A study conducted on over 6,000 "Vinho Verde" wines, including both red and white varieties. Using their biometric measurements, we developed various classifiers providing an analysis of the strengths and weaknesses of each method. Our goal was to achieve the optimal balance between performance and interpretability.

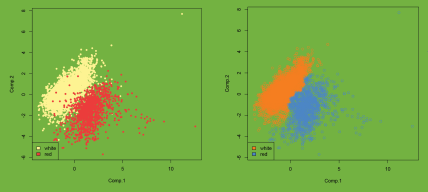
The Path of Vinho Verde:
unveiling quality classifiers

WHITE OR RED?

After performing the PCA, the results reveal that the first two components account for 50% of the variability, this is enough to identify red and white clusters.

Component 1 : type of fermentation Component 2 : quantity of fermentation

Loadings:	Comp 1	Comp 2
fixed acidity	0.387	
volatile acidity	0.293	-0.386
citric acid		0.348
residual sugar	0.138	0.477
chlorides	0.435	-0.183
sulfur ratio	0.141	-0.307
density	0.547	0.200
pH		-0.428
sulphates	0.323	-0.264
alcohol	-0.370	-0.283



Only 4.3% of units would go in the wrong cluster. The high difference in the two types highlighted by this analysis convinced us to treat red and white wines separately.

PRE-PROCESSING STEPS AND METHODS

Removing 6 - rated wines
As a pre-processing step for the classifications, we have determined the need to exclude wines that had received a rating of 6. These wines exhibit characteristics that do not clearly align with either class, both qualitatively and quantitatively.

Chat GPT variables
To gain further insights into the dataset, we sought a method to construct variables that capture relevant information. As a solution, we utilized ChatGPT to assist us in this process.We engineered the problem by dividing the research areas into three main drivers: fermentation, quality of the fermentation, and territory. Subsequently, we posed a series of questions to the artificial intelligence, requesting 15 different indices for each driver.

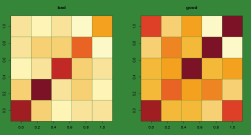
Bootstrap anova for variable selection
Due to the limitations of the ANOVA model in handling a large number of observations, we have made the decision to perform the test on bootstrapped data. Resampling the original dataset to create multiple smaller datasets, allowed us to obtain more reliable and robust statistical results.

Cost of misclassification
As customers ourselves, we consider the cost of purchasing a misclassified bad wine to be twice as high as the cost of not buying a misclassified good wine.

LDA AND QDA

- Prior result in a probability distribution of 2/3 for a good wine and 1/3 for a bad wine;
- In QDA normality assumption are never fully met, the distribution is set as the Student's t-distribution;
- The confusion matrix represents the mean of all the confusion matrices obtained along the cross validation process.

White-raw:
>Some semblance of normality is reconstructed



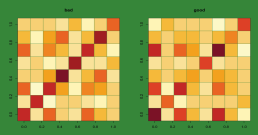
LDA > AER = 22.37%

	true-bad	true-good
predicted-bad	78.93333	30.40000
predicted-good	9.73333	60.93333

QDA > AER = 20.1%

	true-bad	true-good
predicted-bad	83.93333	25.40000
predicted-good	9.93333	60.73333

Red-raw:
>Normality cannot be rejected for almost all covariates



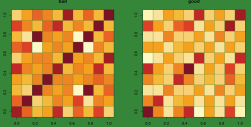
LDA > AER = 14.05%

	true-bad	true-good
predicted-bad	42.20000	7.40000
predicted-good	1.60000	12.86667

QDA > AER = 14.78%

	true-bad	true-good
predicted-bad	41.46667	8.13333
predicted-good	1.33333	13.13333

White-GPT:
>Normality is not verified



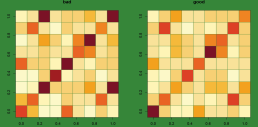
LDA > AER = 20.78%

	true-bad	true-good
predicted-bad	81.33333	28.00000
predicted-good	9.40000	61.26667

QDA > AER = 18.74%

	true-bad	true-good
predicted-bad	86.33333	23.00000
predicted-good	10.73333	59.93333

Red-GPT:
>Normality cannot be rejected for almost all covariates



LDA > AER = 13.5%

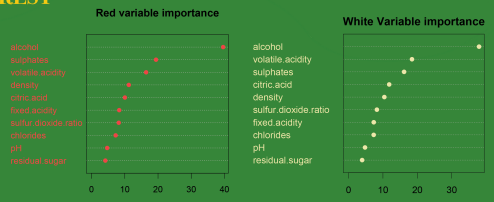
	true-bad	true-good
predicted-bad	42.40000	7.20000
predicted-good	1.46667	13.00000

QDA > AER = 12.6%

	true-bad	true-good
predicted-bad	43.73333	5.86667
predicted-good	2.20000	12.26667

RANDOM FOREST

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



KNN-CLASSIFIER

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

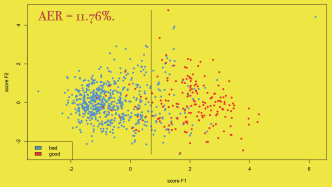
Scaling variables
To ensure uniformity of judgment during the analysis, we recognized that the variables presented significantly different scales. As a result, we made the decision to scale the variables before conducting further analysis.

*For this project, the AER has been estimated using a 15-fold cross-validation approach

FISHER ARGUMENT

Raw-white:	LD1
volatile acidity	-0.4114118
chlorides	-0.06390936
sulphur ratio	0.24747987
density	0.20975647
alcohol	1.32860772

Raw-Red:	LD1
fixed acidity	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



CONCLUSIONS

- The red and white wines are distinctly separated.
- Fermentation plays a pivotal role in differentiating and categorizing the wines effectively.
- Classify red wines is easier because of the importance of the structure of the wine.
- Generally, more complex methods tend to yield higher precision; however, in some cases, we are able to achieve an acceptable level of precision comparable to deep learning techniques, while also providing a clear interpretation of the results.

CRITICISM

- Our analysis is not able to identify the 6-rated wines.
- Regarding the bootstrap anova, usually the hypothesis are not verified.
- Costs of misclassification are easily handled only on LDA & QDA methods.

GROUP #29

- Anna Spelta
- Francesco Maria Mancinelli
- Leonardo Marchesin
- Matteo Saterini