Pràctica 2 – Tipologia i cicle de vida de les dades

Assignatura: M2.951 / Semestre: 2022/23-2 / Data: 10-04-2023

Autors:

- Francisco J. Bastida López – fbastidal@uoc.edu
- Ivan Benaiges Trenchs – ibenaiges@uoc.edu

Resolució dels apartats

1. Descripció del dataset

El conjunt de dades seleccionat "Heart Attack Analysis & Prediction dataset" proporciona informació sobre diferents factors que podrien estar relacionats amb malalties cardiovasculars, més concretament amb la predicció de la probabilitat de patir un atac de cor.

Es tracta d'un conjunt de dades que ens permetrà realitzar un estudi en major profunditat de possibles problemes del cor, centrat en els atacs de cor, al tenir una variable que està directament relacionada amb el resultat que es va obtenir durant l'estudi dels pacients i que ens permet trobar un model per determinar la probabilitat de patir un atac de cor segons els valors de les diferents variables d'estudi.

Per poder facilitar aquesta anàlisi, el dataset conté el fitxer *heart.csv* amb les següents variables/atributs:

- **age**: edat del pacient
- **sex**: gènere del pacient
  - 0 = femení
  - 1 = masculí
- **cp**: tipus de dolor toràcic que experimenta el pacient
  - 0 = angina típica
  - 1 = angina atípica
  - 2 = dolor no relacionat amb angina
  - 3 = asimptomàtic (sense dolor toràcic)
- **trtbps**: pressió arterial en repòs (mm Hg)
- **chol**: nivell de colesterol (md/dl)
- **fbs**: nivell de sucre en sang en dejú
  - 0 = normal

  - 1 = alt)
- **restecg**: resultat de l'electrocardiograma en repòs
  - 0 = normal
  - 1 = anomalies en l'ona ST-T
  - 2 = hipertrofia ventricular esquerra probable o definitiva segons els criteris d'Estes
- **thalachh**: freqüència cardíaca màxima registrada pel pacient durant les proves realitzades
- **exng**: angina provocada per l'exercici
  - 0 = no
  - 1 = sí
- **oldpeak**: canvis en el segment ST de l'electrocardiograma després de l'exercici físic
- **slp**: patró de canvi en el segment ST de l'electrocardiograma durant una prova d'esforç o situacions d'estrès
  - 0 = pendent plana
  - 1 = pendent ascendent (canvi en el patró elèctric del cor)
  - 2 = pendent descendent (canvi en el patró elèctric del cor)
- **caa**: nombre de vasos sanguinis coronaris que mostren obstrucció o estenosi significativa
  - 0 = sense obstrucció detectada
  - 1 = obstrucció en un dels vasos sanguinis
  - 2 = obstrucció en dos dels vasos sanguinis
  - 3 = obstrucció en tres dels vasos sanguinis
  - 4 = obstrucció en els quatre vasos sanguinis
- **thall**: relacionat amb una malaltia hereditària de la sang anomenada talassèmia
  - 1 = no s'ha detectat cap indici
  - 2 = presència d'un defecte fix
  - 3 = presència d'un defecte reversible
- **output**: probabilitat de patir un atac de cor:
  - 0 = sense o poca probabilitat
  - 1 = major probabilitat

Per altra banda, hi ha disponible un segon fitxer, *o2Saturation.csv*, que conté múltiples observacions relacionades amb el nivell de saturació d'oxigen (una única variable). El fitxer conté moltes més observacions (3586 en total) sense cap tipus d'identificador que ens permeti poder realitzar una integració de les dades d'ambdós fitxers per tenir un conjunt més complet. Llavors s'ha descartat aquest fitxer.

2. Integració i selecció de les dades d'interès

Les dades estan integrades en un únic fitxer CSV que hem carregat. El conjunt de dades té unes dimensions de 303 observacions i 14 variables; és, per tant, un fitxer petit i no s'ha considerat reduir la dimensionalitat. No s'ha aplicat cap filtre a les dades.

3. Neteja de les dades

En el fitxer resultat de l'anàlisi (pra2.html) es pot consulta tot una **anàlisi exploratòria** preliminar de les dades per conèixer la distribució de les seves variables. Especialment diferenciant els pacients amb alta i baixa probabilitat d'un atac de cor segons les diferents variables: sexe, edat, dolor toràcic, angina, nivell de sucre, pressió arterial en repòs, colesterol, resultat de l'electrocardiograma, vasos sanguinis obstruïts i talassèmia. També s'ha dibuixat un núvol de punts de totes les parelles de variables per identificar correlacions visuals amb les corbes de densitat de cada variable.

A continuació s'ha procedit a fer una neteja de dades amb els següents resultats:

- No s'han trobat valors absents (NA) ni nuls.
- S'han identificat alguns valors atípics en tres variables; però tots dins rangs possibles i, per tant, no s'han descartat.
- S'ha procedit a fer una normalització de dades numèriques escalant i centrant per obtenir mitjanes zero i desviacions 1. Aquesta tasca ajudarà a l'anàlisi posterior.
- S'han discretitzat les variables numèriques per rangs amb un algorisme de clusterització per poder aplicar posteriorment algorismes de classificació.
- S'ha fet una anàlisi de components principals per considerar la possibilitat de reduir-ne la dimensionalitat; però finalment s'ha descartat.

4. Anàlisi

El procés d'anàlisi ha consistit en les següents tasques:

- **Test de normalitat** sobre les diferents variables numèriques per poder aplicar els contrast d'hipòtesi. Han sortit negatius, no obstant es donen les condicions per poder aplicar el teorema del límit central.
- **Estudi de correlacions** entre les diferents variables, on han aparegut algunes correlacions suaus i poc significatives.
- **Regressió logística**, per poder estimar la probabilitat de tenir un atac de cor a partir de la resta de variables; hem aconseguit un model que explica el 55% de la variabilitat de l'objectiu.
- **Un test d'hipòtesi** que ens ha obligat a acceptar que les dones tenen més probabilitat de tenir un atac de cor que els homes amb un nivell de confiança del 95%. Per fer aquest test ha estat necessari fer un altre previ per confirmar la homocedasticitat dels dos grups: homes i dones.
- A continuació hem aplicat **un algorisme de clusterització no supervisat**, concretament l'algorisme k_means amb dos grups, que amb la interpretació adequada ens permet predir la probabilitat d'un atac de cor amb un 71,3% d'encerts.
- Per últim, hem aplicat un algorisme de classificació supervisat *randomForest*; per al qual ha estat necessari separar el conjunt de dades en dos estratificats per la variable objectiu: un per entrenar el model i l'altre per validar-ho. Aquest model té una taxa d'encert del 80%.

5. Representació dels resultats

L'informe durant tots els punts presenta taules i gràfiques:

- En l'anàlisi exploratori podem trobar diagrames de barres senzill i apilats, núvols de punts, corbes de distribució i *boxplots*.
- Durant la normalització s'han dibuixat gràfics QQnorm per visualitzar la normalitat.
- En la correlació de variables s'ha fet una matriu de colors.
- En el model no supervisat s'han fet diagrames de línies per ajudar a estimar el valor de k i també un núvol de punts per colors.
- En el model supervisat s'ha mostrat la taula de confusió.

6. Resolució del problema

La principal conclusió de l'anàlisi és que es pot predit amb un grau d'encert d'un 80% la probabilitat de tenir un atac de cor amb les dades disponibles. En aquest cas el més important és reduir els falsos negatius (predir baixa probabilitat d'un atac de cor quan és fals) més que els falsos positius (predit un atac de cor quan es fals) ja que els segons són una falsa alarma; però els primers no donen alarma quan en realitat hi ha risc. Segons la taula de confusió, el model té un 10% de casos de falses negatius.

7. Codi

Tot el codi s'ha fer en R i concretament amb R Studio mitjançant R Markdown. El codi es pot descarregar des de GitHub a https://github.com/fbastidal/Practica-2

Contribucions	Signatura
Investigació prèvia	Francisco Bastida López, Ivan Benaiges Trenchs
Redacció de les respostes	Francisco Bastida López, Ivan Benaiges Trenchs
Desenvolupament del codi	Francisco Bastida López, Ivan Benaiges Trenchs
Participació en el vídeo	Francisco Bastida López, Ivan Benaiges Trenchs