Untitled

Introducción al dataset

Para la práctica de reglas de asociación de la asignatura *"Minería de Datos: Detección de Anomálias y Aprendizaje no Supervisado"* se ha utilizado un dataset recuperado de la web de UCI. El objetivo des predecir la nota final G3 de un conjunto de datos de alumnos dado.

Para ello generaramos dos modelos machine learning a partir de un análisis descriptivo y exploratorio utilizando un metodo no supervisado. Los datos los obtenemos en https://archive.ics.uci.edu/ml/datasets/Student+Performance.

Vamos a trabajar únicamente con el dataset de la asignatura de portugués

```
studentMat <- read.table("student/student-mat.csv", row.names=NULL, sep=";", header=TRUE)
studentPor <- read.table("student/student-por.csv", row.names=NULL, sep=";", header=TRUE)
```

Análisis exploratorio

Vamos a comprobar si tenemos datos completos en los datos, ya que los valores perdidos siempre pueden causar problemas.

```
## [1] FALSE
```

Vemos que no hay valores perdidos por lo que el siguiente paso será ver de que tipo son nuestros datos, así como sus distribuciones para comenzar a hacernos una idea de que es lo que tenemos entre manos. Parece que tenemos una combinación entre factores y variables numéricas.

```
## 'data.frame':    649 obs. of  33 variables:
##  $ school    : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
##  $ famsize   : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
##  $ Pstatus   : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
##  $ Fjob      : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3 3 ...
##  $ reason    : Factor w/ 4 levels "course","home",..: 1 1 3 2 2 4 2 2 2 2 ...
##  $ guardian  : Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2 ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ schoolsup : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
##  $ famsup    : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
##  $ paid      : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ activities: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
##  $ nursery   : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
##  $ higher    : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ internet  : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
##  $ romantic  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
```

```
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  4 2 6 0 0 6 0 2 0 0 ...
##  $ G1        : int  0 9 12 14 11 12 13 10 15 12 ...
##  $ G2        : int  11 11 13 14 13 12 12 13 16 12 ...
##  $ G3        : int  11 11 12 14 13 13 13 13 17 13 ...
```

Vemos que tenemos 33, 17 factores que trataremos a continuación y el resto poseen valores enteros. De las 16 variables de tipo enteros, todas ellas son variables categóricas que trataremos más adelante. Los datos de nuestro dataset son extremadamente compatibles con el análisi, ya que la mayoría de las columnas son variables binarias o elementos de un conjunto finito de valor discreto cuyo tamaño oscila alrededor de 5. Por lo tanto, podemos cambiar la clase de la mayoría de Las columnas a factorizar.

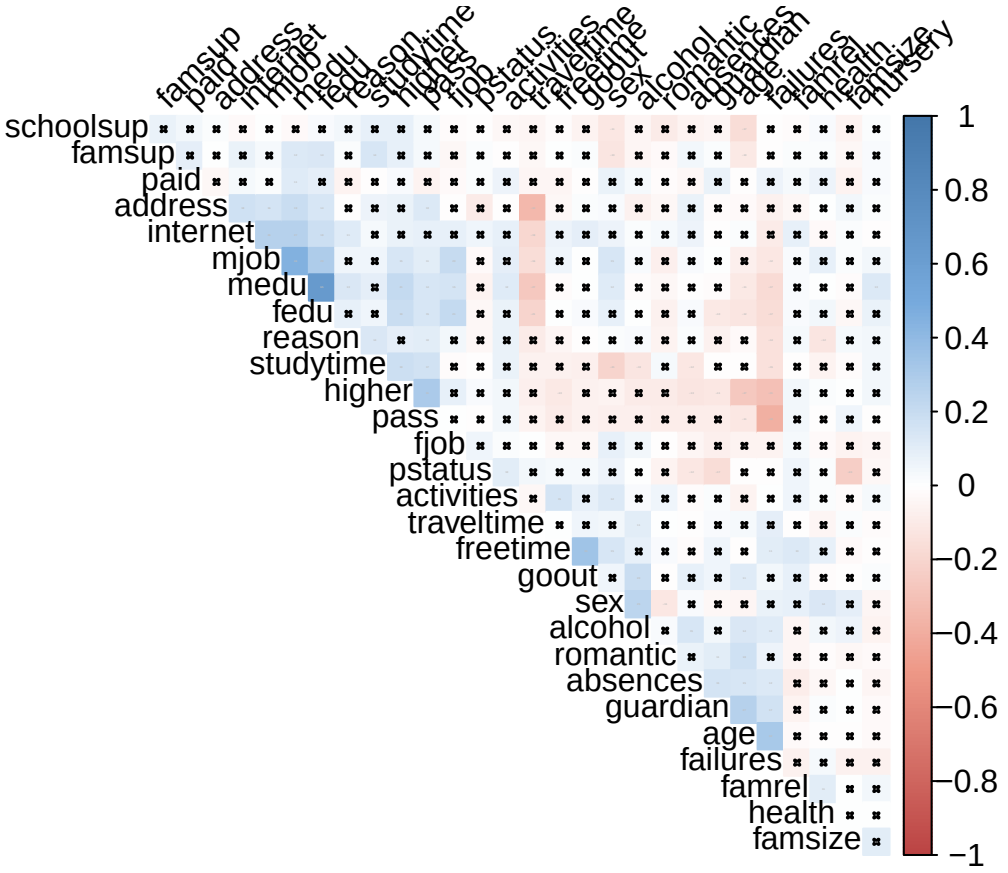
Limpieza de datos

Al igual que solo necesitamos una columna G1 modificada para evaluar el rendimiento, hay varias columnas en el conjunto de datos que no mejoran nuestra comprensión de la situación. Las tres variables de grado (G1 y G2) indican el elemento que queremos predecir, por lo que nos deshacemos de ellas. Además, no necesitamos hacer ninguna distinción entre las dos escuelas y podemos resumir Dalc y Walc como una única variable (alcohol)

Análisis exploratorio

Empecemos con una exploración de datos usando nada más que nuestra intuición. Si creemos en la sabiduría convencional, debería existir una alta correlación positiva entre la variable de tiempo de estudio con la nota final del alumno. Además el estado de la educación de los padres se cita habitualmente como un importante predictor del éxito académico de los estudiantes en muchas publicaciones de analítica de estudiantes

Para ver lo correladas que están unas variables con otras, y también con Pass, podemos calcular y dibujar una matriz de correlación. Primero creamos variables dummies para las variables de tipo factor que queramos incluir en la matriz de correlación.



Aqui ya podemos ver de un simple vistazo que variable poseen un alto nivel de correlación y por tanto serán con las que trabjaremos en los siguientes análisis.

Podemos ver que las relaciones son bastante leves Los alumnos con más failures tienen una nota más baja. *La edad influye de forma negativa en la nota de los estudiantes.* Los alumnos con más tiempo de estudio tienen una nota más alta. *Aquellos que quieren ir a la universidad tienen una nota más alta (higher).* Se parecia una correlación fuerte del arpobado y el nivel educativo de los padres

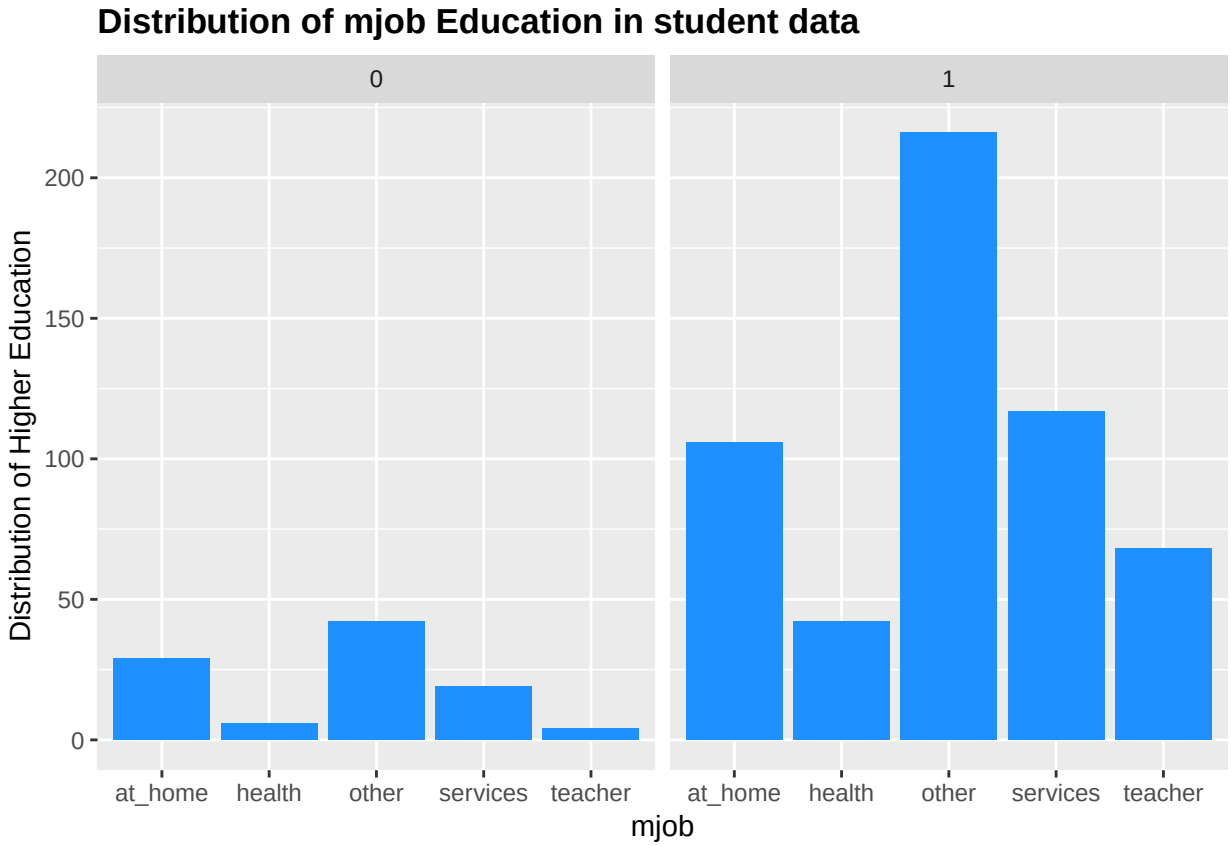
Nivel educativo de los padres

El nivel educativo de los padres tiene un alto nivel de correlación. Vamos a profundizar en el análisis de la distribución de datos.

El trabajo de la madre se agrupa en cinco categorías: ama de casa, sectores de empleo (salud, educación, otros servicios) y otros. La clasificación en salud y educación es realmente importante ya que estos dos son sumamente cruciales en la educación de un niño y una madre que trabaja en estos campos podría ayudar al niño aún más.Además, la distinción de ama de casa, empleada y otra también es importante porque las madres que hacen tareas domésticas a tiempo completo podrían enfocarse más en las necesidades de sus hijos. Todas estas hiposias deben ser probadas para veracidad.

Ahora comparemos el trabajo de la madre y el grado de los estudiantes.

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

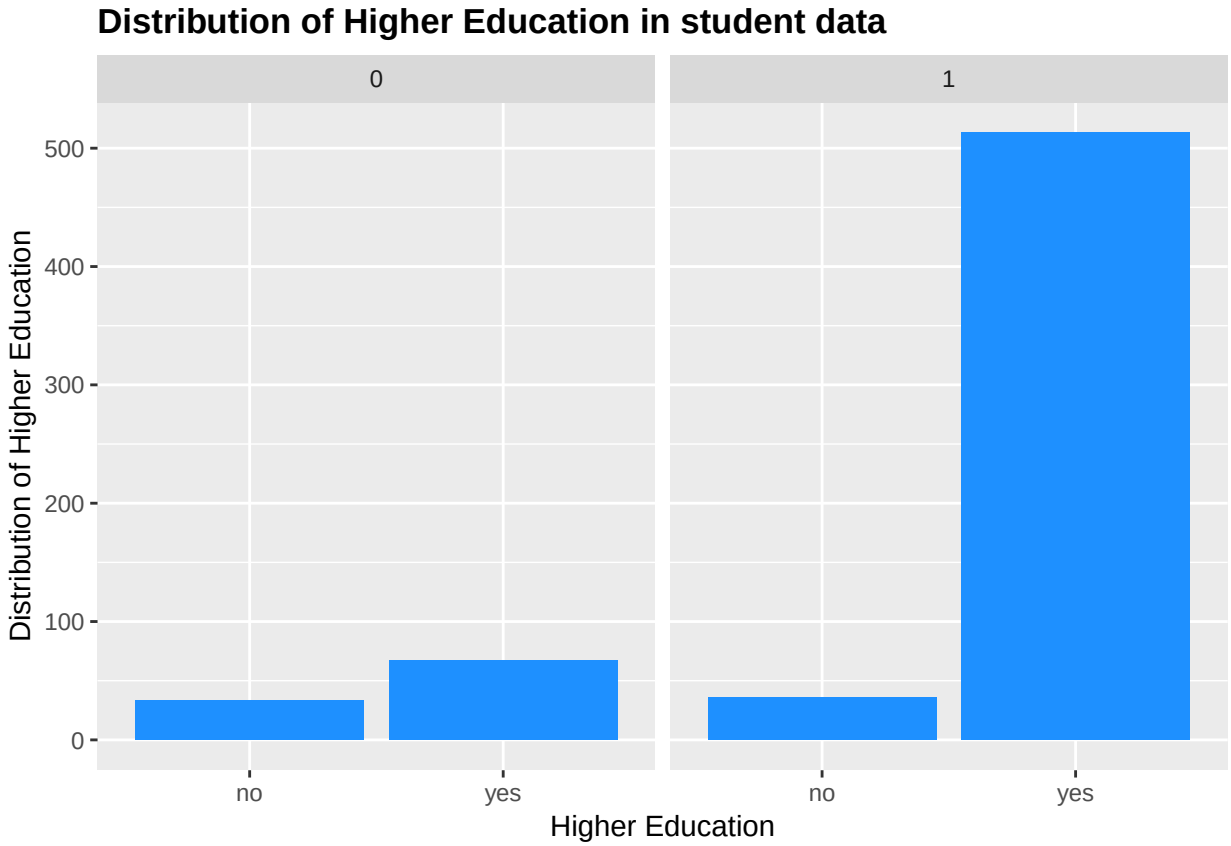


Sería conveniente una clasificación más detallada de las profesiones de las madres de los alumnos. Puede guardar alguna relación con la nota final, seguramente en combinación con otras variables, pero hay muchas instancias de "Other". Los hijos de madres que son amas de casa tienen más probabilidades de tener un desempeño inferior al promedio que otros niños. Mientras tanto, es más probable que los niños cuyas madres trabajan en el sector de la salud tengan un desempeño superior al promedio.

Educación Superior

Pasemos a la distribución de la educación superior. Esta variable representa la disposición de los estudiantes para continuar su educación superior. Está bastante sesgado como se esperaba. Muchos estudiantes han respondido que sí. Es posible que tengamos que igualar la distribución antes de determinar si esto proporciona información significativa sobre las calificaciones.

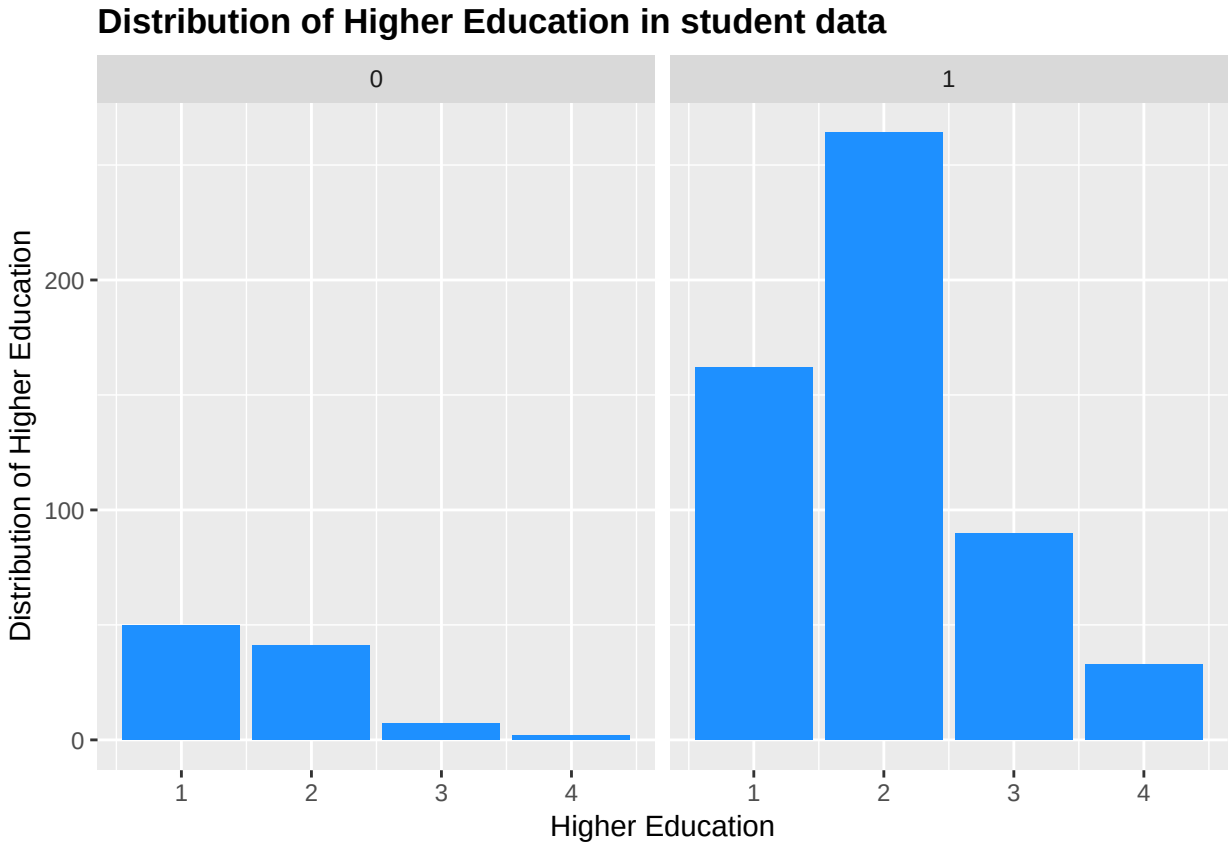
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Dedicación al estudio

Como es lógico la dedicación al estudio influye con el aprobado y el suspendo. Esta variable, al no poseer información relevante, la descartaremos para así evitar que se generen reglas que no aporten valor.

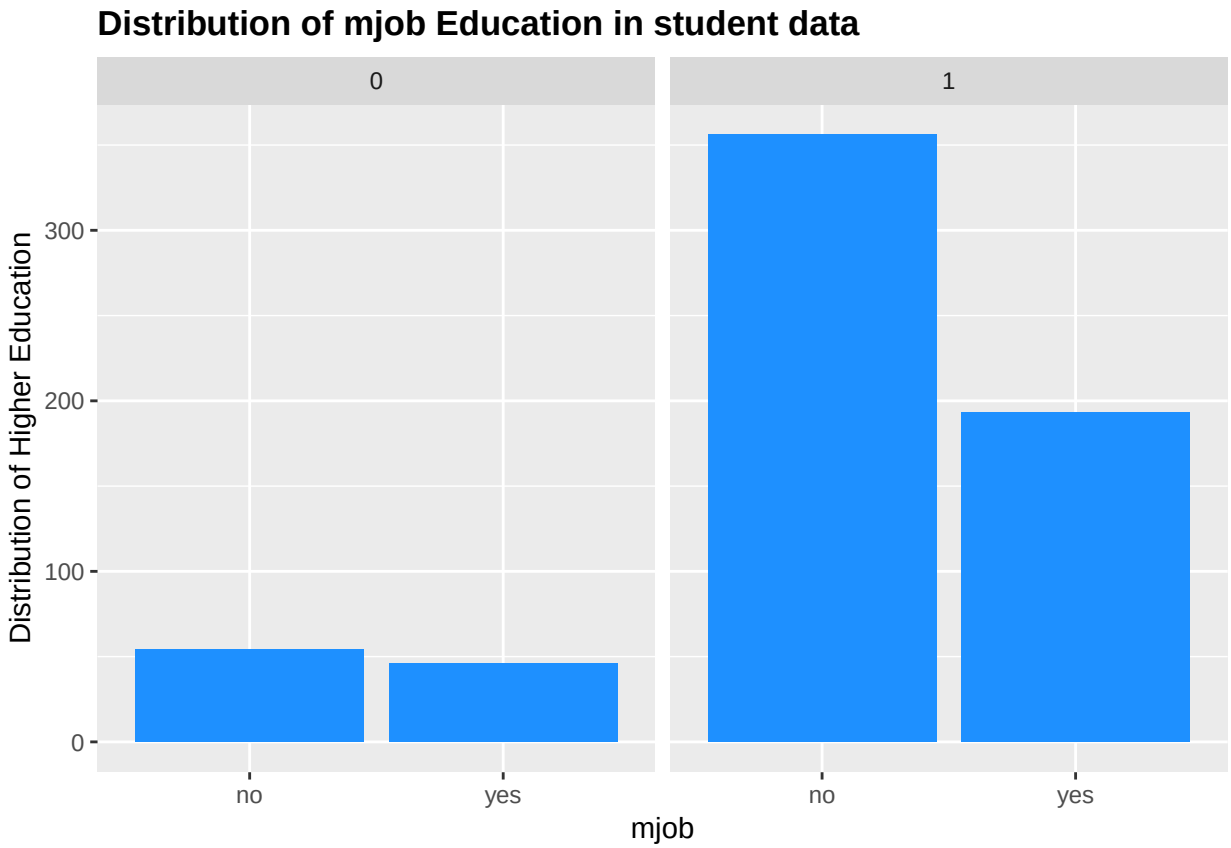
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Relación amorosa

Los alumnos con relación amorosa o no, es determinante a la hora de aprobar. Como se aprecia en el plot, influye negativamente que el alumno se encuentre dentro de una relación amorosa.

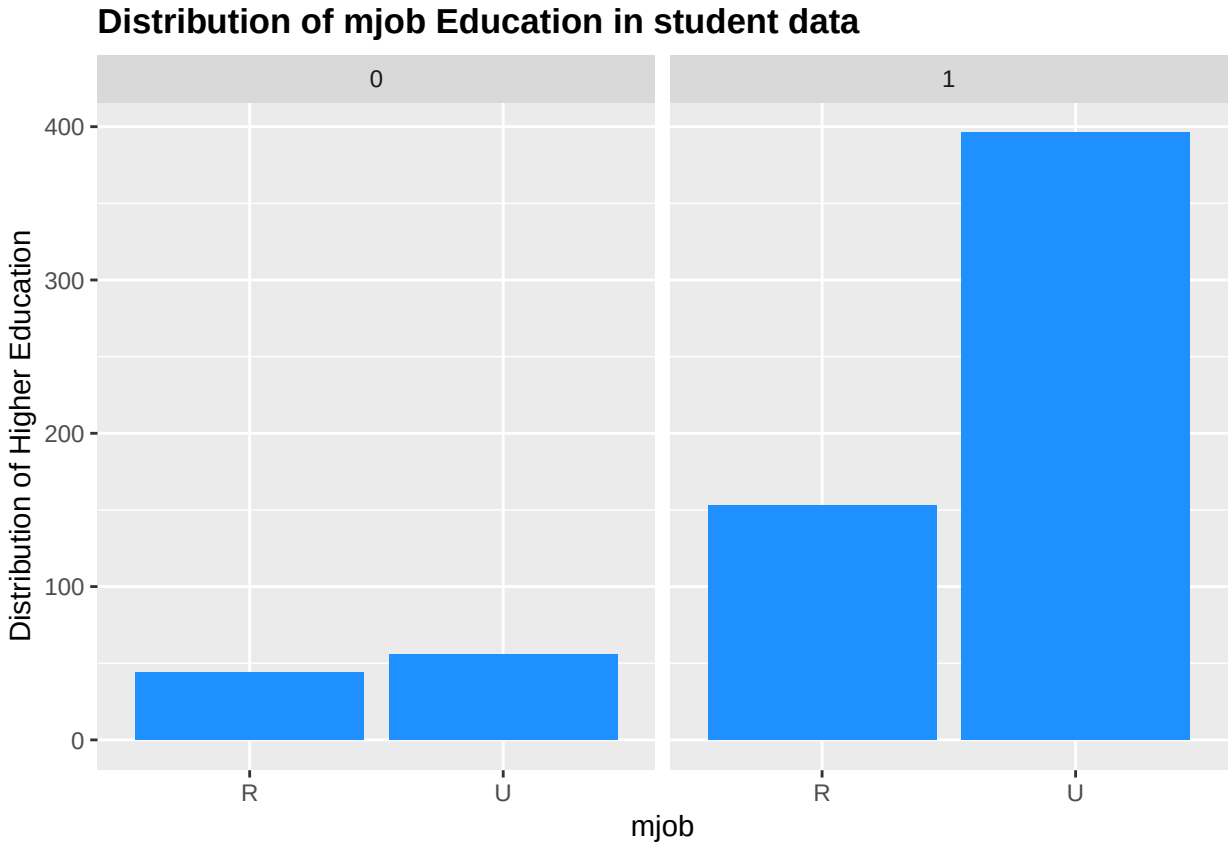
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Dirección del alumno (address)

La dirección (address) del alumno posee cierta correlación con la nota final. A primera vista, es dificil entender dar una explicación de dicha correlación aunque puede que esta influya directamente con otra variable. Por ejemplo, el hecho que viva en el campo o en la ciudad influye bastante en el tipo los estudios de sus padres.

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



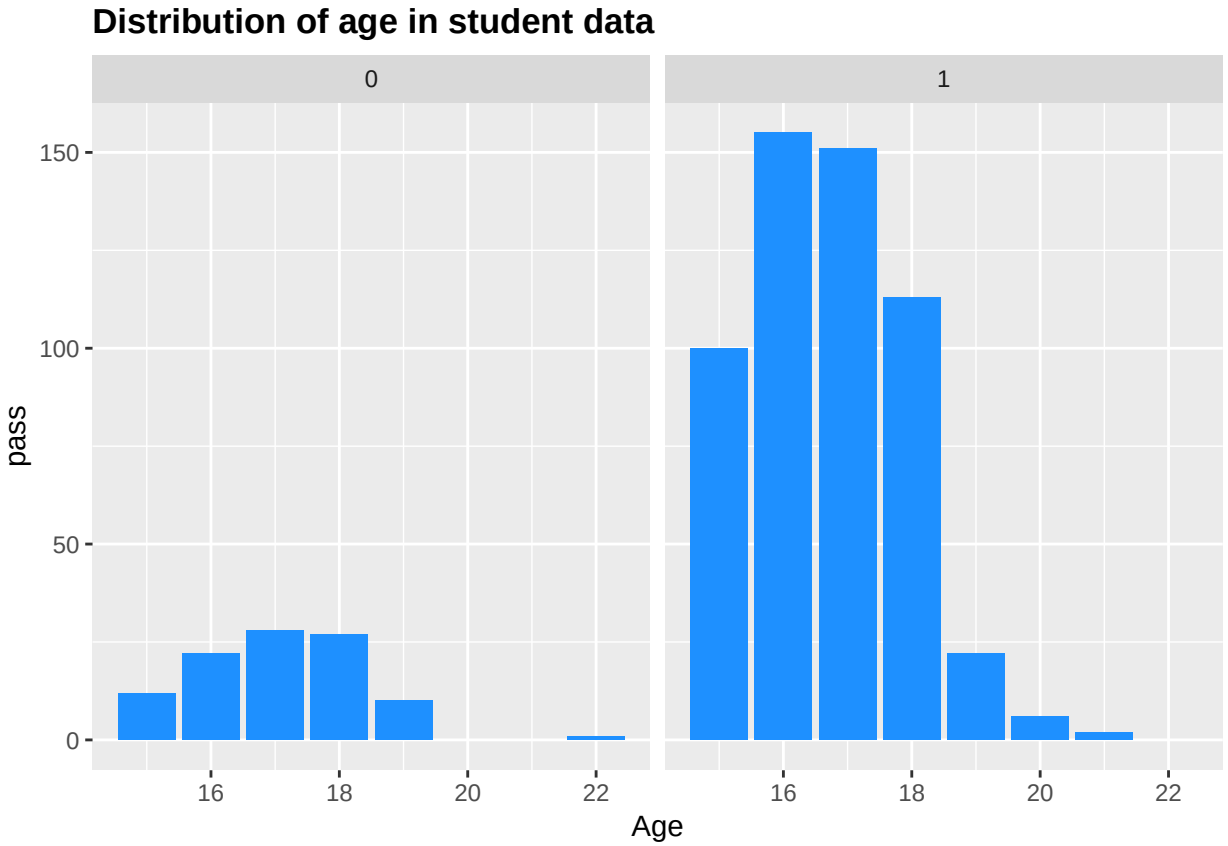
Edad del alumno

La nota media final está por debajo de aprobado en las edades de 19 y 22. En el gráfico siguiente hay pocas observaciones a partir de las edades de 19 años por lo que la media es muy sensible a valores concretos

Media de notas finales por edad:

```
##   age     mean median  sum
## 1  15 12.10714     12 1356
## 2  16 11.99435     12 2123
## 3  17 12.26816     12 2196
## 4  18 11.77143     12 1648
## 5  19  9.53125     10  305
## 6  20 12.00000     11   72
## 7  21 11.00000     11   22
## 8  22  5.00000      5    5
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Reglas de Asociación

Se han aplicado reglas de asociación con el objetivo de detectar asociaciones entre las diferentes variables del dataset, de forma que podamos comprender mejor dicho dataset y qué variables son mas interesantes de cara a clasificar y evaluar los estudiantes.

Las menos correladas están son failures, school, age o walc entre otras. Estas las descartaremos en los siguientes análisis.

En primer lugar aplicamos el algoritmo Apriori, con un soporte minimo de 0.09 y una confianza de 0.8.

```
##     lhs                rhs      support   confidence lift      count
## [1] {higher=yes}   => {pass=1} 0.7904468 0.8844828  1.0455907 513
## [2] {address=U}    => {pass=1} 0.6101695 0.8761062  1.0356884 396
## [3] {romantic=no}  => {pass=1} 0.5485362 0.8682927  1.0264516 356
## [4] {famsup=yes}   => {pass=1} 0.5254237 0.8567839  1.0128466 341
## [5] {famsup=no}    => {pass=1} 0.3204931 0.8286853  0.9796298 208
## [6] {romantic=yes} => {pass=1} 0.2973806 0.8075314  0.9546227 193
```

Como podemos ver, obtenemos un conjunto de 16 reglas. Ordenadas por la medida de soporte, las 3 primeras reglas nos proporcionan gran cantidad de información, puesto que nos indican que en un 80% (soporte=0.80) de los casos de nuestro dataset la nota del alumno será aprobada en caso de que quiera estudiar estudios superiores o que un 50% de los alumnos arpobado son los que no tienen pareja sentimental o tiene soporte educacional por parte de sus padres. Finalmente un 61% de los alumnos aprobados viven el la ciudad.

Esto nos indica que, en general, segun nuestro dataset aquellos alumnos que cumplan alguna o ambas de estas reglas obtendran, probablemente, un aprobado en la notal final. Por tanto, las eliminaremos, de los

posteriores analisis para evitar reglas innecesarias y que se pueden intuir a simple vista

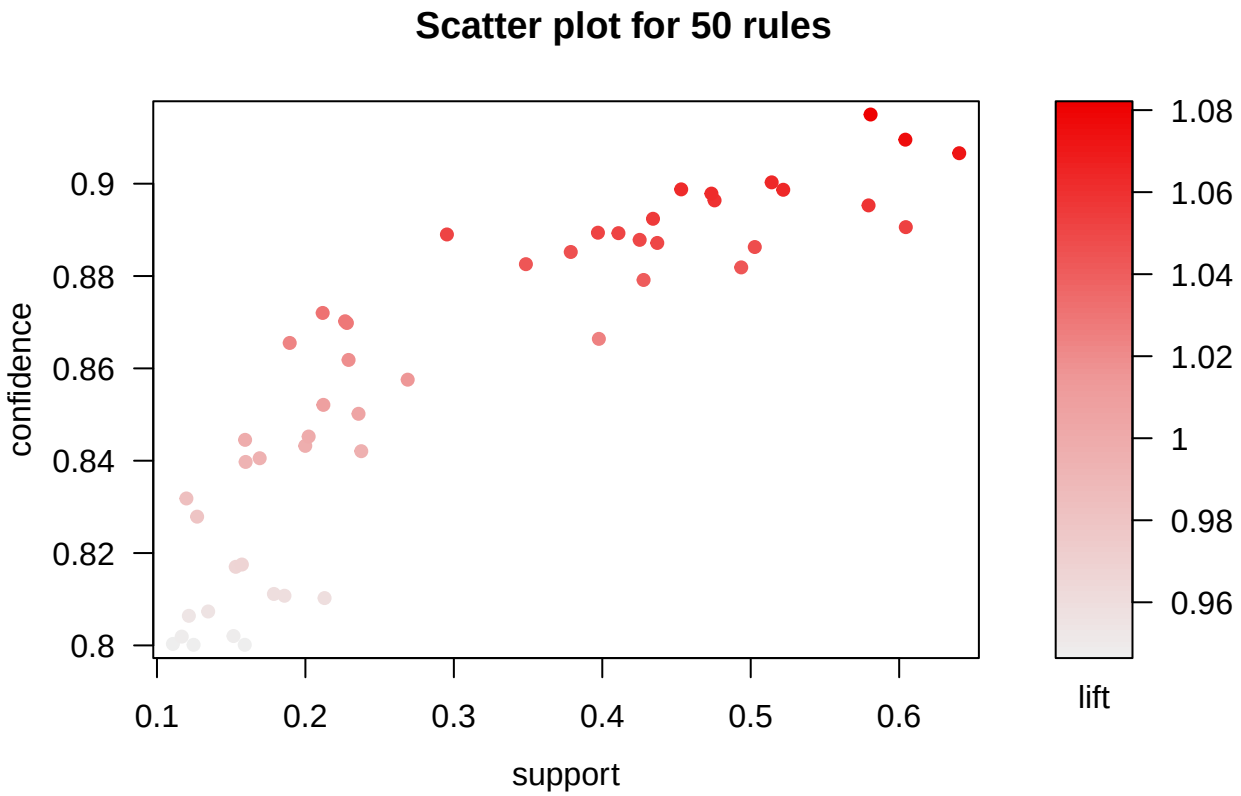
```
##     lhs             rhs      support   confidence lift     count
## [1] {fedu=3}     => {pass=1} 0.1864407 0.9236641  1.091909 121
## [2] {medu=4}     => {pass=1} 0.2480740 0.9200000  1.087577 161
## [3] {age=15}     => {pass=1} 0.1540832 0.8928571  1.055491 100
## [4] {fedu=4}     => {pass=1} 0.1756549 0.8906250  1.052852 114
## [5] {higher=yes} => {pass=1} 0.7904468 0.8844828  1.045591 513
## [6] {address=U}  => {pass=1} 0.6101695 0.8761062  1.035688 396
```

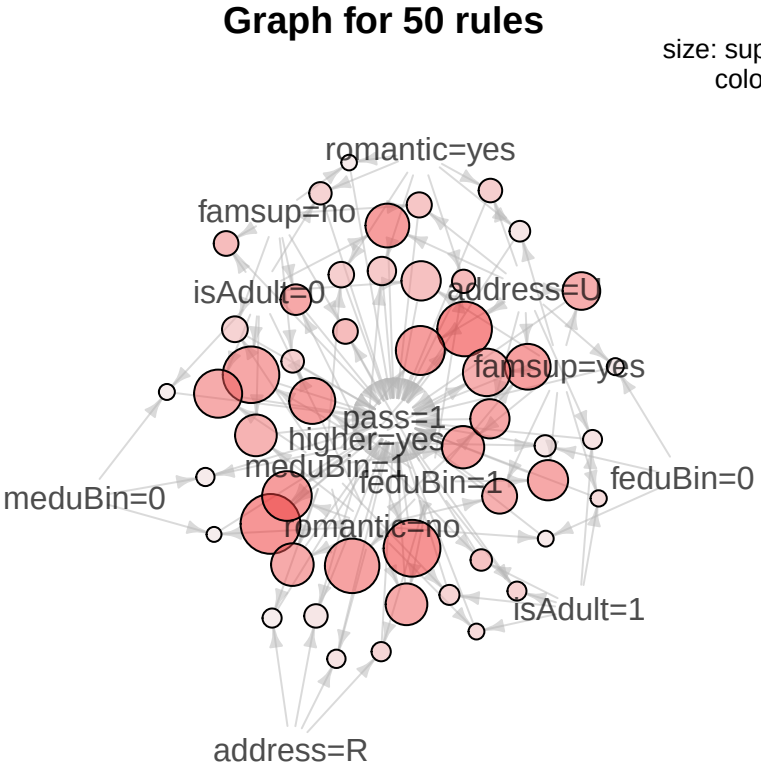
Si ordenamos por *lift* podemos observar que existe una gran dependencia entre la educacón del padreo de la madre con aprobar o no, además de las variaable "romantic" , "higher" o "famsup" anteriormente comentado.

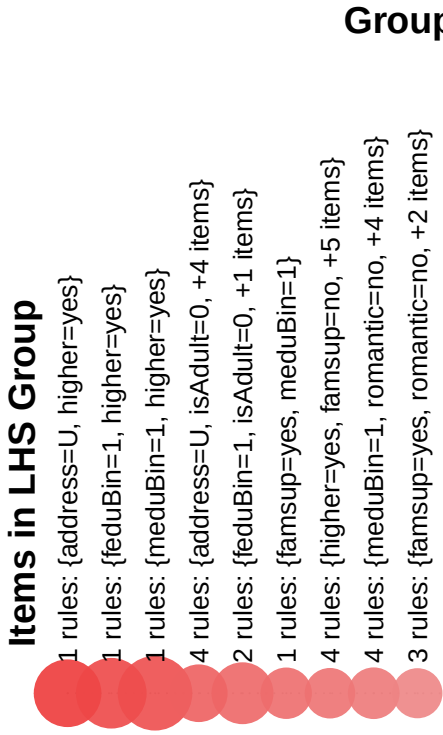
Si queremos acotar y seguir analizando los ítemsets frecuentes, usaremos rangos definidos por las variables **minlen** y **maxlen** para definir cuantos ítems queremos que formen los sets.

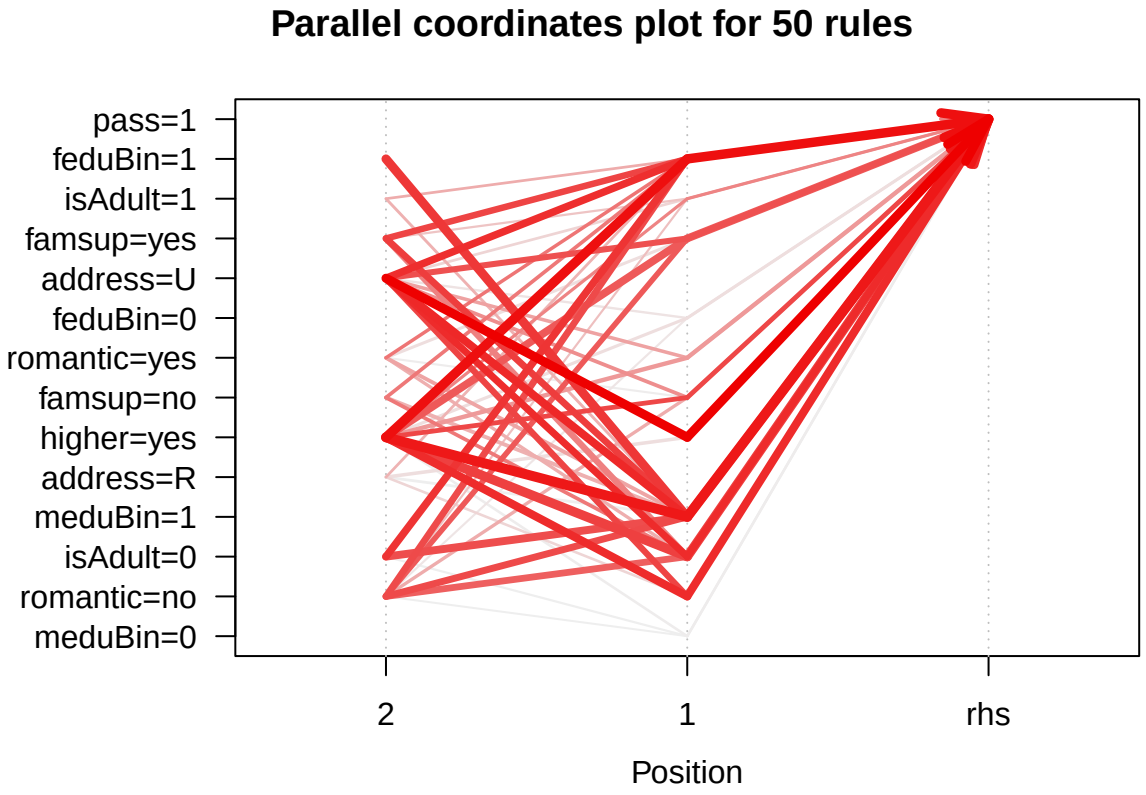
```
##     lhs                          rhs      support   confidence lift
## [1] {higher=yes,address=U}   => {pass=1} 0.5793529 0.9148418  1.081480
## [2] {feduBin=1,higher=yes}   => {pass=1} 0.6055470 0.9097222  1.075428
## [3] {meduBin=1,higher=yes}   => {pass=1} 0.6409861 0.9063181  1.071403
## [4] {meduBin=1,address=U}    => {pass=1} 0.5130971 0.9000000  1.063934
## [5] {romantic=no,higher=yes} => {pass=1} 0.5208012 0.8989362  1.062677
## [6] {isAdult=0,address=U}    => {pass=1} 0.4514638 0.8987730  1.062484
##     count
## [1] 376
## [2] 393
## [3] 416
## [4] 333
## [5] 338
## [6] 293
```

```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```


Scatter plot for 50 rules
confidence
support
lift
0.8
0.82
0.84
0.86
0.88
0.9
0.1
0.2
0.3
0.4
0.5
0.6
1.08
1.06
1.04
1.02
1
0.98
0.96


Graph for 50 rules
size: support (0.111 − 0.641)
color: lift (0.946 − 1.081)
romantic=yes
famsup=no
isAdult=0
address=U
famsup=yes
pass=1
higher=yes
meduBin=1
feduBin=1
feduBin=0
meduBin=0
romantic=no
isAdult=1
address=R


Group
Items in LHS Group
1 rules: {address=U, higher=yes}
1 rules: {feduBin=1, higher=yes}
1 rules: {meduBin=1, higher=yes}
4 rules: {address=U, isAdult=0, +4 items}
2 rules: {feduBin=1, isAdult=0, +1 items}
1 rules: {famsup=yes, meduBin=1}
4 rules: {higher=yes, famsup=no, +5 items}
4 rules: {meduBin=1, romantic=no, +4 items}
3 rules: {famsup=yes, romantic=no, +2 items}


Parallel coordinates plot for 50 rules
pass=1
feduBin=1
isAdult=1
famsup=yes
address=U
feduBin=0
romantic=yes
famsup=no
higher=yes
address=R
meduBin=1
isAdult=0
romantic=no
meduBin=0
2
1
rhs
Position