Universitat Oberta de Catalunya

Inteligencia de Negocio y Big Data Analytics

PEC Global: Un Caso integrador con datos reales

UOC - Master BI - Business Analytics (Nombre Estudiante)

Enero del 2020

1 Introducción

1.1 Descripción de la PEC a realizar

La prueba está estructurada en 9 ejercicios teórico-prácticos que revisan aquellos conceptos fundamentales que se han trabajado durante el semestre.

1.2 Criterios de evaluación

Ejercicios teóricos
Todos los ejercicios deben ser presentados de forma razonada y clara. No se aceptará ninguna respuesta que no esté claramente justificada.

Ejercicios prácticos
Para todas las PEC es necesario documentar en cada ejercicio práctico qué se ha hecho y cómo se ha hecho. Los criterios de valoración se describen en la siguiente tabla:

Pregunta	Criterio de valoración	Peso
1	Respuesta a la pregunta con justificación completa	5%
1	Propuestas de alternativas de tratamiento	5%
2	Gráfico que represente al menos 2 variables cualitativas	5%
2	Gráfico que represente al menos 2 variables cuantitativas	5%
3	Entrenamiento de los distintos modelos KNN	10%
3	Comparación de los resultados entre modelos KNN	5%
4	Entrenamiento y evaluación del modelo Tree	10%
4	Representación del tree	5%
5	Entrenamiento del modelo RandomForest	10%
5	Evaluación de modelo RandomForest entrenado	5%
6	Comparación de los modelos a partir del % de aciertos	5%
6	Comparación de los modelos a partir de la matriz de confusión del paquete “caret” e interpretación de resultados	5%
7	Entrenamiento del modelo Kmeans	10%
7	Identificación del cluster con más clientes	5%
8	Descripcción de los clusters	5%
9	Calcular los porcentajes de clientes con salarios superiores a 50K	5%

1.3 Formato y fecha de entega

El formato de entrega es: studentname-PECGlobal.html
Fecha de Entrega: 02/02/2020
Se debe entregar la PEC en el buzón de entregas del aula

2 Base teórica

Por tratarse de una PEC Global, esta práctica está basada en varias de las técnicas que se han analizado y usado a lo largo de la asignatura, justamente interesa que el estudiante aplique lo aprendido hasta este punto, razón por la cual será menos descriptiva en la teoría vista.

Se ha estructurado en formato de caso, en donde se expondrá un contexto, un problema de negocio plausible y un juego de datos reales precompilados con el que se debe dar solución a las cuestiones planteadas.

Por tratarse de un caso de negocio que debe resolverse mediante técnicas orientadas al descubrimiento de conocimiento, el desarrollo de esta PEC Global y solución se basan, a groso modo, en las fases que se describen en la metodología de gestión CRISP-DM (metodologías y estándares)

3 Problema y Contexto

Usted es un consultor y la empresa de XYZ Sociedad Anónima le contrata para que les asesore con el siguiente problema:

XYZ tiene negocios en Estados Unidos y ha tenido problemas con la segmentación que hace sobre los clientes que desea prospectar en este país y esto ha impactado directamente en su efectividad comercial. Dichos problemas se originan fundamentalmente en la ausencia de una información importante. Es bien sabido que una de las variables más importantes para segmentar o perfilar clientes es su ingreso, XYZ puede obtener los datos del censo de Estados Unidos, sin embargo, esto plantea el gran problema de que el dato es un insumo primario de su negocio que la empresa no controla en absoluto y debe esperar a que la oficina del censo de Estados Unidos actualice la información (lo hacen en múltiplos de 10 años), lo cual, considerando que el próximo censo será en 2020 representa un problema. Se ha intentado usar los datos no actualizados pero la condición de las personas cambia mucho en 10 años, por ejemplo, una persona podría terminar sus estudios y con eso incrementar su ingreso de manera importante.

Ante esta situación, el presidente de la compañía, que tiene algunas nociones básicas del potencial de los modelos predictivos, le contrata para lo siguiente:

La empresa XYZ tiene los datos del censo, la muestra contiene más de 30.000 personas y para cada una de ellas se recopilaron 15 variables distintas. Con base en esta información, se le encomienda que construya un modelo de aprendizaje automático que pueda predecir, en función dichas variables, si una persona gana más de 50.000 dólares al año o si no.

La intención de XYZ es usar este modelo cuando prospecte nuevos clientes para los cuales no se conozca el nivel de ingreso pero si el resto de las variables. Además, esperan en el proceso al menos un 85% de acierto general.

(Note que de manera simplificada y a efectos académicos, ésta sería la primera fase de la metodología CRISP-DM, el entendimiento del negocio y en adelante supondremos que la misma se completó adecuadamente)

4 INSTALACIÓN DE PAQUETES

Instalamos los packages de R que necesitaremos para realizar la práctica:

- install.packages("class")
- install.packages("rpart")
- install.packages("rpart.plot")
- install.packages("randomForest")
- install.packages("caret")
- install.packages("e1071")

Para asegurar que los resultados son reproducibles, fijamos la metodología de generación de semillas.

```
RNGversion('3.5.3')
```

5 Sobre los datos

El base de datos (dataset) de trabajo contiene las siguientes variables:

- fnlwgt: Identificador
- age: Edad en años
- type_employer: Tipo de empleado
- education: Nivel educativo
- education_num: Nivel educativo en años
- marital: Estado marital
- occupation: Ocupación
- relationship: Relación familiar
- race: Etnia
- sex: Género
- capital_gain: Capital ganado en el mercado financiero
- capital_loss: Capital perdido en el mercado financiero
- hr_per_week: Horas laboradas por semana
- country: País de procedencia
- income: Ingreso superior a 50 mil USD (S/N)

Fuente del dataset: UCI Machine Learning Repository [http://archive.ics.uci.edu/ml (http://archive.ics.uci.edu/ml)]. Irvine, CA: University of California, School of Information and Computer Science.

6 Descargar los datos desde internet

Dado que los datos que se requieren se encuentran publicados en internet, es factible efectuar una descarga desde el sitio en donde se encuentran alojados:

```
## Carga del Data Set desde la fuente Original
data = read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.
data",
    sep = ",", header = F, col.names = c("age", "type_employer", "fnlwgt",
        "education", "education_num", "marital", "occupation", "relationship",
        "race", "sex", "capital_gain", "capital_loss", "hr_per_week", "country",
        "income"), fill = FALSE, strip.white = T)

dim(data)
```

```
## [1] 32561    15
```

Mediante la instrucción anterior, se han descargado desde internet y cargado en R 32.561 registros, cada uno con 15 variables diferentes. Sin embargo, aún no cuentan con el nivel de calidad suficiente para su utilización, requieren ser explorados y posteriormente preparados.

7 Exploración y preparación de los datos

La exploración de los datos es la segunda fase de la metodología CRISP-DM y representa la primera aproximación a los datos disponibles, se tratará de entenderlos mejor a efectos de dimensionar su potencial y el posterior tratamiento que requerirán. La preparación de los datos es la tercera fase y a efectos prácticos para el presente caso se verá de manera agrupada con la fase de exploración únicamente como una simplificación.

- Podemos observar la cabecera de los datos (6 primeros registros) y un primer resumen de los mismos a efectos de darnos una primera noción de los datos:

```
head(data[1:6])
```

```
##   age    type_employer fnlwgt education education_num           marital
## 1  39        State-gov  77516 Bachelors            13      Never-married
## 2  50 Self-emp-not-inc  83311 Bachelors            13 Married-civ-spouse
## 3  38          Private 215646   HS-grad             9           Divorced
## 4  53          Private 234721      11th             7 Married-civ-spouse
## 5  28          Private 338409 Bachelors            13 Married-civ-spouse
## 6  37          Private 284582   Masters            14 Married-civ-spouse
```

```
head(data[7:11])
```

```
##          occupation  relationship  race    sex capital_gain
## 1      Adm-clerical Not-in-family White   Male         2174
## 2   Exec-managerial       Husband White   Male            0
## 3 Handlers-cleaners Not-in-family White   Male            0
## 4 Handlers-cleaners       Husband Black   Male            0
## 5    Prof-specialty          Wife Black Female            0
## 6   Exec-managerial          Wife White Female            0
```

```
head(data[12:15])
```

```
##   capital_loss hr_per_week       country income
## 1            0          40 United-States  <=50K
## 2            0          13 United-States  <=50K
## 3            0          40 United-States  <=50K
## 4            0          40 United-States  <=50K
## 5            0          40          Cuba  <=50K
## 6            0          40 United-States  <=50K
```

```
summary(data)
```

```
##       age                 type_employer       fnlwgt
##  Min.   :17.00   Private         :22696   Min.   :  12285
##  1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827
##  Median :37.00   Local-gov       : 2093   Median : 178356
##  Mean   :38.58   ?               : 1836   Mean   : 189778
##  3rd Qu.:48.00   State-gov       : 1298   3rd Qu.: 237051
##  Max.   :90.00   Self-emp-inc    : 1116   Max.   :1484705
##                  (Other)         :  981
##         education     education_num                   marital
##  HS-grad     :10501   Min.   : 1.00   Divorced             : 4443
##  Some-college: 7291   1st Qu.: 9.00   Married-AF-spouse    :   23
##  Bachelors   : 5355   Median :10.00   Married-civ-spouse   :14976
##  Masters     : 1723   Mean   :10.08   Married-spouse-absent:  418
##  Assoc-voc   : 1382   3rd Qu.:12.00   Never-married        :10683
##  11th        : 1175   Max.   :16.00   Separated            : 1025
##  (Other)     : 5134                   Widowed              :  993
##            occupation           relationship                   race
##  Prof-specialty :4140   Husband       :13193   Amer-Indian-Eskimo:  311
##  Craft-repair   :4099   Not-in-family : 8305   Asian-Pac-Islander: 1039
##  Exec-managerial:4066   Other-relative:  981   Black              : 3124
##  Adm-clerical   :3770   Own-child     : 5068   Other              :  271
##  Sales          :3650   Unmarried     : 3446   White              :27816
##  Other-service  :3295   Wife          : 1568
##  (Other)        :9541
##      sex         capital_gain    capital_loss     hr_per_week
##  Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##  Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
##                 Median :    0   Median :   0.0   Median :40.00
##                 Mean   : 1078   Mean   :  87.3   Mean   :40.44
##                 3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
##                 Max.   :99999   Max.   :4356.0   Max.   :99.00
##
##           country        income
##  United-States:29170   <=50K:24720
##  Mexico       :  643   >50K : 7841
##  ?            :  583
##  Philippines  :  198
##  Germany      :  137
##  Canada       :  121
##  (Other)      : 1709
```

Note que para imprimir la cabecera se han usado tres instrucciones, en realidad es innecesario y se hace únicamente a efectos de lograr la mejor visualización de la información en pantalla.

Observando los datos es posible concluir que hay dos variables que podrían ser excluidas: **fnlwgt** por tratarse de un identificador que a efectos de pronóstico no agrega ningón valor y **education_num** por ser una variable que redunda con otra, en este caso **education**. Para eliminarlas podemos usar el siguiente código:

```
data[["education_num"]]=NULL
data[["fnlwgt"]]=NULL
# Comprobamos que se han eliminado las variables indicadas
head(data)
```

```
##   age    type_employer education           marital        occupation
## 1  39        State-gov Bachelors      Never-married      Adm-clerical
## 2  50 Self-emp-not-inc Bachelors Married-civ-spouse   Exec-managerial
## 3  38          Private   HS-grad           Divorced Handlers-cleaners
## 4  53          Private      11th Married-civ-spouse Handlers-cleaners
## 5  28          Private Bachelors Married-civ-spouse    Prof-specialty
## 6  37          Private   Masters Married-civ-spouse   Exec-managerial
##    relationship  race    sex capital_gain capital_loss hr_per_week
## 1 Not-in-family White   Male         2174            0          40
## 2       Husband White   Male            0            0          13
## 3 Not-in-family White   Male            0            0          40
## 4       Husband Black   Male            0            0          40
## 5          Wife Black Female            0            0          40
## 6          Wife White Female            0            0          40
##         country income
## 1 United-States  <=50K
## 2 United-States  <=50K
## 3 United-States  <=50K
## 4 United-States  <=50K
## 5          Cuba  <=50K
## 6 United-States  <=50K
```

- Podemos observar también la estructura del dataset hasta este momento:

```
str(data)
```

```
## 'data.frame':    32561 obs. of  13 variables:
##  $ age          : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ type_employer: Factor w/ 9 levels "?","Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ education    : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10
...
##  $ marital      : Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4
3 5 3 ...
##  $ occupation   : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5
...
##  $ relationship : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2
1 ...
##  $ race         : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ capital_gain : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital_loss : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hr_per_week  : int  40 13 40 40 40 40 16 45 50 40 ...
##  $ country      : Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 6 40 24 40 40 40
...
##  $ income       : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...
```

Lo primero que se puede observar es que el dataset ahora posee 13 variables y no 15, esto como consecuencia de las que ya hemos eliminado en el punto anterior. Llama la atención que algunas variables cualitativas poseen una profundidad alta, es decir, poseen muchos niveles o, dicho de otra forma, la cantidad diferente de opciones posibles que contienen es alta, por ejemplo, observe las variables **education, occupation y country**.

- A continuación, se efectuarón algunas acciones para simplificar el data set en relación a la profundidad de las variables cualitativas:

```
# Se hace una copia de los datos originales
data2<-data

## Se convierten algunas variables para manipularlas adecuadamente
data$type_employer = as.character(data$type_employer)
data$occupation = as.character(data$occupation)
data$country = as.character(data$country)
data$education = as.character(data$education)
data$race = as.character(data$race)
data$marital = as.character(data$marital)

## En tipo de empleador Se agrupan algunas variables que tienen poca frecuencia y son si
milares

data$type_employer = gsub("^Federal-gov","Federal-Govt",data$type_employer)
data$type_employer = gsub("^Local-gov","Other-Govt",data$type_employer)
data$type_employer = gsub("^State-gov","Other-Govt",data$type_employer)
data$type_employer = gsub("^Private","Private",data$type_employer)
data$type_employer = gsub("^Self-emp-inc","Self-Employed",data$type_employer)
data$type_employer = gsub("^Self-emp-not-inc","Self-Employed",data$type_employer)
data$type_employer = gsub("^Without-pay","Not-Working",data$type_employer)
data$type_employer = gsub("^Never-worked","Not-Working",data$type_employer)

## En ocupación se pueden agrupar algunas para restarle profundidad a la variable

data$occupation = gsub("^Adm-clerical","Admin",data$occupation)
data$occupation = gsub("^Armed-Forces","Military",data$occupation)
data$occupation = gsub("^Craft-repair","Blue-Collar",data$occupation)
data$occupation = gsub("^Exec-managerial","White-Collar",data$occupation)
data$occupation = gsub("^Farming-fishing","Blue-Collar",data$occupation)
data$occupation = gsub("^Handlers-cleaners","Blue-Collar",data$occupation)
data$occupation = gsub("^Machine-op-inspct","Blue-Collar",data$occupation)
data$occupation = gsub("^Other-service","Service",data$occupation)
data$occupation = gsub("^Priv-house-serv","Service",data$occupation)
data$occupation = gsub("^Prof-specialty","Professional",data$occupation)
data$occupation = gsub("^Protective-serv","Other-Occupations",data$occupation)
data$occupation = gsub("^Sales","Sales",data$occupation)
data$occupation = gsub("^Tech-support","Other-Occupations",data$occupation)
data$occupation = gsub("^Transport-moving","Blue-Collar",data$occupation)

## En Country logicamente la mayoría de las observaciones son de USA pero hay mucha dive
rsidad de países y es interesante agruparlos por región:

data$country[data$country=="Cambodia"] = "SE-Asia"
data$country[data$country=="Canada"] = "British-Commonwealth"
data$country[data$country=="China"] = "China"
data$country[data$country=="Columbia"] = "South-America"
data$country[data$country=="Cuba"] = "Other"
data$country[data$country=="Dominican-Republic"] = "Latin-America"
data$country[data$country=="Ecuador"] = "South-America"
data$country[data$country=="El-Salvador"] = "South-America"
data$country[data$country=="England"] = "British-Commonwealth"
```

```
data$country[data$country=="France"] = "Euro_1"
data$country[data$country=="Germany"] = "Euro_1"
data$country[data$country=="Greece"] = "Euro_2"
data$country[data$country=="Guatemala"] = "Latin-America"
data$country[data$country=="Haiti"] = "Latin-America"
data$country[data$country=="Holand-Netherlands"] = "Euro_1"
data$country[data$country=="Honduras"] = "Latin-America"
data$country[data$country=="Hong"] = "China"
data$country[data$country=="Hungary"] = "Euro_2"
data$country[data$country=="India"] = "British-Commonwealth"
data$country[data$country=="Iran"] = "Other"
data$country[data$country=="Ireland"] = "British-Commonwealth"
data$country[data$country=="Italy"] = "Euro_1"
data$country[data$country=="Jamaica"] = "Latin-America"
data$country[data$country=="Japan"] = "Other"
data$country[data$country=="Laos"] = "SE-Asia"
data$country[data$country=="Mexico"] = "Latin-America"
data$country[data$country=="Nicaragua"] = "Latin-America"
data$country[data$country=="Outlying-US(Guam-USVI-etc)"] = "Latin-America"
data$country[data$country=="Peru"] = "South-America"
data$country[data$country=="Philippines"] = "SE-Asia"
data$country[data$country=="Poland"] = "Euro_2"
data$country[data$country=="Portugal"] = "Euro_2"
data$country[data$country=="Puerto-Rico"] = "Latin-America"
data$country[data$country=="Scotland"] = "British-Commonwealth"
data$country[data$country=="South"] = "Euro_2"
data$country[data$country=="Taiwan"] = "China"
data$country[data$country=="Thailand"] = "SE-Asia"
data$country[data$country=="Trinadad&Tobago"] = "Latin-America"
data$country[data$country=="United-States"] = "United-States"
data$country[data$country=="Vietnam"] = "SE-Asia"
data$country[data$country=="Yugoslavia"] = "Euro_2"

## En educación, también se agrupan algunos, la idea es restarle profundidad

data$education = gsub("^10th","Dropout",data$education)
data$education = gsub("^11th","Dropout",data$education)
data$education = gsub("^12th","Dropout",data$education)
data$education = gsub("^1st-4th","Dropout",data$education)
data$education = gsub("^5th-6th","Dropout",data$education)
data$education = gsub("^7th-8th","Dropout",data$education)
data$education = gsub("^9th","Dropout",data$education)
data$education = gsub("^Assoc-acdm","Associates",data$education)
data$education = gsub("^Assoc-voc","Associates",data$education)
data$education = gsub("^Bachelors","Bachelors",data$education)
data$education = gsub("^Doctorate","Doctorate",data$education)
data$education = gsub("^HS-Grad","HS-Graduate",data$education)
data$education = gsub("^Masters","Masters",data$education)
data$education = gsub("^Preschool","Dropout",data$education)
data$education = gsub("^Prof-school","Prof-School",data$education)
data$education = gsub("^Some-college","HS-Graduate",data$education)
```

```
## De igual forma se agrupan los estados maritales

data$marital[data$marital=="Never-married"] = "Never-Married"
data$marital[data$marital=="Married-AF-spouse"] = "Married"
data$marital[data$marital=="Married-civ-spouse"] = "Married"
data$marital[data$marital=="Married-spouse-absent"] = "Not-Married"
data$marital[data$marital=="Separated"] = "Not-Married"
data$marital[data$marital=="Divorced"] = "Not-Married"
data$marital[data$marital=="Widowed"] = "Widowed"


## La etnia se cambia para que sea más fácil de leer

data$race[data$race=="White"] = "White"
data$race[data$race=="Black"] = "Black"
data$race[data$race=="Amer-Indian-Eskimo"] = "Amer-Indian"
data$race[data$race=="Asian-Pac-Islander"] = "Asian"
data$race[data$race=="Other"] = "Other"

## Se regresan a factores las variables categóricas

data$marital = factor(data$marital)
data$education = factor(data$education)
data$country = factor(data$country)
data$type_employer = factor(data$type_employer)
data$occupation = factor(data$occupation)
data$race = factor(data$race)
data$sex = factor(data$sex)
data$relationship = factor(data$relationship)

## Se recodifica la variable a predecir a S/N
data$income = as.factor(ifelse(data$income==data$income[1],"N","S"))
```

A este punto se han reducido las variables cualitativas a versiones que agrupan más información al tener menor profundidad:

```
str(data)
```

```
## 'data.frame':    32561 obs. of  13 variables:
##  $ age          : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ type_employer: Factor w/ 6 levels "?","Federal-Govt",..: 4 6 5 5 5 5 5 6 5 5 ...
##  $ education    : Factor w/ 8 levels "Associates","Bachelors",..: 2 2 5 4 2 7 4 5 7 2
...
##  $ marital      : Factor w/ 4 levels "Married","Never-Married",..: 2 1 3 1 1 1 3 1 2
1 ...
##  $ occupation   : Factor w/ 9 levels "?","Admin","Blue-Collar",..: 2 9 3 3 6 9 8 9 6
9 ...
##  $ relationship : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2
1 ...
##  $ race         : Factor w/ 5 levels "Amer-Indian",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ capital_gain : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital_loss : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hr_per_week  : int  40 13 40 40 40 40 16 45 50 40 ...
##  $ country      : Factor w/ 10 levels "?","British-Commonwealth",..: 10 10 10 10 7 10
6 10 10 10 ...
##  $ income       : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 2 2 2 ...
```

Al generar nuevamente la estructura del dataset, se pueden observar variables cualitativas mucho menos amplias, por ejemplo la nueva variable **Country** tiene sólo 10 categorías posibles mientras su versión original tenía 42:

```
as.data.frame(table(data$country))
```

```
##                    Var1  Freq
## 1                     ?   583
## 2  British-Commonwealth   347
## 3                 China   146
## 4                Euro_1   240
## 5                Euro_2   235
## 6         Latin-America  1096
## 7                 Other   200
## 8              SE-Asia   320
## 9         South-America   224
## 10        United-States 29170
```

```
as.data.frame(table(data2$country))
```

```
##                          Var1  Freq
## 1                           ?   583
## 2                    Cambodia    19
## 3                      Canada   121
## 4                       China    75
## 5                    Columbia    59
## 6                        Cuba    95
## 7          Dominican-Republic    70
## 8                     Ecuador    28
## 9                 El-Salvador   106
## 10                    England    90
## 11                     France    29
## 12                    Germany   137
## 13                     Greece    29
## 14                  Guatemala    64
## 15                      Haiti    44
## 16         Holand-Netherlands     1
## 17                   Honduras    13
## 18                       Hong    20
## 19                    Hungary    13
## 20                      India   100
## 21                       Iran    43
## 22                    Ireland    24
## 23                      Italy    73
## 24                    Jamaica    81
## 25                      Japan    62
## 26                       Laos    18
## 27                     Mexico   643
## 28                  Nicaragua    34
## 29 Outlying-US(Guam-USVI-etc)    14
## 30                       Peru    31
## 31                Philippines   198
## 32                     Poland    60
## 33                   Portugal    37
## 34                Puerto-Rico   114
## 35                   Scotland    12
## 36                      South    80
## 37                     Taiwan    51
## 38                   Thailand    18
## 39            Trinadad&Tobago    19
## 40              United-States 29170
## 41                    Vietnam    67
## 42                  Yugoslavia    16
```

- Tratamiento de las variables cuantitativas: Si observamos las variables cuantitativas, en particular **Capital_gain** y **Capital_loss**, dado que tienen un sesgo importante, debido a que no todas las personas juegan en el mercado de valores, es posible simplificarlas si aplicamos un proceso de discretización, mediante el cual se convertirón a categóricas:

```
# La función cut divide el rango de x en intervalos y codifica los valores en x según el intervalo que caen. El intervalo más a la izquierda corresponde al nivel uno, el siguiente más a la izquierda al nivel dos y así sucesivamente.

data[["capital_gain"]] <- ordered(cut(data$capital_gain,c(-Inf, 0,
                                                          median(data[["capital_gain"]]
[data[["capital_gain"]] >0]),
                                                          Inf)),labels = c("None", "Low"
, "High"))
data[["capital_loss"]] <- ordered(cut(data$capital_loss,c(-Inf, 0,
                                                          median(data[["capital_loss"]]
[data[["capital_loss"]] >0]),
                                                        Inf)), labels = c("None", "Low",
"High"))
# Se muestra la distribución de valores
table(data[,"capital_gain"])
```

```
##
##  None   Low  High
## 29849  1559  1153
```

```
table(data[,"capital_loss"])
```

```
##
##  None   Low  High
## 31042   782   737
```

En el caso de las variables **age** y **hr_per_week** pueden ser normalizadas para evitar que las técnicas que usan matrices de distancias se vean afectadas:

```
## La edad y los horas por semana se reescalan, centradas y reducidas
data$age = scale(data$age)
data$hr_per_week = scale(data$hr_per_week)
# Resumen de las variables
summary(data$age)
```

```
##        V1
##  Min.   :-1.5822
##  1st Qu.:-0.7758
##  Median :-0.1160
##  Mean   : 0.0000
##  3rd Qu.: 0.6905
##  Max.   : 3.7696
```

```
summary(data$hr_per_week)
```

```
##        V1
##  Min.   :-3.19398
##  1st Qu.:-0.03543
##  Median :-0.03543
##  Mean   : 0.00000
##  3rd Qu.: 0.36951
##  Max.   : 4.74289
```

- Tratamiento de los valores Nulos: Si no tenemos cuidado a este punto podríamos cometer un error cuando intentamos observar la cantidad de nulos en el data set:

```
# Suma la cantidad de nulos
sum(is.na(data))
```

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

Parece que no hay nulos, sin embargo si prestamos atención veremos que se encuentran “disfrazados”:

```
# Visualizamos un recunto de la variable country
as.data.frame(table(data$country))
```

```
##                    Var1  Freq
## 1                     ?   583
## 2  British-Commonwealth   347
## 3                 China   146
## 4               Euro_1   240
## 5               Euro_2   235
## 6         Latin-America  1096
## 7                 Other   200
## 8              SE-Asia   320
## 9         South-America   224
## 10        United-States 29170
```

Efectivamente, en nuestro juego de datos los nulos han sido reemplazados por un valor centinela, en este caso tenemos 583 registros con “?” en esta variable. Esto resulta problemático ya que no nos permite dimensionar ni lidiar adecuadamente con los nulos, dado esto, procederemos a reemplazarlos:

```
# Reemplazamos "?" con null y posteriormente los contmos
is.na(data) = data=='?'
is.na(data) = data==' ?'
# Se eliminan los registros con datos nulos
data = na.omit(data)
# Se evalua la cantidad de resgistros eliminados
dim(data)
```

```
## [1] 30162    13
```

```
dim(data2)
```

```
## [1] 32561    13
```

```
dim(data2)[1]-dim(data)[1]
```

```
## [1] 2399
```

```
(dim(data2)[1]-dim(data)[1])/dim(data2)[1]
```

```
## [1] 0.0736771
```

En realidad, tenemos más de 2 mil registros con valores nulos, lo cual implica un 7.37% valores perdidos y aunque es posible aplicar técnicas de imputación, orientadas a crear información para rellenar estos valores perdidos, en una simplificación y dado que el dataset contiene bastante información no nula, procederemos a eliminar cualquier registro con valores perdidos, es decir, trabajaremos con un 92.63% de los registros originales.

Tras eliminar los resgistros con nulos nos queda un dataset mucho más preparado para trabajar y aun con una buena cantidad de información:

```
dim(data)
```

```
## [1] 30162    13
```

más de 30 mil sujetos y para describir a cada uno 13 variables distintas ya reducidas.

---

8 Fase de modelado

- Split de los datos: Aleatoriamente seccionamos los datos en dos conjuntos, el primero que denominaremos trainData será el 70% de los datos y lo usaremos a efectos de entrenar los modelos, el segundo, testData será el 30% de los datos que nos reservaremos para poder verificar la calidad predictiva de los modelos:

```
# Split aleatorio de los datos para definir conjunto de entrenamiento / prueba
set.seed(1234)
ind <- sample(2, nrow(data), replace=TRUE, prob=c(0.7, 0.3))
trainData <- data[ind==1,]
dim(trainData)[1]/dim(data)[1]
```

```
## [1] 0.6990253
```

```
testData <- data[ind==2,]
dim(testData)[1]/dim(data)[1]
```

```
## [1] 0.3009747
```

```
str(trainData)
```

```
## 'data.frame':    21084 obs. of  13 variables:
##  $ age          : num [1:21084, 1] 0.0307 0.8371 -0.0426 1.057 -0.116 ...
##  $ type_employer: Factor w/ 6 levels "?","Federal-Govt",..: 4 6 5 5 5 5 6 5 5 5 ...
##  $ education    : Factor w/ 8 levels "Associates","Bachelors",..: 2 2 5 4 7 4 5 7 2 6
...
##  $ marital      : Factor w/ 4 levels "Married","Never-Married",..: 2 1 3 1 1 3 1 2 1
1 ...
##  $ occupation   : Factor w/ 9 levels "?","Admin","Blue-Collar",..: 2 9 3 3 9 8 9 6 9
9 ...
##  $ relationship : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 2 1 2 1
1 ...
##  $ race         : Factor w/ 5 levels "Amer-Indian",..: 5 5 5 3 5 3 5 5 5 3 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 2 1 2 2 ...
##  $ capital_gain : Ord.factor w/ 3 levels "None"<"Low"<"High": 2 1 1 1 1 1 1 3 2 1 ...
##  $ capital_loss : Ord.factor w/ 3 levels "None"<"Low"<"High": 1 1 1 1 1 1 1 1 1 1 ...
##  $ hr_per_week  : num [1:21084, 1] -0.0354 -2.2221 -0.0354 -0.0354 -0.0354 ...
##  $ country      : Factor w/ 10 levels "?","British-Commonwealth",..: 10 10 10 10 10 6
10 10 10 10 ...
##  $ income       : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 2 2 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int  15 28 39 52 62 70 78 94 107 129 ...
##   ..- attr(*, "names")= chr  "15" "28" "39" "52" ...
```

9 Preguntas de evaluación

9.1 Pregunta 1:

¿Considera que es importante efectuar la preparación de los datos? ¿Por qué? Investigue al menos dos alternativas que pudimos haber usado para darle tratamiento a los nulos.

9.2 Respuesta 1:

Escriba su respuesta aquí

```
#Escriba su código aquí
```

9.3 Pregunta 2:

Realizar dos gráficos de los datos en los que pueda realizarse un analisis previo del poder discriminantes sobre tipo de renta del resto de variables en la base de datos. El primero que contenga las dos variables cuantitativas **age** y **hr_per_week** y el segundo que esté asociado a una variable categórica.

9.4 Respuesta 2:

Escriba su respuesta aquí

```
#Escriba su código aquí
```

9.5 Pregunta 3:

Aplicar un modelo KNN al juego de datos de esta práctica utilizando como variables en imput la edad, las horas trabajadas por semana y el sexo. Recordad que para poder añadir el sexo tendréis que expresarla como variable “cuantitativa” de presencia y ausencia. Valorar la predicción utilizando distintos números de vecinos entre 20 y 30. Comentad los resultados.

Pruebe añadir otras variables categóricas en forma de binarias, además del sexo, y analice como mejora la predicción.

9.6 Respuesta 3:

Escriba su respuesta aquí

```
#Escriba su código aquí
```

9.7 Pregunta 4:

Aplicar un modelo de árbol de decisión al juego de datos de esta práctica y evaluar sus predicciones. Representar gráficamente el árbol obtenido.

9.8 Respuesta 4:

Escriba su respuesta aquí

```
#Escriba su código aquí
```

9.9 Pregunta 5:

Aplicar un modelo randomForest al juego de datos de esta práctica y evaluar sus predicciones.

9.10 Respuesta 5:

Escriba su respuesta aquí

```
#Escriba su código aquí
```

9.11 Pregunta 6:

Evalua los modelos anteriores:
1. Calcula el porcentage de acierto de cada modelo.
2. Obtener las respectivas matrices de confusión utilizando la función confusionMatrix() del paquete “caret”.

3. Visualizar los modelos generados.

9.12 Respuesta 6:

Escriba su respuesta aquí

```
#Escriba su código aquí
```

9.13 Pregunta 7:

Considere un dataset data3 formado por las variables capital_gain, capital_loss, hr_per_week y age

```
data3=data2[,c("capital_gain","capital_loss","hr_per_week","age")]
```

Realice una segmentación kmeans para 5 clusters. ¿Tienen todos los clusters el mismo número de clientes?.

Observación: normalice los datos antes de crear la segmentación.

9.14 Respuesta 7:

Escriba su respuesta aquí

```
#Escriba su código aquí
```

9.15 Pregunta 8:

Utilizando la segmentación realizada en el ejercicio anterior. Describa los perfiles de clientes representados por los clusters.

9.16 Respuesta 8:

Para poder interpretar los clusters es habitual representarlos mediante el valor medio de las variables Escriba su respuesta aquí

```
#Escriba su código aquí
```

9.17 Pregunta 9:

Utilizando la segmentación utilizada en el ejercicio anterior. ¿Cuál es el cluster con el mayor porcentaje de clientes con salarios superiores a 50k?

9.18 Respuesta 9:

Escriba su respuesta aquí

```
#Escriba su código aquí
```