Tecnicatura Universitaria en Procesamiento y Explotación de Datos

Minería de Datos

Trabajo Práctico Final

Bioing. Juan Aued

Lopez Melina, Saldaña Guillermo

Introducción

En un mundo donde el comercio electrónico sigue creciendo a pasos agigantados, entender el comportamiento de los usuarios en línea es fundamental para optimizar las estrategias de ventas y mejorar la experiencia del cliente. El presente trabajo analiza los datos de navegación registrados en una tienda online polaca especializada en ropa para embarazadas durante el año 2008, con el objetivo de identificar patrones de comportamiento, categorías de mayor interés y relaciones frecuentes entre productos.

Para ello, se aplica una combinación de técnicas de minería de datos, que incluyen análisis exploratorio, generación de reglas de asociación mediante el algoritmo Apriori, y detección de patrones secuenciales utilizando el algoritmo cSPADE.

Desarrollo

Descripción del dataset

La base contiene 14 variables y 165.474 observaciones, cada una representando un clic realizado por un usuario dentro de una sesión de navegación.

Cada fila en el dataset representa una interacción del usuario con un producto. Las variables están codificadas principalmente en formato numérico o de texto corto, y fueron interpretadas de acuerdo con la documentación oficial del conjunto de datos.

Las variables disponibles son:

- Year (fecha): año del evento.
- Month (fecha): mes del evento.
- Day (fecha): día de la sesión.
- Order (numérica): secuencia de clicks durante la sesión.
- Country (categórico): nombre del país donde reside el cliente.
- Session ID (categórica): ID de la sesión.
- “page 1 (main category)” (categórica): categoría principal del producto.
- “page 2 (clothing model)” (categórica): código de cada producto.
- Colour (categórica): color del producto.
- Location (categórica): ubicación de la foto en la página, la pantalla se ha dividido en seis partes.
- Model photography (categórica): variable binaria.
- Price (numérica): el precio de cada producto por unidad en dólares estadounidenses (USD).
- Price 2 (categórica): Variable que informa de si el precio de un producto concreto es superior al precio medio de toda la categoría de productos.
- Page (categórica): número de página dentro del sitio web de la tienda electrónica.

El archivo no presenta valores nulos y su estructura es adecuada para realizar análisis tanto a nivel exploratorio como avanzado. Las variables relevantes han sido tratadas y transformadas en formatos apropiados para el análisis posterior.

Análisis exploratorio:

Clicks por sesión:

Para este análisis se contó la cantidad total de clics registrados en cada sesión, agrupando por session.ID. Luego, se agrupó en rangos según la cantidad de clics, para visualizar la distribución sin que los valores extremos distorsionen la escala.

Como podemos observar en el Gráfico 1, la mayoría de las sesiones tuvieron entre 1 y 2 clicks, lo que se podría interpretar como que varios usuarios realizaron visitas breves en el sitio. Además, a medida que aumenta el rango de clicks, el número de sesiones disminuye notablemente.

Aproximadamente unas 6000 sesiones se ubicaron en el rango de 3 a 5 clicks y unas 5000 sesiones en el rango de 6 a 10 clicks.

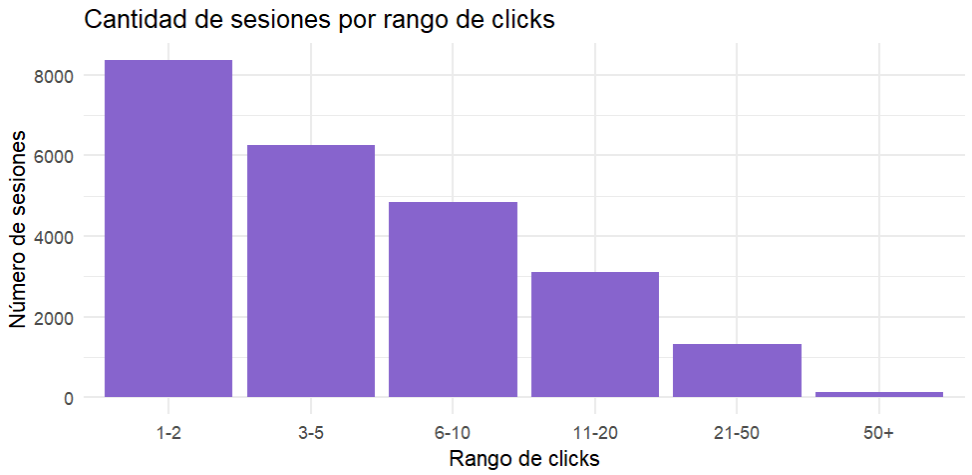

Gráfico N° 1

Fuente de elaboración propia

Sesiones por país:

Se contabilizó la cantidad de sesiones únicas por país de origen, utilizando la variable country. Para evitar que el gráfico se vea sesgado, se excluyó a Polonia (país de origen del e-shop), así como dominios genéricos (como .com, .org, etc.) que no representan ubicaciones geográficas reales.

El Gráfico 2 muestra los 10 países con mayor cantidad de sesiones luego de aplicar este filtro. La República Checa se destaca con diferencia, seguida por Lituania, el Reino Unido e Irlanda.

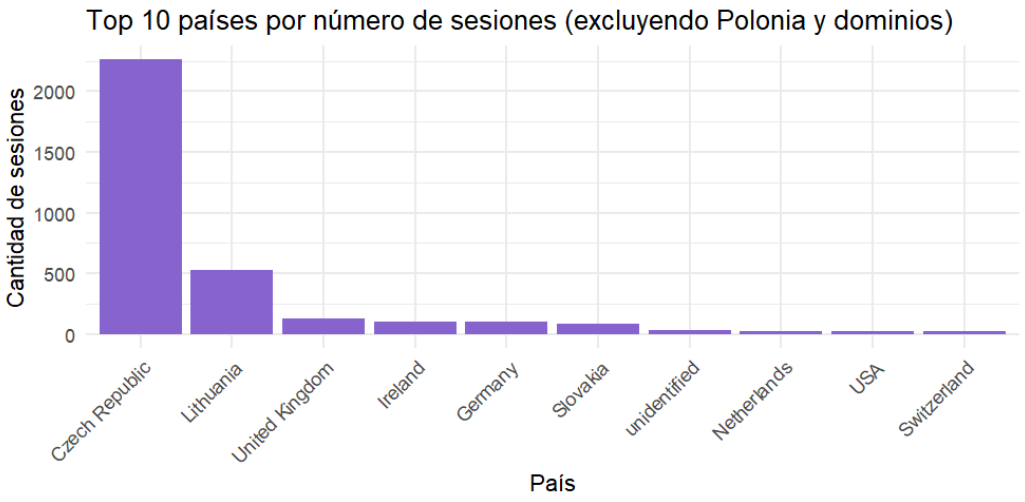

Gráfico N° 2

Fuente de elaboración propia

Productos vistos por sesión:

Las sesiones se agruparon según la cantidad de productos únicos visualizados por los usuarios. Luego, se calculó la cantidad de sesiones dentro de cada rango.

El Gráfico 3 muestra que la mayoría de las sesiones se concentran en los rangos más bajos, especialmente entre 1 y 5 productos vistos, con una disminución progresiva a medida que aumenta la cantidad de productos explorados.

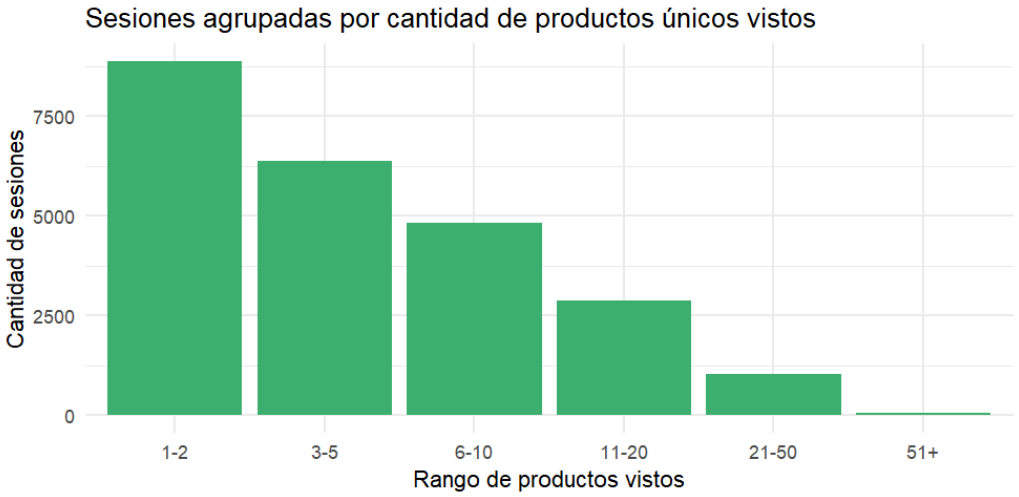

Gráfico N° 3

Fuente de elaboración propia

Categoría de producto por sesión:

Se agruparon las sesiones según la cantidad de categorías distintas exploradas por los usuarios. Luego, se calculó la frecuencia de cada grupo y se representó en un treemap.

El Gráfico 4 muestra que la mayoría de las sesiones corresponden a la exploración de una única categoría de productos, con una disminución progresiva a medida que aumenta la cantidad de categorías vistas.

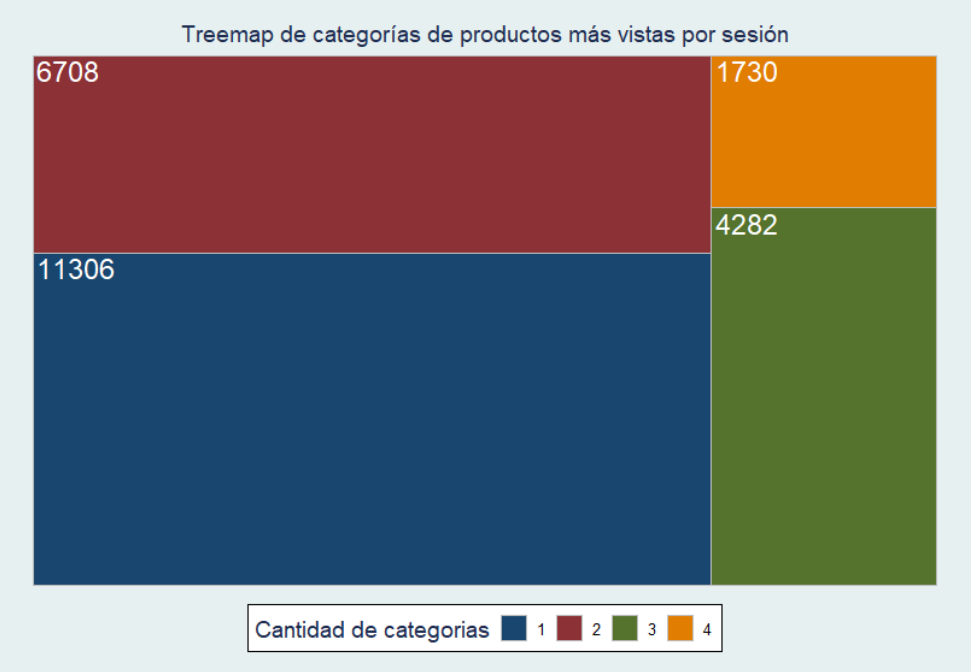

Gráfico N° 4

Fuente de elaboración propia

Categorías de productos más vistas

Las sesiones se agruparon según la categoría de producto visualizada, y se calculó la cantidad de sesiones para cada una.

El Gráfico 5 muestra que la categoría más vista es Trousers (pantalones), seguida de Skirts (polleras), Blouses (blusas) y finalmente Sale (oferta).

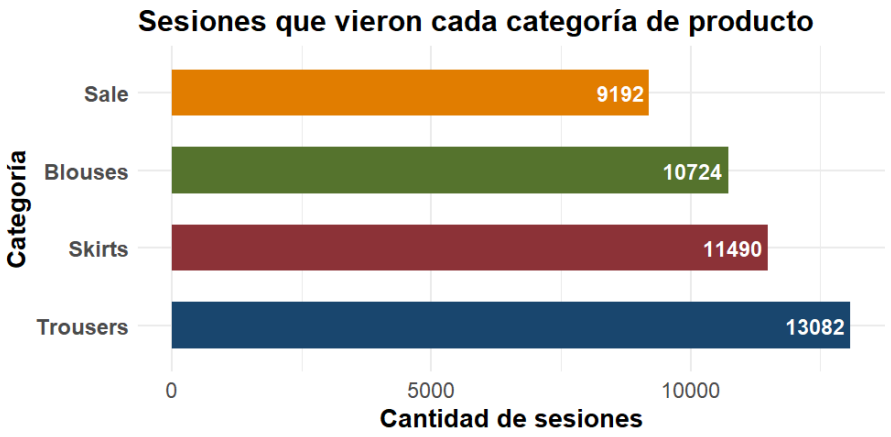

Gráfico N° 5

Fuente de elaboración propia

Evolución de clicks a lo largo de los meses:

El Gráfico N° 6 muestra cómo ha sido la evolución de los clicks de navegación a lo largo de los meses de abril y agosto del año 2008. Abril fue el mes con mayor actividad, registrando 48.199 clicks, lo que indica un periodo de alta interacción. Sin embargo, en mayo los clicks cayeron a 36.654, y en junio descendieron aún más a 32.242. En julio se observó una leve suba al alcanzar 35.231 clicks, pero en agosto hubo una caída abrupta a 14.148.

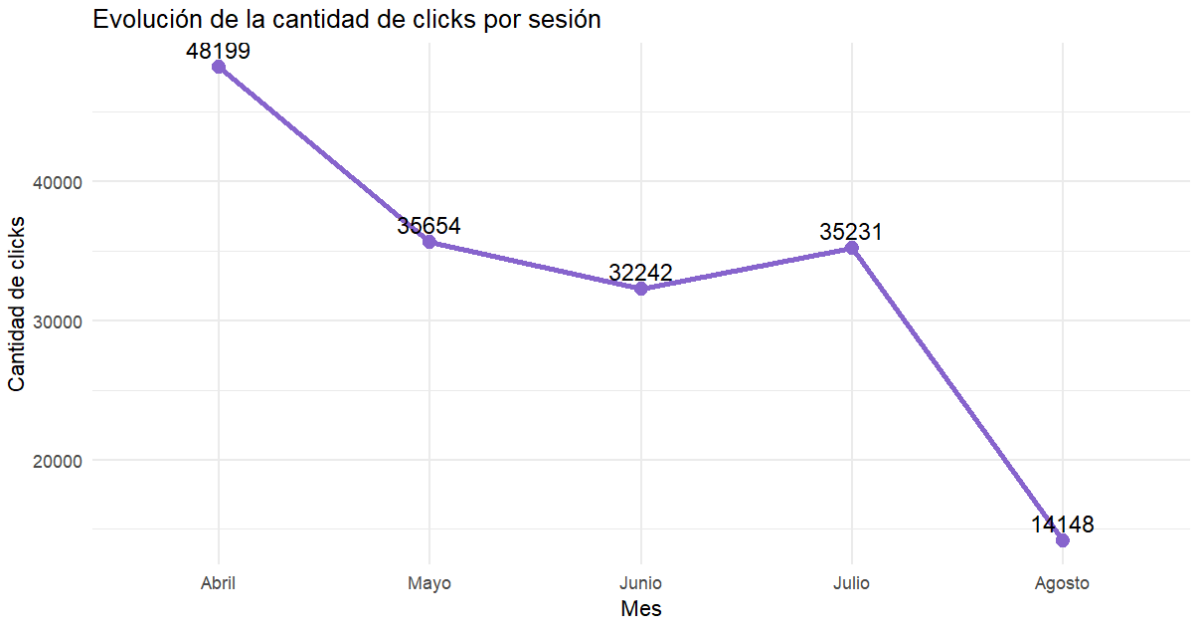

Gráfico N° 6

Fuente de elaboración propia

Análisis de patrones de compra:

Reglas de asociación a nivel general:

Se analizaron 24026 sesiones de navegación en las que se identificaron 217 modelos de ropa distintos. A partir de esta información, se construyó una estructura de transacciones para detectar conjuntos de ítems frecuentemente comprados juntos mediante el algoritmo Apriori, utilizando un soporte mínimo del 2% y limitando los conjuntos a aquellos con al menos dos ítems.

Como resultado obtuvimos que A2, A5, A3, A11 y A1 son los productos centrales en múltiples combinaciones frecuentes. Al analizar los códigos de producto y sus categorías, se observó que las combinaciones más comunes corresponden a productos dentro de la misma categoría: los códigos A corresponden a pantalones, B a polleras, C a blusas y P a productos en oferta (Sale). Esto sugiere que los clientes tienden a explorar y adquirir productos dentro de una misma categoría en lugar de combinar artículos de distintos tipos.

Reglas de asociación por país:

- **Polonia:** Se destacan fuertes asociaciones entre modelos como C12 y C17, y entre C56 y C57, con lifts superiores a 2 y hasta 3.77, indicando relaciones estadísticamente significativas. Además, el ítem C17 aparece como punto central de múltiples reglas.
- **República Checa:** La regla que más se destaca es entre los ítems C57 y C56. El 5,12% de todas las transacciones contienen ambos productos, y en el 42,76% de las sesiones donde se compró el modelo C57 también se compró C56. Otras asociaciones fuertes se dan entre los modelos C50 y C49 y entre C29 y C40.

En República Checa, las reglas no solo presentan mayores valores de soporte y lift, sino que también sugieren comportamientos de compra más consistentes entre ciertos productos. Esto se puede observar entre las fuertes asociaciones de productos específicos como C56, C57, C49 y C50, destacando una posible preferencia por ciertos conjuntos de blusas. Por otro lado, en Polonia, las reglas muestran una mayor variedad y están distribuidas entre diferentes productos. Las asociaciones en este país reflejan un comportamiento más diverso, con preferencias de productos como C5, C12, C17 y C1.

Análisis de secuencias frecuentes:

Se identificaron las secuencias más frecuentes de productos visualizados por los usuarios, considerando aquellas con más de un ítem y un soporte superior al 2%. Para ello, se transformaron los datos en un formato transaccional, asignando identificadores de sesión y orden de visualización. Luego, se aplicó el algoritmo cSPADE para detectar patrones de navegación.

Los resultados muestran que las secuencias más frecuentes suceden dentro de categorías específicas. La secuencia A2 → A5 es la más prominente, seguida de combinaciones como A1 → A2, A2 → A3 y A2 → A11, lo que indica una alta asociación entre distintos modelos de pantalones. Asimismo, la combinación B10 → B13 evidencia patrones de exploración dentro de la categoría de polleras.

Todas las secuencias identificadas corresponden a productos dentro de la misma categoría, lo que sugiere que los usuarios tienden a navegar y comparar variantes de un mismo tipo de prenda en lugar de mezclar productos de diferentes grupos.

Conclusión

El análisis realizado sobre la navegación en la tienda online permitió identificar patrones clave en el comportamiento de los usuarios. Se observó que la mayoría de las sesiones contienen un número reducido de clicks y productos visualizados, lo que sugiere interacciones breves. Además, el análisis de reglas de asociación y secuencias frecuentes evidenció que los consumidores tienden a explorar productos dentro de una misma categoría, sin mezclar diferentes tipos de prendas en sus sesiones de navegación. Estos hallazgos pueden ser aprovechados para mejorar la organización del sitio web y optimizar estrategias de recomendación. Asimismo, el análisis por país mostró diferencias en los patrones de compra, lo que sugiere oportunidades para adaptar estrategias comerciales según el mercado

Referencias

Łapczyński M., Białowąs S. (2013). *Discovering Patterns of Users' Behaviour in an E-shop*. Studia Ekonomiczne, nr 151.