practica_analisis_exploratorio

Jesus Manuel Vicente Garcia

4 de julio de 2017

1. OBJETIVO

EL objetivo es tratar de predecir el valor de la variable G1

2. CARGA DE DATOS

Leemos ambos dataset. El archivo se ha descargado como CSV

Comprobamos que están instalados los paquetes necesarios

```
if(! "readr" %in% installed.packages()) install.packages("readr", depend = TRUE)

currentDir <- getwd()

path <- currentDir

file1 <- "student-mat.csv"

 file2 <- "student-por.csv"

d1=read.table("path /file1", row.names=NULL, header=TRUE, sep = ";")
d2=read.table("path /file2 ", row.names=NULL, header=TRUE, sep = ";")

d3=merge(d1,d2,by=c("school","sex","age","address","famsize","Pstatus","Medu","Fedu","Mjob","Fjob","reason","nursery","internet"))
```

Comprobamos y eliminamos los elementos NA en caso de que los haya. No los hay

```
rowNadatad3 <- d3[rowSums(is.na(d3))>0, ] dim(rowNadatad3) colNadatad3 <- d3[colSums(is.na(d3))>0, ] dim(colNadatad3)
```

Convertimos a minúscula los nombres de las columnas

```
names(d3) <- tolower(names(d3))
```

3. ANÁLISIS DE LA VARIABLE G1.Y

Obtenemos la frecuencia absoluta del valor de a variable G1. Observamos que

los valores más repetidos están entre los valores 10 y 14.

```
table(d3$g1.y)
```

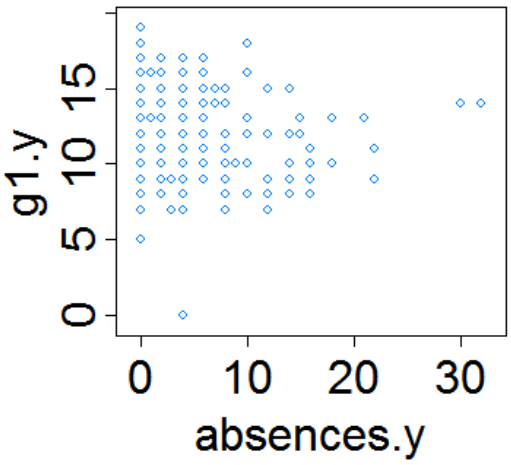
Mostramos el histograma de dichas frecuencias

```
hist(d3$g1.y, nclass = 15, plot = TRUE)
```

Comparamos la variable G1 con las ausencias (variable absences). Vemos que a menor número de ausencias mayor es el valor, por lo que existe una relación.

```
library(gridExtra) library(caret) plot_absences <- xyplot(g1.y ~ absences.y, data = d3,
scales=list(relation="free", x=list(cex=2),
y=list(cex=2)),xlab=list(cex=2),ylab=list(cex=2))
```

```
plot_absences
```



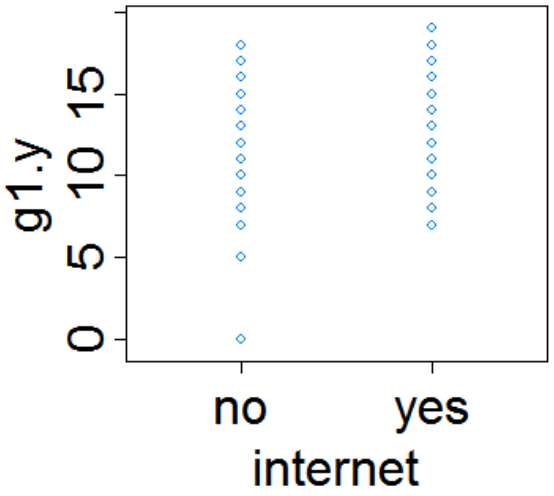
Comparamos la variable G1 con la variable Internet (si el alumno dispone de Internet o no).

Vemos en este caso que los valores están muy repartidos por lo que no hay relacion directa

que permita predecir con precisión el valor de G1.

```
plot_internet <- xyplot(g1.y ~ internet, data = d3, scales=list(relation="free",
x=list(cex=2), y=list(cex=2)),xlab=list(cex=2),ylab=list(cex=2))
```

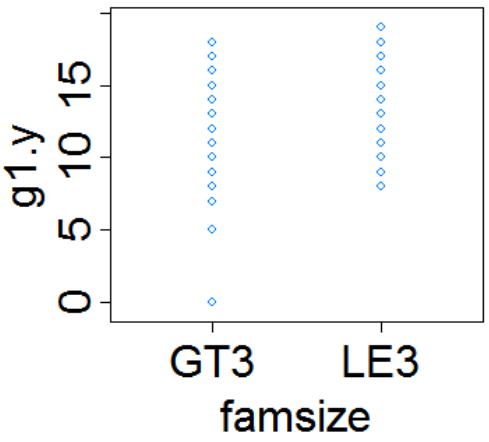
```
plot_internet
```



Comparamos la variable G1 con la variable famsize (miembros de la familia). Los valores también están muy repartidos, por lo que es una variable que tampoco aporta información relevante.

```
plot_famsize <- xyplot(g1.y ~ famsize, data = d3, scales=list(relation="free",
x=list(cex=2), y=list(cex=2)),xlab=list(cex=2),ylab=list(cex=2))
```

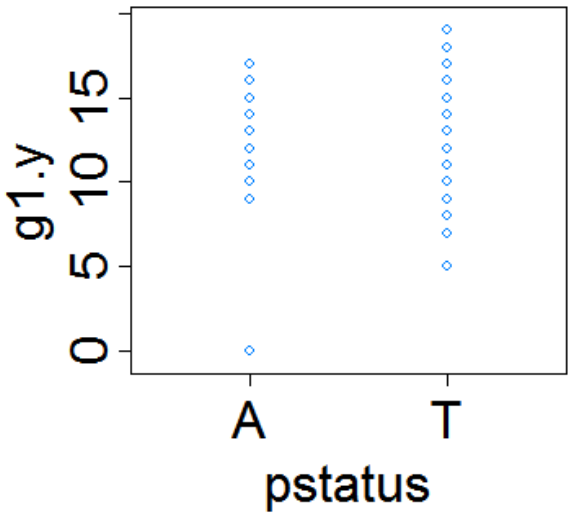
```
plot_famsize
```



Comparamos la variable G1 con la variable pstatus (padres viviendo juntos o separados). Aquí hay cierta relación, pero no la suficiente como para poder realizar una predicción con precisión.

```
plot_pstatus<- xyplot(g1.y ~ pstatus, data = d3, scales=list(relation="free",
x=list(cex=2), y=list(cex=2)),xlab=list(cex=2),ylab=list(cex=2))
```

```
plot_pstatus
```

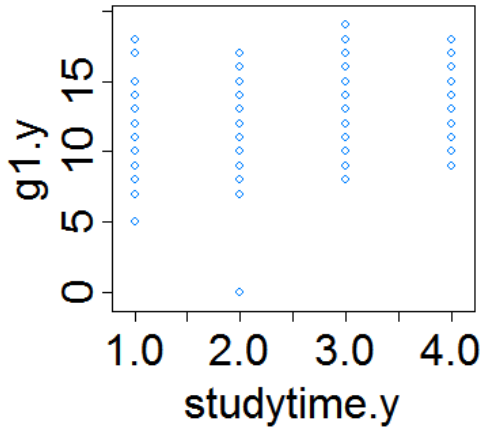


Comparamos la variable G1 con la variable studytime (tiempo de estudio). Se puede observar #que a mayor tiempo de estudio, el valor de la variable es ligeramente mayor. Si embargo, no

se puede predecir con exactitud.

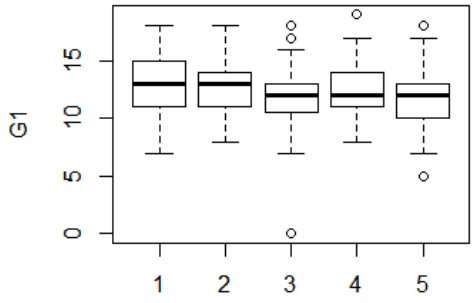
plot_studytime <- xyplot(g1.y ~ studytime.y, data = d3, scales=list(relation="free", x=list(cex=2), y=list(cex=2)),xlab=list(cex=2),ylab=list(cex=2))

plot_studytime



Comparamos la variable G1 con la variable Health (salud). Los valores también están muy repartidos, por lo que es una variable que tampoco aporta información relevante.

boxplot(d3$g1.y $d3health.y ,ylab="G1")



4. CONCLUSION

Habiendo analizado una serie de variables, se observa como la mayoría de ellas no aporta información precisa a la hora de predecir el valor de la variable G1.y.

A priori, y a falta de analizar algunas variables más, podemos decir que este dataset no sería óptimo para el objetivo planteado en la introducción. De todas las variables analizadas, la que más información nos ha aportado para un futuro análisis predictivo ha sido Absences(ausencias), que tiene mucho sentido, ya que normalmente el rendimiento escolar está ligado al número de ausencias a clase.