Data mining project

Alicja Czajkowska

magda leszczynska

Paulina Luty

Anna Skowron

HDSDA

SEP2

Contents

[**Abstract** 2](#_Toc512548606)

[**Data set description and pre-processing** 3](#_Toc512548607)

[**Tools used and data overview** 6](#_Toc512548608)

[**Unsupervised learning – cluster analysis** 9](#_Toc512548609)

[**1.** **Correlations for clustering purposes** 9](#_Toc512548610)

[**2.** **K-means clustering for white wine** 10](#_Toc512548611)

[**3. X- means clustering for white wine** 14](#_Toc512548612)

[**Quality analysis** 18](#_Toc512548613)

[**Naïve Bayes** 18](#_Toc512548614)

[**Decision** **Tree** 22](#_Toc512548615)

[**Quality analysis of white wines – comparison of accuracy** 28](#_Toc512548616)

[**Target variable – level of preservatives** 28](#_Toc512548617)

[**Target attribute Sugar/PH ratio** 31](#_Toc512548618)

[**Target attributes - level of alcohol** 33](#_Toc512548619)

[**Target variable – Balanced wine** 35](#_Toc512548620)

[**Conclusions** 39](#_Toc512548621)

# **Abstract**

White Wines are very specific goods. The quality depends on many factors like barrels, type of grapes, location and ingredients. All that makes attributes of wine. Very important are activities carried over production process. The level of preservatives, level of alcohol and ph ratio and many other attributes depend on the production process.

The level of preservatives, level of alcohol and ph ratio and many other attributes can determine if the manufacturer of the wine will get desired certification and will not loose on quality of wine.

We used the classification technique – clustering to find the similarities between attributes and Decision Tree, Naïve Bayes, Generalised Linear Regression, Random Forest, Gradient Boosted Tree to find dependencies between physiochemical qualities of wine and the target attributes and to predict the target attributes.

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

**Problem Statement**

Preservatives in wines are necessary ingredient but they might be a problem also. How much is too much?

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

How about answering these questions by studying the properties of already made wines? It will certainly cost less than a process of developing a new quality wines. There is literature about manufacturing wines and this expertise is necessary, but we would like to show just small scope for improvement.

The quality management is very important in many industries. Data mining is a nowadays a necessary tool in this field. This project will focus on physiochemical properties of wine and connections and dependencies between them.

And hopefully it will help a little bit more in understanding the dependencies between different attributes. In this project we will take a holistic approach to quality of wine understood not just as a single attribute but an overall blend of various ingredients and attributes.

We decided to use sugar /pH ratio as the wine should not be either too sweet or too dry. 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 used in winemaking mainly because of its anti-oxidative and anti-microbial properties in wine. It is also used for cleaning purposes at the wineries (Carel, 2011). 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.

Balanced wine – we wanted to created perfect balanced wine, with balanced and healthy ratios of ingredients. Wine which will pass the certification of organic wine (low artificial preservatives) etc.

On the market there are many certificates available for wine merchants. Some of regulations are compulsory like European union regulations about percentage of alcohol and remaining sugar content. Some are voluntary like for example CCOF (Organic Vineyards), USDA ( USDA Organic –wines and others. (Organicvineyardalliance, 2017). For organic certificates you need to pass many regulations including level of preservatives and other chemical allowed.

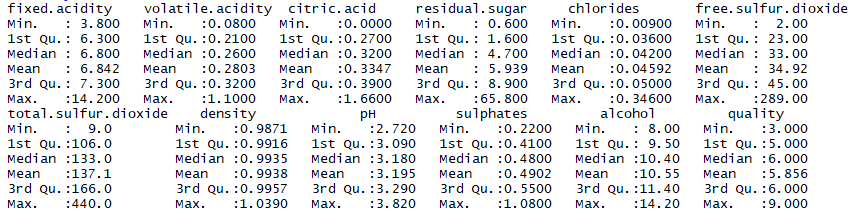
# **Data set description and pre-processing**

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, Cerderira, Almeida, Matos, Reis, 2009). 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 were 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.

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). 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 |
| balanced wine | - | binomial | - | balanced, unbalanced |

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%

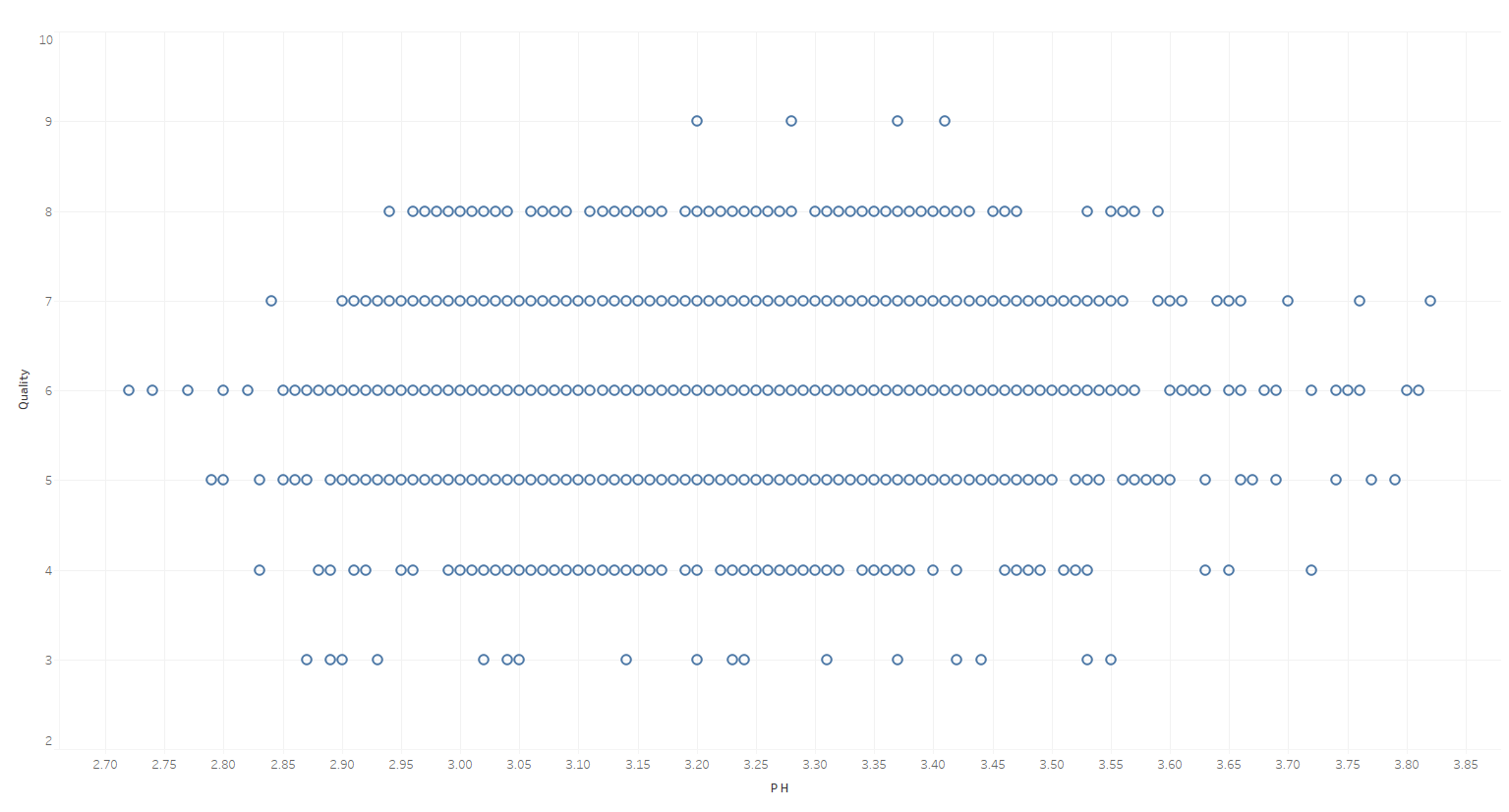
Medium – 9-11%

Low – 9% and less

Balanced wine:

Balanced - If total sulfur dioxide <=200 and ph ratio<=3 and level of alcohol = medium

Unbalanced - others

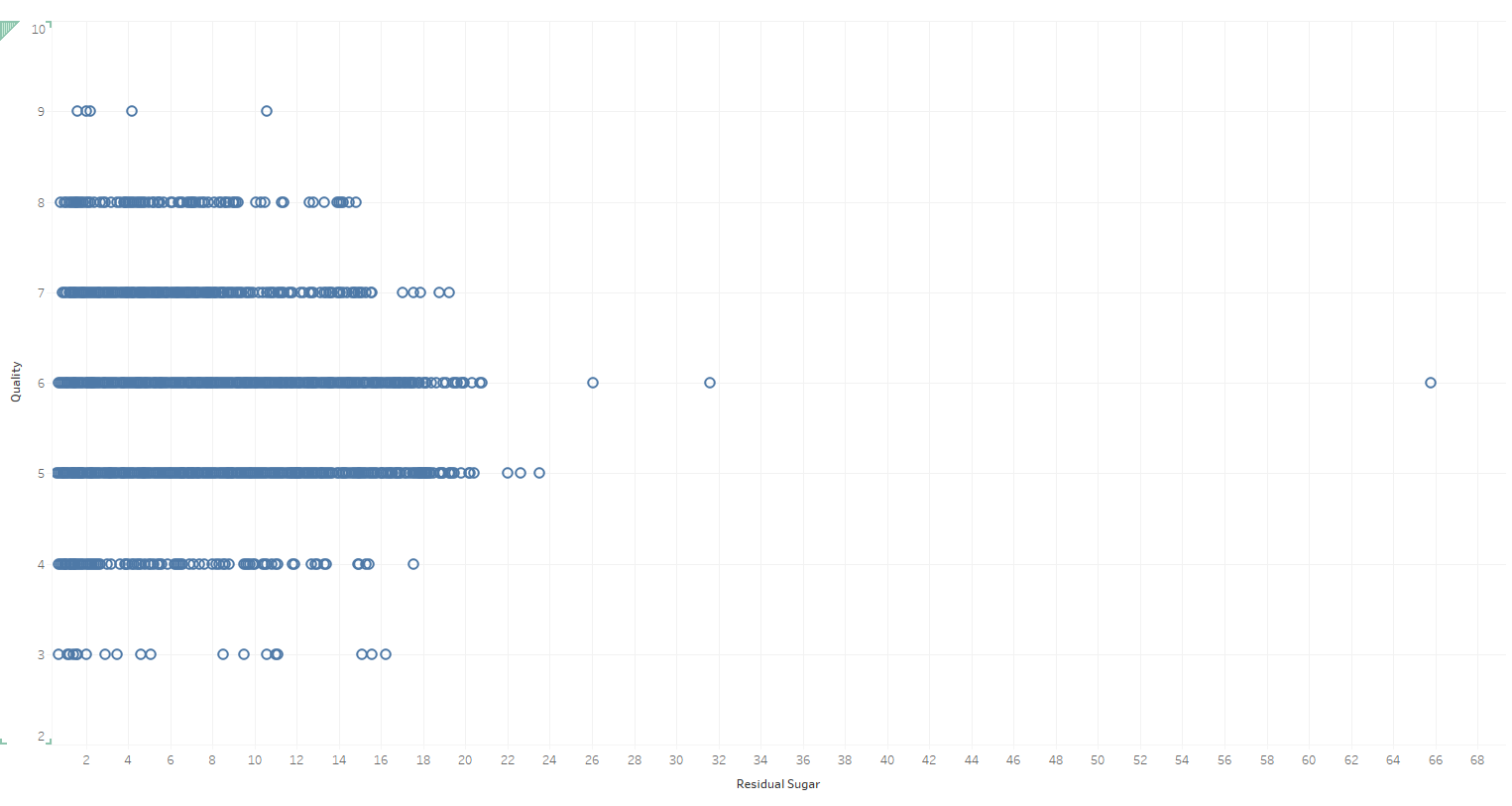


*Figure 1.1. PH to quality scatter plot*

# **Tools used and data overview**

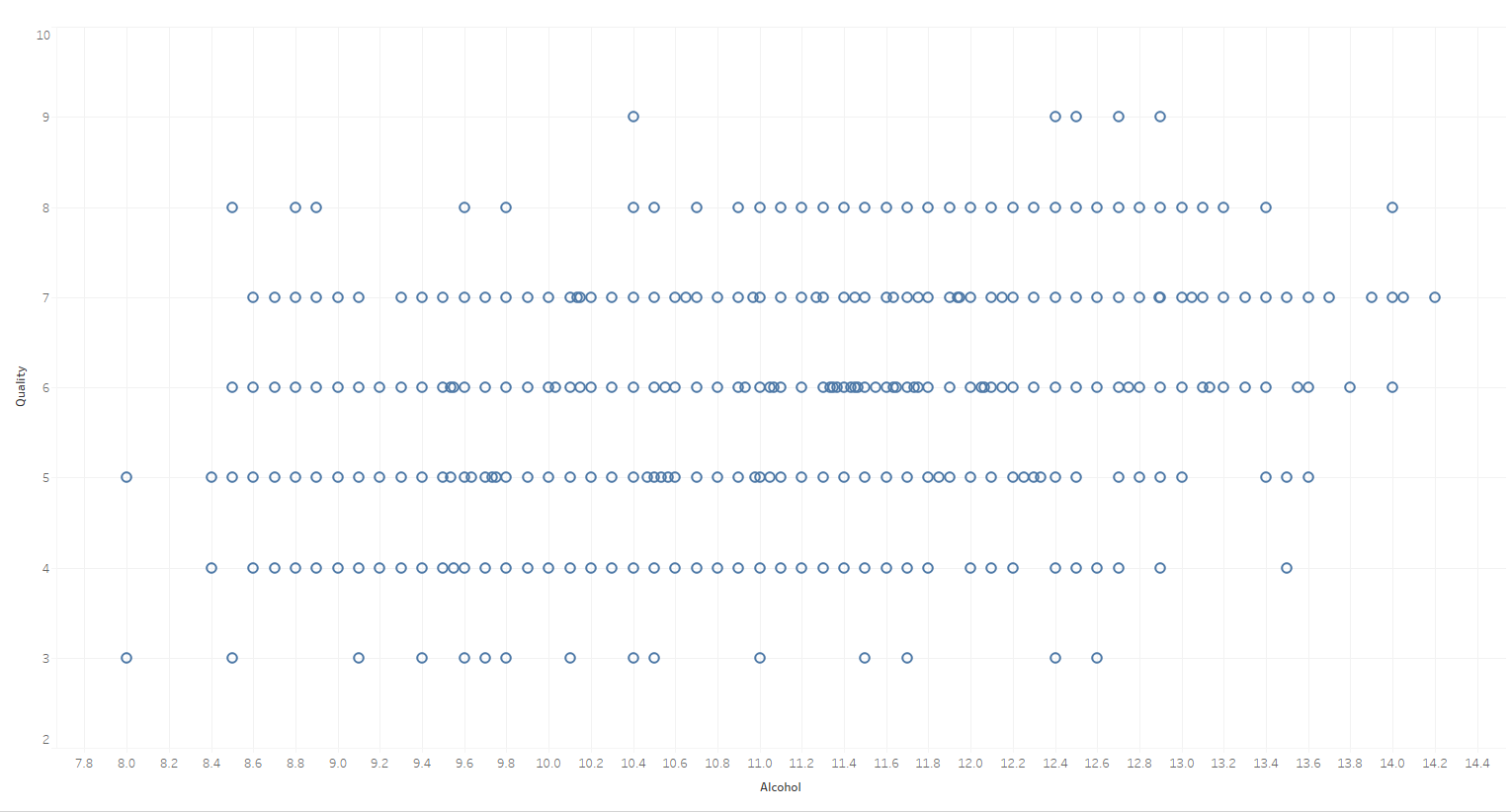
We used Tableau, Excel, Weka, RapidMiner, R and Python in our project.

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

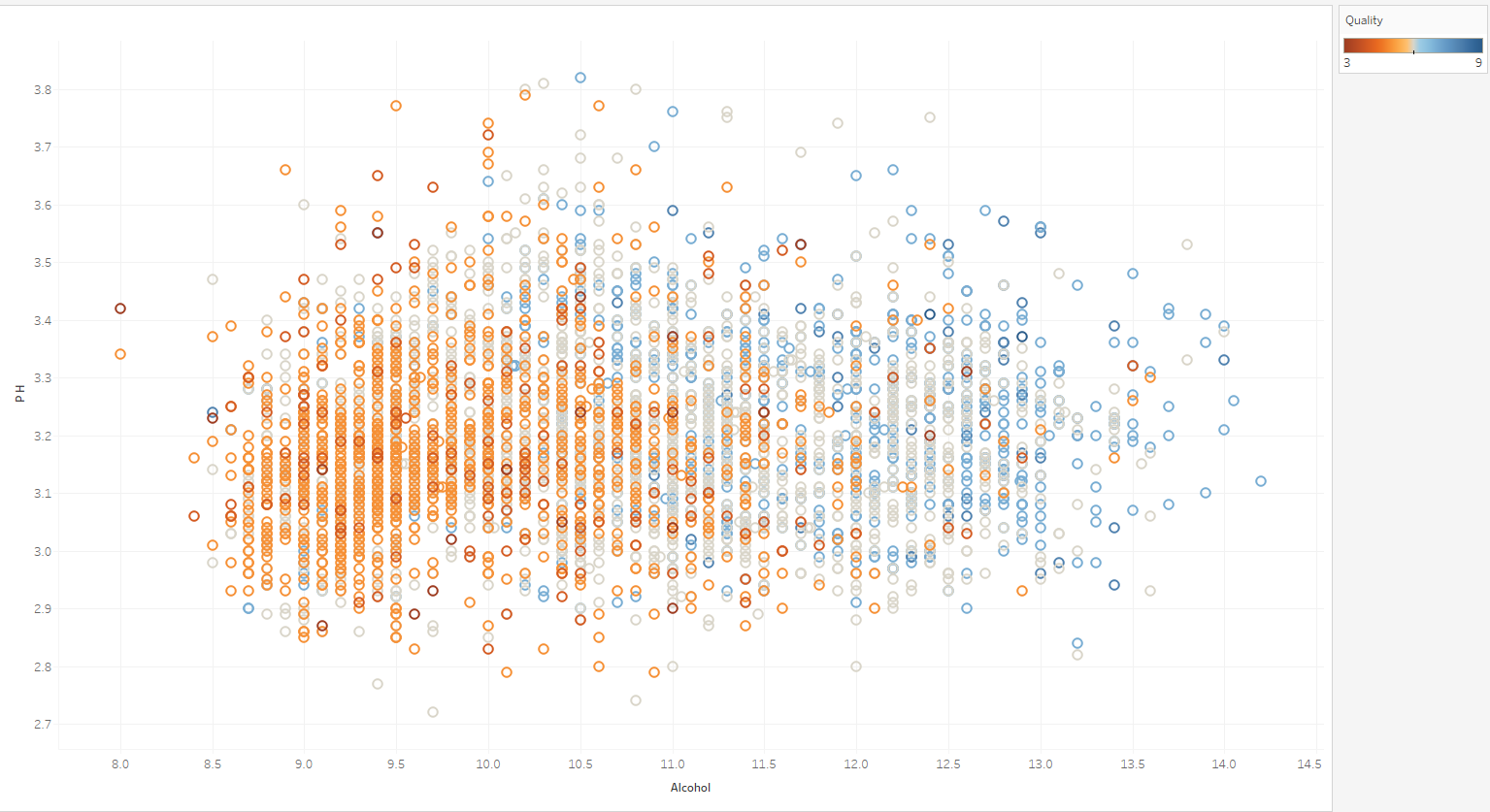


*Figure 1.2. Sugar to quality scatter plot*

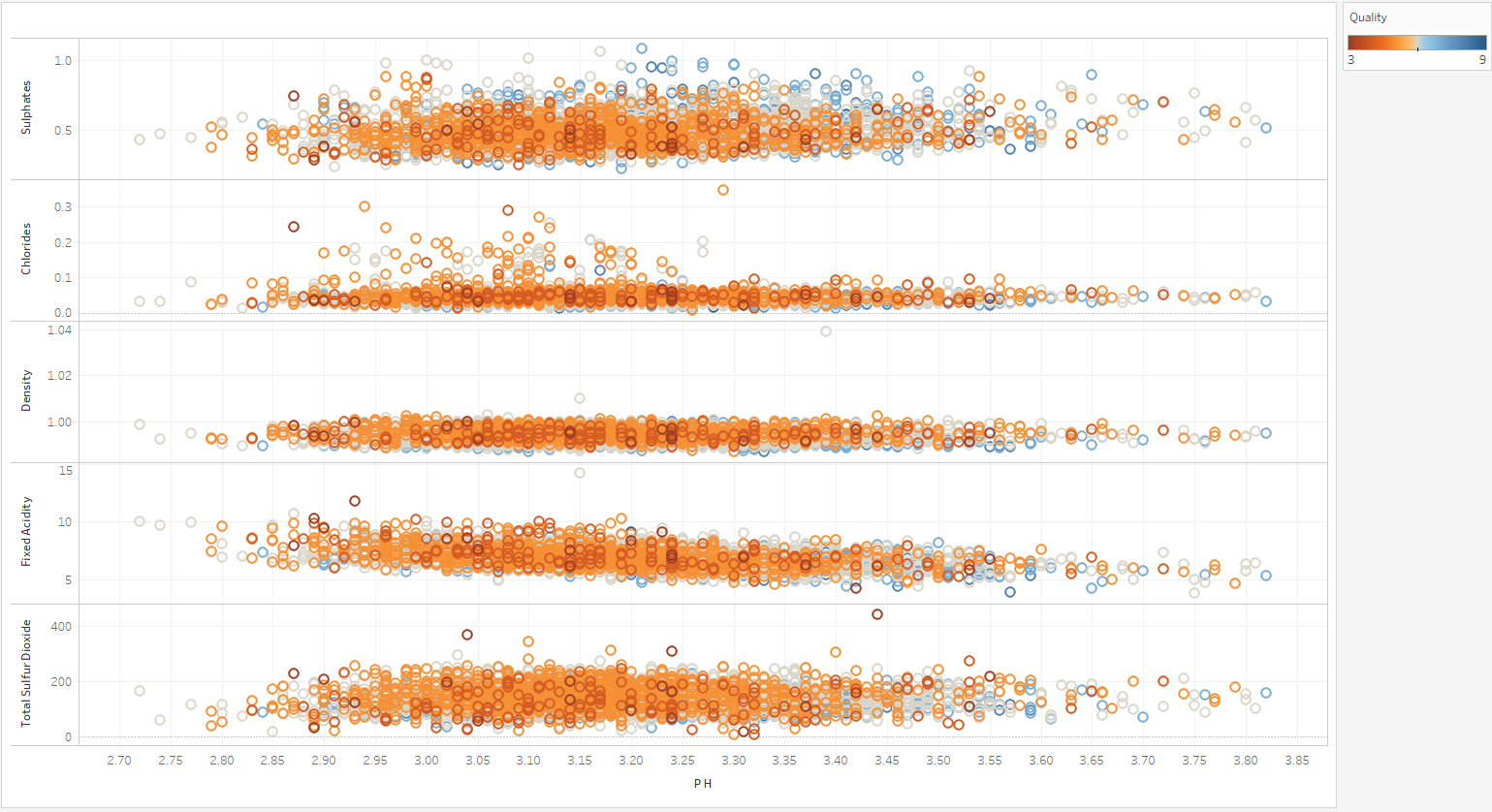
There is an outlier on this graph - a sugar at level of 65.8. We decided to leave that outlier in our model as the amount of sugar in white wines can vary from 0 to 220 grams per litre (WineFolly, 2015). It is possible that one of the tested wines had that amount of sugar.



*Figure 1.3. Alcohol to Quality Scatter plot*



*Figure 1.4. Alcohol to pH with quality as a factor scatter plot*



*Figure1.5. 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).

# **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 (Han, Kamber, Pei, 2011).

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:

## **Correlations for clustering purposes**

*Table 1.1. Correlation table – RapidMiner – white wines*

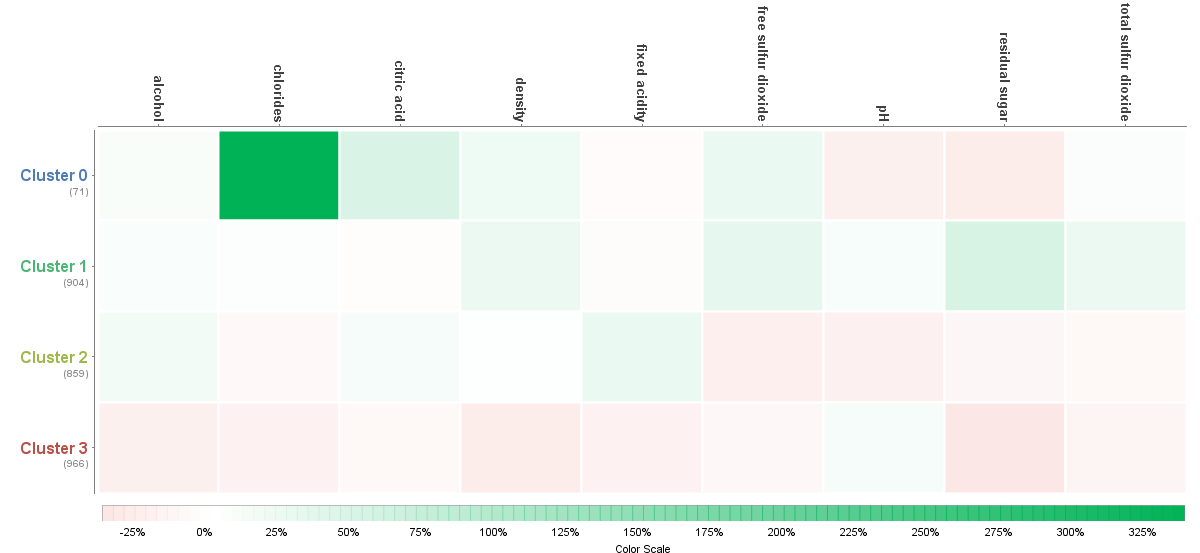
Highest correlations between the attributes in our data set are:

* 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.

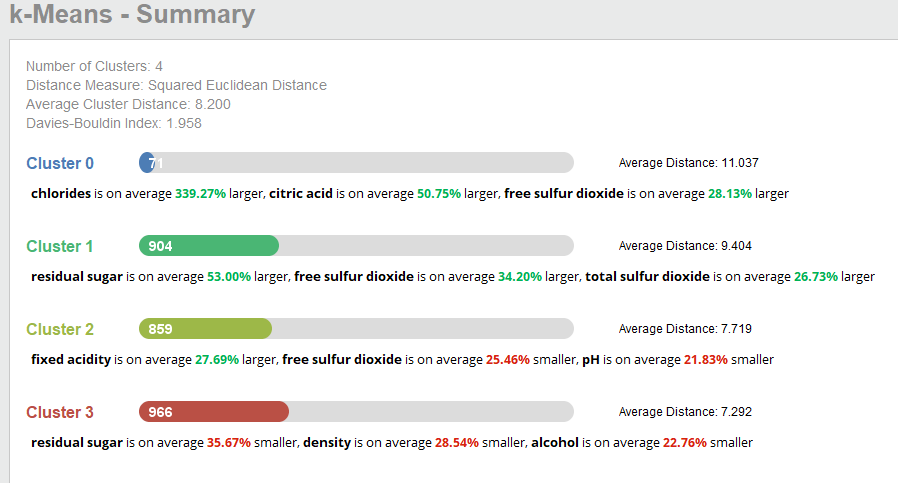
## **K-means clustering for white wine**

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:



*Figure 1.6. 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 natural volume of chlorides.



*Figure 1.7. 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.

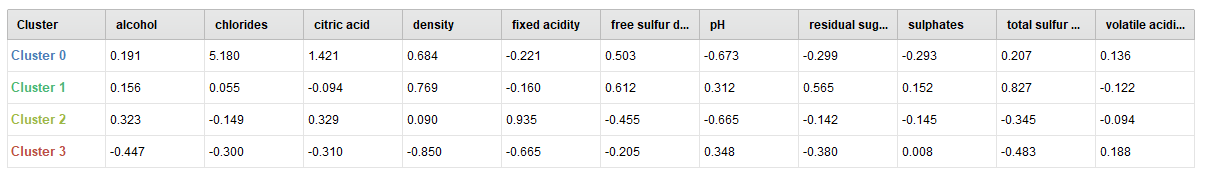
By Davis Bouldin Index we got only 1.958. It is an index for evaluation of cluster algorithms. As closer to zero than the clusters have been better assigned and they are closer to each other.

In cluster 0 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.

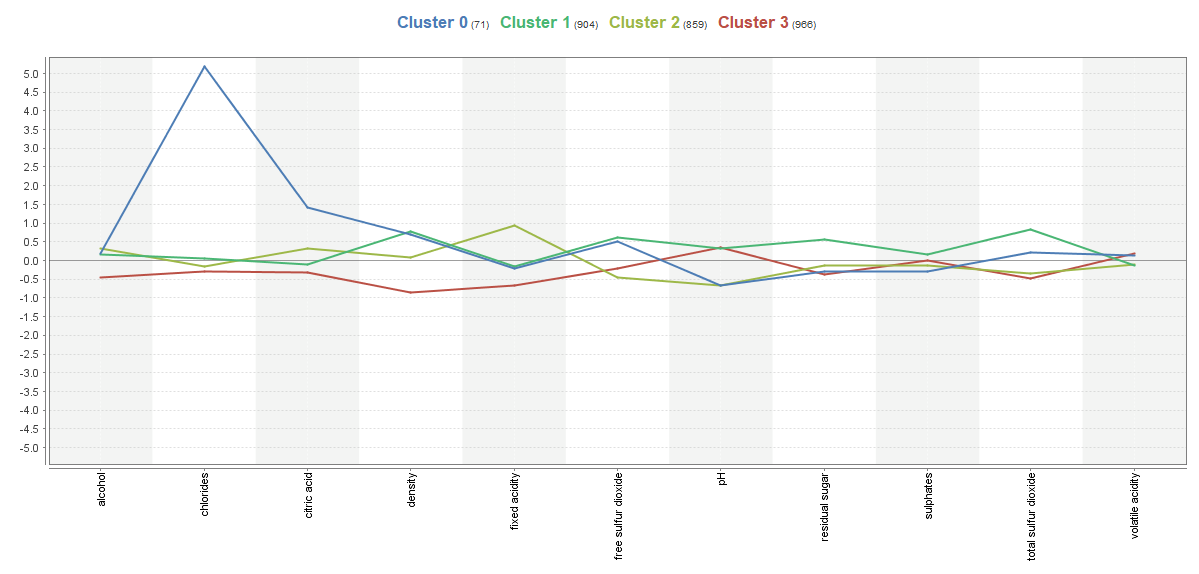
In cluster 1 we can see that residual sugar is on average 53% larger, free sulfur dioxide is on average 34.20% larger and total sulfur dioxide is on average 26.73% larger. There are 904 entries in this cluster but average distance here is 9.4.

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

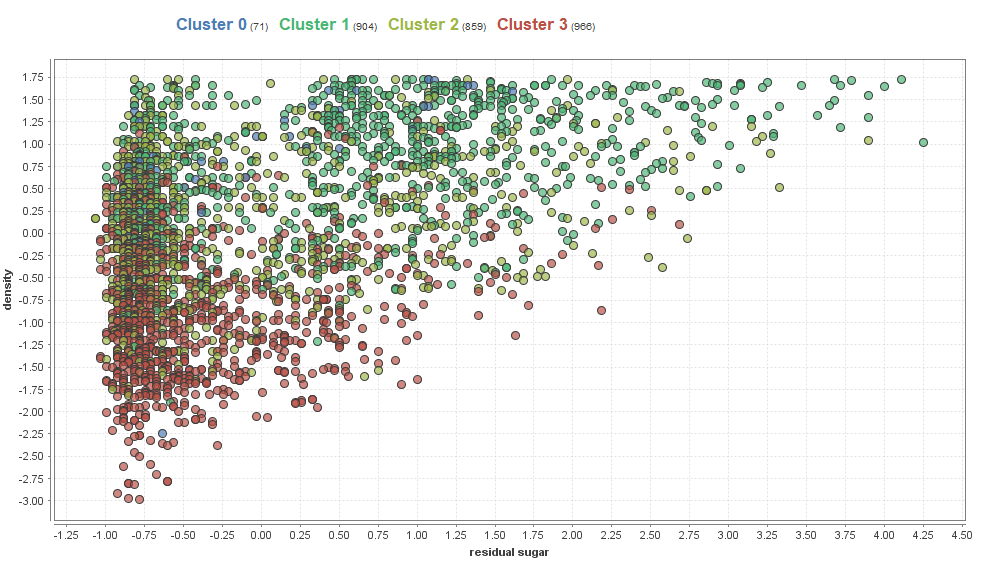
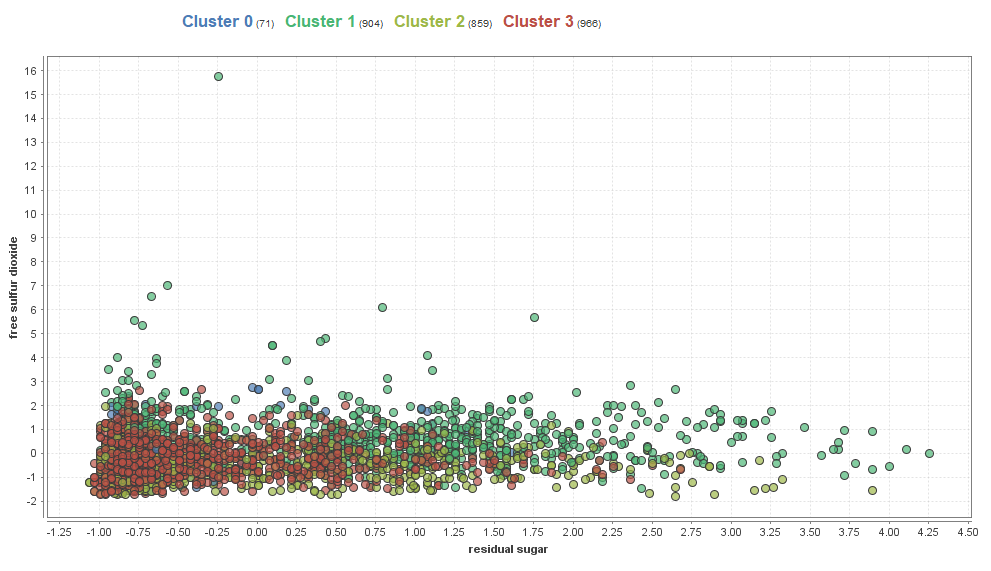
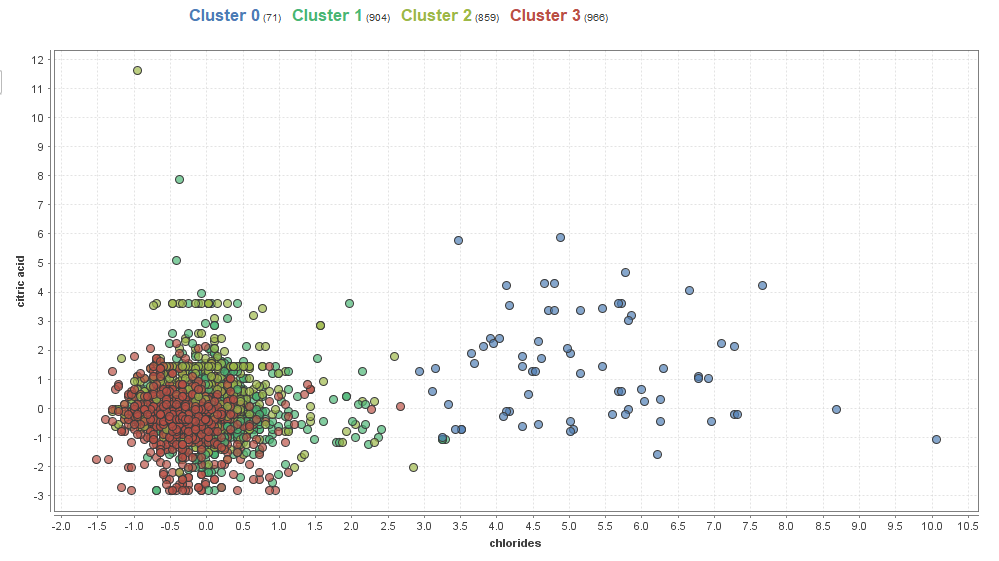
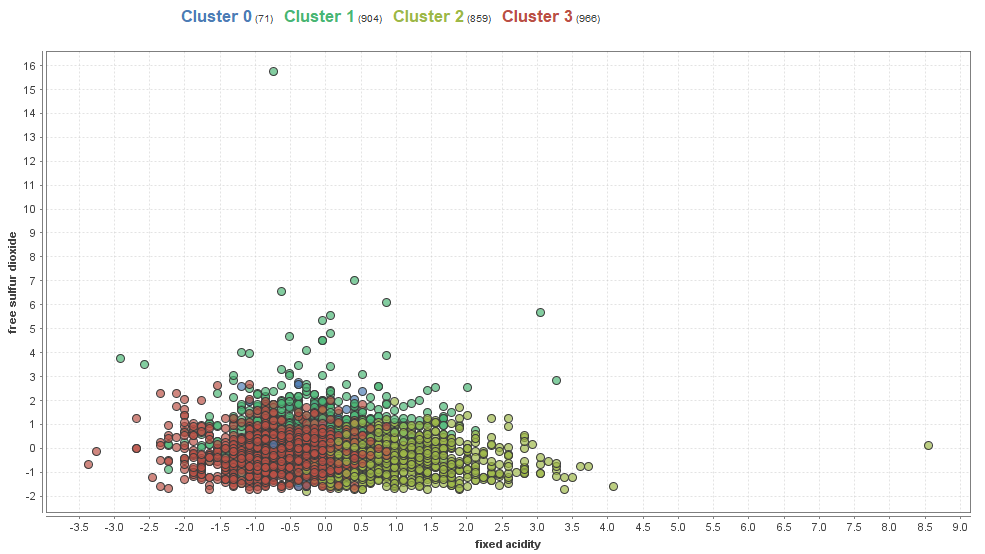
In cluster 3 residual sugar is on average 35.67% smaller, density is on average 28.54% smaller, alcohol is on average 22.76% smaller. That is the biggest cluster with 966 entries and average distance here is 7.2.

*Table 1.2. 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 (Fig. 1.8).

**

*Figure 1.8.* *Line plot – Rapid miner k-means – White wine*

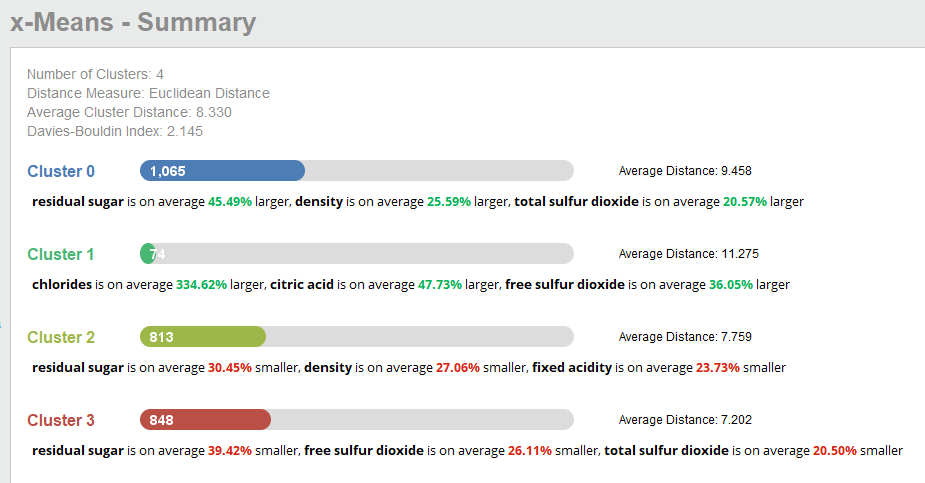
*Figure 1.9. Scatterplots RapidMiner X-mean white wine*

## *\\BOBSRV01\Folder Redirection\magda\Desktop\New folder\DA\DM\CA2\clust\x-means heat.PNG***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:

*Figure 1.10. Heatmap 2 – Rapidminer X-means White wines*

In this algorithm we can see that chlorides have been associated 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.



*Figure 1.11. 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 0 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.

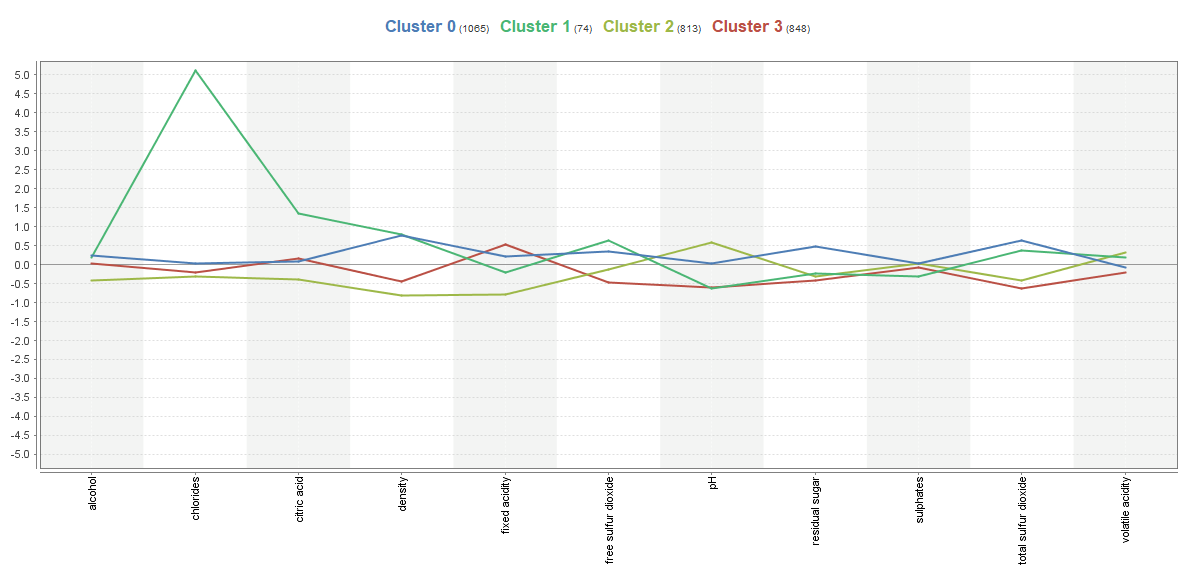
In cluster 1 we can see that chlorides is on average 334.62% larger, citric acid is on average 47.73% larger and free sulfur dioxide is on average 36.05% larger. There are only 74 entries here but average distance here is 11.27.

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

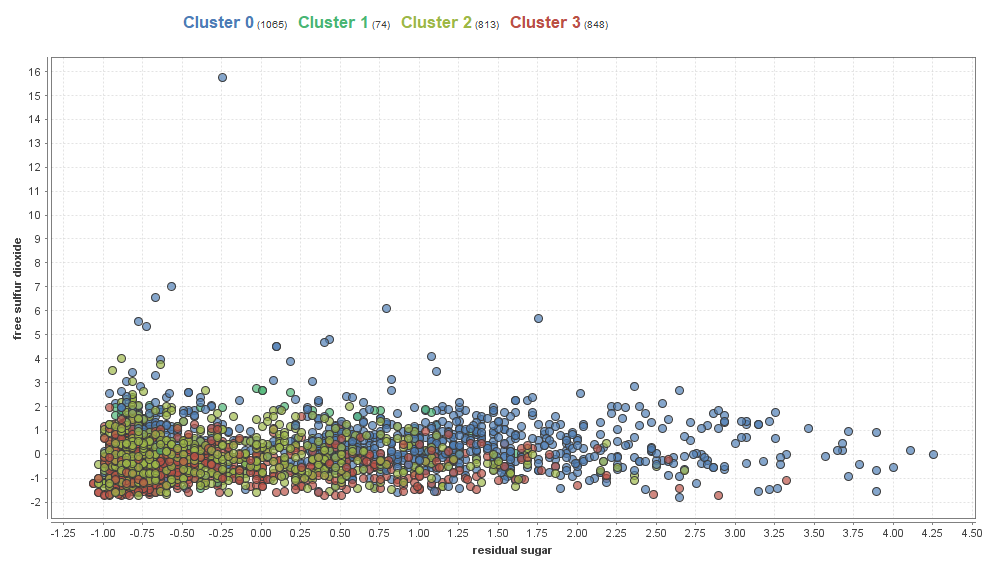
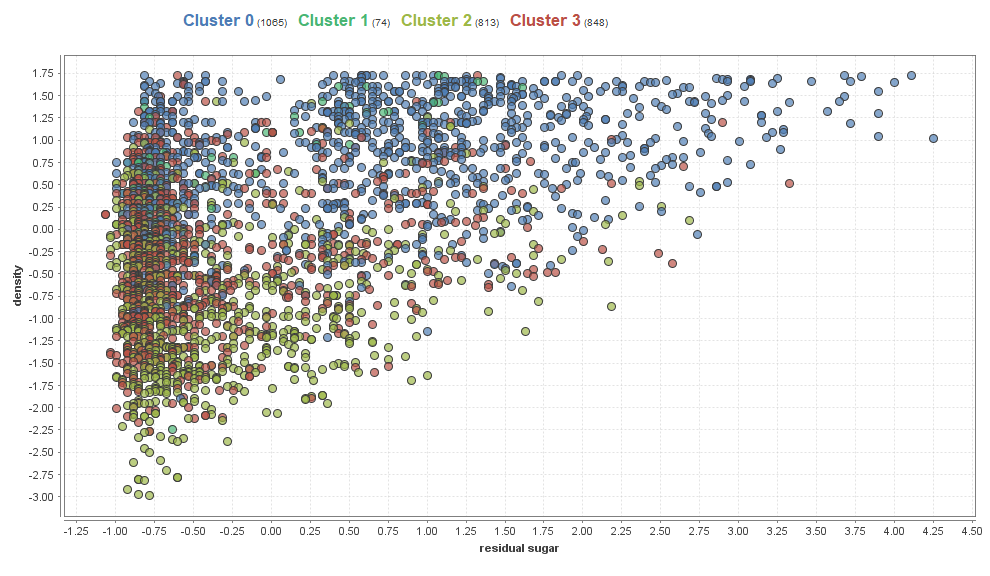
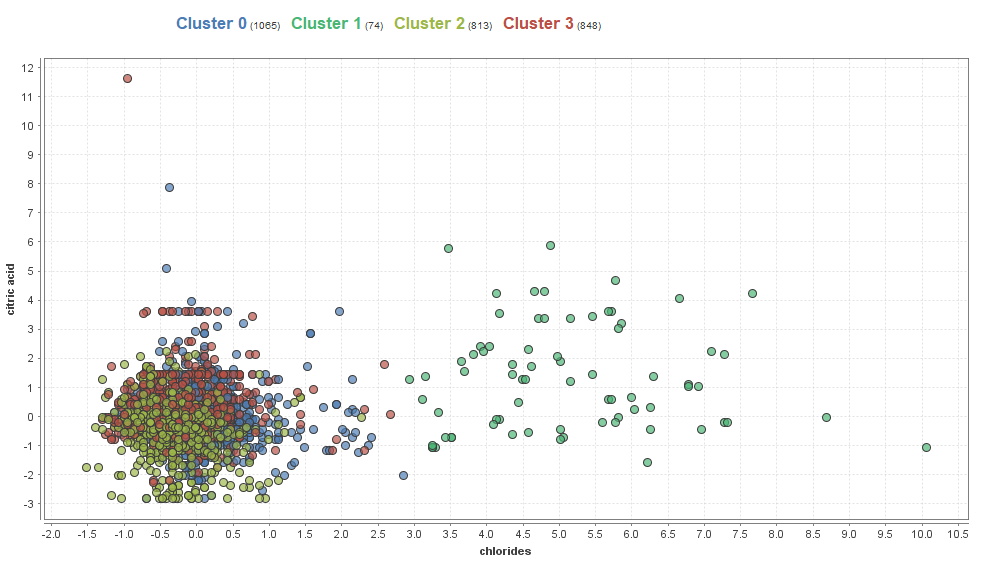
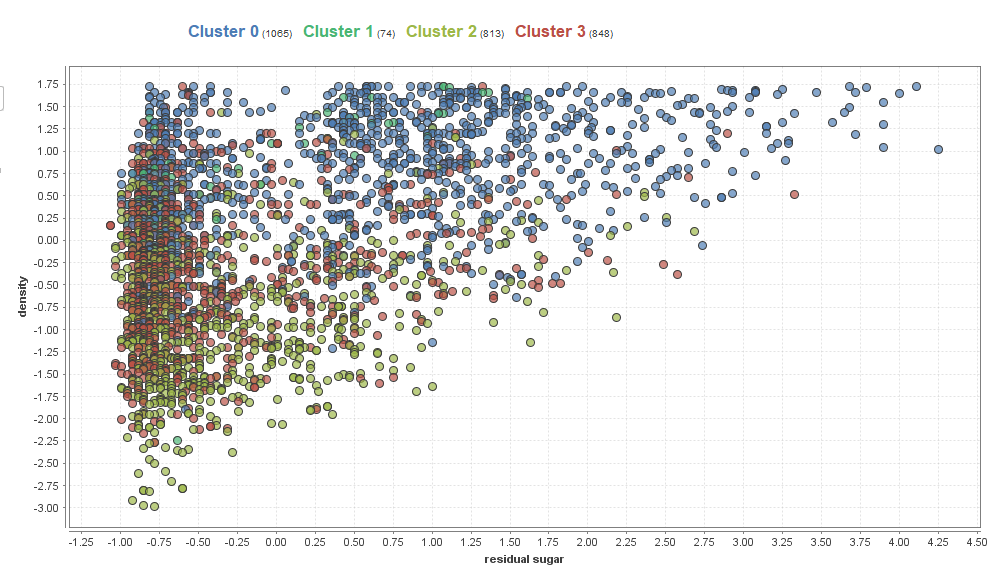
In cluster 3 residual sugar is on average 39.42% smaller, free sulfur dioxide is on average 26.11% smaller, total sulfur dioxide is on average 20.5% smaller. There are 848 entries and average distance here is 7.2.

**

*Table 1.3. Rapidminer – Table of clusters*

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.

*Figure 1.12. Rapidminer – Lineplot – X-means*



*Figure 1.13. 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.

# **Quality analysis**

We will use classification models including decision trees and naïve Bayes and clustering.

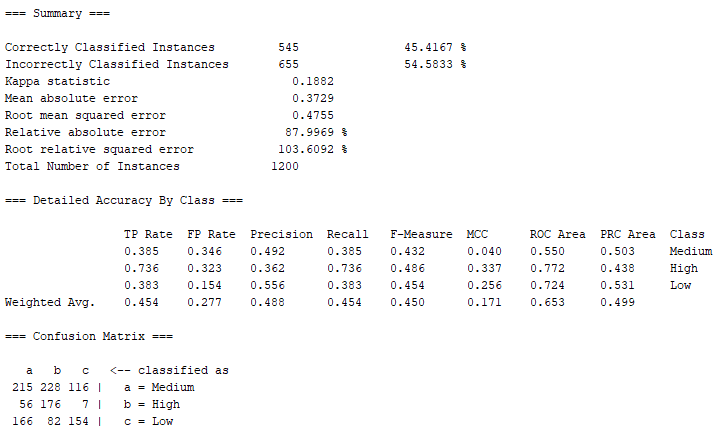
Decision tree is a classification method and a predictive technique. Classification methods use existing data to create a model that will allow to classify new data. In our project we will use decision tree to predict quality and classify alcohol content and sugar/ pH ratio, level of preservatives and wine balance.

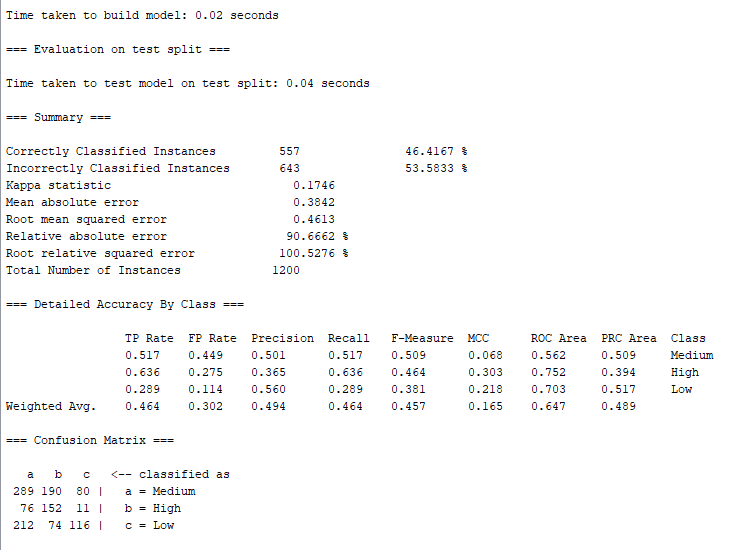
Clustering is a descriptive technique that finds groups of observations (clusters) that share similar characteristics in a data set.

## **Naïve Bayes**

Naïve Bayesmodel was run in Weka using the dataset with 4,000 instances (split 70 /30 - 70.0% of data is for 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 1.14).

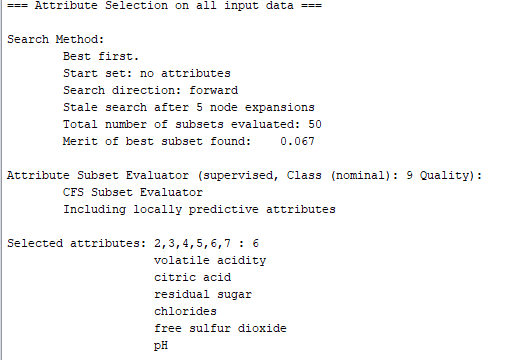




*Figure 1.14. Naïve Bayes classifiers’ output comparison*

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. 1.15).

**

*Figure 1.15. 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%.

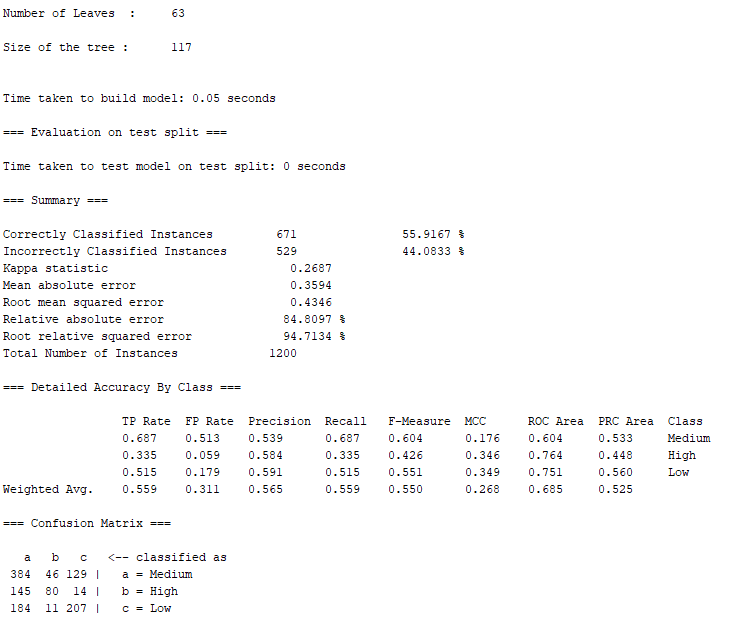
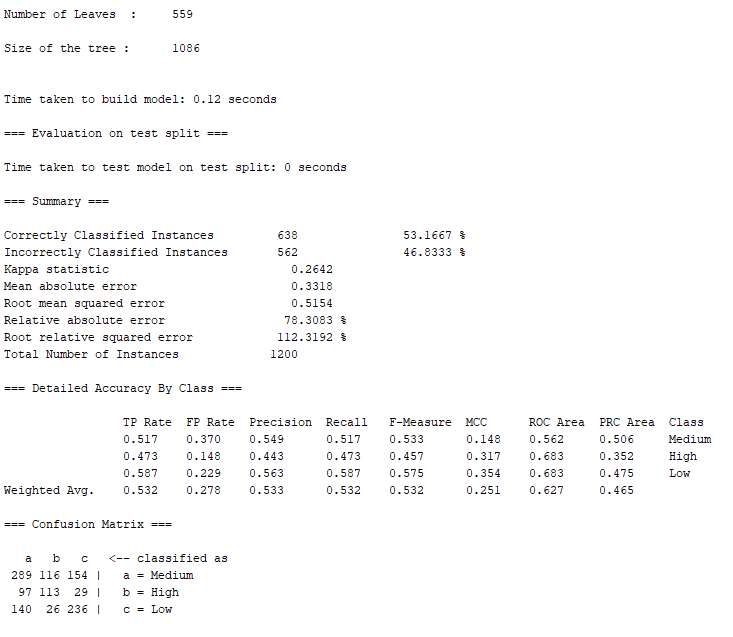
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Class** | |  |  |
| **Attribute** | | **High** | **Medium** | **Low** |
|  | | 0.21 | 0.45 | 0.34 |
| **fixed acidity** | |  |  |  |
| mean | | 6.6968 | 6.8267 | 6.9602 |
| std. dev. | | 0.7846 | 0.8485 | 0.9121 |
| weight sum | | 834 | 1806 | 1360 |
| 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 1.4. Naïve Bayes classifier model*

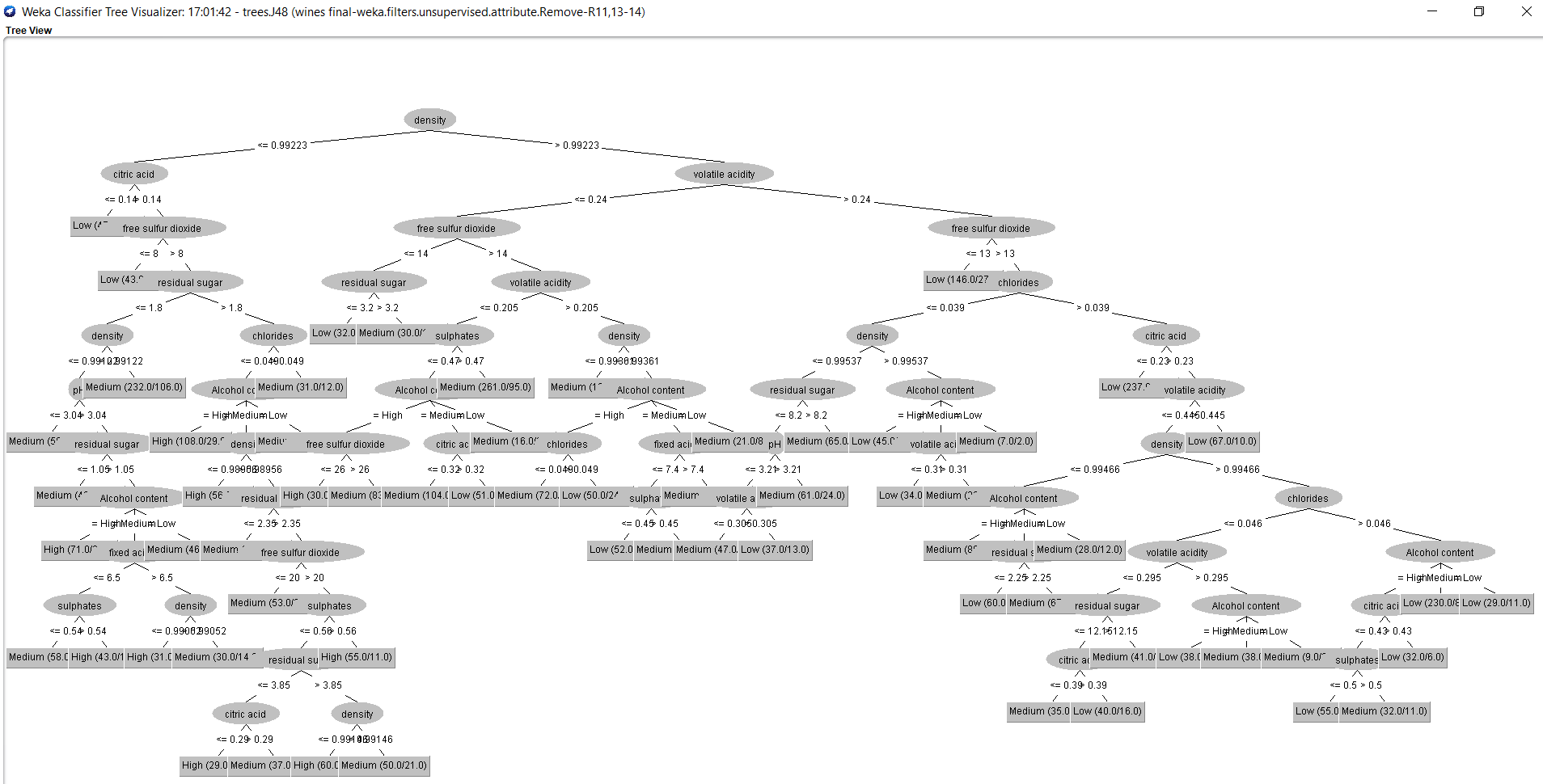
It would appear that the quality is decreasing with higher level of fixed acidity, residual sugar and chlorides and it is improved with higher pH level.

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

Decision Tree J48 model was executed in Weka using the dataset with 4,000 instances (split 70/30) 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) to predict Quality attribute. 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%.



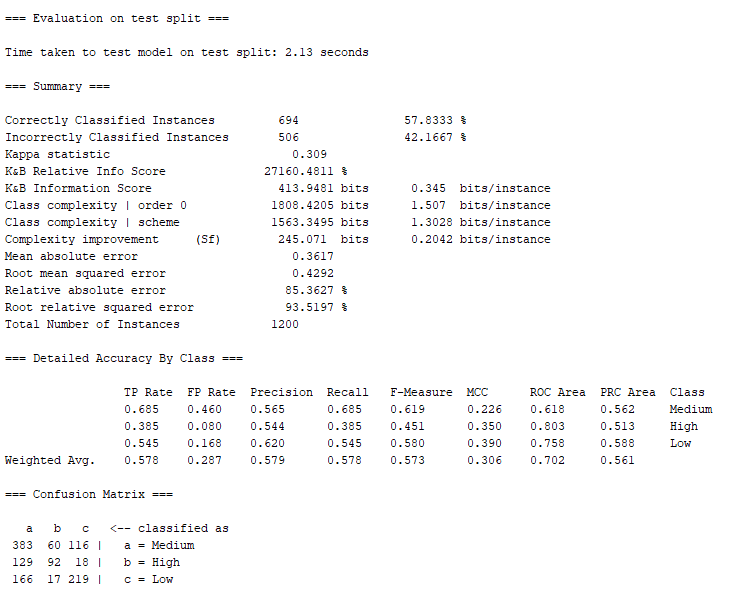
*Figure 1.16. Decision Tree J48 models’ comparison*



*Figure 1.17. Decision Tree J48 model - Weka*

We also used Support Vector Machine model in Weka – SMO to predict the quality attribute. The accuracy of that model was 57.8% (Fig. 1.18).

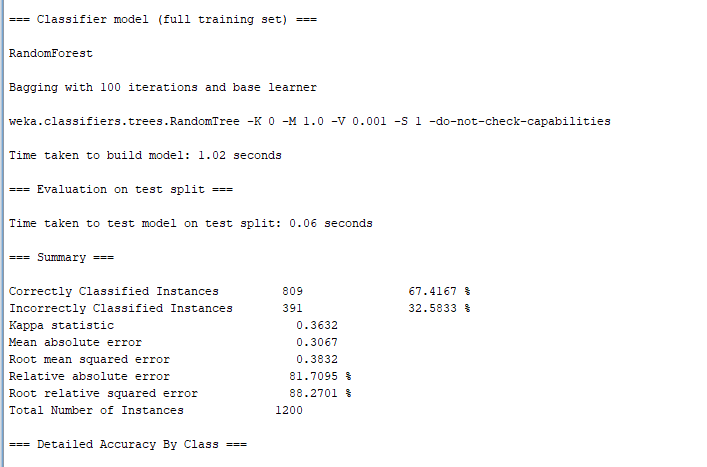
Support vector machine algorithm transforms training data into a higher dimension, where it searches for “decision boundary” (linear optimal separating hyperplane) that separates the data from one class from another. To find this hyperplane SVM model uses support vectors (“essential” training tuples)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



*Figure 1.18. SMO model Weka – evaluation output*

The best accuracy level 67.4% (Fig. 1.19) 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.

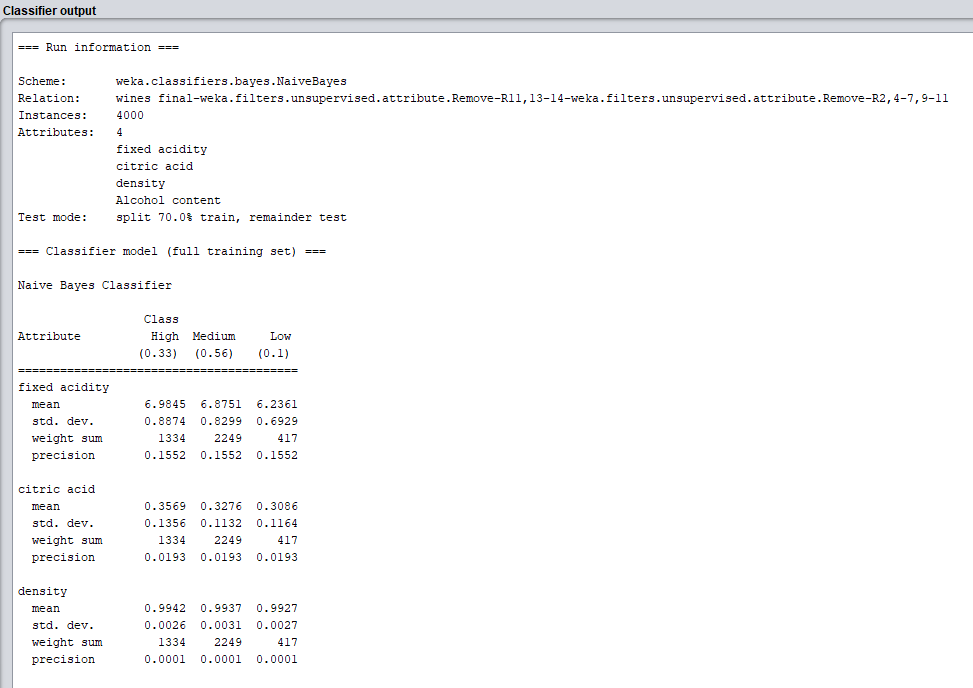


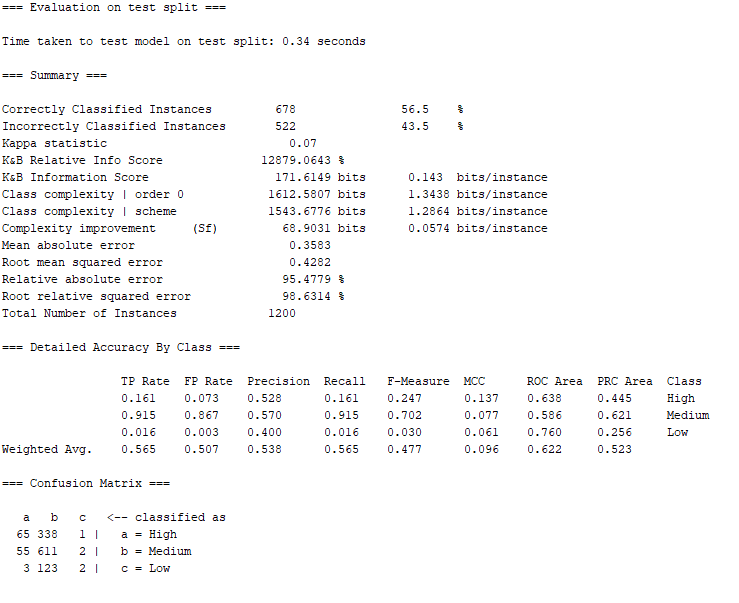
*Figure 1.19. Random Forest model in Weka*

**Classification Models – alcohol content**

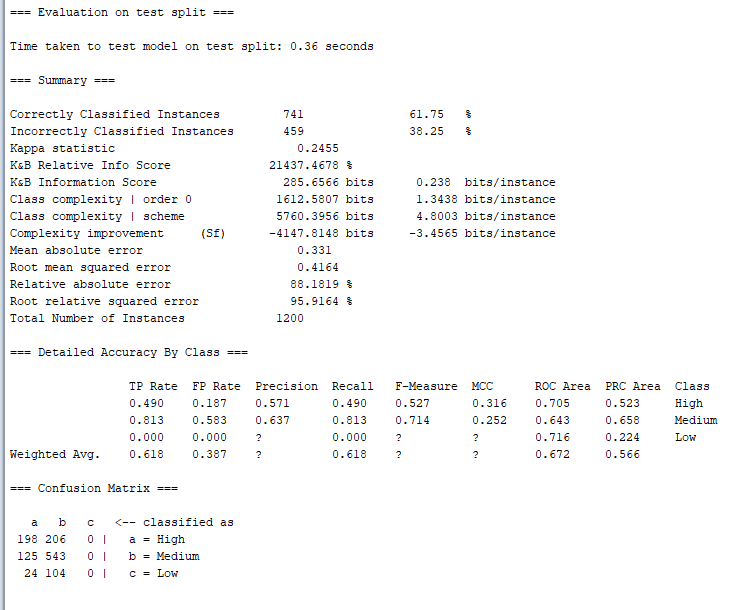
We also used **Naïve Bayes and Decision Tree** models to review alcohol content. The original alcohol values were discretised: values lower than 9% were classed as low, 9 to 11% as medium and above 11% as high. After running the correlation-based feature selection (CFS) algorithm in Weka to we eliminated all attributes but fixed acidity, citric acid and density to classify alcohol content. The accuracy of the model was 56.5% (Fig. 1.20). 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. 1.21).

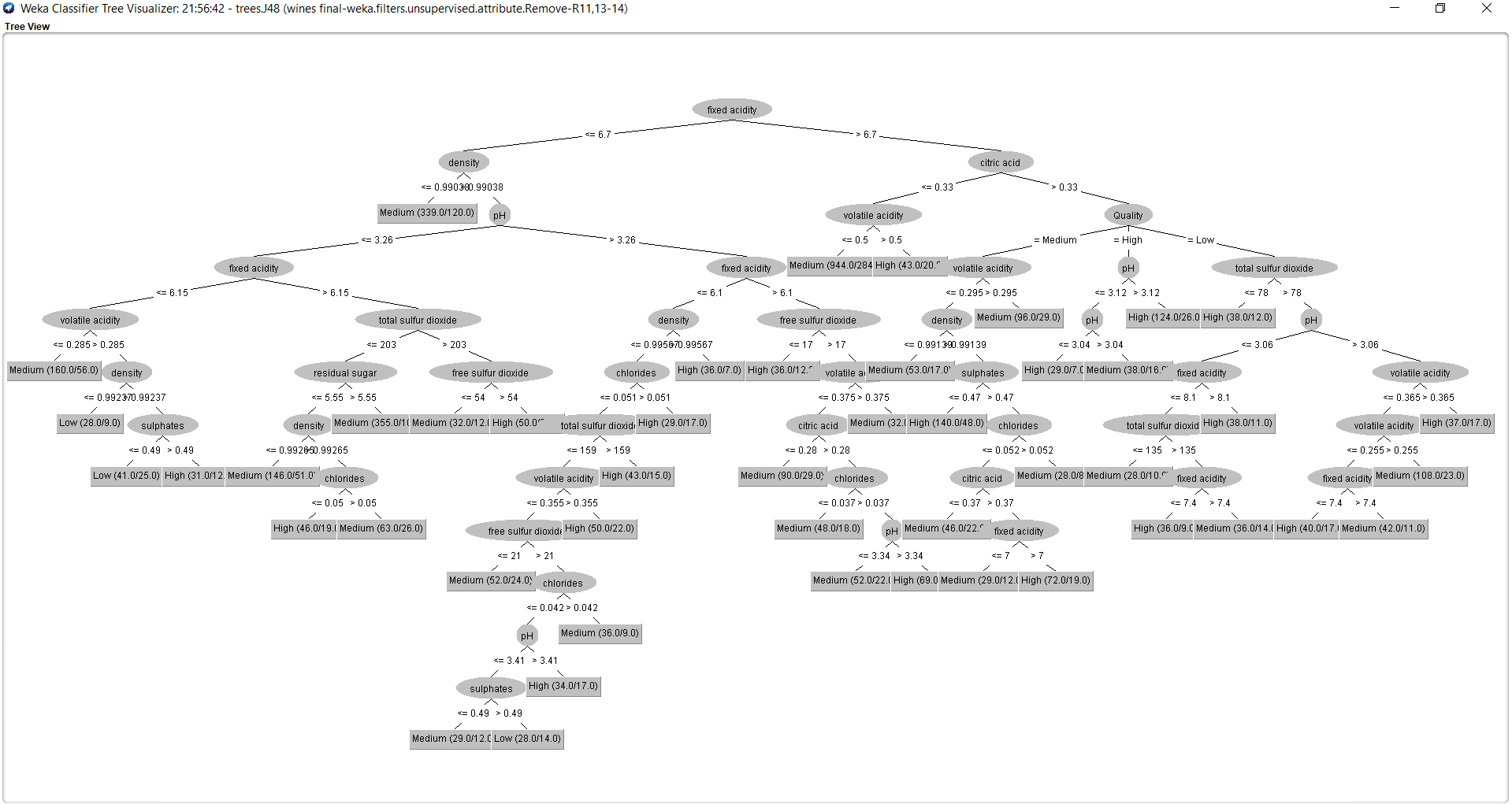




*Figure 1.20. Naïve Bayes model – alcohol content*



*Figure 1.21. Decision Tree J48 – alcohol content prediction – evaluation output Weka*



*Figure 1.22. Decision Tree J48 model – alcohol content prediction (Weka)*

# **Quality analysis of white wines – comparison of accuracy**

**Methods used:**

* Naïve Bayes
* Generalised Linear model
* Decision Tree
* Random Forest
* Gradient Boosted Trees

## **Target variable – level of preservatives**

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

These attributes has been created for this purposes – 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 are 3 types of output Low, Medium and High (Ordinal Polynomial attributes).

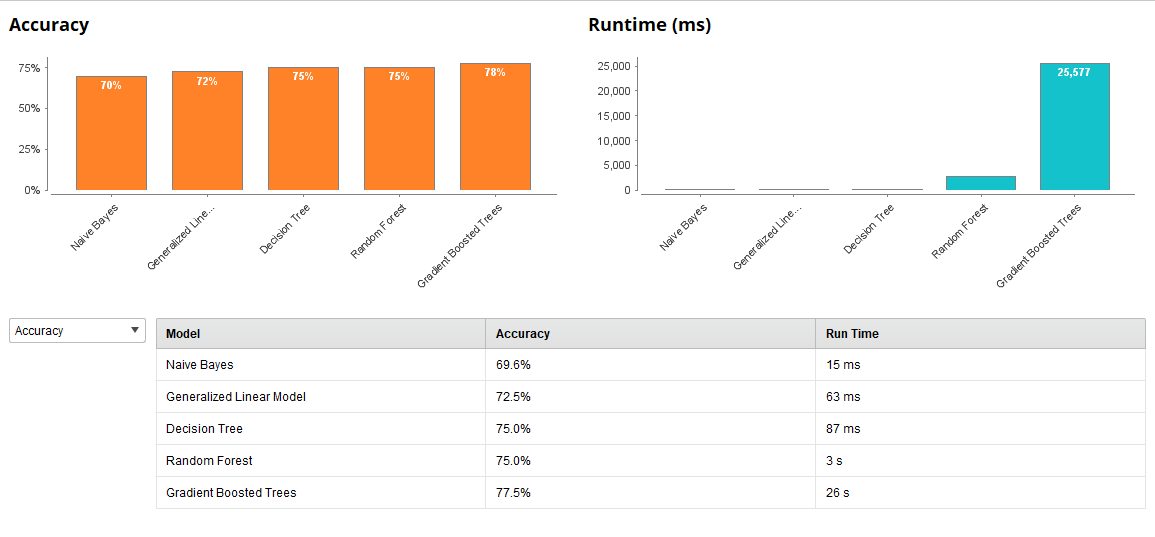
The aim of this analysis is to check what will happen with one variable if we will change or delete the other. In that case we have deleted chlorides.

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

Summary of model outputs.

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

Training:



*Figure 1.23. Training set – level of preservatives – Rapidminer output*

A screenshot of a cell phone

Description generated with very high confidence

*Figure 1.24. Test set - Cross – validation – level of preservatives – Rapidminer output*

Summary of models’ outputs

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.

That’s mean that even without chloride attribute the other attributes of white wines can give us information about level of preservatives at 77.5% chances that we are correct.

Training set had higher score of 89%, but still model is quite good.

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**

This ratio was picked not without discussion and further 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?

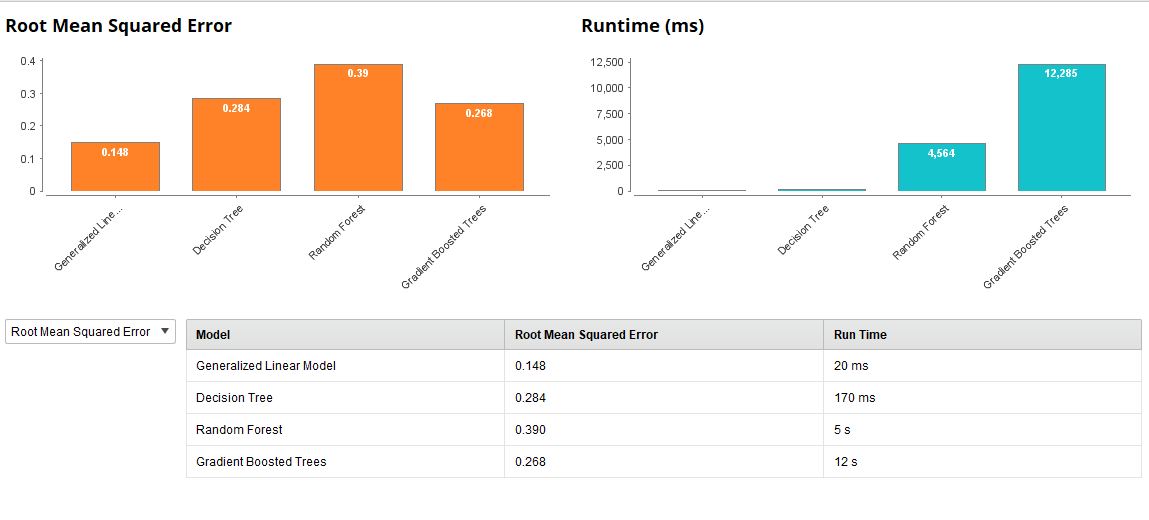
Methodology used as in previous example: training 70%, test 30% and we have deleted PH attribute.

A screenshot of a social media post

Description generated with very high confidenceTraining:

*Figure 1.25. Summary of model outputs with given attribute – root mean squared error*

Cross -validation:

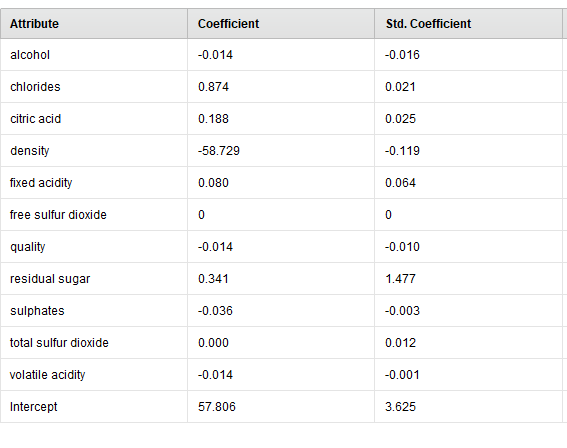
Test:

*Figure 1.26. Summary of model outputs with deleted attribute – Ph*

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

Error in both training 0.043 and test 0.268 is quite small. We can assume that model is quite good.

Output below:



That’s is 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 are 3 types of output Low, Medium and High (Ordinal Polynomial attributes).

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

Training:

Summary of model outputs with given attribute.

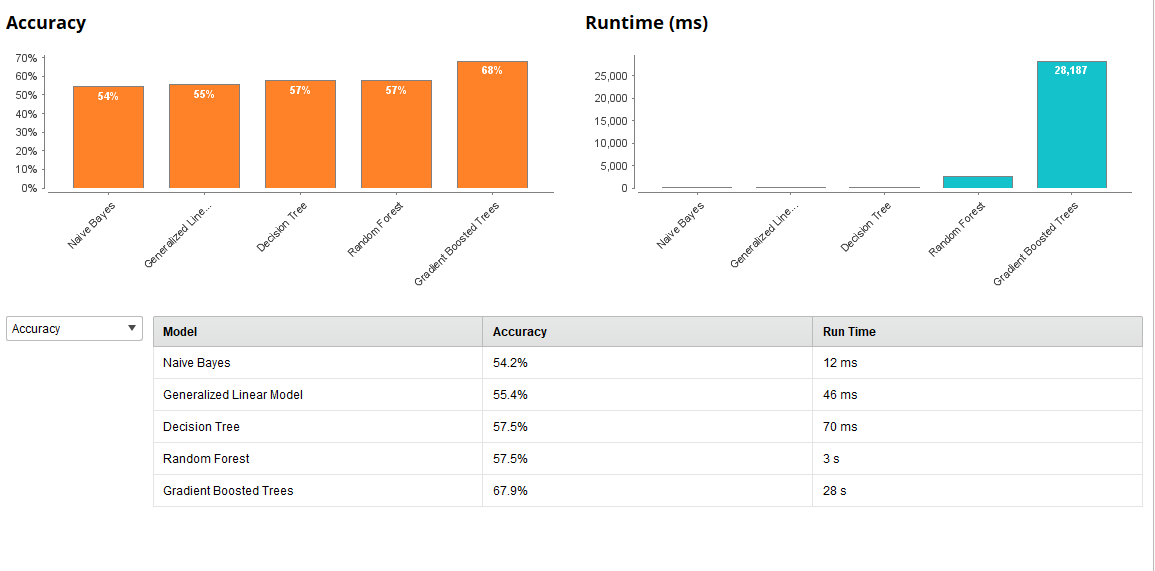
A screenshot of a cell phone

Description generated with very high confidence

*Figure 1.27. Summary of model outputs with deleted attribute – alcohol content*

Test:

Cross-validation:



*Figure 1.28. Summary of model outputs with deleted attribute – alcohol content*

Only Gradient Boosted Trees could give us 68% of accuracy in both training and test set. The conclusion is that even without knowing the alcohol content we still can predict alcohol level in wines when other attributes are known in 68%.

## **Target variable – Balanced wine**

The aim was to classify wines under label balanced and unbalanced. So that is binominal attribute.

Classification rule:

If total sulfur dioxide <=200, and sugar/ph ratio<=3, and level of alcohol = medium

The unknown here is free and total sulfur dioxide.

Is machine able to correctly classify wines with given attributes but unknown total sulfur dioxide?

Methodology used as in previous example: training 70%, test 30%.

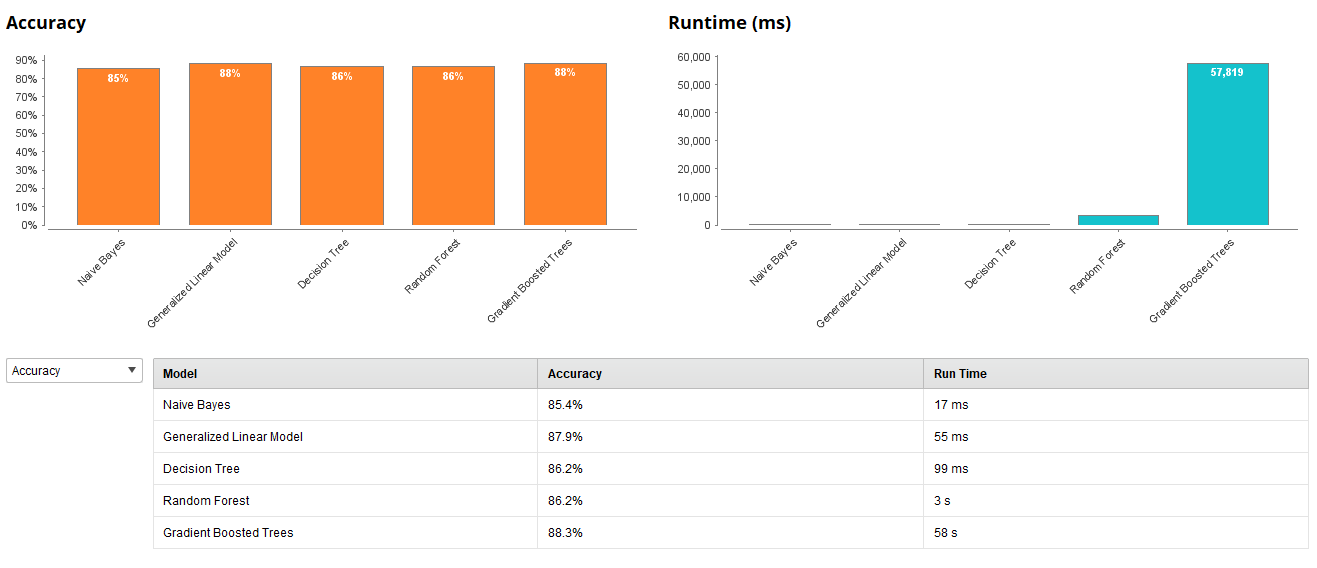
A screenshot of a cell phone

Description generated with very high confidenceTraining:

*Figure 1.29. Rapidminer – accuracy of training model – balanced wine*

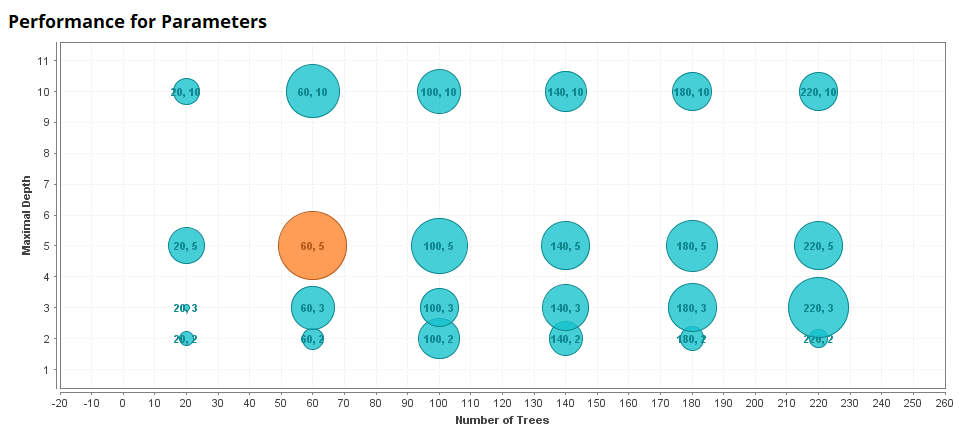
Test

Cross – validation:



*Figure 1.30. Rapidminer – Accuracy of test – balances wine*

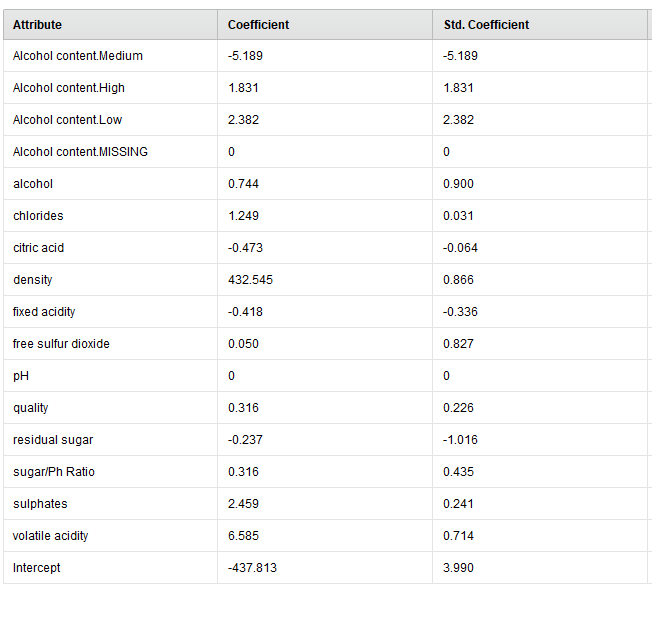
The highest accuracy of models we can see in Gradient Boosted Tree and Generalised Linear model in test set– 88% and 98% in training set. It is quite good classification model.



*Figure 1.31. Rapidminer – GBT – balanced wine – performance for parameters*

Model also shows us that the best result was achieved with 60 trees and 5 level of depth (Fig. 1.31).

Generalised linear model:

*Figure 1.32. Rapidminer - Generalised Linear model outcome – balanced wine*

Results of Generalised Linear models are much clearer for us to understand but it’s not mean it is better even if accuracy is high. We can see many drawbacks of this technique.

As it can see data only in linear position, it can classify wines even with dummy variables.

But overall information which we can take from this model that:

Even without known free and total sulfur dioxide we can classify wines with 88% accuracy. Very important attributes which affects balance of wine will be density, volatile acidity and sulphates.

# **Conclusions**

We can see from the work above that all the ingredients in wine are connected. All of them impact 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. Others in our dataset are purely scientific.

The Clustering analysis showed us that we can deal with 4 groups of clusters, showed us dependencies and correlations for example between 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 tested in isolation.

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

In two cases Gradient Boosted Trees (GBTs) were the best choice and gave us the best accuracy, and in 2 cases simple generalised linear model gave us the best accuracy.

It is not a surprising that GBT is better than Random forest. As in Random forest the trees are built randomly and in GBT the trees are build one by one when next one fixes problems of previous one. But unfortunately, GBT is time consuming for training and expensive technique.

About Generalised Linear models we would have more concerns.

It is built based on linear regression. It is very sensitive to outliers; the outliers can change the model a lot. Accuracy was good in our predictions but still if we will choose something outside of range that prediction could be not accurate. We are aslo not sure that we have linear correlations between these attributes. What if the relationships have different shape?

Giving these weaknesses we would be sceptical choosing this model without more tests.

**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. Carel M. (2011) ‘[Sulfur dioxide (SO2) in wine](https://winobrothers.com/2011/10/11/sulfur-dioxide-so2-in-wine/)’ [Online] Available from: <https://winobrothers.com/2011/10/11/sulfur-dioxide-so2-in-wine/> [Accessed 20 April 2018]
3. Han J., Kamber M. and Pei, J. (2011) Data Mining: Concepts and Techniques. 3rd ed. San Francisco: Morgan Kaufmann Publishers.
4. Organicvineyardalliance (2017) ‘Organic Wine Definitions – Behind The Label [Online] Available from: http://organicvineyardalliance.com/organic-wine-definitions-behind-the-label/ [Accessed 26 April 2018]
5. WARDSCI (2018) ‘Chemistry of Wine’ [Online] Available from: <https://www.wardsci.com/www.wardsci.com/images/Chemistry_of_Wine.pdf> [Accessed 20 April 2018]
6. Water House (2012) ‘What’s in wine’ [Online] Available from: <http://waterhouse.ucdavis.edu/whats-in-wine> [Accessed 20 April 2018]
7. WineFolly (2015) ‘Sugar in Wine Chart’ [Online] Available from: <http://winefolly.com/review/sugar-in-wine-chart/> [Accessed 20 April 2018]