****

**ESCUELA DE NEGOCIOS**

**MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS**

**TÍTULO DE LA INVESTIGACIÓN**

**Aplicación de modelos de predicción para identificar patrones de deserción en los datos de clientes de una entidad financiera.**

**Profesor**

**Ing. Mario Salvador González PH.D**

**Autores**

**Ana Cristina Prado Martínez**

**Gladys Margarita Yambay Ramos**

**2024**

**RESUMEN**

El presente estudio tuvo como objetivo el desarrollar un modelo predictivo para clasificar a los clientes de la entidad financiera según su probabilidad de deserción con la finalidad de segmentar eficazmente a los clientes en función de su probabilidad de deserción. Para lo cual se realizó el análisis de los modelos: Regresión Logística, Arboles de Decisión y Random Forest que según la revisión literaria son los más recomendados para la clasificación y predicción de deserción de clientes.

El análisis mostró que el modelo Random Forest fue el más efectivo en predecir con un rendimiento general del modelo del 84% (AUC). Sin embargo, aunque logró una capacidad de clasificación del 72% (recall) para detectar clientes en riesgo de abandonar la entidad, la precisión fue sólo del 49%, indicando la necesidad de incluir más variables relevantes para mejorar su efectividad. El modelo permitió identificar que las variables como la edad, número de productos contratados, saldo, puntos ganados y salario son las principales variables que influyen en la deserción. En base a estas variables se plantean las estrategias que están orientadas a la personalización de productos y servicios.

Además, con el análisis del modelo Random Survival Forest se identifica a los clientes en riesgo de abandono y su tiempo de permanencia en la entidad, lo que ofrece una ventaja competitiva al anticipar con mayor precisión cuándo intervenir con estrategias de retención.

Entre las recomendaciones, se sugiere la inclusión de variables adicionales como número de reclamos, número de transacciones, ocupación, estado civil y canal de comunicación, esto permitirá obtener mejor resultados de predicción.

Finalmente, se identifica la importancia de actualizar continuamente el modelo de predicción para que éste se ajuste a los cambios en los comportamientos de los clientes y del mercado financiero. Estas acciones aumentarán las probabilidades de retener clientes de manera efectiva y eficiente, y que la entidad disminuya los porcentajes o tasa de deserción.

**ABSTRACT**

This study aimed to develop a predictive model to classify customers of the financial institution based on their churn, facilitating effective segmentation according to their risk levels. For this purpose, the following models were analyzed: Logistic Regression, Decision Trees, and Random Forest, which are identified in the literature as the most suitable for classifying and predicting customer churn.

The analysis indicated that the Random Forest model was the most effective, achieving an overall model performance of 84% (AUC). However, while it attained a classification capability of 72% (recall) for detecting customers at risk of leaving, its precision was only 49%, highlighting the need to incorporate additional relevant variables to enhance its effectiveness. The model identified variables such as age, number of products contracted, balance, points earned and salary as the most relevants. Strategies focused on personalizing products and services were proposed based on these variables.

Additionally, the Random Survival Forest model analysis identified customers at risk of abandonment and estimated the duration time of client, providing a competitive advantage by enabling more precise timing for retention interventions.

Among the recommendations, the inclusion of additional variables such as number of complaints, number of transactions, occupation, marital status, and transaction channels is suggested to improve predictive outcomes.

Finally, the importance of continuously updating the predictive model is emphasized to ensure alignment with changes in customer behavior and the financial market. These actions will increase the likelihood of effectively and efficiently retaining customers, thereby reducing churn rates for bank entity.

**ÍNDICE DEL CONTENIDO**

[1. INTRODUCCIÓN 1](#_Toc179918170)

[2. REVISIÓN DE LITERATURA 2](#_Toc179918171)

[3. IDENTIFICACIÓN DEL OBJETO DE ESTUDIO 7](#_Toc179918172)

[4. PLANTEAMIENTO DEL PROBLEMA 8](#_Toc179918173)

[5. OBJETIVO GENERAL 9](#_Toc179918176)

[6. OBJETIVOS ESPECÍFICOS 9](#_Toc179918177)

[7. JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA 10](#_Toc179918178)

[7.1 RECOLECCION DE DATOS 11](#_Toc179918193)

[7.2 LIMPIEZA, PRE-PROCESAMIENTO Y/O TRANSFORMACIÓN DE DATOS. 12](#_Toc179918194)

[7.3 IDENTIFICACIÓN Y DESCRIPCIÓN DE VARIABLES 16](#_Toc179918195)

[7.4 VISUALIZACIÓN DE VARIABLES 19](#_Toc179918196)

[7.5 SELECCIÓN DE MODELO ESTADÍSTICO 28](#_Toc179918197)

[8. RESULTADOS 33](#_Toc179918198)

[8.1 ANÁLISIS DEL MODELO ESTADÍSTICO 33](#_Toc179918208)

[8.2 INTERPRETACIÓN DE RESULTADOS 58](#_Toc179918209)

[8.3 COMPARACION DE METRICAS ENTRE MODELOS DE REGRESIÓN LOGÍSTICA, ARBOL DE DECISIÓN Y RANDOM FOREST CON BALANCEO E HIPERPARAMETROS. 59](#_Toc179918210)

[8.4 IMPORTANCIA DE LAS CARACTERISTICAS 62](#_Toc179918211)

[9. ESTRATEGIAS ORGANIZACIONALES 64](#_Toc179918212)

[9.1 ESTRATEGIAS 64](#_Toc179918214)

[9.2 INNOVACIÓN Y COMPETITIVIDAD EMPRESARIAL 66](#_Toc179918215)

[10. CONCLUSIONES Y RECOMENDACIONES 70](#_Toc179918216)

[10.1 CONCLUSIONES 70](#_Toc179918218)

[10.2 RECOMENDACIONES 71](#_Toc179918219)

[11. REFERENCIAS 73](#_Toc179918220)

[12. ANEXOS 77](#_Toc179918221)

**ÍNDICE DE TABLAS**

**Tabla 1**. Identificación de datos 11

**Tabla 2.** Datos nulos y duplicados 13

**Tabla 3.** Rangos Edad 14

**Tabla 4**. Transformaciones variables categóricas 15

**Tabla 5.** Variables del dataset 16

**Tabla 6.** Cuadro comparativo entre los Modelos de Clasificación 31

**Tabla 7.** Métricas Regresión Logística 37

**Tabla 8.** Métricas Árbol de Decisión 38

**Tabla 9.** Métricas Random Forest 40

**Tabla 10.** Métricas Regresión Logística Balanceado 42

**Tabla 11.** Métricas Árbol de Decisión Balanceado. 43

**Tabla 12.** Métricas Random Forest Balanceado 45

**Tabla 13.** Métricas Regresión Logística Balanceado e Hiperparámetros 46

**Tabla 14.** Métricas Árbol de Decisión Balanceado e Hiperparámetros 47

**Tabla 15**. Métricas Random Forest Balanceado e Hiperparámetros 49

**Tabla 16.** Comparación Técnicas Aplicadas a modelos 58

**Tabla 17.** Comparación Métricas de Modelos de Clasificación 59

**ÍNDICE DE FIGURAS**

**Figura 1**. Datos atípicos 14

**Figura 2**. Matriz de Correlación Inicial. 17

**Figura 3**. Matriz de Correlación Final. 18

**Figura 4**. Distribución de Deserción 19

**Figura 5.** Distribución de Deserción por Género 20

**Figura 6.** Distribución de Deserción por País 20

**Figura 7.** Distribución de Deserción por Tipo de Tarjeta 21

**Figura 8.** Distribución de Deserción por Antigüedad 22

**Figura 9.** Distribución de Deserción por Rangos de Edad 23

**Figura 10**. Distribución de Deserción por Número de Productos 23

**Figura 11.** Distribución de Deserción por Calificación de Satisfacción 24

**Figura 12.** Distribución de Deserción por Actividad del Cliente 25

**Figura 13**. Distribución de Deserción por Reclamos 25

**Figura 14**. Gráfico dispersión 27

**Figura 15.** Clasificación de variables 33

**Figura 16**. Diagrama de modelamiento 35

**Figura 17**. Matriz de Confusión de Regresión Logística 36

**Figura 18.** Curva ROC Regresión Logística 37

**Figura 19.** Matriz de Confusión Árbol de Decisión 38

**Figura 20.** Curva ROC Árbol de Decisión 39

**Figura 21**. Matriz de Confusión Random Forest 39

**Figura 22.** Curva ROC Random Forest 40

**Figura 23**. Matriz de Confusión Regresión Logística Balanceado 42

**Figura 24.** Curva ROC Regresión Logística Balanceado 42

**Figura 25**. Matriz de Confusión Árbol de Decisión Balanceado 43

**Figura 26**.Curva ROC Árbol de Decisión Balanceado 44

**Figura 27.** Matriz de Confusión Random Forest Balanceado 44

**Figura 28.** Curva ROC Random Forest Balanceado 45

**Figura 29**. Matriz de Confusión Regresión Logística Balanceado e Hiperparámetros 46

**Figura 30**. Curva ROC Regresión Logística Balanceado Hiperparámetros 46

**Figura 31.** Matriz de Confusión Árbol de Decisión Balanceado e Hiperparámetros 47

**Figura 32.** Curva ROC Árbol Decisión Balanceado e Hiperparámetros 48

**Figura 33.** Matriz de Confusión Random Forest Balanceado e Hiperparámetros 48

**Figura 34**. Curva ROC Random Forest Balanceado e Hiperparámetros 49

**Figura 35**. t-SNE de Clientes por Deserción 50

**Figura 36.** Score Modelo RSF 51

**Figura 37.** Curva de Supervivencia de Clientes 52

**Figura 38.** Curvas de Riesgo del Cliente 52

**Figura 39**. Comparación entre modelos 59

**Figura 40.** Curvas ROC comparativas 60

**Figura 41.** Importancia de las características 62

**Figura 42.** Estrategia de Innovación 66

# INTRODUCCIÓN

El sector bancario está evolucionando rápidamente, impulsado por los avances tecnológicos, las nuevas preferencias de los consumidores y la intensa competencia del mercado. Por lo cual las entidades financieras deben ofrecer una experiencia satisfactoria a fin de no correr el riesgo de perder clientes.

El análisis de deserción en una entidad financiera es fundamental porque permite identificar a los clientes que están en riesgo de abandonar a la entidad. Este tipo de análisis se enfoca en entender los factores que llevan a los clientes a dejar de utilizar los productos o servicios ofrecidos, como insatisfacción, competencia, o cambios en sus necesidades financieras.

Aplicando modelos predictivos, se pueden anticipar estos comportamientos de deserción antes de que ocurran. Esto le permite a la entidad tomar medidas proactivas para retener a los clientes, como mejorar la atención, ofrecer productos personalizados.

El reducir la tasa de deserción contribuye al crecimiento de la entidad al mantener una base de clientes sólida, lo que a su vez fortalece la rentabilidad y la competitividad en el mercado.

# REVISIÓN DE LITERATURA

El presente trabajo tiene como objetivo realizar el análisis de deserción o abandono de clientes en una entidad financiera mediante el estudio de la información de la entidad. Comprender el comportamiento de los clientes a partir de sus datos y transacciones, para anticipar su posible deserción y diseñar estrategias de retención.

A continuación, se realiza una revisión de varios casos de estudio sobre deserción de clientes en algunas empresas con diferente giro de negocio, pero cuya necesidad era predecir los patrones y tasas de deserción de sus clientes.

En una administradora de fondos ecuatoriana se identificó una gestión ineficiente en el manejo de información sobre clientes desertores, debido a la falta de análisis del comportamiento de abandono y la ausencia de modelos predictivos para segmentar clientes con precisión. Esto resultó en estrategias de retención y fidelización poco efectivas. Para abordar este problema, se analizaron 105,353 registros de clientes con 20 variables que incluían datos del cotizante y del asesor. Se aplicaron modelos de árboles de decisión, bosques aleatorios y regresión logística, concluyendo que el bosque aleatorio ofreció el mejor rendimiento con un 93% de precisión. Las variables más relevantes fueron el saldo de la cuenta, tipo de sistema de aportación y si el cliente trabaja de forma independiente. Se identificaron 30,577 afiliados con alta probabilidad de deserción, lo que permitirá a la empresa implementar estrategias más focalizadas para retenerlos. (Bohórquez, 2019)

Otro estudio fue el realizado en una entidad bancaria del Ecuador para la elaboración de un modelo predictivo de deserción de clientes, en el cual se desarrolló un modelo de scoring para la detección oportuna de la deserción de clientes de tarjetas de crédito. Los datos de insumo se basan en los indicadores económicos y facturación de tarjetas de crédito de 11 bancos. La metodología empleada permite seleccionar las variables clave para la predicción y determinar el umbral de default, que diferencia entre clientes desertores e inactivos y aquellos que permanecen fieles.(Tatamues, 2021)

En una empresa de créditos en línea, se usó algoritmos de Machine Learning para mejorar el servicio al cliente, mediante analítica de datos, empleando variables de riesgo de crédito, sociodemográficas, y datos históricos de más de 3 millones de entradas de información de contacto, notas de las Respuesta de Voz Interactiva (IVR) y datos de créditos. El área de servicio al cliente prioriza la eficiencia y precisión, evaluada por métricas como la satisfacción del cliente (NPS) y el costo por atención según el canal. Se probaron varios algoritmos de Machine Learning, destacando Random Forest con un Recall de 0.8475 y un desempeño económico del 83.1%. El análisis recomendó no enfocarse en retener clientes con 12 o más créditos, sino concentrar esfuerzos en aquellos con baja recurrencia, quienes representan más del 82% del total y ofrecen mayor potencial de retención. (Fajardo Cruz, 2019)

Predecir la tasa de abandono de clientes en modelos de suscripción utilizando diversos algoritmos de aprendizaje automático, fue el objetivo de otro estudio, el cual tenía como finalidad identificar los modelos más efectivos para ayudar a las empresas de suscripción a predecir y gestionar la pérdida de clientes. En este trabajo se analizó 21 características utilizando 9 algoritmos y, tras un proceso exhaustivo de preparación de datos, selección de variables, construcción y evaluación de modelos, se identificó tres algoritmos destacados: regresión logística, gradient boosting y redes neuronales. El algoritmo de mejor rendimiento es la regresión logística, con una precisión de predicción del 79.6%, lo que lo hace el más recomendable para predecir la tasa de abandono de clientes en modelos de suscripción.(Zhang, 2023)

El análisis de la duración del cliente fue objeto de estudio en una empresa de Telecomunicaciones, en la cual se requería predecir la duración o permanencia del cliente, para ello, evaluaron varias metodologías, destacando el Random Survival Forest por su mayor precisión en la reducción de errores en comparación con métodos tradicionales como Kaplan-Meier y la Regresión de Cox, aunque son fáciles de implementar, pueden ser menos precisos. Sin embargo, Random Survival Forest tiene desventajas como tiempos de procesamiento más largos y menor facilidad de interpretación. Se identificaron variables clave como la experiencia del cliente, intenciones de cancelación, servicios técnicos, desconexiones, facturación promedio, que son críticas para anticipar la deserción. Además, una de las sugerencias fue actualizar la información al menos cada seis meses para incorporar posibles cambios en el comportamiento del mercado que no fueron considerados en el análisis actual.(Moreno, 2023)

En una entidad financiera brasileña examinaron la efectividad de diferentes modelos estadísticos para predecir qué clientes tienden a abandonarla, utilizando un conjunto de datos representativos de 500.000 clientes para la exploración y análisis a fin de documentar nuevos conocimientos sobre los principales determinantes que predicen la futura pérdida de clientes, así como desarrollar estrategias efectivas es esencial para evitar la pérdida de clientes.

El modelo de bosques aleatorios produjo resultados superiores, ya que identificó el 80,2% de los clientes que abandonarían en los siguientes meses. Por otro lado, el 14,8% de los clientes que no abandonaron fueron clasificados como propensos a la deserción.(Lemos, 2022)

Deserción de clientes en el sector asegurador, en este análisis se utilizó machine learning en cuatro modelos distintos como metodología para predecir la probabilidad de deserción en clientes de una aseguradora ecuatoriana en el ramo vehículos. Los modelos usados fueron árboles de decisión, GLM binomial, random forest y xtreme gradient boosting. Como resultados se obtuvo que el modelo xtreme gradient boosting fue el que mejor rendimiento presentó en las métricas usadas para comparar los modelos que fueron precisión y área bajo la curva, con una precisión del 70,77% y además a través del GLM binomial se descubrieron las variables más influyentes en la renovación de los clientes**.** (Centeno Arízaga, 2020)

Deserción de clientes en bancos peruanos, el planteamiento del modelo consiste en que una entidad bancaria pueda fidelizar y retener a sus clientes potenciales. De esta manera, con la tecnología actual del mercado se propuso un Modelo de Análisis Predictivo para la retención de los clientes con tendencias a la deserción en entidades bancarias en el Perú, a través del comportamiento de los clientes en cuanto a sus transacciones, movimientos y actividad financiera. Alcanzando la mejor precisión (93.20%) con el algoritmo R-K-Means, teniendo también mejores resultados en los falsos/positivos y falsos/negativos en comparación con otros algoritmos**.** (Barrueta Meza, 2018)

Otro análisis de deserción fue el realizado en una entidad financiera en base a sus clientes tarjetahabientes, este proyecto analiza la probabilidad de que un cliente o tarjetahabiente cancele su tarjeta de crédito dentro de los 90 días posteriores a su adquisición. La base de datos fue recopilada a partir de conversaciones con técnicos de las áreas de riesgos y del análisis de clientes. Se aplicaron métodos estadísticos, incluyendo la regresión logística y el modelo de Random Forest, para evaluar la probabilidad de deserción. Los resultados del estudio muestran que el modelo de Random Forest proporciona una precisión de clasificación superior, alcanzando un 80.1%, en comparación con el 79% logrado por la regresión logística. (Pulgarín, 2020)

En base a estos casos de estudio sobre la deserción de clientes (*churn*) en diversas empresas, se puede destacar la importancia del análisis predictivo a través del uso de modelos predictivos, los cuales son fundamentales para anticipar el abandono de clientes. La utilización modelos como: árboles de decisión, random forest, regresión logística y otros algoritmos de aprendizaje automático permiten identificar patrones de deserción.

Se destaca el uso de random forest y regresión logística en varios estudios por su precisión para predecir la deserción. Estos modelos son preferidos por su capacidad de manejo de grandes volúmenes de datos y precisión, alcanzando los más altos porcentajes de precisión. Sin embargo, la selección del modelo dependerá de los tipos y características de los datos que se disponga para el análisis.

Los estudios también subrayan la importancia de la segmentación precisa para la retención de clientes. Y plantean como sugerencia, la actualización periódica de datos y modelos predictivos, ya que el comportamiento del cliente puede cambiar con el tiempo y es importante incorporar estas variaciones en las estrategias de retención.

# IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

El propósito del estudio es identificar patrones en los clientes que podrían desertar de la entidad financiera, con el fin de implementar estrategias proactivas de retención y mejorar la fidelización, reduciendo así la pérdida de ingresos y los costos asociados a la adquisición de nuevos clientes (Dirección de Marketing, 2012). Según los autores del libro “Leading on the Edge of Chaos” un aumento del 2% en la retención de clientes es equivalente a una reducción del 10% en los costos (Murphy, 2002).

Esto se logrará mediante el análisis de datos demográficos, financieros y de comportamiento, para detectar las características más relevantes que influyen en el cliente para que tome la decisión de desertar, características tales como: el saldo de cuentas, reclamos, número de productos que poseen en la entidad.

El objetivo es emplear métodos de agrupamiento para identificar grupos en los datos de clientes de la entidad financiera. Esto implica analizar patrones y relaciones en la información para encontrar conjuntos de datos que compartan características similares. Al clasificar los datos que comparten atributos comunes, se pueden obtener información valiosa sobre el comportamiento de los clientes y sus necesidades específicas. (Kuhn, 2018)

Este análisis facilitará una segmentación precisa de los clientes, enfocado a la creación de campañas personalizadas, ofreciendo incentivos específicos para mejorar el servicio a aquellos clientes con mayor riesgo de deserción y maximizando el retorno de la inversión en esfuerzos de retención, antes de que los clientes decidan desertar. Este objetivo fortalecerá la relación con el cliente, mejorará la experiencia del cliente y contribuirá a la estabilidad y crecimiento sostenible de la entidad financiera.

# PLANTEAMIENTO DEL PROBLEMA

Realizar un análisis de la situación actual de deserción de clientes en la entidad financiera del caso de estudio, es una necesidad fundamental para garantizar su estabilidad y crecimiento sostenible. La deserción de clientes no solo representa una pérdida directa de ingresos, sino que también aumenta los costos operativos, ya que adquirir nuevos clientes es significativamente más costoso que retener a los existentes. Identificar y comprender las razones detrás de la deserción permite a quienes toman decisiones en la entidad, anticiparse y mitigar estos riesgos.

Además, este análisis ayudará a segmentar la base de clientes, enfocando recursos y esfuerzos en aquellos segmentos con mayor riesgo de deserción. Esto no solo optimiza el retorno de la inversión en campañas de retención, sino que también mejora la satisfacción y fidelización del cliente, fortaleciendo la relación a largo plazo. Un análisis eficaz de la deserción contribuirá a la mejora continua de productos y servicios, incrementando la competitividad y la resiliencia de la entidad financiera en un mercado dinámico y exigente.



# OBJETIVO GENERAL

Desarrollar un modelo predictivo para clasificar a los clientes de la entidad financiera según su probabilidad de deserción, utilizando algoritmos de machine learning como son: los árboles de decisión, random forest y regresión logística, con la finalidad de segmentar eficazmente a los clientes en función de su riesgo de abandono, optimizando así las estrategias de fidelización y reduciendo costos asociados a las campañas de retención.

# OBJETIVOS ESPECÍFICOS

* Aplicar algoritmos de machine learning como son: los árboles de decisión, random forest y regresión logística, con el fin de segmentar a los clientes en función de su riesgo de deserción.
* Seleccionar el modelo que ofrezca los mejores resultados en términos de precisión predictiva en la deserción de los clientes.
* Identificar y seleccionar las variables más relevantes que afectan la deserción de clientes para desarrollar un modelo de clasificación que sea preciso y robusto. Este proceso incluirá la evaluación de la influencia de cada variable en el abandono y la validación de que el modelo resultante proporciona resultados estables y confiables.
* Analizar los resultados del modelo predictivo seleccionado para descubrir oportunidades de mejora a través de estrategias de retención.
* Proporcionar recomendaciones específicas que permitirán implementar acciones efectivas para reducir la tasa de abandono y mejorar la lealtad del cliente.

# JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

La deserción de clientes en las entidades financieras representa uno de los principales desafíos para la sostenibilidad y el crecimiento a mediano y largo plazo. A medida que el entorno competitivo en el sector bancario se intensifica, las instituciones enfrentan una presión creciente para retener a sus clientes actuales y evitar que migren a la competencia. El costo asociado a la adquisición de nuevos clientes es significativamente más alto que el de mantener a los existentes, lo que convierte la retención de clientes en una estrategia clave para mejorar la rentabilidad y la eficiencia operativa de la entidad financiera.

Es por ello que el desarrollar un modelo de predicción que realice el análisis de la deserción de clientes dentro de la entidad financiera, permitirá identificar clientes con alta probabilidad de deserción, clientes leales o con los que se debe mejorar la relación comercial, para implementar estrategias de retención de clientes, y entender las razones por las cuales se presenta la deserción, personalizar ofertas y comunicaciones para comprender las necesidades específicas de los diferentes grupos o segmentos de clientes.

Un enfoque segmentado y basado en datos permitirá a la entidad financiera diseñar campañas de retención más eficientes, focalizadas en los clientes con mayor riesgo de deserción, y desarrollar incentivos que respondan directamente a sus necesidades y comportamientos específicos, mejorando tanto la satisfacción del cliente como la estabilidad financiera de la entidad.

A continuación, se aplica la siguiente metodología para el análisis a realizar:

* Recolección de datos
* Limpieza, pre-procesamiento y/o transformación de datos.
* Identificación y descripción de variables
* Visualización de variables
* Selección de modelo estadístico



## RECOLECCION DE DATOS

La fuente de datos corresponde a un dataset público (Kaggle, 2023) de una entidad financiera privada con información relacionada a sus clientes y su comportamiento de deserción en la entidad, con variables como: edad, score de crédito, satisfacción, número de productos bancarios, reclamos y otros factores que podrían influir en la retención o salida de los clientes del banco.

El dataset está compuesto de 10000 registros con 18 columnas. De las cuales 14 son variables numéricas y 4 variables categóricas como se indica a continuación en la **Tabla 1**.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Variable** | **Tipo de Dato** | **Tipo Variable** |
| **1** | *NroRegistro* | int64 | Numérica |
| **2** | *ScoreCredito* | int64 | Numérica |
| **3** | *Edad* | int64 | Numérica |
| **4** | *Saldo* | int64 | Numérica |
| **5** | *NroDeProductos* | int64 | Numérica |
| **6** | *Antiguedad* | int64 | Numérica |
| **7** | *SiTieneTarjeta* | int64 | Numérica |
| **8** | *EsMiembroActivo* | int64 | Numérica |
| **9** | *SalarioEstimado* | int64 | Numérica |
| **10** | *Desertor* | int64 | Numérica |
| **11** | *IdCliente* | int64 | Numérica |
| **12** | *NombreCliente* | object | Categórica |
| **13** | *Pais* | object | Categórica |
| **14** | *Genero* | object | Categórica |
| **15** | *Reclamos* | int64 | Numérica |
| **16** | *CalificacionSatisfaccion* | int64 | Numérica |
| **17** | *TipoTarjeta* | object | Categórica |
| **18** | *PuntosGanados* | int64 | Numérica |

**Tabla 1**. Identificación de datos

Estas variables se las agrupa por sus características en: comportamiento, demográfica, financiera y personales.

* **Comportamiento:** Identifican los patrones de comportamiento del cliente con los productos o servicios bancarios.
  + *Antiguedad*
  + *SiTieneTarjeta*
  + *EsMiembroActivo*
  + *Reclamos*
  + *CalificacionSatisfaccion*
* **Información demográfica:** Describen las características básicas de los clientes.
  + *Edad*
  + *Pais*
  + *Genero*
* **Financieras:** Están relacionadas con la situación económica y las transacciones monetarias del cliente.
  + *ScoreCredito*
  + *Saldo*
  + *NroDeProductos*
  + *SalarioEstimado*
  + *TipoTarjeta*
  + *PuntosGanados*
* **Personales:** Datos propios de cada cliente.
  + *NroRegistro*
  + *IdCliente*
  + *NombreCliente*

## LIMPIEZA, PRE-PROCESAMIENTO Y/O TRANSFORMACIÓN DE DATOS.

La limpieza, reprocesamiento y transformación son pasos fundamentales para preparar los datos de manera que sean útiles y efectivos, tanto en el análisis como en la construcción del modelo.

Para la carga y el análisis de los datos se utiliza Anaconda-JupyterLab con sus librerías y lenguaje Phyton.

**Limpieza de Datos**

* Revisión de datos nulos y duplicados.

|  |  |  |
| --- | --- | --- |
| **Proceso de revisión** | **Función de Phyton** | **Resultado** |
| **Datos nulos** | Isna().sum() | No existieron valores nulos |
| **Datos duplicados** | Duplicated() | No existen datos duplicados |

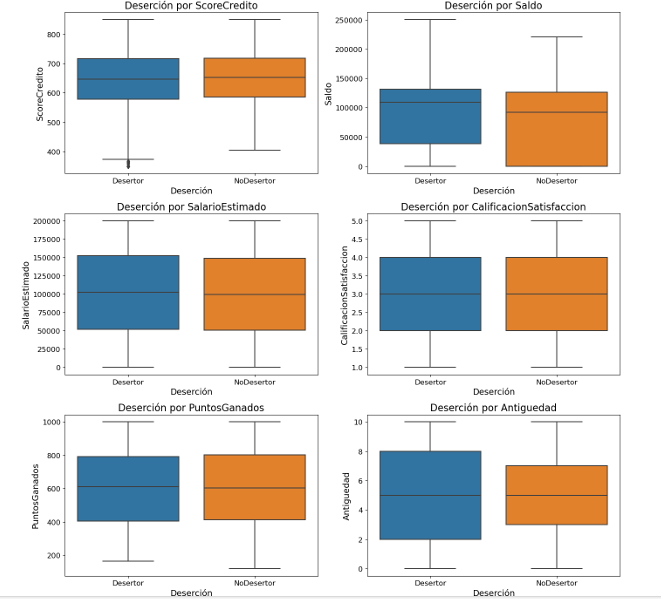
**Tabla 2.** Datos nulos y duplicados

El análisis realizado confirma que el dataset es **completo**, ya que no presenta datos nulos, y no fue necesario realizar ajustes debido a la **ausencia de datos duplicados.**

**Pre-procesamiento de Datos**

* Revisión de datos atípicos aplicado en las variables numéricas o cuantitativas.

Con el uso de diagramas de cajas se puede visualizar las variables que contienen valores atípicos.



**Figura 1**. Datos atípicos

De acuerdo a la **Figura 1**, la variable *ScoreCredito* contiene datos atípicos, los mismos que se procede a eliminarlos dando como resultado 9985 registros, es decir se eliminaron 15 registros.

**Transformación de Datos**

* Generación de rangos de edad

Se genera rangos de edad para facilitar el análisis de patrones y tendencias dentro de cada grupo.

|  |  |
| --- | --- |
| **Rangos de Edad** | **Descripción** |
| 18-30 | Edad del cliente comprendida entre 18 y 30 años |
| 31-40 | Edad del cliente comprendida entre 31 y 40 años |
| 41-50 | Edad del cliente comprendida entre 41 y 50 años |
| 51-60 | Edad del cliente comprendida entre 51 y 60 años |
| 60+ | Edad del cliente mayor a 60 años |

**Tabla 3.** Rangos Edad

* Conversión de variables categóricas

A las variables categóricas *Genero*, *Pais* y *TipoTarjeta* se transforma en numéricas para que éstas sean consideradas en la matriz de correlación y modelo.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Valores Originales** | **Valores  Transformados** | **Tipo de  Codificación** |
| ***Genero*** | Male Female | 0 (Female) 1 (Male) | LabelEncoder |
| ***Pais*** | France Germany Spain | 0 (France) 1 (Germany) 2 (Spain) | LabelEncoder |
| ***TipoTarjeta*** | DIAMOND GOLD SILVER PLATINUM | 0 (DIAMOND) 1 (GOLD) 2 (SILVER) 3 (PLATINUM) | LabelEncoder |

**Tabla 4**. Transformaciones variables categóricas

## IDENTIFICACIÓN Y DESCRIPCIÓN DE VARIABLES

Para el análisis predictivo del abandono de clientes en la entidad, se identifica las siguientes variables dependientes e independientes, que permiten cumplir con el objeto de este estudio:

* **Variable Dependiente**: es aquella que permite identificar si el cliente abandona o no la entidad financiera, para nuestro estudio se la llamará Desertor.
* **Variables Independientes**: son las variables que permiten predecir que cliente va a desertar de la entidad financiera, estas son: antigüedad, si tiene o no tarjeta, si es o no miembro activo, si tiene o no reclamos, calificación de satisfacción, edad, país, género, score de crédito, saldo, número de productos, salario estimado, tipo de tarjeta y puntos ganados.

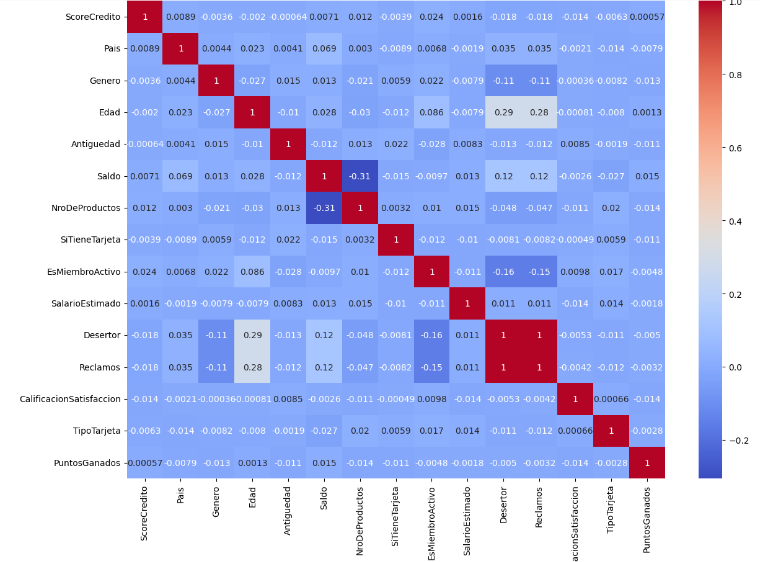
A continuación, se muestra las variables del dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tipo  Variable** | **Clasificación  de Variable** | **Nombre de Variable** | **Descripción  de Variable** |
| **Dependiente** |  | *Desertor* | Si el cliente deserto o no de la Entidad Financiera |
| **Independiente** | Comportamiento | *Antiguedad* | Número de años del cliente en la entidad financiera |
| *SiTieneTarjeta* | Si tiene o no tarjeta |
| *EsMiembroActivo* | Si es miembro Activo |
| *Reclamos* | Cantidad de reclamos realizados |
| *CalificacionSatisfaccion* | Calificación de satisfacción |
| Demográfica | *Edad* | Indica la edad del cliente |
| *Pais* | Ubicación del cliente |
| *Genero* | Género (Masculino, Femenino) |
| Financiera | *ScoreCredito* | Puntaje de Crédito |
| *Saldo* | Saldo actual del cliente |
| *NroDeProductos* | Número de productos del cliente |
| *SalarioEstimado* | Salario estimado |
| *TipoTarjeta* | Tipo de Tarjeta |
| *PuntosGanados* | Puntos ganados por el uso de las tarjetas |
| Personales | *NroRegistro* | Número de Registro |
| *IdCliente* | Identificación del cliente |
| *NombreCliente* | Nombre del cliente |

**Tabla 5.** Variables del dataset

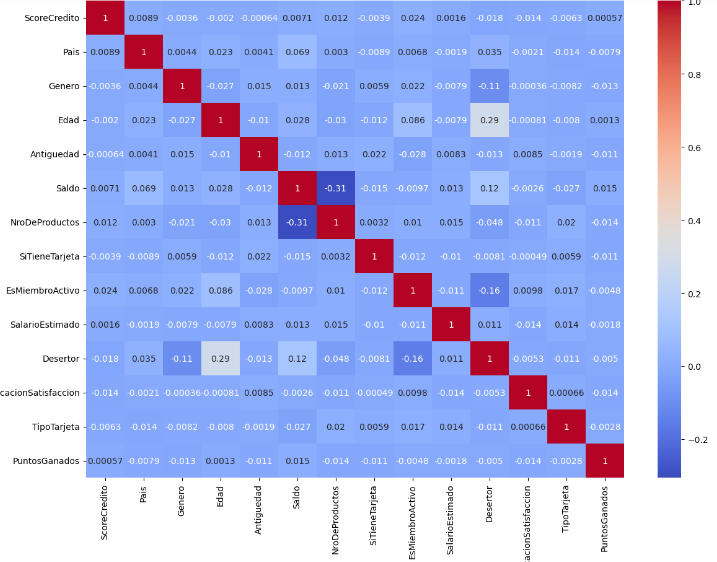
**Matriz de Correlación**

Se analizan las correlaciones entre las variables numéricas para identificar aquellas que están correlacionadas. (Ver código en [**Anexo 1**](#A1)**.**)



**Figura 2**. Matriz de Correlación Inicial.

De acuerdo a la **Figura 2** la variable predictora *Reclamos* por ser altamente correlacionada con la variable a predecir: *Desertor*, se procede a eliminar para evitar que el modelo sobreajuste los datos.



**Figura 3**. Matriz de Correlación Final.

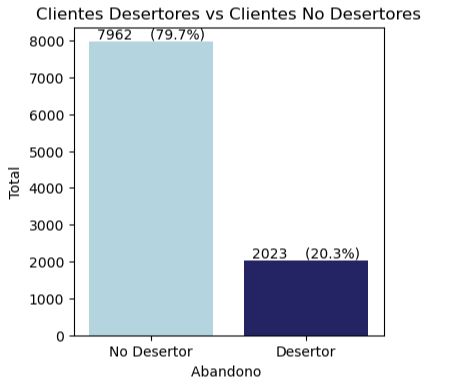
En la matriz de correlación **Figura 3** se identifica varias relaciones entre las variables del dataset, las cuales se indican a continuación:

* **Edad y Desertor**: Existe una correlación positiva considerable (0.29), lo que indica que los clientes de mayor edad tienden a desertar de la entidad financiera con mayor frecuencia.
* **Saldo y Desertor:** Hay una correlación positiva (0.12), sugiriendo que los clientes con mayores saldos tienden a desertar de la entidad financiera en mayor proporción.
* **Es Miembro Activo y Desertor**: Existe una correlación negativa moderada (-0.16), lo que sugiere que los miembros activos tienen menos probabilidades de desertar de la entidad financiera.

Las correlaciones de la matriz indican que las variables como la edad, el saldo, y si es miembro activo tienen mayor impacto en el análisis de deserción de este dataset.

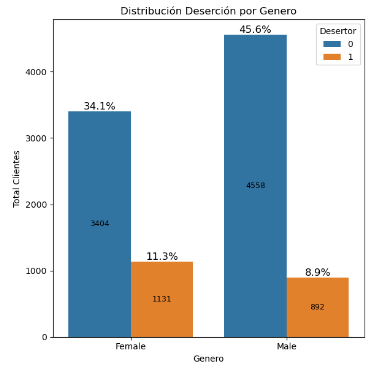
## VISUALIZACIÓN DE VARIABLES

A continuación, se muestra diagramas de barras con el análisis realizado a las variables en base a los datos obtenidos de la entidad financiera con el fin de mostrar cómo se distribuyen los valores, permitiendo identificar diferencias o similitudes entre grupos y detectando posibles patrones. (Ver código en [**Anexo 2.**](#A2))



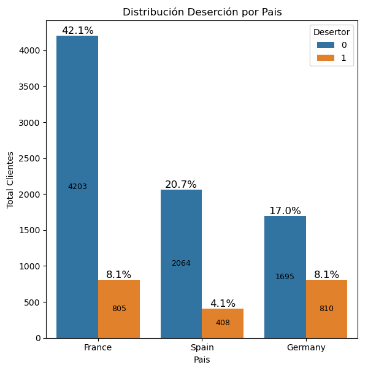
**Figura 4**. Distribución de Deserción

En la **Figura 4**. Se puede observar que del total de 9985 registros que se obtuvo luego de eliminar datos atípicos, el 79.7% son no desertores, y el 20.3% son desertores. Lo que nos indica que, uno de cada cinco clientes desertó de la entidad financiera.



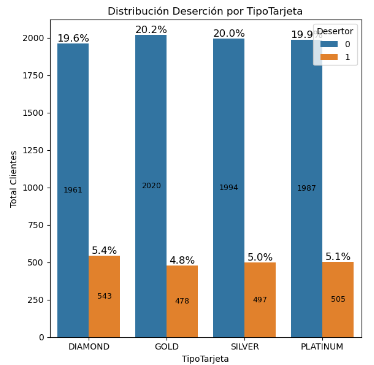
**Figura 5.** Distribución de Deserción por Género

En la **Figura 5**. En el diagrama de barras de Deserción por Género se puede observar que las mujeres (Female) tienen un mayor porcentaje de deserción 11.3% frente al 8.9% de los hombres (Male).



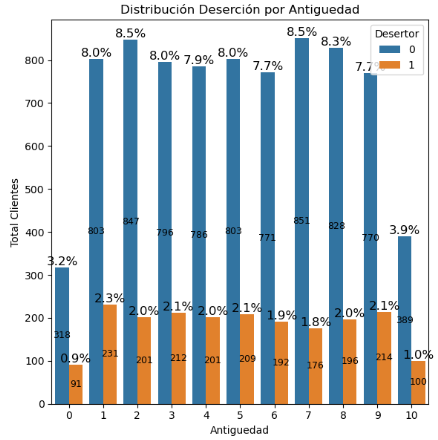
**Figura 6.** Distribución de Deserción por País

En la **Figura 6**. Se puede observar que los Clientes de Alemania son los que más presentan deserción de la entidad financiera con 810 desertores, mientras que en Francia tuvo 805 desertores, se puede ver que en Alemania el porcentaje de desertores es mayor que España y Francia si se compara con el total de clientes por país.



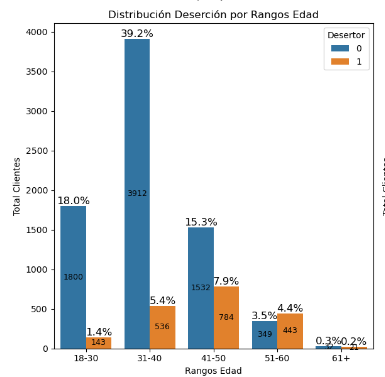
**Figura 7.** Distribución de Deserción por Tipo de Tarjeta

En la **Figura 7.** Se muestra que la mayor cantidad de Desertores se encuentra en los clientes que usan la tarjeta Diamond con 543 desertores, que representa el 5.4% de desertores del total de clientes. Se puede observar también que el tipo de tarjeta no tiene una relevancia importante en la deserción de los clientes ya que los cuatro tipos de tarjeta tienen un porcentaje de deserción similar.



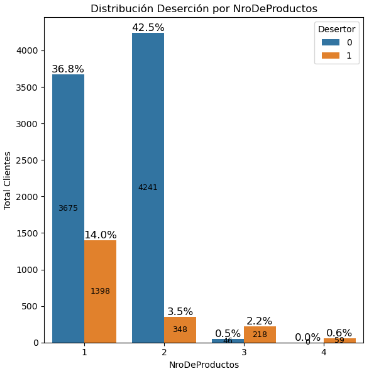
**Figura 8.** Distribución de Deserción por Antigüedad

En el gráfico de la **Figura 8.** se observa que existe mayor cantidad de deserción en clientes en el primer año de antigüedad con 231 desertores, mientras que la menor cantidad de desertores se encuentran en aquellos clientes que tienen menos de un año y los clientes que tienen 10 años de antigüedad.



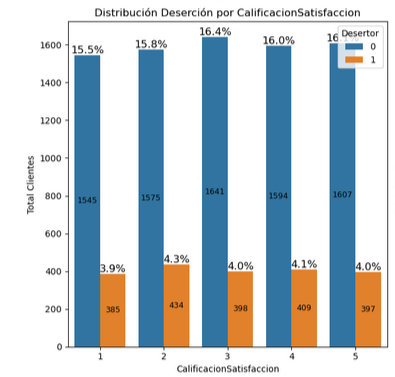
**Figura 9.** Distribución de Deserción por Rangos de Edad

En el análisis de deserción por Rangos de Edad **Figura 9,** se puede observar que los clientes que tienen entre 41 años y 50 años poseen la mayor cantidad de desertores en la entidad financiera. Se puede establecer preliminarmente que la edad es una variable importante en la predicción de la deserción de clientes.



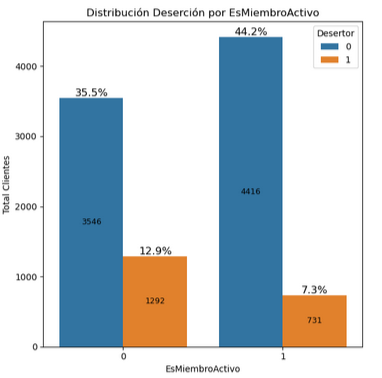
**Figura 10**. Distribución de Deserción por Número de Productos

La **Figura 10.**, nos indica que los clientes que poseen un solo producto tienen mayor porcentaje de deserción.



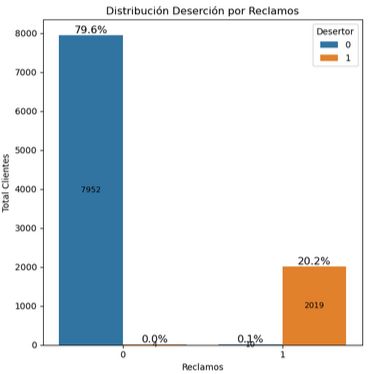
**Figura 11.** Distribución de Deserción por Calificación de Satisfacción

Esta **Figura 11**. muestra que la deserción ocurre más cuando el cliente tiene una calificación de satisfacción de 2, en un rango que va desde el 1 al 5 en la calificación, abarcando un porcentaje de deserción del 4.3% del total de clientes. Sin embargo, no hay una diferenciación considerable entre las distintas calificaciones por lo que se estima preliminarmente que no es una variable de un impacto importante.



**Figura 12.** Distribución de Deserción por Actividad del Cliente

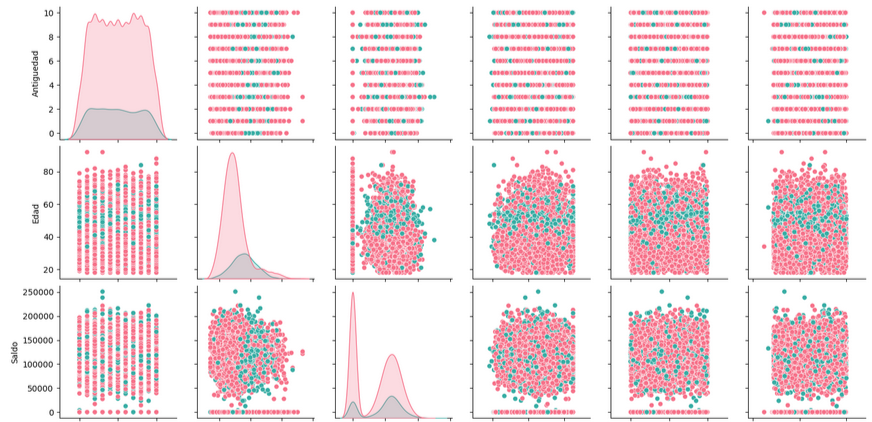
La **Figura 12.** 1 indica si está Activo, 0 si está Inactivo. Según el gráfico el cliente que no presenta actividad, es el que mayor porcentaje de deserción presenta, con un 12.9% frente al 7.3% de deserción de clientes activos.

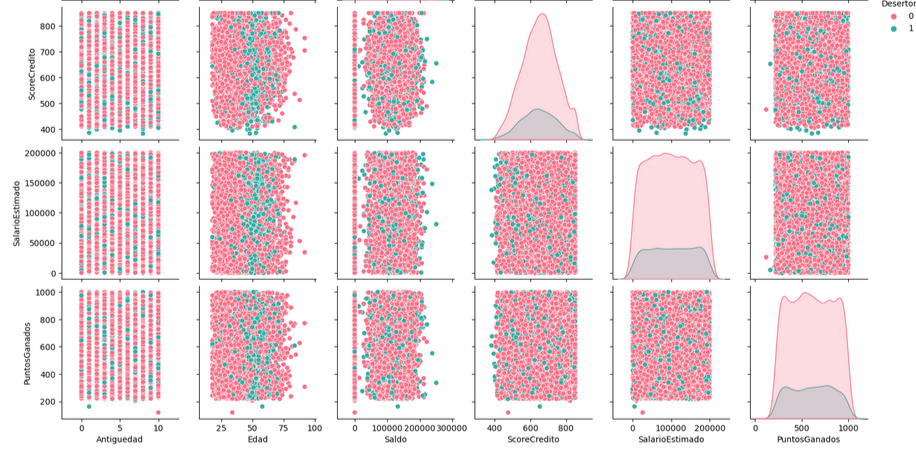


**Figura 13**. Distribución de Deserción por Reclamos

En el diagrama de barras de la **Figura 13** donde se muestra la deserción en base a los clientes que ha realizado reclamos, se observa 2019 desertores y 7952 no desertores, estos valores son similares a los de la **Figura 4** donde se muestra los clientes que efectivamente han desertado 2023 y aquellos que no desertaron 7962. Esto confirma que Reclamos es una variable con alta correlación con Desertor, el considerarla en el modelado puede disminuir su interpretabilidad y por lo tanto no debe ser considerada.

A continuación, se implementan gráficos de dispersión con todas las variables numéricas agrupadas por la variable *Desertor* y en las diagonales se visualiza los KDE (Kernel Density Estimate) de cada variable versus la variable a predecir con la finalidad de evaluar que variables numéricas ofrecen una buena discriminación de la variable objetivo. Esta visualización ofrece una visión integral de las relaciones entre las variables numéricas y su capacidad predictiva.





**Figura 14**. Gráfico dispersión

De la **Figura 14.** Se puede observar en el gráfico de la variable edad una agrupación clara de deserción con los clientes entre 40 y 60 años, que indican que en este grupo los clientes tienen más tendencia a desertar de la entidad financiera que con el resto de los clientes con otras edades, confirmando lo que se pudo mirar con el gráfico de barras de la **Figura 9**.

En el resto de las variables no se identifica agrupaciones claramente diferenciadas entre desertores y no desertores.

## SELECCIÓN DE MODELO ESTADÍSTICO

En base a los datos que se dispone para el análisis: numéricos y categóricos, y el objetivo del estudio, el cual es determinar si hay o no deserción, los algoritmos de acuerdo a la naturaleza de la variable a predecir, que es la deserción (Si/No, 0/1), que van a permitir predecir si un dato pertenece a una clase predeterminada son los algoritmos de clasificación.

Además, según la revisión de la literatura, los modelos de **Regresión Logística, Árbol de Decisión y Random Forest** son los más recomendados para la clasificación y predicción de la deserción de clientes. Este estudio se enfocará en analizar estos tres modelos con el objetivo de compararlos y evaluar su desempeño en la predicción de la deserción.

Para complementar el análisis, se incluye la revisión y aplicación del algoritmo de **Random Survival Forest** para determinar la permanencia del cliente en la entidad financiera.

A continuación, se describe cada uno de los modelos a utilizar y sus respectivas ecuaciones:

La **Regresión Logística** es el modelo que predice la probabilidad de que una muestra pertenezca a una clase determinada, la variable dependiente es binaria. La regresión logística es una prueba predictiva que se utiliza para describir la relación entre un conjunto de variables independientes y una variable binaria dependiente. (Moreno, 2023)

En el caso de este estudio la variable binaria es *Desertor*, donde 0 es No Desertor y 1 es Desertor.

El modelo de regresión logística se puede expresar matemáticamente como:

**P(Y=1|X)= 1/ (1+e−(** **β0+β1X1+β2X2+…+βpXp))**

Donde:

* **P(Y=1|X)**: representa la probabilidad de que la salida **Y** sea igual a 1 dadas las características o variables de entrada **X**.
* **β0, β1, β2,…,βp:** son los coeficientes o pesos asociados con las características de entrada.
* **X1, X2, …, Xp:** denotan las características de entrada.
* **e:** es la base del logaritmo natural, aproximadamente igual a 2,71828.

La función logística también conocida como Sigmoidea generalmente se denota por σ(x) o sig(x) y está dada por:

**σ(X)=1/ (1+e−x)**

Donde:

**X** es la combinación lineal de características de entrada y coeficientes.

**x** = β0+β1X1+β2X2+…+βpXp

Su valor siempre se encuentra entre 0 y 1, por lo que se la implementa para resolver problemas categóricos que tienen dos valores posibles.

Adicionalmente, establece un valor para punto de corte, que en su mayoría se establece en 0,5, por lo que, si se supera el resultado previsto de la curva logística, p > 0.5, la nueva observación se clasificará como y = 1, de lo contrario se clasificará como y = 0.

Los **árboles de decisión** son un modelo de aprendizaje supervisado utilizado tanto en clasificación como en regresión. Estos modelos funcionan dividiendo el conjunto de datos en subconjuntos más pequeños, basándose en características específicas, con el objetivo de tomar decisiones finales.

Cada nodo del árbol representa una pregunta sobre una característica, y las ramas representan las posibles respuestas. Estos árboles se construyen de manera recursiva, dividiendo el conjunto de datos en cada paso hasta alcanzar un criterio de parada, como la pureza de las clases en los nodos o una profundidad máxima.

Se usa el algoritmo de árbol de decisión en situaciones que se tiene varias alternativas posibles con resultados inciertos. Los árboles de decisión se utilizan específicamente para igualar una táctica óptima y alcanzar una meta. (Kozak, 2010)

La ecuación matemática para la construcción de los árboles de decisión es la siguiente (Contreras E., 2017):

Donde:

**n:** es el número de nodos hijos de la partición (número de condiciones de la partición)

**Pj:** es la probabilidad de “caer” en el nodo j, Pj1 es la proporción de elementos de la clase 1 en el nodo j, Pj2 es la proporción de elementos de la clase 2 en el nodo j y así para las c clases.

Otro modelo utilizado para clasificación es el **Random Forest**, éste es un método de aprendizaje de conjuntos (ensemble learning) popular utilizado en tareas de aprendizaje supervisado, especialmente para clasificación y regresión.

Opera construyendo una multitud de árboles de decisión durante la fase de entrenamiento y produce como salida la moda (para clasificación) o la predicción media (para regresión) de los árboles individuales. Random Forest introducen aleatoriedad tanto en la selección de las muestras de datos utilizadas para entrenar cada árbol como en las características consideradas para dividir en cada nodo. Esta aleatoriedad ayuda a mitigar el sobreajuste y mejorar la robustez del modelo, lo que hace que Random Forest sea ampliamente aplicable en diversos ámbitos para tareas de modelado predictivo. (Gonzáles, 2024)

De la revisión teórica de los modelos de clasificación predictivos, para el abandono de los clientes se realiza el siguiente cuadro comparativo a fin de identificar sus características, ventajas y desventajas. (Gonzalez, 2019)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Regresión  Logística** | **Arboles  de decisión** | **Random  Forest** |
| **Características** | • Apropiado para realizar la predicción cuando la variable dependiente es binaria • Clasificación binaria, resultado: probabilidad de que pertenezca a una clase. • Utiliza datos discretos  • Modela probabilidad de eventos. • Predice el valor utilizando datos determinados. | • Estructura jerárquica de decisiones.  • Pueden representar relaciones no lineales para resolver problemas.  • Sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva. | • Combinación de árboles de decisión independientes entre sí para reducir la varianza.  • Aplicable en problemas de clasificación y regresión.  • Robusto  • Crea árboles de decisión a partir de muestras de datos seleccionados al azar. • Obtiene predicciones de cada árbol |
| **Ventajas** | • Fácil de entender y explicar  • Rara vez existe sobreajuste  • Fácil de entrenar sobre grandes cantidades de datos | • Muy fácil de interpretar y entender.  • Rápido.  • Robusto.  • Excelente para aprender relaciones complejas, altamente no lineales.  • Por lo general, pueden lograr un rendimiento bastante alto. | • Puede trabajar en paralelo.  • Rara vez se sobre ajusta.  • Maneja automáticamente los valores perdidos.  • No es necesario transformar ninguna variable.  • No hay necesidad de ajustar parámetros. |
| **Desventajas** | • No adecuado para problemas no lineales.  • Solo funciona con variables binarias  • Puede sufrir con valores atípicos.  • En algunas ocasiones es muy simple para captar relaciones complejas entre variables. | • Propenso al sobreajuste en conjuntos de datos pequeños.  • Pérdida de rendimiento  • Los árboles complejos son difíciles de interpretar.  • Algoritmo sencillo y no tan poderoso para datos complejos. | • Lentos en generar predicciones porque tienen múltiples árboles de decisión.  • Difícil de interpretar en comparación con un árbol de decisión. |

**Tabla 6.** Cuadro comparativo entre los Modelos de Clasificación

**Random Survival Forest (RSF),** es un modelo estadístico utilizado para el análisis de datos de supervivencia en estudios de deserción de clientes, hace referencia al tiempo hasta que ocurre un evento de interés. (Johnson, 2018)

Sus características clave son:

* **Construcción del bosque:** Se construyen múltiples árboles de supervivencia utilizando muestras aleatorias de los datos y subconjuntos de variables. Cada árbol proporciona una estimación de la función de riesgo o de supervivencia, y los resultados se combinan para obtener una predicción más robusta.
* **Medición de la importancia de las variables:** Evalúa la importancia de cada variable en la predicción de la supervivencia, lo cual es útil para identificar factores de riesgo importantes.
* **No paramétrico:** A diferencia de modelos tradicionales como el modelo de Cox, RSF no asume una forma específica de la relación entre las variables y el tiempo hasta el evento, lo que lo hace más flexible.

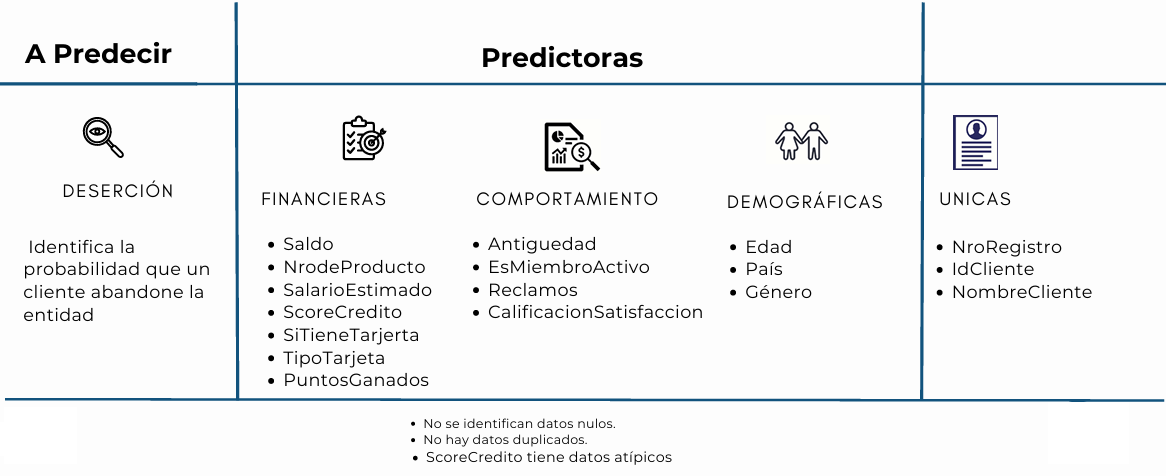
En resumen, el modelo **Random Survival Forest** complementa al **Random Forest**, en el análisis de deserción o abandono de clientes al incluir la variable tiempo, que en el caso de estudio está representada por la antigüedad del cliente. Esto mejora la predicción al no solo determinar si un cliente desertará, sino también cuándo lo hará. De esta manera, es posible diseñar estrategias de retención más precisas al identificar el momento de mayor probabilidad de deserción.

# RESULTADOS



## ANÁLISIS DEL MODELO ESTADÍSTICO

De acuerdo a lo planteado en la sección anterior, para el análisis del dataset de deserción de clientes en la entidad financiera se define las siguientes variables como: predictoras y a predecir.



**Figura 15.** Clasificación de variables

En el dataset no hay datos nulos, no se presentan datos duplicados, la variable *ScoreCredito* es la única que contiene datos atípicos los cuales son tratados para no incluirlos en el modelo. (Ver código en [**Anexo 3**](#A3)**.**)

Se selecciona tres modelos predictivos para realizar el análisis: Regresión Logística, Arboles de Decisión y Random Forest tomando en cuenta el tipo de variable a predecir en este caso Desertor es una variable binaria, esta se adapta perfectamente a la estructura de estos modelos, que están diseñados para clasificar y predecir resultados en un contexto binario.

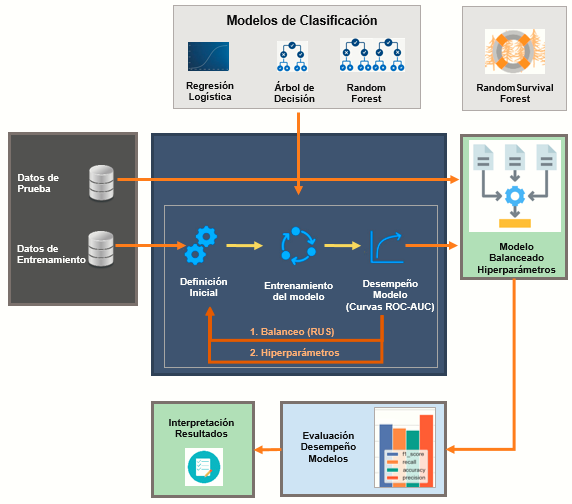
Se desarrolla a continuación cada uno de los modelos seleccionados para este análisis, en cada uno de ellos se genera la matriz de confusión, las métricas para su evaluación de efectividad, así como el diagrama AUC-ROC.

Para los modelos de: Regresión Logística, Arboles de decisión y Random Forest, se obtiene las siguientes métricas para evaluar: f1-score, recall y AUC. Se seleccionaron estas métricas ya que con f1-score se trata de encontrar un balance entre precisión y exhaustividad y con recall obtener la mayor proporción de verdaderos positivos sobre el total de positivos reales. Esto es crucial si el costo de no detectar a un cliente que va a desertar es alto, ya que se requiere identificar la mayor cantidad posible de clientes en riesgo. Y el AUC nos permite obtener una mayor precisión en la predicción.

El modelamiento se realiza con el uso de Phyton, para el Árbol de Decisión y Random Forest se crea y entrena el modelo con un random\_state igual a 42 para controlar la aleatoriedad en el proceso de construcción del modelo. Específicamente, random\_state = 42 fija la semilla del generador de números aleatorios a 42.

Para Random Forest se usa n\_estimators igual 100 lo que indica que el modelo construirá un total de 100 árboles de decisión. El tener 100 árboles ayuda a mejorar la precisión y la estabilidad del modelo.

El modelamiento preliminar se lo realiza sin Balanceo, posteriormente de acuerdo a los valores de las métricas es necesario realizar el Balanceo y el uso de Hiperparámetros. (Ver código en [**Anexo 4.**)](#A4)



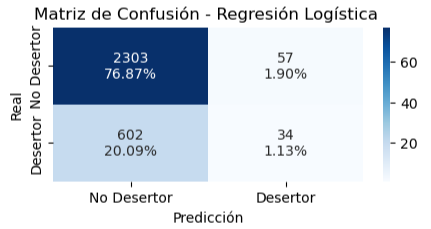
**Figura 16**. Diagrama de modelamiento

**MODELO DE REGRESION LOGISTICA**

A continuación, se muestra la matriz de confusión, métricas y curva ROC del modelo de regresión logística. (Ver código en [**Anexo 5**](#A5)**.**)

**Matriz de Confusión**

Con el modelo de Regresión Logística se obtiene los siguientes datos:



**Figura 17**. Matriz de Confusión de Regresión Logística

Esta matriz de confusión nos indica que:

Predijo correctamente 2303 No Desertores y 34 Desertores.

Predijo incorrectamente 602 No Desertores, siendo estos Desertores.

Predijo incorrectamente 34 Desertores, siendo estos No Desertores.

En conclusión, el modelo está identificando correctamente la mayoría de los casos positivos es decir a los No desertores, pero incorrectamente los Desertores, que es la finalidad de éste análisis. Podría ser necesario ajustar el modelo, con técnicas de balanceo, para mejorar la capacidad de detectar a los Desertores.

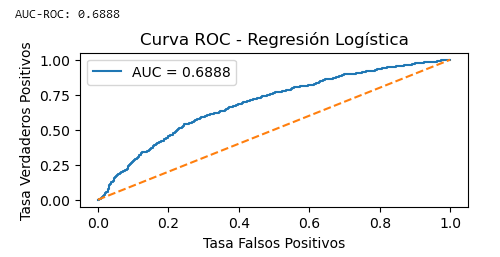
**Métricas**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo Regresión Logística** | **Recall** | **Precision** | **F1-Score** |
| No Desertor (0) | 0.9758 | 0.7928 | 0.8748 |
| Desertor(1) | 0.0535 | 0.3736 | 0.0935 |
| **Accuracy** |  |  | **0.7800** |
| Weighted Avg | 0.7800 | 0.7038 | 0.7090 |

**Tabla 7.** Métricas Regresión Logística

El modelo no es eficiente, con un alto desempeño en la identificación de **no desertores** (0), pero con un rendimiento muy bajo en la identificación de **desertores** (1), que es la que se requiere más efectividad en este análisis. Esto sugiere que el modelo está sesgado hacia la clase mayoritaria, se puede mejorar el rendimiento en la clase minoritaria (desertores) con el uso de técnicas de balanceo de datos.

**Curva ROC**



**Figura 18.** Curva ROC Regresión Logística

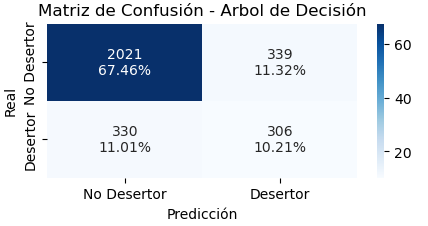
El AUC-ROC de **0.6888** sugiere que el modelo tiene un rendimiento aceptable al ser mayor del 0.5, pero **no es muy óptimo**. Esto implica que, aunque el modelo es capaz de hacer distinciones entre desertores y no desertores, existe un margen significativo de mejora.

**MODELO DE ARBOLES DE DECISION**

**Matriz de Confusión**

Con el modelo de Arboles de Decisión se obtiene los siguientes datos:

(Ver código en [**Anexo 6**](#A6).)



**Figura 19.** Matriz de Confusión Árbol de Decisión

Esta matriz de confusión nos indica que:

Predijo correctamente 2021 No Desertores y 306 Desertores.

Predijo incorrectamente 330 No Desertores, siendo estos Desertores.

Predijo incorrectamente 339 Desertores, siendo estos No Desertores.

En resumen, el modelo está identificando la mayoría de los casos de falsos negativos, es decir los No desertores, pero en cuanto a identificar a los Desertores, si bien éste modelo predijo mayor porcentaje de desertores que el modelo de Regresión Logística, aún la predicción correcta representa un porcentaje bajo.

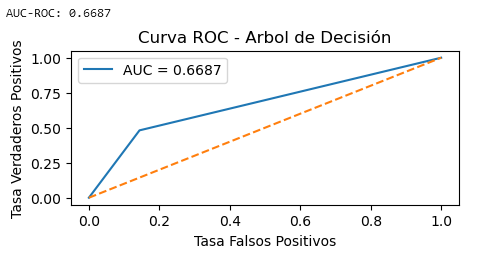
**Métricas**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo Árbol de Decisión** | **Recall** | **Precision** | **F1-Score** |
| No Desertor (0) | 0.8564 | 0.8596 | 0.8580 |
| Desertor(1) | 0.4811 | 0.4744 | 0.4778 |
| **Accuracy** |  |  | **0.7767** |
| Weighted Avg | 0.7767 | 0.7779 | 0.7773 |

**Tabla 8.** Métricas Árbol de Decisión

El modelo tiene un buen desempeño general si consideramos que el accuracy es de 0.7767 sin embargo, el recall de la clase desertora es bajo (0.48).

**Curva ROC**



**Figura 20.** Curva ROC Árbol de Decisión

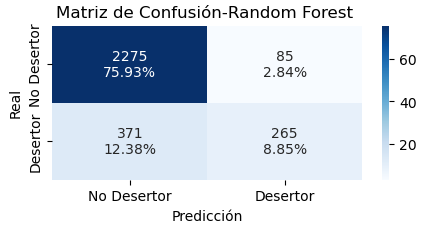
El AUC-ROC que se obtiene con la probabilidad de la variable Desertor es de **0.6687** lo que indica que el modelo tiene un rendimiento razonable pero no eficiente en la discriminación entre desertores y no desertores.

**MODELO DE RANDOM FOREST**

**Matriz de Confusión**

A continuación, se muestra los resultados obtenidos del modelo de Random Forest:

(Ver código en [**Anexo 7**](#A7).)



**Figura 21**. Matriz de Confusión Random Forest

Esta matriz de confusión nos indica que:

Predijo correctamente 2275 No Desertores y 265 Desertores.

Predijo incorrectamente 371 No Desertor, siendo este Desertor.

Predijo incorrectamente 85 Desertores, siendo estos No Desertores.

En conclusión, el modelo está identificando correctamente la mayoría de los No Desertores, no así con los Desertores que es la clase de mayor importancia en el análisis.

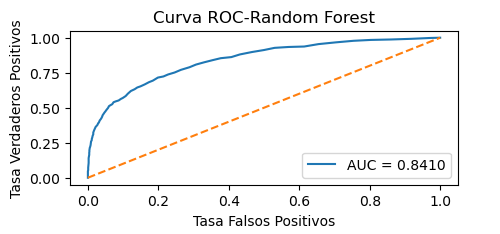
**Métricas**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo Random Forest** | **Recall** | **Precision** | **F1-Score** |
| No Desertor (0) | 0.9640 | 0.8598 | 0.9089 |
| Desertor(1) | 0.4167 | 0.7571 | 0.5375 |
| **Accuracy** |  |  | **0.8478** |
| Weighted Avg | 0.8478 | 0.8380 | 0.8301 |

**Tabla 9.** Métricas Random Forest

El modelo puede identificar correctamente tanto los no **desertores** (0) como los **desertores** (1), los valores de las métricas como: precisión, recall, f1-score están cercanos al 1, lo que nos indica que el modelo es eficiente.

**Curva ROC**



**Figura 22.** Curva ROC Random Forest

El AUC-ROC de **0.8410** sugiere que el modelo tiene un rendimiento eficiente, y es el modelo que obtuvo el valor más cercano al 1.0, aunque el recall de la clase Desertor se mantiene bajo (0.4167).

De acuerdo a los valores obtenidos en cada modelo se puede observar que el modelo que tuvo mejor rendimiento fue el Random Forest, aunque el valor del recall de la clase Desertor es bajo, por lo que se aplica técnicas de balanceo para tratar de obtener mejores resultados.

Para obtener un mejor recall de la clase Desertor en el modelo se aplicó varias técnicas de Balanceo (Zhining, 2021) , entre estas:

**SMOTE (Synthetic Minority Over-sampling Technique)** es un método que genera nuevas muestras sintéticas de la clase minoritaria para abordar el problema de desbalance de clases.

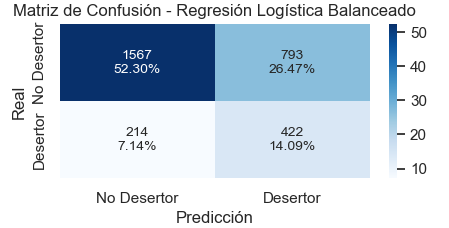
**SVMSMOTE:** es una variante de SMOTE que utiliza un clasificador SVM (Support Vector Machine) para mejorar la creación de las muestras sintéticas, lo que le permite abordar los casos donde las clases están más cercanas o son más difíciles de separar.

**Random Under-Sampling (RUS)** funciona eliminando aleatoriamente ejemplos de la clase mayoritaria hasta que el número de ejemplos en ambas clases sea más balanceado. Esto reduce el sesgo hacia la clase mayoritaria y mejora la capacidad del modelo para identificar correctamente la clase minoritaria.

Siendo Random Under-Sampling con el que se obtuvo mejores resultados, el cual se aplica a continuación en cada uno de los modelos. (Ver código en [**Anexo 8**](#A8).)

**MODELO DE REGRESION LOGÍSTICA CON BALANCEO RUS**

**Matriz de Confusión**



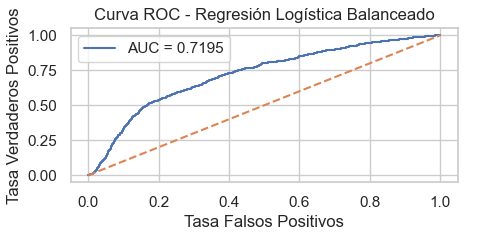
**Figura 23**. Matriz de Confusión Regresión Logística Balanceado

**Métricas**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo Regresión Logística Balanceado** | **Recall** | **Precision** | **F1-Score** |
| No Desertor (0) | 0.6640 | 0.8798 | 0.7568 |
| Desertor(1) | 0.6635 | 0.3473 | 0.4560 |
| **Accuracy** |  |  | **0.6639** |
| Weighted Avg | 0.6639 | 0.7668 | 0.6930 |

**Tabla 10.** Métricas Regresión Logística Balanceado

**Curva ROC**



**Figura 24.** Curva ROC Regresión Logística Balanceado

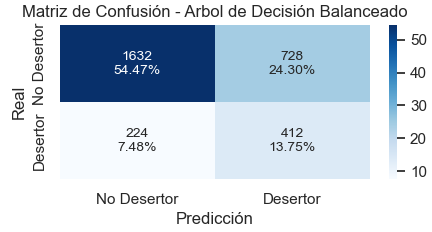
Con el modelo de Regresión Logística con Balanceo Rus mejora la predicción de los Desertores, prediciendo correctamente 422 Desertores respecto a los 34 Desertores que se predijo sin Balanceo.

El recall de la clase Desertora incrementó de 0.0535 a 0.6635 alcanzando una precisión de 0.3473.

El AUC aumentó de 0.6888 a 0.7195.

**MODELO DE ARBOL DE DECISION CON BALANCEO RUS**

**Matriz de Confusión**



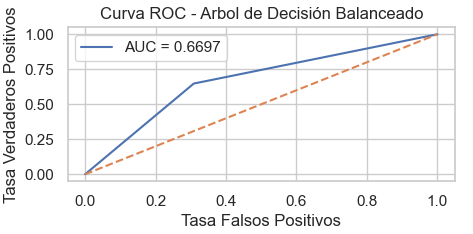
**Figura 25**. Matriz de Confusión Árbol de Decisión Balanceado

**Métricas**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo Árbol de Decisión Balanceado** | **Recall** | **Precision** | **F1-Score** |
| No Desertor (0) | 0.6915 | 0.8793 | 0.7742 |
| Desertor(1) | 0.6478 | 0.3614 | 0.4640 |
| **Accuracy** |  |  | **0.6822** |
| Weighted Avg | 0.6822 | 0.7694 | 0.7083 |

**Tabla 11.** Métricas Árbol de Decisión Balanceado.

**Curva ROC**



**Figura 26**.Curva ROC Árbol de Decisión Balanceado

En resumen, en el modelo de Árbol de Decisión con Balanceo se obtiene:

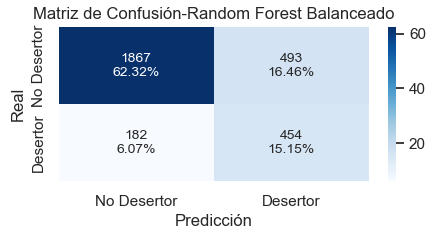
Una mejor predicción de los Desertores de 412 respecto a los 306 Desertores que se predijo sin Balanceo.

El recall de la clase Desertora incrementó de 0.4811 a 0.6478, bajo la precisión de 0.4744 a 0.3614.

El AUC aumentó no significativamente de 0.6687 a 0.6697.

**MODELO RANDOM FOREST CON BALANCEO RUS**

**Matriz de Confusión**



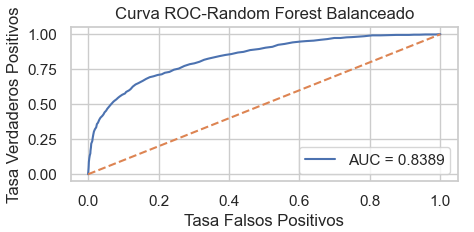
**Figura 27.** Matriz de Confusión Random Forest Balanceado

**Métricas**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo Random Forest Balanceado** | **Recall** | **Precision** | **F1-Score** |
| No Desertor (0) | 0.7911 | 0.9112 | 0.8469 |
| Desertor(1) | 0.7138 | 0.4794 | 0.5736 |
| **Accuracy** |  |  | **0.7747** |
| Weighted Avg | 0.7747 | 0.8195 | 0.7889 |

**Tabla 12.** Métricas Random Forest Balanceado

**Curva ROC**



**Figura 28.** Curva ROC Random Forest Balanceado

Se observa en el modelo de Random Forest con Balanceo lo siguiente:

Se predice mejor los Desertores de 454 respecto a los 265 Desertores que se predijo sin Balanceo.

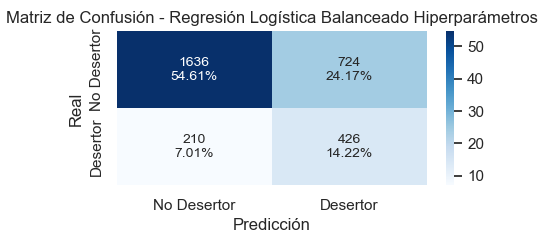
El recall de la clase Desertora incrementó de 0.4167 a 0.7138, bajo la precisión de 0.7571 a 0.4749.

El AUC disminuyó de 0.8410 a 0.8389

En conclusión, si bien con los modelos balanceados se obtuvo mejores resultados en las métricas, la precisión de la clase desertora continua bajo 0.50 por lo que el modelo no está prediciendo correctamente la clase minoritaria que es la clase de interés de este estudio, por lo que a continuación se aplicará hiperparámetros para intentar mejorar el análisis realizado hasta el momento. (Ver código en [**Anexo 9**](#A9).)

**MODELO DE REGRESION LOGÍSTICA CON BALANCEO RUS E HIPERPARAMETROS**

**Matriz de Confusión**



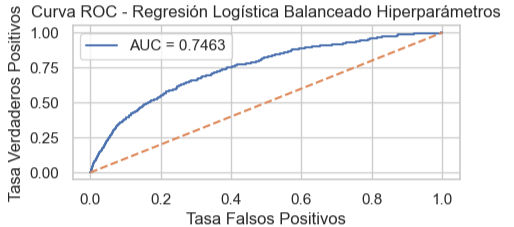
**Figura 29**. Matriz de Confusión Regresión Logística Balanceado e Hiperparámetros

**Métricas**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo Regresión Logística Balanceado e Hiperparámetros** | **Recall** | **Precision** | **F1-Score** |
| No Desertor (0) | 0.6932 | 0.8862 | 0.7779 |
| Desertor(1) | 0.6698 | 0.3704 | 0.4770 |
| **Accuracy** |  |  | **0.6883** |
| Weighted Avg | 0.6883 | 0.7767 | 0.7141 |

**Tabla 13.** Métricas Regresión Logística Balanceado e Hiperparámetros

**Curva ROC**



**Figura 30**. Curva ROC Regresión Logística Balanceado Hiperparámetros

Los resultados que se obtiene con el modelo de Regresión Logística con Balanceo e Hiperparámetros son los siguientes:

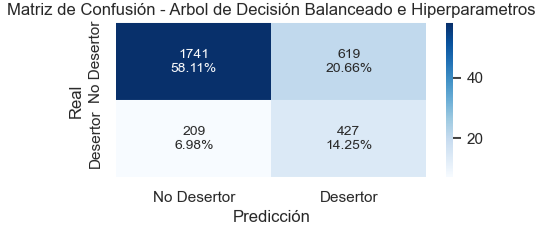
Predice mejor la clase Desertor de 426 respecto a los 422 Desertores que se predijo con Balanceo.

El recall de la clase Desertora incrementó de 0.6698 a 0.7138, bajo la precisión de 0.3473 a 0.3704.

El AUC disminuyó de 0.7195 a 0.7463

**MODELO DE ARBOL DE DECISION CON BALANCEO RUS E HIPERPARAMETROS**

**Matriz de Confusión**



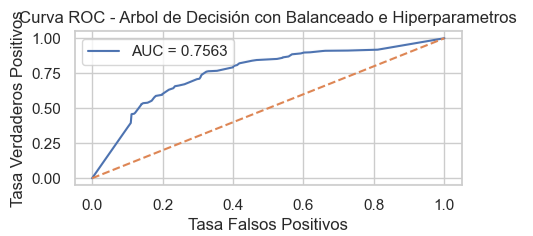
**Figura 31.** Matriz de Confusión Árbol de Decisión Balanceado e Hiperparámetros

**Métricas**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo Árbol de Decisión Balanceado e Hiperparámetros** | **Recall** | **Precision** | **F1-Score** |
| No Desertor (0) | 0.7377 | 0.8928 | 0.8079 |
| Desertor(1) | 0.6714 | 0.4082 | 0.5077 |
| **Accuracy** |  |  | **0.7236** |
| Weighted Avg | 0.7236 | 0.7899 | 0.7442 |

**Tabla 14.** Métricas Árbol de Decisión Balanceado e Hiperparámetros

**Curva ROC**



**Figura 32.** Curva ROC Árbol Decisión Balanceado e Hiperparámetros

En el modelo de Árbol de Decisión con Balanceo e Hiperparámetros, se obtiene:

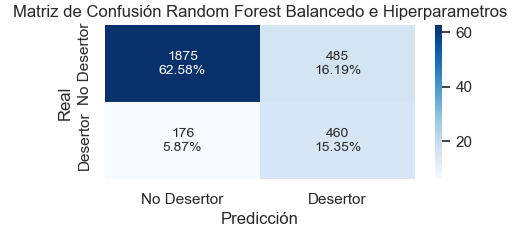
Una mejor predicción de los Desertores de 427 respecto a los 412 Desertores que se predijo con Balanceo.

El recall de la clase Desertora incrementó de 0.6478 a 0.6714, incrementó la precisión de 0.3614 a 0.4082

El AUC aumentó de 0.6697 a 0.7463.

**MODELO RANDOM FOREST CON BALANCEO RUS E HIPERPARAMETROS**

**Matriz de Confusión**



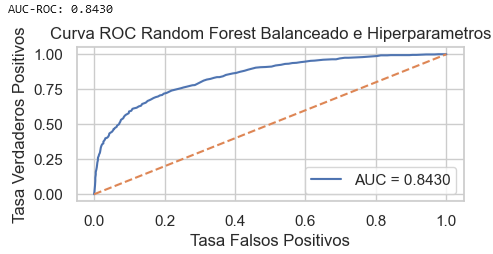
**Figura 33.** Matriz de Confusión Random Forest Balanceado e Hiperparámetros

**Métricas**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo Random Forest Balanceado e Hiperparámetros** | **Recall** | **Precision** | **F1-Score** |
| No Desertor (0) | 0.7945 | 0.9142 | 0.8501 |
| Desertor(1) | 0.7233 | 0.4868 | 0.5819 |
| **Accuracy** |  |  | **0.7794** |
| Weighted Avg | 0.7794 | 0.8235 | 0.7932 |

**Tabla 15**. Métricas Random Forest Balanceado e Hiperparámetros

**Curva ROC**



**Figura 34**. Curva ROC Random Forest Balanceado e Hiperparámetros

Se observa en el modelo de Random Forest con Balanceo e Hiperparámetros lo siguiente:

La predicción de los Desertores mejoró prediciendo 460 desertores respecto a los 454 Desertores que predijo con Random Forest sólo con Balanceo.

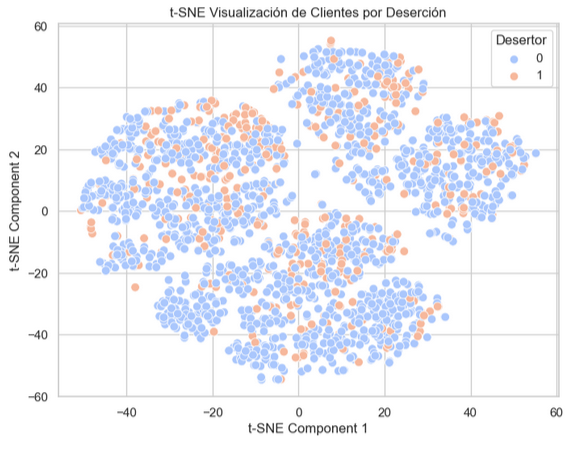
El recall de la clase desertora (1) se incrementó de 0.7138 a 0.7233, la precisión también aumentó de 0.4794 a 0.4868.

El AUC aumentó de 0.8389 a 0.8430.

En conclusión en cuanto a los resultados obtenidos, el modelo **Random Forest con Balanceo e Hiperparámetros** es el que mejor AUC se obtuvo con **0.8430**,  
sin embargo al tener un recall bajo (menor a 0.50) en la clase Desertora de **0.4868**, nos indica que el modelo tiene un bajo rendimiento. Por lo que para complementar este análisis se aplica reducción de dimensionalidad para visualizar la separabilidad de la clase Desertor.

**REDUCCION DE DIMENSIONALIDAD**

El **t-SNE** (t-distributed Stochastic Neighbor Embedding) es una técnica de reducción de dimensionalidad utilizada principalmente en el análisis de datos, particularmente para la visualización de datos de alta dimensión en un espacio de baja dimensión (como 2D o 3D) manteniendo la estructura global de los datos. (Miguel, 2022)



**Figura 35**. t-SNE de Clientes por Deserción

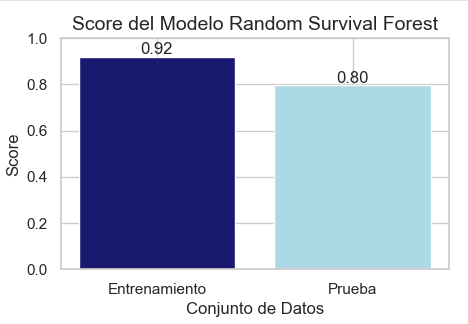
La **Figura 34.** muestra que no existe separabilidad entre las clases desertor y no desertor. Si bien el modelo en general tiene un rendimiento sobre 0.84, la visualización del t-SNE del modelo nos indica que no es eficiente. (Ver código en [**Anexo 10**](#A10).)

La propuesta es enriquecer el dataset con variables relevantes que permitan incrementar este rendimiento, se sugiere como posibles variables a considerar las siguientes: número de reclamos, promedio de tiempo de resolución de reclamos, número de transacciones en un período específico, estado civil, ocupación, número de dependencias. Esta sugerencia de variables se basa en referencias de casos de estudio similares en las que estas variables han sido utilizadas y han dado un buen rendimiento al modelo, estos casos de estudio son citados en la sección de recomendaciones.

Sin embargo, con la información actual del dataset se puede obtener información importante en cuanto al análisis de Supervivencia como se muestra a continuación.

**MODELO RANDOM SURVIVAL FOREST**

El modelo Random Survival Forest amplía al Random Forest al incorporar el factor tiempo, representado por la antigüedad del cliente, en el análisis de deserción.



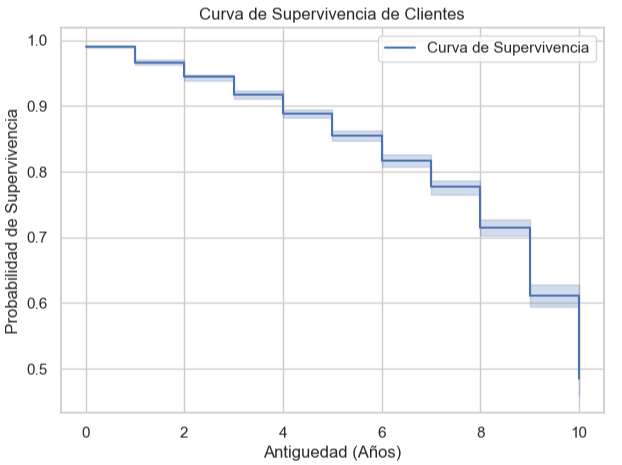
**Figura 36.** Score Modelo RSF

El modelo Random Survival Forest tiene un buen desempeño, aprende de los patrones del conjunto de entrenamiento (0.92) y generaliza a datos nuevos (0.79). Aunque hay diferencia entre el training score y el test score, ambos valores son significativos, lo que indica que el modelo es aceptable. (Ver código en [**Anexo 11**](#A11)**.)**

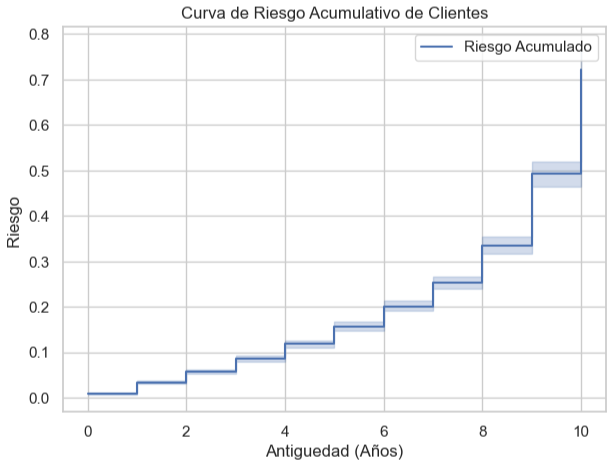
**ANALISIS DE SUPERVIVENCIA**

El análisis de supervivencia se realiza en base a la variable Antigüedad para identificar el tiempo de duración del cliente en la entidad como una técnica complementaria para analizar el comportamiento del cliente.

A continuación, se muestra las curvas de supervivencia y de riesgo del cliente.



**Figura 37.** Curva de Supervivencia de Clientes

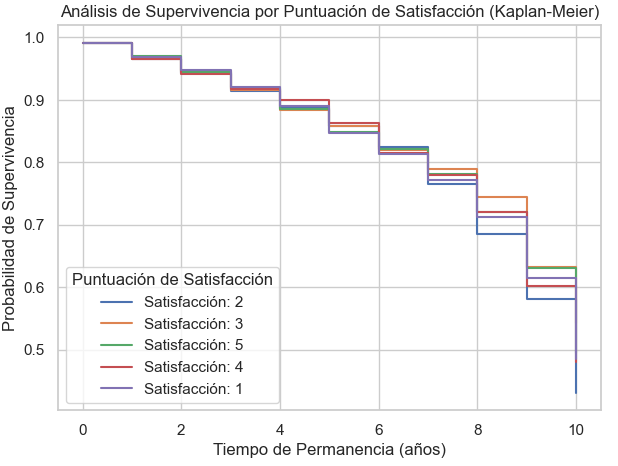
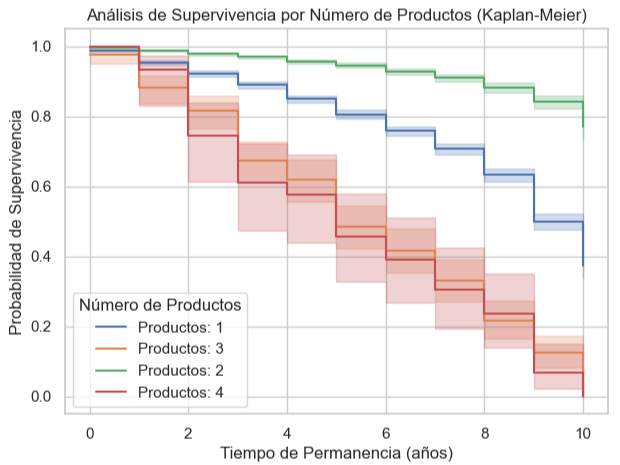
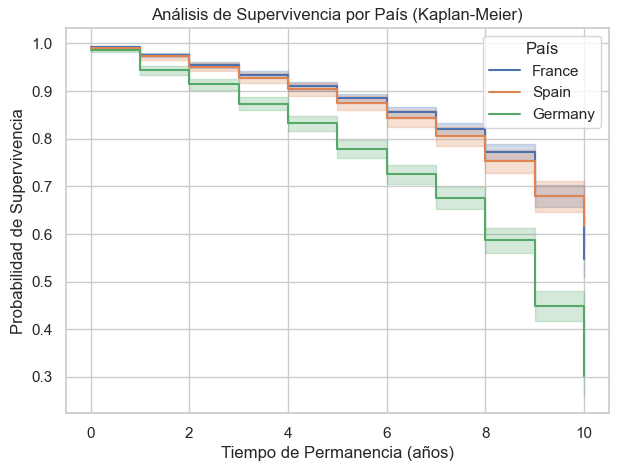
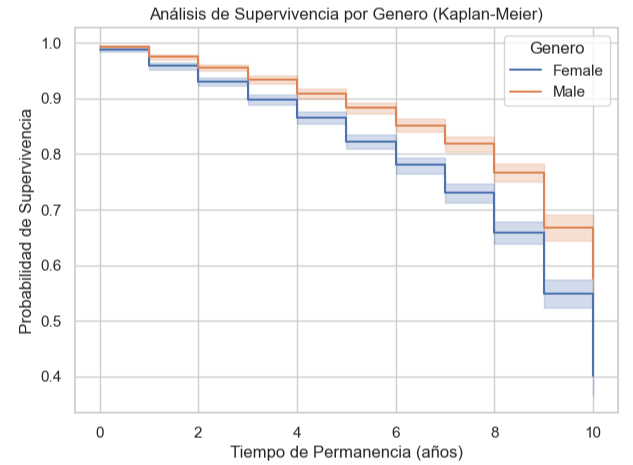


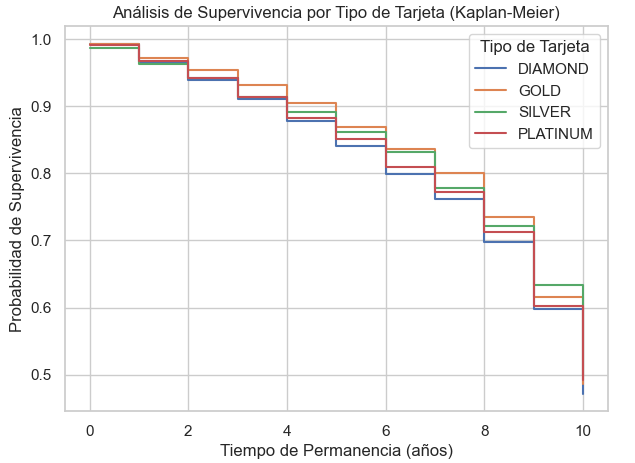
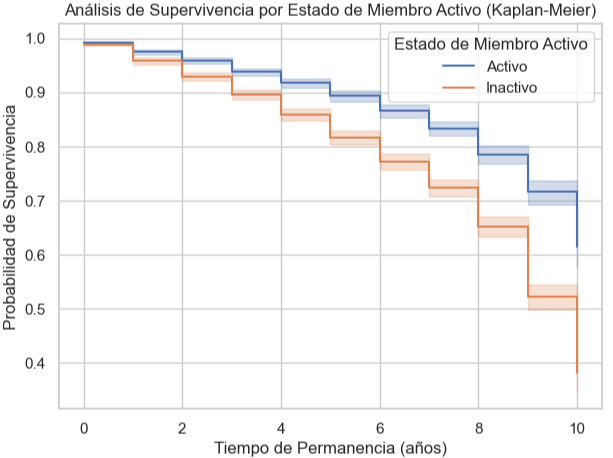
**Figura 38.** Curvas de Riesgo del Cliente

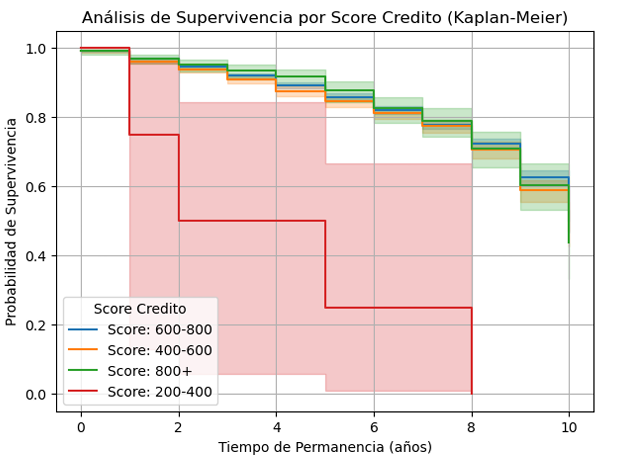
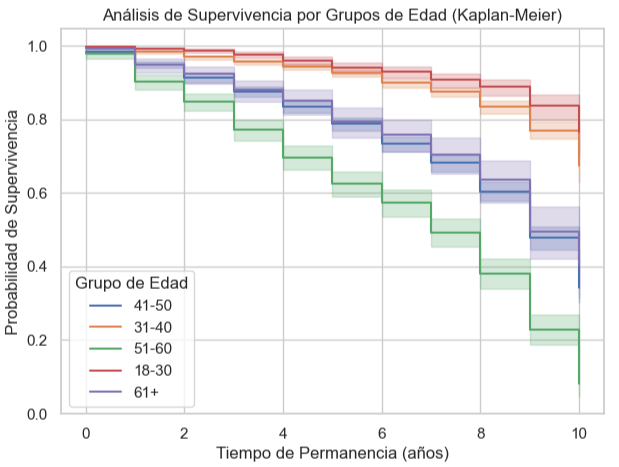
La curva de supervivencia (Kaplan-Meier) de la **Figura 35** nos indica que la probabilidad de supervivencia disminuye mientras más años el cliente permanezca en la entidad y por lo tanto como indica la curva de riesgo (Nelson-Aalen) por su parte confirma que existe más riesgo de que los clientes con más años de antigüedad abandonen la entidad.

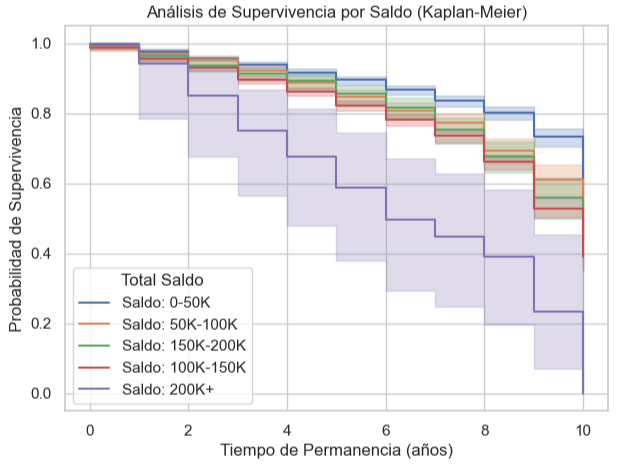
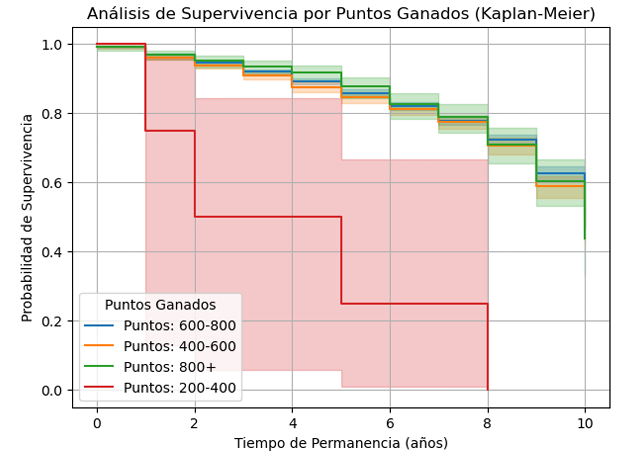
**Curvas de supervivencia**

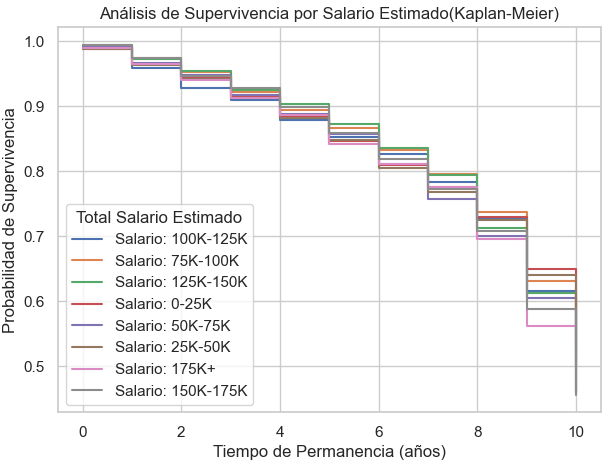
Se muestra a continuación diagramas de supervivencia generadas respecto a las variables del dataset respecto a la antigüedad o tiempo de permanencia del cliente en la entidad.











En cuanto al **Género**, el gráfico de análisis de supervivencia nos muestra que el hombre tiene mayor permanencia en la entidad que las mujeres.

Respecto al **País**, la mayor probabilidad de supervivencia la posee los clientes con residencia en Francia.

En el análisis de supervivencia respecto al **Número de Productos**, aquellos que tienen menos productos tienen mayor duración en la entidad.

Los clientes que son **Miembro Activo** tienen mayor probabilidad de supervivencia.

Respecto a la **Edad**, la curva de supervivencia sugiere que los clientes con mayor edad tienden a permanecer menos tiempo en la entidad, siendo los clientes en el rango de edad de 51-60 años, los que menos probabilidad de supervivencia poseen, mientras que los de 18 a 30 años tienen mayor permanencia.

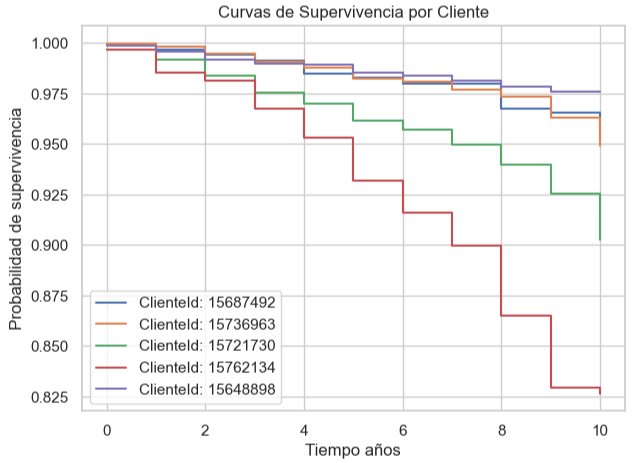
En cuanto a las variables **Score de Crédito y Puntos Ganados,** se puede observar que los clientes que menos permanecen en la entidad financiera son aquellos que tienen un score de crédito y puntos ganados que están en el rango de 200 a 400.

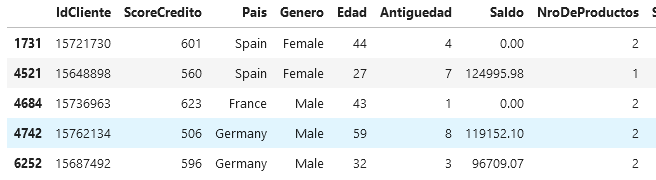
En relación a los **Saldos**, los clientes con saldos bajos tienen una mayor duración en la entidad.

La permanencia o duración en la entidad de clientes respecto a las variables **Calificación de Satisfacción, Tipo de Tarjeta y Salario Estimado,** el gráfico no muestra gran variabilidad, por lo que se concluye que no son variables relevantes para determinar la permanencia del cliente.

**Curvas de Supervivencia por cliente**

Con el uso de Phyton es posible generar curvas de supervivencia por cliente, haciendo más exacto el estudio, lo cual permitirá un manejo más personalizado de los clientes y su posible deserción. (Ver código en [**Anexo 12**](#A12).)





En el gráfico se muestra la supervivencia de 5 clientes, siendo el cliente con ID 15762134 el que menos probabilidad de supervivencia o permanencia presenta en el tiempo analizado de 10 años, mientras que el cliente con ID 15687492 es el que tiene más probabilidad de supervivencia en el mismo lapso de tiempo.

El cliente con ID 15762134 es hombre de 59 años de Alemania con una antigüedad de 8 años, posee 2 productos, con calificación de satisfacción 2, puntos ganados sobre los 800.

El cliente con ID 15648898 es mujer de 27 años de España con una antigüedad de 7 años, posee 1 producto, con calificación de satisfacción 5, puntos ganados entre 400-600.

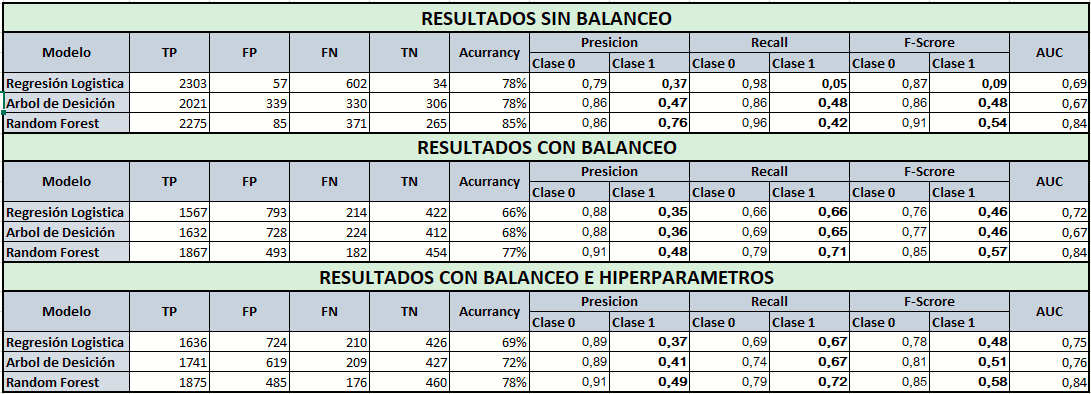
Los dos clientes no tienen reclamos son miembros activos, su saldo se encuentra en el rango de 100K-125K, así como su score de crédito esta entre 400-600.

Esto sugiere que, entre estos dos clientes, las variables más determinantes para la permanencia son la edad y el número de productos. Por ello, el cliente con el ID 15762134 es quien tiene menor probabilidad de continuar, ya que, como se mencionó en el análisis de supervivencia por variable, los clientes con mayor edad y número de productos son los que presentan una menor probabilidad de permanencia en la entidad financiera.

El proceso realizado en este caso de estudio se encuentra en la plataforma Github, en la siguiente referencia bibliográfica. (Prado A. & Yambay M., 2024)

## INTERPRETACIÓN DE RESULTADOS

La técnica de balanceo y la aplicación de hiperparámetros en los modelos de Regresión Logística, Árbol de Decisión y Random Forest mejoraron la capacidad de identificar a los clientes que abandonarán el banco. Los resultados de estas combinaciones presenten a continuación:



**Tabla 16.** Comparación Técnicas Aplicadas a modelos

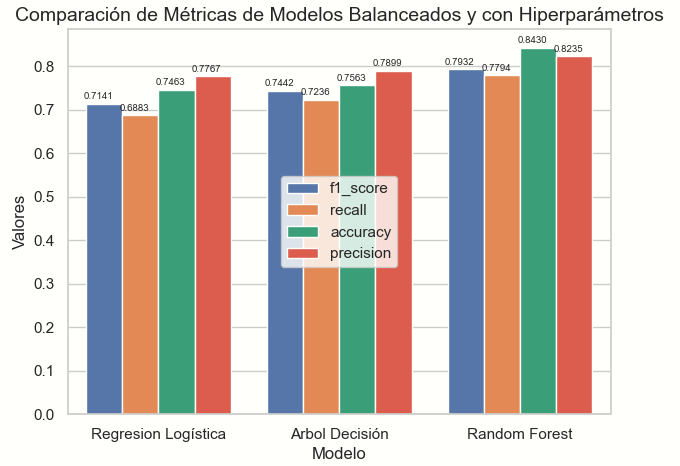
Se identifica un incremento considerable en cuanto al rendimiento (recall) en la predicción de clientes desertores al aplicar la técnica de balanceo Random Under-Sampling, para el modelo de Regresión Logística de 5% a 66%, Arboles de decisión de 48% a 65% y para Random Forest de 42% a 71%.

La aplicación de hiperparámetros, incremento el recall del modelo de regresión logística en 1% y el modelo de Arboles de Decisión en 2%. Para Random Forest se incrementa el recall del 72%. Con respecto a la precisión se ve un leve incremento en los tres modelos, para Regresión Logística un 37%, para Árbol de Decisión 41% y Random Forest una presión de 49%.

## COMPARACION DE METRICAS ENTRE MODELOS DE REGRESIÓN LOGÍSTICA, ARBOL DE DECISIÓN Y RANDOM FOREST CON BALANCEO E HIPERPARAMETROS.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Modelo** | **Acurrancy** | **Precision** | **Recall** | **F-Scrore** | **AUC** |
|
| **Regresión Logística** | 75% | 78% | 69% | 71% | 75% |
| **Árbol de Decisión** | 76% | 79% | 72% | 74% | 76% |
| **Random Forest** | 84% | 82% | 78% | 79% | 84% |

**Tabla 17.** Comparación Métricas de Modelos de Clasificación



**Figura 39**. Comparación entre modelos

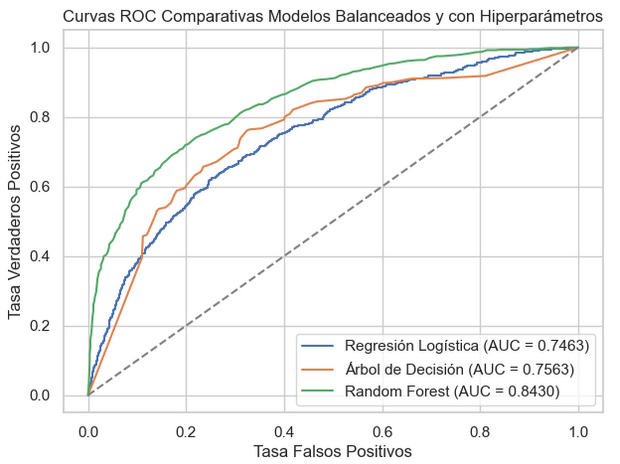
**METRICAS GENERALES**

La capacidad del modelo **Regresión Logística** para identificar a todos los clientes en riesgo de desertar es limitada con un recall del 69% y una precisión del 78%. Lo que significa que es probable que algunos clientes que efectivamente abandonarán no sean detectados, lo que puede llevar a una pérdida de oportunidades para prevenir la deserción.

El Modelo de **Árbol de Decisión** detectó correctamente el 72%(recall) de los clientes que desertaron y de estos el 79% (precisión) efectivamente lo eran, es decir el modelo tiene un buen desempeño.

El Modelo **Random Forest** identifica de forma correcta el 78%(recall) de los clientes que desertaran con una precisión del 82%, un f1-score de 79%. El valor de las métricas permite al banco actuar de manera efectiva sobre una gran cantidad de clientes en riesgo sin generar una gran cantidad de falsos positivos.

**AUC**



**Figura 40.** Curvas ROC comparativas

**Regresión Logística** con un AUC del 75% indica que tiene una capacidad razonable para distinguir entre clientes que desertan y aquellos que no. Aunque su rendimiento es aceptable, tiene una precisión moderada en la clasificación.

Con respecto al **Árbol de decisión** el AUC de 76% indica que aún hay margen para mejorar la capacidad de discriminación del modelo entre desertores y no desertores.

Por su parte **Random Forest** tiene el mejor AUC, al tener un valor de 84%. Lo que indica que este modelo es mejor para diferenciar entre los clientes que van a desertar y los que no. Sus indicadores superan a los otros dos modelos en términos de discriminación entre las clases.

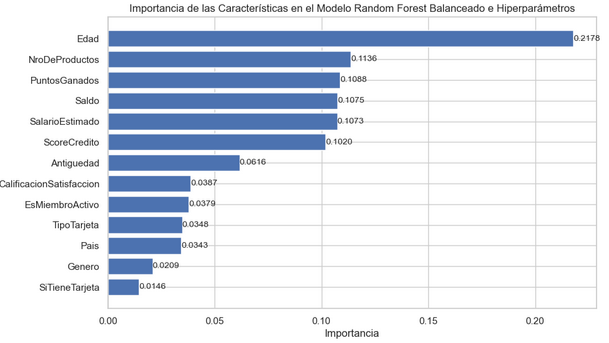
**MEJOR MODELO**

El modelo que mejor desempeño tiene es el del **Random Forest**, de acuerdo a los valores de las métricas de f1-score y recall, pero además su AUC es un poco más alto que el del **Árbol de decisión y Regresión Logística**, con lo que el modelo podrá distinguir mejor entre un desertor y un no desertor.

Random Forest con relación a la **clase Desertora** identifica correctamente el 72% (recall) de todos los clientes que efectivamente desertaron, es decir el modelo tiene un rendimiento razonable para detectar desertores. Sin embargo, de todas los clientes que el modelo predijo como desertores, **el 49%** (precisión) efectivamente lo fueron. Esto indica que la mitad de las predicciones son incorrectas.

Por lo que se sugiere la inclusión de variables relevantes como: Número de reclamos, Número de transacciones en un período específico, Estado civil, Ocupación, Canal, Número de dependencias para mejor la precisión del modelo en la predicción de la clase de estudio. (ver [**Anexo 14**.)](#A14)

## IMPORTANCIA DE LAS CARACTERISTICAS

****

**Figura 41.** Importancia de las características

En la **Figura.41** se puede visualizar cuáles son las variables más influyentes en el modelo Random Forest con Balanceo e Hiperparámetros, se detalla la contribución de cada una de ellas:

* **Edad** con un **21.78%** de importancia es la más influyente, lo que significa que entre mayor edad mayor es la probabilidad de que el cliente abandone la entidad.
* **Número de productos** del cliente es la segunda variable más importante, con una ponderación del **11.36%**. Los clientes que manejan más productos con el banco probablemente tengan menos inclinación a desertar.
* **Puntos ganados** por el cliente tienen un **10.88%** de importancia. Indica que los incentivos de lealtad juegan un rol relevante en la retención de clientes.
* **Saldo** del cliente con una importancia del **10.75%**, indica que aquellos clientes con saldos más altos podrían estar menos propensos a desertar.
* **Salario estimado** con un **10.73%** de importanciatiene un impacto moderado, es decir que los ingresos del cliente pueden influir en su decisión de permanecer con el banco.
* **Score de crédito** con una influencia del **10.20%** puede estar asociado a una mayor estabilidad financiera y menor riesgo de deserción.

Según el porcentaje de importancia las cinco variables más significativas para el modelamiento de deserción de clientes en la entidad financiera son: edad, número de productos, puntos ganados, saldo y salario estimado.

# ESTRATEGIAS ORGANIZACIONALES



## ESTRATEGIAS

Las estrategias propuestas a continuación, se las realiza en base a los factores o variables que de acuerdo a la predicción obtenida tienen mayor incidencia en los porcentajes de deserción como son: edad, número de productos y saldo.

* **Desarrollar productos y servicios personalizados de acuerdo a la edad.**

Considerando las necesidades específicas de cada grupo según su edad, enfocándose sobre todo en el grupo que más deserción presenta.

Analizar y aprovechar los datos de los clientes para ofrecerles servicios y productos personalizados es una de las estrategias clave para reducir la deserción y, en consecuencia, aumentar la rentabilidad de la entidad financiera.

Las decisiones clave que derivan de esta estrategia incluyen la redistribución de recursos hacia los grupos de mayor deserción, la inversión en tecnología para personalización basada en datos y el desarrollo continuo de productos ajustados a las necesidades específicas de cada grupo.

* **Crear o reforzar programas de retención en los clientes.**

De acuerdo al saldo promedio mensual en el que se ofrezca la acumulación de puntos y con ello beneficios a los clientes de acuerdo a su saldo.

Los clientes se pueden segmentar, diferenciándolos por el rango de saldos que mantengan en los productos de la entidad, lo que sería atractivo a clientes con altos saldos, como también incentivaría a otros clientes al ahorro, a usar otros productos financieros o aumentar el uso de servicios digitales.

Esto implica la tomar de decisiones sobre segmentación, diseño de incentivos y la asignación de recursos para minimizar la deserción y fortalecer la fidelización.

* **Incentivar a los clientes a adquirir productos.**

Mediante campañas promocionales en los que se ofrezca descuentos, recompensas por la adquisición de paquetes de productos y servicios.  Aspectos claves como el análisis del comportamiento del cliente, la identificación de productos o servicios complementarios y el enfoque de comunicación personalizada permita maximizar el valor de cada cliente al ofrecerles productos adicionales que sean relevantes y beneficiosos para ellos.

Esto requiere personalizar la comunicación y adaptar las promociones, descuentos, recompensas para que sean relevantes y atractivas para cada segmento de clientes. Decisiones clave, incluyen la selección de productos para los paquetes, el diseño de campañas promocionales, y la asignación de recursos para asegurar que la comunicación llegue a los clientes adecuados, lo que ayudará a incrementar la venta cruzada, la satisfacción del cliente y la retención.

* **Implementar o reforzar encuestas de satisfacción y recomendación.**

En los diversos canales disponibles de la entidad financiera con el fin de capturar y analizar la "voz del cliente", permitiendo que las decisiones estratégicas se alineen con las necesidades y expectativas de los usuarios, mejorando así los productos y servicios ofrecidos.

Al utilizar encuestas de satisfacción y recomendación en canales como: banca móvil, sitio web, atención al cliente, redes sociales, sucursales físicas, la entidad podrá obtener información directa y actualizada sobre la experiencia del cliente.

Esta retroalimentación permitirá una toma de decisiones basada en datos reales y específicos, lo que garantizará que las acciones estén alineadas con las expectativas y deseos del cliente.

## INNOVACIÓN Y COMPETITIVIDAD EMPRESARIAL

La innovación es un factor clave para que la entidad financiera enfrente los desafíos de la banca actual y logre reducir el porcentaje de abandono o deserción de clientes. Para ello, es fundamental implementar estrategias centradas en la mejora de la experiencia del cliente, la optimización de procesos mediante la digitalización, y el desarrollo de productos y servicios financieros personalizados que respondan a las necesidades cambiantes de los clientes. Estas iniciativas no solo fortalecen la lealtad de los clientes, sino que también aumentan la eficiencia operativa, permitiendo a la entidad adaptarse rápidamente a un entorno altamente competitivo y en constante evolución.

A continuación, se propone y detalla la estrategia de innovación:



**Figura 42.** Estrategia de Innovación

* **Generar productos y servicios de alto valor a través de procesos más ágiles e innovación.**

Recoger datos de las interacciones de los clientes con productos digitales y ajustarlos continuamente según las necesidades y preferencias que los datos revelen.

La generación de productos y servicios de alto valor, se deben implementar dentro de un ciclo continuo de innovación basado en las retroalimentaciones y en el análisis de los datos de comportamiento de los clientes, la agilidad en los procesos impulsa la capacidad de respuesta ante las necesidades cambiantes, esto permite a la entidad estar constantemente evolucionando y manteniéndose competitiva.

Para los rangos de edad donde los clientes tienen mayor porcentaje de deserción en este caso de estudio, se propone la generación de los siguientes productos/servicios.

Para jóvenes entre 18 a 30 años, ofrecer productos mediante aplicaciones o plataformas que les permita almacenar de manera segura información financiera, como tarjetas de crédito, débito, cuentas bancarias para realizar pagos electrónicos a través de las denominadas billeteras digitales que funcionan a través de dispositivos móviles como smartphones, tablets o computadoras, lo que elimina la necesidad de llevar dinero en efectivo o tarjetas físicas.

Para clientes cuya edad se encuentra entre 31 años y 40 años, se sugiere ofrecer paquetes de productos como: hipotecas, préstamos personales y asesoría para la planificación financiera ya que son personas en una etapa de crecimiento profesional y familiar, con mayores ingresos y necesidades financieras más complejas.

Para clientes con edad entre 41 y 50 años quienes suelen encontrarse en una etapa de vida estable con responsabilidades familiares y necesidades de ahorro a largo plazo. Se debe generar productos personalizados como planes de ahorro para la jubilación, seguros y opciones de inversión que se alineen con sus objetivos financieros a largo plazo.

Para clientes entre 51 y 60 años se debe diseñar productos para la **tranquilidad financiera**, como planes de ahorro y seguros que minimicen riesgos, mientras se asegura un acceso fácil a información clara y accesible sobre sus finanzas. Proporcionando además asesoría en planificación de jubilación, opciones de inversión de bajo riesgo y seguros de salud complementarios.

* **Mejora continua en la experiencia digital del cliente.**

En un entorno digital en constante evolución, los clientes demandan rapidez, eficiencia e inmediatez en sus interacciones con la entidad financiera. Para sobresalir frente a la competencia, es fundamental que la entidad ofrezca una experiencia digital superior que no solo responda a estas necesidades, sino que también anticipe sus expectativas.

Optimizar la plataforma de banca móvil es clave para proporcionar una experiencia diferenciada. Al ofrecer servicios ágiles, como pagos rápidos, acceso fácil a inversiones y soporte en tiempo real a través de asistentes virtuales (chatbots), la entidad no solo facilitará la vida de los clientes, sino que también incrementará la frecuencia de interacciones. Esta rapidez y eficiencia se traducen en una mayor satisfacción del cliente, lo que fortalece la relación y fomenta la lealtad.

* **Excelencia operacional y digital.**

Mejorar los canales de comunicación y atención al cliente según las preferencias generacionales.

Para clientes jóvenes que evitan generalmente el contacto físico, ofrecer una experiencia digital rápida e intuitiva a través de una aplicación que permita desde realizar pagos hasta recibir asesoramiento financiero virtual, diseñada para atraer a la generación de millennials y generación Z.

Para clientes adultos, que prefieren atención a través de sucursales físicas o llamadas telefónicas, darles una atención personalizada donde se le ofrezca servicios VIP para gestionar inversiones o pensiones.

La combinación de innovación tecnológica con agilidad operativa, no solo mejora la experiencia digital del cliente, sino que también tiene un impacto directo en la retención. Al ofrecer una plataforma intuitiva, rápida, personalizada, segura e integrada, la entidad financiera incrementa la frecuencia de uso y, por tanto, fortalece su relación con el cliente, reduciendo la probabilidad de deserción.

Estas estrategias de innovación se proponen en base a análisis de casos similares realizadas en otras entidades bancarias, cuyo detalle se encuentra en la sección Anexos. (ver [**Anexo 13**](#A13).)

# CONCLUSIONES Y RECOMENDACIONES



## CONCLUSIONES

* El modelo predictivo Random Forest demostró ser el más eficiente en el análisis, alcanzando un AUC del 84%. Este buen rendimiento permite anticipar el comportamiento de los clientes desertores y no desertores., proporcionando una herramienta valiosa para optimizar las estrategias de retención y toma de decisiones en la entidad.
* El modelo Random Forest aplicado para predecir la deserción de clientes en la entidad financiera mostró un rendimiento razonable en la detección de la clase desertora, con un recall del 72%. Sin embargo, la precisión fue baja, alcanzando solo un 49%, lo que indica que la mitad de las predicciones de deserción fueron incorrectas. Debido a esta baja precisión en la clase desertora, se concluye que es necesario incorporar variables adicionales relevantes para mejorar la efectividad general del modelo y su capacidad predictiva.
* El análisis realizado permitió identificar las cinco variables que más influyen en la deserción de clientes, tales como la edad, número de productos, saldo, puntos ganados y salario estimado, las cuales abarcan el 66% del comportamiento de deserción, de éstas la edad es la que mayor porcentaje posee con un 21.78%. Estos hallazgos posibilitan una segmentación más precisa, lo que facilitará la implementación de estrategias personalizadas para maximizar la efectividad de las acciones para disminuir el porcentaje deserción o abandono de clientes. Al enfocarse en estas variables, se puede reducir los costos operativos asociados a campañas masivas menos efectivas, optimizando así los recursos destinados a la retención de clientes.
* La implementación del modelo Random Survival Forest (RSF) no solo permitió identificar a los clientes en riesgo de desertar, sino también estimar el tiempo hasta que ocurra la deserción. Esta capacidad predictiva proporciona una ventaja competitiva significativa, al permitir una anticipación más precisa y una planificación más efectiva de las estrategias de retención, optimizando los recursos y aumentando la probabilidad de éxito en la retención de clientes.

## RECOMENDACIONES

* **Mejorar la precisión del modelo mediante la inclusión de variables adicionales relevantes.** Se recomienda incluir variables como número de reclamos, número de transacciones en un período específico, estado civil, ocupación, número de dependencias y canal de atención. Esto ayudará a mejorar la precisión y capacidad de discriminación del modelo en la predicción de desertores.

Esta recomendación se propone de acuerdo a análisis de casos similares cuyo detalle se encuentra en la sección Anexos. (ver [**Anexo 14**](#A14).)

* **Personalizar productos según las características del cliente, en base a** la segmentación por edad, número de productos y saldo para desarrollar productos y servicios financieros personalizados, enfocándose especialmente en los grupos con mayor tendencia a desertar. Esto permitirá ofrecer soluciones más atractivas y relevantes para cada segmento.
* **Fortalecer programas de retención y venta cruzada, con la implementación de** estrategias propuestas para incentivar a los clientes a adquirir múltiples productos a través de campañas con descuentos y promociones personalizadas, ofreciendo recompensas por el uso de productos financieros, especialmente para clientes con altos saldos.
* **Optimizar la experiencia digital, mejorando** la experiencia digital de los clientes, particularmente en banca móvil, mediante una plataforma ágil, eficiente y personalizada. Ofrecer servicios rápidos y soporte en tiempo real con asistentes virtuales, adaptando la atención a las preferencias de los diferentes grupos generacionales.
* Actualizar y perfeccionar regularmente el modelo de análisis, ya que los comportamientos de los clientes y las condiciones del mercado cambian con el tiempo, por lo que es fundamental actualizar y mejorar continuamente el modelo predictivo. Esto garantizará que siga siendo relevante, preciso y eficaz en la identificación de clientes en riesgo de deserción.

# REFERENCIAS

Avery, J. (Septiembre de 2023). *Harvard Business School*. Obtenido de https://www.hbs.edu/faculty/Pages/item.aspx?num=53301

Ávila, J. (Junio de 2022). *Estrategia de modelado de productos y servicios financieros para la Cooperativa de Ahorro y Crédito Jardín Azuayo*. Obtenido de https://dialnet.unirioja.es/servlet/articulo?codigo=8638016

Barrueta Meza, R. A. (2018). *Modelo de análisis predictivo para determinar clientes con tendencia a la deserción en bancos peruanos.* Obtenido de https://repositorioacademico.upc.edu.pe/handle/10757/626023

Bohórquez, M. T. (2019). *Modelos de Predicción de Deserción de Clientes para una Administradora de Fondos Ecuatoriana*. Obtenido de https://www.dspace.espol.edu.ec/bitstream/123456789/51760/1/T-110027.pdf

Carolina, M. (2018). *Estrategias comerciales para evitar la fuga de clientes del Banco Bilbao Vizcaya Argentaria BBVA de la sucursal Ocaña*. Obtenido de https://www.fesc.edu.co/Revistas/OJS/index.php/mundofesc/article/view/291/453

Centeno Arízaga, J. E. (2020). *Deserción de clientes en el sector asegurador. Evidencia mediante el análisis de modelos de Machine Learning.* Obtenido de https://www.dspace.espol.edu.ec/handle/123456789/53316

Chongqi Wu, L. W. (Diciembre de 2022). *A Comparative Analysis of Churn Prediction Models: A Case Study in Bank Credit Card*. Obtenido de A Comparative Analysis of Churn Prediction Models: A Case Study in Bank Credit Card

Contreras E., F. F. (2017). *Diseño de un modelo predictivo de fuga de clientes utilizando árboles de decisión*. Obtenido de https://revistas.ubiobio.cl/index.php/RI/article/view/3055/3075

Dirección de Marketing. (2012). En P. &. Kotler, *Dirección de Marketing.* México: Pearson.

Fajardo Cruz, D. F. (2019). *Modelo de abandono de clientes en una empresa de créditos en línea*. Obtenido de https://repositorio.uniandes.edu.co/entities/publication/0233f1e9-4a2b-41ec-b1d6-4b27ed508a90

Gonzáles, V. (2024). *Introducción a los Árboles de Decisión y Bosques Aleatorios (Random Forest).* Presentación multimedia- UDLA.

Gonzalez, L. (20 de 12 de 2019). *AprendeIA*. Obtenido de AprendeIA: https://aprendeia.com/ventajas-y-desventajas-de-los-algoritmos-de-clasificacion-machine-learning/

Jennifer, M. (Octubre de 2022). *Estrategias de fidelización de clientes aplicadas a una institución financier en la ciudad de Machala*. Obtenido de https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://dialnet.unirioja.es/descarga/articulo/9018732.pdf&ved=2ahUKEwiKiejxk4yJAxVZQjABHXE8CKAQFnoECBUQAQ&usg=AOvVaw3zOKQ5Tq-Cit7KEVOOu9UB

Johnson, M. K. (2018). *Applied Predictive Modeling.* Groton: Springer.

Kaggle. (2023). *Kaggle*. Obtenido de https://www.kaggle.com/code/shruthiiiee/bank-customer-churn-prediction

Kozak, B. &. (2010). *International Conference on Computational Collective Intelligence*. Obtenido de International Conference on Computational Collective Intelligence: https://link.springer.com/chapter/10.1007/978-3-642-16693-8\_39

Kuhn, M. &. (2018). Applied Predictive Modeling. Springer.

Lemos, R. A. (06 de 03 de 2022). *Propension to customer churn in a financial institution: a machine learning approach*. Obtenido de Propension to customer churn in a financial institution: a machine learning approach: https://link.springer.com/article/10.1007/s00521-022-07067-x

Mayor Lupo, J. (2020). *Aplicación de modelos predictivos para cálculo de probabilidad de churn en importante entidad bancaria*. Obtenido de https://repositorio.utdt.edu/handle/20.500.13098/12104

Miguel, M. Q. (2022). Métodos de reducción de la dimensionalidad: ACP vs t-SNE. Zaragoza.

Moreno, C. A. (2023). *Github*. Obtenido de https://github.com/marsgr6/analitica-online/blob/main/papers\_capstone/Teleco\_churn.pdf

Murphy, E. C. (2002). Leading on the Edge of Chaos. En E. C. Murphy, *Leading on the Edge of Chaos.* Prentice Hall.

Pamela, L. (Octubre de 2023). *BBVA*. Obtenido de https://www.bbva.com/es/pe/salud-financiera/cuales-son-los-productos-financieros-dirigidos-a-jovenes-en-peru/

Prado A. & Yambay M. (Octubre de 2024). *Aplicación de Modelos de Predicción para identificar patrones de deserción en los datos de una Entidad Financiera*. Obtenido de https://github.com/AnaCristinaPrado/CapstoneUDLA/blob/main/Proyecto\_Caso%20Estudio%20Deserci%C3%B3n\_AnaPrado\_MargaritaYambay.ipynb

Pulgarín, R. S. (2020). *Modelo de pronóstico de deserción de los tarjetahabientes al adquirir una tarjeta de crédito y ser cancelada a los 90 días.* Obtenido de https://repository.libertadores.edu.co/server/api/core/bitstreams/73b6fb47-134c-499a-b07d-f8edcf99d686/content

Rodríguez Herrera, V. A. (2013). *Estimación del customer lifetime value a nivel de clientes, de un banco usando variables transaccionales y sociodemográficas.* Obtenido de https://repositorio.uchile.cl/handle/2250/113830

Santander, B. (Junio de 2024). *¿Cómo mide Santander la satisfacción de los clientes?* Obtenido de https://www.santander.com/es/stories/think-customer-nps

Tatamues, F. (18 de 10 de 2021). *Elaboración de un modelo predictivo de deserción de clientes de tarjeta de crédito para una entidad bancaria del Ecuador* . Obtenido de Elaboración de un modelo predictivo de deserción de clientes de tarjeta de crédito para una entidad bancaria del Ecuador : chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://repositorio.uasb.edu.ec/bitstream/10644/8413/1/T3673-MGFARF-Martinez-Elaboracion.pdf

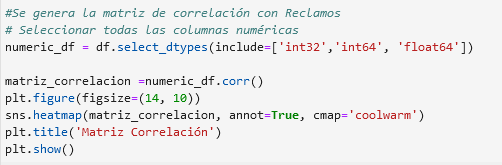
Verma, P. (2020). *Churn Prediction for Savings Bank Customers: A Machine Learning Approach*. Obtenido de https://digitalcommons.aaru.edu.jo/cgi/viewcontent.cgi?article=1314&context=jsap

Zhang, B. (01 de 01 de 2023). *Predictive Analytics on Customer Churn* . Obtenido de Predictive Analytics on Customer Churn : chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://typeset.io/pdf/customer-churn-in-subscription-business-model-predictive-q4qud3zz.pdf

Zhining, L. (2021). *imbalanced-ensemble*. Obtenido de https://imbalanced-ensemble.readthedocs.io/en/latest/api/sampler/\_autosummary/imbens.sampler.SMOTE.html

# ANEXOS

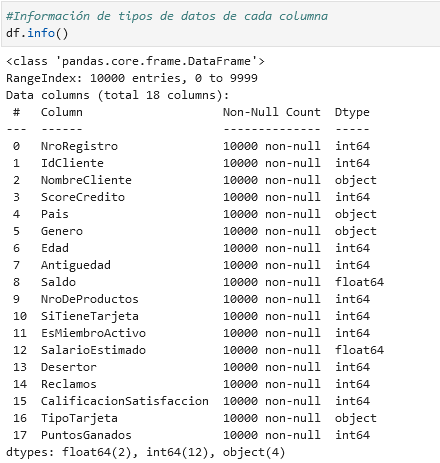
**Anexo 1.** Código Matriz de Correlación



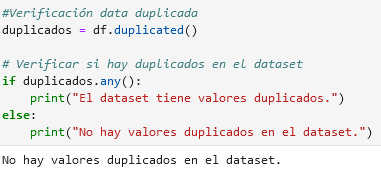
**Anexo 2.** Código Análisis Descriptivo

****

**Anexo 3.** Código Revisión de variables

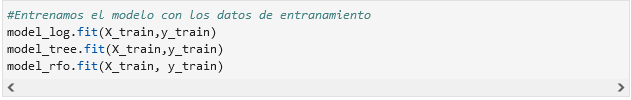
****

****

****

**Anexo 4.** Código Modelamiento entrenamiento train y test.





**Anexo 5.** Código Modelo de Regresión Logística



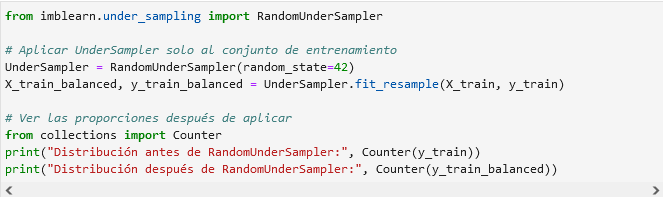
**Anexo 6.** Código Modelo Árbol de Decisión



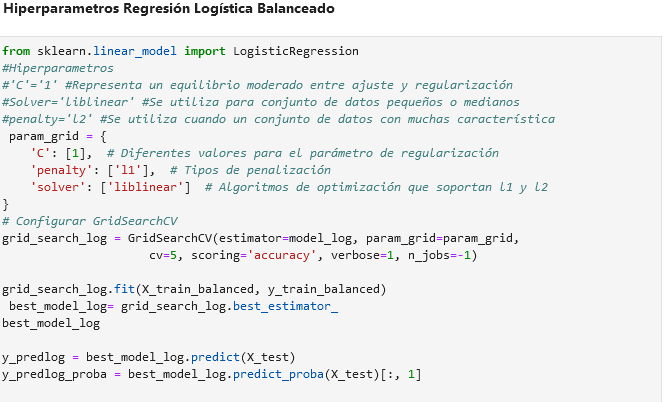
**Anexo 7.** Código Modelo Random Forest

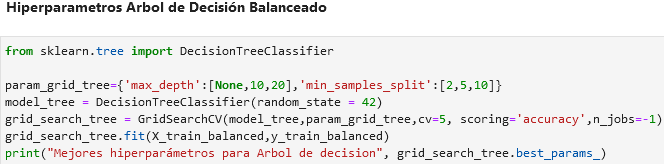


**Anexo 8.** Código Técnica de Balanceo RUS



**Anexo 9.** Código Mejores Hiperparámetros



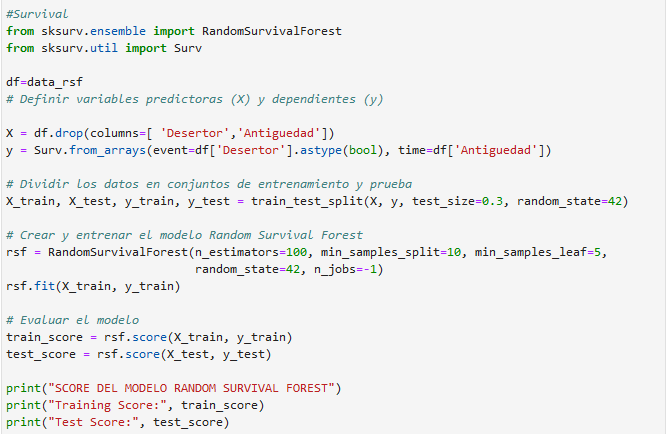




**Anexo 10.** Código T-SNE



**Anexo 11.** Código Modelo Random Survival Forest



**Anexo 12.** Supervivencia por Cliente



**Anexo 13.** Casos de estudio referenciados en Estrategias e Innovación

|  |  |  |  |
| --- | --- | --- | --- |
| **Objetivo de la revisión literaria** | **Autor** | **Fuente** | **Resultado** |
| Estrategia del desarrollo de productos personalizados (de acuerdo a la edad) | López Pamela | https://www.bbva.com/es/pe/salud-financiera/cuales-son-los-productos-financieros-dirigidos-a-jovenes-en-peru/ | Un estudio de la Asociación de Bancos del Perú (Asbanc) y CPI revela que el 47.3% de los jóvenes entre 15 y 25 años considera difícil obtener préstamos de instituciones financieras, y el 30% señala que no cumple los requisitos. Esto representa una oportunidad para que las entidades financieras desarrollen productos adaptados a este segmento, impulsando la inclusión financiera desde una edad temprana.  Las cuentas de ahorro son el primer paso hacia la independencia financiera, ofreciendo beneficios como ausencia de comisiones y acceso a otros productos bancarios. Además, las billeteras digitales han ganado popularidad entre los jóvenes, representando el 52% de las transacciones digitales en el primer trimestre de 2023. Estas billeteras son la "puerta de entrada" a más servicios bancarios, ayudando a crear historial crediticio.  Otro de los productos que BBVA ofrece a sus clientes de hasta 40 años es su producto ”Mi primer auto”, es un crédito vehicular sin entrada, que permite adelantar pagos y que puede vencer hasta en 6 años. (Pamela, 2023) |
| Estrategia de incentivo a los clientes a adquirir productos | Avery Jill | https://www.hbs.edu/faculty/Pages/item.aspx?num=53301 | El caso de referencia es el de Chase Bank el cual implementó exitosos programas de lealtad, particularmente a través de sus tarjetas de crédito Chase Sapphire, para retener y reducir la pérdida de clientes. Un ejemplo clave es la tarjeta de crédito Chase Sapphire Reserve, que atrajo a los clientes con un bono inicial de 100,000 puntos y desarrolló lealtad a través de recompensas específicas, esto captó sobre todo a clientes millennials. Al centrarse en ofertas de productos, como la combinación de tarjetas de crédito con beneficios de viajes y restaurantes, Chase creó un sistema de recompensas flexible que mantenía a los clientes comprometidos. Esto ha sido particularmente efectivo para retener a los clientes. También se centraron en el resto de la cartera de productos Chase Sapphire para evaluar la diferenciación entre los productos e identificar espacios en el mercado que podrían respaldar lanzamientos adicionales de nuevos productos. (Avery, 2023) |
| Estrategia de implementar o reforzar encuestas de satisfacción y recomendación. | Macas Jennifer | https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://dialnet.unirioja.es/descarga/articulo/9018732.pdf&ved=2ahUKEwiKiejxk4yJAxVZQjABHXE8CKAQFnoECBUQAQ&usg=AOvVaw3zOKQ5Tq-Cit7KEVOOu9UB | La estrategia realizada por una entidad bancaria de la ciudad de Machala se basa en fidelizar a los clientes activos mediante ventas cruzadas e incentivar la captación de nuevos clientes a través de la recomendación de familiares o amistades de los clientes. Esta estrategia no sólo retiene a los clientes actuales, sino que también permite ampliar la base de clientes. Los resultados muestran que esta estrategia permite generar otro método de retención del cliente y por otra parte un aumento en la cartera de clientes. (Jennifer, 2022) |
| Aplicación de estrategia de generación de productos y servicios a través de proceso más agiles. | Ávila José | https://dialnet.unirioja.es/servlet/articulo?codigo=8638016 | El caso de estudio tuvo como objetivo analizar la estrategia de modelado de productos y servicios financieros en la Cooperativa Jardín Azuayo cuyos objetivos eran el de agregar valor a la producción de servicio, reducir tiempos al incorporar métodos ágiles de desarrollo, apoyar la multicanalidad y fidelizar a su cliente.  Esta estrategia se basó en una metodología estructurada en tres fases: Fundamentar, Ingeniar e Integrar.  En la fase de Fundamentar, se utiliza el Modelo Canvas para construir la viabilidad del servicio financiero, abordando aspectos como segmentos de mercado, propuestas de valor, relaciones con clientes, y estructura de costos.  En la fase de Ingeniar, se emplean las metodologías de Design Thinking y Lean Startup para desarrollar productos centrados en las necesidades del usuario, reduciendo los ciclos de desarrollo y validando su viabilidad.  Finalmente, la fase de Integrar utiliza la metodología ágil Scrum para la implementación del producto, asegurando su compatibilidad con la infraestructura existente y mejorando la eficiencia mediante la entrega continua de valor.  Esta estrategia, centrada en el cliente y en procesos ágiles, les permitió generar productos financieros que satisfacen las demandas del mercado de forma eficiente y rápida. (Ávila, 2022) |
| Mejoramiento continuo en la experiencia digital del cliente | Banco Santander | <https://www.santander.com/es/stories/think-customer-nps> | De acuerdo a estudios realizado en el Banco Santander (cede principal España) identificaron que es lo que querían los clientes llegando a resumirlas en estas tres necesidades principales:  -Mejores productos y servicios en todos los canales a precios competitivos.  -Experiencia digital completa y fluida.  -Un socio global confiable y receptivo.  Para responder a las expectativas de sus clientes, implementaron estrategias que incluyen: la optimización de su cartera de productos y la mejora de la experiencia del cliente, ofreciendo soluciones competitivas en todos los canales. Para satisfacer la demanda de una experiencia digital fluida, ha desarrollado plataformas tecnológicas avanzadas. Además, ha automatizado y simplificado sus procesos para reducir costos y mejorar la eficiencia, consolidándose como un socio global confiable y receptivo. Estas acciones buscan mejorar la competitividad y satisfacer las necesidades de los clientes de manera efectiva. (Santander, 2024) |
| Excelencia operacional y digital | Marulanda Carolina | https://www.fesc.edu.co/Revistas/OJS/index.php/mundofesc/article/view/291/453 | La estrategia del banco BBVA se centra en capacitar a sus clientes para utilizar portales virtuales, facilitando la realización de transacciones como pagos, transferencias y servicios públicos de manera rápida y sencilla, sin necesidad de acudir a las sucursales. Esta iniciativa busca optimizar el tiempo de los clientes, evitando filas, y aumentar su satisfacción con el servicio digital ofrecido. (Carolina, 2018) |

**Anexo 14.** Casos de estudio de deserción y sus variables relevantes

|  |  |  |  |
| --- | --- | --- | --- |
| **Objetivo de la revisión literaria** | **Autor** | **Fuente** | **Resultado** |
| Uso de variable **número de reclamos** en el modelo del caso de estudio | Mayor Lupo Juana | <https://repositorio.utdt.edu/handle/20.500.13098/12104> | El estudio realizado por Juana Mayor permite identificar la tasa de deserción de clientes de una importante entidad financiera de Argentina, a través de la selección del mejor modelo entre tres modelos predictivos: Modelo Bagging, Random Forest, y Benchmark, se consideró como insumo para este estudio los datos de los clientes a diciembre del 2021. Luego de una pulida minería de datos se incluyó en el grupo de las variables predictoras el número de reclamos, con la probabilidad que un cliente al realizar uno o más reclamos sobre algún producto es probable que al largo o corto plazo pueda dejar la entidad. Como resultado se determinó que el mejor modelo que predice el abandono de los clientes del banco es Random Forest con elección de hiperparámetros a través de la técnica OOB (Out-Of-Bag). (Mayor Lupo, 2020) |
| Relevancia de las variables **ocupación y número de transacciones** en el estudio de la deserción de clientes. | Verma Prashant | <https://digitalcommons.aaru.edu.jo/cgi/viewcontent.cgi?article=1314&context=jsap> | En el artículo investigado que trata de la predicción en la pérdida de clientes de cuentas de ahorro de los bancos de la India. A través de modelos estadísticos como RANDOM FOREST, XG-BOOST y utilizando la precisión, el área bajo la curva (AUC), el coeficiente de Gini, la curva de características operativas del receptor (ROC), los resultados muestran que Random Forest predice la pérdida de clientes con una precisión del 78%, es el modelo más potente para este escenario. Siendo la antigüedad del cliente, la edad del cliente, el saldo medio, el código de ocupación, el tipo de población, el importe medio de débito y el número medio de transacciones las variables con mayor poder predictivo para el modelo de predicción de deserción de clientes. .(Verma, 2020) |
| Impacto de la variable **canal** en el modelo de abandono de clientes en una empresa de créditos en línea | Fajardo Cruz | <https://repositorio.uniandes.edu.co/entities/publication/0233f1e9-4a2b-41ec-b1d6-4b27ed508a90> | Incluir la variable Canal (es decir, el medio por el cual un cliente interactúa con la entidad financiera, como online, en persona, telefónico, etc.) en un análisis de deserción de clientes, puede proporcionar un nivel de detalle importante para mejorar la precisión y la profundidad del análisis. Es así como en el caso de estudio de créditos en línea de una entidad financiera de Colombia cuyo objetivo fue el de crear un modelo de deserción a corto plazo en el que aprovecharon los datos en tiempo real de los distintos canales para identificar a los clientes en alto riesgo y permitir acciones preventivas inmediatas. (Fajardo Cruz, 2019) |
| Importancia de la variable **estado civil** en la predicción de la duración de un cliente bancario | Rodríguez Herrera | <https://repositorio.uchile.cl/handle/2250/113830> | En un estudio realizado en Chile para estimar el valor del ciclo de vida del cliente a través de variables transaccionales y sociodemográficas, se identificó que, además de la edad y el género, factores como la ocupación y el estado civil son variables sociodemográficas relevantes en la predicción de la duración de un cliente bancario. Mediante pruebas de significancia y el cálculo de coeficientes de correlación de Pearson, se redujeron las variables del análisis a las más representativas siendo las variables sociodemográficas más significativas las siguientes: edad, género, ocupación y estado civil. (Rodríguez Herrera, 2013) |
| Relevancia de la variable **número de dependientes** en el modelo de predicción de abandono | Chongqi Wu | <https://www.csupom.com/uploads/1/1/4/8/114895679/n20p7formatted.pdf> | Adicionalmente, un estudio de la Universidad de California que incluyó la variable número de dependientes, concluyó que esta tiene un impacto significativo en los modelos de predicción de abandono. Al calcular la desviación estándar, se observó que esta variable superaba la media, lo que incrementaba la probabilidad de abandono en el modelo, alcanzando un 18,76% de abandono. (Chongqi Wu, 2022) |