Modeling wine preferences from physicochemical properties.

Introduction

- Research question: the research is aimed at discovering what are most important features that determinate the wine quality.
- The purpose of the study is to identify the best predictors of quality from multiple related factors.
- Since consuming alcohol is one of the biggest hobby all around the world, we believe that, in order to create an healthy relationship with alcohol consumption, the biggest difference between different wines is the quality.
- During the pandemic period a lot of people started to drink alcoholic drinks in an inappropriate way and this research could be a method to remodulate alcohol consumption.

Methods

The dataset is related to red and white variants of the Portuguese "Vinho Verde" wine.

Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

The sample included 1599 observations .

Measures

For each observation there are 12 explanatory variables which 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, which 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

Measures

- **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

Measures

- **density**

the density of water is close to that of water depending on the percent alcohol and sugar content

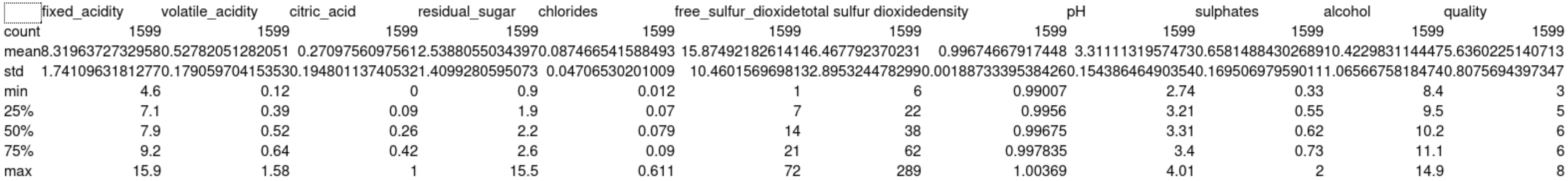
- **pH**

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

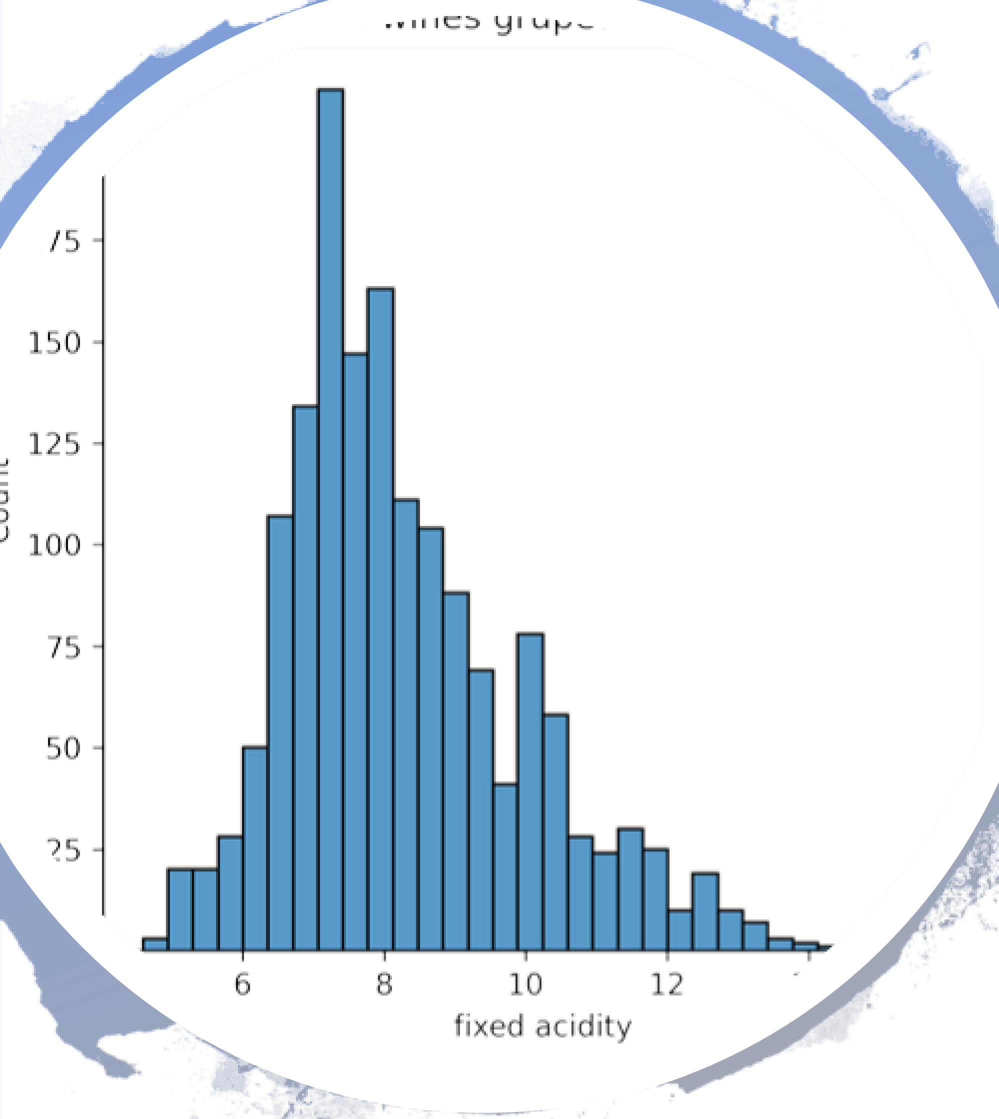
- **sulphates**

a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

Descriptive statistic



	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599
mean	8.3196372732958	0.52782051282051	0.2709756097561	2.5388055034397	0.087466541588493	15.874921826141	46.467792370231	0.99674667917448	3.3111131957473	0.65814884302689	10.422983114447	5.6360225140713
std	1.7410963181277	0.17905970415353	0.19480113740532	1.4099280595073	0.04706530201009	10.46015696981	32.895324478299	0.0018873339538426	0.15438646490354	0.16950697959011	1.0656675818474	0.8075694397347
min	4.6	0.12	0	0.9	0.012	1	6	0.99007	2.74	0.33	8.4	3
25%	7.1	0.39	0.09	1.9	0.07	7	22	0.9956	3.21	0.55	9.5	5
50%	7.9	0.52	0.26	2.2	0.079	14	38	0.99675	3.31	0.62	10.2	6
75%	9.2	0.64	0.42	2.6	0.09	21	62	0.997835	3.4	0.73	11.1	6
max	15.9	1.58	1	15.5	0.611	72	289	1.00369	4.01	2	14.9	8



Analysis

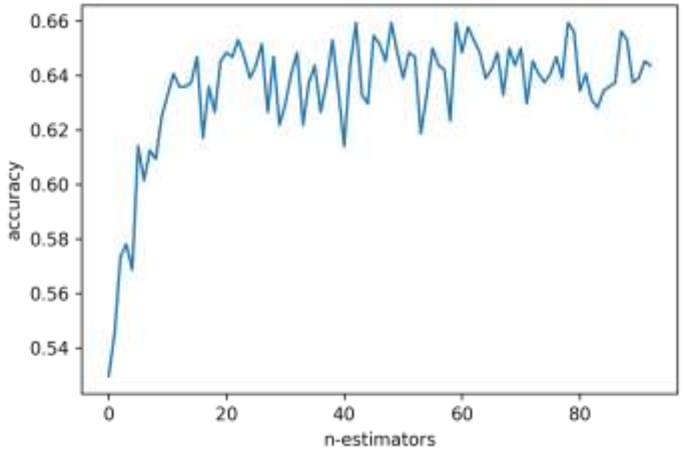
The dataset does not require data management or many transformations since it was appropriately chosen on Kaggle with a particular attention for his structure.

- **Descriptive Statistics**

The distributions for predictors and response variable were evaluated by calculating the mean, standard deviation, minimum, maximum and quantile since they are all quantitative variables.

With a simple graph is simple to understand that no variable is normally distributed which is confirmed by a shapiro-test that presents always a p-value< 0.000001.

Analysis



- **Bivariate analysis**

Bivariate analysis does not provide significant insight since the response acquires values from 3 to 8 .

- **Multivariable Analysis**

I performed a Regression model in which just few variables are included since there is a strong collinearity which is difficult to

indagate.

Alcohol, pH, chlorides, sulphates and volatile acidity seems to be appropriate to describe the response since the p-value is smaller than 0.0001 with coefficients equal to

0.3055 , -0.4173, -1.9288, 0.8474 and -1-.0687.I performed a decision tree regressor with a prediction score of 0.5921 and a Random Forest with a prediction score of 0.5937.

I figured out the best number of estimators =93 which seems to be the best by a visual control

Lasso Regression

At the end I performed a Lasso Regression trying to understand which are the best variables to predict the quality and the resualts showed that:

- alcohol: 0.2919347010934089,
- Fixed acidity: 0.0,
- residual sugar: 0.0,
- free sulfur dioxide: -0.003858720666430753,
- citric acid: 0.0,
- density: 0.0,
- pH: -0.043809882850588996,
- chlorides: -1.437571541957534,
- sulphates: 0.7017492821544051,
- volatile_acidity: -1.0999820265550029

Conclusions

Lasso Regression indicated that 8 of 12 predictor variable were selected.

Alcohol, chlorides, sulphates, volatile acidity have an important role , alcohol and sulphates increasing the quality and chlorides and volatile acidity decreasing the quality which seems pretty reasonable.

It is surprising that residual sugar and density do not contribute to the quality level.

Conclusion

The project successfully developed a predictive model algorithm that provides indications on the quality. However there are some limitations that should be taken into account:First, we analyzed only data from red and white variants of the Portuguese "Vinho Verde" wine but there are a lot of different typology of wine and different factors may emerge as predictors of quality.

Limitations

Therefore we cannot assume that the predictive model developed in this brief study will be valid or useful for predicting quality for all wine typologies .Finally there is a large number of chemical properties that could impact wine quality but the project examined just a few: it is possible that the factors identified as important predictors of quality among the set of predictors analyzed in this project are confounded by other factors not considered in this analysis.Future efforts to develop a solid predictive model algorithm for wine quality should expand the algorithm by adding different types of wine and characteristics .