Final Project: Analysis of Wine Quality Data

Overview

For the final project, each group will be analyzing the wine quality data. This data contains a set of observations on a number of red and white wine varieties from the Vinho Verde region in Portugal. The observations consist of measurements of the wine's chemical properties and rating by tasters. A key factor in wine certification and quality assessment is physicochemical tests which are laboratory-based and takes into account factors like acidity, pH level, presence of sugar and other chemical properties. For the wine market, it would be of interest if human quality of tasting can be related to the chemical properties of wine so that quality assessment can be more objective, rather than subjective.

Two datasets are available, one on red wine having 1,599 different varieties, and the other on white wine having 4,898 varieties. We will only analyze the **red wines**. Data are collected on 12 different properties of the wines one of which is Quality, based on sensory data, and the rest are on chemical properties of the wines including density, acidity, alcohol content etc. All chemical properties of wines are continuous variables. Quality is an ordinal variable with possible rating from 1 (worst) to 10 (best). Each variety of wine is tasted by three independent tasters and the final rating assigned is the median rating given by the tasters.

A predictive model developed on this data is hoped to provide guidance to vineyards regarding the quality expected for their wine without heavy reliance on volatility of wine tasters.

Group Project Tasks

You will first divide the data randomly into two sets, a training set that you will use to fit your models and a test set that will be used to evaluate the methods. Your group should use both regression and classification methods on the data. For the classification methods, you should collapse the ratings into a smaller number of levels, such as low/medium/high, or low/high. If relevant, you can also examine potential unsupervised methods.

Your group report should be between 7-10 pages long (maximum 10 pages, 1 inch margin, 11 pts font or larger Times New Roman, 1.5 line space): describing the data, your methods of analysis, any relevant comparisons between the methods, your findings, and any issues that came up. You should include any plots, graphs, or tables that support your statements, but these should be referenced in the text, and placed at the relevant locations. Each plot, graph and table should be numbered and must have a proper caption. Your main report should not include any raw R code/output -- you may include a separate appendix with relevant R-code that you may deem necessary. Any cover page, references and appendix are not counted towards the 10 pages.

Availability of the Data

The data is available at the UCI Machine Learning Repository, and the direct link to the data is here: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

We will be using the **red wine data** for this analysis.