Práctica 2

Javier Muñoz García

6 de enero, 2020

Contents

Descripción del dataset

Para la realización de esta práctica se ha escogido el conjunto de datos propuesto en el enunciado compuesto por 1.599 observaciones y 12 variables.

Este conjunto de datos está relacionado con las variantes tintas del vino portugués "Vinho Verde". En el conjunto de datos solo están disponibles las variables físico-químicas como entradas y la calidad del vino (vairable sensorial) como salida.

Respecto a diferentes características sobre del vino. En concreto, las características son:

- **fixed acidity:** la mayoría de los ácidos involucrados con el vino o fijos o no volátiles (no se evaporan fácilmente).
- **volatile acidity:** la cantidad de ácido acético en el vino, que en niveles demasiado altos puede producir un sabor desagradable a vinagre ácido cítrico.
- **citric acid:** se encuentra en pequeñas cantidades, puede añadir "frescura" y sabor a los vinos.
- **residual sugar:** la cantidad de azúcar que queda después de la parada de la fermentación, es raro encontrar vinos con menos de 1 gramo/litro y los vinos con más de 45 gramos/litro se consideran dulces.
- **chlorides:** la cantidad de sal en el vino.
- **free sulfur dioxide:** la forma libre de SO2 existe en equilibrio entre el SO2 molecular (como gas disuelto) y el ión bisulfito; impide el crecimiento microbiano y la oxidación del vino.
- **total sulfur dioxide:** cantidad de formas libres y ligadas de S02; en bajas concentraciones, el SO2 es mayormente indetectable en el vino, pero en concentraciones libres de SO2 superiores a 50 ppm, el SO2 se hace evidente en la nariz y el sabor del vino.
- **density:** la densidad del agua se aproxima a la del agua dependiendo del porcentaje de alcohol y del contenido de azúcar.
- **pH:** describe cuán ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico); la mayoría de los vinos están entre 3-4 en la escala de pH.
- **sulphates:** un aditivo para el vino que puede contribuir a los niveles de dióxido de azufre (S02), que actúa como antimicrobiano y antioxidante.

- **alcohol:** el porcentaje de contenido de alcohol del vino.
- **quality:** puntuación subjetiva del vino (valor de 0 a 10).

El dataset está propuesto plantea la posiblidad de conocer la calidad de un vino en base a sus características, así como determinar que variables influyen más en la calidad. Bajo este supuesto, se podría clasificar un vino en base a su calidad sin la necesidad de realizar una cata con su posterior calificación humana.

Integración y selección de los datos de interés a analizar

Para comenzar, leeremos el fichero 'winequality.red' descargado de la web de Kaggle y revisaremos los atributos cargados.

```
# Carga de datos
winequality.red <- read.csv("C:/Users/Javier/Desktop/PRA2/winequality-red.csv")

# Comprobación de datos cargados
head(winequality.red)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

```
# Tipos de variables
types <- sapply(winequality.red,class)
kable(data.frame(Variables = names(types), Clase = as.vector(types)))
```

Variables	Clase
fixed.acidity	numeric
volatile.acidity	numeric
citric.acid	numeric
residual.sugar	numeric
chlorides	numeric
free.sulfur.dioxide	numeric
total.sulfur.dioxide	numeric
density	numeric
pH	numeric
sulphates	numeric

Variables	Clase
alcohol	numeric
quality	integer

Como podemos observar, todos son atributos númericos y válidos para intentar clasificar la calidad del vino, ya que a priori no conocemos que características o combinación de ellas son adecuadas para elevar la calidad del vino.

Limpieza de los datos

Elementos vacíos

Una vez hemos cargado los datos, procedemos a realizar la limpieza del mismo, lo primero será revisar valores nulos o falta de información en algún registro de nuestro conjunto de datos.

```
# Búsqueda de valores nulos
nas <- sapply(winequality.red, function(x) sum(is.na(x)))
kable(data.frame(Variables = names(nas), NAs = as.vector(nas)))
```

Variables	NAs
fixed.acidity	0
volatile.acidity	0
citric.acid	0
residual.sugar	0
chlorides	0
free.sulfur.dioxide	0
total.sulfur.dioxide	0
density	0
pH	0
sulphates	0
alcohol	0
quality	0

Así pues, dado que no hay ningún valor nulo, podemos saltarnos esta parta y revisar si hay algún dato demasiado extremo realizando una búsqueda de los mismos.

Valores extremos

```
# Búsqueda de valores extremos
outliers <- sapply(winequality.red, function(x) paste(boxplot.stats(x)$out,collapse=" "))
kable(data.frame(variables=names(outliers),clase=as.vector(outliers)))
```

variables	clase
fixed.acidity	12.8 12.8 15 15 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8 12.8 14 13.7 13.7 12.7 12.5 12.8 12.6 15
volatile.acidity	1.13 1.02 1.07 1.33 1.33 1.04 1.09 1.04 1.24 1.185 1.02 1.035 1.025 1.115 1.02 1.02 1.58 1.18 1.04
citric.acid	1
residual.sugar	6.1 6.1 3.8 3.9 4.4 10.7 5.5 5.9 5.9 3.8 5.1 4.65 4.65 5.5 5.5 5.5 5.5 7.3 7.2 3.8 5.6 4 4 4 4 7 4 4 6.4 5.6 5
chlorides	0.176 0.17 0.368 0.341 0.172 0.332 0.464 0.401 0.467 0.122 0.178 0.146 0.236 0.61 0.36 0.27 0.039 0.337
free.sulfur.dioxide	52 51 50 68 68 43 47 54 46 45 53 52 51 45 57 50 45 48 43 48 72 43 51 51 52 55 55 48 48 66
total.sulfur.dioxide	145 148 136 125 140 136 133 153 134 141 129 128 129 128 143 144 127 126 145 144 135 165 124 124 13

variables	clase
density	0.9916 0.9916 1.0014 1.0015 1.0015 1.0018 0.9912 1.0022 1.0022 1.0014 1.0014 1.0014 1.0014 1.0032 1.0
pH	3.9 3.75 3.85 2.74 3.69 3.69 2.88 2.86 3.74 2.92 2.92 2.92 3.72 2.87 2.89 2.89 2.92 3.9 3.71 3.69 3.69 3.71
sulphates	1.56 1.28 1.08 1.2 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2 1.08 1.59 1.02 1.03 1.61 1.09 1.26 1.08 1 1.30
alcohol	14 14 14 14 14.9 14 13.6 13.6 13.6 14 14 13.5666666666667 13.6
quality	8 8 8 8 8 3 8 8 8 3 8 3 8 3 3 8 8 8 8 8 3 3 8 8 3 3 3 8

```
summary(winequality.red)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

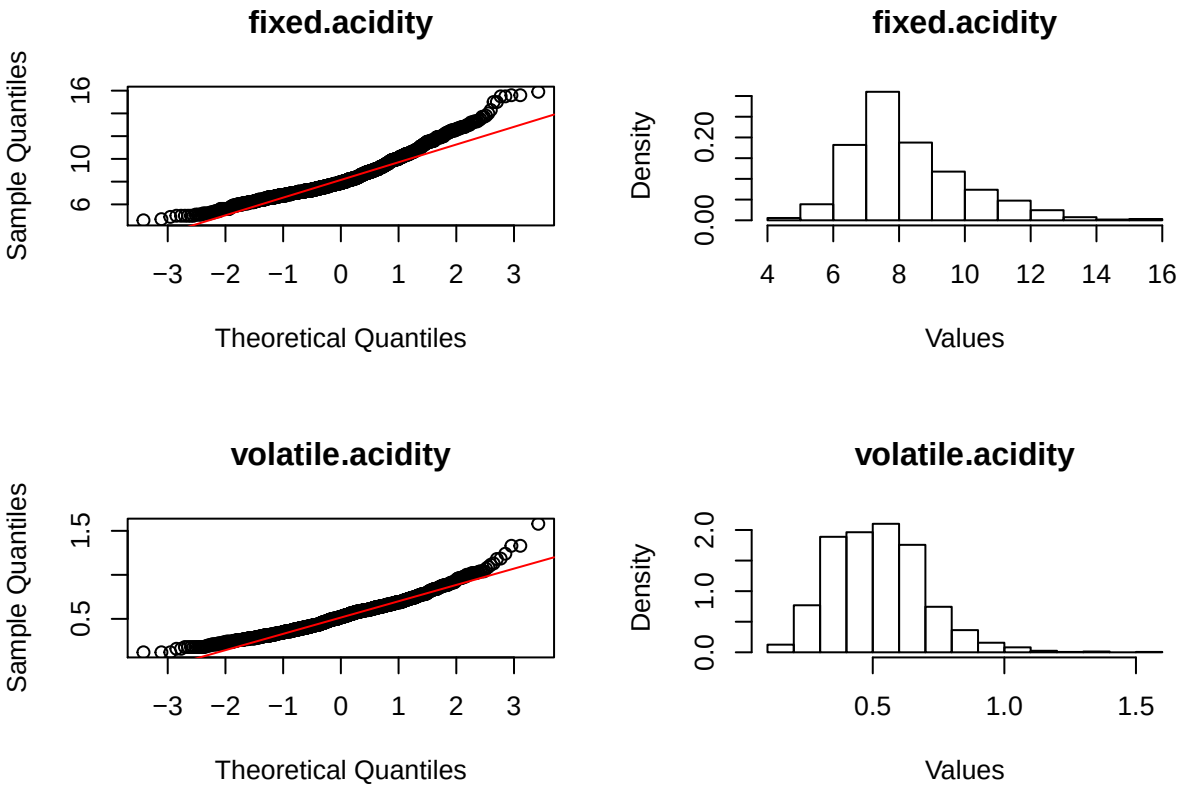
Tras el cálculo de outliers o valores extremos realizados por R, vemos que hay un gran número de ellos en todas las variables, sin embargo, no se trata de valores extremos a eliminar o modificar ya que son valores reales de las características de vinos determinados, aún teniendo valores extremos se trata de valores posibles si lo comparamos con el resumen de los datos extraídos, por lo que estos valoren pueden ser los que afecten a nuestro valor objetivo, decidimos dejarlos.

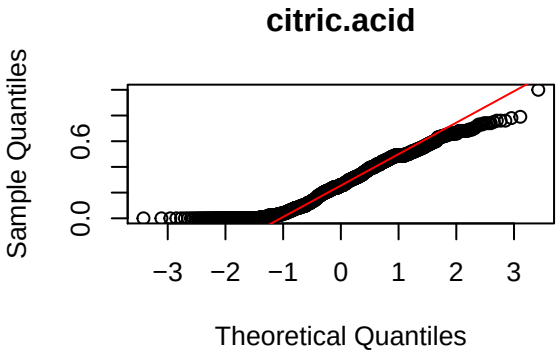
Análisis de los datos

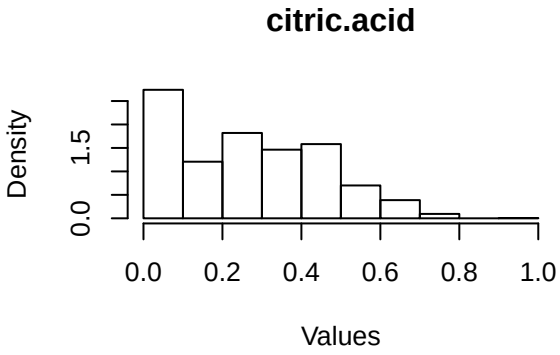
Comprobación de normalidad

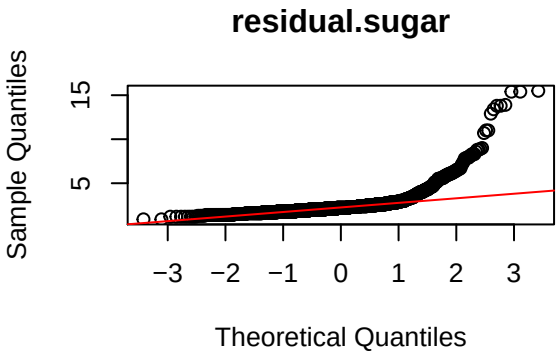
Procedemos a comprobar la normalidad de las características que tenemos, para ello usaremos la prueba de la normalidad de Shapiro Wilk y revisaremos la gráfica quantile-quantile de las características clasificadas como no normales.

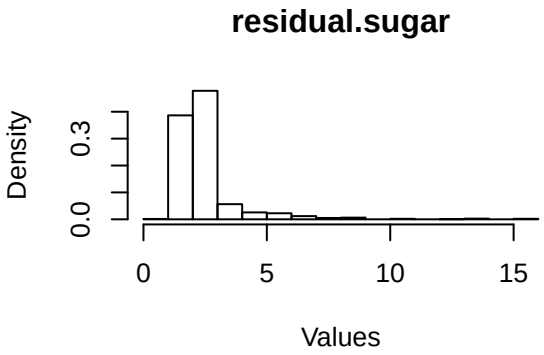
```r
par(mfrow = c(2, 2))
for (i in 1:ncol(winequality.red)) {
        if (shapiro.test(winequality.red[,i])$p.value < 0.05) {
            qqnorm(winequality.red[,i], main = colnames(winequality.red)[i])
            qqline(winequality.red[,i], col = "red")
            hist(winequality.red[,i], main = colnames(winequality.red)[i], xlab = "Values", freq = FALSE
        }
}
```

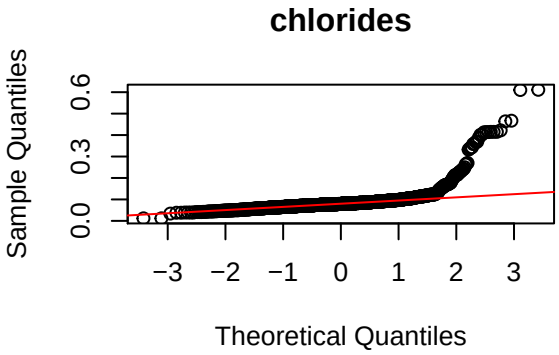


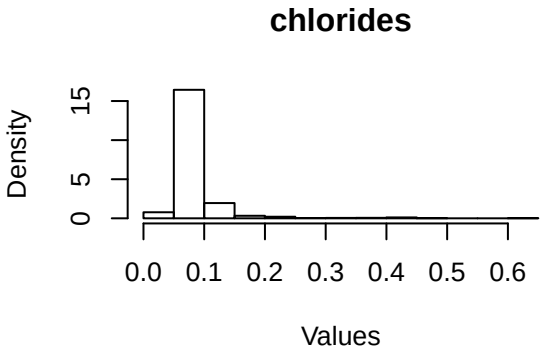

citric.acid
Sample Quantiles
0.0
0.6
−3
−2
−1
0
1
2
3
Theoretical Quantiles

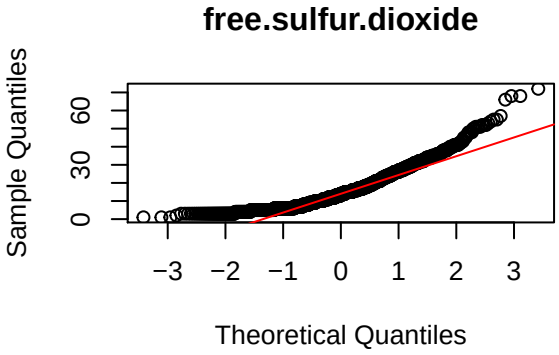

citric.acid
Density
0.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
Values

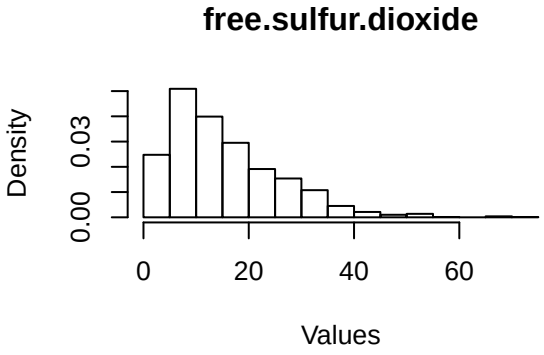

residual.sugar
Sample Quantiles
5
15
−3
−2
−1
0
1
2
3
Theoretical Quantiles

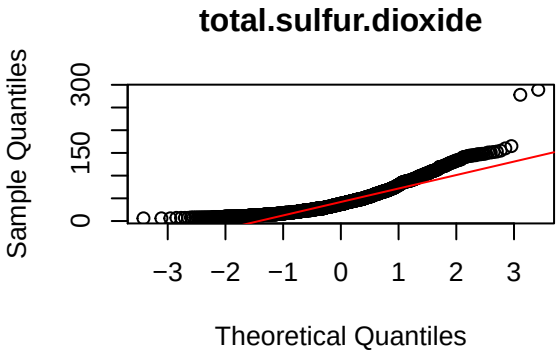

residual.sugar
Density
0.0
0.3
0
5
10
15
Values

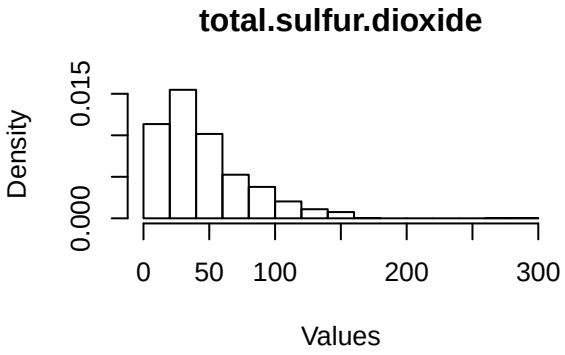

chlorides
Sample Quantiles
0.0
0.3
0.6
−3
−2
−1
0
1
2
3
Theoretical Quantiles

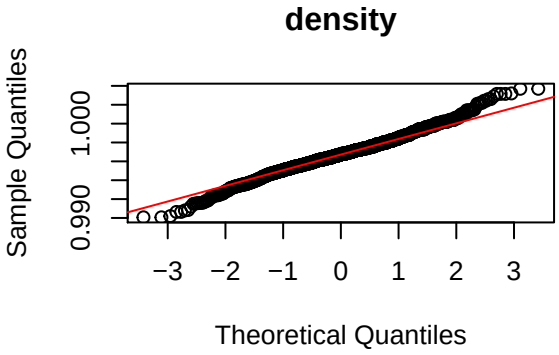

chlorides
Density
0
5
15
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Values

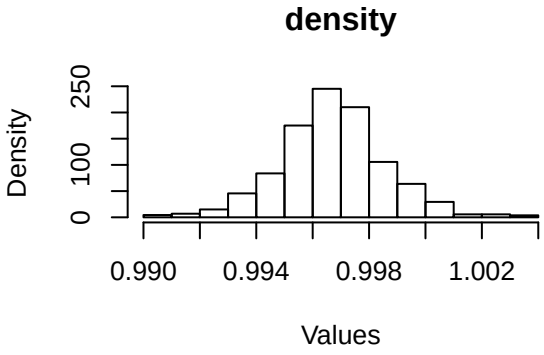

free.sulfur.dioxide
Sample Quantiles
0
30
60
−3
−2
−1
0
1
2
3
Theoretical Quantiles

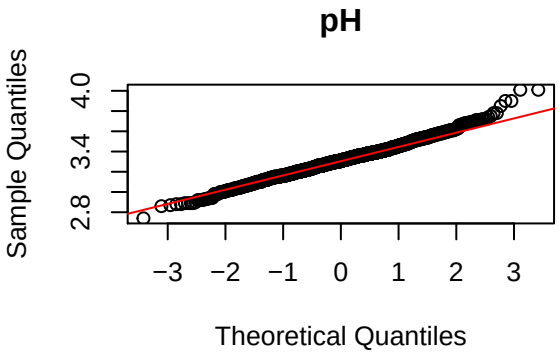

free.sulfur.dioxide
Density
0.00
0.03
0
20
40
60
Values

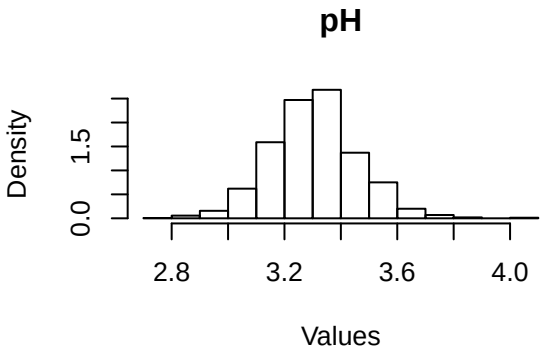

total.sulfur.dioxide
Sample Quantiles
0
150
300
−3
−2
−1
0
1
2
3
Theoretical Quantiles

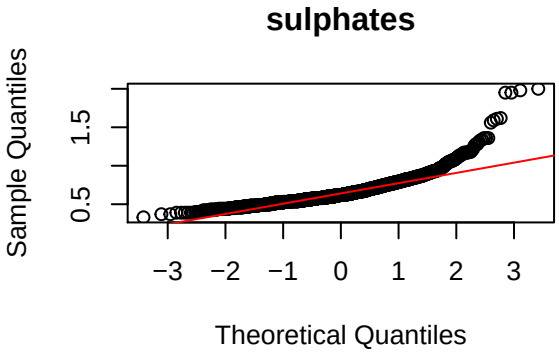

total.sulfur.dioxide
Density
0.000
0.015
0
50
100
200
300
Values

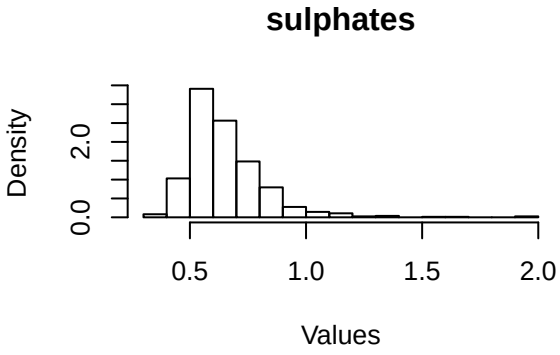

density
Sample Quantiles
0.990
1.000
−3
−2
−1
0
1
2
3
Theoretical Quantiles

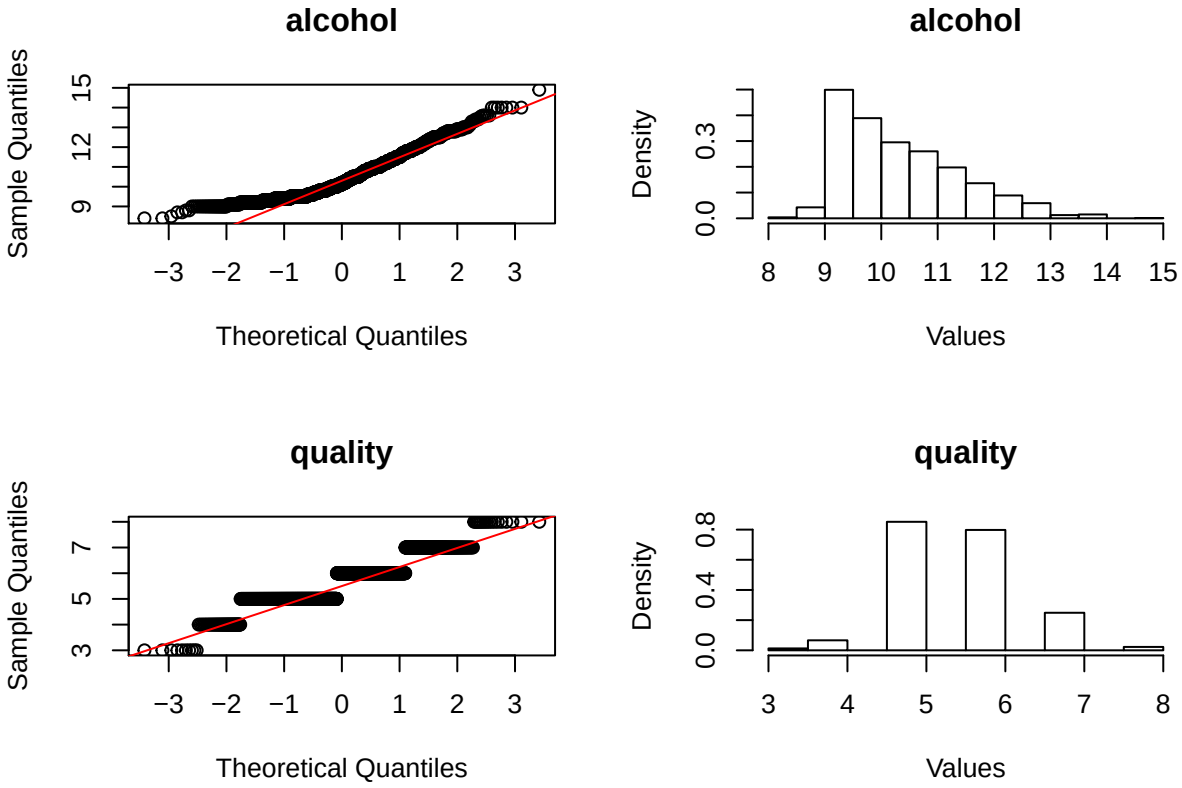

density
Density
0
100
250
0.990
0.994
0.998
1.002
Values


pH
Sample Quantiles
2.8
3.4
4.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles


pH
Density
0.0
1.5
2.8
3.2
3.6
4.0
Values


sulphates
Sample Quantiles
0.5
1.5
−3
−2
−1
0
1
2
3
Theoretical Quantiles


sulphates
Density
0.0
2.0
0.5
1.0
1.5
2.0
Values



Consultando el teorema del limite central, es posible aproximar nuestras variables a una distribución normal de media 0 y desviación 1 cuando se tienen más de 30 elementos, nuestro conjunto de datos tiene 1599 observaciones por lo que podemos aplicar el teorema.

Aplicación de pruebas estadísticas

Nuestro objetivo es intentar conocer la calidad de un vino en base a sus características, pero realmente no conocemos demasiado sobre las mismas, con la descripción del dataset, podemos intuir algunas características *importantes* como el ácido volátil (que puede afectar negativamente al vino en cantidad alta) o el ácido cítrico, que proporciona 'frescura' al vino afectando positivamente al mismo.

Para continuar nuestro estudio, realizaremos una pequeña prueba con el coeficiente de Spearman para ver el grado de correlación de las variables.

```
values <- character(length(colnames(winequality.red))-1)
for (i in 1:(ncol(winequality.red)-1)) {
    values[i] = cor.test(winequality.red[,i], winequality.red[,length(winequality.red)], method = "spea
}
kable(data.frame(Variables = colnames(winequality.red[1:(length(winequality.red)-1)]), Estimate = value
```

Variables	Estimate
fixed.acidity	0.114083673588803
volatile.acidity	-0.380646510425376
citric.acid	0.213480914422136
residual.sugar	0.0320481675290573
chlorides	-0.189922335617307
free.sulfur.dioxide	-0.0569006455015393

Variables	Estimate
total.sulfur.dioxide	-0.196735075441221
density	-0.177074065972
pH	-0.0436719347889284
sulphates	0.377060199102122
alcohol	0.478531687470243

Como se predecía, no solo las variables que se indicaban afectan a la calidad del vino, hemos descubierto que la variable que afecta más a la calidad del mismo es el alcohol y los sulfatos de manera positiva y como bien comentabamos antes, la que más lo perjudica es la ácidez volátil.

Ahora realizaremos una regresión líneal múltiple para determinar si es posible predecir la calidad del vino en base a las variables que disponemos, en primer lugar, separaremos nuestro conjunto de datos en dos, para el entrenamiento y prueba del modelo.

```
# Establecemos una semilla para separar ambos conjuntos de datos aleatoriamente
set.seed(1)
separator = sample(1:nrow(winequality.red), size = nrow(winequality.red)*0.8)
# Separamos ambos conjuntos
wine_train = winequality.red[separator,]
wine_test = winequality.red[-separator,]
```

Y creamos nuestro modelo en función de las *mejores* variables descubiertas anteriormente.

```
# Creamos un modelo de regresión lineal múltiple con las variables de más importancia
regModel = lm(quality~alcohol+volatile.acidity+sulphates, data = wine_train)
summary(regModel)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + sulphates,
##     data = wine_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6900 -0.3893 -0.0743  0.4847  2.2096
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.69345    0.22370  12.041  < 2e-16 ***
## alcohol           0.30471    0.01789  17.028  < 2e-16 ***
## volatile.acidity -1.18643    0.11044 -10.742  < 2e-16 ***
## sulphates         0.58374    0.11274   5.178 2.61e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6636 on 1275 degrees of freedom
## Multiple R-squared:  0.3169, Adjusted R-squared:  0.3153
## F-statistic: 197.2 on 3 and 1275 DF,  p-value: < 2.2e-16
```

Ahora, como segunda aproximación, intentaremos abarcar el mismo problema pero con un método de clasificación, para ello crearemos una variable objetivo categórica e intentaremos clasificar los vinos entre tres etiquetas de calidad con un árbol de decisión simple.

```r
# Creación de variable categórica
winequality.red[,"label"] <- cut(winequality.red$quality, breaks = c(3,5,6,8), labels = c("LOW", "MEDIU

# Separación de variables y variable objetivo
X = winequality.red[,1:11]
y = winequality.red[,13]

# Establecemos una semilla para separar ambos conjuntos de datos aleatoriamente
set.seed(1)
separator = sample(1:nrow(winequality.red), size = nrow(winequality.red)*0.8)
# Separamos ambos conjuntos
trainX = X[separator,]
trainy = y[separator]
testX = X[-separator,]
testy = y[-separator]

#Creación del árbol
library(C50)
```

```
## Warning: package 'C50' was built under R version 3.4.4
```

```r
model <- C50::C5.0(trainX,trainy)

#Extracción de reglas y resultados
summary(model)
```

```
##
## Call:
## C5.0.default(x = trainX, y = trainy)
##
##
## C5.0 [Release 2.07 GPL Edition]      Mon Jan 06 20:42:26 2020
## -------------------------------
##
## Class specified by attribute `outcome'
## *** ignoring cases with bad or unknown class
##
## Read 1273 cases (12 attributes) from undefined.data
##
## Decision tree:
##
## alcohol > 10.5:
## :...citric.acid <= 0.27:
## :   :...sulphates <= 0.58:
## :   :   :...density <= 0.99252: MEDIUM (7)
## :   :   :   density > 0.99252:
## :   :   :   :...total.sulfur.dioxide <= 27:
## :   :   :       :...pH <= 3.3: MEDIUM (3/1)
## :   :   :       :   pH > 3.3: LOW (28/4)
## :   :   :       total.sulfur.dioxide > 27:
## :   :   :       :...alcohol <= 11.1:
## :   :   :           :...alcohol > 10.8: LOW (11)
## :   :   :           :   alcohol <= 10.8:
## :   :   :           :   :...volatile.acidity <= 0.58: LOW (4)
```

```
## :   :   :           :       volatile.acidity > 0.58: MEDIUM (7)
## :   :   :           alcohol > 11.1:
## :   :   :           :...fixed.acidity <= 6: LOW (2)
## :   :   :               fixed.acidity > 6:
## :   :   :               :...volatile.acidity <= 0.755: MEDIUM (18)
## :   :   :                   volatile.acidity > 0.755:
## :   :   :                   :...chlorides <= 0.062: MEDIUM (2/1)
## :   :   :                       chlorides > 0.062: LOW (2)
## :   :   sulphates > 0.58:
## :   :   :...volatile.acidity > 0.9: LOW (5)
## :   :       volatile.acidity <= 0.9:
## :   :       :...total.sulfur.dioxide > 77:
## :   :           :...density <= 0.99294: HIGH (7)
## :   :           :   density > 0.99294:
## :   :           :   :...chlorides > 0.073: LOW (5)
## :   :           :       chlorides <= 0.073:
## :   :           :       :...volatile.acidity <= 0.55: LOW (2)
## :   :           :           volatile.acidity > 0.55: MEDIUM (2)
## :   :           total.sulfur.dioxide <= 77:
## :   :           :...citric.acid > 0.18:
## :   :               :...sulphates > 0.91: LOW (2)
## :   :               :   sulphates <= 0.91:
## :   :               :   :...free.sulfur.dioxide > 14: MEDIUM (9)
## :   :               :       free.sulfur.dioxide <= 14:
## :   :               :       :...total.sulfur.dioxide <= 26: MEDIUM (4)
## :   :               :           total.sulfur.dioxide > 26: LOW (3)
## :   :               citric.acid <= 0.18:
## :   :               :...chlorides > 0.093:
## :   :                   :...total.sulfur.dioxide <= 32: LOW (3)
## :   :                   :   total.sulfur.dioxide > 32: MEDIUM (6)
## :   :                   chlorides <= 0.093:
## :   :                   :...volatile.acidity > 0.665: MEDIUM (18/1)
## :   :                       volatile.acidity <= 0.665:
## :   :                       :...volatile.acidity > 0.645: HIGH (4)
## :   :                           volatile.acidity <= 0.645:
## :   :                           :...residual.sugar > 4.8: HIGH (2)
## :   :                               residual.sugar <= 4.8:
## :   :                               :...density <= 0.99693: MEDIUM (46/9)
## :   :                                   density > 0.99693: HIGH (4/1)
## :   citric.acid > 0.27:
## :   :...alcohol > 11.5:
## :       :...volatile.acidity > 0.565: MEDIUM (7/1)
## :       :   volatile.acidity <= 0.565:
## :       :   :...residual.sugar > 4.6:
## :       :       :...alcohol <= 12.8: HIGH (10)
## :       :       :   alcohol > 12.8: LOW (2)
## :       :       residual.sugar <= 4.6:
## :       :       :...sulphates <= 0.69:
## :       :           :...total.sulfur.dioxide <= 16:
## :       :           :   :...free.sulfur.dioxide > 4: HIGH (11/1)
## :       :           :   :   free.sulfur.dioxide <= 4:
## :       :           :   :   :...density <= 0.99374: HIGH (2)
## :       :           :   :       density > 0.99374:
## :       :           :   :       :...residual.sugar <= 3.9: MEDIUM (7)
```

```
## :       :           :   :           residual.sugar > 3.9: HIGH (2)
## :       :           :   total.sulfur.dioxide > 16:
## :       :           :   :...alcohol > 11.8: MEDIUM (21/2)
## :       :           :       alcohol <= 11.8:
## :       :           :       :...total.sulfur.dioxide > 56: HIGH (3)
## :       :           :           total.sulfur.dioxide <= 56:
## :       :           :           :...alcohol <= 11.7: MEDIUM (6/1)
## :       :           :               alcohol > 11.7:
## :       :           :               :...residual.sugar <= 2.4: MEDIUM (2)
## :       :           :                   residual.sugar > 2.4: LOW (2)
## :       :           sulphates > 0.69:
## :       :           :...residual.sugar > 2.2: HIGH (28/2)
## :       :               residual.sugar <= 2.2:
## :       :               :...citric.acid <= 0.36: HIGH (6)
## :       :                   citric.acid > 0.36:
## :       :                   :...volatile.acidity <= 0.2: HIGH (3)
## :       :                       volatile.acidity > 0.2:
## :       :                       :...free.sulfur.dioxide > 16: MEDIUM (5)
## :       :                           free.sulfur.dioxide <= 16:
## :       :                           :...free.sulfur.dioxide <= 8: MEDIUM (5/1)
## :       :                               free.sulfur.dioxide > 8: HIGH (3)
## :       alcohol <= 11.5:
## :       :...volatile.acidity <= 0.37:
## :           :...total.sulfur.dioxide > 93: LOW (3)
## :           :   total.sulfur.dioxide <= 93:
## :           :   :...pH <= 3.26:
## :           :       :...residual.sugar <= 3.2: HIGH (28/7)
## :           :       :   residual.sugar > 3.2: MEDIUM (3)
## :           :       pH > 3.26:
## :           :       :...pH > 3.39: HIGH (8/1)
## :           :           pH <= 3.39:
## :           :           :...fixed.acidity > 11.5: HIGH (2)
## :           :               fixed.acidity <= 11.5:
## :           :               :...sulphates > 0.7: MEDIUM (19/2)
## :           :                   sulphates <= 0.7:
## :           :                   :...density > 0.99714: LOW (4)
## :           :                       density <= 0.99714:
## :           :                       :...total.sulfur.dioxide <= 10: HIGH (2)
## :           :                           total.sulfur.dioxide > 10:
## :           :                           :...sulphates <= 0.6: MEDIUM (4)
## :           :                               sulphates > 0.6: LOW (4/1)
## :           volatile.acidity > 0.37:
## :           :...alcohol <= 10.75: MEDIUM (7)
## :               alcohol > 10.75:
## :               :...total.sulfur.dioxide > 59: MEDIUM (9)
## :                   total.sulfur.dioxide <= 59:
## :                   :...total.sulfur.dioxide > 57: HIGH (2)
## :                       total.sulfur.dioxide <= 57:
## :                       :...chlorides > 0.106: LOW (6/1)
## :                           chlorides <= 0.106:
## :                           :...citric.acid <= 0.32:
## :                               :...citric.acid <= 0.29: MEDIUM (2)
## :                               :   citric.acid > 0.29: HIGH (4)
## :                               citric.acid > 0.32:
```

```
## :                                :...volatile.acidity > 0.59: LOW (6/1)
## :                                    volatile.acidity <= 0.59:
## :                                    :...alcohol > 11.2: MEDIUM (13/1)
## :                                        alcohol <= 11.2:
## :                                        :...chlorides > 0.097: LOW (3)
## :                                            chlorides <= 0.097:
## :                                            :...sulphates <= 0.59: LOW (2)
## :                                                sulphates > 0.59: [S1]
## alcohol <= 10.5:
## :...volatile.acidity <= 0.315:
##     :...sulphates <= 0.63:
##     :   :...alcohol <= 9.7: LOW (9/1)
##     :   :   alcohol > 9.7:
##     :   :   :...total.sulfur.dioxide <= 34: LOW (2)
##     :   :       total.sulfur.dioxide > 34: MEDIUM (4)
##     :   sulphates > 0.63:
##     :   :...fixed.acidity > 11.7:
##     :       :...residual.sugar <= 2.4: HIGH (7/1)
##     :       :   residual.sugar > 2.4: MEDIUM (2)
##     :       fixed.acidity <= 11.7:
##     :       :...density > 1.0001: LOW (2)
##     :           density <= 1.0001:
##     :           :...alcohol > 10.4: HIGH (3/1)
##     :               alcohol <= 10.4:
##     :               :...density > 0.99654: MEDIUM (29/3)
##     :                   density <= 0.99654:
##     :                   :...fixed.acidity <= 6.7: MEDIUM (2/1)
##     :                       fixed.acidity > 6.7: LOW (2)
##     volatile.acidity > 0.315:
##     :...total.sulfur.dioxide > 98:
##         :...sulphates <= 1.17: LOW (84/3)
##         :   sulphates > 1.17:
##         :   :...alcohol <= 9.8: LOW (2)
##         :       alcohol > 9.8: MEDIUM (3)
##         total.sulfur.dioxide <= 98:
##         :...sulphates <= 0.53:
##             :...alcohol <= 9.5: LOW (66/3)
##             :   alcohol > 9.5:
##             :   :...density <= 0.99578: LOW (19/1)
##             :       density > 0.99578:
##             :       :...volatile.acidity > 0.79: LOW (13)
##             :           volatile.acidity <= 0.79:
##             :           :...chlorides > 0.083:
##             :               :...free.sulfur.dioxide <= 15: LOW (14)
##             :               :   free.sulfur.dioxide > 15: MEDIUM (6/1)
##             :               chlorides <= 0.083:
##             :               :...residual.sugar > 2.55: LOW (6)
##             :                   residual.sugar <= 2.55:
##             :                   :...volatile.acidity <= 0.48: MEDIUM (5)
##             :                       volatile.acidity > 0.48:
##             :                       :...density <= 0.99613: LOW (4)
##             :                           density > 0.99613:
##             :                           :...density <= 0.99745: MEDIUM (7)
##             :                               density > 0.99745: LOW (4/1)
```

```
##             sulphates > 0.53:
##             :...alcohol > 9.8:
##                 :...residual.sugar > 5.2:
##                 :   :...fixed.acidity <= 7.3: LOW (4/1)
##                 :   :   fixed.acidity > 7.3: HIGH (4)
##                 :   residual.sugar <= 5.2:
##                 :   :...pH <= 3.05:
##                 :       :...residual.sugar <= 1.75: HIGH (3)
##                 :       :   residual.sugar > 1.75: LOW (6)
##                 :       pH > 3.05:
##                 :       :...pH > 3.42:
##                 :           :...sulphates > 0.88: MEDIUM (2)
##                 :           :   sulphates <= 0.88:
##                 :           :   :...free.sulfur.dioxide > 16: LOW (10)
##                 :           :       free.sulfur.dioxide <= 16:
##                 :           :       :...total.sulfur.dioxide > 36: MEDIUM (3)
##                 :           :           total.sulfur.dioxide <= 36:
##                 :           :           :...chlorides > 0.076: LOW (11/1)
##                 :           :               chlorides <= 0.076:
##                 :           :               :...sulphates <= 0.63: LOW (2)
##                 :           :                   sulphates > 0.63: MEDIUM (2)
##                 :           pH <= 3.42:
##                 :           :...citric.acid <= 0.04:
##                 :               :...alcohol <= 10.4: MEDIUM (9)
##                 :               :   alcohol > 10.4:
##                 :               :   :...fixed.acidity <= 7.6: MEDIUM (3)
##                 :               :       fixed.acidity > 7.6: HIGH (2)
##                 :               citric.acid > 0.04:
##                 :               :...sulphates > 0.63:
##                 :                   :...chlorides > 0.116: LOW (4/1)
##                 :                   :   chlorides <= 0.116: [S2]
##                 :                   sulphates <= 0.63:
##                 :                   :...total.sulfur.dioxide > 61: [S3]
##                 :                       total.sulfur.dioxide <= 61:
##                 :                       :...sulphates <= 0.54: LOW (6)
##                 :                           sulphates > 0.54:
##                 :                           :...sulphates <= 0.56: MEDIUM (8/1)
##                 :                               sulphates > 0.56:
##                 :                               :...pH > 3.37: MEDIUM (12/3)
##                 :                                   pH <= 3.37: [S4]
##                 alcohol <= 9.8:
##                 :...fixed.acidity > 9.9:
##                     :...free.sulfur.dioxide <= 5: LOW (4)
##                     :   free.sulfur.dioxide > 5:
##                     :   :...free.sulfur.dioxide <= 13:
##                     :       :...sulphates <= 0.77: MEDIUM (18/1)
##                     :       :   sulphates > 0.77: LOW (3/1)
##                     :       free.sulfur.dioxide > 13:
##                     :       :...citric.acid <= 0.48: MEDIUM (3/1)
##                     :           citric.acid > 0.48: LOW (4)
##                     fixed.acidity <= 9.9:
##                     :...fixed.acidity > 9.2: LOW (23/1)
##                         fixed.acidity <= 9.2:
##                         :...fixed.acidity <= 6.5: LOW (16/1)
```

```
##                               fixed.acidity > 6.5:
##                               :...volatile.acidity > 0.645:
##                                   :...alcohol <= 9.05: MEDIUM (4/1)
##                                   :   alcohol > 9.05:
##                                   :   :...fixed.acidity > 9:
##                                   :       :...alcohol <= 9.55: MEDIUM (5)
##                                   :       :   alcohol > 9.55: LOW (3)
##                                   :       fixed.acidity <= 9:
##                                   :       :...alcohol <= 9.4: LOW (30/1)
##                                   :           alcohol > 9.4:
##                                   :           :...alcohol > 9.566667: LOW (25/3)
##                                   :               alcohol <= 9.566667: [S5]
##                                   volatile.acidity <= 0.645:
##                                   :...citric.acid <= 0: MEDIUM (8)
##                                       citric.acid > 0:
##                                       :...total.sulfur.dioxide <= 35:
##                                           :...volatile.acidity > 0.605:
##                                           :   :...chlorides <= 0.077: MEDIUM (2)
##                                           :   :   chlorides > 0.077: LOW (8)
##                                           :   volatile.acidity <= 0.605:
##                                           :   :...density > 0.9971: [S6]
##                                           :       density <= 0.9971:
##                                           :       :...density > 0.9967: HIGH (4/1)
##                                           :           density <= 0.9967:
##                                           :           :...density > 0.9961: LOW (8/1)
##                                           :               density <= 0.9961: [S7]
##                                           total.sulfur.dioxide > 35: [S8]
##
## SubTree [S1]
##
## sulphates <= 0.85: MEDIUM (19/3)
## sulphates > 0.85: LOW (3/1)
##
## SubTree [S2]
##
## total.sulfur.dioxide <= 83: MEDIUM (40/10)
## total.sulfur.dioxide > 83: LOW (2)
##
## SubTree [S3]
##
## total.sulfur.dioxide <= 78: MEDIUM (12)
## total.sulfur.dioxide > 78: LOW (4/1)
##
## SubTree [S4]
##
## volatile.acidity <= 0.53: LOW (12)
## volatile.acidity > 0.53:
## :...residual.sugar > 2.3: MEDIUM (2)
##     residual.sugar <= 2.3:
##     :...alcohol > 10.1: LOW (5)
##         alcohol <= 10.1:
##         :...density <= 0.99765: LOW (3)
##             density > 0.99765: MEDIUM (2)
##
```

```
## SubTree [S5]
##
## chlorides <= 0.08: MEDIUM (4)
## chlorides > 0.08: LOW (10/2)
##
## SubTree [S6]
##
## citric.acid <= 0.24: MEDIUM (12)
## citric.acid > 0.24:
## :...chlorides <= 0.161: LOW (4/1)
##     chlorides > 0.161: MEDIUM (3)
##
## SubTree [S7]
##
## alcohol > 9.7: MEDIUM (4)
## alcohol <= 9.7:
## :...citric.acid <= 0.06: MEDIUM (3)
##     citric.acid > 0.06: LOW (2)
##
## SubTree [S8]
##
## free.sulfur.dioxide <= 6: MEDIUM (4)
## free.sulfur.dioxide > 6:
## :...volatile.acidity <= 0.38:
##     :...pH <= 3.32: LOW (4/1)
##     :   pH > 3.32: MEDIUM (7)
##     volatile.acidity > 0.38:
##     :...pH <= 3.16: LOW (17)
##         pH > 3.16:
##         :...chlorides <= 0.078:
##             :...total.sulfur.dioxide <= 93: LOW (30/1)
##             :   total.sulfur.dioxide > 93: MEDIUM (2)
##             chlorides > 0.078:
##             :...total.sulfur.dioxide > 78: MEDIUM (12/1)
##                 total.sulfur.dioxide <= 78:
##                 :...total.sulfur.dioxide > 64: LOW (9)
##                     total.sulfur.dioxide <= 64:
##                     :...residual.sugar > 2.3: MEDIUM (7/1)
##                         residual.sugar <= 2.3:
##                         :...density > 0.99803: MEDIUM (4)
##                             density <= 0.99803:
##                             :...sulphates > 0.6: LOW (11)
##                                 sulphates <= 0.6:
##                                 :...volatile.acidity <= 0.42: LOW (2)
##                                     volatile.acidity > 0.42: MEDIUM (2)
##
##
## Evaluation on training data (1273 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##     153   96( 7.5%)   <<
```

```
##
##
##     (a)   (b)   (c)    <-classified as
##    ----  ----  ----
##     577    24     5    (a): class LOW
##      27   461    10    (b): class MEDIUM
##       7    23   139    (c): class HIGH
##
##
##  Attribute usage:
##
##  100.00% alcohol
##   95.99% volatile.acidity
##   90.97% sulphates
##   88.69% total.sulfur.dioxide
##   61.74% citric.acid
##   34.56% fixed.acidity
##   34.17% residual.sugar
##   29.62% pH
##   29.14% chlorides
##   27.89% density
##   19.01% free.sulfur.dioxide
##
##
## Time: 0.0 secs
```

Representación de resultados

Para reprentar los resultados, utilizaremos ambos conjuntos de test generados para este momento, aplicaremos los modelos creados con estos nuevos conjuntos y revisaremos si los modelos son fiables.

```
# Aplicación de modelo creado a conjunto de test
predicted = predict(regModel, wine_test, type = "response")
realVsPred = data.frame(real = wine_test$quality,
                        predecido = predicted,
                        diferencia = wine_test$quality - predicted)
head(realVsPred)
```

```
##    real predecido  diferencia
## 6     5  5.101547 -0.10154722
## 21    6  5.606065  0.39393518
## 23    5  5.609205 -0.60920508
## 29    5  5.036388 -0.03638829
## 33    5  5.287736 -0.28773586
## 34    6  5.143451  0.85654855
```

```
# Media de la diferencia en las predicciones
print(mean(realVsPred$diferencia))
```

```
## [1] 0.04322426
```

```
#Prueba del modelo de clasificación
p <- predict(model, testX, type="class")
hit = 0
for (i in 1:length(testX)) {
    if (p[i] == testy[i]) {
```

```
        hit = hit + 1
    }
}
hit/length(testX)
```

```
## [1] 0.7272727
```

Resolución del problema

Como vemos, en ambos modelos tenemos un porcentaje muy alto de aciertos con lo que podemos concluir que si es posible conocer la calidad de un vino (o al menos una nota subjetiva real) a partir de sus características físico-químicas.