Data Set Information

This data set contains records related to red and white variants of the Portuguese Vinho Verde wine. It contains information from 1599 red wine samples and 4898 white wine samples. Input variables in the data set consist of the type of wine (either red or white wine) and metrics from objective tests (e.g. acidity levels, PH values, ABV, etc.), while the target/output variable is a numerical score based on sensory data—median of at least 3 evaluations made by wine experts. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). Due to privacy and logistic issues, there is no data about grape types, wine brand, and wine selling price.

This data set is a combined version of the two separate files (distinct red and white wine data sets) originally shared in the UCI Machine Learning Repository.

The following are some existing data sets on Kaggle from the same source (with notable differences from this data set):

Red Wine Quality (contains red wine data only)

Wine Quality (combination of red and white wine data but with some values randomly removed)

Wine Quality (red and white wine data not combined)

Contents

Input variables:

1 - type of wine: type of wine (categorical: 'red', 'white')

(continuous variables based on physicochemical tests)

2 - fixed acidity: The acids that naturally occur in the grapes used to ferment the wine and carry over into the wine. They mostly consist of tartaric, malic, citric or succinic acid that mostly originate from the grapes used to ferment the wine. They also do not evaporate easily. (g / dm^3)

3 - volatile acidity: Acids that evaporate at low temperatures—mainly acetic acid which can lead to an unpleasant, vinegar-like taste at very high levels. (g / dm^3)

4 - citric acid: Citric acid is used as an acid supplement which boosts the acidity of the wine. It's typically found in small quantities and can add 'freshness' and flavor to wines. (g / dm^3)

5 - residual sugar: The amount of sugar remaining after fermentation stops. It's rare to find wines with less than 1 gram/liter. Wines residual sugar level greater than 45 grams/liter are considered sweet. On the other end of the spectrum, a wine that does not taste sweet is considered as dry. (g / dm^3)

6 - chlorides: The amount of chloride salts (sodium chloride) present in the wine. (g / dm^3)

7 - 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. All else constant, the higher the free sulfur dioxide content, the stronger the preservative effect. (mg / dm^3)

8 - total sulfur dioxide: The 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. (mg / dm^3)

9 - density: The density of wine juice depending on the percent alcohol and sugar content; it's typically similar but higher than that of water (wine is 'thicker'). (g / cm^3)

10 - pH: A measure of the acidity of wine; most wines are between 3-4 on the pH scale. The lower the pH, the more acidic the wine is; the higher the pH, the less acidic the wine. (The pH scale technically is a logarithmic scale that measures the concentration of free hydrogen ions floating around in your wine. Each point of the pH scale is a factor of 10. This means a wine with a pH of 3 is 10 times more acidic than a wine with a pH of 4)

11 - sulphates: Amount of potassium sulphate as a wine additive which can contribute to sulfur dioxide gas (S02) levels; it acts as an antimicrobial and antioxidant agent.(g / dm3)

12 - alcohol: How much alcohol is contained in a given volume of wine (ABV). Wine generally contains between 5–15% of alcohols. (% by volume)

Output variable:

13 - quality: score between 0 (very bad) and 10 (very excellent) by wine experts

Acknowledgements

Source: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Data credit goes to UCI. Visit their website to access the original data set directly: https://archive.ics.uci.edu/ml/datasets/wine+quality

Context

So much about wine making remains elusive—taste is very subjective, making it extremely challenging to predict exactly how consumers will react to a certain bottle of wine. There is no doubt that winemakers, connoisseurs, and scientists have greatly contributed their expertise to the process, but there is still more to be discovered about the art and science of winemaking. Use this data to gain a better understanding of what makes a good quality or bad quality wine according to wine experts' taste-buds and brains.

The data set can be used to solve classification or regression tasks. Note that the classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Moreover, some input variables may or may not be relevant—it could be interesting to test feature selection methods.