Principles of Data Mining

Analysis of White Wine Quality Data with SAS Enterprise Miner and Azure ML Studio

London South Bank University, December 7, 2018
School of Engineering, Data Science MSc

Module: Principles of Data Mining
Lecturer: Dr David D Chen
Author: Patric Oliver Weber

Version: 1.1

Contents

List of Figures

1 Introduction

"Wine is the only artwork you can drink". A drink which exists for more than at least 2000 years, appears still mysterious. The sommelier, a specialist in his territory, has to learn hard to understand this topic which is still not yet fully understood. The sense of the nose and tongue is a mystery of sensation and appears for every wine-drinker entirely different - especially that makes this field so attractive. Having mentioned earlier the quote from Luis Pasteur; art can herby used for its unique craftsmanship and perfection to craft the right fermented grape juice.

With the given knowledge and tools of trade received from lectures by Mr Dr Chen about predictive-and-descriptive-modelling, I am going to gain more in-depth understanding of wine by conducting the exploratory data analysis of the given Dataset with the physico-chemical and quality of the wine and will apply different data mining techniques and gain valuable insights into data processing.

The research of the dataset will be guided through the Cross-industry standard process for data mining, known as CRISP-DM. Most of the data mining process will be conducted in SAS Enterprise Miner v. 14.1 and Azure Machine Learning Studio which recently got introduced by the company I work. Some other Algorithm may also be written in a high-level computing language such as R. In my final section, I will summarise the findings of the wine data set and suggest further course of action.

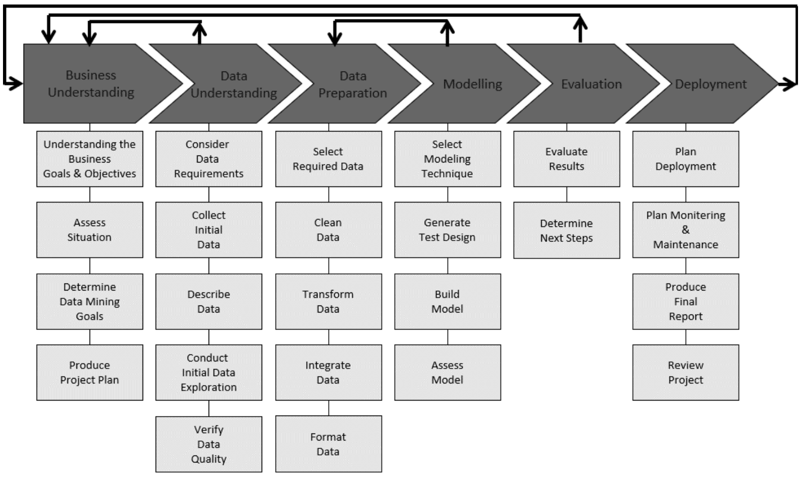


Figure 1: Visual guide to CRISP-DM methodology.(source: crisp-dm.org)

2 Business Understanding

Wine, an alcoholic beverage, also known as a luxury product in the early days, is being consumed by a much wider range of consumers all over the world nowadays.
Portugal, with a market share of 2.5% in 2017, counts as one of the top-nine wine exporting countries [Workman, 2018]. The region Vinho Verde plays a significant part in representing this wine country with its light and crisp white wines. The exports from Vinho Verde are from the northwest region and rapidly increased by 36% from 1997 to 2007. Still, Vinho Verde wine exports continue on an upward trend. In terms of value, the United Kingdom is ranked as the seventh most important market for Vinho Verde [Geoffey, 2016].

To sustain this continuous growth, the wine industry is investing in cutting-edge technology for both the production and sales processes [Kernmode, 2018]. Having the quality assessment as its wine certification plays a significant part within this context.

2.1 Vinho Verde

Vinho Verde[1] is a Portuguese wine which comes from the historic Mihno province, in the far north of the country. In recent years, the region has also accommodated adjacent areas to the south. Vinho Verde does not refer to a grape or blend; it is a DOC, a protected title of its origin for wine, cheeses, butter and many other agricultural products from Portugal since 1984. Vinho Verde is the most prominent DOC in Portugal [Robinson, 2006] The climate is cool, rainy and green in the north-west. The grapes are grown in fertile, granite soils alongside rivers that flow from the mountains into the ocean.
The name also has the literal meaning of "green wine", or also translates as *"young wine"* which means that the wine is released at an early stage from 3 to 6 months after harvesting. Vinho Verde wines may be red, white or rosé and they are usually consumed right after bottling. This is also a wine which can be sparkling, late harvest or even a Brandy. Vinho Verde is one of the best value wines on the market; this is because wine is only stored for a short time [Clouet-Foraison, 2015].
In wine production, historically, the fizz (building a specific gas in the bottle) of a wine steamed from malolactic fermentation was typically considered a wine fault. However, the producers of Vinho Verde found out that this slightly fizzy nature of the wine appealed to consumers.
The Region covers many small growers, in total, around 19,000 in 2004; the wine-producers farm over 51,000 acres of vineyards. 600 wine bottles produce about 85 million litres of Vinho Verde each year, of which 86% is white wine [Clouet-Foraison, 2015]. The locals

[1] `http://winesofvinhoverde.com/`

drink twice as much red Vinho Verde as white and hardly any of the slightly fizzy, very dry, very light red wines which are being exported globally("JancisRobinson.com," noa [a]).

Quinta do Ameal, Quinta de Azevedo, Covelha, Case de Cello, Quinta de Soalheiro are the some of the favourite producers in this region.

2.2 Style

Vinho Verde is a light, refreshing wine with lots of floral and green fruits notes. The wine often contains slight carbonation but does not qualify as semi-sparkling wine (< 1 Bar CO2). The white wines are straw- or bright yellow-coloured and start from 8.5% alcohol up to 11%. The white wine is mostly made from local grape varieties such as Trajadura, Loureiro, Avesso Arinto and Azal. The grape Vinho Alvarinho is the most popular grape which only grows in small selected areas of Melgaco and usually contains more alcohol. Because of the wine's bracing acidity and oceanic minerality, it should be the ultimate partner to raw fish. When food pairing, it goes exceptionally well with shellfish, vegetarian food as well as appetisers and snacks. The character of the body is very light("Portuguese Vinho Verde White | Regional Wine Style," noa [b]).

The style of the red wines is deep red and tannic. The red is mostly made from Borraçal, Vinhão and Amaral grapes. The rosé wines are very fresh and fruity made from Padeiro and Espadeiro grapes.

2.3 Wine Certification and Quality Assessment

As Vinho Verde is a DOC and contains 19,000 small growers, it is essential to set a certain quality standard for the wine produced in the area. Hence, as mentioned above, to sustain growth, the industry is investing in new technologies. Despite their concerns over changes to the status quo, many independent producers are now being more experimental and concentrating their primary goal on the quality of the wine, rather than on quantity. Questioning the status quo has become popular in recent years, take, for example, the Craft Beer revolution, which has been an enormous enabler into many new markets and has attracted new customers. In contrast to the Vinho Verde this is leading to drier, richer and less fizzy styles with much greater ripeness of fruit, resulting in improved export sales, particularly in the US [Clouet-Foraison, 2015].

The Certification ensures that no illegal adulteration of wines will take place. Certification is a safeguard for human wellbeing and assures quality for the global wine market.

2.3.1 Wine Certification

Two aspects are generally measured in the certification of wine; namely, the sensual and physicochemical [Teranishi et al., 1999, pg. 409-422]. The Physicochemical laboratory test is used to characterise wine and determine density, alcohol or pH values. The Sensory tests rely mostly on human experts, a wine sommelier - a subject-matter expert in his field. It should be stressed that taste is still the least understood of the human senses Smith and Margolskee [2006]. Therefore, wine classification is a complicated issue. Moreover, the relationship between chemical and sensory analysis are complex and yet not fully understood [Lengin et al., 2003, pg. 33-34].

2.4 Technology

Groundbreaking technologies in IT have made possible to collect, store and process a massive amount of big data.

Most of this data holds highly valuable information on things such as trends, patterns and correlation. The extracted data, from a data warehouse, can improve decision-making and optimise the chances of success in many business areas. Data mining methods aim to extract high-level knowledge from raw data. There are many algorithms which can fit the needs of any business requirements.

2.5 Business Goals

Data mining models can benefit both certification entities and the producer. They help to improve the oenologist's wine evaluation, improving the quality and accelerating decisions. Furthermore, such physicochemical tests can improve the production process and help to understand the wine quality better. Finally, not to forget target marketing. Applying these similar techniques to the customer's preferences will find new segments such as niche markets and help create a higher turnout.

A breakthrough in the wine-making industry would ultimately be to find solid proof of a correlation between the physicochemical and sensory aspects.

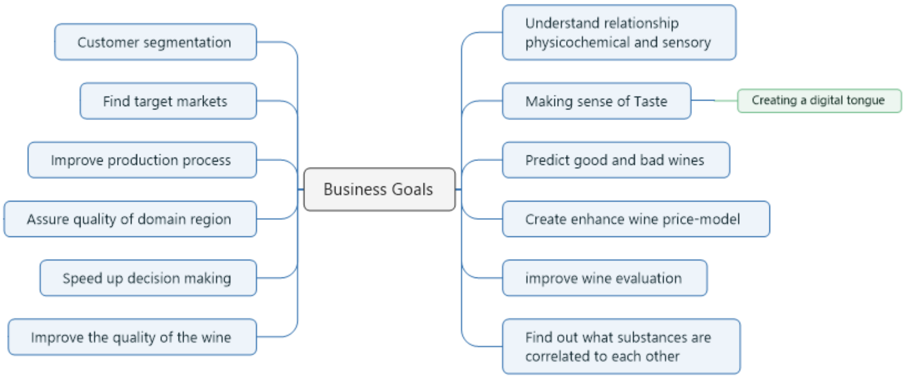

Customer segmentation
Find target markets
Improve production process
Assure quality of domain region
Speed up decision making
Improve the quality of the wine
Business Goals
Understand relationship physicochemical and sensory
Making sense of Taste
Creating a digital tongue
Predict good and bad wines
Create enhance wine price-model
improve wine evaluation
Find out what substances are correlated to each other

Figure 2: Brainstorm of business goals for a wine region

3 Data Understanding

The data of the wine quality and physicochemical data was collected between May 2004 up to 2007 by using only protected designation of origin samples, also known as DOC. These samples were tested at CVRVV, the official wine certification. The CVRVV stands for an inter-professional organisation which aims to improve the quality and marketing of the region Vinho Verde. The dataset is a wine quality dataset which is publicly available for research purposes form the uci.edu[2] homepage. iLap a laboratory computer system has collected this data. iLap automatically handles the process of the wine sample testing. Each entry is a given test. Furthermore, the database is already in a preprocessed stage and there are only common physicochemical test selected.

The red wine dataset contains 1599 observations with eleven features. The white quality dataset covers 4898 instances with the same eleven features. In this project, the white dataset will only be analyzed. The Input tests contain for example pH values and the output is qualitative median sensory data which was tested by blind tastes where experts made at least three evaluations. Each sommelier graded the quality of the wine from zero (very bad) up top ten (excellent) [Cortez et al., 2009].

3.1 Aim

The principal aim of this project is to predict or foresee the qualitative rating of a wine sample by using a range of laboratory properties, such as alcohol, acidity and density. Furthermore, finding hidden patterns in the dataset and understand the correlation between the physicochemical features better.

3.2 Attributes

According to Cortez et al., (2009), “The features include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. pH describes how acidic or basic wine is on a scale from 0 (very acidic) to 14 (very basic). Chloride is the amount of salt in the wine. Alcohol is the per cent alcohol content of the wine”.

```
Input variables (based on physicochemical tests):
   1 - fixed acidity (tartaric acid - g / dm^3)
   2 - volatile acidity (acetic acid - g / dm^3)
```

[2] https://archive.ics.uci.edu/ml/datasets/Wine+Quality

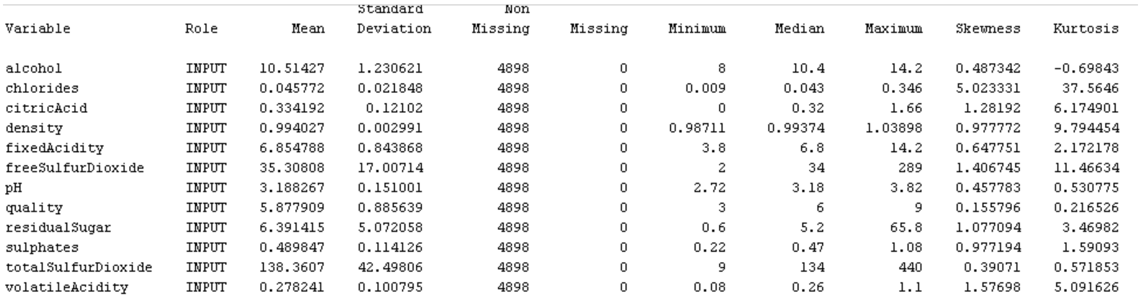
```
    3 — citric acid (g / dm^3)
    4 — residual sugar (g / dm^3)
    5 — chlorides (sodium chloride — g / dm^3
    6 — free sulfur dioxide (mg / dm^3)
    7 — total sulfur dioxide (mg / dm^3)
    8 — density (g / cm^3)
    9 — pH
    10 — sulphates (potassium sulphate — g / dm3)
    11 — alcohol (% by volume)
    Output variable (based on sensory data):
    12 — quality (score between 0 and 10)
```

3.3 Summary

With the gathered information of the dataset, we can conclude that this dataset can be observed as a regression or classification task. However, other tasks can also be applied for a better understanding of the data. The classes are not balanced; this means that there are much more normal than excellent or bad wines. Secondly, outlier detection algorithms should be used to spot poor or excellent wines. Finally, there is no guarantee that all input-variables are relevant. Thus, it might be interesting to test feature selection methods.

3.4 Univariate analysis

At first glance we can see that critic Acid contains zero values, this must be observed to check if these are outliers. Other important indicators are the skewness, standard deviation and kurtosis.



			Standard	Non						
Variable	Role	Mean	Deviation	Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
alcohol	INPUT	10.51427	1.230621	4898	0	8	10.4	14.2	0.487342	-0.69843
chlorides	INPUT	0.045772	0.021848	4898	0	0.009	0.043	0.346	5.023331	37.5646
citricAcid	INPUT	0.334192	0.12102	4898	0	0	0.32	1.66	1.28192	6.174901
density	INPUT	0.994027	0.002991	4898	0	0.98711	0.99374	1.03898	0.977772	9.794454
fixedAcidity	INPUT	6.854788	0.843868	4898	0	3.8	6.8	14.2	0.647751	2.172178
freeSulfurDioxide	INPUT	35.30808	17.00714	4898	0	2	34	289	1.406745	11.46634
pH	INPUT	3.188267	0.151001	4898	0	2.72	3.18	3.82	0.457783	0.530775
quality	INPUT	5.877909	0.885639	4898	0	3	6	9	0.155796	0.216526
residualSugar	INPUT	6.391415	5.072058	4898	0	0.6	5.2	65.8	1.077094	3.46982
sulphates	INPUT	0.489847	0.114126	4898	0	0.22	0.47	1.08	0.977194	1.59093
totalSulfurDioxide	INPUT	138.3607	42.49806	4898	0	9	134	440	0.39071	0.571853
volatileAcidity	INPUT	0.278241	0.100795	4898	0	0.08	0.26	1.1	1.57698	5.091626

Figure 3: The physicochemical and sensory data statistics of white wine

3.4.1 Acidity

The first attributes are Acids; they are playing a significant part regarding the structure of a good wine. Most people will find a high amount of acidity too sour and too tart. In comparison, a low-amount will taste flat and wines are more vulnerable to infection can easily get spoiled by microorganisms. In warmer climates, the grapes contain lover acidity than grapes in mild climates. Acidity shares a correlation with sugar and pH [Pandell, 1999].

Fixed acidity

Fixed acidity is sometimes also referred to titratable acidity. It is a measurement of the total concentration(TA). Most of the acids in the wine are fixed or non-volatile, these acids do not evaporate from itself.

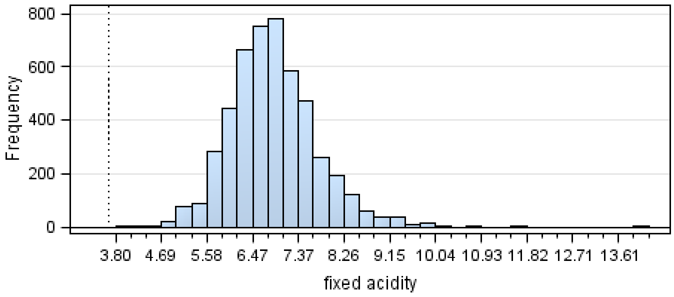


Figure 4: Fixed acidity histogram with x = 35 bins

```
Min. 1st Qu.  Median   Mean    3rd Qu.  Max.
3.800   6.300   6.800   6.855   7.300  14.200
```

According to Figure 4 we can see that fixed acidity looks normal distributed and the skewness is very minor. There are on the right-hand side some outliers.

Volatile acidity

Mentioned above that the tartic acid is non-volatile, this means that it does not evaporate when the wine is heated. Another important indicator is the volatile acidity which represents the acetic acid. Acetic acid will boil off when heated. The higher the VA, the more it can lead to unpleasant vinegar taste [Pandell, 1999].

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

The volatile acidity looks normal distributed. It is positively skewed. There are many outliers detected on the right side.

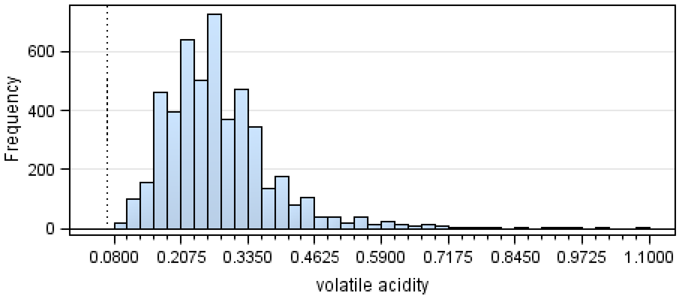


Figure 5: Volatile acidity histogram with x = 40 bins

Citric acid

Critic acid can be found in small quantities, it adds freshness and flavour to wines.

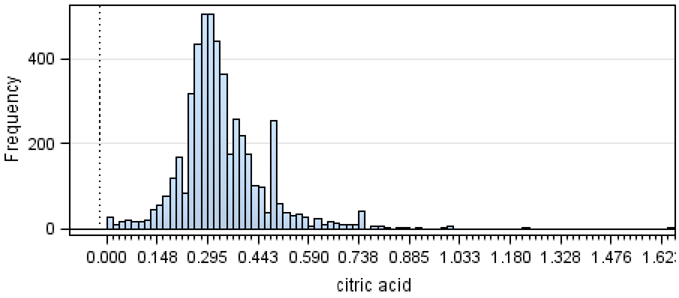


Figure 6: Citric acid histogram with x = 90 bins

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

Figure 6 shoes that the data is normal distributed. It is positively skewed. There is an extreme outlier at 1.66 and there are also zero-values. Another interesting observation is that there are relatively many wines at 0.49 and 0.74.

3.4.2 Residual sugar

The production of sugar is completely the opposite of the acid production mentioned above. The warmer the climate, the higher the content of sugar in the grapes. Residual sugar shows the amount of sugar which remain after the fermentation has stopped. Finding wines with less than 1 gram a liter is rare. Less priced wine (<10$) often contains between 2-15g/L where as more qualitative wine are featuring less residual sugar, because grapes are higher in quality and wines do not need sweetness for its taste of fruitiness (“Sugar in Wine Chart (Calories and Carbs),”noa [2015]).

- Bone Dry <1g/L
- Dry 1 – 10g/L
- Off-Dry 10-35g/L
- Sweet 35-120g/L
- Very Sweet 120-220g/L

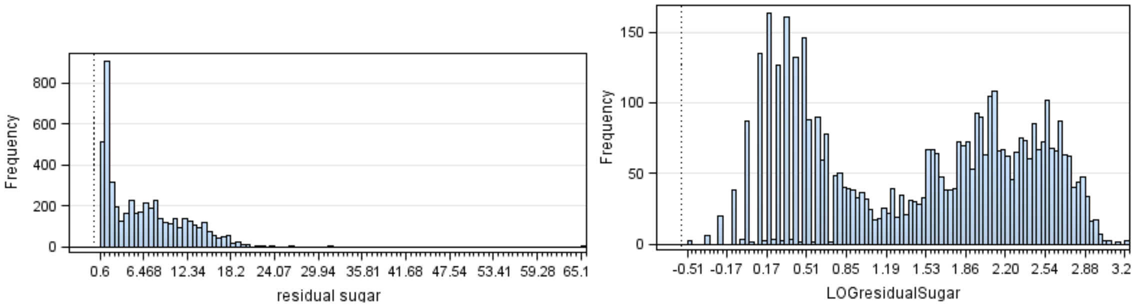


Figure 7: Residual sugar histogram before and after Log transformation

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.600   1.700   5.200   6.391   9.900  65.800
```

Residual sugar is highly positively skewed. Hence, It makes sense to apply a Log 10 Transformation.(“Log Transformations,” n.d.) Most data points are below 3 and half the other part of the values are below 10. There is a huge outlier such as 65.8. After having preprocessed residual sugar, the data loos much more bi-modal distributed.

3.4.3 Density

Density is depending on water with the percent of alcohol and sugar content.

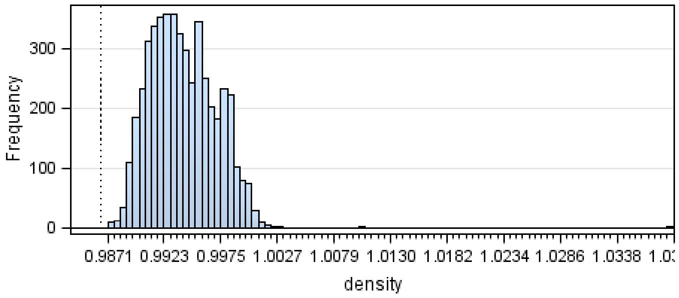


Figure 8: Density histogram with x = 90 bins

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

Most of the data are in a tiny range. For a better understanding and modelling it makes sense to limit the omitting outliers. Even though standard deviation, median and mean is very small there is another peak at 0.48. The dataset does not look like skewed, excluding outliers.

3.4.4 pH

pH is another measurement of the acidity. It shows how sour or how basic the wine is. The scale lasts from 0 (very acidic) to 14 (very basic). pH shares the same scale as the acidic and both are logarithmic. Generally speaking, all wine lie more on the acidic side and most are in the range from 2.5 up to 4.5pH where pH 7 is neutral. Wine with a pH of 3 is 10 times more acidic than a wine with pH 4. The lower the pH, the higher is the acidity [Waterhouse, 2015].

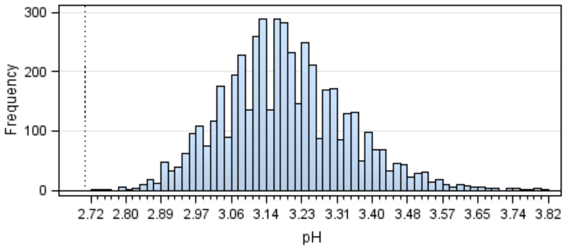


Figure 9: Residual sugar Histogram with x = 100 bins

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.720   3.090   3.180   3.188   3.280   3.820
```

The pH is normally distributed, there are no extreme outliers and there is no skewness.

3.4.5 Alcohol

Alcohol is the percent alcohol content of a wine. It has special characteristics together with wine. Wines with higher alcohol tastes bolder and oily whereas wine with lower alcohol smells lighter-bodied. Most wine are considered in the range between 11-13% [Puckette, 2012].

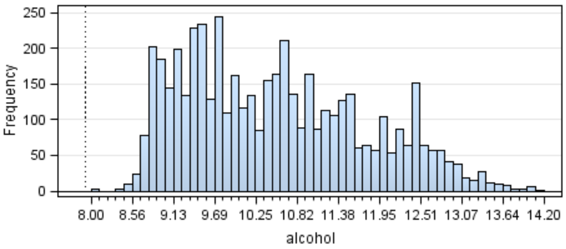


Figure 10: Alcohol histogram with x = 55 bins

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
8.00    9.50  10.40   10.51   11.40   14.20
```

The dataset of the alcohol is positively skewed. There are no outliers. Alcohol may play a curcial part of the analysis as it correlates with many other inputs.

3.4.6 Chlorides

Chlorides are representing the amount of salt in the wine. This is obviously not a common indicator but since recently caught a lot of interest. In fact many wine producing countries have varying legal maximums of sodium chloride. Salinity is a concern in many dry location when frequently irrigation increases the salinity of the soil [szymanskiea, 2015].

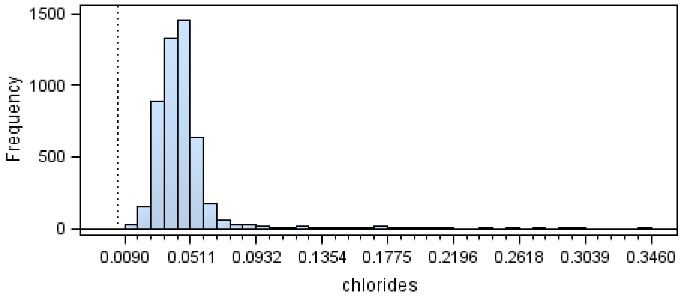


Figure 11: Chlorides histogram with x = 40 bins

```
Min.     1st Qu.   Median.  Mean.     3rd Qu.  Max.
0.009    0.036     0.043   0.04577   0.05     0.346
```

The chlorides are relatively left skewed with a one big outlier at 0.346. Most of the data are below 0.6

3.4.7 Sulfides

Sulfites help to preserve wine and slow the chemical reaction of getting bad. The lower the acidity, the more sulfites are needed. Hence, wines with pH of 3.6 are less stable and need more extra sulfites for storage. Secondly, wine with more sugar contain a higher level of sulfide to stop secondary fermentation of remaining sugar in the wine [Puckette, 2012].

Free sulfur dioxide

This sulfur dioxide is the free form of SO2. It is a liquefied gas and bisulfite ion. Free sulfur dioxide will stop the growth of microbial bacteria and the oxidation of wine [Cortez et al., 2009].

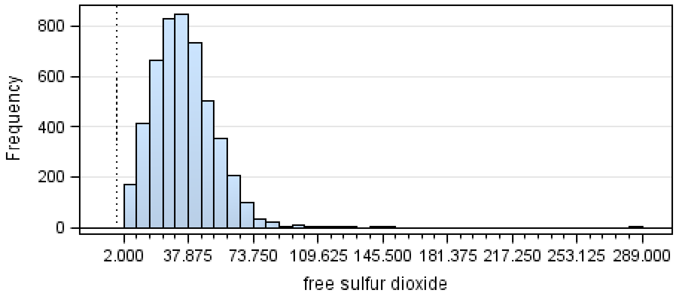


Figure 12: Free sulfur dioxide histogram with x = 40 bins

```
Min.     1st Qu.  Median.  Mean.    3rd Qu.  Max.
2.0      23.0     34.0     35.31    46.0     289.0
```

The sulfur dioxide is left skewed and the range of the data is 287. Most of the data is at the mean of 35.31.

Total sulfur dioxide

It is the total of free and bound forms of SO2. SO2 is generally unnoticeable in low concentration. However, one reached the threshold of 50ppm in free SO2; it becomes noticeable in the nose and can smell similar to citrus and cooked eggs.

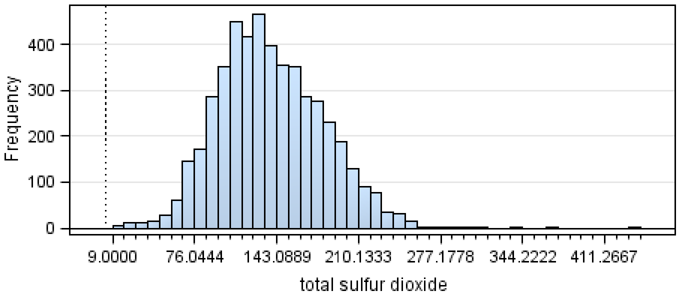


Figure 13: Total sulfur dioxide histogram with x = 44 bins

```
Min.     1st Qu.    Median.  Mean.    3rd Qu.   Max.
9.0      108.00      134.0   138.4    167.0    440.0
```

The data has a high range of 431 but is normal distributed and slightly left-skewed. Outliers on the right side which are twice the distance from the last normal data points.

Sulphates

It's the wine additive that contributes to the level of S02 and helps to stabilise the wine by acting antimicrobial and antioxidant [Puckette, 2012].

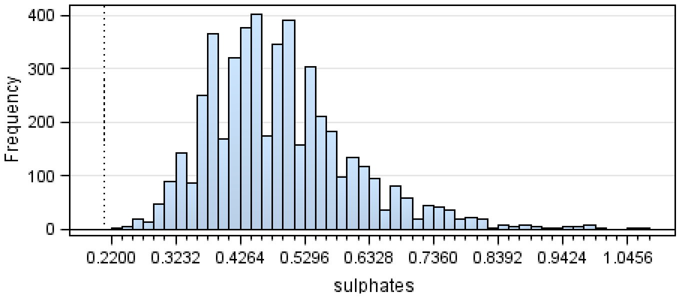


Figure 14: Residual sugar Histogram with x = 100 bins

```
Min.     1st Qu.    Median.  Mean.    3rd Qu.   Max.
0.220     0.410     0.470   0.4898    0.550    1.080
```

Left skewed data in a normal range. Minor outliers on the right side. Special peaks with lows, this could indicate to a special pattern.

3.4.8 Quality

The output Variable of the dataset of sensory data.

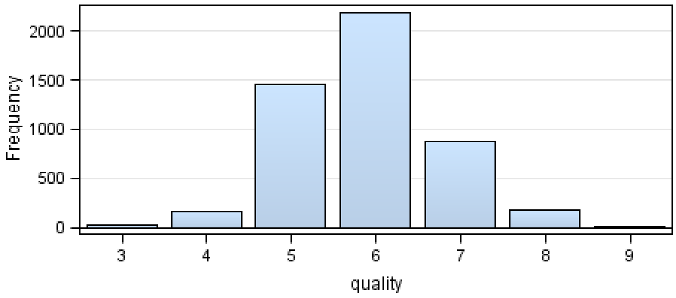


Figure 15: Quality histogram with x = 7 bins

```
Min.     1st Qu.    Median.  Mean.    3rd Qu.   Max.
0.220     0.410     0.470   0.4898    0.550    1.080
```

The Quality dataset is normally distributed with a median of 6. There are no outstanding wines, only a few wines with a score of 3 or 9. This makes clustering more difficult as there are only a few data to train and test with.

3.5 Relationship between Attributes

By summarizing all above mentioned attributes we can draw conclusion that based on research there is a relationship between alcohol, density and sugar as well as acidity and pH and sulphates. The only Attribute which has not shown much association to the others is the attribute chloride. Sulfide might bridge the gap between the both correlated groups.

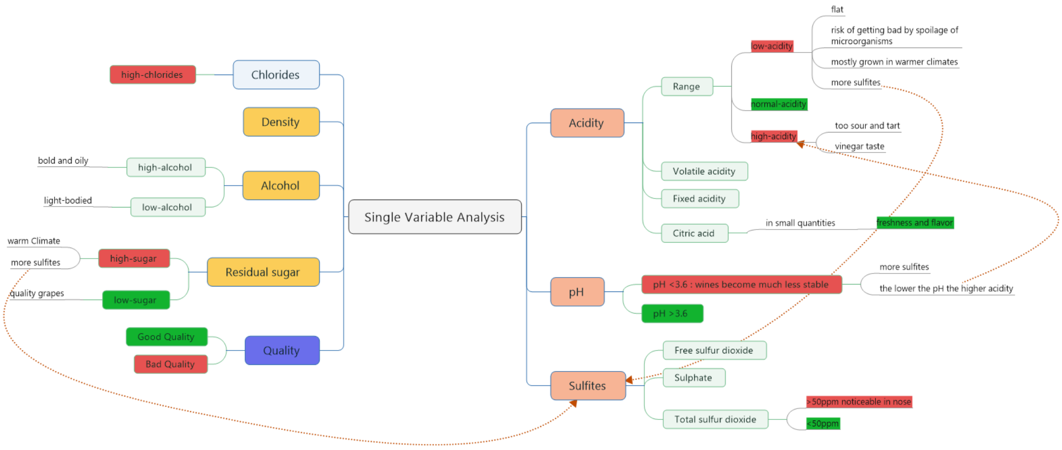


Figure 16: Research based relationship between wine attributes

4 Data Preperation

To look at the data more precisely and understand correlation more clear, further data preprocessing is needed to remove the remaining outliers. Outliers should be addressed as they disturb the processing in modelling and lead to faulty results. Since this collection contains multivariate data, we consider only points which are not outside the limits created by the box plots. Hence, the following standard deviation rule from the mean will be applied by the Filter node in SAS. The formula says: A predictor value is an outlier if it the value is higher than Q3 + 1.5IQR. The thinking behind this rule is that the extreme outliers, mostly in the histograms which are positively skewed, will break down its positive skewness and normalise the dataset. After the normalisation process has run through, the data size shrank from 4898 to 4492.
Figure 17 shows the diagram where the Log Transformation with the node *"Transform Variables"* formula =*'LOG(residualSugar)'* and the outlier removal with the *'Filter'* node was carried out.

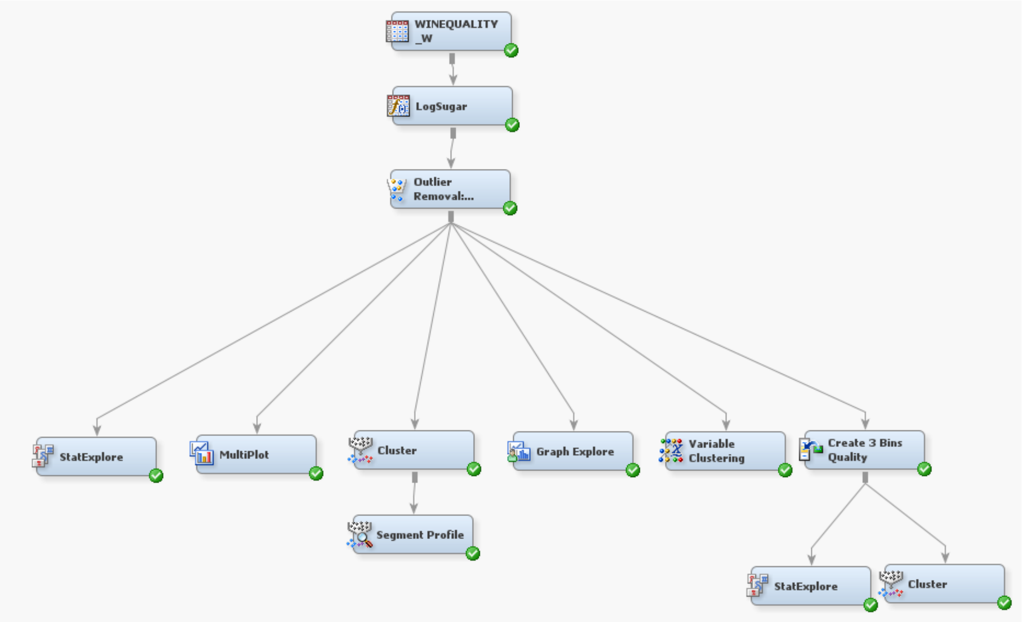
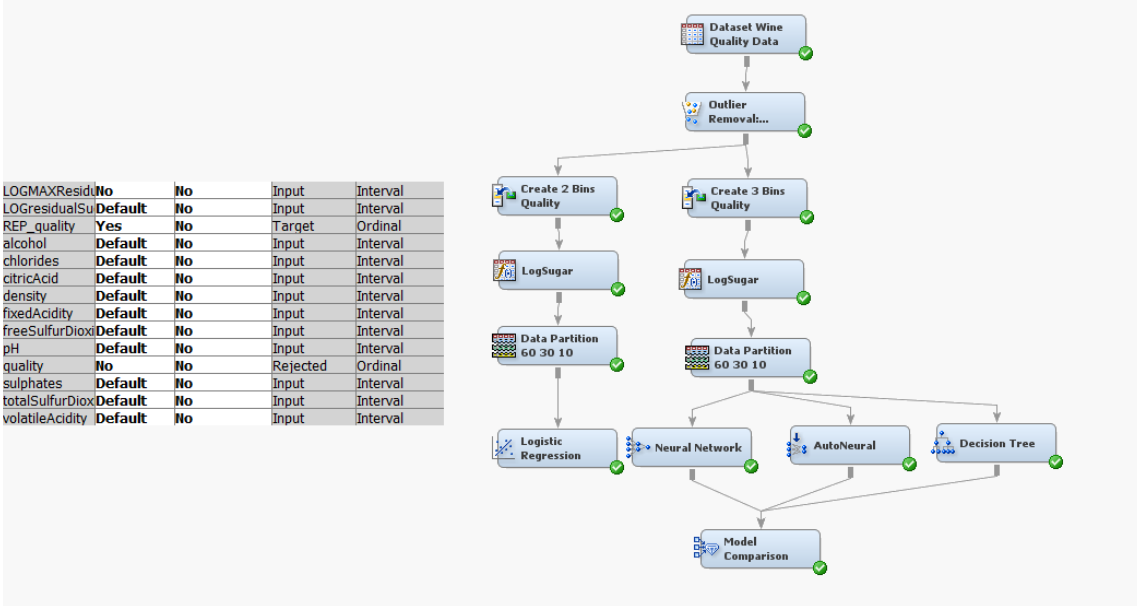


Figure 17: SAS diagram Data Understanding and Data Preparation

5 Modelling

In this section, attributes will be combined and analysed in more detail. An algorithm such as person correlation will observe relationships among the attributes who correlate with the target attribute. Not only StatExplore will show us how the data shows a relationship with others, also Scatterplots and correlation matrixes will help to understand the data better.

Further, more in depth techniques will be applied to the preprocessed data to predict future values on the dataset. Techniques such as Cluster Analysis, Decision Trees, Regression and Neuronal Networks are going to be applied and compared. By applying Regression and Classification tasks the *'Data Partition'* node was selected to sample data into *60% Training, 30% Validation and 10% Test.* In Figure 25 the input variables are delare for the modelling task.



LOGMAXResidu	**No**	**No**	Input	Interval
LOGresidualSu	**Default**	**No**	Input	Interval
REP_quality	**Yes**	**No**	Target	Ordinal
alcohol	**Default**	**No**	Input	Interval
chlorides	**Default**	**No**	Input	Interval
citricAcid	**Default**	**No**	Input	Interval
density	**Default**	**No**	Input	Interval
fixedAcidity	**Default**	**No**	Input	Interval
freeSulfurDioxi	**Default**	**No**	Input	Interval
pH	**Default**	**No**	Input	Interval
quality	**No**	**No**	Rejected	Ordinal
sulphates	**Default**	**No**	Input	Interval
totalSulfurDiox	**Default**	**No**	Input	Interval
volatileAcidity	**Default**	**No**	Input	Interval

Figure 18: Input variables and modelling diagram

5.1 Multivariate Analysis

5.1.1 StatExplore

Stat Explore aims to find an association with three different algorithms. Pearson Chi-Squared is a statistical method that evaluates categorical data, in this case, the target attribute: Quality. The method assumes that inputs are random of independent variables. The result will indicate how far results are from the expected random result (“SAS Enterprise Miner Reference Help v14.1, noa [c]).
Secondly, there is the Cramer’s V algorithm; this is a post-test to determine the strengths of association from the chi-square. Chi-square shows where a strong relationship between variables is. However, it does not say how noteworthy this relationship is. As we can conclude, this is a post-test which provides us with additional information. Carmer’s V varies between zero and one, where one indicates there is a strong association [Starker, 2010].

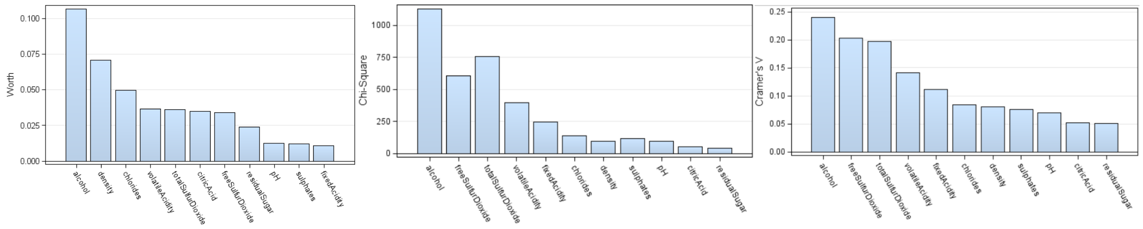


Figure 19: Gini Index, Chi-Square and Craners’V on quality label

5.1.2 Interpretation of results

The observation of Stat Explore in SAS shows that there is a clear correlation with the alcohol variable. This can be observed by looking at any plots. Even though, most of the associations are small alcohol and density must share a strong correlation with the target value.

5.1.3 Correlation Matrix

The correlation matrix shows how strong the independent variables are correlating naturally with each other. Furthermore, the attribute quality is also included. Among all variables, alcohol shows the most substantial relationship with quality. Other physicochemical inputs are mainly negatively correlated. Residual Sugar and density, followed by density and alcohol are the strongest among the independent variables. Density and alcohol are directly related to quality, although the association is weak, besides Residual sugar shares a strong relation to density, the amount of remaining sugar might have a significant influence on Quality.

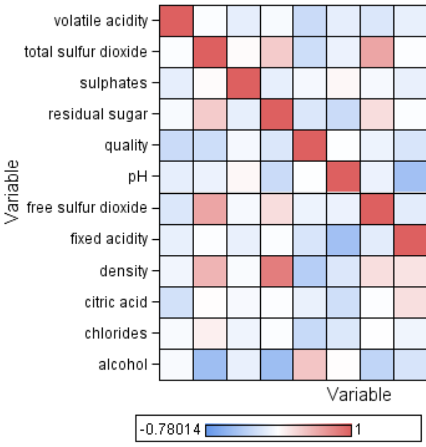


Figure 20: Gini Index, Chi-Square and Craners'V on quality label

5.1.4 Sugar, Density & Alcohol

In the below 3D-scatterplot Figure 21, the quality variable is converted into 3 bins, where one represents the range 3-5, two represents range 6 and three seven until nine. As seen below, less alcohol indicates better quality (range 1-2), but in the range 3 higher alcohol refers to higher quality. Density also has a great effect on wine quality; this can be observed in more higher quality, where the density is significantly lower. As Density and Sugar are strongly correlated, less sugar refers to better quality.

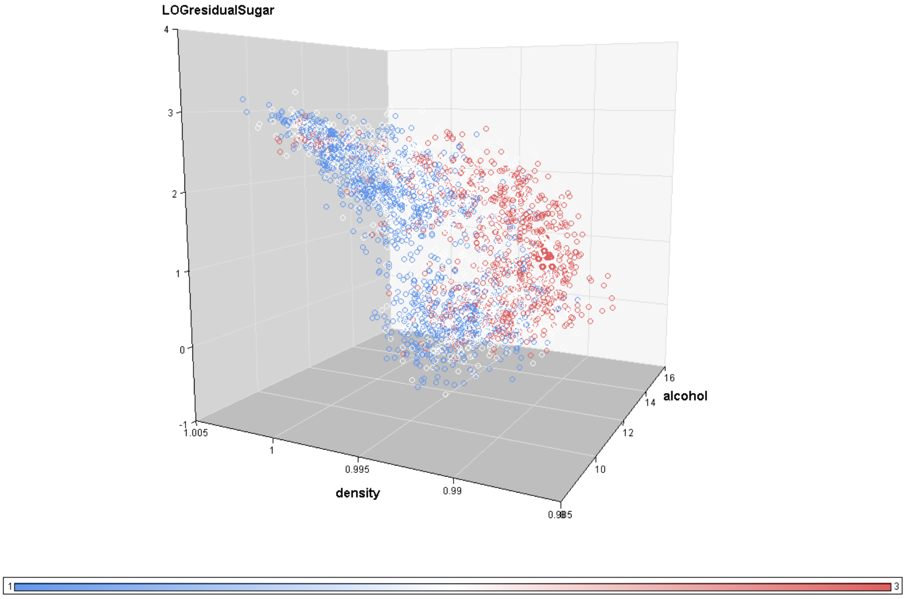


Figure 21: Scatterplot of sugar, density and alcohol

5.1.5 Sulfur, Chlorides & Density

Another interesting graph is the relationship between stabiliser, know as sulfides and the quality. Figure 22 focus on the attribute chlorides which shows how salty the water is. Specifically, this has shown no fundamental relationships among other attributes in research. However, chlorides were selected, because of the high indication in the Cramer's V algorithm. Together with density, it shows that fewer chlorides and fewer sulfides are referred to a higher quality of wine.

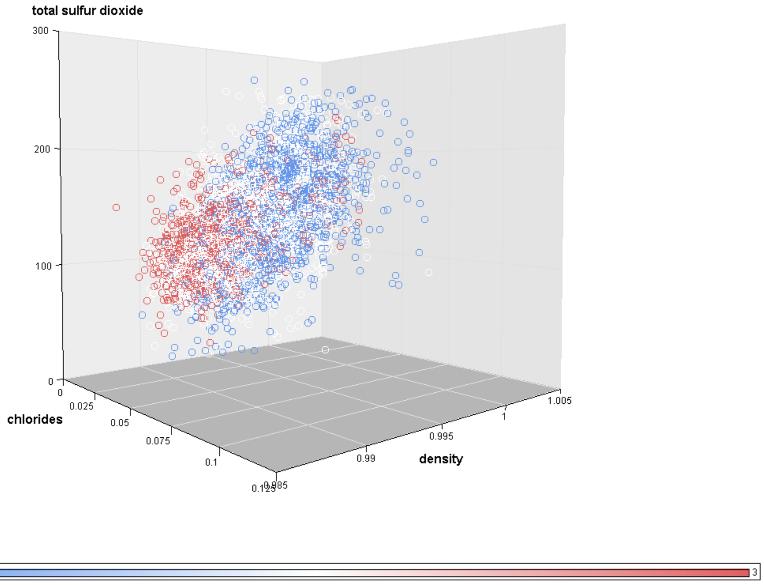


Figure 22: Scatterplot of sulfur, chlorides and density

5.2 Cluster Analysis

To prove the hypothesis mentioned above a cluster analysis model will be applied by only feeding independent variables. According to Figure 23, the cluster created three clusters where cluster one shows wine with a high amount of sugar, segment two wines with lower density and less sugar and segment three wines with an high amount of sugar. For a second proof another 3D-scatterplot will show the segments among with quality, sugar and alcohol. By slicing the data into the quality, the segments cannot correctly represent the quality. However, segment two(white) is the most promising with the indication of a high alcohol volume.

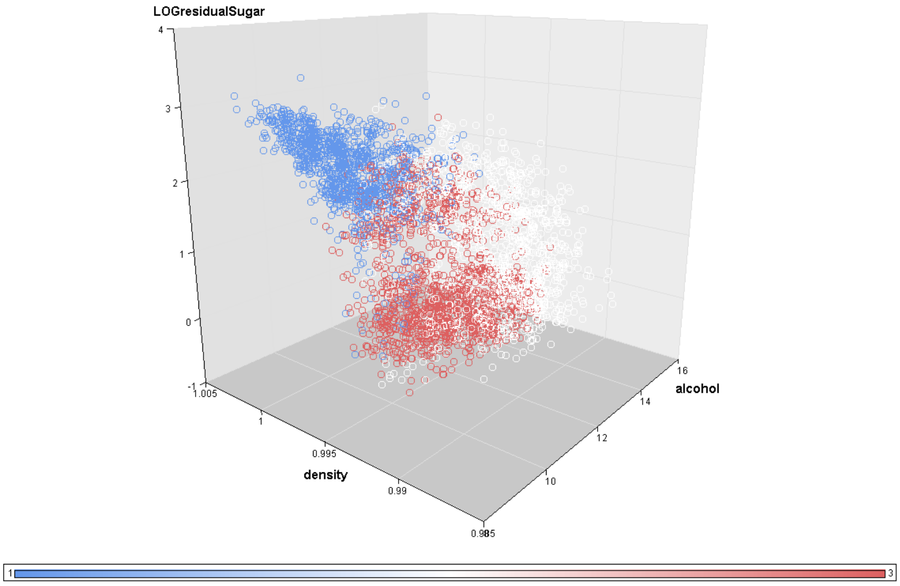


Figure 23: Cluster scatterplot with sugar,density and alcohol

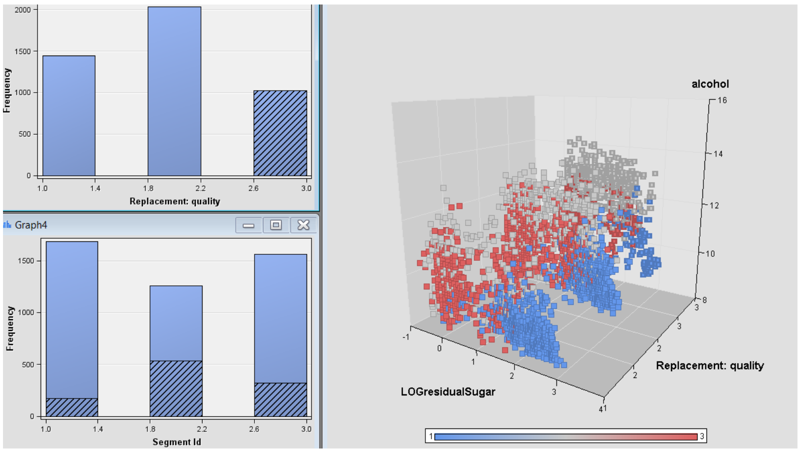


Figure 24: Cluster scatterplot segmented into quality bins

5.3 Logistic Regression

With the multivariate exploration analysis in the previous section, it seems at that there is no simple linear relationship among the label and the independent variables. However, if this hypothesis is correct, it should show that the most techniques does not perform very well. The method *Forward* was selected, where it selects the most significant variables. It begins with the most significant and starts adding. Looking at the results, they are not promising. Furthermore, binning wine data in two buckets makes not much sense.

```
Event Classification Table
 Data Role=TRAIN Target=REP_quality Target Label=Replacement: quality
   False       True        False       True
Negative    Negative    Positive    Positive
     0           0          865        1834
```

5.4 Decision Tree

An autonomous classification decision tree helps to find out if the dataset can generate accurate predictive results. Having applied the Decision Tree node with default settings the only flag which has been changed are *Perform Class Validation = 'Yes'* and *Assessment Measure = 'Decision'*. After the experiment has run through it has generated 8 Leafs and five layers. The results still show a considerable misclassification error of 45% in the test set. However, one niche leaf indicating high-quality wine where alcohol is low and sugar reasonably higher can predict 82% correct. This is an exciting founding, which shows that not every whine with a high amount of alcohol its likely to be high in quality.

```
Event Classification Table
 Data Role=TRAIN Target=REP_quality Target Label=Replacement: quality
  False       True        False       True
Negative    Negative    Positive    Positive
    273        1766         321         339

Data Role=VALIDATE Target=REP_quality Target Label=Replacement: quality
  False       True        False       True
Negative    Negative    Positive    Positive
   141         867         177         165

Predictions on unseen data:
Misclassification Rate:     0.454
Maximum Absolute Error:     0.987
Sum of Squared Errors:      248.566
Average Squared Error:      0.182
Root Average Squared Error: 0.427
```

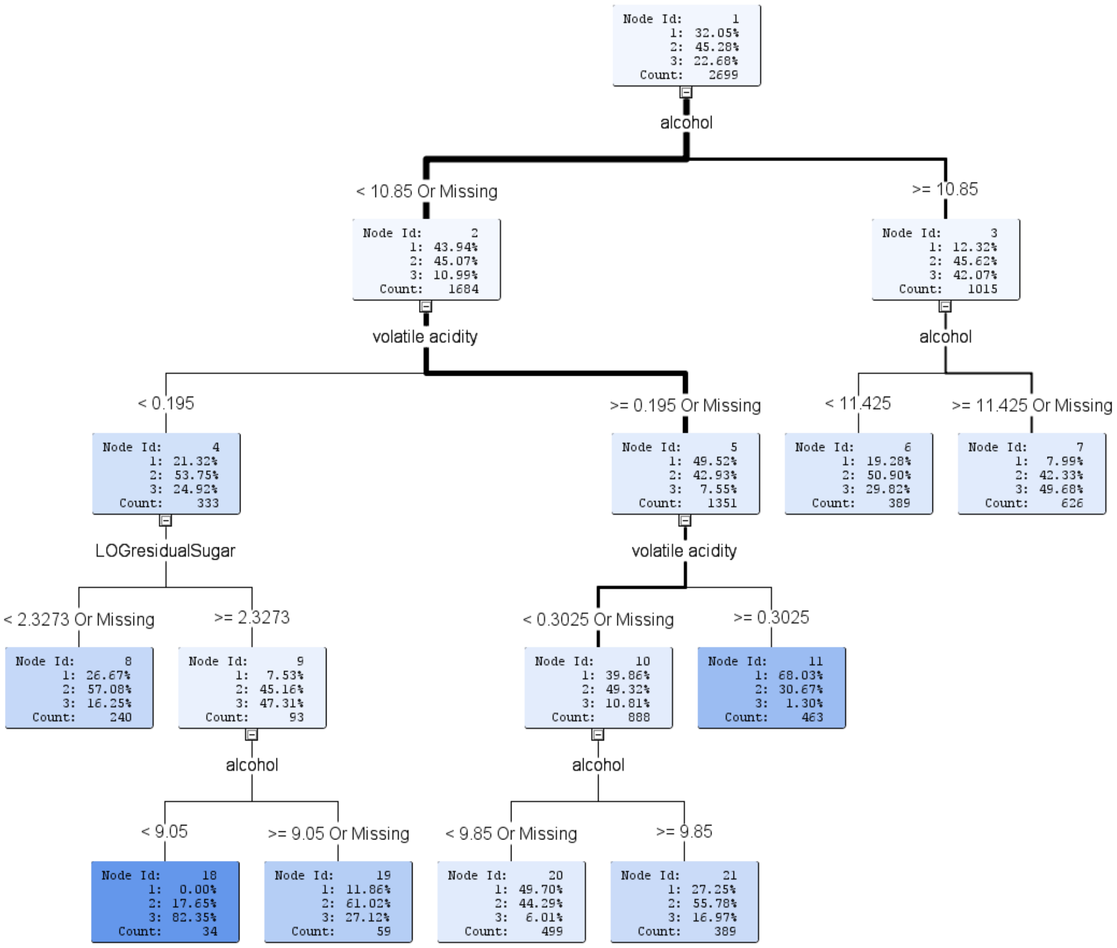
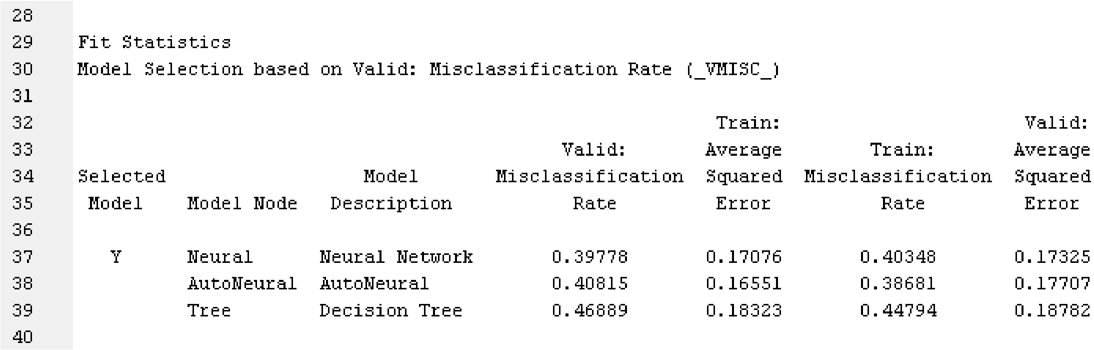


Figure 25: Decision Tree on white wine dataset

5.5 Neural-and-Auto-Neural Networks comparison

Two other classification models are training the data against a neural or auto-networks. The neural network will be trained with the *Model Selection Criterion = 'Misclassification'*, and the *'AutoNeural'* will be trained out-of-the-box without changing any parameters. Additionally, the *Model Comparison* node will be attached to all to see which classifiaction task has performed best.
When looking at the results, we can see that the Neural Network performed best in the overall with an accuracy of around 40%.



```
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

                                                                   Train:                           Valid:
                                                  Valid:          Average           Train:          Average
Selected                     Model           Misclassification    Squared     Misclassification     Squared
 Model     Model Node     Description              Rate             Error            Rate             Error

   Y       Neural        Neural Network          0.39778          0.17076          0.40348          0.17325
           AutoNeural    AutoNeural              0.40815          0.16551          0.38681          0.17707
           Tree          Decision Tree           0.46889          0.18323          0.44794          0.18782
```

Figure 26: Neural Networks and Decision Tree comparison

5.6 Compare Multi-class Classifiers in Azure Machine Learning Studio

One last approach in the modelling section is testing the preprocessed dataset on more advanced classifier algorithms which contains boosting and bagging methods. Some of this advanced algorithms are called out-of-the-box-learner and show with little or no tuning compelling results. In fact, these algorithms are the panacea[3] of all data mining problems and are going to be implemented on almost all datasets these days. For instance, Random Forests a hybrid learner of bagging and random subspace method (RSM) offers many other features such as permutation importance and MDS plots so that with one algorithm many data exploration tasks are covered. With Azure Machine Learning Studio[4] we get the availability to implement this advanced algorithm quickly and compare them against each other. Not only does Azure ML Studio offer great algorithm, it is also easy to handle and provides a web deployment API to set up the model as a service quickly.

The preprocessed wine quality data was exported and uploaded to the Azure cloud, where first all correct columns were selected and the data was split into training (0.8) and testing data (0.2). The data is trained against a Multiclass Neural Network, Multiclass decision

[3] panacea lat. = something that can solve many problems

[4] `https://studio.azureml.net/`

Jungle, Multiclass Logistic Regression, Multiclass Decision Forest and an advanced one vs all Support Vector Machine algorithm. After the evaluation of every model, an R script will select Macro Precision and Marco Recall and will be compared against each other. Further, the models will be trained on the replacement label with three bins and the original label.
As we can see in the Figure 28, the decision Forest shows the highest accuracy of 72% and can predict according to Figure 29 the average class best with an accuracy of 74%. With some little tuning (*mtry and ntree*), the model's accuracy could have been improved by 2%, and class 1 and class 2 show better stability (Class 2: 0%). Regarding the original label the, also the decision forest performed best. However, only class 5 and 6 showed accuracy above 60%. Future Importance were also applied and the least significant dimensions were deselected. However, this shown a decrease in accuracy and therefore all dimension were used during training.

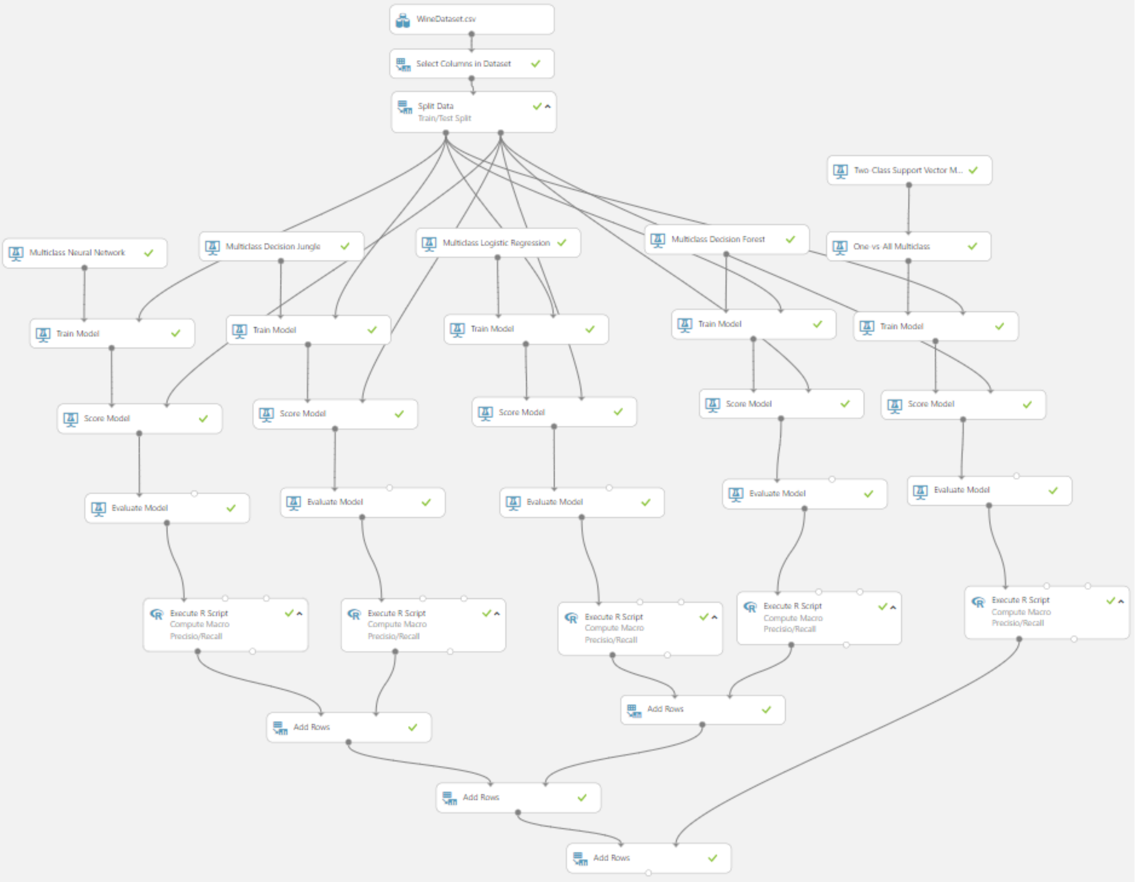
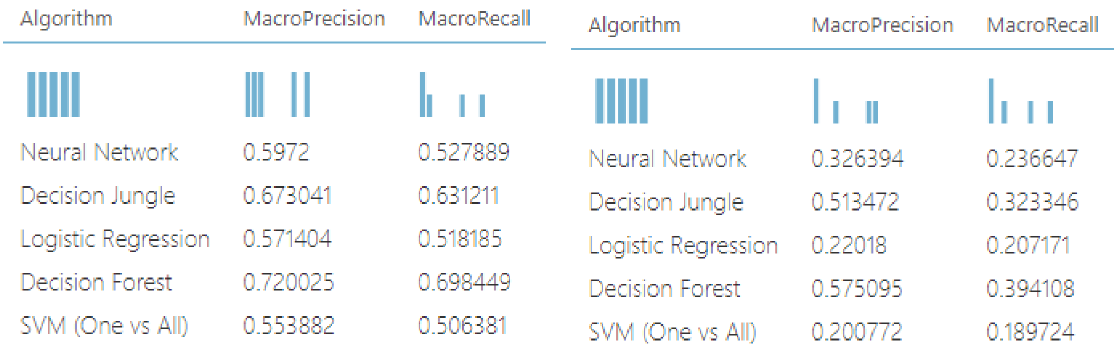


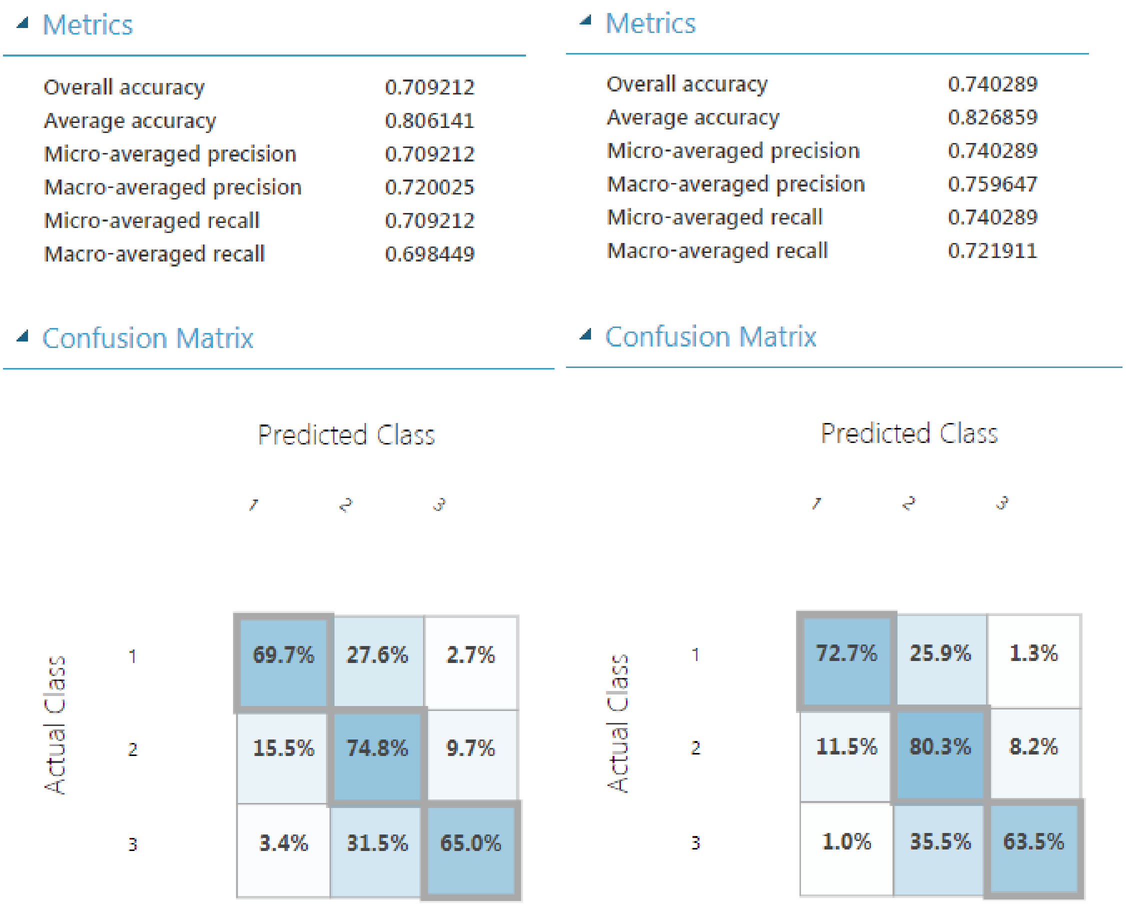
Figure 27: Compare Multi-class Classifiers: White Wine Dataset (Pre-Processed)
5



Algorithm	MacroPrecision	MacroRecall
Neural Network	0.5972	0.527889
Decision Jungle	0.673041	0.631211
Logistic Regression	0.571404	0.518185
Decision Forest	0.720025	0.698449
SVM (One vs All)	0.553882	0.506381

Algorithm	MacroPrecision	MacroRecall
Neural Network	0.326394	0.236647
Decision Jungle	0.513472	0.323346
Logistic Regression	0.22018	0.207171
Decision Forest	0.575095	0.394108
SVM (One vs All)	0.200772	0.189724

Figure 28: Compare Multi-class Classifiers: left with three bins and right with the original quality



Metrics

Overall accuracy	0.709212
Average accuracy	0.806141
Micro-averaged precision	0.709212
Macro-averaged precision	0.720025
Micro-averaged recall	0.709212
Macro-averaged recall	0.698449

Confusion Matrix

Actual Class \ Predicted Class	1	2	3
1	69.7%	27.6%	2.7%
2	15.5%	74.8%	9.7%
3	3.4%	31.5%	65.0%

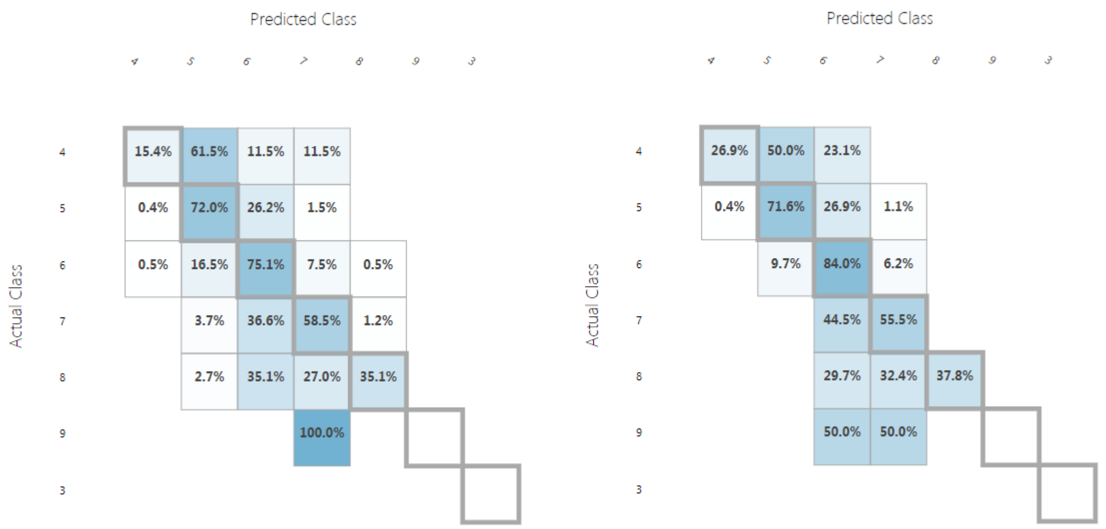
Metrics

Overall accuracy	0.740289
Average accuracy	0.826859
Micro-averaged precision	0.740289
Macro-averaged precision	0.759647
Micro-averaged recall	0.740289
Macro-averaged recall	0.721911

Confusion Matrix

Actual Class \ Predicted Class	1	2	3
1	72.7%	25.9%	1.3%
2	11.5%	80.3%	8.2%
3	1.0%	35.5%	63.5%

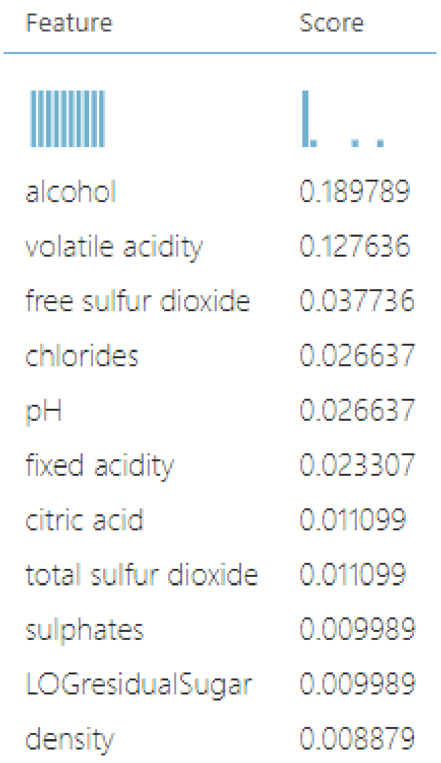
Figure 29: Accuracy on Decision Forest 3 bin out-of-the-box and (right) with tuning *(ntree=300, mtry=64)*



Actual \ Predicted	4	5	6	7	8	9	3
4	15.4%	61.5%	11.5%	11.5%			
5	0.4%	72.0%	26.2%	1.5%			
6	0.5%	16.5%	75.1%	7.5%	0.5%		
7		3.7%	36.6%	58.5%	1.2%		
8		2.7%	35.1%	27.0%	35.1%		
9				100.0%			
3							

Actual \ Predicted	4	5	6	7	8	9	3
4	26.9%	50.0%	23.1%				
5	0.4%	71.6%	26.9%	1.1%			
6		9.7%	84.0%	6.2%			
7			44.5%	55.5%			
8			29.7%	32.4%	37.8%		
9			50.0%	50.0%			
3							

Figure 30: Accuracy on Decision Forest with original quality out-of-the-box and (right) with tuning *(ntree=300, mtry=64)*



Feature	Score
alcohol	0.189789
volatile acidity	0.127636
free sulfur dioxide	0.037736
chlorides	0.026637
pH	0.026637
fixed acidity	0.023307
citric acid	0.011099
total sulfur dioxide	0.011099
sulphates	0.009989
LOGresidualSugar	0.009989
density	0.008879

Figure 31: Permutation Importance from Random Forests techniques, deselecting the last 3 and 5 has shown a decrease in accuracy

6 Evaluation

Having applied visualisations and modelling techniques it can be concluded that the quality of the wine does not go very well with chemical properties. However, there are many attributes which define the sensory data of white wine. However, when it comes to clarification of one quality level the data shows lots of variety which makes it hard to separate the classes in the prediction.
Another problem with the dataset is that the data is very centralised which means most wines are in average quality. Binning them into '*Good'* and *'Bad'* wines would not make much sense as it would bypass the average. Another approach for the classification is binning the sensory data into Good Bad and Average. The most promising algorithm technique on the white wine data set is the out-of-the-box learner such as Random Forests. Advanced hybrid algorithms are considered pancreas as they can be almost applied to any datasets and offer strong results. As we can see in the experiment in Azure Machine Learning Studio, bagging and boosting algorithm performs best in the overall.

Having applied all these techniques, we can conclude:

1. When the percentage of alcohol decreases, the density grows.
2. The higher the quality, the higher the probability of good wine.
3. The better the quality the fewer chlorides the probe contains.
4. When residual sugar increases, the quality will decrease.
5. When the level of alcohol decreases, the residual sugar increase.
6. Residual Sugar and density are strongly positively correlated.
7. Most of the wine data are average classified.
8. When applying cluster techniques, quality cannot be classified

Neither Regression nor Classification Techniques produced any tangible results because most data currently gathered weren't suitable for accurate precision. For a proper Machine Learning process to occur physicochemical data is not enough, additional metadata such as *'age'* and *'grape'* and other big data is required as well.

If Vinho Verde wants to continue working with its current data, it is suggested using an out-of-the-box algorithm as the most realistic one. Other approaches with more metadata would provide more in-depth insights but require more information to be applicable.
It is recommended that Vino Verde and iLab expand their data collection efforts to more accurately represent the sensory data of white wine.

7 Deployment

If Vino Verde wants to continue working with its current data or will have enough metadata to get a more stable prediction a web service to test the physicochemical data against quality would become very handy.

With the third modelling approach in Azure Machine Learning Studio, it is not difficult to deploy a web service to test new data. The deployed web service can be used on every MS Excel or can be implemented as an API on an internal landing website within Vino Verde's network.

The Figure 32 shows how the model can be deployed with the input and output nodes and Figure 33 shows how the new data can be tested via Request and Response or Batch execution.

Having implemented such a web service, we can feed the model to get more accurate data and may identify in early stages how the quality of the wine can be improved.

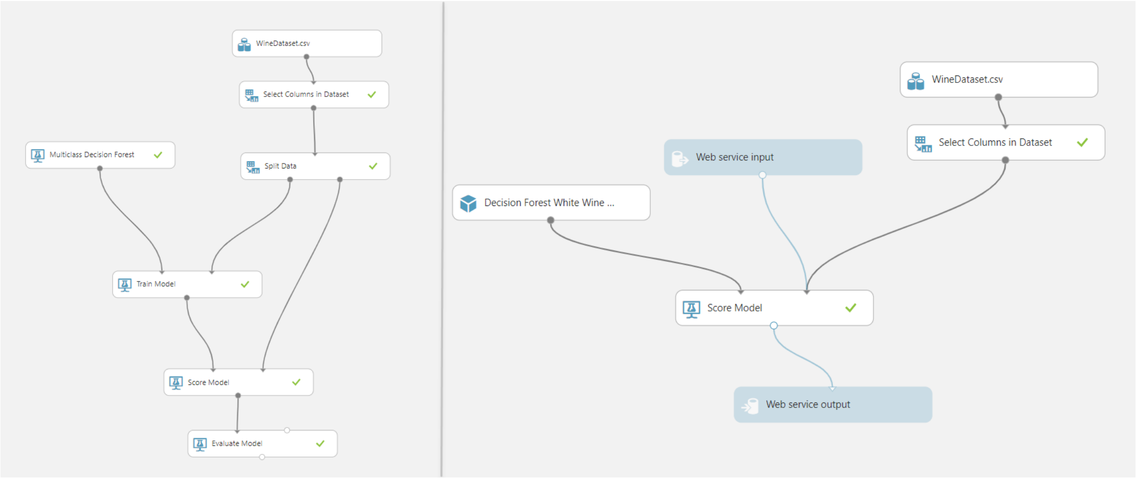
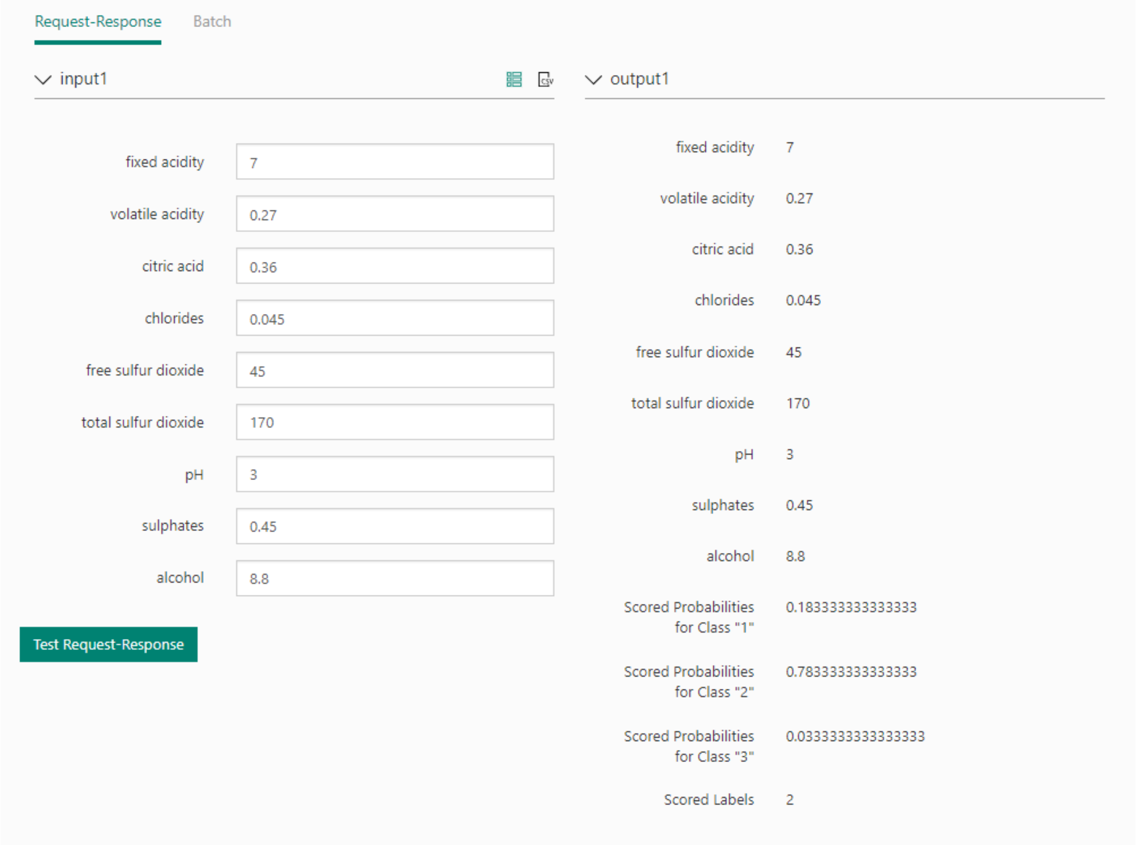


Figure 32: On the left: Final Model for web service and on the right with implemented input nodes



input1	
fixed acidity	7
volatile acidity	0.27
citric acid	0.36
chlorides	0.045
free sulfur dioxide	45
total sulfur dioxide	170
pH	3
sulphates	0.45
alcohol	8.8

output1	
fixed acidity	7
volatile acidity	0.27
citric acid	0.36
chlorides	0.045
free sulfur dioxide	45
total sulfur dioxide	170
pH	3
sulphates	0.45
alcohol	8.8
Scored Probabilities for Class "1"	0.183333333333333
Scored Probabilities for Class "2"	0.783333333333333
Scored Probabilities for Class "3"	0.0333333333333333
Scored Labels	2

Figure 33: The testing API which can be implemented on a website or accessed via MS Excel

8 Appendix

The R Script code to select the accircacy from the trained model.

```
dataset <- maml.mapInputPort(1)


# Compute (macro) average of the precision and recall columns
macroAvgs<-colMeans(dataset[,10:11]) #colums where Macros are stored ...
    (differ on every dataset)

data.set<-data.frame(Algorithm='Neural Network', #name
    MacroPrecision=macroAvgs[1],
    MacroRecall=macroAvgs[2]
)


maml.mapOutputPort("data.set");
```

Bibliography

Daniel Workman. Wine Exports by Country, October 2018. URL `http://www.worldstopexports.com/wine-exports-country/`.

Dean Geoffey. On the road report: Vinho Verde being made in innovative ways with new blends, November 2016. URL `http://www.the-buyer.net/tasting/wine/tradition-innovation-vinho-verde-targeting-uk-growth/`.

David Kernmode. What is Vinho Verde? Understanding Portugal's most famous wine and how Vinho Verde is changing, August 2018. URL `http://www.the-buyer.net/insight/portugals-vinho-verde-black/`.

Jancis Robinson, editor. *The Oxford companion to wine*. Oxford University Press, Oxford ; New York, 3rd ed edition, 2006. ISBN 978-0-19-860990-2. OCLC: ocm70699042.

Edouard Clouet-Foraison. Vinho Verde: 101 | Vinho Verde, 2015. URL `http://winesofvinhoverde.com/vinho-verde-101/`.

JancisRobinson.com, a. URL `https://www.jancisrobinson.com/learn/wine-regions/portugal/minho/vinho-verde`.

Portuguese Vinho Verde White | Regional Wine Style, b. URL `https://www.vivino.com/wine-styles/portuguese-vinho-verde-white`.

Roy Teranishi, Emily L Wick, and Irwin (Eds.) Hornstein. *Flavor Chemistry - Thirty Years of Progress*. Springer, 1 edition, 1999. ISBN 978-1-4615-4693-1.

D. Smith and R Margolskee. Making sense of taste, Scientific American, Special issue. 3 (16):84–92, 2006.

A. Lengin, A. Rudnitskaya, and L. Luvova. Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception, Analytica Chimica Acta 484 (1) (2003) 33–34. *Analytica Chimica Acta*, 484(1): 33–34, 2003.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, November 2009. ISSN 01679236. doi: 10.1016/j.dss.2009.05.016. URL `http://linkinghub.elsevier.com/retrieve/pii/S0167923609001377`.

Alexander Pandell. Wine Acidity, 1999. URL `https://www.wineperspective.com/wine-acidity/`.

Sugar in Wine Chart (Calories and Carbs), May 2015. URL https://winefolly.com/review/sugar-in-wine-chart/.

Andrew Waterhouse. The Power of Understanding Wine Tech Sheets, May 2015. URL https://winefolly.com/review/understanding-wine-tech-sheets/.

Madeline Puckette. The 5 Basic Wine Characteristics, July 2012. URL https://winefolly.com/review/wine-characteristics/.

szymanskiea. When a wine is salty, and why it shouldn't be, October 2015. URL http://wineoscope.com/2015/10/02/when-a-wine-is-salty-and-why-it-shouldnt-be/.

SAS Enterprise Miner Reference Help v14.1, c. URL https://support.sas.com/documentation/onlinedoc/miner/.

David Starker. Cramer's V, 2010. URL http://changingminds.org/explanations/research/analysis/cramers_v.htm.