



PROYECTO FINAL

CLASIFICACION DE OUTLIERS

Andre Sanchez

CONTEXTO

1. Objetivo de negocio

- Detectar transacciones potencialmente fraudulentas antes de que impacten al cliente o a la empresa.

2. Entorno y motivación

- Volumen diario: Miles de transacciones financieras en distintas ciudades.
- Riesgo de fraude: Costos económicos y reputacionales elevados.
- Necesidad: Herramienta automática que ayude a los analistas a enfocar sus esfuerzos.

3. Datos disponibles

- Transacciones: monto, canal (web, app, cajero), fecha y hora, ciudad.
- Usuarios: edad, ocupación, historial de intentos de inicio de sesión.
- Etiqueta: fraude confirmado vs. transacción legítima.

4. Enfoque del proyecto

- Exploración de patrones inusuales en el monto, frecuencia y métodos de pago.
- Modelado estadístico y de machine learning para predecir probabilidad de fraude.
- Evaluación basada en métricas de precisión, sensibilidad y tasa de falsos positivos.

HIPÓTESIS

H_0 (Hipótesis nula)

Las técnicas de clasificación binaria o el dataset no permiten detectar de manera efectiva outliers con una precisión significativa en el dataset de transacciones bancarias.

H_4 (Hipótesis general)

Existen patrones transaccionales inusuales que pueden ser detectados como outliers y representar posibles indicios de fraude.

Hipótesis específicas

- H_1 : Las transacciones con montos extremos, ubicaciones atípicas o canales no habituales presentan mayor probabilidad de ser clasificadas como outliers.
- H_2 : Las transacciones precedidas por múltiples intentos de inicio de sesión tienen una mayor probabilidad de ser detectadas como outliers.
- H_3 : Existen ciertos rangos de edad u ocupaciones en los que se presentan más outliers, debido a variaciones en el comportamiento transaccional típico de esos grupos.

GLOSARIO

Columna	Descripción
TransactionID	Identificador único alfanumérico para cada transacción
AccountID	Identificador único para cada cuenta, con múltiples transacciones por cuenta
TransactionAmount	Valor monetario de cada transacción, que varía desde pequeños gastos diarios hasta compras más grandes
TransactionDate	Marca temporal de cada transacción, capturando fecha y hora
TransactionType	Campo categórico que indica transacciones de "Crédito" o "Débito"
Location	Ubicación geográfica de la transacción, representada por nombres de ciudades de EE.UU.
DeviceID	Identificador alfanumérico para dispositivos utilizados para realizar la transacción
IP Address	Dirección IPv4 asociada con la transacción, con cambios ocasionales para algunas cuentas
MerchantID	Identificador único para comerciantes, mostrando comerciantes preferidos y atípicos para cada cuenta
AccountBalance	Saldo en la cuenta después de la transacción, con correlaciones lógicas basadas en el tipo y monto de la transacción
PreviousTransactionDate	Marca temporal de la última transacción para la cuenta, ayudando a calcular la frecuencia de las transacciones
Channel	Canal a través del cual se realizó la transacción (por ejemplo, en línea, cajero automático, sucursal)
CustomerAge	Edad del titular de la cuenta, con agrupaciones lógicas basadas en la ocupación
CustomerOccupation	Ocupación del titular de la cuenta (por ejemplo, médico, ingeniero, estudiante, jubilado), reflejando patrones de ingresos
TransactionDuration	Duración de la transacción en segundos, variando según el tipo de transacción
LoginAttempts	Número de intentos de inicio de sesión antes de la transacción, con valores más altos que indican posibles anomalías

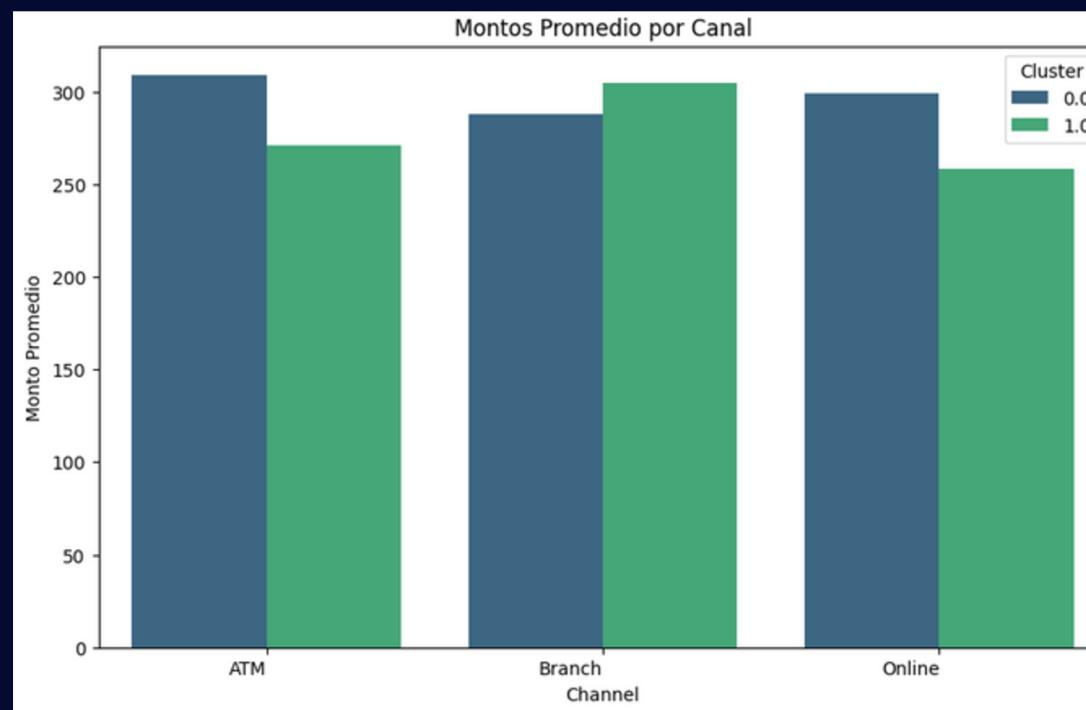
ANALISIS EXPLORATORIO (EDA)

En la sección de Análisis Exploratorio de Datos (EDA) exploraremos de manera visual y descriptiva los patrones y comportamientos más relevantes de las transacciones bancarias. Aquí buscamos comprender cómo se distribuyen los montos, en qué canales y ciudades ocurren la mayoría de las operaciones, y cómo variables como la edad del cliente o los intentos de inicio de sesión previos pueden relacionarse con posibles anomalías. Este paso es fundamental para descubrir tendencias, identificar valores atípicos y sentar las bases de nuestro modelo de detección de fraude de forma clara y accesible, apoyándonos en gráficos sencillos y métricas intuitivas.

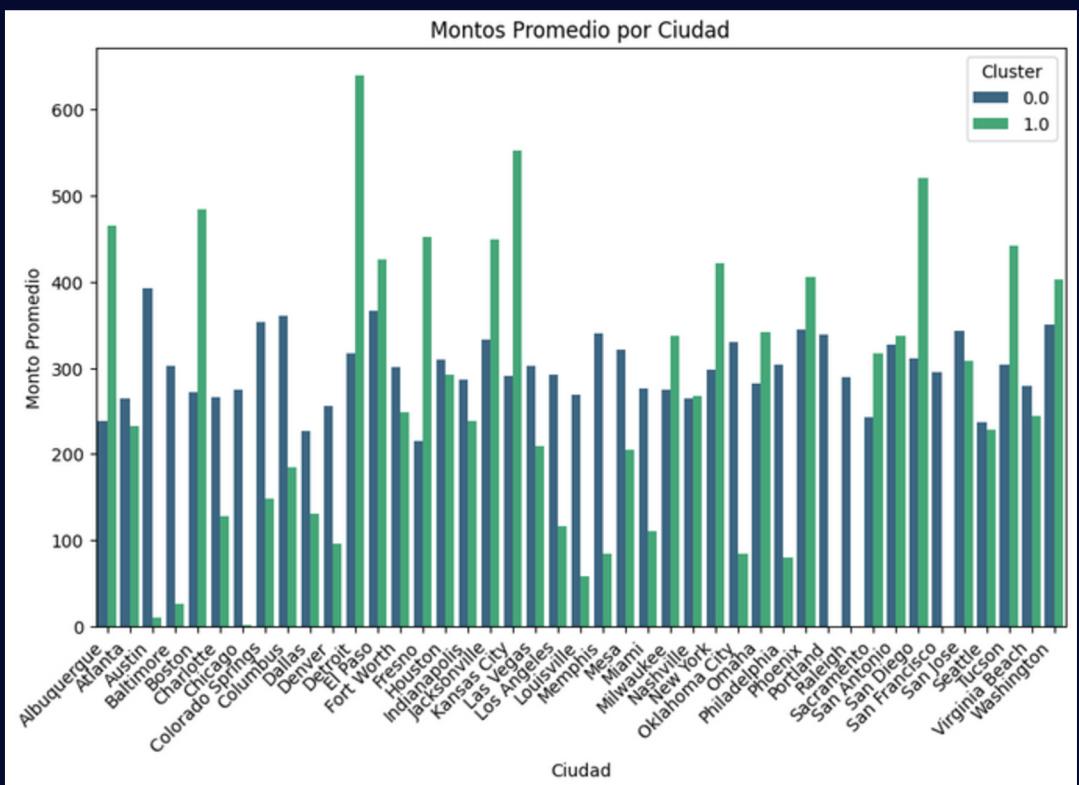


¿QUÉ CARACTERÍSTICAS TRANSACCIONALES (COMO EL MONTO, LA UBICACIÓN O EL CANAL UTILIZADO SON MÁS FRECUENTES EN LAS TRANSACCIONES CONSIDERADAS ANÓMALAS?

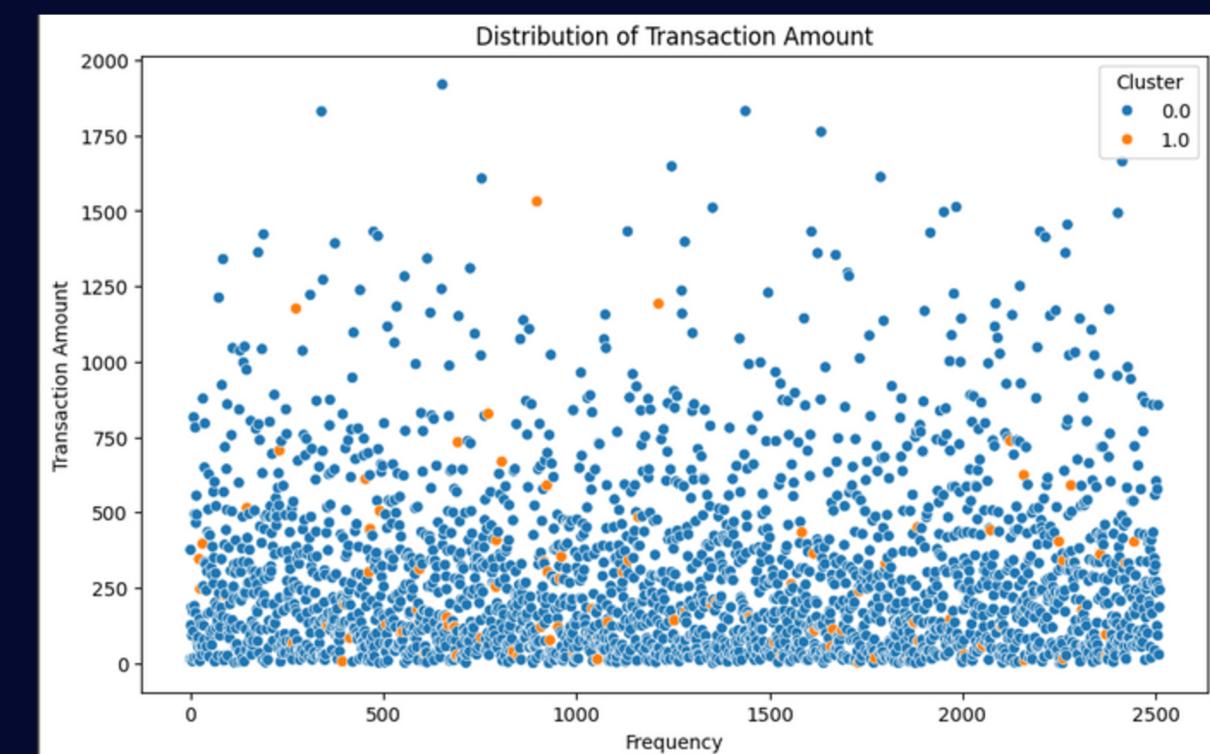
Cuando comparamos los canales, la atención presencial presenta importes de outliers algo superiores a los montos de las transacciones legítimas. Aunque la diferencia es leve, sugiere que incluso en interacciones cara a cara pueden surgir operaciones con valores inusuales.



Al estudiar la distribución por ciudad, encontramos que en lugares como Detroit, Kansas City, Fresno y Albuquerque el monto promedio de las transacciones etiquetadas como outliers supera al de las operaciones estándar. Esto señala que en estas localidades las anomalías financieras tienden a involucrar importes más elevados que en el promedio general.

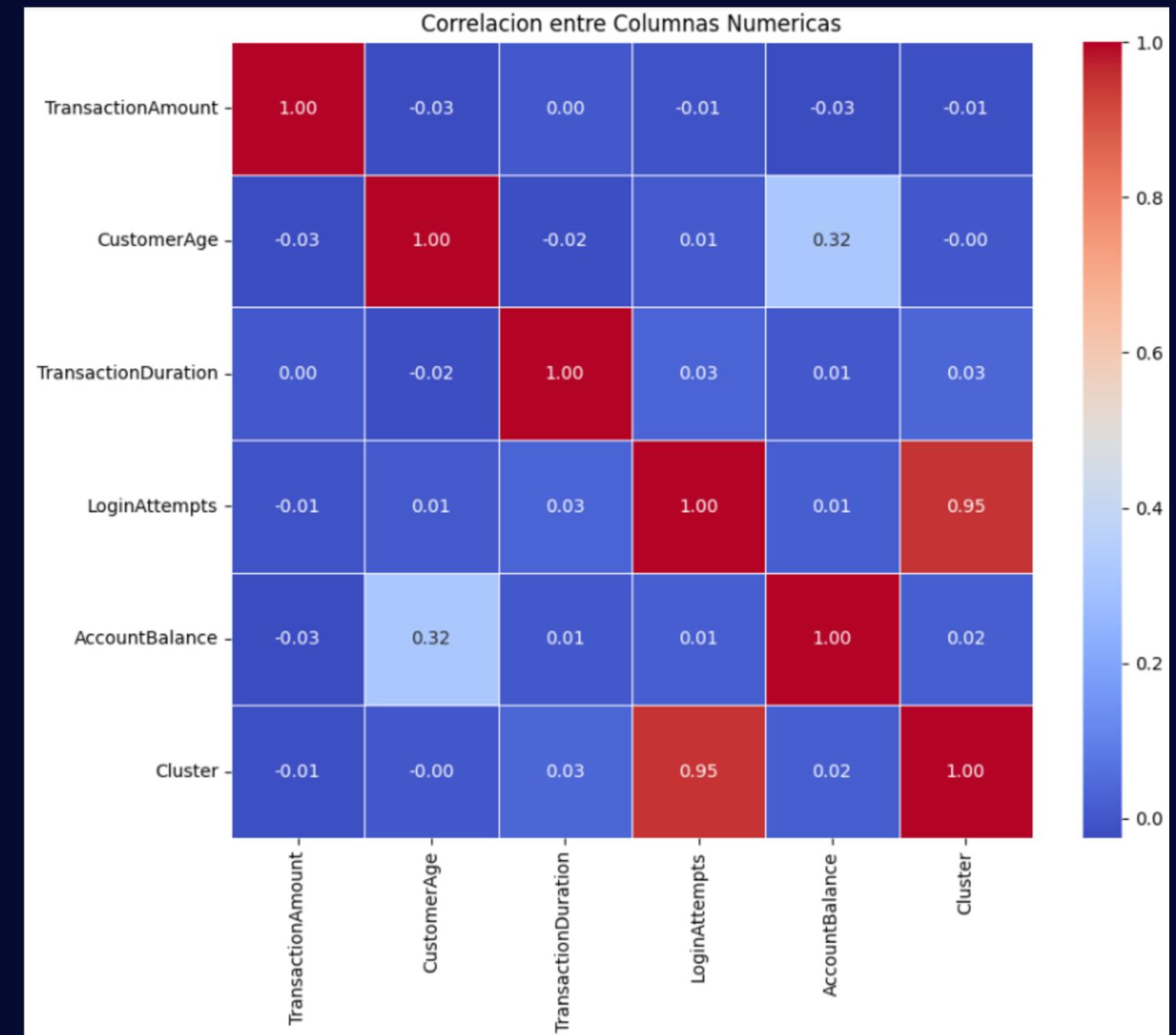


El análisis muestra que no todos los importes elevados se comportan como outliers. Es decir, un monto alto por sí solo no garantiza que la transacción sea atípica; muchos montos significativos se alinean con el patrón habitual de gasto de los usuarios.



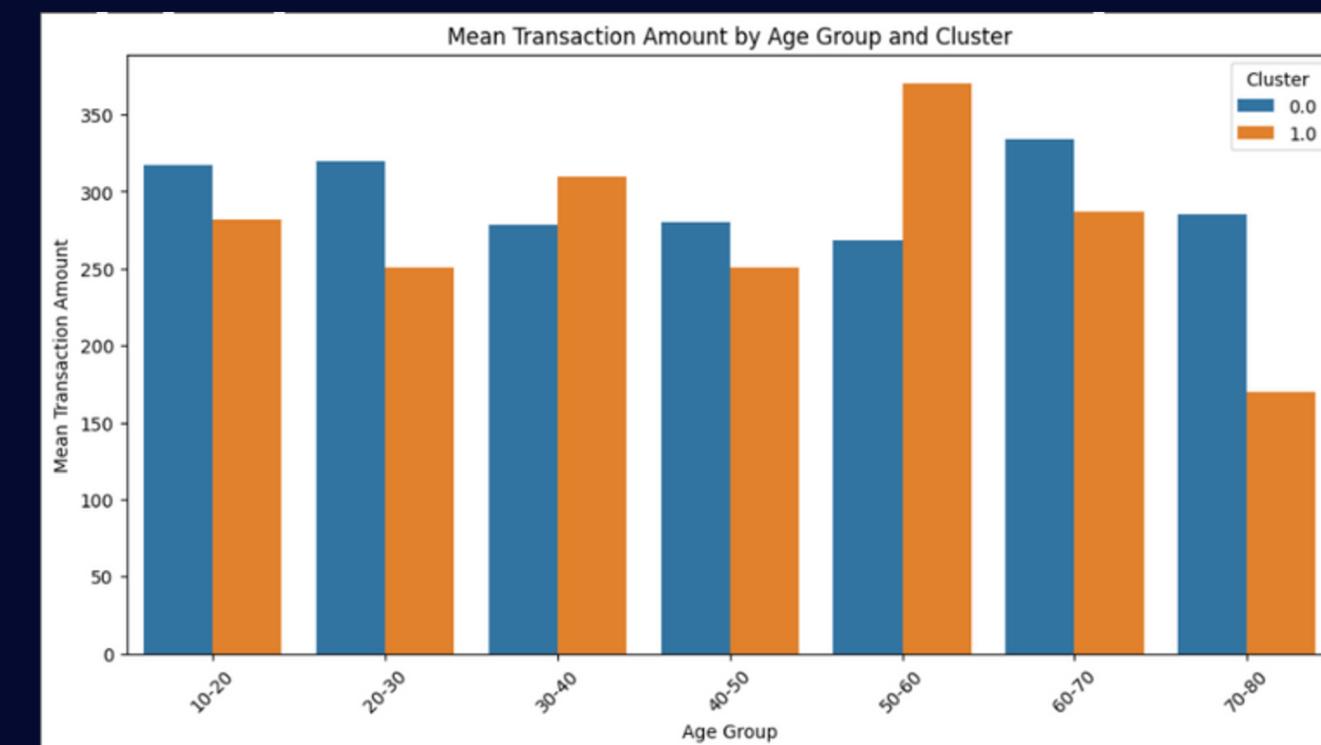
¿LAS TRANSACCIONES REALIZADAS CON MÚLTIPLES INTENTOS DE INICIO DE SESIÓN TIENDEN A COMPORTARSE COMO OUTLIERS?

En la matriz de correlación observamos una relación muy fuerte (≈ 0.95) entre el número de intentos de inicio de sesión y la pertenencia al grupo de transacciones anómalas. Esto indica que, a medida que aumenta la cantidad de intentos fallidos antes de una operación, también crece de forma muy consistente la probabilidad de que dicha transacción sea clasificada como outlier. En otras palabras, múltiples intentos de autenticación fallidos son un claro indicador de comportamientos inusuales en el sistema.

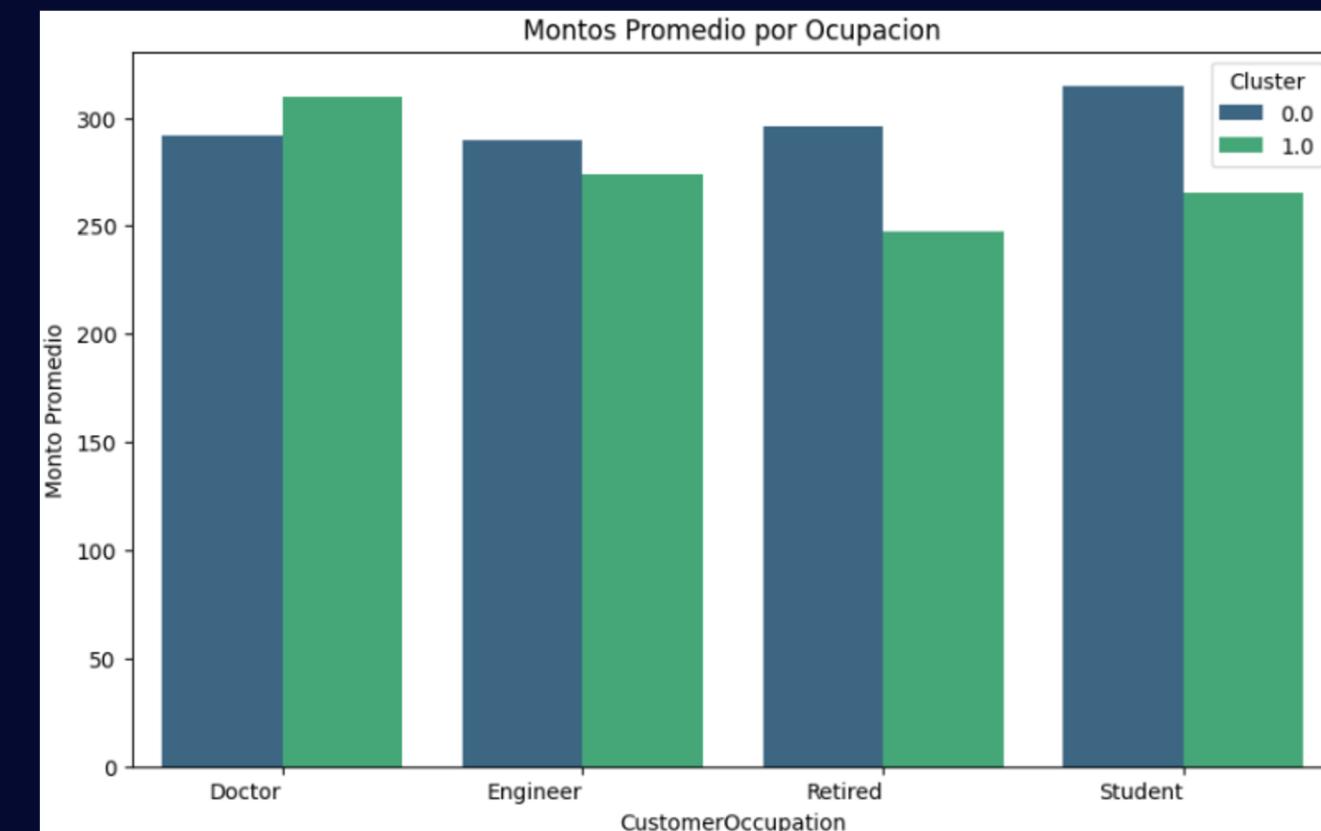


¿HAY OCUPACIONES O RANGOS DE EDAD QUE PRESENTEN UN MAYOR MONTO DE TRANSACCIONES ATÍPICAS?

En el análisis por rango etario, observamos que únicamente los grupos de **30-40** y **50-60** años presentan un **monto promedio de transacciones atípicas superior al de las operaciones legítimas**. Esto sugiere que, en estos tramos de edad, las variaciones en el comportamiento financiero —ya sean compras extraordinarias o movimientos poco comunes— tienden a registrarse con importes más elevados y, por tanto, es más probable que se detecten como outliers.

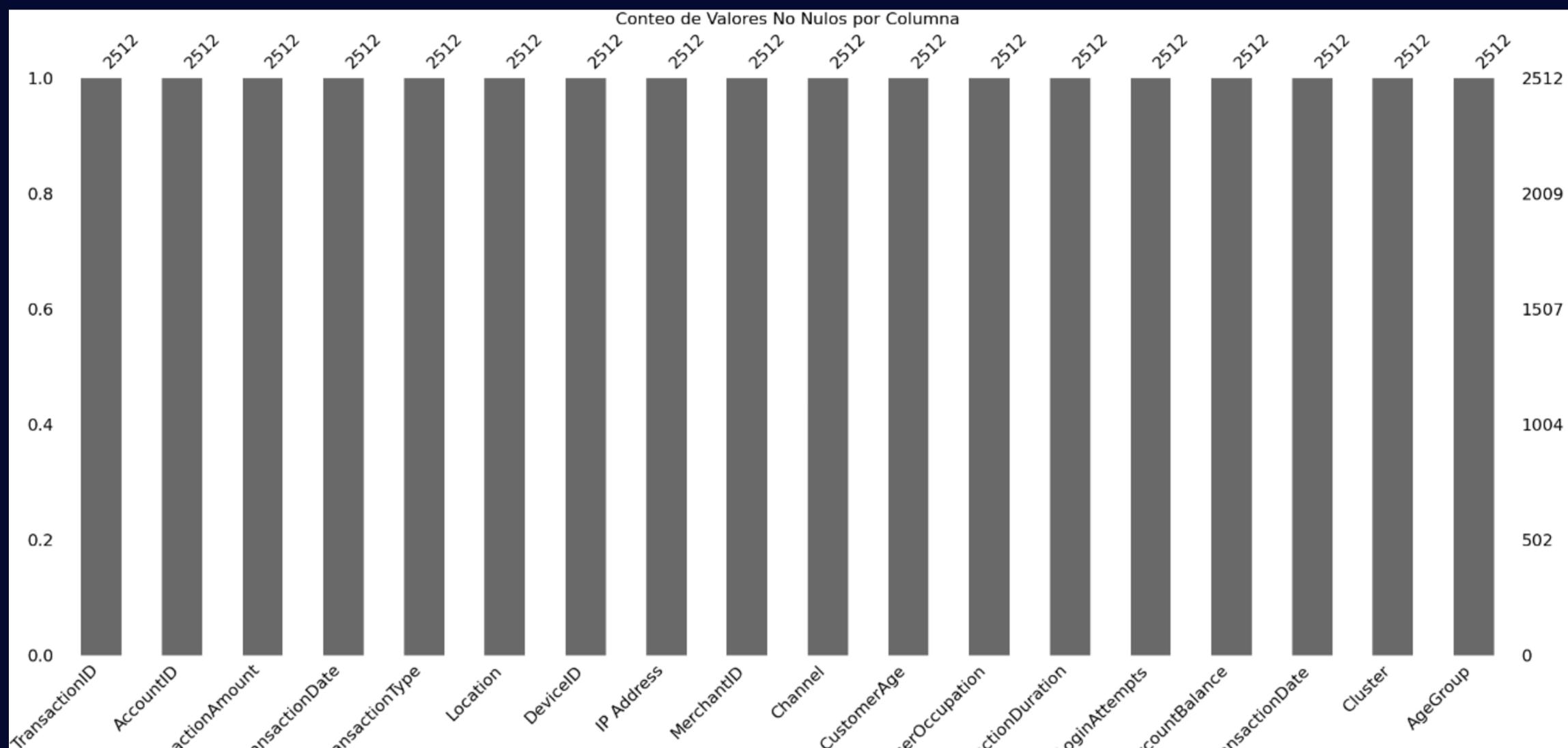


En el análisis según ocupación, únicamente el grupo de **médicos** muestra un monto promedio de transacciones atípicas superior al de sus operaciones legítimas. Esto revela que, entre las distintas profesiones, los médicos realizan con mayor frecuencia movimientos financieros inusuales de mayor cuantía, convirtiéndose en un segmento especialmente relevante para la detección de posibles fraudes.



COMPROBACION DE VALORES FALTANTES

Luego de realizar las comprobaciones correspondientes sobre la integridad del dataset, se verificó que ninguna de las columnas presenta valores vacíos. Esto asegura que la información está completa y lista para ser utilizada en los análisis posteriores, sin necesidad de aplicar técnicas de imputación o limpieza adicional en esta etapa.

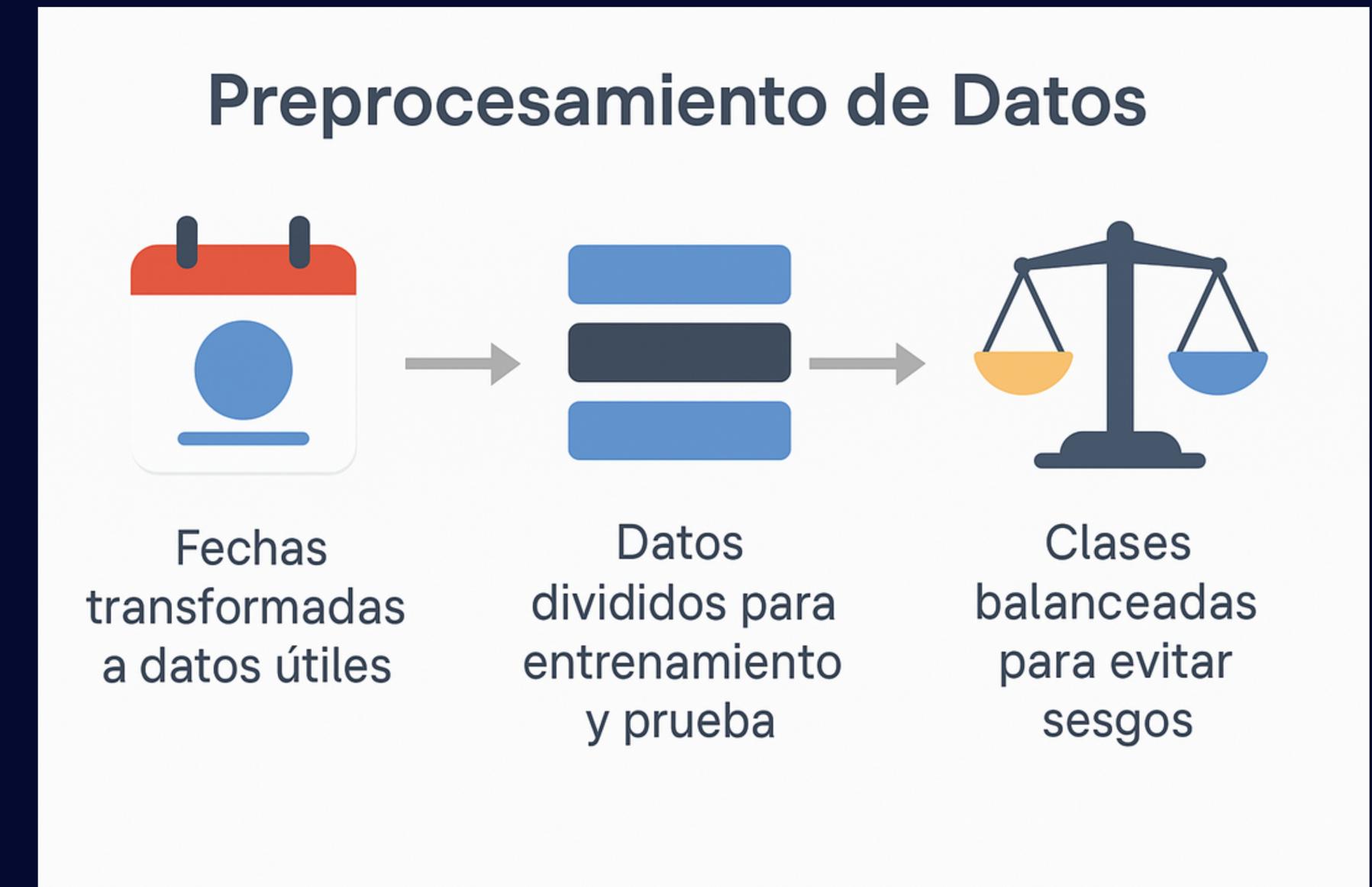


PRE-PROCESAMIENTO DE LOS DATOS

Para el preprocessamiento de los datos, se realizó en primer lugar una transformación de las columnas de tipo fecha, extrayendo componentes relevantes como el día, la hora o la diferencia entre transacciones.

Posteriormente, se procedió a dividir el dataset en conjuntos de entrenamiento y prueba, asegurando una adecuada evaluación del modelo.

Finalmente, se aplicó una técnica de balanceo de clases en el conjunto de entrenamiento con el fin de corregir el desbalance entre transacciones normales y outliers, lo cual es clave para mejorar la capacidad predictiva del modelo frente a casos minoritarios.



SELECCIÓN DE CARACTERÍSTICAS

Para optimizar el rendimiento del modelo, se aplicaron técnicas de selección de variables que permitieron identificar las características más influyentes en la detección de transacciones atípicas.

En el caso de las variables numéricas, se utilizó un método llamado Sequential Feature Selection (SFS), que selecciona de forma progresiva las variables que más aportan al modelo.

Para las variables categóricas, se empleó el coeficiente chi-cuadrado (χ^2), una técnica estadística que mide la relación entre variables, ayudando a identificar aquellas categorías que tienen mayor impacto en la clasificación.

Selección de Variables Relevantes

Variables Numéricas



Sequential Feature Selection

Variables Categóricas

$$\chi^2$$

Coeficiente Chi-Cuadrado

CONSTRUCCION DE PIPELINE

Se diseñó una pipeline automatizada que permite procesar los datos de manera eficiente y preparar el modelo para la predicción.

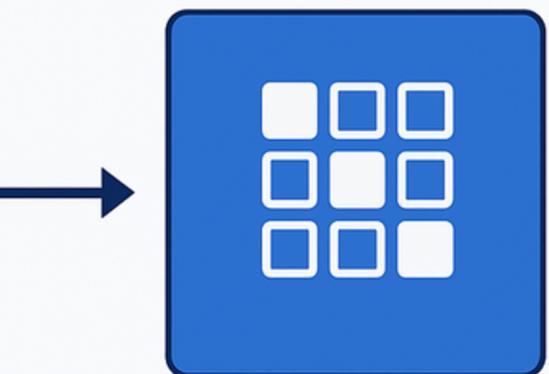
Esta pipeline transforma las variables numéricas mediante un proceso de estandarización (para que todas tengan la misma escala) y convierte las variables categóricas en valores numéricos comprensibles para el modelo mediante una técnica llamada OneHot Encoding.

Finalmente, se incorporó un modelo de clasificación basado en árboles de decisión, conocido como Random Forest, el cual se encarga de detectar patrones complejos y clasificar las transacciones como normales o atípicas.

Construcción de la Pipeline de Modelado



Escalamiento
de Variables
Numéricas



Codificación
OneHot d
Variables
Categóricas



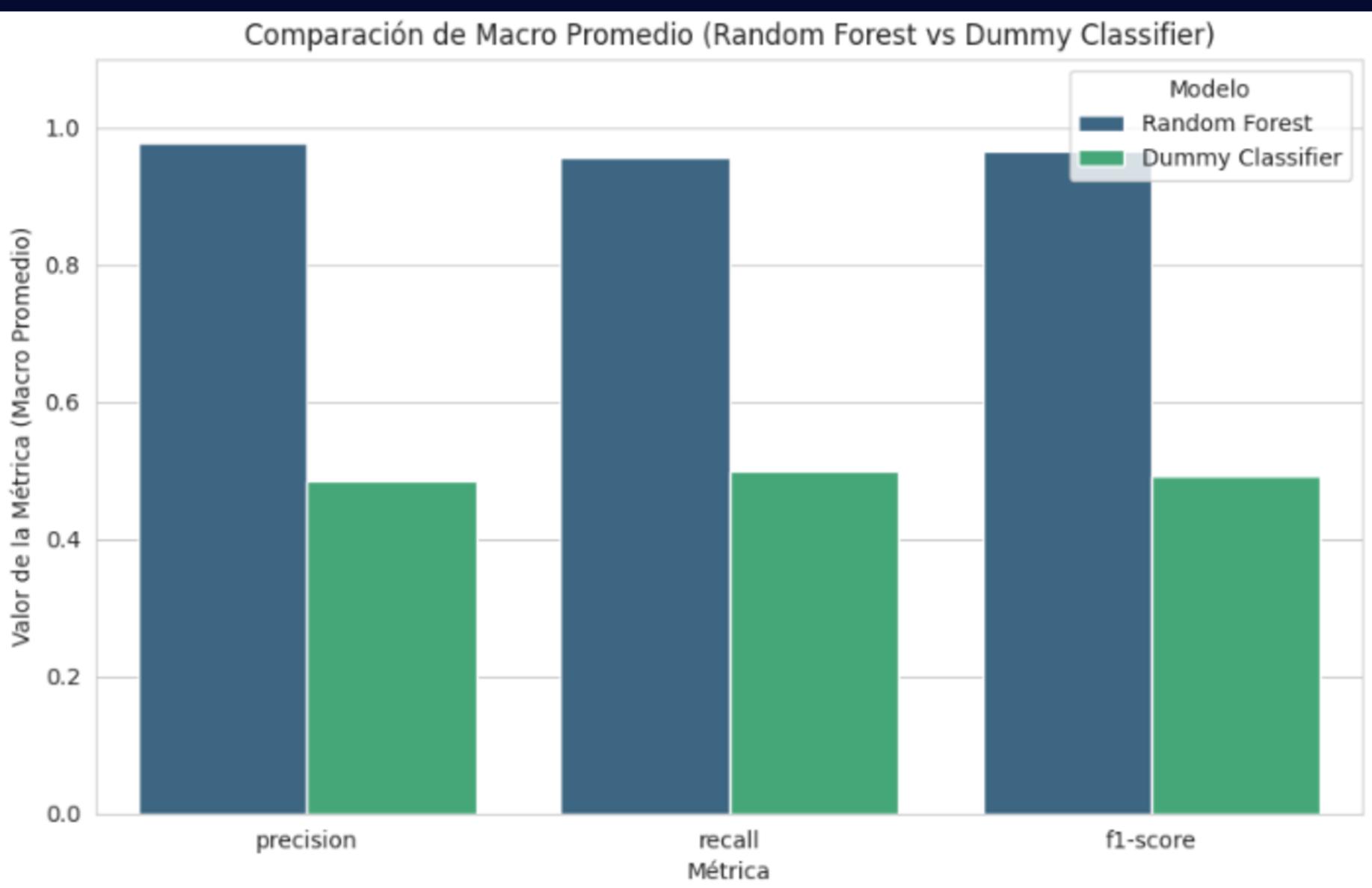
Random
Forest

METRICAS

Para evaluar el desempeño del modelo propuesto, se compararon distintas métricas de clasificación entre un modelo base (modelo dummy) y el modelo de Random Forest entrenado. Las métricas consideradas fueron precisión, recall y F1-score, enfocándose especialmente en la clase minoritaria (outliers).

El modelo dummy, que sirve como referencia, mostró un rendimiento limitado, con baja capacidad para identificar correctamente las transacciones anómalas.

En contraste, el modelo de Random Forest logró mejoras significativas en todas las métricas, evidenciando una mayor capacidad para detectar outliers de forma más precisa y consistente. Esta comparación refuerza la eficacia del enfoque propuesto frente a una estrategia aleatoria o ingenua.



CONCLUSIÓN

1. Cumplimiento de Objetivos

El modelo de Random Forest implementado alcanzó el objetivo de negocio al detectar transacciones potencialmente fraudulentas con una fiabilidad muy superior al modelo de referencia (dummy), permitiendo focalizar los recursos de los analistas en los casos de mayor riesgo.

2. Validación de Hipótesis:

- H_1 (Monto, ubicación y canal): Confirmado; ciertas ciudades y canales presentan montos atípicos más frecuentes.
- H_2 (Intentos de inicio de sesión): Confirmado; existe una fuerte correlación entre múltiples intentos fallidos y la clasificación como outlier.
- H_3 (Edad y ocupación): Confirmado para rangos 30-40 y 50-60 años, y para la profesión de médicos; estos segmentos muestran montos promedio de outliers superiores al de transacciones legítimas.
- H_4 (General): Confirmado; se identificaron patrones transaccionales inusuales que efectivamente sirven como indicios de posible fraude



GRACIAS