Maestría en Explotación de Datos
y Descubrimiento del Conocimiento
Universidad de Buenos Aires

Data Mining

Trabajo Práctico N°2

Reglas de Asociación

Lucas Tomasini

9 de Julio de 2019, Comisión 1

INTRODUCCIÓN

En el presente trabajo se incorpora la técnica de Reglas de Asociación para continuar analizando el comportamiento de precios a lo largo del tiempo de productos de supermercados e hipermercados de la ciudad de Buenos Aires, utilizando el mismo conjunto de datos del trabajo anterior. El objetivo general es abordar problemas puntuales y tratar de entenderlos a partir de observaciones de coocurrencia de factores, aplicando reglas de asociación sobre los datos de precios de productos masivos con el fin de encontrar asociaciones que permitan explicar el comportamiento de la oferta de productos, encontrar novedades y confirmar o ampliar el conocimiento descubierto en el trabajo anterior.

FUENTE DE DATOS

Los archivos de base a partir de los cuales se procesará e integrará toda la información son los mismos que se utilizaron en el trabajo anterior:

- *sucursales.json*: todas las sucursales relevadas de comercios del rubro de supermercados e hipermercados ubicados en CABA.
- *productos.json*: los 1000 productos de diferentes categorías con más frecuencia en el relevamientos de precios en CABA. Es importante aclarar que se agregó a este archivo el campo *categoria* donde, por simple inspección en cada uno de los productos, se determinó la categoría a la que pertenece (frescos, almacén, bebés, limpieza, etc).
- *precios.json*: mediciones de precios de lista sobre los 1000 productos seleccionados en las sucursales de tipo supermercados e hipermercados de CABA.
- *barrios.geojson*: información geoespacial de los barrios de CABA que contiene para cada barrio el conjunto de puntos que lo delimitan (polígono), además de la comuna a la que pertenece, área y perímetro.

PREPROCESAMIENTO E INTEGRACIÓN DE DATOS

Antes de aplicar reglas de asociación realizamos una serie de transformaciones de variables sobre el conjunto de datos de precios, productos y sucursales.

Tratamiento de Precios

Para aplicar reglas de asociación es necesario discretizar las variables numéricas. A partir del dataframe de precios realizamos una serie de transformaciones para extraer características relacionadas con la competitividad de los precios en diferentes períodos y la variación de precios entre éstos. Los períodos los obtenemos agrupando los diez intervalos de mediciones en diferentes meses, donde cada período corresponde a un mes distinto:

Período	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

***Tabla 1**: Períodos asociados a las diferentes mediciones*

Primero pasamos las mediciones de precios de productos por sucursal a formato columnar, donde cada fila representará un producto de una sucursal dada con 10 columnas correspondientes a los precios en cada medición. Del dataframe original que contenía 1,584,661 filas pasamos a tener un nuevo dataframe, que llamaremos *Precios*, de 163,311 filas. Algunos de los valores de precios contienen faltantes, las cuales imputamos tomando el promedio de las mediciones inmediatamente contiguas. Este procedimiento

mantiene faltantes en caso de que ambas mediciones aledañas contengan faltantes. Luego generamos cuatro nuevas columnas de precios con los precios promedios de cada período descripto en la *Tabla 1*. También agregamos una columna extra con el promedio total de todas las mediciones. Si en alguno de los períodos resultantes quedó algún *NA* eliminamos la fila por completo. De esta manera, la tabla *Precios* queda con 157,604 registros sin faltantes (se eliminaron un 5.8% luego de la imputación).

Como segundo paso, calculamos las variaciones porcentuales de precios intra-período y la variación total entre el precio del primero y del último, agregando cuatro nuevas columnas a la tabla de precios (Var12, Var23, Var34 y VarTotal). La variación se calcula como:

$$variación = (precio_{nuevo} - precio_{inicial}) / precio_{inicial}$$

Para tener un valor comparable de precios de un determinado producto en las distintas sucursales calculamos, para cada período, el precio promedio de cada producto promediando en todas las sucursales. Con estos promedios calculamos luego el precio relativo de cada producto en una dada sucursal, por cada período y en total. El precio relativo se calcula como:

$$precio\ relativo = (precio_{producto\ en\ sucursal} - precio_{producto\ promedio}) / precio_{producto\ promedio}$$

Por último, discretizamos las variaciones de precios y los precios relativos en siete categorías cada uno, de acuerdo a los siguientes rangos:

Variación	Rango
Disminución Fuerte	(−∞; −0.05)
Disminución Media	[−0.05; −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; +∞)

***Tabla 2**: Discretización de variaciones de precios*

Precio relativo	Rango
Muy barato	(−∞;−0.1)
Medianamente barato	[−0.1; −0.05)
Levemente barato	[−0.05; −0.01)
Medio	[−0.01; 0.01)
Levemente caro	[0.01; 0.05)
Medio caro	[0.05; 0.1)
Muy caro	[0.1; +∞)

***Tabla 3**: Discretización de precios relativos*

En resumen, luego de aplicar las transformaciones anteriores y dejando únicamente las columnas con valores categóricos, el dataframe de *Precios* queda conformado por las siguientes 11 columnas: *producto, sucursal, Precio_P1, Precio_P2, Precio_P3, Precio_P4, Precio_Total, Var12, Var23, Var34* y *VarTotal*.

Tratamiento de Sucursales

Primero convertimos el dataframe *Sucursales* a tipo *SpaciaPoints* (conjunto de puntos), donde indicamos que los campos *lng* y *lat* serán las coordenadas (longitud y latitud) de cada sucursal representada como un punto con atributos. Establecemos el mismo Sistema de Coordenadas Geográficas en los polígonos de *Barrios* y en los puntos de *Sucursales* para poder generar consultas geoespaciales entre ambos. Para eso utilizamos el paquete *sp* de R, que mediante la función *over* va a encontrar, para cada sucursal representada como un punto, qué polígono (barrio) la contiene. De esa manera ya tenemos integrada la información de sucursales con la de barrios y comunas en una sola tabla: *Sucursales_Barrios*.

Por último, agregamos una columna *AVENIDA* que indica si la sucursal se encuentra o no ubicada sobre avenida. Observamos que del campo *direccion* podemos identificar esto haciendo un match de aquellas direcciones que contengan “Av”. Algunas direcciones que, siendo avenida no contienen “Av” (como Olazábal o Elcano) fueron computadas por inspección.

Tratamiento de Productos

El conjunto de datos *productos.json* contiene 3 campos textuales que describen al producto: *nombre, marca* y *presentacion*. A partir de éstos extraemos palabras que describan a los productos para ser utilizadas en las reglas de asociación. Para eso realizamos primero una serie de transformaciones sobre estos campos.

Primero convertimos todos los textos a minúscula, quitamos los dígitos numéricos, tildes y símbolos de puntuación, eliminamos palabras vacías en español (preposiciones, artículos, etc.) y palabras de un solo caracter, y borramos espacios demás entre palabras y a ambos extremos del texto. Luego obtenemos el listado de unidades de presentación y el listado de marcas, ambos sin repetidos, y los utilizamos para remover del campo *nombre* las unidades de presentación y la marca.

Finalmente, luego de aplicar estas transformaciones, generamos con el paquete *tm* de R un corpus de palabras a partir de los nombres de los productos ya procesados. Se observa que hay 843 términos distintos y 1000 documentos (correspondientes a los nombres de los 1000 productos). Formamos un vocabulario de palabras con aquellos términos cuya frecuencia sea mayor o igual a 15 dentro del corpus y, por cada palabra del vocabulario, generamos una columna de ausencia/presencia, donde *NA* indica ausencia y 'S' presencia. A cada columna le añadimos el prefijo *termino_* antes de cada palabra (ejemplo: *termino_agua*). Como resultado nos quedan agregadas 52 columnas de términos en la tabla *Productos*.

Integración de las tablas

Como último paso antes de aplicar reglas de asociación, integramos mediante *joins* las tablas transformadas de precios, sucursales y productos en un único dataframe que llamaremos *Precios_Full*.

Para las reglas de asociación necesitamos únicamente los campos que contengan información categórica (columnas del tipo *factor*). Por eso no tenemos en cuenta, para la generación de reglas, las columnas *productoId, sucursalId, nombre, presentacion, provincia, banderaId, localidad, sucursalNombre, comercioId, lng, lat, perímetro, area* y *direccion* de la tabla *Precios_Full*.

La tabla final queda de 157,604 filas representando transacciones y 69 columnas que se despliegan en 503 items al transformarse en matriz de transacciones. Las columnas son los 5 precios relativos, las 4 variaciones, marca, categoría, tipo de sucursal, bandera, razón social, barrio, comuna, avenida y los 52 términos (aceite, agua, café, cerveza, chocolate, gaseosa, jugo, mermelada, queso, pack, yogur, etc).

REGLAS DE ASOCIACIÓN

Preliminares

Una vez lista la tabla *Precios_Full*, generamos todas las reglas posibles a partir de la misma. Las 69 columnas se convierten en 503 ítems, lo cual tiende a generar una enorme cantidad de reglas que después iremos filtrando utilizando varios criterios (tener en cuenta que el groso de esta cantidad de ítems son las diferentes marcas que existen de productos, las diferentes categorías en cada período y variación, los barrios porteños y los 52 términos más frecuentes). Ejecutando el algoritmo Apriori con un soporte mínimo de 0.002 (que representa aproximadamente 315 *itemsets* como mínimo para establecer una regla), una confianza mínima del 60% y un largo de regla mínimo y máximo de 2 y 7 ítems respectivamente, se generan 5,335,968 reglas. De esta cantidad obtenemos un subconjunto de reglas aplicando los siguientes criterios de filtrado para eliminar aquellas que sean triviales o de poco interés en cuanto a conocimiento:

- Si la regla contiene la bandera, que no contenga la razón social.
- Si la regla contiene el barrio, que no contenga la comuna, y viceversa.
- Si el antecedente contiene la marca, que el consecuente no contenga ni la categoría ni ningún término.
- Si la regla contiene algún barrio, que sea alguno de los 5 con más cantidad de sucursales: Palermo, Recoleta, Caballito, Balvanera y Belgrano.
- Si aparece algún término en el antecedente, que el consecuente no contenga ni la marca ni la categoría ni ningún otro término.

- Si aparecen precios de períodos adyacentes en el antecedente, que la regla no contenga la variación entre dichos períodos.
- Si la regla contiene el precio total, que no contenga precio individual de ningún período, y viceversa.
- Si el antecedente contiene la variación total, que la regla no contenga ninguna variación individual entre períodos, y viceversa.
- Si el antecedente contiene una variación entre dos períodos y el precio de alguno de esos dos períodos, que el consecuente no contenga el precio del otro período (ej: var12+P1=>P2)
- Si el antecedente contiene una variación entre dos períodos, que el consecuente no contenga variación entre períodos anteriores a los mismos (ej: var23=>var12)
- Si el antecedente contiene dos precios entre períodos, que el consecuente no contenga ni precios ni variaciones.
- Si el consecuente contiene precios de un determinado período, el antecedente puede contener solamente precios del período inmediatamente anterior.

Además de estos filtrados, nos quedamos con un subconjunto de reglas que se encuentren entre un mínimo y un máximo umbral de soporte, de manera tal de no seleccionar reglas triviales con gran soporte o reglas demasiado específicas y de poco interés de bajo soporte. En la *Figura 1* graficamos la distribución del logaritmo natural de los soportes de las 5 millones de reglas generadas (sin el logaritmo no se podría apreciar el decaimiento, ya que la mayoría de los valores estarían pegados al eje y). Tomamos soportes entre -6 y -3 aproximadamente (0.0025 y 0.05) y un *lift* mínimo de 1.3.

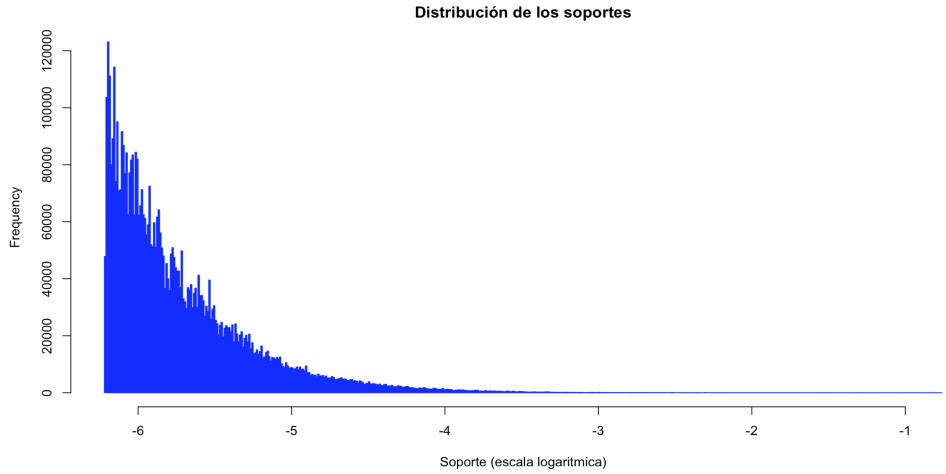

***Figura 1**: Distribución logarítmica de los soportes*

El subconjunto resultante luego de todos los filtrados contiene 24000 reglas. Trabajaremos principalmente con este subconjunto para generar diferentes reglas de interés, aplicando, por supuesto, filtrados posteriores para acotar el problema dependiendo del análisis.

Análisis descriptivo

En esta primera parte se busca analizar mediante reglas de decisión la aceleración/desaceleración de los precios en los últimos períodos y los diferentes factores asociados a este comportamiento.

La *Tabla 4* muestra cómo se distribuyen las variaciones de los precios de los productos entre períodos (cada columna suma 157,604, el total de transacciones). Entre el período 2 y 3 hay 78137 productos que aumentaron sus precios, mientras que entre el período 3 y 4 hay 88373 (un aumento del 13.1%). A su vez, entre el período 2 y 3, hay una reducción del 23.9% de productos que mantienen sus precios respecto de los períodos anteriores (Var23=71521 y Var34=54419). También, como contraparte, en Var34 hay un aumento de la cantidad de productos que disminuyen sus precios, en comparación con Var23 (pasan de 7946 a 14775). En principio parecería ser que predomina más la aceleración de precios en los últimos períodos que la desaceleración.

	Var12	Var23	Var34
Aumento Fuerte	16651	16386	18249
Aumento Leve	26936	35095	45438
Aumento Medio	25126	26656	24723
Disminución Fuerte	7645	3285	5719
Disminución Leve	3339	2716	5023
Disminución Media	4218	1945	4033
Mantiene	73689	71521	54419

***Tabla 4**: Tabla de contingencia de variaciones*

Para un análisis más detallado, generamos reglas simples de la forma Var23 => Var34 y analizamos el soporte y la confianza de cada una de ellas (49 en total, ya que hay 7 categorías por variación). Estas reglas se pueden representar visualmente mediante un diagrama de Sankey (*Figura 2*). El largo de cada barra lateral representa el soporte del ítem asociado a la misma y el ancho de cada línea de flujo entre dos barras representa el soporte de la regla asociada a los ítems de las mismas.

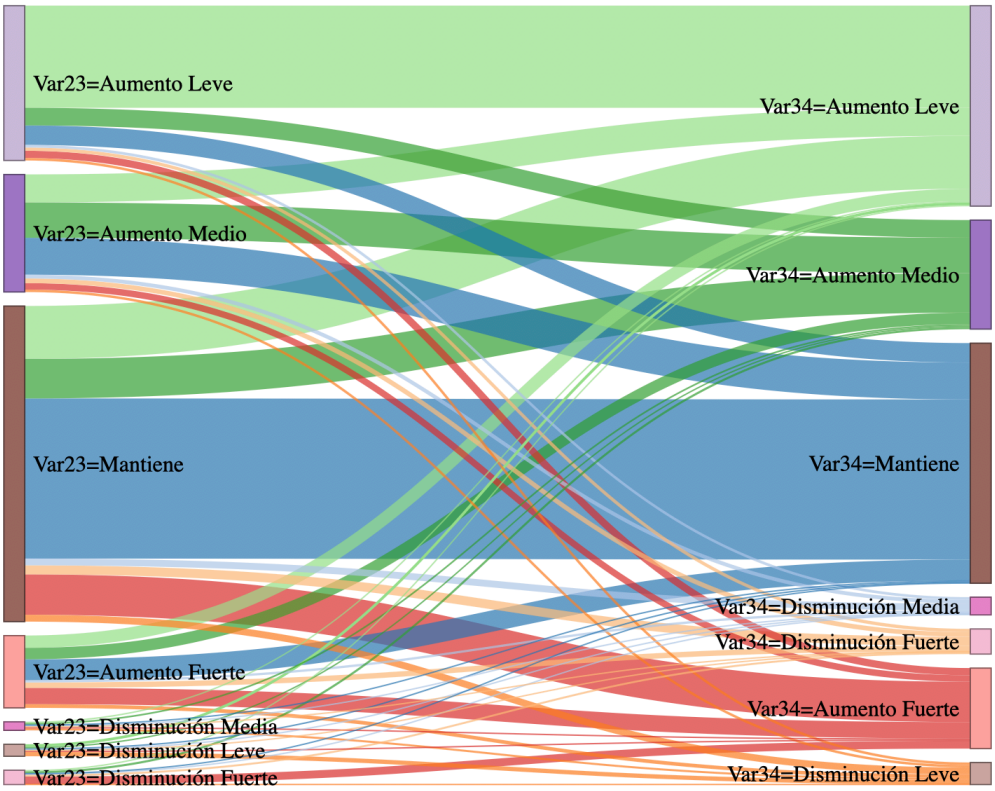

***Figura 2**: Diagrama de Sankey representando flujos de soporte*

A continuación detallamos las observaciones más sobresalientes extraídas de este diagrama y de las métricas asociadas a las 49 reglas (ver *Apéndice*):

- Un gran porcentaje de productos que mantuvieron sus precios entre los períodos 2 y 3, los aumentaron entre el período 3 y 4. Un 12.79% pasó a aumentar fuertemente, un 12.52% medianamente y un 16.77% levemente (ver confianza de las reglas correspondientes en *Apéndice*). Sólo el 7.9% pasó a disminuir de alguna manera. Esto representa aproximadamente 30 mil productos que aceleraron sus precios vs 5600 que desaceleraron.
- El 16.03% de los que aumentaron levemente, pasaron a aumentar medianamente y fuertemente, es decir, aceleraron. En cambio el 17.99% desaceleró, con lo cual se podría decir de manera aproximada que en balance hay una desaceleración de 680 productos.
- El 5.5% de los que aumentaron medianamente, pasaron a aumentar fuertemente (aceleraron). El 64.4% desaceleró, con lo cual en balance hubo 15700 productos que desaceleraron.
- El 77.43% de los que aumentaron fuertemente, desaceleraron, es decir, 12600 productos.

Los casos de productos que disminuyen de 2 a 3 los omitimos ya que tienen soporte bajo en comparación con el resto. Vale mencionar, sin embargo, que el 54% de los productos que disminuyeron fuertemente entre el período 2 y 3, aumentaron fuertemente entre el 3 y el 4, lo cual representa una aceleración abrupta de precios (en el diagrama está representado por la línea roja de abajo de todo).

En base a lo dicho hasta ahora, pareciera que no hay una aceleración o desaceleración general predominante sobre los precios de los últimos períodos, sino más bien grupos de productos que aceleraron y grupos que desaceleraron. Por lo tanto vamos a tomar algunos de estos grupos y buscar reglas que expliquen los factores de aceleración o desaceleración dentro de los mismos. Dentro de cada grupo sólo se muestran las reglas con mayor soporte.

Grupo *Mantiene=>Aumento Leve* (soporte total = 11996)

ID	Antecedente	Consecuente	Confianza	Lift	Soporte
1	{Var23=Mantiene,categoria=BEBIDAS SIN ALCOHOL, termino_jugo=S,termino_polvo=S}	Var34=Aumento Leve	83.5%	2.90	2001
2	{Var23=Mantiene,termino_jugo=S,termino_polvo=S, banderaDescripcion=COTO CICSA}	Var34=Aumento Leve	96.6%	3.35	1282
3	{Var23=Mantiene,categoria=FRESCOS, comercioRazonSocial=Jumbo Retail Argentina S.A.}	Var34=Aumento Leve	65.6%	2.28	1949
4	{Var23=Mantiene,categoria=FRESCOS,termino_pack=S, comercioRazonSocial=Jumbo Retail Argentina S.A.}	Var34=Aumento Leve	97.2%	3.37	486

Según la regla 1, el 83.5% de los jugos en polvo que mantuvieron sus precios entre 2 y 3, luego aumentaron levemente en los últimos períodos. Además, éstos representan el 16.6% (2001 de 11996) del total de productos dentro de este grupo. La regla 2 es un caso particular de la regla 1 en donde bajo una condición extra aumenta la confianza: el 96.6% de los jugos en polvo *de Coto* que mantuvieron sus precios entre 2 y 3, aumentaron levemente en 3 y 4.

Asimismo, la regla 3 establece que el 65.6% de los productos frescos de las sucursales Jumbo, Vea y Disco (razón social Jumbo Retail) que mantuvieron sus precios entre 2 y 3, aumentaron entre 3 y 4. Y éstos representan el 16.2% (1949 de 11996) del total de productos de este grupo. Además, según la regla 4, de estos 1949 hay 486 que vienen en packs y representan casi la totalidad (97.2%) de los productos frescos en packs de Jumbo que mantuvieron los precios entre 2 y 3.

Los productos que representan el soporte de las reglas 1 y 3 no se superponen, ya que si son jugos en polvo no pueden ser de la categoría frescos. Con lo cual podemos afirmar que las reglas 1 y 3 explican el 32.8% de los casos donde se pasó de mantener a aumentar levemente los precios.

Grupo *Mantiene => Aumento Fuerte* (soporte total = 9149)

ID	Antecedente	Consecuente	Confianza	Lift	Soporte
5	{Var23=Mantiene,termino_cafe=S}	Var34=Aumento Fuerte	78.4%	6.8	797
6	{Var23=Mantiene,marca=CUSENIER}	Var34=Aumento Fuerte	99.4%	8.6	478
7	{Var23=Mantiene,categoria=MASCOTAS, banderaDescripcion=COTO CICSA}	Var34=Aumento Fuerte	100%	8.6	360
8	{Var23=Mantiene,marca=KNORR, comercioRazonSocial=Jumbo Retail Argentina S.A.}	Var34=Aumento Fuerte	62.2%	5.4	281
9	{Var23=Mantiene,marca=MAIZENA}	Var34=Aumento Fuerte	100%	8.6	270
10	{Var23=Mantiene,marca=SER,termino_agua=S, banderaDescripcion=COTO CICSA}	Var34=Aumento Fuerte	100%	8.6	247
11	{Var23=Mantiene,marca=HELLMANN'S,categoria=ALMACEN, comercioRazonSocial=Jumbo Retail Argentina S.A.}	Var34=Aumento Fuerte	100%	8.6	242

Este grupo representa productos que aceleraron bastante sus precios, ya que saltan de Mantiene a Aumento Fuerte (3 niveles). La regla 6 es un caso particular de la regla 5, ya que todos los productos marca Cusenier contienen la palabra *café.* Se observa que dentro de este grupo, a diferencia del anterior, hay varias reglas de soporte más bien pequeño.

Según las reglas 7, 9, 10 y 11, *todos* los productos para mascotas y las aguas marca Ser de Coto, junto con productos marca Maizena y Hellman's de Jumbo pasaron de mantener sus precios a aumentar fuertemente en los últimos períodos. Dejando de lado la regla 6, que se superpone con la 5, este conjunto de reglas explica el 24% de los casos que se mantuvieron entre 2 y 3 y aumentaron fuertemente entre 3 y 4.

Grupo *Mantiene => Aumento Medio* (soporte total = 8959)

ID	Antecedente	Consecuente	Confianza	Lift	Soporte
12	{Var23=Mantiene,marca=GRANIX, categoria=ALMACEN}	Var34=Aumento Medio	67.9%	4.33	1196
13	{Var23=Mantiene,marca=GRANIX,termino_galletitas=S, banderaDescripcion=COTO CICSA}	Var34=Aumento Medio	100%	6.37	420
14	{Var23=Mantiene,marca=PAMPERS,categoria=BEBES}	Var34=Aumento Medio	71.7%	4.57	329
15	{Var23=Mantiene,marca=NIVEA, categoria=PERFUMERIA Y CUIDADO PERSONAL}	Var34=Aumento Medio	87.4%	5.57	306

Según la regla 12, el 67.9% de los productos marca Granix pasaron de mantener sus precios a aumentarlos medianamente en los últimos períodos. De esa cantidad el 35% (420 de 1196) son galletitas de Coto, las cuales en su totalidad pasaron de mantener a aumentar medianamente. También, por la regla 14 y 15, tenemos las marcas Pampers y Nivea, cuyos productos pasaron a aumentar medianamente en gran porcentaje (71.7% y 87.4% respectivamente). Todos estos productos representan el 20% del total de los casos.

Grupo *Aumento Medio => Mantiene* (soporte = 8356)

ID	Antecedente	Consecuente	Confianza	Lift	Soporte
16	{Var12=Aumento Medio,Var23=Aumento Medio}	Var34=Mantiene	70.4%	2.0	4850
17	{Var23=Aumento Medio,termino_mermelada=S}	Var34=Mantiene	89.3%	2.6	1055
18	{Var23=Aumento Medio,termino_galletitas=S}	Var34=Mantiene	74.7%	2.2	1834

Este grupo representa productos que desaceleraron sus precios. La regla 16 indica que el 58% (4850 de 8356) de todos los productos que aumentaron medianamente entre 2 y 3 y luego se mantuvieron entre 3 y 4, también habían aumentado medianamente en los primeros períodos. Es decir que la mayoría de estos productos siguieron la secuencia aumento medio => aumento medio => mantiene. A su vez, el 70.4% de los productos con Var12 y Var23 igual a Aumento Medio, mantuvieron sus precios en los últimos períodos.

Las reglas 17 y 18, cuyos soportes no se superponen ya que son mermeladas y galletitas, explican un 34.6% de los productos que desaceleraron en este grupo (1055+1834 de 8356). El 89.3% de las mermeladas que aumentaron medianamente entre 2 y 3, se mantuvieron en los últimos períodos. Lo mismo para el 74.7% de las galletitas.

Grupo *Aumento Fuerte => Aumento Fuerte* (soporte total = 3699)

ID	Antecedente	Consecuente	Confianza	Lift	Soporte
19	{Var12=Mantiene,Var23=Aumento Fuerte, categoria=PERFUMERIA Y CUIDADO PERSONAL, banderaDescripcion=COTO CICSA}	Var34=Aumento Fuerte	71.2%	6.2	742
20	{Var23=Aumento Fuerte,marca=CEPITA, categoria=BEBIDAS SIN ALCOHOL, banderaDescripcion=COTO CICSA}	Var34=Aumento Fuerte	80.9%	7.0	267
21	{Var12=Aumento Leve,Var23=Aumento Fuerte,marca=PATY}	Var34=Aumento Fuerte	95.7%	8.3	157

Este grupo en particular son los productos que aumentaron fuertemente las dos últimas veces. Un 20% corresponde a productos de perfumería y cuidado personal de Coto. Un 11.5% a productos de marca Cepita de Coto y a productos de marca Paty en general. De la regla 21 vemos que prácticamente todos los productos de la marca Paty que aumentaron levemente entre 1 y 2 y fuertemente entre 2 y 3, también aumentaron fuertemente entre 3 y 4: son productos que aumentaron vertiginosamente durante esos meses.

Grupo *Disminución Fuerte => Aumento Fuerte* (soporte total = 1774)

ID	Antecedente	Consecuente	Confianza	Lift	Soporte
	{Var23=Disminución Fuerte, categoria=BEBIDAS SIN ALCOHOL}	Var34=Aumento Fuerte	73.6%	6.4	735

Por último mencionamos que el 73.6% de las bebidas sin alcohol que disminuyeron fuertemente entre 2 y 3, pasaron a aumentar fuertemente entre 3 y 4. Son productos que aceleraron fuertemente en los últimos períodos, teniendo picos negativos seguidos de picos positivos.

Análisis predictivo a partir de reglas

El objetivo de esta sección es buscar reglas no triviales (utilizando características de sucursales y productos en el antecedente) que concluyan sobre los precios de los primeros 3 períodos. Es decir, a partir del subconjunto de 24000 reglas, generamos un nuevo subconjunto bajo la condición de que contenga precios de períodos en el consecuente y no contenga precios ni variaciones entre períodos en el antecedente. De este subconjunto seleccionamos reglas interesantes con buen soporte y confianza mayor a 0.6, y cada una de ellas las evaluamos en el último período. Es decir, buscamos el mismo consecuente de cada regla pero utilizando precios del período 4. En varios casos un mismo antecedente aplicado a diferentes períodos genera reglas con confianza inferior a 0.6 o soporte menor. Para detectar esos casos se permitieron reglas con confianza y soporte bajos, de manera tal de poder incluirlas en el análisis.

En la *Tabla 5* se muestran los resultados. Son 24 reglas en total, 6 por cada período. Las columnas en verde indican la confianza de la regla, en rojo el lift y en azul el soporte. Se marcaron en rojo también aquellos valores cuya confianza es inferior a 0.6.

La *Figura 3* representa la misma información de manera más visual. Es una matriz agrupada donde en el eje horizontal están los *labels* del consecuente y en el vertical los *itemsets* del antecedente. El tamaño de los círculos es proporcional al soporte y la intensidad del color es proporcional a la confianza.

		Precio_P1			Precio_P2			Precio_P3			Precio_P4		
Antecedente	Consecuente	Confianza (%)	Lift	Soporte	Confianza (%)	Lift	Soporte	Confianza (%)	Lift	Soporte	Confianza (%)	Lift	Soporte
{categoria=BEBIDAS SIN ALCOHOL, termino_agua=S,comercioRazonSocial= Jumbo Retail Argentina S.A.}	Levemente caro	79.1	3.5	1281	65.2	3.2	1056	67.5	3.1	1094	58.0	3.0	940
{categoria=BEBIDAS CON ALCOHOL, termino_vino=S, comercioRazonSocial=INC S.A.}	Muy barato	71.0	12.9	2039	72.4	10.8	2080	70.7	12.2	2029	75.2	14.7	2160
{categoria=BEBIDAS CON ALCOHOL, termino_tinto=S, comercioRazonSocial=INC S.A.}	Muy barato	77.6	14.1	1560	77.6	11.6	1560	78.6	13.5	1578	83.2	16.3	1672
{categoria=BEBIDAS SIN ALCOHOL, termino_agua=S, banderaDescripcion=COTO CICSA}	Medio	65.6	2.6	1570	67.4	2.7	1615	49.7	2	1191	54.2	2	1297
{termino_queso=S, banderaDescripcion=COTO CICSA}	Medio	69.3	2.8	1026	54.8	2.2	811	73.5	2.9	1088	64.0	2.3	948
{categoria=FRESCOS, termino_yogur=S, banderaDescripcion=COTO CICSA}	Medio	44.3	1.8	709	92.3	3.6	1478	83.3	3.4	1397	50.1	1.8	801

***Tabla 5**: Reglas con precios en distintos períodos*

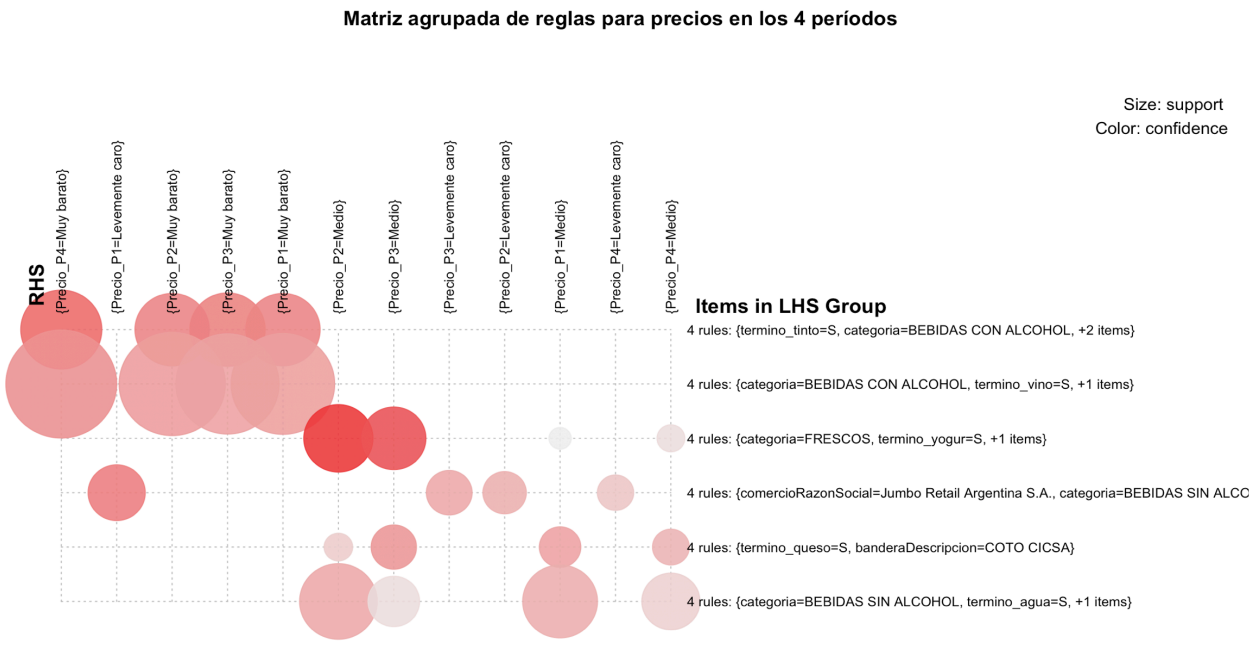

***Figura 3**: Matriz agrupada de reglas para precios en distintos períodos*

De las dos primeras reglas de la *Figura 3* vemos que los soportes son similares en los 4 períodos y la confianza relativamente buena en todos los casos. En particular, en la primera regla, la confianza es mayor en el período 4 que en los anteriores. Estas reglas nos dicen que los vinos de Carrefour (Market, Hipermercados y Express) son muy baratos en los cuatro períodos, con una confianza de 71%, 72.4%, 70.7% y 75.2% en P1, P2, P3 y P4 respectivamente (la confianza aumentó en el último período). Si tomamos el caso particular de vinos tintos, los niveles de confianza suben a 77.6%, 77.6%, 78.6% y 83.2%.

La tercera regla de la *Figura 3* revela que los yogures de Coto son de precio medio en los períodos 2 y 3, con un alto nivel de confianza: 92.3% y 82.3% respectivamente. En cambio la misma regla en los períodos 1 y 4 tiene baja confianza (44.3% y 50.1%) y el soporte se reduce a la mitad. No pasa lo mismo con los quesos de Coto, donde en los períodos 1, 3 y 4 sus precios son levemente caros con una confianza de 69.3%, 73.5% y 64.0% respectivamente y en el período 2 esto ocurre con menor soporte y confianza de 54.8%.

Por último, las reglas que se refieren al precio del agua establecen lo siguiente: el agua en Disco, Vea y Jumbo es levemente cara (sobre todo en los 3 primeros períodos) y el agua en Coto es de precio medio (sobre todo en los períodos 1 y 2). En particular vemos que la confianza y el soporte de la regla del agua en Jumbo son los más bajos en el período 4 y los más altos en el período 1. En el caso de Coto, la regla del agua tiene soporte y confianza más bajos en los últimos dos períodos.

Validación de resultados del TP1

En el trabajo práctico anterior se determinó, mediante dos técnicas por separado (Regresión Lineal y Análisis de Correspondencias) que los productos de perfumería y cuidado personal, limpieza y bebidas sin alcohol son los tres que estaban más asociados a una variabilidad alta en el tiempo, mientras que las bebidas con alcohol y los alimentos frescos son productos de variabilidad más baja en el tiempo. Vamos a validar estos resultados mediante el uso de reglas.

De la *Tabla 4* de frecuencias de variaciones entre períodos, vemos que en todos los casos más del 90% de los productos o mantuvieron sus precios o aumentaron. Con lo cual para comparar la tendencia entre categorías vamos a generar reglas cuyo consecuente contenga ítems de la forma VarXY=Mantiene o Aumenta. Es decir, hay 3 posibles intervalos distintos (Var12, Var23 y Var34) y 4 intensidades de variación (Mantiene, Aumento Leve, Aumento Medio y Aumento Fuerte). En total son 12 ítems para el consecuente. En el antecedente ubicamos las distintas categorías, que son 8 en total, ya que excluimos los productos de la categoría mascotas por tener baja cantidad de ejemplos (no son estadísticamente significativos).

De esta manera quedan conformadas 96 reglas que representamos de nuevo en una matriz agrupada donde en el eje horizontal están las categorías y en el vertical las variaciones. Para que sean comparables

las proporciones entre categorías utilizamos la confianza para representar el tamaño de los círculos, ya que los soporte varían bastante en cada categoría (por ejemplo, la categoría almacén es la que más productos contiene). El soporte lo representamos con la intensidad del color. Observar que las filas están ordenadas, de arriba hacia abajo, de menor a mayor incremento de variabilidad.

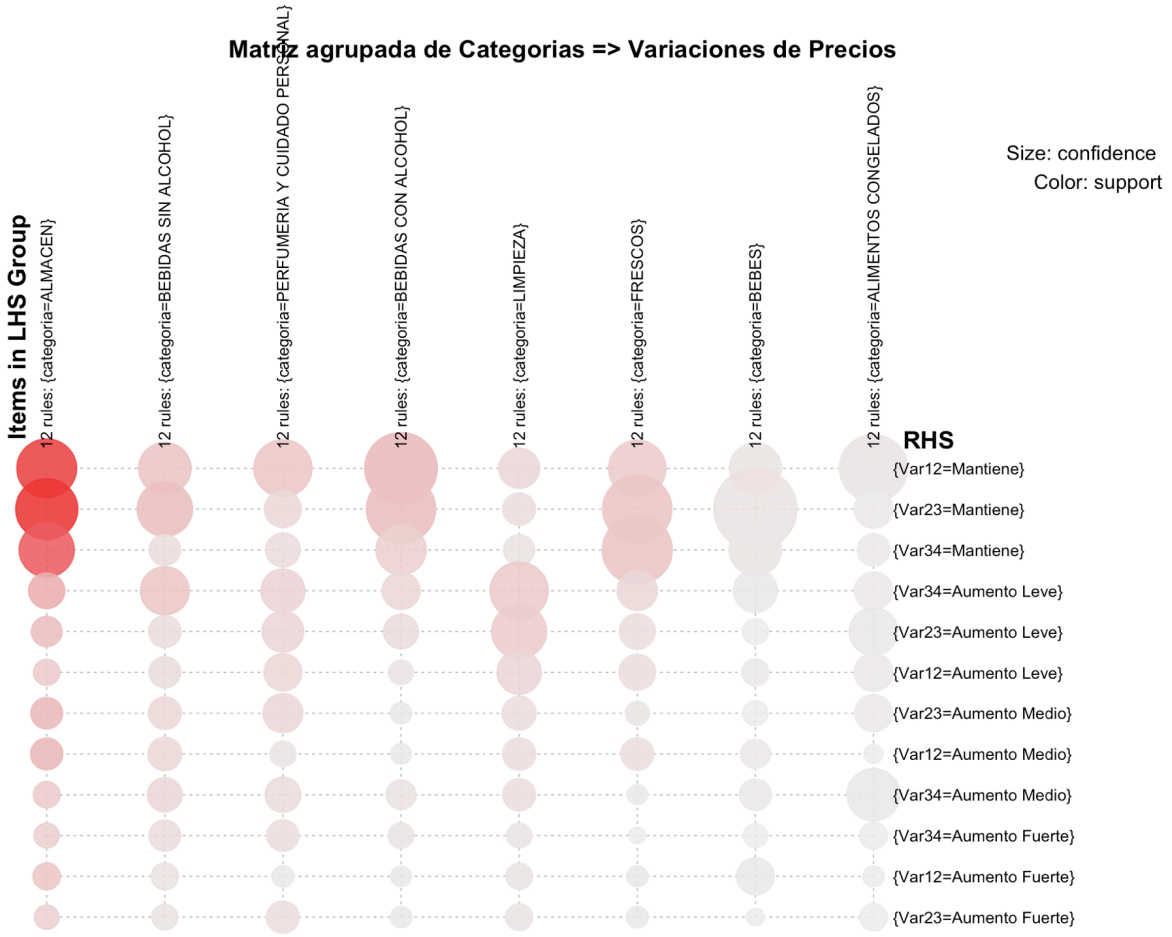

***Figura 4**: Matriz agrupada de reglas tipo categoría => variación de precios*

Se observa que las bebidas con alcohol y los frescos tienen la mayor proporción de sus productos agrupados en las tres primeras filas (correspondientes a Mantiene), mientras que pocos productos (círculos más pequeños) se encuentran en las filas de aumento de precios. En cambio, la categoría limpieza, perfumería y bebidas sin alcohol tienen menos proporción de productos que se mantienen en comparación con las dos anteriores. Además, la categoría limpieza tiene una buena proporción de sus productos en Aumento Leve (en los tres intervalos) y otra, aunque menor, en Aumento Medio. Las categorías de perfumería y cuidado personal y las bebidas sin alcohol tienen la mayoría de sus productos repartidos en Aumento Leve, Medio y Fuerte. La categoría almacén es un intermedio entre estos dos casos. Tiene menos proporción de productos en Mantiene que bebidas con alcohol y frescos pero más que bebidas sin alcohol, perfumería y limpieza. Y lo mismo sucede, pero al revés, con la proporción de productos en Aumento Leve, Medio y Fuerte (por supuesto que si nos referimos al soporte, tiene más productos que cualquier otra categoría, pero acá nos interesa analizar las proporciones, representadas por la confianza). De esta manera podemos concluir que los resultados obtenidos en el trabajo anterior se corresponden con estos.

APÉNDICE

Antecedente	Consecuente	Confianza (%)	Soporte
{Var23=Aumento Leve}	{Var34=Aumento Leve}	65.98	23156
{Var23=Mantiene}	{Var34=Aumento Leve}	16.77	11996
{Var23=Aumento Medio}	{Var34=Aumento Leve}	23.80	6346
{Var23=Aumento Fuerte}	{Var34=Aumento Leve}	17.03	2791

Antecedente	Consecuente	Confianza (%)	Soporte
{Var23=Mantiene}	{Var34=Aumento Fuerte}	12.79	9149
{Var23=Aumento Fuerte}	{Var34=Aumento Fuerte}	22.57	3699
{Var23=Disminución Fuerte}	{Var34=Aumento Fuerte}	54.00	1774
{Var23=Aumento Leve}	{Var34=Aumento Fuerte}	4.65	1633
{Var23=Aumento Medio}	{Var34=Aumento Fuerte}	5.50	1468

Antecedente	Consecuente	Confianza (%)	Soporte
{Var23=Mantiene}	{Var34=Aumento Medio}	12.52	8959
{Var23=Aumento Medio}	{Var34=Aumento Medio}	30.07	8018
{Var23=Aumento Leve}	{Var34=Aumento Medio}	11.38	3995
{Var23=Aumento Fuerte}	{Var34=Aumento Medio}	15.84	2596
{Var23=Disminución Fuerte}	{Var34=Aumento Medio}	14.64	481

Antecedente	Consecuente	Confianza (%)	Soporte
{Var23=Mantiene}	{Var34=Mantiene}	50.72	36282
{Var23=Aumento Medio}	{Var34=Mantiene}	31.34	8356
{Var23=Aumento Fuerte}	{Var34=Mantiene}	28.98	4750
{Var23=Aumento Leve}	{Var34=Mantiene}	12.57	4414

Antecedente	Consecuente	Confianza (%)	Soporte
{Var23=Mantiene}	{Var34=Disminución Leve}	21.78	1558
{Var23=Disminución Leve}	{Var34=Disminución Leve}	35.86	974
{Var23=Aumento Fuerte}	{Var34=Disminución Leve}	50.53	828
{Var23=Aumento Leve}	{Var34=Disminución Leve}	16.15	567
{Var23=Aumento Medio}	{Var34=Disminución Leve}	20.97	559
{Var23=Disminución Media}	{Var34=Disminución Leve}	23.55	458

Antecedente	Consecuente	Confianza (%)	Soporte
{Var23=Mantiene}	{Var34=Disminución Fuerte}	2.86	2046
{Var23=Aumento Fuerte}	{Var34=Disminución Fuerte}	7.42	1216
{Var23=Aumento Medio}	{Var34=Disminución Fuerte}	3.95	1055
{Var23=Aumento Leve}	{Var34=Disminución Fuerte}	2.10	739
{Var23=Disminución Fuerte}	{Var34=Disminución Fuerte}	13.21	434

Antecedente	Consecuente	Confianza (%)	Soporte
{Var23=Mantiene}	{Var34=Disminución Media}	2.14	1531
{Var23=Aumento Medio}	{Var34=Disminución Media}	3.20	854
{Var23=Aumento Leve}	{Var34=Disminución Media}	1.68	591
{Var23=Aumento Fuerte}	{Var34=Disminución Media}	3.08	506