In Vino Veritas

GA Data Science Final Project

3/2/2016

Wen Lu

Data Source: UCI Machine Learning Repository

Wine Quality Data Set
http://mlr.cs.umass.edu/ml/datasets/Wine+Quality

Two datasets: White Wine & Red Wine
       - 12 attributes
       - no missing values
       - sample size >1k

How to discern wine quality?

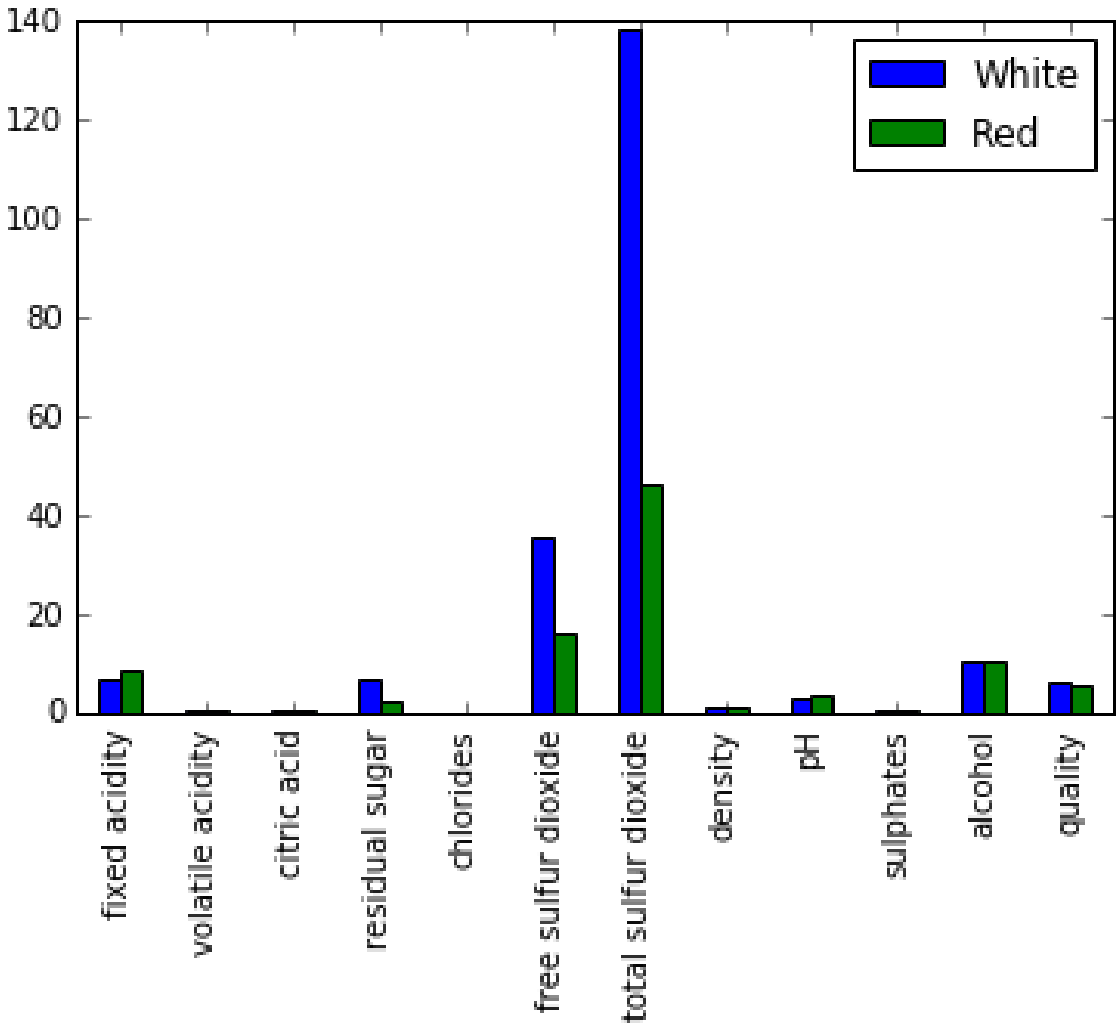*Balance, Length, Depth, Complexity*

Wine making is considered an art.

Is there a formula for a quality wine?
What basic properties are the formula for a good wine?
Do white wine and red wine share the same formula?

	White	Red
fixed acidity	6.854788	8.319637
volatile acidity	0.278241	0.527821
citric acid	0.334192	0.270976
residual sugar	6.391415	2.538806
chlorides	0.045772	0.087467
free sulfur dioxide	35.308085	15.87492
total sulfur dioxide	138.360657	46.46779
density	0.994027	0.996747
pH	3.188267	3.311113
sulphates	0.489847	0.658149
alcohol	10.514267	10.42298
quality	5.877909	5.636023

Next Steps

- Data:

\- Run feature correlations

\- Scaling

- Scenarios:

If two drastic different formulas →
        additional: classification

If not huge difference →
        work out a uniformed formula

- Algorithms:

\- Lasso

\- Random Forest (importance score)

\- PCA

\- SVM

\- Kmeans Clustering