Trabajo práctico 05-03: Reglas de asociación

Ejercicio 1

a)

Soporte y confianza de todos los ítemsets del dataset:

A	B	C	Ítemset	Soporte	Reglas con minsup = 0.3	Confianza	Reglas con minconf = 0.7
0	1	0	{B}	0.5	-	1	-
1	0	1	{AC}	0.4	{A} => {C}	0.571428571	-
0	0	1	{C}	0.6	-	1	-
1	0	0	{A}	0.7	-	1	-
1	1	1	{ABC}	0.2	-	0.666666667	-
0	1	1	{BC}	0.3	{B} => {C}	0.6	-
1	1	0	{AB}	0.3	{A} => {B}	0.428571429	-
1	0	1	{AC}	0.4	{A} => {C}	0.571428571	-
1	0	0	{A}	0.7	-	1	-
1	1	1	{ABC}	0.2	-	0.666666667	-

b)

Reglas con minsup = 0.3:

1. {A} => {C}
2. {B} => {C}
3. {A} => {B}

c)

s({A}) = $\frac{7}{10}$ = 0.7

s({AB}) = 0.3
s({AC}) = 0.4
s({ABC}) = 0.2

Cualquier super-set del 1-ítemset {A} tendrá siempre *menor soporte* que ese 1-ítemset {A}.

Esto se debe a que cuántos más ítems se agreguen al ítemset, la probabilidad de encontrarlos juntos disminuye. En otras palabras, el numerador se achica y el denominador se mantiene constante.

d)

Si se determina que minconf = 0.7 no se encuentra ninguna regla de asociación.

Ejercicio 2

a)

Los parámetros que se pueden modificar previa ejecución de Apriori son especificados a través de la variable `parameter` en forma de lista: `apriori(data, **parameter**, appearance, control).`

Dicha lista permite determinar el nivel de soporte mínimo (minconf) y el nivel de confianza mínimo (minconf) para generar las reglas de asociación. Por defecto, dichos valores son 0.1 y 0.8 respectivamente.

La modificación de los mismos otorga la capacidad de determinar qué tan duras (o blandas) son las reglas de asociación resultantes.

b)

Sí, es posible. Sólo es necesario cargarlo en memoria. No hay que realizar ningún tipo de pre-procesamiento.

c)

Para determinar qué reglas son las más fuertes se utiliza el cálculo *lift*, que permite determinar la relación que hay entre la confianza de cada regla y la proporción de su consecuente.

En otras palabras, *lift* determina cuán propensos (en relación a las demás reglas) son los clientes a comprar el consecuente dado un determinado antecedente.

Proporción:
0 < Lift < 1 => el ítemset antecedente es lift veces perjudicial para la compra del consecuente.
Lift > 1 => el ítemset antecedente es lift veces beneficioso para la compra del consecuente.
Lift = 1 => el ítemset antecedente nada tiene que ver con el consecuente (son independientes).

Se ejecutó el algoritmo Apriori con un soporte = 0.01 y una confianza = 0.3. Las mejores 10 reglas, clasificadas por lift, son las siguientes:

#	Ítemset antecedente		Consecuente	Soporte	Confianza	Lift	Cuenta
1	{citrus fruit,other vegetables}	=>	{root vegetables}	0.01037112	0.3591549	3.295045	102
2	{tropical fruit,other vegetables}	=>	{root vegetables}	0.01230300	0.3427762	3.144780	121
3	{beef}	=>	{root vegetables}	0.01738688	0.3313953	3.040367	171
4	{citrus fruit,root vegetables}	=>	{other vegetables}	0.01037112	0.5862069	3.029608	102
5	{tropical fruit,root vegetables}	=>	{other vegetables}	0.01230300	0.5845411	3.020999	121
6	{other vegetables,whole milk}	=>	{root vegetables}	0.02318251	0.3097826	2.842082	228
7	{whole milk,curd}	=>	{yogurt}	0.01006609	0.3852140	2.761356	99
8	{root vegetables,rolls/buns}	=>	{other vegetables}	0.01220132	0.5020921	2.594890	120
9	{root vegetables,yogurt}	=>	{other vegetables}	0.01291307	0.5000000	2.584078	127
10	{tropical fruit,whole milk}	=>	{yogurt}	0.01514997	0.3581731	2.567516	149

d)

Si ejecutamos el método `summary(Groceries)` nos indica que el dataset posee 9835 transacciones (o ítemsets) y 169 columnas con una densidad de 0.02609146.

Si sabemos que Complejidad = O(N*M*w), entonces:

- N = 9835.
- M = ítemsets candidatos por cada N que haya.
- w cantidad de ítems en un N-ítemset.

Por lo tanto, su complejidad es 2^d siendo d la cantidad de ítemsets candidatos. Es decir, su complejidad depende de M.

Lo que hace Apriori para intentar reducir la complejidad computacional es podar el número de ítemsets candidatos (M). Utiliza el principio antimonotonía del soporte que sostiene que: *Si un itemset es frecuente, todos sus subsets también deben serlo.*

O dicho de otra manera, el soporte de un superset (Y) nunca excede el soporte de sus subsets (X):

$$\forall X,Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y).$$

Así, cada vez que Apriori encuentre un ítemset con un soporte < *minsup* lo descarta, pues su consecuente nunca podrá superar el soporte del suyo.

e)

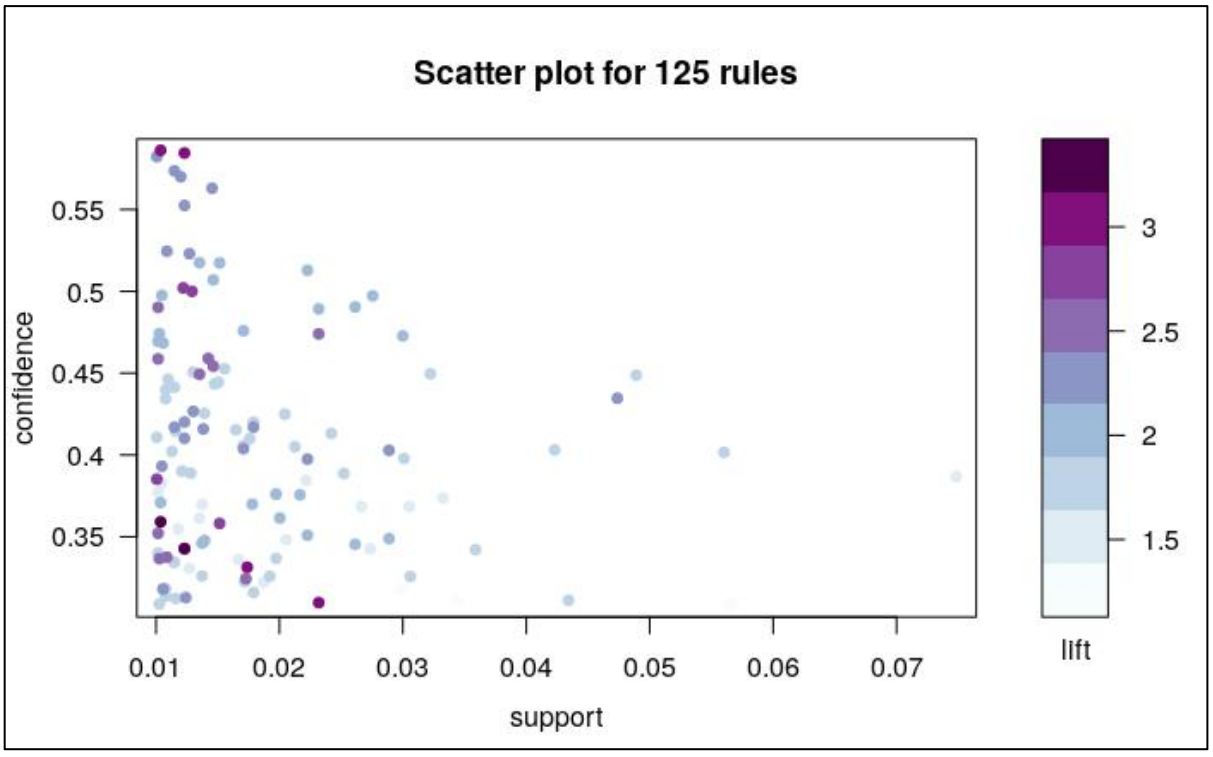
Asociaciones interesantes del top 10:

- Los clientes que compran frutas cítricas y vegetales que no sean de raíz, tienen 3.29 veces más probabilidad de comprar también vegetales de raíz sobre cualquier otro ítem del supermercado.
- Cualquier cliente que compre bife es 3.04 veces más propenso a comprar también vegetales de raíz antes que otro alimento.
- Un cliente que compre vegetales de raíz (zanahora, papa, batata) y yogurt, es 2.58 veces más propenso a comprar también otros vegetales que no sean de raíz.
- Los clientes que compran frutas tropicales y leche entera, son 2.56 veces más propensos a comprar también yogurt.

La política a implementar es sencillamente encontrar una distribución que fomente la ocurrencia de estas reglas encontradas.

Por ejemplo: si sé que cada vez que los clientes que me compran leche entera y ricota también suelen elegir yogurt con una propensidad de 2.7 veces por sobre cualquier otro lácteo o comestible, trato de ubicar las góndolas o estantes que poseen dichos alimentos lo más cerca posible para ayudar a que dicha asociación ocurra.

f)



Observaciones:

- En este scatterplot se pueden distinguir tres dimensiones diferentes. Sobre el eje X se grafica el soporte de cada regla generada mientras que sobre el eje Y se grafica la confianza de cada regla. Por último, la intensidad del color de cada punto determina su valor de lift.

- Sólo hay 5 reglas (y corroborar con la tabla del punto c.) con un *lift* >= 3 y por ende con una tonalidad más intensa que cualquier otra.

- Lo interesante de este grupo de 5 reglas es que se puede apreciar como el valor de confianza de cada regla afecta de manera directa su calificación *lift*.

- Debido a que este dataset se caracteriza por tener un valor de soporte bastante bajo para sus ítemsets, el valor de *lift* se ve directamente afectado por la confianza de los mismos. Es decir, como se espera que el valor de soporte (denomidador en *lift*) sea medianamente "constante" (pues 0.01 < soporte < 0.02) a mayor nivel de confianza (numerador en *lift*) mayor coeficiente *lift* tendrá la regla. No hay una gran variación en el nivel de soporte de cada regla encontrada.

- Adicionalmente, que el dataset posea pocas reglas con tonalidad muy clara es un buen indicio. Significa que los ítems de las reglas encontradas son en su mayoría dependientes (pues *lift* > 1) y es conocimiento que se puede explotar adoptando una política acorde a cada regla.

g)

Al ejecutar el algoritmo con el parámetro `rhs` = 'bottled beer' se encontraron 5 reglas de asociación.

#	Ítemset antecedente		Consecuente	Lift	Cuenta
1	{bottled water}	=>	{bottled beer}	1.7707259	155
2	{soda}	=>	{bottled beer}	1.2092094	167
3	{other vegetables}	=>	{bottled beer}	1.0375464	159
4	{whole milk}	=>	{bottled beer}	0.9932367	201
5	{rolls/buns}	=>	{bottled beer}	0.9198466	134

Observaciones:

- Solo {bottled water} y {soda} marcan la presencia de cerveza ya que poseen un *lift* > 1.

- {other vegetables} y {whole milk} tienen un valor *lift* de 1.037 y 0.993 respectivamente, indicando que ambas variables son independientes.

- El ítem {whole milk} , con un valor *lift* < 1, marca justamente lo opuesto. Es decir, dada su ocurrencia en un carrito de compras es poco probable que un cliente compre también una cerveza.

- En resumen, solo las reglas #1 y #2 son las que marcan la presencia de cerveza y tiene mucho sentido ya que quién probablemente compre alguna cerveza también desee hidratarse con algo que no sea alcohol como una bottela de agua o de soda.

h)

Como se mencionó en el punto a), los parámetros a modificar para ajustar la cantidad de reglas resultantes son el soporte y la confianza. En este caso, tal como dice el punto c), el algoritmo se ejecutó con un minsup = 0.01 y una minconf = 0.3 lo que se traduce en un set de reglas reducido por su elevado nivel de confianza.

A fines de tener un set de reglas de asociación más amplio, lo que se puede hacer es bajar ambos mínimos permitiendo que “ingresen” otros ítemsets y así contar con mayor variedad. Sin embargo, será importante revisar en dichos casos, más que ahora, el valor de *lift* asociado.

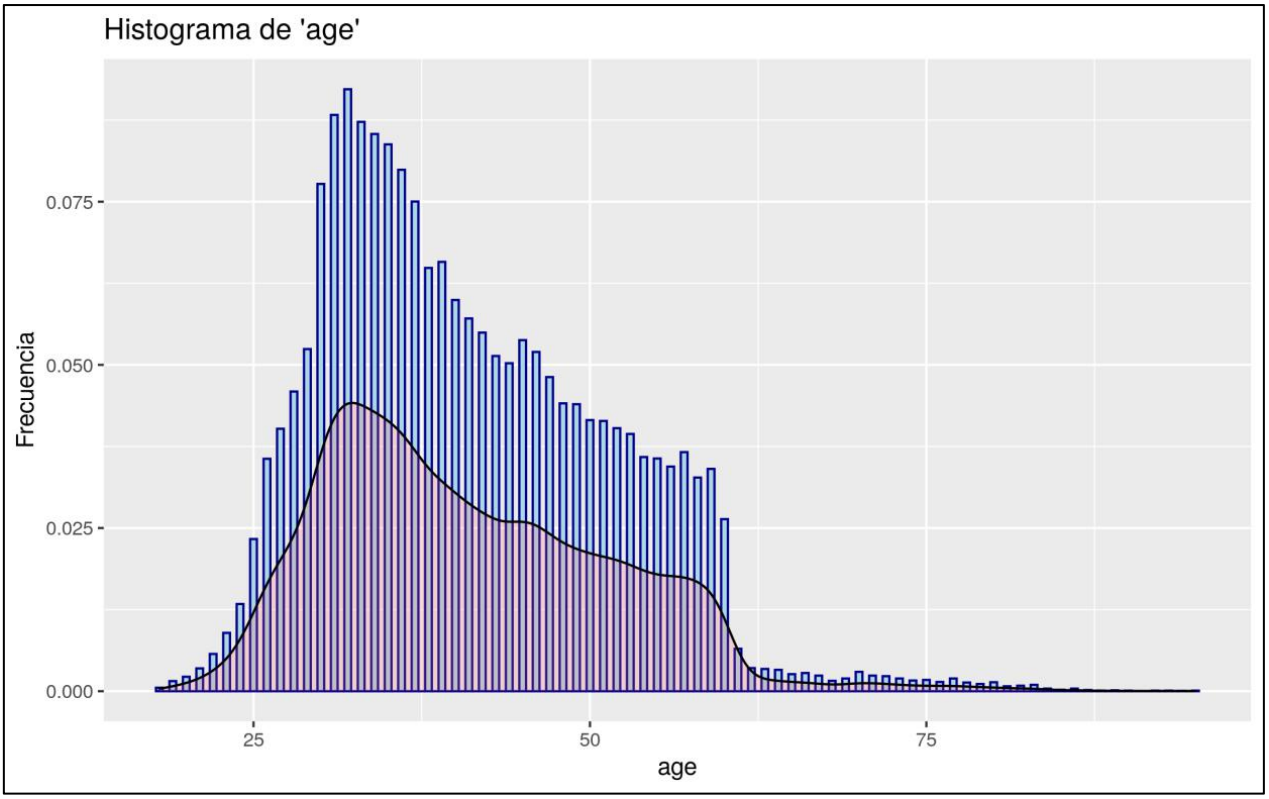
Ejercicio 3

a)

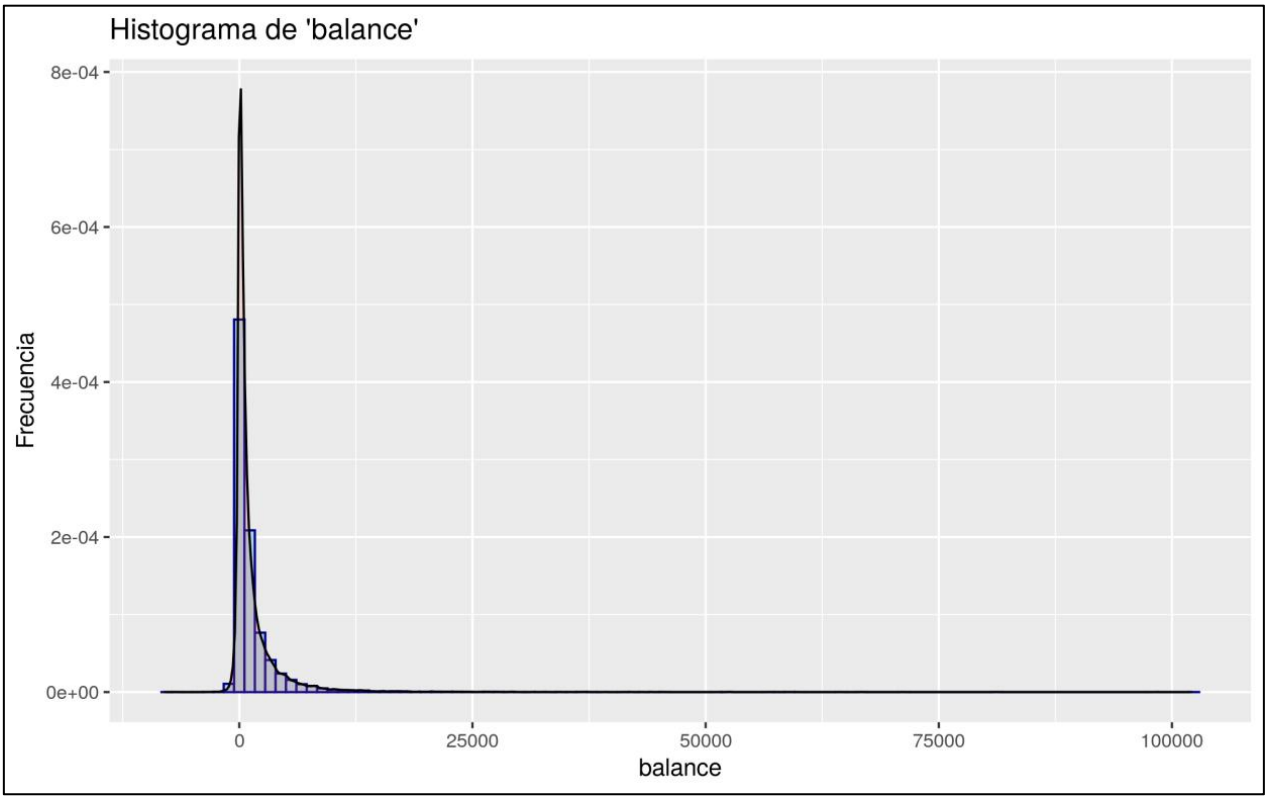
Para poder ejecutar el algoritmo Apriori, el dataset debe encontrarse completamente discretizado y factorizado.

Se convertirán las columnas:

1. *age* (edad), *balance* (promedio anual de dinero en cuenta).
2. *duration* (duración en segundos del último contacto banco-cliente).
3. *day*: último día de contacto.
4. *campaign:* cantidad de contactos durante la campaña de marketing.
5. *pdays:* cantidad de dias que pasaron desde el último contacto perteneciente a una campaña previa.
6. *previous:* cantidad de contactos con el cliente pertenecientes a campañas previas.



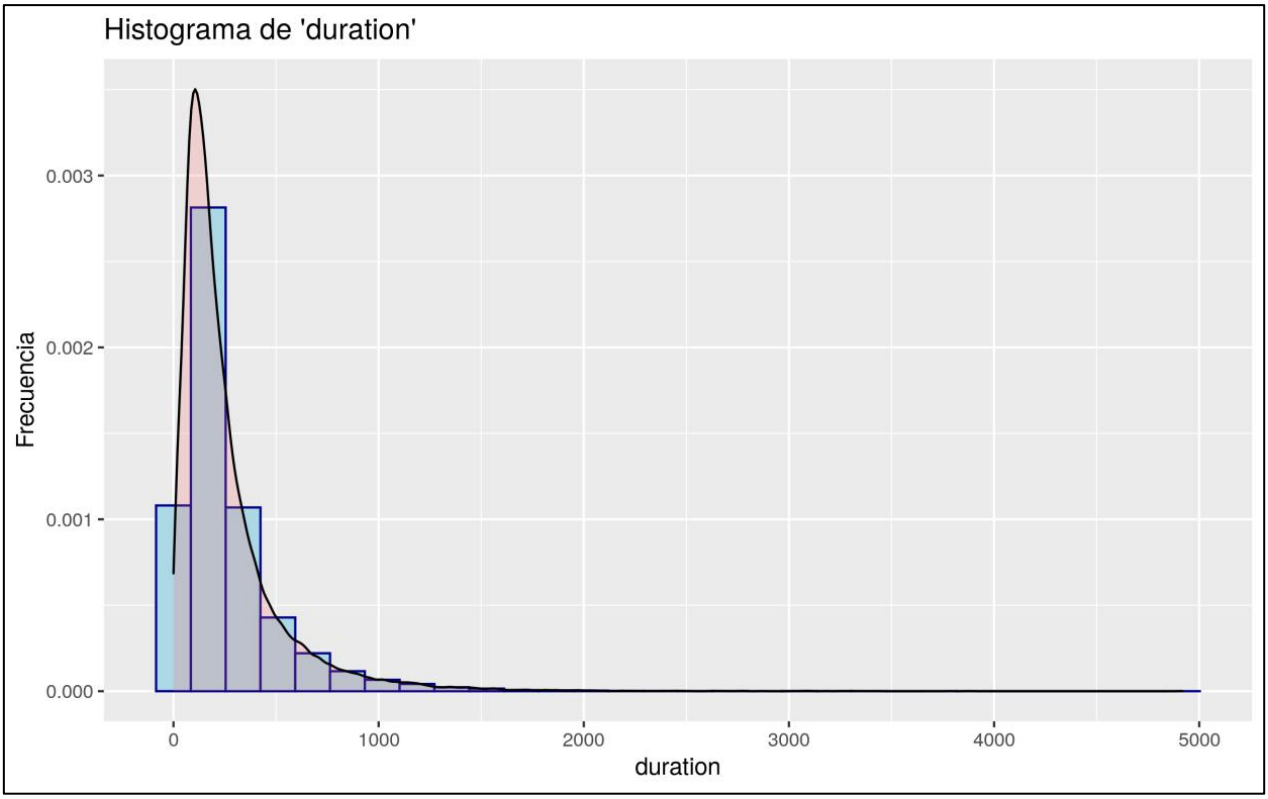
Equivalencias			
19-25	26-39	40-59	60+
Joven	Adulto	Adulto-mayor	Vejez



Observación:

- La presencia de algunos outliers en el extremo derecho del histograma representan personas con muchísimo dinero en su cuenta bancaria.

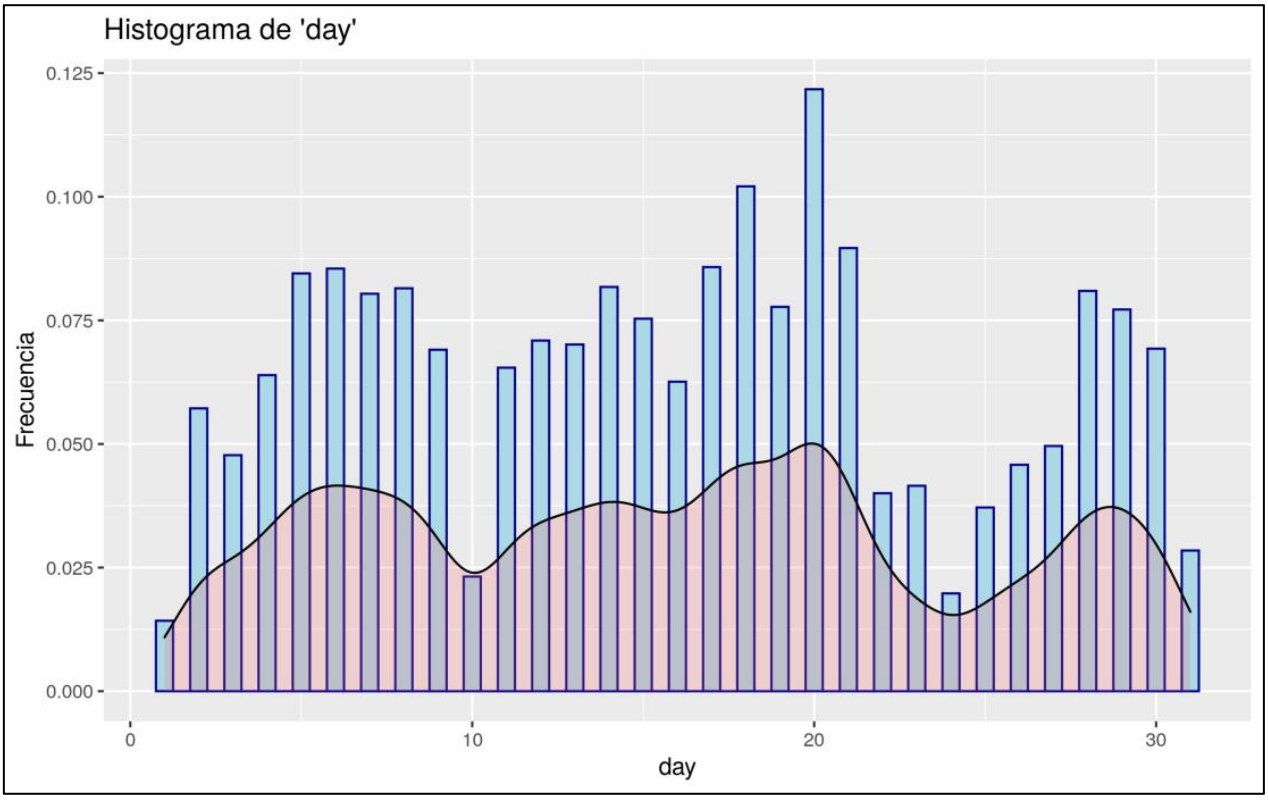
Equivalencias					
-0	1-99	100-249	250-1199	1200-2499	2500+
Negativo	Muy bajo	Bajo	Regular	Alto	Muy alto



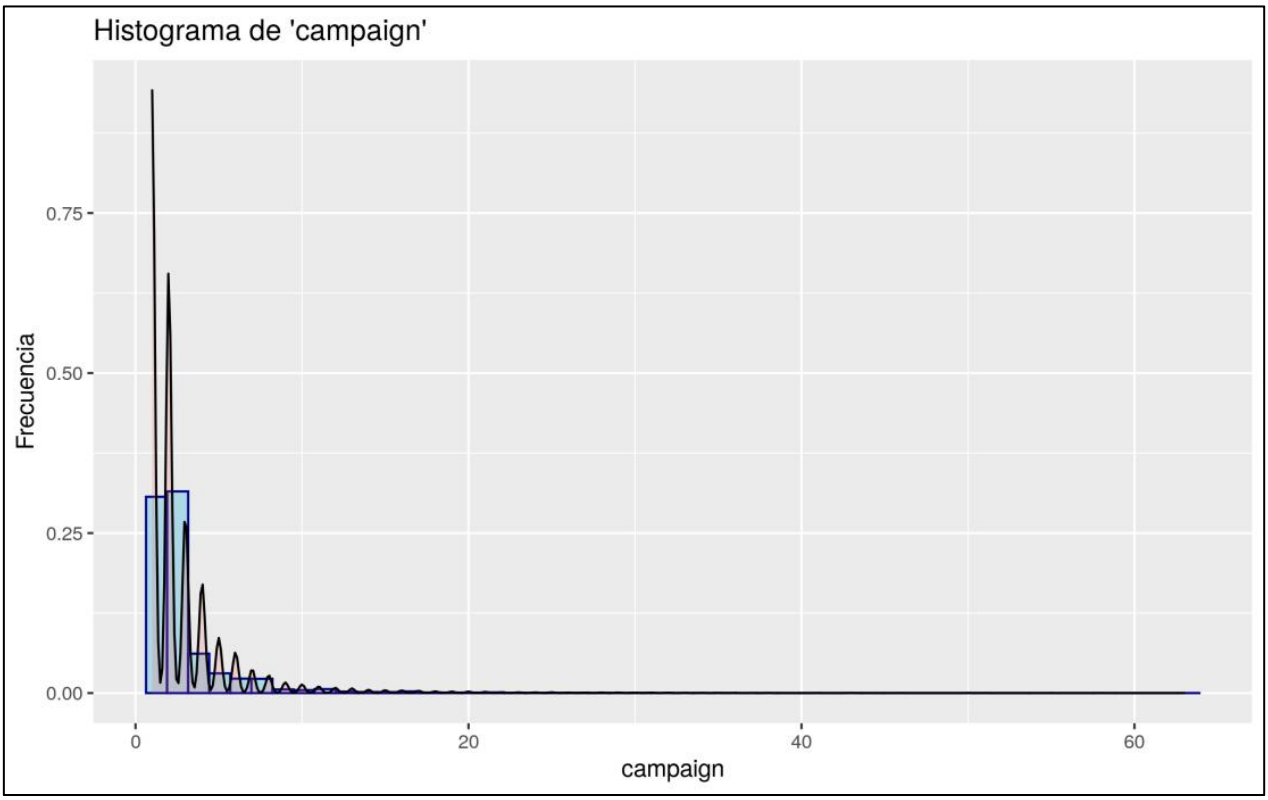
Observación:

- Con el histograma de duration se puede observar que también existen outliers. Esto indica la existencia de contactos de larga duración (5000s = 83minutos) que pueden representar algún teléfono mal colgado o la ocurrencia de una operación bancaria complicada.

Equivalencias									
261	151	76	92	198	139	217	380	50	55
larga	regular	corta	corta	regular	regular	regular	larga	corta	corta



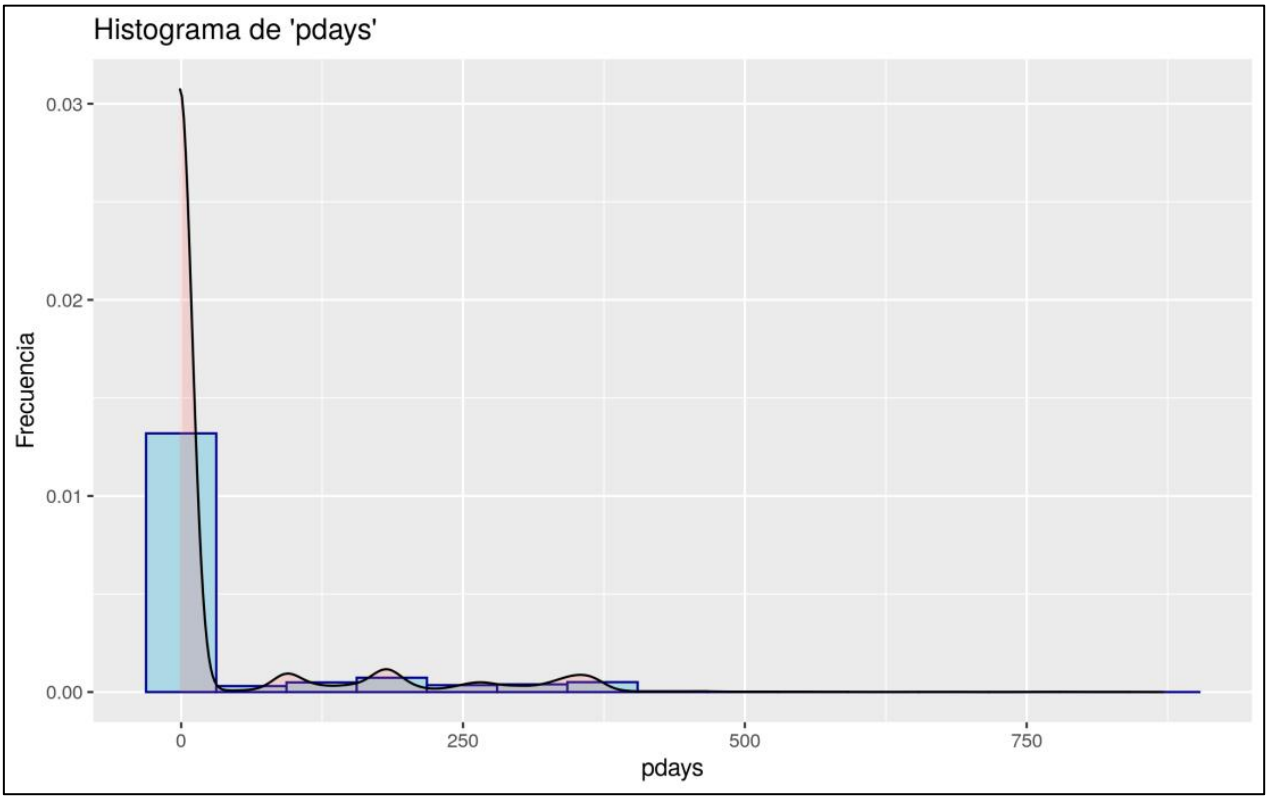
Equivalencias			
1-7	8-14	15-21	22+
Semana_1	Semana_2	Semana_3	Semana_4



Observación:

- Nuevamente se detecta la presencia de outliers. En este caso, algún cliente con el que se contactaron más de 60 veces durante la campaña de marketing.

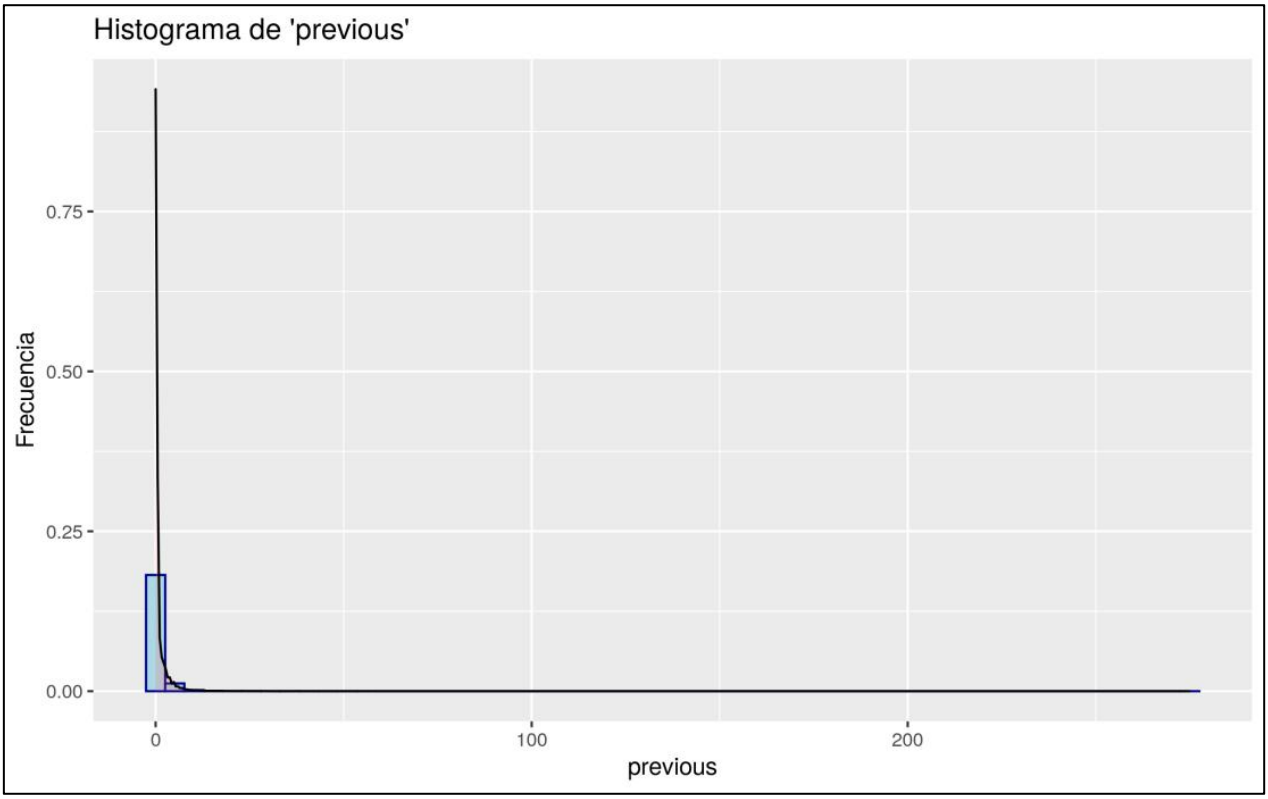
Equivalencias		
1	2-24	25+
Unica	Varias	Muchas



Observación:

- Outliers nuevamente. En esta oportunidad, se observan en el extremo derecho del eje X la cantidad de días que pasaron luego de que los clientes hayan sido contactados por última vez por una campaña previa.
- Respecto a los valores de X = -1, representan clientes que nunca habían sido contactados previamente.

Equivalencias			
-1	0-59	60-119	120+
Nunca	Rapidamente	Lentamente	Extremandamente_lento



Observación:

- Los outliers en esta columna representan a los clientes que más han sido contactados en otras campañas de marketing previas a esta.
- Al igual que en *pdays*, los X = -1 representan el valor nunca.

Equivalencias			
-1-0	1-19	20-49	50+
Nunca	Pocas	Varias	Muchas

Finalmente, se recreó el *dataframe* con estas nuevas columnas discretizadas para luego convertirlo al formato *transactions* y así poder comenzar a trabajar sobre ellas.

b)

Obtenidas las *transactions*, primero se grafica el top 10 de ítems con mayor ocurrencias:

Posteriormente, se ejecutó el algoritmo Apriori con un soporte = 0.1 y una confianza = 0.3. Esta es una selección del top 100 reglas con mayor *lift* (consultar archivo 'ejercicio_3.Rmd' para ver más):

#	Antecedentes		Consecuente	Lift	Cuenta
1	{contact=cellular, previous=pocas, y=no}	=>	{pdays=extremadamente_lento}	5.816212	4722
5	{default=no poutcome=failure}	=>	{previous=pocas}	5.488068	4836
9	{default=no, previous=pocas}	=>	{poutcome=failure}	5.481187	4836
45	{job=management, default=no, loan=no, contact=cellular}	=>	{education=tertiary}	2.915589	5150
56	{education=tertiary, campaign=varias}	=>	{job=management}	2.881837	4955
100	{education=tertiary, default=no, previous=nunca, y=no}	=>	{job=management}	2.827403	5473

Observaciones:

- Al cliente que se lo haya contactado pocas veces, mediante un celular y que haya rechazado el préstamo es un cliente al que se ha demorado demasiado tiempo en volver a contactarlo desde la última vez que lo llamaron.

- Un cliente que no haya defaulteado su deuda y al cual el banco lo haya contactado pocas veces durante esta campaña, es un cliente donde las campañas anteriores del banco han fracasado.

- La regla #9 es una variante de la regla #5 pero el consecuente ahora es *poutcome* en vez de *previous*.

- Un cliente que trabaje en management, no haya defaulteado su deuda, no haya pedido un préstamo personal y su último contacto fué por medio de un celular no aceptará el préstamo.

Conclusión:

1. No siempre las reglas que más aplican al dominio surgen fácilmente.

2. Es importante recordar que este algoritmo calcula las asociaciones más frecuentes y tal vez lo que nos interese no se encuentre en el primer intento o que lo más recurrente sea lo que mayor información nos dé.

3. El siguiente punto es una confirmación de estas conclusiones, ya que hace una búsqueda de reglas que permitan dar con un 1-ítemset consecuente específico.

c)

Para determinar las asociaciones más relevantes que definen si el cliente aceptó o no el préstamo se ejecutaron los siguientes comandos:

```
# Asociaciones por prestamo rechazado
reglas_no <- apriori(tx, parameter = list(support=0.1, confidence=0.3, target = "rules"),
                         appearance = list(rhs='y=no'))
```

Selección del top 10 reglas con mayor *lift* (consultar archivo 'ejercicio_3.Rmd' para ver más):

#	Antecedentes		Consecuente	Lift	Cuenta
1	{marital=married, duration=corta, campaign=varias, previous=nunca}	=>	{y=no}	1.127709	5196
2	{marital=married, duration=corta, campaign=varias, pdays=nunca}	=>	{y=no}	1.127709	5196
3	{marital=married, duration=corta, campaign=varias, poutcome=unknown}	=>	{y=no}	1.127709	5196
9	{marital=married, default=no, duration=corta, campaign=varias, pdays=nunca}	=>	{y=no}	1.127606	5086

Conclusiones:

1. Se detecta 3-itemset formado por {marital=married, duration=corta, campaign=varias} que son claves a la hora de determinar si un cliente rechazará el préstamo.

2. Este ítemset indica que el cliente rechazará el préstamo si:
   a) Está casado, y
   b) La duración del último contacto fué corta, y
   c) Se lo ha contactado varias veces durante esta campaña de marketing.

```
# Asociaciones por prestamo aceptado
reglas_yes <- apriori(tx, parameter = list(support=0.01, confidence=0.3, target = "rules"),
                         appearance = list(rhs='y=yes'))
```

Selección del top 10 reglas con mayor *lift* (consultar archivo 'ejercicio_3.Rmd' para ver más):

#	Antecedentes		Consecuente	Lift	Cuenta
1	{loan=no, duration=larga, poutcome=success}	=>	{y=yes}	6.722424	556
2	{default=no, loan=no, duration=larga, poutcome=success}	=>	{y=yes}	6.722424	556
3	{loan=no, duration=larga, previous=pocas, poutcome=success}	=>	{y=yes}	6.719838	555

5	{loan=no, contact=cellular, duration=larga, poutcome=success}	=>	{y=yes}	6.709669	500
7	{loan=no, contact=cellular, duration=larga, previous=pocas, poutcome=success}	=>	{y=yes}	6.706779	499

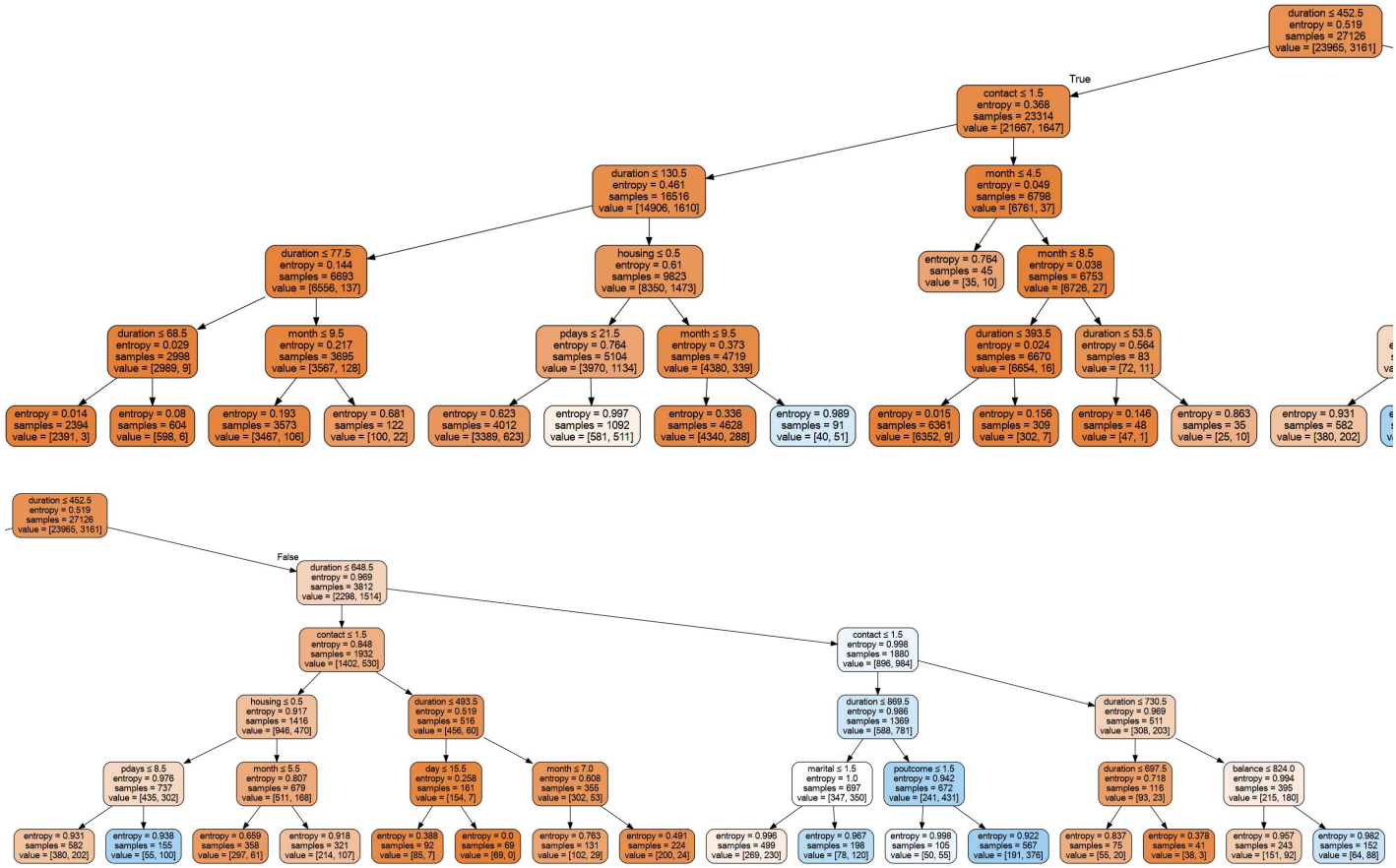
Conclusiones:

1. Al igual que en el caso del *no*, la presencia de un 3-ítemset: {loan=no, duration=larga, poutcome=success} constituye el núcleo de las reglas. En las demás, se lo observa combinado con otros ítems.

2. Esto significa que hay tres factores que juntos determinan la aceptación del préstamo:
   a) Que el cliente no haya solicitado un préstamo personal, y
   b) Que la duración del último contacto haya sido larga (~ 20min.), y
   c) Que la campaña de marketing anterior haya sido exitosa con el cliente.

d)

Se adjunta el árbol utilizado en el TP05-01 con una profundida de 5 niveles (se deja en el .zip la imagen completa para apreciarla mejor).



Conclusiones:

1. Apriori que otorga un 3-ítemset diferente para cada situación (reglas aceptación y reglas rechazo).

2. En cambio, se observa que el árbol saca un factor común de las features que más dividen al dataset (*duration* y *contact*) y las utiliza para generalizar. Esto también se puede ver afectado por la parametrización del árbol con 5 niveles máximos para evitar el overfitting.

3. En resumen, Apriori permite obtener una especificación acerca de los ítems más determinantes para un caso especial mientras que los árboles se abstraen de lo específico para poder generalizar mejor.