I. Data

The two datasets contain two different characteristics which are physico-chemical and sensorial of two different wines (red and white), the product is called "Vinho Verde". The data were taken from the UCI Machine Learning Repository. Both datasets can be used with the permission of Paulo Cortez.

There are 1599 samples of red wine and 4898 samples of white wine in the data sets. Each wine sample (row) has the following characteristics (columns):

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8.
9. Density
10. pH
11. Sulphates
12. Alcohol
13. Quality (score between 0 and 10)

a) Purpose and Workflow :

The work is carried out on different stages, starting with the management of the data, and then we move on to visualization, then analysis, regression modeling and finally Machine learning. Exploration explains the relationships and correlations between the characteristics of the wine and its quality score. The study context is to try different predictive algorithms on the data and examine the results.

b) The work involves the following steps:

1. Data management and visualization
2. Data analysis
3. Regression modeling
4. Machine learning

c) Resources :

- P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

II. Data management and visualization :

As a first step in the project, I checked whether there is any redundancy, missing values in the data in any variable in both datasets. Then, I explore the variable of the frequency of the quality of the wines.

For both wines, the majority of the samples have ranks of quality 5, 6 and 7 (on a scale based between 0 and 10). Then, after the first analysis with the (Plot), it was found that the quality ranks of the white samples are on mean higher than those of the reds. In addition, I applied the countpots to visualize the distribution of quality rankings for both wines.

To further analyze the data, I added the factorplots which show the alcohol levels of the wine samples in each quality rank.

We noticed that there is a less strong but positive correlation between the alcohol level and the quality rank for the two wines. We find that more the quality rank is higher, more the alcohol level is higher too.

III. Data Analysis :

For the second phase of our project, I proceeded to a calculation of the correlation coefficient between the different variables and then plot it as a heat map so that we can easily see what happens.

The heat map shows that there is a positive correlation between density and residual sugar and a strong negative correlation between density and alcohol content. It is also interesting to note that there is a positive correlation between the alcohol content and the quality of the wine.

Then I decreased my analysis to two variables (sulfur dioxide and the quality of the wine).

a) Hypothesis and ANOVA :

Secondly, after investigations on different forums that deal with the win quality dataset, I realized that it was better to add a new value that will contain the brand of wine quality: high if the quality rank is higher Or equal to 8, mean if the rank of quality is equal to 6 or 7 and weak if the rank of quality is less than or equal to 5. Then we carry out an analysis of variance (ANOVA) which will be carried out separately on the two Wine datasets (red and white).

Analysis of the variance of red wine revealed that sulfur dioxide and wine quality were significantly associated, $F(2, 1596) = 45.71$, $p = 4.97e\text{-}20$.

Further point comparisons of the total mean sulfur dioxide in wine by quality marks have revealed that poor quality wine has a very large amount of sulfur dioxide compared to other groups. On the other hand, high quality wine has the smallest mean amount of sulfur dioxide

compared to other groups but not significantly different than that of the wine group with an mean quality.

Similarly, for the analysis of white wine revealed that sulfur dioxide and wine quality are significantly associated, F (2, 4895) = 76.66, p = 1.65e-33. Point comparisons of the total sulfur dioxide mean by quality markings revealed that the three groups had a significantly different amount of sulfur dioxide. In addition, poor quality wine has the highest amount of sulfur dioxide, however, the highest quality wine has the lowest amount.

b) Pearson Correlation

Then, I explore the association between the density of the wine and the amount of residual sugar in the wine.

For this purpose, the Pearson correlation is used to calculate the coefficient of these variables (density is the most explanatory quantitative variable and residual sugar is the quantitative response variable) on each wine set separately.

- The coefficient for red wine is: r = 0.355, p-value = 9.0e-49.
- The coefficient for white wine: r = 0.839, p-value = 0.0.

Otherwise, the coefficients show that there is a positive linear relationship between these variables for both datasets. However, for white wine this relationship is stronger.

In addition, it can be assumed that the percentage variability in the amount of residual sugar in wine is explained by the variation in wine density which indicates that the red wine has a value equal to 12.6% and for wine White, a value equal to 70.4% (using RSquared or Determination Coefficient: $0.355^2 = 0.126$ and $0.839^2 = 0.704$ respectively). It is found that the quantity of sugar is very important in white wine. In other words, this means that the amount of sugar is strongly related to the density of the wine.

Then, I apply scatter-plots of these variables to check the linearity of the relationship.

Both plots show a positive but non-linear correlation.

a) Exploration of statistical interactions

In this part, I check the density of the wine and the quantity of residual sugar if there is a linearity by wine quality range. Otherwise, I divide the data into 3 groups by the quality marks of the wines (low, medium, strong) then I calculate the Pearson correlation coefficient for each group. Then I define the quality marks of the wine as follows: high if the quality rank is greater than or equal to 8, average if the quality rank is equal to 6 or 7 and low if the quality rank is lower or Equal to 5 as in the previous section.

The results of the calculations show that there is a positive linear relationship between the variables for each group in each dataset. This indicates that the quality of the wine does not influence the relationship between the density of the wine and the residual sugar, which is interesting. However, we note in the scatter-plots for each of the groups a rather positive correlation but not exactly linear.

IV. Different regression models

a) Basics of linear regression

In the third part of our study, the association between the amount of volatile acidity in wine and the quality of the wine is explored by applying a simple linear regression model with an explanatory variable (the amount of volatile acidity). I apply it separately for the two sets of wine (red and white).

Scatter-plots reveal a negative association between the variables for the two wine datasets: the more volatile acidity increases, the lower the quality of the wine.

Before testing our linear regression model, the explanatory variable (volatile acidity) was centered by subtracting its mean from each observed value. It is advisable to notify that the average of the volatile acidity is different for each set of wine: for red wine it is 0.53 and for white wine 0.28.

The results of the regression modeling for red wine indicate that the model was constructed using 1599 observations, the RSquared is 0.153, F = 287.4 with p = 2.05e-59, otherwise it means that the Variables are significantly associated. The regression coefficients of the model are: intercept = 5.6360 and slope = -1.7614, which means that we can connect our variables by the formula:

Quality = 5.6360 - 1.7614 * volatile acidity.

The results of the regression modeling for white wine indicate that the model was constructed using 4898 observations, the RSquared is 0.038, F = 193.0 with p = 4.67e-43, otherwise it means that the variables are significantly Associated. The regression coefficients of the model are: interception = 5.8779 and slope = -1.7109, which means that we can connect our variables by the formula:

Quality = 5.8779 - 1.7109 * volatile acidity.

The relationship between the amount of volatile acidity and its quality is almost the same for both wines, although the average amount of volatile acidity in red wine is higher than that in white wine.

In this part, I analyze the association between the amount of sulfates and alcohol in the wine and the quality of the wine using logistic regression. To do this, I recode the explanatory and response variables into binary categorical values. I proceed to the recoding using the following rules:

b) Logistic Regression :

- Quality category: 0 - if the quality of a wine sample is 3, 4, 5 or 6. 1 - if 7, 8 or 9;
- Sulphate category: 0 - if the amount of sulphates in a wine sample is less than the average amount of sulphates in all samples, 1 - if greater;
- Categorical alcohol: 0 - if the amount of alcohol in a wine sample is less than the average amount of alcohol in all samples, 1 - if it is higher.

After recoding the variables, I model the logistic regression and calculate the confidence intervals for the explanatory variables. I perform this procedure separately on both sets of wine.

The results of the logistic regression model for red wine indicate that the two explanatory variables are positively and significantly associated with the response variable (p-values = 0).

None of these variables have a confounding factor. It can be said that after adjusting for the sulphate variable, wine samples with an above average alcohol content are 9.26 times more likely to have a high quality index (> 7). Similarly, after adjusting for the alcohol variable, samples of wine with an above average sulfate content are 3.99 times more likely to have a high quality index (> = 7).

The results of the logistic regression model for white wine indicate that the two explanatory variables are positively and significantly associated with the response variable (p-sulfate value = 0.007, p-alcohol value = 0).

None of these variables is a confounding factor. It can be said that after adjusting for the sulphate variable, wine samples with a value greater than the average quantity of alcohol are 6.00 times more likely to have a high quality index (> = 7). Similarly, after adjusting for the alcohol variable, wine samples with an amount greater than the average quantity of sulphates are 1.22 times more likely to have a high quality index (> = 7).

Consequently, the final results show that for both wines, the sulphate and alcohol variables are positively associated with the quality variable. Moreover, for the two wines, the alcohol variable is stronger than the quality variable because it has a higher regression coefficient than the sulphate variable.

V. Machine Learning :

a) Decision Trees :

In the last section of our study which is machine learning I begin the analysis by decision trees in order to test the nonlinear relations between a set of explanatory variables which are the quantity of residual sugar and alcohol in the wine and a response variable Binary and categorical coded upstream (0 - if the quality of a wine sample is 3, 4 or 5, 1 - if 6, 7, 8 or 9).

In this analysis, the criterion of entropy "goodness of division" was used to cultivate the tree and a cost complexity algorithm was used to carve the complete tree into a final subtree. A classification tree was applied for both datasets (red and white). In each series, 60% of the samples were used for training and 40% for the tests.

The resulting tree precision for red wine is 0.65. The confusion matrix is:

[198,105]

[120,217]

The resulting tree accuracy for white wine is 0.72. The confusion matrix is:

[404,262]

[290,100]

At the end of our simulation, a very large volume of trees was obtained for examination. This may indicate that the selected variables are not suitable for the training of the appropriate trees or that the analysis of the trees is not suitable for these data.

b) KNN :

A KNN Model was applied for both datasets (red and white). In each series, 60% of the samples were used for training and 40% for the tests.

The precision of our resulting model for red wine is 0.65. The confusion matrix is:

Red

[[202 100]

[89 249]]

Score: 0.6546875

The best precision of our resulting model for white wine is 0.70. The confusion matrix is:

White

[[336 320]

[264 1040]]

Score: 0.704081632653

c) Naive Bayes :

Now, we use a Naive Bayes classifier to develop a classification model. Some characteristics show a significant correlation (based on the heat map).

The dataset for the classification is then prepared.

In each series, 60% of the samples were used for training and 40% for the tests.

Then we will train with Naive Bayes classifier and use the test to check the accuracy of the model. We can do this by creating a confusion matrix. We can also create a simple summary statistic to check to see what percentage of the test set finished off the diagonal.

The best precision of our resulting model for red wine is 0.73. The confusion matrix is:

Red

[[230,88]

[83,239]]

Gaus 0.7328125

[[148,170]

[79,243]]

Multi 0.6109375

[[30,288]

[26,296]]

Bernoulli 0.5265625

The best precision of our resulting model for white wine is 0.71. The confusion matrix is:

White

[[388,295]

[272,100]]

Gaus 0.710714285714

[[284,399]

[262 1015]]

Multi 0.662755102041

[[4,679]

[4,1273]]

Bernoulli 0.651530612245

d) Random Forests :

In this section, a random forests analysis is carried out to assess the importance of all explanatory variables in the prediction of wine quality.

The analysis was carried out for each wine set (red and white) separately. In each group, 60% of the sample was used for training and 40% for testing.

The analysis involves two steps. First, I create the random forests model with 25 trees and then look at the results. Second, I run random forests with different numbers of trees (1-100) to see the effect of the number on the accuracy of the prediction.

The results of the Random Forests model with 25 trees for red wine show that the prediction accuracy is 0.778 and alcohol is the most important predictor variable, followed by volatile acidity, sulphates and sulfur dioxide total

Random Forests results with 25 trees for white wine show that the prediction accuracy is 0.816 and alcohol is the most important predictor variable, followed by volatile acidity and density.

Random Forests training with a different number of trees (1-100) shows that after about 20 trees, subsequent growth in the number of trees adds little to the overall accuracy of the forest. The simulation was applied to the two sets of wine: red and white.

e) Lasso regression

A lasso regression analysis was performed to identify a subset of wine characteristics (predictive variables) that are best for predicting wine quality (quantitative response variable). All characteristics of the wine were included as predictors; They are quantitative and have been standardized to have an average of zero and one standard deviation of one.

The data were divided into a training set (70% of observations) and one test (30%). The lowest angle regression algorithm with cross-validation k = 10 times was used to estimate the regression model of the lasso in the training set and the model was validated using l Test set. Cross-validation at each step was used to identify the best subset of predictive variables.

All predictive variables were not selected in the selected models. For red wine, citric acid, fixed acidity and free sulfur dioxide have zero coefficients and will not be taken into account in the prediction. The results of the training indicate that the variables alcohol, volatile acidity and sulphates are the most strongly associated with the quality of the wine and, consequently, the most influential for the prediction. Thus, the mean squared error and RSquared values prove that the model is very vigorous to test new examples. The predictors represent 33% of the variance of the target variable.

For white wine, citric acid and acidity variables have zero coefficients and will not be taken into account in the prediction. The results of the training indicate that the variables alcohol, residual sugar, density and volatile acidity are most strongly associated with the quality of the wine and therefore the most influential for the prediction. Thus, the mean squared error and RSquared values prove that the model is very vigorous to test new examples. The predictors represent 28% of the variance of the target variable.

f) K-Means Cluster Analysis

K-means analysis was carried out to identify the underlying subgroups of wine samples according to their characteristics (quantitative grouping variables): density, alcohol, sulphates, pH, volatile acidity, chlorides, acidity Fixed, citric acid, residual sugar, free sulfur dioxide, and total sulfur dioxide. The analysis was carried out for each wine set (red and

white) separately. All variables were normalized to have an average of 0 and a standard deviation of 1.

The data were randomly distributed in a training set (70%) and a test (30% of the observations). A series of k-means cluster analyzes was conducted on the training set specifying k = 1-9 clusters, using the Euclidean distance. The mean distance between the observations and the cluster centroids was plotted for each of the nine cluster solutions in an elbow curve to provide a guide for choosing the cluster number to interpret.

The elbow curves for both datasets were inconclusive. To choose the best solution, discriminating canonical analyzes were performed for each solution.

The canonical discriminant analysis reduces the 11 clustering variables to 2 canonical variables which represent the majority of the variance in the clustering variables. After mapping the canonical variables for each cluster solution 1-9, the 2-group solution was chosen as the one that divides the data in the best way. For white wine, observations in clusters are densely compressed with relatively low cluster variances and clusters do not overlap with each other. For red wine, the observations in each of the two groups have a greater spread in the cluster variance, but the clusters do not overlap with each other.

For red wine, the mean of the clustering variables show that the values of the first group samples have a higher mean pH, volatile acidity, free sulfur dioxide and total sulfur dioxide than the samples of the second group. The average values of all other cluster variables in the first cluster are lower than in the second cluster.

For white wine, the averages of the grouping variables show that the values of the samples of the first group have an average of the values of alcohol and pH higher than the samples of the second group. The average values of all other cluster variables in the first cluster are lower than in the second cluster.

In order to validate the clusters, an Analysis of Variance (ANOVA) was carried out to test the significant differences between the clusters on the quality of the wine.