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

24 de May 2023

Índice


```
# Se carga el juego de datos
heartAttack <- read.csv('../data/heart_in.csv')
```

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**

---

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`.

---

3 Limpieza de los datos

Contenido del dataset

- **age**: Edad del paciente.
- **sex**: Género del paciente (1 = masculino, 0 = femenino).
- **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 en la onda ST-T (inversiones de onda T y/o elevación o depresión del segmento ST > 0.05 mV).
  - 2: Probable o definitiva hipertrofia ventricular izquierda según los criterios de Estes.
- **thalach**: Ritmo cardíaco máximo alcanzado.
- **exang**: Angina inducida por ejercicio (1 = sí, 0 = no).
- **oldpeak**: Pico anterior
- **slp**: Pendiente.
- **caa**: Número de vasos principales (0-3).
- **thall**: Resultados de prueba de esfuerzo con talio (0-3).
- **output**: 0 = menor probabilidad de ataque al corazón, 1 = mayor probabilidad de ataque al corazón.

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.

```
str(heartAttack)
```

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

FIXME

```
summary(heartAttack)
```

```
##       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
```

FIXME

```
# FIXME
```

FIXME

3.1 ¿Los datos contienen ceros o elementos vacíos?

- **age**: Hay 0 pacientes con edad 0
- **sex**: Es una variable categórica dicotomica de 0 y 1
- **cp**: Es una variable categórica con 0, 1, 2 y 3
- **trtbps**: 0
- **chol**: 0
- **fbs**:
- **restecg**:
- **thalach**: 0
- **exang**:
- **oldpeak**: 99
- **slp**: 21
- **caa**:
- **thall**:
- **output**: Es una variable categórica dicotomica de 0 y 1

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