Predictive Modeling of Red and White Wine Characteristics using Machine Learning Techniques: A Comparative Analysis

Chinedu Okeke & Emre Guvenilir

Introduction & Data

This paper utilizes two datasets of red and white wine samples of the Portuguese "Vinho Verde" wine. The inputs include objective tests (pH values, total sulfur dioxide, residual sugar, etc.) and the output is based on subjective data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). Values are based on 1599 data points for red wine and 1898 data points for white wine.

Objective

Given the two datasets, we are trying to find the reasons/predictors for the subjective scores given by the three experts. By doing so we want to gain an understanding of which of these variables contribute most in a high scoring wine and vice versa for a low scoring wine. We expect to build a model that utilizes the selected variables. Then, the model should recognize the correlation between the quality score and the variables. To do this we built a regression model(s) that can forecast the rating for novel, unseen wines.

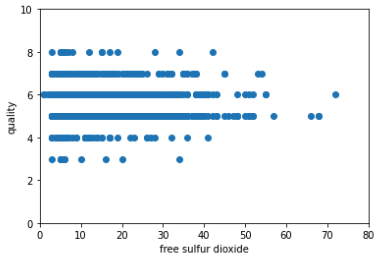
Methods

In this section, we define our approach in detail. The data exploration portion yielded significant results in how we decided to choose our features. We chose our features by utilizing a simple scatter plot where we

plotted the dependent variable, quality, against each independent variable. When looking at the variables, we knew we needed features that resembled normal distributions and lacked a significant number of outliers that could skew the model's results. At a glance and when considering the mean and of the quality of white and red wines, both have a median of 6. Red wines have an average quality score of 5.64, and white has an average quality of 5.88. When looking at the scatter plots of free sulfur dioxide, while there are a couple of outliers, the distribution remains normal (Figure 1).

```
[26]: plt.scatter(fsd_red, zval2)
      plt.axis((0, 80, 0, 10))
      plt.ylabel('quality')
      plt.xlabel('free sulfur dioxide')

[26]: Text(0.5, 0, 'free sulfur dioxide')
```




```
[27]: plt.scatter(fsd_white, zval)
      plt.axis((0, 300, 0, 10))
      plt.ylabel('quality')
      plt.xlabel('free sulfur dioxide')

[27]: Text(0.5, 0, 'free sulfur dioxide')
```

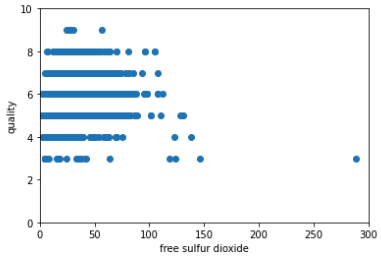


Figure 1: Scatter plot of free sulfur dioxide against quality for red (above) and white (below) wines

The next variable chosen was total sulfur dioxide, which once again had a small number of outliers, and the data was normally distributed, meaning the models did not have to deal with skewed data, as shown in figure 2.

Predictive Modeling of Red and White Wine Characteristics using Machine Learning Techniques: A Comparative Analysis

Chinedu Okeke & Emre Guvenilir

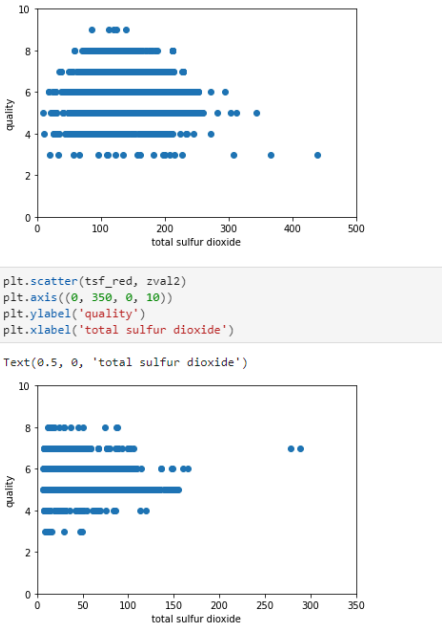


Figure 2: Scatter plot of total sulfur dioxide against quality for red (below) and white (above) wines

The final variable chosen was residual sugar, which resembles the same qualities explained for the previous two variables, as shown in figure 3.

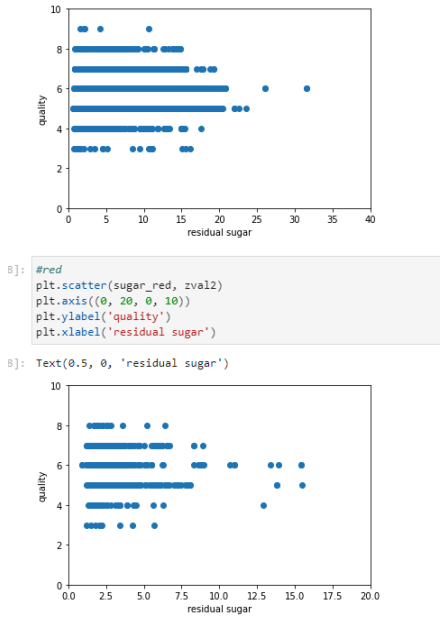


Figure 3: Figure 2: Scatter plot of residual sugar against quality for red (below) and white (above) wines

We decided on two models: a multivariate linear regression, and a decision tree regression model. A linear regression model will allow us to evaluate correlation among the chosen variables and if the combination of those variables are related to the dependent variable, quality. Decision trees essentially answer true or false questions till it reaches the leaf node, or the final node (Gurucharan M K, 2020). It begins with the root node, or the entire sample, and splits into interior nodes, of the features of a data set. The leaf node represents the outcome. Decision trees are also relatively easy to work with as it does not require normalization of data, and is one of the quickest ways to identify relationships between variables (Vadapalli, 2020).

Our hyperparameter choices were .8 for the training set and .2 for the test set. Since these are the widely accepted standard, and our dataset is a publically available, regularly updated dataset, we were confident to use those hyperparameter choices. This is also an attempt to prevent overfitting by not fitting all of the data on the model at one time. Thus attempting, but not guaranteeing, that the model isn't simply just memorizing the dataset itself. It is looking for patterns within the data to make predictions on any data, not just the training set.

Given we have included two different models in our report, the primary model we would use to make further predictions would be the Decision Tree Model, given that it seemed to have fewer accuracy and discrepancy issues in comparison to the

Predictive Modeling of Red and White Wine Characteristics using Machine Learning Techniques: A Comparative Analysis

Chinedu Okeke & Emre Guvenilir

Linear Model overall across both wines. With the hyperparameters we set, it seemed clear to us that the Decision Tree was a much better fit after meticulously going through both.

Results

For both the white and red wine datasets, a Linear Regression Model and Decision Tree were used to make predictions.

Using the Linear regression model for the white wine we had a Mean Squared Error Value of .741, with much of the predicted values concentrated around 5 and 6. With the same model red wine achieved a MSE of .644, with much of the predicted and actual data overlapping one another, with none of the predicted values falling onto non whole numbers (Figure 4).

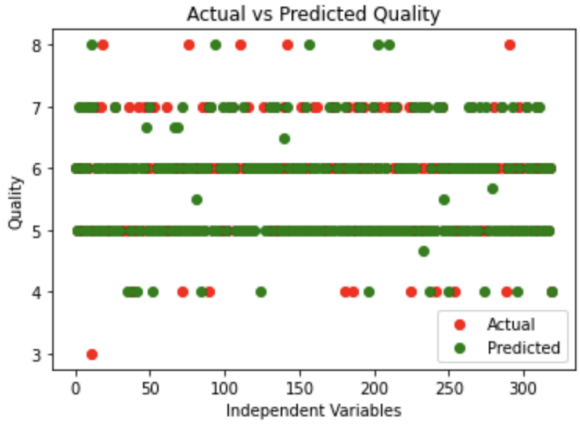


Figure 4: The predicted values for red wine produced by the Linear Regression model, overlapped with the actual values.

For the Decision Trees, a MSE value of .927 was produced for the white wine dataset (Figure 5), while the red wine produced a MSE of 1.019. Both datasets, similar to the Linear Regression Model for the red wine dataset, saw a lot of overlap between their predicted and actual values.

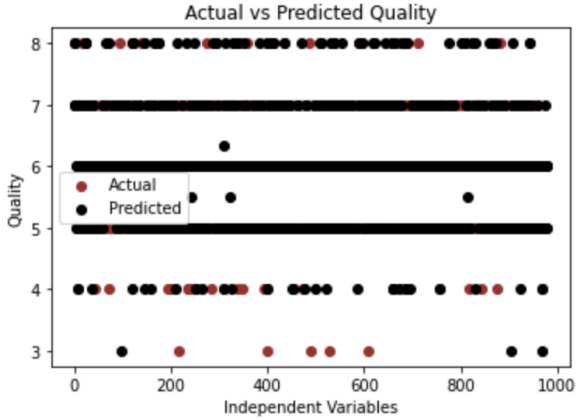


Figure 5: The predicted values for white wine produced by the Linear Regression model, overlapped with the actual values.

Analysis

To create the predicted values, we used two different models. Given the attributes we used as our independent variables (refer to Methods), we determined that the Decision Tree was a better fit for our data in comparison to the Linear Regression Model. This is seen when comparing the results between the two models for our white wine dataset. Despite the Linear Model producing a lower MSE, typically meaning a higher rate in precision and accuracy, the actual predictions were much further off in comparison to the predicted values made by the Decision Tree. This could be due to the presence of many outliers in the data, skewed data, or just simply that this Linear model was not a good enough fit for the presented data due to a lack of features. Given this reality, the fault might also be on us for the lack of accuracy in the model. The three independent variables we chose could have played a role into why the model faltered heavily. So despite our due diligence when picking them, there might be

a better combination of three variables that would have produced a better Linear Model for predictions for this white wine dataset. On the other hand, this massive inaccuracy was not seen when comparing the two models using the red wine data. From a visual standpoint it did not seem to be far off when comparing each of their predicted values to the actual values within the dataset. However, when the MSE of both was calculated, the Linear model produced a substantially lower value in comparison to the Decision Tree Model (.644 > 1.019). This would make one lean towards the decision to pick the Linear model over the Decision Tree Model. However it is also important to know that when calculating MSE, the idea of "low" score is relative to what data is being used and what models are being run, with there being no true universal low. On top of this, an MSE will not always tell the whole story (as seen with the Linear Model for White wine) and can at times be misleading or not be as statistically significant as we think in comparison to other values. This was a major reason as to why we wanted to also plot out the predicted vs. actual data points against each other to see how well they overlapped, as a contingency of sorts (Figure 4 & 5). Regardless, both the Linear Regression Model and the Decision Tree seemed to be good predictors for the red wine data set, when using the three independent variables we selected (total sulfur dioxide, free sulfur dioxide, and residual sugar).

Ethics

Using the NeurIPS Code of Ethics, it is clear that the contribution of this paper falls in line with the standards set. The data used does not include any human subjects or participants, does not have any identifiable information on the wine tasters, and the dataset is publicly available. The dataset used is not vulnerable to misuse as the identities of the wine tasters are unknown, and the names of any wines are not revealed, only the location, Portugal, where the wine was produced. When analyzing the results, it is important to consider that the quality ratings are subjective ratings given by tasters of the wine, and that different people can have their own quality ratings if they taste the wines. It is also important to note that these models do not promote alcohol consumption. These results were produced in a Jupyter environment using Python and libraries such as numpy, matplotlib, and sklearn for the models and regression implementations. All code is available in the 'csc371-machinelearning/project1-chinedu-emre' repository on github.

References

K GM (2020) Machine learning basics: Decision tree regression. In: Medium. https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda#:~:text=It%20can%20be%20used%20to,split%20further%20into%20further%20nodes. Accessed 14 Feb 2024

Vadapalli P (2020) Pros and cons of decision tree regression in machine learning. In: upGrad blog. https://www.upgrad.com/blog/pros-and-cons-of-decision-tree-regression-in-machine-learning/. Accessed 14 Feb 2024