CMTH 642 Data Analytics: Advanced Methods

Assignment 3 (10%)

Tasdeed Aziz

Section:DBH $ ID:500945638

Due:June 17, 2020 11.30PM

```
#install.packages('class')
#install.packages("caret")
#install.packages('e1071')
#install.packages ('gmodels')
library(e1071)
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(class)
library(gmodels)
```

```
wine <- read.csv('http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-whi
```

1. Import to R the following fiel: http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv (The dataset is related to white Portuguese "Vinho Verde" wine. For more info: https://archive.ics.uci.edu/ml/datasets/Wine+Quality) (3 points)

```
str(wine)
```

2. Check the datatypes of the attributes. (3 points)

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
```

```
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```
check <- ifelse(sum(is.na(wine)) == 0, print('No missing values in the dataset'), print('There are missi
```

3. Are there any missing values in the dataset? (4 points)

```
## [1] "No missing values in the dataset"
```

```
#Correlation between variables excluding Quality
cor(wine[-12])
```

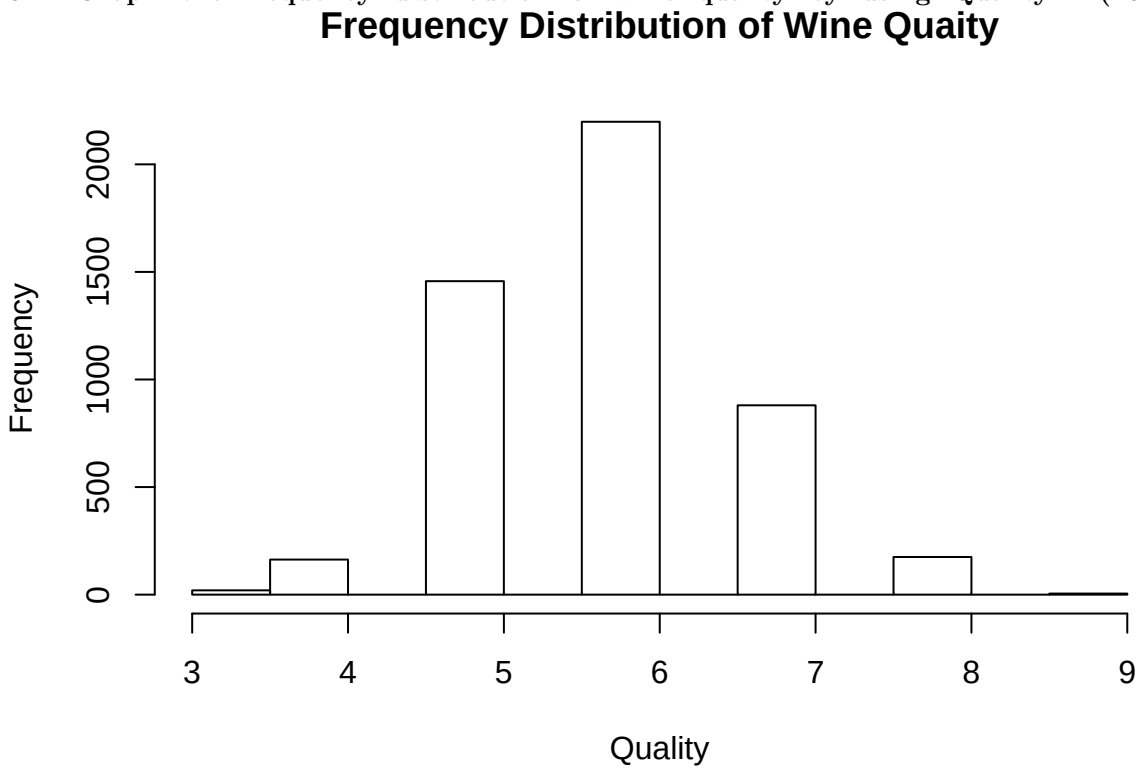
4. What is the correlation between the attributes other than Quality? (10 points)

```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity            1.00000000      -0.02269729  0.28918070     0.08902070
## volatile.acidity        -0.02269729       1.00000000 -0.14947181     0.06428606
## citric.acid              0.28918070      -0.14947181  1.00000000     0.09421162
## residual.sugar           0.08902070       0.06428606  0.09421162     1.00000000
## chlorides                0.02308564       0.07051157  0.11436445     0.08868454
## free.sulfur.dioxide     -0.04939586      -0.09701194  0.09407722     0.29909835
## total.sulfur.dioxide     0.09106976       0.08926050  0.12113080     0.40143931
## density                  0.26533101       0.02711385  0.14950257     0.83896645
## pH                      -0.42585829      -0.03191537 -0.16374821    -0.19413345
## sulphates               -0.01714299      -0.03572815  0.06233094    -0.02666437
## alcohol                 -0.12088112       0.06771794 -0.07572873    -0.45063122
##                        chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.02308564       -0.0493958591          0.091069756
## volatile.acidity      0.07051157       -0.0970119393          0.089260504
## citric.acid           0.11436445        0.0940772210          0.121130798
## residual.sugar        0.08868454        0.2990983537          0.401439311
## chlorides             1.00000000        0.1013923521          0.198910300
## free.sulfur.dioxide   0.10139235        1.0000000000          0.615500965
## total.sulfur.dioxide  0.19891030        0.6155009650          1.000000000
## density               0.25721132        0.2942104109          0.529881324
## pH                   -0.09043946       -0.0006177961          0.002320972
## sulphates             0.01676288        0.0592172458          0.134562367
## alcohol              -0.36018871       -0.2501039415         -0.448892102
##                          density            pH   sulphates     alcohol
## fixed.acidity         0.26533101 -0.4258582910 -0.01714299 -0.12088112
## volatile.acidity      0.02711385 -0.0319153683 -0.03572815  0.06771794
## citric.acid           0.14950257 -0.1637482114  0.06233094 -0.07572873
```

```
## residual.sugar         0.83896645 -0.1941334540 -0.02666437 -0.45063122
## chlorides              0.25721132 -0.0904394560  0.01676288 -0.36018871
## free.sulfur.dioxide    0.29421041 -0.0006177961  0.05921725 -0.25010394
## total.sulfur.dioxide   0.52988132  0.0023209718  0.13456237 -0.44889210
## density                1.00000000 -0.0935914935  0.07449315 -0.78013762
## pH                    -0.09359149  1.0000000000  0.15595150  0.12143210
## sulphates              0.07449315  0.1559514973  1.00000000 -0.01743277
## alcohol               -0.78013762  0.1214320987 -0.01743277  1.00000000
```

```
hist(wine$quality, xlab='Quality', main = 'Frequency Distribution of Wine Quaity')
```

5. Graph the frequency distribution of wine quality by using Quality. (10 points)




```
wine$quality <- ifelse( wine$quality <5, 'Low', ifelse(wine$quality < 7, 'Medium', 'High'))

table(wine$quality)
```

6. Reduce the levels of rating for quality to three levels as high, medium and low. Assign the levels of 3 and 4 to level 0; 5 and 6 to level 1; and 7,8 and 9 to level 2. (10 points)

```
##
##   High    Low Medium
##   1060    183   3655
```

```r
normalize <- function(x){
  return ((x - min(x)) / (max(x) - min(x)))
}
```

```r
wine_n <- as.data.frame(sapply(wine[-12],normalize))
wine_n <- cbind(wine$quality,wine_n)
## Randomly selected a variable to check wether the variable has been normalize
summary(wine_n$alcohol)
```

7. Normalize the data set by using the following function: (12 points)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2419  0.3871  0.4055  0.5484  1.0000
```

```r
set.seed(123)
index <- sample(1:nrow(wine_n), 0.65*nrow(wine_n))
wine_train <- wine_n[index,]
wine_test <- wine_n[-index,]
```

8. Divide the dataset to training and test sets. (12 points)

```r
wine_train_labels <- wine_train[,1]
wine_test_labels <- wine_test[,1]
table(wine_train_labels)
```

9. Use the KNN algorithm to predict the quality of wine using its attributes. (12 points)

```
## wine_train_labels
##   High    Low Medium
##    695    128   2360
```

```r
table(wine_test_labels)
```

```
## wine_test_labels
##   High    Low Medium
##    365     55   1295
```

```r
prediction <- knn(train= wine_train[,2:11], test = wine_test[,2:11], cl = wine_train_labels, k = 10)
```

```r
confusionMatrix(prediction, wine_test_labels)
```

10. Display the confusion matrix to evaluate the model performance. (12 points)

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction High  Low Medium
##     High    133    3    127
##     Low       0    3      2
##     Medium  232   49   1166
##
## Overall Statistics
##
##                Accuracy : 0.7592
##                  95% CI : (0.7382, 0.7793)
##     No Information Rate : 0.7551
##     P-Value [Acc > NIR] : 0.3591
##
##                   Kappa : 0.2706
##
##  Mcnemar's Test P-Value : <2e-16
##
## Statistics by Class:
##
##                      Class: High Class: Low Class: Medium
## Sensitivity              0.36438   0.054545        0.9004
## Specificity              0.90370   0.998795        0.3310
## Pos Pred Value           0.50570   0.600000        0.8058
## Neg Pred Value           0.84022   0.969591        0.5187
## Prevalence               0.21283   0.032070        0.7551
## Detection Rate           0.07755   0.001749        0.6799
## Detection Prevalence     0.15335   0.002915        0.8437
## Balanced Accuracy        0.63404   0.526670        0.6157
```

```r
table(Actual = wine_test_labels, Predict = prediction)
```

11. Evaluate the model performance by computing Accuracy, Sensitivity and Specificity. (12 points)

```
##         Predict
## Actual   High  Low Medium
##   High    133    0    232
##   Low       3    3     49
##   Medium  127    2   1166
```

```
#Accuracy: (TP+TN)/(TN+TP+FN+FP) = (133 + 3 + 1166) /(133+0+232+3+3+49+127+2+1166) = 1302 / 1715 = 0.75
#Sensitivity : (TN) /(FN + TN)
             #High : 0.3644, 36.44%
             #Low : 0.054545, 5.46%
             #Medium: 0.9004, 90%
#Specificity : (TP)/(TP+FP)
             #High: 0.90370, 90.3%
             #Low:  0.998795, 99.9%
             #Medium: 0.3310, 33.10%
```

This is the end of Assignment 3

Ceni Babaoglu, PhD