Project:

Prediction of red wine quality using Machine Learning

Abstract

The quality of red wine is important for consumers as well as the wine industry's point of view. The traditional way of analyzing the quality of wine is time-consuming and expensive. There are various physicochemical features that affect the quality of the wine. Nowadays, many algorithms based on machine learning techniques have been proposed for predicting the quality of wine to replace these human tasks. The main objective of this study was to find out what wine features are important to get a good quality wine. For this purpose, we used a machine learning algorithm decision tree, random forest model, logistic regression, and kNN. We used the Portuguese "Vinho Verde" wine data set for building the model. The wine data set contained 1599 instances with 11 input physicochemical properties and 1 output (quality) variable. Different exploratory data analysis was performed for data visualization. To evaluate the best model, I performed a confusion matrix and evaluated the precision, accuracy, f1-score, ROC curve, and recall of all the models. The logistic regression model displayed a higher accuracy, F1 score, recall, and sensitivity compared to other models. These scores were 96.4, 98.17, 99, and 99 % respectively for the logistic regression model. But the AUC value for logistic regression was only 50%. The most significant physicochemical properties affecting the quality of wine were alcohol, sulfates, and fixed acidity. Though the accuracy was not the highest (77.49%), the random forest was best in detecting different classes of dependent variables as ROC was highest 82.9%.

Keywords: quality of wine, machine learning, algorithms, decision tree, random forest, logistic regression, kNN, accuracy, F1 score, recall, sensitivity, physicochemical properties.

Table of Contents

Introduction

Wine is the most commonly used beverage on the dining table and there is growth in the wine industry as social drinking is on the rise. State of the World Vitivinicultural sector report suggests that worldwide wine consumption is estimated at 234 million hectoliters, where 48% of the world consumption accounts in European countries (OIV, 2021). According to FAO (2008), Portugal is a top ten wine exporting country which contributes 3.17% of the market share. The export of "Vinho Verde" wine increased by 36% from 1997 to 2007 (CVRVV, 2008).

The quality of wine plays important role in the wine industry to be competitive in the market and increase revenue. The chemical composition of wine determines the wine's color, flavor, fragrance, and other characteristics (Sousa et al., 2014). Each chemical component such as volatile compounds provides wine fragrance and phenolic compounds give its flavor (Sousa et al., 2014, Waterhouse et al., 2016). The flavor and fragrance quality of wine plays a big role in its quality and consumers' preference. Consumers are serious about the quality while purchasing wine.

Previously, wine quality testing used to be at the end of the production process, and this process is time-consuming and it requires the resources such as experts which makes the process very expensive. The major challenge for the wine industry is to determine the wine quality based on human experts as every human has their own opinion about the test. The wine industry is searching for new technologies to make better quality wine. Machine Learning has been successfully used in predicting sales, price, etc. Computer modeling is key for improving the efficiency accuracy and efficiency of the prediction. Hence, by developing the predictive model using the Machine Learning Algorithm wine industry can figure out the importance of physicochemical properties for good quality wine and which one to ignore for lowering the wine cost of production.

Research Question

My research question for this study was to find out which wine features are important to get good quality wine by implementing machine learning algorithms. The physicochemical properties of wine affect the quality of the wine. The traditional method of analyzing quality is time-consuming and expensive. So, there is a need to identify the methods to predict the quality of wine to reduce the cost of wine production. I will consider comparing the accuracy of different machine learning algorithm models for predicting the quality of the wine.

Data Appraisal

Data Sources

For this study, the data taken for the analysis is related to Portuguese "Vinho Verde" wine. It is a wine company situated in Portuguese. The dataset is available from the machine learning repository, https://archive.ics.uci.edu/ml/datasets/wine+quality. The wine data set contains 1599 instances with 11 input variables (physicochemical properties): fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and 1 output variable (quality) ranked from 0 to 10. Below are the descriptions of the variables used:

- Fixed acidity: The predominant fixed acids in wine, such as tartaric, succinic, citric, and malic acids.
- Volatile acidity: The high acetic acid present in wine, which causes an unpleasant vinegar taste.
- Citric acid: A weak organic acid used to increase the freshness and flavor of the wine.

- Residual sugar: The amount of sugar left after fermentation.
- Chlorides: The amount of salt in wine. The lower chloride rate creates better quality wines.
- Free sulfur dioxide: SO_2 is used for preventing wine from oxidation and microbial spoilage.
- Total sulfur dioxide: The amount of free and bound forms of SO_2.
- Density: Depends on the alcohol and sugar content. Better wines usually have lower densities.
- pH: Used to check the level of acidity or alkalinity of wine.
- Sulfates: An antibacterial and antioxidant agent added to the wine.
- Alcohol: The percentage of alcohol in wine. A higher concentration leads to better quality.

The output variable, quality score, ranges from 1 to 10, where 1 means the worst score and a score of 10 means the best. In total there are 12 variables, 11 of them consist of input variables and 1 is a target variable. The data types of all the variables are numeric except the target variable (int type). Our model will be predicting which features are the most indicative of good wine quality. The dataset contains 1599 observations. These features are crucial for getting the most accurate and reliable predictions from a machine learning model.

Data Preparation

The data from the source was obtained in .csv format, where all the variables were the column of the dataset, and all the properties of each wine sample are the rows of the dataset. The data were loaded into R, data modeling was performed using different R packages.

All the variables have outliers, variables “residual sugar” and “chlorides” are the variables that have the most outliers. These variables have outliers greater than 5%., i.e., 9.7% and 7% respectively. I chose to change the outlier values with the mean value of the variables because, as we can see in the histogram plot, both variables have a large concentration. Quality greater than equal to 7 was classified as 1 (good quality) and the rest categorized as 0 (bad or mediocre quality) of value near the mean. But for better statistical analysis, I chose to use 2 categorical target variables in the project. I do not have any missing values in the dataset.

While splitting the dataset into the Training set and Test set. The training dataset has 80% of the observations that represent a good quality wine to balance the train set. In other words, the dependent variable will have the same number of observations of 0 and 1 in the training dataset. This was done to remove biases.

Decision Tree Model

Using the library(rpart) in R, the decision tree was created using the input variables alcohol, sulphates, fixed acidity, and output variables alcohol quality.

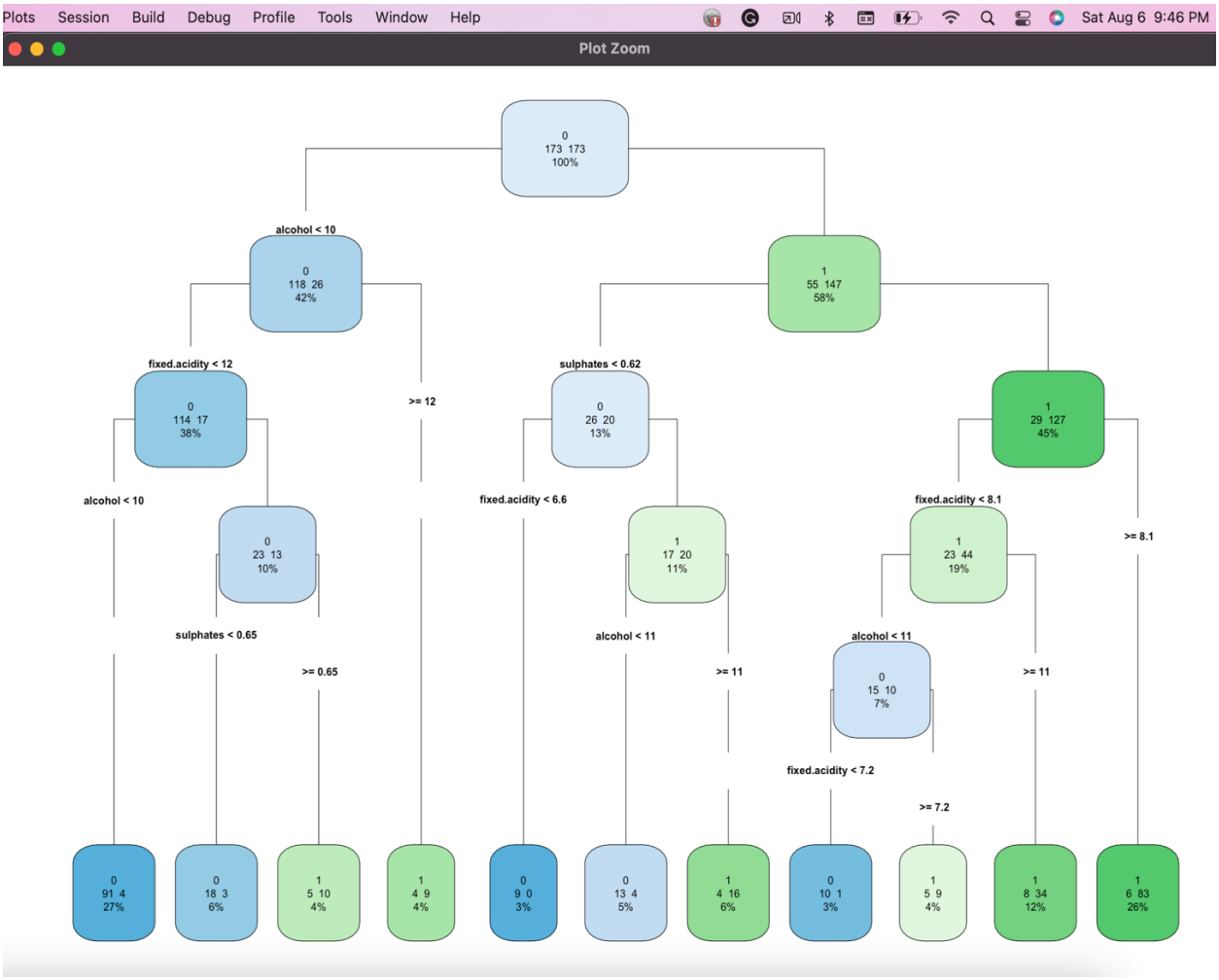
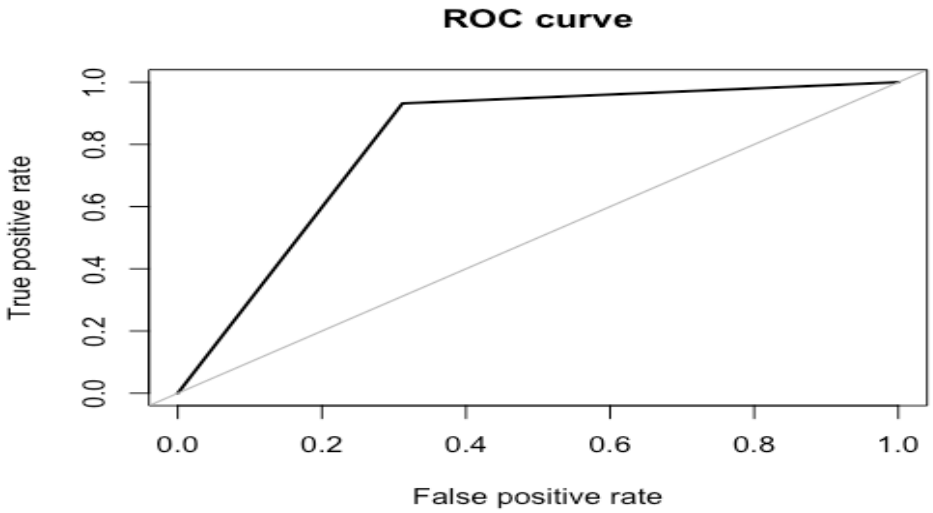


Figure 1. The decision tree model output from R using rpart library. Within each box, the top number represents the quality of alcohol, where 0 represents bad quality whereas 1 represents good quality. The middle numbers represent the number of bad quality wines and numbers of good quality wines respectively, and the bottom number represents the proportion of the total sample at that node.

Starting from the root node (node 1) the decision tree shows for all the training dataset, 173 samples were placed at 0 (bad quality wine), and 173 samples were placed at 1 (good quality). The root node then divides on alcohol. Alcohol value less than 10 (node 2) and alcohol value more than 10(node 3). 58% of the total sample exceeds this amount and 42% did not. Node 2 has predominant bad quality wine (118 samples) whereas node3 had more good quality (147 samples). From this distribution it is reasonable to say that to be good quality wine, alcohol should be more than 10. The node 2 splits on fixed acidity less than 12 (node 4) and fixed acidity more than equal to 12(node 5). Terminal node 4 has 27 % of total data of which 91 samples were bad quality. From this observation, we can conclude that fixed acidity less than 12 is determining factor for having wine as bad quality. Node 5 splits on sulfates less than 0.65 or more than equal to 0.65. Both are leaf nodes and consist of 6% and 4 % data respectively. Node 3 also splits on the basis of same parameters of sulphates. Altogether there are 11 terminal nodes.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 832   3
         1 377  41

               Accuracy : 0.6967
                 95% CI : (0.6704, 0.7221)
    No Information Rate : 0.9649
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1217

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.68817
            Specificity : 0.93182
         Pos Pred Value : 0.99641
         Neg Pred Value : 0.09809
              Precision : 0.99641
                 Recall : 0.68817
                     F1 : 0.81409
             Prevalence : 0.96488
         Detection Rate : 0.66401
   Detection Prevalence : 0.66640
      Balanced Accuracy : 0.81000

       'Positive' Class : 0
```

The accuracy of the decision tree was 69.6%, F1 score-81.4%, recall-68.8%, sensitivity-68.8%.



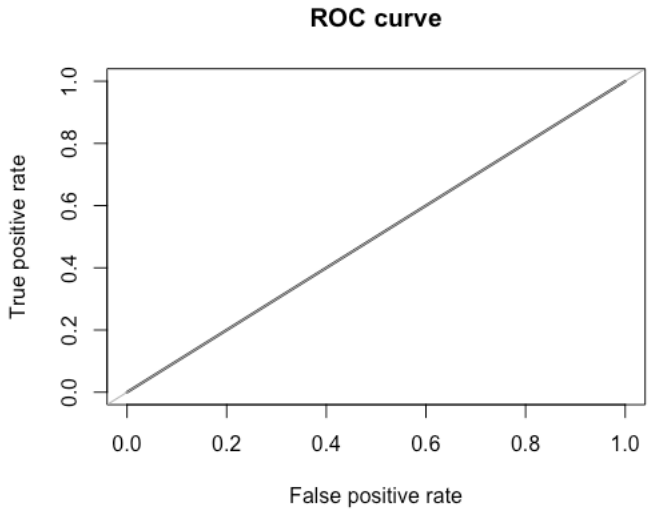

```
> library(ROSE)
> par(mfrow = c(1, 1))
> roc.curve(test_set$quality, y_pred)
Area under the curve (AUC): 0.810
>
```

AUC of the decision tree was 81.1.

Logistic-regression

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 832   3
         1 377  41

               Accuracy : 0.6967
                 95% CI : (0.6704, 0.7221)
    No Information Rate : 0.9649
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1217

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.68817
            Specificity : 0.93182
         Pos Pred Value : 0.99641
         Neg Pred Value : 0.09809
              Precision : 0.99641
                 Recall : 0.68817
                     F1 : 0.81409
             Prevalence : 0.96488
         Detection Rate : 0.66401
   Detection Prevalence : 0.66640
      Balanced Accuracy : 0.81000

       'Positive' Class : 0
```

Logistic regression had an accuracy of 96.4%, which was pretty good. F1 score of 98.17%, recall 99%, sensitivity-99%, specifity-2.2%. But AUC was only 50.0%.

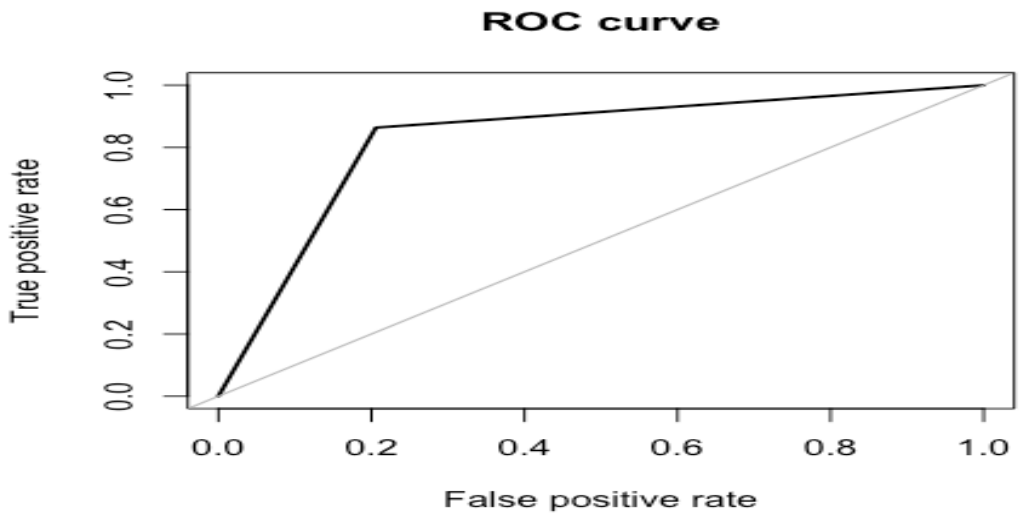



```
> library(ROSE)
> par(mfrow = c(1, 1))
> roc.curve(test_set$quality, y_pred)
Area under the curve (AUC): 0.500
>
```

Random forest:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 934   7
         1 275  37

               Accuracy : 0.7749
                 95% CI : (0.7508, 0.7978)
    No Information Rate : 0.9649
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1559

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7725
            Specificity : 0.8409
         Pos Pred Value : 0.9926
         Neg Pred Value : 0.1186
              Precision : 0.9926
                 Recall : 0.7725
                     F1 : 0.8688
             Prevalence : 0.9649
         Detection Rate : 0.7454
   Detection Prevalence : 0.7510
      Balanced Accuracy : 0.8067

       'Positive' Class : 0
```

Random forest has accuracy-77.49%, F1score-86.6%, sensitivity-77.25%, recall-77.25%



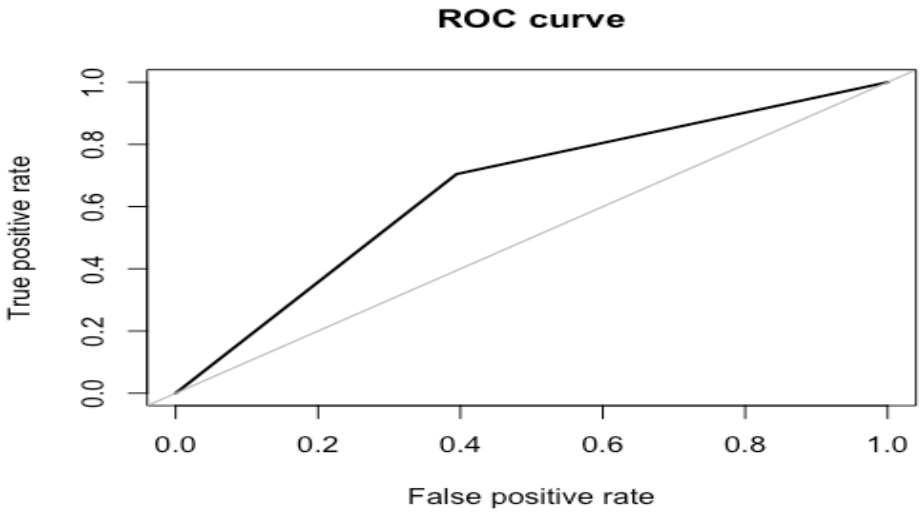

```
> library(ROSE)
> roc.curve(test_set$quality, y_pred)
Area under the curve (AUC): 0.829
>
```

kNN

kNN has accuracy:60.1%, sensitivity:60.5, specificity:70.5, recall:60.5, F1score:74.92

wineknns<- knn(train = train_set1, test=test_set1, cl=wineTrain_label, k=19)

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 732  13
         1 477  31

               Accuracy : 0.6089
                 95% CI : (0.5813, 0.6361)
    No Information Rate : 0.9649
    P-Value [Acc > NIR] : 1

                  Kappa : 0.051

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.60546
            Specificity : 0.70455
         Pos Pred Value : 0.98255
         Neg Pred Value : 0.06102
              Precision : 0.98255
                 Recall : 0.60546
                     F1 : 0.74923
             Prevalence : 0.96488
         Detection Rate : 0.58420
   Detection Prevalence : 0.59457
      Balanced Accuracy : 0.65500

       'Positive' Class : 0

>
```




```
> library(ROSE)
> roc.curve(wineTest_label, wineknns)
Area under the curve (AUC): 0.655
>
```

AUC of kNN is 65.5%

Conclusion

Among the different models I tried, the logistic regression model displayed a higher accuracy in predicting the quality of red wine, an accuracy of 96.4%, this model was able to predict correctly 1,208 values, meaning that the misclassification error of the model was 3.6%. In addition f1 score, recall, and sensitivity were highest in logistic regression in comparison to other models. By evaluating the area under the curve (AUC = 50.0%), I labeled the ROC curve as a fail curve. I concluded that even though the model has good accuracy in predicting the test set values, it has a pitful rate of quickly identifying true positive values. Since logistic regression has the highest accuracy, I have calculated the variable importance of logistic regression

```
> varImp(lr)
                      Overall
fixed.acidity        2.51930944
volatile.acidity     2.37396068
citric.acid          1.57550922
residual.sugar       0.84578158
chlorides            1.21463846
free.sulfur.dioxide  0.87407754
total.sulfur.dioxide 1.94731779
density              0.17070960
pH                   0.08739752
sulphates            3.72332662
alcohol              4.88516494
```

The most significant variable for this model is “alcohol”, followed by the variables “sulphates” and “fixed acidity”. The sulphate is the component of the wine that is responsible for the freshness of the drink, sulphate gives more control over the life of the wine since it helps to ensure the wine will be fresh and clean when opened. The alcohol will help balance the firmer and acid taste of the wine, making an interrelationship of the hard and soft characteristics of the wine. For this reason, the variables “alcohol” and “sulphates” are very significant to the model, if one of these variables changes, the results of the model will be affected strongly.

From the ROC curves, I revealed that the random forest did have the best performance by obtaining an area under the curve of 82.9%, random forest model did not display the highest accuracy among

the four models, it has the best performance by detecting the different classes of the dependent variable better than the logistic regression.

I twisted my training dataset by equally dividing it into the good and bad quality of the target variable in the training dataset to correct the imbalance in the data. Furthermore, the extreme outlier variables residual.sugar and chlorides were changed to their mean value before analyzing the models to remove biases in the dataset. I have changed outliers of those variables to their mean value as most of the values are distributed towards their mean value. In addition, I also calculated f1 scores, recall, precision, and roc curve in this milestone for analyzing the best model.

References

[OIV] International Organisation of Vine and Wine. (2021). State of the World Vitivinicultural sector in 2020.

FAOSTAT — Food and Agriculture Organization Agriculture Trade Domain Statistics (July 2008)

CVRVV. Portuguese Wine — Vinho Verde. Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV), http://www.vinhoverde.pt, July 2008.

Sousa, E. C., Uchôa-Thomaz, A. M. A., Carioca, J. O. B., de Morais, S. M., de Lima, A., Martins, C. G., Alexandrino, C. D., Ferreira, P. A. T., Rodrigues, A. L. M., Rodrigues, S. P., Silva, J. do N., & Rodrigues, L. L. (2014). Chemical composition and bioactive compounds of grape pomace (Vitis vinifera L.), Benitaka variety, grown in the semiarid region of Northeast Brazil. Food Science and Technology, 34(1), 135–142.

Waterhouse, A. L., Sacks, G. L., & Jeffery, D. W. (2016). *Understanding wine chemistry*. John Wiley & Sons.

Kniazieva, Y. (2022, February 15). *Sommelier of the Digital age*. High quality data annotation for Machine Learning. Retrieved August 13, 2022, from https://labelyourdata.com/articles/machine-learning-for-wine-quality-prediction

Data Analysis Reference

Excel data set

R code

EDA Diagram