PEC 4

Adrián Valls Carbó y Javier Herrero Martín

2023-01-07

1.- Descripción del DataSet

Este conjunto de datos contiene información de una muestra extraída a partir de un censo estadounidense, en el que para cada persona (sin datos personales), se registran los salarios aparte de información personal adicional. El conjunto de datos contiene 32.560 registros y 15 variables.

Las variables de esta muestra son:

- `age`: Edad del individuo.
- `workclass`: Categorización del individuo en base al perfil laboral.
- `fnlwgt`: Peso asignado a cada fila, refleja la proporción de datos que se asimilan dentro de la misma línea (misma información)
- `education`: Nivel de formación educativa del individuo.
- `education.num`: Número de años de formación educativa del individuo.
- `marital.status`: Estado civil del individuo.
- `occupation`: Categorización del individuo en base a la tipología de trabajo.
- `relationship`: Estado civil del individuo (a diferencia de marital_status, también hace referencia a hijos)
- `race`: Grupo racial al que pertenece el individuo.
- `sex`: Género del individuo.
- `capital.gain`: Ganancias capitales del individuo €.
- `capital.loss`: Pérdidas capitales del individuo €.
- `native.country`: País de procedencia del individuo
- `hours.per.week`: Horas por semana trabajadas por el individuo.
- `income`: Salario (anual) del individuo, en k€, hace referencia a un umbral de salario.

Vamos a cargar el dataset para un primer vistazo de las variables.

```
adult<-read.delim("./adult.csv", header = TRUE, sep = ",", dec
= ".")

head(adult)
```

```
##   age workclass fnlwgt    education education.num marital.status
## 1  90         ?  77053      HS-grad             9        Widowed
## 2  82   Private 132870      HS-grad             9        Widowed
## 3  66         ? 186061 Some-college            10        Widowed
## 4  54   Private 140359      7th-8th             4       Divorced
## 5  41   Private 264663 Some-college            10      Separated
## 6  34   Private 216864      HS-grad             9       Divorced
##          occupation  relationship  race    sex capital.gain capital.loss
## 1                 ? Not-in-family White Female            0         4356
## 2   Exec-managerial Not-in-family White Female            0         4356
## 3                 ?     Unmarried Black Female            0         4356
## 4 Machine-op-inspct     Unmarried White Female            0         3900
```

```
## 5    Prof-specialty     Own-child White Female            0         3900
## 6     Other-service     Unmarried White Female            0         3770
##   hours.per.week native.country income
## 1             40  United-States  <=50K
## 2             18  United-States  <=50K
## 3             40  United-States  <=50K
## 4             40  United-States  <=50K
## 5             40  United-States  <=50K
## 6             45  United-States  <=50K
```

La idea original del dataset es analizar y predecir cuáles de dichas variables del censo tienen impacto en la probabilidad de que el individuo gane o no más de 50K de salario anual. Si bien el objetivo de la práctica no es específicamente la predicción de probabilidades, la cantidad de variables nos va a permitir realizar el preprocesado de los datos (tanto dentro de las propias variables como eligiendo qué variables son necesarias para el estudio), así como un análisis de la relevancia de dichas variables.