Práctica II
Tipología
y
Ciclo de Vida de los datos

Gregorio Andrés García Menéndez
Manuel Gómez Montero

29 de mayo de 2019

Índice

1. Introducción

1.1. Objetivos

El objetivo de esta práctica es aplicar los conocimientos ... bla bla bla....

1.2. Descripción del Dataset

El dataset escogido ha sido obtenido a través de una encuesta de estudiantes de secundaria que asisten a cursos de matemáticas y portugués. Este conjunto de datos contiene bastante información de interés sobre los estudiantes.

Los datos han sido obtenidos de *Kaggle* y se encuentran en dos ficheros independientes, cada uno de estos ficheros contiene las siguientes variables:

- *school*: Define la escuela del estudiante. Puede ser "GP" (Gabriel Pereira) o "MS" (Mousinho da Silveira).
- *sex*: Indica el sexo del estudiante.
- *age*: Se refiere a la edad del estudiante.
- *address*: Indica si el estudiante vive en zona urbana o rural.
- *famsize*: Define si la familia se compone de menos de tres miembros o de tres o más.
- *Pstatus*: Se refiere al estado de los padres, si viven juntos o no.
- *Medu*: Nivel de educación de la madre (0: Ninguna, 1: Primaria (hasta 4º), 2: Primaria (Desde 5º), 3: Secundaria o 4: Educación superior).
- *Fedu*: Nivel de educación del padre (0: Ninguna, 1: Primaria (hasta 4º), 2: Primaria (Desde 5º), 3: Secundaria o 4: Educación superior).
- *Mjob*: Trabajo de la madre.
- *Fjob*: Trabajo del padre.
- *reason*: Razón por la que se escoge esta escuela.
- *guardian*: Tutor del alumno.

- *traveltime*: Tiempo de camino a la escuela (1: ¡15 min., 2: 15 to 30 min., 3: 30 min. a 1 hora,4: ¿1 hora).
- *studytime*: Tiempo de estudio semanal (1: ¡2 horas, 2: 2 a 5 horas, 3: 5 a 10 horas o 4: ¿10 horas).
- *failures* - Número de fracasos en clases pasadas (4 para 4 o más)
- *schoolsup* - Define si el estudiante recibe o no clases particulares.
- *famsup* - Indica si el estudiante recibe apoyo educativo familiar.
- *paid*: Clases extra pagadas dentro de la asignatura del curso.
- *activities*: Actividades extraescolares.
- *nursery*: Indica si asistió a la guardería.
- *higher*: Define si el estudiante tiene intención de realizar estudios superiores.
- *internet*: Indica si el estudiante posee internet en casa.
- *romantic*: Define si el estudiante tiene una relación amorosa.
- *famrel*: Calidad de las relaciones familiares (De 1 a 5 de peor a mejor).
- *freetime*: Indica la cantidad de tiempo libre del estudiante (De 1 a 5 de menos a más).
- *goout*: Mide la frecuencia con la que el estudiante sale con amigos (De 1 a 5 de menos a más).
- *Dalc* -Consumo de alcohol entre semana (De 1 a 5 de menos a más).
- *Walc* - Consumo de alcohol los fines de semana (De 1 a 5 de menos a más).
- *health* - Mide el estado de salud del estudiante (De 1 a 5 de peor a mejor).
- *absences*: Número de ausencias a clase.
- *G1*: Calificación del primer periodo.
- *G2*: Calificación del segundo periodo.
- *G3*: Calificación final.

2. Limpieza de datos

2.1. Unión de conjuntos

El primer paso consiste en unir ambos conjuntos de datos para tener un único dataset que será el que utilizaremos en el resto del documento.

Según se indica en Kaggle existen 382 estudiantes que asisten a ambos cursos y estos estudiantes pueden ser identificados por las siguientes características: colegio, sexo, edad, dirección, tamaño de familia, trabajo y nivel de educación de los padres, razón por la que ha elegido el colegio, si han asistido a la guardería y si poseen internet en casa.

Sin embargo creemos que además de estas características hay otras que no pueden variar en un mismo estudiante en ambos ficheros como son el tiempo de camino a la escuela, número de fracasos en clases pasadas, si realiza actividades extraescolares, si tiene intención de realizar estudios superiores, si tiene una relación, la calidad de las relaciones, la frecuencia de salidas con amigos, la cantidad consumida de alcohol y el estado de salud. Con esta separación vemos que existen 320 estudiantes que asisten a ambas clases.

Comparando los valores de ambos cursos del resto del variables vemos que también coinciden en ambos cursos el tutor del alumno, si recibe soporte familiar y el tiempo de estudio.

El dataset resultado de la unión posee un total de 724 alumnos 75 de los cuales solo asisten a matemáticas, 329 solo a portugués y 320 a ambos cursos.

2.2. Selección de variables

2.3. Tipos de Variables

3. Análisis Estadístico

3.1. Análisis gráfico inicial

4. Conclusiones

Referencias

[HKP12] Han, Jiawei ; Kamber, Micheline ; Pei, Jian: *Data mining: concepts and techniques*. Elsevier/Morgan Kaufmann, 2012. – Chapter 3

[Osb10] Osborne, Jason W.: Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. In: *Newborn and Infant Nursing Reviews* 10 (2010), Nr. 1, S. 37–43. `http://dx.doi.org/10.1053/j.nainr.2009.12.009`. – DOI 10.1053/j.nainr.2009.12.009

[Squ15] Squire, Megan: *Clean data: save time by discovering effortless strategies for cleaning, organizing, and manipulating your data*. Packt Publishing Ltd, 2015. – Chapters 1 & 2