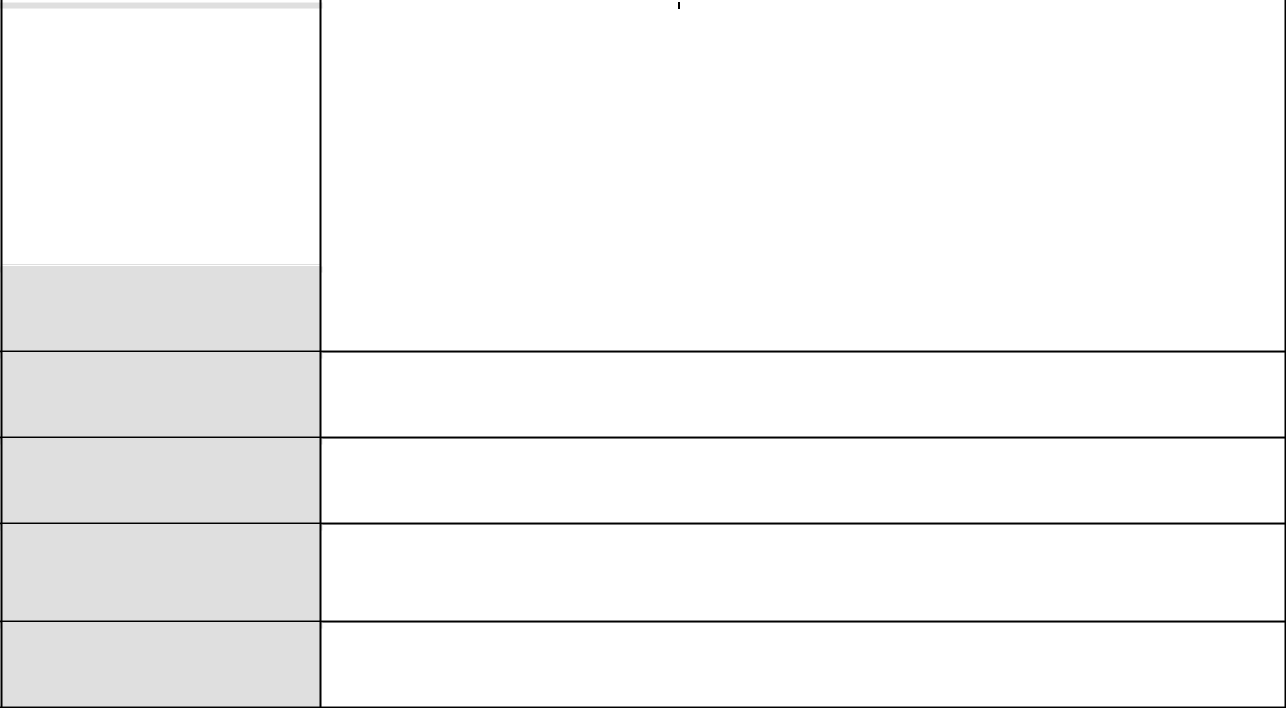


**Computing, Engineering & Mathematics**

**ASSIGNMENT / REPORT COVER SHEET**

**This sheet must be attached to all material being submitted for marking.**

****

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Student name:** |  | **Student name:** |  |
|  | Syed Maaz Ali |  |
|  | **Student number:** | 19637352 | **Student number:** |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  | **Sections completed** | All | **Sections completed** |  |
|  | **individually** |  | **individually** |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**Unit name & number: Data Science**

**Tutorial day and time:**

**Title of Assignment: Significance of chemicals present in Red Wine**

**Student Submitting the**

**Assignment:**

17-Nov-2019

**Date submitted:**

**Student Declaration (must be signed)**

**Declaration:**

1. I hold a copy of this assignment if the original is lost or damaged.
2. I hereby certify that no part of this assignment or product has been copied from any other student’s work or from any other source except where due acknowledgement is made in the assignment.
3. No part of the assignment/product has been written / produced for me by any other person except where collaboration has been authorised by the subject lecturer/tutor concerned
4. I am aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism ***(which may retain a copy on*** ***its database for future plagiarism checking***)***)***
5. I hereby certify that no part of this assignment or product has been submitted by me in another (previous or current) assessment, except where appropriately referenced, and with prior permission from the Lecturer/Tutor/ Unit Co-ordinator for this unit.

**Student signature and date:**

**Student signature and date**

**Note: An examiner or lecturer/tutor has the right to not mark this assignment if the above declaration has not**

**been signed.**

*Significance of chemicals present in Red Wine*

*Research Paper*

Syed Maaz Ali

Western Sydney University

Sydney, Australia

[19637352@student.westernsydney.edu.au](mailto:19637352@student.westernsydney.edu.au)

1. INTRODUCTION

Dramatic increase in wine consumption has been observed in a global scale. This could be attributed to positive effects of wine consumption to cardiovascular health. Consequently, this has led to augmented sales of wine in the last 5 years [1]. In response to the aforementioned growth in wine consumption and its sales, the industry has been innovating new methodologies in wine production to increase its efficiency and daily output. Additionally, innovation and development of strategic marketing tactics are being performed to optimise its sales volume [2].

There is an extensive variety of red wine that is available in the international market. Despite the differences in wine variety, chemical tests indicate that the majority of red wine has similarity in its chemical components. However, its proportions have varied concentrations in different types of red wine. The difference in chemical quantity is a significant aspect in order to classify each type. Due to lack of sophisticated technology, there had been a difficulty in performing wine classification in the past. However, data mining techniques such R-language and Python have enable to not only classify wines, but to determine the significance of each chemical component and its correlation to taste, quality and other features. This would allow wineries to perform exclusion or reduction of chemical components that are unnecessary or could potentially affect the wine quality.

In this report, a public data set of red wine from UCI Repository would be utilised to perform data analysis using R-studio. The research question for this study is to identify the significant chemicals present in red wine which play a critical role in predicting the quality of wine. To answer this question through analysis, different methods such as Linear Regression, Decision Trees and Principal Component Analysis would be used.

1. DATA FOR ANALYSIS

Materials and Methods:

The data set used for this analysis is available at UCI repository website for research purposes under the URL <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>. This data is based on Portuguese “Vinho Verde” red wine. Vinho verde is a wine that is produced in the northwestern region of Portugal called Minho. This wine accounts for 15% of the wine production in Portugal. In this data analysis only red wine has been selected and the white wine variant of vinho verde has been excluded (Corteza, Cerdeira, Almeida, Matos, & Reis, 2009).

A large dataset of red wine comprised of 1599 samples have been utilised, in which 11 independent variables and 1 dependent variable, which is quality have been considered in this data analysis. The independent variables include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH values, sulphates and alcohol (Corteza, Cerdeira, Almeida, Matos, & Reis, 2009). Among the aforementioned 11 independent variables, alcohol and volatile acidity have a profound effect on the aroma and flavour of wine (Reference). Alcohol pertains to the product of fermentation created by the yeast enzymes that convert the grape sugars into ethanol and carbon dioxide. Volatile acidity is a by-product of the oxidation of alcohol to acetic acid. It enhances the wine aroma in small amount; whereas, it causes vinegary taste and odour in excessive quantity (Grainger & Tattersall, 2016). The acidity of wine could be determined in terms of pH value. The pH value of 0 is termed as acidic, while pH value of 14 indicates that the wine is alkaline. On average, wine has a pH value of 3 to 4 (P. Cortez).

The aim of this analysis is to predict the significance of different chemicals present in wine and highlight the elements which affect the quality of wine. Due to confidentiality issues, the brand name of wines, its selling price, type of grapes used and wineries locations are not provided in this dataset. The table below provides the mean and standard deviations of all 11 independent physiochemical properties of red wine.

* 1. DATA MINING APPROACH & TECHNIQUES

Statistical machine learning is a set of powerful algorithms used to predict an outcome variable against multiple predictors. The algorithms improve their performance automatically through learning from the data. The data mining has been divided into two parts: supervised learning and unsupervised learning. Supervised learning captures and processes the required input and conveys output precision and accuracy for the majority of the decision-making problems (Haney, 2015). In supervised learning, one quantitative analysis and one qualitative analysis have been performed. Descriptive statistics has been utilised for quantitative analysis to obtain the mean, median and minimum vallues for each independent variable. In addition, a box plot has been used to indicate the presentation of different parameters of the data (Bruce & Andrew, 2017).

To assess the performance of alcohol in wine using the aforementioned dataset, several tests have been conducted such as multiple linear regression and cross validation approach. Multiple linear regression has been performed to test the correlation between various independent variables and for the dependent variable, which is quality. Multiple linear regression pertains to an extension of simple linear regression in which more than one predictor variables are used to predict an output variable (Bruce & Andrew, 2017). A cross-validation approach has been utilised using alcohol as the only predictor to measure the performance of the resulted model on new test data set. This technique involves dividing the dataset into two subsets known as training dataset and testing dataset. Training dataset is used to train the model while testing dataset is used to test the model by determining the prediction error. There are various cross validation methods to evaluate the performance of a model, such as: validation dataset approach, leave one out cross validation, k-fold cross validation and repeated k-fold cross validation. For this model, the basic validation dataset approach has been used (Bruce & Andrew, 2017).

For this qualitative analysis, decision trees model has been specifically used. It builds a set of decision rules to describe the connection between independent variables and the outcome variable (P. Bruce and Bruce 2017).

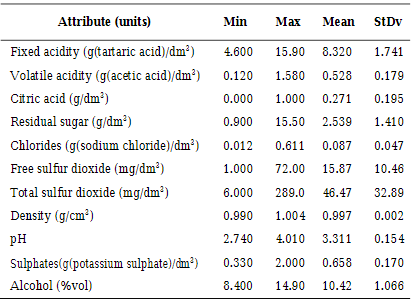
For unsupervised learning, Principal Component Analysis has been performed. This method provides a summary and visualization of the information contained in a dataset having multiple quantitative variables. Each component in the dataset is considered a different dimension; however, if there are more than 3 predictor variables in the dataset, it becomes difficult to visualize a multi-dimensional hyperspace. In this situation, PCA extracts the important information from a multivariate data and expresses it as a new set of fewer variables which are known as principal components (PCs). Newly produced variables correspond to a linear combination of the original variables. The number of PCs is less than or equal to the quantity of original variables. In essence, Principal Component Analysis reduces the dimensions of a data to two or three PCs. Therefore, the data could be visualized graphically without any major losses of data information.

IV. DATA MINING & EXPERIMENTAL RESULTS

From the original data set of wine quality, the red wine data have been used to evaluate the performance of different chemicals and alcohol. The quality of Portuguese red wine called 'Vinho Verde' has been utilised as a response variable. Different data mining methods have been used in this analytical study including multiple linear regression, decision trees, cross validation and principal component analysis.

* Data exploration – Descriptive

The dataset of red wine has been explored to derive its physiochemical baseline statistics. The descriptive analysis has enabled to obtain the mean and standard deviation values of each chemical and its respective min-max values. The table below indicates the descriptive results:

Table1 – Mean and SD of all independent variables

* Linear regression

A multiple linear regression model has been created to predict the quality based on the composition of 11 different independent variables, such as: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates and alcohol.

To interpret the results, the summary of the model has been obtained and the F-statistic and its associated p-values has been examined. Based on the summary, it has been determined that the p-value is < 2.2e-16, which means it is quite significant. This demonstrates that there is at least one independent variable that is highly correlated to the response variable (quality).

To determine the significance of predictor variables, the coefficient table has been examined that shows the regression beta coefficients and t-statistics p-values. The t-statistic assesses the correlation between a given predictor and the outcome variable. From the summary of the model1, it is evident that three elements have high t values. Alcohol and sulphates indicate a strong positive correlation with quality of wine. Whereas, volatile acidity demonstrates a strong negative correlation with quality of wine. To describe the relation of volatile acidity with quality: the lesser the volatile acidity of wine, the better the wine quality (reference).

The model had been redefined using volatile acidity, sulphate and alcohol as predictors that is highly significant with the response variable. Then, the model had been re-tested and named as “Model2”. The summary of model2 illustrates that changes in alcohol level, the composition of sulphates or volatile acidity is significantly associated to changes in quality of wine. The graph for this model is plotted and showed in appendix of this report.

Finally, the model equation has been determined:

Quality (Redwine) = 2.610 + 0.309\*alcohol + 0.679\*sulphates - 1.221\*volatile acidity

Model accuracy: To test the quality of the model, the Adjusted R-squared value and Residual Standard Error of the model have been examined. Based on the Model2 results, the adjusted R-squared = 0.335, meaning that 33.5% of the variance in the measurement of quality could be predicted by changes in sulphates and alcohol levels, and volatile acidity of wine.

RSE is the measure of error of prediction. Lower RSE suggests higher accuracy of the model. For model2, the RSE is 0.659 corresponding to an error rate of 0.117. Error rate is calculated by dividing RSE by the mean of response variable.

* Decision trees & Cross validation

As mentioned earlier, decision trees are one of the automated algorithms that functions via continuously improving its performance through learning from data. It divides the data into multiple subspaces, so that the outcomes in each subspace is homogenous. For this dataset, classification trees has been utilised. Since the response variable in the given dataset is continuous, the quality variable would be changed to a categorical variable. To do this, a new variable known as High quality is be introduced and defined as: if quality is greater than 6, it is High quality.

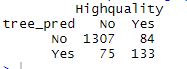
Subsequently, the quality variable from the dataset is removed. A new data frame with “High quality” variable is created. The decision tree model is plotted that indicates the different possible splitting rules that could be used to predict an outcome. The summary illustrates that the misclassification error rate for the decision tree model is 0.099 which means it is not a bad model.

Figure – Misclassification error table

However, having all the predictors in the tree makes it complex. Therefore, it is difficult to interpret with a large data set. A solution to this problem is pruning the tree. Prior the pruning, a cross validation using training and testing data set has to be performed followed by a tree model for training data. The misclassification error rate for training data tree is calculated to be 0.0994.

After cross validation, pruning is performed on the misclassified instances and plot the graph of deviance against size of the nodes. From the graph (Figure6 of appendix), it could be determined that the best size for the decision tree model is 4 as it has the lowest deviance. After pruning, the model is re-tested for accuracy and to facilitate calculation for its misclassification error rate. The new MSE is 0.0512 which shows that misclassification rate of the pruned tree is lower than that of the unpruned tree. Hence, it indicates that pruning has improved the performance of the model.

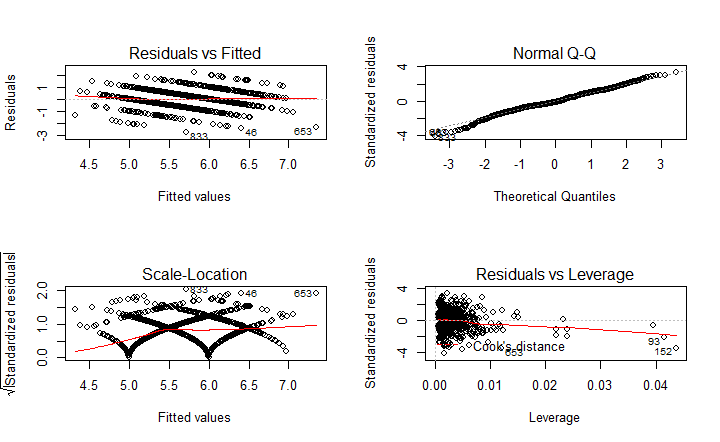
* Principal Component Analysis

PCA is used to extract the important information from the data set containing a large number of predictor variables and summarizes it in a form of new data set containing fewer variables that enables efficient data visualisation and interpretation. To perform PCA, the means and variances of different variables in the dataset is first examined. From the results, it is determined that variables have different means and variances. Subsequently, the PC loading vectors are displayed using $rotation function. When rotation function is performed, it can be seen that there are 11 distinct principal components. Biplot is plotted to determine the dependence of different variables on PC1 and PC2. Since the dataset involves a high number of variables, the graph plotted is a complex one. Therefore, PCA is performed and analysed the biplot graph again after scaling. The new biplot graph indicates that pH, citric acid, fixed acidity and sulphates have high contribution towards PC1 while the rest of the variables have a high contribution towards PC2 as shown in figure 5 of appendix.

The summary of this post PCA data demonstrates the overall perspective of the variance, standard deviation and proportion of variance of all the principal components. The summary illustrates that PC1 encompasses approximately 28% of the data variation; PC2 explains an estimated 17.5% of the variation in the dataset. The first two PCs together explain approximately 46% of variation in the dataset as depicted by the figure 7 in appendix.

1. CONCLUSION

In conclusion, the most significant chemicals in Vigno Verde red wine, such as: alcohol, sulphates and volatile acidity have been determined through quantitative and qualitative analysis with the aid of algorithms such as linear regression, decision trees, cross validation, and principal component analysis. Additionally, it has been validated that alcohol level, composition of sulphates, and volatile acidity play a critical role in the quality of wine. The lesser the volatile acidity, the better the quality of wine. Whereas, other variables do not indicate a high correlation with the quality of wine. Therefore, to produce a high quality of wine, winemakers should give a paramount importance to the wine volatile acidity, level of alcohol present and sulphates. For the best quality, the level of volatile acidity should be maintained as low as possible, while equally focusing on the alcohol and sulphates level present in the wine.



APPENDIX

Figure 1 - Linear regression graph

* Residual vs fitted plot:

Plot 1 does not look like a sky at night and shows multiple patterns, therefore model is not linear.

* Normal Q-Q:

Plot 2 line is not straight, therefore normality assumption of residuals is not valid.

* Scale-location:

Plot 3 does not show any consistency in variance of y from the line, therefore law of homoscedasticity is violated.

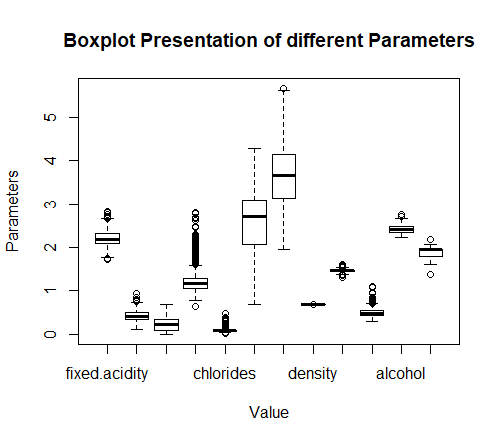


Figure 2 – Boxplot of different variables

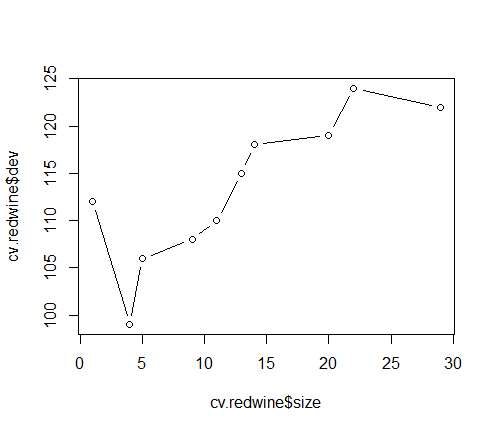
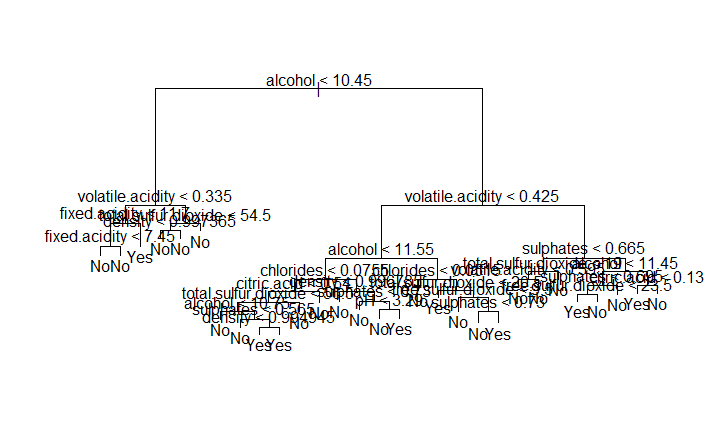


Figure 6 – Graph of deviance vs size of nodes

Figure 3 – plot of decision tree before pruning

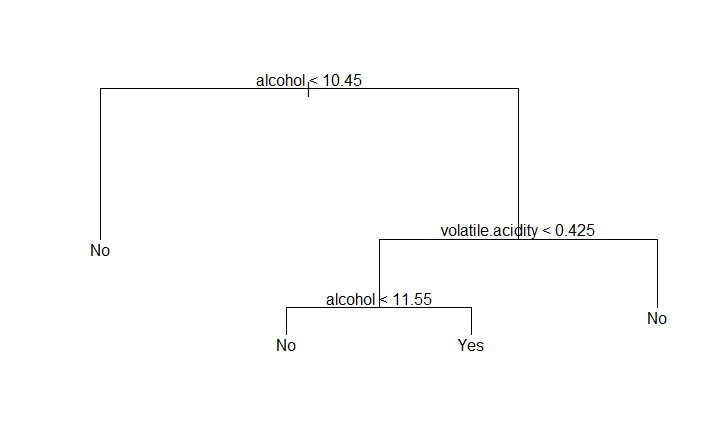


Figure 4 – Plot of decision tree after pruning

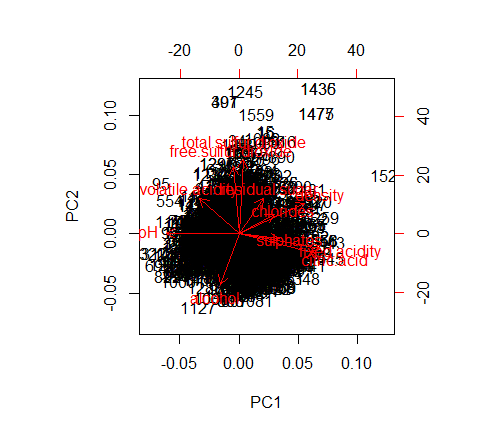


Figure 5 – Biplot of post PCA scaling

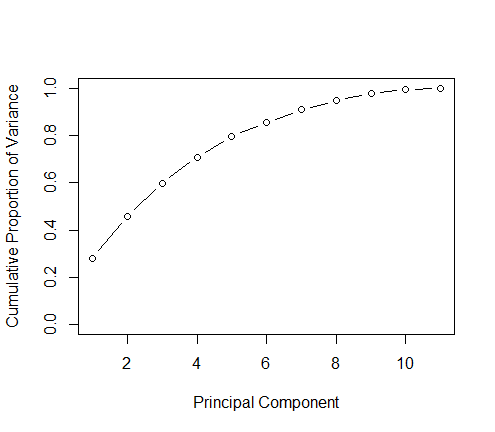


Figure 7 – Graph showing cumulative proportion of variance against PCs

R-SCRIPT

ead.csv("redwine.csv")

redwine = read.csv("redwine.csv", sep = ';')

attach(redwine)

names(redwine)

head(redwine)

dim(redwine)

summary(redwine)

stat.desc(redwine)

summary(redwine$quality)

table(redwine$quality)

redwine1=redwine[,1:12]

redwine2=log(redwine1+1)

boxplot(redwine2,xlab="Value",ylab="Parameters",main="Boxplot Presentation of different Parameters")

model1 = lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcohol, data=redwine)

plot(model1)

summary(model1)

summary(model1)$coefficient

#now considering the elements who have high t-statistic values

model2 = lm(quality~alcohol+volatile.acidity+sulphates, data = redwine)

summary(model2)

summary(model2)$coefficient

par(mfrow=c(2,2))

plot(model2)

anova(model2)

confint(model2)

qf(0.95,1,198)

#Equation ---> Quality = 2.610 +0.309alcohol +0.679sulphates -1.221volatile acidity

sigma(model2)/mean(redwine$quality)

pairs(redwine,panel=panel.smooth)

#covariance

cov(redwine, method = "p")

#correlation

cor(redwine, method = "p")

cor(alcohol, quality)

cor(sulphates, quality)

#continuous variable is alcohol

model2$residuals

plot(predict(model2),model2$residuals)

hist(model2$residuals)

predict(model2)

predict(model2,interval='confidence')

predict(model2,interval='confidence')

boxplot(alcohol~quality, ylab= "alcohol")

#validation set approach

dim(redwine)

set.seed(3)

tr = sample(1:1599,800)

train=redwine[tr , ]

test=redwine[-tr,]

dim(train)

#build the model

modelTr = lm(quality~alcohol,

data=redwine, subset = tr)

coef(modelTr)

mean((quality -predict (modelTr ,redwine))[-tr ]^2)

library(boot)

modelTr1 = glm(quality~alcohol,

data=redwine, subset = tr)

coef(modelTr1)

#decision trees

#changing quality variable from continuous to categorical

Highquality=ifelse(quality<=6,"No","Yes")

redwine=data.frame(redwine,Highquality)

View(redwine)

head(redwine)

str(Highquality)

#Remove Quality variable and create new data frame with Highquality Variable

redwinenew=redwine[,-12]

names(redwinenew)

library(ISLR)

library(tree)

tree\_model=tree(Highquality~.,redwinenew)

plot(tree\_model)

text(tree\_model,pretty=0)

summary(tree\_model)

#model check

tree\_pred=predict(tree\_model,redwinenew,type="class")

table(tree\_pred,Highquality)

#misclassification error rate = 0.09944

#Cross Validation using Training and Testing data Sets

dim(redwinenew)

set.seed(3)

train=sample(1:nrow(redwinenew), 800)

redwine.train=redwinenew[train,]

redwine.test=redwinenew[-train,]

dim(redwine.train)

dim(redwine.test)

head(redwine.train)

head(redwine.test)

#Set the Highquality variable (Target Variable) in training and testing data sets

Highquality.train=Highquality[train]

Highquality.test=Highquality[-train]

list(Highquality.train)

#Build the tree model for training data

tree\_model1=tree(Highquality~.,redwine.train)

plot(tree\_model1)

text(tree\_model1,pretty=0)

summary(tree\_model1)

#testing the model accuracy

tree\_pred1=predict(tree\_model1,redwine.test,type="class")

table(tree\_pred1,Highquality.test)

#missclassification error rate is 0.0512

#cross validation

set.seed(3)

cv.redwine=cv.tree(tree\_model1,FUN=prune.misclass)

names(cv.redwine)

cv.redwine

par(mfrow=c(1,1))

plot(cv.redwine$size, cv.redwine$dev, type = "b")

#pruning the tree

prune.redwine=prune.misclass(tree\_model1,best=4)

plot(prune.redwine)

text(prune.redwine,pretty=0)

table(tree\_pred1,Highquality.test)

yhat<-predict(prune.redwine, newdata = redwine[-train,])

redwine\_test<-redwine[-train,"quality"]

MSE<-mean((yhat-redwine\_test)^2)

MSE

sqrt(MSE)

#PCA

apply(redwine[,1:11], 2, mean)

apply(redwine[,1:11], 2, var)

obj = prcomp(redwine[,1:11]) # perform PCA

par(mfrow=c(1,1))

plot(obj$x[,1:11], col=unclass(redwine$quality)+1,

pch=16, asp=1)

obj

names(obj)

obj$sdev

obj$rotation

obj$center

obj$scale

dim(obj$x)

obj$x

screeplot(obj)

biplot(obj)

obj1 = prcomp(redwine[,1:11], scale. = TRUE)

obj1$rotation

plot(obj1$x[,1:11], col=unclass(redwine$quality)+1,

pch=16, asp=1)

biplot(obj1)

summary(obj1)

obj1$sdev

obj1.var=obj1$sdev^2

obj1.var

pve = obj1.var/sum(obj1.var)

pve

plot(cumsum(pve),xlab="Principal Component",ylab="Cumulative Proportion of Variance",

ylim = c(0,1), type = 'b')

REFERENCES

1. H. Sohaib, A. Bryce, and A. Baranchuk, “Wine and Cardiovascular Health: A Comprehensive Review,” Circulation, 136 (15). DOI:10.1161/CIRCULATIONAHA.117.030387
2. P. Cortez, A. Cerderia, F. Almeida, T. Matos, and J. Reis, “Modelling wine preferences by data mining from physicochemical properties,” In Decision Support Systems, Elsevier, 47 (4): 547-553. ISSN: 0167-9236. J..
3. Corteza, P., Cerdeira, A., Almeida, F., Matos, T., & Reisa, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47, (4). https://doi.org/10.1016/j.dss.2009.05.016
4. Grainger, K. & Tattersall, H. Wine Production and Quality. John Wiley & Sons, Incorporated, 2016. Retrieved from http://ebookcentral.proquest.com/lib/acu/detail.action?docID=4306430
5. Lecture Notes in Computer Science, pp. 77-84, 2010. Available: 10.1007/978-3-642-13498-2\_11.
6. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in Springer.