Haz click en el siguiente material de apoyo, revisa detenidamente la información y resuelve el proyecto: predicción de retorno de clientes.

|  |  |
| --- | --- |
|  |  |
| **Guia** | **Tabla** |

Recursos

* [***Conjunto de datos***](https://www.kaggle.com/datasets/julihocc/customers)***para el proyecto*** https://www.kaggle.com/datasets/julihocc/customers
* [***Proyecto de ejempl***](https://www.kaggle.com/code/julihocc/classification-models-for-titanic-competition)***o (Titanic) https://www.kaggle.com/code/julihocc/classification-models-for-titanic-competition***

**Instrucciones**  
  
**1. Cargar el archivo CSV synthetic\_customer\_data.csv**en un DataFrame usando Pandas y mostrar las primeras filas para comprender la estructura de los datos. **2. Mostrar información básica del DataFrame, como el tipo de datos de cada columna y calcular estadísticas descriptivas para cada variable.**  
  
**3. Visualización de Datos:**

* + - Generar histogramas para analizar la distribución de age y total\_spent.
    - Crear un gráfico de barras para observar la distribución de gender.

**4. Preprocesamiento de los Datos:**

* + - Verificar si existen valores nulos en el conjunto de datos y documentar los resultados.
    - Codificar la variable gender en valores numéricos (0 para Female, 1 para Male).
    - Escalar las columnas numéricas (age, total\_spent, frequency, days\_since\_last\_purchase) usando StandardScaler para asegurar consistencia en las escalas.

**5. Dividir los Datos en Entrenamiento y Prueba:**

* + - Definir las características (X) y la variable objetivo (y), que es returned\_next\_month.
    - Dividir los datos en un conjunto de entrenamiento y otro de prueba (80/20).

**6. Entrenamiento de Modelos:**

* + - Seleccionar tres modelos de clasificación: Regresión Logística, Árbol de Decisión y Bosque Aleatorio.
    - Entrenar cada modelo usando los datos de entrenamiento.

**7. Evaluación Inicial de Modelos:**

* + - Evaluar cada modelo en el conjunto de prueba usando métricas de rendimiento: exactitud, precisión, recall, F1 y AUC-ROC.
    - Comparar los resultados y documentarlos para seleccionar el mejor modelo.

**8. Optimización del Mejor Modelo:**

* + - Realizar optimización de hiperparámetros (Grid Search) en el modelo seleccionado (en este caso, Bosque Aleatorio).
    - Ajustar el modelo con los mejores parámetros encontrados y entrenarlo nuevamente.

**9. Validación y Análisis de Errores:**

* + - Evaluar el rendimiento del modelo optimizado en el conjunto de prueba con las métricas mencionadas.
    - Generar la matriz de confusión para identificar y analizar los errores de clasificación.

**10. Conclusión y Recomendaciones:**

* + - Redactar conclusiones sobre el rendimiento del modelo y propuestas de aplicación en estrategias de retención de clientes.
    - Documentar el proceso y resultados en un Jupyter Notebook con comentarios explicativos y visualizaciones relevantes.