# estudiHepatitis

## Descripció del dataset

El joc seleccionat s'ha obtingut al repositori de machine learning UCI, el seu títol és "Hepatitis Data Set". Aquest dataset és molt interessant per aquesta pràctica, ja que permetrà aplicar algorismes de machine learining tant supervisats com no supervisats. El dataset compta amb la classe a la qual pertany cada observació segons si ha mort o ha sobreviscut, aquesta etiqueta permetrà aplicar algorismes supervisats. El dataset compta amb variables tant qualitatives com quantitatives. Això serà interessant, ja que podrem aplicar tècniques de discretització a les variables quantitatives. És interessant tenir variables dels dos tipus per tant de provar diferents algorismes tant els que necessitin variables qualitatives com els q necessitin variables quantitatives. També hi ha motius d'interès personal per haver seleccionat aquest dataset, ja que em sembla especialment interessant com la ciència de dades pot ajudar al camp mèdic fent estudis per tal de diagnosticar o predir diverses malalties i així poder ajudar a molts humans. Aquest dataset es pot utilitzar amb diversos propòsits, per exemple crear algorismes tant de deep learining com de machine learining per ajudar a fer el seguiment dels pacients amb hepatitis, es pot predir si el pacient viurà o morirà, això pot ajudar a comprovar si el tractament que s'està seguint està funcionant o no. També es poden buscar les relacions que tenen els atributs, quines influeixen més pel diagnòstic.

Aquests són els atributs presents al dataset, els valors de les variables qualitatives estan representats numèricament però en aquesta descripció indicarem el significat d'aquests.

- **Class**: Classe a la qual pertany el pacient, viu o mor (Die/Life).
- **AGE**: Indica l'edat del pacient.
- **SEX**: Indica el sexe del pacient(Male/Female).
- **STEROID**: Indica si el pacient ha pres esteroides (yes/no).
- **ANTIVIRALS**: Indica si el pacient ha pres antivirals (yes/no).
- **FATIGUE**: Indica si el pacient és sent fatigat o no (yes/no).
- **MALAISE**: Indica si el pacient sent malestar (yes/no).
- **ANOREXIA**: Indica si el pacient pateix anorèxia (yes/no).
- **LIVER BIG**: Indica si la mida del fetge ha augmentat (yes/no).
- **LIVER FIRM**: Indica si el fetge és manté ferm (yes/no).
- **SPLEEN PAL**: Indica si el pacient presenta esplenomegàlia, una ampliació de la melsa (yes/no).
- **SPIDERS**: Indica si el pacient presenta aranyes(vasos sanguinis engrandits) visibles.
- **ASCITES**: Pesencia de líquid a la cavitat peritoneal (yes/no).
- **VARICES**: Indica si el pacient presenta varius (yes/no).
- **BILIRUBIN**: Indica el nivell de bilirubina del pacient.
- **ALK PHOSPH**: Indica el valor de la fosfatasa alcalina del pacient.
- **SGOT**: Valor obtingut amb l'anàlisi de sang del pacient després de la prova AST.
- **ALBUMIN**: Indica el valor de la proteïna albúmina del pacient.
- **PROTIME**: Indica el valor de la característica del pacient.

- **HISTOLOGY**: Indica el valor de'estudiar la histología (estudis microscopics)(yes/no).

## Neteja de les dades

Llegim el fitxer.

```
dhep <- read.csv("data/hepatitis.csv")
dattrs <- c("Class","AGE","SEX","STEROID","ANTIVIRALS","FATIGUE","MALAISE","ANOREXIA","LIVER_BIG","LIVE

names(dhep) <- dattrs
```

Com podem veure el dataset compta amb 154 files amb 20 variables (columnes).

```
dim(dhep)
```

```
## [1] 154  20
```

Primer mirem quin tipus s'ha assignat a cada columna. Es pot veure com hi ha variables que s'han llegit amb el tipus erroni, haurem de corregir el tipus assignat.

```
str(dhep)
```

```
## 'data.frame':    154 obs. of  20 variables:
##  $ Class        : int  2 2 2 2 2 1 2 2 2 2 ...
##  $ AGE          : int  50 78 31 34 34 51 23 39 30 39 ...
##  $ SEX          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ STEROID      : Factor w/ 3 levels "?","1","2": 2 3 1 3 3 2 3 3 3 2 ...
##  $ ANTIVIRALS   : int  2 2 1 2 2 2 2 2 2 1 ...
##  $ FATIGUE      : Factor w/ 3 levels "?","1","2": 2 2 3 3 3 2 3 2 3 3 ...
##  $ MALAISE      : Factor w/ 3 levels "?","1","2": 3 3 3 3 3 3 3 3 3 3 ...
##  $ ANOREXIA     : Factor w/ 3 levels "?","1","2": 3 3 3 3 3 2 3 3 3 3 ...
##  $ LIVER_BIG    : Factor w/ 3 levels "?","1","2": 2 3 3 3 3 3 3 3 3 2 ...
##  $ LIVER_FIRM   : Factor w/ 3 levels "?","1","2": 3 3 3 3 3 3 3 3 2 3 2 ...
##  $ SPLEEN_PAL   : Factor w/ 3 levels "?","1","2": 3 3 3 3 3 3 2 3 3 3 3 ...
##  $ SPIDERS      : Factor w/ 3 levels "?","1","2": 3 3 3 3 3 3 2 3 3 3 3 ...
##  $ ASCITES      : Factor w/ 3 levels "?","1","2": 3 3 3 3 3 3 3 3 3 3 3 ...
##  $ VARICES      : Factor w/ 3 levels "?","1","2": 3 3 3 3 3 3 3 3 3 3 3 ...
##  $ BILIRUBIN    : Factor w/ 35 levels "?","0.30","0.40",..: 8 6 6 9 8 1 9 6 9 12 ...
##  $ ALK_PHOSPHATE: Factor w/ 84 levels "?","100","102",..: 19 84 51 1 83 1 1 1 1 73 ...
##  $ SGOT         : Factor w/ 85 levels "?","100","101",..: 54 48 62 31 44 1 1 60 8 46 ...
##  $ ALBUMIN      : Factor w/ 30 levels "?","2.1","2.2",..: 13 18 18 18 18 1 1 22 17 22 ...
##  $ PROTIME      : Factor w/ 45 levels "?","0","100",..: 1 1 42 1 38 1 1 1 1 44 ...
##  $ HISTOLOGY    : int  1 1 1 1 1 1 1 1 1 1 ...
```

Canviem el tipus de les variables numèriques que havien estat llegides com a factor.

```
quantattrs <- c("AGE","BILIRUBIN","ALK_PHOSPHATE","SGOT","ALBUMIN","PROTIME")
for (i in quantattrs){
  dhep[,i] <- as.numeric(dhep[,i])
}
```

Canviem el tipus de les variables factor que han estat llegides incorrectament.

```
factattrs <- c("Class","AGE","SEX")
for (i in factattrs){
  dhep[,i] <- as.numeric(dhep[,i])
}
```

Definim dos vectors amb les variables categòriques i numèriques.

```
quantattrs <- c("AGE","BILIRUBIN","ALK_PHOSPHATE","SGOT","ALBUMIN","PROTIME")
catattrs <- c("Class","SEX","STEROID","ANTIVIRALS","FATIGUE","MALAISE","ANOREXIA","LIVER_BIG","LIVER_FI
```

## Valors desconeguts

En la primera insepecció hem pogut veure com hi ha valors desconeguts representats amb el signe interrogant. Mirarem quines columnes tenen valors desconeguts.

```
colSums(dhep == "?")
```

```
##         Class           AGE           SEX       STEROID    ANTIVIRALS
##             0             0             0             1             0
##       FATIGUE       MALAISE      ANOREXIA     LIVER_BIG    LIVER_FIRM
##             1             1             1            10            11
##    SPLEEN_PAL       SPIDERS       ASCITES       VARICES     BILIRUBIN
##             5             5             5             5             0
## ALK_PHOSPHATE          SGOT       ALBUMIN       PROTIME     HISTOLOGY
##             0             0             0             0             0
```

Mirem si hi ha algun valor NA. Podem veure que no.

```
colSums(is.na(dhep))
```

```
##         Class           AGE           SEX       STEROID    ANTIVIRALS
##             0             0             0             0             0
##       FATIGUE       MALAISE      ANOREXIA     LIVER_BIG    LIVER_FIRM
##             0             0             0             0             0
##    SPLEEN_PAL       SPIDERS       ASCITES       VARICES     BILIRUBIN
##             0             0             0             0             0
## ALK_PHOSPHATE          SGOT       ALBUMIN       PROTIME     HISTOLOGY
##             0             0             0             0             0
```

Canviarem el valor de l'interrogant per NA. Guardem en un vector tots els atributs que contenen valors desconeguts i seguidament apliquem el canvi.

```
missingattrs <- names(which(colSums(dhep == "?") > 0))

for (m in missingattrs){
  dhep[which(dhep[,m] == '?'),m] <- NA
}
```

El primer que farem serà imputar els valors perduts que tenim al dataset, com que aquest dataset no és excessivament gran preferim aproximar els valors en comptes d'eliminar les observacions que compten amb mising values és a dir, assumirem un petit grau d'error a canvi de mantenir més observacions. A l'hora d'aplicar algorismes supervisats és interessant tenir grans datasets ja que es poden dividir en el dataset de train i el de test. En el primer apartat d'anàlisi exploratori hem comprovat com hi ha diversos atributs que tenen valors perduts. Per tal d'imputar els valors perduts utilitzarem el mètode missForest ja que últimament està guanyant popularitat i s'utilitza amb variables mixtes. Encara que el kNN és un dels més populars aquest, és molt sensible a la k que es tria.

Eliminem la columna que no volem utilitzar.

```
#dhep$PROTIME <- NULL
```

Imputem els valors de les variables. Carreguem la llibreria i apliquem la funció.

```
library("missForest")
```

```
## Loading required package: randomForest
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
## Loading required package: foreach
```

```
## Loading required package: itertools
```

```
## Loading required package: iterators
```

```
mf.res <- missForest(dhep, variablewise = TRUE)
```

```
##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
##   missForest iteration 4 in progress...done!
```

Podem veure com tots els valors NA han desaparegut, han estat imputats. També podem obtenir informació sobre l'error, es pot veure com les columnes que no tenien cap valor per imputar tenen un error de 0, les columnes on s'han imputat valors presenten els seus corresponents errors, això pot afectar a l'estudi ja que els valors no són reals simplement són suposicions.

```
dhep <- mf.res$ximp
colSums(is.na(dhep))
```

```
##         Class           AGE           SEX       STEROID    ANTIVIRALS
##             0             0             0             0             0
##       FATIGUE      MALAISE     ANOREXIA     LIVER_BIG    LIVER_FIRM
##             0             0             0             0             0
##    SPLEEN_PAL       SPIDERS       ASCITES       VARICES     BILIRUBIN
##             0             0             0             0             0
## ALK_PHOSPHATE          SGOT       ALBUMIN       PROTIME     HISTOLOGY
##             0             0             0             0             0
```

```
mf.res$OOBerror
```

```
##       MSE       MSE       MSE       PFC       MSE       PFC       PFC       PFC
## 0.0000000 0.0000000 0.0000000 0.4444444 0.0000000 0.1960784 0.2156863 0.1830065
##       PFC       PFC       PFC       PFC       PFC       PFC       MSE       MSE
## 0.1875000 0.2797203 0.2818792 0.2416107 0.1073826 0.1275168 0.0000000 0.0000000
##       MSE       MSE       MSE       MSE
## 0.0000000 0.0000000 0.0000000 0.0000000
```

### Valors extrems - outliers

Mostrem una priemra descripció estadística.

```
summary(dhep)
```

```
##      Class             AGE              SEX        STEROID   ANTIVIRALS
##  Min.   :1.000   Min.   : 7.00   Min.   :1.000   ?: 0    Min.   :1.000
##  1st Qu.:2.000   1st Qu.:32.00   1st Qu.:1.000   1:75    1st Qu.:2.000
##  Median :2.000   Median :39.00   Median :1.000   2:79    Median :2.000
##  Mean   :1.792   Mean   :41.27   Mean   :1.097           Mean   :1.844
##  3rd Qu.:2.000   3rd Qu.:50.00   3rd Qu.:1.000           3rd Qu.:2.000
##  Max.   :2.000   Max.   :78.00   Max.   :2.000           Max.   :2.000
##  FATIGUE MALAISE ANOREXIA LIVER_BIG LIVER_FIRM SPLEEN_PAL SPIDERS ASCITES
##  ?:  0   ?: 0    ?:  0    ?:  0     ?: 0       ?:  0      ?:  0   ?:  0
##  1:100   1:61    1: 32    1: 24     1:63       1: 30      1: 52   1: 21
##  2: 54   2:93    2:122    2:130     2:91       2:124      2:102   2:133
##
##
##
##  VARICES   BILIRUBIN     ALK_PHOSPHATE        SGOT          ALBUMIN
##  ?:  0   Min.   : 1.00   Min.   : 1.00   Min.   : 1.00   Min.   : 1.00
##  1: 18   1st Qu.: 6.00   1st Qu.: 8.25   1st Qu.:29.25   1st Qu.: 9.25
##  2:136   Median : 9.00   Median :40.50   Median :47.50   Median :17.00
##          Mean   :11.45   Mean   :40.13   Mean   :46.60   Mean   :14.70
##          3rd Qu.:14.00   3rd Qu.:71.00   3rd Qu.:66.75   3rd Qu.:20.00
##          Max.   :35.00   Max.   :84.00   Max.   :85.00   Max.   :30.00
##     PROTIME        HISTOLOGY
##  Min.   : 1.00   Min.   :1.000
##  1st Qu.: 1.00   1st Qu.:1.000
##  Median : 3.00   Median :1.000
##  Mean   :13.36   Mean   :1.455
##  3rd Qu.:25.75   3rd Qu.:2.000
##  Max.   :45.00   Max.   :2.000
```

Analitzem els valors extrems, comprovem quins atributs de tipus quantitatiu presenten possibles valors extrems. Com veiem els atributs AGE i BILIRUBIN tenen alguns valors extrems, en el proper apartat valorarem si realment són tan extrems i prendrem decissions sobre com tractar-ho.

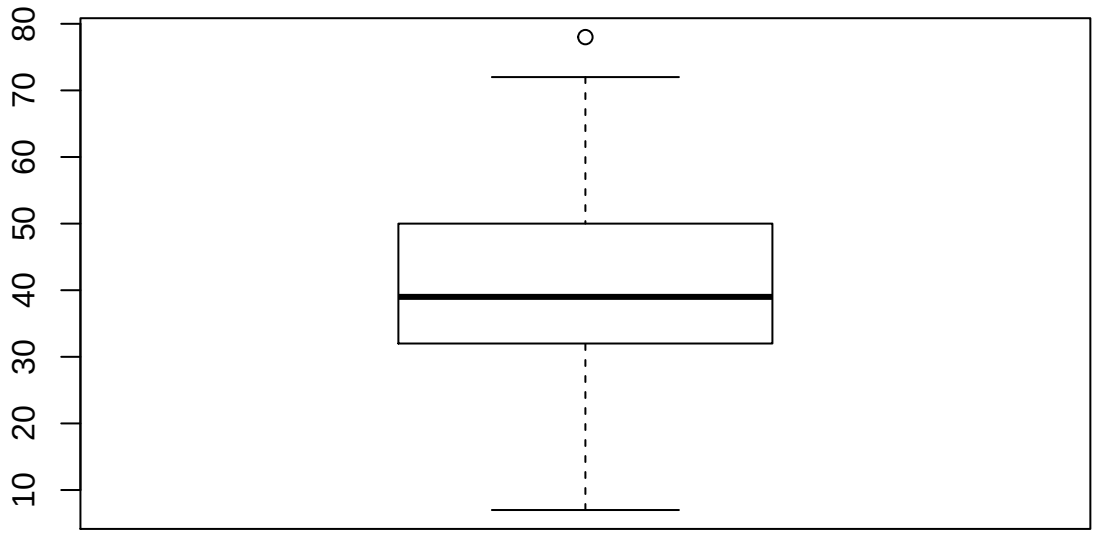
```
for (col in quantattrs){
  cat("col:",col,"Outliers:",length(boxplot.stats(dhep[,col])$out),"\n")
}
```

```
## col: AGE Outliers: 1
## col: BILIRUBIN Outliers: 13
## col: ALK_PHOSPHATE Outliers: 0
## col: SGOT Outliers: 0
## col: ALBUMIN Outliers: 0
## col: PROTIME Outliers: 0
```

Procedim a analitzar cadascun dels valors que es podria considerar outlier.

En aquest primer cas la variable AGE representa l'edat, un valor de 78 anys no es pot considerar un outiler, mantindrem aquest valor.

```r
boxplot(dhep$AGE)
```

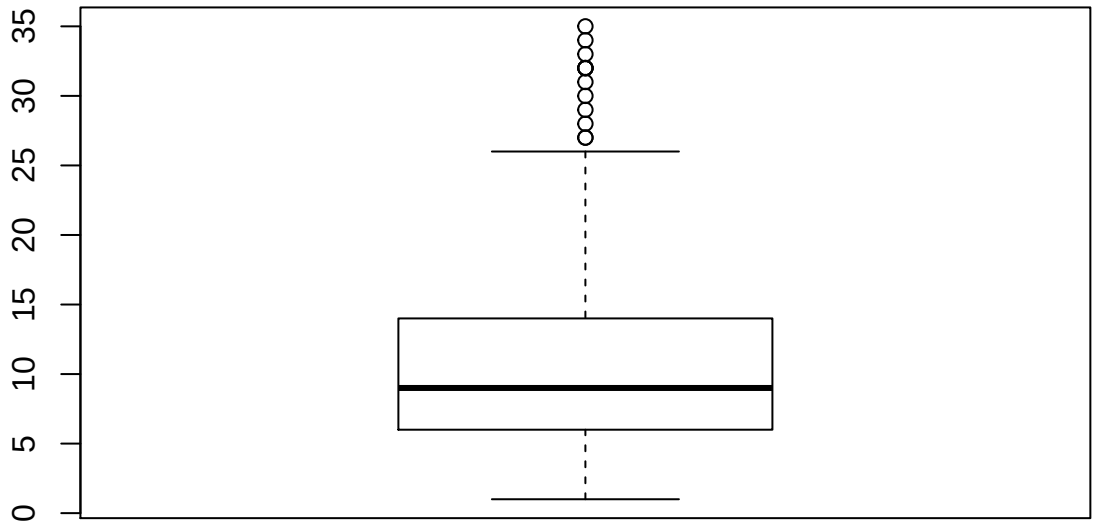



```r
boxplot.stats(dhep$AGE)$out
```

```
## [1] 78
```

En aquest segon cas el nivell de bilirubina és més complex. Cal fer recerca sobre quins nivells màxims i mínims són possibles en pacient d'hepatitis. Segons les fonts consultades els valors més elevats de bilirubina indiquen problemes més greus, en propers apartats buscarem la relació de la bilirubina amb la vida o mort del pacient utilitzant testos d'estadística inferencial. Els nivells normals de bilirubina varien en un rang d'1 fins a 1.2 (mg/dL), a partir dels 2(mg/dL) la pell agafa un color groguenc. Sembla que aquests són valors molt elevats però dintre un rang possible, ja que s'indica que a partir dels 30 mg/dL el pacient es troba en estat molt crític.

```r
boxplot(dhep$BILIRUBIN)
```




```
boxplot.stats(dhep$BILIRUBIN)$out
```

```
##  [1] 32 28 30 32 33 32 27 27 32 35 29 31 34
```

```
summary(dhep$BILIRUBIN)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    6.00    9.00   11.45   14.00   35.00
```

Com hem vist els valors dels nivells de les variables categòriques no són explicatius, actualment es troben representats amb valors numèrics, canviarem aquests valors, ja que pot ser útil quan apliquem futurs algorismes i vulguem extreure conclusions.

```
dhep[dhep$Class == 1,"Class"] <- 'Die'
dhep[dhep$Class == 2,"Class"] <- 'Live'

dhep[dhep$SEX == 1,"SEX"] <- 'Male'
dhep[dhep$SEX == 2,"SEX"] <- 'Female'

dhep$SEX <- as.factor(dhep$SEX)
dhep$Class <- as.factor(dhep$Class)
```

```
# Transformem a caràcter. canviem el valors dels nivells i tornem a transformar a factor, així eliminem
for (i in catattrs[- which (catattrs %in% list("Class","SEX"))] ){
  dhep[,i] <- as.character(dhep[,i])

  dhep[ dhep[,i] == 1,i] <- 'Yes'
  dhep[ dhep[,i] == 2,i] <- 'No'

  dhep[,i] <- as.factor(dhep[,i])
}
```

# Anàlisi de les dades

# Seleccio dels grups a analitzar

En aquest apartat prepararem grups que poden ser d'interés per tal d'analitzar o comparar. En futurs apartats els utilitzarem per tal d'extreure conclusions.

Pacients que han mort d'hepatitis.

```
dhep.die <- dhep[dhep$Class == "Die",]
```

Pacients que han soberviscut a l'hepatitis.

```
dhep.live <- dhep[dhep$Class == "Live",]
```

## Proves estadístiques

### Comprovació de normalitat i la homogenïtat de la variancia

Aplicarem un test de Shapiro a tots els atributs quantitatius per tal de veure en quins casos podem assumir normalitat i en quins no.

$$H_0 : La mostra prové d'una població amb distribució normal$$

$$H_1 : La mostra no prové d'una població amb distribució normal$$

```
lapply(quantattrs,function(x) shapiro.test(dhep[,x]))
```

```
## [[1]]
##
##  Shapiro-Wilk normality test
##
## data:  dhep[, x]
## W = 0.98535, p-value = 0.1034
##
##
## [[2]]
##
##  Shapiro-Wilk normality test
```

```
##
## data:  dhep[, x]
## W = 0.83831, p-value = 9.519e-12
##
##
## [[3]]
##
##  Shapiro-Wilk normality test
##
## data:  dhep[, x]
## W = 0.87412, p-value = 4e-10
##
##
## [[4]]
##
##  Shapiro-Wilk normality test
##
## data:  dhep[, x]
## W = 0.95884, p-value = 0.0001549
##
##
## [[5]]
##
##  Shapiro-Wilk normality test
##
## data:  dhep[, x]
## W = 0.93308, p-value = 1.235e-06
##
##
## [[6]]
##
##  Shapiro-Wilk normality test
##
## data:  dhep[, x]
## W = 0.79285, p-value = 1.724e-13
```

**Variables quantitatives que influeixen més en si el pacient soberviurà o morirà**

**Variables qualitatives que influeixen més en si el pacient soberviurà o morirà**

## Regressió logística