Analysis of Wines Data Sets

Lior Shifman

September 2020

Abstract

In this paper I analyze two large data sets which contain attributes of white and red wine samples, and their quality graded by wine experts. The main focus of this paper is analyze the data using supervised learning methods. The main goal was trying to understand the impact of each attribute on the wine quality, using machine learning. I included my results, mainly using graphs.

1 Introduction

This paper presents the analysis of two data sets, one with 4898 white wine samples and the other with 1599 red wines samples. Each data set has 11 attributes based on physicochemical tests, and one more attribute that is the quality.

These are the 12 total attributes of each wine:

- **FIXED ACIDITY:** The fixed acidity of the wine.
- **VOLATILE ACIDITY:** The volatile acidity of the wine.
- **CITRIC ACID:** How much citric acid in the wine.
- **RESIDUAL SUGAR:** How much residual sugar in the wine.
- **CHLORIDES:** How much chlorides in the wine.
- **FREE SULFUR DIOXIDE:** How much free sulfur dioxide in the wine.
- **TOTAL SULFUR DIOXIDE:** How much total sulfur dioxide in the wine.
- **DENSITY:** The density of the wine.
- **PH:** The pH value of the wine.
- **SULPHATES:** How much sulphates in the wine.
- **ALCOHOL:** How much alcohol in the wine.
- **QUALITY:** The quality of the wine.

2 Supervised Learning Methods - Classification, Regression, Clustering and Dimensionality Reduction

In this section I apply many different supervised learning methods on the data set, and try to visualize the results. Some results are in higher dimensions, so these are either not shown or shown after Dimensionality Reduction.

2.1 Classification and Regression

Since the main attribute we want to learn is the QUALITY, and the QUALITY is being measured as an integer from 0 to 10, it possible to either apply Classification methods or Regression methods. I have decided to apply the SVC method for Classification, and SVR and Random Forest methods for the Regression.

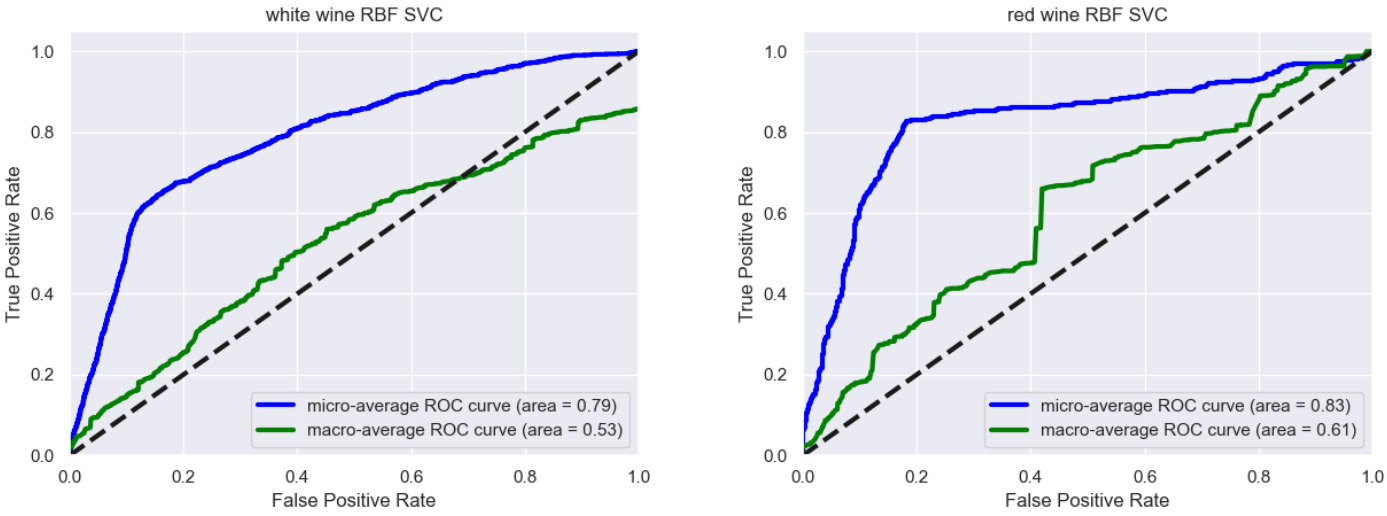

Figure 1: Classification AUC-ROC Curve

In order to check the accuracy of the model, I used the AUC-ROC curve. The Classification consists 11 classes, so in order to sum the results in one curve, I needed an average. I used 2 kinds of averages, a macro average and a micro average. The macro average is calculated as every class has the same weight, and the micro average is calculated as each class has a different weight, that is relative to the class size. Our data set is very unbalanced, so it's no surprise that the micro average has a much better result than the macro average.

	MSE	R2 Score	Explained Variance Score
RBF SVR	0.8	0.25	0.25
Random Forest	0.58	0.46	0.46

	MSE	R2 Score	Explained Variance Score
RBF SVR	0.61	0.35	0.35
Random Forest	0.52	0.44	0.45

Figure 2: Regression Scores (white wine on top, red wine on bottom)

I used the following Regression scores: MSE, R2 Score, and Explained Variance Score.

2.2 Clustering

I have decided to apply the K-Means, Hierarchical Clustering, and GMM methods for the Clustering.

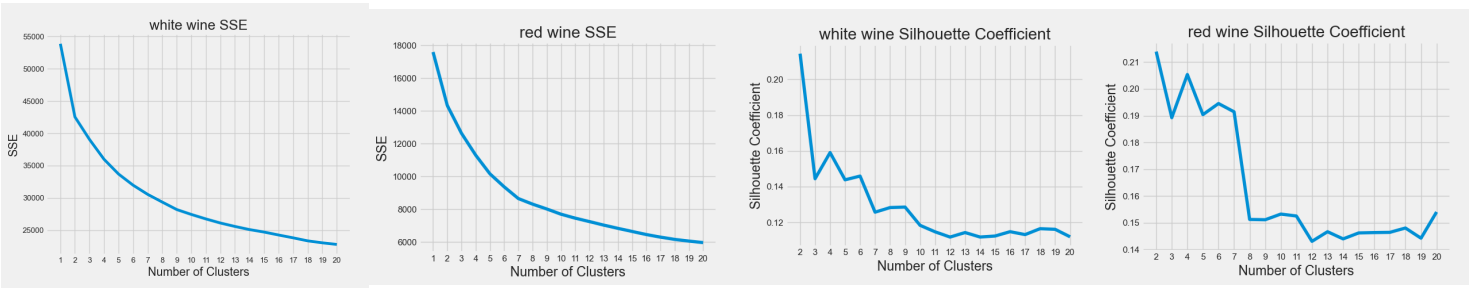

Figure 3: Clustering scores

Using the Distribution score (SSE), I can find the most optimal amount of clusters for my Clustering problem. Here I detect the "elbow" of the graph, to try to minimize the total SSE but still not having to much clusters. The Silhouette coefficient also helps me decide what is the right amount of clusters to use.
Here I used K-Means, Hierarchical Clustering, and GMM Clustering, after I already calculated the optimal number of clusters.

white wine Mean Quality of Each of 6 cluster

K-Means	5.28	5.5	5.63	5.85	6.04	6.46
Hierarchical Clustering	5.44	5.53	5.68	5.92	5.93	6.31
GMM	5.44	5.56	5.6	5.86	5.9	6.37

red wine Mean Quality of Each of 7 cluster

K-Means	5.32	5.32	5.36	5.62	5.86	5.9	6.3
Hierarchical Clustering	5.35	5.36	5.38	5.48	5.91	5.95	6.18
GMM	5.28	5.3	5.38	5.48	5.79	5.97	6.13

Figure 4: Mean QUALITY for each cluster

2.3 Dimensionality Reduction

I have decided to apply the PCA, ICA, and NMF methods for the Dimensionality Reduction.

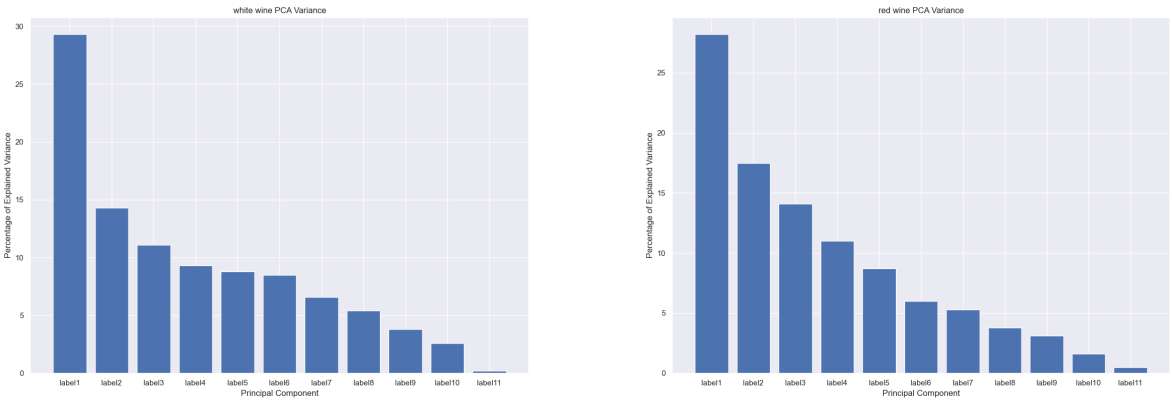

Figure 5: Dimensionality Reduction Percentage of Explained Variance

In this graph I used PCA in order to find the Explained Variance of each dimension, after Dimensionality Reduction. The transformed dimensions from PCA are labeled "label1", "label2", etc.

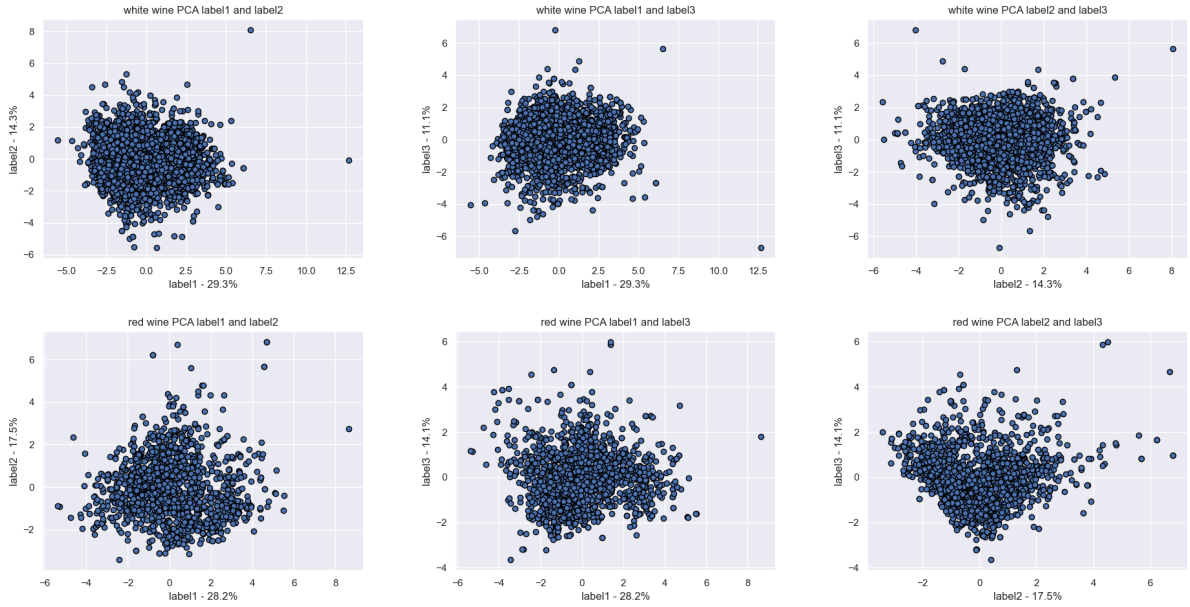

Figure 6: Dimensionality Reduction 12 to 3 using PCA

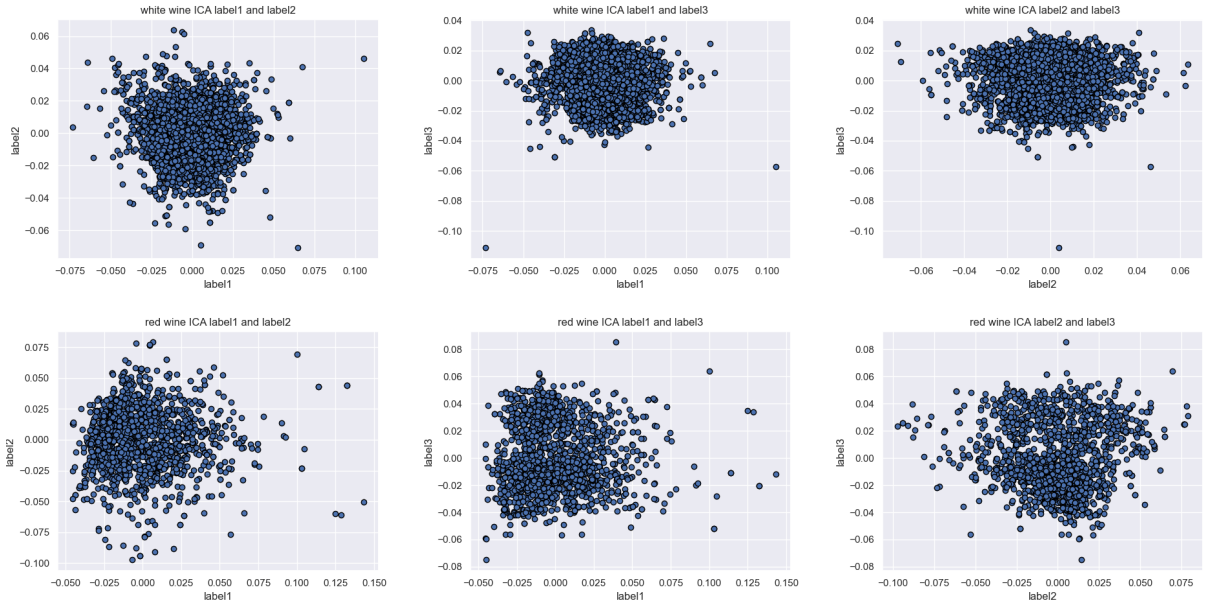

Figure 7: Dimensionality Reduction 12 to 3 using ICA

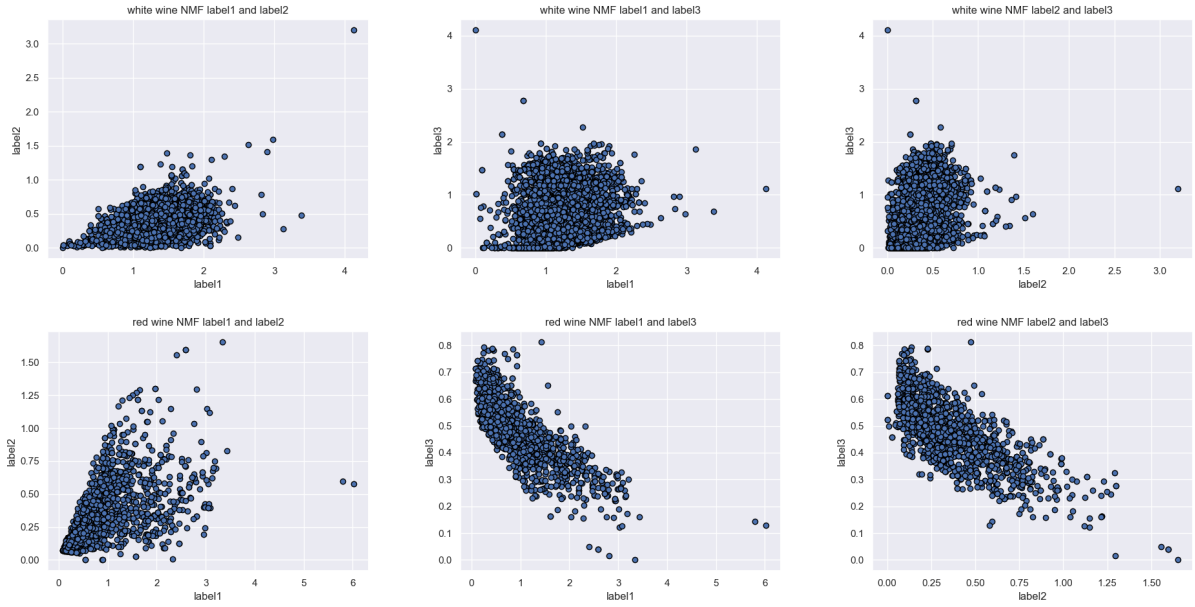

Figure 8: Dimensionality Reduction 12 to 3 using NMF

Since the Dimensionality Reduction is 12 to 3, in order to display the results on a 2 dimansional graph, 3 graphs are needed. After the Dimensionality Reduction, I checked the highest correlation between the projections to the QUALITY. In the white wine they were label3 of ICA, label3 of PCA, and label1 of ICA with the correlations 0.30, 0.21 and 0.20, respectively. In the red wine they were label3 of ICA, label3 of PCA, and label3 of NMF, with the correlations 0.51, 0.39 and 0.27, respectively.

3 Further Analysis

3.1 Feature Learning

In this section I will further analyze the data set, other than understanding the impact on the QUALITY. One of the things we were requested is analyzing which of the columns are the easiest to learn by the other columns. The learning is using the Random Forest Regression.
As we can see in Figure 9, in the white wine data set the ALCOHOL, RESIDUAL SUGAR and DENSITY are the easiest columns to learn. In the red wine data set the DENSITY, FIXED ACIDITY and ALCOHOL are the easiest columns to learn. The score given here is a R2 Score.

white wine R2 Score of Each of 12 attributes

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0.62	0.49	0.5	0.9	0.42	0.6	0.69	0.86	0.62	0.34	0.9	0.45

red wine R2 Score of Each of 12 attributes

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0.86	0.53	0.74	0.45	0.65	0.73	0.71	0.87	0.78	0.53	0.79	0.42

Figure 9: R2 Score of each Reg

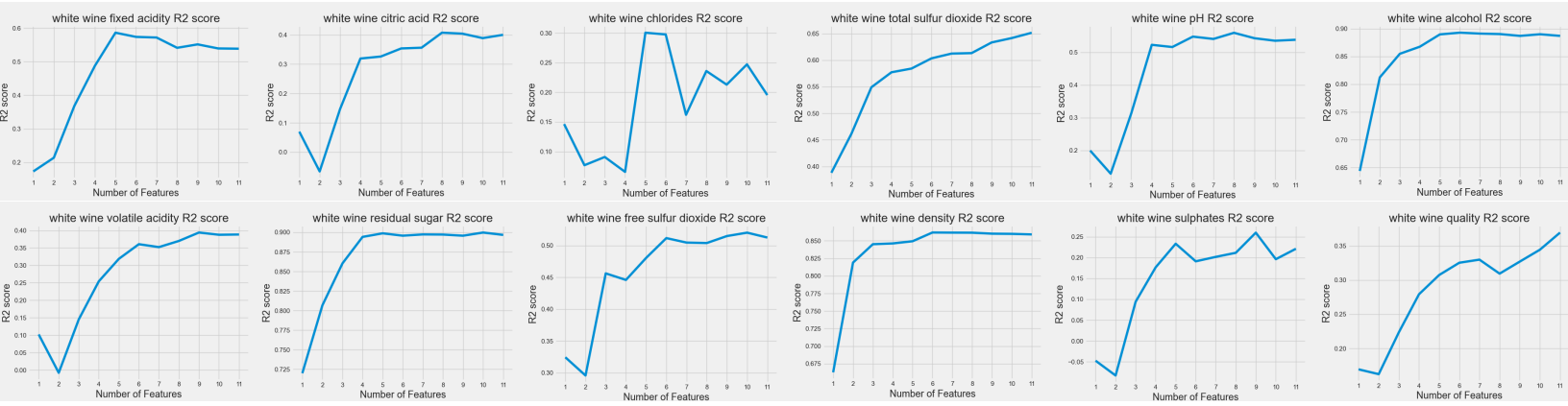

Figure 10: White Wine R2 Scores After Reduced to N Features

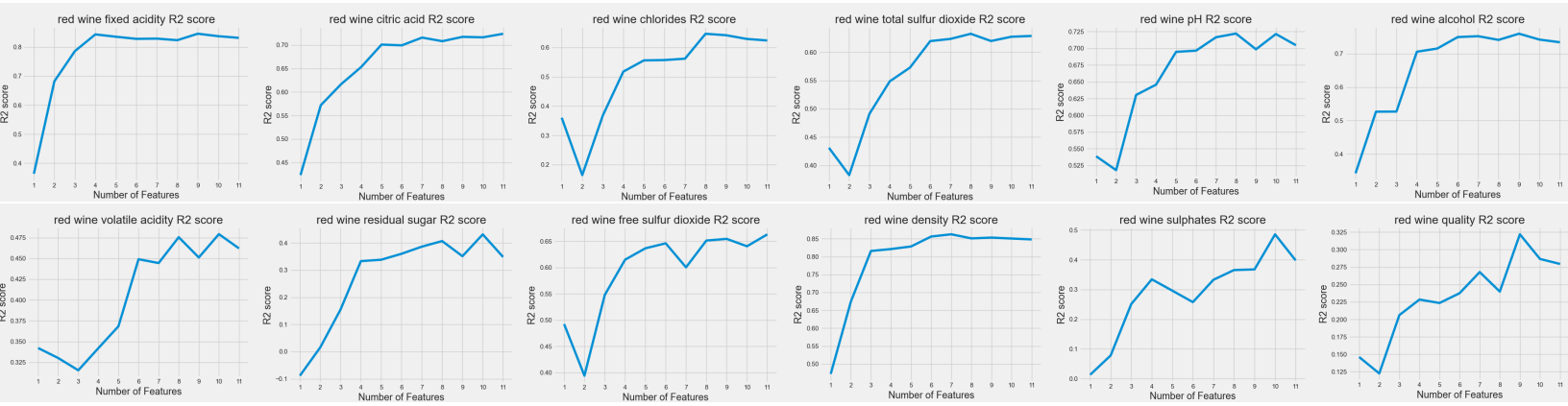

Figure 11: Red Wine R2 Scores Learning from N Features

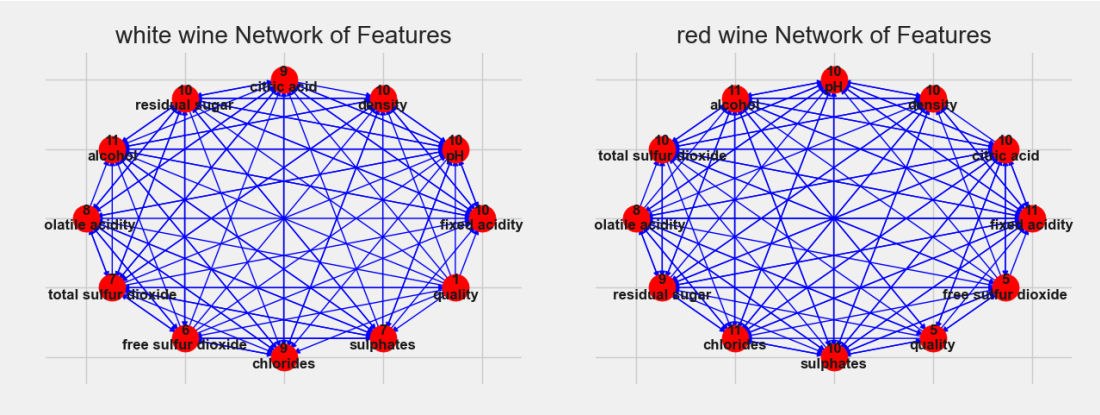

Figure 12: Networks of Features, Showing How Many Features learn from each Feature

In figures 10 and 11 we see how the number of features used in the Regression affected the R2 Score. I calculated this by iterativly adding more and more features, starting at the most contributing feature, and going down. In figure 12 we see each feature and how many features used it in the learning. ALCOHOL, FIXED ACIDITY and CHLORIDES are the most informative in red wine, while ALCOHOL and FIXED ACIDITY are the most informative in white wine.

3.2 Feature Predicting

One of our tasks is to predict the most optimal features, that will most likely give maximun QUALITY. In order to do that, I used Linear Regression for every attribute, and while forcing QUALITY to be 10, getting at the end 11 equations with 11 variables, which is easy to solve. For white wine, the optimal values for QUALITY: 10 are - FIXED ACIDITY: 6.41, VOLATILE ACIDITY: 0.19, CITRIC ACID: 0.33, RESIDUAL SUGAR: 4.09, CHLORIDES: 0.02, FREE SULFUR DIOXIDE: 35.95, TOTAL SULFUR DIOXIDE: 103.8, DENSITY: 0.99, PH: 3.26, SULPHATES: 0.52, ALCOHOL: 13.01. For red wine - FIXED ACIDITY: 9.49, VOLATILE ACIDITY: 0.15, CITRIC ACID: 0.51, RESIDUAL SUGAR: 2.64, CHLORIDES: 0.05, FREE SULFUR DIOXIDE: 13.01, TOTAL SULFUR DIOXIDE: 13.56, DENSITY: 0.99, PH: 3.26, SULPHATES: 0.89, ALCOHOL: 13.17. Estimated accuracy for white and red wine is 0.28 and 0.36.

3.3 Coefficients Analysis

One of our tasks is to analyze and compare the coefficients of Regression and Dimensionality Reduction of the white wine to the red wine. I used the Pearson correlation coefficient. In the Regression the Pearson coefficient is equal to 0.992, comaparing between he white wine and the red. In the Dimensionality Reduction the Pearson coefficient is equal to 0.973.