Prediction of Wine Quality: Comparing Machine Learning Models in R Programming

Olatunde David Akanbi[1*], Taiwo Mercy Faloni[2], Sunday Olaniyi[3]

[1]*Department of Material Science and Engineering, Case Western Reserve University, United States.*
[2]*Department of Biology, Case Western Reserve University, United States.*
[3]*Department of Microbiology, Obafemi Awolowo University.*
**Corresponding Author*

***Abstract*: The consideration of wine quality before consumption or use is not a new decision scheme across ages, fields, and people. Gone were the days when quality of wine solely depended on taste or other physical checks. In this age of data science and machine learning, we can make decisions on the best wine quality with reference to different features/variables. This work was done with in predicting the dependent variable while using existing models to analyze the independent variables. This work utilizes the R programming language for this prediction, while comparing different machine learning models like Linear regression, Neural network, Naive Bayes Classification, Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), k-Nearest Neighbors (kNN), Support Vector Machines (SVM) with a linear kernel, and Random Forest (RF). The provided data was divided into the testing and training portions with parts for validation. It was achieved that Random Forest has a better model for this prediction when cross cross-validated in 10-folds. The accuracy was then used to select the optimal model. Hence, alcohol is the feature variable that contributes more to wine quality while volatile acidity and chloride contribute the least to the quality of wine. This would assist breweries in determining the right additions and subtraction when wine quality is in question.**

***Keywords*: Random Forest, R programming, Machine learning models, Wine quality, Algorithms, RF**

I. INTRODUCTION

Industries, especially the chemical industries have evolved over the years in determining values and analyzing data. Vast collaborations are vast permeating the research atmosphere. Wine is regarded as the commonest beverage across cultures and fields in the global sphere, and its values are weighed by different features in different societies [1]. Determining the quality of wine is always of great interest to both researchers and consumers. Wine quality are generally determined by two basic tests; which are the sensory tests and the physicochemical tests. Considering the fact that physicochemical test is a laboratory test, with no human expertise required and sensory test requires human expertise; researchers come into a difficult terrain in determining the quality of wine [2] [3]. Complex data analysis as wine quality assessment feature therefore requires a better approach for full understanding. Racing through the lane of history, determining wine quality is expensive and time consuming [1].

This century doesn't only allow for collaboration among seemingly contrasting fields, but gives a push for a paradigm shift in technological advancements and tool utilization [4]. Data scientists, computer scientists, chemical engineers, material engineers and others can seamlessly work together for better research in determining the quality of wine today. Utilizing object-oriented programming like python has been of great benefits in the industry [5]. R programming is not only instructive in the academic world, but produces advance results for optimal predictions in the industrial sphere as well. Hence, R is also used for the machine learning prediction and that has been established in this research with the consideration of different algorithms and easy to use packages in CRAN.

II. GENERAL OVERVIEW

Machine learning predictions have become easier through the advent of different algorithms and ML models. With much on supervised and unsupervised learning, researchers can now make right predictions through the right and available tools. Engineers and scientists need the right tools in decision making, proper data interpretation and statistical analysis. R among other programming tools is highly effective in data analytics and machine learning. In the fields, much of the machine learning works have been on the door of Python programming and the available packages. Interestingly however, R provides another strong reinforcement in machine learning and statistics. Statistical data is getting the proper attention through these tools, and statisticians are computer scientists to achieve feats.

The data for this work was provided by Kaggle, as uploaded on 15th January, 2022 [6]. R programming tools were utilized to achieve the correct prediction of the quality of wine. Knowing the right package to install in R for data analysis is very pivotal to the success of any research work with it. R programmers have a bunch of similar packages that perform similar operations. Choice, desire and flexibility are few of the factors that inform the use of a package. Linear regression, Neural network, Naive Bayes Classification,

Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), k-Nearest Neighbors (kNN), Support Vector Machines (SVM) with a linear kernel, and Random Forest (RF) were considered for this project as provided by the R packages in use. Different data visualization scheme was also done.

III. LITERATURE REVIEW

With advent of machine learning models in the recent years, determining wine quality have gone through different analysis and modelling [21]. Some previous works of Moreno et. al was to categorize 54 samples of a particular wine using the popular probabilistic neural network [18]. Authors Yu et al. also worked on something related. 147 bottles of rice wine were analyzed and estimated using spectral measurements [19]. Three different Chilean wine was classified by Beltran et al. [20] using basically three machine leaning models; SVM, linear discriminate analysis and neural network.

Wine has been initially analyzed through the use of electric nose [7]. Different wines could also be classified through the use of taste sensor as applied to neural networks [8]. Larkin used stacked generalization to predict wine preferences. There have also been previous application applications of machine learning in wine prediction, recommendation and classification [10][11][12]. Despite all these works, R as a statistical and machine learning tool has often been neglected in the previous works [13]. According to Hackenberger's research in 2020, there was a general believe that R is unfriendly, but probably the best tool. R is not just open source but easily accessible. It was stated that the power of R is closely based on the availability of packages with functions, algorithms, and flexibility [14]. A major comparison initially done compared R, Python and SAS, and found R efficient but often neglected [15]. R has always been on the frontline for teaching purposes. It seems industries over the years are glued to other tools and would not explore the beauties of R [16]. Hence, bringing the machine learning ability of R is part of the focus of this work.

IV. THE DATASET

This dataset gearing towards the prediction of wine quality is related to red variants of the Portuguese "Vinho Verde" wine [6]. According to a report in 2021, USA consumers have grown to love this type of Portuguese wine, and almost becoming a household name [17]. It basically describes the various chemicals present in wine and their possible effects on the quality

This dataset as obtained from Kaggle has 13 columns and divided into three subsections:

a) The input variables by physicochemical tests:

1. Fixed acidity: This is the non-volatile acid in the wine
2. Volatile acidity: This the amount of acetic acid in the wine
3. Citric acid: This adds flavor to a wine
4. Residual sugar: This is the amount of sugar in the wine
5. Chlorides: This is the amount of salt in the wine
6. Free sulfur dioxide: This is in wine to prevent microbial growth/oxidation.
7. Total sulfur dioxide: This is total presence of sulfur dioxide in the wine
8. Density: The density of the substance
9. PH: This is the measure of acidity and basicity of the wine- from 0-14
10. Sulphates: Added to wine to aid the supply of sulfur dioxide
11. Alcohol: Percentage of alcohol in the wine

b) Output variable by sensory data:

12. Quality: Quality of the wine score between 0 and 10

c) The wine ID:

13. Id: The label on the wine

I. The Dataset at a Glance

This data has a dimension of 1143 x 13. As displayed in Figures 1, 2, 3, and 4, the summary of what the dataset entails could be seen. The datatypes, measures of central tendencies and other features are displayed.

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides
1	7.4	0.70	0.00	1.9	0.076
2	7.8	0.88	0.00	2.6	0.098
3	7.8	0.76	0.04	2.3	0.092
4	11.2	0.28	0.56	1.9	0.075
5	7.4	0.70	0.00	1.9	0.076

	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol
1	11	34	0.9978	3.51	0.56	9.4
2	25	67	0.9968	3.20	0.68	9.8
3	15	54	0.9970	3.26	0.65	9.8
4	17	60	0.9980	3.16	0.58	9.8
5	11	34	0.9978	3.51	0.56	9.4

	quality	Id
1	5	0
2	5	1
3	5	2
4	6	3
5	5	4

Figure 1: The first five lines of the dataset

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides
1139	6.3	0.510	0.13	2.3	0.076
1140	6.8	0.620	0.08	1.9	0.068
1141	6.2	0.600	0.08	2.0	0.090
1142	5.9	0.550	0.10	2.2	0.062
1143	5.9	0.645	0.12	2.0	0.075

	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol
1139	29	40	0.99574	3.42	0.75	11.0
1140	28	38	0.99651	3.42	0.82	9.5
1141	32	44	0.99490	3.45	0.58	10.5
1142	39	51	0.99512	3.52	0.76	11.2
1143	32	44	0.99547	3.57	0.71	10.2

	quality	Id
1139	6	1592
1140	6	1593
1141	5	1594
1142	6	1595
1143	5	1597

Figure 2: The last five lines of the dataset

The Summary of the data

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 4.600   Min.   :0.1200   Min.   :0.0000   Min.   : 0.900
##  1st Qu.: 7.100   1st Qu.:0.3925   1st Qu.:0.0900   1st Qu.: 1.900
##  Median : 7.900   Median :0.5200   Median :0.2500   Median : 2.200
##  Mean   : 8.311   Mean   :0.5313   Mean   :0.2684   Mean   : 2.532
##  3rd Qu.: 9.100   3rd Qu.:0.6400   3rd Qu.:0.4200   3rd Qu.: 2.600
##  Max.   :15.900   Max.   :1.5800   Max.   :1.0000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 21.00       1st Qu.:0.9956
##  Median :0.07900   Median :13.00       Median : 37.00       Median :0.9967
##  Mean   :0.08693   Mean   :15.62       Mean   : 45.91       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 61.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :68.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.205   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6577   Mean   :10.44   Mean   :5.657
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
##        Id
##  Min.   :   0
##  1st Qu.: 411
##  Median : 794
##  Mean   : 805
##  3rd Qu.:1210
##  Max.   :1597
```

Figure 3: The Summary of the dataset

V. DATA PRE-PROCESSING

Id is just a label of the wine and was removed. Hence, it will not be needed as an input variable. The dataset is clean with no NA or voided spaces.

```
##        fixed.acidity     volatile.acidity          citric.acid
##            "numeric"           "numeric"            "numeric"
##      residual.sugar           chlorides  free.sulfur.dioxide
##            "numeric"           "numeric"            "numeric"
## total.sulfur.dioxide             density                   pH
##            "numeric"           "numeric"            "numeric"
##          sulphates             alcohol              quality
##            "numeric"           "numeric"            "integer"
```

Figure 4: Data types of all the input variables

I. Working with the data

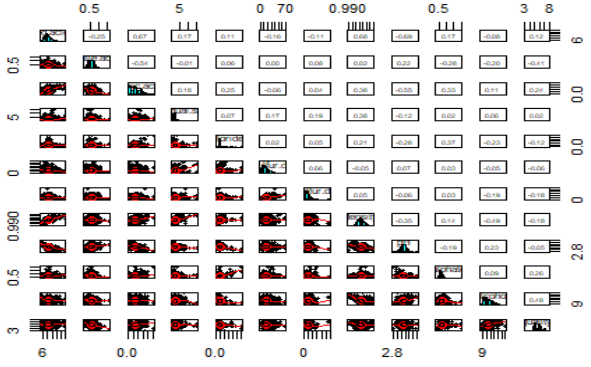

Figure 5: Plot to compare linear relationship

After cleaning the data, it was established a zero linear relationship between quality and other covariants as seen in Figure 5. This indicates that a simple linear regression might not work as an approach. Whenever a linear relationship isn't established between an output and featured inputs, linear models are not fully trusted.

VI. BUILDING A MODEL WITH EXISTING AND SUGGESTED ALGORITHMS

It should be noted that several R packages were utilized for this project aside the normal base functions. These packages contain several in-built functions and algorithms that makes machine learning easy. With just few lines of commands, the dataset was partitioned into the training and testing sets, and cross-validation were done through the proper use of these R packages. These packages include:

- neuralnet for neural networks: This is used to train neural networks using backpropagation, test it and make predictions through existing/supplied data.
- naivebayes for naïve Bayes classification: This is an easy-to-use classification technique that uses the Bayes' Theorem. It works with independence assumptions of independence while predicting.
- ggplot2: This is basically for data representation and visualization. It contains lots of features used to visualize our data in this work.
- lattice: This is also for data visualization. A very powerful tool to visualize data before categorization.
- Caret: This is an indispensable package in R for this project. It is used for training data and encapsulates several machine learning models and algorithm already prepared (It has a great training tool with over 200 models that could be used by simple syntax). This includes the RF model that is suggested in this work.
- dplyr: This is for data manipulation (in tidyverse package)
- psych: This is an important tool for multivariate analysis.

All these tools and packages were effectively utilized for this work because the models have built into some of them.

```
##   freq percentage
## 3    5  0.4854369
## 4   31  3.0097087
## 5  434 42.1359223
## 6  416 40.3883495
## 7  129 12.5242718
## 8   15  1.4563107
##   Shapiro-Wilk normality test
##
## data:  wineDrop$quality
## W = 0.8547, p-value < 2.2e-16

##          (Intercept)        fixed.acidity     volatile.acidity
##         20.293497160          0.012741945         -1.095591018
##          citric.acid       residual.sugar            chlorides
##         -0.075781469          0.005085595         -2.295359928
##  free.sulfur.dioxide total.sulfur.dioxide              density
##          0.004124116         -0.003054712        -15.595314032
##                   pH            sulphates              alcohol
##         -0.541336255          0.987771558          0.269242123

##
## Call:
## lm(formula = quality ~ ., data = trainingset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.41383 -0.36463 -0.04622  0.43338  2.00655
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.029e+01  2.627e+01   0.772 0.440011
## fixed.acidity        1.274e-02  3.228e-02   0.395 0.693151
## volatile.acidity    -1.096e+00  1.469e-01  -7.457 1.90e-13 ***
## citric.acid         -7.578e-02  1.804e-01  -0.420 0.674551
## residual.sugar       5.006e-03  2.003e-02   0.250 0.802667
## chlorides           -2.295e+00  5.355e-01  -4.286 1.99e-05 ***
## free.sulfur.dioxide  4.124e-03  2.683e-03   1.537 0.124532
## total.sulfur.dioxide -3.055e-03  8.657e-04  -3.529 0.000437 ***
## density             -1.560e+01  2.681e+01  -0.582 0.560876
## pH                  -5.413e-01  2.349e-01  -2.304 0.021404 *
## sulphates            9.878e-01  1.473e-01   6.708 3.27e-11 ***
## alcohol              2.692e-01  3.324e-02   8.099 1.57e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6344 on 1017 degrees of freedom
## Multiple R-squared:  0.3834, Adjusted R-squared:  0.3767
## F-statistic: 57.49 on 11 and 1017 DF,  p-value: < 2.2e-16

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.708   5.342   5.640   5.674   6.028   6.751
```

Figure 6: Classifying the Quality variable and creating a partition for data training and testing

From Figure 6, the output variable(quality) is only represented on the scale of 3,4,5,6,7,8. Majority of the data fell between 5 and 6. This is an unbalanced data. The data was partitioned into two. 20% was selected of the data for validation and the remaining 80% of data for training and testing the models.

It is clear that it failed the normal test. The algorithms were run using a-10-fold cross validation approach. This splits the dataset into ten different parts with nine for training and one for testing.

VII. VISUALIZING THE OUTPUT DATA AND INPUT DATA

Figure 7: Histogram of the Quality of wine

I. *Multivariate plots*

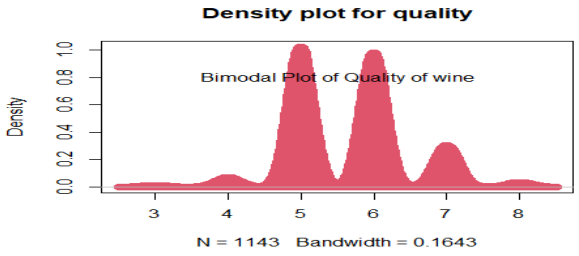

Figure 8: Density plot of quality

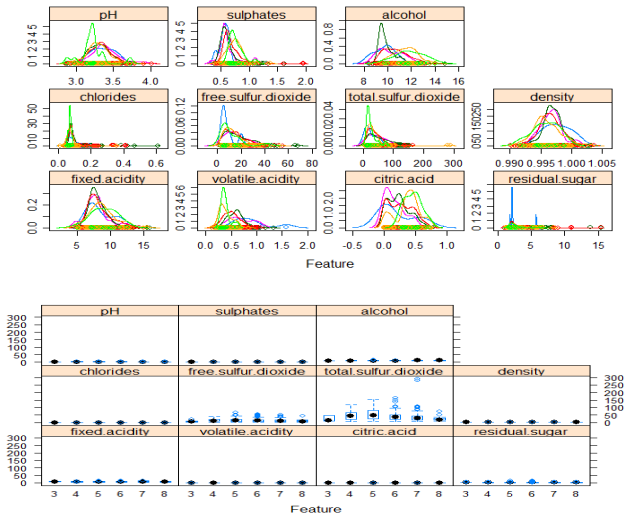

Figure 9 and 10: Feature plots of the dataset

II. *Neural Network*

Using logistic neural network on the training dataset, we have the result as seen in figure 11 below. The predictions of the wine quality as determined by the features in the network.

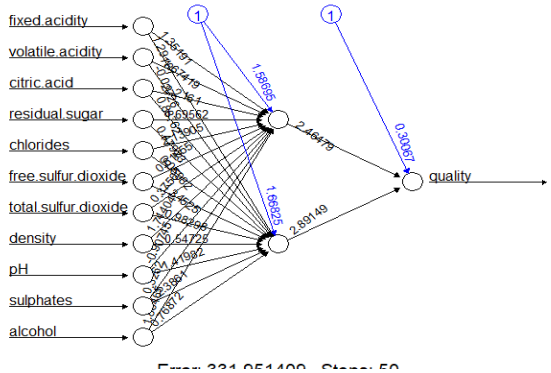

Figure 11: Neural network

III. *Predicting with Naive Bayes Classification*

```
##                3             4             5             6             7             8
## 1 1.181816e-02 0.0091401530 0.8878053 0.09119125 4.512816e-05 4.430916e-10
## 3 1.682561e-04 0.0105054666 0.6303500 0.35862771 3.481745e-04 3.813547e-07
## 4 1.477065e-14 0.0002399467 0.2162965 0.76511154 1.483728e-02 3.514716e-03
## 5 1.181816e-02 0.0091401530 0.8878053 0.09119125 4.512816e-05 4.430916e-10
## 6 4.250727e-02 0.0081991791 0.8552799 0.09394199 7.166352e-05 4.193729e-09
## 8 2.569144e-03 0.1802299063 0.3944818 0.41230263 1.041639e-02 8.196563e-08
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
## 8           7.3             0.65        0.00            1.2     0.065
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
## 8                  15                   21  0.9946 3.39      0.47    10.0
##   quality
## 1       5
## 3       5
## 4       6
## 5       5
## 6       5
## 8       7

## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.

##
## p1    3   4   5   6   7   8
##    3  5   3   2   0   0   0
##    4  0   6   5   6   0   0
##    5  0  11 334 115   3   0
##    6  0   9  78 239  35   1
##    7  0   1  16  54  89   6
##    8  0   0   0   2   2   7

## [1] 0.3391642

##
## p2   3  4  5  6  7  8
##    3  0  0  0  0  0  0
##    4  0  0  0  0  0  0
##    5  0  3 34 15  1  0
##    6  1  0 13 25  3  1
##    7  0  0  1  6  7  1
##    8  0  0  0  0  3  0

## [1] 0.4210526
```

Figure 12: Naïve Bayes classification results

From above, the errors in classification are about 34% and 42%- This isn't good. Naïve Bayes classification couldn't produce the best classification because of this range of error.

IV. *Other models*

From these results, it could be seen that alcohol, sulphates, fixed acidity and citric acid have the greatest impact on the quality of wine.

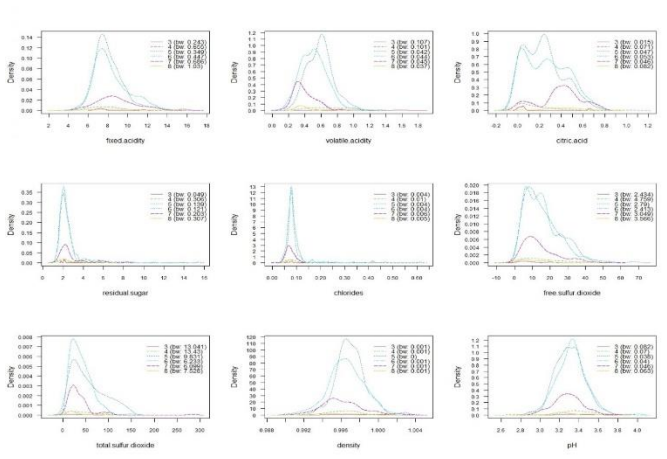

Figure 13: Impact of the input on the Various Output Levels

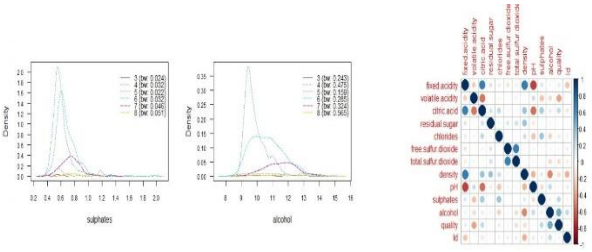

Figure 14: Impact of the inputs and their relationship with outputs

```
## Linear Discriminant Analysis
##
## 1030 samples
##   11 predictor
##    6 classes: '3', '4', '5', '6', '7', '8'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 927, 926, 927, 926, 927, 929, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.5971743  0.3517683

## CART
##
## 1030 samples
##   11 predictor
##    6 classes: '3', '4', '5', '6', '7', '8'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 927, 926, 927, 926, 927, 929, ...
## Resampling results across tuning parameters:
##
##   cp          Accuracy   Kappa
##   0.02013423  0.5475948  0.2438928
##   0.02265101  0.5514783  0.2465860
##   0.23657718  0.5318994  0.2016864
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.02265101.

## k-Nearest Neighbors
##
## 1030 samples
##   11 predictor
##    6 classes: '3', '4', '5', '6', '7', '8'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 927, 926, 927, 926, 927, 929, ...
## Resampling results across tuning parameters:
##
##   k  Accuracy   Kappa
##   5  0.4855390  0.1673174
##   7  0.5068631  0.1896854
##   9  0.5096714  0.1914213
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 9.

## Support Vector Machines with Radial Basis Function Kernel
##
## 1030 samples
##   11 predictor
```

Figure 15: Cross-validation results

```
##    6 classes: '3', '4', '5', '6', '7', '8'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 927, 926, 927, 926, 927, 929, ...
## Resampling results across tuning parameters:
##
##   C     Accuracy   Kappa
##   0.25  0.5913012  0.3050919
##   0.50  0.6028875  0.3391112
##   1.00  0.6107495  0.3579503
##
## Tuning parameter 'sigma' was held constant at a value of 0.09940508
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were sigma = 0.09940508 and C = 1.

## Random Forest
##
## 1030 samples
##   11 predictor
##    6 classes: '3', '4', '5', '6', '7', '8'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 927, 926, 927, 926, 927, 929, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.6571082  0.4421066
##    6    0.6492928  0.4319499
##   11    0.6434863  0.4229817
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.

##
## Call:
## summary.resamples(object = results)
##
## Models: lda, cart, knn, svm, rf
## Number of resamples: 10
##
## Accuracy
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lda  0.5096154 0.5686847 0.5922330 0.5971743 0.6393064 0.6732673    0
## cart 0.4851485 0.5048077 0.5679612 0.5514783 0.5834682 0.6213592    0
## knn  0.4368932 0.4879808 0.5168969 0.5096714 0.5406274 0.5533981    0
## svm  0.5384615 0.5926896 0.6019417 0.6107495 0.6502932 0.6826923    0
## rf   0.5841584 0.5980209 0.6568298 0.6571082 0.6939180 0.7572816    0
##
## Kappa
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lda  0.20048236 0.2982057 0.3533139 0.3517683 0.4233487 0.4707844    0
## cart 0.12144530 0.1860865 0.2669623 0.2465860 0.2924220 0.3809524    0
## knn  0.07751699 0.1535515 0.1996619 0.1914213 0.2508068 0.2654264    0
```

Figure 16: Cross-validation results of other models

```
## svm  0.24346076 0.3255471 0.3409335 0.3579503 0.4240990 0.4764302    0
## rf   0.31624758 0.3495987 0.4393483 0.4421066 0.5004141 0.6031746    0
```

Figure 17: Ten-fold Cross-validation of SVM and RF

V. Ten-fold Cross-validation

Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), k-Nearest Neighbors (kNN), Support Vector Machines (SVM) with a linear kernel, and Random Forest (RF) were considered as seen above, considering the accuracy level, Random Forest is said to be the best. Comparing these models, the confidence level is 95%. Random Forest is a form classification model, and it generates its output based on classified inputs. The training sets in this work was selected and classified to predict the feature that could enhance better quality of wine.

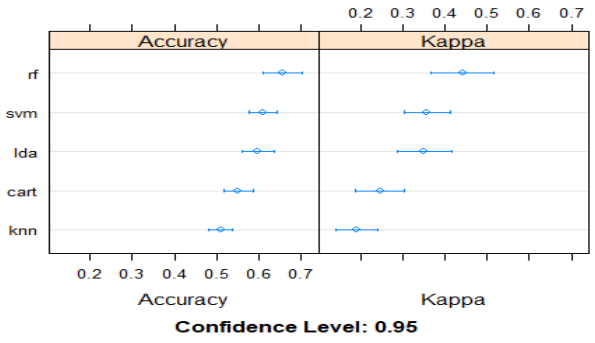

Figure 18: Comparing the models

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction   3   4   5   6   7   8
##          3   5   0   0   0   0   0
##          4   0  31   0   0   0   0
##          5   0   0 434   0   0   0
##          6   0   0   0 416   0   0
##          7   0   0   0   0 129   0
##          8   0   0   0   0   0  15
##
## Overall Statistics
##
##               Accuracy : 1
##                 95% CI : (0.9964, 1)
##    No Information Rate : 0.4214
##    P-Value [Acc > NIR] : < 2.2e-16
##
##                  Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
## Sensitivity          1.000000   1.0000   1.0000   1.0000   1.0000  1.00000
## Specificity          1.000000   1.0000   1.0000   1.0000   1.0000  1.00000
## Pos Pred Value       1.000000   1.0000   1.0000   1.0000   1.0000  1.00000
## Neg Pred Value       1.000000   1.0000   1.0000   1.0000   1.0000  1.00000
## Prevalence           0.004854   0.0301   0.4214   0.4039   0.1252  0.01456
## Detection Rate       0.004854   0.0301   0.4214   0.4039   0.1252  0.01456
## Detection Prevalence 0.004854   0.0301   0.4214   0.4039   0.1252  0.01456
## Balanced Accuracy    1.000000   1.0000   1.0000   1.0000   1.0000  1.00000
```

Figure 19: Overall statistics

VIII. CONCLUSION

The best classification model for this analysis is the Random Forest over the ten-fold classification. Based on the results, the accuracy of the wine quality prediction scores greatly hinges on alcohol level above others. Aside, this accuracy would significantly improve by not only increasing the amount of alcohol level but also the fixed acidity, citric acid, and sulfates. The amount of volatile acidity and chlorides should also be decreased because they contribute the least to the quality of the wine.

The main challenge with this work is that our data was unbalanced. The major values of quality of the wine data were between just the scores 5 and 6. This poses a difficulty in analyzing the plots due to overplotting. The machine learning models themselves could be enhanced for accuracy by using a larger dataset with a greater even distributions of wine quality scores. It's also discovered that R has easy to use model in the caret and psych package that researchers are encouraged to use in future works.

REFERENCES

[1] Dahal, K.R., Dahal, J.N., Banjade, H. and Gaire, S. (2021) Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics, 11, 278-289

[2] Ebeler S. (1999) "Flavor Chemistry — Thirty Years of Progress: chapter Linking flavour chemistry to sensory analysis of wine". Kluwer Academic Publishers, 409–422.

[3] Legin, Rudnitskaya, Luvova, Vlasov, Natale and D'Amico. (2003) "Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception". Analytica Chimica Acta 484 (1): 33–34

[4] Akanbi O, Abegunde O, Design and Implementation of Mobile Information System for Federal Road Safety Corps (FRSC) of Nigeria. Int J Sens Netw Data Commun, 9 (2020).

[5] Ceyhun Ozgur, Taylor Colliau, Grace Rogers, Zachariah Hughes, MatLab vs. Python vs. R, J. data sci. 15(2021), no. 3, 355-372, DOI 10.6339/JDS.201707_15(3).0001 https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

[6] V. Preedy, and M. L. R. Mendez, "Wine Applications with Electronic Noses," in Electronic Noses and Tongues in Food Science, Cambridge, MA, USA: Academic Press, 2016, pp. 137-151

[7] Jr, R.A., de Sousa, H.C., Malmegrim, R.R., dos Santos Jr., D.S., Carvalho, A.C.P.L.F., Fonseca, F.J., Oliveira Jr., O.N. and Mattoso, L.H.C. (2004) Wine Classification by Taste Sensors Made from Ultra-Thin Films and Using Neural Networks. Sensors and Actuators B: Chemical, 98, 77-82. https://doi.org/10.1016/j.snb.2003.09.025

[8] Larkin, T. and McManus, D. (2020) An Analytical Toast to Wine: Using Stacked Generalization to Predict Wine Preference. Statistical Analysis and Data Mining: The ASA Data Science Journal, 13, 451-464. https://doi.org/10.1002/sam.11474

[9] Kunal Thakkar et al, (IJCSIT), "AHP and Machine Learning Techniques for Wine Recommendation" International Journal of Computer Science and Information Technologies, Vol. 7 (5), 2016, 2349-2352

[10] Kavuri N. C. and Madhusree Kundu. "ART1 Network: Application in Wine Classification", International Journal of Chemical Engineering and Applications, Vol. 2, No. 3, June 2011.

[11] N. H. Beltran, M. A. Duarte- MErmound, V. A. S. Vicencio, S. A. Salah, and M. A. Bustos, "Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer," Instrum. Measurement, IEEE Trans., 57: 2421-2436, 2008.

[12] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, "Wineinformatics: applying data mining on wine sensory reviews processed by the computational wine wheel," IEEE International Conference on Data Mining Workshop, pp. 142-149, Dec. 2014.

[13] Hackenberger BK. R software: unfriendly but probably the best. Croat Med J. 2020 Feb 29;61(1):66-68. doi: 10.3325/cmj.2020.61.66. PMID: 32118381; PMCID: PMC7063554.

[14] Brittain, Jim; Cendon, Mariana; Nizzi, Jennifer; and Pleis, John (2018) "Data Scientist's Analysis Toolbox: Comparison of Python, R, and SAS Performance," SMU Data Science Review: Vol. 1: No. 2, Article 7.

[15] Ozgur, Ceyhun & Jha, Sanjeev & Shen, Yiming. (2021). R and Python for Teaching Purposes. https://daily.sevenfifty.com/making-a-case-for-premium-vinho-verde/: accessed August 24th, 2022

[16] Moreno, Gonzalez-Weller, Gutierrez, Marino, Camean, Gonzalez and Hardisson. (2007) "Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks". Talanta 72 263–268.

[17] Yu, Lin, Xu, Ying, Li and Pan. (2008) "Prediction of Enological Parameters and Discrimination of Rice Wine Age Using Least-Squares Support Vector Machines and Near Infrared Spectroscopy". Agricultural and Food Chemistry 56 307–313.

[18] Beltran,Duarte-Mermoud, Soto Vicencio, Salah and Bustos. (2008) "Chilean Wine Classification Using Volatile Organic Compounds Data Obtained With a Fast GC Analyzer". IEEE Transactions on Instrumentation and Measurement 57 2421-2436

[19] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009) Modeling Wine Preferences by Data Mining from Physicochemical Properties. Decision Support Systems, Elsevier, 47, 547-553. https://doi.org/10.1016/j.dss.2009.05.016