Análisis de anomalías en notas débito y crédito utilizando técnicas de aprendizaje automático no supervisado

Maria Camila Duarte Foronda
Cristian Joel Lozano Durán

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesora
María Bernarda Salazar Sánchez, Doctora (PhD) en Ingeniería Electrónica

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2024

Cita	(Duarte y Lozano, 2024)
Referencia **Estilo APA 7 (2020)**	Duarte Foronda, M. C., Lozano Durán, C. J. (2024). *Análisis de anomalías en notas débito y crédito utilizando técnicas de aprendizaje automático no supervisado*. [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte VII.

Grupo de Investigación Intelligent Information Systems Lab – In2Lab.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Dedicatoria

A la Universidad de Antioquia, nuestra casa, que en medio de la adversidad sigue siendo un faro de esperanza. Confiamos en que superará los retos presentes y continuará transformando vidas.

Agradecimientos

Agradecemos a la Universidad de Antioquia, nuestra alma máter, por ser el pilar de nuestra formación académica y profesional. Hoy, más que nunca, reconocemos su invaluable contribución a la educación pública y su papel en el desarrollo de nuestra sociedad. Estamos convencidos que, a pesar de los desafíos actuales, superará las dificultades y continuará iluminando el camino de muchos.

.

Tabla de contenido

Lista de Figuras

Lista de Tablas

Análisis de anomalías en notas débito y crédito utilizando técnicas de aprendizaje automático no supervisado

***Resumen*—** Este estudio desarrolla un modelo basado en aprendizaje no supervisado para detectar transacciones inusuales de ventas (Notas crédito y débito) de una empresa del sector servicios. Se utilizan métodos de reducción de dimensionalidad y UMAP para el análisis y visualización de patrones en los datos transformados. Para la identificación de anomalías, se implementaron técnicas de clustering, lo cual permite agrupar los datos y destacar comportamientos anómalos que requieren un control detallado por parte del equipo auditor de la compañía. Los resultados sustentan un modelo que, a partir de métricas como Cruces Cero y Cruces Neteo, permite identificar patrones de inestabilidad o comportamientos sospechosos; además, el análisis de estadísticas de transacciones, como la media, la desviación estándar y los valores extremos, facilita la detección de outliers, garantizando un monitoreo exhaustivo, continuo y de calidad que contribuye al cumplimiento de la normatividad financiera nacional e internacional, mejorando la gestión de riesgos de la compañía y fortaleciendo la confianza en las muestras revisadas por la auditoría.

Palabras claves — PCA, UMAP, Detección de anomalías, DBSCAN, Auditoria.

I. INTRODUCCIÓN

El control y auditoría de transacciones financieras son elementos fundamentales para la transparencia y cumplimiento normativo en cualquier compañía. Esto es especialmente crítico en empresas que cotizan en mercados financieros, puesto que de ello depende la confianza del inversionista. Como lo menciona Feldsztajn et al. (Feldsztajn, 2024): “Los estados financieros son el lenguaje que revela la salud económica y financiera de una entidad. Y es que estos informes esenciales no son simples registros contables; son la ventana a la posición financiera, rendimiento y flujo de efectivo que define la trayectoria de una organización”. Por ello, garantizar que la información que revelan en sus estados financieros es confiable y transparente, la cual corresponde a operaciones validas registradas en los ERP (Del inglés, Enterprise Resource Planning) de la empresa como notas débito o crédito.
En este sentido una anomalía, es un comportamiento poco común en los datos, en cuanto a la relación de neteo entre los ingresos y las devoluciones. Esto podría indicar errores asociados a los vendedores o un evento que debidamente justificado no representar la materialización de los riesgos expuestos. Considerar la homogeneidad de los datos y el tipo de clientes, puede con llevar a que una anomalía aplique solo a un grupo, por lo tanto, no a todos.

Por ello, para que la compañía pueda brindar confianza sobre la información que revela, crea un sistema de control interno al interior de la empresa que le permite velar por los atributos de la información financiera que requiere asegurar, como: oportunidad, veracidad, confiabilidad y completitud. A nivel estructural las compañías se organizan en tres líneas de actuación; la primera línea corresponde a los lideres de operación, quienes conocen sus procesos y pueden identificar fácilmente riesgos y controles para mitigarlos. Como segunda línea, se encuentran las áreas transversales que brindan a poyo a los líderes en la identificación y gestión de riesgos. Finalmente, está la auditoría interna, quienes aseguran que las líneas anteriores realicen la identificación confiable de riesgos, además de brindar opinión y calificar el grado de madurez de gestión de la compañía.

Así mismo, con el objetivo de brindar un aseguramiento transversal, la auditoría interna tiene grandes retos en términos de confianza y alcance del aseguramiento que

puede brindar al interior de las compañías. Uno de los mayores retos en este proceso de gestión, es la cantidad de operaciones que se realizan día a día, y las cuales deben ser auditadas para garantizar el cumplimiento de políticas y estándares de calidad, tal que se descarte la materialización de algún riesgo de fraude, error operativo y financiero. Esta supervisión se realiza de forma aleatoria a través de tablas militares para poblaciones finitas. Sin embargo, existe la duda de si estas tablas son suficientes para detectar la tendencia en la materialización de errores en las operaciones, priorizando la auditoría den un conjunto de transacciones que se pueden considerar anómalas o inusuales.

En este contexto, los avances de las técnicas de inteligencia artificial en los últimos años han representado una gran oportunidad para la gestión de riesgos y la auditoria en las compañías. Como lo menciona Carlos Roger socio español de EY (Roger, 2023): *"La gran cantidad de datos generados por las compañías en la era digital presenta tantos desafíos como oportunidades para la función de auditoría interna. Gracias al análisis de datos avanzado, los auditores pueden examinar conjuntos masivos de información de manera rápida y efectiva. La identificación de patrones, tendencias y anomalías se ha vuelto más accesible, lo que fortalece al igual que indicábamos previamente la capacidad de los auditores internos de detectar posibles riesgos y fraudes."* Por lo anterior, en este trabajo se busca analizar como el aprendizaje automático no supervisado puede apoyar el análisis y detección de grupos de transacciones con un comportamiento diferente a la mayoría. Lo anterior, puede ser un nuevo método de uso de la auditoría interna, para determinar una muestra significativa y realizar la evaluación de riesgos. Frente a las transacciones que se evalúan en el contexto de este trabajo, se abordan operaciones de débito y crédito que realizan los vendedores de una empresa de servicios, incluyendo la identificación de riesgos por los cuales se genera la necesidad de auditoría. El modelo propuesto en este trabajo se centra en la aplicación de clustering para capturar patrones en las transacciones a lo largo del último año, buscando detectar anomalías de forma rápida y eficiente. Además, se implementa el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos y detectar características relevantes, obteniendo mayor eficiencia en el modelo y mejor comprensión del problema. Para una visualización efectiva y exploración de los datos considerando la alta dimensionalidad, se emplea t-SNE, lo cual facilita la comprensión de la estructura subyacente de los datos. Finalmente, se utiliza un método de clasificación no supervisado con el algoritmo DBSCAN para identificar y agrupar patrones de comportamiento anómalos etiquetando transacciones sospechosas de automatizada.

La implementación de este enfoque integral proporciona una herramienta para el monitoreo proactivo de este proceso financiero, optimizando recursos destinados a auditorías y disminuyendo el tiempo empleado en la gestión/detección del riesgo. Esta estrategia impacta en la eficiencia en las tomas de muestreo para auditoría en grandes volúmenes de datos, proporcionando mayor precisión y centrándose en las anomalías susceptibles de verificación.

II. Materiales y Métodos

A) Descripción de la base de datos

El conjunto de datos contiene 350,957 registros que proviene de una empresa de servicios y está compuesto por transacciones financieras realizadas por los vendedores durante un periodo anual (octubre de 2023 a 2024), el 50% de los datos se distribuye entre diciembre de 2023 y julio de 2024. Estas transacciones incluyen operaciones de ingreso (crédito) y devolución (débito), registradas en los sistemas de la empresa (ver Tabla 1). En promedio el costo de las operaciones está ± 22.14 mil millones, con una media de 1,098,768 COP.

Tabla 1. Descripción *de las variables disponibles en la base de datos*

Variable/Tipo	*Descripción*	*Rango*
Mes / Texto	Indica el mes en que ocurrió la transacción	13 valores, Desde octubre 2023 hasta octubre 2024
cod_sucursal / Categórica	Código numérico que identifica la sucursal donde se realizó la transacción	115 sucursales diferentes
Fecha_Emision / Fecha - Datetime	Fecha exacta de emisión de la operación.	Desde 2023-10-01 al 2024-10-25
tipo_op / Categórica	Tipo de operación realizada, que incluye categorías como: • Anulación de disminución: Operaciones que revierten ajustes previos.	En total son 8 tipos de operación diferentes

	• Cancelación: Transacciones canceladas por diversas razones. • Cobro o aumento: Operaciones de aumento en ingresos.	
Vendedor / Categórica	Identificador numérico único del vendedor asociado con la transacción.	Se cuentan con registros de 3071 vendedores distintos
cod_unidad / Categórica	Código que identifica la unidad organizacional donde se registró la operación.	370 códigos de unidad.
cod_venta / Categórica	Identificador único de la venta asociada con la transacción.	Son 40490 códigos de venta.
cod_producto / Categórica	Código del producto relacionado con la operación.	Para este caso se cuenta con un único código de producto
valor_operación / Numérica	Monto de la transacción	El rango va de (-22,140,160,000) a (+ 22,140,160,000)

B) Preprocesamiento de datos

En función de disminuir los requerimientos de cómputo para el procesamiento del modelo, se realizó Análisis de Componentes Principales (PCA) que permite reducir la dimensionalidad del conjunto de datos, se seleccionan 10 componentes, las cuales del representan el 90% de la varianza de los datos originales (ver Figura 1).

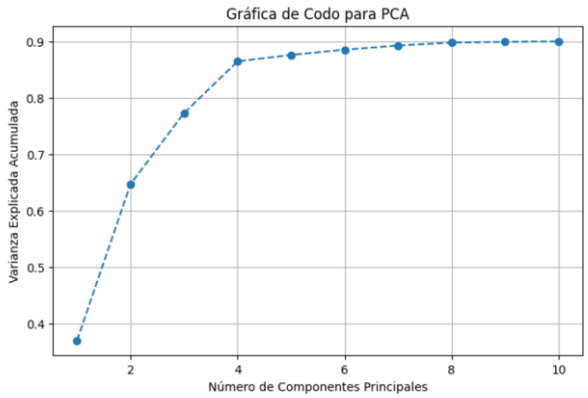

Figura 1. Gráfica de codo para PCA

Con el objetivo de garantizar un análisis integral de los datos y fundamentados en el propósito del estudio que busca identificar transacciones anómalas, se decidió conservar los valores atípicos presentes en el conjunto de datos en lugar de eliminarlos debido a que estos pueden reflejar información valiosa, como posibles errores operativos, casos de fraude o comportamientos excepcionales que podrían requerir acciones específicas. Además, su exclusión habría alterado la estructura estadística de los datos, lo que podría sesgar los resultados y comprometer la generación de insights relevantes para el análisis. Dado que los datos analizados presentan una distribución compleja, los valores extremos no necesariamente representan errores, sino que podrían ser reflejo de comportamientos extremos pero válidos

III. MÉTRICAS DE EVALUACIÓN

Las métricas relacionadas en la Tabla 2, son fundamentales para identificar patrones de comportamiento anómalos y priorizar esfuerzos de auditoría. La métrica de *Cruces_Cero* y *Cruces_Neteo* detectan patrones específicos en las transacciones, las estadísticas resumidas ofrecen un contexto cuantitativo que permite evaluar si un vendedor o producto está operando dentro de los rangos esperados (del conjunto de datos). Este enfoque integral fortalece la capacidad de las empresas para gestionar riesgos y garantizar la integridad de sus procesos financieros.

Tabla 2. Descripción de las métricas *de evaluación.*

Métrica	**Importancia**	**Objetivo**
Cruces_Cero: Cantidad de veces que las transacciones cambian de signo en un período dado por vendedor.	es un indicador de inestabilidad que sugiere problemas en el control de calidad, errores operativos o intentos deliberados de manipulación.	Detectar patrones inestables en vendedores o sucursales y detectar posibilidades de mejora en procesos operativos.
Cruces_Neteo: Identifica el número de veces que las transacciones se compensan entre sí, es decir, cuando la suma acumulada de las transacciones regresa a cero debido a ingresos y devoluciones que se cancelan mutuamente	puede ser un indicador de fraudes, como devoluciones ficticias para ocultar ingresos reales o inflaciones de ventas seguidas por devoluciones. En casos normales, no es habitual que las transacciones se compensen completamente.	Identificar vendedores, productos o sucursales con un comportamiento inusual. Ayudar a priorizar auditorías en las áreas con más riesgos potenciales. Mientras Identificar vendedores, productos o sucursales con un comportamiento inusual. Ayudar a priorizar auditorías en las áreas con más riesgos potenciales.
Estadísticas Resumidas – proporcionan un contexto cuantitativo del comportamiento de las transacciones,		
Media (Media_Valor): Representa el promedio del valor de las transacciones para un vendedor	Ayuda a identificar vendedores con patrones anómalos, como promedios demasiado altos o bajos en comparación con la norma	Establecer un comportamiento base para cada vendedor y detectar outliers en vendedores que realizan transacciones significativamente diferentes.
Desviación Estándar (Std_Valor): Mide la variabilidad en los valores de las transacciones	Una desviación estándar alta indica que las transacciones tienen valores muy diferentes entre sí. Un patrón de variabilidad excesiva puede	Identificar vendedores con comportamientos irregulares y Analizar si la variabilidad se relaciona con tipos específicos de

	sugerir inconsistencias operativas o errores en el registro de transacciones	operaciones o productos.
Valores Mínimos y Máximos (Min_Valor, Max_Valor): Valores mínimo y máximo de las transacciones.	Los valores extremos son clave para detectar transacciones anómalas, como ingresos inflados o devoluciones masivas. Un valor mínimo extremadamente bajo puede reflejar un problema operativo o un error en el registro.	Priorizar transacciones inusuales para auditoría y Establecer rangos normales de operación para cada vendedor.

IV. DESARROLLO DE MODELOS DE CLUSTERING

El desarrollo del modelo aquí planteado y cuya metodología se encuentra resumida en la Figura 2. Busca encontrar grupos con comportamientos similares, asegurándose que todos los puntos de datos del segmento sean del mismo tipo. Es decir, con comportamientos similares. Además, identificar aquellos patrones que hacen que otros datos se ubiquen en grupos diferentes. Para esto, se implementa el uso de métodos de clustering cuyo objetivo de cualquiera de sus modelos es detectar patrones en los datos. Específicamente, para agrupar los datos en distintos grupos, que se componen de puntos de datos, que son muy similares entre sí, pero distintos de los puntos de datos de otros grupos (Google, 2024). Podemos usar esto para la detección de Anomalías y determinar qué datos se parecen mucho a los datos en el clustering y cuales resultan difíciles de asignar a cualquier cluster. Se podrían marcar dichos datos como extraños o sospechosos. Sin embargo, es importante reiterar que no necesariamente significa que sea una transacción fraudulenta o que se haya materializado el riesgo, lo que significa, es que son en comparación con la mayoría de los comportamientos normales, simplemente extraños. En este caso, una vez detectados estos datos sospechosos, los auditores de la compañía deben investigar más a fondo. Permitiendo a la auditoria contar con muestras que contienen los datos más anormales; esto podría brindar mayor confiabilidad en la muestra tomada para las pruebas.

Una vez realizada la exploración inicial de datos y luego de reducir la dimensionalidad de estos por medio de PCA, se aplicó el método UMAP (Del inglés, Uniform Manifold Approximation and Projection) para reducir los datos a 2 dimensiones, esto facilita la visualización de las estructuras subyacentes en el conjunto de datos. Este enfoque se eligió para preservar la estructura local de los datos, asegurando que las relaciones entre puntos cercanos se mantengan, lo anterior, es un aspecto crucial en el análisis donde la proximidad y los clústeres tienen un significado relevante.

Figura 2 . Metodología de desarrollo

Además, UMAP conserva mejor las relaciones globales en comparación con técnicas como t-SNE, permitiendo una interpretación más coherente de la posición relativa entre los clústeres identificados. A esto se suma que su implementación es significativamente más rápida y escalable, lo que resulta especialmente útil en el procesamiento de grandes volúmenes de datos. Finalmente, UMAP ofrece flexibilidad mediante la

personalización de parámetros, permitiendo adaptaciones específicas según las necesidades del análisis.

Se inicia el desarrollo con la implementación de HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), un método diseñado para identificar clústeres basados en la densidad de los datos. Este algoritmo resulta particularmente útil para detectar estructuras en conjuntos complejos, ya que clasifica automáticamente como ruido (-1) aquellos puntos que no pertenecen a ningún clúster (ver Figura 5), eliminando la necesidad de fijar manualmente el número de agrupaciones. Su enfoque jerárquico permitió ajustar los resultados según los niveles de densidad presentes en los datos, lo que facilitó la identificación de clústeres de diversas formas y tamaños. HDBSCAN es una extensión del algoritmo DBSCAN, con la diferencia de que ajusta automáticamente las regiones densas utilizando un modelo jerárquico, lo que lo hace una técnica más robusta frente a datasets con densidades variables. Además, HDBSCAN puede identificar clusters de diferentes tamaños y formas, eliminando la necesidad de decidir un número de clusters previamente, lo que lo hace especialmente útil para análisis exploratorios.

Posteriormente, se aplicó DBSCAN (Density-Based Spatial Clustering of Applications with Noise), un método similar que también se basa en la densidad, pero que requiere la definición manual de los parámetros clave: radio de vecindad (eps) y el número mínimo de puntos requeridos para formar un clúster (min_samples). La elección de estos parámetros fue realizada cuidadosamente, considerando las características específicas del conjunto de datos para asegurar una segmentación precisa. Aunque este método no cuenta con la flexibilidad jerárquica de HDBSCAN, resultó útil como una herramienta complementaria, proporcionando una perspectiva adicional sobre la estructura de los datos y permitiendo validar los resultados obtenidos con el primer enfoque.

Otra forma de tratar los datos anómalos no es solo en función de los valores atípicos de la agrupación. En lugar de tratar las anomalías como datos atípicos, también, se puede usar los
clústeres más pequeños como una indicación de sospechosos.

Para realizar los agrupamientos, existen muchos métodos que se pueden utilizar para la detección de las anomalías. Cada método de agrupación tiene sus pros y contras. DBSCAN es un método de agrupación que funciona bastante bien para datos que no se agrupan en formas redondas normales, significa que genera un agrupamiento espacial de aplicaciones con ruido basada en densidad, por lo cual no se necesita predefinir la cantidad de clústeres. El algoritmo encuentra muestras centrales de alta densidad y expande grupos a partir de ellas (Datacamp, 2024). Este es un tipo de algoritmo que se puede usar para identificar anomalías como grupos pequeños. Dentro de los parámetros de este método se encuentra: la máxima distancia entre los datos dentro del cluster y el número mínimo de puntos de datos en los clústeres.

Por último y para complementar la detección de anomalías, se empleó Isolation Forest, un método específicamente diseñado para identificar puntos atípicos en espacios multidimensionales. Este modelo funciona al aislar iterativamente cada observación del conjunto de datos, aprovechando el hecho de que las anomalías suelen requerir menos divisiones para quedar completamente aisladas en comparación con los puntos normales. A partir de este proceso, se generaron puntuaciones de anomalía que cuantifican el grado de sospecha de cada punto.

V. Resultados y Discusión

A) Reducción de dimensionalidad con UMAP

El método UMAP permitió proyectar los datos originales a un espacio de menor dimensionalidad, preservando tanto la estructura local como global. Esta técnica destacó por su capacidad de capturar relaciones no lineales en los datos, lo que resultó en la identificación de clusters densos y una distribución clara de puntos dispersos, posiblemente asociados con outliers.

En la proyección generada, se observaron zonas densas que sugieren la presencia de clusters bien definidos (ver Figura 3), así como áreas más dispersas que podrían representar outliers o puntos con características atípicas. Esto refuerza la utilidad de UMAP para analizar datasets con estructuras complejas y densidades variables.

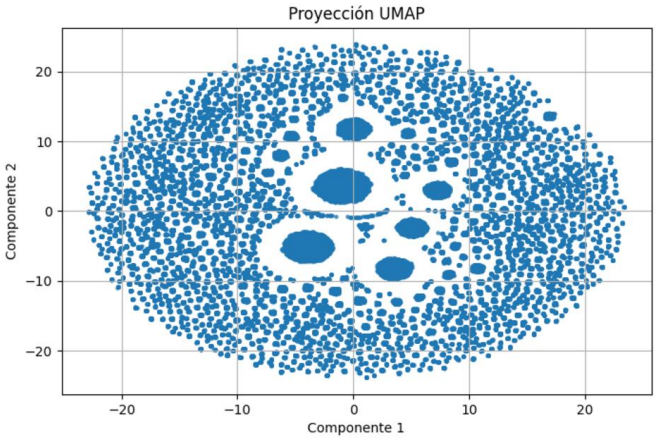

Figura 3. Proyección generada con UMAP

B) Análisis de clustering con HDBSCAN y DBSCAN

La identificación de puntos etiquetados como −1-1−1 (ruido o outliers) fue posible gracias a HDBSCAN, que aprovecha la proyección UMAP para clasificar de manera efectiva tanto clusters como puntos atípicos (ver Figura 4 y Figura 5), que pueden estar asociados a patrones anómalos los cuales son analizados en detalle en la base original.

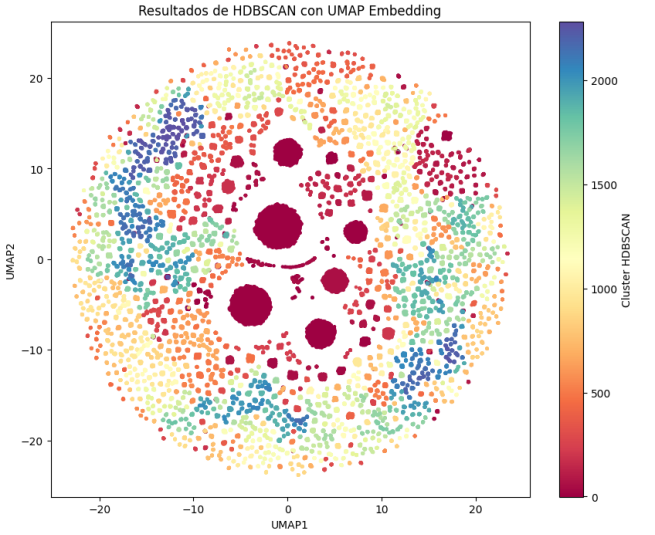

Figura 4. Identificación de atípicos o ruido con el método de HDBSCAN

En este contexto se obtienen dos escenarios de análisis, con altos valores de cruces cero, pero diferenciados en los cruces por neteo (ver Tabla 3). El primer caso, el **Cluster 0,** tiene una gran cantidad de cruces de valor cero (16,831) y un número considerable de cruces por neteo (3,126), lo cual sugiere que el comportamiento de este grupo está más relacionado con operaciones cercanas a un balance cero según se puede observar en la media y desviación presentados en la Tabla 3, es decir operaciones con menor dispersión.

El segundo caso, el **Cluster 3**, presenta muchos menos cruces por neteo (55) en comparación con cruces de valor cero (13,991), lo que sugiere que sus operaciones están altamente polarizadas, cuya desviación estándar es significativamente más alta con respecto a la media (ver Tabla 3), reflejando una gran variabilidad en los valores de las operaciones.

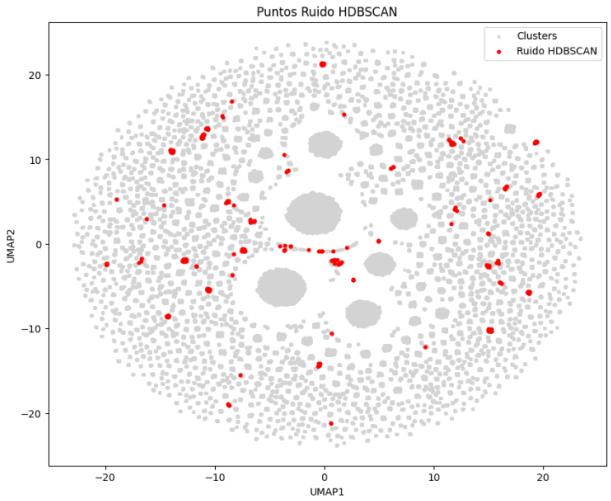

Figura 5.Puntos ruido HDBSCAN

El **Cluster 0** tiene un 24% menos de muestras en comparación con el **Cluster 3**, lo que podría indicar que último agrupa operaciones de alto riesgo o muestras caracterizadas por comportamientos anómalos, relacionados a la alta dispersión y montos extremos.Estas diferencias marcadas entre los clusters, implica que el **Cluster 0** agrupa operaciones más uniformes y cercanas a valores balanceados, sugiriendo que finalmente los datos asignados al **Cluster 3** podrían estar asociados con comportamientos atípicos o riesgosos, lo que lo convierte en un candidato principal para un análisis más detallado.

Tabla 3. Métricas de evaluación de la consistencia de los cluster encontrados con las técnicas DBSCAN y HDBSCAN

Métrica	*HDBSCAN Cluster 0*	*DBSCAN Cluster 1*	*HDBSCAN Cluster 3*	*DBSCAN Cluster 21*
Cruces Cero	16,831.0	16,831.0	13,991.0	13,991.0
Cruces Neteo	3,126.0	3,126.0	55.0	55.0
Media Valor	0.05566	0.05566	513.467	513.467
Desv. Est. Valor	7.229.405	7.229.405	302.473.573	302.473.573
Valor Mínimo	-888,734.875	-888,734.875	-18,625.519	-18,625.519
Valor Máximo	888.596.507	888.596.507	35,346.03	35,346.03
Tamaño Cluster	19,347	19,347	25,469	25,469

Tabla 4. Estadísticas de ruido HDBSCAN

Métrica	*Cruces Cero*	*Cruces Neteo*	*Media Valor*	*Desv. Est. Valor*	*Valor Mínimo*	*Valor Máximo*
Cantidad	172	172	172	172	172	172
Media	25.42	0	14.925	33.581951	-46.590602	186.17102
Deviación estándar	40.619	0	18.264739	32.239505	76.441358	250.39791
Mínimo	0.000000	0	-11.028518	302.473.573	-417.063622	-11.02851
25%	1	0	5.921229	0.000000	-113.243340	30.227502
50%	3	0	9.923213	26.483738	-10.163049	57.312943
75%	23.250000	0	15.591304	60.770284	-2.341800	194.59114
Máximo	131	0	94.731960	209.151114	70.380922	896.89976

Tabla 5. Estadísticas de ruido DBSCAN

Métrica	*Cruces Cero*	*Cruces Neteo*	*Media Valor*	*Desv. Est. Valor*	*Valor Mínimo*	*Valor Máximo*
Cantidad	118	118	118	118	118	118
Media	3.491525	0	7.806585	21.721811	-12.1679	52.160194
Deviación estándar	2.617318	0	7.390121	14.781745	17.5945	47.983081
Mínimo	0.000000	0	-3.889179	1.450536	-78.8292	-0.972655
25%	1	0	3.377431	9.956117	-15.9354	19.640468
50%	3	0	5.66267	18.433323	-4.2244	42.834991
75%	5	0	12.189619	32.624555	-1.8412	65.019269
Máximo	9	0	25.233899	51.937155	2.82487	205.74103

Al aplicar el método de DBSCAN resaltan dos cluster, el **Cluster 1**, muestra características similares en la cantidad de cruces de valor cero y cruces por neteo al **Cluster 1** identificado con el método HDBSCAN (ver Figura 6).

La mayoría de las operaciones de este grupo de muestras están centradas en valores cercanos a cero, pero con una desviación estándar moderada. Y un segundo cluster, el **Cluster 21**, el cual tiene muchas menos cruces por neteo en comparación con cruces de valor cero, lo que lo posiciona como un grupo con operaciones muy particulares. Su media es mayor que la del Cluster 1, con una desviación estándar mucho más alta, indicando que este cluster contiene operaciones de montos considerablemente altos, que podrían ser indicativos de riesgos o anomalías. Este cluster agrupa un mayor número de operaciones en comparación con Cluster 1, probablemente influenciadas por transacciones con valores extremos.

Los resultados entre HDBSCAN y DBSCAN son similares en cuanto al número de cruces (cero y neteo), la media de los valores, la desviación estándar y los valores extremos (mínimos y máximos). Ambos algoritmos detectaron los mismos tamaños de clusters y patrones generales de comportamiento dentro de los datos.

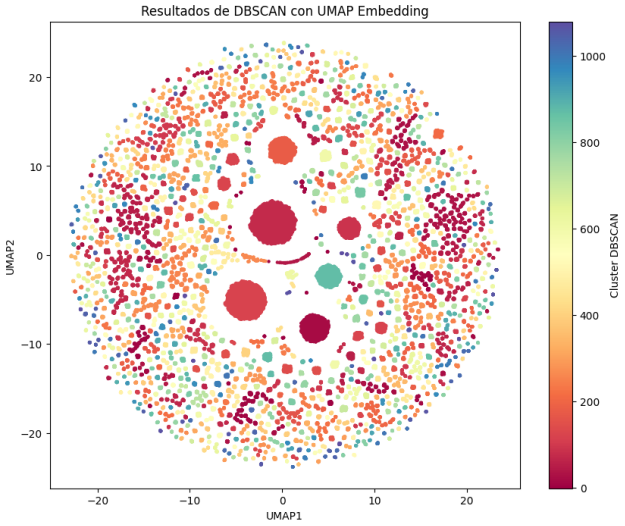

Figura 6. Resultados DBSCAN

HDBSCAN identificó **172 puntos como ruido**, frente a los **118 detectados por DBSCAN**. Esto sugiere que HDBSCAN es más sensible para identificar anomalías en el dataset, incluyendo puntos atípicos con comportamientos más diversos y extremos. Este comportamiento lo hace más adecuado para escenarios donde las anomalías tienen características complejas o donde se espera encontrar variaciones significativas entre los datos.

Los puntos clasificados como ruido por HDBSCAN tienen una media de **25.42 cruces cero**, significativamente mayor que los **3.49 de DBSCAN**. Esto indica que HDBSCAN clasifica como anomalías no solo puntos extremos, sino también aquellos con comportamientos más frecuentes (muchos cruces cero), pero con características adicionales que los hacen distintos del resto. DBSCAN, al ser más conservador, puede no capturar estas anomalías menos evidentes, lo que podría ser una limitación para detectar patrones anómalos complejos.

Los valores de las anomalías detectadas por HDBSCAN tienen una mayor dispersión (StdValor=33) y valores extremos significativamente más amplios (Min=−417.06, Max=896.89) (ver Tabla 4), en comparación con DBSCAN (StdValor=21.72, Min=−78.83, Max=205.74) (ver Tabla 5). Esto podría indicar que HDBSCAN es más efectivo para identificar anomalías con características extremas o distribuciones altamente variables, que suelen ser clave en la detección de fraudes o comportamientos atípicos.

Los resultados muestran que HDBSCAN tiene un mejor desempeño en identificar puntos fuera de los límites normales del comportamiento esperado, capturando anomalías tanto extremas como más sutiles. Por otro lado, DBSCAN podría ser útil si el objetivo es identificar solo las anomalías más obvias, pero tiene una limitación en cuanto a la capacidad de detectar variaciones complejas en los datos.

Se valido el resultado de una de las muestras, para el vendedor identificado con el código 11071, se detectaron inconsistencias en la gestión de ventas que han despertado alertas internas. Estas irregularidades, sumadas a la anomalía detectada en los datos operativos, han llevado a que el caso se encuentre actualmente bajo supervisión. Se está revisando en detalle la información asociada con este vendedor para determinar el origen de las discrepancias y evaluar posibles impactos en los procesos relacionados.

C) Análisis de clustering con Isolation Forest

Al aislar iterativamente cada observación del conjunto de datos, se obtuvo que, aunque se detectaron anomalías, la amplitud de los valores extremos en métricas (ver Tabla 6) como Min_Valor y Max_Valor refleja una alta dispersión, lo que podría indicar que Isolation Forest está capturando valores atípicos amplificados por su enfoque basado en la división iterativa del espacio (ver Figura 7). Además, la baja media de Cruces_Neteo (0.009) y su máximo limitado a 111 sugiere que Isolation Forest podría estar ignorando interacciones complejas en los datos. Estos resultados, aunque útiles para escenarios simples, no parecen ajustarse de manera óptima a las características de este dataset, ya que las anomalías detectadas incluyen muchos valores extremos que no siempre son representativos de patrones reales.

Tabla 6. Resultados Isolation Forest

	Media Valor	*Cruces Cero*	*Cruces Neteo*	*Min Valor*	*Max Valor*
count	2807	2807	2807	2807	2807
mean	15,280	34,471	0,0092	-208,836	515,306
std	33,175	19,089	0,095	680,979	1155,39
min	-6,697	1	0	-5608,66	3,5415
25%	5,071	20	0	-130,293	100,944
50%	7,701	32	0	-58,1923	188,725
75%	12,577	44	0	-30,072	364,904
max	216,151	81	1	-1,73872	7776,46

Lo obtenido con Isolation Forest, aunque útil en algunos contextos, parece sobreestimar ciertas anomalías al enfocarse en divisiones arbitrarias del espacio de características, lo que lleva a resultados menos precisos en este caso particular.

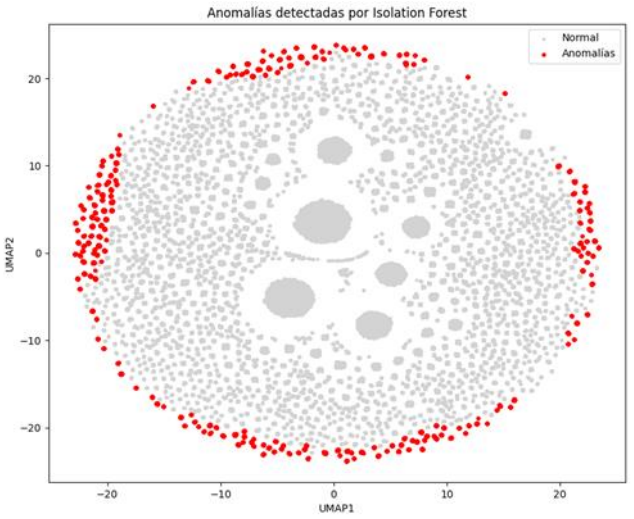

Figura 7.Resultados Isolation Forest

En contraste, HDBSCAN detectó anomalías de manera más consistente, enfocándose en puntos con densidades bajas que realmente se separan de los clusters principales. Esto respalda su capacidad para identificar outliers más relevantes en escenarios con relaciones complejas y distribuciones densas.

HDBSCAN demuestra ser más robusto al capturar anomalías relacionadas con la distribución real de los datos, priorizando la densidad y su separación respecto al resto del dataset.

El detalle de la fase de análisis y desarrollo de modelos fue implementado en lenguaje Python y se encuentra disponible en: https://github.com/maridufo/trabajo_grado_2024

VI. CONCLUSIÓN

Los resultados obtenidos con HDBSCAN son más confiables y representativos del comportamiento del dataset analizado. Esto respalda su uso como método principal para identificar anomalías relevantes en este trabajo. Isolation Forest puede considerarse una herramienta complementaria, pero no ofrece el mismo nivel de precisión en este contexto.

VII. REFERENCIAS

Datacamp. (09 de 29 de 2024). *Datacamp*. Obtenido de Datacamp: https://www.datacamp.com/tutorial/dbscan-clustering-algorithm

Feldsztajn, J. F. (19 de 07 de 2024). *www.esade.edu*. Obtenido de www.esade.edu: https://www.esade.edu/beyond/es/estados-financieros-empresa/

Google. (22 de 07 de 2024). *Developers*. Obtenido de https://developers.google.com/machine-learning/clustering/overview?hl=es-419

Roger, C. (27 de 11 de 2023). *www.ey.com*. Obtenido de www.ey.com: https://www.ey.com/es_es/the-cfo-agenda/transformacion-digital-auditoria-interna-navegando-hacia-eficiencia-precision-era-digital