PROJECT PROPOSAL

**TEAM SHARK**

**Kelsey Day**

**Mohit Rajesh Desai**

**Raymond Hawkins**

**Raghavendran Muthusamy**

Wine Quality Analysis

**Outline of the Report**

1. **Background Information**
2. **Objective**
3. **Description & Source of Data Set**
4. **Nature of Variables**
5. **Data Mining Techniques & Methodology**
6. **Data Exploration & Graphs**
7. **References**

**Background Information**

This project analyzes the various properties of north Portuguese white wine called “Vinho Verde” in order to link these properties to the quality of the wine so that we can help wine manufacturers determine the correct pricing of white wine.

The wine industry is growing by the minute as social drinking is on the rise. One of the challenges that face the pricing of white wine is that it is dependent on abstract and volatile features like wine taste, wine appreciation by wine tasters etc and this highly varies from professional to professional and amateur to amateur. Another factor with regards to wine pricing are physicochemical tests. These include clinical tests which takes into account factors chemical properties of wine like sulphates, amount of acidity, pH levels etc which determine wine quality.

Thus it would be beneficial for wine manufacturers, distributers and traders to quantify and assess the qualitative factors of wine production like human taste measured against chemical properties of wine so that pricing of wine is based on better understanding of human interest. This would in turn will assure that the process wine certification and quality assessment is more controlled and well understood.

**Objective**

The project’s objective is to model white wine quality based on chemical properties, which form the attributes in the given data set. Thus linking quality/human taste to chemical properties helps wine makers make a thoughtful decision with regards to wine production and helps them determine the price expected on their produced wine.

**Description & Source of the Data Set**

The data obtained talks about the chemical properties of north Portuguese white wine called “Vinho Verde”. The owner of the data is “Center for Machine Learning & Intelligent Systems” of University of California at Irvine and the author of the data is Paulo Cortez of University of Minho located in Guimarães, Portugal.

The data set consists of 11 input variables and 1 output variables

Input Variables : fixed house, volatile acidity, citric acid, residual sugar, chlorides, free sulfur oxide, total sulfur oxide, Density, pH, sulphates, alcohol

Output Variable : quality

**Nature of Variables**

| **Sr No** | **Variable Name** | **Data Type** |
| --- | --- | --- |
| 1 | Fixed House | Interval |
| 2 | Volatile Acidity | Interval |
| 3 | Citric Acid | Interval |
| 4 | Residual Sugar | Interval |
| 5 | Chlorides | Interval |
| 6 | Free Sulfur Oxide | Interval |
| 7 | Total Sulfur Oxide | Interval |
| 8 | Density | Interval |
| 9 | pH | Interval |
| 10 | Sulphates | Interval |
| 11 | Alcohol | Interval |
| 12 | Quality | Interval |

**Data Mining Techniques & Methodology**

We chose a data set that analyzes wine quality based on different components and quantities such as acidity levels, sugar amounts, density, pH level, and others to get an idea of how they affect the quality ratings. To better understand the relationship between these variables and the quality rating, we will use the following techniques :

* Multiple Linear Regression
* Regression Tree Analysis.

Multiple Linear Regression.

This method or analysis is helpful for this data set because it dissects and explains the relationship between a dependent variable and many independent variables, which we have in this case. In our analysis, the wine quality rating is the dependent variable and the levels and amounts of ingredients are the independent variable. The multiple linear regression method allows us to forecast trends in wine quality and determine what it is that makes highly rated wine. Our data set contains 11 components that affect the wine quality. This is a good data set for using the complexity of a multiple linear regression model because the higher number of independent variables to a model increases the validity of the statistics. With 11 components, forecasting efforts are significantly more precise.

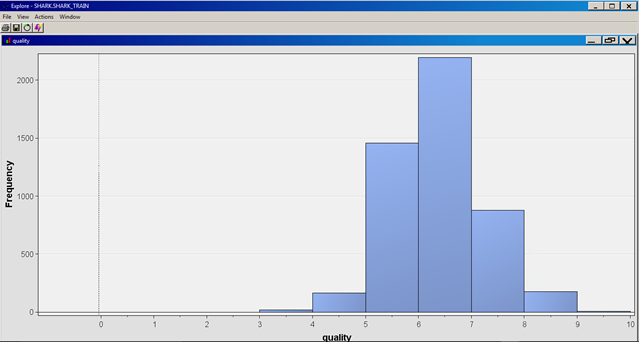
Decision Tree Analysis.

The point of this method is to create a model that predicts the value of the target variable (quality) based on several input variables. Specifically, we will be using the regression tree model because our predicted outcome can be considered a real number. Because we are attempting to predict a target value, we use regression tree and not classification. In this model, we will fit the regression model to the target variable using each of our independent variables (pH level, sugar amount, acidity levels, etc.). Then, for each of these independent variables, we will split the data at several points. This produces an “error” between predicted value and actual value. From there, we are able to get a better understanding of the weight each variable holds in the output of the target variable.

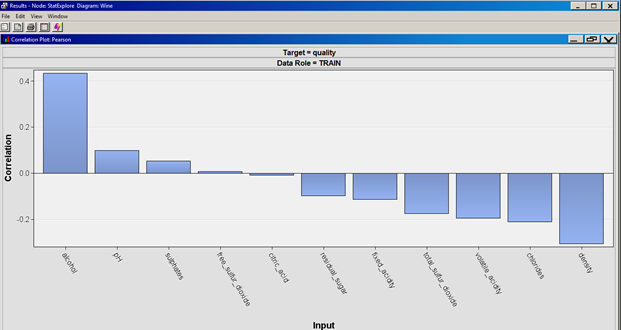
**Data Exploration & Graphs**

**Location of Data Set Analysis in SAS Miner**

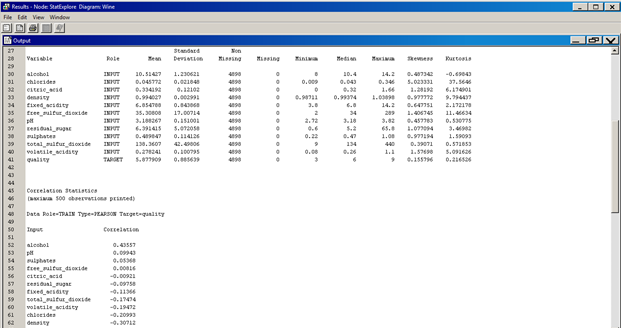
Path: O:\student$\waj245\WineAnalysis



According to this histogram, which reveals the frequency of each rating on a scale of 1-10, the majority of our target variable, quality, falls in the range of 5-7. In fact, 4535/4898 of the observations fall in this range. 183 observations fall below this range and 180 observations exceed 7 for quality. One of the main things we want to achieve in this study is to figure out what chemical property characteristics cause wines to be below average (less than 5) and above average (more than 7).



This graph represents the 11 input variables and their direct correlation with the target variable, quality. Density seems to be the variable that is most negatively correlated with quality and alcohol is most positively correlated. We will have to look further into the individual relationships between each variable to be able to tell a little more about chemical property effects on wine quality.



These are some basic statistics on our data set including numbers such as mean, median, and standard deviation. The numeric correlation values are on the bottom as well. Those variables with relatively high standard deviations (alcohol, residual sugar, and sulfur dioxide), although their ranges are relatively large as well, could provide useful information regarding wine quality because they tend to fluctuate a little more.

**References**

[1] Wine Quality Data Set University of California , Irvine Center for Machine Learning and Intelligent Systems <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>

[2] Link to data set <http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv>