Wine Quality Prediction

Installation of the necessary packages (remove the # to install the packages):

```
#install.packages("class", repos="http://cran.us.r-project.org")
#install.packages("gmodels", repos="http://cran.us.r-project.org")
library("class")
library("gmodels")
#install.packages("corrplot")
library("corrplot")
```

```
## corrplot 0.84 loaded
```

Import the wine quality dataset.

```
wine_quality<-read.csv2(file='C://Users/Doina/Desktop/winequality-white2.csv',
header = T, sep = ";", dec = ".",stringsAsFactors = FALSE)
```

1.Check the data characteristics. Is there missing data?

```
str(wine_quality)
```

```
## '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_quality))
```

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

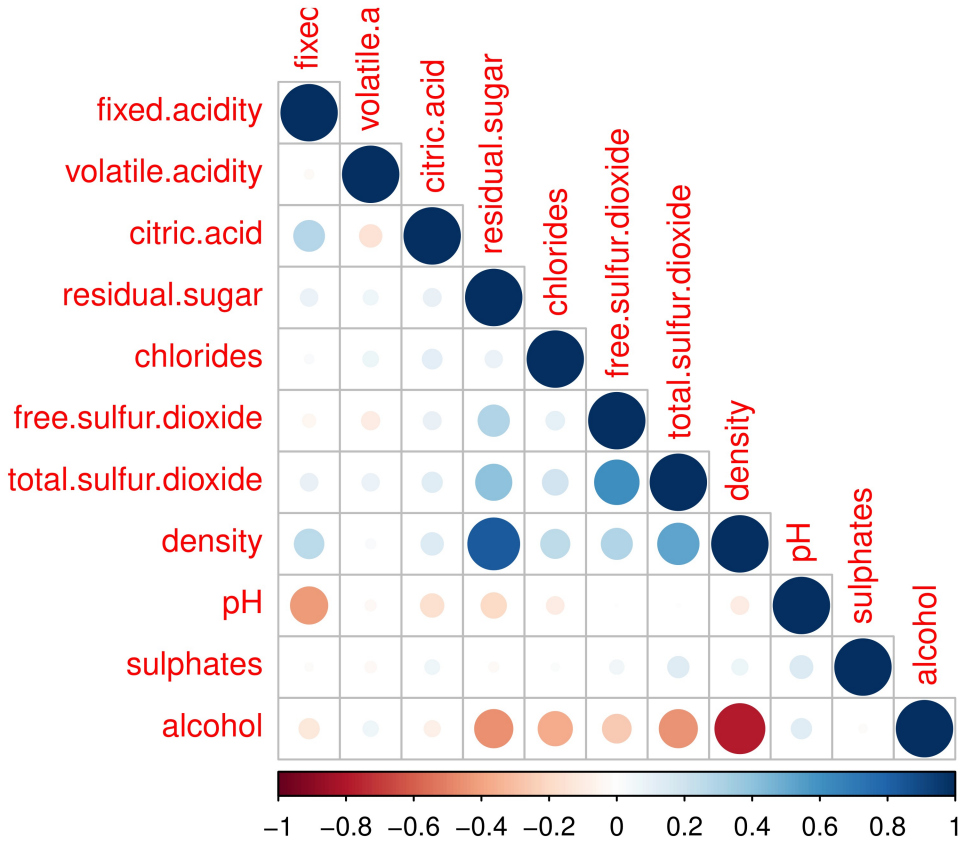
```
# There is no missing values
```

2.What is the correlation between the attributes other than wine quality

```
correlations<-cor(wine_quality[-c(12)])
correlations
```

```
##                       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
```

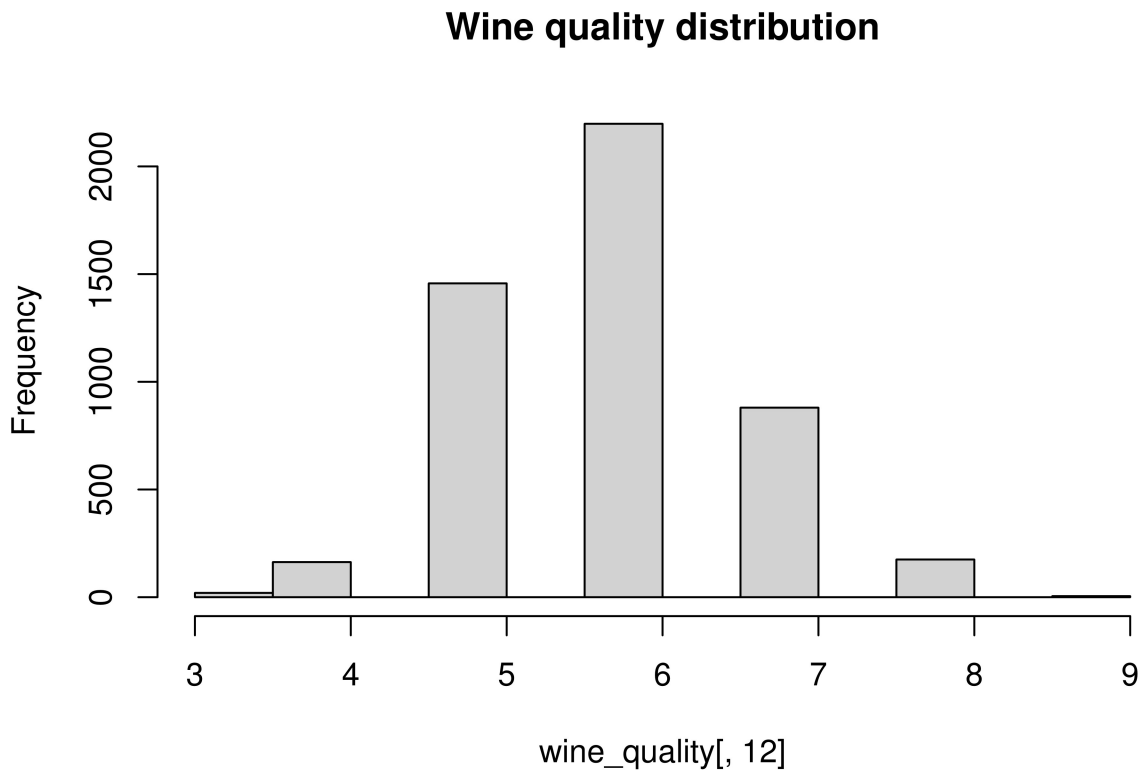
```
corrplot(correlations, type = "lower")
```




```
# There is a very strong correlation between density and residual sugar,
# a strong correlation between total.sulfur.dioxide and free.sulfur.dioxide,
# a moderate correlation between density and total.sulfur.dioxide
# a strong negative correlation between alcohol and density
# a moderate negative correlation between fixed.acidity and ph
# a moderate negative correlation between residual.sugar and alcohol
# a moderate negative correlation between total.sulfur.dioxide and alcohol
# a moderate correlation between total.sulfur.dioxide and residual sugar
```

3.Graph the frequency distribution of wine quality.

```
hist(wine_quality[,12], main="Wine quality distribution"  )
```



#4.Reduce the levels of rating for quality to three levels as high, medium and low.

```
wine_quality$quality<-as.factor(wine_quality$quality)
wine_quality$quality<-factor(wine_quality$quality, levels = c("3",'4','5','6','7','8','9'),
labels = c("low","low","low","medium","medium","high", "high"))
# as the database describes the quality of the the wine according to 10 levels,
#but in the database range between  3 and 9, I considered that levels: [3-5]="low",
#[6-7]="medium", [8-9]="high"
str(wine_quality)
```

```
## '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             : Factor w/ 3 levels "low","medium",..: 2 2 2 2 2 2 2 2 2 2 ...
```

5.Normalize the data set

```
normalize <- function(x) {
return ((x - min(x)) / (max(x) - min(x))) }
wine_quality_n <- as.data.frame(lapply(wine_quality[-c(12)], normalize))
wine_quality_n <- cbind(wine_quality_n,wine_quality$quality)
head(wine_quality_n)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1     0.3076923        0.1862745   0.2168675     0.30828221 0.1068249
## 2     0.2403846        0.2156863   0.2048193     0.01533742 0.1186944
## 3     0.4134615        0.1960784   0.2409639     0.09662577 0.1216617
## 4     0.3269231        0.1470588   0.1927711     0.12116564 0.1454006
## 5     0.3269231        0.1470588   0.1927711     0.12116564 0.1454006
## 6     0.4134615        0.1960784   0.2409639     0.09662577 0.1216617
##   free.sulfur.dioxide total.sulfur.dioxide   density        pH sulphates
## 1          0.14982578            0.3735499 0.2677848 0.2545455 0.2674419
## 2          0.04181185            0.2853828 0.1328321 0.5272727 0.3139535
## 3          0.09756098            0.2041763 0.1540389 0.4909091 0.2558140
## 4          0.15679443            0.4106729 0.1636784 0.4272727 0.2093023
## 5          0.15679443            0.4106729 0.1636784 0.4272727 0.2093023
## 6          0.09756098            0.2041763 0.1540389 0.4909091 0.2558140
##     alcohol wine_quality$quality
## 1 0.1290323               medium
## 2 0.2419355               medium
## 3 0.3387097               medium
## 4 0.3064516               medium
## 5 0.3064516               medium
## 6 0.3387097               medium
```

6.Divide the data to training and testinggroups

```
set.seed(1)
index <- sample(1:nrow(wine_quality_n), 0.7 *nrow(wine_quality_n))
wine_quality_train <- wine_quality_n[index,]
wine_quality_test <- wine_quality_n[-index,]
```

7.Use the KNN algorithm to predict the quality of wine using its attributes.

```
wine_quality_train_labels <- wine_quality_train[,12]
wine_quality_test_labels <- wine_quality_test[,12]

wine_quality_prediction<- knn(train = wine_quality_train[,1:11], test =wine_quality_test[,1:11],
cl = wine_quality_train[,12], k=7)
```

8.Evaluate the model performance

```
CrossTable(x=wine_quality_test_labels, y=wine_quality_prediction, prop.chisq=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1470
##
##
##                          | wine_quality_prediction
## wine_quality_test_labels |       low |    medium |      high | Row Total |
## -------------------------|-----------|-----------|-----------|-----------|
##                      low |       285 |       212 |         0 |       497 |
##                          |     0.573 |     0.427 |     0.000 |     0.338 |
##                          |     0.667 |     0.207 |     0.000 |           |
##                          |     0.194 |     0.144 |     0.000 |           |
## -------------------------|-----------|-----------|-----------|-----------|
##                   medium |       141 |       771 |        14 |       926 |
##                          |     0.152 |     0.833 |     0.015 |     0.630 |
##                          |     0.330 |     0.751 |     0.824 |           |
##                          |     0.096 |     0.524 |     0.010 |           |
## -------------------------|-----------|-----------|-----------|-----------|
##                     high |         1 |        43 |         3 |        47 |
##                          |     0.021 |     0.915 |     0.064 |     0.032 |
##                          |     0.002 |     0.042 |     0.176 |           |
##                          |     0.001 |     0.029 |     0.002 |           |
## -------------------------|-----------|-----------|-----------|-----------|
##             Column Total |       427 |      1026 |        17 |      1470 |
##                          |     0.290 |     0.698 |     0.012 |           |
## -------------------------|-----------|-----------|-----------|-----------|
##
##
```

```
#The accuracy of the model =(TP+TN)/N= (297+757+0)/1470=71.7%.
#We can observe that the model is not very good at predicting the high quality wines,
#but The model could be improved if the dataset whould be balanced.
#To begin with the high quality has very few entries in the test set, so the algorithm
#is bias towards the majority, the medium quality wine.
```