A1 - Preproceso de datos

Enunciado

Semestre 2021.2

Índice

Introducción

En esta actividad realizaremos el preprocesado de un fichero de datos que contiene información de una muestra extraída a partir de un censo, en el que para cada persona, se registran los salarios aparte de información personal adicional. El conjunto de datos contiene 32,560 registros y 14 variables. Los datos se han extraído y modificado parcialmente (por motivos académicos) de la base de datos disponible en la web Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Adult.

Las variables son:

- *CS_ID:* Identificador del individuo.
- *age:* Edad del individuo.
- *workclass*: Categorización del individuo en base al perfil laboral.
- *fnlwgt*: Numero de unidades de la población objetivo que representa la unidad respondiente.
- *education_num*: Numero de años de formación educativa del individuo.
- *marital_status*: Estado civil del individuo.
- *occupation*: Categorización del individuo en base a la tipologia del trabajo.
- *relationship*: Parentesco de la unidad que responde de la familia.
- *race*: Grupo racial al que pertenece el individuo.
- *sex*: Genero del individuo.
- *capital_gain*: Dineros ganados por inversión del individuo.
- *capital_loss*: Dineros perdidos por inversión del individuo.
- *hours_per_week*: Horas por semana trabajadas por el individuo.
- *income*: Salario (anual) del individuo.

El objetivo de esta actividad es preparar el fichero para su posterior análisis. Para ello, se examinará el fichero para detectar y corregir posibles errores, inconsistencias y valores perdidos. Además, se presentará una breve estadística descriptiva.

Criterios de verificación y de normalización de las variables:

A continuación se muestran los criterios con los que deben limpiarse los datos del conjunto:

1. Selección del dataset de estudio a partir del dataset original.
   a) Eliminar las variables ‘fnlwgt‘, ‘capital_gain‘ y ‘capital_loss‘ y los registros con más de 5 valores NAs.
   b) Crear la variable ‘education_cat‘ como categorización de la variable numérica ‘education_num‘ (más detalles en la sección correspondiente)
   c) Cambiar el nombre de la variable ‘sex‘ por ‘gender‘.
2. Verificar si hay registros con el valor de ‘CS_ID‘ duplicado. En caso de duplicación, se han de asignar nuevos códigos no repetidos.
3. Eliminar espacios en blanco en el inicio de los valores en las variables cualitativas.
4. Los valores posibles en la variable ‘marital_status‘ son: M (Married), S (Single), X (Separated), D (Divorced), W (Widowed).
5. Los valores posibles en la variable ‘gender‘ son: f (femeníno) y m (masculíno).

6. En los datos numéricos, el símbolo de separador decimal es el punto y no la coma. Además, si se presenta la unidad de la variable, por ejemplo horas en el caso de hours_per_week, se debe eliminar para convertir la variable a tipo numérico.
7. En la variable 'hours_per_week' se consideran valores atípicos aquellos valores superiores a 80 horas/semana.
8. La variable 'income' se ha de expresar en miles de euros (k€)
9. Las variables 'age' y 'education_num' han de ser de tipos entero, sin decimales.

Nota importante a tener en cuenta para entregar la actividad:

- Es necesario entregar el archivo Rmd y el archivo de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se respetará la misma numeración de los apartados que el enunciado.
- No se pueden realizar listas completas del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden utilizar las funciones **head** y **tail** que sólo muestran unas líneas del archivo de datos.
- Se valora la precisión de los términos utilizados (es necesario utilizar de forma precisa la terminología de la estadística).
- Se valora también la concisión en la respuesta. No se trata de realizar explicaciones muy largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos de forma clara y concisa.

Para realizar el preproceso del fichero, seguir los pasos que se indican a continuación.

---

1. Carga del archivo

Cargar el archivo de datos y examinar el tipo de datos con los que R ha interpretado cada variable. Examinar también los valores resumen de cada tipo de variable.

2. Obtención del dataset para realizar el estudio

En esta fase se seleccionan los registros y/o variables que se utilizarán para realizar el estudio. En muchas ocasiones, no se trabaja con todo el dataset disponible ya que existen variables y/o registros de control o que no es necesario usar para el estudio concreto que se quiere realizar.

En este caso, se quieren eliminar las variables `fnlwgt`, `capital_gain` y `capital_loss` además de los registros con más de 5 valores NAs.

Por otra parte, se pueden crear nuevas variables en función de las disponibles. En este caso, se creará la variable `education_cat` que categoriza la formación académica en formación `primaria` si `education_num` es menor de 7 años, `secundaria` si `education_num` está entre 7 y 9 años, `universitaria` si `education_num` esta entre 10 y 13 años y `postuniversitaria` si `education_num` es mayor de 13 años.

Por último, se quiere cambiar el nombre de la variable `sex` por `gender`.

3. Duplicación de códigos

Verificad la consistencia en la variable `CS_ID`. Si existen registros duplicados, asignad un nuevo código para evitar códigos duplicados. El nuevo código debe ser un valor no usado (valores superiores al máximo valor

numérico contenido en `CS_ID`). Conservad el mismo formato que el resto de códigos, con "CS" delante de la secuencia numérica. Podéis usar la función **duplicated** de R para detectar los duplicados.

4. Normalización de los datos cualitativos

4.1. Eliminación de espacios en blanco

Se ha observado que existen espacios en blanco al inicio de los valores en las variables cualitativas. Por tanto, es necesario eliminar estos espacios en blancos.

4.2. Marital-Status

Cambiar las categorías de la variable marital status actuales por otras que ocupen un carácter. Los valores que se asignaron a la variable `marital_status` son: M por Married, S por Single, X por Separated, D por Divorced, W por Widowed. Representad gráficamente la distribución de los valores de la variable.

4.3. Género

Revisad la consistencia de los valores de la variable `gender` y realice las modificaciones oportunas para indicar las categorías finales como `f` y `m` que corresponde a `femenino` y `masculino`, respectivamente. Representad gráficamente la distribución de los valores de la variable.

5. Normalización de los datos cuantitativos

Inspeccionar los valores de los datos cuantitativos y realizar las normalizaciones oportunas siguiendo los criterios especificados anteriormente. Estas normalizaciones tienen como objetivo uniformizar los formatos. Si hay valores perdidos o valores extremos, se tratarán más adelante.

Al realizar estas normalizaciones, se debe demostrar que la normalización sobre cada variable ha dado el resultado esperado. Por lo tanto, se recomienda mostrar un fragmento del archivo de datos resultante. Para evitar mostrar todo el conjunto de datos, se puede mostrar una parte del mismo, con las funciones **head** y/o **tail**.

Seguid el orden de los apartados.

5.1. Edad

Revisad el formato de la variable `age` y realizad las transformaciones oportunas según los criterios especificados anteriormente. Si existen valores atípicos, se tratarán más adelante.

5.2. Educación

Revisad el formato de la variable `education_num` y realizad las transformaciones oportunas según los criterios especificados anteriormente. Si existen valores atípicos, se tratarán más adelante.

5.3. Horas por semana

Revisad el formato de la variable `hours_per_week` y realizad las transformaciones oportunas según los criterios especificados anteriormente. Si existen valores atípicos, se tratarán más adelante.

5.4. Income

Revisad el formato de la variable `income` y realizad las transformaciones oportunas según los criterios especificados anteriormente. Si existen valores atípicos, se tratarán más adelante.

6. Valores atípicos

Revisad si hay valores atípicos en las variables `age`, `education_num`, `hours_per_week` y `income`. Si se trata de un valor anómalo, es decir anormalmente alto o bajo, substituir su valor por NA y posteriormente, se imputará.

7. Imputación de valores

Buscad si existen valores perdidos en las variables cuantitativas `age`, `education_num`, `hours_per_week` y `income`.

En caso de valores perdidos, aplicad el proceso siguiente:

- Para 'age', aplicad imputación por la media aritmética.
- Para 'income', aplicar imputación por la media aritmética de los registros del mismo género, es decir, separado por género.
- En el resto de variables, aplicad imputación por vecinos más cercanos, usando la distancia de Gower, considerando en el cómputo de los vecinos más cercanos el resto de variables cuantitativas mencionadas en este apartado. Además, considerad que la imputación debe hacerse con registros del mismo género. Por ejemplo, si un registro a imputar es de género "M", se debe realizar la imputación usando las variables cuantitativas de los registros de género "M". Para realizar esta imputación, podéis usar la función "kNN" de la librería VIM con un número de vecinos igual a 11.

Mostrad que la imputación se ha realizado correctamente, mostrando el resultado de los datos afectados por la imputación.

8. Estudio descriptivo

8.1. Funciones de media robustas

Implementad una función en R que, dado un vector con datos numéricos, calcule la media recortada y la media Winsor. Estas funciones se deben definir como sigue:

```
media.recortada <- function( x, perc=0.05){}

media.winsor( x, perc=0.05){}
```

donde x es el vector de datos y perc la fracción de los datos a recortar (por defecto, 0.05). Implementad estas funciones en R y comprobad que funcionan correctamente.

8.2. Estudio descriptivo de las variables cuantitativas

Realizad un estudio descriptivo de las variables cuantitativas `age`, `education_num`, `hours_per_week` y `income`.

Para ello, preparad una tabla con varias medidas de tendencia central y dispersión, robustas y no robustas. Usad, entre otras, las funciones del apartado anterior. Presentad, asimismo gráficos donde se visualice la distribución de los valores de estas variables cuantitativas.

9. Archivo final

Una vez realizado el preprocesamiento sobre el archivo, copiad el resultado de los datos en un archivo llamado **CensusIncome_clean.csv**.

10. Evaluación de la actividad

- Secciones 1, 2 (20 %)
- Secciones 3, 4 (10 %)
- Sección 5 (10 %)
- Sección 6 (10 %)
- Sección 7 (10 %)
- Sección 8 (20 %)
- Seccion 9 (10 %)
- Calidad del informe dinámico (calidad del código, formato y estructura del documento, concisión y precisión en las respuestas) (10 %)