Sarah Ferrero **INFO 8000 Machine Learning Final Project**

Red Wine Quality Machine Learning Project

Deliverable 1:

This dataset is from the UCI Machine learning repository (*https://archive.ics.uci.edu/ml/datasets/wine+quality*). I chose this dataset because it is organized and clean, so the time to manage the dataset and get it ready is minimal or non-existent. In addition to the dataset being ready to be used, the ratio of observations and variables/qualities is over 1:10, which is optimal for creating machine learning algorithms.

As a beginner to machine learning, this dataset was appealing to me because it appeared to be manageable as well as interesting though all of the variables are numeric.

There are 12 variables, which will allow for a more thorough investigation of correlations and statistical visualizations. The attributes or variables included in the datasets are Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur, Total Sulfur, Density, pH, Sulphates, Alcohol, and Quality.

Dataset Source

The dataset is related to red Portuguese Vinho Verde wine.

The original study that utilized and collected this data states:

"The data were collected from May/2004 to February/2007 using only protected designation of origin samples that were tested at the official certification entity (CVRVV). The CVRVV is an inter-professional organization with the goal of improving the quality and marketing of vinho verde. The data were recorded by a computerized system (iLab), which automatically manages the process of wine sample testing from producer requests to laboratory and sensory analysis. Each entry denotes a given test (analytical or sensory) and the final database was exported into a single sheet (.csv)."

https://doi.org/10.1016/j.dss.2009.05.016

Possible Applications

Using this dataset to create a machine learning algorithm may allow predicting the quality rating of a wine before having consumers or testers taste the wine. It could aid in predicting pricing of the wine in determining the quality. Algorithms could be used in quality assurance, marketing, and financial predictions.

Problem To Be Solved

I will use this dataset to predict wine quality rating.