1. Title: Wine Quality

2. Sources
   Created by: Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009

3. Past Usage:

  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. ISSN: 0167-9236.

  In the above reference, two datasets were created, using red and white wine samples.
  The inputs include objective tests (e.g. PH values) and the output is 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). Several data mining methods were applied to model
  these datasets under a regression approach. The support vector machine model achieved the
  best results. Several metrics were computed: MAD, confusion matrix for a fixed error tolerance (T),
  etc. Also, we plot the relative importances of the input variables (as measured by a sensitivity
  analysis procedure).

4. Relevant Information:

  The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine.
  For more details, consult: http://www.vinhoverde.pt/en/ or the reference [Cortez et al., 2009].
  Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables
  are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

  These datasets can be viewed as classification or regression tasks.
  The classes are ordered and not balanced (e.g. there are munch more normal wines than
  excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent
  or poor wines. Also, we are not sure if all input variables are relevant. So
  it could be interesting to test feature selection methods.

5. Number of Instances: red wine - 1599; white wine - 4898.

6. Number of Attributes: 11 + output attribute

  Note: several of the attributes may be correlated, thus it makes sense to apply some sort of
  feature selection.

7. Attribute information:

  For more information, read [Cortez et al., 2009].

  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)

8. Missing Attribute Values: None