Heart Attack Analysis & Prediction Dataset

Autores: Alba Sanz Horcajo y Carlos Santamaría de las Heras

Enero 2023

Contents

1. Detalles de la actividad

1.1. Descripción

Esta prueba de evaluación continua cubre los Módulos 3 (Clasificación: árboles de decisión) y el Módulo 8 (Evaluación de modelos) del programa de la asignatura.

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. 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.

1.3. 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.

2. Resolución

2.1 Descripción del dataset

Para la realiación de esta práctica se ha elegido el dataset proporcionado en la asignatura al ser un data set muy rico en variables y con grandes posibilidades de análisis. A su vez, **se ha optado por** su realización en **R** ya que **la primera práctica la realizamos en Python** y de esta manera ampliamos nuestros conocimientos con otro lenguaje de programación diferente.

El conjunto de datos objeto de análisis se ha obtenido a partir de este enlace en Kaggle y cuyo nombre es: **"Heart Attack Analysis & Prediction dataset".**

Instalamos y cargamos las librerías necesarias

```
# https://cran.r-project.org/web/packages/ggplot2/index.html
if(!require(ggplot2)){
    install.packages('ggplot2', repos='http://cran.us.r-project.org')
    library(ggplot2)
}
# https://cran.r-project.org/web/packages/dplyr/index.html
```

```
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
library(scales)
if (!require('GGally')) install.packages("GGally")
library('GGally')
```

2.1.1 Cargamos y mostramos el fichero de datos.

```
path = 'heart.csv'
datos_brutos <- read.csv(path, row.names=NULL)
```

2.1.2 Exploración del conjunto de datos

Verificamos la estructura del juego de datos principal. Vemos el número de columnas que tenemos y ejemplos de los contenidos de las filas.

```
str(datos_brutos)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ trtbps  : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exng    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slp     : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ caa     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thall   : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
dim(datos_brutos)
```

```
## [1] 303  14
```

```
filas <- dim(datos_brutos)[1]
print(paste("El número de filas o registros es:", filas))
```

```
## [1] "El número de filas o registros es: 303"
```

```
variables <- dim(datos_brutos)[2]
print(paste("El número de variables o atributos es:", variables))
```

```
## [1] "El número de variables o atributos es: 14"
```

Vemos que tenemos **14** variables y **303** registros.

Revisamos la descripción de las variables contenidas en el fichero y si los tipos de variables se corresponden con las que hemos cargado. Las organizamos lógicamente para darles sentido y construimos un pequeño diccionario de datos utilizando la documentación auxiliar.

Las variables consideradas son:

- **age** : edad del paciente
- **sex** : género del paciente (según el foro)
  - 0: femenino
  - 1: masculino
- **exang**: angina inducida por el ejercicio
  - 0: no
  - 1: sí
- **caa**: número de vasos principales (0-3)
- **cp**: tipo de dolor en el pecho:
  - 0: angina típica
  - 1: angina atípica
  - 2: dolor no anginal
  - 3: asintomático
- **trtbps** : presión arterial en reposo (en mm Hg)
- **chol** : colesterol en mg/dl obtenido a través del sensor de IMC
- **fbs** : (glucemia en ayunas $>$ 120 mg/dl)
  - 0: falso
  - 1: verdadero
- **oldpeak**: depresión del segmento ST inducida por el ejercicio en relación con el reposo
- **slp**: pendiente del segmento ST de ejercicio máximo:
  - 0: pendiente descendente
  - 1: plana
  - 2: pendiente ascendente
- **thall**: talasemia:
  - 0: nulo
  - 1: defecto fijo
  - 2: normales
  - 3: defecto reversible
- **rest_ecg** : resultados electrocardiográficos en reposo
  - 0: normal

  - 1: con anormalidades en la onda ST-T (inversiones de onda T y/o elevación o depresión de ST > 0,05 mV)
  - 2: con hipertrofia ventricular izquierda probable o definitiva siguiendo los criterios de Estes
- **thalach** : frecuencia cardíaca máxima alcanzada
- **output**: diagnóstico de enfermedad cardíaca (estado de enfermedad angiográfico)
  - 0: menor posibilidad de ataque al corazón (<50% estrechamiento del diámetro)
  - 1: mayor posibilidad de ataque al corazón (>50% estrechamiento del diámetro)

A través del dataset comentado se pretende determinar cuales son las variables de entre todas las que disponemos que tienen una mayor influencia sobre la posibilidad de tener un ataque al corazón. Un segundo objetivo sería identificar mediante contrastes de hipótesis diferencias que puedan inferirse a la población. Por último, otro objetivo sería crear modelos de clasificación que nos permita diferenciar a aquellas personas que por sus características es más probable que tengan un ataque al corazón.

2.2 Importancia y objetivos de los análisis

A partir de este conjunto de datos se plantea la problemática de determinar qué variables influyen más en los ataques al corazón o si tienen las mujeres menos ataques al corazón que los hombres. Además, se podrá proceder a crear modelos de regresión que permitan responder a estas preguntas y contrastes de hipótesis que ayuden a identificar propiedades interesantes en las muestras que puedan ser inferidas con respecto a los pacientes.

Las preguntas objetivo del análisis son:

- ¿Cuáles son las variables cuantitativas que tienen una mayor influencia en los ataques al corazón?
- ¿Tienen las mujeres más ataques al corazón que los hombres?
- ¿Podemos clasificar a los pacientes en alta y baja probabilidad de ataque cardiaco?

Este análisis puede tener gran relevancia para los profesionales de la salud, más concretamente para los especialistas en cardiología. Sabiendo cuales son las variables que tiene mayor influencia en los ataques al corazón, se podría mejorar la prevención y tratamiento de esta patología. Si se integrase el modelo de clasificación en su software de trabajo y pudiesen análizar toda su base de datos de pacientes extrayendo un listado de los pacientes considerados de riesgo, podrían llevarse a cabo campañas de prevención enfocada a esta población.

2.3 Preprocesamiento y gestión de características

2.3.1 Renombramos las variables

Procedemos a renombrar los campos para una mejor comprensión de los mismos.

```
datos_renombrados <- rename(datos_brutos, edad = age, sexo = sex, dolor_pecho = cp, pa_reposo  = trtbps
str(datos_renombrados)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ edad                          : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sexo                          : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ dolor_pecho                   : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ pa_reposo                     : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ colesterol                    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ glucemia_ayunas_mayor_120m_dl: int  1 0 0 0 0 0 0 0 1 0 ...
##  $ ecg_reposo                    : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ frec_cardiaca_max             : int  150 187 172 178 163 148 153 173 162 174 ...
##  $ angina_por_ejercicio          : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ depresionST_ejercicioVSreposo: num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ segmentoST_ejercicioMAX       : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ num_vasos_principales         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ talasemia                     : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ salida                        : int  1 1 1 1 1 1 1 1 1 1 ...
```

Las variables quedan finalmente de la siguiente forma:

- **edad**: edad del paciente
- **sexo** : género del paciente (según el foro)
  - 0: femenino
  - 1: masculino
- **angina_por_ejercicio**: angina inducida por el ejercicio
  - 0: no
  - 1: sí
- **num_vasos_principales**: número de vasos principales (0-3)
- **dolor_pecho**: tipo de dolor en el pecho:
  - 0: asintomático
  - 1: angina típica
  - 2: angina atípica
  - 3: dolor no anginal
- **pa_reposo**: presión arterial en reposo (en mm Hg)
- **colesterol**: colesterol en mg/dl obtenido a través del sensor de IMC
- **glucemia_ayunas_mayor_120mg/dl**: (glucemia en ayunas > 120 mg/dl)
  - 0: falso
  - 1: verdadero
- **depresionST_ejercicioVSreposo**: depresión del segmento ST inducida por el ejercicio en relación con el reposo
- **segmentoST_ejercicioMAX**: pendiente del segmento ST de ejercicio máximo:
  - 0: pendiente descendente
  - 1: plana

  - 2: pendiente ascendente

- **talasemia**: talasemia:
  - 0: nulo
  - 1: defecto fijo
  - 2: normales
  - 3: defecto reversible

- **ecg_reposo** : resultados electrocardiográficos en reposo
  - 0: normal
  - 1: con anormalidades en la onda ST-T (inversiones de onda T y/o elevación o depresión de ST > 0,05 mV)
  - 2: con hipertrofia ventricular izquierda probable o definitiva siguiendo los criterios de Estes

- **frec_cardiaca_max** : frecuencia cardíaca máxima alcanzada

- **salida**: diagnóstico de enfermedad cardíaca (estado de enfermedad angiográfico)
  - 0: menor posibilidad de ataque al corazón (<50% estrechamiento del diámetro)
  - 1: mayor posibilidad de ataque al corazón (>50% estrechamiento del diámetro)

2.3.2 Mostramos las estadisticas básicas

```
summary(datos_renombrados)
```

```
##       edad            sexo         dolor_pecho      pa_reposo
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0
##    colesterol    glucemia_ayunas_mayor_120m_dl   ecg_reposo
##  Min.   :126.0   Min.   :0.0000                Min.   :0.0000
##  1st Qu.:211.0   1st Qu.:0.0000                1st Qu.:0.0000
##  Median :240.0   Median :0.0000                Median :1.0000
##  Mean   :246.3   Mean   :0.1485                Mean   :0.5281
##  3rd Qu.:274.5   3rd Qu.:0.0000                3rd Qu.:1.0000
##  Max.   :564.0   Max.   :1.0000                Max.   :2.0000
##  frec_cardiaca_max angina_por_ejercicio depresionST_ejercicioVSreposo
##  Min.   : 71.0     Min.   :0.0000       Min.   :0.00
##  1st Qu.:133.5     1st Qu.:0.0000       1st Qu.:0.00
##  Median :153.0     Median :0.0000       Median :0.80
##  Mean   :149.6     Mean   :0.3267       Mean   :1.04
##  3rd Qu.:166.0     3rd Qu.:1.0000       3rd Qu.:1.60
##  Max.   :202.0     Max.   :1.0000       Max.   :6.20
##  segmentoST_ejercicioMAX num_vasos_principales   talasemia         salida
##  Min.   :0.000           Min.   :0.0000        Min.   :0.000   Min.   :0.0000
##  1st Qu.:1.000           1st Qu.:0.0000        1st Qu.:2.000   1st Qu.:0.0000
```

```
##  Median :1.000           Median :0.0000        Median :2.000   Median :1.0000
##  Mean   :1.399           Mean   :0.7294        Mean   :2.314   Mean   :0.5446
##  3rd Qu.:2.000           3rd Qu.:1.0000        3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :2.000           Max.   :4.0000        Max.   :3.000   Max.   :1.0000
```

Mostramos las estadísticas básicas:

```
summary(datos_renombrados)
```

```
##       edad            sexo         dolor_pecho      pa_reposo
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0
##    colesterol    glucemia_ayunas_mayor_120m_dl   ecg_reposo
##  Min.   :126.0   Min.   :0.0000                Min.   :0.0000
##  1st Qu.:211.0   1st Qu.:0.0000                1st Qu.:0.0000
##  Median :240.0   Median :0.0000                Median :1.0000
##  Mean   :246.3   Mean   :0.1485                Mean   :0.5281
##  3rd Qu.:274.5   3rd Qu.:0.0000                3rd Qu.:1.0000
##  Max.   :564.0   Max.   :1.0000                Max.   :2.0000
##  frec_cardiaca_max angina_por_ejercicio depresionST_ejercicioVSreposo
##  Min.   : 71.0     Min.   :0.0000       Min.   :0.00
##  1st Qu.:133.5     1st Qu.:0.0000       1st Qu.:0.00
##  Median :153.0     Median :0.0000       Median :0.80
##  Mean   :149.6     Mean   :0.3267       Mean   :1.04
##  3rd Qu.:166.0     3rd Qu.:1.0000       3rd Qu.:1.60
##  Max.   :202.0     Max.   :1.0000       Max.   :6.20
##  segmentoST_ejercicioMAX num_vasos_principales   talasemia         salida
##  Min.   :0.000           Min.   :0.0000        Min.   :0.000   Min.   :0.0000
##  1st Qu.:1.000           1st Qu.:0.0000        1st Qu.:2.000   1st Qu.:0.0000
##  Median :1.000           Median :0.0000        Median :2.000   Median :1.0000
##  Mean   :1.399           Mean   :0.7294        Mean   :2.314   Mean   :0.5446
##  3rd Qu.:2.000           3rd Qu.:1.0000        3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :2.000           Max.   :4.0000        Max.   :3.000   Max.   :1.0000
```

De las cuales podemos destacar lo siguiente:

- La **edad mínima** de los pacientes es de **29 años** y la **máxima** de **77 años**.
- Tenemos información de **96 personas de género femenino y 207 de género masculino**.
- Respecto al dolor de pecho, lo más común es ser **asintomático** lo que representa el **47,19%** del total de casos, seguido de **angina atípica** con un **28,71%.**
- La **media** de **presión arterial en reposo** se sitúa en los **131.6mm Hg.**
- El **colesterol máximo** registrado es de **564mg/dl** y la **mínimo 126mg/dl.**
- Un total de **45** personas han registrado **glucemia en ayunas > 120 mg/dl**, mientras que 258 no.
- La prueba de **electrocardiografo en reposo** ha resultado **normal** en **147 personas** mientras que en 156 se han encontrado resultados anormales.

- La **frecuencia cardíaca máxima** alcanzada es de **202 lat/min**, la **media** de **150 lat/min** y la **mínima** de **71lat/min.**
- Se ha producido una **angina inducida por el ejercicio** en **99 personas** (32,67%).
- La depresión del segmento ST inducida por el ejercicio en relación con el reposo es de 1,04.
- La pendiente del segmento ST de ejercicio máximo es ascendente en 142 personas (46,87%), plana en 140 (46,20%) y descendente en 21 (6,93%).
- La media de vasos principales es de 1 (0,73).
- Un total de 166 personas no tienen trastorno de la sangre hereditario (talasemia normal)
- Un total de 165 personas tienen una mayor posibilidad de ataque al corazón (54,46%)

2.3.3 Incoherencias

- Se observa que en la página Kaggle de descarga del dataset se indica que el número de vasos principales (num_vasos_principales) puede ser entre 0 y 3, sin embargo se detectan registros con un valor de 4.

2.4 Limpieza

Tras haber realizado en el apartado anterior una primera medida de acondicionado de los datos (pasar las variables cualitativas de int a factor) para poder realizar un mejor análisis exploratorio de los datos, procedemos a terminar de realizar la limpieza y acondicionado de datos para poder ser usado en procesos de modelado.

2.4.1 Gestión de valores nulos/vacíos

El siguiente paso será la limpieza de datos, para lo cual primero comprobamos si hay valores vacíos o nulos en el conjunto de datos

```
print('NA')
```

```
## [1] "NA"
```

```
colSums(is.na(datos_renombrados))
```

```
##                         edad                         sexo
##                            0                            0
##                  dolor_pecho                    pa_reposo
##                            0                            0
##                   colesterol glucemia_ayunas_mayor_120m_dl
##                            0                            0
##                   ecg_reposo             frec_cardiaca_max
##                            0                            0
##         angina_por_ejercicio depresionST_ejercicioVSreposo
##                            0                            0
##      segmentoST_ejercicioMAX         num_vasos_principales
##                            0                            0
##                     talasemia                       salida
##                            0                            0
```

```
print('Blancos')
```

```
## [1] "Blancos"
```

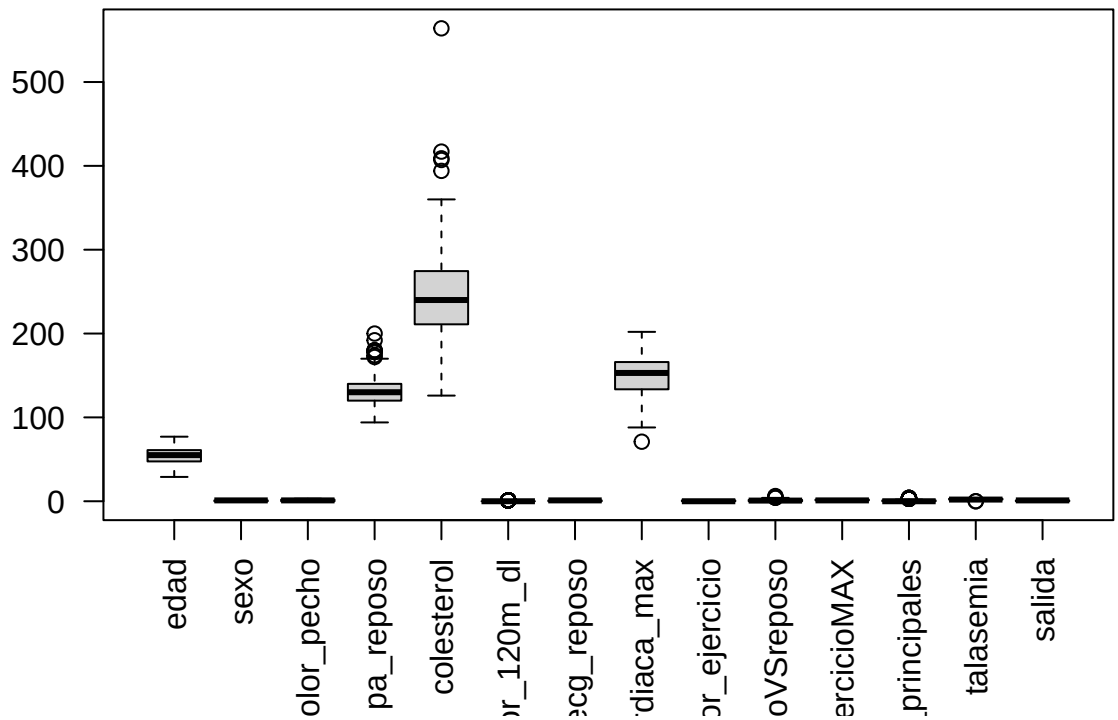
```
colSums(datos_renombrados=="")
```

```
##                          edad                          sexo
##                             0                             0
##                   dolor_pecho                     pa_reposo
##                             0                             0
##                    colesterol glucemia_ayunas_mayor_120m_dl
##                             0                             0
##                    ecg_reposo             frec_cardiaca_max
##                             0                             0
##          angina_por_ejercicio depresionST_ejercicioVSreposo
##                             0                             0
##       segmentoST_ejercicioMAX         num_vasos_principales
##                             0                             0
##                     talasemia                        salida
##                             0                             0
```

Vemos que no hay valores nulos ni vacíos en los datos, por lo cual no tendremos que realizar ninguna acción de eliminar o modificar registros.

2.4.2 Mostramos valores anómalos/outliers

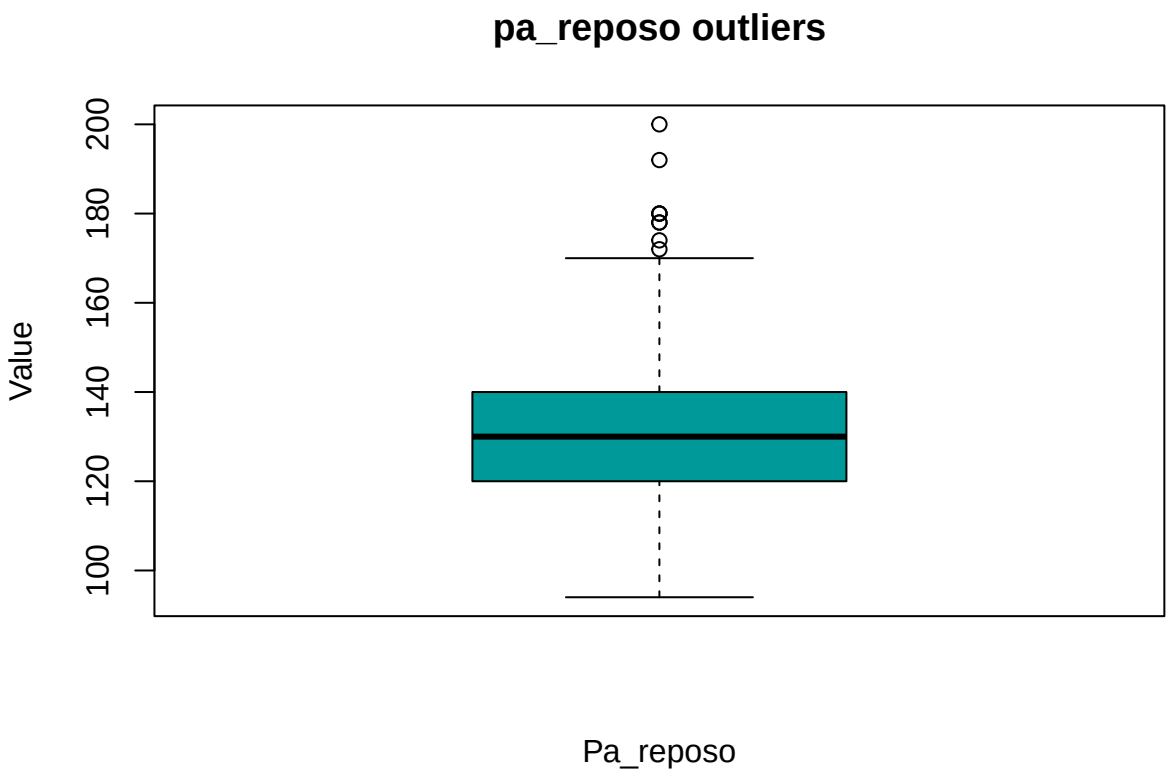
```
boxplot(datos_renombrados, las = 2)
```



Como podemos ver, encontramos valores anómalos, sobretodo en la variable pa_reposo y colesterol.

- Mostramos los valores anómalos de pa_reposo:

```
pa_reposo_outliers <- boxplot(datos_renombrados$pa_reposo, col = "#009999", main = "pa_reposo outliers"
```




```
pa_reposo_outliers$out
```

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

Mostramos los valores anómalos de **colesterol**:

```
colesterol_outliers <- boxplot(datos_renombrados$colesterol, col = "#009999", main = "Colesterol outlie
```


```
colesterol_outliers$out
```

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

No se van a eliminar los valores anómalos para preservar la mayor variabilidad posible en los datos. Sin embargo, estos se tendrán en cuenta para intentar disminuir la influencia de estos datos en futuros análisis.

2.5 Análisis de los datos

2.5.1 Discretizamos las variables numéricas

Las variables se están considerando como variables continuas cuando algunas de ellas son categóricas, por este motivo, antes de comenzar con el análisis, vamos a crear el dataframe **datos_categoricos** en el que tengamos correctamente representado las variables continuas y categóricas.

A su vez, crearemos el dataframe **datos_discretizados** discretizando las variables numéricas: *edad, pa_reposo y colesterol, frec_cardiaca_max*:

```
datos_categoricos <- datos_renombrados

datos_categoricos$sexo = factor(datos_categoricos$sexo, levels = c(0:1), labels = c("femenino", "masculi

datos_categoricos$angina_por_ejercicio = factor(datos_categoricos$angina_por_ejercicio, levels = c(0:1)
```

```r
datos_categoricos$dolor_pecho = factor(datos_categoricos$dolor_pecho, levels = c(0:3), labels = c("asint

datos_categoricos$glucemia_ayunas_mayor_120m_dl = factor(datos_categoricos$glucemia_ayunas_mayor_120m_dl

datos_categoricos$segmentoST_ejercicioMAX = factor(datos_categoricos$segmentoST_ejercicioMAX, levels = c

datos_categoricos$talasemia = factor(datos_categoricos$talasemia, levels = c(0:3), labels = c("nulo", "c

datos_categoricos$ecg_reposo = factor(datos_categoricos$ecg_reposo, levels = c(0:2), labels = c("normal

datos_categoricos$salida = factor(datos_categoricos$salida, levels = c(0:1), labels = c("menor posibilid
```

```r
datos_discretizados <- datos_categoricos
datos_discretizados$edad <- cut(datos_discretizados$edad, breaks = c(25,47,62,86), labels = c("jovenes"
datos_discretizados$pa_reposo <- cut(datos_discretizados$pa_reposo, breaks = c(50,120,139,159,179,240),
datos_discretizados$colesterol <- cut(datos_discretizados$colesterol, breaks = c(0,170,199,240,600), lab
datos_discretizados$frec_cardiaca_max <- cut(datos_discretizados$frec_cardiaca_max, breaks = c(0,110,160

head(datos_discretizados)
```

```
##            edad      sexo      dolor_pecho         pa_reposo
## 1 edad avanzada masculino dolor no anginal hipertension leve
## 2       jovenes masculino   angina atípica       normal-alta
## 3       jovenes  femenino    angina típica       normal-alta
## 4    media edad masculino    angina típica            normal
## 5    media edad  femenino     asintomático            normal
## 6    media edad masculino     asintomático hipertension leve
##                   colesterol glucemia_ayunas_mayor_120m_dl
## 1                       alto                            si
## 2 hipercolesterolemia severa                            no
## 3                       alto                            no
## 4                       alto                            no
## 5 hipercolesterolemia severa                            no
## 6             valores limite                            no
##                      ecg_reposo frec_cardiaca_max angina_por_ejercicio
## 1                        normal           normal                   no
## 2 anormalidades en la onda ST-T       taquicardia                  no
## 3                        normal       taquicardia                  no
## 4 anormalidades en la onda ST-T       taquicardia                  no
## 5 anormalidades en la onda ST-T       taquicardia                  si
## 6 anormalidades en la onda ST-T            normal                  no
##   depresionST_ejercicioVSreposo segmentoST_ejercicioMAX num_vasos_principales
## 1                           2.3   pendiente descendente                     0
## 2                           3.5   pendiente descendente                     0
## 3                           1.4    pendiente ascendente                     0
## 4                           0.8    pendiente ascendente                     0
## 5                           0.6    pendiente ascendente                     0
## 6                           0.4                   plana                     0
##     talasemia                                 salida
## 1 defecto fijo mayor posibilidad de ataque al corazón
## 2     normales mayor posibilidad de ataque al corazón
## 3     normales mayor posibilidad de ataque al corazón
## 4     normales mayor posibilidad de ataque al corazón
```

```
## 5     normales mayor posibilidad de ataque al corazón
## 6 defecto fijo mayor posibilidad de ataque al corazón
```

```
str(datos_discretizados)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ edad                          : Factor w/ 3 levels "jovenes","media edad",..: 3 1 1 2 2 2 2 1 2 2
##  $ sexo                          : Factor w/ 2 levels "femenino","masculino": 2 2 1 2 1 2 1 2 2 2 ...
##  $ dolor_pecho                   : Factor w/ 4 levels "asintomático",..: 4 3 2 2 1 1 2 2 3 3 ...
##  $ pa_reposo                     : Factor w/ 5 levels "normal","normal-alta",..: 3 2 2 1 1 3 3 1 4 3
##  $ colesterol                    : Factor w/ 4 levels "bueno","valores limite",..: 3 4 3 3 4 2 4 4 2 3
##  $ glucemia_ayunas_mayor_120m_dl: Factor w/ 2 levels "no","si": 2 1 1 1 1 1 1 1 2 1 ...
##  $ ecg_reposo                    : Factor w/ 3 levels "normal","anormalidades en la onda ST-T",..: 1 1
##  $ frec_cardiaca_max             : Factor w/ 3 levels "bradicardia",..: 2 3 3 3 3 2 2 3 3 3 ...
##  $ angina_por_ejercicio          : Factor w/ 2 levels "no","si": 1 1 1 1 2 1 1 1 1 1 ...
##  $ depresionST_ejercicioVSreposo: num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ segmentoST_ejercicioMAX       : Factor w/ 3 levels "pendiente descendente",..: 1 1 3 3 3 2 2 3 3 3
##  $ num_vasos_principales         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ talasemia                     : Factor w/ 4 levels "nulo","defecto fijo",..: 2 3 3 3 3 2 3 4 4 3 .
##  $ salida                        : Factor w/ 2 levels "menor posibilidad de ataque al corazón",..: 2 1
```

2.5.2 Normalizamos las variables numéricas

Creamos el dataframe **datos_normalizados** normalizando las variables numéricas: *edad, pa_reposo y colesterol, frec_cardiaca_max:*

```
library(scales)
datos_normalizados <- datos_renombrados
datos_normalizados$sexo = factor(datos_normalizados$sexo, levels = c(0:1), labels = c("femenino", "mascu
datos_normalizados$angina_por_ejercicio = factor(datos_normalizados$angina_por_ejercicio, levels = c(0:
datos_normalizados$dolor_pecho = factor(datos_normalizados$dolor_pecho, levels = c(0:3), labels = c("as
datos_normalizados$glucemia_ayunas_mayor_120m_dl = factor(datos_normalizados$glucemia_ayunas_mayor_120m_
datos_normalizados$segmentoST_ejercicioMAX = factor(datos_normalizados$segmentoST_ejercicioMAX, levels =
datos_normalizados$talasemia = factor(datos_normalizados$talasemia, levels = c(0:3), labels = c("nulo",
datos_normalizados$ecg_reposo = factor(datos_normalizados$ecg_reposo, levels = c(0:2), labels = c("norma
datos_normalizados$edad <- rescale(datos_normalizados$edad)
datos_normalizados$pa_reposo <- rescale(datos_normalizados$pa_reposo)
datos_normalizados$colesterol <- rescale(datos_normalizados$colesterol)
datos_normalizados$frec_cardiaca_max <- rescale(datos_normalizados$frec_cardiaca_max)

head(datos_normalizados)
```

```
##        edad      sexo      dolor_pecho pa_reposo colesterol
## 1 0.7083333 masculino dolor no anginal 0.4811321  0.2442922
## 2 0.1666667 masculino   angina atípica 0.3396226  0.2831050
## 3 0.2500000  femenino    angina típica 0.3396226  0.1780822
## 4 0.5625000 masculino    angina típica 0.2452830  0.2511416
## 5 0.5833333  femenino     asintomático 0.2452830  0.5205479
## 6 0.5833333 masculino     asintomático 0.4339623  0.1506849
##   glucemia_ayunas_mayor_120m_dl                    ecg_reposo frec_cardiaca_max
## 1                            si                        normal         0.6030534
## 2                            no anormalidades en la onda ST-T         0.8854962
```

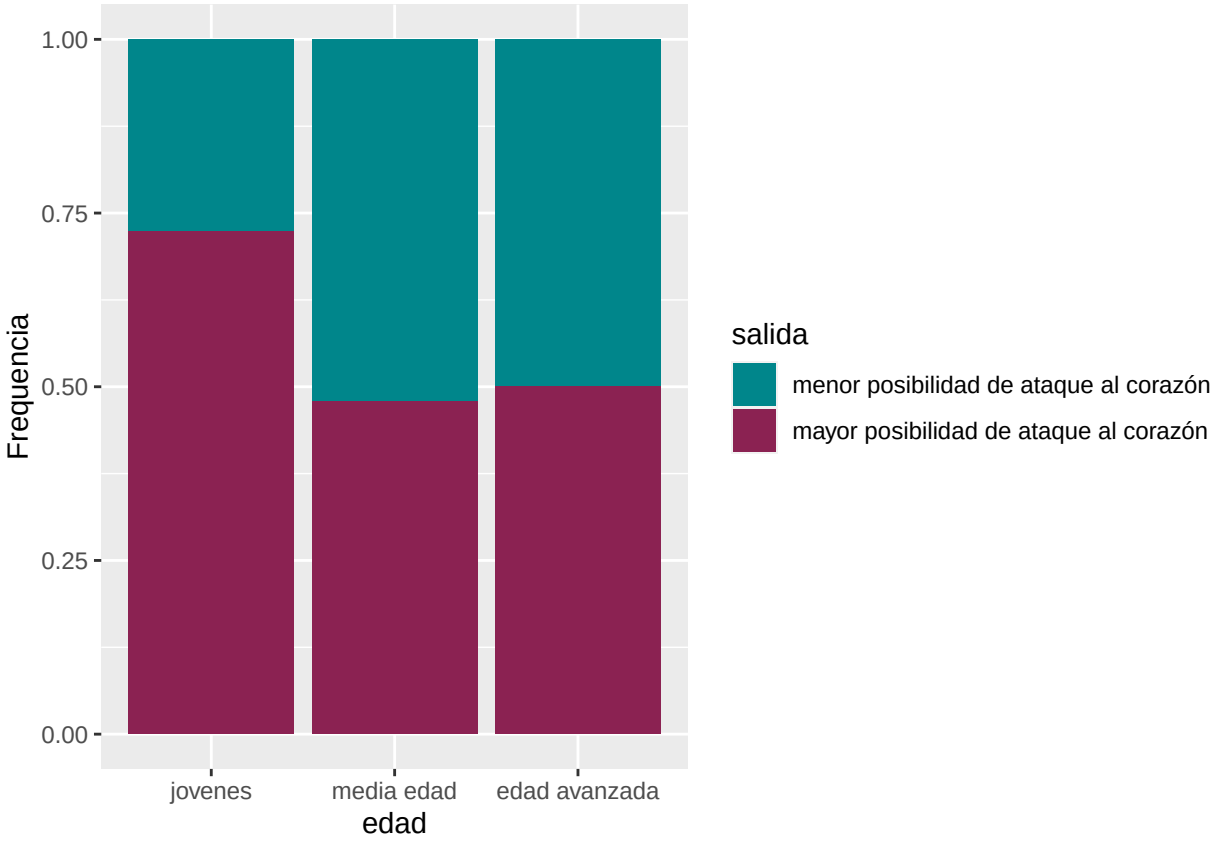
```
## 3                           no                       normal         0.7709924
## 4                           no anormalidades en la onda ST-T         0.8167939
## 5                           no anormalidades en la onda ST-T         0.7022901
## 6                           no anormalidades en la onda ST-T         0.5877863
##   angina_por_ejercicio depresionST_ejercicioVSreposo segmentoST_ejercicioMAX
## 1                   no                          2.3   pendiente descendente
## 2                   no                          3.5   pendiente descendente
## 3                   no                          1.4    pendiente ascendente
## 4                   no                          0.8    pendiente ascendente
## 5                   si                          0.6    pendiente ascendente
## 6                   no                          0.4                   plana
##   num_vasos_principales    talasemia salida
## 1                     0 defecto fijo      1
## 2                     0     normales      1
## 3                     0     normales      1
## 4                     0     normales      1
## 5                     0     normales      1
## 6                     0 defecto fijo      1
```

2.5.3 Análisis de la relación de las variables con target

Vamos a analizar la relación entre las diferentes variables y target, que en este caso es la variable **salida**.

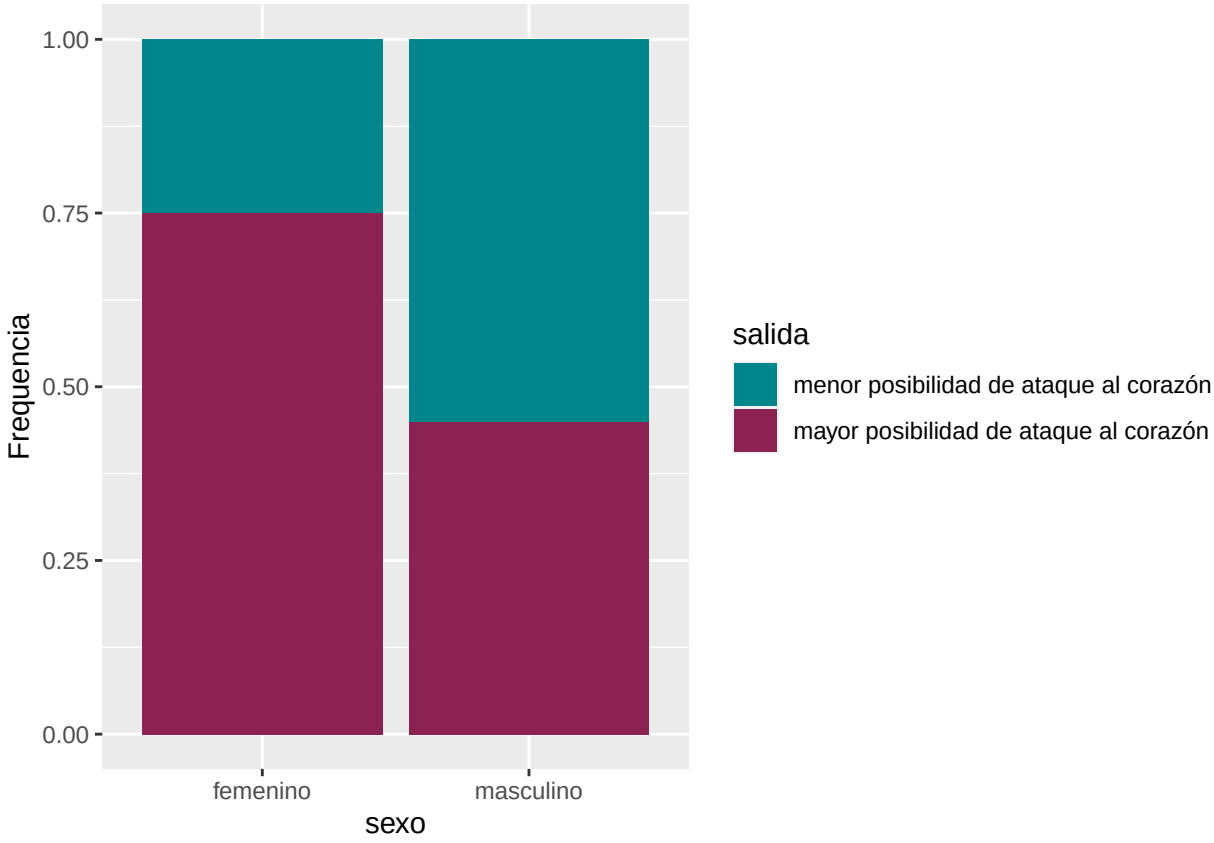
- **edad vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=edad,fill=salida))+geom_bar(position="fill")+ylab("Frequencia
```



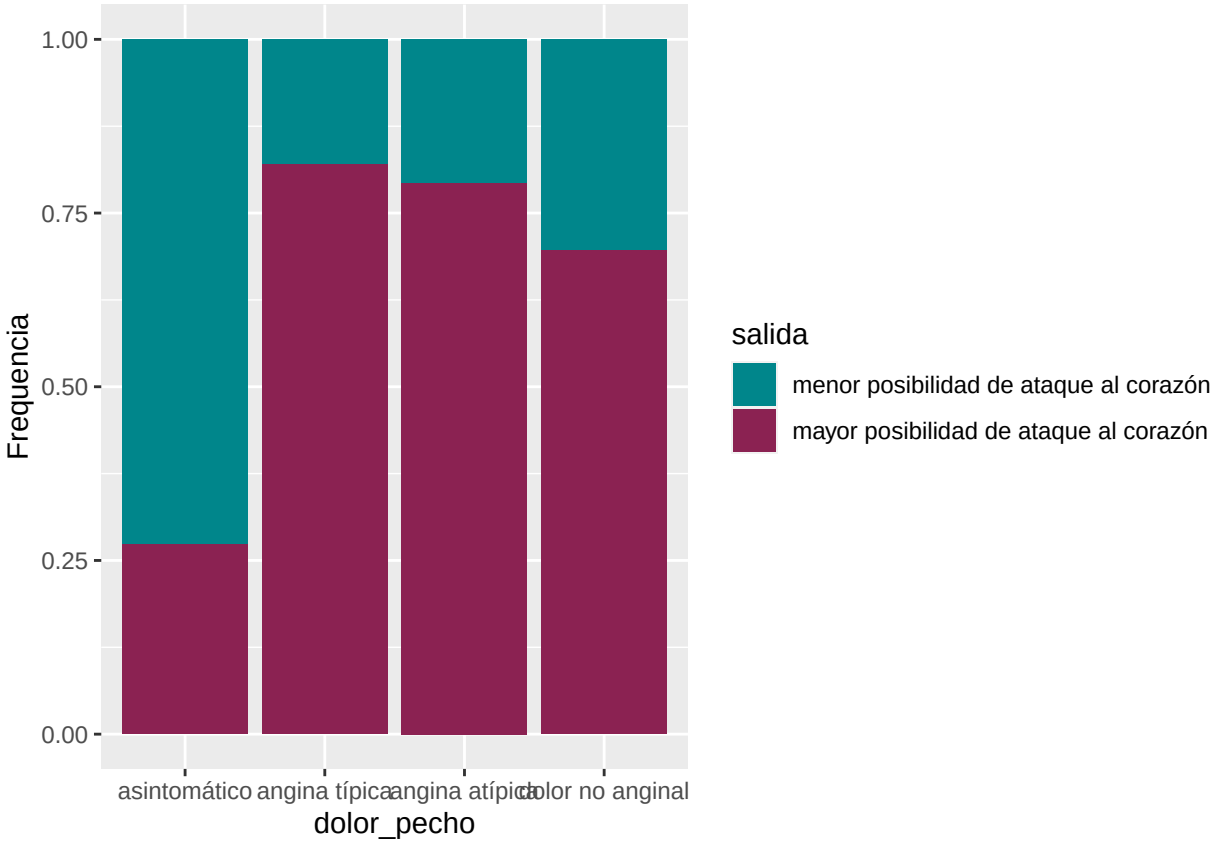
- **sexo vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=sexo ,fill=salida))+geom_bar(position="fill")+ylab("Frequenc
```



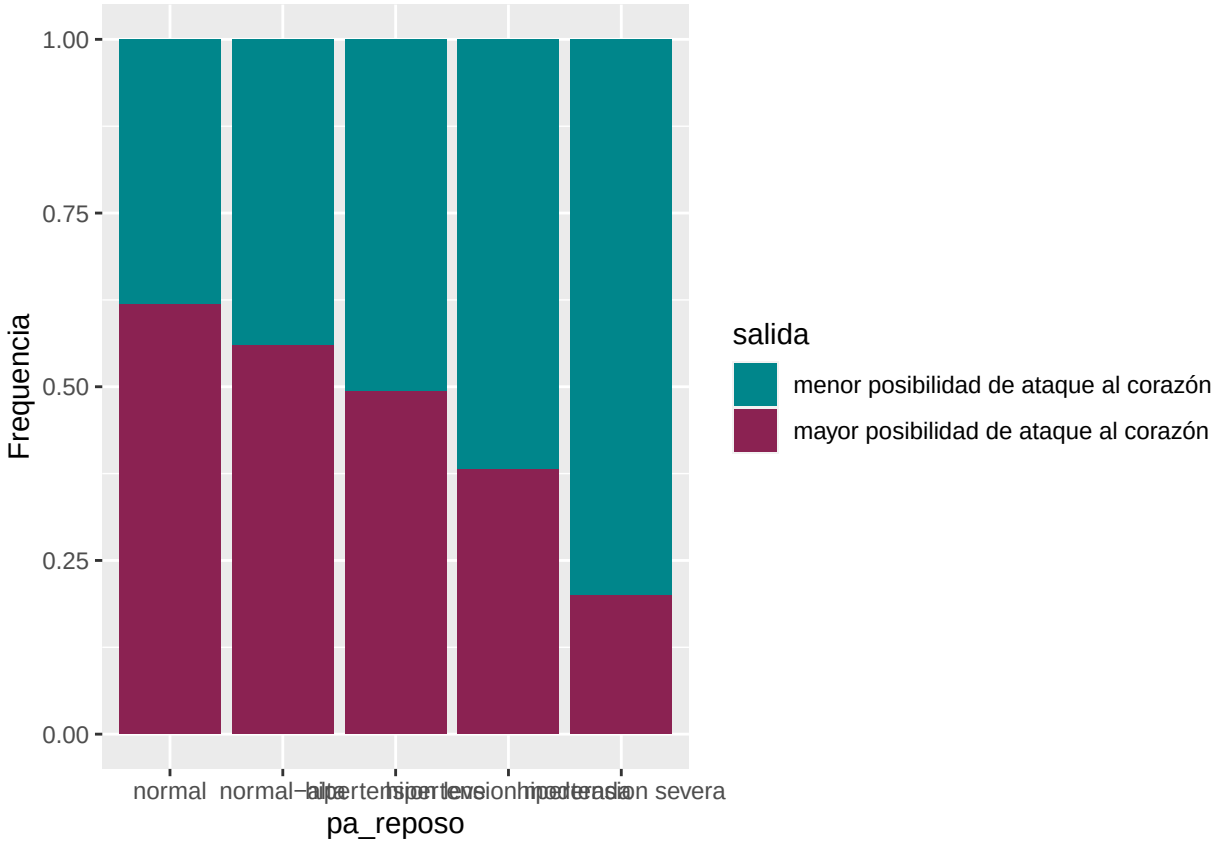
- **dolor_pecho vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=dolor_pecho,fill=salida))+geom_bar(position="fill")+ylab("Fr
```



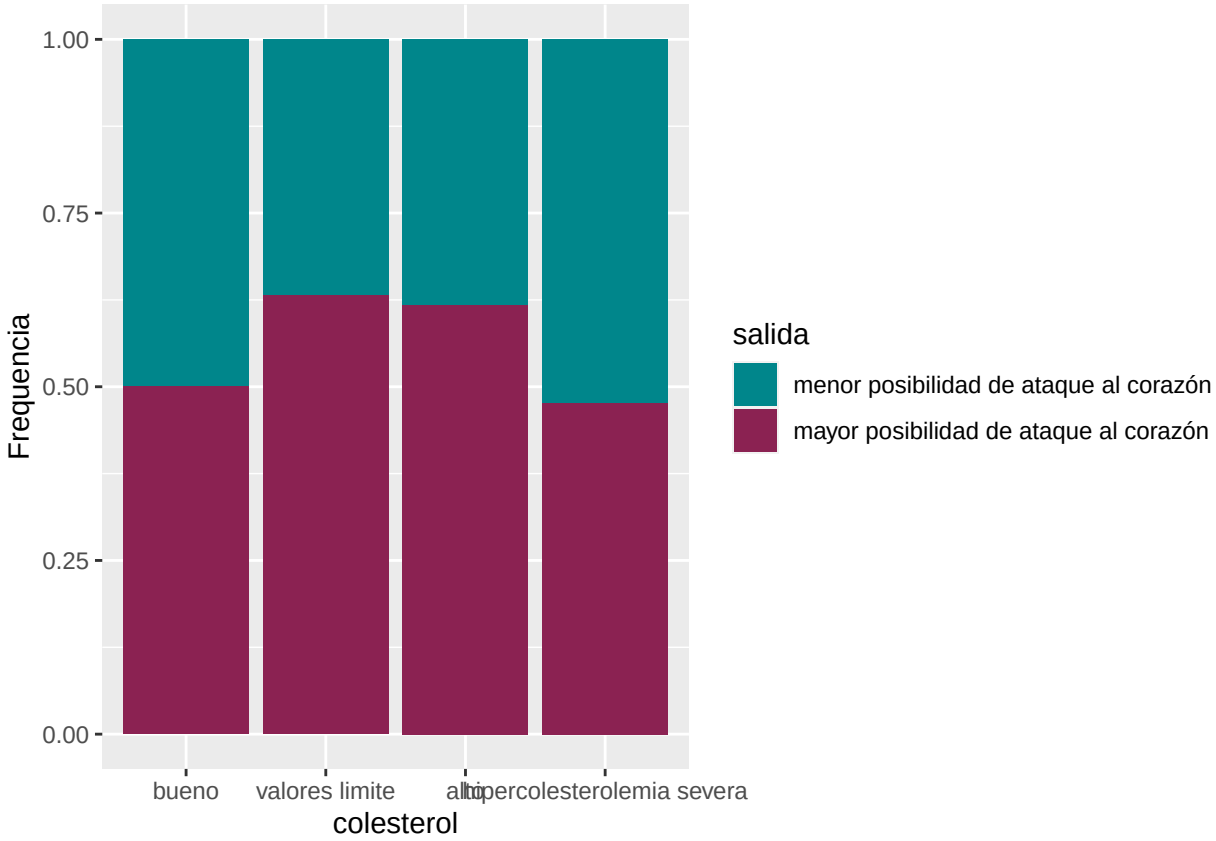
- **pa_reposo vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=pa_reposo,fill=salida))+geom_bar(position="fill")+ylab("Freq
```



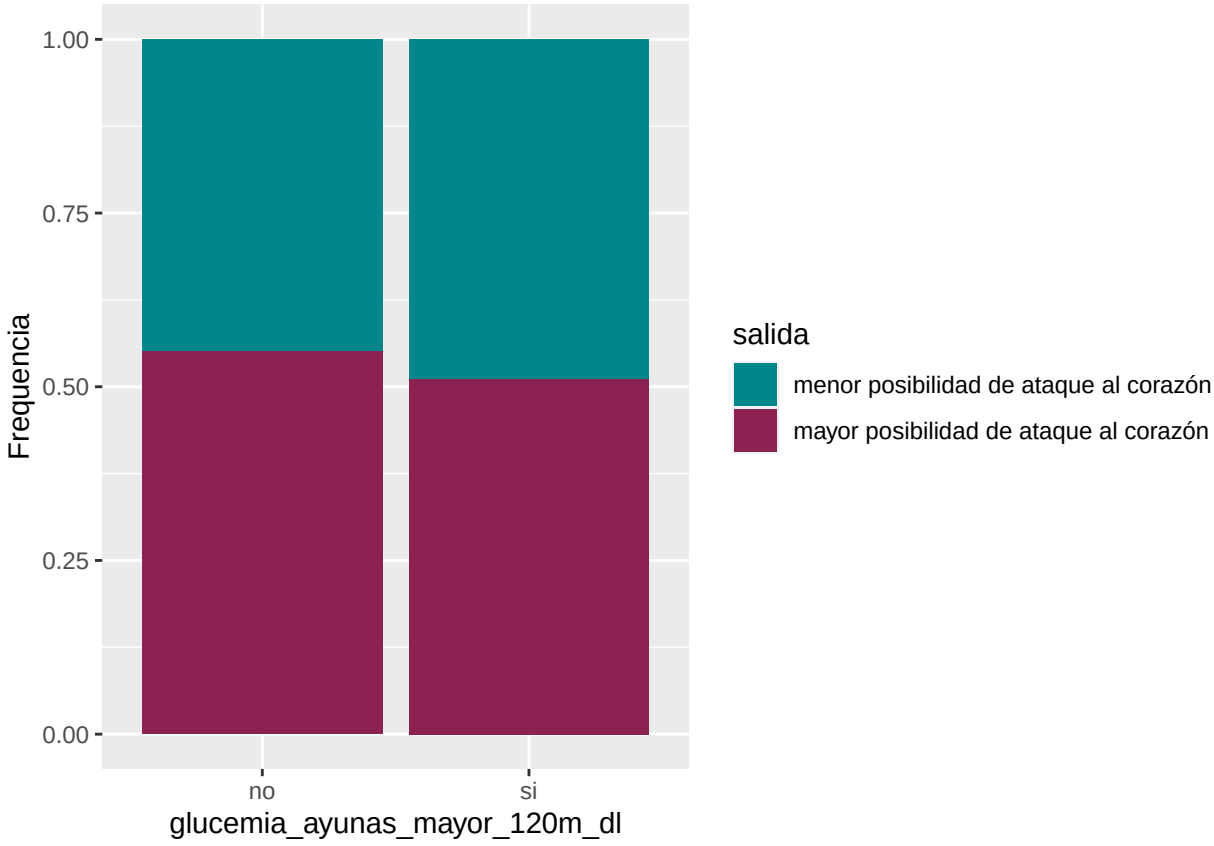
- **colesterol vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=colesterol ,fill=salida))+geom_bar(position="fill")+ylab("Fr
```



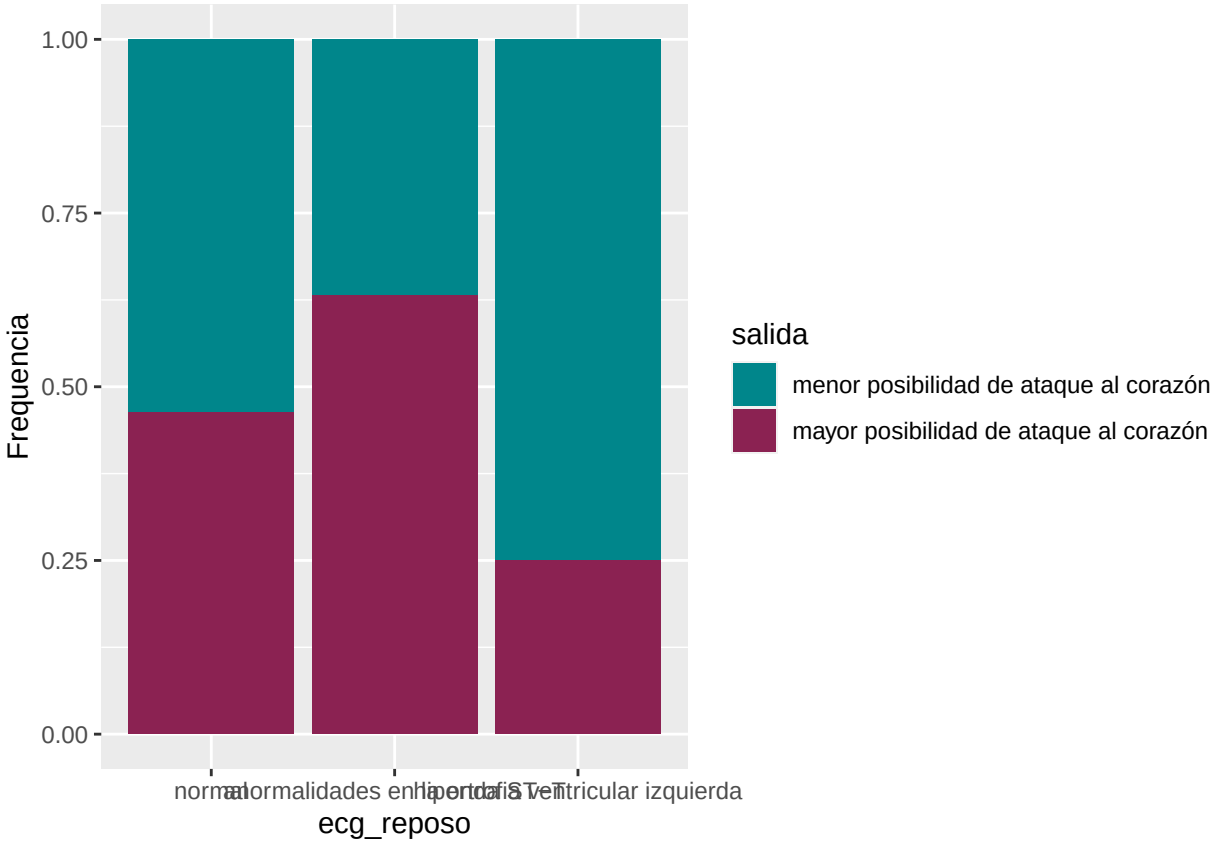
- **glucemia_ayunas_mayor_120m_dl vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=glucemia_ayunas_mayor_120m_dl ,fill=salida))+geom_bar(positi
```



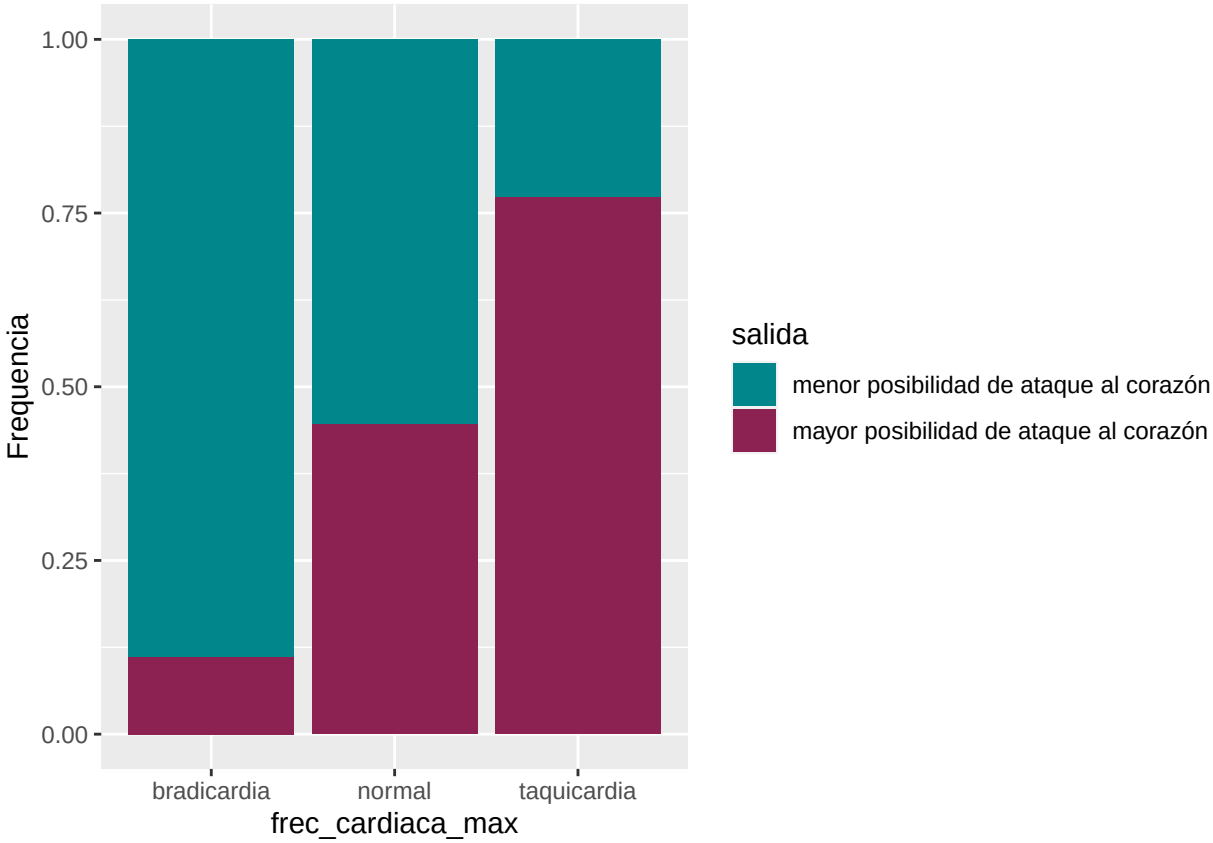
- **ecg_reposo vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=ecg_reposo ,fill=salida))+geom_bar(position="fill")+ylab("Fr
```



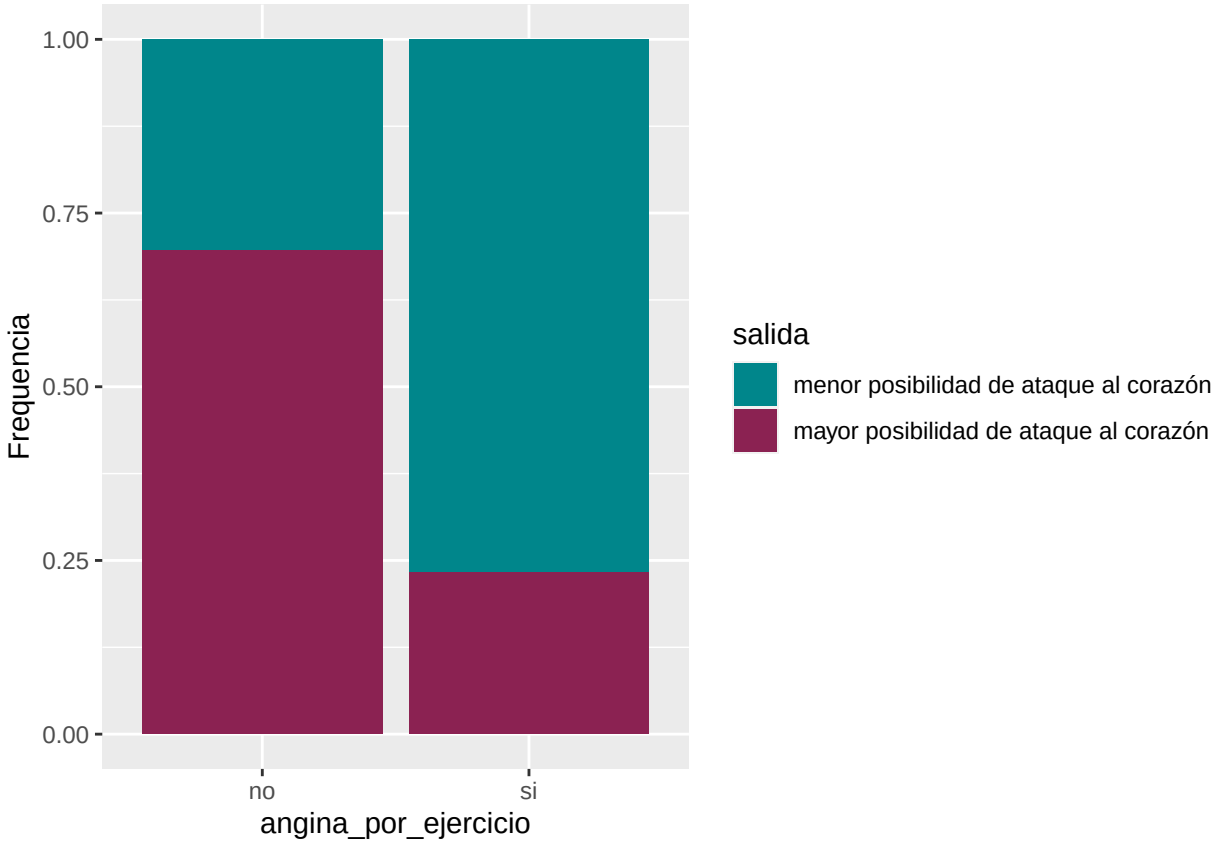
- **frec_cardiaca_max vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=frec_cardiaca_max ,fill=salida))+geom_bar(position="fill")+y
```



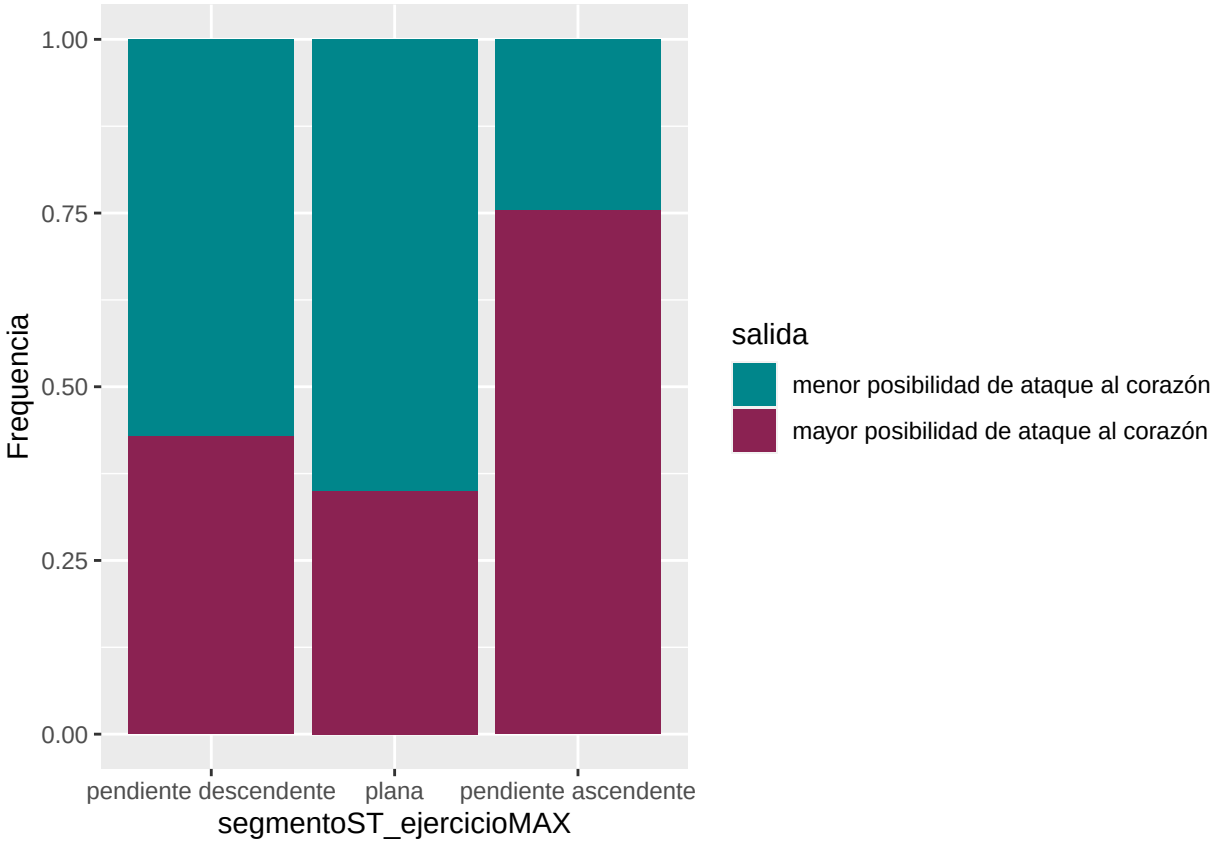
- **angina_por_ejercicio vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=angina_por_ejercicio,fill=salida))+geom_bar(position="fill")+
```



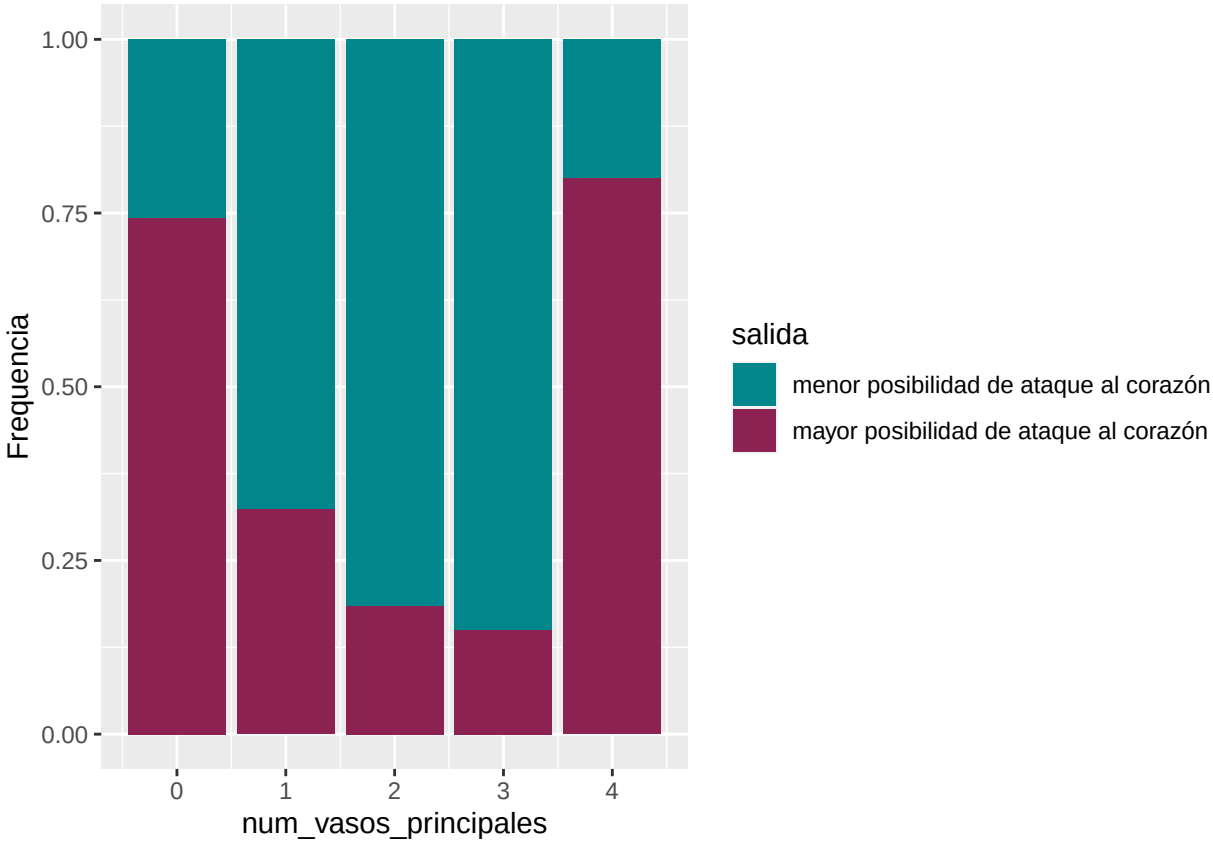
- **segmentoST_ejercicioMAX vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=segmentoST_ejercicioMAX,fill=salida))+geom_bar(position="fil
```



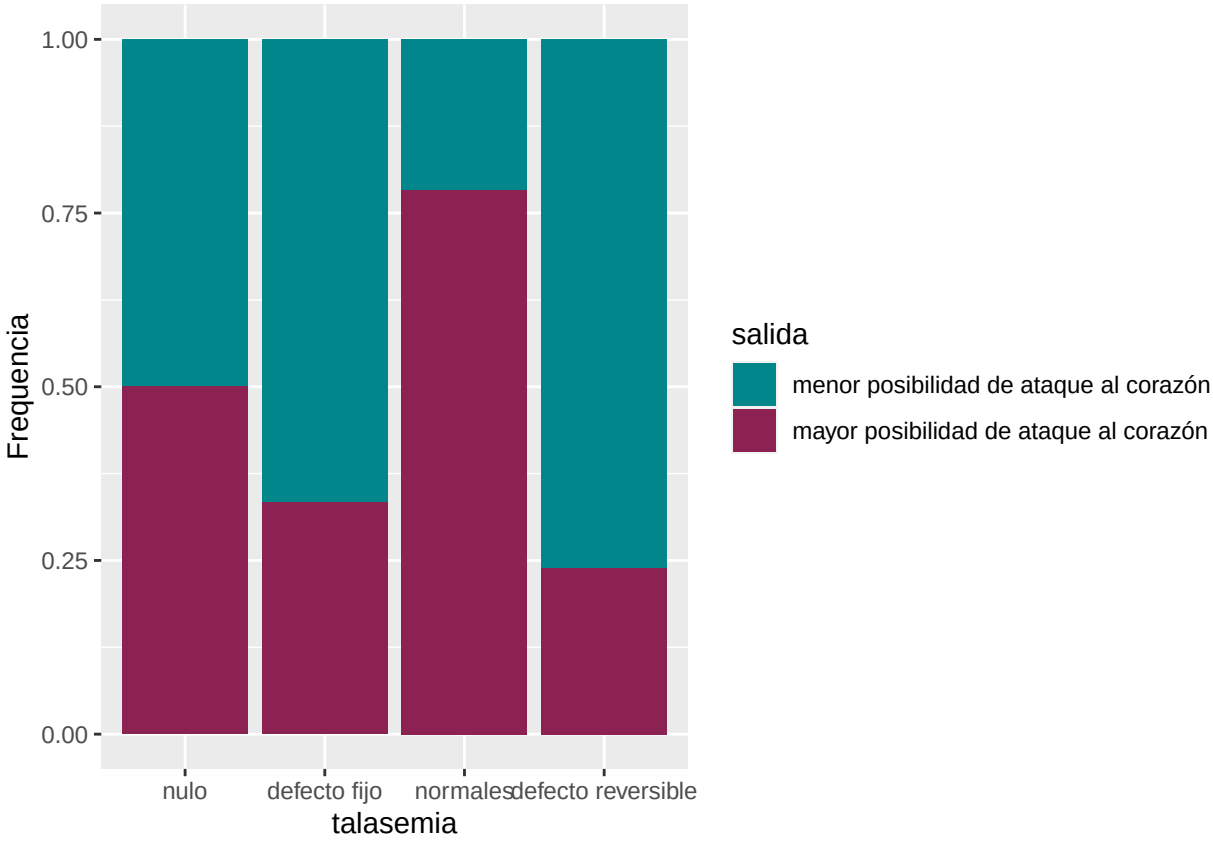
- **num_vasos_principales vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=num_vasos_principales,fill=salida))+geom_bar(position="fill"
```



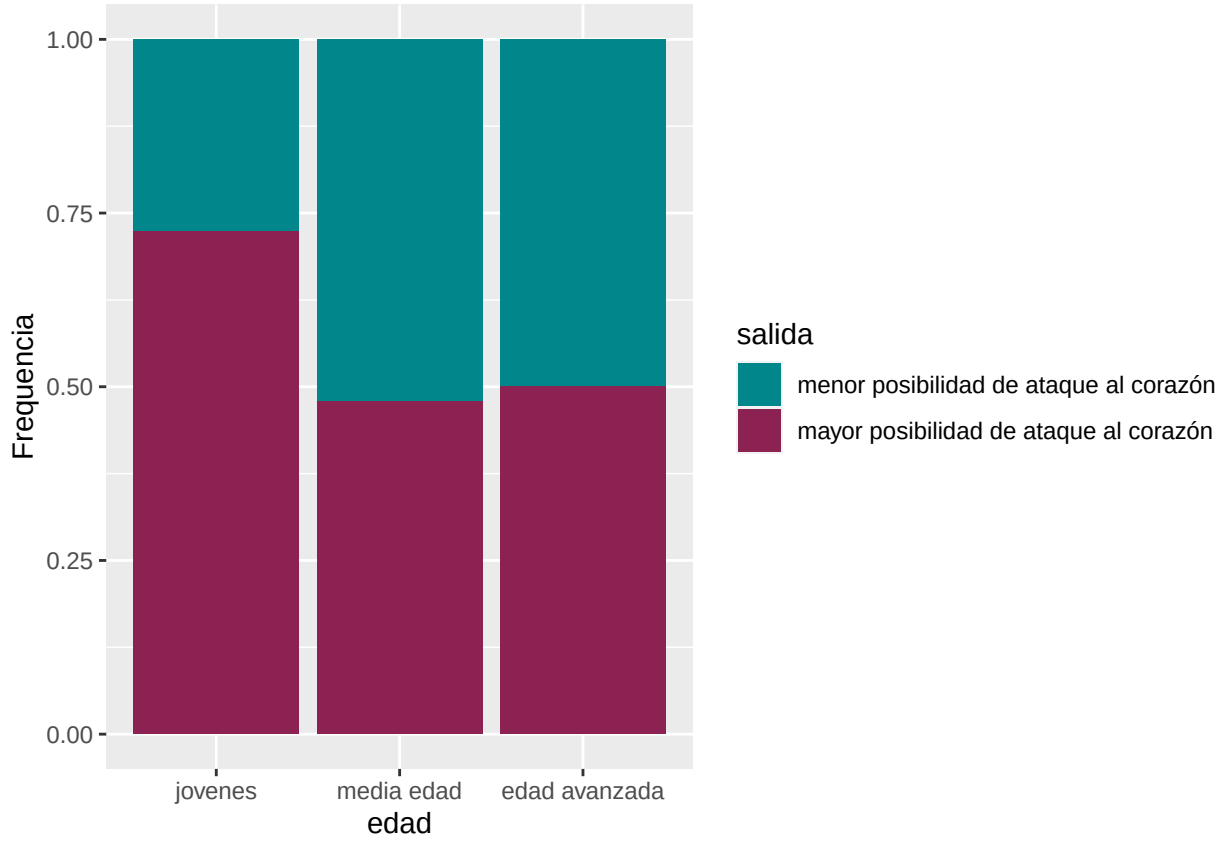
- **talasemia vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=talasemia,fill=salida))+geom_bar(position="fill")+ylab("Frequ
```



- **edad vs salida**

```
ggplot(datos_discretizados[1:filas,],aes(x=edad,fill=salida))+geom_bar(position="fill")+ylab("Frequencia
```

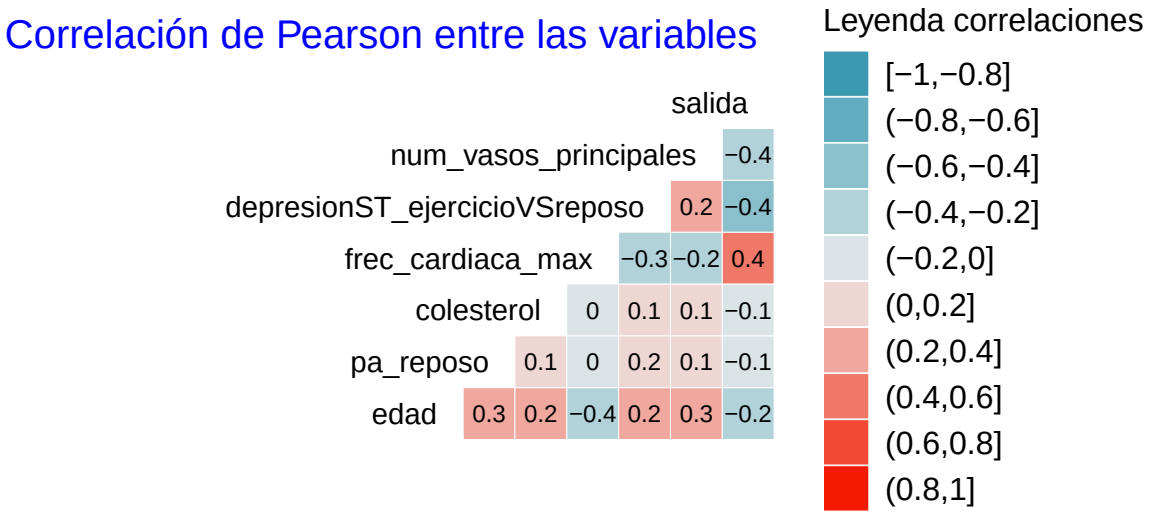


2.5.4 Análisis descriptivo y de correlaciones

Vamos a llevar a cabo un estudio de la correlación entre las variables numéricas. Para eso vamos a calcular la **correlación de Pearson** sobre el dataframe datos_normalizados ya que se ignoraran aquellas columnas que no sean numéricas:

```
ggcorr(datos_normalizados[,c(1:14)], method = c("everything", "pearson"), nbreaks = 10, name = "Leyenda
```

```
## Warning in ggcorr(datos_normalizados[, c(1:14)], method = c("everything", :
## data in column(s) 'sexo', 'dolor_pecho', 'glucemia_ayunas_mayor_120m_dl',
## 'ecg_reposo', 'angina_por_ejercicio', 'segmentoST_ejercicioMAX', 'talasemia' are
## not numeric and were ignored
```



Centrándonos en las relaciones que nos interesan para el objetivo del proyecto, es decir, **la relación de salida con el resto de variables**, podemos observar que la mayor correlación positiva encontrada entre salida y el resto de variables es entre salida y frec_cardiaca_max con una correlación aproximada de 0,4. En cuanto a las correlaciones negativas encontramos con un valor de -0,4 la correlación entre salida y num_vasos_principales y la correlación entre salida y depresionST_ejercicioVSreposo. Entre el resto de variables encontramos una correlación moderada entre edad y frec_cardiaca_max (-0,4), entre edad y pa_reposo (0,3) y entre edad y num_vasos_principales (0,3).

2.5.5 Grupos de datos a analizar

Vamos a analizar el grupo de mujeres con respecto al grupo de hombres. Vamos a realizar un contraste de hipótesis entrre estos dos grupos para estudiar si las mujeres tienen estadísticamente más probabilidad de ataques al corazon que los hombres. Además vamos a analizar al grupo con output 1(mayor posibilidad de ataque al corazón) con respecto al grupo con output 0 (menor posibilidad de ataque al corazón) para generar un modelo de clasificación que permita determinar si un registro pertenece a un grupo o a otro.

2.5.5 Comprobación de la normalidad y homogeneidad de la varianza

Se va a llevar a cabo un contraste unilateral de dos muestras independientes sobre la media. Al ser el tamaño de las muestras mayor de 30 asumimos normalidad y como las varianzas poblacionales son desconocidas tenemos que comprobar si son desconocidas iguales o diferentes realizando el test de homoscedasticidad.

Realizamos el test de homoscedasticidad

La hipótesis nula y alternativa son:

```r
print("H0:  12 =  22")
```

```
## [1] "H0:  12 =  22"
```

```r
print("H1:  12   22")
```

```
## [1] "H1:  12   22"
```

```r
femenino <- datos_renombrados$salida[datos_renombrados$sex==0]
masculino <- datos_renombrados$salida[datos_renombrados$sex==1]
var.test(femenino, masculino)
```

```
##
##  F test to compare two variances
##
## data:  femenino and masculino
## F = 0.76208, num df = 95, denom df = 206, p-value = 0.1343
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5455293 1.0885394
## sample estimates:
## ratio of variances
##          0.7620767
```

Dado que p es mayor que el nivel de significancia o alfa (0.05) podemos aceptar la hipótesis nula (varianzas iguales).

2.5.6 Contraste de hipótesis

Como se ha aceptado la hipótesis de varianzas iguales, aplicamos el test unilateral sobre la media de dos poblaciones independientes con varianza desconocida igual y asumimos normalidad. Se elige muestras independientes porque los datos no tienen relación entre ellos y proceden de poblaciones diferentes (como podemos ver vienen de registros distintos).

Las hipótesis serían:

```r
print("H0:  sigma m = sigma h")
```

```
## [1] "H0:  sigma m = sigma h"
```

```r
print("H1:  sigma m > sigma h")
```

```
## [1] "H1:  sigma m > sigma h"
```

La hipótesis nula indica que no hay diferencias estadísticamente significativas entre la media de la probabilidad de ataques al corazón en mujeres y en hombres. La hipótesis alternativa indica que hay diferencias estadísticamente significativas entre la media de la probabilidad de ataques al corazón en mujeres y en hombres, siendo las mujeres las que tienen una mayor probabilidad media de ataques al corazón.

```r
t.test(x=femenino, y=masculino, var.equal=TRUE, alpha=0.05, alternative="greater")
```

```
##
##  Two Sample t-test
##
## data:  femenino and masculino
## t = 5.0786, df = 301, p-value = 3.339e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.2030255       Inf
## sample estimates:
## mean of x mean of y
## 0.7500000 0.4492754
```

Dado que p es menor que el nivel de significancia o alfa(0.05) podemos rechazar la hipótesis nula a favor de la alternativa.Por tanto, concluimos que exiten diferencias estadísticamente signiticativas en la media de la probabilidad de ataques al corazón de forma que las mujeres tienen más probabilidad de ataques al corazón que los hombres.

2.5.7 Modelo de clasificación

Dividimos los datos entre el conjunto de entrenamiento y de prueba diferenciando entre la variable objetivo y el resto de variables

```r
set.seed(20)
y <- datos_discretizados[,14]
X <- datos_discretizados[,c(1:13)]
split_prop <- 3
indexes = sample(1:nrow(datos_discretizados), size=floor(((split_prop-1)/split_prop)*nrow(datos_discreti
trainX<-X[indexes,]
trainy<-y[indexes]
testX<-X[-indexes,]
testy<-y[-indexes]
```

Mostramos los conjuntos de datos creados:

```r
summary(trainX)
```

```
##             edad            sexo               dolor_pecho
##  jovenes      : 53   femenino : 66   asintomático    :101
##  media edad   :108   masculino:136   angina típica   : 28
##  edad avanzada: 41                   angina atípica  : 57
##                                      dolor no anginal: 16
##
##
##                  pa_reposo                       colesterol
##  normal               :61   bueno                      :  8
##  normal-alta          :79   valores limite             : 23
##  hipertension leve    :46   alto                       : 64
##  hipertension moderada:14   hipercolesterolemia severa:107
##  hipertension severa  : 2
```

```
##
##  glucemia_ayunas_mayor_120m_dl                              ecg_reposo
##  no:175                        normal                           :103
##  si: 27                        anormalidades en la onda ST-T    : 98
##                                hipertrofia ventricular izquierda:  1
##
##
##
##    frec_cardiaca_max angina_por_ejercicio depresionST_ejercicioVSreposo
##  bradicardia: 14     no:136               Min.   :0.000
##  normal     :115     si: 66               1st Qu.:0.000
##  taquicardia: 73                         Median :0.600
##                                          Mean   :1.006
##                                          3rd Qu.:1.600
##                                          Max.   :6.200
##           segmentoST_ejercicioMAX num_vasos_principales
##  pendiente descendente:14         Min.   :0.0000
##  plana                :95         1st Qu.:0.0000
##  pendiente ascendente :93         Median :0.0000
##                                   Mean   :0.7921
##                                   3rd Qu.:1.0000
##                                   Max.   :4.0000
##               talasemia
##  nulo               :  1
##  defecto fijo       : 10
##  normales           :104
##  defecto reversible: 87
##
##
```

```
summary(trainy)
```

```
## menor posibilidad de ataque al corazón mayor posibilidad de ataque al corazón
##                                     97                                     105
```

```
summary(testX)
```

```
##             edad           sexo              dolor_pecho
##  jovenes      :23   femenino :30   asintomático    :42
##  media edad   :59   masculino:71   angina típica   :22
##  edad avanzada:19                  angina atípica  :30
##                                    dolor no anginal: 7
##
##
##                  pa_reposo                       colesterol
##  normal               :36   bueno                     : 4
##  normal-alta          :30   valores limite            :15
##  hipertension leve    :25   alto                      :38
##  hipertension moderada: 7   hipercolesterolemia severa:44
##  hipertension severa  : 3
##
##  glucemia_ayunas_mayor_120m_dl                              ecg_reposo
##  no:83                         normal                           :44
```

```
##  si:18                          anormalidades en la onda ST-T    :54
##                                 hipertrofia ventricular izquierda: 3
##
##
##
##    frec_cardiaca_max angina_por_ejercicio depresionST_ejercicioVSreposo
##  bradicardia: 4      no:68                Min.   :0.000
##  normal     :60      si:33                1st Qu.:0.000
##  taquicardia:37                           Median :0.900
##                                           Mean   :1.107
##                                           3rd Qu.:1.900
##                                           Max.   :5.600
##           segmentoST_ejercicioMAX num_vasos_principales              talasemia
##  pendiente descendente: 7         Min.   :0.000         nulo               : 1
##  plana                 :45        1st Qu.:0.000         defecto fijo       : 8
##  pendiente ascendente :49         Median :0.000         normales           :62
##                                   Mean   :0.604         defecto reversible:30
##                                   3rd Qu.:1.000
##                                   Max.   :4.000
```

```
summary(testy)
```

```
## menor posibilidad de ataque al corazón mayor posibilidad de ataque al corazón
##                                     41                                     60
```

Se crea el árbol de decisión usando los datos de entrenamiento y se gráfica

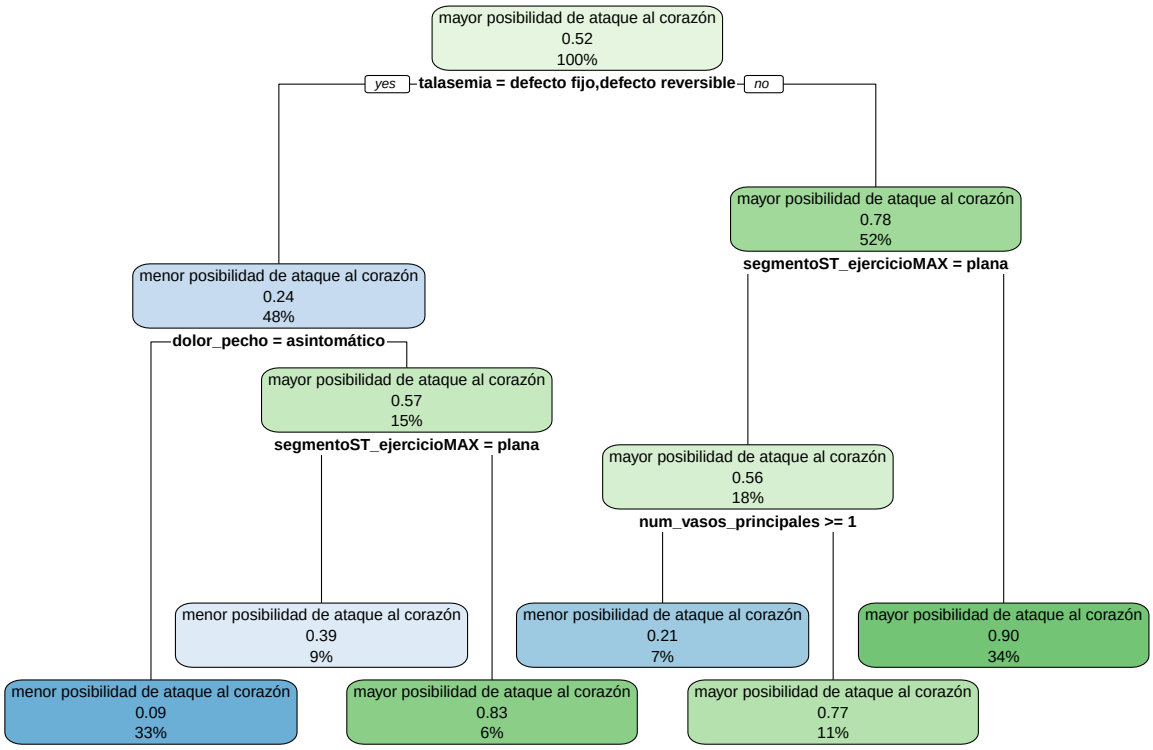
```
#install.packages('rpart.plot');
library('rpart.plot')
```

```
## Warning: package 'rpart.plot' was built under R version 4.2.2
```

```
## Loading required package: rpart
```

```
model <- rpart(trainy ~ ., data = trainX)

rpart.plot(model)
```



Se muestran los datos del modelo:

```
summary(model)
```

```
## Call:
## rpart(formula = trainy ~ ., data = trainX)
##   n= 202
##
##           CP nsplit rel error    xerror       xstd
## 1 0.52577320      0 1.0000000 1.1030928 0.07313180
## 2 0.04123711      1 0.4742268 0.5154639 0.06323527
## 3 0.01000000      5 0.3092784 0.4226804 0.05893285
##
## Variable importance
##                     talasemia                   dolor_pecho
##                            27                            19
##       segmentoST_ejercicioMAX         num_vasos_principales
##                            15                            12
##                         sexo          angina_por_ejercicio
##                            10                             9
## depresionST_ejercicioVSreposo            frec_cardiaca_max
##                             3                             1
## glucemia_ayunas_mayor_120m_dl                    colesterol
##                             1                             1
##                         edad
##                             1
```

```
##
## Node number 1: 202 observations,    complexity param=0.5257732
##   predicted class=mayor posibilidad de ataque al corazón  expected loss=0.480198  P(node) =1
##     class counts:    97   105
##    probabilities: 0.480 0.520
##   left son=2 (97 obs) right son=3 (105 obs)
##   Primary splits:
##       talasemia                     splits as  RLRL,     improve=29.82499, (0 missing)
##       dolor_pecho                   splits as  LRRR,     improve=27.81188, (0 missing)
##       num_vasos_principales         < 0.5  to the right, improve=22.65798, (0 missing)
##       depresionST_ejercicioVSreposo < 0.7  to the right, improve=19.93987, (0 missing)
##       angina_por_ejercicio          splits as  RL,       improve=18.56039, (0 missing)
##   Surrogate splits:
##       dolor_pecho             splits as  LRRR,     agree=0.683, adj=0.340, (0 split)
##       sexo                    splits as  RL,       agree=0.678, adj=0.330, (0 split)
##       angina_por_ejercicio    splits as  RL,       agree=0.668, adj=0.309, (0 split)
##       num_vasos_principales   < 0.5  to the right, agree=0.649, adj=0.268, (0 split)
##       segmentoST_ejercicioMAX splits as  LLR,      agree=0.644, adj=0.258, (0 split)
##
## Node number 2: 97 observations,    complexity param=0.04123711
##   predicted class=menor posibilidad de ataque al corazón  expected loss=0.2371134  P(node) =0.480198
##     class counts:    74    23
##    probabilities: 0.763 0.237
##   left son=4 (67 obs) right son=5 (30 obs)
##   Primary splits:
##       dolor_pecho                   splits as  LRRR,     improve=9.434077, (0 missing)
##       depresionST_ejercicioVSreposo < 0.7  to the right, improve=5.914405, (0 missing)
##       num_vasos_principales         < 0.5  to the right, improve=5.914405, (0 missing)
##       angina_por_ejercicio          splits as  RL,       improve=3.358940, (0 missing)
##       colesterol                    splits as  LLRL,     improve=2.732207, (0 missing)
##   Surrogate splits:
##       angina_por_ejercicio splits as  RL, agree=0.701, adj=0.033, (0 split)
##
## Node number 3: 105 observations,    complexity param=0.04123711
##   predicted class=mayor posibilidad de ataque al corazón  expected loss=0.2190476  P(node) =0.519802
##     class counts:    23    82
##    probabilities: 0.219 0.781
##   left son=6 (36 obs) right son=7 (69 obs)
##   Primary splits:
##       segmentoST_ejercicioMAX       splits as  RLR,      improve=5.566322, (0 missing)
##       dolor_pecho                   splits as  LRRL,     improve=5.470754, (0 missing)
##       num_vasos_principales         < 0.5  to the right, improve=4.961921, (0 missing)
##       depresionST_ejercicioVSreposo < 0.85 to the right, improve=4.052597, (0 missing)
##       angina_por_ejercicio          splits as  RL,       improve=3.429557, (0 missing)
##   Surrogate splits:
##       depresionST_ejercicioVSreposo < 0.85 to the right, agree=0.743, adj=0.250, (0 split)
##       dolor_pecho                   splits as  LRRR,     agree=0.733, adj=0.222, (0 split)
##       angina_por_ejercicio          splits as  RL,       agree=0.695, adj=0.111, (0 split)
##       frec_cardiaca_max             splits as  LRR,      agree=0.686, adj=0.083, (0 split)
##       colesterol                    splits as  LRRR,     agree=0.676, adj=0.056, (0 split)
##
## Node number 4: 67 observations
##   predicted class=menor posibilidad de ataque al corazón  expected loss=0.08955224  P(node) =0.33168
##     class counts:    61     6
```

```
##    probabilities: 0.910 0.090
##
## Node number 5: 30 observations,    complexity param=0.04123711
##   predicted class=mayor posibilidad de ataque al corazón  expected loss=0.4333333  P(node) =0.148514
##     class counts:    13    17
##    probabilities: 0.433 0.567
##   left son=10 (18 obs) right son=11 (12 obs)
##   Primary splits:
##       segmentoST_ejercicioMAX       splits as  RLR,      improve=2.8444440, (0 missing)
##       num_vasos_principales         < 0.5  to the right, improve=1.5206640, (0 missing)
##       depresionST_ejercicioVSreposo < 1.1  to the right, improve=1.4318980, (0 missing)
##       colesterol                    splits as  RRRL,     improve=0.9000000, (0 missing)
##       pa_reposo                     splits as  RLRRL,    improve=0.8333333, (0 missing)
##   Surrogate splits:
##       frec_cardiaca_max             splits as  LLR,      agree=0.767, adj=0.417, (0 split)
##       depresionST_ejercicioVSreposo < 0.3  to the right, agree=0.700, adj=0.250, (0 split)
##       pa_reposo                     splits as  LLRLR,    agree=0.667, adj=0.167, (0 split)
##       glucemia_ayunas_mayor_120m_dl splits as  LR,       agree=0.667, adj=0.167, (0 split)
##       colesterol                    splits as  RLLL,     agree=0.633, adj=0.083, (0 split)
##
## Node number 6: 36 observations,    complexity param=0.04123711
##   predicted class=mayor posibilidad de ataque al corazón  expected loss=0.4444444  P(node) =0.178217
##     class counts:    16    20
##    probabilities: 0.444 0.556
##   left son=12 (14 obs) right son=13 (22 obs)
##   Primary splits:
##       num_vasos_principales         < 0.5  to the right, improve=5.336219, (0 missing)
##       angina_por_ejercicio          splits as  RL,       improve=4.425051, (0 missing)
##       sexo                          splits as  RL,       improve=3.336219, (0 missing)
##       depresionST_ejercicioVSreposo < 0.75 to the right, improve=3.073016, (0 missing)
##       dolor_pecho                   splits as  LRRL,     improve=2.777778, (0 missing)
##   Surrogate splits:
##       depresionST_ejercicioVSreposo < 1.7  to the right, agree=0.722, adj=0.286, (0 split)
##       glucemia_ayunas_mayor_120m_dl splits as  RL,       agree=0.694, adj=0.214, (0 split)
##       edad                          splits as  RRL,      agree=0.667, adj=0.143, (0 split)
##       sexo                          splits as  RL,       agree=0.667, adj=0.143, (0 split)
##       colesterol                    splits as  RRRL,     agree=0.667, adj=0.143, (0 split)
##
## Node number 7: 69 observations
##   predicted class=mayor posibilidad de ataque al corazón  expected loss=0.1014493  P(node) =0.341584
##     class counts:     7    62
##    probabilities: 0.101 0.899
##
## Node number 10: 18 observations
##   predicted class=menor posibilidad de ataque al corazón  expected loss=0.3888889  P(node) =0.089108
##     class counts:    11     7
##    probabilities: 0.611 0.389
##
## Node number 11: 12 observations
##   predicted class=mayor posibilidad de ataque al corazón  expected loss=0.1666667  P(node) =0.059405
##     class counts:     2    10
##    probabilities: 0.167 0.833
##
## Node number 12: 14 observations
```

```
##   predicted class=menor posibilidad de ataque al corazón  expected loss=0.2142857  P(node) =0.0693069
##     class counts:    11     3
##    probabilities: 0.786 0.214
##
## Node number 13: 22 observations
##   predicted class=mayor posibilidad de ataque al corazón  expected loss=0.2272727  P(node) =0.1089109
##     class counts:     5    17
##    probabilities: 0.227 0.773
```

Se muestran las reglas:

```
reglas <- rpart.rules(model)
reglas
```

```
##  trainy
##    0.09 when talasemia is defecto fijo or defecto reversible
##    0.21 when talasemia is                    nulo or normales & segmentoST_ejercicioMAX is
##    0.39 when talasemia is defecto fijo or defecto reversible & segmentoST_ejercicioMAX is
##    0.77 when talasemia is                    nulo or normales & segmentoST_ejercicioMAX is
##    0.83 when talasemia is defecto fijo or defecto reversible & segmentoST_ejercicioMAX is pendiente
##    0.90 when talasemia is                    nulo or normales & segmentoST_ejercicioMAX is pendiente
```

Se calcula la precisión del modelo:

```
predicted_model <- predict( model, testX, type="class" )
print(sprintf("La precisión del árbol es: %.4f %%",100*sum(predicted_model == testy) / length(predicted_
```

```
## [1] "La precisión del árbol es: 78.2178 %"
```

Analizamos mediante una matriz de confusión los tipos de errores cometidos:

```
mat_conf<-table(testy,Predicted=predicted_model)
mat_conf
```

```
##                                        Predicted
## testy                                   menor posibilidad de ataque al corazón
##   menor posibilidad de ataque al corazón                                     27
##   mayor posibilidad de ataque al corazón                                      8
##                                        Predicted
## testy                                   mayor posibilidad de ataque al corazón
##   menor posibilidad de ataque al corazón                                     14
##   mayor posibilidad de ataque al corazón                                     52
```

2.6 Conclusiones de los análisis y modelos realizados

Al analizar las correlaciones de las variables de interés para el estudio hemos podido observar que la mayor correlación encontrada entre salida y el resto de variables es entre salida y frec_cardiaca_max con una correlación aproximada de **0,4**. Entre el resto de variables encontramos una correlación moderada entre edad y pa_reposo **0,3** y entre edad y num_vasos_principales **0,3**.

El contraste de hipótesis nos ha permitido concluir que existen diferencias estadísticamente significativas (al **95%**) en la media de la probabilidad de ataques al corazón de forma que las mujeres tienen más probabilidad de ataques al corazón que los hombres.

En cuanto al modelo de clasificación generado podemos ver que ha alcanzado una precisión del **78.22%**.

En la tabla se puede observar como el modelo:

- Ha clasificado correctamente **79 casos** (la suma de los valores diagonales)
- Ha clasificado erróneamente **22** (suma del resto de valores).

De los 22 casos erróneamente clasificados 8 corresponden a falsos negativos. Es decir, un 7.92% de los casos serán clasificados con una probabilidad menor de ataque al corazón cuando en realidad tienen una mayor probabilidad.

En un área como el de la salud, a pesar de haber obtenido una precisión moderada-alta, un 7,92% de falsos negativos es un porcentaje elevado al poder afectar a la salud de las personas.

Por todo lo expuesto, podemos concluir que un **sistema clasificador como este debería tomarse como una recomendación o ayuda para un profesional y no como una herramienta de diagnóstico**.

3. Documentación consultada

- *Calvo M, Subirats L, Pérez D (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.*
- https://scientistcafe.com/ids/regression-and-decision-tree-basic.html
- *Árboles de decisión. Ramon Sangüesa i Solé*
- *Tutorial de Github (https://guides.github.com/activities/hello-world/)*
- *Squire, Megan (2015). Clean Data. Packt Publishing Ltd.*
- *Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.*
- *Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews.*
- *Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.*
- *Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.*

4. Contribuciones

Alba Sanz Horcajo: ASH

Carlos Santamaría de las Heras: CSH

- Investigación previa: CSH, ASH
- Redacción de las respuestas: CSH, ASH
- Desarrollo del código: CSH, ASH
- Participación en el vídeo: CSH, ASH