**Introduction**

The starting point of the project is a dataset made of 6497 different wines, their quantitative chemical features, and two factor variables, namely the rating of the wine (evaluated by three professional tasters) and the macro type of wine (white or red). The one and only transformation we have made is the ratio between free sulphur dioxide and total sulphur dioxide, in order to have a measurement of the consumption of the chemical entity which is a measurement of the fermentation of the wine.

**Cluster**

As a first introductive step, we performed the principal component analysis, resulting in a 50% of variability explained by the first two components. The plot on the first two pcs highlights the differences between red and white wines, more than the differences in the ratings. (Plot line 15 file Cluster.R).

Looking at the loadings:

Loadings:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10

fixed.acidity 0.387 0.465 0.117 0.311 0.383 0.246 0.206 0.382 0.355

volatile.acidity 0.293 -0.386 -0.175 0.386 0.306 0.274 -0.103 -0.338 -0.532

citric.acid 0.348 0.536 -0.211 -0.244 0.296 -0.430 -0.458

residual.sugar 0.138 0.477 -0.349 -0.291 0.194 -0.410 -0.356 0.148 0.435

chlorides 0.425 -0.183 0.114 0.206 -0.152 -0.624 -0.423 0.362

sulfur ratio 0.141 -0.307 -0.730 0.500 -0.281 -0.125

density 0.547 0.200 -0.225 -0.116 0.241 0.121 -0.718

pH -0.428 -0.311 -0.247 -0.473 0.297 0.479 -0.200 0.182 0.212

sulphates 0.323 -0.264 0.252 -0.213 -0.485 0.126 -0.555 0.346 -0.180

alcohol -0.370 -0.283 0.347 -0.115 0.126 0.331 -0.339 -0.437 0.333 -0.326

A green and red dots

Description automatically generated

We can see how the first principal component describe the type of fermentation, in fact counterbalanced by the alcohol, all the features with a significant loading described how structured the process is, in particular sulphates and density represents this characteristics, and in fact from the plot of the red/white wines we can see that, despite having a higher value in the alcohol, the red wines have a higher score on pc1. The second one describe the quantity of fermentation, indeed all the features that characterize a high fermented wine are counterbalance by the ones that have the opposite effect; the red wines, which ferment more than the whites, have a lower score on this component.

A graph with blue and green bubbles

Description automatically generatedA graph of a number of individuals

Description automatically generatedThe differences in the two are so relevant that a hierarchical clustering algorithm can be performed, in particular using Euclidean distance and Ward linkage the clustering structure is reconstructed, misclassifying only 4.3% of wines with the wrong type.

Given this clear distinction, we have decided to treat the two types in two different ways, and to describe their differences having performed two different analyses.

**Pre-process classification**

As a pre-processing step for the classifications, we need to get rid of the 6-evaluated wines, in fact both from a qualitative and a quantitative point of view, these kinds of wines are not clearly in one class or in the other, being that a 6-evvaluation is not completely good neither bad. For the quantitative method, we sampled 1000 datasets with the bootstrap method, and we saw that the distribution of our variables is not significant for the wine to be 6. Regarding the qualitative approach, the plot used was the different colour on the first two component of the pca.

Eliminating six-rated wines is in line with the general ethic of our research too: in fact they represent a wine which is not as good to be good, as bad to be bad therefore we can conclude that we can eliminate them.

**Data engineering**

In order to have more information on the dataset, we search for a way to construct these variables, and we decided to use ChatGPT: we engineered the problem dividing the research areas in three main drivers, and then asked the artificial intelligence to give us 15 different indices for each driver, so we asked the same question looking for indices for the three themes: fermentation, quality of the fermentation, territory.

At this point, we needed a tool to see if these indices are significant for the classification in good wines and bad wines; the problem is that the anova model is not good for this problem which have a high dimension, this technique tends to overestimate the significance, given that the statistics is divided by the number of rows of the dataset. So we came up with a fusion between the bootstrap and the anova, where we sampled multiple times from each new variables the values and performed an anova each time, having as result the fraction of times when the p value of the test was below a certain threshold (0.05), we didn’t use the mean of the p value because, being the bootstrapped data not independent, the mean loses some of its power. This method is not perfect; indeed, the assumptions are impossible to be met at each time, and moreover almost every variable didn’t come from a normal distribution, but we decided to use it anyway in order to have some indices to rank the variables and select the most significant. (See critic points for more).

So, the indices that we choose as most significant are the following:

For the wites:

|  |
| --- |
| "TotalAcidity" "SugarAlcohol" "SulphurDioxideIndex" "pHSulphates"  "DensityClarity"  "Fermentability"  "OxidationPotential" "FlavorIntensity" "BalanceAroma" "SugarAcid"  "SweetnessBody" "SulphurInteraction" "AlcohlpH" "Complexity" |
|  |
| |  | | --- | |  | |

For the reds:

"SulfurDioxide2ChlorideRatio" "Alcohol2Density.Ratio" "Sulphates2pH.Ratio"

[4] "Chloride2Sulfate.Ratio" "SugarAlcohol" "BalanceAroma"

[7] "AlcohlpH" "AcidRatio" "CitricAcidRatio"

"AlcoholSugarSquare" "pHSulphateInteraction" "AcidAlcoholInteraction"

Regarding our variables, this procedure was applied even to these features, resulting in the following:

“volatile acidity” "chlorides" "sulphur.ratio" "density" "alcohol"

For the reds:

"volatile.acidity" "citric.acid" "sulphur.ratio" "sulphites" "alcohol"

**Classification**

The variables present very different scales, and in the classification context, where the distances are fundamental (we have used Euclidean distance overall), this is a problem, so all the analysis below will be made on scaled variables, to have uniformity of judgment.

In order to estimate the error rate, we used always the same code for the evaluation of the AER through a 15-fold cross validation, while for the confusion matrix, we will represent the mean of the numbers that appear in that square of the matrix along all the steps of the cross validation.

KNN – This is the first classifier used in order to tackle the problem, these are the results:

|  |  |  |
| --- | --- | --- |
| KNN | k | AER |
| White – Raw | 37 | 14.67% |
| Red – Raw | 50 | 9.83% |
| White – GPT | 11 | 13.84% |
| Red - GPT | 10 | 9.58% |

The error rate gives good promises to find a good classifier, being this the simplest one we do not set the costs of misclassification, being this a rudimental tool used only for an introductive analysis.

LOGISTIC REGRESSION

|  |  |
| --- | --- |
| logit | AER |
| White – Raw | 20.37 % |
| Red – Raw | 10.4% |
| White – GPT | 20% |
| Red - GPT | 9.89 % |

The threshold chosen for deciding in which class a unit should be set to 1/3, in order to consider the costs of misclassification (need to be changed the reds)

White-Raw: we see how the probability for a wine to be good is mainly influenced by the alcohol, counterbalanced by the volatile acidity, so for a white wine the acid content is not appreciated.

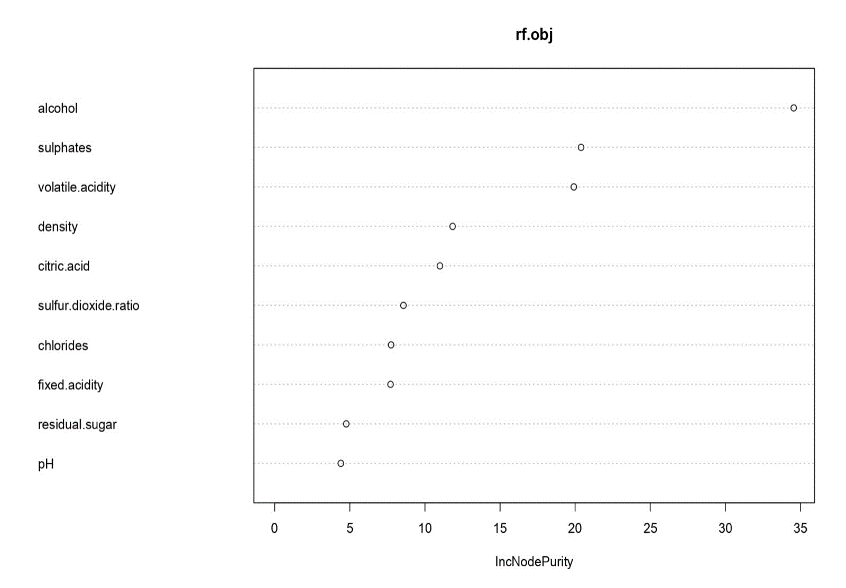
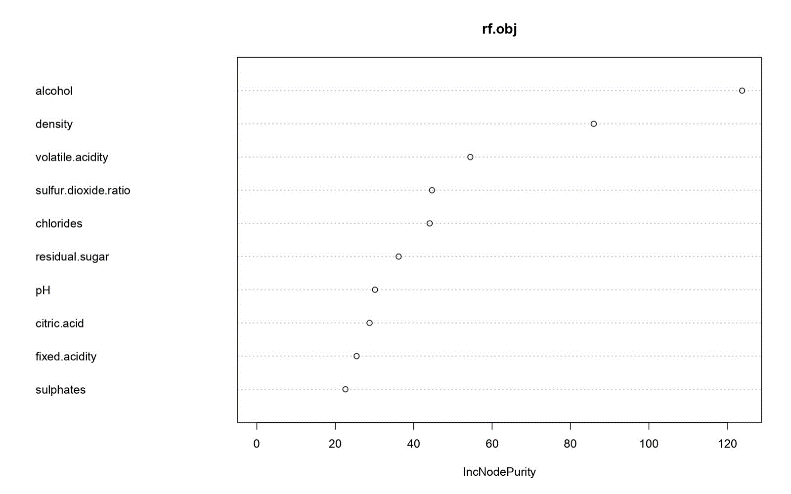
Red-Raw : here again alcohol is the main actor, influencing positively the quality, counterbalanced again by the acids, this time the sulphur ratio is relevant, given that this feature is associated with stability, an important feature with the reds.

Red-gpt: Quality in red wines is mainly determined by their alcohol content together with some territory characteristics of the wine. In particular we found that one significant factor is the relation between sulphates and ph , qualities caused by the type of soil substrate where the vineyards grow that are determinant in characterizing the chemical stability and the acid taste of the wine.

SVM

|  |  |  |  |
| --- | --- | --- | --- |
| SVM | Cost | Gamma | AER |
| Raw – White | 1 | 5 | 13.9% |
| Top – White | 10 | 1 | 10.5% |
| Raw – Red | 10 | 1 | 8.4% |
| Top - Red | 1 | 1 | 9.56% |

RANDOM FOREST



RED WHITE

This is the most deep learning method we will use in a total black box way. It performs very well, at the cost of the lack of interpretation. The only thing that we can say about it is the importance that the method gives to every variables, and in this way we assess the quality of our method bootstrap anova, in fact the variables which are the most used are the same that our bootstrap anova measure as the most significant.

Results:

|  |  |
| --- | --- |
| Random Forest | AER |
| Raw – White | 10.6% |
| Top – White | 8.8% |
| Raw – Red | 5.9% |
| Top - Red | 6.3% |

LDA/QDA

The costs of misclassification for a bad wine classified as a good one was set to be the double of the other one, using the techniques used on the Bayes classifier this can be included in the evaluation of the prior probabilities (settled to 0.5 both), resulting in a vector of priors: 1/3 and 2/3.

1. A screenshot of a grid

   Description automatically generatedRaw variables - White: The best variables, namely those which we used to built the classifier, are not strongly following a gaussian distribution, from the histograms the normality cannot be assessed, using again the bootstrap method some sort of normality is reconstructed, and we performed the analysis. The equality of covariance matrices is not verified; indeed this is the plot of the two matrices. Nevertheless, from the boxM test we have seen that the statistics on the eigenvalues are not so different and so we performed even lda, which indeed gives not bad results. For the qda the distribution to be followed was set to be the Student’s t, being that the normality assumptions are not met, with the feature method in the R command, this change of distribution can be used.

LDA 🡪 AER = (18% without priors) 22.29% confusion:

true-bad true-good

predicted-bad 78.933333 30.40000

predicted-good 9.733333 60.93333

QDA 🡪 AER = 19.63% confusion:

true-bad true-good

predicted-bad 83.933333 25.40000

predicted-good 9.933333 60.73333

A screenshot of a graph

Description automatically generated

1. TopGPT – White: the normality assumptions in this case are not verified at all, maybe some of the variables can be thought to be so but the results will be affected by this loss, and even the equality of covariance matrices cannot be assumed. Even in this case the Student’s t distribution is imposed to the qda to reduce the problems of the assumptions.

LDA 🡪 AER = 20.78% confusion:

true-bad true-good

predicted-bad 81.33333 28.00000

predicted-good 9.40000 61.26667

QDA 🡪 AER = 18,74% confusion:

true-bad true-good

predicted-bad 86.33333 23.00000

predicted-good 10.73333 59.93333

1. A screenshot of a graph

   Description automatically generatedRaw variables – Red: In this case the assumptions of normality cannot be rejected for almost all the variables, and even the assumption on the equality of covariance matrices can be accepted. Nevertheless, in order to make qda more robust, the Student’s t distribution is again imposed.

LDA 🡪 AER = 14.05% (9.99% without priors ) confusion:

true-bad true-good

predicted-bad 42.2 7.40000

predicted-good 1.6 12.86667

QDA 🡪 AER = 14.78% confusion:

true-bad true-good

predicted-bad 41.466667 8.133333

predicted-good 1.333333 13.133333

1. A screenshot of a graph

   Description automatically generatedTopGPT – Red: The normality of the covariates cannot be rejected for all the features, and even some statistical tests results in the acceptance of the hypothesis, event the equality of the covariance matrices can be accepted, both for the ratio of the statistics of the two matrices, and even qualitatively with the plot, not clearly as in the previous case.

LDA 🡪 AER = 13.5% confusion:

true-bad true-good

predicted-bad 42.400000 7.2

predicted-good 1.466667 13.0

QDA 🡪 AER = 12.6% confusion:

true-bad true-good

predicted-bad 43.73333 5.866667

predicted-good 2.20000 12.266667

so, overall, these Bayesian classifiers behave quite well in classify this dataset, in particular with the red wines, it is possible to see how in the whites the error rate of the qda is much less than the rate in the lda, because of the important difference in the covariance matrices, while for the raw variables in the reds, the lda outperformed the qda, being that the variables are with the same covariance matrix. So, this is presented as an overview on these methods; the great problem with this is the representation of the results, and this is the reason why we will perform the Fisher analysis.

FISHER

We choose this approach to represent the variables and the classification region. We performed the Fisher’s analysis on the directions of maximum differentiation, and we will represent the results on the first two fisher’s directions, evaluating the AER only on the first, because in all the cases the first explains almost all the differentiation.

1. Raw – white : AER = 17.9%

A chart showing a number of dots

Description automatically generatedWe can see how the error is smaller than the one evaluated by the lda, stressing again the non-normality of the data, remembering that this approach does not require normality.

The first direction is the following:

LD1

volatile -0.41146118

chlorides -0.06390936

sulphur.ratio 0.24747987

density 0.20975647

alcohol 1.32860772

almost all the score is given by the alcohol, again highlighting the fact that this feature is crucial in the quality of the wine, counterbalanced by the volatile and chlorides, overall, this is a sort of index of the fermentation because the longer is the fermentation the higher is this component. The sulphur ratio in this context represent the quality of the fermentation and how it is well behaved.

1. ­­­­TopGPT – white: AER = 18.2% the considerations on normality apply here again.

A purple and yellow dots

Description automatically generatedThe first direction is the following:

LD1

AlcoholContent 1.42895200

SulfurDioxide2ChlorideRatio 0.20855342

Chloride2Sulfate.Ratio -0.22609769

AcidityInteraction -0.45731397

AlcohlpH -0.22323491

SugarAlcohol 0.25145026

DensityClarity 0.58782429

ChloridePHRatio 0.84923728

AcidAlcoholInteraction 0.11268657

SugarChlorideInteraction -0.01521984

It is almost the same thing as before, in particular it is a balance between the alcoholic part and the acid/chloride part.

1. A purple and yellow dots

   Description automatically generatedRaw – red: AER = 11.76%.Here, having assumed the normality the error rate is higher than the one with lda, nevertheless this is a massive result consider that this error rate is obtained using only one direction.

The direction is the following:

LD1

fixed 0.37578716

volatile.acidity -0.31901157

citric.acid 0.09383091

chlorides -0.20598750

sulphur.ratio 0.23890361

density -0.30261509

sulphites 0.39789250

alcohol 0.87839495

here again alcohol is the most important, but it is less important than before, in fact in the red wines the stability and the body of the wine is fundamental, here represented by the sulphites and sulphur ratio.

1. A graph showing a number of dots

   Description automatically generatedTop – red: AER = 12.2%, in this case our variables outperforme the result of the selected variables, again the results on gaussianity apply as before.

The direction is the following:

LD1

SulfurDioxide2ChlorideRatio 0.280423205

Alcohol2Density.Ratio 0.942352680

Sulphates2pH.Ratio 0.180696303

SugarAlcohol 0.009660895

SulphurDioxideIndex 0.004853731

AcidRatio 0.527632775

CitricAcidRatio 0.046463371

AlcoholSugarSquare 0.058444728

pHSulphateInteraction 0.127633967

It is an index of fermentation, indeed all the features that represents a good fermentation are at the numerator, the opposite for a low fermentation.

**Conclusions**

The main concern of our analysis is the trade-off between interpretability and representability versus precision, in fact we present many different types of classifiers with the aim of evaluating their differences. The more the method is complicated the more it is precise, overall. But, in some cases we are able to reconstruct a precision which is acceptable with respect to the one obtained with the deep learning techniques, and there is even a clear interpretation of the result. In this scenario the interpretation of the results, together with the importance of the assumptions, which are not always met, are crucial.

The results of the analysis are that whites and reds are clearly divided, and in particular, to classify red wines is easier because of the importance of the structure of the wine, that gives us another driver to assess the quality. The common factor between them and the whites is the alcohol, which mainly represents the fermentation, and which together with the other fermentation indices gives a decent classifier.

**Criticism**

Our analysis is not able to identify the 6-rated wines, from a fermentation point of view they are very variable along all the ranges of both good and bad wines.

Regarding the bootstrap anova, the hypothesis (gaussianity and equality of covariance matrices) usually are not verified, we needed a tool to select, within a reasonable time, many variables, and so we performed this test, relying on the robustness of the anova on its hypothesis. More importantly, we set a very low threshold in order to select the variables, aiming to counterbalance the lack of these hypothesis, as we suggest doing if you were to repeat this analysis.

Another point which is important to stress is that for the raw variables with the white wines, the AER for the Fisher’s discriminant score is smaller than the one for the lda. This fact sounds a bit strange, indeed the classifier we built on the Fisher’s score is a simplified lda. Nevertheless, these are the numbers that result, we must consider the fact that the sample is not gaussian at all in this context, and this lack of assumption may affect he goodness of the classifier, but the main difference in the two method is the kind of distance used: the lda algorithm uses the Mahalanobis distance while we used in the one-dimensional scenario the Euclidean distance. The covariance matrix is scaled on the original variables, and so the differences in the variances are not so big, but the main differences in the means are on features which are not so significant for our analysis, resulting in a tiny improvement for the Euclidean distance in one dimension.