

PRÁCTICA 2: Limpieza y validación de los datos

Autores: Anna Llorens Roig, Carlos Villar Robles

Dataset: Adults

Contents

1. Introducción	2
2. Descripción del dataset	2
3. Integración y selección de los datos de interés.	2
4. Limpieza de los datos.	3
5. Análisis de los datos.	5
6. Representación de los resultados a partir de tablas y gráficas	7
7. Resolución del problema.	7
8. Contribuciones	7

1. Introducción

El conjunto de datos objeto de análisis es de la base de datos ‘Adult’ del Censo de 1994 en Estados Unidos. Los detalles de este conjunto de datos se pueden encontrar en el repositorio de UCI Machine Learning: <https://archive.ics.uci.edu/ml/datasets/Adult>.

Durante el desarrollo de esta práctica trataremos de construir un modelo para predecir si el ingreso de cualquier individuo en los Estados Unidos es mayor o menor que USD 50000 según la información disponible sobre ese individuo en los datos del censo. Nos interesa conocer qué tan bien se puede predecir si el ingreso anual de una persona supera los 50000\$ utilizando el conjunto de variables en este conjunto de datos. La pregunta se inspecciona en dos enfoques: técnicas de aprendizaje automático, visualización de datos y modelado estadístico tradicional.

2. Descripción del dataset

Se trata del ‘Adult data set’ el cual consta de 15 atributos y 32561 observaciones. Entre los campos del conjunto de datos encontramos las siguientes variables dependientes:

- **age**: edad del individuo
- **type_employer**: tipo de empleador que tiene el individuo. Ya sean gubernamentales, militares, privados, etc.
- **fnlwgt**: El # de personas que los encuestados creen que representa la observación. Ignoraremos esta variable
- **education**: nivel más alto de educación alcanzado para esa persona
- **education_num**: nivel más alto de educación en forma numérica
- **marital**: estado civil del individuo
- **occupation**: ocupación del individuo
- **relationship**: contiene valores de relaciones familiares como marido, padre, etc., pero solo contiene uno por observación
- **race**: descripciones de la raza individual. Negro, blanco, esquimal,
- **sex**: sexo del individuo
- **capital_gain**: ganancias de capital registradas
- **capital_loss**: pérdidas de capital registradas
- **hr_per_week**: horas trabajadas por semana
- **country**: país de origen del individuo

Cómo variable dependiente del dataset tenemos:

- **income**: variable booleana. Representa si la persona gana o no más de \$ 50,000 por año de ingresos.

3. Inegración y selección de los datos de interés.

Antes de comenzar con la limpieza de los datos, procedemos a realizar la lectura del fichero en formato CSV en el que se encuentran. El resultado devuelto por la llamada a la función `read.csv()` será un objeto `data.frame` el cual mostraremos su cabecera:

```
# Lectura de datos
data <- read.csv('../data/adults.csv', header = TRUE)

# Comprobamos que los datos se han importado correctamente
head(data[,1:5])

##   age      workclass fnlwgt education education_num
## 1  39      State-gov  77516 Bachelors             13
## 2  50 Self-emp-not-inc 83311 Bachelors             13
```

```
## 3 38 Private 215646 HS-grad 9
## 4 53 Private 234721 11th 7
## 5 28 Private 338409 Bachelors 13
## 6 37 Private 284582 Masters 14
```

Un resumen de alto nivel de los datos se encuentra a continuación. Todas las variables han sido leídas en sus clases esperadas.

```
str(data)
```

```
## 'data.frame': 32561 obs. of 15 variables:
## $ age : int 39 50 38 53 28 37 49 52 31 42 ...
## $ workclass : Factor w/ 9 levels "?","Federal-gov",...: 8 7 5 5 5 5 7 5 5 ...
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
## $ education : Factor w/ 16 levels "10th","11th",...: 10 10 12 2 10 13 7 12 13 10 ...
## $ education_num : int 13 13 9 7 13 14 5 9 14 13 ...
## $ maritalStatus : 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 "F","M": 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 ...
## $ hour_per_week : int 40 13 40 40 40 40 16 45 50 40 ...
## $ native_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 ...
```

En primer lugar seleccionaremos las variables que las que nos centraremos para realizar el análisis. Por lo que eliminaremos las variables *education*, *fnlwgt*, *relationship*, *capital_gain*, *capital-loss*, *race*

```
# Eliminamos variables
data$education <- NULL
data$fnlwgt <- NULL
data$relationship <- NULL
data$capital_gain <- NULL
data$capital_loss <- NULL
data$native_country <- NULL
data$race <- NULL
```

4. Limpieza de los datos.

Una vez tenemos el conjunto de datos con el que trabajaremos observamos el resumen del conjunto y pasaremos a la limpieza de estos.

```
summary(data)
```

```
##      age      workclass      education_num
## Min.   :17.00 Private      :22696 Min.   : 1.00
## 1st Qu.:28.00 Self-emp-not-inc: 2541 1st Qu.: 9.00
## Median :37.00 Local-gov    : 2093 Median :10.00
## Mean   :38.58 ?           : 1836 Mean   :10.08
## 3rd Qu.:48.00 State-gov    : 1298 3rd Qu.:12.00
## Max.   :90.00 Self-emp-inc  : 1116 Max.   :16.00
##      (Other)           : 981
##      maritalStatus      occupation      sex
## Divorced                : 4443 Prof-specialty :4140 F:10771
## Married-AF-spouse       : 23 Craft-repair    :4099 M:21790
```

```
## Married-civ-spouse :14976 Exec-managerial:4066
## Married-spouse-absent: 418 Adm-clerical :3770
## Never-married :10683 Sales :3650
## Separated : 1025 Other-service :3295
## Widowed : 993 (Other) :9541
## hour_per_week income
## Min. : 1.00 <=50K:24720
## 1st Qu.:40.00 >50K : 7841
## Median :40.00
## Mean :40.44
## 3rd Qu.:45.00
## Max. :99.00
##
```

4.1 Ceros y elementos vacíos

Tal y como se indica en la descripción del dataset se utiliza el carácter ‘?’ para denotar un valor desconocido. Así, se procede a conocer a continuación qué campos contienen elementos vacíos:

```
# Números de valores desconocidos por campo
colSums(data=="?")
```

```
##          age      workclass education_num maritalStatus      occupation
##          0         1836             0             0             1843
##          sex hour_per_week      income
##          0             0             0
```

Llegados a este punto debemos decidir cómo manejar estos registros que contienen valores desconocidos para algún campo. Al disponer de un conjunto de datos relativamente grande (más de 32000 observaciones) procederemos a eliminar las columnas con registros nulos.

```
data<-data[!(data$workclass=="?"),]
data<-data[!(data$occupation=="?"),]
```

```
# Comprobamos valores nulos
colSums(data=="?")
```

```
##          age      workclass education_num maritalStatus      occupation
##          0             0             0             0             0
##          sex hour_per_week      income
##          0             0             0
```

4.2 Identificación y tratamiento de valores extremos

Los valores extremos o outliers son aquellos que parecen no ser congruentes sin los comparamos con el resto de los datos. Para identificarlos, podemos hacer uso de dos vías: (1) representar un diagrama de caja por cada variable y ver qué valores distan mucho del rango intercuartílico (la caja) o (2) utilizar la función `boxplots.stats()` de R, la cual se emplea a continuación. Así, se mostrarán sólo los valores atípicos para aquellas variables que los contienen:

```
boxplot.stats(data$age)$out
```

```
## [1] 79 76 90 77 76 81 78 90 88 90 77 90 77 78 80 90 81 81 76 80 90 76 79
## [24] 76 81 76 90 76 90 80 90 90 79 78 79 84 90 77 80 77 90 81 83 84 79 76
## [47] 85 82 79 77 90 76 90 84 78 78 76 80 90 90 77 76 84 76 90 76 90 76 77
## [70] 81 90 77 78 77 81 78 82 81 77 76 80 90 80 84 82 78 79 76 90 84 90 83
## [93] 78 80 77 78 76 79 80 79 80 90 90 90 90 81 76 83 90 90 81 80 80 90 79
## [116] 77 77 80 76 82 85 80 79 90 76 76 77 76 79 81 77 88 90 82 76 88 76 77
```

```
## [139] 83 76 77 79 77 86 90 77 82 83 81 76 79 76 84 78 76 76 76 78 84 79 78
## [162] 90 80 81 78 81 90 80 82 90 90 85
```

```
boxplot.stats(data$education_num)$out
```

```
## [1] 1 2 2 2 2 1 2 2 2 2 2 2 2 2 1 1 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2
## [36] 1 1 1 2 2 2 2 2 1 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 2 2 2 2 1 2 1 2 2
## [71] 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 1 2 2 2 2 2 1 2 2
## [106] 2 2 2 2 2 2 2 2 1 2 2 2 1 1 2 2 1 1 2 2 1 2 2 1 2 2 2 2 2 1 2 1 1
## [141] 2 1 1 1 2 1 2 1 2 2 1 1 2 2 2 2 1 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2
## [176] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 1 2 1 2 2 1
```

No obstante, si revisamos los anteriores datos para varios adultos escogidos aleatoriamente, comprobamos que son valores que perfectamente pueden darse (La edad de un individuo puede ser de 90 años y los niveles de educación pueden ser 1 o 2). Es por ello que el manejo de estos valores extremos consistirá en simplemente dejarlos como actualmente están recogidos.

5. Análisis de los datos.

5.1 Selección de los grupos de datos que se quieren analizar

Para simplificar el conjunto de valores de las variables categóricas vamos agrupar por categorías los siguientes atributos: *workclass*, *maritalStatus*, *occupation*

- Para la variable **workclass** diferenciaremos entre: government, public, public, self-employed, other

```
# Observamos atributos originales para la variable workclass
summary(data$workclass)
```

```
##           ?           Federal-gov           Local-gov           Never-worked
##           0           960           2093           0
##      Private      Self-emp-inc Self-emp-not-inc           State-gov
##      22696           1116           2541           1298
##      Without-pay
##           14
```

```
levels(data$workclass)[1] <- 'Unknown'
# combine into Government job
data$workclass <- gsub('^Federal-gov', 'Government', data$workclass)
data$workclass <- gsub('^Local-gov', 'Government', data$workclass)
data$workclass <- gsub('^State-gov', 'Government', data$workclass)
data$workclass <- gsub('^Government', 'Public', data$workclass)
# combine into Sele-Employed job
data$workclass <- gsub('^Self-emp-inc', 'Self-Employed', data$workclass)
data$workclass <- gsub('^Self-emp-not-inc', 'Self-Employed', data$workclass)
# combine into Other/Unknown
data$workclass <- gsub('^Never-worked', 'Other', data$workclass)
data$workclass <- gsub('^Without-pay', 'Other', data$workclass)
data$workclass <- gsub('^Other', 'Other', data$workclass)
data$workclass <- gsub('^Unknown', 'Other', data$workclass)
data$workclass <- as.factor(data$workclass)
```

```
# Observamos nuevos atributos para la variable workclass
summary(data$workclass)
```

```
##      Other      Private      Public Self-Employed
##      14      22696      4351      3657
```

- Para la variable **maritalStatus** diferenciaremos entre: married, single

```
# Observamos atributos originales para la variable maritalStatus
summary(data$maritalStatus)
```

```
##           Divorced      Married-AF-spouse      Married-civ-spouse
##           4258             21             14339
## Married-spouse-absent      Never-married      Separated
##           389             9912             959
##           Widowed
##           840
```

```
data$maritalStatus <- gsub('Married-AF-spouse', 'Married', data$maritalStatus)
data$maritalStatus <- gsub('Married-civ-spouse', 'Married', data$maritalStatus)
data$maritalStatus <- gsub('Married-spouse-absent', 'Married', data$maritalStatus)
data$maritalStatus <- gsub('Never-married', 'Single', data$maritalStatus)
data$maritalStatus <- gsub('Widowed', 'Single', data$maritalStatus)
data$maritalStatus <- gsub('Divorced', 'Single', data$maritalStatus)
data$maritalStatus <- gsub('Separated', 'Single', data$maritalStatus)
data$maritalStatus <- as.factor(data$maritalStatus)
```

```
# Observamos nuevos atributos para la variable maritalStatus
summary(data$maritalStatus)
```

```
## Married  Single
##   14749   15969
```

- Para la variable **occupation** diferenciaremos entre: Adm-clerical, Blue-Collar, Other/Unknown, Professional, Sales, Service, White-Collar

```
# Observamos atributos originales para la variable occupation
summary(data$occupation)
```

```
##           ?      Adm-clerical      Armed-Forces      Craft-repair
##           0      3770             9             4099
## Exec-managerial      Farming-fishing      Handlers-cleaners      Machine-op-inspct
##           4066             994             1370             2002
##      Other-service      Priv-house-serv      Prof-specialty      Protective-serv
##           3295             149             4140             649
##           Sales      Tech-support      Transport-moving
##           3650             928             1597
```

```
levels(data$occupation)[1] <- 'Unknown'
data$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', 'Service', data$occupation)
data$occupation <- gsub('Tech-support', 'Service', data$occupation)
data$occupation <- gsub('Transport-moving', 'Blue-Collar', data$occupation)
data$occupation <- gsub('Unknown', 'Other/Unknown', data$occupation)
data$occupation <- gsub('Armed-Forces', 'Other/Unknown', data$occupation)
data$occupation <- as.factor(data$occupation)
```

```
# Observamos nuevos atributos para la variable occupation
summary(data$occupation)
```

```
## Adm-clerical   Blue-Collar Other/Unknown Professional      Sales
##           3770           10062              9           4140           3650
##      Service White-Collar
##           5021           4066
```

```
# Observamos el conjunto de datos una vez han sido procesados
head(data)
```

```
##   age      workclass education_num maritalStatus  occupation sex
## 1  39      Public           13      Single Adm-clerical   M
## 2  50 Self-Employed          13    Married White-Collar   M
## 3  38      Private           9      Single Blue-Collar   M
## 4  53      Private           7    Married Blue-Collar   M
## 5  28      Private          13    Married Professional   F
## 6  37      Private          14    Married White-Collar   F
##  hour_per_week income
## 1           40  <=50K
## 2           13  <=50K
## 3           40  <=50K
## 4           40  <=50K
## 5           40  <=50K
## 6           40  <=50K
```

5.1.1 Exportación de los datos

Llegados a este punto, vamos a exportar el conjunto de datos con el vamos a realizar el análisis en un nuevo fichero al que denominaremos 'adults_clean.csv'.

```
# Exportación de los datos limpios en .csv
write.csv(data, "../data/adults_clean.csv")
```

6. Representación de los resultados a partir de tablas y gráficas

7. Resolución del problema.

8. Contribuciones

Contribuciones	Firma
Investigación previa	AL, CV
Introducción	AL, CV
Descripción	AL, CV
Integración	AL, CV
Limpieza de datos	AL, CV
Análisis de datos	AL, CV