PROYECTO "OBSERVATORIO LAFT: ANALÍTICA DE DATOS PARA LA PREVENCIÓN DE LAFT"

Equipo 22 – Aprendizaje No Supervisado

Descripción breve

- Este informe se enfoca en la exploración y preparación de datos financieros para su uso en modelos de aprendizaje no supervisado, la selección preliminar de algoritmos de clustering, y el desarrollo y calibración de modelos para identificar patrones anómalos relacionados con LAFT, junto con la propuesta metodológica como parte de Documento con propuesta inicial.

INTEGRANTES

- Beraún Chamorro, Christian
- Gomez Fernández, Diana Ximena
- Molina Marriott, Alberto Alejandro
- Zambrano Franco, Marlon Jose

1. RESUMEN

El proyecto aborda el desafío de clasificar a los clientes de una entidad financiera que ofrece billeteras digitales, con el objetivo de cumplir con las regulaciones destinadas a prevenir el Lavado de Activos y la Financiación del Terrorismo (LAFT). Estas regulaciones requieren que los clientes sean agrupados en función de diversos factores de riesgo, de modo que cada grupo sea homogéneo interna y claramente diferenciado de los otros. Este proceso es esencial para identificar comportamientos inusuales en las transacciones.

Para lograrlo, se aplicarán técnicas de análisis de datos avanzadas, específicamente el aprendizaje no supervisado, utilizando el algoritmo de *K-means clustering*. La metodología seguirá los siguientes pasos: primero, se realizará un preprocesamiento de los datos transaccionales y de cuentas proporcionados por la entidad, lo que incluirá la limpieza de datos, la imputación de valores faltantes y la normalización de las variables. A continuación, se aplicará el algoritmo de *K-means* para agrupar a los clientes en clústeres basados en la similitud de sus características, como el saldo total, la frecuencia de transacciones, y la antigüedad de la cuenta. Finalmente, se evaluará la calidad de los clústeres mediante métodos como el análisis del codo y la silueta, para asegurar que el modelo segmenta efectivamente a los clientes según su nivel de riesgo.

El principal resultado esperado de este trabajo es el desarrollo de un modelo que permita a la entidad financiera detectar de manera temprana y eficaz posibles riesgos, garantizando así el cumplimiento normativo y mejorando la seguridad de sus operaciones. Este enfoque no solo clasificará a los clientes utilizando *K-means*, sino que también permitirá una evaluación continua del comportamiento de las cuentas, asegurando una adaptación dinámica frente a nuevas amenazas y proporcionando una protección constante contra los riesgos asociados al LAFT.

2. INTRODUCCIÓN

¿Cómo puede una entidad financiera que ofrece billeteras digitales segmentar eficazmente a sus clientes para cumplir con las regulaciones de prevención de Lavado de Activos y la Financiación del Terrorismo (LAFT)? Esta pregunta surge no solo por la necesidad de adherirse a las normativas establecidas por la Superintendencia Financiera de Colombia (SFC), sino también debido al contexto histórico y social en el que estas regulaciones operan. En las últimas décadas, Colombia ha enfrentado graves desafíos relacionados con el lavado de activos y la financiación del terrorismo, fenómenos que han evolucionado en complejidad junto con las técnicas empleadas para infiltrarse en el sistema financiero.

En respuesta a estos desafíos, las entidades financieras buscan continuamente optimizar sus sistemas de prevención, adaptándose a los nuevos medios de pago que han emergido con el avance tecnológico. Un ejemplo de esta adaptación es la implementación del Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT) en Fintech y billeteras digitales. En este contexto, el presente proyecto tiene como objetivo desarrollar un modelo de segmentación que permita a una entidad financiera identificar de manera temprana y eficiente posibles riesgos LAFT, garantizando así el cumplimiento normativo y fortaleciendo la seguridad financiera.

El cliente potencial de este proyecto es una entidad financiera que ofrece productos de billeteras digitales y que debe cumplir estrictamente con las normativas LAFT. Dado que los medios de pago digitales están en constante evolución, la segmentación precisa de los clientes es crucial para mitigar los riesgos asociados al LAFT. Este proyecto se basa en el uso de técnicas de aprendizaje no supervisado, específicamente en la segmentación de datos, para identificar patrones de comportamiento similares entre los clientes. Este enfoque no solo facilitará la detección de transacciones inusuales, sino que también permitirá categorizar el riesgo asociado a cada cliente, ayudando a la entidad a mejorar sus procesos de cumplimiento normativo y a mantener su estabilidad y reputación en un entorno financiero en constante cambio.

3. REVISIÓN PRELIMINAR DE LA LITERATURA

Se exploraron estudios que abordan la detección de fraudes bancarios y el lavado de activos mediante técnicas avanzadas de análisis de datos y aprendizaje automático. A nivel internacional, Alexandre y Balsa (2016) presentan un sistema multiagente para la detección de lavado de dinero, que utiliza *clustering* y minería de datos para perfilar clientes y generar reglas de clasificación aplicables a operaciones financieras sospechosas. Este enfoque destaca por su arquitectura distribuida y la automatización de tareas críticas en el proceso de detección.

Por otro lado, a nivel nacional, Rangel Quiñonez, Barrera Gómez y Gómez Sánchez (2021) desarrollan un modelo de segmentación basado en el factor de jurisdicción para clasificar el riesgo de lavado de activos y financiación del terrorismo en los municipios de Colombia. Utilizando técnicas de análisis de conglomerados y estadísticas multivariadas, logran identificar ciudades con mayor exposición a estos riesgos. Este enfoque es particularmente relevante debido a su aplicación en el contexto colombiano, donde el lavado de activos ha tenido un impacto significativo en la estabilidad económica y social.

Aunque ambos enfoques comparten el uso de técnicas de agrupamiento para la detección de actividades ilícitas, difieren en la aplicación práctica y el contexto geográfico. Mientras que el trabajo de Alexandre y Balsa (2016) se centra en la integración de sistemas multiagente para mejorar la eficiencia en la detección de fraudes a nivel institucional, el estudio de Rangel Quiñonez et al. (2021) se enfoca en la evaluación del riesgo a nivel municipal en Colombia, lo que refleja la adaptabilidad de estas técnicas a diferentes escalas y contextos.

Este análisis comparativo proporciona una base sólida para justificar la relevancia y novedad del enfoque propuesto en este trabajo, destacando las posibles mejoras en la precisión y eficiencia en la detección de actividades ilícitas en entornos financieros complejos como lo son las Fintech, debido a su auge y a la naturaleza de los productos financieros que brindan.

4. DESCRIPCIÓN DE LOS DATOS

Como parte del proyecto, la entidad financiera “Confidencial” proporcionó dos conjuntos de datos en formato separado por comas:

- BASE_CUENTAS: que expresa de manera única la caracterización de las cuentas. Se compone inicialmente de 159,665 filas y 10 columnas.
- BASE_TRX: que expresa de manera transaccional cada ingreso (consignación) o egreso de la cuenta. Se compone inicialmente de 1,220,698 filas y 14 columnas.

Preprocesamiento:

Se establecieron criterios para obtener un solo Dataset que permita hacer un análisis de cuentas y transferencias de cara al problema. Estos criterios fueron:

1. Se analizan las transacciones con una antigüedad de un año, específicamente en el periodo 30 de junio 2023 al 30 de junio 2024.
2. Se analiza el producto financiero “PR001”.
3. Se analiza los valores faltantes en todos los campos con excepción de Jurisdicción_usuario, ciudad_corresponsal, y departamento_corresponsal ya que no impacta en la evaluación.
4. Se realizó un agrupamiento tomando como llave primaria el tipo y número de identificación de tal manera que en el *dataframe* agrupado se consideran las siguientes variables: frecuencia y proporción de retiros, consignaciones, promedios de transacciones por cada tipo de canal.

Dataset resultante:

BASE_TRX: que expresa de por cada cliente, las variables de interés que permitan identificar grupos de clientes con comportamientos similares, facilitando la detección de transacciones inusuales y la categorización del riesgo asociado a cada cliente. Se compone inicialmente de 27,362 filas (clientes) y 24 columnas.

A partir de este Dataset resultante, se realiza el análisis descriptivo:

Tabla 1

Tipo de datos

N	Variable	Tipo	N	Variable	Tipo
1	Negocio	object	13	prop_consignacion	float64
2	Saldo total	float64	14	prop_retiro	float64
3	Saldo disponible	float64	15	promedio_valor_consignacion	float64
4	Acumulado retiro	float64	16	promedio_valor_retiro	float64
5	Ciudad	object	17	AP	float64
6	Departamento	object	18	BS	float64
7	antiguedad_cliente_dias	int64	19	DF	float64
8	total_transacciones	int64	20	EM	float64
9	corresponsales_diferentes	int64	21	PD	float64
10	usuarios_diferentes	int64	22	PS	float64
11	cantidad_consignacion	int64	23	RM	float64
12	cantidad_retiro	int64	24	TI	float64

Nota. Esta tabla muestra las veinticuatro (24) variables y su tipología.

Dentro de las 24 variables listadas, se identifican tres variables de tipo object (categóricas): “Negocio”, “Ciudad” y “Departamento”, que representan descripciones textuales de negocios y ubicaciones geográficas. El resto de las variables son numéricas, distribuyéndose en dos tipos: int64, que incluye variables como antiguedad_cliente_dias, total_transacciones, corresponsales_diferentes, usuarios_diferentes, cantidad_consignacion y cantidad_retiro, y float64, que abarca variables relacionadas con saldos, proporciones y promedios, como “Saldo total”, prop_consignacion, promedio_valor_consignacion, entre otras. Las variables categóricas pueden utilizarse para segmentar los datos, mientras que las variables numéricas permiten la evaluación de métricas financieras y así realizar el análisis de transacciones y evaluar el comportamiento financiero de clientes.

Tabla 2

Análisis descriptivo

N	VAR	count	unique	top	freq	mean	std	min	25%	50%	75%	max
1	Negocio	27362	9	WW005	26053	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Saldo total	27362	NaN	NaN	NaN	162100.8	593671.44	0	0	100	3500	10156685
3	Saldo disponible	27362	NaN	NaN	NaN	53479.1	283672.33	0	0	0	500	10156685
4	Acumulado retiro	27362	NaN	NaN	NaN	146884.32	454872.36	0	0	0	0	9824955.1
5	Ciudad	27362	458	Bogotá	19256	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	Departamento	27362	38	Bogotá	19281	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	antiguedad_cliente_dias	27362	NaN	NaN	NaN	420.57	255	0	252	368	589	2036
8	total_transacciones	27362	NaN	NaN	NaN	20.21	34.48	1	4	7	20	477
9	corresponsales_diferentes	27362	NaN	NaN	NaN	1.86	3.47	0	0	1	2	53
10	usuarios_diferentes	27362	NaN	NaN	NaN	2.9	3.32	0	1	2	3	72
11	cantidad_consignacion	27362	NaN	NaN	NaN	8.04	13.59	0	2	2	8	289
12	cantidad_retiro	27362	NaN	NaN	NaN	12.17	22.71	0	2	4	13	328
13	prop_consignacion	27362	NaN	NaN	NaN	0.43	0.17	0	0.33	0.44	0.5	1
14	prop_retiro	27362	NaN	NaN	NaN	0.57	0.17	0	0.5	0.56	0.67	1
15	promedio_valor_consignacion	27362	NaN	NaN	NaN	374581.21	576091.74	0	115000	282000	503750	9952800
16	promedio_valor_retiro	27362	NaN	NaN	NaN	273874.79	496221.22	0	92500	180000	330000	9906488
17	AP	27362	NaN	NaN	NaN	85654.51	247879.99	0	0	0	89856.25	9000000
18	BS	27362	NaN	NaN	NaN	35038.86	426142.94	0	0	0	0	9906488
19	DF	27362	NaN	NaN	NaN	12209.91	183529.57	0	0	0	0	10052189
20	EM	27362	NaN	NaN	NaN	113001.31	255817.99	0	0	0	0	4961534
21	PD	27362	NaN	NaN	NaN	148118.43	218707.33	0	0	6000	252993.42	7000000
22	PS	27362	NaN	NaN	NaN	165230.49	561826.49	0	0	0	169854.47	9952800
23	RM	27362	NaN	NaN	NaN	1727.82	51989.63	0	0	0	0	4298888
24	TI	27362	NaN	NaN	NaN	203815.78	302946.95	0	0	114000	288839.7	9216135.4

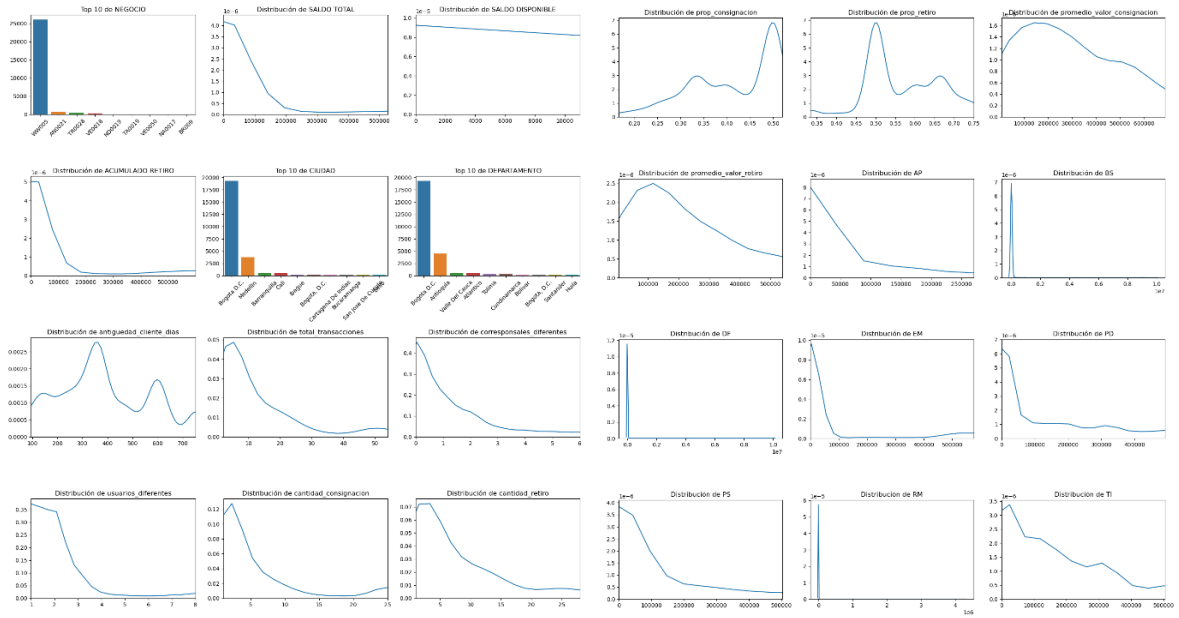
Nota. Esta tabla muestra las veinticuatro (24) variables y su tipología.

La variable “Negocio” es categórica, con 27,362 observaciones y 9 categorías únicas, destacando el negocio “WW005” como el más frecuente y representando la mayoría de los registros (26,053 observaciones). Las variables numéricas, como “Saldo total”, “Saldo disponible” y “Acumulado retiro”, presentan grandes variaciones en sus valores, evidenciando una alta dispersión con desviaciones estándar elevadas y valores máximos que alcanzan cifras multimillonarias, lo que sugiere la existencia de *outliers* significativos. Las variables “Ciudad” y “Departamento” son también categóricas, siendo Bogotá D.C. como la categoría más frecuente, lo que podría indicar una concentración de datos en esta región. Las métricas de transacciones y actividades financieras, como total_transacciones, corresponsales_diferentes, y usuarios_diferentes, reflejan una tendencia central hacia valores bajos, pero con una dispersión considerable que sugiere variabilidad en el comportamiento de los usuarios.

Las proporciones de consignación y retiro (prop_consignacion y prop_retiro) muestran una distribución balanceada alrededor de sus medias, mientras que los valores promedio de consignación y retiro presentan rangos amplios, destacando nuevamente la presencia de transacciones de alto valor. Finalmente, las variables AP, BS, DF, entre otras, presentan características similares, con medias relativamente bajas pero con valores máximos extremadamente altos, reforzando la idea de una alta variabilidad y la posible influencia de casos atípicos en el conjunto de datos.

Figura 1

Distribución de las variables N1 A N24



Nota. Este gráfico presenta veinticuatro (24) mini gráficos que representan la distribución de todas las variables.

El análisis general de los gráficos revela que las variables del *dataset*, en su mayoría relacionadas con métricas financieras y transacciones, presentan distribuciones sesgadas hacia la derecha. Esto indica que mientras la mayoría de los valores son bajos, existe una minoría de casos con valores extremadamente altos, lo que es típico en datos financieros donde unos pocos individuos o transacciones concentran una parte significativa de los recursos.

Además, algunas variables muestran distribuciones multimodales, lo que sugiere la existencia de subgrupos dentro del dataset con comportamientos diferenciados. Variables como prop_consignacion y prop_retiro ilustran esto, indicando que los clientes se agrupan en torno a distintas proporciones de sus transacciones. También se destaca la fuerte concentración en ciertos valores categóricos, como en “Negocio”, donde una entidad específica domina el conjunto de datos.

Finalmente, las variables relacionadas con la ubicación geográfica, como “Ciudad” y “Departamento”, muestran una dominancia clara de Bogotá D.C., lo que sugiere una concentración de las actividades económicas en esta región. Este análisis subraya la importancia de considerar la variabilidad y la presencia de outliers al realizar inferencias y tomar decisiones basadas en estos datos.

5. PROPUESTA METODOLÓGICA

En este proyecto, planeamos incorporar técnicas de aprendizaje no supervisado para descubrir patrones ocultos y estructuras subyacentes en nuestros datos financieros y transaccionales. Dado que los datos no están etiquetados, el enfoque no supervisado permitirá explorar las relaciones entre las variables y segmentar el conjunto de datos de manera efectiva, sin necesidad de un conocimiento previo exhaustivo sobre la distribución de las categorías.

Figura 2

Propuesta Metodológica

Nota. Este gráfico presenta seis (6) etapas que permiten abordar el problema analítico para la propuesta de solución.

El algoritmo preliminar que planeamos utilizar es ***K-means clustering***. Este método es ampliamente utilizado en el aprendizaje no supervisado debido a su capacidad para agrupar datos en función de la similitud entre puntos, creando clústeres que pueden revelar segmentos naturales dentro del conjunto de datos. K-means es particularmente útil para nuestro caso, donde deseamos identificar diferentes comportamientos financieros entre los clientes, agrupándolos en clústeres basados en características como el saldo total, la cantidad de transacciones, la antigüedad del cliente, entre otros. A través de este algoritmo, esperamos poder identificar grupos de clientes con comportamientos similares, lo que podría ser crucial para la creación de estrategias personalizadas de marketing o para la detección de perfiles de riesgo.

Sin embargo, reconocemos que este es un enfoque preliminar. A medida que avancemos en el curso y aprendamos nuevas herramientas y técnicas, es posible que modifiquemos nuestro enfoque, considerando otros algoritmos de clustering como ***DBSCAN*** o ***Gaussian Mixture Models (GMMs)***, que podrían ofrecer ventajas en la detección de patrones más complejos o en la gestión de datos con distribuciones más irregulares. Además, evaluaremos la idoneidad de la cantidad de clústeres mediante métodos como el análisis del codo (*Elbow Method*) y la silueta, para asegurar que nuestro modelo capture la estructura real de los datos de la manera más efectiva posible. Toda esta evaluación se realizará en el ajuste del modelo y la presentación de hallazgos y conclusiones.

6. BIBLIOGRAFIA Y REFERENCIAS

Rangel Quiñonez, H. S., Barrera Gómez, G., & Gómez Sánchez, O. M. (2021). Clasificación del riesgo de lavado de activos y financiación del terrorismo en Colombia en 2019. *Cuadernos De Contabilidad*, *22*. https://doi.org/10.11144/Javeriana.cc22.crla

Lo, S.-C., & Li, T.-S. (2016). Using big data analytics for money laundering detection: A case study. *Unpublished manuscript*. Retrieved from https://www.researchgate.net/publication/369854750_Using_Big_Data_Analytics_for_Money_Laundering_Detection_-A_Case_Study

Singh, K., & Best, P. (2019). Anti-money laundering: Using data visualization to identify suspicious activity. *International Journal of Accounting Information Systems*, *34*, 100418. https://doi.org/10.1016/j.accinf.2019.06.001

Ameijeiras Sánchez, D., Valdés Suárez, O., & González Diez, H. (2021). Algoritmos de detección de anomalías con redes profundas: Revisión para detección de fraudes bancarios. *Revista Cubana de Ciencias Informáticas*, *15*(4, Supl. 1), 244-264. Advance online publication. Retrieved on September 1, 2024, from http://ref.scielo.org/v7rrcd