



Trabajo Practico Entregable 2

Data Mining

**Maestría en Explotación de Datos y Descubrimiento del
Conocimiento**

**Facultad de Ciencias Exactas y Naturales
Departamento de Computación
Universidad de Buenos Aires**

**Martín Llada
Juan Jose Iguarán Fernández
9 de Julio de 2019**

Tabla de Contenido

<u>Introducción</u>	<u>2</u>
<u>Preprocesamiento</u>	<u>3</u>
<u>Análisis</u>	<u>5</u>
<u>Reglas de carácter descriptivo</u>	<u>6</u>
<u>Descripciones generales</u>	<u>6</u>
<u>Desaceleración del ultimo periodo</u>	<u>10</u>
<u>Comportamiento de los productos</u>	<u>11</u>
<u>Reglas de carácter predictivo</u>	<u>14</u>
<u>Visualización</u>	<u>17</u>
<u>Relación con análisis previos</u>	<u>19</u>
<u>Conclusiones</u>	<u>20</u>
<u>Bibliografía</u>	<u>21</u>

Introducción

Mediante el uso de reglas de asociación, es posible evidenciar cuán frecuente es la ocurrencia simultanea de los valores de distintas variables para un set de observaciones dado, de manera que pueda establecerse cierta correlación en la ocurrencia de dichas variables. Para evaluar cuan relevante es el patrón encontrado se disponen de las métricas soporte, confianza y lift. El soporte calcula el porcentaje del total de las observaciones en los que se evidencia la ocurrencia conjunta de los valores de las variables. La confianza es el porcentaje del total de ocurrencias de los valores de una variable, en los que se observó la ocurrencia del valor de otra. Finalmente, el lift calcula la independencia de las dos ocurrencias, es decir, la probabilidad de la intersección dividida entre la multiplicación de la probabilidad de ocurrencia de cada variable por separado. Un ejemplo el siguiente:

computer \Rightarrow *antivirus_software* [*support* = 2%, *confidence* = 60%]

Figure 1: Ejemplo de regla de asociación

En el anterior ejemplo, se observa que un 2% del total de las transacciones incluyeron la compra de un computador y el 60% por ciento de las veces que se compró un computador también se compró un antivirus.

Preprocesamiento

Se disponen de datos que contienen información sobre productos precios y sucursales de supermercados en CABA para 10 mediciones diferentes. Inicialmente se discretiza la variable fecha como se muestra a continuación:

Periodo	Mes predominante	Mediciones	Fecha inicio	Fecha fin
1	Noviembre 2018	1-2-3	2018-11-05	2018-12-08
2	Diciembre 2018	4-5	2018-12-10	2018-12-26
3	Enero 2019	6-7	2018-12-31	2019-01-30
4	Febrero 2019	8-9-10	2019-02-04	2019-03-02

Figure 2: Discretización aplicada a las fechas

Luego el precio particular con relación a la media del mismo producto:

Categorías	Rango
Muy caro	$[\infty; 0.1)$
Medio caro	$[0.1; 0.05)$
Levemente caro	$[0.05; 0.01)$
Medio	$[-0.01; 0.01)$
Levemente barato	$[-0.05; -0.01)$
Medianamente barato	$[-0.1; -0.05)$
Muy barato	$[-\infty; 0.1)$

Finalmente, su variación con respecto al periodo anterior.:

Categorías	Rango
Disminución Fuerte	$[-\infty; -0.05)$
Disminución Media	$[-0.5; -0.02)$
Disminución Leve	$[-0.02; -0.005)$
Mantiene	$[-0.005; 0.005)$
Aumento Leve	$[0.005; 0.05)$
Aumento Medio	$[0.05; 0.1)$
Aumento Fuerte	$[0.1; \infty]$

Figure 4: Discretización aplicada a las variaciones

Los nombres de los productos están originalmente en el siguiente formato: “Ron Blanco Capitán Morgan 750 MI”. Se selecciona únicamente la primera palabra, con previa remoción de tildes y mayúsculas, pues esta contiene el nombre del producto en cuestión, para luego construir una nueva columna con dicha palabra. Se realizó la siguiente grafica en donde el eje de las abscisas representa la cantidad de veces que aparece la palabra y el eje de las ordenadas representa cuantas palabras existen para determinada frecuencia:

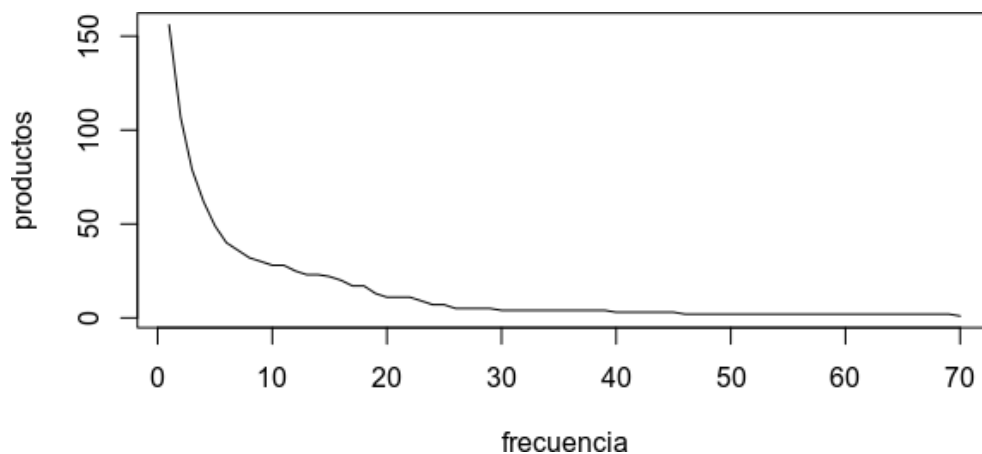


Figure 5: Frecuencia de aparición de los productos en la lista

Como puede observarse, la frecuencia de aparición de termino tiene un decrecimiento exponencial, estabilizándose la pendiente entre las 20 y 30 ocurrencias, por lo que se procederá a seleccionar los 28 términos con mayor ocurrencia para el presente análisis.

Se determina a partir de las coordenadas los barrios de las sucursales y finalmente, para intentar concentrar la información y obtener características más generales, se dividieron los 44 barrios en 4 zonas geográficas: Noreste, noroeste, sureste y suroeste.

Análisis

Para la elaboración de reglas de asociación, se utilizará la librería arules del software R,

específicamente la función apriori cuya utilidad radica en ejecutar el algoritmo homónimo para el descubrimiento de reglas de asociación. A partir de este se pretenden descubrir reglas que arrojen información tanto en lo que se refiere a la descripción de los datos, así como reglas de carácter predictivo. Finalmente, utilizando la librería arulesVIZ, se incluirán algunas representaciones graficas.

Reglas de carácter descriptivo

Para abordar el análisis descriptivo, se dividirá el tema en cuatro tres principales: Descripciones generales, reglas asociadas a la desaceleración del incremento de precios del último periodo y finalmente reglas ligadas al comportamiento de los productos.

Descripciones generales

Al ejecutar el algoritmo, se evidencia que existen una gran cantidad de reglas que no aportan información, como es el caso de las relacionadas con el tipo de sucursal, ya que al ser la gran mayoría supermercados, este valor estará profundamente relacionado con todo lo demás lo cual es bastante evidente. También es el caso de las variaciones y precios para cada periodo pues se obtienen reglas tales como al presentar en el periodo 1, 2, 3 y 4 precios caros, el precio promedio es también caro. Así mismo, las características generales de los precios van más allá de su relación con los productos, por lo que las columnas relacionadas con los términos previamente elaboradas no aportan mayor información. Por estas razones, las columnas tenidas en cuenta para la extracción de reglas han sido variación total discreta, el precio relativo medio discreto, razón social y zona.

A partir de las reglas con mayor lift se pueden apreciar algunas relaciones interesantes, a saber:

- Aquellos bienes calificados como muy baratos, pero con incrementos fuertes, se han encontrado con más frecuencia en Carrefour. Aquellos bienes calificados como muy caros y que así mismo presentan aumentos fuertes En Jumbo con más frecuencia (ver figura 6):

Reglas		
{variacion.total.discret=Aumento fuerte,relativo.medio.discret=Muy barato} => {comercioRazonSocial=INC S.A.}		
{variacion.total.discret=Aumento fuerte,relativo.medio.discret=Muy caro} => {comercioRazonSocial=Jumbo Retail Argentina S.A.}		
{comercioRazonSocial=Jumbo Retail Argentina S.A.,variacion.total.discret=Aumento fuerte} => {relativo.medio.discret=levemente barato}		
{variacion.total.discret=Aumento fuerte,relativo.medio.discret=levemente barato} => {comercioRazonSocial=Jumbo Retail Argentina S.A.}		
Support	confidence	Lift
0.012	0.786	3.002
0.013	0.805	2.979
0.010	0.092	0.366
0.010	0.085	0.315

Figure 6: Reglas generales y sus respectivas métricas 1

- Los productos cuyos precios se encontraban en el rango medio y se mantuvieron relativamente estables podían encontrarse con más probabilidad en COTO (ver figura 7):

Reglas	
{variacion.total.discret=Mantiene,relativo.medio.discret=Medio} => {comercioRazonSocial=Coto Centro Integral de ComercializaciÃ³n S.A.}	
{variacion.total.discret=Mantiene,relativo.medio.discret=levemente barato} => {comercioRazonSocial=Coto Centro Integral de ComercializaciÃ³n S.A.}	

Support	confidence	Lift
0.036	0.737	1.851
0.019	0.575	1.444

Figure 7: Reglas generales y sus respectivas

- Cuando se analiza aquellos productos que han reflejado disminuciones en sus precios, las reglas permiten inferir que estos bienes han sido ofrecidos en la cadena Carrefour (ver figura 8):

Reglas	Support	confidence	Lift
{variacion.total.discret=Disminucion media} => {comercioRazonSocial=INC S.A.}	0.012	0.522	1.974
{comercioRazonSocial=INC S.A.} => {variacion.total.discret=Disminucion media}	0.012	0.045	1.974
{variacion.total.discret=Disminucion fuerte} => {comercioRazonSocial=INC S.A.}	0.017	0.615	2.323
{comercioRazonSocial=INC S.A.} => {variacion.total.discret=Disminucion fuerte}	0.017	0.065	2.323

Figure 8: Reglas generales y sus respectivas métricas

- En la figura 9 se aprecia que los barrios agrupados como zona Noreste, se encuentran precios muy caros. A su vez se encontró que los precios aumentaron fuertemente.

En estos barrios Carrefour (INC S.A.) registro leves reducciones de precios, mientras que Jumbo reporto grandes aumentos. COTO reporto estabilidad en sus precios para productos de precios medios. Es poco frecuente hallar sucursales de la cadena Día en barrios de la zona Noreste.

Se puede inferir que la zona Noreste se asocia a barrios con cadenas que han fijado precios muy por encima del promedio. En particular, dentro de este barrio, Recoleta y Palermo reflejan las mayores diferencias de precios entre sucursales:

Reglas
1 {comercioRazonSocial=Jumbo Retail Argentina S.A.,zona=Noreste} => {relativo.medio.discret=Muy caro}
2 {zona=Noreste,relativo.medio.discret=Muy barato} => {comercioRazonSocial=INC S.A.}
3 {zona=Noreste,variacion.total.discret=Aumento fuerte,relativo.medio.discret=Medio caro} => {comercioRazonSocial=Jumbo Retail Argentina S.A.}
4 {comercioRazonSocial=Coto Centro Integral de Comercializaci3n S.A.,zona=Noreste,variacion.total.discret=Mantiene} => {relativo.medio.discret=Medio}
5 {comercioRazonSocial=DIA Argentina S.A} => {zona=Noreste}
6 {zona=Noreste} => {comercioRazonSocial=DIA Argentina S.A}
7 {comercioRazonSocial=Jumbo Retail Argentina S.A.,zona=Noreste} => {relativo.medio.discret=Medio}
8 {zona=Noreste,relativo.medio.discret=Levemente caro} => {comercioRazonSocial=INC S.A.}
9 {barrio_zona=RECOLETA.Noreste,relativo.medio.discret=Medio caro} => {comercioRazonSocial=Jumbo Retail Argentina S.A.}
10 {barrio_zona=PALERMO.Noreste,relativo.medio.discret=Medio caro} => {comercioRazonSocial=Jumbo Retail Argentina S.A.}

	Support	confidence	Lift
1	0.019	0.140	3.444
2	0.014	0.782	2.985
3	0.013	0.748	2.769
4	0.012	0.433	1.672
5	0.020	0.308	0.761
6	0.020	0.049	0.761
7	0.020	0.154	0.595
8	0.014	0.144	0.549
9	0.012	0.883	3.270
10	0.010	0.717	2.655

Figure 9: Reglas generales y sus respectivas

- En la zona Suroeste se encuentran precios levemente baratos. Estos productos se han encontrado de forma frecuente en la empresa Día Argentina S.A. En esta cadena pueden encontrarse aumentos fuertes Ver figura 10.

En esta zona en Jumbo es más frecuente encontrar productos medio caros, mientras que en COTO es más frecuente hallar precios medios que se mantienen. Además, es poco frecuente hallar sucursales de Carrefour en esta zona.

Se puede apreciar, por ejemplo, que en Mataderos las sucursales más frecuentes corresponden a la cadena DÍA o bien, que los barrios Monte Castro, Flores y Villa Lugano son abastecidos por COTO:

Reglas	
1	{zona=Suroeste,relativo.medio.discret=levemente barato} => {comercioRazonSocial=DIA Argentina S.A}
2	{comercioRazonSocial=Jumbo Retail Argentina S.A.,zona=Suroeste} => {relativo.medio.discret=Medio caro}
3	{zona=Suroeste,relativo.medio.discret=Medio caro} => {comercioRazonSocial=Jumbo Retail Argentina S.A.}
4	{zona=Suroeste,variacion.total.discret=Aumento fuerte} => {comercioRazonSocial=DIA Argentina S.A}
5	{zona=Suroeste,variacion.total.discret=Mantiene,relativo.medio.discret=Medio} => {comercioRazonSocial=Coto Centro Integral de Comercialización S.A.}
6	{comercioRazonSocial=INC S.A.} => {zona=Suroeste}

	Support	confidence	Lift
1	0.012	0.183	2.814
2	0.012	0.231	2.496
3	0.012	0.577	2.138
4	0.015	0.132	2.023
5	0.011	0.798	2.005
6	0.046	0.175	0.697

Figure 10: Reglas generales y sus respectivas métricas 5

- En Sureste parecen asociarse mayoritariamente COTO, Día y Carrefour. A su vez, los productos hallados en COTO y Carrefour poseen precios estables. En estos barrios es frecuente hallar precios estables y poco frecuente que encontrar sucursales de Jumbo (ver figura 11).

El barrio Parque Patricios es abastecido por Día. En Mataderos las sucursales más frecuentes corresponden a la cadena DÍA; los barrios Monte Castro, Flores y Villa Lugano son abastecidos por COTO:

Reglas

- 1 {zona=Sureste} => {comercioRazonSocial=DIA Argentina S.A.}
- 2 {comercioRazonSocial=INC S.A.,zona=Sureste} => {relativo.medio.discret=levemente barato}
- 3 {zona=Sureste,relativo.medio.discret=Medio} => {comercioRazonSocial=Coto Centro Integral de Comercializaci3n S.A.}
- 4 {zona=Sureste,variacion.total.discret=Mantiene} => {comercioRazonSocial=Coto Centro Integral de Comercializaci3n S.A.}
- 5 {zona=Sureste,relativo.medio.discret=levemente barato} => {comercioRazonSocial=INC S.A.}
- 6 {relativo.medio.discret=Medianamente barato} => {zona=Sureste}
- 7 {zona=Sureste} => {relativo.medio.discret=Medianamente barato}
- 8 {relativo.medio.discret=Levemente caro} => {zona=Sureste}
- 9 {zona=Sureste} => {comercioRazonSocial=Jumbo Retail Argentina S.A.}
- 10 {comercioRazonSocial=Jumbo Retail Argentina S.A.} => {zona=Sureste}
- 11 {barrio_zona=PARQUE PATRICIOS.Sureste} => {comercioRazonSocial=DIA Argentina S.A.}

	Support	confidence	Lift
1	0.015	0.134	2.054
2	0.011	0.469	1.874
3	0.022	0.684	1.717
4	0.012	0.662	1.663
5	0.011	0.337	1.288
6	0.012	0.142	1.267
7	0.012	0.103	1.267
8	0.020	0.088	0.782
9	0.012	0.105	0.389
10	0.012	0.044	0.389
11	0.010	0.619	9.513

Figure 11: Reglas generales y sus respectivas métricas

- En la zona Noroeste e (ver figura 12) se encuentran con frecuencia Carrefour, Jumbo y COTO. Se encuentra con más frecuencia precios baratos en Carrefour y COTO, mientras que caros en Jumbo. Belgrano se posiciona como el barrio con más dispersión de precios entre sucursales:

Reglas

- 1 {zona=Noroeste,relativo.medio.discret=Medio caro} => {comercioRazonSocial=Jumbo Retail Argentina S.A.}
- 2 {zona=Noroeste,relativo.medio.discret=Medianamente barato} => {comercioRazonSocial=INC S.A.}
- 3 {comercioRazonSocial=Jumbo Retail Argentina S.A.,zona=Noroeste} => {relativo.medio.discret=Medio caro}
- 4 {comercioRazonSocial=INC S.A.,zona=Noroeste} => {relativo.medio.discret=Medianamente barato}
- 5 {zona=Noroeste,variacion.total.discret=Aumento fuerte,relativo.medio.discret=Levemente caro} => {comercioRazonSocial=Jumbo Retail Argentina S.A.}
- 6 {zona=Noroeste,variacion.total.discret=Aumento fuerte,relativo.medio.discret=Medio} => {comercioRazonSocial=Coto Centro Integral de Comercializaci3n S.A.}
- 7 {comercioRazonSocial=INC S.A.,zona=Noroeste} => {relativo.medio.discret=Medio}
- 8 {comercioRazonSocial=Jumbo Retail Argentina S.A.,zona=Noroeste} => {relativo.medio.discret=Medio}
- 9 {barrio_zona=BELGRANO.Noroeste,relativo.medio.discret=Levemente caro} => {comercioRazonSocial=Jumbo Retail Argentina S.A.}

	Support	confidence	Lift
1	0.01616611	0.73716381	2.72903524
2	0.01336452	0.69769069	2.66359993
3	0.01616611	0.22156899	2.39848469
4	0.01336452	0.18554015	2.27804279
5	0.02821697	0.38673526	1.67785074
6	0.0133109	0.58053201	1.45796018
7	0.01227874	0.17046618	0.65837602
8	0.01213129	0.1662686	0.64216411
9	0.01136722	0.74190726	2.74659583

Figure 12: Reglas generales y sus respectivas métricas 7

Desaceleración del último periodo

La única variable relacionada con variación que se mantuvo en el análisis fue la de variación 3 discretizada, es decir, la que indica de una forma discreta la variación de precios del último periodo con respecto al periodo anterior. Así mismo, en cuanto a precios relativos, solamente permaneció precios relativo medio. Las variables binarias de los productos tampoco fueron incluidas. Partiendo de esta preselección, se generan reglas cuyos resultados se filtran principalmente por tres factores: aquellas que presentan un valor más elevado de lift, aquellas que arrojan como consecuencia la variable de variación 3 discretizada y dentro de estas, aquellas cuyo valor fueron disminuciones, mantiene o incremento leve, obteniendo las siguientes reglas interesantes:

- Carrefour, es decir INC S.A., como se aprecia en las siguientes reglas, exhibe una relación entre los valores disminución fuerte, disminución leve y disminución media para la variación en cuestión (ver figura 13):

Reglas		
{comercioRazonSocial=INC S.A.} => {variacion3.discret=Disminucion fuerte}		
{comercioRazonSocial=INC S.A.} => {variacion3.discret=Disminucion leve}		
{comercioRazonSocial=INC S.A.} => {variacion3.discret=Disminucion media}		
Support	Confidence	Lift
0.021	0.079	2.192
0.015	0.057	1.804
0.012	0.045	1.785

Figure 13: Reglas relativas a la desaceleración del último periodo y sus respectivas métricas 1

Puede evidenciarse que la regla con mejores métricas es la asociada a disminuciones fuertes y el peso de las mismas va decreciendo con las siguientes reglas, lo que evidencia que la política de reducción de precios de Carrefour fue bastante perceptible.

- Jumbo presento una disminución leve en general, sin embargo, los productos cuyos precios se encuentran en el rango medio y levemente caro presentaron aumentos leves (ver figura 14):

Reglas
{comercioRazonSocial=Jumbo Retail Argentina S.A.} => {variacion3.discret=Disminucion leve}
{relativo.medio.discret=Medio,comercioRazonSocial=Jumbo Retail Argentina S.A.} => {variacion3.discret=Aumento leve}
{relativo.medio.discret=Levemente caro,comercioRazonSocial=Jumbo Retail Argentina S.A.} => {variacion3.discret=Aumento leve}

Support	Confidence	Lift
0.012	0.045	1.420
0.020	0.468	1.617
0.040	0.388	1.342

Figure 14: Reglas relativas a la desaceleración del último periodo y sus respectivas métricas 2

La regla de disminución leve es mucho menor que la observada para las demás, por lo que se puede afirmar que se observa con mayor frecuencia aumentos leves para esta cadena.

- Las reglas asociadas a COTO permiten apreciar que dicha cadena mantuvo sus precios para aquellos productos catalogados como levemente baratos durante el periodo bajo análisis (ver figura 15).

Reglas
{relativo.medio.discret=levemente barato,comercioRazonSocial=Coto Centro Integral de Comercialización S.A.} => {variacion3.discret=Mantiene}

Support	Confidence	Lift
0.044	0.451	1.317

Figure 15: Reglas relativas a la desaceleración del último periodo y sus respectivas métricas

Puede concluirse así que la desaceleración del aumento de precios está fuertemente ligada a la empresa de estudio con diferentes comportamientos en cada una, siendo mucho más evidente en la empresa Carrefour.

Comportamiento de los productos

Se tuvieron en cuenta las mismas variables que para descripción general además de las 28 variables binarias de los términos más frecuentes. Se ordenan los resultados según su lift y se filtra a partir de la palabra termino, pues así se denominan las columnas de los productos (por ejemplo, la de vino se denomina termino_vino).

Se seleccionaron un total de 19 reglas que serán presentadas en grupos de 5 ordenadas según su lift:

- La cadena Carrefour, el producto vino, esta ligado a precios muy baratos (figura 16). Para cadena Coto, este mismo lote de productos presenta un precio relativo medio caro. En la siguiente regla se observa que los vinos para la cadena COTO en el total del periodo de estudio, mantuvieron los precios. Luego, se aprecia que la sal mantuvo su precio. Finalmente se aprecia que los jugos levemente caros, tuvieron un aumento medio.

Reglas
1 {comercioRazonSocial=INC S.A.,termino_vino=S} => {relativo.medio.discret=Muy barato}
2 {comercioRazonSocial=Coto Centro Integral de Comercialización S.A.,termino_vino=S} => {relativo.medio.discret=Medio caro}
3 {comercioRazonSocial=Coto Centro Integral de Comercialización S.A.,termino_vino=S} => {variacion.total.discret=Mantiene}
4 {termino_sal=S} => {variacion.total.discret=Mantiene}
5 {relativo.medio.discret=Levemente caro,termino_jugo=S} => {variacion.total.discret=Aumento medio}

	Support	Confidence	Lift
1	0.013	0.736	16.119
2	0.013	0.460	5.021
3	0.014	0.496	3.378
4	0.020	0.427	2.906
5	0.012	0.675	2.845

Figure 16: Reglas relativas a los productos y sus respectivas métricas 1

- Se puede observar en la figura 17 que el aperitivo tuvo un aumento medio de precios indistintamente de las demás variables. Así mismo, que los jugos de mayor preponderancia en Coto poseen precios levemente caros. El yogur y el vino también mantuvieron mayoritariamente sus precios. Por último, se observa que los productos de leche encontrados en Coto poseen principalmente precios medios. Se puede apreciar que las métricas siguen siendo relevantes pues todos los lift se encuentran alejados considerablemente de 1, así como confianzas superiores al 50%.

Reglas			
1	{termino_aperitivo=S} => {variacion.total.discret=Aumento medio}		
2	{comercioRazonSocial=Coto Centro Integral de Comercialización S.A.,termino_jugo=S} => {relativo.medio.discret=Levemente caro}		
3	{termino_yogur=S} => {variacion.total.discret=Mantiene}		
4	{termino_vino=S} => {variacion.total.discret=Mantiene}		
5	{comercioRazonSocial=Coto Centro Integral de Comercialización S.A.,termino_leche=S} => {relativo.medio.discret=Medio}		
	Support	Confidence	Lift
1	0.011	0.658	2.776
2	0.011	0.576	2.512
3	0.010	0.363	2.472
4	0.026	0.359	2.446
5	0.011	0.563	2.167

Figure 17: Reglas relativas a los productos y sus respectivas métricas

- En la figura 18 se aprecia un aumento medio para el queso y el jugo y aumento fuerte tanto para las bebidas como para los desodorantes. Así mismo, se observa que las galletitas disponibles mayoritariamente en Coto, poseen un precio relativo medio barato. Aun cuando los valores de lift para estas reglas son inferiores, siguen estando lo suficientemente alejados de uno para considerar que existe dependencia entre las variables provistas por las reglas.

Reglas	
1	{termino_queso=S} => {variacion.total.discret=Aumento medio}
2	{termino_bebida=S} => {variacion.total.discret=Aumento fuerte}
3	{termino_desodorante=S} => {variacion.total.discret=Aumento fuerte}
4	{termino_jugo=S} => {variacion.total.discret=Aumento medio}
5	{comercioRazonSocial=Coto Centro Integral de Comercialización S.A.,termino_galletitas=S} => {relativo.medio.discret=levemente barato}

	Support	Confidence	Lift
1	0.012	0.466	1.966
2	0.013	0.852	1.928
3	0.018	0.780	1.764
4	0.018	0.400	1.687
5	0.012	0.421	1.674

Figure 18: Reglas relativas a los productos y sus respectivas métricas

- El jabón, así como la gaseosa tuvieron aumentos fuertes durante el periodo bajo análisis, mientras que la leche tuvo un aumento medio (ver figura 19). Se evidencia también que las galletitas disponibles mayoritariamente en Jumbo tuvieron un aumento fuerte. Se evidencia que los valores de lift se estabilizan alrededor de 1.5 para las reglas, lo cual les confiere el mismo grado de relevancia en adelante, poseyendo estas a su vez valores semejantes de soporte

Reglas			
1	{termino_leche=S} => {variacion.total.discret=Aumento medio}		
2	{termino_jabon=S} => {variacion.total.discret=Aumento fuerte}		
3	{variacion.total.discret=Aumento fuerte,comercioRazonSocial=Jumbo Retail Argentina S.A.} => {termino_galletitas=S}		
4	{termino_gaseosa=S} => {variacion.total.discret=Aumento fuerte}		

	Support	Confidence	Lift
1	0.019	0.392	1.651
2	0.016	0.703	1.591
3	0.012	0.107	1.539
4	0.014	0.669	1.515

Figure 19: Reglas relativas a los productos y sus respectivas métricas 4

Con esto puede concluirse que en general las bebidas, a excepción de las bebidas alcohólicas, son los productos comunes más susceptibles a ser afectados por el incremento de precios para los productos analizados en el presente estudio. Respecto a esto último, los productos de cosmética personal también presentan esta característica.

Reglas de carácter predictivo

Para evidenciar patrones predictivos se generarán reglas de asociación mediante el algoritmo apriori sobre los datos en 4 etapas, las primeras tres consistirán en encontrar reglas que permitan explicar los precios relativos de los tres primeros periodos y en la última se validaran estas reglas con las encontradas para el último periodo, encontrando así cuan prevalentes siguen siendo con fines de análisis predictivos. No se seleccionarán todas las columnas, en su lugar solo permanecerán en cuanto a precio la respectiva al periodo que se esté analizando, la razón social de la sucursal, así como las variables concernientes a ubicación. Una vez ejecutado el algoritmo se organizan los resultados de acuerdo al lift, para luego filtrar a partir de la palabra relativo como consecuencia de la regla. Los resultados se exponen en la figura 20, 21 y 22.

Para todos los periodos de estudio, en general la razón social Jumbo está asociada a precios muy caros, Carrefour a precios muy baratos, y Dia a medianamente barato. Estos resultados se exponen

en la figura 20. Para la cadena Jumbo en la mayoría de las zonas se observan precios muy caros, teniendo por otro lado precios medio caros para la zona suroeste, así como para el barrio de recoleta. La cadena Día presenta precios levemente baratos en la zona suroeste aun cuando en general se encuentran precios medianamente baratos.

Reglas	
1	{comercioRazonSocial=Jumbo Retail Argentina S.A.,zona=Noreste} => {relativo1.discret=Muy caro}
2	{comercioRazonSocial=DIA Argentina S.A.} => {relativo1.discret=Medianamente barato}
3	{comercioRazonSocial=INC S.A.,zona=Noroeste} => {relativo1.discret=Muy barato}
4	{barrio=RECOLETA,comercioRazonSocial=Jumbo Retail Argentina S.A.} => {relativo1.discret=Medio caro}
5	{comercioRazonSocial=Jumbo Retail Argentina S.A.} => {relativo2.discret=Muy caro}
6	{comercioRazonSocial=INC S.A.} => {relativo2.discret=Muy barato}
7	{comercioRazonSocial=INC S.A.,zona=Noreste} => {relativo2.discret=Muy barato}
8	{comercioRazonSocial=Jumbo Retail Argentina S.A.,zona=Noroeste} => {relativo3.discret=Muy caro}
9	{comercioRazonSocial=Jumbo Retail Argentina S.A.,zona=Suroeste} => {relativo3.discret=Medio caro}
10	{comercioRazonSocial=DIA Argentina S.A.,zona=Suroeste} => {relativo3.discret=levemente barato}

	Support	Confidence	Lift
1	0.022	0.165	3.692
2	0.016	0.243	2.678
3	0.011	0.141	2.642
4	0.011	0.225	2.282
5	0.048	0.179	3.007
6	0.048	0.180	2.719
7	0.020	0.166	2.505
8	0.011	0.158	2.947
9	0.011	0.204	2.046
10	0.012	0.394	1.671

Figure 20: reglas predictivas y sus respectivas métricas 1

Como era de esperarse, las reglas previamente encontradas siguen estando presentes para el último periodo, sin embargo, existe diferencia en cuanto a sus métricas (ver figura 21):

Reglas	
1	{comercioRazonSocial=Jumbo Retail Argentina S.A.,zona=Noreste} => {relativo4.discret=Muy caro}
2	{comercioRazonSocial=DIA Argentina S.A.} => {relativo4.discret=Medianamente barato}
3	{comercioRazonSocial=INC S.A.,zona=Noroeste} => {relativo4.discret=Muy barato}
4	{barrio=RECOLETA,comercioRazonSocial=Jumbo Retail Argentina S.A.} => {relativo4.discret=Medio caro}
5	{comercioRazonSocial=Jumbo Retail Argentina S.A.} => {relativo4.discret=Muy caro}
6	{comercioRazonSocial=INC S.A.} => {relativo4.discret=Muy barato}
7	{comercioRazonSocial=INC S.A.,zona=Noreste} => {relativo4.discret=Muy barato}
8	{comercioRazonSocial=Jumbo Retail Argentina S.A.,zona=Noroeste} => {relativo4.discret=Muy caro}
9	{comercioRazonSocial=Jumbo Retail Argentina S.A.,zona=Suroeste} => {relativo4.discret=Medio caro}
10	{comercioRazonSocial=DIA Argentina S.A.,zona=Suroeste} => {relativo4.discret=levemente barato}

	Support	Confidence	Lift
1	0.022	0.166	3.235
2	0.011	0.170	2.217
3	0.010	0.135	2.709
4	0.010	0.208	2.087
5	0.042	0.157	3.048
6	0.035	0.133	2.673
7	0.015	0.129	2.590
8	0.011	0.151	2.929
9	0.011	0.204	2.051
10	0.012	0.406	1.620

*Figure 21: reglas predictivas
y sus respectivas métricas 2*

Para evaluar esta diferencia, a continuación, se evalúa la diferencia porcentual del lift respecto a las encontradas en los periodos anteriores. Puede apreciarse que las reglas tuvieron cambios de lift tanto en sentido negativos como positivos. Es apreciable gran descenso para la regla que asocia Jumbo y zona noreste con muy caro dando a entender que los precios se abarataron para el último periodo de estudio para esta zona para esta cadena. Para la cadena Carrefour la regla que la asocia a muy barato aumento su lift con respecto a la observada anteriormente. La cadena Jumbo para el barrio recoleta también cambio negativamente para los precios medianamente caro lo cual obedece a la tendencia de disminución de precios en el noreste. La cadena Día tuvo una variación significativa para los precios medianamente barato. Aquellas reglas que tuvieron la menor variación estuvieron relacionadas con la cadena Jumbo para las zonas noroeste y suroeste por lo que se intuye que los precios en dichas zonas para esta cadena permanecieron relativamente estables.

	Diferencia Lift
1	-12.36%
2	-17.22%
3	2.54%
4	-8.56%
5	1.35%
6	-1.72%
7	3.42%
8	-0.62%
9	0.22%
10	-3.04%

*Figure 22: Diferencia entre lift para las
reglas escogidas VS las encontradas en el
periodo 4*

Visualización

En pos de construir gráficos que sean claros y comuniquen información relevante, sólo nos concentramos en crear reglas que asocien cadenas, precios medios, variaciones totales y la zona. A

su vez, solo graficaremos las siguientes 10 reglas pues son las que cuentan con mayor métrica lift. Estas reglas se exponen en la figura 23.

	Lhs	Rhs	Support	Confidence	Lift	Count
1	{comercioRazonSocial=Jumbo Retail Argentina S.A., variacion.total.discret=Aumento medio}	=> {relativo.medio.discret=Muy caro}	0.010	0.140	3.453	1513
2	{comercioRazonSocial=Jumbo Retail Argentina S.A., zona=Noreste}	=> {relativo.medio.discret=Muy caro}	0.019	0.140	3.444	2777
3	{zona=Noreste, relativo.medio.discret=Muy caro}	=> {comercioRazonSocial=Jumbo Retail Argentina S.A.}	0.019	0.916	3.390	2777
4	{variacion.total.discret=Aumento medio, relativo.medio.discret=Muy caro}	=> {comercioRazonSocial=Jumbo Retail Argentina S.A.}	0.010	0.879	3.255	1513
5	{relativo.medio.discret=Muy caro}	=> {comercioRazonSocial=Jumbo Retail Argentina S.A.}	0.035	0.867	3.211	5260
6	{comercioRazonSocial=Jumbo Retail Argentina S.A., variacion.total.discret=Aumento fuerte, relativo.medio.discret=Muy barato}	=> {comercioRazonSocial=Jumbo Retail Argentina S.A.}	0.035	0.131	3.211	5260
7	{zona=Noreste, relativo.medio.discret=Muy barato}	=> {comercioRazonSocial=INC S.A.}	0.012	0.786	3.002	1804
8	{variacion.total.discret=Aumento fuerte, relativo.medio.discret=Muy caro}	=> {comercioRazonSocial=INC S.A.}	0.014	0.782	2.985	2083
9	{comercioRazonSocial=Jumbo Retail Argentina S.A., variacion.total.discret=Aumento fuerte, relativo.medio.discret=Muy caro}	=> {comercioRazonSocial=Jumbo Retail Argentina S.A.}	0.013	0.805	2.979	1909
10	{comercioRazonSocial=Jumbo Retail Argentina S.A., variacion.total.discret=Aumento medio}	=> {relativo.medio.discret=Medio caro}	0.019	0.265	2.865	2853

Figure 23: Reglas a ser representadas gráficamente

La figura 24 muestra un gráfico de coordenadas paralelas en el cual la intensidad del color de las flechas representa la métrica confianza, mientras que el ancho de estas representa el soporte. A su vez, el eje de las abscisas de este grafico representa los ítems en el antecedente de la regla (lhs), por orden (1 y 2) y el ítem correspondiente al consecuente (rhs), mientras que en el eje de las ordenadas se representan los ítems. La quinta regla exhibida ({Relatico: muy caro} → {Jumbo}) en posee el mayor soporte relativo representado por la flecha gruesa y una intensidad de color fuerte que responde al elevado nivel de confianza relativo. Sin embargo, la regla 8 ({Noreste, Relatico: muy barato} → {INC S.A.}) queda levemente relegada teniendo en cuenta el nivel de confianza relativo de la regla, ya que la flecha parece imperceptible. En resumen, se pueden inferir relaciones similares a las expuestas durante este trabajo: Jumbo parece consolidarse como cadena que exhibió productos con precios más caros con las variaciones más fuertes, mientras que en las cadenas de Carrefour es más frecuente hallar productos con precios bajos, incluso aquellos bienes que exhibieron variaciones porcentuales fuertes durante el periodo.

Otra forma de visualizar las mismas reglas es a través de una figura que se concentre en la métrica lift. En este sentido, la figura 25 pinta con más intensidad conforme el valor del lift de la regla sea mayor. En esta figura, en el eje de las abscisas se representa los antecedentes, mientras que en el eje de las ordenadas los consecuentes. Como es de esperar, la primera regla ({Jumbo, Variación: aumento medio} → {Relativo: muy caro}), al cual posee mayor lift está pintada con el rojo más intenso en este gráfico (punto (1,4)). Al igual que se desprende de la figura uno y la tabla exhibida debajo, las conclusiones a las cuales se arriba mediante este grafico coinciden con las expuestas a lo largo de este trabajo.

Figure 24: Gráfico coordenadas paralelas

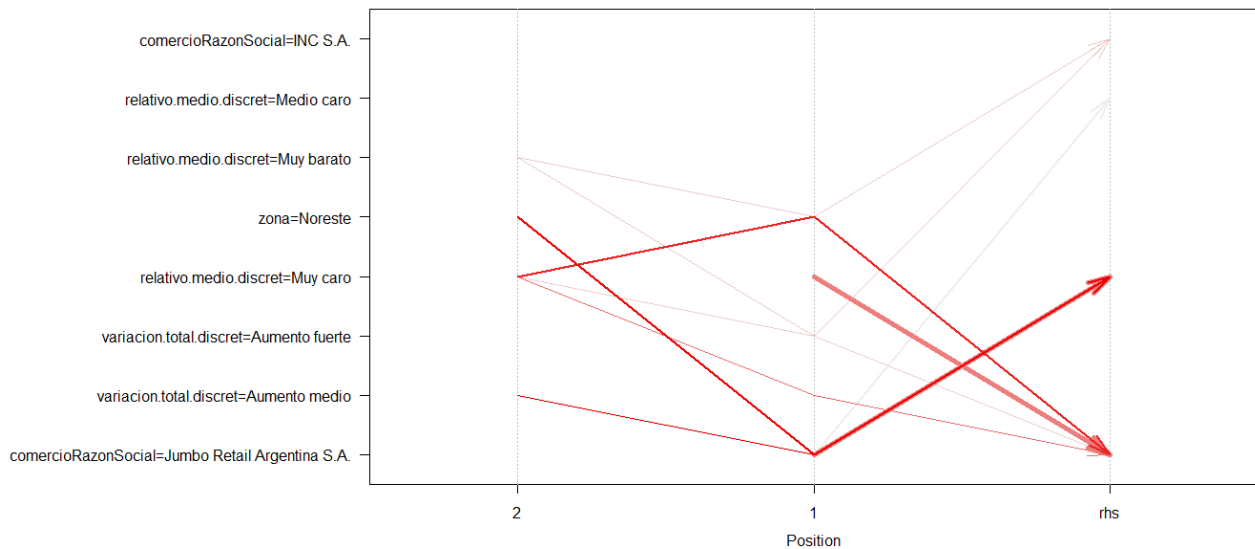
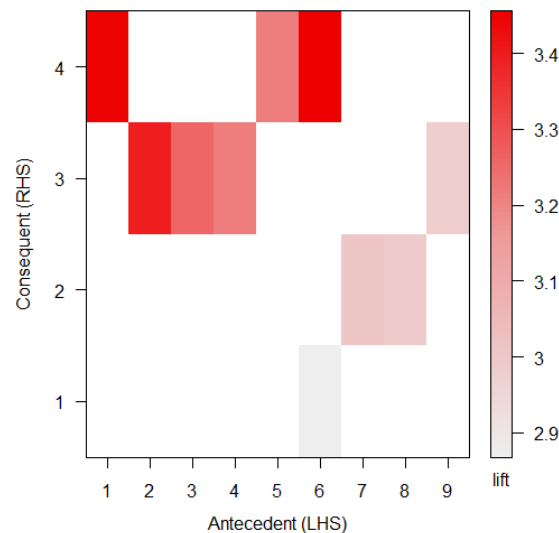


Figure 25: Gráfico basado en matrices



Relación con análisis previos

Teniendo en cuenta las conclusiones encontradas en el anterior trabajo, a partir de las asociaciones reportadas a lo largo de este trabajo, se pueden reafirmar la evolución del costo de una canasta de bienes de primera necesidad o esenciales promedio para CABA, por barrio y por sucursal. Esta canasta se construyó utilizando como base la lista de 64 productos del plan Precios Esenciales. En dicho trabajo se reportó evidencia consistente con dispersión en el costo de la canasta de bienes entre barrios y al interior de los barrios. Además, se encontró que Jumbo y COTO son las cadenas donde acceder a dicha canasta es más caro, mientras que en Carrefour y Día el acceso a la misma canasta es más barato. Por último, se reportó heterogeneidades entre los barrios asociadas al costo de la canasta. Los barrios más baratos, en términos de dicha canasta, se asocian con: Colegiales, Mataderos, Monserrat, Nuñez, Parque Chacabuco, Paternal, San Nicolas, San Telmo, Versalles, Villa del Parque y Villa Pueyrredon; y los más caros son: Boedo, Chacarita, Constitución, Monte Castro, Puerto Madero, Retiro, Saavedra, San Cristobal, Villa Crespo, Villa Gral. Mitre y Villa Santa Rita.

En este sentido, este trabajo exhibe asociaciones que están en línea con los resultados expuestos anteriormente: Jumbo se consolida como la cadena con precios relativos más caros respecto del promedio general, mientras que Carrefour y Día parecen exhibir los precios relativos más baratos y, por último, los supermercados de la cadena COTO parecen permanecer relativamente estables durante el periodo de análisis. Por otro lado, los resultados hallados en este trabajo parecen identificar como barrios caros aquellos agrupados en la zona de Noreste (con mayor lift aparecieron Palermo y Recoleta) y Noroeste (Belgrano) en los cuales Jumbo parece consolidarse con mayor presencia, mientras que los barrios donde se reportaron precios relativos más bajos se asocian con las zonas Sureste (Parque Patricios aparece con mayor lift) y Suroeste donde Día y COTO parecen ganar peso. Estos resultados son en gran medida consistentes con los reportados oportunamente en el primer trabajo practico.

Conclusiones

- Las variables seleccionadas, así como las discretizaciones aplicadas son determinantes en el tipo de información que arrojan las reglas de decisión.
- La relevancia de las reglas se mide a partir de las métricas que poseen, sin embargo, existe un componente humano en la selección de lo interesante que son.
- Al ser las reglas de decisión un método de análisis basado en probabilidades de incidencia, tendrán mayor protagonismo las reglas provenientes de ocurrencias específicas mayores, es decir, comportamiento consistente.
- Los mayores precios, así como las mayores variaciones en la cadena JUMBO, reglas que poseen las mejores métricas con respecto a las demás. Asimismo, la cadena CARREFOUR, se encontró ligada a precios muy baratos, y en sus precios. La cadena coto por otra parte presenta precios en el rango medio, así como poca variación de sus precios.
- La zona norte es la que presenta los precios más elevados de la ciudad, específicamente la zona noreste, presentando la zona noroeste precios un poco más bajos con respecto a esta.
- En general las bebidas, a excepción de las bebidas alcohólicas las cuales no varían significativamente en el tiempo, son los productos con mas variaciones. Así mismo, los productos de cosmética personal también presentan esta característica.
- Los comportamientos observados relacionados al incremento de precios suelen mantener su prevalencia en el tiempo, pues las reglas encontradas para los primeros periodos siguen siendo significantes según sus métricas para el ultimo aun cuando existan ciertas variaciones al respecto.

Bibliografía

1. <https://www.preciosclaros.gob.ar/>
2. Jiawei Han, Micheline Kamber, Jian Pei. 2012. Tercera edición. Data Mining: Concepts and Techniques.