Práctica 2

Edita Talledo

Dic 2021

Contents

1 DESCRIPCION DEL DATASET

Este dataset [Cortez et al., 2009] contiene caracteristicas fisícoquimicas del vino portugues, como son el pH, la densidad, la acidez entre otros que permiten determinar la calidad de un buen vino.

LECTURA DE DATOS

```
data<-read.csv("./winequality-red.csv",header=T,sep=",")
```

ATRIBUTOS

```
str(data)
```

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

Descripción de variables

- fixed.acidity : ácidez fija o total (ácido tartárico y málico) provenientes de la uva.
- volatile.acidity : ácidez volátil o ácido acético adquirida en la vinificación.
- citric.acid : ácido citrico procedende de la uva.
- residual.sugar : azucar residual.
- chlorides : Cloruros.
- free.sulfur.dioxide : Dioxido de azufre Libre incluidos en la conservación.
- total.sulfur.dioxide: Dioxido de azufre total de incluidos en la conservación.
- density : Densidad.
- pH : medida de acidez-pH.
- sulphates : Sulfatos incluidos en la conservación.
- alcohol : Alcohol.
- quality : Calidad.

Este dataset contiene 1599 observaciones y 12 atributos, los cuales uno de ellos provee la calidad del vino en 10 categorias.

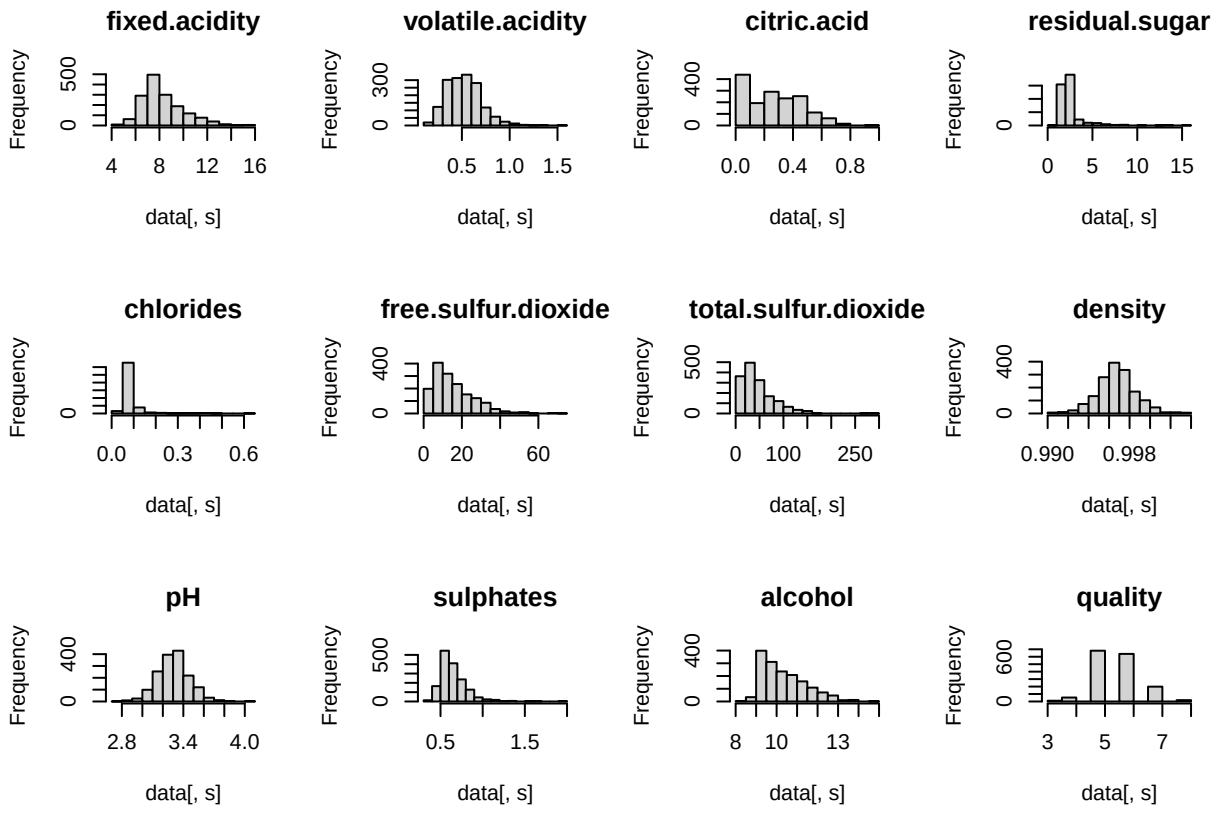
1.1 ¿Porque es importante y que pregunta/problema pretende responder/resolver?

El conocimiento de la calidad de un vino es parte importante en procesos de certificación los cuales buscan asegurar un buen producto, de calidad y seguros para la salud de las personas. El conjunto de datos permitira construir un modelo de clasificación a partir de sus atributos fisicoquímicos, ya que estos tienen asignados una calificación en la variable objeto (*quality*) que permitira al modelo aprender a clasificar. Para este fin se empleará el método de regresión logística.

2 INTEGRACIÓN Y SELECIÓN DE DATOS DE INTERES

```
titulos = c("fixed.acidity","volatile.acidity","citric.acid",
            "residual.sugar","chlorides", "free.sulfur.dioxide",
            "total.sulfur.dioxide","density","pH", "sulphates",
            "alcohol", "quality")

par(mfrow=c(3,4))
pl = lapply(X=titulos, FUN=function(s)
   hist(data[, s], main=paste(s)))
```



Por ahora cada variable contenida en el dataset es importante, la eliminación de alguna de ellas dependera de su aporte al modelo y esto será evaluado en el apartado 4. Ahora analizemos la variable *quality*, como se observa de los histogramas es de tipo categórico y sus categorias son:

```
unique(data$quality)
```

```
## [1] 5 6 7 4 8 3
```

Segun el documento de Cortez este atributo esta clasificado en 10 categorias, de las cuales en este dataset estan presentes solo 6 de ellas. Por tal, para un mejor análisis reagrupamos la variable en 2 categorias y la guardamos en una nueva variable *quality_c*:

```
data$quality_c[0 < data$quality & data$quality < 7] <- "baja"
data$quality_c[7 <= data$quality & data$quality < 10] <- "alta"
```

Lo convertimos a factor

```
data[c(13)] <- lapply(data[c(13)], factor)
```

3 LIMPIEZA DE DATOS

3.1 ¿Los datos contienen ceros o elementos vacios

```
missing(data)
```

```
## [1] FALSE
```

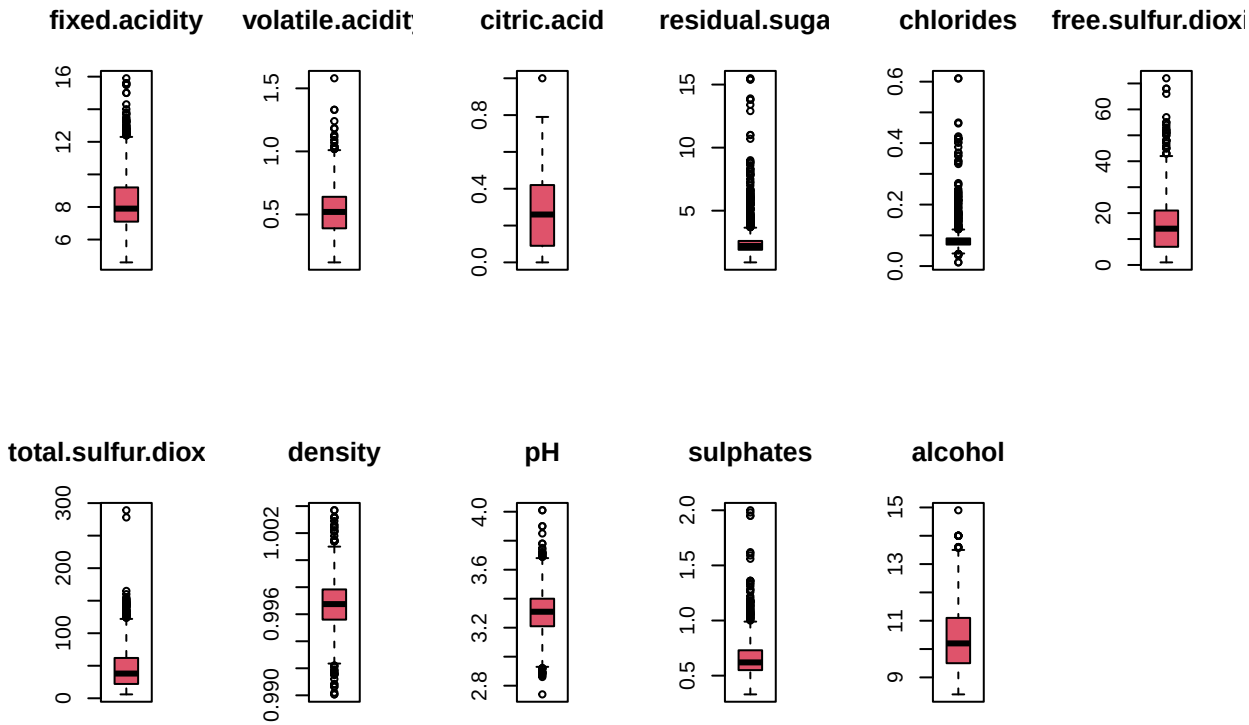
Como se puede ver, del comando anterior el dataset no contiene datos nulos o vacios

3.2 ¿Como gestionarias cada uno de estos casos?

En caso se hubieran encontrado NA, primero emepezariamos por conocer el porcentaje de NA, luego decidir si se eliminan, se conservan, se imputan por la media, o por algún otro método como el kNN (vecinos mas cercanos).

3.3 Identificación y tratamiento de valores extremos

Para identificar valores extremos podemos ayudarnos de los boxplot.



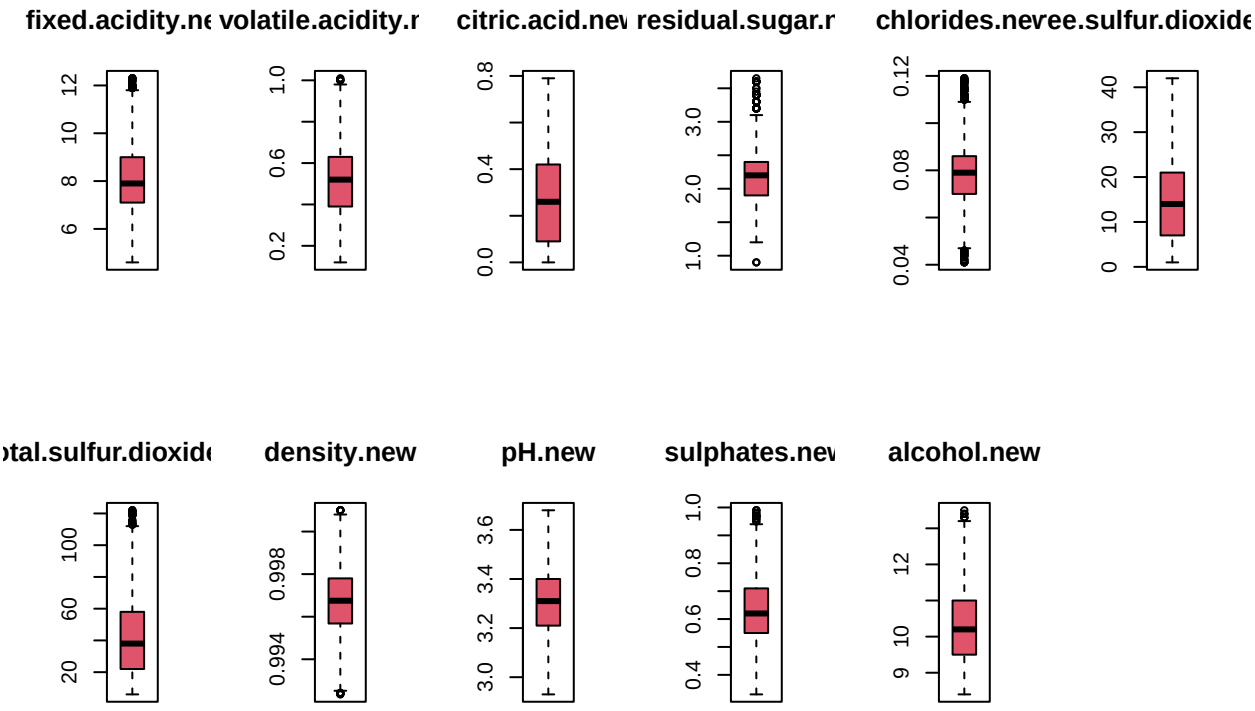
Se aprecia valores extremos en muchas variables. Una situación ideal sería conocer los rangos permitidos de cada una de las variables para poder tomar una decisión adecuada sobre los datos. En esta oportunidad como no se cuenta con esta información se procedera a reemplazar los outliers por sus medianas, luego se verificará su pertinencia al evaluar los modelos.

Creamos una función que sustituira los outliers con la mediana de los datos.

```
remove_outliers <- function(x) {
  qnt <- quantile(x, probs=c(.25, .75))
  lim <- 1.5 * IQR(x)
  y <- x
  me = median(x)
  y[x < (qnt[1] - lim)] <- me
  y[x > (qnt[2] + lim)] <- me
  y
}
```

Ahora lo aplicamos sobre cada uno de los datos.

Verificamos



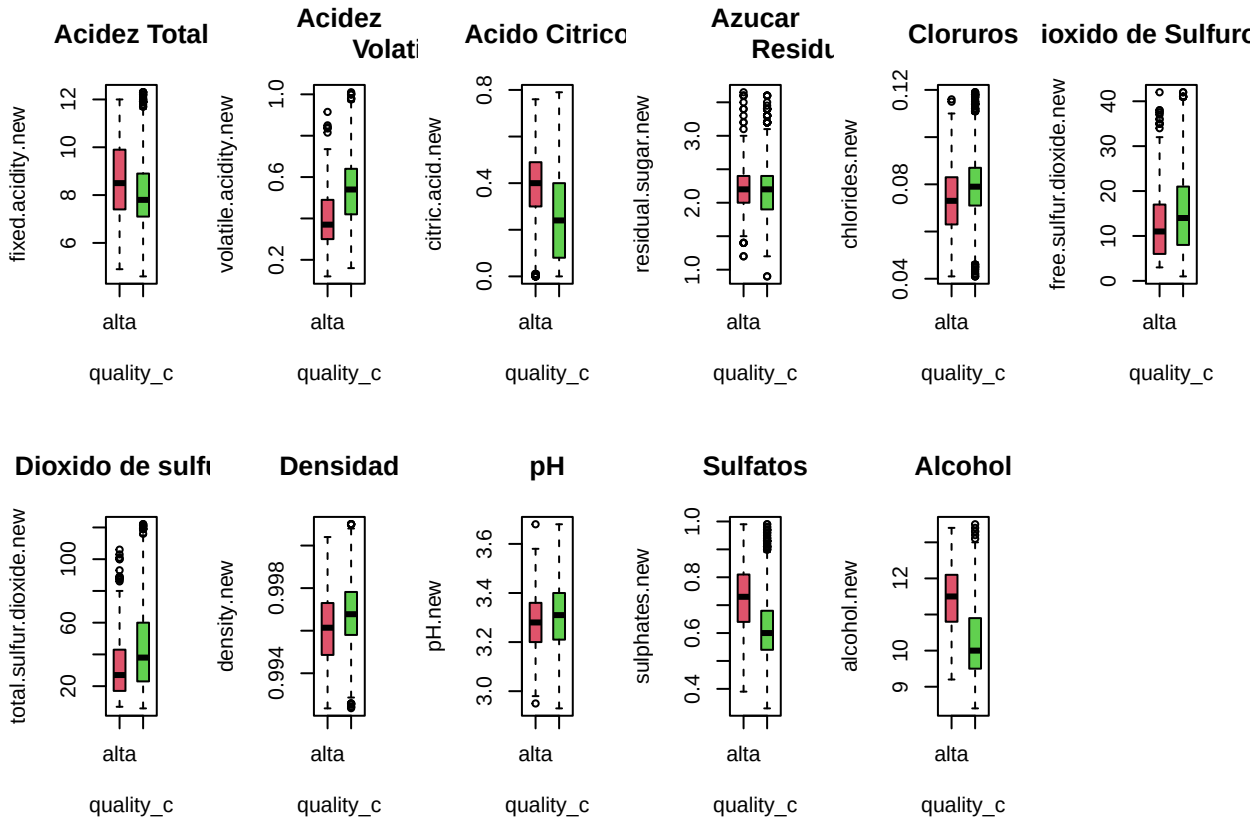
Se ha logrado suprimir varios outliers.

4 ANÁLISIS DE LOS DATOS

Como nos interesa evaluar la calidad del vino en función de cada una de las caracteristicas, se analizará las variables cuantitativas con respecto a las dos categorias de calidad: alta (calificación mayor o igual 7) y baja (calificación menor a 7). Luego aplicaremos las pruebas de normalidad, homocedasticidad y finalmente el test de medias entre los diferentes grupos.

4.1 Selección de los grupos de datos que se quieren comparar.

Exploración descriptiva según grupos de calidad



Para vinos de alta calidad, se observo que la acidez total, el acido cítrico, los sulfatos y el alcohol tienen medianas mayores en comparación de la mediana de vinos de baja calidad. Por el contrario, valores menores de la mediana se observa para la ácidez volatil, los cloruros, los dioxidos de sulfuros, la densidad y el pH en altas calidades de vinos. Estos comportamientos coinciden con lo establecido en la literatura, ya que se conoce que vinos de alta calidad suelen tener alta acidez fija y baja acidez volatil que suele ser la acidez final después del proceso de vinificación, también se espera altos valores de alcohol y bajo pH.

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Se llevaran a cabo las pruebas estadísticas de normalidad y homocedasticidad también considerando las dos categorias de vino

PRUEBAS ESTADITICAS DE NORMALIDAD

```
shapiro.test(data$fixed.acidity.new)$p.value
```

```
## [1] 3.291794e-19
```

```
shapiro.test(data$volatile.acidity.new)$p.value
```

```
## [1] 3.034688e-09
```

```
shapiro.test(data$citric.acid.new)$p.value
```

```
## [1] 7.153016e-22
```

```
shapiro.test(data$residual.sugar.new)$p.value
```

```
## [1] 1.314813e-17
```

```
shapiro.test(data$chlorides.new)$p.value
```

```
## [1] 1.411583e-11
```

```
shapiro.test(data$free.sulfur.dioxide.new)$p.value
```

```
## [1] 8.830049e-27
```

```
shapiro.test(data$total.sulfur.dioxide.new)$p.value
```

```
## [1] 3.925572e-29
```

```
shapiro.test(data$density.new)$p.value
```

```
## [1] 5.227266e-05
```

```
shapiro.test(data$pH.new)$p.value
```

```
## [1] 0.00349143
```

```
shapiro.test(data$sulphates.new)$p.value
```

```
## [1] 2.722575e-18
```

```
shapiro.test(data$alcohol.new)$p.value
```

```
## [1] 4.811127e-26
```

Como se puede observar de los resultados anteriores, las pruebas de normalidad dan un p-valor menor al valor de significancia (0.05), se rechaza la hipotesis nula y se concluye que las variables no siguen una distribución normal.

PRUEBAS DE HOMOCEDASTICIDAD

```
#include = FALSE, echo = FALSE
fligner.test(fixed.acidity.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.0004305772
```

```
fligner.test(volatile.acidity.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.003099794
```

```
fligner.test(citric.acid.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.9359677
```

```r
fligner.test(residual.sugar.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.5703455
```

```r
fligner.test(chlorides.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.001265754
```

```r
fligner.test(free.sulfur.dioxide.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.0452541
```

```r
fligner.test(total.sulfur.dioxide.new ~ quality_c, data = data)$p.value
```

```
## [1] 5.606058e-06
```

```r
fligner.test(density.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.0004310362
```

```r
fligner.test(pH.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.4084443
```

```r
fligner.test(sulphates.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.2874236
```

```r
fligner.test(alcohol.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.116183
```

De las pruebas de homocedasticidad se obtuvo dos grupos:

- El primero, con p-valores menores a 0.05, se rechaza la hipotesis nula, por lo tanto se afirma que las varianzas de los grupos son estadisticamente diferentes y estos son: fixed.acidity, volatile.acidity, chlorides, free.sulfur.dioxide, total.sulfur.dioxide y density.
- El segundo, donde las pruebas de homocedasticidad dan un p-valor mayor al valor de significancia, se acepta la hipotesis nula y se concluye que los grupos son homocedasticos es decir tienen varianzas estadisticamente iguales con un nivel de confianza del 95% y estos son: citric.acid, residual.sugar, pH, sulphates y alcohol.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Habiendo obtenido los resultados de normalidad y homocedasticidad y observando que las variables no cumplen los criterios pasaremos a aplicar la prueba de Wilcoxon para determinar si los grupos *alta y baja* tienen o no distribuciones estadisticamente diferentes para cada una de los atributos del vino.

TEST DE WILCOXON

```
wilcox.test(fixed.acidity.new ~ quality_c, data = data)$p.value
```

```
## [1] 2.976862e-06
```

```
wilcox.test(volatile.acidity.new ~ quality_c, data = data)$p.value
```

```
## [1] 1.503246e-30
```

```
wilcox.test(chlorides.new ~ quality_c, data = data)$p.value
```

```
## [1] 8.851699e-09
```

```
wilcox.test(free.sulfur.dioxide.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.0002339916
```

```
wilcox.test(total.sulfur.dioxide.new ~ quality_c, data = data)$p.value
```

```
## [1] 6.676489e-11
```

```
wilcox.test(density.new ~ quality_c, data = data)$p.value
```

```
## [1] 1.490243e-08
```

```
wilcox.test(citric.acid.new ~ quality_c, data = data)$p.value
```

```
## [1] 2.198429e-17
```

```
wilcox.test(residual.sugar.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.1300983
```

```
wilcox.test(pH.new ~ quality_c, data = data)$p.value
```

```
## [1] 0.002593688
```

```
wilcox.test(sulphates.new ~ quality_c, data = data)$p.value
```

```
## [1] 4.432786e-32
```

```
wilcox.test(alcohol.new ~ quality_c, data = data)$p.value
```

```
## [1] 8.074494e-50
```

Se observa, excepto *residual.sugar*, en todas las pruebas que el p-valor es menor que el valor de significancia, luego se rechaza la hipotesis nula y se concluye que los grupos en análisis poseen distribuciones estadisticamente diferentes.

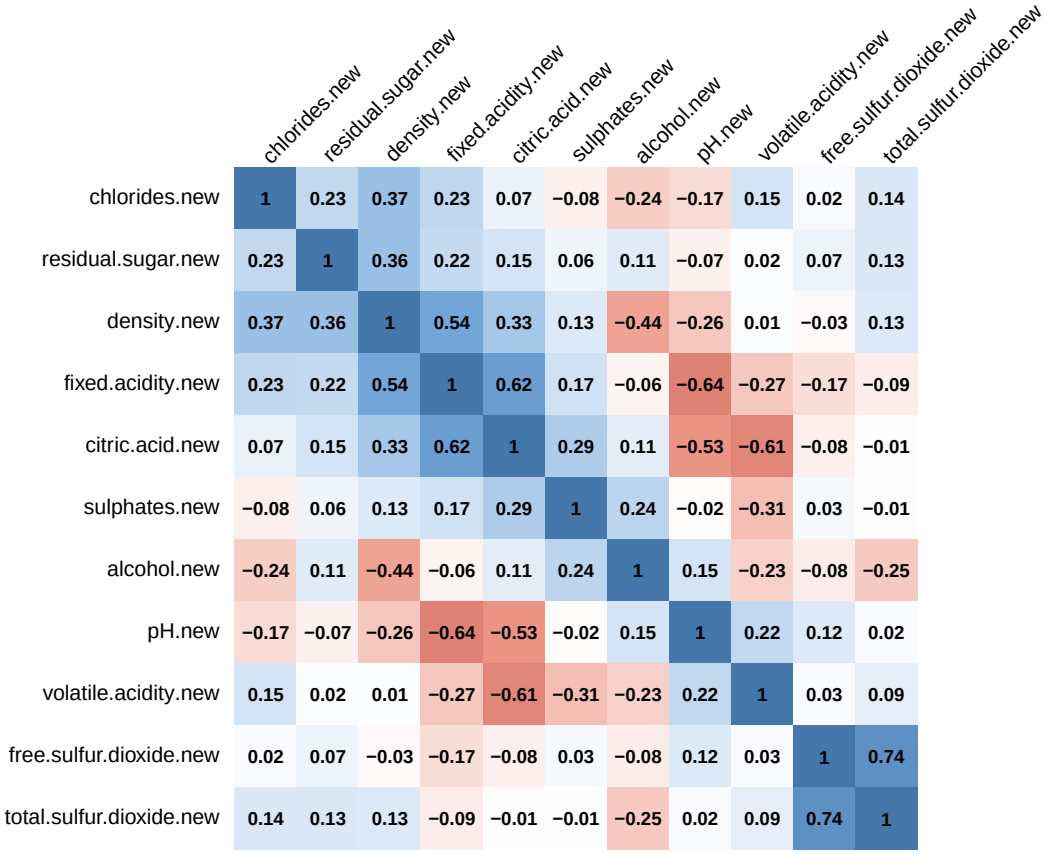
CORRELACIÓN

Deseamos conocer la relación entre las variables a manera de poder seleccionar aquellas que aportar mayor información al modelo de predicción. Para ello usaremos el comando **cor** y ya que las

variables no cumplen con el criterio de normalidad emplearemos la correlación de **Spearman**.

```
library(corrplot)

mcor_mean <- cor(na.omit(data[c(14:24)]), method = c("spearman"))
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD",
                          "#4477AA"))
corrplot(mcor_mean, method = "shade", shade.col = NA,
         tl.col = "black", tl.srt = 45, col = col(200),
         addCoef.col = "black",cl.pos = "n", order = "AOE",
         tl.cex=0.7, number.cex=0.6)
```



De la matriz de correlación se observa que existe una relación lineal moderada tanto positiva como negativa entre algunas variables. Para las variables con r>=|0.3| se evaluará si estas son significativamente diferente de cero. Emplearemos *cor.test*

```
# message= FALSE, warning=FALSE, included=FALSE,
cor.test(data$fixed.acidity, data$citric.acid, method = "spearman")$p.value
```

```
## [1] 5.385745e-202
```

```
cor.test(data$fixed.acidity, data$density, method = "spearman")$p.value
```

```
## [1] 1.268199e-172
```

```r
cor.test(data$fixed.acidity, data$pH, method = "spearman")$p.value
```

```
## [1] 3.182584e-242
```

```r
cor.test(data$volatile.acidity, data$citric.acid, method = "spearman")$p.value
```

```
## [1] 9.013427e-164
```

```r
cor.test(data$citric.acid, data$density, method = "spearman")$p.value
```

```
## [1] 6.301241e-48
```

```r
cor.test(data$citric.acid, data$sulphates, method = "spearman")$p.value
```

```
## [1] 3.29505e-42
```

```r
cor.test(data$citric.acid, data$pH, method = "spearman")$p.value
```

```
## [1] 5.089354e-126
```

```r
cor.test(data$residual.sugar, data$density, method = "spearman")$p.value
```

```
## [1] 3.701523e-70
```

```r
cor.test(data$chlorides, data$density, method = "spearman")$p.value
```

```
## [1] 2.428135e-66
```

```r
cor.test(data$free.sulfur.dioxide, data$total.sulfur.dioxide, method = "spearman")$p.value
```

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

```r
cor.test(data$fixed.acidity, data$citric.acid, method = "spearman")$p.value
```

```
## [1] 5.385745e-202
```

```r
cor.test(data$density, data$pH, method = "spearman")$p.value
```

```
## [1] 1.882939e-37
```

```r
cor.test(data$density, data$alcohol, method = "spearman")$p.value
```

```
## [1] 1.559709e-85
```

A partir de los test de correlación se obtuvo un p-valor menor al nivel de significación, se rechaza la hipotesis nula y se concluye que la correlación entre las variables es significativamente diferente de cero. Entre estas variables destacan el acido citrico y la densidad que tienen alta correlación con muchas variables.

REGRESIÓN LOGÍSTICA

Habiendo determinado que la relación entre las variables cuantitavas, construiremos el modelo de regresión logistica que determinará la relación de nuestra variable dependiente en este caso

quality_c y las explicativas. Empezaremos construyendo el modelo con una sola variable, luego se iran agregando la demás variables y observaremos el valor de información de Akike (AIC), que nos dira cuan bien se relaciona nuestra variable objeto con las variables explicativas. Antes recodificamos la variable *quality_c*.

```
data$quality_c = as.integer(ifelse(data$quality_c=="alta", 1, 0))
```

```
rlg1 = glm(quality_c ~ fixed.acidity.new, data = data, family = "binomial")
rlg1$aic
```

```
## [1] 1255.592
```

```
rlg11 = glm(quality_c ~ fixed.acidity.new+volatile.acidity.new+citric.acid.new+
              residual.sugar.new+chlorides.new+free.sulfur.dioxide.new+
              total.sulfur.dioxide.new+density.new+pH.new+sulphates.new+alcohol.new,
            data = data, family = "binomial")
summary(rlg11)
```

```
##
## Call:
## glm(formula = quality_c ~ fixed.acidity.new + volatile.acidity.new +
##     citric.acid.new + residual.sugar.new + chlorides.new + free.sulfur.dioxide.new +
##     total.sulfur.dioxide.new + density.new + pH.new + sulphates.new +
##     alcohol.new, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0655  -0.4491  -0.2289  -0.1282   2.9198
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               33.730936  80.444213   0.419   0.6750
## fixed.acidity.new          0.002972   0.084336   0.035   0.9719
## volatile.acidity.new      -3.238227   0.779429  -4.155 3.26e-05 ***
## citric.acid.new            0.061355   0.783388   0.078   0.9376
## residual.sugar.new         0.151508   0.220721   0.686   0.4924
## chlorides.new             -8.596443   6.600459  -1.302   0.1928
## free.sulfur.dioxide.new    0.002345   0.012849   0.183   0.8552
## total.sulfur.dioxide.new  -0.013885   0.005543  -2.505   0.0123 *
## density.new              -42.021938  81.114845  -0.518   0.6044
## pH.new                    -1.617750   0.885665  -1.827   0.0678 .
## sulphates.new              5.450911   0.747088   7.296 2.96e-13 ***
## alcohol.new                0.945433   0.115778   8.166 3.19e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
```

```
## Residual deviance:  884.57  on 1587  degrees of freedom
## AIC: 908.57
##
## Number of Fisher Scoring iterations: 6
```

Se observo que el AIC fue disminuyendo conforme se ingresaban las variables, para finalmente ir de 1252.4 a 908.57. También se sabe que a menor valor del AIC se tiene un mejor modelo ya que este valor considera la bondad de ajuste, como la complejidad del modelo. Por tal nos quedaremos con el último modelo y es el que contiene todas las variables. Por no extender mas el documento no se muestra los resultados de los modelos intermedios, solo se muestra el primer y el último modelo.

Ahora analizamos los coeficientes del modelo seleccionado. A vista de los datos se observa que los p-valor asociados a las variables **volatile.acidity, total.sulfur.dioxide, sulphates y alcohol** son menores al nivel de significancia (0.05), por tal se rechaza la hipotesis nula y concluimos que solo estas variables son estadisticamente significativas.

Ahora probemos construyendo el modelo solo con las variables significativas y retiramos las variables no significativas:

```
rlg12 = glm(quality_c ~ volatile.acidity.new+total.sulfur.dioxide.new+
              sulphates.new+alcohol.new,data = data, family = "binomial")
#summary(rlg12)
rlg12$aic
```

```
## [1] 901.6816
```

El valor de AIC ha disminuido ligeramente retirando las variables que son no significativas. Nos quedaremos con este modelo *rlg12*

5 REPRESENTACIÓN DE LOS RESULTADOS A PARTIR DE TABLAS O GRÁFICAS.

PREDICCIÓN DEL MODELO

Ahora evaluaremos nuestro modelo

```
library(caret)
prediccion = predict(rlg12, newdata = data[14:24], type = "response")
data$prediccion = as.integer(ifelse(test=prediccion>0.5, yes = 1, no=0))
# Variable quality_c
data[c(13)] <- lapply(data[c(13)], factor)
# Variable predicción
data[c(25)] <- lapply(data[c(25)], factor)
confusionMatrix(data$quality_c, data$prediccion)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
```

```
##          0 1337   45
##          1  153   64
##
##                Accuracy : 0.8762
##                  95% CI : (0.859, 0.8919)
##     No Information Rate : 0.9318
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.332
##
##  Mcnemar's Test P-Value : 2.868e-14
##
##             Sensitivity : 0.8973
##             Specificity : 0.5872
##          Pos Pred Value : 0.9674
##          Neg Pred Value : 0.2949
##              Prevalence : 0.9318
##          Detection Rate : 0.8361
##    Detection Prevalence : 0.8643
##       Balanced Accuracy : 0.7422
##
##        'Positive' Class : 0
##
```

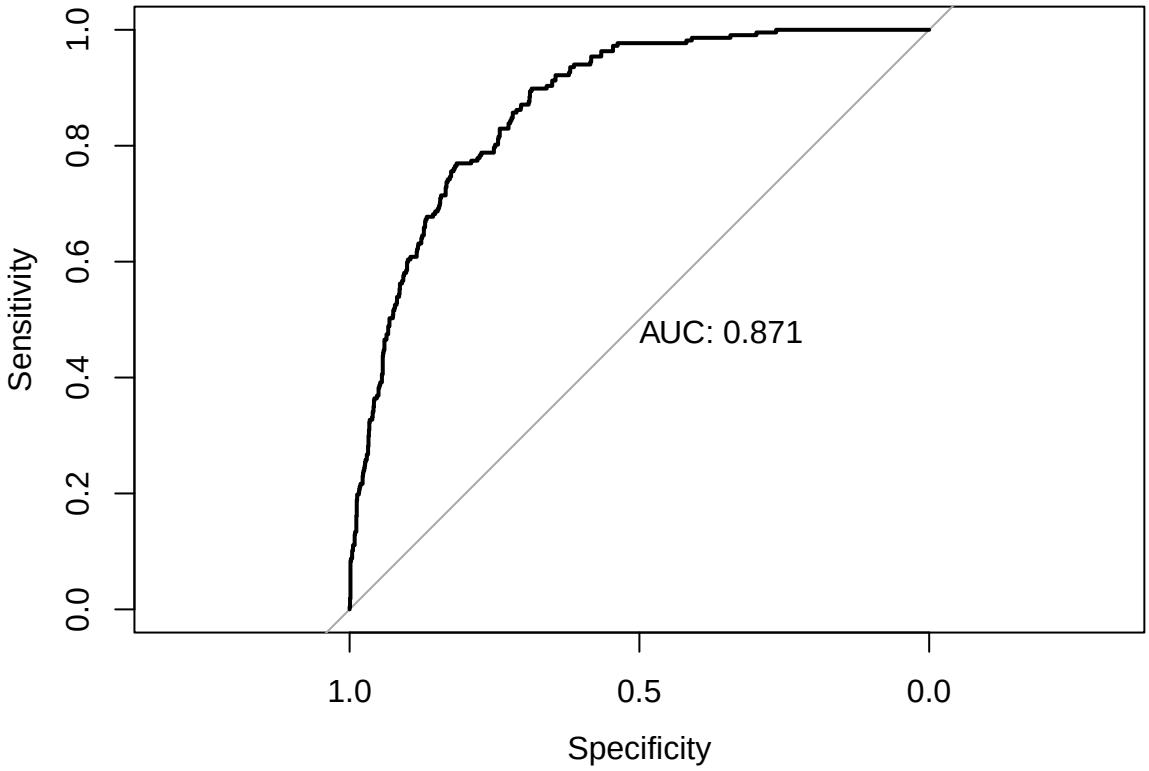
A vista de los datos el modelo tiene una exactitud de 88%, pero mas nos interesa al sensibilidad cuyo valor es de 90%, que nos indica que el modelo tiene capacidad de clasificar correctamente el 90% de los registros de vino a partir de sus atributos fisicoquímicos.

CURVA ROC

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.1.2
```

```
prob = predict(rlg12,data, type = "response")
r=roc(data$quality_c, prob, data=data)
plot(r, print.auc = TRUE)
```



El valor del area bajo la curva (AUC) nos da información sobre la calidad del modelo, aqui obtuvimos un AUC = 0.871, por lo tanto se concluye que el modelo discrimina de forma excelente la calidad del vino.

6 RESOLUCIÓN DEL PROBLEMA.

6.1 A partir de los resultados obtenidos, ¿cuáles son las conclusiones?

- Del diagrama de cajas se observo valores esperados en las medianas de las caracteristicas fisicoquímicas del vino para cada uno de los grupos de alta y baja calidad. Podriamos decir que se realizó una adecuada partición del grupo de calidad, así mismo que se cuenta con datos representativos ya que estos se comportan como se establece en la literatura.
- Se obtuvo a partir de las pruebas estadísticas para cada una de las variables, que los grupos diferenciados de alta y baja calidad tienen distribuciones estadísticamente diferentes.
- Se construyo el modelo de regresión logística considerando solo las siguientes variables:la acidez volatil, el dioxido de azufre total y el alcohol, ya que solo estas resultaron ser estadisticamente significativas, pero estos resultados son coherentes ya que son estas variables las que se tienen que controlar para dar el grado de calidad esperada. El grado de acidez volatil es la acidez

final producida la cual no debe ser muy alta pero la suficiente para dar frescura al vino, el dioxido de azufre incoorporado en algunos vinos para la conservación puede dar mal sabor al vino y el grado de alcohol es importante para dar el equilibrio de calidez al vino.

- El modelo construido fue evaluado con el mismo conjunto de datos que se construyo el modelo. Se obtuvo una exactitud de 88%, una sensibilidad del 90% y un AUC del 88%, valores que indican el buen desempeño del modelo para clasificar.

6.2 ¿Los resultados permiten responder al problema?

El modelo construido, dados los resultados de exactitud, sensibilidad y auc nos permiten tener cierta confianza para clasificar vinos, dadas sus especificaciones fisicoquímicas. Una buena estrategia para evaluar mejor el modelo puedo ser, dividir el modelo en dos grupos uno primer grupo serviría para construir el modelo y el segundo para evaluarlo, en lugar de usar el mismo conjunto de datos como se hizo en esta práctica. Esto daría un resultado mas confiable ya que serían datos nuevos los que estarian evaluando al modelo.