Propuesta de Proyecto: Detección de Fraude en Transacciones Bancarias

# 1. Comprensión del Negocio (Business Understanding)

Objetivo del negocio:  
El banco desea reducir sus pérdidas financieras derivadas de transacciones fraudulentas en un 80% en el primer año de implementación del sistema. Actualmente, las pérdidas ascienden a $10 millones anuales, por lo que el éxito se medirá si se logra reducir estas pérdidas a $2 millones o menos al año. Asimismo, el banco busca mejorar la experiencia del cliente, disminuyendo en un 50% los falsos positivos (transacciones legítimas marcadas erróneamente como fraudulentas).

Métrica de éxito del negocio:  
- Reducción de las pérdidas por fraude en un 80%.  
- Reducción de los falsos positivos en un 50%.

Objetivo del proyecto:  
Desarrollar un modelo de Machine Learning que logre:  
- Precisión superior al 95%.  
- Recall superior al 90%, para garantizar una alta detección de fraudes y minimizar los falsos negativos.

Duración: 1 semana  
Fecha estimada: 1 - 7 octubre  
Actividades:  
- Reuniones con stakeholders para identificar necesidades y definir métricas clave.  
- Documentación del impacto financiero actual del fraude y los objetivos financieros.

# 2. Comprensión de los Datos (Data Understanding)

Fuentes de datos:  
- Historial de transacciones (almacenadas en bases de datos locales o en la nube).  
- Datos en tiempo real provenientes de sistemas de transacciones en línea (streaming de eventos).  
- Información adicional de clientes (edad, historial de uso, dispositivos utilizados).

Duración: 2 semanas  
Fecha estimada: 8 - 21 octubre  
Actividades:  
- Ingesta de datos históricos usando Azure Data Factory (ADF) y almacenamiento en Azure Data Lake Storage (ADLS).  
- Ingesta de datos en tiempo real mediante Azure Event Hubs.  
- Exploración de los datos para entender la distribución de las transacciones fraudulentas.  
- Identificación de valores nulos, duplicados y atípicos.

# 3. Preparación de los Datos (Data Preparation)

Tareas principales:  
- Limpieza de datos: eliminación de duplicados y tratamiento de valores faltantes.  
- Codificación de variables categóricas (p. ej., tipo de dispositivo, ubicación de la transacción).  
- Normalización de variables numéricas (p. ej., monto de la transacción).  
- Balanceo de datos: utilización de técnicas como SMOTE para corregir el desbalance en las clases (fraudes vs. transacciones legítimas).

Duración: 3 semanas  
Fecha estimada: 22 octubre - 11 noviembre  
Actividades:  
- Procesamiento y transformación de datos utilizando Azure Databricks.  
- Almacenamiento de los datos procesados en ADLS para su uso posterior en la fase de modelado.

# 4. Modelado (Modeling)

Elección de algoritmos:  
- Modelos de clasificación como Árboles de Decisión, Random Forest, Gradient Boosting, y Redes Neuronales.  
Entrenamiento:  
- Dividir el dataset en conjunto de entrenamiento (80%) y prueba (20%).  
- Utilización de Azure Machine Learning (AML) para entrenar los modelos.  
- Ajuste de hiperparámetros mediante técnicas como la búsqueda en cuadrícula y validación cruzada.

Duración: 4 semanas  
Fecha estimada: 12 noviembre - 9 diciembre  
Actividades:  
- Entrenamiento del modelo en clusters de AML, optimizando el rendimiento del modelo en cuanto a precisión y recall.  
- Versionado y almacenamiento de los modelos en el Azure ML Model Registry para facilitar su posterior despliegue.

# 5. Evaluación (Evaluation)

Evaluación del modelo:  
- Medir el rendimiento del modelo en el conjunto de datos de prueba.  
- Evaluar las métricas clave: precisión, recall, F1-score, tasa de falsos positivos y falsos negativos.  
- Seleccionar el modelo que mejor balance entre precisión y recall.

Duración: 2 semanas  
Fecha estimada: 10 - 23 diciembre  
Actividades:  
- Comparar los resultados obtenidos por diferentes modelos (Random Forest, Gradient Boosting, etc.).  
- Seleccionar el mejor modelo en base a su desempeño en el conjunto de prueba.

# 6. Despliegue (Deployment)

Despliegue del modelo en producción:  
- Desplegar el modelo en Azure Kubernetes Service (AKS) o Azure Container Instances (ACI) para su ejecución en tiempo real.  
Integración:  
- Exponer el modelo a través de una API gestionada con Azure API Management, para que el sistema bancario pueda consumir el servicio de predicción de fraude en tiempo real.

Duración: 3 semanas  
Fecha estimada: 24 diciembre - 13 enero  
Actividades:  
- Despliegue de contenedores que contengan el modelo entrenado.  
- Integración del modelo con el sistema de transacciones bancarias a través de la API.

# 7. Monitorización y Mantenimiento (Monitoring and Maintenance)

Monitorización del rendimiento del modelo en producción:  
- Utilización de Azure Application Insights para monitorear la precisión y recall del modelo en tiempo real.  
- Configuración de alertas si el rendimiento del modelo disminuye (p. ej., si la precisión cae por debajo del 90%).  
Actualización del modelo:  
- Automatización de pipelines para reentrenar el modelo cuando se reciban nuevos datos a través de Azure DevOps.

Duración: 2 semanas  
Fecha estimada: 14 - 27 enero  
Actividades:  
- Configuración de métricas y alertas con Application Insights.  
- Creación de pipelines automáticos para reentrenar el modelo si cambian los patrones de fraude.

# Presupuesto en Azure

1. Azure Data Factory (ADF): $150/mes.  
2. Azure Event Hubs: $100/mes.  
3. Azure Data Lake Storage (ADLS): $200/mes.  
4. Azure Databricks: $280/mes.  
5. Azure Machine Learning (AML): $250/mes.  
6. Azure Kubernetes Service (AKS): $360/mes.  
7. Azure API Management: $50/mes.  
8. Azure Application Insights: $13.80/mes.

Costo total aproximado mensual: $1,403.80/mes.

# Estimación Económica o Financiera del Impacto en el Negocio

1. Reducción de pérdidas por fraude:  
Actualmente, el banco pierde $10 millones anualmente por fraudes. Si el sistema reduce estas pérdidas en un 80%, las pérdidas disminuirán a $2 millones, lo que representa un ahorro de $8 millones anuales.  
2. Reducción de falsos positivos:  
Los falsos positivos afectan negativamente la experiencia del cliente. Si se reducen en un 50%, se espera que la retención de clientes mejore, lo que podría generar ingresos adicionales de $500,000 anuales debido a la mayor lealtad y uso de productos bancarios.  
Total de beneficios anuales estimados:  
- Ahorro por reducción de fraudes: $8 millones.  
- Impacto en ingresos por mejora en la experiencia del cliente: $500,000.  
- Beneficio total estimado: $8.5 millones anuales.