Tipología de datos

Janneth Chicaiza, Jaime Velandia

June 2019

Descripción de las fuentes de datos

Origen y estructura de datasets

De forma concreta se ha elegido el dataset Heart Disease el cual comprende varios archivos de datos proporcionados por diferentes instituciones de salud. La base de datos contiene 14 atributos y corresponde a información de pacientes recolectada por instituciones como: Cleveland Clinic Foundation, Hungarian Institute of Cardiology, University Hospital of Zurich, y University Hospital de Basel. Para la presente práctica, se han tomando como base los datos de las dos primeras instituciones.

El dataset contiene una serie de observaciones de pacientes clasificados de acuerdo a si tienen alguna enfermedad cardíaca o no. Los datos puede ser descargados desde UCI Machine Learning Repository y contiene la siguiente información:

Atributo	Descripción	Tipo de dato
age	edad del paciente en año	entero
sex	sexo del paciente (1 = male; 0 = female)	nominal
cp	tipo de dolor torácico (1 = angina típica, 1 = angina atípica, 2 = dolor no anginal, 4 = asintomático)	nominal
trestbps	presión arterial en reposo (en mm Hg al ingreso en el hospital)	entero
chol	serum cholestoral in mg/dl	entero
fbs	azúcar en la sangre en ayunas. Si es > 120 mg/dl, entonces 1 = verdadero sino 0 = falso	nominal
restecg	resultados electrocardiográficos en reposo (0 = normal, 1 = existe anomalía en la onda ST-T; 2 =muestra una hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes)	nominal
thalach	ritmo cardíaco máximo alcanzado	entero
exang	angina inducida por el ejercicio (1 = sí; 0 = no)	nominal
oldpeak	depresión ST inducida por el ejercicio en relación con el	decimal

	descanso.	
slope	pendiente del segmento pico del ejercicio ST (1 = pendiente ascendente, 2 = plano, 3 = pendiente descendente)	nominal
ca	número de vasos principales (0-3) coloreados por fluoroscopia	entero
thal	talasemia: trastorno sanguíneo hereditario que hace provoca que el cuerpo produzca menos hemoglobina o la hemoglobina sea anormal. (3 = normal, 6 = defecto fijo, 7 = defecto reversible)	nominal
target	indica la presencia de enfermedad cardíaca en el paciente o no (0 = no se evidencia presencia, 1 = si)	nominal

De las 13 variables independientes, 5 son enteras, 1 es decimal y 7 son nominales discretas. La variable target contiene 4 valores.

Importancia

El dataset seleccionado para esta práctica es ***Heart Disease***. Algunos aspectos se pueden destacar de este conjunto de datos:

- El conjunto de datos contiene una serie de variables numéricas y nominales lo que facilita lo que constituye un reto y a la vez una oportunidad para aplicar diferentes técnicas específicas para cada tipo de dato.
- Puesto que se conoce la clase a predecir es posible aplicar diferentes técnicas de aprendizaje supervisado y no supervisado. Por ejemplo, se pueden crear árboles o un conjunto de reglas de decisión que modelen la relación entre las características que los pacientes y su posible impacto en una dolencia cardiaca.
- Los dos archivos elegidos suman más de 5 centenas de observaciones y 14 atributos, es decir, tiene una dimensión adecuada como para separarlo en dos conjuntos de datos y proceder así a aplicar diferentes técnicas de análisis.
- El dataset puede ser utilizado en múltiples estudios puesto que ha sido donado a UCI para este fin. Esto facilita la comparación entre múltiples algoritmos.
- El dataset corresponde al área de la salud. Estudiar los factores que influyen en la aparición o complicación del funcionamiento del corazón puede resultar positivo para apoyar al personal médico a detectar factores tempranos de riesgo en sus pacientes. También, a partir de un modelo robusto de predicción, se puede alertar a las personas para que conozcan los factores de riesgo y eviten así problemas del corazón.

Pregunta/problemas que pretende responder

Tres preguntas de análisis han sido establecidas para desarrollar la presente práctica.

1. ¿Cuáles son las variables que más influyen en una enfermedad relacionada al corazón?
2. ¿El tipo de dolor torácico del paciente (cp) es un indicativo de posible problema del corazón?

3. ¿Qué tan efectivo es un modelo de regresión logística para predecir padecimiento del corazón en el caso de hombres y de mujeres?

Integración y selección de datos

En este apartado se procede a cargar las dos fuentes seleccionadas. Luego se aplican algunas operaciones para unificar y seleccionar los atributos que puedan ayudar a responder las preguntas planteadas.

```
# Carga de datos:
  ## Hungarian Institute of Cardiology:
data1 <- read.csv("data/processed.hungarian.csv" , header=T, sep=",",
encoding = "UTF-8", dec = ".")
colnames(data1) <- c("age", "sex", "cp", "trestbps", "chol", "fbs",
"restecg", "thalach", "exang", "oldpeak", "slope", "ca", "thal",
"target")

  ##Cleveland Clinic Foundation:
data2 <- read.csv("data/processed.cleveland.csv" , header=T, sep=",",
encoding = "UTF-8", dec = ".")
colnames(data2) <- c("age", "sex", "cp", "trestbps", "chol", "fbs",
"restecg", "thalach", "exang", "oldpeak", "slope", "ca", "thal",
"target")

# Integración de datos:
dataH <- rbind(data1, data2)
dim(dataH)
```

```
## [1] 595  14
```

```
# Presentar las primeras 5 observaciones.
ej.dataH <- data.frame(cbind(t(head(dataH, 5))))
names(ej.dataH) <- c("Obs. 1", "Obs. 2", "Obs. 2", "Obs. 4", "Obs. 5")

ej.dataH
```

```
##          Obs. 1 Obs. 2 Obs. 2 Obs. 4 Obs. 5
## age          29     29     30     31     32
## sex           1      1      0      0      0
## cp            2      2      1      2      2
## trestbps    120    140    170    100    105
## chol        243      ?    237    219    198
## fbs           0      0      0      0      0
## restecg       0      0      1      1      0
## thalach     160    170    170    150    165
## exang         0      0      0      0      0
## oldpeak       0      0      0      0      0
```

```
## slope         ?      ?      ?      ?      ?
## ca            ?      ?      ?      ?      ?
## thal          ?      ?      6      ?      ?
## target        0      0      0      0      0
```

Según se observa en la tabla anterior, atributos como, *slope*, *ca* y *thal* tienen valores desconocidos, identificados por el símbolo "?". Se procede a verificar el porcentaje de elementos desconocidos por cada atributo, con el objetivo de seleccionar la mejor decisión al momento de procesarlos:

```
propSI <- function(x) {
  r <- NULL
  r[1] <- sum(ifelse(as.character(x)=="?", 1, 0))/NROW(x)
  return(r)
}

NA.data <- apply(X = dataH, MARGIN=2, FUN = function(x) propSI(x))

NA.data

##         age         sex          cp    trestbps        chol
fbs
## 0.000000000 0.000000000 0.000000000          NA          NA
0.013445378
##     restecg     thalach       exang     oldpeak       slope
ca
## 0.001680672          NA 0.001680672 0.000000000 0.317647059
0.494117647
##        thal      target
## 0.448739496 0.000000000
```

Del resumen presentado en la tabla anterior se puede observar que las variables *slope*, *ca* y *thal* tienen entre el 30% y 45% de valores desconocidos. Considerando este hecho, se procede a descartar estos atributos del dataset. El procedimiento para tratar el resto de atributos que tienen menos del 1% de valores desconocidos se explica en el siguiente apartado.

```
# Seleccionar los primeros 10 atributos y la clase del dataset integrado
original
dataH <- cbind(dataH[, 1:10], dataH$target)
names(dataH)[11] <- "target"
```

Antes de continuar con la siguiente etapa del procesamiento de datos, se establecen dos valores para la variable dependiente *target*. La descripción del dataset original proporciona información de dos valores para la clase (0 = sin dolencia cardiaca y 1 = con presencia de problema cardiaco), sin embargo, no se especifica qué tipo de problema representan 2 y 3. Por tanto, para entender mejor los resultados del análisis se transformará la variable target como una variable binaria:

```
dataH$target <- ifelse(dataH$target != "0", "1", dataH$target)
```

Limpieza de datos

Asignación de tipos de datos

De acuerdo al análisis de la estructura del dataset, ahora se verifica si el archivo se cargó con los tipos de datos adecuados:

```
str(dataH) #  11 variables resultantes y 585 observaciones
```

```
## 'data.frame':    595 obs. of  11 variables:
##  $ age     : num  29 29 30 31 32 32 32 33 34 34 ...
##  $ sex     : num  1 1 0 0 0 1 1 1 0 1 ...
##  $ cp      : num  2 2 1 2 2 2 2 3 2 2 ...
##  $ trestbps: Factor w/ 32 levels "?","100","105",..: 11 21 27 2 3 6 14
11 16 24 ...
##  $ chol    : Factor w/ 153 levels "?","100","117",..: 69 1 64 49 30 55
78 115 8 44 ...
##  $ fbs     : Factor w/ 3 levels "?","0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ restecg : Factor w/ 4 levels "?","0","1","2": 2 2 3 3 2 2 2 2 2 3
...
##  $ thalach : Factor w/ 72 levels "?","100","102",..: 46 53 53 39 49 60
43 61 63 52 ...
##  $ exang   : Factor w/ 3 levels "?","0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ oldpeak : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ target  : chr  "0" "0" "0" "0" ...
```

Como se puede observar en el resumen anterior, algunas variables se cargaron con tipos de datos numéricos (int) en lugar de nominal (factor). Se procede a asignar los datos correctos.

```
dataH$sex <- factor(dataH$sex)
dataH$cp <- factor(dataH$cp)
dataH$fbs <- factor(dataH$fbs)
dataH$restecg <- factor(dataH$restecg)
dataH$exang <- factor(dataH$exang)
dataH$target <- factor(dataH$target)
```

Así mismo, tres variables numéricas se cargaron como factores: trestbps, chol, y thalach. En este caso, las variables contienen valores desconcocidos “?”. Por este motivo, primero se aplicará el tratamiento correspondiente para valores nulos o desconocidos y luego se procederá asignar el tipo de dato correcto.

Tratamiento de valores nulos y desconocidos:

Como se puede observar en la siguiente tabla, 3 atributos: trestbps, chol y thalach tienen valores desconocidos:

```
# Verificar si existen variables con valores desconocidos
sapply(dataH, function(x) sum(is.na(x)))
```

```
##      age      sex       cp trestbps     chol      fbs  restecg
thalach
##        0        0        0       39       81        0        0
94
##    exang  oldpeak   target
##        0        0        0
```

Para tratar estos atributos se aplica la técnica de imputación de valores mediante la función kNN():

```
library(VIM)

# Asignar tipos de datos correctos:
dataH$trestbps[dataH$trestbps == "?"] <- NA
dataH$trestbps <- as.integer(dataH$trestbps)

dataH$chol[dataH$chol == "?"] <- NA
dataH$chol <- as.integer(dataH$chol)

dataH$thalach[dataH$thalach == "?"] <- NA
dataH$thalach <- as.integer(dataH$thalach)

# Imputar aplicando kNN:
dataH$trestbps <- kNN(dataH)$trestbps
dataH$chol <- kNN(dataH)$chol
dataH$thalach <- kNN(dataH)$thalach
```

Para las variables nominales con valores desconocidos "?" se eliminan las filas puesto que como se puede ver corresponden a 10 observaciones, lo que equivale al menos del 2% del total del dataset.

```
NA.nominal.count <- sum(ifelse(dataH$fbs=="?" | dataH$restecg=="?" |
dataH$exang=="?", 1,0)) ; NA.nominal.count  # cantidad de observaciones
que se eliminarían
```

```
## [1] 10
```

```
dataH <- subset(dataH, dataH$fbs!="?" & dataH$restecg!="?" &
dataH$exang!="?")
```

Tratamiento de valores atípicos

Para identificar los valores atípicos de las variables numéricas se realiza la inspección mediante boxplot.stats.

```
boxplot.stats(dataH$age)$out
```

```
## numeric(0)
```

```r
boxplot.stats(dataH$trestbps)$out
```

```
## integer(0)
```

```r
boxplot.stats(dataH$chol)$out
```

```
## integer(0)
```

```r
boxplot.stats(dataH$thalach)$out
```

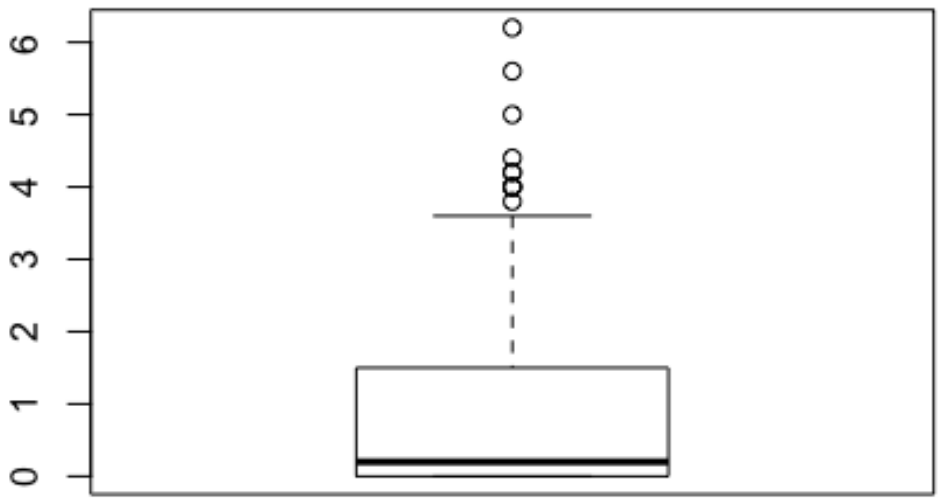
```
## integer(0)
```

```r
boxplot.stats(dataH$oldpeak)$out
```

```
##  [1] 4.0 5.0 6.2 4.0 5.6 4.0 4.2 4.2 3.8 4.4 4.0
```

Según la verificación realizada, la única variable que tendría valores atípicos es oldpeak (depresión ST inducida por el ejercicio en relación con el descanso).

```r
boxplot(dataH$oldpeak)
```



Como se puede observar en la gráfica anterior, son pocos valores atípicos y la diferencia con los valores “normales” no es significativa, por tanto, se opta por no aplicar ningún tratamiento para estos datos.

Análisis de datos

Selección de los grupos de datos que se quieren analizar/comparar

Puesto que el dataset tiene variables cualitativas y cuantitativas, primero se verifican los tipos de datos y se crean dos subconjuntos con cada tipo de dato.

```
typeD <-  data.frame(names(dataH), sapply(dataH, class))  # obtener tipos de datos de cada variable
colnames(typeD) <- c("variable", "typeR")
num.var <- typeD[typeD$typeR == "integer" | typeD$typeR == "numeric", ] # variables numéricas
nom.var <- typeD[typeD$typeR == "factor", ] # variables nominales

#Sub-datasets de variables numéricas y nominales:
num.data <- dataH[, rownames(num.var)]  # obtener datos numéricos
nom.data <- dataH[, rownames(nom.var)]  # obtener datos numéricos
```

En este punto se procede a clasificar los datos con el objetivo de prepararlos para su análisis y dar contestación así las preguntas planteadas.

Comprobación de la normalidad y homogeneidad de la varianza: Normalización de datos

Se utiliza el test de Anderson-Darling para verificar si los valores de las variables cuantitativas provienen de una población distribuida normalmente. Para cada variable numérica se obtiene el p-value, si este valor es superior al nivel de significación ($\alpha = 0.05$), entonces, aceptamos la hipótesis nula: la variable sigue una distribución normal.

```
# Función tomada desde: Teguayco Gutiérrez González (Práctica 2: Limpieza y validación de los datos)

alpha=0.05 # prueba con nivel de significación del 5%
col.nombres = colnames(dataH)
col.nombres

##  [1] "age"      "sex"      "cp"       "trestbps" "chol"     "fbs"
##  [7] "restecg"  "thalach"  "exang"    "oldpeak"  "target"

for(nodo in 1:ncol( dataH ))
{
  if (nodo == 1) cat("Atributos que no sigue una distribucion normal:
\n")
```

```
  if (is.integer(dataH[, nodo]) | is.numeric(dataH[, nodo]))
  {
    p_val = ad.test(dataH[,nodo])$p.value
    if(p_val < alpha)
    {
      cat(" ")
      cat(col.nombres[nodo])
      if(nodo < ncol(dataH) - 1) cat(";  ")
        if(nodo %% 3 == 0) cat("\n")
    }
  }
}

## Atributos que no sigue una distribucion normal:
##  age;   trestbps;   chol;   thalach;   oldpeak
```

Como se puede observar, todas las variables numéricas no siguen una distribución normal.

En cuanto a la comprobación de la homogeneidad de la varianza se aplica el test de Levene, este tipo de test es ideal cuando no se tiene certeza de la distribución de normalidad y permite elegir la medida de tendencia central.

```
library(car)
leveneTest(y = dataH$age, group = dataH$target, center = "median")

## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value Pr(>F)
## group   1  1.8163 0.1783
##       583

leveneTest(y = dataH$trestbps, group = dataH$target, center = "median")

## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value Pr(>F)
## group   1  0.2791 0.5975
##       583

leveneTest(y = dataH$chol, group = dataH$target, center = "median")

## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value Pr(>F)
## group   1  0.0055 0.9412
##       583

leveneTest(y = dataH$thalach, group = dataH$target, center = "median")

## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value  Pr(>F)
## group   1  3.4921 0.06217 .
##       583
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(y = dataH$oldpeak, group = dataH$target, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value    Pr(>F)
## group   1  102.42 < 2.2e-16 ***
##       583
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A partir del test realizado se puede concluir que se puede aceptar la homogeneidad de la varianza para los atributos: presión arterial en reposo (trestbps), colesterol (chol) y ritmo cardíaco máximo alcanzado (thalach).

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.

Primer método de análisis: ¿Cuáles son las variables que más influyen en una enfermedad relacionada al corazón?

En este apartado, se realizan dos tipos de análisis para determinar la correlación entre variables. Considerando que el dataset consta de dos tipos de variables: numéricas y nominales se realiza el análisis determinando la correlación y el análisis de correspondencia.

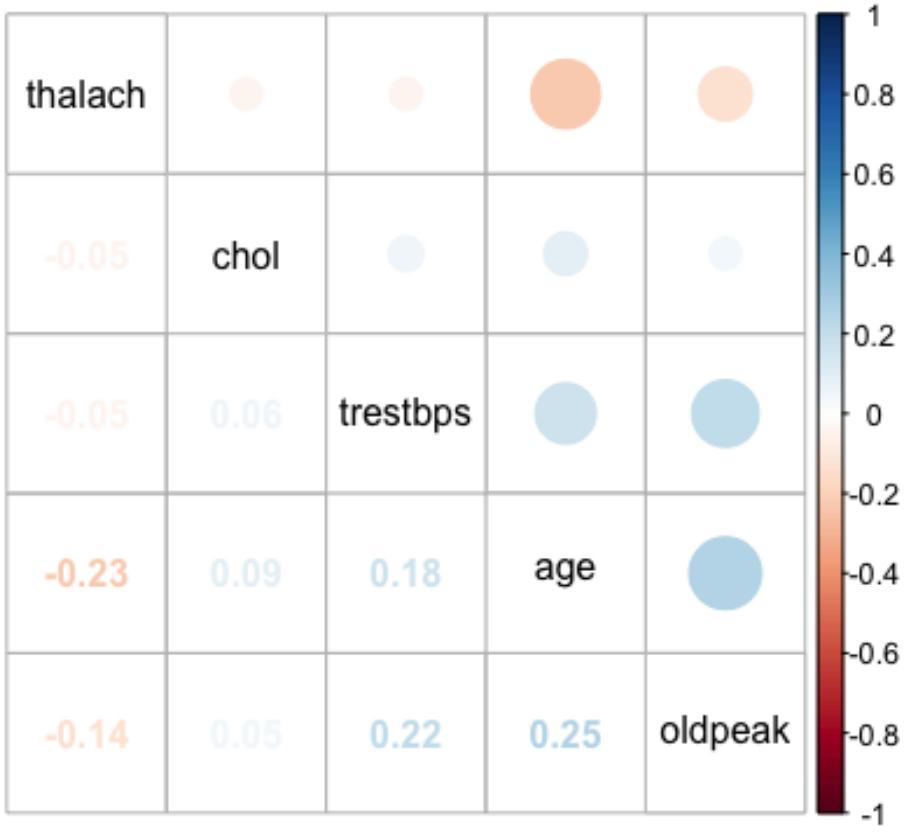
Análisis de correlación

Se mide el nivel de correlación entre las variables independientes:

```
correl <- round(cor(num.data), 3); correl  # Calcular el coeficiente de correlación
```

```
##             age trestbps   chol thalach oldpeak
## age       1.000    0.178  0.090  -0.230   0.252
## trestbps  0.178    1.000  0.059  -0.050   0.216
## chol      0.090    0.059  1.000  -0.047   0.048
## thalach  -0.230   -0.050 -0.047   1.000  -0.137
## oldpeak   0.252    0.216  0.048  -0.137   1.000
```

```
corrplot.mixed(correl, order="hclust", tl.col="black") #Graficar la matriz de correlación
```



En la matriz se puede observar que entre las variables numéricas independientes, el valor más alto de correlación positiva (0.252), se da entre la edad (age) y la depresión ST (oldpeak) del paciente. Por el contrario, el valor más alto de correlación negativa (-0.230) ocurre entre edad (age) y el ritmo cardiaco máximo (thalach) de una persona. Es conclusión, no existe ninguna correlación significativa entre las diferentes variables numéricas, pues ningún valor de la matriz de correlación es superior a 0.7 a inferior a -0.7. Esto se comprueba calculando el p-value para un valor de significación de 0.05.

Ahora se analiza la influencia de las variables independientes en la clase.

```
num.data$targetN <- as.numeric(dataH$target)-1  #  Crear un atributo
numérico equivalente al factor
correl <- cor(num.data[, c(1,2,3,4,5,6)], num.data$targetN) ; correl

##                [,1]
## age        0.2209140
## trestbps   0.1146745
## chol       0.1241796
## thalach   -0.1745995
## oldpeak    0.4867670
## targetN    1.0000000

round(correl^2 * 100, 1)  # coeficiente de determinación r^2
```

```
##           [,1]
## age        4.9
## trestbps   1.3
## chol       1.5
## thalach    3.0
## oldpeak   23.7
## targetN  100.0
```

Luego de presentar la matriz de correlación, se puede observar que existe relación positiva entre todas las variables numéricas, excepto una, y la clase, es decir:

- La edad (age), la presión arterial en reposo (trestbps), el colesterol (chol), y la depresión ST (oldpeak) que se observan en el paciente, influyen positivamente en la existencia de un padecimiento cardíaco. En otras palabras, el riesgo de padecer una enfermedad cardíaca aumenta a medida que se incrementa el nivel de las variables mencionadas (age, trestbps, chol y oldpeak).
- Mientras que el ritmo cardíaco máximo (thalach) alcanzado por una persona disminuye la posibilidad de un mal cardiaco.
- Según el coeficiente de determinación, la variable que más explica la variación de la clase es la depresión ST (oldpeak). Por el contrario, la variable que menos influye en la clase es el colesterol (chol).

Ahora se verifica qué tan significativa es la influencia de las variables sobre la clase:

```
cor.test <- psych::corr.test(num.data[, c(1,2,3,4,5,6)],
num.data$targetN)
cor.test$p # p-value
```

```
##                  [,1]
## age      2.686403e-07
## trestbps 5.489014e-03
## chol     5.246796e-03
## thalach  6.510376e-05
## oldpeak  1.966573e-35
## targetN  0.000000e+00
```

Analizando el p-value, en todos los casos, excepto para chol, su valor es cercano a (0). Siendo este valor menor al nivel de significancia (5%), se puede concluir que la correlación es estadísticamente significativa, excepto para chol.

Análisis de correspondencia

En este punto se aplica esta técnica para identificar la influencia entre cada variable independiente y la clase del dataset. Se utiliza el test chi^2 para verificar si la correspondencia es significativa o no.

```
corr.nom <- function(x){
  r <- NULL
  nom.ct <- table(x, nom.data$target) # tabla de contingencia
  nom.dt <- as.table(as.matrix(nom.ct))
```

```
  chisq <- chisq.test(nom.dt)
  r[1] <- chisq$p.value  # p-value
  return(r)
}

nom.cor <- apply(X = nom.data[, 1:5], MARGIN=2, FUN = function(x)
corr.nom(x))
round(nom.cor, 3) # p-value
```

```
##     sex      cp     fbs restecg   exang
##   0.000   0.000   0.031   0.002   0.000
```

Para todas las cinco variables nominales, el p-value es inferior al nivel de significación 0.05. Es decir, cada variable (sex, cp, FBS, restecg y exang) explican la presencia de un mal cardiaco.

En conclusión, luego del análisis de correlación y correspondencia, la única variable que no explicaría de manera significativa la variabilidad de la clase es chol (colesterol).

Segundo método de análisis: ¿El tipo de dolor torácico del paciente (cp) es un indicativo de posible problema del corazón?

```
# selección de grupos de datos
prop.table(table(dataH$cp, dataH$target))
```

```
##
##              0          1
##   1 0.03760684 0.01709402
##   2 0.22905983 0.02905983
##   3 0.18632479 0.04957265
##   4 0.13333333 0.31794872
```

En la tabla de proporción de casos se puede observar que el tipo de dolor torácico si tiene incidencia en problemas de corazón.

```
cp.anginaTipica <- dataH[dataH$cp =="1", ]$target
prop.table(table(cp.anginaTipica))
```

```
## cp.anginaTipica
##      0      1
## 0.6875 0.3125
```

```
cp.anginaAtipica <- dataH[dataH$cp =="2", ]$target
prop.table(table(cp.anginaAtipica))
```

```
## cp.anginaAtipica
##         0         1
## 0.8874172 0.1125828
```

```
cp.dolorNoAnginal <- dataH[dataH$cp =="3", ]$target
prop.table(table(cp.dolorNoAnginal))
```

```
## cp.dolorNoAnginal
##         0         1
## 0.7898551 0.2101449
```

```
cp.asintomatico <- dataH[dataH$cp =="4", ]$target
prop.table(table(cp.asintomatico))
```

```
## cp.asintomatico
##         0         1
## 0.2954545 0.7045455
```

En conclusión, un paciente sin dolor (cp = 4 o asintomático) tiene menos probabilidad de tener una dolencia cardíaca.

Tercer método de análisis: ¿Qué tan efectivo es un modelo de regresión logística para predecir padecimiento del corazón en el caso de hombres y de mujeres?

En este caso se utilzará la Regresión Logística Simple para estimar la probabilidad en la que influye el sexo en la presencia o no de un padecimiento cardiaco. Al aplicar este método, además se podrá verificar cuáles de los valores de las variables que parecen tener mayor incidencia en una enfermedad cardíaca.

Primero se crean dos subconjuntos de datos para cada género y se obtienen los respectivos conjuntos de datos para entrenamiento y prueba del modelo de regresión lineal. Luego de crear el modelo logístico se probará qué tan efectivo es para predecir casos de hombres y de mujeres.

```
h.heartd <- dataH[dataH$sex =="1", ]
m.heartd <- dataH[dataH$sex =="0", ]

#Elegir conjunto de datos para la creación del modelo logístico
set.seed(123)
#selección de casos de sexo = hombre:
indexesH = sample(1:nrow(h.heartd), size=floor((0.9)*nrow(h.heartd)))
trainH<-h.heartd[indexesH,]
testH<-h.heartd[-indexesH, ]

#selección de casos de sexo = mujer:
indexesM = sample(1:nrow(m.heartd), size=floor((0.9)*nrow(m.heartd)))
trainM<-m.heartd[indexesM,]
testM<-m.heartd[-indexesM, ]

#Union de datasets
train <- rbind(trainH, trainM)
test <- rbind(testH, testM)

#Crear modelo
model.lg = glm(target ~., family = binomial(logit), data = train)
summary(model.lg)
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial(logit), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6538  -0.5523  -0.2528   0.5415   2.4121
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.317693   1.205970  -4.409 1.04e-05 ***
## age          0.028867   0.015429   1.871 0.061342 .
## sex1         1.325453   0.299060   4.432 9.33e-06 ***
## cp2         -0.228724   0.576852  -0.397 0.691734
## cp3          0.037254   0.539847   0.069 0.944983
## cp4          1.798826   0.537731   3.345 0.000822 ***
## trestbps    -0.006543   0.019594  -0.334 0.738412
## chol         0.004472   0.003819   1.171 0.241687
## fbs1         0.545094   0.395794   1.377 0.168446
## restecg1    -0.193830   0.481182  -0.403 0.687080
## restecg2     0.396903   0.290274   1.367 0.171519
## thalach      0.008049   0.008796   0.915 0.360143
## exang1       1.069873   0.287457   3.722 0.000198 ***
## oldpeak      0.860146   0.147897   5.816 6.03e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 715.26  on 524  degrees of freedom
## Residual deviance: 424.83  on 511  degrees of freedom
## AIC: 452.83
##
## Number of Fisher Scoring iterations: 5
```

A partir del resumen del modelo creado se pueden destacar los siguientes hechos:

- Siendo el p-value menor a 0.05 para **age, cp = {4}, fbs = {1}, restecg1 = {2}, y oldpeak**, por tanto, éstas variables y los respectivos dominios mencionados tienen una influencia significativa en el padecimiento de corazón de un paciente.
- Por el contrario, el género (age), cp = {1,2, 3}, presión arterial en reposo(trestbps), colesterol (chol), azúcar en la sangre en ayunas menor a 120 mg/dl (fbs=0), resultados electrocardiográficos en reposo anormal (restecg={1}), ritmo cardíaco máximo alcanzado (thalach), y angina inducida por el ejercicio (exang=1) no tienen tanta influencia en el padecimiento del corazón.

Ahora se pone a prueba el modelo, verificando su precisión para cada sexo:

```
#Prediciendo para hombres:
predict.lgH = predict(model.lg, testH,type='response')
```

```r
predict.lgH <- ifelse(predict.lgH > 0.5,1,0)
misClasificErrorH <- mean(predict.lgH != testH$target)
print(paste('Precisión para hombres:',1-misClasificErrorH))
```

```
## [1] "Precisión para hombres: 0.833333333333333"
```

```r
#Prediciendo para mujeres:
predict.lgM = predict(model.lg, testM,type='response')
predict.lgM <- ifelse(predict.lgM > 0.5,1,0)
misClasificErrorM <- mean(predict.lgM != testM$target)
print(paste('Precisión para mujeres: ',1-misClasificErrorM))
```

```
## [1] "Precisión para mujeres:  0.944444444444444"
```

Como se puede observar el modelo de regresión logística simple resulta mucho más efectivo para predecir si una mujer padecerá una enfermedad cardiaca (100% de precisión) en base a todas las variables independientes consideradas.

Ahora de lo analizado previamente, 3 variables tienen una influencia significativa en el padecimiento cardiaco: age, cp, restecg y oldpeak. Se creará un modelo más simple basado en estas variables y considerando únicamente datos de mujeres.

```r
model.lgM = glm(target ~ age+cp+restecg+oldpeak, family =
binomial(logit), data = trainM)
summary(model.lgM)
```

```
##
## Call:
## glm(formula = target ~ age + cp + restecg + oldpeak, family =
binomial(logit),
##     data = trainM)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.77268  -0.56602  -0.32298  -0.00011   2.40568
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -20.98054 1164.54873  -0.018  0.98563
## age            0.03930    0.02751   1.428  0.15318
## cp2           16.12264 1164.54755   0.014  0.98895
## cp3           16.40284 1164.54752   0.014  0.98876
## cp4           17.86487 1164.54743   0.015  0.98776
## restecg1      -0.38093    0.78158  -0.487  0.62598
## restecg2       0.02750    0.55285   0.050  0.96032
## oldpeak        0.96951    0.30701   3.158  0.00159 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 168.02  on 154  degrees of freedom
## Residual deviance: 114.20  on 147  degrees of freedom
## AIC: 130.2
##
## Number of Fisher Scoring iterations: 16
```

A partir del resumen del modelo, se evalúa la capacidad del modelo para predecir:

```
# Probar el modelo con datos de mujeres:
predict.lgM = predict(model.lgM, testM, type='response')

predict.lgM <- ifelse(predict.lgM > 0.5,1,0)
table(testM$target, predict.lgM)

##    predict.lgM
##      0  1
##   0 16  1
##   1  0  1

misClasificErrorM <- mean(predict.lgM != testM$target)
print(paste('Precisión para mujeres: ',1-misClasificErrorM))

## [1] "Precisión para mujeres:  0.944444444444444"
```

Como se puede observar, considerando únicamente cuatro predictores (age, cp, restecg y oldpeak) se puede determinar si una mujer padecerá alguna dolencia cardiaca con el 100% de efectividad. En conclusión, si una mujer tiene una edad avanzada (age), mayor depresión ST inducida por el ejercicio (oldpeak) es más suceptible de padecer una dolencia cardiaca.

Se pone a prueba este modelo, efectivo para mujeres, para determinar su rendimiento en el conjunto de datos de prueba de hombres:

```
# Probar el modelo con datos de hombres:
predict.lgH = predict(model.lgM, testH,type='response')
predict.lgH <- ifelse(predict.lgH > 0.5,1,0)
misClasificErrorH <- mean(predict.lgH != testH$target)
print(paste('Precisión para hombres: ',1-misClasificErrorH))

## [1] "Precisión para hombres:  0.69047619047619"
```

La reducción de la precisión para predecir la clase en el caso de hombres, indica que se deben tomar en cuenta otras variables de las utilizadas en este segundo modelo. Aunque el rendimiento no podría ser mejorado significativamente porque como se vió previamente, considerando una mayor cantidad de datos, máximo se podría alcanzar el 83% de la efectividad para hombres.

Representación de los resultados a partir de tablas y gráficas

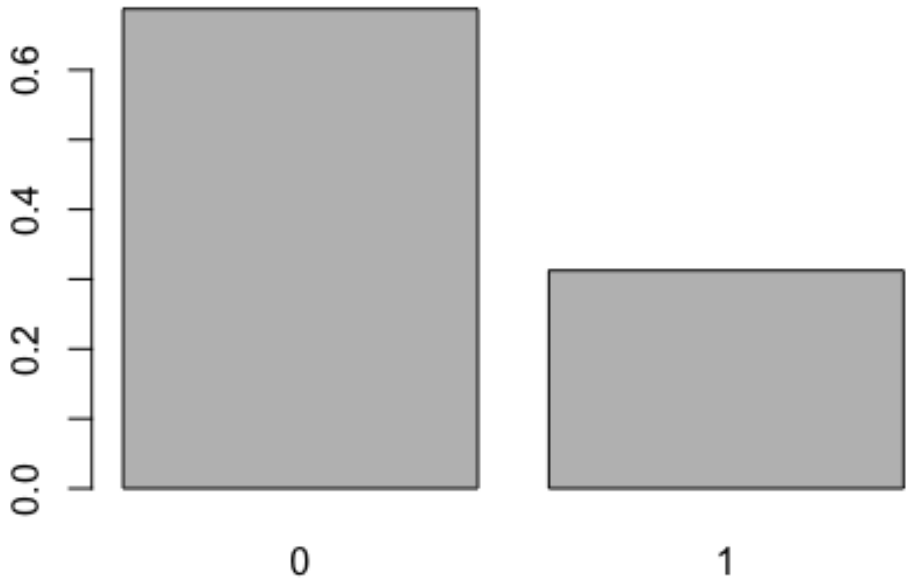
---

Influencia del dolor torácico (cp)

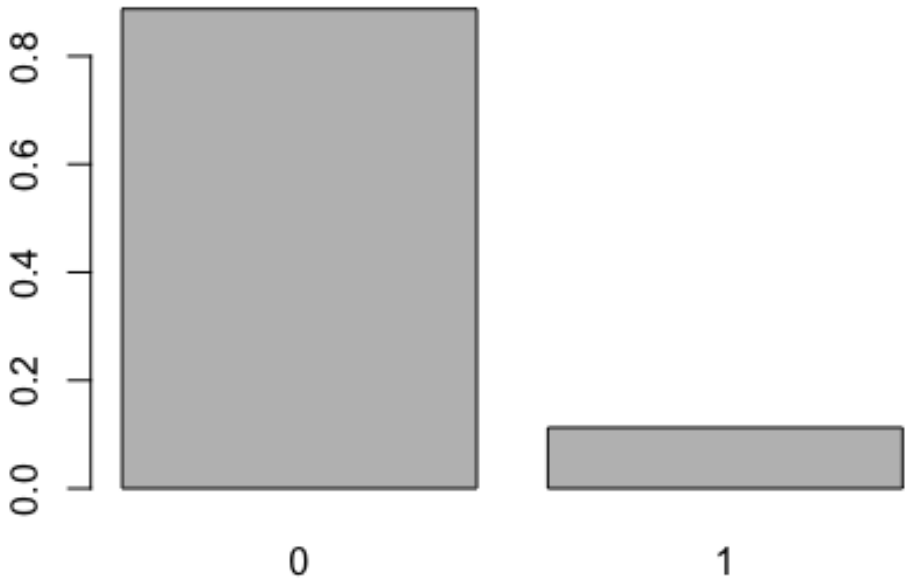
Como se analizó previamente, el tipo de dolor que padezca el paciente explica en cierta medida la presencia o no de un padecimiento del corazón.

En las siguientes gráficas se puede observar que el hecho de que un paciente no presente dolor (cp=4) tiene menos incidencia en la clase = 1.

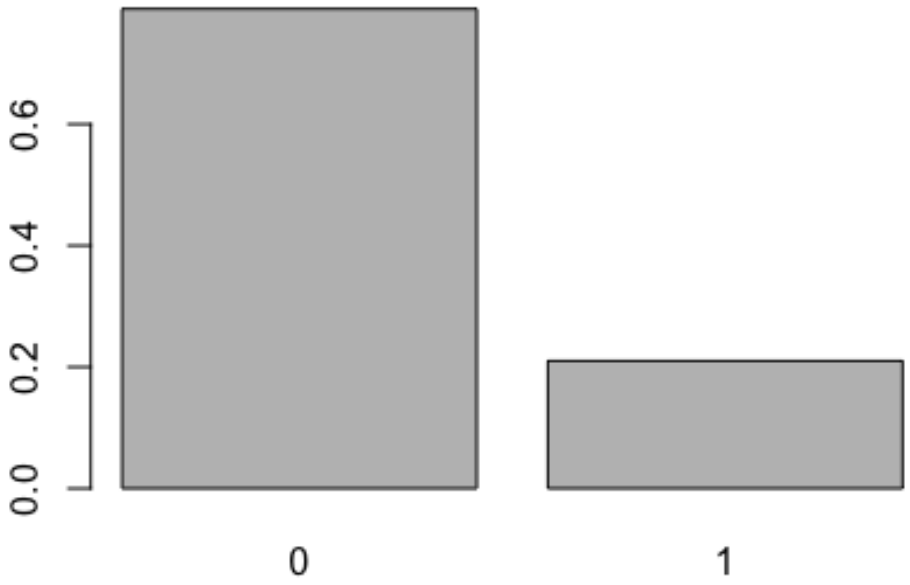
```
barplot(prop.table(table(cp.anginaTipica)))
```



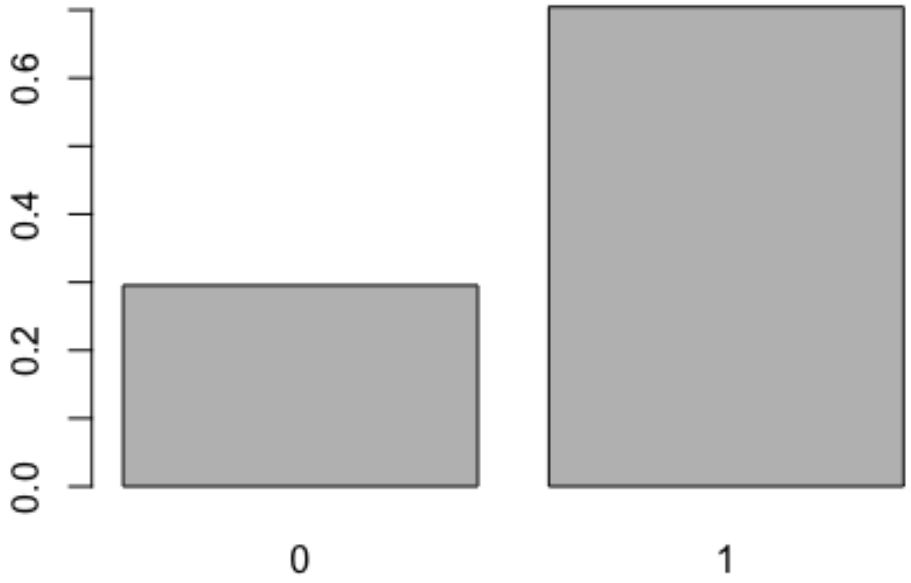

```
barplot(prop.table(table(cp.anginaAtipica)))
```




```
barplot(prop.table(table(cp.dolorNoAnginal)))
```



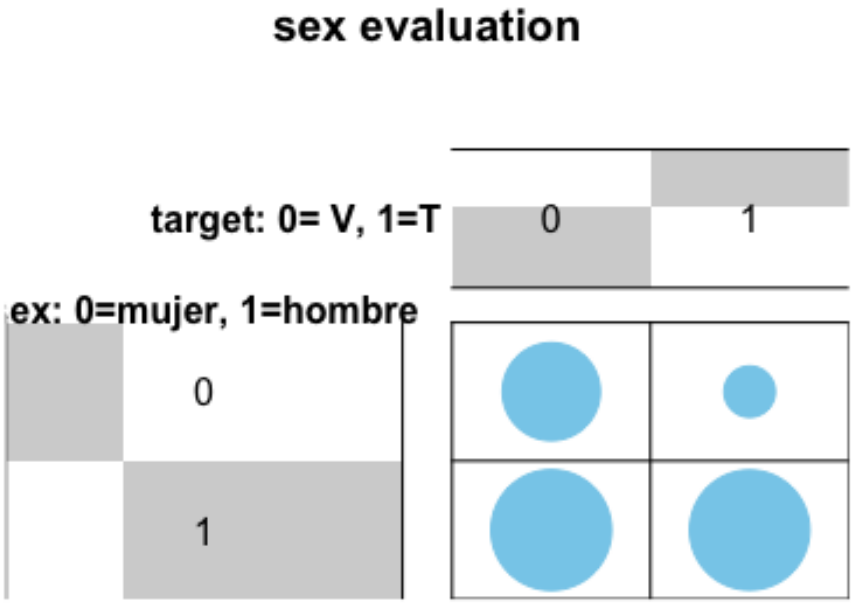

```
barplot(prop.table(table(cp.asintomatico)))
```



Influencia del sexo (sex)

A continuación, se verifica visualmente la influencia entre el sexo y la clase del dataset.

```
# sex
sex.ct <- table(nom.data$sex, nom.data$target)
gplots::balloonplot(t(as.table(as.matrix(sex.ct))), main ="sex
evaluation", xlab ="target: 0= V, 1=T", ylab="sex: 0=mujer, 1=hombre",
            label = FALSE, show.margins = FALSE)
```



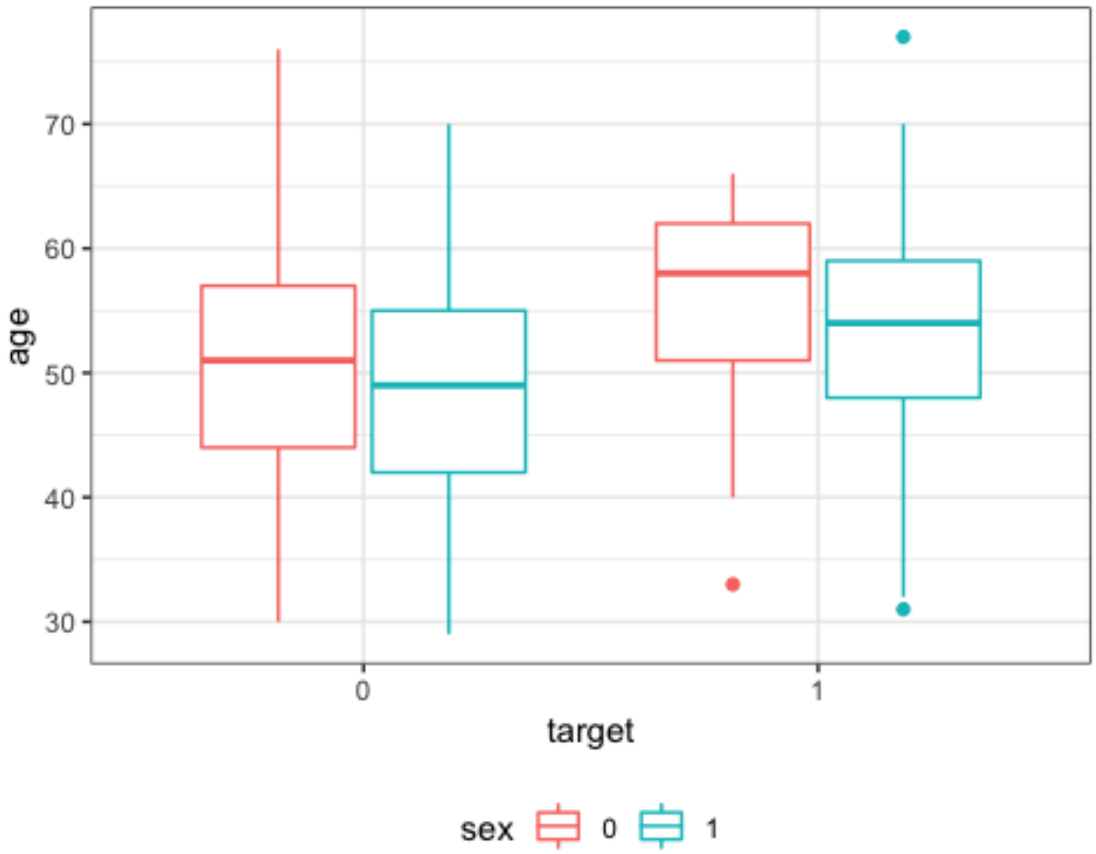

```
 prop.table(table(nom.data$sex, nom.data$target))
```

```
##
##               0          1
##   0 0.23247863 0.06324786
##   1 0.35384615 0.35042735
```

Influencia de edad (age) y sexo (sex)

En la siguiente gráfica se puede comprobar que un padecimiento del corazón (target = 1) influye de diferente manera a hombres y mujeres, dependiendo de su edad.

```
ggplot(data = dataH, aes(x = target , y = age, colour = sex)) +
  geom_boxplot() +
  theme_bw() +
  theme(legend.position = "bottom")
```



En el caso de los hombres (sex=1) con padecimiento de enfermedad cardiaca (target = 1), el espectro de su edad es más amplio que en el caso de las mujeres. En el caso de las mujeres se concentra el mal entre los 56 y 62 años, mientras que en el de los hombres entre los 53 años a 63.

Resolución del problema

A partir de las tres preguntas planteadas, se puede concluir que:

1. ¿Cuáles son las variables que más influyen en una enfermedad relacionada al corazón?. A partir del análisis de correlación y correspondencia se puede establecer que cada uno de los 10 atributos explican en cierta proporción la variabilidad de la clase. El único atributo menos influyente sería el colesterol.
2. ¿El tipo de dolor torácico del paciente (cp) es un indicativo de posible problema del corazón?. A través del análisis de cada tipo de dolor (cp) que adolece el paciente, se comprobó que si el paciente no presenta dolor (asintomático) existen menos probabilidades de que esté padeciendo de un problema del corazón.

3. ¿Qué tan efectivo es un modelo de regresión logística para predecir el padecimiento del corazón en el caso de hombres y de mujeres? A través de un modelo de regresión se pudo probar la efectividad del modelo para predecir la clase por sexo. En el caso de las mujeres, el modelo alcanza el 100% de precisión, mientras que en el caso de hombres llega a un poco más del 80%, es decir, en el caso de los hombres pueden existir otras variables que no se tomaron en cuenta en el análisis y que inciden en el estado de su salud cardiaca.

Código

En el sitio GitHub: https://github.com/jachicaiza/prac02 está disponible el código, conjuntos de datos de origen utilizados para la presente práctica.

Además, los datos preprocesados también pueden ser descargados y guardados:

```
write.csv(dataH, file = "data/heart.processed.csv")
```

References