A1 - Preproceso de datos

Enunciado

Semestre 2022.1

Índice

Introducción

El conjunto de datos está en el archivo `gpa_row.csv`, contiene la nota media de estudiantes universitarios después del primer semestre de clases (GPA: grade point average, en inglés), así como información sobre la nota de acceso, la cohorte de graduación en el instituto y algunas características de los estudiantes.

Este conjunto de datos surge de una encuesta realizada a una muestra representativa de estudiantes de una universidad de EEUU (por razones de confidencialidad el conjunto de datos no incluye el nombre de la universidad). Las variables incluidas en el conjunto de datos son:

- sat: nota de acceso (medida en escala de 400 a 1600 puntos)
- tothrs: horas totales cursadas en el semestre
- hsize: numero total de estudiantes en la cohorte de graduados del bachillerato (en cientos)
- hsrank: ranking del estudiante, dado por la nota media del bachillerato, en su cohorte de graduados del bachillerato
- hsperc: ranking relativo del estudiante (hsrank/hsize).
- colgpa: nota media del estudiante al final del primer semestre (medida en escala de 0 a 4 puntos)
- athlete: indicador de si el estudiante practica algún deporte en la universidad
- female: indicador de si el estudiante es mujer
- white: indicador de si el estudiante es de raza blanca o no
- black: indicador de si el estudiante es de raza negra o no

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

Por otra parte, se realizará un informe ejecutivo que resumirá lo realizado. Constará de dos partes:

1. Documentar todos los cambios realizados en los datos originales.
2. Breve resumen de las características más destacables de cada variable.

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. Verificar que las variables de tipo indicador deben tener sólo el valor TRUE o FALSE (mayúsculas y sin espacios en blanco) y deben codificarse como variables categóricas ("factor"). En caso de que no se cumpla, es necesario corregirlo.
2. En los datos de naturaleza numéricas, el símbolo de separador decimal es el punto y no la coma. Además, si se presenta la unidad de la variable es necesario eliminarla para convertir la variable a tipo numérico.
3. Comprobar si se cumple el rango de valores posibles en las variables donde se tiene esta información:
   - ‘sat‘ : escala de 400 a 1600 puntos
   - ‘colgpa‘ : escala de 0 a 4 puntos
4. Revisar si los valores de la variable ‘hsperc‘ se ha calculado correctamente a partir de ‘hsrank / hsize‘ con tres decimales de precisión. En caso contrario, modificarlo.

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

Cargue el archivo de datos y examine el tipo de datos con el que R ha interpretado cada variable.

Indique qué variables son de naturaleza numérica, aunque R lo haya podido interpretar de forma distinta. En caso de que el tipo de variable que ha otorgado R no coincida con el tipo que le correspondería, deberá aplicar la transformación correspondiente cuando realice la normalización de la variable (apartado siguiente).

2 Normalización de las variables cualitativas

2.1 Athlete

Normalizar la variable `Athlete` según las indicaciones proporcionadas.

2.2 Female

Normalizar la variable `Female` según las indicaciones proporcionadas.

2.3 Black

Normalizar la variable `Black` según las indicaciones proporcionadas.

2.4 White

Normalizar la variable `white` según las indicaciones proporcionadas.

3 Normalización de las variables cuantitativas

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.

3.1 Nota de acceso

Revise el formato de la variable `sat` y realice las revisiones o transformaciones oportunas según los criterios especificados anteriormente.

3.2 Horas totales cursadas al semestre

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

3.3 Nota media del estudiante al final del primer semestre

Revise el formato de la variable `colgpa` y realice las revisiones o transformaciones oportunas según los criterios especificados anteriormente.

3.4 Número total de estudiants en la cohorte de graduados del bachillerato

Revise el formato de la variable `hsize` y realice las revisiones o transformaciones oportunas según los criterios especificados anteriormente.

3.5 Ranking relativo del estudiante

Revise si la variable `hsperc` se ha obtenido correctamente según los criterios especificados anteriormente.

4 Valores atípicos

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

5 Imputación de valores

Busque si hay valores perdidos en las variables cuantitativas. En el caso de detectar algún valor perdido es necesario realizar una imputación de valores en estas variables. Aplique imputación por vecinos más cercanos, utilizando la distancia de Gower, considerando en el cómputo de los vecinos más cercanos el resto de variables cuantitativas. Además, considere que la imputación debe realizarse con registros del mismo género. Por ejemplo, si un registro a imputar es mujer, se debe realizar la imputación usando sólo las variables cuantitativas de los registros de mujeres.

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.

6 Creación de una nueva variable

La variable `colgpa` contiene la nota numerica del alumnado. Crear una variable categorica denominada `gpaletter`, que indique la nota en letra de cada estudiante de la siguiente forma: A, de 3.50 a 4.00; B, de 2.50 a 3.49; C, de 1.50 a 2.49; D, de 0 a 1.49.

7 Estudio descriptivo

7.1 Estudio descriptivo de las variablas cualitativas

Represente en un primer gráfico, la variable `athlete` en porcentaje de atletas y un segundo gráfico, la variable `athlete` en función del sexo donde se muestre visualmente si el porcentaje de hombres y mujeres cambia al ser atleta o no.

7.2 Estudio descriptivo de las variables cuantitativas

Haga un estudio descriptivo de las variables cuantitativas "sat", "tothrs", "hsize", "hsrank".

Para ello, prepare una tabla con diversas medidas de tendencia central y dispersión, robustas y no robustas. Presente los gráficos donde se visualice la distribución de los valores de "sat" y "sat" en función del sexo.

8 Archivo final

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

9 Informe ejecutivo

Está formado por dos partes:

- Tabla resumen de los cambios realizados en el preprocesamiento.
- Breve explicación de las características estadísticas básicas de cada variable por separado.

9.1 Tabla resumen del preprocesamiento

Documentar de forma resumida, en forma de tabla, los cambios introducidos en el archivo original durante su preprocesamiento. Hay que explicar el detalle del preproceso aplicado. Por ejemplo, no basta con decir "se ha normalizado la variable hsize". En todo caso, debería indicarse si se ha reemplazado la coma por el punto decimal, o si se han redondeado decimales, etcétera y a que observaciones. Debe ser específicos, ya que el informe a de ser útil como documentación de los cambios realizados.

La primera y última fila de la tabla debe indicar el número de observaciones, el número de variables cuantitativas, el número de variables cualitativas y el total de variables al inicio del preprocesamiento y al final, respectivamente.

9.2 Resumen estadístico

A partir de la información obtenida en los apartados anteriores haga un breve comentario de cada variable destacando el más relevante y característico. El resumen no debe ocupar más de una página.

10 Evaluación de la actividad

- Secciones 1, 2 (10%)
- Secciones 3, 4 (20%)
- Sección 5 (10%)
- Sección 6 (10%)
- Secciones 7, 8 (20%)
- Sección 9 (20%)

- Calidad del informe dinámico (calidad del código, formato y estructura del documento, concisión y precisión en las respuestas) (10%)

11 Referencias

Quick-R

Cookbook for R

LaTex tables

Data Visualization with R