PRÀCTICA (PART I)

Key Indicators of Heart Disease

MÀSTER DE CIÈNCIA DE DADES
UOC
VISUALITZACIÓ DE DADES

Marc Lladó i Maldonado

El conjunt de dades "**Key Indicators of Heart Disease**" és una recopilació de dades de l'enquesta anual del **CDC de 2022**. Inclou informació de més de 400.000 adults relacionada amb el seu estat de salut. Aquest conjunt de dades està orientat a proporcionar indicadors clau sobre les malalties cardíaques i els seus factors de risc associats.

Les malalties cardíaques són una de les principals causes de mort als Estats Units, segons el CDC. Aproximadament la meitat de tots els americans tenen almenys un dels tres principals factors de risc per a les malalties cardíaques: hipertensió arterial, colesterol alt i tabaquisme. Altres indicadors clau inclouen l'estat de diabetis, l'obesitat (alt IMC), no fer prou activitat física o beure massa alcohol. Identificar i prevenir els factors que tenen el major impacte en les malalties cardíaques és molt important en la sanitat pública.

El conjunt de dades té el seu origen en el CDC i forma part important del Behavioral Risk Factor Surveillance System (BRFSS), que realitza enquestes telefòniques anuals per recopilar dades sobre l'estat de salut dels residents als Estats Units. Amb una llarga història des del seu establiment el 1984, el BRFSS recull dades en tots els 50 estats, al Districte de Columbia i a tres territoris dels Estats Units.

Aquest conjunt de dades ofereix una visió àmplia de les condicions de salut dels adults als Estats Units i és una de les investigacions de salut més extenses del món.

Hi ha una versió del conjunt de dades, penjat a Kaggle, que ja ha estat preprocessada i n'han reduït el nº de variables, concretament de 300 a 40 variables, seleccionant les més imprescindibles. El conjunt de dades ha estat sotmès a diverses etapes de tractament per fer-lo més apte per a l'anàlisi i la predicció. Aprofitarem doncs el dataset penjat a Kaggle ja que està més polit i preparat per dur-ne a terme anàlisis. Aquest es pot trobar al següent enllaç:

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

A més de l'anàlisi exploratòria de dades (EDA), aquest conjunt de dades pot ser utilitzat per aplicar una sèrie de mètodes d'aprenentatge automàtic, especialment models de classificació com la regressió logística, SVM, , etc.

La variable "HadHeartAttack" s'ha de tractar com binària ("Sí" - el participant ha tingut malalties cardíaques; "No" - el participant no ha tingut malalties cardíaques), i cal tenir en compte que les classes estan desequilibrades.

El propietari del dataset ja preprocessat penjat a Kaggle**, Kamil Pytlak** adjunta el link al seu repositori de github per tal de poder comprovar quines són les transformacions que s'han dut a terme des del dataset original al dataset simplificat:

Repositori Github de Kamil Pytlak

A la publicació de Kaggle es disposen tres fitxer CSV:

- heart_2020_cleaned.csv
- heart_2022_no_nans.csv
- heart_2022_with_nans

Com que volem les dades més recents posssibles, ens quedarem amb el fitxer **heart_2022_no_nans.csv** per realitzar la pràctica. La part bona d'aquest dataset és que no inclou els registres que tenen molts nuls.

El fitxer **heart_2022_with_nans.csv** sí que inclou els valors nuls i al fer-ne una ullada als missing values:

```{r}
datamissing <-read.csv("./heart_2022_with_nans.csv",header=T,sep=",")

```

```{r}
colSums(is.na(datamissing))
```

```
                   State                      Sex
                       0                        0
           GeneralHealth        PhysicalHealthDays
                       0                    10927
        MentalHealthDays           LastCheckupTime
                    9067                        0
      PhysicalActivities               SleepHours
                       0                     5453
            RemovedTeeth            HadHeartAttack
                       0                        0
               HadAngina                HadStroke
                       0                        0
               HadAsthma             HadSkinCancer
                       0                        0
                 HadCOPD     HadDepressiveDisorder
                       0                        0
        HadKidneyDisease              HadArthritis
                       0                        0
             HadDiabetes       DeafOrHardOfHearing
                       0                        0
  BlindOrVisionDifficulty   DifficultyConcentrating
                       0                        0
       DifficultyWalking DifficultyDressingBathing
                       0                        0
       DifficultyErrands              SmokerStatus
                       0                        0
         ECigaretteUsage                 ChestScan
                       0                        0
   RaceEthnicityCategory               AgeCategory
                       0                        0
         HeightInMeters         WeightInKilograms
                   28652                    42078
                     BMI           AlcoholDrinkers
                   48806                        0
             HIVTesting              FluVaxLast12
                       0                        0
          PneumoVaxEver         TetanusLast10Tdap
                       0                        0
        HighRiskLastYear                 CovidPos
                       0                        0
```

S'ha optat per partir del fitxer que no conté nuls ja que heart_2022_with_nans.csv té algunes variables amb molts nuls, com s'ha pogut comprovar.

El fitxer **heart_2022_no_nans.csv,** el qual conté el dataset amb el que realitzrem la pràctica conté 246022 observacions (files) i 40 variables (columnes)

Per justificar la selecció d'aquest conjunt de dades, cal considerar primer la naturalesa crítica de les malalties cardíaques com a problema de salut pública. Les malalties cardíaques, segons el CDC, representen una de les principals causes de mortalitat en els Estats Units, afectant una àmplia franja de la població, independentment de la raça o l'ètnia. Això posa de manifest la importància de comprendre els factors de risc associats a aquestes malalties per poder dissenyar estratègies de prevenció i intervenció efectives.

En aquest sentit, la selecció d'un conjunt de dades que aborda aquest tema proporciona una oportunitat per aprofundir en la comprensió de les causes i els factors de risc relacionats amb les malalties cardíaques.

Pel que fa a la **rellevància** del conjunt de dades en el context, cal destacar que aquesta prové del CDC, una font fiable i autoritzada de dades de salut als Estats Units. Aquest conjunt de dades està basat en enquestes anuals realitzades a adults, cobrint una àmplia gamma de variables relacionades amb la salut cardíaca, com ara pressió arterial, nivells de colesterol, hàbits de vida i antecedents mèdics. D'aquesta manera, ofereix una visió completa dels factors de risc associats a les malalties cardíaques, proporcionant una base de dades rica i diversa per a la investigació en aquest àmbit.

A més, tenint en compte que partirem de la versió preprocessada que hi ha a Kaggle, es considera una opció fiable en la qual no s'han adulterat els registres, ja que el dataset està "premiat" per Kaggle amb una medalla d'or, la qual només es pot obtenir rebent moltes valoracions positives per part de la comunitat.

En quant a la complexitat del conjunt de dades, cal destacar que aquest conté una gran quantitat de registres, amb **més de 400.000 entrevistes d'adults**, **i 40 variables seleccionades**. Aquesta gran quantitat de dades permet una anàlisi detallada dels factors de risc associats a les malalties cardíaques, tant des d'una perspectiva estadística com epidemiològica. Les variables seleccionades inclouen tant dades categòriques com quantitatives, proporcionant una base de dades completa i diversa per a la investigació en aquest àmbit.

Pel que fa a l'originalitat del conjunt de dades, cal destacar que, malgrat que les malalties cardíaques són un tema àmpliament estudiades, la selecció d'aquest conjunt de dades respon a la necessitat d'una aproximació focalitzada i específica a la identificació de factors de risc. La reducció del conjunt de dades original de més de 300 variables a 40 variables rellevants per part de Kamil Pytlak reflecteix un enfocament precís en els factors clau associats a les malalties cardíaques. A més el conjunt de dades és força recent doncs els registres són de l'any 2022.

Finalment, quant a les preguntes que es poden abordar amb la visualització de dades, és possible identificar quins factors tenen un impacte significatiu en la probabilitat de desenvolupar malalties cardíaques. Això inclou explorar les relacions entre les variables seleccionades i la variable objectiu, "HadHeartAttack", així com avaluar la importància relativa de cada variable i com poden interactuar per influir en el risc de malalties cardíaques. Aquest enfocament permet generar coneixements útils per a la prevenció i el tractament de les malalties cardíaques, alhora que contribueix a la investigació en salut pública i epidemiologia.

En definitiva, **el conjunt de dades és adequat** per la visualització de dades, ja que disposa de 246022 registres i 40 variables de diferents tipus (categòriques i numèriques).

Diccionari de variables

- **State:** estat dels Estats Units on es va realitzar l'enquesta. Categòrica
- **Sex:** Sexe del participant. Categòrica (Male/Female)
- **GeneralHealth:** Percepció de la salut en general. Categòrica:
  1: "Excellent",
  2: "Very good",
  3: "Good",
  4: "Fair",
  5: "Poor"
- **PhysicalHealthDays: Nombre de dies durant els últims 30 dies en què la salut física no ha** estat bona. Quantitativa
- **MentalHealthDays:** Nombre de dies durant els últims 30 dies en què la salut mental no ha estat bona. Quantitativa
- **LastCheckupTime:** Temps transcorregut des de la darrera visita al metge per a una revisió rutinària. Categòrica:
  1: "Within past year (anytime less than 12 months ago)",
  2: "Within past 2 years (1 year but less than 2 years ago)",
  3: "Within past 5 years (2 years but less than 5 years ago)",
  4: "5 or more years ago"
- **PhysicalActivities:** Participació en activitats físiques durant el mes passat. Categòrica (Yes/No)
- **SleepHours:** Nombre mitjà d'hores de son en un període de 24 hores. Quantitativa
- **RemovedTeeth: Nom**bre de dents permanents eliminades a causa de càries o malalties de les genives. Categòrica:
  1: "1 to 5",
  2: "6 or more, but not all",
  3: "All",
  8: "None of them"
- **HadHeartAttack:** Ha tingut algun atac de cor. Categòrica (Yes/No)
- **HadAngina:** Ha tingut angines. Categòrica (Yes/No)
- **HadStroke:** Ha tingut algun accident cerebrovascular. Categòrica (Yes/No)
- **HadAsthma:** Ha tingut asma. Categòrica (Yes/No)
- **HadSkinCancer:** Historial de càncer de pell no melanoma. Categòrica (Yes/No)
- **HadCOPD:** Historial de malaltia pulmonar obstructiva crònica (MPOC), enfisema o bronquitis crònica. Categòrica (Yes/No)
- **HadDepressiveDisorder:** Historial de trastorn depressiu. Categòrica (Yes/No)
- **HadKidneyDisease:** Historial de malaltia renal. Categòrica (Yes/No)
- **HadArthritis:** Historial d'artritis. Categòrica (Yes/No)
- **HadDiabetes:** Historial de diabetis**.** Categòrica:
  1: "Yes",
  2: "Yes, but only during pregnancy (female)",
  3: "No",
  4: "No, pre-diabetes or borderline diabetes"
- **DeafOrHardOfHearing:** Pèrdua auditiva o dificultats serioses d'audició. Categòrica (Yes/No)

- **BlindOrVisionDifficulty:** Ceguesa o dificultats serioses de visió. Categòrica (Yes/No)
- **DifficultyConcentrating:** Dificultats serioses de concentració, record o presa de decisions. Categòrica (Yes/No)
- **DifficultyWalking:** Dificultats serioses per caminar o pujar escales. Categòrica (Yes/No)
- **DifficultyDressingBathing:** Dificultats per vestir-se o banyar-se. Categòrica (Yes/No)
- **DifficultyErrands:** Dificultats per realitzar tasques com fer recados sense ajuda. Categòrica (Yes/No)
- **SmokerStatus:** És fumador? Categòrica:
  1: "Current smoker - now smokes every day",
  2: "Current smoker - now smokes some days",
  3: "Former smoker",
  4: "Never smoked"
- **ECigaretteUsage:** Ús de cigarrets electrònics o productes de vapor electrònic. Categòrica:
  1: "Never used e-cigarettes in my entire life",
  2: "Use them every day",
  3: "Use them some days",
  4: "Not at all (right now)"
- **ChestScan**: Històrial de tomografia computaritzada (TC) o escàner de tòrax.
- **RaceEthnicityCategory:** Raça. Categòrica
  1: "White only, Non-Hispanic",
  2: "Black only, Non-Hispanic",
  3: "Other race only, Non-Hispanic",
  4: "Multiracial, Non-Hispanic",
  5: "Hispanic"
- **AgeCategory:** Categoria d'edat. Categòrica
  1: "Age 18 to 24",
  2: "Age 25 to 29",
  3: "Age 30 to 34",
  4: "Age 35 to 39",
  5: "Age 40 to 44",
  6: "Age 45 to 49",
  7: "Age 50 to 54",
  8: "Age 55 to 59",
  9: "Age 60 to 64",
  10: "Age 65 to 69",
  11: "Age 70 to 74",
  12: "Age 75 to 79",
  13: "Age 80 or older"
- **HeightInMeters:** Alçada en metres. Quantitativa
- **WeightInKilograms:** Pes en quilograms. Quantitativa
- **BMI:** Índex de massa corporal (IMC). Quantitativa
- **AlcoholDrinkers:** Consum d'alcohol en els últims 30 dies. Categòrica (Yes/No)
- **HIVTesting:** Prova de VIH realitzada. Categòrica (Yes/No)
- **FluVaxLast12:** Vacuna contra la grip rebuda en els últims 12 mesos. Categòrica (Yes/No)

- **PneumoVaxEver:** Vacuna contra la pneumònia rebuda en algun moment. Categòrica (Yes/No)
- **TetanusLast10Tdap:** Vacuna contra el tètanus rebuda en els últims 10 anys. Categòrica:
  1: "Yes, received Tdap",
  2: "Yes, received tetanus shot, but not Tdap",
  3: "Yes, received tetanus shot but not sure what type",
  4: "No, did not receive any tetanus shot in the past 10 years",
- **HighRiskLastYear:** Risc alt de malalties transmeses sexualment o conductes de risc. Categòrica (Yes/No)
- **CovidPos:** Història de resultat positiu en la prova de la COVID-19. Categòrica:
  1: "Yes",
  2: "No",
  3: "Tested positive using home test without a health professional"

Com es pot observar hi ha una gran quantitat de variables que ens permetran estudiar amb detall el pes que tenen a l'hora de determinar si una persona es propensa a tenir atacs de cor.

Algunes d'aquestes variables són:

HadHeartAttack: Aquesta variable indica si el participant ha tingut un atac de cor en el passat, el qual és un fort indicador de risc cardiovascular.

HadAngina: L'angina és un símptoma de malaltia coronària que pot indicar un risc de futurs problemes cardíacs.

HadStroke: Els accidents cerebrovasculars estan relacionats amb factors de risc similars als de les malalties cardíaques, i la seva presència pot ser un indicador important de salut cardiovascular.

HadDiabetes: La diabetis és un factor de risc ben conegut per a les malalties cardíaques i altres problemes de salut.

BMI: L'Índex de Massa Corporal és una mesura de l'índex de pes corporal, i un valor elevat pot indicar obesitat, que és un factor de risc important per a les malalties cardíaques.

AlcoholDrinkers: El consum d'alcohol en excés pot augmentar el risc de malalties cardíaques i altres problemes de salut.

SmokerStatus: El tabaquisme és un dels principals factors de risc per a les malalties cardíaques i altres malalties relacionades amb el tabac.

PhysicalActivities: L'exercici físic regular pot reduir el risc de malalties cardíaques i millorar la salut cardiovascular.

SleepHours: El patró de son inadequat pot estar associat amb un augment del risc de malalties cardíaques i altres problemes de salut.