CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

TRABAJO PRÁCTICO ENTREGABLE I

Preprocesamiento de datos, integración y gestión de datos mediante una DB NOSQL

INTRODUCCIÓN

En este primer Trabajo Práctico entregable del curso, se integrarán los conocimientos relacionados con el preprocesamiento de datos, integración, tratamiento de datos faltantes, construcción de variables y gestión de datos mediante una Base de Datos NoSQL orientada a documentos.

Para la exploración de estos temas, se utilizará el IDE R-Studio del lenguaje de programación R y la Base de Datos MongoDB con el objetivo de ejercitar los conceptos abordados en las clases teóricas.

El conjunto de datos consiste en un relevamiento reciente por medio de crawling de Precios Claros[1]. Se seleccionaron precios de supermercados e hipermercados de Ciudad Autónoma de Buenos Aires (CABA), sobre los 1000 productos más frecuentes, existen precios desde noviembre de 2018 hasta finales de febrero de este año.

OBJETIVO GENERAL

El objetivo general de este trabajo es realizar un análisis exploratorio del dataset y el posterior preprocesamiento, de acuerdo a las técnicas vistas en clase, a efectos entender el diferentes comportamientos de los precios en los distintos puntos de ventas de CABA.

Se deberán formular algunas preguntas como primer paso del proceso de KDD para poder guiar el trabajo a partir de esa necesidad de generación de conocimiento. Las mismas deben ser respondidas a través de las diferentes estrategias de análisis.

FUENTE DE DATOS

El proceso de relevamiento de precios fue generado de manera automática mediante la técnica de web crawling. El proceso consistió en la generación de consultas sobre la APP Precios Claros para relevar información de locales, productos y precios. El período para llevar a cabo un relevamiento completo lleva entre una a tres semanas debido a las gran cantidad de datos y consultas que se deben hacer sobre la APP. Este proceso fue repetido sucesivamente a lo largo del tiempo, generando una secuencia de diez mediciones de

[1] www.preciosclaros.gob.ar

precios sobre los productos comercializados. A continuación se detallan las fechas correspondientes a cada una de estas mediciones:

Medición	Fecha de inicio de relevamiento
1	05/11/2018
2	12/11/2018
3	26/11/2018
4	10/12/2018
5	17/12/2018
6	31/12/2018
7	21/01/2019
8	04/02/2019
9	11/02/2019
10	25/02/2019

Tabla 1. *Fechas asociadas a mediciones de precios obtenidas por web crawler.*

CONJUNTO DE DATOS

Se presentan tres archivos en formato JSON Lines[2]:

- **sucursales.json**: Contiene información de todas las sucursales relevadas de comercios del rubro de supermercados, hipermercados y autoservicios ubicados en la ciudad de Buenos Aires.
- **productos.json**: Contienen información de productos de tipo alimentos, bebidas, limpieza, higiene personal y alimentos y productos para mascotas. Está limitado para los 1000 productos con más frecuencia en el relevamientos de precios en CABA.
- **precios.json**: Contiene información de medición de precios sobre los 1000 productos seleccionados en las sucursales de tipo supermercados e hipermercados de CABA (se excluye a la categoría autoservicios). Los precios corresponden a precios de listas, es decir que no contemplan promociones especiales de ningún tipo, por ejemplo promociones bancarias, ni promociones por cantidad. Por lo tanto, corresponden al precio de venta de una unidad en pago en efectivo.

[2] http://jsonlines.org/

```
{
        "_id" : ObjectId("5cbc698b7af152186c0cd14e"),
        "id" : "12-1-108",
        "sucursalTipo" : "Supermercado",
        "direccion" : "Av. Cabildo 4125",
        "provincia" : "AR-C",
        "banderaId" : 1,
        "localidad" : "Cap. Fed",
        "banderaDescripcion" : "COTO CICSA",
        "lat" : -34.5458941,
        "comercioRazonSocial" : "Coto Centro Integral de Comercialización S.A.",
        "lng" : -58.4705622,
        "sucursalNombre" : "CABILDO ",
        "comercioId" : 12,
        "sucursalId" : "108"
}
```


Figura 1. *Documento JSON con información de sucursal (línea de archivo sucursales.json)*


```
{
        "_id" : ObjectId("5cbc69be7af152186c0cd682"),
        "id" : "7790895000232",
        "nombre" : "Coca Cola Sabor Original en Lata 354 Cc",
        "marca" : "COCA COLA",
        "presentacion" : "354.0 cc"
}
```


Figura 2. *Documento JSON con información de producto (línea de archivo productos.json)*


```
{
        "_id" : ObjectId("5cba84a27af1523c358c5457"),
        "producto" : "7790895000232",
        "sucursal" : "12-1-108",
        "precio" : 30.0,
        "fecha" : ISODate("2018-12-12T21:36:36.869Z"),
        "medicion" : 4
}
```


Figura 3. *Documento JSON con información de precio (línea de archivo precios.json)*

TÓPICOS

El trabajo tiene una componente importante de integración de datos, donde se proveerán resultados del crawling. En función de las preguntas planteadas, uds deberán considerar incorporar otras fuentes de información como por ejemplo: comunas y barrios de CABA, cotización del dólar y cualquier otra información que se considere pertinente para el análisis. Una fuente interesante de datos de CABA puede es el portal de datos abiertos de la ciudad[3] . Para integrar los diferentes orígenes de datos se debe utilizar MongoDB.

Las tareas de preprocesamiento vistas en clase le servirán para mejorar el análisis de precios, contemple la posibilidad de generar nuevos features a partir de diferentes transformaciones.

Identificación de atributos con presencia de faltantes. Se debe evaluar para esta situación y cuál es la factibilidad de realizar imputaciones con las técnicas vistas en clase u otra propuesta por uds. Para realizar esta tarea analice las fuentes posibles de faltantes, plantee escenarios teniendo en cuenta el dominio del problema y la técnica de recolección de datos. Analice si es posible identificar o deducir estas fuentes a partir de relaciones entre los atributos o instancias.

Otro análisis de interés en el contexto de un análisis de precios es ver valores extremos o casos atípicos. Tenga en cuenta que las mediciones atípicas podrían estar asociadas a múltiples factores en casos de recolección de datos mediante crawling.

Teniendo en cuenta las preguntas de interés que guíen el proceso de KDD, analice realizar reducción de dimensionalidad sobre aquellas variables o casos que no aporten información adicional al dominio del problema.

PREGUNTAS

A continuación se presentan algunas preguntas que podrían ser respondidas con los datos:

- ¿Existe una variación de precios entre los diferentes barrios porteños?
- ¿Cuáles son los productos que sufrieron mayores y menores variaciones de precios en el tiempo?
- ¿Cuáles son los productos con mayores y menores variaciones de precios entre los distintos puntos de ventas?
- ¿Qué empresa de supermercado ofrece productos más baratos, o más caros? ¿Sobre qué productos?

[3] data.buenosaires.gob.ar

- ¿Los hipermercados ofrecen precios más baratos o más caros que los supermercados?
- ¿Cuáles marcas de productos sufrieron mayores modificaciones de precios?
- ¿A partir de la evolución de precios, se puede detectar productos complementarios o suplementarios?
- ¿Cuáles son los productos que presentan mayor incremento de precios luego de una devaluación del peso?
- ¿En qué período existió un mayor aumento de precios? ¿Cuándo se mantuvieron más estables?
- ¿Concuerdan los resultados obtenidos con las estadísticas del INDEC?

Proponga entre 5 y 10 preguntas de investigación para ser respondidas a partir de los datos provistos y datos adicionales incorporados en el procesos de integración.

INFORME

Se deberá entregar un informe describiendo brevemente cómo se utilizaron y combinaron las técnicas y resultados en las etapas de análisis exploratorio, integración de datos, reducción de dimensionalidad, tratamiento de faltantes, valores extremos y preprocesamiento de datos. Comente las consideraciones tomadas en cada una de estas etapas.

El informe deberá contener los resultados de análisis obtenidos a partir de las preguntas planteadas. Concluya sobre los resultados y descubrimiento de conocimiento obtenido.

No será necesaria la entrega del código desarrollado para resolver el TP. La evaluación del TP se realizará a partir del contenido del informe y no por el código fuente desarrollado.