Contexto del Negocio

Una empresa digital quiere mejorar la detección de fraude y optimizar la retención de usuarios. Se cuenta con tres fuentes principales de datos: información de usuarios, eventos de plataforma y transacciones de pagos. El objetivo es analizar el comportamiento, generar modelos predictivos y proponer soluciones accionables.

Archivo	Descripción	N° Registros
usuarios.csv	Información demográfica y de adquisición de usuarios	5,000
eventos.csv	Eventos registrados en la plataforma (login, compra, cancelación)	25,293
pagos.csv	Histórico de pagos con etiqueta de fraude (0=no, 1=sí)	20,000

1.**SQL & Modelado de Datos**

Tablas disponibles:

- usuarios(user_id, pais, fecha_registro, canal_adquisicion)
- eventos(user_id, fecha_evento, tipo_evento, dispositivo) *(tipo_evento ∈{login, compra, cancelación})*
- pagos(pago_id, user_id, fecha_pago, monto, metodo_pago)

Ejercicios:

1. Construye una consulta que calcule el LTV (lifetime value) promedio de los usuarios por canal de adquisición.
2. Encuentra los usuarios activos en los últimos 90 días que tienen más de un método de pago distinto.
3. Diseña una consulta que devuelva la tasa de conversión semanal: usuarios que hicieron login usuarios que hicieron compra en la misma semana.
4. Explica cómo normalizarías este esquema de datos si tuvieras que escalar a 50M de usuarios y 2B de eventos.

2. **Feature Engineering + EDA**

Dataset: Proporcionar un dataset.

Tareas:

1. Genera nuevas features a partir de la información temporal
2. Detecta variables altamente correlacionadas y explica si deberías eliminarlas o transformarlas.
3. Identifica sesgos en los datos.
4. Propón 3 hipótesis de negocio que tu modelo debería poder responder.

3. **Machine Learning + Interpretabilidad**

Caso: Predecir probabilidad de fraude en pagos.

Tareas:

1. Divide los datos en train/test y aplica estratificación por clase.
2. Entrena 3 algoritmos distintos
3. Compara resultados con métricas adecuadas para clases desbalanceadas.
4. Usa técnicas de interpretabilidad, para explicar las 5 variables más influyentes.
5. Diseña un pipeline reproducible (con scikit-learn Pipelines o MLflow).

4. **Puesta en Producción + Escalabilidad**

Caso: Tu modelo de fraude funcionó bien offline. Ahora debe integrarse en tiempo real.

Tareas:

1. Explica cómo diseñarías una arquitectura en AWS/GCP para que el modelo se ejecute en streaming sobre pagos en tiempo real.
2. ¿Cómo manejarías el drift de datos y modelos en producción?
3. Diseña un esquema básico de monitoreo de métricas de negocio (ejemplo: ratio de fraude detectado vs. pérdidas por falsos positivos).

5. **Comunicación de Resultados**

Caso: Debes presentarle al CFO de la empresa los resultados del modelo de fraude.

Tareas:

- Explica en un reporte ejecutivo de máximo 1 página:
    - Riesgos detectados.

- Impacto económico estimado.
- Recomendaciones accionables.