OPTIMIZING WINE QUALITY PRODUCTION

Data analysis for quality products

Derek Osawaguan Ahamioje

Table of Contents

Derek O. Ahamioje
Email: ahamiojederek@gmail.com

1 INTRODUCTION

The subject of quality has always been a major discussion and an important part of every product. In medieval Europe, the importance placed on quality resulted in the formation of guilds of skilled craftsmen (Montgomery, 2020). Quality, which can be best viewed from the perspective of 'best suited for its purpose is the fulcrum for sustaining a business as well as boosting business. The monitoring of the quality of a product has been a long practice, but the use of statistical techniques/principles to control quality is a modern practice (Mitra, 2016). In this study, datasets on the physicochemical composition of red wine products are to be analysed to identify key features (physicochemical components) that can enhance the quality of wine products.

1.1 DATASET

The wine dataset used for this analysis can be viewed both as a regression problem and a classification problem (Wine Quality, 2018). Table 1 shows the features of the dataset which is composed of eleven (11) input variables, in this case, the physicochemical parameters and one (1) output variable which is the quality of the wine product. The data set has a total of 1599 data points (or wine samples). See the dataset here (Wine Quality, 2018).

Table 1: Dataset showing the features and first 5 datapoints

```
[4] # represent data
    wine.head()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Derek O. Ahamioje
Email: ahamiojederek@gmail.com

2.0 METHOD

The analytical method adopted in this study is Exploratory data analysis, Data pre-processing, statistical techniques, and visualisation. All of these were done using google colab python notebook.

2.1 EXPLORATORY DATA ANALYSIS

In the first step, the dataset having been read to the dataframe was inspected to see if there is any form of anomaly like null values or outliers (knowing that these will impact the analysis negatively). To get this done, the following functions and methods were us

i. Describe function: There was the need to understand the statistical distribution of the dataset, this will help us to understand the summary of the central tendencies and dispersion of the dataset.

ii. Quality class: Having our target of the Quality feature, we also attempted to view the various quality class and the total count of each parameter in each class. The reason for this is to find out if there is an even distribution of data points across the quality class because if it is not even there will be an imbalance in our statistical test which will result in a false result. The result obtained shows that there is wide variation in all the quality class. The output was visualized using a bar graph and pie chart to show the percentage distribution of the data for each quality class. Figure 1 shows a bar graph of the observed distribution.

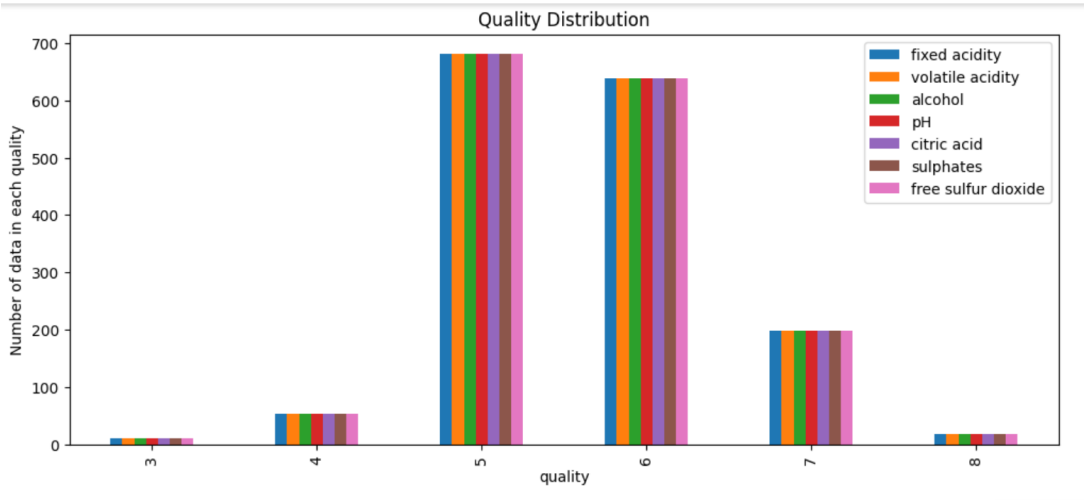

Figure 1: Bar graph showing the total data point distribution per quality class

iii. Correlation function: There was also the need to understand the level of interaction among the features. This will help identify the features without correlation and hence would not be needed in our analysis. For clarification purposes, Heatmap

Derek O. Ahamioje
Email: ahamiojederek@gmail.com

generated from the correlation function was generated. It is observed as shown in Figure 2 that all features share at least one correlation.

iv. Feature variation with quality was another attempt to see the interaction between each feature and the quality output. Visualising this output gives us an idea of the various features that will give the best quality. Figure 3

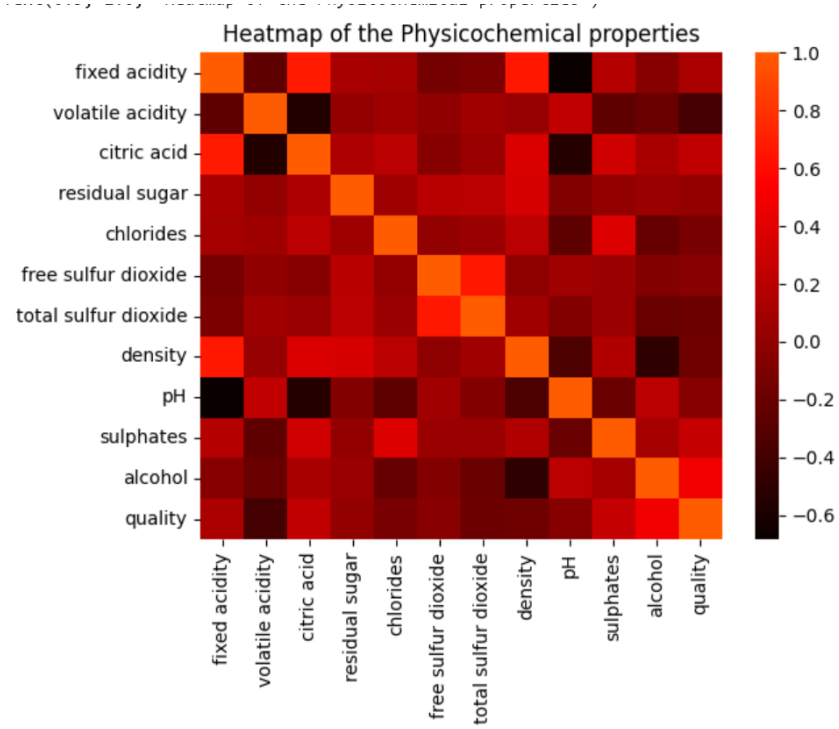

Figure 2: A heatmap showing the correlation among features.

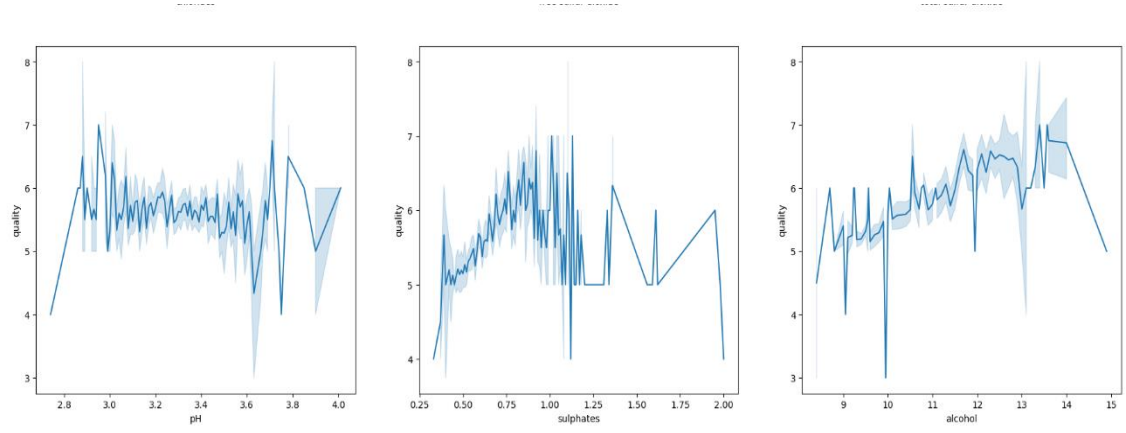

Figure 3a. A plot of the relation between quality and concentration of each physicochemical property (pH, Sulphates, and Alcohol)

Derek O. Ahamioje
Email: ahamiojederek@gmail.com

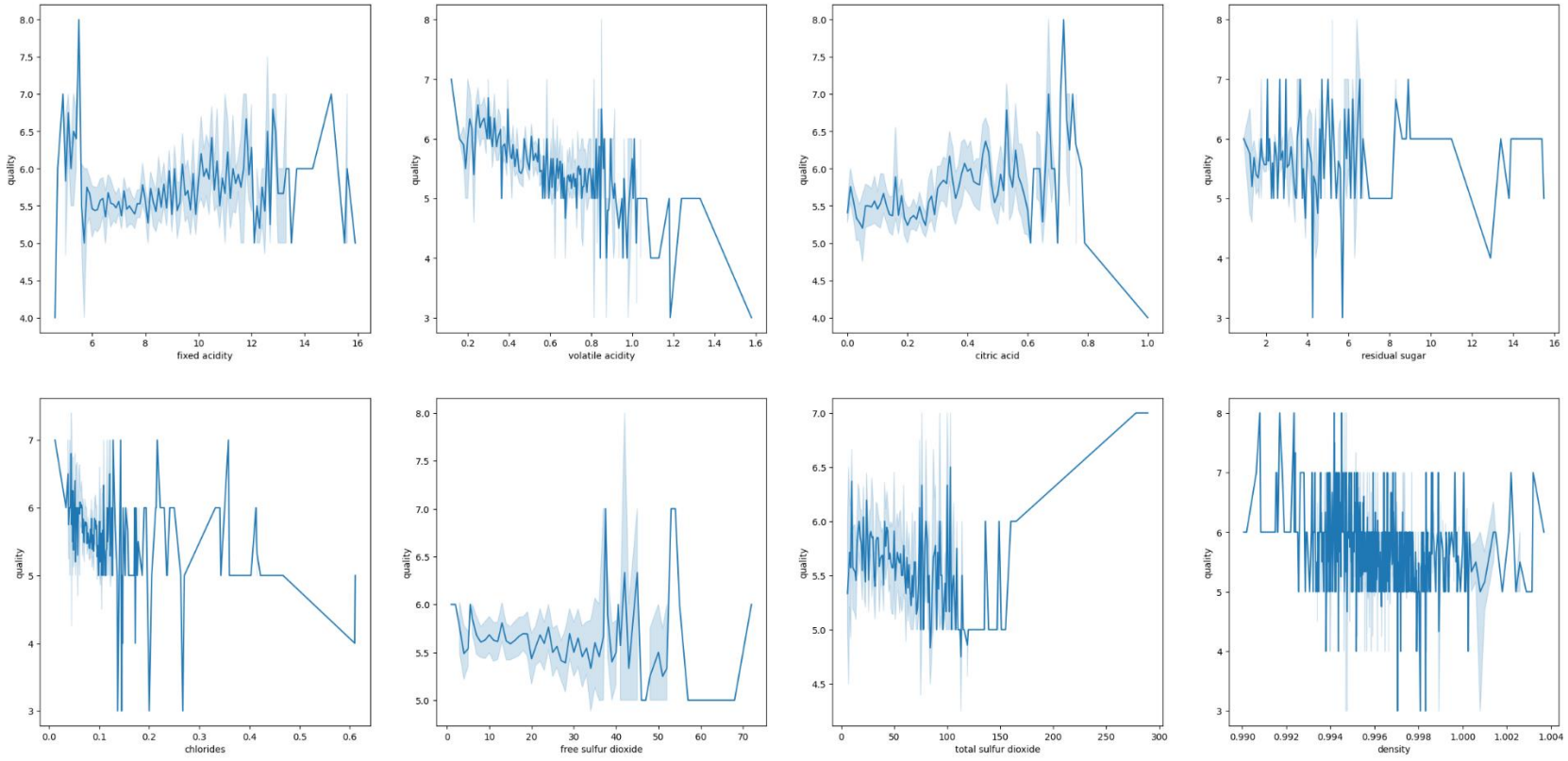

Figure 3: Plot of the relationship between quality and concentration of each physicochemical property

Derek O. Ahamioje
Email: ahamiojederek@gmail.com

2.2 DATA PREPROCESSING

The observations from the exploratory data analysis have satisfied the fact that there are interactions among the features of the dataset there is the need to enhance the quality of the result by further ensuring that the dataset is equally of quality through the following standard techniques or steps, Data cleaning, Data integration, Data transformation, Data reduction and Data discretisation (Yang, 2018). Given the dataset, certain specific pre-processing tasks were chosen and bearing in mind that some pre-processing, like removing null values have been addressed at the data exploratory stage. However, the following steps were taken to put the dataset at best for further analysis:

i. SMOTE Function: The discrepancy observed at the exploratory analysis phase as shown in table 3a, where the number of data points for each quality class varied differently, needs to be addressed and so the Synthetic Minority Oversampling Technique function was used to increase the number of cases in the dataset in a balanced way. Table 3b shows the output after resolving the inconsistency with the SMOTE (). The SMOTE function successfully blew the datapoints from 1599 datapoints to 4086 datapoints by raising other points' values, up to the value of the maximum points found in quality class five, which is 681. However,

Table 3a: datapoints for each quality class before using SMOTE.

quality	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
5	681	681	681	681	681	681	681	681	681	681	681
6	638	638	638	638	638	638	638	638	638	638	638
7	199	199	199	199	199	199	199	199	199	199	199
4	53	53	53	53	53	53	53	53	53	53	53
8	18	18	18	18	18	18	18	18	18	18	18
3	10	10	10	10	10	10	10	10	10	10	10

Table 3b: After resolving the inconsistency.

Derek O. Ahamioje
Email: ahamiojederek@gmail.com

```
x.count()

fixed acidity           4086
volatile acidity        4086
citric acid             4086
residual sugar          4086
chlorides               4086
free sulfur dioxide     4086
total sulfur dioxide    4086
density                 4086
pH                      4086
sulphates               4086
alcohol                 4086
dtype: int64
```

ii. MinMaxScalar(): To further ensure consistency for better model results, the minmax scalar was used to normalise the scale of the dataset to a uniform scale, such that they are all within the same scale.

3.0 PREDICTION MODEL

Now that the quality of the dataset has been sufficiently improved, the next step is the building of machine learning models that will help extract more useful information like the possibility of predicting the quality of wine from the input features, determine the extent to which this can be done and also identifying key features that would play an active role in determining or enhancing the quality of the wine product. This is very important because it is the core interest of this study that will help improve the business of wine production.

Considering that the dataset can be viewed as regression and classification model. So, we performed regression and classification analysis on the dataset to see how well our training dataset will predict the test set.

3.1 MULTIPLE REGRESSION

Multiple regression models were used to predict the quality of wine, compared to other models:

i. Ridge regression (which can best handle challenges resulting from the multicollinearity of the dataset). This is necessary, especially having observed from the heatmap at the exploratory stage that all the features correlate. So, if that is the case, we will use the Ridge regressor to see if there can be an improvement in the accuracy score

ii. Lasso regression – We used the Lasso regression to see if we can further enhance the prediction by regularizing the datapoints towards a central point

3.2 FEATURE SELECTION

Furthermore, the selectkbest was sued to select features according to the highest k score, this way features of importance can become a key focus for prediction.

Other algorithms used in the analysis are, Decision tree model to determine maximum tree depth for the best R2 score and XGBoost regression.

Other models used include Decision tree and XGBoost.

3.3 CLASSIFICATION MODEL

The second approach to the prediction is classification model. Three classification models were used and they include Logistic regression, Support vector machine and k nearest neighbour and their results compared.

4.0. EVALUATION OF RESULT

Reviewing the results obtained from the analysis of the dataset, figure 3a and Figure 3b has shown how all the eleven input features interact with the output feature by showing the quality variations with various adjustment in the concentration of the input features, this is also in agreement with the heatmap plot shown in figure 2. This affirms the fact that it is possible to adjust the input variables for optimum-quality products.

Furthermore, the result obtained from the statistical model shows that the multiple regression model has a 74 percent R^2 score which is good, but it also has a high Mean Square Error (MSE) of 0.74, which means although the independent variable can predict 74 percent of the output, the risk factor with the model is too high. Further attempts to improve check the performance with Ridge did not yield any considerable improvement, while Lasso regression performed very badly. These values are shown in Figure 4.

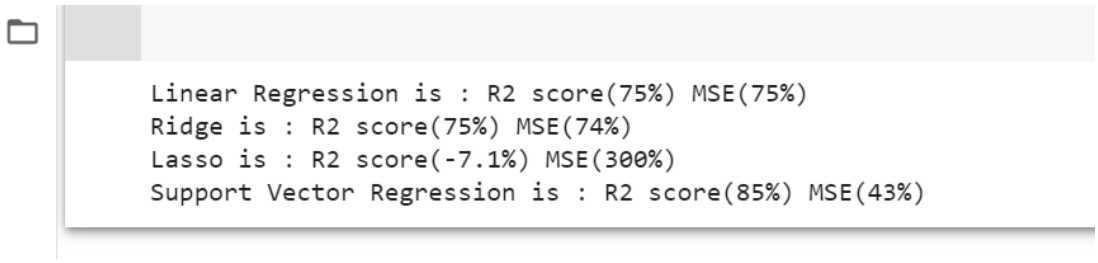

Figure 4: Regression model evaluation

With the Feature selection regression, there was a reduction in the R2 score from 74% to 70% and an increase in Mean Square Error from 74% to 87%. This shows that selecting features of importance (Figure 5) for predicting output did not improve the model but rather gave a poorer performance when compared with the multiple regression model. However, The Decision Tree model performed better than the multiple regression model when the max depth parameter was set to default, resulting in an 85% R2 score and a huge drop in MSE to 44%. Attempt to boost the model performance further with XGBoost yielded a positive result as seen in Figure 6, where the R2 score increased to 95% and MSE dropped further to 15%

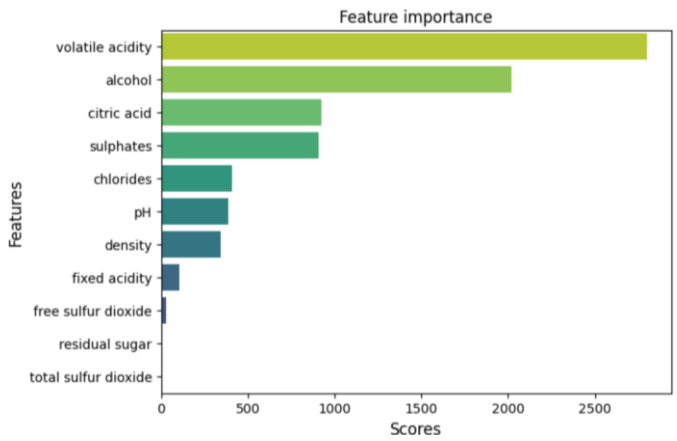

Figure 5: Identified features in order of their importance

Table 3: Results obtained from different models used

	model	r2_score	mean_squared_error
0	Multiple regression	0.74	0.74
1	Regression (feature selection)	0.70	0.87
2	Decision Tree	0.85	0.44
3	XGBoost	0.95	0.15

The results obtained from the classification model shows that a test set accuracy of 60% was obtained from the logistic regression model, and 64% with SVM, on applying GridsearchCv, the SVM test set accuracy did not improve further. From the K nearest neighbour, the test set accuracy was 61%. Table 4 Shows the results obtained.

Table 4: Test accuracy obtained from classification models

	model	metrics.accuracy_score
0	LR	0.60
1	SVC()	0.64
2	SVC()	0.64
3	KNeighborsClassifier()	1.00

Derek O. Ahamioje
Email: ahamiojederek@gmail.com

5. CONCLUSION

At the end of the analysis, it is however seen that even though some features of the data set such as volatile acidity, Alcohol, Citric acid, and Sulphates had scores above 1000 on features importance gradient, they could not be isolated for optimum quality predictions. On the other hand, both regression and classification models worked for the prediction but the regression model had a prediction higher score than when classification model was used.

Derek O. Ahamioje
Email: ahamiojederek@gmail.com

REFERENCES

Mitra, A., 2016. *Fundamentals of quality control and improvement*. John Wiley & Sons. Pg 4-6

Montgomery, D.C., 2020. *Introduction to statistical quality control*. John Wiley & Sons. Pg 9

Wine Quality (2018). Available at: https://www.kaggle.com/datasets/rajyellow46/wine-quality.

Yang, H., 2018. Data pre-processing. *Pennsylvania State University: Citeseer*.