Group 9: Yirui Guo | Pei-Shin Su |Chieh-Hsin Wu | Saurabh Yadav

Dataset proposal

Dataset: Wine Quality Data Set

Source: second-hand http://archive.ics.uci.edu/ml/datasets/Wine+Quality

Data summary:
Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests.

Data Set Characteristics:	Multivariate	**Number of Instances:**	4898	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	12	**Date Donated**	2009-10-07
Associated Tasks:	Classification, Regression	**Missing Values?**	None	**Number of Web Hits:**	1632787

What we can do:

- Correlation between different attributes (e.g., alcohol, density, pH, quality)
- Can use data to predict wine quality using multivariant regression
- 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.

Other details:
Number of Instances: red wine - 1599; white wine - 4898.
Number of Attributes: 11 + output attribute
Missing Attribute Values: None
Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)