

Análisis de Transacciones Sospechosas (AML)

Soluciones Inteligentes para Retos Empresariales Reales



25 de julio de 2025

Equipo 12

Ayala Zavala, Damaris
Escobedo Valenzuela, Ana Karen
Esquivel Aragón, Francisco Javier
González Peña, Gustavo
Montaño Romero, Jonathan Alec
Narváez Mantilla, Efrén
Velázquez Domínguez, Felipe
Zayas Espejel, Rodrigo Iván

DESCRIPCIÓN DEL PROYECTO

Este proyecto se centra en la detección de **transacciones sospechosas de lavado de dinero** mediante análisis exploratorio y modelado predictivo usando la plataforma KNIME. Se trabajó con un conjunto de datos sintético realista, orientado a simular escenarios complejos de AML (Anti-Money Laundering).

SECTOR SELECCIONADO

- **Finanzas y Banca**
- Subsector: Prevención de Lavado de Dinero (AML, Anti Money Laundering).

CONTEXTO EMPRESARIAL

El lavado de dinero representa una amenaza grave para la integridad del sistema financiero global. A pesar de las regulaciones actuales, muchas instituciones bancarias siguen teniendo dificultades para detectar transacciones ilegales debido a:

- Limitaciones legales en el acceso a datos.
- Insuficiencia de etiquetas y tipologías reales.
- Modelos de monitoreo poco eficientes o desactualizados.

PROBLEMA DETECTADO

Las técnicas tradicionales de monitoreo transaccional resultan ineficientes para identificar patrones complejos de lavado de dinero, principalmente debido a la **baja frecuencia de transacciones etiquetadas como sospechosas** y a la **alta variabilidad en los métodos y tipologías utilizados por los delincuentes**. Esta situación dificulta el entrenamiento de modelos efectivos y retrasa la detección temprana de actividades ilícitas, poniendo en riesgo la integridad de los sistemas financieros.

SOLUCIÓN PROPUESTA

Dataset utilizado:

- **SAML-D (Synthetic AML Dataset).**
- Contiene más de **9.5 millones de transacciones**, de las cuales solo el **0.1039%** son etiquetadas como sospechosas.
- Aporta **28 tipologías de transacción** (11 normales, 17 sospechosas), inspiradas en entrevistas con especialistas AML y fuentes académicas.

(Consulta la [presentación visual en Canva](#) y accede al [dataset original en Kaggle](#).)

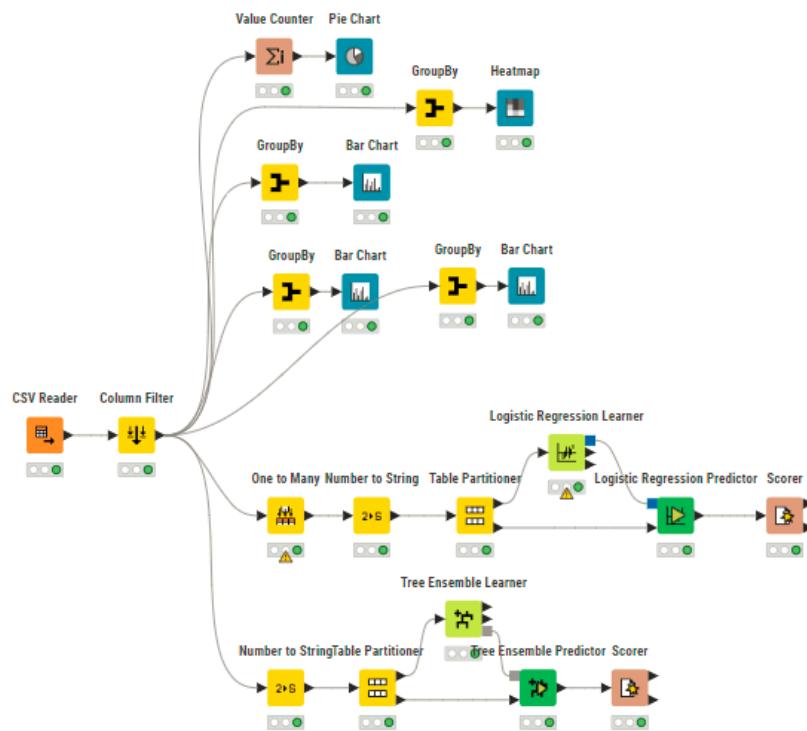
Herramientas aplicadas:

- **Python** (datos preprocesados con Pandas)
- **KNIME Analytics Platform** (enfoque low-code / no-code).

Enfoque metodológico:

- Análisis Exploratorio de Datos (EDA)
- Modelado supervisado de clasificación (Regresión Logística y Tree Ensemble)

FLUJO DE TRABAJO EN KNIME



Fase 1: Preparación y limpieza

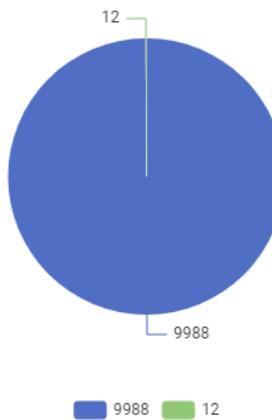
- **CSV Reader**: carga de datos preprocesados con python (10,000 transacciones aleatorias)
- **Column Filter**: eliminación de columnas irrelevantes (Time, Date, Sender_account, Receiver_account)
- **Number to String**: corrección de tipos para algoritmos

Fase 2: EDA (Exploración de Datos)

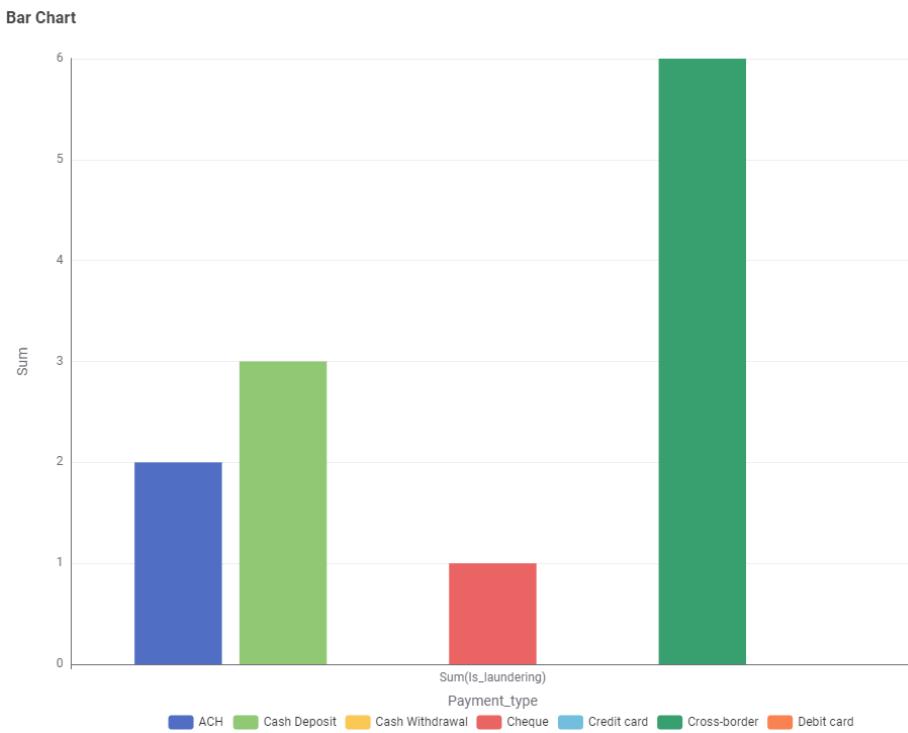
Variables clave analizadas:

- Amount
 - Payment_currency
 - Sender_bank_location
 - Receiver_bank_location
 - Payment_type
 - Laundering_type
 - Is_Laundering
-
- **Value Counter**: conteo de clases Is_laundering (Legales 9988, Sospechosas 12).

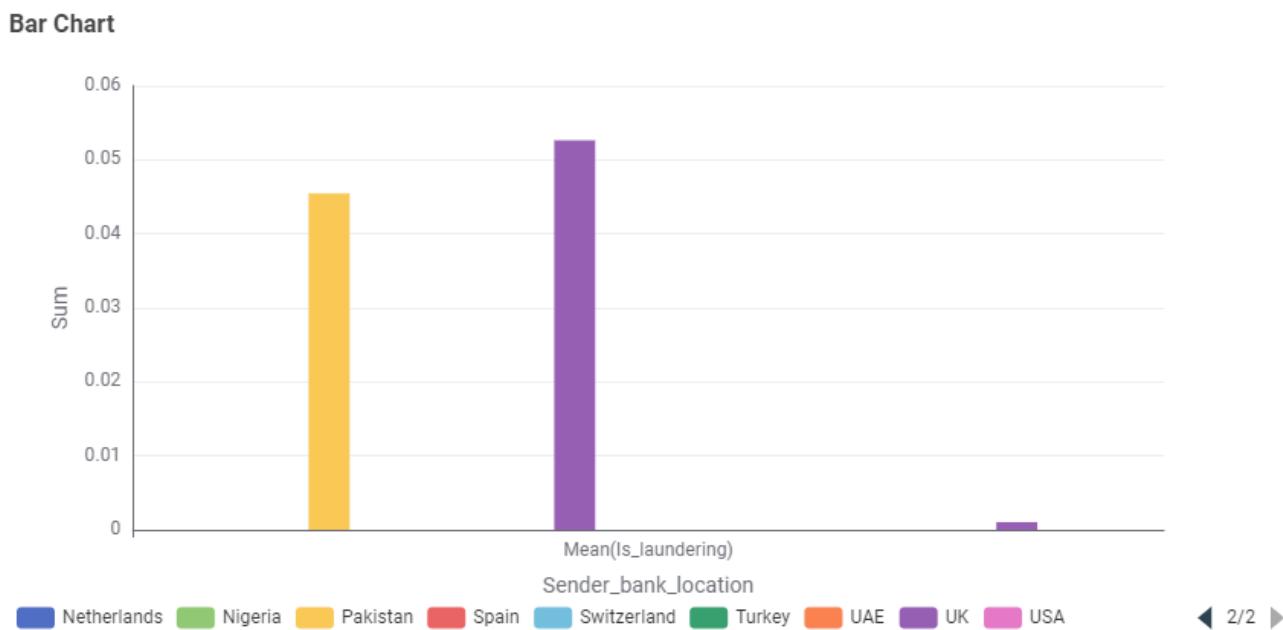
Pie Chart



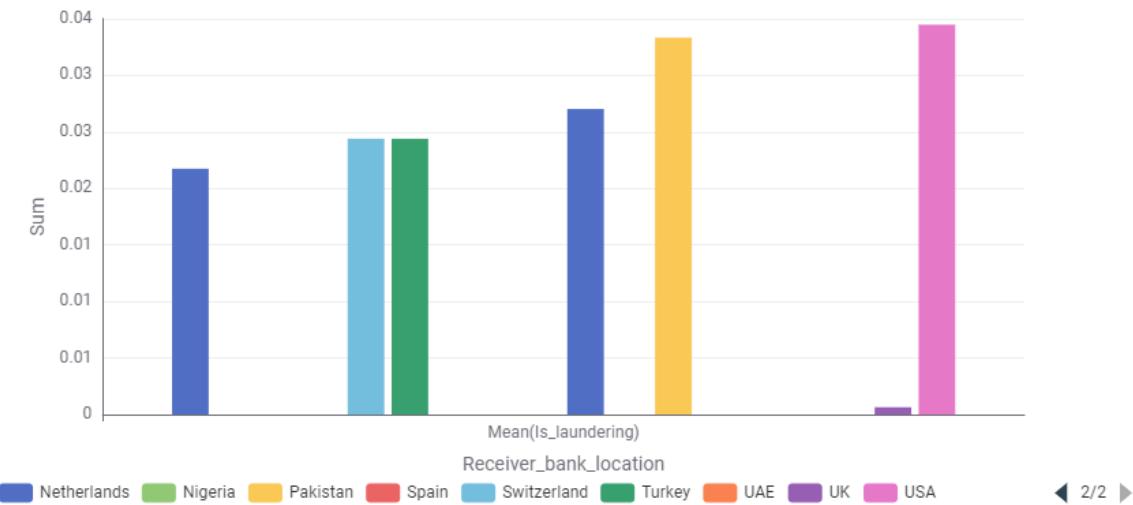
- **Bar Chart, Pie Chart:** visualización de distribución por país y tipo de pago.



- **GroupBy:** análisis por ubicación emisor-receptor.

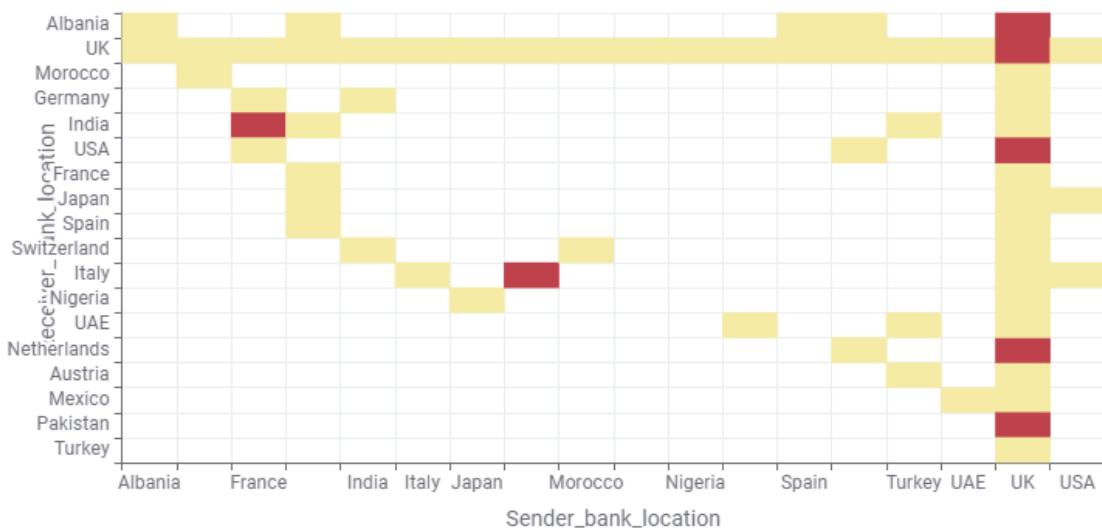


Bar Chart



- **Heatmap:** detección de combinaciones geográficas con mayor riesgo.

Heatmap



Fase 3: Modelado Supervisado

A. Regresión Logística

- **One to Many:** codificación de variables categóricas
- **Table Partitioning:** 70% entrenamiento / 30% prueba
- **Logistic Regression Learner** y **Predictor**
- **Scorer:** métricas de evaluación

B. Tree Ensemble (Random Forest)

- **Table Partitioning:** 70% entrenamiento / 30% prueba
- **Tree Ensemble Learner** y **Predictor**
- **Scorer:** análisis de desempeño

RESULTADOS OBTENIDOS

Interpretación de resultados Regresión Logística

- Excelente rendimiento general (accuracy $\approx 99.99\%$).
- Detecta el 60% de los fraudes reales (recall = 0.6), sin falsos positivos (precisión = 1.0).
- Buen F1 score (0.75) pero menor que modelos tipo árbol.

Accuracy statistics (Table)													
	RowID	TruePositive Number (Inte...)	FalsePositive Number (Inte...)	TrueNegative Number (Inte...)	FalseNegative Number (Inte...)	Recall Number (Floa...)	Precision Number (Floa...)	Sensitivity Number (Floa...)	Specificity Number (Floa...)	F-measure Number (Floa...)	Accuracy Number (Floa...)	Cohen's kappa Number (Floa...)	
□	1	0	2995	2	3	0	1	0.999	1	0.6	1	0.75	
□	2	1	3	0	2995	2	0.6	1	0.6	1	0.75	0.75	
□	3	Overall	?	?	?	?	?	?	?	?	0.999	0.75	

Interpretación de resultados Tree Ensemble (Random Forest)

- Mejor rendimiento que regresión logística.
- Detecta el 80% de los fraudes reales.
- Alto F1 score (0.88) y excelente Cohen's Kappa (≈ 0.889).
- Recomendado para este caso por su robustez y capacidad de interpretar relaciones complejas.

Accuracy statistics (Table)													
	RowID	TruePositive Number (Inte...)	FalsePositive Number (Inte...)	TrueNegative Number (Inte...)	FalseNegative Number (Inte...)	Recall Number (Floa...)	Precision Number (Floa...)	Sensitivity Number (Floa...)	Specificity Number (Floa...)	F-measure Number (Floa...)	Accuracy Number (Floa...)	Cohen's kappa Number (Floa...)	
□	1	0	2995	1	4	0	1	1	0.8	1	0.889	0.889	
□	2	1	4	0	2995	1	0.8	1	0.8	1	0.889	0.889	
□	3	Overall	?	?	?	?	?	?	?	?	1	0.889	

IMPACTO ESPERADO

- Mejora en la detección temprana de transacciones sospechosas mediante modelos automatizados con alta precisión.
- Mayor confiabilidad y respaldo analítico para los equipos de cumplimiento normativo (*compliance*), permitiéndoles tomar decisiones mejor informadas.
- Potencial para extenderse en sistemas de monitoreo en tiempo real, integrándose a plataformas bancarias existentes.
- Contribución al desarrollo de sistemas antifraude open-source, mediante el uso de datos sintéticos accesibles como el dataset SAML-D.
- Cumplimiento regulatorio proactivo, alineado con normas internacionales como la Ley Antilavado, la GAFI (FATF), y estándares del sector financiero.

CONCLUSIÓN

Este proyecto demuestra que es posible predecir transacciones sospechosas de lavado de dinero mediante modelos de clasificación aplicados a datos estructurados. Usando KNIME como herramienta principal, se construyó un flujo de análisis claro, eficiente y sin necesidad de programar, complementado en algunos puntos con Python.

Los modelos alcanzaron altos niveles de precisión y sensibilidad, especialmente el Tree Ensemble, facilitando su aplicación en entornos reales. Además, se identificaron patrones geográficos y operativos relevantes que pueden apoyar políticas de prevención más efectivas.

En conjunto, el flujo desarrollado representa una solución escalable, interpretable y útil para fortalecer los sistemas de monitoreo en instituciones financieras.