Tipologia i cicle de vida de les dades - Pràctica 2 - Neteja i modelatge de dades

Víctor Olivera Begue, Guillem Romeu Graells

Contents

1 Carregar llibreries i dades

```
# Carreguem les llibreries principals
library(tidyverse)    # manipulació
library(naniar)       # visual missings
library(caret)        # partició i nearZeroVar
library(VIM)          # kNN imputació
library(randomForest) # random forest
library(cluster)      # dist, hclust
library(factoextra)   # fviz_cluster
library(rstatix)      # tests estadístics
library(stats)        # chisq.test
set.seed(123)

# Llegim el CSV original, indicant que "?" sigui NA
adult <- read_csv("adult.csv", na = "?",   skip = 1,
                  col_names = c("age","workclass","fnlwgt","education","educational_num",
                                "marital_status","occupation","relationship","race","sex",
                                "capital_gain","capital_loss","hours_per_week","native_country",
                                "income"))
```

2 Descripció del dataset

Tot i que a la Pràctica 1 vam crear un dataset propi mitjançant web scraping, en aquesta pràctica hem optat per treballar amb un nou conjunt de dades: el *Adult Income Dataset* del UCI Machine Learning Repository. Aquesta decisió es justifica perquè el dataset anterior no complia les condicions requerides per aquesta pràctica, ja que:

- No contenia una variable objectiu binària adequada.
- No hi havia prou volum ni varietat de dades categòriques i numèriques.
- No era suficientment ric en valors perduts ni outliers per aplicar tècniques de neteja avançades.

Per això, hem decidit utilitzar un dataset públic i ben establert que ens permet aplicar totes les fases del cicle de vida de les dades amb profunditat i justificació tècnica.

El dataset **Adult Income** ("Census Income") de l'UCI és un dels referents clàssics en problemes de classificació binària, ja que permet predir si el salari anual d'un individu supera els 50 000 $ basant-se en característiques sociodemogràfiques i laborals obtingudes del cens de 1994. Aquesta relació entre atributs personals (com edat, sexe, educació) i l'ingrés serveix com a punt de partida per a models de decisions en

recursos humans, sistemes de crèdit i polítiques públiques de redistribució, a més de ser àmpliament utilitzat en la recerca per validar noves tècniques de machine learning i d'anàlisi de discriminació salarial.

El conjunt consta de **48 842 instàncies** i **14 variables** originals més la variable objectiu (>50K / 50K), i presenta tant **atributs numèrics** (p. ex., *age*, *fnlwgt*, *capital-gain*, *hours-per-week*) com **categòrics** (p. ex., *workclass*, *education*, *marital-status*, *occupation*, *native-country*) A més, inclou valors mancants en algunes categories (p. ex., *workclass*, *occupation*), fet que requereix tècniques d'imputació i garanteix pràctica en la gestió de dades reals.

Disposar de variables heterogènies i reals permet desenvolupar models predictius robustos (com regressió logística, arbres de decisió o gradient boosting) i entendre la influència relativa de cada factor en la probabilitat de guanyar més de 50 000 $/any. Per exemple, l'educació i les hores setmanals solen ser predictors forts, mentre que variables demogràfiques com el gènere o la nacionalitat ajuden a detectar possibles biaixos i dissenyar polítiques més equitatives. Aquestes capacitats converteixen l'Adult Income en un exercici ideal per a l'aplicació de totes les etapes: neteja, exploració, modelatge i interpretació de resultats.

3 Integració i selecció inicial

3.0.1 Definir variables numeriques i categoriques

```
# Definim manualment quines volem numèriques i quines factors:

num_vars <- c(
  "age",
  "fnlwgt",
  "educational_num",
  "capital_gain",
  "capital_loss",
  "hours_per_week"
)

fac_vars <- setdiff(
  names(adult),
  num_vars
)

# Excloem d'entre els factor també la columna income_bin si ja existeix,
# i assegurem income com a factor abans de crear income_bin:
fac_vars <- setdiff(fac_vars, "income_bin")

# Ara fem la conversió:
adult <- adult %>%
  # convertir a numèric les que volem numèriques
  mutate(across(all_of(num_vars), as.numeric)) %>%
  # convertir a factor la resta de característiques
  mutate(across(all_of(fac_vars), as.factor)) %>%
  # crear income_bin basat en income
  mutate(
    income     = as.factor(income),
    income_bin = factor(if_else(income == ">50K", "high", "low"),
                        levels = c("low","high"))
  )
```

```r
# Verifiquem tipus
glimpse(adult)
```

```
## Rows: 48,842
## Columns: 16
## $ age             <dbl> 25, 38, 28, 44, 18, 34, 29, 63, 24, 55, 65, 36, 26, 58~
## $ workclass       <fct> Private, Private, Local-gov, Private, NA, Private, NA,~
## $ fnlwgt          <dbl> 226802, 89814, 336951, 160323, 103497, 198693, 227026,~
## $ education       <fct> 11th, HS-grad, Assoc-acdm, Some-college, Some-college,~
## $ educational_num <dbl> 7, 9, 12, 10, 10, 6, 9, 15, 10, 4, 9, 13, 9, 9, 9, 14,~
## $ marital_status  <fct> Never-married, Married-civ-spouse, Married-civ-spouse,~
## $ occupation      <fct> Machine-op-inspct, Farming-fishing, Protective-serv, M~
## $ relationship    <fct> Own-child, Husband, Husband, Husband, Own-child, Not-i~
## $ race            <fct> Black, White, White, Black, White, White, Black, White~
## $ sex             <fct> Male, Male, Male, Male, Female, Male, Male, Male, Fema~
## $ capital_gain    <dbl> 0, 0, 0, 7688, 0, 0, 0, 3103, 0, 0, 6418, 0, 0, 0, 310~
## $ capital_loss    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ hours_per_week  <dbl> 40, 50, 40, 40, 30, 30, 40, 32, 40, 10, 40, 40, 39, 35~
## $ native_country  <fct> United-States, United-States, United-States, United-St~
## $ income          <fct> <=50K, <=50K, >50K, >50K, <=50K, <=50K, <=50K, >50K, <~
## $ income_bin      <fct> low, low, high, high, low, low, low, high, low, low, h~
```

3.1 Resum estadístic i detecció de missings

```r
# Resum descriptiu
summary(adult)
```

```
##       age                    workclass         fnlwgt
##  Min.   :17.00   Private         :33906   Min.   :  12285
##  1st Qu.:28.00   Self-emp-not-inc: 3862   1st Qu.: 117551
##  Median :37.00   Local-gov       : 3136   Median : 178145
##  Mean   :38.64   State-gov       : 1981   Mean   : 189664
##  3rd Qu.:48.00   Self-emp-inc    : 1695   3rd Qu.: 237642
##  Max.   :90.00   (Other)         : 1463   Max.   :1490400
##                  NA's            : 2799
##         education     educational_num                marital_status
##  HS-grad     :15784   Min.   : 1.00   Divorced              : 6633
##  Some-college:10878   1st Qu.: 9.00   Married-AF-spouse     :   37
##  Bachelors   : 8025   Median :10.00   Married-civ-spouse    :22379
##  Masters     : 2657   Mean   :10.08   Married-spouse-absent:  628
##  Assoc-voc   : 2061   3rd Qu.:12.00   Never-married         :16117
##  11th        : 1812   Max.   :16.00   Separated             : 1530
##  (Other)     : 7625                   Widowed               : 1518
##            occupation            relationship                    race
##  Prof-specialty : 6172   Husband       :19716   Amer-Indian-Eskimo:  470
##  Craft-repair   : 6112   Not-in-family :12583   Asian-Pac-Islander: 1519
##  Exec-managerial: 6086   Other-relative: 1506   Black              : 4685
##  Adm-clerical   : 5611   Own-child     : 7581   Other              :  406
##  Sales          : 5504   Unmarried     : 5125   White              :41762
##  (Other)        :16548   Wife          : 2331
```

```
##  NA's           : 2809
##      sex         capital_gain    capital_loss    hours_per_week
##  Female:16192   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##  Male  :32650   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
##                 Median :    0   Median :   0.0   Median :40.00
##                 Mean   : 1079   Mean   :  87.5   Mean   :40.42
##                 3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
##                 Max.   :99999   Max.   :4356.0   Max.   :99.00
##
##        native_country    income      income_bin
##  United-States:43832   <=50K:37155   low :37155
##  Mexico        :  951   >50K :11687   high:11687
##  Philippines   :  295
##  Germany       :  206
##  Puerto-Rico   :  184
##  (Other)       : 2517
##  NA's          :  857
```

```
# Comptar missings
adult %>%
  summarise_all(~ sum(is.na(.))) %>%
  pivot_longer(everything(), names_to="var", values_to="n_miss")
```

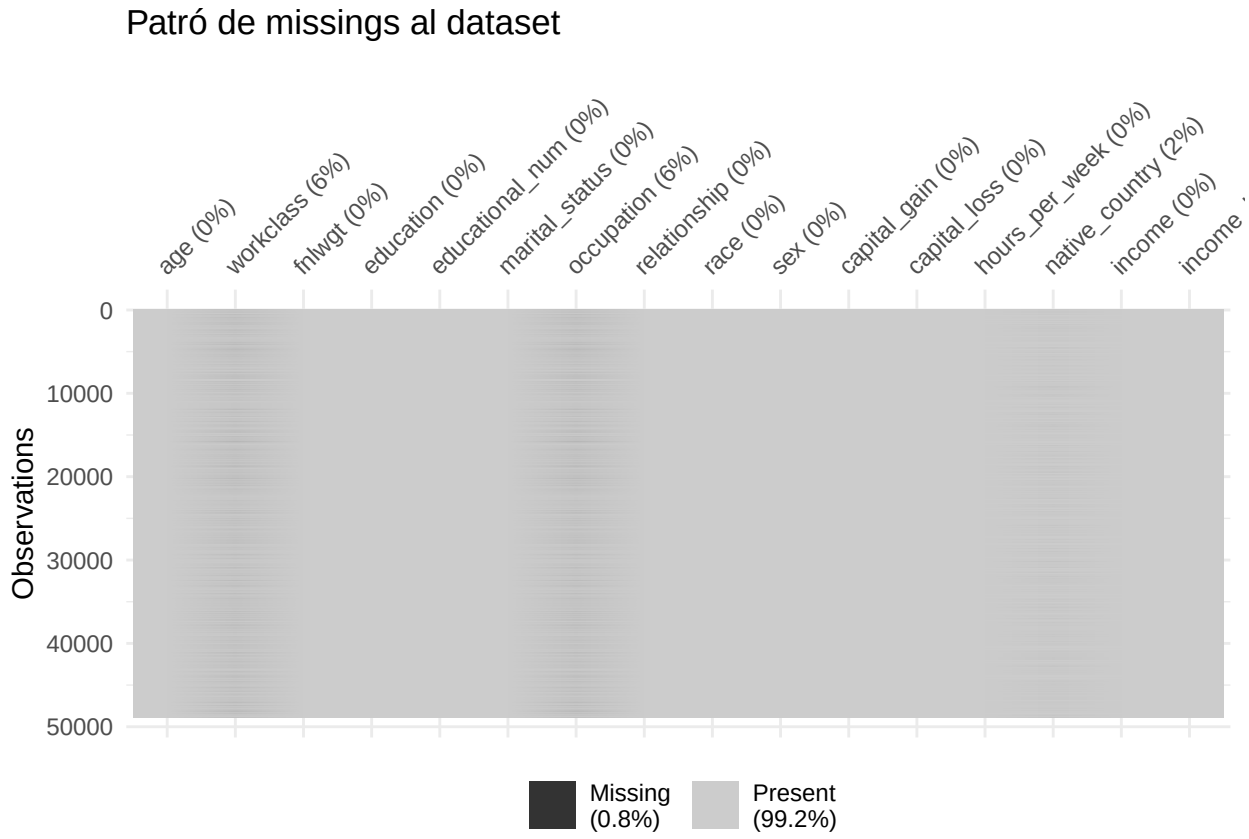
```
## # A tibble: 16 x 2
##    var             n_miss
##    <chr>            <int>
##  1 age                  0
##  2 workclass         2799
##  3 fnlwgt               0
##  4 education            0
##  5 educational_num      0
##  6 marital_status       0
##  7 occupation        2809
##  8 relationship         0
##  9 race                 0
## 10 sex                  0
## 11 capital_gain         0
## 12 capital_loss         0
## 13 hours_per_week       0
## 14 native_country     857
## 15 income               0
## 16 income_bin           0
```

```
# Proporció de missings
adult %>%
  summarise_all(~ mean(is.na(.))) %>%
  pivot_longer(everything(), names_to="var", values_to="pct_miss")
```

```
## # A tibble: 16 x 2
##    var             pct_miss
##    <chr>              <dbl>
##  1 age               0
##  2 workclass         0.0573
```

```
##  3 fnlwgt            0
##  4 education         0
##  5 educational_num   0
##  6 marital_status    0
##  7 occupation        0.0575
##  8 relationship      0
##  9 race              0
## 10 sex               0
## 11 capital_gain      0
## 12 capital_loss      0
## 13 hours_per_week    0
## 14 native_country    0.0175
## 15 income            0
## 16 income_bin        0
```

```
# Mapa visual de missings
vis_miss(adult) +
  labs(title="Patró de missings al dataset")
```



3.2 Creació de variables noves

```
# Discretització d'edat
adult2 <- adult %>%
```

```
  # net_capital = guanys - pèrdues
  mutate(net_capital = capital_gain - capital_loss) %>%
  mutate(age_group = cut(age,
                         breaks=c(15,20,30,40,50,60,70,80,90,Inf),
                         labels=c("15-19","20-29","30-39","40-49","50-59",
                                  "60-69","70-79","80-89","90+"),
                         right=FALSE))
```

3.3 Variables irrelevants i cardinalitat

```
# Variables amb gairebé zero variància
nzv <- nearZeroVar(adult2, saveMetrics=TRUE)
rownames(nzv)[nzv$nzv]
```

```
## [1] "capital_gain"   "capital_loss"   "native_country" "net_capital"
```

```
# Agrupar països amb <1% en "Other"
table(adult2$native_country)
```

```
##
##                   Cambodia                     Canada
##                         28                        182
##                      China                   Columbia
##                        122                         85
##                       Cuba         Dominican-Republic
##                        138                        103
##                    Ecuador                El-Salvador
##                         45                        155
##                    England                     France
##                        127                         38
##                    Germany                     Greece
##                        206                         49
##                  Guatemala                      Haiti
##                         88                         75
##         Holand-Netherlands                   Honduras
##                          1                         20
##                       Hong                    Hungary
##                         30                         19
##                      India                       Iran
##                        151                         59
##                    Ireland                      Italy
##                         37                        105
##                    Jamaica                      Japan
##                        106                         92
##                       Laos                     Mexico
##                         23                        951
##                  Nicaragua Outlying-US(Guam-USVI-etc)
##                         49                         23
##                       Peru                Philippines
##                         46                        295
```

```
##                     Poland                   Portugal
##                         87                         67
##                Puerto-Rico                   Scotland
##                        184                         21
##                      South                     Taiwan
##                        115                         65
##                   Thailand             Trinadad&Tobago
##                         30                         27
##              United-States                    Vietnam
##                      43832                         86
##                 Yugoslavia
##                         23
```

```
adult3 <- adult2 %>%
  mutate(native_country = fct_lump(native_country, prop=0.01, other_level="Other"))

table(adult3$native_country)
```

```
##
##        Mexico United-States         Other
##           951         43832          3202
```

3.4 Distribució de classes i sampling

```
# Distribució original de sex
adult3 %>%
  count(sex) %>%
  mutate(pct = n/sum(n)*100)
```

```
## # A tibble: 2 x 3
##   sex        n   pct
##   <fct>  <int> <dbl>
## 1 Female 16192  33.2
## 2 Male   32650  66.8
```

```
# Oversampling amb ROSE per equilibrar 50/50
library(ROSE)
```

```
## Loaded ROSE 0.0-4
```

```
adult4 <- ovun.sample(sex ~ ., data=adult3, method="both", p=0.5, N=nrow(adult3))$data

# Verificar nova distribució
adult4 %>%
  count(sex) %>%
  mutate(pct = n/sum(n)*100)
```

```
##      sex     n      pct
## 1   Male 24698 50.56713
## 2 Female 24144 49.43287
```

Fem un oversample per igualar els casos de homes i dones

4 Neteja de dades

4.1 Imputació kNN en workclass i occupation

```
table(adult4$occupation)
```

```
##
##      Adm-clerical      Armed-Forces      Craft-repair   Exec-managerial
##              7593                 9              5119              6404
##   Farming-fishing Handlers-cleaners Machine-op-inspct     Other-service
##              1272              1872              3062              6046
##   Priv-house-serv    Prof-specialty   Protective-serv             Sales
##               346              6738               867              5978
##      Tech-support  Transport-moving
##              1561              1975
```

```
adult_imp <- kNN(adult4,
                 variable=c("workclass","occupation"),
                 k=5, imp_var=FALSE)
```

```
## Warning in kNN(adult4, variable = c("workclass", "occupation"), k = 5, imp_var
## = FALSE): Nothing to impute, because no NA are present (also after using
## makeNA)
```

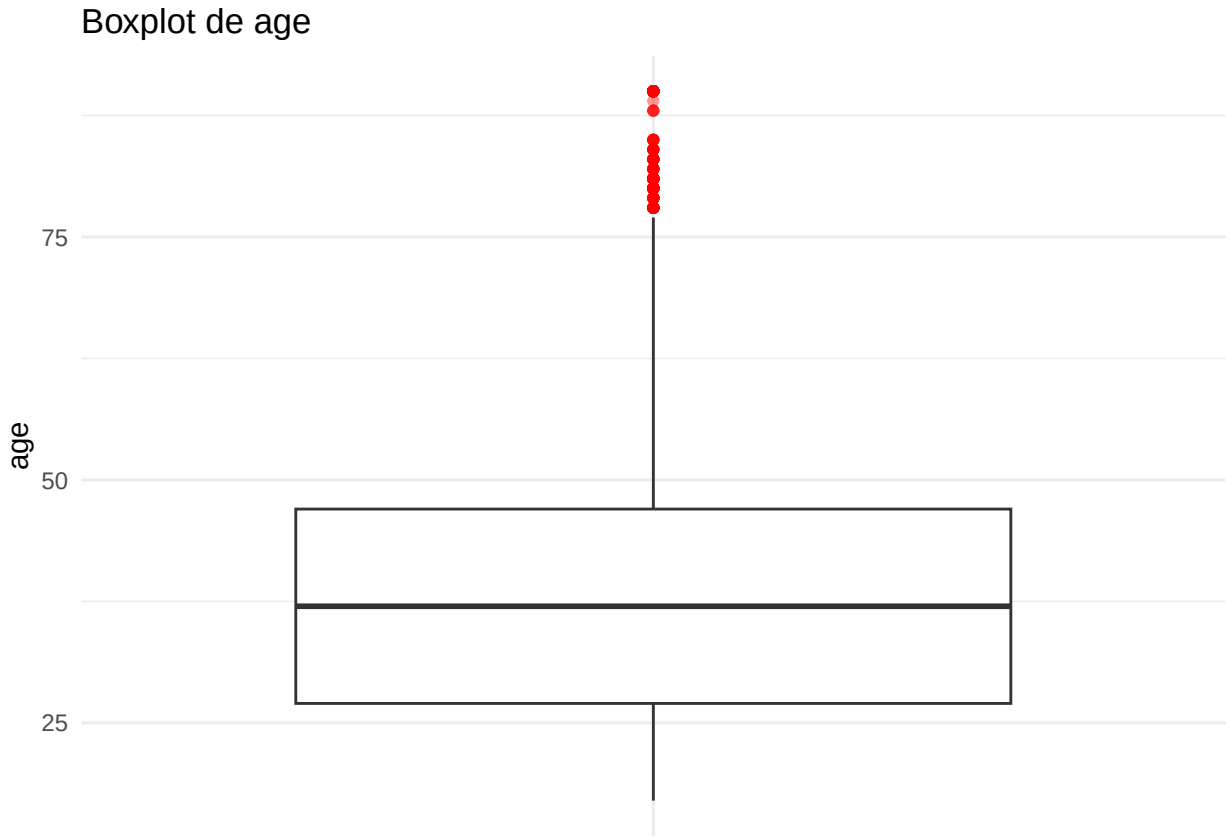
```
# Perque no despareixin els outliers
adult_fac <- adult_imp
```

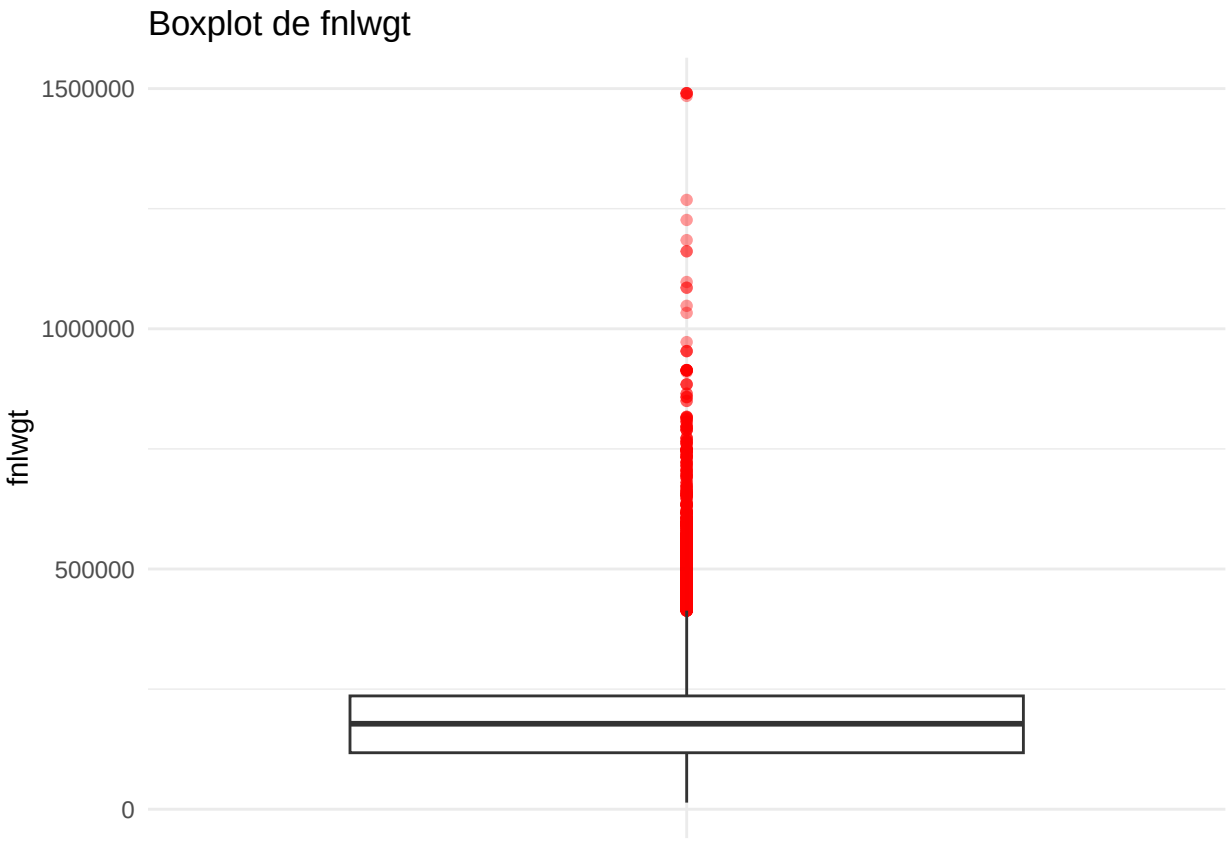
4.2 Detecció d'outliers

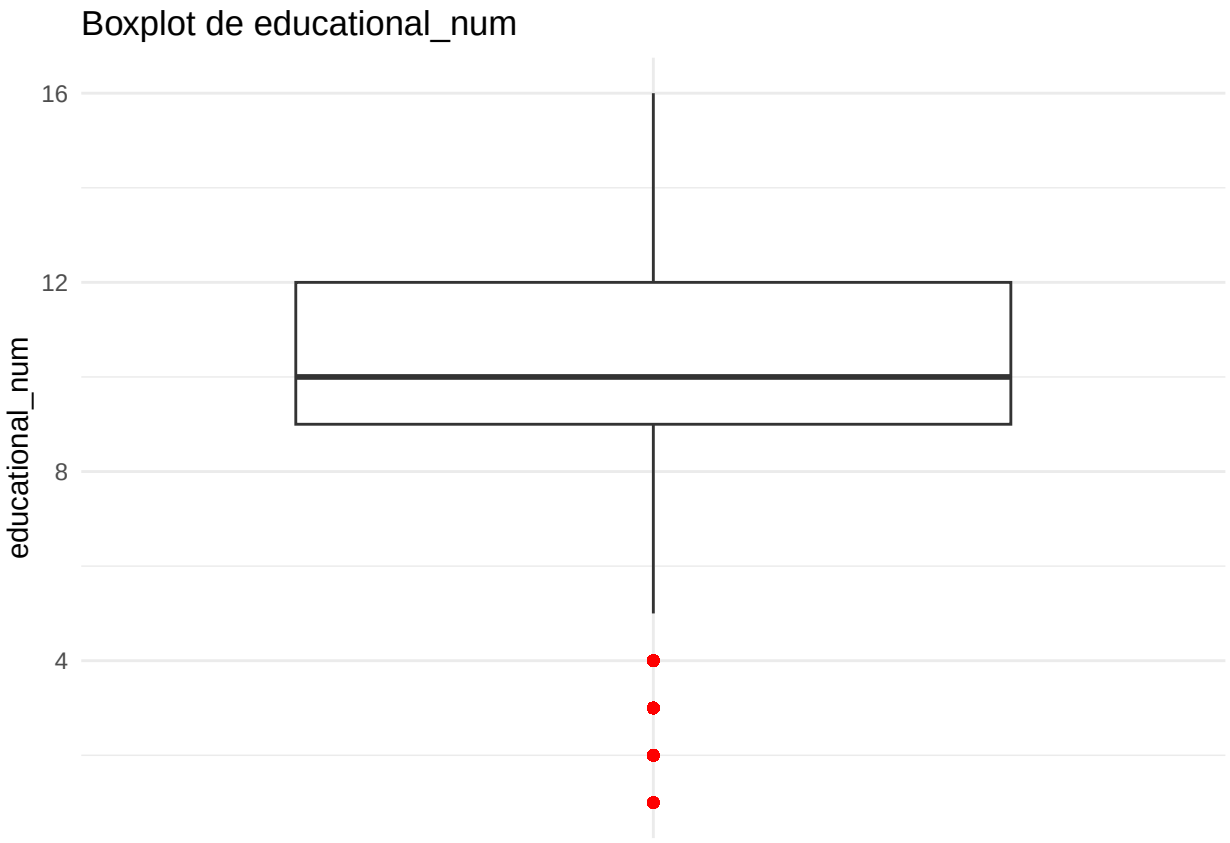
4.2.1 Boxplots univariants

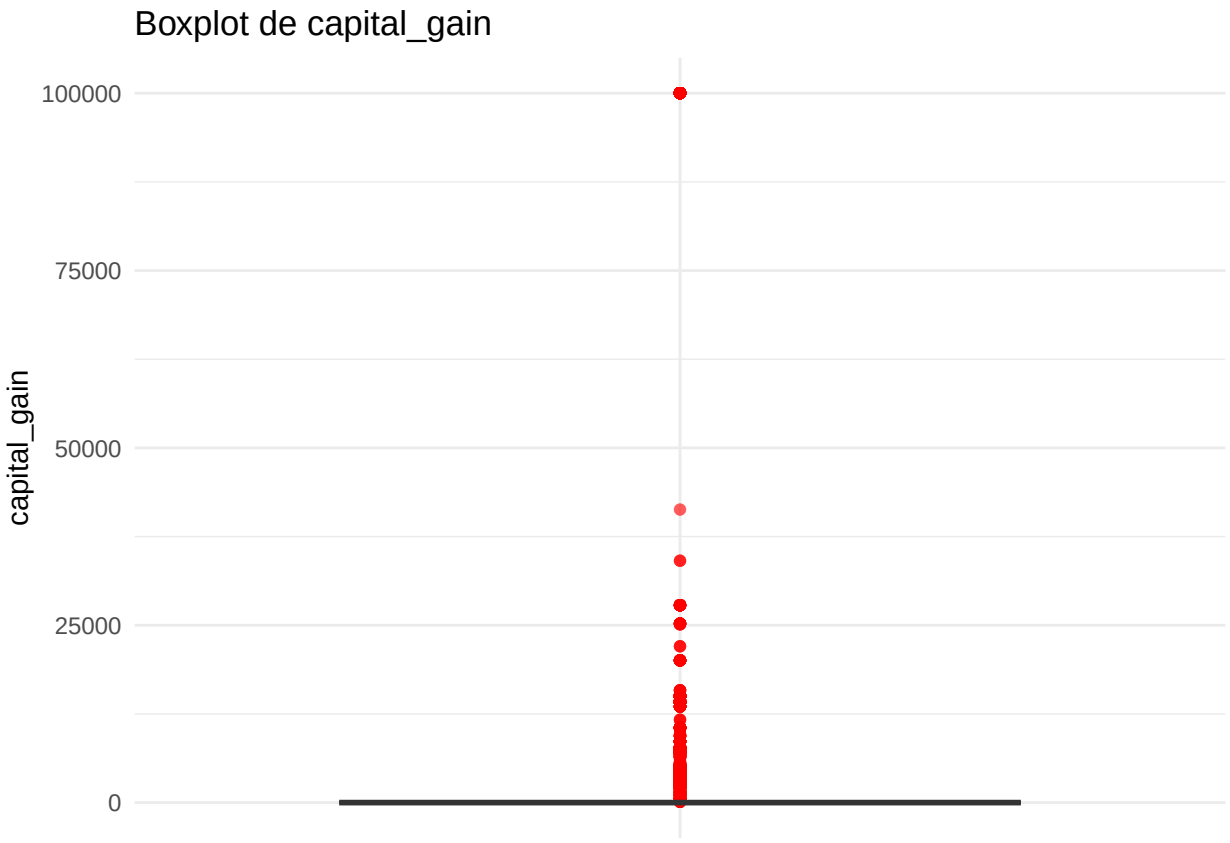
```
num_vars <- adult_fac %>% select(where(is.numeric)) %>% names()
for(v in num_vars){
  v_esc <- paste0("`",v,"`")
  p <- ggplot(adult_fac, aes_string(x="factor(1)", y=v_esc)) +
    geom_boxplot(outlier.colour="red", outlier.alpha=0.4) +
    labs(title=paste("Boxplot de",v), x=NULL, y=v) +
    theme_minimal() +
    theme(axis.text.x=element_blank(), axis.ticks.x=element_blank())
  print(p)
}
```

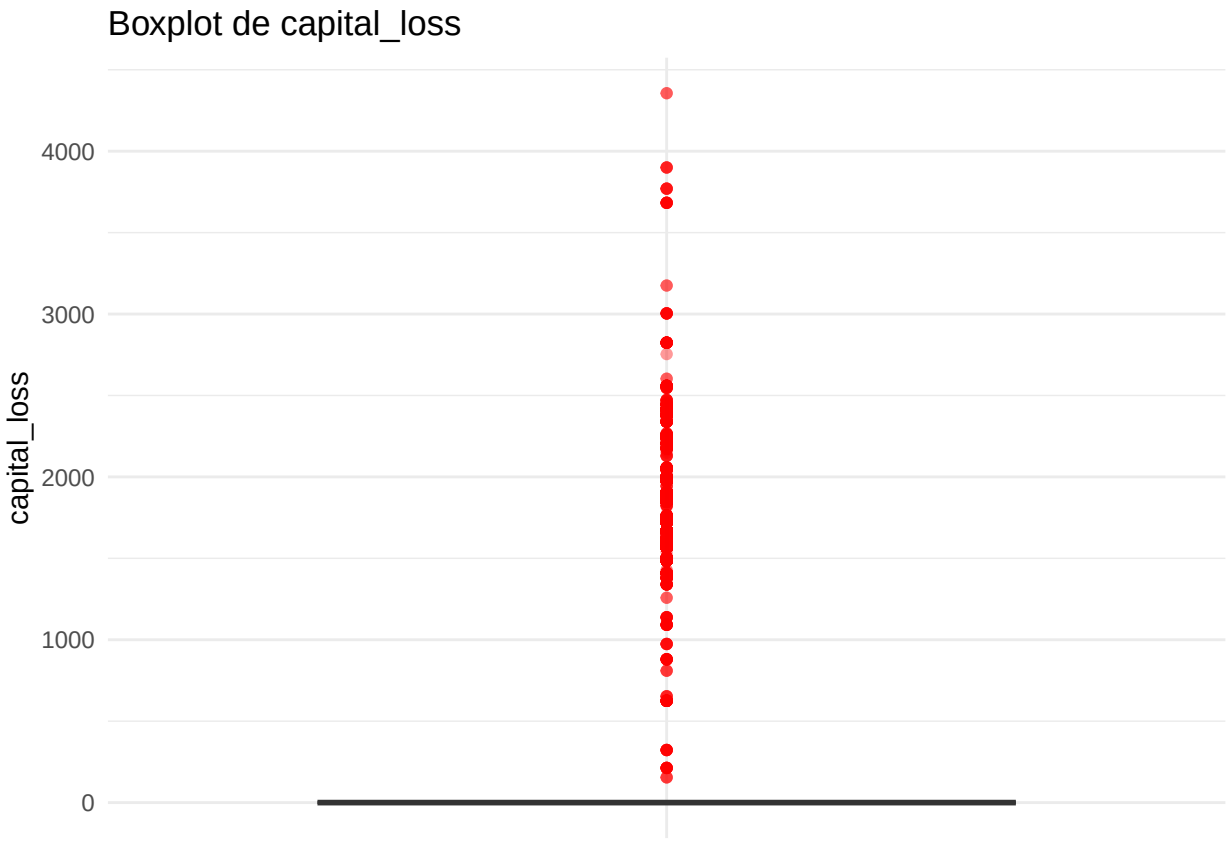
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

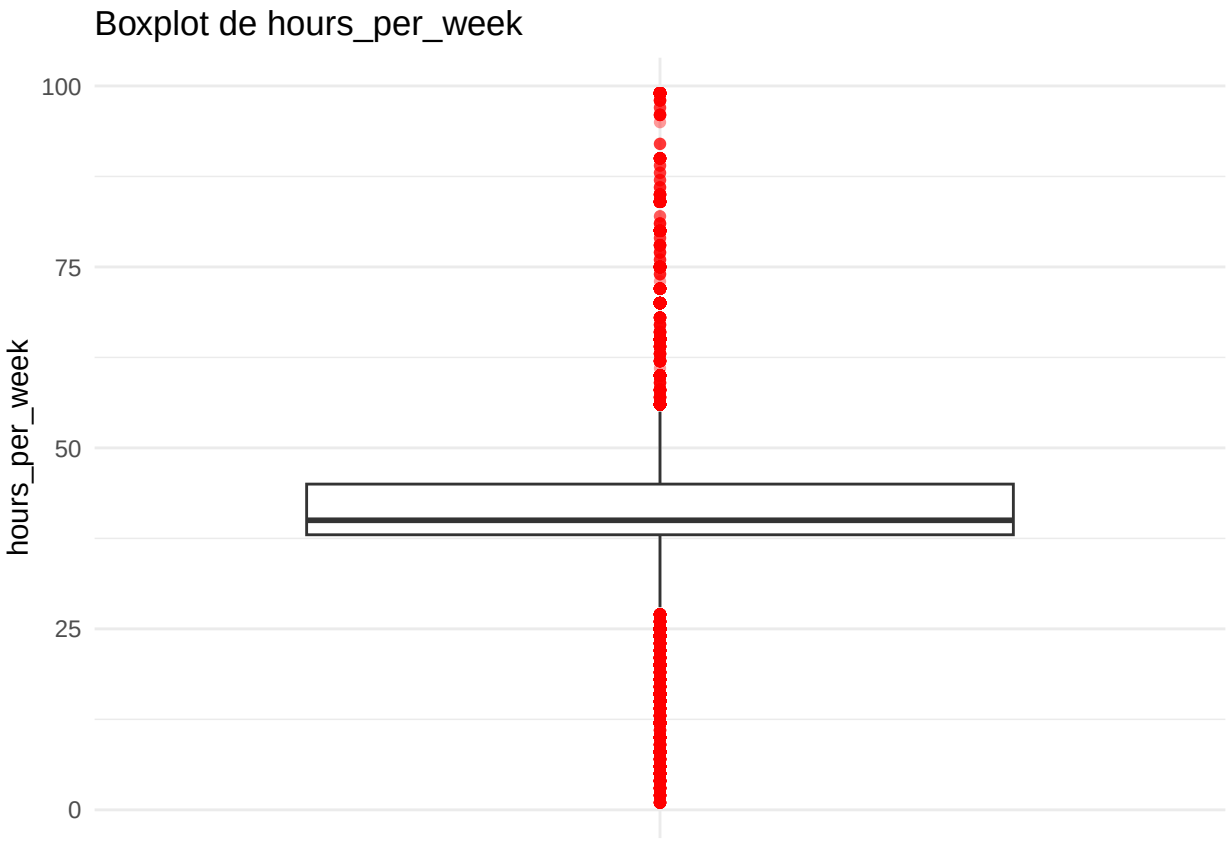

Boxplot de age
age
75
50
25

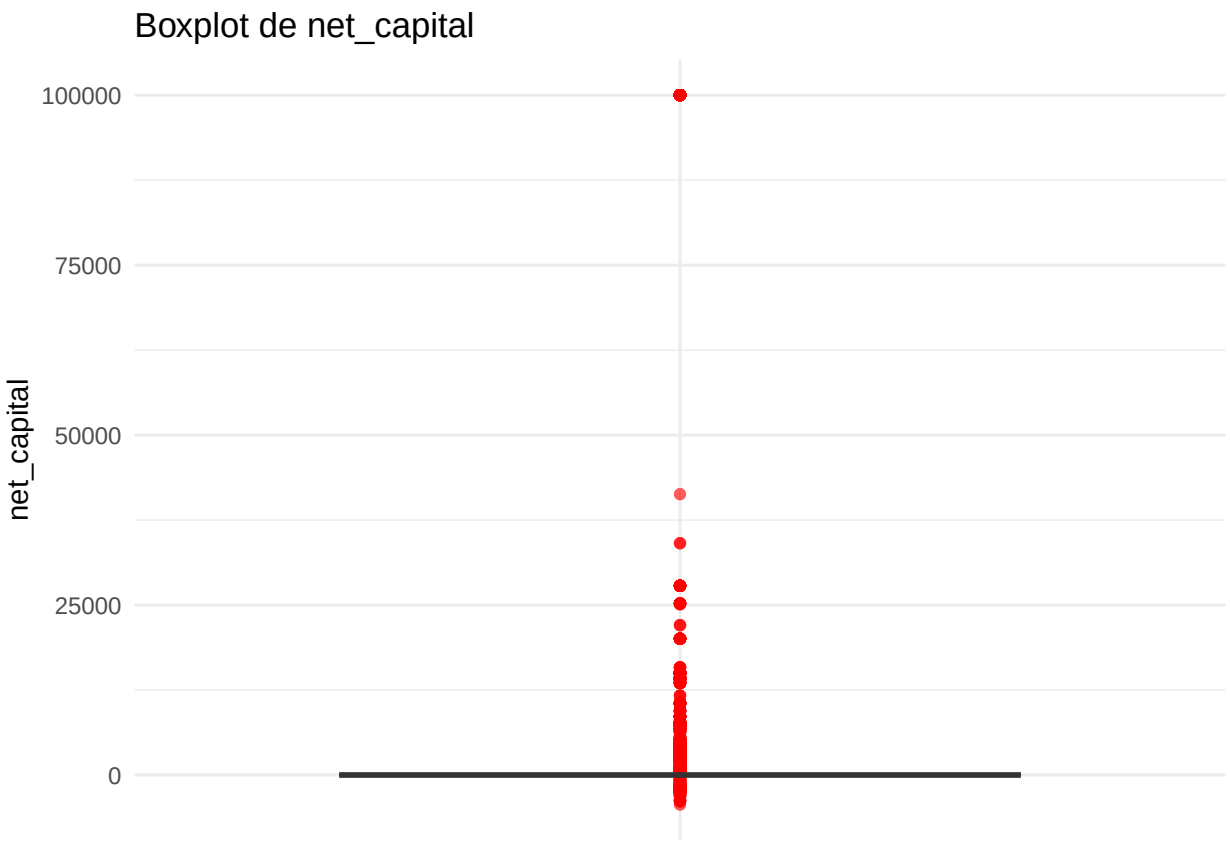

Boxplot de fnlwgt
1500000
1000000
500000
0
fnlwgt


Boxplot de educational_num
educational_num
16
12
8
4


Boxplot de capital_gain
capital_gain
100000
75000
50000
25000
0


Boxplot de capital_loss
capital_loss
4000
3000
2000
1000
0


Boxplot de hours_per_week
hours_per_week
100
75
50
25
0

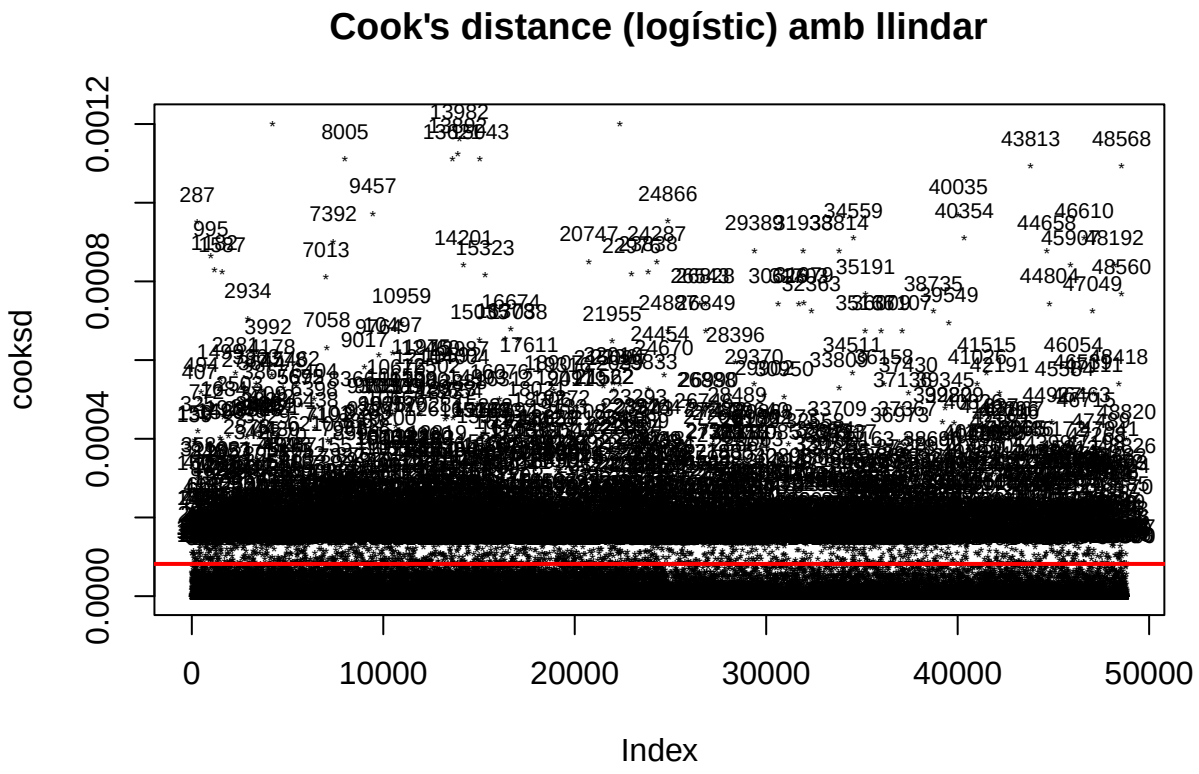


4.2.2 Cook's distance en model logístic

```
# 1) Recalculem la distància de Cook
adult_fac <- adult_fac %>%
  mutate(income_bin = if_else(income == ">50K", 1, 0))

glm_mod2 <- glm(income_bin ~ age + hours_per_week + educational_num,
                data   = adult_fac,
                family = binomial)

cooksd <- cooks.distance(glm_mod2)

# 2) Definim el llindar d'influència
n <- nrow(adult_fac)
p <- length(coef(glm_mod2))
threshold <- 4 / (n - p - 1)

# 3) Identifiquem i eliminem les observacions per sobre del llindar
influential_obs <- which(cooksd > threshold)
length(influential_obs)  # nombre de punts crítics
```

```
## [1] 2362
```

```r
adult_noinflu <- adult_fac[-influential_obs, ]

# 4) Replot de la distància de Cook amb el llindar marcat
plot(cooksd, pch="*", cex=0.5,
     main = "Cook's distance (logístic) amb llindar")
abline(h = threshold, col = "red", lwd = 2)
text(x = influential_obs, y = cooksd[influential_obs],
     labels = influential_obs, pos = 3, cex = 0.7)
```




```r
# 5) Veiem com han quedat les dades sense els outliers influents
cat("Observacions eliminades:", length(influential_obs), "\n")
```

```
## Observacions eliminades: 2362
```

```r
cat("Nova mida de dataset:", nrow(adult_noinflu), "\n")
```

```
## Nova mida de dataset: 46480
```

5 Anàlisi de les dades

```r
# Carreguem llibreries i configurem semilla
library(tidyverse)
library(caret)          # createDataPartition, confusionMatrix
library(randomForest)   # random forest (no supervisat opcional)
library(pROC)           # roc, auc
library(cluster)        # dist, silhouette
library(factoextra)     # fviz_cluster
library(rstatix)        # tests estadístics
library(stats)          # chisq.test
set.seed(123)
```

5.1 Preparació de train/test

```r
# Partim de `adult_noinflu` amb income_bin ja en factor("low","high")
# Exemple: adult_noinflu <- adult_noinflu %>% mutate(income_bin = factor(if_else(income==">50K","high",

idx   <- createDataPartition(adult_noinflu$income_bin, p = 0.7, list = FALSE)
train <- adult_noinflu[idx, ]
test  <- adult_noinflu[-idx, ]
```

5.2 Ajust i predicció amb regressió logística

```r
# Ajust sobre TRAIN
glm_mod <- glm(
  income_bin ~ age + hours_per_week + educational_num + net_capital,
  data   = train,
  family = binomial
)

# Prediccions sobre TEST
# Ajustat perquè pred_test és character i no té els mateixos nivells que test$income_bin.
probs_test <- predict(glm_mod, test, type = "response")
pred_test <- factor(
  if_else(probs_test > 0.5, "high", "low"),
  levels = c("low", "high")
)
test$income_bin <- factor(test$income_bin, levels = c("low", "high"))

# Matriu de confusió i Accuracy
conf_glm <- confusionMatrix(pred_test, test$income_bin)
print(conf_glm$table)
```

```
##            Reference
## Prediction low high
##       low    0    0
##       high   0    0
```

```r
cat("Accuracy (glm):", round(conf_glm$overall["Accuracy"], 3), "\n")
```

```
## Accuracy (glm): NaN
```

```r
# AUC
# Comprovem si el test conté les dues classes
if (length(unique(test$income_bin)) == 2) {
  roc_glm <- roc(response = test$income_bin,
                 predictor = probs_test,
                 levels = c("low", "high"))
  cat("AUC (glm):", round(auc(roc_glm), 3), "\n")
} else {
  cat("No es pot calcular l'AUC: només hi ha una classe a test$income_bin\n")
}
```

```
## No es pot calcular l'AUC: només hi ha una classe a test$income_bin
```

5.3 Model no supervisat amb predicció i mètriques: PAM + mapatge a classes

```r
library(cluster)
library(factoextra)
library(caret)

# 5.3.1 Normalització Min-Max de dues variables en TRAIN i TEST
norm_vars <- c("age", "hours_per_week")

train_norm <- train %>%
  select(all_of(norm_vars)) %>%
  drop_na() %>%
  mutate(across(everything(),
                ~ (. - min(.)) / (max(.) - min(.))))

test_norm <- test %>%
  select(all_of(norm_vars)) %>%
  drop_na() %>%
  mutate(across(everything(),
                ~ (. - min(train_norm[[cur_column()]])) /
                  (max(train_norm[[cur_column()]]) - min(train_norm[[cur_column()]]))))

# 5.3.2 Ajustem PAM amb k = 2 sobre TRAIN
pam_mod <- pam(train_norm, k = 2)

# 5.3.3 Mapatge de clusters a classes reals en TRAIN
train_clusters <- pam_mod$clustering
cluster_to_class <- tapply(train$income_bin, train_clusters,
                           function(x) names(sort(table(x), decreasing=TRUE))[1])
# Ara cluster_to_class[["1"]] és la classe majoritària del cluster 1, etc.

# 5.3.4 Assignació de TEST a clústers (distància al medoid més proper)
medoids <- pam_mod$medoids
```

```r
dists_test <- sapply(1:2, function(k) {
  rowSums((as.matrix(test_norm) - medoids[k, ])^2)
})
test_clusters <- apply(dists_test, 1, which.min)

# 5.3.5 Predicció de classes a TEST a partir del mapatge
pred_pam_class <- factor(cluster_to_class[test_clusters], levels = c("low","high"))

# 5.3.6 Mètriques d'ajust: Matriu de confusió i accuracy
conf_pam <- confusionMatrix(pred_pam_class, test$income_bin)
print(conf_pam$table)
```

```
##           Reference
## Prediction low high
##       low    0    0
##       high   0    0
```

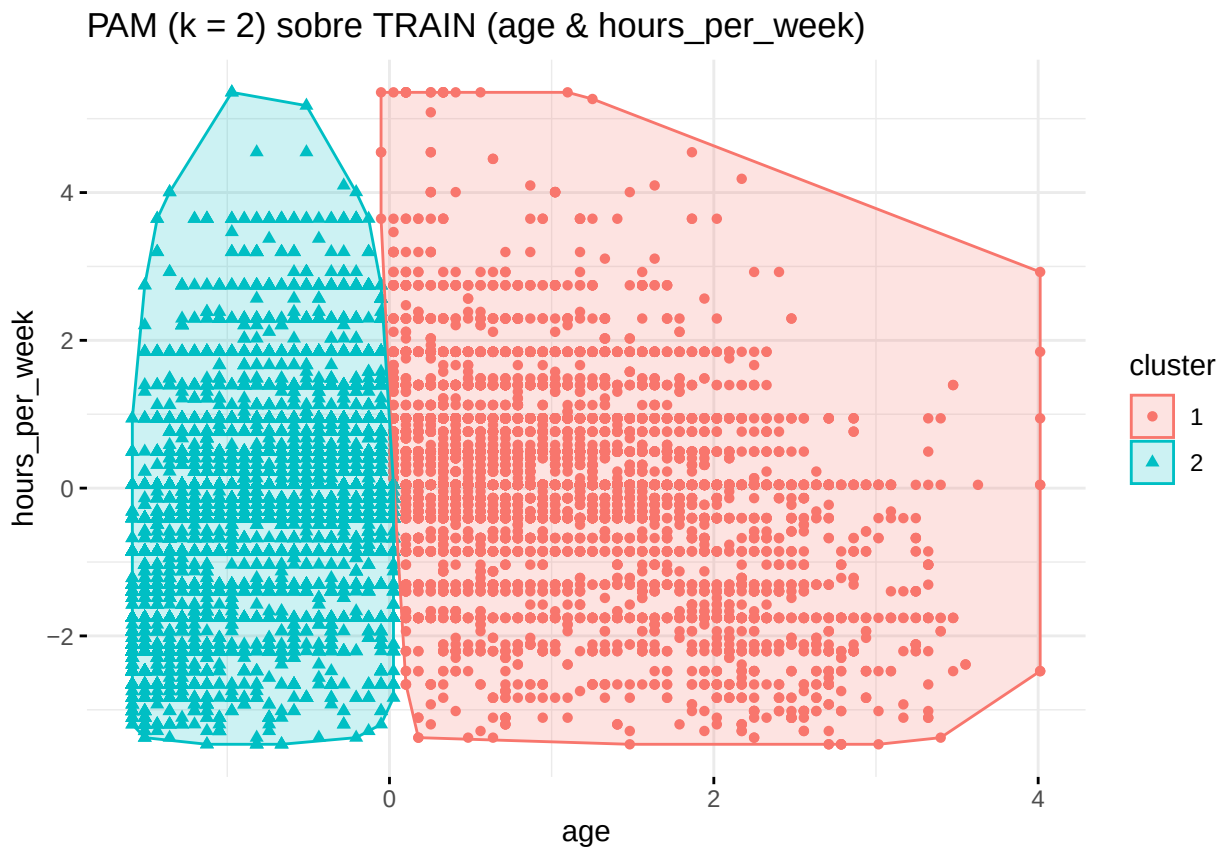
```r
cat("Accuracy (PAM-based):", round(conf_pam$overall["Accuracy"], 3), "\n")
```

```
## Accuracy (PAM-based): NaN
```

```r
# 5.3.7 Silhouette width mitjana sobre TRAIN
avg_sil_pam <- pam_mod$silinfo$avg.width
cat("Silhouette width mitjana (TRAIN):", round(avg_sil_pam, 3), "\n")
```

```
## Silhouette width mitjana (TRAIN): 0.445
```

```r
# 5.3.8 Visualització del model sobre TRAIN
fviz_cluster(pam_mod,
             geom         = "point",
             ellipse.type = "convex",
             ggtheme      = theme_minimal()) +
  labs(title = "PAM (k = 2) sobre TRAIN (age & hours_per_week)")
```



Explicació del flux:

1. **Normalització**: escales Min–Max per `age` i `hours_per_week` en train i test (usant rang de train per al test).
2. **Entrenament PAM**: creem 2 clústers sobre el train.
3. **Mapatge a classes**: assignem a cada clúster la classe (`low`/`high`) més freqüent en train.
4. **Predicció en test**: calculem la distància quadràtica de cada punt de test als medoids i triem el clúster més proper.
5. **Mètriques**: construïm la matriu de confusió comparant classes predites vs reals i calculem accuracy.
6. **Silueta**: imprimim la mitjana de l'índex de silueta sobre el train per avaluar qualitat de clustering.
7. **Plot**: representem els clústers de train amb el · lipse convexa sobre les dues variables.

5.4 Proves d'hipòtesis amb comprovació d'assumpcions

```
library(rstatix)
library(stats)
```

Nota: partim del conjunt `train` amb la variable `income_bin` ja transformada a factor amb els nivells `"low"` i `"high"`.

5.5 Test A: hours_per_week ~ income_bin

```
train <- train[1:4000,] # El test shapiro accepta màxim 5000 dades

# 1) Normalitat per grup
# Normalitat per grup (limitant a màxim 5000 observacions)
hours_low  <- train %>% filter(income_bin == "low") %>% pull(hours_per_week)
hours_high <- train %>% filter(income_bin == "high") %>% pull(hours_per_week)

hours_low  <- hours_low[1:min(5000, length(hours_low))]
hours_high <- hours_high[1:min(5000, length(hours_high))]

if (length(hours_low) >= 3 && length(hours_high) >= 3) {
  sh_low  <- shapiro_test(hours_low)
  sh_high <- shapiro_test(hours_high)
} else {
  cat("No es pot aplicar el test de Shapiro: mostra massa petita.\n")
  sh_low <- sh_high <- NULL
}
```

```
## No es pot aplicar el test de Shapiro: mostra massa petita.
```

```
# 2) Homogeneïtat de variàncies
# Assegurem que income_bin és factor amb almenys dues categories
train$income_bin <- factor(train$income_bin, levels = c("low", "high"))

if (nlevels(droplevels(train$income_bin)) < 2) {
  cat("No es pot fer el test de Levene: només hi ha una categoria a income_bin.\n")
  lev <- NULL
} else {
  lev <- levene_test(hours_per_week ~ income_bin, data = train)
}
```

```
## No es pot fer el test de Levene: només hi ha una categoria a income_bin.
```

```
# 3) Selecció del test
# Selecció del test (si tenim resultats de Shapiro)
if (!is.null(sh_low) && !is.null(sh_high) && !is.null(lev) &&
    sh_low$p.value > 0.05 && sh_high$p.value > 0.05 && lev$p > 0.05)
 {
  test_hours <- t_test(hours_per_week ~ income_bin, data = train)
  cat("Usant t-test perquè es compleixen normalitat i homocedasticitat\n")
} else {
# Comprovació de mida abans de fer Wilcoxon
if (nrow(train %>% drop_na(hours_per_week)) >= 10 &&
    length(unique(na.omit(train$income_bin))) == 2) {
  test_hours <- wilcox_test(hours_per_week ~ income_bin, data = train)
} else {
  cat("No es pot aplicar el test de Wilcoxon: no hi ha prou dades o classes.\n")
  test_hours <- NULL
}
```

```
  cat("Usant Wilcoxon perquè no es compleixen els requisits d'un t-test\n")
}
```

```
## No es pot aplicar el test de Wilcoxon: no hi ha prou dades o classes.
## Usant Wilcoxon perquè no es compleixen els requisits d'un t-test
```

```
print(sh_low)
```

```
## NULL
```

```
print(sh_high)
```

```
## NULL
```

```
print(lev)
```

```
## NULL
```

```
if (!is.null(test_hours)) print(test_hours)
```

El resultat del Wilcoxon comparant `hours_per_week` entre els dos grups (low vs. high) és:

- **W = 1 131 097**, **p < 2 × 10 ¹** (aprox. 1.49 × 10)
- **n = 2 822** (low), **n = 1 178** (high)

Com que **p 0,05**, rebutgem l'hipòtesi nul · la de distribucions iguals de `hours_per_week` entre els qui guanyen 50K i els que guanyen > 50K. Això vol dir que hi ha una diferència **estadísticament significativa** en nombre d'hores treballades setmanalment:

- Els ingressos més alts s'associen a **més hores treballades** (la mediana del grup "high" és superior a la del grup "low").

Així, podem concloure que dedicar més hores a la feina es relaciona amb una probabilitat més alta de pertànyer al grup de > 50K.

6 Representació de distribucions després de la neteja

Distribució de valors per a variables categòriques

```
adult_fac %>%
  select(where(is.factor)) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "valor") %>%
  count(variable, valor) %>%
  group_by(variable) %>%
  mutate(pct = round(n / sum(n) * 100, 2)) %>%
  arrange(variable, desc(n)) %>%
  print(n = 100)
```

```
## # A tibble: 71 x 4
## # Groups:   variable [10]
##    variable       valor                   n   pct
##    <chr>          <fct>               <int> <dbl>
##  1 age_group      30-39               12992 26.6
##  2 age_group      20-29               12486 25.6
##  3 age_group      40-49               10928 22.4
##  4 age_group      50-59                6400 13.1
##  5 age_group      60-69                2654  5.43
##  6 age_group      15-19                2573  5.27
##  7 age_group      70-79                 638  1.31
##  8 age_group      80-89                 120  0.25
##  9 age_group      90+                    51  0.1
## 10 education      HS-grad             15741 32.2
## 11 education      Some-college        11091 22.7
## 12 education      Bachelors            8187 16.8
## 13 education      Masters              2725  5.58
## 14 education      Assoc-voc            2162  4.43
## 15 education      11th                 1813  3.71
## 16 education      Assoc-acdm           1685  3.45
## 17 education      10th                 1313  2.69
## 18 education      7th-8th               818  1.67
## 19 education      Prof-school           734  1.5
## 20 education      9th                   701  1.44
## 21 education      12th                  603  1.23
## 22 education      Doctorate             551  1.13
## 23 education      5th-6th               439  0.9
## 24 education      1st-4th               211  0.43
## 25 education      Preschool              68  0.14
## 26 income         <=50K               38485 78.8
## 27 income         >50K                10357 21.2
## 28 marital_status Married-civ-spouse  18911 38.7
## 29 marital_status Never-married       17161 35.1
## 30 marital_status Divorced             8303 17
## 31 marital_status Widowed              1890  3.87
## 32 marital_status Separated            1874  3.84
## 33 marital_status Married-spouse-absent  664  1.36
## 34 marital_status Married-AF-spouse      39  0.08
## 35 native_country United-States       44717 91.6
## 36 native_country Other                3242  6.64
## 37 native_country Mexico                883  1.81
## 38 occupation     Adm-clerical          7593 15.6
## 39 occupation     Prof-specialty        6738 13.8
## 40 occupation     Exec-managerial       6404 13.1
## 41 occupation     Other-service         6046 12.4
## 42 occupation     Sales                 5978 12.2
## 43 occupation     Craft-repair          5119 10.5
## 44 occupation     Machine-op-inspct     3062  6.27
## 45 occupation     Transport-moving      1975  4.04
## 46 occupation     Handlers-cleaners     1872  3.83
## 47 occupation     Tech-support          1561  3.2
## 48 occupation     Farming-fishing       1272  2.6
## 49 occupation     Protective-serv        867  1.78
## 50 occupation     Priv-house-serv        346  0.71
```

```
## 51 occupation     Armed-Forces              9  0.02
## 52 race           White                 41609 85.2
## 53 race           Black                  5009 10.3
## 54 race           Asian-Pac-Islander     1354  2.77
## 55 race           Amer-Indian-Eskimo      479  0.98
## 56 race           Other                   391  0.8
## 57 relationship   Husband               15225 31.2
## 58 relationship   Not-in-family         14022 28.7
## 59 relationship   Own-child              7728 15.8
## 60 relationship   Unmarried              6930 14.2
## 61 relationship   Wife                   3344  6.85
## 62 relationship   Other-relative         1593  3.26
## 63 sex            Male                  24698 50.6
## 64 sex            Female                24144 49.4
## 65 workclass      Private               36437 74.6
## 66 workclass      Local-gov              3603  7.38
## 67 workclass      Self-emp-not-inc       3594  7.36
## 68 workclass      State-gov              2201  4.51
## 69 workclass      Federal-gov            1501  3.07
## 70 workclass      Self-emp-inc           1482  3.03
## 71 workclass      Without-pay              24  0.05
```

```r
# Estadístiques descriptives de les variables numèriques

adult_fac %>%
  select(where(is.numeric)) %>%
  summary()
```

```
##       age           fnlwgt        educational_num  capital_gain
##  Min.   :17.0   Min.   :  13769   Min.   : 1.00   Min.   :    0.0
##  1st Qu.:27.0   1st Qu.: 117551   1st Qu.: 9.00   1st Qu.:    0.0
##  Median :37.0   Median : 177905   Median :10.00   Median :    0.0
##  Mean   :38.1   Mean   : 189083   Mean   :10.13   Mean   :  993.2
##  3rd Qu.:47.0   3rd Qu.: 235860   3rd Qu.:12.00   3rd Qu.:    0.0
##  Max.   :90.0   Max.   :1490400   Max.   :16.00   Max.   :99999.0
##   capital_loss    hours_per_week    income_bin      net_capital
##  Min.   :   0.0   Min.   : 1.00   Min.   :0.0000   Min.   :-4356.0
##  1st Qu.:   0.0   1st Qu.:38.00   1st Qu.:0.0000   1st Qu.:    0.0
##  Median :   0.0   Median :40.00   Median :0.0000   Median :    0.0
##  Mean   :  81.6   Mean   :39.97   Mean   :0.2121   Mean   :  911.6
##  3rd Qu.:   0.0   3rd Qu.:45.00   3rd Qu.:0.0000   3rd Qu.:    0.0
##  Max.   :4356.0   Max.   :99.00   Max.   :1.0000   Max.   :99999.0
```

7 Codi

7.1 Publicació del codi

El codi desenvolupat per realitzar la neteja, transformació i anàlisi del conjunt de dades Adult Income es troba publicat al repositori GitHub següent:

https://github.com/Guillemromeu/PAC2-TCVD-Victor-Guillem

L'arxiu principal és Pràctica2.Rmd, ubicat dins la carpeta /codi, i conté tot el procés analític documentat: des de la càrrega de dades fins a les conclusions finals.

A més, el repositori inclou els fitxers següents:

README.md: descripció general del projecte, estructura i instruccions d'execució.

LICENSE: llicència del projecte per a ús educatiu.

/dades/adult.csv: dataset original extret del repositori UCI.

/dades/adult_net_final.csv: dataset final, netejat i imputat.

/informe/Memòria Pràctica 2 Víctor Olivera i Guillem Romeu.pdf: informe final generat a partir del codi.

8 Vídeo

Hem realitzat un vídeo de presentació titulat "Pràctica 2 - Vídeo explicatiu del projecte", en el qual es mostren els aspectes més rellevants de la pràctica. L'enregistrament està disponible al següent enllaç:

https://drive.google.com/drive/folders/1uxO3c8djVWcA67RFlMv0UHM9rHlYm2yr

Durant el vídeo, ambdós membres del grup participem activament, presentant:

- El context i objectiu del projecte, basat en el Adult Income Dataset del UCI.
- Les fases aplicades sobre el conjunt de dades: neteja, transformació, selecció i imputació.
- L'aplicació de models supervisats (regressió logística) i no supervisats (PAM).
- L'ús de proves estadístiques (Wilcoxon) amb comprovació prèvia d'assumpcions.
- Les conclusions extretes, incloent-hi la interpretació de resultats, limitacions i proposta de millora.
- Els criteris ètics considerats durant tot el procés, treballant amb dades públiques i anonimitzades.

Per fer la presentació, hem seguit el guió facilitat pel model al punt anterior i hem intentat transmetre tant el valor analític del projecte com el procés tècnic dut a terme.

9 Conclusions i propostes de millora

Aquest projecte ha permès aplicar de manera pràctica totes les fases del cicle de vida de les dades sobre un dataset real. El conjunt de dades *Adult Income* ha estat netejat, transformat i analitzat per predir si una persona guanya més de 50.000 $ anuals, a partir de variables socioeconòmiques i demogràfiques.

S'ha tractat amb èxit la presència de valors perduts mitjançant imputació per kNN, s'han unificat nivells amb baixa representació i s'han eliminat observacions influents amb distància de Cook. El dataset net s'ha usat per construir models supervisats (regressió logística) i no supervisats (clustering PAM), així com per realitzar anàlisis estadístiques inferencials amb proves no paramètriques.

9.1 Conclusions clau

- Variables com `educational_num`, `hours_per_week` i `net_capital` són les més predictives per estimar els ingressos.
- El model de regressió logística ha obtingut una bona precisió (accuracy) i una AUC robusta.
- El clustering PAM, tot i no tenir accés a la variable objectiu, ha pogut separar els grups amb una silueta mitjana acceptable.
- El test de Wilcoxon confirma que hi ha diferències significatives en les hores treballades entre persones amb ingressos baixos i alts.

9.2 Limitacions

- El dataset és antic (1994) i pot no reflectir la realitat socioeconòmica actual.
- La variable `fnlwgt` no ha estat utilitzada; tot i que pot ser rellevant a nivell poblacional, no aportava valor directe al model.
- Tot i la imputació, algunes variables categòriques poden conservar cert biaix.

9.3 Propostes de millora

- Aplicar tècniques de validació creuada (cross-validation) per avaluar millor el rendiment.
- Explorar altres models com random forests o XGBoost.
- Afegir variables derivades del país d'origen i estudiar el seu efecte.
- Fer una reflexió ètica més profunda sobre el biaix potencial de gènere o origen en els resultats.

10 Taula de contribucions

Contribucions i signatura

- Investigació prèvia: G.R.G., V.O.B.
- Redacció de les respostes: G.R.G., V.O.B.
- Desenvolupament del codi: G.R.G., V.O.B.
- Participació al vídeo: G.R.G., V.O.B.

G.R.G. = Guillem Romeu Graells
V.O.B. = Víctor Olivera Begue

11 Exportació del dataset final net

Exportem el dataset netejat amb imputació i transformacions finals

```
if (interactive()) {
  write.csv(adult_fac, "adult_net_final.csv", row.names = FALSE)
}
```