

PRA 2: Neteja, validació i anàlisi de les dades

Álvaro Díaz i David Leiva

23/12/2020

Contents

1	Detalls de l'activitat	2
1.1	Presentació	2
1.2	Competències	2
1.3	Objectius	2
2	Resolució	3
2.1	Descripció del dataset	3
2.2	Importància i objectius de l'anàlisi	5
2.3	Integració i selecció de les dades d'interès a analitzar	5
2.3.1	Preparació del dataset	5
2.3.2	Descripció del dataset	6
2.4	Neteja de les dades	7
2.4.1	Valors nuls o buits	7
2.4.2	Correcció valors nuls de variables categòriques	9
2.4.3	Correcció valors o extrems de variables quantitatives	10
2.4.4	Exportació del dataset netejat	10
2.5	Anàlisi de les dades	10
2.5.1	Variables numèriques	10
2.5.2	Variables qualitatives	22
3	Versio Prova	23
3.0.1	Comprobació de la normalitat i homogeneïtat de la variància	33
3.1	Proves estadístiques	35
3.1.1	Comparació entre grups de la classe	35
3.1.2	Correlació entre les variables seleccionades	36
3.1.3	Creació de noves variables	38
3.1.4	Selecció de variables significatives	39
3.2	Resolució del problema, model amb regressió logística	40

List of Figures

1	Symptoms	24
2	Symptoms	25

List of Tables

1	Variables amb NA	8
2	Variables amb una difència significativa de NA entre els grups de la classe	9
3	Variables que no segueixen una distribució normal	34
4	Homogeneïtat de la varinça a partir del test de Levene.	34
5	Variables categòriques amb $p < 0.10$ entre la classe	35
6	Variables numèriques amb $p < 0.10$ entre la classe	36
7	Correlació entre variables, valors entre -1 i -0.8 o 0.8 i 1.	37

1 Detalls de l'activitat

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 d'aquestes.

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 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.

2 Resolució

2.1 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 depèn de l'estadi tumoral sinó que també depèn 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] Simptomàtic
- **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] Estadi 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 varis esofàgiques 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):** Nivells 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

2.2 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 probabilitat de sobreviure, i en canvi, optant per teràpies pal·liatives o de confort per pacients amb pitjor pronòstic

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

En el nostre cas, només tenim una base de dades, pel que no farem cap integració amb altres fonts i seleccionarem totes les dades a analitzar, ja que a priori totes en són importants. Quan fem l'anàlisi de les dades ja seleccionarem aquelles que considerem més importants per tal de donar resposta a l'objectiu d'aquest anàlisi.

2.3.1 Preparació del dataset

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

# 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

#Mirar de quin tipus son les variables carregades
tipVar <- c()
for (i in 1:ncol(hcc)) tipVar <- c(tipVar,is(hcc[,i])[1])
tipVar <- table(tipVar)
tipVar

## tipVar
## integer numeric
##      32      18
```

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

2.3.2 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. Es 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)
```

2.4 Neteja de les dades

En aquest apartat ens encarregarem de determinar la presència de valors nuls o buits i com els tractem.

2.4.1 Valors nuls o buits

La distribució de valors desconeguts per cada pacient és:

```
#Valors nuls per pacient
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 valors buits per variable és:

```
#Funció que calcula percentatge
funNA <- function(a, n){
  a = round(100*a/n,1)
}

#Calcul dels valors NA per variable
totNA <- hcc %>% dplyr::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`)

#Taula en dos columnes
tau <- cbind(tauNA[1:25,], tauNA[26:50,])

#Taula amb resultats
kable(x = tau, format = "latex", caption = "Variables amb NA",
      booktabs = TRUE, linesep = '') %>%
  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(Clase = i, p_value = test$p.value,
                  prob_0=(casos/long)[1], prob_1=(casos/long)[2],
                  numNA_0=casos[1], numNA_1=casos[2]))
  }
}
```

```

}

#Filtrar aquelles que rebutjen l'hip nul·la
testSig <- probTest %>% filter(p_value <= 0.05) %>%
  mutate_at(.vars = c("p_value", "prob_0", "prob_1", "numNA_0", "numNA_1"), as.numeric)

#Presentació de resultats
kable(x = testSig, format = "latex",
      caption = "Variables amb una difència significativa de NA entre els grups de la classe",
      booktabs = TRUE, digits = 4, linesep = '') %>%
  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 succeïx a les variables ‘Hemochromatosis’ i ‘Esophageal Varices’. Els pacients afectats es registren i probablement presenten tasses més altes de mortalitat. En canvi, molts pacients es desconixerà si presenten hemocromatosis o varius, però probablement no la patiran, i tindran tasses de supervivència superior.

Corregir aquests 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.

2.4.2 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 standars, k = 5
hcc <- kNN(hcc, variable = factorNames) %>% subset(select = Gender:`Class Attribute`)

```

2.4.3 Correcció valors o extrems de variables quantitatives

Existeixen dues variables amb valors estranys, incompatibles amb la vida; són `Leukocytes` i `Platelets`. La gran majoria dels valors a la variable `Leukocytes` estan per sota de 100, que és l'esperat. Valors majors són 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'aquests valors.

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

Es corregeix els errors i s'assigna la mediana 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])  
}
```

Per a la resta de valors, com es veurà a continuació en l'anàlisi de les variables numèriques, els valors extrems o outliers, els considerarem factibles, ja que biològicament tenen sentit i no farem cap tractament especial més.

2.4.4 Exportació del dataset netejat

Una vegada hem fet la neteja de la nostra base de dades, la guardarem en format csv.

```
# Exportación de les dades netes en .csv  
hcc_clean <- hcc  
write.csv(hcc_clean, file = "../data/hcc_data_clean.csv")
```

2.5 Anàlisi de les dades

2.5.1 Variables numèriques

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

```
# Estudi distribució variables Numeriques  
ind_CA <- which(colnames(hcc) == "Class Attribute")  
  
# emparejar els grafics de 2 en 2. Si son senars, deixar una sola  
# grafic 1 - histograma  
# grafic 2 - boxplot  
ncols <- length(hcc_num)  
if (ncols%%2 == 1){  
  last_col = ncols  
} else{  
  last_col = ncols + 1
```

```

}

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 = "Frecuencia", x =name_var[1] ) +
      theme(legend.position = "bottom")

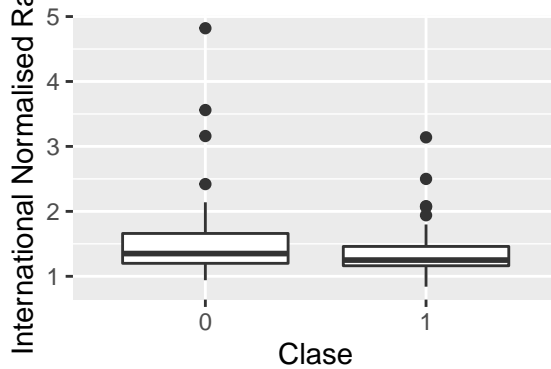
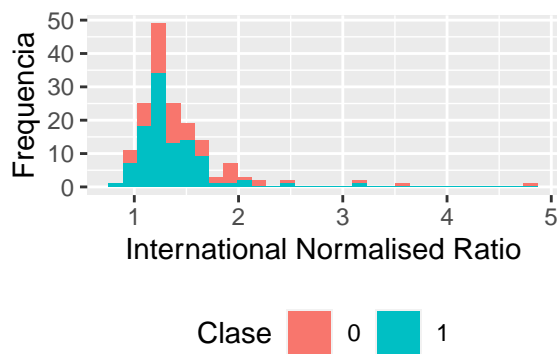
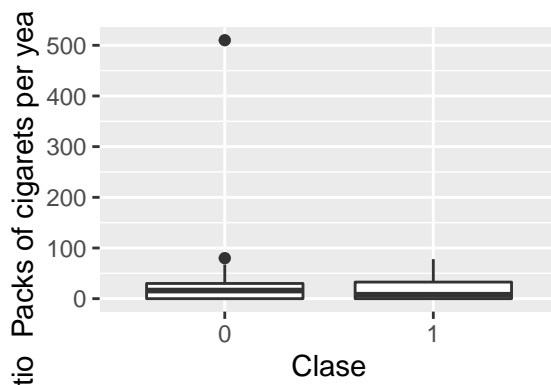
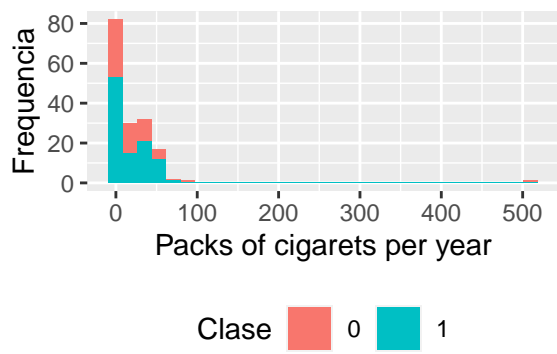
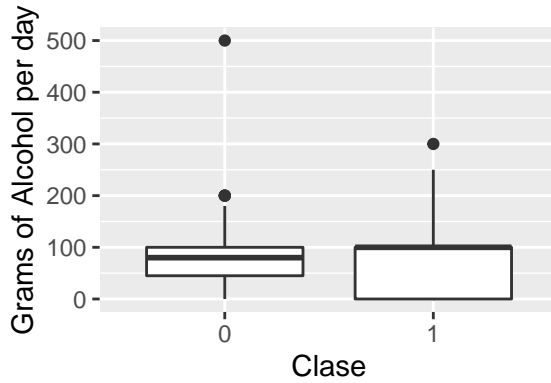
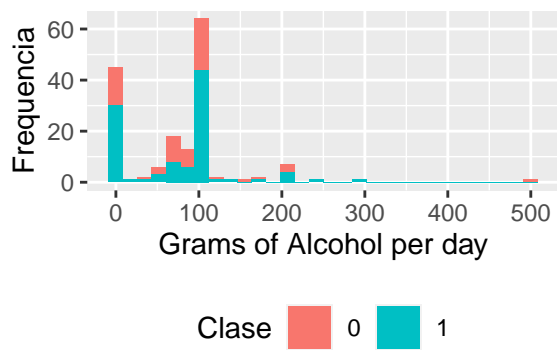
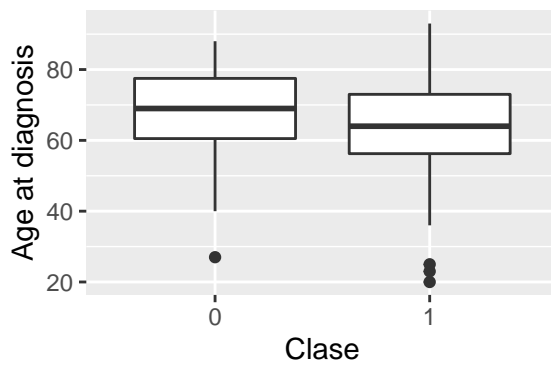
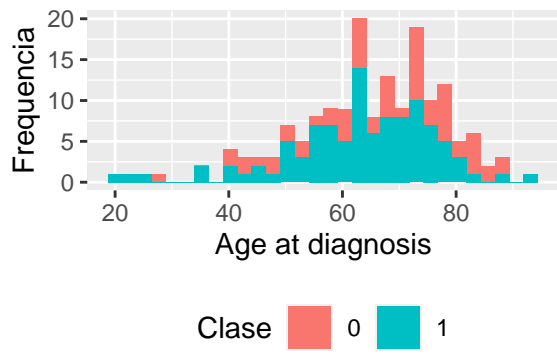
    a2<-data %>%
      ggplot(aes(x=`Class Attribute`,y=data[,1])) +
      geom_boxplot() +
      labs(x = "Clase", y = name_var[1])

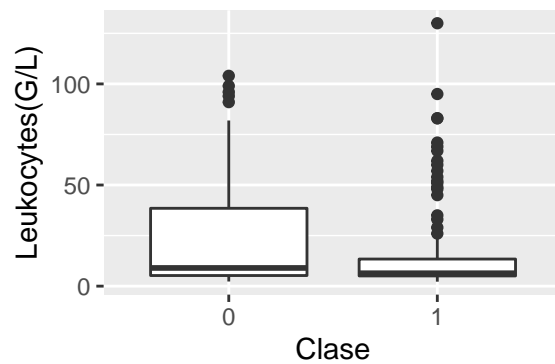
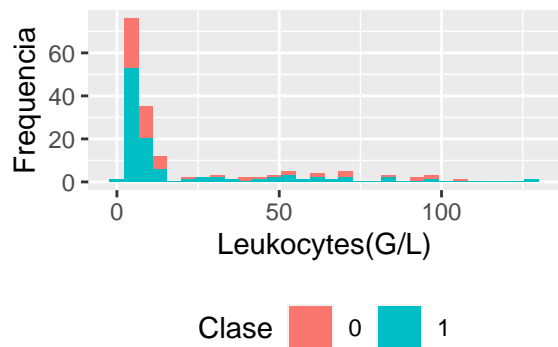
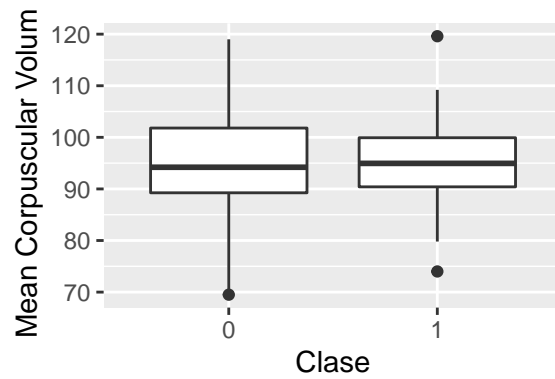
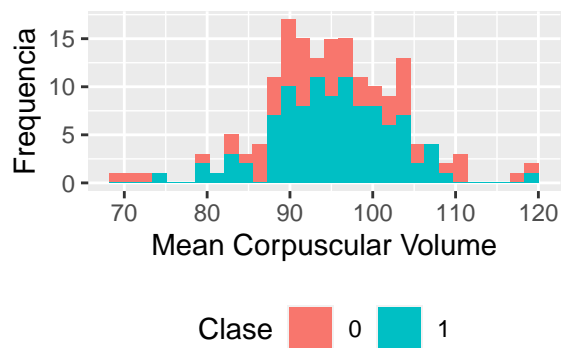
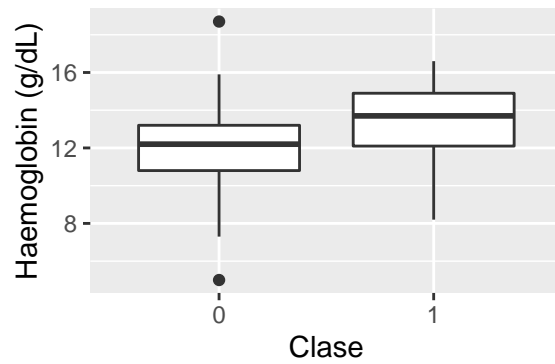
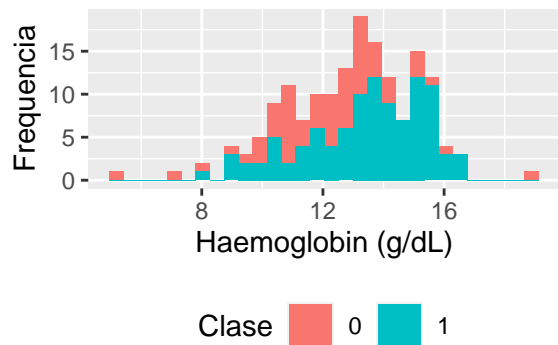
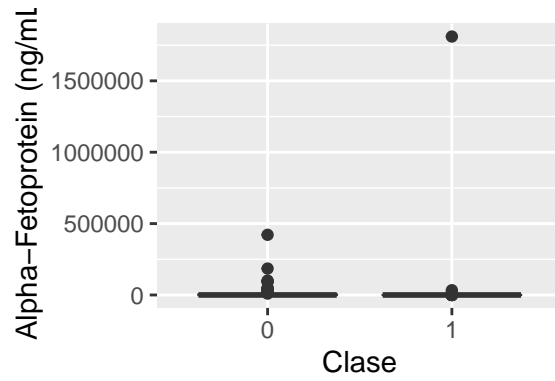
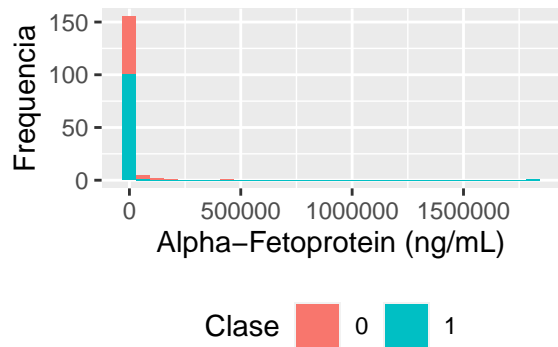
    if ( i != last_col){
      a3<-data %>%
        ggplot(aes(x=data[,2], fill=`Class Attribute`))+
        geom_histogram() +
        labs(fill="Clase", y = "Frecuencia", x =name_var[2] ) +
        theme(legend.position = "bottom")

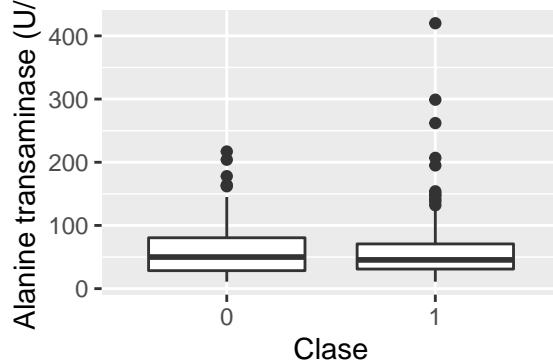
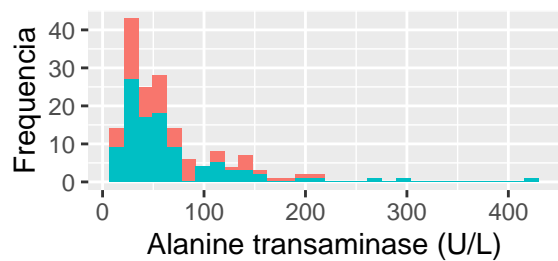
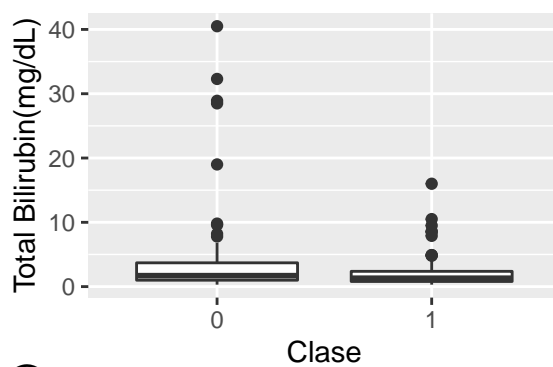
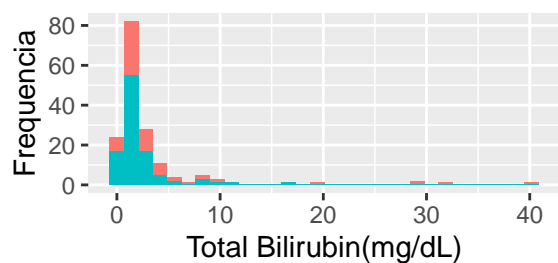
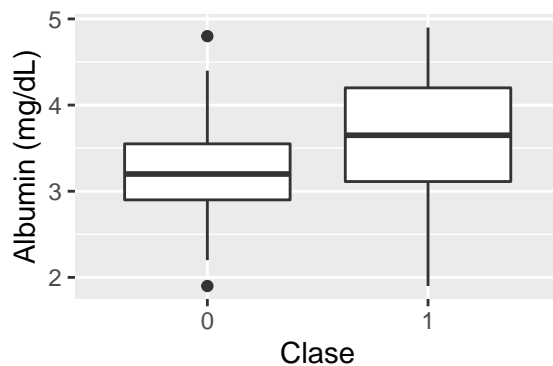
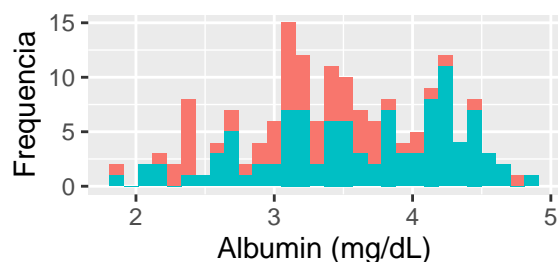
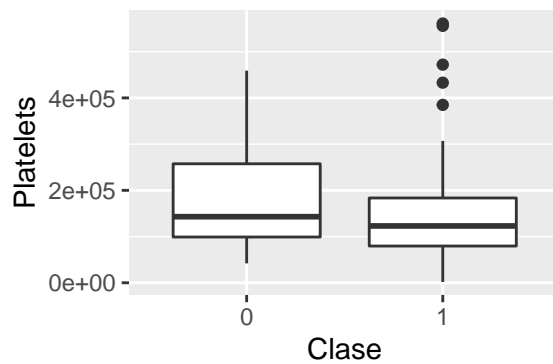
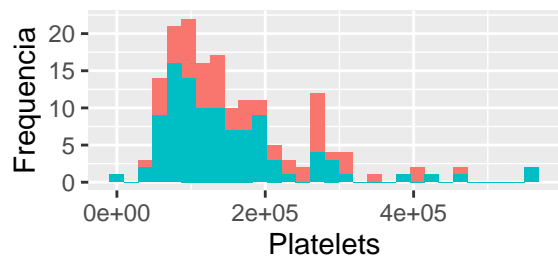
      a4<-data %>%
        ggplot(aes(x=`Class Attribute`,y=data[,2])) +
        geom_boxplot() +
        labs(x = "Clase", y = name_var[2])

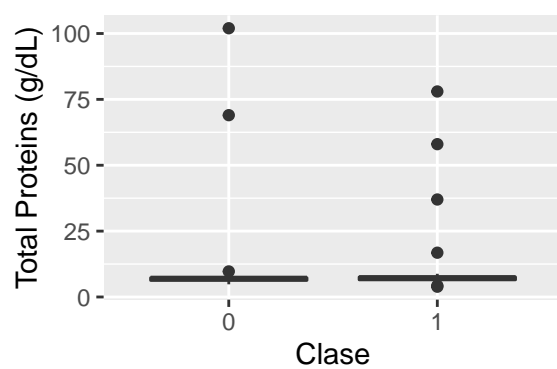
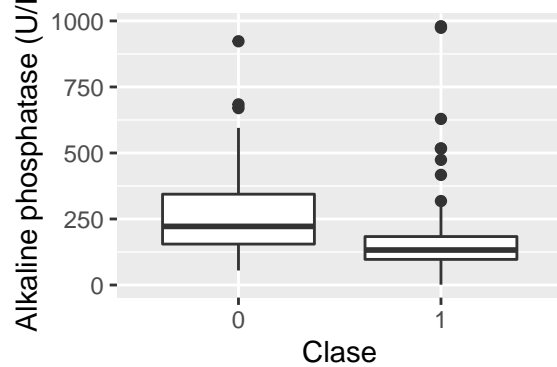
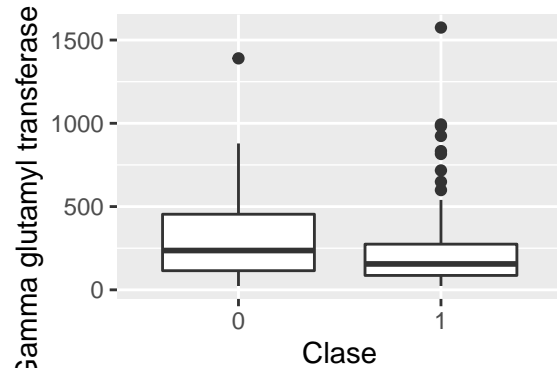
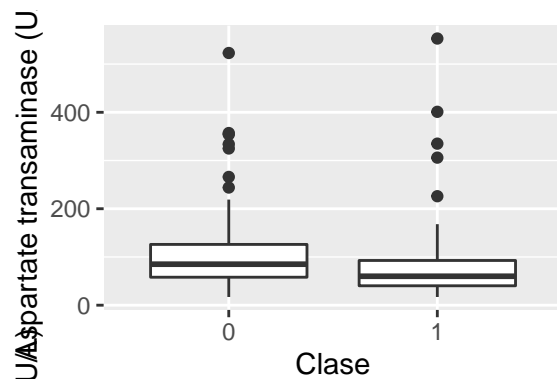
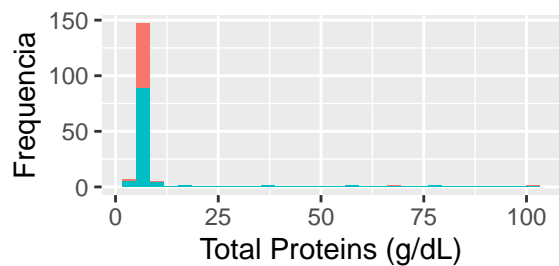
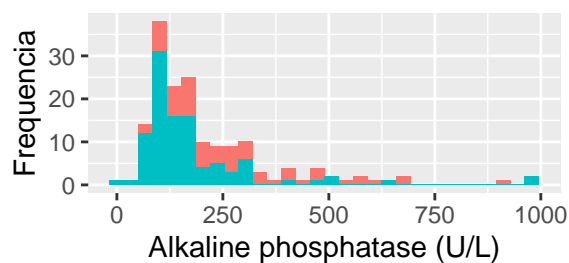
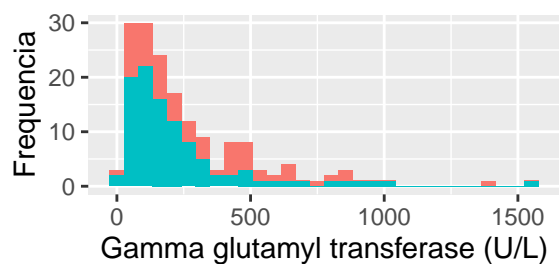
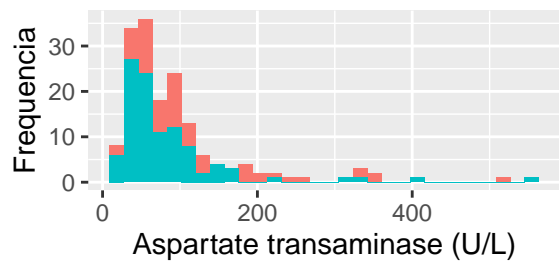
      grid.arrange(a1,a2,a3,a4,nrow=2)
    } else{
      grid.arrange(a1,a2,nrow=1 )
    }
  }
}

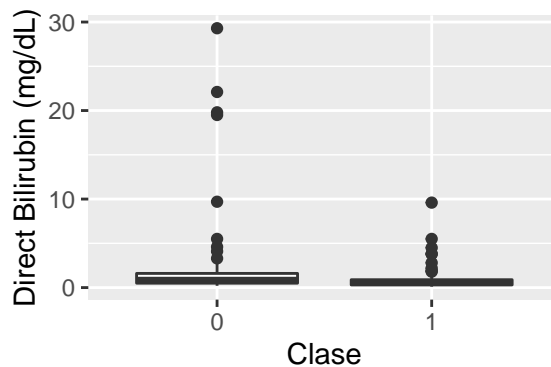
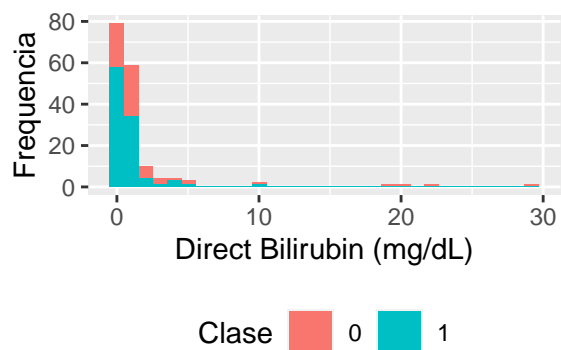
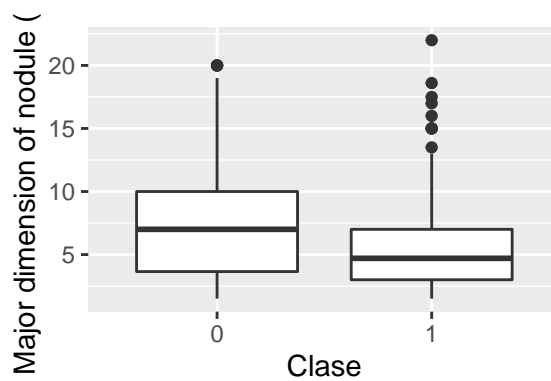
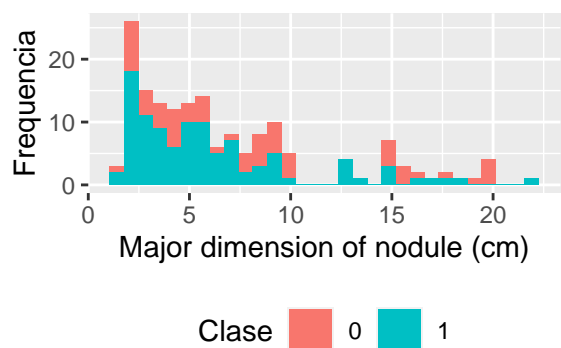
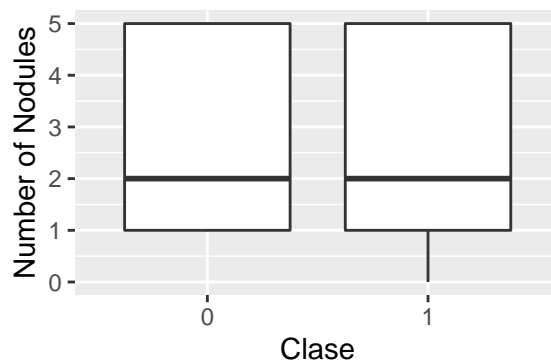
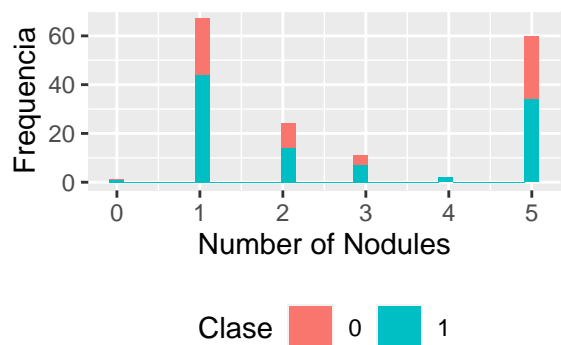
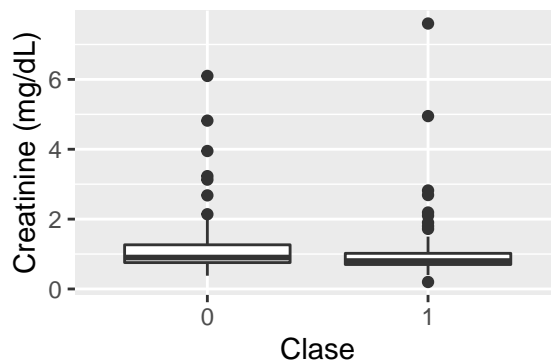
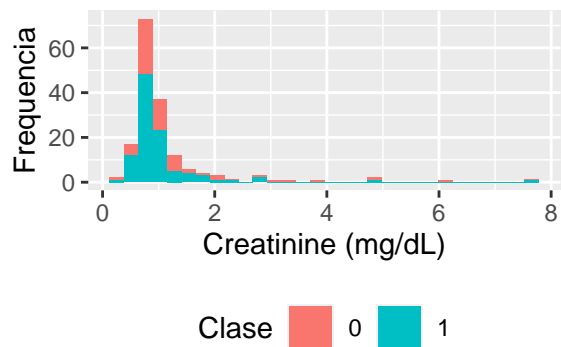
```

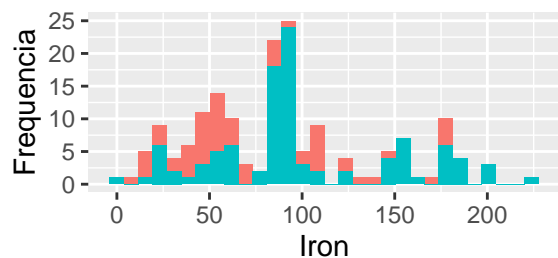




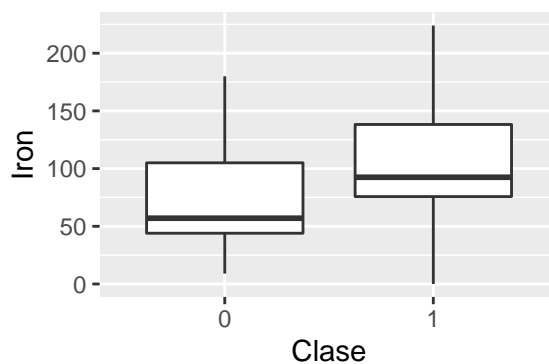




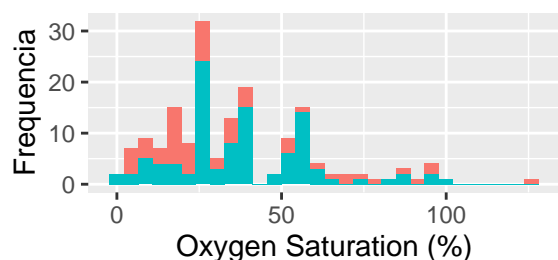




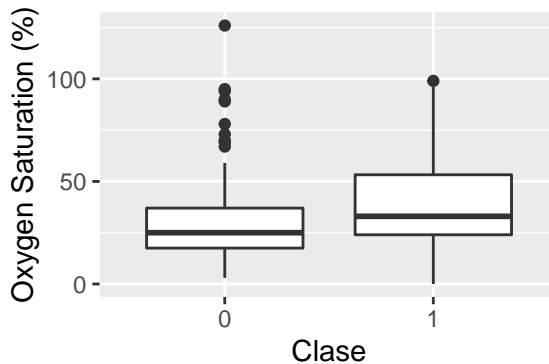
Clase 0 1



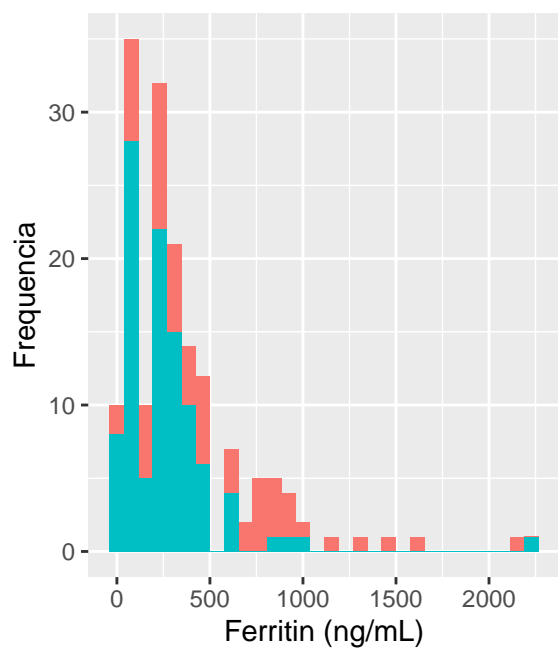
Clase



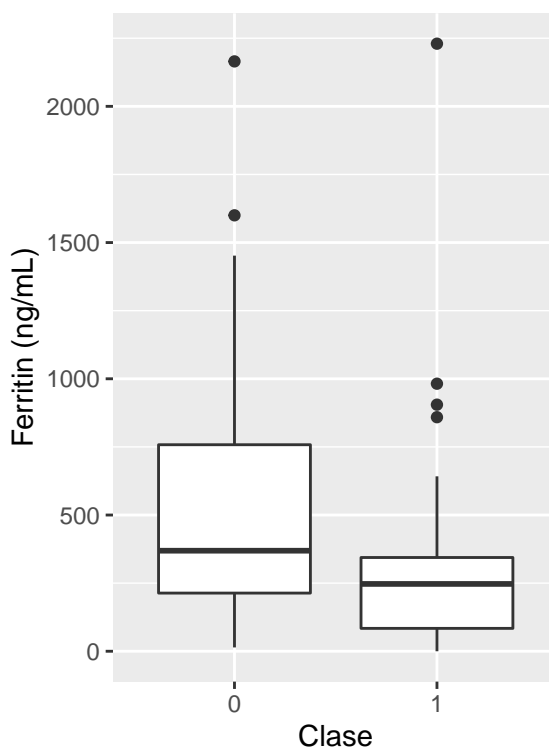
Clase 0 1



Clase



Clase 0 1



Clase

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

- 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)

#emparejar els grafics de 2 en 2. Si son senars, deixar una sola
# grafic 1 - histograma
# grafic 2 - boxplot
ncols <- length(hcc_log)
if (ncols%%2 == 1){
  last_col = ncols
} else{
  last_col = ncols + 1
}

for (i in 1:ncols){
  #transformacio logaritmica
  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 = "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 %>%

```

```

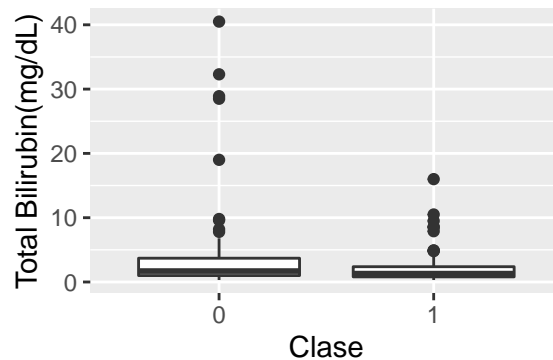
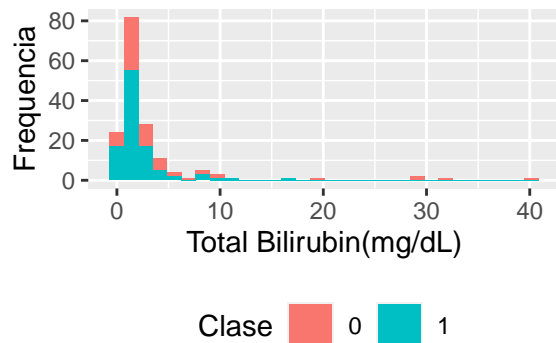
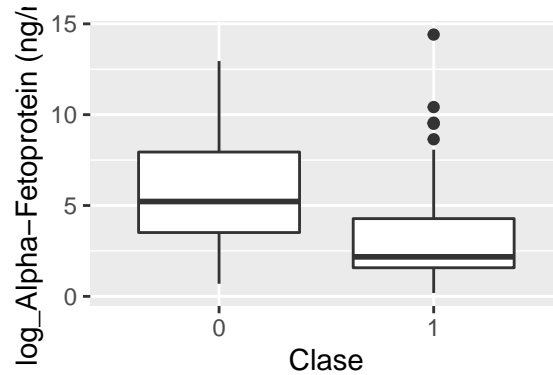
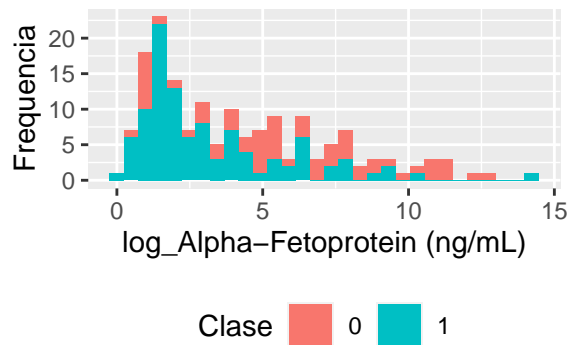
ggplot(aes(x=data[,2], fill=`Class Attribute`))+
  geom_histogram() +
  labs(fill="Clase", y = "Frecuencia", x =name_var[2] ) +
  theme(legend.position = "bottom")

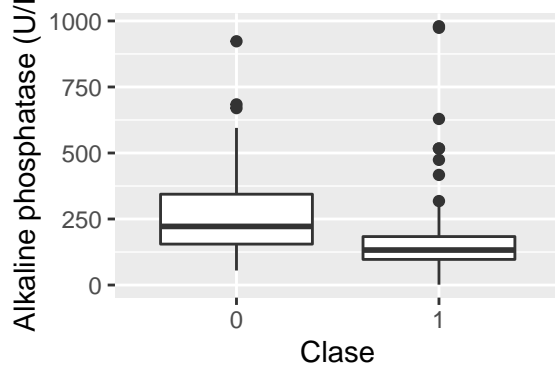
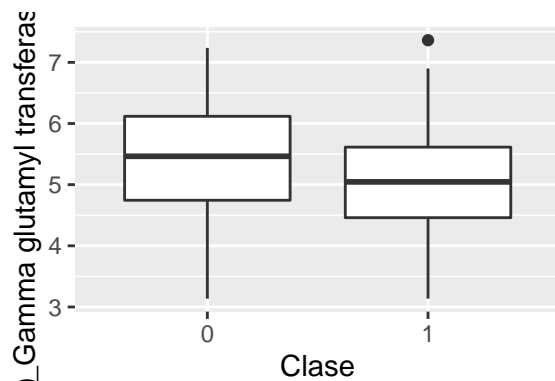
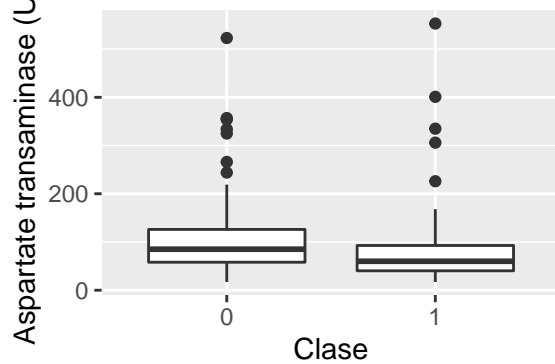
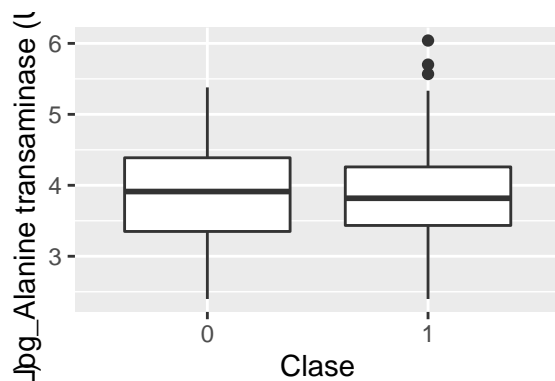
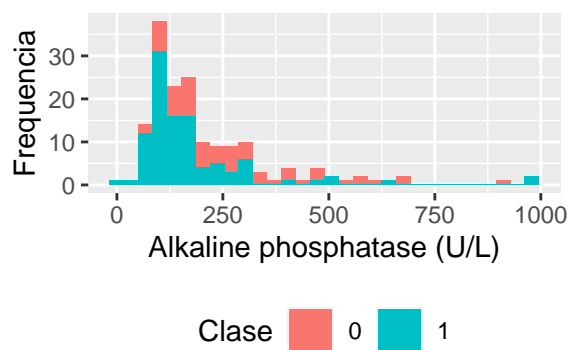
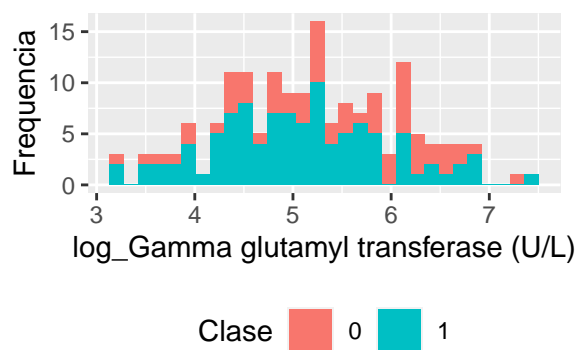
a4<-data %>%
  ggplot(aes(x=`Class Attribute`,y=data[,2])) +
  geom_boxplot() +
  labs(x = "Clase", y = name_var[2])

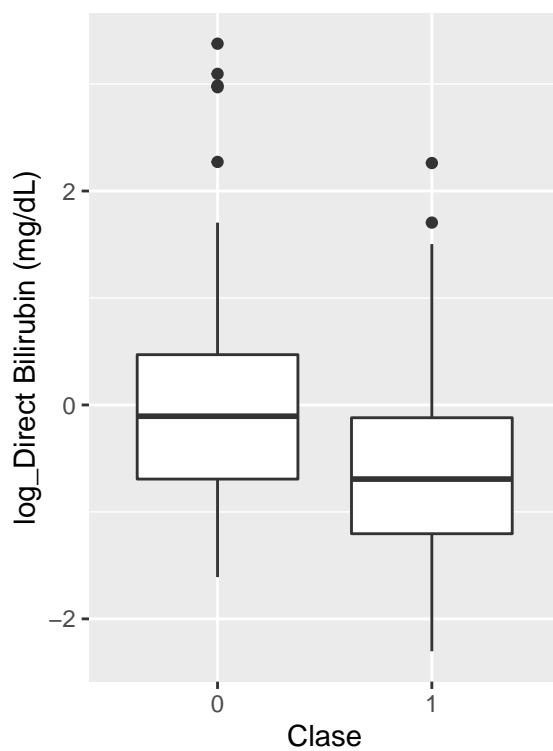
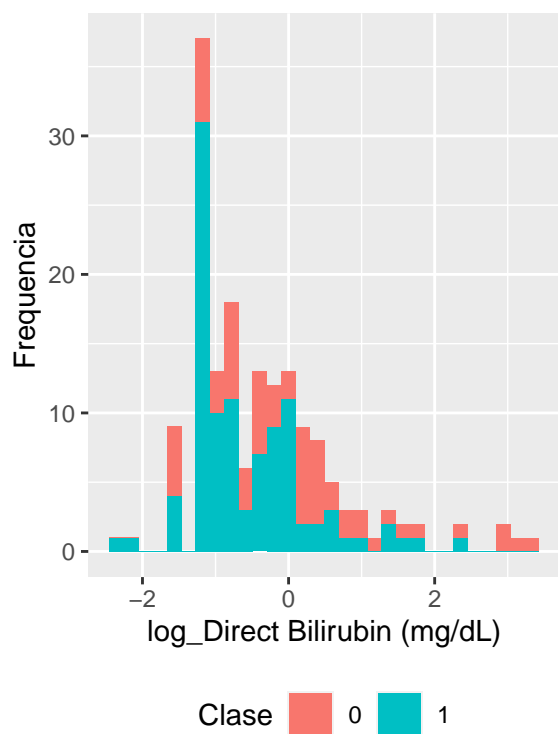
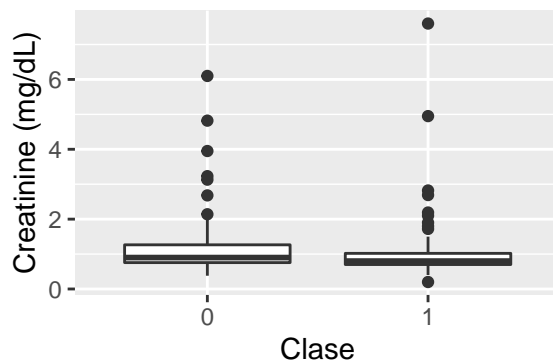
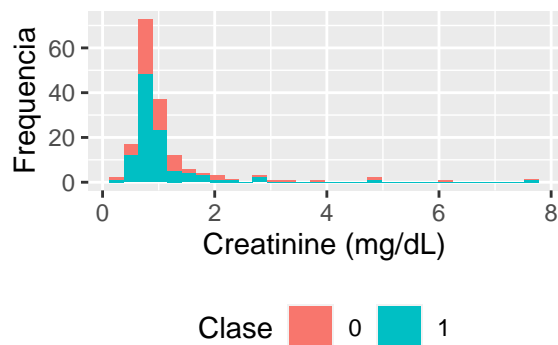
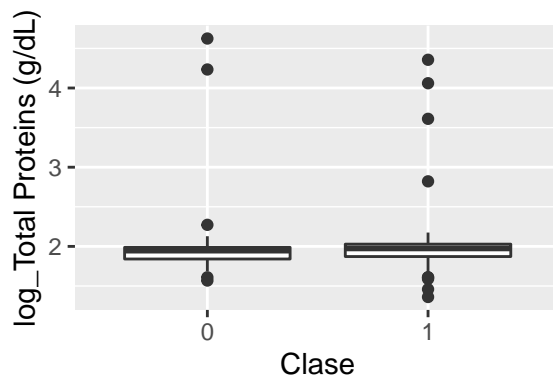
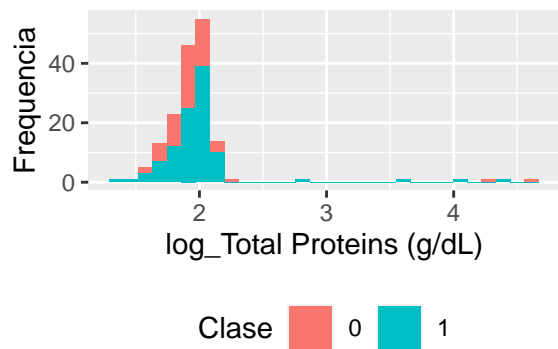
grid.arrange(a1,a2,a3,a4,nrow=2)

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

```







2.5.2 Variables qualitatives

Fem una primera exploració visual d'aquest tipus de variable:

3 Versio Prova

```
captions <- c("Caption 1",
             "Caption 2",
             "Caption 3",
             "Caption 4: a very very very very very very very very very long one")

hcc_factorT<-c(1:2)

listOfPlots <- list()
name_plot <- names(hcc)[hcc_factorT]
name_plot2 <- c()
for( i in 1:length(name_plot)){
  np <- c(paste(name_plot[i], "Frequencia", sep = ' '),paste(name_plot[i], "Proporcio", sep = ' '))
  name_plot2 <- c(name_plot2, np)
}

#hcc_factorT<-c(1:23,27:29)
hcc_factorT<-c(1:2)

listOfPlots <- list()
name_plot <- names(hcc)[hcc_factorT]

#emparejar els grafics de 2 en 2. Si son senars, deixar una sola
# grafic 1 - diagrama barres
# grafic 2 - diagrama barres proporció
for (i in 1:length(hcc_factorT) ) {
  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 = "Frequencia") +
    theme(legend.position = "bottom",
          plot.caption = element_text(hjust = 0.5))

  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",
          plot.caption = element_text(hjust = 0.5))

  print(a1)
  print(a2)

  cat('\n\n')

  # grid.arrange(a1, a2, ncol=2,
  #              top = textGrob("Figure 2: this time with no bottom caption",
```

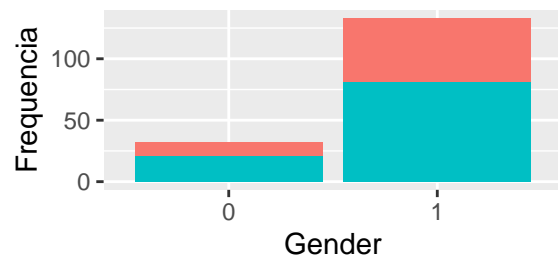


```

#           x = 0, y = 0.5,
#           just = "left",
#           gp = gpar(fontsize = 14)))

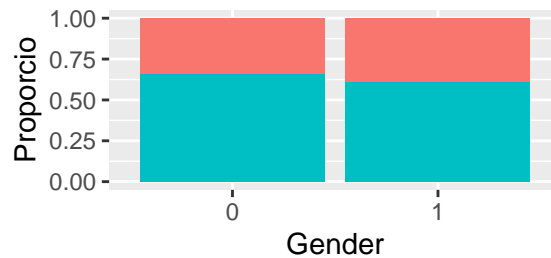
name1 <- paste(name_var[1],1, sep = "")
name2 <- paste(name_var[1],2, sep = "")
listOfPlots[[name1]] <- a1
listOfPlots[[name2]] <- a2
}

```



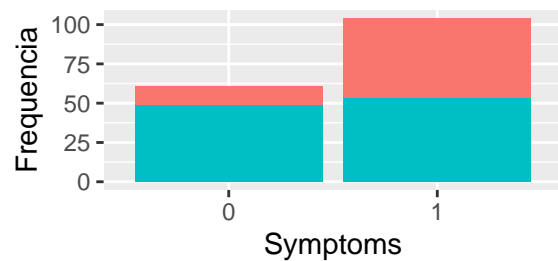
Clase ■ 0 ■ 1

(a) Gender Frecuencia



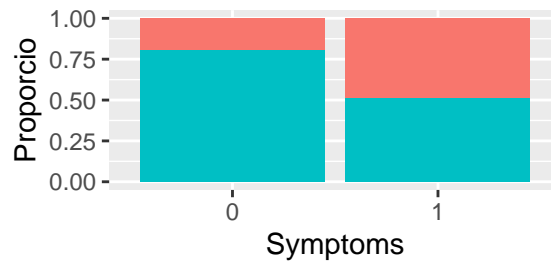
Clase ■ 0 ■ 1

(b) Gender Proporcio



Clase ■ 0 ■ 1

(c) Symptoms Frecuencia



Clase ■ 0 ■ 1

(d) Symptoms Proporcio

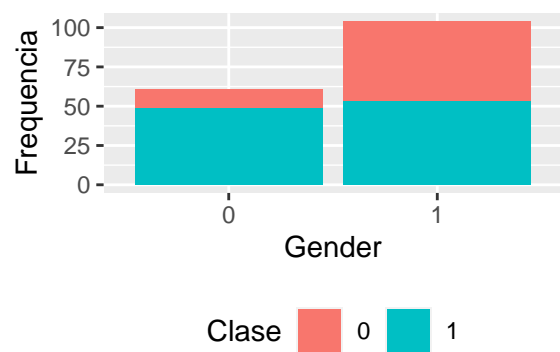
Figure 1: Symptoms

```

len <- length(listOfPlots)/2

for(i in 1:len){
  print(listOfPlots[[(2*i-1)]])
  print(listOfPlots[[(2*i)]]])
  cat('\n\n')
}

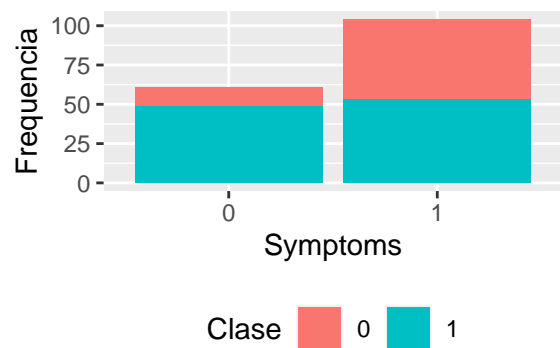
```



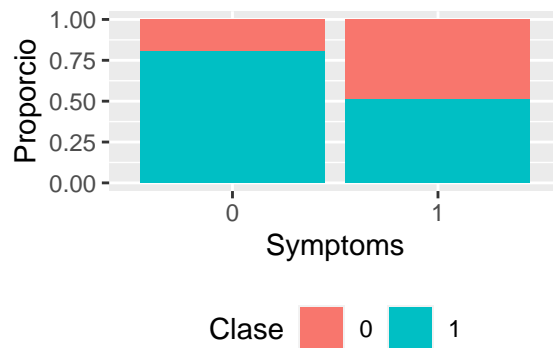
(a) Gender Frecuencia



(b) Gender Proporcio



(c) Symptoms Frecuencia



(d) Symptoms Proporcio

Figure 2: Symptoms

```

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

#emparejar els grafics de 2 en 2. Si son senars, deixar una sola
# grafic 1 - diagrama barres
# grafic 2 - diagrama barres proporció
ncols <- length(hcc_factorT)
if (ncols%%2 == 1){
  last_col = ncols
} else{
  last_col = ncols + 1
}

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)

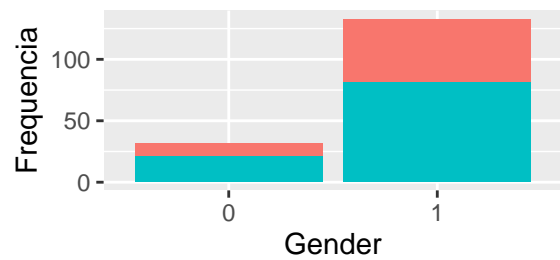
    a1<-data %>%
      ggplot(aes(x=data[,1],fill=`Class Attribute`))+
      geom_bar() +
      labs(fill="Clase", x = name_var[1], y = "Frecuencia") +
      theme(legend.position = "bottom")

    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")

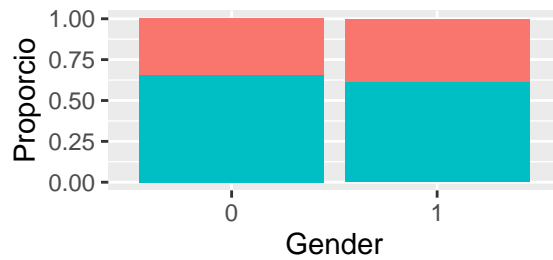
      a4<- data %>%
        ggplot(aes(x=data[,2],fill=`Class Attribute`))+
        geom_bar(position = "fill") +
        labs(fill="Clase", x = name_var[2], y = "Proporcio") +
        theme(legend.position = "bottom")

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

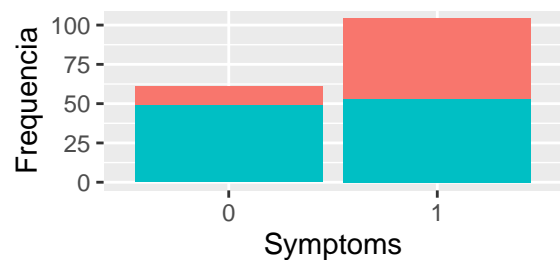
```



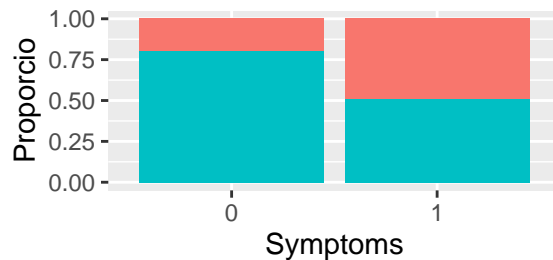
Clase 0 1



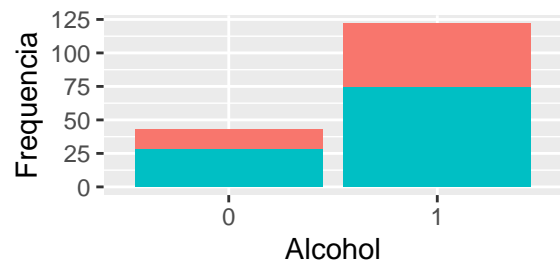
Clase 0 1



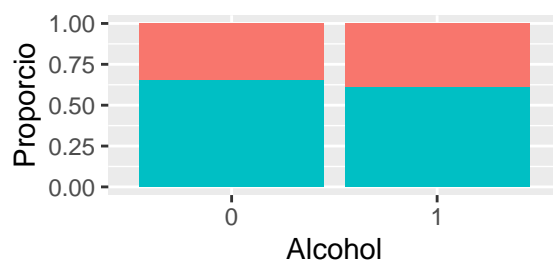
Clase 0 1



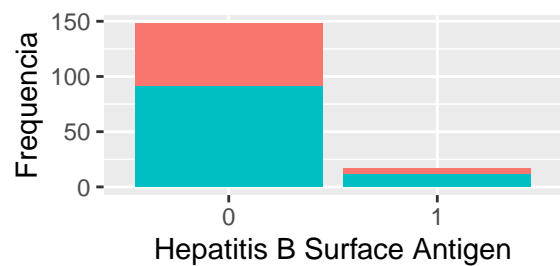
Clase 0 1



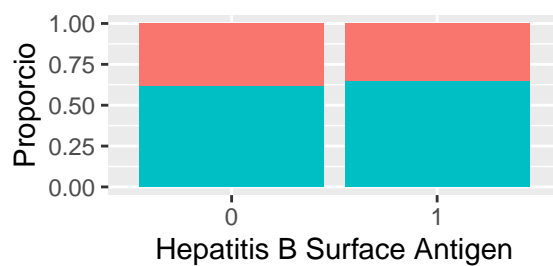
Clase 0 1



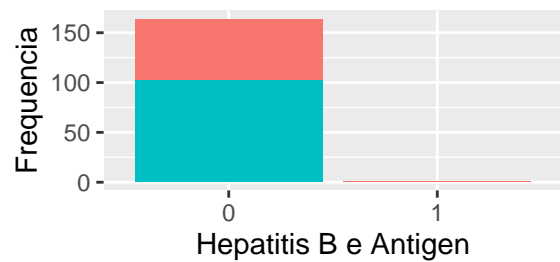
Clase 0 1



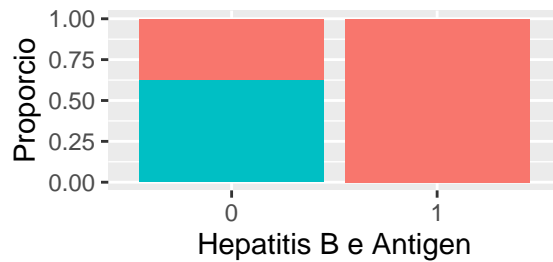
Clase 0 1



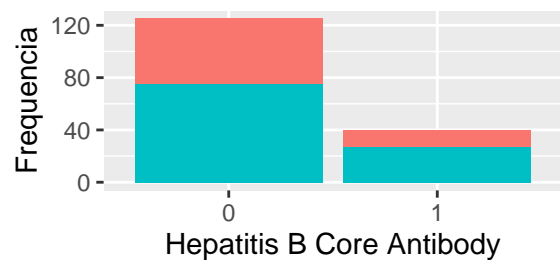
Clase 0 1



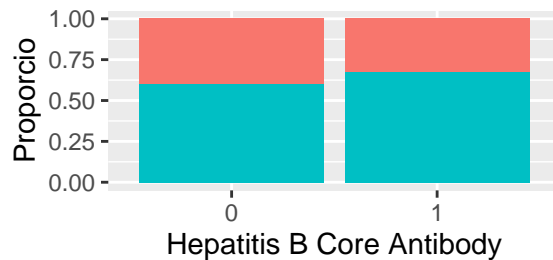
Clase 0 1



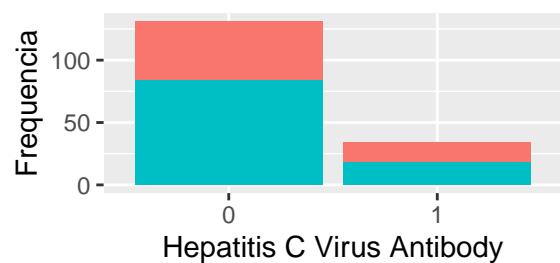
Clase 0 1



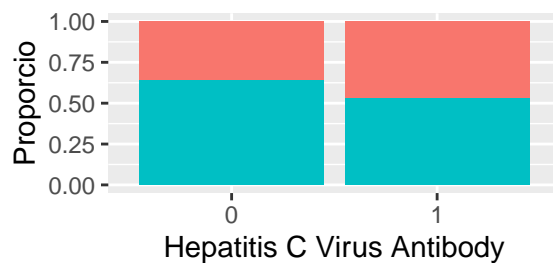
Clase 0 1



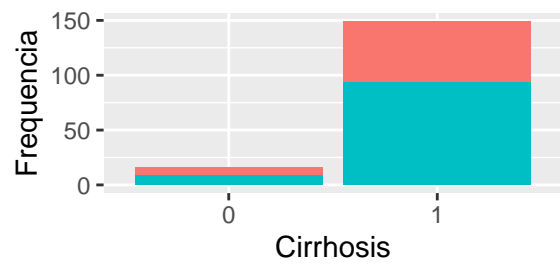
Clase 0 1



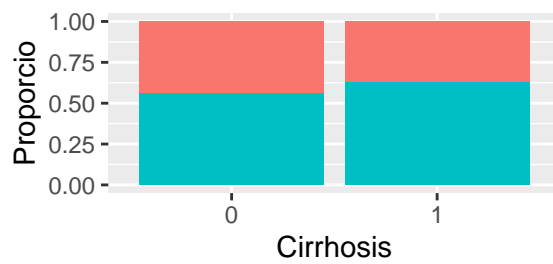
Clase 0 1



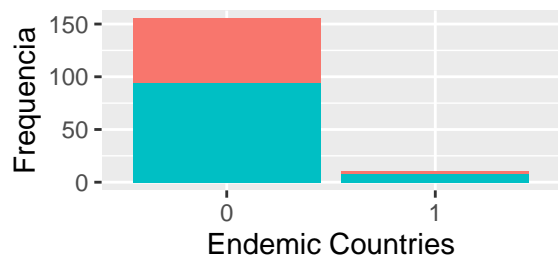
Clase 0 1



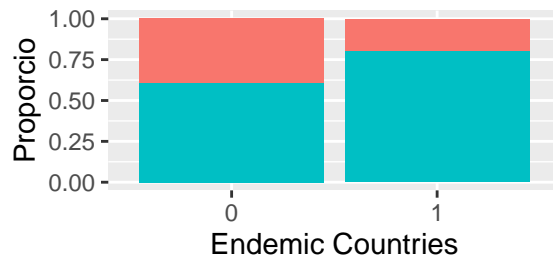
Clase 0 1



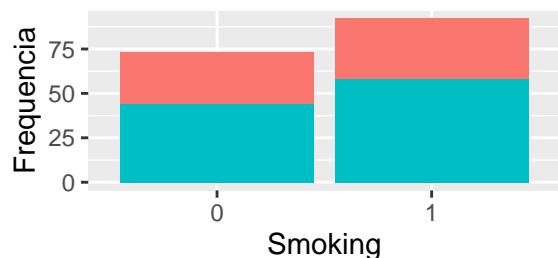
Clase 0 1



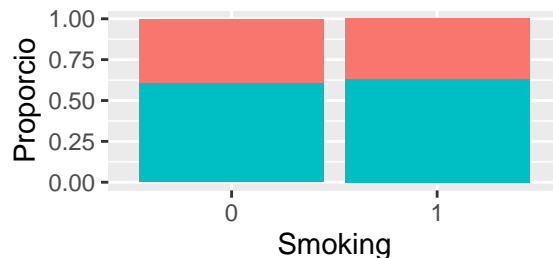
Clase 0 1



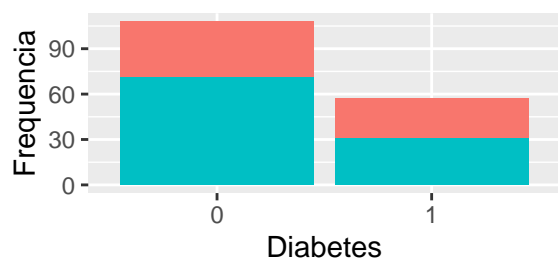
Clase 0 1



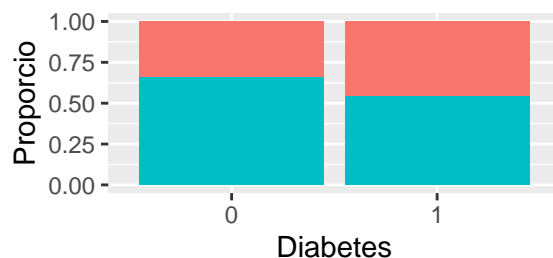
Clase 0 1



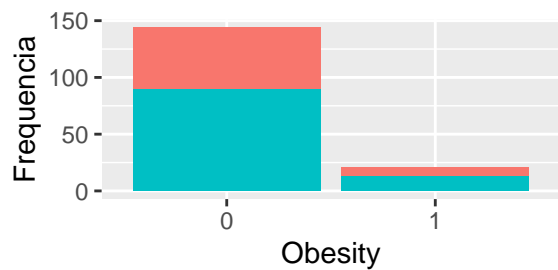
Clase 0 1



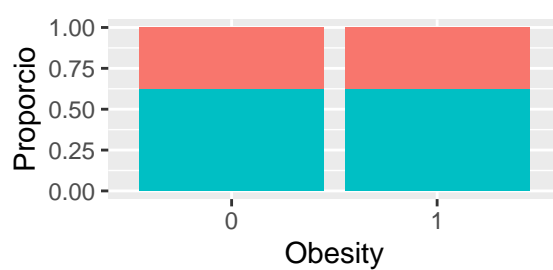
Clase 0 1



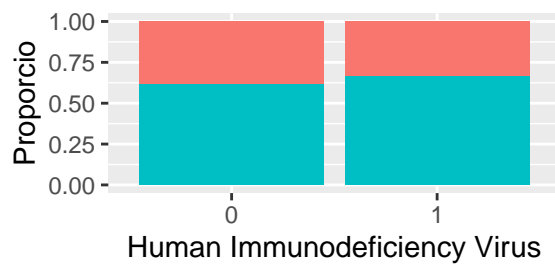
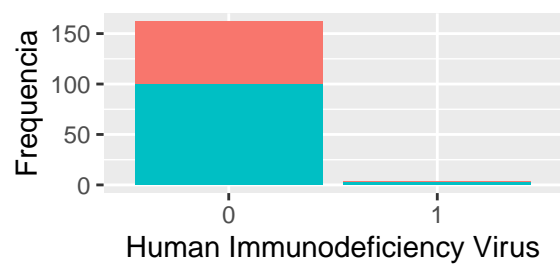
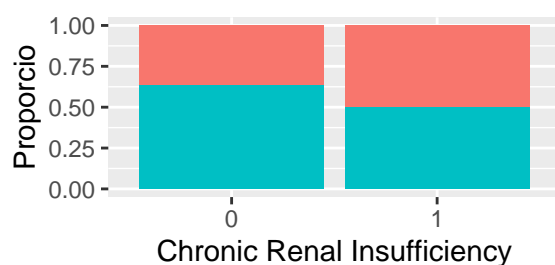
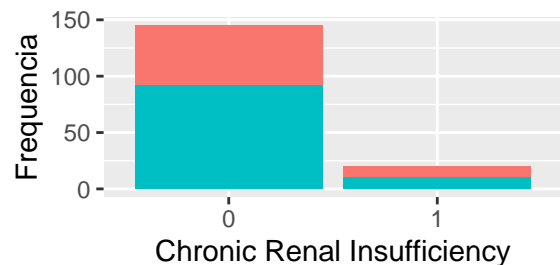
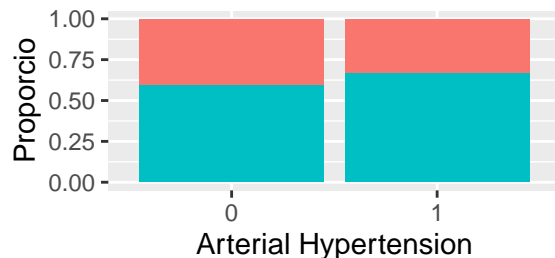
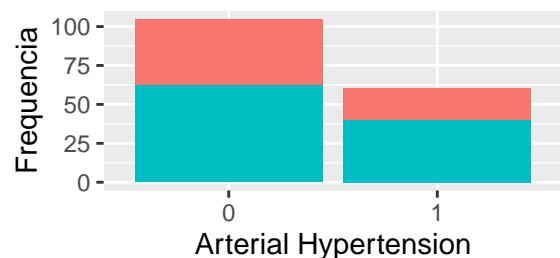
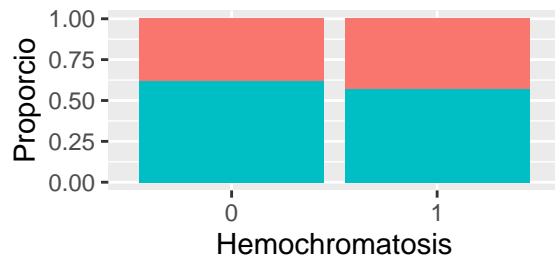
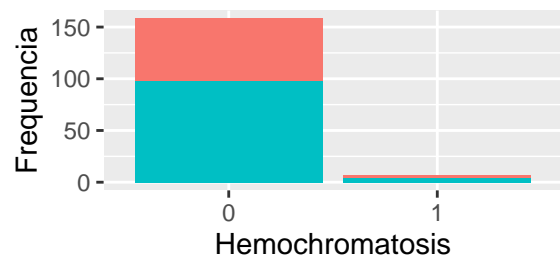
Clase 0 1

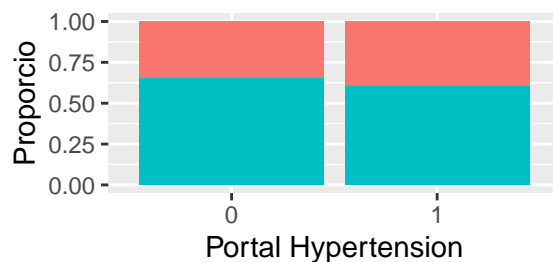
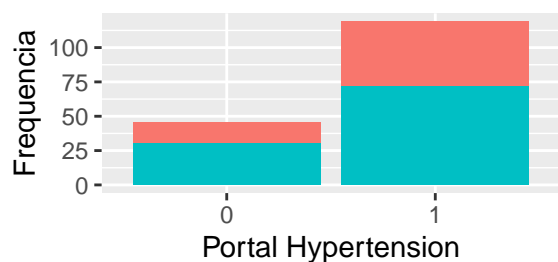
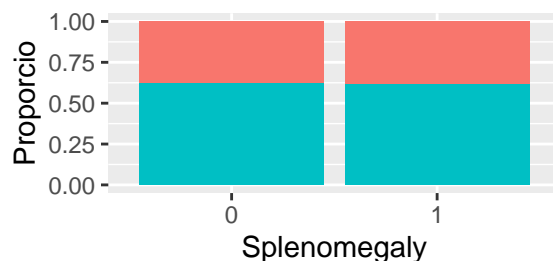
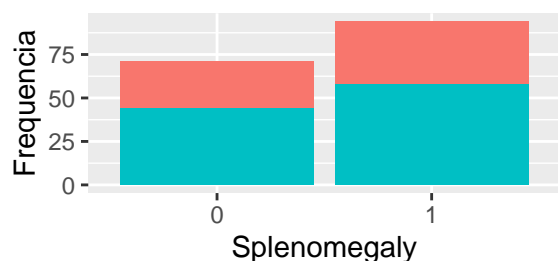
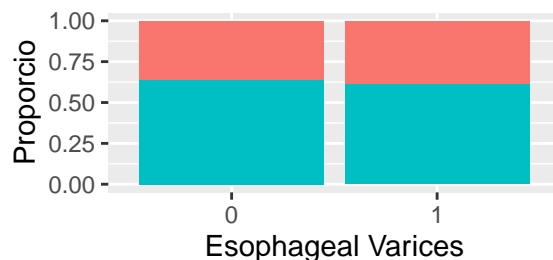
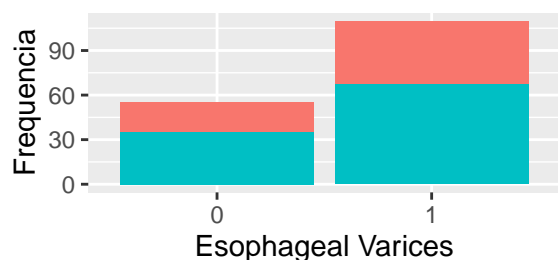
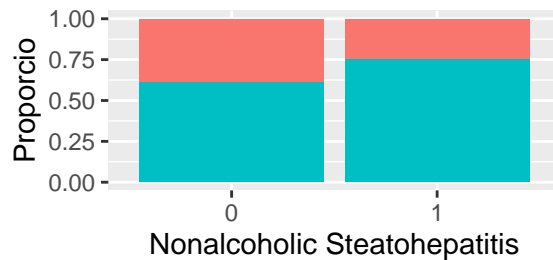
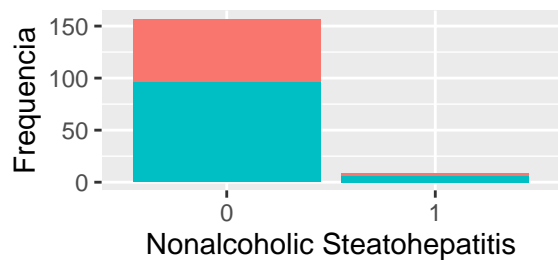


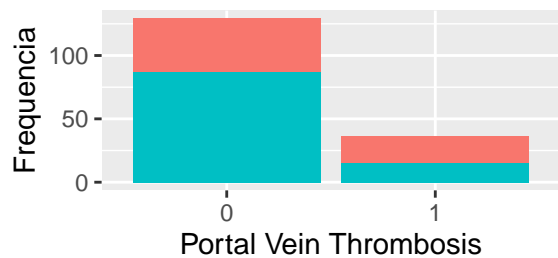
Clase 0 1



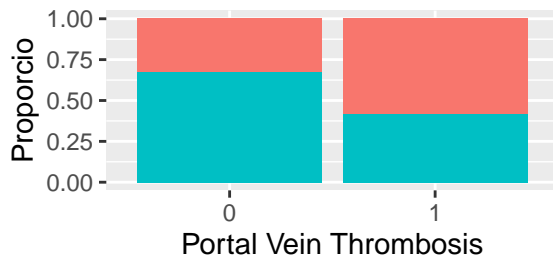
Clase 0 1



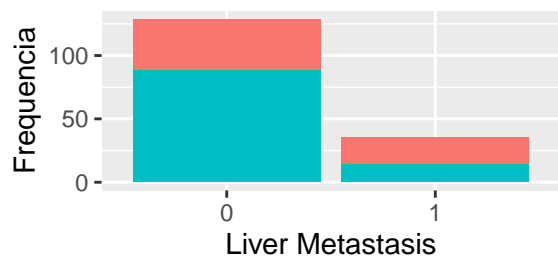




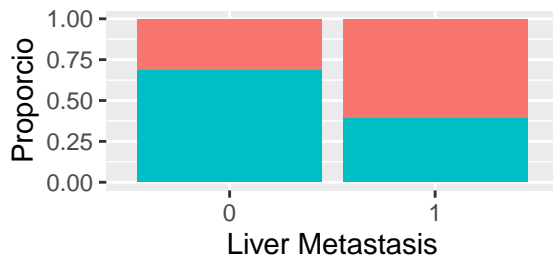
Clase 0 1



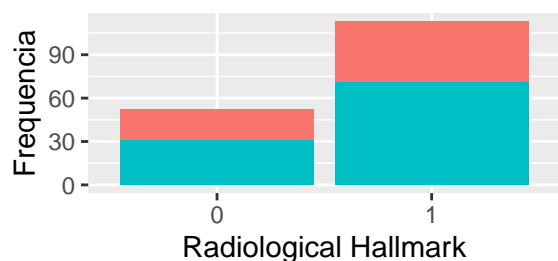
Clase 0 1



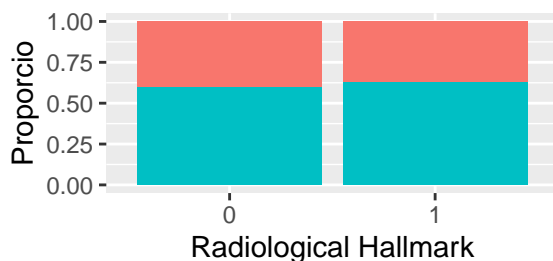
Clase 0 1



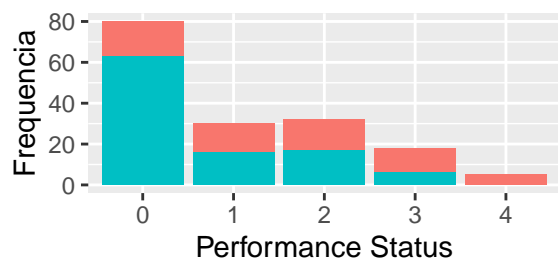
Clase 0 1



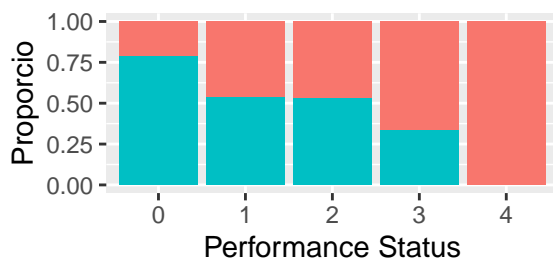
Clase 0 1



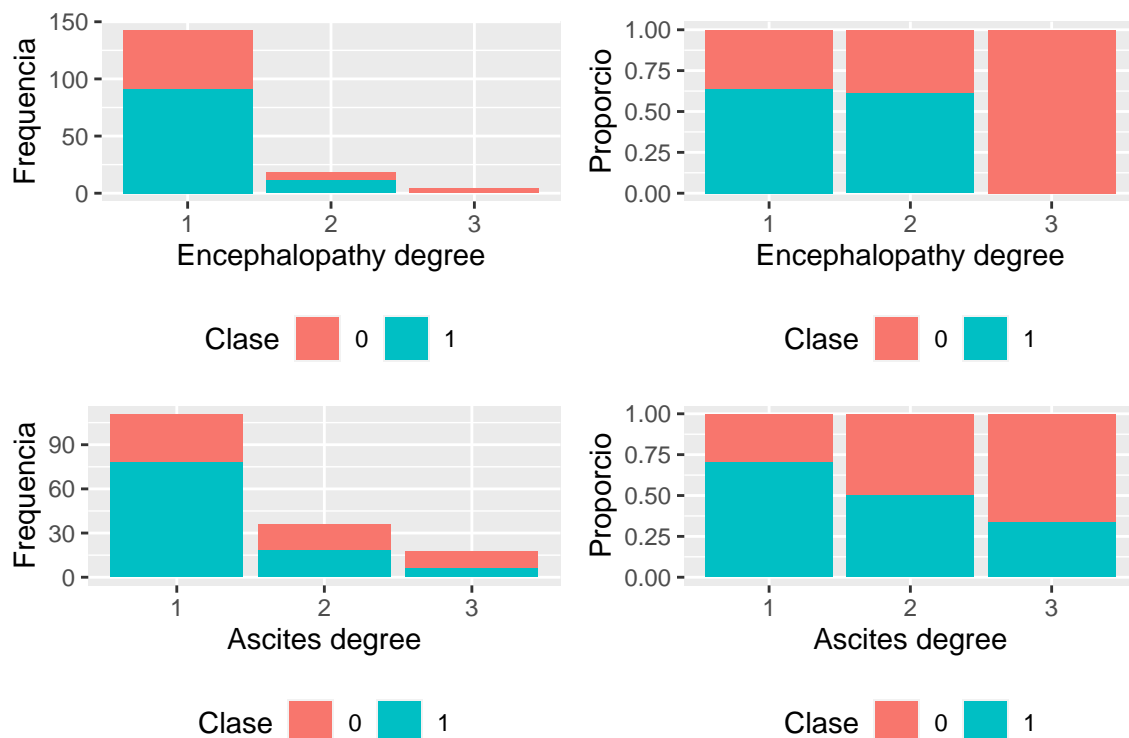
Clase 0 1



Clase 0 1



Clase 0 1



De les variables quantitatives podem veure que entre elles hi ha força diferència en la proporció de persones que sobreviuen segons el valor de la variable. Per exemple, per a la variable Smoking aquesta diferència no és notable, però per Ascites degree sí que ho és. És a dir, podem preveure que sí que hi haurà variables que influiran en la mortalitat d'aquests pacients i d'altres que semblen que no.

3.0.1 Comprobació de la normalitat i homogeneïtat de la variància

Per la comprovació de la distribució normal dels valors quantitius farem servir la prova de normalitat d'Anderson-Darling.

```
#Test Anderson-Darling per a la normalitat
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))
  }
}

#Filtrar les que no rebutgen l'hipotesis de normalitat
var_normales<- tNorm %>% filter(as.numeric(p_value)>0.05) %>% pull(Variable)
tau <- tNorm %>% filter(p_val < 0.05) %>%
  mutate_at(.vars = c("p_value"), as.numeric)
tau <- cbind(tau[1:12,],tau[13:24,])

#Correïó d'ultima fila columna dreta
tau[12,3 ] <- ""
tau[12,4 ] <- 0
```

```
#Presentacio resultats
kable(x = tau, format = "latex", caption = "Variables que no segueixen una distribuci3 normal",
      booktabs = TRUE, digits = 4, linesep = '') %>%
  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 numèriques que es distancien significativament de la distribuci3 normal, pel que es faran servir test no paramètrics (Mann–Whitney–Wilcoxon) per la comparativa amb relació a la supervivència.

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

```
#Test Levene per homogeneïtat de la variància
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))
}

tVar <- tVar %>%mutate_at(.vars = c("p_value"), as.numeric)

#presentacio de resultats
kable(x = tVar, format = "latex", caption = "Homogeneïtat de la varinça a partir del test de Levene.",
      booktabs = TRUE, digits = 4, linesep = '') %>%
  kable_styling(latex_options = c("HOLD_position"))
```

Table 4: Homogeneïtat de la varinça a partir del test de Levene.

Variable	p_value
Haemoglobin (g/dL)	0.9502
Mean Corpuscular Volume	0.0189
Albumin (mg/dL)	0.0217
log_Gamma glutamyl transferase (U/L)	0.4290

Com es pot 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 variàncies entre la població que sobreviu i la que no.

3.1 Proves estadístiques

3.1.1 Comparació entre grups de la classe

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

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

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

#Filtrar les que rebutgen hipotesis nul·la
tau <- testChi %>% filter(p_value < 0.1) %>%
  mutate_at(.vars = c("p_value"), as.numeric)

#Guardar variables significatives
varSig <- tau

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

Table 5: 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

```
#Test Mann-Whitney-Wilcoxon per a variables quantitatives
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, conf.int = 0.95)
```

```

testWil <- testWil %>%
  bind_rows(c(Class = i, Name=names(hcc[i]),Categorica="0",p_value = wil$p.value))
}

#Filtrar les que rebutgen hipotesis nul·la
tau <- testWil %>% filter(p_value < 0.1) %>%
  mutate_at(.vars = c("p_value"), as.numeric)

#Guardar variables significatives
varSig <- varSig %>% bind_rows(tau)

#Resultats en dos columnes
tau <- cbind(tau[1:6,],tau[7:12,])

#Presentacio de resultats
kable(x = tau, format = "latex", caption = "Variables numèriques amb p<0.10 entre la classe",
      booktabs = TRUE, digits = 4, linesep = '') %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))

```

Table 6: Variables numèriques 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 se seleccionarà per a la creació d'un model de regressió logística, però abans, valorarem les correlacions entre elles per tal de seleccionar variables que estiguin poc relacionades entre elles.

3.1.2 Correlació entre les variables seleccionades

Respecte a la correlació entre les variables seleccionades, veiem la seva matriu de correlacions. Donada l'existència de variables categòriques, aquestes és consideraran numèriques i usarem la correlació no paramètrica de Spearman per a la seva valoració. Prèviament es normalitzaran totes les variables quantitatives.

```

# Correlació variables independents.
hcc_norm <- hcc
atr<-names(hcc)

#normalitzacio variables quantitatives
hcc_norm[,hcc_num]<-scale(hcc_norm[,hcc_num])

#variables quantitatives tipus numeric
hcc2<-as.data.frame(lapply(hcc_norm,as.numeric))
names(hcc2)<-atr

#correlacio Spearman + filtrar alta correlació
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 %>%
  dplyr::select(variable, variable2, value) %>%
  rename(corr = value) %>%
  filter(variable != variable2) %>%
  filter(abs(corr) >= 0.8) %>%
  arrange(-corr) %>%
  group_by(corr) %>%
  mutate(vars = paste(variable, collapse = '-')) %>%
  dplyr::select(vars, corr) %>%
  unique()

#Presentació variables altament correlades
kable(x = hcc_corTop, format = "latex",
      caption = "Correlació entre variables, valors entre -1 i -0.8 o 0.8 i 1.",
      booktabs = TRUE, digits = 4, linesep = '') %>%
  kable_styling(latex_options = c("HOLD_position"))

```

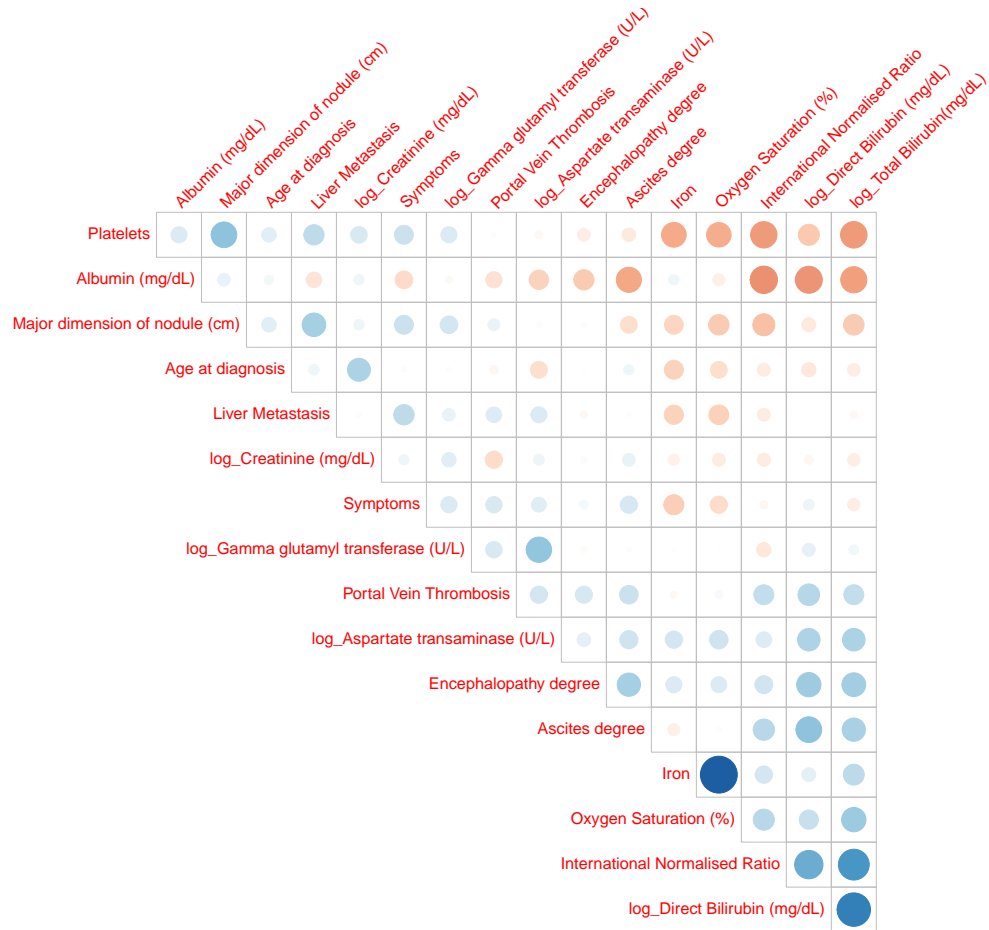
Table 7: Correlació entre variables, valors entre -1 i -0.8 o 0.8 i 1.

vars	corr
Iron-Oxygen Saturation (%)	0.828

```

#Grafica de totes les correlacions
corrplot(hcc_cor, cl.pos='n',tl.srt = 45,
         tl.cex = 1.25, type="upper",method = "circle",
         order="FPC", diag=FALSE)

```



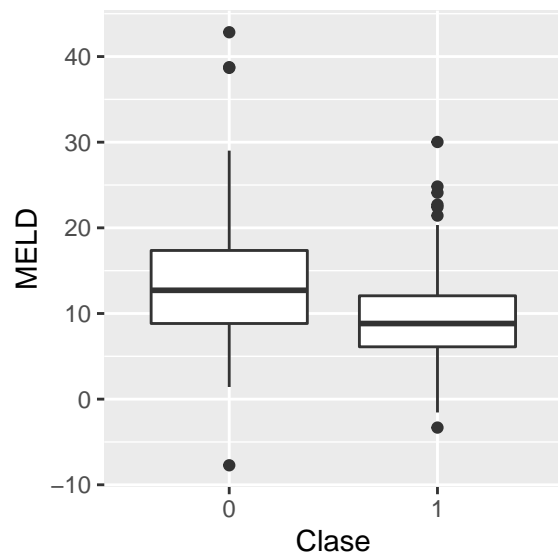
Com es pot veure, només dues variables tenen una molt alta correlació (**Iron** i **Oxygen Saturation (%)**). De les dues, seleccionarem **Iron** que és la que presenta una menor probabilitat de significació.

3.1.3 Creació de noves variables

Hi ha un valor conegut per a valorar la probabilitat de mort al cap de tres mesos de pacients amb hepatopatia que depèn de la creatinina, la bilirubina total i de l'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 tres mesos, no per a l'any, com és 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 )

#Boxplot per veure la distribució repsecte classe
hcc %>%
  ggplot(aes(x=`Class Attribute`,y=MELD)) +
  geom_boxplot() +
  labs(x = "Classe", y = "MELD")
```



```
#Test Mann-Whitney-Wilcoxon
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
```

Es 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 bilirubina total, l'INR i la creatinina.

3.1.4 Selecció de variables significatives

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

3.2 Resolució del problema, 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 numèrica.

```
#Model logit
varElim <- c("Oxygen Saturation (%)", "log_Total Bilirubin(mg/dL)",
            "International Normalised Ratio", "log_Creatinine (mg/dL)")

#Variables seleccionades
var_selecc<- setdiff(varSig$Name, varElim)

#Afegir les que falten
hcc_sel <- hcc[c(var_selecc, "MELD", "Class Attribute")] %>%
  mutate_at(vars("Ascites degree", "Encephalopathy degree"), as.numeric)

#Model logistic
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 Pr(>|z|)
## (Intercept)      4.338e+00  2.493e+00   1.740  0.08181
## Symptoms1       -7.260e-01  4.690e-01  -1.548  0.12169
## `Portal Vein Thrombosis`1 -4.829e-01  5.045e-01  -0.957  0.33840
## `Liver Metastasis`1      -3.996e-01  5.269e-01  -0.758  0.44821
## `Encephalopathy degree`  -7.901e-02  5.591e-01  -0.141  0.88761
## `Ascites degree`       -1.814e-01  3.164e-01  -0.574  0.56631
## `Age at diagnosis`      -1.757e-02  1.667e-02  -1.054  0.29193
## Platelets           4.129e-07  2.260e-06   0.183  0.85502
## `Albumin (mg/dL)`       4.739e-01  3.490e-01   1.358  0.17459
## `log_Aspartate transaminase (U/L)` -5.985e-01  3.439e-01  -1.740  0.08182
## `log_Gamma glutamyl transferase (U/L)` -1.090e-01  2.603e-01  -0.419  0.67547
## `Major dimension of nodule (cm)` -1.021e-01  4.602e-02  -2.219  0.02647
## `log_Direct Bilirubin (mg/dL)`    -3.561e-01  2.925e-01  -1.217  0.22342
## Iron                1.533e-02  5.152e-03   2.976  0.00292
## MELD               -7.455e-02  4.042e-02  -1.844  0.06514
##
## (Intercept)      .
```

```

## Symptoms1
## `Portal Vein Thrombosis`1
## `Liver Metastasis`1
## `Encephalopathy degree`
## `Ascites degree`
## `Age at diagnosis`
## Platelets
## `Albumin (mg/dL)`
## `log_Aspartate transaminase (U/L)`      .
## `log_Gamma glutamyl transferase (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: 152.68  on 150  degrees of freedom
## AIC: 182.68
##
## Number of Fisher Scoring iterations: 5

```

```
#Prediccions de resultats + matriu de confusió
```

```

pred1 <- ifelse(test = modelo$fitted.values > 0.50, yes = 1, no = 0)
matConf <- table(hcc_sel$`Class Attribute`, pred1,
                 dnn = c("observacions", "prediccions"))
matConf

```

```

##              prediccions
## observacions  0  1
##              0 44 19
##              1 13 89

```

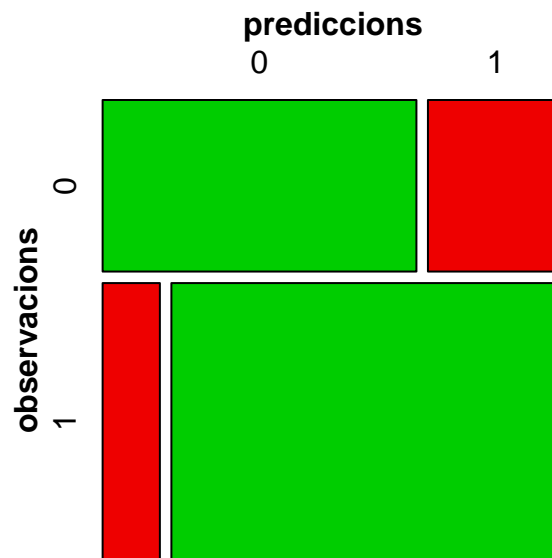
```
#Grafic de matriu de confusió
```

```

mosaic(matConf, shade = T, colorize = T,
       main = "Matriu de confusió del primer model",
       gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))

```

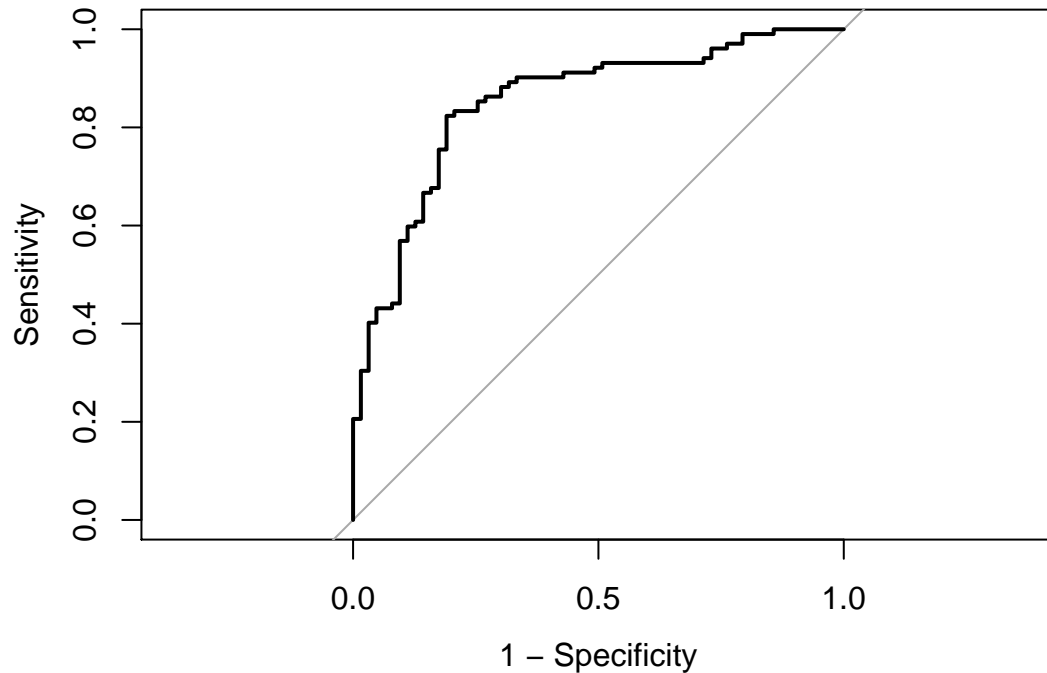
Matriu de confusió del primer model



```
#Èxit del model  
presPred <- 100*(matConf[1]+matConf[4])/nrow(hcc_sel)
```

Aquest model amb totes les variables té un 80.6% d'èxit en la predicció de la supervivència del pacient. Pintem la corba ROC i auc del model:

```
#Predir els valors del model  
prob <- predict(modelo,type = "response")  
hcc$prob <- prob  
  
#Pintar corba ROC, legacy.axes = TRUE per tal de pintar 1- Specificity, si no pinta Specificity.  
plot.roc(`Class Attribute` ~ prob,data = hcc,legacy.axes = TRUE)
```



```
#Area sota la corba
auc(roc(`Class Attribute` ~ prob, data = hcc))
```

```
## Area under the curve: 0.8511
```

Podem veure que tenen valors molts bons.

El nostre objectiu és construir un segon model de regressió logística fent servir menys variables, evitant així haver de conèixer moltes variables d'un pacient i amb unes miques menys poder també fer la predicció. 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").

```
#Model logit mètode backward
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 *
## ---
## 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
```

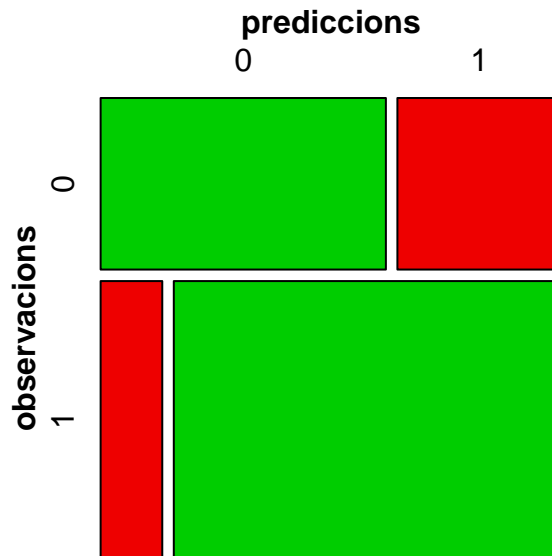
Podem veure com aquest model selecciona fins a 7 variables. Podem veure com l'AIC és menor en aquest model. Calculem la matriu de confusió i el percentatge d'encert en la predicció:

```
#predicció resultat + matriu de confusió
pred2 <- ifelse(test = modback$fitted.values > 0.50, yes = 1, no = 0)
matConf<- table(hcc_sel$`Class Attribute`, pred2,
                dnn = c("observacions", "prediccions"))
matConf
```

```
##                prediccions
## observacions  0  1
##              0 40 23
##              1 14 88
```

```
#grafica matriu de confusió
mosaic(matConf, shade = T, colorize = T,
       main = "Matriu de confusió del primer model",
       gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))
```

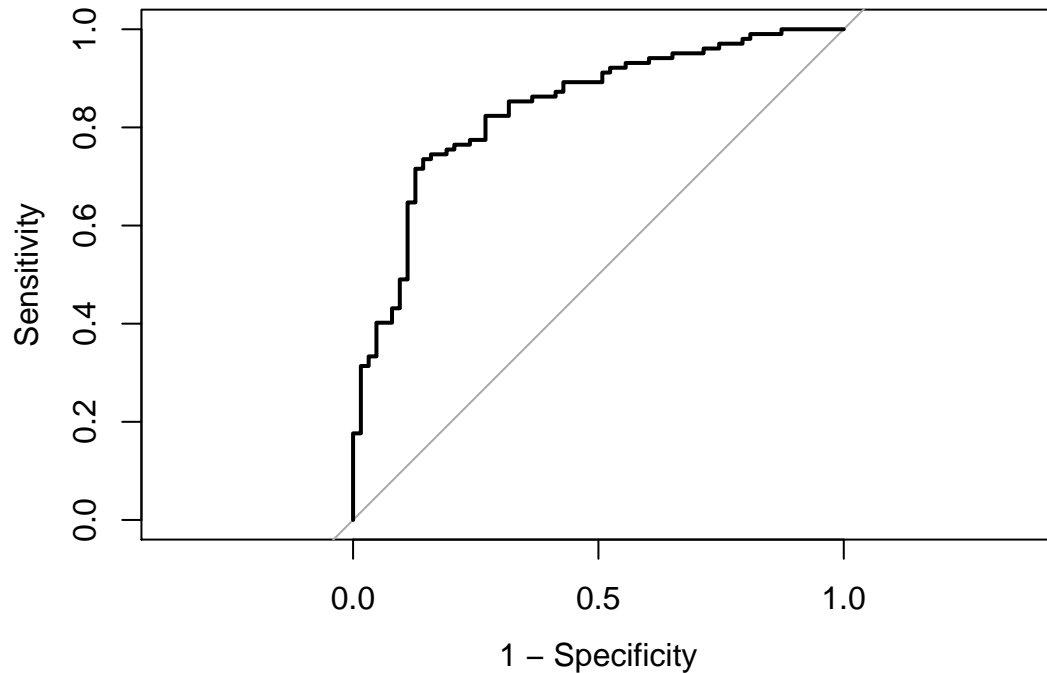
Matriu de confusió del primer model



```
#Èxit del model  
presPred2 <- 100*(matConf[1]+matConf[4])/nrow(hcc_sel)
```

El model generat té només 7 variables i obté un 77.6% de precisió en les prediccions. Pel que fa a l'especificitat i sensibilitat del model:

```
#predicció variables  
prob2 <- predict(modback,type = "response")  
hcc$prob2 <- prob2  
  
#Pintar corba ROC, legacy.axes = TRUE per tal de pintar 1- Specificity, si no pinta Specificity.  
plot.roc(`Class Attribute` ~ prob2,data = hcc,legacy.axes = TRUE)
```



```
#Area sota la corba
auc(roc(`Class Attribute` ~ prob2,data = hcc))
```

```
## Area under the curve: 0.8392
```

Podem veure com el valor de AUC és també una mica pitjor.

Si poguéssim tenir totes les dades possibles, ens quedariem amb el primer model, ja que té una millor predicció de les dades i un valor AUC més elevat. Tanmateix, reduint quasi a la meitat el nombre de variables podem obtenir uns resultats molt bons.

Per tant, hem construït un model logístic, que ens determina la supervivència o no d'un pacient d'acord amb uns indicadors mèdics amb un 80% aproximadament d'èxit. És un molt bon punt per anar refinant el model i millorar el seu ajust.

4 Conclusions

Aquest treball s'ha realitzat sobre dades reals sense processar i, com hem pogut comprovar, és necessari un primer pas de neteja per a poder crear models útils. Les dades en brut tenen valors desconeguts, erronis, i extrems els quals s'han de tractar. En el nostre cas, només 8 dels 165 pacients registrats i només 5 dels 49 atributs disposaven de totes les dades. Eliminar atributs o pacients amb valors desconeguts condicionaria una pèrdua molt important de dades, per al que vam decidir corregir els valor desconeguts amb diferents tècniques d'imputació de valors com són kNN i l'assignació de la mitjana. També hem pogut comprovar que conèixer l'àmbit de les dades fa que sigui possible detectar i corregir algunes de les dades introduïdes erròniament.

Amb el contrast d'hipòtesi hem pogut seleccionar les variables que més influeixen en la supervivència d'aquests pacients i amb l'estudi de correlacions entre aquests atributs eliminar els que tenen alta correlació, per finalment crear models de regressió logística per intentar predir la supervivència. Inclús seleccionant un model de només 7 atributs s'aconsegueix una alta precisió en les prediccions.