Análisis de alumnado

Alvaro Armero

27 de marzo de 2016

# Establecemos espacio de trabajo y descargamos los datos

setwd("/Volumes/HDD/alvaroarmero/Documents/Master\_BigData \_UPSA/dataAnalysisPractice")  
if (!file.exists("students"))   
{  
 dir.create("students")  
 fileURL <- "https://archive.ics.uci.edu/ml/machine-learning-databases/00320/student.zip"  
 download.file(fileURL,destfile="./students/student.zip", method="curl")  
 unzip("students/student.zip", exdir="./students/")  
 list.files("./students")  
}  
  
library(knitr)

# Cargamos los estudiantes de portugués y de español

studentMat <- read.table("./students/student-mat.csv",   
 row.names=NULL, sep=";", header=TRUE)  
studentPor <- read.table("./students/student-por.csv",   
 row.names=NULL, sep=";", header=TRUE)  
#kable(studentPor) Para mostrar todos  
kable(head(studentMat[,1:5]))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| school | sex | age | address | famsize |
| GP | F | 18 | U | GT3 |
| GP | F | 17 | U | GT3 |
| GP | F | 15 | U | LE3 |
| GP | F | 15 | U | GT3 |
| GP | F | 16 | U | GT3 |
| GP | M | 16 | U | LE3 |

names(studentMat)

## [1] "school" "sex" "age" "address" "famsize"   
## [6] "Pstatus" "Medu" "Fedu" "Mjob" "Fjob"   
## [11] "reason" "guardian" "traveltime" "studytime" "failures"   
## [16] "schoolsup" "famsup" "paid" "activities" "nursery"   
## [21] "higher" "internet" "romantic" "famrel" "freetime"   
## [26] "goout" "Dalc" "Walc" "health" "absences"   
## [31] "G1" "G2" "G3"

# Limpieza de los datos

names(studentMat) <- tolower(names(studentMat))  
names(studentPor) <- tolower(names(studentPor))  
studentMat$mjob <- gsub("\_", "", studentMat$mjob)  
studentPor$mjob <- gsub("\_", "", studentPor$mjob)  
studentMat$mjob <- gsub("\_", "", studentMat$fjob)  
studentPor$mjob <- gsub("\_", "", studentPor$fjob)  
  
  
columnas <- c("school","sex","age", "freetime","internet",  
 "address","famsize",  
 "pstatus","medu",  
 "fedu","mjob",  
 "fjob","reason",  
 "nursery","internet")  
columnas

## [1] "school" "sex" "age" "freetime" "internet" "address"   
## [7] "famsize" "pstatus" "medu" "fedu" "mjob" "fjob"   
## [13] "reason" "nursery" "internet"

# Unimos ambos conjuntos

studentMatPor <- merge(studentMat,studentPor,  
 by=columnas,  
 all=FALSE, suffixes=c("mat","por"))  
kable(studentMatPor[1:5, 1:7])

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| school | sex | age | freetime | internet | address | famsize |
| GP | F | 15 | 1 | yes | R | GT3 |
| GP | F | 15 | 1 | yes | R | LE3 |
| GP | F | 15 | 1 | yes | U | GT3 |
| GP | F | 15 | 2 | yes | R | GT3 |
| GP | F | 15 | 2 | yes | R | GT3 |

## Mostramos edades

edades <- studentMatPor$age  
kable(edades[1:5])

edadesReves <- sort(edades, decreasing=TRUE)  
kable(edadesReves[1:5])

## Estudiamos sus notas

Los alumnos de matematicas con mayor tiempo libre son capaces de sacar mejores calificaciones.

library(reshape)  
mediaFreeMat <- cast(studentMat, freetime~internet, mean, value=c("g3"))   
mediaFreeMat

## freetime no yes  
## 1 1 8.666667 10.384615  
## 2 2 10.538462 11.823529  
## 3 3 8.500000 9.992593  
## 4 4 9.235294 10.632653  
## 5 5 11.000000 11.375000

Los alumnos de portugués parecen sacarle menos partido a tener internet para mejorar sus calificaciones. Podrían necesitar alguna ayuda para poder enseñarles cómo sacarle mayor partido.

mediaFreePor <- cast(studentPor, freetime~internet, mean, value=c("g3"))   
mediaFreePor

## freetime no yes  
## 1 1 11.812500 11.68966  
## 2 2 11.428571 13.16456  
## 3 3 11.018182 12.35204  
## 4 4 10.942857 11.90210  
## 5 5 9.823529 10.98039