Wine Quality Prediction

HarvardX Capstone Project
Marcus Vinicius Goulart Gonzaga Junior

09/04/2019

Table of Contents

Summary

The determination of the sensorial quality of wines is of great interest to the entire wine industry. From producers to consumers, there is high interest in the subject. There are several motivations, ranging from the cost of certification to the definition of market prices. In this study, we propose to use machine learning techniques to predict wine taste preferences based on physicochemical properties from wine analyses. We use data obtained from the UCI Machine Learning Repository, [Cortez et al., 2009], a significant dataset, with physicochemical properties as expert evaluation as well. White Vinho Verde samples were obtained from Minho, a northwest region of Portugal. The results suggest that the field of sensorial taste prediction is reliable.

Introduction

Wine quality classification is an exciting task since taste is the least understood of the human senses. The paper titled Modeling wine preferences by data mining from physicochemical properties, [Cortez et al., 2009], inspired this work. At the Cortez study, the objective was focused on the certification process that the wine industry needs to carry out. This process is currently carried out by experts and is usually slow and expensive, and prediction models is a process to reduce costs and increase efficiency. Here, we use a different approach, changing the focus from industrial certification processes, and looking for a market view. In that context, the relevant question is to answer if we can predict if wine is excellent ou very poor, the extremes of the curve is what matter in this perspective. So The main goal is to know what level of precision a model can predict the quality of the very best and inferior wines. Another Relevant aspect is in trying to understand some relationship among data, to obtain some insights that may be important to improve the knowledge of the wine quality. Unlike the original work, we used a classification strategy rather than regression. We also re-arrange dataset in just three category groups: Best, Normal, and Bad. The result was a simplified data structure and easy comprehension since we are looking for extremes of quality levels, seeking excellent and inferior wines. Finally, we use some different models that the original work made, trying to discover the best model applies in this context.

Dataset:

The dataset used can be obteined by the link: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/ and the name of file is winequality-white.csv

The Input Variables ((based on physicochemical tests) are: - **fixed acidity:** most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

- **volatile acidity:** the amount of acetic acid in wine,hich at too high of levels can lead to an unpleasant, vinegar taste
- **citric acid:** found in small quantities, citric acid can add 'freshness' and flavor to wines
- **residual sugar:** the amount of sugar remaining after fermentation stops , it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
- **chlorides:** the amount of salt in the wine
- **free sulfur dioxide:** the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
- **total sulfur dioxide:** amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
- **density:** the density of water is close to that of water depending on the percent alcohol and sugar content
- **pH:** describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic),most wines are between 3-4 on the pH scale
- **sulphates:** a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
- **alcohol:** the percent alcohol content of the wine.

The output Variable (based in sensory data):

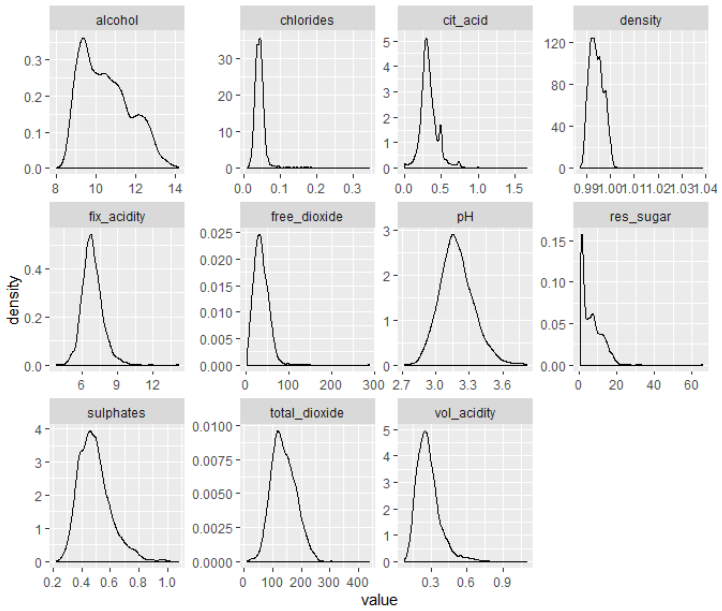
quality: score between 0 to 10 maded in blind mode by experts

Exploratory Data Analysis

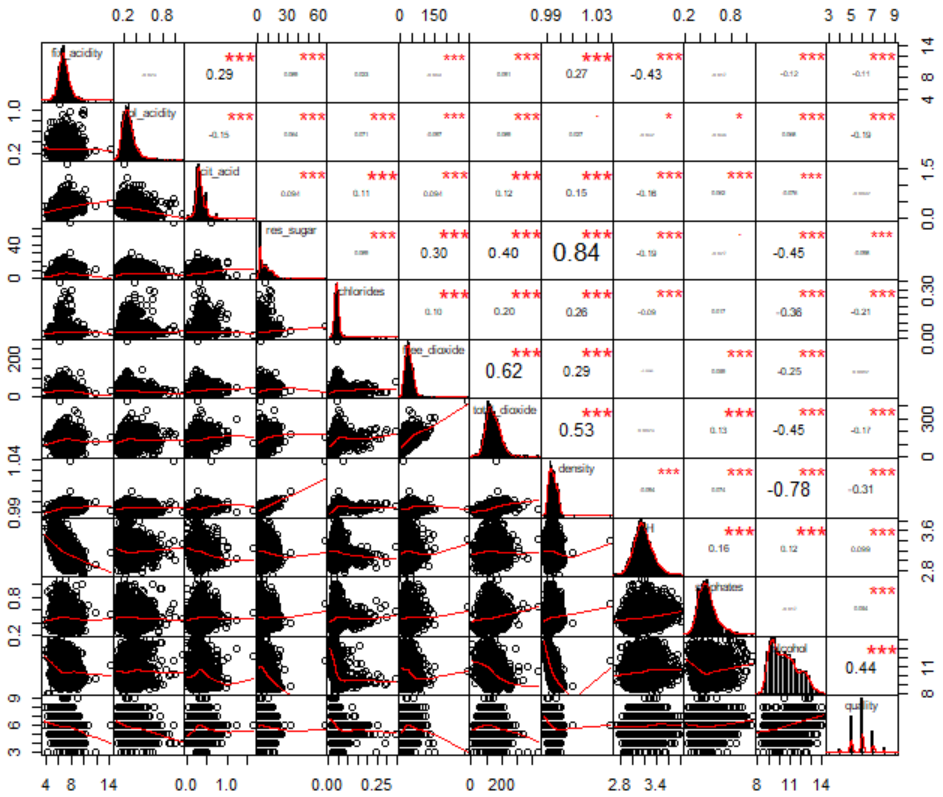
Summary Dataset

```
##   fix_acidity      vol_acidity       cit_acid       res_sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides        free_dioxide    total_dioxide      density
##  Min.   :0.00900   Min.   :  2.00   Min.   :  9.0   Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00   1st Qu.:108.0   1st Qu.:0.9917
##  Median :0.04300   Median : 34.00   Median :134.0   Median :0.9937
##  Mean   :0.04577   Mean   : 35.31   Mean   :138.4   Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00   3rd Qu.:167.0   3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00   Max.   :440.0   Max.   :1.0390
##        pH           sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

It seems that there is symmetry in several distributions since the majority of features have very similar median and mean. Let´s see on the graph:



Except for residual sugar and alcohol, they looks have symmetrical distribution, although some have outliers in the positive side. Nothing special has been found so far, let's start a bivariate analysis with the correlation matrix.



The relationships among quality and independent variables, alcohol has the strongest correlation, the second is density, and the third is chlorides. Since the correlation between density and alcohol is -0.78, the fact that they together seemed natural.

correlations among quality

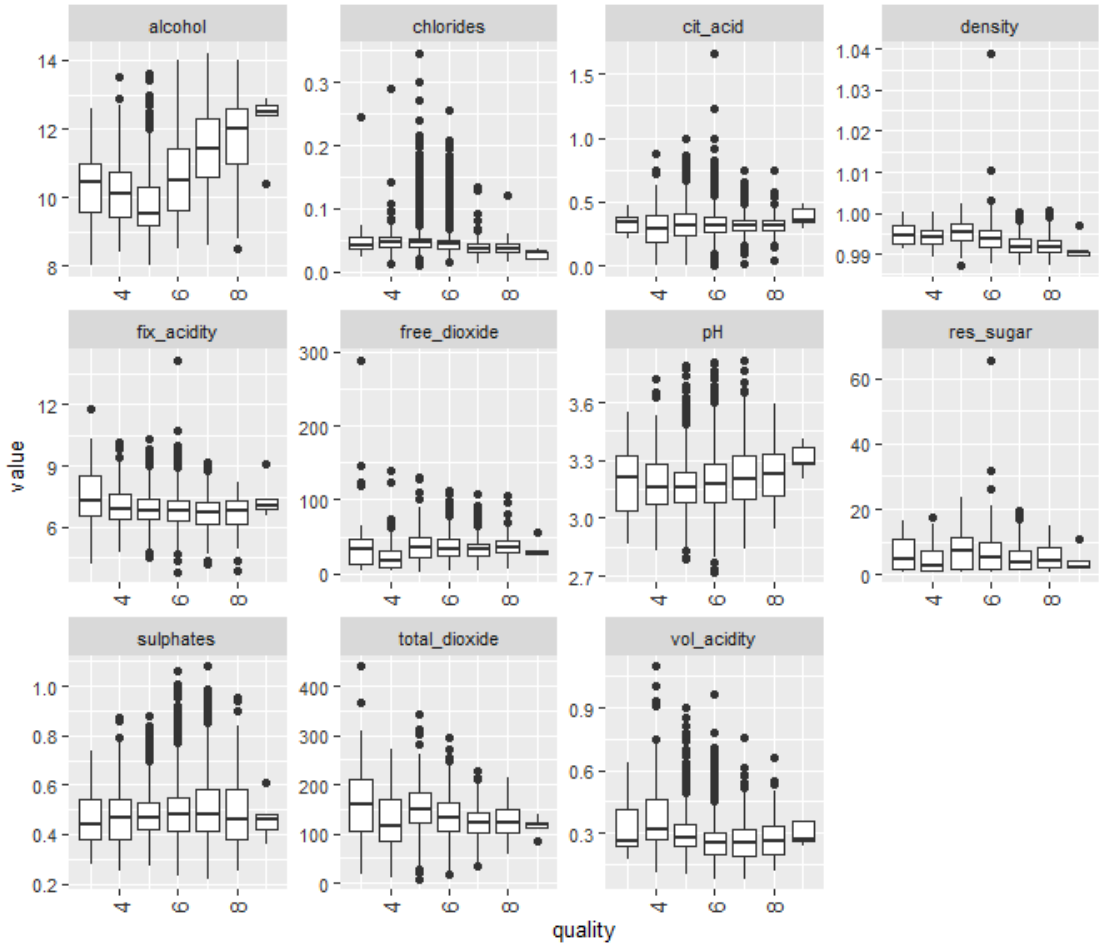
variable	correlation
alcohol	0.44
density	-0.31
chlorides	-0.21

In terms of relationships between independent variables, some strong correlations are observed:

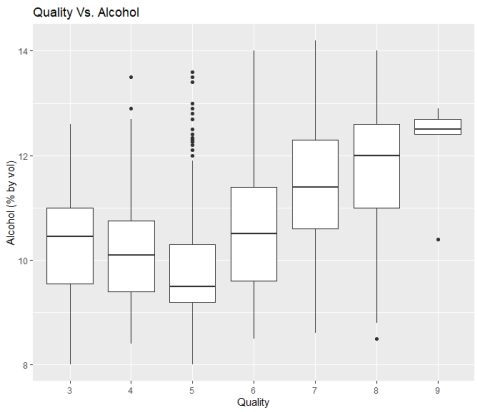
high correlated independent variables

variables	correlation
residual sugar - density	0.84
free sulfur dioxide - total sulfur dioxide	0.62
total sulfur dioxide - density	0.53

Plotting variables distribution relative as quality

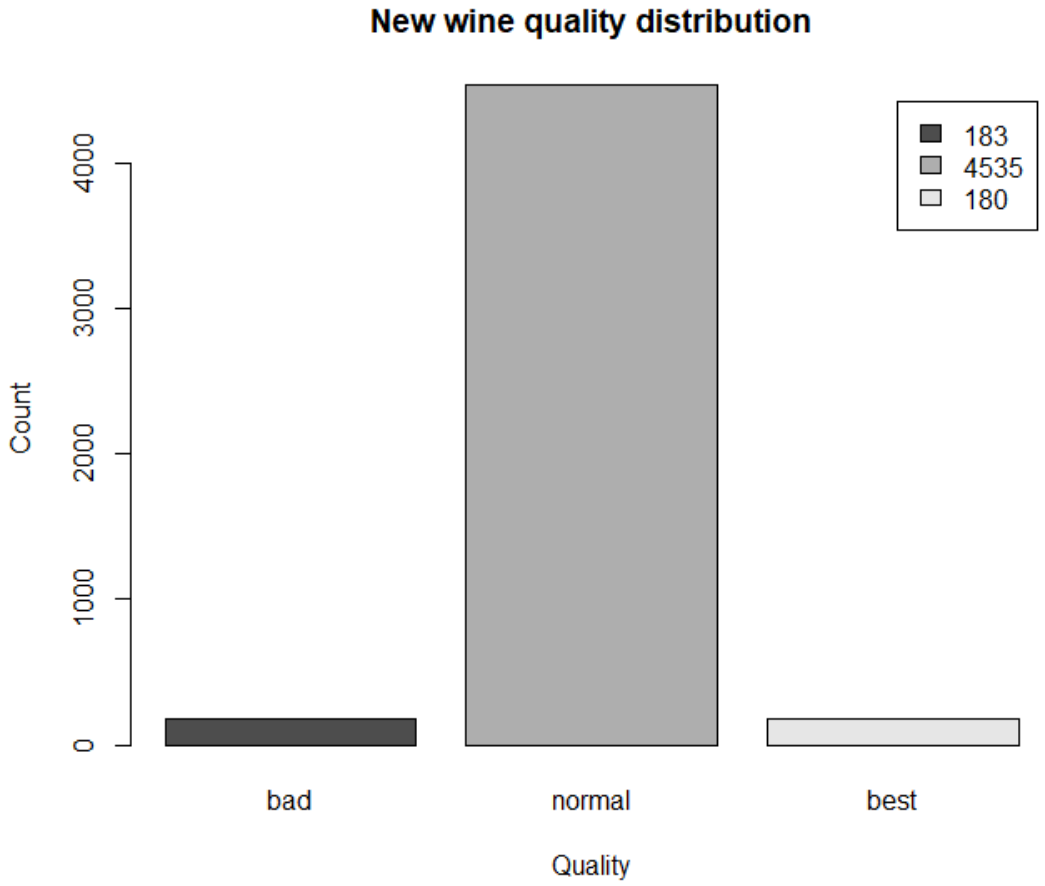


Alcohol seems to have the most variation among quality. Density goes down when quality improves. Plot zooming to see it in detail:

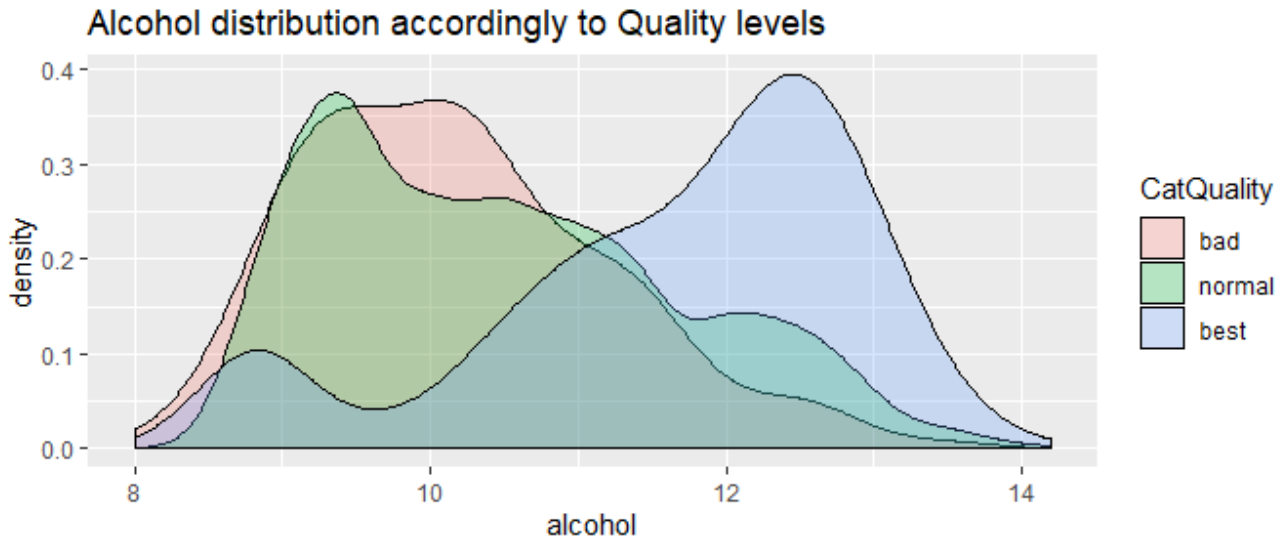


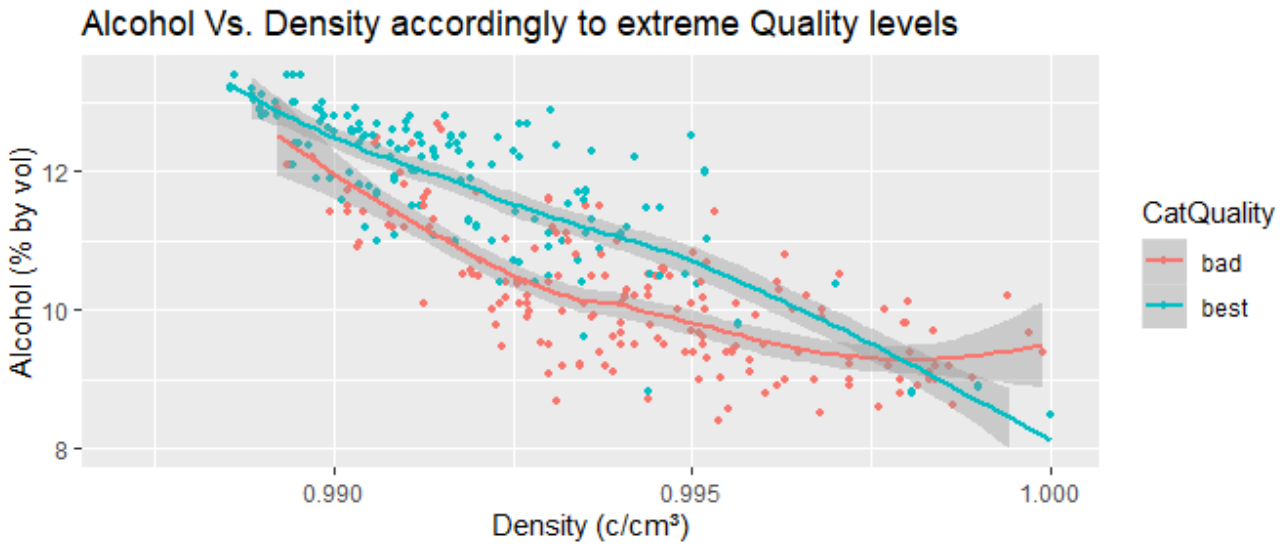
Data Wrangling

As we plan, we transform the quality parameter more simply. Bad, for grades 3 and 4, Best for grades 8 and 9, and Normal to wines that obtained grading 5,6 or 7. After transformation, only remains 3 categories. As expected, since the dataset is very unbalanced, the Best wines and Worst corresponds to about 3.5% each of Total.



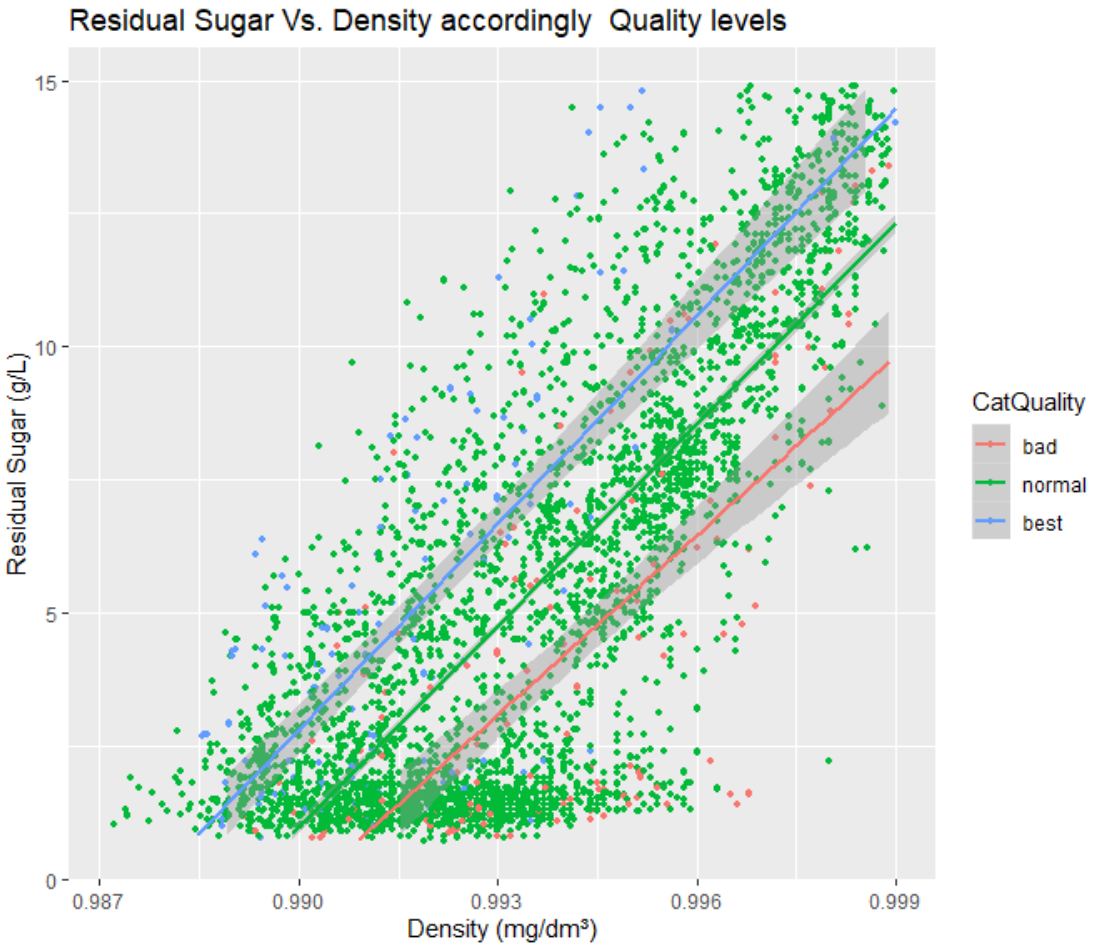
More Analysis



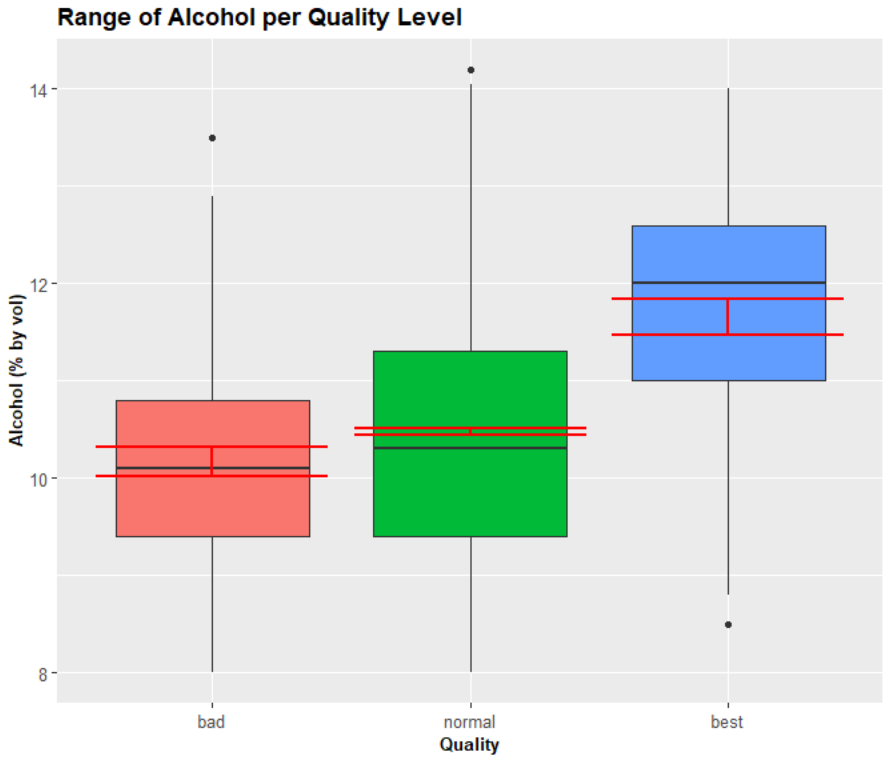


Points to more Alcohol and more density comes to better quality

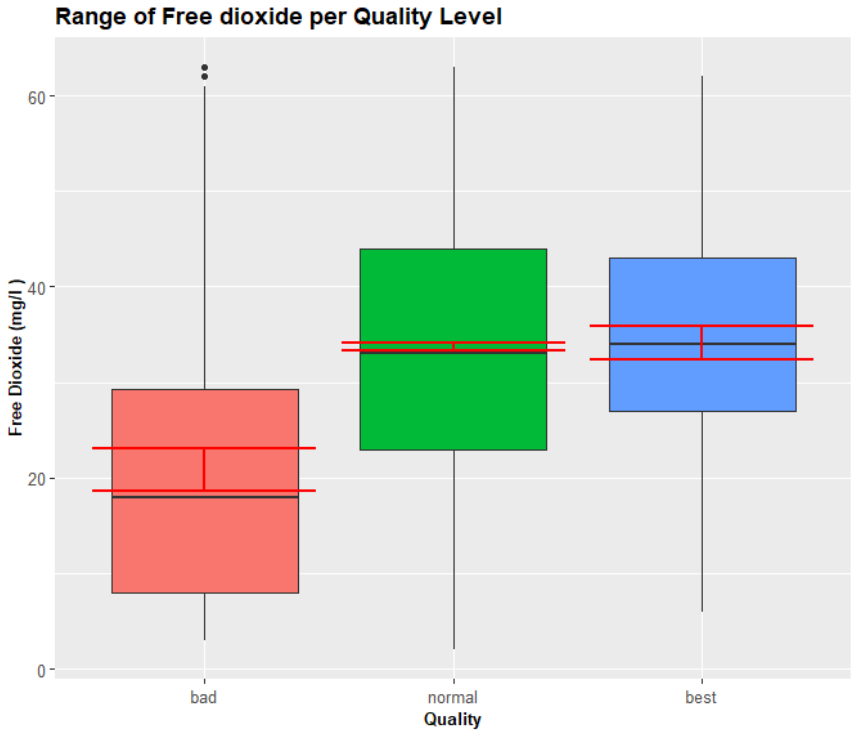
Another view of the Relationship of Alcohol and density relationship quality



For the same density, higher residual sugar seems to have better quality.



Red Bars shows 95% confidence, indicates that more alcohol tends to more quality.



Free Dioxide tends to less quality, especially to bad.

Modeling

Divide Dataset

First we divide a dataset in two parts, Training with 80% of data, and Testing wine with 20% of dataset.

Now we aplly some of best Machine Learning classification algorithms. We will try SVM - Support-vector machine, GBM - Gradient boost Machine, RF - Random Forest and XGB - XGBoost

SVM Support-vector machine

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad normal best
##     bad      0      0    0
##     normal  36    907   36
##     best     0      0    0
##
## Overall Statistics
##
##                Accuracy : 0.9265
##                  95% CI : (0.9083, 0.942)
##     No Information Rate : 0.9265
##     P-Value [Acc > NIR] : 0.5313
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: bad Class: normal Class: best
## Sensitivity             0.00000        1.0000     0.00000
## Specificity             1.00000        0.0000     1.00000
## Pos Pred Value              NaN        0.9265         NaN
## Neg Pred Value          0.96323           NaN     0.96323
## Prevalence              0.03677        0.9265     0.03677
## Detection Rate          0.00000        0.9265     0.00000
## Detection Prevalence    0.00000        1.0000     0.00000
## Balanced Accuracy       0.50000        0.5000     0.50000
```

The result seems strange since the prediction SMV model classificate all wines as Normal. Lets check in Original dataset to try deep understand

```
## Confusion Matrix and Statistics
##
##           Reference
```

```
## Prediction   3   4   5   6   7   8   9
##          3   0   0   0   0   0   0   0
##          4   0   0   0   0   0   0   0
##          5   1  25 153  97  12   4   0
##          6   1   9 130 357 158  31   1
##          7   0   0   0   0   0   0   0
##          8   0   0   0   0   0   0   0
##          9   0   0   0   0   0   0   0
##
## Overall Statistics
##
##                Accuracy : 0.5209
##                  95% CI : (0.4891, 0.5526)
##     No Information Rate : 0.4637
##     P-Value [Acc > NIR] : 0.0001915
##
##                   Kappa : 0.1858
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
## Sensitivity          0.000000  0.00000   0.5406   0.7863   0.0000  0.00000
## Specificity          1.000000  1.00000   0.8003   0.3714   1.0000  1.00000
## Pos Pred Value            NaN      NaN   0.5240   0.5197      NaN      NaN
## Neg Pred Value       0.997957  0.96527   0.8108   0.6678   0.8264  0.96425
## Prevalence           0.002043  0.03473   0.2891   0.4637   0.1736  0.03575
## Detection Rate       0.000000  0.00000   0.1563   0.3647   0.0000  0.00000
## Detection Prevalence 0.000000  0.00000   0.2983   0.7017   0.0000  0.00000
## Balanced Accuracy    0.500000  0.50000   0.6705   0.5789   0.5000  0.50000
##                      Class: 9
## Sensitivity          0.000000
## Specificity          1.000000
## Pos Pred Value            NaN
## Neg Pred Value       0.998979
## Prevalence           0.001021
## Detection Rate       0.000000
## Detection Prevalence 0.000000
## Balanced Accuracy    0.500000
```

The confusion matrix aplied at all categorical quality variables helps to clarify. The prediction tends to go center, maybe because the dataset is so unbaleced. The best wines (quality 8 and 9) as predicted as (6), and bad quality wines(quality 3 and 4) was predicted by model as (5 and 6).

GBM Gradient Boost Machine

```
## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction bad normal best
##     bad      9      4    0
##     normal  27    899   31
##     best     0      4    5
##
## Overall Statistics
##
##                Accuracy : 0.9326
##                  95% CI : (0.915, 0.9475)
##     No Information Rate : 0.9265
##     P-Value [Acc > NIR] : 0.2534
##
##                   Kappa : 0.2793
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: bad Class: normal Class: best
## Sensitivity            0.250000        0.9912    0.138889
## Specificity            0.995758        0.1944    0.995758
## Pos Pred Value         0.692308        0.9394    0.555556
## Neg Pred Value         0.972050        0.6364    0.968041
## Prevalence             0.036772        0.9265    0.036772
## Detection Rate         0.009193        0.9183    0.005107
## Detection Prevalence   0.013279        0.9775    0.009193
## Balanced Accuracy      0.622879        0.5928    0.567324
```

RF Random Forest

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad normal best
##     bad      7      1    0
##     normal  29    906   23
##     best     0      0   13
##
## Overall Statistics
##
##                Accuracy : 0.9459
##                  95% CI : (0.9298, 0.9592)
##     No Information Rate : 0.9265
##     P-Value [Acc > NIR] : 0.009451
##
##                   Kappa : 0.4155
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
```

```
##
##                      Class: bad Class: normal Class: best
## Sensitivity            0.194444        0.9989     0.36111
## Specificity            0.998940        0.2778     1.00000
## Pos Pred Value         0.875000        0.9457     1.00000
## Neg Pred Value         0.970134        0.9524     0.97619
## Prevalence             0.036772        0.9265     0.03677
## Detection Rate         0.007150        0.9254     0.01328
## Detection Prevalence   0.008172        0.9785     0.01328
## Balanced Accuracy      0.596692        0.6383     0.68056
```

XGB Extreme Gradient Boost

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad normal best
##     bad      6      2    0
##     normal  30    904   24
##     best     0      1   12
##
## Overall Statistics
##
##                Accuracy : 0.9418
##                  95% CI : (0.9252, 0.9556)
##     No Information Rate : 0.9265
##     P-Value [Acc > NIR] : 0.03458
##
##                   Kappa : 0.3714
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: bad Class: normal Class: best
## Sensitivity            0.166667        0.9967     0.33333
## Specificity            0.997879        0.2500     0.99894
## Pos Pred Value         0.750000        0.9436     0.92308
## Neg Pred Value         0.969104        0.8571     0.97516
## Prevalence             0.036772        0.9265     0.03677
## Detection Rate         0.006129        0.9234     0.01226
## Detection Prevalence   0.008172        0.9785     0.01328
## Balanced Accuracy      0.582273        0.6233     0.66614
```

Model Benchmark

Accuracy commonly used to overall evaluate classification models. In this case, this measurement is not the most appropriate, since the data are very unbalanced concerning quality, the vast majority of the data are of medium quality, which implies a possible distortion of interpretation. Then we use Precision and Balanced Accuracy as the primary evaluation criterion. Note that False positives are less desirable than false negatives in our context. It means that it is preferable to not predict some Best Wines as Best than predict some Bad or Normal Wines as Best. The same apply to prediction Bad Wines

Best Wines

Best Wines predictor Model Benchmark

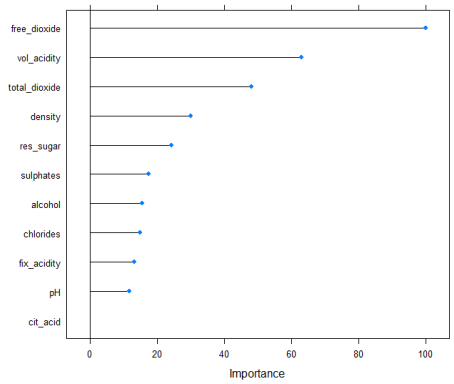
Model	Sensitivity	Specificity	Balanced	Precision
SVM	0.000	1.000	0.500	
GBM	0.139	0.996	0.567	0.556
RF	0.361	1.000	0.681	1.000
XGB	0.333	0.999	0.666	0.923

Bad Wines

Bad Wines predictor Model Benchmark

Model	Sensitivity	Specificity	Balanced	Precision
SVM	0.000	1.000	0.500	
GBM	0.250	0.996	0.623	0.692
RF	0.194	0.999	0.597	0.875
XGB	0.167	0.998	0.582	0.750

RF - Random Forest, outperformed the others, in best Wines category and also Bad Wines. The Variable ranking importance considering the Random Forest Model:



Results

The result showed that machine learning algorithms could capture, from physicochemical data, several aspects of sensory analysis made by humans, and then make useful predictions. Although this study is not able to causality levels, we believe that it is enough to have the confidence to continuos investigation in this field. Other datasets, different models, data engineering, hyperparameters tunning have to be considered. The main challenge is increasing Sensitivity levels at the same time maintain high levels of Specificity since usually that means trade-off.

Considering Random Forest, the best performed of Machine Learning Model used in this analysis, 100% considered as excellent coincided with the expert's assessment. For the lowest quality wines, the result was 87%. Even though this result in terms of precision may be considering very significant, the other side is the low Sensitivity rate shows that only 36% of Best Wines as recognized as is, and for Bad wines result is 19%. The others predicted as Normal. Translating this in simple words, we say:

"If Algorithm tells that a Wine is Good or some Wine is Bad, you can trust. However, you have to know that many Good Wines and Bad Wines will be out of the list since the model predicts most of them as Normal quality." Cheers?

References

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, ISSN: 0167-9236. https://www.sciencedirect.com/science/article/pii/S0167923609001377

https://archive.ics.uci.edu/ml/datasets/wine+quality

Rafael A. Irizarry, Introduction to Data Science Data Analysis and Prediction Algorithms with R. 2019-04-22. https://rafalab.github.io/dsbook/

Hadley Wickham & Garret Grolemund, R for Daqta Science. O´REILLY 2016