INFORME ENTREGA 1

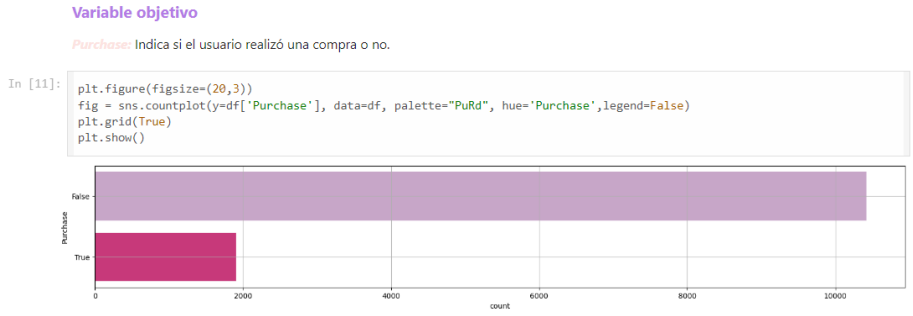
EXPLORACIÓN DE DATOS

Inicialmente realizamos un análisis exploratorio inicial de los datos conociendo el tipo de datos que contiene cada variable, de los cuales identificamos 2 variables booleanas, 2 categóricas y 14 numéricas.

Mediante el **Análisis Univariado** de las **variables numéricas** podemos identificar que el tiempo promedio de duración de las visitas a páginas relacionadas con productos es de aprox. 1194.75 seg (cerca de 19,91minutos), con una desviación estándar alta de 1913.67 seg. Esto indica una gran variabilidad en cuanto al tiempo que los usuarios pasan en estás páginas, entre visitas cortas y prolongadas. Los índices de rebote (Bounce Rates) y salida (Exit Rates) son indicadores importantes de la experiencia del usuario en un sitio web. En este conjunto de datos, el promedio de rebote es bajo, alrededor del 2.22%, lo que sugiere que la mayoría de los usuarios no abandonan el sitio inmediatamente después de acceder a él, el promedio de salida es del 4.31%, lo que indica que un porcentaje considerable de usuarios abandonan el sitio después de visitar múltiples páginas.

Por otro lado, analizando la variable 'PageValues', observamos que el valor promedio de las páginas visitadas es de aproximadamente 5.89, lo que puede interpretarse como la contribución de una página en el alcance de objetivos específicos por ejemplo ventas. Un valor alto indica que las páginas tienen un impacto significativo en los resultados del sitio.

Respecto a las **variables categóricas** observamos que el mes que presenta más frecuencia en las observaciones es mayo, seguido por noviembre, marzo y diciembre; el visitante más común es “Returning visitor” con 10551 veces, lo que indica que la mayoría de los visitantes son usuarios que han regresado al sitio web; también se observa que la mayoría de las visitas (9462 veces) no ocurrieron en un fin de semana, lo que sugiere que el sitio web tiene más tráfico durante los días de semana que los fines de semana; y a partir de estar visitas analizamos la variable “Purchase”, que nos indica si se realizó compra durante la visita al sitio web, notamos que el valor más frecuente es “False”, porque la mayoría de las visitas no finalizaron en una compra.



De la **Variable Objetivo** ¨Purchase”, nos indica si el usuario realizó compra o no, en el gráfico es evidente la diferencia entre la cantidad de usuarios que realizaron compra y los que no, lo que nos lleva a concluir que nos encontramos frente a un problema de clases desbalanceadas.

REDEFINICIÓN DE VARIABLES

Estas variables parecen ser numéricas, pero, al analizar sus valores únicos, se descubre que en realidad contienen un conjunto limitado de valores que se repiten. Por lo tanto, se decidió que estas variables serán tratadas como variables categóricas en lugar de numéricas.

Según lo descrito en el diccionario, son características categóricas.

OperatingSystems	Sistema operativo usado por el usuario para navegar en el sitio web
Browser	Navegador usado por el usuario para navegar en el sitio web
Region	Región (ubicación geográfica personalizada) desde la cual el usuario navega en el sitio web
TrafficType	Variable que indica el tipo de trafico al cual pertenece el usuario que navega en el sitio web (por ejemplo, si llegó al sitio desde un anunció o a través de una búsqueda)

Las variables categóricas (Mes, Tipo de Visitante, Fin de Semana) revelaron relaciones importantes con las compras. Por ejemplo, las compras son más frecuentes entre visitantes que regresan y durante los días laborables en comparación con los fines de semana.

Las variables numéricas (Reviews, Duración de Reviews, Duración de Páginas Relacionadas con Productos, Tasas de Rebote, Tasas de Salida, Valor de Página, Día Especial) mostraron patrones relevantes. Por ejemplo, las compras suelen asociarse con un bajo número de reviews y tasas de rebote más bajas, indicando mayor compromiso de los usuarios antes de realizar una compra.

Matriz de correlación (ver Pretratamiento.ipynb)

Se realizó un análisis mediante la matriz de correlación entre las variables numéricas del DataFrame **df**. , de lo cual se observan **correlaciones positivas** significativas entre las variables como 'Reviews' y 'ProductRelated', 'Reviews_Duration' y ¨'ProductRelated_Duration', indicando que a medida que una variable aumenta, la otra también tiende a aumentar. Esto podría indicar que los usuarios que interactúan más con las revisiones de productos también tienden a pasar más tiempo en páginas relacionadas con productos. Se observa una **correlación negativa** entre 'BounceRates' y 'PageValues', lo que sugiere que a medida que la tasa de rebote disminuye (es decir, los usuarios permanecen más tiempo en el sitio), el valor de la página tiende a aumentar. Esto indica una relación inversa entre la tasa de rebote y la contribución de las páginas al valor del sitio.

Algunas variables muestran **correlaciones débiles o nulas** entre sí, como 'SpecialDay' con otras variables. Esto indica que los días especiales no tienen una correlación fuerte con otras métricas analizadas en este conjunto de datos.

LIMPIEZA Y TRANSFORMACIÓN DE LOS DATOS

Se identificaron los datos faltantes en el DataFrame df, y se encontró que no había ninguno, es decir, todas las celdas estaban completas.

Se identificaron y eliminaron 125 filas duplicadas en el DataFrame. Después de eliminar los duplicados, se verificó que ya no quedaban filas duplicadas en el conjunto de datos.

Se identificaron datos atípicos en las variables numéricas utilizando diagramas de caja (boxplots) y el rango intercuartílico (IQR). Las variables con datos atípicos fueron: 'Reviews', 'Reviews_Duration',

'Informational', 'Informational_Duration', 'ProductRelated', 'ProductRelated_Duration', 'BounceRates', 'ExitRates', 'PageValues'. Se calculó el rango intercuartílico (RIC), el mínimo y máximo aceptables, y se identificaron la cantidad y el porcentaje de outliers para cada variable.

Se generó una nueva matriz de correlación entre las variables numéricas relacionadas con la interacción de los usuarios en el sitio web ('Reviews', 'Reviews_Duration', 'Informational', 'Informational_Duration', 'ProductRelated', 'ProductRelated_Duration'). La correlación muestra la fuerza y la dirección de la relación entre estas variables. Tras notar la gran cantidad de outliers y la alta correlación entre las variables, decidimos no tener en cuenta las variables que respecten al tiempo: ..._Duration en el tratamiento de datos atípicos.

Después de identificar y tratar los datos atípicos en el conjunto de datos, se realizaron las siguientes acciones: Identificación de Datos Atípicos Extremos, tratamiento de datos atípicos, efecto del tratamiento y finalmente la visualización de los resultados. El DataFrame final después del tratamiento de datos atípicos tiene 12205 entradas y 18 columnas.

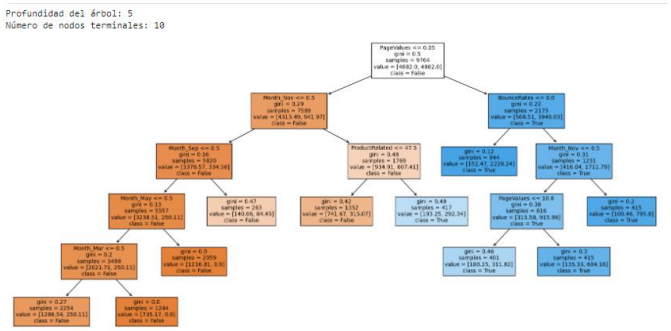
Selección de Variables por Métodos de Wrapper:

Se aplicaron dos enfoques de selección de características: Recursive Feature Elimination (RFE) y Sequential Feature Selector (SFS) utilizando regresión logística. En RFE se seleccionaron 10 características, pero no tuvo en cuenta la importancia de 'PageValues' para la variable objetivo. En SFS, se seleccionaron las variables 'PageValues', 'Month_Dec', 'Month_Mar', 'Month_May' y 'VisitorType_New_Visitor' debido a su relevancia con la variable objetivo y se utilizó un enfoque de selección secuencial.

Desbalanceo de Clases:

Se identificó un desbalanceo en las clases de la variable objetivo 'Purchase', con una gran diferencia entre las clases 0 y 1. Se realizó un submuestreo aleatorio utilizando RandomUnderSampler para equilibrar las clases, resultando en un dataset con igual cantidad de ejemplos para ambas clases (1908 ejemplos de cada clase).

Se aplican técnicas de modelado Decision Tree Classifier ,al realizar la comparación de datos originales con outliers y sin outliers, vemos que al tratamiento de estos no tiene mayor incidencia, de echo reduce un poco las métricas de desempeño, por lo tanto, se toma la decisión de seguir el resto de modelos con la base de datos original sin tratamiento de outliers extremos.



Modelo 1: Decision Tree Classifier con datos originales sin balancear, el modelo de Árbol de Decisión entrenado con datos originales sin balancear muestra un desempeño aceptable en términos de precisión y recall tanto en el conjunto de entrenamiento como en el conjunto de pruebas. Aunque la precisión es buena para la clase mayoritaria, el recall para la clase minoritaria es relativamente bajo, lo que indica que el modelo puede tener dificultades para identificar correctamente los casos positivos de esa clase.

Modelo 2: Random Forest Classifier, el modelo Random Forest Classifier entrenado con datos balanceados presenta un buen desempeño en términos de precisión y recall tanto en el conjunto de entrenamiento como en el de pruebas. Aunque la precisión es alta para la clase mayoritaria, el modelo también logra un recall significativo para la clase minoritaria, lo que indica una capacidad razonable para identificar correctamente los casos positivos de esa clase.

Modelo 3: Tunning de Hiperparámetros, el proceso de ajuste de hiperparámetros utilizando RandomizedSearchCV ha mejorado el desempeño del modelo RandomForestClassifier en términos de precisión, recall y f1-score tanto en el conjunto de entrenamiento como en el de pruebas. El mejor modelo obtenido después de la búsqueda aleatoria muestra una mejora significativa en la capacidad de generalización y en la identificación de casos positivos de la clase minoritaria, lo que demuestra la eficacia de la optimización de hiperparámetros en la mejora del rendimiento del modelo.

Modelo 4 Gradient Boosting Classifier, el modelo Gradient Boosting Classifier ha demostrado un buen desempeño tanto en el conjunto de entrenamiento como en el de pruebas, con métricas sólidas de precisión, recall y f1-score. Esto indica que el modelo es capaz de predecir con precisión tanto los casos positivos como los negativos, lo que sugiere su eficacia en la clasificación de los datos.

Modelo 5: Gradient Boosting Classifier con Tuning de Hiperparámetros, el modelo Gradient Boosting Classifier con tuning de hiperparámetros ha logrado mejorar significativamente su desempeño en comparación con el modelo base. Con la optimización de los hiperparámetros, se ha obtenido un aumento en las métricas de precisión, recall y f1-score tanto en el conjunto de entrenamiento como en el de pruebas. Esto indica que el modelo ajustado es más efectivo en la clasificación de los datos y presenta una mayor capacidad para generalizar a nuevas instancias.

Modelo 6: XGB Classifier, el modelo XGB Classifier ha demostrado un buen desempeño en la clasificación de los datos, con una precisión y recall equilibrados tanto en el conjunto de entrenamiento como en el de pruebas. Si bien su f1-score podría mejorarse, su capacidad para identificar correctamente ambas clases es notable. Esto sugiere que el modelo tiene potencial para mejorar aún más con ajustes adicionales en los hiperparámetros o técnicas de optimización.

Modelo 7: XGB Classifier Tuning de hiperparámetros, el modelo XGB Classifier con ajuste de hiperparámetros ha logrado un rendimiento sobresaliente en la clasificación de los datos, con una precisión y recall notables tanto en el conjunto de entrenamiento como en el de pruebas. Destaca especialmente su capacidad para identificar correctamente la clase positiva (1), con un recall del 100% en el conjunto de pruebas. Esto sugiere que el modelo optimizado es altamente efectivo para detectar casos positivos en la predicción.

Basándonos en la métrica de "Recall" para la clase positiva, se presentan las siguientes análisis y conclusión final, después de realizar los modelos anteriores:

- GradientBoostingClassifier obtuvo la puntuación de recall más alta tanto en el conjunto de entrenamiento (aproximadamente 89.89%) como en el conjunto de prueba (aproximadamente 89.88%). Esto sugiere que este modelo generaliza bien y tiene un buen rendimiento tanto en datos conocidos como en datos desconocidos, lo que indica una capacidad robusta de generalización.
- XGB Classifier también mostró un rendimiento sólido, con una puntuación de recall cercana en ambos conjuntos de datos (alrededor del 88% en el conjunto de entrenamiento y 87.65% en el conjunto de prueba). Aunque ligeramente inferior al GradientBoostingClassifier, sigue siendo una opción prometedora.
- DecisionTreeClassifier (DTC) y Tunning de Random FC (RandomForestClassifier) también demostraron un rendimiento decente en ambas métricas de recall, aunque ligeramente inferiores a los modelos anteriores. Esto sugiere que estos modelos pueden ser considerados como opciones viables, especialmente si se tienen en cuenta otros factores como la interpretabilidad del modelo y los recursos computacionales requeridos.
- Tunning GBC (GradientBoostingClassifier) y Tunning XGB (XGB Classifier), muestra una puntuación de recall muy alta de 1.0 tanto en el conjunto de entreamiento como de prueba, por lo que se podria decir que el modelo esta sobreajustado (overfitting).

Por lo tanto concluimos que los modelos ajustados con hiperparámetros no siempre garantizan una mejora en el rendimiento del modelo y que el mejor modelo fue con GradientBoostingClassifier, siendo este el que arrojo una mejor metrica en sencibilidad pudiendo identicar en buena proporción las instancias positivas tanto con los datos de entrenamiento y como en los de prueba. Sin embargo, vimos que los modelos tienden a presentar dificultades para la clase minoritaria, siendo esta la de interes, por lo que se sugiere recopilar mas datos donde se tengan mas valores sobre la clase positiva y poder demerminar unas caracteristicas mucho mas completas y determinar patrones mas acertados para lograr el objetivo de precedir con buenos resultados el comportamiento en el comercio electronico.