**CURSO: MINERÍA DE DATOS**

**MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO**

TRABAJO PRÁCTICO ENTREGABLE 2

Reglas de Asociación

# INTRODUCCIÓN

En este Trabajo Práctico, se incorpora la técnica de reglas de asociación para abordar un problema puntual y tratar de entenderlo a partir observaciones de coocurrencia de factores.

Para la exploración de este tema, se utilizará el IDE R-Studio del lenguaje de programación R con el objetivo de ejercitar los conceptos abordados en las clases teóricas.

El dominio del problema a analizar es el comportamiento de precios temporal de productos de supermercados e hipermercados de la ciudad de Buenos Aires. Para esto se utilizará el mismo conjunto de datos provisto para el TP1.

# OBJETIVO GENERAL

El objetivo general de este trabajo es aplicar la técnica de reglas de asociación sobre los datos de precios de productos de consumo masivo, 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 TP1.

# PREPROCESAMIENTO

Antes de aplicar reglas de asociación sobre el conjunto de datos se plantean las siguientes pautas de preprocesamiento y transformación de variables. Considere estas pautas como una línea de base e incorpore otras transformaciones en el caso de ser necesario.

## Tratamiento de precios

Para aplicar reglas de asociación es necesario discretizar las variables numéricas. A partir del conjunto de datos **precios.json** se proponen un conjunto de transformaciones para extraer características relacionadas con la competitividad de los precios en diferentes períodos, y la variación de precios entre estos.

Se proponen cuatro períodos derivados de los intervalos de mediciones realizados en el relevamiento de precios.

| 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*

## Pautas de transformación de precios

1. Pasar las mediciones de precios de productos por sucursal a formato columnar. En este formato cada fila representará un producto de una sucursal con 10 columnas asociadas a los precios en cada medición[1].
2. Luego de aplicar el paso anterior, algunos valores de precios contendrán faltantes. Se propone imputar los valores tomando, por ejemplo, el promedio de las mediciones inmediatamente contiguas a la medición del faltante. Este procedimiento mantendría faltantes en caso que ambas mediciones aledañas contengan faltantes[2].
3. Generar cuatro nuevas columnas de precios con los precios promedios de cada período descripto en la **Tabla 1**. Además se podría generar una columna con el promedio total de todas las mediciones.
4. Eliminar a los casos que contengan faltantes en las nuevas columnas de precios calculadas.
5. Calcular variaciones porcentuales intra-período y la variación total entre el precio del primer y último período. Se tomará como variación de precio a la siguiente fórmula:

$$variación = \frac{precio_{nuevo} - precio_{inicial}}{precio_{inicial}}$$

[1] Inspeccione el paquete *reshape*, o las función *dcast()*, por ejemplo.
[2] Verifique la función *rowMeans.*

6. Discretizar las variaciones calculadas en el paso anterior, por ejemplo según los rangos propuestos en la **Tabla 2**[3].

| 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 | $[0.1; \infty]$ |

***Tabla 2:*** *Ejemplo de discretización de variaciones*

7. Para tener un valor comparable de precios por producto en distintas sucursales se calculará la media por productos en cada una de las variables de precios generadas en el paso 3[4].

8. Utilizando los promedios calculados en el paso anterior se deberá calcular el precio relativo sobre las variables de precios generadas en el paso 3[5]. El precio relativo se calculará siguiendo la siguiente fórmula:

$$precio\ relativo\ = \frac{precio_{\ producto\ en\ sucursal} - precio_{producto\ promedio}}{precio_{producto\ promedio}}$$

Discretizar las variaciones calculadas en el paso anterior según los rangos propuestos en la siguiente tabla:

[3] Puede verificar la función *discretize* tanto del paquete *arules* como de *infotheo.*
[4] Verifique el paquete *dplyr: group_by, summarize.*
[5] Puede ayudarse con la función *inner_join* de *dplyr.*

| 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 | $[-\infty;\ 0.1)$ |

***Tabla 3:*** *Intervalos propuestos para discretización de precios*

En resumen, luego de seguir las pautas anteriores, el conjunto de datos de precios quedaría conformado por las columnas descritas en la Tabla 4:

| Variables |
|---|
| Producto |
| Sucursal |
| Precio relativo período 1 discretizado |
| Precio relativo período 2 discretizado |
| Precio relativo período 3 discretizado |
| Precio relativo período 4 discretizado |
| Precio relativo medio discretizado |
| Primera variación de precio discretizada |
| Segunda variación de precio discretizada |
| Tercera variación de precio discretizada |
| Variación total de precio discretizada |

***Tabla 4:*** *Atributos del conjunto de datos para la aplicación de Reglas de Asociación*

## Tratamiento de coordenadas geográficas

Utilizando las coordenadas geográficas presentes en el conjunto de datos **sucursales.json** se podría obtener el barrio en donde se sitúa cada sucursal. Para esto deberá integrarse los datos de sucursales con datos externos con las delimitaciones geográficas de los barrios de Buenos Aires[6].

## Tratamiento de textos de descripciones de productos

El conjunto de datos **productos.json** contiene 3 campos textuales que describen al producto: el nombre, la marca y la presentación. A partir de estos campos se deberán extraer palabras que describan a los productos. A continuación se proponen una serie de pautas para realizar esta extracción de características.

### Extracción de términos por productos

1. Generar una copia de las columnas nombre, presentación y marca para realizar este proceso.
2. Convertir las tres columnas a minúsculas[7].
3. Quitar dígitos numéricos[8]
4. Quitar símbolos de puntuación[9].
5. Quitar tildes a las vocales[10].
6. Borrar espacios en campos marca y presentación[11]
7. Obtener el listado de unidades de presentación sin repetidos. Utilizar este listado para quitar unidades de presentación en el campo nombre[12].
8. Repetir el paso anterior, pero para marcas. De esta forma el campo nombre no contendrá marcas.
9. Eliminar palabras vacías en español[13] (preposiciones, artículos, etc).
10. Finalmente luego de aplicar todas estas transformaciones sobre el campo de nombres, separar en palabras y realizar conteos. Formar un vocabulario de palabras

---

[6] Verifique el sitio data.buenosaires.gob.ar y alguno de los paquetes de R para mapas y geolocalización como por ejemplo *leaflet, ggmap, etc.*
[7] Investigue la función *tolower.*
[8] Sugerencia: *tm::removeNumbers.*
[9] Sugerencia: *tm::removePunctuation*.
[10] Sugerencia: *stringi::stri_trans_general(texto,"Latin-ASCII")*.
[11] Sugerencia: *tm::stripWhitespace*.
[12] Sugerencia: *unique() tm::removeWords.*
[13] Sugerencia: *tm::stopwords*.

con aquellas palabras que tengan una mínima frecuencia en los textos de los productos[14].

11. Por cada palabra de vocabulario seleccionada, generar una columna de presencia ausencia. En el caso de ausencia marcar como NA, y en el caso de presencia algún caracter (Ej 'S'). Es recomendable para luego aplicar reglas de asociación, utilizar un prefijo en cada una de de estas columnas seguido por la palabra en cuestión[15] (Ej: termino_yerba, termino_leche).

### Otras fuentes de datos y transformaciones

Considere las transformaciones y fuentes adicionales de datos para generar nuevas características que sean de utilidad para la extracción de reglas de asociación.

## Reglas de asociación

Una vez que cuenta con el dataset preprocesado, avance sobre la obtención de reglas que le permitan concluir respecto de las siguientes consignas:

**Consignas:**

- *Análisis descriptivo*: Escoja entre 10 y 20 reglas que le permitan describir el conjunto de datos.
    - Explique qué aportes se obtienen a partir de las mismas y explique cuál es el peso de cada una de acuerdo a las métricas asociadas.
    - Genere reglas de decisión que expliquen los factores más asociados a la desaceleración del aumento de precios en las últimas mediciones. Explique cuáles son las reglas más robustas, justifique y explique el conocimiento que estas aportan.
    - Escoja un lote de productos de su interés y explique qué factores asociados a su comportamiento obtiene a partir de reglas de asociación.
- *Análisis predictivo a partir de las reglas*.

[14] Sugerencia: *findFreqTerms* o *apriori*.
[15] Sugerencia: *stringr::str_detect*.

- Seleccione 10 reglas sobre los precios de los primeros 3 períodos, y evalúelas utilizando métricas sobre los precios del último período.
- En el caso que corresponda: ¿Puede validar los resultados obtenidos en el TP1 a partir de reglas? En el caso afirmativo incorpore algunos ejemplos.

- Documente las iteraciones realizadas entre preprocesamiento y la construcción de las reglas y haga referencia a cuáles son las transformaciones con las que obtuvo una mejor configuración.
- Complemente su análisis utilizando herramientas de visualización que contribuyan al entendimiento de las reglas generadas[16].

[16] Investigue el paquete *arulesViz.*