ANÁLISIS DE CLUSTERING Y MODELADO DE DESERCIÓN DE CLIENTES BANCARIOS

Lilian Marcela Ramírez Ramírez
Hans Silva Rodríguez
Omar Javier Ramírez Rodríguez

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
Cristian Alejandro Torres

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, OCTUBRE 13 DE 2025

FICHA RESUMEN
PROYECTO DE TRABAJO DE GRADO

ANÁLISIS DE CLUSTERING Y MODELADO DE DESERCIÓN DE CLIENTES BANCARIOS

1. **ÁREA DE TRABAJO:** Ciencia de Datos e Inteligencia Artificial aplicada al sector financiero
2. **TIPO DE PROYECTO:** Investigación
3. **ESTUDIANTE(S):**
   - ✓ Lilian Marcela Ramírez Ramírez
   - ✓ Hans Silva Rodríguez
   - ✓ Omar Javier Ramírez Rodríguez
4. CORREO ELECTRÓNICO:
   - ✓ lilians@javerianacali.edu.co
   - ✓ hsilvar88@javerianacali.edu.co
   - ✓ fisher12quest@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO: Banco Mundo Mujer Carrera 11 #5-56 Barrio Valencia – Popayán
6. DIRECTOR: Cristian Alejandro Torres
7. VINCULACIÓN DEL DIRECTOR: Docente Pontificia Universidad Javeriana - Cali / Departamento de Electrónica y Ciencias de la Computación / Facultad de Ingeniería y ciencias
8. CORREO ELECTRÓNICO DEL DIRECTOR: cristian.torres@javerianacali.edu.co
9. PALABRAS CLAVE: Churn bancario, IRFC, índice compuesto, data leakage, K-Prototypes, clustering para datos mixtos, Regresión Logística, Random Forest, XGBoost, AUC-ROC, AUC-PR, Recall Top Decile, Lift, Gain Chart, retención de clientes.
10. FECHA DE INICIO: 20 mayo 2025
11. DURACIÓN ESTIMADA: 12 meses

12. RESUMEN:

El proyecto aplicado "Análisis de Clustering y Modelado de Deserción de Clientes Bancarios" desarrolla un enfoque integral para el análisis del churn bancario, articulando técnicas de ingeniería de variables, segmentación de clientes y aprendizaje supervisado con el propósito de comprender, priorizar y predecir el riesgo de abandono bajo criterios de validez metodológica y aplicabilidad analítica.

Inicialmente, se planteó la construcción de un Índice de Riesgo Financiero-Comportamental (IRFC) como herramienta para sintetizar múltiples dimensiones financieras y comportamentales del cliente. No obstante, durante el proceso de evaluación se identificaron métricas anormalmente altas, lo que condujo a una revisión crítica del modelo y a la detección de fuga de información asociada a la variable complain. La exclusión de esta variable permitió construir escenarios más realistas y metodológicamente más sólidos, evidenciando la importancia de cuestionar resultados aparentemente ideales en proyectos de ciencia de datos aplicada.

De manera complementaria, se implementó un proceso de segmentación mediante clustering, orientado a identificar perfiles diferenciados de clientes según patrones financieros y comportamentales. Esta segmentación permitió estructurar grupos con distintas exposiciones relativas al riesgo, fortaleciendo procesos de perfilamiento, priorización y comprensión estratégica del churn.

En la etapa de modelado supervisado se evaluaron múltiples escenarios comparativos, concluyendo que el IRFC no resulta adecuado como predictor directo debido a colinealidad con variables base, aunque conserva valor como herramienta de interpretación y ranking. Asimismo, la segmentación mostró capacidad para aportar valor analítico adicional en escenarios comparativos; sin embargo, su principal contribución se orienta al fortalecimiento de procesos de segmentación estratégica más que a su adopción automática como predictor final.

En conjunto, los resultados demuestran que el principal valor del proyecto no radica únicamente en la obtención de métricas predictivas elevadas, sino en la construcción de un enfoque analítico metodológicamente sólido, capaz de diferenciar entre desempeño aparente, validez metodológica e interpretación estratégica.

En conclusión, esta investigación evidencia que la integración de analítica avanzada permite no solo mejorar la comprensión del churn, sino también fortalecer la toma de decisiones mediante herramientas complementarias de predicción, segmentación y priorización, facilitando estrategias de retención más focalizadas y metodológicamente defendibles en contextos financieros reales.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

En el contexto actual del sector financiero, la retención de clientes se ha consolidado como un factor estratégico para la sostenibilidad y competitividad de las organizaciones. La deserción de clientes (*churn*) no solo representa pérdida de ingresos futuros, sino también un incremento en los costos asociados a adquisición, fidelización y reposición de clientes, lo que hace fundamental el desarrollo de herramientas analíticas capaces de anticipar este fenómeno y apoyar decisiones estratégicas de retención [1, 26].

En este escenario, las técnicas de analítica de datos y aprendizaje automático han adquirido una relevancia creciente como soporte para la comprensión y predicción del comportamiento del cliente [19, 20]. No obstante, la construcción de modelos predictivos metodológicamente sólidos enfrenta desafíos importantes, entre ellos la calidad de los datos, la colinealidad entre variables, la redundancia informativa y la posible presencia de fuga de información, factores que pueden generar resultados artificialmente optimistas y comprometer la validez de los modelos en escenarios reales. Bajo esta perspectiva, el presente trabajo desarrolla un enfoque integral para el análisis del churn bancario, articulando tres dimensiones complementarias: (i) la construcción de un Índice de Riesgo Financiero-Comportamental (IRFC) como herramienta de síntesis e interpretación del riesgo, (ii) la segmentación de clientes mediante técnicas de clustering orientadas al perfilamiento estratégico y (iii) la evaluación de modelos supervisados de clasificación para analizar capacidad predictiva bajo distintos escenarios metodológicos [11, 20].

Durante el desarrollo del estudio surgió un hallazgo crítico asociado a la variable *complain*, cuya influencia generaba métricas predictivas artificialmente altas. Este resultado condujo a una revisión profunda del proceso de modelado, permitiendo identificar un problema de fuga de información y reorientando el estudio hacia una lógica centrada en validez metodológica, realismo analítico y aplicabilidad práctica, más allá del desempeño aparente.

En este contexto, la segmentación de clientes adquiere relevancia como herramienta de comprensión estructural, al permitir identificar patrones latentes de comportamiento financiero y comercial que facilitan procesos de perfilamiento, priorización y análisis estratégico del riesgo. De esta manera, el clustering se incorpora no únicamente como mecanismo de mejora comparativa, sino como una capa analítica orientada a fortalecer la interpretación del churn [14, 18].

El objetivo general de esta investigación consiste en desarrollar y validar un enfoque analítico integral para el estudio de la deserción de clientes bancarios, combinando técnicas de ingeniería de variables, segmentación y modelado supervisado bajo criterios de validez metodológica, interpretación estratégica y potencial aplicabilidad en contextos reales. La relevancia del estudio radica en que no se limita a abordar el churn desde una perspectiva exclusivamente predictiva, sino que incorpora una visión crítica sobre la construcción, evaluación y uso de modelos analíticos, incluyendo aspectos como detección de fuga de información, análisis de colinealidad,

segmentación estratégica y priorización operativa. En este sentido, el trabajo contribuye al desarrollo de enfoques más confiables, interpretables y estratégicamente útiles para la toma de decisiones en el sector financiero [20, 26, 28].

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1 Definición y Planteamiento del problema

En Colombia, la deserción de clientes bancarios (*churn*) constituye un desafío estratégico para la sostenibilidad del sector financiero. Estudios como el de **Boston Consulting Group** señalan que el 56 % de los clientes mantiene relaciones con más de un banco, reflejando alta movilidad y riesgo de pérdida [1]. Además, perder un cliente existente resulta significativamente más costoso que retenerlo, lo que convierte al churn en un problema crítico [2].

Las tasas de rotación superan el 24 % anual, generando pérdidas en ingresos y costos de adquisición hasta siete veces mayores que los de retención [1]. Este escenario se agrava por la irrupción de *fintechs* y la demanda de experiencias personalizadas, factores que reducen la lealtad y aumentan la competencia.

Actualmente, los modelos predictivos utilizados por muchas entidades financieras son genéricos y asumen homogeneidad en la base de clientes, lo que limita la identificación de patrones diferenciados y reduce la efectividad de las estrategias de fidelización. Esta aproximación genera intervenciones poco focalizadas, desperdicio de recursos y bajo retorno sobre la inversión (ROI).

La problemática central radica en la ausencia de metodologías que integren segmentación avanzada y modelado predictivo, permitiendo anticipar la probabilidad de abandono y comprender la heterogeneidad de los clientes para priorizar acciones según perfiles de riesgo. Sin esta visión, las entidades enfrentan un círculo vicioso: altos niveles de churn, costos crecientes y pérdida de competitividad.

En este contexto, surge la necesidad de desarrollar un enfoque basado en ciencia de datos que combine técnicas de clustering y modelos supervisados, con el fin de generar perfiles de riesgo diferenciados y mejorar la precisión en la predicción del churn, optimizando recursos y contribuyendo a la sostenibilidad del sector financiero colombiano.

1.1.1 Formulación del problema

Pregunta Central:

¿Cómo puede un enfoque basado en clustering no supervisado y modelado predictivo supervisado mejorar la identificación y gestión del riesgo de deserción de clientes bancarios en el sector financiero colombiano, en comparación con los modelos genéricos actuales?

1.1.2 Sistematización

Subpreguntas especificas:

1. ¿Qué variables financieras y comportamentales son más relevantes para construir un índice de riesgo de churn que represente adecuadamente la propensión al abandono?
2. ¿Qué técnica de clustering para datos mixtos permite obtener una segmentación coherente y accionable de clientes bancarios, considerando costo, Silhouette, Calinski-Harabasz, Davies-Bouldin y distancia de Gower?
3. ¿En qué medida la inclusión de variables derivadas como el IRFC o el cluster mejora el desempeño predictivo frente a un modelo base sin complain, y qué riesgos metodológicos genera frente a colinealidad, leakage y aplicabilidad real?
4. ¿Cómo garantizar que el modelo propuesto sea equitativo, evitando sesgos por género, región u otras características sensibles?
5. ¿Qué estrategias de retención pueden proponerse a partir de la probabilidad de churn y de perfiles de riesgo financiero, y cómo podrían validarse mediante un piloto comercial controlado?

1.2. Objetivos

1.2.1 Objetivo General

Identificar y caracterizar perfiles de riesgo de deserción en clientes bancarios mediante la construcción de un índice de riesgo financiero - comportamental, técnicas de clustering para datos mixtos y modelos supervisados de churn, con el fin de evaluar su aporte analítico, comparativo y estratégico en la priorización de clientes en riesgo.

1.2.2. Objetivos Específicos

1. Preprocesar el dataset de churn bancario mediante limpieza de datos, codificación de variables categóricas y escalado de atributos numéricos, para garantizar la calidad y consistencia de la información antes del análisis.

2. Construir y evaluar un Índice de Riesgo Financiero-Comportamental mediante variables financieras y comportamentales, diferenciando versiones con y sin variables potencialmente asociadas a fuga de información, con el fin de analizar su utilidad como herramienta de interpretación, ranking y segmentación.

3. Aplicar clustering para datos mixtos mediante K-Prototypes, evaluando el número óptimo de grupos con costo del modelo, Silhouette, Calinski-Harabasz, Davies-Bouldin y distancia de Gower, para caracterizar perfiles de riesgo accionables.

4. Evaluar modelos supervisados de churn mediante Regresión Logística, Random Forest y XGBoost, comparando escenarios con y sin variables potencialmente asociadas a fuga de información, y seleccionando el enfoque que ofrezca mejor equilibrio entre desempeño, estabilidad metodológica y aplicabilidad real.

5. Analizar las implicaciones prácticas y de fairness derivadas de los perfiles de riesgo identificados, evaluando posibles sesgos por género o geografía y proponiendo estrategias de retención segmentadas que optimicen el ROI de las campañas.

1.3 Marco de Referencia

1.3.1 Marco Teórico y Conceptual

La deserción de clientes (*customer churn*) se define como la pérdida voluntaria o involuntaria de clientes en un periodo determinado, fenómeno que afecta directamente la rentabilidad de las organizaciones. En el sector financiero, este problema se intensifica por la alta competencia y la aparición de actores digitales (*fintech*), lo que obliga a las entidades a implementar estrategias basadas en analítica avanzada para anticipar el abandono y diseñar acciones preventivas.

1.3.1.1 Aprendizaje automático aplicado al churn

El aprendizaje automático (*machine learning*) constituye un conjunto de técnicas que permiten construir modelos capaces de identificar patrones en los datos y realizar predicciones sin ser programados explícitamente para cada regla. En el contexto del churn, estos modelos buscan estimar la probabilidad de que un cliente abandone el servicio, a partir de variables financieras, comportamentales y demográficas.

Existen dos enfoques principales: aprendizaje supervisado y no supervisado.

Aprendizaje Supervisado: El aprendizaje supervisado se basa en el uso de un conjunto de datos etiquetado, donde existe una variable objetivo (por ejemplo, Exited), que indica si un cliente abandonó o no. El objetivo es aprender una función:

$$f(X) \rightarrow Y$$

Donde:

- X: conjunto de variables explicativas
- Y: variable objetivo (churn)

Es decir, la probabilidad de que un cliente abandone dado su comportamiento.

Regresión Logística: La regresión logística es uno de los modelos más utilizados en problemas de churn debido a su interpretabilidad. Este modelo estima la probabilidad de abandono mediante la función logística:

$$P(Y = 1) = 1 / (1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)})$$

Donde:

- β_i: coeficientes del modelo
- X_i: variables explicativas

En el contexto del churn, cada coeficiente indica cómo una variable influye en la probabilidad de abandono. Por ejemplo, un coeficiente positivo en una variable de riesgo financiero incrementa la probabilidad de churn.

Su principal ventaja es la interpretabilidad; sin embargo, su capacidad para capturar relaciones no lineales es limitada.

Modelos basados en árboles: Los modelos basados en árboles permiten capturar relaciones no lineales y complejas entre variables, lo que los hace especialmente útiles en problemas de churn.

- **Random Forest**: Random Forest es un modelo de ensamble que construye múltiples árboles de decisión a partir de subconjuntos aleatorios de datos y variables. La predicción final se obtiene mediante votación o promedio:

  $$\hat{Y} = (1 / T) \Sigma_{t=1}^{T} h_t(X)$$

  Donde:

  - $h_t(X)$: predicción del árbol t

- T: número total de árboles

En churn, este modelo permite identificar patrones complejos como combinaciones de variables que incrementan el riesgo de abandono. Su estructura reduce el sobreajuste y mejora la generalización.

- **XGBoost**: Es un algoritmo de *gradient boosting* que construye árboles de manera secuencial, donde cada nuevo árbol corrige los errores del anterior. El modelo busca minimizar una función de pérdida:

$$L = \Sigma_{i=1}^{n} l(y_i, \hat{y}_i) + \Sigma_{k=1}^{K} \Omega(f_k)$$

Donde:
- l: función de pérdida (por ejemplo, log-loss)
- Ω: término de regularización
- f_k: árboles del modelo

En problemas de churn, XGBoost resulta útil en este tipo de problemas por:

- Su capacidad para manejar datos desbalanceados
- Su robustez frente a ruido
- Su alto poder predictivo

Métricas de evaluación: Dado que el churn es un problema desbalanceado, no es suficiente evaluar el modelo con precisión (*accuracy*). Se utilizan métricas como:

- **AUC-ROC**: mide la capacidad del modelo para discriminar entre clases
- **AUC-PR**: relevante cuando la clase positiva es minoritaria
- **Recall**: proporción de clientes churn correctamente identificados
- **F1-score**: balance entre precisión y recall

Estas métricas permiten evaluar el desempeño del modelo en escenarios reales, donde la correcta identificación de clientes en riesgo es crítica.

Datos → Modelo supervisado → Probabilidad de churn → Decisión (retención)

Aprendizaje no supervisado: El aprendizaje no supervisado busca identificar patrones o estructuras en los datos sin utilizar una variable objetivo. En el contexto del churn, se emplea principalmente para segmentar clientes en grupos homogéneos.

- **K-Means**: K-Means es un algoritmo que agrupa observaciones minimizando la varianza intra-cluster:

$$min\ \Sigma_{k=1}{}^{K}\Sigma_{xi\in C_k}\ //x_i - \mu_k//^2$$

Donde:

- C_k: cluster k
- μ_k: centroide del cluster

En churn, permite identificar segmentos de clientes con comportamientos similares, aunque presenta limitaciones al trabajar con variables categóricas.

- **Clustering jerárquico (Ward)**: Este método construye una estructura jerárquica de clusters mediante la minimización de la varianza dentro de los grupos. Su resultado se representa mediante un dendrograma, lo que permite visualizar la formación de clusters.

En el contexto de churn, facilita la exploración de relaciones entre segmentos de clientes.

- **DBSCAN**: DBSCAN es un algoritmo basado en densidad que identifica clusters como regiones densas de puntos y detecta outliers:

  - No requiere definir el número de clusters
  - Es robusto ante ruido
  - En churn, puede ser útil para detectar clientes atípicos o comportamientos extremos.

Aplicación al churn

Las técnicas no supervisadas permiten:

- Identificar segmentos de clientes con diferente nivel de riesgo
- Detectar patrones de comportamiento no evidentes
- Complementar modelos supervisados

En particular, la combinación de segmentación y modelado predictivo permite enriquecer el análisis, integrando tanto patrones globales como probabilidades individuales de abandono.

Segmentación → Perfiles → Integración con modelo de churn

1.3.1.2 Índices compuestos y análisis factorial

La construcción de un Índice de Riesgo Financiero-Comportamental (IRFC) busca sintetizar múltiples variables (financieras y de comportamiento) en un indicador único. Inicialmente se explora mediante análisis factorial exploratorio (EFA), que evalúa la adecuación de los datos (KMO, Bartlett) y las cargas factoriales. Cuando no se cumple la estructura latente, se recurre a métodos compuestos y validación cruzada para garantizar desempeño predictivo.

1.3.1.3 Métricas de evaluación

La evaluación de modelos en problemas de churn requiere un enfoque multidimensional, ya que no basta con medir únicamente la precisión global. En contextos financieros, donde existe desbalance entre clientes que abandonan y clientes que permanecen, así como restricciones operativas para intervenir toda la base, resulta fundamental incorporar métricas estadísticas, métricas de segmentación y métricas de priorización comercial. [19, 20].

Métricas de Modelos Supervisados

- **AUC-ROC (Area bajo la curva ROC)**: La curva ROC (Receiver Operating Characteristic) representa la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) para distintos umbrales de clasificación [16].

$$TPR = TP / (TP + FN); FPR = FP / (FP + TN)$$

  El área bajo esta curva (AUC-ROC) se define como:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

Esta métrica mide la capacidad del modelo para discriminar entre clientes que abandonan y aquellos que permanecen, independientemente del umbral seleccionado.

En el contexto de churn, un valor cercano a 1 indica una excelente capacidad del modelo para ordenar correctamente a los clientes según su riesgo de abandono [18].

- **AUC-PR (Area bajo la curva precisión Recall)**: En problemas de churn, donde la clase positiva suele ser minoritaria, la curva Precision-Recall resulta especialmente relevante, ya que se concentra en el desempeño sobre los clientes que efectivamente abandonan [17].

$$Precisión = TP / (TP + FP); Recall = TP / (TP + FN)$$

El área bajo esta curva se expresa como:

$$AUC\_PR = \int_0^1 Precision\ (Recall)\ d(Recall)$$

Esta métrica permite evaluar qué tan bien el modelo prioriza clientes realmente en riesgo, minimizando falsas alarmas.

- **F1-score**: El F1-score sintetiza el balance entre precisión y recall.

$$F1 = 2 * (Precision * Recall) / (Precision + Recall)$$

Es particularmente útil cuando se requiere equilibrio entre detección de churn y eficiencia operativa

- **Average Precision (AP):** Average Precision resume la curva Precision-Recall ponderando la precisión en cada incremento del recall.

$$AP = \Sigma_n\ (Recall_n - Recall_{n-1}) \times Precision_n$$

Esta métrica resulta especialmente valiosa en escenarios de clasificación desbalanceada, ya que refleja la capacidad del modelo para mantener alta precisión a medida que amplía cobertura sobre clientes churn.

Métricas para clustering y segmentación

Dado que este trabajo incorpora segmentación de clientes mediante clustering para datos mixtos, se emplean métricas específicas para evaluar la cohesión, separación y calidad estructural de los grupos.

- **K-Prototypes:** El algoritmo K-Prototypes extiende K-Means para datos mixtos, combinando variables numéricas y categóricas en una única función objetivo [22].

$$J = \Sigma_{i=1}^{n} \Sigma_{k=1}^{K} \delta\ (z_i, k)\ [\ \Sigma_j \in Num\ (x_{ij} - \mu_{kj})^2 + \gamma\ \Sigma_j \in Cat\ I\ (x_{ij} \neq \mu_{kj})]$$

Donde:

- $\delta(z_i, k)$: asignación de observación al cluster
- μ_{kj}: centroide o moda del cluster
- γ: parámetro de balance entre variables numéricas y categóricas

Este enfoque permite capturar simultáneamente señales financieras, demográficas y comportamentales.

- **Distancia de Gower:** La distancia de Gower permite medir similitud entre observación de variables mixtas [21].

$$D_{ij} = (\Sigma_{k=1}{}^{p} w_{ijk}\, d_{ijk}) / (\Sigma_{k=1}{}^{p} w_{ijk})$$

Donde:
- d_{ijk}: distancia parcial por variable
- w_{ijk}: peso o disponibilidad de la variable

Su uso resulta especialmente útil para validar clustering con datos heterogéneos.

- **Silhouette Score:** El índice de Silhouette evalúa simultáneamente cohesión interna y separación externa [23].

$$S(i) = (b(i) - a(i)) / max(a(i), b(i))$$

Donde:
- $a(i)$: distancia promedio dentro del cluster
- $b(i)$: distancia promedio al cluster más cercano

Valores cercanos a 1 indican mejor agrupamiento.

- **Índice de Calinski-Harabasz:** Evalúa la relación entre dispersión entre clusters y dispersión interna [24].

$$CH = [Tr(B_k) / (k - 1)] / [Tr(W_k) / (n - k)]$$

Valores altos indican clusters mejor definidos.

- **Índice de Davies-Bouldin:** Mide similitud promedio entre clusters, considerando dispersión y separación [25].

$$DB = (1 / K)\, \Sigma_{i=1}{}^{K} max_{j,\, j \neq i} [\, (S_i + S_j) / M_{ij}]$$

Valores más bajos representan mejor calidad del clustering.

- **Costo del clustering (Elbow Method):** En K-Prototypes, el costo representa la suma total de distancias intra-cluster.

$$Cost = \Sigma_{i=1}^{n} d\,(x_i, C_i)$$

La identificación del "codo" permite seleccionar un número de clusters que equilibre simplicidad y calidad.

Métricas de priorización comercial

En escenarios reales, una entidad financiera no puede intervenir a todos sus clientes, por lo que resulta necesario evaluar qué tan bien el modelo prioriza los casos de mayor riesgo.

- **Recall Top Decile:** Mide la proporción de churn capturado dentro del 10% de clientes con mayor riesgo predicho:

$$Recall\_TopDecile = Churners\ en\ Top\ 10\% \,/\, Total\ Churners$$

- **Lift Top Decile:** Compara la efectividad del top decil frente a una selección aleatoria.

$$Lift = Recall\_TopDecile \,/\, 0.10$$

Un Lift > 1 indica que el modelo prioriza mejor que el azar.

- **Gain Chart:** permite evaluar la capacidad del modelo para capturar clientes con churn a medida que se prioriza progresivamente un porcentaje de la población según el riesgo predicho. En términos prácticos, muestra qué proporción acumulada de clientes que desertan puede identificarse al intervenir primero a los clientes con mayor probabilidad estimada de abandono [20].

$$Gain(x) = Churners\ capturados\ en\ el\ Top\ x\% \,/\, Total\ de\ churners$$

Donde:

- **Churners capturados en el Top x%:** número acumulado de clientes churn correctamente identificados dentro del porcentaje superior priorizado.
- **Totalde churners:** número total de clientes que desertan en el conjunto evaluado

Un modelo con mayor capacidad predictiva presentará una curva de ganancia más pronunciada que la línea base aleatoria, lo que implica una mayor eficiencia en campañas de retención focalizadas [18].

- **Lift Chart:** Muestra cuánto mejora el modelo respecto a una estrategia aleatoria en cada segmento priorizado.

$$Lift(x) = Gain(x) / x$$

Interpretación:

- $Lift = 1$: desempeño equivalente al azar
- $Lift > 1$: el modelo supera una selección aleatoria
- $Lift < 1$: desempeño inferior al azar

Por ejemplo, un Lift de 4 en el primer decil indica que el modelo identifica cuatro veces más clientes churn que una estrategia aleatoria al intervenir el 10% superior de clientes priorizados.

Importancia integral de las métricas:

En conjunto, estas métricas permiten evaluar tres dimensiones críticas del proyecto:

- **Desempeño predictivo** (AUC-ROC, AUC-PR, F1, AP)
- **Calidad de segmentación** (Silhouette, CH, DB, Gower, costo)
- **Valor de negocio** (Recall Top Decile, Lift, Gain)

Esta combinación asegura que el modelo no solo sea estadísticamente sólido, sino también metodológicamente defendible y operacionalmente útil.

1.3.2 Marco Conceptual

- Churn: Tasa de abandono de clientes en un periodo específico.
- Índice de riesgo financiero-comportamental: Medida compuesta que integra variables financieras (ej. CreditScore) y de comportamiento (ej. satisfacción, quejas).
- Clustering: Técnica no supervisada para agrupar observaciones en función de su similitud.
- Modelos supervisados: Algoritmos que predicen una variable objetivo (churn) a partir de variables explicativas.
- AUC-ROC: Métrica que evalúa la capacidad discriminativa de un modelo.
- Recall: Proporción de clientes que en realidad desertan y son correctamente identificados por el modelo.

1.3.3 Marco Legal y Normativo

En Colombia, la gestión y tratamiento de datos personales se encuentra regulada principalmente por la **Ley 1581 de 2012**, conocida como Régimen General de Protección de Datos Personales, y por el **Decreto 1377 de 2013**, compilado posteriormente en el Decreto Único 1074 de 2015. Estas disposiciones establecen los principios, derechos y procedimientos que deben cumplir las organizaciones para garantizar el tratamiento legal, seguro y transparente de la información personal [30, 31, 32].

Entre los principios más relevantes para el desarrollo de modelos analíticos aplicados al sector financiero se encuentran la legalidad, finalidad, libertad, veracidad, transparencia, acceso restringido, seguridad y confidencialidad. En este sentido, cualquier proceso de análisis de datos orientado a predicción, segmentación o perfilamiento de clientes debe asegurar que la información utilizada responda a fines legítimos, cuente con medidas adecuadas de protección y respete los derechos del titular de los datos.

Adicionalmente, el sector financiero colombiano opera bajo lineamientos regulatorios complementarios emitidos por entidades como la **Superintendencia Financiera de Colombia (SFC)** y la **Superintendencia de Industria y Comercio (SIC)**, particularmente en temas relacionados con administración de riesgo, seguridad de la información y protección del consumidor financiero. Esto implica que el uso de modelos predictivos, incluyendo herramientas de churn, debe considerar no solo precisión analítica, sino también criterios de gobernanza, trazabilidad, explicabilidad y uso responsable de la información. Desde una perspectiva aplicada, el desarrollo de modelos de churn y segmentación debe evitar prácticas que comprometan derechos fundamentales, como el uso indebido de información sensible, decisiones automatizadas opacas o incorporación de variables que puedan generar sesgos injustificados. Por ello, conceptos como minimización de datos, calidad de información, validación metodológica y uso ético de modelos adquieren relevancia dentro de la implementación de soluciones analíticas en entornos financieros [33, 34].

En consecuencia, aunque el presente trabajo se desarrolla en un contexto académico y sobre una base de datos de referencia, su enfoque metodológico reconoce la importancia de la normativa colombiana en protección de datos y la necesidad de que cualquier implementación futura en escenarios reales se encuentre alineada con estándares legales, regulatorios y éticos de tratamiento responsable de información.

1.3.4 Antecedentes

La problemática del customer churn ha sido ampliamente estudiada en diferentes industrias, especialmente en telecomunicaciones y banca, debido a su impacto directo en la rentabilidad. A nivel internacional, destacan que la predicción del churn mediante modelos supervisados como la Regresión Logística y árboles de decisión ha sido una práctica común, aunque limitada por la falta de segmentación de clientes [15].

En el contexto financiero, Burez y Van den Poel, demostraron que la combinación de modelos predictivos con estrategias de retención personalizadas puede reducir significativamente la tasa de abandono. Asimismo, incorporaron técnicas de clustering para segmentar clientes antes de aplicar modelos supervisados, logrando mejoras en métricas como el AUC-ROC y la precisión [15].

En América Latina, la transformación digital del sector financiero ha impulsado el uso de analítica avanzada para fortalecer la retención, la personalización y la gestión del riesgo de clientes. Estudios aplicados a instituciones financieras en economías emergentes, como el de de Lima Lemos, Silva y Tabak, muestran que los modelos de aprendizaje automático pueden mejorar la identificación de clientes con propensión al abandono cuando incorporan variables financieras, transaccionales y comportamentales. De forma complementaria, reportes regionales sobre fintech y digitalización financiera evidencian que la interacción digital y los patrones de uso se han convertido en insumos relevantes para comprender la relación entre cliente y entidad financiera. [29].

En Colombia, aunque existen avances importantes en transformación digital, inclusión financiera y analítica comercial, la producción académica específica sobre churn bancario con enfoques integrales de clustering, modelos supervisados y evaluación de equidad aún es limitada. Esta situación abre una oportunidad para proponer metodologías que integren segmentación no supervisada, modelado predictivo y criterios de aplicabilidad estratégica en el sector financiero colombiano [27, 28].

2. Preparación y Comprensión de Datos

2.1 Fuente, alcance y estructura de los datos

Se utilizó el conjunto “Bank Customer Churn Prediction” (Kaggle), con 10.000 observaciones y 18 variables que incluyen atributos financieros (ejemplo: *CreditScore, Balance, Estimated Salary*), comportamentales (ejemplo: Is Active *Member, Satisfaction Score, Complain, Card Type, Point Earned)* y la variable objetivo-binaria *Exited* (1 = abandonó, 0 = permanece). La carga del dataset se realizó de forma robusta, detectando automáticamente el archivo CSV principal y validando el schema (dimensiones, tipos, nulos, duplicados).

El conjunto de datos utilizado corresponde a una base de referencia pública disponible en Kaggle, ampliamente empleada en ejercicios académicos de predicción de churn bancario. Aunque representa un entorno simulado y anonimizado, incluye variables financieras, demográficas y comportamentales consistentes con escenarios reales del sector financiero. Debido a la ausencia de documentación detallada sobre la generación temporal de algunas variables, los resultados del estudio se interpretan desde una perspectiva metodológica y comparativa, más que como una implementación directa en una entidad específica.

Tabla 1. *Estructura básica del dataset (primeras filas y tipos de datos)*
Fuente: *Elaboración propia con Python.*

	RowNumer	Customerid	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance
0	1	15634602	Hargrave	619	France	Female	42	2	0,00
1	2	15647311	Hill	608	Spain	Female	41	1	83807,86
2	3	15619304	Onio	502	France	Female	42	8	159660,80
3	4	15701354	Bóni	699	France	Female	39	1	0,00
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510,82

La **Tabla 1** muestra las primeras filas del dataset utilizado, evidenciando la estructura básica de los datos. Se incluyen identificadores (*RowNumber, CustomerId, Surname*), atributos financieros (*CreditScore*, Balance, *EstimatedSalary*), demográficos (*Geography*, *Gender, Age*), y variables relacionadas con la relación del cliente con el banco (*Tenure*). Esta visualización permite confirmar que el archivo se cargó correctamente, que las columnas están bien definidas y que no existen valores nulos ni duplicados en las variables mostradas. Además, sirve como referencia para comprender el tipo de información disponible para el análisis posterior.

Hallazgo inicial sobre el target. La distribución de *Exited* evidencia desbalance de clases (≈ 79,62 % no-churn y 20,38 % churn).

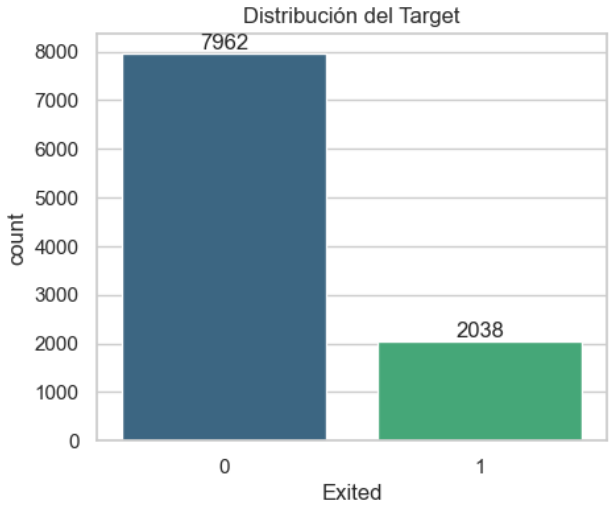

Figura 1. Distribución de la variable objetivo (Exited)
Fuente: Elaboración propia con Python.

La **Figura 1** presenta un gráfico de barras con la distribución de la variable objetivo (*Exited*), mostrando el conteo de observaciones por clase (0 = permanecen, 1 = abandonan). Se observa un desbalance, con 7.962 clientes que permanecen (79,62%) y 2.038 que abandonan (20,38%).

Este comportamiento es típico en problemas de churn y tiene implicaciones en el modelado, ya que puede sesgar los resultados hacia la clase mayoritaria. Por ello, se hace necesario el uso de métricas como AUC-PR y *recall*, así como estrategias de ponderación de clases para evaluar adecuadamente el modelo.

Tabla 2. *Proporciones de la variable objetivo*
Fuente: *Elaboración propia con Python.*

Proporciones:	***Exited***
0	*79.62%*
1	*20.38%*
Name:	*Proportion, dtype: object*

El comportamiento de la **Tabla 2** confirma el desbalance en la variable objetivo, lo cual debe considerarse en el modelado para evitar sesgos hacia la clase mayoritaria. En este contexto, los resultados posteriores muestran que la incorporación del IRFC y la segmentación mediante clustering mejora la identificación de clientes en riesgo, reflejándose en un mejor desempeño en métricas como AUC-PR y *recall*. [9].

2.2 Preparación y limpieza

Se normalizaron los nombres de variables, se identificaron columnas de identificación (*RowNumber, CustomerId, Surname*) y se separaron las variables numéricas y categóricas para el análisis. No se detectaron valores faltantes; esto simplifica la preparación.

Tabla 3. *Resumen de tipos y conteo de nulos por variable*
Fuente: *Elaboración propia con Python.*

Shape: (10000, 18)		***Nulos por columna***	***0***
Tipos:		***Tipos:***	
RowNumber	*Int64*	*RowNumber*	*0*
CustomerId	*Int64*	*CustomerId*	*0*
Surname	*object*	*Surname*	*0*
CreditScore	*Int64*	*CreditScore*	*0*
Geography	*object*	*Geography*	*0*
Gender	*object*	*Gender*	*0*
Age	*int64*	*Age*	*0*
Tenure	*Int64*	*Tenure*	*0*
Balance	*Float64*	*Balance*	*0*
NumOfProducts	*Int64*	*NumOfProducts*	*0*
HasCrCard	*Int64*	*HasCrCard*	*0*
IsActiveMember	*Int64*	*IsActiveMember*	*0*
EstimatedSalary	*Float64*	*EstimatedSalary*	*0*
Exited	*Int64*	*Exited*	*0*
Complain	*Int64*	*Complain*	*0*
Satisfaction Score	*Int64*	*Satisfaction Score*	*0*
Card Type	*object*	*Card Type*	*0*
Point Earned	*Int64*	*Point Earned*	*0*
dtype: object		*dtype: int64*	*0*
		Duplicados (filas):	*0*

La **Tabla 3** presenta un resumen de los tipos de datos y la verificación de valores nulos por variable. Se observa que el dataset está compuesto por 10.000 observaciones y 18 variables, incluyendo atributos numéricos y categóricos relevantes para el análisis. No se identifican valores nulos ni registros duplicados, lo que indica una adecuada calidad de los datos para el desarrollo del modelo.

2.3 Calidad de datos y Outliers

Para la revisión de valores atípicos en variables continuas se utilizó el criterio del rango intercuartílico (IQR), que identifica observaciones ubicadas por debajo de $Q1-1.5\times IQR$ o por encima de $Q3+1.5\times IQR$. Este criterio se aplicó únicamente a variables numéricas continuas, dado que en variables binarias puede generar falsos valores extremos por la propia naturaleza 0/1 de la variable.

***Tabla 4**. Conteo de atípicos por IQR, número de valores únicos y faltantes*
***Fuente:** Elaboración propia con Python.*

	n_outliers_IQR	***n_unique***	***missing***
Complain	*2044*	*2*	*0*
Age	*359*	*70*	*0*
NumOfProducts	*60*	*4*	*0*
CreditScore	*15*	*460*	*0*
Tenure	*0*	*11*	*0*
HasCrCard	*0*	*2*	*0*
Balance	*0*	*6382*	*0*
IsActiveMember	*0*	*2*	*0*
EstimatedSalary	*0*	*9999*	*0*
Satisfaction Score	*0*	*5*	*0*

En la **Tabla 4** se aplicó tamiz IQR para variables continuas; el conteo de atípicos confirma que *Age* presenta valores extremos plausibles en el negocio (clientes muy jóvenes y de mayor edad). En cambio, para variables binarias (p. ej., *Complain*), el IQR no es aplicable porque induce "falsos extremos" (la dispersión se explica por la naturaleza 0/1). Se documentó el criterio: mantener los valores extremos plausibles y no tratar outliers en binarias, priorizando la estandarización y el uso de métricas robustas en los modelos.

3. Análisis Exploratorio de Datos (EDA) y Señales frente a Churn

3.1 Exploración Univariada

Se caracterizaron las distribuciones de las variables numéricas mediante histogramas / KDE y las categóricas con diagramas de barras (frecuencias).

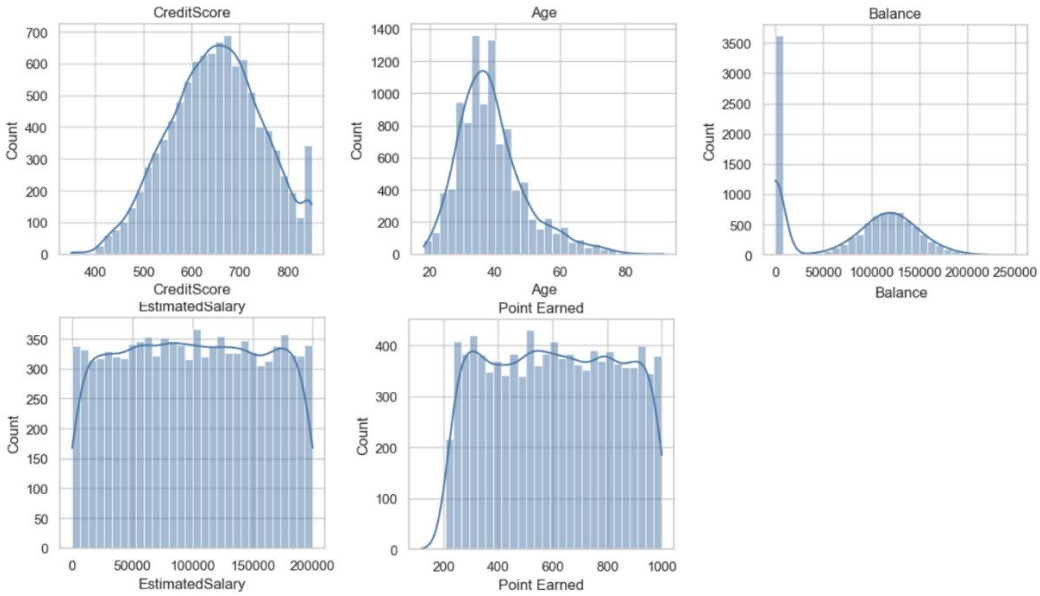

Figura 2. *Distribución de variables numéricas Continuas (histogramas)*
Fuente: *Elaboración propia con Python.*

La **Figura 2** presenta la distribución de variables numéricas continuas como *CreditScore, Age, Balance*, *EstimatedSalary y PointEarned*. Se observan distribuciones relativamente concentradas en algunas variables y asimetrías en otras, como Balance, donde existe una acumulación importante de valores bajos. Estas características justifican el uso posterior de procesos de estandarización y permiten identificar variables con potencial aporte al análisis del churn.

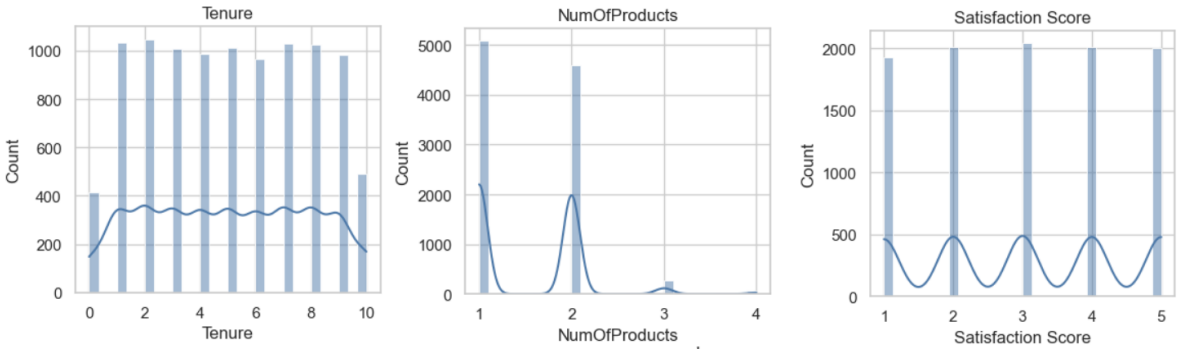

Figura 3. *Distribución de variables numéricas Discretas (histogramas)*
Fuente: *Elaboración propia con Python.*

La **Figura 3** presenta variables como *Tenure, NumOfProducts y Satisfaction Score*. Se evidencia una concentración en valores bajos para *NumOfProducts* y una distribución más uniforme en *Tenure y Satisfaction Score*. Estas variables reflejan el nivel de vinculación y satisfacción del cliente, siendo factores relevantes en la probabilidad de abandono.

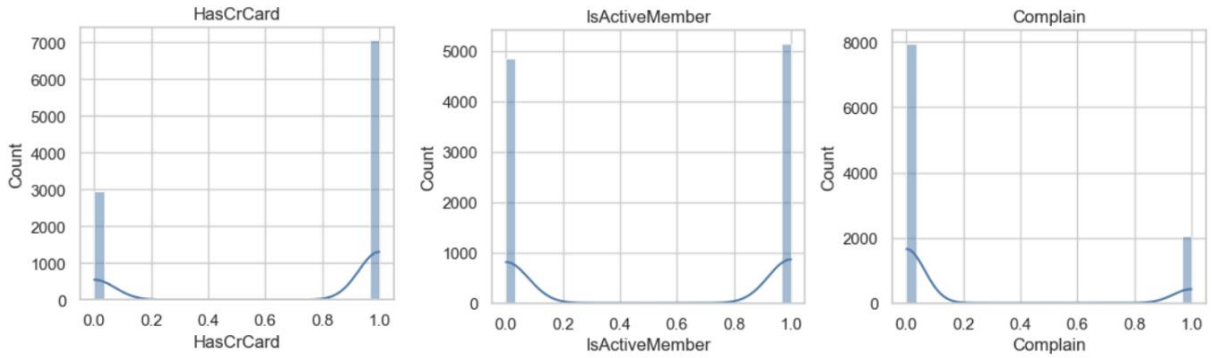

Figura 4. *Distribución de variables numéricas Binarias (histogramas)*
Fuente: *Elaboración propia con Python.*

La **Figura 4** muestra variables como *HasCrCard, IsActiveMember y Complain*. Se observa una alta proporción de clientes con tarjeta y activos, mientras que *Complain* presenta menor frecuencia, no obstante, esta última mostró alta correlación con la variable objetivo, evidenciando un posible problema de fuga de información, lo cual fue considerado en el ajuste del modelo.

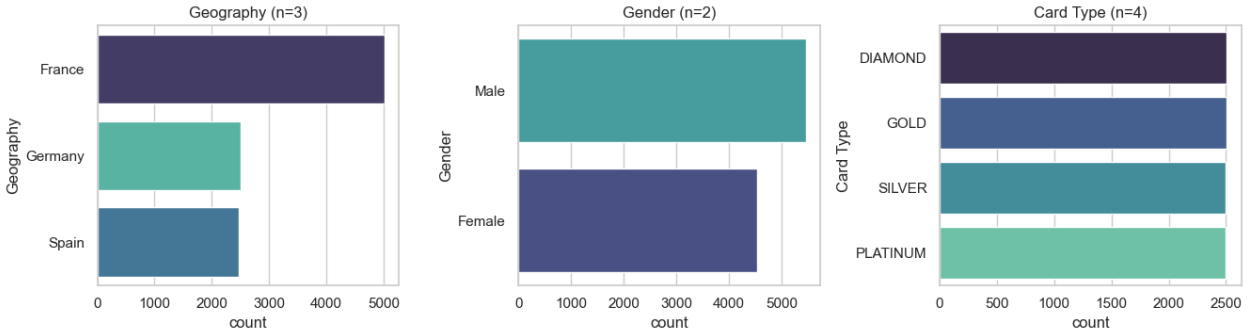

Figura 5*. Distribución de variables categóricas (conteos)*
Fuente: *Elaboración propia con Python.*

La **Figura 5** presenta la distribución de variables categóricas como *Geography*, *Gender* y *Card Type*. Se observa que la mayoría de los clientes se concentran en Francia, seguida de Alemania y España, lo que indica una distribución geográfica no uniforme.

En cuanto al género, se evidencia una mayor proporción de clientes masculinos frente a femeninos. Por su parte, la variable *Card Type* muestra una distribución relativamente equilibrada entre las distintas categorías, sin una concentración dominante. Estas variables aportan contexto

sobre la composición de la base de clientes y pueden influir en el comportamiento de churn, especialmente en factores relacionados con ubicación geográfica y características del producto.

3.2 Relaciones entre variables (matriz de correlación)

Se calculó la correlación de Spearman entre variables numéricas. Las dependencias lineales son, en general, bajas, lo que sugiere relaciones no lineales y efectos de interacción modestos entre predictores financieros y de actividad.

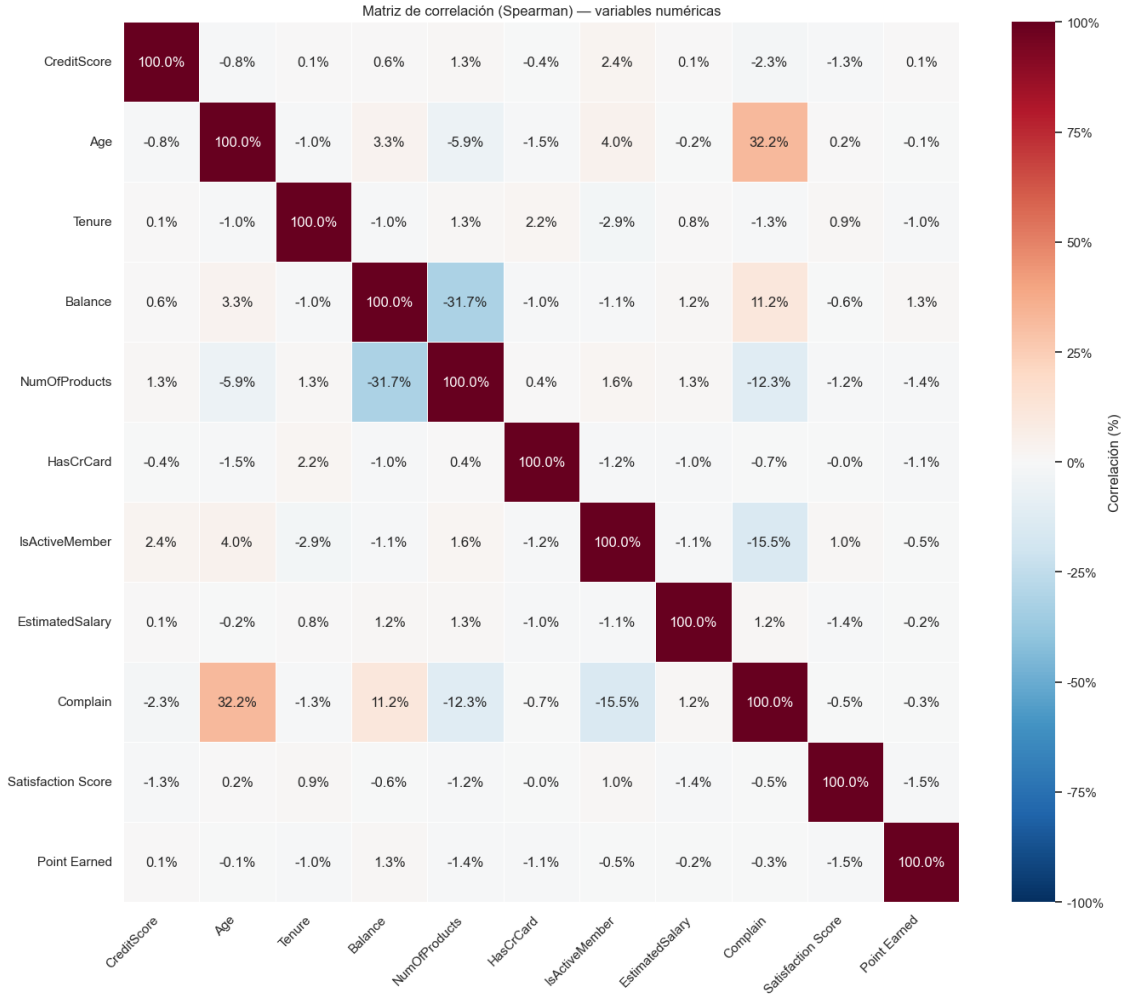

***Figura 6**. Matriz de correlación (Spearman) entre numéricas*
***Fuente:** Elaboración propia con Python.*

La **Figura 6** presenta la matriz de correlación de Spearman entre variables numéricas. En general, se observan asociaciones bajas o moderadas entre los predictores, lo que sugiere ausencia de redundancia severa en esta etapa del análisis. Algunas variables, como Age y Balance, muestran señales de relación con Exited, aunque no suficientemente fuertes para explicar por sí solas el fenómeno de churn. Este resultado respalda el uso de enfoques multivariados y modelos capaces de capturar relaciones no lineales.

3.3 Señales Bivariadas frente a Churn

Se exploraron boxplots (numéricas vs *Exited*) y (tasa de churn por categóricas).

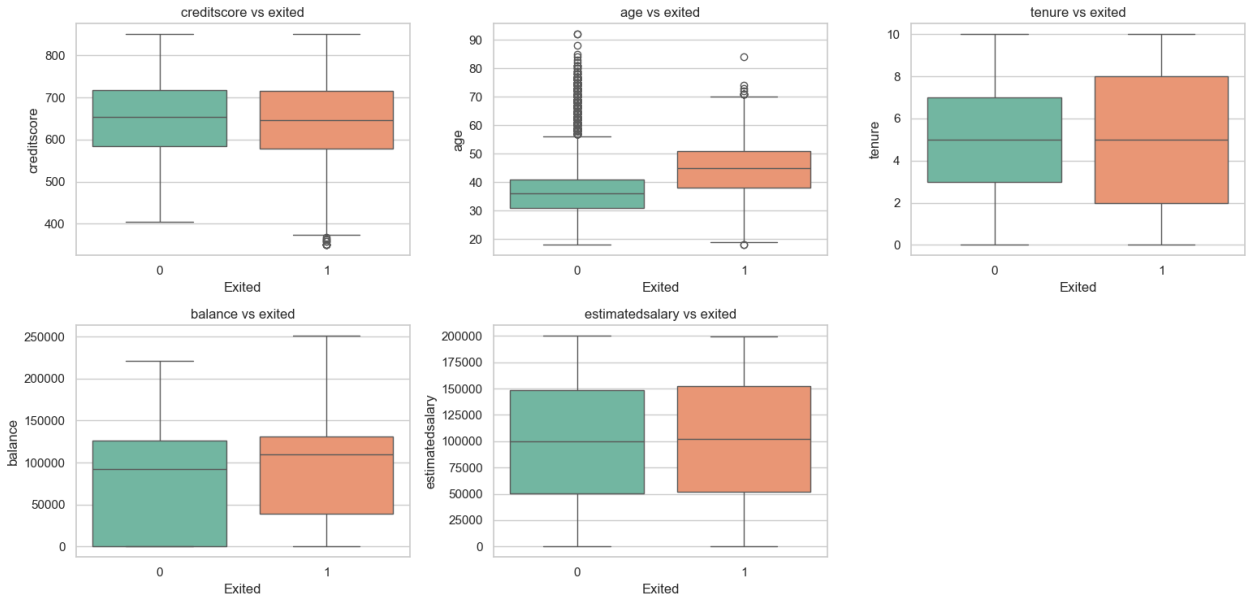

***Figura 7**. Boxplots de variables numéricas vs Exited*
***Fuente:** Elaboración propia con Python.*

La **Figura 7** presenta la relación entre variables numericas y la variable objetivo (*Exited*) mediante diagramas de caja. Se observan diferencias relevantes en variables como *Age y Balance*, donde los clientes que abandonan tienden a presentar valores más altos, lo que sugiere una posible asociación con el churn.

Por su parte, variables como *CreditScore*, *EstimatedSalary* no evidencian diferencias marcadas entre clases, indicando un menor poder discriminativo cuando se analiza de manera individual. En cuanto a *Tenure*, se observan variaciones moderadas, aunque no concluyentes de forma aislada.

En conjunto, estos resultados sugieren que el comportamiento del churn responde a la interacción de múltiples factores financieros y demográficos, lo que motiva el uso de enfoque multivariados y técnicas avanzadas de analítica para comprender mejor el riesgo de abandono.

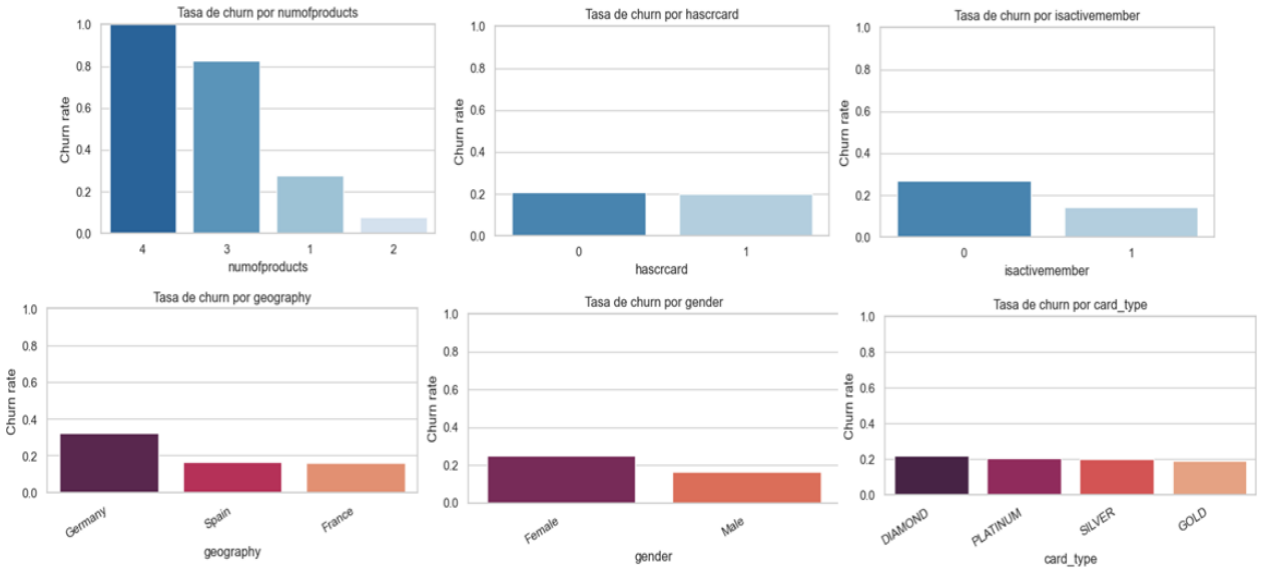

Figura 8. *Tasa de churn por Variables categóricas y discretas*
Fuente: *Elaboración propia con Python.*

La **Figura 8** muestra la tasa promedio de churn para variables categóricas, binarias y discretas, permitiendo analizar como varia la proporción de clientes que abandonan la entidad según diferentes caraterizticas del cliente.

Se observa diferencias relevantes en variables como *Geography,* donde Alemania presenta una tasa de churn superior a España y Francia, sugiriendo que la ubicación geográfica puede influir en la probabilidad de abandono. De igual manera, la variable *ISActiveMember* evidencia que los clientes activos presentan menores niveles de churn, lo que resalta la importancia de la vinculación y el uso de los servicios financieros.

Por su parte variables como *CardType* y *HasCrCard* muestran diferencias menos marcadas entre categorías, indicando un menor poder discriminativo cuando se analizan de forma aislada. En el caso de *NumOfProducts*, se identifican variaciones en la tasa de churn entre grupos, lo que sugiere que el nivel de vinculación comercial podría estar asociado al riesgo de abandono.

En conjunto, estos resultados muestran que el churn está influenciado por factores demográficos, comerciales y comportamentales, reforzando la necesidad de utilizar enfoques multivariados para comprender la complejidad del fenómeno.

3.4 Ranking de importancia Univariada (punto -biserial)

El coeficiente punto-biserial (r) entre cada numérica y Exited arrojó una asociación prácticamente perfecta para *Complain* (r≈0,996), seguida por Age (r≈0,285) y *Balance* (r≈0,119). Efectos muy pequeños o nulos para *EstimatedSalary, Satisfaction Score*, *HasCrCard, Tenure, CreditScore y NumOfProducts*.

***Tabla 5**. Correlación punto-biserial entre numéricas y Exited*
***Fuente:** Elaboración propia con Python.*

	feature	*Pointbiserial_r*	*p_value*
8	*Complain*	*0.995693*	*0.000000e+00*
1	*Age*	*0.285296*	*1.346716e-186*
3	*Balance*	*0.118577*	*1.209208e-32*
7	*EstimatedSalary*	*0.012490*	*2.117146e-01*
10	*Point Earned*	*-0.004628*	*6.435350e-01*
9	*Satisfaction Score*	*-0.005849*	*5.586474e-01*
5	*HasCrCard*	*-0.006976*	*4.8554722e-01*
2	*Tenure*	*-0.013656*	*1.721045e-01*
0	*CreditScore*	*-0.026771*	*7.422037e-06*
4	*NumOfProducts*	*-0.047611*	*1.905777e-06*

La **Tabla 5** presenta la correlación punto-biserial entre la variable objetivo *Exited* y las variables numéricas. La variable *Complain* muestra una asociación prácticamente perfecta con el churn (*r≈0.996*), lo cual constituye una señal metodológica relevante, ya que sugiere que esta variable podría estar registrada después del evento de abandono o estar estrechamente relacionada con su definición. Por esta razón, en los capítulos posteriores se evaluarán escenarios con y sin *Complain*, con el fin de medir su impacto sobre el desempeño y la validez del modelo. *Age* presenta una asociación positiva de magnitud débil a moderada, lo que indica que los clientes de mayor edad tienden a mostrar una mayor proporción de churn. *Balance* también presenta una relación positiva, aunque de menor magnitud. En contraste, variables como *CreditScore* y *NumOfProducts* muestran asociaciones negativas pequeñas, lo que sugiere que mayores valores en estas variables se relacionan levemente con menor probabilidad de abandono. Las demás variables presentan efectos reducidos o no significativos desde el punto de vista práctico en este análisis univariado.

3.5 Implicaciones para la evaluación de modelos

Dado el desbalance observado en la variable objetivo, la evaluación de los modelos no se basará únicamente en métricas globales como *accuracy*. Se priorizarán AUC-ROC, AUC-PR, *recall* y métricas de ranking como *Recall Top Decile* y *Lift*, ya que permiten valorar mejor la capacidad del modelo para identificar clientes con mayor probabilidad de abandono. Este enfoque es más adecuado para escenarios de retención, donde las intervenciones comerciales suelen concentrarse en segmentos priorizados. [11].

4. Diseño y Validación del Índice de Riesgo Financiero-Comportamental (IRFC)

4.1 Enfoque exploratorio inicial: Análisis Factorial Exploratorio (EFA)

Con el propósito de construir un indicador sintético que representara la propensión al abandono de los clientes, se exploró inicialmente el uso de Análisis Factorial Exploratorio (EFA), una técnica orientada a identificar estructuras latentes subyacentes a partir de la correlación entre variables, reduciendo dimensionalidad y facilitando la construcción de factores interpretables.

Este enfoque constituyó una aproximación metodológica inicial para evaluar si las variables financieras y comportamentales del cliente podían explicarse mediante una estructura factorial común. No obstante, más allá de su utilidad exploratoria, el análisis también permitió evaluar la viabilidad estadística de este enfoque como base para la construcción del índice. Posteriormente, las limitaciones observadas en términos de adecuación factorial condujeron a replantear el proceso hacia una construcción compuesta más orientada al desempeño predictivo.

Para este análisis se seleccionaron variables transformadas con orientación al riesgo, entre ellas *credit_risk*, *satisfaction_risk* y *complaint_risk*, diseñadas para capturar dimensiones financieras y comportamentales relevantes en la explicación del churn.

Previo a la estimación del modelo factorial, se evaluó la adecuación estadística de los datos para determinar si las variables seleccionadas presentaban condiciones suficientes para sustentar una estructura latente común. Para ello, se utilizó el índice KMO (Kaiser-Meyer-Olkin), orientado a medir la proporción de varianza compartida entre variables, y la prueba de esfericidad de Bartlett, cuya finalidad es contrastar si la matriz de correlación difiere significativamente de una matriz identidad. De manera complementaria, se analizaron las cargas factoriales, las comunalidades y el coeficiente alfa de Cronbach, con el propósito de evaluar tanto la capacidad explicativa de los factores extraídos como la consistencia interna del índice propuesto.

***Tabla 6**. Adecuación muestral (KMO), esfericidad (Bartlett), cargas y comunalidades*
***Fuente:** Elaboración propia con Python.*

Bartlett X^2=9.2, P-value = 0.02714 (p<0.05 → covarianzas ≠ 0	
KMO global=0.497 (≥0.6 aceptable, ≥0.7 bueno)	
Cargas (loadings) del factor IRFC:	
Credit_risk	*-0.997*
Satisfaction_risk	*0.013*
Complaint_risk	*-0.027*

Name: loading, dtype:	*Float64*
Comunalidades:	
Credit_risk	*0.995*
Satisfaction_risk	*0.000*
Complain_risk	*0.001*
Name: communality, dtype:	*Float64*
Cronbach α ≈ 0.019	
AUC (IRFC → Exited) = 0.48 (↑mejor)	

La **Tabla 6** presenta los resultados del análisis factorial muestran una baja adecuación muestral, reflejada en un valor de KMO inferior al umbral recomendado. Si bien la prueba de Bartlett resulta significativa, indicando la existencia de correlaciones distintas de cero, las cargas factoriales evidencian que solo una de las variables aporta de manera relevante al factor común. Este comportamiento sugiere que las variables consideradas no comparten una estructura latente sólida, limitando la viabilidad de construir un índice unifactorial representativo.

4.2 Limitaciones del enfoque factorial

Los resultados obtenidos evidencian que el enfoque basado en EFA no es adecuado para la construcción del IRFC en este contexto.

En particular:

- El valor de **KMO** se encuentra por debajo del nivel aceptable (≥ 0.6), lo que indica una baja correlación común entre las variables.
- El coeficiente de Cronbach **(α ≈ 0.019)** evidencia una consistencia interna prácticamente nula.
- Las cargas factoriales muestran una alta concentración en una sola variable, mientras que las demás presentan contribuciones marginales.
- El índice derivado presenta un desempeño predictivo bajo **(AUC ≈ 0.48)**.

Los resultados sugieren que la estructura de correlación observada no respalda de manera suficiente la existencia de un constructo latente único, lo que limita la aplicabilidad del enfoque factorial como mecanismo principal de construcción del índice.

4.3 Construcción del IRFC compuesto (pipeline final)

Dado que el enfoque factorial exploratorio no ofreció evidencia estadística suficiente para sustentar una estructura latente robusta, se replanteó la construcción del índice hacia un enfoque compuesto, orientado a sintetizar múltiples dimensiones financieras y comportamentales del riesgo de churn bajo criterios de capacidad discriminativa, interpretabilidad y aplicabilidad analítica.

En este contexto, el Índice de Riesgo Financiero-Comportamental (IRFC) no fue concebido exclusivamente como una variable predictiva para modelado supervisado final, sino como una herramienta analítica integral destinada a condensar señales relevantes de riesgo en una sola medida interpretable. Este enfoque permite ordenar clientes según su propensión relativa al abandono, apoyar procesos de priorización estratégica, enriquecer ejercicios de segmentación y facilitar la comprensión del fenómeno de churn desde una perspectiva multidimensional. En otras palabras, el IRFC se diseñó como una capa de interpretación y ranking del riesgo, útil para identificar patrones de comportamiento, comparar perfiles de clientes y fortalecer decisiones analíticas, más que como un predictor definitivo para implementación productiva. Esta distinción resulta metodológicamente relevante, ya que permite separar el valor explicativo y estratégico del índice de su eventual uso dentro de modelos supervisados, donde factores como colinealidad, redundancia o fuga de información deben evaluarse con mayor rigor.

Bajo esta lógica, se desarrolló un pipeline estructurado de construcción del índice compuesto, conformado por las siguientes etapas:

4.3.1 Ingeniería de Variables

Con el propósito de capturar señales de riesgo más complejas que las observables mediante variables originales aisladas, se diseñaron variables derivadas orientadas a representar relaciones financieras y comportamentales potencialmente asociadas al churn.

Entre las principales variables construidas se encuentran:

- **balance_to_salary**: mide la relación entre el saldo disponible del cliente y su ingreso estimado, funcionando como una aproximación a presión financiera relativa o uso intensivo de recursos financieros frente a capacidad económica.

- **num_products_dev**: representa la desviación del número de productos contratados respecto a un patrón esperado, actuando como indicador de comportamiento atípico o posible desconexión con la oferta comercial de la entidad.

Estas variables derivadas permiten incorporar relaciones no lineales y señales compuestas que difícilmente emergen en análisis descriptivos tradicionales, fortaleciendo así la sensibilidad analítica del índice.

4.3.2 Estandarización y orientación a riesgo

Una vez definidas las variables base y derivadas, se aplicó un proceso de estandarización mediante z-score, con el objetivo de garantizar comparabilidad entre variables originalmente medidas en escalas heterogéneas.

$$Z = (X - \mu) / \sigma$$

Donde:

- X: valor observado
- μ: media de la variable
- σ: desviación estándar

Posteriormente, todas las variables fueron orientadas bajo una lógica interpretativa común:

$$\textit{Valores más altos del IRFC} \Rightarrow \textit{Mayor riesgo de churn}$$

Este proceso aseguró coherencia semántica y consistencia analítica, permitiendo que el índice final pudiera interpretarse de manera directa como una escala ordenada de riesgo relativo.

4.3.3 Búsqueda de combinaciones optimas

En lugar de asumir una estructura fija, se implementó un proceso sistemático de búsqueda exhaustiva de combinaciones de variables, evaluando múltiples configuraciones entre tres y seis variables, con el fin de identificar estructuras compactas capaces de maximizar desempeño sin sacrificar interpretabilidad.

Cada combinación fue evaluada mediante:

- Regresión Logística con *class_weight='balanced'*
- Validación cruzada estratificada de 5 folds

Este procedimiento permitió controlar el desbalance de clases y reducir el riesgo de sobreajuste, asegurando una evaluación más robusta.

Las métricas utilizadas fueron:

- AUC-ROC (capacidad discriminativa global)

- AUC-PR (desempeño sobre la clase minoritaria (clientes churn)

El criterio de selección no se limitó a maximizar métricas, sino también a privilegiar configuraciones parsimoniosas, interpretables y metodológicamente defendibles.

Tabla 7. Top combinaciones y métricas (AUC-ROC/AUC-PR)
Fuente: *Elaboración propia con Python.*

3-vars: 100% | 120/120 [00:24<00:00, 5.72it/s]				
4-vars: 100% | 210/210 [00:41<00:00, 6.39it/s]				
5-vars: 100% | 252/252 [00:51<00:00, 6.48it/s]				
6-vars: 100% | 210/210 [00:43<00:00, 6.19it/s]				
=== Mejores combinaciones para IRFC===				
	N_VARS	**AUC_ROXC**	**AUC_PR**	**KMO**
55	3	0.998012	0.988045	0.501386
46	3	0.997400	0.981707	0.497987
51	3	0.995194	0.971425	0.498524
91	3	0.985810	0.967404	0.505089
231	4	0.984454	0.956925	0.505815

Los resultados presentados en la **Tabla 7** evidencian que combinaciones relativamente simples, particularmente aquellas construidas con tres variables, alcanzaron niveles de discriminación extremadamente altos. La recurrencia de variables como *balance_to_salary* y *num_products_dev* dentro de las mejores configuraciones sugiere que estas capturan dimensiones estructurales importantes del riesgo de abandono.

Sin embargo, aunque estos resultados iniciales reflejan una capacidad predictiva sobresaliente, también constituyen un punto de alerta metodológica. Métricas cercanas a la perfección en problemas reales de churn requieren una revisión crítica, ya que podrían estar influenciadas por variables excesivamente correlacionadas con la variable objetivo o por señales de fuga de información.

Por esta razón, la construcción del IRFC no se interpretó como un punto final, sino como un proceso iterativo de validación, refinamiento y contraste metodológico. Este enfoque permitió que el índice evolucionara desde una herramienta de alto desempeño aparente hacia un instrumento más robusto para interpretación, priorización y segmentación, priorizando validez analítica sobre precisión superficial.

4.4 Evaluación del desempeño del IRFC

El índice compuesto mostró inicialmente niveles de discriminación extremadamente altos, reflejados en métricas AUC-ROC y AUC-PR cercanas a la perfección. Aunque estos resultados evidenciaban una fuerte capacidad de separación entre clases, también motivaron una revisión

crítica orientada a validar si dicho desempeño respondía a señales genuinas o a posibles problemas metodológicos asociados a variables con alta proximidad al evento de churn.

- **AUC-ROC ≈ 0.995 – 0.998**
- **AUC-PR ≈ 0.97 – 0.99**

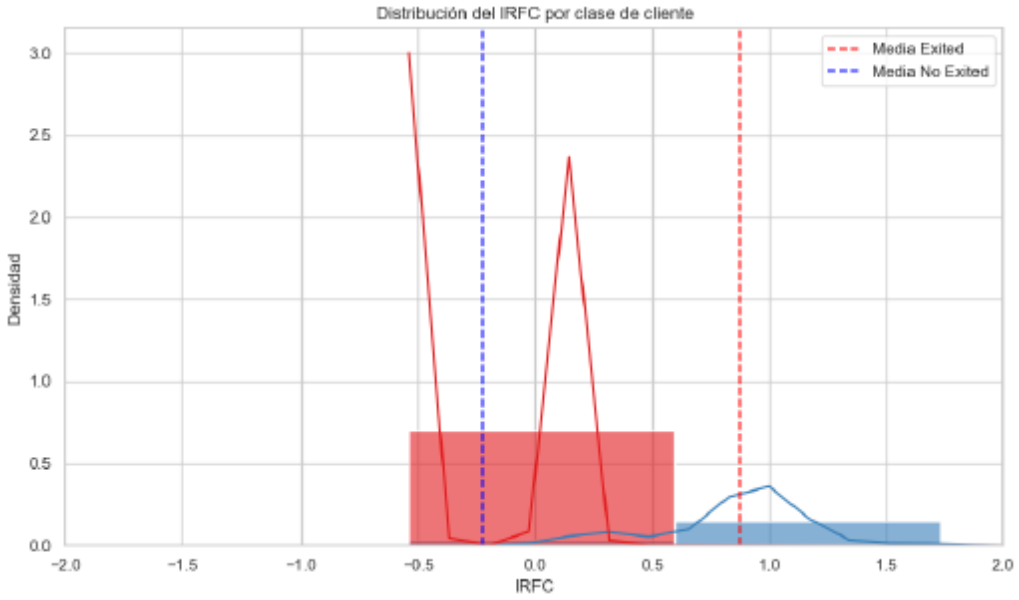

Figura 9. Distribución del IRFC por clase (Exited)
Fuente: Elaboración propia con Python.

La **Figura 9** muestra la distribución del Índice de Riesgo Financiero-Comportamental (IRFC) según la condición de abandono (*Exited*), evidenciando una diferenciación marcada entre clientes que desertan y clientes que permanecen. Se observa que los clientes con churn tienden a concentrarse en valores más altos del índice, mientras que los clientes no churn presentan una mayor concentración en rangos inferiores. Esta separación confirma que el IRFC logra sintetizar señales relevantes de riesgo financiero y comportamental, permitiendo ordenar a los clientes según su propensión relativa al abandono.

No obstante, aunque este comportamiento evidencia una alta capacidad discriminativa en esta etapa analítica, los resultados deben interpretarse con cautela debido a la influencia potencial de variables altamente correlacionadas, particularmente *Complain*, aspecto que posteriormente conduce a una revisión metodológica del índice para evitar sobreestimaciones derivadas de fuga de información.

__Tabla 8__. Ejemplo de clientes con IRFC y Percentiles
Fuente: *Elaboración propia con Python.*

IRFC construido con 3 variables clave			
AUC-ROC: 0.9980			
AUC-PR: 0.9880			
Ejemplo de clientes con IRFC:			
	irfc	Irfc_pctl_0_100	Exited
0	0.941263	85.1	1
1	0.943744	89.8	0
2	0.945932	94.2	1
3	-0.539312	12.6	0
4	0.119959	67.8	0

La **Tabla 8** presenta ejemplos de clientes clasificados mediante el IRFC y sus percentiles asociados, ilustrando cómo el índice permite jerarquizar el riesgo de abandono a nivel individual. En términos generales, clientes ubicados en percentiles superiores muestran una mayor propensión al churn, mientras que aquellos en percentiles inferiores presentan menor riesgo relativo.

Este resultado respalda la utilidad del IRFC como herramienta de priorización, interpretación y segmentación, al facilitar la identificación de clientes que requieren mayor atención desde una perspectiva analítica. Sin embargo, dado que parte de su desempeño inicial puede estar influenciado por variables con riesgo de *leakage*, su principal valor dentro del proyecto se orienta al ordenamiento estratégico del riesgo y no a su incorporación directa como predictor final en modelos supervisados productivos.

Tabla 9. Métricas IRFC_2f
Fuente: *Elaboración propia con Python.*

KMO total=0.508 (≥0.6 aceptable) | Bartlett p=4.714e-243 (p<0.05 ideal)			
AFE: KMO / Bartlett no apoyan un 1-factor; intentaré 2-factores o composite.			
IRFC_2f ->AUC-ROC=0.671 | AUC-PR=0.386			
=== Resumen de variantes IRFC ===			
IRFC_2f ->AUC-ROC=0.671 | AUC-PR=0.386			
0	*0.153528*	*61.8*	*1*
1	*0.078247*	*56.9*	*0*
2	*1.147473*	*99.9*	*1*
3	*0.087944*	*57.6*	*0*
4	*-0.674206*	*14.4*	*0*

Como se observa en la **Tabla 9**, como parte del proceso de construcción del índice se evaluó una alternativa basada en dos factores (financiero y comportamental), denominada IRFC_2f. Los resultados muestran que este enfoque alcanza un desempeño moderado (**AUC-ROC ≈ 0.671 y AUC-PR ≈ 0.386**), lo cual indica que, si bien el índice captura cierta señal predictiva, no resulta suficientemente robusto para discriminar adecuadamente entre clientes que abandonan y aquellos que permanecen.

Desde el punto de vista estadístico, el valor de **KMO ≈ 0.508**, aunque cercano al umbral mínimo aceptable, continúa indicando una débil estructura latente común. Por su parte, la prueba de Bartlett significativa confirma la existencia de correlaciones, pero no suficientes para sostener un modelo factorial sólido.

En conjunto, estos resultados evidencian que el enfoque basado en factores latentes no es adecuado para este conjunto de datos, lo que justifica la transición hacia un índice compuesto optimizado por desempeño predictivo.

4.5 Riesgo de fuga de información (data leakage)

Uno de los hallazgos metodológicos más relevantes del proceso de construcción del IRFC fue la identificación de un posible problema de fuga de información (*data leakage*), asociado principalmente a la variable *Complain*. Durante el análisis exploratorio y de correlación, esta variable presentó una relación prácticamente perfecta con la variable objetivo *Exited*, evidenciando un comportamiento atípico para un problema real de predicción de churn.

Desde una perspectiva analítica, una correlación de esta magnitud constituye una señal crítica, ya que puede indicar que la variable utilizada como predictor contiene información demasiado cercana al evento de abandono, comprometiendo la capacidad del modelo para generalizar adecuadamente en escenarios reales.

En términos metodológicos, la fuga de información ocurre cuando el modelo incorpora variables que, de forma directa o indirecta, contienen información que no estaría disponible en el momento real de la predicción o que reflejan consecuencias posteriores al fenómeno que se busca anticipar. Esto puede suceder, por ejemplo, cuando una variable:

- Es registrada después de ocurrido el churn o como consecuencia del proceso de salida del cliente
- Está estrechamente vinculada con la definición operativa del abandono
- Refleja decisiones, reclamaciones o eventos cuya ocurrencia depende del propio desenlace de churn

En este contexto, la variable *Complain* planteó una alerta metodológica significativa, ya que su

inclusión generaba métricas extraordinariamente altas cercanas a la perfección en diversas configuraciones del índice. Aunque en una primera lectura esto podría interpretarse como un desempeño sobresaliente, en la práctica representa un riesgo importante: el modelo podría estar "aprendiendo" información filtrada o demasiado próxima al resultado, en lugar de identificar patrones genuinos de riesgo previos al abandono.

Este hallazgo fue particularmente importante porque transformó la lógica del proyecto: más allá de buscar únicamente el mejor desempeño aparente, fue necesario priorizar la validez metodológica y la aplicabilidad real del índice. En otras palabras, el análisis pasó de una lógica centrada en maximización de métricas a una lógica de validación crítica del modelo.

Como respuesta a este problema, se decidió construir y comparar explícitamente dos versiones del IRFC:

- **IRFC-A:** Incluye la variable *Complain*, alcanzando métricas de desempeño extremadamente altas, pero con riesgo significativo de fuga de información y menor confiabilidad para escenarios productivos.

- **IRFC-B:** Excluye la variable *Complain*, reduciendo el riesgo de *leakage* y priorizando una estructura más prudente desde el punto de vista metodológico, aun cuando esto implique métricas menos espectaculares, pero más realistas.

Esta diferenciación permitió convertir un problema potencial en un aporte central del estudio, ya que evidenció cómo variables aparentemente poderosas pueden distorsionar artificialmente la evaluación del modelo si no son analizadas críticamente.

En consecuencia, el análisis posterior del proyecto diferencia explícitamente entre desempeño aparente y validez real. La versión IRFC-B se conserva como referencia metodológicamente más robusta para procesos de interpretación, ranking de riesgo y segmentación, mientras que la versión con *Complain* se utiliza principalmente como evidencia del impacto que puede tener la fuga de información sobre la evaluación de modelos predictivos.

Desde una perspectiva académica y aplicada, este hallazgo refuerza una lección fundamental en ciencia de datos: métricas extremadamente altas no siempre representan mejores modelos, y en problemas de churn resulta indispensable cuestionar la temporalidad, naturaleza y origen de las variables antes de considerar su uso en escenarios de implementación.

En síntesis, la detección del riesgo de fuga de información no solo permitió depurar el IRFC, sino que fortaleció la rigurosidad metodológica del trabajo, consolidando un enfoque más confiable, interpretable y defendible para el análisis del churn bancario.

4.6 Definición final del índice y uso en el modelo

Una vez identificado y documentado el riesgo metodológico asociado a variables con potencial fuga de información, la definición final del IRFC se orientó bajo un criterio de equilibrio entre capacidad discriminativa, interpretabilidad y validez analítica.

Más allá de privilegiar exclusivamente el desempeño aparente, esta etapa implicó seleccionar una versión del índice que pudiera conservar utilidad estratégica sin comprometer su aplicabilidad en escenarios reales. En este sentido, aunque algunas configuraciones iniciales alcanzaron métricas extraordinariamente altas, el análisis crítico desarrollado en la sección anterior evidenció que resultados excesivamente ideales debían interpretarse con cautela. Por esta razón, para las fases posteriores del estudio se priorizó una versión metodológicamente más prudente del índice, construida a partir de variables capaces de capturar señales relevantes de riesgo sin depender de información potencialmente filtrada o excesivamente cercana al evento de churn.

Con el propósito de documentar la evolución metodológica del índice, la **Tabla 10** resume los principales escenarios evaluados durante el proceso de construcción del IRFC, incluyendo las variables consideradas, el objetivo analítico y los principales resultados obtenidos.

***Tabla 10**. Escenarios evaluados para la construcción del IRFC*
***Fuente:** Elaboración propia.*

Escenario	Variables principales	Objetivo	Resultado
EFA unifactorial	credit_risk, satisfaction_risk, complain_risk	Identificar una estructura latente común	No soportado (KMO bajo)
EFA dos factores	Variables orientadas al riesgo	Evaluar una estructura multifactorial	Desempeño moderado
IRFC-A	balance_to_salary, num_products_dev, complain_bin	Maximizar capacidad discriminativa	Alto desempeño con riesgo de leakage
IRFC-B	Escenario sin influencia directa de complain	Evaluar robustez metodológica	Más prudente para análisis posteriores

Los resultados evidenciaron que configuraciones que incorporaban la variable complain alcanzaban métricas excepcionalmente altas. Sin embargo, el análisis posterior sugirió que parte de este comportamiento podía estar asociado a mecanismos de fuga de información, razón por la cual dichos escenarios fueron interpretados con cautela.

En consecuencia, el IRFC no se define en este estudio como una fórmula única y definitiva para modelado supervisado, sino como un constructo analítico evaluado bajo diferentes escenarios. Su principal utilidad radica en sintetizar señales financieras y comportamentales para apoyar procesos de interpretación, ranking y segmentación de clientes.

Entre los escenarios evaluados, las variables balance_to_salary y num_products_dev demostraron capturar señales relevantes asociadas al riesgo de churn:

- **balance_to_salary:** refleja presión financiera relativa o posibles desequilibrios entre capacidad económica y comportamiento financiero.
- **num_products_dev:** captura desviaciones respecto a patrones comerciales esperados, funcionando como señal de comportamiento atípico o desconexión con la oferta de productos.

Estas variables aportan interpretabilidad al análisis y contribuyen a la caracterización del riesgo desde una perspectiva financiera y comportamental.

Es importante señalar que, aunque esta versión del IRFC mantiene capacidad discriminativa relevante, su principal valor dentro del proyecto no radica en constituirse como predictor final para modelos supervisados productivos. Debido a su naturaleza compuesta y a su dependencia de variables base, el índice se entiende principalmente como una herramienta de:

- Interpretación del riesgo
- Ranking o priorización de clientes
- Segmentación analítica
- Apoyo a procesos de clustering

En consecuencia, el IRFC se incorpora en el flujo metodológico como una capa analítica complementaria, útil para enriquecer la comprensión del comportamiento del cliente y facilitar decisiones estratégicas, más que como una variable definitiva para implementación supervisada.

Este enfoque resulta especialmente importante, ya que permite separar dos dimensiones distintas del proyecto:

- **Valor analítico del índice:** Capacidad para sintetizar múltiples señales de riesgo en una medida interpretable.
- **Validez predictiva productiva:** Capacidad de integrarse a modelos supervisados finales sin generar colinealidad, redundancia o dependencia metodológica.

Bajo esta lógica, los escenarios asociados al IRFC se utilizaron principalmente como referencia para procesos de clustering, perfilamiento y evaluación comparativa, donde aportan una representación estructurada del riesgo y fortalecen la interpretación de perfiles de clientes.

En síntesis, la definición final del índice no responde únicamente a la búsqueda del mayor desempeño estadístico, sino a una decisión metodológica orientada a robustez, generalización y utilidad estratégica, consolidando al IRFC como una herramienta analítica de alto valor para comprensión, priorización y segmentación del churn.

4.7 Discusión y conclusiones del IRFC

El proceso de diseño, evaluación y ajuste del Índice de Riesgo Financiero-Comportamental (IRFC) constituye uno de los principales aportes metodológicos de este trabajo, no únicamente por los resultados obtenidos, sino por las decisiones analíticas derivadas de su construcción. Más allá de la creación de un indicador compuesto, el desarrollo del IRFC permitió identificar limitaciones, cuestionar resultados, ideales y fortalecer la rigurosidad metodológica del proyecto.

4.7.1 Aprendizaje metodológico: desempeño alto no siempre implica validez

Uno de los aprendizajes más importantes del proceso fue reconocer que métricas predictivas extremadamente altas deben interpretarse críticamente, especialmente en problemas reales de churn. Durante las primeras etapas de construcción, ciertas configuraciones del índice alcanzaron valores de AUC-ROC y AUC-PR cercanos a la perfección, lo que inicialmente podría sugerir una capacidad de predicción excepcional.

Sin embargo, desde una perspectiva metodológica, este comportamiento también constituye una señal de alerta. En escenarios aplicados, resultados casi perfectos suelen ser poco frecuentes y pueden indicar sobreajuste, colinealidad excesiva o fuga de información. En este sentido, el IRFC permitió evidenciar que la evaluación de modelos no debe limitarse a maximizar métricas, sino que debe incorporar una revisión crítica sobre el origen, temporalidad y naturaleza de las variables utilizadas.

Así, el principal aprendizaje no fue simplemente construir un índice “preciso”, sino demostrar que la validez analítica requiere cuestionar incluso los mejores resultados cuando estos no son coherentes con el contexto real del problema.

4.7.2 Data leakage: identificación crítica del impacto de *Complain*

El hallazgo más determinante en esta etapa fue la identificación de la variable *Complain* como una fuente potencial de fuga de información. Su correlación casi perfecta con la variable objetivo generó un efecto de inflación artificial sobre el desempeño del índice, elevando sus métricas a niveles metodológicamente sospechosos. Este comportamiento evidenció que parte del desempeño inicial del IRFC no necesariamente respondía a una mejor representación estructural del riesgo, sino a la incorporación de una variable excesivamente próxima al evento de churn. En consecuencia, el análisis permitió demostrar cómo una variable aparentemente poderosa puede comprometer la generalización del modelo si no se examina críticamente.

La exclusión de *Complain* no solo redujo el riesgo de leakage, sino que transformó el proyecto en un ejercicio más sólido de validación metodológica, priorizando robustez sobre desempeño superficial. Este ajuste representa una de las contribuciones más relevantes del estudio, al reforzar la importancia de distinguir entre precisión aparente y capacidad predictiva genuina.

4.7.3 Valor real del IRFC: herramienta analítica, no solo predictiva

A pesar de las limitaciones identificadas, el IRFC conserva un valor significativo dentro del proyecto. Su principal utilidad no reside exclusivamente en servir como predictor, sino en su capacidad para sintetizar múltiples dimensiones del riesgo en una medida interpretable y estratégicamente accionable.

En su versión refinada, el índice permite:

- Ordenar clientes según nivel relativo de riesgo
- Priorizar poblaciones objetivo para estrategias de retención
- Apoyar procesos de segmentación y clustering
- Facilitar interpretación del riesgo financiero-comportamental

En este sentido, el IRFC funciona como una herramienta de ranking y comprensión del churn, aportando una representación estructurada del riesgo que facilita tanto análisis exploratorios como decisiones estratégicas.

Este enfoque resulta especialmente valioso porque convierte al índice en una capa analítica complementaria, capaz de enriquecer la lectura del comportamiento del cliente más allá de la predicción binaria tradicional.

4.7.4 Decisión final: exclusión como predictor principal en modelos supervisados

A partir de los hallazgos obtenidos, se concluye que el IRFC no debe adoptarse como variable principal en el modelo supervisado final. Esta decisión no implica que el índice carezca de utilidad, sino que su naturaleza compuesta, dependencia de variables base y potencial riesgo de redundancia lo hacen más apropiado como herramienta de interpretación y segmentación que como predictor productivo directo.
Desde una perspectiva metodológica, separar estas funciones fortalece la arquitectura general del proyecto:

- **Modelos supervisados finales:** Priorizan estabilidad, generalización y ausencia de colinealidad.
- **IRFC:** Aporta interpretación, priorización, segmentación y análisis estratégico.

Esta diferenciación permite evitar una sobre dependencia de variables derivadas, manteniendo una estructura de modelado más defendible y aplicable en contextos reales.

En síntesis, el IRFC representa una contribución central al proyecto no porque constituya el modelo final de predicción, sino porque su construcción permitió recorrer un proceso analítico

completo: desde la exploración factorial inicial, pasando por la optimización compuesta, hasta la detección crítica de fuga de información y la redefinición metodológica del índice. El verdadero valor del IRFC radica en haber demostrado que la ciencia de datos aplicada al churn requiere no solo construir modelos predictivos, sino también validar rigurosamente sus fundamentos, cuestionar resultados aparentemente ideales y diferenciar entre herramientas de predicción, interpretación y segmentación.

Por tanto, el IRFC se consolida como una herramienta analítica útil y metodológicamente informativa para ranking, comprensión y priorización del riesgo, mientras que su exclusión como predictor final supervisado responde a una decisión metodológica orientada a garantizar mayor realismo, estabilidad y potencialmente aplicable tras validación en escenarios reales.

5. Segmentación de clientes mediante Clustering para perfilamiento y priorización estratégica del riesgo

5.1 Enfoque metodológico para segmentación

Con el fin de complementar el análisis del churn desde una perspectiva estratégica, se implementó un proceso de segmentación mediante técnicas de aprendizaje no supervisado. A diferencia del modelado supervisado, cuyo propósito principal es predecir la variable objetivo, este enfoque busca identificar estructuras subyacentes, patrones de comportamiento y perfiles diferenciados de clientes sin utilizar directamente la variable Exited.

En este contexto, el clustering se incorpora como una herramienta de perfilamiento y priorización analítica, orientada a comprender mejor la heterogeneidad de la base de clientes, identificar segmentos con características de riesgo diferenciadas y facilitar la definición de estrategias de retención más focalizadas.

5.2 Preparación de datos para clustering

El procedimiento completo se documenta en el Anexo A. En el desarrollo metodológico, la preparación de datos para clustering incluyó la estandarización de variables numéricas, la conversión de variables categóricas al formato requerido por K-Prototypes y la selección de variables relevantes para la segmentación:

- Estandarización de variables numéricas mediante StandardScaler
- Conversión de variables categóricas en formato requerido por K-Prototypes
- Selección de variables relevantes alienadas con el análisis del IRFC

En coherencia con los hallazgos del capítulo 4, se excluyó la variable *Complain* del proceso de segmentación, evitando así la introducción de información potencialmente filtrada (data leakage).

Se definieron dos configuraciones de análisis:

- Escenario A: Clustering sin variables derivadas de riesgo
- Escenario B: Clustering incorporando el índice IRFC (sin la variable *Complain*)

El objetivo de esta comparación es evaluar si la incorporación de un índice de riesgo mejora la calidad de la segmentación.

5.3 Selección del número óptimo de clusters (K)

Para determinar el número adecuado de clusters, se evaluaron valores de K en el rango de 2 a 10, utilizando una combinación de criterios:

5.3.1 Curva de Costos (K-Prototypes)

Se analizó la curva de costo (distorsión) del modelo K-Prototypes, identificando el punto a partir del cual la reducción del error se vuelve marginal.

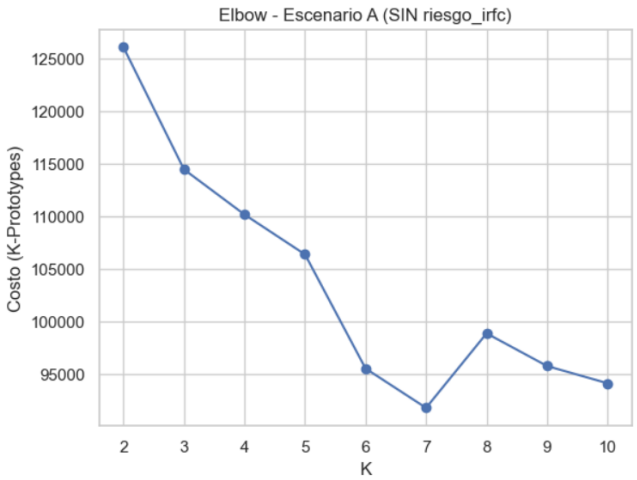

***Figura 10**. Curva de costo (K-Prototypes) – Escenario A (SIN riesgo_irfc)*
***Fuente:** Elaboración propia con Python.*

La **Figura 10** presenta la curva de costo (distorsión) del modelo K-Prototypes para el Escenario A, mostrando una disminución progresiva del costo a medida que aumenta el número de clusters, lo que refleja una mejora inicial en la estructuración del agrupamiento. Sin embargo, a partir de ciertos valores de K, la reducción del costo se vuelve menos pronunciada, evidenciando rendimientos marginales decrecientes. Este comportamiento permite identificar un punto de equilibrio entre complejidad, interpretabilidad y calidad analítica, donde incrementar el número de clusters deja de generar mejoras sustanciales. En este sentido, la curva funciona como una referencia preliminar para seleccionar valores razonables de K y comparar posteriormente si la incorporación del IRFC mejora la capacidad de segmentación.

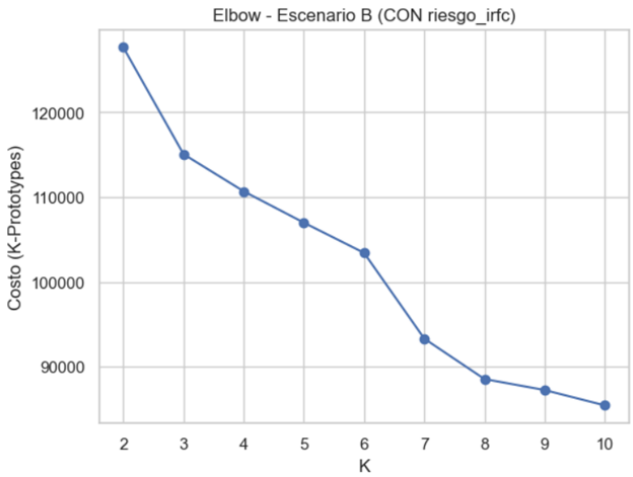

***Figura 11**. Curva de costo (K-Prototypes) – Escenario B (CON riesgo_irfc)*
***Fuente:** Elaboración propia con Python.*

La **Figura 11** presenta el escenario que incorpora la variable de riesgo, la curva presenta una caída más pronunciada en los primeros valores de K, lo que indica una mayor capacidad del modelo para capturar estructura en los datos. Esto sugiere que la inclusión del IRFC contribuye a una mejor organización de los clientes en grupos diferenciados.

5.3.2 Métricas internas de validación

Se evaluó la calidad de los clusters mediante:

- Silhouette Score (alcanza su valor máximo en K=4, indicando mejor cohesión y separación)
- Índice de Calinski–Harabasz (también presenta su valor más alto en K=4)
- Índice de Davies–Bouldin (es mínimo en K=4, lo que indica mejor calidad del clustering)

Dado el carácter mixto de los datos, la similitud entre observaciones se calculó utilizando distancia de Gower, aplicada sobre una muestra aleatoria estratificada para optimizar el costo computacional.

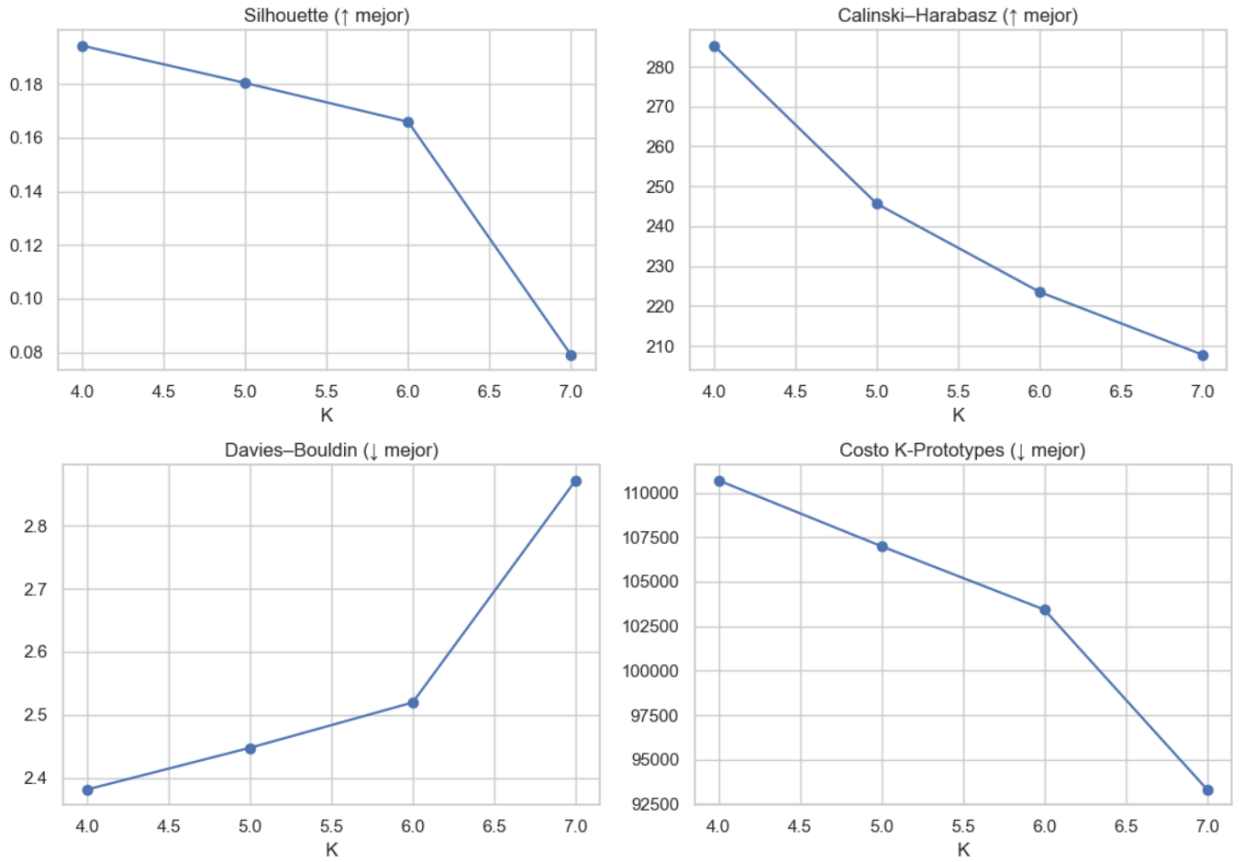

***Figura 12**. Métricas internas por K – Silhouette, Calinski–Harabasz y Davies–Bouldin (Escenario B)*
***Fuente:** Elaboración propia con Python.*

Como se observa en la **Figura 12**, las métricas internas presentan un comportamiento consistente que permite identificar un rango adecuado de valores de K. En particular, el índice de Silhouette alcanza su valor más alto en K=4 y disminuye progresivamente a medida que aumenta el número de clusters, lo que sugiere una mejor cohesión y separación en valores bajos de K.

De manera similar, el índice de Calinski–Harabasz muestra su valor máximo en K=4, decreciendo conforme aumenta K, lo que indica una menor capacidad de separación entre grupos al incrementar la cantidad de clusters. Por su parte, el índice de Davies–Bouldin presenta su valor más bajo en K=4 y aumenta progresivamente, lo que confirma que la calidad del agrupamiento se deteriora para valores mayores de K.

Finalmente, la curva de costo de K-Prototypes muestra una disminución continua, aunque con mejoras marginales a partir de K=4, lo que es consistente con el criterio del "codo". En conjunto, estos resultados sugieren que el valor óptimo de clusters se encuentra alrededor de K=4, ya que ofrece el mejor equilibrio entre cohesión interna, separación entre grupos y complejidad del modelo.

5.4 Comparación de escenarios y selección final

Se compararon los resultados de los dos escenarios definidos:

- **Escenario A (sin riesgo_irfc)**
- **Escenario B (con riesgo_irfc)**

Los resultados evidencian que el **Escenario B** presenta mayor capacidad de estructuración analítica, reflejado en:

- Mayor separación entre clusters
- Mayor cohesión interna
- Mejor comportamiento en métricas de validación

Esto indica que la incorporación de la variable de riesgo permite estructurar mejor los datos, generando grupos más interpretables y coherentes. En consecuencia, se selecciona el Escenario B como base para la segmentación final, junto con el valor de K óptimo identificado en las métricas.

5.5 Asignación de clusters y perfilamiento

Una vez definido el número óptimo de clusters, se procedió a asignar a cada cliente una etiqueta de grupo (*cluster_kproto_B*), permitiendo estructurar la base de datos en segmentos

diferenciados según patrones compartidos de comportamiento financiero, vinculación comercial y exposición relativa al riesgo.

Esta etapa constituye el núcleo estratégico del proceso de clustering, ya que el objetivo no se limita a clasificar observaciones en grupos estadísticos, sino a transformar dichos agrupamientos en perfiles interpretables y accionables desde una perspectiva analítica y de negocio. En este sentido, cada cluster se entiende como una representación estructurada de comportamientos diferenciados, útil para comprender heterogeneidad dentro de la base de clientes y orientar estrategias segmentadas de priorización.

Para el perfilamiento de los grupos se analizaron múltiples dimensiones, entre ellas:

- Promedios de variables numéricas relevantes
- Categorías predominantes o patrones de comportamiento comercial
- Distribución del IRFC como representación compuesta del riesgo
- Tasa de churn observada (Exited), utilizada como referencia interpretativa posterior

Este enfoque permitió pasar de una segmentación puramente algorítmica a una lectura estratégica de perfiles de cliente.

***Tabla 11**. Caracterización de Cluster kproto B (Sin Complain)*
***Fuente:** Elaboración propia con Python.*

cluster	*creditScore*	*age*	*tenure*	*balance*	*Num of products*	*Estimated salary*	*Is active mebers*	*Irfc*
0	*650.470*	*37.674*	*5.002*	*101061.460*	*1.000*	*99010.342*	*0.565*	*0.126*
1	*655.283*	*38.721*	*4.975*	*48720.980*	*2.012*	*100318.572*	*1.00*	*-0.521*
2	*645.230*	*44.924*	*4.938*	*92113.552*	*1471*	*101779.502*	*0.359*	*0.883*
3	*650.409*	*35.301*	*5.156*	*4747.901*	*2009*	*100098.839*	*0.000*	*-0.525*

En la **Tabla 11** se evidencia que, a partir de los resultados obtenidos, se identifican diferencias relevantes entre clusters, particularmente en variables como *balance*, *numofproducts*, *is_active_members* e *IRFC*, lo que evidencia que la segmentación logró capturar estructuras diferenciadas de comportamiento más allá de simples agrupaciones matemáticas.

Interpretación estratégica de perfiles:

- **Cluster 0:** Muestra balances altos, menor número de productos y niveles intermedios de actividad, lo que sugiere una exposición moderada al riesgo. Este segmento podría

representar clientes financieramente relevantes, pero con oportunidades de fortalecimiento en vinculación comercial.

- **Cluster 1 y 3:** Ambos presentan valores bajos o negativos del IRFC y una mayor vinculación comercial reflejada en el número de productos financieros, lo que sugiere una menor exposición relativa al churn. Sin embargo, aunque comparten niveles similares de riesgo compuesto, difieren en aspectos operativos relevantes. El Cluster 1 concentra clientes activos, mientras que el Cluster 3 agrupa clientes predominantemente inactivos. Esta diferencia sugiere que la actividad reciente del cliente constituye un factor diferenciador importante para el diseño de estrategias de retención, ya que clientes con perfiles financieros similares pueden exhibir comportamientos relacionales distintos frente a la entidad.

- **Cluster 2:** Presenta el valor más alto de IRFC, mayor edad promedio, balances elevados y menor proporción de clientes activos, configurando un perfil de mayor exposición relativa al riesgo. Este grupo puede interpretarse como un segmento potencialmente más vulnerable, caracterizado por condiciones financieras y comportamentales asociadas a mayor probabilidad de abandono o menor estabilidad relacional.

Estos resultados muestran que el clustering permitió identificar segmentos con combinaciones diferenciadas de:

- Presión financiera
- Diversificación de productos
- Nivel de actividad
- Riesgo compuesto

Más allá de clasificar clientes, este proceso facilita comprender cómo distintas configuraciones financieras y comerciales se relacionan con patrones de riesgo heterogéneos.

En este sentido, el valor principal del perfilamiento no consiste en generar una predicción individual definitiva, sino en construir una herramienta de segmentación estratégica que permita:

- Priorizar grupos de mayor exposición relativa al riesgo
- Diseñar estrategias de retención diferenciadas
- Orientar campañas comerciales según perfil
- Complementar el análisis supervisado con una visión estructural del cliente

Implicación para el negocio

Desde una perspectiva aplicada, los clusters permiten traducir resultados analíticos complejos en segmentos comprensibles para toma de decisiones. Por ejemplo:

- Clientes de alto IRFC y baja actividad → estrategias de retención prioritaria
- Clientes con alta vinculación y bajo IRFC → estrategias de fidelización
- Clientes intermedios → acciones preventivas o de cross-selling

De esta manera, el clustering se consolida como una herramienta de perfilamiento y priorización, capaz de conectar analítica avanzada con decisiones operativas más focalizadas.

5.6 Definición de perfiles de Riesgo

A partir del análisis de clusters y los percentiles del IRFC, se definieron perfiles de riesgo:

- Alto riesgo: Clientes con valores elevados de IRFC y características asociadas a mayor exposición al churn
- Riesgo medio: Comportamiento intermedio
- Bajo riesgo: Clientes estables con baja probabilidad de abandono

Estos perfiles no constituyen una predicción individual determinística, sino una herramienta de clasificación estratégica que permite priorizar poblaciones, orientar campañas diferenciadas y apoyar decisiones de retención bajo una lógica segmentada.

5.7 Discusión

Un hallazgo central del análisis es que la segmentación permitió identificar estructuras consistentes en los datos sin depender directamente de la variable objetivo, fortaleciendo su utilidad como herramienta de comprensión y perfilamiento. La exclusión de *Complain* garantizó que los grupos obtenidos respondieran a patrones financieros y comportamentales más genuinos, reduciendo el riesgo de agrupaciones artificialmente influenciadas por fuga de información.

Asimismo, la incorporación del IRFC enriqueció la capacidad interpretativa del clustering, al aportar una representación compuesta del riesgo que facilitó una segmentación más estructurada y coherente. Desde una perspectiva metodológica, el clustering no debe entenderse como sustituto directo del modelo supervisado final, sino como una capa analítica complementaria que fortalece procesos de interpretación, priorización y diseño de estrategias segmentadas de retención.

5.8 Conclusiones del Clustering

El proceso de clustering permitió identificar perfiles diferenciados de clientes, estructurando segmentos coherentes con distintas exposiciones relativas al riesgo financiero-comportamental.

La selección de K=4 ofreció un equilibrio adecuado entre simplicidad, interpretabilidad y calidad analítica, mientras que la incorporación del IRFC fortaleció la capacidad de perfilamiento sin depender de variables con riesgo de fuga de información. En conjunto, el clustering se consolida no como un modelo predictivo final, sino como una herramienta estratégica de segmentación, priorización y comprensión del churn, útil para enriquecer procesos de toma de decisiones y orientar futuras estrategias de retención.

6. Modelado Supervisado del Churn

Con el fin de evaluar el aporte del Índice de Riesgo Financiero-Comportamental (IRFC) y de la segmentación de clientes en el análisis predictivo de la deserción (churn), se definieron distintos escenarios de modelado supervisado que permiten comparar configuraciones basadas en variables originales, variables derivadas y herramientas de segmentación.

Más allá de medir únicamente mejoras en desempeño predictivo, este enfoque busca analizar de manera crítica si la incorporación de capas analíticas adicionales como el IRFC o el clustering aporta valor metodológico real en términos de discriminación, interpretación y aplicabilidad, o si, por el contrario, introduce riesgos asociados a redundancia, colinealidad o fuga de información.
En este contexto, dado que el IRFC constituye una variable compuesta construida a partir de variables del propio dataset, se desarrolló un análisis específico orientado a validar su uso dentro de modelos supervisados, evaluando cuidadosamente su impacto sobre la estabilidad metodológica del modelo y su pertinencia como predictor frente a su posible rol como herramienta complementaria de interpretación, ranking y segmentación.

6.1 Modelo Supervisado y Escenarios de comparación

El propósito del modelado supervisado fue evaluar la capacidad predictiva del churn bajo diferentes configuraciones de variables, contrastando escenarios con variables base, variables derivadas y segmentación. Este análisis permitió diferenciar entre escenarios con alto desempeño aparente y escenarios con mayor validez metodológica para una posible aplicación real.

Se definieron los siguientes escenarios de análisis:

- **Escenario A (Baseline):** Incluye variables financieras y comportamentales originales, sin incorporar el IRFC ni información de segmentación.
- **Escenario B (+IRFC):** Incorpora el índice IRFC como variable adicional, con el propósito de evaluar si esta medida sintética aporta capacidad discriminativa al modelo.
- **Escenario C (+IRFC + Clustering):** incorpora la variable de segmentación (*cluster_kproto_B*) para analizar si los perfiles obtenidos mediante clustering aportan información adicional.
- **Escenario D (Exploratorio):** Combina variables base, IRFC y segmentación, con el fin de evaluar el efecto conjunto de las variables derivadas.

Estos escenarios no se entienden como alternativas equivalentes para implementación inmediata, sino como configuraciones comparativas que permiten analizar el efecto de variables potencialmente problemáticas, como *Complain*, y de herramientas complementarias, como el

IRFC y el cluster. Posteriormente, el diagnóstico de colinealidad y la evaluación de fuga de información permiten determinar cuáles escenarios resultan más defendibles desde una perspectiva metodológica.

6.2 Modelos Evaluados

Se utilizaron tres modelos representativos de clasificación supervisada:

- **Regresión Logística:** Modelo base interpretable, útil para evaluar relaciones directas entre variables explicativas y churn.
- **Random Forest:** Modelo de ensamble capaz de capturar relaciones no lineales y reducir varianza.
- **XGBoost:** Modelo de gradient boosting con alta capacidad predictiva y buen desempeño en problemas de clasificación desbalanceada.

Para garantizar la reproducibilidad del estudio, la **Tabla 12** presenta la configuración de hiperparametros utilizados para el entrenamiento del modelo XGBoost. Los valores fueron definidos a partir de las pruebas experimentales orientadas a mantener un equilibrio entre capacidad predictiva y control del sobreajuste.

***Tabla 12**. Hiperparámetros del modelo XGBoost*
***Fuente:** Elaboración propia con Python.*

Hiperparámetros	Valor	Descripción general
n_estimators	200	Número de árboles que construye el modelo. Más árboles pueden mejorar el aprendizaje, pero aumentan el costo computacional.
max_depth	4	Profundidad máxima de cada árbol. Controla qué tan complejas pueden ser las reglas aprendidas.
learning_rate	0.05	Ritmo con el que el modelo aprende en cada iteración. Un valor bajo permite un aprendizaje más gradual y estable.
Subsample	0.8	Porcentaje de observaciones usadas para entrenar cada árbol. Ayuda a reducir sobreajuste.
colsample_bytree	0.8	Porcentaje de variables usadas en cada árbol. Ayuda a mejorar la generalización del modelo.
eval_metric	logloss	Métrica interna usada por XGBoost para evaluar el error durante el entrenamiento.
random_state	42	Semilla para garantizar reproducibilidad de los resultados.

6.3 Criterios de Evaluación (Métricas)

Dado que la variable objetivo presenta desbalance de clases, con aproximadamente 20% de clientes en condición de churn, la evaluación no se basó únicamente en métricas globales. Se utilizaron métricas orientadas a la correcta identificación de clientes en riesgo:

- AUC-ROC
- AUC-PR
- Recall
- F1-score
- Recall Top Decile
- Lift Top Decile

Estas métricas permiten evaluar el desempeño del modelo tanto desde una perspectiva estadística como operativa, considerando que en un contexto bancario el interés principal es priorizar adecuadamente a los clientes con mayor probabilidad de abandono para focalizar acciones de retención.

6.4 Diagnóstico de Colinealidad y Selección de Escenarios

Antes de comparar los diferentes escenarios de modelado, se realizó un diagnóstico de colinealidad con el fin de identificar posibles relaciones de dependencia entre las variables explicativas. Este análisis resulta especialmente relevante en el presente estudio debido a la incorporación de variables derivadas, como el IRFC y las variables construidas durante la fase de ingeniería de características, las cuales podrían contener información redundante respecto a las variables originales. La identificación temprana de estos problemas permite seleccionar configuraciones metodológicamente válidas y evitar que el desempeño de los modelos se vea afectado por redundancia o sobreposición de información.

***Tabla 13**. Variables VIF*
***Fuente:** Elaboración propia con Python.*

12	***Balance to salary***	***inf***
13	*Num products dev*	*inf*
14	*irfc*	*inf*
9	*complain*	*inf*
15	*Cluster kproto b*	*3.91113*
5	*numofproducts*	*2.1381*

7	*isactivemember*	*1.6360*
4	*balance*	*1.1683*
2	*age*	*1.1096*
8	*estimatedsalary*	*1.0040*
3	*tenure*	*1.0021*
1	*crediscore*	*1.0018*
6	*hascrcard*	*1.0014*
10	*Satisfaction score*	*1.0013*
11	*Point earned*	*1.0012*
0	*const*	*0.0000*

La **Tabla 13** presenta los resultados del análisis de colinealidad mediante el Factor de Inflación de la Varianza (VIF), aplicado a las variables consideradas en los escenarios de modelado. Se observa que variables derivadas como *balance_to_salary*, *num_products_dev*, *irfc* y *complain* presentan valores de VIF infinitos, lo que evidencia colinealidad perfecta. Este comportamiento se explica por la dependencia directa entre variables construidas y variables base incluidas en el modelo. En contraste, la variable de segmentación (*cluster_kproto_B*) presenta un VIF inferior a 5, lo que indica que no genera colinealidad significativa dentro del conjunto evaluado. Sin embargo, aunque este resultado respalda su estabilidad estadística, su uso como predictor final debe interpretarse con cautela y validarse en escenarios reales.

Este diagnóstico permite establecer una primera decisión metodológica: el IRFC no debe utilizarse como variable explicativa directa en modelos supervisados finales, debido a su redundancia con variables base. Su rol dentro del proyecto se conserva como herramienta analítica de interpretación, ranking y segmentación.

6.5 Pipeline de procesamiento

El pipeline de modelado se construyó con el fin de garantizar un proceso reproducible y consistente entre entrenamiento y evaluación. Para ello se utilizaron las siguientes transformaciones:

- Estandarización de variables numéricas mediante StandardScaler
- Codificación One-Hot Encoding para variables categóricas

- Integración de transformaciones mediante ColumnTransformer

Este diseño permite aplicar las transformaciones de forma controlada dentro del flujo de modelado, reduciendo el riesgo de inconsistencias y evitando fugas de información durante el proceso de entrenamiento y evaluación.

6.6 Resultados Iniciales del Modelo

Antes de profundizar en el análisis de los distintos escenarios de modelado, se realizó una evaluación inicial del desempeño predictivo mediante curvas ROC y Precision–Recall. Estas métricas permiten analizar la capacidad de discriminación de los modelos y detectar posibles comportamientos atípicos que requieran una revisión más detallada de las variables utilizadas

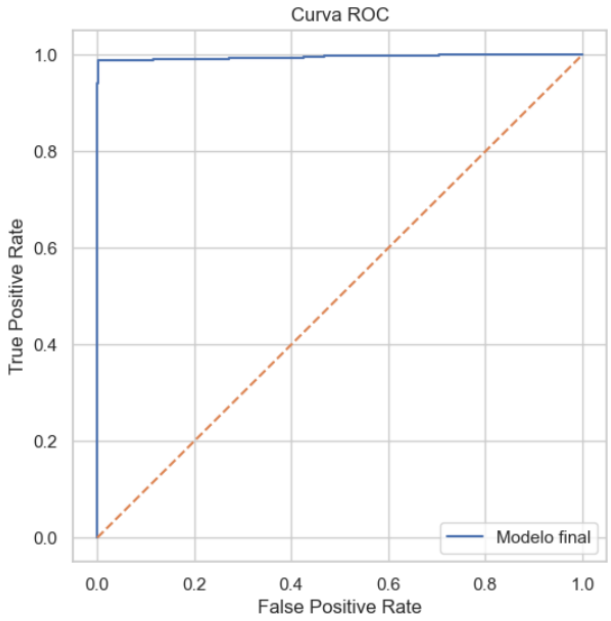

Figura 13. *Curva ROC*
Fuente: *Elaboración propia con Python.*

La **Figura 13** presenta la curva ROC de los modelos evaluados en los escenarios iniciales. Se observa una alta capacidad de discriminación, con curvas cercanas al vértice superior izquierdo. Aunque este comportamiento indica una fuerte separación entre clientes que abandonan y permanecen, también exige una lectura crítica, dado que desempeños excesivamente altos

pueden ser señal de variables con información demasiado cercana al evento de churn.

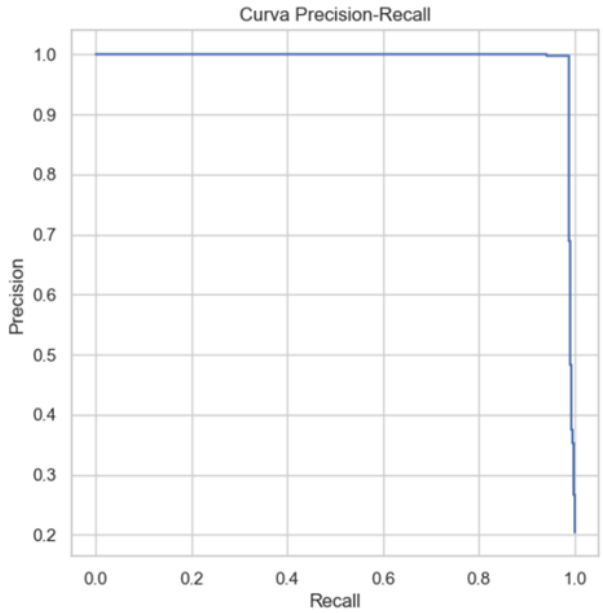

Figura 14. *Curva Precision Recall*
Fuente: *Elaboración propia con Python.*

La Figura 14 presenta la curva Precision-Recall, especialmente relevante en problemas desbalanceados. Los resultados muestran un desempeño elevado en la identificación de la clase minoritaria. Sin embargo, al igual que en la curva ROC, la magnitud de las métricas obliga a revisar si el modelo está capturando patrones genuinos o si se encuentra influenciado por variables con riesgo de fuga de información.

6.7 Evaluación operativa del modelo

Con el fin de complementar la evaluación del desempeño, se analizaron las matrices de confusión y métricas de clasificación bajo diferentes umbrales de decisión. Este análisis permite examinar el comportamiento operativo del modelo y verificar la estabilidad de sus resultados frente a distintos criterios de clasificación.

Tabla 14*. Matriz de Confusión*
Fuente: *Elaboración propia con Python.*

Threshold: 0.3				
Matriz de confusión:				
[[1591 1]				
[2 406]]				
Classification Report:				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	1592

1	1.00	1.00	1.00	408
accuracy			1.00	2000
Macro avg	1.00	1.00	1.00	2000
Weighted avg	1.00	1.00	1.00	2000
Threshold: 0.4				
Matriz de confusión:				
[[1591 1]				
[2 406]]				
Classification Report:				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	1592
1	1.00	1.00	1.00	408
accuracy				
Macro avg	1.00	1.00	1.00	2000
Weighted avg	1.00	1.00	1.00	2000
Threshold: 0.5				
Matriz de confusión:				
[[1591 1]				
[2 406]]				
Classification Report:				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	1592
1	1.00	1.00	1.00	408
accuracy			1.00	2000
Macro avg	1.00	1.00	1.00	2000
Weighted avg	1.00	1.00	1.00	2000
Threshold: 0.6				
Matriz de confusión:				
[[1591 1]				
[2 406]]				
Classification Report:				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	1592
1	1.00	1.00	1.00	408
accuracy			1.00	2000
Macro avg	1.00	1.00	1.00	2000
Weighted avg	1.00	1.00	1.00	2000

La **Tabla 14** muestra las matrices de confusión obtenidas bajo distintos umbrales de decisión. En todos los casos se observa un desempeño prácticamente perfecto, con muy pocos errores de clasificación tanto para clientes que permanecen como para clientes que abandonan. Aunque este resultado podría interpretarse inicialmente como favorable, en problemas reales de churn un desempeño tan cercano a la perfección resulta poco usual. Por esta razón, la matriz de confusión no se interpreta como una validación definitiva del modelo, sino como una evidencia adicional que refuerza la necesidad de revisar la presencia de fuga de información, especialmente por la influencia de la variable *Complain*.

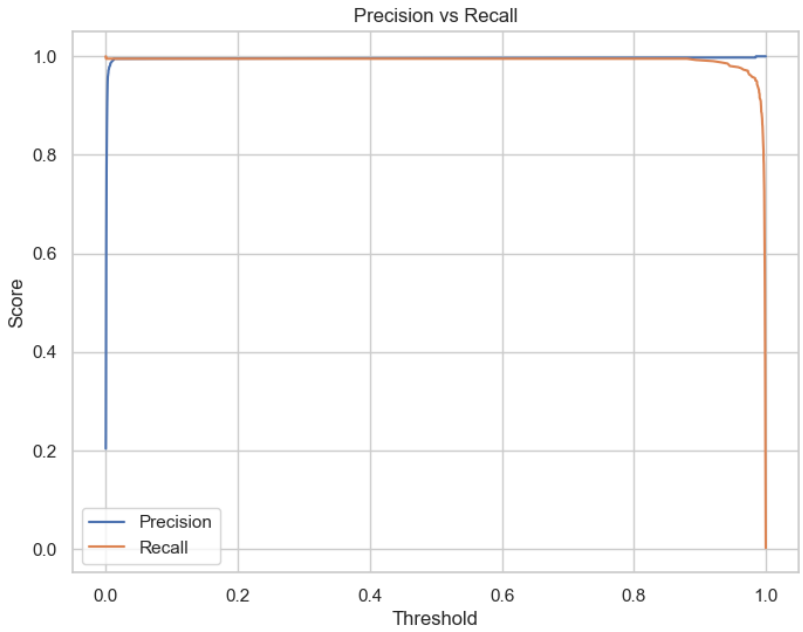

***Figura 15**. Curva Precision Recall vs threshold*
***Fuente:** Elaboración propia con Python.*

La **Figura 15** permite analizar cómo varían la precisión y el recall a medida que cambia el umbral de clasificación. En los resultados obtenidos, ambas métricas se mantienen cercanas a 1.0 en un rango amplio de umbrales, lo que evidencia una estabilidad inusual del modelo. En condiciones reales, suele existir un intercambio entre precisión y recall: al aumentar la cobertura de clientes en riesgo, normalmente se incrementa también el número de falsos positivos. La ausencia de este comportamiento refuerza la hipótesis de que el modelo inicial estaba influenciado por información demasiado cercana al evento de abandono.

6.8 Hallazgo Critico: Fuga de información asociada a Complain

El análisis evidenció que la variable *Complain* introduce un sesgo significativo en el modelo, generando métricas artificialmente altas debido a su cercanía con el evento de churn. Este

comportamiento ya había sido identificado en el análisis exploratorio, donde la variable mostró una asociación prácticamente perfecta con *Exited*. Al excluir *Complain*, el desempeño del modelo base disminuye a valores más realistas, con AUC-ROC aproximado entre 0.85 y 0.87. Este cambio confirma que los escenarios iniciales estaban sobreestimados y que el uso de *Complain* compromete la validez del modelo en escenarios reales.

Por esta razón, los modelos con *Complain* se interpretan como una evidencia metodológica del impacto de la fuga de información, pero no como alternativas válidas para implementación productiva.

6.9 Modelo sin *Complain*

Tras identificar el impacto de *Complain* en la generación de métricas artificialmente altas, se construyeron nuevos escenarios de modelado excluyendo dicha variable. Esta evaluación permite observar el desempeño del modelo en condiciones más realistas, cercanas a un escenario en el que no se dispone de información posterior o excesivamente relacionada con el evento de abandono.

En esta etapa se compararon dos configuraciones principales: el modelo base sin *Complain* y el modelo que incorpora la variable de segmentación (*cluster_kproto_B*) sin utilizar *Complain*.

***Tabla 15**. Resultado A_nc*
***Fuente:** Elaboración propia con Python.*

AUC ROC	AUC PR	Recall Top Decile	Lift Top Decile	Model	Scenario
0	0.847814	0.631411	0.715	3.504902	Logistic
1	0.867413	0.703909	0.825	4.044118	RandomForest
2	0.876293	0.728921	0.855	4.191176	XGBoost

La **Tabla 15** muestra los resultados del modelo base tras excluir *Complain*. Se observa una reducción importante en el desempeño respecto a los escenarios iniciales, con valores de AUC-ROC entre 0.84 y 0.87 y Recall Top Decile entre 0.71 y 0.85. Este comportamiento no debe interpretarse como una debilidad del modelo, sino como una evaluación más realista del problema. En este escenario, XGBoost alcanza el mejor desempeño entre los modelos evaluados, con AUC-ROC de 0.876 y AUC-PR de 0.729, lo cual representa una capacidad discriminativa adecuada sin depender de variables con riesgo de fuga de información.

***Tabla 16**. Resultado C_nc*
***Fuente:** Elaboración propia con Python.*

AUC ROC	AUC PR	Recall Top Decile	Lift Top Decile	Model	Scenario
0	0.946619	0.887807	0.98	4.803922	Logistic

1	0.992892	0.992787	1.00	4.901961	RandomForest
2	0.995109	0.992940	1.00	4.901961	XGBoost

La **Tabla 16** muestra los resultados del escenario que incorpora la variable de cluster, manteniendo excluida la variable *Complain*. Se observa una mejora relevante en todas las métricas, con AUC-ROC superior a 0.94 y *Recall* Top *Decile* cercano a 1.0. Este resultado evidencia que la segmentación captura patrones estructurales del comportamiento de los clientes que aportan información adicional al modelo. Sin embargo, dada la magnitud de las métricas, este escenario debe interpretarse como una alternativa comparativa de alto potencial, cuya adopción productiva requeriría validar la estabilidad temporal de los clusters y su independencia frente a variables derivadas.

Tabla 17. Comparación Completa
Fuente: *Elaboración propia con Python.*

AUC ROC	AUC PR	Recall Top Decile	Lift Top Decile	Model	Scenario
1	0.999567	0.998887	1.000	4.901961	RandomForest
0	0.999340	0.998290	1.000	4.901961	Logistic
1	0.999258	0.998133	1.000	4.901961	RandomForest
0	0.998648	0.997465	1.000	4.901961	Logistic
2	0.998194	0.997234	1.000	4.901961	XGBoost
2	0.997019	0.996636	1.000	4.901961	XGBoost
2	0.995109	0.992940	1.000	4.901961	XGBoost
1	0.992892	0.992787	1.000	4.901961	RandomForest
0	0.946619	0.887807	0.980	4.803922	Logistic
2	0.876293	0.728921	0.855	4.191176	XGBoost
1	0.867413	0.703909	0.825	4.044118	RandomForest
0	0.847814	0.631411	0.715	3.504902	Logistic

La **Tabla 17** resume los escenarios evaluados y permite distinguir tres comportamientos:

- Los modelos con *Complain* presentan métricas cercanas a 1.0, lo que evidencia sobreestimación por fuga de información.
- El modelo base sin *Complain* presenta un desempeño más moderado, pero metodológicamente más realista y defendible.
- El modelo con cluster sin *Complain* recupera capacidad predictiva, lo que muestra el valor analítico de la segmentación, aunque requiere validación adicional antes de considerarse como opción productiva final.

En consecuencia, el modelo base sin *Complain* se considera el escenario más prudente para una primera validación real, mientras que el escenario con cluster se conserva como una alternativa analítica prometedora para exploración y validación posterior.

6.10 Evaluación avanzada del modelo

Con el fin de complementar la evaluación mediante métricas globales, se incorporaron herramientas gráficas que permiten analizar el modelo desde una perspectiva operativa. En particular, se revisaron las curvas *ROC* y *Precision-Recall*, así como las curvas de ganancias y *lift*, orientadas a evaluar la capacidad del modelo para priorizar clientes con mayor probabilidad de churn.

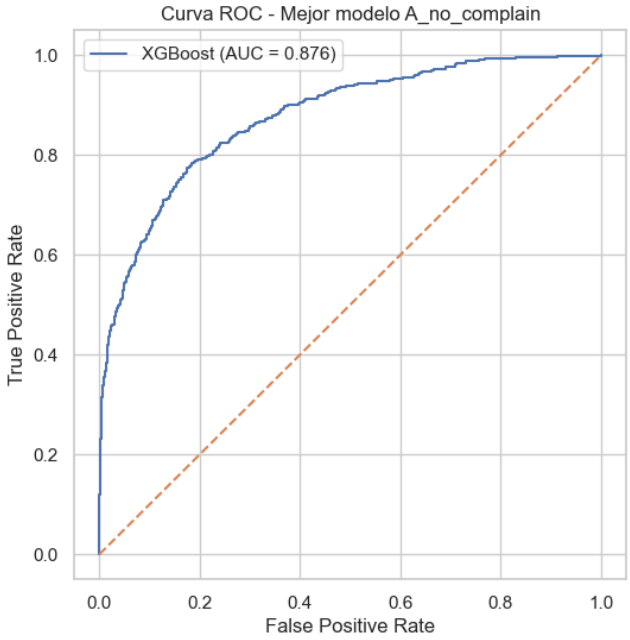

***Figura 16**. Curva ROC-Mejor Modelo*
***Fuente:** Elaboración propia con Python.*

La **Figura 16** muestra la curva ROC del mejor modelo evaluado en el escenario sin *Complain*. Se observa una adecuada capacidad de discriminación entre clientes que abandonan y clientes que permanecen. Este resultado confirma que, incluso sin la variable problemática, el modelo mantiene capacidad predictiva relevante para ordenar clientes según su riesgo relativo de abandono.

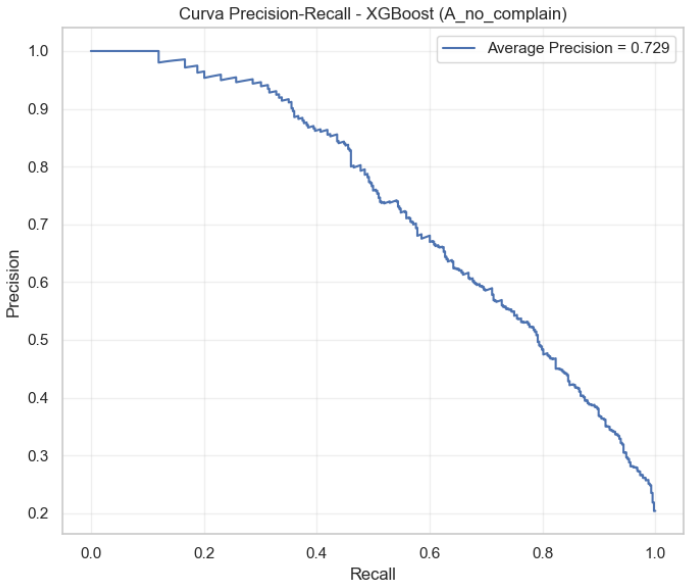

***Figura 17**. Curva Precision Recall -XGBoost*
***Fuente:** Elaboración propia con Python.*

La **Figura 17** presenta la curva *Precision-Recall* del modelo XGBoost en el escenario sin *Complain*. Esta curva resulta especialmente importante porque evalúa el desempeño sobre la clase minoritaria, es decir, los clientes que efectivamente presentan churn. El valor de Average Precision cercano a 0.729 indica que el modelo logra mantener un equilibrio razonable entre precisión y *recall*, siendo útil para priorizar clientes en riesgo sin depender de la variable *Complain*.

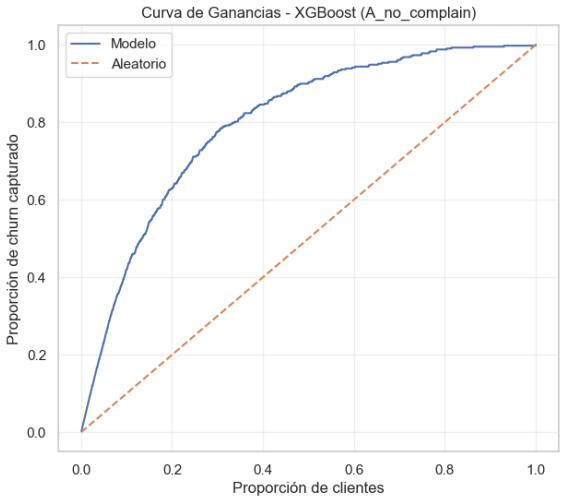

***Figura 18**. Curva de Ganancia – XGBoost*
***Fuente:** Elaboración propia con Python.*

La **Figura 18** muestra la curva de ganancia acumulada del modelo. Se observa que el modelo logra concentrar una proporción importante de clientes con churn en los primeros percentiles de la población priorizada. Este resultado tiene una lectura operativa relevante: en lugar de intervenir

toda la base de clientes, la entidad podría focalizar acciones en los segmentos con mayor probabilidad estimada de abandono, optimizando recursos comerciales.

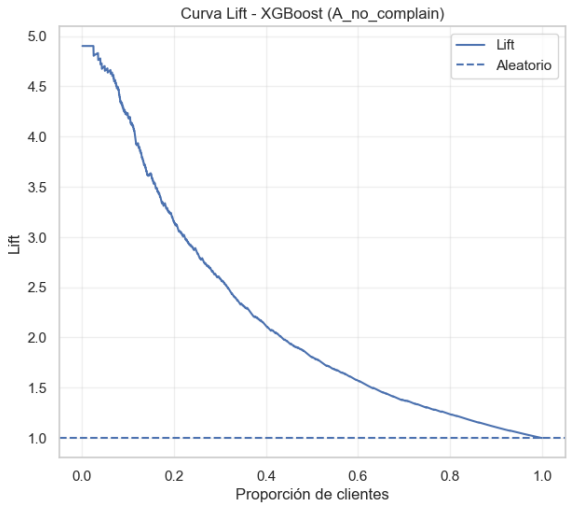

***Figura 19**. Curva Lift – XGBoost*
***Fuente:** Elaboración propia con Python.*

La **Figura 19** presenta la curva de lift, que compara el desempeño del modelo frente a una selección aleatoria. Los valores superiores a 1 en los primeros percentiles indican que el modelo prioriza clientes en riesgo de manera más eficiente que una estrategia no informada. Este resultado refuerza el valor operativo del modelo como herramienta de priorización, especialmente en contextos donde la capacidad de contacto o intervención comercial es limitada.

6.11 Comparación Final de Escenarios

Con el fin de consolidar los hallazgos obtenidos a lo largo del proceso de modelado, se presenta una comparación de los principales escenarios evaluados. Este análisis permite contrastar el efecto de variables como *Complain* y *cluster* sobre el desempeño del modelo, así como valorar sus implicaciones desde una perspectiva metodológica y aplicada.

***Tabla 18**. Comparación clave de escenarios*
***Fuente:** Elaboración propia con Python.*

Escenario	Características	Desempeño	Interpretación
Modelo A (base)	Variables originales (incluye *Complain*)	Métricas extremadamente altas (AUC ≈ 0.99, Recall = 1.0)	Resultado sobreestimado debido a posible fuga de información
Modelo A (sin ***Complain***)	Variables originales sin *Complain*	Desempeño moderado (AUC ≈ 0.85–0.87)	Representa un escenario más realista
Modelo C (sin ***Complain*** + cluster)	Variables originales + segmentación	Alto desempeño (AUC ≈ 0.99, Recall ≈ 1.0)	Escenario comparativo de alto desempeño; requiere validación temporal y metodológica antes de considerarse productivo

La **Tabla 18** sintetiza los principales escenarios evaluados y permite establecer una lectura metodológica de los resultados. El modelo con *Complain* alcanza métricas extremadamente altas, pero estas no se consideran representativas de un escenario real debido al riesgo de fuga de información. El modelo base sin *Complain* presenta un desempeño más moderado, pero constituye el escenario más realista y defendible desde el punto de vista metodológico. Por su parte, el modelo con cluster sin *Complain* muestra una recuperación importante del desempeño, lo cual confirma el valor de la segmentación como capa analítica complementaria.

Sin embargo, para efectos de selección final, se adopta una posición conservadora: el modelo base sin *Complain* se plantea como la alternativa inicial más prudente para validación comercial, mientras que el escenario con cluster se conserva como una línea de profundización y validación futura.

6.12 Conclusión del modelo e Insight del Negocio

El análisis desarrollado permite concluir que el desempeño predictivo no debe evaluarse únicamente a partir de métricas elevadas, sino también considerando la validez metodológica de las variables utilizadas. En este sentido, los escenarios con *Complain* demostraron cómo una variable altamente cercana al evento de churn puede producir resultados artificialmente altos y comprometer la generalización del modelo. De igual manera, el diagnóstico de colinealidad mostró que el IRFC no debe incorporarse como predictor directo en modelos supervisados finales, debido a su naturaleza compuesta y su dependencia de variables base. Su valor principal dentro del proyecto se mantiene como herramienta de interpretación, ranking y segmentación. La variable de cluster mostró un aporte importante en los escenarios comparativos, al capturar patrones estructurales de comportamiento y riesgo. No obstante, dado que su desempeño es muy alto, su uso como predictor productivo debe validarse con pruebas adicionales de estabilidad, temporalidad y aplicación en contexto real. En este sentido, el escenario con cluster se interpreta como evidencia del potencial analítico de la segmentación, más que como una adopción automática para implementación inicial

Por tanto, el modelo base sin *Complain* se considera el escenario final más prudente para una primera validación comercial, especialmente con XGBoost, dado que ofrece un desempeño realista y defendible sin depender de variables asociadas a fuga de información. A su vez, el IRFC y el clustering se conservan como capas analíticas complementarias, útiles para comprender el riesgo, perfilar clientes y orientar estrategias diferenciadas de retención.

Desde una perspectiva de negocio, el principal aporte del modelo no es reemplazar el criterio comercial, sino ofrecer una herramienta de priorización. Esto permite identificar clientes con mayor probabilidad de abandono y enfocar esfuerzos de retención en segmentos donde la intervención puede generar mayor impacto.

6.13 Propuesta de validación comercial

Con el propósito de trasladar los hallazgos del modelo a un entorno aplicado, se propone una validación comercial controlada de tres meses, orientada a evaluar la utilidad real del escenario base sin *Complain* como herramienta de priorización en estrategias de retención. La población objetivo estaría conformada por clientes activos evaluados con el modelo base, quienes serían ordenados según su probabilidad estimada de churn, priorizando inicialmente el top decil o percentiles superiores según la capacidad operativa de la entidad.

El piloto debería estructurarse mediante un grupo tratamiento y un grupo control comparable, con el fin de medir de manera objetiva el impacto incremental de las acciones comerciales implementadas. Esta aproximación permitiría validar no solo el desempeño analítico del modelo, sino también su capacidad para generar resultados medibles en términos de retención, eficiencia comercial y rentabilidad operativa.

Con el fin de traducir el análisis predictivo en decisiones comerciales diferenciadas, se propone complementar esta validación con una matriz estratégica de priorización que combine la probabilidad estimada de churn con el perfil de riesgo financiero del cliente. Esta herramienta permite ajustar las estrategias según el equilibrio entre probabilidad de abandono, valor comercial y nivel de exposición financiera, evitando intervenciones homogéneas sobre poblaciones con características distintas. Los indicadores sugeridos para evaluar el piloto incluyen tasa de retención incremental, tasa de aceptación de oferta, permanencia a 90 días, costo por cliente retenido, activación o uso de productos, así como diferencia de resultados frente al grupo control. En conjunto, estos indicadores permitirían evaluar la viabilidad comercial del modelo bajo condiciones reales.

***Tabla 19**. Matriz estratégica de priorización comercial según probabilidad de churn y perfil de riesgo financiero*
***Fuente:** Elaboración propia.*

Probabilidad de churn	Perfil de riesgo financiero	Lectura de negocio	Estrategia recomendada	KPI
Alta probabilidad de churn	Bajo riesgo crediticio / buen perfil financiero	Cliente valioso y defendible comercialmente	Retención prioritaria: contacto ejecutivo, oferta personalizada, beneficios de permanencia, mejora de experiencia, campaña de fidelización.	Tasa de retención, aceptación de oferta, permanencia a 90 días
Alta probabilidad de churn	Riesgo medio	Cliente recuperable con control de exposición	Retención selectiva: beneficios no crediticios, educación financiera, revisión de productos, acompañamiento comercial.	Retención, uso de productos, activación
Alta probabilidad de churn	Alto riesgo crediticio	Cliente con riesgo comercial y financiero	No incentivar endeudamiento. Gestión de servicio, escucha de quejas, alternativas de normalización, monitoreo preventivo.	Reducción de quejas, deterioro financiero, permanencia sin aumento de exposición
Probabilidad	Bajo riesgo	Oportunidad de	Campanas de vinculación, venta	Activación, venta

media de churn		crecimiento preventivo	cruzada, beneficios por uso, comunicaciones personalizadas.	cruzada, uso de canales
Baja probabilidad de churn	Bajo riesgo	Cliente estable y potencialmente rentable	Fidelización ligera, programas de lealtad, ofertas de valor sin intervención intensiva.	Valor de Vida del Cliente (CLV), permanencia, NPS
Baja probabilidad de churn	Alto riesgo	Cliente estable pero financieramente sensible	Monitoreo, alertas tempranas y gestión prudente de cupos o productos.	Mora temprana, uso responsable, estabilidad

La **Tabla 19** transforma los resultados del modelo en una herramienta práctica de decisión comercial, al combinar la probabilidad de churn con el perfil de riesgo financiero para definir estrategias diferenciadas de intervención. Su principal valor consiste en permitir que la organización no solo identifique clientes con mayor probabilidad de abandono, sino que ajuste sus acciones según el nivel de riesgo y valor comercial de cada segmento. En este sentido, la matriz evidencia que una alta probabilidad de churn no siempre implica la misma respuesta: clientes con alto churn y bajo riesgo financiero pueden justificar estrategias intensivas de retención, mientras que aquellos con alta probabilidad de abandono y mayor riesgo financiero requieren enfoques más prudentes, orientados a contención y sostenibilidad.

De igual forma, segmentos con probabilidad media o baja permiten diseñar acciones de crecimiento, fidelización o monitoreo preventivo, ampliando la utilidad del modelo más allá de la retención reactiva. En conjunto, esta matriz fortalece la aplicabilidad del trabajo al conectar analítica predictiva con decisiones comerciales concretas, demostrando que el verdadero valor del modelo radica no solo en predecir churn, sino en orientar estrategias segmentadas, eficientes y financieramente coherentes.

7. Conclusiones y Trabajos futuros

7.1 Conclusiones

El desarrollo del presente trabajo permitió diseñar y validar un enfoque integral para el análisis del churn bancario, articulando técnicas de ingeniería de variables, segmentación de clientes y aprendizaje supervisado dentro de una misma arquitectura metodológica. Más allá de la construcción de un modelo predictivo, el proyecto permitió desarrollar un proceso analítico orientado a comprender el riesgo de deserción desde múltiples dimensiones, integrando capacidad predictiva, interpretación y aplicabilidad estratégica.

Uno de los hallazgos más relevantes fue la identificación de métricas excesivamente altas en las etapas iniciales del modelado, situación que condujo a una revisión crítica de las variables utilizadas y a la detección de un problema de fuga de información asociado a la variable *complain*. La exclusión de esta variable permitió obtener resultados más realistas, metodológicamente más sólidos y potencialmente más generalizables, evidenciando que en proyectos de ciencia de datos el desempeño aparente no debe interpretarse automáticamente como sinónimo de validez. Este resultado refuerza la importancia de evaluar críticamente la naturaleza, temporalidad y relación estructural de las variables antes de su incorporación en escenarios productivos.

En relación con el Índice de Riesgo Financiero-Comportamental (IRFC), se concluye que su principal aporte no radica en su uso como predictor final dentro de modelos supervisados, dado que su naturaleza compuesta introduce colinealidad y redundancia con variables base. Sin embargo, el IRFC sí representa una herramienta analítica valiosa para sintetizar múltiples dimensiones del riesgo del cliente, facilitando procesos de interpretación, ranking, priorización y segmentación. En este sentido, su valor metodológico se ubica principalmente en la comprensión estructurada del riesgo más que en la predicción directa.

Por su parte, la segmentación mediante clustering permitió identificar grupos de clientes con patrones diferenciados de comportamiento financiero y comercial, aportando una capa adicional de comprensión estratégica sobre la heterogeneidad del churn. En particular, la variable de cluster funcionó como una representación estructurada del riesgo, al condensar patrones de comportamiento que no resultan evidentes al observar únicamente variables individuales. Esta capacidad permitió enriquecer el análisis del cliente desde una perspectiva de perfilamiento y priorización.

Adicionalmente, se evidenció que la incorporación de la segmentación en escenarios supervisados comparativos puede contribuir a recuperar parte del desempeño predictivo incluso en ausencia de variables problemáticas como *complain*. No obstante, este hallazgo debe interpretarse con prudencia metodológica: más que asumir automáticamente al cluster como predictor productivo definitivo, su principal valor se encuentra en fortalecer procesos de segmentación estratégica, priorización y diseño de intervenciones diferenciadas, cuya implementación final requiere validación adicional en escenarios reales.

En consecuencia, el principal aporte del trabajo no consiste únicamente en alcanzar métricas elevadas, sino en haber construido un enfoque integral metodológicamente sólido, capaz de diferenciar entre variables problemáticas, herramientas de interpretación y mecanismos de segmentación estratégicamente útiles. Bajo esta lógica, el escenario basado en variables originales sin *complain* constituye la alternativa más prudente como base para validación real, mientras que el IRFC y el clustering se consolidan como herramientas complementarias de interpretación, perfilamiento y priorización.

Desde una perspectiva aplicada, este trabajo demuestra que la integración de técnicas de analítica avanzada puede fortalecer significativamente la toma de decisiones en contextos financieros, no solo al mejorar la identificación de clientes en riesgo, sino también al facilitar estrategias de retención más focalizadas, eficientes y estratégicamente accionables.

Finalmente, la investigación evidencia que el verdadero valor de un modelo de churn no radica exclusivamente en su precisión estadística, sino en su capacidad para sostenerse metodológicamente, generalizar en contextos reales y traducirse en decisiones útiles para negocio. En este sentido, el proyecto aporta una visión más amplia de la analítica predictiva, donde predicción, interpretación y estrategia convergen como componentes complementarios para la gestión integral del riesgo de deserción.

7.2 Trabajos Futuros

Si bien los resultados obtenidos permiten construir una aproximación metodológicamente sólida para el análisis del churn bancario, existen limitaciones y oportunidades de profundización que pueden fortalecer futuras investigaciones y facilitar una eventual implementación aplicada.

En primer lugar, el análisis se desarrolló sobre un dataset específico, por lo que la generalización de los hallazgos a otras entidades financieras, segmentos de clientes o contextos de mercado debe realizarse con cautela. El comportamiento del churn puede variar significativamente según factores como tipo de producto, nivel de digitalización, perfil socioeconómico del cliente y condiciones competitivas del entorno. Por ello, una línea futura relevante consiste en validar el modelo en bases de datos de distintas instituciones o contextos geográficos, con el fin de evaluar su estabilidad externa y capacidad de adaptación.

En segundo lugar, la identificación de la variable *complain* como fuente potencial de fuga de información evidenció la importancia crítica de la temporalidad de los datos en proyectos predictivos. Aunque este hallazgo permitió depurar el modelo, no se contó con trazabilidad temporal completa para todas las variables, lo que limita la posibilidad de garantizar de forma absoluta la ausencia de señales futuras o variables registradas posterior al evento de churn. En este sentido, futuros trabajos podrían incorporar estructuras de datos longitudinales o series temporales que permitan validar explícitamente el momento de generación de cada variable y

construir modelos más cercanos a escenarios reales de predicción anticipada.

En relación con el IRFC, aunque el índice mostró valor como herramienta de interpretación, priorización y segmentación, su naturaleza compuesta limita su incorporación directa en modelos supervisados finales debido a colinealidad y redundancia. A partir de ello, investigaciones futuras podrían explorar metodologías alternativas de construcción de índices, incluyendo técnicas de reducción dimensional supervisada, regularización o enfoques basados en aprendizaje representacional, orientados a conservar interpretabilidad sin comprometer estabilidad estadística.

Por otra parte, el proceso de clustering demostró utilidad para el perfilamiento estratégico de clientes; sin embargo, la segmentación obtenida depende de decisiones metodológicas como selección de variables, número de clusters y parámetros del algoritmo. Esto implica que los perfiles identificados no deben considerarse estructuras definitivas, sino aproximaciones analíticas sujetas a refinamiento. Como trabajo futuro, sería valioso incorporar análisis de estabilidad temporal de clusters, validación con datos externos o comparación con otros enfoques de segmentación para datos mixtos. Asimismo, aunque la segmentación mostró potencial para fortalecer escenarios predictivos, su adopción como componente productivo requiere validación adicional. En este sentido, una extensión natural del trabajo consiste en desarrollar pilotos comerciales controlados que permitan medir el impacto real del modelo sobre estrategias de retención, incluyendo indicadores como retención incremental, costo por cliente retenido, permanencia posterior a la intervención y retorno sobre inversión.

Desde una perspectiva ética y de gobernanza analítica, futuras investigaciones también podrían profundizar en análisis de equidad algorítmica, evaluando el desempeño del modelo por subgrupos como género, geografía o segmento comercial. Esto permitiría identificar posibles diferencias en priorización o cobertura, fortaleciendo el uso responsable de modelos de churn en contextos financieros.

Finalmente, el modelo desarrollado no fue implementado en un entorno productivo real, por lo que no fue posible evaluar su comportamiento ante cambios dinámicos en patrones de clientes, deriva de datos (*data drift*) o necesidad de actualización continua. Por ello, una línea futura especialmente relevante consiste en diseñar arquitecturas de monitoreo, recalibración y mantenimiento del modelo que permitan su adaptación sostenida en escenarios operativos. En conjunto, estos trabajos futuros refuerzan que el presente estudio constituye una base metodológica robusta, pero también una plataforma inicial para desarrollos posteriores orientados a escalabilidad, validación externa, implementación comercial y gobernanza analítica.

Más que cerrar el problema del churn, esta investigación abre una ruta estructurada para su evolución hacia soluciones predictivas más maduras, transferibles y estratégicamente accionables.

8. Referencias Bibliográficas

[1] Boston Consulting Group, *Global Payments Report 2023*. [En línea]. Disponible en: https://www.bcg.com/publications/2023/bcg-global-payments-report-2023.

[2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324

[3] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626–4636, 2009. doi: 10.1016/j.eswa.2008.05.027

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794. doi: 10.1145/2939672.2939785

[5] Congreso de la República de Colombia, *Ley 1581 de 2012*. [En línea]. Disponible en: http://www.secretariasenado.gov.co/senado/basedoc/ley_1581_2012.html

[6] Presidencia de la República de Colombia, *Decreto 1377 de 2013*. [En línea]. Disponible en: https://sedeelectronica.sic.gov.co/transparencia/normativa/decreto-1377.

[7] K. Coussement and D. Van den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," *Expert Systems with Applications*, vol. 34, no. 1, pp. 313–327, 2008. doi: 10.1016/j.eswa.2006.09.038

[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD-96)*, 1996, pp. 226–231.

[9] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. doi: 10.1016/j.patrec.2005.10.010

[10] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 8th ed. Cengage, 2019.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[12] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.

[13] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion?," *Journal of Classification*, vol. 31, pp. 274–295, 2014. doi: 10.1007/s00357-014-9161-z

[14] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*, 2nd ed. Pearson, 2019.

[15] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit-driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012. doi: 10.1016/j.ejor.2011.09.031

[16] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

[17] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE, 10(3).

[18] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann.

[19] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly.

[20] Provost, F., & Fawcett, T. (2013). Data Science for Business. O'Reilly.

[21] Gower, J. C. (1971). *A general coefficient of similarity and some of its properties*. Biometrics, 27(4), 857–871.

[22] Huang, Z. (1997). *Clustering large data sets with mixed numeric and categorical values*. Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining.

[23] Rousseeuw, P. J. (1987). *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 20, 53–65.

[24] Calinski, T., & Harabasz, J. (1974). *A dendrite method for cluster analysis*. Communications in Statistics.

[25] Davies, D. L., & Bouldin, D. W. (1979). *A cluster separation measure*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[26] Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). *Propensity to customer churn in a financial institution: A machine learning approach. Neural Computing and Applications, 34, 11751–11768.*

[27] Superintendencia Financiera de Colombia. (2023). *Reporte de inclusión financiera y transformación digital del sistema financiero colombiano. Bogotá, Colombia.*

[28] Asobancaria. (2024). *Informe de innovación, analítica y experiencia del cliente en el sector bancario colombiano. Bogotá, Colombia.*

[29] Inter-American Development Bank (IDB). (2023). *Fintech in Latin America and the Caribbean: A consolidated ecosystem for digital financial transformation. Washington, DC: IDB.*

[30] Congreso de Colombia. (2012). Ley Estatutaria 1581 de 2012. *Por la cual se dictan disposiciones generales para la protección de datos personales. Diario Oficial No. 48.587.*

[31] Presidencia de la República de Colombia. (2013). Decreto 1377 de 2013. *Por el cual se reglamenta parcialmente la Ley 1581 de 2012. Diario Oficial No. 48.834.*

[32] Presidencia de la República de Colombia. (2015). *Decreto 1074 de 2015. Decreto Único Reglamentario del Sector Comercio, Industria y Turismo.*

[33] Superintendencia de Industria y Comercio. (2020). *Guía para la implementación del principio de responsabilidad demostrada (accountability) en protección de datos personales.*

[34] Superintendencia Financiera de Colombia. (2021). *Circular Básica Jurídica – Instrucciones relacionadas con gestión de riesgos y seguridad de la información.*

9. Anexos

Anexos 1. Notebook “Análisis de Clustering y Modelado de Deserción de Clientes Bancarios”

El presente anexo contiene el notebook desarrollado en Python, el cual integra el proceso completo de análisis, preparación de datos, construcción del índice IRFC, segmentación mediante clustering y modelado supervisado para la predicción del churn. Este material constituye el soporte técnico del estudio y permite la reproducción de los resultados presentados en el documento principal.

Para fortalecer la reproducibilidad del trabajo, el código fuente y las dependencias requeridas se encuentran disponibles en el siguiente repositorio de Google Colab:

https://colabboxplo.research.google.com/drive/1LMYgX42VSkc40W21GcG4brx1Wraiq4DJ?usp=sharing

Anexos 2. Dataset “Bank Customer Churn Prediction”

El presente anexo corresponde al dataset utilizado en el desarrollo del proyecto, obtenido de la plataforma Kaggle.

- Nombre: Bank Customer Churn Prediction
- Fuente: Kaggle
- Enlace: https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn/data

El dataset contiene información relacionada con variables demográficas, financieras y de comportamiento de clientes bancarios, las cuales fueron utilizadas para la construcción del modelo de predicción de churn.

Anexos 3. Diccionario de Variables originales y derivadas

Diccionario de variables del dataset original

Variable	Tipo de dato	Categoría	Descripción	Rango/Categorías
RowNumber	Entero	Identificación	Índice secuencial del registro	1–10000
CustomerId	Entero	Identificación	Identificador único del cliente	Valor único
Surname	Texto	Identificación	Apellido del cliente	Texto
CreditScore	Numérica	Financiera	Puntaje crediticio del cliente	350–850

Geography	Categórica	Demográfica	País de residencia del cliente	France, Germany, Spain
Gender	Categórica	Demográfica	Género del cliente	Male, Female
Age	Numérica	Demográfica	Edad del cliente	18–92 años
Tenure	Numérica discreta	Relación con el banco	Antigüedad del cliente en años	0–10
Balance	Numérica	Financiera	Saldo disponible en la cuenta	≥ 0
NumOfProducts	Numérica discreta	Productos	Número de productos contratados	1–4
HasCrCard	Binaria	Productos	Indica si posee tarjeta de crédito	0 = No, 1 = Sí
IsActiveMember	Binaria	Comportamiento	Indica si el cliente es activo	0 = No, 1 = Sí
EstimatedSalary	Numérica	Financiera	Ingreso estimado del cliente	> 0
Exited	Binaria	Objetivo	Indica si el cliente abandonó la entidad	0 = Permanece, 1 = Abandona
Complain	Binaria	Comportamiento	Registro de reclamación o queja del cliente	0 = No, 1 = Sí
Satisfaction Score	Numérica discreta	Experiencia	Nivel de satisfacción del cliente	1–5
Card Type	Categórica	Productos	Tipo de tarjeta del cliente	Silver, Gold, Platinum, Diamond
Point Earned	Numérica	Fidelización	Puntos acumulados por el cliente	Variable

Diccionario de variables derivadas construidas en el estudio

Variable	**Tipo de dato**	**Descripción**	**Método de construcción**
balance_to_salary	Numérica	Relación entre saldo e ingreso estimado	Balance / EstimatedSalary
num_products_dev	Numérica	Desviación del número de productos respecto a un valor esperado	Transformación derivada
credit_risk	Numérica	Puntaje crediticio orientado al riesgo	Transformación de CreditScore
satisfaction_risk	Numérica	Satisfacción orientada al riesgo	Inversión o transformación de Satisfaction Score
complain_risk	Numérica	Variable de queja orientada al riesgo	Transformación de Complain
complain_bin	Binaria	Codificación binaria de reclamaciones	Conversión de Complain
IRFC	Numérica	Índice de Riesgo Financiero-Comportamental	Combinación de señales financieras y comportamentales
riesgo_irfc	Categórica/Ordinal	Nivel de riesgo derivado del IRFC	Segmentación por niveles
cluster_kproto_B	Categórica	Etiqueta del cluster generado mediante K-Prototypes	Algoritmo K-Prototypes

Las variables derivadas fueron construidas como parte del proceso de ingeniería de características y análisis exploratorio. Su finalidad principal fue apoyar tareas de interpretación, segmentación y evaluación metodológica. En particular, variables asociadas al IRFC y a la segmentación no fueron necesariamente incorporadas como predictores finales en el modelo supervisado, priorizando criterios de validez metodológica y aplicabilidad en escenarios reales.