

# 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:

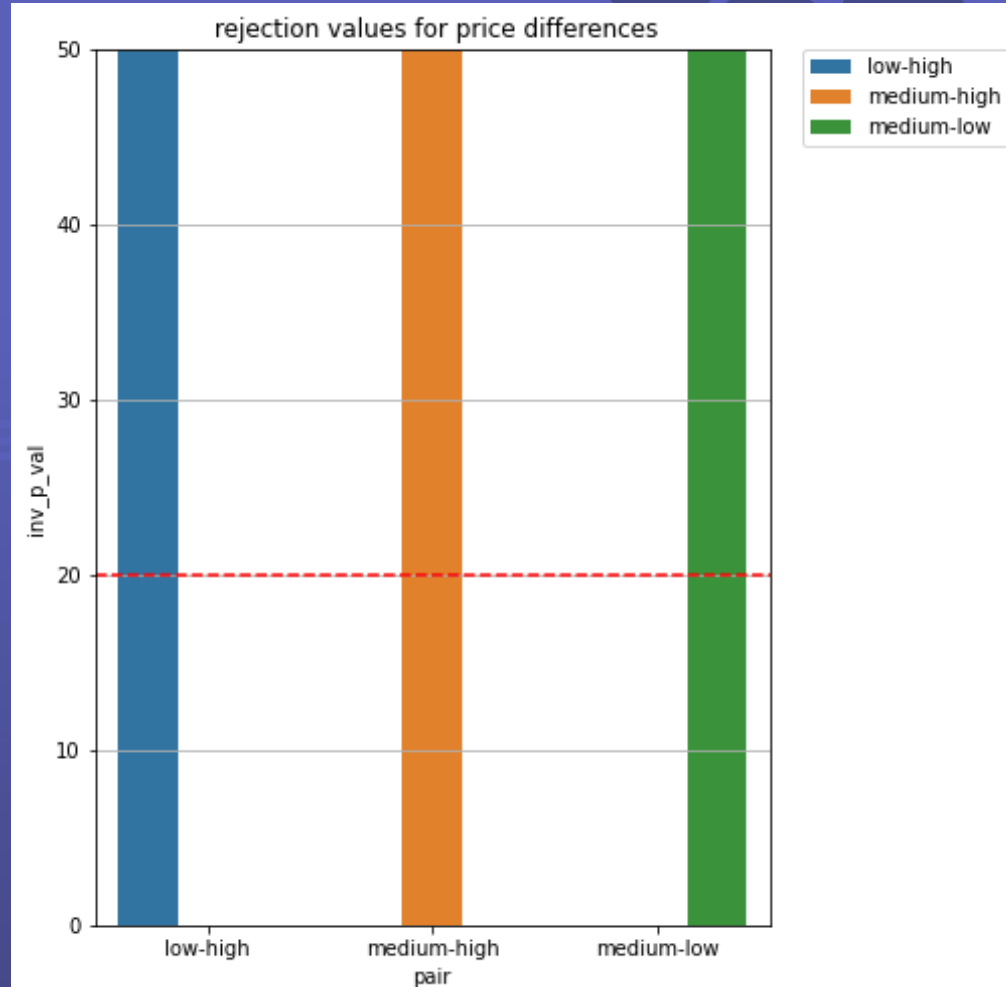
- Data about physicochemical properties is sourced from a study carried by Cortez et al (2009).
- The sample includes data of 4898 white wines and 1599 red wines.
- Sample contains physicochemical data and quality grade given by wine experts.
- Data about relation between price and quality is sourced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

# Key Insights

- 1) Different quality labels have a statistically significant difference in price
- 2) Low quality wine has a median price of 9eur.
- 3) Medium quality wine has a median price of 11eur.
- 4) High quality wine has a median price of 16eur
- 5) All red wine physicochemical features ave statistically significant difference between the 3 quality labels except for pH and residual sugars which have no signifcant difference between the label.
- 6) All white wine physicochemical features have statistically significant difference between at least one of the 3 quality wine\_labels except for free sulfur dioxide and citric acid which have no significant difference between the labels.
- 7) The quality label of red wine could be predicted with 85% accuracy
- 8) The quality label of white wine could be predicted with 87% accuracy

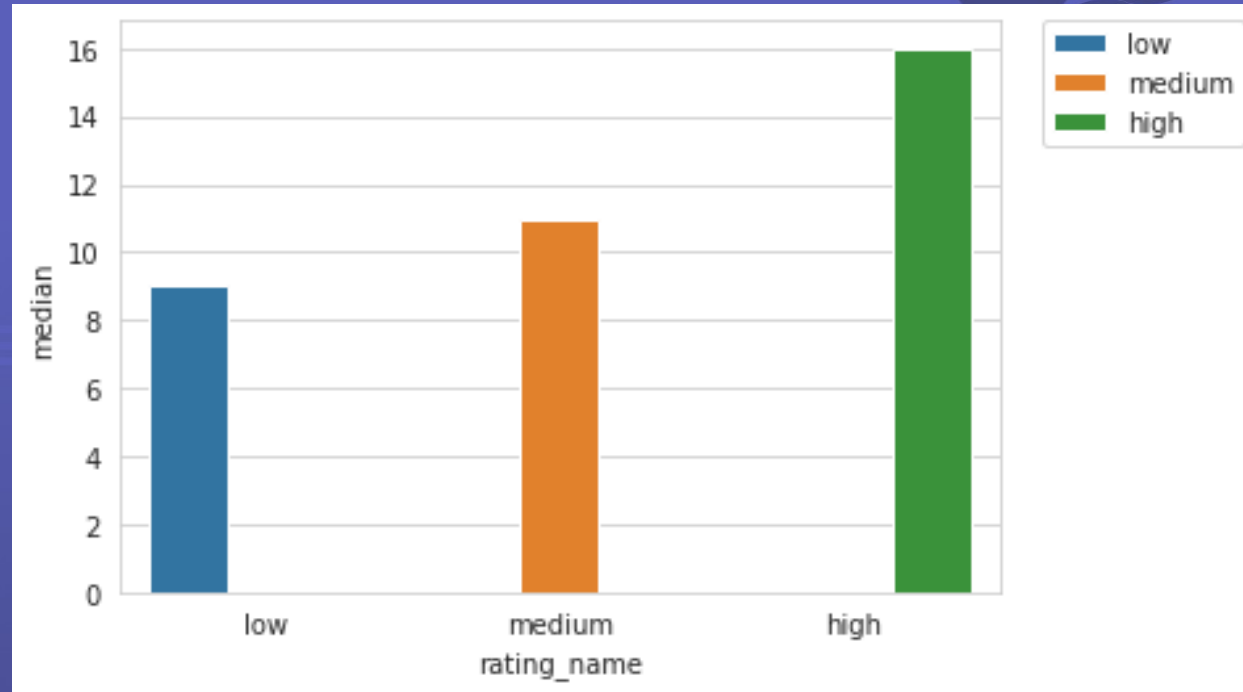
## #1 Statistics Significance in different quality labels

- 1) Different quality labels have a statistically significant difference in price



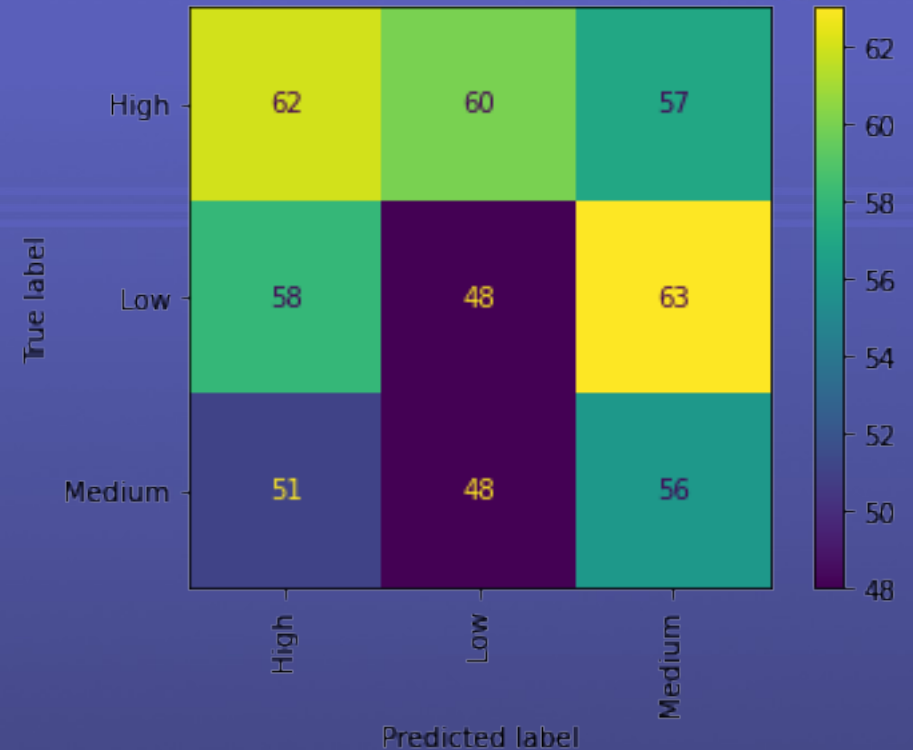
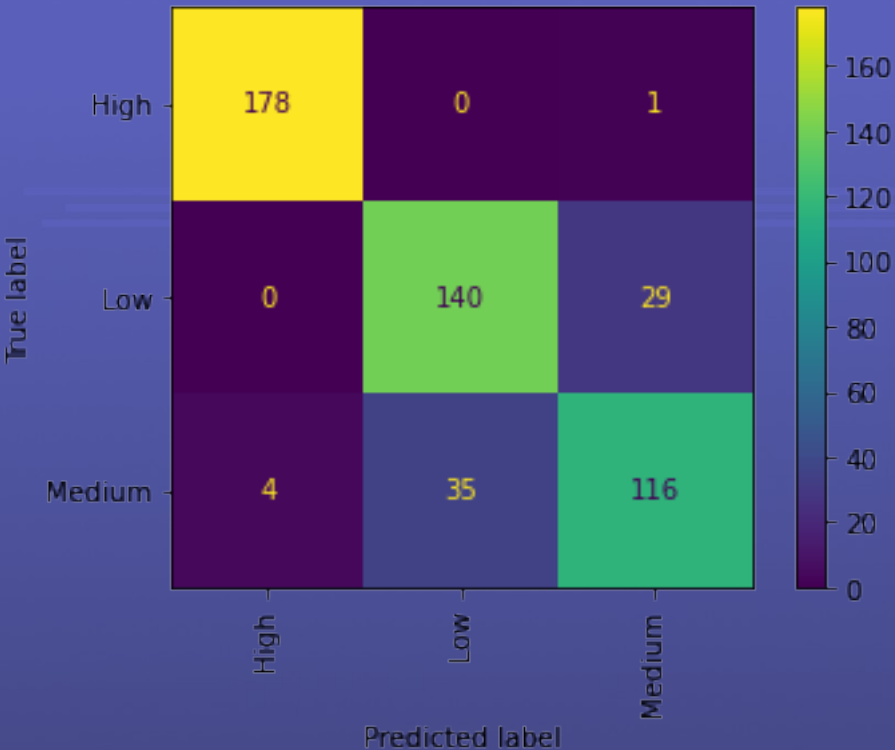
## #2 Median Price of quality labels

The median statistic was chosen as it is more robust against outliers in wine price than average.



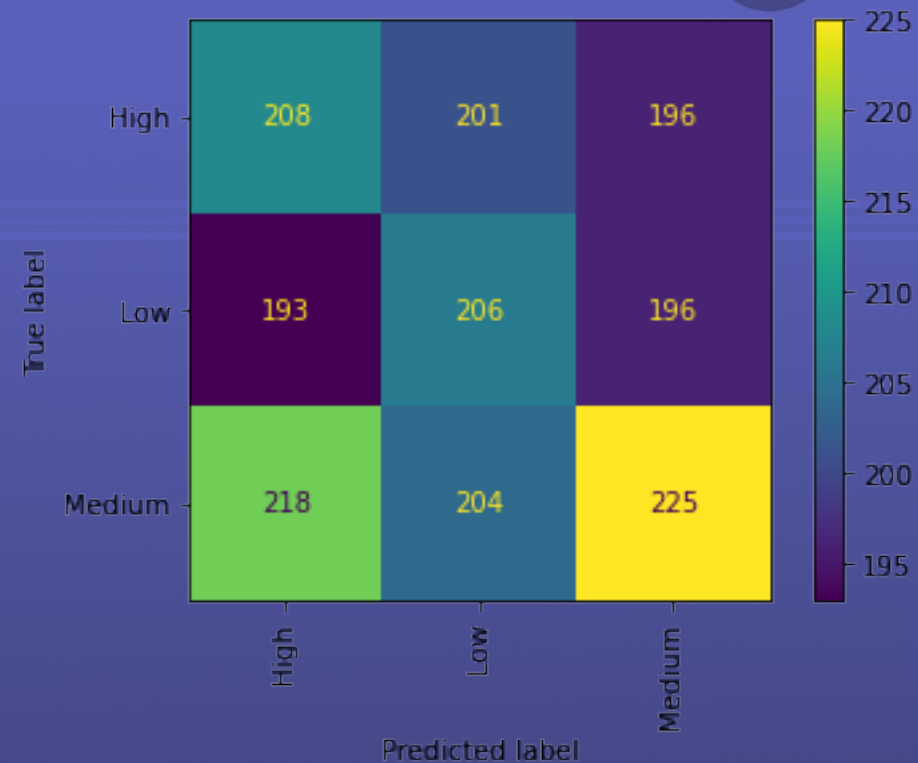
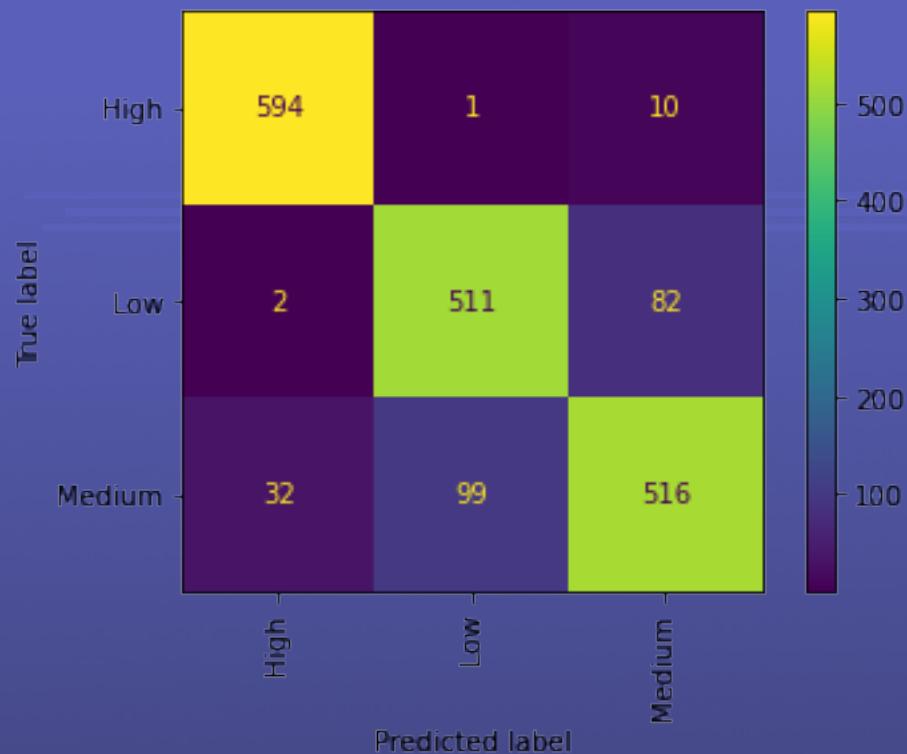
# Model results for red wine

**Model accuracy:** 86% vs 33% generic **CKS** ~ 81%



# Model results for white wine

Model accuracy: 88% vs 35% generic CKS ~ 83%





# 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.

# 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.

# 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!

