# ANÁLISIS PREDICTIVO

## Determinación de la base de datos

Para este proyecto se seleccionó la base de datos SAML-D (Suspicious Activity Monitoring in Laundering - Dataset), una muestra representativa de transacciones financieras que incluye atributos relevantes para la detección de posibles casos de lavado de dinero. Esta muestra contiene variables clave como montos, monedas, ubicaciones bancarias, tipo de pago y una clasificación binaria que indica si la transacción es sospechosa o no, así como su tipología asociada en caso de serlo.

El motivo principal de esta elección radica en mi interés y proyección profesional en el área de análisis de riesgo financiero. Actualmente estoy por iniciar un puesto relacionado directamente con esta disciplina, por lo que consideré fundamental trabajar con un conjunto de datos que represente escenarios reales de detección de operaciones inusuales, con el fin de fortalecer mis habilidades en análisis predictivo, minería de datos y prevención de delitos financieros.

Además, este dataset ofrece el desafío adicional de enfrentar un problema altamente desbalanceado, lo cual es común en situaciones reales de riesgo financiero, y me ha permitido aplicar técnicas avanzadas de preprocesamiento, balanceo de clases y modelos supervisados, reforzando así mi preparación para asumir responsabilidades analíticas en entornos profesionales.

## Pre-procesamiento y limpieza

El archivo original contenía 9,504,852 transacciones, de las cuales aproximadamente el 0.1039 % estaban etiquetadas como sospechosas, simulando de forma realista el fuerte desbalance que ocurre en entornos financieros reales.

Dado el tamaño del archivo original y las limitaciones prácticas de procesamiento en entornos de escritorio, se realizó un recorte del conjunto de datos, extrayendo una muestra representativa del 1 % mediante muestreo estratificado, asegurando que se conservara la proporción original de transacciones sospechosas. Esta versión reducida conserva la estructura y el comportamiento estadístico del problema original, facilitando el análisis exploratorio, predictivo y visual.

Posteriormente, se llevó a cabo un proceso exhaustivo de limpieza y transformación del dataset, con el objetivo de asegurar su integridad y prepararlo para el modelado predictivo. Las principales etapas del pre-procesamiento fueron:

### Estandarización de nombres de columnas:

Todos los nombres de columnas fueron convertidos a minúsculas y se eliminaron espacios, utilizando guiones bajos para asegurar una sintaxis uniforme.

### Conversión de fecha y hora a formato datetime:

Se integraron las columnas date y time en una nueva variable datetime, en formato estándar, para permitir análisis cronológicos y temporales.

### Eliminación de columnas duplicadas:

Se identificaron y eliminaron columnas redundantes.

### Revisión de valores nulos:

Se comprobó que no existían datos nulos en el conjunto de datos, por lo que no fue necesario aplicar técnicas de imputación.

### Codificación de variables categóricas:

Variables como payment\_type, sender\_bank\_location, receiver\_bank\_location y laundering\_type fueron transformadas numéricamente mediante Label Encoding para poder ser procesadas por los algoritmos de aprendizaje automático.

### Balanceo de clases con SMOTE:

Dado el fuerte desbalance de clases, se aplicó SMOTE (Synthetic Minority Over-sampling Technique) para generar nuevas muestras sintéticas de la clase minoritaria (transacciones sospechosas). Esto permitió entrenar modelos sin sesgo hacia la clase mayoritaria.

### División en conjuntos de entrenamiento y prueba:

Finalmente, se dividieron los datos balanceados en conjuntos de entrenamiento (70%) y prueba (30%) para asegurar una evaluación justa y representativa del desempeño de los modelos.

## Análisis Descriptivo

El análisis exploratorio de datos (EDA) permitió conocer la estructura general del conjunto de datos, así como identificar patrones relevantes en la distribución de transacciones y posibles comportamientos anómalos.

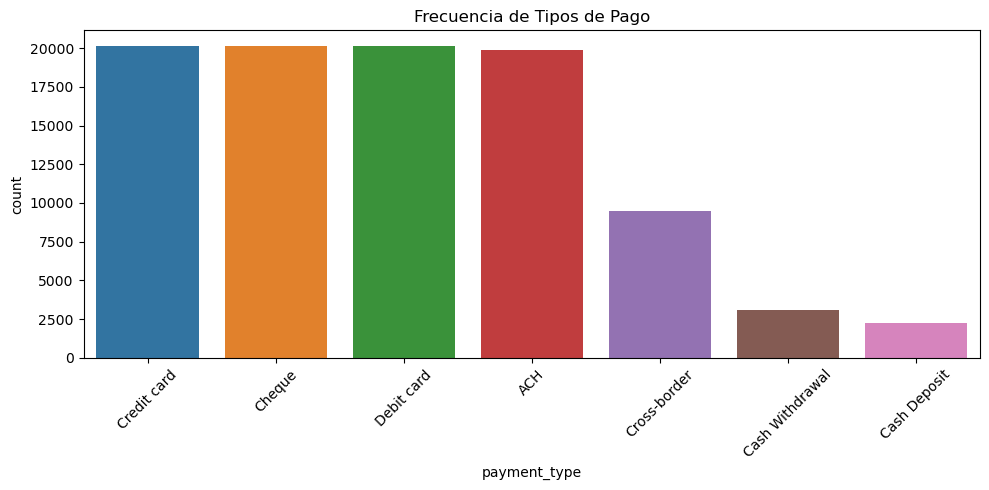
A continuación, se destacan los principales hallazgos:

### Distribución de Monto de Transacciones

La variable amount presenta una distribución fuertemente asimétrica hacia la derecha (distribución sesgada positiva), como se observa en la gráfica anterior. La mayoría de las transacciones se concentran en rangos bajos probablemente menores a 10,000 unidades monetarias mientras que una cantidad muy reducida de transacciones presenta montos extremadamente altos, alcanzando incluso cifras cercanas al millón.

Esta concentración masiva en la parte izquierda del eje X indica una alta frecuencia de transacciones de bajo valor, mientras que los valores atípicos (outliers) en la parte derecha representan posibles transacciones de alto riesgo, especialmente relevantes en contextos de lavado de dinero.

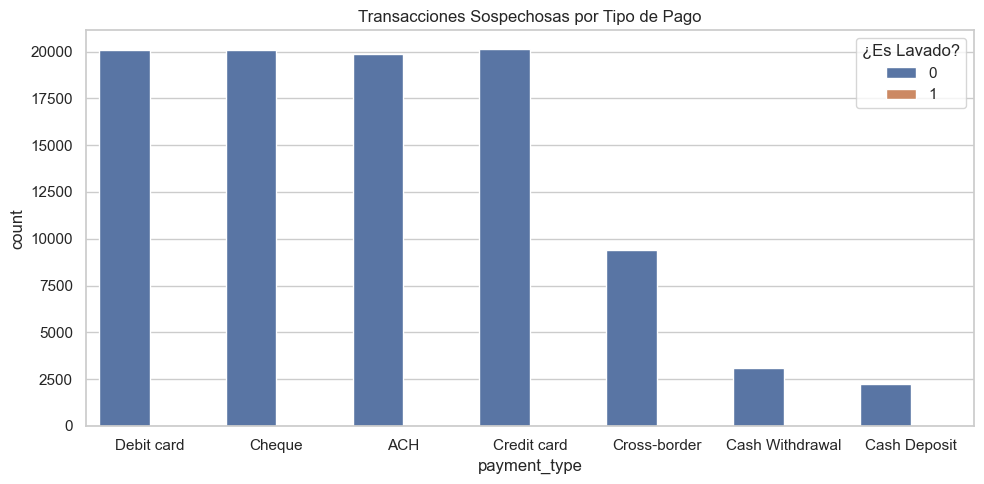
### Tipos de pagos mas frecuentes



El tipo de pago fue una de las variables categóricas más relevantes en el análisis descriptivo. Como se observa en la gráfica, los métodos más utilizados fueron Credit Card, Cheque, Debit Card y ACH, todos con volúmenes similares cercanos a las 20,000 transacciones. Esto sugiere que estos métodos constituyen los canales más comunes dentro de las operaciones financieras representadas en el dataset.

En contraste, los métodos Cash Withdrawal y Cash Deposit fueron los menos frecuentes, lo cual podría estar asociado a controles más estrictos sobre operaciones en efectivo o a la automatización digital del sistema financiero.

### Transacciones sospechosas por tipo de pagos

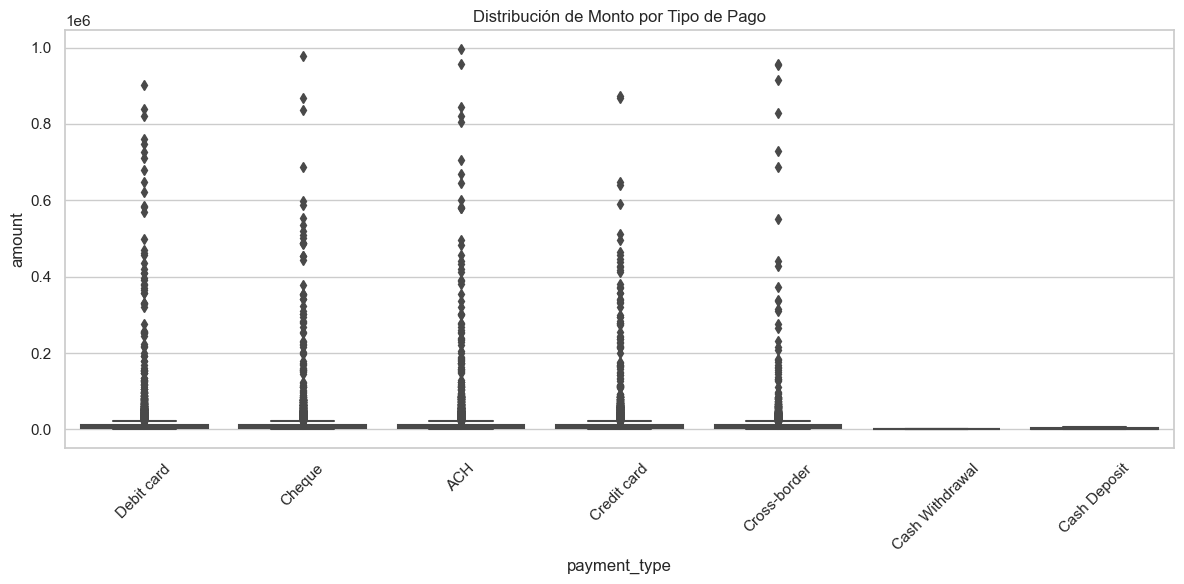


La gráfica comparativa por tipo de pago, segmentada por la variable is\_laundering, muestra que las transacciones sospechosas (valor 1) son una proporción extremadamente baja en comparación con las no sospechosas (valor 0) en todos los métodos de pago.

Aunque los tipos más frecuentes como Debit Card, Cheque, ACH y Credit Card presentan altos volúmenes de transacciones, no se observan picos significativos de actividades sospechosas en ninguna de estas categorías. Esto confirma el desbalance natural del dataset, donde las operaciones inusuales son escasas y no dominan ningún método de pago en particular.

Sin embargo, métodos como Cash Deposit y Cash Withdrawal, aunque menos frecuentes, podrían ser de interés en investigaciones futuras, ya que en otros contextos suelen estar más relacionados con actividades no reguladas. Sería útil analizar estas categorías de forma aislada en modelos de análisis más avanzados o mediante técnicas no supervisadas.

### Dsitribución de monto por tipo de pago



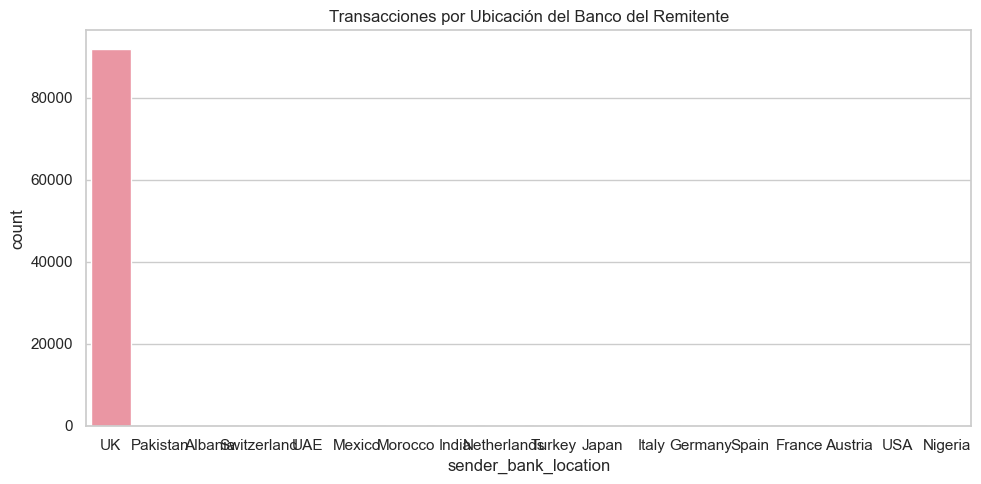
La gráfica de cajas (boxplot) por tipo de pago permite observar con claridad la dispersión de los montos en cada categoría. Se identificó que todos los métodos presentan una gran concentración de valores bajos, acompañados de múltiples valores atípicos (outliers) con montos elevados, algunos incluso cercanos al millón.

Aunque la mediana de los montos es relativamente baja en todas las categorías, los métodos como ACH, Cross-border y Cheque muestran una mayor amplitud en la distribución, lo cual podría indicar un uso más frecuente en operaciones de alto valor.

Este comportamiento sugiere que ciertos métodos de pago no solo son más utilizados, sino también más propensos a facilitar movimientos financieros de gran magnitud, un elemento que puede ser relevante al momento de evaluar el riesgo de lavado de dinero.

Este hallazgo refuerza la decisión de incluir payment\_type y amount como variables clave en los modelos predictivos del proyecto.

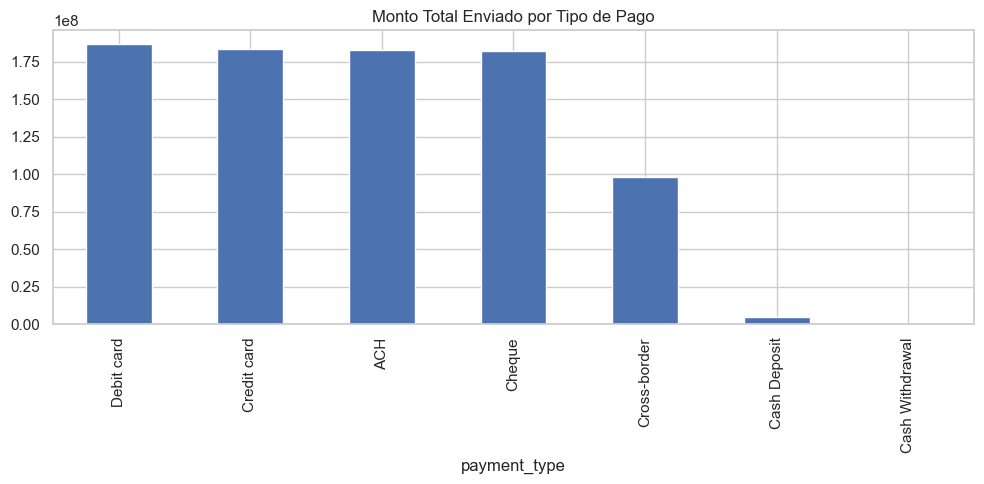
### Transacciones por ubicación del banco remitente



Al analizar la variable sender\_bank\_location, se observó una altísima concentración de transacciones originadas en el Reino Unido (UK), con más de 90,000 registros en la muestra analizada. Esto refleja que la base de datos tiene un fuerte enfoque en el sistema financiero británico, posiblemente debido a su rol central en redes financieras internacionales.

El resto de los países aparece con una frecuencia considerablemente menor, incluyendo ubicaciones como Pakistán, Suiza, India, Alemania, USA y España, entre otros. Aunque su volumen es reducido, la diversidad geográfica sigue siendo relevante, ya que en contextos de análisis de riesgo, transacciones internacionales pueden tener características diferentes en cuanto a regulación, vigilancia y patrones operativos.

### Monto total enviado por tipo de pago

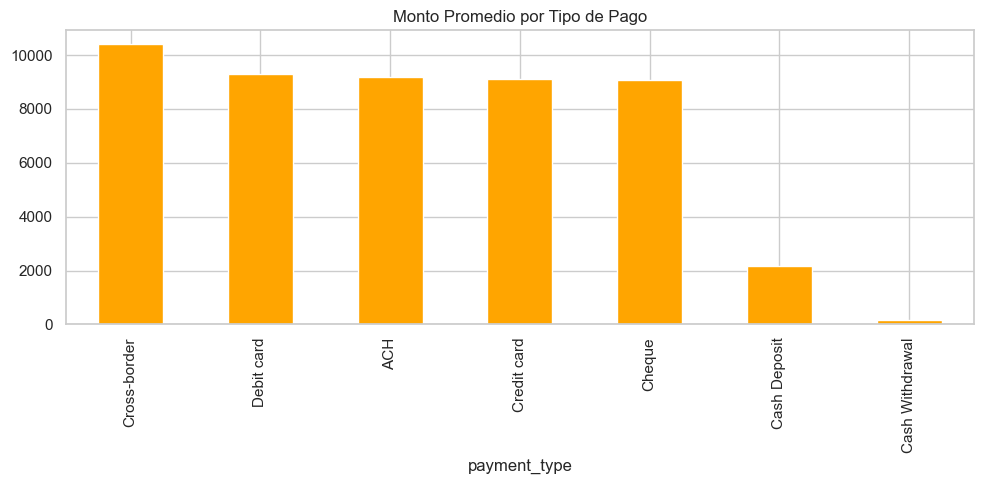


Esta gráfica muestra el monto total acumulado que fue enviado mediante cada tipo de método de pago. Se evidencia que los métodos Debit Card, Credit Card, ACH y Cheque representan el grueso de las transacciones en términos de volumen monetario, cada uno superando los 180 millones de unidades monetarias.

Este comportamiento está alineado con su frecuencia de uso en la base de datos, pero también resalta su importancia desde una perspectiva operativa y de riesgo: estos métodos podrían ser el canal preferido tanto para transacciones legítimas como para posibles maniobras de lavado si no se supervisan adecuadamente.

En contraste, métodos como Cash Deposit y Cash Withdrawal presentan montos totales mucho menores. Aunque podrían parecer irrelevantes en volumen, su baja trazabilidad y uso esporádico pueden convertirlos en canales discretos para actividades ilícitas.

### Monto promedio por tipo de pago



Esta segunda gráfica refleja el monto promedio por transacción para cada tipo de pago, revelando patrones menos evidentes en la simple frecuencia.

Destaca que las transacciones Cross-border (internacionales) presentan el mayor monto promedio, superando los 10,000 por operación. Este dato es coherente con lo esperado, ya que las operaciones internacionales suelen involucrar grandes sumas asociadas a comercio exterior, inversiones o transferencias de alto valor.

En segundo lugar, métodos como Debit Card, ACH, Credit Card y Cheque mantienen montos promedio elevados y similares entre sí (cerca de los 9,000), lo que los convierte en categorías relevantes tanto por volumen total como por comportamiento unitario.

Finalmente, Cash Deposit y Cash Withdrawal tienen los montos promedio más bajos, reflejando operaciones de menor escala, aunque potencialmente repetitivas. Estos métodos pueden no ser significativos a nivel agregado, pero sí podrían tener importancia en análisis individuales o por tipología.

## Selección de variables

La selección de variables se basó en el análisis descriptivo previo, criterios de relevancia operativa en contextos financieros y la viabilidad técnica para el modelado predictivo. Se eligieron aquellas variables que presentaban una relación potencial con el comportamiento sospechoso y que podían capturar características clave del patrón transaccional.

Las variables seleccionadas fueron:

1. amount

Representa el monto de la transacción. Su distribución altamente sesgada y la presencia de valores extremos lo convierten en un indicador relevante para identificar operaciones inusuales. En contextos reales de lavado de dinero, los montos suelen jugar un papel crucial, ya sea por su valor absoluto o por su comportamiento atípico respecto al historial del cliente.

1. payment\_type

Método de pago utilizado. Se seleccionó por su valor informativo: algunos métodos están más asociados a grandes volúmenes o baja trazabilidad (como cash o transferencias internacionales), mientras que otros reflejan operaciones más regulares.

1. sender\_bank\_location

Ubicación del banco que emite la transacción. Se eligió como posible indicador geográfico de riesgo, ya que algunas jurisdicciones tienen mayor supervisión financiera que otras, lo cual puede influir en el comportamiento de las transacciones.

1. receiver\_bank\_location

Similar a la variable anterior, pero desde la perspectiva del destinatario. Al analizar la red de origen y destino, se puede tener una mejor comprensión del flujo de fondos y posibles rutas de evasión.

1. laundering\_type (solo informativa en el análisis de sospechosas)

Aunque no se usó como predictor en los modelos supervisados para evitar fugas de información, fue clave en el análisis exploratorio para entender qué tipos de patrones (fan-in, fan-out, estructuras circulares) estaban presentes en los casos sospechosos.

Las variables seleccionadas fueron codificadas numéricamente mediante Label Encoding para permitir su uso en modelos supervisados como Random Forest, XGBoost y Regresión Logística. No se aplicó reducción de dimensionalidad, ya que el número total de variables era manejable y no presentaba colinealidad significativa.

Esta selección permitió construir modelos robustos sin sobrecarga de información, manteniendo el enfoque en variables con sentido financiero, analítico y de riesgo.

## Selección de modelos

La elección de los modelos predictivos se realizó considerando las características del conjunto de datos, el desbalance de clases presente en el problema y la necesidad de contar con algoritmos que ofrecieran una buena capacidad de discriminación, especialmente para detectar transacciones sospechosas en un entorno financiero realista.

Los modelos seleccionados fueron los siguientes:

### Regresión Logística

Se utilizó como modelo base debido a su simplicidad, interpretabilidad y bajo costo computacional. Permite establecer una relación probabilística entre las variables predictoras y la probabilidad de que una transacción sea sospechosa. Aunque no captura relaciones no lineales, fue útil como punto de comparación con modelos más complejos.

### **Random Forest**

Este modelo de ensamble basado en árboles de decisión fue elegido por su robustez ante datos con variables categóricas codificadas, su capacidad de capturar relaciones no lineales y su buen rendimiento en problemas desbalanceados. Además, ofrece interpretabilidad a través de la importancia de variables y tiene una alta tolerancia al ruido.

### **XGBoost (Extreme Gradient Boosting)**

Se incorporó como modelo avanzado por su excelente desempeño en competiciones de ciencia de datos y su capacidad de manejo eficiente de grandes volúmenes de datos. XGBoost optimiza la función de pérdida mediante gradiente y regularización, lo que reduce el sobreajuste y mejora la precisión. Este modelo resultó ser el de **mejor desempeño global** según las métricas ROC AUC y matriz de confusión.

## Balanceo de Clases y Preparación del Dataset

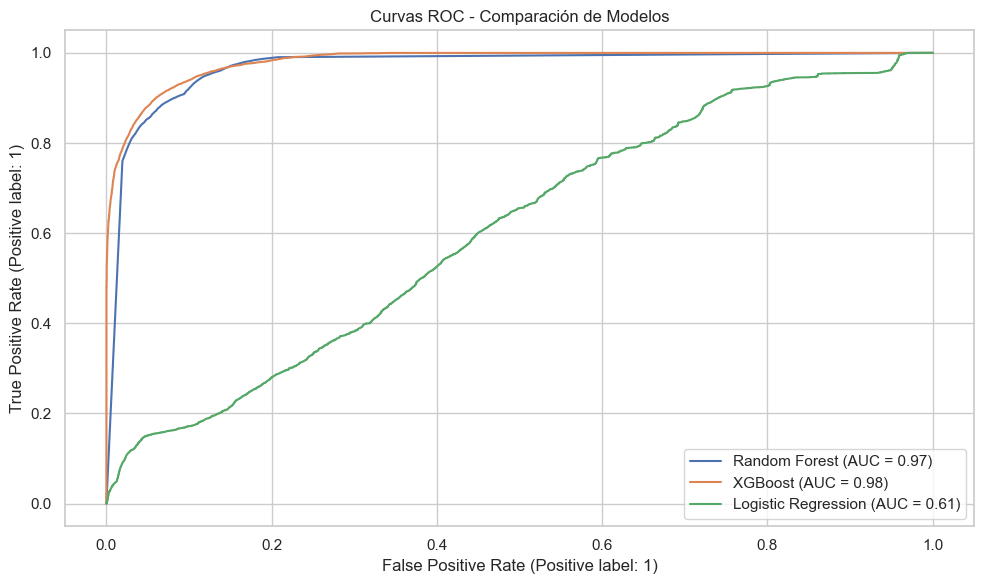
Dado el marcado desbalance en la variable objetivo (is\_laundering), se utilizó la técnica SMOTE (Synthetic Minority Over-sampling Technique) para crear ejemplos sintéticos de la clase minoritaria antes del entrenamiento. Esto permitió que los modelos no se sesgaran hacia la clase mayoritaria, mejorando así la sensibilidad del sistema de detección.

## Evaluación y Métricas

### Área bajo la curva ROC (AUC-ROC)

Permite medir la capacidad del modelo para distinguir entre clases. Cuanto mayor es el valor (cercano a 1.0), mejor es el modelo clasificando correctamente las transacciones sospechosas sin incurrir en falsos positivos.

Fue la métrica principal para comparar modelos.



La gráfica muestra las curvas ROC (Receiver Operating Characteristic) para los tres modelos entrenados: Random Forest, XGBoost y Regresión Logística, evaluados sobre el mismo conjunto de datos.

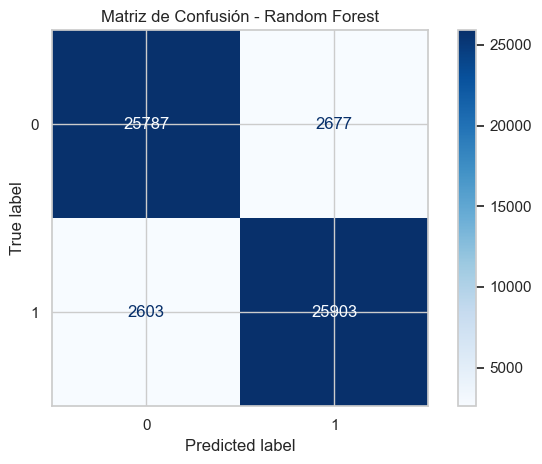
**Resultados:**

|  |  |  |
| --- | --- | --- |
| Modelo | AUC ROC | Interpretación |
| XGBoost | **0.98** | Excelente. Detecta transacciones sospechosas con alta sensibilidad y muy pocos falsos positivos. |
| Random Forest | **0.97** | Muy bueno. Solo ligeramente por debajo de XGBoost. Es una opción robusta y confiable. |
| Regresión Logística | **0.61** | Desempeño pobre. Apenas supera el azar (AUC = 0.5), lo que indica que no es adecuada para este problema tan complejo y no lineal. |

### Matriz de Confusión

Se utilizó para observar el número de verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN). Esto permitió identificar el tipo de error más común en cada modelo.

Matriz de Confusión de Random Forest



* TN (True Negatives) = 25,787

Transacciones legítimas correctamente clasificadas como no sospechosas.

* FP (False Positives) = 2,677

Transacciones legítimas que el modelo marcó incorrectamente como sospechosas.

* FN (False Negatives) = 2,603

Transacciones sospechosas que el modelo no detectó.

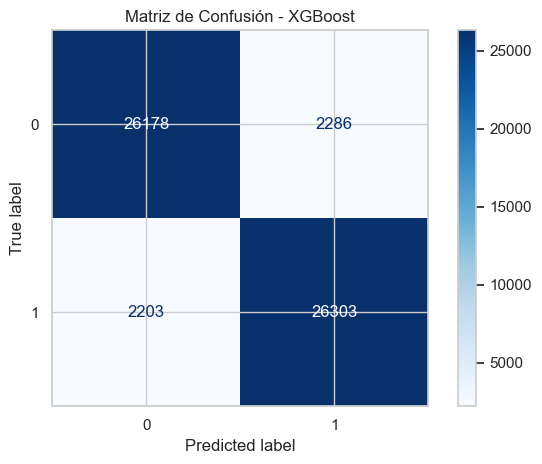
* TP (True Positives) = 25,903

Transacciones sospechosas correctamente identificadas.

**Análisis:**

* El modelo tiene un muy buen balance: identifica la gran mayoría de transacciones sospechosas (TP alto) y comete relativamente pocos errores (FN y FP bajos).
* La recall (sensibilidad) para la clase sospechosa es alta: detecta correctamente muchas de las transacciones de riesgo.
* La precisión también es buena: no marca demasiadas transacciones legítimas como fraude.
* El modelo es confiable y útil para priorizar investigaciones, minimizando tanto los falsos negativos como las falsas alarmas.

Matriz de Confusión de XGBoost



* TN (True Negatives) = 26,178

Transacciones legítimas correctamente clasificadas como normales.

* FP (False Positives) = 2,286

Transacciones legítimas que fueron incorrectamente clasificadas como sospechosas.

* FN (False Negatives) = 2,203

Transacciones sospechosas que el modelo no logró identificar.

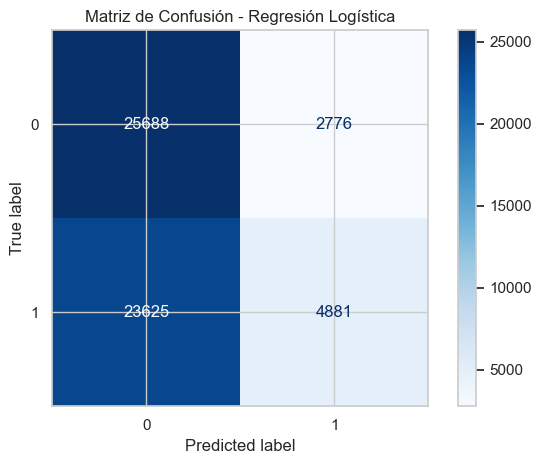
* TP (True Positives) = 26,303

Transacciones sospechosas correctamente detectadas.

**Análisis:**

* Mejor precisión que Random Forest: menos falsos positivos (2,286 vs. 2,677).
* Menor número de falsos negativos (FN): detecta más casos sospechosos correctamente.
* Mayor recall y precisión para la clase positiva (fraude).
* El modelo XGBoost logra el mejor equilibrio entre sensibilidad y especificidad entre los tres modelos probados.
* Este rendimiento lo convierte en un modelo ideal para sistemas de monitoreo de lavado de dinero, donde tanto el subdiagnóstico como el sobrediagnóstico tienen costos importantes.

Matriz de Confusión de Regresión Logística



* N (True Negatives) = 25,688

Transacciones normales correctamente clasificadas.

* FP (False Positives) = 2,776

Transacciones normales clasificadas erróneamente como sospechosas.

* FN (False Negatives) = 23,625

Transacciones sospechosas que el modelo NO detectó. Este es el error más grave en este contexto.

* TP (True Positives) = 4,881

Transacciones sospechosas correctamente detectadas.

**Análisis:**

* El modelo tiene una tasa muy alta de falsos negativos (FN). Esto significa que la mayoría de las transacciones sospechosas pasan desapercibidas, lo que lo hace poco confiable para la detección de lavado de dinero.
* Aunque la precisión sobre las transacciones normales (TN) es aceptable, el recall (sensibilidad) sobre la clase sospechosa es muy bajo.
* Esto concuerda con el AUC de apenas 0.61 visto en la curva ROC.

1. **F1-score** y **recall:**

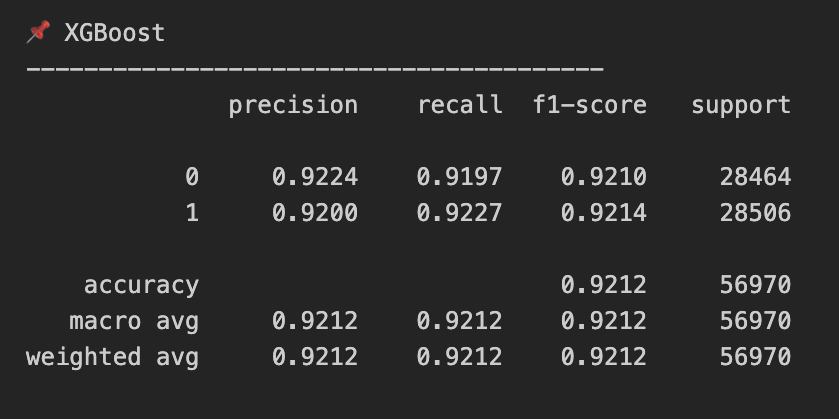
Dado que el problema del lavado de dinero presenta un alto desbalance de clases (menos del 0.11 % de las transacciones son sospechosas), fue necesario utilizar métricas que permitieran evaluar el rendimiento del modelo en la clase minoritaria (transacciones sospechosas).

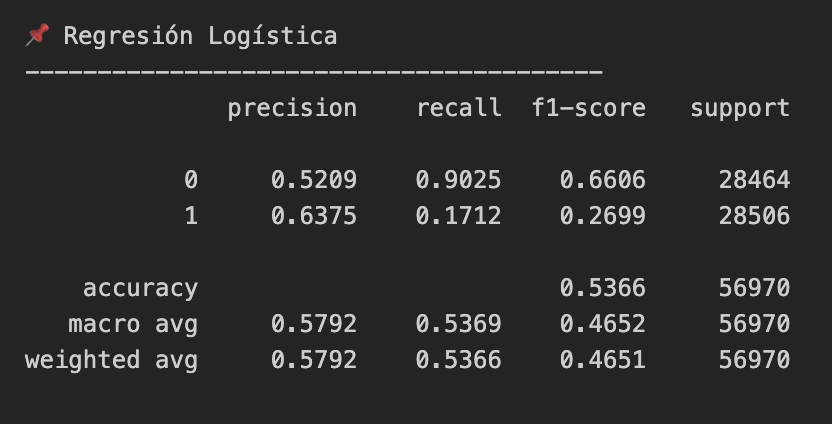
**Recall**

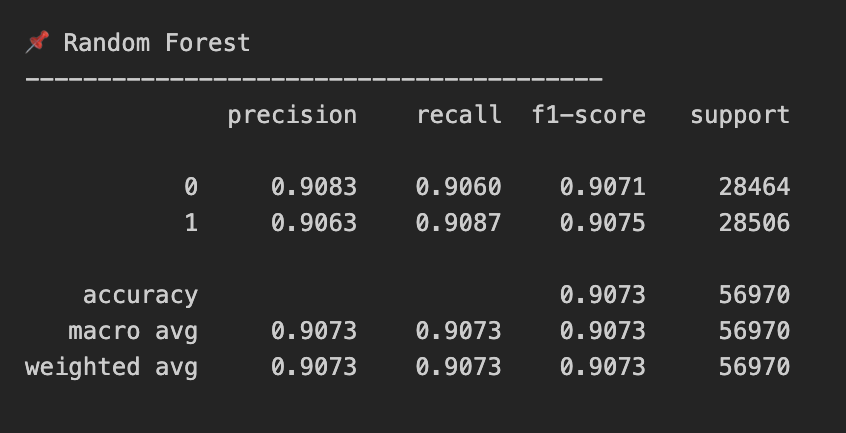
* Se utilizó para medir la capacidad del modelo de detectar correctamente las transacciones sospechosas.
* Un alto recall significa que el modelo logra encontrar la mayoría de los casos verdaderamente sospechosos, reduciendo el riesgo de que operaciones ilegales pasen desapercibidas (falsos negativos).
* Es clave en contextos de riesgo, donde no detectar un fraude puede tener consecuencias graves.

**F1-score**

* Se usó para encontrar un equilibrio entre precisión y recall.
* Fue importante porque no solo queremos detectar muchos casos (recall alto), sino también evitar generar falsas alarmas (falsos positivos).
* El F1-score permite evaluar el rendimiento general del modelo sobre la clase positiva, especialmente útil en escenarios desbalanceados.







**Resultados:**

XGBoost fue el mejor modelo en todas las métricas:

* Alta precisión, recall y F1-score tanto en clase positiva (sospechosa) como negativa.
* Mejor capacidad de detección sin comprometer la precisión.

Random Forest también tuvo un desempeño muy sólido, solo ligeramente por debajo de XGBoost.

* Es una excelente opción si se busca un modelo interpretativo y robusto.

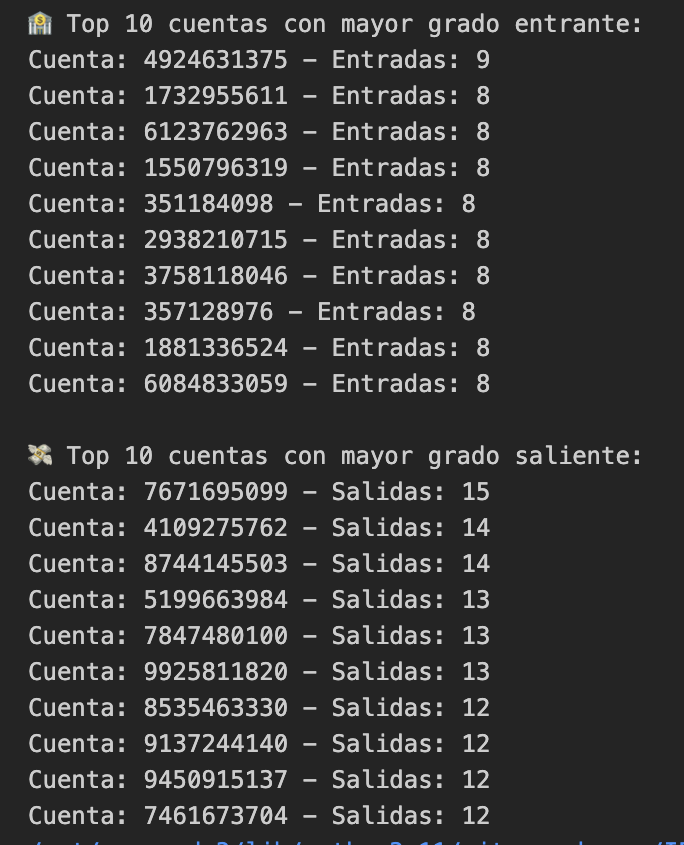
Regresión Logística presentó bajo rendimiento, especialmente en recall (0.17) y F1-score (0.26) en la clase 1.

* Esto significa que detectó muy pocas transacciones sospechosas, y no es adecuado para problemas de desbalance extremo como este.

1. **Análisis de Grafos:**

Como parte del análisis estructural, se construyó un grafo dirigido utilizando las relaciones entre cuentas emisoras (sender\_account) y receptoras (receiver\_account) en las transacciones. Esto permitió analizar el comportamiento de nodos (cuentas) dentro de la red financiera simulada.

Se calcularon los grados de entrada y salida para identificar nodos con alta centralidad, los cuales podrían tener un rol clave en posibles esquemas de lavado de dinero.



Top 10 cuentas con mayor grado entrante:

Estas cuentas recibieron múltiples transacciones desde distintas fuentes. Este patrón puede indicar una posible cuenta de acumulación o nodo central en esquemas de tipo fan-in:

Cuenta 4924631375 → 9 entradas

Cuentas como 1732955611, 6123762963, 1550796319, entre otras → 8 entradas cada una.

Top 10 cuentas con mayor grado saliente:

Estas cuentas enviaron fondos a muchas otras, lo que puede representar un comportamiento de dispersión, típico en esquemas fan-out:

Cuenta 7671695099 → 15 salidas

Cuentas como 4109275762, 8744145503 → 14 salidas

Varias más con 12 o 13 conexiones salientes

Se generó un subgrafo con las 20 cuentas más activas, visualizando sus conexiones directas. Este gráfico permitió detectar subredes densas y posibles patrones repetitivos o cíclicos que podrían representar rutas estructuradas para mover dinero entre múltiples cuentas.

