CMTH 642 Data Analytics: Advanced Methods

Assignment 3 (10%)

[Ilhak Park]

[DJ0 & 501072432]

```
wine <- read.csv(file="http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality
                 header = TRUE, sep =";")
```

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)

```
sapply (wine, class)
```

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

```
##        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"
```

```
str(wine)
```

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

```
sum(is.na(wine))
```

```
## [1] 0
```

```
#there is no missing values in the dataset
```

```
corWine <- cor(wine[-12])
corWine
```

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

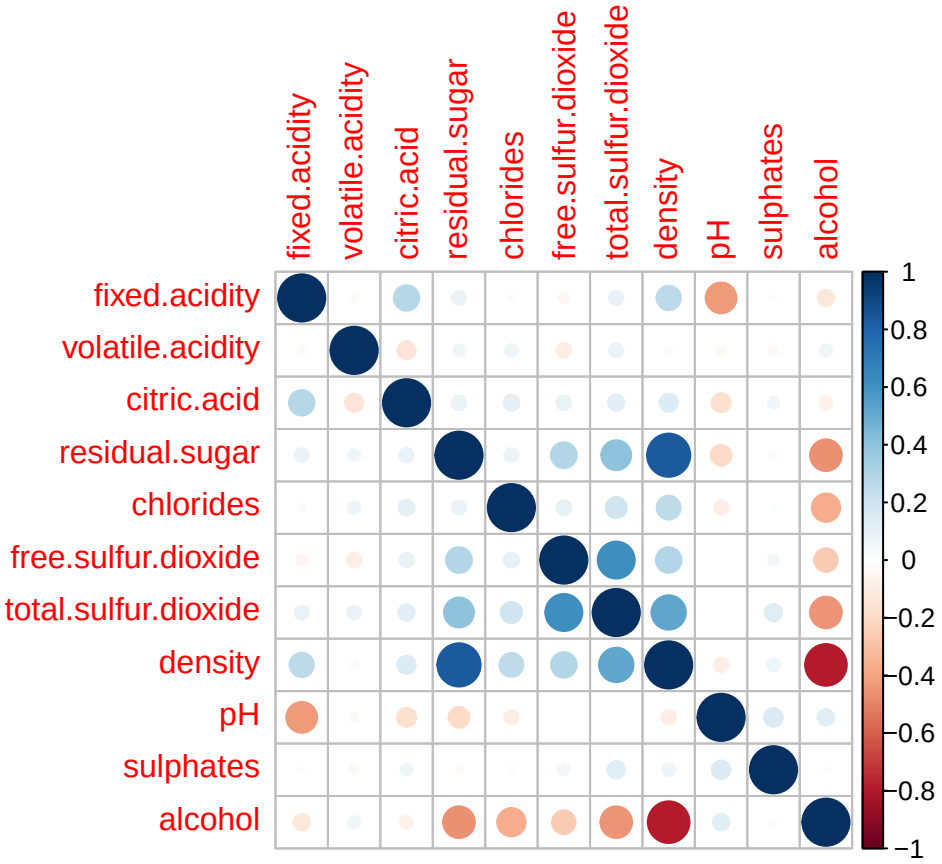
```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity            1.00000000      -0.02269729  0.28918070     0.08902070
## volatile.acidity        -0.02269729       1.00000000 -0.14947181     0.06428606
## citric.acid              0.28918070      -0.14947181  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
```

```
library(corrplot)
```

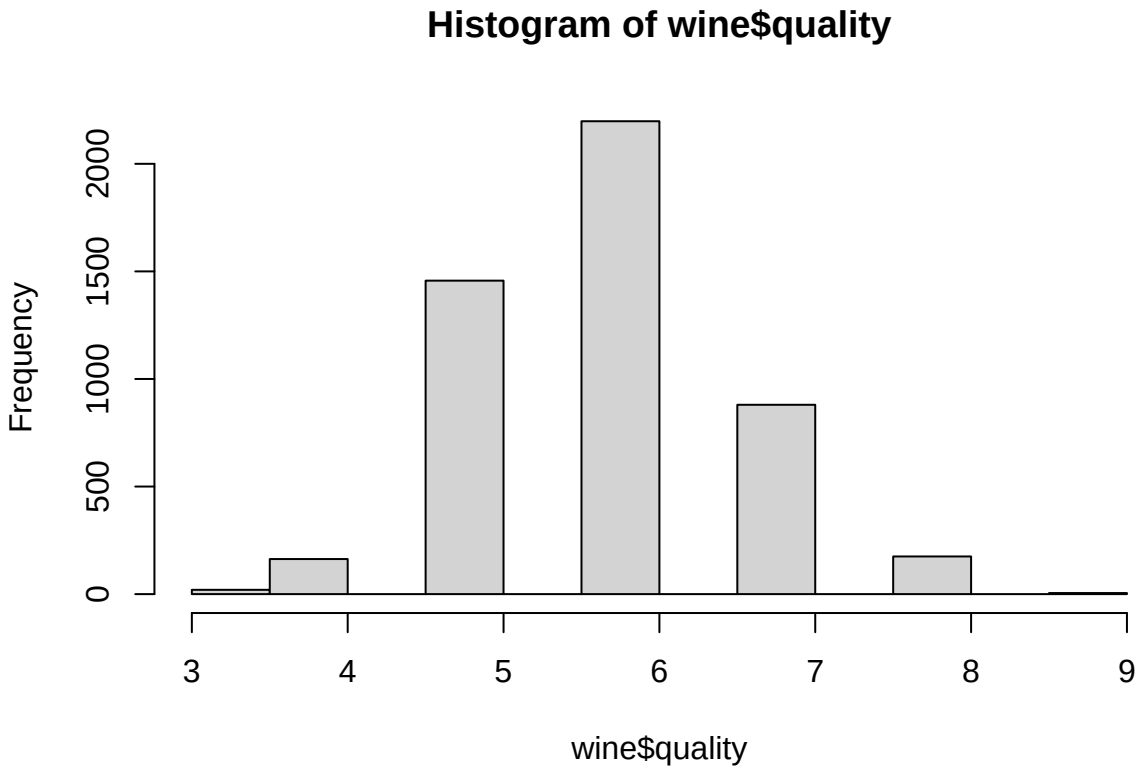
```
## corrplot 0.84 loaded
```

```
corrplot(corWine, method="circle")
```




```
hist(wine$quality)
```

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




```
for (i in  1:4898){ if (wine$quality[i]==3 | wine$quality[i]==4) {wine$quality[i] = 'Level 0: Low'}
  else if (wine$quality[i]==5 | wine$quality[i]==6) {wine$quality[i] = 'Level 1: Medium'}
  else if (wine$quality[i]==7 | wine$quality[i]==8 | wine$quality[i]==9)
    {wine$quality[i] = 'Level 2: 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)

```
##
##    Level 0: Low Level 1: Medium   Level 2: High
##             183            3655            1060
```

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

```
wine_n <- as.data.frame(lapply(wine[-12], normalize))
wine_n <- cbind(wine_n, wine$quality)
summary(wine_n)
```

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

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.2404   1st Qu.:0.1275   1st Qu.:0.1627   1st Qu.:0.01687
##  Median :0.2885   Median :0.1765   Median :0.1928   Median :0.07055
##  Mean   :0.2937   Mean   :0.1944   Mean   :0.2013   Mean   :0.08883
##  3rd Qu.:0.3365   3rd Qu.:0.2353   3rd Qu.:0.2349   3rd Qu.:0.14264
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00000   Min.   :0.00000     Min.   :0.0000       Min.   :0.00000
##  1st Qu.:0.08012   1st Qu.:0.07317     1st Qu.:0.2297       1st Qu.:0.08892
##  Median :0.10089   Median :0.11150     Median :0.2900       Median :0.12782
##  Mean   :0.10912   Mean   :0.11606     Mean   :0.3001       Mean   :0.13336
##  3rd Qu.:0.12166   3rd Qu.:0.15331     3rd Qu.:0.3666       3rd Qu.:0.17332
##  Max.   :1.00000   Max.   :1.00000     Max.   :1.0000       Max.   :1.00000
##        pH           sulphates         alcohol       wine$quality
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Length:4898
##  1st Qu.:0.3364   1st Qu.:0.2209   1st Qu.:0.2419   Class :character
##  Median :0.4182   Median :0.2907   Median :0.3871   Mode  :character
##  Mean   :0.4257   Mean   :0.3138   Mean   :0.4055
##  3rd Qu.:0.5091   3rd Qu.:0.3837   3rd Qu.:0.5484
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

```
set.seed(1)
index <- sample (1:nrow(wine_n), 0.65 *nrow(wine_n))
wine_train <- wine_n[index,]
wine_test <- wine_n[-index,]
wine_train_labels <- wine_train[,12]
wine_test_labels <- wine_test[,12]
table(wine_train_labels)
```

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

```
## wine_train_labels
##    Level 0: Low Level 1: Medium   Level 2: High
##             127            2372             684
```

```
table(wine_test_labels)
```

```
## wine_test_labels
##    Level 0: Low Level 1: Medium   Level 2: High
##              56            1283             376
```

```
library(class)
wine_test_pred <- knn(train = wine_train[,1:11], test = wine_test[,1:11], cl = wine_train[,12], k=10)
table(wine_test_pred)
```

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

```
## wine_test_pred
##    Level 0: Low Level 1: Medium   Level 2: High
##               3            1428             284
```

```
CM <- table(Actual = wine_test_labels, Predicted = wine_test_pred)
CM
```

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

```
##                   Predicted
## Actual             Level 0: Low Level 1: Medium Level 2: High
##   Level 0: Low                1              54             1
##   Level 1: Medium             2            1172           109
##   Level 2: High               0             202           174
```

```
library(e1071)
library(caret)
```

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

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

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

```
confusionMatrix(CM)
```

```
## Confusion Matrix and Statistics
##
##                   Predicted
## Actual             Level 0: Low Level 1: Medium Level 2: High
##   Level 0: Low                1              54             1
##   Level 1: Medium             2            1172           109
##   Level 2: High               0             202           174
##
## Overall Statistics
##
##                Accuracy : 0.7854
```

```
##                  95% CI : (0.7652, 0.8046)
##     No Information Rate : 0.8327
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3702
##
##  Mcnemar's Test P-Value : <2e-16
##
## Statistics by Class:
##
##                      Class: Level 0: Low Class: Level 1: Medium
## Sensitivity                    0.3333333                 0.8207
## Specificity                    0.9678738                 0.6132
## Pos Pred Value                 0.0178571                 0.9135
## Neg Pred Value                 0.9987945                 0.4074
## Prevalence                     0.0017493                 0.8327
## Detection Rate                 0.0005831                 0.6834
## Detection Prevalence           0.0326531                 0.7481
## Balanced Accuracy              0.6506036                 0.7170
##                      Class: Level 2: High
## Sensitivity                        0.6127
## Specificity                        0.8588
## Pos Pred Value                     0.4628
## Neg Pred Value                     0.9178
## Prevalence                         0.1656
## Detection Rate                     0.1015
## Detection Prevalence               0.2192
## Balanced Accuracy                  0.7358
```

This is the end of Assignment 3

Ceni Babaoglu, PhD