PR 2. Tipología y ciclo de vida de los datos

Tratamiento del dataset Heart Attack Analysis & Prediction

Patricia Luengo Carretero

Tipología y ciclo de vida de los datos

Máster de Ciencia de Datos

1. Descripción del dataset

El conjunto de datos que vamos a tratar recoge una serie de variables relacionadas con el ataque al corazón.

El objetivo es tratar de clasificar si una persona tiene más o menos posibilidades de tener un ataque al corazón en función de las características estudiadas.

Tenemos un dataset con 14 variables y 303 observaciones.

→ **age**: Edad del paciente **sex**: Sexo del paciente **cp**: Tipo de dolor en el pecho, categorizado en:
- 1: typical angina (dolor típico)
- 2: atypical angina (dolor atípico)
- 3: non-anginal pain (dolor no relacionado con la angina de pecho)
- 0: asymptomatic (asintomático)

→ **trtbps**: presión arterial en reposo (in mm Hg)
→ **chol**: colesterol mg/dl obtenido a través del sensor BMI
→ **thall**: - 1: fixed defect - 2: normal - 3: reversable defect
→ **fbs**: (glucemia en ayunas > 120 mg/dl) (1 = true; 0 = false)
→ **restecg** : resultados electrocardiográficos en reposo
- 1: normal
- 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) tener anomalías en la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST > 0,05 mV)
- 0: showing probable or definite left ventricular hypertrophy by Estes' criteria (que muestra hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes)

→ **thalachh**: frecuencia cardíaca máxima alcanzada
→ **exng**: exercise induced angina (1 = yes; 0 = no)
→ **oldpeak**: Depresión del ST inducida por el ejercicio en relación con el reposo
→ **slp**: La pendiente del segmento ST de ejercicio máximo
→ **caa**: Número de vasos principales (0-4)
→ **output**:
→ - 0 = menos posibilidades de ataque al corazón
→ - 1 = más posibilidades de ataque al corazón

Para discernir cuáles eran las variables y sus valores se consultó la siguiente discusión

2. Integración y selección

Tenemos tres variables que desconocemos de que se tratan, así que no las vamos a utilizar, son las variables thall, oldpeak y slp.

Posteriormente se renombraron las columnas para facilitar la comprensión a 'edad', 'sexo', 'dolor', 'presion_arterial', 'colesterol', 'glucemia', 'electro', 'frec_cardiaca', 'ejercicio', 'vasos', 'ataque'.

También se renombraron los datos categóricos:

- → Sexo (hombre, mujer)
- → Dolor (típico, atipico, no relacionado, asintomatico)
- → Electro (normal, anomalías, hipertrofia)
- → Ejercicio (si,no)
- → Glucemia (tiene glucemia, no tiene glucemia)
- → Vasos (cero, uno, dos, tres, cuatro)

Y se modificaron a su tipo correspondiente que es factor.

3 Limpieza de los datos

Observamos que el dataset no tiene NA's.

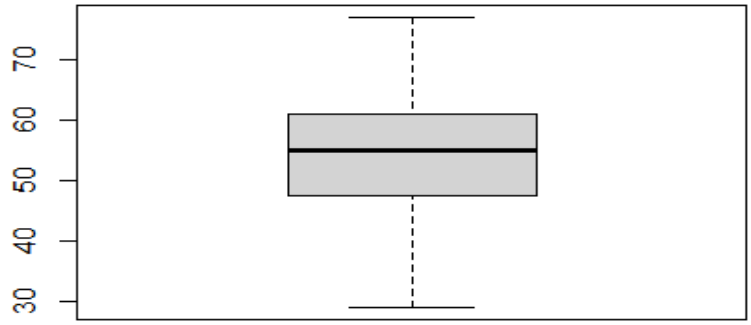
```
summary(heart)
```

```
##       edad           sexo                dolor     presion_arterial
##  Min.   :29.00   hombre:207   asintomatico  :143   Min.   : 94.0
##  1st Qu.:47.50   mujer : 96   atipico       : 87   1st Qu.:120.0
##  Median :55.00                no_relacionado: 23   Median :130.0
##  Mean   :54.37                tipico        : 50   Mean   :131.6
##  3rd Qu.:61.00                                     3rd Qu.:140.0
##  Max.   :77.00                                     Max.   :200.0
##    colesterol                 glucemia          electro    frec_cardiaca
##  Min.   :126.0   no_tiene_glucemia:258   anomalias  :  4   Min.   : 71.0
##  1st Qu.:211.0   tiene_glucemia   : 45   hipertrofia:147   1st Qu.:133.5
##  Median :240.0                           normal     :152   Median :153.0
##  Mean   :246.3                                             Mean   :149.6
##  3rd Qu.:274.5                                             3rd Qu.:166.0
##  Max.   :564.0                                             Max.   :202.0
##  ejercicio    vasos         ataque
##  no:204    cero  :175   Min.   :0.0000
##  si: 99    cuatro:  5   1st Qu.:0.0000
##            dos   : 38   Median :1.0000
##            tres  : 20   Mean   :0.5446
##            uno   : 65   3rd Qu.:1.0000
##                         Max.   :1.0000
```

En cuanto a los nulos vemos que tampoco hay, ya que las variables numéricas, edad, frecuencia cardiaca, colesterol y presión arterial no tienen como valor mínimo cero.

Valores extremos de la edad

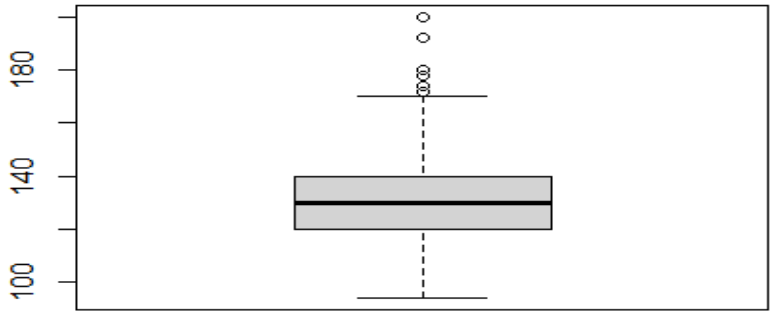
Dibujamos un diagrama de caja para estudiar la variable presión arterial:



Observamos que la variable no presenta valores extremos.

Valores extremos de la presión arterial en reposo

Dibujamos un diagrama de caja para estudiar la variable presión arterial:



Observamos que la variable presenta valores extremos, así que los visualizamos:

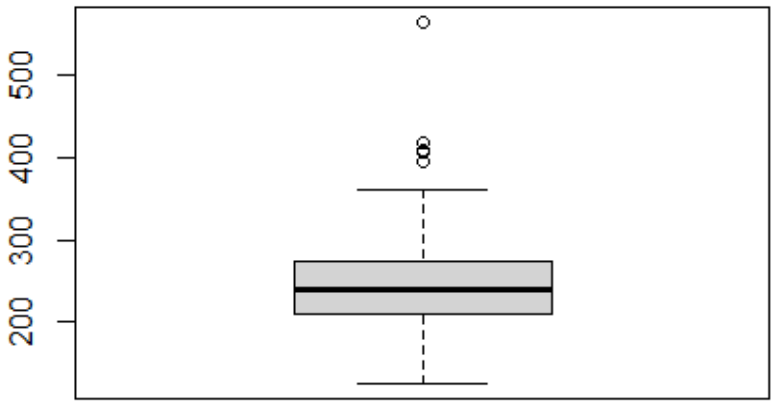
```
outliersPres
```

```
## [1] 172 178 180 180 200 174 192 178 180
```

Aunque son valores extremos se dan por legítimos ya que quizá sea un indicador de ataque al corazón.

Valores extremos del colesterol

Dibujamos un diagrama de caja para estudiar la variable colesterol:



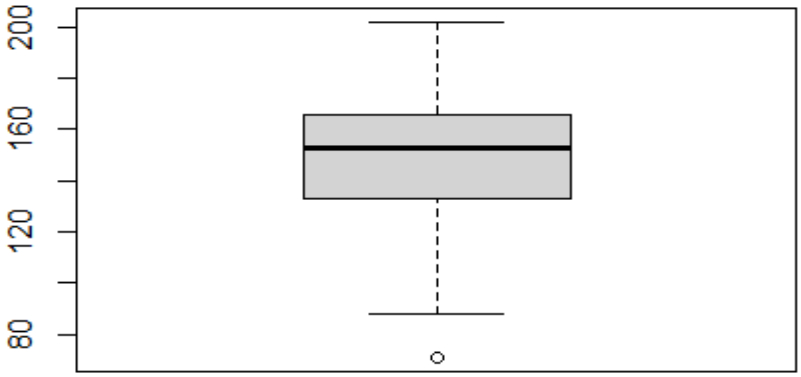

```
outliersCol
```

```
## [1] 417 564 394 407 409
```

Aunque son valores extremos se dan por legítimos ya que quizá sea un indicador de ataque al corazón, salvo el dato de 564 que sí parece ser un dato atípico y se elimina.

Valores extremos de la frecuencia cardíaca

Dibujamos un diagrama de caja para estudiar la variable colesterol:




```
outliersFrec
```

```
## [1] 71
```

Eliminamos el dato

4. Análisis de los datos

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

Agrupamos por un lado los datos numéricos y por otro los categóricos y factorizamos en este último la variable respuesta ataque.

```
heartNumerico <- select(heart, edad, presion_arterial, colesterol, frec_cardiaca, ataque
)
```

```
heartCategorico <- select(heart, sexo, dolor, glucemia, electro, ejercicio, vasos, ataqu
e)
heartCategorico$ataque[heartCategorico$ataque==1] <- "Si"
heartCategorico$ataque[heartCategorico$ataque==0] <- "No"
```

4.2. Variables numéricas.

4.2.1. Análisis univariante

Estudio de la normalidad de las variables

Estudiamos la normalidad de las variables, para ello se aplica el test de Shapiro-Wilk donde:

→ H_0: la población sigue una distribución normal > 0.05

→ H_1: la población no sigue una distribución normal < 0.05

```
shapiro.test(heartNumerico$edad)

##
##  Shapiro-Wilk normality test
##
## data:  heartNumerico$edad
## W = 0.98685, p-value = 0.007635

shapiro.test(heartNumerico$presion_arterial)

##
##  Shapiro-Wilk normality test
##
## data:  heartNumerico$presion_arterial
## W = 0.96646, p-value = 1.882e-06

shapiro.test(heartNumerico$frec_cardiaca)

##
##  Shapiro-Wilk normality test
##
## data:  heartNumerico$frec_cardiaca
## W = 0.97755, p-value = 0.0001159

shapiro.test(heartNumerico$colesterol)

##
##  Shapiro-Wilk normality test
##
## data:  heartNumerico$colesterol
## W = 0.9829, p-value = 0.001174
```

Todas las variables han tenido como resultado valores de pvalue < 0.05 así que se rechaza la hipótesis nula de que las variables siguen una distribución normal.

Estudio de la homocedasticidad de las variables

Como hemos asumido que las variables siguen una distribución normal aplicamos el test de Levene.

→ H_0: igualdad de varianzas entre los grupos > 0.05
→ H_1: diferencias significativas < 0.05

```
leveneTest(edad~factor(ataque), data=heartNumerico)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   1  7.7356 0.005758 **
##       299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(presion_arterial~factor(ataque), data=heartNumerico)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  1.9294 0.1659
##       299
```

```
leveneTest(frec_cardiaca~factor(ataque), data=heartNumerico)
```

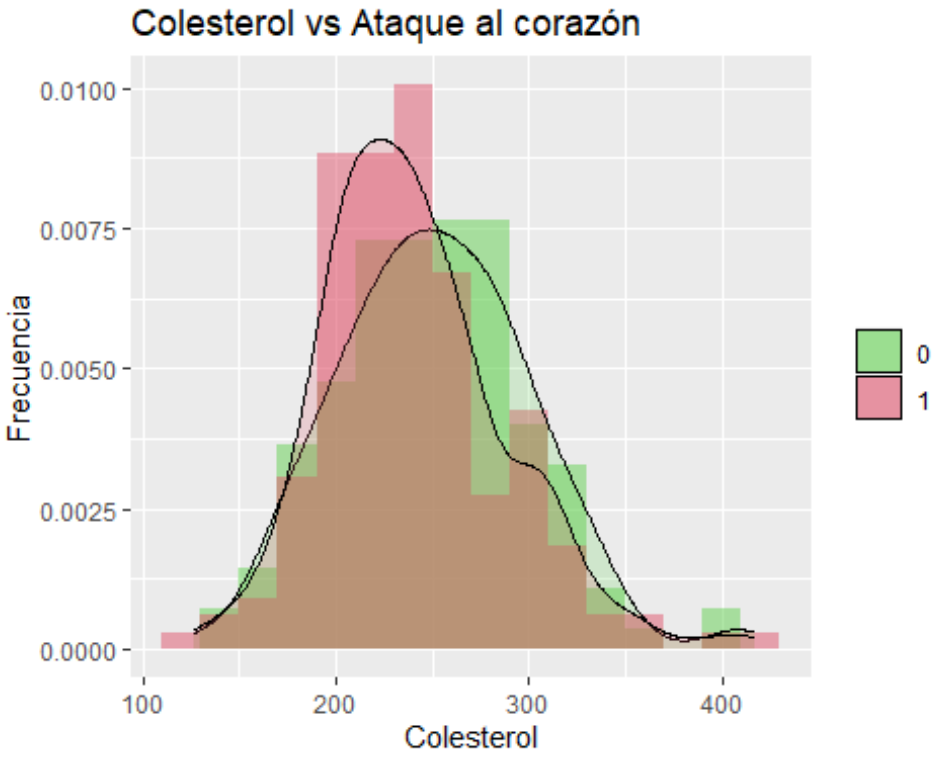
```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  4.0669 0.04463 *
##       299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

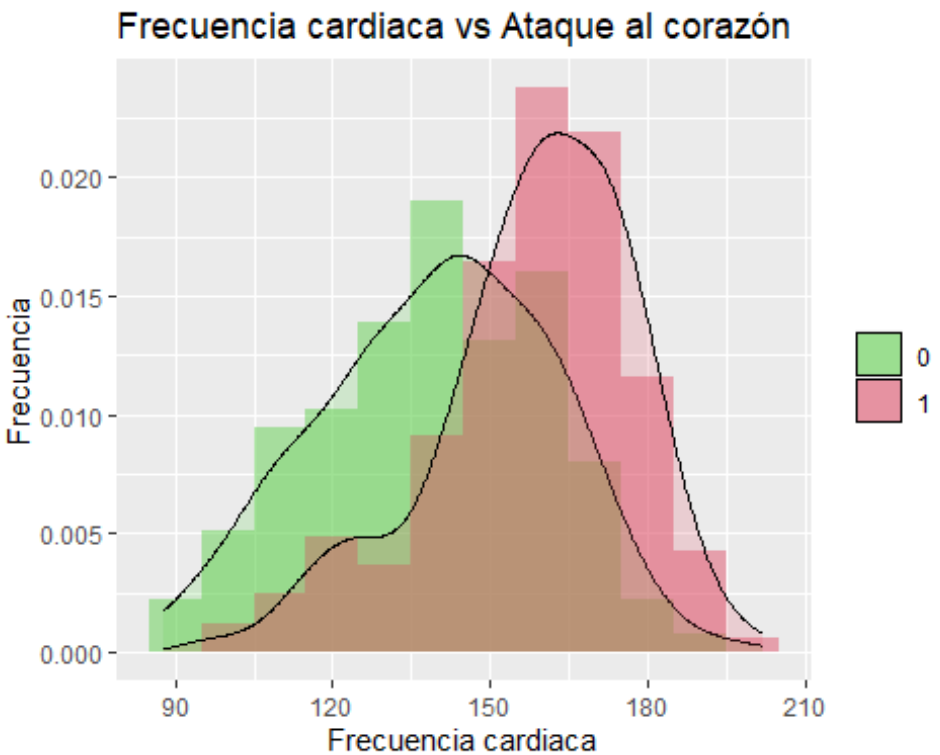
```
leveneTest(colesterol~factor(ataque), data=heartNumerico)
```

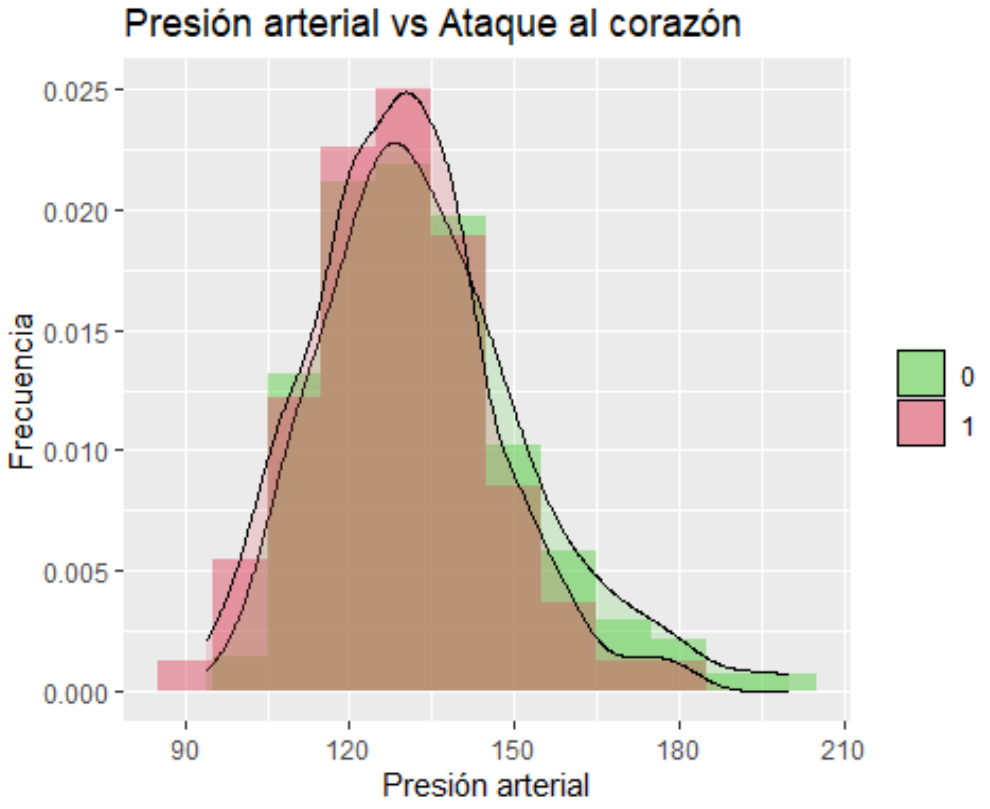
```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.8475  0.358
##       299
```

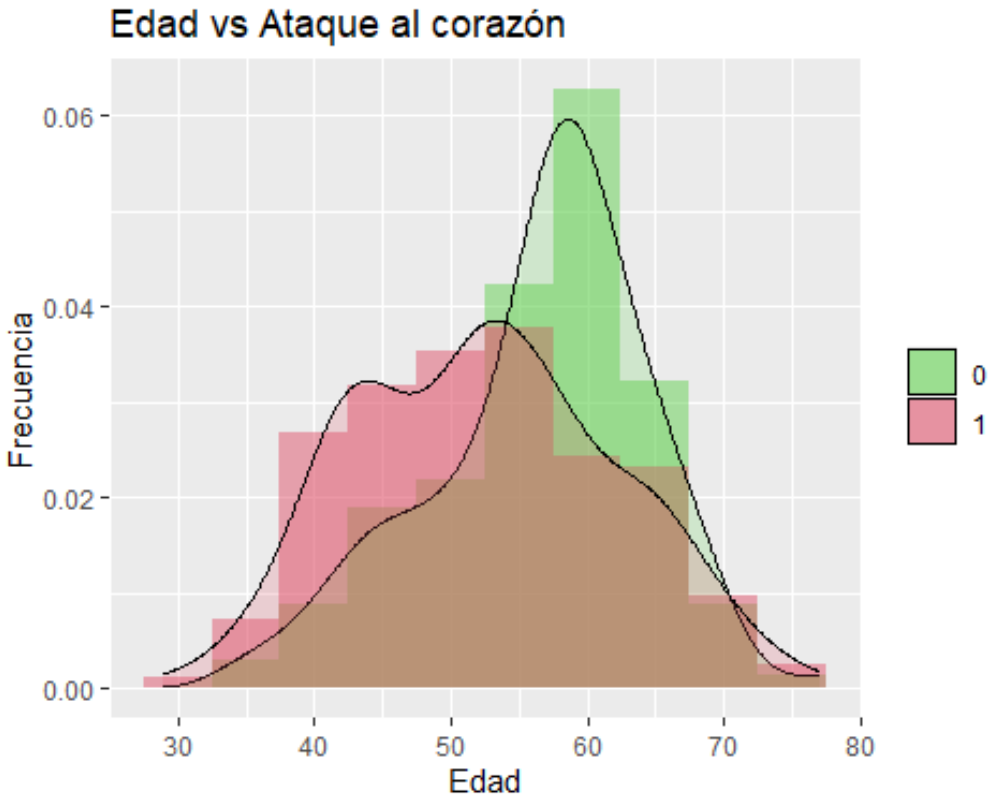
La presión arterial y el colesterol han dado resultados de pvalue superiores al nivel de significancia por tanto se acepta la hipótesis nula de homocedasticidad y se concluye que la variable de la presión arterial y el colesterol no presentan varianzas estadísticamente diferentes para los grupos de ataque al corazón, sin embargo, con la frecuencia cardíaca y la edad, sucede todo lo contrario.

Visualización de las variables numéricas frente a la variable respuesta


Colesterol vs Ataque al corazón
Frecuencia
0.0100
0.0075
0.0050
0.0025
0.0000
100
200
300
400
Colesterol
0
1


Frecuencia cardiaca vs Ataque al corazón
Frecuencia
0.020
0.015
0.010
0.005
0.000
90
120
150
180
210
Frecuencia cardiaca
0
1


Presión arterial vs Ataque al corazón
Frecuencia
0.025
0.020
0.015
0.010
0.005
0.000
90
120
150
180
210
Presión arterial
0
1


Edad vs Ataque al corazón
Frecuencia
0.06
0.04
0.02
0.00
30
40
50
60
70
80
Edad
0
1

4.2.3. Conclusiones

Los pvalores han dado como resultado valores menor al nivel de significancia por tanto se rechaza la hipótesis nula y se concluye que **los datos no siguen una distribución normal.** No obstante, como el conjunto de datos se compone de una muestra de registros suficientemente grande, por el **teorema central del límite, se puede considerar que los datos siguen una distribución normal.**

La **presión arterial y el colesterol** han dado resultados de pvalue superiores al nivel de significancia por tanto se acepta la hipótesis nula de **homocedasticidad** y se concluye que la variable de la presión arterial y el colesterol no presentan varianzas estadísticamente diferentes para los grupos de ataque al corazón, sin embargo, con la **frecuencia cardíaca y la edad, sucede todo lo contrario**.

En la **edad** se observa que al contrario de lo que se suele pensar, la edad con mayor factor de riesgo a la hora de tener un ataque al corazón se encuentra entre los 35 y 55 años.

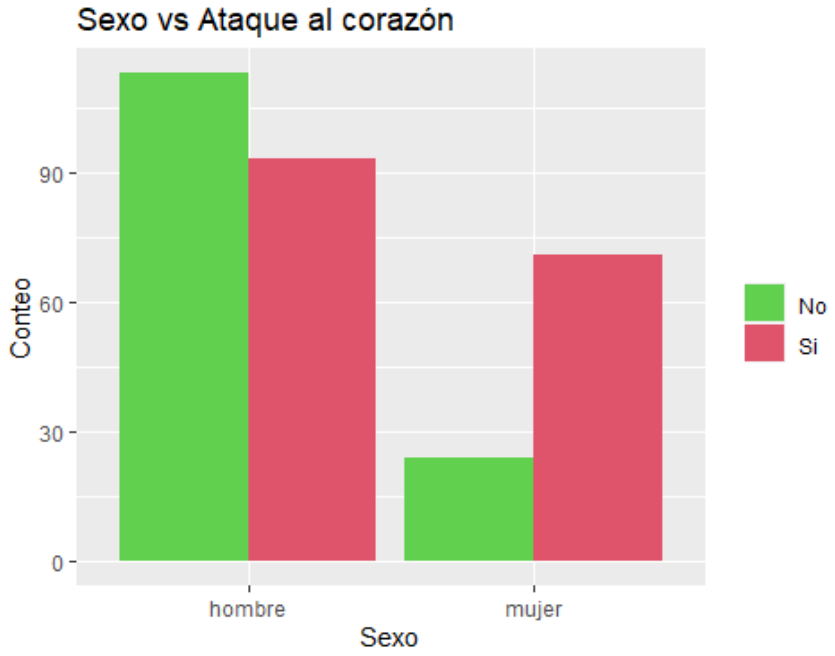
En el caso de la **presión arterial** es difícil decir si tienen influencia este dato con respecto a tener un ataque al corazón o no, ya que ambas gráficas son prácticamente iguales.

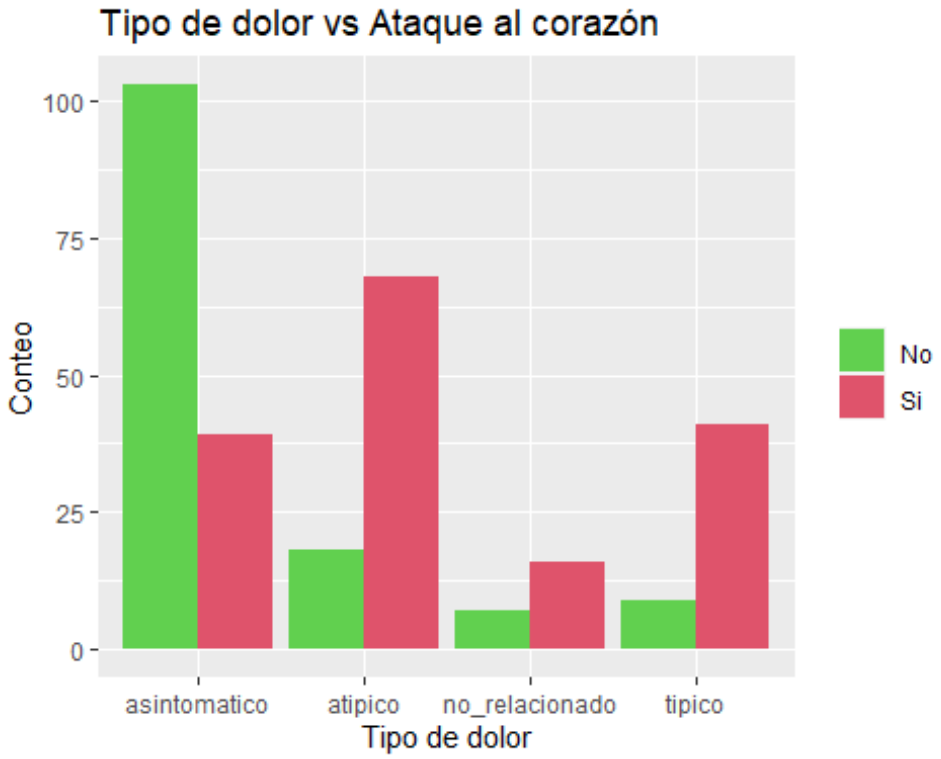
En el caso del nivel de **colesterol** sí que encontramos un rango entre 200 y 250 que es más propenso a tener un ataque al corazón, aunque la diferencia entre las gráficas es pequeña.

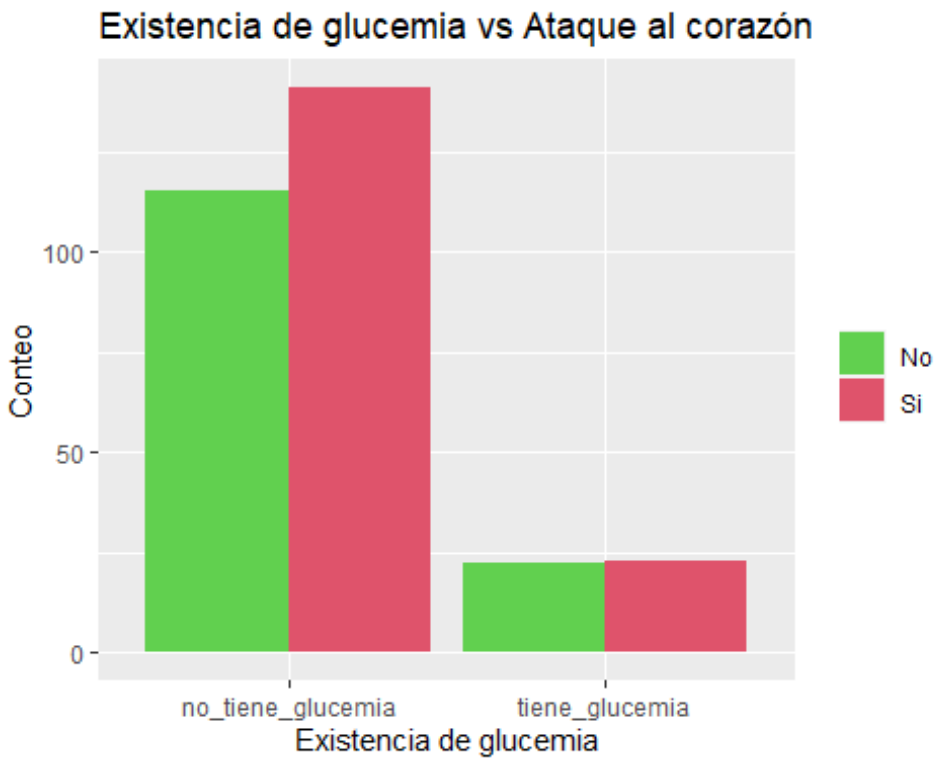
En el caso de la **frecuencia cardíaca** se observa como a medida que aumenta la frecuencia cardíaca la posibilidad de tener un infarto también lo hace.

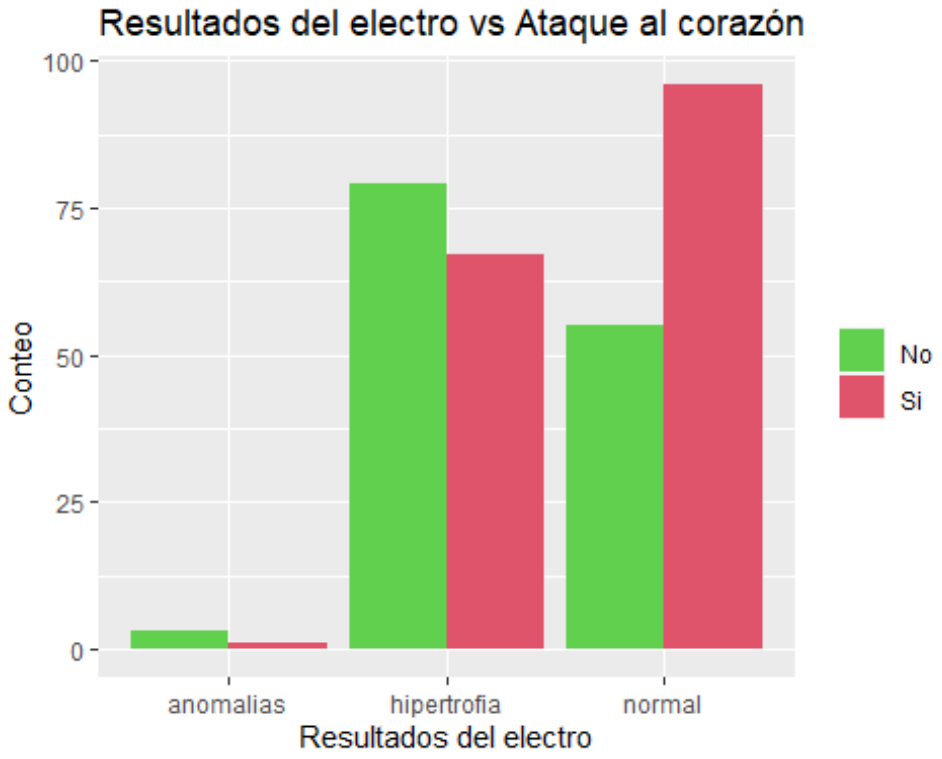
4.3. Variables categóricas.

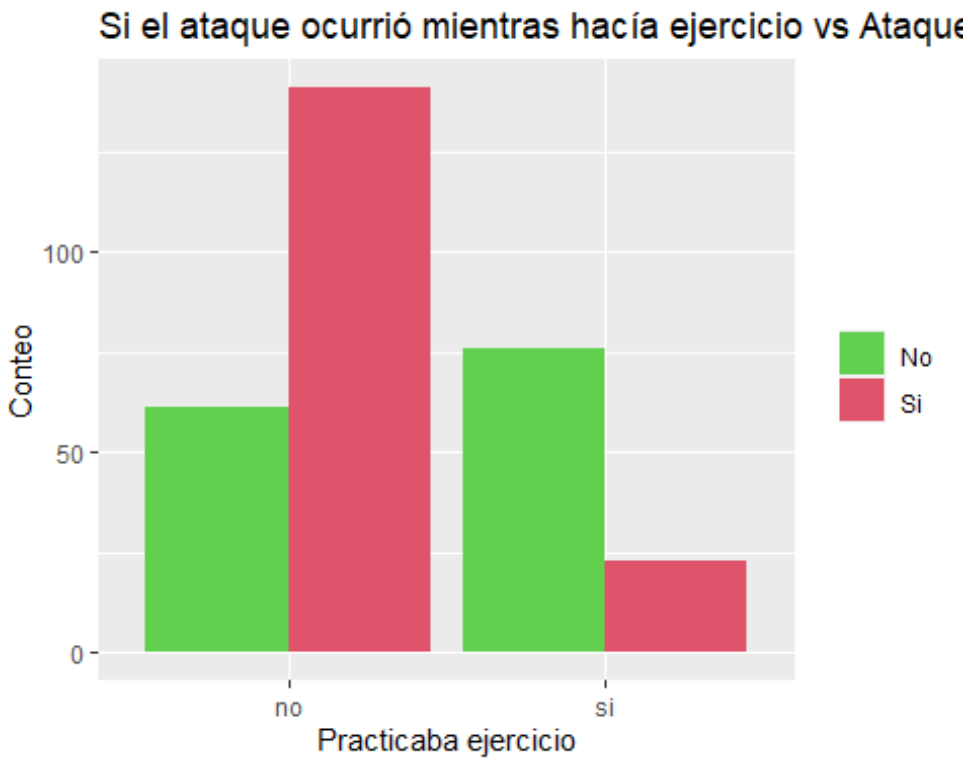
Visualización de las variables categóricas frente a la variable respuesta

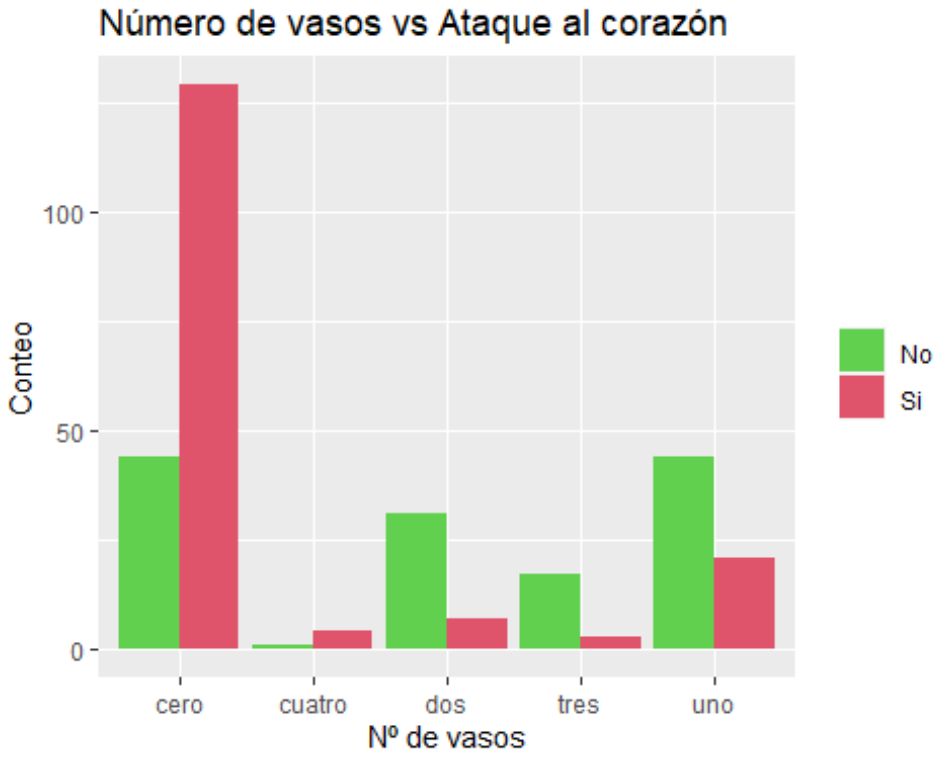



Tipo de dolor vs Ataque al corazón
Conteo
100
75
50
25
0
asintomatico
atipico
no_relacionado
tipico
Tipo de dolor
No
Si


Existencia de glucemia vs Ataque al corazón
Conteo
100
50
0
no_tiene_glucemia
tiene_glucemia
Existencia de glucemia
No
Si


Resultados del electro vs Ataque al corazón
100
75
50
25
0
Conteo
anomalias
hipertrofia
normal
Resultados del electro
No
Si


Si el ataque ocurrió mientras hacía ejercicio vs Ataque
100
50
0
Conteo
no
si
Practicaba ejercicio
No
Si



Conclusiones

Los **hombres son quienes más sufren ataques al corazón frente a las mujeres**, sin embargo ellas son las que más probabilidad tienen en este caso de sufrir un ataque al corazón.

Aquellos pacientes que no presentan **dolor** alguno son los que menos probabilidades tienen de sufrir un infarto, sin embargo esto cambia totalmente cuando se presenta algún dolor atípico, típico y no relacionado en orden de importancia.

El riesgo de ataque cardiaco es ligeramente superior en aquellos pacientes que tienen **glucemia**.

Aquellos pacientes que tuvieron un resultado del **electro** normal tienen más riesgo de sufrir un ataque cardiaco.

Los pacientes que sufrieron un ataque al corazón mientras realizaban **ejercicio** es muy inferior a los que no, así que este hecho no afecta a que se produzca o no un ataque al corazón.

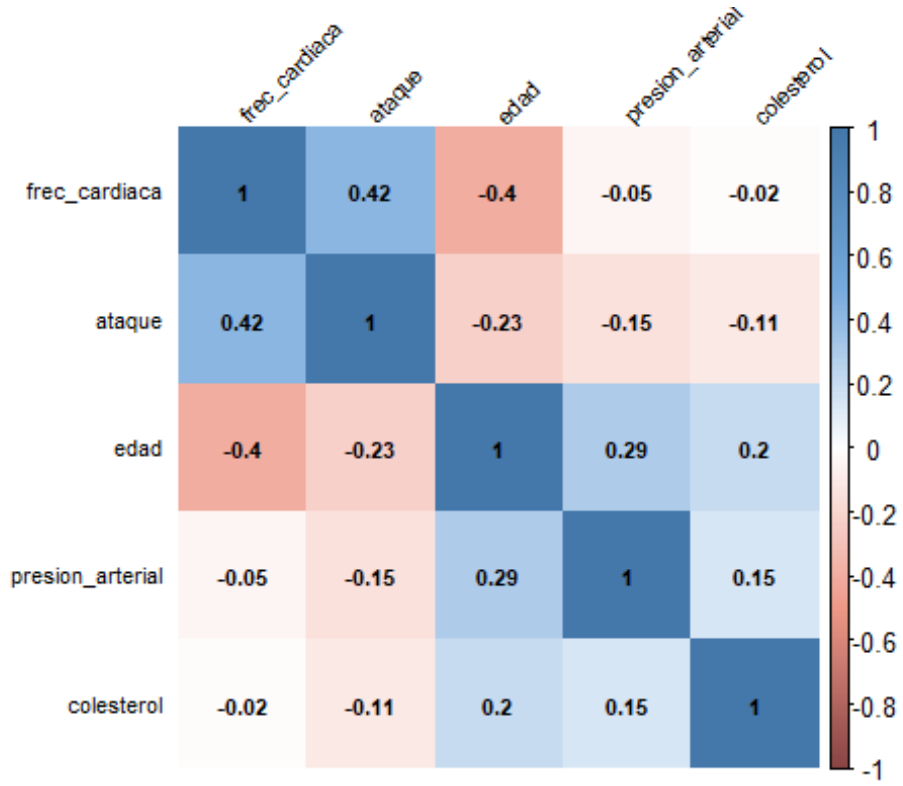
Por último observamos que en la variable **vasos** también hay una gran incidencia de ataque al corazón cuando los vasos tiene un valor igual a cero.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos

4.3.1 Correlaciones y Frecuencias

Comparación entre los datos numéricos con la variable respuesta

Realizamos una correlación de Pearson



Observamos que la mayor correlación con el ataque cardíaco se encuentra en la **frecuencia cardiaca**, siguiendo una correlación positiva, lo que significa que a medida que aumenta la frecuencia cardiaca también se incrementa la posibilidad de sufrir ataques al corazón.

Por contra la **edad** está correlacionada en sentido inverso, es decir, a medida que se incrementa la edad menos posibilidades existen de que se tenga un ataque al corazón.

Comparación entre los datos categóricos con la variable respuesta

Para las variables categóricas aplicamos un chi cuadrado test. En este caso las hipótesis son:

→ H_0: igualdad de frecuencias entre los grupos > 0.05
→ H_1: diferencias significativas < 0.05

Es decir, que cuando el pvalor sea mayor al nivel de significancia estaremos ante grupos de variables cuyas frecuencias son iguales.

Si observamos los resultados de nuestra variable respuesta con respecto al resto de variables, vemos que la **glucemia** la única variable categórica que presenta igualdad de frecuencias con respecto al ataque al corazón.

4.3.2 Regresión lineal para los valores numéricos

Como la variable de salida es una variable en realidad dicotómica usaremos una regresión logística.

```
modelo <- glm(ataque ~ edad + presion_arterial + colesterol + frec_cardiaca,
              heartNumerico, family=binomial)
summary(modelo)
```

```
##
## Call:
## glm(formula = ataque ~ edad + presion_arterial + colesterol +
##     frec_cardiaca, family = binomial, data = heartNumerico)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0460  -1.0340   0.5423   0.9320   1.9866
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.970012   1.674826  -1.773   0.0762 .
## edad             -0.006422   0.016468  -0.390   0.6966
## presion_arterial -0.014601   0.007868  -1.856   0.0635 .
## colesterol       -0.004239   0.002694  -1.573   0.1157
## frec_cardiaca     0.043249   0.007073   6.114  9.7e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 414.85  on 300  degrees of freedom
## Residual deviance: 349.92  on 296  degrees of freedom
## AIC: 359.92
##
## Number of Fisher Scoring iterations: 3
```

A raíz de los resultados ofrecidos por el modelo observamos que la variable explicativa colesterol y edad tienen un pvalue superior al nivel de significancia por tanto no están aportando información al modelo, aunque el colesterol se encuentra cerca del nivel de significancia.

Incluimos en el modelo un dato categórico y estudiamos sus niveles:

```
heart$vasos<-relevel(heart$vasos, ref='cero')
modelo <- glm(ataque ~ frec_cardiaca + vasos, heart, family=binomial)
summary(modelo)
```

```
##
## Call:
## glm(formula = ataque ~ frec_cardiaca + vasos, family = binomial,
##     data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1673  -0.7902   0.4481   0.7266   2.3315
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.83779    1.10521  -4.377 1.20e-05 ***
## frec_cardiaca  0.03893    0.00731   5.326 1.00e-07 ***
## vasoscuatro    0.06681    1.16176   0.058 0.954144
## vasosdos      -2.56183    0.47615  -5.380 7.44e-08 ***
## vasostres     -2.30178    0.67442  -3.413 0.000643 ***
## vasosuno      -1.53898    0.34005  -4.526 6.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 414.85  on 300  degrees of freedom
## Residual deviance: 302.90  on 295  degrees of freedom
## AIC: 314.9
##
## Number of Fisher Scoring iterations: 4
```

Observamos que hay significancia en casi todos sus niveles. Tiene el AIC más bajo de todos los modelos que se han probado y que se encuentran en el código ya que en este informe solo hemos incluido el mejor modelo de regresión logista, formado por la frecuencia cardiaca como variable numérica y el número de vasos como variable categórica.

4.3.3 Modelo de aprendizaje supervisado

En este caso queremos clasificar en base a los datos si se produce un ataque al corazón o no, por tanto estamos ante una variable respuesta dicotómica. Entre las variables explicativas tenemos variables de tipo numérico y categórico, así que vamos a utilizar un Random Forest.

```
heart$presion_arterial <- scale(heart$presion_arterial)
heart$frec_cardiaca <- scale(heart$frec_cardiaca)
heart$colesterol <- scale(heart$colesterol)
heart$edad <- scale(heart$edad)
```

Creamos un conjunto de entrenamiento y otro de test

```
# fijamos la semilla
set.seed(123)
# Se crean los índices de las observaciones de entrenamiento
ind <- createDataPartition(y=heart$ataque,
                                  p=0.8, list = FALSE, times = 1)

train <- heart[ind, ]
test <- heart[-ind, ]
```

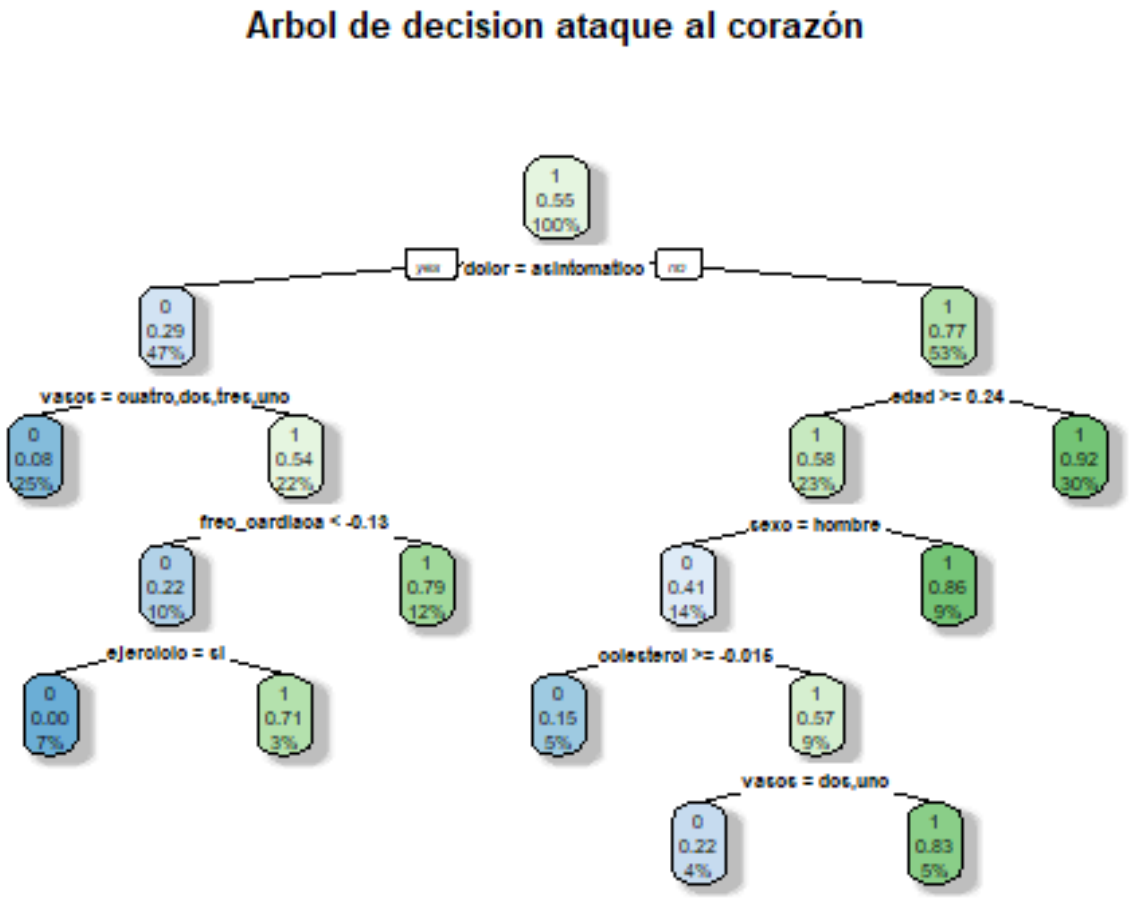
Verificamos que la distribución de la variable respuesta ataque es similar en el conjunto de entrenamiento y en el de test.

```
#datos
prop.table(table(heart$ataque))

##
##         0         1
## 0.4551495 0.5448505

#entrenamiento
prop.table(table(train$ataque))

##
##         0         1
## 0.4522822 0.5477178

#test
prop.table(table(test$ataque))

##
##         0         1
## 0.4666667 0.5333333
```

Creamos el árbol de decisión:

```
set.seed(1234)
rf <- rpart(train$ataque ~ .,
                      data = train[-11],
                      method = "class", cp = .01)
```

```
rpart.plot(rf, fallen.leaves = FALSE,
                   main = "Arbol de decision ataque al corazón",
                   shadow.col = "gray")
```



Generamos las predicciones sobre el conjunto de test:

```
prediccion <- predict(rf, newdata = test, type = "class")

matrizConfusion <- confusionMatrix(factor(prediccion), factor(test$ataque), posi
tive = "1")
matrizConfusion
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 20  3
##          1  8 29
##
##                Accuracy : 0.8167
##                  95% CI : (0.6956, 0.9048)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 4.344e-06
##
##                   Kappa : 0.6275
```

```
##
##  Mcnemar's Test P-Value : 0.2278
##
##             Sensitivity : 0.9062
##             Specificity : 0.7143
##          Pos Pred Value : 0.7838
##          Neg Pred Value : 0.8696
##              Prevalence : 0.5333
##          Detection Rate : 0.4833
##    Detection Prevalence : 0.6167
##       Balanced Accuracy : 0.8103
##
##        'Positive' Class : 1
##
```

5. Resolución del problema

Creemos que los análisis estadísticos realizados, visualización de variables, análisis de correlaciones y análisis de frecuencias, regresión lineal logística y árbol de decisión Random Forest han permitido llegar a ciertas conclusiones.

Hemos visto que la creencia de que a mayor edad más riesgo hay de tener un ataque al corazón no es cierta, al igual que sucede con el tema del colesterol, que hemos visto que no tenía tanto peso como podríamos creer, algo similar sucede con el tipo de dolor en el que aunque los tipos de dolor son relevantes el mayor ha resultado ser el asintomático. El número de vasos y la frecuencia cardíaca han resultado ser factores relevantes en el estudio y en la aplicación de un algoritmo de decisión como es el Random Forest se ha obtenido una precisión de 0.81 con buenos datos de clasificación para las dos clases.

6. Código

El código se ha realizado en R y se ha incluido el archivo en el apartado correspondiente en el repositorio GitHub.

7. Video

https://drive.google.com/file/d/19r8OtFubYenZlq5IX4yArPJkFssBmqSH/view?usp=sharing

8. Contribuciones

Contribuciones	Firma
Investigación previa	PLC
Redacción de las respuestas	PLC
Desarrollo del código	PLC
Participación en el vídeo	PLC