

Universidad Internacional San Isidro Labrador

Curso: Data Science

Tema:

Modelo predictivo acerca de la deserción de clientes

(Bank Customer Churn Prediction)

PROYECTO I

Profesor: Samuel Saldaña Valenzuela

Estudiante:

Melanie Joana Moreira Sánchez

402450025

Noviembre, 2024

Contenido

Introducción.....	3
Objetivo General	4
Objetivo Específicos	4
Marco Teórico	5
Desarrollo Teórico:.....	6
• Fase Conceptual del Modelo CRISP-DM (Cross Industry Standard Process for Data Mining).....	6
1. Comprensión del Negocio.....	6
2. Comprensión de los Datos	6
3. Preparación de los Datos.....	7
4. Análisis Exploratorio de Datos (EDA).....	7
5. Preparación para el Modelado Predictivo	7
6. Importancia del EDA	8
• Técnicas de Data Mining Aplicadas.....	8
1. Minería de reglas de asociación:.....	8
2. Clasificación:.....	8
3. Agrupación en clústeres (Clustering).....	9
4. Secuencias y Trayectorias – Análisis.....	9
• Data Mining	9
1. Minería de procesos.....	9
2. Minería de textos	9
3. Minería predictiva.....	10
• Análisis Exploratorio de Datos / Exploratory Data Analysis (EDA)	11
• Código:.....	12
• Conclusiones y recomendaciones.....	20
Conclusiones:.....	20
Recomendaciones:.....	21

Introducción

La deserción de clientes en el sector bancario representa un desafío significativo que impacta directamente en la rentabilidad y la sostenibilidad de las instituciones financieras. En un entorno cada vez más competitivo, la capacidad de predecir y prevenir la pérdida de clientes se ha convertido en una prioridad estratégica. Este proyecto aborda esta problemática desde la perspectiva del Data Science, aplicando un conjunto de técnicas y herramientas analíticas para desarrollar un modelo predictivo de deserción.

Por consiguiente y con el objetivo de abordar esta problemática, el presente proyecto se centra en el análisis de datos provenientes de los clientes del banco, según el Dataset brindado. Este proceso incluye la depuración de inconsistencias en cuanto a las variables y de ahí la transformación necesaria para preparar la información de cara a un modelo predictivo.

La creciente disponibilidad de grandes volúmenes de datos, junto con el avance de las tecnologías de la información, ha impulsado el desarrollo de soluciones basadas en datos para resolver problemas complejos en diversos ámbitos. En el caso del sector bancario, el Data Science ofrece una oportunidad única para aprovechar la información histórica de los clientes y construir modelos predictivos capaces de identificar patrones y tendencias que permitan anticipar la probabilidad de deserción.

En última instancia, este proyecto no solo busca contribuir a una mejor toma de decisiones en la planificación estratégica del banco, sino también fortalecer su capacidad de respuesta ante los desafíos del mercado, ofreciendo soluciones prácticas para mejorar la experiencia del cliente y consolidar relaciones a largo plazo.

Objetivo General

Preparar un dataset sobre la salida de clientes en una institución bancaria, llevando a cabo un análisis exploratorio detallado y un proceso de limpieza riguroso para asegurar que los datos estén en condiciones óptimas para entrenar un modelo de machine learning.

El resultado esperado de esta etapa será un conjunto de datos limpio y optimizado, adecuado para entrenar modelos de machine learning que permitan predecir la probabilidad de deserción de clientes. Este enfoque no solo garantizará la calidad de los datos, sino que también sienta las bases para desarrollar un modelo predictivo robusto en las etapas posteriores del proyecto.

Objetivo Específicos

1. Recolectar y preparar un conjunto de datos representativo, para ello se recopilarán datos históricos de clientes bancarios, incluyendo información demográfica y de comportamiento, para construir una base de datos sólida y confiable.
2. Identificar patrones y características clave que expliquen el comportamiento de deserción, brindando herramientas para diseñar estrategias que fomenten la retención.
3. Desarrollar y evaluar modelos predictivos, además de construir y evaluar diversos modelos de clasificación, como regresión logística, para así seleccionar el mejor modelo, aquel que brinde el mejor rendimiento en términos de precisión y capacidad de generalización y con ello, interpretar los resultados del modelo para identificar los factores más importantes asociados con la deserción de clientes y generar recomendaciones para la toma de decisiones.

Marco Teórico

La deserción de clientes (“churn”), es uno de los principales desafíos para las instituciones bancarias, ya que afecta directamente la estabilidad financiera y la competitividad. Este fenómeno se resume en la pérdida de ingresos y obliga a las organizaciones a invertir más recursos para captar nuevos clientes, lo cual es significativamente más costoso que retener a los actuales. Según diversos estudios, la alta competencia y la amplia disponibilidad de opciones financieras han incrementado la complejidad de este problema, haciendo imprescindible que los bancos desarrollen estrategias proactivas para entender y prevenir la deserción de clientes.

En nuestro país, el sector bancario desde hace algunos años atrás decidió apoyarse de las distintas herramientas tecnológicas de análisis predictivo para no solo identificar patrones de deserción, sino también para tomar decisiones informadas sobre cómo retener a los clientes y mejorar la satisfacción del cliente.

La aplicación de EDA, técnicas de minería de datos y el uso de CRISP-DM permite garantizar un enfoque riguroso y orientado a resultados. Este marco teórico respalda el análisis y preparación del conjunto de datos, asegurando que los resultados obtenidos sean confiables y útiles para construir un modelo predictivo de deserción de clientes. Esto no solo facilita la identificación de patrones relevantes, sino que también contribuye al diseño de estrategias efectivas para mejorar la retención y la experiencia del usuario.

Predecir la deserción de clientes permite a los bancos identificar tempranamente a aquellos clientes con mayor probabilidad de abandonar sus servicios. Esta previsión facilita la implementación de estrategias de retención personalizadas, ayuda a optimizar los recursos destinados a la atención al cliente y mejora las iniciativas de marketing y fidelización.

El modelo CRISP-DM proporciona consigo una estructura ampliamente aceptada para abordar proyectos de la minería de datos, abordando desde la comprensión del problema hasta la implementación de soluciones. En el contexto de la deserción de clientes bancarios, este enfoque facilita un análisis riguroso y eficiente, asegurando que cada etapa del proceso esté alineada con los objetivos comerciales. Para abordar lo anterior tenemos y requerimos de las fases explicadas y desarrolladas a continuación:

Desarrollo Teórico:

- **Fase Conceptual del Modelo CRISP-DM (Cross Industry Standard Process for Data Mining)**

1. Comprensión del Negocio

La primera fase de CRISP-DM se centra en identificar los objetivos específicos del negocio y definir las metas analíticas del proyecto. En el caso de los bancos, el problema de la deserción de clientes implica entender las razones detrás de esta conducta para anticiparla y/o prevenirla. Esto incluye explorar factores como la frecuencia de uso de servicios, el nivel de satisfacción, y los patrones de interacción con el banco.

La deserción representa una pérdida financiera directa y, además, el costo de adquisición de nuevos clientes suele ser significativamente más alto que el costo de retención, tal y como se menciona anteriormente, es por ello que identificar los patrones y comportamientos previos a la deserción permitirá diseñar programas de fidelización más efectivos.

2. Comprensión de los Datos

Como bien se sabe, la comprensión de datos es una fase fundamental en cualquier proyecto de Data Science y consiste en el análisis inicial y detallado de los datos con los que vamos a trabajar; En palabras sencillas es como conocer a fondo a tu material antes de comenzar a construir algo con él.

En esta etapa, se recopilan y analizan los datos para identificar patrones, relaciones y posibles problemas que puedan afectar su uso. El Análisis Exploratorio de Datos (EDA) es una herramienta central que permite:

- Identificar variables.
- Detectar datos faltantes o valores atípicos.
- Evaluar correlaciones entre características relevantes, como ingresos, historial transaccional y tipo de producto adquirido.

3. Preparación de los Datos

La preparación de los datos es una etapa fundamental que garantiza que la información esté lista para el modelado. Este proceso implica:

- **Limpieza de datos:** Eliminación de valores nulos, duplicados o inconsistentes.
- **Transformación de variables:** Codificación de datos categóricos, escalamiento de variables numéricas y normalización para mejorar el rendimiento del modelo.
- **Manejo de valores atípicos:** Uso de técnicas estadísticas para mitigar su impacto.

Dentro de lo que se espera obtener en dicha fase es un conjunto de datos limpio, balanceado (si así lo requiere el proyecto) y adecuado para entrenar el modelo predictivo, para lograr así identificar las más relevantes y necesarias para el modelo.

4. Análisis Exploratorio de Datos (EDA)

El EDA permite una comprensión profunda del comportamiento de los datos y su relación con la deserción de clientes. Técnicas como gráficos de dispersión, histogramas y diagramas de caja ayudan a visualizar las características principales del conjunto de datos. Estas herramientas facilitan:

- Detectar patrones inesperados.
- Identificar relaciones significativas entre las variables independientes y la deserción como variable objetivo.
- Evaluar posibles sesgos o desequilibrios en los datos.

5. Preparación para el Modelado Predictivo

Tras una limpieza y transformación exhaustiva, los datos están listos para el modelado predictivo. Esto incluye:

- Selección de características relevantes mediante técnicas como correlación o reducción de dimensionalidad (PCA).

- Dividir los datos en conjuntos de entrenamiento y prueba para evaluar la eficacia del modelo.
- Validar la integridad y consistencia de los datos transformados.

6. Importancia del EDA

El Análisis Exploratorio de Datos (EDA) no solo revela patrones ocultos en el conjunto de datos, sino que también proporciona una base sólida para seleccionar las técnicas estadísticas y los modelos más adecuados.

Es decir, en palabras sencillas, no solo ayuda a entender los datos, sino que también, guía el enfoque hacia las técnicas de modelado convenientes y asegura que el análisis se realice sobre datos confiables y bien preparados.

• Técnicas de Data Mining Aplicadas

Acá encontramos conceptos tales como:

1. Minería de reglas de asociación:

Básicamente, acá se logran determinar relaciones cercanas en cuanto a las características de los usuarios que puedan desertar y aquellos que sean probables de permanecer o continuar los servicios adquiridos en la institución bancaria determinada

2. Clasificación:

Se ejecuta para identificar las posibilidades de que un cliente sea probable en cancelar su cuenta, teniendo como opción única “sí o no” en función de datos como:

- Saldo
- Edad
- Ingresos
- Puntaje crediticio y otros factores.

3. Agrupación en clústeres (Clustering)

Se define como una técnica utilizada para organizar datos en “grupos”, conocidos como “clústeres”, estas agrupaciones se ejecutan según la similitud encontrada en los datos. El objetivo principal es que los puntos dentro de un mismo clúster sean muy parecidos entre sí, mientras que los puntos en diferentes clústeres sean lo más distintos posible.

4. Secuencias y Trayectorias – Análisis

Básicamente, esta parte en vez de analizar los datos estáticos, esta técnica busca identificar patrones y tendencias en datos secuenciales

- **Data Mining**

1. Minería de procesos

En resumen, la minería de procesos es una herramienta poderosa para entender y mejorar los procesos de negocio. Al analizar los datos de los sistemas de información, podemos descubrir oportunidades para optimizar los procesos, reducir costos y mejorar la eficiencia.

2. Minería de textos

En esta etapa se suelen visualizar y/o aplicar los siguientes pasos:

- **Obtención de información:** Se recolectan textos provenientes de diferentes fuentes.
- **Preparación de los datos:** Se realiza una limpieza profunda, eliminando palabras con significado poco relevantes o bien “iguales” al formato y segmentando del texto.
- **Transformación numérica:** Los textos se convierten en estructuras matemáticas (vectores) que los algoritmos de Machine Learning (ML) pueden interpretar.
- **Procesamiento analítico:** Se utilizan técnicas como clasificación, agrupamiento, identificación de entidades, entre otras, para extraer conocimiento relevante.
- **Visualización:** Los hallazgos se visualizan de forma que su interpretación sea sencilla y práctica.

Por ejemplo, si el banco dispone de datos textuales como reseñas de clientes sobre los servicios, grabaciones de las conversaciones de soporte técnico, el análisis de texto (minería) podría detectar patrones recurrentes, como los temas más mencionados o las emociones predominantes en los comentarios.

En caso de que no haya información en formato de texto disponible en el dataset, esta metodología no sería aplicable. Sin embargo, si en el futuro se incorporaran datos de encuestas o comentarios de usuarios, se podría emplear esta técnica para comprender mejor las razones detrás de la pérdida de clientes y evaluar sus opiniones de manera más detallada.

3. Minería predictiva

En términos generales, esta parte es el enfoque principal del proyecto debido a que se basa en la predicción analítica, empleando algoritmos para estimar la probabilidad de churn de los clientes. Esto le permite al banco clasificar a su base de clientes en categorías de "alto, medio y bajo riesgo" según corresponda, ajustando sus estrategias de acuerdo con el perfil de cada grupo.

La importancia de este enfoque radica en que los modelos de clasificación, serán utilizados para construir un sistema capaz de identificar qué usuarios tienen mayor probabilidad de desertar. Este sistema proporcionará al banco una herramienta eficaz para diseñar estrategias específicas para poder retenerlos.

Con este modelo predictivo, la institución podrá anticiparse y actuar de manera proactiva sobre los clientes con mayor probabilidad de abandono. Por ejemplo, podrán implementar campañas personalizadas, ofrecer beneficios exclusivos o proporcionar un servicio más enfocado en las necesidades individuales, aumentando las posibilidades de fidelización.

Adicionalmente, aunque el foco principal del proyecto está en la predicción de deserción, otras técnicas, como el análisis de procesos o la minería de textos, pueden enriquecer el modelo. Estas herramientas permiten identificar patrones de comportamiento o temáticas recurrentes en las quejas y sugerencias, brindando una comprensión más detallada de las causas de insatisfacción.

En resumen, esta fase prioriza la predicción del riesgo de deserción, pero no se limita a ella, integrando otros enfoques para ofrecer al banco una visión más completa y estratégica de las dinámicas que influyen en la lealtad de sus clientes. Esta perspectiva integral facilita la toma de decisiones informadas, enfocadas en retener a los clientes y optimizar la relación con ellos.

- **Análisis Exploratorio de Datos / Exploratory Data Analysis (EDA)**

Su objetivo principal es comprender los datos, de modo que se logre evaluar la estructura, tamaño, características y datos generales del Dataset, con esto se logran detectar patrones iniciales, identificar las tendencias y las relaciones entre las variables, además de los datos incongruentes que están fuera de lo esperado y que por supuesto podrían afectar el análisis general.

Dentro de las técnicas comunes tenemos:

- Visualizaciones: Gráficos como histogramas, diagramas de caja (boxplots), gráficos de dispersión y mapas de calor ayudan a detectar patrones y distribuciones.
- Estadísticas descriptivas: Resumen de los datos con medidas como la media, mediana, moda, rango, desviación estándar y correlación.
- Análisis de datos faltantes: Identificación de la cantidad y ubicación de valores ausentes.
- Estudio de distribuciones: Evaluación de cómo están distribuidas las variables, verificando su normalidad u otras propiedades.

Ejemplo

“Predecir el abandono en una institución bancaria”, un análisis correctamente ejecutado podría:

- Analizar cuántos clientes desertaron en comparación con los que permanecieron.
- Estudiar la relación entre variables como edad, ingresos o número de productos contratados con la probabilidad de deserción.
- Visualizar la distribución de ingresos para identificar si hay grupos bien diferenciados.

- Detectar valores atípicos, como clientes con ingresos anormalmente altos o bajos.

En dicho proyecto tenemos que tener en cuenta que nuestro análisis se encuentra enfocado en:

Distribución de variables: Se analizará la distribución de cada variable clave, como el saldo de cuenta, antigüedad, y número de productos adquiridos.

Correlación: Las correlaciones entre variables como “saldo” y “deserción” ayudarán a identificar las más predictivas para la creación del modelo.

Transformación de datos: Variables categóricas serán transformadas en dummies, y los valores nulos serán tratados para asegurar consistencia en el modelo.

- **Código:**

Explicación:

Primeramente, damos inicio a la creación del proyecto en Google Colab, en primera instancia se importan las librerías, seguidamente se procede a la lectura del dataset (el cual fue previamente cargado a un archivo Drive nombrado “PROYECTO I”, para así facilitar la ubicación del mismo al momento de cargar los datos), esta visualización nos permite tener un contexto claro y amplio del tipo de información con el cual estamos trabajando.

```
[ ] # Importar librerías
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler

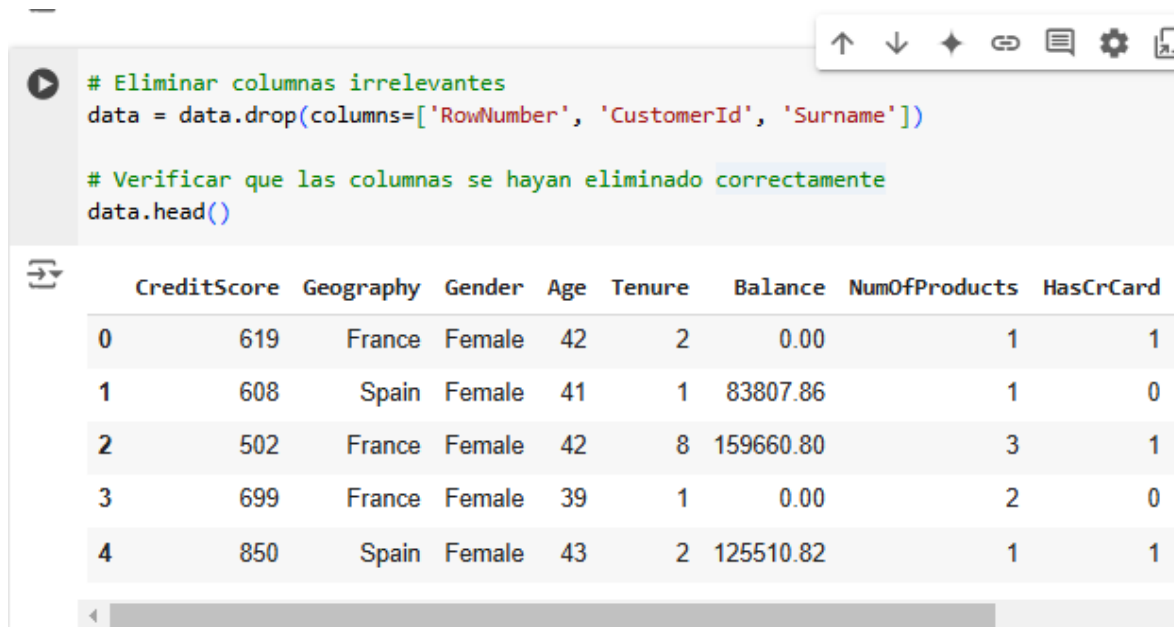
[ ] # Lectura del dataset
from google.colab import drive
drive.mount('/content/drive')

ruta = "/content/drive/MyDrive/PROYECTO I/Churn_Modelling.csv"
data = pd.read_csv(ruta)
```

Posteriormente, visualizamos los tipos de variables que tenemos, esto con el fin de observar si nuestras variables son de texto o numéricas, teniendo en cuenta que los modelos de Machine Learning suelen aceptar en su gran mayoría únicamente valores numéricos o booleanos.

Una vez realizado este paso, se procede a realizar una limpieza profunda y exhaustiva de los datos cargados previamente del Dataset y procedemos a la eliminación de las columnas que se consideran irrelevantes, por ejemplo.

- RowNumber
- CustomerID
- Surname



```
# Eliminar columnas irrelevantes
data = data.drop(columns=['RowNumber', 'CustomerId', 'Surname'])

# Verificar que las columnas se hayan eliminado correctamente
data.head()
```

	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard
0	619	France	Female	42	2	0.00	1	1
1	608	Spain	Female	41	1	83807.86	1	0
2	502	France	Female	42	8	159660.80	3	1
3	699	France	Female	39	1	0.00	2	0
4	850	Spain	Female	43	2	125510.82	1	1

Luego, se procede a la verificación de que las columnas se hayan eliminado de manera correcta, y verificando que los datos que quedan dentro son los que nos permiten entrar en contexto con el tema principal del proyecto, dentro de los datos que sí se necesitan tenemos:

- CreditScore
- Balance
- Tenure
- NumofProducts

Y la más relevante en cuestión, “Exited” entre otros...

De igual manera y a modo de verificación se procede nuevamente a verificar si existen filas duplicadas en el dataset, sin embargo, en este paso, tenemos como resultado “0”, lo que nos indica que no existe la presencia de filas iguales dentro del conjunto de datos.

```
# Verificar si hay filas duplicadas en el dataset
data.duplicated().sum()
```

⇒ 0

Seguidamente, se realiza la “conversión” de las variables categóricas en variables Dummy, como se menciona anteriormente, el modelo únicamente permite números y nos encontramos con que la variable “geography” y “gender” se encuentran en texto, de modo que, se deben convertir dichas columnas categóricas en variables dummy, que son básicamente representaciones numéricas entre “0 y 1”; Indicando, por ejemplo, las columnas a transformar:

Geography en valores como:

- Spain – France – Germany

Gender en otros tantos como:

- Male – Female

⇒	Geography_Spain	Gender_Male
0	False	False
1	True	False
2	False	False
3	False	False
4	True	False

Seguidamente se continua con el escalado de las variables, el objetivo de este código es "escalar" o "ajustar" los valores de ciertas columnas numéricas de un conjunto de datos para que todas estén en la misma escala.

Esto resulta muy útil en dicho modelo de Machine Learning, ya que algunos algoritmos funcionan mejor si los valores numéricos no varían demasiado entre sí, posterior a ello, verificar que las variables hayan sido escaladas correctamente.

```
# Definir las columnas numéricas que necesitamos escalar
numerical_columns = ['CreditScore', 'Age', 'Balance', 'NumOfProducts', 'EstimatedSalary']

# Crear el objeto scaler
scaler = StandardScaler()

# Escalar las variables numéricas
data[numerical_columns] = scaler.fit_transform(data[numerical_columns])

# Verificar que las variables hayan sido escaladas correctamente
print(data[numerical_columns].head())
```

	CreditScore	Age	Balance	NumOfProducts	EstimatedSalary
0	-0.326221	0.293517	-1.225848	-0.911583	0.021886
1	-0.440036	0.198164	0.117350	-0.911583	0.216534
2	-1.536794	0.293517	1.333053	2.527057	0.240687
3	0.501521	0.007457	-1.225848	0.807737	-0.108918
4	2.063884	0.388871	0.785728	-0.911583	-0.365276

Importante mencionar que cada vez que se ejecutaban los comandos se ponían a prueba y se comprobaban a modo de “seguro” para evitar posibles errores durante su ejecución.

Como siguiente función vamos a crear la representación gráfica de las columnas desarrolladas, se procede a la creación de gráficos tipo Boxplot para visualizar algunas de las columnas numéricas de nuestro DataSet... Así logramos comprender de una manera más sencilla como están distribuidos los datos, identificar los valores atípicos y también analizar las tendencias

En primera instancia, definimos una lista denominada “numeric_columns” con los nombres de las columnas que se graficarán, es decir:

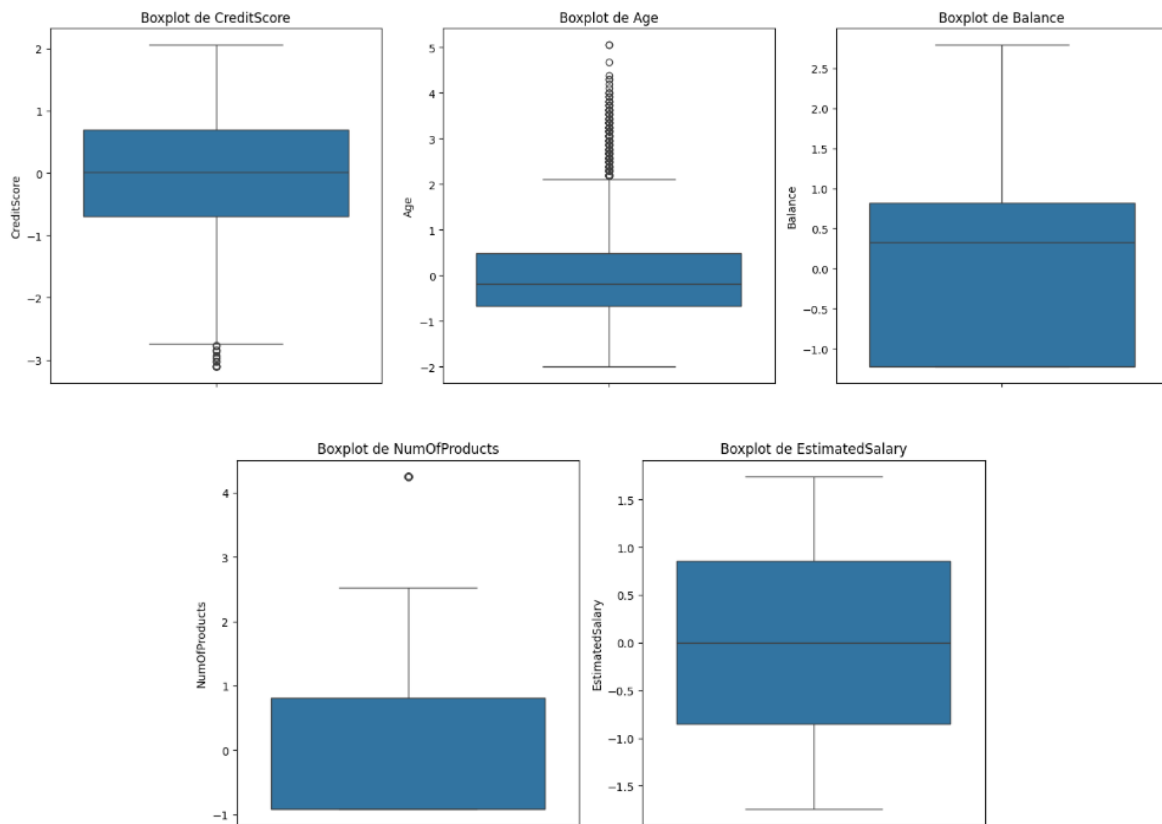
- CreditScore, Age, Balance, NumofProducts y EstimatedSalary.

Después, se configuró el tamaño general de la figura con `plt.figure(figsize=(15, 10))`, lo que asegura que las gráficas se vean “bien” es decir, organizadas y claras.

Posteriormente, vamos a crear un bucle para crear un boxplot para cada una de las columnas.

Se ajustará el espacio entre los gráficos y se dará el comando para mostrar los boxplots

El resultado es una visualización clara y ordenada de los datos...



Una vez obtenidas las gráficas logramos observar la presencia de outliers en “CreditScore, Age y NumofProducts”, por ende, debemos eliminar dichos valores atípicos.

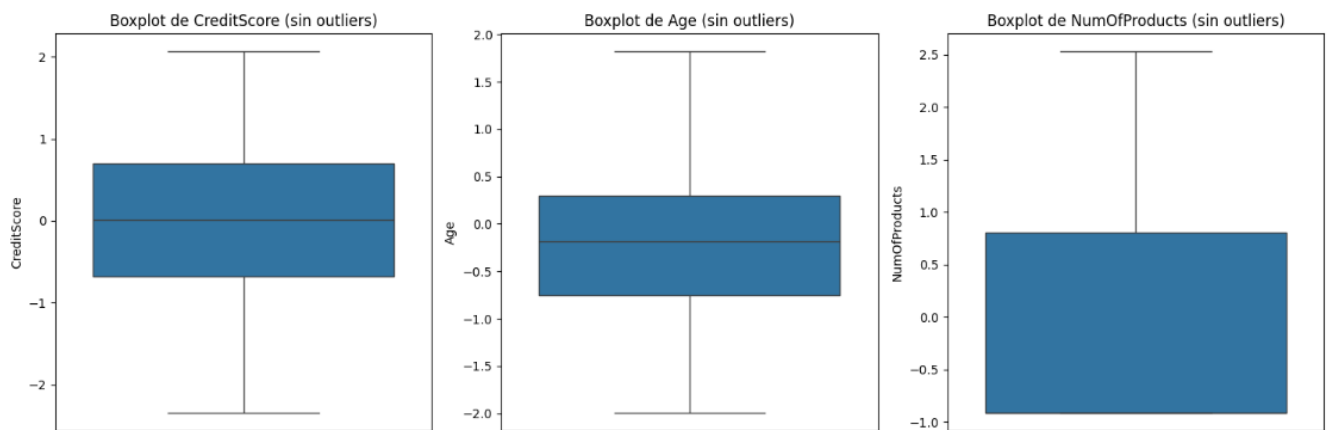
Para lograr este objetivo, se creó una primera lista llamada “columns to check”, la cual incluye los nombres de las columnas que vamos a revisar:

- CreditScore: Calificación crediticia del cliente.
- Age: Edad del cliente.
- NumOfProducts: Número de productos contratados por el cliente.

Enfocarnos en estas columnas nos permite centrar nuestra atención en aquellas que consideramos más importantes o más propensas a contener valores fuera de lo normal. El siguiente paso fue detectar y eliminar esos outliers, asegurándonos de que nuestros datos sean más confiables para el análisis o modelado.

Una vez realizado esto, tenemos como resultado, un boxplot limpio y preciso, tal y como se muestran en las gráficas a continuación, que vienen siendo las mismas mostradas

anteriormente, con la diferencia de que en estas logramos comprobar la eliminación de los outliers:



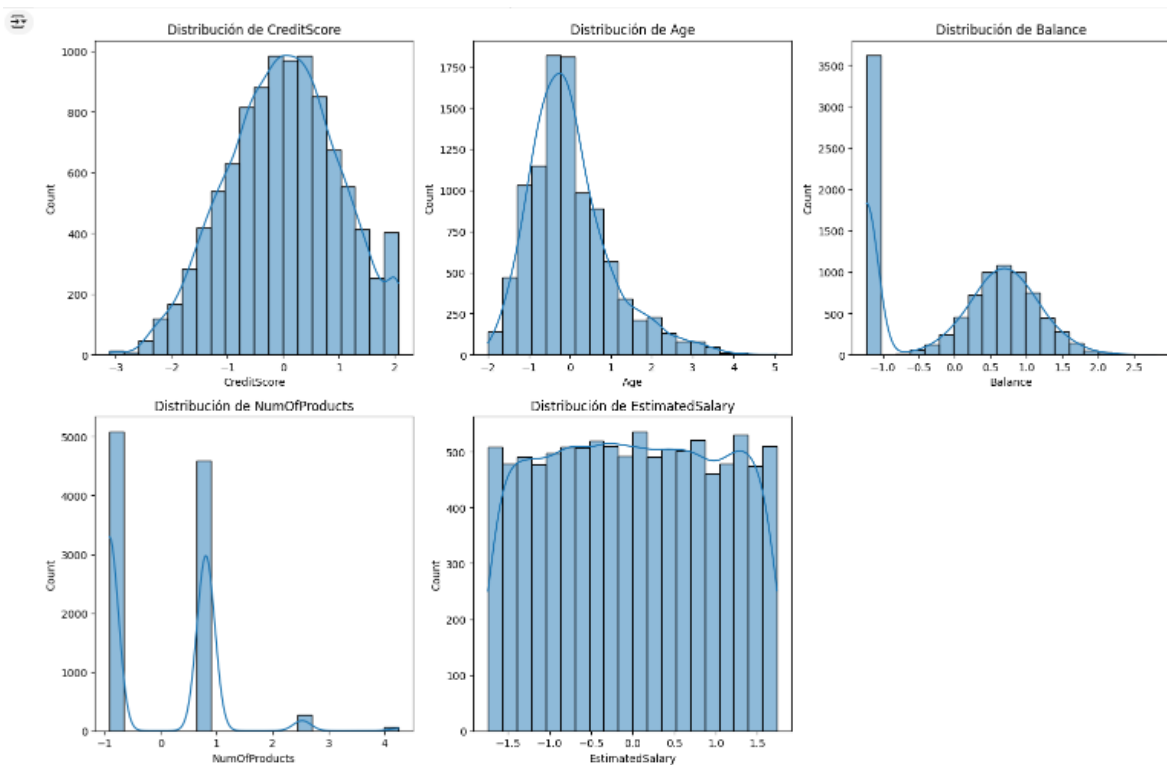
Luego, se procede a verificar si existen valores nulos, y se crea la verificación del número de filas antes y después de eliminar los outliers, sin embargo, contamos con la dicha de que nuevamente obtenemos 0 resultados nulos por columna.

```
⇒ Valores nulos por columna:
CreditScore      0
Age              0
Tenure           0
Balance          0
NumOfProducts    0
HasCrCard        0
IsActiveMember   0
EstimatedSalary  0
Exited           0
Geography_Germany 0
Geography_Spain  0
Gender_Male      0
```

Para ir finalizando, se crea un código acerca la lista de las columnas numéricas a analizar, entre ellas:

- CreditScore – Age – Balance – Numofproducts - Salary

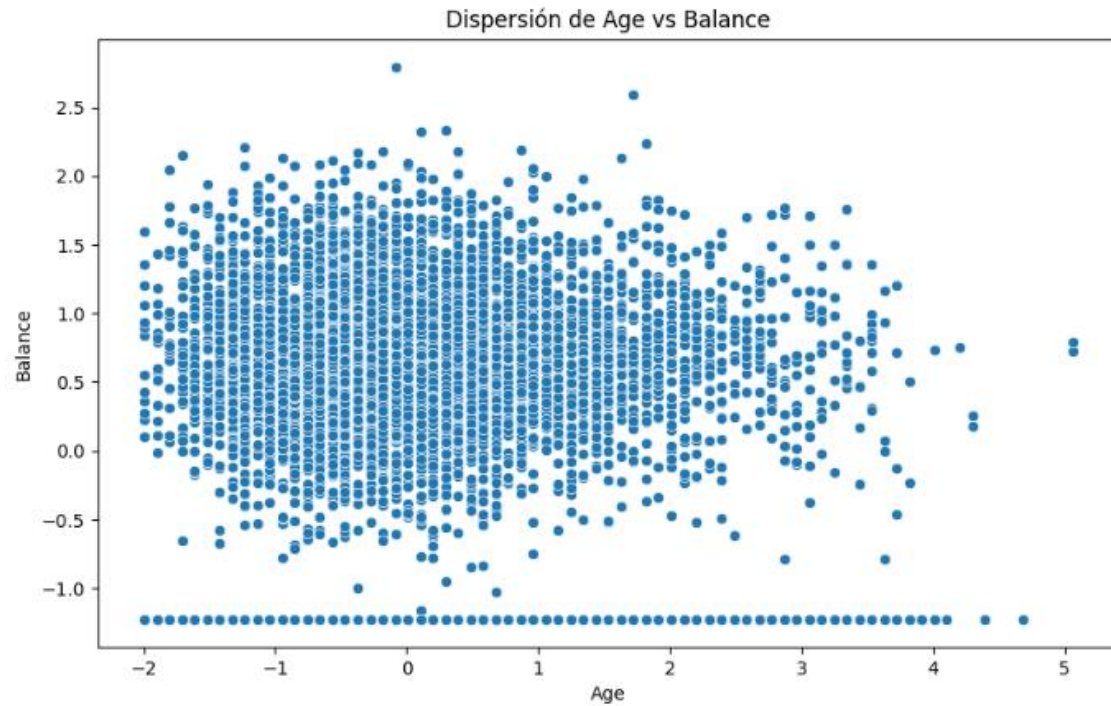
Y se ejecuta el comando para dar inicio a la creación de los histogramas, quedando representadas de la siguiente manera:



Como parte final del código, creamos un gráfico de dispersión, para poder visualizar la relación existente entre las variables “Age” y “Balance”, Finalmente, y después de todos los comandos utilizados en el trayecto mostramos el gráfico para visualizar la distribución de los datos.

Este tipo de gráfico es muy útil para analizar si hay alguna relación o patrón entre dos variables. Por ejemplo, podemos observar si a medida que la edad aumenta, el saldo también lo hace, o si no existe una relación aparente.

En este caso propiamente logramos interpretar que las personas mayores generalmente suelen tener un saldo más alto o que las personas jóvenes.



Ahora bien, teniendo en cuenta la explicación cerramos el código con el hecho de la exportación del nuevo dataset, mismo que ya se encuentra limpio y en las condiciones óptimas para ser trabajado.

```
[ ] # Exportamos el nuevo dataset limpio

ruta = "/content/drive/MyDrive/PROYECTO1/Dataset_Nuevo.csv"
data.to_csv(ruta, index=False)
```

- **Conclusiones y recomendaciones**

Conclusiones:

1. La conversión de las variables categóricas en el conjunto de datos resultó vital para la preparación de los datos a utilizar en el modelo. Este proceso garantiza que los datos estén en las mejores condiciones posibles, de manera que, simplifique la aplicación de algoritmos de machine learning que necesitan información clara, o bien, coherente y estandarizada para mejorar la precisión del objetivo principal del proyecto.
2. La limpieza y estandarización de los datos han sido cruciales para garantizar la calidad del dataset, eliminando inconsistencias y errores. Esto no solo facilita el desarrollo de modelos predictivos más confiables, sino que también asegura que las decisiones basadas en estos modelos sean más acertadas. Un dataset de alta calidad permite hacer análisis más precisos y mantener la fiabilidad del modelo a lo largo del tiempo, al reducir los riesgos de sesgos y resultados erróneos.
3. El análisis exploratorio de datos desempeña un papel fundamental en la comprensión de los factores que influyen en el churn, al identificar modelos y similitudes significativas entre variables... Esta etapa no solo mejora la calidad y la precisión del modelo al enfocarse en las características clave que determinan el comportamiento del cliente, optimiza la capacidad del modelo para predecir la deserción de manera efectiva. Es entonces que se puede mencionar que en términos generales el EDA se posiciona como una herramienta crucial para construir modelos robustos y confiables que puedan generar resultados prácticos y orientados a la toma de decisiones estratégicas.

Recomendaciones:

1. Implementar incentivos para clientes de alto riesgo: Para reducir la tasa de deserción en segmentos identificados como de alto riesgo, se recomienda diseñar incentivos específicos y personalizados.
 - Estos incentivos podrían incluir beneficios como descuentos, recompensas en función de la permanencia, programas de puntos, o servicios adicionales gratuitos durante un tiempo limitado. La efectividad de estos incentivos podría monitorearse mediante análisis continuos, ajustando las estrategias según el impacto observado.
2. Actualizar periódicamente el modelo predictivo: Con el fin de asegurar que el modelo predictivo se mantenga preciso y efectivo, es recomendable actualizarlo periódicamente con nuevos datos. Esto permitirá que el modelo refleje los cambios en el comportamiento de los clientes y los factores de riesgo emergentes, mejorando la capacidad del banco para predecir con mayor exactitud las posibilidades de deserción. La actualización continua también ayuda a mantener la competitividad del modelo frente a las dinámicas cambiantes del mercado financiero y las necesidades de los clientes.