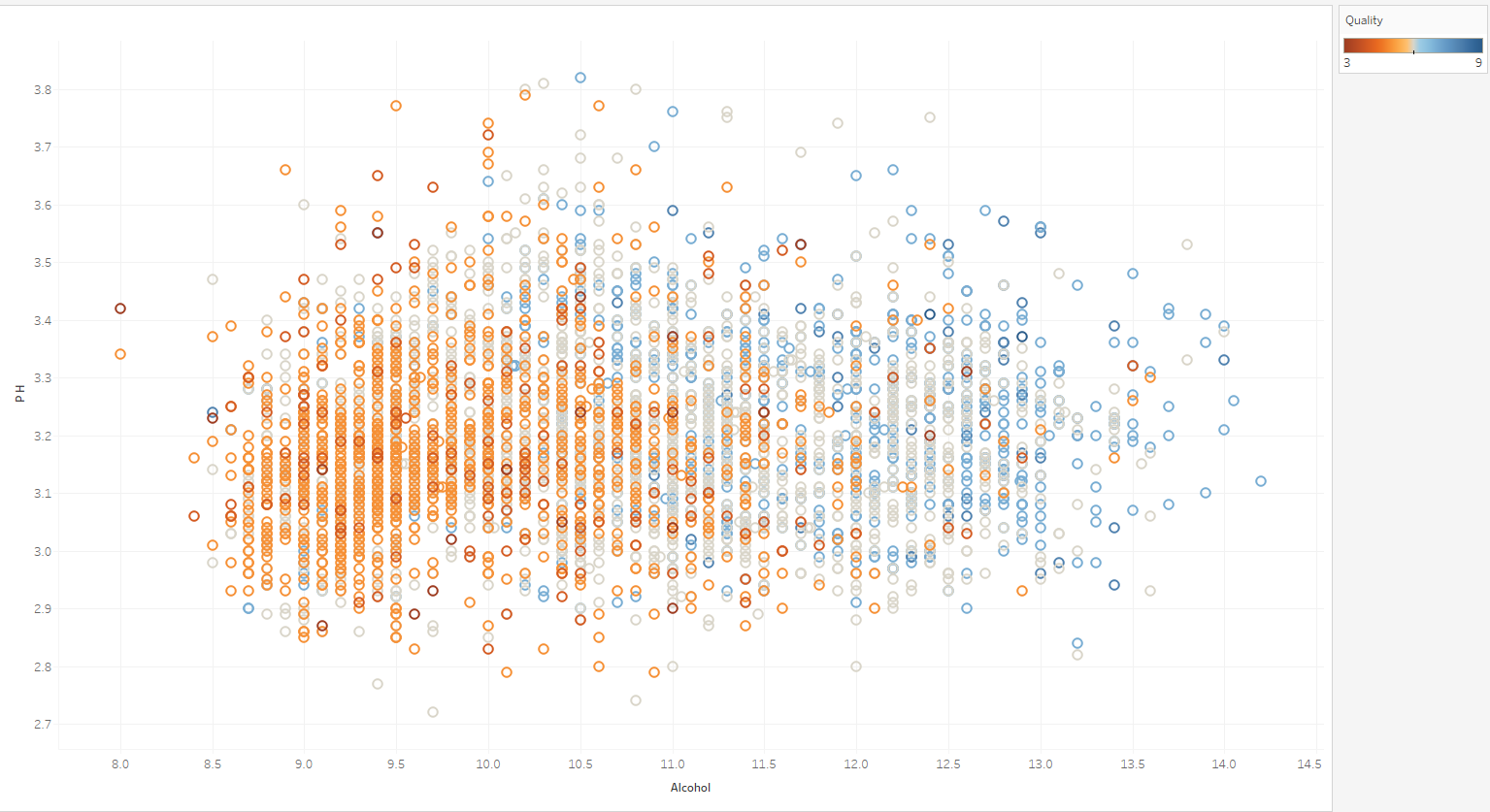


*Graph 2. Sugar to quality scatter plot*

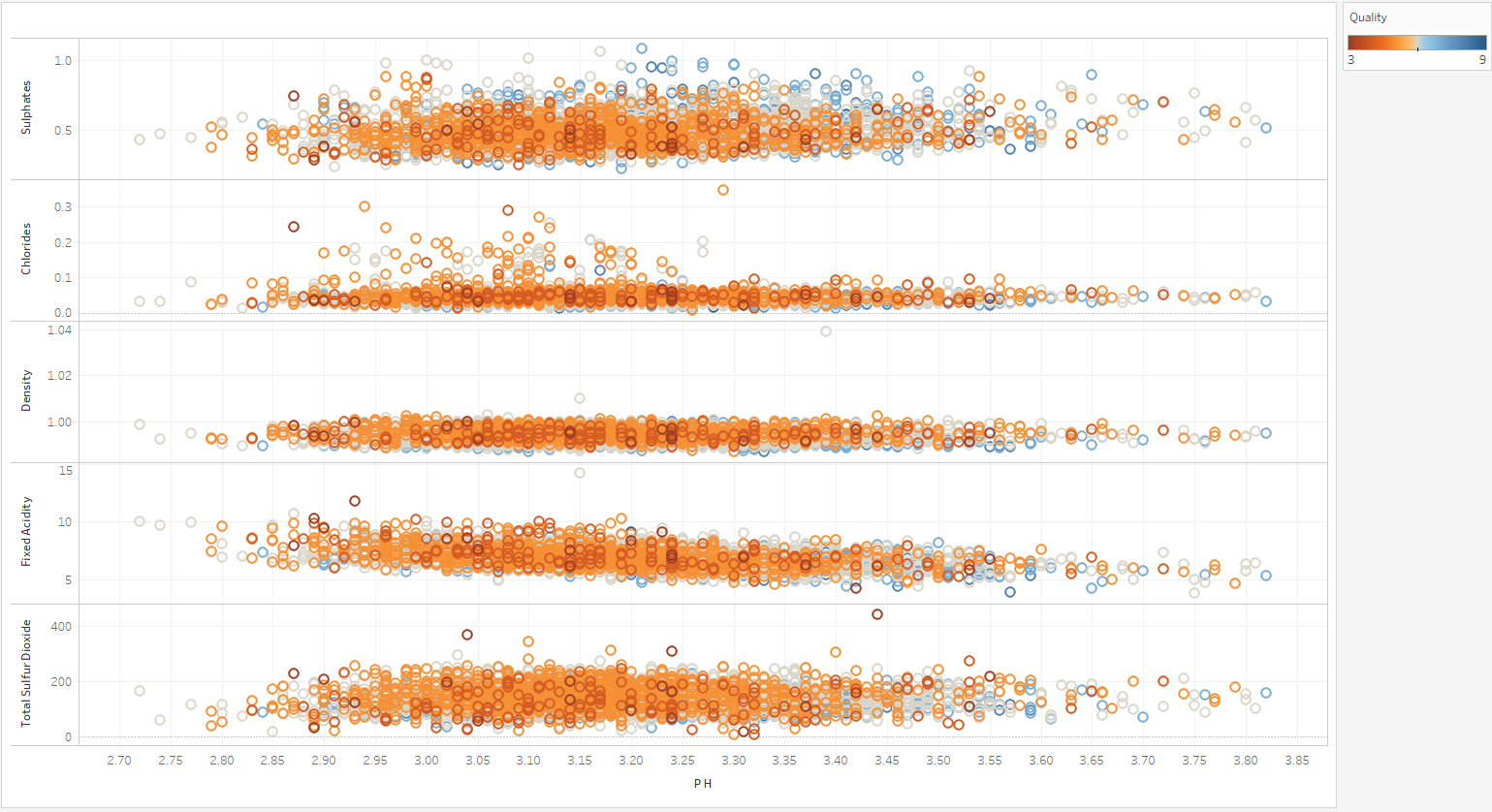
There is an outlier on this graph - a sugar at level of 65.8. We decided to leave that outlier in our model as the amount of sugar in white wines can vary from 0 to 220 grams per litre[[4]](#footnote-4). It is possible that one of the tested wines had that amount of sugar.



*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 is taking points from wine quality (excluding alcohol content – if less alcohol in wine, then it is less probable that it will good quality wine).

# **Target attributes analysis of white wines**

We used the following methods for attributes analysis:

1. Naïve Bayes
2. Generalised Linear model
3. Decision Tree
4. Random Forest
5. Gradient Boosted Trees
6. Target variable – level of preservatives

In this particular exercise we did analysis to figure out how changing the average of one of measurement affects another.

We created a new attribute - Level of preservatives in white wine. It is categorical attribute.

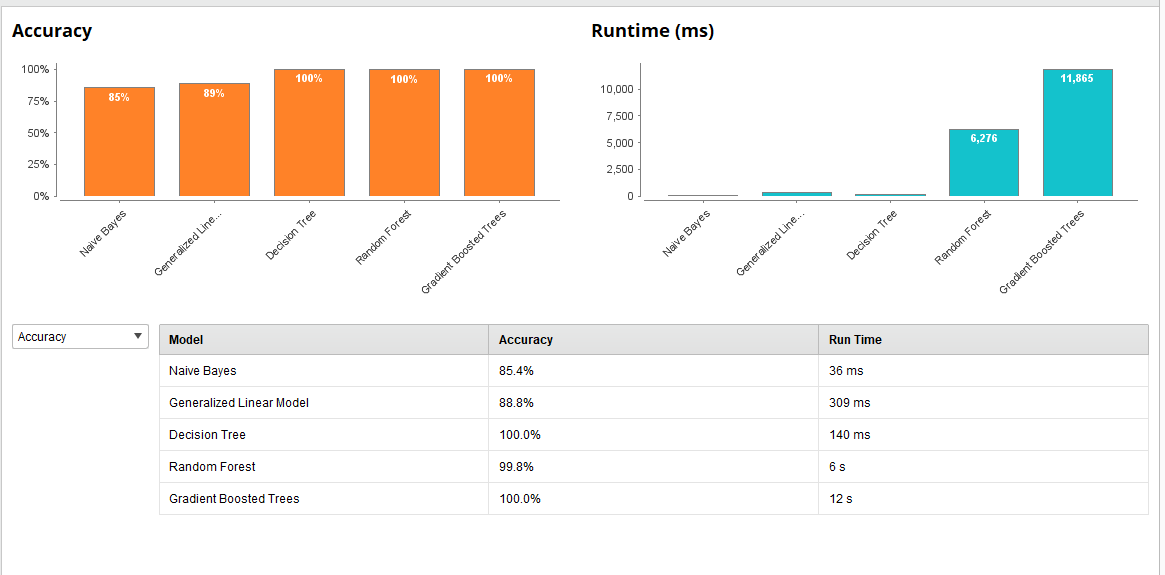
Rule : =IF(AND(J44<=0.45,E44<=0.045),"Low",IF(AND(J44<=0.6,E44<=0.06),"Medium","High"))

There is 3 kind of output Low, Medium and High. Ordinal, polynomial attributes.

The aim of this analysis is to check what will happen with one variable if we will change or delete the other.

Training set was extracted at 70% of dataset and 30% was a test set.

Training:

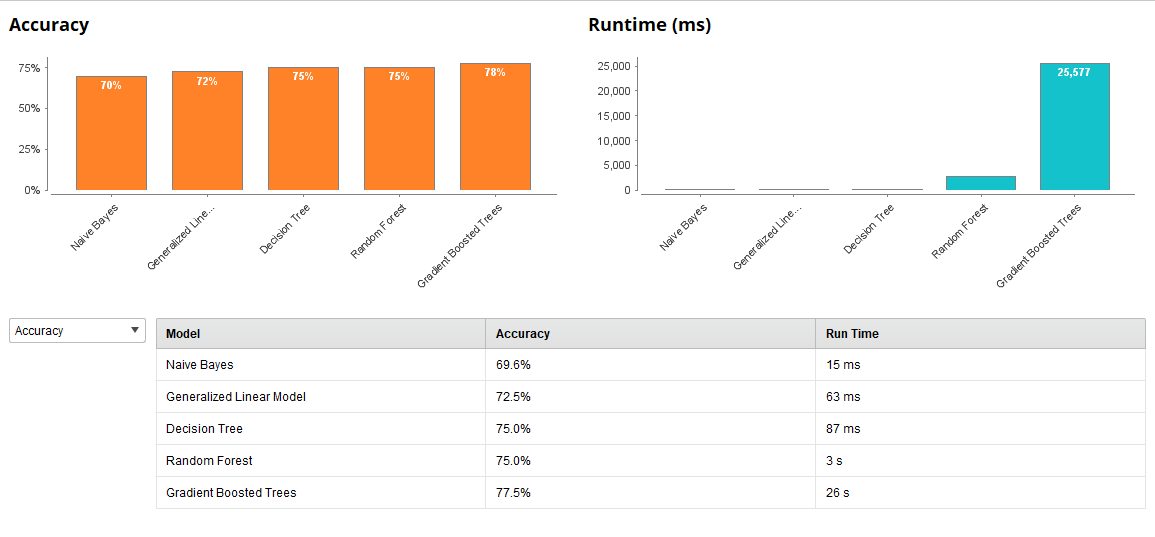


Graph 5. Summary of model outputs with given attribute.

In test set we have delete chloride attribute and then try to predict the outcome of target variables based only on other attributes.

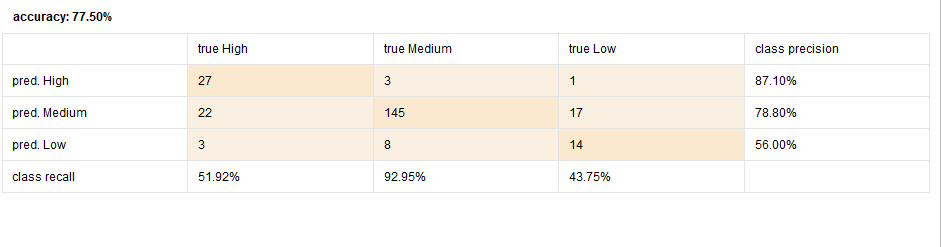
Cross – validation:

Test:



Graph 6. Summary of model outputs with deleted attribute - chloride.

As we can see the Gradient Boosted Trees method had the highest accuracy at 77.5%. But it was the slow as it takes 26 seconds compared to others.

It means that even without chloride value the other values of white wines can give us information about level of preservatives at 77.5% chances that we are correct. 

Let’s see the outcome of Gradient Boosted Trees method.

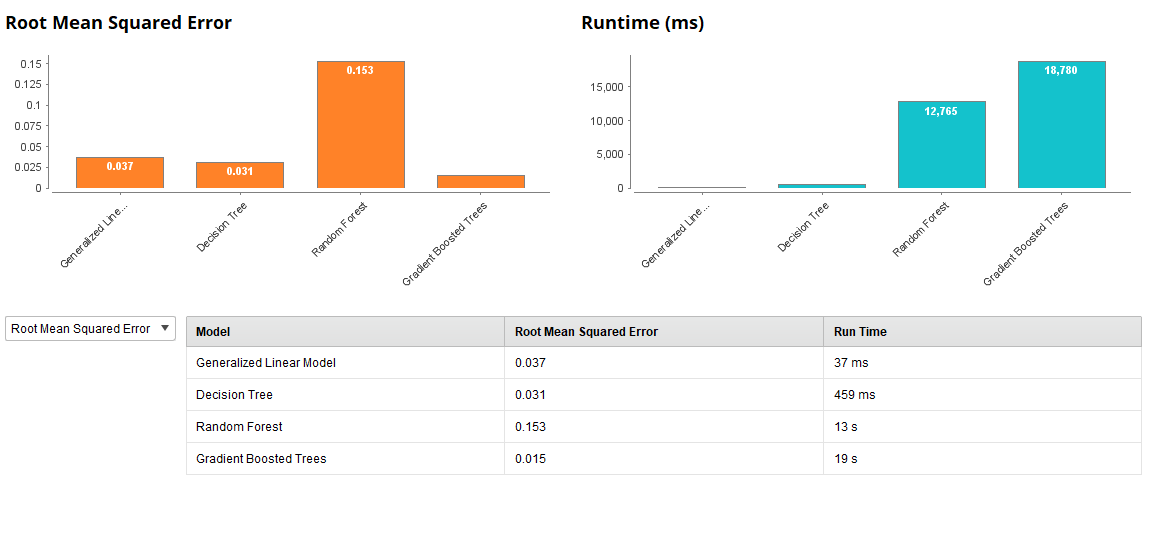
We can see that model had no problem with medium level of preservatives but had problem with the High and Low level.

Target attribute Sugar/PH ratio.

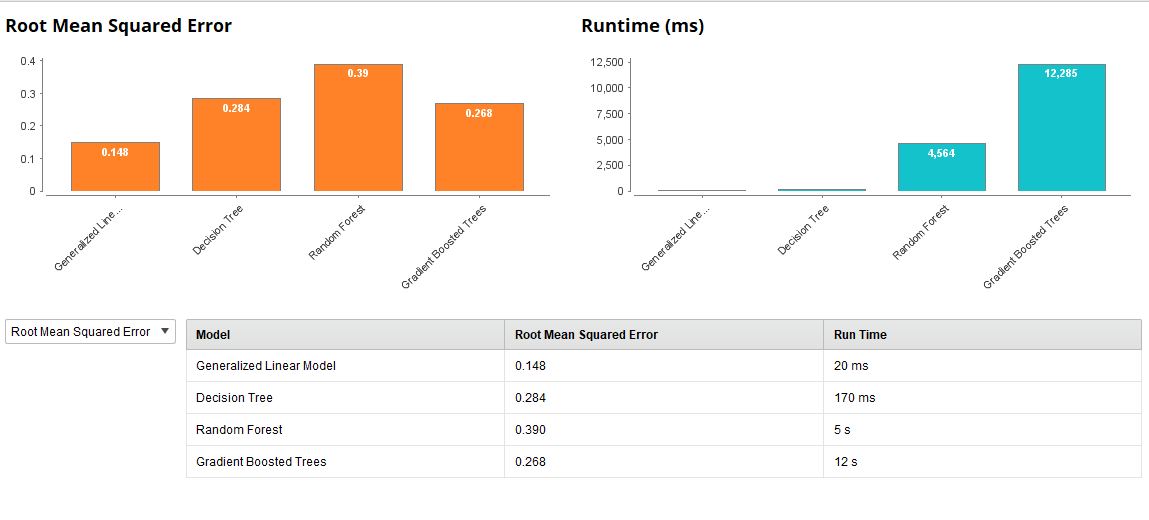
We chose to create and examine this ratio after discussion and some research. It is proven that this ratio should be the lowest to make sure that wine will be tasty.

Are we able to prove it with our sample of wines?

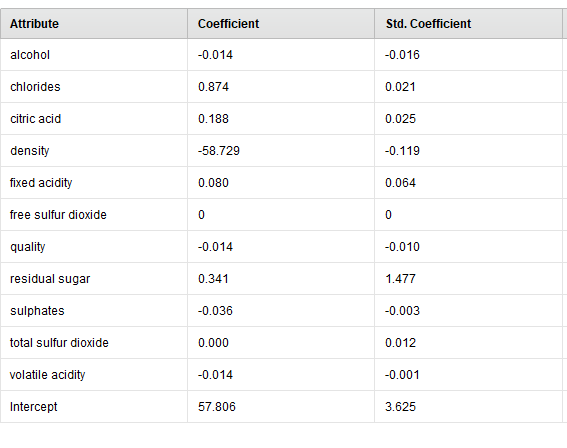
Methodology used as in previous example: training 70%, test 30% and for test we have deleted pH attribute.

Training

Graph 7. Summary of model outputs with given attribute. – root mean squared error

Cross -validation: Test:

Graph 8. Summary of model outputs with deleted attribute – pH.

In this case the best was simple Generalised Linear Model with the smallest error. Output below:

This is a very interesting finding. We can see here that even without given pH of wine we are able to predict sugar/pH ratio in almost 90% accuracy.

The biggest impact on that ratio will have density -58.72. But few others are also very important like chlorides or residual sugar.

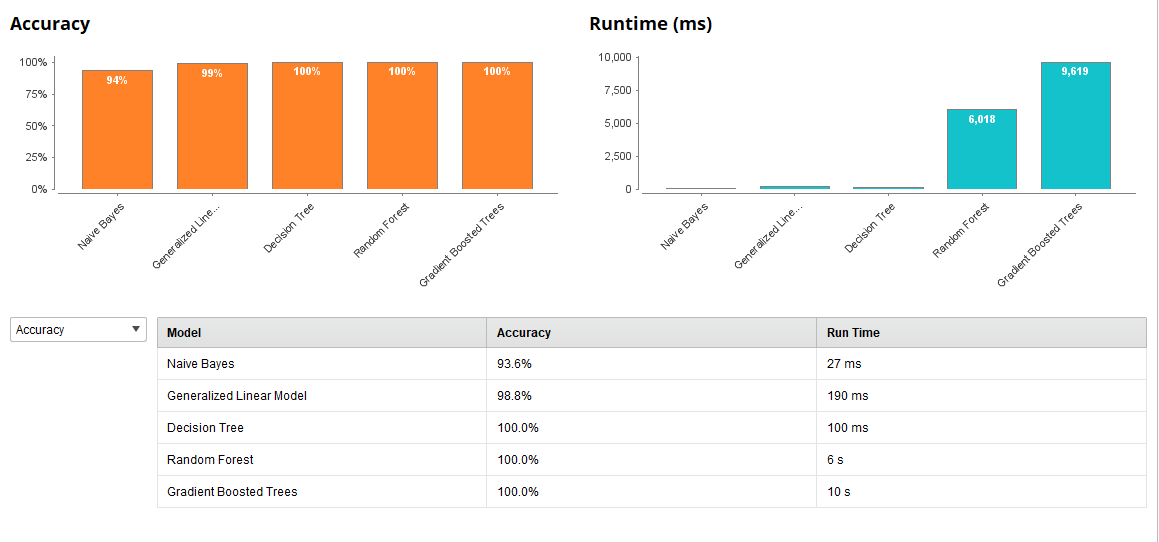
Target attributes - level of alcohol.

That is a very interesting one. Are we able to predict level of alcohol without given alcohol content?

Level of alcohol was set up under rule:

=IF(K2<=9,"Low",IF(K2<=11,"Medium","High"))

There is 3 kind of output Low, Medium and High. It is ordinal, polynomial attribute.

Methodology used as in previous example: training 70%, test 30% and for test we have deleted PH attribute.

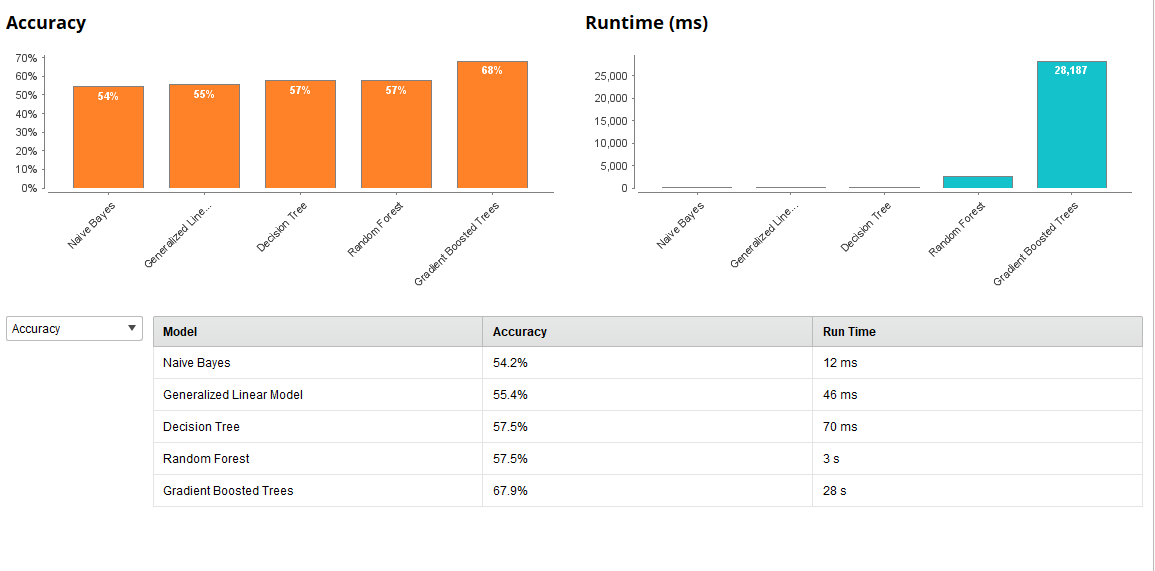
Training:

Graph 9 Summary of model outputs with given attribute.

Test:

Cross-validation:

Summary of model outputs with deleted attribute – alcohol content.



Only Gradient Boosted Trees could give us 68% of accuracy.

# **Models used**

We will use decision trees, clustering and association models.

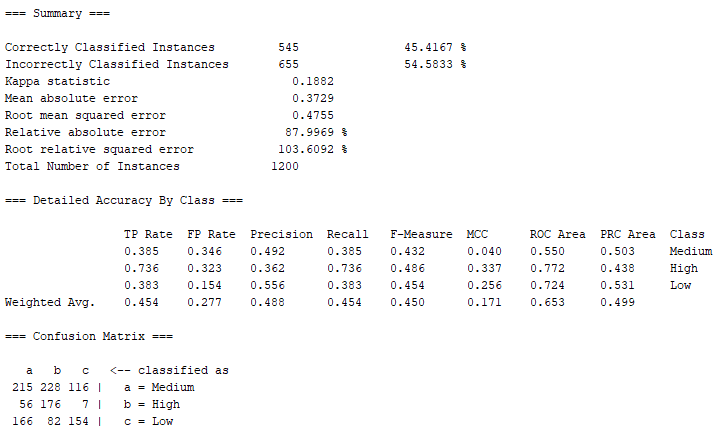
Decision tree is a classification method and is also a predictive method. Classification methods use existing data to create a model that will allow to classify new data. In our project we will use decision tree to predict xxxxxx.

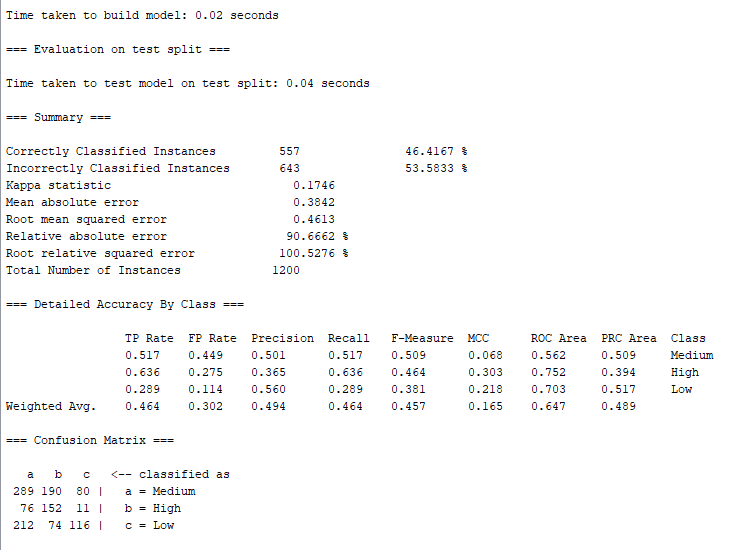
Clustering is a descriptive technique that finds groups of observations (clusters) that share similar characteristics in a data set.

## **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).

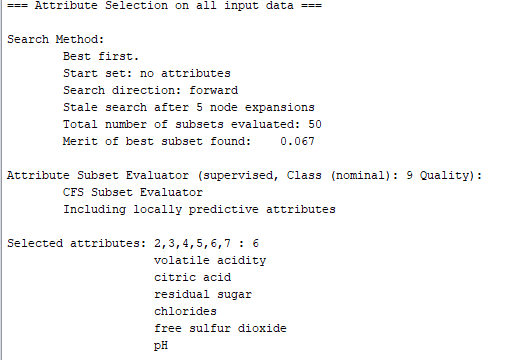




*Figure 1.1 Naïve Bayes classifiers’ output comparison*

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%.

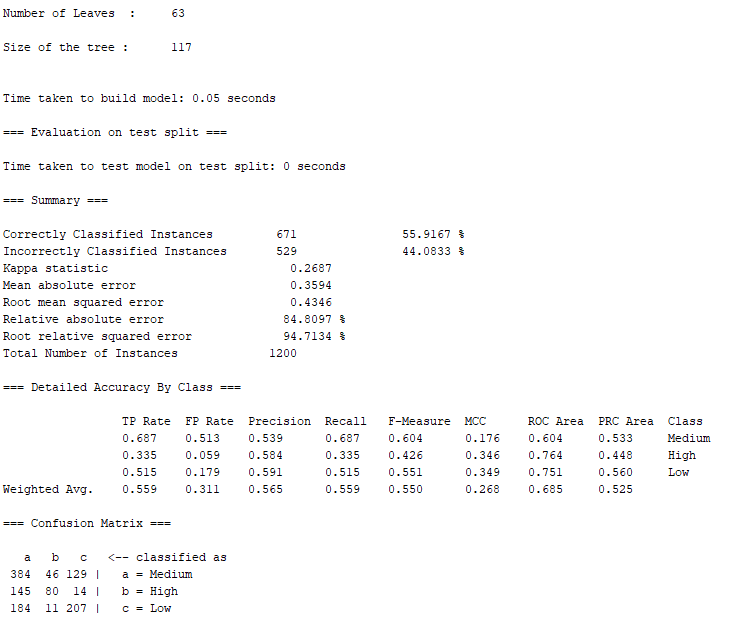
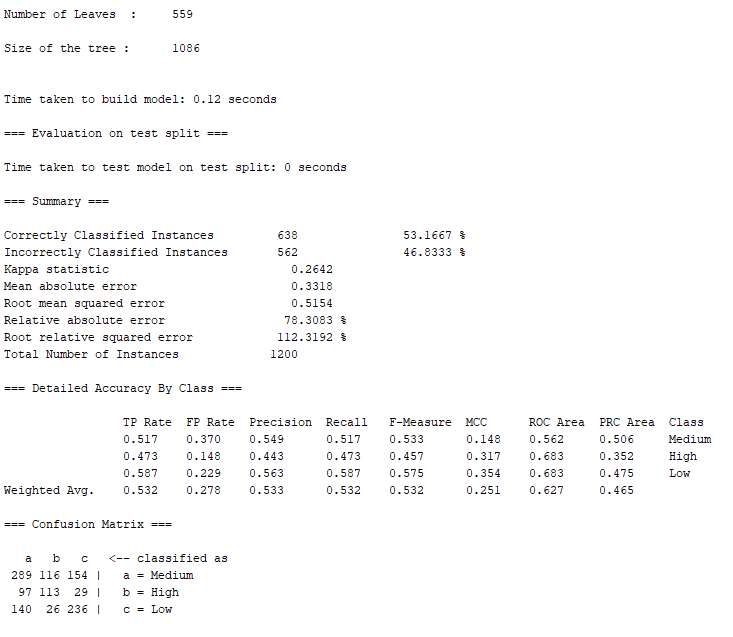
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Class** | |  |  |
| **Attribute** | | **High** | **Medium** | **Low** |
|  | | 0.21 | 0.45 | 0.34 |
| **fixed acidity** | |  |  |  |
| mean | | 6.6968 | 6.8267 | 6.9602 |
| std. dev. | | 0.7846 | 0.8485 | 0.9121 |
| weight sum | | 834 | 1806 | 1360 |
| 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**

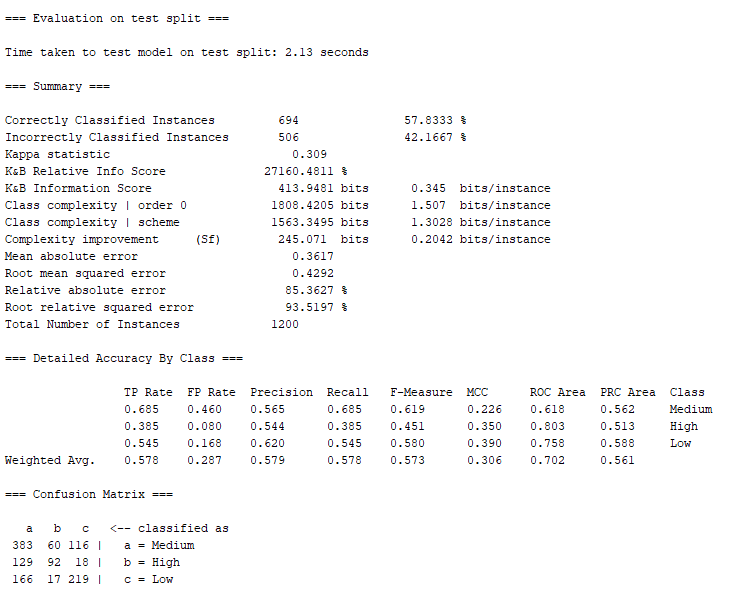
Decision Tree J48 model was executed in Weka using the dataset with 4,000 instances (split 70.0% train, remainder test) and 12 attributes (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, Quality and Alcohol content). 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%.



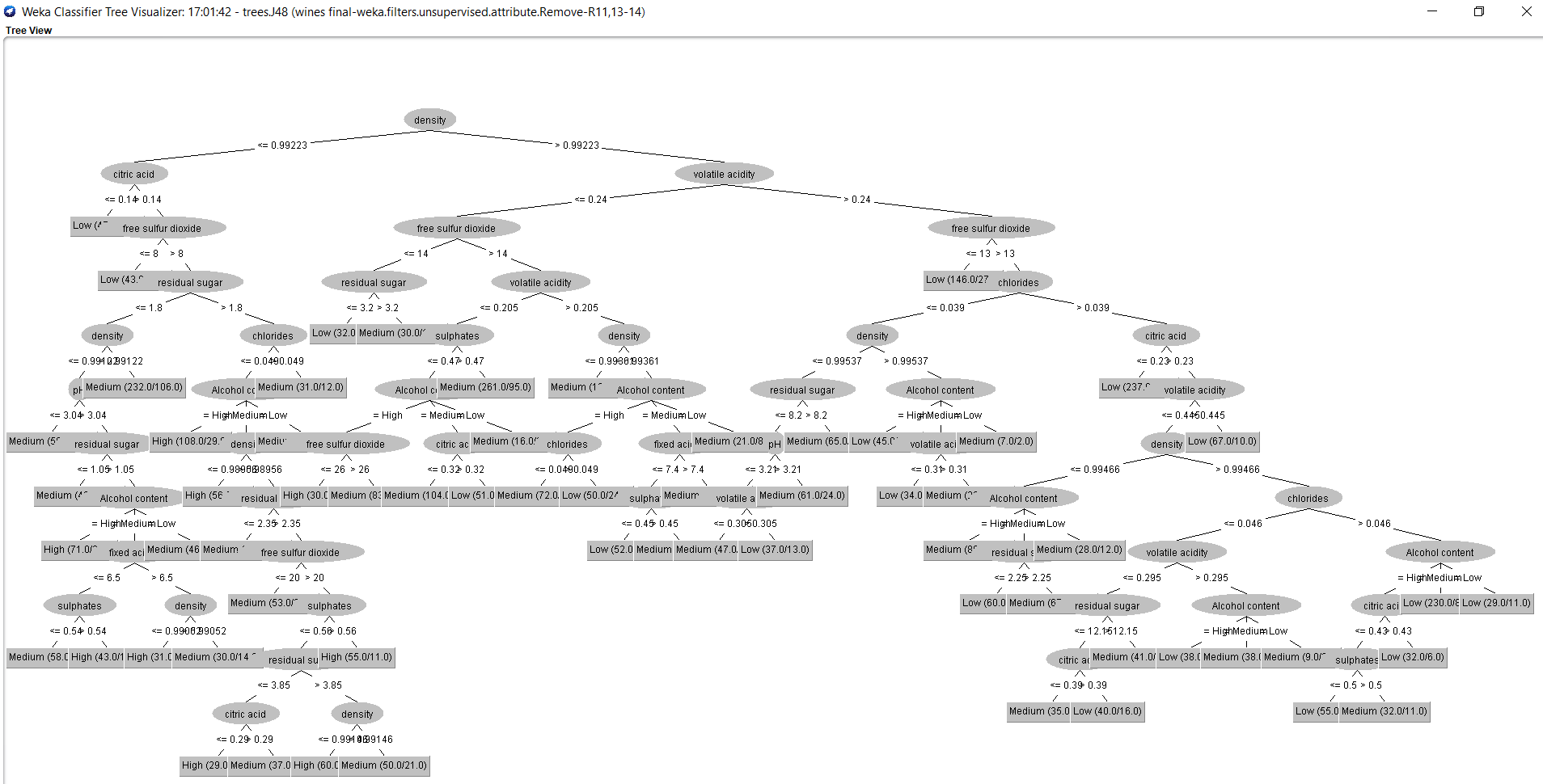
*Figure 1.3. Decision Tree J48 models’ comparison*

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).

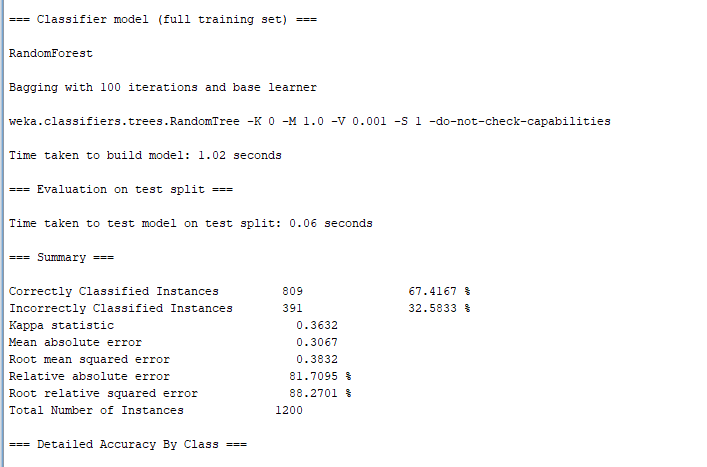


*Figure 1.4. SMO model Weka – evaluation output*



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.

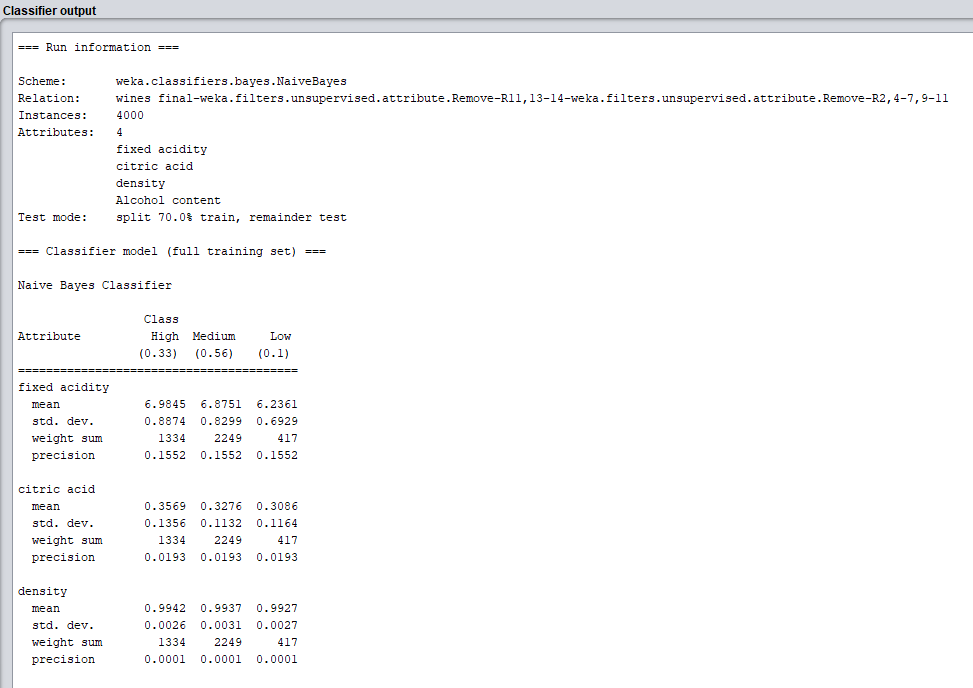


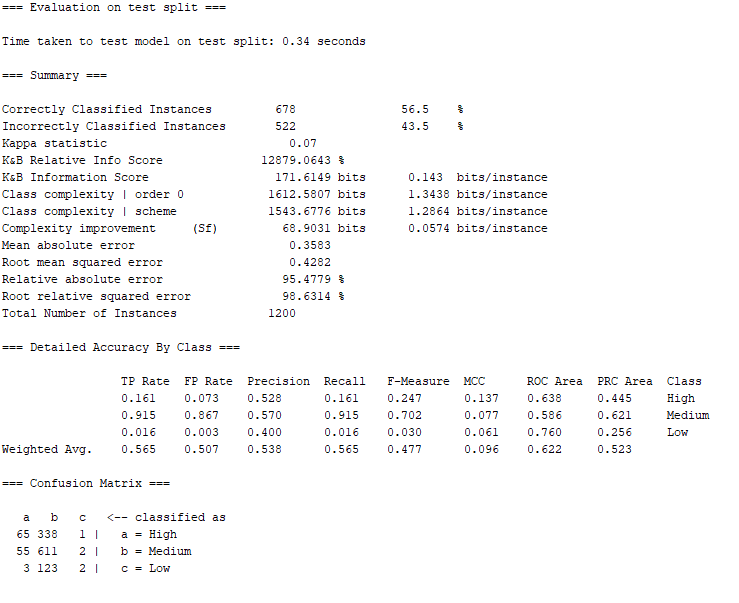
*Figure 1.6. Random Forest model in Weka*

**Classification Models – alcohol content**

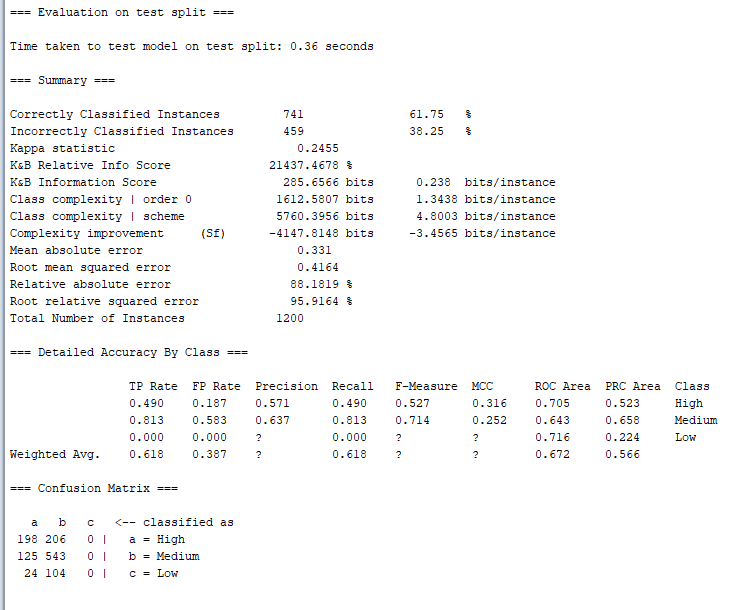
We also used **Naïve Bayes and Decision Tree** models to review alcohol content. The original alcohol values were discretised: values lower than 9% were classed as low, 9 to 11% as medium and above 11% as high. After running the correlation-based feature selection (CFS) algorithm in Weka to we eliminated 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)

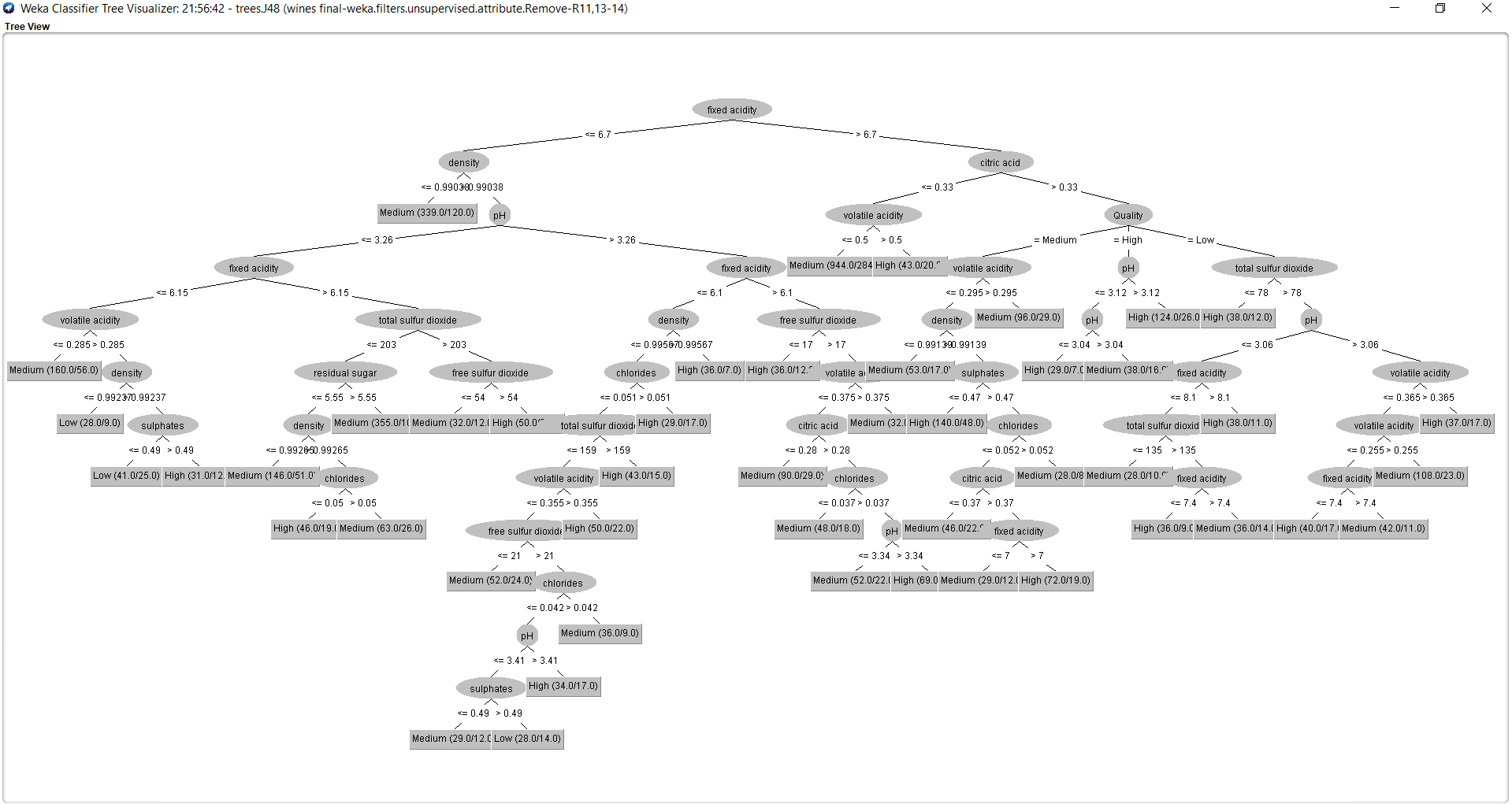




*Figure 1.7. Naïve Bayes model – alcohol content*



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



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

# **Unsupervised learning – cluster analysis**

Clustering is a descriptive method of data mining used to find groups of observations (clusters) that share similar characteristics in a data set. In our project we decided to use cluster analysis to group together attributes of white wine. We wanted to find similarities between attributes as a step to build a good model.

It is a form of unsupervised learning – that means machine is learning from a raw data. All target attributes have been removed for the moment of training[[5]](#footnote-5).

In that particular case we decided to use partitioning approach. We have used k-means and x-means and a variation of k-means to compare these two methods. K-means is the best known method in data science. We also chose x-means as in x-means we do not need to specify the number of clusters.

K-means calculating again the centroid after every assignment and repeat that step until there is no change. That is why it is sensitive for changes.

In clustering we have be very careful with outliers. That is why we were extremely careful when pre-processing data. We took the following steps:

1. Correlations for clustering purposes

 Correlation table – RapidMiner – white wines

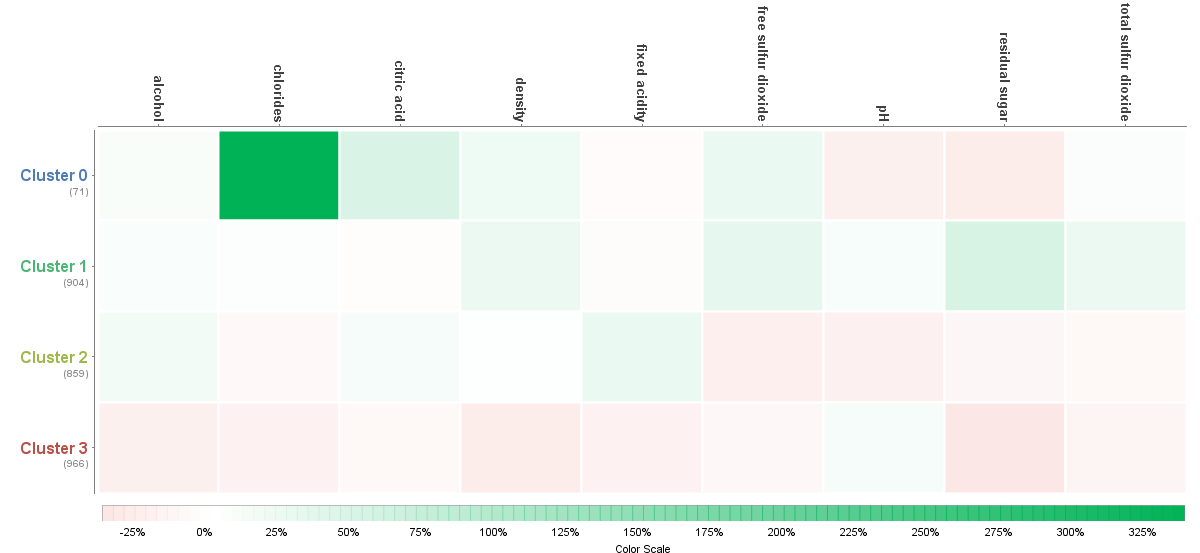
Highest correlations between the attributes in our data set are:

* Total sulfur dioxide and free sulfur dioxide - 0.585
* Fixed acid and pH - 0.480
* Residual sugar and density - 0.443

This information can be very helpful in later stage of project when building a prediction model. This is also a basis for clustering analysis. That is also our basis for choosing the right number of clusters in k-means cluster analysis.

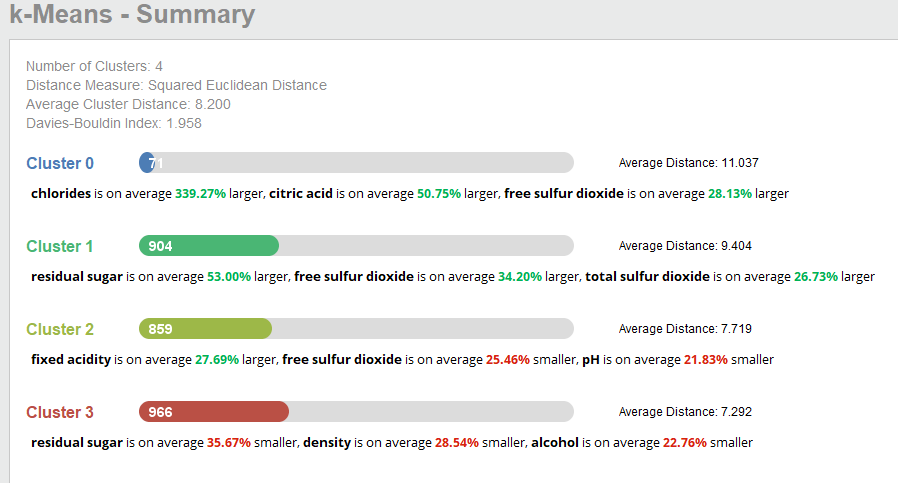
2. K-means clustering for white wine.

We started that part by examining the scatter plots from data exploration part of project, and we decided to use 4 clusters. We did analysis in RapidMiner:



Heat map 1 – Rapidminer studio k-means– White wines attributes

Chlorides could look here as outliers as we can see very intensive green. But in reality there chlorides in wine depends on the kind of grapes. There are types of grapes with very high nutural volume of chlorides.



Summary 1 – K-means RapidMiner studio – white wines

In above outcome it is clearly visible that 4 clusters were a good choice. Although it is still not perfect distribution.

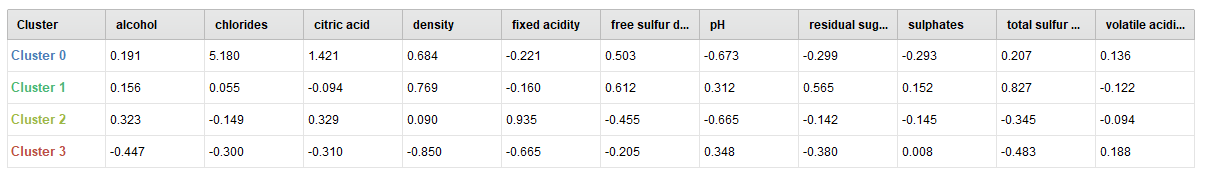
By Davis Bouldin Index we got only 1.958. It is an index for evaluation of cluster algorithms. As closer to zero then the clusters has been better assigned and they are closer to each other.

In cluster zero we can see that chlorides are on average 339.27% larger than others, citric acid is on average 50.75% larger and free sulfur dioxide is on average 28.13% larger then in others. There are only 71 entries in this category but average distance here is 11.

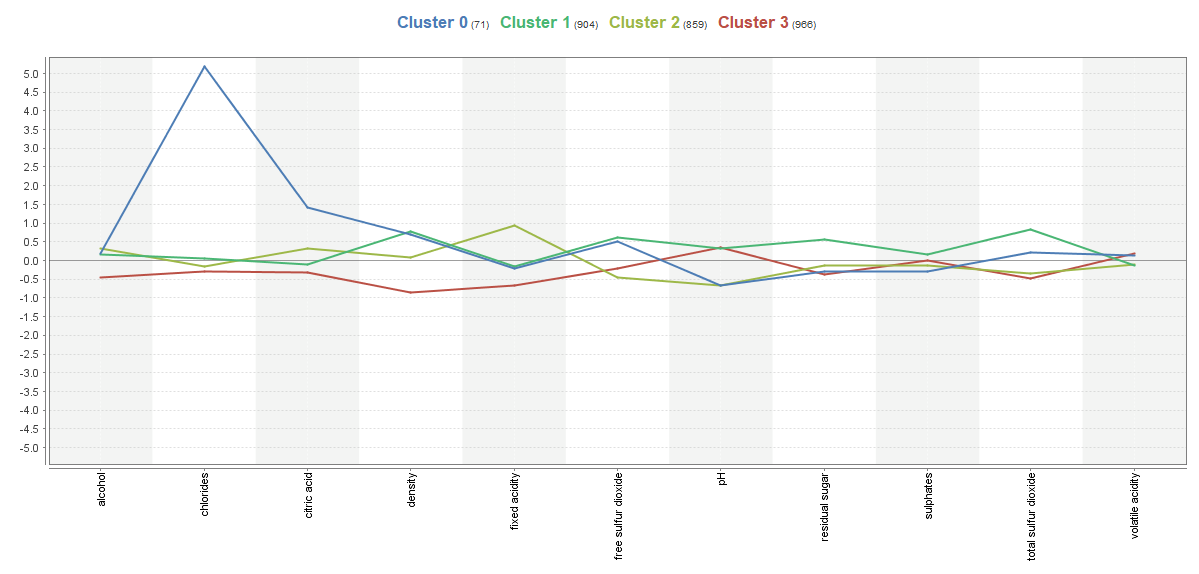
In cluster 1 we can see that residual sugar is on average 53% larger, free sulfur dioxide is on average 34.20% larger and total sulfur dioxide is on average 26.73% larger. There are 904 entries in this cluster but average distance here is 9.4.

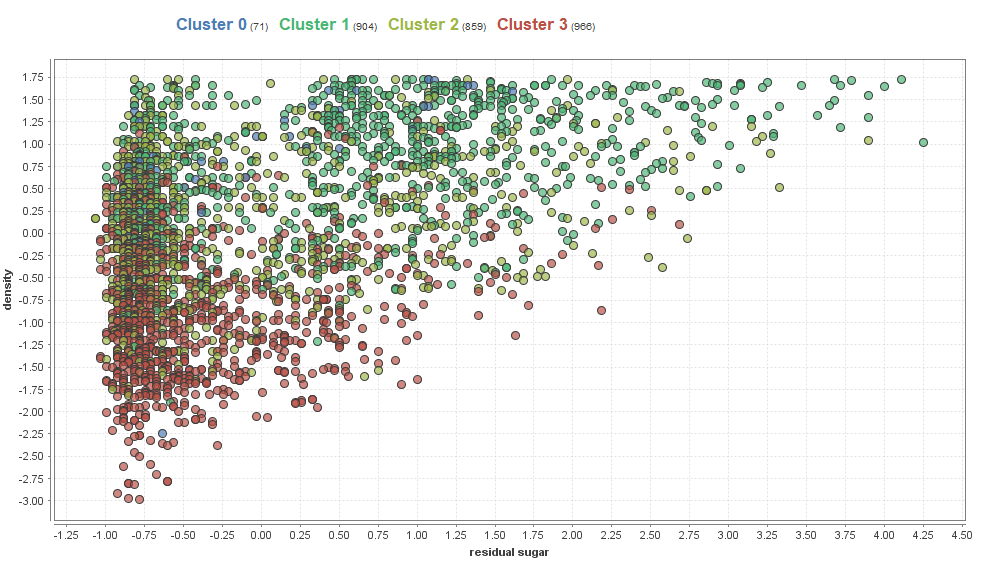
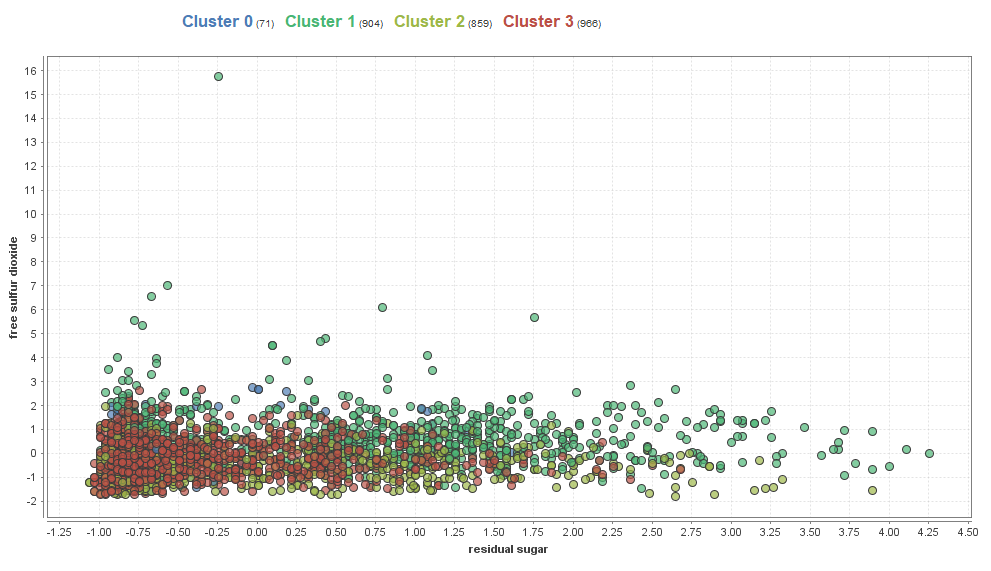
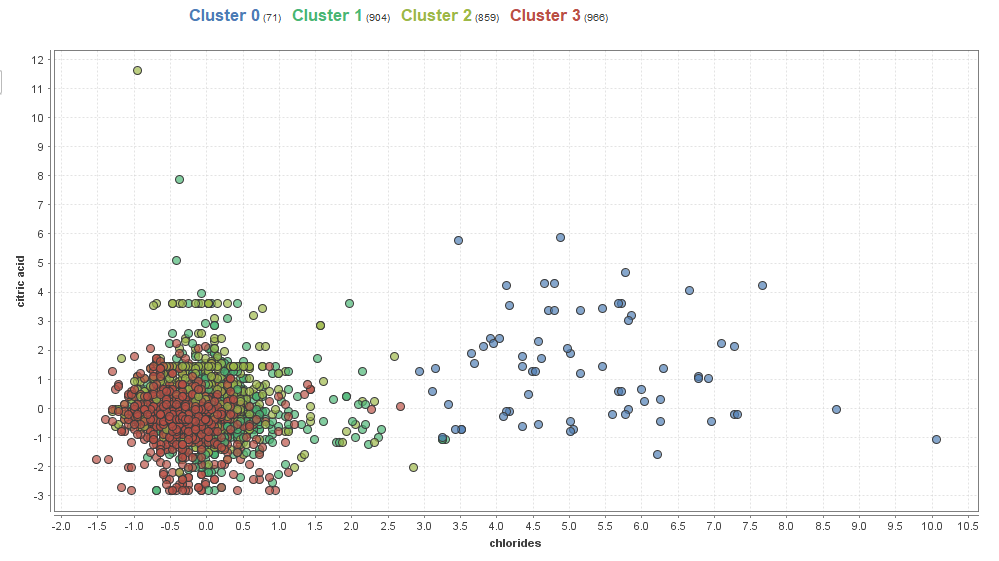
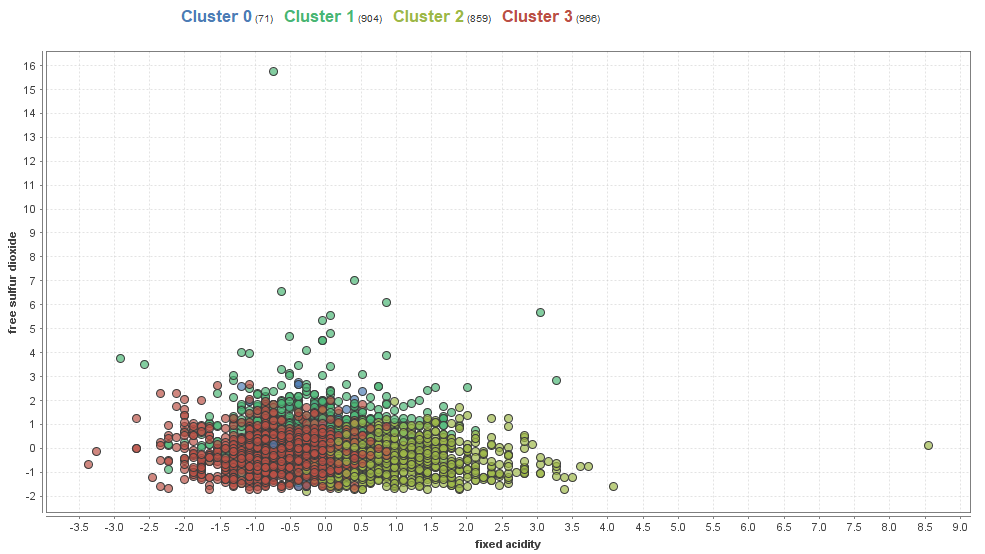
In cluster 2 fixed acidity is on average 27.69% larger free sulfur dioxide is on average 25.46% smaller and pH is on average 21.83% smaller. There are 859 entries and average distance here is 7.7

In cluster 3 residual sugar is on average 35.67% smaller, density is on average 28.54% smaller, alcohol is on average 22.76% smaller. That is the biggest cluster with 966 entries and average distance here is 7.2.

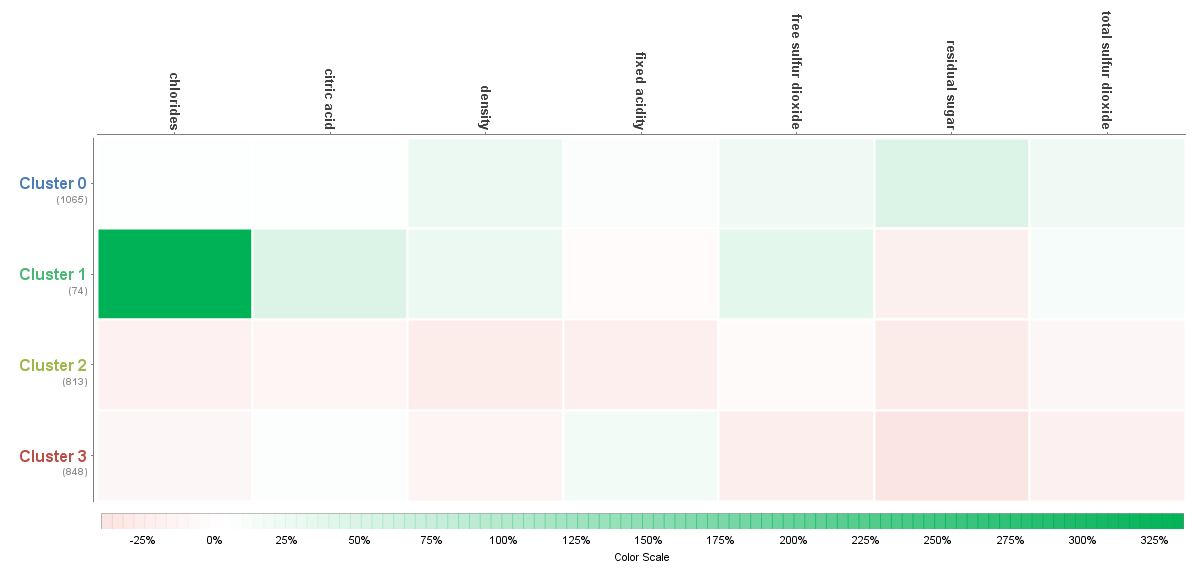
 Table 1 – Rapidminer, k-means white wine

In tables we can see the differences between clusters. Below there is line plot with 4 clusters and scatter plots for all 4 clusters separately.

Line plot – Rapid miner k-means – White wine

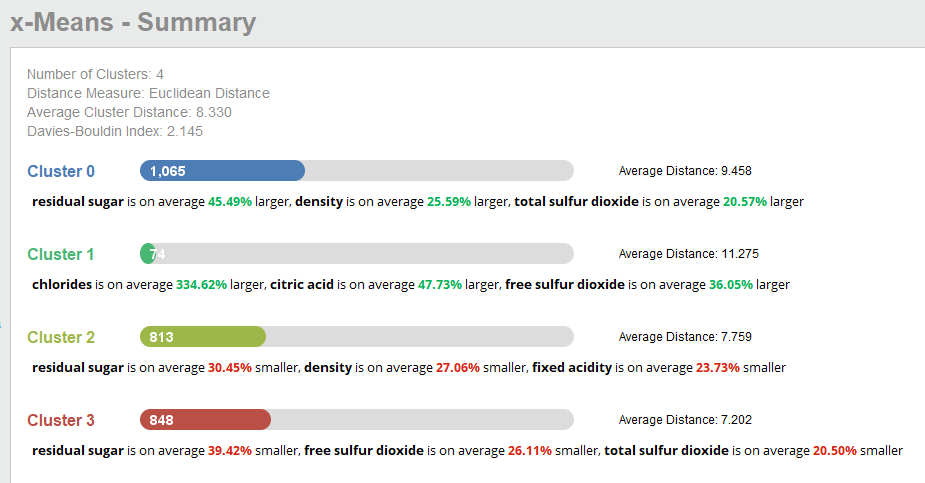
Scatterplots RapidMiner X-mean white wine

3. X- means clustering for white wine.

**In this analysis we did not need to choose number of clusters. The algorithm did it itself. X-means it is a variation of K-means. The results are:

Heatmap 2 – Rapidminer X-means White wines

In this algorithm we can see that chlorides have been associates in different cluster than the previous k-means algorithm. Interesting fact is those algorithms also choose doing 4 clusters. That means we made similar assumptions.



Summary Rapidminer X-means – white wines

By Davis Bouldin Index we got only 2.145. So its means that k-means was better choice in that case.

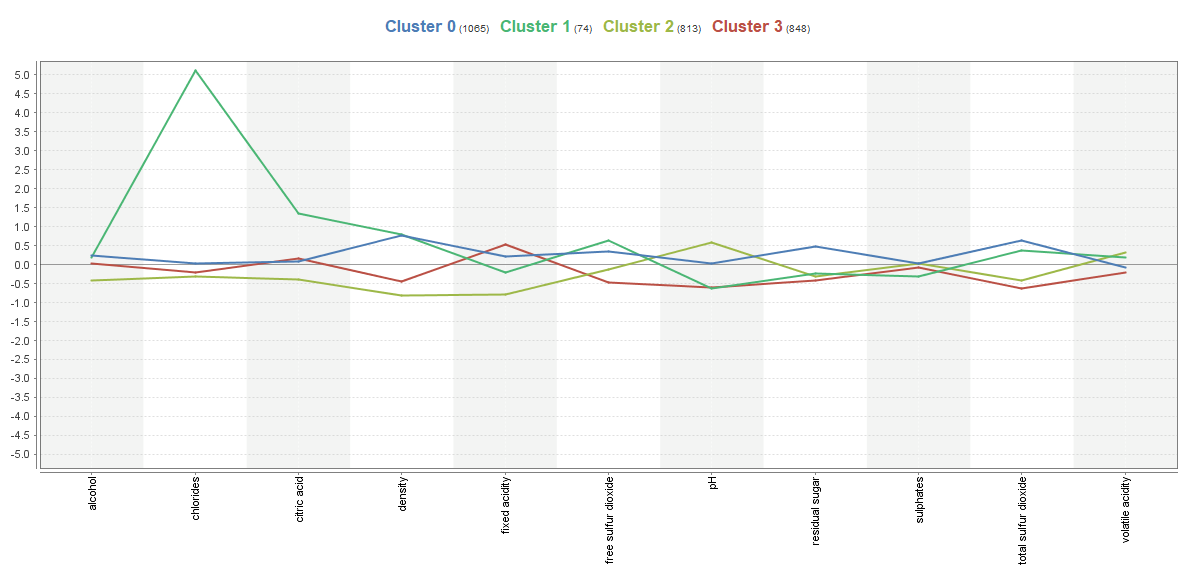
We can see that we have 4 clusters too.

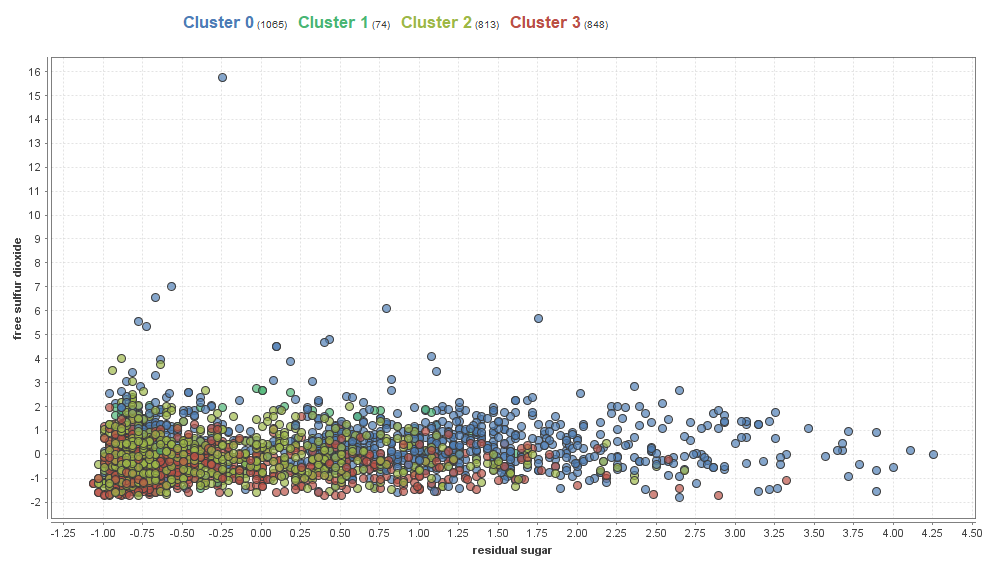
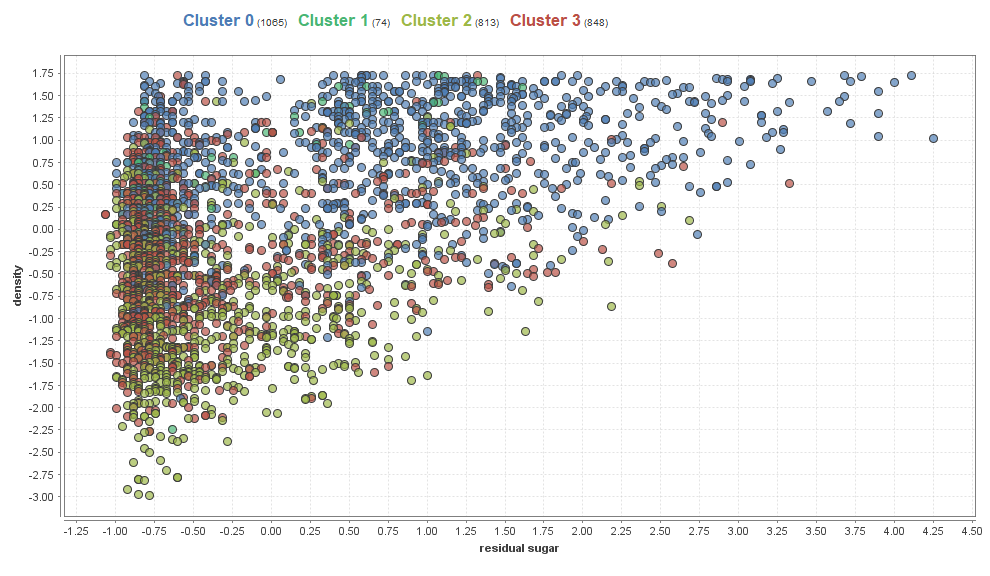
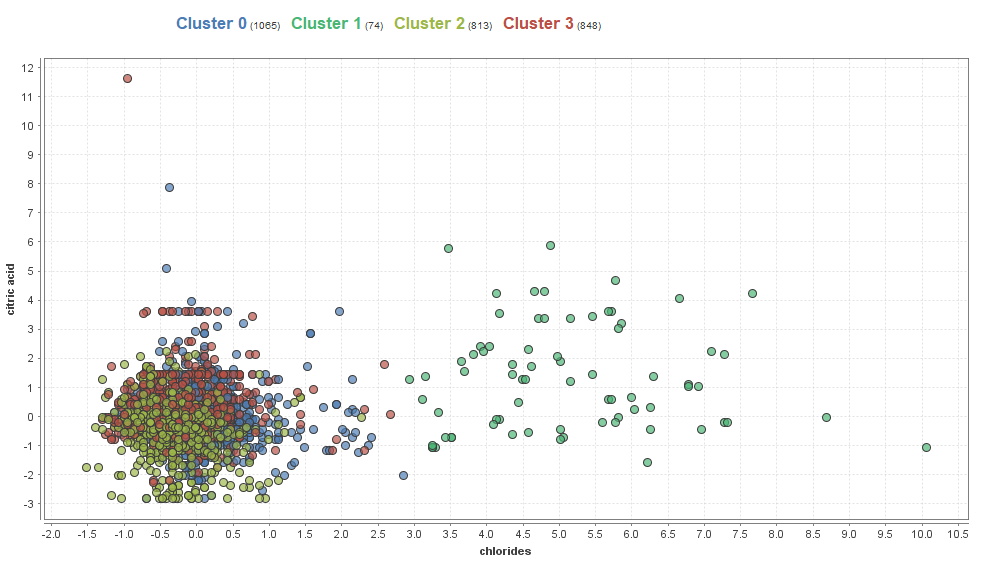
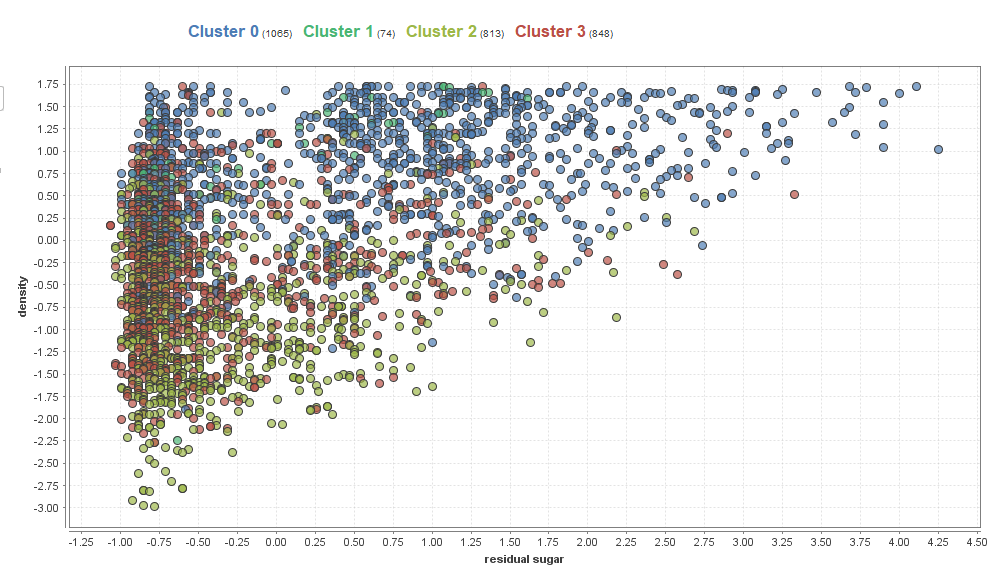
In cluster zero we can see that residual sugar are on average 45.49% larger than others, density is on average 25.59% larger and total sulfur dioxide is on average 20.57% larger then in others. It has large number of entries 1065 and average distance here is 9.45.

In cluster 1 we can see that chlorides is on average 334.62% larger, citric acid is on average 47.73% larger and free sulfur dioxide is on average 36.05% larger. There are only 74 entries here but average distance here is 11.27.

In cluster 2 residual sugar is on average 30.45% smaller density is on average 27.06% smaller and fixed acidity is on average 23.73% smaller. There are 813 entries in this cluster and average distance here is 7.7

In cluster 3 residual sugar is on average 39.42% smaller, free sulfur dioxide is on average 26.11% smaller, total sulfur dioxide is on average 20.5% smaller. There are 848 entries and average distance here is 7.2.

In tables we can see the differences between clusters. Below there is line plot with 4 clusters and scatter plots for all 4 clusters separately.



Scatter plots of 4 clusters of white wines x-means - Rapidminer

We found the following weaknesses in clustering:

* We cannot use it for categorical data as it applicable for continuous attributes only.
* In k-means we need to specify number of clusters
* We need to remove outliers.
* Only standard shape of clusters
* X-means seems to be variation of k-means but as in our example k-means was better choice in this particular case.

# **Conclusions**

New technology can be used in majority of industries. Perfect example here is manufactory of white wines.

We can see from the work above that all the ingredients in wine are connected. All of them have impact on each other.

Although it is still hard to predict the taste of wine and quality. We have to remember that attribute quality it is a subjective attribute when compare to others – scientific.

The Clustering analysis showed us that we can deal with 4 groups of clusters, showed us dependencies and correlations Total sulfur dioxide and free sulfur dioxide, Fixed acid and PH, Residual sugar and density.

Target attributes analysis of targets such us level of alcohol, sugar/Ph ratio, level of preservatives and Key Performance Indicator – Balanced wine prove that other ingredients of wine influence the important attribute and they cannot be checked in isolation.

By eliminating some attributes we just created our model weaker. They are all necessary.

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)
5. Han J, Kamber M., Pei J. 2012 *Data mining Concepts and Techniques,* p. 19 ISBN 978-0-12-381479-1 [↑](#footnote-ref-5)