Reglas de Asociación

Materia Minería de Datos
Maestría en Minería de Datos y Descubrimiento del Conocimiento, UBA
Primer Cuatrimestre 2019 - Comisión 2

Federico Moreno - Bianca Picchetti

Tabla de contenido

1. Introducción

En línea con el trabajo presentado previamente, *"Estudio de precios en la Ciudad Autónoma de Buenos Aires en el periodo Noviembre 2018 a Febrero 2019"* (1), en este trabajo se analiza el mismo *dataset* (*Apéndice*, *Tabla A.1*) aplicando la metodología de Reglas de Asociación. En el presente estudio se incorpora la técnica mencionada para abordar el análisis a partir de observaciones de coocurrencia de factores con el fin de ampliar el conocimiento descubierto anteriormente.

Uno de los objetivos de este estudio es realizar un análisis descriptivo del *dataset*. Se busca, mediante un número acotado de reglas, poder describir el comportamiento de los precios durante el total del periodo (*Tabla 1*). Además, se realiza un análisis en mayor detalle de los productos comprendidos en el grupo de alimentos recomendados para el consumo (2) (*Apéndice, Tabla A.2*).

Otro de los objetivos principales es aplicar la técnica con fines predictivos. Se busca conocer la variación de precios en el último periodo (*Tabla 1*) a partir de los factores asociados a la variación de precios en los tres primeros. A su vez, se busca responder algunas de las preguntas de investigación planteadas en el estudio anterior para así confirmar las respuestas a través de esta técnica.

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

Tabla 1: *Periodos asociados a las diferentes mediciones.*

2. Estructura del trabajo

El análisis del presente trabajo se divide en dos partes:

- *Análisis descriptivo*
    - *Descripción del dataset a partir de reglas de asociación*. Se describe el *dataset* utilizado a partir de esta técnica; se muestran las métricas de las reglas para explicar el peso de cada una de ellas.
    - *Desaceleración del aumento de precios*. Se determinan los factores que se asocian con una disminución de los precios en el cuarto periodo.
    - *Factores asociados al comportamiento del grupo de alimentos recomendados*. Se estudian las reglas de asociación que describen el comportamiento de este grupo de alimentos.
- *Análisis predictivo*
    - *Predicción de precios*. Se comparan las reglas de los primeros tres periodos contra las reglas del cuarto para explicar el comportamiento de los mismos.
    - *Validación de resultados*. Se exponen las reglas de asociación que responden las preguntas de investigación planteadas en el trabajo anterior.

3. Materiales y métodos

Origen de los datos

Se utilizaron para este trabajo datos de precios, sucursales y productos del programa "Precios Claros". El proceso de relevamiento de precios fue generado de manera automática mediante la técnica de *web crawling*. Se muestran en la *Tabla A.1 del apéndice* los atributos de cada *dataset*.

Software utilizado

Los datos fueron cargados en el IDE (*Integrated Development Environment,* entorno de desarrollo integrado) *R-Studio*. Para el análisis de los mismos se utilizó el lenguaje *R*. Dentro de los paquetes más utilizados de este lenguaje se encuentran *Geojsonio*, para realizar gráficos a partir de datos de polígonos; *dplyr* para manipular datos; *ggplot2*, para graficar; *wesanderson* para las paletas de colores; *Arules* para generar las reglas de asociación; *ArulesViz* para visualización de las reglas y *TM* para minería de texto.

Preprocesamiento

Para la aplicación de Reglas de Asociación es necesario contar con variables discretas. Se detalla en la siguiente sección las transformaciones aplicadas a los datos.

Transformación de precios

Previo a la transformación de los precios se identificaron las mediciones con cada uno de los periodos, según lo mostrado en la *Tabla 1*; dicha información se agregó como variable al *dataset*. Luego, se pasaron las mediciones de precios -de productos por sucursal- a formato columnar, obteniendo así una fila por cada uno con 10 columnas asociadas, una para cada medición.

Al realizar esta transformación quedan en evidencia los datos faltantes en los precios. En el caso de faltantes en la primera medición se imputó el valor de la segunda. En los faltantes comprendidos entre las mediciones 2 y 9 se tomó la media de sus dos mediciones adyacentes. Finalmente, en el caso de faltantes de la última medición, se imputó el valor de la medición número 9. Los faltantes remanentes tras este procedimiento fueron eliminados.

Una vez completos todos los datos de precio se adicionaron cinco variables: el precio promedio en cada periodo y el precio promedio total. También, se calculó la variación intra-periodo y la variación total, calculados según la *Ecuación 1*. Luego, fueron discretizadas según la *Tabla 2.*

$$variación = \frac{precio_{nuevo} - precio_{inicial}}{precio_{inicial}} \qquad \textit{(Ecuación 1)}$$

Para tener un valor comparable de precios por producto en distintas sucursales se calculó la media de cada uno. Este valor fue utilizado para conocer el precio relativo, calculado según la *Ecuación 2* para cada periodo y para el total, adicionándose como variables al *dataset* de trabajo. Estos valores fueron discretizados según la *Tabla 2*.

$$precio\ relativo = \frac{precio_{producto\ en\ sucursal} - precio_{producto\ promedio}}{precio_{producto\ promedio}} \qquad \textit{(Ecuación 2)}$$

Tratamiento de coordenadas geográficas de las sucursales

Del portal de datos abiertos de CABA[1] se obtuvo información sobre los barrios de la ciudad.(*Tabla A.3 del apéndice*). A partir de estos datos geográficos y los datos de longitud y latitud de cada sucursal se obtuvo el barrio en el que se ubica cada una. Este dato fue adicionado como variable al *dataset* de trabajo.

[1] *data.buenosaires.gob.ar*

Tratamiento de textos de descripciones de productos

El conjunto de datos sobre productos contiene tres campos textuales que los describen: nombre, marca y presentación. Para extraer estas palabras descriptivas, dichas variables fueron convertidas a minúscula, se quitaron los dígitos numéricos, los símbolos de puntuación, las tildes y se borraron los espacios excedentes. Se obtuvo el listado de unidades y el listado de marcas para quitarlo del campo "nombre". Finalmente, se quitaron las palabras vacías de español.

Figura 1: *Nube de palabras frecuentes. El color y el tamaño representan la frecuencia de la palabra.*

Tras aplicar estas transformaciones fue posible separar en palabras para realizar conteos, se muestra en la *Figura 1* la nube de palabras coloreadas y dimensionadas según su frecuencia. Se armó un vocabulario con aquellas cuya frecuencia es mayor a 20, obteniéndose las 32 palabras más presentes en el *dataset*; se consideran estos términos buenos descriptores de los productos relevados, en particular de los alimentos, rubro de mayor interés en este estudio. Las mismas se detallan en la *Tabla A.4 del apéndice*. Por cada una de ellas se generó una columna en la cual se indica la presencia o ausencia de la misma en la fila (producto por sucursal).

Categorías de variación de precios	**Rango de variación de precios**
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 ; +∞)
Categorías de precios relativos	**Rango de precios relativos**
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 ; +∞)
Categoría de precios del metro cuadrado	**Rango de precios del metro cuadrad (USD)**
Barato	(-∞ ; 1000)
Medianamente barato	[1000 ; 2000)
Medio	[2000 ; 4000)
Medianamente caro	[4000 ; 5000)
Caro	[5000 ; 6000)
Muy caro	[6000 ; +∞)

Tabla 2: *categorías y rangos utilizados para la discretización de las variaciones de precios, los precios relativos y el precio del metro cuadrado de los barrios.*

Otras fuentes de datos

Durante el trabajo anterior fueron consultadas del portal de compras *CotoDigital*[2] las categorías de productos para clasificarlos en los grupos de alimentos a estudiar. De la página oficial de INDEC[3] se obtuvieron los datos de inflación para los periodos de interés (*Tabla 3*). Por último, se obtuvo de la página web *Properati*[4] el valor del metro cuadrado (en dólares) de cada barrio. Se muestran en la *Tabla A.3 del apéndice* los atributos de los *dataset* de fuentes externas consultadas.

Los datos de estas fuentes fueron adicionados al *dataset* de trabajo. El precio del metro cuadrado de cada barrio se discretizó según la *Tabla 2,* mientras que la inflación no fue integrada al *dataset*, sino que se mantuvo como fuente de consulta para las transformaciones detalladas en la sección *4.2.2.*

[2] *https://www.cotodigital3.com.ar/sitios/cdigi/*
[3] https://www.indec.gob.ar
[4] *https://www.properati.com.ar*

Otras transformaciones

Finalmente, se agregaron dos variables al *dataset* de trabajo: "canasta", según la categoría de productos se asignó el grupo de alimentos al que pertenece (*Tabla A.2 del apéndice*) y "ubicación", en la que se clasificó cada caso en "avenida" o "calle" según dónde se encuentra la sucursal; estos valores fueron asignados en base a la dirección.

Mes	Inflación mensual	Inflación de periodo
Noviembre 2018	3.4 %	14.9 %
Diciembre 2018	1.7 %	
Enero 2019	3.4 %	
Febrero 2019	5.7 %	

Tabla 3: *Inflación de alimentos en los meses estudiados. Se muestra también la inflación del periodo completo. Fuente: INDEC.*

Reglas de asociación

Antes de proceder con la búsqueda de conocimiento a partir de reglas de asociación, se realizó un breve estudio exploratorio de las mismas para decidir si todas las variables presentes en el *dataset* de trabajo aportaban información relevante.

Se observó que la variable "Razón Social" (de los comercios) daba lugar a reglas de asociación fuertemente vinculadas con "Bandera". De esta manera si permanecen ambas variables se obtienen reglas con ítems fuertemente asociados que no aportan nuevo conocimiento. Por un lado, se observa que la Razón Social *Coto* presenta una única bandera; dado que el 40% de los datos corresponden a esta cadena (*Figura 2*), quedaría el 60% de los datos repartido entre 14 banderas diferentes correspondientes a otras cadenas (*Tabla 4*). Por otro lado, la variable "Bandera" da indicios del tamaño de la sucursal y el tipo de oferta; además se observa que la distribución de precios podría tener relación con la bandera de las cadenas (*Sección 4.1.1*). Con el fin de no perder información relevante y evitar asociaciones erróneas, se utilizó una u otra variable según el objetivo de estudio. Se aclara en cada sección el objetivo del análisis y la variable que se conserva para realizarlo (ver *más adelante*).

De manera similar, se encontró que las reglas cuyos componentes eran "Categoría" y alguna de las palabras frecuentes tampoco brindaban información nueva. Asociaciones como "Categoría=Aguas" y "Término=Agua" no son relevantes. La variable "Categoría" fue excluida para la generación de reglas focalizando el estudio en "Canasta", variable que indica el grupo de alimentos al que pertenece el producto.

Una vez seleccionadas las variables de interés para el estudio se generaron las reglas de asociación. El *dataset* final a utilizar para aplicar esta técnica cuenta con 153.378 observaciones y 51 variables (*Tabla 4*). Considerando este número, se eligió un *soporte* de 0.01; este parámetro indica la proporción de transacciones que contienen a los ítems que componen la regla. El parámetro *confianza* elegido fue 0.25 para filtrar aquellas reglas que son “ciertas” un 25% de las veces. Se filtraron las reglas con *Lift* mayor a 1.25 o menor a 0.75, considerándose estos valores condición de dependencia. Todas las reglas fueron generadas con el *Algoritmo A priori*.

Marca (291)	variacion1 (7)	precio_rel2 (7)	Ubicación (2)
Sucursal (173)	variacion2 (7)	precio_rel3 (7)	Columnas de términos
sucursalTipo (3)	variacion3 (7)	precio_rel4 (7)	
banderaDescripcion (15)	variacion4 (7)	precio_rel_medio (7)	
comercioRazonSocial (10)	variacionT (7)	precio_m2 (6)	
Barrio (41)	precio_rel1 (7)	Canasta (3)	

Tabla 4: *Atributo del dataset final. Todas las variables son de tipo factor. Se muestra entre paréntesis la cantidad de niveles de cada uno.*

4. Resultados

4.1. Análisis descriptivo

Como parte del análisis descriptivo se generaron, como primer paso, reglas cuyo *itemset* consta de un solo elemento. De esta manera quedan en evidencia los elementos más frecuentes del *dataset*. Se muestran los 20 items cuya frecuencia es mayor al 25% en la *Figura 2*. Se observa que los items *{variacion2=mantiene}* y *{precio_m2=medio}* están presentes en más del 80% de los datos.

Posteriormente, se generaron 2 lotes de reglas (uno conservando "Bandera" y otro conservando "Razón social") sin límite de componentes de su *itemset* para analizar las asociaciones entre elementos. Se obtuvieron, en ambos casos, alrededor de 140.000 reglas, las cuales fueron filtradas para su análisis. Se generaron diversos *subsets* filtrando según los elementos presentes en el antecedente y en el consecuente, según el objetivo de estudio.

4.1.1. Descripción del dataset a partir de reglas de asociación

Es de interés poder describir el comportamiento de los precios en el periodo de estudio a partir de las características de los productos y las sucursales. Dada la segmentación de precios según la bandera (*Tabla 5*) para este análisis se trabajó conservando la variable "Bandera".

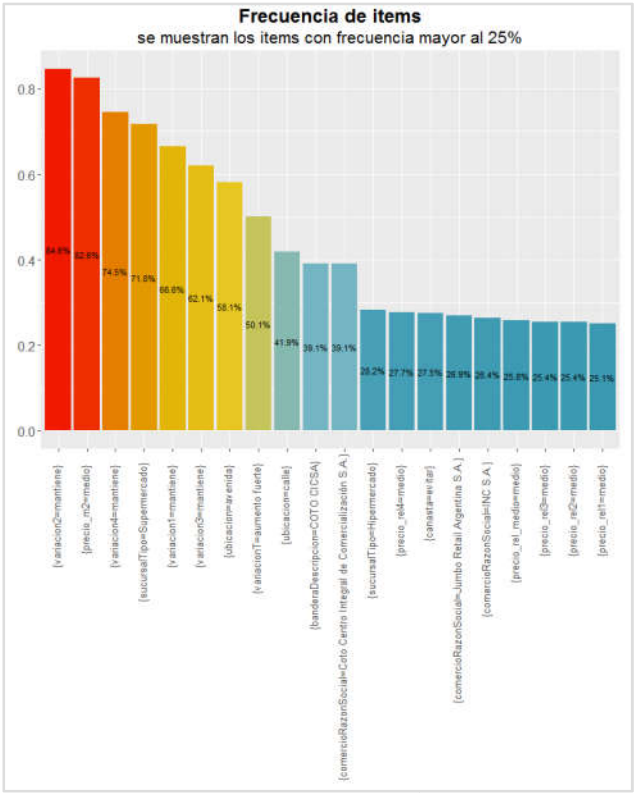

Figura 2: *frecuencia de items en el dataset de trabajo.*

De esta manera, se generaron aquellas reglas cuyo antecedente contiene ítems que describen a los productos (marca, termino, canasta) y a las sucursales (sucursal, bandera, barrio, ubicación, tipo), mientras que el consecuente sea un descriptor del precio (variación, precio relativo). Se excluyeron las reglas cuyo antecedente se compone de los items de frecuencia mayor al 80% (*Figura 2*). De los *subset* generados se seleccionaros 17 reglas para describir los datos (*Tabla 5*).

Observando las reglas de dos *ítems* se aprecia segmentación entre la bandera y el precio relativo medio. Las banderas de la empresa *Cencosud* (*Jumbo, Vea y Disco*) se asocian con precios relativos levemente y medianamente caros. En cambio la empresa *Carrefour* (*Carrefour Express, Carrefour Market, Hipermercado Carrefour y Supermercados Día*), a excepción de la bandera Express asociada con precios levemente caros, se asocian con precios relativos levemente baratos. La empresa *Coto*, que cuenta con una única bandera, se asocia con precios relativos medios.

En base a esta información se generan reglas de más *ítems*, filtrando aquellas en las que se asocia la bandera con la variación total del periodo. En este caso todas las banderas se asocian con una aumento fuerte a excepción de *Coto*.

En la *Figura 3* se observan sobre el mapa de la ciudad las sucursales relevadas. Se aprecia que las sucursales de la razón social *Jumbo* presentan precios relativos mayores a las de *INC S.A* y *Día S.A*. Las correspondientes a la razón social *Coto* parecieran ubicarse en el medio. Se observa que la mayor concentración de puntos de precio más alto pareciera ubicarse en zonas aledañas a las líneas de Subte B y C.

Independientemente de la bandera, se observa que aquellos productos que en el primer mes relevado contaban con precios relativos baratos sufrieron un aumento fuerte en el total del periodo estudiado.

Bandera	LHS	RHS	Soporte	Confianza	Lift	Cantidad
Jumbo	{banderaDescripcion=Jumbo}	{precio_rel_medio= levemente caro}	0.02	0.44	1.83	3653
Vea	{banderaDescripcion=Vea}	{precio_rel_medio= levemente caro}	0.02	0.34	1.43	3431
Carrefour Express	{banderaDescripcion=Express}	{precio_rel_medio= levemente caro}	0.02	0.39	1.63	2743
Disco	{banderaDescripcion=Disco}	{precio_rel_medio= medianamente caro}	0.04	0.28	3.07	6317
Coto	{banderaDescripcion=COTO CICSA}	{precio_rel_medio= medio}	0.16	0.40	1.54	23889
Día	{banderaDescripcion=Supermercados DIA}	{precio_rel_medio= levemente barato}	0.02	0.39	1.57	3685
Hiper Carre-four	{banderaDescripcion=Hipermercado Carrefour}	{precio_rel_medio= levemente barato}	0.02	0.43	1.76	3669
Carrefour Market	{banderaDescripcion=Market}	{precio_rel_medio= levemente barato}	0.08	0.47	1.89	11587
Jumbo	{banderaDescripcion=Jumbo, variacionT=aumento fuerte}	{precio_rel_medio= levemente caro}	0.01	0.46	1.92	1895
Vea	banderaDescripcion=Vea, variacionT=aumento fuerte}	{precio_rel_medio= levemente caro}	0.01	0.33	1.39	1695
Carrefour Express	No se encontraron reglas asociadas a la variación total del periodo					
Disco	{sucursalTipo=Supermercado, banderaDescripcion=Disco, variacionT=aumento fuerte}	{precio_rel_medio= levemente caro}	0.03	0.45	1.87	5215
Coto	{banderaDescripcion=COTO, ,variacionT=aumento leve}	{precio_rel_medio= medio}	0.01	0.47	1.84	1714
Día	{banderaDescripcion=Supermercados DIA, variacionT=aumento fuerte}	{precio_rel_medio= levemente barato}	0.02	0.42	1.70	2372
Hiper Carre-four	banderaDescripcion=Hipermercado Carrefour, variacionT=aumento fuerte}	{precio_rel_medio= levemente barato}	0.01	0.46	1.86	1796
Carrefour Market	{sucursalTipo=Supermercado, banderaDescripcion=Market, variacionT=aumento fuerte}	{precio_rel_medio= levemente barato}	0.03	0.51	2.06	5288
	{precio_rel1=medianamente barato}	{variacionT=aumento fuerte}	0.06	0.62	1.23	8417
	{precio_rel1=levemente barato}	{variacionT=aumento fuerte}	0.14	0.61	1.22	20997

Tabla 5: *Reglas de asociación seleccionadas para describir el dataset de trabajo.*

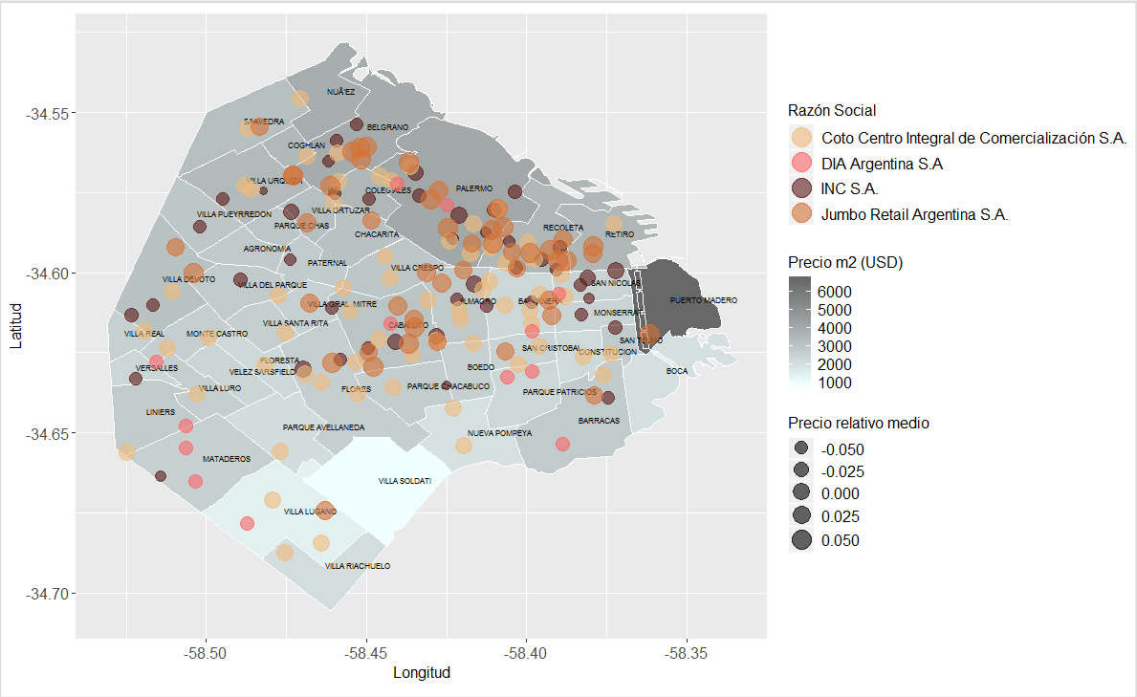

Figura 3: *Mapa de C.A.B.A. con las sucursales relevadas y su precio relativo medio*

4.1.2. Desaceleración del aumento de precios

En este caso se busca encontrar los factores asociados a la desaceleración de precios. Se conservó para este análisis la variable "Razón Social", dado que conservando "Bandera" sólo se obtenían reglas de *Coto* y por ende de una única cadena.

Se filtraron aquellas reglas cuyo consecuente corresponde al detenimiento en el aumento de precios en el cuarto periodo; esto es: mantiene, disminución leve, disminución media o disminución fuerte. En la mayoría de este *subset* el consecuente fue *{variacion4=mantiene},* ítem que está presente en aproximadamente el 75% de los datos (*Figura 2*).

Las reglas de asociación seleccionadas se detallan en la *Tabla A.5 del apéndice* y se esquematizan en la *Figura 4*. La cadena de supermercados *INC S.A* (*Carrefour*) es la única asociada a una disminución fuerte en los precios durante el cuarto periodo; esta variación ocurre sobre los productos de precio relativo medio muy barato. Esta regla es la que presenta mayor asociación de todas las presentadas en esta sección (*Lift* = 5.60). La misma regla tiene como consecuente "mantiene", pero con un valor de *Lift* por debajo de 1 (0.44), indicando que este *itemset* no aumentan la probabilidad de mantenerse el precio sino que la disminuye.

La cadena de supermercados *Coto* se asocia en todos los casos con una invariancia en sus precios durante el periodo. Se observa que estos comercios mantuvieron los precios en productos incluidos en el grupo de alimentos de consumo moderado, observándose el caso particular de los fideos. El mismo comportamiento se asocia con la cadena *Jumbo*, pero en este caso con las galletitas (comprendido entre los alimentos a evitar su consumo) y con la canasta de alimentos recomendados. En los casos de esta cadena el *lift* es menor a uno.

Figura 4: *Esquema representativo de las reglas de asociación seleccionadas para explicar la desaceleración de precios. Cada círculo de color indica una regla.*

4.1.3. Factores asociados al comportamiento del grupo de alimentos recomendados

En línea con el trabajo anterior, es de particular interés conocer el comportamiento del grupo de alimentos recomendados para el consumo. En este caso se busca mediante la técnica de reglas de asociación encontrar los factores que influyen en el comportamiento de precios de esa canasta. Las reglas presentadas en esta sección fueron filtradas del lote que conserva la variable "Razón Social".

En este caso se obtuvieron pocas reglas representativas del comportamiento, destacando las mostradas en la *Tabla 6*. Se observa que la cadena *Jumbo* se asocia con un precio relativo medio levemente caro, mientras que la cadena *Coto* con un precio de valor medio. La variación de precio de esta cadena durante todo el periodo de estudio es de aumento. Esto es consistente con lo enunciado en la *Sección 4.1.1,* pudiéndose especificar esta regla para la canasta en estudio. En cuanto a la ubicación de la sucursal, no se observa diferencia entre calle y avenida.

LHS	RHS	Soporte	Confianza	Lift	Cantidad
{comercioRazonSocial=Coto, canasta=recomendado}	{precio_rel_medio=medio}	0.02	0.59	2.30	2928
{comercioRazonSocial=Jumbo, canasta=recomendado}	{precio_rel_medio=levemente caro}	0.01	0.46	1.91	1551
{comercioRazonSocial=Coto, canasta=recomendado}	{variacionT=aumento medio}	0.01	0.38	1.92	1897
{canasta=recomendado, ubicacion=avenida}	{precio_rel_medio=medio}	0.02	0.34	1.33	2491
{canasta=recomendado, ubicacion=calle}	{precio_rel_medio=medio}	0.01	0.39	1.51	2061

Tabla 6: *Reglas de asociación seleccionadas para el análisis del comportamiento de precios del grupo de alimentos recomendados.*

4.2. Análisis predictivo

En esta sección se busca evaluar la capacidad predictiva de la técnica en uso. Para ello se intentan verificar las reglas obtenidas durante los primeros tres periodos con las reglas correspondientes al cuarto. También se busca comprobar los resultados obtenidos en el trabajo anterior (1).

4.2.1. Predicción de precios

Se seleccionaron aquellas reglas correspondientes a los tres primeros periodos cuyo consecuente sea alguna de las variables relacionadas al comportamiento del precio (precio relativo y variación). Luego, se buscaron las mismas reglas durante el periodo 4. Se seleccionaron 10 de ellas relacionadas a cada grupo de alimentos en estudio, priorizando aquellas que permitan mayor generalización.

Se muestran dichas reglas en la *Figura 5*. Sobre el grupo de alimentos recomendados para el consumo se observa que en la cadena *Coto* el precio relativo medio se mantuvo durante los cuatro periodos, contando con los valores de *Lift* más altos de este muestreo. Para las sucursales ubicadas en calle el comportamiento fue similar, exceptuando durante el periodo 2 que presentan un precio levemente mayor.

El grupo de alimentos de consumo moderado son los más presentes en las reglas; se observa que la cadena *Jumbo* presenta un rango de precios levemente caro para esta canasta durante los cuatro periodos, contando con valores de *Lift* por encima del mínimo exigido (1.25). En cambio, la cadena *Coto* presenta precios de valor medio y constancia en el precio. La cadena *Carrefour* presenta precios levemente baratos. Los alimentos comprendidos en este grupo que se distribuyen en locales ubicados sobre una calle no presentan esta constancia, observando un aumento durante el periodo 4.

Los productos del grupo de alimentos a evitar el consumo presentan precios levemente caros en la cadena *Jumbo*; esta regla parece ser generalizable a todas las sucursales ubicadas en calle.

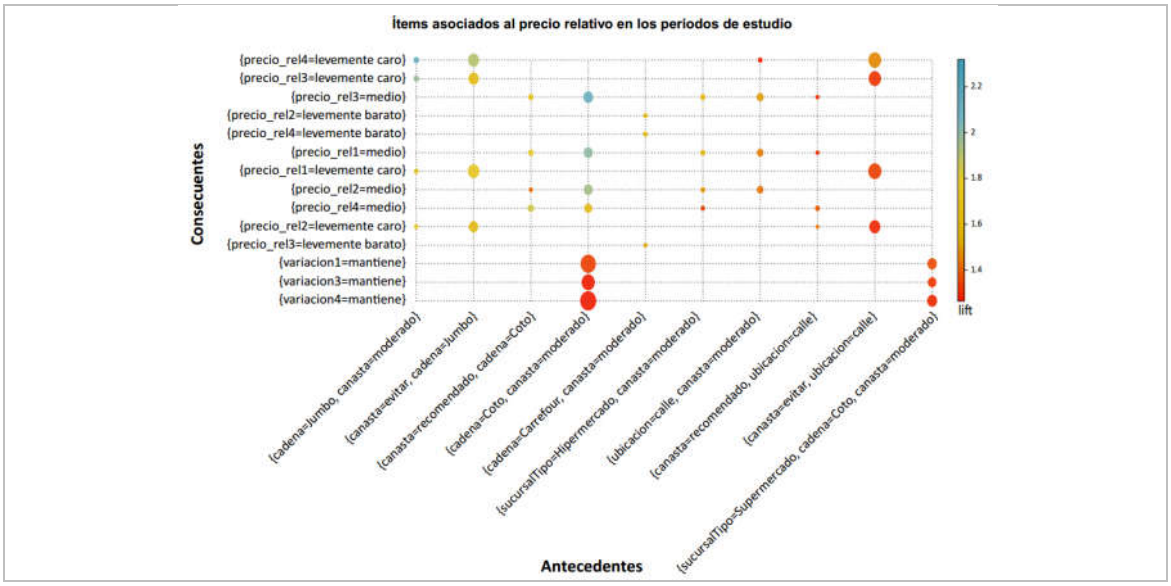

Figura 5: *Matriz de relación entre antecedentes y consecuentes. La escala de color indica el valor del lift, el tamaño indica el soporte.*

4.2.2. Validación de resultados obtenidos

Durante el estudio anterior (1) se observaron tendencias y comportamiento de precios en relación a ciertos productos y grupos de alimentos. Se busca en esta sección obtener los mismos resultados a través de la técnica de reglas de asociación.

Aumento de precios comparado a la inflación del periodo

Se observó previamente que los productos comprendidos en la categoría "Agua" y "Leche" presentaron un aumento total mayor a la inflación informada por entes oficiales. Quitando estos productos, el aumento general intra periodo fue por debajo de la inflación.

Para realizar una comparación con respecto a la inflación se adicionaron columnas al *dataset* de trabajo indicando si la variación de precios en cada periodo se ubica por encima o por debajo de la inflación, según lo indicado en la *Tabla 3*. Lo mismo se realizó para la variación total. Se generaron las reglas de asociación considerando estas variables en lugar de las variaciones de precio.

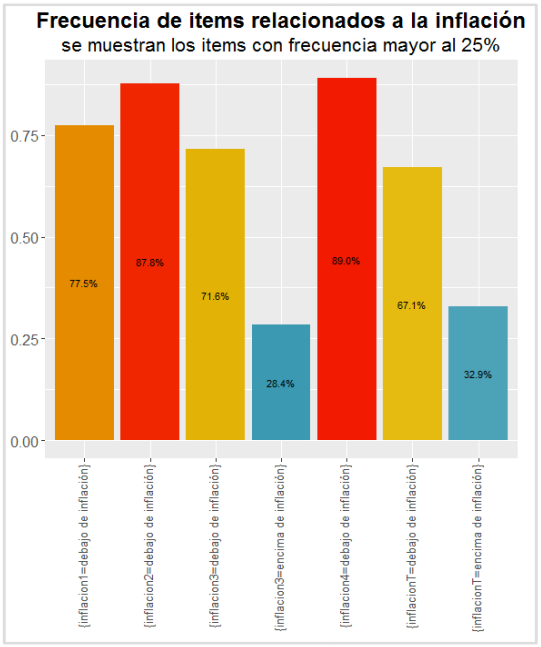

Figura 6: *Frecuencia de ítems relacionados a la inflación.*

Al observar la frecuencia de los ítems (*Figura 6*) se observa que para los cuatro periodos el elemento que indica "Debajo de la inflación" presenta una frecuencia mayor al 70%. Esto es consistente con lo observado en el trabajo anterior. Sin embargo, para el caso particular de Agua y Leche, se obtienen verificaciones parciales. Se observa en la *Figura 7* que los precios por encima de la inflación se asocian a estos términos sólo en determinados periodos (periodo 3 para Agua, periodo 2 para leche, en los que efectivamente el aumento fue superior a la inflación (1)), pero no en la totalidad del periodo de estudio. Las reglas y sus métricas se encuentran detalladas en la *Tabla A.6 del apéndice*.

Aumento y precio medio de grupos de alimentos

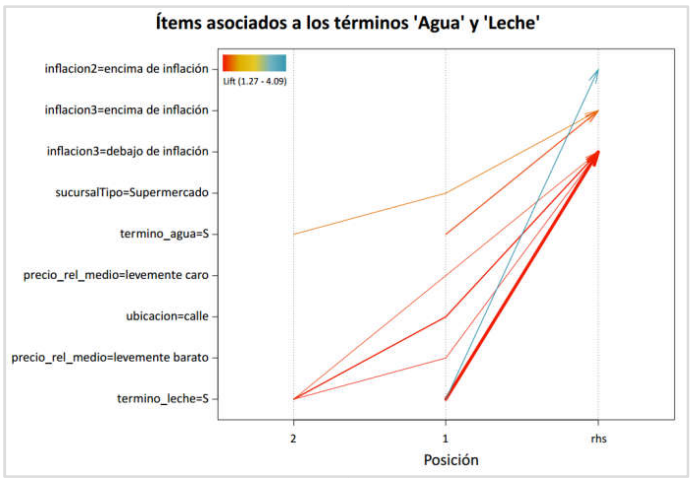

Figura 7: *Reglas asociadas a los términos "Agua" y "Leche" e inflación. El espesor de la línea indica el soporte (rango: 0.01-0.05) y el color el Lift.*

Otra de las observaciones realizadas en el estudio anterior fue que la canasta de alimentos de consumo moderado presentaba menor aumento en el total del periodo y el precio medio mayor. A diferencia, la canas de alimentos a evitar el consumo presenta los precios más baratos. Se seleccionaros las reglas más generales obtenidas para verificar estos hechos.

Se observa en la *Figura 8* que el grupo de alimentos de consumo moderado (Antecedentes 1, 2 y 3) presenta asociación con una invariancia en su precio; si bien esto no brinda información sobre la magnitud del aumento, al compararla con las otras dos canastas queda en evidencia: la canasta de alimentos a evitar el consumo (Antecedentes 8, 9, 10 y 11) se asocia con un aumento fuerte, mientras que la de productos recomendados (Antecedentes 4, 6 y 7) con un aumento medio. Se puede inferir, de esta manera, que la canasta de alimentos de consumos moderado fue la que menos aumentó en el total del periodo.

Para la canasta de productos a evitar el consumo no se verifica que sus precios sean los más baratos, sino al contrario, aquellos productos comercializados en supermercados se asocian con precios levemente caros (Antecedente 10). Las reglas representadas en la figura pueden encontrarse en la *Tabla A.7 del apéndice.*

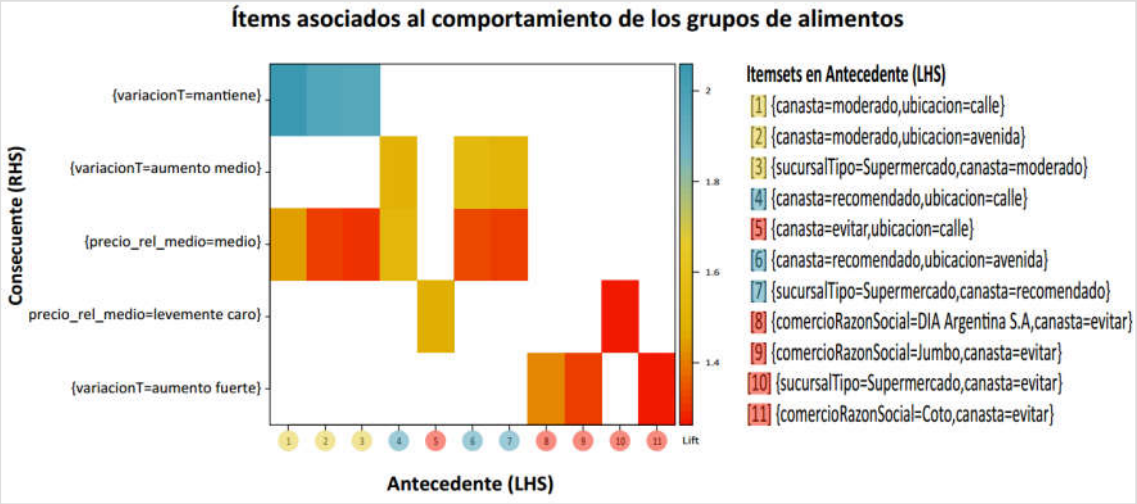

Figura 8: *Reglas de asociación seleccionadas para el análisis del comportamiento de precios de las canastas de alimentos de consumo moderado y a evitar el consumo*

5. Discusión

Tras los estudios y resultados presentados en este informe se pudieron realizar nuevas observaciones y ampliar el conocimiento descubierto previamente gracias a los análisis descriptivos y predictivos realizados con la técnica de reglas de asociación.

Durante el análisis descriptivo fue posible segmentar los datos por su relación Bandera-Precio relativo medio. Se observó que las sucursales de la empresa *Cencosud* son las que presentan un precio levemente caro mientras que las de la empresa *Carrefour* se asocian con un precio levemente barato. La empresa *Coto*, con su única bandera, se asocia a un precio medio. Esto podría explicarse por la cantidad de empleados por metro cuadrado (*Tabla 7*) de cada una. Las tres cadenas manejan cantidad de proveedores similar (3) con lo cual el número de empleados por unidad de área pasa a ser de importancia en el costo de comercialización. Así, a menor cantidad de empleados por metro cuadrado es posible mantener el margen de ganancia con precios de venta al público menores. Cabe mencionar, que un mayor número de empleados se entiende como una mejor atención al público; esto podría validar la estrategia de *Cencosud* y la continuidad en la demanda de productos a pesar de los precios.

Lo enunciado anteriormente queda en evidencia en la Figura *3* en la que se aprecia que cada empresa se asocia a un rango de precios. Además se observa cierta correspondencia entre estos precios y la cercanía con las líneas de subte. Queda para futuras investigaciones ahondar en este tema relacionando el precio medio de cada sucursal con la distancia a la estación de subte más cercana.

Empresa	Razón social	Bandera	Bocas	Superficie	Empleados	Empleados/ metros	Facturación anual/ empleados
Carrefour	INC S.A.	Express Market Hiper	200	571,473 m²	18,750	1 empleado cada 43.8 m²	$ 852,970
	Dia S.A	Dia	400	250,200 m²			
Cencosud	Jumbo	Vea	121	184,011 m²	23,018	1 empleado cada 22.7 m²	$ 421,409
		Jumbo	16	139,719 m²			
		Disco	143	198,862 m²			
Coto	Coto CICSA	Coto	113	476,500 m²	13,353	1 empleado cada 35.7 m²	$ 404,312

Tabla 7: *Cantidad de bocas, superficie, empleados y facturación de cada empresa. Fuente: Federación Argentina de Empleados de Comercio y Afines.*

La asociación observada entre producto de precio relativo barato en el primer periodo y el aumento total fuerte se podría explicar como un atraso en la actualización de los precios en Noviembre 2018, en comparación a Febrero 2019.

Se buscaron los factores que influyen en la desaceleración del aumento de precios en el cuarto periodo, obteniendo la mayor asociación entre los productos de precio relativo barato comercializados en la cadena *INC S.A*. (*Carrefour*) y una disminución fuerte de precios. Además de que dicha cadena se ubica dentro de las más baratas *(Sección 4.1.1*), este hecho podría estar ligado a una fuerte campaña de precios llamada "*Precios Corajudos*" en la que prometen mantener los precios de Abril 2018 hasta Diciembre 2019; la misma campaña fue relanzada en enero, fijando los precios en diciembre 2019 para mantenerlos hasta marzo del mismo año (4). La cadena Coto se asocia con una invariancia en los precios para el periodo en productos del grupo de alimentos recomendados. Esto podría deberse a que los productos de esta canasta son los de mayor demanda en el mercado minorista (5).

Fue de particular interés en los dos trabajos presentados el comportamiento de los precios del grupo de alimentos recomendados para el consumo. En línea con lo observado al describir el *dataset* de trabajo, la cadena *Coto* presenta precios moderados mientras que *Jumbo* precios levemente más caros. La ubicación de la sucursal no afecta apreciablemente los precios de esta canasta, siendo indistinto los ubicados en calle o en avenida.

En cuanto a la capacidad predictiva de esta técnica, podría afirmarse que funciona correctamente con predicciones moderadamente específicas, como es evaluar el comportamiento de cierta canasta en determinada cadena. Si se intenta generalizar el comportamiento de precios para, por ejemplo, supermercados ubicados sobre calle o avenida, el comportamiento no es constante como para generalizar. Sería adecuado para sustentar esta afirmación repetir el estudio con mas periodos.

Por último, se buscó validar los resultados obtenidos en el estudio anterior. Los postulados propuestos fueron validados parcialmente, no pudiendo obtener la generalización buscada y obteniendo a cambio resultados más específicos.

6. Conclusiones

Mediante la técnica estudiada pudieron ordenarse las tres empresas de supermercados principales según su asociación al precio medio relativo: *Cencosud* (*Jumbo, Disco* y *Vea*) como la más cara, Coto con precios moderados y *Carrefour* (Hipermercado, *Market* y *Supermercados Día*) como la más barata. Con respecto al precio relativo de la canasta de alimentos recomendados para el consumo, se asocia de igual manera a los supermercados ubicados en calles y en avenidas. Finalmente, se observaron mejores resultados predictivos para supuestos específicos que para generales.

Se considera que con la técnica de reglas de asociación pudo describirse en detalle el *dataset* de trabajo, obteniendo conocimiento específico que no se descubrió con las técnicas exploratorias del anterior estudio. Creemos que es una técnica potente y de gran utilidad para encontrar información específica que no se aprecia en una visión general de los datos.

7. Referencias

1. ***Federico Moreno, Bianca Picchetti.*** Estudio de precios en la Ciudad Autónoma de Buenos Aires en el periodo Noviembre 2018 a Febrero 2019. *s.l. : Minería de Datos, 2019.*

2 **Manual para la aplicación de las guías alimentarias para la población Argentina.** *Ciudad Autónoma de Buenos Aires. Dirección de Promoción de La Salud y Control de Enfermedades No Transmisibles. 2018978-950-38-0267-0*

3. ***Federación Argentina de Empleados de Comercio y Afines.*** Relevamiento sobre supermercados en Argentina. *2019.*

4. Ámbito Financiero. [En línea] 03 de 01 de 2019. [Citado el: 08 de 07 de 2019.] https://www.ambito.com/carrefour-relanza-su-plan-precios-corajudos-tres-meses-n5008475.

5. ***INDEC.*** Encuesta de supermercados y autoservicios mayoristas. *2019.*

Apéndice

Productos	Sucursales	Precios
- Id de relevamiento	-Id de relevamiento	- id de relevamiento
- Nombre (*char*)	- SucursalTipo(*char*)	**- id de producto**
- Marca (*char*)	- Provincia(*char*)	**- id de sucursal**
- Presentación (*char*)	- BanderaId (*cat*)	- Precio (*num*)
- Id de producto	- Localidad(*char*)	- Fecha (*date*)
	- Latitud(*num*)	- Medición (*cat*)
	- ComercioRazonSocial(*char*)	
	- Longitud (*num*)	
	- SucursalNombre(*char*)	
	- CcomercioId (*cat*)	
	- SucursalId (*cat*)	
	- id de sucursal	

Tabla A.1: *Atributos de los dataset utilizados. Se muestra el tipo de variable (char: texto, num: numérica, cat: categórica; date: formato fecha y hora) y a través de cuál de ellas se relacionan entre ellas (negrita) (Sección 3).*

Grupo 1: Recomendados	Grupo 2: Moderado	Grupo 3: Evitar
- Agua	- Conservas	- Gaseosas
- Infusiones de hierba	- Quesos	- Jugos en polvo
- Leche	- Pastas	- Bebidas deportivas
- Leche en polvo	- Yogurt	- Carnes congeladas
- Arroz	- Aceite de oliva	- Otros congelados
- Harinas de trigo y maíz	- Aceites (otros)	- Salsas a base de tomate
- Verduras congeladas	- Manteca	- Cereales azucarados
	- Endulzantes	- Galletitas y panificados

Tabla A.2: *Categorías de productos comprendidos en cada grupo de alimentos estudiado (Sección 3).:*

Barrios	Inflación	Precio m^2
-WKT (*polígono*)	- Medición (*cat*)	-Barrio (*char*)
-Barrio (*char*)	- Periodo (*char*)	- Precio m^2 USD (*num*)
-Comuna (*cat*)	- Nacional (*num*)	
-Área (*num*)	- Nacional del periodo (*num*)	
Perímetro (*num*)	- Alimentos (*num*)	
	- Alimentos del periodo (*num*)	
	- Alcohol (*num*)	
	- Alcohol del periodo (*num*)	

Tabla A.3: *Atributos de los datasets obtenidos de fuentes externas (Sección 3).*

Palabra	*Frecuencia*
vino	*72*
galletitas	*69*
polvo	*52*
pack	*50*
leche	*49*
tinto	*49*
vainilla	*45*
jugo	*45*
agua	*43*
light	*42*
chocolate	*41*
naranja	*39*
gas	*38*
doypack	*32*
aerosol	*31*
crema	*29*
yogur	*27*
malbec	*25*
fideos	*25*
cafe	*24*
saborizada	*24*
queso	*24*
manzana	*23*
desodorante	*23*
dulce	*22*

limon	*22*
gaseosa	*22*
liquido	*22*
jabon	*22*
mate	*21*
blanco	*20*
frutilla	*20*

Tabla A.4: *Palabras más frecuentes (Sección 3).*

LHS	RHS	Soporte	Confianza	Lift	Cantidad
{comercioRazonSocial=Coto, termino_fideos=S}	{variacion4=mantiene}	0.01	1.00	1.34	1579
{comercioRazonSocial=INC S.A., precio_rel_medio=muy barato}	{variacion4=disminucion fuerte}	0.01	0.27	5.60	1602
{comercioRazonSocial=INC S.A., precio_rel_medio=muy barato}	{variacion4=mantiene}	0.01	0.33	0.44	1945
{comercioRazonSocial=Coto, canasta=moderado}	{variacion4=mantiene}	0.04	0.95	1.27	6799
{comercioRazonSocial=Jumbo, canasta=recomendado}	{variacion4=mantiene}	0.01	0.52	0.69	1757
{comercioRazonSocial=Jumbo, termino_galletitas=S}	{variacion4=mantiene}	0.01	0.53	0.71	1565

Tabla A.5: *Reglas seleccionadas para el análisis de la desaceleración de precios (Sección 4.1.2)*

LHS	RHS	Soporte	Confianza	Lift	Cantidad
{termino_agua=S}	{inflacion3=encima de inflación}	0.02	0.38	1.35	2490
{sucursalTipo=Supermercado, termino_agua=S}	{inflacion3=encima de inflación}	0.01	0.41	1.44	1893
{termino_leche=S}	{inflacion2=encima de inflación}	0.01	0.29	2.40	2272
{termino_leche=S}	{inflacion3=debajo de inflación}	0.05	0.91	1.27	7073
{precio_rel_medio=levemente caro, termino_leche=S}	{inflacion3=debajo de inflación}	0.01	0.95	1.33	1597
{precio_rel_medio=levemente barato, termino_leche=S}	{inflacion3=debajo de inflación}	0.01	0.91	1.28	1687
{ubicacion=calle, termino_leche=S}	{inflacion3=debajo de inflación}	0.02	0.92	1.29	2970

Tabla A.6: *Reglas de asociación seleccionadas para verificar observaciones sobre el precio del agua y de la leche del trabajo anterior (Sección 4.2.2).*

LHS	RHS	Soporte	Confianza	Lift	Cantidad
{canasta=recomendado, ubicacion=calle}	{precio_rel_medio=medio}	0.01	0.39	1.51	2061
{canasta=recomendado, ubicacion=avenida}	{precio_rel_medio=medio}	0.02	0.34	1.33	2491
{sucursalTipo=Supermercado, canasta=recomendado}	{precio_rel_medio=medio}	0.02	0.34	1.30	3022
{canasta=moderado, ubicacion=calle}	{precio_rel_medio=medio}	0.02	0.37	1.44	2822
{canasta=moderado, ubicacion=avenida}	{precio_rel_medio=medio}	0.02	0.34	1.31	3588
{sucursalTipo=Supermercado, canasta=moderado}	{precio_rel_medio=medio}	0.03	0.34	1.30	4401
{sucursalTipo=Supermercado, canasta=evitar}	{precio_rel_medio=levemente caro}	0.06	0.30	1.26	9106
{canasta=evitar, ubicacion=calle}	{precio_rel_medio=levemente caro}	0.04	0.35	1.46	6154
{canasta=recomendado, ubicacion=calle}	{variacionT=aumento medio}	0.01	0.30	1.49	1572
{canasta=recomendado, ubicacion=avenida}	{variacionT=aumento medio}	0.01	0.31	1.53	2233
{sucursalTipo=Supermercado, canasta=recomendado}	{variacionT=aumento medio}	0.02	0.30	1.50	2708
{canasta=moderado, ubicacion=calle}	{variacionT=mantiene}	0.02	0.32	2.05	2408
{canasta=moderado, ubicacion=avenida}	{variacionT=mantiene}	0.02	0.30	1.96	3201
{sucursalTipo=Supermercado, canasta=moderado}	{variacionT=mantiene}	0.03	0.30	1.94	3921
{comercioRazonSocial=Jumbo, canasta=evitar}	{variacionT=aumento fuerte}	0.05	0.65	1.31	7276
{comercioRazonSocial=DIA, canasta=evitar}	{variacionT=aumento fuerte}	0.01	0.70	1.40	1894
{comercioRazonSocial=Coto, canasta=evitar}	{variacionT=aumento fuerte}	0.07	0.63	1.25	10384

Tabla A.7: *Reglas de asociación seleccionadas para verificar observaciones sobre el precio de las canastas (Sección 4.2.2).*