Trabajo Práctico Especial

Online Shoppers Purchasing

Curso: Fundamentos de la Ciencia de Datos

Integrantes

- **Felipe Costa**
- **Micaela Soledad Fernández**

Fecha de entrega: 16/11/2025

Análisis de datos sobre el dataset:
Online Shoppers Purchasing Intention

- *Análisis Exploratorio y Limpieza*

El siguiente análisis exploratorio de datos (EDA) tiene como objetivo comprender el comportamiento de los usuarios dentro del *Online Shoppers Purchasing Intention Dataset*. Este conjunto de datos contiene información sobre 12.330 sesiones de navegación registradas en un sitio de comercio electrónico a lo largo de un año, donde cada sesión corresponde a un usuario distinto.

El dataset incluye variables tanto numéricas como categóricas que describen la interacción de los visitantes con diferentes tipos de páginas, la duración de la navegación, las tasas de rebote y salida, el origen del tráfico, la condición del visitante, la época del año y si la sesión culminó o no en una compra.

Este análisis permite identificar la estructura general del dataset, detectar patrones, evaluar la distribución de las principales variables y comprender relaciones potenciales entre ellas.

Se comenzó observando tipos de datos, existencia de valores nulos, repetidos o extremos que podrían afectar el análisis posterior del dataset.

- Al realizar las pruebas correspondientes se verificó la no presencia de *valores nulos* así como de *NaNs*. Esto es importante a la hora de generar tests o gráficas a fin de ver el comportamiento del dataset.
- Por otro lado se vió, que en las columnas que representan variables numéricas existían valores atípicos que escapaban de la varianza esperada, aunque al considerar qué hacer con estos se llegó a la conclusión de que eliminarlos supondría perder información importante. Por lo tanto, y para que los análisis no fuesen sesgados por estos valores se tomaron precauciones de ser necesarias.
- Finalmente al revisar existencia de duplicados, se encontraron *125 filas duplicadas* (el 1% aproximadamente). Si bien no existe un ID único para confirmar, la probabilidad de que 18 variables coincidan exactamente por azar es muy baja. Con esto en mente, se consideran posibles errores de la recolección de datos y fueron eliminados.

- *Análisis Univariado*

Para este paso se tomaron dos hipótesis que se consideraron, su análisis podía ser de gran utilidad para tomar decisiones en un hipotético marco empresarial o en donde esta información sea de peso.

HIPÓTESIS 1:

Descripción

- “La mayoría de los visitantes son 'Returning_Visitors.”

  De ser probada, supondría que los usuarios que vuelven al sitio tienen más probabilidad de comprar y aportar valor a largo plazo. Además, retener visitantes recurrentes puede ser más fácil que adquirir nuevos, lo que mejoraría el marketing.

Estrategia de abordaje

- Para evaluar esta hipótesis se realizó un análisis univariado sobre la variable categórica *VisitorType*, que clasifica cada sesión en tres tipos de usuario: *Returning_Visitor*, *New_Visitor* y *Other*. La estrategia consistió en calcular la frecuencia relativa de cada categoría y visualizar su distribución mediante un gráfico de tortas, lo que permite observar rápidamente la composición del tráfico según tipo de visitante.

Resultados

El análisis mostró que:

- 85.5% de las sesiones corresponden a Returning_Visitors,
- 13.9% a New_Visitors,
- 0.7% a usuarios clasificados como Other.

La marcada diferencia entre la proporción de visitantes recurrentes y la de nuevos visitantes evidencia que la mayoría del tráfico está formado por usuarios que ya han interactuado anteriormente con la plataforma.

HIPÓTESIS 2:

Descripción

“La mayoría de compras se realizan cuando ProductRelated es alto”

Para evaluar esta hipótesis se analizó la variable ProductRelated, que representa la cantidad de páginas relacionadas a productos que un usuario visitó durante una sesión. La pregunta es si las sesiones que finalizan en compra (Revenue = True) tienden a presentar valores altos en esta variable.

Conocer si las ventas aumentan o no cerca de días especiales es importante para entender el comportamiento del consumidor y evitar suposiciones equivocadas a la hora de, por ejemplo, realizar estrategias de marketing o mercado.

Estrategia de abordaje

- Filtrar únicamente las sesiones que finalizaron en compra, ya que la hipótesis se refiere exclusivamente a ellas.
- Definir qué se considera un valor “alto” de ProductRelated definiendo un umbral.
- Clasificar cada sesión en dos grupos:
    - “Menos de umbral”
    - “Umbral o más” (considerado *ProductRelated alto*)
- Visualizar y comparar la frecuencia de ambos grupos mediante un gráfico de barras.

Resultados

El conteo de sesiones arrojó:

- 1424 sesiones con ProductRelated por debajo del umbral,
- 484 sesiones con ProductRelated por encima del umbral.

Solo el 25% de las sesiones con compra presentan valores de *ProductRelated* “altos”.

La gran mayoría de las compras (74%) ocurrió en sesiones donde *ProductRelated* estuvo por debajo del valor considerado alto.

Finalmente, los datos permiten refutar la hipótesis de que *“la mayoría de las compras se realizan cuando ProductRelated es alto”*.

- *Análisis Bivariado*

HIPÓTESIS 1:

Descripción

“La mayor cantidad de ventas ocurre cerca de un día especial.”

Conocer si las ventas aumentan o no cerca de días especiales es clave para entender el comportamiento real del consumidor y evitar suposiciones equivocadas.

Estrategia de abordaje

- Separación del dataset según Revenue
- Se generaron dos subconjuntos:
    - df_ventas: sesiones que terminaron en compra (Revenue = True)
    - df_no_ventas: sesiones que NO terminaron en compra (Revenue = False)

- Clasificación de sesiones según proximidad a un día especial
    - Se tomó como criterio:
        - Cerca de día especial: SpecialDay ≥ 0.5
        - Lejos de día especial: SpecialDay < 0.5

- Visualización de proporciones con gráficos de torta
    - Se generó un gráfico para ventas y otro para no ventas.
    - Esto permitió observar la distribución de cada grupo respecto a los días especiales.

- Prueba estadística para validar la hipótesis
    - Se evaluó normalidad y homocedasticidad → ambas rechazadas. Eso justificó el uso de un test no paramétrico acorde.
    - Se aplicó Kruskal-Wallis para comparar la distribución de SpecialDay entre sesiones con y sin ventas.

Resultados

Distribución visual

- Entre las sesiones con venta:
    - Solo 2.6% ocurrió cerca de un día especial.
    - El 97.4% ocurrió lejos de un día especial.
- Entre las sesiones sin venta:
    - 7.6% ocurrió cerca de un día especial.
    - 92.4% lejos de un día especial.

A simple vista, las ventas no se concentran cerca de días especiales.

Al contrario, incluso parecen ser más frecuentes lejos de ellos.

Test estadístico aplicado

Kruskal–Wallis para SpecialDay entre ventas y no ventas:

- Estadístico = 95.178
- p-valor = 0.000

Finalmente, se rechaza la hipótesis nula, por lo que existe una diferencia significativa en la variable *SpecialDay* entre sesiones con venta y sin venta.

HIPÓTESIS 2:

Descripción

"Las sesiones que ocurren durante el fin de semana presentan un comportamiento de compras distinto al de las sesiones realizadas en días de semana."

Esta información puede resultar útil a la hora de saber en qué momento promover ofertas, rebajas o publicidad aumentando la eficiencia de lo invertido.

Estrategia de abordaje

- Primero se plantearon gráficos de barras dobles para observar el comportamiento de sesiones que resultaron en compra y las que no, según si fué un día de semana o un fin de semana.
- Posteriormente y dado que los atributos analizados eran categóricos, se avanzó con la realización de la prueba chi-cuadrado la cual genera un valor que indica si existe una asociación estadísticamente significativa entre dos variables categóricas.
  Si el p valor era menor que el nivel de significancia (0.05), se rechazaría la hipótesis nula de independencia, concluyendo que existe una relación.

Resultados

- El resultado obtenido de la prueba fue un p = 0.00241. Lo cual ratificó, que como podía suponerse del gráfico de barras, ambas distribuciones no eran iguales.

Por lo tanto se comprueba la hipótesis planteada al inicio, y como puede verse en el gráfico, el tráfico de usuarios en días de semana, es mucho mayor que el de las ocurridas los fines de semana.

HIPÓTESIS 3:

Descripción

"A mayor valor de las páginas visitadas (PageValues), más frecuente es que una sesión termine en una compra."

Si se confirmara que PageValues altos están fuertemente asociados a compras, un sitio web o empresa podría identificar qué tipos de páginas realmente generan valor y optimizar el recorrido del usuario hacia ellas (mejorando su accesibilidad, diseño o contenido).

Estrategia de abordaje

- Se realizó un gráfico de violín, el cual es similar a un boxplot pero muestra una densidad de probabilidad. Esto para visualizar las densidades de Page Values para los grupos de Revenue (si se realizó o no la compra).
- A su vez, y para comprobar lo que se viera en el gráfico, se planteó el uso de un test. Como las distribuciones resultaron no normales y sus varianzas resultaron no ser estadísticamente iguales (homocedásticas) se usó el test de Kruskal-Wallis.

Resultados

El resultado arrojado por el test de Kruskal-Wallis rechazó la hipótesis nula, por lo tanto ambas distribuciones (y como se puede ver en el gráfico de violín) no son iguales.
Con esta información se fue capaz de validar la hipótesis planteada en un principio.

- *Análisis Multivariado*

HIPÓTESIS:

Descripción

"Es posible diferenciar perfiles de comportamiento de usuario basándose únicamente en sus métricas numéricas de navegación (actividad, duración, tasas de rebote, valor de página y día especial)."

Analizar si las métricas numéricas de navegación permiten diferenciar perfiles de usuarios es importante para entender patrones de comportamiento que no son visibles a simple vista. Sabiendo esto, se podría mejorar la personalización, la optimización del sitio y las estrategias de venta.

Estrategia de abordaje

Se optó por analizar la hipótesis a través del uso de un algoritmo de aprendizaje no supervisado Kmeans, el cual agrupa los puntos de datos de manera que los puntos dentro del mismo clúster sean lo más similares posible, minimizando la distancia entre cada punto y el centroide (o media) de su clúster.

Para esto:

- Se creó un dataset nuevo, el cual solo guarda los atributos numéricos del dataset y los mismos se estandarizaron.

- El siguiente paso fué, mediante un elbow-plot visualizar cuál sería el k(número de clusters) a utilizar para el análisis.
  A medida que se aumenta el número de clusters. Llega un punto donde agregar más clusters ya no mejora significativamente la compactación.
  A través de la visualización del mismo se llegó a la conclusión de que un número adecuado sería k=5.

- Una vez obtenido el valor k, se usó Kmeans para trabajar con el dataset refinado.

- Finalmente, mediante t-sne y como se tenía una alta dimensionalidad, se buscó reducir la dimensionalidad del problema con el fin de trasladar los resultados a dos dimensiones.

Resultados

El resultado de t-sne fueron 5 clusters, cercanos entre sí, excepto por los clusters 2 y 4. Aunque visualmente parecía observarse 5 comportamientos de usuarios marcados.
Se procedió a analizar los centroides de estos y posteriormente visualizar los mismos mediante gráficos de radar a fin de obtener una visualización de lo que estaba ocurriendo.

En base a esto se pudieron plantear las siguientes conclusiones:

Sí hay diferencias entre grupos, pero no de manera perfectamente separada.

El t-SNE deja ver que algunos clusters están muy definidos y otros forman nubes más solapadas. Lo que puede significar que existen perfiles más marcados de usuarios, y otros más mezclados.

- **Cluster 0 (azul):**

  Usuarios de baja interacción.
  En t-SNE aparecen en una zona bastante contenida.
  Radar: casi todas las métricas bajas, excepto ExitRates y BounceRates (alta).

  Interpretación: Usuarios que entran, miran poco y se van rápido.
  Es decir, no tienen una interacción activa con los sitios.

- **Cluster 1 (naranja):**

  Este es el grupo "promedio" de no compradores.
  Tienen métricas bajas en todo, pero no tan extremas como el Cluster 0.

  Interpretación: Parecen ser usuarios que simplemente navegan un poco, tienen tasas de rebote y salida moderadas (0.16 y 0.40), y se van sin generar valor (PageValues casi 0).

- **Cluster 2 (Verde):**

  En t-SNE es el grupo más marcado seguido por el cluster 0.

  SpecialDay es 1, lo cual es por lejos, el que mayor valor tiene para este atributo. A su vez poseen valores de Product Related y Product Related Duration de 0.38 y 0.36 respectivamente.

Interpretación: Este grupo representa a los usuarios que visitan el sitio cerca de una festividad o día especiales.

- **Cluster 3 (Rojo):**

  Su PageValues, Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, ProductRelated_Duration: Todos son los más altos con un valor de 1.
  A su vez tienen tasas de BounceRates y ExitRates muy bajas.

  Interpretación: Estos son los usuarios más comprometidos. Pasan tiempo en sus cuentas (admin), investigan (informational) y ven productos (product related). Su alto PageValues confirma que son los que generan ingresos.

- **Cluster 4 (Violeta):**

  Su PageValues es de 0.77 (El segundo más alto).

  BounceRates y ExitRates son los más bajos de todos.

  Interpretación: Son usuarios muy comprometidos, que no rebotan y pasan mucho tiempo en el sitio. Su PageValues es alto, lo que indica que están muy cerca de comprar, pero algo les falta para convertir. Son el objetivo perfecto para campañas de retargeting. La cual es una estrategia de publicidad digital que consiste en mostrar anuncios a usuarios que ya han interactuado previamente con una marca, como visitar un sitio web o una aplicación, pero sin haber completado una acción deseada, como una compra.

- *Conclusiones finales*

La realización de este trabajo nos permitió atravesar de principio a fin el proceso completo de un análisis de datos real, enfrentándonos a desafíos propios de la práctica profesional y desarrollando criterios para tomar decisiones fundamentadas. A lo largo del proyecto aprendimos no solo a aplicar técnicas estadísticas y de machine learning, sino también a interpretar sus resultados, cuestionar supuestos iniciales y validar hipótesis con rigor.

Explorar el dataset nos mostró la importancia de no dar por sentado que las hipótesis “intuitivas” se van a cumplir: varios supuestos que parecían razonables fueron finalmente refutados por los datos. Esto nos enseñó el valor de dejar que la evidencia guíe las conclusiones y no al revés. También pudimos comprobar que la limpieza, la detección de outliers, el manejo de duplicados y la elección entre métodos paramétricos y no paramétricos son pasos fundamentales que pueden cambiar completamente la interpretación de un análisis.

Otra experiencia valiosa fue trabajar con técnicas multivariadas como K-Means y t-SNE. Allí comprendimos que no basta con correr un algoritmo: es necesario justificar la elección de parámetros, interpretar visualizaciones y evaluar cuándo un clustering es útil o cuándo sus resultados tienen limitaciones. Aprendimos también que el análisis no supervisado no garantiza separaciones perfectas, pero aun así puede revelar patrones importantes para la toma de decisiones.

Finalmente, el hecho de integrar distintos enfoques (exploratorio, univariado, bivariado y multivariado) nos dio una visión más completa del proceso analítico y reforzó la idea de que la ciencia de datos combina herramientas técnicas con pensamiento crítico. En conjunto, este proyecto nos dejó como aprendizaje principal que trabajar con datos requiere curiosidad, disciplina, criterio y la capacidad de iterar hasta encontrar interpretaciones sólidas y fundamentadas.