Intelligent Data análisis

Basic exploration of a data set

Introduction

We have chosen to use the wines dataset, this dataset has 12 variables based on physicochemical tests on different wines plus a categorical variable grading the wine quality by experts between 0 (very bad) and 10(very excellent) and a binary variable, type: (1 = white wine and 2 = red wine). In order to understand better the dataset, we have research information about the different variables:

* Fixed acidity: a non-volatile acid(also known as a fixed acid or metabolic acid) is an acid produced from sources other than carbon dioxide. They are produced from e.g. an incomplete metabolism of carbohydrates, fats, and proteins.
* Volatile acidity: In the context of wine making, volatile acidity is a form of wine spoilage that is by-product of microbial metabolism that can occur during the wine making process with the introduction of harmful bacteria. Dimensionality not appreciated
* Citric acid has many uses in wine production. Citric acid is a weak organic acid, which is often used as a natural preservative or additive to food or drink to add a sour taste to food. It can also be used to neutralize surfaces that have been treated with basic substances. Dimensionality not appreciated
* Residual sugar, percentage of sugar not fermented (sugar fermented= alcohol)
* Chloride: Cl-, don’t know too much but high concentrations makes the wine salty (pH >7)
* Free sulphur dioxide: Sulphur dioxide (SO(2)) is used as a preservative and stabilizer in wine production to prevent undesired biochemical processes in the must and the final product. The concentration of SO(2) is restricted by national regulations. There are two main forms of SO(2) in wine-free (inorganic forms) and bound (fixed to organic compounds, e.g. aldehydes)
* Total sulphur dioxide
* Density: mass/volume
* pH(acid, salinity meter)
* sulphates: turns into sulphur dioxide (S-)
* alcohol percentage
* quality: rated from 0 to 10
* type (1= white wine or 2)

1. Choose a quantitative variable and explore its distribution in terms of descriptive measures of center, dispersion, skewness and kurtosis. Is a normal model a plausible one for its distribution? If the answer is no, can you think of a transformation of the variable that improves normality. Are there any outliers?