Analysis and prediction of wine quality

of the “Vinho Verde” wines in Portugal

by Boris Y. Nedyalkov
11/05/2023

Based on data by Cortez et al., 2009

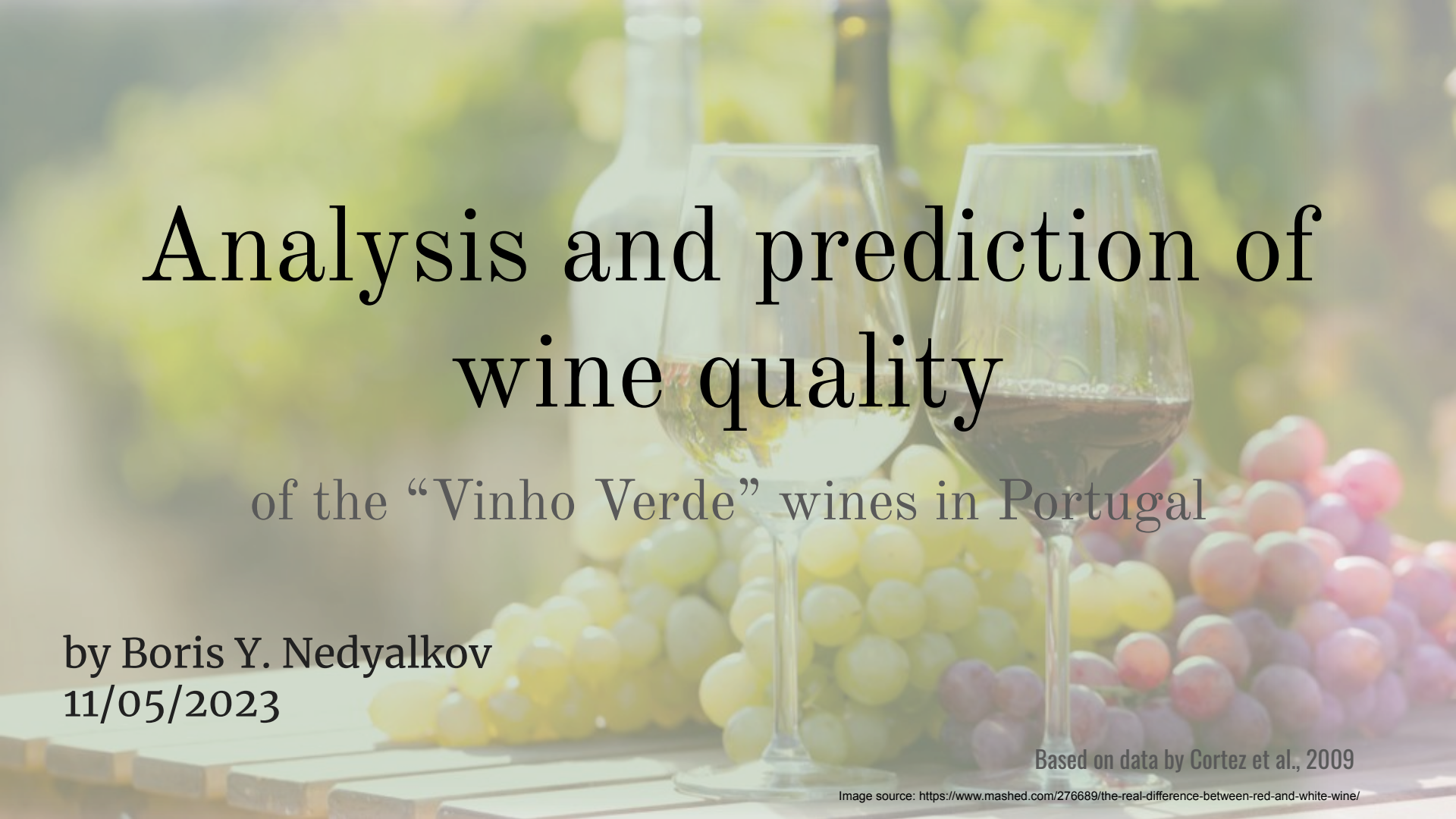
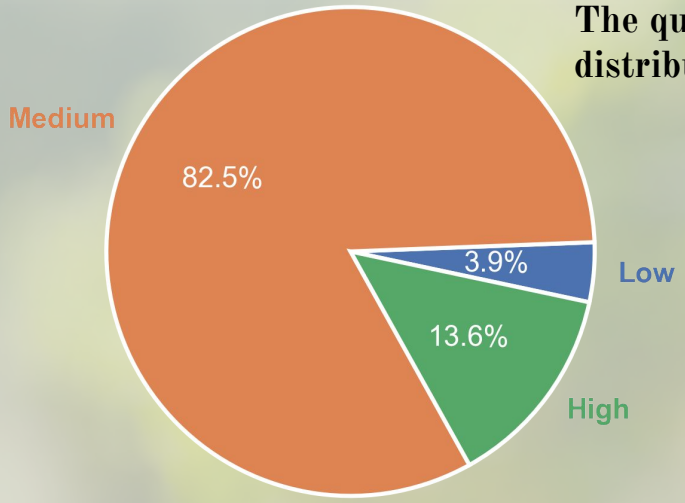
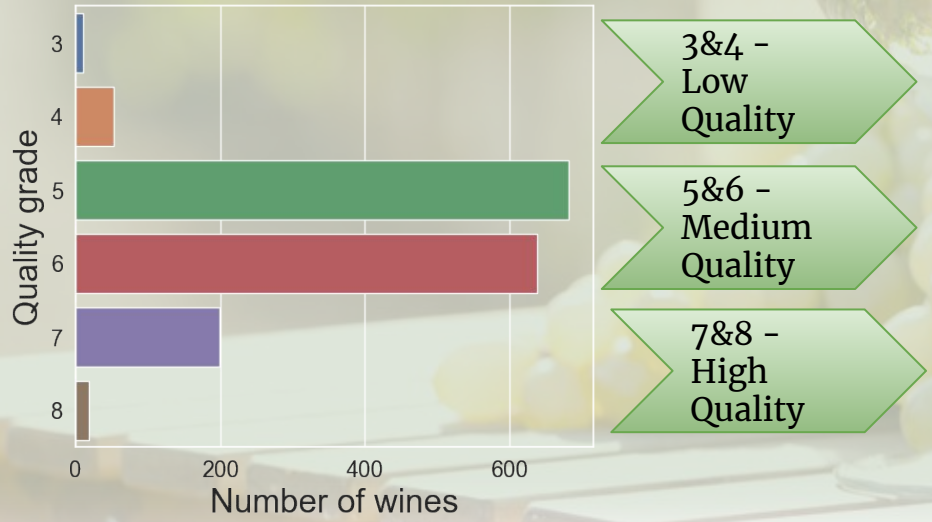


Image source: https://www.mashed.com/276689/the-real-difference-between-red-and-white-wine/

Analysis

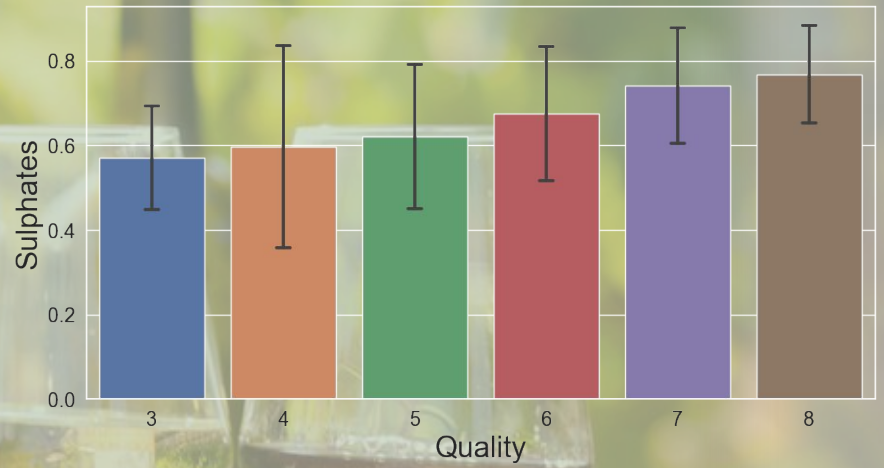




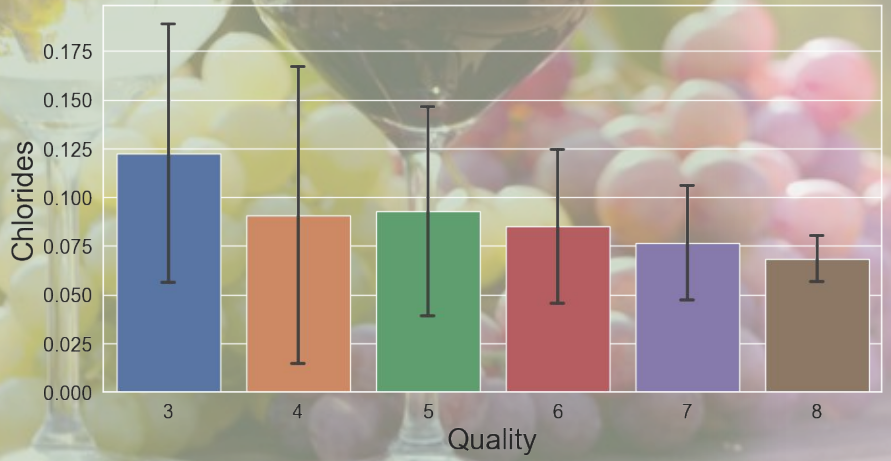
- The Vinho Verde wines are DOC wines native to the old Minho province of Portugal
- It is characterized by slight carbonation and is consumed normally not long after bottling
- 1599 different wines of the label were recorded, produced b/w 1998 and 2007
- The records contain:
  - The chemical parameters of each wine (acidity, sugars, alcohol, etc.)
  - A mark of the quality of the wine (being graded on a scale of 3 to 8)

Analysis

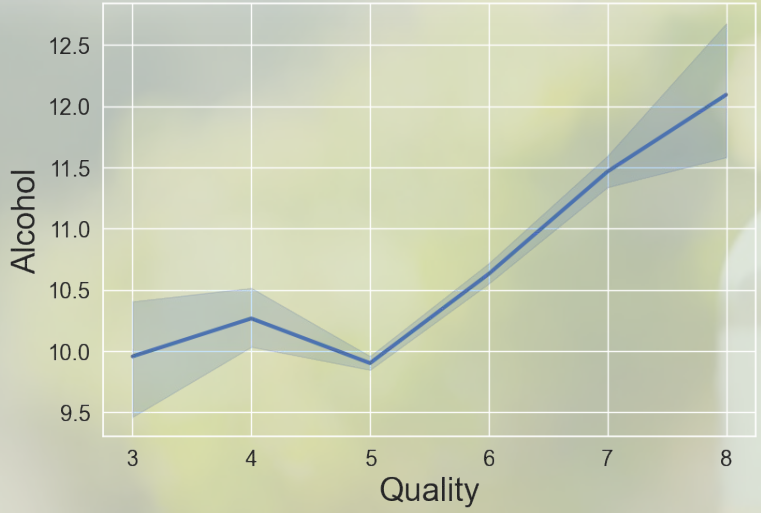
- **Sulphates:**
  - seems to on average, linearly influence the quality of the wine
  - increase in sulphates -> increase in quality

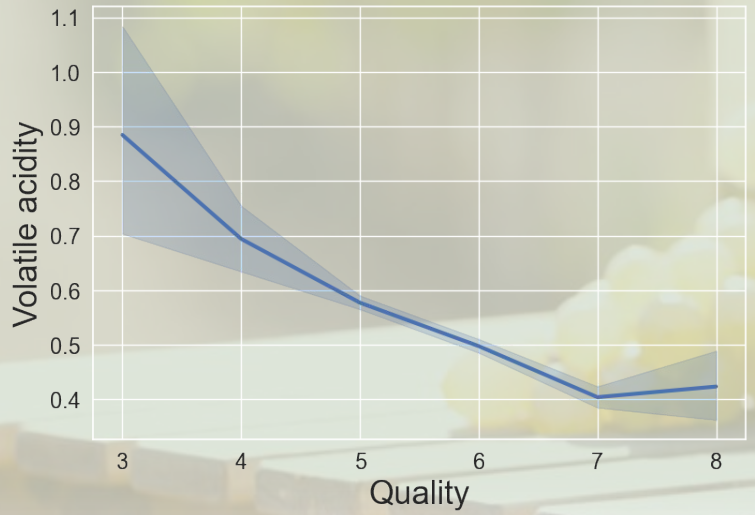


- **Chlorides:**
  - seem to have an inverse linear relationship with quality on average
  - decrease in chlorides -> increase in quality
  - note that the variance of the chloride saturation decreases with increase of quality



Analysis





- **Alcohol:**
  - There is a clear increase in quality with the increase of alcohol
  - Especially for the medium and high end wines

- **Volatile acidity:**
  - A decrease in volatile acidity can be seen with the increased quality
  - This is mostly true for the low and medium end wines

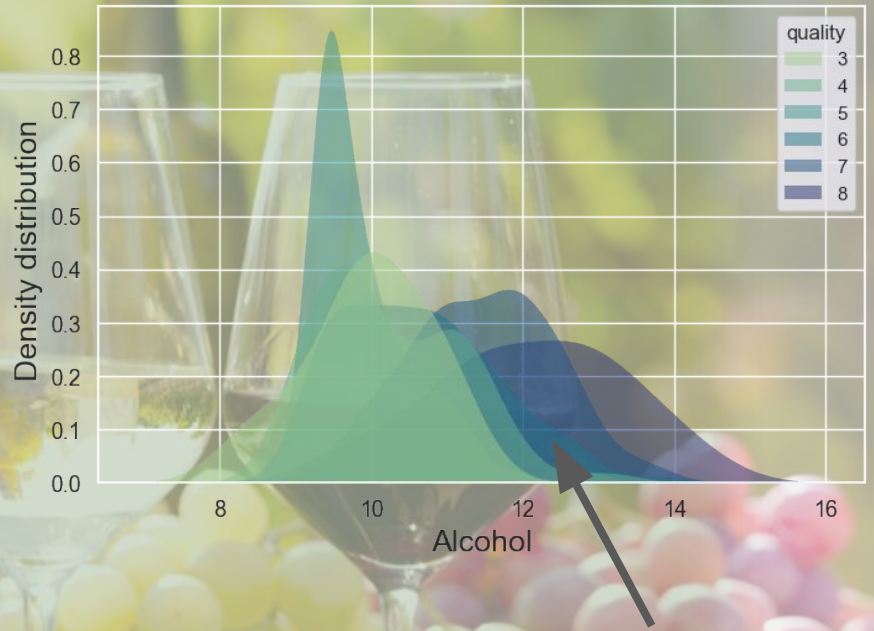
The Machine Learning Model

- Big fraction of the high and low quality wines are chemically quite similar
- Thus none of the macro factors like:
  - density,
  - alcohol,
  - acidity, etc.

,could be used to conclusively determine whether the wine is of high quality or not

- Other parameters, like:
  - aroma
  - different kinds of minerals
  - different kinds of sulphates

,in the wine may play big role as well



★ Big overlap of the wine features.

★ Irrespective of quality.

★ => Difficult to predict dependencies

The Machine Learning Models

- The ML model is trained to differentiate 2 kinds of wine
  - Of high quality (grades 7 & 8) - ‘1’
  - Of lower quality (grades 3,4,5 & 6) - ‘0’

- The best ML models generated were by optimized SVM (Support vector machine) algorithms

- Because of the high overlap of the data two different models for identifying high quality wines were trained:
  - Model 1: High Precision model - Precision = 83%
  - Model 2: Low Recall model - Recall = 77%