Práctica II
Tipología
y
Ciclo de Vida de los datos

Gregorio Andrés García Menéndez
Manuel Gómez Montero

30 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:

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

13. *traveltime*: Tiempo de camino a la escuela (1: ¡15 min., 2: 15 to 30 min., 3: 30 min. a 1 hora,4: ¿1 hora).

14. *studytime*: Tiempo de estudio semanal (1: ¡2 horas, 2: 2 a 5 horas, 3: 5 a 10 horas o 4: ¿10 horas).

15. *failures* - Número de fracasos en clases pasadas (4 para 4 o más)

16. *schoolsup* - Define si el estudiante recibe o no clases particulares.

17. *famsup* - Indica si el estudiante recibe apoyo educativo familiar.

18. *paid*: Clases extra pagadas dentro de la asignatura del curso.

19. *activities*: Actividades extraescolares.

20. *nursery*: Indica si asistió a la guardería.

21. *higher*: Define si el estudiante tiene intención de realizar estudios superiores.

22. *internet*: Indica si el estudiante posee internet en casa.

23. *romantic*: Define si el estudiante tiene una relación amorosa.

24. *famrel*: Calidad de las relaciones familiares (De 1 a 5 de peor a mejor).

25. *freetime*: Indica la cantidad de tiempo libre del estudiante (De 1 a 5 de menos a más).

26. *goout*: Mide la frecuencia con la que el estudiante sale con amigos (De 1 a 5 de menos a más).

27. *Dalc* -Consumo de alcohol entre semana (De 1 a 5 de menos a más).

28. *Walc* - Consumo de alcohol los fines de semana (De 1 a 5 de menos a más).

29. *health* - Mide el estado de salud del estudiante (De 1 a 5 de peor a mejor).

30. *absences*: Número de ausencias a clase.

31. *G1*: Calificación del primer periodo.

32. *G2*: Calificación del segundo periodo.

33. *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.

```
school   sex          age        address famsize   Pstatus      Medu            Fedu            Mjob           Fjob            reason
GP:485   F:417   Min.   :15.00   R:218   GT3:508   A: 90   Min.   :0.000   Min.   :0.000   at_home :150   at_home : 48   course    :312
MS:239   M:307   1st Qu.:16.00   U:506   LE3:216   T:634   1st Qu.:2.000   1st Qu.:1.000   health  : 52   health  : 26   home      :171
                 Median :17.00                             Median :2.000   Median :2.000   other   :283   other   :407   other     : 78
                 Mean   :16.81                             Mean   :2.485   Mean   :2.285   services:164   services:205   reputation:163
                 3rd Qu.:18.00                             3rd Qu.:4.000   3rd Qu.:3.000   teacher : 75   teacher : 38
                 Max.   :22.00                             Max.   :4.000   Max.   :4.000

nursery   internet    traveltime       failures      schoolsup activities higher    romantic      famrel         freetime       goout
no :150   no :169   Min.   :1.000   Min.   :0.0000   no :648   no :378    no : 82   no :452   Min.   :1.000   Min.   :1.0   Min.   :1.000
yes:574   yes:555   1st Qu.:1.000   1st Qu.:0.0000   yes: 76   yes:346    yes:642   yes:272   1st Qu.:4.000   1st Qu.:3.0   1st Qu.:2.000
                    Median :1.000   Median :0.0000                                           Median :4.000   Median :3.0   Median :3.000
                    Mean   :1.565   Mean   :0.3453                                           Mean   :3.913   Mean   :3.2   Mean   :3.195
                    3rd Qu.:2.000   3rd Qu.:0.0000                                           3rd Qu.:5.000   3rd Qu.:4.0   3rd Qu.:4.000
                    Max.   :4.000   Max.   :3.0000                                           Max.   :5.000   Max.   :5.0   Max.   :5.000

     Dalc           Walc          health       guardian   famsup      studytime    paid.mat    absences.mat       G1.mat
Min.   :1.000   Min.   :1.000   Min.   :1.000   father:169   no :287   Min.   :1.00   no  :214   Min.   : 0.000   Min.   : 3.00
1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   mother:491   yes:437   1st Qu.:1.00   yes :181   1st Qu.: 0.000   1st Qu.: 8.00
Median :1.000   Median :2.000   Median :4.000   other : 64             Median :2.00   NA's:329   Median : 4.000   Median :11.00
Mean   :1.519   Mean   :2.311   Mean   :3.552                          Mean   :1.92              Mean   : 5.709   Mean   :10.91
3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000                          3rd Qu.:2.00              3rd Qu.: 8.000   3rd Qu.:13.00
Max.   :5.000   Max.   :5.000   Max.   :5.000                          Max.   :4.00              Max.   :75.000   Max.   :19.00
                                                                                                 NA's   :329      NA's   :329
    G2.mat          G3.mat      paid.por    absences.por       G1.por          G2.por          G3.por
Min.   : 0.00   Min.   : 0.00   no  :610   Min.   : 0.000   Min.   : 0.0   Min.   : 0.00   Min.   : 0.00
1st Qu.: 9.00   1st Qu.: 8.00   yes : 39   1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00   1st Qu.:10.00
Median :11.00   Median :11.00   NA's: 75   Median : 2.000   Median :11.0   Median :11.00   Median :12.00
Mean   :10.71   Mean   :10.42              Mean   : 3.659   Mean   :11.4   Mean   :11.57   Mean   :11.91
3rd Qu.:13.00   3rd Qu.:14.00              3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00   3rd Qu.:14.00
Max.   :19.00   Max.   :20.00              Max.   :32.000   Max.   :19.0   Max.   :19.00   Max.   :19.00
NA's   :329     NA's   :329                NA's   :75       NA's   :75     NA's   :75      NA's   :75
```

Figura 1: Summary dataset conjunto

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. Tratamiento de NAs

En primer lugar vamos a encargarnos de los valores perdidos que se dan unicamente en los casos de que los alumnos no hayan asistido a alguno de los cursos, una opción sería eliminar los datos de cualquier alumno que no posea datos en algunos de los cursos, pero implicaría quedarnos solo con un 44 % de los datos.

Por lo tanto la opción escogida para los valores perdidos es la siguiente:

- *paid*: Unificaremos las variables de ambos cursos tomando un valor "yes" cuando el alumno de clases extras pagadas de alguno de los cursos.
- *absences*: Viendo los números parece que las ausencias son mayores en matemáticas. Como lo que nos interesa es saber los alumnos que han faltado más o menos y queremos evitar que tenga un mayor peso los alumnos de matemáticas. Normalizaremos estos valores y en el caso de que los alumnos hayan asistido a ambos cursos utilizaremos la media de ambos valores.
- *calificaciones*: En este caso todas las calificaciones se mueven en el mismo rango por lo que en el caso de los alumnos que hayan asistido a ambos cursos aplicaremos la media de ambos.

En el notebook se pueden ver los detalles de estos procesos en la figura 2 podemos ver la salida del comando *summary* para este nuevo dataset.

2.3. Selección de variables/Reducción de dimensionalidad

Observando el dataset vemos que, aún tras la unión de los datasets, tenemos una gran cantidad de dimensiones, en concreto 33, lo que puede dificultar nuestro análisis. En esta primera fase vamos a hacer una primera aproximación para reducir la dimensionalidad. Posteriormente en los primeros apartados del análisis estadístico reduciremos aún más esta dimensionalidad.

En primer lugar existen variables que, a priori, tienen poca significancia para nuestro análisis como puede ser la escuela a la que asiste el alumno, por lo que podemos eliminar esta variable.

```
school   sex          age        address famsize   Pstatus      Medu            Fedu            Mjob
GP:485   F:417   Min.   :15.00   R:218   GT3:508   A: 90   Min.   :0.000   Min.   :0.000   at_home :150
MS:239   M:307   1st Qu.:16.00   U:506   LE3:216   T:634   1st Qu.:2.000   1st Qu.:1.000   health  : 52
                 Median :17.00                             Median :2.000   Median :2.000   other   :283
                 Mean   :16.81                             Mean   :2.485   Mean   :2.285   services:164
                 3rd Qu.:18.00                             3rd Qu.:4.000   3rd Qu.:3.000   teacher : 75
                 Max.   :22.00                             Max.   :4.000   Max.   :4.000
      Fjob           reason    nursery   internet    traveltime       failures      schoolsup activities higher
at_home : 48   course    :312   no :150   no :169   Min.   :1.000   Min.   :0.0000   no :648   no :378    no : 82
health  : 26   home      :171   yes:574   yes:555   1st Qu.:1.000   1st Qu.:0.0000   yes: 76   yes:346    yes:642
other   :407   other     : 78                       Median :1.000   Median :0.0000
services:205   reputation:163                       Mean   :1.565   Mean   :0.3453
teacher : 38                                       3rd Qu.:2.000   3rd Qu.:0.0000
                                                   Max.   :4.000   Max.   :3.0000
romantic      famrel         freetime       goout           Dalc            Walc           health        guardian
no :452   Min.   :1.000   Min.   :1.0   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   father:169
yes:272   1st Qu.:4.000   1st Qu.:3.0   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   mother:491
          Median :4.000   Median :3.0   Median :3.000   Median :1.000   Median :2.000   Median :4.000   other : 64
          Mean   :3.913   Mean   :3.2   Mean   :3.195   Mean   :1.519   Mean   :2.311   Mean   :3.552
          3rd Qu.:5.000   3rd Qu.:4.0   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000
          Max.   :5.000   Max.   :5.0   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
famsup      studytime     paid           G1              G2              G3           absences
no :287   Min.   :1.00   no :515   Min.   : 2.50   Min.   : 0.00   Min.   : 0.00   Min.   :0.0000
yes:437   1st Qu.:1.00   yes:209   1st Qu.: 9.00   1st Qu.: 9.00   1st Qu.: 9.50   1st Qu.:0.0000
          Median :2.00             Median :11.00   Median :11.00   Median :11.00   Median :0.0625
          Mean   :1.92             Mean   :10.94   Mean   :10.97   Mean   :11.07   Mean   :0.1057
          3rd Qu.:2.00             3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:0.1527
          Max.   :4.00             Max.   :18.50   Max.   :18.50   Max.   :18.50   Max.   :0.8733
```

Figura 2: Summary dataset sin NAs

Por otro lado podemos agrupar las calificaciones en una única variable que nos indique la calificación media del alumno.

Con estos detalles que pueden observarse en el notebook vemos que hemos conseguido reducir a 30 la dimensionalidad.

```
school   sex          age        address famsize   Pstatus      Medu            Fedu            Mjob
GP:485   F:417   Min.   :15.00   R:218   GT3:508   A: 90   Min.   :0.000   Min.   :0.000   at_home :150
MS:239   M:307   1st Qu.:16.00   U:506   LE3:216   T:634   1st Qu.:2.000   1st Qu.:1.000   health  : 52
                 Median :17.00                             Median :2.000   Median :2.000   other   :283
                 Mean   :16.81                             Mean   :2.485   Mean   :2.285   services:164
                 3rd Qu.:18.00                             3rd Qu.:4.000   3rd Qu.:3.000   teacher : 75
                 Max.   :22.00                             Max.   :4.000   Max.   :4.000
      Fjob           reason    nursery   internet    traveltime       failures      schoolsup activities higher
at_home : 48   course    :312   no :150   no :169   Min.   :1.000   Min.   :0.0000   no :648   no :378    no : 82
health  : 26   home      :171   yes:574   yes:555   1st Qu.:1.000   1st Qu.:0.0000   yes: 76   yes:346    yes:642
other   :407   other     : 78                       Median :1.000   Median :0.0000
services:205   reputation:163                       Mean   :1.565   Mean   :0.3453
teacher : 38                                       3rd Qu.:2.000   3rd Qu.:0.0000
                                                   Max.   :4.000   Max.   :3.0000
romantic      famrel         freetime       goout           Dalc            Walc           health        guardian
no :452   Min.   :1.000   Min.   :1.0   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   father:169
yes:272   1st Qu.:4.000   1st Qu.:3.0   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   mother:491
          Median :4.000   Median :3.0   Median :3.000   Median :1.000   Median :2.000   Median :4.000   other : 64
          Mean   :3.913   Mean   :3.2   Mean   :3.195   Mean   :1.519   Mean   :2.311   Mean   :3.552
          3rd Qu.:5.000   3rd Qu.:4.0   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000
          Max.   :5.000   Max.   :5.0   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
famsup      studytime     paid           G1              G2              G3           absences
no :287   Min.   :1.00   no :515   Min.   : 2.50   Min.   : 0.00   Min.   : 0.00   Min.   :0.0000
yes:437   1st Qu.:1.00   yes:209   1st Qu.: 9.00   1st Qu.: 9.00   1st Qu.: 9.50   1st Qu.:0.0000
          Median :2.00             Median :11.00   Median :11.00   Median :11.00   Median :0.0625
          Mean   :1.92             Mean   :10.94   Mean   :10.97   Mean   :11.07   Mean   :0.1057
          3rd Qu.:2.00             3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:0.1527
          Max.   :4.00             Max.   :18.50   Max.   :18.50   Max.   :18.50   Max.   :0.8733
```

Figura 3: Summary dataset reducido

2.4. Tipos de Variables

En los tipos mostrados hay algunos que no es correcto el tipo. En la educación tanto de la madre como del padre consideramos que aunque se use un número para la representación debería ser un factor. Lo mismo ocurre para el tiempo de viaje o el tiempo de estudio. Podemos encontrar más detalles en el notebook.

3. Análisis Estadístico

3.1. Análisis gráfico inicial

Observando las gráficas vemos que hay variables que parece que sí tienen influencia a simple vista en el consumo de alcohol tanto a diario como los fines de semana como pueden ser el sexo, el estado de los padres, sin embargo vemos otros que, a priori, no parece que tengan influencia así que en un primer momento vamos a dejar fuera del análisis si el alumno tiene internet, si tiene una relación, el tutor, la dirección, si realiza actividades extraescolares, si recibe clases de pago o si recibe el apoyo de su familia en el estudio.

TODO meter algunas de las gráficas

```
sex          age        famsize   Pstatus      Medu           Fedu            Mjob           Fjob
F:417   Min.   :15.00   GT3:508   A: 90   Min.   :0.000   Min.   :0.000   at_home :150   at_home : 48
M:307   1st Qu.:16.00   LE3:216   T:634   1st Qu.:2.000   1st Qu.:1.000   health  : 52   health  : 26
        Median :17.00                     Median :2.000   Median :2.000   other   :283   other   :407
        Mean   :16.81                     Mean   :2.485   Mean   :2.285   services:164   services:205
        3rd Qu.:18.00                     3rd Qu.:4.000   3rd Qu.:3.000   teacher : 75   teacher : 38
        Max.   :22.00                     Max.   :4.000   Max.   :4.000
       reason      traveltime       failures      schoolsup higher        famrel         freetime
course    :312   Min.   :1.000   Min.   :0.0000   no :648   no : 82   Min.   :1.000   Min.   :1.0
home      :171   1st Qu.:1.000   1st Qu.:0.0000   yes: 76   yes:642   1st Qu.:4.000   1st Qu.:3.0
other     : 78   Median :1.000   Median :0.0000                       Median :4.000   Median :3.0
reputation:163   Mean   :1.565   Mean   :0.3453                       Mean   :3.913   Mean   :3.2
                 3rd Qu.:2.000   3rd Qu.:0.0000                       3rd Qu.:5.000   3rd Qu.:4.0
                 Max.   :4.000   Max.   :3.0000                       Max.   :5.000   Max.   :5.0
    goout           Dalc            Walc           health        studytime       absences
Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :0.0000
1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.00   1st Qu.:0.0000
Median :3.000   Median :1.000   Median :2.000   Median :4.000   Median :2.00   Median :0.0625
Mean   :3.195   Mean   :1.519   Mean   :2.311   Mean   :3.552   Mean   :1.92   Mean   :0.1057
3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:2.00   3rd Qu.:0.1527
Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :4.00   Max.   :0.8733
```

Figura 4: Summary dataset final

En la figura 4 podemos ver el resumen del dataset. Tras la fase de limpieza y este pequeño análisis hemos pasado de tener dos datasets con 33 dimensiones a **un único dataset con 21 dimensiones** pasando por un dataset conjunto de 41 dimensiones.

3.2. Estadística Inferencial

En este apartado vamos a ver si hay diferencia significativas en el consumo de alcohol a diario o fines de semana entre estudiantes de distinto sexo o entre estudiantes con diferente situación de los padres.

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