ENTREGA FINAL DEL 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 colombiana que ofrece billeteras digitales, con el fin de cumplir con las regulaciones de prevención de Lavado de Activos y Financiación del Terrorismo (LAFT). Estas normativas exigen que los clientes sean agrupados en función de factores de riesgo, de manera que los grupos resultantes sean homogéneos interna y claramente diferenciados entre sí. Este proceso es clave para identificar comportamientos inusuales en las transacciones financieras.

Para alcanzar este objetivo, se implementarán técnicas avanzadas de análisis de datos, con un enfoque en el aprendizaje no supervisado, empleando el algoritmo DBSCAN. El proyecto seguirá una metodología estructurada que comienza con el preprocesamiento de los datos transaccionales y de cuentas proporcionados por la entidad, lo que incluirá la limpieza, imputación de valores faltantes y normalización de variables. Posteriormente, se aplicará DBSCAN para identificar clústeres basados en densidades, lo que permitirá detectar grupos de clientes con patrones de comportamiento similares, excluyendo a aquellos considerados "ruido" por no ajustarse a ningún grupo definido.

La elección de la segmentación adecuada se realizará en base a la caracterización de los grupos encontrados, donde el resultado esperado es encontrar distintos grupos de clientes según su naturaliza transaccional y un grupo de clientes con características potencialmente sospechosas en materia de lavado de activos y fraude. El resultado principal de este trabajo será un modelo que permita a la entidad financiera identificar de manera proactiva posibles riesgos y cumplir con las normativas de manera eficiente, asegurando además una protección dinámica y continua frente a nuevas amenazas relacionadas con el 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. MATERIALES Y MÉTODOS

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 (159,665 filas y 10 columnas), y BASE_TRX, que expresa de manera transaccional cada ingreso (consignación) o egreso de la cuenta (1,220,698 filas y 14 columnas).

Preprocesamiento:

Con el objetivo de capturar el comportamiento transaccional de los clientes de la entidad financiera, se realizó un exhaustivo preprocesamiento de los datos proporcionados, que resultó en un dataset consolidado a nivel de cliente, agrupando toda la información financiera relevante. En primer lugar, se seleccionaron las transacciones del último año, comprendidas entre el 30 de junio de 2023 y el 30 de junio de 2024. En segundo lugar, se filtró el producto financiero "PR001", ya que es el de mayor interés por representar las transferencias de dinero entre clientes. A continuación, se descartaron las variables relacionadas con la jurisdicción del usuario, la ciudad y el departamento del corresponsal, debido a que en su mayoría contenían valores faltantes. No obstante, en las variables numéricas restantes, los valores perdidos se imputaron como ceros, ya que su ausencia indicaba que el cliente no realizó transacciones o retiros en el periodo analizado (clientes inactivos). Finalmente, los datos se agruparon utilizando como clave el tipo y número de identificación del cliente, incorporando variables como la frecuencia y proporción de retiros y consignaciones, así como los promedios de transacciones por tipo de canal.

Dataset resultante:

Luego de realizar el preprocesamiento de datos, se obtuvo el dataset final, compuesto por un total de 27,362 filas (clientes) y 29 columnas (características), las cuales se enlistan en la tabla adjunta en el anexo n.° 1. Estas variables fueron construidas para extraer el comportamiento transaccional de los clientes y así agruparlos e identificar **comportamientos atípicos** relacionados con lavado de activos y financiación del terrorismo.

Aunque las variables categóricas como *Negocio, Ciudad y Departamento* no se usarán en la segmentación por su baja interpretabilidad o calidad, se incluyen para complementar la caracterización de los grupos. Se analizaron variables como Saldo total y Saldo disponible para conocer la cantidad de dinero que posee el cliente, observándose que muchos tienen $0 COP en sus cuentas y que el saldo disponible suele ser menor que el saldo total. Variables como *Acumulado retiro, total_transacciones, cantidad_consignacion, cantidad_retiro, prop_consignacion, prop_retiro, promedio_valor_consignacion y promedio_valor_retiro* se incluyeron para perfilar el comportamiento transaccional, mostrando concentraciones en valores bajos y medianas cercanas a 0, aunque una minoría realiza transacciones de alto valor (hasta 10 millones COP). Estudios previos, como Zhang et al. (2022), indican que clientes fraudulentos presentan valores atípicos en estas variables. Además, se consideraron variables recomendadas por expertos, como *antiguedad_cliente_dias* para cuantificar la relación con la institución, y *corresponsales_diferentes* y *usuarios_diferentes* para representar la diversidad de interacciones, esperando que clientes fraudulentos tengan números atípicos en estas variables. Finalmente, se incluyeron variables de comportamiento transaccional por canal: App (AP), Bank as a Service (BS), Distribución de fondos (DF), Portal Empresas (EM), Corresponsal no bancario (PD), Pasarela de pago web (PS), Remesas (RM) y Traslados internos (TI), que muestran valores de $0 COP hasta la mediana, ya que los clientes no suelen usar todos los canales. Debido a la correlación entre variables, se realizará un análisis de componentes principales (PCA) para generar componentes no correlacionados útiles para la segmentación.

Algoritmo utilizado:

Como se mencionó anteriormente, se utilizó Análisis de Componentes Principales (PCA) para abordar el problema de la correlación esperada entre las variables debido a que fueron construidas a partir de la misma fuente de información. El PCA es una técnica estadística que permite reducir la dimensionalidad de un conjunto de datos al transformar las variables originales correlacionadas

en un nuevo conjunto de variables no correlacionadas llamadas **componentes principales**. Estas componentes capturan la mayor parte de la variabilidad presente en los datos originales, facilitando el análisis y visualización. Para obtener estos componentes principales se estandarizaron los datos y se calculó la matriz de covarianza $S = VAR(X)$, la cual se utilizó para encontrar los eigenvalores λ_i y sus correspondientes eigenvectores δ_i que maximicen la expresión $\delta_i S \delta_i{}' = \lambda_i$. En otras palabras, la búsqueda del primer componente se reduce a encontrar el mayor eigenvalor y su correspondiente eigenvector, vector el cual contiene los pesos de cada variable en este nuevo componente y servirá para proyectar los datos originales. Así se realizó sucesivamente hasta encontrar la cantidad de componentes que acumulen al menos el 90% de la varianza.

Una vez reducida la dimensionalidad, se aplicó el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) sobre las componentes principales obtenidas. DBSCAN es un método de clustering basado en densidad que identifica agrupaciones de puntos densamente conectados y puede manejar eficientemente formas de clusters arbitrarias y presencia de ruido. De hecho, el motivador principal detrás de la elección de este algoritmo reside en este último aspecto, la facilidad que tiene para encontrar ruido se alinea con nuestro objetivo de identificar clientes "atípicos" que puedan presentar algún comportamiento sospechoso. Este algoritmo requiere dos parámetros esenciales: ε (radio máximo para considerar que dos puntos son vecinos) y $min\_samples$ (número mínimo de puntos necesarios para formar un clúster denso). Siguiendo la recomendación de Sander et al. (1998), se estableció $\text{min\_samples} = 2 * dim$, y para determinar el valor óptimo de ε se utilizó el enfoque de Rahmah y Sitanggang (2016), que consiste en calcular la distancia a los min_samples más cercanos de cada punto y posteriormente ordenar los resultados de menor a mayor para identificar el punto donde se encuentre la mayor curvatura.

Una vez determinados, el algoritmo *DBSCAN* clasifica los puntos de datos en tres categorías: puntos núcleo, que tienen al menos min_samples vecinos dentro del radio ε; puntos frontera, que están dentro del vecindario ε de un punto núcleo, pero no cumplen con min_samples; y puntos de ruido, que no están suficientemente cerca de otros puntos para pertenecer a un clúster. El algoritmo comienza seleccionando un punto no visitado y, si es un punto núcleo, inicia un nuevo clúster y expande iterativamente este clúster agregando todos los puntos densamente conectados (aquellos alcanzables desde puntos núcleo dentro de ε), hasta que no queden puntos adicionales para agregar; este proceso se repite hasta que todos los puntos hayan sido visitados y clasificados.

4. RESULTADOS Y DISCUSIÓN

4.1. Resultados obtenidos al aplicar el algoritmo

Aplicando el 90% de la varianza de los datos, se decidió seleccionar 15 componentes principales para realizar la clusterización. Al analizar los gráficos, observamos lo siguiente:

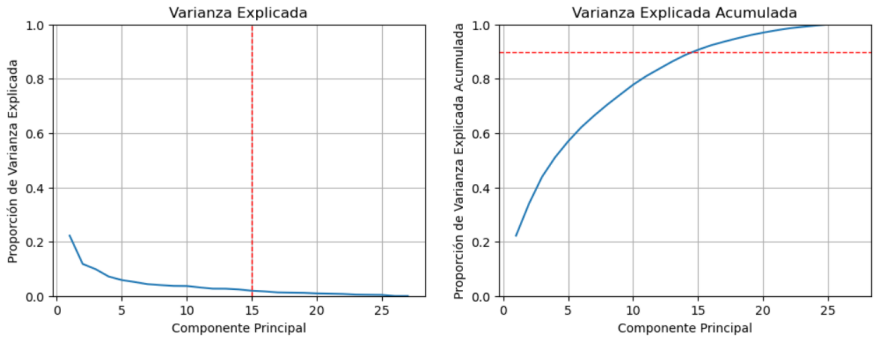

Gráfico 1. Varianza Explicada y acumulada por componentes principales

- En la gráfica de **varianza explicada** (izquierda), se puede apreciar que a partir del componente 15, el aporte de varianza se estabiliza, indicando que los primeros 15 componentes son suficientes para explicar una parte significativa de la variabilidad en los datos.
- En la gráfica de **varianza explicada acumulada** (derecha), el umbral del 90% se alcanza con los primeros 15 componentes, lo que justifica la decisión de usar este número de componentes para los posteriores análisis de clustering.

Al aplicar el algoritmo *DBSCAN*, utilizando los parámetros *eps* = 1.738 y *min_samples* = 30, se observó la segmentación de los datos en diferentes grupos, los cuales están representados por distintos colores en el gráfico. El valor de *eps* y el número mínimo de muestras fue ajustado para optimizar la detección de grupos de interés, logrando así una correcta clasificación de las observaciones. Sin embargo, algunos puntos fueron clasificados como ruido, lo cual es común en este tipo de algoritmos cuando se encuentran puntos que no encajan claramente en ningún clúster.

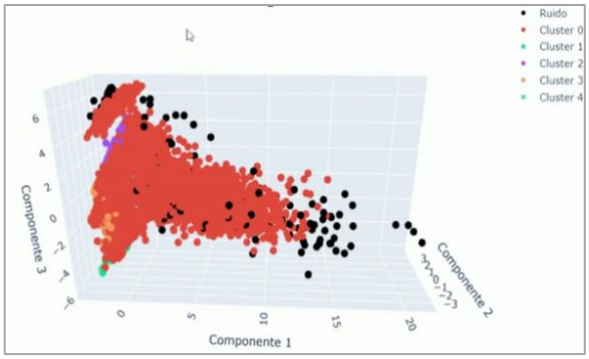

Gráfico 2. Clústeres generados

A continuación se describen los clústeres identificados:

- **Cluster 0:** Agrupa cuentas con saldos mayormente bajos, poca actividad y una gran variabilidad en los montos. Podría representar cuentas de clientes con poca interacción o balances pequeños.
- **Clusters 1,2,3,4**: agrupan cuentas con saldos bajos a moderados, menos transacciones y menor variabilidad, posiblemente representando cuentas menos activas o de clientes con fines específicos.
- **Outliers** son particularmente sospechosos y sus patrones de comportamiento generan indicios de LAFT, ya que muestra características típicas de cuentas que podrían estar involucradas en actividades ilícitas, como lavado de dinero, debido a sus altos saldos, grandes retiros y alta actividad transaccional.

4.2. Comparativo de los resultados con otros algoritmos

Como parte de la evaluación, se implementaron los algoritmos K-Means y Clustering Jerárquico. Sin embargo, estos no lograron identificar de manera óptima la cantidad adecuada de clústeres ni detectar el ruido (clientes con transacciones sospechosas). Esta comparativa puede consultarse en el código adjunto al presente informe, cuyos resultados se detallan en el anexo 2."

4.3. Posibles limitaciones

El algoritmo *DBSCAN (Density-Based Spatial Clustering of Applications with Noise)*, aunque es muy eficaz para detectar clústeres en datos de alta densidad, consideramos que presenta ciertas limitaciones, identificándose principalmente tres (3) en el contexto financiero:

- ✓ Sobre la densidad: *DBSCAN* asume que todos los clústeres tienen una densidad similar (definida por los parámetros eps y min_samples). Si los clústeres varían significativamente

en densidad, el algoritmo puede fragmentar algunos de ellos o clasificarlos como ruido, lo que afecta la detección de agrupamientos heterogéneos.

- ✓ Sobre el ruido: Aunque *DBSCAN* es eficiente en identificar el ruido, un conjunto de datos con una gran cantidad de puntos atípicos, como cuentas o transferencias irregulares, puede llevar a clasificar a demasiados clientes como ruido.
- ✓ Sobre el análisis no lineal: El algoritmo no considera patrones no lineales o secuenciales comunes en actividades como el lavado de dinero, donde transacciones aparentemente aisladas pueden tener un impacto relevante cuando se analizan en conjunto.

Las limitaciones en la variabilidad de densidades, la posible clasificación excesiva de ruido y la falta de análisis no lineal hace que sea necesario complementarlo con otras técnicas o ajustar sus parámetros para adaptarlo mejor a los patrones complejos y dinámicos característicos de las transacciones financieras.

4.4. Potenciales estudios por realizarse

Además del uso del algoritmo *DBSCAN*, es posible complementar el algoritmo con otros enfoques destinados a la detección y prevención de Lavado de Activos y Financiamiento del Terrorismo (LAFT) con los siguientes candidatos:

- ✓ Análisis de Redes Sociales (ARS): Permite estudiar las relaciones y conexiones entre entidades, como cuentas bancarias, tanto de personas como de empresas. ARS identifica cómo están conectadas las entidades participantes en transacciones financieras, revelando patrones y colusiones ocultas que no son detectables con métodos tradicionales.
- ✓ Análisis de series temporales con *LSTM (Long Short-Term Memory): LSTM* es útil para detectar cambios graduales o abruptos en las transacciones, identificando movimientos inusuales que los métodos no supervisados podrían pasar por alto. Un ejemplo es cuando un cliente, que suele realizar transacciones regulares, comienza a hacer grandes movimientos de dinero.
- ✓ Algoritmo de *Label Propagation*: Este algoritmo etiqueta nodos basándose en sus conexiones con otros nodos sospechosos, ayudando a identificar transacciones que inicialmente parecen normales pero podrían estar vinculadas a actividades ilícitas.

En conjunto, estas técnicas serían materia de estudio para mejorar la detección de actividades sospechosas y la prevención del LAFT y así poder capturar patrones ocultos, tanto en términos de relaciones entre entidades como en comportamientos temporales, fortaleciendo así los mecanismos de prevención y detección en el contexto financiero.

5. CONCLUSIONES Y RECOMENDACIONES

En este proyecto, se ha abordado el reto de clasificar clientes de una entidad financiera colombiana que ofrece billeteras digitales, con el objetivo de cumplir con las normativas de prevención de LAFT. Utilizando el algoritmo *DBSCAN*, se logró agrupar a los clientes de acuerdo con sus comportamientos transaccionales, identificando grupos homogéneos y detectando aquellos puntos considerados como ruido, que podrían indicar actividad sospechosa o anómala.

Las principales conclusiones derivadas de este análisis incluyen la capacidad de *DBSCAN* para identificar grupos relevantes y detectar comportamientos inusuales. Sin embargo, también se observaron limitaciones, especialmente en la variabilidad de densidades entre clústeres y la posibilidad de clasificar a demasiados clientes como ruido, lo que podría llevar a la omisión de importantes casos de estudio. A pesar de estas limitaciones, el algoritmo cumplió con su propósito principal de segmentación.

Entre las recomendaciones, se sugiere complementar el uso de *DBSCAN* con otras técnicas, como el Análisis de Redes Sociales (ARS), modelos LSTM y el algoritmo de *Label Propagation*, para mejorar la identificación de patrones ocultos, secuenciales y no lineales en las transacciones financieras.

6. BIBLIOGRAFIA Y REFERENCIAS

Superintendencia Financiera de Colombia. (2022). *Circular Básica Jurídica (C.E. 029/14), Parte I, Título IV, Capítulo IV (C.E. 011/22)*. Superintendencia Financiera de Colombia.

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. Recuperado el 1 de setiembre de 2024, desde: http://ref.scielo.org/v7rrcd

Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery, 2*(2), 169–194. https://doi.org/10.1023/A:1009745219419

Rahmah, N., & Sitanggang, I. S. (2016). Determination of optimal epsilon (Eps) value on DBSCAN algorithm to clustering data on peatland hotspots in Sumatra. *IOP Conference Series: Earth and Environmental Science, 31*, 012012. https://doi.org/10.1088/1755-1315/31/1/012012

ANEXOS

Anexo 1

Dataset Resultante

N	*VAR*	*count*	*unique*	*top*	*freq*	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
1	Negocio	27,558	170	WW005	26088	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Saldo total	27,558	NaN	NaN	NaN	161,048.44	591,745.10	0	0	92.00	3,377.25	10,156,684.88
3	Saldo disponible	27,558	NaN	NaN	NaN	53,199.30	282,802.46	0	0	0	500.00	10,156,684.88
4	Acumulado retiro	27,558	NaN	NaN	NaN	147,254.43	457,160.77	0	0	0	0	9,824,955.10
5	Ciudad	27,558	459	Bogota D.C	19413	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	Departamento	27,558	38	Bogota D.C	19439	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	antiguedad_cliente_dias	27,558	NaN	NaN	NaN	421.82	256.25	0	253.00	368.00	589.00	2,036.00
8	total_transacciones	27,558	NaN	NaN	NaN	20.17	34.42	1	4.00	6.00	20.00	477.00
9	corresponsales_diferentes	27,558	NaN	NaN	NaN	1.86	3.47	0	0	1.00	2.00	53.00
10	usuarios_diferentes	27,558	NaN	NaN	NaN	2.55	3.61	0	1.00	1.00	2.00	75.00
11	cantidad_consignacion	27,558	NaN	NaN	NaN	8.02	13.57	0	2.00	2.00	8.00	289.00
12	cantidad_retiro	27,558	NaN	NaN	NaN	12.15	22.66	0	2.00	4.00	13.00	328.00
13	prop_consignacion	27,558	NaN	NaN	NaN	0.43	0.17	0	0.33	0.44	0.50	1.00
14	prop_retiro	27,558	NaN	NaN	NaN	0.57	0.17	0	0.50	0.56	0.67	1.00
15	promedio_valor_consignacion	27,558	NaN	NaN	NaN	373,841.29	574,416.38	0	115,000.00	281,819.59	503,750.00	9,952,800.00
16	promedio_valor_retiro	27,558	NaN	NaN	NaN	273,664.36	494,856.95	0	92,349.22	180,000.00	330,000.00	9,906,488.00
17	AP_Consignación	27,558	NaN	NaN	NaN	15,805.42	71,689.68	0	0	0	0	2,400,000.00
18	AP_Retiro	27,558	NaN	NaN	NaN	80,534.57	247,188.90	0	0	0	60,000.00	9,000,000.00
19	BS_Consignación	27,558	NaN	NaN	NaN	9,962.92	223,080.33	0	0	0	0	9,902,000.00
20	BS_Retiro	27,558	NaN	NaN	NaN	33,999.90	420,868.51	0	0	0	0	9,906,488.00
21	DF_Retiro	27,558	NaN	NaN	NaN	12,199.54	182,946.15	0	0	0	0	10,052,189.00
22	EM_Consignación	27,558	NaN	NaN	NaN	112,891.35	255,598.16	0	0	0	0	4,961,534.00
23	PD_Consignación	27,558	NaN	NaN	NaN	36,357.38	147,576.62	0	0	0	0	7,000,000.00
24	PD_Retiro	27,558	NaN	NaN	NaN	135,301.58	208,348.50	0	0	0	222,181.22	3,000,000.00
25	PS_Consignación	27,558	NaN	NaN	NaN	164,342.89	560,038.42	0	0	0	166,628.57	9,952,800.00
26	PS_Retiro	27,558	NaN	NaN	NaN	65.90	7,265.65	0	0	0	0	1,046,000.00
27	RM_Consignación	27,558	NaN	NaN	NaN	1,715.53	51,804.61	0	0	0	0	4,298,888.00
28	TI_Consignación	27,558	NaN	NaN	NaN	99,301.00	204,015.86	0	0	0	140,000.00	8,836,358.69
29	TI_Retiro	27,558	NaN	NaN	NaN	115,675.38	285,247.27	0	0	0	113,390.59	9,216,135.41

Anexo 2
Comparativa con otros algoritmos