Tipologia i cicle de vida de les dades

PRÀCTICA 2 - Neteja i anàlisi de dades

Sergio Costa i Lorena Casanova

Gener 2020

Índex

1 Introducció

1.1 Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes.

1.2 Competències

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science: * Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l. * Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi.

1.3 Objectius

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contexts més amplis o multi-disciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.
- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera auto-dirigida o autònoma.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

1.4 Descripció de la pràctica a realitzar

En aquesta pràctica es realitzarà una cerca d'un conjunt de dades en la qual ens sigui possible aplicar diferents tècniques de tractament, processat i anàlisi de dades mitjançant la implementació de funcions i algoritmes apresos durant l'assignatura de tipologia i cicle de vida de les dades i altres que composen el màster de Ciència de Dades.

El conjunt de dades serà extret de la plataforma Kaggle, un repositori de dades de diferents tipologies pensada per a que els usuaris publiquin els seus projectes i entrin en competicions entre ells, de manera que es converteixi en una entorn en el qual científics de dades el puguin utilitzar com a recurs per a desenvolupar el seu treball.

2 Pràctica

2.1 Descripció del dataset

El conjunt de dades triat per aquesta pràctica és el *Indian Liver Patient Records* (o *Indian Liver dataset*) extret del repositori de *Kaggle* i accessible a través de l'enllaç *https://www.kaggle.com/uciml/indian-liver-patient-records*.

El joc de dades conté els resultats d'anàlisi de mostres biològiques d'un total de 583 individus, 416 pacients amb malalties hepàtiques i 167 sense patologia, procedents del nord-est d'Andhra Pradesh, Índia. En total inclou 10 atributs, tots numèrics menys el sexe de l'individu i un 11è atribut classe binari que indica si l'individu presenta o no patologia de fetge.

Amb les dades disponibles fixem l'objectiu d'esbrinar l'influencia de les diferents variables en el desenvolupament de patologies de fetge i desenvolupar un model que ens permeti de predir, és a dir, classificar, si un individu pateix una patologia de fetge o no en base als diferents atributs de què disposem, en concret: edat del pacient, sexe, diferents valors de les mostres biològiques.

Atès que les malalties hepàtiques acostumen a ser greus l'interès del nostre objectiu rau en avaluar la capacitat que tenim de detectar aquestes patologies en pacients al més aviat possible. Plantegem la hipòtesis que els valors obtinguts a través de les anàlisis de sang de què disposem poden ser suficients per a detectar individus malalts malgrat encara no hagin desenvolupat una simptomatologia evident o, si més no, podem desenvolupar un model que ens permeti fer una primera cribra dels pacients candidats a anàlisis posteriors. Volem donar resposta, doncs, a les següents preguntes:

- Existeix una diferencia estadísticament significativa en la proporció de malalts de fetge en funció del sexe?
- Existeix una diferència estadísticament significativa en la mitjana d'edat entre els malalts i els individus sans?
- Quina és la influència dels diferents factors (atributs de què disposem) en la probabilitat de desenvolupar una patologia de fetge?
- Quina capacitat tenim de predir si un individu pateix o no una malaltia de fetge a partir dels atributs que disposem?

Per tal de donar resposta a aquestes preguntes s'apliquen diferents mètodes d'anàlisi estadístic sobre el conjunt de dades tot comprovant si es compleixen les condicions necessàries per a la validesa dels resultats. Atès que els objectius definits giren al voltant d'un problema de classificació i disposem d'una variable classe de control (*Desease*) farem ús d'algoritmes d'aprenentatge supervisat per tal de construir el model final de classificació.

Els diferents mètodes que s'utilitzaran per a realitzar l'anàlisi estadístic són els següents:

- Test d'hipòtesis -> Comparar mostres d'Homes/Dones i Malalts-Sans per donar resposta a les preguntes plantejades.
- Regressió logística -> Avaluar la influència de cada variable en la probabilitat de tenir afecció de fetge i predir el valor resultant.
- Random Forest -> Construir un model de classificació que ens permeti de predir la variable Disease a partir de la resta d'atributs.

Els atributs de què disposem són:

1. Age: Edat del pacient
2. Sex: Sexe del pacient
3. Tot_bil: Nivell de bilirrubina total
4. Dir_bil: Nivell de bilirrubina directa
5. Alkphos: Nivell de fosfotasa alcalina
6. Alamine: Nivell d'aminotransferasa alanina
7. Aspartate: Nivell d'aminotransferasa aspartat
8. Tot_Prot: Nivell de proteïnes totals
9. Albumin: Nivell d'albunima
10. A_G_Ratio: Ratio albumina-globulina
11. Disesase: Estat de malaltia

2.2 Integració i selecció de les dades d'interès a analitzar

2.2.1 Càrrega del joc de dades

Procedim a carregar i comprovar el joc de dades seleccionat:

```r
# Llegim el dataset
liver_data <- read.csv("indian_liver_patient.csv", header = TRUE)

# Assignem el nom de les variables segons la informació del dataset
colnames(liver_data) <- c("Age", "Sex", "Tot_Bil", "Dir_Bil", "Alkphos", "Alamine",
    "Aspartate", "Tot_Prot", "Albumin", "A_G_Ratio", "Disease")

# Comprovem l'estructura del dataset resultant
str(liver_data)
```

```
## 'data.frame':    583 obs. of  11 variables:
##  $ Age      : int  65 62 62 58 72 46 26 29 17 55 ...
##  $ Sex      : chr  "Female" "Male" "Male" "Male" ...
##  $ Tot_Bil  : num  0.7 10.9 7.3 1 3.9 1.8 0.9 0.9 0.9 0.7 ...
##  $ Dir_Bil  : num  0.1 5.5 4.1 0.4 2 0.7 0.2 0.3 0.3 0.2 ...
##  $ Alkphos  : int  187 699 490 182 195 208 154 202 202 290 ...
##  $ Alamine  : int  16 64 60 14 27 19 16 14 22 53 ...
##  $ Aspartate: int  18 100 68 20 59 14 12 11 19 58 ...
##  $ Tot_Prot : num  6.8 7.5 7 6.8 7.3 7.6 7 6.7 7.4 6.8 ...
##  $ Albumin  : num  3.3 3.2 3.3 3.4 2.4 4.4 3.5 3.6 4.1 3.4 ...
##  $ A_G_Ratio: num  0.9 0.74 0.89 1 0.4 1.3 1 1.1 1.2 1 ...
##  $ Disease  : int  1 1 1 1 1 1 1 1 2 1 ...
```

```r
# Mostrem la distribució per sexe i malaltia
table(liver_data$Disease)
```

```
##
##   1   2
## 416 167
```

Aquests resultats corroboren la informació que s'havia trobat a Kaggle juntament en la descàrrega del fitxer de dades: un total de 583 mostres procedents de 11 variables diferents.

S'observa l'estructura del joc de dades i els diferents atributs que inclou, la majoria dels quals són numèrics. Per tal de facilitar la lectura i el posterior anàlisi de les dades es mapeja la columna sexe i s'abrevien els valors de Male/Female a M/F.

Quant a l'atribut que informa sobre la presència de malaltia hepàtica retorna un 1 o un 2, sabent que el dataset conté un total de 416 pacients amb malaltia hepàtica i 167 sans, es mapejen els valors a Y/N (Sí o No presenta malaltia).

```r
liver_data$Sex <- as.factor(ifelse(liver_data$Sex == "Male", "M", "F"))
liver_data$Disease <- as.factor(ifelse(liver_data$Disease == 2, "N", "Y"))
```

2.2.2 Anàlisi exploratori inicial

En l'anàlisi descriptiu es pretén visualitzar les dades i analitzar la seva estructura i tendència a través de gràfiques i resums. El que es vol és tenir el coneixement de les variables del nostre conjunt de dades, per tal de saber com netejar-les, extraure les característiques més representatives i interpretar-les posteriorment.

```r
# Observem les dades
summary(liver_data)
```

```
##       Age        Sex        Tot_Bil          Dir_Bil          Alkphos
##  Min.   : 4.00   F:142   Min.   : 0.400   Min.   : 0.100   Min.   :  63.0
##  1st Qu.:33.00   M:441   1st Qu.: 0.800   1st Qu.: 0.200   1st Qu.: 175.5
##  Median :45.00           Median : 1.000   Median : 0.300   Median : 208.0
##  Mean   :44.75           Mean   : 3.299   Mean   : 1.486   Mean   : 290.6
```

```
##  3rd Qu.:58.00           3rd Qu.: 2.600   3rd Qu.: 1.300   3rd Qu.: 298.0
##  Max.   :90.00           Max.   :75.000   Max.   :19.700   Max.   :2110.0
##
##     Alamine          Aspartate         Tot_Prot        Albumin
##  Min.   :  10.00   Min.   :  10.0   Min.   :2.700   Min.   :0.900
##  1st Qu.:  23.00   1st Qu.:  25.0   1st Qu.:5.800   1st Qu.:2.600
##  Median :  35.00   Median :  42.0   Median :6.600   Median :3.100
##  Mean   :  80.71   Mean   : 109.9   Mean   :6.483   Mean   :3.142
##  3rd Qu.:  60.50   3rd Qu.:  87.0   3rd Qu.:7.200   3rd Qu.:3.800
##  Max.   :2000.00   Max.   :4929.0   Max.   :9.600   Max.   :5.500
##
##    A_G_Ratio      Disease
##  Min.   :0.3000   N:167
##  1st Qu.:0.7000   Y:416
##  Median :0.9300
##  Mean   :0.9471
##  3rd Qu.:1.1000
##  Max.   :2.8000
##  NA's   :4
```

Veiem que per la variable Age tenim valors des dels 4 anys fins als 90 (sabem de la informació extreta del repositori que tots els individus de més de 89 anys estan marcats com a 90). Tenim una mitjana d'edat de 44.75 anys i el 50% dels casos es troben entre els 33 i els 58 anys.

Observem també que la variable A_G_Ratio presenta 4 valors perduts, en tindrem cura més endavant.

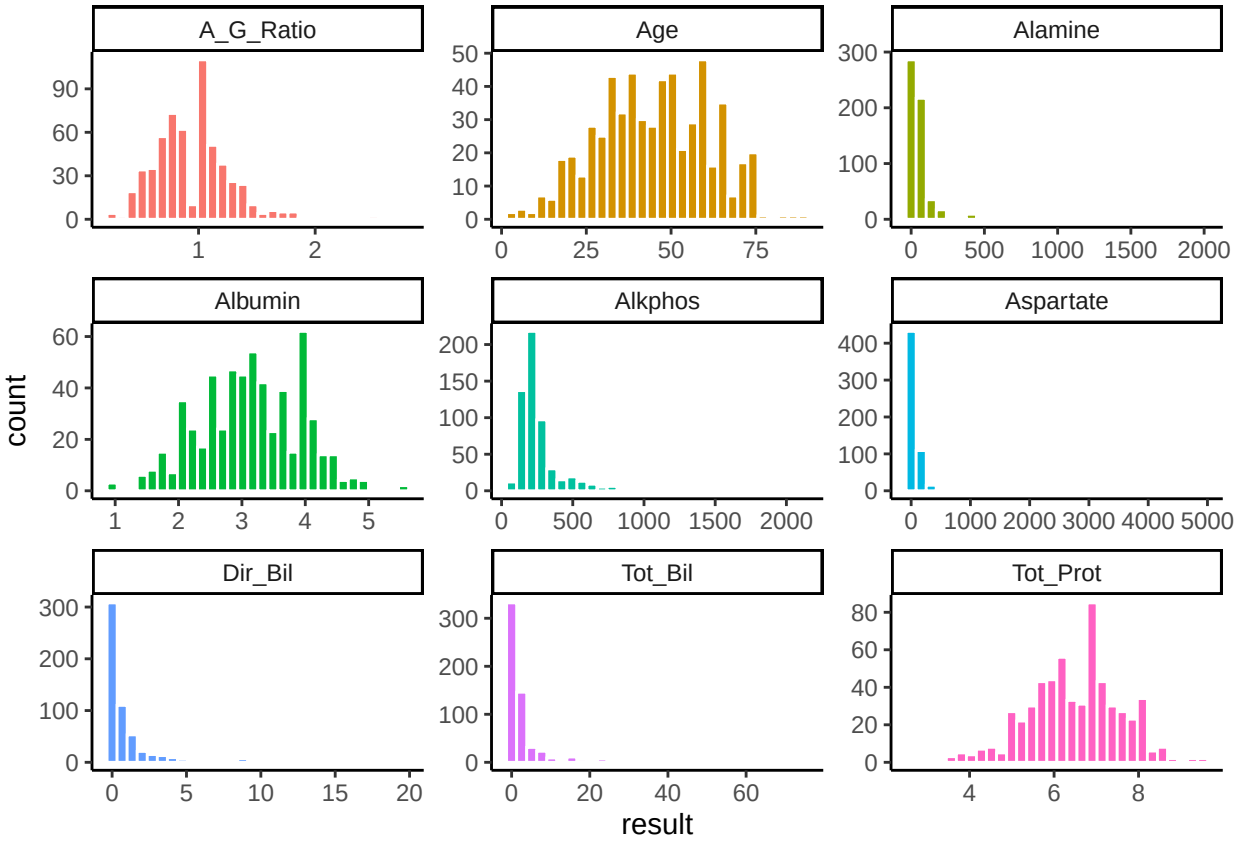
En el cas del gènere, tal com ens indicava la descripció del dataset, tenim una distribució de 2/3 d'homes en front 1/3 de dones, clarament sota-representades.

Quant a les anàlisi de diferents atributs sobten els valors màxims per als nivells de Bilirubina, fosfatasa alcalina, alamina, aspartat i A_G_Ratio, els haurem d'estudiar amb més cura.

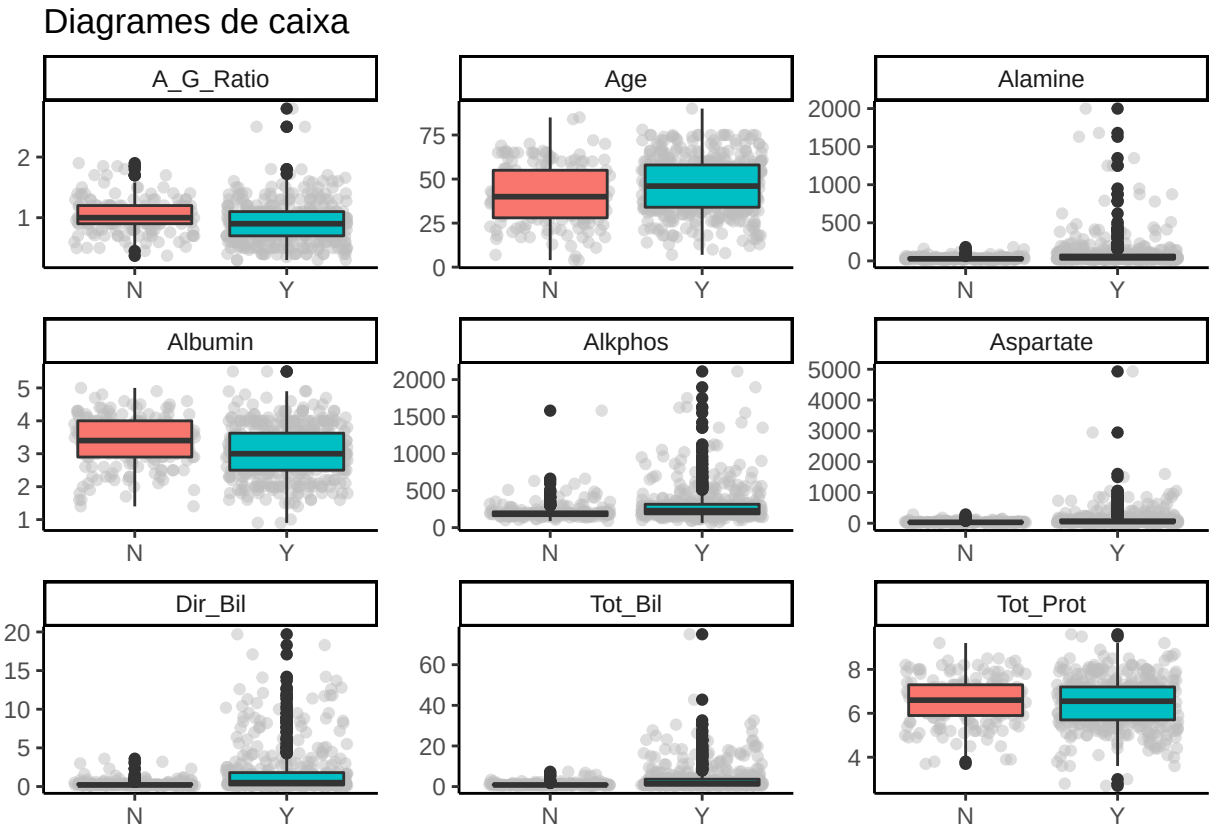
Quant a l'atribut classe Disease veiem que la distribució és 416 malalts en front 167 individus sans. Cal notar que ens trobem davant un joc de dades poc equilibrat respecte a l'atribut classe, fet que caldrà tenir en compte a l'avaluar el rendiment dels models de classificació, atès que tenim per tant una probabilitat per defecte del 71%.

Continuarem l'anàlisi visualitzant la distribució de les diferents variables en un histograma i en diagrames de caixes.

```
# Mostrem el panell d'histogrames
liver_data %>% gather(c(1, 3:10), key = "variables", value = "result") %>% ggplot(aes(result)) +
    geom_histogram(aes(fill = variables), color = "white", bins = 30) + theme_classic() +
    facet_wrap(. ~ variables, scale = "free") + theme(legend.position = "none")
```




```
# Mostrem el panell de diagrames de caixa
liver_data %>% pivot_longer(c(1, 3:10), names_to = "variables", values_to = "result") %>%
    ggplot(aes(Disease, result, fill = Disease)) + geom_jitter(color = "grey", alpha = 0.5) +
    geom_boxplot() + labs(x = element_blank(), y = element_blank()) + theme_classic() +
    facet_wrap(. ~ variables, scale = "free") + ggtitle("Diagrames de caixa") + theme(legend.position = "none
```



- **A_G_Ratio**: La variable presenta una distribució propera a la normal lleugerament asimètrica cap a la dreta. El diagrama de caixes ens mostra la presència de valors possiblement atípics a l'extrem superior.
- **Age**: La variable presenta una distribució força simètrica si bé hi ha escassos registres per sobre dels 75 anys. Cal tenir en compte que la descripció del joc de dades indica que per a individus de més de 90 anys l'edat consta com 90, tot i això no veiem una acumulació de valors en 90. La variable no sembla contenir valors atípics.
- **Albumin**: La variable presenta una distribució força simètrica i no destaquen valors atípics.
- **Tot_Protein**: Els valors de proteïnes totals descriuen una distribució simètrica si bé es detecten valors extrems en ambdós costats, s'avaluarà posteriorment si es consideren valors atípics.

Les variables *Alamine*, *Alkphos*, *Aspartate*, *Dir_Bil* i *Tot_Bil* presenten distribucions fortament asimètriques cap a la dreta amb cues molt llargues amb distribucions similars a una logarítmica. La identificació de valors atípics ens aquests casos s'ha de fer amb cura, si bé en alguns casos (*Alamine*, *Aspartate* i *Tot_Bil*) s'identifiquen alguns valors aïllats a l'extrem superior.

Haurem de tenir en compte la distribució d'aquestes variables a l'hora d'aplicar els mètodes estadístics per no violar les assumpcions de normalitat, si fora el cas, fet que invalidaria les estimacions de sifgnificància de les inferències i, en definitiva, la validesa dels resultats.

Si fora necessari podem aplicar transformacions a les variables, bé una transformació logarítmica o bé aplicar el mètode de BoxCox per assajar un conjunt de transformacions exponencials, en qualsevol cas, una transformació aplicada a la variable en dificultarà la interpretació dels resultats; les aplicarem, doncs, només en cas que sigui necessari per a garantir la validesa de les assumpcions.

2.3 Neteja de les dades

És freqüent que en l'adquisició de dades es produeixin errors els quals poden desvirtuar la interpretació d'aquestes. La neteja de les dades consisteix en detectar aquestes anomalies de les dades com poden ser la presència de zeros o elements buits i la identificació dels valors extrems. Hi ha diferents maneres de tractar amb aquest tipus d'errors i segons el cas s'estudiarà la millor sol · lució.

Presència de zeros o elements buits

Analitzem la presència de valors nuls.

```
colSums(is.na(liver_data))
```

```
##       Age       Sex   Tot_Bil   Dir_Bil   Alkphos   Alamine Aspartate  Tot_Prot
##         0         0         0         0         0         0         0         0
##   Albumin A_G_Ratio   Disease
##         0         4         0
```

En aquest cas s'observa que tan sols la variable A_G_Ratio conté valors nuls. Analitzem la presència de valors buits o zeros.

```
colSums(na.omit(liver_data) == "" | na.omit(liver_data) == "0")
```

```
##       Age       Sex   Tot_Bil   Dir_Bil   Alkphos   Alamine Aspartate  Tot_Prot
##         0         0         0         0         0         0         0         0
##   Albumin A_G_Ratio   Disease
##         0         0         0
```

No hi ha presència de valors buits ni zeros, la qual cosa voldrà dir que tan sols s'hauran de tractar els Nuls de la variable A_G_Ratio. S'observen alguns valors estadístics de la variable.

```
summary(liver_data$A_G_Ratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.3000  0.7000  0.9300  0.9471  1.1000  2.8000       4
```

Atès que, com hem vist abans, la variable presenta una distribució força normal, probablement la millor sol · lució en tan sols 4 mostres amb valors nuls seria aplicar la mitjana sobre aquests valors per tal de no desvirtuar els resultats ni perdre informació. No obstant, es fa un anàlisi més a fons observant la variable per tal de veure el seu comportament en els següents apartats i per prendre una decisió.

Identificació i tractament de valors extrems

Procedim ara a identificar i tractar els valors extrems presents en les dades. A l'analitzar els valors atípics hem de tenir en compte que aquests poden ser de naturalesa diversa, poden ser valors incorrectes fruit d'un error a les dades o representar casos extrems, però reals. El tractament que en farem dependrà en molts casos del seu orígen, així com en el primer cas és clar que haurem d'actuar sobre els valors, bé eliminant-los o bé imputant un altre valor, en el segon cas haurem de tenir més cura atès que cal tenir un coneixement profund del domini de les dades.

- **A_G_Ratio**:

Al panell de diagrames de caixa anterior hem identificat valors extrems a la variable A_G_Ratio, dos dels quals tenen uns valors mol allunyats de la mitjana. Observem la resta de valors quan aquesta variable pren un valor superior a 2.

```
filter(liver_data, A_G_Ratio > 2)
```

```
##   Age Sex Tot_Bil Dir_Bil Alkphos Alamine Aspartate Tot_Prot Albumin A_G_Ratio
## 1  42   M    11.1     6.1     214      60       186      6.9     2.8       2.8
## 2  32   M    15.6     9.5     134      54       125      5.6     4.0       2.5
## 3  32   M    25.0    13.7     560      41        88      7.9     2.5       2.5
##   Disease
## 1       Y
## 2       Y
## 3       Y
```

En dos dels casos que hi ha valors extrems el valor de la variable coincideix amb el de l'Albúmina. Per tant, es pot intuir que aquest valor pot ser un duplicat de la variable Albúmina. Si es sumen aquests tres valors extrems a els 4 nuls que s'han observat en l'apartat anterior representa l'1.2% de totes les dades, per tant es decideix imputar el valor de la mitjana.

```
liver_data$A_G_Ratio[liver_data$A_G_Ratio > 2] <- mean(liver_data$A_G_Ratio[liver_data$A_G_Ratio <
    2], na.rm = T)
liver_data$A_G_Ratio[is.na(liver_data$A_G_Ratio)] <- mean(liver_data$A_G_Ratio, na.rm = T)
sum(is.na(liver_data$A_G_Ratio))
```

```
## [1] 0
```

- **Variables *Alamine*, *Alkphos*, *Aspartate*, *Dir_Bil* i *Tot_Bil***:

En el cas de les variables on hem identificat una forta asimetria cap a la dreta l'anàlisi dels valors extrems és més complex. Si bé es cert que les variables presenten valors extrems força elevats car tenir en compte que: 1. En molts casos aquests valors elevats no són valors aïllats sinó que segueixen la distribució de la variable. 2. La majoria de casos es donen en pacients classificats com a malalts de fetge, per tant, és possible que els anàlisis presentin valors disparats en aquestes variables.

Amb el coneixement que tenim del domini no podem descartar que es tracti de valors reals de casos extrems (no valors erronis). D'altra banda la presència de valors extrems, tot i representar casos reals, ens pot afectar els anàlisis que volem realitzar, atès que l'objectiu de la pràctica no és l'anàlisi dels casos atípics sinó l'estudi de la tendència general de les dades i la modelització predictiva.

Tenint en compte allò exposat anteriorment prendrem un enfoc relativament conservador i només tractarem aquells valors extrems més aïllats. Considerem que en el cas de pacients malalts i, tenint en compte la distribució fortament asimètrica de les variables, la imputació de valors de tendència central no seria adient. Podríem realitzar una imputació mitjançant un algoritme de k-veïns propers, tanmateix, aquests valors representen un petit percentatge del conjunt de dades i, a més, corresponen a la classe majoritària i, eliminar-los o imputar-los tindrà un efecte minso en els resultats. Decidim eliminar-los.

```
# Identifiquem els valors extrems
sum(liver_data$Aspartate > 2500)
```

```
## [1] 2
```

```
sum(liver_data$Tot_Bil > 40)
```

```
## [1] 2
```

```
sum(liver_data$Alkphos > 1000 & liver_data$Disease == "N")
```

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

Eliminem un total de 5 registres que correspon a un 0.86% de les dades.

```
# Eliminem els registres identificats
liver_data[liver_data$Tot_Bil > 40, ]$Tot_Bil <- NA
liver_data[liver_data$Aspartate > 2500, ]$Aspartate <- NA
liver_data[liver_data$Alkphos > 1000 & liver_data$Disease == "N", ]$Alkphos <- NA
liver_data <- liver_data[complete.cases(liver_data), ]
```

Finalment exportem el joc de dades netejat:

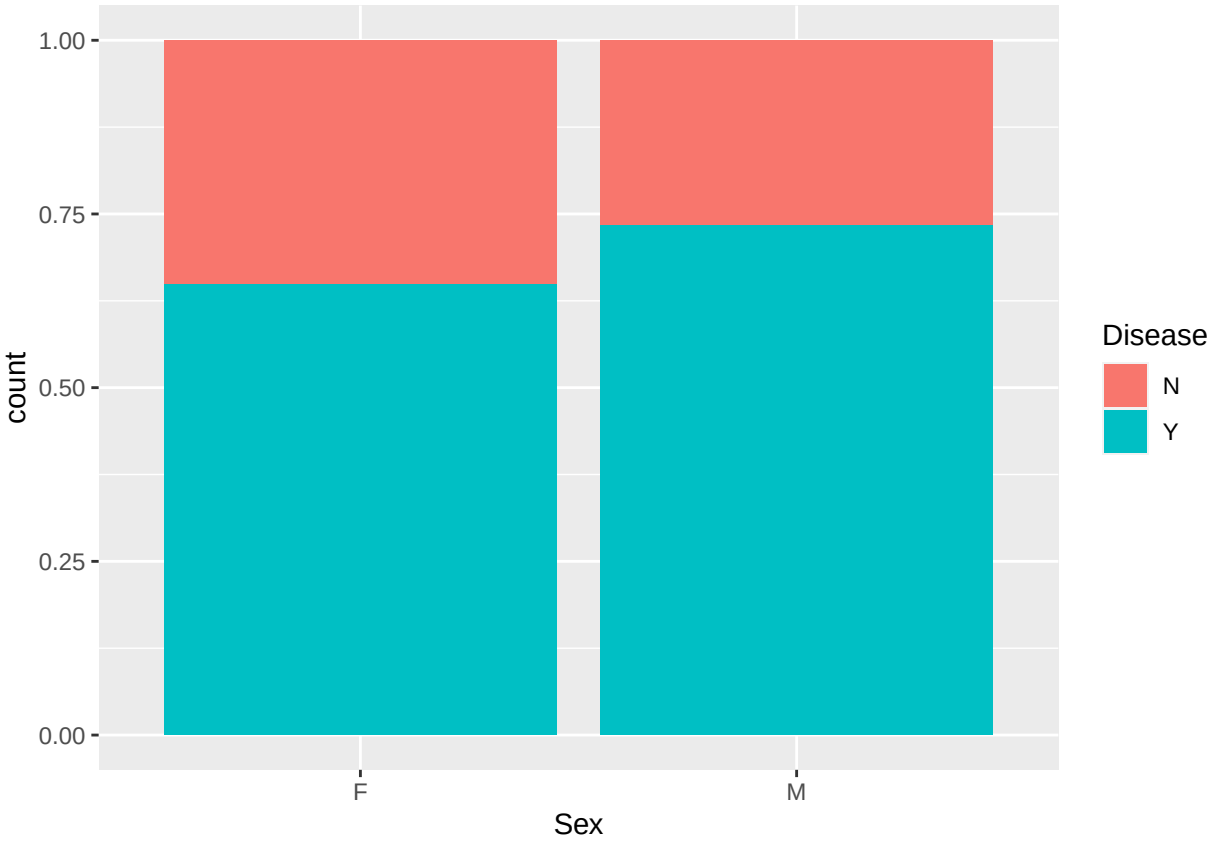
```
write.csv(liver_data, "processed_liver.csv", row.names = FALSE)
```

2.4 Anàlisi de les dades

2.4.1 Anàlisi de la diferència en la proporció de casos de malalts entre homes i dones.

Observem la proporció de casos de malalts segregat per sexe en el joc de dades

```
ggplot(liver_data, aes(x = Sex, fill = Disease)) + geom_bar(position = "fill")
```



Veiem que en la mostra que tenim la proporció de malalts és lleugerament superior en els homes. Volem saber, però, si aquesta diferències és estadísticament significativa per al conjunt de la població (en aquest cas amb població ens referim al conjunt d'anàlisis dels quals s'ha extret la mostra de què disposem, assumint atès que no tenim altra informació, que la mostra s'ha extret amb mètodes que n'assegurin la representativitat).

Per a contestar la pregunta de si existeixen diferències significatives en la proporció de malalts de fetge entre homes i dones a partir de la mostra que ens ofereix el joc de dades aplicarem un contrast d'hipòtesis de dues mostres independents sobre la proporció de malalts.

Per a fer-ho considerarem que a cada individu li correspon una variable de Bernoulli que val 1 si és malalt o 0 altrament. La suma de n les distribucions de Bernoulli de paràmetre p és una distribució binomial(n,p) que, en el cas de n gran, com és el cas que ens trobem, és aproximadament una distribució normal de mitjana np i variància $np(1-p)$.

Plantegem la següent hipòtesis per al contrast:

$$H_0 : p_1 - p_2 = 0$$
$$H_1 : p_1 - p_2 > 0$$

essent p_1 la proporció d'homes malalts i p_2 la proporció de dones malaltes. És, per tant, un test contrast d'hipòtesis unilateral de dues mostres independents sobre la diferència de proporcions.

Per a que les estimacions que farem siguin vàlides necessitem que la variable sobre la que fem la proporció provingui d'una distribució de Bernoulli i que la mida de les mostres sigui gran ($n > 30$), es compleixen tots dos requeriments per tant podem aplicar el test.

L'estadístic de contrast que utilitzarem pel test, sota la hipòtesis nul · la (que no hi ha diferència en les proporcions) serà:

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

on $\hat{p}$ és l'estimació de la proporció poblacional comuna calculada de la següent manera:

$$\hat{p} = \frac{n_1\hat{p_1} + n_2\hat{p_2}}{n_1 + n_2}$$

Procedim al càlcul del test:

```
# Comptem els casos de malalts segregats per sexe
m.patient <- sum(liver_data$Disease == "Y" & liver_data$Sex == "M")
f.patient <- sum(liver_data$Disease == "Y" & liver_data$Sex == "F")

# Definim els paràmetres del test
alpha <- 0.05
n1 <- nrow(liver_data[liver_data$Sex == "M", ])
n2 <- nrow(liver_data[liver_data$Sex == "F", ])
p1 <- m.patient/n1
p2 <- f.patient/n2
success <- c(p1 * n1, p2 * n2)
nn <- c(n1, n2)
prop.test(success, nn, conf.level = 1 - alpha, alternative = "greater", correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 3.875, df = 1, p-value = 0.0245
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.01150363 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.7339450 0.6478873
```

Obtenim un valor p inferior al nivell de significança del 0.05 i, per tant, podem rebutjar la hipòtesis nul · la i afirmar amb un nivell del confiança del 95% que la proporció de malalts de fetge és significativament superior en homes que en dones.

2.4.2 Analísi de la diferència d'edat entre individus sans i malalts.

Calculem la mitjana d'edat segregada per la variables Disease en la mostra de què disposem.

```
# Mitjana d'edat d'individus sans
mean(liver_data[liver_data$Disease == "N", "Age"])
```

```
## [1] 41.18675
```

```r
# Mitjana d'edat d'individus malalts
mean(liver_data[liver_data$Disease == "Y", "Age"])
```

```
## [1] 46.16748
```

Volem saber si la mitjana d'edat de la població de malalts es significativament superior tal com indica el joc de dades.

Procedim a seleccionar un test d'hipòtesis que ens permeti donar resposta, amb un cert nivell de confiança, a la pregunta.Per a realitzar el test considerarem les següents hipòtesis:

$$H_0 : \mu_1 - \mu_2 = 0$$

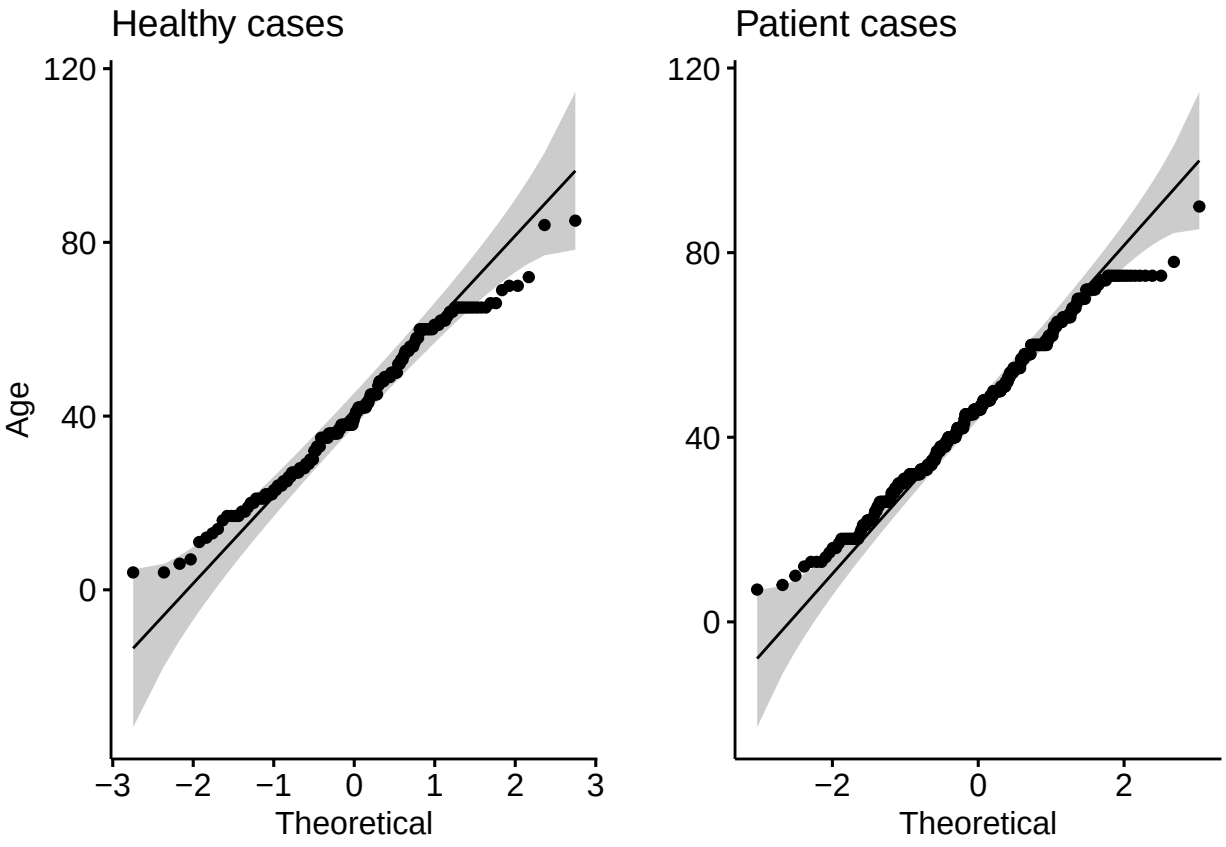
$$H_1 : \mu_1 - \mu_2 > 0$$

essent μ_1 la mitjana poblacional de l'edat en malalts i μ_2 la mitjana poblacional de l'edat en individus sans. És, per tant, un test d'hipòtesis unilateral sobre la diferència de mitjanes de dues poblacions independents.

Començarem revisant les assumpcions que ens permetran escollir l'estadístic de contrast més adient per a realitzar el test. En el cas d'un contrast d'hipòtesis sobre la mitjana de dues poblacions independents ens assumim que les dos poblacions provenen de distribucions normals i que existeix igualtat de variàncies (homocedasticitat).

```r
# Separem les dues mostres per fer el codi més net
agePatient <- liver_data[liver_data$Disease == "Y", "Age"]
ageHealthy <- liver_data[liver_data$Disease == "N", "Age"]

# Mostrem el gràfic Q-Q per ambdues mostres:
p1 <- ggqqplot(ageHealthy, title = "Healthy cases", ylab = "Age")
p2 <- ggqqplot(agePatient, title = "Patient cases", ylab = "")
grid.arrange(p1, p2, nrow = 1)
```



```r
# Realitzem el test de normalitat per les dues mostres
print("Test de normalitat per la distribució de la mostra d'individus sans")
```

```
## [1] "Test de normalitat per la distribució de la mostra d'individus sans"
```

```r
shapiro.test(ageHealthy)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ageHealthy
## W = 0.98276, p-value = 0.03711
```

```r
print("Test de normalitat per la distribució de la mostra d'individus malalts")
```

```
## [1] "Test de normalitat per la distribució de la mostra d'individus malalts"
```

```r
shapiro.test(agePatient)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  agePatient
## W = 0.99032, p-value = 0.008237
```

Veiem en els gràfics Q-Q com en ambdós casos la distribució de les poblacions s'aproxima a una normal si bé en els extrems els quartils s'allunyen de la distribució teòrica. Veiem també com en ambdós casos el test de Shapiro indica que rebutgem la hipòtesis nul · la de distribució normal.

Tanmateix, en virtut del Teorema del Límit Central sabem que per mostres de n gran (n > 30) independentment de la distribució de la variable original, la distribució de les mitjanes de les mostres obtingudes sí segueix una distribució normal, per tant, validem l'assumpció de normalitat per al test d'hipòtesis.

Passem a fer el test d'homocedasticitat per avaluar la igualtat de variàncies. Ho farem amb la funció var.test.

```r
# Apliquem el test sobre les mitjanes mostrals:
var.test(ageHealthy, agePatient)
```

```
##
##  F test to compare two variances
##
## data:  ageHealthy and agePatient
## F = 1.1856, num df = 165, denom df = 411, p-value = 0.1806
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9242191 1.5426527
## sample estimates:
## ratio of variances
##           1.185608
```

El test ens retorna un valor p de 0.18, superior al nivell de significació de 0.05 i, per tant, no rebutgem la hipòtesis nul · la i considerem que existeix homocedasticitat. Ens trobem, per tant en un cas de test d'hipòtesis de dues mostres diferents amb distribucions normal (distribució de les mitjanes) i variàncies desconegudes però iguals. L'estadístic de contrast a utilitzar és el següent:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

que segueix una distribució t d'Student amb $n_1 + n_2 - 2$ graus de llibertat i S és la desviació típica comuna.

Procedim al càlcul de test

```
# Definim els paràmetres inicials
alpha <- 0.05

# Realitzem el test amb la funció ja implementada en R
t.test(agePatient, ageHealthy, alternative = "greater", var.equal = TRUE, conf.level = 1 -
    alpha)
```

```
##
##  Two Sample t-test
##
## data:  agePatient and ageHealthy
## t = 3.3741, df = 576, p-value = 0.0003952
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.548743      Inf
## sample estimates:
## mean of x mean of y
##  46.16748  41.18675
```

```
# Netegem la memòria
rm(agePatient, ageHealthy)
```

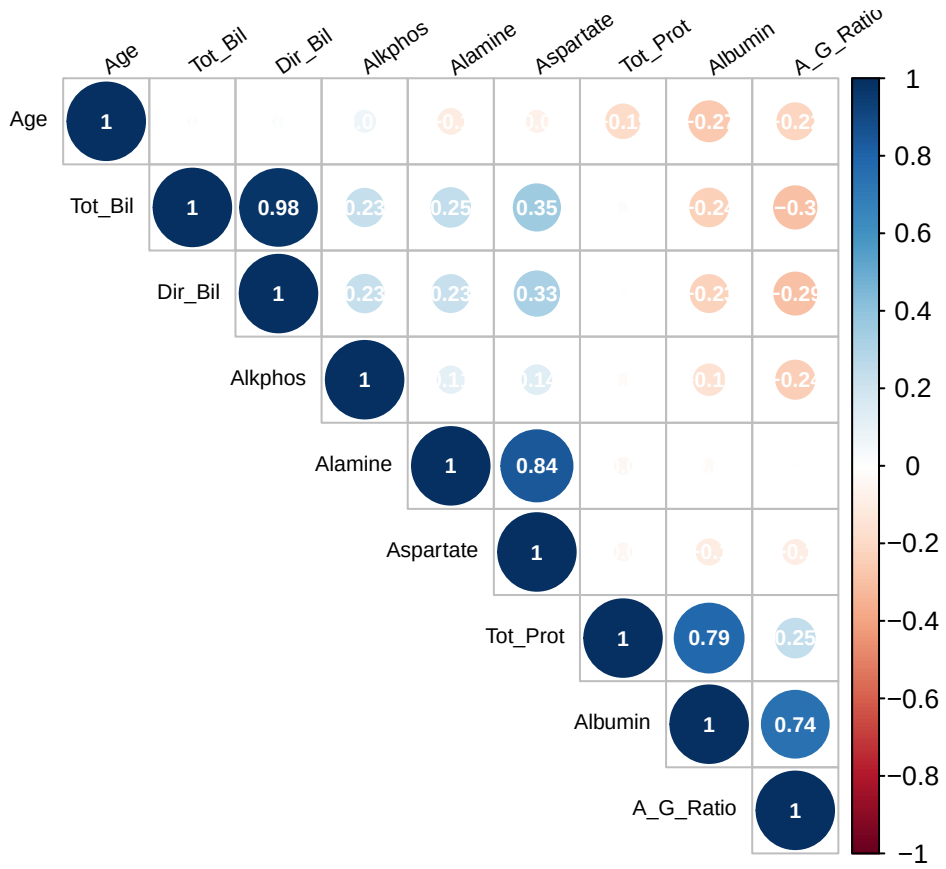
El p-valor resultant proper a 0 ens permet rebutjar la hipòtesis nul · la i afirmar amb una confiança del 95% que la mitjana d'edat per als malalts de fetge es significativament superior a la mitjana d'edat per als individus sans.

2.4.3 Anàlisi de la covariància entre variables

Per al desenvolupament d'aquesta pràctica hem plantejat la construcció de dos models predictius sobre la variable *Disease*, un basat en un regressió logística i l'altre basat en l'algoritme de *Random Forest*. Ni la regressió logística ni el *Random Forest* fan assumpcions sobre la normalitat ni homocedasticitat (si bé en el cas de la regressió logística sí s'assumeix una associació lineal entre les variables independents i el logit de la probabilitat d'ocurrència), però un requeriment bàsic per a la construcció de la majoria de models és que no existeixi colinearitat entre les variables explicatives. Per tal d'assegurar aquest requeriment i de seleccionar les variables més adients per a la construcció dels models realitzarem un anàlisis de la correlació de les variables de què disposem.

El primer que farem serà estudiar-ne la covariància mitjançant una matriu on enfrontem les variables dos a dos:

```
# Mostrem la matriu de correlació
corrplot(cor(liver_data[-c(2, 11, 12)]), method = "circle", type = "upper", addCoef.col = "white",
    number.cex = 0.7, tl.col = "black", tl.srt = 35, tl.cex = 0.7)
```



Si examinem la matriu de covariàncies veiem com tenim una alta covariància entre els atributs Tot_Bil i Dir_Bil, entre Alamine i Aspartate, i entre Tot_Prot i Albumin i Albumin i A_G_Ratio, fet que ens indica que tenim certa redundància en les dades i que possiblement podem seleccionar només alguns d'aquests atributs.

Si estudiem en deteniment aquestes correlacions veurem que són valors esperables: - Tot_Bil i Dir_Bil: són dos mesures del la concentració de bilirubina, per tant és lògic que tinguin una alta correlació. - Tot_Prot i Albumin: L'Albúmina representa un gran percentatge del total de proteïnes en sang, per tant, és lògic que tingui una bona correlació amb les proteïnes totals. - A_G_Ratio i Albumine: l'albúmina forma part del càlcul de l'A_G_Ratio i veiem que això n'afecta la correlació.

Tindrem en compte aquests resultat a l'hora de fer la selecció d'atributs per als models predictius.

Entre Tot_Bil i Dir_Bil atès que tenen una molt alta correlació i tots dos presenten la mateixa informació, la selecció de l'atribut és indiferent, escollirem Tot_Bil.

Entre Tot_Prot, Albumin i A_G_Ratio, si triem Albumin deixem de banda Tot_Prot i A_G_Ratio, per tant definirem fos grups de variables amb totes dues opcions i compararem els resultats.

2.4.4 Model de regressió logística

Procedim a desenvolupar un model de regressió logística que ens permeti analitzar l'efecte de les diferents variables de què disposem en la probabilitat que un individu tingui problemes hepàtics. Si bé la construcció del model ens permetrà també de predir el valor de probabilitat per a nous casos i, fixant un llindar de probabilitat, classificar aquests casos entre malalts i sans, no serà aquest l'objectiu del model, atès que desenvolupament el model predictiu mitjançant l'algoritme de *Random Forest*. És per això que construirem el model amb totes les dades disponibles i no dividirem el joc de dades en entrenament i prova.

Començarem construint un model de regressió logística amb les variables *Age* i *Sex*

```
# Construïm el model indicat
m1 <- glm(formula = Disease ~ Age + Sex, data = liver_data, family = binomial(link = logit))
```

Procedim a interpretar el model, comencem mostrant-ne el resum:

```r
# Mostrem el resum
summary(m1)
```

```
##
## Call:
## glm(formula = Disease ~ Age + Sex, family = binomial(link = logit),
##     data = liver_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8627  -1.3617   0.7373   0.8583   1.2023
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.190428   0.301674  -0.631  0.52789
## Age          0.018857   0.005844   3.227  0.00125 **
## SexM         0.373393   0.208559   1.790  0.07340 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 693.16  on 577  degrees of freedom
## Residual deviance: 678.73  on 575  degrees of freedom
## AIC: 684.73
##
## Number of Fisher Scoring iterations: 4
```

Analitzem els resultats i veiem que el test de Wald d'hipòtesi de nul · litat dels paràmetres ens indica que les variables *Age* i *Sex* contribueixen significativament a predir la probabilitat d'ocurrència de la variable dicotòmica *Disease*, si bé és cert que el nivell de significació de la variable SexM (la probabilitat base considera que l'individu sigui una dona), és a dir, el fet de ser home, influeix en la probabilitat de tenir malaltia amb una significació inferior a 0.1, i no 0.05.

Calculem els odd ratio de cada variable amb un interval d e confiança del 95%

```r
# Intervals de confiança per als odd ratios
exp(confint(m1))
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %   97.5 %
## (Intercept) 0.4571558 1.494805
## Age         1.0075242 1.030907
## SexM        0.9613449 2.180257
```

```r
exp(m1$coefficients)
```

```
## (Intercept)         Age        SexM
##   0.8266053   1.0190360   1.4526556
```

Observem els intervals de confiança i veiem que en el cas de la variable *Age* l'odds ratio és lleugerament superior a 1, això ens indicaria que l'edat és una variable de risc, tanmateix amb un valor tant proper a 1 d'increment de la probabilitat per augment de l'edat és força petit.

Quant a la variable *Sex* veiem que l'interval de confiança per a l'odds ratio és més ampli que en el cas anterior, ens ho indicava ja el resum del model. L'estimació per a l'odds ratio és de 1,45 el que ens indica que la probabilitat de patir malaltia de fetge augmenta 1,45 vegades pel fet de ser home, veiem però que l'interval de confiança avarca valors inferiors

a 1 per a l'odd ratio. Amb les dades de que disposem no s'evidencia clarament la influència del factor edat mitjançant aquest model.

Procedirem ara a incloure més variables i estudiar el seu efecte en el model. Construirem dos models amb la selecció d'atributs realitzada a partir de l'anàlisi de covariància.

m2: Age + Sex + Tot_Bil + Alkphos + Alamine + Albumin

m3: Age + Sex + Tot_Bil + Alkphos + Aspartate + Tot_Prot + A_G_Ratio

```
# Construïm els dos models
m2 <- glm(formula = Disease ~ Age + Sex + Tot_Bil + Alkphos + Alamine + Albumin,
    data = liver_data, family = binomial(link = logit))
m3 <- glm(formula = Disease ~ Age + Sex + Tot_Bil + Alkphos + Aspartate + Tot_Prot +
    A_G_Ratio, data = liver_data, family = binomial(link = logit))
```

Mostrem el resum d'ambdós models:

```
# Mostrem el resum
summary(m2)
```

```
##
## Call:
## glm(formula = Disease ~ Age + Sex + Tot_Bil + Alkphos + Alamine +
##     Albumin, family = binomial(link = logit), data = liver_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2446  -1.0999   0.4006   0.9107   1.4627
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.3110320  0.6872146  -1.908 0.056424 .
## Age          0.0193408  0.0064334   3.006 0.002644 **
## SexM         0.0229867  0.2303286   0.100 0.920504
## Tot_Bil      0.3229638  0.1034080   3.123 0.001789 **
## Alkphos      0.0024878  0.0009916   2.509 0.012116 *
## Alamine      0.0128907  0.0038594   3.340 0.000838 ***
## Albumin     -0.1180023  0.1365491  -0.864 0.387492
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 693.16  on 577  degrees of freedom
## Residual deviance: 571.81  on 571  degrees of freedom
## AIC: 585.81
##
## Number of Fisher Scoring iterations: 7
```

```
summary(m3)
```

```
##
## Call:
## glm(formula = Disease ~ Age + Sex + Tot_Bil + Alkphos + Aspartate +
##     Tot_Prot + A_G_Ratio, family = binomial(link = logit), data = liver_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -3.2177  -1.0830   0.4143   0.9131   1.5666
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.6792576  0.8625016  -1.947  0.05154 .
## Age          0.0193090  0.0064215   3.007  0.00264 **
## SexM         0.0760750  0.2304935   0.330  0.74136
## Tot_Bil      0.3003661  0.1036387   2.898  0.00375 **
## Alkphos      0.0025954  0.0009976   2.602  0.00928 **
## Aspartate    0.0076416  0.0025684   2.975  0.00293 **
## Tot_Prot     0.0888299  0.0987934   0.899  0.36857
## A_G_Ratio   -0.4972291  0.3867483  -1.286  0.19856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 693.16  on 577  degrees of freedom
## Residual deviance: 574.49  on 570  degrees of freedom
## AIC: 590.49
##
## Number of Fisher Scoring iterations: 7
```

Veiem com en els dos casos se'ns indica que la variable *Sex* no és significativa, tanmateix en estudis relatius a paràmetres biològics el gènere és potencialment una variable de confusió. Tenint en compte que hem vist que hi havia diferències significatives en la proporció de malalts en funció del gènere, mantindrem la variable en el model independentment del nivell de significació per a controlar-ne l'efecte.

A l'm2 la variable Albumin tampoc és significativa i en m3 les variables Tot_Prot i A_G_Ratio tampoc són significatives. Tornem a justar els models eliminant aquestes variables i mostrem novament el resum:

```
# Construïm els dos models
m2 <- glm(formula = Disease ~ Age + Sex + Tot_Bil + Alkphos + Alamine, data = liver_data,
    family = binomial(link = logit))
m3 <- glm(formula = Disease ~ Age + Sex + Tot_Bil + Alkphos + Aspartate, data = liver_data,
    family = binomial(link = logit))

# Mostrem el resum
summary(m2)
```

```
##
## Call:
## glm(formula = Disease ~ Age + Sex + Tot_Bil + Alkphos + Alamine,
##     family = binomial(link = logit), data = liver_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2350  -1.0905   0.3963   0.9073   1.4163
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.7909864  0.4084749  -4.385 1.16e-05 ***
## Age          0.0207165  0.0062367   3.322 0.000895 ***
## SexM         0.0317086  0.2299476   0.138 0.890324
## Tot_Bil      0.3380267  0.1024043   3.301 0.000964 ***
## Alkphos      0.0025559  0.0009889   2.585 0.009745 **
## Alamine      0.0127104  0.0038320   3.317 0.000910 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 693.16  on 577  degrees of freedom
## Residual deviance: 572.56  on 572  degrees of freedom
## AIC: 584.56
##
## Number of Fisher Scoring iterations: 7
```

```
summary(m3)
```

```
##
## Call:
## glm(formula = Disease ~ Age + Sex + Tot_Bil + Alkphos + Aspartate,
##     family = binomial(link = logit), data = liver_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2334  -1.1147   0.4236   0.9192   1.4009
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.6545847  0.4026381  -4.109 3.97e-05 ***
## Age          0.0195168  0.0062208   3.137  0.00170 **
## SexM         0.0400428  0.2283422   0.175  0.86079
## Tot_Bil      0.3155570  0.1039761   3.035  0.00241 **
## Alkphos      0.0028989  0.0009984   2.904  0.00369 **
## Aspartate    0.0074188  0.0025494   2.910  0.00361 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 693.16  on 577  degrees of freedom
## Residual deviance: 576.49  on 572  degrees of freedom
## AIC: 588.49
##
## Number of Fisher Scoring iterations: 7
```

Veiem ara com el test de Wald d'hipòtesi de nul · litat dels paràmetres ens indica que totes les variables són significatives amb uns p-valors en tots els casos inferior al nivell de significació de 0.01.

D'entre els dos models desenvolupats en quedarem amb el primer, m2, atès que presenta una desviació residual lleugerament inferior a més d'un valor pel Criteri d'Informació d'Akaike també inferior, el que ens indicaria un millor ajust del model.

Procedim ara a comprovar l'assumpció de linealitat entre les variables independent contínues i el logit de la probabilitat d'ocurrència de la variable objectiu. Per a fer-ho utilitzarem el mètode de Box-Tidell. Inclourem doncs una variable d'interacció entre cada variable contínua i el logaritme natural d'aquella variable contínua i comprovarem si aquesta variable d'interacció és significativa per al model. Si és significativa indicarà que la variable no té una relació lineal amb el logit.

```
# Comprobació d'assumpció de linealitat mitjançant el mètode de Box-Tidewell:
m2.BTtest <- glm(formula = Disease ~ Age + Sex + Age:log(Age) + Tot_Bil + Tot_Bil:log(Tot_Bil) +
    Alkphos + Alkphos:log(Alkphos) + Alamine + Alamine:log(Alamine), data = liver_data,
    family = binomial(link = logit))
summary(m2.BTtest)
```

```
##
```

```
## Call:
## glm(formula = Disease ~ Age + Sex + Age:log(Age) + Tot_Bil +
##     Tot_Bil:log(Tot_Bil) + Alkphos + Alkphos:log(Alkphos) + Alamine +
##     Alamine:log(Alamine), family = binomial(link = logit), data = liver_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2240  -1.0269   0.3943   0.9090   1.5749
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.6224407  1.3917829  -2.603  0.00925 **
## Age                   0.1988964  0.1168807   1.702  0.08881 .
## SexM                  0.0253295  0.2357441   0.107  0.91444
## Tot_Bil               0.3752814  0.3780987   0.993  0.32093
## Alkphos               0.0067430  0.0163838   0.412  0.68066
## Alamine               0.0267459  0.0340974   0.784  0.43281
## Age:log(Age)         -0.0377457  0.0246966  -1.528  0.12642
## Tot_Bil:log(Tot_Bil) -0.0273193  0.1718840  -0.159  0.87372
## Alkphos:log(Alkphos) -0.0005657  0.0023671  -0.239  0.81113
## Alamine:log(Alamine) -0.0027007  0.0062973  -0.429  0.66802
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 693.16  on 577  degrees of freedom
## Residual deviance: 569.93  on 568  degrees of freedom
## AIC: 589.93
##
## Number of Fisher Scoring iterations: 9
```

Comprovem que cap dels termes d'interacció introduïts és significatiu, per tant, validem la hipòtesis de linealitat entre les variables independents i el logit i procedim a interpretar els odds ratios del model.

Comencem mostrant les estimacions per als odd ratios de cada variable:

```
##                   2.5 %    97.5 %
## (Intercept) 0.07326162 0.3644874
## Age         1.00864739 1.0336543
## SexM        0.65504959 1.6155039
## Tot_Bil     1.18064486 1.7589424
## Alkphos     1.00082015 1.0046789
## Alamine     1.00599459 1.0211875
```

```
## (Intercept)         Age        SexM     Tot_Bil     Alkphos     Alamine
##   0.1667956   1.0209325   1.0322167   1.4021779   1.0025592   1.0127915
```

A partir de l'estimació dels odds ratio de les variables estudiades veiem com la variable més rellevant és el nivell de bilirrubina (Tot_Bil) amb un efecte d'augment de la probabilitat de tenir patologia de fetge d'entre 1.2 i 1.76 vegades superior per a cada unitat que s'incrementa. La resta de variables significatives totes presenten intervals de confiança per a l'odds ratio superiors a 1, el que indica que són factors de risc, tanmateix, el seu valor és molt proper a 1 i, per tant, l'efecte en l'augment de la probabilitat d'ocurrència és minso, especialment en el cas d'*Alkphos*.

Podem concloure a partir de l'estudi del model logístic de regressió que els nivells de bilirrubina és el factor que exerceix un major efecte sobre la possibilitat de patir patologies de fetge i, és un factor clau a controlar en el diagnòstic d'un pacient.

De la resta de variables de què disposem i en base a la mostra estudiada em obtingut uns odds ratio molt propers a la unitat, fet que indica que el seu efecte sobre la variable objectiu és minso, tot i ser factors de risc. Podem concloure que el conjunt de variables de que disposem es poc rellevant quan a la inferència de patologies de fetge fet que limita la capacitat de diagnòstic.

2.4.5 Model predictiu mitjançant Random Forest

Random Forests és un classificador combinat i està basat en arbres de decisió com a classificadors base. En aquest cas s'utilitza un mostreig tant dels elements del conjunt original d'entrenament com de les seves variables.

L'algoritme consisteix en generar versions diferents del conjunt d'entrenament utilitzant mostreig per reemplaçament, mètode conegut també com bagging. Durant el procés de construcció de cada arbre de decisió es selecciona aleatòriament un subconjunt de les variables del conjunt de dades, donant opcions a variables que normalment quedarien eclipsades per altres que tinguessin major rellevància.

Anteriorment s'ha observat que algunes variables tenen una alta colinearitat entre elles, per tant es procedirà a entrenar el model eliminant algunes variables que presenten colinearitat del conjunt original.

Com s'ha comentat anteriorment, l'objectiu de la implementació d'aquest model és realitzar una predicció del valor de la variable Disease, per tant en aquest cas si que entrenarem el model amb el 70% del conjunt de les dades i després testarem el model amb el 30% restant, d'aquesta manera obtindrem un accuracy de la capacitat de predicció del model.

```
# Partició de les dades amb train i test amb una proporció 70/30
set.seed(1234)
train <- sample(nrow(liver_data), 0.7 * nrow(liver_data), replace = FALSE)
TrainSet <- liver_data[train, ]
TestSet <- liver_data[-train, ]
```

Com que hem comprovat anteriorment que les dades es troben poc equilibrades, anem a mirar la proporció d'observacions segons la variable Disease que, en aquest cas, és la que volem predir.

```
prop.table(table(liver_data$Disease))
```

```
##
##         N         Y
## 0.2871972 0.7128028
```

```
prop.table(table(TrainSet$Disease))
```

```
##
##         N         Y
## 0.2871287 0.7128713
```

```
prop.table(table(TestSet$Disease))
```

```
##
##         N         Y
## 0.2873563 0.7126437
```

Es pot observar que tant en el conjunt original, com en el d'entrenament i test la proporció és bastant igual. Factoritzem, ara, la variable sex per a que el model la pugui incloure amb les dades d'entrenament i la de Disease perquè és la que volem predir.

```
TrainSet$Disease <- factor(TrainSet$Disease)
TestSet$Disease <- factor(TestSet$Disease)
TrainSet$Sex <- factor(TrainSet$Sex)
TestSet$Sex <- factor(TestSet$Sex)
```

Generalment, quants més arbres utilitzi el model més robust serà el classificador i per tant s'obtindrà més precisió. No obstant, s'ha de tenir en compte que el cost computacional augmenta en conseqüència i també es pot donar el cas de sobre-entrenar el conjunt de dades, per tant no estaríem obtenint una bona precisió.

Primerament s'entrena un model amb els paràmetres per defecte de la funció i amb només algunes variables.

```r
# Create a Random Forest model with default parameters
set.seed(1234)
model1 <- randomForest(Disease ~ Age + Sex + Tot_Bil + Alkphos + Aspartate + Tot_Prot +
    A_G_Ratio, data = TrainSet, importance = TRUE)
model1
```

```
##
## Call:
##  randomForest(formula = Disease ~ Age + Sex + Tot_Bil + Alkphos +      Aspartate + Tot_Prot + A_G_Ratio, d
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 27.23%
## Confusion matrix:
##    N   Y class.error
## N 38  78   0.6724138
## Y 32 256   0.1111111
```

```r
# Create a Random Forest model with default parameters
set.seed(1234)
model2 <- randomForest(Disease ~ Age + Sex + Tot_Bil + Alkphos + Alamine + Albumin,
    data = TrainSet, importance = TRUE)
model2
```

```
##
## Call:
##  randomForest(formula = Disease ~ Age + Sex + Tot_Bil + Alkphos +      Alamine + Albumin, data = TrainSet,
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 30.69%
## Confusion matrix:
##    N   Y class.error
## N 39  77   0.6637931
## Y 47 241   0.1631944
```

El primer model ens dóna una millor precisió. S'observa que el model ha utilitzat un total de 500 arbres de decisió i el nombre de variables provades en cada divisió és 3 en aquest cas. La taxa d'error és del 26.43%, el que és un valor bastant alt. Anem a mirar quina és la precisió del model predint els valors que s'han reservat per a testar.

```r
library(caret)
```

```
## Loading required package: lattice
```

```r
predictions <- predict(model1, TestSet)
confusionMatrix(predictions, TestSet$Disease, positive = "Y")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   N   Y
##          N  15  14
##          Y  35 110
##
```

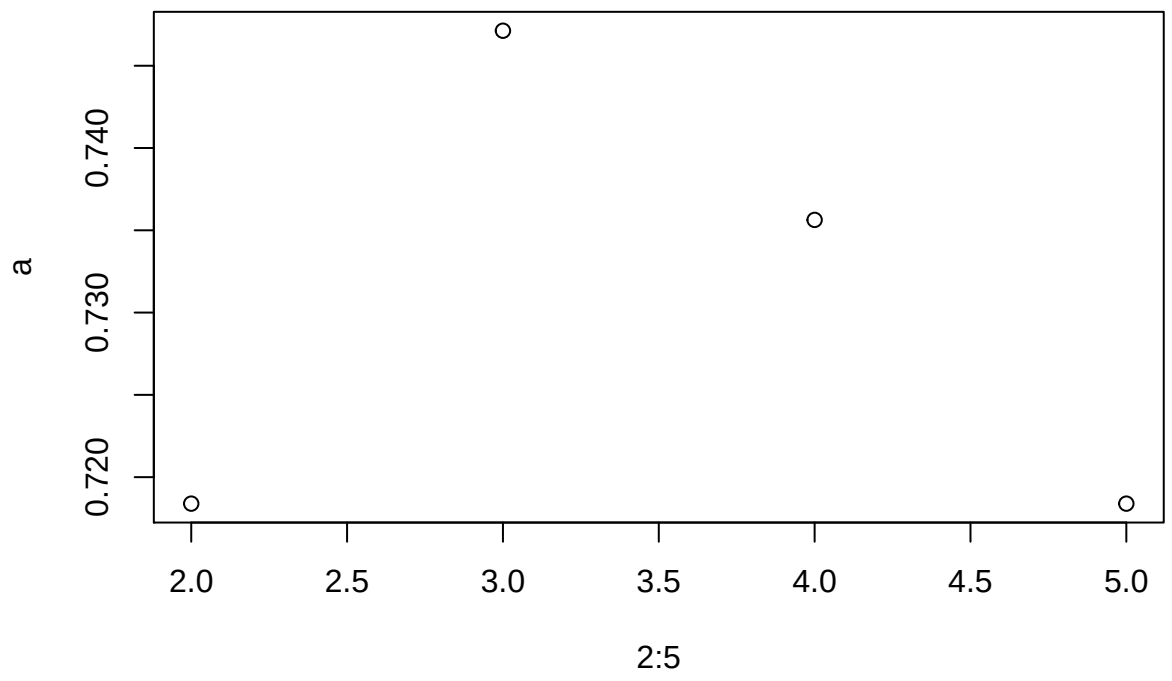
```
##                Accuracy : 0.7184
##                  95% CI : (0.6453, 0.7838)
##     No Information Rate : 0.7126
##     P-Value [Acc > NIR] : 0.471348
##
##                   Kappa : 0.2139
##
##  Mcnemar's Test P-Value : 0.004275
##
##             Sensitivity : 0.8871
##             Specificity : 0.3000
##          Pos Pred Value : 0.7586
##          Neg Pred Value : 0.5172
##              Prevalence : 0.7126
##          Detection Rate : 0.6322
##    Detection Prevalence : 0.8333
##       Balanced Accuracy : 0.5935
##
##        'Positive' Class : Y
##
```

S'obté una predicció bastant alta, del 70%. Es torna a entrenar el model modificant alguns paràmetres com el nombre d'arbres i el nombre de variables provades en cadascun dels nodes de partició. Es defineix un nombre d'arbres igual a 600 i es mirarà la millor predicció provant diferents valors en quant al nombre de variables.

```
a = c()
i = 4
for (i in 2:6) {
    set.seed(1234)
    model3 <- randomForest(Disease ~ Age + Sex + Tot_Bil + Alkphos + Alamine + Albumin,
        data = TrainSet, ntree = 500, mtry = i, importance = TRUE)
    predictions <- predict(model3, TestSet)
    a[i - 2] = mean(predictions == TestSet$Disease)
}
a
```

```
## [1] 0.7183908 0.7471264 0.7356322 0.7183908
```

```
plot(2:5, a)
```



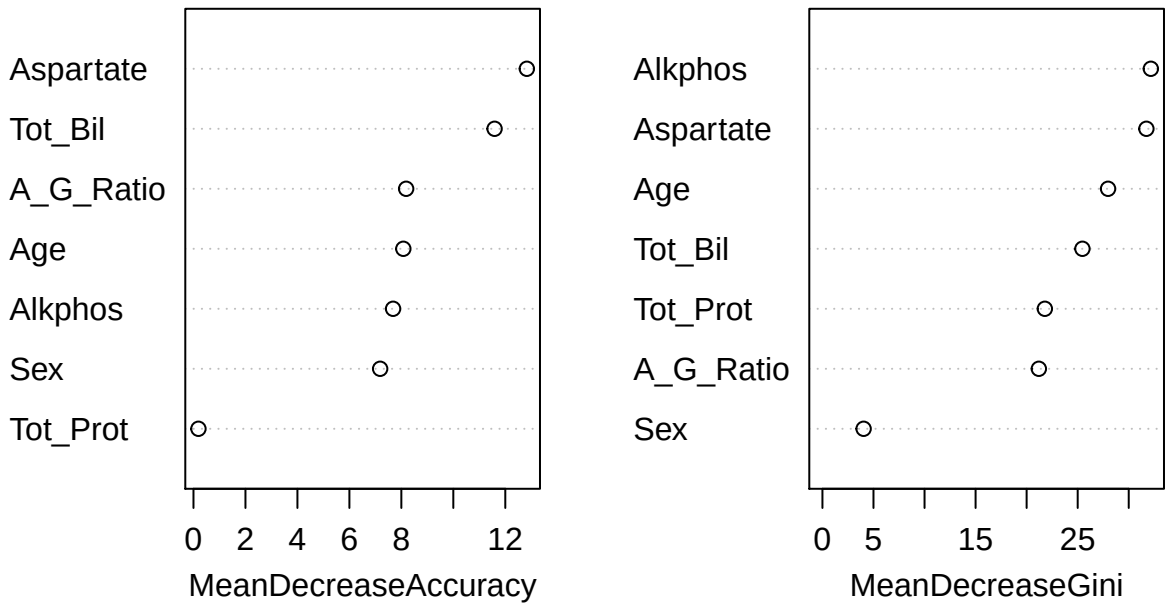
S'observa que la precisió del model, fixant un nombre d'arbres de decisió igual a 600, va variant en funció del nombre de variables en cada node. La millor precisió s'obté amb un nombre de variables igual a 3, després s'observa una fluctuació tendint a la baixa a mesura que van augmentant el nombre de variables.

Les següents funcions mostren la caiguda de la precisió mitjana en cadascuna de les variables. Això permet veure la importància relativa de cada variable, estimant l'error comès pel Random Forest quan s'altera alguna d'aquestes variables, permutant aleatòriament els seus valors en el conjunt de test. Aquest error es mesura per a cadascun dels arbres classificadors, fent el promig de l'error comès en tots ells per a cadascuna de les variables. El percentatge d'error es compara a l'estimat amb el conjunt de test sense aquesta permutació aleatòria, de manera que és possible mesurar l'impacte d'aquesta variable, ja que si l'error augmenta quan es permuta una variable voldrà dir que la variable és rellevant per al problema que s'està resolent.

```
importance(model1)
```

```
##                    N          Y MeanDecreaseAccuracy MeanDecreaseGini
## Age        6.8404929  5.0424011            8.0757924        27.974322
## Sex        0.7779884  8.3868896            7.1837936         4.031996
## Tot_Bil   16.7490960  1.7526901           11.5876300        25.462955
## Alkphos   13.1913881 -0.6211510            7.6810663        32.165376
## Aspartate 11.9031449  6.8486836           12.8296218        31.729359
## Tot_Prot   1.4712436 -0.8470926            0.1903271        21.782766
## A_G_Ratio  7.2841623  4.9510167            8.1837582        21.195704
```

```
varImpPlot(model1)
```



En aquest cas s'observa com la variable Tot_Prot és bastant irrellevant per a l'entrenament del model, ja que la predicció tan sols disminuiria un 0.5 en el cas de que s'eliminés aquesta variable. Per contra, en el cas de la variable Aspartate, l'eliminació provocaria una classificació errònia adicional d'aproximadament 16 observacions en promig. El cas de la MeanDecreaseGini el que mostra és el desordre, és a dir, la variància o puresa dels nodes. Els valors més baixos indiquen que el node conté majoritàriament observacions d'una sola classe. Per tant, es pot dir que en el cas del Sex seria una variable la qual la majoria d'observacions corresponen a una de les dues opcions de la variable Disease i, per contra, en el cas de Alkphos els seus valors poden pertànyer en una variabilitat bastant alta a qualsevol de les dues classes.

2.5 Representació dels resultats

La representació dels resultats s'ha fet a cada apartat per atès que s'ha considerat que és més aclaridor i millora el discurs de la pràctica.

2.6 Resolució del problema

Les conclusions de cada anàlisi han estat exposades a cada apartat corresponent, en fem ara un resum a mode de conclusió final.

- Existeix una diferencia estadísticament significativa en la proporció de malalts de fetge en funció del sexe? Hem realitzat un test d'hipòtesis sobre la diferència de proporcions entre la mostra de homes i dones i em pogut concloure amb un nivell de confiança del 95% que la proporció de malalts en homes és significativament superior a les dones. Si bé això no indica que sigui realment la distinció de gènere la que genera la diferència, sí que és un factor a tenir en compte per al diagnòstic.
- Existeix una diferència estadísticament significativa en la mitjana d'edat entre els malalts i els individus sans? Hem realitzat un test d'hipòtesis sobre la diferència de mitjanes d'edat entre la mostra d'individus sans i malalts i hem pogut concloure amb un nivell de confiança del 95% que la mitjana d'edat dels malalts és significativament superior a la dels individus sans el que indica certa rellevància del factor edat en el diagnòstic de patologies de fetge. Aquesta relació s'ha estudiat amb més deteniment durant la interpretació del model de regressió logística.

- Quina és la influència dels diferents factors (atributs de què disposem) en la probabilitat de desenvolupar una patologia de fetge? Podem concloure a partir de l'estudi del model logístic de regressió que els nivells de bilirrubina és el factor que exerceix un major efecte sobre la possibilitat de patir patologies de fetge i, és un factor clau a controlar en el diagnòstic d'un pacient.

De la resta de variables de què disposem i en base a la mostra estudiada em obtingut uns odds ratio molt propers a la unitat, fet que indica que el seu efecte sobre la variable objectiu és minso, tot i ser factors de risc. Podem concloure que el conjunt de variables de que disposem es poc rellevant quan a la inferència de patologies de fetge fet que limita la capacitat de diagnòstic.

- Quina capacitat tenim de predir si un individu pateix o no una malaltia de fetge a partir dels atributs que disposem? Per a predir l'existència de patologia de fetge hem construït un model de classificació a partir de l'algorisme *Random Forest* i n'hem estudiat el rendiment sobre un conjunt de dades de prova no utilitzar per a l'entrenament del model.

Els resultats del model corroboren que la capacitat per a predir la patologia de fetge a partir de les dades disponibles és força limitada. Ajustant els paràmetres del model em aconseguit una precisió (Accuracy) del 71,84% vers el 71,26% que ens ofereix si seleccionem directament sempre l'individu com a malalt (No Information Rate). Comprovem que, tot i tenir una sensibilitat força alta (~88%) la sensibilitat del model és molt baixa (~30%) el que indica un taxa alta del falsos positius.

Podem destacar dos factors principals que juguen un paper important en els resultats obtinguts a la predicció, d'una banda el joc de dades tractat és força desequilibrat amb molt més nombre de malalts i d'altra banda, com hem vist a partir del model logístic, les variables de què disposem no són especialment rellevants per a l'objectiu.

En resum, al llarg d'aquesta pràctica hem realitzar diferents anàlisis sobre el joc de dades seleccionat que ens han permès d'extreure'n coneixement i donar resposta les preguntes plantejades. Com a continuació del treball considerem que podria ser interessant l'aplicació de tècniques específiques per contrarestat el desequilibri present en les dades mitjançant un submostreig dels casos de malalts i una sobre-mostreig per repetició aleatòria dels casos d'individus sans amb l'objectiu d'equiparar ambdós classes i verificar-ne l'efecte sobre el model.

---

3 Bibliografia

Greenwell, B. (2019, October 19). Hands-On Machine Learning with R. Retrieved December 4, 2019, from https://bradleyboehmke.github.io/HOML/.

I. H. Witten, Data mining: practical machine learning tools and techniques / Ian H. Witten, Eibe Frank, Mark A. Hall. Burlington, MA : Morgan Kaufmann Publishers/Elsevier, 2011.

J. Gironés Roig, Data mining / Jordi Gironès Roig ; [el encargo y la creación de este material docente han sido coordinados por los profesores: Jordi Conesa Caralt, David Masip Rodó]. Barcelona : Universitat Oberta de Catalunya, 2013.

J. Gironès Roig, Minería de datos : modelos y algoritmos / Jordi Gironés Roig [i 3 més]. Barcelona : Editorial UOC, 2017.

J. Long i P. Teetor, R Cookbook, 2nd Edition [recurs electrònic]; J. Long i P. Teetor, 2019. https://rc2e.com

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168755/

https://rpubs.com/bpoulin-CUNY/338004

https://www.kaggle.com/uciml/indian-liver-patient-records

http://www.statisticalassociates.com/logistic10.htm

https://event-mohs.gov.mm/wp-content/uploads/2019/12/Logistics-regression.pdf

https://www.statology.org/assumptions-of-logistic-regression/

Cómo implementar bosques aleatorios en R | R-bloggers. (n.d.). Retrieved January 5, 2021, from https://www.r-bloggers.com/2018/01/how-to-implement-random-forests-in-r/

Indian Liver Patient Records | Kaggle. (n.d.). Retrieved January 5, 2021, from https://www.kaggle.com/uciml/indian-liver-patient-records

Lorena96/PRAC2_Neteja-i-Analisi-de-Dades. (n.d.). Retrieved January 5, 2021, from https://github.com/Lorena96/PRAC2_Neteja-i-Analisi-de-Dades

RPubs - Árboles de decisión y métodos de ensemble. (n.d.). Retrieved January 5, 2021, from https://rpubs.com/Cristina_Gil/arboles_ensemble Tune Machine Learning Algorithms in R (random forest case study). (n.d.). Retrieved January 5, 2021, from https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/