Exploratory analysis of the White Wine dataset.

Bidish Basu
Dublin City University

Sachin Mahesh
Dublin City University

***Abstract*—Analysis of a dataset on various attributes that describe the quality of the Wine. The goal of this assignment is to best describe the variation in different types of wine available to the consumers. We find the correlation of each attribute in the data to pose a hypothesis and perform feature selection and feature relevance, further we performed A/B Hypothesis testing where data is chosen at random to determine which variation gives a better result.**

Feature scaling was implemented to convert the raw data into a normal distributed one using natural logarithm scaling, few outliers were detected during the process and had to be removed. Further feature transformation was implemented using "Principal Component Analysis" to draw conclusions about the underlying structure, which dimensions about the data best maximizes the variance of features involved and also *explains the variance ratio* of each dimension. Findings are visualized using various plots such as Biplot and Clusters.

Keywords
Correlation, Hypothesis Testing, Feature Relevance, Feature Scaling, Principal Component Analysis, Dimensionality Reduction, Clustering, Gaussian Mixture Model, Biplot

I. Introduction

The following Paper contains the steps enumerated below for analyzing characteristics of white variants of the Portuguese "Vinho Verde" wine. Quality is based on sensory scores (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

Wine is one of the largest industries in the world, Portuguese being one of the top exporters of wine, popularity of wine has skyrocketed in the past few years [6]. New and advanced tools and techniques are being developed to gauge the quality of the wine, regular quality assessment is the key stone of the wine manufacturing process which helps in improving the quality of wine. Physicochemical research facility tests are used to distinguish wine based on pH, chlorides or density values but sensory tests depend mainly on human experts. Humans are prone to make errors thus wine characterization is a troublesome task. Moreover, striking a balance between sensory and physicochemical analysis are complex and not fully discovered. Quality of wine depends on various aspects and is complex mixture of hundreds of components, but main components are alcohol content, PH value, amount of residual sugar, citric acid and sulphates.

Some of the implementation techniques for exploratory data analysis include, Feature Relevance i.e. Is it possible to determine whether some characteristics of one category of wine will necessarily impact other characteristics in the mix, in order to find such features we trained a supervised regression learner on a subset of the data with one feature removed, and then score how well that model can predict the removed feature. Lot of preprocessing had to be done in order to normalize the data, once the data was normalized Feature Scaling was implemented using Natural Logarithm, few plots were skewed due to outliers and had to be dealt. Now the data has been scaled to a more normal distribution and has had any necessary outliers removed Feature transformation was implemented using PCA to reduce the dimensionality of the data which in turn reduces the complexity of the problem.

II. Related Work

There has been enormous work in this spectrum to reduce the cost related to processing grade of a wine and reduces the need for certification, various machine learning techniques such as support vector machines, neural networks and linear or Multivariate Regression are used to determine the quality of the wine based on certain attributes [5].

Data mining also plays a part in this sector which is coupled with neural networks to find the best fit and to determine trends and patterns which can be used to predict the quality of wine down the lane. Linear regression model builds the relationship between two or more independent variables and a dependent variable, it is basically used to predict the value of a variable depending on other variables. It also determines the overall fit of the model and is used to predict the values of independent variables relative to their contribution [1].

Support Vector Machines are also implemented which acts as an upgrade from regression and neural networks which outperforms both the methods, the aim of a SVM is to find the best linear separating hyperplane with small errors when fitting model to the data but SVM's are supervised models that can be used for classification [5]

III. Dataset And Exploratory Analysis

The dataset has been procured from the UCI Machine Learning Repository. [1] The white wine dataset is related to

the “Vinho Verde” wine of the Portuguese. [2] The dataset contains 4898 rows and 12 columns. The following are the attributes present in the dataset, fixed acidity (this refers to the nonvolatile acids in the wine), volatile acidity(this refers to the volatile acid, specifically acetic acid which in high amounts in the wine can make it taste unpleasant), citric acid(this attribute refers to the amount of citric acid present in the wine which provides additional flavor to the wine), residual sugar(this attribute refers to the quantity of sugar that is left behind in the wine after the fermentation process has been stopped), chlorides(this refers to the quantity of salt in the wine), free sulfur dioxide(this refers to the amount of Sulphur dioxide present in the wine as a dissolved gas), total sulfur dioxide (this refers to the total amount of Sulphur dioxide present in the wine), density(the density of the wine based on the alcohol percentage and the quantity of sugar), pH(this tells us the acidity amount of the wine), sulphates(this is an additive which acts as an antioxidant and antimicrobial), alcohol(this refers to the amount of alcohol present in the wine), quality (this refers to the score between 0 and 10 used to grade the wine).[3] Each of the rows gives us information about the various attributes of the wine. The first step in the exploratory analysis was to clean the datasets to see if there were any missing values or duplicates in the dataset. After checking for the above-mentioned parameters, the dataset had 4898 rows and 12 columns. For the sake of the hypothesis we had to drop two columns, citric acid and quality as they were not relevant to the analysis that was performed.

IV. Hypothesis

For the assignment, we are trying to find the relationship between the quality of wine and the amount of sulphates present in the wine. Therefore,

Null Hypothesis: The quality of the wine gets better with less amount of sulphates in it.

Alternate Hypothesis: The quality of the wine gets better with more amount of sulphates in it.

V. Proposed Methodology and Findings

The method used in order to assess the quality of wine is A/B Hypothesis testing. A/B testing is used in order to check if making changes to any existing product will affect it in a positive way or in a negative way. In this assignment, we are trying to find out if the amount of sulphates present affects the quality of the white wine. First, the white wine data is preprocessed and certain attributes which are irrelevant for our analysis are removed. Next, a sample of the dataset is taken and is explored, then the feature relevance was considered in order to find out which feature was relevant to the model. The coefficient of determination was then calculated to find out which was the feature that we would be considering for the hypothesis. After that the feature normalizing was performed using the natural logarithms and then the dimensions were reduced using the Principal Component Analysis in order to visualize and understand how the features relate to each other and where the variance is maximum. Once this was done, the white wine was clustered using the Gaussian Mixture Model to identify the relationship of the sulphates to the quality of the wine.

A. Feature Relevance

The quality of wine is determined by multiple factors, such as alcohol content in percentage, the amount of sugar that gets left behind after the fermentation process is finished but

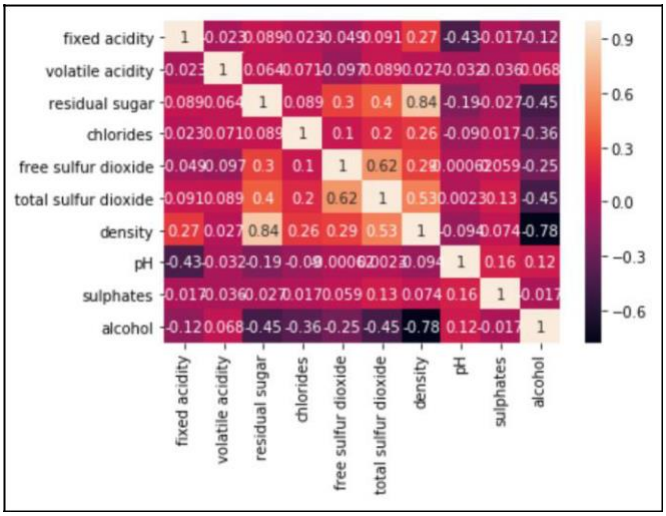

Fig 1: Correlation Matrix of the attributes

there are also some features like the volatile acidity which tells us about the amount of SO2 that is present in gas form which has no relevance to the quality of the wine. Now, in order to see which feature is relevant, we first plotted a heatmap of the correlation of all the features as seen in Fig 1. From this we see that there is a strong correlation among density and residual sugar and total sulfur dioxide and free sulfur dioxide. This implies that if we have either of the data points in the respective cases, then predicting its correlated feature will be easy. On the contrary, a feature such as sulphates has very less correlation with the other features which tells us that it is an important feature and will become difficult to predict its values with the help of the other features.

B. Feature Scaling

If the data is not distributed normally, then, it can mislead the analysis by moving the median and the mean, i.e., a skewed dataset can cause immense change in the final analysis. One of the most common way to achieve normalization of the data is by applying natural logarithm. The Fig 2 shows us the results after we applied the natural

logarithm on the dataset and how it reduced the skewness. We see from the figure that sulphates have not much correlation with the other features in the dataset.

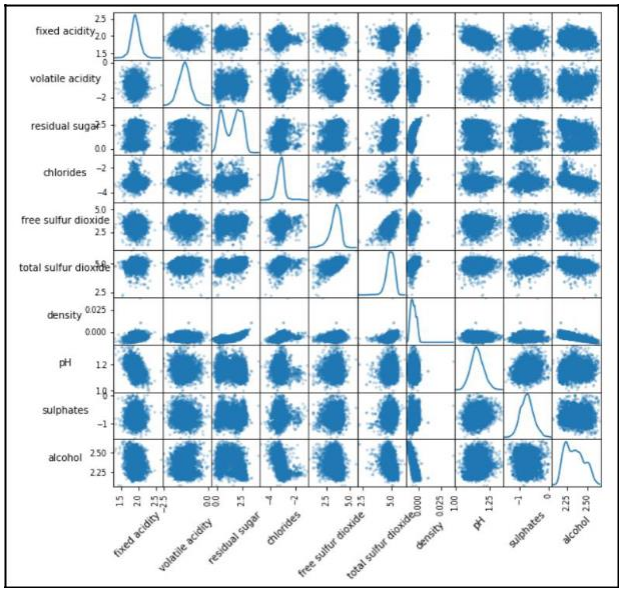

Fig 2: Plot of the dataset after applying natural log to reduce skewness.

C. Dimensionality Reduction Using Principal Component Analysis

Now that we have the data that has been normalized, the next step was to find out which dimension about the data best maximizes the variance of the features present in the dataset. This was achieved by applying the Principal Component Analysis on the dataset. By reducing the number of dimensions, it becomes easier for the analysis as the complexity reduces from multiple dimensions, in this case we have 10 features so the number of dimensions to be reduced are 10, to a few manageable dimensions. In this assignment we have reduced the number of dimensions to 2.

	Dimension 1	Dimension 2
0	-1.617566	0.100302
1	1.149611	0.350929
2	-0.371721	0.280795
3	-0.817820	-0.298501
4	-0.817820	-0.298501

Fig 3: The reduced dimensions helps us understand the data easily by reducing its complexity.

But the issue with the dimension reduction is that lesser the dimensions we have less variance to work with. Once the dimensions were reduced, we created a biplot which helped visualize the dimensions. The biplot here shows us the original features along the components. The axes here are the reduced dimensions. The biplot helps us discover relationships between the original data features and the principal components.

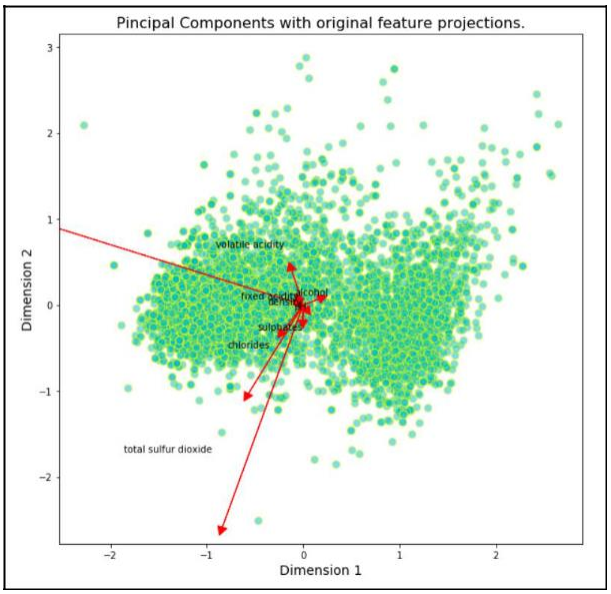

Fig 4: Biplot showing the relationship between the Principal Components and the original features.

From the biplot above we can now easily interpret the relative position of each of the datapoints as we have the original features projected on top of it. For example, if we choose a point on the lower left corner then, we can conclude that that specific wine has a large amount of sulfur dioxide.

D. Clustering

Once the dimensions were reduced, we used the Gaussian Mixture Model clustering algorithm to identify the good and bad wine based on the sulphate quantity. The Gaussian Mixture Model was chosen because, it can incorporate the covariance among the data points into the model and identify complex clusters. From Fig 4, we see that a lot of the data points are quite uniformly distributed. This tells us that the data points don't belong to a specific cluster. A Gaussian Mixture model can be used on a non-spherical cluster which seems to be the case here.

Now, one we know which clustering algorithm we were going to use, we had to decide on the number of clusters needed. In order to identify the number of clusters needed, we calculated the silhouette coefficient [4]. The silhouette coefficient is used to measure how similar a data point is to the cluster to which it is assigned. This is measured between -1 to 1. -1 refers to the point being dissimilar and 1 refers to the point being similar to the cluster. Once that is done, the average of the silhouette coefficient is calculated which provides us a score for the given number of clusters. For the dataset at hand, the optimal number of cluster calculated was 2 and the silhouette score was 0.52.

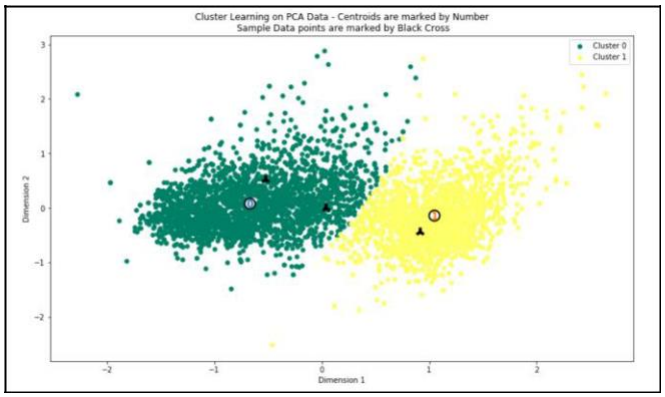

Fig 5: Clustering based on the amount of sulphate present and the sample data points selected.

Now that we had the optimal number of clusters, we plotted the data points along with the sample which we had selected in order to find out how the sulphate content affects the quality of the wine. From Fig 5, we see that there are two clusters formed based on the sulphate content, the green cluster refers to the one with the less amount of sulphates and the yellow cluster has larger quantities of sulphates. Now, the Sample data as seen in Fig 6, which we had selected to fit on to clusters on prediction shows us that the lesser the amount of the sulphate content in wine, better is its quality.

	fixed acidity	volatile acidity	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	6.2	0.47	8.3	0.029	24.0	142.0	0.99200	3.22	0.45	12.3
1	5.8	0.30	1.7	0.014	45.0	104.0	0.98914	3.40	0.56	12.6
2	7.9	0.11	4.5	0.048	27.0	133.0	0.99460	3.24	0.42	10.6

```
Sample point 0 predicted to be in Cluster 0
Sample point 1 predicted to be in Cluster 1
Sample point 2 predicted to be in Cluster 0
```

Fig 6: Sample data points selected and the predicted results.

VI. CONCLUSION

After performing the analysis on the dataset and clustering the data points based on the amount of sulphates, we conclude that the lesser the amount of sulphates in the wine better is its quality. Hence, we go and accept the Null Hypothesis and reject the Alternate Hypothesis.

REFERENCES

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547– 553. doi:10.1016/j.dss.2009.05.016

[2] Vinho Verde, Available Online : https://www.vinhoverde.pt/en/homepage.

[3] Senna Panizzo, Daniel. “Red and White Wine Quality EDA,” Available Online : https://www.kaggle.com/danielpanizzo/red-and-white-wine-quality/data.

[4] USING SILHOUETTE ANALYSIS FOR SELECTING THE NUMBER OF CLUSTER FOR K-MEANS CLUSTERING., n.d. https://kapilddatascience.wordpress.com/2015/11/10/using-silhouette-analysis-for-selecting-the-number-of-cluster-for-k-means-clustering/.

[5] Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science, 125, 305–312. doi:10.1016/j.procs.2017.12.041

[6] Escande, S., 2019. Le vinho verde. Wine territories, Available online at :https://www.nytimes.com/2013/07/03/dining/vinho-verde-portuguese-for-cheap-and-cheerful.html