¿Cómo realizar la limpieza y análisis de datos?

Miguel Ángel Quesada Fernández

17 de junio, 2023

Contents

En esta práctica se va a abordar las primeras fases de un proyecto de minería de datos, para ello se va a utilizar la metodología CRISP-DM que proporcionado un modelo estructurado basado en 6 etapas con dependencia entre sí. En esta práctica se van a abordar únicamente las cuatro primeras etapas, haciendo especialmente enfases en las tres primeras, y abordando una pequeña modelización de un modelo de regresión, sin profundizar en sus resultados como parte del análisis de los datos:

- ETAPA 1 – Determinar los objetivos de negocio y análisis de la situación actual
- ETAPA 2 – Comprensión de los datos
- ETAPA 3 – Preparación de los datos
- ETAPA 4 – Modelado

1 Determinar los objetivos de negocio y análisis de la situación actual

La primera fase trata de entender el negocio, estableciendo los requisitos y los objetivos del proyecto tanto a nivel empresarial como técnico para posteriormente trasladarlos a objetivos técnicos y un plan de proyecto.

Como resultado de esta fase:

- Se realizará un análisis de la situación actual
- Se establecerán los requisitos, restricciones y asunciones del proyecto
- Se establecerá la pregunta objetivo y los resultados esperados del proyecto

1.1 Análisis de la situación actual

Se dispone de un conjunto de datos fechados en 1988 que contienen información de pacientes de Cleveland, Hungria, Suiza y Long Beach V. El conjunto original de datos disponía inicialmente 76 variables, pero todos los estudios posteriores se han referido exclusivamente a un subconjunto de sólo 14 variables. Una de estas variables es la variable objetivo, que determina la presencia de alguna enfermedad de corazon en el paciente.

El conjunto inicial de datos puede encontrarse en la plataforma kaggle: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

1.2 Requisitos, restricciones y asunciones

A continuación, se retallan los requisitos, restricciones y asunciones que se van a realizar.

- Los datos necesarios serán los obtenidos en el dataset de kaggle proporcionado para la práctica: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.
- Se ha seleccionado un dataset que tuviese capacidades de aplicar un modelo supervisado y no supervisado, y además con garantías de poder generar un modelo que pueda predecir con precisión.

1.3 ¿Qué pregunta se está intentando responder?

Conocidas una serie de variables del paciente queremos determinar **si el paciente presenta una enfermedad del corazon o no**.

En concreto, queremos analizar la presencia de enfermedades cardíacas atendiendo a variables como la edad, sexo, presión arterial en reposo, y colesterol.

1.4 ¿Cuáles son los resultados esperados del proyecto?

En esta práctica se van a abordar principalmente cuatro objetivos:

- Se va a plantear un problema de analítica de datos detallando los objetivos analíticos y explicando la metodologia adecuada.
- Se va a acondicionar el juego de datos.
- Se va a intentar contrastar diversas hipotesis.

Como parte del análisis se realizará los siguientes puntos:

- Selección de algoritmo de regresión
- Generación de modelo de regresión capaz de generalizar los datos

Queda fuera del alcance de esta práctica los siguientes puntos que se abordarían siguiendo la metodología CRISP-DM: - Evaluación del modelo - Implantación

2 Comprensión de los datos y preparación de los datos

En esta fase se realiza la recolección y exploración inicial de los datos a fin de tener un primer contacto con el problema. Un mal entendimiento de los datos puede tener como consecuencia un aumento del tiempo del proyecto y una peor calidad del modelo generado.

Como consecuencia de esta fase se realizarán las siguientes tareas:

2.1 Captura de los datos

```
# Leemos el fichero de datos
heartDiseaseDS <- read.csv('heart.csv')
```

2.2 Descripción de los datos

Queremos hacer una primera aproximación al conjunto de los datos realizando una descripción de las variables, su tipología, como se distribuyen los valores...

```
# Obtenemos la estructura del DataSet
estructura = str(heartDiseaseDS, list.len=ncol(heartDiseaseDS))
```

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

Observamos que disponemos de 303 observaciones con 14 variables, siendo una de esas variables la variable objetivo.

Echamos un vistazo preliminar a los datos para ver su contenido y estructura:

```
head(heartDiseaseDS, 5)
```

```
##   age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output
## 1  63   1  3    145  233   1       0      150    0     2.3   0   0     1      1
## 2  37   1  2    130  250   0       1      187    0     3.5   0   0     2      1
## 3  41   0  1    130  204   0       0      172    0     1.4   2   0     2      1
## 4  56   1  1    120  236   0       1      178    0     0.8   2   0     2      1
## 5  57   0  0    120  354   0       1      163    1     0.6   2   0     2      1
```

A continuación, identificamos las observaciones y construimos un pequeño diccionario de datos utilizando la documentación proporcionada en la página de kaggle del dataset.

El conjunto de datos dispone de:

- 5 variables númericas
- 4 variables binarias
- 4 variables categoricas
- 1 variable objetico de tipo binaria

Se describen las variables:

- **AGE**: Edad del paciente *(Variable númerica)*
- **SEX**: Sexo del paciente *(Variable binaria)*
  - 0 - Mujer
  - 1 - Hombre
- **CP**: Tipo de dolor en el pecho del paciente. *(Variable categórica)*
  - 0 - Angina típica
  - 1 - Angina atípica
  - 2 - Dolor no anginal
  - 3 - Asíntomática
- **TRTBPS**: Presión sanguínea en reposo (en mg Hg en admisión al hospital) *(Variable númerica)*
- **CHOL**: Colesterol (mg/dl) *(Variable númerica)*
- **FBS**: Azucar en sangre en ayunas ($>$ 120mg/dl) *(Variable binaria)*
  - 0 - No
  - 1 - Si
- **RESTECG**: Resultados del electrócardiograma en reposo. *(Variable binaria)*
  - 0 - Normal
  - 1 - Onda ST-T anormal
  - 2 - Hipertrofía ventriculo izquierdo

- **THALACHH**: Pulso corazón máximo registrado *(Variable númerica)*
- **EXNG**: Angina inducida por ejercicio *(Variable binaria)*
  - 0 - No
  - 1 - Si
- **OLDPEAK**:Depresión ST inducida por el ejercicio en relación con el descanso *(Variable númerica)*
- **SLP**: Pendiente del segmento ST de ejercicio máximo *(Variable categórica)*
  - 0 - Pendiente de subida
  - 1 - Plano
  - 2 - Pendiente de bajada
- **CAA**: Número de vasos principales en el corazón observados con Fluoroscopia *(Variable categorica)*
  - 0
  - 1
  - 2
  - 3
- **THALL**: *(Variable categórica)*
  - 0 - Normal
  - 1 - Defecto permanente
  - 2 - Defecto reversible
- **OUTPUT**: *(Variable binaria)*
  - 0 - Presenta enfermedad cardiaca
  - 1 - No presenta enfermedad cardiaca

2.3 Exploración de los datos mediante técnicas descriptivas

Determinación de la consistencia, cantidad y distribución de valores nulos o fuera de rango. Con esta información se obtendrá un informe de calidad de los datos que listará los resultados de las verificaciones realizadas, y sugerirá los posibles tratamientos para mejorar la calidad de los datos. Estos tratamientos se realizarán en la siguiente fase.

Empezamos obteniendo una serie de datos estadísticos simples (media, moda, quartiles...). Como podemos observar todas las variables se han identificado de tipo númerico, calculandose valores como la media que no tienen sentido en las variables de tipo categórico.

```
# Observamos variables estadísticas del conjunto de datos
summary(heartDiseaseDS)
```

```
##       age             sex              cp            trtbps
##  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
##       chol            fbs            restecg          thalachh
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0
```

```
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##       exng           oldpeak          slp             caa
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000
##      thall           output
##  Min.   :0.000   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:0.0000
##  Median :2.000   Median :1.0000
##  Mean   :2.314   Mean   :0.5446
##  3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.0000
```

2.4 Outliers

Tras una primera inspección podemos observar como algunos valores están fuera del rango contemplado para el dataset, presuponiendo que se trata de muestras de mala calidad, vamos a proceder a su eliminación:

```
# Eliminamos aquellas muestras cuyo valor de "ca" no está en el rango de 0 a 3
heartDiseaseDS<-heartDiseaseDS[(heartDiseaseDS$caa>=0 & heartDiseaseDS$caa<=3),]

# Eliminamos aquellas muestras cuyo valor de "thall" no está en el rango de 0 a 2
heartDiseaseDS<-heartDiseaseDS[(heartDiseaseDS$thall>=1 & heartDiseaseDS$thall<3),]
```

2.5 Nulos y blancos

El siguiente paso será la limpieza de datos, mirando si hay valores vacíos o nulos.

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

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

```
colSums(is.na(heartDiseaseDS))
```

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0        0        0        0
##     exng  oldpeak      slp      caa    thall   output
##        0        0        0        0        0        0
```

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

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

```
colSums(heartDiseaseDS=="",)
```

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0        0        0        0
##     exng  oldpeak      slp      caa    thall   output
##        0        0        0        0        0        0
```

Observamos que no hay valores nulos ni blancos en el conjunto de datos. Por lo que no es preciso realizar ningún tipo de acción para eliminar afectados.

2.6 Factorización

A continuación, procedemos a factorizar aquellas variables que se han identificado de tipo númerico pero que realmente son variables de tipo categorico.

```
# Reemplazamos las variables categóricas / binarias por su significado

# Sexo
heartDiseaseDS$sex <- replace(heartDiseaseDS$sex, heartDiseaseDS$sex =="0", "Mujer")
heartDiseaseDS$sex <- replace(heartDiseaseDS$sex, heartDiseaseDS$sex =="1", "Hombre")

# Dolor en pecho
heartDiseaseDS$cp <- replace(heartDiseaseDS$cp, heartDiseaseDS$cp =="0",
                             "Angina tipica")
heartDiseaseDS$cp <- replace(heartDiseaseDS$cp, heartDiseaseDS$cp =="1",
                             "Angina atipica")
heartDiseaseDS$cp <- replace(heartDiseaseDS$cp, heartDiseaseDS$cp =="2",
                             "Dolor no anginal")
heartDiseaseDS$cp <- replace(heartDiseaseDS$cp, heartDiseaseDS$cp =="3",
                             "Asíntomática")

# Azúcar en sangre
heartDiseaseDS$fbs <- replace(heartDiseaseDS$fbs, heartDiseaseDS$fbs =="0", "No")
heartDiseaseDS$fbs <- replace(heartDiseaseDS$fbs, heartDiseaseDS$fbs =="1", "Si")

# Resultados electrócardiograma en reposo
heartDiseaseDS$restecg <- replace(heartDiseaseDS$restecg, heartDiseaseDS$restecg =="0",
                                  "Normal")
heartDiseaseDS$restecg <- replace(heartDiseaseDS$restecg, heartDiseaseDS$restecg =="1",
                                  "Onda ST-T anormal")
heartDiseaseDS$restecg <- replace(heartDiseaseDS$restecg, heartDiseaseDS$restecg =="2",
                                  "Hipertrofía ventriculo izquierdo")

# Angina inducida por el ejercicio
heartDiseaseDS$exng <- replace(heartDiseaseDS$exng, heartDiseaseDS$exng =="0", "No")
heartDiseaseDS$exng <- replace(heartDiseaseDS$exng, heartDiseaseDS$exng =="1", "Si")

# Pendiente segmento ST de ejercicio máximo
heartDiseaseDS$slp <- replace(heartDiseaseDS$slp, heartDiseaseDS$slp =="0",
                              "Pendiente de subida")
heartDiseaseDS$slp <- replace(heartDiseaseDS$slp, heartDiseaseDS$slp =="1",
                              "Plano")
heartDiseaseDS$slp <- replace(heartDiseaseDS$slp, heartDiseaseDS$slp =="2",
                              "Pendiente de bajada")
```

```
# Thal
heartDiseaseDS$thall <- replace(heartDiseaseDS$thall, heartDiseaseDS$thall =="1",
                                "Normal")
heartDiseaseDS$thall <- replace(heartDiseaseDS$thall, heartDiseaseDS$thall =="2",
                                "Defecto permanente")
heartDiseaseDS$thall <- replace(heartDiseaseDS$thall, heartDiseaseDS$thall =="3",
                                "Defecto reversible")

# Target
heartDiseaseDS$output <- replace(heartDiseaseDS$output, heartDiseaseDS$output =="0",
                                 "Con enfermedad cardiaca")
heartDiseaseDS$output <- replace(heartDiseaseDS$output, heartDiseaseDS$output =="1",
                                 "Sin enfermedad cardiaca")

# Convertimos las variables categoricas
heartDiseaseDS$sex <- as.factor(heartDiseaseDS$sex)
heartDiseaseDS$cp <- as.factor(heartDiseaseDS$cp)
heartDiseaseDS$fbs <- as.factor(heartDiseaseDS$fbs)
heartDiseaseDS$restecg <- as.factor(heartDiseaseDS$restecg)
heartDiseaseDS$exng <- as.factor(heartDiseaseDS$exng)
heartDiseaseDS$caa <- as.factor(heartDiseaseDS$caa)
heartDiseaseDS$thall <- as.factor(heartDiseaseDS$thall)
heartDiseaseDS$slp <- as.factor(heartDiseaseDS$slp)
heartDiseaseDS$output <- as.factor(heartDiseaseDS$output)
```

Volvemos a representar la estructura, ahora las variables categoricas / binarias se representan correctamente, con el número de elementos de cada clase, mientras que las variables númericas se representan con sus datos estadísticos de media, desviación, percentiles...

```
# Observamos variables estadísticas del conjunto de datos
summary(heartDiseaseDS)
```

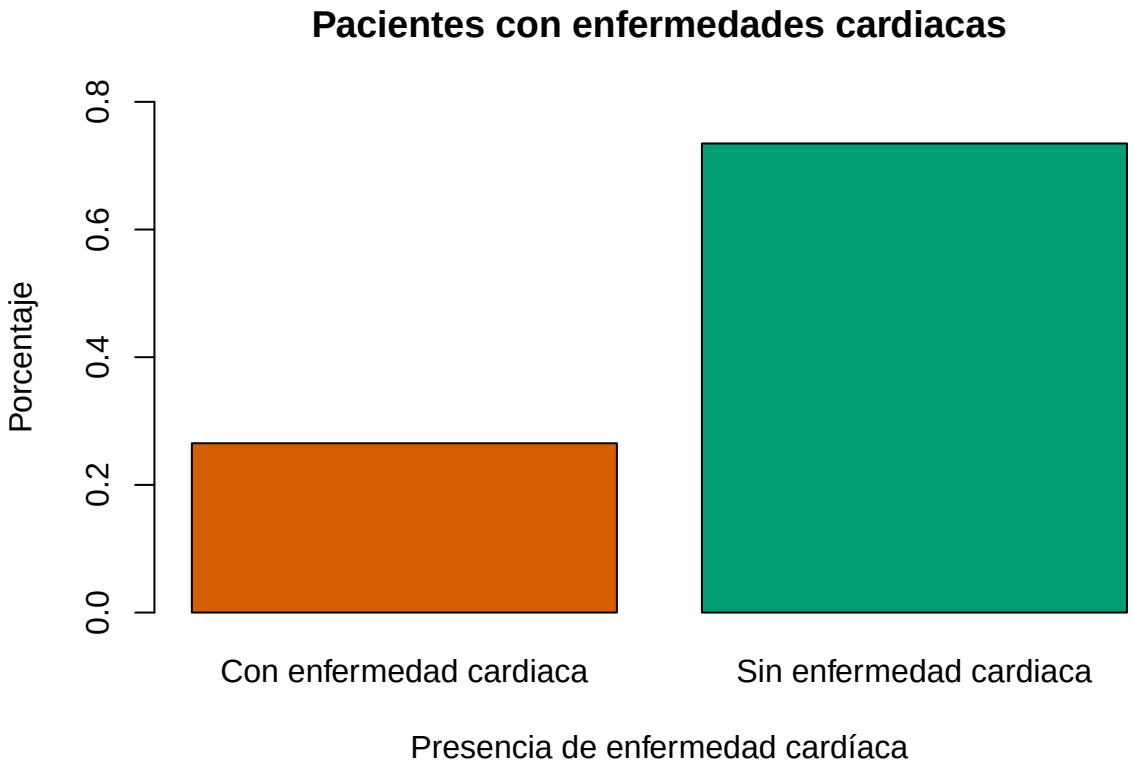
```
##       age            sex                      cp         trtbps
##  Min.   :29.00   Hombre:101   Angina atipica  :41   Min.   : 94
##  1st Qu.:45.00   Mujer : 80   Angina tipica   :64   1st Qu.:120
##  Median :54.00                Asíntomática    :15   Median :130
##  Mean   :53.76                Dolor no anginal:61   Mean   :130
##  3rd Qu.:61.00                                      3rd Qu.:140
##  Max.   :77.00                                      Max.   :180
##       chol       fbs                                  restecg      thalachh
##  Min.   :141.0   No:156   Hipertrofía ventriculo izquierdo: 3   Min.   : 71.0
##  1st Qu.:210.0   Si: 25   Normal                           :88   1st Qu.:142.0
##  Median :242.0            Onda ST-T anormal                :90   Median :158.0
##  Mean   :245.1                                                   Mean   :153.5
##  3rd Qu.:273.0                                                   3rd Qu.:170.0
##  Max.   :417.0                                                   Max.   :202.0
##  exng        oldpeak                        slp      caa
##  No:142   Min.   :0.0000   Pendiente de bajada:101   0:122
##  Si: 39   1st Qu.:0.0000   Pendiente de subida: 11   1: 33
##           Median :0.4000   Plano              : 69   2: 18
##           Mean   :0.7773                             3:  8
##           3rd Qu.:1.4000
##           Max.   :4.4000
```

```
##                 thall                          output
##  Defecto permanente:163   Con enfermedad cardiaca: 48
##  Normal             : 18   Sin enfermedad cardiaca:133
##
##
##
##
```

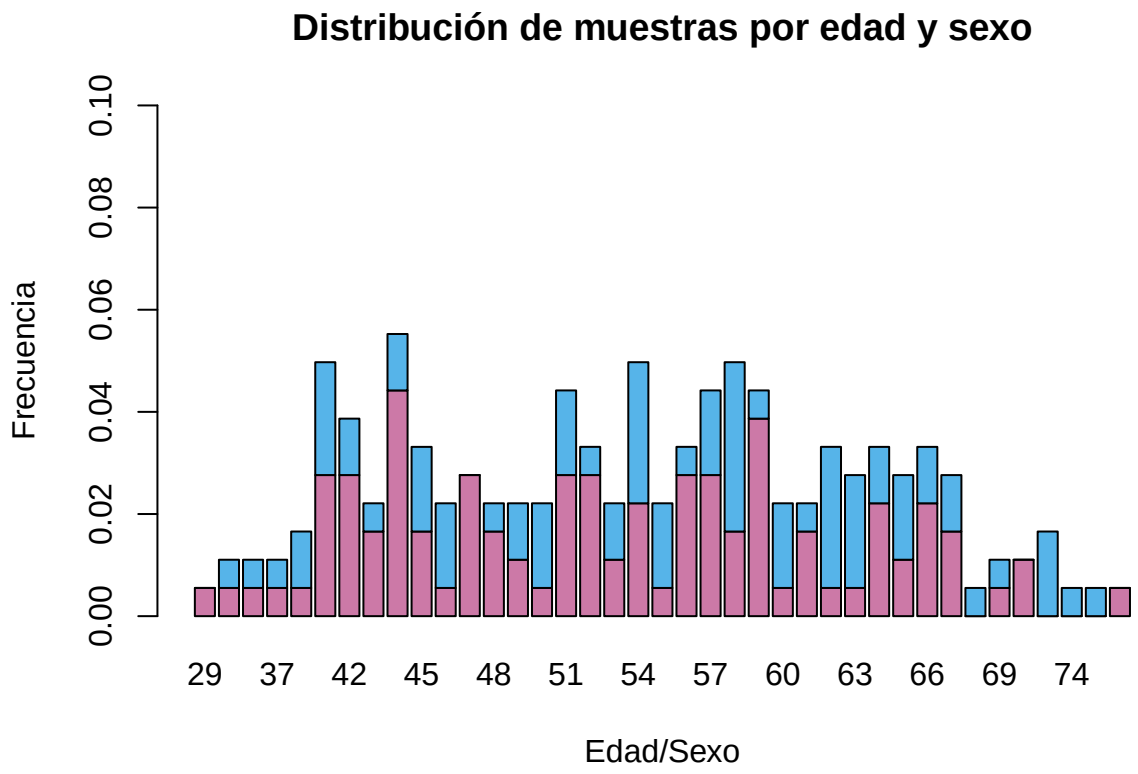
2.7 Estudio de los datos mediante gráficos

En este apartado vamos a generar histogramas y gráficos que ayuden a entender mejor el significado de los datos.
En primer lugar, observamos que la variable destino NO esta balanceada:

```
barplot(prop.table(table(heartDiseaseDS$output)),col=c("#D55E00", "#009E73"),
        main="Pacientes con enfermedades cardiacas",
        legend.texto=c("0 - Sin enfermedad cardíaca","1 - Con enfermedad cardíaca"),
        xlab ="Presencia de enfermedad cardíaca", ylab = "Porcentaje",ylim=c(0,0.8) )
```



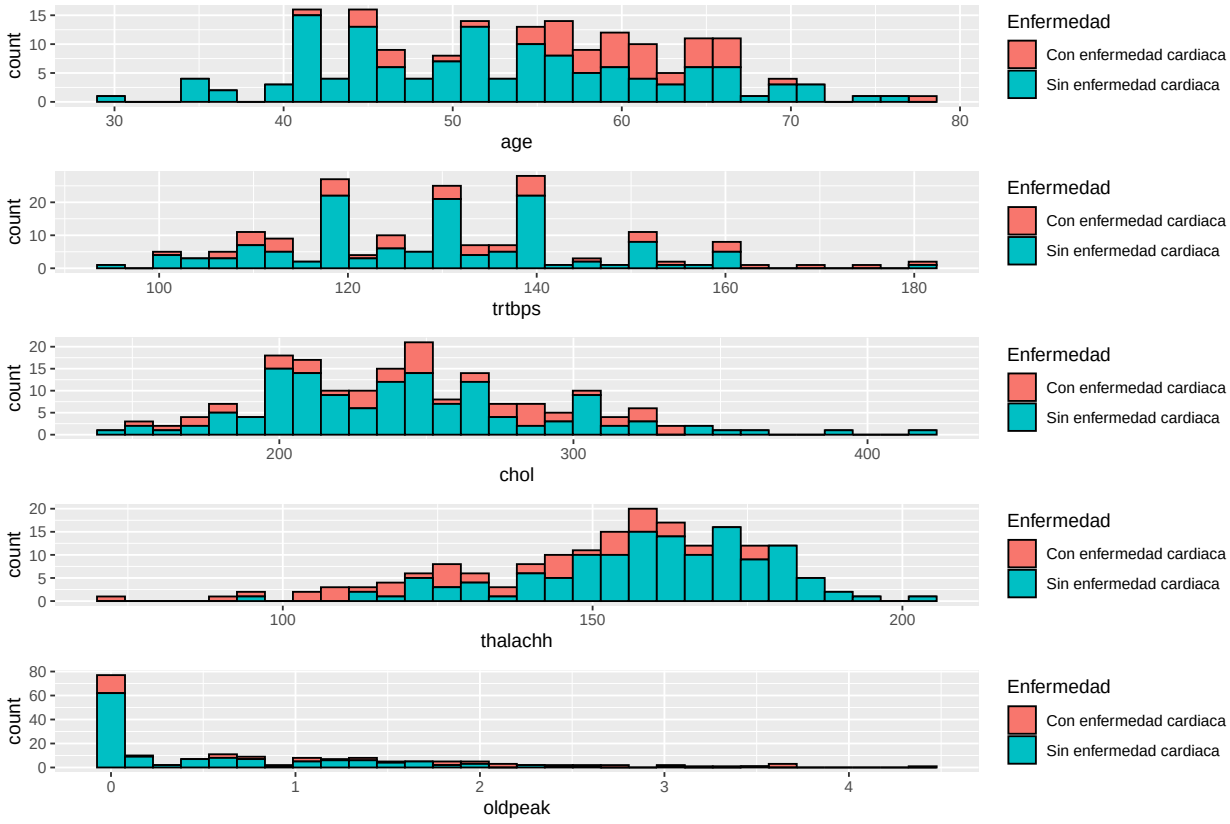

```
# Distribución de muestras por edad y sexo
barplot(prop.table(table(heartDiseaseDS$sex, heartDiseaseDS$age)),
        ylim = c(0, 0.1),
        main = "Distribución de muestras por edad y sexo",
        xlab = 'Edad/Sexo', ylab='Frecuencia',
        col = c("#CC79A7", "#56B4E9"))
```



Como podemos observar tenemos una buena representación por grupos de edad-sexo, con un poco de sobrerepresentación de la clase de sexo hombre.

Continuamos estudiando el resto de las variables númericas en relación a la variable objetivo.

```
histList<- list()
heartDiseaseDSAUX= heartDiseaseDS %>% select(all_of(c('age','trtbps','chol','thalachh',
                                                      'oldpeak', 'output')))

for(y in 1:5){
  col <- names(heartDiseaseDSAUX)[y]
  ggp <- ggplot(heartDiseaseDSAUX, aes_string(x = col, fill=heartDiseaseDSAUX$output)) +
  geom_histogram(bins = 30, color = "black")+labs(fill = "Enfermedad")
  histList[[y]] <- ggp # añadimos cada plot a la lista vacía
}

multiplot(plotlist = histList, coles = 1)
```




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

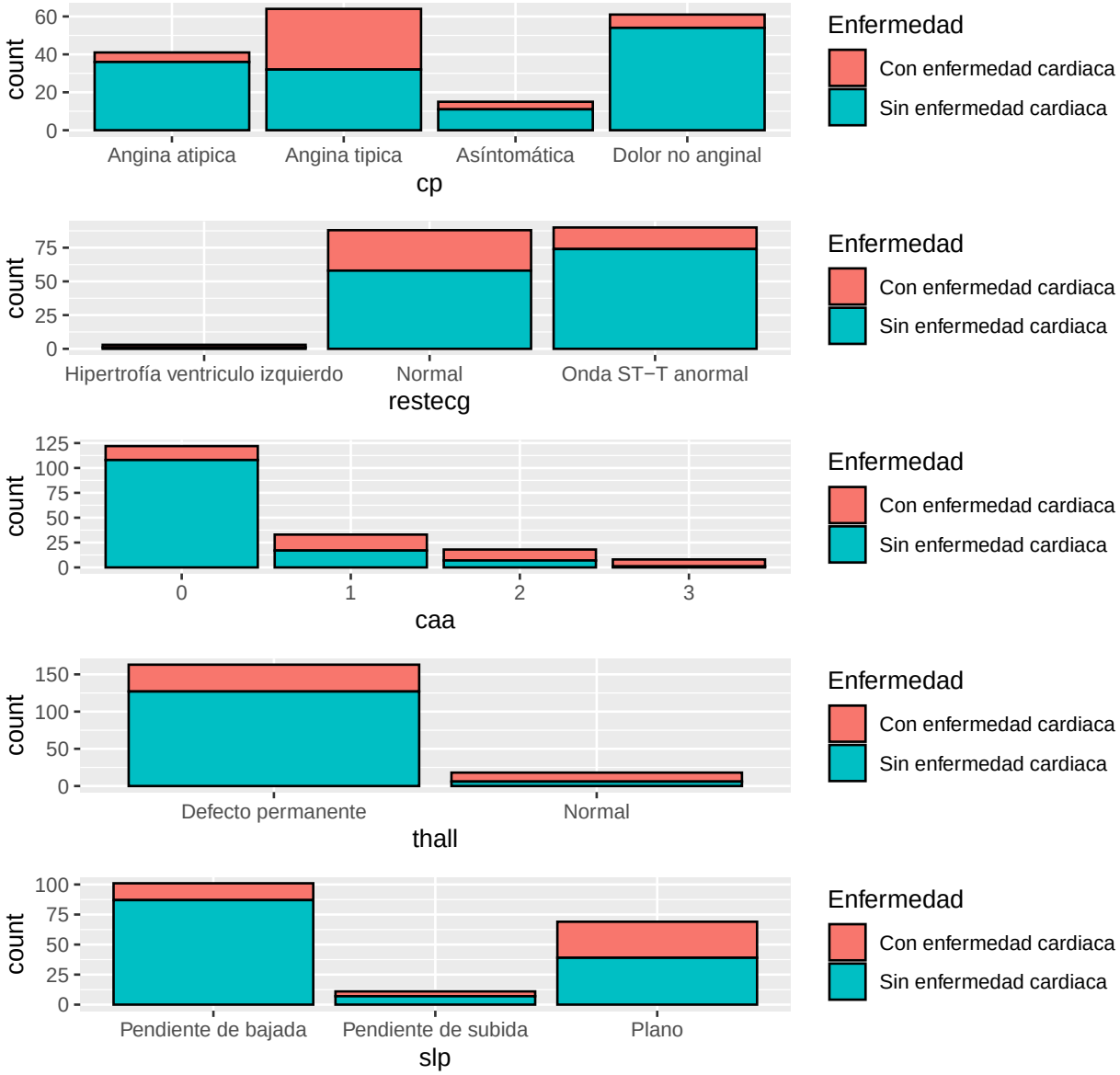
Algunas observaciones que podemos realizar:

- **Edad (age):** A mayor edad la posibilidad de tener una enfermedad cardiaca aumenta. En edades tempranas el número de casos con / sin enfermedad se mueven en porcentajes similares.
- **Presión sanguínea en reposo (trtbps):** Tener una presión sanguinea en reposo elevada (> 120) aumenta las opciones de tener enfermedad cardíaca.
- **Colesterol (chol):** Un colesterol elevado (> 200) también dispara las opciones de tener una enfermedad cardíaca
- **Thalachh (thal):** Tener un valor superior a 140 disminuye las opciones a tener una enfermedad cardíaca.

Revisando la relación de las variables categóricas con respecto al target:

```
histList<- list()
heartDiseaseDSAUX= heartDiseaseDS %>% select(all_of(c('cp','restecg','caa','thall','slp',
                                                      'output')))

for(y in 1:5){
```

```
  col <- names(heartDiseaseDSAUX)[y]
  ggp <- ggplot(heartDiseaseDSAUX, aes_string(x = col, fill=heartDiseaseDSAUX$output)) +
  geom_bar(color = "black")+labs(fill = "Enfermedad")
  histList[[y]] <- ggp # añadimos cada plot a la lista vacía
}
multiplot(plotlist = histList, coles = 1)
```



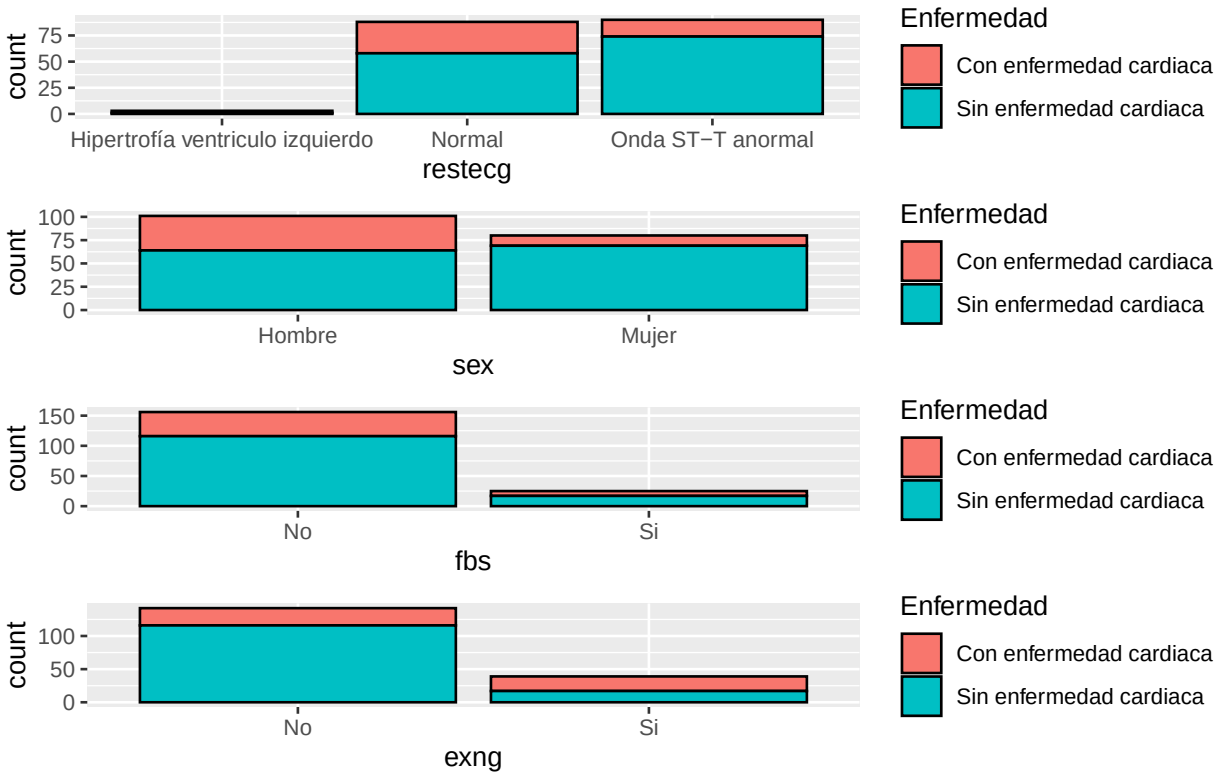

```
## [1] 1
```

De estas primeras variables categóricas podemos extraer las siguientes conclusiones:

- **Dolor en el pecho (cp):** Personas con un dolor angína tipica son más propensos a tener una enfermedad cardíaca
- La clase de "Hipertrofia ventriculo izquierdo" de la variable restecg está infrarepresentada en comparación con las otras dos clases.
- Igualmente, la clase "Pendiente de subida" de la variable slope también está infrarepresentada en comparación con las otras clases de la misma variable.

Esto puede provocar que los algoritmos ignoren dicha clase.

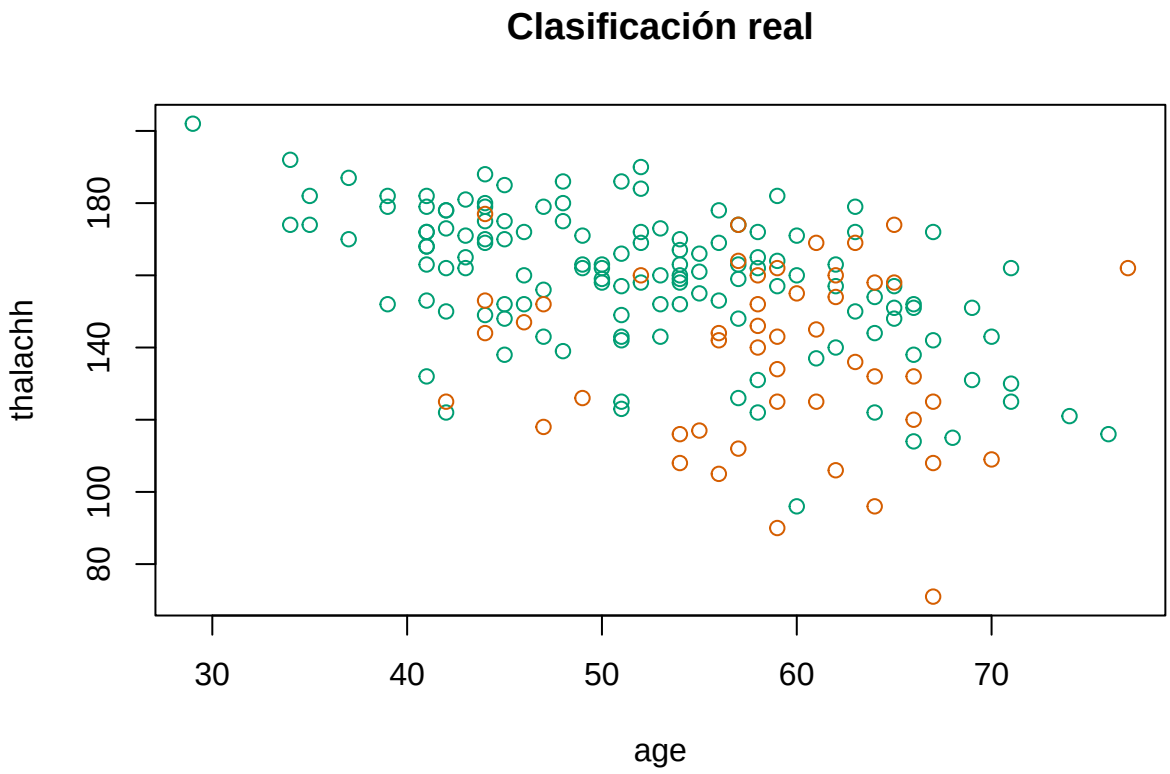
```
histList<- list()
heartDiseaseDSAUX= heartDiseaseDS %>% select(all_of(c('restecg', 'sex','fbs','exng',
                                                      'output')))

for(y in 1:4){
  col <- names(heartDiseaseDSAUX)[y]
  ggp <- ggplot(heartDiseaseDSAUX, aes_string(x = col, fill=heartDiseaseDSAUX$output)) +
  geom_bar(color = "black")+labs(fill = "Enfermedad")
  histList[[y]] <- ggp # añadimos cada plot a la lista vacía
}
multiplot(plotlist = histList, coles = 1)
```



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

Puede verse también la frecuencia cardiaca máxima en contraste con la edad.

```
plot(heartDiseaseDS[c("age", "thalachh")],
     col=c("#D55E00", "#009E73")[heartDiseaseDS$output],
     main="Clasificación real")
```



2.8 Búsqueda de correlaciones

De forma previa a la búsqueda de correlaciones entre las variables, vamos a realizar una decodificación de las variables de tipo categórico en forma ONE-HOT.

Debe tenerse en cuenta, que al realizar la codificación ONE-HOT las variables binarias han pasado a ser dos variables. De las dos variables obtenidas por cada variable binaria, se procederá a eliminar una de ellas, ya que por propia definición, ya están en forma ONE-HOT, y al incorporarse como dos variables se va a introducir una alta correlación inversa.

Adicionalmente, en el resto de variables categoricas no binarias se procederá también a eliminar una de las variables, ya que la información queda igualmente representada por las otras variables generadas haciendo = 0.

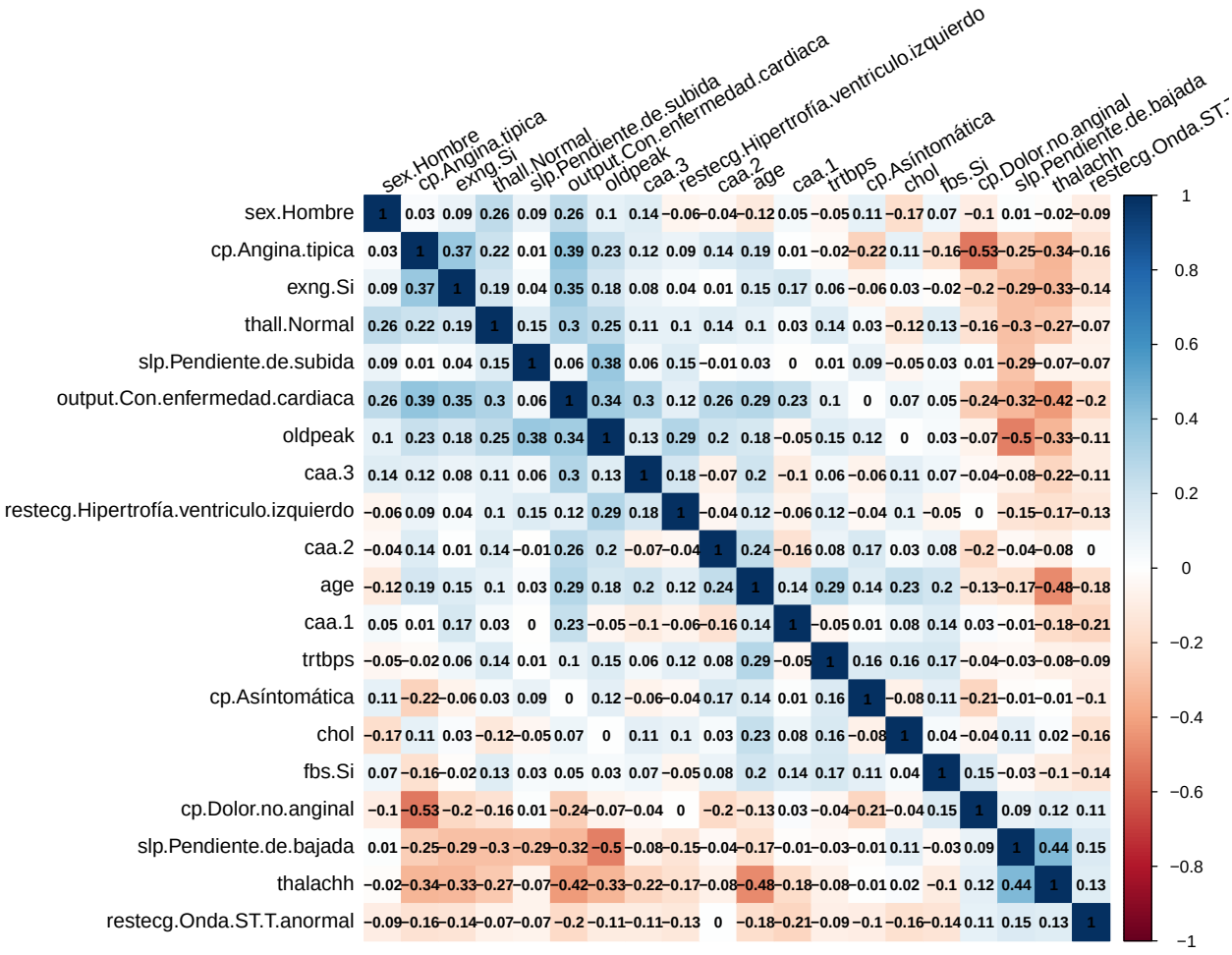
```
dummy <- dummyVars(" ~ .", data=heartDiseaseDS)
heartDiseaseDSOH <- data.frame(predict(dummy, newdata=heartDiseaseDS))
```

```r
#Eliminamos una de las variables binarias, por ser redundantes
heartDiseaseDSOH <- select(heartDiseaseDSOH, -c("sex.Mujer", "exng.No", "fbs.No",
                                                "output.Sin.enfermedad.cardiaca"))

# También eliminamos las variables categóricas con sufijo, ya que la información está
# igualmente representada por las otras variables obtenidas de la variable original
heartDiseaseDSOH <- select(heartDiseaseDSOH, -c("cp.Angina.atipica", "restecg.Normal",
                                                "slp.Plano", "caa.0",
                                                "thall.Defecto.permanente"))

str(heartDiseaseDSOH, list.len=ncol(heartDiseaseDSOH))
```

```
## 'data.frame':    181 obs. of  20 variables:
##  $ age                                    : num  63 37 41 56 57 57 56 57 54 48 ...
##  $ sex.Hombre                             : num  1 1 0 1 0 1 0 1 1 0 ...
##  $ cp.Angina.tipica                       : num  0 0 0 0 1 1 0 0 1 0 ...
##  $ cp.Asíntomática                        : num  1 0 0 0 0 0 0 0 0 0 ...
##  $ cp.Dolor.no.anginal                    : num  0 1 0 0 0 0 0 1 0 1 ...
##  $ trtbps                                 : num  145 130 130 120 120 140 140 150 140 130 ...
##  $ chol                                   : num  233 250 204 236 354 192 294 168 239 275 ...
##  $ fbs.Si                                 : num  1 0 0 0 0 0 0 0 0 0 ...
##  $ restecg.Hipertrofía.ventriculo.izquierdo: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ restecg.Onda.ST.T.anormal              : num  0 1 0 1 1 1 0 1 1 1 ...
##  $ thalachh                               : num  150 187 172 178 163 148 153 174 160 139 ...
##  $ exng.Si                                : num  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 1.6 1.2 0.2 ...
##  $ slp.Pendiente.de.bajada                : num  0 0 1 1 1 0 0 1 1 1 ...
##  $ slp.Pendiente.de.subida                : num  1 1 0 0 0 0 0 0 0 0 ...
##  $ caa.1                                  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ caa.2                                  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ caa.3                                  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ thall.Normal                           : num  1 0 0 0 0 1 0 0 0 0 ...
##  $ output.Con.enfermedad.cardiaca         : num  0 0 0 0 0 0 0 0 0 0 ...
```

Con las variables que hemos mantenido y nos interesan, realizamos una matriz de correlación:

```r
if(!require("corrplot")) install.packages("corrplot"); library("corrplot")
res<-cor(heartDiseaseDSOH)
corrplot(res,method="color",tl.col="black", tl.srt=30, order = "AOE",
number.cex=0.75,sig.level = 0.01, addCoef.col = "black")
```



De la matriz podemos extraer las siguientes observaciones:

- *Existe alta correlación NEGATIVA entre:*
  - **Dolor de pecho:** dolor no anginal y angina tipica.

A pesar de lo anterior, ninguna variable tiene una correlación positiva o negativa superior a 0.8

3 Análisis de los datos

3.1 Selección de los grupos de datos que se quieren analizar

Se va a analizar de una parte los datos relativos a los pacientes con enfermedad cardíaca y los que no presentan enfermedades cardíacas, atendiendo a términos de variables como la edad, sexo, presión arterial

en reposo, y colesterol.

3.2 Comprobación de la normalidad y homogeneidad de la varianza

```
# Grupos
heartDiseaseDSConEnfermedad <- heartDiseaseDS %>% filter(output == "Con enfermedad cardiaca")
heartDiseaseDSSinEnfermedad <- heartDiseaseDS %>% filter(output == "Sin enfermedad cardiaca")

# Comprobación de la normalidad
ShapiroConEnfermedad <- shapiro.test(heartDiseaseDSConEnfermedad$age)
ShapiroSinEnfermedad <- shapiro.test(heartDiseaseDSSinEnfermedad$age)

print(ShapiroConEnfermedad)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  heartDiseaseDSConEnfermedad$age
## W = 0.95684, p-value = 0.07524
```

```
print(ShapiroSinEnfermedad)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  heartDiseaseDSSinEnfermedad$age
## W = 0.98325, p-value = 0.1017
```

Según los resultados de la prueba de normalidad de Shapiro-Wilk, para el grupo de datos con enfermedad cardíaca se obtiene un valor de p = 0.07524, y para el grupo de datos sin enfermedad cardíaca se obtiene un valor de p = 0.1017.

En ambos casos, los valores de p son mayores que el nivel de significancia comúnmente utilizado de 0.05, por lo que no hay evidencia suficiente para rechazar la hipótesis nula de normalidad en ninguno de los dos grupos. Por lo tanto, podemos asumir que las variables "age" en ambos grupos siguen una distribución normal.

```
# Comprobación de homogeneidad de varianza
levene_test <- leveneTest(age ~ output, data = heartDiseaseDS)

# Imprimir resultado
print(levene_test)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   1   7.225 0.007868 **
##       179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.3 Aplicación de pruebas estadísticas

3.3.1 La edad como factor influyente en la presencia de enfermedades cardíacas

En primer lugar, vamos a realizar una prueba de t de Student para determinar si hay una diferencia significativa entre las medias de edad de dos grupos. Se va a comparar la edad promedio de los pacientes que sufrieron un ataque al corazón contra la edad promedio de los pacientes que no lo hicieron. La hipótesis nula sería que no hay ninguna diferencia entre las medias de edad de estos dos grupos.

```
t.test(age ~ output, data = heartDiseaseDS)
```

```
##
##  Welch Two Sample t-test
##
## data:  age by output
## t = 4.5533, df = 109.06, p-value = 1.381e-05
## alternative hypothesis: true difference in means between group Con enfermedad cardiaca and group Sin
## 95 percent confidence interval:
##  3.528865 8.968942
## sample estimates:
## mean in group Con enfermedad cardiaca mean in group Sin enfermedad cardiaca
##                              58.35417                              52.10526
```

La media de edad de las personas con enfermedad cardiaca es de 58.35 años, mientras que la media de edad de las personas sin enfermedad cardiaca es de 52.10 años. El valor p es extremadamente pequeño 1.381e-05, menor que 0.05, esto significa que es poco probable que los resultados que estamos viendo se deban al azar. Por otro lado, el intervalo de confianza indica que, si hiciéramos la misma prueba muchas veces, esperaríamos que la diferencia entre las medias de las edades de los dos grupos se encuentre entre 3.53 y 8.97 años en el 95% de las veces.

Como conclusión podemos afirmar que la prueba de t de Student sugiere que hay una diferencia significativa en las medias de las edades de las personas con enfermedad cardiaca y las personas sin enfermedad cardiaca. Podríamos interpretar esto como que la edad puede ser un factor influyente en la presencia de una enfermedad cardiaca.

3.3.2 El sexo como factor influyente en la presencia de enfermedades cardíacas

A continuación, vamos a comprobar la influencia del sexo del paciente en la presencia de enfermedades cardíacas del corazón. Para ello vamos a realizar una prueba de Chi-cuadrado, que es un análisis estadístico que se utiliza para determinar si hay una asociación significativa entre dos variables categóricas en una población. En este caso se va a revisar la asociación entre las variables de 'sex' y 'output' en el conjunto de datos.

```
chisq.test(heartDiseaseDS$sex, heartDiseaseDS$output)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  heartDiseaseDS$sex and heartDiseaseDS$output
## X-squared = 10.851, df = 1, p-value = 0.0009876
```

El valor p es 0.0009876, que es mucho menor que el umbral común de 0.05, lo que indica que la hipótesis nula de que las variables son independientes puede ser rechazada con confianza. Atendiendo a los resultados de la prueba podemos observar que hay una asociación significativa entre las variables 'sex' y 'output'.

3.3.3 Regresión

```
modeloRegresion <- glm(output ~ age + sex + cp + trtbps + chol, data = heartDiseaseDS,
                       family = binomial())
summary(modeloRegresion)
```

```
##
## Call:
## glm(formula = output ~ age + sex + cp + trtbps + chol, family = binomial(),
##     data = heartDiseaseDS)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5273  -0.4523   0.3700   0.6217   2.0495
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          8.316016   2.223244   3.740 0.000184 ***
## age                 -0.086515   0.026545  -3.259 0.001117 **
## sexMujer             1.977221   0.495787   3.988 6.66e-05 ***
## cpAngina tipica     -2.041044   0.604413  -3.377 0.000733 ***
## cpAsíntomática      -0.274403   0.831227  -0.330 0.741311
## cpDolor no anginal  -0.103443   0.671177  -0.154 0.877514
## trtbps              -0.011609   0.012317  -0.942 0.345954
## chol                -0.002971   0.004413  -0.673 0.500815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 209.39  on 180  degrees of freedom
## Residual deviance: 150.62  on 173  degrees of freedom
## AIC: 166.62
##
## Number of Fisher Scoring iterations: 5
```

Según se observa, las variables 'age', 'sex' y tener el valor de 'Angina tipica' son variables estadísticamente significativas en el modelo. Sin embargo, otras variables como 'trtbps' y 'chol' no son estadísticamente significativas en este modelo, lo que significa que no hay suficiente evidencia para sugerir que estas variables tengan un efecto en la presencia o no de enfermedades cardíacas.

Algunas observaciones:

- **Edad (Age):** La variable "edad" tiene una relación negativa y significativa con la variable dependiente, ya que su coeficiente es -0.086515 y el valor p es 0.001117 (menor que 0.05). Esto significa que por cada aumento unitario en la edad, la variable dependiente disminuiría en 0.086515 unidades, manteniendo constantes las otras variables.
- **Tipo de dolor en el pecho (cp):** Las diferentes categorías de dolor en el pecho parecen tener un efecto diferente en la variable dependiente. En particular, tener 'angina típica' disminuye significativamente la variable dependiente (coeficiente de -2.041044, valor p de 0.000733). Sin embargo, las categorías 'asintomática' y 'dolor no anginal' no son significativas en el modelo, lo que sugiere que, con los datos disponibles, no podemos afirmar que tengan un efecto en la variable dependiente.

- **Presión arterial en reposo (trtbps) y colesterol (chol)**: Estas variables no son significativas en el modelo (valores p de 0.345954 y 0.500815, respectivamente).

4 Conclusiones

Como conclusiones de esta práctica podemos extraer que hay determinadas variables que parecen tener una influencia en la presencia de enfermedades cardíacas, tal es el caso de la edad o el sexo del paciente, así como la presencia o no de dolor de agína tipica. Por otro lado, se ha visto que con la información que se dispone, las variables como el colesterol en sangre, o la presión sanguínea en reposo no parecen estar ligadas a la variable objetivo.