RED WINE QUALITY ANALYSIS

BIA 652 – Multivariate Data Analysis

Project by: Shreyas Sawant, Siddhesh Powar, and Neel Nath

Under Guidance of Prof. David Belanger

INTRODUCTION

The wine quality is a crucial factor which makes a wine unique and popular. In this analysis, we will try to comprehend the relationship of various parameters that impact the quality ratings for the wine. We determine the factors responsible for the change in the quality of the wine and finding the correlation of all the factors.

This analysis uses different classification models which can predict the quality of the wine if a new dataset is added based on our current dataset. The dataset we have used contains the data of all the components which are used to make a better-quality wine. Finding a correlation between different components and relating these components to the quality variable will determine how the quality gets affected.

This analysis classifies the data based on the rating of the wine quality which we have classified as good, average and bad. We have used different classification Models for predicting the quality if a new dataset is introduced. Also, we have found out the accuracy of each model. Using this predictive analysis, we have predicted which model is best for analyzing the quality of the wine.

DATA & DATA PREPARATION

We chose the Red Wine Quality analysis dataset from Kaggle.com. The dataset contains 11 columns showing the content value of every factor in a Red Wine and there is a column showing the quality of the Red wine ranging from 1 to 8.

The goal here is to determine the factors affecting the quality of the Red wine. Also, build and compare classification models that give the best accuracy while predicting the Red Wine Quality which is dependent on all the other factors.

Variables are explained in the below:

Variable Numbers	**Variable Type/ Description**	**Variable Names**
1-3	Red wine acid composition and concentration	fixed_acidity, volatile_acidity and citric_acid
4-10	Other Components of the Red wine.	Residual_sugar, Chlorides, free,Sulphur_di-oxide, total_sulphur_dioxide, density, pH, and sulphates
11	The alcohol content of the red wine	alcohol
12	Quality of the wine(dependent variable)	Quality
13	Rating is the additional variable that we created.	Rating (1-5 = bad wine, 5-7 = average wine, 8 = good wine)

Further, we cleaned the data and split it into two parts. One was the training dataset which contained 1120 observations which is the 70% of the data and the other was the test dataset containing 479 observations which are 30% of the data.

Below we can see the top rows of our dataset:

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide
1	7.4	0.70	0.00	1.9	0.076	11
2	7.8	0.88	0.00	2.6	0.098	25
3	7.8	0.76	0.04	2.3	0.092	15
4	11.2	0.28	0.56	1.9	0.075	17
5	7.4	0.70	0.00	1.9	0.076	11
6	7.4	0.66	0.00	1.8	0.075	13

	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
1	34	0.9978	3.51	0.56	9.4	5
2	67	0.9968	3.20	0.68	9.8	5
3	54	0.9970	3.26	0.65	9.8	5
4	60	0.9980	3.16	0.58	9.8	6
5	34	0.9978	3.51	0.56	9.4	5
6	40	0.9978	3.51	0.56	9.4	5

ANALYSIS & RESULTS

CORRELATION:

Firstly, we wanted to check what factors affect the quality of the wine and how related are these factors to each other and how it affects the quality which is a dependent factor.

Thus, we plotted the correlation matrix using the following R code:

```
#Loading the csv file and viewing the data
wine <- read.csv('wine.csv')
head(wine)
View(wine)

#correlation
res <- cor(wine)
View(res)
round(res, 2)
```

Following is the output:

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
fixed_acidity	1.00000000	-0.256130895	0.67170343	0.114776724	0.093705186	-0.153794193	-0.11318144	0.66804729	-0.68297819	0.183005664	-0.06166827	0.12405165
volatile_acidity	-0.25613089	1.000000000	-0.55249568	0.001917882	0.061297772	-0.010503827	0.07647000	0.02202623	0.23493729	-0.260986685	-0.20228803	-0.39055778
citric_acid	0.67170343	-0.552495685	1.00000000	0.143577162	0.203822914	-0.060978129	0.03553302	0.36494718	-0.54190414	0.312770044	0.10990325	0.22637251
residual_sugar	0.11477672	0.001917882	0.14357716	1.000000000	0.055609535	0.187048995	0.20302788	0.35528337	-0.08565242	0.005527121	0.04207544	0.01373164
chlorides	0.09370519	0.061297772	0.20382291	0.055609535	1.000000000	0.005562147	0.04740047	0.20063233	-0.26502613	0.371260481	-0.22114054	-0.12890656
free_sulfur_dioxide	-0.15379419	-0.010503827	-0.06097813	0.187048995	0.005562147	1.000000000	0.66766645	-0.02194583	0.07037750	0.051657572	-0.06940835	-0.05065606
total_sulfur_dioxide	-0.11318144	0.076470005	0.03553302	0.203027882	0.047400468	0.667666450	1.00000000	0.07126948	-0.06649456	0.042946836	-0.20565394	-0.18510029
density	0.66804729	0.022026232	0.36494718	0.355283371	0.200632327	-0.021945831	0.07126948	1.00000000	-0.34169933	0.148506412	-0.49617977	-0.17491923
pH	-0.68297819	0.234937294	-0.54190414	-0.085652422	-0.265026131	0.070377499	-0.06649456	-0.34169933	1.00000000	-0.196647602	0.20563251	-0.05773139
sulphates	0.18300566	-0.260986685	0.31277004	0.005527121	0.371260481	0.051657572	0.04294684	0.14850641	-0.19664760	1.000000000	0.09359475	0.25139708
alcohol	-0.06166827	-0.202288027	0.10990325	0.042075437	-0.221140545	-0.069408354	-0.20565394	-0.49617977	0.20563251	0.093594750	1.00000000	0.47616632
quality	0.12405165	-0.390557780	0.22637251	0.013731637	-0.128906560	-0.050656057	-0.18510029	-0.17491923	-0.05773139	0.251397079	0.47616632	1.00000000

By looking at the above correlation matrix we could observe that there were few of the factors such as fixed acidity and density and few other factors which were strongly correlated according to the correlation coefficient while rest were not that strongly correlated.

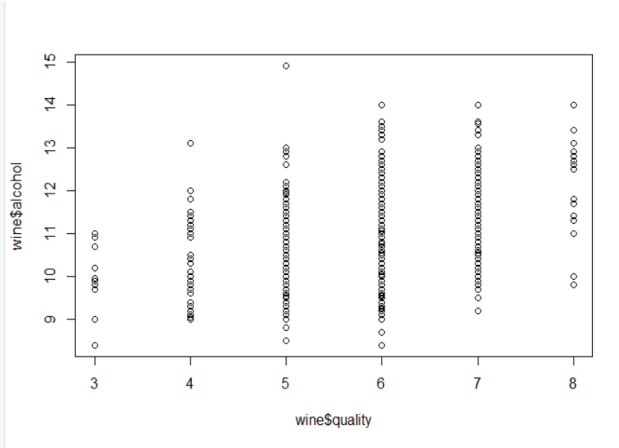
Further, we wanted to check about how every factor affects the quality of the wine. To check on the same we tried plotting graphs with quality on the x-axis and every other factor on the y-axis.

Following is the code for plotting the graphs:

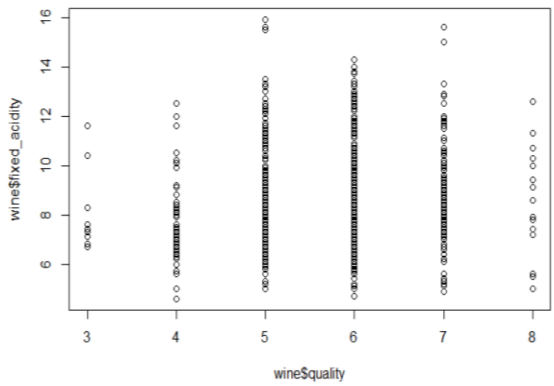
```
#Plotting various factors against quality
plot(x= wine$quality, y = wine$alcohol)
plot(x= wine$quality, y = wine$fixed_acidity)
plot(x= wine$quality, y =wine$citric_acid)
plot(x= wine$quality, y =wine$residual_sugar)
plot(x= wine$quality, y =wine$chlorides)
plot(x= wine$quality, y =wine$free_sulfur_dioxide)
plot(x= wine$quality, y =wine$total_sulfur_dioxide)
plot(x= wine$quality, y =wine$density)
plot(x= wine$quality, y =wine$pH)
plot(x= wine$quality, y =wine$sulphates)
plot(x= wine$quality, y =wine$volatile_acidity)
```

Below are the plots (outputs) of the above code:

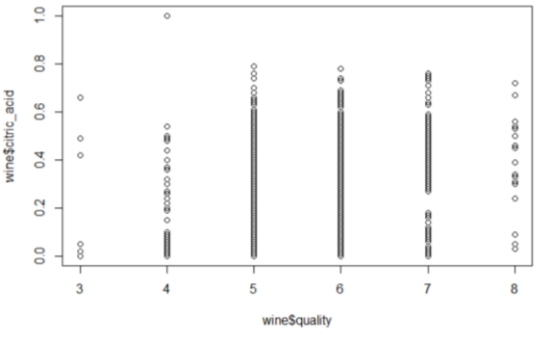
Alcohol VS Quality



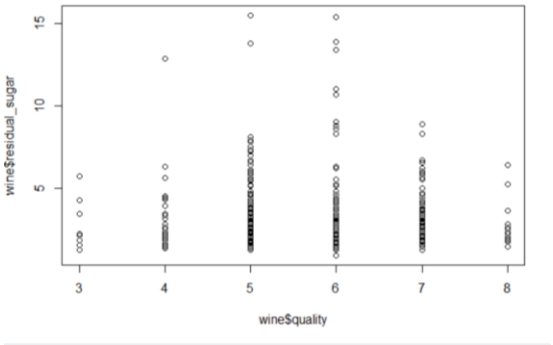
Fixed acidity VS Quality



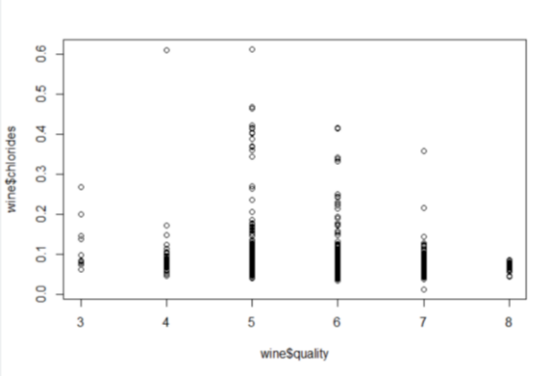
Citric acid VS Quality



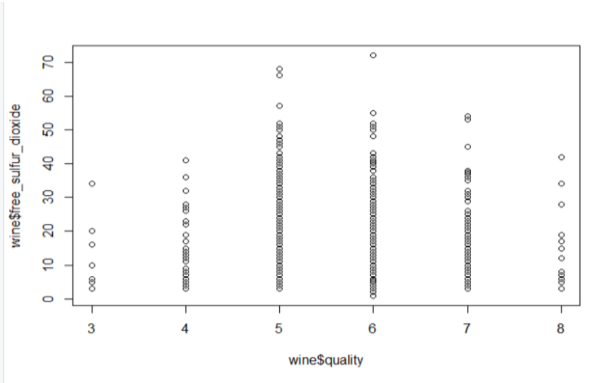
Residual Sugar VS Quality



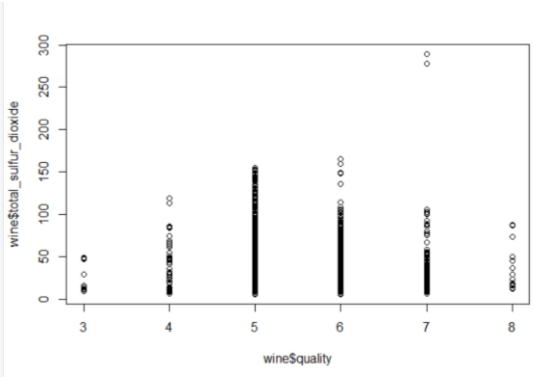
Chlorides VS Quality



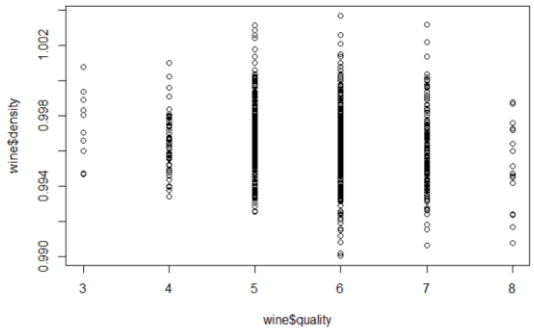
Free Sulphur dioxide VS Quality



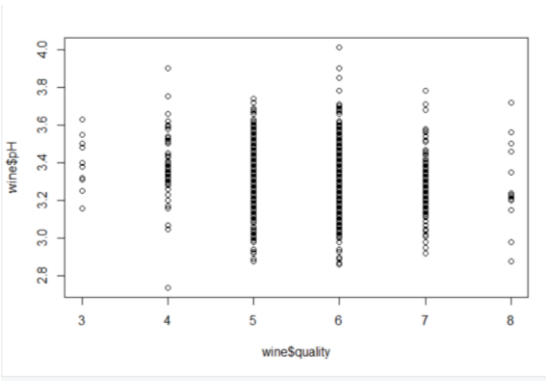
Total Sulphur dioxide VS Quality



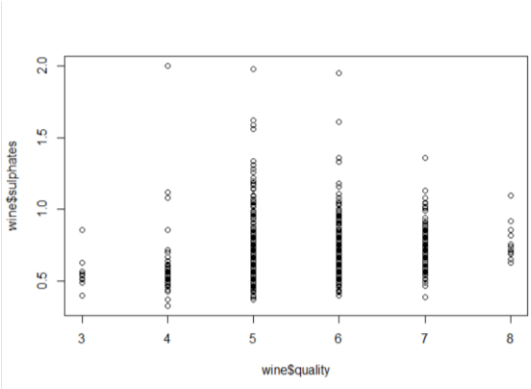
Density VS Quality



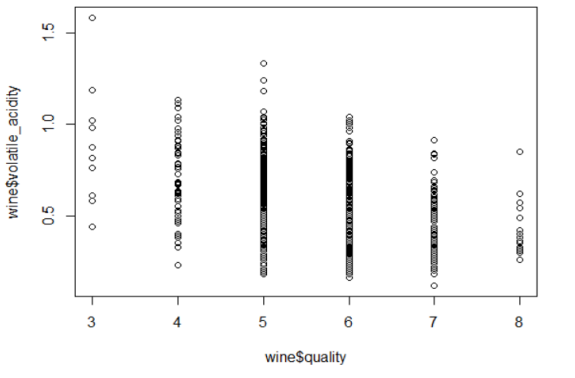
Ph VS Quality



Sulphates VS Quality



Volatile Acidity VS Quality



Thus, after observing the above graphs we could conclude the following:

- Comparison of Fixed Acidity, Residual Sugar and Free Sulphur Oxide with Quality did not give us any clear view.
- Volatile Acidity and Chloride levels depreciate as Quality increases.
- Alcohol and Citric Acid show an Upwards trend (Directly proportional) when compared to the quality.

We further wanted to try and compare different classification models which could be used to predict the quality if a new dataset is introduced to the model which is trained using our dataset and compare their accuracy.

To try different classification models, we had to create a new column “Rating” which was further used to classify our data according to the wine quality.

In the rating column we could see that the wine quality ranging from 1-5 was given a poor rating, 5-7 was average and 8 was a good quality wine.

To perform the above we used the following R-Code:

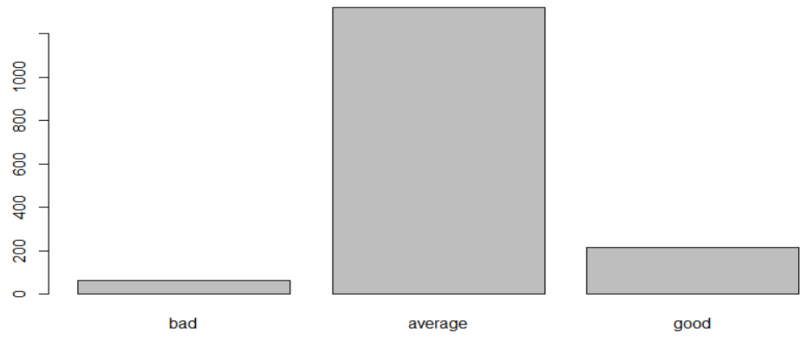
```
#Transforming Quality from an Integer to a Factor
wine$quality <- factor(wine$quality, ordered = T)
#Creating a new Factored Variable called 'Rating'
wine$rating <- ifelse(wine$quality < 5, 'bad', ifelse(wine$quality < 7, 'average', 'good'))
View (wine)
wine$rating <- ordered(wine$rating,levels = c('bad', 'average', 'good'))
wine$rating
#PLOTTING TO CLASSIFY DIFFERENT CATEGORIES OF THE QUALITY OF THE WINE
barplot(table(wine$rating))
```

The new dataset with the additional column “Rating” is given below:

fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality	rating
7.4	0.700	0.00	1.90	0.076	11	34	0.9978	3.51	0.56	9.4	5	average
7.8	0.880	0.00	2.60	0.098	25	67	0.9968	3.20	0.68	9.8	5	average
7.8	0.760	0.04	2.30	0.092	15	54	0.9970	3.26	0.65	9.8	5	average
11.2	0.280	0.56	1.90	0.075	17	60	0.9980	3.16	0.58	9.8	6	average
7.4	0.700	0.00	1.90	0.076	11	34	0.9978	3.51	0.56	9.4	5	average
7.4	0.660	0.00	1.80	0.075	13	40	0.9978	3.51	0.56	9.4	5	average
7.9	0.600	0.06	1.60	0.069	15	59	0.9964	3.30	0.46	9.4	5	average
7.3	0.650	0.00	1.20	0.065	15	21	0.9946	3.39	0.47	10.0	7	good
7.8	0.580	0.02	2.00	0.073	9	18	0.9968	3.36	0.57	9.5	7	good
7.5	0.500	0.36	6.10	0.071	17	102	0.9978	3.35	0.80	10.5	5	average
6.7	0.580	0.08	1.80	0.097	15	65	0.9959	3.28	0.54	9.2	5	average
7.5	0.500	0.36	6.10	0.071	17	102	0.9978	3.35	0.80	10.5	5	average
5.6	0.615	0.00	1.60	0.089	16	59	0.9943	3.58	0.52	9.9	5	average
7.8	0.610	0.29	1.60	0.114	9	29	0.9974	3.26	1.56	9.1	5	average
8.9	0.620	0.18	3.80	0.176	52	145	0.9986	3.16	0.88	9.2	5	average
8.9	0.620	0.19	3.90	0.170	51	148	0.9986	3.17	0.93	9.2	5	average
8.5	0.280	0.56	1.80	0.092	35	103	0.9969	3.30	0.75	10.5	7	good

After this, we wanted to classify and check how is our dataset divided and what type of wines have a dominance in our dataset.

Thus, we plotted a bar graph which is given below:



Here we could see that most of the wines in our dataset were Average wines.

Now based on this, we further created a train and test dataset which was then used in our models to predict their accuracy.

We used the following R-code to create the Train (70%) and Test (30%) datasets.

```
#Creating Training and testing datasets
training = wine[0:1120,]
test = wine[1121:1599,]
#TRAINING AND TEST LABEL FOR KNN
train_label = wine[0:1120, 13]
test_label = wine[1121:1599, 13]
```

CLASSIFICATION MODELS

Now we developed and compared following Classification Models for Prediction:

Linear Discriminant Analysis:

- Linear Discriminant Analysis is used to classify individual objects into two or more groups based on measurements.

Following is the R- Code for this model:

```
#LDA
library("MASS")
#TRAINING THE MODEL
lda_dataset <- lda(formula = rating ~ ., data=training)
lda_dataset
#TESTING THE MODEL
lda_PredictVal <- predict(lda_dataset, test)
lda_PredictVal

lda_classValues <- lda_PredictVal$class
lda_classValues

lda_table <- table(lda_classValues, test$rating)
lda_table
#ACCURACY
accuracy_LDA <- sum(diag(lda_table)/sum(lda_table))*100
accuracy_LDA
```

Using the above code, we got the following output:

CONFUSION MATRIX:

lda_classValues	bad	average	good
bad	19	0	0
average	5	405	4
good	2	0	44

ACCURACY:

```
> accuracy_LDA
[1] 97.70355
>
```

Considering our comparatively small dataset we got a higher accuracy as we predicted using the test dataset. The accuracy of predicting the quality of wine if new data is added 97.70355%.

To get a better accuracy on our small set of data we tried few more models.

Naïve Bayes:

- Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

Following is the R-code for this model:

```
# Naive bayes
library(MASS)
#TRAINING THE MODEL
nBayes_all <- naiveBayes(rating ~., data =training)
#TESTNG THE MODEL
category_all<-predict(nBayes_all,test)

#Creating a confusion matrix
table(category_all,test$rating)
str(category_all)
#CONVERTING IN THE ORDERED FORMAT
test$rating <- as.ordered(test$rating)
category_all <- as.ordered(category_all)
# Calculating the error rate
NB_wrong<-sum(category_all!=test$rating)
NB_error_rate<-NB_wrong/length(category_all)
print("Error rate is")
NB_error_rate
accuracy <- 1- NB_error_rate
print("Accuracy is")
accuracy*100
```

Using this we got the following output:

Confusion Matrix

```
category_all bad average good
     bad      23       5    0
     average   1     398    0
     good      2       2   48
```

Accuracy:

```
[1] "Error rate is"
> NB_error_rate
[1] 0.02087683
> accuracy <- 1- NB_error_rate
> print("Accuracy is")
[1] "Accuracy is"
> accuracy*100
[1] 97.91232
```

According to our analysis, the error rate for Naïve Bayes is 0.02087% and accuracy rate is 97.91232%.

Random Forest:

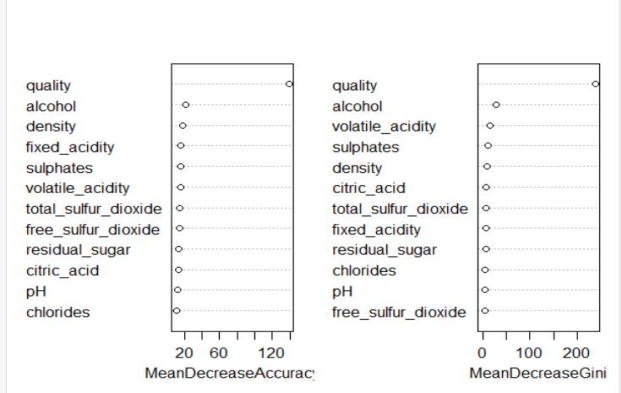
- Random forests or random decision forests are an ensemble learning method for classification, regression, and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

We used the following code for trying this model:

```
#Random Forest
#install.packages("randomForest")
library(randomForest)
#TRAINING AND TESTING THE MODEL
prediction <- randomForest(rating~., data = training,ntree = 1000,importance = TRUE)
summary(prediction)
#training$rating <- as.factor(training$rating)
###Calculate the feature importance for feature selection
importance(prediction)
varImpPlot(prediction)
prediction$predicted
category_all <- predict(prediction, test)
category_all
#CONVERTING IN THE ORDERED FORMAT
test$rating <- as.ordered(test$rating)
category_all <- as.ordered(category_all)
RF_wrong<-sum(category_all!=test$rating)
RF_error_rate<-RF_wrong/length(category_all)
print("Error rate is")
RF_error_rate*100
accur<- 1- RF_error_rate
print("Accuracy is")
accur*100
```

The output of the code is given below:

Feature Importance:



ACCURACY:

```
> accur*100
[1] 99.79123
>
```

The accuracy that we got using Random forest was 99.79% which is the highest when compared to all the previous models.

We could have tried Logistic regression, but we already built our models on 3 different rating classifications and got an accuracy of 99.79%. Thus, again reconsidering and classifying the dataset into two categories “good” and “bad” was not a feasible solution as we already achieved a higher accuracy using Random Forest.

PERFORMANCE MEASUREMENTS & CONCLUSION

- We performed several different analyses to predict the quality of wine if a new dataset is added using different classification models.
- After classifying our Dataset based on the quality we observed that majority of the wines fall into the average category.
- Based on our observations we could observe that Alcohol and citric acid are directly proportional to the wine quality, whereas factors like Volatile acidity and Chlorides are inversely proportional.

- To which we can conclude that comparatively increasing the alcohol and citric acid content in the wine can give us a better quality of the wine. Also, this works inversely for the factors like Volatile acidity and Chlorides.
- After comparing different classifiers to predict the accuracy if a new data is added, we could conclude that Random Forest works best for predicting the wine quality based on our dataset.

REFERENCES

- https://www.kaggle.com/
- http://www.rdatamining.com/docs/regression-and-classification-with-r