Propuesta de Proyecto Final de Ciencia de Datos

Gabriela Mariel Lopez Armenta
Proyecto: Predicción de Deserción de Clientes Bancarios (Churn)

4 de Noviembre de 2025

---

1. Título del Proyecto

Predicción de la Deserción de Clientes Bancarios (*Churn*) mediante Modelos de Machine Learning para la Optimización de Estrategias de Retención.

1.1. Problemática y Justificación

La deserción de clientes (*Churn*) es el principal desafío de rentabilidad en el sector financiero. Este proyecto busca no solo predecir qué clientes abandonarán el banco, sino **identificar los factores subyacentes** que impulsan esta decisión, proporcionando información accionable para campañas de retención. El problema es de naturaleza de **clasificación binaria**.

2. Pregunta de Investigación (Q.I.)

¿Cuáles son los **factores demográficos, financieros y de compromiso** que tienen el mayor impacto en la probabilidad de deserción de un cliente bancario, y cuál es el modelo de *Machine Learning* más eficiente (medido por *AUC* y *Recall*) para identificar a los clientes de alto riesgo?

3. Hipótesis

La deserción de clientes está impulsada principalmente por un **bajo nivel de compromiso financiero** con la entidad. Específicamente, se postula que la **combinación de un Saldo bajo, una baja Puntuación Crediticia y un número reducido de productos** será el predictor más robusto de *Churn*.

4. Fuente de Datos

- **Dataset:** Datos de *Churn Modelling* (10,000 entradas) de una entidad bancaria, disponible en plataformas públicas como Kaggle.

- **Variables Clave:** `CreditScore`, `Age`, `Balance`, `NumOfProducts`, `IsActiveMember` y la variable objetivo `Exited` (1=Churn, 0=No Churn).

5. Metodología de Desarrollo

El proyecto se desarrollará bajo un enfoque de modelado de **Clasificación**, utilizando el lenguaje Python y las librerías estándar de Ciencia de Datos (`Pandas`, `Scikit-learn`).

5.1. Fases del Proyecto

1. **Preparación de Datos:** Eliminación de identificadores (`CustomerID`). **Codificación** (*One-Hot Encoding*) de variables categóricas (`Geography`, `Gender`).
2. **Preprocesamiento:** División de los datos en Entrenamiento/Prueba (80/20). **Escalado** de variables numéricas mediante `StandardScaler` para evitar el sesgo por magnitud.
3. **Modelado y Comparación:** Se implementarán y compararán tres modelos de diversa complejidad:
   - **Modelo Base:** Regresión Logística.
   - **Modelos Avanzados:** Random Forest y **XGBoost**.

5.2. Criterios de Éxito y Evaluación

El desbalance de clases (*approx.* 80 % No-Churn vs. 20 % Churn) exige priorizar métricas que midan la capacidad del modelo para identificar a la minoría:

- **Métrica Primaria: AUC** (Área bajo la curva ROC), que mide la capacidad general del modelo para distinguir las dos clases.
- **Métrica Crítica: Recall** para la Clase 1 (*Churn*), que mide qué porcentaje de clientes en riesgo real fue identificado.
- **Criterio de Éxito:** Un **AUC superior a 0.85** y un **Recall** para la clase *Churn* de al menos **0.50**, superando el modelo base de Regresión Logística.

6. Anticipación de Resultados (Resultados Preliminares)

La experimentación inicial de los modelos arrojó los siguientes resultados comparativos:

Cuadro 1: Tabla Comparativa de Métricas de Modelos (Clase 1: Churn)

Modelo	**Accuracy**	**Recall**	**Precision**	**AUC Score**
Regresión Logística	0.81	0.20	0.55	0.7789
Random Forest	0.87	0.47	0.76	**0.8653**
XGBoost (Ganador)	**0.87**	**0.55**	0.72	0.8502

Conclusión Anticipada: El modelo **XGBoost** será seleccionado como la solución final, ya que su **Recall de 0.55** es el más alto, superando el criterio de éxito (*Recall* $\geq$ 0,50). Esto indica la mayor capacidad para identificar clientes en riesgo, combinada con una alta *Accuracy*.

Factores de Riesgo Anticipados: La **Edad** sigue siendo el factor más dominante en la predicción, pero la validación de la hipótesis se centrará en la fuerte influencia conjunta de `Balance`, `NumOfProducts` y `CreditScore`. Esto demuestra que los factores de compromiso y riesgo financiero son los principales impulsores del *Churn*.