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

26 de May 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
- **thalach**: 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

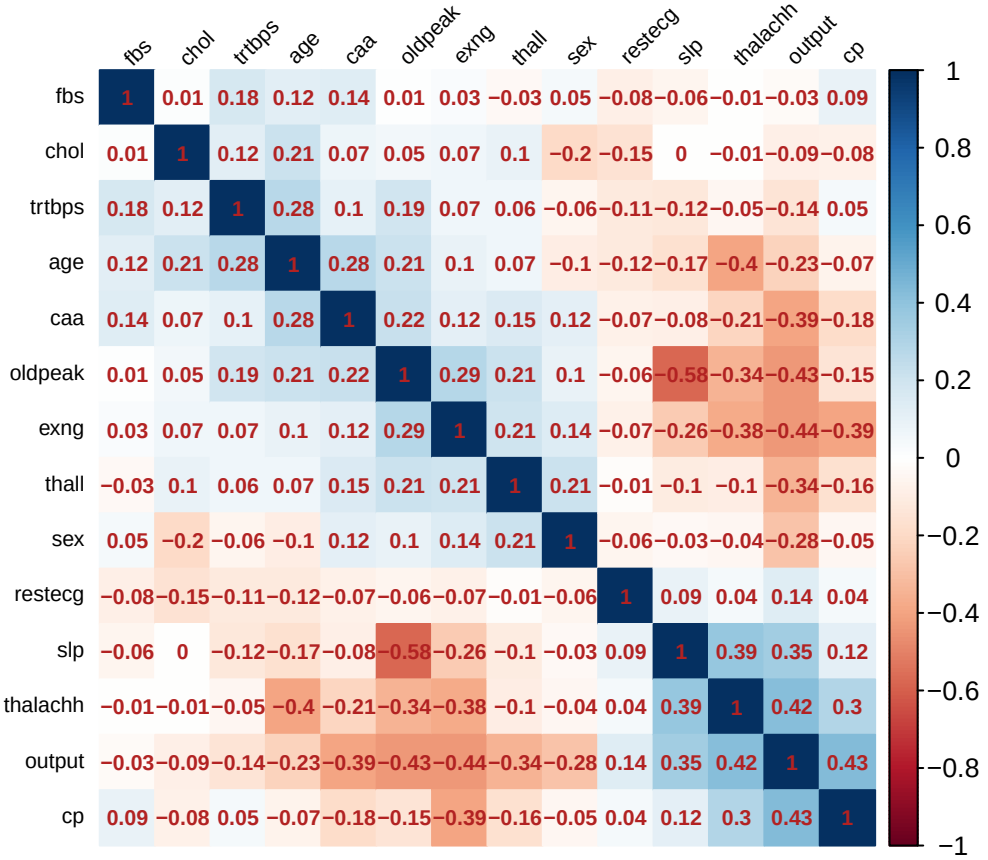
---

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, aprovechando de que todas las variables son numéricas se puede comprobar la correlación entre ellas



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: age = -0.23, sex = -0.28, cp = 0.43, trtbps = -0.14, chol = -0.09, fbs = -0.03, restecg = 0.14, thalachh = 0.42, exng = -0.44, oldpeak = -0.43, slp = 0.35, caa = -0.39 y thall = -0.34.

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 `exng`, `oldpeak`, `cp`, `thalachh`, `caa`, `slp`, `thall`, `sex` y `age` como variables aptas y `trtbps`, `restecg`, `chol` y `fbs` como poco importantes. Sin embargo para las variables categóricas numéricas seria más apropiado hacer un test de `Fisher` o un `Chi-squared`.

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

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

Se obtiene que:

- **sex** tiene un p-value de 1.0422376×10^{-6} que es menor a 0.05, con lo que es estadisticamente significativa
- **cp** tiene un p-value de $1.2079934 \times 10^{-18}$ 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 $1.7599142 \times 10^{-14}$ que es menor a 0.05, con lo que es estadisticamente significativa
- **slp** tiene un p-value de $1.1654794 \times 10^{-11}$ que es menor a 0.05, con lo que es estadisticamente significativa
- **caa** tiene un p-value de $1.3056069 \times 10^{-16}$ que es menor a 0.05, con lo que es estadisticamente significativa
- **thall** tiene un p-value de $2.2664765 \times 10^{-20}$ 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     : 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

```r
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

```r
# 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