Laboratorio 03 – Reglas de Asociación

Fabio Luis Buitrago Ochoa   
*Departamento de Ingeniería de Sistemas*  
*Pontificia Universidad Javeriana*Bogotá, Colombia  
fabio\_buitrago@javeriana.edu.co

David Mateo Henao Prieto   
*Departamento de Ingeniería de Sistemas*  
*Pontificia Universidad Javeriana*Bogotá, Colombia  
davidmhenaop@javeriana.edu.co

I. OBJETIVOS

* Familiarizar al estudiante ante una situación similar a la real en donde se requiere analizar un conjunto de datos. En particular, transacciones de compra de una tienda de conveniencia (*retail*).
* Explorar, describir y argumentar el comportamiento de distintas implementaciones de algoritmos distintos que resuelven el mismo problema. En particular, algoritmos de reglas de asociación.
* Elaborar reportes académicos posteriores a la experimentación, medición y observación de fenómenos estudiados. Medición de rendimiento de dos implementaciones, elaboración de apoyo a la explicación del fenómeno, y argumentación basada en la observación, razonamiento y estudio del fenómeno.

II. INTRODUCCIÓN

El presente estudio se enfoca en el análisis de transacciones de una tienda de conveniencia para el día sábado. Se emplea Python como herramienta principal para la manipulación y análisis de los datos, los cuales se encuentran en formato .DBF. El objetivo es realizar una exploración inicial de los datos, llevar a cabo la limpieza necesaria, generar estadísticas descriptivas para comprender el comportamiento de las transacciones y, finalmente, aplicar los algoritmos de generación de reglas de asociación a-priori y FP-growth para identificar patrones de compra y relaciones entre los productos adquiridos.

El análisis de transacciones comerciales es fundamental para comprender el comportamiento de los clientes y mejorar la gestión de inventario y la estrategia de precios. En este contexto, el presente estudio se centra en las transacciones de una tienda de conveniencia durante los días sábados, un período de alta actividad comercial en este tipo de establecimientos.

La metodología empleada en este estudio incluye una exploración inicial de los datos en formato .DBF para identificar tipos y calidad de datos, seguida de una limpieza para garantizar su coherencia. Posteriormente, se generan estadísticas descriptivas para comprender el comportamiento de las transacciones. Se plantea una hipótesis sobre el comportamiento de compra de los clientes, la cual se evalúa utilizando los algoritmos de generación de reglas de asociación a-priori y FP-growth en Python. Finalmente, se analiza el rendimiento de los algoritmos y se responde a las preguntas planteadas en el estudio.

III. EXPLORACIÓN INICIAL DE LOS DATOS

Para realizar la exploración inicial de los datos se crean dos clases iniciales para encapsular la información brindada por ambos archivos .DBF, estas son *LineItem* y *Factura*:

* *LineItem*: Contiene el nombre del producto, cantidad de ítems vendidos y el valor subtotal de la compra.
* *Factura*: Contiene el número o consecutivo de la factura, total de factura, total impuesto y una colección de *LineItems* asociados a la factura.

Para el almacenamiento de la información se utiliza un diccionario que contiene una llave asociada, que es el número de la factura, y un objeto tipo Factura que contiene toda la información antes mencionada.

Después de realizar el almacenamiento de la información, se realiza una exploración inicial de los datos, dentro de la exploración se encuentra que la digitación de algunos productos se encuentra incorrecta, por lo que se tiene que utilizar la librería re para realizar la limpieza de los datos, eliminando espacios y caracteres que no aporten información necesaria.

Después, se realiza la recopilación de la información más importante, entre ellas la cantidad de registros total, la cantidad de facturas representadas en los registros para ese día, cantidad promedio de ítems y la cantidad máxima y mínima de ítems de una factura.

TABLA 01 – ESTADÍSTICAS GENERALES

|  |  |
| --- | --- |
| **Estadística** | **Dato asociado** |
| Cantidad de registros total | 18214 |
| Cantidad de facturas | 8647 |
| Cantidad promedio de ítems | 5.5 |
| Cantidad máxima de ítems por factura | 1538 |
| Cantidad mínima de ítems por factura | 1 |

Además, se realiza un análisis de la distribución a partir de la cantidad de productos o ítems por factura y el rango de precios establecido.

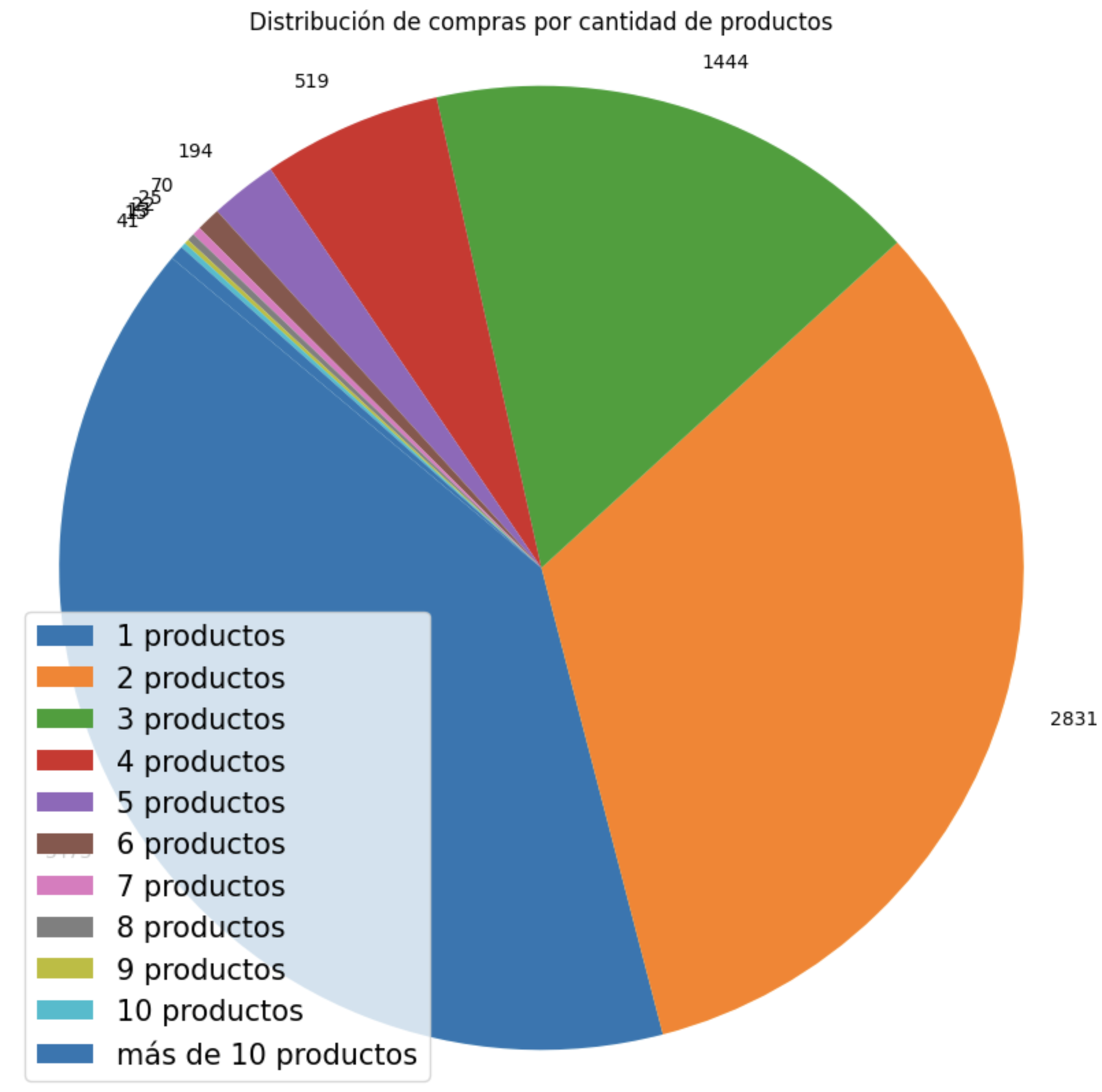


FIGURA 01 – DISTRIBUCIÓN DE COMPRAS POR CANTIDAD DE PRODUCTO

A partir de la *Figura 01*, se puede establecer que el **40.1%** de las facturas posee solamente un ítem, el **32.7%** son facturas que poseen dos ítems y el **27.2%** de las facturas aproximadamente posee 3 artículos o más, por lo que la factura con 1538 ítems es un dato atípico que no demuestra el comportamiento general de las transacciones de esos días, y puede afectar la cantidad promedio de ítems calculada anteriormente.

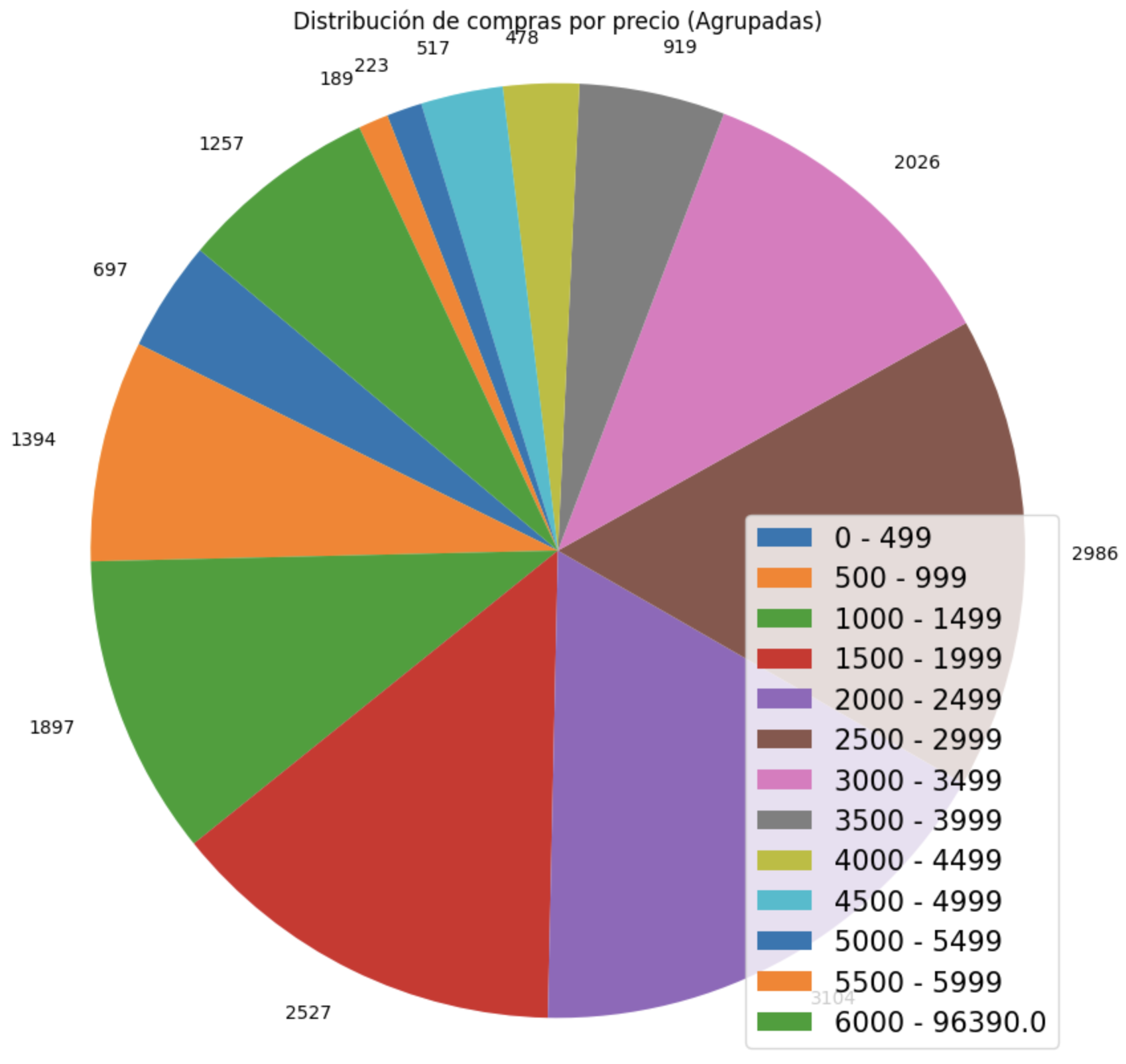


FIGURA 02 – DISTRIBUCIÓN DE COMPRAS POR PRECIO

Ahora bien, a partir de la *Figura 02* se puede concluir que aproximadamente el 85.4% de las compras cuestan $6000 COL o menos, lo cual está correlacionado con el rango de precios de los productos de la tienda de conveniencia, además, está directamente relacionado ocn la distribución de cantidad, con menos productos por factura, el costo de las facturas es menor.

Por otra parte, se realiza el estudio de los productos más vendidos junto con su cantidad asociada:

TABLA 02 – ITEMS MÁS VENDIDOS

|  |  |
| --- | --- |
| **Ítem** | **Cantidad** |
| Empanadas | 2713 |
| Pasteles (Palitos) | 1339 |
| Jugos Hit | 1073 |
| Desechable x Unidad | 1072 |
| Sándwich sencillo | 726 |
| Coca Cola Flexi | 596 |
| CandyRanch 13gr | 527 |
| Pastillas Chao Cereza | 457 |
| Cerveza Poker | 403 |
| Rothmans Azul | 370 |

Se puede inferir a partir de la tabla que los elementos más vendidos tienen una mayor probabilidad de aparecer en la regla de asociación, dado que son los que más frecuentemente se encuentran en las transacciones. Esto sugiere una posible alta relación entre la venta de productos de comida y la de bebidas o dulces. Además, el producto más vendido (la empanada) aparece el **29.25%** de las veces respecto a los ítems más vendidos, esto sugiere que este elemento puede aparecer dentro de las principales reglas de asociación.

IV. Hipótesis

Dada la alta frecuencia de venta de empanadas en las transacciones de la tienda de conveniencia, se plantea que existe una asociación significativa entre la compra de empanadas y ciertos complementos. Este vínculo se ve reforzado por el comportamiento de los clientes, que en su mayoría adquieren uno o dos productos por compra. Se espera que los productos más vendidos, como las empanadas, estén vinculados a otros elementos populares, lo que podría reflejar preferencias de consumo específicas entre los clientes. Además, se anticipa que la identificación de estas asociaciones permitirá la elaboración de combos o recomendaciones de productos que impulsen las ventas y mejoren la experiencia del cliente en la tienda.

V. Desarrollo

Para la comparación entre las implementaciones de Apriori y FP-Growth, se utiliza la biblioteca `mlxtend` en Python. Primero, se cargan y preparan los datos contenidos en los archivos .dbf. Luego, se establecen los parámetros de soporte y umbral para cada ejecución. Estos parámetros son valores numéricos que determinan la frecuencia mínima con la que un conjunto de elementos debe aparecer en los datos para ser considerado relevante. El umbral es un valor que establece el límite mínimo de confianza para que una regla sea considerada válida.

Se itera sobre todas las combinaciones posibles de estos parámetros, lo que significa que se ejecuta el algoritmo con diferentes valores de soporte y umbral. Para cada combinación, se realiza lo siguiente:

1. Se ejecuta el algoritmo Apriori con los datos y los parámetros definidos.

2. Se ejecuta el algoritmo FP-Growth con los mismos datos y parámetros.

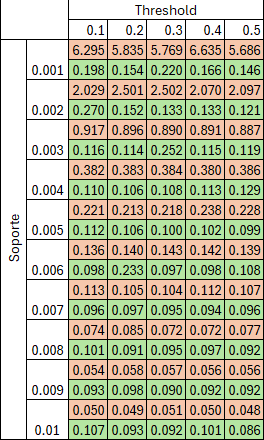
3. Se mide el tiempo de ejecución de cada algoritmo.

VI. Respuesta a preguntas del Taller

1. Liste reglas predecibles entre los productos vendidos k=1. Ejemplo: avena → pandebono.
   1. Tapabocas→Descongel Gripa
   2. Bombas cromada→Feston Felíz cumpleaños
   3. Yogurt Griego→Kumis Natural
   4. Cerveza Poker→Rothmans Azul
   5. Cerveza Poker→Cerveza Aguila
2. ¿Existen diferencias entre las ejecuciones de *apriori* y *fp-growth*?

Existen diferencias significativas entre los algoritmos Apriori y FP-Growth en términos de complejidad computacional, estructura de datos, generación de conjuntos de ítems frecuentes y uso de memoria. Apriori requiere múltiples pasadas sobre la base de datos, lo que aumenta la carga computacional, mientras que FP-Growth es más eficiente gracias a su estructura de árbol de prefijos condensada. En cuanto a la estructura de datos, Apriori utiliza una matriz de conteo de ítems frecuentes, lo que puede ocupar más memoria, mientras que FP-Growth utiliza un árbol de prefijos más compacto. Además, Apriori genera candidatos basados en conjuntos de ítems frecuentes anteriores y escanea la base de datos repetidamente, mientras que FP-Growth construye un FP-Tree y realiza una sola pasada para extraer conjuntos de ítems frecuentes, lo que reduce la complejidad. Por último, Apriori tiende a requerir más memoria, especialmente para conjuntos de datos grandes, en comparación con FP-Growth, que es más eficiente en términos de uso de memoria.

En cuanto a tiempos de ejecución, a continuación, se presentan los tiempos en segundos que tarda cada algoritmo para ejecutarse con los siguientes parámetros de cada uno:

1. ¿Cómo afectan los parámetros, *e.g*. soporte, tanto en la ejecución como en el resultado?

El soporte y el threshold son parámetros esenciales en la minería de reglas de asociación. El soporte determina la frecuencia mínima para considerar un conjunto de ítems como relevante, influyendo en la especificidad y cantidad de patrones identificados. Un soporte alto revela patrones más específicos, mientras que uno bajo muestra patrones más generales. Por otro lado, el threshold actúa como un umbral para decidir la relevancia de los resultados. Un umbral alto garantiza resultados más fiables pero menos abundantes, mientras que uno bajo ofrece más resultados, aunque algunos podrían ser menos confiables. Ajustar adecuadamente estos parámetros es crucial para obtener reglas de asociación significativas y útiles.

1. ¿Encuentra reglas espurias? ¿Con cuales parámetros?

Para determinar la presencia de reglas espurias, se evalúan el lift y el soporte asociado. El lift indica la fuerza de la asociación entre los elementos, mientras que el soporte refleja la frecuencia de esa asociación en el conjunto de datos. Una regla con alto lift y bajo soporte podría ser considerada espuria, ya que aunque muestra una conexión significativa entre los elementos, esta asociación es poco frecuente en el contexto general. Un ejemplo es la regla que relaciona "ARTESANAL MARACUYA 70GR" y "CHOCO CONO 90GR", con un lift extraordinariamente alto (45.41) pero un soporte bajo (0.001156), lo que sugiere una posible relación casual. Los parámetros empleados fueron soporte: 0.001 y threshold: 0.1.

1. Liste reglas con más de 1 producto, k > 1.

La salida de los resultados revela que no se encontraron reglas que cumplieran con el criterio de tener más de un producto tanto en el antecedente como en el consecuente. Este hallazgo puede atribuirse a la tendencia de compra de pocos productos por factura en el conjunto de datos analizado. La falta de reglas con múltiples productos en ambos lados sugiere que los clientes tienden a comprar productos de forma individual o en combinaciones más simples en lugar de realizar compras complejas con múltiples artículos

1. ¿Cómo calcula la probabilidad de que al azar se compren estos dos productos?

Para calcular la probabilidad de que al azar se compren dos productos específicos, podemos utilizar el concepto de probabilidad conjunta. La probabilidad conjunta de que dos eventos ocurran juntos se calcula multiplicando las probabilidades de cada evento individual.

Si tenemos dos productos A y B, la probabilidad de que ambos productos se compren al azar se puede calcular como:

P(A∩B)=P(A)×P(B)

Donde P(A) es la probabilidad de comprar el producto A al azar y P(B) es la probabilidad de comprar el producto B al azar.

1. ¿Encuentra reglas interesantes?

La regla más interesante para el equipo de trabajo es la que asocia la compra de Yogurt Griego Alpina 150G con la compra de Empanadas. Esto es por la naturaleza de los alimentos. Ya que uno de estos es considerado como Alimento saludable y el otro como comida rápida o de bajo contenido nutricional.

El resto de las reglas en cuestión presenta un comportamiento similar que relaciona comida rápida o productos empacados salados (Empanadas, papas, etc) con ya sea una bebida (Gaseosa, jugo, etc) o con un producto dulce (Chocolate, Ponky, avena, etc). Por ende, resalta mucho más la regla mencionada anteriormente.

1. ¿Cómo explica el comportamiento del consumidor para estas reglas?

Las reglas que relacionan alimentos rápidos o productos empacados salados con bebidas o productos dulces pueden ser explicadas por la tendencia de los consumidores a combinar alimentos salados con bebidas refrescantes o productos dulces para equilibrar los sabores. Por ejemplo, es común que las personas acompañen empanadas o papas fritas con gaseosas, jugos o bebidas energéticas para contrarrestar el sabor salado de los alimentos. Del mismo modo, la combinación de productos salados con productos dulces puede deberse a preferencias personales por sabores contrastantes o a la búsqueda de una experiencia sensorial más completa.

En resumen, el comportamiento del consumidor detrás de estas reglas sugiere que las decisiones de compra pueden estar influenciadas por factores como preferencias de sabor, conveniencia y contexto de consumo, lo que lleva a asociaciones específicas entre diferentes tipos de productos en las transacciones de compra.

VII. Conclusiones

●El taller contrastó los algoritmos a-priori y FP-growth, destacando diferencias en complejidad y eficiencia. Mientras a-priori realiza múltiples pasadas y utiliza una matriz de conteo, FP-growth emplea un árbol de prefijos en una sola pasada, resultando en tiempos de ejecución más rápidos y menor uso de memoria. Esta comprensión permite seleccionar el algoritmo más adecuado según las necesidades computacionales y los datos disponibles.

●Se exploró cómo ajustar el soporte y el umbral afecta la generación de reglas de asociación. Un mayor soporte genera patrones más específicos pero menos abundantes, mientras que un umbral más bajo produce más resultados pero potencialmente menos fiables. Ajustar estos parámetros adecuadamente es crucial para obtener insights significativos y relevantes en el análisis de datos.

●El análisis reveló reglas interesantes sobre el comportamiento del consumidor en la tienda de conveniencia, como la asociación entre alimentos saludables y comida rápida. Estos hallazgos proporcionan información valiosa para la gestión de inventario y la elaboración de estrategias de marketing. Además, la detección de reglas espurias resalta la importancia de evaluar la significancia de los resultados para evitar interpretaciones erróneas y tomar decisiones informadas.