Práctica 2: Limpieza y análisis de datos

Autor: Ainara Romero y Fidel Romero

junio 2022

Contents

Resolución práctica

1.1- Descripción del dataset

El conjunto de datos se ha obtenido del repositorio *kaggle* mediante el enlace https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009. El *dataset* recoge características fisico-químicas de 1599 diferentes variantes de vino tinto y blanco del vino "Vinho Verde" portugués. Además incluye la calidad de cada muestra basada en datos sensoriales.

En primer lugar se carga el fichero de datos:

```
# Conjunto de datos:
df<-read.csv("winequality-red.csv",header=TRUE)
```

```
# Estructura de los datos
str(df)
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
```

```
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

El conjunto de datos está compuesto por 12 variables y 1599 observaciones. Los campos que constituyen el *dataset* son los siguientes, en ellos se hace una breve descripción de cada característica para una mejor comprensión:

Características físico- químicas:

- **fixed.acidity:** cantidad de ácido fijo. La acidez fija es el conjunto de ácidos naturales del vino que hacen que se preservan las cualidades naturales del vino, así como el color. Variable numérica.
- **volatile.acidity:** cantidad de ácido volátil. La acidez volátil es el ácido acético de un vino que en niveles demasiado altos puede dar lugar a un desagradable sabor a vinagre. Variable numérica.
- **citric.acid:** cantidad de ácido cítrico. Variable numérica.
- **residual.sugar:** cantidad de azúcar residual que queda tras la fermentación. Variable numérica.
- **chlorides:** cantidad de cloruro que contiene el vino. Variable numérica.

- **free.sulfur.dioxide:** cantidad de forma libre de SO2 que impide el crecimiento microbiano y la oxidación del vino. Variable numérica.
- **total.sulfur.dioxide:** cantidad de formas libres y ligadas de S02. En concentraciones altas, es evidente en el olfato y gusto del vino. Variable numérica.
- **density:** indica la densidad del agua. Variable numérica.
- **pH:** El ph describe el grado de acidez o base de un vino en una escala de 0 (muy ácido) a 14 (muy básico). Variable numérica.
- **sulphates** cantidad de sulfatos. Son un tipo de aditivo para el vino que actúan como antimicrobial y antioxidante. Variable numérica.
- **alcohol:** porcentaje de alcohol. Variable numérica.

Datos sensoriales:

- **quality:** variable numérica que cuantifica en una escala de 0 a 10 la calidad del vino basada en datos sensoriales.

Se observa que las variables que se han cargado corresponden con el tipo de variables del conjunto de datos. En este caso, todas las variables son cuantitativas. Se renombran para una mejor comprensión:

```
colnames(df)<-c("Acidez fija", "Acidez volátil", "Ácido cítrico", "Azúcar",
                "Cloruros", "SO2 libre", "Total SO2", "Densidad", "pH", "Sulfatos",
                "Alcohol", "Calidad")
```

Es interesante obtener una primera aproximación de la distribución del conjunto de datos para análisis posteriores:

```
# Distribución
summary(df)
```

```
##   Acidez fija    Acidez volátil   Ácido cítrico       Azúcar
##  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
##     Cloruros         SO2 libre       Total SO2         Densidad
##  Min.   :0.01200   Min.   : 1.00   Min.   :  6.00   Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00   1st Qu.: 22.00   1st Qu.:0.9956
##  Median :0.07900   Median :14.00   Median : 38.00   Median :0.9968
##  Mean   :0.08747   Mean   :15.87   Mean   : 46.47   Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00   3rd Qu.: 62.00   3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00   Max.   :289.00   Max.   :1.0037
##        pH           Sulfatos         Alcohol         Calidad
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

Se obtienen las siguientes conclusiones relevantes sobre la distribución de las variables:

- **Total SO2**: los valores de este atributo se encuentran en el rango 6 y 289. La media está en 46.47 y el 50% de las muestras su sulfato total es menor o igual a 38. Existe diferencia entre la media y la mediana, por lo que los valores pueden estar dispersos y será interesante el análisis de *outliers* de esta variable.
- En general, las demás variables no presentan grandes diferencias entre la media y mediana, por lo que sus valores no se encuentran muy dispersos y la presencia de *outliers* disminuirá, aún así se analizará más detalladamente en el siguiente apartado.
- **pH:** las muestras de vino analizadas son ácidas, ya que todas toman valores entre 2 y 4.
- **Calidad:** ningún vino ha obtenido una calidad máxima y sus valores se encuentran entre 3 y 8.

1.2- Objetivo analítico

El sector vinícola es de gran relevancia en Portugal, no solo desde el punto de vista cultural o social, sino que también económico ya que actualmente es uno de los productores de vino más importante a nivel internacional por su calidad y originalidad. Por ello, el objetivo será crear un modelo matemático que en función de las características físico-químicas del vino prediga el éxito que tendrá entre sus consumidores, además de obtener las características que más influyen en la calidad del vino y así los enólogos podrán producir un vino de máxima calidad.

2- Integración y selección de los datos de interés a analizar

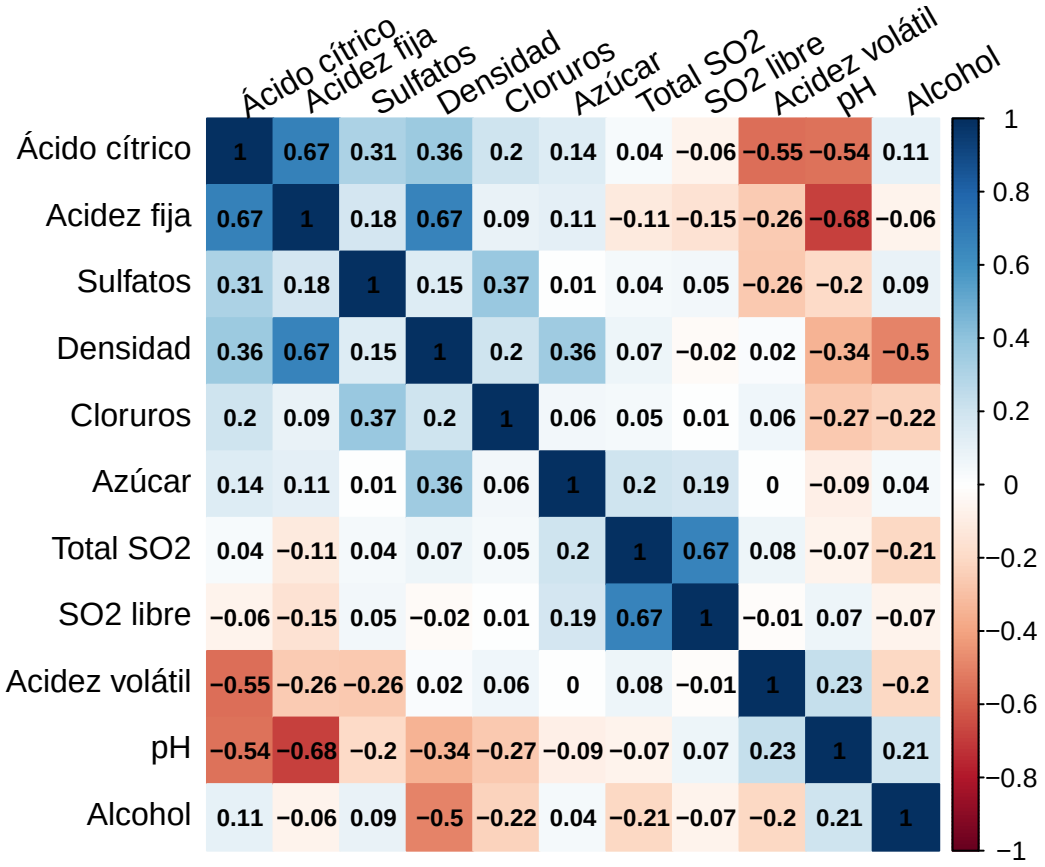
Se analiza la correlación entre las variables predictoras y se visualiza utilizando la función corrplot:

```
# https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
if(!require("corrplot")) install.packages("corrplot"); library("corrplot")
```

```
## Loading required package: corrplot
```

```
## corrplot 0.92 loaded
```

```
rel<-cor(df[,-12])
# Matriz de correlación
corrplot(rel,method="color",tl.col="black", tl.srt=30, order = "AOE",
number.cex=0.75,sig.level = 0.01, addCoef.col = "black")
```



No se observa que haya una correlación positiva o negativa muy alta entre dos variables, en caso contrario, una de las dos variables debería eliminarse ya que darían casi exactamente la misma información en el modelo de regresión y al incluir las dos variables se debilitaría. En cuanto a los resultados obtenidos, la relación positiva más alta se encuentra entre el ácido cítrico y la acidez fija, entre la densidad y acidez fija y entre el SO2 libre y total de SO2. En estos casos, la magnitud de relación es de 0.67. Por otra parte, la relación negativa más alta se encuentra entre la acidez fija y pH y su magnitud es de 0.68.

En este caso, el análisis se hará con todas las características físico- químicas y no se descartará ninguna.

3- Limpieza de los datos

En este apartado se gestionan los errores que pueden tener los datos antes de iniciar el estudio analítico, entre ellos los valores nulos y valores extremos.

3.1- Elementos vacíos o nulos

En primer lugar se procederá a analizar los datos perdidos. Pueden tener distintos formatos, típicamente " " o NA (*Not Available* en inglés):

```
# NA
colSums(is.na(df))
```

```
##    Acidez fija Acidez volátil  Ácido cítrico         Azúcar       Cloruros
##              0              0              0              0              0
##      SO2 libre      Total SO2       Densidad             pH       Sulfatos
##              0              0              0              0              0
##        Alcohol        Calidad
##              0              0
```

```r
# Valores vacíos
colSums(df=="")
```

```
##    Acidez fija Acidez volátil  Ácido cítrico         Azúcar       Cloruros
##              0              0              0              0              0
##      SO2 libre      Total SO2       Densidad             pH       Sulfatos
##              0              0              0              0              0
##        Alcohol        Calidad
##              0              0
```

Se puede observar que no existen valores nulos ni vacíos, por lo que no es necesario realizar ningún tipo de tratamiento.

En caso contrario, existen diversas técnicas para imputar datos perdidos, entre ellas la imputación por la media o mediana, imputación por regresión o mediante el método de kNN (por sus siglas en inglés, *k-Nearest Neighbours*). No siempre es buena práctica aplicar técnicas para reemplazar los valores perdidos, ya que en ocasiones la cantidad de valores perdidos es elevada, por lo que no hay información suficiente sobre la distribución del atributo y la aplicación de técnicas para su imputación nos conduciría a errores y los resultados obtenidos no serían reales.

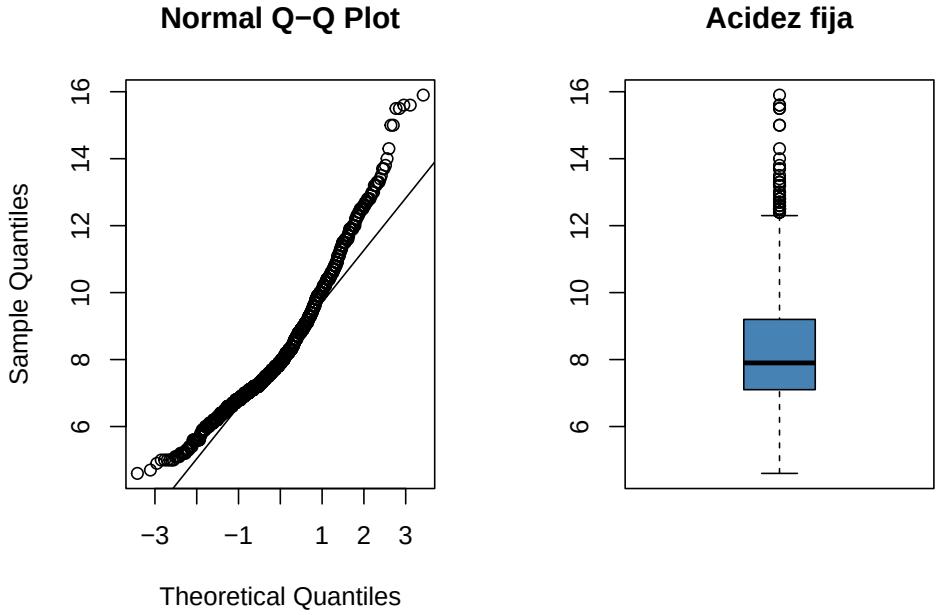
Además, Los valores perdidos pueden tomar valores numéricos como 0 o 999, pero la detección de éstos es más fácil cuando se analiza la distribución de cada atributo.

En el caso de que cada muestra tuviese un código identificador, en este apartado se analizarían los registros duplicados, mediante la función *duplicated* de R.

3.2- Valores extremos

Los valores extremos o *outliers* son datos que se encuentran en los extremos de la distribución normal de una variable o población y pueden influir en los resultados de los análisis ya que incrementan el error en la varianza de los datos y producen estimaciones significativamente sesgadas. Existen diferentes técnicas para la detección de *outliers*, en este caso se utilizarán los gráficos de cajas (*boxplot*):

```r
# Acidez fija
par(mfrow = c(1, 2))
qqnorm(df$`Acidez fija`)
qqline(df$`Acidez fija`)

boxplot(df$"Acidez fija",
        main = "Acidez fija",
        boxwex = 0.5,col="steelblue")
```



En la gráfica Q-Q se observa que los elementos de la muestra no se distribuyen según los cuantiles teóricos de la distribución normal y los puntos situados fuera de la línea representan los valores extremos observados en el *boxplot* de la derecha. El atributo toma valores dentro del intervalo:

```
range(df$`Acidez fija`)
```

```
## [1]  4.6 15.9
```

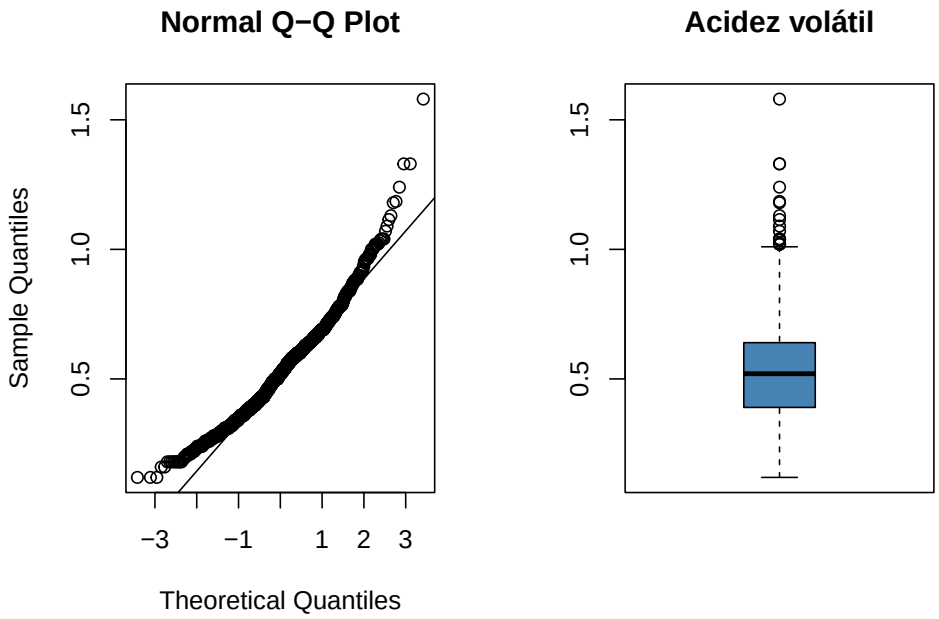
Mediante la función *boxplot.stats$out* se obtienen 49 *outliers*:

```
boxplot.stats(df$"Acidez fija")$out
```

```
##  [1] 12.8 12.8 15.0 15.0 12.5 13.3 13.4 12.4 12.5 13.8 13.5 12.6 12.5 12.8 12.8
## [16] 14.0 13.7 13.7 12.7 12.5 12.8 12.6 15.6 12.5 13.0 12.5 13.3 12.4 12.5 12.9
## [31] 14.3 12.4 15.5 15.5 15.6 13.0 12.7 13.0 12.7 12.4 12.7 13.2 13.2 13.2 15.9
## [46] 13.3 12.9 12.6 12.6
```

El vinho verde portugués es característico por su alta acidez fija ya que le aporta frescura, por lo que no se considerarán valores extremos.

```
# Acidez volátil
par(mfrow = c(1, 2))
qqnorm(df$`Acidez volátil`)
qqline(df$`Acidez volátil`)

boxplot(df$`Acidez volátil`,
        main = "Acidez volátil",
        boxwex = 0.5,col="steelblue")
```



Se observa que la distribución no es normal y que existen presencia de valores extremos. El atributo toma valores entre:

```
range(df$`Acidez volátil`)
```

```
## [1] 0.12 1.58
```

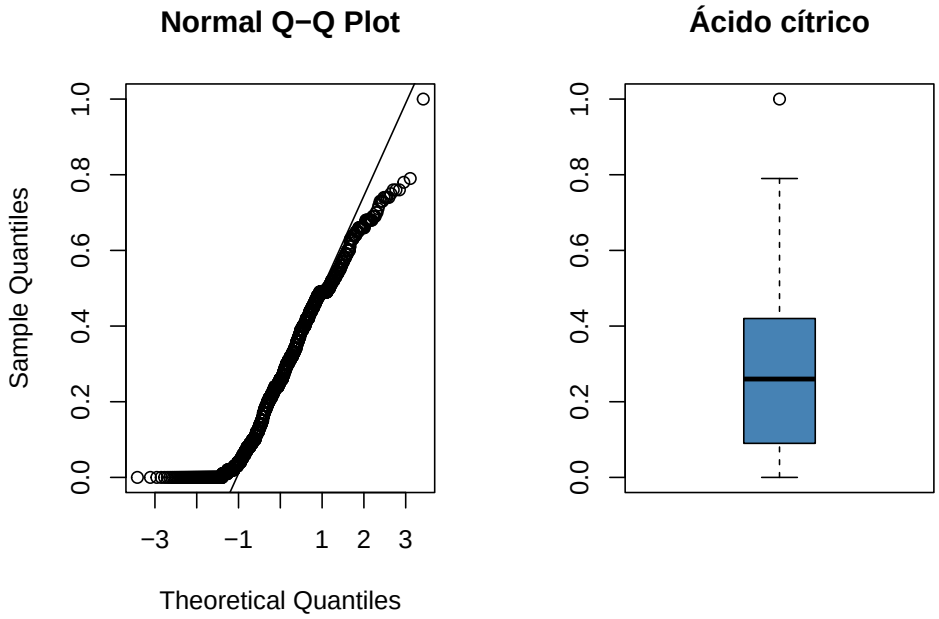
Se registran los siguientes 19 *outliers* para esta variable:

```
boxplot.stats(df$`Acidez volátil`)$out
```

```
##  [1] 1.130 1.020 1.070 1.330 1.330 1.040 1.090 1.040 1.240 1.185 1.020 1.035
## [13] 1.025 1.115 1.020 1.020 1.580 1.180 1.040
```

Normalmente, el acidez volátil estará entre 0,20 y 0,70 según el tipo de vino y proceso de elaboración. El dataset contiene diferentes muestras, por lo que es posible que en alguna este atributo tome valores más altos. Los valores no se alejan notablemente, por lo que no se considerarán valores extremos.

```
# Ácido cítrico
par(mfrow = c(1, 2))
qqnorm(df$`Ácido cítrico`)
qqline(df$`Ácido cítrico`)

boxplot(df$`Ácido cítrico`,
        main = "Ácido cítrico",
        boxwex = 0.5,col="steelblue")
```




```
range(df$`Ácido cítrico`)
```

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

La cantidad máxima legal en vino de ácido cítrico es 1 g/l, por lo que todas las muestras cumplen tal requisito. Mediante el diagrama de cajas se detecta un único valor extremo que toma el valor de uno, pero al encontrarse dentro del rango permitido, no se imputará.

Además, el ácido cítrico es de origen natural y está presente en vinos y uvas en concentraciones entre 0,1-1 g/l. El *dataset* contiene valores nulos e inferiores a 0.1, por lo que se considera que esas muestras tienen valores incorrectos:

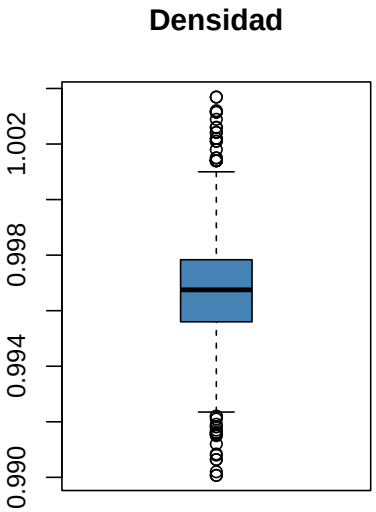
```
length(which(df$`Ácido cítrico`<0.1))
```

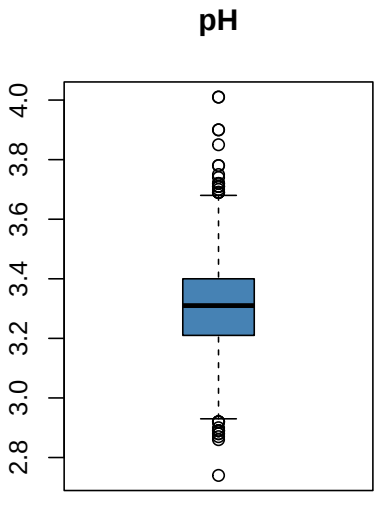
```
## [1] 403
```

El 25% de las muestras son incorrectas y tales valores se imputarán mediante la media, ya que no hay presencia de valores extremos y el resultado se encontrará dentro del rango requerido:

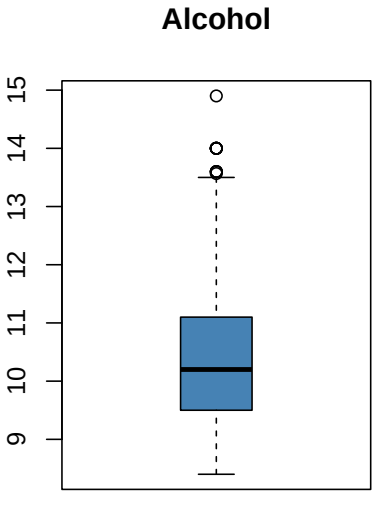
```
# Media
media<-mean(df[-which(df$`Ácido cítrico`<0.1),"Ácido cítrico"])

# Imputación por la media
df[which(df$`Ácido cítrico`<0.1),"Ácido cítrico"]<-media
```

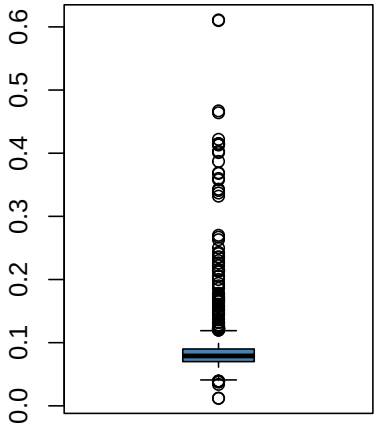
```
par(mfrow = c(1, 2))
# Densidad
boxplot(df$"Densidad",
        main = "Densidad",
        boxwex = 0.5,col="steelblue")
# pH
boxplot(df$pH,
        main = "pH",
        boxwex = 0.5,col="steelblue")
```






```r
par(mfrow = c(1, 2))
# Alcohol
boxplot(df$Alcohol,
        main = "Alcohol",
        boxwex = 0.5,col="steelblue")
# Cloruros
boxplot(df$Cloruros,
        main = "Cloruros",
        boxwex = 0.5,col="steelblue")
```



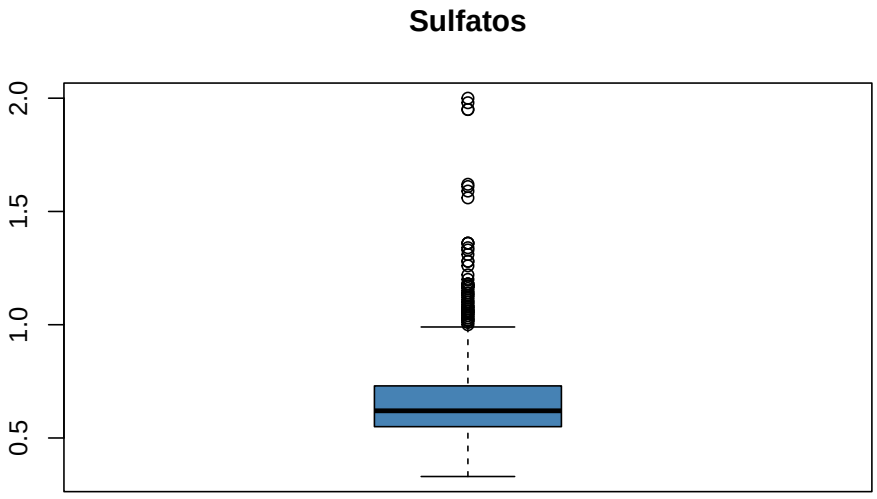


Se observan *outliers* en los cuatro atributos, pero son valores que pueden darse, por lo que no se realizará ninguna modificación. Se ha revisado la documentación sobre los límites máximos permitidos para el cloruro y el máximo aceptado es 1g/L y se observa que las muestras se encuentran dentro de ese intervalo:

```
range(df$Cloruros)
```

```
## [1] 0.012 0.611
```

```
# Sulfatos
boxplot(df$Sulfatos,
        main = "Sulfatos",
        boxwex = 0.5,col="steelblue")
```



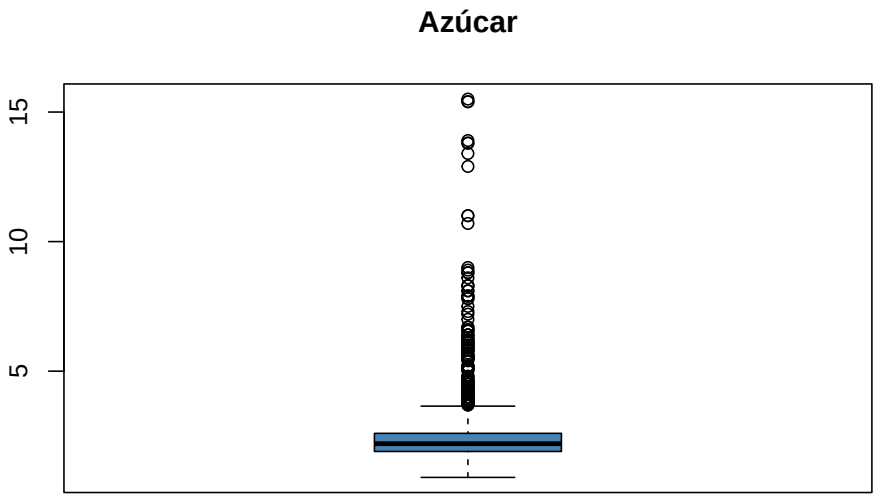

```
boxplot.stats(df$Sulfatos)$out
```

```
##  [1] 1.56 1.28 1.08 1.20 1.12 1.28 1.14 1.95 1.22 1.95 1.98 1.31 2.00 1.08 1.59
## [16] 1.02 1.03 1.61 1.09 1.26 1.08 1.00 1.36 1.18 1.13 1.04 1.11 1.13 1.07 1.06
## [31] 1.06 1.05 1.06 1.04 1.05 1.02 1.14 1.02 1.36 1.36 1.05 1.17 1.62 1.06 1.18
## [46] 1.07 1.34 1.16 1.10 1.15 1.17 1.17 1.33 1.18 1.17 1.03 1.17 1.10 1.01
```

Tras analizar la distribución de los sulfatos, se observan *outliers* donde su valor es superior a 1.5g/L, límite máximos aceptados por la OIV (*Organización Internacional de la Viña y el Vino*) para sulfatos en vinos. En este caso, los valores superiores a 1.5 g/L se imputarán mediante la media:

```
# Media
media<-mean(df[-which(df$Sulfatos>1.5),"Sulfatos"])

# Imputación por la media
df[which(df$Sulfatos>1.5),"Sulfatos"]<-media
```

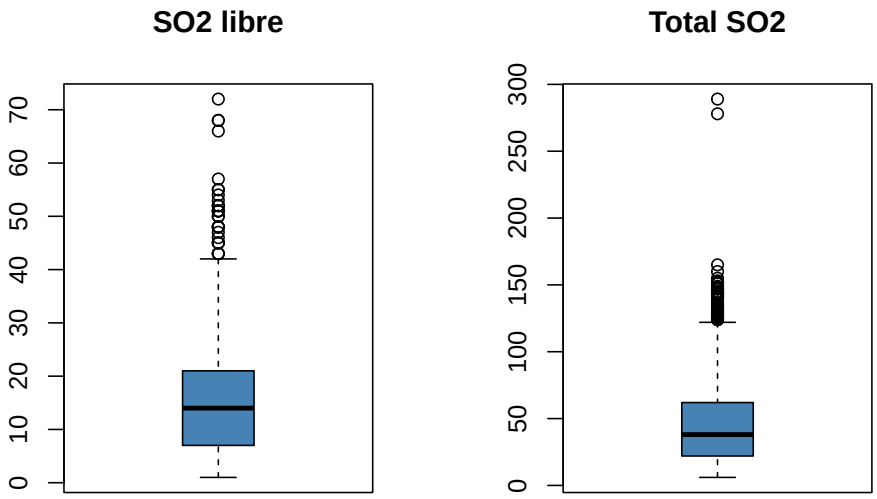
```
# Azúcar
boxplot(df$Azúcar,
        main = "Azúcar",
        boxwex = 0.5,col="steelblue")
```



Tras analizar la distribución de la variable *Azúcar*, los *outliers* que se observan son valores que pueden darse y en función de ellos es posible clasificar los vinos. La clasificación se desarrolla en el siguiente apartado.

A continuación se analizan los *outliers* de las variables *SO2 libre* y *Total SO2* y para su identificación se utilizan también los límites máximos aceptados por la OIV. En este caso son 70mg/L y 200mg/L para el SO2 Libre y Total SO2 respectivamente. Los valores que exceden dichos umbrales se imputarán por la media:

```
par(mfrow = c(1, 2))
# SO2 Libre
boxplot(df$`SO2 libre`,
        main = "SO2 libre",
        boxwex = 0.5,col="steelblue")
# Total SO2
boxplot(df$`Total SO2`,
        main = "Total SO2",
        boxwex = 0.5,col="steelblue")
```




```
# Media SO2 libre y Total SO2
media_libre<-mean(df[-which(df$`SO2 libre`>70),"SO2 libre"])
media_total<-mean(df[-which(df$`Total SO2`>200),"Total SO2"])

# Imputación por la media para ambas
df[which(df$`SO2 libre`>70),"SO2 libre"]<-media_libre
df[which(df$`Total SO2`>200),"Total SO2"]<-media_total
```

4- Análisis de los datos

Después de obtener un *dataset* de calidad, se realiza el análisis de los datos con el fin de resolver el objetivo analítico planteado.

4.1- Selección de grupos de datos y tipo de análisis

En función de la cantidad de azúcar, los vinos se clasifican como:

- *Seco*: entre 0 y 5g/L de azúcar residual.
- *Semiseco*: entre 5g/L y 12g/L de azúcar residual.
- *Semidulce*: entre 12g/L y 23g/L de azúcar residual.
- *Dulce*: azúcar residual superior a 23g/L. Este tipo no está en el dataset.

Se aplica dicha clasificación en nuestro dataset y se obtiene la variable *tipo*:

```
df$tipo <- ifelse(df$Azúcar <= 5, "Seco",
                  ifelse(df$Azúcar>5 & df$Azúcar<=12,"Semiseco","Semidulce"))
```

El análisis de hipótesis se centrará en comparar la calidad sobre los diferentes tipos de vinos en función de la cantidad de azúcar que tienen. Para el análisis se descartarán los vinos semidulces ya que en el conjunto de datos solo hay 8 muestras. Para ello, creamos un dataset sin añadir los vinos semidulces:

```
df_2 <- subset(df, tipo!="Semidulce")
```

Además, se realizará un análisis de correlación para ver la relación de cada característica físico- química y la calidad del vino. Por último, se obtendrá un modelo de regresión logística que en función de sus características físico- químicas nos permitirá saber si tendrá éxito o no entre sus consumidores.

4.2- Comprobación de la normalidad y homogeneidad de la varianza

Normalidad:

En primer lugar se comprobará el supuesto de normalidad mediante el test de normalidad Lilliefors, ya que la muestra es superior a 50 observaciones. La hipótesis nula asume que la variable proviene de una población con distribución normal y la hipótesis alternativa asume que la variable proviene de una población con una distribución diferente a la normal:

```
# Es necesario el paquete: nortest
library(nortest)

col<-colnames(df[1:12])
for (i in col){
  pvalue = lillie.test(df[,i])$p.value
  if (pvalue < 0.05) cat("p-value:", pvalue, "< 0.05 ->",i, "no cumple el supuesto de normalidad.\n")
}
```

```
## p-value: 6.982456e-53 < 0.05 -> Acidez fija no cumple el supuesto de normalidad.
## p-value: 4.489084e-12 < 0.05 -> Acidez volátil no cumple el supuesto de normalidad.
## p-value: 2.35934e-105 < 0.05 -> Ácido cítrico no cumple el supuesto de normalidad.
## p-value: 3.981712e-309 < 0.05 -> Azúcar no cumple el supuesto de normalidad.
## p-value: 1.260107e-306 < 0.05 -> Cloruros no cumple el supuesto de normalidad.
## p-value: 7.536048e-53 < 0.05 -> SO2 libre no cumple el supuesto de normalidad.
## p-value: 4.786681e-60 < 0.05 -> Total SO2 no cumple el supuesto de normalidad.
## p-value: 6.251707e-08 < 0.05 -> Densidad no cumple el supuesto de normalidad.
## p-value: 2.244048e-06 < 0.05 -> pH no cumple el supuesto de normalidad.
## p-value: 1.76401e-56 < 0.05 -> Sulfatos no cumple el supuesto de normalidad.
## p-value: 2.391501e-64 < 0.05 -> Alcohol no cumple el supuesto de normalidad.
## p-value: 1.951455e-283 < 0.05 -> Calidad no cumple el supuesto de normalidad.
```

En todos casos se rechaza la hipótesis nula con un nivel de confianza de 95% ya que el valor-p < 0.05. Por lo tanto, ninguna variable proviene de una población que tenga una distribución normal.

Para la aplicación del contraste de hipótesis, se estudia el supuesto de la normalidad para la calidad en función del tipo de vino. Para ello, se utiliza el test de normalidad Lilliefors, ya que las muestras son superiores a 50 observaciones:

```
# Calidad en función del tipo de vino
lillie.test(df$Calidad[df$tipo=="Semiseco"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df$Calidad[df$tipo == "Semiseco"]
## D = 0.26445, p-value = 1.906e-14
```

```
lillie.test(df$Calidad[df$tipo=="Seco"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df$Calidad[df$tipo == "Seco"]
## D = 0.24968, p-value < 2.2e-16
```

En ambos casos se rechaza la hipótesis nula con un nivel de confianza de 95% ya que el valor-p < 0.05.

Homogeneidad de la varianza:

Se estudia si existe igualdad en las varianzas entre la calidad de los vinos secos y la calidad de los vinos semisecos mediante el test de Fligner-Killeen. Es el test de homocedasticidad no paramétrico donde la hipótesis nula acepta igualdad en las varianzas y la hipótesis alternativa no:

```
fligner.test(Calidad ~ tipo, data=df_2)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Calidad by tipo
## Fligner-Killeen:med chi-squared = 5.588, df = 1, p-value = 0.01808
```

Se rechaza la hipótesis nula ya que p-valor < 0.05. Por lo tanto, se concluye que la variable calidad presenta varianzas estadísticamente diferentes para los diferentes tipos de vinos en función de la cantidad de azúcar residual.

4.3- Aplicación de pruebas estadísticas

4.3.1- Contraste de hipótesis

¿Tienen la misma calidad los vinos secos y semisecos?

Para ello, se plantea la siguiente hipótesis:

- H0: $\mu_s = \mu_{ss}$
- H1: $\mu_s \neq \mu_{ss}$

donde μ_s es el valor medio de la calidad de los vinos secos y μ_{ss} es el valor medio de la calidad de los vinos semisecos.

Para poder aplicar pruebas por contraste de hipótesis de tipo paramétrico, se debe comprobar el supuesto de normalidad para la variable 'Calidad'. Se ha comprobado que no tiene distribución normal, pero aplicando el teorema del límite central, la distribución de las medias muestrales de muestras suficientemente grandes (n>30) es aproximadamente normal por lo que se puede asumir normalidad.

En este caso la varianza de la población es desconocida y se ha visto que no se puede asumir igualdad en la varianza de ambos tipos de vinos. Por lo tanto, utilizaremos el test t-student de dos muestras independientes sobre la media con varianza desconocida y diferente:

```
t.test(df$Calidad[df$tipo=="Semiseco"],df$Calidad[df$tipo=="Seco"],
alternative="two.sided",var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$Calidad[df$tipo == "Semiseco"] and df$Calidad[df$tipo == "Seco"]
## t = 1.0339, df = 80.054, p-value = 0.3043
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1094186  0.3460522
## sample estimates:
## mean of x mean of y
##  5.750000  5.631683
```

El valor p > 0.05, por lo que se acepta la hipótesis nula y se puede asumir que la media de la calidad de los vinos secos y semisecos es estadísticamente igual.

4.3.2- Análisis de correlación

En primer lugar, se analizará la correlación entre las variables físico- químicas y la variable *Calidad* para determinar cuáles de ellas ejercen una mayor influencia sobre la calidad del vino.

Para ello, se utilizará el coeficiente de correlación de Spearman, puesto que las variables no cumplen con el supuesto de normalidad y este método no conlleva ninguna suposición sobre la distribución de los datos:

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")

# Calcular el coeficiente de correlación para cada variable cuantitativa
# con respecto al campo "calidad"
for (i in 1:(ncol(df[-12]))) {
      spearman_test = cor.test(df[,i],
      df[,"Calidad"],
      method = "spearman")
      corr_coef = spearman_test$estimate
      p_val = spearman_test$p.value
      # Add row to matrix
      pair = matrix(ncol = 2, nrow = 1)
      pair[1][1] = corr_coef
      pair[2][1] = p_val
      corr_matrix <- rbind(corr_matrix, pair)
      rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(df)[i]
}

print(corr_matrix)
```

```
##                   estimate      p-value
## Acidez fija     0.11408367 4.801220e-06
## Acidez volátil -0.38064651 2.734944e-56
## Ácido cítrico   0.20277375 2.672179e-16
## Azúcar          0.03204817 2.002454e-01
## Cloruros       -0.18992234 1.882858e-14
## SO2 libre      -0.05751547 2.144792e-02
## Total SO2      -0.19983854 7.246668e-16
## Densidad       -0.17707407 9.918139e-13
## pH             -0.04367193 8.084594e-02
## Sulfatos        0.38302273 4.983890e-57
## Alcohol         0.47853169 2.726838e-92
## Calidad         1.00000000 0.000000e+00
```

En todos casos, el p-valor es significativo. El coeficiente de Spearman evalúa la relación monótona entre dos variables, es decir, analiza si las variables cambian al mismo tiempo, pero no necesariamente a un ritmo constante.

Se observa que no hay variables con una relación monótona fuerte. Por una parte, la variable *Alcohol* es la que mayor relación positiva presenta respecto a la calidad del vino, es decir, cuando el porcentaje de alcohol aumenta, también incrementa la calidad del vino. Por otra parte, es la acidez volátil la que mayor relación negativa presenta respecto a la calidad del vino, es decir, a medida que la cantidad de acidez volátil disminuye, incrementa la calidad del vino.

4.3.2 Modelo de Regresión Logística

Por último, se pretende predecir si el vino producido tendrá o no éxito entre sus consumidores en función de sus características físico químicas. Para ello, se utilizará un modelo de regresión logística cuya variable

dependiente dicotómica será calidad sí/calidad no y sus variables independientes serán las características físico- químicas. En este caso, se trata de un tipo de modelo supervisado que se estima a partir de un conjunto de datos de entrenamiento y posteriormente es validado por un conjunto de datos de test.

En primer lugar, se recodifica la variable *Calidad* y se obtiene la variable dicotómica *calidad_rec*, donde 0 indica calidad no y 1 calidad si:

```
# Recodificación variable calidad
df$calidad_rec <- ifelse(df$Calidad<=5, 0,1)
```

Para nuestro estudio consideraremos que **calidad si** serán aquellos valores de la variable **calidad** que se encuenetren en el rango de 6 a 10, y **calidad no** serán aquellos valores que se encuentren en el rango de 0 a 5.

Se verifica que la variable dependiente es balanceada:, de esa manera los resultados obtenidos no estarán sesgados:

```
prop.table(table(df$calidad_rec))
```

```
##
##         0         1
## 0.4652908 0.5347092
```

El número de muestras para cada grupo de la variable *calidad_rec* es similar, por lo que estaría balanceado y de esa manera los resultados obtenidos no estarán sesgados.

Se genera los conjuntos de datos para entrenar el modelo y para testarlo. Para ello, el tamaño muestral para el conjunto de entrenamiento será el 80% del original:

```
#Partición de los datos
set.seed(123) # fijamos semilla
size<-nrow(df)*0.8 # tamaño muestra entrenamiento (80% original)
position<-sample(1:nrow(df),size)
train<-df[position, ]
test<-df[-position, ]
```

Se comprueba que los conjuntos obtenidos tienen una proporción similar de muestras que el conjunto de datos principal:

```
#Comprobar que están igual distribuídos
prop.table(table(train$calidad_rec))
```

```
##
##         0         1
## 0.4566067 0.5433933
```

```
prop.table(table(test$calidad_rec))
```

```
##
##   0   1
## 0.5 0.5
```

Se observa que ambos subconjuntos tienen similar distribución que en el dataset inicial.

Una vez obtenido los conjuntos de datos de entrenamiento y test, se estima el modelo de regresión lineal:

```
Model_logit_1<-glm(calidad_rec ~ `Acidez fija`+`Acidez volátil`+`Ácido cítrico`+ Azúcar +
                     Cloruros + `SO2 libre`+`Total SO2`+ Densidad + pH + Sulfatos +
                     Alcohol, data=train,family=binomial)
summary(Model_logit_1)
```

```
##
```

```
## Call:
## glm(formula = calidad_rec ~ `Acidez fija` + `Acidez volátil` +
##     `Ácido cítrico` + Azúcar + Cloruros + `SO2 libre` + `Total SO2` +
##     Densidad + pH + Sulfatos + Alcohol, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2529  -0.8577   0.3251   0.8098   2.2328
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       43.113015  89.336346   0.483  0.62939
## `Acidez fija`      0.080361   0.109135   0.736  0.46152
## `Acidez volátil`  -2.403795   0.454126  -5.293 1.20e-07 ***
## `Ácido cítrico`   -0.616574   0.641245  -0.962  0.33629
## Azúcar            -0.014284   0.063287  -0.226  0.82143
## Cloruros          -4.642152   1.690691  -2.746  0.00604 **
## `SO2 libre`        0.025291   0.009462   2.673  0.00752 **
## `Total SO2`       -0.019953   0.003340  -5.974 2.31e-09 ***
## Densidad         -49.115553  91.198427  -0.539  0.59019
## pH                -0.928678   0.802860  -1.157  0.24739
## Sulfatos           3.302631   0.568211   5.812 6.16e-09 ***
## Alcohol            0.843287   0.116153   7.260 3.87e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1763.4  on 1278  degrees of freedom
## Residual deviance: 1312.3  on 1267  degrees of freedom
## AIC: 1336.3
##
## Number of Fisher Scoring iterations: 4
```

Con un nivel de confianza de 95% se afirma que las variables *acidez volátil*, *cloruros*, *SO2 libre*, *total SO2*, *Sulfatos* y *Alcohol* son significativas (valor- p <0.05) respecto a la probabilidad de que el vino tenga calidad (calidad si). Se analiza por separado las variables significativas en función del coeficiente obtenido:

- *Acidez volátil*, *cloruros* y *total SO2*: el coeficiente de estas variables es negativo, por lo que a medida que aumenta la cantidad de estas variables, disminuye la probabilidad de tener un vino que tenga calidad.
- *SO2 libre*, *Sulfatos* y *Alcohol*: el coeficiente de estas variables es positivo, por lo que a medida que aumenta la cantidad de estas variables, aumenta la probabilidad de tener un vino que tenga calidad.

Se obtiene que AIC (criterio de información de Akaike) es 1336.3. Es la medida que se utiliza para evaluar la bondad del ajuste y se pretende conseguir un valor bajo de ésta.

Matriz de confusión

A continuación, se analiza la precisión del modelo, comparando la predicción del modelo contra el conjunto de prueba (test). Se asume que la predicción del modelo es 1 (calidad si) si la probabilidad del modelo de regresión logística es superior o igual a 0.5 y 0 de lo contrario:

```
respuesta <- ifelse(predict.glm(Model_logit_1, newdata=test, type="response")>=0.5,1,0)
observado <-ifelse(test$calidad_rec==1,1,0)
# Matriz de confusión
library(caret)
```

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

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

```
confusionMatrix(table(respuesta,observado),positive='1')
```

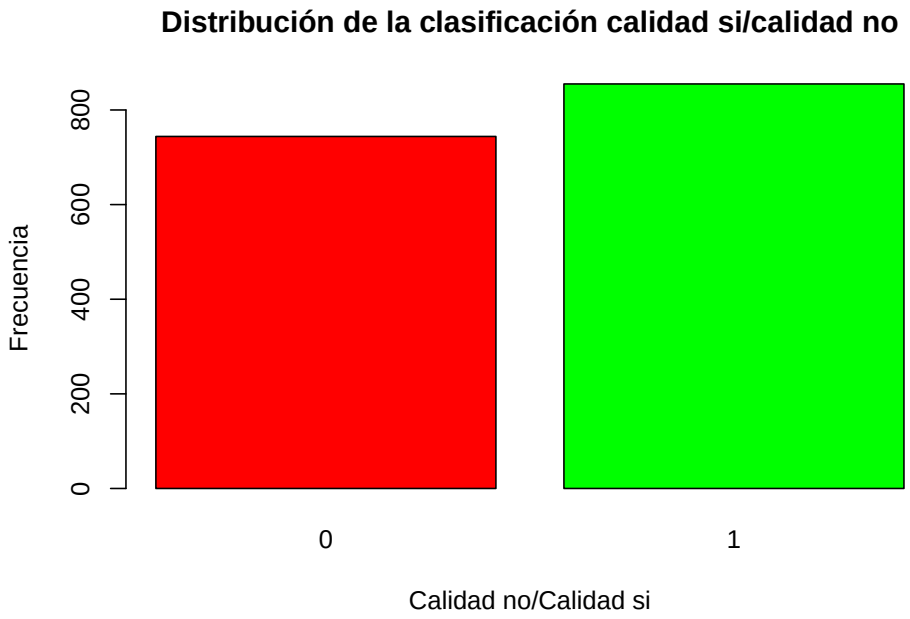
```
## Confusion Matrix and Statistics
##
##          observado
## respuesta   0   1
##         0 122  46
##         1  38 114
##
##                Accuracy : 0.7375
##                  95% CI : (0.6857, 0.7849)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.475
##
##  Mcnemar's Test P-Value : 0.445
##
##             Sensitivity : 0.7125
##             Specificity : 0.7625
##          Pos Pred Value : 0.7500
##          Neg Pred Value : 0.7262
##              Prevalence : 0.5000
##          Detection Rate : 0.3563
##    Detection Prevalence : 0.4750
##       Balanced Accuracy : 0.7375
##
##        'Positive' Class : 1
##
```

La sensibilidad del modelo ajustado es de 0.7125, por lo que el modelo predice correctamente el 71.25% de las muestras que tienen buena calidad. Ademas, la especificidad es de 0.7625 y así el modelo predice correctamente el 76.25% de las muestras que su calidad es mala. En este caso, una mala práctica sería clasificar un vino con calidad mala como calidad buena, ya que no tendría éxito entre los consumidores y económicamente no sería rentable sacarlo a la venta a sus productores. Para este caso hay 38 muestras con baja calidad clasificadas incorrectamente, es decir, el porcentaje de falsos positivos es de 11.87%.

5- Representación resultados

Por una parte, la precisión del modelo obtenido se ha representado en el apartado anterior mediante su matriz de confusión. El barplot de la distribución de la variable dicotómica *calidad_rec* es el siguiente:

```
barplot(table(df$calidad_rec), main="Distribución de la clasificación calidad si/calidad no",ylab="Frecu
```



Los demás resultados se han desarrollado en su respectivo apartado.

6- Conclusiones

Para obtener el dataset final utilizado en el análisis estadístico se realizó un preprocesamiento aplicando diferentes técnicas como: análisis de valores nulos y vacíos, análisis de valores extremos (outliers), análisis de normalidad y homogeneidad, discretización y normalización. Se han utilizado los límites máximos aceptados por la OIV (*Organización Internacional de la Viña y el Vino*) como criterio para descartar valores en el análisis de *outliers*.

En el análisis de los datos, se ha hecho una clasificación en función de la cantidad de azúcar y se han obtenido dos grupos: vinos secos y semisecos. Mediante un análisis de hipótesis se ha obtenido que estadísticamente no existen diferencias en la calidad de los diferentes tipos de vinos.

Mediante el estudio de correlación entre la variable calidad y las características físico-químicas se ha analizado qué variables influyen más en la calidad del vino. Para ello, la relación se ha cuantificado mediante el coeficiente de spearman y se puede concluír que la variable *Alcohol* es la que mayor relación positiva presenta respecto a la calidad del vino y la acidez volátil la que mayor relación negativa. Por ello, los productores de vino deberían añadir bajas cantidades de acidez volátil y aumentar el porcentaje de alcohol.

Por último, con el fin de poder predecir el éxito (calidad si/calidad no) que tendrá entre los consumidores un vino en función de sus variables físico- químicas, se ha creado un modelo de regresión logística que determina el índice de importancia que cada una de las características representa sobre la calidad del vino. Se concluye que las variables Acidez fija, Ácido cítrico, azúcar, densidad y ph no son estadísticamenteson significativas en la buena calidad del vino, por lo que los productores no deberían darle relevancia. Sin embargo, las variables acidez volátil, cloruros, SO2 libre, total SO2, Sulfato* y Alcohol son estadísticamente significativas. Entre ellas, a medida que aumenta las cantidades de *Acidez volátil*, *cloruros* y *total SO2*, disminuye la probabilidad de tener un vino que tenga calidad, por lo que los productores deberían añadir bajas cantidades de estas. Por otro lado, cantidades altas de *SO2 libre*, *Sulfatos* y *Alcohol* aumentan la probabilidad de tener un vino que tenga calidad y por ello se debe aumentar la presencia de éstas en el diseño de los vinos.

7-Código:

El código de R se encuentra en github en la siguiente dirección:

https://github.com/FidelRZ/PRA2_Limpieza_Validacion_Vinos/tree/main/Codigo

El dataframe inicial y final se encuentran en github en la siguiente dirección:

https://github.com/FidelRZ/PRA2_Limpieza_Validacion_Vinos/tree/main/DataSet

Contribuciones	Firma
Investigación previa	A.R.R / F.R.Z
Redación de las resp.	A.R.R / F.R.Z
Desarrollo código	A.R.R / F.R.Z