

Predicting Wine Quality




Predicting wine quality by their physicochemical properties

Created by: Alon Parag, 20.05.2021 for
BlueBerry Wineries

Data by: Modeling wine preferences by data mining from physicochemical properties by Cortez et al.(2009), wine review data (CODE Analytics), date unknown.

Is it possible to predict the quality of wine?



- **Discussed matters:**

- Do wines of different quality have statistically significant difference in physicochemical features?
- Could a predictive ML model have better accuracy in predicting wine quality than a random guess?
- Is there a significant price difference between quality labels?

- **Methods:**


- Explorative data analysis
- Comparative prediction
- Model optimization

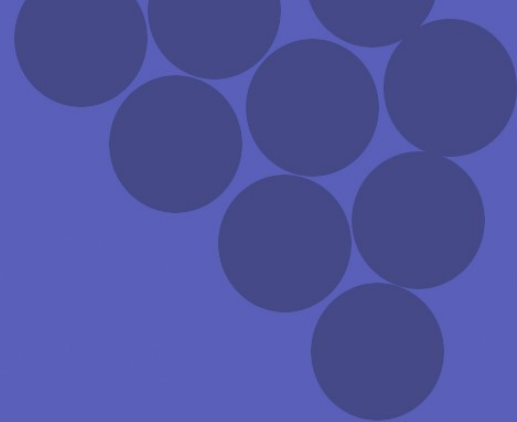
Data:

- Data about physicochemical properties is sourced from a study carried by Cortez et al (2009).
- The sample includes data of 4898 white wines and 1599 red wines.
- Sample contains physicochemical data and quality grade given by wine experts.
- Data about relation between price and quality is sourced from CODE analytics, it contains 146 relevant entries about Vinho Verde wines.

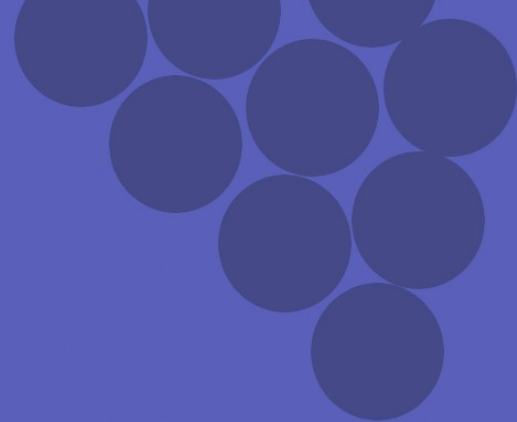
Details about data:

- **Fixed acidity**: acids are major wine properties and contribute greatly to the wine's taste. Usually, the total acidity is divided into two groups: the volatile acids and the nonvolatile or fixed acids. Among the fixed acids that you can find in wines are the following: tartaric, malic, citric, and succinic. This variable is expressed in g(tartaricacid)/l in the data sets.
- **Volatile acidity**: the volatile acidity is basically the process of wine turning into vinegar. In the U.S, the legal limits of Volatile Acidity are 1.2 g/L for red table wine and 1.1 g/L for white table wine. In these data sets, the volatile acidity is expressed in g(aceticacid)/l. Excessive volatile acidity may result in lower wine quality
- **Citric acid**: one of the fixed acids that you'll find in wines. It's expressed in g/l in the two data sets.
- **Residual sugar**: typically refers to the sugar remaining after fermentation stops, or is stopped. It's expressed in g/l in the red and white data.

- 
- **Chlorides**: Chlorides can be a significant contributor to saltiness in wine. Here, you'll see that it's expressed in g(sodiumchloride)/l.
 - **Free sulfur dioxide**: the part of the sulfur dioxide that is added to a wine and that is lost into it is said to be bound, while the active part is said to be free. The winemaker will always try to get the highest proportion of free sulfur to bind. This variable is expressed in mg/l in the data.
 - **Total sulfur dioxide**: the sum of the bound and the free sulfur dioxide (SO₂). Here, it's expressed in mg/l. There are legal limits for sulfur levels in wines: in the EU, red wines can only have 160mg/L, while white and rose wines can have about 210mg/L. Sweet wines are allowed to have 400mg/L. For the US, the legal limits are set at 350mg/L, and for Australia, this is 250mg/L.



- **Density**: used as a measure of the conversion of sugar to alcohol. Here, it's expressed in g/ml.
- **pH** or the potential of hydrogen is a numeric scale to specify the acidity or basicity the wine. As you might know, solutions with a pH less than 7 are acidic, while solutions with a pH greater than 7 are basic. With a pH of 7, pure water is neutral. Most wines have a pH between 2.9 and 3.9 and are therefore acidic.
- **Sulfates**: regular part of winemaking around the world and are considered necessary.
In this case, they are expressed in $\frac{\text{g}(\text{potassiumsulphate})}{\text{l}}$.

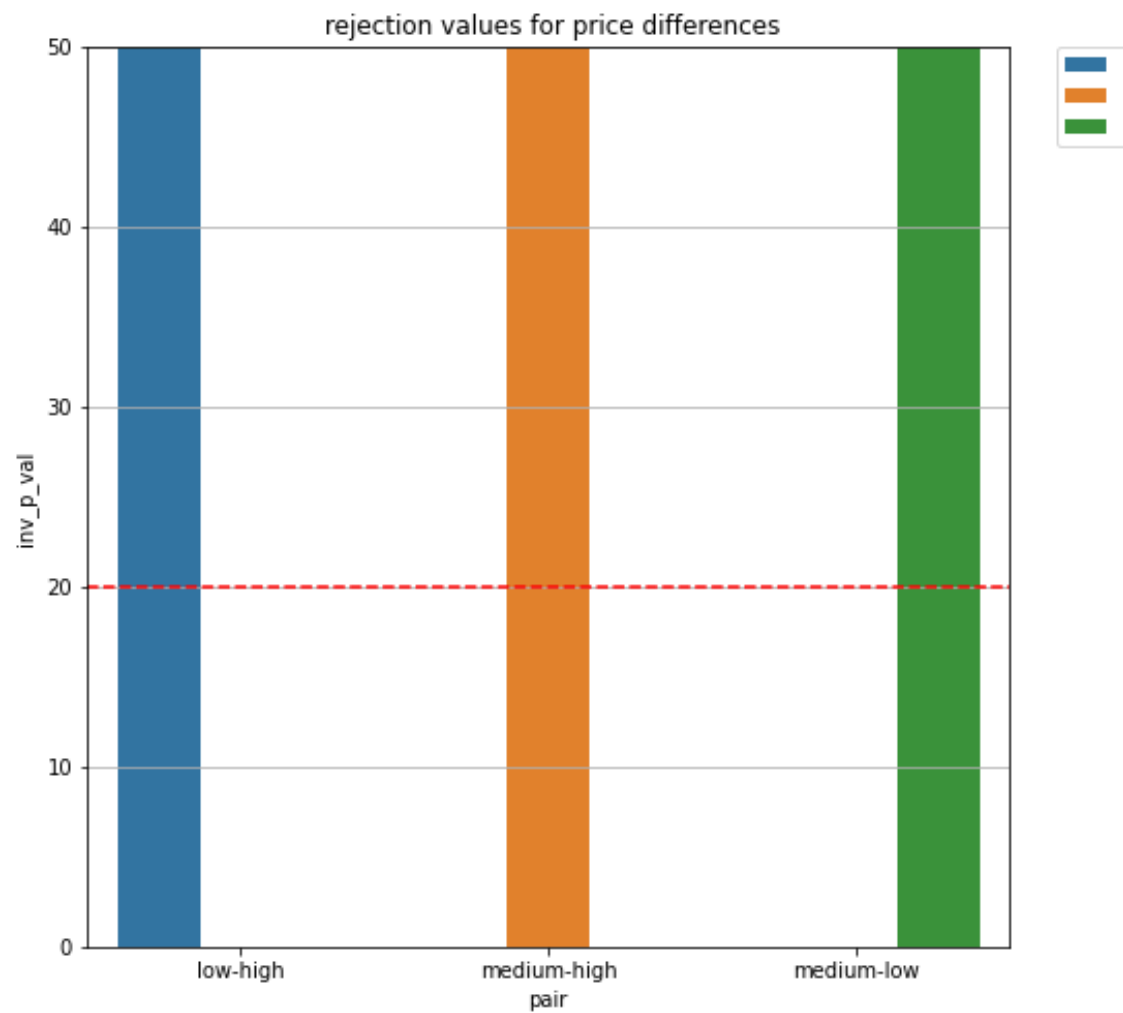


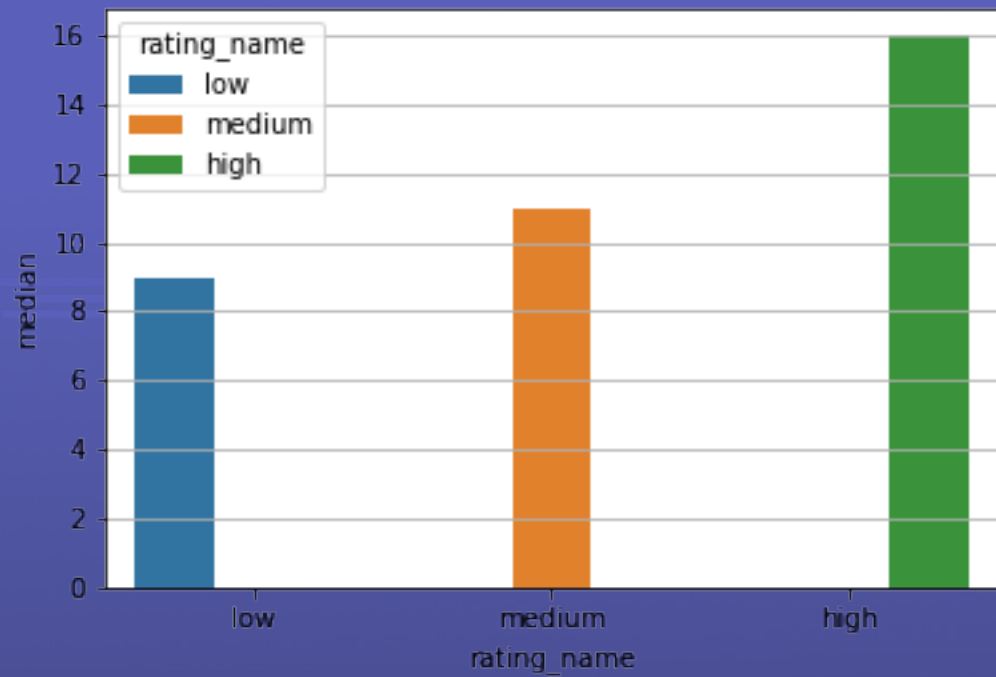
- **Alcohol**: wine is an alcoholic beverage, and as you know, the percentage of alcohol can vary from wine to wine. It shouldn't be surprised that this variable is included in the data sets, where it's expressed in % vol. Concentration above 15% halts the fermentation process of wine
- **Quality**: wine experts graded the wine quality between 0 (very bad) and 10 (very excellent). The eventual number is the median of at least three evaluations made by those same wine experts.

Insights from data of price and quality

- **Summary:**

- Different quality labels have a statistically significant difference in price
- Low quality wine has a median price of 9eur.
- Medium quality wine has a median price of 11eur.
- High quality wine has a median price of 16eur.





Analysis of physicochemical properties



Steps Involved: Preparation of data

- Since most of the samples (3915 wines out of 6497 has a quality score of 6 or 7, I added another property named “Quality label”, with the categories “low”(wine rated under 6), “medium”(wine rated 6\7), and “high”(wine rated above 7) due to the abovementioned imbalanced distribution of data, and because predictive models perform better with lower number of categories.
- White wines and red wines have slightly different physicochemical features, therefore I analyzed and modeled them separately
- Data was tested with statistical methods (Pairwise tukey’s method HSD) to check for significant difference between quality label and feasibility of prediction.

Insights summary:

- **Red Wines:**

- except for pH and residual sugars, all other features have statistically significant difference between the 3 quality wine_labels
- The following features have approximately normal distribution of values: fixed acidity, density, ph.
- the following feature are sqewed towards higher values: residual sugars, chlorides, free sulfur dioxide, total sulfur dioxide, sulphates, alcohol
- the distribution of citric acid values is tri-modal and sqewed to higher values.

- **White Wines:**

- except for free sulfur dioxide and citric acid, all other features have statistically significant difference between at least one of the 3 quality wine_labels
- The following features have approximately normal distribution of values: fixed acidity, chlorides ph.
- the following feature are sqewed towards higher values: volatile acidity, residual sugars, free sulfur dioxide, total sulfur dioxide, density, sulphates, alcohol

Predictive Model

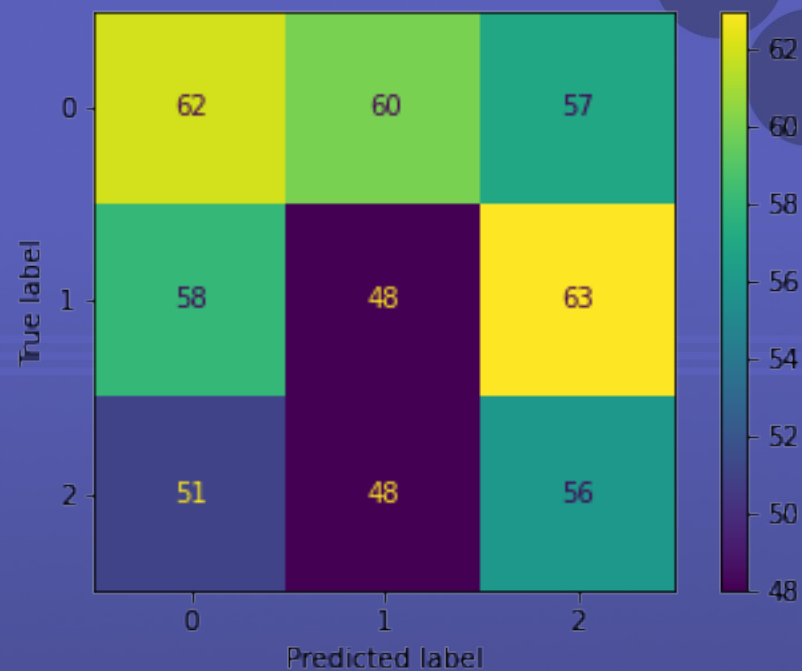
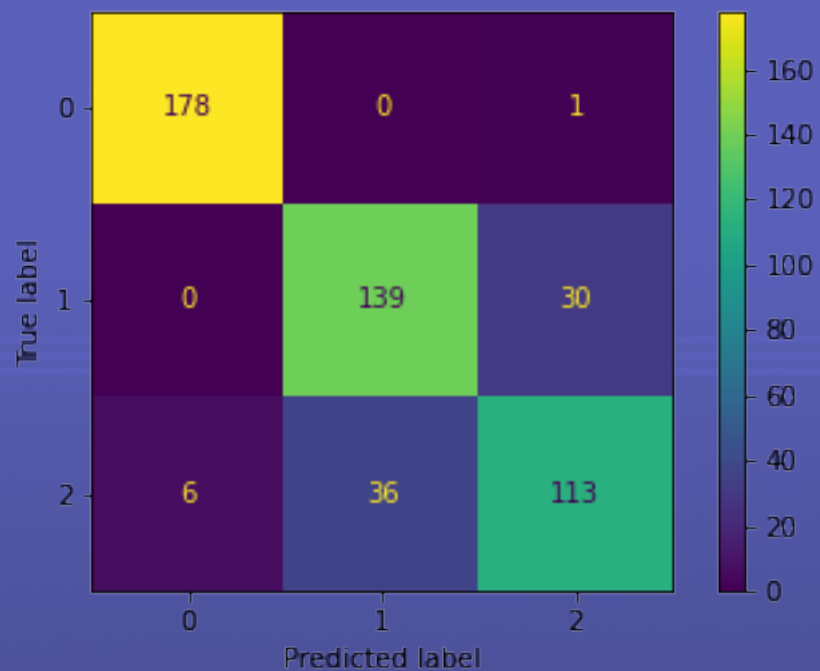
- **Steps Involved:**

- Different models were compared using different sampling methods and scaling methods, in order to address acute data imbalance between quality labels. The goal was to select a method with low bias towards each quality label.
- Once a model was selected, it's parameters were optimized to maximize its accuracy and minimize bias towards the different quality labels.
- Finally, the models predictions were compared to predictions made by a dummy model.

Red Wines prediction



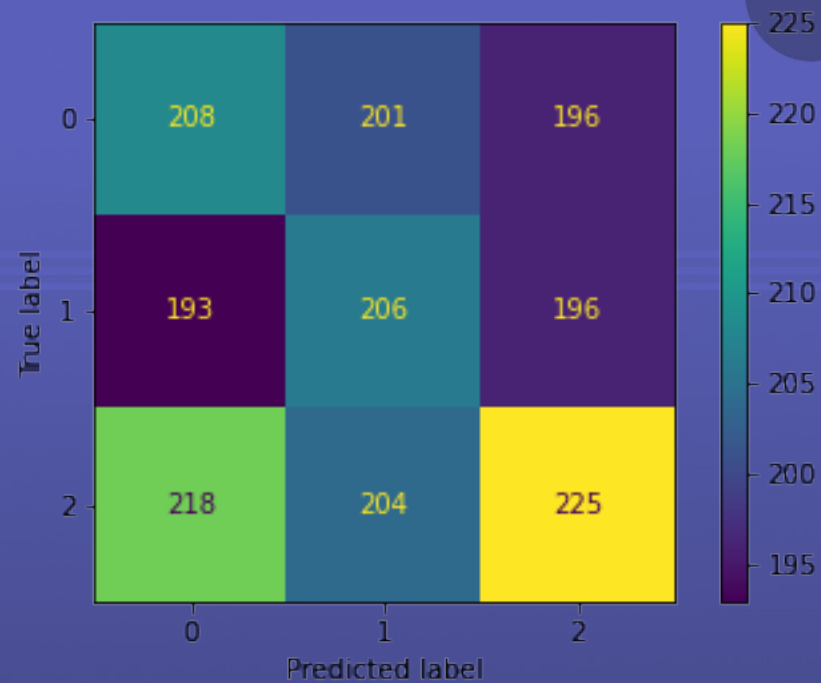
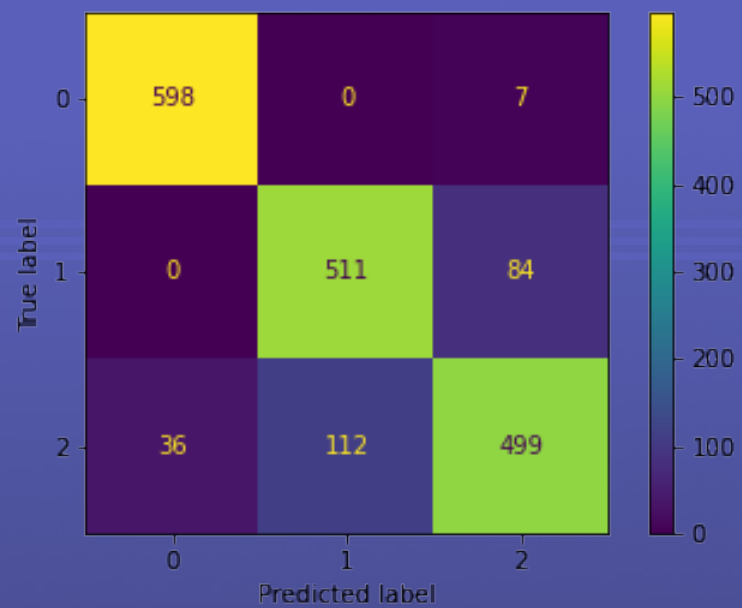
- Accuracy: 85%
- Label legend:
 - 1 – Low Quality
 - 2 – Medium Quality
 - 0 – High Quality



White Wines prediction



- Accuracy: 87%
- Label legend:
 - 1 – Low Quality
 - 2 – Medium Quality
 - 0 – High Quality



Recommendations

- More physicochemical data is needed for wine with quality rating 8 or above
- Data about how many Vinho Verde bottles were sold with information about the sale date, price and quality\rating of the wine is needed in order to gain information about the expected sales and revenue. This information is a crucial piece of the puzzle.
- An analysis of sales with regard to quality label is needed in order to optimize wine production goals

Thank you for listening!

