Extracción Automática del Conocimiento R-PL2

Zamar Elahi Fazal Roura

March 12, 2018

1 Análisis asociación estudiantes

1.1 Ejercicio con dataset Kaggle

Desarrollo un ejercicio en el que se realice un análisis con R de asociación.

Obtención de los datos

UsaMOS el fichero **student-por.csv** que hemos encontrado en Kaggle [2]. Es un dataset que representa un estudio sobre los alumnos de una escuela de secundaria, contiene datos interesantes de temática social, de género y relacionada con los estudios. Como el dataset es muy grande, hemos decidido acotarlo en número de filas, ademáss de elegir solo unas cuantas columnas. La descripción de los datos se encuentra en `https://www.kaggle.com/uciml/student-alcohol-consumption`.

Usaremos 'arulesViz' [3], un paquete de R que proporciona funciones gráficas al paquete 'arules'

```
dat = read.csv("student-por.csv")
```

```
summary(dat)
```

```
##  school   sex          age        address famsize   Pstatus      Medu
##  GP:423   F:383   Min.   :15.00   R:197   GT3:457   A: 80   Min.   :0.000
##  MS:226   M:266   1st Qu.:16.00   U:452   LE3:192   T:569   1st Qu.:2.000
##                   Median :17.00                             Median :2.000
##                   Mean   :16.74                             Mean   :2.515
##                   3rd Qu.:18.00                             3rd Qu.:4.000
##                   Max.   :22.00                             Max.   :4.000
##       Fedu             Mjob           Fjob            reason      guardian
##  Min.   :0.000   at_home :135   at_home : 42   course    :285   father:153
##  1st Qu.:1.000   health  : 48   health  : 23   home      :149   mother:455
##  Median :2.000   other   :258   other   :367   other     : 72   other : 41
##  Mean   :2.307   services:136   services:181   reputation:143
##  3rd Qu.:3.000   teacher : 72   teacher : 36
##  Max.   :4.000
##    traveltime      studytime        failures      schoolsup famsup     paid
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :581   no :251   no :610
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 68   yes:398   yes: 39
##  Median :1.000   Median :2.000   Median :0.0000
##  Mean   :1.569   Mean   :1.931   Mean   :0.2219
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
##  Max.   :4.000   Max.   :4.000   Max.   :3.0000
##  activities nursery   higher    internet  romantic      famrel         freetime
##  no :334    no :128   no : 69   no :151   no :410   Min.   :1.000   Min.   :1.00
##  yes:315    yes:521   yes:580   yes:498   yes:239   1st Qu.:4.000   1st Qu.:3.00
##                                                     Median :4.000   Median :3.00
##                                                     Mean   :3.931   Mean   :3.18
##                                                     3rd Qu.:5.000   3rd Qu.:4.00
##                                                     Max.   :5.000   Max.   :5.00
##      goout            Dalc            Walc          health         absences
##  Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   : 0.000
```

```
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:2.000   1st Qu.: 0.000
##  Median :3.000   Median :1.000   Median :2.00   Median :4.000   Median : 2.000
##  Mean   :3.185   Mean   :1.502   Mean   :2.28   Mean   :3.536   Mean   : 3.659
##  3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd Qu.:5.000   3rd Qu.: 6.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :32.000
##        G1              G2              G3
##  Min.   : 0.0   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:10.0   1st Qu.:10.00   1st Qu.:10.00
##  Median :11.0   Median :11.00   Median :12.00
##  Mean   :11.4   Mean   :11.57   Mean   :11.91
##  3rd Qu.:13.0   3rd Qu.:13.00   3rd Qu.:14.00
##  Max.   :19.0   Max.   :19.00   Max.   :19.00
```

Objetivo

Queremos realizar un análisis de asociación para investigar que hechos ocurren en común entre un grupo de estudiantes. Realizaremos el análisis también para solo las mujeres estudiantes.

También estudiaremos la visualización del concepto de soporte y el comportamiento del "mineo" de reglas de asociación. Para ello trataremos 3 conjuntos:

- Estudiantes (sin diferenciar sexo)
- Estudiantes (diferenciando sexo).
- Mujeres Estudiantes (subconjunto de estudiantes).

Por último veremos que soporte y confianza tienen las reglas que indican que *los estudiantes beben mucho alcohol los fines de semana*

Limpieza de datos

Cambiamos las variables nominales para que tengan significado.

```
head(dat$health,n=10)
```

```
##  [1] 3 3 3 5 5 5 3 1 1 5
```

Transformamos:

```
#Escogemos las columnas que queremos pasar a "mine"
mine <- dat[,c("Walc","Dalc","failures","health","Pstatus")]
mine$Walc <- factor(mine$Walc, labels = c("very low", "low","normal","high","very high"))
mine$Dalc <- factor(mine$Dalc, labels = c("very low", "low","normal","high","very high"))
mine$failures <- factor(mine$failures, labels = c("cero", "uno","dos","tres"))
mine$health <- factor(mine$health, labels = c("very bad","bad","normal","good","very good"))
mine$Pstatus <- factor(mine$Pstatus, labels = c("apart","living together"))
```

Resultado:

```
head(mine$health, n=10)
```

```
##  [1] normal    normal    normal    very good very good very good normal    very bad
##  [9] very bad  very good
## Levels: very bad bad normal good very good
```

Cambiamos las variables binarias para que tengan significado.

```
summary(dat$sex)

##   F   M
## 383 266
```

Transformamos:

```
mine$male <- dat$sex == "M"
mine$female <-dat$sex =="F"
mine$notRomantic <- dat$romantic == "no"
mine$romantic <-dat$romantic =="yes"
```

Resultado:

```
summary(mine$female)

##    Mode   FALSE    TRUE
## logical     266     383
```

Para estudiar a los alumnos como estudiantes sin tener en cuenta el género, creamos una tabla donde no guardamos la columna.

```
students <- mine
students$female <- NULL
students$male <- NULL
```

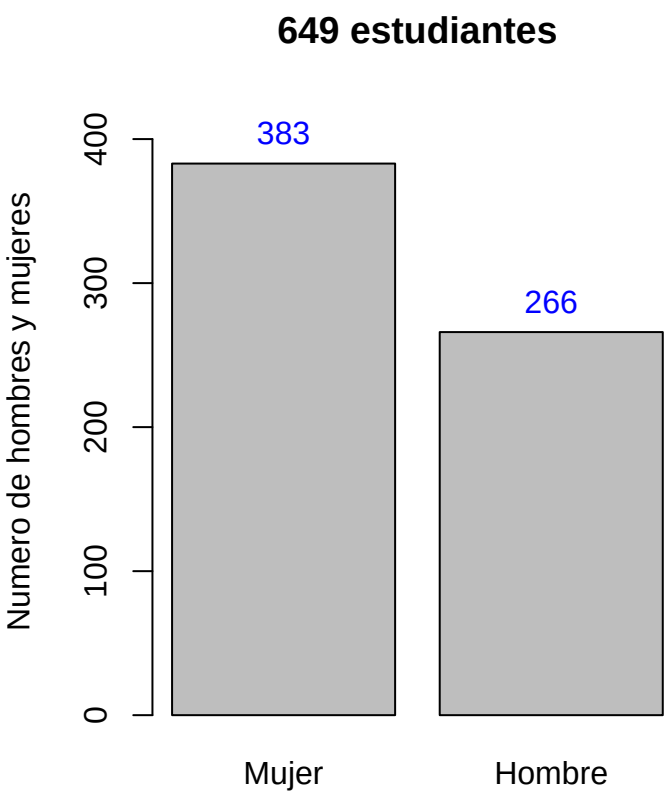
Todas las variables que no introducimos son eliminadas. Comprobamos el resumen de los datos **limpios**

```
summary(mine)

##         Walc            Dalc     failures         health               Pstatus
##  very low :247   very low :451   cero:549   very bad : 90   apart          : 80
##  low      :150   low      :121   uno : 70   bad      : 78   living together:569
##  normal   :120   normal   : 43   dos : 16   normal   :124
##  high     : 87   high     : 17   tres: 14   good     :108
##  very high: 45   very high: 17              very good:249
##     male           female        notRomantic      romantic
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:383       FALSE:266       FALSE:239       FALSE:410
##  TRUE :266       TRUE :383       TRUE :410       TRUE :239
##
##
```

Entender el contexto

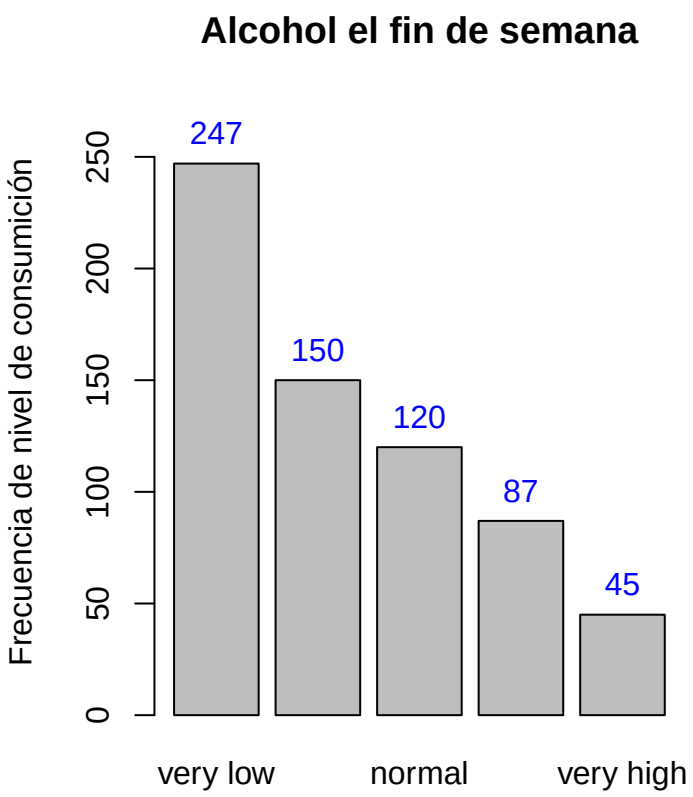
Número de estudiantes

```
tbl <- table(factor(dat$sex, labels = c("Mujer","Hombre")))
ylim <- c(0, 1.1*max(table(dat$sex)))
xx <- barplot(
    tbl,ylab="Numero de hombres y mujeres",ylim=ylim,main=paste(
        nrow(dat),"estudiantes",sep=" "))
text(x=xx,y=table(dat$sex),label=table(dat$sex), pos = 3, col="blue")
```



Frecuencia de beber alcohol el fin de semana

```
tbl <- table(mine$Walc)
ylim <- c(0, 1.1*max(table(mine$Walc)))
xx <- barplot(tbl,ylab="Frecuencia de nivel de consumición",ylim=ylim,main="Alcohol el fin de semana")
text(x=xx,y=table(mine$Walc),label=table(mine$Walc), pos = 3, col="blue")
```



Podemos observar que la moda (247) es beber muy poco alcohol. Esta frecuencia tan alta nos da una pista sobre cómo el mayor soporte (entre reglas de alcohol) va a estar en reglas donde casi no se bebe alcohol entre semana.

Tratamiento de datos

En los apartados anteriores hemos discretizado la información, esto es, separar los datos convirtiendolos en únicos y distintos. De esta manera podemos transformar nuestros data frames a transacciones, en la cual se guarda nuestra información en una matriz de incidencias binaria (TRUE O FALSE)

```
#Tratar como estudiantes
trans_students <- as(students,"transactions")
#Diferenciamos género, en el objetivo hemos especificado que estudiamos el femenino.
trans <- as(mine,"transactions")
#Subconjunto
trans_female <- subset(trans, items %in% "female")
```

Por un lado guardamos a los estudiantes, en otro separamos por hombre y mujer (mine) y en el otro solo las mujeres.

```
summary(trans)

## transactions as itemMatrix in sparse format with
##  649 rows (elements/itemsets/transactions) and
##  25 columns (items) and a density of 0.28
```

```
##
## most frequent items:
## Pstatus=living together           failures=cero           Dalc=very low
##                     569                     549                     451
##             notRomantic                  female                  (Other)
##                     410                     383                    2181
##
## element (itemset/transaction) length distribution:
## sizes
##   7
## 649
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       7       7       7       7       7       7
##
## includes extended item information - examples:
##          labels variables   levels
## 1 Walc=very low      Walc very low
## 2      Walc=low      Walc      low
## 3   Walc=normal      Walc   normal
##
## includes extended transaction information - examples:
##   transactionID
## 1             1
## 2             2
## 3             3
```

En el "summary" de la transacción de mujeres debería haber solo 383 filas.

```
summary(trans_female)
```

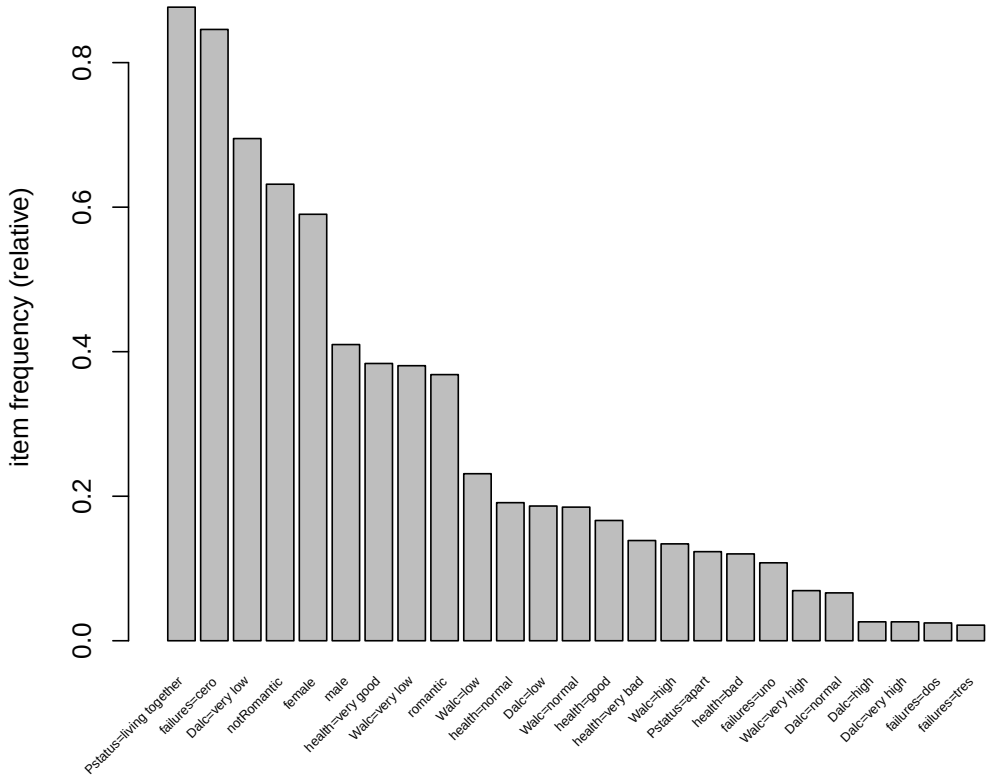
```
## transactions as itemMatrix in sparse format with
##  383 rows (elements/itemsets/transactions) and
##  25 columns (items) and a density of 0.28
##
## most frequent items:
##                   female           failures=cero Pstatus=living together
##                      383                     329                     329
##           Dalc=very low             notRomantic                  (Other)
##                      305                     225                    1110
##
## element (itemset/transaction) length distribution:
## sizes
##   7
## 383
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       7       7       7       7       7       7
##
## includes extended item information - examples:
##          labels variables   levels
## 1 Walc=very low      Walc very low
## 2      Walc=low      Walc      low
## 3   Walc=normal      Walc   normal
##
## includes extended transaction information - examples:
##   transactionID
## 1             1
## 2             2
## 3             3
```

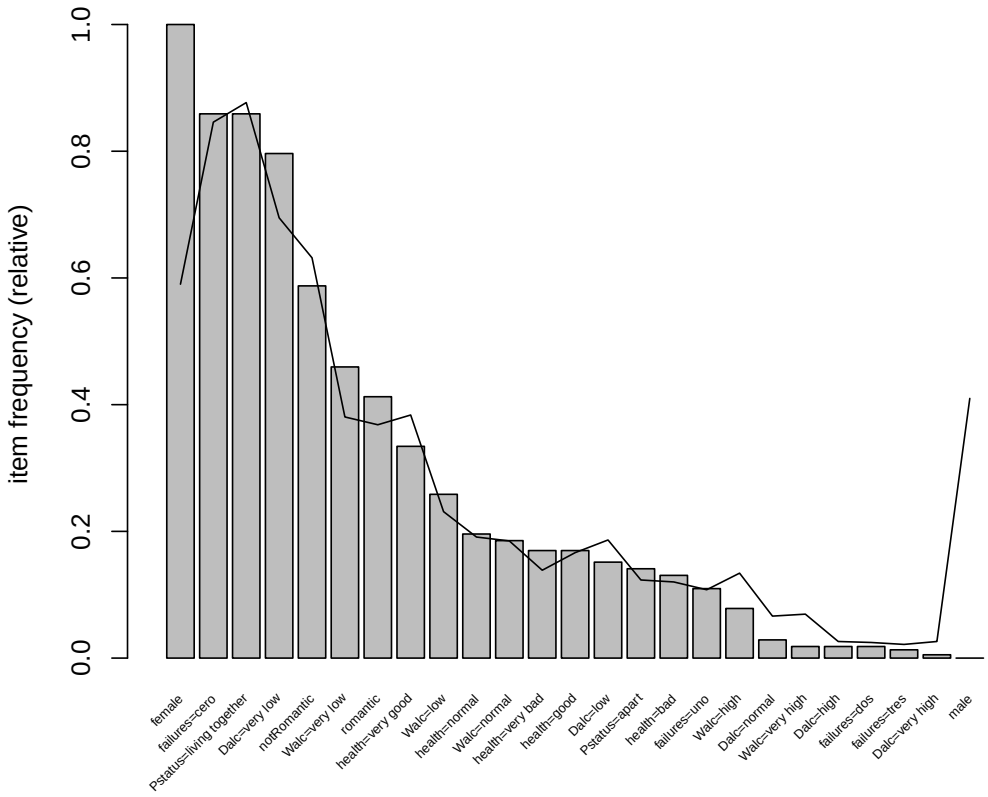
En la primera gráfica vemos la frecuencia de "aparición" de los hechos en la transacción, en la segunda gráfica la frecuencia de aparición de hechos para las mujeres.

```
itemLabels(trans)
```

```
##  [1] "Walc=very low"          "Walc=low"               "Walc=normal"
##  [4] "Walc=high"              "Walc=very high"         "Dalc=very low"
##  [7] "Dalc=low"               "Dalc=normal"            "Dalc=high"
## [10] "Dalc=very high"         "failures=cero"          "failures=uno"
## [13] "failures=dos"           "failures=tres"          "health=very bad"
## [16] "health=bad"             "health=normal"          "health=good"
## [19] "health=very good"       "Pstatus=apart"          "Pstatus=living together"
## [22] "male"                   "female"                 "notRomantic"
## [25] "romantic"
```

```
itemFrequencyPlot(trans,topN = 30,cex.names=.5)
itemFrequencyPlot(trans_female,population=trans,topN = 30,cex.names=.5)
```





En la ultima gráfica la linea da el valor de la frecuencia del "item" en el conjunto total (Lo que llamamos población en la función)

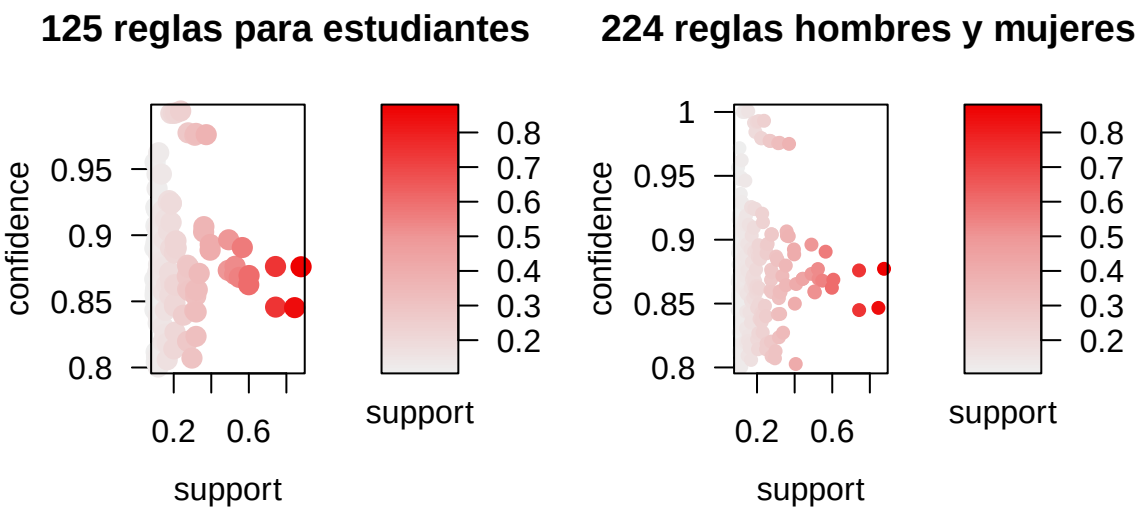
Análisis de Asociación

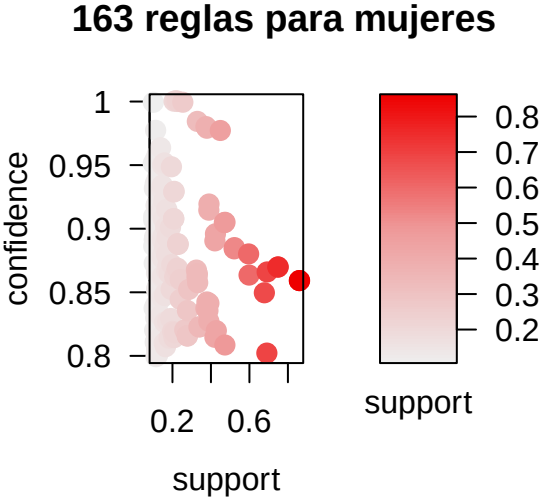
Con el objetivo claro, los datos estructurados, limpios y tratados comenzamos a realizar el "mineo" para sacar los hechos frecuentes.

```
total_rules_students<-apriori(trans_students)
total_rules <- apriori(trans)
#Ya sabemos que todo son mujeres, sacamos a las mujeres.
total_rules_female <- apriori(trans_female, appearance =list(none=c("female")))
```

«results='hide'»= para ocultar los resultados pero dejar las funciones.

```
plot(total_rules_students,shading='support',control=list(main = paste(length(total_rules_students),"reglas para est
plot(total_rules,shading='support',control=list(main = paste(length(total_rules),"reglas hombres y mujeres",sep=" "
plot(total_rules_female,shading='support',control=list(main = paste(length(total_rules_female),"reglas para mujeres
```





Aquí podemos visualizar el número de reglas que salen con el mismo soporte y confianza para estudiantes, hombres y mujeres, y sólo para las mujeres.

Comparando **estudiantes** y **estudiantes diferenciados por sexo** observamos que hay menos reglas para el primero pero las reglas que comparten entre uno y otro tienen el mismo soporte y confianza. Esto es debido a que hay el **mismo número de sucesos** para cada uno.La única diferencia es que en los sucesos en los que diferenciamos hombre y mujer existirá un elemento más dentro del suceso. En un carrito de comidas donde hay [Pan, Leche] también incluiríamos quién ha comprado el carrito, esto es, [Pan,Leche,Mujer] .Y como hemos incluido dos nuevos sucesos elementales en nuestro espacio muestral saldrán más reglas.

Comparando **estudiantes (diferenciados por sexo)** y **Subconjunto de estudiantes que son mujer**sabemos que el número de sucesos que existe dentro del subconjunto de mujeres es menor que el total de estudiantes [383 < 649]. El soporte ahora se calcula para un número menor de sucesos y saldrá un soporte acorde para el comportamiento dentro de las mujeres. Si no excluyeramos a [female] dentro de las reglas de asociación saldrían muchas reglas con confianza 1 ya que los hechos que pasan el soporte nos indican que somos una mujer, por eso lo hemos excluido.

Curiosidades:

Si en la función de plot escribimos como argumento (engine='interactive') podemos movernos a traves de las

reglas por regiones y ver cuales son haciendo doble click y aprentando el botón "inspect"

Por defecto el color de los puntos es dado por "lift", para nuestro entendimiento hemos cambiado el coloreo por nivel de soporte con shading='support'

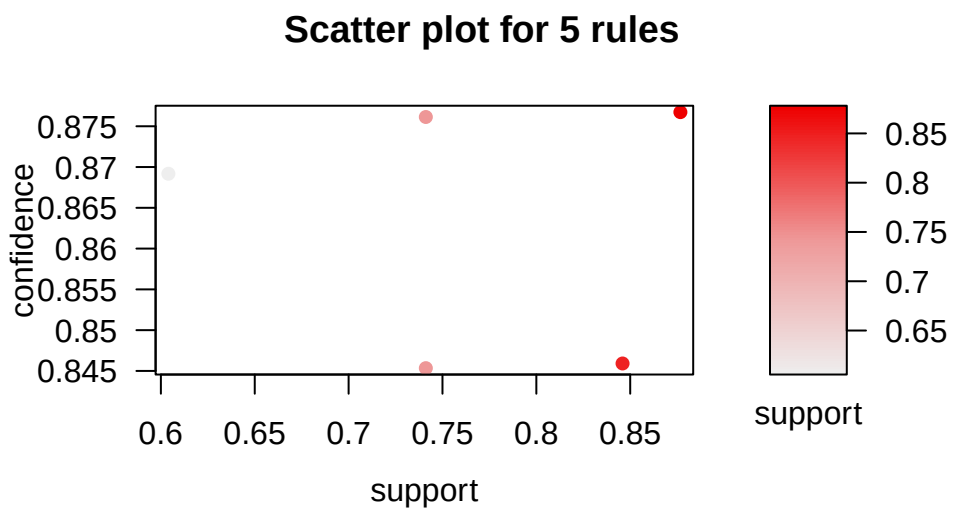
Estudiantes

Este es el caso en el que no diferenciamos hombres y mujeres, dividimos estos casos para que se aprecie Creamos una función donde establecemos el soporte y confianza deseados (0.6 y 0.8 respectivamente)

```
apriori_p <- function (x){
    apriori(x,parameter=list(supp=0.6,conf=0.8))
}
```

Observamos que reglas salen con los umbrales establecidos.

```
rules_selection_students <- apriori_p(trans_students)
plot(rules_selection_students,shading='support')
```




```
inspect(rules_selection_students)
```

```
##     lhs                            rhs                          support   confidence lift
## [1] {}                          => {failures=cero}              0.8459168 0.8459168  1.0000000
## [2] {}                          => {Pstatus=living together} 0.8767334 0.8767334  1.0000000
## [3] {Dalc=very low}             => {failures=cero}              0.6040062 0.8691796  1.0275001
## [4] {failures=cero}             => {Pstatus=living together} 0.7411402 0.8761384  0.9993213
## [5] {Pstatus=living together} => {failures=cero}              0.7411402 0.8453427  0.9993213
##     count
## [1] 549
## [2] 569
## [3] 392
## [4] 481
## [5] 481
```

Con estos umbrales podemos decir, siempre que:

- Se bebe muy poco alcohol entre semana se aprueba todo
- Si apruebas todo tus padres están juntos
- Si tus padres están juntos apruebas todo

Hombres y Mujeres

Observamos que reglas salen con los umbrales establecidos.

```
rules_selection <- apriori_p(trans)
plot(rules_selection,shading='support')
```

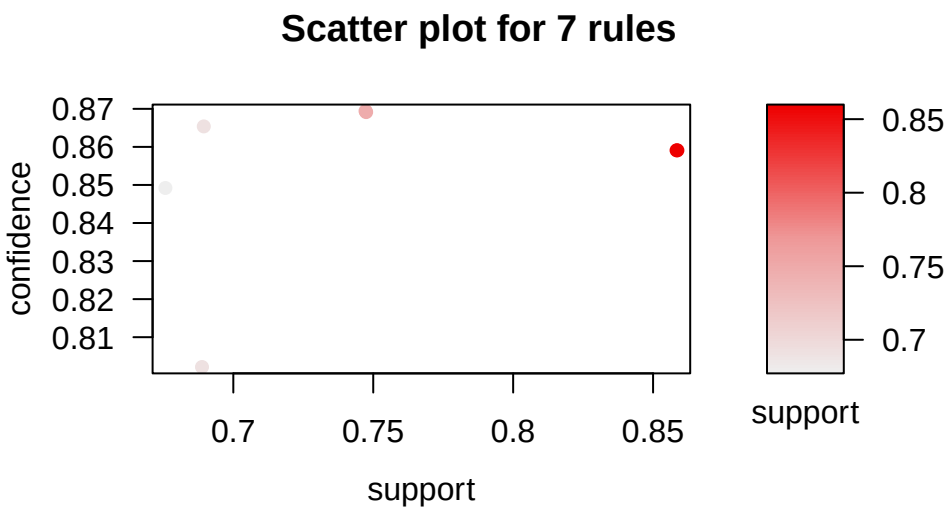

```
inspect(rules_selection)

##     lhs                           rhs                          support   confidence lift
## [1] {}                         => {failures=cero}              0.8459168 0.8459168  1.0000000
## [2] {}                         => {Pstatus=living together} 0.8767334 0.8767334  1.0000000
## [3] {Dalc=very low}            => {failures=cero}              0.6040062 0.8691796  1.0275001
## [4] {failures=cero}            => {Pstatus=living together} 0.7411402 0.8761384  0.9993213
## [5] {Pstatus=living together} => {failures=cero}              0.7411402 0.8453427  0.9993213
##     count
## [1] 549
## [2] 569
## [3] 392
## [4] 481
## [5] 481
```

Corroboramos lo que hemos mencionado anteriormente, las reglas que comparten ambos conjuntos (En este caso las mismas pasan el umbral) tienen el mismo soporte y confianza.

Mujeres

Dentro de las reglas de mujeres aparecerá el suceso elemental [Female] en todos los casos, por lo tanto no queremos que se muestren las reglas que indiquen si X => [Female].

```
#Apriori distinto ya que no puede aparecer female en las reglas, ya sabemos que son todo mujeres.
rules_selection_female <- apriori(trans_female,parameter=list(supp=0.6,conf=0.8),appearance =list(none=c("female")))
plot(rules_selection_female,shading='support')
```




```
inspect(rules_selection_female)
```

```
##     lhs                           rhs                          support   confidence lift
## [1] {}                         => {failures=cero}              0.8590078 0.8590078  1.0000000
## [2] {}                         => {Pstatus=living together} 0.8590078 0.8590078  1.0000000
## [3] {Dalc=very low}            => {failures=cero}              0.6892950 0.8655738  1.0076436
## [4] {failures=cero}            => {Dalc=very low}              0.6892950 0.8024316  1.0076436
## [5] {Dalc=very low}            => {Pstatus=living together} 0.6762402 0.8491803  0.9885595
## [6] {failures=cero}            => {Pstatus=living together} 0.7467363 0.8693009  1.0119825
## [7] {Pstatus=living together} => {failures=cero}              0.7467363 0.8693009  1.0119825
##     count
## [1] 329
## [2] 329
## [3] 264
## [4] 264
## [5] 259
## [6] 286
## [7] 286
```

- Si bebe muy poco alcohol entre semana se aprueba todo y viceversa
- Si bebe muy poco alcohol entre semana sus padres están juntos
- Si apruebas todo tus padres están juntos y viceversa

Observar reglas específicas

Por último queremos ver que soporte y confianza tienen las reglas de asociación que indican que los estudiantes beben mucho alcohol los fines de semana.

Usaremos el conjunto de estudiantes sin diferenciar sexo.

```
all_rules_students <- apriori(trans_students, parameter=list(support=0.01))
rules_alcohol_subset <- subset(all_rules_students, items %in% "Walc=very high")
```

```
inspect(head(sort(rules_alcohol_subset,by="support"),n=10))
```

```
##      lhs                          rhs                           support confidence      lift count
## [1]  {Walc=very high}          => {Pstatus=living together} 0.06471495  0.9333333 1.0645577    42
## [2]  {Walc=very high}          => {failures=cero}           0.05546995  0.8000000 0.9457195    36
## [3]  {Walc=very high,
##       failures=cero}           => {Pstatus=living together} 0.05084746  0.9166667 1.0455477    33
## [4]  {Walc=very high,
##       notRomantic}             => {Pstatus=living together} 0.03852080  0.9259259 1.0561088    25
## [5]  {Walc=very high,
##       notRomantic}             => {failures=cero}           0.03389831  0.8148148 0.9632328    22
## [6]  {Walc=very high,
##       failures=cero,
##       notRomantic}             => {Pstatus=living together} 0.03081664  0.9090909 1.0369069    20
## [7]  {Walc=very high,
##       Pstatus=living together,
##       notRomantic}             => {failures=cero}           0.03081664  0.8000000 0.9457195    20
## [8]  {Walc=very high,
##       romantic}                => {Pstatus=living together} 0.02619414  0.9444444 1.0772310    17
## [9]  {Walc=very high,
##       health=very good}        => {Pstatus=living together} 0.02619414  0.8947368 1.0205346    17
## [10] {Walc=very high,
##       health=very good}        => {failures=cero}           0.02465331  0.8421053 0.9954942    16
```

Podemos cuales son las reglas con mayor soporte para que se cumpla este hecho, sin embargo el soporte es demasiado bajo para que lo contemplemos.

Aún así, hemos visto otra forma interesante de observar reglas, atacando la regla específica que nos interesa.

1.2 Anexo

Mini-Chunk

Si queremos escribir código que esté incluido como texto y no como bloque usamos:

```
\Sexpr {'Codigo a incluir'}
Por lo tanto si queremos calcular la media del vector [2,3]:
La media del vector es Sexpr{mean(c(2,3))}. #\Sexpr
```

Ejecución:

La media del vector es 2.5.

Knitr

Knitr es un paquete de R [5] que añade opciones de presentación a Sweave. Nos permite entre otras cosas:

- Colorear el código dentro de los chunks
- Cambiar el tamaño de la fuente dentro del chunk
- Cambiar los margenes en los chunks para crear saltos de línea
- No mostrar warnings de R
- Manejar el tamaño de las figuras

Toda la información de las nuevas opciones se encuentran en : `https://yihui.name/knitr/options/`

References

[1] Package 'arules' *https://cran.r-project.org/web/packages/arules/arules.pdf*, Michael Hahsler

[2] Student Alcohol Consumption, *https://www.kaggle.com/uciml/student-alcohol-consumption*, Kaggle.

[3] Package 'arulesViz' *https://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf*, Michael Hahsler

[4] arules: Association Rule Mining with R – A Tutorial. *http://michael.hahsler.net/research/arules_RUG_-2015/demo/*, Michael Hahsler

[5] Package 'knitr' *https://cran.r-project.org/web/packages/knitr/knitr.pdf*, Yihui Xie