STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Red Wine Quality Analysis

Multivariate Data Analysis
Prof. David Belanger

Siddhesh Powar
Shreyas Sawant
Neel Nath

Contents

1) Overview
2) Project Summary
3) Data Understanding
4) Data Quality Check & Cleaning
5) Correlation Matrix
6) Classification
7) Plot and Plot analysis of factors
8) Linear Discriminant Analysis
9) Naïve Bayes
10) Random Forest
11) Conclusion
12) References

Overview

- Analyze the quality of the wine.
- The overall scope of this analysis is to comprehend the relationship of various parameters that impact the quality ratings for the wine.
- In this analysis we are trying to understand and perform the following:
- Determining the factors responsible for the change in the quality of the wine.
- Correlation of all the factors.
- Creating, training and testing different classification models which can predict the quality of the wine if a new dataset is added based on our current dataset.

Project Summary

➢ **Data Source:**

- The dataset contains information of the factors which determine the quality of the Red Wine.
- Dataset available on Kaggle.com

➢ **Project Objective:**

- The objective of this project is to determine how the quality of the red wine gets affected by the different parameters.
- Finding correlation between different parameters which affect the wine quality.
- Classification of our dataset to categorize the quality of wine between good, average and bad.
- Building, training, testing and comparing different classification models to better predict the quality of the wine.

Data understanding

- The dataset has 1599 observations and 12 variables.
- Quality is the Dependent variable.
- All other 11 variables are independent variables based on which wine quality is tested.
- Following is a list of variables in the dataset:

Variables 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 Red Wine.	residual_sugar, chlorides, free_sulfur_dioxide, total_sulfur_dioxide, density, pH and sulphates.
11	The alcohol content of the red wine	alcohol
12	Quality of the wine	quality

Data Quality Check and Cleaning

- **Data Cleaning**
  - We removed the null values in the dataset.
  - Tried locating the inaccurate data or an outlier.
  - Removed the redundancies and considered only the unique rows.

- **Data Splitting**

  - We randomly split the dataset into training and testing datasets.
  - Training dataset contains 1120 observations(70%)
  - Testing dataset contains 479 observations(30%)

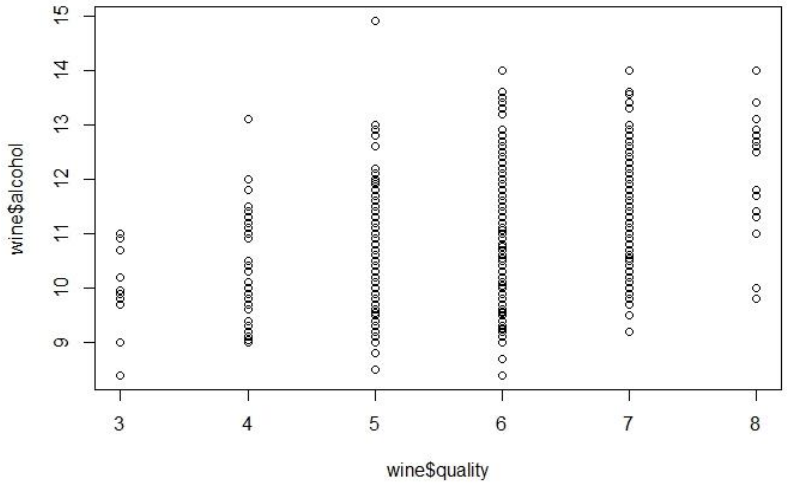
Correlation Matrix

	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

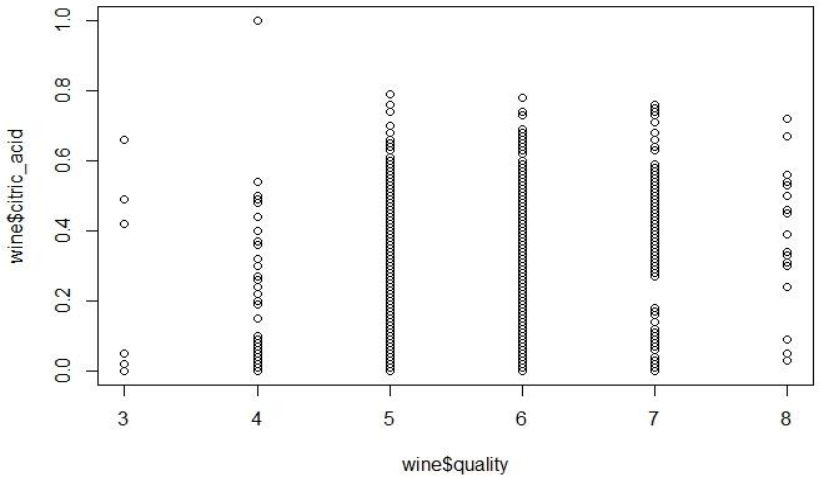
- This is the correlation matrix for the factors of the dataset.
- Certain factors such as density and fixed acidity were strongly correlated while few others were not.

Directly Proportional

Alcohol

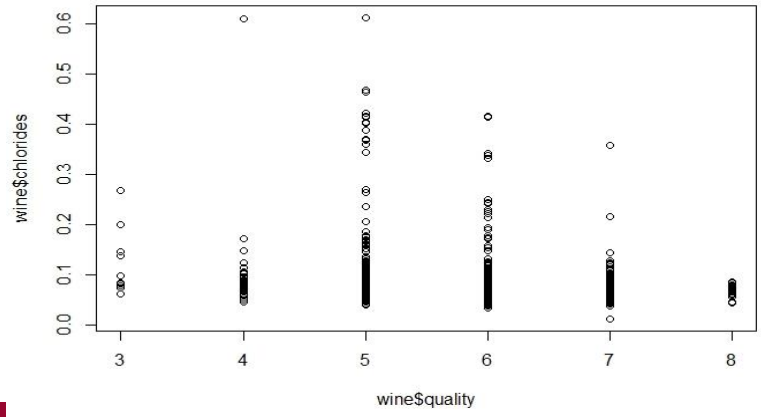


Citric Acid

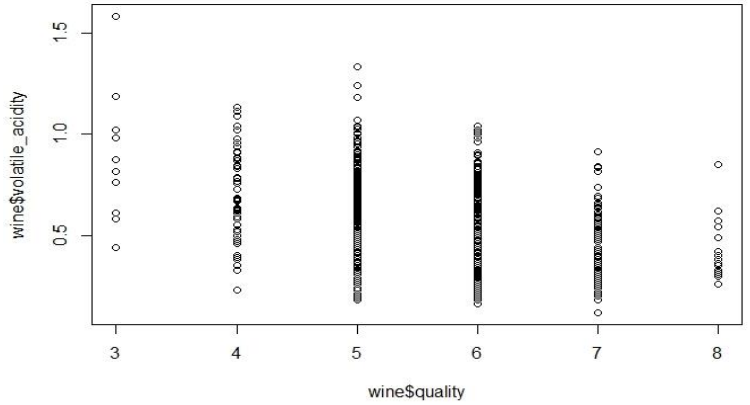


Inversely Proportional

Chlorides



Volatile Acidity

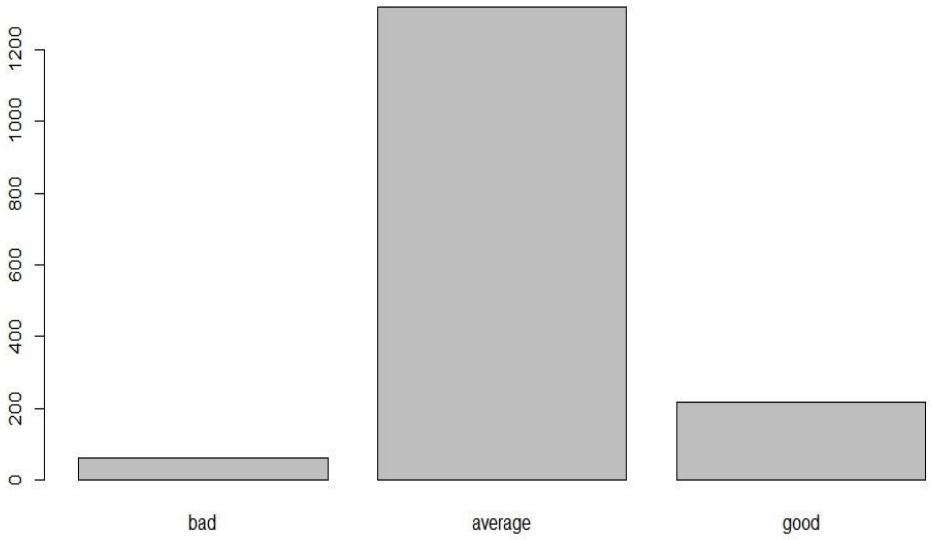


Plot Analysis (OBSERVATIONS)

- 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.

Classification of the wine dataset based on the quality

- Quality ranges from 1 to 8
- Thus, we classified the data into three categories based on the quality which are good, bad and average.
- The wines with quality lesser than 5 are classified as bad, lesser than 7 but greater than 5 are classified as average and quality 8 is classified as a good wine.
- We have created a bar plot to observe the number of good, bad and average quality of wines.
- We can observe that average quality of wines is significantly more than the good and the bad ones.



Linear Discriminant Analysis

- Linear Discriminant Analysis is used to classify individual objects into two or more groups on the basis of measurements.
- We built a model on our Training dataset to classify the quality of wine as bad, average and good.
- 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%.

```
Coefficients of linear discriminants:
                              LD1           LD2
fixed_acidity         -0.143519549  -0.245075790
volatile_acidity      -0.166581224  -2.427814428
citric_acid           -0.625620820  -0.949498064
residual_sugar        -0.104637659  -0.073851997
chlorides              0.381468801  -1.267066073
free_sulfur_dioxide    0.004161471  -0.010871988
total_sulfur_dioxide   0.000504987   0.008917571
density              132.035440231 239.379394495
pH                    -0.920553112  -1.930447044
sulphates             -0.511502461  -0.788526435
alcohol               -0.103215002   0.037711194
quality.L             -9.040389409   6.455084465
quality.Q              5.994491749   3.087005867
quality.C             -4.015783824   1.867160312
quality^4              5.191382138   2.673425502
quality^5             -0.701102324   0.652550737

Proportion of trace:
   LD1    LD2
0.8866 0.1134
```

```
> lda_table <- table(lda_classvalues, test$rating)
> lda_table

lda_classvalues bad average good
        bad      19       0    0
        average   5     405    4
        good      2       0   44
>
> accur <- sum(diag(lda_table)/sum(lda_table))*100
> accur
[1] 97.70355
>
> plot(lda_dataset)
```

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.
- According to our analysis, the error rate for Naïve Bayes is 0.02087% and accuracy rate is 97.91232%.
- An advantage of Naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.
- We can also see the confusion matrix.

```
[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
```

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

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.
- Accuracy = 99.79 %
- So far Random Forest is the best model for prediction of wine quality depending on the factors.

```
> summary(prediction)
                Length Class   Mode
call               5   -none-  call
type               1   -none-  character
predicted       1120   factor  numeric
err.rate        4000   -none-  numeric
confusion         12   -none-  numeric
votes           3360   matrix  numeric
oob.times       1120   -none-  numeric
classes            3   -none-  character
importance        60   -none-  numeric
importanceSD      48   -none-  numeric
localImportance    0   -none-  NULL
proximity          0   -none-  NULL
ntree              1   -none-  numeric
mtry               1   -none-  numeric
forest            14   -none-  list
y               1120   ordered numeric
test               0   -none-  NULL
inbag              0   -none-  NULL
terms              3   terms   call
> RF_error_rate*100
[1] 0.2087683
> accur<- 1- RF_error_rate
> print("Accuracy is")
[1] "Accuracy is"
> accur*100
[1] 99.79123
```

Conclusion

- We performed several different analysis 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 in 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

Thank You