UOC - Tipologia y ciclo de vida de los datos - PRA2

Limpieza y Preprocesado: Heart Attack Analysis & Prediction Dataset

Vanessa Moreno González, Manuel Ernesto Martínez Martín

27 de mayo 2023

Índice

1. Descripción del dataset

Este dataset trae dos ficheros `heart.csv` y `o2Saturation.csv` y **es importante porque proporciona información sobre factores relacionados con enfermedades cardíacas**, como edad, sexo, síntomas otros datos médicos. Ya que con el **se puede entender mejor la enfermedad y hacer un análisis para detectar cuando se puede estar en riesgo de ataque cardíaco**, sabiendo esto se pueden desarrollar modelos predictivos que tomen decisiones para ayudar a prevenir un ataque cardíaco.

El dataset es el propuesto en el enunciado de la práctica y se ha extraído de kaggel: **Heart Attack Analysis & Prediction Dataset**

Contenido del dataset

Las variables que tiene el dataset son: age, sex, cp, trtbps, chol, fbs, restecg, thalachh, exng, oldpeak, slp, caa, thall y output. Siendo `output` la variable objetivo. A continuación se detallan más en profundidad.

- **age**: Edad del paciente.
- **sex**: Género del paciente.
  - *0*: Femenino
  - *1*: Masculino
- **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 medido mediante un sensor BMI.
- **fbs**: Nivel de azúcar en sangre en ayunas (> 120 mg/dl).
  - *1*: Verdadero
  - *0*: Falso
- **restecg**: Resultados electrocardiográficos en reposo.
  - *0*: Normal
  - *1*: Anormalidad con inversiones de onda ST-T y/o alteraciones del segmento ST > 0.05 mV
  - *2*: Hipertrofia ventricular izquierda
- **thalachh**: Ritmo cardíaco máximo alcanzado.
- **exng**: Angina inducida por ejercicio.
  - *1*: Sí
  - *0*: No
- **oldpeak**: Diferencia entre la depresión del segmento ST durante el ejercicio y durante el descanso en un electrocardiograma.
- **slp**: Pendiente del segmento ST durante el ejercicio en la prueba de esfuerzo.
  - *1*: Ascendente
  - *2*: Plana
  - *3*: Descendente
- **caa**: Número de vasos principales (0-3).
- **thall**: Talasemia, trastorno hereditario de la sangre caracterizado por un menor nivel de hemoglobina.
  - *0*: Ausencia
  - *1*: Talasemia normal
  - *2*: Talasemia fija defectuosa
  - *3*: Talasemia Reversible defectuosa
- **output**: Variable objetivo.
  - *0*: Menor probabilidad de ataque al corazón
  - *1*: Mayor probabilidad de ataque al corazón

Análisis inicial

Verificamos la estructura del juego de datos principal y el tipo de datos con los que R ha interpretado cada variable, y si, corresponde a la descripción de las variables del fichero original:

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

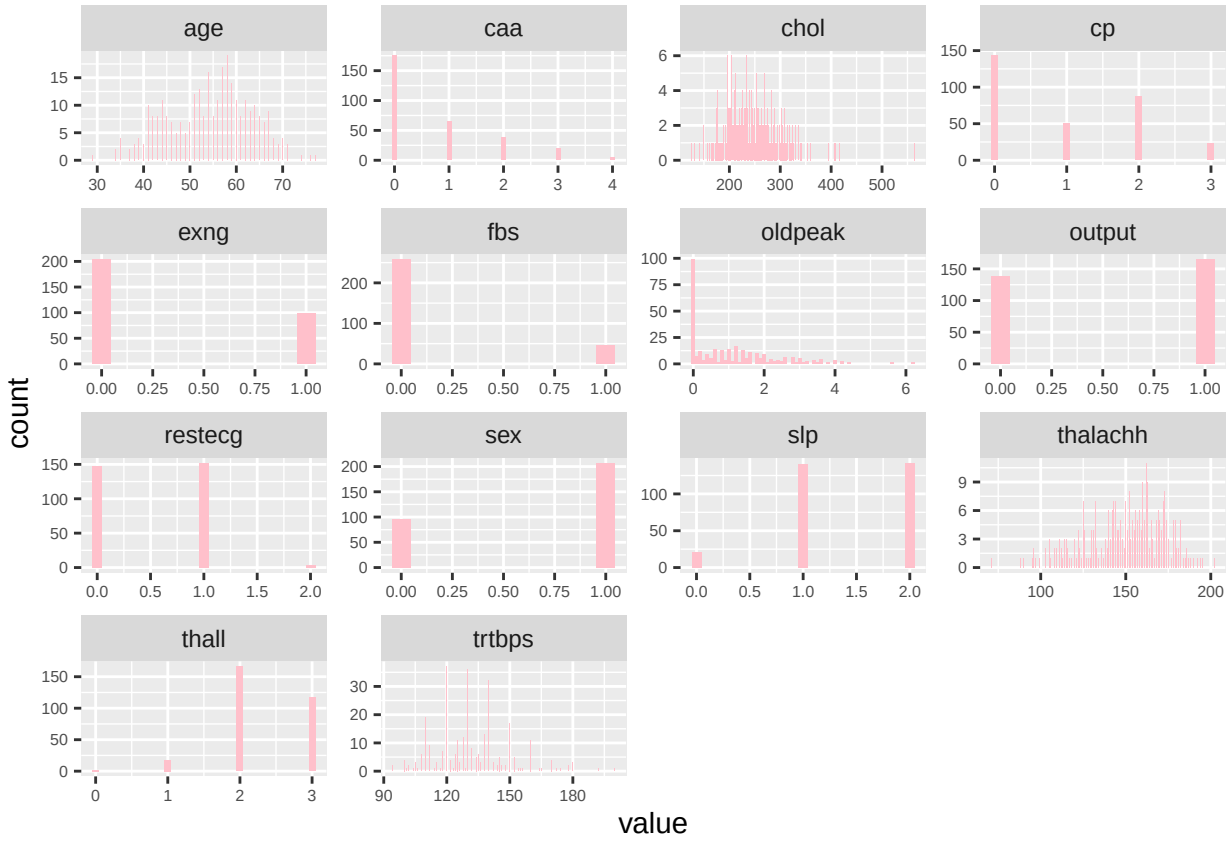
Observamos que todas las variables se han cargado como númerica discreta a excepción de **oldpeak** que se ha cargado como númerica continua.

A continuación realizaremos una visión general del dataset.

```
glimpse(heartAttack)
```

```
## Rows: 303
## Columns: 14
## $ age      <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 5~
## $ sex      <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1~
## $ cp       <int> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3, 0~
## $ trtbps   <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 1~
## $ chol     <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2~
## $ fbs      <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0~
## $ restecg  <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1~
## $ thalachh <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1~
## $ exng     <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
## $ oldpeak  <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0~
## $ slp      <int> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1~
## $ caa      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0~
## $ thall    <int> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3~
## $ output   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

```
library(dplyr)
library(tidyr)
library(ggplot2)
heartAttack %>%
  select_if(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  geom_bar(fill = "pink") +
  facet_wrap(~ key, scales = 'free') +
  theme(axis.text = element_text(size = 6))
```



Observamos como **sex, caa, cp, fbs, restecg, exng, slp y thall** contienen un número limitado de valores únicos, por lo que, probablemente, esten representando variables categóricas.

Comprobaremos, según la descripción oficial del dataset del punto anterior, que és cada variable y si nuestro análisis inicial es correcto.

Según la descripción oficial, observamos que nuestra suposición es correcta, y que,a excepción de caa, **sex, cp, fbs, restecg, exng, slp y thall** son variables categóricas. Por lo tanto, las convertiremos:

```
# Definimos las variables que hemos indentificado como categoricas
categorical_var <- c("sex", "cp", "fbs", "restecg", "exng", "slp", "thall")

# Iterar sobre cada variable y la convertimos a factor
for (variable_name in categorical_var) {
  heartAttack[[variable_name]] <- as.factor(heartAttack[[variable_name]])
}
```

Ahora definiremos otra variable que contenga el nombre de las variables identificadas como númericas:

```
numerical_var <- c("age", "trtbps", "chol", "thalachh", "oldpeak", "caa")
```

---

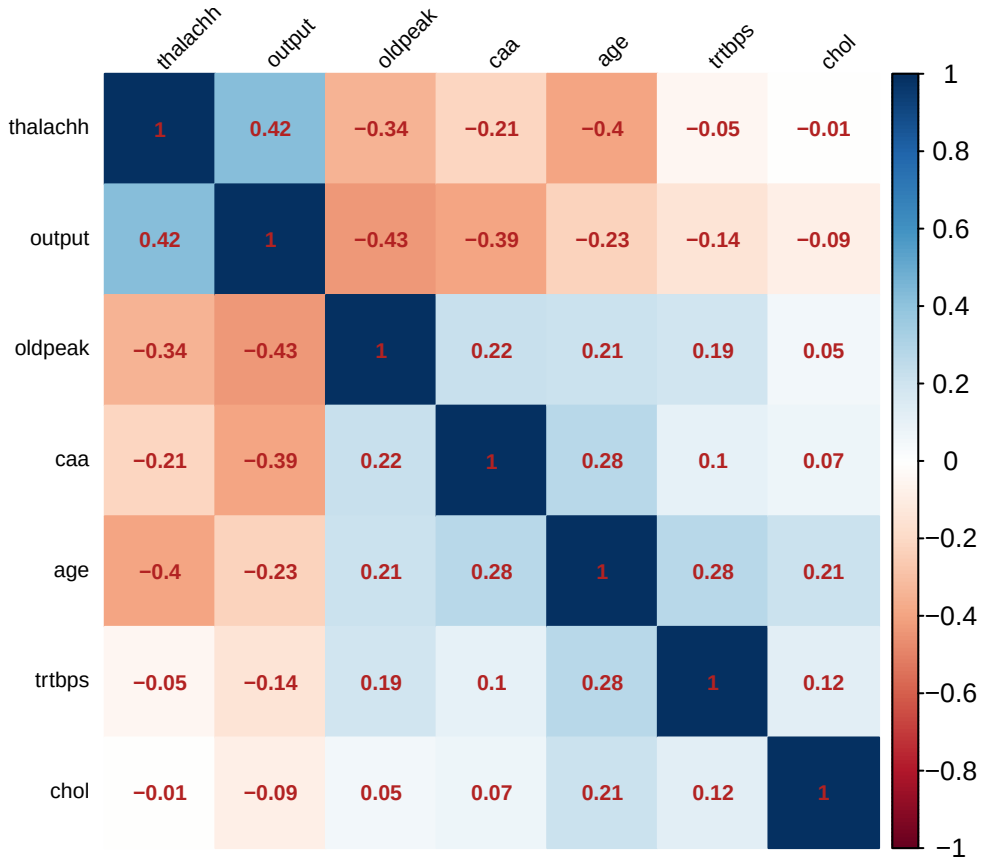
2. Integración y selección de variables

Observando los dos ficheros csv, **`heart.csv`** tiene **14 variables** y **303 registros** mientras que **`o2Saturation.csv`** con **1 variable** y **3585 registros**.

Aunque el nivel de saturación de oxigeno pueda ser importante para los ataques cardíacos, no hay manera de

juntar los dos conjuntos de datos en uno solo debido a que no hay un identificador de paciente, por lo que solo usaremos `heart.csv`.

Para la selección de los datos, comprobaremos la correlación entre ellas. En el caso de las variables numericas, realizaremos la correlación de Pearson:



Tanto una correlación positiva como una muy negativa son interesantes para la selección de variables. Centrándonos en la fila de la variable objetivo `output` se tienen los siguientes valores:

```
# Visualizamos la correlación de Pearson de la variable "output" versus el resto
cor_pearson_output
```

```
##      age   trtbps     chol thalachh  oldpeak      caa   output
##    -0.23    -0.14    -0.09     0.42    -0.43    -0.39     1.00
```

Se puede tomar como referencia 0.15 como umbral para comprobar las variables que no son necesarias para el estudio, siempre en valor absoluto. En este caso para el coeficiente de correlación de pearson se tienen `age`, `trtbps`, `thalachh`, `oldpeak` y `caa` como variables aptas y `chol` como poco importante.

Para las variables categóricas seria más apropiado hacer un test de `Fisher` o un `Chi-squared`.

Se va a proceder a hacer uso del test de `Fisher`:

$p = \frac{\binom{a+b}{a} \cdot \binom{c+d}{c}}{\binom{n}{a+c}}$

```
# Creamos una lista vacía para almacenar los resultados de las pruebas de Fisher
fisher_results <- list()

# Iteramos sobre cada variable categórica
for (var in categorical_var) {
```

```
  fisher_result <- fisher.test(heartAttack[[var]], heartAttack$output)
  fisher_results[[var]] <- fisher_result
}
```

Visualizamos los resultados obtenidos del test de Fisher:

```
fisher_results
```

```
## $sex
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value = 1.042e-06
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1519598 0.4783553
## sample estimates:
## odds ratio
##  0.2731136
##
##
## $cp
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value < 2.2e-16
## alternative hypothesis: two.sided
##
##
## $fbs
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value = 0.6308
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4308961 1.6975867
## sample estimates:
## odds ratio
##  0.8544825
##
##
## $restecg
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value = 0.003629
## alternative hypothesis: two.sided
##
##
## $exng
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value = 1.76e-14
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.07259027 0.23708719
## sample estimates:
## odds ratio
##   0.133146
##
##
## $slp
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value = 1.165e-11
## alternative hypothesis: two.sided
##
##
## $thall
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value < 2.2e-16
## alternative hypothesis: two.sided
```

Se obtiene que:

- **sex** tiene un p-value de 10422376×10^{-6} que es menor a 0.05, con lo que es estadisticamente significativa
- **cp** tiene un p-value de 12079934×10^{-18} que es menor a 0.05, con lo que es estadisticamente significativa
- **fbs** tiene un p-value de 0.6308003 que es mayor a 0.05, con lo que no es estadisticamente significativa
- **restecg** tiene un p-value de 0.0036292 que es menor a 0.05, con lo que es estadisticamente significativa
- **exng** tiene un p-value de 17599142×10^{-14} que es menor a 0.05, con lo que es estadisticamente significativa
- **slp** tiene un p-value de 11654794×10^{-11} que es menor a 0.05, con lo que es estadisticamente significativa
- **thall** tiene un p-value de 22664765×10^{-20} que es menor a 0.05, con lo que es estadisticamente significativa

---

3. Limpieza de los datos

FIXME

```
str(heartAttack)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 2 ...
##  $ cp      : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
##  $ 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     : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
##  $ restecg : Factor w/ 3 levels "0","1","2": 1 2 1 2 2 2 1 2 2 2 ...
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exng    : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slp     : Factor w/ 3 levels "0","1","2": 1 1 3 3 3 2 2 3 3 3 ...
##  $ caa     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thall   : Factor w/ 4 levels "0","1","2","3": 2 3 3 3 3 2 3 4 4 3 ...
##  $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

FIXME

```
summary(heartAttack)
```

```
##       age        sex     cp          trtbps           chol       fbs
##  Min.   :29.00   0: 96   0:143   Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:47.50   1:207   1: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45
##  Median :55.00           2: 87   Median :130.0   Median :240.0
##  Mean   :54.37           3: 23   Mean   :131.6   Mean   :246.3
##  3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:274.5
##  Max.   :77.00                   Max.   :200.0   Max.   :564.0
##  restecg    thalachh     exng       oldpeak     slp          caa
##  0:147   Min.   : 71.0   0:204   Min.   :0.00   0: 21   Min.   :0.0000
##  1:152   1st Qu.:133.5   1: 99   1st Qu.:0.00   1:140   1st Qu.:0.0000
##  2:  4   Median :153.0           Median :0.80   2:142   Median :0.0000
##          Mean   :149.6           Mean   :1.04           Mean   :0.7294
##          3rd Qu.:166.0           3rd Qu.:1.60           3rd Qu.:1.0000
##          Max.   :202.0           Max.   :6.20           Max.   :4.0000
##  thall       output
##  0:  2   Min.   :0.0000
##  1: 18   1st Qu.:0.0000
##  2:166   Median :1.0000
##  3:117   Mean   :0.5446
##          3rd Qu.:1.0000
##          Max.   :1.0000
```

FIXME

```
# FIXME
```

FIXME

3.1. ¿Los datos contienen ceros o elementos vacíos?

Tenemos algunas variables categóricas en formato numérico en nuestro conjunto de datos. Estas variables no se pueden considerar en la búsqueda de ceros, ya que el valor 0 es una de las posibles categorías para cada una de ellas. Las variables categóricas en formato numérico son `sex`, `cp`, `fbs`, `restecg`, `exng`, `slp`,`caa`, `thall` y `output`. De las cuales son dicotomicas `sex`, `fbs`, `exng` y `output`.

FIXME

- **age**: 0
- **trtbps**: 0
- **chol**: 0
- **thalach**: 0
- **oldpeak**: 99

FIXME

3.2. Identifica y gestiona los valores extremos

FIXME

---

4. Análisis de los datos

FIXME

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

FIXME

4.2. Comprobación de la normalidad y homogeneidad de la varianza

FIXME

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

FIXME

---

5. Representación de los resultados

FIXME

---

6. Resolución del problema

FIXME

---

7. Código

FIXME

---

8. Vídeo

FIXME