# Práctica 2: Limpieza y análisis de datos

*Luis Martin*

1 de June, 2021

## Contents

# 1 Detalles de la actividad

## 1.1 Descripción

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

## 1.2 Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

## 1.3 Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

# 2 Resolución

## 2.1 Descripción del dataset

El conjunto de datos a analizar consiste en un set de atributos fisicoquímicos de muestras de vino rojo de variantes del vino portugués *"Vinho Verde"*. A cada una de estas muestras se le atribuye una puntuación en función de su calidad, que va de 0 a 10. El dataset se puede descargar desde este enlace en *Kaggle*, está formado por 11 atributos mas la puntuación asociada a la calidad, y por un total de 1599 muestras o registros.

A continuación se detallan las diferentes columnas del dataset:

- **fixed.acidity** (Acidez fija): ácidos involucrados con el vino o fijos o no volátiles (no se evaporan fácilmente). Principalmente ácido tartárico. Expresada en $g/dm^3$.
- **volatile.acidity** (Acidez volátil): ácidos volátiles presentes en el vino, principalmente ácido acético. A niveles altos puede provocar un sabor avinagrado en el vino. Expresada en $g/dm^3$.
- **citric.acid** (Ácido cítrico): Se usa en pequeñas cantidades para agregar frescura y sabor al vino. Expresado en $g/dm^3$.
- **residual.sugar** (Azúcar residual): Cantidad de azúcar residual una vez termina la fermentación. Expresado en $g/dm^3$.
- **chlorides** (Cloruros): Cantidad de cloruro sódico (sal) en el vino. Expresado en $g/dm^3$.
- **free.sulfur.dioxide** (Dióxido de azufre libre): Fracción de dióxido de azufre libre. Previene la proliferación de bacterias y la oxidación del vino. Expresado en $mg/dm^3$.
- **total.sulfur.dioxide** (Dióxido de azufre total): En concentraciones altas resalta la olor y el sabor del vino. Expresado en $mg/dm^3$.
- **density** (Densidad): Depende del porcentaje de alcohol y la concentración de azúcar. Expresada en $g/cm^3$.
- **pH** (pH): Describe cuán ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico).
- **sulphates** (Sulfatos): Aditivo que contribuye a los niveles de dióxido de azufre. Expresado en $g/dm^3$.
- **alcohol** (Alcohol): Porcentaje de alcohol en el vino.
- **quality** (Calidad): Calidad de la muestra en una escala de 0 a 10, basada en la percepción sensorial.

## 2.2 Importancia y objetivos de los análisis

Mediante este estudio se pretende averiguar cuáles son las principales características que influyen en la calidad de un vino rojo de la familia de vinos *"Vinho Verde"*. Por otro lado, elaborar y ajustar modelos de predicción que permitan determinar si un vino es de calidad a partir de sus propiedades fisicoquímicas, así como extraer características interesantes del dataset extrapolables al resto de la población. Los resultados de un estudio de este estilo pueden ser interesantes de cara a mejorar los procesos de elaboración de estos vinos para producir vinos de mejor calidad.

## 2.3 Limpieza de datos

Empezamos cargando el dataset y comprobando las dimensiones y la estructura de los datos. Además, podemos realizar un primer vistazo a los datos mediante un pequeño resumen estadístico.

```
# Cargar los datos
df.wines <- read.csv("winequality-red.csv")
```

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

```
## '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 ...
```

```
# Resumen estadístico
summary(df.wines)
```

```
##  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    density
##  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          sulphates         alcohol         quality
##  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
```

### 2.3.1 Selección de los datos de interés

Dentro de nuestro conjunto de datos separamos los atributos en dos categorías, los atributos de entrada o explicativos, asociados a las 11 características fisicoquímicas de las muestras de vino, y un atributo de salida o dependiente, asociado a la variable que se pretende explicar, es decir, la calidad del vino. En un principio, todas las características del vino son susceptibles a provocar un impacto en la calidad, por lo tanto nos interesa conservar todos los atributos para estudiar y extraer información del conjunto de datos. En el momento de elaborar el modelo predictivo es posible que renunciemos a alguno de estos atributos debido a la colinealidad entre diferentes variables, pero esto es algo que dejaremos para mas adelante, cuando estudiemos las correlaciones.

De cara a los análisis posteriores, nos interesaría incluir una nueva variable dicotómica que nos permita clasificar los vinos por calidad de una manera más sencilla, dividiéndolos entre buenos y malos. De este modo dividimos el conjunto de datos en dos grandes subgrupos y damos pié a un posible modelo de regresión logística.

```
# Variable calidad binaria
df.wines$bin_quality <- with(df.wines, ifelse(quality <= 6, 0, 1))
```

### 2.3.2 Ceros y elementos vacíos

En el resumen estadístico, donde podemos observar el rango de valores que toma cada variable y la cantidad de valores nulos, vemos que todos los atributos caen dentro del dominio esperado y que no existen valores centinela para indicar la presencia de información perdida. El único atributo que contiene ceros en su rango es el ácido cítrico, pero se trata de un valor completamente plausible en cuanto a la composición de un vino. Por otro lado, tampoco detectamos la presencia de elementos vacíos, al parecer se trata de una dataset ya limpio y preparado para el análisis.

En el caso hipotético de que hubiésemos detectado información perdida se debería haber reemplazado estos datos con medidas de tendencia central o imputar los valores implementando métodos de predicción como *K-Nearest Neighbours* o *missForest*.

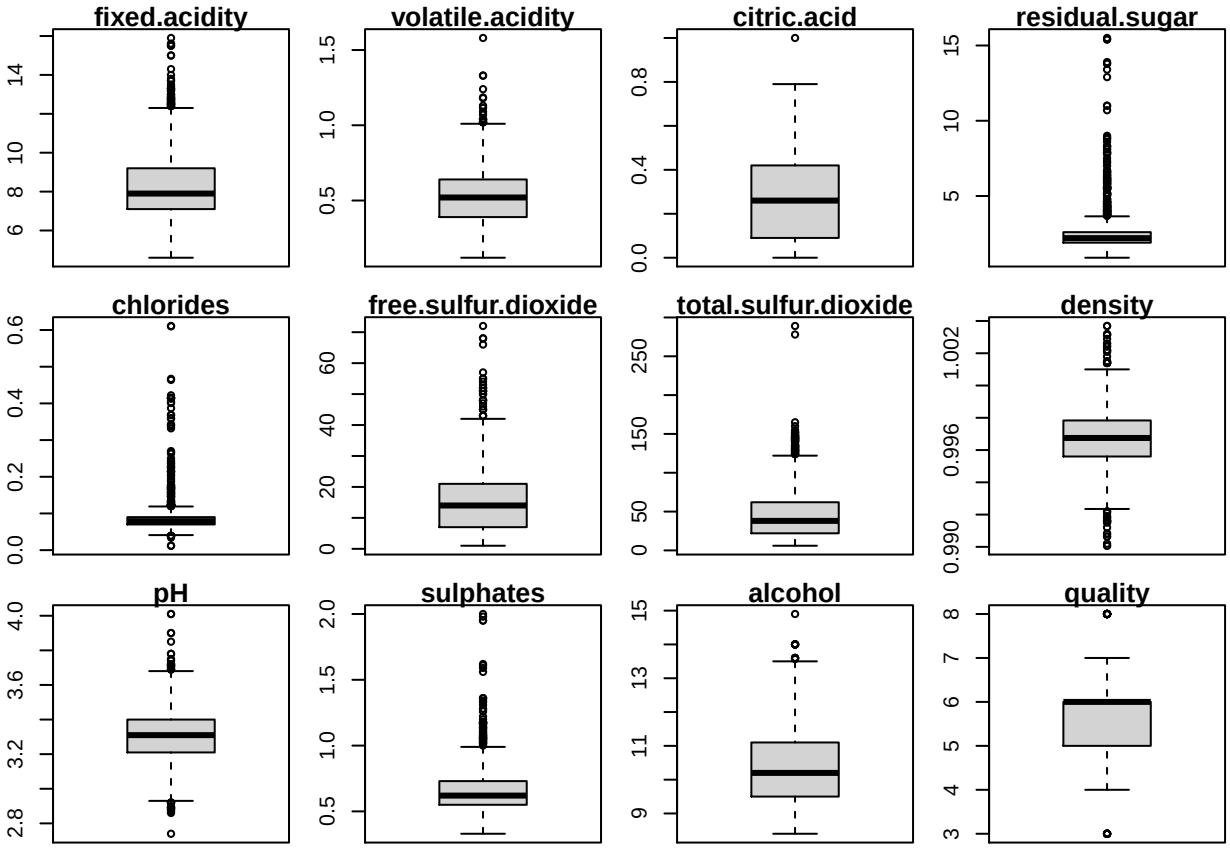
```
# Números de valores desconocidos por campo
sapply(df.wines, function(x) sum(is.na(x)))
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0
##          bin_quality
##                    0
```

### 2.3.3 Valores extremos

Se considera *outliers* todos aquellos valores que se encuentran muy alejados de la media de la población. En función de su origen deben tratarse de un modo u otro. En ocasiones son debido a una desviación sistemática en la obtención de los datos y es fácilmente remediable mediante una operación sencilla. En otras ocasiones son medidas completamente legítimas de valores atípicos de la población, en ese caso deben contemplarse dentro del análisis. Los diagramas de caja a continuación muestran la presencia de una gran cantidad de *outliers* en los diferentes atributos, aún así, se trata de valores legítimos que entran dentro del rango de posibles valores que podrían tomar estas características. Deben dejarse como están y tenerse en consideración en el posterior análisis.

```
# BoxPlots
par(mfrow=c(3,4), mar=c(1,2,1,1))
boxplot(df.wines$fixed.acidity, main="fixed.acidity")
boxplot(df.wines$volatile.acidity, main="volatile.acidity")
boxplot(df.wines$citric.acid, main="citric.acid")
boxplot(df.wines$residual.sugar, main="residual.sugar")
boxplot(df.wines$chlorides, main="chlorides")
boxplot(df.wines$free.sulfur.dioxide, main="free.sulfur.dioxide")
boxplot(df.wines$total.sulfur.dioxide, main="total.sulfur.dioxide")
boxplot(df.wines$density, main="density")
boxplot(df.wines$pH, main="pH")
boxplot(df.wines$sulphates, main="sulphates")
boxplot(df.wines$alcohol, main="alcohol")
boxplot(df.wines$quality, main="quality")
```



### 2.3.4 Exportación de los datos

Tras validar y limpiar los datos guardamos el dataset con la nueva columna que etiqueta los vinos en buenos y malos en función de su calidad.

```
# Guardar los datos
write.csv(df.wines, "winequality-red-clean.csv")
```

## 2.4 Análisis de los datos

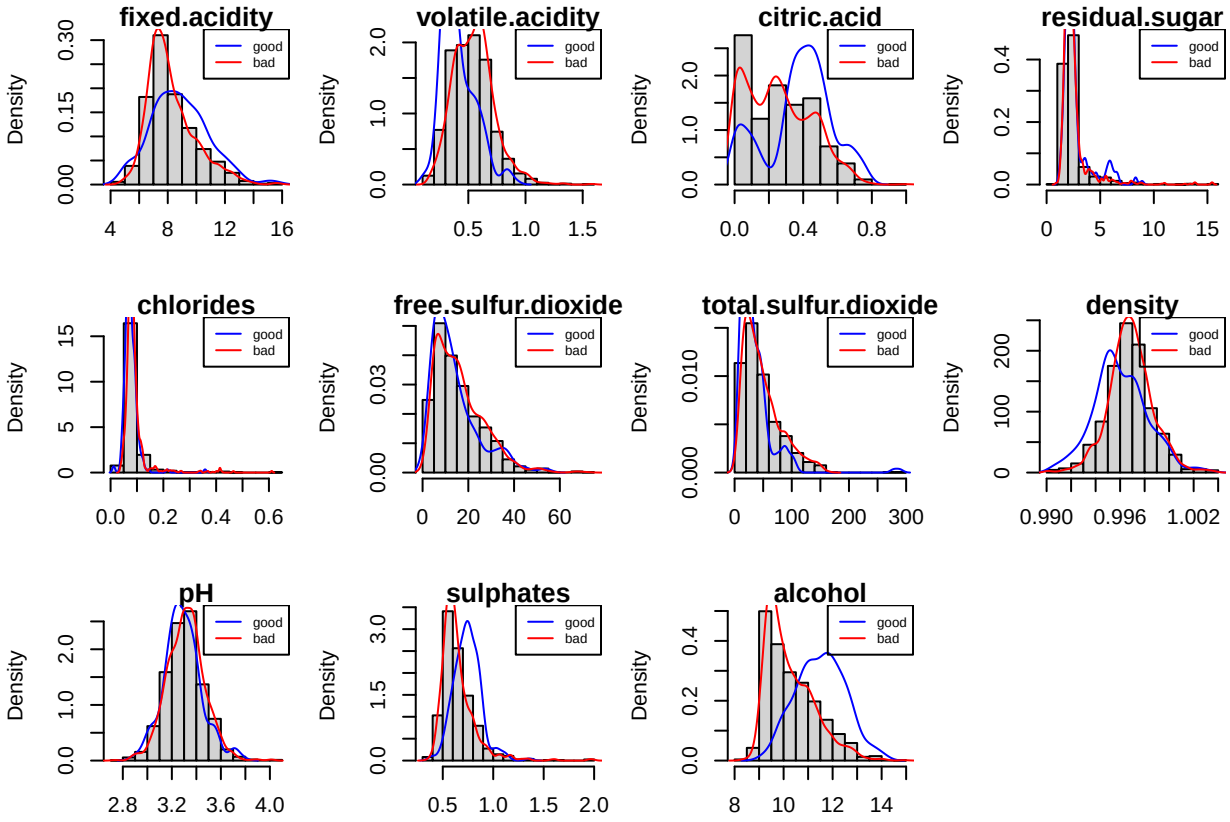
### 2.4.1 Selección de los grupos de datos a analizar

Dividimos las muestras en dos subgrupos utilizando el atributo *bin_quality* para poder comparar las propiedades fisicoquímicas de los vinos de buena calidad con los de mala calidad.

```
# Agrupación por calidad
df.wines.good <- df.wines[df.wines$bin_quality==1,]
df.wines.bad <- df.wines[df.wines$bin_quality==0,]
```

### 2.4.2 Distribución de las variables

Realizamos los histogramas de las variables con las distribuciones de densidad de las dos poblaciones para un primer acercamiento a la distribución de los datos.



### 2.4.3 Comprobación de la normalidad

Utilizaremos el test de *Shapiro-Wilk* con un nivel de significancia $\alpha = 0.05$ para comprobar la normalidad de los atributos fisicoquímicos del vino. En esta prueba la hipótesis nula asume normalidad, por tanto, un valor p menor al nivel de significancia rechaza la normalidad en los datos. El principal objetivo de esta sección es conocer la distribución de valores de cada atributo para determinar qué tests estadísticos será conveniente utilizar durante el análisis. Estudiar cada población por separado será más efectivo de cara a escoger los tests para los contrastes de hipótesis.

```
# Shapiro-Wilk test good wines
alpha <- 0.05
col.names <- colnames(df.wines.good)

#Variables que siguen una distribución normal
normal = c()
for (i in 1:(ncol(df.wines.good)-2)) {
  p_val <- shapiro.test(df.wines.good[,i])$p.value
  if (p_val > alpha) {
    normal <- c(normal, col.names[i])
  }
}

print(normal)
```

```
## [1] "density" "alcohol"
```

```
# Shapiro-Wilk test bad wines
alpha <- 0.05
col.names <- colnames(df.wines.bad)

#Variables que siguen una distribución normal
normal = c()
for (i in 1:(ncol(df.wines.bad)-2)) {
  p_val <- shapiro.test(df.wines.bad[,i])$p.value
  if (p_val > alpha) {
    normal <- c(normal, col.names[i])
  }
}

print(normal)
```

```
## NULL
```

En general las variables no siguen una distribución normal. Aún así, debemos destacar que ambas poblaciones están formadas por centenares de muestras, y podemos asumir gracias al **Teorema del Limite Central** que las medias muestrales sí que siguen una distribución normal.

### 2.4.4 Comprobación de la homogeneidad de la varianza

Dado que los datos que queremos comparar no siguen una distribución normal aplicaremos el test *Fligner-Killeen* para comprobar si existe homocedasticidad entre los atributos de ambas poblaciones. En este test la hipótesis nula asume igualdad de varianzas, por lo tanto, un valor p menor a $\alpha = 0.05$ concluye que la diferencia en las varianzas es estadísticamente significativa.

```
# Fligner-Killeen test
alpha <- 0.05
col.names <- colnames(df.wines)

#Variables con igualdad de varianzas
eq.varianzas = c()
for (i in 1:(ncol(df.wines)-2)) {
  p_val <- fligner.test(df.wines[,i] ~ df.wines$bin_quality)$p.value
  if (p_val > alpha) {
    eq.varianzas <- c(eq.varianzas, col.names[i])
  }
}

print(eq.varianzas)
```

```
## [1] "citric.acid"    "residual.sugar" "chlorides"      "pH"
## [5] "sulphates"      "alcohol"
```

El ácido cítrico, el azúcar residual, los cloruros, el pH, los sulfatos y la concentración de alcohol comparten varianza entre ambas poblaciones.

## 2.5 Pruebas estadísticas

### 2.5.1 Contraste de hipótesis

Queremos comprobar si la graduación, la concentración de sulfatos y la concentración de cítricos es mayor en los vinos de buena calidad. Para ello disponemos de dos poblaciones independientes de tamaño suficiente para aplicar el **Teorema del Limite Central** y con homocedasticidad. El test estadístico correspondiente es un T-student unilateral y la formulación de las hipótesis es:

$$H_0 : \mu_{bueno} = \mu_{malo}$$

$$H_1 : \mu_{bueno} > \mu_{malo}$$

```
# Contraste de hipótesis para el porcentaje de alcohol
 t.test(df.wines.good$alcohol, df.wines.bad$alcohol,
        alternative = "greater", var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  df.wines.good$alcohol and df.wines.bad$alcohol
## t = 17.823, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.150012      Inf
## sample estimates:
## mean of x mean of y
##  11.51805  10.25104
```

```r
# Contraste de hipótesis para la concentración de sulfatos
 t.test(df.wines.good$sulphates, df.wines.bad$sulphates,
        alternative = "greater", var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  df.wines.good$sulphates and df.wines.bad$sulphates
## t = 8.1354, df = 1597, p-value = 4.081e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.07873467        Inf
## sample estimates:
## mean of x mean of y
## 0.7434562 0.6447540
```

```r
# Contraste de hipótesis para la concentración de ácido cítrico
 t.test(df.wines.good$citric.acid, df.wines.bad$citric.acid,
        alternative = "greater", var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  df.wines.good$citric.acid and df.wines.bad$citric.acid
## t = 8.7855, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.09921937        Inf
## sample estimates:
## mean of x mean of y
## 0.3764977 0.2544067
```

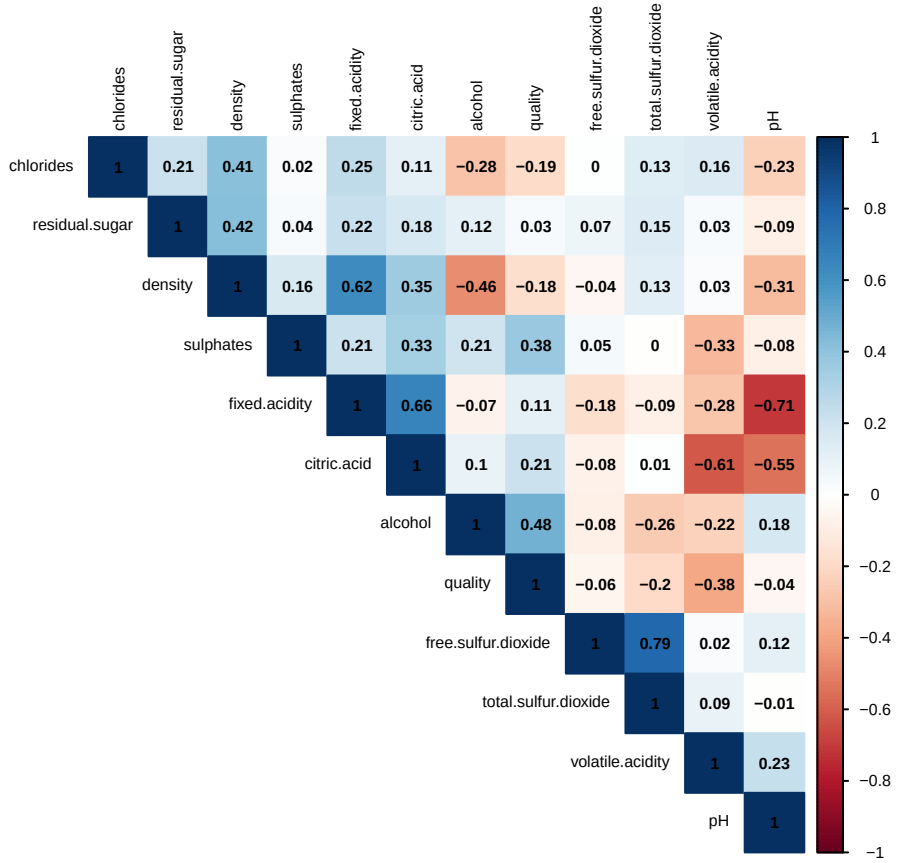
En los tres casos el valor p es inferior a $\alpha = 0.05$ y por tanto podemos rechazar la hipótesis nula con un 95% de confianza. Podemos afirmar con un nivel de significancia $\alpha = 0.05$ que los vinos de buena calidad tienen un mayor porcentaje de alcohol, una mayor concentración de sulfatos y mayor concentración de ácido cítrico, responsables de añadir más frescura y sabor y evitar su oxidación.

### 2.5.2 Correlaciones

Para estudiar la correlación entre las variables no podemos recurrir al coeficiente de correlación lineal de *Pearson* debido que los datos no siguen una distribución normal. El coeficiente de correlación de *Spearman* es la alternativa no paramétrica. Mide el grado de dependencia entre dos variables sin asumir ningún tipo de suposición sobre la distribución de los datos.

Tras calcular la matriz de correlación vemos que los factores que más influyen positivamente en la calidad del vino son el porcentaje de alcohol y la concentración de sulfatos, mientras que la acidez volátil lo hace negativamente. También observamos la relación que mantienen las variables explicativas entre si: la densidad depende en gran medida de otras variables como el alcohol, los ácidos fijos y el azúcar residual; el pH depende sobretodo de el ácido cítrico y los ácidos fijos; y el dióxido de azufre total está altamente relacionado con el dióxido de azufre libre. Deberíamos tener en cuenta estas relaciones si queremos obtener información útil de los regresores al elaborar un modelo de regresión logística debido a la multicolinealidad entre las variables explicativas. De todos modos, en general no se trata de correlaciones con un coeficiente demasiado alto.

```
M <- cor(df.wines[,-length(df.wines)], method = "spearman")

corrplot(M, method="color", type="upper", cl.cex=0.5,
         order="hclust", addCoef.col = "black",
         number.cex = .5, tl.col="black", tl.cex=0.5)
```



### 2.5.3 Modelo de regresión logística

Finalmente, y aprovechando los conocimientos que hemos deducido de los demás apartados, crearemos un modelo de regresión logística para predecir si un vino es de bueno calidad en función de las características que tienen más impacto. Dividmios los datos en train y test para posteriormente evaluar el modelo.

```
# Test y Train sets
set.seed(123)
training.samples <- createDataPartition(df.wines$bin_quality, p = 0.7, list = FALSE)
train.data  <- df.wines[training.samples, ]
test.data <- df.wines[-training.samples, ]

# Ajuste de un modelo logístico.
model <- glm(bin_quality ~ alcohol + sulphates + volatile.acidity,
                        data = train.data, family = "binomial")

# Resumen
summary(model)
```

```
##
## Call:
## glm(formula = bin_quality ~ alcohol + sulphates + volatile.acidity,
##     family = "binomial", data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4331  -0.5105  -0.2857  -0.1821   2.8221
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -12.35089    1.19767 -10.312  < 2e-16 ***
## alcohol            0.97549    0.09114  10.703  < 2e-16 ***
## sulphates          2.28957    0.50566   4.528 5.96e-06 ***
## volatile.acidity  -3.33412    0.64381  -5.179 2.23e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 939.9  on 1119  degrees of freedom
## Residual deviance: 708.2  on 1116  degrees of freedom
## AIC: 716.2
##
## Number of Fisher Scoring iterations: 6
```

Todos los regresores son estadísticamente significativos y el valor de AIC ronda los 700. Cuando los regresores son positivos los factores se denominan factores de riesgo. Estos factores influyen en la desviación hacia el factor que no es el de referencia, en nuestro caso, el vino de calidad. Recuperamos, por tanto, los resultados obtenidos en el apartado anterior sobre la correlación, que nos decía que el porcentaje de alcohol y la concentración de sulfatos afectan positivamente a la calidad del vino.

Para calcular la capacidad de diagnóstico del modelo comparamos los valores estimados del conjunto de test con los valores registrados. Obtenemos que el modelo clasifica correctamente el 90% de los datos cuando el umbral de discriminación es 0.5.

```
# Make predictions
probabilities <- predict(model, test.data, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, 1, 0)

# Model accuracy
mean(predicted.classes == test.data$bin_quality)
```

```
## [1] 0.8997912
```

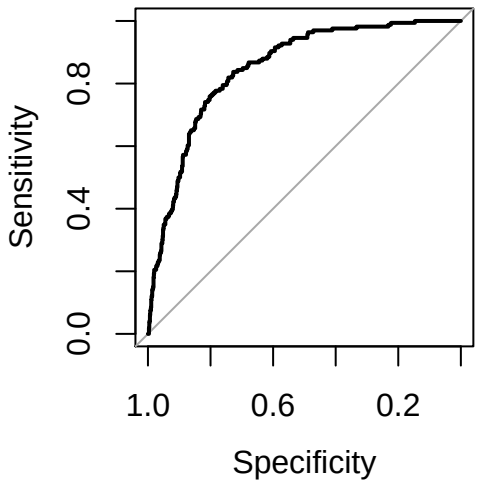
Para calcular la bondad del ajuste también podemos realizar un test *Hosmer-Lemeshow*, que se usa en modelos lineales generalizados con distribución binomial. La hipótesis nula de este test asume que no existe diferencia estadística entre los valores observados y los valores estimados. El valor p es mayor al nivel de significancia $\alpha = 0.05$, por tanto no podemos rechazar la hipótesis nula. No existe diferencia estadística entre los valores observados y los valores estimados, lo que significa que el modelo está bien ajustado.

```
#Test Hosmer-Lemeshow
hoslem.test(train.data$bin_quality,fitted(model))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  train.data$bin_quality, fitted(model)
## X-squared = 10.405, df = 8, p-value = 0.2378
```

Análogamente, podemos dibujar la curva ROC, que muestra la relación entre la sensibilidad y la especificidad para diferentes umbrales de discriminación. Cuanto mayor sea el área bajo la curva mejor será el valor de diagnóstico del modelo.

```
# Curva Roc
r=roc(train.data$bin_quality, model$fitted.values)
plot(r)
```




```
# Area bajo la curva
auc(r)
```

```
## Area under the curve: 0.8484
```

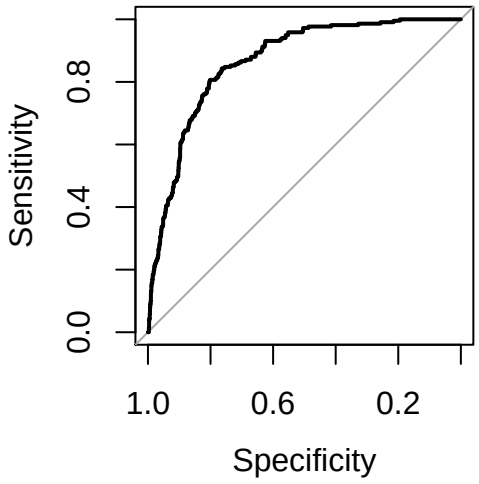
### 2.5.4 Cross Validation

Podemos intentar entrenar el modelo utilizando validación cruzada. En este método todas las observaciones se usan tanto para entrenar como testear el modelo la misma cantidad de veces.

```
# Train control
train_control <- trainControl(method = "cv", number = 10)

# Entrenamos el modelo
modelCV <- train(bin_quality ~ alcohol + sulphates + volatile.acidity,
                 data = df.wines, method = "glm",
                 trControl = train_control)

# Make predictions
probabilities <- predict(modelCV)
predicted.classes <- ifelse(probabilities > 0.5, 1, 0)

# Model accuracy
mean(predicted.classes == df.wines$bin_quality)
```

```
## [1] 0.8686679
```

```
# Curva Roc
r=roc(df.wines$bin_quality, probabilities)
plot(r)
```




```
# Area bajo la curva
auc(r)
```

```
## Area under the curve: 0.8656
```

Obtenemos un modelo muy similar al anterior. La precisión del modelo para los datos de la muestra es ligeramente menor, pero gracias a la validación cruzada conseguimos un modelo con una capacidad de diagnóstico ligeramente superior.

### 2.6 Conclusiones

Hemos estudiado cuáles son las principales características que influyen en la calidad de un vino tinto de la variedad *"Vinho Verde"*. Tras limpiar el conjunto de datos se han realizado un conjunto de pruebas estadísticas. Para conocer las diferencias entre los vinos de buena y mala calidad se han ejecutado un seguido de contrastes de hipótesis que nos permiten extrapolar los resultados extraídos de nuestra muestra al resto de la población. A continuación, se ha estudiado las relaciones entre la calidad y las variables explicativas, así como las relaciones entre las diferentes variables independientes, mediante la correlación de *Spearman*, test que se puede aplicar sin asumir una distribución en los datos. Finalmente, utilizando los conocimientos previos se han ajustado un par de modelo de regresión logística para predecir si un vino es de buena calidad basándose en su composición y se ha comprobado que estén bien ajustados. El modelo entrenado mediante validación cruzada consigue una capacidad de diagnóstico ligermaente superior.

## 3 Recursos


- P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
- Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
- Bernadó E. (2020). Contrastes de hipótesis. Editorial UOC.
- Guillén M., Alonso M. (2020). Modelos de regresión logística. Editorial UOC.