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`.

Contenido del dataset

- **Age**: Edad del paciente.
- **Sex**: Género del paciente (1 = masculino, 0 = femenino).
- **exang**: Angina inducida por ejercicio (1 = sí, 0 = no).
- **ca**: Número de vasos principales (0-3).
- **cp**: Tipo de dolor en el pecho.
  - 1: Angina típica.
  - 2: Angina atípica.
  - 3: Dolor no anginal.
  - 4: 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).
- **rest_ecg**: 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.
- **target**: 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.

```
heartAttack_summary <- capture.output(str(heartAttack))
kable(heartAttack_summary, format = "html")
```

x

'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

---

3 Limpieza de los datos

```
# FIXME
```

FIXME

3.1 ¿Los datos contienen ceros o elementos vacíos?

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