Trabajo Integrador Final

Introducción

El objetivo del trabajo integrador es aplicar los conocimientos teóricos vistos en la materia, principalmente la exploración y preparación de datos, para la aplicación de reglas de asociación y detección de secuencias.

Entregables

Parte I – Resolución del trabajo práctico:

En equipos **de hasta dos (2) integrantes** se debe entregar un script con el código utilizado para resolver el trabajo (correctamente comentado) y un informe en formato PDF con las respuestas a las actividades planteadas.

Fecha límite de entrega: viernes 07/06 a las 23:59 hs.

Parte II - Defensa del trabajo realizado:

Las lecturas y comentarios se realizarán por medio del Aula virtual entre 10 y el 12 de junio de 2024 para coordinar un día y horario para realizar la defensa del trabajo. La defensa del mismo será individual entre el 14 al 19 de junio.

La defensa será referida a una consulta sobre los contenidos teóricos de la materia aplicados en el trabajo y sobre las características del script presentado, los resultados obtenidos, las dificultades presentadas y diferentes propuestas que pudieran realizar para mejorar y/o modificar la propuesta.

Contexto del problema

El comercio electrónico se ha convertido en un nuevo canal para apoyar el desarrollo de las empresas. A través del comercio electrónico, las empresas pueden acceder y establecer una mayor presencia en el mercado proporcionando canales de distribución más baratos y eficientes para sus productos o servicios. El comercio electrónico también ha cambiado la forma en que la gente compra y consume productos y servicios. Muchas personas recurren a sus ordenadores o dispositivos inteligentes para pedir productos, que pueden ser entregados fácilmente en sus hogares.

En este caso, una e-shop polaca de ropa para embarazadas quiere dirigirse a los clientes con sugerencias sobre el conjunto de artículos que es más probable que un cliente compre, lo que permitirá aumentar el compromiso de los clientes, mejorar su experiencia e identificar su comportamiento.

Datos

Se trata de un conjunto de datos de que contiene el flujo de clics durante distintas sesiones, en una tienda en línea con sede en Polonia, durante 5 meses del año 2008. Los datos de navegación provienen de clientes procedentes de distintos países.

El conjunto de datos contiene más de 165.000 filas y 14 columnas entre el 01/04/2008 y el 13/08/2008. 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.

Enlace de descarga del dataset:
https://archive.ics.uci.edu/dataset/553/clickstream+data+for+online+shopping

Nota: Junto a los datos propiamente dichos, hay un archivo *.txt que contiene la información de las variables y sus categorías.

Actividades

a) Explore los datos y presente sus características principales.
b) Utilizando las gráficas que considere adecuadas, muestre: clicks por sesión, sesiones por país, productos vistos por sesión y por categoría de producto.
c) ¿Cómo ha sido la evolución de los clicks de navegación a lo largo de los meses estudiados?
d) Encuentre el número de transacciones e ítems (pensando cada sesión como una transacción).
e) Encuentre un conjunto de itemsets frecuentes para un soporte mínimo de 2% y con una longitud mínima de 2 ítems.
f) Encuentre las reglas de asociación para los datos de navegación correspondientes a Polonia, en la categoría “blusas”. Para un soporte mínimo de 2% y una confianza de 20%. Muestre las 10 reglas de mayor soporte.
g) Encuentre las reglas para la República Checa, en la misma categoría del ítem anterior, pero para un soporte mínimo de 4% y una confianza de 25%. Muestre las 10 reglas de mayor soporte.
h) Encuentre las secuencias más frecuentes que tienen más de un elemento (ítem) y un soporte mayor a 3%.