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.

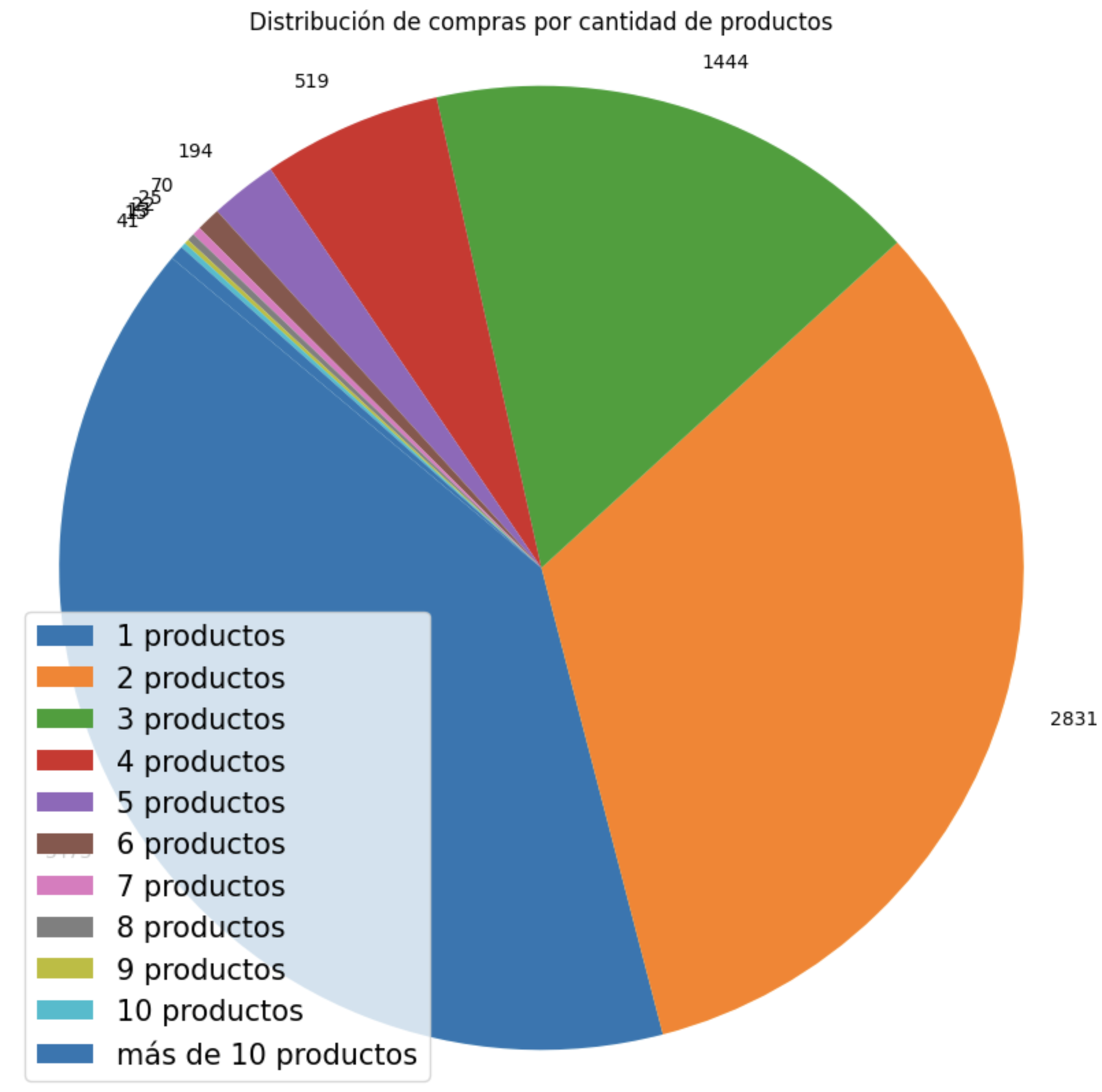


FIGURA 01 – DISTRIBUCIÓN DE COMPRAS POR CANTIDAD DE PRODUCTO

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.

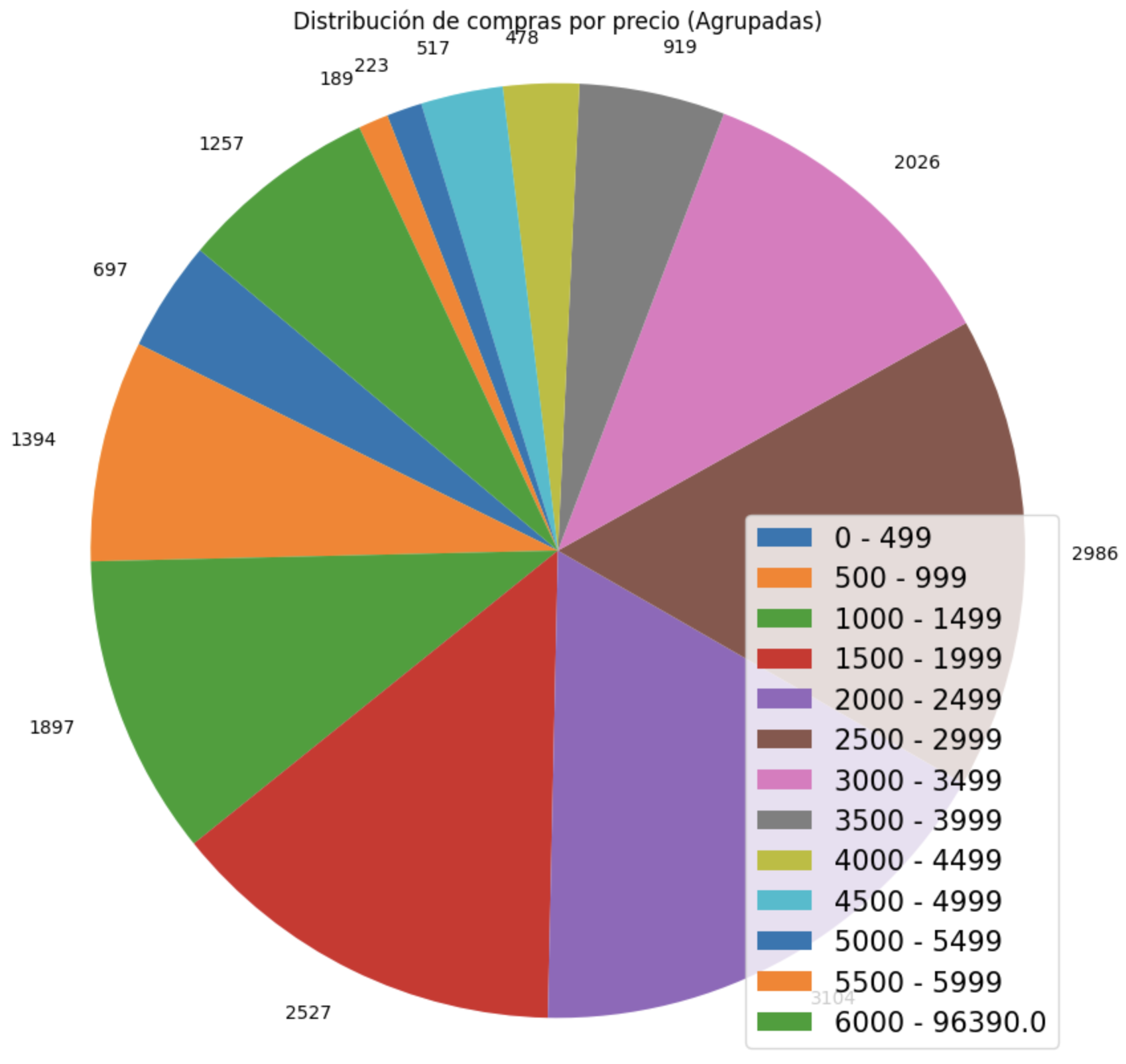


FIGURA 02 – DISTRIBUCIÓN DE COMPRAS POR PRECIO

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

V. Desarrollo

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
2. ¿Existen diferencias entre las ejecuciones de *apriori* y *fp-growth*?

Existen ciertas diferencias en las ejecuciones de los dos algoritmos, algunas de ellas se listan a continuación:

Complejidad computacional: Apriori se caracteriza por requerir múltiples pasadas sobre la base de datos para generar los conjuntos de ítems frecuentes, lo que puede resultar en una mayor carga computacional. En contraste, FP-Growth tiende a ser más eficiente al construir una estructura de árbol de prefijos condensada, lo que reduce la cantidad de pasadas necesarias y hace que sea más eficiente en términos de tiempo de ejecución.

Estructura de datos utilizada: Apriori se apoya en una matriz de conteo de ítems frecuentes para mantener información sobre la frecuencia de los ítems y conjuntos de ítems, lo que puede ocupar más memoria. Por otro lado, FP-Growth utiliza una estructura de árbol de prefijos condensada para representar la base de datos de manera más compacta, lo que reduce el uso de memoria y hace que sea más eficiente en el almacenamiento de datos.

Generación de conjuntos de ítems frecuentes: Mientras Apriori genera candidatos basándose en los conjuntos de ítems de tamaño (k-1) frecuentes y luego escanea la base de datos para contar su soporte, FP-Growth construye un árbol de prefijos (FP-Tree) y realiza una sola pasada sobre la base de datos para extraer los conjuntos de ítems frecuentes, lo que reduce la complejidad y hace que sea más rápido.

Uso de memoria: Apriori puede requerir una cantidad considerable de memoria, especialmente para conjuntos de datos grandes, debido a la necesidad de mantener la matriz de conteo. Por el contrario, FP-Growth tiende a utilizar menos memoria, ya que la estructura de árbol de prefijos es más compacta, lo que lo hace más adecuado para conjuntos de datos extensos y limitados recursos de memoria.

En cuanto a tiempos de ejecución, a continuación, se presentan los tiempos que tarda cada algoritmo para ejecutarse:

|  |  |
| --- | --- |
| Apriori |  |
| Fp-growth |  |

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

El soporte es un parámetro fundamental en la minería de reglas de asociación, ya que determina la frecuencia mínima con la que un conjunto de ítems debe aparecer en la base de datos para considerarse relevante. Establecer un valor adecuado de soporte permite identificar conjuntos de ítems que ocurren con suficiente frecuencia como para ser considerados significativos en el análisis de asociación. Un soporte más alto conduce a la identificación de patrones más específicos y significativos, mientras que un soporte más bajo puede revelar patrones más generales y numerosos. Ajustar el soporte adecuadamente es crucial para obtener reglas de asociación relevantes y útiles.

Por otra parte, el threshold, o umbral, es como una línea de corte que ayuda a decidir qué resultados son lo suficientemente importantes para considerarlos. Si el resultado supera este umbral, lo consideramos relevante y lo conservamos. Un umbral más alto nos da resultados más seguros, pero menos abundantes, mientras que uno más bajo nos da más resultados, aunque algunos podrían ser menos fiables. En resumen, el threshold nos ayuda a filtrar y seleccionar los resultados más relevantes para nuestro análisis.

1. ¿Encuentra reglas espurias? ¿Con cuales parámetros?
2. Liste reglas con más de 1 producto, k > 1.
3. ¿Cómo calcula la probabilidad de que al azar se compren estos dos productos?
4. ¿Encuentra reglas interesantes?
5. ¿Cómo explica el comportamiento del consumidor para estas reglas?

VII. Conclusiones