**Projecte: Malaltia Cardiovascular**

La idea és explicar la presència o absència de malaltia cardiovascular donat uns factors de risc, per això, es disposa del conjunt de dades **cardio\_train.csv** obtingut de Kaggle[[1]](#footnote-1).

Aquest conjunt de dades està format per 70 000 dades de pacients, de les quals es disposa d’una variable objectiu i 11 variables amb característiques o “factors de risc”.

La variable objectiu és una variable binària que indica la presència o absència de malaltia cardiovascular:



De característiques (features) n’hi ha de tres tipus; objectives, resultat d’alguna prova i subjectives, i són les següents:

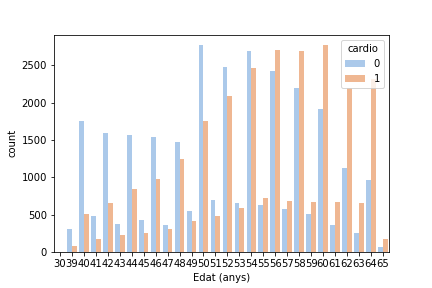


El conjunt de dades no disposa de cap valor mancant, si bé, si que hi ha alguna dada que sembla errònia. Per una banda, possibles errors de transcripció en les variables relatives a les pressions, la magnitud de la pressió sistòlica ha de ser superior al de la pressió diastòlica, però no sempre passa. Per l’altra, en les 4 variables continues (pes, alçada, pressió diastòlica i pressió sistòlica) hi valors molt extrems i incoherents: alçada màxima de 250 cm, pes mínim de 10 Kg (dades d’adults), pressions negatives o valors molt alts.... Eliminem del nostre conjunt de dades aquest casos, ho fem a partir dels valors dels percentils, casos amb valors inferiors al percentil 2.5 o superior al percentil 97.5. Dels 70.000 casos inicials, ens quedem amb 60.141 casos.

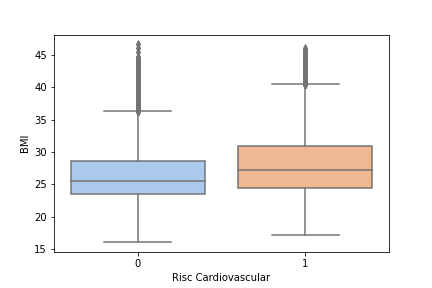
De les 11 variables característiques o “factors de risc” inicials, en calculem dues més: l’edat en anys i l’índex de massa corporal, que engloba les variables pes i alçada.

La presència o absència de risc cardiovascular està relacionada amb l’edat, a mesura que augmenta l’edat l’exposició al risc cardiovascular augmenta (figura 1), amb l’índex de massa corporal (figura 2), les pressions sistòlica i diastòlica (figura 3) i el colesterol (figura 4).

***Figura 1****. Presència o absència de risc cardiovascular per edat*



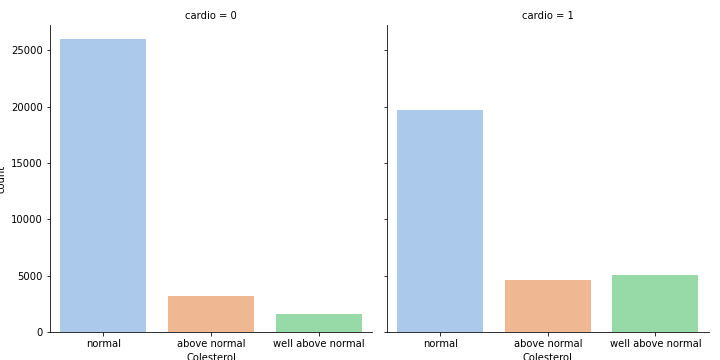
***Figura 2****. Presència o absència de risc cardiovascular per índex de massa corporal*



***Figura 3****. Presència o absència de risc cardiovascular per pressió sistòlica i diastòlica*

|  |  |
| --- | --- |
| Boxplot Risc cardiovascular x Sistòlica.png | Boxplot Risc cardiovascular x Diastòlica.png |

***Figura 4****. Presència o absència de risc cardiovascular per colesterol*



Abans d’ajustar els diferents algoritmes de machine learning d’aprenentatge supervisat, models de classificació per tal de predir la presència o absència de risc cardiovascular, hem de fer algunes modificacions al conjunt de dades.

* Estandarditzem les variables contínues per evitar problemes en les diferents magnituds de les variables
* Creem variables dummies de les variables categòriques no binàries
* Dividim el conjunt de dades en train-test, el Train té el 67% de les observacions i el Test el 33% restant

D’algoritmes de classificació, ajustem quatre: Regressió Logística, Arbre de decisió, K – Nearest Neightbours i Support Vector Machine. Amb les següents variables: gender2 (dummy variable inicial gender), age\_years (edat en anys) , BMI (calculada a partir de height o weight), ap\_hi, ap\_lo, cholesterol (2 dummies), gluc (2 dummies), smoke, alco i active.

Les mètriques per comprar els diferents models es troben a la taula 1.

***Taula 1.*** *Mètriques comparació d’algoritmes supervisats de Classificació*



Dels quatre models ajustats, el que té un major valor d’accuracy correspon al model de Support Vector Machine, que és d’un 73%. Aquests models contenen totes les variables, ajustant els models amb només les variables que veiem gràficament que tenien relació amb la presència o absència de risc cardiovascular no aconseguíem millorar els models. De la mateixa manera que tampoc ho fèiem si excloíem una de les variables de pressió per evitar problemes de colinealitat.

A la vista de les variables amb característiques o “factors de risc” disponibles, sorprèn que l’accuracy del model només sigui del 73%. Ara bé, en el moment de fer l’anàlisi descriptiu ja s’havia vist la falta de relació entre algunes d’elles, quan teòricament no hauria de ser així. Falta informació per part de l’autor del conjunt de dades de la procedència d’aquestes, no s’especifica res, només s’explica les variables que formen el conjunt de dades.

Arxiu Jupyter Notebook disponible a: <https://github.com/ntlbcn/ProjecteDS.git>

1. <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset> [↑](#footnote-ref-1)