

PRA 2: Neteja, validació i anàlisi de les dades

Álvaro Díaz i David Leiva

13/12/2020

Contents

Detalls de l'activitat	1
Presentació	1
Competències	1
Objectius	1
Resolució	2
Descripció del dataset	2
Importància i objectius de l'anàlisi	4
Neteja de les dades	4
Preparació del dataset	4
Descripció del dataset	7
Valors nuls	8
Correcció valors nuls de variables categòriques	10
Correcció valors nuls de variables quantitatives	10
Anàlisi de les dades	11
Comprobació de la normalitat	29
Proves estadístiques	31
Comparació entre grups de la classe	31
Correlació entre les variables seleccionades	32
Creació de noves variables	33
Model amb regressió logística	35

Detalls de l'activitat

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

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.

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 contextos més amplis o multidisciplinaris.
 - 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 autodirigida o autònoma.
 - Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.
-

Resolució

Descripció del dataset

Es realitzarà l'anàlisi exploratori (EDA) del dataset <https://archive.ics.uci.edu/ml/datasets/HCC+Survival> on es recopila informació de pacients amb carcinoma hepatocel·lular (HCC) i la seva supervivència a l'any.

L'HCC és el tumor hepàtic més comú en els pacients amb hepatopatia crònica. La supervivència d'aquests pacients no només depen de l'estadi tumoral si no que també depen de l'estat funcional del fetge.

El conjunt de dades HCC es va obtenir en l'Hospital Universitari de Coïmbra (Portugal) i contenia diversos factors demogràfics, de risc, de laboratori i de supervivència global de 165 pacients reals diagnosticats de HCC. El conjunt de dades conté 49 funcions seleccionades segons les directrius de pràctica clínica EASL-EORTC (Associació Europea per a l'Estudi del Fetge - Organització Europea per a la Recerca i el Tractament del Càncer), que són els habituals en la gestió de HCC.

Es tracta d'un conjunt de dades heterogeni, amb 23 variables quantitatives i 26 variables qualitatives. La variable objectiu és la supervivència a 1 any i es va codificar com a variable binària: 0 (mor) i 1 (viu).

Les variables del dataset són:

- **Gender (Gen):** [1=Home;0=Dona] Sexe del pacient
- **Symptoms (Sym):**[1=Si;0=No] Sintomàtic
- **Alcohol (Alc):** [1=Si;0=No] Hepatopatia alcohòlica
- **Hepatitis B Surface Antigen (HBS):** [1=Si;0=No] Antigen de superfície de l'hepatitis B present a la sang
- **Hepatitis B e Antigen (HBe):**[1=Si;0=No] Antigen e de l'hepatitis B present a la sang
- **Hepatitis B Core Antibody (HBC):** [1=Si;0=No] Anticòs per l'hepatitis B present a la sang
- **Hepatitis C Virus Antibody (HCV):** [1=Si;0=No] Anticòs per l'hepatitis C present a la sang
- **Cirrhosis (Cir):** [1=Si;0=No] Estadio avançat d'hepatopatia crònica

- **Endemic Countries (End):** [1=Si;0=No] Pacient procedent de països amb alta prevalença d'hepatitis vírica
- **Smoking (Smo):** [1=Si;0=No] Fumador
- **Diabetes (Dia):** [1=Si;0=No] Diabètic
- **Obesity (Obe):** [1=Si;0=No] Obesitat
- **Hemochromatosis (Hem):**[1=Si;0=No] Hemocromatosi
- **Arterial Hypertension (HyA):** [1=Si;0=No] Hipertensió arterial
- **Chronic Renal Insufficiency(CRI):** [1=Si;0=No] Insuficiència renal
- **Human Immunodeficiency Virus (HIV):** [1=Si;0=No] Infecció per HIV
- **Nonalcoholic Steatohepatitis (Ste):** [1=Si;0=No] Esteatosis hepàtica de origen no alcohòlic
- **Esophageal Varices (Eso):** [1=Si;0=No] Presència de varius esofàgics com indicador d'hipertensió portal
- **Splenomegaly (Spl):** [1=Si;0=No] Augment del tamany de la melsa com indicador d'hipertensió portal
- **Portal Hypertension (PHT):** [1=Si;0=No] Pacient amb hipertensió arterial coneguda
- **Portal Vein Thrombosis (PVT):** [1=Si;0=No] Presència de trombosi venosa portal
- **Liver Metastasis (Met):** [1=Si;0=No] Metàstasi hepàtica
- **Radiological Hallmark (Rad):** [1=Si;0=No] Comportament radiològic típic per HCC
- **Age at diagnosis (Age):** Anys d'edat al moment del diagnòstic de HCC
- **Grams of Alcohol per day (gAl):** Grams d'alcohol ingerit de mitjana al dia
- **Packs of cigarets per year (PCi):** Número de paquets de cigarrets consumits per any
- **Performance Status (PSt):** [0=Actiu;1=Restringit;2=Assistència ocasional;3=Assistència parcial;4=Assistència total;5=Mort] Escala de l'estat general del pacient oncològic
- **Encephalopathy degree (Enc):** [1=Cap;2=Grau I/II; 3=Grau III/IV] Grau d'afectació mental de l'hepatopatia
- **Ascites degree (Asc):** [1=Cap;2=Lleu;3=Moderada a Severa] Grau d'ascitis com a indicador indirecte d'hipertensió portal
- **International Normalised Ratio (INR):** Temps de protrombina
- **Alpha-Fetoprotein (ng/mL) (AFe):** Nivells del marcador tumoral a la sang
- **Haemoglobin (g/dL) (Hae):** Nivells de Hemoglobina a la sang
- **Mean Corpuscular Volume (MCV):** Volum corpuscular mig dels eritrocits
- **Leukocytes(G/L) (Leu):** Concentració de cèl·lules blanques en sang
- **Platelets (Pla):** Concentració de plaquetes en sang
- **Albumin (mg/dL) (Alb):** Nivells d'albumina en sang
- **Total Bilirubin(mg/dL) (BiT):** Nivells de Bilirrubina Total en sang
- **Alanine transaminase (U/L) (ALT):** Nivells d'ALT en sang
- **Aspartate transaminase (U/L) (AST):** Nivells d'ASP en sang
- **Gamma glutamyl transferase (U/L) (GGT):** Nivells gamma-GT en sang

- **Alkaline phosphatase (U/L) (ALP)**: Nivel·ls de fosfatasa alcalina en sang
- **Total Proteins (g/dL) (Pro)**: Concentració total de proteïnes en sang
- **Creatinine (mg/dL) (Crea)**: Concentració de creatinina en sang
- **Number of Nodules (Nod)**: Número de nòduls d'HCC visualitzats
- **Major dimension of nodule (cm) (DiN)**: Tamany major dels nòduls d'HCC
- **Direct Bilirubin (mg/dL) (BiD)**: Nivells de Bilirrubina Directe en sang
- **Iron (Iro)**: Concentració de ferro en sang
- **Oxygen Saturation (%) (OxS)**: Saturació d'oxigen de la sang
- **Ferritin (ng/mL) (Fer)**: Nivells de ferritina en sang
- **Class Attribute (Class)**: [0=Mort; 1=Viu] Supervivent a l'any del diagnòstic d'HCC

Importància i objectius de l'anàlisi

Amb les dades recopilades al dataset podem intentar saber quines variables estan més relacionades amb la supervivència a l'any. Podem conèixer el grau de correlació entre les variables independents per finalment definir un model de regressió logística per tal d'intentar predir la mortalitat a l'any amb les variables seleccionades.

Poder conèixer la probabilitat de supervivència d'un pacient amb diagnòstic recent d'HCC podrà fer que s'adaptin millors les opcions de tractament, sent més agressius en pacients amb alta probilitat de sobreviure, i en canvi, optant per teràpies pal·liatives o de confort per pacients amb pitjor pronòstic.

Neteja de les dades

Preparació del dataset

```
# Importació del dataset
download.file("https://archive.ics.uci.edu/ml/machine-learning-databases/00423/hcc-survival.zip", "mydata.zip")
unzip("mydata.zip")
hcc<-read.csv ("hcc-survival/hcc-data.txt",header = F, dec = ".", stringsAsFactors = F, na.strings = "?")

# Noms de les variables
hcc_header<-c("Gender", "Symptoms", "Alcohol", "Hepatitis B Surface Antigen",
              "Hepatitis B e Antigen", "Hepatitis B Core Antibody",
              "Hepatitis C Virus Antibody", "Cirrhosis",
              "Endemic Countries", "Smoking", "Diabetes", "Obesity",
              "Hemochromatosis", "Arterial Hypertension",
              "Chronic Renal Insufficiency", "Human Immunodeficiency Virus",
              "Nonalcoholic Steatohepatitis", "Esophageal Varices",
              "Splenomegaly", "Portal Hypertension",
              "Portal Vein Thrombosis", "Liver Metastasis",
              "Radiological Hallmark", "Age at diagnosis",
              "Grams of Alcohol per day", "Packs of cigarets per year",
              "Performance Status", "Encephalopathy degree", "Ascites degree",
              "International Normalised Ratio", "Alpha-Fetoprotein (ng/mL)",
              "Haemoglobin (g/dL)", "Mean Corpuscular Volume",
              "Leukocytes(G/L)", "Platelets", "Albumin (mg/dL)",
              "Total Bilirubin(mg/dL)", "Alanine transaminase (U/L)")
```

```

, "Aspartate transaminase (U/L)",
"Gamma glutamyl transferase (U/L)",
"Alkaline phosphatase (U/L)", "Total Proteins (g/dL)",
"Creatinine (mg/dL)", "Number of Nodules",
"Major dimension of nodule (cm)",
"Direct Bilirubin (mg/dL)", "Iron", "Oxygen Saturation (%)",
"Ferritin (ng/mL)", "Class Attribute" )

colnames(hcc)<-hcc_header

tipVar <- c()
for (i in 1:ncol(hcc)) tipVar <- c(tipVar,is(hcc[,i])[1])
tipVar <- table(tipVar)

```

Al nostre dataset, tenim 32 variables de tipus integer i 18 variables de tipus numeric.

```
summary(hcc)
```

```
##      Gender      Symptoms      Alcohol
## Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
## 1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
## Median :1.0000   Median :1.0000   Median :1.0000
## Mean   :0.8061   Mean   :0.6395   Mean   :0.7394
## 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
## Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
## NA's   :18
## Hepatitis B Surface Antigen Hepatitis B e Antigen
## Min.   :0.0000           Min.   :0.00000
## 1st Qu.:0.0000           1st Qu.:0.00000
## Median :0.0000           Median :0.00000
## Mean   :0.1081           Mean   :0.00794
## 3rd Qu.:0.0000           3rd Qu.:0.00000
## Max.   :1.0000           Max.   :1.00000
## NA's   :17           NA's   :39
## Hepatitis B Core Antibody Hepatitis C Virus Antibody   Cirrhosis
## Min.   :0.0000           Min.   :0.0000           Min.   :0.000
## 1st Qu.:0.0000           1st Qu.:0.0000           1st Qu.:1.000
## Median :0.0000           Median :0.0000           Median :1.000
## Mean   :0.2695           Mean   :0.2179           Mean   :0.903
## 3rd Qu.:1.0000           3rd Qu.:0.0000           3rd Qu.:1.000
## Max.   :1.0000           Max.   :1.0000           Max.   :1.000
## NA's   :24           NA's   :9
## Endemic Countries      Smoking      Diabetes      Obesity
## Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
## 1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
## Median :0.00000   Median :1.0000   Median :0.0000   Median :0.000
## Mean   :0.07937   Mean   :0.5081   Mean   :0.3457   Mean   :0.129
## 3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000
## Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000
## NA's   :39       NA's   :41       NA's   :3       NA's   :10
## Hemochromatosis Arterial Hypertension Chronic Renal Insufficiency
## Min.   :0.0000   Min.   :0.0000           Min.   :0.0000
## 1st Qu.:0.0000   1st Qu.:0.0000           1st Qu.:0.0000
## Median :0.0000   Median :0.0000           Median :0.0000

```

```

## Mean      :0.0493      Mean      :0.3642      Mean      :0.1227
## 3rd Qu.   :0.0000      3rd Qu.   :1.0000      3rd Qu.   :0.0000
## Max.      :1.0000      Max.      :1.0000      Max.      :1.0000
## NA's      :23          NA's      :3          NA's      :2
## Human Immunodeficiency Virus Nonalcoholic Steatohepatitis
## Min.      :0.00000      Min.      :0.00000
## 1st Qu.   :0.00000      1st Qu.   :0.00000
## Median    :0.00000      Median    :0.00000
## Mean      :0.01987      Mean      :0.05594
## 3rd Qu.   :0.00000      3rd Qu.   :0.00000
## Max.      :1.00000      Max.      :1.00000
## NA's      :14          NA's      :22
## Esophageal Varices Splenomegaly Portal Hypertension
## Min.      :0.0000      Min.      :0.00      Min.      :0.0000
## 1st Qu.   :0.0000      1st Qu.   :0.00      1st Qu.   :0.0000
## Median    :1.0000      Median    :1.00      Median    :1.0000
## Mean      :0.6106      Mean      :0.56      Mean      :0.7143
## 3rd Qu.   :1.0000      3rd Qu.   :1.00      3rd Qu.   :1.0000
## Max.      :1.0000      Max.      :1.00      Max.      :1.0000
## NA's      :52          NA's      :15      NA's      :11
## Portal Vein Thrombosis Liver Metastasis Radiological Hallmark
## Min.      :0.0000      Min.      :0.0000      Min.      :0.000
## 1st Qu.   :0.0000      1st Qu.   :0.0000      1st Qu.   :0.000
## Median    :0.0000      Median    :0.0000      Median    :1.000
## Mean      :0.2222      Mean      :0.2236      Mean      :0.681
## 3rd Qu.   :0.0000      3rd Qu.   :0.0000      3rd Qu.   :1.000
## Max.      :1.0000      Max.      :1.0000      Max.      :1.000
## NA's      :3          NA's      :4          NA's      :2
## Age at diagnosis Grams of Alcohol per day Packs of cigarets per year
## Min.      :20.00      Min.      : 0.00      Min.      : 0.00
## 1st Qu.   :57.00      1st Qu.   : 0.00      1st Qu.   : 0.00
## Median    :66.00      Median    : 75.00      Median    : 0.00
## Mean      :64.69      Mean      : 71.01      Mean      : 20.46
## 3rd Qu.   :74.00      3rd Qu.   :100.00      3rd Qu.   : 30.50
## Max.      :93.00      Max.      :500.00      Max.      :510.00
## NA's      :          NA's      :48      NA's      :53
## Performance Status Encephalopathy degree Ascites degree
## Min.      :0.000      Min.      :1.000      Min.      :1.000
## 1st Qu.   :0.000      1st Qu.   :1.000      1st Qu.   :1.000
## Median    :1.000      Median    :1.000      Median    :1.000
## Mean      :1.018      Mean      :1.159      Mean      :1.442
## 3rd Qu.   :2.000      3rd Qu.   :1.000      3rd Qu.   :2.000
## Max.      :4.000      Max.      :3.000      Max.      :3.000
## NA's      :          NA's      :1          NA's      :2
## International Normalised Ratio Alpha-Fetoprotein (ng/mL)
## Min.      :0.840      Min.      :1.20e+00
## 1st Qu.   :1.170      1st Qu.   :5.20e+00
## Median    :1.300      Median    :3.30e+01
## Mean      :1.422      Mean      :1.93e+04
## 3rd Qu.   :1.530      3rd Qu.   :6.15e+02
## Max.      :4.820      Max.      :1.81e+06
## NA's      :4          NA's      :8
## Haemoglobin (g/dL) Mean Corpuscular Volume Leukocytes(G/L)
## Min.      : 5.00      Min.      : 69.50      Min.      : 2.20

```

## 1st Qu.:11.43	1st Qu.: 89.78	1st Qu.: 5.10
## Median :13.05	Median : 94.95	Median : 7.20
## Mean :12.88	Mean : 95.12	Mean : 1473.96
## 3rd Qu.:14.60	3rd Qu.:100.67	3rd Qu.: 19.52
## Max. :18.70	Max. :119.60	Max. :13000.00
## NA's :3	NA's :3	NA's :3
## Platelets	Albumin (mg/dL)	Total Bilirubin(mg/dL)
## Min. : 1.7	Min. :1.900	Min. : 0.300
## 1st Qu.: 255.8	1st Qu.:3.000	1st Qu.: 0.800
## Median : 93000.0	Median :3.400	Median : 1.400
## Mean :113206.4	Mean :3.446	Mean : 3.088
## 3rd Qu.:171500.0	3rd Qu.:4.050	3rd Qu.: 2.925
## Max. :459000.0	Max. :4.900	Max. :40.500
## NA's :3	NA's :6	NA's :5
## Alanine transaminase (U/L)	Aspartate transaminase (U/L)	
## Min. : 11.00	Min. : 17.00	
## 1st Qu.: 31.00	1st Qu.: 46.25	
## Median : 50.00	Median : 71.00	
## Mean : 67.09	Mean : 96.38	
## 3rd Qu.: 78.00	3rd Qu.:110.25	
## Max. :420.00	Max. :553.00	
## NA's :4	NA's :3	
## Gamma glutamyl transferase (U/L)	Alkaline phosphatase (U/L)	
## Min. : 23.00	Min. : 1.28	
## 1st Qu.: 91.25	1st Qu.:108.25	
## Median : 179.50	Median :162.00	
## Mean : 268.03	Mean :212.21	
## 3rd Qu.: 345.25	3rd Qu.:261.50	
## Max. :1575.00	Max. :980.00	
## NA's :3	NA's :3	
## Total Proteins (g/dL)	Creatinine (mg/dL)	Number of Nodules
## Min. : 3.900	Min. :0.200	Min. :0.000
## 1st Qu.: 6.300	1st Qu.:0.700	1st Qu.:1.000
## Median : 7.050	Median :0.850	Median :2.000
## Mean : 8.961	Mean :1.127	Mean :2.736
## 3rd Qu.: 7.575	3rd Qu.:1.100	3rd Qu.:5.000
## Max. :102.000	Max. :7.600	Max. :5.000
## NA's :11	NA's :7	NA's :2
## Major dimension of nodule (cm)	Direct Bilirubin (mg/dL)	Iron
## Min. : 1.500	Min. : 0.10	Min. : 0.0
## 1st Qu.: 3.000	1st Qu.: 0.37	1st Qu.: 40.5
## Median : 5.000	Median : 0.70	Median : 83.0
## Mean : 6.851	Mean : 1.93	Mean : 85.6
## 3rd Qu.: 9.000	3rd Qu.: 1.40	3rd Qu.:118.0
## Max. :22.000	Max. :29.30	Max. :224.0
## NA's :20	NA's :44	NA's :79
## Oxygen Saturation (%)	Ferritin (ng/mL)	Class Attribute
## Min. : 0.00	Min. : 0	Min. :0.0000
## 1st Qu.: 16.00	1st Qu.: 84	1st Qu.:0.0000
## Median : 27.00	Median : 295	Median :1.0000
## Mean : 37.03	Mean : 439	Mean :0.6182
## 3rd Qu.: 56.00	3rd Qu.: 706	3rd Qu.:1.0000
## Max. :126.00	Max. :2230	Max. :1.0000
## NA's :80	NA's :80	

Descripció del dataset

El dataset està compost de 165 observacions de 49 atributs de pacients amb una variable de classe que registra la supervivència a l'any del diagnòstic. Dels atributs dels pacients, existeixen 26 categòrics, 3 dels quals són ordinals (*Performance Status*, *Encephalopathy degree*, *Ascites degree*), sent la resta numèrics. És pot veure que existeixen valors nuls codificats com *NA*.

```
# Definició de tipus de dades per columna
hcc_factor <-c(1:23,50)
hcc_order <- c(27:29)
hcc_factorT<-c(1:23,27:29,50)
hcc_num<-c(24:26,30:49)

# Factorització de les columnes categòriques
hcc <- hcc %>% mutate_at(vars(c(1:23,50)), as.factor)
hcc <- hcc %>% mutate_at(vars(c(27:29)), as.factor)
```

Valors nuls

La distribució de valors desconeguts per pacient és:

```
table(apply(hcc, 1, function(x) sum(is.na(x))))

##
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 16 18 22 23
##  8 16 15 28 19 21 15  7 11 12  4  2  1  1  1  1  1  1  1
```

El percentatge de NAs per variable és:

```
funNA <- function(a, n){
  a = round(100*a/n,1)
}

totNA <- hcc %>% select(everything()) %>% #
  summarise_all(funs(sum(is.na(.))))
perNA <- totNA %>% mutate_all(funNA, n= nrow(hcc))
tauNA <- totNA %>% bind_rows(perNA)
tauNA <- as_tibble(t(tauNA), rownames = "Variable") %>%
  rename(`total NA` = V1, `NA` = V2) %>%
  arrange(-`total NA`)

tau <- cbind(tauNA[1:25,],tauNA[26:50,])

kable(x = tau, format = "latex", caption = "Variables amb NA", booktabs = TRUE) %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))
```


Table 1: Variables amb NA

Variable	total NA	%NA	Variable	total NA	%NA
Oxygen Saturation (%)	80	48.5	Total Bilirubin(mg/dL)	5	3.0
Ferritin (ng/mL)	80	48.5	Liver Metastasis	4	2.4
Iron	79	47.9	International Normalised Ratio	4	2.4
Packs of cigarets per year	53	32.1	Alanine transaminase (U/L)	4	2.4
Esophageal Varices	52	31.5	Diabetes	3	1.8
Grams of Alcohol per day	48	29.1	Arterial Hypertension	3	1.8
Direct Bilirubin (mg/dL)	44	26.7	Portal Vein Thrombosis	3	1.8
Smoking	41	24.8	Haemoglobin (g/dL)	3	1.8
Hepatitis B e Antigen	39	23.6	Mean Corpuscular Volume	3	1.8
Endemic Countries	39	23.6	Leukocytes(G/L)	3	1.8
Hepatitis B Core Antibody	24	14.5	Platelets	3	1.8
Hemochromatosis	23	13.9	Aspartate transaminase (U/L)	3	1.8
Nonalcoholic Steatohepatitis	22	13.3	Gamma glutamyl transferase (U/L)	3	1.8
Major dimension of nodule (cm)	20	12.1	Alkaline phosphatase (U/L)	3	1.8
Symptoms	18	10.9	Chronic Renal Insufficiency	2	1.2
Hepatitis B Surface Antigen	17	10.3	Radiological Hallmark	2	1.2
Splenomegaly	15	9.1	Ascites degree	2	1.2
Human Immunodeficiency Virus	14	8.5	Number of Nodules	2	1.2
Portal Hypertension	11	6.7	Encephalopathy degree	1	0.6
Total Proteins (g/dL)	11	6.7	Gender	0	0.0
Obesity	10	6.1	Alcohol	0	0.0
Hepatitis C Virus Antibody	9	5.5	Cirrhosis	0	0.0
Alpha-Fetoprotein (ng/mL)	8	4.8	Age at diagnosis	0	0.0
Creatinine (mg/dL)	7	4.2	Performance Status	0	0.0
Albumin (mg/dL)	6	3.6	Class Attribute	0	0.0

Només hi ha 8 pacients amb les dades completes, faltant a la majoria de pacients entre 2 i 9 dades. Fins i tot hi ha pacients 13 pacients amb més de 10 dades desconegudes.

Estudiant la distribució dels valors desconeguts per variable veiem que només hi ha 6 variables amb totes les dades íntegres. Amb més de l'10% de dades desconegudes hi ha 16 de les 50 variables (un 32%), destacant 9 variables amb entre el 20 i el 50% de les seves dades desconegudes, com són la saturació d'oxigen o els nivells de ferritina en sang. Els valors NA poden seguir una distribució a l'atzar de manera que la proporció dels esperats en cada classe hauria de ser similar. En cas contrari, la correcció dels valors desconeguts podria provocar un biaix cap a un dels dos grups. Vegem com es distribueixen en les variables els valors missing i si hi ha diferències significatives depenent de la classe.

```
# Valoració de la distribució dels NA a les variables amb respecte la variable classe
cero <- hcc %>% filter(`Class Attribute` == 0)
uno <- hcc %>% filter(`Class Attribute` == 1)

probTest <- tibble()
for (i in names(hcc[,hcc_factorT])) {
  if (sum(is.na(hcc[,i]))>0){
    casos<-c(sum(is.na(cero[,i])),sum(is.na(uno[,i])))
    long<-c(length(cero[,i]),length(uno[,i]))
    test<-prop.test(x=casos,n=long)
    probTest <- probTest %>%
      bind_rows(c(Class = i, p_value = test$p.value,
                  prob_0=(casos/long)[1], prob_1=(casos/long)[2],
```

```

    numNA_0=casos[1], numNA_1=casos[2]))
  }
}

testSig <- probTest %>% filter(p_value <= 0.05) %>%
  mutate_at(.vars = c("p_value", "prob_0", "prob_1", "numNA_0", "numNA_1"), as.numeric)

kable(x = testSig, format = "latex",
      caption = "Variables amb una difència significativa de NA entre els grups de la classe",
      booktabs = TRUE, digits = 4) %>%
  kable_styling(latex_options = c("HOLD_position"))

```

Table 2: Variables amb una difència significativa de NA entre els grups de la classe

Clase	p_value	prob_0	prob_1	numNA_0	numNA_1
Symptoms	0.0058	0.0159	0.1667	1	17
Hemochromatosis	0.0019	0.2540	0.0686	16	7
Esophageal Varices	0.0084	0.4444	0.2353	28	24

A la variable **Symptoms** s'observen molts NA entre els pacients que sobreviuen. Els pacients que no sobreviuen solen presentar molta simptomatologia i aquesta es registra. En canvi, els pacients que no presenten símptomes, poden no registrar-se com a negatiu a aquesta variable, existint un biaix d'informació.

Igualment succeix a les variables **Hemochromatosis** i **Esophageal Varices**. Els pacients afectats es registren i probablement presenten tases més altes de mortalitat. En canvi, molts pacients es desconeixerà si presenten hemocromatosis o varius, però probablement no la patiran, i tindran tases de supervivència superiors.

Corregir aquest valors desconeguts cap a la moda condicionarà un biaix. Per tant, per corregir els valors NA es farà:

- A totes les variables qualitatives s'assignarà el valor més pròxim utilitzant l'algoritme kNN
- A les variables quantitatives s'assignarà la mitjana de la variable. Per tal de no tenir una mitjana condicionada per valors erronis extrems, es corregiran abans de l'assignació del valor mitjà als valors desconeguts.

Correcció valors nuls de variables categòriques

S'assigna el valor més pròxim d'entre k=5 als valor NA de les variables categòriques.

```

#computem tots els valors NA de variables factor
factorNames<- colnames(hcc[hcc_factorT,])

# Computar per kNN, abm valors standards, k = 5
hcc <- kNN(hcc, variable = factorNames) %>% subset(select = Gender:`Class Attribute`)

```

Correcció valors nuls de variables quantitatives

Existeixen dues variables amb valors estranys, incompatibles amb la vida; son **Leukocytes** i **Platelets** La gran majoria dels valors a la variable **Leukocytes** estan per sota de 100, que és l'esperat. Valors majors son pràcticament impossibles. Els valors d'aquesta variable es solen expressar sobre mm3 pel que solen tenir

valors múltiples de 1000, d'aquí la probable confusió amb els valors extrems trobats. Es corregiran modificant les unitats d'aquest valors.

Amb respecte `Platelets`, l'error és similar al trobat a l'anterior variable.

Es corregeix els errors i s'assigna la mitjana als valors desconeguts

```
# Correcció valors leucocitosi i plaquetes + assignació mediana als valors NA de variables numeriques
hcc <- hcc %>%
  mutate(`Leukocytes(G/L)` = ifelse(`Leukocytes(G/L)` > 100, `Leukocytes(G/L)`/100, `Leukocytes(G/L)`),
         Platelets = ifelse(Platelets < 1000, Platelets*1000, Platelets))

for (i in hcc_num){
  hcc[,i] <- ifelse(is.na(hcc[,i]), median(hcc[,i], na.rm = T), hcc[,i])
}

#parlem de mitjana o de mediana? median es mediana, mitjana es mean.
```

Anàlisi de les dades

Per fer l'anàlisi de les dades numèriques, farem 2 plots per a cada variable. El primer correspon a un anàlisi dels valors numèrics i el segon es el boxplot. Sempre separant la variable estudiada en dos classes, les que moren i les que sobreviuen.

```
# Estudi distribució variables Numeriques
ind_CA <- which(colnames(hcc) == "Class Attribute")

ncols <- length(hcc_num)
if (ncols%%2 == 1){
  last_col = ncols
} else{
  last_col = ncols + 1
}

#for (i in hcc_num) {
for (i in 1:ncols) {
  if(i%%2 == 1){
    if ( i != last_col){
      data <- hcc[,c(hcc_num[i], hcc_num[i+1], ind_CA )]
    } else{
      data <- hcc[,c(hcc_num[i], ind_CA )]
    }
    name_var <- names(data)

    a1<-data %>%
      ggplot(aes(x=data[,1], fill=`Class Attribute`))+
      geom_histogram() +
      labs(fill="Clase", y = "Frequencia", x =name_var[1] ) +
      theme(legend.position = "bottom")

    a2<-data %>%
      ggplot(aes(x=`Class Attribute`, y=data[,1])) +
      geom_boxplot() +
      labs(x = "Clase", y = name_var[1])
  }
}
```

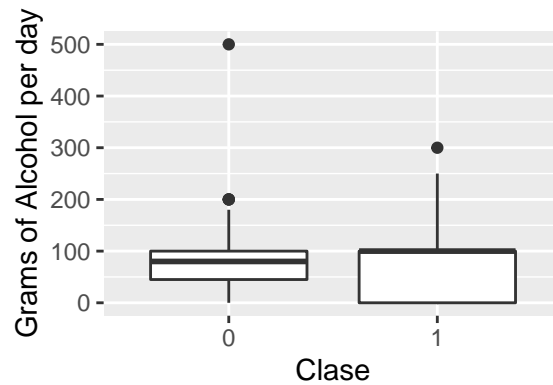
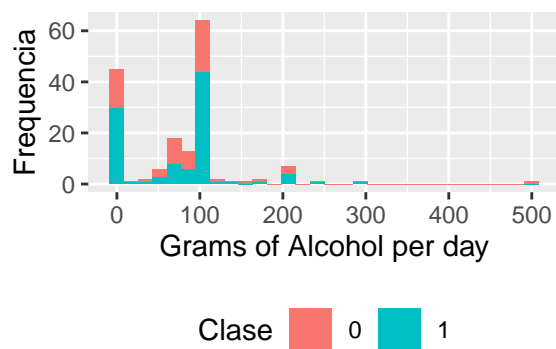
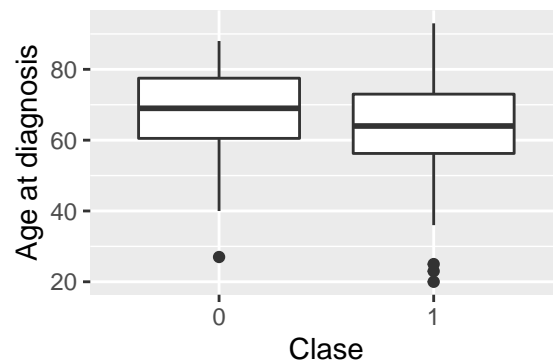
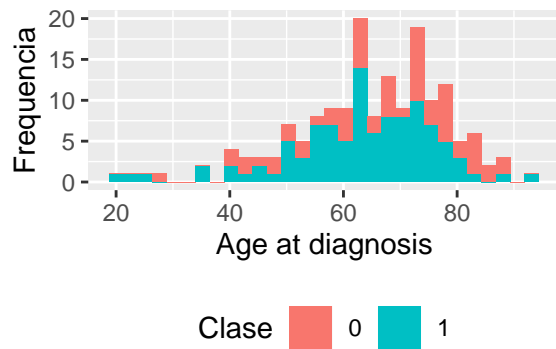
```

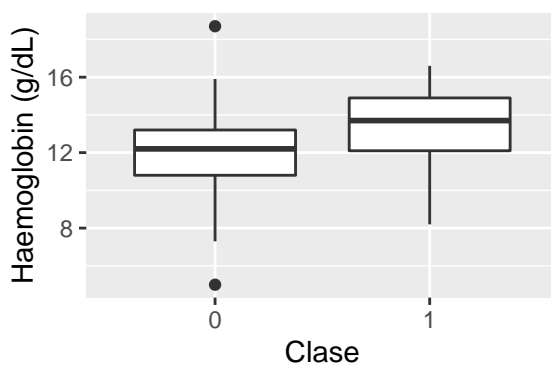
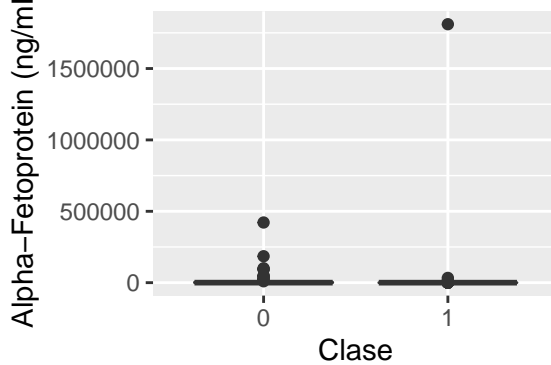
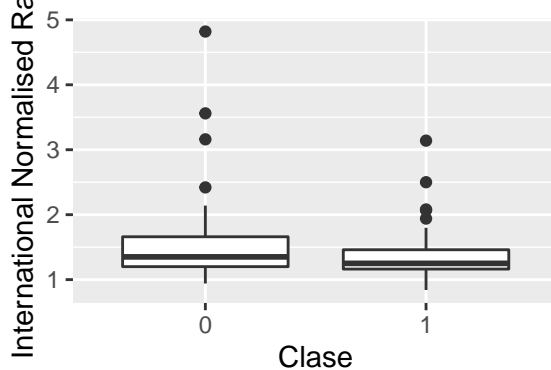
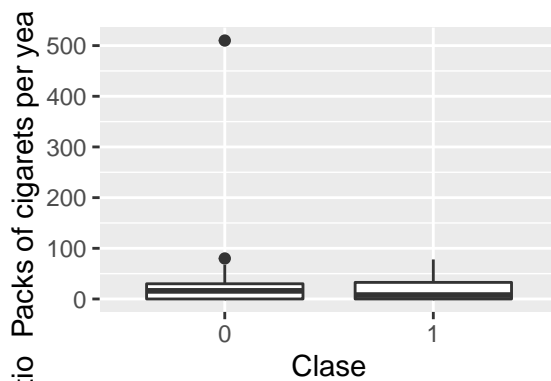
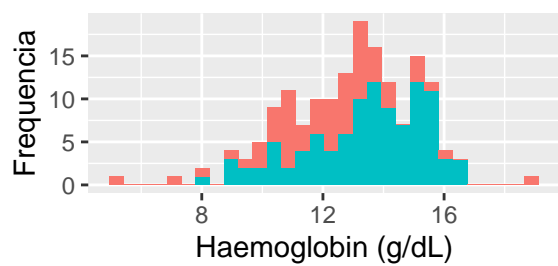
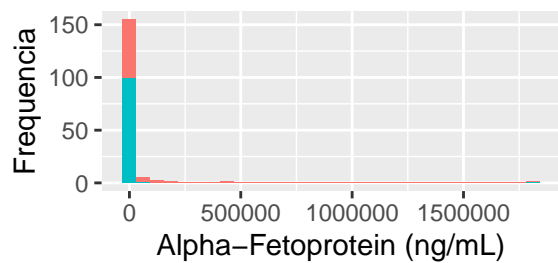
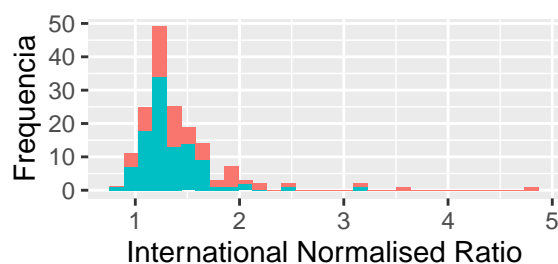
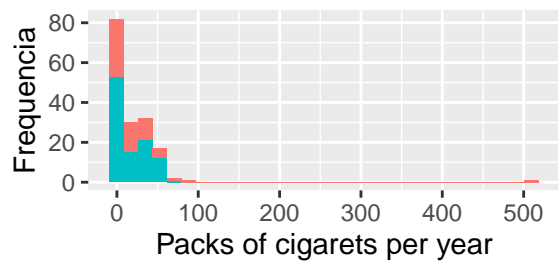
if ( i != last_col){
  a3<-data %>%
    ggplot(aes(x=data[,2], fill=`Class Attribute`))+
    geom_histogram() +
    labs(fill="Clase", y = "Frecuencia", x =name_var[2] ) +
    theme(legend.position = "bottom")

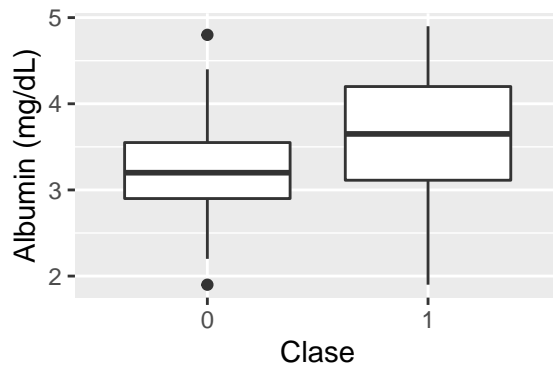
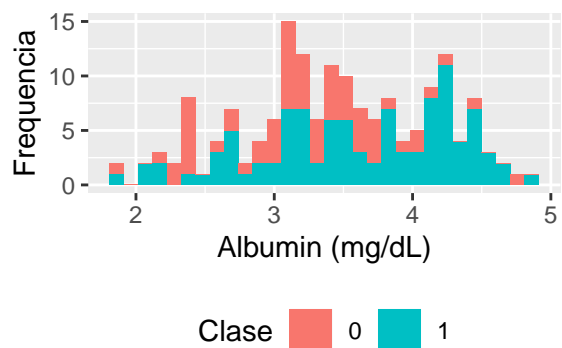
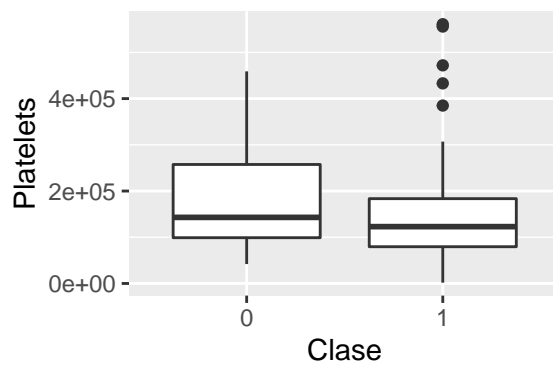
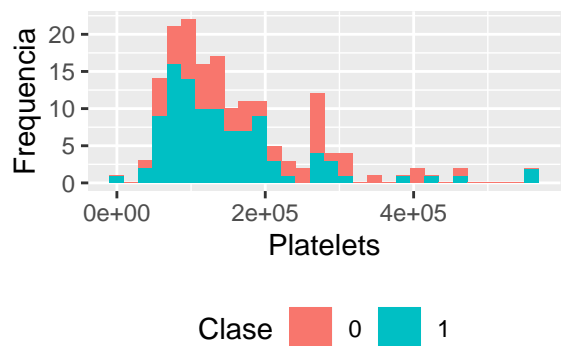
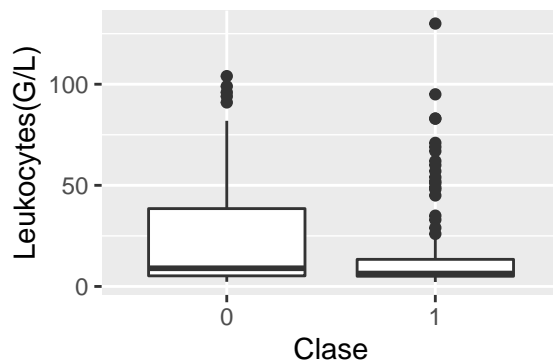
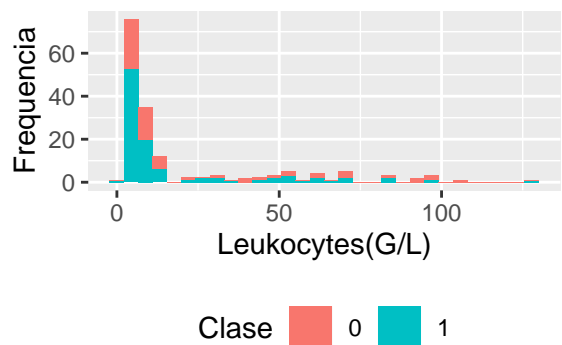
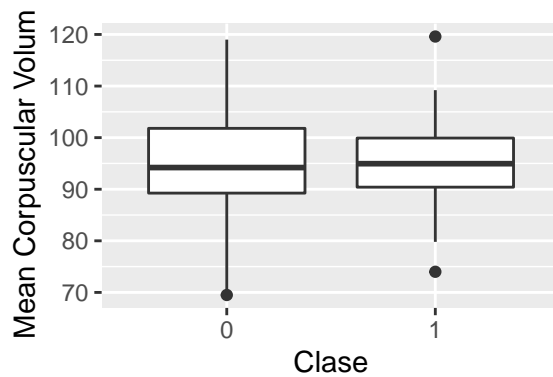
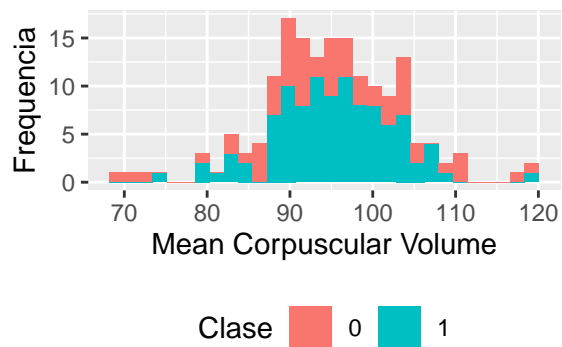
  a4<-data %>%
    ggplot(aes(x=`Class Attribute`,y=data[,2])) +
    geom_boxplot() +
    labs(x = "Clase", y = name_var[2])

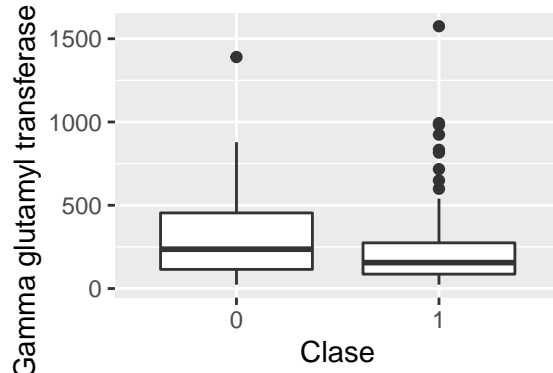
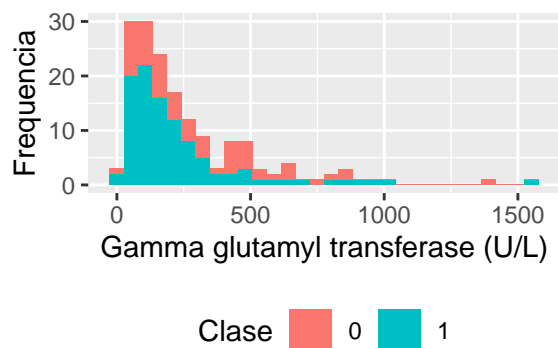
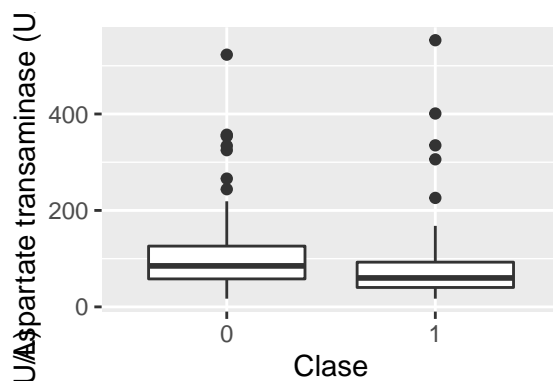
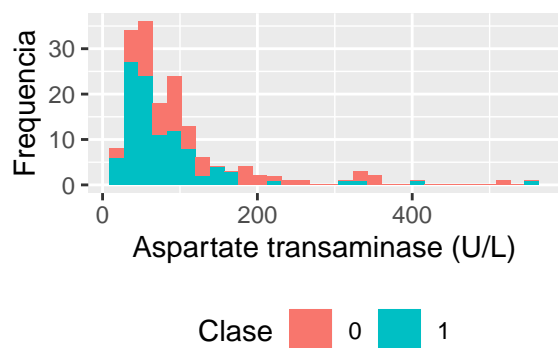
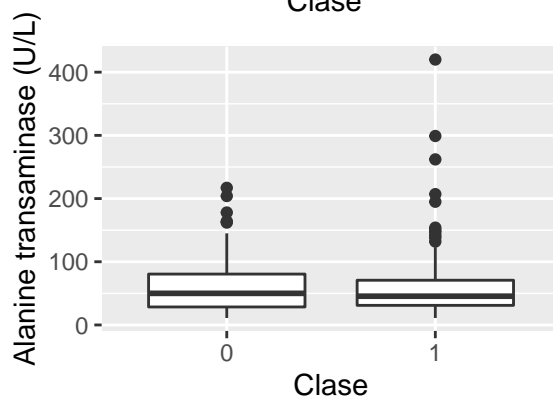
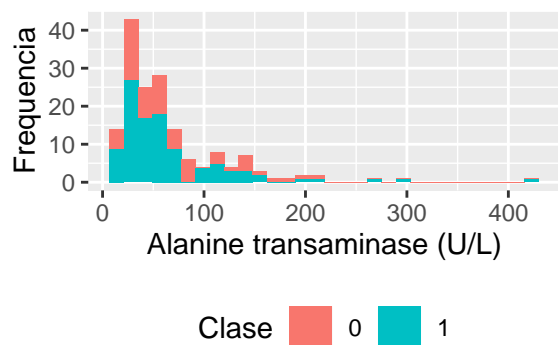
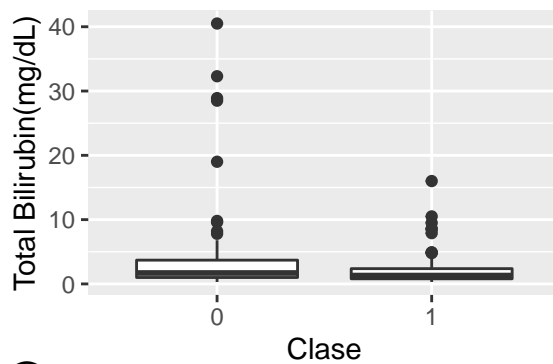
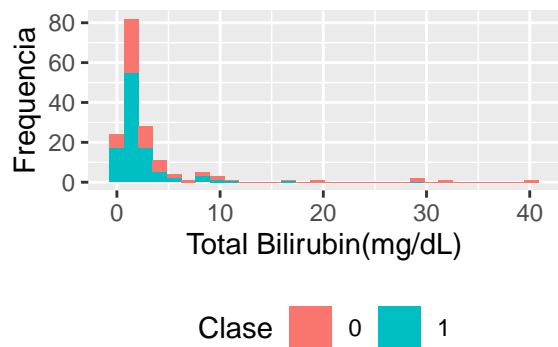
  grid.arrange(a1,a2,a3,a4,nrow=2)
} else{
  grid.arrange(a1,a2,nrow=1)
}
}
}

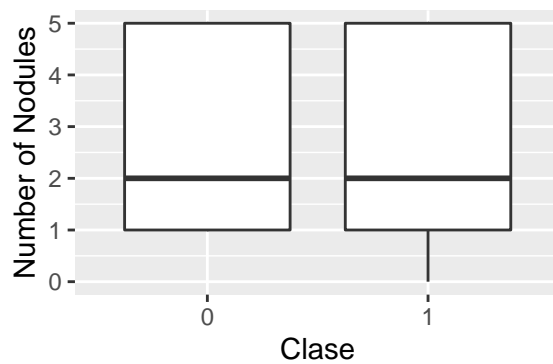
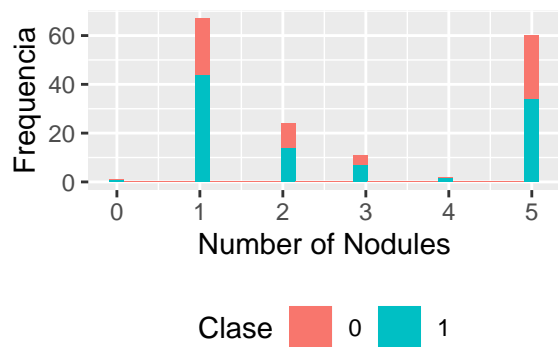
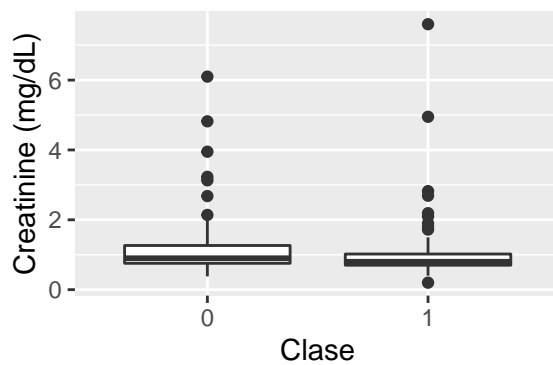
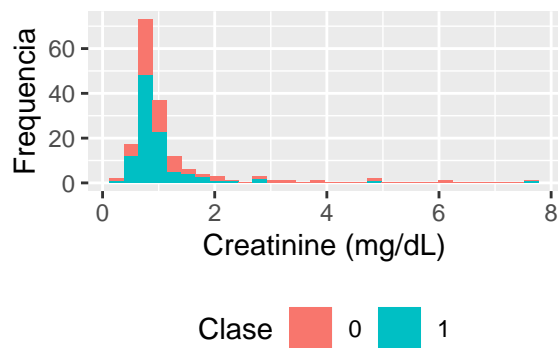
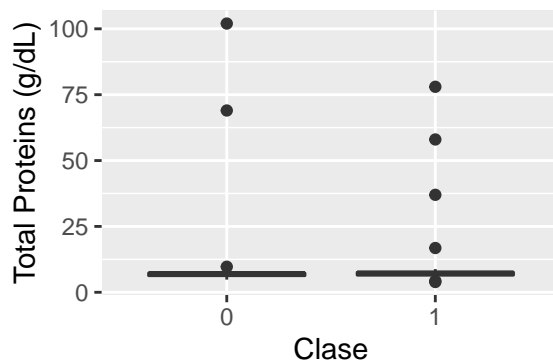
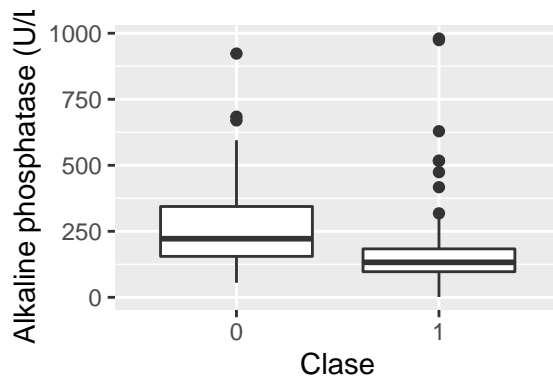
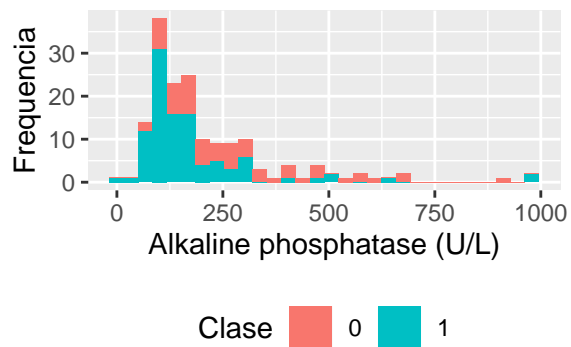
```

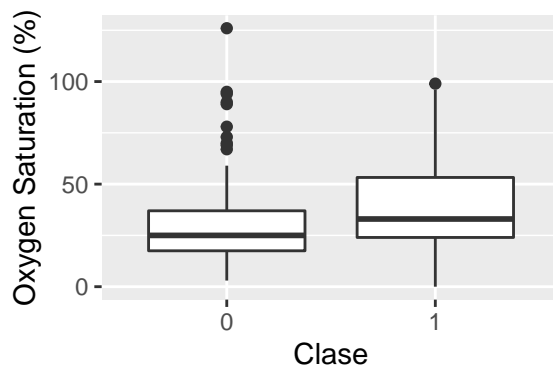
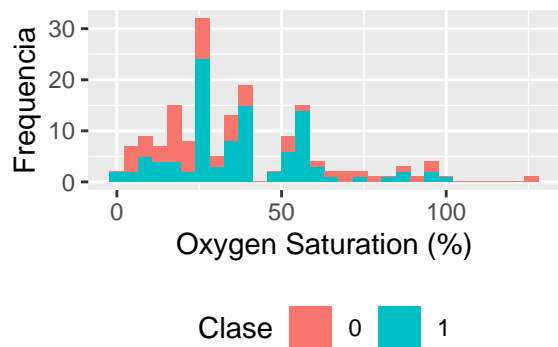
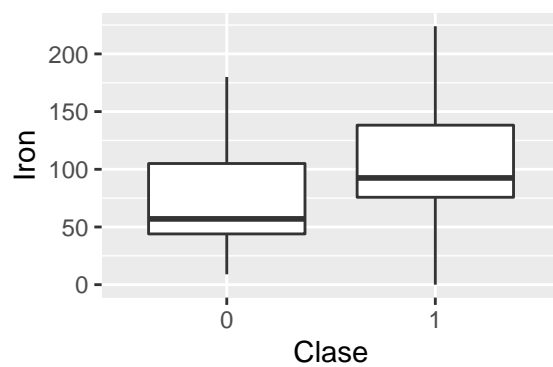
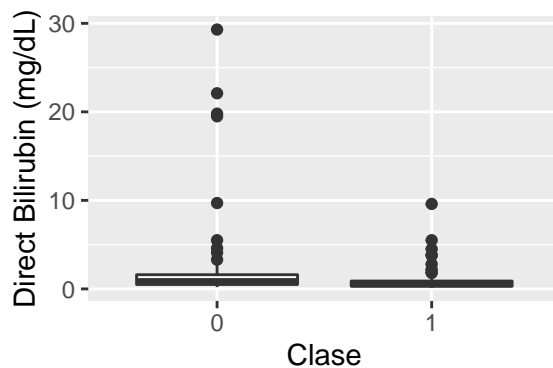
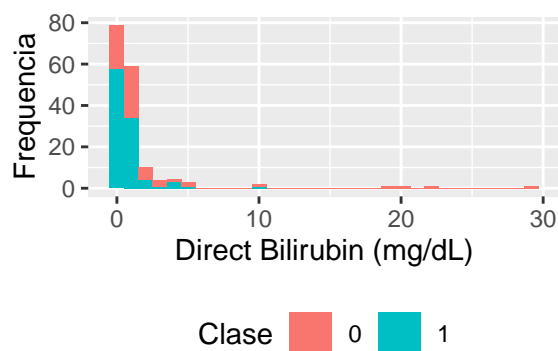
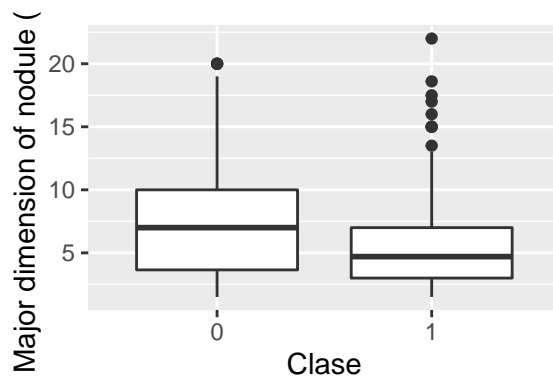
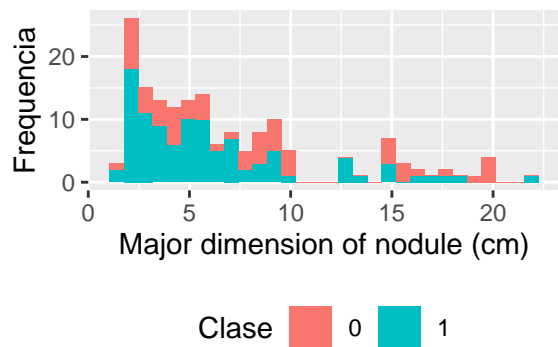


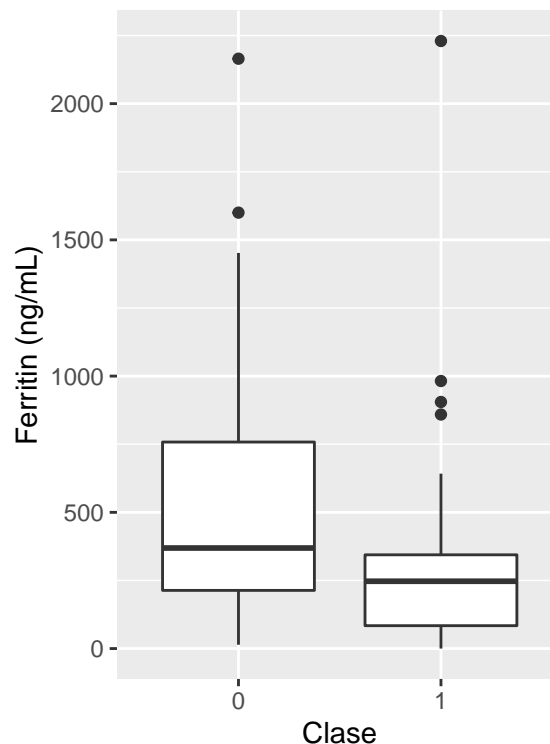
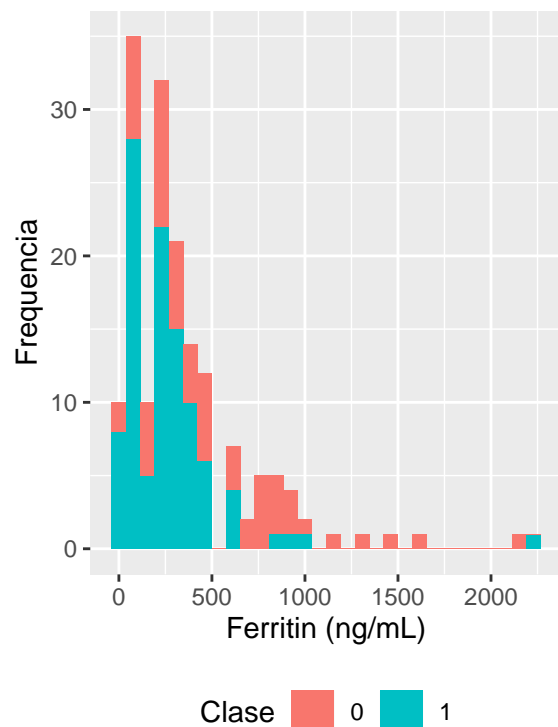












Crida l'atenció varies variables amb una distribució molt desplaçada cap a valors baixos però amb valors extrems alts. Molts son valors de laboratori, i sembla que s'adapten més a distribucions logarítmiques, per el que es modificaran. Aquestes variables son:

- Alpha-Fetoprotein (ng/mL)
- Total Bilirubin(mg/dL)
- Alanine transaminase (U/L)
- Aspartate transaminase (U/L)
- Gamma glutamyl transferase (U/L)
- Alkaline phosphatase (U/L)
- Total Proteins (g/dL)
- Creatinine (mg/dL)
- Direct Bilirubin (mg/dL)

```
# Transformació logarítmica de variables numèriques
ind_CA <- which(colnames(hcc) == "Class Attribute")
hcc_log<-c(31,37:43, 46)

ncols <- length(hcc_log)
if (ncols%%2 == 1){
  last_col = ncols
} else{
  last_col = ncols + 1
}

for (i in 1:ncols){
```

```

#transformacio
hcc[,hcc_log[i]]<- log(hcc[,hcc_log[i]])
colnames(hcc)[hcc_log[i]] <- paste("log_", colnames(hcc)[hcc_log[i]], sep = '')

if(i%%2 == 1){
  if ( i != last_col){
    data <- hcc[,c(hcc_log[i],hcc_log[i+1],ind_CA )]
  } else{
    data <- hcc[,c(hcc_log[i],ind_CA )]
  }

  name_var <- names(data)

  a1<-data %>%
    ggplot(aes(x=data[,1], fill=`Class Attribute`))+
    geom_histogram() +
    labs(fill="Clase", y = "Frecuencia", x =name_var[1] ) +
    theme(legend.position = "bottom")

  a2<-data %>%
    ggplot(aes(x=`Class Attribute`,y=data[,1])) +
    geom_boxplot() +
    labs(x = "Clase", y = name_var[1])
  if( i != last_col){
    a3<-data %>%
      ggplot(aes(x=data[,2], fill=`Class Attribute`))+
      geom_histogram() +
      labs(fill="Clase", y = "Frecuencia", x =name_var[2] ) +
      theme(legend.position = "bottom")

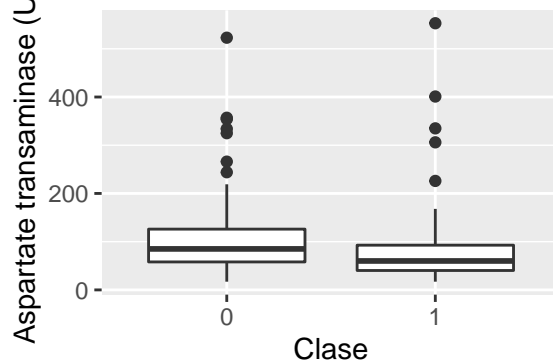
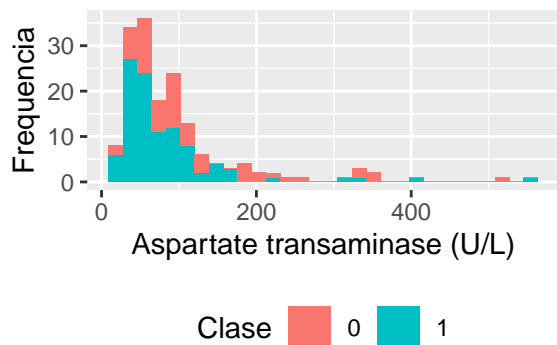
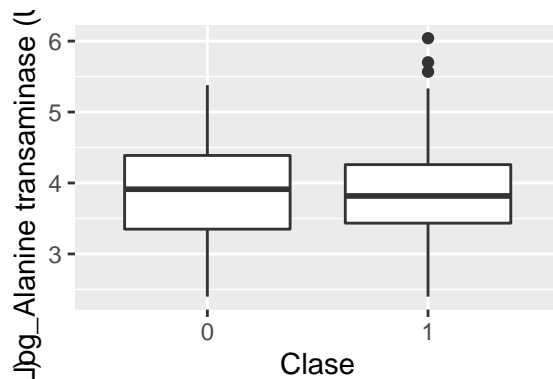
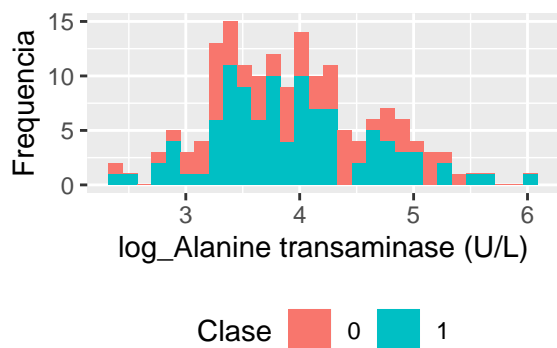
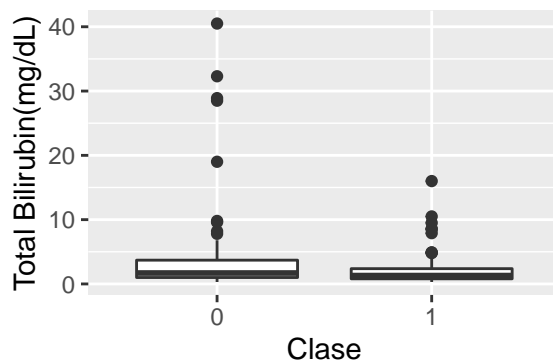
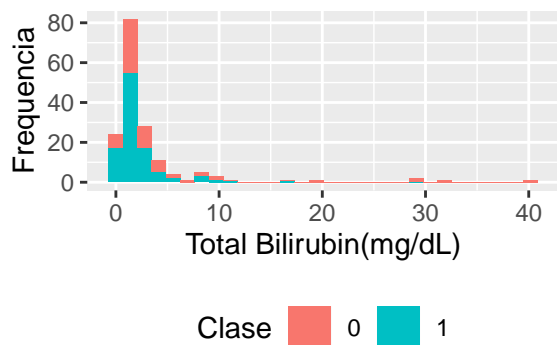
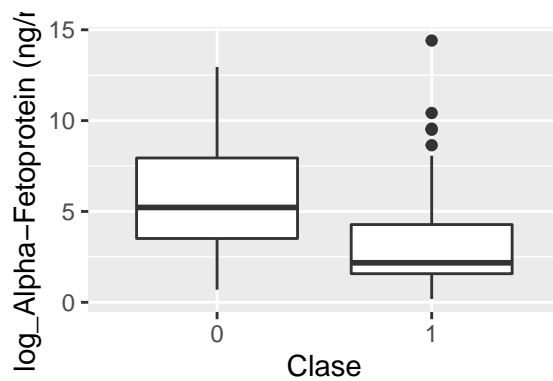
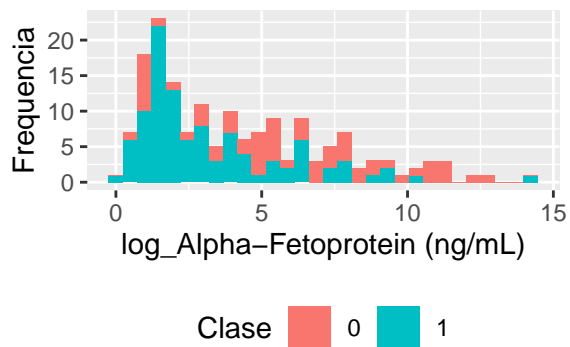
    a4<-data %>%
      ggplot(aes(x=`Class Attribute`,y=data[,2])) +
      geom_boxplot() +
      labs(x = "Clase", y = name_var[2])

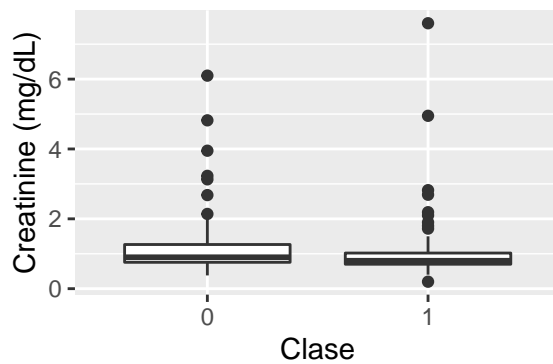
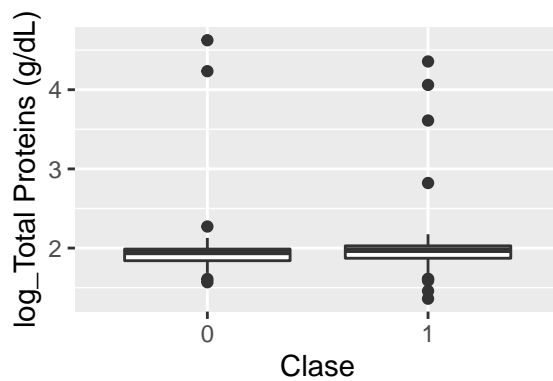
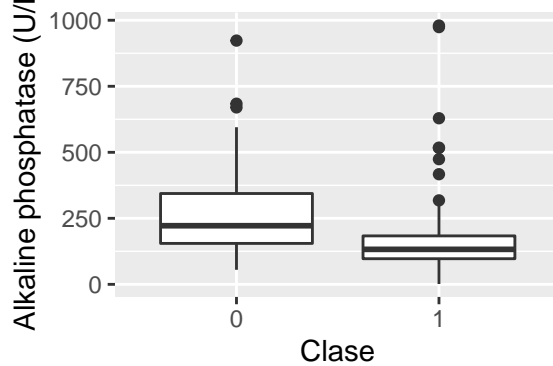
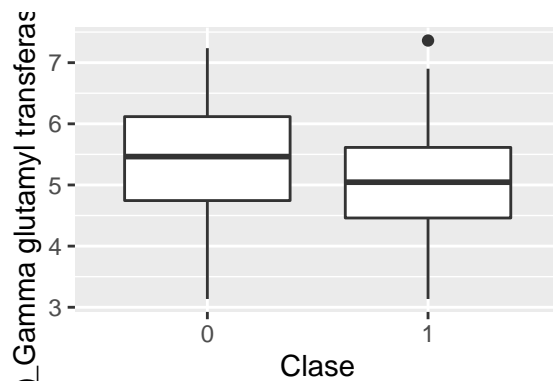
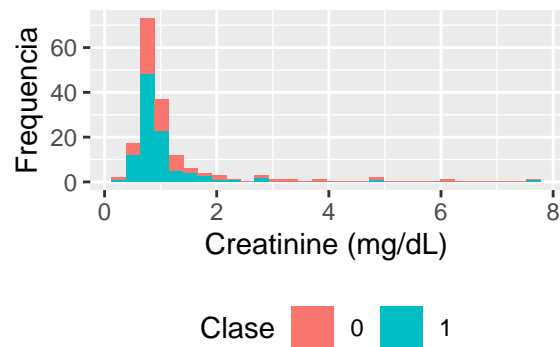
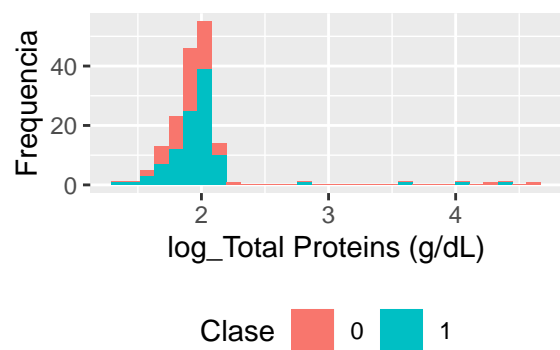
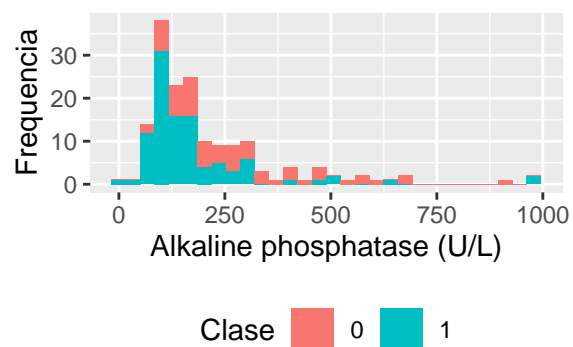
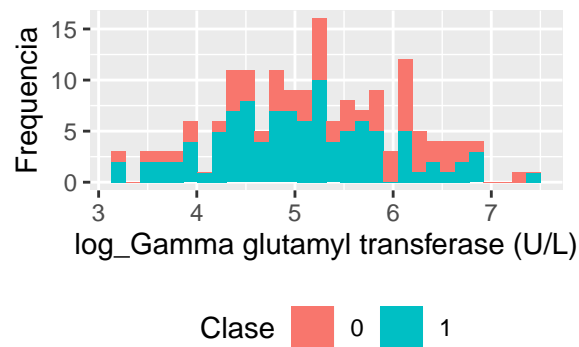
    grid.arrange(a1,a2,a3,a4,nrow=2)

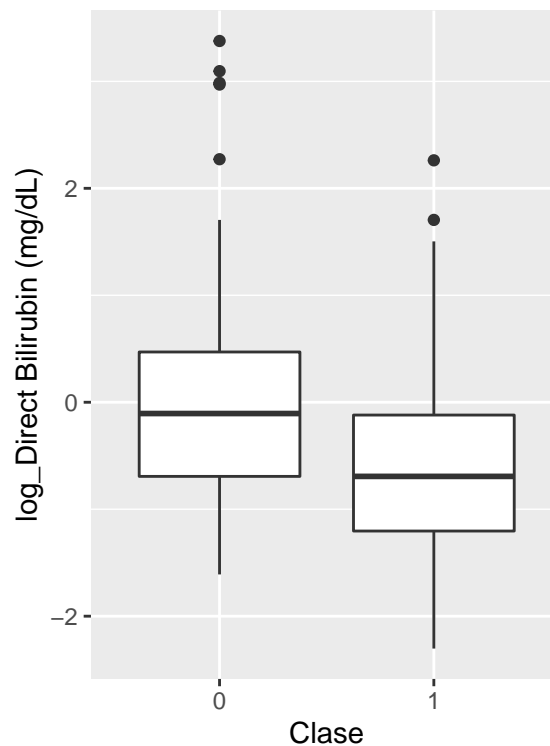
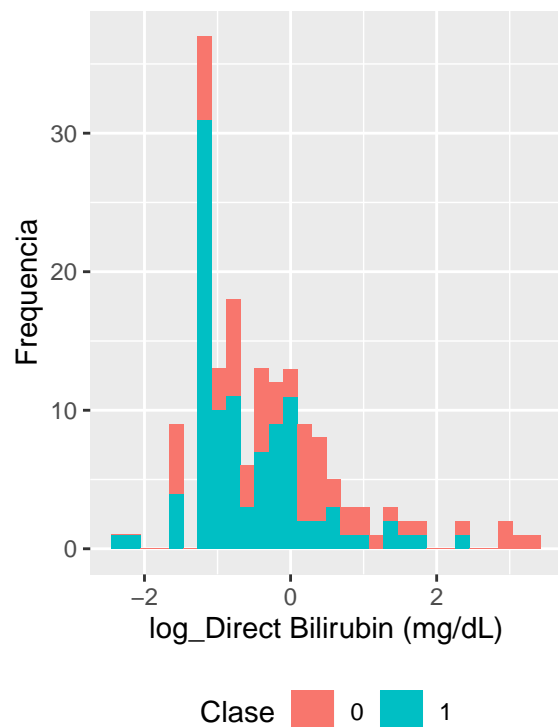
  } else{
    grid.arrange(a1,a2,nrow=1)

  }
}
}

```







```
# Distribució variables quantitatives
hcc_factorT<-c(1:23,27:29)

ncols <- length(hcc_factorT)
if (ncols%%2 == 1){
  last_col = ncols
} else{
  last_col = ncols + 1
}

#for (i in hcc_num) {
for (i in 1:ncols) {
  if(i%%2 == 1){
    if ( i != last_col){
      data <- hcc[,c(hcc_factorT[i],hcc_factorT[i+1],ind_CA )]
    } else{
      data <- hcc[,c(hcc_factorT[i],ind_CA )]
    }
    name_var <- names(data)

    a1<-data %>%
      ggplot(aes(x=data[,1],fill=`Class Attribute`))+
      geom_bar() +
      labs(fill="Clase", x = name_var[1], y = "Frecuencia") +
      theme(legend.position = "bottom")

    a2<- data %>%
      ggplot(aes(x=data[,1],fill=`Class Attribute`))+
      geom_bar(position = "fill") +
```

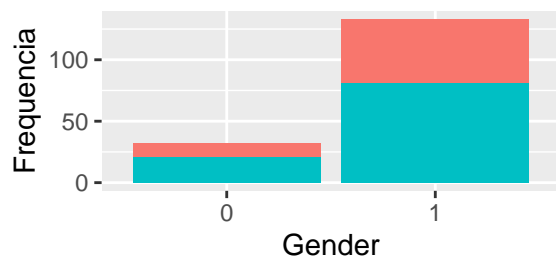
```

labs(fill="Clase", x = name_var[1], y = "Proporcio") +
theme(legend.position = "bottom")
if( i != last_col){
a3<-data %>%
  ggplot(aes(x=data[,2],fill=`Class Attribute`))+
  geom_bar() +
  labs(fill="Clase", x = name_var[2], y = "Frecuencia") +
  theme(legend.position = "bottom")

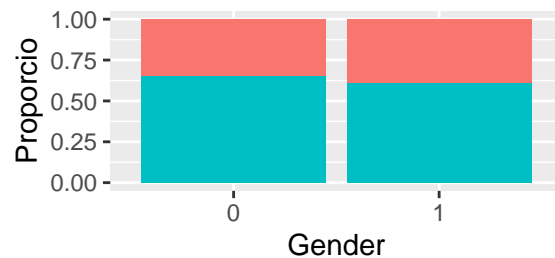
a4<- data %>%
  ggplot(aes(x=data[,2],fill=`Class Attribute`))+
  geom_bar(position = "fill") +
  labs(fill="Clase", x = name_var[2], y = "Proporcio") +
  theme(legend.position = "bottom")

grid.arrange(a1,a2,a3,a4,nrow=2)
} else{
  grid.arrange(a1,a2,nrow=1)
}
}
}

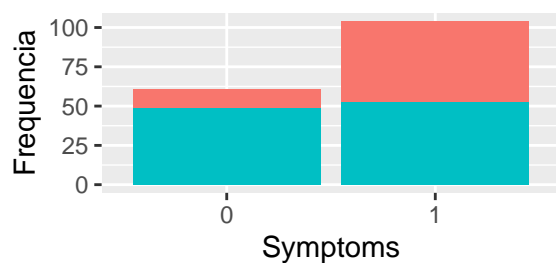
```



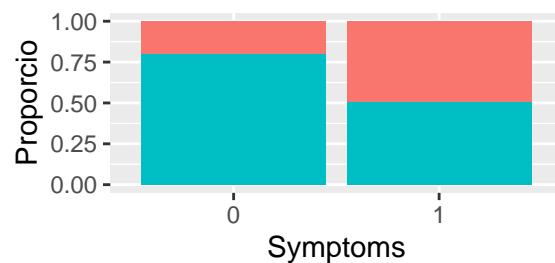
Clase 0 1



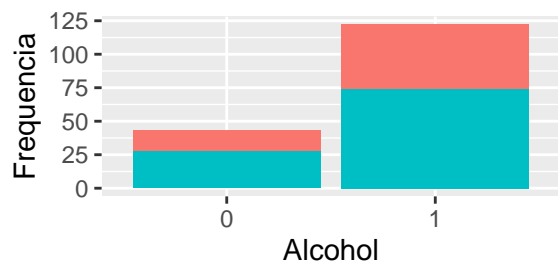
Clase 0 1



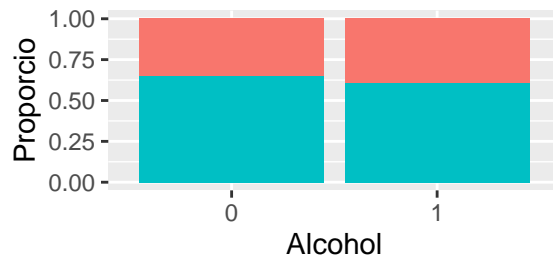
Clase 0 1



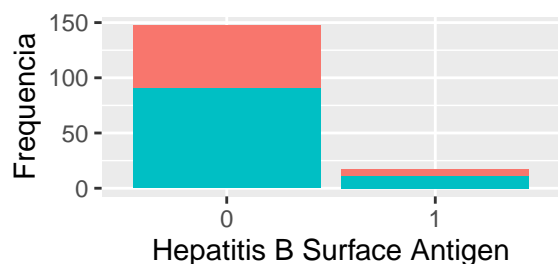
Clase 0 1



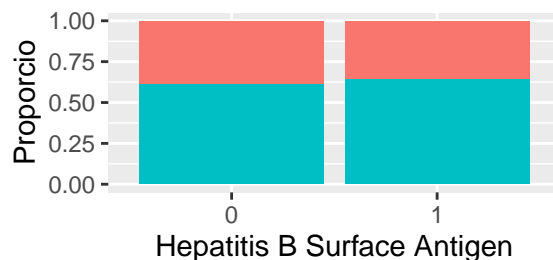
Clase 0 1



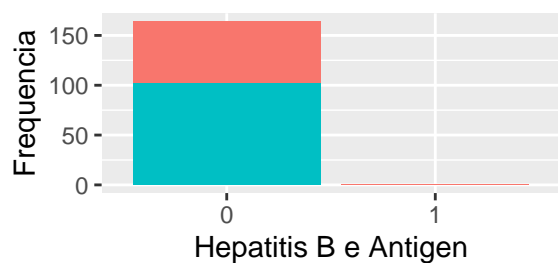
Clase 0 1



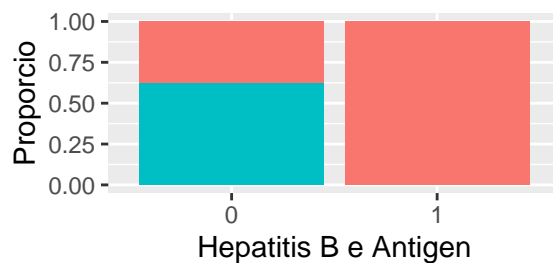
Clase 0 1



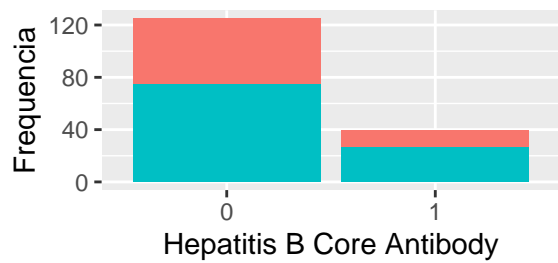
Clase 0 1



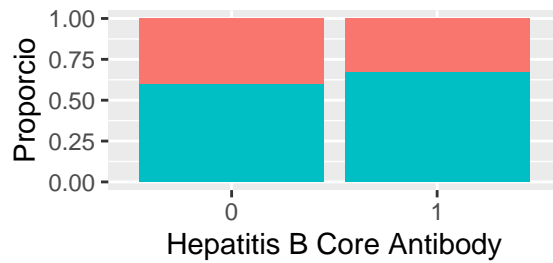
Clase 0 1



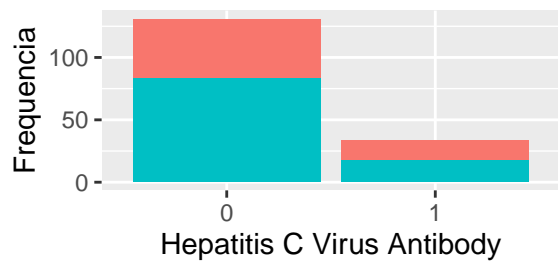
Clase 0 1



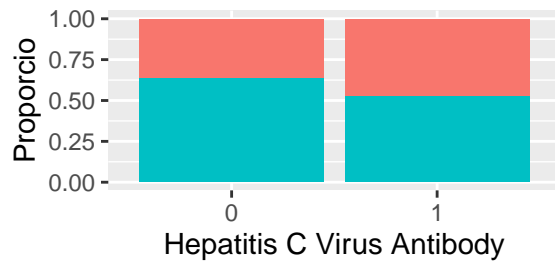
Clase 0 1



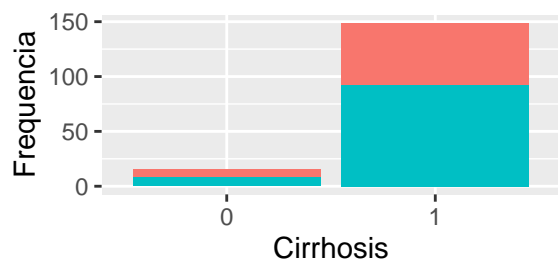
Clase 0 1



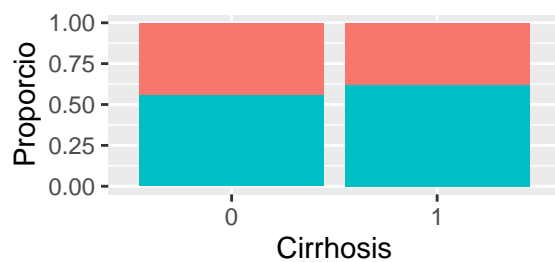
Clase 0 1



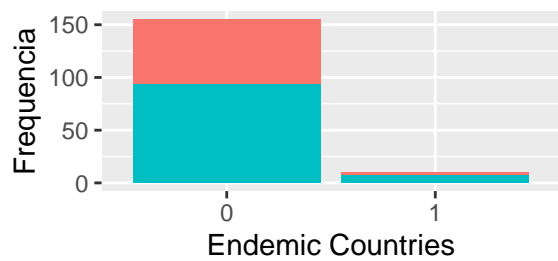
Clase 0 1



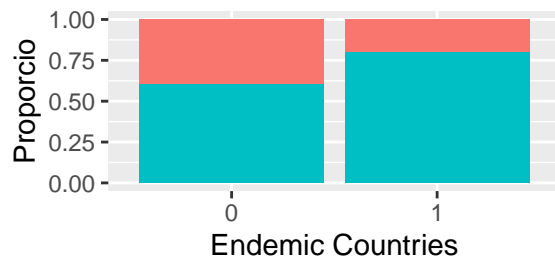
Clase 0 1



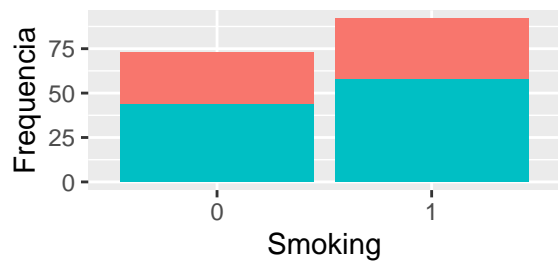
Clase 0 1



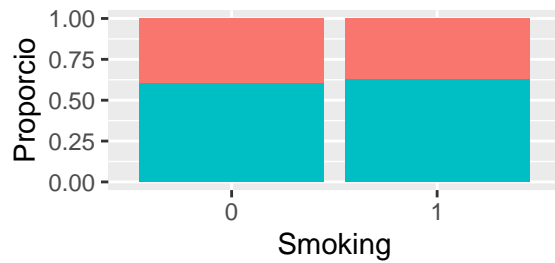
Clase 0 1



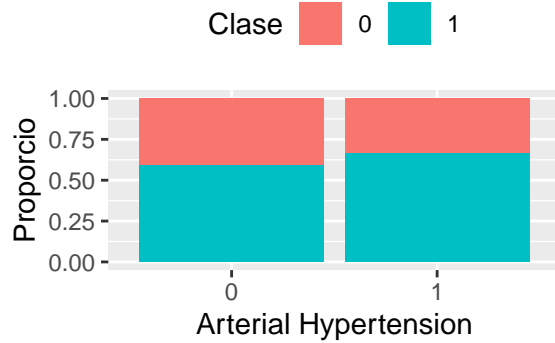
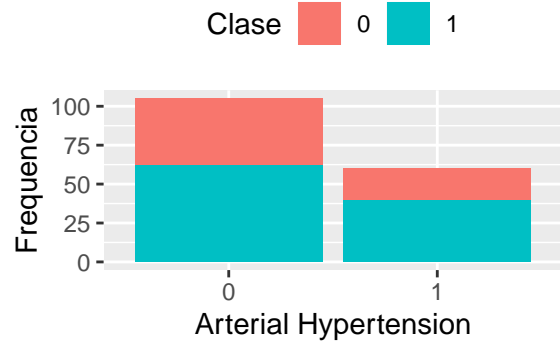
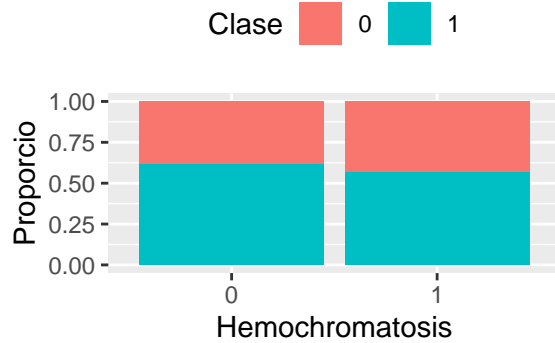
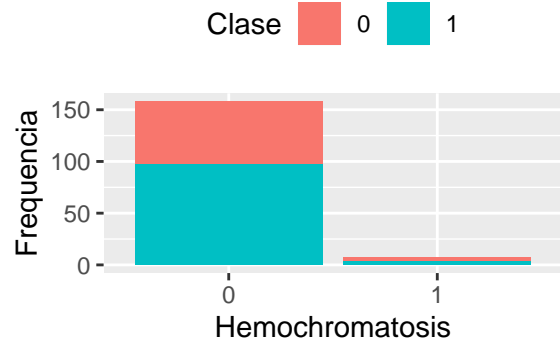
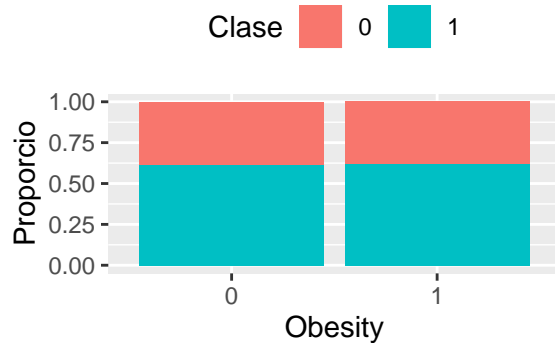
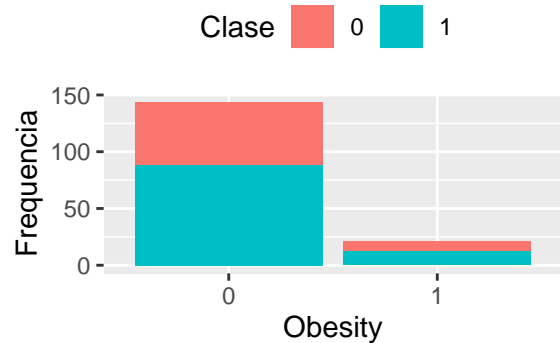
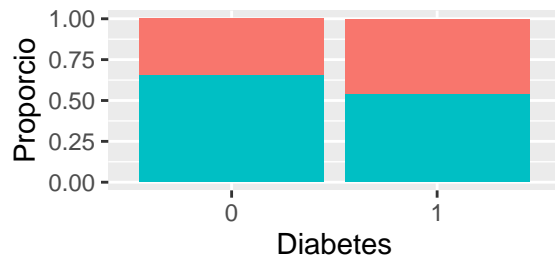
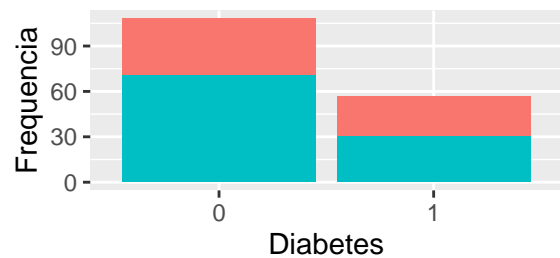
Clase 0 1

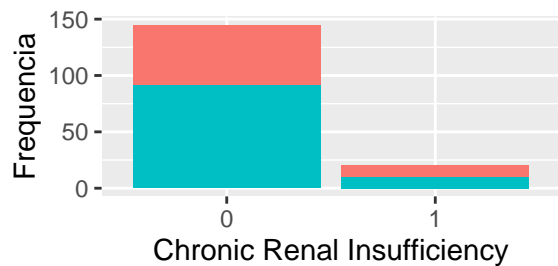


Clase 0 1

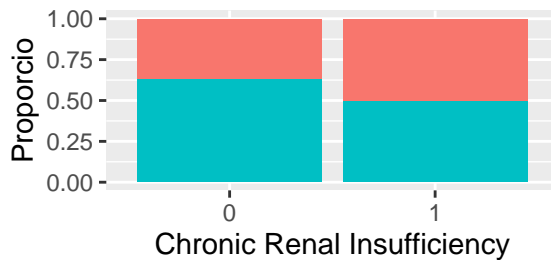


Clase 0 1

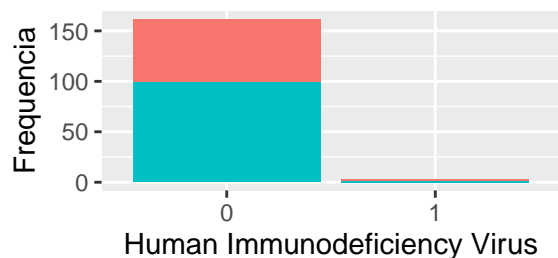




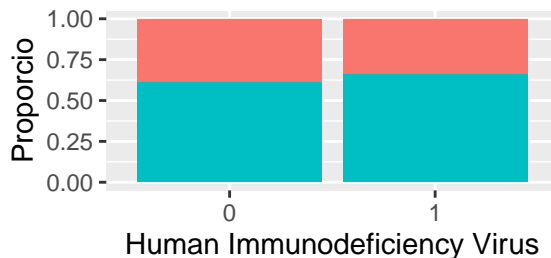
Clase 0 1



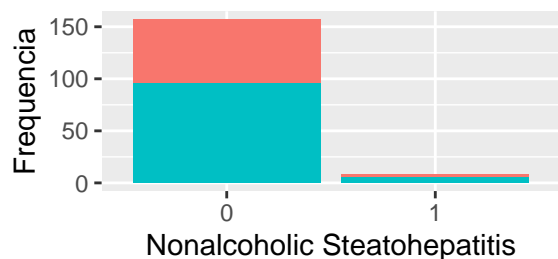
Clase 0 1



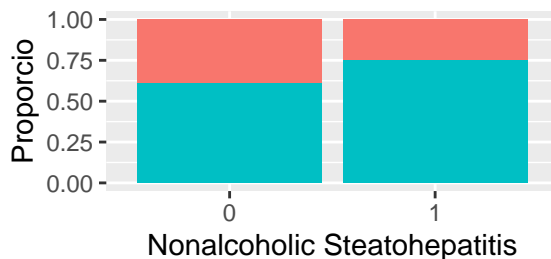
Clase 0 1



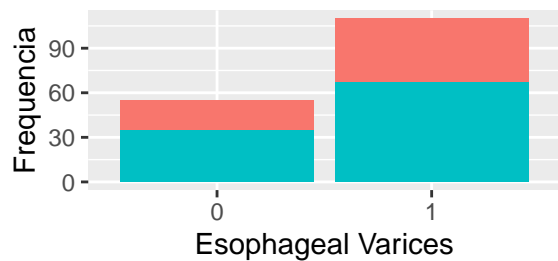
Clase 0 1



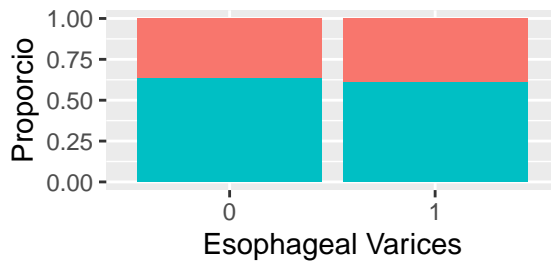
Clase 0 1



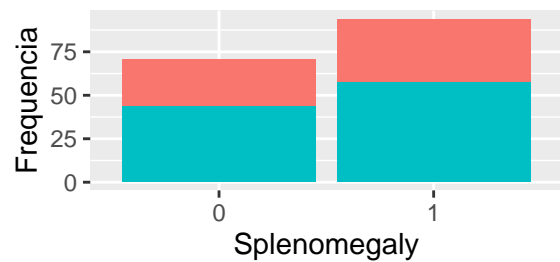
Clase 0 1



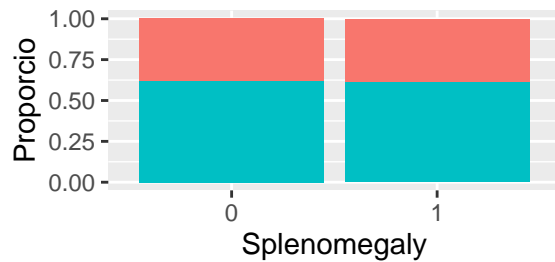
Clase 0 1



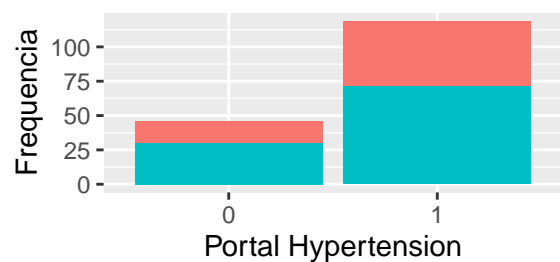
Clase 0 1



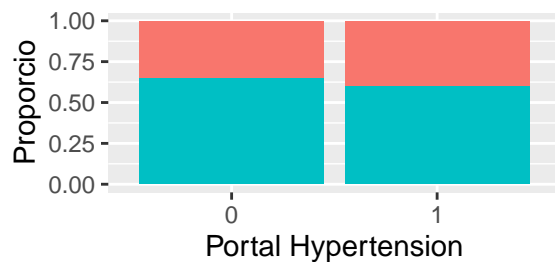
Clase 0 1



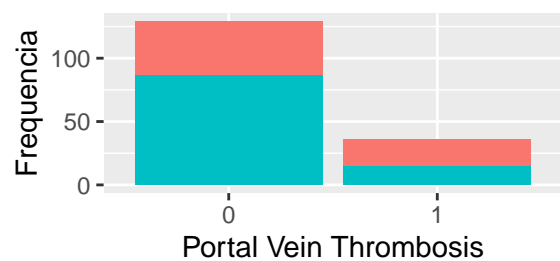
Clase 0 1



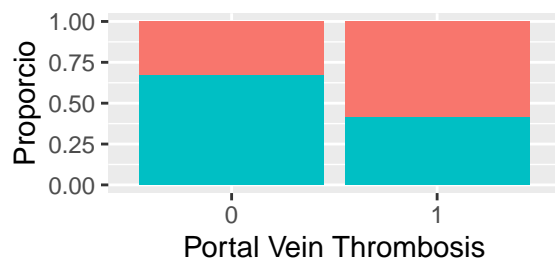
Clase 0 1



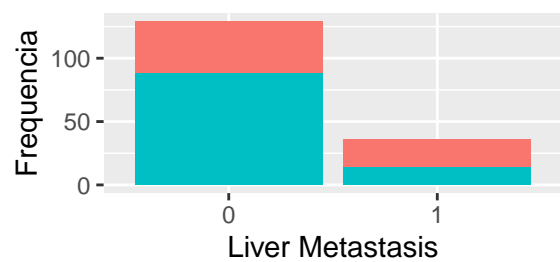
Clase 0 1



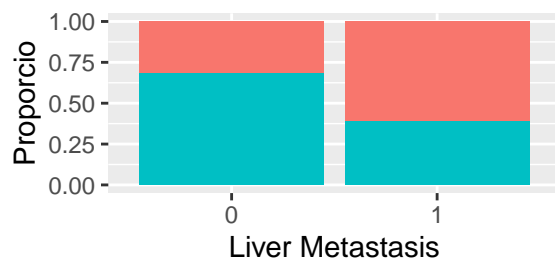
Clase 0 1



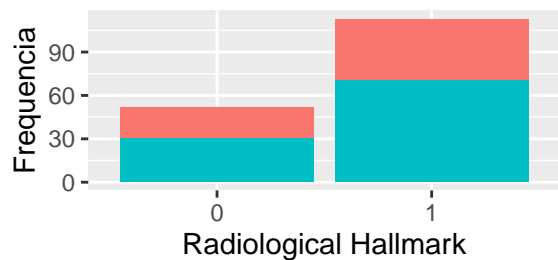
Clase 0 1



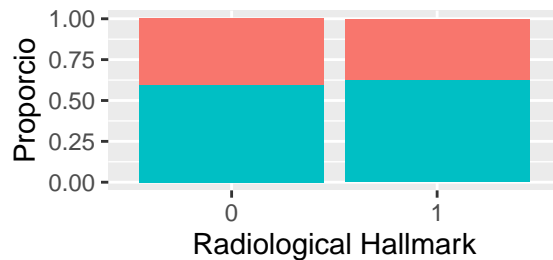
Clase 0 1



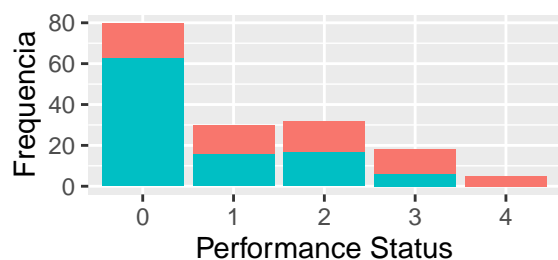
Clase 0 1



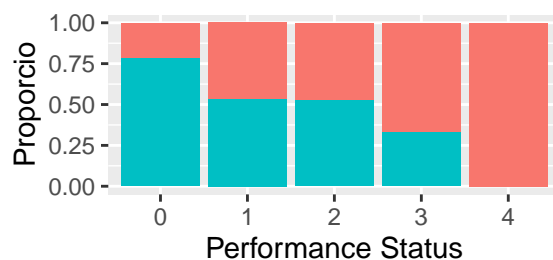
Clase 0 1



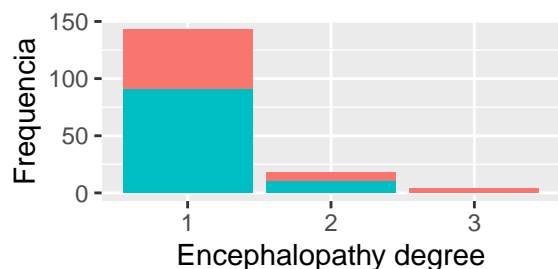
Clase 0 1



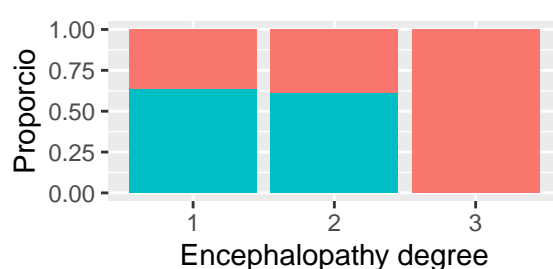
Clase 0 1



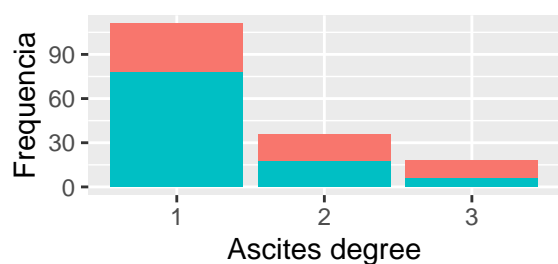
Clase 0 1



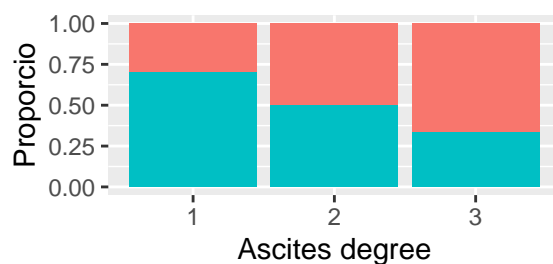
Clase 0 1



Clase 0 1



Clase 0 1



Clase 0 1

Comprobació de la normalitat

Per la comprovació de la distribució normal dels valors quantitativs utilitzarem la prova de normalitat de Anderson-Darling.

```

col.names = colnames(hcc)
tNorm <- tibble()
for (i in 1:ncol(hcc)) {
  if (is.integer(hcc[,i]) | is.numeric(hcc[,i])) {
    p_val = ad.test(hcc[,i])$p.value
    tNorm <- tNorm %>% bind_rows(c("Variable" = col.names[i], "p_value" = p_val))
  }
}

var_normales<-tNorm$Variable[as.numeric(tNorm$p_value)>0.05]

tau <- tNorm %>% filter(p_val < 0.05) %>%
  mutate_at(.vars = c("p_value"), as.numeric)
tau <- cbind(tau[1:12,],tau[13:24,])

tau[12,3 ] <- ""
tau[12,4 ] <- 0

kable(x = tau, format = "latex", caption = "Variables que no segueixen una distribuci3 normal",
      booktabs = TRUE, digits = 4) %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))

```

Table 3: Variables que no segueixen una distribuci3 normal

Variable	p_value	Variable	p_value
Age at diagnosis	0.0015	log_Aspartate transaminase (U/L)	0.0187
Grams of Alcohol per day	0.0000	log_Gamma glutamyl transferase (U/L)	0.6474
Packs of cigarets per year	0.0000	log_Alkaline phosphatase (U/L)	0.0002
International Normalised Ratio	0.0000	log_Total Proteins (g/dL)	0.0000
log_Alpha-Fetoprotein (ng/mL)	0.0000	log_Creatinine (mg/dL)	0.0000
Haemoglobin (g/dL)	0.1003	Number of Nodules	0.0000
Mean Corpuscular Volume	0.1435	Major dimension of nodule (cm)	0.0000
Leukocytes(G/L)	0.0000	log_Direct Bilirubin (mg/dL)	0.0000
Platelets	0.0000	Iron	0.0000
Albumin (mg/dL)	0.0560	Oxygen Saturation (%)	0.0000
log_Total Bilirubin(mg/dL)	0.0000	Ferritin (ng/mL)	0.0000
log_Alanine transaminase (U/L)	0.0310		0.0000

Com es pot veure, hi ha moltes variables num3riques que es distancien significativament de la distribuci3 normal, per el que s'usaran test no param3trics (Mann-Whitney-Wilcoxon) per la comparativa en relaci3 a la superviv3ncia.

Nom3s hi ha quatre variables que no es pot assegurar que no segueixin una distribuci3 normal. Comprovarem si aquestes mantenen una varian3a similar entre els grups formats per la classe amb el test de Levene.

```

library(car)
tVar <- tibble()

for (i in var_normales){
  p_val = leveneTest(group=hcc$`Class Attribute`, y=hcc[,i])$`Pr(>F)`[1]

```

```
tVar <- tVar %>% bind_rows(c("Variable" = i, "p_value" = p_val))
}
```

FALTA PONER LAS TABLAS CON KABLE.

Com es por veure, només dues de totes les variables quantitatives (**Haemoglobin (g/dL)** i **log_Gamma glutamyl transferase (U/L)**) segueixen una distribució normal i no tenen diferències significatives entre les variances entre la població que sobreviu i la que no.

Proves estadístiques

Comparació entre grups de la classe

Per tal de valorar quines variables es comporten diferents entre en que sobreviuen i els que no, es realitzarà els diferents test estadístics:

- Per les variables quantitatives, donada les seves distribucions majoritària diferent a la normalitat, es realitzara el test no paramètric de Mann-Whitney-Wilcoxon
- Per les variables qualitatives es realitzarà un test chi-quadrat.

```
testChi <- tibble()

for (i in hcc_factorT) {
  tau=table(hcc[,i], hcc$`Class Attribute`)
  chi=chisq.test(tau)
  testChi <- testChi %>% bind_rows(c(Class = i, Name=names(hcc[i]),Categorica="1",p_value = chi$p.value))
}

tau <- testChi %>% filter(p_value < 0.1) %>%
  mutate_at(.vars = c("p_value"), as.numeric)

varSig <- tau

kable(x = tau, format = "latex", caption = "Variables categòriques amb p<0.10 entre la classe",
      booktabs = TRUE, digits = 4) %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))
```

Table 4: Variables categòriques amb p<0.10 entre la classe

Clase	Name	Categorica	p__value
2	Symptoms	1	0.0003
21	Portal Vein Thrombosis	1	0.0088
22	Liver Metastasis	1	0.0026
28	Encephalopathy degree	1	0.0354
29	Ascites degree	1	0.0029

```

testWil <- tibble()
for (i in hcc_num) {
  wil=wilcox.test(hcc[hcc$`Class Attribute`==1,i],hcc[hcc$`Class Attribute`==0,i], mu = 0,paired = FALSE)
  testWil <- testWil %>% bind_rows(c(Class = i, Name=names(hcc[i]),Categorica="0",p_value = wil$p.value))
}

tau <- testWil %>% filter(p_value < 0.1) %>%
  mutate_at(.vars = c("p_value"), as.numeric)

varSig <- varSig %>% bind_rows(tau)

tau <- cbind(tau[1:6,],tau[7:12,])

kable(x = tau, format = "latex", caption = "Variables num riques amb p<0.10 entre la classe",
      booktabs = TRUE, digits = 4) %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))

```

Table 5: Variables num riques amb p<0.10 entre la classe

Clase	Name	Categorica	p_value	Clase	Name	Categorica	p_value
24	Age at diagnosis	0	0.0357	40	log_Gamma glutamyl transferase (U/L)	0	0.0166
30	International Normalised Ratio	0	0.0227	43	log_Creatinine (mg/dL)	0	0.0967
35	Platelets	0	0.0385	45	Major dimension of nodule (cm)	0	0.0137
36	Albumin (mg/dL)	0	0.0001	46	log_Direct Bilirubin (mg/dL)	0	0.0004
37	log_Total Bilirubin(mg/dL)	0	0.0195	47	Iron	0	0.0005
39	log_Aspartate transaminase (U/L)	0	0.0011	48	Oxygen Saturation (%)	0	0.0408

Aquestes variables seran les que es seleccionar  per a la creaci  d'un model de regressi  log stica, per  abans, valorarem les correlacions entre elles per tal de seleccionar variables que estiguin poc relacionades entre elles.

Correlaci  entre les variables seleccionades

Amb respecte a la correlaci  entre les variables seleccionades, veiem la seva matriu de correlacions. Donada l'exist ncia de variables cat g riques, aquestes es consideraran num riques i usarem la correlaci  no param trica de Spearman per a la seva valoraci . Previament es normalitzaran totes les variables quantitatives.

```

# Correlaci  variables independents.
hcc_norm <- hcc
atr<-names(hcc)
hcc_norm[,hcc_num]<-scale(hcc_norm[,hcc_num])

hcc2<-as.data.frame(lapply(hcc_norm,as.numeric))

names(hcc2)<-atr

hcc_cor<-cor(hcc2[,varSig$Name], method = "spearman")

col.names <- colnames(hcc_cor)
hcc_corTop <- as.tibble(hcc_cor) %>% melt()
hcc_corTop$variable2 <- col.names
hcc_corTop <- hcc_corTop %>%
  select(variable, variable2, value) %>%

```



```

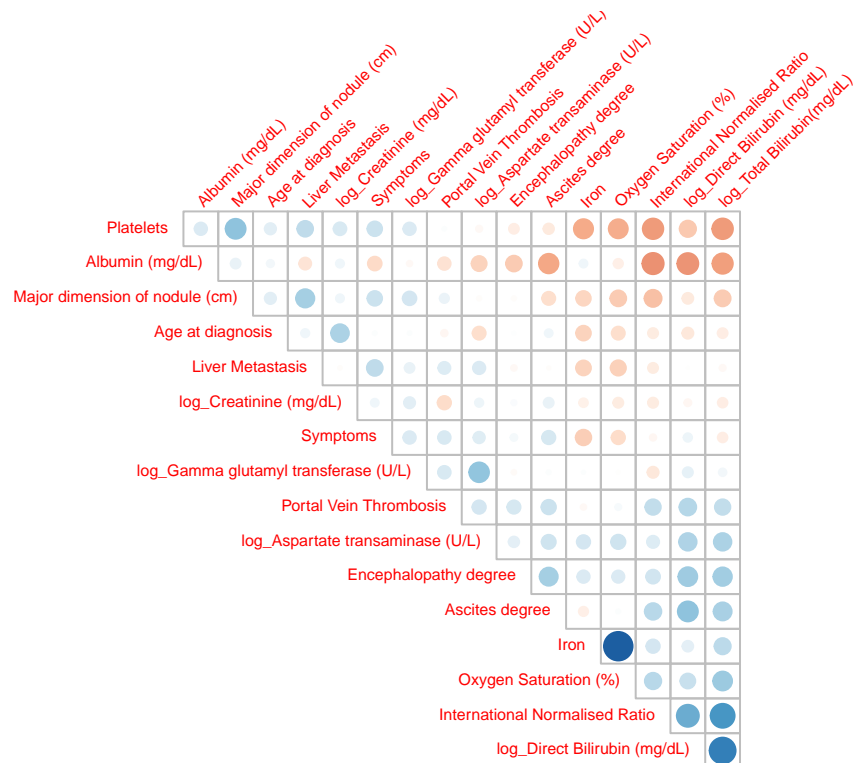
rename(corr = value) %>%
filter(variable != variable2) %>%
filter(abs(corr) >= 0.4) %>%
arrange(-corr) %>%
group_by(corr) %>%
mutate(vars = paste(variable, collapse = '-')) %>%
select(vars, corr) %>%
unique()

```

```

corrplot(hcc_cor, cl.pos='n', tl.srt = 45, tl.cex = 0.5, type="upper", method = "circle", order="FPC", di

```



Per intentar reduir el número de variables del model, d'entre les que han tingut difències significatives entre els que sobreviuen i els que no, seleccionarem les que tinguin poca correlació amb altres variables, i amb les que tenen molta correlació amb d'altres, només seleccionarem una com a representativa.

Com es pot veure, només dues variables tenen una molt alta correlació (**Iron** i **Oxygen Saturation (%)**). De les dues, seleccionarem **Iron** que es la que presenta un menor probabilitat de significació.

Creació de noves variables

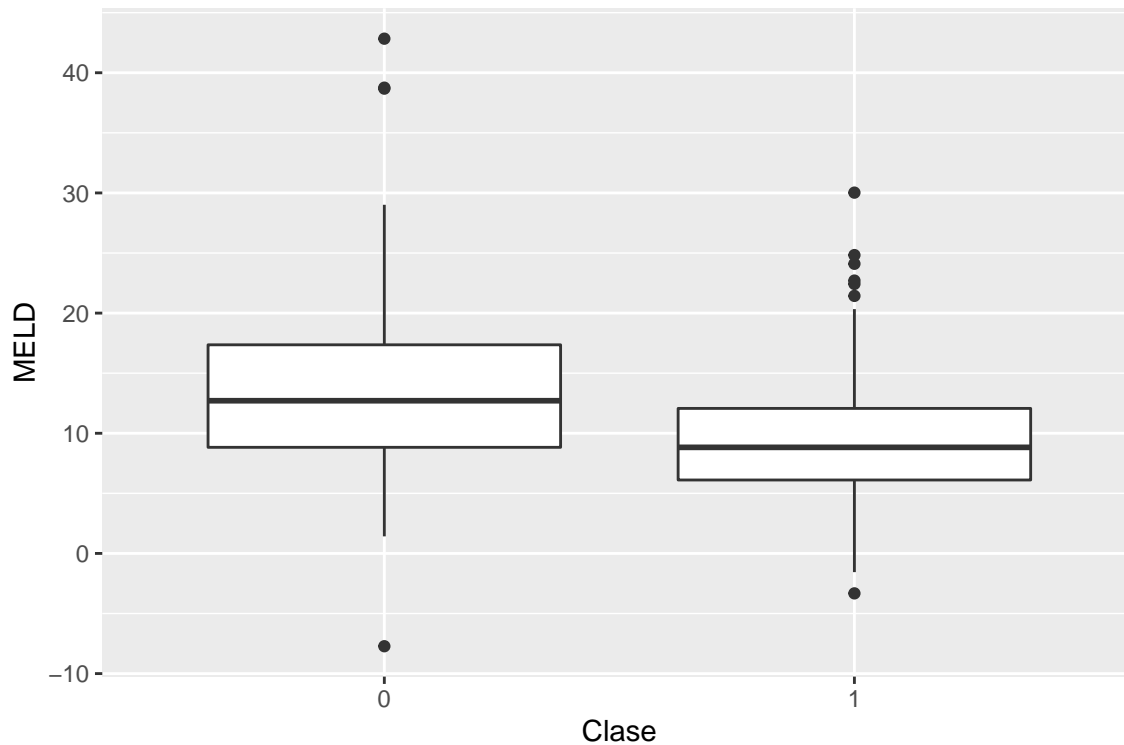
Hi ha un valor extès per a valorar la probabilitat de mort als tres mesos de pacients amb hepatopatia que depend de la creatinina, la bilirrubina total i del INR, amb un valor calculat denominat MELD. Aquest valor és una estimació de probabilitat de fallida hepàtica. A majors valors major probabilitat de mort. Aquest valor està validat pels 3 mesos, no per a l'any, com es el nostre cas.

```

# MELC calculat vs classe
hcc <- hcc %>%

```

```
mutate(MELD = 3.78*`log_Total Bilirubin(mg/dL)`+
      11.2*log(`International Normalised Ratio`)+
      9.57*hcc$log_Creatinine (mg/dL)`+
      6.43 )
hcc %>%
  ggplot(aes(x=`Class Attribute`,y=MELD)) +
    geom_boxplot() +
    labs(x = "Clase", y = "MELD")
```



```
#Hi ha diferències en la supervivència depenent del MELD
x = hcc %>% filter(`Class Attribute`==1) %>% pull(MELD)
y = hcc %>% filter(`Class Attribute`==0) %>% pull(MELD)
wilcox.test(x,y, mu = 0,paired = FALSE, conf.int = 0.95)
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data: x and y
## W = 2139.5, p-value = 0.0003196
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
## -5.697309 -1.735361
## sample estimates:
## difference in location
## -3.697976
```

És pot veure que els valors de MELD són significativament diferents entre el grup de pacients que sobreviuen i els que no. Amb aquesta combinació lineal agrupem en una única variable la bilirrubina total, l'INR i la creatinina.

Per tant, les variables seleccionades per estudiar amb regressió logística seran:

- Symptoms
- Portal Vein Thrombosis
- Liver Metastasis
- Age at diagnosis
- Performance Status (Aquest valor no esta entre les varSig)
- log_Alpha-Fetoprotein (ng/mL) (Aquest valor no esta entre les varSig)
- Haemoglobin (g/dL) (Aquest valor no esta entre les varSig)
- log_Aspartate transaminase (U/L)
- log_Gamma glutamyl transferase (U/L)
- Major dimension of nodule (cm)
- Iron
- Ferritin (ng/mL) (Aquest valor no esta entre les varSig)
- MELD

Model amb regressió logística

Amb les variables seleccionades, es crearà un model de regressió logística per tal de predir la supervivència a l'any del diagnòstic d'HCC. La variable ordinal es considerarà numérica.

```
var_selecc<-varSig$Name[-c(7,10,13,17)]

hcc_sel<-hcc[c(var_selecc, "MELD", "Class Attribute")]

hcc_sel$`Ascites degree`<- as.numeric(hcc_sel$`Ascites degree`)
hcc_sel$`Encephalopathy degree`<- as.numeric(hcc_sel$`Encephalopathy degree`)

modelo <- glm(`Class Attribute` ~ ., data = hcc_sel, family = "binomial")
summary(modelo)
```

```
##
## Call:
## glm(formula = `Class Attribute` ~ ., family = "binomial", data = hcc_sel)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3256  -0.7560   0.3625   0.7034   2.1115
##
## Coefficients:
##              Estimate Std. Error z value
## (Intercept)    4.338e+00  2.493e+00  1.740
## Symptoms1      -7.260e-01  4.690e-01 -1.548
## `Portal Vein Thrombosis`1 -4.829e-01  5.045e-01 -0.957
## `Liver Metastasis`1      -3.996e-01  5.269e-01 -0.758
## `Encephalopathy degree`  -7.901e-02  5.591e-01 -0.141
## `Ascites degree`        -1.814e-01  3.164e-01 -0.574
## `Age at diagnosis`      -1.757e-02  1.667e-02 -1.054
## Platelets           4.129e-07  2.260e-06  0.183
## `Albumin (mg/dL)`       4.739e-01  3.490e-01  1.358
## `log_Aspartate transaminase (U/L)` -5.985e-01  3.439e-01 -1.740
## `log_Gamma glutamyl transferase (U/L)` -1.090e-01  2.603e-01 -0.419
```

```
## `Major dimension of nodule (cm)`      -1.021e-01  4.602e-02  -2.219
## `log_Direct Bilirubin (mg/dL)`        -3.561e-01  2.925e-01  -1.217
## Iron                                  1.533e-02  5.152e-03   2.976
## MELD                                  -7.455e-02  4.042e-02  -1.844
##                                         Pr(>|z|)
## (Intercept)                          0.08181 .
## Symptoms1                             0.12169
## `Portal Vein Thrombosis`1             0.33840
## `Liver Metastasis`1                   0.44821
## `Encephalopathy degree`               0.88761
## `Ascites degree`                      0.56631
## `Age at diagnosis`                    0.29193
## Platelets                             0.85502
## `Albumin (mg/dL)`                     0.17459
## `log_Aspartate transaminase (U/L)`     0.08182 .
## `log_Gamma glutamyl transferase (U/L)` 0.67547
## `Major dimension of nodule (cm)`      0.02647 *
## `log_Direct Bilirubin (mg/dL)`        0.22342
## Iron                                  0.00292 **
## MELD                                  0.06514 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 219.43  on 164  degrees of freedom
## Residual deviance: 152.68  on 150  degrees of freedom
## AIC: 182.68
##
## Number of Fisher Scoring iterations: 5
```

Per tal de seleccionar un model amb menys atributs sense disminuir excessivament l'error del model, es realitzarà un estudi iteratiu eliminant a cada pas la variable menys significativa ("backward").

```
library(MASS)
```

```
modback <- stepAIC(modelo, trace=FALSE, direction="backward")
```

```
summary(modback)
```

```
##
## Call:
## glm(formula = `Class Attribute` ~ Symptoms + `Albumin (mg/dL)` +
##      `log_Aspartate transaminase (U/L)` + `Major dimension of nodule (cm)` +
##      `log_Direct Bilirubin (mg/dL)` + Iron + MELD, family = "binomial",
##      data = hcc_sel)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3839  -0.7748   0.3944   0.7448   2.0297
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        2.476225    1.908340   1.298 0.194431
## Symptoms1                          -0.870723    0.457772  -1.902 0.057159
```

```

## `Albumin (mg/dL)`          0.514117    0.331922    1.549 0.121404
## `log_Aspartate transaminase (U/L)` -0.639796    0.303998   -2.105 0.035325
## `Major dimension of nodule (cm)`  -0.121036    0.041407   -2.923 0.003466
## `log_Direct Bilirubin (mg/dL)`    -0.408255    0.271661   -1.503 0.132887
## Iron                          0.016260    0.004648    3.499 0.000468
## MELD                         -0.090825    0.036977   -2.456 0.014040
##
## (Intercept)
## Symptoms1                    .
## `Albumin (mg/dL)`
## `log_Aspartate transaminase (U/L)` *
## `Major dimension of nodule (cm)`  **
## `log_Direct Bilirubin (mg/dL)`
## Iron                          ***
## MELD                         *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 219.43  on 164  degrees of freedom
## Residual deviance: 156.83  on 157  degrees of freedom
## AIC: 172.83
##
## Number of Fisher Scoring iterations: 5

```

```

library(vcd)
predicciones <- ifelse(test = modback$fitted.values > 0.50, yes = 1, no = 0)
matriz_confusion <- table(hcc_sel$`Class Attribute`, predicciones,
                          dnn = c("observaciones", "predicciones"))
matriz_confusion

```

```

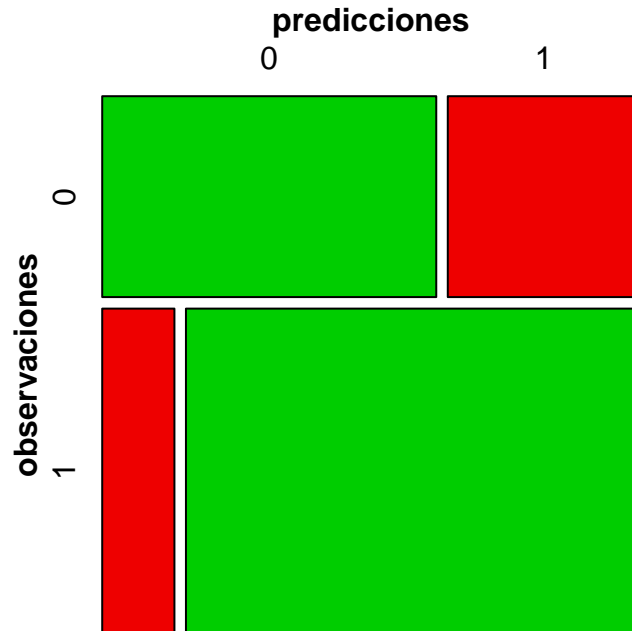
##                predicciones
## observaciones  0  1
##                0 40 23
##                1 14 88

```

```

mosaic(matriz_confusion, shade = T, colorize = T,
        gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))

```



```
cat("\n")
```

```
cat("Exactitut del model:\n")
```

```
## Exactitut del model:
```

```
(matriz_confusion[1]+matriz_confusion[4])/nrow(hcc_sel)
```

```
## [1] 0.7757576
```

El model generat te només 7 variables i obté un 78% de precisió en les prediccions.