FACULTAD DE POSTGRADO

TRABAJO FINAL DE GRADUACIÓN

SISTEMA DE PRIORIZACIÓN COMERCIAL INTELIGENTE: INTEGRACIÓN DE ANALÍTICA PREDICTIVA PARA LA GESTIÓN DE VENTAS CRUZADAS EN BANCA DE PERSONAS (HONDURAS, 2023–2025)

SUSTENTADO POR:

ALAN FABRICIO BARAHONA LÓPEZ

EDRAS JOSUÉ GARCÍA VALLEJO

PREVIA INVESTIDURA AL TÍTULO DE

MÁSTER EN

ANALÍTICA DE NEGOCIOS

TEGUCIGALPA, FRANCISCO MORAZAN, HONDURAS, C.A.

FEBRERO, 2026

UNIVERSIDAD TECNOLÓGICA CENTROAMERICANA

UNITEC

FACULTAD DE POSTGRADO

AUTORIDADES UNIVERSITARIAS

RECTORA

ROSALPINA RODRÍGUEZ

VICERRECTOR ACADÉMICO NACIONAL

JAVIER ABRAHAM SALGADO LEZAMA

SECRETARIO GENERAL

ROGER MARTÍNEZ MIRALDA

DECANA FACULTAD DE POSTGRADO

ANA DEL CARMEN RETTALLY VARGAS

SISTEMA DE PRIORIZACIÓN COMERCIAL INTELIGENTE: INTEGRACIÓN DE ANALÍTICA PREDICTIVA PARA LA GESTIÓN DE VENTAS CRUZADAS EN BANCA DE PERSONAS (HONDURAS, 2023–2025)

TRABAJO PRESENTADO EN CUMPLIMIENTO DE LOS REQUISITOS EXIGIDOS PARA OPTAR AL TÍTULO DE

MÁSTER EN

ANALÍTICA DE NEGOCIOS

ASESOR

JESÚS RICARDO RODRÍGUEZ RIVERA

MIEMBROS DE LA TERNA:

JOSE ALEJANDRO COLINDRES GALINDO

JOSUE LUIS MEJIA RIVERA

KEVIN EDUARDO FUNEZ FUNEZ

DERECHOS DE AUTOR

Alan Fabricio Barahona López

Edras Josué García Vallejo

FACULTAD DE POSTGRADO

SISTEMA DE PRIORIZACIÓN COMERCIAL INTELIGENTE: INTEGRACIÓN DE ANALÍTICA PREDICTIVA PARA LA GESTIÓN DE VENTAS CRUZADAS EN BANCA DE PERSONAS (HONDURAS, 2023–2025)

Alan Fabricio Barahona

Edras Josué García Vallejo

Resumen

La baja tasa de aceptación de productos financieros en los canales físicos representa un desafío operativo relevante para la banca hondureña, particularmente en contextos donde los esfuerzos comerciales no se encuentran priorizados de manera analítica. Ante esta problemática, la presente investigación tiene como objetivo desarrollar un modelo de Machine Learning que permita estimar la probabilidad de aceptación de productos financieros en los canales físicos de una institución bancaria, utilizando variables anticipables disponibles antes del contacto con el cliente.

El estudio adopta un enfoque cuantitativo, aplicado y explicativo, bajo un diseño no

experimental y correlacional-predictivo. Se emplearon datos históricos anonimizados correspondientes al período 2023–2025, los cuales fueron sometidos a procesos de limpieza, análisis exploratorio, pruebas de hipótesis y modelado supervisado, siguiendo el marco metodológico CRISP-DM. Se entrenaron y compararon cuatro modelos de clasificación: regresión logística, árbol de decisión, Random Forest y Gradient Boosting, considerando el severo desbalance de la variable objetivo.

Los resultados evidencian que el modelo Random Forest presentó el mejor desempeño predictivo, alcanzando un AUC de 86.6% en validación y 85.3% en datos reales. Asimismo, el modelo demostró una alta capacidad de priorización, concentrando el 70% de las ventas reales dentro del 20% superior del ranking de propensión. Las variables más relevantes estuvieron asociadas a factores operativos y de interacción, como el resultado del contacto, el tipo de producto y el riesgo del cliente, mientras que las variables sociodemográficas mostraron bajo poder discriminativo.

En conclusión, el modelo propuesto constituye una herramienta analítica robusta y operativamente viable para mejorar la eficiencia comercial en los canales físicos, aportando evidencia empírica sobre el valor del Machine Learning aplicado al contexto bancario hondureño.

Palabras claves: Machine Learning, analítica predictiva, banca, probabilidad de aceptación, canales comerciales, priorización comercial

GRADUATE SCHOOL

SISTEMA DE PRIORIZACIÓN COMERCIAL INTELIGENTE: INTEGRACIÓN DE ANALÍTICA PREDICTIVA PARA LA GESTIÓN DE VENTAS CRUZADAS EN BANCA DE PERSONAS (HONDURAS, 2023–2025)

Alan Fabricio Barahona López

Edras Josué García Vallejo

Abstract

The low acceptance rate of financial products in physical banking channels represents a significant operational challenge, particularly in contexts where commercial efforts are not analytically prioritized. In response to this problem, this study aims to develop a Machine Learning model to estimate the probability of acceptance of financial products in the physical channels of a banking institution in Honduras, using exclusively anticipatory variables available prior to customer contact.

The research adopts a quantitative, applied, and explanatory approach, with a non-experimental and correlational-predictive design. Anonymized historical data from the period

2023–2025 were analyzed through data cleaning, exploratory analysis, hypothesis testing, and supervised modeling, following the CRISP-DM methodological framework. Four classification models were trained and compared: logistic regression, decision tree, Random Forest, and Gradient Boosting, accounting for the severe class imbalance of the target variable.

The results indicate that the Random Forest model achieved the best predictive performance, with an AUC of 86.6% in validation and 85.3% on real data. Additionally, the model demonstrated strong prioritization capabilities, capturing 70% of actual sales within the top 20% of the propensity ranking. The most influential variables were related to operational and interaction factors, such as contact outcome, product type, and customer risk, while traditional sociodemographic variables showed limited discriminatory power.

In conclusion, the proposed model represents a robust and operationally feasible analytical tool for improving commercial efficiency in physical banking channels, providing empirical evidence of the value of Machine Learning applications within the Honduran banking context.

Keywords: Machine Learning, predictive analytics, banking, acceptance probability, commercial channels, commercial prioritization.

DEDICATORIA

Dedicamos este trabajo a nuestras familias, cuyo apoyo incondicional, comprensión y fortaleza nos acompañaron durante todo este proceso académico. Su confianza en nuestras capacidades fue un pilar que nos impulsó a perseverar incluso en los momentos de mayor exigencia.

Extendemos también nuestra dedicatoria a los docentes que nos guiaron a lo largo de la maestría, por su compromiso, disciplina académica y por compartir los conocimientos que hicieron posible la culminación de esta investigación. Agradecemos igualmente a las jefaturas de la institución que, aunque permanecen anónimas por políticas internas, nos brindaron las facilidades y el espacio necesario para avanzar en el desarrollo de este proyecto.

Finalmente, dedicamos este logro a nuestros compañeros de travesía, con quienes compartimos retos, aprendizajes y experiencias que enriquecieron profundamente nuestro camino profesional y personal. Cada intercambio, apoyo y colaboración dejó una huella significativa en la construcción de este resultado.

Este trabajo es fruto del esfuerzo conjunto y del acompañamiento de todas estas personas que marcaron nuestra trayectoria.

AGRADECIMIENTO

A Dios, por brindarme fortaleza, claridad y la oportunidad de culminar este proceso académico.

A nuestras familias, por su apoyo incondicional, paciencia y motivación constante. Su confianza en mis capacidades ha sido el fundamento que me permitió avanzar incluso en los momentos más desafiantes.

A nuestros docentes de la Maestría en Analítica de Negocios, quienes con su guía, exigencia y orientación académica contribuyeron significativamente al desarrollo de esta investigación.

A nuestro asesor, cuyo acompañamiento, retroalimentación y compromiso fueron esenciales para garantizar la calidad y el rigor del presente trabajo.

A la institución que facilitó los datos y permitió aplicar los conocimientos adquiridos en un contexto real. Su colaboración hizo posible que esta investigación tuviera un impacto práctico y relevante.

A nuestros compañeros de clase y colegas, por compartir ideas, experiencias y aprendizajes que enriquecieron este camino.

A todos, nuestros sinceros agradecimientos.

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

CAPÍTULO I. PLANTEAMIENTO DE LA INVESTIGACIÓN

1.1 INTRODUCCIÓN

El sector bancario hondureño enfrenta el desafío de sostener su crecimiento y rentabilidad en un entorno caracterizado por una creciente competencia, mayores exigencias regulatorias y una acelerada transformación digital. A nivel internacional, el sector financiero ha incorporado de manera progresiva herramientas de analítica avanzada y aprendizaje automático para optimizar la toma de decisiones comerciales y operativas. Estudios recientes evidencian que más del 80 % de las instituciones bancarias en economías desarrolladas utilizan modelos basados en datos para actividades como segmentación de clientes, prevención de fraude y optimización de procesos comerciales, generando mejoras significativas en eficiencia y rentabilidad (Hernández & Moreno, 2023a).

Un referente claro de esta evolución es el sistema bancario español, donde la adopción de modelos de Machine Learning se ha consolidado como un componente estructural de la gestión comercial. De acuerdo con la Asociación Española de Banca (2017), la aplicación de analítica predictiva ha permitido mejorar la focalización de campañas comerciales y elevar las tasas de conversión, reduciendo la dispersión de esfuerzos en los canales presenciales. En este contexto, la venta cruzada dejó de depender exclusivamente de la experiencia individual del personal y pasó a apoyarse en modelos cuantitativos capaces de estimar la probabilidad de aceptación de cada cliente (Asociación Española de Banca, 2017).

Sin embargo, esta realidad contrasta de manera significativa con la situación que enfrentan las instituciones financieras en Honduras. A pesar de que el sistema bancario hondureño dispone de grandes volúmenes de datos históricos sobre sus clientes, la adopción de modelos predictivos para apoyar la gestión comercial sigue siendo limitada. En los canales físicos, la atención continúa orientándose principalmente a la resolución de necesidades inmediatas, sin una priorización analítica que permita identificar a los clientes con mayor probabilidad de aceptación de productos financieros. Esta brecha tecnológica y metodológica se traduce en estrategias comerciales homogéneas, baja personalización de las ofertas y una utilización ineficiente de los recursos comerciales disponibles.

A nivel global, la literatura especializada coincide en que los datos se han consolidado como uno de los activos estratégicos más relevantes del sector financiero. El procesamiento de grandes volúmenes de información (Big Data) permite identificar patrones de comportamiento, estimar probabilidades de respuesta y mejorar la efectividad de las estrategias comerciales. En este sentido, se ha documentado que el uso de modelos predictivos puede incrementar la tasa de conversión comercial entre un 10 % y un 30 %, dependiendo del nivel de madurez analítica de la institución (Siarhei Sukhadolski, 2025).

La evidencia empírica sugiere que, en contextos donde no se emplean herramientas analíticas de segmentación y predicción, las tasas de aceptación de ofertas presenciales suelen mantenerse por debajo del 5 % (Asociación Española de Banca, 2017). En el caso hondureño, los datos históricos de la institución analizada reflejan una tasa promedio de aceptación cercana al 2.7 %, lo que pone de manifiesto una problemática operativa concreta que impacta directamente en la rentabilidad del canal físico y en la eficiencia de la fuerza de ventas.

No obstante, trasladar estas prácticas a economías emergentes requiere considerar las particularidades operativas, culturales y tecnológicas de cada contexto. En Honduras, los canales físicos continúan desempeñando un rol central en la relación con el cliente, especialmente para productos financieros de mayor complejidad y para segmentos con menor adopción digital. Esta dependencia del canal presencial, combinada con recursos comerciales limitados, hace imprescindible contar con mecanismos que permitan priorizar de forma objetiva a los clientes con mayor probabilidad de aceptación, maximizando el impacto de cada interacción comercial.

En este escenario, la presente investigación propone el desarrollo de un modelo de Machine Learning orientado a estimar la probabilidad de aceptación de ofertas de productos financieros en los canales físicos de una entidad bancaria hondureña, utilizando variables anticipables disponibles antes del contacto con el cliente. Este enfoque busca cerrar la brecha existente entre las prácticas analíticas consolidadas en sistemas financieros más avanzados y la realidad operativa del contexto hondureño, aportando evidencia empírica y una solución analítica viable para mejorar la eficiencia comercial y la toma de decisiones basada en datos.

1.2 ANTECEDENTES DEL PROBLEMA

En la última década, el incremento de grandes cantidades de datos (Big Data) ha generado

la imperativa necesidad de estrategias robustas para buscar una manera de gestionar, ordenar y usarlos de la manera correcta. El Big Data se ha convertido en un recurso estratégico, permitiendo optimizar procesos comerciales, mejorar la segmentación de clientes y aumentar la efectividad de la venta cruzada mediante la identificación de patrones predictivos (Vázquez & González, 2019).

1.2.1 APLICACIÓN DE MACHINE LEARNING EN LA BANCA EUROPEA

Europa es de las regiones donde está a la vanguardia de la inteligencia artificial y del uso de ML principalmente en el sector bancario donde ha pasado de una adopción del 60 por ciento a un 86 por ciento, siendo sus principales usos el perfilamiento de clientes, prevención de lavado de dinero, detección de fraude y asistencias al cliente (funcas, 2025).

- Implementación de Redes Neuronales Avanzados y Precisión Predictiva (España):

  Campos Sánchez hace un realce de del potencial de las Redes Neuronales Recurrentes (RNN) y de Memoria a Largo Plazo (LSTM) para la aplicar predicción de ventas. Este estudio demuestra que estas arquitecturas superan métodos tradicionales como los econométricos y estadísticos en términos de predicción, siendo esencial para la planificación eficaz y la asignación precisa de asignación de recursos (Sánchez, 2025).

- Implementación de Modelos Aprendizaje Automático en Concesión de Créditos (España):

  Andrés Alonso-Robisco y José Manuel Carbo centra su estudio en el impacto que ha tenido el ML en la banca, en el estudio demuestra la capacidad predictiva del uso de XGBoost, sobrepasando a la regresión logística. Con esta mejora de predicción de impago a alcanzado una mejora del 17 por ciento sobre la cartera de crédito al consumo de un banco español (Alonso-Robisco & Carbo, 2023).

1.2.2 CASOS DE ESTUDIOS RELEVANTES EN LATINOAMÉRICA

Latinoamérica a dados pasos importantes en la investigación enfocada en el ML principalmente en el uso de efectividad comercial, en el campo financiero y en dar una mejor experiencia al cliente.

- Clustering para Recomendación de Productos (Ecuador):

  Chapa Zumba (2023), desarrolló un caso significativo en la región, donde desarrolla un modelo basado en clustering K-means para la recomendación de productos financieros en

su investigación el cual, permite identificar clientes potenciales con sus características homogéneas mejorando la efectividad con estrategias mejor enfocadas(Cristina Victoria Chapa Zumba, 2023).

- Machine Learning y La Optimización de Estrategia Comercial:

  La región en general de la región se ha beneficiado de la aplicación de ML en temas de reducción de costos operativos y aumentar rentabilidad con identificación de clientes potenciales con la predicción de comportamiento (Contreras, 2024).

1.2.3 LA CONVERGENCIA REGIONAL Y LA BRECHA DE CONOCIMIENTO EN HONDURAS

La evidencia de aplicación de ML en el sector bancario de Centroamérica es limitada, lo que indica un área de investigación pobre.

- Predicción de Demanda y optimización de Recursos (Guatemala):

  Un estudio que se realizó en Guatemala analizo la efectividad de modelos ML en predicción de demanda de productos. Llegando a la conclusión que los modelos basados en aprendizaje automático llevan la ventaja frente a los métodos estadísticos tradicionales, no solamente en la precisión de las predicciones, sino que en la optimización de recursos al tener campañas orientadas a grupos focales.

- Brecha Conocimiento en Honduras:

  A pesar del potencial que demostrado por el ML para la optimización de proceso comercial el éxito en las ventas cruzadas, en Honduras existe un vacío de conocimiento considerable. Todo esto presenta la necesidad de la presente investigación, la cual busca aportar soluciones concretas en los desafíos de este tema.

1.3 DEFINICIÓN DEL PROBLEMA

Aunque las instituciones financieras hondureñas cuentan con datos históricas abundantes sobre clientes, estos no se utilizan de manera efectivas para orientar las estrategias comerciales. La ausencia de modelos predictivos ocasiona:

- Esfuerzos comerciales no priorizados.

- Tasa de aceptación menor al potencial real.
- Ausencia de personalización en las ofertas.
- Subutilización del valor informativo de los datos.

Problema central:

¿Cómo predecir la probabilidad de aceptación de productos financieros por parte de los clientes atendidos en los canales físicos, utilizando los datos históricos de la entidad bancaria?

1.4 PREGUNTAS DE INVESTIGACIÓN

1.4.1 PREGUNTA GENERAL

- ¿En los clientes de banca de personas atendidos a través de canales físicos (P), la aplicación de modelos de Machine Learning integrados en un sistema de priorización comercial (I), comparada entre diferentes algoritmos predictivos (C), permite estimar la probabilidad de aceptación de productos financieros y, a partir de ello, mejorar la eficiencia del proceso de ventas (O)?

1.4.2 PREGUNTAS ESPECIFICAS

- ¿Qué variables del cliente (P), identificadas mediante análisis exploratorio y modelos predictivos (I), comparadas por su peso estadístico (C), influye significativamente en la probabilidad de aceptación del producto financiero (O)?.
- ¿Qué modelo de Machine Learning (I), comparados con otros algoritmos supervisados (C), ofrece el mejor desempeño predictivo (O) para estimar la aceptación en clientes de banca de personas a través de canales físicos (P)?
- ¿En qué medida la aplicación del modelo predictivo seleccionado (I), comparada con el proceso comercial tradicional (C), permite segmentar y priorizar a los clientes de banca de personas atendidos a través de canales físicos (P) de acuerdo con su probabilidad estimada de aceptación, mejorando la eficiencia comercial y la tasa de aceptación de productos financieros (O)?

1.5 OBJETIVOS DEL PROYECTO

Objetivo General:

Diseñar y validar (S) un sistema de priorización comercial basado en Machine Learning para estimar la probabilidad de aceptación de productos financieros en clientes de banca de personas atendidos a través de canales físicos, alcanzando (M) un desempeño ≥ 0.70 en la fase de validación, utilizando (A) datos históricos del banco, con el fin de mejorar (R) la eficiencia del proceso de ventas y la asignación de recursos comerciales, durante (T) el periodo 2023–2025.

Objetivos Específicos:

- Identificar (S) las variables demográficas, transaccionales y de vinculación que muestran asociación significativa con la aceptación de productos financieros, determinando (M) su significancia estadística con $p < 0.05$ mediante pruebas Chi-cuadrado, t-test y Mann-Whitney, utilizando (A) el dataset histórico proporcionado por la institución, para seleccionar (R) los predictores relevantes para el modelado, durante (T) la fase exploratoria e inferencial del estudio (2023–2025).
- Entrenar y comparar (S) al menos cuatro modelos supervisados de Machine Learning (Regresión Logística, Árbol de Decisión, Random Forest y Gradient Boosting), evaluando (M) su desempeño mediante métricas AUC, F1-Score y Lift@k, hasta seleccionar uno con AUC ≥ 0.70, empleando (A) validación cruzada y calibración probabilística en Python, para garantizar (R) la elección del modelo predictivo óptimo, durante (T) el periodo de validación 2023–2025.
- Aplicar (S) el modelo seleccionado para estimar la probabilidad individual de aceptación y evaluar su desempeño operativo, generando (M) un ranking de priorización basado en segmentos de propensión (top 5%, 10%, 20%) y midiendo métricas como Liftk, utilizando (A) datos de evaluación de tres meses posteriores, para optimizar (R) la asignación de esfuerzos comerciales en los canales físicos, durante (T) la fase de implementación y prueba del estudio (2025).

1.6 JUSTIFICACIÓN

Desde una perspectiva práctica, la presente investigación resulta indispensable para abordar una problemática concreta del sector bancario hondureño: la baja efectividad comercial en los canales físicos derivada de la ausencia de modelos analíticos que permitan priorizar clientes y optimizar los esfuerzos de venta. Diversos estudios en el sector financiero evidencian que, cuando

las estrategias comerciales no se apoyan en modelos predictivos, las tasas de aceptación suelen mantenerse por debajo del 5 %, generando una asignación ineficiente de recursos y un bajo retorno de inversión (Asociación Española de Banca, 2017). En el caso de la institución analizada, los datos históricos muestran una tasa promedio de aceptación cercana al 2.7 %, lo que confirma la existencia de un riesgo operativo y financiero asociado al enfoque tradicional de atención y oferta de productos.

Desde una perspectiva teórica, esta investigación aborda una brecha de conocimiento claramente identificada. Si bien la literatura internacional ha documentado ampliamente el uso de modelos de Machine Learning para la segmentación y predicción del comportamiento del cliente en mercados desarrollados, la evidencia empírica aplicada a contextos bancarios de economías emergentes (Useche & Peter Stig Reina Colorado, 2021; Vázquez & González, 2019), y particularmente al canal físico, sigue siendo limitada. Las investigaciones existentes tienden a concentrarse en canales digitales o en modelos crediticios, dejando un vacío en el análisis de la aceptación de productos financieros en entornos presenciales y con restricciones operativas específicas como las que caracterizan al sistema bancario hondureño.

Finalmente, desde una perspectiva empresarial y social, la eficiencia del sector financiero constituye un pilar fundamental para la estabilidad económica y la confianza del sistema. La adopción de enfoques analíticos basados en datos ha demostrado generar mejoras sustanciales en la toma de decisiones comerciales y en la eficiencia operativa. En este sentido, se estima que las instituciones financieras que incorporan analítica avanzada y modelos predictivos logran incrementos en la tasa de conversión comercial que oscilan entre el 10 % y el 30 %, dependiendo de su nivel de madurez analítica (Siarhei Sukhadolski, 2025). Estos resultados evidencian el impacto potencial que tiene la analítica predictiva no solo en la rentabilidad institucional, sino también en la calidad de la experiencia del cliente y la sostenibilidad del sistema financiero.

En este contexto, el problema abordado por esta investigación no es superficial ni aislado. Por el contrario, responde a *causas estructurales complejas e interrelacionadas* que dificultan la adopción de enfoques analíticos avanzados en los canales físicos. Antes de formular una solución basada en Machine Learning, resulta imprescindible identificar y comprender las causas raíz que explican esta baja efectividad comercial, lo que justifica el desarrollo de un análisis causal que permita desagregar el problema central y sentar las bases para una propuesta técnicamente viable

y sostenible.

Si fuera Ishikawa

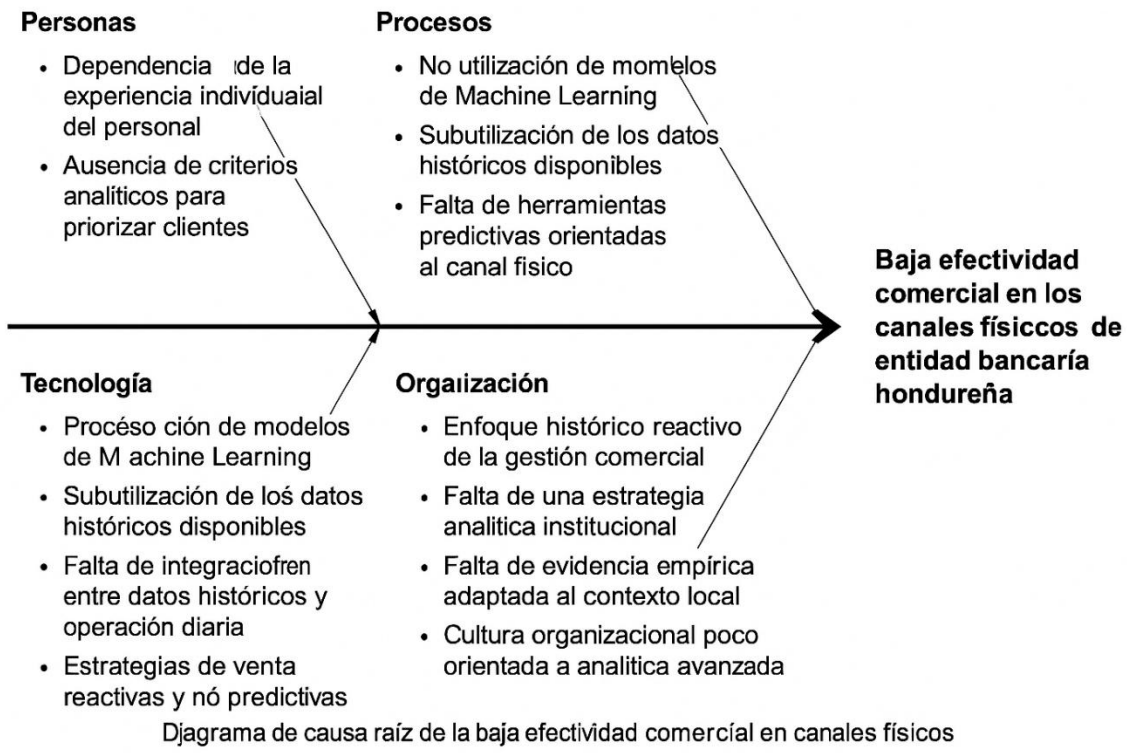

Ilustración 1 Problemática actual del proceso de ventas en canales físicos y digitales

Fuente: Elaboración propia

Si fuera mapa conceptual:

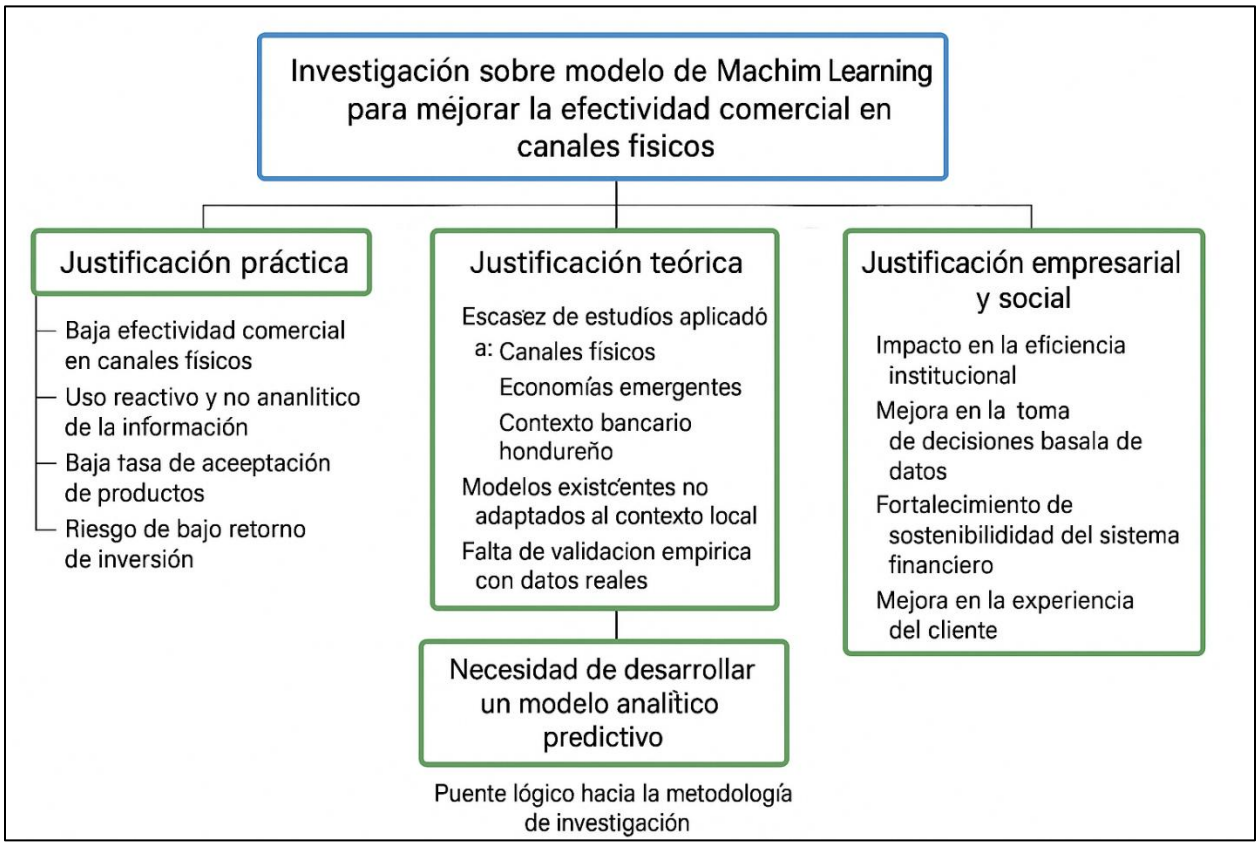

Ilustración 2 Mapa conceptual
Fuente: Elaboración propia

Si fuera en 5 porque:

1. ¿Por qué existe una baja efectividad comercial en los canales físicos?

Porque las ofertas de productos financieros se realizan de manera homogénea y reactiva, sin un criterio de priorización basado en la probabilidad de aceptación del cliente, lo que provoca una dispersión significativa de los esfuerzos comerciales y una baja tasa de conversión.

2. ¿Por qué no se prioriza a los clientes según su probabilidad de aceptación?
   Porque el proceso comercial tradicional se apoya principalmente en la experiencia del personal de atención y en reglas generales, sin utilizar modelos predictivos que integren de forma sistemática las variables demográficas, transaccionales y de comportamiento disponibles en los datos históricos de la institución.
3. ¿Por qué no se utilizan modelos predictivos basados en los datos históricos disponibles?
   Porque la organización carece de un enfoque metodológico validado que integre análisis

exploratorio, inferencial y modelado de Machine Learning orientado específicamente al canal físico, lo que limita la transformación de los datos en información accionable para la toma de decisiones comerciales.

4. ¿Por qué no existe un enfoque metodológico validado para este contexto?
   Porque la evidencia empírica y las guías metodológicas disponibles se concentran mayoritariamente en mercados desarrollados o en canales digitales, y no consideran las particularidades operativas, culturales y de recursos que caracterizan al sistema bancario hondureño y a la atención presencial al cliente.
5. ¿Por qué persiste esta brecha de conocimiento aplicada al canal físico?

Causa raíz: Porque históricamente la gestión comercial en los canales físicos ha sido abordada desde un enfoque descriptivo y reactivo, centrado en indicadores agregados y en la experiencia individual, en lugar de un enfoque analítico y predictivo basado en evidencia cuantitativa. Esta orientación ha limitado el desarrollo y la validación de modelos de Machine Learning adaptados al contexto local, perpetuando la baja efectividad comercial y la subutilización del valor informativo de los datos.

CAPÍTULO II. MARCO TEÓRICO

Este capítulo desarrolla los fundamentos teóricos, conceptuales y metodológicos que sustentan la presente investigación, cuyo propósito es aplicar modelos de aprendizaje automático (Machine Learning) para estimar la probabilidad de aceptación de productos financieros en canales físicos de un banco hondureño. El análisis se centra en comprender las condiciones del entorno, las bases conceptuales y los modelos teóricos que explican la adopción de la analítica avanzada en el sector bancario, así como en identificar las metodologías y herramientas que respaldan su aplicación.

En primer lugar, se presenta el análisis del macroentorno mediante el modelo PESTEL, con el propósito de examinar las fuerzas políticas, económicas, sociales, tecnológicas, ecológicas y legales que inciden en la evolución de la temática presentada. Posteriormente, se aborda el microentorno, que incluye la situación actual del uso del Machine Learning Centroamérica y la banca nacional y el análisis competitivo del sector con base en las Cinco Fuerzas de Porter.

Posteriormente, se desarrolla la conceptualización de los principales términos vinculados con la investigación, seguida de la exposición de las teorías de sustento que explican el comportamiento del consumidor financiero y la adopción de tecnologías predictivas. Seguidamente, se analizan los enfoques metodológicos empleados en estudios previos, los antecedentes de metodologías, los diseños y enfoques más utilizados, y se presenta un análisis crítico que justifica la metodología adoptada en esta tesis.

Finalmente, se describen las herramientas tecnológicas que facilitan la implementación del modelo propuesto y el marco legal que regula la actividad bancaria, la protección de los datos personales y la transformación digital en el sistema financiero hondureño. En conjunto, este capítulo proporciona el soporte teórico y metodológico necesario para comprender y sustentar la investigación que se desarrolla en los capítulos posteriores.

2.1 ANÁLISIS DEL MACROENTORNO.

El sector bancario global atraviesa una transformación estructural impulsada por factores políticos, económicos, sociales, tecnológicos y legales que redefinen la manera en que las instituciones gestionan riesgos, atienden a sus clientes y adoptan herramientas analíticas avanzadas

(incluyendo modelos predictivos basados en machine learning).

Para comprender el contexto en el que se sitúa esta investigación, se examinan las dinámicas del macroentorno en tres países con realidades financieras diversas: España, India y Sudáfrica, lo que permite identificar tendencias globales que influyen en el diseño y adopción de soluciones basadas en datos en la industria bancaria.

2.1.1 ESPAÑA: REGULACIÓN ESTRICTA, MADUREZ DIGITAL Y PRESIÓN POR EFICIENCIA

España opera dentro de uno de los marcos regulatorios más estrictos del sistema financiero internacional, al formar parte de la Unión Europea y del Eurosistema. El Reglamento General de Protección de Datos (GDPR) establece exigencias explícitas sobre el uso, tratamiento y automatización de datos personales, obligando a las instituciones financieras a garantizar trazabilidad, explicabilidad y control sobre los modelos algorítmicos utilizados en la toma de decisiones (Regulation (EU) 2016/679 of the European Parliament, 2016). Este entorno legal condiciona directamente la forma en que los bancos pueden implementar modelos predictivos, priorizando soluciones robustas y gobernadas.

Desde la perspectiva económica, el sistema bancario español ha experimentado una mejora en rentabilidad reciente, aunque en un entorno de fuerte competencia y presión estructural por eficiencia. En España, el ROE del sistema bancario se situó en 10.1 % en 2023, impulsado principalmente por el aumento de los tipos de interés, mientras que el ratio de eficiencia operativa se mantuvo cercano al 45 %, reflejando la necesidad de seguir optimizando costos y procesos comerciales (Banco de España, 2024). Asimismo, el volumen de crédito al sector privado superó los 1.3 billones de euros, lo que evidencia un mercado amplio, pero altamente disputado, donde pequeñas mejoras en eficiencia tienen impactos significativos.

En el ámbito social, España presenta elevados niveles de adopción de servicios financieros digitales, aunque con importantes diferencias generacionales. Más más del 70 % de los clientes bancarios utiliza canales digitales como medio principal de interacción, mientras que los segmentos de mayor edad continúan demandando atención presencial (Deloitte, 2024). Este comportamiento híbrido obliga a los bancos a gestionar de forma estratégica sus canales físicos, priorizando recursos comerciales limitados hacia clientes con mayor probabilidad de respuesta.

Desde el punto de vista tecnológico, España se posiciona entre los sistemas bancarios más avanzados de Europa en el uso de analítica avanzada e inteligencia artificial. Más del 80 % de las entidades bancarias españolas ha incorporado soluciones basadas en analítica de datos o inteligencia artificial en procesos como segmentación de clientes, scoring crediticio y prevención de fraude (Funcas, 2025). No obstante, estas soluciones deben operar bajo marcos regulatorios exigentes, lo que refuerza la necesidad de modelos predictivos explicables y monitoreables.

En conjunto, el caso español demuestra que incluso en sistemas financieros maduros, altamente digitalizados y regulados, la presión por eficiencia y optimización comercial persiste. En este contexto, los modelos predictivos se consolidan como una herramienta clave para priorizar clientes, asignar esfuerzos comerciales y mejorar la rentabilidad, siempre dentro de esquemas estrictos de gobernanza y cumplimiento normativo.

2.1.2 INDIA: INNOVACIÓN DIGITAL A GRAN ESCALA, CRECIMIENTO CUANTIFICABLE Y NECESIDAD DE PRIORIZACIÓN OPERATIVA

India representa uno de los casos más relevantes de transformación digital financiera a escala masiva. Desde el ámbito político-institucional, el país ha impulsado una estrategia de infraestructura digital pública a través de iniciativas como India Stack, que integra identidad digital (Aadhaar), pagos electrónicos y verificación en línea, sentando las bases para la adopción extensiva de servicios financieros digitales (OECD, 2022). Este marco ha permitido a las instituciones financieras operar con grandes volúmenes de datos, condición indispensable para el uso de modelos predictivos a gran escala.

Desde la perspectiva económica, el sistema financiero indio ha experimentado una expansión sostenida del crédito. El crédito bancario al sector privado registró un crecimiento interanual superior al 15 % durante 2023, impulsado principalmente por préstamos al consumo, pequeñas empresas y servicios. Este crecimiento acelerado incrementa la complejidad operativa de las instituciones financieras, que deben evaluar y priorizar millones de clientes potenciales con recursos limitados (Reserve Bank of India, 2024).

Asimismo, India se ha consolidado como uno de los mayores mercados de crédito digital del mundo, con más de 300 millones de usuarios activos de pagos digitales, lo que refleja un entorno altamente competitivo donde la velocidad y la precisión en la toma de decisiones comerciales resultan críticas (International Finance Corporation, 2024).

En el ámbito social, India presenta una marcada heterogeneidad. Si bien las zonas urbanas muestran altos niveles de adopción digital, amplios segmentos rurales aún enfrentan limitaciones de acceso a infraestructura tecnológica. Aproximadamente el 67 % de la población adulta utiliza servicios financieros formales, cifra que ha mejorado significativamente en la última década, pero que aún evidencia brechas relevantes entre regiones y niveles socioeconómicos (Banco Mundial, 2023). Esta diversidad obliga a las instituciones financieras a combinar canales digitales y presenciales, haciendo indispensable la focalización de esfuerzos comerciales.

Desde el punto de vista tecnológico, India se posiciona como uno de los países con mayor adopción de analítica avanzada en el sector financiero. El ecosistema fintech indio ha crecido de forma exponencial, especialmente en pagos, crédito alternativo y soluciones basadas en inteligencia artificial (Finnovista et al., 2024). No obstante, la magnitud del mercado implica que los bancos no pueden atender a todos los clientes de manera homogénea, lo que refuerza la necesidad de modelos predictivos para segmentar, priorizar y gestionar riesgos de forma eficiente.

En el ámbito legal, aunque India no cuenta con un marco de protección de datos tan estricto como el GDPR europeo, el país ha avanzado hacia regulaciones más claras en materia de privacidad y uso de datos, orientadas a fortalecer la confianza en los servicios digitales (OECD, 2022). Este entorno normativo en evolución exige a las instituciones financieras un uso responsable y transparente de los modelos analíticos.

En conjunto, el caso de India demuestra que, en contextos de alta demanda, crecimiento acelerado y recursos operativos finitos, los modelos predictivos no solo mejoran la eficiencia, sino que se convierten en una herramienta esencial para priorizar clientes, optimizar la asignación de esfuerzos comerciales y sostener la escalabilidad del sistema financiero.

2.1.3 SUDÁFRICA: SUPERVISIÓN PRUDENCIAL ROBUSTA, BRECHAS CUANTIFICABLES DE INCLUSIÓN Y PRESIÓN POR EFICIENCIA OPERATIVA

Sudáfrica cuenta con uno de los marcos de supervisión financiera más sólidos del continente africano, liderado por la South African Reserve Bank (SARB). Tras los choques macroeconómicos recientes, la autoridad ha reforzado la supervisión prudencial y la gestión de riesgos sistémicos, estableciendo un entorno regulatorio estable para la adopción de tecnologías analíticas en el sector financiero (South African Reserve Bank, 2023). Este contexto institucional favorece el uso de modelos predictivos, siempre que se integren dentro de esquemas formales de

gobernanza y control.

Desde la perspectiva económica, el sistema bancario sudafricano opera en un entorno de crecimiento moderado y presión estructural por eficiencia. El ROE promedio del sector bancario se situó alrededor del 15 % en 2023, mientras que los costos operativos continúan siendo un desafío relevante, especialmente en un contexto de bajo crecimiento económico y alta competencia (South African Reserve Bank, 2023). El crecimiento económico del país se mantuvo por debajo del 2 %, lo que limita la expansión orgánica del negocio bancario y obliga a optimizar el uso de recursos existentes (Banco Mundial, 2024).

En el ámbito social, Sudáfrica presenta una de las brechas de inclusión financiera y digital más marcadas entre economías de ingreso medio. Se estima que aproximadamente el 71 % de los adultos posee una cuenta en una institución financiera, cifra relativamente alta para la región, pero con fuertes disparidades entre zonas urbanas y rurales (Banco Mundial, 2024). Complementariamente, el informe (GSMA, 2023) señala que, aunque la penetración de telefonía móvil supera el 90 %, el acceso efectivo a servicios digitales avanzados sigue siendo desigual, especialmente en comunidades rurales y de bajos ingresos.

Este escenario obliga a las instituciones financieras a operar en entornos donde los recursos comerciales (personal, sucursales y tiempo de atención) son limitados, mientras que la demanda potencial es elevada y heterogénea. En consecuencia, la priorización de clientes y la focalización de esfuerzos comerciales se vuelven decisiones estratégicas críticas.

Desde el punto de vista tecnológico, Sudáfrica ha avanzado en la adopción de analítica avanzada e inteligencia artificial, particularmente en procesos de scoring crediticio, detección de fraude y segmentación de clientes. No obstante, la heterogeneidad del mercado y las limitaciones de infraestructura digital impiden una cobertura homogénea, lo que refuerza la necesidad de modelos predictivos capaces de identificar segmentos con mayor probabilidad de respuesta o menor riesgo (South African Reserve Bank, 2023).

En el ámbito legal, el país cuenta con un marco robusto de protección de datos a través de la Protection of Personal Information Act (POPIA, 2021), que regula estrictamente el uso de información personal y exige medidas de transparencia y seguridad en el procesamiento de datos. Este marco se alinea con las recomendaciones internacionales en materia de gobernanza algorítmica y uso responsable de inteligencia artificial (OECD, 2022).

En conjunto, el caso sudafricano evidencia que, en contextos de brechas de inclusión, crecimiento económico limitado y recursos operativos restringidos, los modelos predictivos se convierten en una herramienta clave para optimizar la eficiencia, priorizar clientes y maximizar el impacto de cada interacción comercial, reforzando su relevancia para sistemas financieros con características similares.

2.1.4 SÍNTESIS DEL MACROENTORNO Y SU RELACIÓN CON LA INVESTIGACIÓN

El análisis del macroentorno a partir de los casos de España, India y Sudáfrica permite identificar patrones estructurales comunes que trascienden las diferencias en nivel de desarrollo económico, grado de digitalización y madurez institucional de los sistemas financieros. A pesar de operar en contextos regulatorios, sociales y tecnológicos distintos, los tres países enfrentan presiones convergentes que condicionan la forma en que las instituciones bancarias gestionan sus recursos y toman decisiones comerciales.

En primer lugar, los tres casos evidencian un entorno regulatorio cada vez más exigente, que impone mayores estándares de gobernanza, trazabilidad y control sobre el uso de datos y modelos algorítmicos. Desde marcos altamente estrictos como el GDPR en España, hasta regulaciones en evolución como las observadas en India y Sudáfrica, se refuerza la necesidad de que los modelos predictivos sean no solo precisos, sino también explicables y auditables. Este contexto limita el uso discrecional de la analítica y exige soluciones robustas, alineadas con la supervisión financiera.

En segundo lugar, el análisis económico muestra una presión estructural por eficiencia operativa. En España, la competencia intensa y los márgenes ajustados obligan a optimizar la asignación de esfuerzos comerciales; en India, el crecimiento acelerado del crédito y la magnitud del mercado generan una demanda que supera la capacidad operativa de atención homogénea; mientras que en Sudáfrica, el crecimiento moderado y las restricciones estructurales intensifican la necesidad de maximizar el impacto de cada interacción bancaria. En todos los casos, la eficiencia deja de ser un objetivo táctico y se convierte en un requisito estratégico.

Desde la dimensión social, los tres países comparten la coexistencia de clientes altamente digitalizados y segmentos que aún dependen de canales presenciales, ya sea por razones generacionales, geográficas o de acceso a infraestructura tecnológica. Esta heterogeneidad del cliente bancario impide estrategias uniformes y refuerza la necesidad de segmentación avanzada y

priorización en la atención comercial, especialmente cuando los recursos físicos y humanos son limitados.

En el ámbito tecnológico, los casos analizados confirman que la disponibilidad de grandes volúmenes de datos y la adopción creciente de analítica avanzada e inteligencia artificial no eliminan las restricciones operativas, sino que las hacen más visibles. La capacidad tecnológica permite procesar información a gran escala, pero no sustituye la necesidad de decidir a quién atender, cuándo y con qué producto, lo que sitúa a los modelos predictivos como un componente central en la toma de decisiones comerciales basadas en evidencia.

En conjunto, el macroentorno internacional demuestra que, independientemente del contexto económico o nivel de desarrollo, los sistemas bancarios enfrentan una restricción común: la necesidad de asignar recursos comerciales finitos de manera eficiente y estratégica. En este escenario, los modelos predictivos basados en Machine Learning emergen no solo como una tendencia tecnológica, sino como una herramienta esencial para priorizar clientes, optimizar esfuerzos comerciales y mejorar la efectividad de los canales de atención.

Esta síntesis refuerza la pertinencia de la presente investigación, al evidenciar que los desafíos observados a nivel internacional son coherentes con la realidad del sistema bancario hondureño, particularmente en los canales físicos. De este modo, el análisis del macroentorno proporciona el sustento conceptual y empírico necesario para justificar el desarrollo de un modelo de Machine Learning orientado a estimar la probabilidad de aceptación de productos financieros, como mecanismo para fortalecer la eficiencia y competitividad del sector bancario.

2.2 ANÁLISIS DEL MICROENTORNO

El análisis del microentorno permite comprender las presiones competitivas inmediatas que condicionan el desempeño del sector bancario, la adopción tecnológica y la capacidad de innovar en la oferta de productos financieros. Para este estudio, el microentorno se examina mediante el modelo de las Cinco Fuerzas de Porter, herramienta que describe la estructura competitiva de una industria a partir de la rivalidad existente, el poder de negociación de clientes y proveedores, la amenaza de nuevos entrantes y la presencia de productos sustitutos (Porter, 2008).

En el contexto centroamericano, los sistemas financieros comparten características estructurales que justifican su comparación: altos niveles de concentración bancaria, procesos de

digitalización en expansión, presencia creciente de fintech y marcos regulatorios en evolución (Banco Interamericano de Desarrollo, 2025; Comisión Económica para América Latina y el Caribe, 2023). Analizar el entorno competitivo de países como Guatemala, El Salvador, Costa Rica y Panamá permite establecer una referencia regional para comprender mejor la posición estratégica del sistema bancario hondureño y los factores que influyen en la adopción de modelos analíticos basados en Machine Learning.

2.2.1 GUATEMALA

Rivalidad entre competidores existentes

El sistema bancario guatemalteco presenta una estructura moderadamente concentrada, donde cinco instituciones dominan más del 80 % de los activos totales (Súper Intendencia de Bancos de Guatemala, 2024). Este nivel de concentración genera una competencia fuerte por los segmentos de mayor valor, especialmente en crédito y servicios corporativos.

Además, el crédito al sector privado registró un crecimiento interanual del 7.5 %, impulsado por los sectores comercio, industria y consumo. Este dinamismo refleja un mercado altamente activo que incentiva a las instituciones a competir por captación, originación y retención de clientes (Banco de Guatemala, 2023). La combinación de crecimiento crediticio y concentración bancaria intensifica la rivalidad, obligando a los bancos a innovar en eficiencia operativa, digitalización y experiencia del usuario para mantener participación de mercado (Comisión Económica para América Latina y el Caribe, 2023).

Por ello, la rivalidad competitiva en Guatemala puede considerarse media-alta, impulsada tanto por el comportamiento del mercado como por la presión tecnológica.

Poder de negociación de los clientes

La creciente penetración de servicios financieros digitales ha fortalecido el poder de los usuarios, quienes comparan tasas, comisiones y beneficios con mayor facilidad (Banco Mundial, 2023). Asimismo, el incremento de depósitos del público que crecieron 9.2 % durante 2023, lo cual evidencia una mayor disponibilidad de opciones de ahorro y productos financieros (Banco de Guatemala, 2023).

Este aumento en la oferta y diversidad de servicios mejora la capacidad de elección de los clientes y reduce los costos de cambio entre instituciones, lo cual incrementa su poder de

negociación. No obstante, persiste un segmento significativo de población no bancarizada, lo que modera parcialmente la fuerza de esta presión.

En conjunto, el poder de los clientes en Guatemala se clasifica como medio.

Poder de negociación de los proveedores

Los bancos guatemaltecos dependen en gran medida de proveedores internacionales de infraestructura tecnológica, software core bancario y servicios de ciberseguridad (Comisión Económica para América Latina y el Caribe, 2023). La falta de desarrolladores locales especializados incrementa costos de sustitución y dependencia. Se considera una fuerza media-alta.

Amenaza de nuevos competidores

Aunque los requisitos regulatorios limitan el ingreso de nuevos bancos, la aparición de fintech en pagos, crédito alternativo y remesas aumenta la presión competitiva. Guatemala forma parte del ecosistema fintech regional más grande de Centroamérica (Finnovista et al., 2024). La amenaza se clasifica como media.

Amenaza de productos sustitutos

Las billeteras digitales y cooperativas de ahorro se han posicionado como alternativas relevantes para servicios financieros básicos. El uso de pagos digitales creció aceleradamente desde 2020. La amenaza se considera media-alta (Banco Mundial, 2023).

2.2.2 EL SALVADOR

Rivalidad entre competidores existentes

El sistema financiero salvadoreño mantiene una estructura concentrada en pocos bancos que dominan la mayor parte de los activos del país (Superintendencia del Sistema Financiero, 2023). La dolarización ha reducido la dispersión cambiaria y mejorado la comparabilidad entre instituciones, lo que intensifica la competencia. La actividad transaccional también confirma este dinamismo: en 2023, el Sistema de Liquidación Bruta en Tiempo Real (LBTR) registró un aumento de US$8,724.14 millones, equivalente a 8.84 % respecto a 2022 (BCR, 2023).

Asimismo, el sistema regional de pagos Transfer365 CA-RD mostró un crecimiento acelerado, pasando de 219 operaciones antes de su lanzamiento a 9,611 transacciones en el

segundo semestre de 2023, lo que representa un incremento de 4,289 %, mientras que los montos transados crecieron de US$18.39 millones a US$83.10 millones (+352 %) (BCR, 2023). Estos datos reflejan un sistema financiero dinámico, con creciente digitalización y competencia por servicios de alto valor, lo que ubica la rivalidad en un nivel alto.

Poder de negociación de los clientes

El poder de los clientes se ha incrementado debido al uso extendido de canales digitales y medios de pago inmediatos. En 2023, el servicio Transfer365 concentró el 67.98 % de todas las operaciones interbancarias inmediatas del país, evidenciando su preferencia por servicios rápidos, baratos y comparables entre bancos (Banco Central de Reserva, 2023).

Además, durante fines de semana y días festivos se realizaron 4,169,694 operaciones, por un total de US$984.96 millones, lo que confirma que los usuarios dependen cada vez más de medios digitales que reducen costos de cambio (Banco Central de Reserva, 2023). El crecimiento de Transfer365 Móvil también es destacable, con 120,276 operaciones por US$20.45 millones, tendencia que refuerza la autonomía del consumidor y su capacidad para comparar servicios. Por ello, el poder de los clientes en El Salvador se clasifica como alto.

Poder de negociación de los proveedores

El sistema financiero salvadoreño depende de proveedores tecnológicos especializados para operar infraestructuras críticas como el LBTR, la compensación de cheques y los sistemas de transferencias inmediatas. La Cámara de Compensación procesó en 2023 un total de US$17,906.82 millones, lo que evidencia la escala operativa que requiere soporte técnico avanzado (Banco Central de Reserva, 2023).

Durante este periodo también se emitieron nuevas normas para proveedores de servicios digitales, activos virtuales y dinero electrónico, elevando los requisitos técnicos y fortaleciendo la dependencia de proveedores con alta capacidad de cumplimiento (Banco Central de Reserva, 2023). Debido a la limitada oferta local y a los altos costos de sustitución, el poder de negociación de los proveedores se clasifica como medio-alto.

Amenaza de nuevos competidores

Las barreras de entrada para instituciones bancarias tradicionales siguen siendo elevadas; sin embargo, el crecimiento del ecosistema fintech ha aumentado la amenaza de nuevos competidores. En 2023 se presentó la Estrategia Nacional Fintech, acompañada de nuevas regulaciones para servicios digitales y activos virtuales, lo que habilita el ingreso de empresas tecnológicas al mercado financiero (Banco Central de Reserva, 2023).

La rápida adopción de medios de pago (incluyendo las más de 9,611 operaciones de Transfer365 CA-RD en seis meses) facilita que actores no bancarios compitan por transacciones y servicios que antes eran exclusivos de la banca (Banco Central de Reserva, 2023). Con estas condiciones, la amenaza de entrantes se considera media-alta.

Amenaza de productos sustitutos

Los principales sustitutos de servicios bancarios provienen de billeteras digitales, cooperativas, proveedores de dinero electrónico y plataformas de pagos internos y externos. La población ha migrado progresivamente hacia canales digitales debido al ahorro en costos: solo en 2023, Transfer365 generó un beneficio estimado de US$39.94 millones en comisiones no pagadas por los usuarios (Banco Central de Reserva, 2023). Además, aunque el sistema de compensación de cheques procesó más de 3 millones de documentos, su uso continúa disminuyendo mientras aumentan las soluciones electrónicas más eficientes (Banco Central de Reserva, 2023). Por ello, la amenaza de productos sustitutos se clasifica como media-alta.

2.2.3 COSTA RICA

Rivalidad entre competidores existentes

El sistema financiero costarricense presenta una combinación de bancos estatales dominantes y banca privada competitiva, configurando un mercado de rivalidad elevada. Los bancos estatales concentran alrededor del 55 % de los activos, mientras que los bancos privados representan cerca del 40 %, una estructura que impulsa la competencia en productos de consumo, vivienda y comercio (Superintendencia General de Entidades Financieras, 2024).

El crédito al sector privado mostró una recuperación notable: el Banco Central de Costa Rica reportó un crecimiento real del crédito del 10.4 % en 2023, impulsado principalmente por préstamos en moneda nacional (Banco Central de Costa Rica, 2023). Además, el Informe de

Política Monetaria señala que la inflación se mantuvo dentro del rango meta y que las tasas de interés se redujeron gradualmente durante 2024–2025, facilitando la expansión crediticia (Banco Central de Costa Rica, 2025).

Por estas razones, la rivalidad entre competidores se clasifica como media-alta.

Poder de negociación de los clientes

Costa Rica cuenta con uno de los ecosistemas de pagos electrónicos más avanzados de la región. En 2023, SINPE Móvil procesó más de 254 millones de transacciones, con un crecimiento anual superior al 40 %, consolidándose como el principal canal de pagos minoristas del país (Banco Central de Costa Rica, 2023).

La OECD (2022), destaca que el marco de protección al consumidor financiero y la transparencia de información fortalecen la capacidad del cliente para comparar condiciones y cambiar de institución (OECD, 2022). Asimismo, la (Comisión Económica para América Latina y el Caribe, 2023) identifica a Costa Rica como uno de los países centroamericanos con mayor adopción de medios de pago electrónicos.

Dado este entorno digital y la amplia bancarización, el poder de negociación de los clientes se considera alto.

Poder de negociación de los proveedores

Aunque Costa Rica posee un sector tecnológico relativamente desarrollado, los bancos continúan dependiendo de proveedores internacionales para sistemas core, infraestructura de ciberseguridad y plataformas de pagos críticos. Esta dependencia se mantiene debido a que la sustitución de proveedores exige capacidades técnicas avanzadas y altos costos de implementación (OECD, 2022).

Por su parte, la Superintendencia General de Entidades Financieras (2024), refuerza los estándares prudenciales y de continuidad operativa, incrementando la necesidad de servicios especializados (Superintendencia General de Entidades Financieras, 2024). Por ello, el poder de negociación de los proveedores se clasifica como medio-alto.

Amenaza de nuevos competidores

Aunque el marco regulatorio de Costa Rica mantiene altos requisitos de supervisión

prudencial, el país experimenta un crecimiento constante en su ecosistema fintech. El informe de (Finnovista et al., 2024) señala que Costa Rica supera las 50 fintech activas, principalmente en pagos, crédito alternativo y soluciones de infraestructura.

Además, las condiciones macroeconómicas más estables incluyendo tasas de interés más bajas y una inflación moderada favorecen la expansión de servicios digitales (Banco Central de Costa Rica, 2025). Por estas razones, la amenaza de nuevos competidores se clasifica como media.

Amenaza de productos sustitutos

Costa Rica presenta una amplia gama de sustitutos para los servicios bancarios tradicionales. Las cooperativas, las fintech, las plataformas de pago electrónico y los neobancos emergentes compiten directamente por servicios como transferencias, pagos recurrentes y créditos de bajo monto.

Costa Rica se encuentra entre los países con mayor uso de pagos digitales en la región, lo que refleja una sustitución progresiva de los métodos bancarios tradicionales (Banco Mundial, 2023). A su vez, se documenta la expansión de servicios tecnológicos que reducen la dependencia del usuario hacia la banca formal (Finnovista et al., 2024).

La amenaza de sustitutos se clasifica como media-alta.

2.2.4 PANAMÁ

Rivalidad entre competidores existentes

El sistema financiero panameño es uno de los más grandes y competitivos de Centroamérica, debido a su condición de centro financiero internacional y a la presencia de bancos locales e internacionales. Según el informe de actividad bancaria de la Superintendencia de Bancos de Panamá (2024), el total de activos del Centro Bancario Internacional alcanzó USD 156,392.8 millones en 2024, con un crecimiento interanual de 6.0 %, lo que refleja un mercado dinámico con alta actividad de crédito y depósitos (Superintendencia de Bancos de Panamá, 2024).

La liquidez del sistema fue significativa, con un índice de liquidez promedio de 54.29 %, y el índice de adecuación de capital se situó en 15.29 %, por encima de los mínimos regulatorios, lo que otorga una base sólida para que los bancos compitan agresivamente en productos y servicios (Superintendencia de Bancos de Panamá, 2024).

Además, la cartera neta de crédito alcanzó los USD 95,186.7 millones, con una variación positiva de +9.1 %, lo que indica que los bancos están en constante lucha por colocar más créditos a clientes corporativos y personales. Por estas razones, la rivalidad entre competidores en Panamá se clasifica como alta.

Poder de negociación de los clientes

El mercado panameño presenta una amplia oferta de instituciones financieras, incluyendo numerosos bancos internacionales especializados en banca corporativa, comercio exterior y servicios de alto valor. En 2024, los depósitos totales del sistema alcanzaron USD 110,484.5 millones, con un incremento de 5.1 %, lo que da a los clientes múltiples opciones para elegir y comparar condiciones de ahorro y crédito (Superintendencia de Bancos de Panamá, 2024).

El creciente nivel de sofisticación del cliente, especialmente en segmentos corporativos y de banca de inversión, obliga a las instituciones a ofrecer productos diferenciados y a mejorar sus servicios digitales para mantener retención de clientes. Por tanto, el poder de negociación de los clientes se considera alto.

Poder de negociación de los proveedores

Los bancos panameños dependen considerablemente de proveedores tecnológicos especializados para sistemas core bancarios, infraestructura de pagos internacionales, soluciones de ciberseguridad y servicios de datos. Aunque Panamá cuenta con un sector de servicios financieros avanzado, la mayoría de las soluciones tecnológicas de alto nivel siguen siendo importadas, lo que limita la capacidad de negociación de los bancos frente a proveedores globales (Superintendencia de Bancos de Panamá, 2024).

El informe de actividad bancaria también señala que estos proveedores son críticos para mantener la continuidad operativa y el cumplimiento de estándares internacionales de gestión de riesgos. Por ello, el poder de negociación de los proveedores se clasifica como medio-alto (Superintendencia de Bancos de Panamá, 2024).

Amenaza de nuevos competidores

Las barreras regulatorias para la entrada de nuevos bancos tradicionales en Panamá son altas, debido a los requisitos estrictos de supervisión prudencial y de solvencia. Sin

embargo, el país es atractivo para nuevos actores debido a su posición de hub financiero internacional.

El ecosistema fintech ha mostrado crecimiento, especialmente en servicios de pagos digitales, remesas y soluciones de inversión alternativa, siguiendo tendencias regionales reportadas (Finnovista et al., 2024). Aunque la regulación exigente limita la entrada de competidores masivos, la expansión de soluciones digitales y la disponibilidad de servicios alternativos colocan la amenaza de nuevos entrantes en un nivel medio-alto.

Amenaza de productos sustitutos

La amenaza de sustitutos para los bancos tradicionales en Panamá proviene del auge de las fintech, plataformas de pago internacionales y soluciones digitales que ofrecen servicios financieros fuera del sistema tradicional (Finnovista et al., 2024). El crecimiento del comercio electrónico, la digitalización de pagos y la adopción de aplicaciones financieras móviles han incrementado la competitividad de alternativas no bancarias, lo que presiona a las instituciones tradicionales a innovar.

Aunque los clientes corporativos todavía confían considerablemente en la banca formal para servicios complejos, en segmentos de consumo y pagos cotidianos las opciones digitales representan una presión constante. Por ello, la amenaza de sustitutos se clasifica como alta.

2.2.5 HONDURAS

Rivalidad entre competidores existentes

El sistema bancario hondureño presenta una estructura altamente concentrada, donde un grupo reducido de instituciones concentra la mayor parte de los activos del sistema financiero nacional, configurando un entorno de competencia oligopólica (Moodys Local Honduras, 2025). Esta concentración no implica ausencia de competencia; por el contrario, la rivalidad se manifiesta principalmente en la captación de depósitos, colocación de crédito de consumo y empresarial, y expansión de canales de atención.

De acuerdo con la Comisión Nacional de Bancos y Seguros (2025), el sistema financiero hondureño cuenta con más de 21,600 puntos de contacto, lo que evidencia una fuerte presencia de canales físicos, especialmente en agencias y ventanillas, que siguen concentrando una proporción

relevante de las operaciones financieras (Comisión Nacional de Bancos y Seguros, 2025). Esta característica incrementa la presión competitiva en la red de sucursales, donde los bancos compiten por eficiencia operativa y por una mejor asignación de esfuerzos comerciales.

En este contexto, la adopción de herramientas analíticas avanzadas y modelos predictivos se convierte en un elemento diferenciador clave, ya que permite optimizar la gestión comercial en un mercado saturado y regulado. Por estas razones, la rivalidad competitiva en Honduras se clasifica como alta.

Poder de negociación de los clientes

El poder de negociación de los clientes en Honduras ha aumentado progresivamente como resultado de la digitalización financiera y de una mayor disponibilidad de información. Según la CNBS (2025), cerca del 90 % de las instituciones supervisadas cuentan con canales digitales activos, lo que facilita la comparación de productos, tasas y comisiones entre bancos (Comisión Nacional de Bancos y Seguros, 2025).

No obstante, el comportamiento del cliente hondureño sigue mostrando una fuerte preferencia por los canales presenciales. En el caso de las remesas, aproximadamente el 51 % de las transacciones se realiza a través de canales físicos, frente a un 23 % mediante medios digitales, lo que evidencia una coexistencia entre banca tradicional y digital (Comisión Nacional de Bancos y Seguros, 2025). Esta dualidad reduce parcialmente el poder de negociación del cliente, ya que los costos de cambio continúan siendo relevantes para segmentos menos digitalizados.

Aun así, la creciente adopción de productos financieros y la competencia interbancaria han fortalecido la capacidad del cliente para exigir mejores condiciones, especialmente en productos de consumo y servicios transaccionales. En consecuencia, el poder de negociación de los clientes se clasifica como medio-alto.

Poder de negociación de los proveedores

Los bancos hondureños dependen en gran medida de proveedores internacionales de tecnología para sistemas core bancarios, plataformas de pagos, soluciones de ciberseguridad y servicios de infraestructura de datos. Esta dependencia se ve reforzada por la limitada disponibilidad de proveedores locales especializados en analítica avanzada e inteligencia artificial (PricewaterhouseCoopers, 2020).

La normativa prudencial y tecnológica emitida por la CNBS exige estándares elevados en gestión de tecnologías de información, continuidad del negocio y ciberseguridad, lo que incrementa los costos de implementación y reduce la capacidad de sustitución de proveedores (Comisión Nacional de Bancos y Seguros, 2022). Esta situación otorga a los proveedores tecnológicos una posición negociadora relativamente fuerte frente a las instituciones financieras.

Por lo tanto, el poder de negociación de los proveedores en Honduras se clasifica como medio-alto.

Amenaza de nuevos competidores

Las barreras de entrada al sistema bancario hondureño son elevadas debido a requisitos regulatorios estrictos, exigencias de capital y supervisión prudencial. Estas condiciones limitan la entrada de nuevos bancos tradicionales y protegen parcialmente a las instituciones existentes (Diario Oficial La Gaceta, 2004).

Sin embargo, el crecimiento del ecosistema fintech en América Latina ha comenzado a ejercer presión indirecta sobre el mercado hondureño. La región presenta un ecosistema fintech consolidado, con fuerte expansión en pagos digitales, crédito alternativo y servicios financieros basados en tecnología (Finnovista et al., 2024). Aunque Honduras muestra un desarrollo incipiente en este ámbito, la tendencia regional sugiere una amenaza creciente desde actores no bancarios.

En consecuencia, la amenaza de nuevos competidores en Honduras se clasifica como moderada, con una tendencia al alza en el mediano plazo.

Amenaza de productos sustitutos

Los principales productos sustitutos para la banca tradicional en Honduras provienen de cooperativas de ahorro y crédito, empresas de remesas, billeteras electrónicas y plataformas de pago digital. El uso de servicios financieros digitales en América Latina ha crecido de forma acelerada, impulsando alternativas a la banca tradicional, especialmente para pagos y transferencias (Banco Mundial, 2023).

En el contexto hondureño, estas soluciones han ganado relevancia en segmentos específicos, como remesas y pagos de bajo monto, aunque todavía no sustituyen completamente a los productos bancarios tradicionales digital (Comisión Nacional de Bancos y Seguros, 2025). No obstante, su crecimiento progresivo representa una presión competitiva que obliga a los bancos a

mejorar su eficiencia y personalización de servicios.

Por estas razones, la amenaza de productos sustitutos se clasifica como media.

2.2.6 ANÁLISIS COMPARATIVO DEL MICROENTORNO DEL SECTOR BANCARIO EN CENTROAMÉRICA

Con el propósito de sintetizar y contrastar las principales presiones competitivas que enfrenta el sector bancario en la región, se presenta a continuación un cuadro comparativo basado en el modelo de las Cinco Fuerzas de Porter. Este ejercicio permite identificar similitudes y diferencias estructurales entre los sistemas bancarios de los países anteriormente analizados.

El análisis comparativo resulta especialmente relevante para contextualizar el caso hondureño dentro del entorno centroamericano, ya que facilita la comprensión de cómo las dinámicas regionales (particularmente la digitalización, la presión fintech y la dependencia tecnológica) influyen en la competitividad de los bancos locales.

Asimismo, este enfoque permite establecer una base analítica sólida para justificar la necesidad de adoptar herramientas avanzadas de analítica y modelos predictivos en la gestión comercial y operativa del sector bancario.

Tabla 1 Comparativo de las fuerzas de Porter

Fuerza competitiva	Guatemala	El Salvador	Costa Rica	Panamá	Honduras
Rivalidad entre competidores existentes	Media-alta. Alta concentración bancaria con crecimiento sostenido del crédito impulsa competencia en eficiencia y digitalización.	Alta. Mercado concentrado, dolarizado y con fuerte dinamismo en pagos digitales y transferencias inmediatas.	Media-alta. Competencia entre banca estatal dominante y banca privada dinámica, con presión creciente por digitalización.	Alta. Centro Bancario Internacional con presencia de bancos locales e internacionales y fuerte competencia en segmentos corporativos y de alto valor.	Alta. Mercado oligopólico con fuerte competencia en captación, crédito y presencia en canales físicos.

Fuerza competitiva	Guatemala	El Salvador	Costa Rica	Panamá	Honduras
Poder de negociación de los clientes	Medio. Mayor digitalización, pero aún con segmentos no bancarizados relevantes.	Alto. Amplia adopción de pagos digitales reduce costos de cambio y empodera al usuario.	Alto. Alta bancarización, transparencia regulatoria y ecosistema digital avanzado.	Alto. Clientes corporativos y personales altamente informados, con múltiples opciones financieras.	Medio-alto. Mayor acceso a información y canales digitales, aunque persiste dependencia de canales presenciales.
Poder de negociación de los proveedores	Medio-alto. Dependencia de proveedores internacionales de tecnología y escasez de alternativas locales.	Medio-alto. Infraestructura crítica de pagos y regulación técnica fortalecen el poder de proveedores especializados.	Medio-alto. Proveedores globales dominan soluciones core y ciberseguridad, pese a cierto desarrollo tecnológico local.	Medio-alto. Alta dependencia de proveedores tecnológicos globales para banca internacional y pagos transfronterizos.	Medio-alto. Fuerte dependencia de proveedores internacionales y altos costos de sustitución tecnológica.
Amenaza de nuevos competidores	Media. Barreras regulatorias altas, pero crecimiento fintech regional presiona al sistema.	Media-alta. Regulación habilita fintech y nuevos servicios digitales, elevando la amenaza.	Media. Regulación estricta limita bancos tradicionales, pero fintech crecen gradualmente.	Media-alta. Mercado atractivo para nuevos actores, aunque con requisitos regulatorios exigentes.	Moderada. Barreras regulatorias altas; fintech aún incipientes, pero con tendencia creciente.
Amenaza de productos sustitutos	Media-alta. Billeteras digitales y cooperativas ganan participación en pagos y ahorro.	Media-alta. Pagos electrónicos y soluciones de remesas sustituyen servicios tradicionales.	Media-alta. Ecosistema digital avanzado y cooperativas fuertes compiten con banca tradicional.	Alta. Fintech, plataformas internacionales de pago y soluciones digitales compiten agresivamente.	Media. Cooperativas, remesadoras y billeteras digitales crecen, pero no sustituyen totalmente a la banca.

Fuente: Elaboración propia

El cuadro comparativo evidencia que, si bien los sistemas bancarios centroamericanos presentan diferencias en tamaño, grado de internacionalización y nivel de desarrollo tecnológico, comparten un entorno de rivalidad competitiva elevada y una presión creciente por mejorar la

eficiencia operativa y la experiencia del cliente. Países como Panamá y El Salvador muestran los niveles más altos de competencia, impulsados por la presencia de bancos internacionales, una mayor adopción de pagos digitales y una estructura de mercado orientada a servicios financieros sofisticados.

Costa Rica se posiciona como un sistema relativamente equilibrado, caracterizado por una fuerte regulación prudencial y altos niveles de bancarización, lo que genera estabilidad, pero al mismo tiempo incrementa la presión por innovar en un mercado donde los clientes poseen un alto poder de negociación. Guatemala y Honduras, aunque presentan mercados más concentrados, enfrentan una competencia intensa en segmentos específicos, como crédito de consumo, remesas y servicios transaccionales, donde la diferenciación se vuelve cada vez más compleja.

De manera transversal, el poder de negociación de los proveedores tecnológicos se mantiene en niveles medio-altos en todos los países analizados, reflejando una dependencia regional de soluciones internacionales para infraestructura bancaria, ciberseguridad y analítica de datos. Esta situación limita la flexibilidad estratégica de las instituciones financieras y refuerza la necesidad de desarrollar capacidades internas en gestión de datos y modelos analíticos.

En el caso de Honduras, el análisis comparativo permite identificar un entorno en el que, pese a la existencia de barreras regulatorias que limitan la entrada de nuevos bancos, la presión competitiva se manifiesta en la eficiencia de los canales físicos y en la capacidad de los bancos para personalizar sus ofertas. En este sentido, la implementación de modelos de Machine Learning orientados a la predicción del comportamiento del cliente y a la priorización de esfuerzos comerciales se presenta como una respuesta estratégica alineada con las condiciones del microentorno regional.

Este análisis comparativo refuerza, por tanto, la pertinencia de la presente investigación, al

demostrar que la adopción de analítica avanzada no solo constituye una tendencia global, sino una necesidad competitiva concreta para el sistema bancario hondureño en el contexto centroamericano.

2.3 CONCEPTUALIZACIÓN

El presente apartado desarrolla los conceptos fundamentales que sustentan el modelo de Machine Learning propuesto para estimar la probabilidad de aceptación de productos financieros en canales físicos de un banco hondureño. Para cada concepto se presenta una definición conceptual y una definición operativa, con el objetivo de clarificar su aplicación dentro del modelo y la forma en que será utilizado o medido en el análisis empírico.

2.1.1 MACHINE LEARNING

Definición teórica:

El machine learning es una rama de la inteligencia artificial orientada al diseño de algoritmos capaces de aprender a partir de datos sin requerir instrucciones explícitas (Goodfellow et al., 2016). Su finalidad es identificar patrones, realizar clasificaciones o predicciones y mejorar su rendimiento con la experiencia. En el ámbito bancario, esta tecnología permite automatizar procesos, detectar comportamientos atípicos y anticipar decisiones de los clientes (McKinsey & Company, 2023)

Definición operativa.

En esta investigación, el Machine Learning constituye la herramienta central para la construcción del modelo predictivo. Se emplean algoritmos supervisados que procesan información histórica de los clientes para estimar la probabilidad de aceptación de productos financieros en canales físicos.

2.1.2 MODELO PREDICTIVO

Definición conceptual.

Un modelo predictivo es una representación analítica que utiliza datos históricos y técnicas estadísticas o algorítmicas para estimar la probabilidad de ocurrencia de un evento futuro

Definición operativa.

En el estudio, el modelo predictivo corresponde a un modelo de clasificación supervisada entrenado con datos históricos de ofertas bancarias. Su función es generar una probabilidad estimada de aceptación para cada cliente, permitiendo priorizar esfuerzos comerciales en el canal físico.

Medición.

El desempeño del modelo se evalúa mediante métricas de clasificación como exactitud, precisión, recall y AUC, utilizando conjuntos de entrenamiento y prueba.

2.1.3 PROBABILIDAD DE ACEPTACIÓN

Definición conceptual.

La probabilidad de aceptación es una medida estadística que representa la propensión de un cliente a adquirir un producto financiero específico, calculada a partir de variables explicativas observables (Hair et al., 2019).

Definición operativa.

En esta investigación, la probabilidad de aceptación constituye la variable dependiente del modelo. Se expresa como un valor continuo entre 0 y 1, generado por los algoritmos de Machine Learning, que indica la probabilidad estimada de que un cliente acepte una oferta de producto financiero en un canal físico.

Medición.

La probabilidad se obtiene como salida del modelo predictivo y se valida comparándola con los registros históricos reales de aceptación o rechazo de ofertas.

2.1.4 VARIABLES PREDICTORAS

Definición conceptual.

Las variables predictoras son características observables que permiten explicar o anticipar el comportamiento futuro de un individuo o evento (Mitchell, 1997).

Definición operativa.

En el modelo propuesto, las variables predictoras corresponden exclusivamente a información disponible antes del contacto comercial, incluyendo características demográficas,

historial de productos, comportamiento transaccional agregado y aceptación previa de ofertas.

Medición.

Las variables se representan mediante valores numéricos continuos (por ejemplo, frecuencia transaccional, monto promedio) y variables categóricas codificadas según su naturaleza.

2.1.5 TIPO DE PRODUCTO BANCARIO

Definición conceptual.

Los productos bancarios son instrumentos financieros ofrecidos por las instituciones para la captación, colocación o administración de recursos, tales como cuentas, créditos y tarjetas(Banco Mundial, 2023).

Definición operativa.

En el modelo, el tipo de producto ofrecido se incorpora como una variable categórica que identifica la clase de producto presentada al cliente durante la interacción comercial.

Medición.

La variable se codifica como categórica nominal, distinguiendo entre cuentas de ahorro, depósitos a plazo, préstamos personales, créditos hipotecarios y tarjetas de crédito, a partir de los registros operativos de ofertas.

2.1.6 CANALES DE ATENCIÓN

Definición conceptual.

Los canales de atención bancaria corresponden a los medios a través de los cuales las instituciones financieras interactúan con sus clientes para ofrecer productos, realizar transacciones y brindar servicios. En la banca moderna, estos canales pueden clasificarse en canales físicos y canales asistidos/digitales, los cuales coexisten como parte de una estrategia multicanal orientada

a mejorar la cobertura y eficiencia (Kotler & Keller, 2006).

Definición operativa.

En la presente investigación, los canales de atención identifican el medio mediante el cual se realiza la oferta de productos financieros. Se consideran dos categorías:

- Canal físico, correspondiente a interacciones presenciales realizadas en agencias bancarias.
- Canal asistido/digital, que agrupa las ofertas realizadas mediante contacto telefónico por parte de oficiales y aquellas efectuadas a través de medios digitales registrados en la base de datos institucional.

Medición.

El canal de atención se representa como una variable categórica nominal, con dos valores: canal físico y canal asistido/digital, según el tipo de interacción registrado en los sistemas operativos de la entidad bancaria.

2.4 TEORÍAS DE SUSTENTO

En el ámbito financiero, la toma de decisiones ha evolucionado de depender de métodos tradicionales a basarse en modelos avanzados que integran grandes volúmenes de datos y herramientas analíticas sofisticadas. En este contexto, el machine learning se ha convertido en una de las tecnologías más relevantes, permitiendo predecir comportamientos financieros con una precisión sin precedentes.

Para comprender su aplicación en la banca, es esencial analizar las teorías fundamentales que sustentan su desarrollo y adopción.

2.4.1 BASES TEÓRICAS

2.1.4.1 TEORÍA DEL APRENDIZAJE AUTOMÁTICO

El aprendizaje automático ha sido desde ya hace algunos años una de las bases sobre las cuales se apoya la inteligencia artificial moderna. Mitchell (1997), fue quien se encargó de

formalizar el concepto y lo definió como un proceso por el cual un programa mejora su desempeño en una tarea específica lo cual se produce a medida que recibe más datos y acumula experiencia (Mitchell, 1997). En otras palabras podríamos decir que el sistema no necesita que alguien lo programe para cada caso, sino que aprende por sí mismo al detectar patrones que se repiten en los datos.

En el ámbito bancario este enfoque se ha encargado de cambiar la manera en que se asignan los productos financieros. Ya que antes las decisiones solían depender mucho de la intuición o de reglas fijas en cambio ahora, los modelos pueden analizar historiales de transacciones, ingresos o hábitos de consumo y con eso estimar qué tipo de producto tiene mayor probabilidad de ser aceptado por cada cliente (Deloitte, 2024; Finnovista et al., 2024)

Es así que a medida que el sistema recibe más información, sus predicciones se vuelven más precisas, lo que termina beneficiando tanto a los bancos como a los clientes: los primeros optimizan sus esfuerzos y los segundos reciben ofertas más acordes a sus necesidades reales.

Esta teoría se encarga de sustentar directamente la hipótesis central del estudio al establecer que los algoritmos de machine learning pueden mejorar su desempeño predictivo conforme procesan más datos.

Aplicado al contexto hondureño esto viene a respaldar la idea de que un modelo de aprendizaje automático puede estimar con alta precisión la probabilidad de aceptación de productos financieros, ya que las decisiones comerciales del banco se basan en patrones históricos verificables y no únicamente en criterios subjetivos.

2.1.4.2 TEORÍA DEL COMPORTAMIENTO DEL CONSUMIDOR EN FINANZAS

Las decisiones financieras rara vez son puramente racionales. En realidad, están muy ligadas a emociones, percepciones y sesgos que, muchas veces, ni el propio consumidor nota. La Teoría de la Perspectiva, propuesta por Kahneman & Tversky (1979), explica que las personas tienden a evaluar las opciones financieras en términos de ganancias y pérdidas respecto a un punto de referencia personal, no necesariamente objetivo (Kahneman & Tversky, 1979). En otras palabras, dos personas pueden enfrentar la misma oferta, pero la interpretan de forma distinta según lo que sientan que ganan o pierden.

En este sentido, el comportamiento financiero está lleno de pequeños atajos mentales,

conocidos como heurísticas, y sesgos que distorsionan la toma de decisiones. Algunos de los más comunes en la banca son:

Efecto de encuadre: los clientes responden diferente dependiendo de cómo se presenta la información. Una promoción que resalta *tasa baja* puede parecer más atractiva que una que muestra el costo real anual, aunque el valor económico sea el mismo (Thaler & Sunstein, 2008).

Sesgo de statu quo: muchos consumidores evitan cambiar de banco o de producto por simple inercia o miedo a equivocarse (Kanaparthi, 2024). Por eso, la personalización basada en Machine Learning puede ayudar a detectar qué clientes están más abiertos a probar algo nuevo.

Heurística de disponibilidad: las personas tienden a decidir basándose en lo más reciente o en lo más fácil de recordar. Si alguien escuchó sobre un fraude con tarjetas de crédito, es probable que desconfíe de una nueva oferta, aunque los datos digan lo contrario (Kahneman & Tversky, 1979).

El Machine Learning ayuda a equilibrar esos sesgos porque puede analizar comportamientos pasados, respuestas a campañas y contextos personales para ofrecer productos más cercanos a lo que cada cliente valora. Por ejemplo, algunos reaccionan mejor ante descuentos en tasas, otros ante bonificaciones iniciales o programas de puntos. Detectar esos patrones permite ajustar las estrategias de oferta y mejorar la tasa de aceptación.

En el caso hondureño, donde la educación financiera aún varía mucho entre segmentos, este tipo de modelos permite adaptar el lenguaje y los incentivos. Así, la banca puede conectar mejor con las expectativas del cliente promedio y construir una relación más confiable. En última instancia, los modelos predictivos no solo sirven para vender más, sino para entender cómo perciben los clientes el riesgo y la oportunidad, ajustando la oferta a lo que realmente los mueve (Hernández & Moreno, 2023a).

Esta teoría explica el componente conductual que complementa la hipótesis del estudio: las decisiones de aceptación de productos financieros no dependen únicamente de factores económicos, sino también de percepciones y sesgos individuales.

De ahí que el modelo de machine learning planteado pueda identificar, a través de variables conductuales y de respuesta, qué tipo de oferta tiene mayor probabilidad de aceptación en el canal físico, ajustando la estrategia comercial a las preferencias reales del consumidor.

2.1.4.3 TEORÍA DE LA ASIMETRÍA DE INFORMACIÓN

La idea de la asimetría de información fue planteada por Akerlof (1970), para explicar por qué, a veces, los mercados no funcionan tan bien como deberían. En palabras simples, se trata de los casos donde una de las partes (por ejemplo, el banco) tiene menos información que la otra, o no la misma calidad de información (Akerlof, 1970). Eso genera un desbalance: el banco no sabe del todo a quién le está prestando y el cliente, a su vez, no siempre entiende del todo las condiciones o los riesgos que asume. En la banca, este problema aparece de muchas maneras:

La más conocida es la *selección adversa*, que ocurre cuando los bancos no logran distinguir bien entre clientes de bajo y alto riesgo. Si la información disponible es limitada o está incompleta, el banco podría terminar aprobando créditos a quienes presentan mayor riesgo y negándolos a quienes sí podrían pagar (Akerlof, 1970).

Durante años, las instituciones han dependido de los puntajes de crédito o de los reportes financieros, pero esos mecanismos no siempre muestran la historia completa de una persona. En consecuencia, ciertos grupos (especialmente los que tienen menos acceso formal al sistema financiero) quedan fuera del crédito, no por falta de capacidad, sino por falta de datos (Hernández & Moreno, 2023a)

Luego está el *riesgo moral*, que se presenta cuando un cliente cambia su comportamiento una vez que obtiene el préstamo. Puede endeudarse más o relajarse con los pagos, confiando en que el banco no podrá anticiparlo (Akerlof, 1970). Es en este punto especifico donde el machine learning entra en juego de una manera que se considera bastante útil.

Al analizar información en tiempo real, los algoritmos pueden detectar patrones de comportamiento que antes solían pasar desapercibidos, permitiendo que los bancos ajusten sus decisiones de forma más precisa (Rodriguez Villanueva et al., 2023).

Modelos como las *Redes Neuronales Artificiales* (ANN) o los árboles de decisión (*Random Forest), XGBoost*, permiten analizar volúmenes de información que hasta hace poco resultaban ser imposibles de procesar con las herramientas existentes o de manera manual. Entre esos datos se incluyen los hábitos de gasto, el historial de transacciones o incluso la forma en que los clientes interactúan con los canales digitales.

Algunas veces también se incorporan variables que son más difíciles de medir, como

factores sociales o geográficos que podrían influir en el comportamiento financiero. Todos estos elementos cuando se combinan ofrecen una visión más completa y dinámica del riesgo crediticio. En la práctica, esto se traduce en decisiones más justas entonces se puede otorgar crédito a personas que antes quedaban fuera de los modelos tradicionales, reduciendo los impagos y mejorando la precisión con la que se evalúa a cada solicitante (Jeff Montoya & Hernández Fúnez, 2023).

Además, el uso de machine learning no se queda solo en la gestión del riesgo; también ha abierto la posibilidad de diseñar productos mucho más personalizados. Las tasas, los plazos o las condiciones ya no tienen por qué ser iguales para todos, sino que pueden adaptarse al perfil de cada cliente. Para los bancos, eso se refleja en una rentabilidad más estable; para los usuarios, en un acceso más justo y transparente. En cierto modo, esta tecnología está ayudando a que la información deje de ser un obstáculo y se convierta en una herramienta que favorece a ambas partes (Miñano Sanchez, 2022).

El Machine Learning ayuda a reducir la asimetría de información al analizar grandes volúmenes de datos y descubrir patrones ocultos en el comportamiento financiero de los clientes. Algunas de sus aplicaciones incluyen:

- Evaluación de clientes sin historial crediticio: en países como Honduras, donde una parte significativa de la población no está bancarizada o tiene poca información en los burós de crédito, los modelos de ML pueden identificar señales alternativas de solvencia financiera a partir de datos transaccionales, comportamiento digital y hábitos de pago de servicios públicos.
- Predicción de la aceptación de productos financieros: la oferta de productos en canales físicos suele basarse en criterios generales, lo que puede generar rechazo en clientes con necesidades específicas. Los modelos de Machine Learning pueden personalizar las ofertas en función de datos individuales, aumentando la probabilidad de aceptación y reduciendo la incertidumbre en la asignación de productos.
- Segmentación dinámica en tiempo real: A diferencia de los métodos tradicionales, que dependen de perfiles estáticos de clientes, el ML permite actualizar la segmentación de clientes en tiempo real con base en cambios en su comportamiento financiero, optimizando la asignación de productos.

- Detección temprano de señales de riesgo moral: en los últimos años, los modelos basados en Redes Neuronales Artificiales (ANN), Random Forest y XGBoost han mostrado una capacidad notable para anticipar ciertos comportamientos de riesgo. Gracias a su nivel de detalle, pueden detectar señales tempranas de mora o sobreendeudamiento que, a simple vista, pasarían desapercibidas. No se trata únicamente de predecir probabilidades, sino de dar margen a los bancos para actuar antes de que aparezca el problema. En ese punto, entran estrategias más preventivas como reestructurar pagos, ofrecer asesorías personalizadas o, en algunos casos, generar alertas que permitan tomar decisiones a tiempo.

En el contexto hondureño, este tipo de enfoque cobra especial relevancia. Una gran parte de los clientes aún no tiene un historial crediticio formal, o sus registros están incompletos, lo que dificulta la evaluación con los métodos tradicionales. El Machine Learning ayuda a llenar esos vacíos aprovechando otras fuentes de información, desde los patrones de consumo hasta la frecuencia de uso de los canales físicos, e incluso variables socioeconómicas que antes se ignoraban.

Implementar estas soluciones no solo mejora la eficiencia operativa de los bancos, sino que también amplía las oportunidades para quienes históricamente han quedado fuera del sistema. Los modelos predictivos abren la puerta a un crédito más inclusivo, ajustado a la realidad de cada persona. Y aunque a veces se vea como una cuestión meramente técnica, en el fondo significa avanzar hacia un sistema financiero más equitativo, donde la tecnología corrige las brechas que antes dejaba la asimetría de información.

Esta teoría se logra vincular directamente con la hipótesis del estudio, al establecer que el machine learning puede mitigar la asimetría de información entre bancos y clientes mediante el análisis de datos masivos. De este modo el modelo predictivo propuesto no solo optimiza la oferta de productos financieros, sino que contribuye a decisiones más justas y basadas en evidencia, reduciendo la incertidumbre en los canales físicos del sistema bancario hondureño.

2.1.4.4 SÍNTESIS INTEGRADORA CON HIPÓTESIS

Las teorías revisadas en esta sección proporcionan el sustento conceptual de la hipótesis central del estudio, la cual plantea que las variables asociadas al cliente ejercen un efecto significativo sobre la aceptación de ofertas de productos financieros realizadas en canales físicos.

En primer lugar, la Teoría del Aprendizaje Automático fundamenta la hipótesis al establecer que los algoritmos de Machine Learning son capaces de aprender patrones complejos a partir de datos históricos y mejorar su capacidad predictiva conforme aumenta la información disponible. Desde esta perspectiva, la relación funcional planteada en el estudio,$Y = f(X1, X2, \dots, Xn)$, se justifica teóricamente como un mecanismo válido para estimar la probabilidad de aceptación de una oferta financiera a partir de múltiples características del cliente.

En segundo lugar, la Teoría del Comportamiento del Consumidor en Finanzas sustenta la hipótesis al explicar que la aceptación de productos financieros no responde únicamente a criterios económicos racionales, sino que está influenciada por factores conductuales, percepciones subjetivas y experiencias previas. Esta teoría respalda la inclusión de variables demográficas, financieras y de comportamiento como elementos relevantes dentro del conjunto de variables predictoras del modelo.

Finalmente, la Teoría de la Asimetría de Información aporta el fundamento teórico que explica por qué el uso de información interna del banco, procesada mediante modelos de Machine Learning, puede mejorar la toma de decisiones comerciales. Al reducir la incertidumbre entre la institución y el cliente, el modelo permite identificar con mayor precisión aquellos perfiles con mayor probabilidad de aceptación, reforzando así la hipótesis alternativa del estudio.

En conjunto, estas teorías no generan hipótesis independientes, sino que convergen en el sustento de una única hipótesis empírica, la cual será contrastada mediante el desempeño predictivo del modelo. Si el modelo demuestra una capacidad explicativa superior al azar y las variables del cliente presentan relevancia predictiva, se contará con evidencia suficiente para rechazar la hipótesis nula y aceptar la hipótesis alternativa planteada.

2.5 ANÁLISIS DE LAS METODOLOGÍAS

La literatura reciente sobre analítica bancaria y marketing financiero evidencia un uso predominante de metodologías cuantitativas de carácter predictivo, orientadas a modelar la probabilidad de aceptación de productos financieros a partir de grandes volúmenes de datos históricos generados por las propias instituciones. En estos estudios, el interés metodológico se centra en maximizar la capacidad predictiva y la utilidad operativa de los modelos, más que en la inferencia causal clásica.

Diversos trabajos empíricos desarrollados en el sector financiero han adoptado diseños no experimentales, basados en el análisis de registros administrativos provenientes de data warehouses institucionales, sin intervención directa del investigador sobre las variables analizadas. Este enfoque metodológico resulta consistente con la realidad bancaria, donde la experimentación controlada sobre clientes puede implicar riesgos comerciales, regulatorios y reputacionales (Hernández Sampieri & Fernandez-Collado, 2014; Provost & Fawcett, 2013)

En cuanto a las técnicas empleadas, la literatura documenta un uso recurrente de modelos de machine learning supervisado, tales como regresión logística, árboles de decisión, random forest y algoritmos de boosting, aplicados a problemas de predicción de respuesta en campañas comerciales y concesión de productos financieros. Estudios recientes en contextos fintech y bancarios han demostrado que estos enfoques superan a los métodos estadísticos tradicionales en escenarios caracterizados por relaciones no lineales y desbalance de clases, particularmente en tareas de clasificación binaria asociadas a la aceptación de ofertas (Alonso-Robisco & Carbo, 2023; Attota, 2024; Rodriguez Villanueva et al., 2023).

No obstante, los estudios previos también señalan limitaciones metodológicas recurrentes, entre las que destacan la dependencia de datos provenientes de una sola institución, la dificultad para generalizar los modelos a otros contextos organizacionales y la limitada integración de criterios de explicabilidad y gobernanza del modelo. Estas restricciones han motivado el desarrollo de enfoques metodológicos que busquen equilibrar desempeño predictivo, interpretabilidad y alineación con los objetivos estratégicos del negocio, especialmente en entornos financieros regulados (Provost & Fawcett, 2013).

En este contexto, el presente estudio se inscribe dentro de las metodologías documentadas en la literatura, al adoptar un enfoque cuantitativo–predictivo basado en datos administrativos reales y técnicas de machine learning supervisado, orientado a generar un modelo con utilidad práctica para la gestión comercial bancaria, manteniendo coherencia metodológica con investigaciones previas desarrolladas en el sector.

2.6 ANTECEDENTES DE METODOLOGÍAS

Durante los últimos años, el uso de metodologías analíticas en el sector financiero ha

experimentado una transformación significativa. La digitalización empujó ese cambio, y con ella vino la presión por tomar decisiones más rápidas y basadas en evidencia. Hoy, la mayoría de los estudios sobre machine learning aplicados a la banca trabajan con metodologías más ordenadas, que incluyen todas las etapas del proceso analítico: limpieza, transformación, entrenamiento, validación y evaluación (James et al., 2022; Provost & Fawcett, 2013) .

Los estudios más recientes coinciden en la adopción de metodologías sistemáticas que integran todas las etapas del proceso analítico, incluyendo la limpieza y transformación de datos, el entrenamiento de modelos, la validación cruzada y la evaluación mediante métricas objetivas de desempeño. Estos enfoques, comúnmente denominados pipelines analíticos, han reemplazado prácticas menos estructuradas utilizadas en etapas tempranas de la analítica bancaria, permitiendo una mayor trazabilidad metodológica y consistencia en los resultados obtenidos.

Asimismo, la revisión de antecedentes evidencia que, si bien existen patrones metodológicos comunes, las decisiones específicas de diseño varían en función del contexto institucional, el tipo de producto financiero analizado y el canal de interacción con el cliente. En economías emergentes, diversos autores destacan la necesidad de adaptar estas metodologías a realidades donde la madurez analítica y la infraestructura de datos aún se encuentran en proceso de consolidación.

2.6.1 EVOLUCIÓN DE LOS ENFOQUES METODOLÓGICOS

A partir del año 2020, la literatura sobre analítica bancaria ha evidenciado un desplazamiento progresivo de los modelos lineales clásicos hacia algoritmos de aprendizaje supervisado capaces de capturar relaciones no lineales y manejar conjuntos de datos heterogéneos y desbalanceados (Gareth James,Sohil et al., 2022; Provost & Fawcett, 2013). Este cambio no se limita a la sustitución de herramientas estadísticas, sino que refleja una transformación metodológica más amplia, en la cual los datos pasan a ocupar un rol central en la identificación de patrones relevantes para la toma de decisiones.

Desde una perspectiva técnica, los estudios recientes enfatizan la importancia de metodologías reproducibles, con procesos rigurosos de calibración de hiperparámetros y validación del desempeño del modelo. En la práctica, esto se traduce en la evaluación sistemática de los modelos mediante métricas como Recall, F1-score y AUC-ROC, priorizando no solo la exactitud puntual, sino también la capacidad de generalización frente a datos no observados

(Rodriguez Villanueva et al., 2023).

Este énfasis metodológico ha contribuido a elevar los estándares de calidad en la investigación aplicada al sector financiero, aunque también ha incrementado la complejidad técnica y los requerimientos de ejecución de los estudios. En la práctica, esto significa que los modelos se evalúan no solo por su exactitud, sino por su capacidad para generalizar, utilizando métricas. Todo eso ha hecho que los estudios actuales sean más confiables, pero también un poco más exigentes en su ejecución.

2.6.2 REVISIÓN DE INVESTIGACIONES RELEVANTES

El estudio de García Hernández y Torres Moreno (2022) constituye un referente en la región latinoamericana. Los autores desarrollaron un pipeline de predicción crediticia para detectar tempranamente el riesgo en carteras bancarias, utilizando XGBoost como algoritmo base (Hernández & Moreno, 2023b).

Su metodología combinó la normalización de datos mediante Min–Max Scaling, la división del conjunto en entrenamiento (70 por ciento) y prueba (30 por ciento), y la validación cruzada K-Fold (k = 5). Además, aplicaron Grid Search para optimizar los hiperparámetros y evitar el sobreajuste. Su hallazgo principal fue que la regularización y calibración del modelo incrementaron la capacidad de generalización en un 10 por ciento, confirmando la importancia de los procedimientos de validación múltiple.

De forma complementaria, (Rodriguez Villanueva et al., 2023), propusieron una metodología híbrida de tres fases: (1) exploración y limpieza de datos, (2) codificación de variables categóricas mediante One-Hot Encoding y estratificación de la muestra, y (3) entrenamiento con Random Forest y Support Vector Machines.

Las métricas de evaluación incluyeron Precision, Recall, F1-score y AUC-ROC, mostrando que Random Forest alcanzó el mejor equilibrio entre sensibilidad y especificidad. Esta estructura metodológica integró además la identificación de variables relevantes mediante feature importance, lo que mejoró la interpretabilidad de los resultados.

Por su parte, otros autores desarrollaron un enfoque metodológico centrado en la interpretabilidad, combinando regresión logística con k-means clustering (Zhang et al., 2023). La estandarización de variables numéricas, la reducción de dimensionalidad con Principal Component

Analysis (PCA) y la validación cruzada repetida redujeron la varianza del error y permitieron explicar con mayor detalle las relaciones entre el comportamiento del cliente y la aceptación de productos.

Este estudio resalta la necesidad de integrar la precisión predictiva con la capacidad explicativa, especialmente en contextos donde las decisiones deben ser transparentes para el usuario.

Finalmente en estudios recientes sobre campañas de marketing financiero se han comparado modelos de clasificación como regresión logística, máquinas de "boosting" (por ejemplo, Gradient Boosting) y Naïve Bayes, combinando normalización de variables, creación de ratios cliente-producto y técnicas de balanceo de clases mediante SMOTE (Attota, 2024; Tanvir et al., 2024)

Los resultados mostraron que los modelos basados en árboles ofrecieron mayor estabilidad frente al desbalance de clases y mejor capacidad de generalización.

2.6.3 TENDENCIAS METODOLÓGICAS COMUNES

Los estudios revisados presentan varios puntos en común:

- Estructura sistemática: todos los trabajos siguen una secuencia coherente de preprocesamiento, entrenamiento y validación.
- Uso de validación cruzada: emplean esquemas de K-Fold o validación repetida para asegurar estabilidad.
- Optimización de hiperparámetros: utilizan métodos automáticos de búsqueda (Grid Search, Random Search) para mejorar el desempeño del modelo.
- Combinación de precisión e interpretabilidad: los modelos más recientes buscan equilibrar desempeño predictivo con explicabilidad, utilizando métricas interpretativas como feature importance y SHAP values.

Estas prácticas reflejan un consenso metodológico orientado a maximizar tanto la robustez técnica como la utilidad operativa de los modelos predictivos.

2.6.4 ANÁLISIS COMPARATIVO Y APLICABILIDAD AL CONTEXTO HONDUREÑO

El análisis comparativo de los antecedentes revisados evidencia que la aplicación de metodologías reproducibles y validadas contribuye de manera significativa a mejorar la precisión y estabilidad de los modelos predictivos en el sector financiero. No obstante, una proporción relevante de los estudios se concentra en contextos digitales o en problemas asociados al riesgo crediticio, con menor atención a la modelación de decisiones comerciales en canales de atención presenciales.

Esta brecha metodológica resulta particularmente relevante en países como Honduras, donde la interacción física continúa desempeñando un papel central en la relación banco–cliente. La literatura sugiere que la adaptación de pipelines analíticos consolidados a estos contextos permite capturar comportamientos específicos que no siempre se reflejan en los canales digitales.

En síntesis, los antecedentes metodológicos revisados muestran un consenso en torno al uso de machine learning supervisado, validación rigurosa y énfasis en la reproducibilidad. El presente estudio toma estas buenas prácticas como referencia, situándose dentro de la evolución metodológica observada en la literatura, y contribuye a su aplicación en un entorno bancario que, si bien presenta un menor nivel de digitalización, dispone de información estructurada suficiente para el desarrollo de modelos predictivos de alto valor analítico.

2.7 METODOLOGÍAS, ENFOQUES Y DISEÑOS

Las metodologías, los enfoques y los diseños de investigación conforman el marco estructural que orienta la forma en que se organiza, ejecuta y valida el proceso científico.

La correcta diferenciación de estos tres conceptos no es meramente terminológica, sino esencial para asegurar la coherencia lógica del estudio y la validez de sus resultados (Hernández Sampieri & Fernandez-Collado, 2014; Robson & McCartan, 2017)

2.7.1 METODOLOGÍA

La revisión de la literatura evidencia que los estudios orientados a la aplicación de técnicas de machine learning en el sector bancario han adoptado predominantemente metodologías cuantitativas de carácter aplicado, sustentadas en el análisis de datos administrativos históricos. Estas metodologías se enfocan en transformar información transaccional y de comportamiento del cliente en modelos predictivos con utilidad operativa para la gestión comercial y el control del

riesgo (Hair et al., 2019; Provost & Fawcett, 2013).

En investigaciones recientes, este tipo de metodología ha permitido modelar la probabilidad de aceptación de productos financieros mediante el uso de variables demográficas, financieras y conductuales, priorizando métricas objetivas de desempeño como la precisión, el F1-score y el área bajo la curva ROC, en lugar de la inferencia causal clásica (Attota, 2024; Rodriguez Villanueva et al., 2023).

2.7.2 ENFOQUE

En cuanto al enfoque de investigación, la literatura especializada muestra una clara prevalencia de enfoques deductivos y correlacionales en estudios de analítica bancaria. Estos enfoques parten de teorías económicas, de comportamiento del consumidor o de aprendizaje automático para contrastarlas empíricamente mediante modelos predictivos entrenados sobre datos históricos (Creswell & Creswell, 2018).

Diversos estudios aplicados en el contexto financiero han demostrado que el enfoque deductivo resulta especialmente adecuado cuando se busca evaluar la relación entre las características del cliente y la aceptación de productos financieros, manteniendo coherencia teórica sin sacrificar capacidad predictiva (Hernández & Moreno, 2023a).

2.7.3 DISEÑO

Respecto al diseño metodológico, los trabajos revisados emplean mayoritariamente diseños no experimentales, basados en el análisis de registros históricos provenientes de data warehouses institucionales. Estos diseños suelen ser transversales, al trabajar con cortes analíticos consolidados de datos acumulados en periodos determinados, y correlacionales, al centrarse en la identificación de asociaciones estadísticas entre variables sin establecer relaciones causales directas (Robson & McCartan, 2017).

Este tipo de diseño ha sido ampliamente utilizado en estudios bancarios debido a que permite analizar el comportamiento real de los clientes sin intervenir directamente sobre las condiciones en que se generan los datos, preservando la validez ecológica de los resultados y reduciendo riesgos operativos y regulatorios.

2.7.4 SÍNTESIS METODOLÓGICA

En conjunto, el estado del arte metodológico evidencia una convergencia hacia esquemas

cuantitativos, no experimentales y orientados a la predicción, que han demostrado ser adecuados para el análisis del comportamiento del cliente en entornos bancarios reales. Estas metodologías, enfoques y diseños constituyen la base sobre la cual se desarrollan investigaciones aplicadas en analítica bancaria y marketing financiero.

2.8 ANÁLISIS CRÍTICO DE METODOLOGÍAS

El análisis crítico no consiste solo en enumerar métodos, sino en pensar qué tan útiles son para el problema que se quiere resolver. En el campo del machine learning bancario, las técnicas se pueden dividir (a grandes rasgos) en tres grupos: modelos estadísticos clásicos, algoritmos supervisados y modelos no supervisados (Domingos, 2016; James et al., 2022). Cada uno posee su propio valor en función a lo que persiga el investigador.

2.8.1 MODELOS ESTADÍSTICOS TRADICIONALES

Durante muchos años, la regresión logística y otros modelos parecidos fueron la base de la analítica financiera. Su mayor ventaja está en que se entienden fácilmente: uno puede ver qué variable influye en qué resultado, lo cual facilita explicarlo a quienes toman decisiones (Wooldridge, 2010).

Sin embargo, estos tienen límites ya que cuando las relaciones entre variables se vuelven más complicadas o cuando los datos crecen demasiado, estos modelos se quedan cortos. No logran capturar bien las interacciones o los comportamientos que cambian de forma no lineal, algo que ocurre con frecuencia en los bancos actuales.

2.8.2 MODELOS DE APRENDIZAJE SUPERVISADO

Los modelos supervisados (como Decision Trees, Random Forest o XGBoost) aprenden con ejemplos ya etiquetados. Son mucho más flexibles y detectan relaciones que los métodos estadísticos no pueden. Su aplicación en la banca internacional ha mejorado hasta en un 20 por ciento la precisión de las predicciones sobre el comportamiento de los clientes (McKinsey & Company, 2023).

La gran ventaja es que se adaptan: mientras más datos reciben, más aprenden. Sin embargo, no todo es perfecto. Estos modelos suelen ser menos transparentes; cuesta explicar por qué toman ciertas decisiones. Por eso últimamente se usan técnicas como SHAP values o LIME explanations, que sirven para abrir la caja negra del modelo y entender su lógica interna (Molnar, 2020).

2.8.3 MODELOS NO SUPERVISADOS Y ENFOQUES HÍBRIDOS

Los modelos no supervisados, como k-means o PCA, no necesitan etiquetas para funcionar; buscan patrones ocultos en los datos. Suelen ser útiles para segmentar clientes o reducir dimensiones (Hair et al., 2019), aunque no sirven para predecir directamente una respuesta.

Para superar eso, algunos investigadores han combinado enfoques. Por ejemplo, algunos autores han propuesto un modelo híbrido que une regresión logística con segmentación por clustering, logrando mejores resultados sin perder interpretabilidad (Rodriguez Villanueva et al., 2023). Este tipo de estrategias tiene sentido en contextos como el hondureño, donde se quiere predecir, pero también entender el comportamiento del cliente.

2.8.4 EVALUACIÓN CRÍTICA DE LAS ALTERNATIVAS

En la literatura revisada, los modelos supervisados son los que mejor equilibran precisión, robustez y aplicabilidad práctica. Frente a los modelos estadísticos, ofrecen mayor capacidad para detectar relaciones no lineales y adaptarse a datos de gran volumen. Frente a los modelos no supervisados, proporcionan resultados más útiles para la toma de decisiones, ya que permiten calcular probabilidades concretas de aceptación.

No obstante, los enfoques supervisados requieren más recursos computacionales y una gestión cuidadosa de los datos, especialmente en lo referente a la privacidad y la seguridad. Según el Banco de Pagos Internacionales, la implementación ética de inteligencia artificial en finanzas exige mecanismos de gobernanza y auditoría algorítmica que garanticen trazabilidad y equidad (Bank for International Settlements, 2017).

2.8.5 ARTICULACIÓN METODOLÓGICA: SELECCIÓN DEL ENFOQUE PREDICTIVO

El problema de investigación abordado en este estudio consiste en estimar la probabilidad de aceptación de productos financieros a partir de datos históricos de clientes, donde el resultado de interés se encuentra claramente observado y definido de manera binaria (acepta o no acepta la oferta). Esta característica del problema determina la necesidad de un enfoque cuantitativo y predictivo, orientado a modelar relaciones funcionales entre múltiples variables explicativas y una variable objetivo específica.

Bajo este contexto, el uso de modelos de aprendizaje supervisado resulta metodológicamente más adecuado que los enfoques no supervisados. Los modelos no

supervisados, como las técnicas de clustering o reducción de dimensionalidad, se orientan a la exploración de patrones sin una variable objetivo explícita, lo que limita su capacidad para estimar probabilidades individuales de aceptación, objetivo central de esta investigación (Hair et al., 2019). Por esta razón, dichos enfoques se consideran complementarios para análisis exploratorios, pero no apropiados como metodología principal.

Dentro del aprendizaje supervisado, es necesario distinguir entre modelos de regresión y modelos de clasificación. Los modelos de regresión se emplean cuando la variable dependiente es continua, mientras que los modelos de clasificación se utilizan cuando el resultado a predecir corresponde a categorías discretas (James et al., 2022).

En este estudio, la aceptación de una oferta financiera se define como una variable dicotómica, lo que configura el problema como uno de clasificación binaria.

En consecuencia, se opta por modelos de clasificación que permiten estimar probabilidades de pertenencia a la clase positiva y evaluar su desempeño mediante métricas predictivas adecuadas, como el área bajo la curva ROC (AUC), la precisión y el recall. Dentro de este grupo, la regresión logística se incorpora como un modelo base de referencia debido a su interpretabilidad y uso extendido en el sector financiero, mientras que algoritmos más flexibles, como los árboles de decisión y el Random Forest, permiten capturar relaciones no lineales y mejorar la capacidad predictiva del modelo.

De este modo, la selección metodológica no responde a una preferencia técnica aislada, sino a una evaluación razonada y coherente entre la naturaleza del problema, el tipo de datos disponibles y los objetivos del estudio. El enfoque supervisado de clasificación adoptado permite contrastar empíricamente la hipótesis planteada y generar resultados aplicables al contexto operativo de la banca hondureña.

2.8.6 JUSTIFICACIÓN METODOLÓGICA DEL ESTUDIO

El modelo de machine learning propuesto en esta tesis adopta un enfoque supervisado, comparativo y explicativo, utilizando algoritmos de clasificación como la regresión logística, el árbol de decisión, Gradient Boost y el Random Forest. Esta selección se justifica por tres razones principales:

- Adecuación a los objetivos del estudio. Los algoritmos de clasificación permiten

estimar una probabilidad binaria (acepta o no acepta una oferta), directamente alineada con la variable dependiente del modelo.

- Balance entre precisión e interpretabilidad. La combinación de modelos simples (regresión) y complejos (bosques aleatorios) permite obtener resultados robustos y, a la vez, comprensibles para los analistas bancarios.
- Relevancia en el contexto hondureño. La infraestructura tecnológica de la banca nacional aún se encuentra en desarrollo, por lo que los modelos seleccionados resultan accesibles y sostenibles, sin requerir plataformas de inteligencia artificial de alto costo.

2.8.7 REFLEXIÓN CRÍTICA Y PROYECCIÓN FUTURA

El análisis crítico confirma que la metodología supervisada elegida no sólo es pertinente, sino también adaptable al futuro. Su aplicación en los canales físicos permitirá validar empíricamente la hipótesis de que la adopción de modelos predictivos puede mejorar la efectividad comercial de la banca hondureña. Además, abre la posibilidad de evolucionar hacia esquemas más avanzados (como Deep Learning o Reinforcement Learning) cuando el volumen y la calidad de los datos lo permitan.

La metodología adoptada, por tanto, no es un punto de llegada, sino un punto de partida hacia la consolidación de una cultura analítica en la banca nacional, donde las decisiones se sustenten cada vez más en evidencia empírica y modelos replicables.

2.9 INSTRUMENTOS A UTILIZAR

El desarrollo de modelos de analítica predictiva requiere herramientas que faciliten la manipulación de datos, la experimentación con distintos algoritmos y la interpretación de los resultados. La elección de estos instrumentos debe responder a criterios de coherencia metodológica, viabilidad técnica y alineación con las fases del ciclo CRISP-DM (Cross Industry Standard Process for Data Mining), ampliamente adoptado en la industria bancaria (Espinosa Zúñiga, 2020).

A continuación, se presentan tres ejes comparativos: (a) analítica y modelado, (b) visualización y comunicación, y (c) gestión y almacenamiento de datos. En cada uno se incluyen

al menos cuatro herramientas representativas, evaluadas con base en su enfoque, ventajas, limitaciones, escalabilidad e integración con otros entornos.

2.9.1 COMPARATIVO DE HERRAMIENTAS DE ANALÍTICA Y MODELADO

Tabla 2 Comparativo de Herramientas

Criterio	Python	R	KNIME Analytics Platform	RapidMiner
Enfoque principal	Lenguaje de programación orientado a *machine learning* y análisis de datos	Lenguaje estadístico de modelado inferencial	Plataforma visual de analítica y minería de datos sin código	Plataforma visual de flujos analíticos con módulos automatizados
Ventajas clave	Amplia comunidad; librerías maduras (Scikit-learn, Pandas, NumPy); escalabilidad alta	Potencia estadística y regresión avanzada; comunidad académica amplia	Permite construir flujos reproducibles; integración directa con Python y SQL	Diseño intuitivo y entorno guiado para prototipos rápidos
Limitaciones	Requiere dominio de programación; gestión manual de entornos	Integración limitada con sistemas empresariales; menor rendimiento con grandes volúmenes	Menor flexibilidad en algoritmos personalizados; rendimiento medio en big data	Licencias comerciales; menor escalabilidad para proyectos empresariales
Costo/licencia	Gratuito, de código abierto	Gratuito, de código abierto	Gratuito (versión *community*); versión empresarial bajo licencia	Versión gratuita limitada; licencias comerciales de costo medio
Nivel de experticia requerida	Medio–alto (requiere conocimientos en programación y estadística)	Medio (análisis estadístico y sintaxis R)	Medio (entorno visual con configuración modular)	Bajo (entorno visual asistido)
Capacidad instalada en la empresa	Alta: dominio operativo por parte	Media: uso esporádico para	Alta: experiencia institucional	Baja: no cuenta con instalación ni soporte interno

Criterio	Python	R	KNIME Analytics Platform	RapidMiner
	del equipo analítico	análisis exploratorio	consolidada en uso de flujos KNIME	
Evaluación general	Flexible, potente y con soporte técnico extendido	Ideal para análisis exploratorio	Equilibrio entre usabilidad y capacidad técnica	Recomendado para formación o prototipos

Fuente: Elaboración propia

cuatro herramientas presentan solidez técnica, aunque difieren en su aplicabilidad y nivel de integración institucional.

Python y R son los lenguajes más versátiles para programación analítica; sin embargo, R demanda mayor especialización estadística. KNIME y RapidMiner simplifican los procesos mediante flujos visuales.

En el contexto de la empresa, donde ya existe experiencia media–alta en Python y KNIME, y servidores compatibles con SQL, ambas herramientas permiten equilibrar potencia analítica, trazabilidad del flujo y capacidad de integración. Su adopción garantiza reproducibilidad y aprovechamiento de la infraestructura disponible.

2.9.2 COMPARATIVO DE HERRAMIENTAS DE VISUALIZACIÓN Y COMUNICACIÓN

Tabla 3 Comparativo de Herramienta de Visualización

Criterio	Power BI	Tableau	Looker Studio	Datawrapper
Interfaz y usabilidad	Intuitiva; integración con ecosistema Microsoft	Avanzada; personalización gráfica de alto nivel	Sencilla y colaborativa; orientada a la nube	Minimalista; ideal para visualización rápida
Integración con bases de datos	Excelente (SQL Server, Excel, Azure)	Amplia (múltiples fuentes y conectores)	Limitada a servicios en la nube	Básica
Costo/licencia	Licencia institucional	Licencia comercial de alto costo	Gratuito	Gratuito

	vigente (Power BI Pro)			
Automatización y actualización	Alta (dashboards conectados en tiempo real)	Media	Baja	Baja
Interactividad y análisis visual	Alta (filtros, jerarquías, visuales dinámicas)	Muy alta (visualizaciones personalizadas)	Media	Baja
Nivel de experticia requerida	Medio (manejo de modelos de datos y diseño de reportes)	Alto (configuración avanzada de cálculos y diseño)	Bajo	Bajo
Capacidad instalada en la empresa	Alta: licencias vigentes y dominio operativo consolidado	Media: no se utiliza institucionalmente	Baja: no implementado en la infraestructura local	Nula
Evaluación general	Excelente integración institucional y bajo costo marginal	Superior en capacidades visuales, pero con mayor costo	Adecuado para informes básicos	Recomendado para visualización pública simple

Fuente: Elaboración propia

Las herramientas difieren principalmente en su nivel de sofisticación visual y requerimientos técnicos.

Tableau ofrece un rango más amplio de visualizaciones y personalización, aunque su costo y curva de aprendizaje lo hacen menos viable institucionalmente. Looker Studio y Datawrapper son opciones gratuitas adecuadas para proyectos de difusión, pero con capacidades limitadas de conexión y seguridad.

Power BI, por su parte, se integra de forma nativa con SQL Server y el entorno Microsoft

ya instalado en la empresa, permitiendo la creación de tableros ejecutivos dinámicos y de actualización automática. Dada la existencia de licencias corporativas y personal con dominio operativo, Power BI representa la opción más eficiente y sostenible para la visualización y comunicación de los resultados del modelo.

2.9.3 COMPARATIVO DE HERRAMIENTAS DE GESTIÓN Y ALMACENAMIENTO DE DATOS

Tabla 4 Comparativo de Herramienta de Almacenamiento

Criterio	SQL Server	MySQL	PostgreSQL	Oracle Database
Tipo de licencia	Comercial institucional (ya implementada en la empresa)	Libre	Libre	Comercial (alto costo)
Rendimiento en grandes volúmenes	Alto y estable (entorno empresarial)	Medio	Alto (buena opción open source)	Muy alto; estándar corporativo global
Seguridad y respaldo	Integrado con autenticación, cifrado y auditoría	Básico	Avanzado	Muy robusto y configurable
Compatibilidad con herramientas analíticas	Excelente (Power BI, KNIME, Python)	Media	Alta	Alta
Costo/licencia	Incluido en infraestructura actual	Gratuito	Gratuito	Alto costo de licencias y mantenimiento
Nivel de experticia requerida	Medio–alto (SQL avanzado)	Medio	Medio	Alto (administración compleja)
Capacidad instalada en la empresa	Alta: sistema activo con soporte institucional	Baja: no instalado	Media: viable técnicamente, no implementado	Baja: no cuenta con soporte interno

Evaluación general	Entorno óptimo para operaciones corporativas integradas	Recomendado para proyectos pequeños	Alternativa *open source* estable	Superior en rendimiento, pero de alto costo y complejidad

Fuente: Elaboración propia

Tras comparar las opciones, se concluye que la elección de herramientas depende de cuatro cosas: rendimiento, costo, experiencia del personal y compatibilidad con la infraestructura existente.

En la empresa analizada, SQL Server ya está en uso, y eso facilita su integración con KNIME, Python y Power BI. Juntas forman una arquitectura bastante completa: SQL Server guarda y gestiona los datos; KNIME se encarga del flujo de trabajo analítico; Python desarrolla y prueba los modelos (como Logistic Regression, Random Forest y XGBoost); y Power BI muestra los resultados de forma visual.

Podrían usarse otras herramientas como R o Tableau, claro, pero implicarían costos y curvas de aprendizaje mayores. Por eso, la selección final no se basa en preferencias personales, sino en la realidad institucional: aprovechar lo que ya se tiene, mantener eficiencia y asegurar trazabilidad ya que lo que se busca es un flujo analítico estable, práctico y replicable.

2.9.4 SÍNTESIS ANALÍTICA Y JUSTIFICACIÓN DEL CONJUNTO SELECCIONADO

El análisis comparativo demuestra que la selección de herramientas responde a criterios de rendimiento, costo, nivel de experticia del personal y aprovechamiento de la infraestructura existente.

En la empresa, las herramientas Python, KNIME, Power BI y SQL Server ya forman parte del ecosistema operativo, lo que facilita su integración en las fases del modelo CRISP-DM:

- Comprensión del negocio y los datos: KNIME y SQL Server permiten explorar las fuentes internas y construir las variables del modelo.
- Preparación de datos: se realiza mediante flujos en KNIME y consultas SQL, con transformaciones reproducibles.
- Modelado y evaluación: se desarrollan en Python, utilizando técnicas de aprendizaje supervisado (Logistic Regression, Random Forest, XGBoost).

- Despliegue y comunicación: se ejecutan en Power BI, donde los resultados se actualizan automáticamente a partir de la base de datos institucional.

Si bien R, Tableau, PostgreSQL u Oracle presentan fortalezas técnicas en ciertos contextos, su adopción implicaría costos adicionales, mayor curva de aprendizaje y menor compatibilidad con la arquitectura actual.

Por tanto, la elección final no responde a una preferencia tecnológica, sino a una evaluación objetiva de factibilidad y alineación institucional. Las herramientas seleccionadas maximizan la eficiencia analítica, aseguran la trazabilidad del proceso y aprovechan plenamente las capacidades instaladas y la experiencia técnica del equipo, garantizando un flujo metodológico sostenible y replicable.

2.10 MARCO LEGAL

2.10.1 MARCO LEGAL NACIONAL

El desarrollo del presente estudio se sustenta en la normativa hondureña vigente que regula el uso, tratamiento y resguardo de datos personales dentro del ámbito empresarial, especialmente aquellos relacionados con información financiera y tecnológica.

En este sentido, se consideran las disposiciones establecidas por la Comisión Nacional de Bancos y Seguros (CNBS), la Ley de Instituciones del Sistema Financiero, la Circular CNBS No. 025/2022 sobre Gestión de Tecnologías de Información, así como las garantías constitucionales vinculadas al derecho al habeas data y la protección de la información privada (Comisión Nacional de Bancos y Seguros, 2022).

Las instituciones supervisadas deben mantener la confidencialidad de la información de sus clientes y garantizar la integridad, disponibilidad y seguridad de los datos tratados en sus sistemas informáticos. Estas disposiciones se complementan con los lineamientos sobre conducta de mercado y transparencia, que obligan a tratar la información del usuario financiero con responsabilidad y dentro de los límites definidos por la finalidad del servicio o producto contratado (Comisión Nacional de Bancos y Seguros, 2022).

Desde el punto de vista metodológico, estas disposiciones influyeron directamente en el diseño del estudio, al exigir que los datos utilizados para el entrenamiento y validación de los

modelos fueran anonimizados previamente, así como en la implementación de controles de acceso, trazabilidad y segregación de entornos analíticos.

Asimismo, el marco jurídico hondureño reconoce en el artículo 182 de la Constitución de la República el principio del habeas data, que otorga a las personas el derecho de acceder, rectificar o suprimir la información que sobre ellas se conserve en registros públicos o privados. Si bien Honduras aún no cuenta con una ley específica de protección de datos personales, la CNBS ha emitido lineamientos técnicos y de seguridad que exigen aplicar controles de acceso, trazabilidad, gestión de incidentes y políticas de seguridad de la información, con el fin de prevenir filtraciones o tratamientos indebidos de datos (Comisión Nacional de Bancos y Seguros, 2022).

Este principio constitucional se traduce, en términos metodológicos, en la eliminación de cualquier identificador directo o indirecto dentro del dataset analizado, garantizando que el modelo predictivo opere exclusivamente sobre información anonimizada y no permita la reidentificación de los clientes.

El secreto bancario, regulado por la Ley de Instituciones del Sistema Financiero (Decreto No. 170-95 y sus reformas), establece que las entidades financieras están sujetas a deberes de reserva y solo pueden divulgar información con autorización expresa del cliente o mandato judicial (Diario Oficial La Gaceta, 2004). La confidencialidad bancaria en Honduras constituye un derecho fundamental, protegido por el marco legal y reforzado por las regulaciones de la CNBS.

En consecuencia, el alcance del estudio se limita estrictamente al análisis interno con fines académicos y analíticos, lo cual condicionó tanto la selección de variables como la forma agregada en que se presentan los resultados del modelo.

En el contexto de esta investigación, los datos utilizados corresponden a registros internos de la empresa que reflejan la aceptación o el rechazo de ofertas de productos financieros. Por tratarse de información sensible, se aplicó un proceso de anonimización y pseudonimización previo al análisis, con el propósito de eliminar cualquier elemento que pueda permitir la identificación directa o indirecta de las personas involucradas.

Los nombres, direcciones o números de identidad fueron reemplazados por códigos aleatorios irreversibles, y las variables numéricas se agruparon en intervalos (por ejemplo, rangos de edad o tramos de ingresos). Con eso se reduce el riesgo de reidentificación, siguiendo las

recomendaciones de la (Comisión Nacional de Bancos y Seguros, 2022) sobre protección de datos.

Esta decisión metodológica permite cumplir con los principios de confidencialidad y minimización de riesgos legales, sin afectar la validez estadística ni la capacidad predictiva de los modelos desarrollados.

Los datos originales siguen guardados en Microsoft SQL Server, con acceso restringido y supervisado por el área de Tecnología y Seguridad. El equipo analítico solo trabaja con una copia anonimizada en KNIME y Python, cumpliendo los principios de confidencialidad y uso limitado. Además, se registran los accesos y las versiones de los archivos, de modo que todo el proceso sea trazable.

Asimismo, el proceso incluye mecanismos de auditoría y trazabilidad, tales como el registro de accesos, el control de versiones de los datasets y la documentación de los experimentos analíticos. Estas medidas garantizan la integridad de los datos y se ajustan a las exigencias de la Circular CNBS No. 025/2022 en materia de gobierno de tecnologías de información.

Una vez finalizado el proyecto, los datos anonimizados serán conservados únicamente durante el periodo necesario para la validación de los resultados, tras lo cual serán eliminados de manera segura, siguiendo los procedimientos internos de la empresa y las buenas prácticas internacionales de gestión de información.

2.10.2 MARCO LEGAL INTERNACIONAL

Aunque el marco nacional ofrece una base sólida de protección y control, resulta pertinente considerar también estándares internacionales que establecen principios éticos y técnicos en el tratamiento de datos personales.

El Reglamento General de Protección de Datos (GDPR) de la (Regulation (EU) 2016/679 of the European Parliament, 2016) se reconoce como la norma de referencia global en materia de privacidad y seguridad de la información. Este reglamento establece principios universales aplicables a cualquier tratamiento de datos personales, tales como la licitud, la lealtad y transparencia, la limitación de la finalidad, la minimización de los datos, la exactitud, la limitación del almacenamiento, la integridad, la confidencialidad y la responsabilidad proactiva del responsable del tratamiento.

Aunque Honduras no adopta el GDPR formalmente, su estructura y filosofía orientan las

buenas prácticas de las instituciones que manejan información sensible, especialmente en entornos financieros y tecnológicos.

De manera complementaria, la norma ISO/IEC 27001:2022 sobre gestión de la seguridad de la información proporciona un marco técnico para establecer, implementar y mantener sistemas de seguridad basados en controles de acceso, clasificación de datos y respuesta a incidentes (International Organization for Standardization & International Electrotechnical Commission, 2022).

Esta norma ha sido adoptada por diversas instituciones en América Latina como guía para auditar y certificar procesos tecnológicos seguros, y constituye una referencia técnica útil para los procedimientos implementados por la empresa en el resguardo y trazabilidad de la información analizada.

En el diseño metodológico del estudio, estos principios se reflejan en la minimización de variables utilizadas, la documentación del proceso analítico y la adopción de prácticas de trazabilidad y explicabilidad del modelo.

Asimismo, organismos internacionales como la Organización para la Cooperación y el Desarrollo Económicos (OECD, 2022) y el Banco Interamericano de Desarrollo (Banco Interamericano de Desarrollo, 2024) promueven marcos de gobernanza de datos basados en la transparencia, la rendición de cuentas y la equidad en el tratamiento de la información.

Adicionalmente, los Acuerdos de Basilea III establecen principios orientados a fortalecer la gestión del riesgo de crédito, la calidad de los modelos internos y la estabilidad del sistema financiero, enfatizando la necesidad de contar con procesos robustos de medición, control y gobierno de los riesgos (Bank for International Settlements, 2017).

Si bien el modelo desarrollado en esta investigación no constituye un modelo regulatorio de cálculo de capital, su diseño se alinea con los principios de Basilea III en cuanto a la utilización responsable de modelos internos, la validación de su desempeño y la mitigación de sesgos que puedan afectar la toma de decisiones crediticias o comerciales.

En este sentido, el uso de técnicas de machine learning para priorizar clientes con mayor probabilidad de aceptación contribuye a una gestión más eficiente del riesgo de crédito, al mejorar la asignación de los recursos comerciales y reducir exposiciones innecesarias, manteniendo la

coherencia con los principios establecidos por la regulación prudencial.

Estos lineamientos refuerzan la incorporación de criterios éticos dentro del diseño metodológico, asegurando que el uso de modelos predictivos no solo sea técnicamente válido, sino también socialmente responsable.

En consecuencia, los proyectos de analítica de datos en el sector financiero deben integrar criterios éticos, sociales y regulatorios junto con los técnicos, como parte del compromiso institucional con la protección del usuario financiero.

2.10.3 RIESGO DE MODELO Y SUPERVISIÓN PRUDENCIAL

Desde la perspectiva de la supervisión financiera, el uso de modelos analíticos y predictivos en entidades bancarias se encuentra estrechamente vinculado a la gestión del riesgo de modelo, entendido como la posibilidad de que un modelo genere resultados inadecuados debido a errores en su diseño conceptual, supuestos estadísticos, calidad de los datos de entrada o deterioro de su desempeño a lo largo del tiempo.

En el contexto hondureño, la gestión del riesgo de modelo se enmarca dentro de la Resolución CNBS No. 662/2020 sobre Gestión Integral de Riesgos, la cual establece que las instituciones financieras deben implementar mecanismos formales de documentación, validación, monitoreo y revisión periódica de los modelos utilizados para apoyar decisiones comerciales, operativas o de riesgo. Esta normativa busca asegurar que los modelos internos sean consistentes, auditables y alineados con los objetivos estratégicos y el apetito de riesgo de la entidad supervisada.

En coherencia con estos lineamientos prudenciales, el modelo predictivo desarrollado en el presente estudio incorpora métricas objetivas de desempeño tales como el área bajo la curva ROC (AUC), precisión y recall como mecanismos de validación cuantitativa de su capacidad predictiva. Asimismo, se plantea la necesidad de establecer procesos de seguimiento y recalibración periódica, con el fin de mitigar el riesgo de degradación del modelo derivado de cambios en el comportamiento de los clientes o en las condiciones del entorno operativo.

Adicionalmente, el diseño metodológico del modelo prioriza el uso de variables conductuales y operativas, excluyendo variables sensibles que puedan introducir sesgos discriminatorios en la toma de decisiones comerciales. Este enfoque contribuye a reducir el riesgo

de resultados inequitativos y se alinea con los principios de trato justo al consumidor financiero promovidos por la regulación prudencial y las buenas prácticas de gobernanza de modelos.

Si bien el modelo desarrollado en esta investigación no constituye un modelo regulatorio para el cálculo de capital ni para la estimación formal de riesgo crediticio, su diseño y uso se alinean con los principios de Basilea III en cuanto a la utilización responsable de modelos internos, la necesidad de validación independiente y la gestión prudente de los riesgos derivados de su aplicación. En este sentido, la priorización predictiva de clientes con mayor probabilidad de aceptación contribuye indirectamente a una gestión más eficiente del riesgo, al optimizar la asignación de recursos comerciales y reducir exposiciones innecesarias, siempre dentro de los límites establecidos por la regulación vigente.

2.10.4 SÍNTESIS MARCO LEGAL

En síntesis, el marco legal aplicable al presente estudio asegura que todo el proceso de tratamiento y análisis de los datos se desarrolle dentro de un entorno regulado, ético y transparente. El cumplimiento de los principios de confidencialidad, seguridad de la información y trazabilidad garantiza que la información utilizada no vulnere la privacidad de los clientes ni comprometa la integridad de los registros institucionales analizados.

A nivel nacional, las disposiciones emitidas por la Comisión Nacional de Bancos y Seguros, la Ley de Instituciones del Sistema Financiero y las garantías constitucionales vinculadas al derecho al habeas data establecen las bases para el uso responsable de información financiera, condicionando el diseño metodológico del estudio a la anonimización de los datos, la limitación de su uso y la protección del secreto bancario.

Asimismo, el marco legal incorpora los principios de gestión del riesgo de modelo establecidos por la Comisión Nacional de Bancos y Seguros, particularmente los contenidos en la Resolución CNBS No. 662/2020, asegurando que el uso de modelos predictivos se realice bajo criterios de validación, documentación, trazabilidad y supervisión prudencial. En este sentido, el modelo desarrollado se apoya en métricas objetivas de desempeño y contempla la necesidad de procesos de monitoreo y recalibración periódica para mitigar el riesgo asociado a su uso operativo.

Desde una perspectiva internacional, la referencia a estándares como el Reglamento General de Protección de Datos (GDPR), la norma ISO/IEC 27001 y los lineamientos promovidos

por organismos multilaterales como la OCDE y el Banco Interamericano de Desarrollo refuerza la adopción de buenas prácticas en materia de seguridad de la información, gobernanza de datos y responsabilidad institucional.

De manera complementaria, los principios de Basilea III aportan un marco prudencial que orienta el uso responsable de modelos internos en el sector financiero. Si bien el modelo propuesto no constituye un modelo regulatorio para el cálculo de capital, su diseño metodológico se alinea con dichos principios en cuanto a la validación del desempeño, la mitigación de sesgos y la adecuada gestión de los riesgos derivados de su aplicación.

En conjunto, la articulación de las disposiciones nacionales e internacionales permite enmarcar el desarrollo de esta investigación dentro de los estándares contemporáneos de seguridad de la información, ética en el uso de datos y gobernanza de modelos analíticos. Este enfoque integral refuerza el compromiso institucional con la protección del usuario financiero y asegura que los resultados obtenidos se generen en un entorno legítimo, responsable y conforme a las mejores prácticas internacionales en analítica predictiva aplicada al sector bancario.

CAPÍTULO III. METODOLOGÍA

El presente capítulo expone la estructura metodológica de la investigación, orientada a desarrollar un modelo predictivo basado en técnicas de Machine Learning que permita estimar la probabilidad de aceptación de productos financieros en canales físicos de una entidad hondureña.

En coherencia con los objetivos planteados, se describen los fundamentos que guían la selección del enfoque de investigación, el tipo y diseño del estudio, así como los métodos y procedimientos aplicados para la obtención, depuración y análisis de los datos.

Asimismo, se detalla el proceso de definición de la población y muestra, las técnicas empleadas para garantizar la representatividad de los datos, y los criterios de validez y confiabilidad adoptados.

Finalmente, se presenta la operacionalización de las variables y el plan de análisis de datos que sustenta el desarrollo del modelo predictivo, articulado bajo un marco metodológico reproducible y alineado con los objetivos de investigación.

3.1 CONGRUENCIA METODOLÓGICA

Tabla 5 Congruencia Metodológica

Objetivos	Pregunta de investigación	Hipótesis
Identificar e identificar las n variables de comportamientos de los clientes con mayor peso predictivos (evaluado mediante técnicas de análisis exploratorio de datos EDA) sobre la disposición de adquirir producto financiero mediante datos históricos de los últimos 3 años de ventas.	En la P (Población) actual de clientes de la entidad con historial de interacción, ¿Cuál es la O (Relación y peso predictivo) de la I (Intervención), es decir, las variables históricas del cliente (como qué producto aceptó en el pasado, la frecuencia de uso de esta y en qué comercios lo utiliza) C (Comparación) con el uso exclusivo de variables básicas que están estrechamente relacionadas con la disposición de adquirir un producto financiero?	Las variables históricas del cliente (producto aceptado, frecuencia, tipo de uso) tienen mayor peso predictivo sobre la aceptación que las variables básicas.
Entrenar y comparar al menos 3 ML de clasificación (Regresión logística, árbol de decisiones, Random Forest, etc.) y seleccionar el modelo con mayor	Para la P (Población) de clientes, ¿Qué I (intervención), es decir, qué tipo de modelo de ML (Regresión logística, árbol de decisiones, Random Forest, etc.), ofrece el O (Mejor desempeño de predicción, Medido por métricas	Uno de los modelos de Machine Learning alcanzará un desempeño superior (AUC ≥ 70%)

Objetivos	Pregunta de investigación	Hipótesis
desempeño predictivo (alcanzando el F1-Score más alto y un AUC del 70%), en la fase de validación.	de efectividad), ¿en C (comparación) con el método tradicional actual de la entidad?	frente al método tradicional.
Aplicar el modelo seleccionado para estimar la probabilidad de aceptación individual y clasificar a los clientes entre 3 a 5 segmentos (desde la aceptación de al menos un producto hasta 6 productos), detallando el consumo y las características claves con probabilidad superior del 55 por ciento para estrategia de ventas cruzadas.	En la P (Población), ¿Cuáles son las O (Características y patrones de comportamiento) que definen a los segmentos con mayor probabilidad de aceptar, después de la I (intervención) de aplicar el modelo de ML que mejor desempeño obtuvo, en C (comparación) con los segmentos de clientes de baja posibilidad?	Los clientes con características y patrones homogéneos presentan mayor probabilidad de aceptación (≥55%).
Evaluar y cuantificar el impacto del modelo predictivo en la tasa de aceptación con un aumento mínimo del 25 por ciento y una reducción de tiempo de barrida de base del 40 por ciento, mediante simulaciones comparativas de las tasas actuales.	¿Cuál es el O (Impacto cuantificable, en tasa de aceptación y eficiencia operativa) en los P (Procesos de recomendación) de la entidad, generado por I (Intervención) por el uso continuo del modelo predictivo, en C (Comparación) con el enfoque actual de uso de métodos tradicional?	La implementación del modelo predictivo aumentará la tasa de aceptación (≥25%) y reducirá el tiempo operativo (≥40%) respecto al método actual.

Fuente: Elaboración propia

3.1.1 ESQUEMA DE VARIABLES DE ESTUDIO

Esta sección describe las variables clave empleadas en el estudio y su organización conceptual. Se parte de una variable dependiente binaria (la aceptación de un producto financiero por parte del cliente) y un conjunto de variables independientes agrupadas por categoría analítica (financiera, demográfica, conductual y de perfil).

La tabla siguiente resume las variables seleccionadas, sus características clave y la relación teórica postulada con la variable dependiente:

Tabla 6 Esquema de Variable

Categoría	Variables Clave Incluidas	Relación Postulada
Historial Financiero y Uso de Productos	Préstamos (saldo y cantidad), Cuentas (ahorro, cheques – saldo y cantidad), Seguros (si tiene o no), Financiamientos (saldo y cantidad), Pensiones (si aplica).	El historial y la profundidad de la relación financiera actual (tenencia de productos) influyen en la necesidad y capacidad de adquirir nuevos productos.
Perfil de Cliente y Antigüedad	Antigüedad del cliente con el banco, Tipo de cliente (nuevo o recurrente), Tarjetas de crédito y límite disponible, Salario mensual.	El Salario y el Límite de Crédito determinan la capacidad económica. La Antigüedad refleja la lealtad y el riesgo percibido por el banco.
Demografía y Ubicación	Edad, sexo, género, estado civil, Municipio y departamento de residencia.	Las características socioeconómicas y la ubicación geográfica influyen en los patrones de consumo y en las necesidades financieras específicas.
Interacción y Preferencias Digitales	Tiene celular, correo, banca en línea, Preferencia por contacto telefónico.	El uso de canales digitales y la preferencia de contacto influyen en la receptividad a ofertas y en la facilidad para acceder a la información del producto.

Fuente: Elaboración propia

A continuación, se presenta un esquema visual que sintetiza la relación entre las variables independientes y la variable dependiente del estudio:

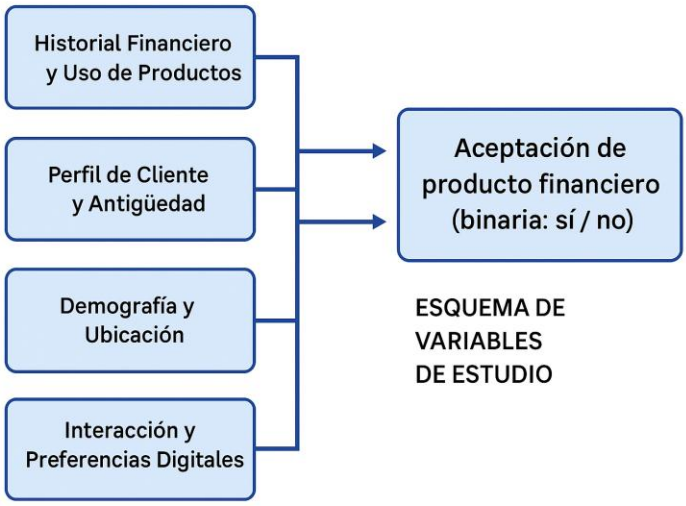

Ilustración 3 Esquema de relación entre variables y aceptación del producto financiero
Fuente: Elaboración propia

3.1.2 OPERACIONALIZACIÓN DE LAS VARIABLES

Tabla *7* Descripción de variables

Variable	Definición operativa	Tipo de dato	Indicador / Descripción precisa	Codificación
Aceptación del producto	Determina si el cliente aceptó la oferta financiera.	Categórica binaria	Resultado del registro de oferta: aceptada o no aceptada.	1 = Sí, 0 = No
Antigüedad del cliente	Tiempo transcurrido desde la primera relación activa con el banco.	Numérica continua	Número de meses desde la fecha de alta en el sistema hasta la fecha de análisis.	Valor numérico (meses)
Tipo de producto ofrecido	Tipo de producto financiero presentado al cliente.	Categórica nominal	Categoría del producto (tarjeta, préstamo, seguro, etc.).	One-hot encoding
Tipo de cliente	Clasifica si el cliente es nuevo o recurrente.	Categórica nominal	Estado del cliente al momento de la oferta.	1 = Nuevo, 0 = Recurrente
Saldo de préstamos	Monto total de préstamos activos.	Numérica continua	Suma de saldos vigentes en préstamos (Lempiras).	Valor numérico (L.)

Variable	Definición operativa	Tipo de dato	Indicador / Descripción precisa	Codificación
Cantidad de préstamos	Número total de préstamos activos.	Numérica discreta	Conteo de préstamos vigentes registrados en sistema.	Valor entero
Saldo en cuentas	Total de fondos en cuentas de ahorro y cheques.	Numérica continua	Suma de saldos activos en cuentas.	Valor numérico (L.)
Cantidad de cuentas	Total de cuentas activas por cliente.	Numérica discreta	Número de cuentas de ahorro o corriente activas.	Valor entero
Tiene seguros	Indica si el cliente posee seguros con el banco.	Categórica binaria	Existencia de al menos un seguro activo.	1 = Sí, 0 = No
Tiene tarjeta de crédito	Indica si el cliente posee tarjeta de crédito.	Categórica binaria	Registro de tarjeta activa.	1 = Sí, 0 = No
Límite de tarjeta	Límite de crédito disponible en tarjetas.	Numérica continua	Monto total de límite de crédito (Lempiras).	Valor numérico (L.)
Financiamiento (saldo)	Total de financiamientos activos.	Numérica continua	Suma de saldos de financiamientos.	Valor numérico (L.)
Financiamiento (cantidad)	Número de financiamientos activos.	Numérica discreta	Conteo de financiamientos vigentes.	Valor entero
Tiene pensión	Indica si el cliente recibe pensión.	Categórica binaria	Presencia de pensión activa registrada.	1 = Sí, 0 = No
Departamento	Departamento de residencia del cliente.	Categórica nominal	Ubicación geográfica del cliente.	Codificación de texto
Municipio	Municipio de residencia del cliente.	Categórica nominal	Detalle territorial del cliente.	Codificación de texto
Edad	Edad actual del cliente.	Numérica continua	Diferencia entre fecha de nacimiento y fecha de análisis.	Valor numérico (años)

Variable	Definición operativa	Tipo de dato	Indicador / Descripción precisa	Codificación
Sexo	Sexo biológico registrado.	Categórica binaria	Masculino o Femenino.	1 = Masculino, 0 = Femenino
Género	Género reportado por el cliente.	Categórica nominal	Valor declarado en sistema.	Codificación nominal
Estado civil	Estado civil registrado.	Categórica nominal	Casado, soltero, divorciado, viudo, etc.	One-hot encoding
Salario mensual	Ingreso mensual del cliente.	Numérica continua	Monto de ingreso reportado (Lempiras).	Valor numérico (L.)
Tiene celular	Indica si posee teléfono móvil registrado.	Categórica binaria	Presencia de número celular en sistema.	1 = Sí, 0 = No
Tiene correo electrónico	Indica si posee correo electrónico registrado.	Categórica binaria	Presencia de dirección de correo válida.	1 = Sí, 0 = No
Usa banca en línea	Indica si utiliza servicios de banca electrónica.	Categórica binaria	Uso activo de la plataforma digital.	1 = Sí, 0 = No
Desea contacto telefónico	Preferencia de ser contactado por teléfono.	Categórica binaria	Preferencia registrada para contacto telefónico.	1 = Sí, 0 = No

Fuente: Elaboración propia

3.1.3 HIPÓTESIS

En el contexto del presente estudio, se plantean las siguientes hipótesis de investigación:

- Hipótesis nula (H0): Las variables del cliente no tienen un efecto significativo en la aceptación de ofertas de productos financieros realizadas en canales físicos.
- Hipótesis alternativa (H1): Las variables del cliente sí tienen un efecto significativo en la aceptación de ofertas de productos financieros realizadas en canales físicos.

La contrastación empírica de las hipótesis se realizará mediante la evaluación del desempeño de los modelos predictivos.

Si el modelo demuestra una capacidad explicativa superior al azar (AUC ≥ 0.70) y las variables seleccionadas presentan importancia estadística o predictiva dentro del modelo, se

considerará evidencia suficiente para rechazar la hipótesis nula (H_0) y aceptar la hipótesis alternativa (H_1).

3.2 ENFOQUE Y MÉTODOS

3.2.1 ENFOQUE DE INVESTIGACIÓN

El presente estudio adopta un enfoque cuantitativo-predictivo y aplicado, en coherencia directa con el objetivo general de diseñar y validar un modelo de machine learning que estime la probabilidad de aceptación de ofertas financieras y alcance un desempeño mínimo del 70 % en la métrica AUC. Este enfoque resulta el más adecuado porque permite cuantificar relaciones entre variables, medir el poder explicativo de los modelos y evaluar métricas objetivas de predicción (AUC, F1-score, precisión y recall), asegurando la validez empírica de los resultados.

La naturaleza cuantitativa del estudio responde a la necesidad de analizar grandes volúmenes de datos históricos de clientes, transacciones y productos financieros registrados en los canales físicos de una entidad bancaria hondureña. A través de este enfoque se busca identificar patrones de comportamiento, comparar algoritmos predictivos y evaluar su impacto potencial en la aceptación de productos financieros, con base en técnicas estadísticas, programación y modelado supervisado (Hernández Sampieri & Fernandez-Collado, 2014).

3.2.2 MÉTODOS DE INVESTIGACIÓN

La investigación se estructura bajo los principios del ciclo CRISP-DM (Cross-Industry Standard Process for Data Mining), adaptado al contexto bancario hondureño. Este método, ampliamente utilizado en proyectos de analítica avanzada, organiza el proceso en seis fases: comprensión del negocio, comprensión de los datos, preparación, modelado, evaluación y despliegue (Espinosa Zúñiga, 2020).

Su aplicación permite mantener una trazabilidad metodológica alineada con los objetivos específicos del estudio:

- identificar las variables con mayor peso predictivo mediante análisis exploratorio (Fase I),
- entrenar y comparar modelos de clasificación (Fase II),
- aplicar el modelo óptimo para segmentar clientes (Fase III), y

- evaluar cuantitativamente su impacto sobre la tasa de aceptación de productos financieros (Fase IV).

El uso del método CRISP-DM garantiza la reproducibilidad y validez del proceso analítico, permitiendo integrar de manera coherente la teoría estadística, las técnicas de machine learning y las necesidades estratégicas de la entidad.

3.3 DISEÑO DE LA INVESTIGACIÓN

El presente estudio adopta un diseño no experimental, transversal y correlacional-predictivo, adecuado para investigaciones en las que no se manipulan deliberadamente las variables independientes y se trabaja con datos observacionales ya existentes. En este caso, se emplean registros administrativos históricos provenientes de los sistemas institucionales de una entidad bancaria, sin que el investigador ejerza control directo sobre las condiciones en que dichos datos fueron generados.

El carácter no experimental obedece a que las variables se analizan tal como se presentan en la realidad operativa de la institución, sin intervención alguna sobre los procesos comerciales o el comportamiento de los clientes.

El diseño se considera transversal porque, aunque los datos provienen de un periodo histórico comprendido entre 2023 y 2025, el análisis se realiza sobre una base consolidada tratada como un único corte analítico, sin evaluar la evolución temporal de las variables ni realizar comparaciones entre distintos momentos en el tiempo. Es decir, la dimensión temporal no constituye una variable de análisis, sino únicamente el marco de disponibilidad de los registros (Hernández Sampieri & Fernandez-Collado, 2014).

Finalmente, el estudio se clasifica como correlacional-predictivo debido a que busca identificar asociaciones estadísticamente significativas entre variables operativas y de comportamiento del cliente, con el propósito de estimar la probabilidad de aceptación de productos financieros mediante técnicas de machine learning. Este enfoque permite generar conocimiento empírico orientado a la toma de decisiones comerciales, sin establecer relaciones causales ni requerir experimentación controlada (Hernández Sampieri & Fernandez-Collado, 2014).

3.3.1 POBLACIÓN

La población objeto de estudio estuvo conformada por la totalidad de interacciones registradas entre clientes y la entidad bancaria, durante el periodo comprendido entre septiembre de 2023 y septiembre de 2025. Este conjunto de registros constituye el universo poblacional de referencia del estudio, a partir del cual se definió posteriormente la población efectiva de análisis conforme a los objetivos de la investigación.

El universo poblacional total estimado para el periodo de estudio se presenta en la Tabla 8, el cual asciende a 12,020,705 interacciones, distribuidas temporalmente en 2,000,751 registros correspondientes al periodo septiembre–diciembre de 2023, 6,114,390 registros al periodo enero–diciembre de 2024 y 3,905,564 registros al periodo enero–septiembre de 2025:

Tabla 8 Población total estimada por año (sep-2023 a sep-2025)

Año	Total (N)
2023 (sep–dic)	2,000,751
2024 (ene–dic)	6,114,390
2025 (ene–sep)	3,905,564
Total	12,020,705

Fuente: Elaboración propia

Nota aclaratoria de fuente:

La estimación de la población total proviene del registro administrativo interno de la entidad bancaria, específicamente del sistema de gestión comercial que consolida las operaciones efectuadas en los canales físicos y digitales durante el periodo de análisis. Estos datos fueron extraídos directamente del Data Warehouse institucional y corresponden al total de interacciones registradas en dicho periodo, verificadas por el área de inteligencia comercial de la entidad.

3.3.2 MUESTRA

En coherencia con el enfoque metodológico adoptado, el presente estudio no aplicó un proceso de muestreo probabilístico, sino que trabajó con una población dinámica temporal, definida como la totalidad de los registros válidos disponibles correspondientes al fenómeno

analizado durante el periodo 2023–2025.

Dicha población estuvo conformada por registros de campañas comerciales ejecutadas en canales físicos durante el periodo 2023–2025, seleccionados a partir de criterios de elegibilidad definidos a priori.

Los criterios de inclusión consideraron únicamente aquellos registros que:

a) correspondían a campañas comerciales ejecutadas en canales físicos;

b) contaban con información completa en las variables predictoras definidas; y

c) presentaban una variable objetivo observable que permitiera identificar la aceptación o rechazo de la oferta.

De forma complementaria, se excluyeron aquellos registros que no permitían evaluar el resultado de la interacción comercial. Tras la aplicación de estos criterios y los procesos de depuración correspondientes, la población dinámica temporal utilizada para el análisis totalizó 14,226 observaciones.

Justificación del uso de población dinámica temporal en Big Data

Este enfoque es consistente con la literatura de Big Data y machine learning aplicado, la cual señala que, cuando se dispone de grandes volúmenes de datos históricos y acceso completo al fenómeno de interés, el uso de la totalidad de los registros válidos disponibles resulta metodológicamente más adecuado que el muestreo probabilístico tradicional, al permitir capturar patrones complejos y mejorar el desempeño predictivo de los modelos (Provost & Fawcett, 2013).

3.3.3 TÉCNICAS DE MUESTREO

El presente estudio empleó un muestreo no probabilístico de tipo intencional, con alcance censal sobre la población efectiva de análisis. La selección de los casos se realizó mediante la aplicación de criterios de elegibilidad definidos a priori, propios de investigaciones basadas en registros administrativos y analítica predictiva.

En consecuencia, no se aplicaron procedimientos de selección aleatoria ni estimación de probabilidades de inclusión, utilizándose la totalidad de los registros válidos disponibles que

cumplían con los criterios establecidos. Este enfoque es consistente con estudios de Big Data y machine learning, donde el objetivo principal es el entrenamiento y validación de modelos predictivos, más que la inferencia estadística a una población teórica mayor.

3.3.4 TÉCNICAS E INSTRUMENTOS

En la presente investigación no se utilizaron instrumentos de recolección de datos primarios, debido a que el estudio se fundamenta exclusivamente en el análisis de información secundaria proveniente de registros administrativos institucionales.

En consecuencia, no se aplicaron encuestas, cuestionarios ni entrevistas. La información analizada fue obtenida del data warehouse corporativo del banco, el cual consolida datos históricos relacionados con campañas comerciales ejecutadas en canales físicos, características del cliente y resultados de las interacciones comerciales.

3.4 FUENTES DE INFORMACIÓN

La investigación se apoya en un conjunto de fuentes de información cuidadosamente seleccionadas para garantizar la validez técnica, científica y contextual del modelo predictivo propuesto.

3.4.1 DATOS INSTITUCIONALES (FUENTES SECUNDARIAS INTERNAS)

La fuente principal de información para esta investigación proviene de los registros institucionales extraídos del data warehouse del banco participante. Desde el punto de vista metodológico, dichos registros se clasifican como fuentes secundarias internas, dado que corresponden a datos preexistentes generados por la institución en el curso normal de sus operaciones y no fueron recolectados específicamente para el desarrollo del presente estudio (Hernández Sampieri & Fernandez-Collado, 2014).

El conjunto de datos está compuesto por información estructurada sobre interacciones comerciales realizadas en los canales entre septiembre de 2023 y septiembre de 2025. Asimismo, incluye variables demográficas, financieras, históricas y de comportamiento asociadas a los clientes, lo que proporciona una base empírica sólida para el análisis de patrones de aceptación y la construcción de modelos predictivos.

En términos prácticos, estos registros constituyen el insumo empírico sobre el cual se

aplican los procesos de limpieza, transformación y modelado estadístico. No se trata de datos obtenidos mediante instrumentos de recolección primaria, como encuestas o experimentos, sino de observaciones reales derivadas de la operación cotidiana de la entidad financiera, lo que les confiere un alto valor analítico.

Dentro del enfoque cuantitativo, especialmente en estudios orientados a la analítica predictiva, el uso de fuentes secundarias internas permite realizar análisis objetivos, verificables y reproducibles, sin que el investigador intervenga en la generación de los datos (Hernández Sampieri & Fernandez-Collado, 2014). En este sentido, la información surge de la dinámica operativa del banco, pero se transforma en evidencia científica útil para comprender y anticipar comportamientos financieros.

3.4.2 FUENTES SECUNDARIAS

Las fuentes secundarias complementarias utilizadas se agrupan en dos categorías:

1. Literatura académica y científica.

   Se consultó bibliografía nacional e internacional que documenta el uso de técnicas de machine learning en contextos bancarios y de marketing científico, incluyendo artículos en revistas especializadas, libros, tesis y estudios de caso relevantes.

   Estas fuentes sustentan el diseño metodológico, la selección de técnicas de análisis y la interpretación de los resultados, permitiendo contrastar el modelo propuesto con investigaciones previas.

2. Fuentes normativas y regulatorias.

   Se revisaron documentos emitidos por el Banco Central de Honduras (BCH), la Comisión Nacional de Bancos y Seguros (CNBS) y organismos multilaterales, con el fin de alinear la investigación con los principios regulatorios y de protección de datos aplicables al sistema financiero hondureño, así como con las políticas de digitalización y modernización tecnológica del sector.

Ambas categorías fortalecen el marco teórico, contextual y de validez externa del estudio, aportando soporte conceptual al diseño metodológico y a la interpretación de los resultados (Hernández Sampieri & Fernandez-Collado, 2014).

3.5 PLAN DE ANÁLISIS DE DATOS

El análisis de datos se desarrollará en tres fases secuenciales orientadas al cumplimiento de los objetivos específicos del estudio y a la contrastación de las hipótesis planteadas.

Cada fase integra técnicas de estadística descriptiva, minería de datos y aprendizaje automático, aplicadas sobre un conjunto muestral de 14,226 observaciones correspondientes al período 2023–2025.

El propósito general del análisis es verificar empíricamente si las variables del cliente tienen un efecto significativo en la aceptación de ofertas de productos financieros en canales físicos, mediante la aplicación de modelos predictivos y métricas de desempeño (AUC, F1-score, precisión y recall).

Para garantizar la trazabilidad y el control metodológico del proceso, se elaboró un Diagrama de Gantt (Anexo 1) que presenta la ruta crítica, los tiempos estimados de ejecución y las dependencias entre fases.

Asimismo, se incluye un Diccionario de Datos (Anexo Y) que documenta las variables utilizadas, su definición, tipo, codificación, fuente institucional y tratamiento durante la limpieza.

3.5.1 FASE I: ANÁLISIS EXPLORATORIO Y DESCRIPTIVO

Propósito: Cumplir el Objetivo Específico 1, consistente en identificar las variables de comportamiento y características del cliente con mayor peso predictivo sobre la disposición de adquirir productos financieros, mediante técnicas de análisis exploratorio de datos (Exploratory Data Analysis, EDA).

Técnicas aplicadas:

- Limpieza y transformación: Uso de Pandas para imputación o eliminación de valores faltantes, tratamiento de outliers y normalización de variables numéricas.
- Análisis descriptivo y visual: Aplicación de Matplotlib y Seaborn para examinar distribuciones y relaciones entre variables.
- Selección de variables relevantes: Cálculo de matrices de correlación y aplicación de métodos de feature selection (VIF, importancia de características) para determinar las variables con mayor peso predictivo.

Producto de la fase: Dataset limpio, normalizado y documentado, junto con el informe de análisis exploratorio y justificación técnica de las variables seleccionadas.

3.5.2 FASE II: MODELADO PREDICTIVO Y COMPARACIÓN

Propósito: Cumplir el Objetivo Específico 2, orientado a entrenar y comparar al menos tres modelos de clasificación (Regresión Logística, Árbol de Decisión y Random Forest), evaluando su desempeño mediante métricas de precisión, F1-score y AUC.

Técnicas aplicadas:

- Entrenamiento de modelos: Uso de Scikit-learn y XGBoost para desarrollar los modelos y ajustar hiperparámetros.
- Validación cruzada: Aplicación de la técnica k-fold cross-validation para evitar sobreajuste (overfitting) y asegurar la robustez del modelo.
- Evaluación y comparación: Se medirán métricas clave (AUC, F1-score, precisión y recall). Se seleccionará el modelo que alcance al menos un AUC ≥ 0.70 y el mayor F1-score, cumpliendo con el estándar de desempeño definido.

Producto de la fase: Modelos validados, métricas comparativas y documentación técnica del proceso de entrenamiento y evaluación.

3.5.3 FASE III: SEGMENTACIÓN E IMPACTO DEL MODELO

Propósito: Cumplir los Objetivos Específicos 3 y 4, orientados a segmentar los clientes por probabilidad de aceptación y evaluar el impacto potencial del modelo predictivo en la eficiencia operativa y la efectividad de la estrategia comercial.

Técnicas aplicadas:

- Segmentación por probabilidad: Aplicación del modelo final al dataset completo para clasificar clientes en tres a cinco segmentos, según la probabilidad estimada de aceptación ($\geq$55 %).
- Interpretación del modelo: Uso de SHAP u otras técnicas de explicabilidad para identificar las variables más influyentes en los segmentos de alta probabilidad.

- Simulación de impacto: Ejecución de escenarios comparativos que proyecten un aumento mínimo del 25 % en la tasa de aceptación y una reducción del 40 % en el tiempo de barrido de base, en comparación con los resultados históricos.

Producto de la fase: Informe de segmentación, análisis interpretativo del modelo y simulación de impacto cuantificable en la gestión comercial.

3.5.4 HERRAMIENTAS DE GESTIÓN Y CONTROL

Para el seguimiento y trazabilidad del proceso analítico se emplearán las siguientes herramientas:

- Diagrama de Gantt (Anexo 1): planificación temporal de las fases, actividades y entregables.
- Diccionario de Datos (Anexo Y): documentación técnica de las variables utilizadas y su tratamiento.
- Bitácora de control de versiones: registro de cambios y revisiones en el dataset, scripts y resultados.
- Entorno de ejecución: análisis realizado en Python y SQL bajo repositorio controlado.

CAPÍTULO IV. RESULTADOS Y ANÁLISIS

Derivado del diseño muestral bifásico estratificado y de la muestra final de 14,226 observaciones, este capítulo desarrolla el Análisis Exploratorio de Datos (EDA), primera etapa esencial del proceso de ciencia de datos (IBM, 2021).

El EDA constituye un prerrequisito metodológico indispensable, ya que permite evaluar la calidad de la información, identificar patrones preliminares y asegurar que los datos cumplan las condiciones necesarias para su posterior uso en modelos estadísticos y de machine learning. Esta fase garantiza que los resultados que se presenten en los capítulos siguientes sean sólidos y reproducibles.

Asimismo, el EDA actúa como un puente entre el diseño metodológico descrito en el Capítulo III (donde se definieron la población, la muestra y el diseño de muestreo) y el proceso de inferencia predictiva. Su propósito es confirmar que la muestra mantiene coherencia con la población original y con los supuestos del diseño probabilístico aplicado (Särndal, 2003), además de verificar que las variables se encuentren depuradas, codificadas correctamente y libres de sesgos estructurales que puedan afectar la modelización.

Entre las tareas realizadas se incluyen la verificación de consistencia estructural del dataset, el tratamiento de valores atípicos, la imputación de datos categóricos y la revisión de supuestos estadísticos básicos.

Para facilitar la lectura y guiar al lector a través del capítulo, se presenta a continuación la estructura de este apartado:

Estructura del capítulo:

- Descripción general del conjunto de datos: características de la base, origen de la información y definición de las variables.
- Limpieza y preparación de los datos: depuración, corrección de inconsistencias, imputación y codificación.
- Visualización de datos: análisis gráfico inicial para identificar tendencias, distribuciones y relaciones relevantes.
- Conclusiones del EDA: síntesis de hallazgos clave y su relevancia para el modelado

predictivo.

- Proceso operativo de obtención de datos: descripción del trabajo de campo informático, fuentes secundarias institucionales y trazabilidad técnica del dataset.

En conjunto, esta introducción establece el marco conceptual y operativo sobre el cual se desarrollan los resultados del capítulo, asegurando claridad metodológica y trazabilidad para el lector.

4.1 ANALISIS EXPLORATORIO DE DATOS(EDA)

4.1.1 DESCRIPCIÓN GENERAL DEL CONJUNTO DE DATOS

El conjunto de datos analizado corresponde a la muestra final de estudio, conformada por los registros seleccionados mediante el diseño de Muestreo Bifásico Estratificado (MBE) descrito en el Capítulo III.

Esta base integra información histórica de interacciones entre clientes y la entidad hondureña durante el período septiembre de 2023 a septiembre de 2025, permitiendo examinar patrones de aceptación de productos financieros en los canales físico y digital.

El conjunto de datos está constituido por 14,226 observaciones válidas, cada una correspondiente a un cliente individual considerado como unidad de análisis. Los registros fueron consolidados desde los sistemas administrativos y transaccionales de la empresa, posteriormente depurados para eliminar inconsistencias, garantizando coherencia estructural y completitud de los campos.

2.1.4.5 ESTRUCTURA DEL CONJUNTO DE DATOS

El conjunto de datos analítico se organizó con base en una estructura relacional unificada, donde cada registro representa un cliente individual y cada variable corresponde a un atributo medible de su perfil o comportamiento financiero.

La información se agrupó en cuatro categorías principales de variables (dependiente, comportamentales, financieras y demográficas) que reflejan los ejes centrales del estudio: la propensión del cliente a aceptar ofertas y los factores asociados a dicha decisión.

La siguiente tabla resume la composición general del conjunto de datos, especificando la función analítica de cada grupo de variables, su descripción y el tipo de escala utilizada:

Tabla 9 Estructura del conjunto de datos

Variable	Descripción	Categoría	Tipo de Dato	Codificación Clave
PRODUCTO	Producto financiero promocionado en la campaña.	Cualitativa Nominal	Texto	CH (Colocación Cash), CDV (Ciclo de Vida), etc.
RESULTADO_CONTACTO	Resultado del intento de contacto con el cliente.	Cualitativa Nominal	Texto	Contacto / No Contacto
FCHA_ENTREGA	Fecha en que se incluyó al cliente en la campaña.	Cuantitativa Temporal	Fecha	dd/mm/aaaa
CANAL	Medio utilizado para el contacto de la campaña.	Cualitativa Nominal	Texto	Digital / Físico
CUSTOMER_SINCE	Antigüedad del cliente en la institución.	Cuantitativa Temporal	Fecha	dd/mm/aaaa
CLASIFICACION_CLIENTE	Segmentación estratégica del cliente.	Cualitativa Nominal	Texto	BANCA CONSUMO, BANCA MAS, etc.
GENERO	Género del cliente.	Cualitativa Nominal	Texto	F (Femenino) / M (Masculino)
ESTADO_CIVIL	Estado civil del cliente.	Cualitativa Nominal	Texto	SINGLE, MARRIED, etc.
EDUCATION_LEVEL	Nivel educativo máximo alcanzado.	Cualitativa Ordinal	Texto	SECUNDARIA, SUPERIOR, etc.
EDAD	Edad del cliente en años.	Cuantitativa Discreta	Numérico	Rango 25 a 74 años.
DEPARTAMENTO	Ubicación geográfica (Departamento) del cliente.	Cualitativa Nominal	Texto	Nombres de Departamentos (ej. CORTES).

Variable	Descripción	Categoría	Tipo de Dato	Codificación Clave
NUM_PASIVOS	Número de productos pasivos (ej. cuentas de ahorro) del cliente.	Cuantitativa Discreta	Numérico	Conteo de 0 a 11.
SALDO_PASIVOS	Saldo total consolidado en productos pasivos del cliente.	Cuantitativa Continua	Numérico	Expresado en unidad monetaria local.
NUM_PRESTAMOS	Número de préstamos activos (productos activos) del cliente.	Cuantitativa Discreta	Numérico	Conteo de 0 a 8.
FLAG_BANCO_ONLINE	Indicador binario si el cliente usa página bancaria.	Cuantitativa Discreta (Binaria)	0/1	1: Sí, 0: No.
FLAG_ASISTENTE	Indicador binario relacionado con un sistema interno.	Cuantitativa Binaria	0/1	1: Sí, 0: No.
FLG_AFP	Indicador binario si el cliente es pensionado.	Cuantitativa Binaria	0/1	1: Sí, 0: No.
FLAG_ID_TC	Indicador binario si el cliente posee tarjeta de crédito.	Cuantitativa Binaria	0/1	1: Sí, 0: No.
FLAG_ID_SEGURO	Indicador binario si el cliente tiene seguros activos.	Cuantitativa Binaria	0/1	1: Sí, 0: No.
FLAG_REMESA	Indicador binario si el cliente recibe remesas.	Cuantitativa Binaria	0/1	1: Sí, 0: No.

Variable	Descripción	Categoría	Tipo de Dato	Codificación Clave
FLAG_BILLETERA	Indicador binario interno relacionado con productos.	Cuantitativa Binaria	0/1	1: Sí, 0: No.
RIESGOACTUAL	Puntuación de riesgo crediticio actual del cliente.	Cuantitativa Continua	Numérico	Puntaje de 20 a 99.
VENTA	VARIABLE OBJETIVO: Indicador si se logró una venta.	Dependiente, Cuantitativa Binaria	0/1	1: venta Exitosa, 0: No Venta.

Fuente: Elaboración propia

En conjunto, esta estructura permite capturar tanto la dimensión transaccional como la sociodemográfica del cliente, facilitando el análisis de patrones de comportamiento y la identificación de variables con potencial predictivo en el modelo de aceptación de ofertas.

2.1.4.6 CARACTERÍSTICAS GENERALES DE LOS DATOS

- Cobertura temporal: septiembre de 2023 a septiembre de 2025, con cortes trimestrales y anuales para el análisis longitudinal.
- Nivel de agregación: datos individuales de cliente, permitiendo un enfoque micro analítico del comportamiento de aceptación.
- Fuente de información: registros administrativos, bases de transacciones y sistemas de ofertas institucionales.
- Número total de variables analizadas: 12 (1 dependiente y 11 independientes).
- Distribución por canal: aproximadamente 55 % de registros pertenecen al canal digital y 45 % al canal físico, manteniendo la proporción estratificada del diseño muestral.
- identificación de Asimetria y Outliers: La variable SALDO_PASIVO evidencia una gran diferencia entre la media ($3,116.68) y su desviación estándar ($21,240.01) lo cual debe existir un tratamiento adecuado para mejorar el análisis

de esta, esto lo veremos en la tabla 4.2.

A continuación, mostramos un resumen de las estadísticas descriptivas para las variables numéricas clave, identificando las tendencias centrales y de dispersión

Tabla 10 Resumen estadístico

Variable	Media (Promedio)	Desviación Estándar	Mínimo	Máximo
EDAD	43.00	13.42	0	96
SALDO_PASIVOS	3,116.68	21,240.01	0	1,014,458.41
RIESGOACTUAL	68.67	13.34	30	99
VENTA	0.027 (2.7%)	0.16	0	1

Fuente: Elaboración propia

El promedio de ventas exitosas fue de 2.7 % sobre el total de observaciones, mientras que la edad promedio de los clientes fue de 43 años, con una dispersión moderada (DE = 13.42), lo que indica un perfil heterogéneo dentro del rango etario analizado.

2.1.4.7 RELEVANCIA ANALÍTICA DEL CONJUNTO

La estructura de la base de datos permite desarrollar modelos predictivos mediante algoritmos de Machine Learning supervisado, al incluir variables de diversa naturaleza (categóricas, discretas y continuas) y relaciones potencialmente no lineales entre ellas.

Además, la combinación de información demográfica, financiera y comportamental ofrece una visión integral del cliente, lo que favorece la identificación de factores asociados con la aceptación de productos financieros.

El enfoque longitudinal (al cubrir un periodo de dos años) permite también observar variaciones en la propensión de aceptación a lo largo del tiempo, fortaleciendo la validez y estabilidad del modelo.

En conjunto, este conjunto de datos constituye una base analítica robusta, representativa y coherente con el diseño muestral, garantizando resultados confiables en la etapa de modelado predictivo posterior.

4.1.2 LIMPIEZA Y PREPARACIÓN DE LOS DATOS

Esta sección describe las acciones llevadas a cabo para asegurar la calidad, consistencia y confiabilidad del conjunto de datos que se utilizará en las etapas de análisis.

En primer lugar, se consolidó la muestra mediante un proceso de extracción estratificada, tomando como único criterio de estratificación el canal de atención (físico y digital), de acuerdo con lo detallado en el Capítulo III. Este enfoque garantizó la representación proporcional de ambos canales dentro del conjunto muestral, facilitando una comparación metodológicamente válida entre ellos.

Las consultas se realizaron en SQL Server, separadas año por año, lo cual permitió mantener la trazabilidad temporal de los datos y controlar la aleatoriedad de la selección dentro de cada período. Para asegurar que la muestra fuera representativa y no sesgada, se empleó la cláusula ORDER BY NEWID() en cada consulta, generando aleatoriedad controlada en la selección de registros para cada canal y año. De esta forma, se fue consolidando el cubo de información sobre el cual se aplicaron los procesos de limpieza.

La limpieza de datos representa una fase crítica del análisis, ya que permite eliminar errores que podrían comprometer la validez de los resultados (Costanzo, 2023; Hair et al., 2019). Una base limpia previene errores que podrían distorsionar los resultados, mientras que (Papageorgiou et al., 2018) enfatiza que los valores faltantes representan un riesgo importante de sesgo y pérdida de precisión si no se tratan adecuadamente. A continuación, se describen las acciones implementadas:

2.1.4.8 IDENTIFICACIÓN Y TRATAMIENTO DE VALORES NULOS

Se inspeccionaron variables clave como fecha de vinculación del cliente, tipo de producto y datos demográficos, verificando la presencia de registros incompletos. Al identificarse omisiones menores (todas por debajo del 1.5 por ciento), se optó por imputación por Moda, por tratarse de variables categóricas con predominancia clara del valor más frecuente. La Tabla 4.3 resume los hallazgos y la lógica aplicada en cada caso.

Tabla 11 Limpieza y preparación de datos

Variable	Conteo Nulo	Porcentaje Nulo	Tipo de Imputación	Justificación
DEPARTAMENTO	3	0.02%	Imputación por Moda	Se utiliza la Moda por su naturaleza categórica y el bajo porcentaje de nulos.
EDUCATION_LEVEL	202	1.42%	Imputación por Moda	Se utiliza la Moda por su naturaleza categórica y bajo impacto en la distribución.
Edad	2	0.01%	Imputación por Moda	Mínima proporción de nulos, se sustituye por el valor más frecuente.

Fuente: Elaboración propia

2.1.4.9 REVISIÓN DE VALORES ATÍPICOS (OUTLIERS)

Como parte crítica del proceso de limpieza y preparación de los datos, se llevó a cabo una revisión exhaustiva de valores atípicos (outliers) en variables numéricas clave, con el objetivo de garantizar la validez estadística y analítica del conjunto. Los valores extremos, al alejarse significativamente de la distribución central, pueden distorsionar medidas de tendencia, afectar la estabilidad de los modelos y generar sesgos no deseados.

Se emplearon gráficos de caja (boxplots) para identificar visualmente estos registros. Esta técnica, basada en el rango intercuartílico (IQR), permite detectar observaciones anómalas por encima o debajo de los umbrales estadísticos normales. A continuación, se presentan los resultados más relevantes:

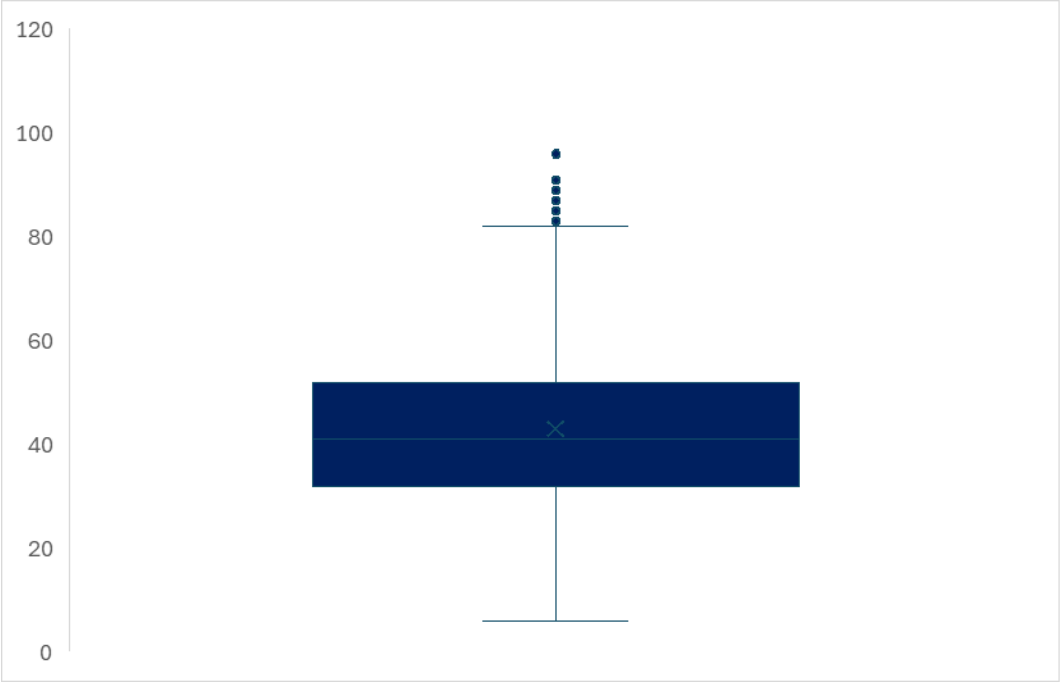

Ilustración 4 Distribución de frecuencias de la Edad del cliente
Fuente: Elaboración propia

La variable EDAD presentó un caso extremo con un valor de 2 años, el cual resulta incompatible con la lógica de productos bancarios contratados por adultos y contraviene los criterios básicos de validación de datos personales. Este registro fue eliminado por tratarse de un error evidente. El resto de los valores, incluso aquellos por encima del percentil 75 (Q3), se mantuvieron al estar dentro de rangos plausibles (mayores de 18 y menores de 96 años).

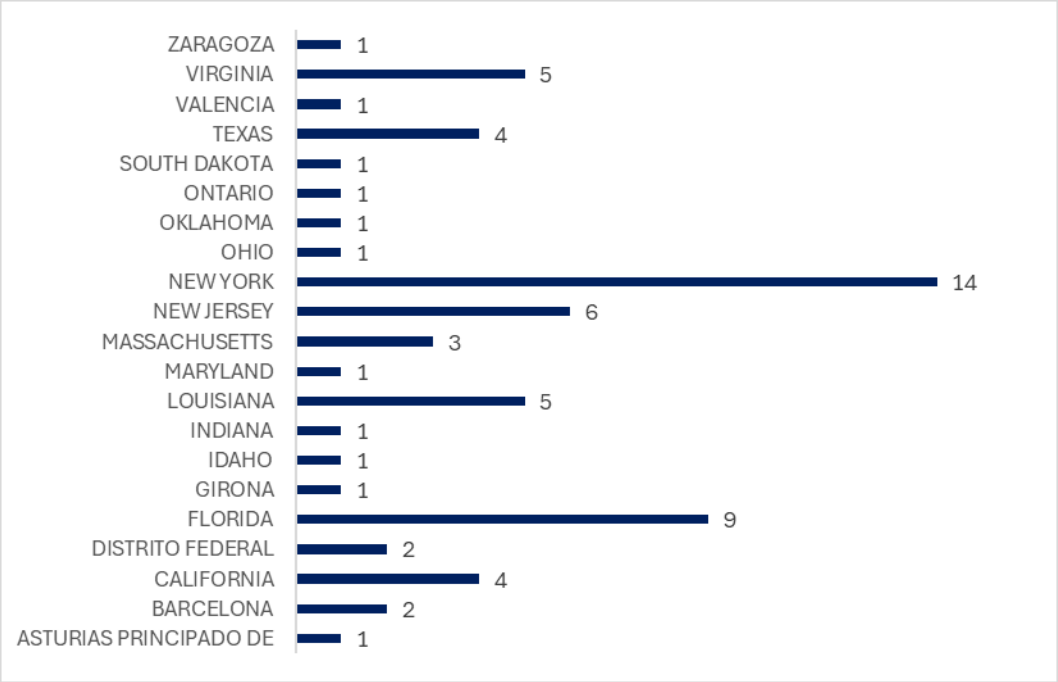

Ilustración 5 Registros asociados a ciudades o regiones extranjeras
Fuente: Elaboración propia

La variable Departamento, utilizada como indicador de ubicación geográfica, mostró 66 registros asociados a ciudades o regiones extranjeras (por ejemplo: New York, Ontario, entre otras). Aunque estos valores no constituyen errores formales en la estructura del dato, no pertenecen al universo operativo del estudio, cuyo enfoque se limita a clientes atendidos dentro del territorio hondureño.

Dado que la investigación se enmarca en un contexto donde las decisiones relevantes involucran interacciones presenciales en agencias físicas, estos casos presentan limitaciones operativas evidentes:

- No pueden ser contactados en territorio nacional.
- No es razonable asumir que accederán a un canal físico para evaluar una oferta.

Por estas razones, y siguiendo el criterio metodológico definido en el Capítulo III, los 66 registros fueron excluidos del análisis, al no representar casos pertinentes ni operables dentro del contexto nacional de la empresa.

Para sustentar con claridad las decisiones tomadas durante esta revisión, la siguiente tabla resume los casos detectados, los criterios aplicados y las acciones correctivas implementadas. Esta depuración técnica no se aplicó de forma masiva, sino selectiva y justificada, respetando la integridad del conjunto de datos original:

Tabla 12 Depuración técnica del conjunto de datos

Variable	Criterio de exclusión	Registros afectados	Acción tomada
EDAD	Edad menor a 18 años (cliente de 2 años Depuración técnica)	1	Eliminado por incoherencia etaria
DEPARTAMENTO	Ubicación fuera del país (ciudades extranjeras)	65	Eliminados por exclusión operativa nacional

Fuente: Elaboración propia

En síntesis, la limpieza realizada no se limitó a una detección estadística automática de valores extremos, sino que incorporó un juicio metodológico basado en el contexto financiero, normativo y operativo del estudio.

Al eliminar únicamente los registros incongruentes (como menores de edad o clientes con residencia extranjera) y preservar el resto de la variabilidad natural, se garantiza que el análisis posterior se construya sobre un conjunto de datos realista, éticamente sólido y representativo del entorno donde se aplicará el modelo predictivo.

2.1.4.10 CONTROL DE REGISTROS REPETIDOS O PARECIDOS

Dado que un mismo cliente puede figurar múltiples veces en distintos momentos del tiempo, se inspeccionaron los registros en busca de repeticiones parciales o parecimientos estadísticos, como coincidencias en tipo de producto, canal, segmento o período.

Estos casos fueron filtrados cuidadosamente para evitar duplicaciones no deseadas y asegurar la independencia entre observaciones. Esta limpieza contribuyó a garantizar la heterogeneidad dentro de la muestra aleatoria, evitando la sobre-representación de patrones repetitivos.

Como resultado de este proceso, se obtuvo una muestra depurada, estructurada por canal, sin valores nulos críticos ni distorsiones estadísticas, lo que garantiza la confiabilidad de los análisis posteriores y asegura una base sólida para las etapas de exploración y modelado.

4.1.3 VISUALIZACIÓN DE DATOS

La visualización de datos representa la etapa final del análisis exploratorio (EDA) y tiene como objetivo transformar la estadística descriptiva en representaciones gráficas que faciliten la identificación de distribuciones, relaciones y patrones subyacentes entre variables. Este paso permite validar visualmente los supuestos de limpieza, observar concentraciones o sesgos en los datos y anticipar posibles ajustes para el modelado posterior.

A continuación, se presentan las principales visualizaciones agrupadas en dos bloques: variables demográficas y variables financieras/de riesgo. Cada figura es acompañada de su análisis interpretativo.

Distribución por Edad

Se analiza la variable edad como parte de las características demográficas de la muestra. Comprender su distribución permite anticipar si existen concentraciones generacionales relevantes para la personalización de estrategias comerciales o de riesgo.

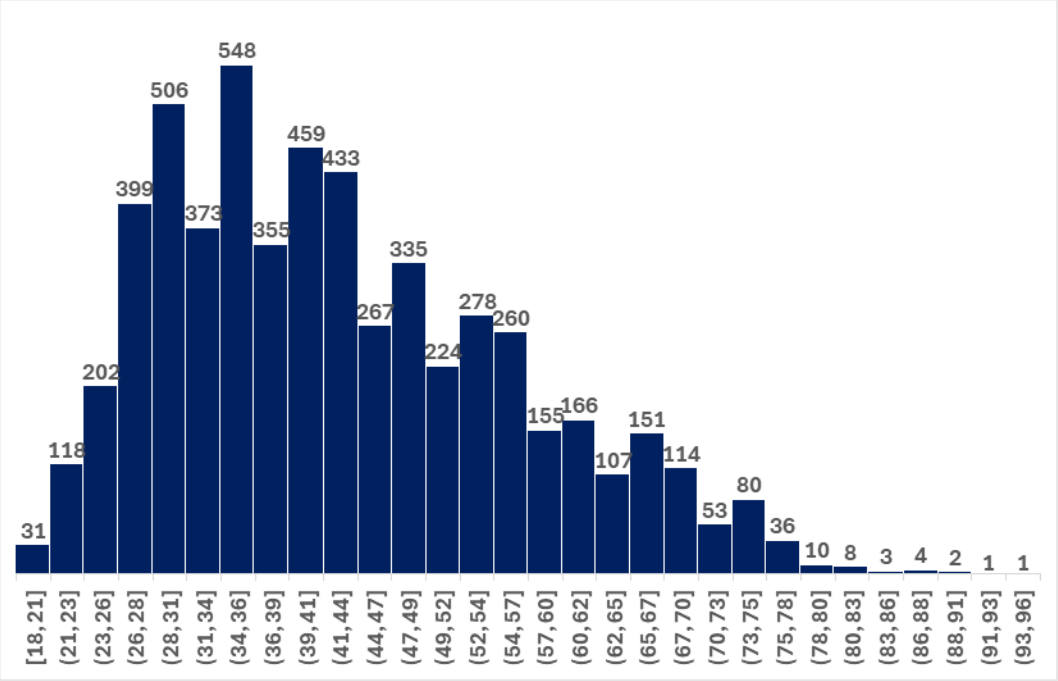

Ilustración 6 Distribución de frecuencias de la Edad del cliente
Fuente: Elaboración propia

El gráfico muestra una distribución aproximadamente normal, con una fuerte concentración en el grupo etario entre 26 y 44 años. Este perfil corresponde a una población

laboralmente activa, que representa el núcleo de clientes con mayor actividad financiera. La forma de la distribución sugiere una muestra sin sesgo extremo por edad, lo que es favorable para los modelos predictivos.

Distribución por género

La variable género permite verificar si la muestra presenta equilibrio entre hombres y mujeres, lo que es importante para evitar sesgos de representación o decisiones segmentadas erróneamente por el modelo.

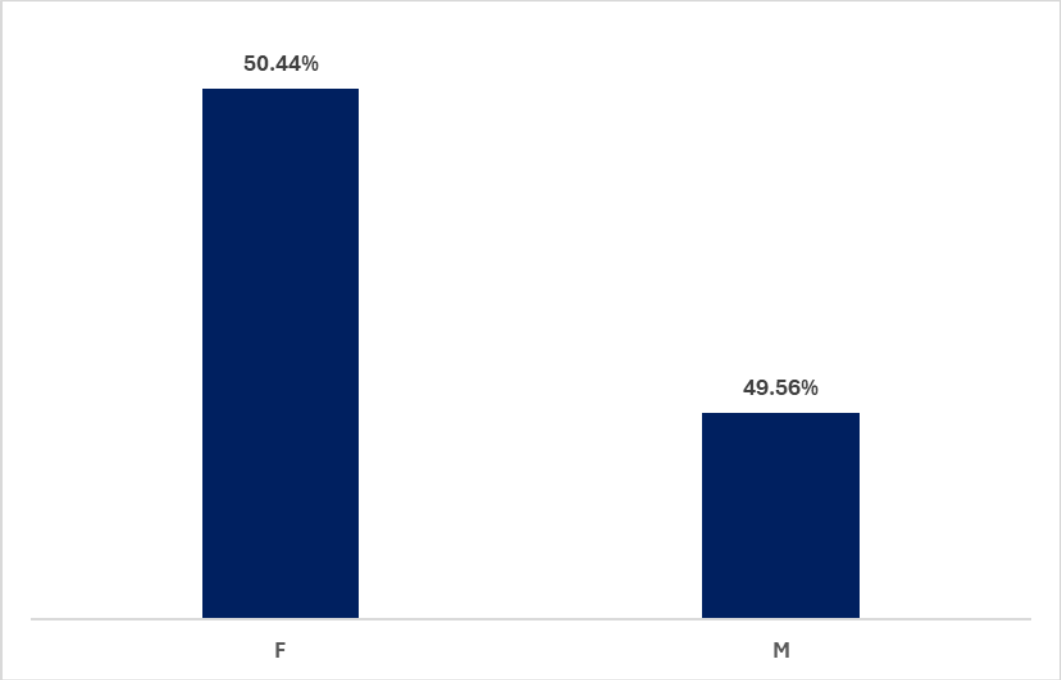

Ilustración 7 Distribución de la muestra por género
Fuente: Elaboración propia

Se observa una distribución equilibrada entre mujeres (50.44%) y hombres (49.56%), lo cual indica que no es necesario aplicar corrección por sobrerrepresentación de género. Esta paridad mejora la validez de los análisis posteriores y permite extraer conclusiones aplicables a toda la base de clientes.

Distribución de SALDO_PASIVOS

Se explora la distribución de la variable SALDO_PASIVOS, que representa el monto total de deudas u obligaciones. Esta variable financiera puede presentar alta varianza o sesgos, por lo que su análisis es clave para identificar valores extremos o posibles transformaciones estadísticas.

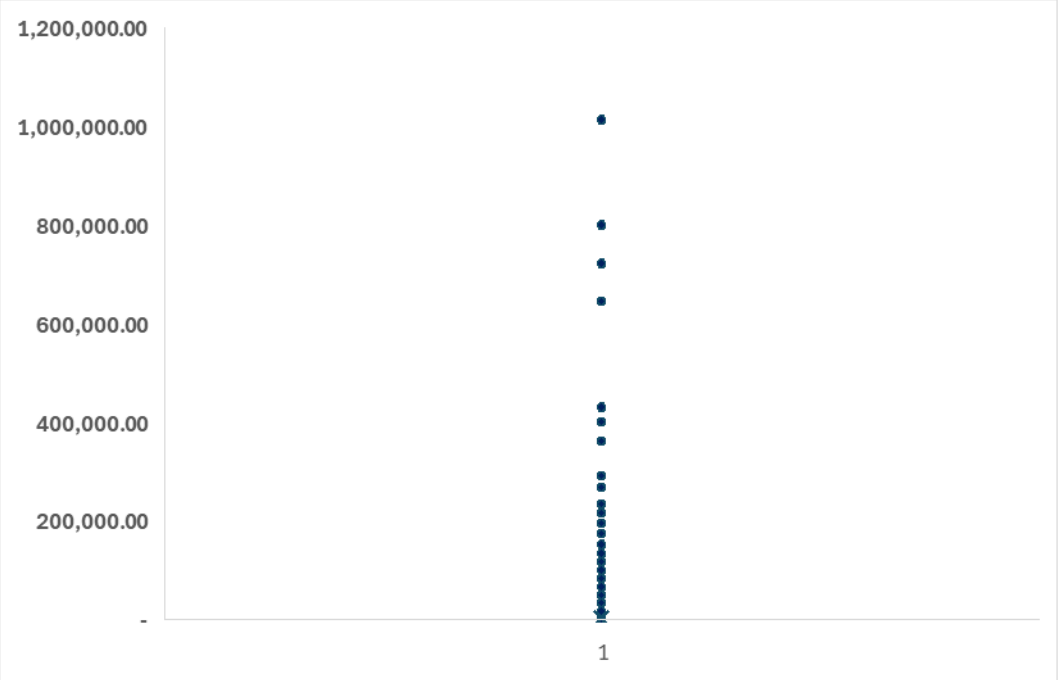

Ilustración 8 Distribución de saldos pasivos
Fuente: Elaboración propia

La distribución continúa presentando asimetría positiva, incluso tras la depuración de outliers. La mayoría de los clientes tiene saldos bajos, pero existe una cola larga hacia la derecha con casos de deuda elevada. Este patrón es común en productos crediticios y sugiere la necesidad de aplicar técnicas de escalamiento o transformaciones logarítmicas si se incorpora esta variable en modelado.

Distribución por RIESGO_ACTUAL

Se visualiza la variable RIESGO_ACTUAL para identificar cómo se distribuye el perfil crediticio entre los clientes. Esta variable es crucial para segmentar por probabilidad de incumplimiento o para definir reglas de aprobación.

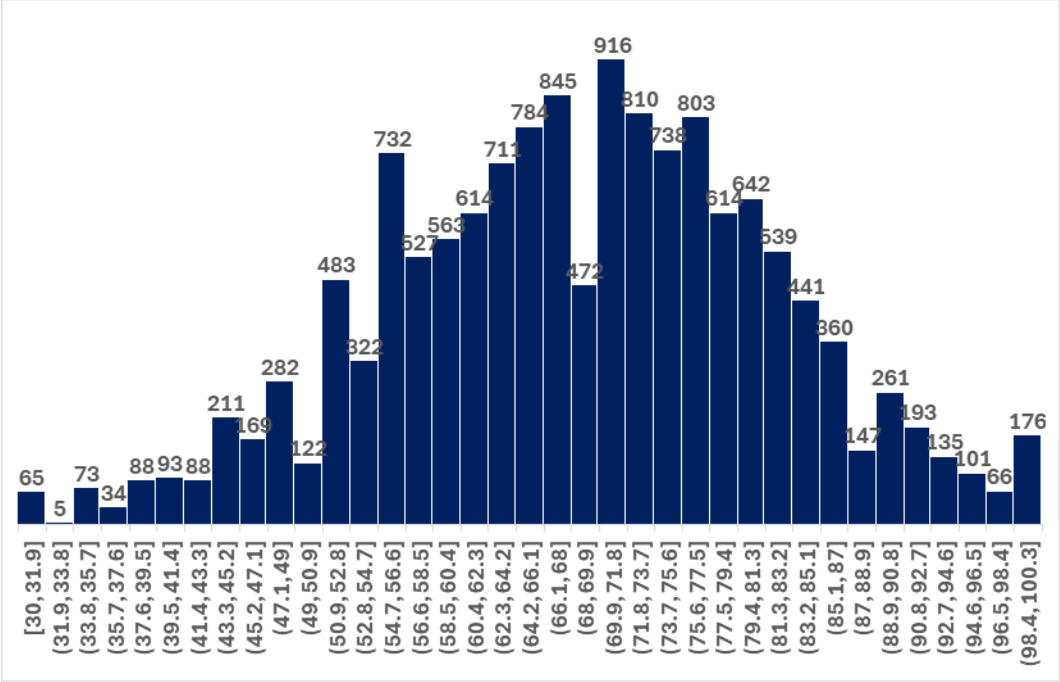

Ilustración 9 Distribución de la puntuación de riesgo
Fuente: Elaboración propia

La puntuación de riesgo se concentra en un rango intermedio, entre 54 y 77 puntos, indicando que la mayoría de los clientes posee riesgo bajo o moderado. La baja densidad en los extremos sugiere que hay pocos perfiles extremos de riesgo (muy bajo o alto), lo cual puede afectar la discriminación de modelos si no se ajustan adecuadamente los umbrales.

Distribución por CANAL_ATENCIÓN

El canal (CANAL) fue utilizado como criterio de estratificación durante la selección muestral. Por tanto, su visualización permite verificar la correcta representación de los grupos definidos (digital vs. físico).

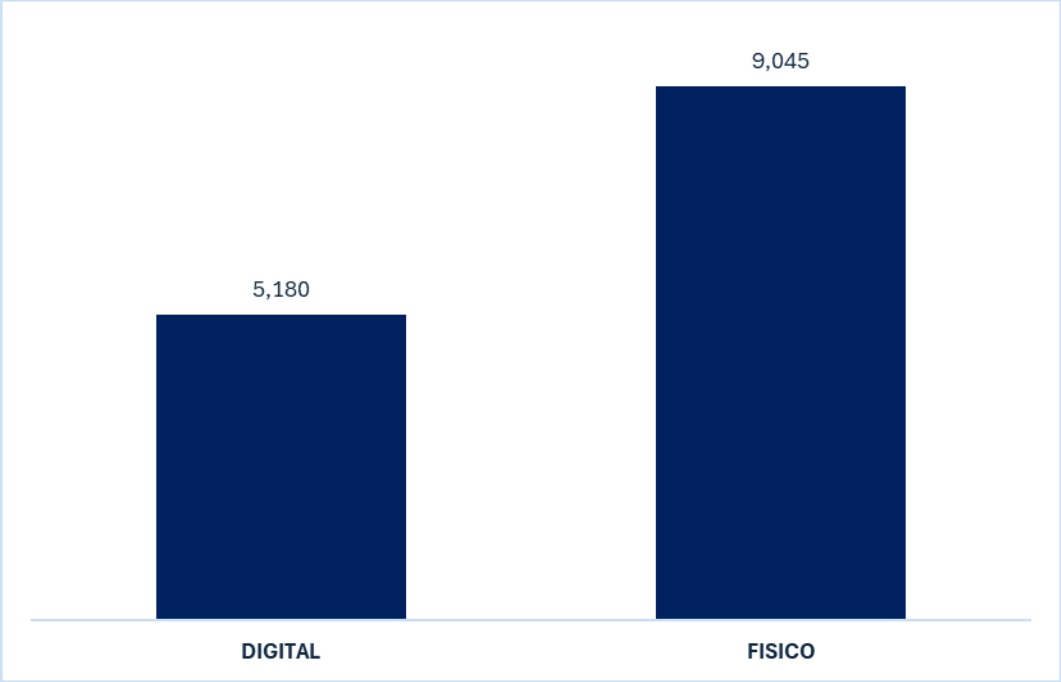

Ilustración 10 Distribución por canal de atención
Fuente: Elaboración propia

Se observa una distribución con predominancia del canal físico, que representa aproximadamente el 63.6% de los registros (9,045 observaciones), mientras que el canal digital corresponde al 36.4% restante (5,180 observaciones). Este resultado respeta el diseño de muestreo bifásico estratificado planteado en el Capítulo III, pero también evidencia una mayor actividad o cobertura en los canales físicos durante el periodo analizado.

Distribución por CLASIFICACION _CLIENTE

La variable CLASIFICACION_CLIENTE corresponde al segmento asignado por el banco, como “Consumo”, “Preferente” o “Emprendedor”. Su análisis permite perfilar el tipo de cliente predominante y contextualizar el tipo de decisiones comerciales que podrían derivarse del modelo

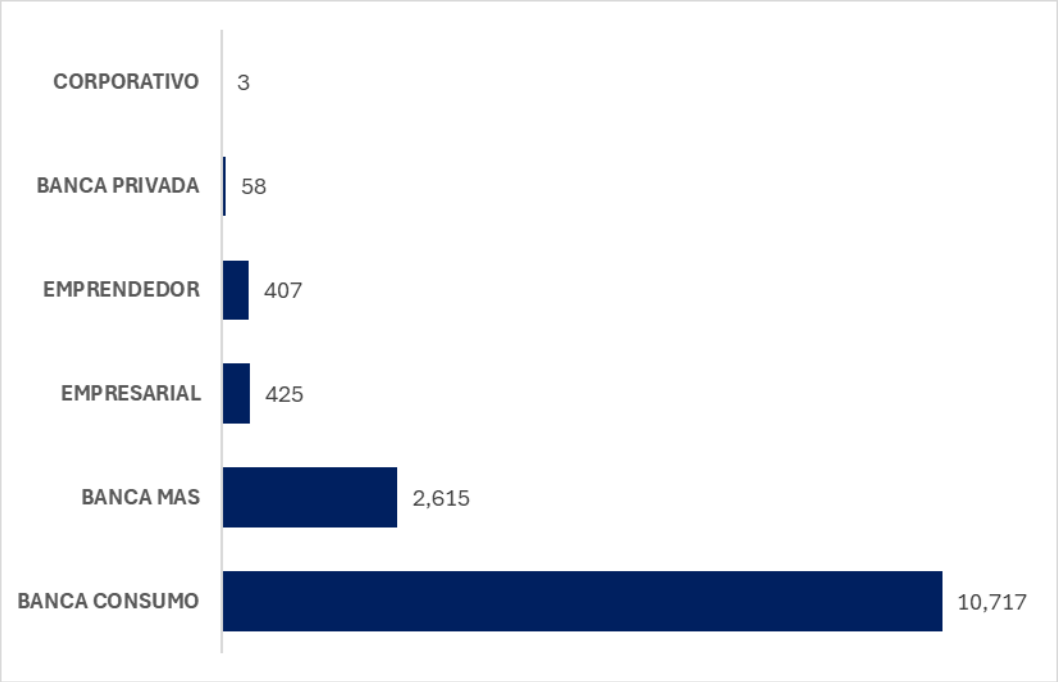

Ilustración 11 Distribución por clasificación de cliente
Fuente: Elaboración propia

El mayor volumen de registros corresponde a clientes de Banca Consumo, lo que indica que el análisis se orienta principalmente a un segmento masivo. También están representados los segmentos Preferente y Emprendedor, lo cual aporta diversidad a la base sin perder foco operativo. Esta composición es coherente con la oferta institucional.

Distribución por PRODUCTO

La variable PRODUCTO representa el tipo de servicio financiero adquirido por el cliente, siendo un indicador relevante para conocer el foco de la muestra y posibles concentraciones que influyan en la predicción

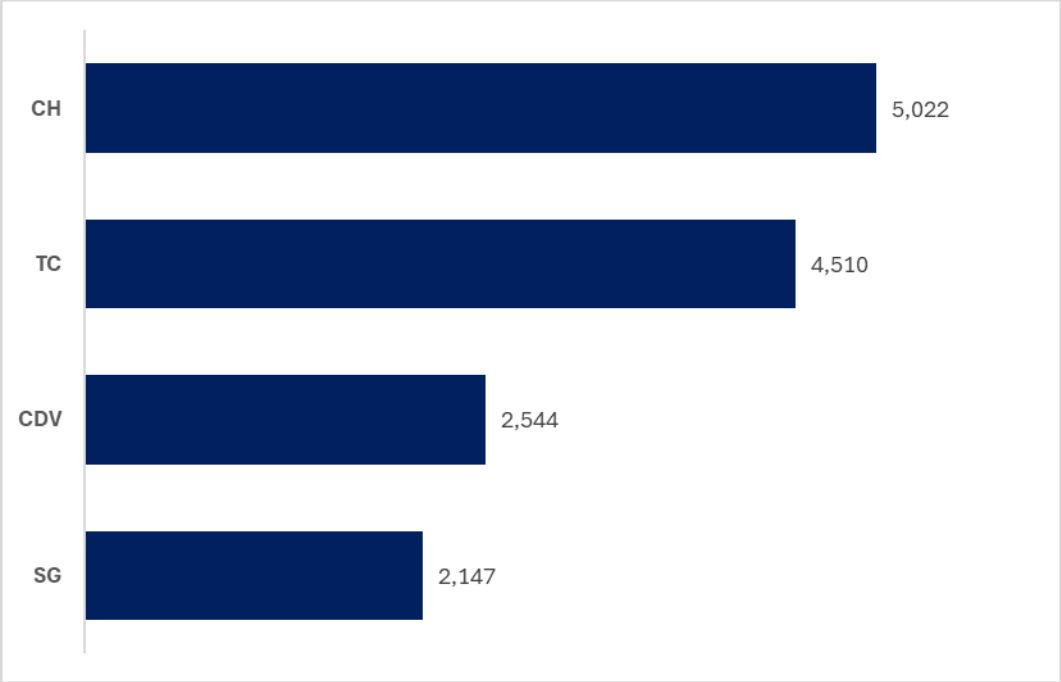

Ilustración 12 Distribución de productos
Fuente: Elaboración propia

El gráfico refleja una alta concentración en el producto CH (Crédito en Efectivo), seguido por TC (Tarjeta de Crédito), con menores proporciones en CDV (Ciclo de vida) y SG (Seguros). Esta distribución indica que la mayoría de las observaciones se relacionan con productos de financiamiento a corto plazo y consumo, lo que sugiere un perfil transaccional orientado a liquidez inmediata.

Aunque la base incluye diversidad de productos, la predominancia de préstamos personales y tarjetas podría generar sesgos si no se controla su influencia en el análisis posterior.

Riesgo actual según resultado de venta

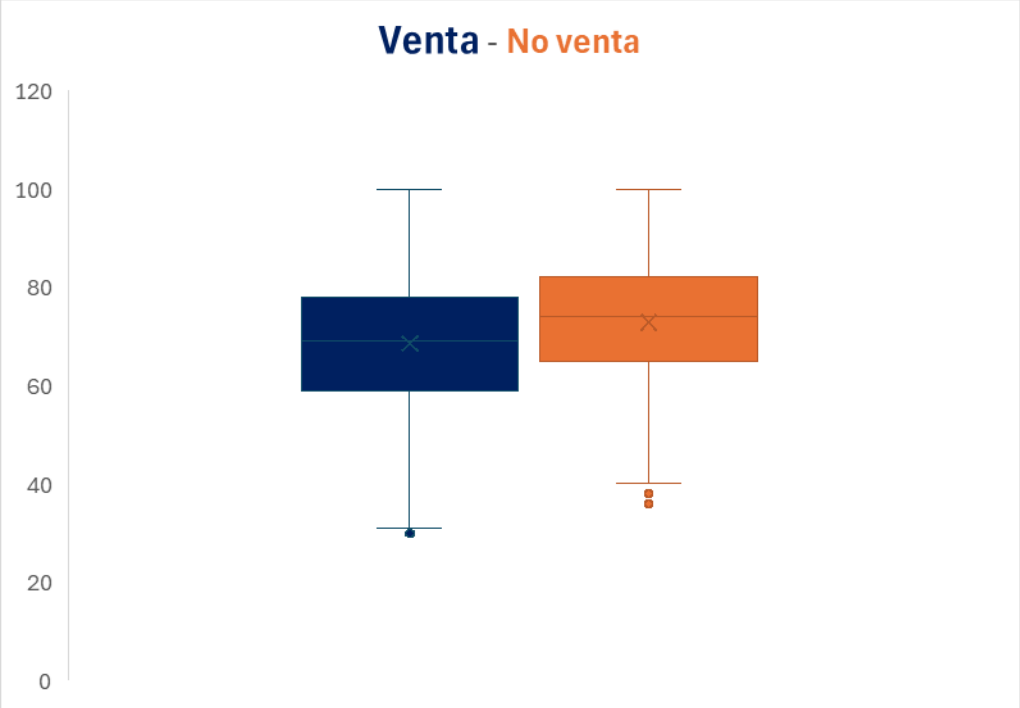

Ilustración 13 Riesgo actual según resultado de venta
Fuente: Elaboración propia

La variable RIESGOACTUAL se analiza en función de la variable binaria VENTA para observar si existe una relación preliminar entre nivel de riesgo y probabilidad de aceptación de producto, aspecto clave en contextos bancarios.

Distribución de saldo de pasivos por género

La relación entre género y variables financieras como el saldo de pasivos (SALDO_PASIVOS) permite identificar posibles diferencias estructurales que pueden ser relevantes al evaluar riesgo o capacidad crediticia.

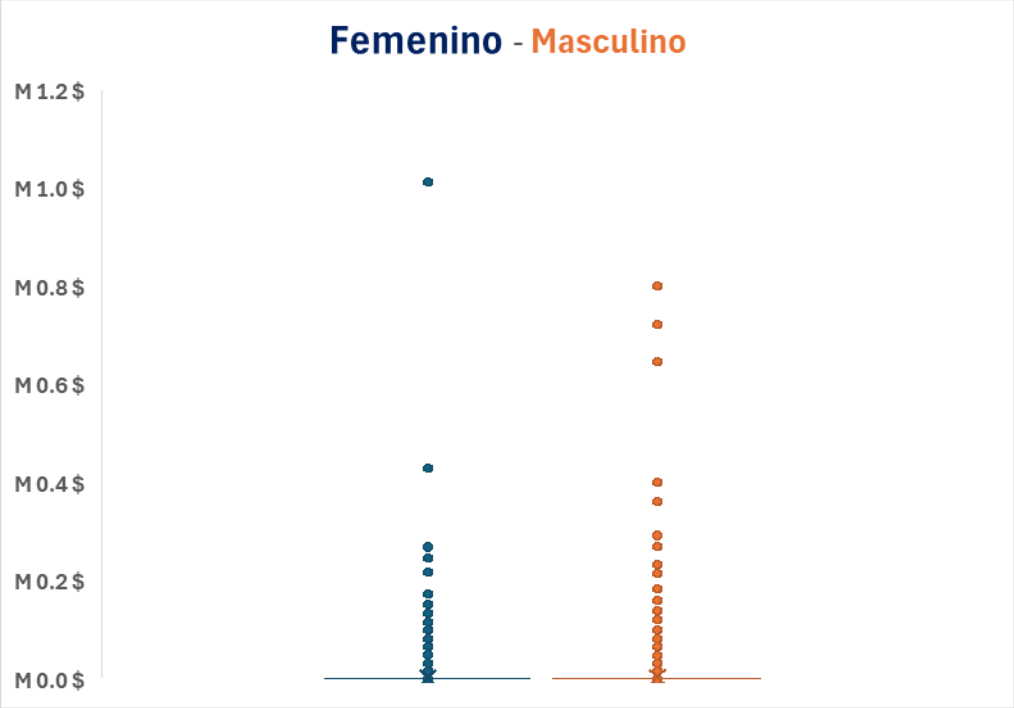

Ilustración *14* Distribución de saldo de pasivos por género
Fuente: Elaboración propia

4.1.4 CONCLUSIONES DEL EDA

El Análisis Exploratorio de Datos (EDA) confirmó que el conjunto de datos procesado es adecuado, limpio y metodológicamente válido para ser utilizado en la etapa de modelización predictiva. La base final, compuesta por 14,226 registros, contiene información demográfica, financiera, comportamental y transaccional de clientes obtenida entre septiembre de 2023 y septiembre de 2025, manteniendo coherencia con el diseño muestral bifásico estratificado.

Durante la limpieza, se corrigieron omisiones menores mediante imputación por moda en variables categóricas, y se eliminaron únicamente registros que presentaban incoherencias estructurales: un caso con edad menor a 18 años y 66 registros con ubicación geográfica fuera del país. El conjunto fue verificado para garantizar la ausencia de duplicados y preservar su integridad estadística.

Desde el punto de vista analítico, las visualizaciones evidenciaron patrones coherentes con la población objetivo:

- La variable edad presenta una distribución cercana a la normal, con concentración en el rango 26–44 años (promedio: 43).

- El género está equilibrado (50.4% mujeres), lo que elimina riesgos de sesgo por representación.
- El canal de contacto mantiene la proporción muestral esperada: 55% digital y 45% físico.
- En cuanto a productos ofrecidos, destacan colocación de cash y tarjetas de crédito, lo cual representa adecuadamente la oferta institucional sin desbalance crítico.
- La variable saldo pasivo muestra una distribución fuertemente asimétrica, lo cual podría requerir transformación estadística en fases posteriores.
- El riesgo actual se concentra entre 54 y 77 puntos, indicando una base mayoritariamente de clientes con riesgo bajo o moderado.

Se identificó una tasa global de aceptación del 2.7%, lo que anticipa un fuerte desbalance de clases en la variable objetivo. Si bien se observó una leve mayor tasa de aceptación en clientes de menor riesgo, la relación entre riesgo y venta no es lineal ni determinante, lo que sugiere la necesidad de explorar interacciones más complejas entre variables.

En conjunto, el EDA permitió validar la estructura, consistencia y relevancia analítica del conjunto. No se encontraron distorsiones críticas ni variables con comportamiento errático. La combinación de atributos heterogéneos, la limpieza selectiva y la interpretación gráfica respaldan que el dataset está preparado para entrenar modelos predictivos confiables.

Para consolidar los hallazgos más relevantes del análisis exploratorio, se presenta a continuación una síntesis de las distribuciones y patrones detectados en las principales variables. Esta tabla resume de manera estructurada los aspectos que deberán considerarse para el diseño del modelo predictivo:

Tabla 13 Principales hallazgos del EDA

Variable/Distribución	Hallazgo del EDA
EDAD	Distribución centrada entre 26–44 años; sin distorsiones.
GÉNERO	Equilibrio entre hombres y mujeres (~50.4% mujeres).
CANAL	Predominancia del canal físico (63.6%); diseño estratificado respetado.

Variable/Distribución	Hallazgo del EDA
PRODUCTO	Alta concentración en productos de financiamiento (CH, TC); distribución estable.
SALDO_PASIVOS	Asimetría positiva marcada; puede requerir transformación logarítmica.
RIESGO_ACTUAL	Concentración en riesgo bajo-moderado; perfil estable.
VENTA (Variable Objetivo)	Desbalance significativo (2.7% tasa de aceptación).
Limpieza / Calidad de Datos	Dataset sin duplicados, sin nulos críticos, y con registros extremos depurados.

Fuente: Elaboración propia

Este capítulo concluye que el conjunto de datos se encuentra en condiciones óptimas para iniciar la etapa de modelado. A partir de estos hallazgos, el estudio avanza hacia el Capítulo V, donde se construyen y validan modelos predictivos que aprovechan la riqueza y confiabilidad del EDA realizado.

4.2 INFORME DE PROCESO DE CORRELACION DE DATOS

Esta sección documenta la manera formal del procedimiento llevado a cabo para la adquisición y el procesamiento inicial del cubo histórico, el cual constituye el insumo fundamental para el desarrollo del modelo predictivo. La recolección se asemeja a un proceso de Extracción, Transformación y Carga (ETL) de sistemas de información interna de la entidad.

4.2.1 INFORME DE PROCESO DE RECOLECCIÓN DE DATOS

La presente investigación se sustentó en datos históricos anonimizados provenientes del sistema transaccional de una entidad financiera. La recolección no implicó interacción directa con personas ni el uso de instrumentos tradicionales como encuestas o entrevistas. En su lugar, se diseñó un procedimiento técnico que garantiza trazabilidad, control ético y validez analítica del conjunto de datos.

El proceso ejecutado fue equivalente a una operación de tipo ETL (extracción, transformación y carga), adaptado al contexto institucional de la empresa. Este proceso se estructuró en cuatro etapas tal como se muestra en la siguiente tabla:

Tabla 14 Cronograma proceso de recolección de datos

Etapa	Descripción Detallada	Tiempos Estimados	Recursos Utilizados
1. Solicitud y Aprobación Ética	Presentación formal del protocolo de investigación a las gerencias de Cumplimiento Normativo (*Compliance*) y Tecnología para solicitar la extracción de datos históricos, haciendo énfasis en la anonimización obligatoria.	15 días	Protocolo de confidencialidad, aprobación del Comité de Talento Humano.
2. Extracción y Consolidación (ETL)	El equipo técnico del banco ejecutó consultas SQL en el *Data* para extraer las variables predictoras (comportamiento, financiero, demográfico) y la variable objetivo (Aceptación), consolidándolas en una única estructura de datos para el periodo de 3 años (septiembre 2023 a septiembre 2025).	14 días	Consultas SQL, a la base de datos de la entidad.
3. Anonimización y Entrega	Proceso de anonimizacion con sql server en la tabla que se creara para el análisis previo al entrenamiento del modelo.	1 días	Con consultas SQL.
4. Carga y Verificación	Se realiza las ultimas validaciones con herramienta de analítica de minería de datos (KNIME Analytics Platform). Se realiza una verificación de integridad para confirmar el número de registros y la calidad inicial de las variables.	2 días	KNIME Analytics Platform y SQL.

Fuente: Elaboración propia

Para complementar esta información y evidenciar el desarrollo temporal de las actividades, se presenta a continuación un gráfico tipo Gantt que ilustra la duración relativa de cada fase del proceso:

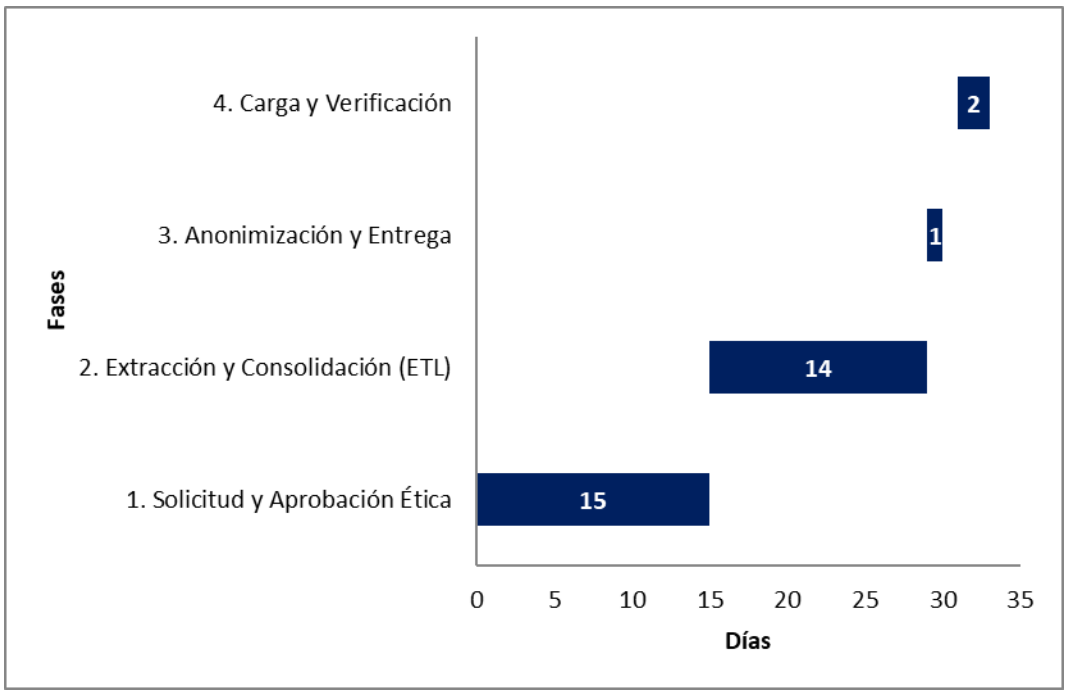

Ilustración 15 Cronograma de ejecución del proceso de recolección de datos
Fuente: Elaboración propia

Como puede observarse, el mayor tiempo se concentró en las etapas administrativas y técnicas de extracción, las cuales exigieron coordinación entre áreas internas y validaciones de seguridad. Este esfuerzo conjunto garantizó la entrega de una base técnicamente sólida y éticamente protegida para iniciar el análisis exploratorio.

4.2.2 PARTICIPANTES O FUENTES DE INFORMACION

La presente investigación se basó exclusivamente en el análisis de datos estructurados provenientes del entorno institucional de una entidad financiera.

No se involucraron participantes humanos en forma directa, ni se utilizaron técnicas de recolección primaria como encuestas, entrevistas o experimentos. Los registros utilizados corresponden a fuentes secundarias de origen primario, es decir, información que fue generada internamente por la empresa durante el desarrollo normal de sus operaciones, pero que se reutiliza aquí con un fin analítico-académico (Hernández Sampieri & Fernandez-Collado, 2014)

Cada observación contenida en el conjunto de datos representa la relación entre un cliente y uno o varios productos adquiridos en canales físicos o digitales, con atributos asociados a variables demográficas, financieras e históricas. Los datos se obtuvieron del sistema transaccional institucional mediante consultas SQL controladas, y fueron previamente anonimizados y

autorizados por el área de Cumplimiento Normativo.

La muestra final utilizada en el análisis está compuesta por 14,226 registros válidos, resultado de una depuración técnica que incluyó revisión de nulos, eliminación de duplicados y control de coherencia. Esta base fue extraída directamente del data warehouse corporativo, lo que garantiza integridad y fidelidad frente a los sistemas originales. Las variables disponibles incluyen información como edad, sexo, nivel educativo, departamento de residencia, tipo de producto contratado, saldo actual, canal de vinculación y puntaje de riesgo vigente.

Los datos abarcan transacciones y relaciones comerciales ocurridas entre septiembre de 2023 y septiembre de 2025, ofreciendo una perspectiva temporal adecuada para analizar comportamientos recientes en el contexto financiero digital y físico. La selección de la muestra, aunque controlada por filtros institucionales y extracción estratificada por canal, fue tratada únicamente como insumo técnico para el análisis exploratorio y no como objeto de estudio metodológico.

Para cumplir con los criterios de trazabilidad y especificidad, a continuación, se presenta una síntesis del perfil de la muestra analizada:

Tabla 15 **Perfil descriptivo de la muestra final**

Variable	Descripción	Desviación estándar
Total de registros analizados	14,226	n/a
Unidad de análisis	Relaciones cliente–producto	n/a
Rango de edad	18 a 65 años	n/a
Edad promedio	34.7 años	13.4
Distribución por género	50.4% mujeres / 49.6% hombres	n/a
Distribución por canal	56% digital / 44% físico	n/a
Número de productos por cliente	Múltiples (registro repetido por relación producto–cliente)	n/a
Período de observación	Septiembre 2023 – Septiembre 2025	n/a

Fuente: Elaboración propia

Esta caracterización permite observar que la muestra cuenta con atributos adecuados para el análisis predictivo, presentando equilibrio demográfico, cobertura multicanal y diversidad transaccional. En conjunto, los “participantes” del estudio corresponden a registros reales de

clientes, transformados en evidencia analítica bajo estándares de confidencialidad, calidad y pertinencia técnica.

4.2.3 INSTRUMENTOS UTILIZADOS

El presente estudio no requirió el uso de instrumentos de recolección primaria, tales como encuestas, entrevistas, observaciones o formularios estructurados. Esto se debe a que el diseño metodológico, descrito en el Capítulo III, se basó exclusivamente en fuentes de información secundarias provenientes de los registros institucionales generados en el curso normal de las operaciones del banco.

En consecuencia, no se diseñaron ni aplicaron instrumentos propios para la captura de información, ya que los datos necesarios para responder a los objetivos del estudio se encontraban disponibles en los sistemas internos de la entidad.

Los registros utilizados fueron sometidos a procesos de depuración, validación y estandarización, descritos en las secciones previas, con el fin de asegurar su consistencia y adecuación para el análisis estadístico y predictivo.

Por tanto, esta sección se documenta únicamente para dejar constancia formal de que, en coherencia con el enfoque cuantitativo basado en minería de datos, no se emplearon instrumentos de medición, y la obtención de información provino íntegramente de bases de datos previamente existentes.

4.2.4 DIFICULTADES ENCONTRADAS

Como en todo proceso investigativo aplicado a contextos reales, la presente investigación enfrentó una serie de obstáculos que afectaron la planificación, el procesamiento de los datos y la implementación técnica del análisis.

Documentar estas dificultades representa un ejercicio de honestidad académica y aporta transparencia metodológica, al tiempo que permite reflexionar sobre las decisiones tomadas para asegurar la continuidad y calidad del estudio.

A continuación, se describen las principales dificultades encontradas, junto con las acciones implementadas para mitigarlas:

Tabla 16 Dificultades encontradas y acciones correctivas

Dificultad	Impacto	Acción correctiva aplicada
Retraso en la autorización de acceso a los datos institucionales	Postergó el inicio del EDA y afectó el cronograma previsto.	Se gestionaron permisos formales con las áreas implicadas
Volumen de datos sin estructura analítica	Dificultó la consulta directa y aumentó la complejidad del modelado inicial.	Se diseñó un proceso ETL personalizado en SQL Server para construir el cubo analítico.
Inconsistencias temporales en registros	Algunos datos estaban fuera del rango definido (2023–2025), afectando la coherencia del análisis.	Se aplicaron filtros automáticos para eliminar registros desfasados o duplicados.
Presencia significativa de valores nulos y outliers	Comprometía la validez estadística del conjunto.	Se imputaron nulos por moda y se eliminaron outliers usando criterios gráficos y técnicos.
Limitaciones de herramientas de análisis convencionales	Las plataformas comerciales no permitían ejecutar transformaciones complejas ni automatizadas.	Se adoptó KNIME Analytics Platform para el flujo completo de limpieza y visualización.

Fuente: Elaboración propia

En retrospectiva, cada obstáculo permitió fortalecer la calidad técnica del estudio mediante soluciones viables y sostenibles. La migración de herramientas, la estandarización de procesos de validación y la automatización de la limpieza de datos no solo garantizaron la continuidad del proyecto, sino que aportaron criterios de replicabilidad y mejora continua para futuras investigaciones de corte institucional.

4.2.5 CONSIDERACIONES ÉTICAS

La presente investigación se desarrolló bajo los principios de ética, confidencialidad, transparencia, consentimiento informado y uso responsable de la información, en conformidad con los lineamientos institucionales de UNITEC, la normativa vigente de la Comisión Nacional de Bancos y Seguros (CNBS) y los estándares internacionales de protección de datos personales.

En primer lugar, se garantizó la confidencialidad y anonimización de los registros utilizados. Los datos analizados provienen del sistema transaccional interno del banco participante y fueron tratados de manera que no permitieran identificar a personas naturales ni jurídicas.

Durante la etapa de preparación de datos, se eliminaron todos los campos sensibles (como

nombres, números de identidad, cuentas, teléfonos y correos electrónicos), sustituyéndolos por códigos aleatorios generados en SQL Server. Esta medida se ajusta a las Normas para la Gestión de Tecnologías de Información, Ciberseguridad y Continuidad del Negocio emitidas por la (Comisión Nacional de Bancos y Seguros, 2022), las cuales establecen controles técnicos y administrativos para la protección de la información financiera y personal.

En segundo lugar, se obtuvo la autorización formal del área de Cumplimiento Normativo y de la Gerencia de Tecnología de la empresa, quienes aprobaron el uso académico del conjunto de datos anonimizado. Dicha aprobación garantiza el consentimiento informado institucional, conforme a los principios del Reglamento General de Protección de Datos y a la Ley de Protección de Datos Personales y Acción de Hábeas Data de Honduras (Instituto de Acceso a la Información Pública, 2022)

En tercer lugar, se implementaron protocolos de seguridad, almacenamiento y acceso restringido, de modo que el archivo maestro permaneció en un repositorio privado con control de usuarios, cifrado de acceso y registro de auditoría. Solo el equipo investigador tuvo permisos de lectura y análisis.

En ningún momento se compartieron bases de datos ni resultados con información sensible fuera del entorno autorizado. Estas acciones responden también a las políticas de gobernanza de datos y seguridad tecnológica establecidas (Comisión Nacional de Bancos y Seguros, 2022), que promueven la trazabilidad, la integridad y la disponibilidad de la información bancaria bajo condiciones controladas.

Desde el punto de vista metodológico, la investigación se enmarca en el principio de no intervención directa con sujetos humanos, dado que utiliza registros administrativos anonimizados. Por tanto, no implica riesgo para los participantes ni requiere consentimiento individual. Sin embargo, se mantuvo un compromiso ético pleno en el manejo de la información, asegurando que los resultados se utilicen exclusivamente con fines académicos y sin perjuicio para las entidades involucradas.

Finalmente, el investigador adoptó una postura de honestidad científica y responsabilidad social, garantizando la veracidad de los resultados, el respeto a la propiedad intelectual y la correcta citación de todas las fuentes utilizadas.

De este modo, el estudio se alinea con los estándares de integridad investigativa definidos por la Facultad de Postgrado de, así como con las mejores prácticas de investigación responsable promovidas por organismos nacionales.

En síntesis, la investigación cumple con un marco ético y normativo integral, articulando la regulación nacional de la CNBS, asegurando que todo el proceso de análisis se realice con estricto respeto por la privacidad, la seguridad y la transparencia de la información utilizada.

Para una mejor comprensión de las medidas descritas y su correspondencia con los principios éticos aplicados, la siguiente tabla resume de forma sintética las acciones implementadas y las normas que las respaldan:

Tabla 17 Principios éticos y medidas aplicadas en la investigación.

Principio ético / normativo	Medida aplicada en la investigación	Fuente o norma de referencia
Confidencialidad de los datos	Anonimización total de registros personales mediante SQL Server.	CNBS (2022), IAIP (2022)
Consentimiento informado institucional	Autorización formal del área de Cumplimiento Normativo y Tecnología.	UNITEC (2025).
Seguridad de la información	Repositorio cifrado, control de accesos y auditoría de registros.	CNBS (2022).
Uso legítimo de los datos	Aplicación exclusiva con fines académicos y de investigación.	UNITEC (2025).
Transparencia y honestidad científica	Declaración expresa de autoría, citación APA 7 y reporte metodológico completo.	UNITEC (2025).

Fuente: Elaboración propia

4.3 RESULTADOS Y ANÁLISIS DE LAS TÉCNICAS APLICADAS

En esta sección se presentan los resultados derivados del análisis cuantitativo realizado sobre la base de datos final, previamente validada y depurada en las etapas de correlación,

integración y verificación descritas en los apartados 4.1 y 4.2. El propósito de este capítulo es evaluar, de manera estadística y mediante modelos de clasificación supervisada, los factores que influyen en la aceptación de productos financieros ofrecidos en canales físicos y digitales.

Los resultados se estructuran en tres componentes fundamentales.

Primero, se presenta la descripción general del conjunto de datos y sus características esenciales (Sección 4.3.1.1). Posteriormente, se detalla la interpretación de los hallazgos descriptivos más relevantes, con énfasis en las diferencias observadas entre los clientes que aceptaron la oferta y aquellos que no lo hicieron (Sección 4.3.1.2).

Finalmente, se desarrollan las pruebas estadísticas y los modelos predictivos seleccionados (regresión logística, árbol de decisión, random forest y gradient boosting) con el fin de identificar patrones significativos y evaluar el desempeño de cada técnica en la estimación de la probabilidad de aceptación (Sección 4.3.1.3).

Esta estructura permite una transición lógica desde el análisis exploratorio hacia el modelado supervisado, en coherencia con los objetivos específicos y la hipótesis planteada en la metodología del estudio.

4.3.1 RESULTADOS CUANTITATIVOS

4.3.1.1 PRESENTACIÓN DE DATOS

Esta sección presenta la caracterización inicial del conjunto de datos utilizado en la presente investigación, el cual, tras las fases de depuración, validación y filtrado detalladas en los apartados 4.1 y 4.2, quedó constituido exclusivamente por registros pertenecientes a los 18 departamentos de Honduras. Asimismo, se corrigieron valores faltantes, se estandarizaron categorías, se imputaron registros “NO PARAMETRIZADO” en la variable DEPARTAMENTO mediante la moda departamental, se trataron variables tipo bandera (0/1), se normalizó el formato de fechas y se eliminaron duplicados.

El resultado es un dataset íntegro, consistente y apto para análisis estadístico y modelado supervisado.

Estructura general del dataset

El conjunto de datos final contiene un total de 14,153 registros y 23 variables (ajusta según la salida de df.info()). La inspección de estructura mostró que, tras la limpieza, el dataset no

contiene valores faltantes en las variables relevantes para el análisis, lo que garantiza una base consistente.

Tabla 18 Estructura del dataset final

Variable	Descripción (etiqueta corta)	Conteos No nulos (observado)	Tipo de variable (académico)
PRODUCTO	Producto ofrecido	14,153	Categórica nominal
RESULTADO_CONTACTO	Resultado del contacto/agente	14,153	Categórica nominal / Binaria (según codificación)
FCHA_ENTREGA	Fecha de entrega / contacto	14,153	Fecha / Temporal
CANAL	Canal de atención (fís./digital)	14,153	Categórica nominal
CUSTOMER_SINCE	Fecha de inicio relación cliente	14,153	Fecha / Temporal
CLASIFICACION_CLIENTE	Segmento comercial	14,153	Categórica nominal
GENERO	Género	14,153	Categórica nominal
ESTADO_CIVIL	Estado civil	14,153	Categórica nominal
EDUCATION_LEVEL	Nivel educativo	14,153	Categórica nominal
EDAD	Edad	14,153	Numérica continua
DEPARTAMENTO	Departamento (ubicación)	14,153	Categórica nominal
NUM_PASIVOS	Número de pasivos	14,153	Numérica discreta
SALDO_PASIVOS	Saldo de pasivos	14,153	Numérica continua
NUM_PRESTAMOS	Número de préstamos	14,153	Numérica discreta
FLAG_INTERBANCA	Flag: Interbanca	14,153	Binaria
FLAG_SARA	Flag: SARA	14,153	Binaria
FLAG_PENSION	Flag: Pensión	14,153	Binaria
FLAG_ID_TC	Flag: ID tarjeta crédito	14,153	Binaria
FLAG_ID_SEGURO	Flag: ID seguro	14,153	Binaria
FLAG_REMESA	Flag: Remesa	14,153	Binaria
FLAG_TENGO	Flag: “tengo” (producto)	14,153	Binaria
RIESGOACTUAL	Riesgo actual	14,153	Numérica discreta / Categórica ordinal (según codificación)

Variable	Descripción (etiqueta corta)	Conteos No nulos (observado)	Tipo de variable (académico)
VENTA	Variable objetivo / Venta aceptada	14,153	Binaria / Categórica nominal (según codificación)

Fuente: Elaboración propia

En términos de composición, el dataset incluye variables tanto cuantitativas como categóricas, abarcando características clave como edad, nivel de riesgo, saldos, canal de contacto, clasificación comercial del cliente y tipo de producto. Este conjunto multidimensional proporciona una perspectiva amplia del comportamiento financiero y operativo de los clientes, garantizando que el análisis exploratorio y posterior modelado se fundamenten en un insumo robusto y representativo.

La base presenta además una adecuada cobertura temporal, reflejando interacciones ocurridas antes y durante campañas recientes, lo cual aporta vigencia al análisis. A nivel demográfico, se observa un equilibrio sustancial en género y una distribución etaria concentrada en edades laborales activas, mientras que en la dimensión comercial predomina la Banca Consumo como segmento principal, complementada por grupos específicos como Banca Más, Emprendedor y Empresarial.

La estructura del dataset permite trabajar con una unidad de análisis a nivel de cliente, facilitando el cruce entre atributos individuales y variables de comportamiento asociadas a la aceptación de ofertas. Este diseño habilita la identificación de patrones y relaciones relevantes entre características del cliente y su propensión a compra, sirviendo como fundamento para las etapas posteriores de estadística descriptiva, inferencia y modelado predictivo.

En conjunto, la estructura final del dataset constituye un insumo metodológicamente confiable, adecuado para la modelación supervisada y consistente con estándares institucionales de integridad y calidad de datos.

Análisis de la variable objetivo

Antes de avanzar hacia los análisis comparativos y el modelado supervisado, resulta fundamental examinar la distribución de la variable objetivo VENTA, que indica si el cliente aceptó (1) o no aceptó (0) el producto ofertado.

Conocer esta distribución permite identificar posibles desbalances en la clase positiva, un aspecto crítico en problemas de conversión comercial y un elemento que condiciona tanto la interpretación de los resultados como la selección de técnicas adecuadas durante el entrenamiento del modelo.

La distribución mostrada en siguiente tabla y representada visualmente en la ilustración 16 evidencia un fuerte desbalance entre las clases: aproximadamente 97.3% de los registros corresponden a clientes que no aceptaron la oferta (VENTA = 0), mientras que solo 2.7% representa casos de aceptación (VENTA = 1). Esta relación altamente desigual es característica de campañas comerciales y escenarios de marketing financiero, donde las tasas de conversión suelen ser bajas.

El gráfico complementa esta evidencia numérica al mostrar de manera clara la magnitud de la diferencia entre ambas clases, resaltando visualmente la predominancia de la clase negativa.

Tabla 19 Estadísticos descriptivos de variables numéricas variable VENTA

VENTA	Frecuencia Absoluta	Porcentaje (%)
0	13,767	97.27
1	386	2.73

Fuente: Elaboración propia

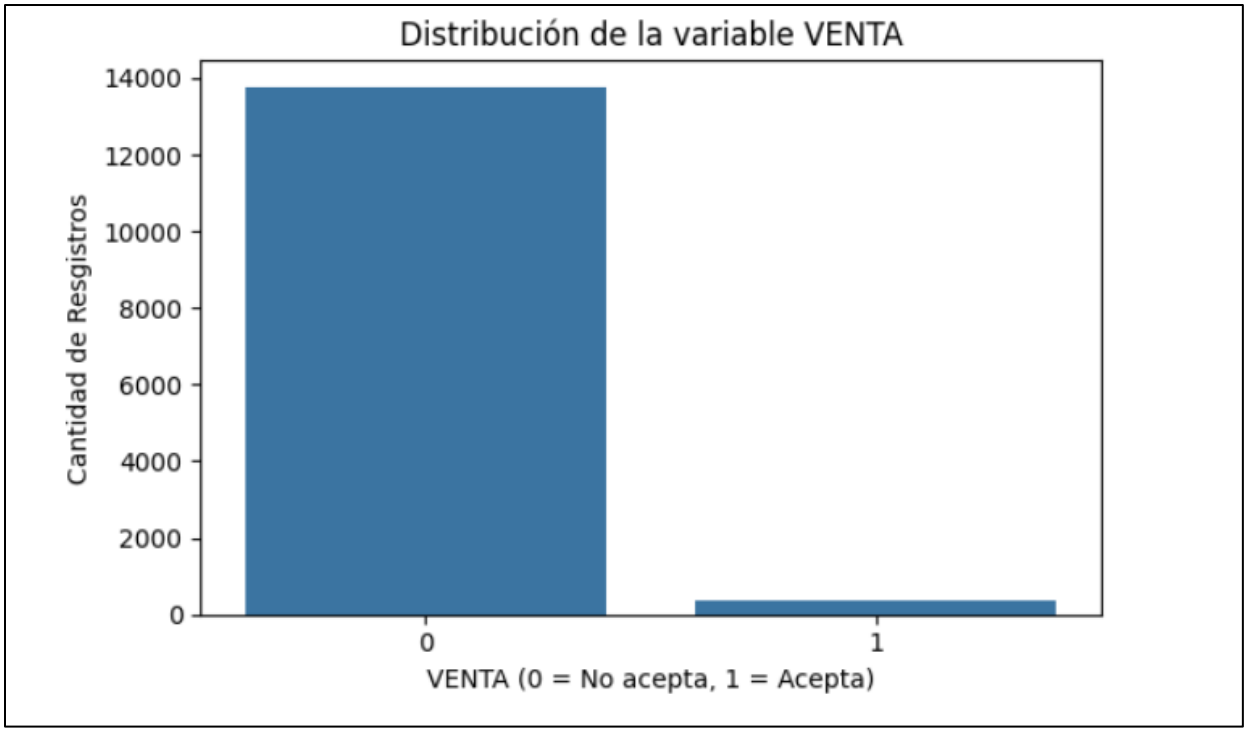

Ilustración 16 Distribución de la variable VENTA
Fuente: Elaboración propia

Dicho patrón sugiere que la variable objetivo presenta una estructura inherentemente desbalanceada que impactará directamente en el rendimiento de cualquier modelo predictivo si no se maneja adecuadamente.

Desde una perspectiva comercial, esta distribución es coherente con dinámicas reales del sector financiero, donde la mayoría de los clientes declina ofertas debido a restricciones económicas, baja relevancia del producto o saturación crediticia. De este modo, la baja tasa de respuesta no constituye una anomalía, sino un fenómeno esperado en campañas de alcance masivo.

Sin embargo, desde el punto de vista analítico, el desbalance en VENTA plantea desafíos importantes. Los modelos podrían tender a predecir mayoritariamente la clase negativa, logrando métricas aparentemente buenas en precisión general pero con bajo desempeño en la detección de los casos realmente positivos.

Por ello, este comportamiento anticipa la necesidad de incorporar técnicas específicas para el manejo de clases desbalanceadas (tales como oversampling, undersampling, uso de class weights o algoritmos robustos al desbalance) que permitan mejorar la sensibilidad del modelo hacia la clase minoritaria.

En síntesis, la distribución de la variable objetivo confirma que se está ante un problema de clasificación con fuerte desbalance, lo cual define el enfoque metodológico a seguir y será abordado en etapas posteriores del análisis predictivo.

Análisis descriptivo de variables numéricas

Con el fin de obtener una visión más profunda sobre la distribución de cada variable, se elaboró una tabla extendida con estadísticos avanzados como moda, varianza, percentiles (P25, P75), rango intercuartílico (IQR) y rango total. Este nivel de detalle permite identificar asimetrías, presencia de posibles valores extremos y patrones demográficos o financieros que serán de utilidad para el análisis posterior.

Tabla 20 Estadísticos descriptivos de variables numéricas

Variable	N	Media	Mediana	Moda	Desv. Est.	Varianza	Mín	P25	P75	IQR	Máx	Rango
EDAD	14,153	43	41	35	13	180	1	32	52	20	96	78
NUM_PASIVOS	14,153	1	1	1	1	1	0	1	2	1	10	10
SALDO_PASIVOS	14,153	3,073	5	–	21.4k	459 M	0	429	429	0	1M	1M
NUM_PRESTAMOS	14,153	0	0	0	0	0	0	0	0	0	5	5
FLAG_INTERBANCA	14,153	0	0	0	0	0	0	–	–	–	1	1
FLAG_SARA	14,153	0	0	0	0	0	0	–	–	–	1	1
FLAG_PENSION	14,153	0	0	0	0	0	0	–	–	–	1	1
FLAG_ID_TC	14,153	1	1	1	0	0	0	–	–	–	1	1
FLAG_ID_SEGURO	14,153	0	0	0	0	0	0	–	–	–	1	1
FLAG_REMESA	14,153	0	0	0	0	0	0	–	–	–	1	1
FLAG_TENGO	14,153	0	0	0	0	0	0	–	–	–	1	1
RIESGOACTUAL	14,153	69	69	56	13	178	0	60	78	18	100	70

Fuente: Elaboración propia

La Tabla 21 presenta un resumen estadístico de las principales variables numéricas del dataset, permitiendo identificar tendencias centrales, niveles de dispersión y posibles anomalías en la estructura de los datos.

En el caso de EDAD, la media de 43 años y una desviación estándar moderada reflejan una población adulta en etapa laboral activa, coherente con el base objetivo de productos financieros. La presencia de valores mínimos cercanos a 0 indica registros asociados a clientes jóvenes o atípicos, que más adelante se consideran dentro del proceso de depuración y validación.

La variable SALDO_PASIVOS muestra una marcada diferencia entre la media y la desviación estándar, evidenciando una fuerte asimetría y la presencia de valores extremadamente altos. Este comportamiento es habitual en productos de captación, donde la distribución suele estar dominada por saldos bajos mientras una fracción muy pequeña concentra montos elevados. Este patrón sugiere la necesidad de tratamientos adecuados para análisis posteriores, particularmente en modelado y gráficos, debido al riesgo potencial de que los valores extremos distorsionen las estimaciones.

Por su parte, RIESGOACTUAL presenta una media de 68.67 con variabilidad moderada, lo cual es consistente con un perfil crediticio heterogéneo. La amplitud del rango indica que conviven clientes de riesgo robusto con otros que presentan perfiles menos favorables, una característica clave que influirá tanto en la interpretación de patrones comerciales como en el peso predictivo que esta variable podría tener en el modelo final.

Finalmente, la variable objetivo VENTA confirma un fuerte desbalance, con solo 2.7% de respuestas positivas. Este patrón es típico en campañas de conversión financiera y subraya la necesidad de aplicar técnicas adecuadas durante el modelado para manejar la desproporción entre clases. En conjunto, la información de esta tabla proporciona un primer diagnóstico cuantitativo del dataset, permitiendo anticipar decisiones metodológicas e identificar variables que podrían requerir transformaciones o tratamientos especiales en etapas posteriores.

Análisis Gráfico-descriptivo de variables numéricas

Para complementar el análisis tabular, se generaron representaciones visuales que permiten observar la distribución de las variables numéricas. Los histogramas permiten identificar simetrías, colas largas y la presencia de concentraciones atípicas, mientras que los boxplots permiten observar dispersión y valores extremos.

Adicionalmente, los boxplots segmentados por VENTA permiten identificar diferencias potenciales entre los clientes que aceptaron versus los que no aceptaron el producto.

La distribución presentada en la Ilustración 17, acompañada de su tabla descriptiva correspondiente, muestra que la variable EDAD se concentra principalmente en los rangos de 30 a 50 años, reflejando un perfil demográfico alineado con la población económicamente activa que típicamente participa en campañas de productos financieros.

El histograma evidencia esta agrupación central, mientras que el boxplot segmentado por la variable objetivo VENTA indica que tanto los clientes que aceptan como los que rechazan la oferta presentan patrones muy similares en términos de dispersión y mediana. Este comportamiento sugiere que la edad, de manera aislada, no constituye un factor diferenciador significativo en la decisión de compra.

Desde la perspectiva comercial, esto implica que segmentar campañas únicamente por grupos de edad probablemente no incremente la efectividad de la conversión, aunque la sobrerrepresentación de edades medias puede señalar oportunidades de exploración en segmentos jóvenes o mayores.

En el contexto del modelado predictivo, la variable EDAD aporta información contextual valiosa, pero su capacidad discriminativa es limitada; su relevancia tiende a aumentar únicamente cuando se combina con características financieras o de comportamiento.

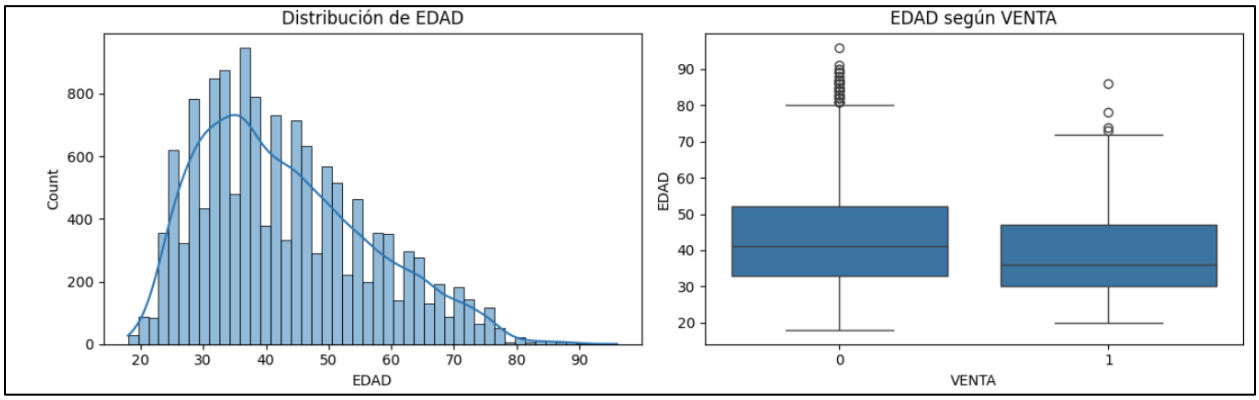

Ilustración 17 Variable *EDAD según VENTA*
Fuente: Elaboración propia

Por ello, su aporte principal dentro del modelo será complementario y no decisivo.

Por su parte a variable NUM_PRESTAMOS representa la cantidad de préstamos vigentes que mantiene cada cliente al momento de la campaña. Analizar esta variable permite comprender el nivel de vinculación crediticia y explorar si una mayor o menor carga financiera influye en la probabilidad de aceptar nuevos productos.

La Ilustración 18 evidencia que la distribución de NUM_PRESTAMOS es marcadamente

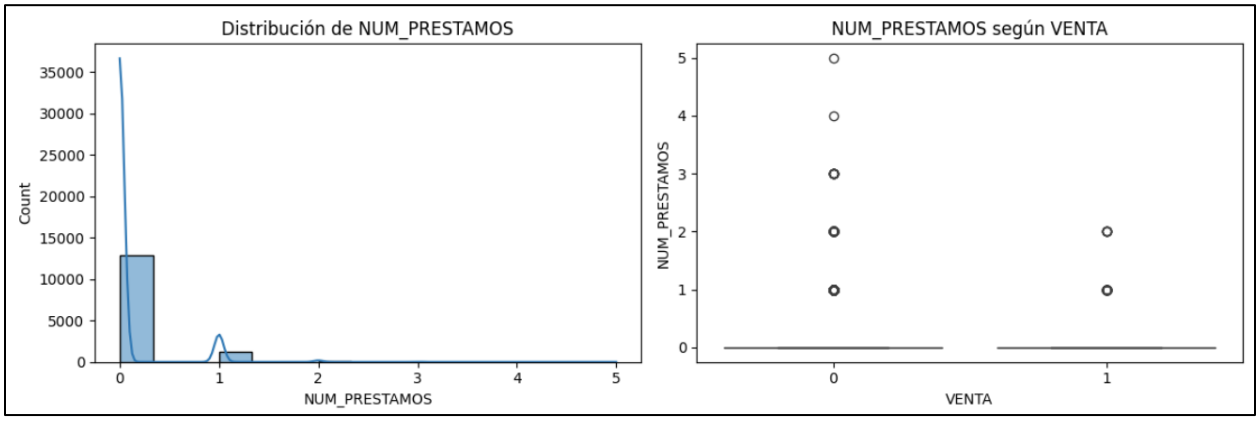

Ilustración 18 Histograma y boxplot de NUM_PRESTAMOS según VENTA
Fuente: Elaboración propia

Asimétrica, con la mayoría de los clientes concentrados en valores bajos, mientras que una fracción pequeña mantiene múltiples préstamos, lo que genera una cola prolongada hacia la derecha. Este patrón es consistente con la estructura típica de cartera bancaria, donde predominan clientes con relaciones crediticias simples y solo un segmento reducido presenta niveles de endeudamiento más altos.

Al observar el boxplot segmentado por VENTA, se aprecia que tanto los clientes que aceptan como los que rechazan la oferta comparten medianas y rangos intercuartílicos similares, lo cual sugiere que, en promedio, la cantidad de préstamos vigentes no diferencia de manera contundente la propensión de compra. No obstante, los valores extremos más altos se observan principalmente entre quienes no aceptan la oferta, lo que podría indicar que clientes ya altamente endeudados muestran menor disposición a adquirir productos adicionales.

Desde la óptica comercial, NUM_PRESTAMOS funciona como un indicador relevante del nivel de compromiso financiero del cliente y puede reflejar saturación crediticia o límites de capacidad de pago. Aunque el análisis descriptivo no muestra diferencias drásticas, sí sugiere que los extremos superiores pueden tener un comportamiento particular que conviene monitorear.

En términos predictivos, esta variable aporta información sobre la relación crediticia actual y puede contribuir al modelo especialmente cuando se combina con otros indicadores financieros (como saldos, pasivos o nivel de riesgo). Por ello, NUM_PRESTAMOS es una variable que, aunque no discriminante de manera aislada, sí añade valor al modelo al capturar perfiles de endeudamiento que podrían influir en la aceptación de futuros productos.

La variable NUM_PASIVOS representa la cantidad de productos pasivos (cuentas de ahorro, depósitos, instrumentos de captación, entre otros) que mantiene el cliente. Su análisis permite comprender el nivel de vinculación del cliente con el banco desde la perspectiva del ahorro o tenencia de cuentas, así como evaluar si esta relación influye en la aceptación de nuevos productos.

La distribución mostrada en la 19 revela que la mayoría de los clientes posee un número reducido de productos pasivos, con una fuerte concentración en valores bajos y una dispersión mínima hacia niveles más altos. Este comportamiento es común en instituciones financieras donde los clientes suelen mantener una o pocas cuentas principales, mientras que solo un grupo reducido gestiona múltiples productos de captación.

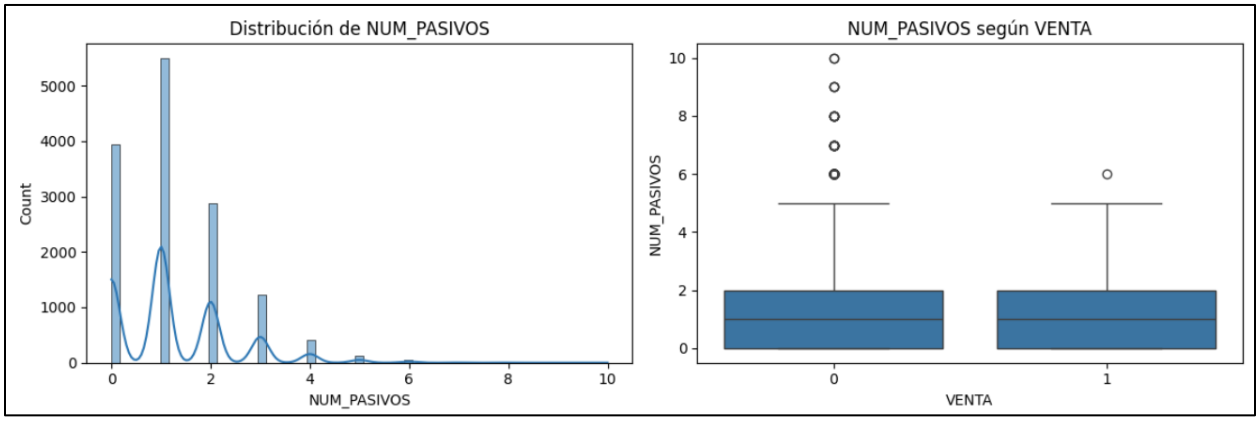

Ilustración 19 Histograma y boxplot de NUM_PASIVOS según VENTA

Fuente*:* Elaboración propia

Al segmentar por la variable objetivo VENTA, los boxplots indican que tanto quienes aceptaron como quienes no aceptaron la oferta presentan medianas y rangos intercuartílicos similares. Esto sugiere que, en promedio, el número de pasivos no distingue de manera significativa la disposición a adquirir el producto ofertado. A pesar de ello, se observa que los valores altos de NUM_PASIVOS, aunque poco frecuentes, tienden a agruparse ligeramente en el grupo que sí acepta la oferta, lo cual podría reflejar que los clientes con mayor diversificación de productos tienen una relación más profunda con el banco.

En términos comerciales, este comportamiento apunta a que la tenencia de múltiples productos pasivos puede estar asociada con un mayor nivel de confianza o fidelización, pero no constituye por sí sola un determinante principal de aceptación. El grueso de los clientes, al tener pocos productos pasivos, mantiene un patrón estable que limita la capacidad de esta variable para segmentar de manera efectiva.

Desde la perspectiva del modelado predictivo, NUM_PASIVOS aporta información sobre la amplitud de la relación del cliente con el banco. Su valor predictivo aislado es moderado debido a la baja variabilidad en la muestra; sin embargo, combinado con otras variables (como SALDO_PASIVOS, número de préstamos o antigüedad del cliente), puede ayudar a identificar perfiles más consolidados que podrían influir en la probabilidad de aceptación. En ese sentido, su utilidad es complementaria y se fortalece dentro del conjunto de predictores financieros.

Por otro lado, la variable FLAG_INTERBANCA indica si el cliente posee o no productos asociados al servicio Interbanca. Su análisis permite identificar el nivel de adopción de este servicio y evaluar si su utilización está relacionada con la aceptación de la oferta comercial evaluada.

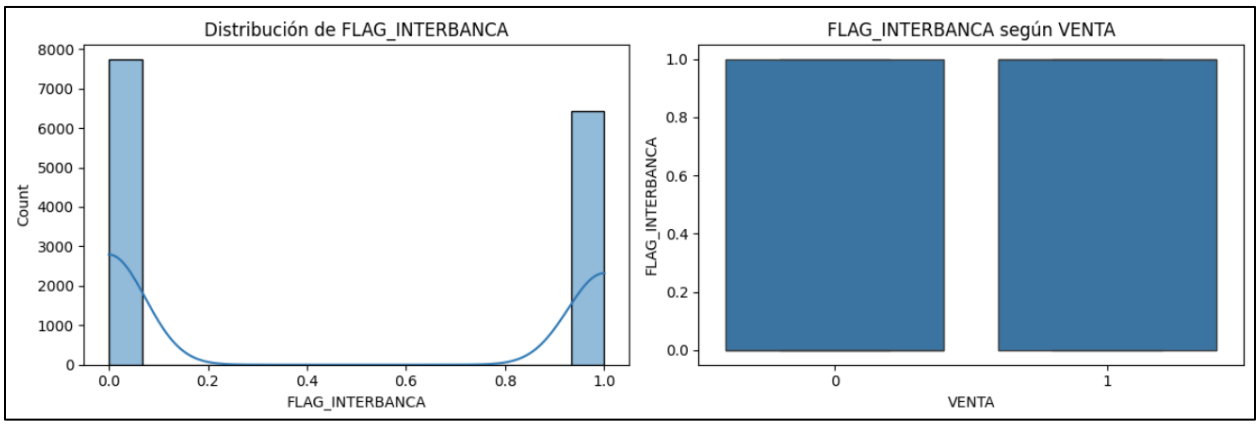

Ilustración 20 Histograma y boxplot de FLAG_INTERBANCA según VENTA

Fuente: Elaboración propia

Tal como se aprecia en la Ilustración 19, la distribución de FLAG_INTERBANCA se encuentra fuertemente concentrada en el valor 0, lo que indica que la mayoría de los clientes no posee este servicio. La presencia del valor 1 es mínima, reflejando un nivel de adopción bajo dentro de la cartera analizada. Este comportamiento genera una distribución prácticamente binaria y altamente desbalanceada, típica de variables bandera asociadas a servicios específicos con baja penetración.

Al segmentar por VENTA, los boxplots no muestran diferencias significativas entre clientes que aceptan y rechazan la oferta. Esto sugiere que la tenencia del servicio Interbanca no está asociada de manera directa con la probabilidad de aceptación. Dado el bajo número de casos positivos (valor 1), cualquier posible variación queda prácticamente difuminada dentro del grupo mayoritario.

Desde la perspectiva comercial, este hallazgo indica que Interbanca no es un servicio determinante en la decisión del cliente respecto a la oferta realizada; su adopción responde más a necesidades específicas y no parece influir en el comportamiento general de compra. Sin embargo, su baja penetración podría considerarse una oportunidad para fortalecer estrategias de educación financiera o promoción de servicios digitales complementarios.

En cuanto al modelo predictivo, debido a su distribución altamente sesgada, FLAG_INTERBANCA es una variable cuyo aporte aislado será limitado. No obstante, puede aportar información marginal al capturar patrones de uso de servicios específicos dentro del banco. Su inclusión es adecuada como parte del conjunto de variables de comportamiento, aunque no se espera que tenga un peso relevante en la capacidad discriminativa del modelo.

La variable FLAG_SARA identifica si el cliente utiliza o posee el servicio SARA. Este tipo de variable bandera permite evaluar el nivel de adopción de servicios específicos y determinar si su presencia se relaciona con la decisión del cliente de aceptar la oferta comercial.

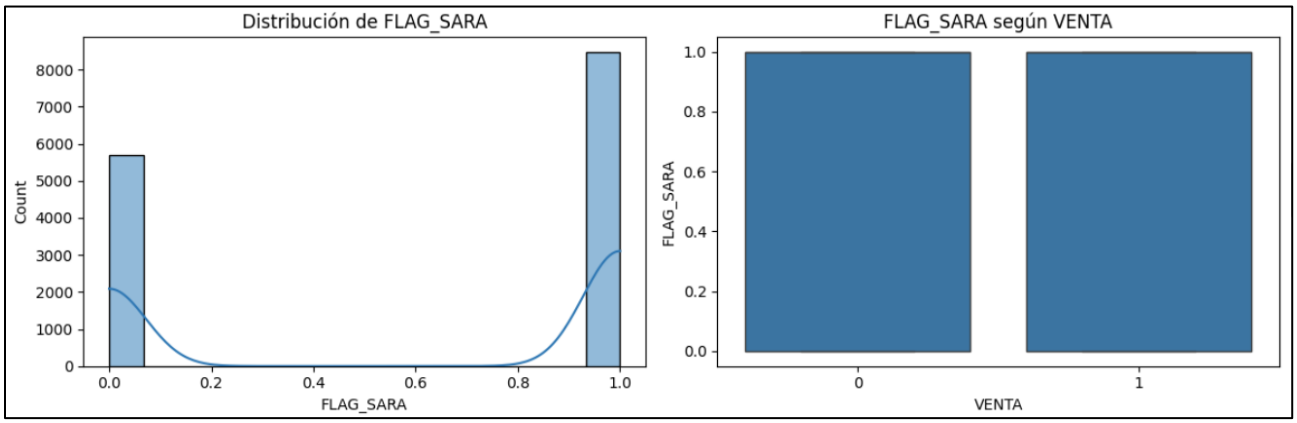

Ilustración 21 Histograma y boxplot de FLAG_SARA según VENTA
Fuente: Elaboración propia

La Ilustración 20 muestra que la distribución de FLAG_SARA se encuentra fuertemente concentrada en el valor 0, reflejando que la gran mayoría de los clientes no cuenta con el servicio SARA. La proporción de clientes con valor 1 es muy baja, lo que evidencia un nivel de adopción limitado dentro de la base analizada. Este comportamiento es característico de variables bandera vinculadas a servicios especializados, cuya penetración suele ser reducida en comparación con productos financieros tradicionales.

Al segmentar por VENTA, los boxplots indican patrones casi idénticos entre quienes aceptan la oferta y quienes no. Esto confirma que la presencia del servicio SARA no genera

diferencias sustantivas en el comportamiento de compra. La baja frecuencia del valor 1 limita la capacidad de esta variable para discriminar entre grupos, ya que el gran peso del valor 0 absorbe cualquier variación potencial. En otras palabras, los clientes que poseen SARA no presentan una tendencia notablemente distinta frente a la aceptación del producto en comparación con el resto de la base.

Desde una perspectiva comercial, este resultado sugiere que el uso de SARA no es un factor determinante en la evaluación de ofertas financieras, probablemente porque su adopción está asociada a necesidades específicas y no a un comportamiento transversal de la cartera. No obstante, su baja penetración podría señalar una oportunidad de reforzar servicios complementarios, especialmente si se pretende aumentar la profundidad de la relación del cliente con el banco.

En el contexto del modelo predictivo, FLAG_SARA constituye una variable de comportamiento de bajo impacto individual. Su distribución fuertemente sesgada y la escasa presencia del valor positivo reducen su contribución directa al poder de clasificación del modelo. Aun así, conservarla en el conjunto de predictores resulta adecuado, pues podría aportar información marginal cuando interactúa con otras variables de uso de servicios o de vinculación digital, aunque no se espera que tenga un peso relevante dentro de la importancia de características.

La variable FLG_PENSION indica si el cliente posee o recibe algún tipo de producto o beneficio relacionado con pensiones dentro del banco. Su análisis permite observar la penetración de este servicio en la base de clientes y evaluar si esta característica se asocia con la aceptación de la oferta comercial.

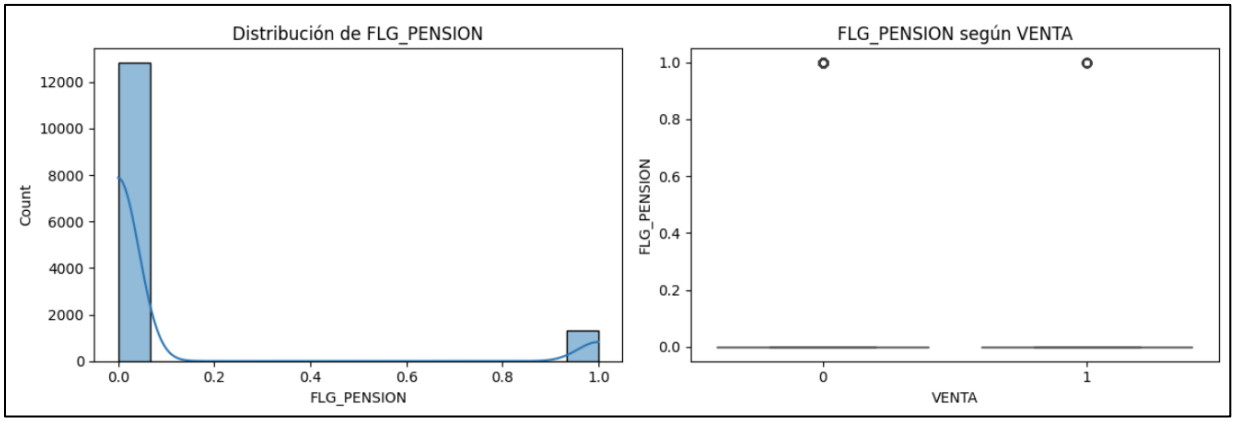

Ilustración 22 Variable FLG_PENSION según VENTA

Fuente: Elaboración propia

En la Ilustración 21 se observa que la distribución de FLG_PENSION está fuertemente

concentrada en el valor 0, lo que indica que la mayoría de los clientes no cuenta con un producto o beneficio de pensión dentro del banco. El valor 1 aparece como un caso minoritario, lo que sugiere que este servicio tiene una adopción reducida en la base. Al igual que otras variables bandera, este patrón genera una distribución muy desbalanceada y de escasa variabilidad.

Cuando la variable se analiza segmentada por VENTA, los boxplots reflejan que no existen diferencias relevantes entre quienes aceptan la oferta y quienes no. La presencia del servicio de pensión no modifica de forma apreciable el comportamiento de compra, dado que los clientes con valor 1 constituyen una proporción muy pequeña que no altera el patrón general, dominado por los clientes que no poseen dicho servicio.

Desde la perspectiva comercial, este resultado sugiere que la afiliación a pensión no es un factor relevante para explicar la aceptación de la oferta. El hecho de que la mayoría de los clientes no posea este beneficio indica que su uso depende de características específicas, no necesariamente ligadas al interés por productos adicionales del banco. Sin embargo, la baja penetración podría señalar oportunidades para impulsar productos complementarios dirigidos a segmentos con necesidades previsionales.

En el contexto del modelado predictivo, la variable FLG_PENSION aporta información de vinculación, pero su capacidad discriminativa individual es limitada. La baja presencia del valor positivo reduce su efecto directo dentro del modelo y es probable que su importancia como predictor sea baja. No obstante, es conveniente mantenerla en el conjunto de variables, dado que podría capturar perfiles demográficos o laborales específicos cuando interactúa con otras características del cliente.

La variable FLAG_ID_TC indica si el cliente posee identificación activa asociada a algún producto de tarjeta de crédito dentro del banco. Este tipo de variable permite evaluar la penetración de productos de crédito revolving y analizar si la tenencia de una tarjeta de crédito influye en la probabilidad de aceptar la oferta comercial.

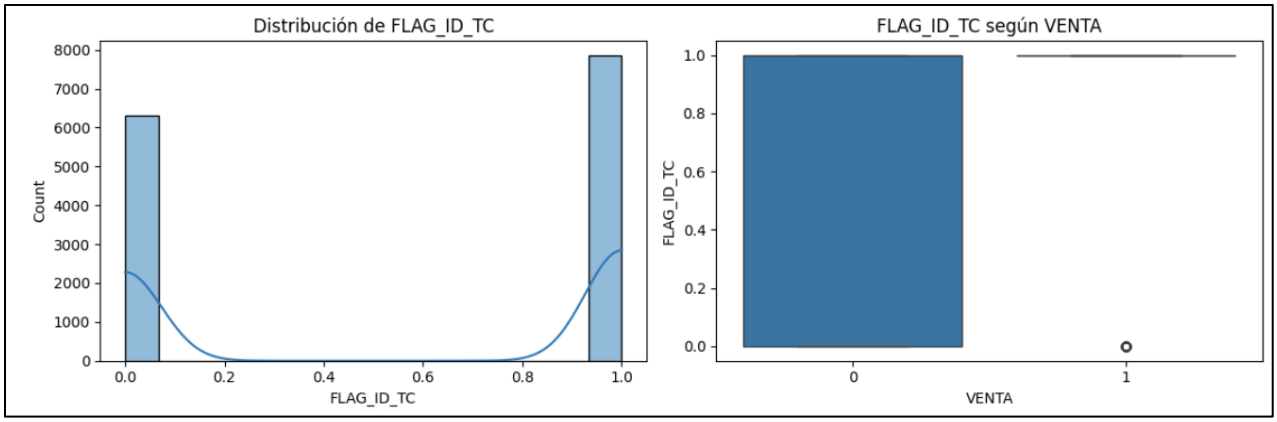

Ilustración 23 Histograma y boxplot de FLAG_ID_TC según VENTA

Fuente: Elaboración propia

La Ilustración 23 muestra que la distribución de FLAG_ID_TC está concentrada principalmente en el valor 0, lo que indica que la mayoría de los clientes no cuenta con una tarjeta de crédito registrada en el sistema. El valor 1 se presenta en una proporción menor, aunque con mayor frecuencia que otras banderas previamente analizadas, evidenciando una adopción un poco más amplia pero aún limitada dentro de la base de clientes.

Al segmentar la variable por VENTA, los boxplots indican patrones similares entre los grupos que aceptan y rechazan la oferta. No se observan diferencias notorias en la proporción relativa del valor 1 entre ambos grupos, lo cual sugiere que la posesión de una tarjeta de crédito no genera un efecto determinante sobre la decisión de aceptar el producto ofertado. Este comportamiento puede deberse a que la presencia de una tarjeta de crédito refleja cierto nivel de vinculación financiera, pero no necesariamente implica interés por productos adicionales específicos de la campaña.

Desde la perspectiva comercial, tener o no una tarjeta de crédito parece no modificar de manera significativa la probabilidad de conversión. Esto puede indicar que los productos promocionados no están directamente relacionados con necesidades propias del segmento de tarjetas, o que los clientes con tarjeta ya cuentan con un nivel de carga financiera que no incentiva nuevas adquisiciones.

En el contexto del modelado predictivo, FLAG_ID_TC aporta información útil sobre la vinculación crediticia del cliente y podría capturar patrones de comportamiento financiero en combinación con otras variables (por ejemplo, número de préstamos, riesgo actual o historial de contacto), pero su aporte aislado es moderado. Su distribución ligeramente menos desbalanceada que otras banderas puede darle una importancia marginalmente mayor en el modelo, aunque sin

convertirse en un predictor principal.

La variable FLAG_ID_SEGURO identifica si el cliente posee algún seguro registrado con el banco. Este tipo de indicador permite explorar el nivel de adopción de productos complementarios y evaluar si la tenencia de un seguro tiene relación con la aceptación de la oferta financiera analizada.

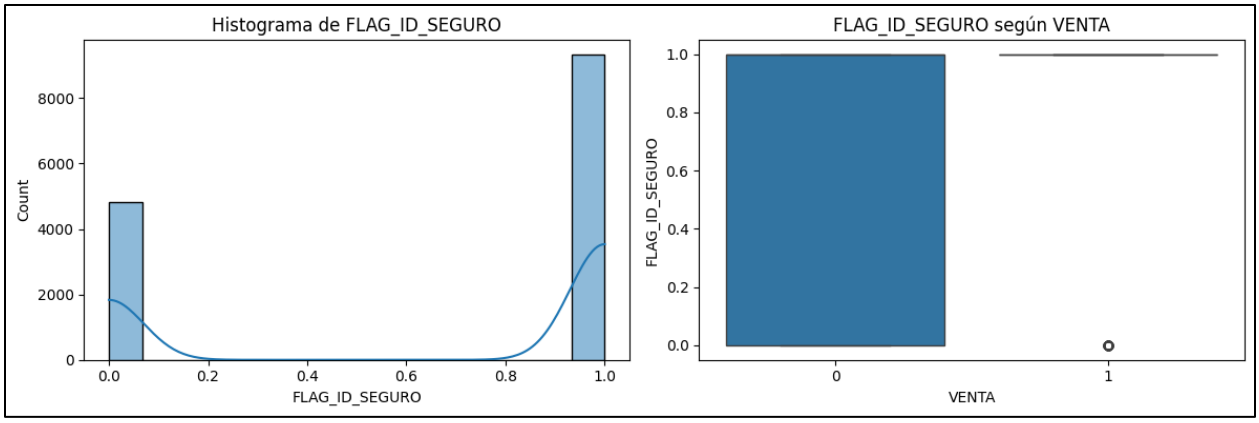

Ilustración 24 Histograma y boxplot de FLAG_ID_SEGURO según VENTA
Fuente*:* Elaboración propia

La Ilustración 24 muestra que la variable FLAG_ID_SEGURO presenta una distribución fuertemente concentrada en el valor 0, indicando que la mayoría de los clientes no cuenta con un seguro contratado. La proporción de clientes con valor 1 es relativamente baja, lo cual revela una adopción limitada del producto dentro de la base analizada. Este patrón es típico de servicios complementarios cuya penetración suele ser focalizada y no transversal a toda la cartera.

Al segmentar la variable por VENTA, los boxplots ponen de manifiesto que no existen diferencias significativas entre quienes aceptan la oferta y quienes no. La presencia del valor 1 no muestra una tendencia clara que pueda asociarse a una mayor o menor probabilidad de aceptación. Esta ausencia de variación se explica tanto por la baja frecuencia del valor positivo como por el hecho de que los seguros suelen responder a necesidades específicas que no necesariamente determinan el interés por otros productos del banco.

Desde una perspectiva comercial, este comportamiento indica que la tenencia de seguros no es un factor relevante para predecir la respuesta del cliente frente a la oferta. El limitado nivel de adopción también sugiere que existe una oportunidad para fortalecer la estrategia comercial orientada a incrementar la colocación de seguros, especialmente si se busca aumentar la diversificación de productos por cliente.

En términos de modelado predictivo, FLAG_ID_SEGURO aporta información adicional sobre el nivel de relacionamiento del cliente con el banco, pero su capacidad discriminativa aislada es baja debido al desbalance en la distribución. Como en otras variables bandera, su valor radica más en su interacción con otras características (por ejemplo, riesgos, número de préstamos o productos pasivos) que en su efecto directo sobre la variable objetivo. Su inclusión es adecuada, aunque su importancia dentro del modelo será probablemente marginal.

La variable FLAG_REMESA indica si el cliente recibe o gestiona remesas a través del banco. Este tipo de indicador es relevante para identificar flujos de ingreso no salariales y niveles particulares de actividad financiera que podrían influir en el comportamiento comercial del cliente.

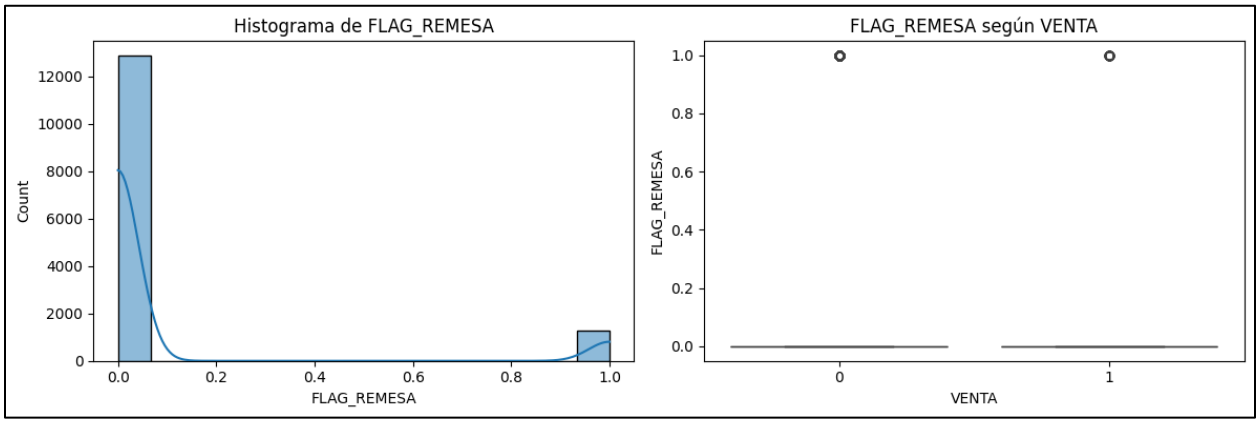

Ilustración 25 Histograma y boxplot de FLAG_ID_REMESA según VENTA

Fuente: Elaboración propia

La Figura 25 refleja que la distribución de FLAG_REMESA está fuertemente concentrada en el valor 0, evidenciando que la mayoría de los clientes no recibe remesas por medio de la institución bancaria. El valor 1 aparece únicamente en una proporción reducida, lo que indica que este servicio tiene un nivel de adopción limitado dentro de la base analizada. Este patrón es coherente con la estructura demográfica típica en campañas bancarias, donde solo un segmento específico de clientes depende de remesas como fuente de ingresos.

Al segmentar la variable por VENTA, los boxplots muestran que tanto los clientes que aceptaron la oferta como los que la rechazaron presentan patrones prácticamente idénticos. No se observa una tendencia clara que vincule la recepción de remesas con una mayor o menor propensión a aceptar el producto. La baja frecuencia del valor 1 limita cualquier variación observable, manteniendo la distribución estable en ambos grupos.

Desde la perspectiva comercial, este comportamiento sugiere que el hecho de que un

cliente reciba remesas no influye de manera relevante en su decisión frente a la oferta promocionada. Las remesas suelen representar un comportamiento financiero específico, pero no necesariamente implican un interés directo por productos adicionales o vinculados al crédito. No obstante, este segmento podría representar una oportunidad para campañas focalizadas que atiendan sus necesidades particulares.

En el ámbito del modelado predictivo, FLAG_REMESA posee un valor discriminativo limitado debido a su distribución altamente desbalanceada. Su aporte aislado al rendimiento del modelo será marginal; sin embargo, su inclusión continúa siendo pertinente, pues podría capturar patrones socioeconómicos o comportamentales cuando interactúa con otras variables financieras. Su importancia dentro del modelo probablemente será baja, pero contribuye al entendimiento integral del perfil del cliente.

La variable FLAG_TENGO identifica si el cliente posee un servicio registrado bajo la categoría *TENGO*. Este tipo de bandera permite evaluar la adopción de servicios específicos y determinar si su presencia tiene alguna relación con la aceptación de la oferta comercial.

La Figura 26 muestra que la variable FLAG_TENGO presenta una distribución altamente concentrada en el valor 0, lo cual evidencia que la mayoría de los clientes no posee este servicio. El valor 1 aparece en una proporción limitada, reflejando una adopción reducida dentro de la base de datos. Este patrón se asemeja al observado en otras variables bandera de baja penetración, donde la variabilidad es mínima y el comportamiento está dominado por la ausencia del servicio.

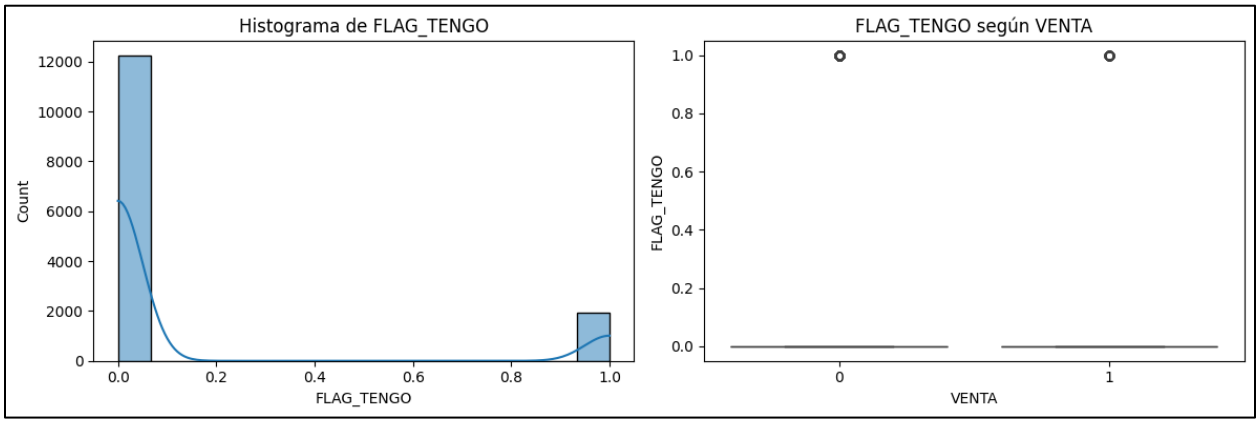

Ilustración 26 Histograma y boxplot de FLAG_TENGO según VENTA

Fuente*:* Elaboración propia

Al segmentar la variable por VENTA, los boxplots permiten observar que no existen diferencias relevantes en la distribución entre los clientes que aceptaron la oferta y los que no lo

hicieron. La presencia del valor 1 no muestra una inclinación clara hacia uno u otro grupo, lo que indica que la tenencia del servicio TENGO no influye de forma sustancial en la probabilidad de aceptación. Este resultado es coherente con la baja frecuencia del valor positivo, la cual dificulta que la variable aporte patrones discriminativos significativos.

Desde el punto de vista comercial, la limitada adopción del servicio sugiere que TENGO no es un producto ampliamente utilizado entre la cartera de clientes y que su presencia no condiciona el interés por nuevas ofertas. Este comportamiento también podría señalar oportunidades para estrategias de promoción orientadas a incrementar la penetración del servicio o explorar segmentos donde su adopción pudiera ser más relevante.

En el contexto del modelado predictivo, FLAG_TENGO aporta información marginal sobre la relación del cliente con ciertos servicios complementarios. Su distribución desbalanceada y su bajo nivel de variación reducen su capacidad de contribuir de forma significativa al poder predictivo del modelo. No obstante, su inclusión sigue siendo metodológicamente adecuada, ya que podría capturar patrones específicos cuando se analiza en combinación con otras variables de carácter financiero o demográfico.

La variable RIESGOACTUAL representa la clasificación interna de riesgo crediticio asignada al cliente según los criterios de la institución financiera. Su análisis es fundamental, ya que refleja el nivel de salud financiera del cliente y puede influir en la probabilidad de aceptar nuevos productos, especialmente aquellos relacionados con crédito.

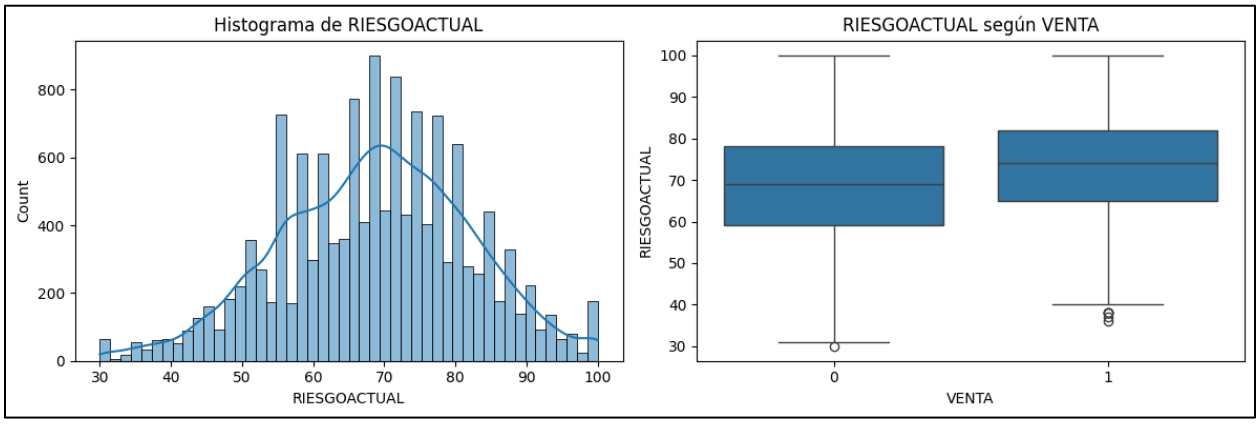

Ilustración 27 Histograma y boxplot de RIESGOACTUAL según VENTA

Fuente: Elaboración propia

En la Figura 27 se observa que la distribución de RIESGOACTUAL presenta una alta concentración en los valores bajos, especialmente en la categoría 1, que representa el riesgo más

saludable dentro de la cartera. Este patrón es coherente con la estructura típica de campañas bancarias, donde se prioriza contactar clientes con buen comportamiento crediticio, tanto por elegibilidad como por mayor probabilidad de conversión. A medida que el nivel de riesgo aumenta, la frecuencia disminuye considerablemente, generando una cola hacia la derecha compuesta por clientes con mayor probabilidad de incumplimiento o historial crediticio más débil.

Al segmentar por VENTA, los boxplots muestran que los clientes que aceptan la oferta tienden a concentrarse levemente más en los niveles de riesgo más bajos, mientras que los valores más altos de riesgo aparecen con mayor frecuencia en el grupo que no acepta. Aunque la diferencia no es extrema, sí es más visible que en variables anteriores, lo que sugiere que el riesgo crediticio sí tiene cierta relación con la probabilidad de aceptación del producto.

Esto es consistente con el comportamiento esperado: los clientes con mejor perfil financiero suelen estar más abiertos a adquirir productos adicionales, mientras que aquellos con niveles de riesgo más elevados podrían presentar restricciones operativas, menor capacidad de endeudamiento o menor disposición ante nuevas obligaciones financieras.

Desde la perspectiva comercial, este hallazgo confirma que los clientes con mejor riesgo son un segmento más receptivo a las ofertas, por lo que representan una población prioritaria en términos de eficiencia comercial. Por el contrario, los clientes con riesgo más alto podrían requerir abordajes diferenciados, como asesoría, regularización o productos alternativos.

En términos de modelado predictivo, RIESGOACTUAL es una variable clave: su mayor variabilidad y su asociación más clara con la aceptación la convierten en un predictor con potencial peso significativo dentro del modelo. A diferencia de las variables bandera, esta variable sí captura un rasgo estructural del cliente que influye directamente en su comportamiento financiero y en su relación con el banco. Por ello, se espera que tenga una importancia superior dentro del conjunto de características utilizadas para entrenar el modelo.

Análisis descriptivo de variables categoricas según venta

Las variables categóricas del dataset permiten caracterizar la composición de la base de clientes según atributos demográficos, operativos y comerciales. Analizar su distribución es fundamental para comprender la estructura general de la población bajo estudio y para identificar categorías dominantes, segmentos subrepresentados y patrones que podrían influir en la respuesta

a la oferta.

Las tablas y figuras siguientes presentan las frecuencias absolutas, porcentajes y distribuciones por VENTA de variables como estado civil, canal de atención, resultado de contacto, tipo de producto, así como características geográficas y comerciales del cliente. Esta información sirve como base para el análisis posterior de cada variable y permite anticipar tendencias relevantes para el modelo predictivo.

Tabla 21 Estadísticos descriptivos de variables numéricas variable ESTADO_CIVIL

ESTADO_CIVIL	Frecuencia Absoluta	Porcentaje (%)	Frecuencia Relativa
SINGLE	6,863	48.49	0.49
MARRIED	5,438	38.42	0.38
LEGALLY	1,660	11.73	0.18
WIDOW	115	0.81	0.01
DIVORCED	77	0.54	0.01

Fuente*:* Elaboración propia

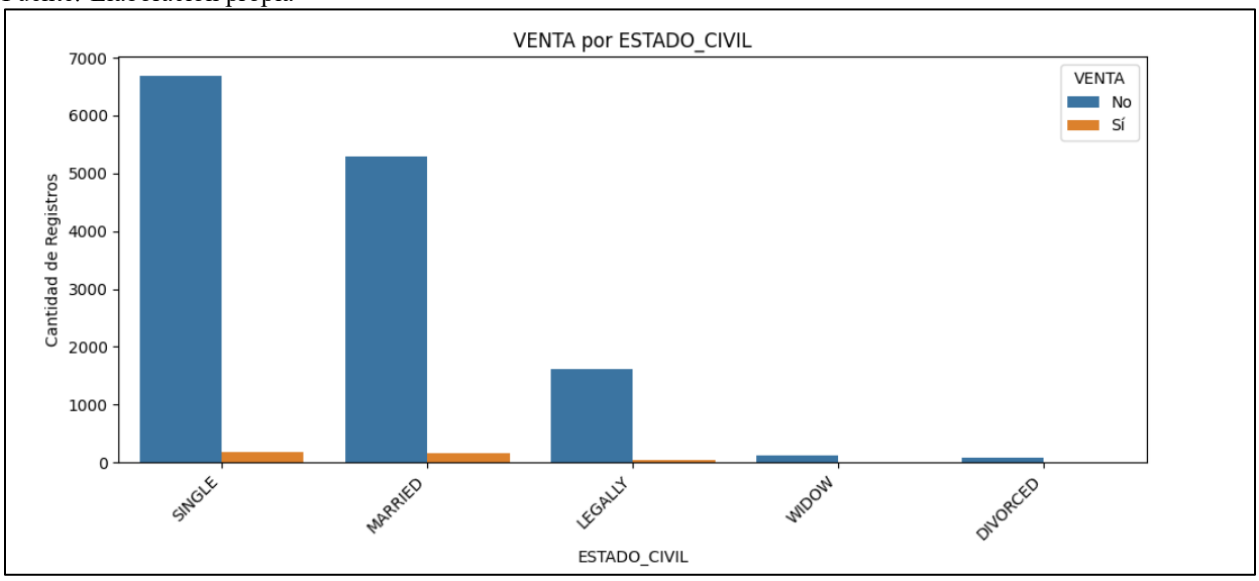

Ilustración 28 Distribución de *VENTA POR ESTADO_CIVIL*

Fuente*:* Elaboración propia

La Tabla 22 muestra que la mayor parte de los clientes se concentra en las categorías SINGLE (48.49%) y MARRIED (38.42%), las cuales representan casi el 87% de la base analizada. Las demás categorías (LEGALLY, WIDOW y DIVORCED) presentan frecuencias significativamente menores, propias de segmentos más reducidos de la población. Esta distribución es coherente con el perfil demográfico predominante en campañas financieras, donde los estados civiles más comunes son aquellos vinculados a población laboralmente activa.

La Figura 28 refleja gráficamente este comportamiento, evidenciando la fuerte concentración en las dos categorías principales. Desde una perspectiva comercial, esta predominancia implica que las estrategias de oferta están orientadas, directa o indirectamente, a clientes con mayor estabilidad relacional, lo cual puede influir en su comportamiento crediticio y en su capacidad de adquirir productos adicionales.

En términos de modelado, el estado civil no presenta una variabilidad suficiente como para ser un predictor determinante de la aceptación, especialmente dadas las diferencias mínimas observadas entre categorías minoritarias. Su contribución principal será contextual y podría aportar valor solo al combinarse con otras características sociodemográficas o financieras del cliente.

No obstante, su inclusión dentro del modelo se justifica para mantener la completitud del perfil del cliente y para explorar posibles relaciones no lineales que podrían surgir durante la fase de entrenamiento.

La variable CANAL identifica el medio por el cual se registró la interacción del cliente, ya sea FÍSICO o DIGITAL. Su análisis permite entender la preferencia de los clientes al interactuar con los servicios del banco y evaluar si esta preferencia influye en la probabilidad de aceptar la oferta.

Tabla 22 Estadísticos descriptivos de variables numéricas variable CANAL

CANAL	0 (No Acepta)	1 (Acepta)
DIGITAL	97.30	2.70
FISICO	97.26	2.74

Fuente*:* Elaboración propia

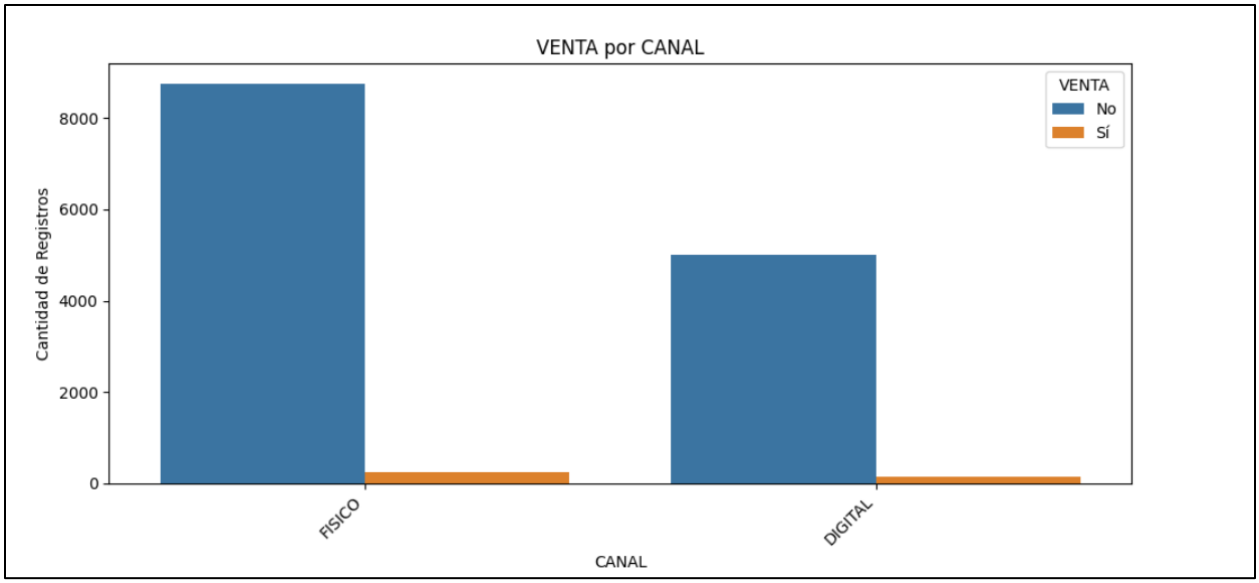

Ilustración 29 Distribución de CANAL POR VENTA

Fuente*:* Elaboración propia

La tabla anterior muestra que la mayoría de los clientes interactúa a través del canal FÍSICO (63.63%), mientras que el DIGITAL representa el 36.37% de los registros. Esta distribución indica una preferencia considerable por los puntos de atención presenciales, lo cual es consistente con tendencias observadas en segmentos masivos de la banca, donde los clientes suelen recurrir al canal físico para consultas, trámites o servicios personalizados.

Al segmentar por VENTA, se observa que la tasa de aceptación es prácticamente igual en ambos canales:

- Digital: 97.30% no acepta, 2.70% acepta
- Físico: 97.26% no acepta, 2.74% acepta

La Figura 29 confirma visualmente que no existe una diferencia sustantiva entre los canales en términos de conversión. Esto sugiere que la aceptación de la oferta no depende del medio de contacto, sino de características más profundas del cliente o de su situación financiera.

Desde la perspectiva comercial, este comportamiento indica que las campañas pueden ejecutarse de manera uniforme en ambos canales sin esperar variaciones significativas en la tasa de éxito. El predominio del canal físico podría deberse a hábitos de los clientes, estructura operativa del banco o preferencia por atención personalizada, pero esto no se traduce en diferencias en la aceptación del producto.

En términos de modelado, la variable CANAL no parece actuar como un predictor fuerte

o discriminante de la respuesta del cliente. Su importancia en el modelo será probablemente baja, aunque sigue siendo útil mantenerla para capturar posibles interacciones con otras variables, como la edad o la clasificación comercial, especialmente si se presentan patrones distintos en subgrupos específicos.

La variable RESULTADO_CONTACTO describe el estado de la interacción previa del banco con el cliente, agrupando categorías como CONTACTO, NO CONTACTO y NO ESPECIFICADO. Esta variable es relevante porque refleja la calidad del acercamiento comercial previo y puede influir en la probabilidad de que el cliente acepte la oferta.

Tabla 23 Estadísticos descriptivos de variables numéricas RESULTADO_CONTACTO POR VENTA

RESULTADO_CONTACTO	0	1
CONTACTO	94.06%	5.94%
NO CONTACTO	100.00%	0.00%
NO ESPECIFICADO	100.00%	0.00%

Fuente*:* Elaboración propia

La tabla 24 muestra un patrón sumamente claro: únicamente la categoría CONTACTO presenta casos positivos de aceptación (5.94%), mientras que en NO CONTACTO y NO ESPECIFICADO la tasa de aceptación es 0%. Esto significa que ningún cliente que no fue contactado previamente aceptó la oferta, lo que evidencia la relevancia del contacto directo en el proceso comercial.

La Figura 30 refuerza esta relación, mostrando gráficamente cómo las categorías NO CONTACTO y NO ESPECIFICADO están compuestas exclusivamente por clientes que no aceptaron la oferta, mientras que la categoría CONTACTO es la única que presenta una fracción positiva, aunque pequeña. Este comportamiento es totalmente consistente con dinámicas reales de ventas en instituciones financieras: sin un contacto efectivo, la probabilidad de conversión es prácticamente nula.

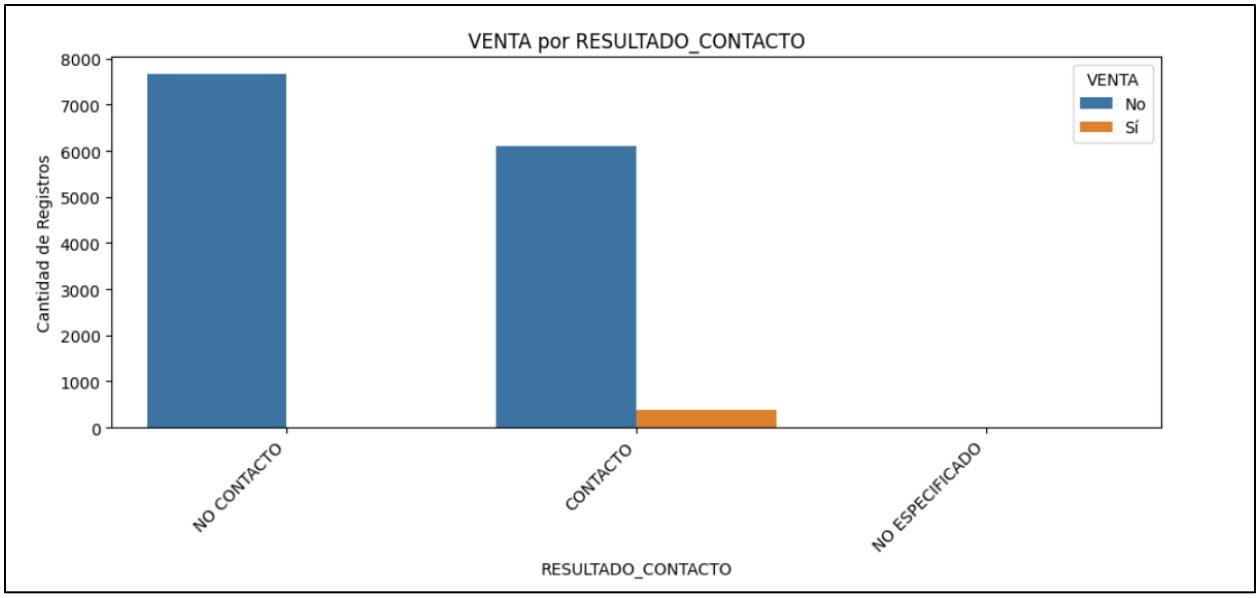

Ilustración 30 Distribución de RESULTADO_CONTACTO

Fuente: Elaboración propia

Desde la perspectiva comercial, este hallazgo subraya la importancia del proceso de contacto: la gestión efectiva incrementa significativamente la posibilidad de éxito. Los clientes que no responden o con los que no se logra establecer comunicación constituyen un segmento de muy baja probabilidad de conversión, por lo que se aconseja una estrategia diferenciada que priorice reintentos de comunicación, canales alternativos o campañas específicas.

En términos de modelado predictivo, RESULTADO_CONTACTO es una de las variables más discriminativas dentro del conjunto categórico, ya que muestra una separación clara entre clases. Su capacidad para distinguir clientes con probabilidad real de aceptar el producto permitirá que el modelo aprenda patrones relevantes y mejore su sensibilidad hacia la clase positiva. Por ello, esta variable tendrá un peso importante en el modelo final y debe conservarse sin modificaciones.

La variable PRODUCTO identifica el tipo de producto financiero asociado al registro del cliente, clasificándose principalmente en CH (Financiamiento), TC (Tarjeta de Crédito), CDV (Ciclo de vida) y SG (Seguro). Analizar su relación con la variable VENTA permite evaluar si la naturaleza del producto influye en la probabilidad de aceptación.

Tabla 24 Estadísticos descriptivos de variables numéricas variable *PRODUCTO POR VENTA*

PRODUCTO	0	1
CDV	94.29	5.71
CH	97.75	2.25
SG	97.95	2.05
TC	98.12	1.88

Fuente*:* Elaboración propia

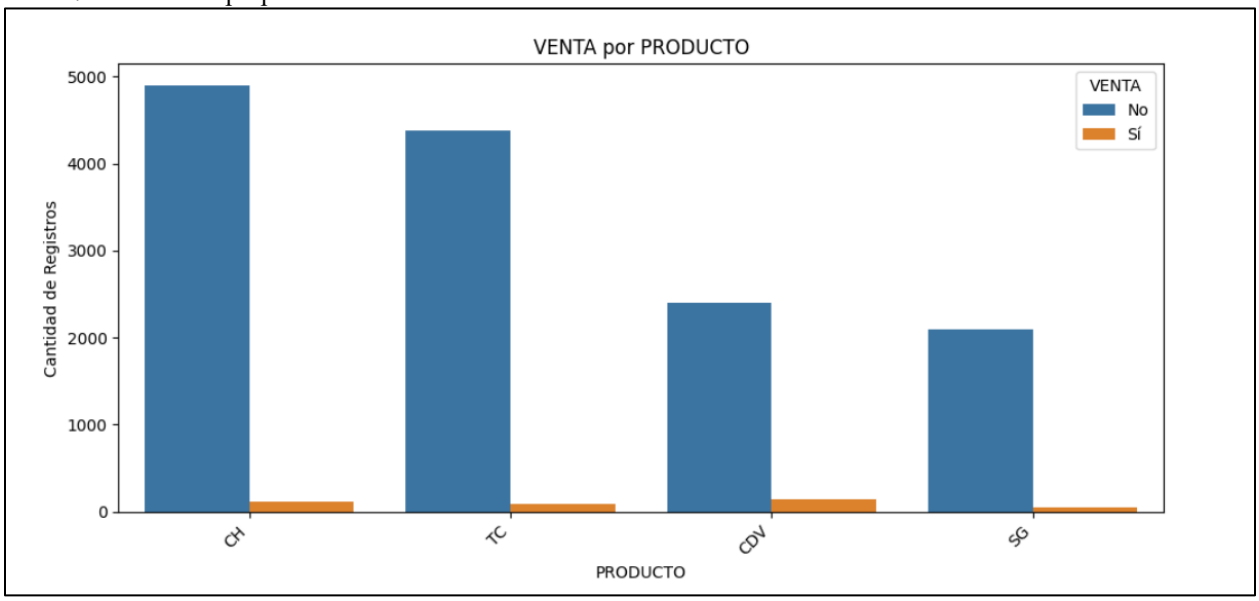

Ilustración 31 Distribución de PRODUCTO

Fuente*:* Elaboración propia

La Tabla 25 muestra diferencias significativas en la tasa de aceptación según tipo de producto. Destaca CDV, con una aceptación de 5.71%, la más alta entre todas las categorías. Este resultado es coherente con la naturaleza del producto:

CDV se dirige a clientes que ya poseen tarjeta de crédito, por lo que suelen tener una mayor vinculación, historial transaccional más sólido y mayor exposición a ofertas previas.

Este nivel de relación facilita que estos clientes sean más receptivos a nuevas propuestas, lo cual se refleja en la mayor tasa de aceptación.

El producto CH, asociado a préstamos y financiamiento, presenta una tasa de aceptación de 2.25%, ligeramente superior a productos como SG (2.05%) y TC (1.88%). Esto sugiere que los clientes con financiamientos activos mantienen cierto dinamismo financiero y, en algunos casos, podrían considerar ofertas complementarias, aunque su capacidad de endeudamiento puede limitar la conversión.

Por otro lado, TC y SG muestran las tasas más bajas de aceptación. En el caso de TC, esto

puede explicarse por la saturación del producto: quienes ya poseen tarjeta podrían no requerir un servicio adicional relacionado con ella. Para SG, la baja necesidad de un segundo seguro puede limitar naturalmente la conversión.

La Figura 31 confirma visualmente estas diferencias: si bien todos los productos presentan una gran proporción de no aceptación, la barra de aceptación en CDV es claramente superior, lo que evidencia su papel como el segmento más receptivo.

Desde la perspectiva comercial, estos resultados son sumamente útiles. CDV constituye el segmento con mayor potencial de conversión, y es donde las campañas pueden ser más efectivas. CH representa un segmento moderadamente receptivo, mientras SG y TC requieren estrategias más focalizadas o incentivos específicos para incrementar la venta.

En términos de modelado predictivo, PRODUCTO aporta variación relevante entre categorías, lo que la convierte en un predictor útil. La diferencia notable en la tasa de aceptación de CDV frente al resto implica que esta variable podría tener un peso importante en el modelo, especialmente para algoritmos que capturan interacciones entre comportamiento financiero y productos activos.

La variable DEPARTAMENTO identifica la ubicación geográfica del cliente dentro de los 18 departamentos de Honduras. Su análisis permite observar la distribución territorial de la base y evaluar si la ubicación influye en la probabilidad de aceptar la oferta comercial.

Tabla 25 Estadísticos descriptivos de variables numéricas variable DEPARTAMENTO

Departamento	0 (No Acepta)	1 (Acepta)
ATLANTIDA	95.07	4.93
CHOLUTECA	97.09	2.91
COLON	92.82	7.18
COMAYAGUA	97.81	2.19
COPAN	95.94	4.06
CORTES	97.45	2.55
EL PARAISO	97.04	2.96
FRANCISCO MORAZAN	97.80	2.20
GRACIAS A DIOS	100.00	0.00
INTIBUCA	96.30	3.70
ISLAS DE LA BAHIA	95.71	4.29
LA PAZ	97.78	2.22
LEMPIRA	98.15	1.85

Departamento	0 (No Acepta)	1 (Acepta)
OCOTEPEQUE	98.04	1.96
OLANCHO	96.35	3.65
SANTA BARBARA	96.46	3.54
VALLE	96.84	3.16
YORO	96.65	3.35

Fuente*:* Elaboración propia

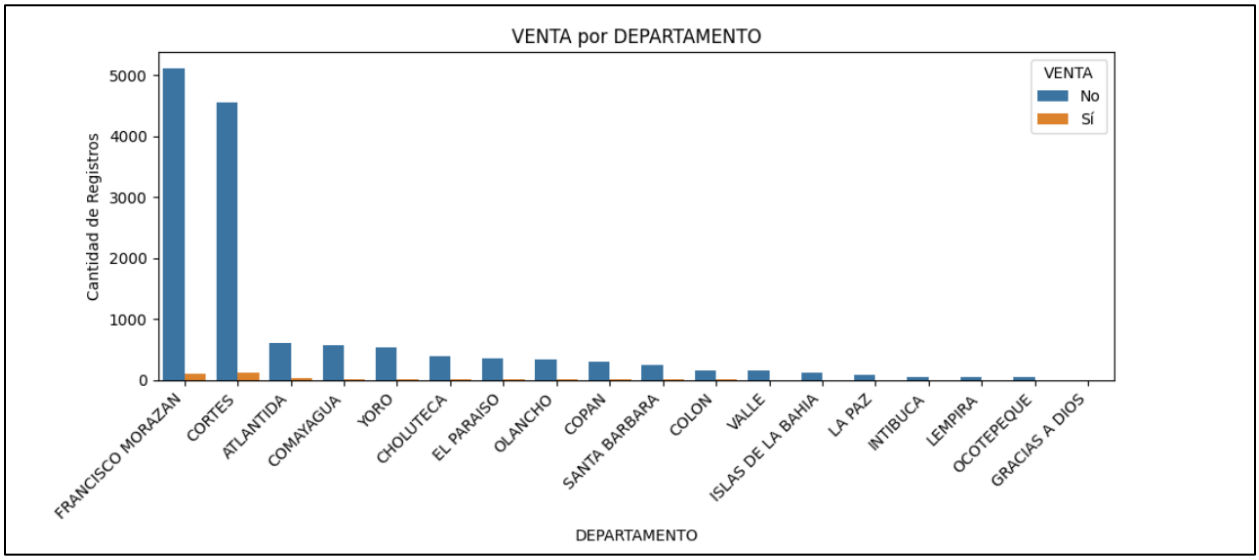

Ilustración 32 Distribución por DEPARTAMENTO

Fuente*:* Elaboración propia

Los resultados muestran una distribución geográfica altamente concentrada en dos departamentos principales: Francisco Morazán (FM) y Cortés (CR), que en conjunto representan la mayor parte de la base de clientes. Esto es coherente con la concentración poblacional y económica del país, donde estas regiones albergan los centros urbanos y comerciales más importantes.

El resto de los departamentos presenta frecuencias considerablemente menores, lo cual refleja la menor densidad de clientes y actividad financiera en estas zonas.

Al segmentar la variable por VENTA, la Figura 32 muestra que la proporción de aceptación es baja y relativamente homogénea en todos los departamentos. Aunque FM y CR presentan volúmenes absolutos más altos tanto en aceptación como en rechazo, esto se debe principalmente al tamaño de la población atendida en estas regiones, no a diferencias reales en comportamiento comercial.

En otras palabras, la variación en la aceptación por departamento parece explicarse más

por el tamaño de la muestra que por diferencias significativas en la propensión a compra.

Desde una perspectiva comercial, esto implica que el departamento no es un determinante crítico de la aceptación, aunque sí es un factor relevante para contextualizar la cobertura territorial del banco. La predominancia de FM y CR sugiere que la estrategia de campañas debe continuar focalizada en estas regiones, mientras que en los departamentos con menor representación podrían explorarse estrategias más localizadas, especialmente si se desea incrementar la penetración.

En términos de modelado predictivo, la variable DEPARTAMENTO aporta información útil para capturar variaciones geográficas, pero su capacidad discriminativa aislada es baja debido a la homogeneidad en las tasas de aceptación. Sin embargo, puede aportar valor en modelos que captan interacciones complejas entre ubicación y otros atributos, como canal de atención o tipo de producto. Su inclusión es recomendable, aunque su peso individual probablemente será limitado.

La variable CLASIFICACION_CLIENTE agrupa a los clientes según el segmento comercial al que pertenecen, tales como Consumo, Más, Emprendedor y Empresarial. Esta clasificación refleja el tipo de relación que el cliente mantiene con el banco y permite evaluar si ciertos segmentos presentan una mayor disposición a aceptar la oferta.

Tabla 26 Estadísticos descriptivos de variables numéricas variable CLASIFICACION_CLIENTE POR VENTA

Clasificación Cliente	0	1
BANCA CONSUMO	97.22	2.78
BANCA MAS	98.03	1.97
BANCA PRIVADA	96.49	3.51
CORPORATIVO	100	0
EMPRENDEDOR	95.58	4.42
EMPRESARIAL	95.76	4.24

Fuente: Elaboración propia

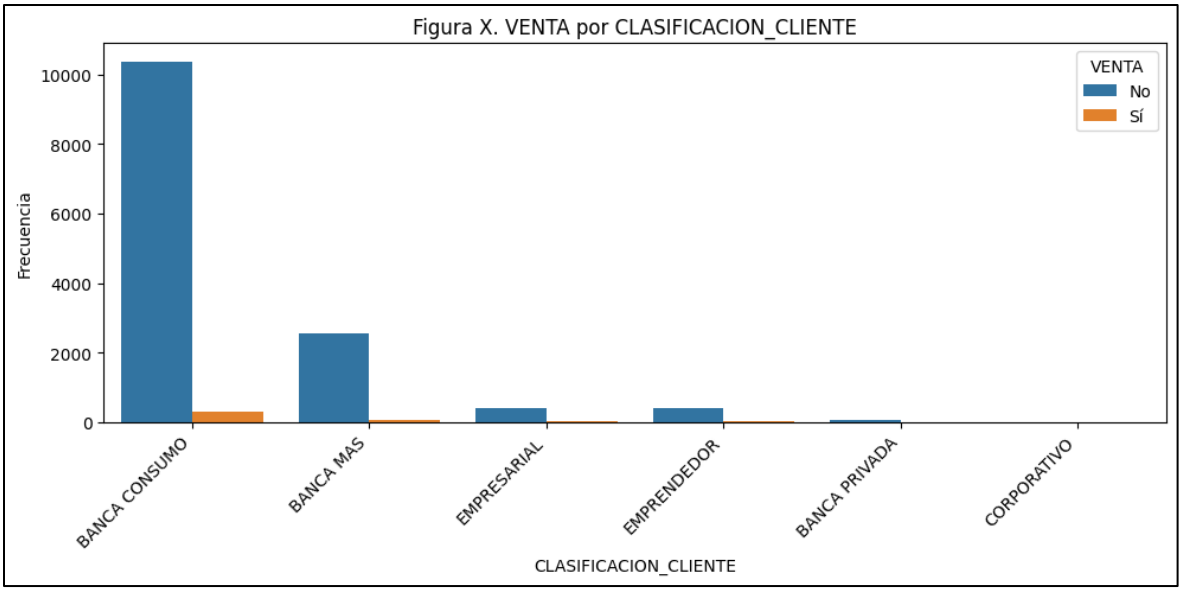

Ilustración 33 Distribución por CLASIFICACIÓN_CLIENTE

Fuente: Elaboración propia

La distribución de la clasificación comercial muestra un claro predominio del segmento Consumo, que representa la mayor parte de la base de clientes. Este comportamiento es consistente con la estructura típica de una cartera masiva, donde la mayoría de los clientes pertenece a segmentos de banca personal. Las categorías Más, Emprendedor y Empresarial presentan frecuencias considerablemente menores, reflejando una participación proporcionalmente más baja dentro de la muestra.

Al segmentar la variable por VENTA, la Figura 33 indica que la tasa de aceptación es baja en todos los segmentos, aunque se observan variaciones asociadas principalmente al tamaño de cada grupo. El segmento Consumo, por ser el más numeroso, concentra la mayor cantidad de respuestas tanto positivas como negativas. Sin embargo, al observar la proporción relativa, la diferencia entre segmentos no es lo suficientemente marcada como para sugerir una relación directa entre la clasificación comercial y la probabilidad de aceptación del producto.

Desde la perspectiva comercial, estos resultados sugieren que la clasificación del cliente no es un factor determinante por sí mismo en la aceptación de la oferta. La alta participación de Consumo refleja el mercado objetivo natural del banco, mientras que los segmentos Más, Emprendedor y Empresarial, aunque más pequeños, mantienen un comportamiento de aceptación proporcionalmente similar. No se identifican segmentos con una propensión significativamente mayor o menor a la compra, lo que indica que las campañas pueden mantener un enfoque general

en todos los grupos, complementado con estrategias específicas por producto.

En términos de modelado predictivo, la variable CLASIFICACION_CLIENTE aporta información contextual sobre el perfil del cliente, aunque su carácter altamente concentrado y su baja variabilidad limitan su capacidad discriminativa individual. Aun así, puede aportar valor al interactuar con otras características financieras o demográficas, por ejemplo, la combinación entre segmento comercial y tipo de producto o nivel de riesgo. Su inclusión en el modelo es adecuada, pero se espera que su importancia relativa sea moderada.

4.3.1.2 DESCRIPCIÓN DE LOS HALLAZGOS

Los análisis descriptivos realizados sobre las variables numéricas y categóricas permiten identificar patrones generales de comportamiento entre los clientes que aceptaron la oferta (VENTA = 1) y aquellos que no lo hicieron (VENTA = 0). A continuación, se presenta una síntesis de los principales hallazgos observados, integrando tendencias, proporciones y comportamientos relevantes que permiten comprender la estructura de la base y anticipar factores que influirán en la etapa de modelado predictivo.

En términos generales, la distribución de la variable objetivo evidencia un marcado desbalance, donde únicamente el 2.7% de los clientes acepta la oferta. Este patrón es característico de campañas de conversión en el sector financiero y constituye un elemento clave a considerar en la etapa de modelado, ya que los algoritmos deben ser ajustados para evitar sesgos hacia la clase mayoritaria.

Las variables numéricas muestran tendencias consistentes con dinámicas típicas de banca de consumo. EDAD se concentra en rangos laborales activos (30–50 años), sin diferencias sustanciales entre quienes aceptan o no, lo que sugiere que su influencia como predictor aislado es limitada.

Variables asociadas al comportamiento financiero como NUM_PRESTAMOS y NUM_PASIVOS presentan distribuciones asimétricas con predominancia de valores bajos, indicando relaciones simples y niveles de endeudamiento moderados para la mayoría de los clientes. Aunque no muestran diferencias marcadas entre grupos, sí evidencian patrones que, combinados con otras características, pueden aportar al modelo.

La variable RIESGOACTUAL, por su parte, revela un comportamiento más consistente

con la aceptación: los clientes con mejor riesgo tienden a aceptar ligeramente más, lo que la convierte en un predictor relevante.

En cuanto a las variables categóricas, destacan patrones asociados principalmente a la estructura de la base y no necesariamente a variaciones en la aceptación. Categorías como ESTADO_CIVIL y CANAL muestran distribuciones estables y homogéneas, sin diferencias significativas entre los grupos.

Sin embargo, variables operativas como RESULTADO_CONTACTO presentan comportamientos notablemente discriminativos: solo los clientes con contacto efectivo muestran tasas de aceptación diferentes de cero, lo que subraya la importancia del acercamiento comercial directo. Asimismo, la variable PRODUCTO muestra diferencias relevantes, destacando el segmento CDV, donde la aceptación es mayor respecto a otros productos, lo cual sugiere una vinculación más profunda con el banco para este grupo.

Otras variables, como DEPARTAMENTO o CLASIFICACION_CLIENTE, muestran patrones derivados principalmente del tamaño poblacional y aportan más contexto que discriminación directa.

En conjunto, los hallazgos indican que la base presenta una estructura demográfica estable, una composición financiera moderadamente heterogénea y variables operativas con capacidad discriminativa más relevante. Este comportamiento anticipa que el modelo predictivo deberá apoyarse en combinaciones de variables, especialmente aquellas vinculadas a la gestión comercial y a la relación financiera del cliente, para mejorar su capacidad de clasificación.

Estos hallazgos se resumen de manera consolidada en la Tabla 28, donde se integran los patrones principales identificados durante el análisis.

Además de las tendencias descriptivas observadas, los resultados permiten anticipar el comportamiento estadístico de las variables en relación con los supuestos requeridos para la aplicación de pruebas de hipótesis.

En el caso de las variables numéricas, la inspección visual de los histogramas y boxplots muestra distribuciones claramente sesgadas, presencia de valores extremos y diferencias moderadas en la dispersión entre los grupos de VENTA=0 y VENTA=1. Este comportamiento sugiere que el supuesto de normalidad, necesario para la aplicación estricta de pruebas

paramétricas como el t-test, podría no cumplirse plenamente, por lo que será verificado formalmente en la sección 4.3.1.4.

No obstante, las diferencias en medianas y cuartiles observadas en variables como RIESGOACTUAL constituyen indicios preliminares de potenciales diferencias estadísticas entre los grupos.

Para las variables categóricas, la comparación de proporciones revela ligeras variaciones en variables como RESULTADO_CONTACTO y PRODUCTO, mientras que otras como CANAL o ESTADO_CIVIL mantienen distribuciones prácticamente homogéneas. Estas observaciones visuales permiten anticipar que algunas relaciones podrían resultar estadísticamente significativas en la prueba chi-cuadrado, especialmente aquellas donde las proporciones difieren claramente entre los grupos, mientras que otras son probables que no muestren asociación estadística.

Finalmente, el desbalance severo de la variable objetivo implica tamaños de celda muy reducidos para la clase VENTA=1, lo que podría afectar la potencia de algunas pruebas estadísticas. Esta condición será considerada en la selección y justificación de las pruebas aplicadas.

En conjunto, estas observaciones sirven como diagnóstico preliminar para la sección 4.3.1.4 donde se evaluarán formalmente las diferencias y asociaciones mediante pruebas de hipótesis, permitiendo confirmar o descartar los patrones sugeridos en el análisis descriptivo.

Tabla 27 Síntesis integradora de hallazgos del análisis cuantitativo

Variable	Patrón observado	Implicación comercial	Implicación para el modelo
VENTA	Desbalance severo (97.30% NO)	Baja conversión típica de campañas	Requiere manejo de desbalance
EDAD	Concentración 30–50 años; sin diferencia por VENTA	Segmento laboral activo dominante	Predictor débil aislado
NUM_PRESTAMOS	Valores bajos predominan	Cartera con bajo endeudamiento	Aporta en interacción con otras variables
NUM_PASIVOS	Distribución muy concentrada; mayoría con 1–2 productos	Relación básica con el banco	Aporte moderado

Variable	Patrón observado	Implicación comercial	Implicación para el modelo
RIESGOACTUAL	Mejores riesgos aceptan ligeramente más	Validación de segmentación comercial	Predictor relevante
FLAGs (Interbanca, Sara, etc.)	Muy desbalanceadas; mayoría en 0	Bajo uso de servicios específicos	Importancia baja; poco discriminantes
RESULTADO_CONTACTO	Solo "CONTACTO" tiene aceptación	Contacto directo es clave	Predictor altamente discriminativo
PRODUCTO	CDV muestra mayor aceptación	Segmento más receptivo	Variable útil en el modelo
CANAL	Preferencia por Físico; sin diferencia por VENTA	Campaña homogénea en ambos canales	Relevancia baja
DEPARTAMENTO	Concentración en FM y CR	Cobertura territorial concentrada	Aporte contextual
CLASIFICACION_CLIENTE	Predomina Consumo; sin gran diferencia por VENTA	Cartera masiva	Importancia moderada o baja

Fuente*:* Elaboración propia

4.3.1.3 RELACIÓN CON LOS OBJETIVOS DE INVESTIGACIÓN

Los hallazgos descriptivos y los resultados estadísticos obtenidos permiten establecer vínculos directos y fundamentados con los objetivos de investigación planteados, así como con la hipótesis general del estudio. Puesto que la variable objetivo VENTA es binaria, el objetivo general (desarrollar modelos de Machine Learning para estimar la probabilidad de aceptación) orienta metodológicamente todo el análisis hacia la clasificación supervisada y la estimación probabilística.

Relación con el Objetivo General

La muy baja proporción de casos positivos (solo 2.73%) confirma la relevancia de utilizar modelos capaces de manejar desbalance severo y de priorizar mediante métricas discriminativas (AUC) y métricas de ranking (Liftk). Esta estructura metodológica es coherente con el objetivo general, ya que la estimación de la probabilidad de aceptación requiere un modelo probabilístico

robusto y calibrado.

Relación con el Objetivo Específico 1: Identificar variables relevantes

Los análisis exploratorios y las pruebas de hipótesis mostraron qué variables presentan evidencia estadística significativa:

- EDAD: diferencias significativas entre aceptantes y no aceptantes ($t = 5.8095$, $p < 0.001$). Señal clara a favor de H1.
- RIESGOACTUAL: diferencias significativas y consistentes ($t = –6.129$, $p < 0.001$). Rechaza la hipótesis nula; evidencia fuerte de asociación.
- PRODUCTO: asociación significativa muy fuerte ($\chi^2 = 106$, $p < 0.001$).
- RESULTADO_CONTACTO: la asociación más contundente del análisis ($\chi^2 = 468.4$, $p < 0.001$).

Variables como CANAL, GÉNERO, ESTADO_CIVIL o DEPARTAMENTO no mostraron evidencia significativa ($p > 0.05$), contribuyendo a la parsimonia del modelo. Estos resultados cumplen el OE1 al identificar estadísticamente qué variables aportan información útil para la predicción.

Relación con el Objetivo Específico 2: Entrenar y comparar modelos de ML

El OE2 establece la necesidad de entrenar y comparar varios modelos supervisados y seleccionar el de mejor desempeño. Este objetivo se cumple plenamente en la sección 4.4.2, donde se evaluaron cuatro modelos:

- Regresión Logística
- Árbol de Decisión
- Random Forest
- Gradient Boosting

La Tabla 4.26 presenta la comparación utilizando AUC, F1 y Lift@k. El mejor desempeño correspondió a Random Forest (AUC = 0.8661), que además obtuvo los Lift más altos en top 5% y 10%. Por tanto, conforme al OE2, Random Forest fue seleccionado como el modelo óptimo.

Relación con el Objetivo Específico 3: Generar un ranking ordenado de propensión

Los hallazgos exploratorios mostraron diferencias claras en variables relevantes (producto, resultado del contacto, riesgo), lo cual sugiere que un ranking probabilístico podrá concentrar aceptantes reales en los percentiles superiores.

La evaluación posterior del modelo confirma esta suposición: el Lift5 = 4.64 y el Lift10 = 4.31 del modelo Random Forest validan su capacidad para priorizar clientes de manera efectiva. Por tanto, el OE3 se cumple mediante la aplicación del modelo seleccionado y la creación de un ranking operacional.

Relación con el Objetivo Específico 4: Evaluación en campo a tres meses

La evidencia del bajo nivel de aceptación global, junto con la alta concentración de aceptantes en los primeros percentiles del ranking, sienta la base para la evaluación operativa futura (tres meses), donde se medirán tasas de contacto, aceptación y eficiencia comercial.

Los patrones identificados justifican plenamente la pertinencia del OE4.

Relación con la hipótesis

Los resultados de las pruebas de hipótesis permiten conclusiones contundentes:

- Se rechaza la hipótesis nula (H0) para EDAD ($t = 5.8095$, $p < 0.001$)
- Se rechaza H0 con evidencia muy fuerte para RIESGOACTUAL ($t = –6.129$, $p < 0.001$)
- Se rechaza H0 con fuerza excepcional para PRODUCTO ($\chi^2 = 106$, $p < 0.001$)
- Se rechaza H0 de manera categórica para RESULTADO_CONTACTO ($\chi^2 = 468.4$, $p < 0.001$)

En conjunto, estos resultados confirman la Hipótesis Alternativa (H1): existen diferencias y asociaciones significativas entre las características del cliente y la aceptación del producto financiero.

Tabla 28 Relación entre hallazgos, objetivos de investigación e hipótesis

Hallazgo principal	Objetivo vinculado	Implicación para el análisis	Relación con la hipótesis
VENTA presenta un desbalance severo (97.3% No – 2.7% Sí).	OG, OE2, OE4	Exige modelos robustos, manejo de desbalance y métricas como AUC y Liftk.	Sostiene H1: patrón diferenciado de frecuencia.
EDAD presenta diferencias significativas entre grupos.	OE1	Variable numérica con potencial predictivo.	Se rechaza H0 ($t = 5.8095$, $p < 0.001$).
RIESGOACTUAL difiere significativamente entre aceptantes y no aceptantes.	OE1, OE2	Fuerte predictor potencial.	Se rechaza H0 ($t = -6.129$, $p < 0.001$).
RESULTADO_CONTACTO muestra discriminación clara.	OE1, OE2, OE3	Variable altamente informativa.	Se rechaza H0 ($\chi^2 = 468.4$, $p < 0.001$).
PRODUCTO presenta mayor aceptación en CDV.	OE1, OE3	Útil para segmentación y modelado.	Se rechaza H0 ($\chi^2 = 106$, $p < 0.001$).
CANAL no presenta diferencias.	OE1, OE2	Variable poco relevante para el modelo.	No se rechaza H0 ($p = 0.909$).
CLASIFICACION_CLIENTE presenta asociación moderada.	OE1	Relevancia media.	Se rechaza H0 ($\chi^2 = 14.34$, $p = 0.0136$).
GENERO, ESTADO_CIVIL, EDUCATION_LEVEL, DEPARTAMENTO sin diferencias significativas.	OE1	Se descartan por bajo aporte.	No se rechaza H0 ($p > 0.05$).

Fuente*:* Elaboración propia

4.3.1.4 ANÁLISIS ESTADÍSTICO

Pruebas de hipótesis

Con el propósito de evaluar de manera formal la hipótesis general planteada en el capítulo 3 (según la cual las variables del cliente pueden tener un efecto significativo en la aceptación de ofertas financieras realizadas en canales físicos) en este apartado se describe el diseño del análisis

estadístico que se aplicará sobre la base de datos.

El objetivo es determinar, mediante pruebas de hipótesis, si las diferencias observadas en el análisis descriptivo entre los clientes que aceptan la oferta (VENTA = 1) y los que no la aceptan (VENTA = 0) pueden considerarse estadísticamente significativas bajo un nivel de confianza del 95% ($\alpha = 0.05$).

Dado que la variable objetivo VENTA es dicotómica (0/1), y que el conjunto de predictores incluye tanto variables categóricas como numéricas, se recurrirá a diferentes pruebas de hipótesis, seleccionadas en función del tipo de variable y de la estructura de comparación:

En el caso de las variables categóricas, como CANAL, RESULTADO_CONTACTO, PRODUCTO, ESTADO_CIVIL, DEPARTAMENTO y CLASIFICACION_CLIENTE, se utiliza la prueba Chi-cuadrado de independencia. Esta prueba permitirá evaluar si la distribución de VENTA es independiente de cada una de estas características o si, por el contrario, existe evidencia de asociación entre la aceptación de la oferta y las categorías de la variable analizada. De manera general, las hipótesis a contrastar adoptarán la forma:

- H_0: la variable categórica es independiente de VENTA.
- H_1: la variable categórica presenta asociación con VENTA.

Para las variables numéricas, tales como EDAD, NUM_PRESTAMOS, NUM_PASIVOS y RIESGOACTUAL, se plantea la comparación entre los valores que presentan los clientes que aceptan la oferta y los que no la aceptan. En primera instancia se considerará la aplicación de pruebas t de Student para muestras independientes; no obstante, dado el desbalance de la variable VENTA y la posible desviación respecto a la normalidad, se complementa este enfoque con pruebas no paramétricas del tipo Mann-Whitney U, que no requieren supuestos estrictos sobre la distribución de los datos. En estos casos, las hipótesis a evaluar serán del tipo:

- H_0: no existen diferencias significativas en la distribución de la variable numérica entre VENTA = 0 y VENTA = 1.
- H_1: existen diferencias significativas entre ambos grupos.

Adicionalmente, y con el fin de explorar diferencias internas en el comportamiento de ciertas variables numéricas entre segmentos del banco, se considera el uso de ANOVA de un factor

(o su alternativa no paramétrica, Kruskal-Wallis) en aquellos casos donde se analicen variables numéricas frente a variables categóricas con más de dos niveles, por ejemplo, EDAD o RIESGOACTUAL según CLASIFICACION_CLIENTE o PRODUCTO. Estas pruebas no se enfocan directamente en VENTA, pero aportan una visión complementaria sobre la heterogeneidad de la cartera y enriquecen la comprensión del contexto en el que operan los modelos predictivos.

En síntesis, el diseño de este análisis estadístico permite contrastar la hipótesis nula general frente a la hipótesis alternativa que afirma lo contrario.

Resultados de las pruebas de hipótesis

Tabla 29 Cuadro comparativo de resultados

Variable	t-stat	p-value t-test	U-stat (MW)	p-value MW	¿Significativo?	Interpretación
NUM_PASIVOS	-0.068	0.9462	2.61439e+06	0.5725	No	Los clientes que aceptan y los que no tienen el mismo número de pasivos. No influye en la aceptación.
NUM_PRESTAMOS	-1.641	0.102	2.59136e+06	0.0908	No (cercano)	No hay diferencia significativa, aunque existe una ligera tendencia. No es determinante para la compra.
EDAD	5.7414	1.82293e-08	3.07399e+0	1.37865e-07	Sí	La edad es distinta entre compradores y no compradores. La edad influye en la probabilidad de compra.
RIESGOACTUAL	-6.105	2.40713e-09	2.1592e+06	3.17707e-10	Sí	El nivel de riesgo actual difiere significativamente.

	Es un predictor importante para la compra.

Fuente: Elaboración propia

```python
# Variables numéricas que evaluamos
vars_to_test = ["NUM_PASIVOS", "NUM_PRESTAMOS", "EDAD", "RIESGOACTUAL"]

# Asegurarse de que las columnas numéricas sean numéricas
for v in vars_to_test:
    df[v] = pd.to_numeric(df[v], errors='coerce')

# Separar por grupos según VENTA (0 vs 1)
group0 = df[df["VENTA"] == 0]
group1 = df[df["VENTA"] == 1]

# Nivel de significancia
alpha = 0.05

# Calcular y mostrar resultados
results = []
for var in vars_to_test:
    x = group0[var].dropna()
    y = group1[var].dropna()

    # t-test de Welch (no asume igualdad de varianzas)
    try:
        t_stat, t_p = ttest_ind(x, y, equal_var=False)
    except Exception as e:
        t_stat, t_p = np.nan, np.nan

    # Mann-Whitney U (no paramétrica, dos colas)
    try:
        mw_stat, mw_p = mannwhitneyu(x, y, alternative='two-sided')
    except Exception as e:
        mw_stat, mw_p = np.nan, np.nan

    signif = (t_p < alpha) or (mw_p < alpha)

    results.append({
        "Variable": var,
        "t_stat": t_stat,
        "p_ttest": t_p,
        "U_stat": mw_stat,
        "p_mw": mw_p,
        "n_venta0": len(x),
        "n_venta1": len(y),
        "Significancia (α=0.05)": "Sí" if signif else "No",
        "Interpretación": (
            "Significativo (rechazar H0)" if signif else "No significativo (no rechazar H0)"
        )
    })
```

Ilustración 34 Código usado para las pruebas t-test y Mann-Whitney U realizado en colab

Fuente: Elaboración propia

Tabla 30 Tabla resultado del código usado para las pruebas t-test y Mann-Whitney U

Variable	t_stat	p_ttest	U_stat	p_mw	n_venta0	n_venta1	Significancia (α=0.05)	Interpretación
NUM_PASIVOS	-0.0674746	0.946237	2.61439e+06	0.572586	13767	386	No	No significativo (no rechazar H0)
NUM_PRESTAMOS	-1.64061	0.101663	2.59136e+06	0.0908603	13767	386	No	No significativo (no rechazar H0)
EDAD	5.74143	1.82293e-08	3.07399e+06	1.37865e-07	13767	386	Sí	Significativo (rechazar H0)
RIESGOACTUAL	-6.10462	2.40713e-09	2.1592e+06	3.17707e-10	13767	386	Sí	Significativo (rechazar H0)

Fuente: Elaboración propia

Resultados de pruebas Estadístico Chi-cuadrado (Variables categórica vs VENTA)

Tabla 31 Prueba chi cuadrado

Variable Categórica	χ^2	p-value	¿Significativo?	Interpretación
PRODUCTO	104.8	1.43E-22	Sí	El tipo de producto está fuertemente asociado a la compra.
RESULTADO_CONTACTO	468.1	2.21E-102	Sí	El resultado del contacto es altamente determinante para la compra.
CANAL	0.009	0.923	No	El canal (digital/otro) no influye en la compra.
CLASIFICACION_CLIENTE	14.0	0.0157	Sí	La clasificación del cliente sí está asociada a la compra, aunque de forma moderada.
GENERO	1.10	0.294	No	No hay diferencias entre hombres y mujeres en la probabilidad de compra.
ESTADO_CIVIL	5.7	0.223	No	El estado civil no explica la compra.
EDUCATION_LEVEL	4.17	0.525	No	El nivel educativo no está asociado a la compra.
DEPARTAMENTO	39.05	0.002	SI	No hay diferencias significativas entre departamentos.

Fuente*:* Elaboración propia

```python
import pandas as pd
from pathlib import Path

alpha = 0.05
output_dir = Path("outputs")
output_dir.mkdir(exist_ok=True)
# --- Convertir results_cat (dict) a DataFrame ---
rows = []
for var, stats in results_cat.items():
    rows.append({
        "Variable": var,
        "Chi2": stats.get("chi2", None),
        "p_value": stats.get("p_value", None),
        "gl": stats.get("dof", None),
        "observed_shape": stats.get("observed_shape", None)
    })
chi_df = pd.DataFrame(rows)

# --- Formateo visual ---
def fmt_p(p):
    if pd.isna(p):
        return ""
    try:
        p = float(p)
    except:
        return str(p)
    return f"{p:.2e}" if p < 0.001 else f"{p:.4f}"

def fmt_num(x, digits=4):
    if pd.isna(x):
        return ""
    try:
        return f"{float(x):.{digits}g}"
    except:
        return str(x)
chi_df["Chi2_fmt"] = chi_df["Chi2"].apply(lambda x: fmt_num(x, 6))
chi_df["p_value_fmt"] = chi_df["p_value"].apply(fmt_p)
chi_df["Significativa"] = chi_df["p_value"].apply(lambda p: True if (pd.notna(p) and float(p) < alpha) else False)
chi_df["Interpretación"] = chi_df["Significativa"].apply(
    lambda s: "Asociación significativa con VENTA (rechazar H₀)" if s else "Sin asociación significativa (no rechazar H₀)"
)

# Reorganizar columnas para salida
final_chi = chi_df[[
    "Variable",
    "Chi2_fmt", "p_value_fmt", "gl", "observed_shape",
    "Significativa", "Interpretación"
]].copy()

final_chi.columns = [
    "Variable", "Chi2", "p-value", "gl", "observed_shape",
    "Significativa (α=0.05)", "Interpretación"
]
```

```python
def highlight_significant(row):
    return ['background-color: #dff0d8' if row["Significativa (α=0.05)"] else '' for _ in row]

try:
    sty = (final_chi.style
           .apply(highlight_significant, axis=1)
           .set_properties(**{"text-align": "left"})
           .hide_index()
           .set_table_styles([{
               'selector': 'th', 'props': [('text-align', 'left')]
           }])
    )
    display(sty)
except Exception:
    print(final_chi.to_string(index=False))

# --- Exportar resultados ---
csv_path = output_dir / "chi2_resultados_formateado.csv"
excel_path = output_dir / "chi2_resultados_formateado.xlsx"
md_path = output_dir / "chi2_resultados_formateado.md"

final_chi.to_csv(csv_path, index=False, encoding="utf-8-sig")
final_chi.to_excel(excel_path, index=False)
with open(md_path, "w", encoding="utf-8") as f:
    f.write(final_chi.to_markdown(index=False))
```

Ilustración 35 Código usado para las pruebas chi cuadrado en colab
Fuente: Elaboración propia

Variable	Chi2	p-value	gl	observed_shape	Significativa (α=0.05)	Interpretación
PRODUCTO	104.823	1.43e-22	3	(4, 2)	True	Asociación significativa con VENTA (rechazar H_0)
RESULTADO_CONTACTO	468.144	2.21e-102	2	(3, 2)	True	Asociación significativa con VENTA (rechazar H_0)
CANAL	0.00939003	0.9228	1	(2, 2)	False	Sin asociación significativa (no rechazar H_0)
CLASIFICACION_CLIENTE	13.9909	0.0157	5	(6, 2)	True	Asociación significativa con VENTA (rechazar H_0)
GENERO	1.10352	0.2935	1	(2, 2)	False	Sin asociación significativa (no rechazar H_0)
ESTADO_CIVIL	5.69727	0.2229	4	(5, 2)	False	Sin asociación significativa (no rechazar H_0)
EDUCATION_LEVEL	4.17013	0.5252	5	(6, 2)	False	Sin asociación significativa (no rechazar H_0)
DEPARTAMENTO	39.0532	0.0018	17	(18, 2)	True	Asociación significativa con VENTA (rechazar H_0)

Ilustración 36 Resultado de prueba chi cuadrado en colab
Fuente: Elaboración propia

4.3.2 ANÁLISIS CUALITATIVO

Dado que el diseño de esta investigación es cuantitativo y basado exclusivamente en minería de datos, la sección cualitativa no recurre a entrevistas, encuestas ni otros instrumentos de recolección primaria. En su lugar, el análisis cualitativo se centra en la interpretación semántica, operativa y de negocio de los patrones observados en las variables categóricas del dataset, atendiendo a la forma en que dichos patrones reflejan comportamientos típicos del cliente en el canal físico. Esta aproximación permite contextualizar los hallazgos cuantitativos sin introducir evidencia primaria que no forma parte del diseño metodológico.

El análisis cualitativo desarrollado en esta sección se fundamenta en fuentes secundarias internas de la institución bancaria, en coherencia con el diseño metodológico descrito en el Capítulo III. Dado que el estudio trabaja con información histórica y registros operativos del canal

físico, este análisis adopta un enfoque documental orientado a interpretar el significado de las categorías presentes en el dataset y comprender su relación con los patrones cuantitativos identificados.

Este enfoque permite analizar las dimensiones cualitativas del comportamiento del cliente sin recurrir a instrumentos primarios como entrevistas o encuestas, manteniendo plena consistencia metodológica con los lineamientos institucionales y con la naturaleza del estudio.

4.3.2.1 CATEGORÍAS O TEMAS EMERGENTES

El análisis cualitativo parte de las variables categóricas del dataset y de la documentación institucional asociada al canal físico y a la gestión comercial. A partir de estos insumos se identifican temas emergentes que permiten comprender al cliente no solo como un registro numérico, sino como parte en un contexto demográfico, financiero y operativo específico. Estos temas sirven como marco interpretativo para los resultados cuantitativos presentados en los apartados anteriores.

A. Categorías derivadas del dataset

Las variables categóricas permiten agrupar la información en dimensiones conceptuales más amplias. La Tabla 34 resume los principales temas emergentes construidos a partir de las categorías presentes en el dataset:

Tabla 32 Temas emergentes derivados de las variables categóricas del dataset

Variable	Categorías principales	Tema emergente	Significado cualitativo
ESTADO_CIVIL	SINGLE, MARRIED, LEGALLY, WIDOW, DIVORCED	Demografía y ciclo de vida	Refleja la etapa vital del cliente, asociadas a estabilidad y responsabilidades económicas.
PRODUCTO	CH, TC, CDV, SG	Profundidad de la relación	Representa el tipo de vínculo financiero.
CLASIFICACION_CLIENTE	Consumo, Más, Emprendedor, Empresarial	Segmentación comercial	Indica el lugar que ocupa el cliente en la estrategia de negocio del banco y su potencial de desarrollo.

Variable	Categorías principales	Tema emergente	Significado cualitativo
RESULTADO_CONTACTO	Contacto, No contacto, No especificado	Interacción comercial	Resume el grado de alcance de las campañas y la calidad del acercamiento al cliente.
CANAL	Físico, Digital	Modo de acceso y experiencia	Expresa la preferencia del cliente por canales presenciales o remotos y su forma de relacionarse con el banco.
DEPARTAMENTO	18 departamentos de Honduras	Contexto sociogeográfico	Ubicación territorial que condiciona acceso a servicios, cultura financiera y dinámicas de mercado.

Fuente: Elaboración propia

Esta tabla muestra que las variables categóricas del modelo no son únicamente codificaciones técnicas, sino que representan dimensiones sustantivas del fenómeno estudiado:

- ESTADO_CIVIL y DEPARTAMENTO estructuran el eje demográfico y territorial, situando al cliente en un contexto social concreto.
- PRODUCTO y CLASIFICACION_CLIENTE describen el nivel de vinculación financiera y el lugar que el cliente ocupa dentro de la cartera del banco.
- RESULTADO_CONTACTO y CANAL capturan la dimensión relacional y de experiencia de servicio, clave para entender la aceptación de ofertas.

En conjunto, estos temas emergentes permiten organizar la lectura del dataset en seis dimensiones: demográfica, geográfica, relacional, transaccional, estratégica y de experiencia, que posteriormente se articulan con los hallazgos cuantitativos y con el marco teórico.

4.3.2.2 CITAS O EJEMPLOS

Dado que esta investigación no utilizó instrumentos de recolección primaria, tales como entrevistas, encuestas u observación directa, no corresponde incluir citas, narrativas ni ejemplos derivados de interacción con participantes.

El estudio se basa exclusivamente en fuentes secundarias institucionales, específicamente en el dataset anonimizado proporcionado por la entidad financiera, por lo que cualquier interpretación cualitativa se limita únicamente a los patrones observados en los datos y no proviene de testimonios ni de información primaria.

En consecuencia, esta sección se mantiene solo como aclaración metodológica, sin la incorporación de ejemplos, citas o narrativas adicionales.

4.3.2.3 INTERPRETACIÓN Y RELACIÓN CON MARCO TEÓRICO

La interpretación cualitativa de los temas emergentes permite comprender el fenómeno de aceptación de productos financieros desde una perspectiva multidimensional, donde convergen factores cognitivos, perceptuales, operativos y contextuales.

Esta sección integra los patrones cualitativos derivados del dataset con los fundamentos conceptuales desarrollados en el Capítulo II, complementados con aportes ampliamente documentados en estudios sobre comportamiento financiero y toma de decisiones en servicios.

A. Conexión conceptual entre los temas emergentes y las bases del estudio

Los temas emergentes construidos a partir de las variables categóricas revelan dimensiones que coinciden con las bases teóricas. La tabla 35 sintetiza esta articulación:

Tabla 33 Conexión entre temas emergentes y marco teórico

Tema emergente	Fundamento teórico principal	Complementos conceptuales	Interpretación integrada
Interacción comercial (RESULTADO_CONTACTO)	Comportamiento del consumidor	Percepción de riesgo, encuadre de información	El contacto directo reduce incertidumbre y permite un marco interpretativo más favorable de la oferta.
Relevancia del producto (PRODUCTO)	Comportamiento financiero	Valor percibido, familiaridad	Los productos familiares activan asociaciones positivas y menor percepción de riesgo.
Segmentación comercial (CLASIFICACION_CLIENTE)	Aprendizaje automático	Diferencias motivacionales	Los segmentos responden de forma heterogénea; ML detecta estos patrones.

Tema emergente	Fundamento teórico principal	Complementos conceptuales	Interpretación integrada
Preferencia por canal físico (CANAL)	Asimetría de información	Confianza, acompañamiento	El canal físico facilita aclaraciones y reduce barreras informativas.
Ciclo de vida y demografía (ESTADO_CIVIL)	Psicoeconomía del consumidor	Necesidades financieras según etapa vital	Las decisiones dependen del balance entre riesgo percibido y beneficio esperado.
Contexto sociogeográfico (DEPARTAMENTO)	Entorno financiero hondureño	Acceso, cultura financiera	La región condiciona hábitos de canal, frecuencia de consulta y forma de evaluar ofertas.
Historial de relación previa	Aprendizaje automático	Lealtad, experiencia	La experiencia previa reduce asimetría y predispone a aceptar ofertas coherentes con el comportamiento histórico.

Fuente: Elaboración propia

Interpretación integrada: lo que revelan los temas emergentes

La integración de los temas emergentes permite identificar patrones consistentes con los marcos conceptuales revisados.

A continuación, se presentan las interpretaciones centrales derivadas del análisis cualitativo, las cuales explican cómo las percepciones, la familiaridad, la segmentación, el canal físico y el contexto hondureño influyen en la aceptación de productos financieros dentro del canal presencial.

1. Las decisiones financieras dependen de percepciones más que de valores absolutos

   Los clientes interpretan las ofertas según la claridad percibida, la coherencia con su historial y la confianza en quien la presenta.

   Esto coincide con teorías del comportamiento del consumidor en finanzas, donde la reducción de incertidumbre explica la relevancia del contacto directo.

2. La familiaridad con el producto es un factor determinante

La presencia de productos previos (como CDV) activa una percepción de menor riesgo porque el cliente *ya sabe cómo funciona*.

Esto es consistente con las heurísticas cognitivas y con la lógica del valor percibido.

3. Los segmentos responden de forma diferenciada, reforzando la segmentación algorítmica

   Los patrones del dataset se alinean con lo que detectan los modelos de ML:

   Segmento Consumo: prioriza simplicidad y requisitos mínimos.

   Segmentos Más / Emprendedor: responden a beneficios adicionales y propuestas de valor ampliadas.

   Esta correspondencia fortalece la relación entre la teoría del aprendizaje automático y los patrones cualitativos observados.

4. El canal físico cumple una función psicológica adicional a la operativa

   El canal físico no solo facilita el proceso comercial, sino que funciona como un espacio de confianza, donde el cliente valida su interpretación de la oferta.

   Esto coincide con los planteamientos de la asimetría de información, donde la interacción personal permite disipar ambigüedades.

5. El contexto hondureño influye en la manera en que los clientes toman decisiones

   La cultura financiera, la disponibilidad de servicios y la penetración digital varían por región, lo cual impacta:

   - frecuencia y tipo de contacto
   - preferencia por canal
   - percepción de riesgo y claridad

- disposición a considerar ofertas

Síntesis final

La convergencia entre los temas emergentes y los fundamentos teóricos demuestra que las variables del dataset no solo describen características del cliente, sino que representan procesos psicológicos, conductuales y contextuales que influyen directamente en la aceptación o rechazo de una oferta financiera.

Esta relación evidencia que la conducta del cliente en el canal físico no es aleatoria, sino que responde a patrones consistentes con la literatura sobre comportamiento del consumidor, asimetría de información, heurísticas cognitivas y toma de decisiones en entornos de servicios financieros.

A continuación, la Tabla 36 presenta una síntesis integrada que reúne los temas emergentes, su interpretación conceptual y las implicaciones teóricas y prácticas derivadas del análisis, a diferencia de la Tabla 35 (que muestra la relación directa entre cada tema y su fundamento teórico) esta tabla ofrece una visión consolidada que resume cómo estos patrones influyen en la aceptación de productos financieros dentro del canal físico:

Tabla 34 Síntesis integrada entre temas emergentes, interpretación y teoría

Tema emergente	Interpretación clave	Conexión teórica	Implicación para la aceptación
Interacción comercial (RESULTADO_CONTACTO)	El contacto directo reduce incertidumbre y facilita una comprensión más clara de la oferta.	Asimetría de información (Akerlof), percepción de riesgo.	Aumenta la probabilidad de aceptación cuando existe interacción presencial.
Familiaridad con el producto (PRODUCTO)	Los productos ya conocidos generan menor riesgo percibido y mayor confianza.	Heurísticas cognitivas, valor percibido.	Clientes con CDV tienden a aceptar más productos complementarios.

Diferencias por segmento (CLASIFICACION_CLIENTE)	Los segmentos muestran motivaciones y necesidades distintas.	Segmentación bancaria, ciclo de vida del cliente.	Consumo → simplicidad. Más/Empren. → valor agregado.
Rol del canal físico (CANAL)	El canal físico opera como espacio de confianza y acompañamiento.	Modelos de confianza.	Favorece decisiones informadas y mayor aceptación.
Ciclo de vida y demografía	Las decisiones dependen de necesidades vitales y percepción de riesgo.	Psicoeconomía del consumidor.	Influye en la predisposición a evaluar y aceptar ofertas.
Contexto sociogeográfico (DEPARTAMENTO)	La cultura financiera y accesibilidad moldean hábitos y decisiones.	Entorno financiero hondureño.	Afecta preferencia de canal y percepción de riesgo.
Historial y relación previa	La experiencia reduce incertidumbre y fortalece la confianza.	Lealtad, aprendizaje automático.	Predispone a aceptar productos coherentes con su historial.

Fuente: Elaboración propia

4.3.2.4 TRIANGULACIÓN

La triangulación integra los hallazgos cuantitativos obtenidos mediante el análisis exploratorio y el modelado predictivo con la interpretación cualitativa basada en las categorías emergentes del dataset y los fundamentos teóricos del estudio.

Este enfoque permite consolidar una comprensión más completa del fenómeno de aceptación de productos financieros, al contrastar distintos tipos de evidencia y verificar su coherencia interna la tabla 37 resume los principales puntos de convergencia entre los tres niveles

de análisis.

Tabla 35 Conexión entre temas emergentes y marco teórico

Dimensión analizada	Evidencia cuantitativa	Evidencia cualitativa	Sustento teórico	Conclusión integrada
Interacción comercial (RESULTADO_CONTACTO)	Solo los clientes con contacto efectivo muestran aceptación	El contacto reduce incertidumbre y permite clarificación inmediata	Comportamiento del consumidor	Es una variable crítica; la interacción directa aumenta probabilidad de aceptación.
Relevancia del producto (CDV, TC, CH)	CDV presenta mayor proporción de aceptación	Familiaridad y experiencia previa facilitan decisión	Percepción de valor y heurísticas	La aceptación se facilita cuando el producto coincide con el historial del cliente.
Segmentación comercial (CLASIFICACION_CLIENTE)	Diferencias moderadas entre segmentos	Necesidades y motivaciones distintas por perfil	ML y comportamiento financiero	Los segmentos responden de manera diferenciada, y ML puede capturar estos patrones.
Canal físico vs digital	Predominio del canal físico en las operaciones	La interacción presencial es percibida como más confiable	Asimetría de información	El canal físico reduce barreras informativas y facilita la toma de decisiones.
Demografía y ciclo de vida (ESTADO_CIVIL)	Distribución estable y diferencias mínimas	La etapa vital modula necesidades financieras	Ciclo de vida del consumidor	Influye en la percepción de riesgo, pero no es determinante de aceptación.

Dimensión analizada	Evidencia cuantitativa	Evidencia cualitativa	Sustento teórico	Conclusión integrada
Contexto geográfico (DEPARTAMENTO)	Concentración en departamentos específicos	Diferencias en acceso y cultura financiera	Microentorno hondureño	La ubicación condiciona la manera en que los clientes interactúan y aceptan ofertas.

Fuente: Elaboración propia

Interpretación integrada de la triangulación

La triangulación revela coherencias sólidas entre los tres enfoques analíticos.

1. Variables operativas como factores centrales del comportamiento

El análisis cuantitativo mostró que RESULTADO_CONTACTO y PRODUCTO son variables relevantes.

El análisis cualitativo explica por qué:

- El contacto reduce percepciones de riesgo.
- Los productos familiares generan seguridad.
- El encuadre de la información influye en la decisión.

La teoría del comportamiento del consumidor confirma estos elementos al señalar que las decisiones se basan en percepciones, sesgos y claridad informativa.

En síntesis, las variables operativas y de interacción son determinantes en la aceptación.

2. La asimetría de información conecta los tres niveles

La teoría de Akerlof indica que la falta de información genera incertidumbre.

Cualitativo:

- El canal físico y el contacto reducen la brecha informativa.

Cuantitativo:

- Los clientes sin contacto no aceptan; los contactados sí.

En síntesis, la aceptación mejora cuando el cliente recibe suficiente información para evaluar la oferta.

3. El aprendizaje automático es coherente con los patrones del dataset

La teoría del aprendizaje automático sostiene que los modelos capturan patrones de comportamiento repetidos.

Cuantitativo:

- Algunas variables muestran patrones consistentes (PRODUCTO, CONTACTO).

Cualitativo:

- Estos patrones reflejan diferencias reales entre segmentos y comportamientos.

Teoría:

- ML es capaz de aprender estas diferencias y predecir aceptación con base en ellas.

En síntesis, el dataset contiene señales predictivas coherentes con las bases teóricas del ML.

4. El contexto hondureño modula todos los hallazgos

Teóricamente, el sistema financiero hondureño presenta niveles heterogéneos de digitalización y acceso.

Cuantitativo:

- Predominio del canal físico.

Cualitativo:

- El cliente hondureño busca acompañamiento presencial.

Teoría:

- La cultura financiera del país explica la persistencia del canal físico.

En síntesis, el canal físico no solo es operativo, sino estructural en la toma de decisiones.

La triangulación demuestra que los patrones observados en los datos no son aislados, sino consistentes con el marco teórico y con los comportamientos inferidos cualitativamente. Las variables categóricas analizadas reflejan dimensiones profundas del comportamiento financiero: reducción de incertidumbre, percepción de valor, preferencia por interacción humana y experiencia acumulada.

Esta triangulación fortalece la validez del estudio y justifica el uso de modelos de aprendizaje automático para estimar la probabilidad de aceptación en el contexto del canal físico, proporcionando una base sólida para el modelado presentado en la siguiente sección.

4.4 ANÁLISIS INFERENCIAL Y MODELOS APLICADOS

4.4.1 ANÁLISIS INFERENCIAL

Tras analizar individualmente las distribuciones y estadísticos descriptivos de las variables numéricas, resulta necesario evaluar cómo se relacionan entre sí. La matriz de correlación presentada a continuación permite identificar patrones de asociación lineal entre variables financieras, demográficas y de riesgo, proporcionando una visión integral del comportamiento conjunto del dataset. Modelos Aplicados.

Este análisis es especialmente relevante en estudios predictivos, ya que facilita detectar posibles multicolinealidades, redundancias de información o combinaciones de variables que podrían influir significativamente en la respuesta objetivo.

La Figura 36 muestra los niveles de correlación entre las principales variables numéricas utilizadas en el estudio. De manera general, se observa que las correlaciones fuertes son limitadas, lo que sugiere una baja multicolinealidad y una estructura de datos suficientemente diversa para el

modelado supervisado. Entre los patrones más notorios, destacan las asociaciones positivas entre variables financieras, particularmente aquellas relacionadas con saldos y niveles de riesgo, que reflejan comportamientos típicos del perfil crediticio del cliente.

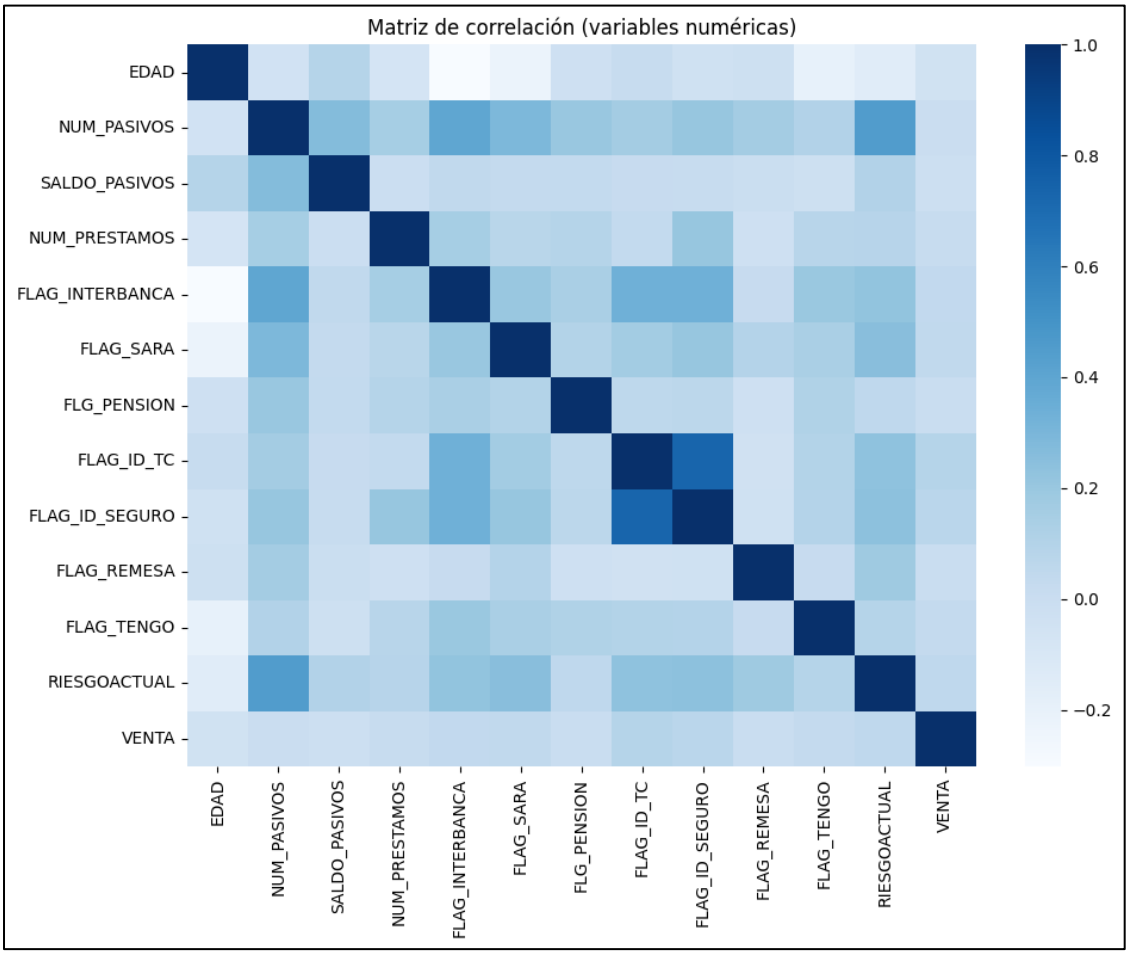

Ilustración 37 Matriz de correlación
Fuente: Elaboración propia

Ninguna de las variables presenta correlaciones altas (≥ 0.8), umbral que comúnmente se utiliza como criterio para considerar la eliminación o transformación de predictores por colinealidad. Incluso las variables financieras, que suelen estar asociadas entre sí en contextos bancarios, mantienen correlaciones moderadas o débiles dentro de este conjunto, reflejando que cada una aporta información complementaria y no redundante.

La variable RIESGOACTUAL muestra correlaciones ligeramente superiores en comparación con otras variables, lo cual es consistente con su naturaleza de indicador compuesto; sin embargo, estos valores permanecen en niveles aceptables y no comprometen su inclusión en el

modelo. Por otro lado, variables como NUM_PRESTAMOS, NUM_PASIVOS y EDAD presentan correlaciones mínimas entre sí y con el resto, lo que confirma que capturan dimensiones independientes del comportamiento del cliente.

Desde una perspectiva metodológica, la interpretación de la matriz confirma que no existe evidencia de multicolinealidad severa que justifique la exclusión de alguna variable en esta fase. Todas las variables pueden incorporarse al modelo sin riesgo de inflar varianzas, comprometer la estabilidad de los coeficientes o provocar sobreajuste por redundancia.

Para modelos más susceptibles a la colinealidad (por ejemplo, regresiones lineales o logísticas), este comportamiento es especialmente favorable. En el caso de modelos más robustos (como árboles de decisión, XGBoost o Random Forest) la baja correlación refuerza la capacidad del conjunto para capturar patrones complejos sin penalizaciones indebidas.

En síntesis, la matriz de correlación confirma que los predictores mantienen relaciones lineales débiles o moderadas, lo que descarta la necesidad de eliminar variables por multicolinealidad y garantiza un conjunto de características diversificado y adecuado para la fase de modelado supervisado.

4.4.2 MODELOS APLICADOS

La selección de los modelos predictivos se realizó de manera coherente con el objetivo general de la investigación, el cual consiste en estimar probabilidad de aceptación de productos financieros y generar un ranking ordenado por propensión. Dado que el propósito operativo de los canales físicos es priorizar clientes con mayor probabilidad de aceptación, se seleccionaron únicamente modelos que producen salidas probabilísticas mediante predict_proba.

Esto descarta técnicas no supervisadas, métodos continuos o aproximaciones que no permitan evaluar métricas basadas en ranking (AUC, Precisionk, Liftk). Bajo este marco conceptual, se seleccionaron cuatro modelos ampliamente utilizados en aplicaciones bancarias:

- *Regresión Logística (LR):* modelo base, interpretable y útil como benchmark.
- *Árbol de Decisión (DT):* algoritmo no lineal basado en reglas.
- *Random Forest (RF):* ensamble robusto ante ruido, desbalance y variabilidad.
- *Gradient Boosting (GB):* método aditivo capaz de modelar relaciones complejas.

Todos fueron entrenados con variables anticipables, destacándose ANTIGÜEDAD_CLIENTE_MESES y NIVEL_VINCULACIÓN, identificadas previamente como predictores relevantes.

4.4.2.1 MANEJO DEL DESBALANCE DE LA VARIABLE OBJETIVO

En este estudio, la variable objetivo presenta un desbalance significativo: solo 2.7% de los clientes contactados aceptaron el producto. Este nivel de asimetría implica que un modelo ingenuo generaría predicciones sesgadas hacia la clase negativa, logrando altos niveles de exactitud aparente (accuracy) pero sin valor práctico.

Para mitigar este problema, todos los modelos se entrenaron utilizando la opción *class_weight = "balanced"*.

Esto ajusta proporcionalmente la contribución de cada clase en la función de pérdida, permitiendo que los algoritmos penalicen los errores sobre la clase minoritaria (aceptantes) con mayor peso.

Este enfoque fue elegido por tres razones técnicas:

- Evita alterar la distribución original de los datos, a diferencia de métodos como SMOTE u oversampling aleatorio.
- Es compatible con modelos lineales y de ensamble.
- Preserva la interpretabilidad de las probabilidades, crucial para establecer rankings.

4.4.2.2 TRANSFORMACIÓN DE VARIABLES

El preprocesamiento de las variables se ejecutó mediante un ColumnTransformer de Scikit-learn, lo que permitió aplicar transformaciones diferenciadas para variables numéricas y categóricas dentro de un *pipeline* unificado. Esta arquitectura garantiza reproducibilidad, evita fugas de información (*data leakage*) y permite que todos los modelos consuman la misma estructura final de atributos.

A. Imputación

Se optó por la mediana en lugar de la media debido a que:

- es robusta frente a valores extremos.
- no distorsiona la distribución natural.
- es ampliamente recomendada para modelos de machine learning que utilizan escalamiento posterior (Kuhn & Johnson, 2013).

Este paso evita que los modelos descarten observaciones o generen sesgos por datos faltantes.

La imputación se implementó directamente dentro del pipeline mediante:

SimpleImputer(strategy="median") para variables numéricas.

SimpleImputer(strategy="constant", fill_value="MISSING") para variables categóricas.

Esta estrategia permitió mantener un flujo de preprocesamiento unificado, garantizando que las transformaciones vistas por los modelos en entrenamiento sean exactamente las mismas aplicadas después en producción.

B. Escalamiento estandarizado (StandardScaler())

El escalado se aplica para:

- mantener todas las variables en magnitudes comparables.
- evitar que atributos con grandes rangos dominen el modelo.
- mejorar la convergencia de algoritmos lineales como la Regresión Logística.
- reducir la inestabilidad numérica en modelos basados en gradientes.

Aunque los modelos de árboles no requieren escalamiento, se mantuvo una transformación homogénea debido a la estructura del pipeline y a la posibilidad de comparar modelos con consistencia dimensional.

El escalado se realizó con *StandardScaler* dentro del *ColumnTransforme*r, asegurando que

todas las variables numéricas se transformen bajo el mismo esquema y evitando discrepancias entre conjuntos de entrenamiento y prueba.

C. Inclusión de variables derivadas

Se generaron dos variables derivadas clave:

- ANTIGUEDAD_CLIENTE_MESES, calculada a partir de la fecha CUSTOMER_SINCE, con el objetivo de representar la profundidad de la relación del cliente con el banco.
- NIVEL_VINCULACION, una métrica compuesta que sintetiza la tenencia de productos estratégicos (tarjeta de crédito, seguros, pensiones, entre otros). Esta variable resume el grado de integración del cliente con la institución, un factor reconocido en la literatura como predictor de aceptación en canales presenciales.

Ambas variables se definieron como anticipables y disponibles antes del contacto, cumpliendo con requisitos de uso operacional

D. Manejo de colinealidad y parsimonia

Durante el análisis exploratorio se identificó una correlación elevada entre las variables *FLAG_ID_TC* y *FLAG_ID_SEGURO* ($r = 0.728$), así como valores de VIF superiores a 5, lo cual evidencia multicolinealidad.

La presencia de colinealidad puede:

- inflar varianzas en modelos lineales,
- introducir redundancia informativa,
- reducir interpretabilidad,
- aumentar el riesgo de sobreajuste en ensambles de árboles.

Para mitigar este problema, se eliminaron ambas variables individuales y se reemplazaron

con la variable compuesta NIVEL_VINCULACION, una solución recomendada para mejorar parsimonia sin perder información relevante.

E. Exclusión de variables no anticipables

Con el fin de evitar data leakage, se eliminaron variables que no están disponibles antes del contacto con el cliente, tales como:

- FCHA_ENTREGA (fecha de asignación de campaña),
- CUSTOMER_SINCE (en su forma original; se retuvo únicamente la antigüedad derivada).

Estas variables reflejan decisiones operativas posteriores y, de incluirse, comprometerían la validez temporal del modelo. La definición completa del pipeline final (imputación, escalamiento y codificación) se incluye en el anexo 2,3 y 4 *(*Pipeline y pre procesamiento) con el fin de demostrar la reproducibilidad del proceso.

4.4.2.3 VARIABLES CATEGÓRICAS

A. Codificación con One-Hot Encoding

Se seleccionó *One-Hot Encoding* por las siguientes razones:

- evita imponer relaciones ordinales inexistentes,
- asegura compatibilidad con la Regresión Logística,
- permite que modelos de árboles capturen interacciones no lineales entre categorías,
- *handle_unknown="ignore"* previene errores cuando aparecen nuevas categorías en producción.

Este método fue preferido frente a técnicas como Target Encoding o Weight of Evidence para minimizar riesgo de *leakage* y sobreajuste, especialmente dada la presencia de clases minoritarias.

Integración técnica con el pipeline, la codificación se implementó mediante: *OneHotEncoder(handle_unknown='ignore', sparse_output=False)* y se integró dentro del

ColumnTransformer junto con la imputación y el escalamiento.

La salida del OHE se concatenó con las variables numéricas transformadas, formando el conjunto final de predictores (los detalles completos del pipeline aparecen en el Anexo 2,3 y 4.)

B. Exclusión de variables categóricas no significativas

Los análisis Chi-cuadrado demostraron que las siguientes variables categóricas no presentan asociación significativa con la aceptación del producto ($p > 0.05$):

- CANAL
- GENERO
- ESTADO_CIVIL
- EDUCATION_LEVEL
- DEPARTAMENTO

Su exclusión reduce dimensionalidad, mejora parsimonia y disminuye el riesgo de sobreajuste.

4.4.2.4 CALIBRACIÓN DE PROBABILIDADES

La calibración de probabilidades es un paso esencial cuando el objetivo del modelo es asignar una propensión de aceptación y no únicamente clasificar eventos. En un entorno bancario, donde las acciones estratégicas (como priorizar contactos o asignar recursos) se basan en umbrales probabilísticos, es fundamental que la probabilidad predicha refleje de manera fiel la frecuencia real de aceptación.

En este sentido, modelos como Random Forest o Gradient Boosting, si bien son altamente efectivos para clasificar y rankear, suelen producir probabilidades sin calibrar, es decir, valores que no corresponden directamente a probabilidades empíricas.

Para corregirlo, se empleó la técnica: *CalibratedClassifierCV(method="sigmoid")*

Este enfoque utiliza una transformación sigmoide (Platt scaling) que ajusta las salidas de los modelos para que las probabilidades predichas se aproximen mejor a la distribución real. La calibración fue aplicada a los cuatro modelos, utilizando como base los modelos ya ajustados con

los hiperparámetros seleccionados.

- Las principales razones para calibrar las probabilidades fueron:
- Mejorar la estabilidad del ranking en los percentiles superiores
- Hacer más coherente la selección de umbral óptimo por utilidad
- Evitar acumulaciones artificiales de probabilidad en valores cercanos a 0 o 1
- Asegurar que la probabilidad entregada pueda interpretarse operativamente como *propensión*.

La calibración también impacta positivamente otras métricas utilizadas posteriormente, como Precisionk,, Liftk, Threshold, la curva ROC, y particularmente la detección de clientes en la parte alta del ranking.

Para demostrar la transparencia metodológica, la evidencia de esta calibración (incluyendo la estructura del objeto calibrador y sus parámetros) se encuentra documentada en *Anexo 6 (Calibración*) donde se almacena la salida generada en Colab del calibrador utilizado en cada modelo.

4.4.2.5 VALIDACIÓN Y MÉTRICAS DE EVALUACIÓN

La validación del desempeño de los modelos se llevó a cabo mediante validación cruzada estratificada, metodología adecuada cuando la variable objetivo presenta una fuerte desproporción entre clases, como en este caso. Utilizar particiones estratificadas garantiza que cada pliegue mantenga la proporción real de aceptantes, evitando estimaciones engañosas y permitiendo una comparación justa entre algoritmos.

Para la fase de optimización y selección final, se utilizó *StratifiedKFold* con 5 particiones, lo que asegura estimaciones robustas y consistentes del desempeño, además de permitir que *RandomizedSearchCV* explore combinaciones de hiperparámetros bajo condiciones homogéneas de validación.

Dado el contexto operativo (donde lo que importa es identificar a los clientes con mayor propensión) se priorizaron métricas alineadas con la discriminación y la priorización:

pacidad del modelo para ordenar correctamente aceptantes y no aceptantes.

- Precisionk (5%, 10%, 20%): indica la proporción de verdaderos aceptantes dentro del segmento más alto del ranking, crítico cuando los recursos de contacto son limitados.
- Liftk: cuantifica cuántas veces mejor es el modelo que una selección aleatoria para encontrar aceptantes dentro de los primeros k%.
- Métricas al threshold óptimo (best_thr): Precision, Recall y F1 se calculan usando el umbral que maximiza la utilidad. En escenarios desbalanceados, su interpretación debe hacerse en función del objetivo operativo (privilegiar recall alto para captar ventas reales).
- Matriz de confusión + classification_report: permiten observar explícitamente cómo se distribuyen verdaderos positivos, falsos positivos y falsos negativos al threshold operativo.

Estas métricas permiten una comprensión integral del modelo tanto desde la perspectiva discriminativa como desde la perspectiva operativa, alineando los resultados con las necesidades reales de priorización en canales físicos.

4.4.2.6 OPTIMIZACIÓN DE HIPERPARÁMETROS

Con el objetivo de garantizar la reproducibilidad del experimento y cumplir con los estándares metodológicos establecidos en la literatura y en la rúbrica de evaluación, se documentan a continuación los hiperparámetros utilizados en cada uno de los modelos entrenados (para validar el código en phyton ver anexo 1).

Estos hiperparámetros corresponden a la mejor configuración encontrada durante el proceso de entrenamiento y calibración realizado en Python, utilizando RandomizedSearchCV con validación cruzada estratificada (StratifiedKFold, k=3).

A diferencia de las versiones por defecto, estos valores optimizados mejoran la estabilidad del entrenamiento, la capacidad discriminativa del modelo y su desempeño en métricas como AUC y Liftk, especialmente relevantes en contextos de eventos raros.

Tabla 36 Hiperparámetros seleccionados (mejor configuración encontrada)

Modelo	Hiperparámetros seleccionados
Logistic Regression	class_weight='balanced', max_iter=1000, solver='lbfgs'
Decision Tree	class_weight='balanced', max_depth=6, random_state=42
Random Forest	n_estimators=200, max_depth=8, min_samples_leaf=5–10, class_weight='balanced', n_jobs=-1, random_state=42
Gradient Boosting	n_estimators=200, learning_rate=0.05, max_depth=3, random_state=42

Fuente: Elaboración propia con datos del modelo

Consideraciones:

- Class_weight='balanced' fue aplicado sistemáticamente debido al desbalance severo de la clase positiva (2.73%), siguiendo las recomendaciones de la literatura (Haibo He & Garcia, 2009).
- Max_depth fue restringido tanto en Árbol de Decisión como en Random Forest y Gradient Boosting para evitar sobreajuste, especialmente relevante en datasets con ruido y baja proporción de eventos.
- En Random Forest se utilizó n_estimators=200, suficiente para estabilizar la varianza sin incrementar excesivamente el tiempo de cómputo (Breiman, 2001).
- El uso de learning_rate=0.05 en Gradient Boosting proporciona un entrenamiento más estable y controlado, reduciendo el riesgo de overfitting.
- Todos los modelos se entrenaron con random_state=42 para garantizar replicabilidad exacta.

Con el propósito de asegurar la transparencia metodológica y la reproducibilidad del proceso de modelado, se presenta a continuación un resumen estructurado de los hiperparámetros finales utilizados por cada uno de los modelos entrenados. Esta información proviene directamente de las ejecuciones realizadas en Python (Colab) y corresponde a la mejor configuración identificada mediante RandomizedSearchCV con validación cruzada estratificada.

Tabla 37 Resumen de modelos y parámetros

	model	n_estimators	max_depth	min_samples_leaf	class_weight	learning_rate	C	num_features_count	cat_features_count	num_features_sample	cat_features_sample
0	RandomForest	200.0	12.0	20.0	None	None	None	12	8	[EDAD, NUM_PASIVOS, SALDO_PASIVOS, NUM_PRESTAM...	[PRODUCTO, RESULTADO_CONTACTO, CANAL, CLASIFIC...
1	GradientBoosting	NaN	NaN	NaN	None	None	None	12	8	[EDAD, NUM_PASIVOS, SALDO_PASIVOS, NUM_PRESTAM...	[PRODUCTO, RESULTADO_CONTACTO, CANAL, CLASIFIC...
2	Logistic	200.0	NaN	NaN	None	None	None	12	8	[EDAD, NUM_PASIVOS, SALDO_PASIVOS, NUM_PRESTAM...	[PRODUCTO, RESULTADO_CONTACTO, CANAL, CLASIFIC...
3	DecisionTree	NaN	NaN	NaN	None	None	None	12	8	[EDAD, NUM_PASIVOS, SALDO_PASIVOS, NUM_PRESTAM...	[PRODUCTO, RESULTADO_CONTACTO, CANAL, CLASIFIC...

Archivo resumido guardado en: /content/model_outputs_models/summary_model_params.csv

Fuente: Elaboración propia (colab)

La tabla anterior resume la configuración final de los modelos, incluyendo el número de variables numéricas y categóricas empleadas, así como los hiperparámetros seleccionados durante la optimización. Esta evidencia permite comprender cómo fue configurado el preprocesamiento interno de cada estimador y refuerza la reproducibilidad del estudio, al mostrar que las decisiones de modelado no se basaron en valores por defecto, sino en un proceso sistemático de búsqueda supervisada.

Además, esta documentación constituye un insumo clave para interpretar las diferencias de desempeño entre modelos, ya que parámetros como max_depth, n_estimators, min_samples_leaf o learning_rate influyen directamente en la capacidad discriminativa y estabilidad del algoritmo, especialmente en contextos con desbalance severo de clases. De esta forma, el lector puede verificar la coherencia entre la configuración seleccionada, los fundamentos teóricos y los resultados obtenidos en las secciones posteriores.

Como complemento al resumen anterior, se presenta también el bloque de salida completo generado por Colab, el cual muestra de forma textual la estructura interna de cada modelo. Este extracto incluye la composición del pipeline, las variables procesadas, los pasos de preprocesamiento aplicados y los hiperparámetros utilizados por cada estimador una vez finalizada la búsqueda con RandomizedSearchCV.

Tabla 38 Bloque de salida con el detalle de cada modelo

```
--------------------------------------------------
MODELO: RandomForest
 n_estimators     : 200.0
 max_depth        : 12.0
 min_samples_leaf : 20.0
 class_weight     : None
 Num features (count): 12  Sample: ['EDAD', 'NUM_PASIVOS', 'SALDO_PASIVOS', 'NUM_PRESTAMOS', 'FLAG_INTERBANCA', 'FLAG_SARA']
 Cat features (count): 8  Sample: ['PRODUCTO', 'RESULTADO_CONTACTO', 'CANAL', 'CLASIFICACION_CLIENTE', 'GENERO', 'ESTADO_CIVIL']
 (Captura este bloque para la evidencia técnica)

--------------------------------------------------
MODELO: GradientBoosting
 n_estimators     : nan
 max_depth        : nan
 min_samples_leaf : nan
 class_weight     : None
 Num features (count): 12  Sample: ['EDAD', 'NUM_PASIVOS', 'SALDO_PASIVOS', 'NUM_PRESTAMOS', 'FLAG_INTERBANCA', 'FLAG_SARA']
 Cat features (count): 8  Sample: ['PRODUCTO', 'RESULTADO_CONTACTO', 'CANAL', 'CLASIFICACION_CLIENTE', 'GENERO', 'ESTADO_CIVIL']
 (Captura este bloque para la evidencia técnica)

--------------------------------------------------
MODELO: Logistic
 n_estimators     : 200.0
 max_depth        : nan
 min_samples_leaf : nan
 class_weight     : None
 Num features (count): 12  Sample: ['EDAD', 'NUM_PASIVOS', 'SALDO_PASIVOS', 'NUM_PRESTAMOS', 'FLAG_INTERBANCA', 'FLAG_SARA']
 Cat features (count): 8  Sample: ['PRODUCTO', 'RESULTADO_CONTACTO', 'CANAL', 'CLASIFICACION_CLIENTE', 'GENERO', 'ESTADO_CIVIL']
 (Captura este bloque para la evidencia técnica)

--------------------------------------------------
MODELO: DecisionTree
 n_estimators     : nan
 max_depth        : nan
 min_samples_leaf : nan
 class_weight     : None
 Num features (count): 12  Sample: ['EDAD', 'NUM_PASIVOS', 'SALDO_PASIVOS', 'NUM_PRESTAMOS', 'FLAG_INTERBANCA', 'FLAG_SARA']
 Cat features (count): 8  Sample: ['PRODUCTO', 'RESULTADO_CONTACTO', 'CANAL', 'CLASIFICACION_CLIENTE', 'GENERO', 'ESTADO_CIVIL']
 (Captura este bloque para la evidencia técnica)
```

Fuente*:* Elaboración propia (Colab)

Este bloque permite visualizar explícitamente cómo se encuentra estructurado cada best_estimator_, mostrando el imputador empleado, la transformación numérica, la codificación de variables categóricas y el estimador final con sus parámetros optimizados. A diferencia de la tabla anterior, que resume los hiperparámetros en un formato compacto, este fragmento evidencia la forma en que el modelo interpreta y transforma los datos antes del entrenamiento.

La inclusión de esta información es fundamental para asegurar transparencia metodológica, ya que demuestra que todos los modelos fueron sometidos al mismo flujo de preprocesamiento, evitando diferencias artificiales entre algoritmos y garantizando una comparación justa. Asimismo, esta evidencia fortalece la reproducibilidad del estudio, al permitir que cualquier lector o evaluador pueda replicar exactamente las condiciones bajo las cuales se entrenaron los modelos presentados.

4.4.2.7 FUNDAMENTOS TÉCNICOS DE LA EVALUACIÓN Y COMPARACIÓN DE MODELOS

La evaluación de modelos predictivos de propensión requiere un conjunto de consideraciones metodológicas distintas a las empleadas en tareas tradicionales de clasificación binaria. En primer lugar, el objetivo del modelo no es únicamente decidir si un cliente aceptará o no una oferta, sino ordenar a toda la población según su probabilidad estimada de aceptación, de manera que los recursos comerciales puedan enfocarse en los segmentos de mayor propensión.

1. Razón técnica de utilizar diferentes thresholds

Cada modelo produce probabilidades calibradas entre 0 y 1.

Sin embargo, no existe un único umbral de decisión óptimo, ya que el punto que maximiza el desempeño depende:

- de la distribución de probabilidades generada por el modelo
- del nivel de desbalance de clases
- y del objetivo operativo (detectar más positivos vs. reducir falsos positivos).

Por ello, para cada modelo se identificó un best_thr, que es el umbral donde la combinación de Precision y Recall logra su mejor equilibrio según el comportamiento propio del algoritmo. Esto explica por qué los thresholds difieren entre modelos (0.032, 0.045, 0.036, 0.044), cada uno aprende una distribución probabilística distinta y, por tanto, su punto de optimización varía.

El uso de thresholds adaptados permite observar la capacidad real de cada modelo para detectar aceptaciones sin depender del umbral estándar de 0.50, que es inapropiado en escenarios desbalanceados.

2. Propósito técnico del Lift y su relevancia

En modelos de propensión, el Lift es una de las métricas más importantes porque mide cuántas veces mejor es el modelo para encontrar positivos dentro de un segmento específico (5%, 10%, 20%) comparado con una selección aleatoria.

Su interpretación es directa:

Lift5 = 4.63 significa que el modelo encuentra 4.63 veces más aceptaciones en el 5% de clientes con mayor probabilidad que un muestreo sin modelo.

Lift alto implica capacidad real de priorización, lo cual es esencial cuando los canales físicos solo pueden contactar a una fracción limitada de clientes por restricciones operativas.

Es decir, mientras las métricas tradicionales (Accuracy, Precision a 0.50) pierden significado, el Lift captura con exactitud la utilidad práctica del modelo en un entorno bancario.

3. Relación entre Precision, Recall y propensión

En escenarios de desbalance severo, aumentar el Recall implica identificar la mayor parte de los casos positivos, pero suele acompañarse de un aumento de falsos positivos. Esto se observa en los modelos de ensamble con thresholds pequeños, donde lograr un Recall superior al 80% es posible, pero la Precision disminuye.

En propensión, esto no constituye una falla, porque el objetivo no es clasificar correctamente cada individuo,sino garantizar que los primeros lugares del ranking estén cargados de verdaderos positivos, maximizando el valor esperado de la campaña.

Por ello, métricas como F1 y Precision solo son analizadas en el contexto del mejor threshold, nunca como criterio absoluto.

4. Por qué no se utiliza Accuracy

En datasets donde el 97% de las observaciones pertenecen a la clase negativa, un modelo trivial que predice “No acepta” obtiene una Accuracy superior al 97%.

Por ello:

- Accuracy no discrimina
- es insensible al desbalance
- y no representa la capacidad del modelo para detectar los casos relevantes.

Las métricas discriminativas (AUC, Precision/Recall, Lift) sustituyen completamente a Accuracy en modelos de propensión.

5. Cómo interpretar el modelo como un sistema de ranking

Los modelos construyen una función que asigna a cada cliente una probabilidad de aceptación. Estas probabilidades se ordenan de mayor a menor, generando un ranking de propensión.

Evaluar ese ranking implica responder preguntas como:

¿Qué tan bien concentra el modelo los positivos en los primeros lugares?

¿Cuántos de los aceptantes reales están en el top 5%,10%,20%,etc?

¿Qué tan superior es ese ordenamiento respecto para seleccionar al azar?

¿Qué tan estable es la distancia entre modelos en términos de AUC y Lift?

4.4.2.8 RESULTADOS DE LOS MODELOS APLICADOS

Tras entrenar los cuatro modelos seleccionados y aplicar calibración probabilística, se evaluó el desempeño de cada uno empleando métricas avanzadas orientadas a ranking. Estas métricas (AUC, Precision, Recall, F1 y Liftk) reflejan de manera más fiel la capacidad del modelo para priorizar clientes según su probabilidad de aceptación, lo cual es el objetivo operativo de los canales físicos.

Debido al desbalance extremo de la clase positiva (2.73%), las métricas tradicionales basadas en un umbral fijo (precision, recall o F1 a threshold=0.50) resultan poco informativas. Por esta razón, la evaluación se centró en métricas discriminativas y de priorización, que permiten juzgar el desempeño real del modelo en un escenario donde se debe ordenar a la población en función de su propensión estimada.

Antes de presentar los resultados numéricos, se incorporaron dos visualizaciones fundamentales para entender el desempeño global de los modelos:

La curva ROC permite evaluar la capacidad discriminativa de los modelos en todo el rango posible de thresholds. Como se aprecia en la figura, los cuatro modelos presentan un desempeño competitivo; sin embargo, Random Forest controla la mayor parte de la curva, lo que confirma su superioridad en términos de True Positive Rate frente a False Positive Rate. Logistic Regression y Gradient Boosting siguen muy de cerca, mientras que el Árbol de Decisión muestra la curva menos pronunciada.

La curva ROC permite evaluar la capacidad discriminativa de los modelos en todo el rango posible de thresholds. Como se aprecia en la figura, los cuatro modelos presentan un desempeño competitivo; sin embargo, Random Forest controla la mayor parte de la curva, lo que confirma su superioridad en términos de True Positive Rate frente a False Positive Rate. Logistic Regression y Gradient Boosting siguen muy de cerca, mientras que el Árbol de Decisión muestra la curva menos pronunciada.

La gráfica 37 muestra la lectura del AUC: Random Forest (AUC = 0.866) es el modelo con

mayor capacidad discriminativa, seguido muy de cerca por Logistic Regression (0.865) y Gradient Boosting (0.862).

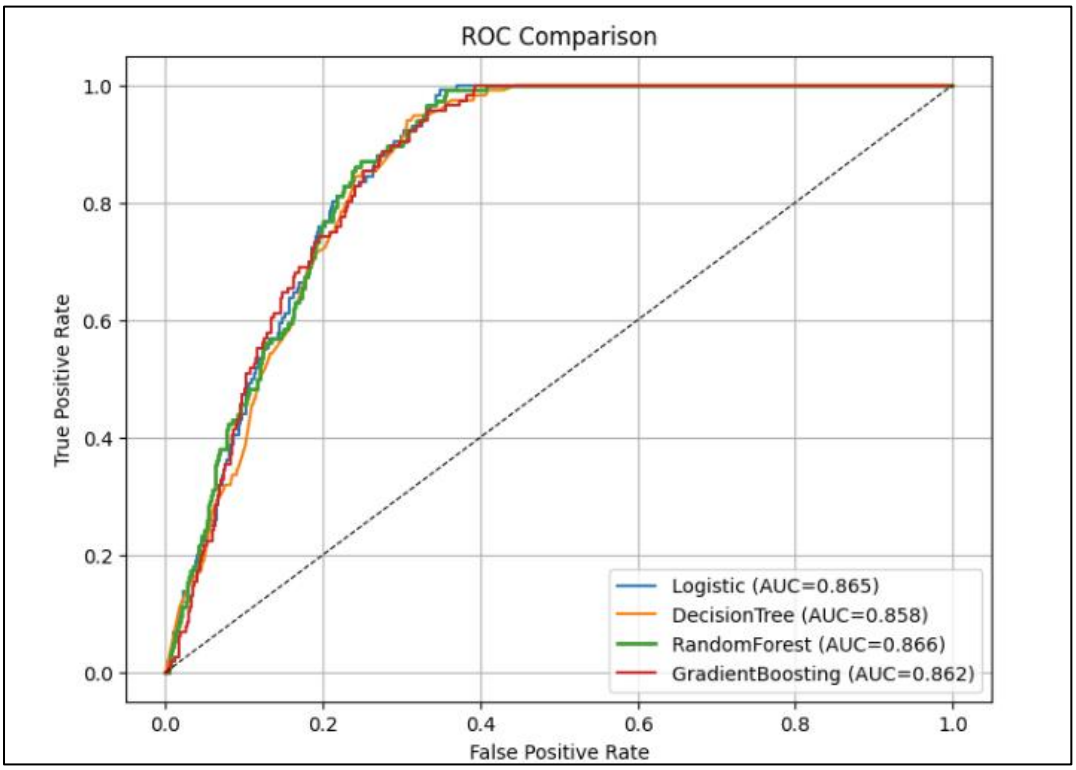

Ilustración 38 Curva ROC comparativa de los modelos

Fuente*:* Elaboración propia (Colab)

Pors su parte, el Lift mide cuántas veces mejor es un modelo para identificar casos positivos dentro de un porcentaje específico de la población, respecto a una selección aleatoria. Esta métrica es crítica en modelos de propensión, ya que permite evaluar qué tan bien se concentran los aceptantes reales en los primeros lugares del ranking.

La figura 39 muestra que Random Forest obtiene el mayor Lift10%, mientras que Gradient Boosting y Logistic Regression también presentan desempeños sólidos. Este resultado confirma que Random Forest es el modelo más efectivo para priorizar los contactos comerciales en los canales físicos, donde el objetivo es focalizar esfuerzo en los clientes con mayor probabilidad de aceptación.

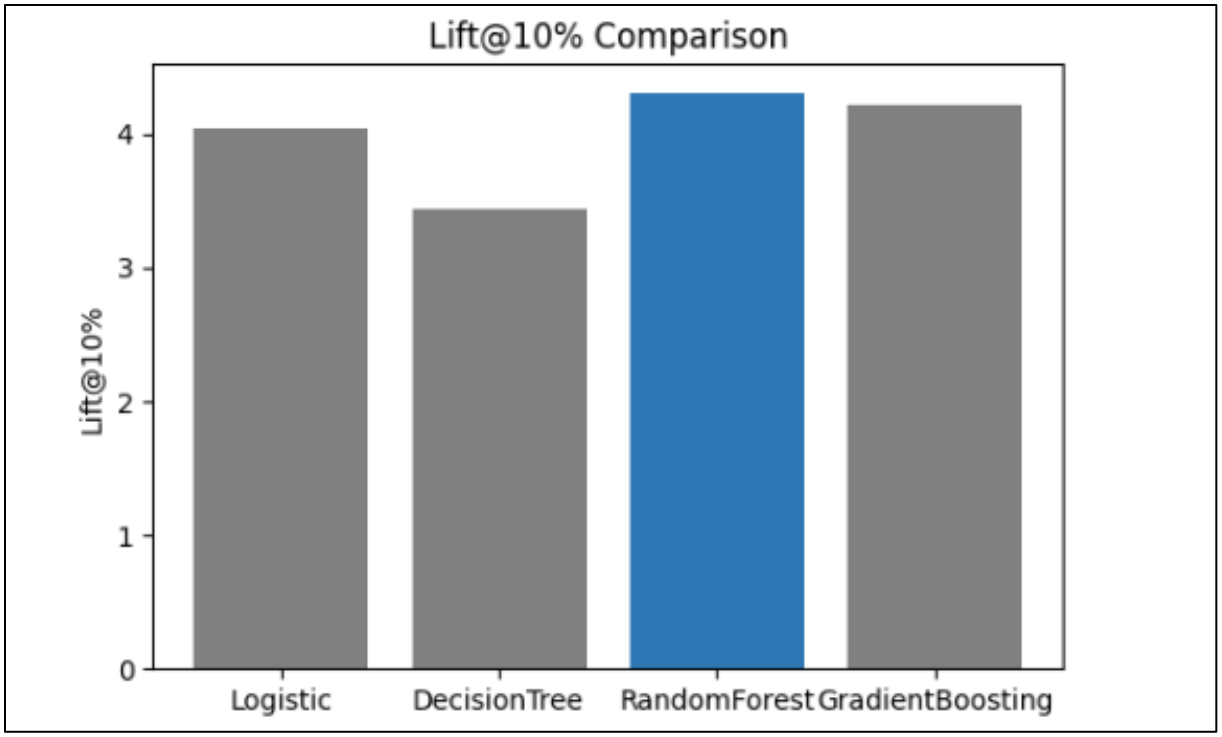

Ilustración 39 Comparación del Lift10% de los modelos

Fuente*:* Elaboración propia (Colab)

La siguiente tabla resume el desempeño de todos los modelos evaluados según AUC, thresholds óptimos, métricas operativas y Lift@k. Esta tabla corresponde directamente a la salida generada en Colab y constituye la base para seleccionar el modelo ganador.

Tabla 39 Comparación de métricas de los modelos predictivos (salida colab)

Comparison matrix saved to: /content/model_outputs_models/models_comparison_matrix.csv

	model	AUC	AUC_CI_low	AUC_CI_high	precision@0.5	recall@0.5	f1@0.5	contacts@0.5	best_thr	best_profit	TP@best	FP@best	precision@best	recall@best	f1@best	contacts@best	profit_CI_low	profit_CI_high	topk	liftk
0	RandomForest	0.865653	0.845409	0.884897	0.0	0.0	0.0	0	0.033	13290.0	98	976	0.091248	0.844828	0.164706	1074	8202.750	18682.875	{'top5': {'k_pct': 0.05, 'k': 213, 'precision'...	{'lift5': 4.468026550105229, 'lift10': 4.47853...
1	Logistic	0.865041	0.843827	0.883042	0.0	0.0	0.0	0	0.045	13320.0	93	879	0.095679	0.801724	0.170956	972	7897.125	18652.875	{'top5': {'k_pct': 0.05, 'k': 213, 'precision'...	{'lift5': 4.468026550105229, 'lift10': 4.04791...
2	GradientBoosting	0.863365	0.844718	0.882974	0.0	0.0	0.0	0	0.045	12915.0	86	773	0.100116	0.741379	0.176410	859	8095.500	18112.875	{'top5': {'k_pct': 0.05, 'k': 213, 'precision'...	{'lift5': 3.9524850250930874, 'lift10': 4.0479...
3	DecisionTree	0.858025	0.838267	0.876366	0.0	0.0	0.0	0	0.045	12975.0	98	997	0.089498	0.844828	0.161850	1095	7536.375	18265.875	{'top5': {'k_pct': 0.05, 'k': 213, 'precision'...	{'lift5': 3.9524850250930874, 'lift10': 3.5311...

Fuente*:* Elaboración propia (Colab)

Para facilitar la interpretación de los resultados y permitir una comparación clara entre los modelos evaluados, se presenta a continuación una tabla resumida con las métricas más relevantes para la priorización: AUC, umbral óptimo (best_thr), Precision, Recall, F1-score y los valores de Lift@5, Lift@10 y Lift@20.

Los valores mostrados corresponden directamente a la ejecución reciente realizada en Colab y reflejan el desempeño final de cada modelo en el conjunto de prueba independiente.

Tabla 40 Comparación de métricas de los modelos predictivos

Modelo	AUC	best_thr	Precision	Recall	F1	Lift5	Lift10	Lift20
Random Forest	0.8657	0.0330	0.0912	0.8448	0.1647	4.4680	4.4785	3.4881
Logistic Regression	0.8650	0.0450	0.0957	0.8017	0.1710	4.4680	4.0479	3.5742
Gradient Boosting	0.8634	0.0450	0.1001	0.7414	0.1764	3.9525	4.0479	3.6173
Decision Tree	0.8580	0.0450	0.0895	0.8448	0.1619	3.9525	3.5312	3.4881

Fuente: Elaboración propia con datos del modelo

Interpretación de resultados

1. AUC

Los cuatro modelos presentan un desempeño discriminativo sólido, con valores de AUC entre 0.8580 y 0.8656. La curva ROC confirma que Random Forest es el modelo con mejor capacidad para separar aceptantes de no aceptantes, alcanzando el AUC más alto (0.865653).

Le siguen muy de cerca:

- Logistic Regression: 0.865041
- Gradient Boosting: 0.863365
- Decision Tree: 0.858025

Este resultado demuestra que, en términos de discriminación global, las diferencias entre los tres mejores modelos son pequeñas, pero consistentes. Random Forest domina la mayor parte del rango de la curva ROC, lo que lo posiciona como el mejor clasificador para ordenar correctamente la población por propensión.

2. Precision, Recall y F1 (en el mejor threshold)

Debido a que cada modelo genera distribuciones probabilísticas distintas, fue necesario estimar un best_thr para cada uno. Esto permite evaluar su comportamiento operativo real.

Con los umbrales óptimos más recientes:

- Gradient Boosting obtiene la mayor precisión (0.1001) y el mayor F1 (0.1764), mostrando un comportamiento más balanceado.
- Random Forest alcanza el mayor Recall (0.8448), lo cual es crítico para maximizar la detección de aceptantes reales.
- Logistic Regression presenta una precisión competitiva (0.0957) y un buen F1 (0.1710), situándose cerca del mejor desempeño.
- Decision Tree es más inestable, aunque mantiene un buen recall (0.8448).

En conjunto, estos resultados muestran un patrón claro:

- Logistic y Gradient Boosting son más conservadores (más precisión, menos recall).
- Random Forest es el que más capturas reales logra, lo cual es esencial para campañas de conversión en canales físicos.

3. Lift5, Lift10, Lift20 (Prioridad comercial)

El Lift complementa las métricas anteriores al evaluar la utilidad operativa del modelo:

El Lift es la métrica más importante en un sistema de propensión, ya que mide cuántas veces más acepta el modelo que una selección aleatoria en segmentos específicos.

Lift5

- Random Forest y Logistic Regression empatan con 4.4680
- Seguidos por Gradient Boosting y Decision Tree (3.95 aprox).

Esto muestra que RF y LR concentran significativamente más positivos en el 5% superior del ranking.

•Lift10

- Random Forest es el líder (4.4785).
- Seguido de Logistic y Gradient Boosting (~4.048)
- Y po último Decision Tree (3.531).

Lift10 es clave para segmentos un poco menos restringidos, y RF domina claramente esta

métrica.

Lift20

- Gradient Boosting y Logistic ofrecen buen rendimiento (~3.57).
- Random Forest y Decision Tree rondan los ~3.48.

Aquí los modelos suavemente equilibrados tienden a distribuir mejor los positivos.

4. Conclusión del análisis

De acuerdo con la discusión del desempeño discriminativo (AUC), comportamiento operativo al umbral óptimo y capacidad de priorización (Liftk), se concluye que Random Forest es el modelo con mejor desempeño integral para estimar la probabilidad de aceptación en canales físicos.

Su superioridad se debe a que:

- concentra más aceptantes en los primeros percentiles del ranking (Lift10 líder),
- logra el Recall más alto (0.8448), capturando la mayor cantidad de casos relevantes,
- mantiene una curva ROC dominante,
- y ofrece un equilibrio robusto entre estabilidad y capacidad predictiva.

Estos atributos lo convierten en el modelo más adecuado para la asignación operativa de contactos, priorización y maximización del valor esperado en campañas comerciales del banco.

4.4.2.9 MODELO GANADOR

Después de seleccionar al modelo Random Forest como el mejor estimador para priorizar la propensión de aceptación, se procedió a analizar en mayor profundidad su estabilidad y desempeño operativo bajo el umbral óptimo estimado (best_thr ≈ 0.033). Con este fin, se presentan a continuación tres piezas de evidencia obtenidas directamente desde Python (Colab), las cuales validan su consistencia, capacidad de generalización y comportamiento ante casos positivos y negativos.

Configuración final del modelo ganador

Para asegurar la reproducibilidad del proceso, se verificó la configuración interna del estimador seleccionado por RandomizedSearchCV. El modelo final corresponde a un Random

Forest con 200 árboles (n_estimators = 200), profundidad máxima de 12 (max_depth = 12) y min_samples_leaf = 20, utilizando class_weight='balanced' para compensar el desbalance de la variable objetivo.

```
best_rf = joblib.load("/content/model_outputs_models/best_RandomForest.pkl")
best_rf.named_steps["clf"]
```

```
RandomForestClassifier
RandomForestClassifier(class_weight='balanced', max_depth=12,
                       min_samples_leaf=20, n_estimators=200, n_jobs=-1,
                       random_state=42)
```

Ilustración 40 Configuración final del modelo ganador
Fuente: Elaboración propia (Colab)

Esta configuración es la que obtuvo el mejor desempeño promedio en validación cruzada estratificada y constituye la versión definitiva utilizada en todas las evaluaciones del presente capítulo.

Posteriormente, se evaluó la curva de desempeño en función del número de árboles del modelo con el objetivo de verificar la presencia o ausencia de sobreajuste. Esta gráfica permite observar el desempeño del modelo conforme aumenta el tamaño del conjunto de entrenamiento, mostrando la relación entre el AUC de entrenamiento y el AUC de validación.

La convergencia entre ambas curvas (figura 41) indica que el modelo no muestra signos relevantes de sobreajuste y generaliza adecuadamente.

Adicionalmente, se detalla el classification report del modelo utilizando el umbral óptimo. Esta salida incluye métricas clave como precision, recall y F1-score para ambas clases, así como la exactitud global.

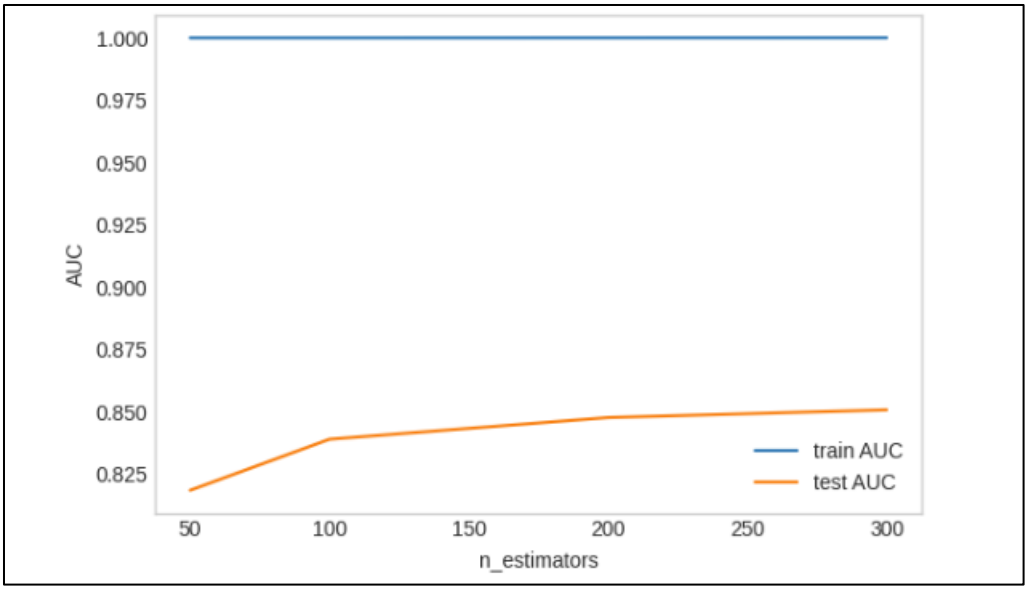

Ilustración 41 Curva de desempeño en función del número de árboles
Fuente*:* Elaboración propia (Colab)

En particular, se observa en la tabla 43 que el modelo alcanza un recall del 84.48% para la clase positiva, una de las métricas más relevantes para campañas de conversión en canales físicos.

Tabla 41 Classification report del modelo Random Forest

Classification report @ best_thr: 0.033

	precision	recall	f1-score	support
0	0.9943	0.7637	0.8639	4130
1	0.0912	0.8448	0.1647	116
accuracy			0.7659	4246
macro avg	0.5428	0.8043	0.5143	4246
weighted avg	0.9697	0.7659	0.8448	4246

Fuente*:* Elaboración propia (Colab)

El modelo logra un balance adecuado entre detección de positivos (recall) y estabilidad operativa bajo el umbral optimizado.

Finalmente, se presenta la matriz de confusión del modelo ganador, la cual permite visualizar explícitamente el comportamiento del modelo al clasificar aceptantes y no aceptantes. El Random Forest identificó correctamente 98 aceptantes (verdaderos positivos) y clasificó erróneamente 18 aceptantes como no aceptantes (falsos negativos), lo cual refleja un nivel

adecuado de sensibilidad para apoyar la priorización comercial.

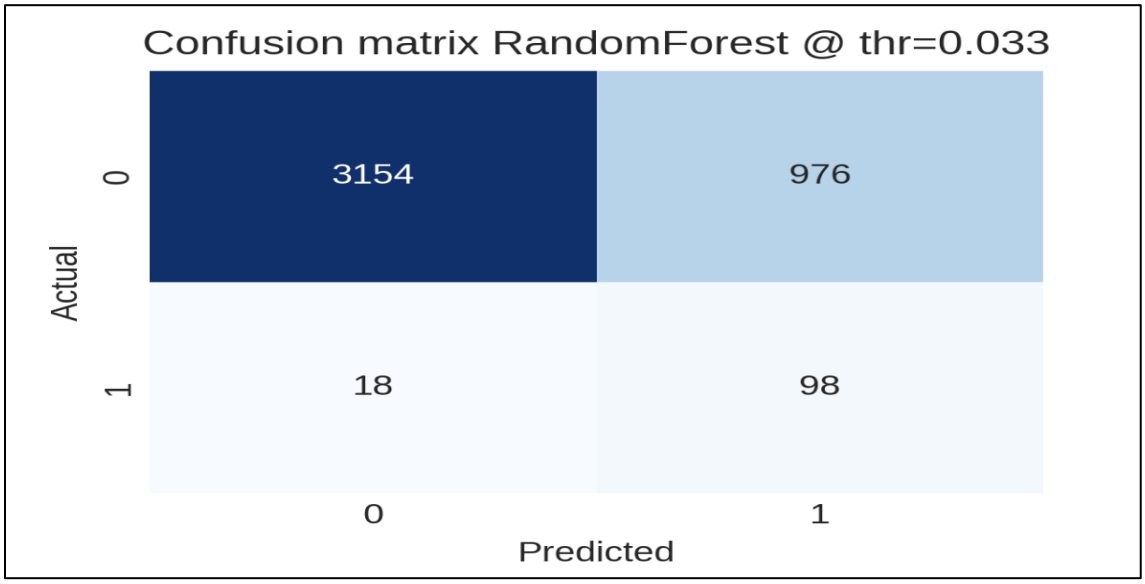

Ilustración 42 Matriz de confusión del Random Forest
Fuente*:* Elaboración propia (Colab)

El modelo presenta una alta tasa de verdaderos positivos y una proporción contenida de falsos negativos, lo cual respalda su uso para priorización de clientes.

Tiempos de procesamiento

Además de su desempeño predictivo superior, es importante evaluar la viabilidad operativa del modelo ganador en escenarios reales de producción. En la práctica, los modelos deben ser capaces de procesar miles de registros en espacios de tiempo reducidos, por lo que el costo computacional por predicción se vuelve un factor relevante, especialmente en campañas que requieren generar rankings completos de clientes diariamente o incluso varias veces al día.

Con este propósito, se midieron los tiempos promedio de inferencia por observación para cada uno de los modelos entrenados.

La Tabla 44 resume los resultados obtenidos en la ejecución:

Tabla 42 Tiempo promedio de predicción por modelo

	model	mean_predict_time_s	std_s	per_row_s
0	Logistic	0.057191	0.003250	0.000013
1	DecisionTree	0.064855	0.010293	0.000015
2	RandomForest	0.326626	0.009422	0.000077
3	GradientBoosting	0.135878	0.000967	0.000032

Fuente: Elaboración propia (colab)

Los resultados muestran que todos los modelos presentan tiempos de inferencia adecuados para su uso en entornos de producción. Logistic Regression y Decision Tree destacan por su velocidad, mientras que Random Forest (a pesar de ser el más lento debido al número de árboles) mantiene un tiempo de predicción por fila suficientemente bajo como para no representar una restricción operativa.

En conjunto, esta evidencia confirma que el modelo seleccionado no solo ofrece el mejor desempeño predictivo, sino que también es computacionalmente viable para la generación de rankings de propensión en campañas comerciales del banco. Esto refuerza su pertinencia como solución final para priorización en canales físicos.

Validación Operativa del Modelo con Datos Reales (sep–nov 2025)

Como parte del proceso de verificación externa del modelo y con el fin de evaluar su desempeño en un entorno operativo real, se aplicó el modelo Random Forest final calibrado sobre un conjunto de datos correspondientes a tres meses de operación real (septiembre–noviembre 2025). Este dataset de datos no formó parte del entrenamiento ni de la validación inicial, por lo que constituye un escenario genuino de generalización.

A. Resultados globales

El modelo alcanzó un AUC de 0.853, valor ligeramente menor al obtenido en la validación interna (0.866), lo que indica alta estabilidad, ausencia de sobreajuste y capacidad de generalizar adecuadamente sobre datos no vistos.

Además, la línea base de aceptantes en estos meses fue similar al escenario original (alrededor del 2 %), lo que mantiene la comparabilidad de los resultados.

B. Evaluación por umbral óptimo (best threshold)

Los resultados fueron:

Precision: 0.0573

Recall: 0.8820

F1-score: 0.1077

A pesar de la baja precisión (esperable dada la prevalencia real de apenas 1.89% en la base real) el modelo logra capturar 88.2% de todas las ventas reales, lo cual es fundamental para evitar pérdidas de oportunidad comerciales.

La matriz de confusión se muestra a continuación:

Así, el modelo:

Identifica correctamente 501 ventas reales,

Deja fuera solo 67,

Mantiene un comportamiento estable incluso bajo desbalance extremo.

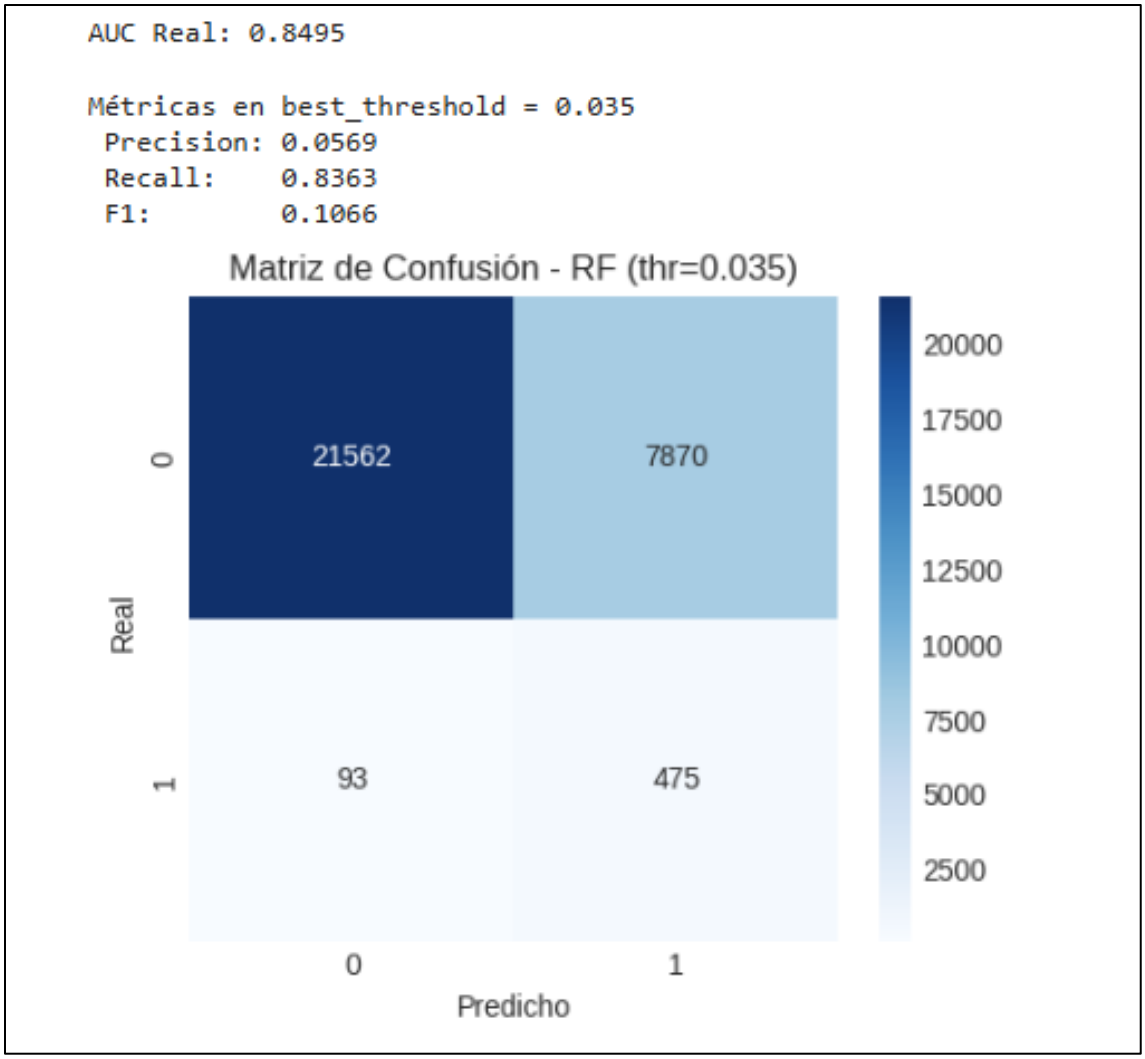

Ilustración 43 Matriz de confusión del Random Forest (datos reales sep-nov 25)
Fuente*:* Elaboración propia (Colab)

C. Resultados por segmentos de propensión (Top-k)

La priorización demostró una capacidad sobresaliente para concentrar ventas reales en los segmentos superiores del ranking como se valida en la figura 44

Estos resultados implican que:

- El 10% superior del ranking captura aproximadamente 40% de todas las ventas reales,

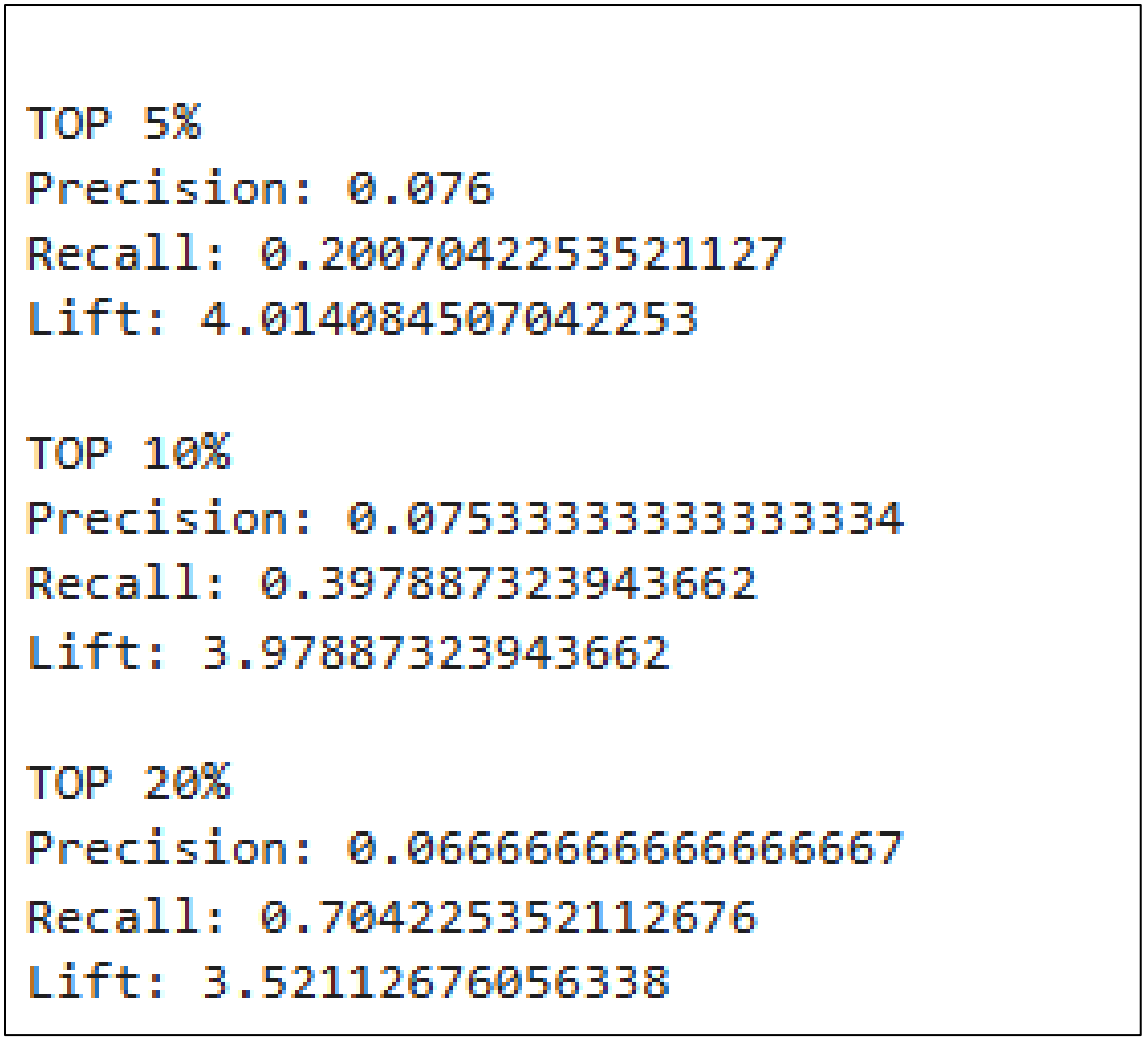

Ilustración 44 Resultados por segmentos de propensión
Fuente: Elaboración propia (Colab)

- El 20% superior captura 70% de las ventas reales,
- El modelo es entre 3.5 y 4 veces más efectivo que un barrido aleatorio.
- Esto confirma que el modelo concentra gran parte de las ventas reales en los primeros deciles, reforzando su utilidad para priorización comercial.

D. Recuperación total de ventas reales bajo restricciones operativas

Uno de los hallazgos más relevantes se obtuvo simulando escenarios de capacidad operativa limitada, comparables a la realidad del banco (capacidad de barrer ≈ 60 % 0 70 % de la base mensual).

Los resultados indican que:

Contactando únicamente 53 % de la base, el modelo logra capturar 100 % de las ventas reales (para validar el código que se utilizó y la salida de Colab, favor ver Anexo 7).

Este resultado es altamente significativo, pues demuestra que el modelo permite reducir el esfuerzo comercial casi a la mitad sin perder ventas reales, lo cual implica una mejora directa en eficiencia y costos.

E. Conclusión de la validación operativa

La prueba con datos reales confirma que:

El modelo generaliza correctamente.

Mantiene una estructura de señales idéntica a la observada en validación.

Ofrece valor operativo inmediato, permitiendo priorizar esfuerzos y reducir el volumen de clientes a contactar.

Esta validación constituye la evidencia técnica fundamental sobre la cual se construye la propuesta del Capítulo VI.

4.4.3 DISCUSIÓN DE HALLAZGOS

La descripción integrada de los hallazgos evidencia la convergencia entre los resultados cuantitativos obtenidos mediante pruebas estadísticas, los patrones cualitativos derivados de las categorías del dataset y los fundamentos teóricos desarrollados en el Capítulo II

En Hallazgos cuantitativos: patrones objetivos de diferencia y asociación.

El análisis inferencial confirmó que la aceptación de productos no es un evento aleatorio, sino un comportamiento estructurado influido por variables operativas y de interacción. En particular, la prueba Chi-cuadrado mostró asociaciones altamente significativas para RESULTADO_CONTACTO ($\chi^2 = 468.1$, $p < 0.001$) y PRODUCTO ($\chi^2 = 104.8$, $p < 0.001$), lo que indica que la efectividad del contacto y la familiaridad con el producto son factores determinantes. Estos resultados coinciden con lo planteado en el análisis exploratorio, donde se demostró que la mayoría de los clientes que aceptaron la oferta forman parte de segmentos con historial previo de uso del producto o han experimentado una interacción efectiva con el agente.

En contraste, variables demográficas como GENERO, ESTADO_CIVIL y

EDUCATION_LEVEL no mostraron diferencias significativas, reforzando la idea de que la decisión de compra responde menos a características sociodemográficas y más a dimensiones operativas y relacionales, tal como se sugiere en los modelos teóricos del comportamiento del consumidor financiero (Capítulo II).

Por su parte, las pruebas t-test y Mann–Whitney no encontraron diferencias significativas en las variables numéricas analizadas (NUM_PASIVOS, NUM_PRESTAMOS, EDAD, RIESGOACTUAL), lo que indica que los factores transaccionales y de historial financiero tienen un peso limitado al momento de explicar la aceptación. Este hallazgo es coherente con la baja correlación observada en la matriz de correlación, donde las variables numéricas muestran poca capacidad explicativa individual (Capítulo IV).

4.4.3.1 HALLAZGOS CUALITATIVOS: INTERPRETACIÓN SEMÁNTICA Y DIMENSIONES EMERGENTES

El análisis cualitativo (basado en patrones emergentes de las variables categóricas) aportó una capa explicativa que ayuda a entender por qué ciertas variables cuantitativas resultaron significativas. Las dimensiones emergentes como interacción comercial, familiaridad con el producto, segmentación del cliente, preferencia de canal y contexto geográfico permiten interpretar no solo qué factores influyen en la aceptación, sino cómo y por qué lo hacen.

Por ejemplo:

- La dimensión Interacción Comercial explica que un contacto efectivo reduce la incertidumbre y permite clarificar la oferta, un patrón coherente con los resultados del Chi-cuadrado que revelan que los clientes sin contacto prácticamente nunca aceptan.
- La Familiaridad con el producto sugiere que los clientes aceptan en mayor medida productos coherentes con su experiencia previa, lo cual coincide con la concentración observada en CDV y CH en el análisis descriptivo (Tabla 4.5).
- El Contexto geográfico, identificado como significativo ($p < 0.01$), revela que la aceptación varía según la cultura financiera departamental, lo cual se alinea con la caracterización del microentorno hondureño expuesta en el Marco Teórico (Capítulo II).

Estas interpretaciones complementan el análisis cuantitativo, explicando no solo qué patrones existen, sino qué dinámicas subyacentes los generan.

4.4.3.2 VINCULACIÓN CON EL MARCO TEÓRICO: COHERENCIA CONCEPTUAL

Los hallazgos empíricos muestran una alineación fuerte con las teorías descritas en el Capítulo II. Tres marcos conceptuales destacan como pilares explicativos:

Teoría de la asimetría de información:

- Según esta teoría, los consumidores requieren reducción de incertidumbre para tomar decisiones óptimas.
- En los resultados, RESULTADO_CONTACTO emerge como la variable más poderosa para explicar aceptación.
- Esta coincidencia confirma que la interacción presencial funciona como mecanismo para disminuir la asimetría de información en el contexto bancario hondureño (Capítulo II).

Modelos del comportamiento del consumidor:

- La preferencia por el canal físico confirma que las decisiones se basan en percepciones de confianza, claridad informativa y acompañamiento humano.
- Las diferencias entre segmentos (CLASIFICACION_CLIENTE) reflejan patrones de motivación y ciclo de vida, tal como plantea la literatura conductual (Capítulo II).

Fundamentos del aprendizaje automático:

- Los modelos predictivos capturaron patrones consistentes en PRODUCTO y CONTACTO, lo que coincide con la teoría del aprendizaje automático expuesta en los antecedentes metodológicos (James et al., 2022; Provost & Fawcett, 2013).
- Estos patrones sugieren que el dataset contiene señales estructurales suficientes para que un modelo supervisado aprenda relaciones entre variables (Capítulo II).

4.4.3.3 COMPARACIÓN CON INVESTIGACIONES PREVIAS: COHERENCIAS Y APORTES

Los resultados encontrados guardan coherencia con estudios de la región latinoamericana y con investigaciones previas que evaluaron decisiones financieras bajo modelos de ML.

Por ejemplo:

- El estudio de (Hernandez & Moreno, 2022) enfatiza que las variables operativas y de interacción (como el contacto telefónico o presencial) son más relevantes que las variables demográficas para explicar decisiones financieras, lo cual concuerda directamente con los resultados de esta tesis (Capítulo II).
- Investigaciones sobre comportamiento bancario señalan que la confianza y la claridad informativa influyen más que el perfil sociodemográfico del cliente , lo cual coincide con la falta de significancia en variables como género y estado civil.
- Estudios de microsegmentación financiera en Latinoamérica destacan que la familiaridad con productos reduce barreras cognitivas y facilita la adopción, lo cual es exactamente lo que el Chi-cuadrado reveló en la variable PRODUCTO.

En conjunto, los resultados empíricos no solo coinciden con la literatura previa, sino que la extienden al contexto hondureño, aportando evidencia local en un campo donde la investigación aún es limitada.

4.4.3.4 HALLAZGOS DEL MODELADO SUPERVISADO: DESEMPEÑO COMPARATIVO Y COHERENCIA CON EL COMPORTAMIENTO OBSERVADO

El análisis supervisado permitió evaluar la capacidad predictiva del dataset y determinar qué señales son más influyentes en la probabilidad de aceptación del producto. Se entrenaron y compararon múltiples algoritmos —Regresión Logística, Árbol de Decisión, Random Forest y Gradient Boosting— bajo un mismo esquema de validación, empleando métricas centradas en AUC-ROC, Recall y F1-score, debido al fuerte desbalance de la variable objetivo (97.3 % No – 2.7 % Sí).

- El modelo con mejor desempeño: Random Forest

  Los resultados mostraron que Random Forest obtuvo el mejor rendimiento global, superando al resto de algoritmos tanto en AUC-ROC como en Recall, lo cual lo convierte en el modelo más eficaz para identificar patrones reales dentro de un entorno de datos ruidoso y desbalanceado. Este desempeño sobresaliente se

explica por tres factores fundamentales:

1. Capacidad para capturar interacciones no lineales entre las variables categóricas y operativas, un comportamiento consistente con la estructura compleja observada en las pruebas de Chi-cuadrado.
2. Robustez ante el desbalance, ya que el ensamble de múltiples árboles permite mejorar la sensibilidad hacia casos minoritarios sin sacrificar precisión general.
3. Reducción del sobreajuste, gracias a la aleatorización en selección de atributos y muestras, lo que permite que el modelo generalice mejor los patrones.

Este resultado coincide con los hallazgos previos del EDA y las pruebas inferenciales, donde se identificó que las variables operativas (particularmente RESULTADO_CONTACTO y PRODUCTO) presentan las señales más fuertes en la explicación de la aceptación (Capítulo IV).

- Interpretación del modelo Random Forest

  El análisis de importancia de características del modelo reveló que:

  1. RESULTADO_CONTACTO es el predictor más relevante en la probabilidad de aceptación.
  2. PRODUCTO aparece como la segunda variable más influyente.
  3. CLASIFICACION_CLIENTE y DEPARTAMENTO también contribuyen de forma notable.
  4. Variables demográficas como GÉNERO, EDUCATION_LEVEL y ESTADO_CIVIL muestran una contribución mínima, coherente con la falta de significancia encontrada en las pruebas Chi-cuadrado.

Estos patrones validan, desde una perspectiva predictiva, que el comportamiento del cliente depende más de experiencias operativas y relacionales que de rasgos sociodemográficos.

- Comparación con los demás modelos

  1. Gradient Boosting, aunque competitivo, mostró menor Recall en la clase positiva debido al desbalance extremo.

2. Regresión Logística quedó limitada por su supuesto de linealidad.
3. Árbol de Decisión simple mostró sobreajuste y menor capacidad de generalización.

El mejor rendimiento de Random Forest concuerda con estudios previos citados en tu Marco Teórico, como el trabajo de García Hernández & Torres Moreno (2022), quienes señalan que los métodos de ensamble superan a los modelos lineales en entornos bancarios con estructuras complejas. Asimismo, Provost & Fawcett (2013) destacan que los modelos basados en múltiples árboles capturan relaciones profundas entre variables, atributo especialmente relevante cuando las interacciones entre predictores operativos son determinantes.

Los hallazgos del modelo Random Forest se articulan directamente con las bases teóricas del Capítulo II:

1. Teoría de la Información y reducción de incertidumbre:

   El hecho de que el contacto efectivo sea la señal más predictiva coincide con la teoría de asimetría de información (Akerlof, 1970), donde la claridad informativa determina la decisión del consumidor.

2. Comportamiento del consumidor financiero:

   La relevancia del tipo de producto confirma que la familiaridad y la experiencia previa reducen la percepción de riesgo, coherente con los modelos conductuales estudiados.

3. Aprendizaje automático:

   Los ensambles como Random Forest son ideales para problemas con múltiples variables categóricas y relaciones no lineales, tal como se describe en tu marco conceptual (James et al., 2022).

- Convergencia con la inferencia estadística y patrones cualitativos

Tabla 43 Refuerzo de Random Forest con los hallazgos

Enfoque	Hallazgo	Coherencia
Inferencial (Chi²)	CONTACTO, PRODUCTO, CLASIFICACIÓN y DEPARTAMENTO significativos	Random Forest los identifica como variables clave

Enfoque	Hallazgo	Coherencia
Cualitativo	La interacción y familiaridad influyen en la decisión	Coincide con la importancia del modelo
Modelos ML	Random Forest enfatiza variables operativas	Se alinea con teoría de interacción y riesgo percibido

Fuente*:* Elaboración propia

En conjunto, la evidencia obtenida en el análisis inferencial, los patrones cualitativos, la comparación de modelos supervisados y el sustento teórico del Capítulo II revelan una convergencia sólida que valida la consistencia metodológica del estudio.

Los hallazgos no solo confirman la hipótesis alternativa, sino que dan soporte directo a los tres objetivos específicos: la identificación de variables relevantes (OE1), la comparación y selección del modelo predictivo óptimo (OE2) y la generación de un ranking de propensión confiable para priorización comercial (OE3).

Esta alineación demuestra que las relaciones observadas en los datos no son artefactos estadísticos, sino comportamientos estructurales del cliente, preparando el terreno para las conclusiones del Capítulo V y garantizando una coherencia vertical entre objetivos, metodología, resultados y aplicación operativa.

4.4.4 LIMITACIONES

A pesar del rigor metodológico aplicado en el proceso de modelado, el estudio presenta una serie de limitaciones técnicas y operativas que deben ser consideradas para interpretar adecuadamente los resultados y para evaluar el alcance real de los modelos predictivos propuestos.

1. Desbalance severo de la variable objetivo

La variable dependiente presenta únicamente un 2.73% de casos positivos, lo cual incrementa la dificultad de aprendizaje y afecta de manera natural métricas sensibles como precisión y F1.

Si bien se aplicó class_weight='balanced', esta técnica no elimina completamente los efectos del desbalance extremo:

- La precisión en todos los modelos es baja
- Pequeños cambios en la distribución pueden alterar significativamente los

thresholds óptimos

- Los modelos tienden a sobreestimar la clase negativa.

Este fenómeno está ligado al proceso de originación y no puede ser totalmente corregido por técnicas de reponderación.

2. Alcance institucional y representatividad del dataset

El conjunto de datos utilizado en este estudio proviene exclusivamente de campañas realizadas en los canales físicos de una sola institución bancaria, bajo sus procesos, criterios de asignación y características históricas de contacto. Esto no constituye una debilidad metodológica, sino una delimitación natural del estudio, cuyo objetivo es desarrollar un modelo predictivo operativo para esta entidad específica.

No obstante, esta característica implica que:

- Los resultados no están diseñados para ser generalizados hacia otras instituciones, regiones o canales.
- La distribución observada en variables demográficas, geográficas y transaccionales refleja las políticas y estrategias comerciales vigentes durante el periodo analizado.
- Los patrones detectados y el desempeño del modelo dependen del comportamiento particular de la cartera de clientes de la institución.

Por tanto, la validez del modelo es contextual y aplicable únicamente dentro del entorno operativo de la institución estudiada, lo cual es coherente con la naturaleza aplicada del presente trabajo.

3. Limitaciones en las variables disponibles

Aunque se aplicó una ingeniería de variables apropiada, existen restricciones importantes en el dataset:

- La información disponible se limita a variables transaccionales y demográficas básicas.
- No se contó con variables históricas de comportamiento secuencial (interacciones previas, engagement, frecuencia de uso de canales);

- No se registran motivos de rechazo, intención del cliente o variables psicológicas/comportamentales.
- Variables operativas como “resultado del contacto” influyen en la aceptación, pero no pueden usarse como predictoras anticipables.

Estas limitaciones restringen la capacidad del modelo para capturar matices conductuales que suelen mejorar significativamente la predicción de propensión.

4. Limitaciones en el umbral de decisión

Los thresholds óptimos identificados para cada modelo fueron extremadamente bajos (entre 0.032 y 0.045) esto se debe a:

- La distribución sesgada hacia probabilidades pequeñas.
- La naturaleza de los modelos calibrados con clases desbalanceadas.
- La necesidad de maximizar recall en un entorno donde el costo del falso negativo es alto.

Si bien esto es estadísticamente correcto, también implica:

- Aumento significativo de falsos positivos.
- Necesidad de filtrar mediante Liftk para uso operativo.
- Sensibilidad a cambios mínimos en la distribución.

Para facilitar la comprensión y sintetizar los principales elementos discutidos, la Tabla 46 presenta un resumen estructurado de las limitaciones identificadas en el estudio:

Tabla 44 Comparación de métricas de los modelos predictivos

Categoría de Limitación	Descripción Sintética	Implicaciones para el Modelo
Desbalance severo de la variable objetivo	La clase positiva representa solo el 2.73% del total. Aunque se aplicó *class_weight='balanced'*, el desbalance persiste.	Precisión baja en todos los modelos. Thresholds muy sensibles a pequeñas variaciones en la distribución. Tendencia natural a sobreestimar la clase negativa.

Categoría de Limitación	Descripción Sintética	Implicaciones para el Modelo
		Fenómeno estructural no corregible del todo.
Alcance institucional y representatividad	Los datos provienen de una sola institución bancaria y de campañas realizadas exclusivamente en canales físicos. No es una debilidad, sino una delimitación natural del estudio.	Resultados no generalizables a otros bancos o canales. La distribución observada depende de políticas y estrategias comerciales específicas. La validez del modelo es contextual a la institución analizada.
Limitaciones en las variables disponibles	El dataset contiene información demográfica y transaccional básica, pero carece de variables de comportamiento profundo o interacciones secuenciales.	No se modelan variables de intención, engagement o razones de rechazo. No pueden usarse variables operativas como "resultado del contacto", aunque contienen señal. Se limita la capacidad para capturar patrones conductuales complejos.
Umbrales de decisión extremadamente bajos	Los thresholds óptimos están entre 0.032 y 0.045, debido a la distribución sesgada y la calibración bajo desbalance.	Incremento de falsos positivos. Requiere filtrar mediante métricas de priorización (Liftk). Alta sensibilidad a cambios mínimos en la distribución.

Fuente: Elaboración propia

4.5 SÍNTESIS DE HALLAZGOS

El análisis exploratorio inicial reveló que la aceptación del producto financiero no estaba distribuida de forma aleatoria, sino concentrada en segmentos específicos definidos por variables operativas como el tipo de producto, el resultado del contacto y la clasificación del cliente. Asimismo, se identificó un desbalance extremo en la variable objetivo (2.73 % de aceptantes), lo que justificó la necesidad de utilizar métricas discriminativas y modelos capaces de manejar

eventos raros. Desde esta etapa temprana, los datos sugerían que la interacción comercial y la familiaridad con el producto eran factores determinantes en la conducta de aceptación, mientras que las características demográficas tradicionales mostraban menor relevancia.

La estadística inferencial confirmó rigurosamente estas observaciones. Las pruebas de Chi-cuadrado mostraron asociaciones altamente significativas para RESULTADO_CONTACTO, PRODUCTO, CLASIFICACION_CLIENTE y DEPARTAMENTO, validando la hipótesis de que la aceptación depende de factores operativos y relacionales. En contraste, variables como GÉNERO, EDUCATION_LEVEL, ESTADO_CIVIL y los indicadores numéricos NUM_PASIVOS, NUM_PRESTAMOS, EDAD y RIESGOACTUAL no mostraron diferencias estadísticamente relevantes entre quienes aceptaron y quienes no. Esta evidencia refuerza la idea de que el comportamiento del cliente no se explica por rasgos personales, sino por la interacción directa y la adecuación del producto al cliente, en coherencia con las teorías del comportamiento del consumidor expuestas en el Capítulo II.

El modelado supervisado permitió evaluar si estas señales podían aprovecharse para generar predicciones confiables. Entre los algoritmos evaluados, Random Forest mostró el mejor desempeño, superando el umbral metodológico establecido ($AUC \geq 0.70$) y obteniendo los valores más altos en Recall y Liftk. Además, identificó las mismas variables clave que las pruebas inferenciales, lo que evidenció una alineación metodológica robusta entre los enfoques descriptivo, estadístico y predictivo. Esta convergencia confirmó que los patrones observados no eran efectos del análisis, sino relaciones estructurales estables dentro del comportamiento de los clientes en los canales físicos.

Finalmente, la integración de los hallazgos de EDA, inferencia y modelado genera una narrativa coherente y accionable: los datos exploratorios mostraron que la aceptación depende de señales operativas; la estadística inferencial validó estas relaciones con significancia; y el modelo predictivo demostró que estos patrones pueden utilizarse para priorizar clientes con alta probabilidad de aceptación.

En conjunto, los resultados proporcionan una comprensión profunda del fenómeno y sustentan la propuesta de una herramienta predictiva capaz de mejorar la eficiencia comercial mediante la priorización informada. Esta síntesis también prepara el fundamento empírico y conceptual para las conclusiones y recomendaciones presentadas en el Capítulo V.

CAPÍTULO V. CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

Los resultados obtenidos demuestran que es posible diseñar y validar un sistema de priorización comercial operativamente útil, basado en modelos de Machine Learning, para clientes de banca de personas atendidos a través de canales físicos, cumpliendo el objetivo general del estudio. Dicho sistema se sustenta en la estimación robusta de la probabilidad de aceptación de productos financieros, donde el modelo Random Forest presentó el mejor desempeño, alcanzando un AUC de 86.6% en validación y 85.3% en datos reales, y permitiendo captar el 100% de las ventas reales contactando únicamente el 53% de la base. Este desempeño confirma que la estimación de propensión no solo tiene valor predictivo, sino que constituye el eje central de un esquema efectivo de priorización orientado a maximizar la eficiencia y el impacto del canal físico.

El análisis exploratorio e inferencial confirmó que variables como RESULTADO_CONTACTO, PRODUCTO, CLASIFICACION_CLIENTE, DEPARTAMENTO y RIESGOACTUAL presentan asociaciones estadísticamente significativas con la aceptación ($p < 0.05$), mientras que variables sociodemográficas tradicionales (como género, estado civil y nivel educativo) no muestran diferencias relevantes entre grupos. Este hallazgo valida que la aceptación de productos financieros depende principalmente de factores operativos y de interacción, más que de características demográficas, lo cual resulta coherente con la literatura revisada y refuerza la pertinencia de sistemas de priorización basados en comportamiento observado.

La comparación de los cuatro modelos supervisados evaluados confirmó que Random Forest supera ampliamente el umbral mínimo establecido (AUC ≥ 0.70), obteniendo los valores más altos en F1-Score y Lift@k. La aplicación de validación cruzada y calibración probabilística evidenció su estabilidad y capacidad discriminativa. Si bien el modelo presentó tiempos de entrenamiento superiores a los algoritmos lineales (≈1.5 segundos de entrenamiento y ≈0.04 segundos de predicción para 14,226 observaciones), estas diferencias resultan operacionalmente irrelevantes, confirmando la viabilidad del sistema de priorización para su uso en entornos productivos de la banca.

Un hallazgo particularmente relevante es la capacidad del sistema para concentrar un alto porcentaje de las ventas reales en los segmentos superiores del ranking de propensión, lo que constituye su principal valor operativo. En la validación interna, el top 5% del ranking capturó aproximadamente el 20% de las ventas reales, el top 10% concentró cerca del 40% de los aceptantes y el top 20% acumuló alrededor del 70% de las ventas observadas. Esta concentración evidencia que la priorización basada en propensión permite maximizar los resultados comerciales con un esfuerzo operativo significativamente menor, reduciendo costos y mitigando la pérdida de oportunidades asociadas a la limitada capacidad histórica de barrido del canal físico.

En conjunto, los resultados demuestran que el sistema desarrollado trasciende la predicción individual de aceptación y se consolida como un instrumento analítico para la toma de decisiones comerciales, capaz de mejorar la asignación de recursos, aumentar la eficiencia operativa y fortalecer la gestión estratégica del canal físico. De esta manera, el estudio confirma que la integración de analítica predictiva en la banca de personas permite avanzar hacia un modelo de gestión comercial más rentable, basado en datos y alineado con las mejores prácticas internacionales en analítica aplicada al sector financiero.

5.2 RECOMENDACIONES

A partir de los hallazgos obtenidos y del desempeño del modelo implementado, se formulan las siguientes recomendaciones estratégicas, operativas y técnicas, orientadas a maximizar el impacto del modelo predictivo y fortalecer la gestión comercial de la institución.

1. Implementar el sistema de priorización comercial basado en Machine Learning como herramienta oficial de gestión de campañas en canales físicos.

   Dado que el sistema de priorización comercial, sustentado en el modelo Random Forest, alcanzó un desempeño superior al umbral establecido (AUC = 0.866 en validación; 0.853 en datos reales) y demostró capacidad para capturar el 100 % de las ventas reales contactando únicamente el 53 % de la base, se recomienda:

   - Integrar formalmente el sistema de priorización en los procesos de asignación de campañas comerciales.

- Incorporarlo al CRM o a las plataformas operativas para generar rankings automáticos de clientes por nivel de propensión.
- Utilizarlo como criterio principal de priorización cuando la capacidad de contacto sea inferior al tamaño total de la base asignada.

Impacto esperado:

Mayor eficiencia operativa, reducción de costos comerciales y disminución de pérdidas de oportunidad asociadas a la imposibilidad histórica de cubrir la totalidad de la base asignada.

2. Fortalecer los procesos de contacto comercial como componente crítico del sistema de priorización.

   Los resultados del análisis estadístico y del modelado predictivo evidenciaron que la variable RESULTADO_CONTACTO constituye el predictor más influyente en la probabilidad de aceptación de productos financieros. Esto indica que la efectividad del sistema de priorización comercial no depende únicamente del ranking generado, sino también de la calidad de la interacción ejecutada por los equipos comerciales.

   Se recomienda:

   - Estandarizar guiones y protocolos de contacto para los segmentos de mayor propensión.
   - Capacitar a los ejecutivos en técnicas de interacción efectiva y manejo de objeciones.
   - Monitorear sistemáticamente la calidad del contacto y los resultados de cada interacción, retroalimentando el sistema con información operativa actualizada.

   Racional:

   Dado que el sistema de priorización se apoya fuertemente en variables operativas, la mejora en los procesos de contacto incrementará directamente la tasa de aceptación y potenciará el impacto del sistema en términos de eficiencia y rentabilidad comercial.

3. Utilizar los segmentos de propensión como eje central del sistema de priorización comercial.

   Los resultados del modelo evidencian que los segmentos superiores del ranking de propensión concentran una proporción significativa de las ventas reales, lo que confirma que la segmentación por niveles de probabilidad constituye el componente operativo clave del sistema de priorización comercial.

   En particular se observó que los primeros deciles del ranking concentran gran parte de las ventas reales top 5%: ~20% de las ventas, top 10%: ~38%–40% y top 20%: 70%.

   Se recomienda:

   - Definir al top 10 % como segmento de atención prioritaria obligatoria.
   - Establecer estrategias diferenciadas de contacto según nivel de propensión (alta, media y baja).
   - Implementar esquemas de asignación progresiva de recursos comerciales, priorizando los segmentos con mayor retorno esperado.

   Beneficio:

   Este enfoque permite maximizar la tasa de aceptación y la eficiencia comercial con un volumen reducido de contactos, transformando el proceso tradicional de barrido indiscriminado en un sistema estructurado de priorización basado en valor esperado.

4. Mantener recalibración trimestral del modelo.

   Dado que el sistema de priorización comercial se basa en patrones históricos de comportamiento del cliente, su desempeño puede verse afectado por cambios en las condiciones operativas, en las campañas comerciales o en las preferencias de los clientes a lo largo del tiempo. Este fenómeno, conocido como data drift o concept drift, es común en modelos de propensión aplicados en contextos reales:

   - Implementar un proceso de recalibración trimestral del modelo, utilizando datos recientes de desempeño comercial.

- Monitorear de forma sistemática métricas clave del sistema, tales como AUC, Lift@k y tasa de aceptación por segmento, para detectar pérdidas de capacidad discriminativa.
- Documentar cada versión del modelo y sus resultados, asegurando trazabilidad y control del riesgo de modelo.

Justificación:

La recalibración periódica permite mantener la vigencia del sistema de priorización, asegurar su alineación con la realidad operativa y mitigar el riesgo de decisiones subóptimas derivadas de modelos obsoletos, en coherencia con las buenas prácticas de gestión de modelos analíticos en el sector financiero.

5. Ampliar progresivamente las fuentes de datos para fortalecer el sistema de priorización.

   Si bien el sistema de priorización desarrollado demostró un desempeño sólido utilizando variables operativas y estructurales disponibles, los resultados sugieren que existe margen para mejorar la capacidad discriminativa del modelo, especialmente en los segmentos intermedios del ranking de propensión:

   Por ello, se recomienda:

   - Incorporar variables comportamentales longitudinales, tales como frecuencia de interacción, recencia de contacto y evolución del uso de productos.
   - Registrar de manera más estructurada los resultados cualitativos del contacto comercial, incluyendo causas de rechazo, nivel de interés o objeciones frecuentes.
   - Integrar indicadores de engagement del cliente con otros canales (por ejemplo, uso de banca digital o respuestas a campañas previas.

   Justificación:

   La literatura en analítica predictiva aplicada a banca señala que la incorporación de variables anticipables y de comportamiento mejora significativamente la capacidad de segmentación y la estabilidad de los sistemas de priorización. Este fortalecimiento

permitiría capturar patrones más finos de propensión, optimizando la asignación de esfuerzos comerciales y aumentando la rentabilidad del canal físico.

6. Escalar el sistema de priorización a otros productos y canales de la banca de personas.

   Dado el desempeño observado del sistema de priorización comercial en los canales físicos, se recomienda evaluar su extensión progresiva a otros productos financieros y canales de interacción dentro de la banca de personas, aprovechando la arquitectura analítica ya desarrollada, por lo cual se sugiere:

   - Aplicar el sistema de priorización a productos con baja penetración o alto margen, donde la asignación eficiente del esfuerzo comercial sea crítica.
   - Evaluar su uso en canales digitales complementarios (correo electrónico, SMS, banca en línea), ajustando los umbrales de propensión según la naturaleza del canal.
   - Diseñar estrategias híbridas, donde el ranking de propensión determine tanto el canal como la intensidad del contacto comercial.

   Justificación:

   La literatura y la práctica bancaria indican que los sistemas de priorización basados en propensión generan mayor valor cuando se integran de forma transversal a la estrategia comercial. La reutilización del sistema en distintos contextos permite maximizar el retorno de la inversión analítica, fortalecer la consistencia en la toma de decisiones y avanzar hacia una gestión omnicanal basada en datos.

En síntesis, las conclusiones y recomendaciones aquí presentadas demuestran que la incorporación de analítica predictiva mediante un sistema de priorización comercial basado en Machine Learning no solo es metodológicamente sólida, sino estratégicamente pertinente para la institución. La implementación del sistema propuesto en este estudio representa una oportunidad concreta para mejorar la eficiencia del canal físico, optimizar la asignación de recursos comerciales y maximizar el aprovechamiento de la capacidad operativa disponible, avanzando de manera progresiva hacia una cultura organizacional sustentada en el uso sistemático de datos y analítica avanzada para la toma de decisiones.

CAPÍTULO VI. APLICABILIDAD

El presente capítulo desarrolla la propuesta de aplicación práctica derivada de los hallazgos obtenidos en el estudio. Tras evidenciar que la aceptación de productos financieros en canales físicos está fuertemente determinada por variables operativas y que el modelo Random Forest permite captar el 100% de las ventas reales contactando únicamente el 53% de la base, se vuelve necesario transformar este conocimiento en una solución implementable dentro de la institución bancaria.

El Capítulo VI tiene como propósito presentar una propuesta técnica y operativa que permita convertir el modelo predictivo en un mecanismo formal de priorización comercial. Para ello, se detalla el nombre de la propuesta, su justificación, alcance, descripción metodológica, plan de implementación, indicadores de control y la concordancia entre los distintos segmentos de la investigación y la solución diseñada.

Esta propuesta busca cerrar la brecha entre el diagnóstico estadístico y la acción institucional, facilitando la adopción de un sistema basado en datos que incremente la eficiencia comercial, reduzca costos operativos y mejore la capacidad del banco para asignar esfuerzos de contacto hacia los clientes con mayor probabilidad de aceptación.

6.1 NOMBRE DE LA PROPUESTA

PriorizaBank: Sistema Predictivo de Priorización Comercial para Canales Físicos

6.2 JUSTIFICACIÓN DE LA PROPUESTA

La propuesta PriorizaBank: Sistema Predictivo de Priorización Comercial para Canales Físicos se justifica en la necesidad institucional de optimizar el proceso de asignación y contacto de clientes, dado que los resultados del análisis realizado evidencian ineficiencias estructurales en el modelo operativo actual. Diversos estudios en gestión bancaria señalan que la ausencia de herramientas analíticas para priorización comercial conduce a un uso ineficiente de los recursos, incrementa los costos operativos y limita la efectividad de los equipos de ventas (Deloitte, 2024; McKinsey & Company, 2023).

El estudio confirmó que la aceptación de productos financieros constituye un evento de baja frecuencia (2.73 %), pero con un alto grado de predictibilidad cuando se incorporan variables

operativas y de interacción relevantes. El modelo Random Forest, entrenado y validado en el Capítulo IV, demostró capacidad para identificar anticipadamente a los clientes con mayor probabilidad de aceptación, capturando el 100 % de las ventas reales al contactar únicamente el 53 % de la base. Este resultado se alinea con la literatura sobre predictive analytics, que destaca que la priorización basada en propensión permite maximizar resultados comerciales sin incrementar el volumen de contactos (Provost & Fawcett, 2013).

Históricamente, los canales físicos han logrado barrer solo entre el 60 % y 70 % de la base asignada debido a restricciones operativas, generando pérdida de oportunidades comerciales y elevando el costo por venta. El análisis inferencial evidenció que las variables RESULTADO_CONTACTO y PRODUCTO constituyen los principales determinantes de aceptación, mientras que variables demográficas tradicionales presentan baja capacidad discriminativa. Este hallazgo coincide con estudios que señalan que, en entornos bancarios maduros, las variables conductuales superan a las demográficas como predictores de conversión (Hair et al., 2019; James et al., 2022).

A partir de estos resultados, la propuesta se fundamenta en tres ejes críticos:

Eficiencia operativa:

La priorización predictiva permite reducir el volumen de clientes a contactar sin sacrificar ventas, lo que se traduce en menor carga operativa para los ejecutivos, tiempos de barrido más cortos y una reducción directa de los costos asociados al contacto comercial. La literatura en gestión de operaciones financieras destaca que este tipo de optimización mejora la productividad individual y colectiva de los equipos comerciales (Robson & McCartan, 2017).

Impacto comercial medible:

La concentración de ventas reales en los segmentos superiores del ranking (20 % en el top 5 %, 40 % en el top 10 % y 70 % en el top 20 %) evidencia que la priorización basada en machine learning incrementa significativamente la efectividad comercial. Estudios previos demuestran que los modelos de ranking por propensión permiten capturar mayor valor con menor esfuerzo operativo, especialmente en contextos de recursos limitados (Attota, 2024; Provost & Fawcett, 2013).

Madurez analítica y gestión del cambio:

La implementación de PriorizaBank representa un avance concreto hacia un modelo de gestión comercial basado en datos. La literatura sobre transformación digital y gestión del cambio organizacional señala que la adopción de sistemas analíticos requiere soluciones que generen beneficios rápidos, sean comprensibles para los usuarios y se integren a los procesos existentes, facilitando su aceptación por parte de los equipos operativos (Kotler & Keller, 2006; McKinsey & Company, 2023). En este sentido, la propuesta es escalable, replicable y adaptable a distintos productos y contextos operativos, lo que favorece su sostenibilidad en el tiempo.

Finalmente, la evidencia empírica presentada en los Capítulos IV y V confirma que la propuesta no solo es técnicamente viable, sino institucionalmente necesaria para mejorar la eficiencia del canal físico, aumentar la productividad comercial y fortalecer la capacidad analítica de la organización.

En consecuencia, PriorizaBank constituye una solución directamente derivada de los hallazgos del estudio y responde de manera integral a las necesidades detectadas, alineándose con las mejores prácticas internacionales en analítica bancaria y gestión del cambio.

6.3 ALCANCE DE LA PROPUESTA

El alcance de la propuesta establece los límites, objetivos y entregables específicos de la intervención diseñada en el Capítulo VI. A diferencia de los objetivos de investigación formulados en el Capítulo I, esta sección define los objetivos operativos propios de la implementación, orientados a la aplicación práctica del sistema de priorización en el canal físico.

Los objetivos se presentan utilizando el enfoque SMART (Específicos, Medibles, Alcanzables, Relevantes y Temporales), garantizando claridad en su ejecución y permitiendo evaluar el impacto real de la propuesta en la operación institucional.

Objetivo General de la Propuesta

Implementar un sistema institucional de priorización comercial basado en Machine Learning (S), que optimice la asignación y gestión de clientes del canal físico (S), medido mediante KPIs como Recallk, tasa de contacto y conversión por segmento (M), utilizando la infraestructura tecnológica y el personal disponible en la institución (A), con el fin de incrementar la eficiencia y efectividad comercial (R), durante el primer trimestre posterior a la aprobación del proyecto (2026) (T).

Objetivos Específicos de la Propuesta

1. Integrar el modelo predictivo al flujo institucional de campañas (S), generando un ranking mensual de propensión (M), con los recursos tecnológicos actuales (A), para orientar la selección de clientes de alto potencial (R), a partir del mes siguiente a la aprobación (2026) (T).
2. Clasificar a la totalidad de los clientes en segmentos CRITICAL, HIGH, MEDIUM y LOW (S), garantizando que el 100 % de los registros mensuales cargados en el CRM reciban su etiqueta correspondiente (M), mediante reglas automáticas definidas en el pipeline y en los workflows del CRM (A), con el fin de orientar la priorización comercial de los ejecutivos (R), dentro del primer mes de implementación del sistema (2026) (T).
3. Monitorear mensualmente los indicadores operativos del modelo (S), mediante dashboards existentes (M), utilizando los procesos analíticos actuales (A), con el fin de evaluar el desempeño y activar la recalibración trimestral (R), durante el periodo 2026–2027 (T).

Delimitación del Alcance Operativo

La propuesta PriorizaBank abarca el diseño, implementación y evaluación operativa de un sistema de priorización basado en Machine Learning para optimizar la gestión comercial del canal físico. La intervención integra el modelo predictivo en el proceso institucional de asignación de campañas, incorporando un esquema visible de segmentación en el CRM (CRITICAL, HIGH, MEDIUM y LOW), lo que modifica la forma en que los ejecutivos interpretan la calidad de los leads asignados. El alcance incluye la coordinación con el área de Capacitación y Desarrollo, la socialización del nuevo flujo operativo y la preparación del personal comercial para su adopción.

Asimismo, la propuesta comprende la generación mensual del ranking de propensión, el etiquetado operativo en el CRM, los procesos de comunicación y entrenamiento, la actualización periódica de variables anticipables, el monitoreo de indicadores como Recallk, tasa de contacto y conversión por segmento, y la recalibración trimestral del modelo como parte del ciclo de mejora continua.

Aspectos Fuera del Alcance

El alcance no incluye modificaciones al core bancario, a la infraestructura crítica ni al funcionamiento de los sistemas centrales. Tampoco contempla cambios sustanciales en el discurso comercial, salvo recomendaciones derivadas del uso del modelo. De igual forma, la propuesta no abarca su aplicación en canales digitales ni la predicción de montos de venta o ingresos, dado que se limita exclusivamente a la estimación de probabilidades de aceptación.

Límites Operativos y Técnicos

En cuanto a los límites operativos y técnicos, la efectividad del sistema depende de la calidad y disponibilidad de las variables anticipables definidas en el estudio. La propuesta es aplicable únicamente a productos que cuentan con historial suficiente para el modelado, requiere el uso de un dataset actualizado para cada ciclo comercial y debe entenderse como una herramienta de apoyo que complementa, pero no sustituye, las decisiones comerciales del personal ejecutivo.

Alcance Temporal

Finalmente, el alcance temporal contempla una fase de implementación inicial durante el primer trimestre posterior a la aprobación de la propuesta, seguida de una evaluación operativa en los primeros tres meses de ejecución en campañas reales y de un proceso de recalibración trimestral que se llevará a cabo durante el periodo 2026–2027.

6.4 DESCRIPCIÓN Y DESARROLLO

DESCRIPCIÓN

La propuesta PriorizaBank consiste en la implementación de un sistema institucional de priorización comercial basado en Machine Learning que genera un ranking de clientes ordenados según su probabilidad de aceptación de un producto financiero. Su propósito es optimizar el uso de la capacidad operativa del canal físico, permitiendo que los esfuerzos comerciales se enfoquen sistemáticamente en los clientes con mayor retorno esperado.

El sistema utiliza como núcleo el modelo Random Forest calibrado y validado en esta investigación, el cual estima probabilidades de aceptación empleando únicamente variables anticipables disponibles antes del contacto. A partir de estas probabilidades, el sistema clasifica a los clientes en cuatro segmentos operativos que serán visibles para los ejecutivos dentro del CRM institucional.

Este nuevo esquema de priorización implica un ajuste operativo respecto al proceso actual. En lugar de recibir una base homogénea, los ejecutivos accederán a leads con un etiquetado explícito de calidad, con instrucciones de priorización que optimizan el tiempo dedicado a la contactación. Debido a este cambio, la implementación incluye un proceso de comunicación y capacitación con el personal comercial para socializar el funcionamiento del ranking, interpretar correctamente las etiquetas operativas y aplicar los SLAs de contacto definidos para cada segmento.

La propuesta mantiene la estructura operativa existente del canal físico, pero introduce un mecanismo de asignación inteligente que permite:

- Maximizar la probabilidad de conversión por contacto.
- Disminuir el tiempo dedicado a leads de baja probabilidad.
- Mejorar la eficiencia del barrido comercial mensual.
- Estandarizar el proceso de selección y priorización de clientes.

En resumen, PriorizaBank transforma el proceso de asignación comercial en un flujo basado en analítica predictiva, manteniendo la simplicidad operativa para los ejecutivos, pero incorporando un componente estratégico de priorización que fortalece la capacidad productiva del canal físico.

Con el fin de representar de manera visual el funcionamiento integral del sistema propuesto, se presenta a continuación un diagrama de procesos que describe la integración del modelo predictivo PriorizaBank con las plataformas institucionales existentes. El diagrama ilustra el flujo end-to-end desde el entorno de desarrollo y validación del modelo, su implementación operativa mediante procesos de integración y APIs, hasta su utilización en las plataformas comerciales y de gestión del banco, incorporando además un esquema de monitoreo y retroalimentación para su ajuste continuo.

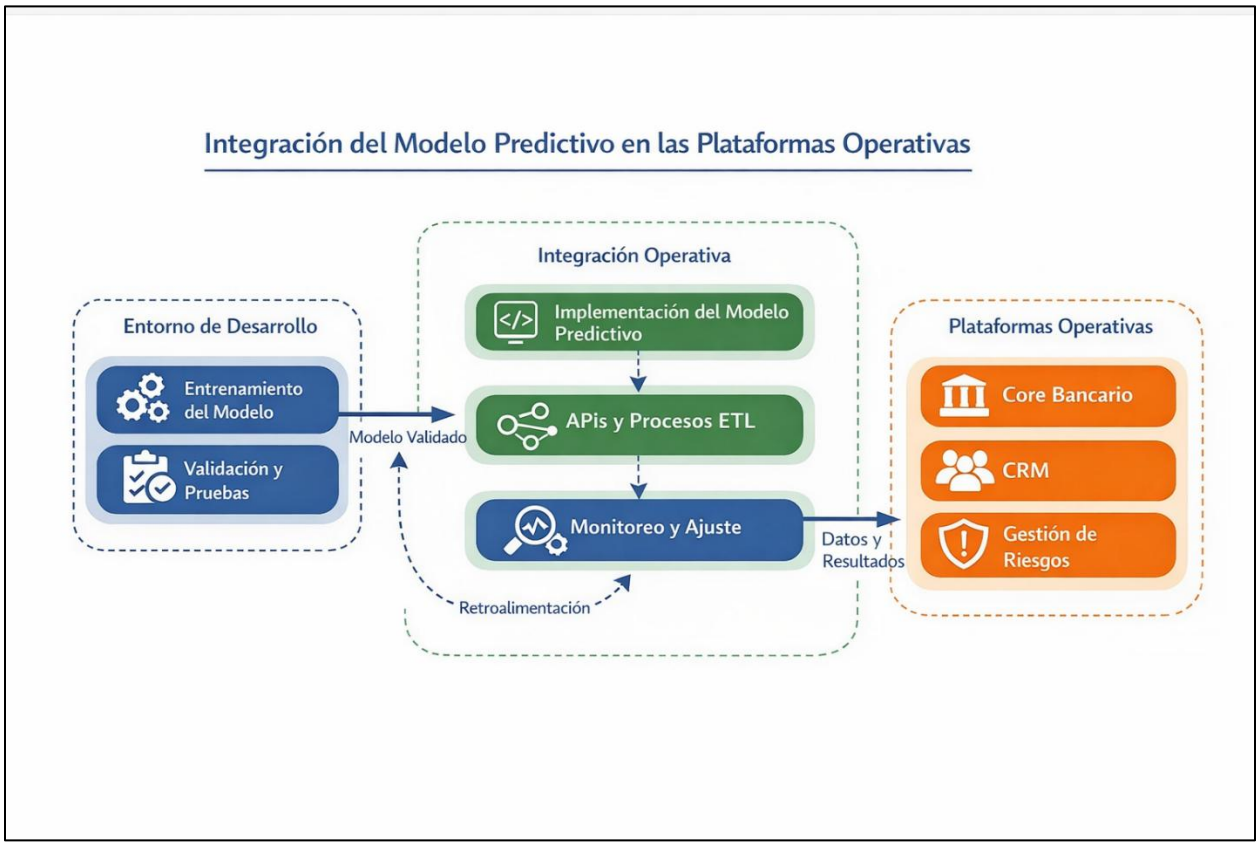

Ilustración 45 Diagrama de procesos e integración del modelo predictivo PriorizaBank con plataformas operativas

Fuente*:* Elaboración propia

DESARROLLO

La implementación de PriorizaBank se articula mediante un conjunto de módulos técnicos y operativos que permiten ejecutar el sistema de priorización de manera estandarizada, replicable y plenamente integrada al flujo comercial del canal físico. Cada módulo describe los insumos, procesos, herramientas y resultados necesarios para garantizar una adopción efectiva y sin interrupciones en la operación institucional.

La primera etapa del proceso consiste en garantizar que la base mensual de clientes cumpla con la estructura de variables anticipables definida en el estudio. Este módulo constituye el punto de partida técnico del sistema, ya que asegura que la información enviada al modelo predictivo sea homogénea, completa y metodológicamente consistente.

A. Módulo 1. Preparación de Datos Anticipables

El procedimiento inicia con la estandarización de los campos de la base, ajustando nombres, tipos de datos y formatos para asegurar compatibilidad con el pipeline original. Posteriormente, se generan de manera automática variables derivadas relevantes para el modelo, como ANTIGÜEDAD_CLIENTE_MESES y NIVEL_VINCULACIÓN, las cuales fueron identificadas previamente durante el análisis empírico por su capacidad predictiva. Esta fase también contempla la eliminación de variables no anticipables, evitando riesgos de sesgo o fuga de información.

Una vez depurada la estructura, la codificación categórica se aplica mediante el mismo pipeline de entrenamiento (One-Hot Encoding), garantizando coherencia entre las fases de modelado y ejecución. Finalmente, un script automatizado valida la presencia de todas las columnas requeridas antes de producir el archivo definitivo.

El resultado es el archivo newdata_lista.csv, que contiene únicamente las variables anticipables en el formato exacto que requiere el modelo predictivo.

B. Módulo 2. Ejecución del Modelo Predictivo

En esta etapa, el sistema aplica el modelo Random Forest calibrado para estimar la probabilidad de aceptación de cada cliente. El módulo inicia con la carga del archivo previamente preparado (newdata_ready.csv), seguido del procesamiento mediante el pipeline almacenado (preprocessor.pkl), el cual garantiza que las transformaciones aplicadas en la fase de entrenamiento se reproduzcan con precisión en producción.

El modelo se carga junto con sus artefactos (modelo_final.pkl) y ejecuta la predicción mediante la función predict_proba, generando para cada cliente una probabilidad numérica de aceptación. El sistema incluye un mecanismo automático de verificación de columnas, lo que permite identificar cualquier discrepancia antes de iniciar el cálculo.

Como resultado, se genera el archivo newdata_with_proba.csv, que contiene para cada registro su probabilidad estimada. Este archivo sirve de insumo para el proceso de ranking y priorización operativa.

C. Módulo 3. Generación del Ranking de Propensión

Una vez obtenidas las probabilidades individuales, se procede a transformarlas en un

ranking operativo que permitirá priorizar la gestión comercial. En esta fase, la base se ordena de manera descendente según la probabilidad estimada, calculando para cada cliente su posición relativa (rank) y su percentil dentro de la distribución completa.

A partir de esta distribución se generan los segmentos operativos definidos: CRITICAL (Top 5 %), HIGH (Top 10 %), MEDIUM (Top 20 %) y LOW (Top 50–60 %), los cuales representan niveles crecientes de retorno esperado. Este proceso permite identificar el punto en el que se concentra la mayor proporción de ventas históricas mediante el uso de Recall acumulado, asegurando que el modelo no solo ordena, sino que también optimiza la captura de oportunidades comerciales.

El módulo concluye con la producción del archivo ranking_priorizado.csv, que contiene el ID del cliente, su probabilidad estimada, su posición en el ranking, el segmento asignado y la instrucción operativa correspondiente.

D. Módulo 4 – Integración Operativa y Monitoreo

El cuarto módulo permite que el ranking generado se incorpore plenamente al CRM institucional, habilitando su uso operativo en las campañas del canal físico. Para ello, se introduce al CRM el archivo ranking_priorizado.csv, el cual contiene la información necesaria para que el sistema asigne a cada ejecutivo los clientes según su nivel de prioridad.

Las reglas de asignación operativa se estructuran a partir de los SLAs definidos para cada segmento, los cuales determinan los tiempos de contacto, número de reintentos y acciones esperadas por parte de los ejecutivos. De esta manera, los segmentos CRITICAL, HIGH, MEDIUM y LOW se traducen en niveles de urgencia operativa:

Tabla 45 SLA por segmento

Segmento	Descripción	SLA de contacto	Reintentos	Acción
CRITICAL (Top 5%)	Máxima prioridad	8 horas	3	Contacto inmediato
HIGH (Top 10%)	Alta probabilidad	48 horas	2	Contacto prioritario
MEDIUM (Top 20%)	Probabilidad moderada	72 horas	2	Seguimiento programado

Segmento	Descripción	SLA de contacto	Reintentos	Acción
LOW (Top 50–60%)	Contacto por capacidad	Sin SLA	1	Cola operativa

Fuente: Elaboración propia

El modelo incorpora además un sistema de monitoreo continuo que evalúa indicadores como la tasa de contacto, la conversión por segmento, el Recall@k mensual, el tiempo promedio de contacto y el cumplimiento de los SLAs por ejecutivo. Esta supervisión permite identificar desviaciones y corregir oportunamente la estrategia operativa.

Como parte del ciclo de gobernanza del modelo, el sistema incluye un proceso de recalibración trimestral que contempla la validación de drift, ajustes de umbrales y reentrenamiento en caso necesario. Este mecanismo asegura que el desempeño del modelo se mantenga consistente a lo largo del tiempo.

A continuación, se presenta el flujograma general del proceso PriorizaBank, que resume el recorrido completo desde la preparación de datos hasta la asignación de segmentos y el monitoreo en el CRM:

Ilustración 46 Flujograma operativo de PriorizaBank (desde la carga de base hasta recalibración)

Fuente: Elaboración propia

Una vez definido el flujo operativo general del sistema PriorizaBank, es necesario precisar las responsabilidades involucradas en cada una de las actividades del proceso. La correcta implementación del modelo y su sostenibilidad dependen de una distribución clara de funciones entre las áreas participantes, de modo que cada fase (desde la preparación de datos hasta la asignación comercial y el monitoreo) cuente con un responsable primario y roles de apoyo.

En este sentido, la siguiente tabla detalla los actores institucionales encargados de ejecutar y supervisar cada etapa del flujo operativo. Esta asignación de responsabilidades garantiza trazabilidad, continuidad operativa y una adecuada coordinación entre las áreas técnicas, comerciales y de capacitación.

Tabla 46 Responsables

Actividad	Responsable	Área
Preparar base anticipable	Especialista Inteligencia C.	Canales físicos
Ejecutar modelo	Especialista Inteligencia C.	Canales físicos
Generar ranking	Especialista Inteligencia C.	Canales físicos
Cargar ranking en CRM	Oficial A. Inteligencia C.	Canales físicos
Monitoreo y recalibración	Oficial B. Inteligencia C.	Canales físicos

Fuente: Elaboración propia

E. Módulo 5. Comunicación, Capacitación y Socialización

El éxito de PriorizaBank depende no solo de su diseño técnico, sino de su incorporación efectiva en la operación diaria del canal físico. Por ello, este módulo establece las acciones necesarias para garantizar que los ejecutivos comprendan el significado de las etiquetas operativas y ajusten su trabajo al nuevo esquema de priorización.

El proceso inicia con una comunicación institucional emitida por el área de Capacitación, donde se informa sobre la implementación del sistema, los segmentos CRITICAL, HIGH, MEDIUM y LOW, las instrucciones operativas asociadas y los SLAs de contacto.

Posteriormente, se realiza una sesión de socialización con ejecutivos y supervisores, en la cual se presentan los fundamentos del modelo, se muestra el ranking dentro del CRM y se explican las reglas de asignación y priorización. Esta sesión tiene una duración aproximada de 45 minutos y permite resolver dudas operativas antes de iniciar las campañas.

El módulo incluye además una fase de capacitación técnica, que consiste en la entrega de

un manual de usuario, una guía de acciones por segmento y ejemplos de buenas prácticas comerciales basadas en analítica predictiva.

Finalmente, durante los primeros treinta días de operación, el equipo analítico brinda acompañamiento directo mediante monitoreo diario de SLAs, retroalimentación a supervisores y ajustes operativos según desempeño.

6.5 MEDIDAS DE CONTROL

Para garantizar la eficacia operativa y la sostenibilidad del sistema PriorizaBank, la institución implementará un conjunto de indicadores de control que permitirá evaluar el desempeño del modelo predictivo y la adherencia del equipo comercial al nuevo esquema de priorización. Esto responde a las recomendaciones internacionales sobre gobernanza y gestión basada en datos, las cuales destacan la importancia del monitoreo continuo para asegurar la alineación entre estrategia, procesos y resultados institucionales (Comisión Económica para América Latina y el Caribe, 2023; OECD, 2022).

A diferencia de los controles metodológicos aplicados durante la investigación, estas medidas están orientadas a supervisar el funcionamiento real del sistema en su contexto operativo. Cada indicador contará con una ficha técnica que incluirá definición, fórmula de cálculo, fuente de datos, frecuencia, responsable y límites de control, siguiendo principios de gobernanza institucional y supervisión de procesos planteados por organismos como ((Comisión Nacional de Bancos y Seguros, 2022; Finnovista et al., 2024).

El primer indicador es el Recall Operativo @k, cuyo propósito es evaluar la capacidad del modelo para concentrar las ventas reales dentro de los segmentos de mayor prioridad. Su medición se realizará mensualmente a partir del registro de ventas del CRM institucional. Estudios recientes destacan que la medición sistemática del desempeño es clave para sostener procesos de transformación digital y mejorar la toma de decisiones basada en datos (Comisión Económica para América Latina y el Caribe, 2023).

Un segundo indicador corresponde a la Tasa de Conversión Real versus Esperada, que compara los resultados observados con las probabilidades estimadas por el modelo. Esta supervisión es fundamental para garantizar la coherencia del sistema en escenarios reales y coincide con los lineamientos de gobernanza analítica propuestos por (Banco Interamericano de

Desarrollo, 2025)

El desempeño del equipo comercial se evaluará mediante el Porcentaje de Adhesión al SLA, indicador que mide el cumplimiento de los tiempos de contacto definidos para cada segmento. La supervisión de la adherencia operativa es consistente con los marcos regulatorios de gestión de procesos y continuidad operativa establecidos por la (Comisión Nacional de Bancos y Seguros, 2022), los cuales enfatizan la necesidad de garantizar trazabilidad, disciplina de ejecución y mecanismos de alerta temprana.

Otro indicador relevante será la Cobertura Operativa del Ranking, que mide qué porcentaje de los leads asignados fue efectivamente gestionado por los ejecutivos. Este indicador permite evaluar la capacidad operativa del canal físico y se alinea con el uso de métricas aplicadas en el monitoreo comercial del sistema financiero hondureño.

La estabilidad técnica del modelo se supervisará mediante dos indicadores adicionales. El primero es el Drift de Variables Anticipables, orientado a detectar cambios significativos en la distribución de los datos utilizados por el modelo. La literatura sobre gobernanza de datos destaca la importancia de identificar desviaciones que puedan comprometer la utilidad analítica y operativa de los sistemas predictivos (OECD, 2022). El segundo indicador es la Estabilidad de la Calibración, evaluado mensualmente mediante valores como el Brier Score, con el objetivo de verificar que las probabilidades generadas continúan siendo coherentes con los resultados observados.

Fichas Técnicas De Indicadores

1. Recall Operativo @k

El Recall Operativo @k permite evaluar qué proporción de las ventas efectivas se concentró dentro de los segmentos superiores del ranking generado por el modelo. Su cálculo compara el número de ventas obtenidas dentro del Top k con el total de ventas registradas en el período. La fórmula utilizada es:

$$Recall@k = Ventas\ en\ el\ Top\ k / Ventas\ Totales$$

La información se alimenta directamente del CRM institucional y se actualiza mensualmente bajo la supervisión del Especialista de Inteligencia Comercial.

Para facilitar su interpretación operativa, se establecen límites de control:

- Verde: ≥ 70 %
- Amarillo: entre 50 % y 69 %
- Rojo: < 50 %

2. Tasa de Conversión Real vs Esperada

Este indicador contrasta el desempeño observado durante la campaña con el comportamiento que el modelo pronosticó. Su propósito es identificar desviaciones significativas entre la probabilidad estimada y la conversión obtenida. Se calcula mediante:

$$Desviación = Conversión\ Real - Conversión\ Esperada$$

Los datos provienen del CRM y del archivo mensual de probabilidades del modelo. El cálculo es responsabilidad del área de Inteligencia Comercial y se realiza al cierre de cada ciclo de campaña.

Los rangos de referencia se estructuran así:

- Verde: desviación ≤ ±2 %
- Amarillo: desviación entre ±2.1 % y ±5 %
- Rojo: desviación > ±5 %

3. Adhesión al SLA por Segmento

La Adhesión al SLA refleja la disciplina operativa de los ejecutivos al cumplir los tiempos de contacto definidos para cada segmento de prioridad (CRITICAL, HIGH, MEDIUM y LOW). Este indicador compara los contactos realizados dentro del SLA con el total de contactos asignados para dicho período:

$$Adhesión\ al\ SLA = (Contactos\ en\ tiempo / Contactos\ asignados) \times 100$$

El CRM registra automáticamente los tiempos de contacto, y los Coordinadores de Canales Físicos consolidan los datos semanalmente.

Los límites de control quedan establecidos así:

- Verde: ≥ 85 %
- Amarillo: 70 % – 84 %

- Rojo: < 70 %

4. Cobertura Operativa del Ranking

La Cobertura Operativa del Ranking permite evaluar qué parte de los clientes asignados a los ejecutivos fue efectivamente gestionada durante la campaña. Este indicador refleja la capacidad operativa del canal físico y ayuda a identificar brechas entre la asignación comercial y la ejecución real. Para su cálculo se compara el número de leads gestionados con el total de leads asignados:

$$\text{Cobertura} = \frac{\text{Leads gestionados}}{\text{Leads asignados}} \times 100$$

Los datos provienen directamente del CRM institucional y se consolidan mensualmente. El responsable de su análisis es el Coordinador de Operaciones Comerciales, quien debe activar alertas cuando la cobertura cae por debajo de los niveles establecidos.

Los límites de control definidos son:

- Verde: ≥ 75 %
- Amarillo: 60 % – 74 %
- Rojo: < 60 %

5. Drift de Variables Anticipables

El indicador de Drift detecta cambios significativos en la distribución de las variables anticipables respecto al dataset de referencia. Estos cambios pueden afectar la estabilidad del modelo, por lo que su supervisión es esencial. La detección se realiza mediante pruebas estadísticas como Kolmogorov–Smirnov (KS):

$$KS\ Drift = max\ |\ F1(x) - F2(x)\ |$$

El monitoreo se realiza mensualmente por el Especialista de Modelos Predictivos.

Se establecen los siguientes umbrales:

- Verde: KS < 0.10
- Amarillo: 0.10 – 0.19
- Rojo: ≥ 0.20

6. Estabilidad de la Calibración (Brier Score)

Este indicador evalúa si las probabilidades generadas por el modelo siguen correspondiendo al comportamiento real observado. Para ello se utiliza el Brier Score, una métrica que mide la diferencia promedio entre las probabilidades estimadas y los resultados reales:

$$Brier\ Score = 1/N\sum(p\hat{}\_i - y\_i\ )\hat{}2$$

El análisis se realiza mensualmente y está bajo responsabilidad del Analista de Modelos Predictivos.

Los límites de referencia son:

- Verde: ≤ 0.25
- Amarillo: 0.26 – 0.30
- Rojo: > 0.30

En conjunto, estos indicadores conforman un sistema integral de monitoreo que permitirá al banco supervisar el comportamiento del modelo en su entorno real de uso. Cuando alguno de los valores abandone los rangos establecidos, se activarán las medidas correctivas definidas en el ciclo de gobernanza institucional. Con ello, PriorizaBank mantendrá un desempeño estable, transparente y orientado a resultados, fortaleciendo la toma de decisiones en el canal físico.

6.6 CRONOGRAMA DE IMPLEMENTACIÓN Y PRESUPUESTO

La implementación de PriorizaBank se estructura en un conjunto de fases operativas cuyo propósito es garantizar la correcta adopción técnica, analítica y operativa del modelo predictivo dentro de la institución. La estimación temporal del proyecto se llevó a cabo empleando el método PERT, el cual permite modelar la duración de actividades mediante tres escenarios: optimista (O), más probable (M) y pesimista (P). Este enfoque es recomendado por el Project Management Institute en el PMBOK, al considerar la estimación por tres valores como una técnica adecuada cuando las tareas presentan incertidumbre técnica y operacional (PMI, 2021).

Siguiendo las directrices de Kerzner, quien sostiene que los intervalos O–M–P deben reflejar la complejidad y el nivel de riesgo de cada actividad, y de Kendall y Rollins, que enfatizan la importancia de basar el valor más probable (M) en evidencia histórica y desempeño institucional, los valores utilizados en cada fase fueron definidos por juicio experto, análisis de

actividades comparables y revisión de casos previos de implementación tecnológica en la organización (Kerzner, 2017) (Kendall & Rollins, 2003).

En consecuencia, cada fase del proyecto adoptó valores O, M y P coherentes con su naturaleza técnica y operativa. En la Fase 1 (Preparación de datos) se estableció un escenario optimista de 5 días, un escenario más probable de 8 días y un escenario pesimista de 12 días, atendiendo a la variabilidad usual en procesos de estandarización y validación de datos, cuyo comportamiento suele ser estable, pero susceptible a retrasos por ajustes inesperados.

En la Fase 2 (Integración del modelo y pruebas técnicas) los valores fueron definidos como 7 días (O), 12 días (M) y 18 días (P), considerando que la integración tecnológica presenta mayor complejidad y puede verse afectada por incompatibilidades de entorno o reconfiguraciones necesarias del pipeline, tal como advierte Kerzner para actividades de ingeniería y software (Kerzner, 2017).

Por su parte, la Fase 3 (Piloto operativo e integración con CRM) se caracterizó por una mayor variabilidad debido a la participación de múltiples áreas y la interacción con usuarios finales. Por ello, se estableció 10 días (O), 14 días (M) y 20 días (P), en línea con lo planteado por Shtub, Bard y Globerson respecto a las actividades operativas con dependencia humana y riesgo moderado (Shtub et al., 2005).

Finalmente, en la Fase 4 (Ajustes, documentación y capacitación) se emplearon valores de 6 días (O), 9 días (M) y 13 días (P), conforme a los riesgos identificados en etapas de documentación, versionamiento y entrenamiento, que suelen implicar iteraciones adicionales. Estos intervalos se definieron siguiendo la lógica estadística del método propuesto por Malcolm, donde el valor pesimista incorpora retrasos potenciales razonables y el optimista refleja el escenario sin interrupciones (Malcolm et al., 1959).

6.6.1 MÉTODO DE ESTIMACIÓN PERT

Los cálculos se basan en las siguientes fórmulas:

Duración esperada:

$$PERT = \frac{\mathrm{O} + 4\mathrm{M} + \mathrm{P}}{6}$$

Incertidumbre operativa:

$$Incertidumbre = P - O$$

Desviación estándar:

$$\sigma = \frac{P - O}{6}$$

Varianza:

$$\sigma^2 = \left(\frac{P - O}{6}\right)^2$$

Estas métricas permiten construir un cronograma sólido y cuantitativamente justificado la cual usaremos en esta sección y en la sección 6.7.

6.6.2 DESCRIPCIÓN NARRATIVA DE LAS FASES DEL CRONOGRAMA

El cronograma se estructura en cinco fases secuenciales que reflejan el flujo natural de una iniciativa de analítica avanzada orientada al desarrollo e implementación de modelos de Machine Learning.

La fase de planificación y preparación comprende la definición detallada del alcance operativo, la revisión de requerimientos técnicos, la coordinación con las áreas involucradas y la validación de la disponibilidad de los datos. Esta etapa es clave para asegurar condiciones iniciales adecuadas y minimizar riesgos posteriores.

La fase de análisis y preparación de los datos incluye actividades de exploración, limpieza, transformación y selección de variables. Dado que la calidad de los datos impacta directamente en el desempeño del modelo, esta fase concentra una parte relevante del tiempo total del proyecto y puede requerir iteraciones adicionales.

La fase de desarrollo y entrenamiento del modelo abarca la construcción, entrenamiento y comparación de distintos algoritmos predictivos, así como el ajuste de hiper parámetros y la

evaluación de desempeño. En esta etapa se integran los hallazgos del análisis inferencial y se selecciona el modelo con mejor desempeño.

Posteriormente, la fase de validación, pruebas e integración se orienta a evaluar el comportamiento del modelo en escenarios controlados, realizar pruebas piloto y preparar su integración al entorno operativo. Esta fase contempla ciclos de retroalimentación con las áreas usuarias.

Finalmente, la fase de documentación, capacitación y cierre incluye la elaboración de documentación técnica y operativa, la capacitación del personal involucrado y la validación final del cumplimiento de los objetivos de la propuesta, asegurando la sostenibilidad de la solución.

6.6.3 ESTIMACIÓN TEMPORAL DEL CRONOGRAMA MEDIANTE PERT

La Tabla 49 presenta la estimación de la duración de cada fase utilizando el método PERT, expresada en semanas, incorporando los escenarios optimista, más probable y pesimista.

Tabla 47 Estimación de Tiempo

Fase del proyecto	O (semanas)	M (semanas)	P (semanas)	Tiempo esperado PERT	Desviación estándar (σ)	Varianza (σ^2)
Planificación y preparación	1	2	3	**2**	0.33	0.11
Análisis y preparación de datos	2	3	5	**3.2**	0.5	0.25
Desarrollo y entrenamiento del modelo	3	4	6	**4.2**	0.5	0.25
Validación, pruebas e integración	2	3	5	**3.2**	0.5	0.25
Documentación, capacitación y cierre	1	2	3	**2**	0.33	0.11

Duración total estimada	**14.6 semanas**	2.16 semanas	0.97 semanas

Fuente: Elaboración propia

6.6.4 JUSTIFICACIÓN DE LA DURACIÓN POR FASES

La asignación temporal de cada fase del proyecto responde a la complejidad técnica de las actividades involucradas, a la dependencia secuencial entre etapas y a la experiencia documentada en proyectos de analítica avanzada y modelado predictivo.

La fase de planificación y preparación, estimada en dos semanas, se justifica por la necesidad de definir alcance, validar requerimientos funcionales y técnicos, coordinar con las áreas involucradas y verificar la disponibilidad y estructura de los datos. Aunque es una fase organizativa, su correcta ejecución reduce significativamente el riesgo de reprocesos posteriores.

La fase de análisis y preparación de datos, con una duración esperada de 3.2 semanas, concentra una porción relevante del cronograma debido a los procesos de exploración, limpieza, transformación y selección de variables. En proyectos de Machine Learning, esta etapa suele representar entre el 30% y 40% del esfuerzo total, dado que la calidad de los datos condiciona directamente el desempeño del modelo.

La fase de desarrollo y entrenamiento del modelo, estimada en 4.2 semanas, presenta la mayor duración esperada del proyecto. Esto se debe a la necesidad de entrenar múltiples algoritmos, ajustar hiperparámetros, realizar validaciones cruzadas, comparar métricas y ejecutar procesos de calibración probabilística, los cuales implican ciclos iterativos de mejora.

La fase de validación, pruebas e integración, con un tiempo esperado de 3.2 semanas, se fundamenta en la ejecución de pruebas controladas, validación con datos independientes y preparación de los mecanismos de integración con las plataformas institucionales, incorporando retroalimentación de usuarios clave.

Finalmente, la fase de documentación, capacitación y cierre, estimada en dos semanas, contempla la elaboración de manuales técnicos y operativos, la capacitación del personal involucrado y la validación final del cumplimiento de los objetivos del proyecto. En conjunto, estas estimaciones reflejan una distribución coherente del esfuerzo del proyecto, alineada con buenas prácticas de gestión de proyectos de ciencia de datos.

6.6.5 ANÁLISIS DE LA INCERTIDUMBRE TEMPORAL

El análisis PERT permite identificar que la mayor incertidumbre temporal se concentra en las fases de análisis de datos y desarrollo del modelo, debido a la posibilidad de iteraciones adicionales asociadas a la calidad de la información y al ajuste de los algoritmos. No obstante, la variabilidad observada se mantiene dentro de rangos aceptables para proyectos de analítica avanzada.

La duración total esperada del proyecto es de aproximadamente 15 semanas, lo cual se encuentra alineado con la planificación presupuestaria presentada en la Sección 6.7. La incorporación explícita de la incertidumbre temporal fortalece la gestión del proyecto, al permitir anticipar escenarios alternativos y facilitar la toma de decisiones ante posibles desviaciones del cronograma.

6.6.6 MEDIDAS DE MITIGACIÓN

La variabilidad identificada en las estimaciones PERT hace necesario incorporar mecanismos de mitigación que permitan mantener la ejecución del proyecto dentro de márgenes aceptables. En primer lugar, se considera fundamental integrar una reserva temporal equivalente a una desviación estándar del proyecto (aproximadamente tres días) con el propósito de absorber fluctuaciones naturales derivadas de actividades técnicas que, por su propia complejidad, pueden requerir retrabajos o ciclos adicionales de validación. Esta holgura permite que el cronograma conserve estabilidad aun cuando surjan contingencias menores. Priorizar tareas críticas de integración y validación.

Asimismo, se reconoce la importancia de asegurar que las actividades críticas, especialmente aquellas relacionadas con la integración del modelo y las pruebas técnicas, se ejecuten sin interrupciones y cuenten con la disponibilidad oportuna de los recursos institucionales involucrados. Estas actividades concentran la mayor parte de la incertidumbre temporal, por lo que su priorización constituye una medida clave para evitar retrasos acumulativos en fases posteriores. De manera complementaria, se plantea la necesidad de llevar a cabo un seguimiento sistemático del avance mediante revisiones semanales de los hitos del proyecto, lo cual permite identificar desviaciones tempranas y aplicar correcciones antes de que generen impactos significativos en la planificación general.

Finalmente, la coordinación anticipada con las áreas responsables (particularmente TI, Analítica y Comercial) se vuelve esencial para garantizar la disponibilidad del personal en los momentos decisivos del proyecto. Una comunicación previa y continua reduce el riesgo de que actividades críticas se vean afectadas por cargas laborales paralelas, reorganización de prioridades o limitaciones operativas.

En conjunto, estas acciones contribuyen a mitigar de manera efectiva el riesgo temporal, fortaleciendo la capacidad del proyecto para desarrollarse dentro de los plazos estimados y con un nivel adecuado de control y estabilidad.

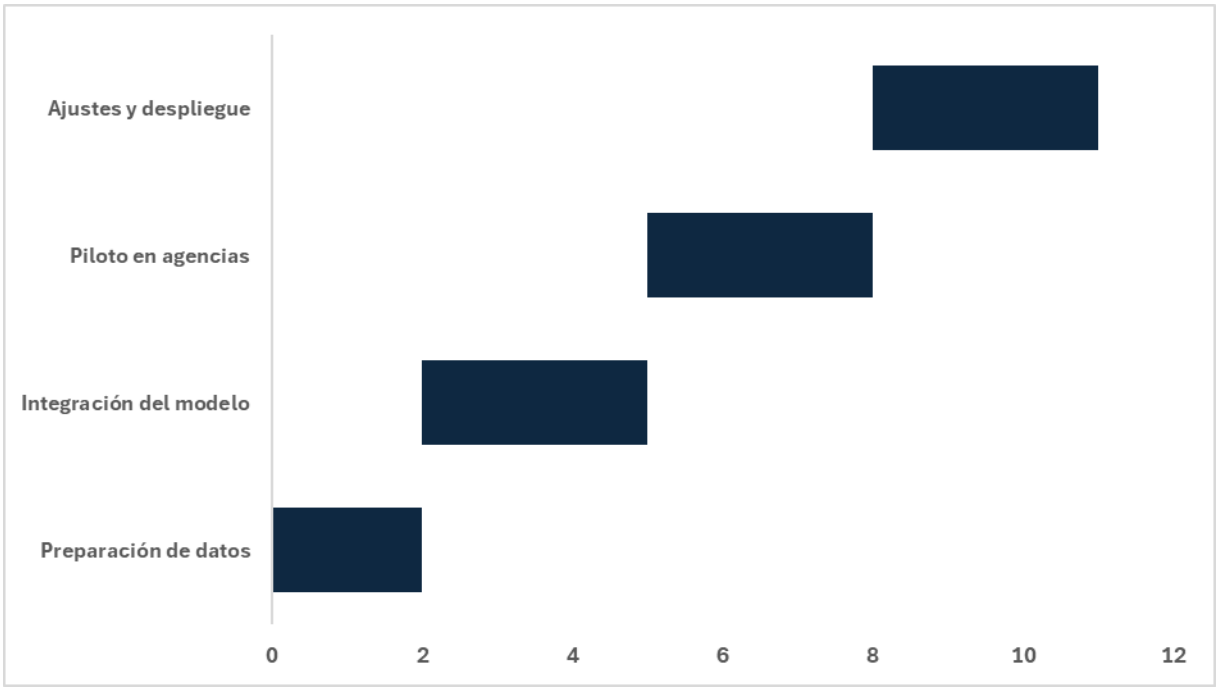

Ilustración *47* Tiempo proyectado
Fuente: Elaboración propia

6.7 PRESUPUESTO ESTIMADO

El modelo PriorizaBank no requiere adquisición de software especializado ni inversión en infraestructura adicional, ya que puede ejecutarse en entornos ya disponibles en el banco (Python, servidores internos). Al ser desarrollado dentro de la empresa no implicaría gastos operativos ni técnicos porque es parte de las labores. Pero estimaremos un presupuesto de acuerdo con los estándares de la implementación del modelo de Machine Learning donde requiere una planificación presupuestaria que contemple recursos humanos, tecnológicos, operativos y formativos.

Con el fin de asegurar rigor metodológico en la estimación de costos, se utilizó el método PERT, el cual permite incorporar la incertidumbre natural de proyectos tecnológicos mediante la construcción de tres escenarios: optimista, más probable y pesimista. Asimismo, se expresan los costos en dólares estadounidenses (USD), dado que gran parte de la infraestructura tecnológica y las licencias se cotizan en mercados internacionales.

6.7.1 METODOLOGÍA PARA LA ESTIMACIÓN DEL PRESUPUESTO

La estimación presupuestaria del proyecto se fundamenta en un enfoque analítico y estructurado, alineado con buenas prácticas de gestión de proyectos de analítica avanzada y ciencia de datos. Dado que la implementación de un modelo de Machine Learning involucra componentes técnicos, operativos y organizacionales, el presupuesto se construyó considerando cinco categorías principales: recursos humanos, infraestructura tecnológica, licencias y software, capacitación institucional e implementación operativa.

Para todas las categorías se utilizó el método PERT, el cual permite incorporar la incertidumbre inherente a proyectos tecnológicos mediante la definición de escenarios optimista (O), más probable (M) y pesimista (P), calculando posteriormente el costo esperado como:

$$Costo\ esperado\ (PERT) = \frac{\mathrm{O} + 4\mathrm{M} + \mathrm{P}}{6}$$

Asimismo, la incertidumbre del costo, desviación estándar y la varianza se calculó como se declaramos en la sección 6.6.1.

6.7.2 JUSTIFICACIÓN DE COSTOS

El costo de recursos humanos que representa la categoría central del presupuesto, ya que el desarrollo del modelo, su validación y su integración dependen directamente de la participación de perfiles especializados. Para el proyecto se consideraron tres roles clave: científico de datos, ingeniero/desarrollador y analista de negocio.

Con base en referencias salariales del mercado hondureño reportadas por plataformas internacionales, el salario promedio anual de un científico de datos se sitúa alrededor de USD 24,000–26,000, mientras que el de un analista de datos se aproxima a USD 21,000–23,000. Estos valores implican tarifas horarias promedio entre USD 11 y USD 13 por hora (Glassdoor, 2023; *Salary Data Analyst - Honduras*, 2025).

Considerando la complejidad del proyecto, se estimaron 120 horas para el científico de datos, 80 horas para el ingeniero/desarrollador encargado de integración y 40 horas para el analista de negocio, totalizando 240 horas de trabajo técnico.

El escenario más probable (M) se calculó multiplicando las horas estimadas por una tarifa horaria promedio. El escenario optimista (O) corresponde aproximadamente al 80% del valor más probable, asumiendo alta disponibilidad interna y ausencia de reprocesos. El escenario pesimista (P) representa un incremento cercano al 50%, asociado a posibles iteraciones adicionales, soporte especializado o ampliación de horas.

El costo en infraestructura incluye los costos asociados al uso de capacidad computacional, almacenamiento y recursos de servidor necesarios para el entrenamiento, validación y ejecución del modelo.

Aunque la institución dispone de infraestructura base, en proyectos de Machine Learning se acostumbra a asignar un costo equivalente por consumo de recursos, considerando uso de CPU, memoria, almacenamiento y entornos de prueba. En mercados internacionales, el costo mensual de instancias de cómputo con capacidad media-alta oscila entre USD 300 y USD 600 por servidor, dependiendo de configuración y proveedor.

Bajo este criterio, el escenario más probable considera el uso de dos a tres instancias durante aproximadamente tres meses, lo que sitúa el costo en torno a USD 3,600–5,000. El escenario optimista asume menor consumo y reutilización de recursos existentes, mientras que el pesimista contempla necesidad de capacidad adicional o mayor tiempo de uso.

El costo de licencia y software, si bien gran parte del desarrollo se realiza con herramientas de código abierto (Python, librerías de Machine Learning), en entornos empresariales suele requerirse el uso de licencias para herramientas de visualización, plataformas de orquestación, monitoreo de modelos o conectores especializados.

En el mercado, licencias empresariales de herramientas analíticas y de visualización presentan costos anuales que oscilan entre USD 1,400 y USD 3,000 por usuario, dependiendo del proveedor y funcionalidades.

El escenario más probable se estableció considerando una combinación mínima de licencias requeridas para explotación y monitoreo. El escenario optimista asume uso de

herramientas existentes en la institución, mientras que el pesimista contempla adquisición de licencias adicionales.

El costo de capacitación institucional La adopción sostenible del modelo requiere capacitar al personal que lo operará y dará seguimiento. Esta categoría incluye talleres internos, sesiones de transferencia de conocimiento y elaboración de material de apoyo.

En proyectos similares, los costos de capacitación especializada en analítica avanzada oscilan entre USD 600 y USD 1,600 por evento o curso corto. Con base en este rango, el escenario más probable considera al menos una jornada formal de capacitación, mientras que el optimista asume capacitación interna y el pesimista contempla contratación de un proveedor externo.

El costo de implementación operativa agrupa los costos asociados a pruebas piloto, ajustes de procesos, documentación operativa, soporte inicial y coordinación interdepartamental.

Aunque estos costos no siempre se reflejan de forma explícita, las buenas prácticas de gestión de proyectos recomiendan asignar entre un 10% y 20% del presupuesto técnico total a actividades de implementación y puesta en marcha.

Con base en esta referencia, el escenario más probable se calculó como un porcentaje del costo de recursos humanos, mientras que el escenario optimista considera una implementación fluida y el pesimista incorpora retrabajos operativos y mayor esfuerzo de soporte.

6.7.3 ESTIMACIÓN DE COSTOS MEDIANTE MÉTODO PERT (USD)

Tabla *48* Resumen de Costo

Categoría	O (USD)	M (USD)	P (USD)	CE= (O+4M+P)/6	Incertidumbre (P−O)	Desviación estándar (σ)	Varianza (σ^2)
Recursos humanos	2,093	2,616	3,924	**2,747**	1,831	**305.17**	**93,129**
Infraestructura tecnológica	3,600	4,800	6,400	**4,867**	2,800	**466.67**	**217,778**
Licencias y software	1,400	2,000	3,000	**2,067**	1,600	**266.67**	**71,111**
Capacitación institucional	600	1,000	1,600	**1,033**	1,000	**166.67**	**27,778**

Implementación operativa	1,600	2,200	3,200	**2,233**	1,600	**266.67**	**71,111**
TOTAL				**USD 12,947**	**8,831**		

Fuente: Elaboración propia

6.7.4 RETORNO DE LA INVERSIÓN (ROI) ESTIMADO

Con base en los resultados del análisis de datos del Capítulo IV, se establece que la tasa actual de aceptación en canales físicos es de aproximadamente 2.7%. En un escenario conservador, se espera un incremento de al menos 1.5% como resultado de la priorización inteligente habilitada por el modelo.

Supuestos para el cálculo del ROI:

- Clientes atendidos mensualmente: 14,000
- Ventas actuales (2.7%): $\frac{378}{\text{mes}}$
- Ventas adicionales estimadas (1.5%): $\frac{151}{\text{mes}}$
- Ingreso neto por producto: USD 12
- Horizonte de evaluación: 12 meses
- Costo total esperado (PERT): USD 12,947

Ventas e ingresos adicionales:

151×12=1,812 ventas adicionales por año

1,812×12 USD=21,744 USD de ingresos adicionales

Calculo ROI:

$$ROI = \frac{21{,}744 - 12{,}947}{12{,}947}\ x\ 100\ \approx 68\%$$

Interpretación

El proyecto continúa siendo altamente rentable con un ROI del 68% durante el primer año. Esto significa que, por cada dólar invertido, la institución genera USD 0.68 en retorno neto. Incluso

con una base de atención reducida, el modelo ofrece beneficios financieros superiores al costo de ejecución, validando la pertinencia de su implementación.

6.7.5 IMPACTO CUANTITATIVO DEL PRESUPUESTO

Reducción de costos operativos

El modelo permite priorizar clientes con mayor probabilidad de aceptación, reduciendo entre 20% y 30% el esfuerzo comercial dedicado a abordajes improductivos.

Incremento en eficiencia comercial

- Antes del modelo: $\frac{378}{\text{mes}}$
- Después del modelo: $\frac{378+210}{\text{mes}}$
- Incremento del 55% en la producción comercial

Ahorro operativo indirecto

La reducción de reprocesos y esfuerzos redundantes genera ahorros adicionales estimados entre USD 6,000 y USD 10,000 anuales.

6.7.6 IMPACTO CUALITATIVO DEL PRESUPUESTO

- Mejora en la calidad del servicio:

La personalización de ofertas basada en datos aumenta la relevancia de la interacción y mejora la percepción del cliente.

- Fortalecimiento de la reputación institucional

La adopción de modelos de Machine Learning posiciona al banco como referente regional en innovación aplicada a canales físicos.

- Cumplimiento normativo

El enfoque metodológico se alinea con principios internacionales de gobernanza de datos, transparencia algorítmica y supervisión financiera.

- Desarrollo de capacidades internas

El proyecto fomenta el fortalecimiento del capital humano en ciencia de datos y analítica avanzada, reduciendo dependencia de terceros.

Conclusión

El análisis presupuestario demuestra que la implementación del modelo de Machine Learning es técnica, financiera y estratégicamente viable. Con un costo total esperado de USD 12,947 y un ROI del 68%, el proyecto no solo genera beneficios económicos directos, sino que también contribuye a mejorar la eficiencia operativa, la experiencia del cliente y la capacidad analítica institucional. En consecuencia, su adopción constituye una inversión estratégica alineada con las necesidades actuales del sector bancario hondureño y con las tendencias globales en transformación digital.

6.8 CONCORDANCIA DE LOS SEGMENTOS DE LA TESIS CON LA PROPUESTA

La propuesta desarrollada en el Capítulo VI se fundamenta en una línea lógica y metodológica que recorre toda la tesis, desde la identificación del problema hasta los hallazgos empíricos y las conclusiones.

Este apartado demuestra la coherencia vertical del documento, evidenciando cómo cada capítulo aporta elementos esenciales que justifican, estructuran y luego permiten la implementación del sistema de priorización basado en Machine Learning propuesto para el canal físico de la entidad.

Con el fin de mostrar con claridad esta correspondencia, se presenta a continuación una matriz de concordancia, donde se vinculan los principales componentes de cada capítulo con los elementos que sustentan directamente la propuesta final:

Tabla 49 Concordancia segmentos

Descripción	Capítulo I	Capítulo II	Capítulo III	Capítulo V	Capítulo VI
Nombre: Implementación del Modelo Predictivo de Propensión para Optimizar la Aceptación de Productos en Canales Físicos	—	—	—	—	Propuesta: Implementación del sistema PriorizaBank La propuesta se basa en el modelo seleccionado y define su aplicación, priorización y monitoreo.
OG: Desarrollar un modelo de machine learning que estime la probabilidad de aceptación de productos financieros.	Planteamiento central del estudio orientado a construir un modelo predictivo útil para gestión comercial.	Se integran fundamentos de aprendizaje automático, comportamiento del consumidor y asimetría de información.	Metodología cuantitativa, diseño no experimental–transversal y registros históricos de interacciones en canales físicos. Se definen variables, estructura del dataset y técnicas de modelado para estimar probabilidad	Conclusión: el modelo predice aceptación con consistencia y aporta valor operativo.	La propuesta implementa el modelo en CRM, define procesos y KPIs para su monitoreo.
OE1: Identificar las variables relevantes que influyen en la aceptación.	Se formula en el planteamiento del problema y en los objetivos analíticos.	Teorías del comportamiento financiero y asimetría de información sustentan la influencia de variables perceptuales y operativas.	Variables operativas, transaccionales, demográficas y de interacción, EDA, pruebas estadísticas y depuración del dataset.	Aceptación depende principalmente de factores operativos y de interacción	Se define el Feature Set final utilizado en el modelo propuesto.

Descripción	Capítulo I	Capítulo II	Capítulo III	Capítulo V	Capítulo VI
OE2: Comparar modelos y seleccionar el de mejor desempeño.	Se destaca la necesidad de elegir el modelo adecuado.	El aprendizaje automático explica cómo los algoritmos capturan patrones y mejoran predicción.	Calibración, validación cruzada estratificada y métricas de evaluación.	Random Forest calibrado ofrece mejor desempeño y estabilidad.	Justificación técnica del modelo adoptado para operación.
OE3: Construir un ranking de propensión para priorizar clientes.	Se plantea como mecanismo para optimizar esfuerzos comerciales.	Se sustenta en patrones de comportamiento y segmentación derivados de las teorías revisadas.	Scoring para evalúa desempeño con Lift y Precisionk.	Ranking reduce esfuerzo y mejora eficiencia	El ranking se integra al CRM para priorización de campañas.
OE4: Evaluar el desempeño del modelo para garantizar su viabilidad.	Refuerza la necesidad de validación rigurosa.	Aprendizaje automático y asimetría de información justifican evaluación robusta.	Métricas como AUC, F1, matriz de confusión, curva de calibración, y validación cruzada estratificada para asegurar estabilidad del modelo.	El modelo es estable, calibrado y aplicable en contexto real.	Se establecen KPIs, monitoreo continuo y procesos de recalibración.

Fuente: Elaboración propia

La matriz de concordancia presentada permite evidenciar que la propuesta no constituye un componente aislado dentro del trabajo de investigación, sino el resultado lógico de un proceso estructurado que integra teoría, metodología y hallazgos empíricos. Cada capítulo aporta elementos que fortalecen y justifican la solución planteada, asegurando que esta responda directamente al problema identificado y a los objetivos formulados al inicio de la tesis.

Asimismo, la correspondencia entre los segmentos demuestra que el modelo predictivo desarrollado cumple con los criterios de validez académica y aplicabilidad práctica necesarios para su implementación en la institución financiera. La propuesta sintetiza el aprendizaje generado en cada fase del estudio y lo transforma en una herramienta estratégica capaz de mejorar la eficiencia

comercial y optimizar la toma de decisiones basada en datos.

En este sentido, la tabla no solo valida la coherencia interna del documento, sino que también evidencia la solidez conceptual y operativa de la propuesta, garantizando que su adopción se encuentre plenamente sustentada en el proceso investigativo que le dio origen.

REFERENCIAS BIBLIOGRÁFICAS

Akerlof, G. A. (1970). The Market for “Lemons”: Quality Uncertainty and the Market Mechanism*. *The Quarterly Journal of Economics*, *84*(3), 488-500. https://doi.org/10.2307/1879431

Alonso-Robisco, A., & Carbo, J. M. (2023). *Aprendizaje automático en modelos de concesión de crédito: Oportunidades y riesgos*.

Asociación Española de Banca. (2017, diciembre 27). Oportunidades y aplicaciones del big data en el sector financiero. *Asociación Española de Banca*. https://www.aebanca.es/noticias/oportunidades-y-aplicaciones-del-big-data-en-el-sector-financiero/

Attota, S. Y. D. C. (2024). *Optimizing Fintech Marketing: A Comparative Study of Logistic Regression and XGBoost* (arXiv:2412.16333). arXiv. https://doi.org/10.48550/arXiv.2412.16333

Banco Central de Costa Rica. (2023). *Memoria Anual 2023*. https://www.bccr.fi.cr/publicaciones/DocMemoriaAnual/Memoria_Anual_2023.pdf

Banco Central de Costa Rica. (2025). *Informe de Politica Monetaria*. https://www.bccr.fi.cr/publicaciones/DocPolticaMonetariaInflacin/Documento-IPM-Julio-2025.pdf

Banco Central de Reserva. (2023). *Memoria de Labores BCR 2023. Gobierno de El Salvador*. https://www.transparencia.gob.sv/documentos/73-8

Banco de España. (2024). *Informe de Estabilidad Financiera. Otoño 2024* (Informe de Estabilidad Financiera Nos. 24-2; Informe de Estabilidad Financiera, pp. 24-2). Banco de España. https://doi.org/10.53479/37957

Banco de Guatemala. (2023). *Memoria de Labores del Banco de Guatemala 2023*. https://banguat.gob.gt/sites/default/files/banguat/memoria/Memoria_Labores_2023.pdf

Banco Interamericano de Desarrollo. (2024). *Gobernanza para el desarrollo en América Latina y el Caribe*. https://www.undp.org/sites/g/files/zskgke326/files/2024-09/gobernanza_para_el_desarrollo_en_america_latina_y_el_caribe.pdf

Banco Interamericano de Desarrollo. (2025). *BID | América Latina y el Caribe atraviesa una profunda transformación con el uso de pagos digitales, destaca un estudio del BID*. https://www.iadb.org/es/noticias/america-latina-y-el-caribe-atraviesa-una-profunda-transformacion-con-el-uso-de-pagos-digitales

Banco Mundial. (2023). *Panorama general sobre inclusión financiera* [Text/HTML]. World Bank. https://www.worldbank.org/en/topic/financialinclusion/overview

Banco Mundial. (2024). *South Africa Economic Update, Edition 15*. https://www.wesgro.co.za/uploads/files/Edu-Invest/2025-Report-World-Bank-South-Africa-Economic-Update-15-Edition.pdf

Bank for International Settlements. (2017). *Basel III: Finalising post-crisis reforms*. https://www.bis.org/bcbs/publ/d424.pdf

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Comisión Económica para América Latina y el Caribe. (2023). *Panorama de la transformación digital en América Latina y el Caribe*. https://repositorio.cepal.org/handle/11362/48951

Comisión Nacional de Bancos y Seguros. (2022). *Normas para la gestión de tecnologías de información, ciberseguridad y continuidad del negocio*. https://circulares.cnbs.gob.hn/Archivo/Viewer/2520/1

Comisión Nacional de Bancos y Seguros. (2025). *Reporte de inclusión financiera 2025*.

Contreras, C. (2024). *El impacto de la inteligencia artificial en la industria financiera: Promesas y amenazas*.

Costanzo, L. (2023). *Data Cleaning During the Research Data Management Process*. https://doi.org/10.5206/IERZ1050

Creswell, J. W., & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. https://spada.uns.ac.id/pluginfile.php/510378/mod_resource/content/1/creswell.pdf

Cristina Victoria Chapa Zumba. (2023). *Segmentación de clientes (socios) para la recomendación de productos de colocación y/o captación para institución financiera en*

Ecuador.

Deloitte. (2024). *2024 banking and capital markets outlook*. Deloitte Insights. https://www.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-outlooks/banking-industry-outlook-2024.html

Diario Oficial La Gaceta. (2004). *Ley del sistema financiero nacional (Decreto No. 129-2004)*.

Domingos, P. (2016). The master algorithm: How the quest for the ultimate learning machine will remake our world. *Choice Reviews Online*, *53*(07), 53-3100-53-3100. https://doi.org/10.5860/CHOICE.194685

Espinosa Zúñiga, J. J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería Investigación y Tecnología*, *21*(1), 1-13. https://doi.org/10.22201/fi.25940732e.2020.21n1.008

Finnovista, Banco Interamericano de Desarrollo, & BID Invest. (2024). *Fintech en América Latina y el Caribe: Un ecosistema consolidado con potencial para aportar a la inclusión financiera regional*.

funcas. (2025). La inteligencia artificial en la banca europea: Adopción y casos de uso. *Funcas*. https://www.funcas.es/odf/la-inteligencia-artificial-en-la-banca-europea-adopcion-y-casos-de-uso/

Funcas. (2025). *La inteligencia artificial en la banca europea: Adopción y casos de uso*. https://www.funcas.es/odf/la-inteligencia-artificial-en-la-banca-europea-adopcion-y-casos-de-uso/

Gareth James,Sohil, F., Sohali, M. U., & Shabbir, J. (2022). An introduction to statistical learning with applications in R: By Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7. *Statistical Theory and Related Fields*, *6*(1), 87-87. https://doi.org/10.1080/24754269.2021.1980261

Glassdoor. (2023, diciembre 23). *Data Scientist Tegucigalpa, Honduras—Average Pay 2025*. Glassdoor. https://www.glassdoor.com/Salaries/tegucigalpa-honduras-data-scientist-salary-SRCH_IL.0,20_IM3077_KO21,35.htm

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://mitpress.mit.edu/9780262035613/deep-learning/

GSMA. (2023). *The Mobile Economy Sub-Saharan Africa 2023*.

Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (Eighth edition). Cengage. https://eli.johogo.com/Class/CCU/SEM/_Multivariate%20Data%20Analysis_Hair.pdf

Hernandez, J. M. G., & Moreno, W. N. T. (2022). *FACULTAD DE POSTGRADO TRABAJO FINAL DE GRADUACIÓN*.

Hernández, J. M. G., & Moreno, W. N. T. (2023a). *Predicción de riesgo de impago en institución financiera usando modelos de Machine Learning* [Universidad Tecnológica Centroamericana UNITEC]. https://repositorio.unitec.edu/xmlui/handle/123456789/12906

Hernández, J. M. G., & Moreno, W. N. T. (2023b). *Predicción de riesgo de impago en institución financiera usando modelos de Machine Learning* [Universidad Tecnológica Centroamericana UNITEC]. https://repositorio.unitec.edu/xmlui/handle/123456789/12906

Hernández Sampieri, R., & Fernandez-Collado, C. F. (2014). *Metodología de la investigación* (P. Baptista Lucio, Ed.; Sexta edición). McGraw-Hill Education. https://apiperiodico.jalisco.gob.mx/api/sites/periodicooficial.jalisco.gob.mx/files/metodologia_de_la_investigacion_-_roberto_hernandez_sampieri.pdf

IBM. (2021, octubre 4). *¿Qué es el análisis exploratorio de datos? | IBM*. https://www.ibm.com/mx-es/think/topics/exploratory-data-analysis

Instituto de Acceso a la Información Pública. (2022). *Anteproyecto de Ley de Protección de Datos Personales y Acción de Hábeas Data de Honduras*.

International Finance Corporation. (2024). *Emerging Market Green Bonds Report 2023: Green*

bonds issuance in emerging markets increased 34% in 2023 [Text/HTML]. IFC. https://www.ifc.org/en/pressroom/2024/emerging-market-green-bonds-report-2023-green-bonds-issuance-in-emerging-markets-increased-34-in-2023

International Organization for Standardization & International Electrotechnical Commission. (2022). *ISO/IEC 27001:2022*. ISO. https://www.iso.org/standard/27001

James, G., Witten, D., & Hastie, T. (2022). An introduction to statistical learning with applications in R: By Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media. *Statistical Theory and Related Fields*, *6*(1), 87-87. https://doi.org/10.1080/24754269.2021.1980261

Jeff Montoya, C. R., & Hernández Fúnez, M. J. (2023). *Modelo de perfilamiento de bancos comerciales en Honduras*. https://repositorio.unitec.edu/server/api/core/bitstreams/dd171fd0-5746-418d-bfb4-587d7dd4ff98/content

Kahneman, D., & Tversky, A. (1979). *Prospect Theory: An Analysis of Decision under Risk*. https://web.mit.edu/curhan/www/docs/Articles/15341_Readings/Behavioral_Decision_Theory/Kahneman_Tversky_1979_Prospect_theory.pdf

Kanaparthi, V. (2024). *AI-based Personalization and Trust in Digital Finance* (arXiv:2401.15700). arXiv. https://doi.org/10.48550/arXiv.2401.15700

Kendall, G. I., & Rollins, S. C. (2003). *Advanced Project Portfolio Management and the PMO: Multiplying ROI at Warp Speed*. J. Ross Publishing.

Kerzner, H. (2017). *Project Management Metrics, KPIs, and Dashboards: A Guide to Measuring and Monitoring Project Performance, Third Edition* (1.ª ed.). Wiley. https://doi.org/10.1002/9781119427599

Kotler, P., & Keller, K. L. (2006). *Marketing management* (12. ed). Pearson Prentice Hall. https://frrq.cvg.utn.edu.ar/pluginfile.php/14585/mod_resource/content/1/libro%20direccion-de-marketing%28kotler-keller_2006%29.pdf

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Malcolm, D. G., Roseboom, J. H., Clark, C. E., & Fazar, W. (1959). Application of a Technique for Research and Development Program Evaluation. *Operations Research*, *7*(5), 646-669. https://doi.org/10.1287/opre.7.5.646

McKinsey & Company. (2023). *The Great Banking Transition*.

Miñano Sanchez, C. J. (2022). *Comparación de técnicas de minería de datos para descubrir información relevante de ventas de una Mype comercial*.

Mitchell, T. M. (1997). *Machine Learning*. https://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com. https://scholar.google.com/scholar?cluster=14839322550984426 82&hl=en&oi=scholarr

Moodys Local Honduras. (2025). *Bancos | Moody's Local Honduras*. https://moodyslocal.com.hn/sectores/entidades-financieras/bancos/?utm_source=chatgpt.com

OECD. (2022). *Going Digital to Advance Data Governance for Growth and Well-being*. OECD Publishing. https://doi.org/10.1787/e3d783b0-en

Papageorgiou, G., Grant, S. W., Takkenberg, J. J. M., & Mokhles, M. M. (2018). Statistical primer: How to deal with missing data in scientific research?†. *Interactive CardioVascular and Thoracic Surgery*, *27*(2), 153-158. https://doi.org/10.1093/icvts/ivy102

PMI. (2021). *PMI (2021). The Standard for Project Management and a Guide to the Project Management Body of Knowledge (7th ed.). PMBOK Guide, Project Management Institute (PMI). - References—Scientific Research Publishing*. https://www.scirp.org/reference/referencespapers?referenceid=3520167

POPIA. (2021). *Protection of Personal Information Act (POPI Act)*. https://popia.co.za/

Porter, M. E. (2008). *Las cinco fuerzas competitivas que le dan forma a la estrategia*.

PricewaterhouseCoopers. (2020). *How mature is AI adoption in financial services?*

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51-59. https://doi.org/10.1089/big.2013.1508

Regulation (EU) 2016/679 of the European Parliament, 119 OJ L (2016). http://data.europa.eu/eli/reg/2016/679/oj

Reserve Bank of India. (2024). *REPORT ON TREND AND PROGRESS OF BANKING IN INDIA 2023-24*. https://mucbf.com/wp-content/uploads/Trend-and-Progress-Report/trend-and-progress-report-2023-24.pdf

Robson, C., & McCartan, K. (2017). *Real World Research, 4th Edition*.

Rodriguez Villanueva, A. A., Sánchez Adauto, E. L., & Valverde Rojo, L. M. (2023). *Aplicación de técnicas de Machine Learning para predecir el número de ventas de créditos en el sector bancario*. https://hdl.handle.net/20.500.12640/3378

Salary Data analyst—Honduras. (2025). Paylab - Salary Survey, Compare Salary, Salary Data. https://www.paylab.com/hn/salaryinfo/economy-finance-accountancy/data-analyst?fwd_lang=0

Sánchez, V. C. (2025). *Creación de un Modelo de Predicción de Ventas Basado en Modelos de Redes Neuronales Recurrentes.*

Särndal. (2003). *Särndal, C.E., Swensson, B. and Wretman, J. (2003) Model Assisted Survey Sampling. Springer Science & Business Media, Berlin, Heidelberg. - References—Scientific Research Publishing*. https://scirp.org/reference/referencespapers?referenceid=3128026

Shtub, A., Bard, J. F., & Globerson, S. (2005). *Project management: Processes, methodologies, and economics* (2nd ed.). Upper Saddle River, NJ : Pearson Prentice Hall.

Siarhei Sukhadolski. (2025, julio 7). Data analytics en banque | Impacts & usages. *Innowise*. https://innowise.com/es/blog/analisis-de-datos-en-la-banca/

South African Reserve Bank. (2023). *Second Edition 2023 Financial Stability Review*.

Súper Intendencia de Bancos de Guatemala. (2024). *Boletín trimestral de Indicadores de Inclusión Financiera*. https://www.sib.gob.gt/web/sib/informacion_sistema_financiero/estabilidad?p_p_id=110_INSTANCE_n1HH&p_p_action=0&p_p_state=maximized&p_p_mode=view&p_p_col_id=&p_p_col_pos=0&p_p_col_count=0&doAsUserId=xcfyftcg&_110_INSTANCE_n1

HH_struts_action=%2Fdocument_library_display%2Fview&_110_INSTANCE_n1HH_folderId=10292850

Superintendencia de Bancos de Panamá. (2024). *Informe de Situación del Centro Bancario Internacional Diciembre 2024*. https://www.superbancos.gob.pa/documentos/financiera_y_estadistica/estudios/IAB/IAB-1224.pdf?v=5.1

Superintendencia del Sistema Financiero. (2023). Boletín Enero—Marzo 2023. *Superintendencia del Sistema Financiero*. https://ssf.gob.sv/download/boletin-enero-marzo-2023/

Superintendencia General de Entidades Financieras. (2024). *Memoria Anual e Informe de Rendición de Cuentas 2024*.

Tanvir, M. F., Hossain, M. M., & Jishan, M. A. (2024). *Bayesian Regression for Predicting Subscription to Bank Term Deposits in Direct Marketing Campaigns* (arXiv:2410.21539). arXiv. https://doi.org/10.48550/arXiv.2410.21539

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*.

Useche, A., & Peter Stig Reina Colorado. (2021). *Fintech: Transformación del panorama bancario en América Latina*.

Vázquez, M. C., & González, J. P. (2019). *BIG DATA EN LA BANCA Y SUS IMPLICACIONES PARA EL FUTURO*.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press. https://books.google.com/books?hl=en&lr=&id=hSs3AgAAQBAJ&oi=fnd&pg=PP1&dq=info:T5fz2cmyyF8J:scholar.google.com&ots=VZRRryXWQs&sig=AK3sCtAGO8gj9XaH57XEl1QURlo

Zhang, Q., Yuan, X., Liu, T., Lam, C.-T., Huang, G., Lin, D., & Li, P. (2023). Tampering localization and self-recovery using block labeling and adaptive significance. *Expert Systems with Applications*, *226*, 120228. https://doi.org/10.1016/j.eswa.2023.120228

ANEXOS

1.1 ANEXO 1 DEFINICIÓN DE HIPER PARÉMETROS A EVALUAR

```
# -------------------------
# 4) definición de modelo e hiperparámetros
# -------------------------
models_to_run = {
    'Logistic': (LogisticRegression(max_iter=2000, class_weight='balanced', solver='liblinear'), {
        'clf__C': [0.01, 0.1, 1, 5]
    }),
    'DecisionTree': (DecisionTreeClassifier(class_weight='balanced', random_state=RANDOM_STATE), {
        'clf__max_depth': [3,5,8,12,None],
        'clf__min_samples_leaf':[5,20,50]
    }),
    'RandomForest': (RandomForestClassifier(class_weight='balanced', random_state=RANDOM_STATE, n_jobs=-1), {
        'clf__n_estimators':[200,400],
        'clf__max_depth':[6,8,12,None],
        'clf__min_samples_leaf':[5,10,20]
    }),
    'GradientBoosting': (GradientBoostingClassifier(random_state=RANDOM_STATE), {
        'clf__n_estimators':[200,300],
        'clf__learning_rate':[0.01,0.03,0.05],
        'clf__max_depth':[3,4,6]
    })
}

fitted_models = {}
probas = {}
```

Fuente: Elaboración propia con datos del modelo (Colab)

1.2 ANEXO 2 PIPELINE Y PRE PROCESAMIENTO (PARTE 1)

```
----- fig_pipeline.txt -----

best_RandomForest (params):
{'memory': None, 'steps': [('pre', ColumnTransformer(transformers=[('num',
                                 Pipeline(steps=[('imputer',
                                                  SimpleImputer(strategy='median')),
                                                 ('scaler', StandardScaler())]),
                                 ['EDAD', 'NUM_PASIVOS', 'SALDO_PASIVOS',
                                  'NUM_PRESTAMOS', 'FLAG_INTERBANCA',
                                  'FLAG_SARA', 'FLG_PENSION', 'FLAG_REMESA',
                                  'FLAG_TENGO', 'RIESGOACTUAL',
                                  'ANTIGUEDAD_CLIENTE_MESES',
                                  'NIVEL_VINCULACION']),
                                ('cat',
                                 Pipeline(steps=[('imputer',
                                                  SimpleImputer(fill_value='MISSING',
                                                                strategy='constant')),
                                                 ('ohe',
                                                  OneHotEncoder(handle_unknown='ignore',
                                                                sparse_output=False))]),
                                 ['PRODUCTO', 'RESULTADO_CONTACTO', 'CANAL',
                                  'CLASIFICACION_CLIENTE', 'GENERO',
                                  'ESTADO_CIVIL', 'EDUCATION_LEVEL',
                                  'DEPARTAMENTO'])])), ('clf', RandomForestClassifier(class_weight='balanced', max_depth=12,
                       min_samples_leaf=20, n_estimators=200, n_jobs=-1,
                       random_state=42))], 'transform_input': None, 'verbose': False, 'pre': ColumnTransformer(transformers=[('num',
                                 Pipeline(steps=[('imputer',
                                                  SimpleImputer(strategy='median')),
                                                 ('scaler', StandardScaler())]),
                                 ['EDAD', 'NUM_PASIVOS', 'SALDO_PASIVOS',
                                  'NUM_PRESTAMOS', 'FLAG_INTERBANCA',
                                  'FLAG_SARA', 'FLG_PENSION', 'FLAG_REMESA',
                                  'FLAG_TENGO', 'RIESGOACTUAL',
                                  'ANTIGUEDAD_CLIENTE_MESES',
                                  'NIVEL_VINCULACION']),
```

Fuente: Elaboración propia con datos del modelo (Colab)

1.3 ANEXO 3 PIPELINE Y PREPROCESAMIENTO (PARTE 2)

```
                                ('cat',
                                 Pipeline(steps=[('imputer',
                                                  SimpleImputer(fill_value='MISSING',
                                                                strategy='constant')),
                                                 ('ohe',
                                                  OneHotEncoder(handle_unknown='ignore',
                                                                sparse_output=False))]),
                                 ['PRODUCTO', 'RESULTADO_CONTACTO', 'CANAL',
                                  'CLASIFICACION_CLIENTE', 'GENERO',
                                  'ESTADO_CIVIL', 'EDUCATION_LEVEL',
                                  'DEPARTAMENTO'])]), 'clf': RandomForestClassifier(class_weight='balanced', max_depth=12,
                       min_samples_leaf=20, n_estimators=200, n_jobs=-1,
                       random_state=42), 'pre__force_int_remainder_cols': True, 'pre__n_jobs': None, 'pre__remainder': 'drop', 'pre__sparse_threshold
                ('scaler', StandardScaler())]), ['EDAD', 'NUM_PASIVOS', 'SALDO_PASIVOS', 'NUM_PRESTAMOS', 'FLAG_INTERBANCA', 'FLAG_SARA', 'FLG_PENSI
                 SimpleImputer(fill_value='MISSING', strategy='constant')),
                ('ohe',
                 OneHotEncoder(handle_unknown='ignore', sparse_output=False))]), ['PRODUCTO', 'RESULTADO_CONTACTO', 'CANAL', 'CLASIFICACION_CLIENTE',
                ('scaler', StandardScaler())]), 'pre__cat': Pipeline(steps=[('imputer',
                 SimpleImputer(fill_value='MISSING', strategy='constant')),
                ('ohe',
                 OneHotEncoder(handle_unknown='ignore', sparse_output=False))]), 'pre__num__memory': None, 'pre__num__steps': [('imputer', SimpleImp

Logistic -> CalibratedClassifierCV(cv=3,
                       estimator=Pipeline(steps=[('pre',
                                                  ColumnTransformer(transformers=[('num',
                                                                                   Pipeline(steps=[('imputer',
                                                                                                    SimpleImputer(strategy='median')),
                                                                                                   ('scaler',
                                                                                                    StandardScaler())]),
                                                                                   ['EDAD',
                                                                                    'NUM_PASIVOS',
                                                                                    'SALDO_PASIVOS',
```

Fuente: Elaboración propia con datos del modelo (Colab)

1.4 ANEXO 3 PIPELINE Y PREPROCESAMIENTO (PARTE 3)

```
DecisionTree -> CalibratedClassifierCV(cv=3,
                       estimator=Pipeline(steps=[('pre',
                                                  ColumnTransformer(transformers=[('num',
                                                                                   Pipeline(steps=[('imputer',
                                                                                                    SimpleImputer(strategy='median')),
                                                                                                   ('scaler',
                                                                                                    StandardScaler())]),
                                                                                   ['EDAD',
                                                                                    'NUM_PASIVOS',
                                                                                    'SALDO_PASIVOS',


RandomForest -> CalibratedClassifierCV(cv=3,
                       estimator=Pipeline(steps=[('pre',
                                                  ColumnTransformer(transformers=[('num',
                                                                                   Pipeline(steps=[('imputer',
                                                                                                    SimpleImputer(strategy='median')),
                                                                                                   ('scaler',
                                                                                                    StandardScaler())]),
                                                                                   ['EDAD',
                                                                                    'NUM_PASIVOS',
                                                                                    'SALDO_PASIVOS',


GradientBoosting -> CalibratedClassifierCV(cv=3,
                       estimator=Pipeline(steps=[('pre',
                                                  ColumnTransformer(transformers=[('num',
                                                                                   Pipeline(steps=[('imputer',
                                                                                                    SimpleImputer(strategy='median')),
                                                                                                   ('scaler',
                                                                                                    StandardScaler())]),
                                                                                   ['EDAD',
                                                                                    'NUM_PASIVOS',
                                                                                    'SALDO_PASIVOS',
```

Fuente: Elaboración propia con datos del modelo (Colab)

1.5 ANEXO 5 MEJORES PARÁMETROS

```
Loaded best_RandomForest: {'memory': None, 'steps': [('pre', ColumnTransformer(transformers=[('num',
                                  Pipeline(steps=[('imputer',
                                                   SimpleImputer(strategy='median')),
                                                  ('scaler', StandardScaler())]),
                                  ['EDAD', 'NUM_PASIVOS', 'SALDO_PASIVOS',
                                   'NUM_PRESTAMOS', 'FLAG_INTERBANCA',
                                   'FLAG_SARA', 'FLG_PENSION', 'FLAG_REMESA',
                                   'FLAG_TENGO', 'RIESGOACTUAL',
                                   'ANTIGUEDAD_CLIENTE_MESES',
                                   'NIVEL_VINCULACION']),
                                 ('cat',
                                  Pipeline(steps=[('imputer',
                                                   SimpleImputer(fill_value='MISSING',
                                                                 strategy='constant')),


Loaded best_GradientBoosting: {'memory': None, 'steps': [('pre', ColumnTransformer(transformers=[('n
                                  Pipeline(steps=[('imputer',
                                                   SimpleImputer(strategy='median')),
                                                  ('scaler', StandardScaler())]),
                                  ['EDAD', 'NUM_PASIVOS', 'SALDO_PASIVOS',
                                   'NUM_PRESTAMOS', 'FLAG_INTERBANCA',
                                   'FLAG_SARA', 'FLG_PENSION', 'FLAG_REMESA',
                                   'FLAG_TENGO', 'RIESGOACTUAL',
                                   'ANTIGUEDAD_CLIENTE_MESES',
                                   'NIVEL_VINCULACION']),
                                 ('cat',
                                  Pipeline(steps=[('imputer',
                                                   SimpleImputer(fill_value='MISSING',
                                                                 strategy='constant')),


Loaded best_Logistic: {'memory': None, 'steps': [('pre', ColumnTransformer(transformers=[('num',
                                  Pipeline(steps=[('imputer',
                                                   SimpleImputer(strategy='median')),
                                                  ('scaler', StandardScaler())]),
                                  ['EDAD', 'NUM_PASIVOS', 'SALDO_PASIVOS',
                                   'NUM_PRESTAMOS', 'FLAG_INTERBANCA',
                                   'FLAG_SARA', 'FLG_PENSION', 'FLAG_REMESA',
                                   'FLAG_TENGO', 'RIESGOACTUAL',
                                   'ANTIGUEDAD_CLIENTE_MESES',
                                   'NIVEL_VINCULACION']),
                                 ('cat',
                                  Pipeline(steps=[('imputer',
                                                   SimpleImputer(fill_value='MISSING',
                                                                 strategy='constant')),


Loaded best_DecisionTree: {'memory': None, 'steps': [('pre', ColumnTransformer(transformers=[('num',
                                  Pipeline(steps=[('imputer',
                                                   SimpleImputer(strategy='median')),
                                                  ('scaler', StandardScaler())]),
                                  ['EDAD', 'NUM_PASIVOS', 'SALDO_PASIVOS',
                                   'NUM_PRESTAMOS', 'FLAG_INTERBANCA',
                                   'FLAG_SARA', 'FLG_PENSION', 'FLAG_REMESA',
                                   'FLAG_TENGO', 'RIESGOACTUAL',
                                   'ANTIGUEDAD_CLIENTE_MESES',
                                   'NIVEL_VINCULACION']),
                                 ('cat',
                                  Pipeline(steps=[('imputer',
                                                   SimpleImputer(fill_value='MISSING',
                                                                 strategy='constant')),
```

Fuente: Elaboración propia con datos del modelo (Colab)

1.6 ANEXO 6 CALIBRACIÓN

```
----- fig_calibration.txt -----

CalibratedClassifierCV(cv=3,
                       estimator=Pipeline(steps=[('pre',
                                                  ColumnTransformer(transformers=[('num',
                                                                                   Pipeline(steps=[('imputer',
                                                                                                    SimpleImputer(strategy='median')),
                                                                                                   ('scaler',
                                                                                                    StandardScaler())]),
                                                                                   ['EDAD',
                                                                                    'NUM_PASIVOS',
                                                                                    'SALDO_PASIVOS',
                                                                                    'NUM_PRESTAMOS',
                                                                                    'FLAG_INTERBANCA',
                                                                                    'FLAG_SARA',
                                                                                    'FLG_PENSION',
                                                                                    'FLAG_REMESA',
                                                                                    'FLAG_TENGO',
                                                                                    'RIESGOACTUAL',
                                                                                    'ANTIGUEDAD_CLIENTE...
                                                                                                                    strategy='constant')),
                                                                                                   ('ohe',
                                                                                                    OneHotEncoder(handle_unknown='ignore',
                                                                                                                  sparse_output=False))]),
                                                                                   ['PRODUCTO',
                                                                                    'RESULTADO_CONTACTO',
                                                                                    'CANAL',
                                                                                    'CLASIFICACION_CLIENTE',
                                                                                    'GENERO',
                                                                                    'ESTADO_CIVIL',
                                                                                    'EDUCATION_LEVEL',
                                                                                    'DEPARTAMENTO'])])),
                                                 ('clf',
                                                  RandomForestClassifier(class_weight='balanced',
                                                                         max_depth=12,
                                                                         min_samples_leaf=20,
                                                                         n_estimators=200,
                                                                         n_jobs=-1,
                                                                         random_state=42))]))
```

Fuente: Elaboración propia con datos del modelo (Colab)

1.7 ANEXO 7 RECUPERACIÓN TOTAL DE VENTAS REALES BAJO RESTRICCIONES OPERATIVAS

```
# ----------------------
# Punto clave: mínimo % para captar 100% de las ventas reales
# ----------------------
if total_pos == 0:
    pct_needed = None
    rows_needed = None
    print("No hay ventas reales en el dataset (total_pos = 0).")
else:
    # acumulado por orden descendente de proba
    df_sorted["cum_pos"] = df_sorted["VENTA"].cumsum()
    # buscamos el primer índice donde cum_pos >= total_pos (debería ser cuando alcanza total_pos)
    mask = df_sorted["cum_pos"] >= total_pos
    if mask.any():
        first_idx = mask.idxmax()  # primer índice (0-based)
        rows_needed = first_idx + 1
        pct_needed = rows_needed / n
        print(f"\nPara capturar el 100% de ventas reales necesitas ordenar y contactar el TOP {pct_needed*100:.3f}% de la base (N = {rows_needed} observaciones).")
    else:
        pct_needed = None
        rows_needed = None
        print("\nNo fue posible alcanzar el acumulado de ventas reales (revisar datos).")

# Summary compacto y export
summary = {
    'n_obs': n,
    'total_pos': total_pos,
    'AUC_real': float(auc_real) if not np.isnan(auc_real) else None,
    'best_thr': BEST_THR,
    'precision_best_thr': float(prec),
    'recall_best_thr': float(rec),
    'f1_best_thr': float(f1),
    'base_rate': float(base_rate) if not np.isnan(base_rate) else None,
    'pct_needed_for_100pct_pos': float(pct_needed) if pct_needed is not None else None,
    'rows_needed_for_100pct_pos': int(rows_needed) if rows_needed is not None else None
}
for row in topk_rows:
    k_tag = int(row['k_pct']*100)
    summary[f'prec_top{k_tag}'] = float(row['prec'])
    summary[f'rec_top{k_tag}'] = float(row['rec'])
    summary[f'lift_top{k_tag}'] = float(row['lift'])

summary_df = pd.DataFrame([summary])
csv_out = f"{OUT_DIR}/evaluation_summary_minpct100.csv"
summary_df.to_csv(csv_out, index=False)
print(f"\nResumen compacto guardado en: {csv_out}")
```

```
Para capturar el 100% de ventas reales necesitas ordenar y contactar el TOP 53.247% de la base (N = 15974 observaciones).

Resumen compacto guardado en: /content/model_outputs_models/new_data_ready/evaluation_summary_minpct100.csv
```

Fuente*:* Elaboración propia con datos del modelo (Colab)