Universidad Icesi Cali
Facultad de Ingeniería, Diseño y Ciencias Aplicadas.
Maestría Ciencia de Datos
Proyecto de Grado.

Aplicación de Inteligencia Artificial y Machine Learning para la segmentación en GIRO buscando identificar el riesgo de LA/FT según Circular Básica Jurídica de la Superintendencia Financiera de Colombia, numeral 4.1.1.1, Título IV, Parte I

Laura Daniela Espinosa
Carlos Enrique Jaramillo
Andrea Estefania Timaran

10 Diciembre 2024

Resumen

Este proyecto tiene como objetivo la implementación de un módulo de segmentación dentro del aplicativo GIRO, diseñado para mejorar la identificación y gestión de riesgos asociados con el Lavado de Activos y la Financiación del Terrorismo (LA/FT), en conformidad con las normativas colombianas establecidas por la Superintendencia Financiera. Utilizando técnicas de Inteligencia Artificial (IA) y Machine Learning (ML), se busca clasificar factores de riesgo, como clientes, productos, canales de distribución y jurisdicciones, con el fin de optimizar la eficiencia operativa y reducir los riesgos financieros y reputacionales.

La metodología aplicada incluye un análisis exploratorio de datos, seguido de la implementación de modelos de aprendizaje no supervisado: K-means y Clustering Jerárquico, tanto con reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA) como sin esta técnica. La evaluación del desempeño de los modelos se realiza a través de métricas robustas, lo que permite generar segmentaciones precisas que mejoren la toma de decisiones y fortalezcan el cumplimiento de las normativas regulatorias en el sector financiero colombiano.

Palabras clave: Segmentación, Lavado de Activos, Financiación del Terrorismo, Machine Learning, K-means, Hierarchical Clustering, PCA, Principal Components Analysis, Cumplimiento Normativo.

Abstract

This project focuses on the implementation of a segmentation module for the GIRO application, aimed at optimizing the identification and management of risks related to Money Laundering and Terrorism Financing (ML/TF), in compliance with Colombian regulations established by the Financial Superintendence. Using Artificial Intelligence (AI) and Machine Learning (ML) techniques, the project categorizes risk factors such as clients, products, distribution channels, and jurisdictions to enhance operational efficiency and minimize financial and reputational risks. The methodology involves data exploration, the implementation of unsupervised learning models (K-means and Hierarchical Clustering), and the evaluation of model performance using robust metrics. This approach enables precise segmentations, improves decision-making, and strengthens regulatory compliance in the Colombian financial sector.

Keywords: Segmentation, Money Laundering, Terrorism Financing, Machine Learning, Regulatory Compliance.

Índice general

Índice de figuras

Índice de tablas

Introducción

Dado el comportamiento de las cooperativas en Colombia en la década de 1990, las cuales comenzaron a recibir ingresos basado en transacciones sospechosas, lo que culminó en la crisis de las cooperativas; Se creó la superintendencia de economía solidaria, basado en esto, fue expedida la Ley 454 de 1998, que buscaba regular la economía solidaria y se dictaron normas sobre la actividad financiera de naturaleza cooperativa, entre otras disposiciones; De igual manera, con el estatuto orgánico del estatuto financiero, se creó el SIPLA (por sus siglas, sistema integral de prevención de lavado de activos), aplicable para el sector solidario inicialmente aplicable para cooperativas con actividad financiera.

Como consecuencia del atentado a las torres gemelas en estados unidos y la solicitud de Colombia el ingreso a la OCDE (por sus siglas de Organización para la Cooperación y el Desarrollo Económico) desde 2014; Se agregó la validación de financiación al terrorismo (FT), con las condiciones y normas regulatorias que se extienden a otros sectores económicos y la generalidad del sector solidario.

El Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT), debe ser implementado por las entidades financieras y otros sectores en Colombia para prevenir, detectar, y gestionar los riesgos asociados con el lavado de activos y la financiación del terrorismo.

Con el fin de apoyar a las diferentes entidades en la detección, identificación y gestión de los riesgos asociados con el lavado de activos y la financiación del terrorismo, Óptima Corporation desarrolló la suite de herramientas de software para la Gestión Integral de Riesgo Organizacional (GIRO), el cual es un conjunto de aplicaciones tecnológicas que soportan la adecuada gestión de los diferentes sistemas de administración de riesgos exigidos por la ley.

Esta gestión se realiza a través de diferentes módulos: Administración de riesgo, Monitoreo de Transacciones y Gestión Integral de Listas de Riesgos; es así como GIRO permite realizar una identificación oportuna de actividades sospechosas en organizaciones de manera oportuna y preventiva, dada la última actualización [1] de la Circular Básica Jurídica, con el fin de permitir una optimización en el proceso de Identificación de factores de riesgo en las compañías, se requiere generar un módulo de segmentación basado en las buenas prácticas recomendadas por la Superintendencia financiera de Colombia, que ayude a mejorar el proceso de gestión de riesgo de lavado de activo y financiación del terrorismo.

Este documento se abordará la problemática previamente mencionada aplicando técnicas de Inteligencia Artificial (IA) y Machine Learning (ML) con el fin de realizar la segmentación establecida por la ley buscando optimización en el proceso de Identificación de transacciones, personas sospechosas en la organización que use GIRO como facilitador de dicho proceso.

Capítulo 1

Descripción del Problema

1.1. Planteamiento del Problema

Desde el año 2014, la Circular Básica Jurídica 029 de 2014 [2] de la Superintendencia Financiera de Colombia (SFC) establece que las entidades sometidas a su inspección y vigilancia deben implementar un Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT). Este sistema está diseñado para prevenir las pérdidas o daños que las entidades puedan sufrir debido a su vulnerabilidad de ser utilizadas como instrumentos para el LA/FT, lo cual representa una amenaza significativa para la estabilidad del sistema financiero y la integridad de los mercados, debido a su naturaleza global.

A pesar de ello, según el artículo 27 de la Ley 1121 de 2006 [3], no todas las entidades están obligadas a implementar un sistema de administración de riesgos LA/FT. No obstante, este artículo establece que, en cualquier proceso de contratación, se debe identificar tanto a personas naturales como jurídicas, así como el origen de sus recursos, con el fin de prevenir actividades delictivas.

En este contexto, diversas compañías de desarrollo software, buscan suplir la necesidad exigida por ley a las organizaciones de los diferentes sectores en Colombia por medio de herramientas tecnológicas que permita la adecuada gestión de riesgos con el fin de minimizar la materialización de riesgos relacionados con el lavado de activos y financiación del terrorismo en su interacción con los diferentes actores, así como para evitar sanciones impuestas por la superintendencia financiera de Colombia.

Optima Corporation, una entidad con 15 años de experiencia en el mercado , partner Oracle Corporation y Open Smartflex, en respuesta a esta necesidad, ha desarrollado un software para la Gestión Integral de Riesgo Organizacional (GIRO). Giro tiene como objetivo permitir a los usuarios y clientes la gestión oportuna en la identificación de posibles riesgos relacionados con el Lavado de Activos y la Financiación del terrorismo, pues es una herramienta diseñada para la gestión de diversos Sistemas de Administración de Riesgos, validación y debida diligencia de contrapartes y monitoreo de transacciones sospechosas.

Esta gestión se realiza por medio de diferentes módulos:

- **Administración de riesgos**: Integra y facilita la gestión del riesgo organizacional como Lavado de Activos, Financiación del terrorismo, Continuidad del negocio, liquidez, entre otros.

- **Monitoreo de Transacciones**: Diseñado para detectar en linea, operaciones inusuales y sospechosas.
- **Gestión Integral de Listas de Riesgos**: Valida clientes, usuarios, proveedores, empleados, accionistas y terceros de la organización en listas restrictivas y de medios, para detectar posibles vinculados con actividades de lavado de activos y financiación del terrorismo.

Adicional a los módulos que actualmente se tienen desarrollados en el aplicativo GIRO, con el fin de implementar herramientas a la vanguardia se requiere investigar, desarrollar e implementar un nuevo módulo que permita la segmentación de los diferentes factores de riesgos, descritos a continuación:

- Clientes - Actividad económica, volumen o frecuencia de sus transacciones y monto de ingresos, egresos y patrimonio.
- Productos - Naturaleza, características y nicho de mercado o destinatarios
- Canales de distribución - Naturaleza y características.
- Jurisdicción - Ubicación, características y naturaleza de las transacciones

Este desarrollo del módulo de segmentación, busca optimizar la gestión y costos operativos designados en la etapa de identificación de Riesgos LA/FT, de igual manera, minimizar el impacto generado por los errores operativos por la intervención manual en la identificación oportuna de los mismos factores de riesgo.

1.1.1. Formulación

¿Cómo realizar la segmentación de los factores de riesgo (Clientes, Productos, canales de distribución y/o Jurisdicción) en los datos del periodo 2023-2024 del aplicativo GIRO de Optima Corporation para un Cliente en Particular, cumpliendo con las normativas vigentes, para optimizar del proceso de identificación de riesgos llevado a cabo por los revisores fiscales u oficiales de cumplimiento?

1.1.2. Sistematización

- ¿Cómo realizar la segmentación de factores de riesgo usando técnicas de ciencia de datos?
- ¿Cuáles son los patrones de comportamiento en el conjunto de datos recolectados por GIRO en el periodo 2023 - 2024?
- ¿Cómo seleccionar los modelos de machine learning para la segmentación por conglomerados que permita el agrupamiento de los factores de riesgo en los datos recolectados?
- ¿Cómo medir la precisión y eficacia de los modelos de segmentación para identificación de factores de riesgo?

1.2. Objetivos

1.2.1. Objetivo General

Implementar un modelo para la segmentación de los factores de riesgo organizacional, empleando técnicas de Inteligencia Artificial (IA) o Machine Learning (ML) en los datos de GIRO para el periodo 2023-2024 (**FIDUCOLDEX**), con el fin de optimizar el proceso en la identificación de los riesgos LAFT, el cual se lleva a cabo por los oficiales de cumplimiento, según la normativa legal vigente.

1.2.2. Objetivos Específicos

- Realizar un análisis exploratorio de datos (EDA) exhaustivo para identificar patrones de comportamiento y relaciones relevantes en los datos y variables de GIRO para el periodo 2023-2024, que puedan influir en la segmentación de factores de riesgo LAFT.
- Implementar y entrenar diferentes modelos de Inteligencia Artificial (IA) o Machine Learning (ML), para la segmentación de factores de riesgo LAFT, evaluando su desempeño mediante métricas, con el fin de determinar el modelo más efectivo.
- Realizar la comparación de los resultados obtenidos en los modelos entrenados y determinar cuál modelo ofrece el mejor equilibrio entre las métricas de rendimiento, optimizando su implementación en el proceso de identificación de riesgos LAFT para mejorar la eficiencia del cumplimiento normativo.

1.3. Justificación

El lavado de activos y la financiación del terrorismo representan una gran amenaza para la estabilidad del sistema financiero y la integridad de los mercados debido a su carácter global y las complejas redes utilizadas para manejar estos recursos ilícitos. En Colombia, la Superintendencia Financiera (SFC, 2020)[4], a través de la Circular Externa Nro. 027, Capítulo IV, define el riesgo de lavado de activos y financiación del terrorismo (riesgo LA/FT) como: la posibilidad de pérdida o daño que puede sufrir una entidad vigilada por su propensión a ser utilizada, ya sea directamente o a través de sus operaciones, como instrumento para el lavado de activos, la canalización de recursos hacia actividades terroristas, la financiación de armas de destrucción masiva, o para ocultar activos provenientes de dichas actividades. Los riesgos asociados incluyen riesgos legales, reputacionales, operativos y de contagio, los cuales pueden tener un impacto económico negativo significativo en la estabilidad financiera de la entidad si se ve involucrada en actividades ilícitas.

Para combatir estos riesgos, los gobiernos y organismos internacionales han implementado regulaciones estrictas. A nivel global, el Grupo de Acción Financiera Internacional (GAFI) desempeña un papel fundamental al establecer estándares y promover la implementación de medidas legales,

regulatorias y operativas efectivas para prevenir el lavado de activos y la financiación del terrorismo. GAFI recomienda, entre otras cosas, la implementación de procedimientos para conocer a los clientes y para gestionar los riesgos asociados con LA/FT.

La segmentación de clientes es una herramienta esencial para cumplir con estas recomendaciones y obligaciones regulatorias. Al clasificar a los clientes según su perfil de riesgo, las instituciones financieras pueden aplicar medidas de diligencia debida adecuadas, asegurando que los clientes de alto riesgo reciban una vigilancia más rigurosa. Esto permite a las instituciones centrar sus recursos en áreas de mayor riesgo, optimizando así la vigilancia y el cumplimiento normativo, y detectando actividades sospechosas con mayor efectividad y rapidez.

Adoptar un enfoque proactivo en la segmentación de clientes no solo ayuda a prevenir que los servicios financieros sean utilizados para actividades ilícitas, sino que también mejora la toma de decisiones informadas sobre la aceptación y manejo de los clientes. Una segmentación adecuada permite diseñar políticas y procedimientos específicos para cada segmento de riesgo, protegiendo la reputación de la institución y demostrando un compromiso sólido con la integridad y seguridad financiera.

Finalmente, implementar una segmentación de clientes no solo minimiza el riesgo de sanciones, multas y otras penalidades por incumplimiento, sino que también protege la imagen corporativa y fortalece la confianza de los clientes y accionistas en la organización. Un enfoque riguroso en la gestión de riesgos contribuye a la sostenibilidad a largo plazo y promueve un crecimiento saludable y libre de actividades ilícitas, asegurando un entorno financiero seguro y confiable.

Capítulo 2

Marco de Referencia

2.1. Áreas Temáticas

2.1.1. Ciencia de Datos

A grandes rasgos [5], la ciencia de datos [5] es un conjunto de principios fundamentales que sustentan y guían la extracción de información y conocimientos a partir de datos. Posiblemente, el concepto más estrechamente relacionado con la ciencia de datos es la minería de datos: la extracción de conocimiento a partir de datos mediante tecnologías que incorporan estos principios. Existen cientos de algoritmos diferentes de minería de datos y una gran cantidad de detalles sobre los métodos de este campo.

La ciencia de datos engloba principios, procesos y técnicas para comprender fenómenos mediante análisis estadístico (automatizado) de datos, esto, con el fin de mejorar la toma de decisiones basadas en datos para compañías.

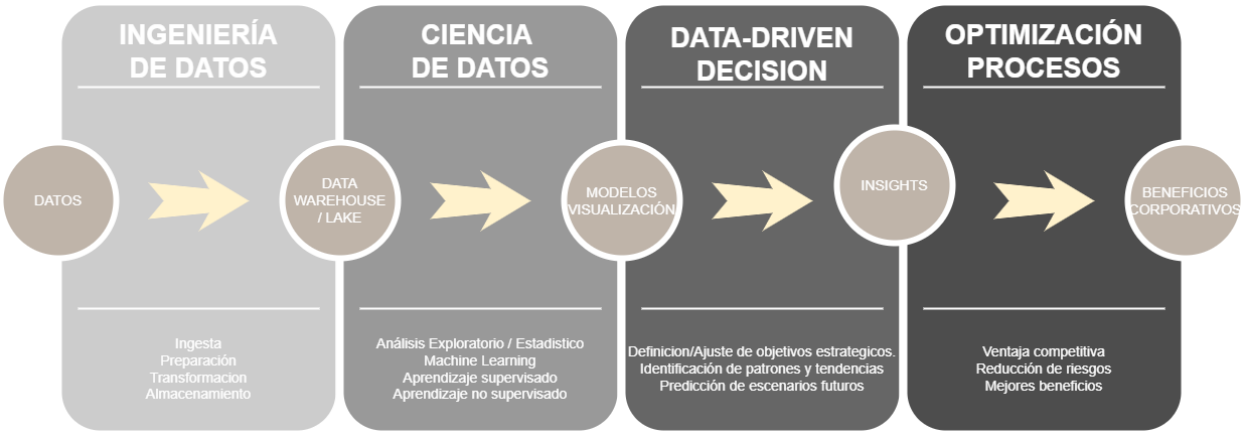


Figura 2.1: Flujo de Trabajo en Ciencia de Datos

Como se muestra en la Figura 2.1, el ciclo de vida de los datos, parte del proceso de **Ingeniería de datos**, en donde se toman desde las diferentes fuentes de información, y se pasan por procesos o pipelines que realiza transformación, limpieza, ingeniería de características, eliminación de variables, valores nulos, tratamiento de valores perdidos entre otras operaciones; Con el fin de estandarizar

y almacenar de manera centralizada (o descentralizada) la información para posteriormente ser consumida o usada para las demás fases.

El siguiente proceso, **Ciencia de datos**, basado en el análisis exploratorio de datos (EDA, por sus siglas en ingles) realizado para entender los patrones de información, comportamiento y tipos de datos, permite aplicar técnicas y metodologías estadísticas que para el entrenamiento y prueba de modelos de machine learning e inteligencia artificial enfocados en la necesidad de negocio con el fin de entender: ¿que ha sucedido?, ¿que puede suceder? o ¿que hacer para que suceda? cierto escenario entendiendo la información; De la misma manera, permite describir los datos usando técnicas de clasificación, agrupamiento entre otras.

Estos insight recibidos por los diferentes modelos desarrollados, entrenados y medidos en la paso previo, son usados para tomar **decisiones basadas en datos** para áreas estratégicas de la compañía con el fin de mitigar riesgos operativos, tácticos o estratégicos para **optimizar procesos** y mejorar los resultados de dichas actividades, al igual que estar a la vanguardia en relación a las demás compañías del medio.

2.2. Marco Teórico

2.2.1. Glosario

- **ASUM-DM:** Analytics Solutions Unified Method for Data Mining
- **BI :** Business Intelligence
- **EDA :** Análisis Exploratorio de Datos.
- **GAFI :** Grupo de Acción Financiera Internacional.
- **GIRO :** Gestión Integral de Riesgo Organizacional
- **ML :** Machine Learning (Aprendizaje de Máquina).
- **PCA:** Análisis de Componentes Principales.
- **SARLAFT:** Sistema de Administración del Riesgo de Lavado de Activos y Financiación al Terrorismo.
- **SFC:** Superintendencia Financiera de Colombia.
- **UIAF:** Unidad de Información y Análisis Financiero.
- **ROS :** Reporte de Operación Sospechosa

2.2.2. Normativa

2.2.2.1. Lavado de activos

Según el artículo 323 del Código Penal Colombiano, establecido en la Ley 599 de 2000[1] y actualizado por la Ley 1121 de 2006[2], el lavado de activos es una actividad criminal que busca ocultar el verdadero origen ilícito de los recursos que obtienen los criminales mediante operaciones financieras y no financieras, en las que usan a sectores de la economía de los países -incluido el comercio exterior y el mercado de capitales- para hacerlos parecer lícitos, lo que vuelve a esta una actividad transnacional en la que se involucra a varios países y sus economías.

Figura 2.2: Diagrama sobre el lavado de activos. Imagen tomada del Ministerio de Justicia de Colombia.

[1]Ley 599 de 2000: Código Penal Colombiano. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=6388

[2]Ley 1121 de 2006: Medidas para prevenir el lavado de activos y la financiación del terrorismo. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=21843

2.2.2.2. Financiación del terrorismo

Según el artículo 345 del Código Penal Colombiano, establecido en la Ley 599 de 2000[3], complementado por la Ley 1121 de 2006[4], la financiación del terrorismo se define como el acto de proporcionar, recaudar, distribuir o administrar recursos económicos, sabiendo que serán utilizados para la comisión de actos terroristas o destinados a grupos armados ilegales. Este delito es una amenaza significativa para la seguridad y la estabilidad financiera de las naciones.

2.2.2.3. SARLAFT

El Sistema de Administración del Riesgo de Lavado de Activos y Financiación al Terrorismo, es un mecanismo desarrollado por el Banco de la República para dar cumplimiento a la Circular Basica Juridica 029 de 2014 de la Superintendencia Financiera de Colombia. SARLAFT se estructura como un marco integral de gestión de riesgos que abarca las políticas, procedimientos y mecanismos que las entidades deben adoptar para mitigar estos riesgos y proteger la estabilidad del sistema financiero. Estas regulaciones suelen ser implementadas por los gobiernos y las empresas de cada sector.

SARLAFT se compone de cuatro etapas esenciales que permiten a las entidades financieras identificar, medir, controlar y monitorear los riesgos asociados al lavado de activos y la financiación del terrorismo (LA/FT):

1. Identificación:En esta fase, las entidades deben identificar los riesgos de LA/FT inherentes a su actividad, considerando factores como productos, mercados, jurisdicciones y tecnologías. Esto incluye la segmentación de los factores de riesgo y el análisis de cómo puede presentarse el riesgo en cada uno de ellos.

2. Medición o Evaluación: Una vez identificados los riesgos, se evalúa la probabilidad de que se materialicen y el impacto que tendrían. Este análisis puede ser cualitativo o cuantitativo y permite a las entidades determinar su perfil de riesgo inherente.

3. Control: En esta etapa, las entidades implementan medidas para reducir el riesgo inherente. El objetivo es minimizar la probabilidad de que el riesgo se materialice y su impacto en caso de ocurrir. Esto implica el diseño, evaluación y ajuste de las medidas de control.

4. Monitoreo: Las entidades realizan un seguimiento continuo del riesgo inherente y residual, comparando su evolución. Se implementa una matriz de riesgo y reportes periódicos para garantizar la efectividad de los controles y realizar ajustes si es necesario.

[3]Ley 599 de 2000: Código Penal Colombiano. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=6388

[4]Ley 1121 de 2006: Medidas para prevenir el lavado de activos y la financiación del terrorismo. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=21843

De igual forma, cuenta con ocho elementos cuyo objetivo es permitir el diseño, desarrollo e implementación o instrumentación, ordenada y metódica, de cada una de las etapas que componen el sistema de administración de riesgo.

1. Políticas: Directrices claras para la gestión de riesgos de LA/FT.
2. Procedimientos: Métodos detallados para llevar a cabo las políticas.
3. Documentación: Registro formal de las políticas y procedimientos.
4. Estructura organizacional: Definición de roles y responsabilidades.
5. Órganos de control: Supervisión interna para garantizar el cumplimiento.
6. Infraestructura tecnológica: Herramientas tecnológicas para soportar la gestión de riesgos.
7. Divulgación de información: Comunicación efectiva de las políticas a toda la organización.
8. Capacitación: Formación continua para el personal sobre la prevención de LA/FT.

2.2.3. Analítica de datos

La analitica de datos se ha convertido en un componente crucial en la toma de decisiones y la inteligencia de negocio para las empresas de diferentes gremios, con el fin de generar optimización de proceso, reducción de gastos o cumplimiento de normativas; Dentro de este último alcance, se dados los requerimientos actuales por las diferentes entidades regulatorias, se busca generar el entendimiento actual del negocio (Analitica Descriptiva) y de las transacciones que actualmente están siendo soportadas en las diferentes plataformas core de las compañías.

Con la analítica de datos, se busca realizar la revisión detallada de la información recolectada, el análisis estadístico y cuantitativo, modelos explicativos y predictivos, con el fin de generar gestión basada en hechos para impulsar las decisiones y acciones de la compañía.

2.2.3.1. Análisis Descriptivo

El análisis descriptivo/exploratorio se enfoca en revisar y resumir datos históricos para identificar patrones de comportamiento, o tendencias que ayudan a responder a la pregunta ¿Qué ha sucedido? . Este tipo de análisis es el punto de partida y fundamental para cualquier tipo de análisis de datos, ya que proporciona una visión clara del comportamiento pasado de los datos.

El análisis descriptivo y exploratorio cuenta con diferentes técnicas, Estadísticas Descriptivas, que involucran medidas como la media, mediana, moda, sesgos, desviaciones estándar entre otras para resumir las características que se definan como principales en los datos, por otro lado, los informes de gestión o resúmenes, buscan generar reportes que condensan información crítica para la toma de decisiones, estos informes suelen incluir tablas, gráficos entre otras herramientas visuales,

como gráficos, mapas de calor y dashboards elaborados en herramientas como Power BI, OAC, entre otras.

2.2.3.2. Análisis Diagnóstico

El análisis diagnóstico, busca explorar las razones detrás de los eventos, respondiendo la pregunta ¿Por qué sucedió?. Este tipo de profundizar en los datos para buscar las caídas de los patrones observados utilizando herramientas estadísticas más avanzadas, como el análisis de correlación, entre las diferentes variables para determinar si existe asociación significativa entre ellas, el análisis de la varianza, que permite comparar las diferentes variables para determinar diferencias entre al menos uno de ellos.

2.2.3.3. Análisis Predictivo

El análisis predictivo, usa los datos históricos combinados con modelos estadísticos y técnicas de aprendizaje de máquina (ML por sus siglas en inglés Machine Learning) para hacer proyecciones sobre eventos futuros, respondiendo a la pregunta ¿ Que podría suceder?, este análisis busca generar la toma de decisiones basadas en datos teniendo en cuenta el comportamiento de los mismos, permitiendo a las compañías anticiparse a posibles escenarios.

Entre otras técnicas, modelos y estrategias de aprendizaje de maquina, aprendizaje profundo o inteligencia artificial, se encuentran, regresión que estima la relación entre la variable dependiente y una o más variables independientes para predecir valores futuros, modelos de series de tiempo que usan datos secuenciales para predecir valores futuros basado en patrones (tendencias, estacionalidad, entre otras) históricos o también algoritmos como árboles de decisiones , redes neuronales convolucionales o clustering que pueden aprender de los datos para mejorar la predicción de las predicciones.

2.2.3.4. Análisis Prescriptivo

El análisis prescriptivo, es la siguiente fase del predictivo, no sólo anticipa lo que podría suceder sino que también sugiere que acciones hacer para optimizar los resultados futuros, este tipo de análisis busca contestar a la pregunta ¿Que debería suceder? y se basa en técnicas avanzadas como optimización que consisten en encontrar la mejor solución posible para un problema determinado usando técnicas como programación lineal, o algoritmos genéticos, sistema de recomendación , que usan grandes cantidades de datos para sugerir acciones que maximicen el beneficio o minimicen el riesgo.

2.2.4. Aprendizaje no supervisado

El aprendizaje no supervisado es un enfoque del Machine Learning que se centra en analizar y comprender los datos sin la necesidad de etiquetas o categorías predefinidas. En este tipo de aprendizaje, el objetivo no es realizar predicciones, sino descubrir patrones, relaciones o estructuras

ocultas en los datos, proporcionando una nueva perspectiva, una simplificación o un resumen de la información disponible. A través del análisis no supervisado, es posible identificar estructuras subyacentes que revelan nuevas agrupaciones, asociaciones entre las variables, o resúmenes que ayuden a comprender mejor los datos.

Un método estadístico útil es el análisis en clusters, el cual agrupa las observaciones en conjuntos o çlústeres"distintos, de manera que las observaciones dentro de cada grupo sean más similares entre sí, basándose en las variables medidas, que aquellas pertenecientes a otros grupos.

2.2.4.1. ¿Qué es Clustering/Segmentación?

El clustering como técnica de aprendizaje no supervisado juega un papel importante en segmentar clientes, productos basados en comportamiento basados en la similitud de los mismos.

Segmentar es el proceso de dividir o categorizar un conjunto de elementos en grupos más pequeños, basándose en características o criterios específicos que estos elementos tienen en común. El objetivo de la segmentación es identificar y entender mejor las diferencias dentro de un grupo más amplio, para poder dirigir estrategias, decisiones o acciones de manera más efectiva y personalizada hacia cada segmento identificado.

Con base en esto, la segmentación de clientes es un proceso estratégico que implica dividir la base de clientes de una organización en grupos distintos, basándose en características específicas que son relevantes para identificar y gestionar riesgos. En el contexto de la prevención de riesgos de Lavado de Activos y Financiamiento del Terrorismo (LA/FT), la segmentación de clientes se utiliza para clasificar a los clientes según su perfil de riesgo en relación con actividades ilícitas.

Clustering - El análisis de clustering o conglomerados es un método estadístico de aprendizaje automático que a partir de datos y las diferentes iteraciones sobre los mismos, busca generar homogeneidad entre los integrantes de cada grupo y heterogeneidad entre los diferentes grupos, es decir, los individuos que sean considerados similares podrían pertenecer al mismo grupo y los individuos que sean considerados distintos sean asignados a grupos diferentes.

Principios de análisis de clustering o conglomerados:

1. Se puede realizar el análisis usando múltiples variables (Cualitativas y Cuantitativas) para la clasificación de acuerdo a las características.

2. Muestra la concentración de datos para su agrupamiento eficiente de acuerdo a su homogeneidad.

3. La selección del principio de distancia (Ecuclidianta, descenso de gradiente u otros), permite la adecuación del modelo de acuerdo a las características de los datos o componentes con los que se cuente.

4. La cantidad de grupos es desconocida pero se puede establecer de acuerdo a las características de los datos o componentes.

5. La identificación de valores atípicos es un resultado alterno que brinda valor al cliente logrando identificar esos individuos que son objetivo de análisis posterior por parte del cliente.

2.2.4.2. Algoritmos de Clustering

Análisis de Componentes Principales

K-Means, que busca dividir los individuos en diferentes grupos, donde cada uno pertenece al grupo (cluster) con la media más cercana. Pseudocódigo k-means.

1. Se eligen k puntos iniciales como centroides

2. Se asigna cada punto al centroide más cercano.

3. Se recalculan los centroides como la media de todos los puntos asignados a ese cluster.

4. Se repiten los pasos 2 y 3 hasta que los centroides no cambien significativamente.

k-Nearest Neighbors o k-NN, no paramétrico para clasificacion y regresion, busca generar grupos basados en su distancia de datos, es decir, KNN se basa en la idea de que los objetos similares tienden a estar cerca unos de otros en el espacio de características. Pseudocódigo k-nn.

1. Para un punto de datos dado, se calculan las distancias* a todos los otros puntos en el conjunto de datos.

2. Se seleccionan los k puntos más cercanos (donde k es un parámetro definido por el usuario).

3. En clasificación, se asigna la clase mayoritaria entre estos K vecinos.

4. En regresión, se calcula el promedio (o mediana) de los valores de los K vecinos.

Hierarchical Clustering, o agrupamiento jerárquico, es un método de agrupamiento que busca construir una jerarquía de clusters. Este enfoque no requiere un número predefinido de clusters, lo que lo diferencia de métodos como K-Means. Existen dos tipos principales de clustering jerárquico:

- **Aglomerativo (bottom-up):** comienza con cada punto como un cluster individual y los fusiona gradualmente hasta formar un único cluster que agrupe a todos los puntos.
- **Divisivo (top-down):** comienza con todos los puntos en un único cluster y los divide sucesivamente en clusters más pequeños.

En ambos casos, la elección de qué puntos o clusters agrupar o dividir depende de una métrica de distancia (como distancia euclidiana o Manhattan) y un criterio de enlace (como enlace simple, completo o promedio).

Pseudocódigo para clustering jerárquico aglomerativo:

1. Comenzar asignando cada punto de datos a un cluster individual.
2. Calcular las distancias entre todos los clusters (utilizando una métrica de distancia y un criterio de enlace).
3. Combinar los dos clusters más cercanos en uno solo.
4. Repetir los pasos 2 y 3 hasta que todos los puntos estén en un único cluster.
5. Representar los resultados mediante un dendrograma, que muestra la estructura jerárquica de los clusters.

Búsqueda del mejor número de clusters (k) y parámetros: Aunque el clustering jerárquico no requiere especificar k previamente, es importante determinar el número óptimo de clusters al interpretar el dendrograma. Esto se puede realizar utilizando métodos como:

- **Corte del dendrograma:** Identificar un nivel en el dendrograma donde los clusters sean homogéneos y estén bien separados.
- **Índice de Silueta:** Calcular el índice de silueta para diferentes valores de k. Este índice mide qué tan bien se encuentra un punto en su cluster en comparación con otros clusters. Un valor cercano a 1 indica una buena separación.
- **Coeficiente de cohesión y separación:** Evaluar métricas como cohesión (qué tan cercanos están los puntos dentro de un cluster) y separación (qué tan lejanos están los clusters entre sí).
- **Elbow Method:** Aunque más común en K-Means, también puede usarse aquí midiendo cómo cambia la inercia o la suma de distancias dentro de los clusters a medida que aumenta k.

Ajuste de parámetros: En clustering jerárquico, los resultados dependen de:

- **Métrica de distancia:** Elegir una métrica adecuada según las características de los datos. Por ejemplo, la distancia euclidiana es útil para datos continuos, mientras que la distancia de Jaccard es más adecuada para datos binarios.
- **Criterio de enlace:** Seleccionar un criterio que defina cómo medir la distancia entre clusters. Por ejemplo:
  - **Enlace simple:** Utiliza la distancia mínima entre puntos de dos clusters.
  - **Enlace completo:** Usa la distancia máxima entre puntos de dos clusters.
  - **Enlace promedio:** Calcula la distancia promedio entre puntos de dos clusters.

La combinación de estas técnicas permite encontrar una configuración óptima para obtener clusters significativos y representativos.

El dendrograma permite visualizar cómo se forman los clusters a diferentes niveles de similitud, facilitando la identificación de un número adecuado de clusters al cortar el árbol en un nivel específico.

2.3. Trabajos Relacionados

En seguida se detalla de manera breve algunas investigaciones que se han desarrollado en el marco de los riesgos asociados con el lavado de activos y la financiación del terrorismo LA/FT. Dentro de este grupo de investigaciones hay algunas que tratan los riesgos asociados a LA/FT empleando en su metodología variadas técnicas. Otras investigaciones desarrollan la técnica de segmentación en casos aplicados de LA/FT.

- *Segmentación de clientes y definición de alertas para la prevención de riesgos de lavado de activos y financiación del terrorismo (SARLAFT): un estudio económico aplicado a entidad financiera colombiana en 2017 (Amaya Molina, 2017)* [6].

  El trabajo destaca la importancia que tiene el priorizar la deteccion y fortalecer los controles de crimenes financieros tales como el lavado de activos y la financiacion del terrorismo puesto que, son los crimenes que se cometen con mas frecuencia alrededor de todo el mundo. Estos vulneran el equilibrio económico y a su vez constituyen una problemática cuando de relaciones comerciales internacionales se habla.

  Además se destaca la importancia de la aplicación de metodologías avanzadas como los son la minería de datos ya que, en la actualidad crean un panorama completamente nuevo que ayuda mucho en la generación de controles para prevenir y detectar este tipo de actividades.

  Al igual que en el trabajo relacionado (Amaya Molina, 2017) [6], una de las metodologías de aprendizaje no supervisado que se implementará en este proyecto será la de K-means. Esta técnica nos ayudará a identificar de manera efectiva, a los individuos que son bastante homogéneos dentro de cada grupo y diferentes entre grupos, permitiendo así caracterizarlos adecuadamente, definir perfiles y construir controles.

- *Modelo Matemático para Estimar el Riesgo de Lavado de Activos en Clientes de Pequeñas Instituciones Financieras (Enriquez Sanchez, 2019)* [7].

  Siguiendo el enfoque utilizado en trabajo relacionado (Enriquez Sanchez, 2019)[7], se inició con una fase de entendimiento del contexto funcional, que incluyó la participación en varios eventos organizados por Optima Corporation. Además, se llevaron a cabo reuniones con el Director de Operaciones y Tecnología y el Director Administrativo de la compañía, con el objetivo de comprender el estado actual de GIRO y su capacidad para manejar la información relevante.

  Como resultado de estas sesiones, se obtuvo una autorización escrita y un acuerdo de confidencialidad para el uso de una base de datos extraída de GIRO, que abarca un período de un año y medio, con información sobre clientes, productos, jurisdicciones y canales de distribución de GIRO.

  Al igual que en trabajo relacionado (Enriquez Sanchez, 2019)[7], para asegurar la adecuada selección de datos y el modelo matemático, se revisaron diversas técnicas considerando la

cantidad y calidad de las variables presentes en las bases de datos proporcionadas por Optima Corporation. Se llevó a cabo una selección de características (feature selection) basándose en criterios de expertos, como el Oficial de Cumplimiento, priorizando una alta interpretabilidad de la información. Esta aproximación permite contrastar las entradas de diferentes modelos de clustering que podrían ser utilizados para la segmentación de clientes. Por esta razón, se decidió emplear una metodología como el Análisis de Componentes Principales (PCA) para identificar las características que mejor se ajusten al modelo, considerando las entradas de los modelos de clasificación seleccionados.

- *Análisis de segmentación y alertamiento transaccional para la gestión de riesgos SARLAFT en el sector financiero* [8].

  El trabajo se centra en la identificación de patrones transaccionales inusuales mediante la segmentación de clientes, utilizando el enfoque CRISP-DM. Este método es clave para que las instituciones financieras puedan detectar actividades sospechosas, especialmente relacionadas con el lavado de activos y la financiación del terrorismo. A través del análisis de una muestra de más de 217.000 clientes asalariados, los autores lograron clasificar a los usuarios en diferentes grupos, basándose en características financieras y transaccionales, lo que facilitó la emisión de alertas preventivas ante comportamientos atípicos.

  El modelo K-Means fue la herramienta principal utilizada para la segmentación de clientes, junto con el método del codo, que permitió determinar que ocho clusters eran óptimos para representar los patrones financieros de la población estudiada. Esta técnica fue complementada con el uso de métricas como el Z-score y el rango intercuartílico (IQR), que sirvieron para identificar transacciones anómalas. El enfoque no solo optimiza la detección de riesgos financieros, sino que refuerza el cumplimiento de las normativas SARLAFT, contribuyendo a mitigar riesgos asociados a actividades ilegales.

  En nuestro proyecto, se tomarán las recomendaciones del trabajo de Correa Correa y Montoya Gómez para mejorar nuestra estrategia. Al igual que el trabajo mencionado, se implementará el modelo K-Means como herramienta de segmentación, de ser necesario se emplea el método del codo para optimizar el número de clusters.

- *Fraud Detection and Prevention* [9].

  Enmarcado en el contexto de Sistema de Administrador del Riesgo de Lavado de Activos y Financiación del terrorismo, se encuentra la definición de Fraude el cual según [9], se describe como Engaño ilícito o delictivo destinado a obtener un beneficio económico o personal.

  Los fraudes se cometen con la intención de no ser detectados. A veces el motivo es la codicia de algunos individuos por falsificar registros y/o modificar las transacciones para obtener su propio beneficio, la mayoría de los fraudes contables tienen ánimo de lucro y este beneficio se obtiene utilizando transacciones falsas.

Las organizaciones (bajo requisito obligatorio) deben establecer controles tales como SARLAFT con el fin de prevenir tener relaciones o negociaciones con personas involucradas con actividades ilícitas o que tengan fines de financiación a actividades ilícitas, de igual manera El fraude contable consiste en manipular deliberadamente la información contable de la compañía, ya sea compartiendo información errónea con los inversores omitiendo cierta información para reducir los beneficios y la carga fiscal o beneficiar a un tercero o a uno mismo.

Los auditores o revisores fiscales realizan controles internos para supervisar e informar de lo que ocurre en la empresa según lo establecido por la ley colombiana, con el fin evaluar su eficacia para evitar errores y comprobar la fiabilidad de la información facilitada a las partes interesadas.

En [9] se describen los mecanismo para combatir el fraude y se centran en algunos en particular para poder hacerlo tales como análisis predictivo sobre la información recolectada sobre auditores externos en el país de Líbano en el cual se realizaron encuestas las cuales incluían sexo, edad, nivel de estudios y años de experiencia en auditoría, esta información fue analizada usando regresión Lineal, en donde se buscaba probar ciertas hipótesis en donde determinaban la necesidad de establecer o fortalecer los controles para regular el fraude en el país Líbano.

En nuestro caso buscaremos centrarnos el control establecido que tiene como base normativa la Circular Básica Jurídica de la Superintendencia Financiera de Colombia, numeral 4.1.1.1, Título IV, Parte I en Colombia, el cual describe el control y las buenas prácticas para poder detectar de manera oportuna actores que tengan esta relación con el fin de minimizar el impacto en el ámbito financiero y de imagen de la compañía, esta Circular Básica Jurídica establece también que para poder tener un etiquetado adecuado a los riesgos se requiere establecer la segmentación con el fin de clasificar de manera adecuada de acuerdo al comportamiento de los clientes , productos, canales de comunicación y/o jurisdicción.

Contrastando contra [9], en Colombia se cuentan con regulaciones y aplicativos como GIRO que permiten reconocer de manera oportuna el control al lavado de activos y posible financiación del terrorismo, para nuestro trabajo, buscaremos analizar los datos recolectados en GIRO que permitan identificar y clasificar las variables descritas en la circular básica jurídica, utilizando técnicas de aprendizaje automático no supervisado en donde, de acuerdo a las características de los datos, se pueda establecer una segmentación de datos adecuada.

- *El perfil financiero: una estrategia para detectar el lavado de activos*[10].

  El perfilamiento de clientes tanto en el entorno financiero como en otros sectores económicos funge como herramienta para prevenir y detectar el lavado de activos, este proceso consiste en la recopilación, almacenamiento y análisis de los clientes y sus movimientos financieros para establecer perfiles, que acompañado de las características de cada transacciones y de acuerdo a lo que se espera de cada uno de ellos se realiza la comparación con transacciones que se consideran normales.

Dentro de lo analizado por Lazono Vila [10], se describe que de acuerdo a las transacciones y operaciones normales, se usan herramientas estadísticas y de minería de datos que permiten predecir el comportamiento de los clientes conociendo sus características socioeconómicas y demográficas , esta información se espera encontrar en los datos recolectados por GIRO con el fin de realizar un análisis exploratorio, moldeamiento y pruebas de modelos estadísticos y de aprendizaje automático con el fin de determinar los modelos que más se acomoden a los datos.

Según lo descrito por Lozano Vila [10], pueden existir dos grupos de herramientas tecnológicas que apoyen la labor de detección y perfilamiento de clientes, Generación de conocimiento la cual permite describir y perfilar los clientes de acuerdo a las características socio económicas y demográficas; Y un segundo grupo es la generación de alertas de acuerdo al comportamiento o detección de anomalías.

Particularmente GIRO cuenta con herramientas que permiten a los revisores fiscales y oficiales en el proceso de generación de conocimiento el análisis de la información para poder así determinar posibles sospechas para poder reportar ante la Unidad de Información y Análisis Financiero, del Ministerio de Hacienda y Crédito Público (UIAF) posible un posible Reporte de Operación Sospechosa (ROS).

De igual manera GIRO, permite realizar un monitorio constante de transacciones que genera alertas (En caso de habilitar el módulo o suscripción correspondiente) sobre las transacciones que se consideran estar fuera de los normales para que el revisor fiscal o oficial de cumplimiento ejecute el proceso de revisión correspondiente.

Dentro de estos módulos, y de acuerdo a lo descrito por la Circular Básica Jurídica (Anexo 1), el grupo de herramientas (Dentro de la que está enmarcada GIRO) que realiza la gestión del conocimiento, requiere un proceso de segmentación automática de acuerdo a las diferentes características de clientes a través de la aplicación de métodos estadísticos y procesos de aprendizaje automático.

- *El delito de lavado de activos: su complejidad y las dificultades de su investigación* [11].

  En su artículo, Herrera (2018)[11] explora cómo la complejidad inherente de delitos como el lavado de activos y la financiación del terrorismo genera desafíos significativos en su investigación, judicialización y obtención de sentencias. Estas dificultades representan un reto que debe ser afrontado en Colombia por la Fiscalía General de la Nación.

  Menciona que la literatura sobre el lavado de activos no es muy rica, ni se ha abordado hasta ahora este fenómeno de manera profunda, solo las entidades oficiales que de una manera u otra están vinculadas con la lucha contra este delito, dan luces por medio de sus informes acerca de la naturaleza del lavado de activos.

  Además, comenta unos resultados interesantes, en una investigación de lavado de activos en

Colombia se tienen asignados en total 1806 radicados, de los cuales el 93 % están en etapa de indagación, el 1.9 % están en etapa de investigación, el 4.4 % están en etapa de juicio y un 0.05 % están en ejecución de penas. Llama la atención el gran número de procesos en la etapa de indagación, lo cual permite concluir que hay dificultades en determinar la existencia del hecho delictivo y esto obstaculiza el avance a etapas posteriores de la investigación criminal.

Lo anterior parece confirmado por la inmensa cantidad de investigaciones sobre lavado de activos que son archivadas. Menciona que entre el año 2005 y el año 2017 se han archivado 1347 investigaciones, mientras que 212 han culminado en sentencia condenatoria. La mayoría han sido archivada por la inexistencia o atipicidad del delito.

Herrera concluye que entre las principales dificultades para la investigación de delitos como el lavado de activos se encuentran: La dificultad de vincular el lavado de activos con la ilicitud del delito subyacente (es una dificultad de índole probatoria, hay ausencia de prueba directa del delito subyacente). Otra dificultad es la carga de trabajo de los funcionarios de la Fiscalía General de la Nación, al asignarles muchos casos provoca que no terminen y que no cumplan con ninguna de las órdenes.

Por último, hay dificultades de orden organizacional que también son obstáculos para la investigación. Es necesario tener un equipo multidisciplinario y suficiente para adelantar las investigaciones del lavado de activos (vinculación y capacitación) Todos estos desafíos subrayan la importancia de realizar este trabajo de grado ya que, el análisis y detección de este tipo de amenazas como lo son el lavado de activos y financiación del terrorismo son algo que no se tiene del todo estandarizado, ademas de que la información disponible es escasa. Con este estudio, esperamos obtener conclusiones valiosas que puedan ayudar a futuras investigaciones en este campo.

- *Elaboración de un modelo de segmentación de jurisdicciones que aporte a la identificación de riesgos de Lavado de Activos y Financiación del terrorismo por este factor en una institución microfinanciera de la ciudad de Popayán* [12].

  El trabajo de Nayibe Lizzeth Daza Hoyos (2019)[12] desarrolla un modelo para segmentar riesgos de LA/FT utilizando la metodología CRISP-DM, que estructura el análisis de datos y ajusta estrategias de prevención según las características de cada región. A través de la construcción de matrices de riesgo geográfico, se identifican patrones anómalos en las transacciones, mejorando la detección de operaciones sospechosas y asegurando el cumplimiento de las normativas de la Superintendencia Financiera de Colombia.

  El uso de algoritmos de segmentación permite definir perfiles de riesgo precisos, optimizando la gestión de riesgos en instituciones financieras. Este enfoque ofrece una visión detallada de los riesgos potenciales, lo que facilita la implementación de controles específicos y la toma de decisiones informadas.

Los resultados de esta investigación son relevantes, ya que proporcionan un marco metodológico adaptable para segmentar riesgos en diferentes contextos financieros, como el módulo GIRO. La aplicación de estas técnicas avanzadas fortalece el diseño de estrategias de prevención más efectivas, contribuyendo a una mejor identificación de riesgos y al cumplimiento regulatorio en el sector financiero colombiano.

- *Desglose de Factores de Riesgo en LA/FT: Un Enfoque de Segmentación* [13].

  La segmentación de factores de riesgo es una estrategia clave para gestionar eficazmente los riesgos asociados al lavado de activos y la financiación del terrorismo (LA/FT). Este enfoque permite a las entidades financieras identificar, evaluar, clasificar y mitigar los riesgos de manera estructurada. Entre los principales factores de riesgo se destacan, el tipo de cliente o contraparte, la ubicación geográfica y la eficacia de los controles internos. Clientes con transacciones complejas o ubicados en regiones con altos niveles de corrupción y crimen organizado representan mayores riesgos, al igual que las entidades con controles internos débiles o inadecuados.

  El proceso de segmentación consta de varias etapas. La primera es la identificación de factores de riesgo, mediante la recopilación de información de diversas fuentes para obtener una visión integral. Luego, se realiza una evaluación que considera la probabilidad de que un factor se materialice y su impacto potencial, utilizando métodos cuantitativos y cualitativos. Posteriormente, los factores se clasifican en segmentos según su nivel de riesgo, empleando técnicas como el análisis de datos. Finalmente, se desarrollan estrategias de mitigación adaptadas a cada segmento, basadas en mejores prácticas y recursos disponibles.

  La segmentación facilita una comprensión más profunda de los riesgos, permitiendo analizar cada factor en detalle y tomar decisiones informadas. Además, ayuda a identificar patrones y tendencias que pueden no ser evidentes al observar los factores de riesgo de manera aislada. Por ejemplo, ciertos tipos de clientes o productos podrían compartir características de riesgo similares, lo que permite desarrollar controles específicos, como mejoras en la diligencia debida del cliente o en los sistemas de monitoreo de transacciones.

  El enfoque de esta investigación no solo mejora la efectividad de las estrategias de mitigación, sino que también optimiza la eficiencia operativa. Al enfocar los recursos en los segmentos de mayor riesgo, las entidades financieras pueden priorizar sus esfuerzos y gestionar mejor sus capacidades. Así, la segmentación de riesgos no solo reduce la exposición al LA/FT, sino que también fortalece la gestión integral del riesgo y la sostenibilidad operativa de las organizaciones.

- *Metodología para segmentación de un SARLAFT* [14].

  El documento titulado Metodología para segmentación de un SARLAFT de Lincoln Pérez (2020) ofrece un marco integral para abordar la segmentación en el contexto de los sistemas de administración del riesgo de lavado de activos y financiación del terrorismo (SARLAFT). En el

trabajo se desarrollan métodos estadísticos y de aprendizaje automático no supervisado como herramientas clave para clasificar grupos homogéneos y heterogéneos según variables relevantes dentro del ámbito financiero, en cumplimiento de la normativa colombiana establecida en la Circular Básica Jurídica de la Superintendencia Financiera de Colombia.

La investigación destaca la aplicabilidad de técnicas como K-means, K-medoids y CLARA, evaluadas mediante índices robustos de validación de clusters para garantizar una segmentación precisa. Asimismo, propone soluciones prácticas que combinan algoritmos y métodos optimizados para la realidad de las entidades financieras, superando limitaciones previas en términos de tiempo y recursos computacionales. Un valor adicional del estudio es su capacidad de adaptación a distintos escenarios mediante simulaciones que consideran datos anonimizados y perturbados, asegurando la privacidad y la aplicabilidad generalizada.

En el contexto de nuestro proyecto de grado, esta metodología es relevante porque establece una estructura analítica robusta que podemos adaptar para el desarrollo de nuestro módulo en GIRO. La implementación de técnicas de segmentación basadas en aprendizaje automático nos permitirá identificar patrones en clientes, productos y canales, alineándonos con las regulaciones nacionales para minimizar riesgos de lavado de activos y financiación del terrorismo. Además, el enfoque propuesto por Pérez Lincon sobre índices de validación de clusters nos ofrece criterios concretos para evaluar la eficacia y confiabilidad de las agrupaciones generadas.

- *Aplicación de Técnicas de Minería de Datos en la Detección de Fraude* [15]

  El artículo examina detalladamente cómo las herramientas de minería de datos han evolucionado para enfrentar desafíos relacionados con la identificación de actividades fraudulentas en diversos contextos financieros. En su análisis, se presentan métodos como el clustering, los árboles de decisión y las redes neuronales, además de estrategias híbridas que integran varias técnicas para incrementar la precisión y efectividad en la detección de patrones anómalos. Se resalta el papel fundamental de los datos de calidad y la capacidad de los algoritmos para aprender de conjuntos de datos complejos y adaptarse a nuevos escenarios, lo que resulta clave para garantizar resultados consistentes.

  La investigación enfatiza que la elección de la técnica más adecuada depende del entorno y las características del problema a resolver. Por ejemplo, las redes neuronales pueden ser útiles en situaciones donde se necesita identificar relaciones no lineales, mientras que los métodos de agrupamiento son efectivos para descubrir grupos naturales en datos no etiquetados. Asimismo, se destacan los beneficios de los enfoques híbridos, los cuales combinan diferentes metodologías para aprovechar las fortalezas individuales de cada una, maximizando así su capacidad de identificar comportamientos fraudulentos.

  Este artículo resulta especialmente relevante para nuestro proyecto de grado, que busca desarrollar un módulo de segmentación para el aplicativo GIRO. Las estrategias de minería de datos descritas en el documento son esenciales para implementar un sistema robusto que per-

mita clasificar clientes y operaciones según su nivel de riesgo. La incorporación de enfoques híbridos y métodos de aprendizaje automático no supervisado ofrece una oportunidad para mejorar la identificación de comportamientos sospechosos. Además, la capacidad adaptativa mencionada en el artículo es fundamental para garantizar que el sistema pueda responder eficazmente a nuevas amenazas y desafíos en el contexto de la prevención del lavado de activos y la financiación del terrorismo. Esto asegura un enfoque dinámico, eficiente y alineado con las necesidades normativas y operativas.

Capítulo 3

Metodología

En este proyecto se ha adoptado la metodología ASUM-DM (Analytic Solutions Unified Method for Data Mining), la cual estructura el desarrollo de soluciones analíticas en diferentes fases, desde el análisis y el diseño hasta la optimización y el despliegue. Adicionalmente, se ha implementado la Medallion Architecture como referencia para la gestión y el procesamiento de datos. Esta arquitectura desarrollada en el contexto de Delta Lake, ofrece una estructura escalable y confiable para la integración de datos en proyectos de ciencia de datos.

3.1. Metodología ASUM-DM

La metodología ASUM-DM de IBM [16] [17], la cual se encuentra establecida con base en la metodología **CRISP DM** por sus siglas en inglés, Cross Industry Standard Process for Data Mining [18], es una metodología que presenta un proceso jerárquico en donde se tienen unas tareas para hacer minería de datos de una manera ágil y con mejores resultados.

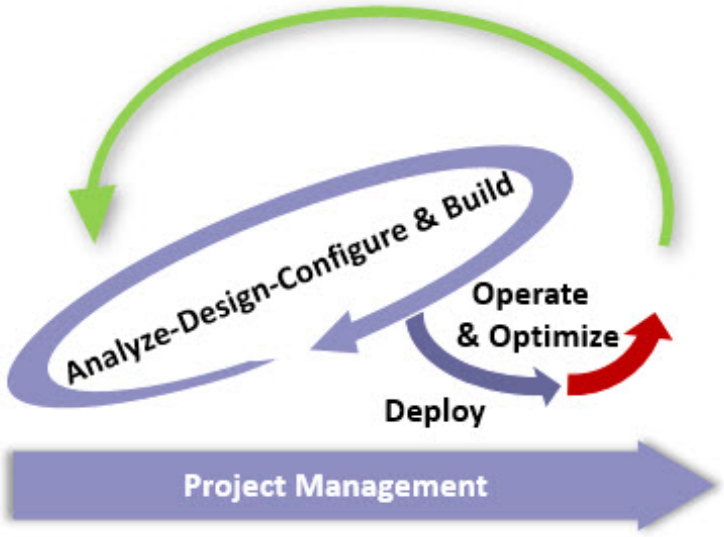


Figura 3.1: Fases de la metodología ASUM-DM
[16]

Particularmente ASUM-DM es un enfoque estructurado para el desarrollo y despliegue de so-

luciones de minería de datos y analítica avanzada. Enfatiza en varias de las nuevas prácticas en la ciencia de datos, como el uso de volúmenes de datos muy grandes, la incorporación de análisis de texto en el modelado predictivo y la automatización de algunos procesos. A continuación se detallan sus respectivas fases:

3.1.1. Análisis, diseño, configuración y construcción

La fase de Análisis busca comprender los objetivos y necesidades desde una perspectiva empresarial, para convertirlos en un problema abordable con técnicas analíticas. Esto implica aplicar enfoques estadísticos, minería de datos o aprendizaje automático para abordar tareas como predicción, segmentación, clasificación o detección de relaciones entre datos. Una vez finalizado el entendimiento del negocio, se inicia la fase de entendimiento de los datos, que implica varias actividades clave:

- Recolección de datos iniciales: Se identifican y recopilan los datos disponibles dentro del negocio, evaluando si es necesario adquirir nuevos datos externos para completar el análisis.
- Análisis exploratorio de datos: Se exploran los datos utilizando gráficos, tablas, diagramas y otras visualizaciones con el fin de identificar patrones, características importantes y relaciones entre variables.
- Verificación de calidad de los datos: Se revisa si los datos están completos, identificando posibles inconsistencias, valores faltantes o anomalías que puedan afectar el análisis.
- Generación de un reporte: Finalmente, se elabora un informe detallado que resume los hallazgos sobre la calidad, disponibilidad y características de los datos, proporcionando una base sólida para el siguiente paso en el proyecto.

En la fase de diseño se llevan a cabo varias actividades clave relacionadas con la infraestructura técnica y de seguridad. Primero, se diseña y valida toda la arquitectura necesaria para los ambientes de pruebas y producción, asegurando que la infraestructura técnica cumpla con los requisitos del proyecto. Esto incluye la planificación y creación de estrategias sólidas de autenticación y autorización para garantizar la seguridad de los datos y el acceso a los sistemas. Una vez completado el diseño, se elabora un documento que valida toda la infraestructura diseñada, el cual es revisado y aprobado por los interesados. Finalmente, se procede a la configuración del ambiente, ya sea en un entorno local o en la nube, preparando todo el sistema para su implementación.

En la fase de configuración, se lleva a cabo la preparación de los datos, una actividad crucial para asegurar que los datos sean aptos para el análisis. En primer lugar, se seleccionan los datos que se utilizarán, basándose en criterios como la calidad de la información, restricciones de los datos y los objetivos del negocio. A continuación, se realiza una limpieza de los datos para corregir errores, eliminar valores inconsistentes o duplicados, e imputar valores faltantes si es necesario. También puede ser necesario agregar nuevas columnas para mejorar el análisis.

Si los datos provienen de múltiples fuentes, se realiza la integración de los datos, teniendo en cuenta procesos de agregación para computar valores nuevos a partir de valores que ya se encuentren en las tablas. Posteriormente, los datos se formatean, unificando los formatos (como fechas, tipos de datos o unidades) para asegurar consistencia en la presentación. Finalmente, se elabora un informe detallado que documenta todo el proceso de preparación de los datos, describiendo las decisiones tomadas y las transformaciones aplicadas.

En la fase de construcción se lleva a cabo la etapa de construcción del modelo, que se realiza de forma iterativa y, en ocasiones, puede requerir volver a la etapa de preparación de datos para realizar ajustes. Esta etapa incluye tres actividades principales:

- **Selección de técnicas de modelado:** Se eligen las técnicas más adecuadas para el análisis, basándose en las metas del proyecto y las métricas de evaluación, con el objetivo de identificar el modelo que mejor se ajuste a los objetivos del negocio.
- **Generación y diseño de pruebas:** Generar un procedimiento para evaluar la calidad y validez del modelo. Esta actividad está enfocada más en modelos supervisados, en el cual se divide el conjunto de datos en dos partes: una para entrenamiento y otra para prueba. El modelo se construye con el conjunto de entrenamiento y su calidad se estima utilizando el conjunto de prueba separado.
- **Construcción del modelo:** Con los datos ya preparados, se procede a la implementación del modelo. En caso de que se detecten problemas de calidad en los datos o inconsistencias, se puede iterar y ajustar el proceso, volviendo a la fase de preparación de datos de ser necesario.

Una vez finalizada la construcción del modelo, se procede a su evaluación para determinar si los resultados obtenidos son adecuados para el proyecto. En esta etapa, se verifica que los modelos cumplan con los criterios previamente establecidos y se identifican posibles hallazgos nuevos. Asimismo, se revisa el proceso completo, evaluando las tareas realizadas, los errores detectados y los pasos que no fueron ejecutados correctamente o resultaron innecesarios. Finalmente, se puede elaborar un informe que abarque los procesos implementados, las fallas encontradas y las oportunidades de mejora.

Finalmente, se lleva a cabo la fase de configuración y prueba en el entorno de QA. En primer lugar, se cargan todos los archivos necesarios y se valida que el entorno de QA, la configuración de seguridad y los sistemas de respaldo hayan sido correctamente establecidos. Posteriormente, se migra el modelo al ambiente de pruebas, se realizan las pruebas necesarias obteniendo la retroalimentación de los usuarios. Con base en los resultados de las pruebas y la aprobación del cliente, se toma la decisión final sobre el despliegue del modelo en producción.

3.1.2. Despliegue

La fase de **Despliegue** es el momento en que se lleva a ambiente produccion, la configuracion, datos y demás consideraciones que permitan entregar al usuario final, stakeholders y sponsors la

solución o modelo que resuelva el requerimiento inicial establecido y tomado en la fase previa.

De igual manera, se realiza la documentación de cierre del proyecto, el cual busca generar un consolidado detallado de la solución dada a los objetivos planteados inicialmente.

Las actividades clave de esta fase se encuentran enmarcadas en los diferentes momentos de la instalación en producción del producto, modelo o solución al requerimiento solicitado por el cliente, como primer paso, se debe dar el contexto y la transferencia de conocimiento no analitico de la solución, al grupo de usuario/s encargado/s de manejar la solución una vez se entregue.

Como siguiente paso, se debe establecer el plan hora ahora de la instalación tanto los pasos como los responsables de los mismos, este plan también comprende los puntos clave (Check Point) en donde se ejecutan las pruebas previamente planeadas para garantizar que la solución está siendo instalada de manera adecuada, o también los puntos clave en donde se determina hacer rollback de la misma solución.

Como última fase, se busca establecer de forma detallada en el informe de cierre los resultados del proyecto, las lecciones aprendidas y las lecciones futuras a ejecutar, al igual que los canales de comunicación con el equipo de soporte para dar solución a requerimientos mínimos (que no estén fuera del alcance) de los usuarios finales del requerimiento.

3.1.3. Optimización

La fase de Operar y Optimizar es crucial, ya que se centra en la implementación y mejora continua de la solución analítica dentro de un entorno de producción. En esta fase, se llevan a cabo actividades de implementación, monitoreo y optimización de la solución desplegada, garantizando no solo que funcione correctamente, sino también que se adapte a los cambios en los datos, las necesidades del negocio y las condiciones del entorno. Esto permite maximizar la efectividad y eficiencia de la solución a largo plazo.

La fase abarca varias actividades clave, estas actividades se centran en evaluar la calidad del modelo, preparar el entorno de despliegue, implementar la solución, monitorear su rendimiento y ajustar los modelos según sea necesario. A continuación, se detallan estas actividades y consideraciones esenciales:

1. **Evaluación del Modelo:** El primer paso es evaluar el modelo analítico para entender su calidad y asegurarse de que aborda de manera adecuada y completa el problema del negocio. Esto incluye el cálculo de diversas medidas de diagnóstico para validar su precisión y efectividad.
2. **Preparación para el Despliegue:** Antes de implementar el modelo, es crucial asegurarse de que todos los componentes de la solución estén listos para la producción, lo que incluye software, hardware y datos. Además, es necesario documentar y configurar el entorno de producción, estableciendo procedimientos claros para la migración y despliegue de los modelos.

3. **Despliegue de la Solución:** Una vez que el modelo ha demostrado resultados satisfactorios, se despliega en el entorno de producción o en un entorno de prueba comparable. Durante esta fase, se implementan los modelos analíticos y se integran con los sistemas existentes de la organización, asegurando la compatibilidad y el flujo adecuado de datos.

4. **Monitoreo de Rendimiento:** Tras el despliegue, es fundamental establecer indicadores clave de rendimiento (KPIs) para medir la efectividad de la solución. Se implementan herramientas y procesos de monitoreo para rastrear el rendimiento de los modelos en tiempo real, lo que permite detectar posibles problemas o caídas en la eficacia.

5. **Gestión de Alertas y Problemas:** Es esencial configurar sistemas de alerta que puedan detectar anomalías o caídas en el rendimiento del modelo. Además, se deben establecer procedimientos para responder rápidamente a los problemas identificados, minimizando el impacto en las operaciones del negocio.

6. **Evaluación y Ajuste de Modelos:** De manera periódica, se realizan evaluaciones del rendimiento del modelo utilizando datos actualizados. Según los resultados de estas evaluaciones, los modelos se ajustan para asegurar que continúen proporcionando valor y abordando eficazmente los problemas del negocio.

7. **Documentación y Retroalimentación:** La documentación de todos los procesos de monitoreo y optimización es esencial para referencia futura y mejora continua. Además, se debe recoger retroalimentación de los usuarios finales y partes interesadas, lo que ayuda a identificar áreas de mejora y refinar la solución analítica.

En conjunto, estas actividades y consideraciones permiten a la organización no solo implementar soluciones analíticas efectivas, sino también adaptarse a cambios en los datos y necesidades del negocio, optimizando así la eficacia y eficiencia a largo plazo.

Estas características, combinadas con el enfoque de ASUM-DM, proporcionan un marco de trabajo robusto para el desarrollo y la ejecución de proyectos de ciencia de datos.

3.2. Modelo de Diseño de Datos

El *modelo de diseño de datos* o mejor conocido como arquitectura de datos, se refiere a la estructura y organización de los datos dentro de una compañía, estableciendo los principios, prácticas, tecnologías y herramientas que guían el almacenamiento, la integración, la gestión y la distribución de los mismos. Su objetivo principal es optimizar la disponibilidad, calidad y seguridad de los datos, permitiendo que los usuarios y sistemas puedan acceder a ellos de manera eficiente y alineada con los objetivos empresariales.

Una arquitectura de datos bien definida facilita la toma de decisiones informadas, mejora la interoperabilidad entre sistemas, y asegura que los datos sean tratados con los más altos estándares

de calidad y seguridad. Además, permite la integración de fuentes de datos dispares, la creación de procesos automatizados y la generación de informes y análisis complejos.

Ejemplos de Modelos de Diseño de Datos

En el ámbito de la ingeniería de datos, existen múltiples modelos de diseño que se utilizan para gestionar, procesar y almacenar datos de manera eficiente. A continuación, se explican algunos de los modelos más comunes, destacando sus características, ventajas y casos de uso:

- **Arquitectura de Datos Centralizada**
  Este modelo organiza todos los datos en una única base de datos centralizada que actúa como la "única fuente de verdad" para toda la organización. Es ideal para empresas con una estructura jerárquica simple, donde los procesos y flujos de datos están altamente estandarizados. Al centralizar los datos:

  - Se facilita el control de acceso y la seguridad, ya que toda la información se encuentra en un solo lugar.
  - Permite una toma de decisiones basada en datos consistentes y actualizados.

  Sin embargo, puede presentar limitaciones en términos de escalabilidad y resiliencia, especialmente para organizaciones grandes o distribuidas geográficamente.

- **Arquitectura de Datos Distribuida**
  En este modelo, los datos se distribuyen entre múltiples sistemas o bases de datos ubicados en diferentes regiones o incluso en diferentes infraestructuras. Este enfoque es útil para organizaciones grandes o diversificadas, como aquellas con varias líneas de negocio o presencia global. Sus beneficios incluyen:

  - Mayor disponibilidad y redundancia, ya que los datos están replicados en diferentes ubicaciones.
  - Reducción de la latencia en el acceso a datos al acercarlos a los usuarios finales.

  Por otro lado, puede implicar desafíos adicionales en la sincronización, consistencia y manejo de conflictos entre diferentes bases de datos.

- **Arquitectura de Datos en la Nube**
  Con la adopción masiva de servicios en la nube, muchas organizaciones han migrado hacia arquitecturas que aprovechan la flexibilidad y escalabilidad de plataformas como AWS, Azure o Google Cloud. Este modelo ofrece:

  - Escalabilidad dinámica, permitiendo a las organizaciones manejar grandes volúmenes de datos sin necesidad de inversión inicial en infraestructura.
  - Accesibilidad global, permitiendo a los usuarios acceder a los datos desde cualquier lugar con conexión a Internet.

- Reducción de costos operativos mediante modelos de pago por uso.

Sin embargo, puede implicar dependencias de proveedores (*vendor lock-in*) y preocupaciones de seguridad y cumplimiento normativo en sectores regulados.

- **Arquitectura Lambda**
  Este modelo combina dos enfoques para el procesamiento de datos: en tiempo real (o flujo) y por lotes. Es especialmente popular en sistemas que requieren manejar grandes volúmenes de datos en aplicaciones como IoT, análisis de comportamiento en línea o sistemas de recomendación. Sus características incluyen:

  - Procesamiento en tiempo real, permitiendo reaccionar de manera inmediata a eventos o datos entrantes.
  - Procesamiento por lotes, para realizar análisis más complejos o generar reportes basados en datos históricos.

  A pesar de sus beneficios, puede ser un modelo complejo de implementar y mantener debido a la necesidad de gestionar dos rutas de procesamiento distintas.

- **Arquitectura de Datos Híbrida**
  Este modelo combina soluciones de almacenamiento locales (on-premises) con almacenamiento en la nube, permitiendo a las organizaciones aprovechar lo mejor de ambos mundos. Es ideal para empresas que necesitan cumplir con requisitos específicos de regulación o privacidad de datos, pero que también buscan flexibilidad y escalabilidad. Sus beneficios incluyen:

  - Flexibilidad para manejar diferentes cargas de trabajo en entornos locales y en la nube.
  - Optimización de costos, utilizando almacenamiento local para datos críticos o sensibles y la nube para datos menos utilizados o necesidades de escalamiento.

  Sin embargo, requiere una integración cuidadosa para garantizar la sincronización y consistencia entre ambos entornos.

El modelo de diseño de datos seleccionado para este proyecto se puede considerar híbrido debido a la forma en que se organizan y gestionan los datos a lo largo de su ciclo de vida. Inicialmente, los datos tienen como fuente principal la base de datos de GIRO, pero, dependiendo de la etapa del ciclo de vida, pueden transferirse entre diferentes buckets[1].

Este enfoque permite garantizar tanto la operatividad del aplicativo GIRO para los distintos usuarios como la manipulación controlada y autorizada de los datos con fines científicos. Además, facilita la identificación y adopción de los modelos de datos más adecuados para el desarrollo del proyecto.

[1] En AWS (Amazon Web Services), un bucket es un contenedor de nivel superior en el servicio Amazon Simple Storage Service (S3), que se utiliza para almacenar objetos como archivos y datos. Cada bucket tiene un nombre único globalmente y actúa como una unidad de almacenamiento dentro de S3.

3.2.1. Medallion Architecture

La *arquitectura Medallion de Delta Lake* es un modelo de diseño de datos utilizado en plataformas de procesamiento de datos distribuidos, como Apache Spark con Delta Lake, que organiza los datos en tres capas distintas para facilitar la gestión de grandes volúmenes de información a lo largo de su ciclo de vida. Estas capas son comúnmente denominadas **Bronze**, **Silver** y **Gold**, y representan los distintos niveles de calidad, procesamiento y refinamiento de los datos.

- **Bronze Layer**: Contiene los datos crudos o ingestados, tal como se reciben de las fuentes. Esta capa garantiza la disponibilidad de datos históricos y una trazabilidad completa.
- **Silver Layer**: En esta capa se almacenan datos enriquecidos y limpios. Aquí se realizan procesos de transformación y filtrado que estandarizan los datos para su análisis.
- **Gold Layer**: Incluye los datos agregados y listos para consumo por parte de los modelos analíticos o dashboards. Esta capa está diseñada para maximizar la eficiencia y la accesibilidad de los datos.

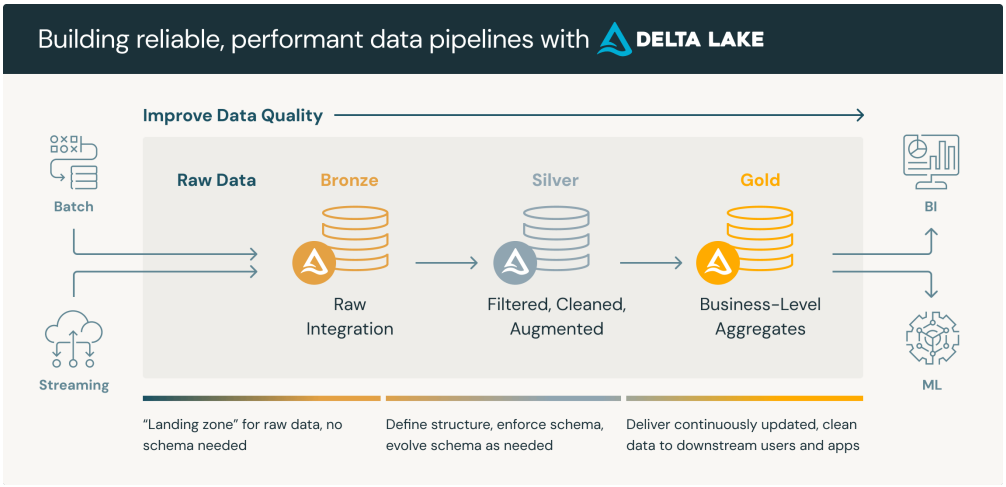


Figura 3.2: Arquitectura Medallion
[19]

La combinación de estas capas permite crear pipelines de datos robustos que soportan casos de uso complejos, como la implementación de modelos de machine learning, reportes en tiempo real y análisis predictivos.

Además de facilitar la organización y el refinamiento de los datos, la arquitectura Medallion promueve la modularidad y la escalabilidad en el diseño de soluciones de datos. Al separar los datos en capas bien definidas, las organizaciones pueden implementar estrategias de control de calidad, auditoría y monitoreo en cada etapa del pipeline. Esto resulta especialmente útil en entornos

colaborativos, donde múltiples equipos trabajan en paralelo sobre diferentes fases del ciclo de vida de los datos, garantizando que las operaciones aguas abajo no se vean comprometidas por cambios o errores en los datos crudos. De esta forma, Medallion Architecture no solo optimiza el rendimiento y la trazabilidad, sino que también establece una base sólida para el cumplimiento normativo y la gobernanza de datos.

Ventajas de Delta Lake y Medallion Architecture

La implementación de la arquitectura Medallion ofrece varias ventajas, entre las cuales se destacan:

- **Escalabilidad**: Facilita el procesamiento y almacenamiento eficiente de grandes volúmenes de datos a medida que fluyen a través de las diferentes capas.
- **Calidad de Datos**: Garantiza la calidad de los datos mediante el refinamiento progresivo desde la capa Bronze hasta la capa Gold.
- **Flexibilidad**: Permite integrar datos tanto estructurados como no estructurados, lo que facilita el análisis y procesamiento de datos de diversas fuentes.
- **Reproducibilidad y Auditoría**: Dado que cada capa mantiene un historial de las transformaciones aplicadas, la arquitectura permite la trazabilidad y auditoría completa de los datos, facilitando la detección de problemas y la mejora continua.

En el marco de un enfoque holístico, la adopción de la **Medallion Architecture** se presenta como una solución integral para la gestión y diseño de proyectos de datos. Como se mencionó antes [3.2.1] esta arquitectura es reconocida por su capacidad para segmentar y estructurar la información en capas claras y definidas, se ajusta perfectamente a las necesidades de evolución e iteración inherentes al desarrollo de proyectos complejos.

El uso para este proyecto de esta aproximación no solo se limita a la organización de datos en fases, tales como las capas de bronze, silver y gold [3.2.1], que permiten un flujo lógico y escalable desde la ingesta hasta la obtención de datos listos para análisis. También se extrapola al diseño metodológico del proyecto, alineándose con los principios de **ASUM-DM** (Analytical Solutions Unified Method for Data Mining) [3.1]. Esto implica que la Medallion Architecture sirve como una base no solo para estructurar datos, sino también para orquestar la colaboración efectiva entre equipos y tareas a lo largo de las diferentes fases de trabajo.

En este sentido, la implementación de la Medallion Architecture no solo se limita a ser una herramienta técnica, sino que se convierte en una filosofía de trabajo que garantiza la paralelización y la asincronía en la ejecución de actividades. Al dividir claramente las responsabilidades y fases de trabajo, el equipo puede abordar diferentes partes del proyecto simultáneamente, sin crear dependencias que puedan ralentizar el avance general. Esto es especialmente relevante en proyectos con múltiples participantes, donde la coordinación es esencial para alcanzar los objetivos dentro de los

plazos establecidos.

Además, este enfoque facilita la trazabilidad y la transparencia del proyecto, ya que cada capa de datos y fase metodológica se encuentra documentada y definida de forma clara, lo que permite una evolución continua del diseño de la solución. De esta manera, se asegura que cada iteración del proyecto no solo refine los resultados obtenidos, sino que también aporte mejoras significativas al proceso general.

En conclusión, al extrapolar la **Medallion Architecture** a todos los niveles del proyecto combinado con **ASUM-DM**, se establece un marco robusto que no solo gestiona la complejidad de los datos, sino que también fomenta un entorno de trabajo ágil y colaborativo. Este diseño garantiza la capacidad de escalar y adaptarse a las necesidades cambiantes del proyecto, al tiempo que promueve una ejecución ordenada y eficiente de cada etapa. Así, se proporciona una base sólida para el desarrollo de soluciones, no solo desde la perspectiva técnica, sino también desde el diseño organizativo y metodológico, asegurando el éxito integral del proyecto.

Capítulo 4

Diseño

En el diseño de la solución desarrollada, se ha abordado una estructura integral que combina diversas estrategias y herramientas para garantizar un desarrollo alineado con los objetivos propuestos.

En primer lugar, se definió una **arquitectura de la solución** que integra los aspectos de desarrollo, almacenamiento y despliegue. Este diseño holístico asegura la interoperabilidad entre los distintos componentes, desde la fuente de datos hasta la presentación de los resultados finales, alineándose con los objetivos de negocio y las mejores prácticas de ingeniería de datos.

Por otro lado, la **estructura del proyecto** en *GitHub* fue diseñada para facilitar el control de versiones y la colaboración. Se implementó una jerarquía de ramas que permite gestionar de manera ordenada las diferentes fases y entregables según la asignación de cada persona involucrada en el proyecto, asegurando la trazabilidad y la integración continua.

Adicionalmente, para la **arquitectura de datos**, se adoptó la *Medallion Architecture*, esto con el fin de realizar el procesamiento y gestión de datos en la nube utilizando los servicios de AWS. Esta arquitectura, basada en los principios de Delta Lake, estructura los datos en capas (bronze, silver y gold), lo que asegura escalabilidad, confiabilidad y calidad en los procesos analíticos.

Finalmente, se analizó la *estructura de datos en GIRO*, la aplicación origen del proyecto. Esto permitió identificar los requerimientos y optimizar la extracción, transformación y carga (ETL) de los datos hacia la arquitectura definida.

4.1. Arquitectura de Solución

En esta sección se describe la arquitectura general del proyecto desarrollado. La arquitectura se centra en la integración de diversas herramientas y servicios, principalmente utilizando GitHub para el control de versiones y AWS para el almacenamiento y procesamiento de datos. A continuación, se detalla cada componente de la arquitectura representada en la Figura 4.1.

GIRO (Aplicación de Optima Corporation) actúa como el origen de datos principal. Los datos se extraen de un sistema de gestión de bases de datos (DBMS) utilizando una herramienta de cliente Oracle para consultar las tablas fuentes en batch y extraer para cargar en AWS. Esta herramienta se encarga de extraer los datos necesarios y enviarlos a la siguiente etapa del proceso.

El proceso de ELT (Extract, Load, Transform) se realiza de manera batch, es decir, en lotes periódicos. Este proceso se encarga de extraer los datos desde Oracle, cargarlos en los buckets de

AWS y transformarlos según las necesidades del proyecto.

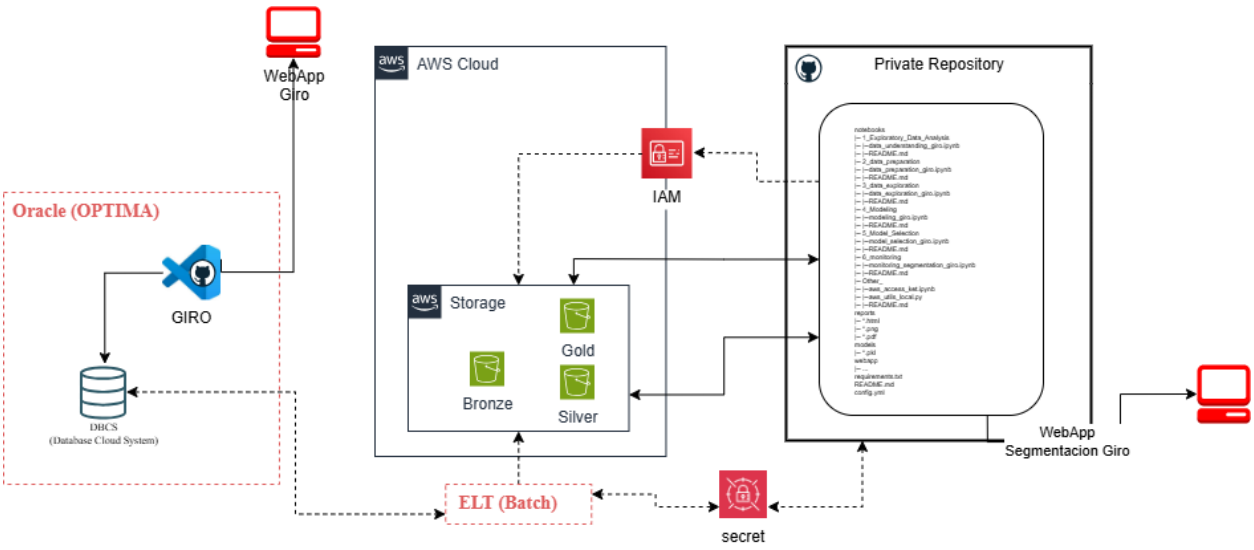


Figura 4.1: Arquitectura de Solución

Los datos extraídos son transferidos a la nube de AWS, donde se almacenan en diferentes buckets de S3 (Simple Storage Service) 4.3.1. Para gestionar el acceso a los recursos en AWS se utiliza IAM (Identity and Access Management). Este servicio permite definir políticas de acceso y roles, asegurando que solo los usuarios autorizados puedan acceder a los datos y servicios en la nube.

El repositorio privado en GitHub contiene todo el código fuente del proyecto, incluyendo scripts de ETL, análisis de datos y configuraciones de AWS. Este repositorio es accesible únicamente por los miembros del equipo de desarrollo, asegurando la integridad y confidencialidad del código.

Para gestionar las credenciales y secretos necesarios para acceder a los diferentes servicios y recursos, se utiliza AWS Secrets Manager. Este servicio asegura que las credenciales sean almacenadas y gestionadas de manera segura.

4.2. Estructura de proyecto

4.2.1. GitHub

En el desarrollo del presente proyecto, en donde se emplea la metodología ASUM-DM adaptada para cubrir las necesidades del mismo omo se menciono en subsecciones previas 3.1, ASUM-DM es una metodología diseñada para procesos de ciencia de datos y toma decisiones basada en la gestión ágil de proyectos. En el contexto de este trabajo, se seleccionaron y adaptaron algunas fases clave de la metodología, mientras que otras se omiten debido a la naturaleza específica del proyecto.

El uso de GitHub [1] para el manejo de versiones y colaboración fue crucial en la implementación de esta metodología. El repositorio de GitHub fue configurado para reflejar las fases del proyecto, permitiendo un desarrollo ágil y estructurado. De esta manera, se estableció una jerarquía en el repositorio que facilitó la organización y gestión de las distintas etapas del proyecto, como se muestra en la Figura 4.2.

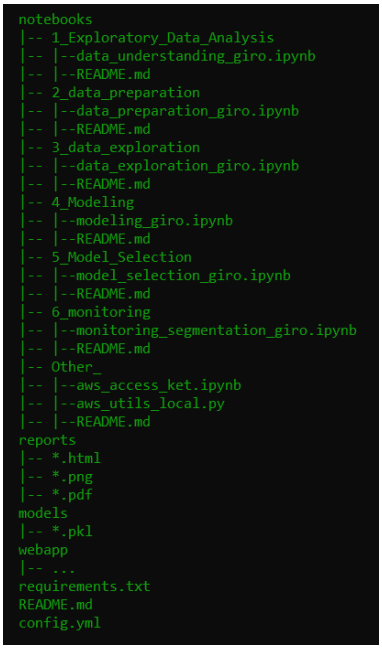


Figura 4.2: Estructura de proyecto GitHub

4.2.2. Ventajas de Usar GitHub como Repositorio para el Proyecto

GitHub, como plataforma de desarrollo colaborativo, presenta varias ventajas que contribuyen al éxito del proyecto:

- **Control de versiones**: GitHub utiliza el sistema de control de versiones Git, lo que permite realizar un seguimiento detallado de los cambios en el código y facilita la recuperación de versiones anteriores si es necesario.

- **Colaboración en tiempo real**: Nuestro equipo de trabajo puede colaborar de manera eficiente, realizando cambios en paralelo sin interferir unos con otros.

[1] GitHub es una plataforma de desarrollo colaborativo basada en la web que utiliza el sistema de control de versiones Git. Permite a los usuarios almacenar y gestionar sus proyectos de software, realizar control de versiones, y colaborar con otros usuarios mediante repositorios. GitHub proporciona características como el seguimiento de problemas, la integración continua, y la gestión de solicitudes de extracción (pull requests), facilitando el desarrollo colaborativo de software.

- **Gestión de ramas**: La estructura de ramas en GitHub permite gestionar diferentes versiones o tareas del proyecto asignadas a cada uno de los integrantes, facilitando la segregación de tareas y el control de los cambios aplicados a cada fase.

- **Integración continua**: GitHub facilita la integración continua mediante herramientas como GitHub Actions, lo que permite ejecutar pruebas y procesos de validación de manera automatizada.

- **Seguimiento de problemas y solicitudes de extracción (Pull Requests)**: GitHub permite gestionar las nuevas funcionalidades mediante un sistema de seguimiento eficiente. Las solicitudes de extracción (pull requests) aseguran que los cambios realizados por diferentes miembros del equipo sean revisados antes de ser integrados al repositorio principal 4.3.

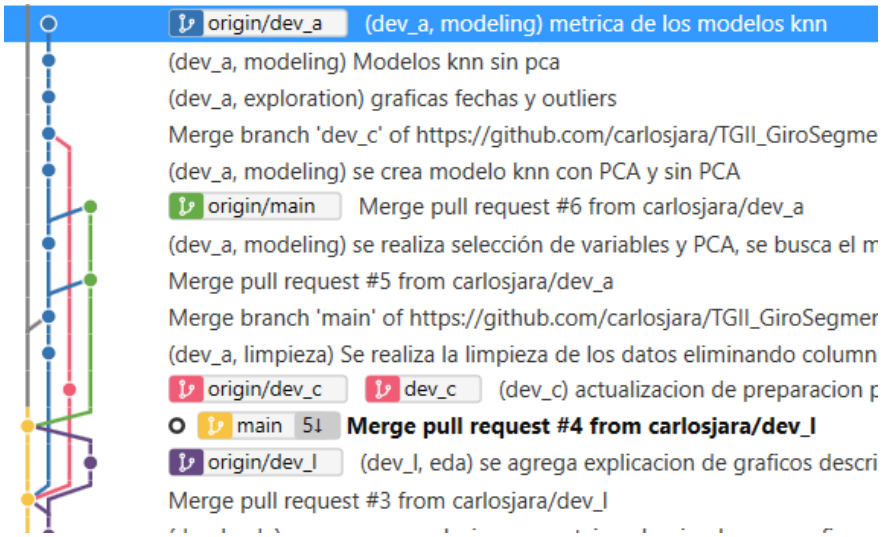


Figura 4.3: commits y pull-request en repositorio de GitHub.

4.3. Arquitectura de datos

4.3.1. Integración con AWS

La integración de la arquitectura con AWS permitió aprovechar servicios como Amazon S3 para el almacenamiento de datos, la orquestación de los procesos de extracción, transformación y carga (ETL) fue realizada con el proyecto desarrollado en python dispuesto en el repositorio en github [4.2.1]. Estos servicios, aseguraron que los datos fueran gestionados de manera eficiente, con procesos automatizados que redujeron los tiempos de desarrollo y minimizaron los errores.

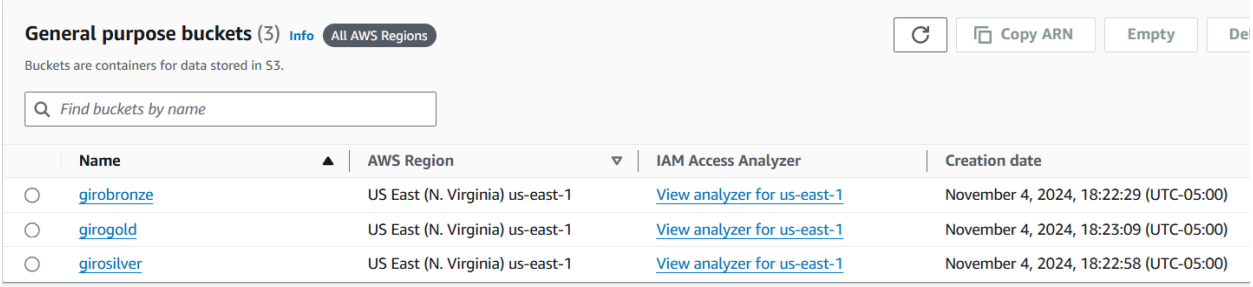


Figura 4.4: Medallion Architecture implementada en AWS.

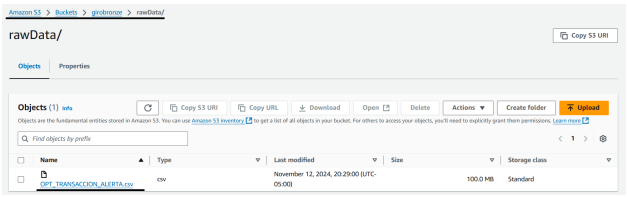
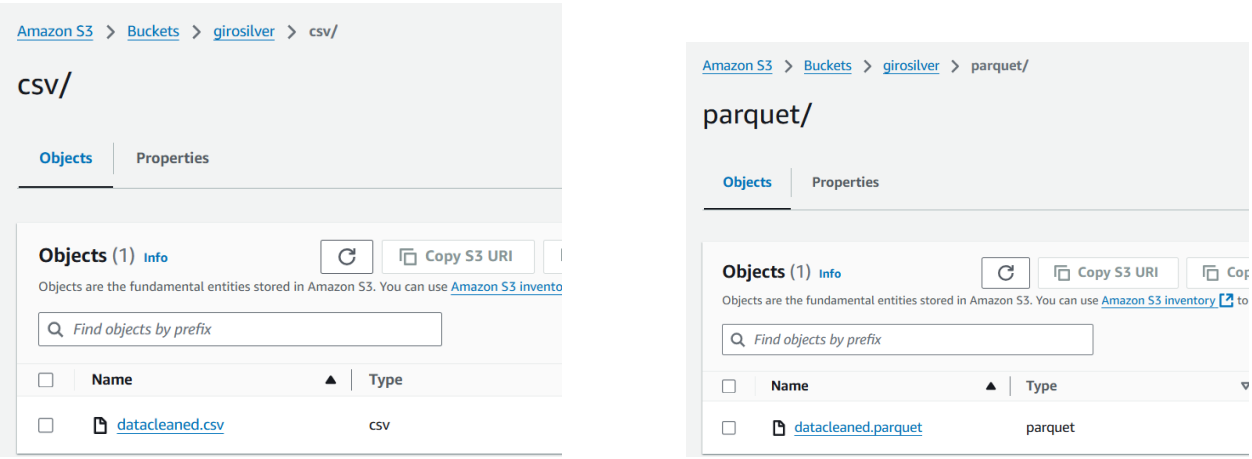


Figura 4.5: Bronze



(a) Silver : CSV

(b) Silver : Parquet

Figura 4.6

El manejo de los archivos según su estado de procesamiento, como se observa en las figuras 4.5, 4.6a y 4.6b, permite un aislamiento efectivo de los datos procesados y asegura la independencia de las diferentes fases o roles involucrados en la manipulación y explotación de la información, ya sea en estado bruto o procesado. Este enfoque modular garantiza que el flujo de datos sea consistente, eficiente y fácilmente reproducible, lo que resulta en análisis finales de alta calidad y mayor confiabilidad en los resultados.

4.4. Estructura de datos Fuente

4.4.1. Modelo Entidad-Relación de GIRO

La base de datos fuente del proyecto se basa en un modelo entidad-relación (MER) desarrollado en Oracle, el cual sirvió como punto de partida para definir los procesos de ETL. La Figura 4.7 presenta el esquema del modelo.

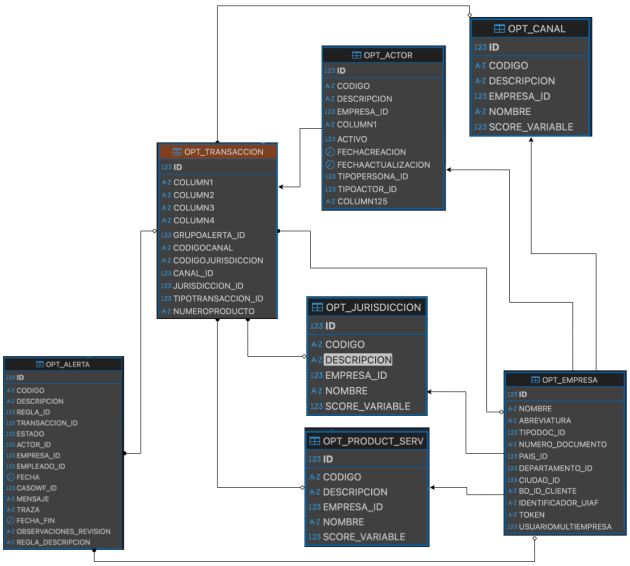


Figura 4.7: Modelo Entidad-Relación de GIRO.

Este análisis permitió identificar los requerimientos específicos para optimizar la extracción y carga hacia la arquitectura definida.

Capítulo 5

Ciencia de Datos

5.1. Dataset

5.1.1. Dataset inicial

Para el presente trabajo se utilizaron dos conjuntos de datos extraídos de las bases de datos del aplicativo GIRO por medio de consultas SQL (batch-load[1]).

1. El primer conjunto de datos corresponde a las transacciones y alertas del año 2023 con un total de 151.366 registros y 67 atributos.

2. El segundo conjunto de datos corresponde a las transacciones y alertas del año 2024 con un total de 140.436 registros y 67 atributos.

5.2. Análisis exploratorio de datos

A continuación, se explora la base de datos para obtener un primer vistazo de los datos y realizar un análisis exploratorio inicial.

Después de unificar las dos bases de datos y eliminar aquellas columnas y filas que presentaban valores faltantes en su totalidad, se cuenta para el análisis, exploración y modelado con una base de 291.802 registros y 38 atributos. Cabe mencionar que los datos corresponden a fechas entre Enero de 2023 y Noviembre de 2024.

- **Variables categóricas**

En la figura 5.1, se destaca que el 91 % de los registros en la base de datos corresponden a movimientos realizados por personas jurídicas, mientras que solo el 0.6 % fueron realizados por personas naturales.

El 99 % de los registros de movimientos se realizaron en oficinas, y únicamente el 0.71 % a través de Internet.

[1]Batch-Load: Técnica de extracción de datos en la cual se realiza una consulta programada que no utiliza replicación en tiempo real, sino que toma un *snapshot* o captura de los datos disponibles en el momento exacto de la ejecución de la consulta. Esto garantiza la consistencia de los datos procesados en un estado puntual del tiempo.

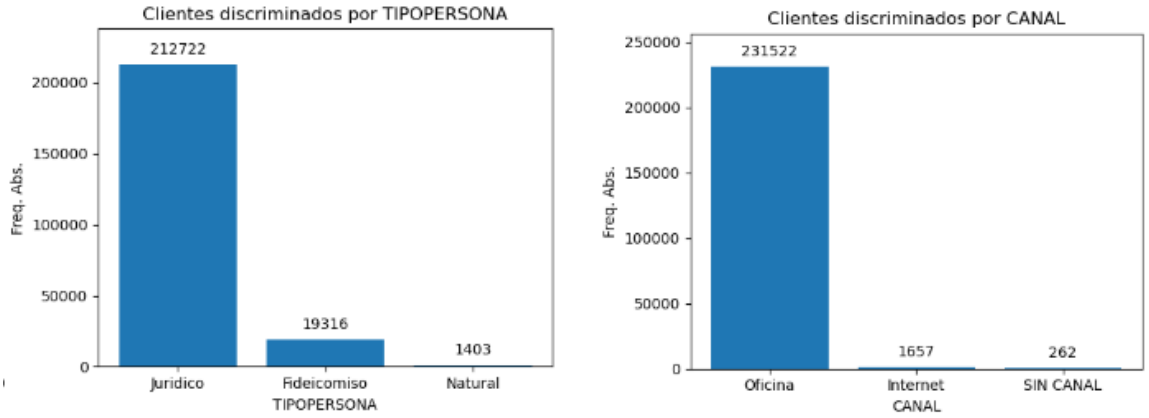


Figura 5.1: Gráficos de barras variables categóricas

En la figura 5.2, se observa que el tipo de transacción o movimiento predominante fueron los egresos, representando el 95 %.

Del total de movimientos o transacciones en la base de datos, solo el 0.81 % presentaron una alerta.

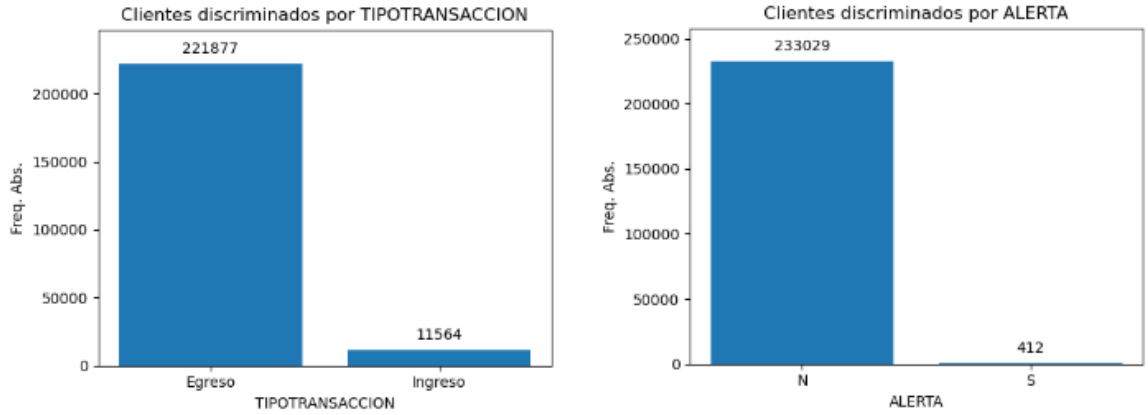
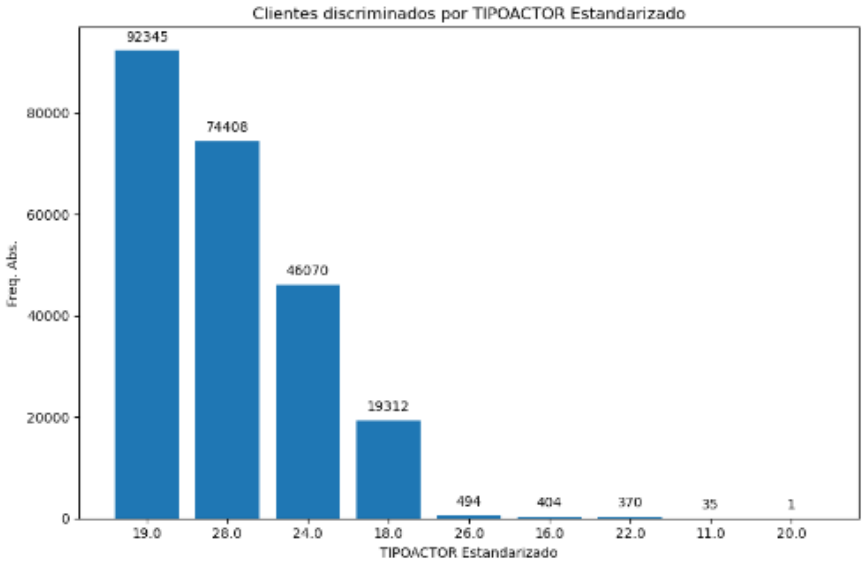


Figura 5.2: Gráficos de barras variables categóricas

Debido a que los valores dentro de algunas columnas eran muy largos se presentan a continuación las gráficas con su valor estandarizado para una mejor visualización, ademas de una tabla con la correspondiente estandarización.

Como se puede observar en la figura 5.3 para la variable **TIPOACTOR** aproximadamente el 40 % de los registros corresponden a clientes Fideicomitentes que son aquellos que transfieren o destinan ciertos bienes o derechos a un fideicomiso. Aproximadamente el 32 % de los registros corresponden a un Tipo Tercero General, este tipo de actor se refiere a una categoría amplia que

incluye a personas o entidades externas con las que una organización mantiene relaciones comerciales o transaccionales. Aproximadamente el 20 % de los registros corresponden a Proveedores y el restante porcentaje se reparte en las categorías Fideicomiso, Sociedades Fiduciarias, Empleado, Mixto, Beneficiario/No autorizado y Gobierno. En nuestro set de datos solo se presentó un registro para este último.



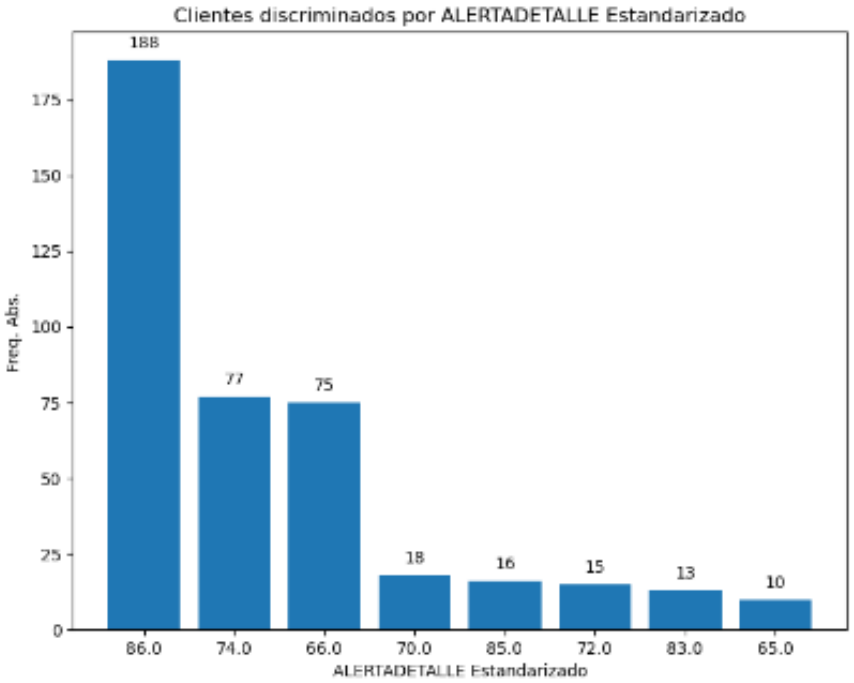
Descripción	Código
Tipo Tercero General	28.0
Fideicomitente	19.0
Fideicomiso	18.0
Proveedor	24.0
Empleado	16.0
Mixto	22.0
Sociedades Fiduciarias	26.0
Beneficiario / No Autorizado	11.0
Gobierno	20.0

Figura 5.3: Gráfico y tabla con los valores estandarizados para la variable tipo de actor

Para la variable **ALERTADETALLE** (figura 5.4), se observa que es la que presenta menos registros en la base de datos, lo cual podría deberse a que no todos los clientes generaron una alerta al momento de revisar sus transacciones. Aproximadamente el 46 % de los registros corresponden a alertas del tipo 'Concentración de pagos a un usuario o terceros', seguido por un 19 % relacionado con 'Transacciones en jurisdicciones de alto riesgo' y un 18 % con alertas de 'Monitoreo de clientes con actividades económicas meritorias o entidades sin ánimo de lucro.

Por otro lado, para la variable **PRODUCTOSERVICIO** (figura 5.5), que indica el tipo de

cuenta en la que se registró el movimiento, aproximadamente el 53 % de los registros corresponden a 'Fiducia de Administración', el 35 % a 'Recursos del Sistema General de Seguridad Social - Pasivos Pensionales', y el 12 % a 'Carteras colectivas'

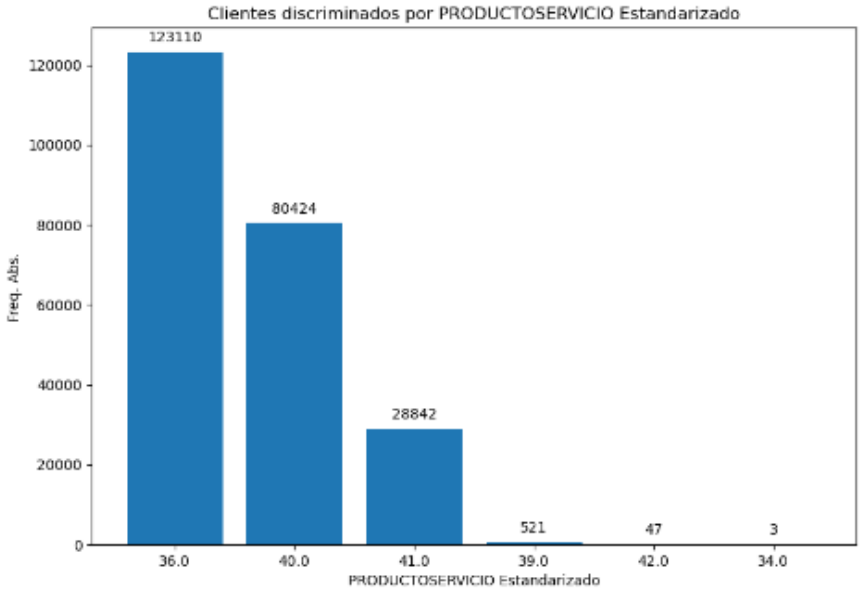


Descripción	Código
Monitoreo Entidades sin Ánimo de Lucro	66.0
Concentración de pagos a un usuario o tercero	86.0
Transacciones Jurisdicciones de Alto Riesgo (I)	74.0
Transacciones (E) Fideicomisos	85.0
Transacciones realizadas por PEP (I)	72.0
Transacciones (I) Fideicomisos	70.0
Transacciones realizadas por PEP (E)	83.0
Monitoreo Actividades - Sector Salud	65.0

Figura 5.4: Gráfico y tabla con los valores estandarizados para la variable alerta detalle

La variable JURISDICCION (tabla 5.1) hace referencia al país o región donde una entidad, transacción o cliente opera o está registrado. Esta variable es clave, ya que las diferentes regiones tienen regulaciones, normativas fiscales y políticas de cumplimiento específicas, lo que influye directamente en el riesgo asociado a las transacciones y actividades financieras realizadas en dichos lugares.

Aproximadamente el 78 % de los registros corresponden a movimientos en la ciudad de Bogotá, mientras que un 8 % no tienen jurisdicción asignada. El resto de los registros se distribuyen entre otros departamentos y municipios.



Descripción	Código
Fiducia de Administración	36.0
Recursos Del Sistema General De Seguridad Social - Pasivos Pensionales	40.0
Carteras Colectivas De General	41.0
Fiducia En Garantia - Fuente De Pagos	39.0
Fondos De Capital Privado	42.0
Fiducia Inmobiliaria - Administración Y Pagos	34.0

Figura 5.5: Gráfico y tabla con los valores estandarizados para la variable producto servicio

JURISDICCION	Freq. Abs.	Freq. Rel.
Bogotá D.C	183221	78.49 %
SIN JURISDICCION	17793	7.62 %
Santander	13580	5.82 %
Boyacá	7229	3.10 %
Antioquia	3486	1.49 %
Risaralda	2130	0.91 %
Cauca	1797	0.77 %
Valle Del Cauca	1424	0.61 %
Cesar	796	0.34 %
Chocó	657	0.28 %
Otros	1328	0.57 %

Tabla 5.1: Variable JURISDICCION

- **Variables numéricas**

Algunas variables presentaban valores extremadamente altos que dificultaban la interpretación de su distribución. Por ello, se han excluido estos valores atípicos para proporcionar una visualización más clara y detallada del comportamiento de las variables numéricas.

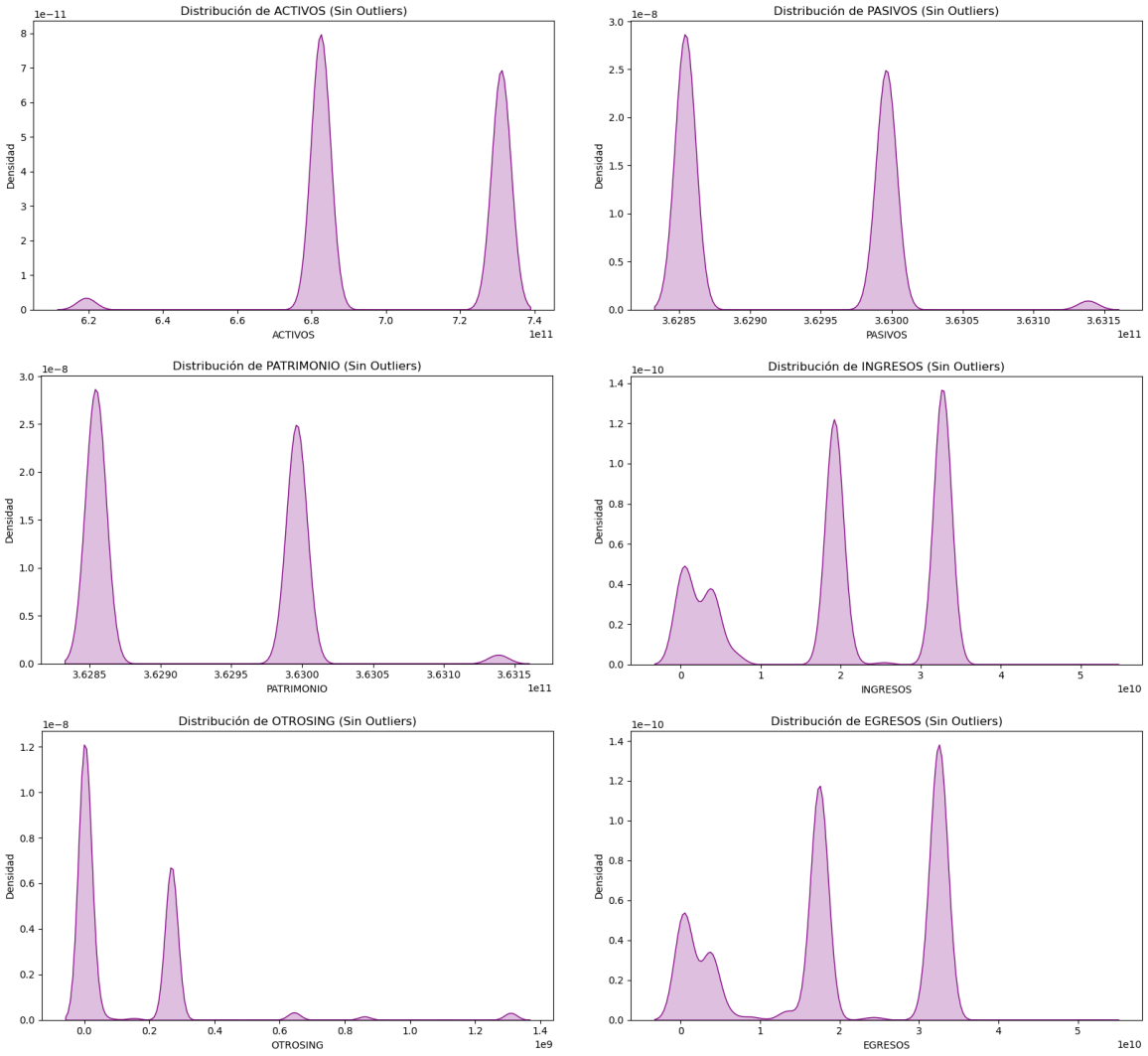


Figura 5.6: Gráficos de densidad para las variables numéricas

En las gráficas de la figura 5.6, se observa que la mayoría de las variables numéricas tienden a presentar picos, lo que indica una concentración significativa alrededor de dos o tres valores. Estas distribuciones multimodales podrían sugerir la existencia de diferentes grupos o patrones de comportamiento dentro de los datos.

En las variables **ACTIVOS**, **PASIVOS** y **PATRIMONIO**, parece haber una mayor concentración en ciertos rangos específicos, con un pico principal y una baja dispersión alrededor de este. Esto podría indicar que la mayoría de los registros tienen valores similares en estas categorías.

Por otro lado, las variables **INGRESOS** y **EGRESOS** presentan distribuciones multimodales, lo que sugiere la posible existencia de subgrupos en los datos. Estas diferencias podrían estar relacionadas con atributos como el tipo de cliente, la ubicación geográfica u otros factores relevantes.

En el caso de la variable **OTROSING**, aunque muestra una fuerte concentración inicial, también se observa cierta dispersión en su cola derecha. Esto sugiere la presencia de registros con valores más altos, aunque estos sean menos frecuentes.

- **Matriz de correlación**

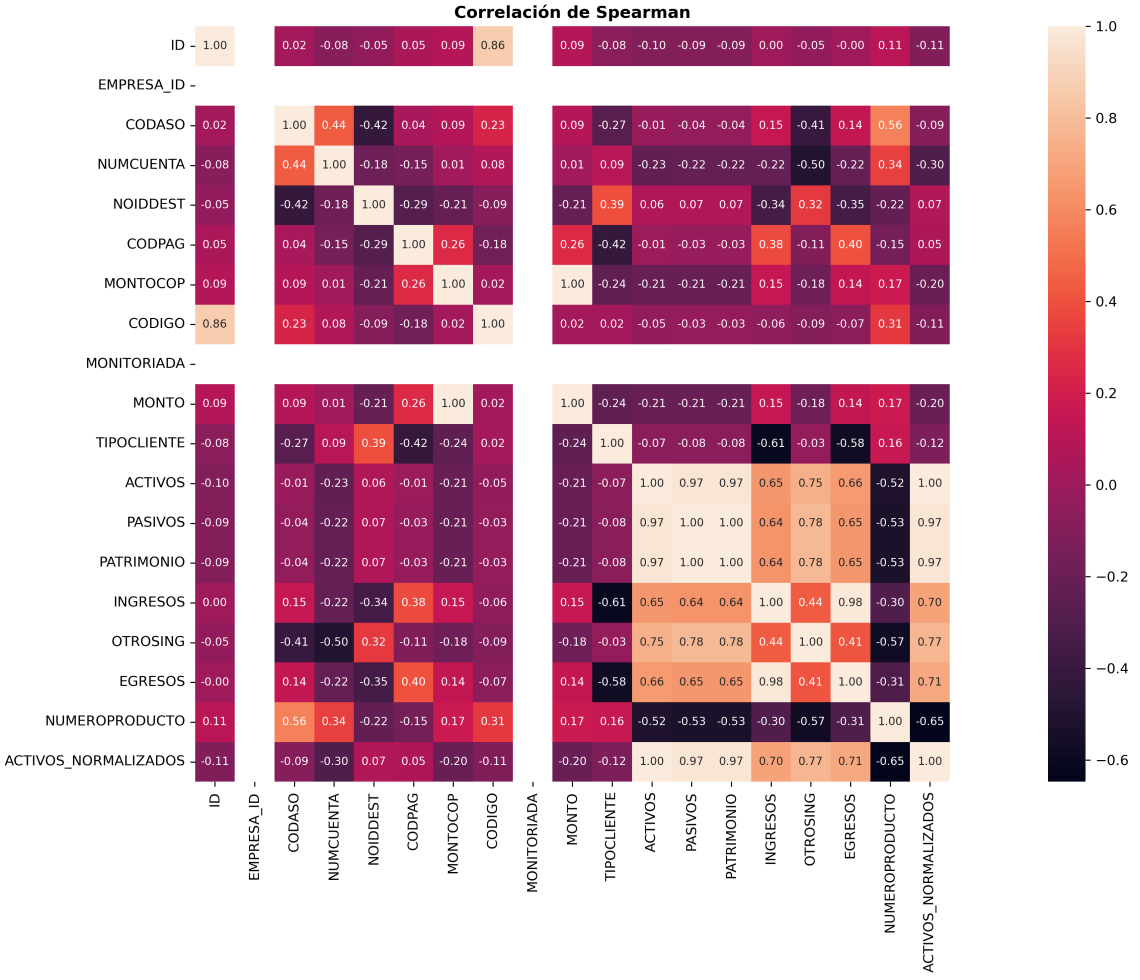


Figura 5.7: Matriz de correlación

Se elaboró la matriz de correlación para las variables numéricas (figura 5.7), en la cual se identificaron algunas correlaciones fuertes. Estas se observan principalmente entre los ingresos y los egresos, así como entre otros tipos de ingresos y el patrimonio. También destacan las correlaciones entre el patrimonio, los activos y los pasivos, así como entre los ingresos, los activos y los pasivos. Por otro lado, se identificó una relación negativa entre el tipo de cliente y sus ingresos, además de entre otros tipos de ingresos y el número del producto.

Estos hallazgos son relevantes, ya que pueden ser útiles tanto para la selección de variables que aporten mayor información al momento de aplicar nuestros modelos, como para identificar clientes con características similares y patrones comunes.

- **Valores Atípicos**

En las figuras 5.8 y 5.9 se observan los diagramas de cajas de las variables numéricas, en ellos se puede identificar variabilidad en los datos, por ejemplo, las variables como **NUMCUENTA**, **CODPAG** y **TIPOCLIENTE** muestran concentraciones claras alrededor de ciertos rangos, lo que indica que la mayoría de los datos se encuentran dentro de estos intervalos específicos, con pocos valores extremos. Esto puede ser útil para segmentar registros o priorizar análisis en los rangos más comunes.

Algunas variables como **CODIGO** y **CODPAG** parecen tener distribuciones uniformes con pocos outliers o valores atípicos. Esto podría implicar que estas variables son más consistentes en los datos y, por lo tanto, menos propensas a influencias externas o errores de registro.

Las variables como **SUMA_BENEFICIARIO**, **MONTOCOP** y **MONTO** presentan valores atípicos más marcados y en mayor cantidad. Estos valores pueden ser indicadores de comportamientos anómalos, errores de registro o casos especiales que merecen un análisis más detallado.

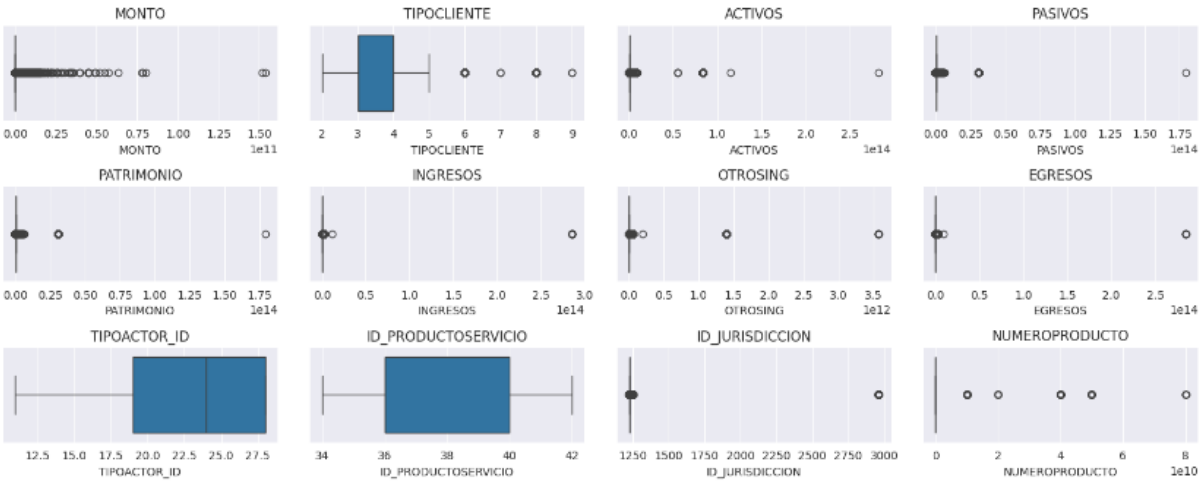


Figura 5.8: Diagramas de cajas

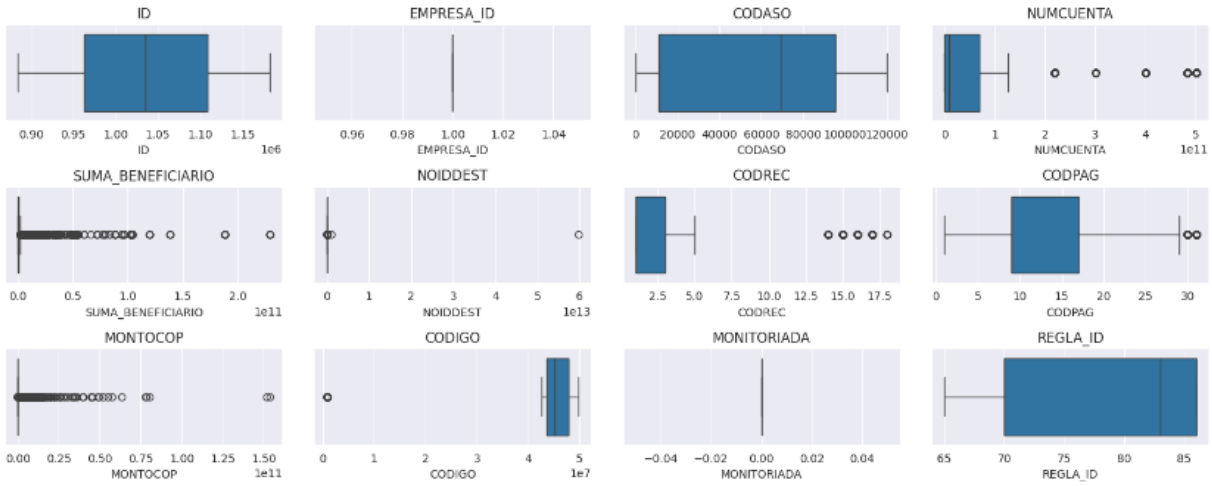


Figura 5.9: Diagramas de cajas

La variable **EMPRESA_ID** tienen una variación mínima, indicando que casi todos los valores están muy cercanos entre sí. Para otras variables no es posible ver a detalle su distribución, pero se puede apreciar que tienen presencia de valores atípicos.

Además del análisis exploratorio presentado anteriormente, en la figura 5.10 se muestra la cantidad de registros en la base de datos desglosada por mes y año. Se observa que los meses de noviembre y diciembre presentan el menor volumen de movimientos, mientras que los meses de abril, mayo y, en particular, marzo concentran la mayor cantidad de registros.

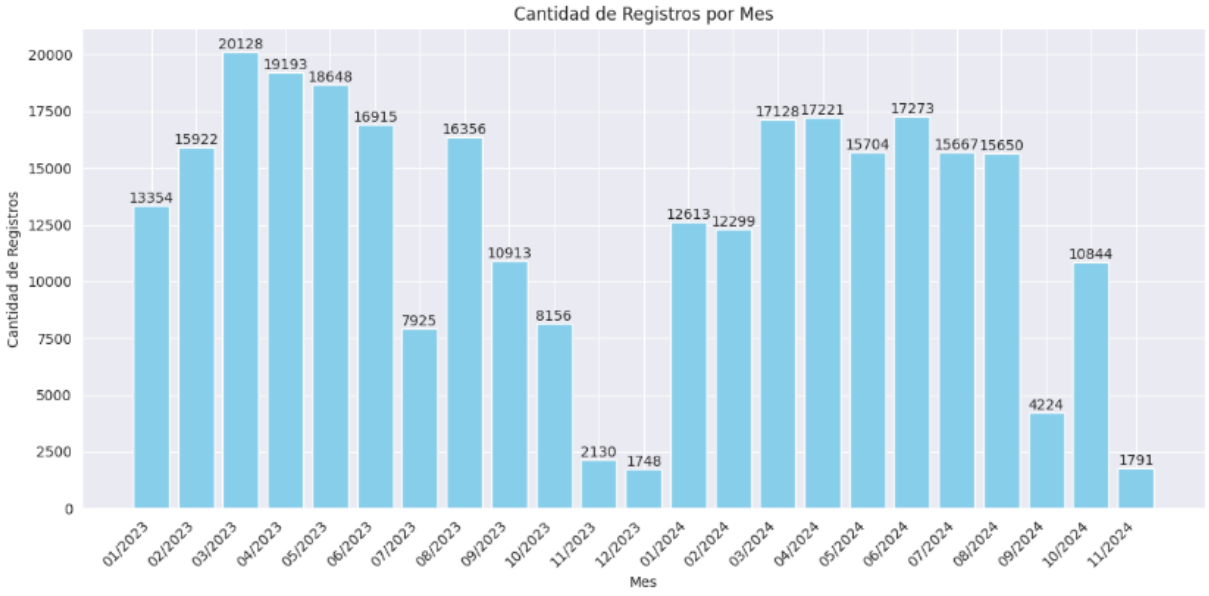


Figura 5.10: Cantidad de registros por fecha

Este comportamiento podría estar relacionado con patrones estacionales o factores específicos que afectan la frecuencia de los movimientos durante estos periodos.

- **Valores Nulos o faltantes**

A continuación, se presentan las correspondientes gráficas que ilustran los valores faltantes tanto en las variables numéricas como en las categóricas, encontrando lo siguiente:

En el caso de las variables numéricas, en la figura 5.11 se observa que **REGLA_ID**, **SUMA_BENEFICIARIO** y **CODREC** son las que presentan la mayor cantidad de valores faltantes, con aproximadamente el 98 % de sus datos ausentes. En contraste, las demás variables con valores faltantes presentan un porcentaje significativamente menor, alrededor del 7 %. Por lo tanto, más adelante se tomará la decisión de imputar los valores faltantes de estas variables utilizando diferentes metodologías.

En cuanto a las variables categóricas, en la figura 5.12 se observa que **ALERTADETALLE** es la que presenta la mayor cantidad de valores faltantes, con aproximadamente el 99 % de sus datos ausentes. Por otro lado, las demás variables con valores faltantes tienen un porcentaje considerablemente menor, alrededor del 0.2 %.

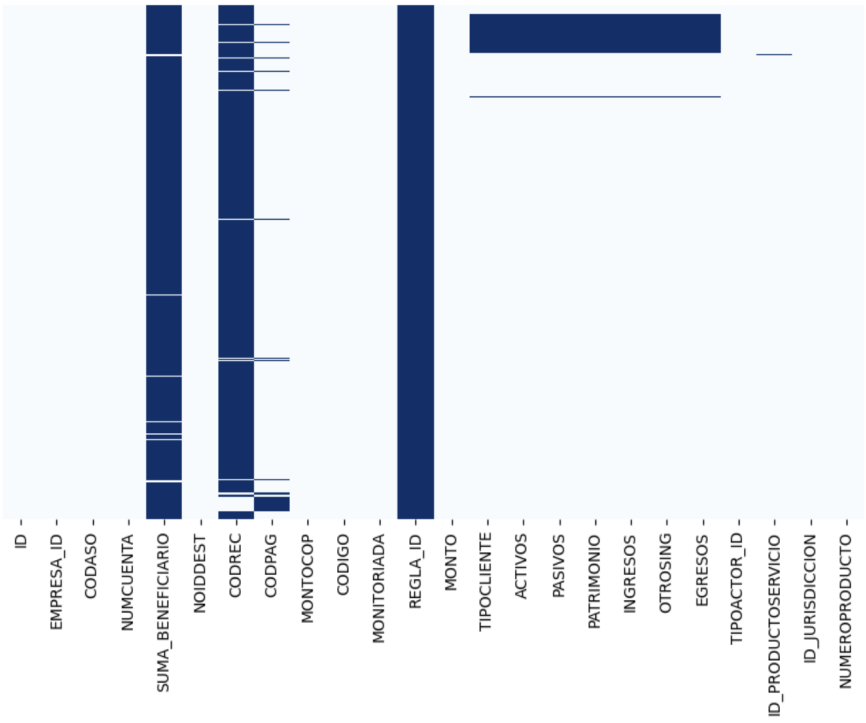


Figura 5.11: Valores faltantes variables numéricas

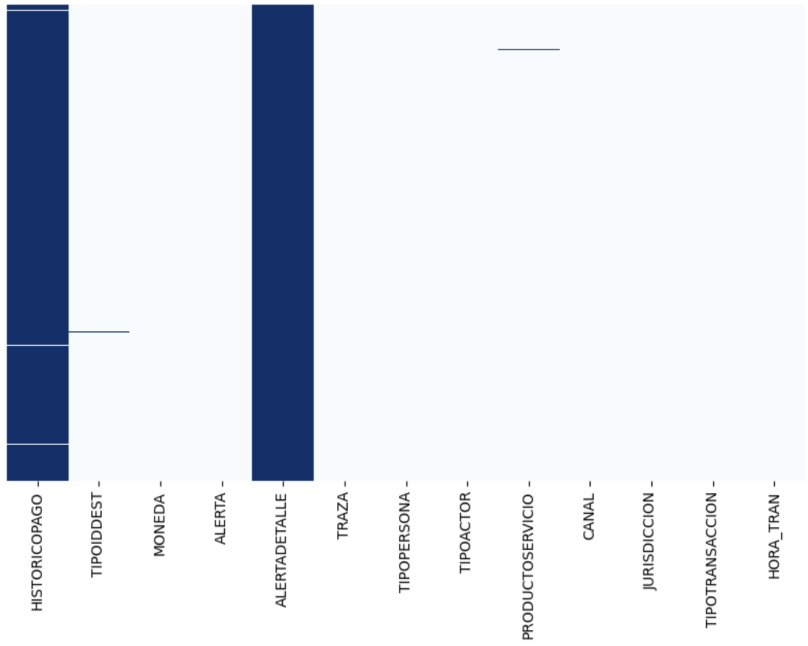


Figura 5.12: Valores faltantes variables categóricas

5.3. Preparación datos

La limpieza y preparación de los datos es un paso esencial en el desarrollo, este proceso asegura que los datos estén en un formato adecuado para su análisis y procesamiento, eliminando inconsistencias y redundancias que puedan afectar la calidad del modelo de segmentación. La limpieza garantiza que los algoritmos operen de manera eficiente y efectiva, proporcionando resultados confiables.

5.3.1. Limpieza de datos

Tras realizar un análisis exploratorio de los datos disponibles, se obtuvo un conjunto con un total de 291,802 registros y 38 columnas. Estas se clasifican inicialmente en 24 columnas numéricas, 14 categóricas y ninguna de tipo fecha. Sin embargo, se identificaron 5 columnas, **REGLA_ID**, **TIPOCLIENTE**, **TIPOACTOR_ID**, **ID_PRODUCTOSERVICIO** y **ID_JURISDICCION** que deberían corresponder a datos de tipo categóricos pero están incorrectamente definidas.

Al finalizar este ajuste, el conjunto de datos quedó compuesto por:

Tipo de columna	Cantidad
Numéricas	19
Categóricas	19

En esta etapa, se identificaron columnas que no aportan valor significativo al análisis debido a su naturaleza de valores únicos o casi únicos. Estas columnas, como IDs, códigos de transacciones, números de cuenta o identificadores de trazas, generalmente no representan características generales ni ayudan a identificar patrones relevantes. Por esta razón, se decidió eliminarlas.

- **Descartes por irrelevancia:** Las columnas **NUMCUENTA** y **TRAZA**, **NUMEROPRODUCTO**, **CODPAG**, **CODASO**, **NOIDDEST**, que representan datos específicos del cliente o la transacción, no contribuyen al objetivo del análisis. Por tanto, se procedió a eliminarlas. Además, se descartaron las columnas **EMPRESA_ID**, **MONEDA** y **MONITORIADA**, que contenían valores únicos o constantes que no aportan información útil para el análisis.
- **Validación de columnas con alta diversidad de valores:** Se identificaron columnas con más del 70 % de datos únicos, como **ID** y **CODIGO**, que tampoco ofrecían valor analítico. Estas columnas fueron descartadas por no alinearse con los objetivos del proyecto.

En el siguiente paso, se realizó una validación de las variables que contenían más del 70 % de registros nulos. La alta cantidad de valores nulos en las columnas **REGLA_ID**, **ALERTA** y **ALERTADETALLE** podría deberse a que la transacción no se categorizó como una alerta. Dado que estos campos pueden justificar su nulidad en esos casos, se decide mantenerlas.

A diferencia de otras variables, como se puede observar en las figuras, 5.11 y 5.12 **HISTORICOPAGO**, **SUMA_BENEFICIARIO** y **CODREC** presentaban altos porcentajes de valores nulos sin justificación relacionada con la clasificación de alertas. Debido a la falta de consistencia en los registros de estas columnas y su escasa relevancia para el análisis, se decidió eliminarlas. Imputar una cantidad significativa de nulos en estas variables podría introducir sesgos en el análisis, afectar la calidad de descubrir estructuras inherentes en los datos y comprometer la estabilidad del modelo. Por esta razón, la eliminación fue considerada la estrategia más adecuada para mantener la calidad del conjunto de datos.

En el conjunto de datos se identificaron columnas que representan el ID y la descripción asociada a este mismo ID. Para nuestro objetivo de entrenar modelos de *machine learning*, que se benefician principalmente de datos numéricos, estas columnas redundantes pueden generar complejidad innecesaria en el análisis. Por esta razón, se decidió conservar únicamente las columnas numéricas que representan los IDs, eliminando las columnas descriptivas correspondientes.

En concreto, se detectaron las siguientes relaciones redundantes:

- **ALERTADETALLE** y su columna numérica **REGLA_ID**.
- **JURISDICCION** y su columna numérica **JURISDICCION_ID**.
- **PRODUCTOSERVICIO** y su columna numérica **PRODUCTOSERVICIO_ID**.
- **TIPOACTOR** y su columna numérica **TIPOACTOR_ID**.

Dado que las descripciones en estas columnas categóricas no aportan valor adicional para los modelos y están directamente relacionadas con sus equivalentes numéricos, se procedió a eliminarlas del conjunto de datos. Mantener únicamente las columnas numéricas ayuda a simplificar el modelo, reducir la dimensionalidad y mejorar la eficiencia computacional, sin sacrificar información relevante para el análisis.

Este enfoque asegura que los datos utilizados para el entrenamiento de los modelos sean más compactos y específicos, optimizando su desempeño y reduciendo el riesgo de redundancias que podrían afectar la calidad del modelo.

5.3.2. Imputación de Datos

Tras realizar la limpieza de los datos y la ingeniería de características, el conjunto de datos final cuenta con un total de 291,802 registros y 18 columnas, de las cuales 10 son de tipo categórico y 8 de tipo numérico. La imputación de datos es un paso fundamental para manejar valores faltantes de forma adecuada, asegurando que el conjunto sea completo y funcional para el análisis.

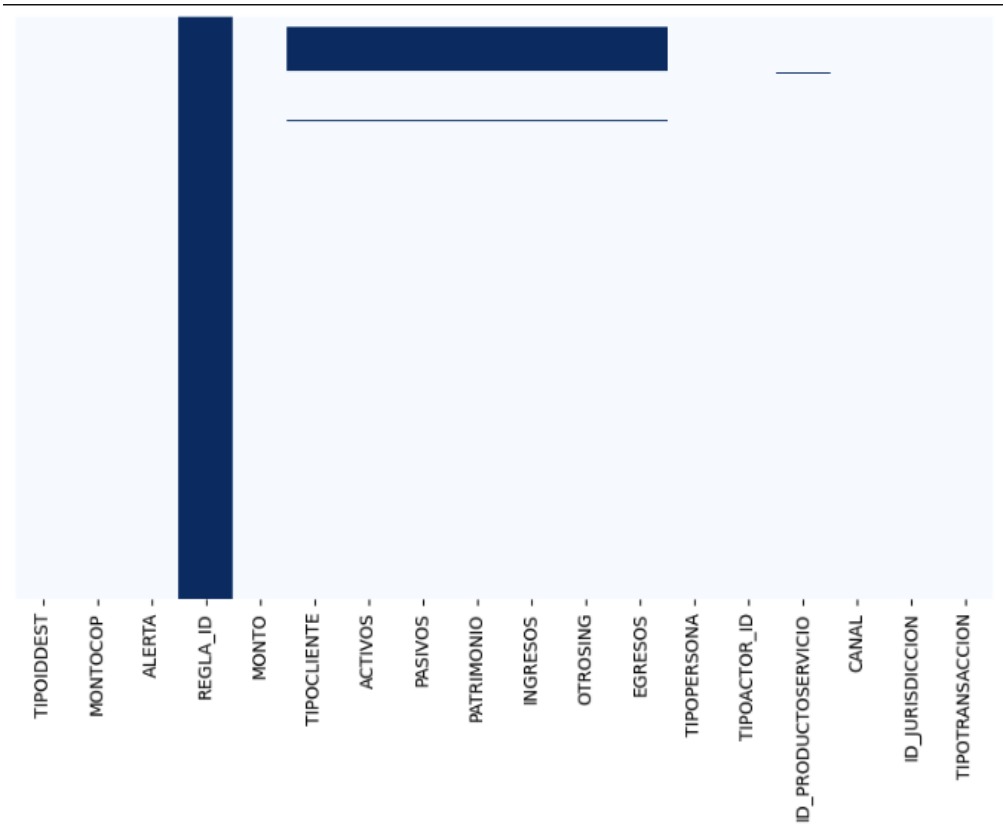


Figura 5.13: Valores faltantes después de limpieza e ingeniería de características

Columna	Cantidad de nulos
REGLA_ID	291295
TIPOCLIENTE	23144
ACTIVOS	23144
PASIVOS	23144
PATRIMONIO	23144
INGRESOS	23144
OTROSING	23144
EGRESOS	23144
TIPOIDDEST	693
ID_PRODUCTOSERVICIO	629
TIPOACTOR_ID	3

Tabla 5.2: Valores faltantes después de limpieza e ingeniería de características

De acuerdo con la figura 5.13 y la tabla 5.2, la variable con mayor cantidad de valores faltantes es **REGLA_ID**. Sin embargo, como se comentó anteriormente, y alineado con la lógica del negocio, se imputó el valor `-1` para indicar que la transacción no generó ninguna alerta.

Por último paso, se realizaron imputaciones específicas para las variables restantes. Para las variables categóricas con datos faltantes, **ID_PRODUCTOSERVICIO**, **TIPOCLIENTE**, **TIPOIDDEST** y **TIPOACTOR_ID**, se utilizó la imputación por moda, asignando el valor más frecuente de la columna a los registros nulos.

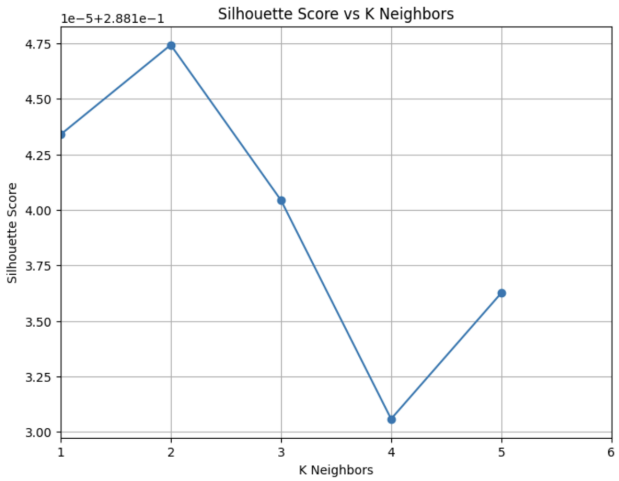


Figura 5.14: Mejor k de acuerdo a Silhouette Score

Para las variables numéricas con una mayor cantidad de datos faltantes, como **ACTIVOS**, **PASIVOS**, **PATRIMONIO**, **INGRESOS**, **OTROSING** y **EGRESOS**, se utilizó imputación basada en el método de *K-Nearest Neighbors* (KNN). Este enfoque requirió un preprocesamiento

en el que las variables categóricas fueron transformadas a formato numérico para su compatibilidad con el algoritmo. A continuación, se llevó a cabo un análisis para determinar el valor óptimo de k utilizando la métrica *Silhouette Score*, y de acuerdo con la gráfica 5.14, el mejor número de vecinos más cercanos fue $k = 2$ con un score de 0.288147.

Con $k = 2$ identificado como el valor óptimo, se procedió a imputar los datos faltantes en las variables mencionadas. Finalmente, se realizó una verificación de los datos faltantes mediante un mapa de calor (figura 6.4) que confirmó que todas las imputaciones se realizaron de manera exitosa y que no existían valores nulos restantes en el conjunto de datos, quedando con un total de 291.802 registros y 84 caracteríasticas. Este resultado asegura un conjunto de datos completo y listo para ser utilizado en el entrenamiento de modelos de *machine learning*.

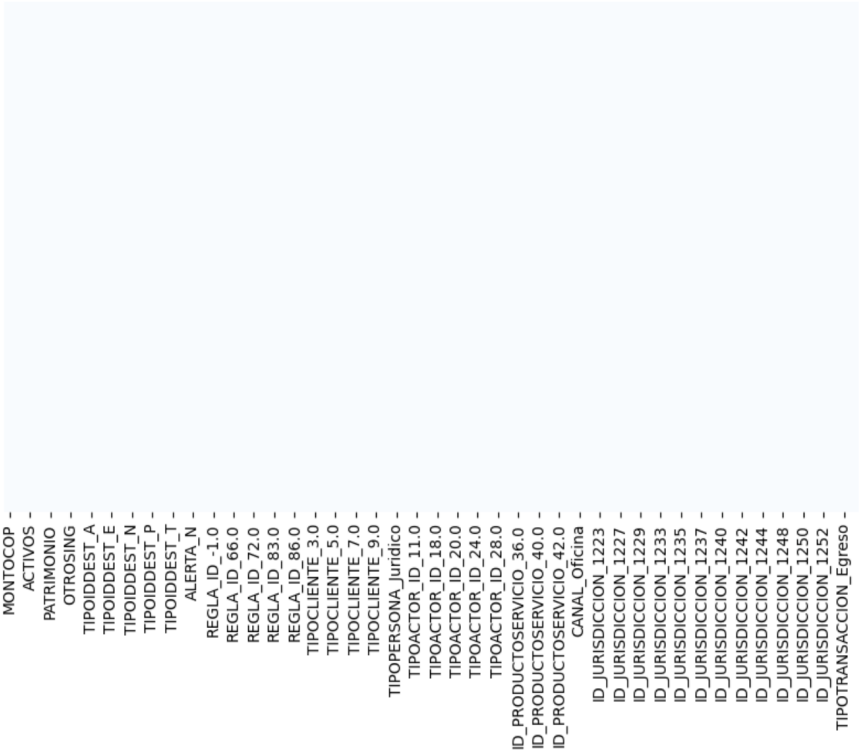


Figura 5.15: Mapa de calor sin datos faltantes

En cuanto al manejo de *outliers*, se decidió no eliminarlos inicialmente, ya que estos pueden representar grupos pequeños que definen patrones interesantes en el análisis, particularmente al identificar clústeres. En lugar de eliminarlos, se exploró la combinación de modelos utilizando el conjunto de datos tanto con los *outliers* como sin ellos, empleando técnicas como el Análisis de Componentes Principales (PCA). Este enfoque permitió comparar el impacto de los *outliers* en los resultados y evaluar su relevancia en la segmentación. Al dejar los *outliers* intactos, se asegura que el modelo pueda capturar características únicas que podrían ser clave para identificar segmentos

importantes en el conjunto de datos.

Se obtuvo al final un total de 84 características con 291,802 registros en el conjunto de datos original. Al aplicar la eliminación de los *outliers*, el conjunto de datos se redujo a 218,926 registros manteniendo las mismas 84 características. Este ajuste permite analizar tanto los datos completos como una versión filtrada, proporcionando perspectivas complementarias sobre la segmentación y el impacto de los *outliers* en la formación de los clústeres.

Capítulo 6

Implementación

6.1. Selección de Características

El primer paso en la construcción de los modelos consistió en seleccionar las variables o características más relevantes mediante dos metodologías: análisis de varianza y análisis de correlación. Estos métodos permitieron reducir la dimensionalidad del conjunto de datos, eliminando redundancias y centrándonos en las características más significativas.

6.1.1. Selección por Varianza

Mediante el análisis de varianza, se seleccionaron las siguientes características:

- MONTOCOP
- MONTO
- ACTIVOS
- PASIVOS
- PATRIMONIO
- INGRESOS
- OTROSING
- EGRESOS
- TIPOIDDEST_A
- TIPOIDDEST_C
- TIPOCLIENTE_3.0
- TIPOCLIENTE_4.0
- TIPOACTOR_ID_19.0
- TIPOACTOR_ID_24.0
- TIPOACTOR_ID_28.0

- ID_PRODUCTOSERVICIO_36.0
- ID_PRODUCTOSERVICIO_40.0
- ID_PRODUCTOSERVICIO_41.0
- ID_JURISDICCION_1227

6.1.2. Selección por Correlación

Por otro lado, el análisis de correlación identificó las siguientes características relevantes:

- TIPOIDDEST_E, TIPOIDDEST_F, TIPOIDDEST_N, TIPOIDDEST_O, TIPOIDDEST_P, TIPOIDDEST_R, TIPOIDDEST_T, TIPOIDDEST_X
- REGLA_ID_65.0, REGLA_ID_66.0, REGLA_ID_70.0, REGLA_ID_72.0, REGLA_ID_74.0, REGLA_ID_83.0, REGLA_ID_85.0, REGLA_ID_86.0
- TIPOCLIENTE_2.0, TIPOCLIENTE_3.0, TIPOCLIENTE_5.0, TIPOCLIENTE_6.0, TIPOCLIENTE_7.0, TIPOCLIENTE_9.0
- TIPOPERSONA_Natural
- TIPOACTOR_ID_11.0, TIPOACTOR_ID_16.0, TIPOACTOR_ID_19.0, TIPOACTOR_ID_20.0, TIPOACTOR_ID_22.0, TIPOACTOR_ID_24.0, TIPOACTOR_ID_26.0
- ID_PRODUCTOSERVICIO_34.0, ID_PRODUCTOSERVICIO_36.0, ID_PRODUCTOSERVICIO_39.0, ID_PRODUCTOSERVICIO_40.0, ID_PRODUCTOSERVICIO_42.0
- CANAL_SIN CANAL
- ID_JURISDICCION_1223, ID_JURISDICCION_1226, ID_JURISDICCION_1227, ID_JURISDICCION_1228, ID_JURISDICCION_1229, ID_JURISDICCION_1230, ID_JURISDICCION_1233, ID_JURISDICCION_1234, ID_JURISDICCION_1235, ID_JURISDICCION_1236, ID_JURISDICCION_1237, ID_JURISDICCION_1238, ID_JURISDICCION_1240, ID_JURISDICCION_1241, ID_JURISDICCION_1242, ID_JURISDICCION_1243, ID_JURISDICCION_1244, ID_JURISDICCION_1245, ID_JURISDICCION_1248, ID_JURISDICCION_1249, ID_JURISDICCION_1250, ID_JURISDICCION_1251, ID_JURISDICCION_1252

Ambas selecciones se probaron en el análisis, pero se determinó que la reducción basada en el análisis de varianza ofrecía mejores resultados. Por ello, se continuó con las variables seleccionadas mediante varianza para los pasos posteriores.

6.1.3. Reducción de Dimensionalidad

Vamos a buscar una mejor representación de los datos que nos permita conservar la mayor cantidad de información a través de la transformación de las variables originales en componentes principales.

Dado que todas las variables seleccionadas son numéricas, se escaló el conjunto de datos utilizando un método de normalización. Posteriormente, se aplicó Análisis de Componentes Principales para reducir la dimensionalidad del conjunto. De acuerdo a la gráfica 6.9, se determinó que el número de componentes necesarios para alcanzar el 80 % de la varianza explicada era 6.

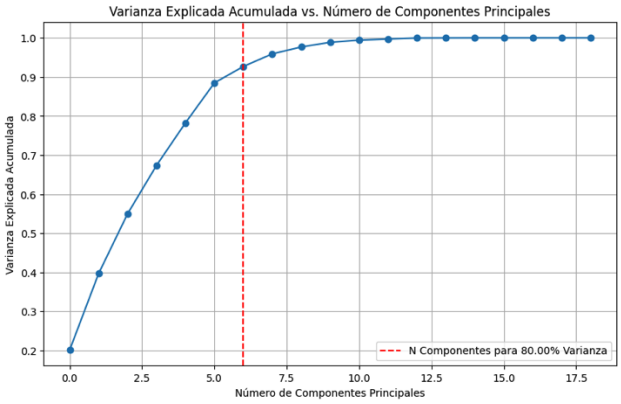


Figura 6.1: Varianza Explicada acumulada vs número de componentes principales

Los valores obtenidos para la proporción de varianza explicada por cada componente principal fueron los siguientes:

$$\texttt{pca.explained\_variance\_ratio\_} = \begin{bmatrix} 0{,}2014, & 0{,}1956, & 0{,}1528, & 0{,}1238, & 0{,}1089, & 0{,}1022, & 0{,}0421, \\ 0{,}0322, & 0{,}0177, & 0{,}0119, & 0{,}0056, & 0{,}0029, & 0{,}0025, & 0{,}0002, \\ 0{,}0002, & 0{,}00001, & 0{,}0000, & 0{,}0000, & 0{,}0000 & & \end{bmatrix}$$

La combinación de PCA y la selección de variables permite reducir la complejidad del modelo manteniendo una representación adecuada de la varianza presente en el conjunto de datos. Este enfoque asegura que los datos sean más manejables y efectivos para los algoritmos de aprendizaje automático utilizados en la segmentación.

	0	1	2	3	4	5	6
3	PASIVOS	0.293397	0.014071	0.470751	-0.055889	-0.074143	0.014679
4	PATRIMONIO	0.293397	0.014071	0.470751	-0.055889	-0.074143	0.014679
2	ACTIVOS	0.307715	0.025962	0.453264	-0.07198	-0.093124	0.023796
18	ID_JURISDICCION_1227	-0.213096	-0.099423	0.185527	-0.171308	0.400323	0.06396
14	TIPOACTOR_ID_28.0	-0.1825	0.379775	0.176234	-0.049532	-0.17049	0.036354
11	TIPOCLIENTE_4.0	-0.264805	0.359292	0.126517	-0.180295	-0.020274	0.097953
16	ID_PRODUCTOSERVICIO_40.0	0.008989	-0.440626	0.078377	-0.205193	-0.016276	0.103944
9	TIPOIDDEST_C	-0.272873	-0.080771	0.048009	-0.439841	-0.029319	0.182724
0	MONTOCOP	0.021933	0.014331	0.004173	0.263216	0.064206	0.651851
1	MONTO	0.021933	0.014331	0.004173	0.263216	0.064206	0.651851
13	TIPOACTOR_ID_24.0	0.062915	-0.03558	0.000926	0.20024	0.526081	-0.132547
15	ID_PRODUCTOSERVICIO_36.0	-0.025329	0.402461	-0.000104	-0.046401	0.323489	-0.039788
8	TIPOIDDEST_A	0.277884	0.093888	-0.018495	0.385688	0.147748	-0.165174
10	TIPOCLIENTE_3.0	0.228196	-0.336812	-0.043364	-0.054157	0.303781	-0.009151
12	TIPOACTOR_ID_19.0	-0.03577	-0.402844	-0.079555	-0.049154	-0.256761	0.05544
17	ID_PRODUCTOSERVICIO_41.0	0.024258	0.026247	-0.112174	0.361828	-0.467109	-0.087389
6	OTROSING	0.398159	0.149322	-0.182951	-0.274657	-0.034971	0.104149
7	EGRESOS	0.327588	0.148111	-0.321416	-0.271501	-0.01356	0.104654
5	INGRESOS	0.327417	0.148106	-0.321645	-0.271503	-0.013604	0.104667

Figura 6.2: Importancia componentes

En la figura 6.2 se observa el aporte de cada variable original a cada uno de los componentes principales, de aquí se puede denotar que:

- Al observar las tres variables con mayor contribución, se resalta que la primera componente principal se caracteriza por una participación positiva de las variables **OTROSING**, **EGRESOS** y **INGRESOS**. Esto sugiere que esta componente caracteriza a clientes con altos ingresos provenientes de diversas fuentes, como salarios, ventas, inversiones o alquileres. Además, estos clientes también presentan altos egresos, lo que refleja elevados niveles de gastos o inversiones realizadas.

- Se destaca que la segunda componente principal se caracteriza por una participación positiva de las variables **ID_PRODUCTOSERVICIO_36**, **TIPOACTOR_ID_28** y **TIPOCLIENTE_4**. Esto sugiere que esta componente caracteriza a clientes con una fiducia de

administración, es decir, un tipo de contrato fiduciario en el cual una persona o entidad transfiere a una fiduciaria ciertos bienes o derechos con el propósito de que esta los administre en beneficio del propio fideicomitente o de un tercero. También, esta componente nos agrupa a los clientes en la categoría "tercero general", probablemente se refiere a una categoría amplia o genérica que agrupa a actores que no encajan claramente en ninguna de las demás categorías específicas. Es posible que este tipo incluya a terceros que interactúan con la entidad, pero cuya relación o rol no se clasifica como fiduciario, proveedor, empleado, o gobierno. Por último, son clientes de tipo 4.

- Al observar la tercera componente principal, esta se destaca por la participación positiva de las variables **PASIVOS**, **PATRIMONIO** y **ACTIVOS**. Esto sugiere que esta componente caracteriza a entidades o individuos con una estructura financiera robusta. Los altos valores en los activos reflejan la posesión de recursos significativos, mientras que el patrimonio elevado sugiere una fuerte capacidad de respaldo financiero. Sin embargo, los altos pasivos muestran que estas entidades también tienen compromisos o deudas considerables, lo que podría ser indicativo de un modelo financiero apalancado o de una estrategia de inversión y financiamiento activa.

- Al observar las tres variables con mayor contribución en la cuarta componente principal, se destaca una participación positiva de las variables **ID_PRODUCTOSERVICIO_41**, **TIPOIDDEST_A** y **MONTO**. Esto sugiere que esta componente está asociada a clientes vinculados con las "Carteras Colectivas De General". Estas carteras son vehículos financieros que agrupan los recursos de varios inversionistas para invertirlos de manera conjunta en diferentes instrumentos financieros, como acciones, bonos, inmuebles, entre otros. La administración de estas carteras está a cargo de una entidad profesional que gestiona los recursos según los objetivos y políticas de inversión definidos. Además, los montos asociados en esta componente son considerablemente elevados, lo que refuerza la idea de una participación significativa en este tipo de instrumentos financieros.

- Se destaca que la quinta componente principal se caracteriza por una participación positiva de las variables **ID_PRODUCTOSERVICIO_36**, **ID_JURISDICCION_1227** y **TIPOACTOR_ID_24**. Esto sugiere que esta componente está asociada a clientes con una fiducia de administración, caracterizada por la gestión de activos o recursos específicos. Los movimientos se concentran mayoritariamente en la ciudad de Bogotá D.C., lo que podría indicar un enfoque geográfico clave para este tipo de transacciones. Además, el tipo de actor identificado como proveedor refuerza la idea de que los recursos administrados están vinculados a relaciones comerciales, como pagos o contratos asociados a la provisión de bienes o servicios.

- Por último, al observar la sexta componente principal, esta se destaca por la participación positiva de las variables **MONTOCOP**, **MONTO**, **INGRESOS** y **TIPOIDDEST_C**. Esto sugiere que esta componente caracteriza a entidades o individuos con altos niveles de ingresos y montos significativos en sus transacciones. Estas características pueden estar asociadas a

actividades financieras de gran envergadura, como inversiones o movimientos de capital relevantes. Adicionalmente, la variable **TIPOIDDEST_C** podría indicar un tipo específico de destinatario o entidad con un rol definido en estas operaciones, lo que refuerza el perfil financiero robusto asociado a esta componente.

6.2. Modelos

6.2.1. Modelo K-means

Para el análisis de segmentación, se seleccionó el modelo K-means debido a su simplicidad, eficiencia y su capacidad para identificar patrones en conjuntos de datos grandes y complejos. Además, este modelo fue estudiado en detalle durante la maestría, lo que refuerza su relevancia y aplicabilidad en el contexto del proyecto. Dado que K-means es particularmente sensible a la escala y a los datos atípicos, se realizaron diversas combinaciones para evaluar su desempeño: con y sin PCA, y con y sin *outliers*. Este enfoque permitió comparar el impacto de estas configuraciones en los resultados del modelo.

6.2.1.1. K-means con PCA

El valor óptimo de k, que representa el número de clústeres, fue determinado utilizando tres métodos diferentes para la configuración con PCA y *outliers*:

- **Método del codo:** Según la gráfica 6.3, el punto de inflexión del codo indicó que el valor óptimo de k es 3.

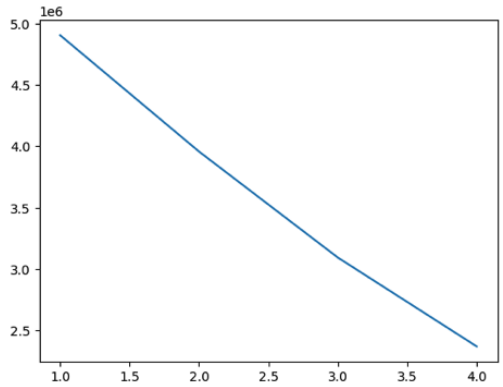
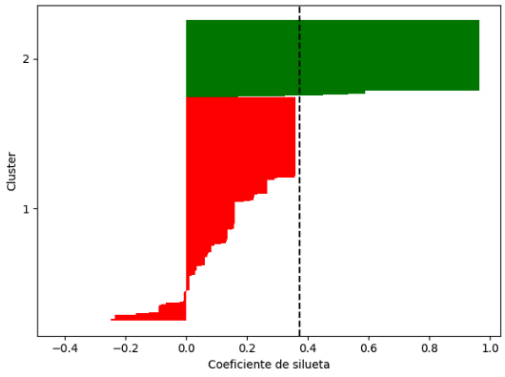
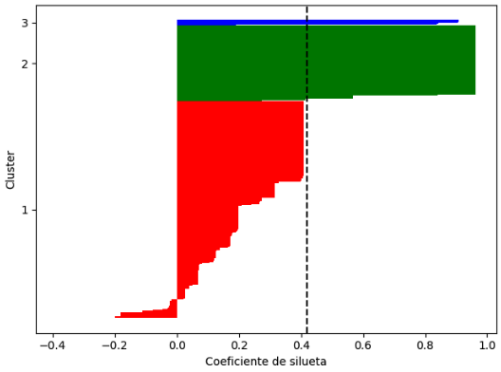


Figura 6.3: Método del codo

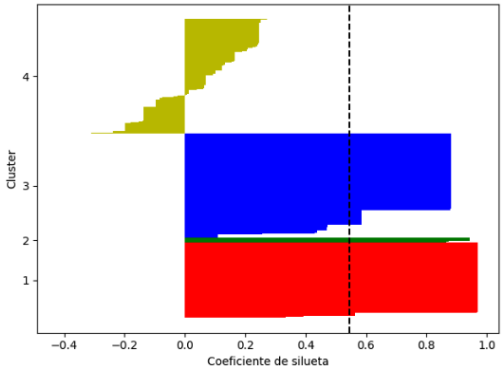
- **Método de la silueta:** Se evaluaron los valores $k = 2, 3, 4, 5$ (figura 6.4), y el mejor resultado correspondió a $k = 2$.



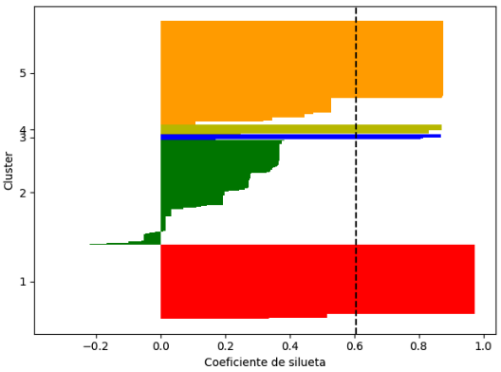
(a) Silueta con k = 2



(b) Silueta con k = 3



(c) Silueta con k = 4



(d) Silueta con k = 5

Figura 6.4

- **Índice de Calinski-Harabasz:** Aunque valores mayores de k siguen mostrando una mejora, los aumentos se vuelven menos significativos, y agregar más clusters puede sobreajustar el modelo sin aportar beneficios significativos, este criterio sugirió un valor óptimo de $k = 3$ (figura 6.9).

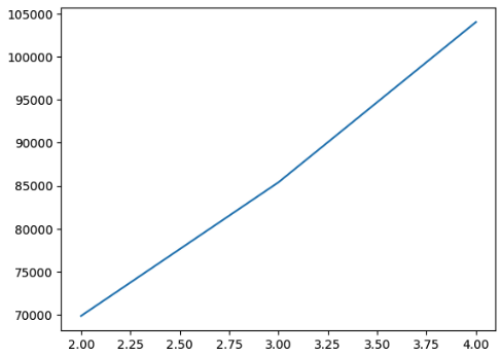


Figura 6.5: Método de Calinski Harabasz

En conclusión, considerando los resultados de los métodos del codo y de Calinski-Harabasz, se seleccionó $k = 3$ por mayoría como el número de clústeres para este análisis.

Evaluación del Modelo

Tras ajustar el modelo K-means con $k = 3$, se procedió a evaluar la distribución de los registros dentro de los clústeres resultantes. Los datos se agruparon de la siguiente manera:

```
Counter({0: 212,324, 1: 74,358, 2: 5,120})
```

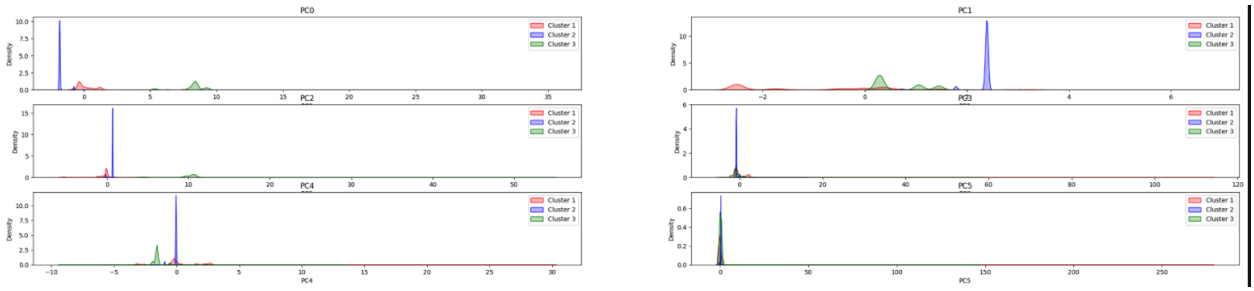


Figura 6.6: Distribución de los cluster

Estos resultados reflejan una distribución razonable entre los clústeres, con un balance adecuado que permite una interpretación clara y significativa de los segmentos. La combinación de PCA y *outliers* en este caso demostró ser una configuración efectiva, respaldada por la consistencia en la selección del número de clústeres y la calidad de los resultados obtenidos.

En la figura 6.6, se observa que la primera componente principal ($PC0$) permite una buena separación de los puntos de datos correspondientes a los tres clústeres. Por otro lado, la segunda componente principal ($PC1$) destaca en la identificación de los puntos pertenecientes al clúster 2.

Sin embargo, cabe mencionar que ninguno de los tres clústeres puede separarse de forma directa y clara de los otros al considerar la tercera ($PC2$) y la quinta ($PC4$) componente principal.

6.2.1.2. K-means sin PCA

- **Método del codo:** Según la gráfica 6.7, el punto de inflexión del codo indicó que el valor óptimo de k es 3.

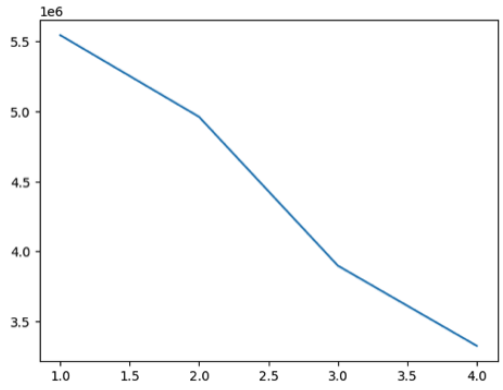
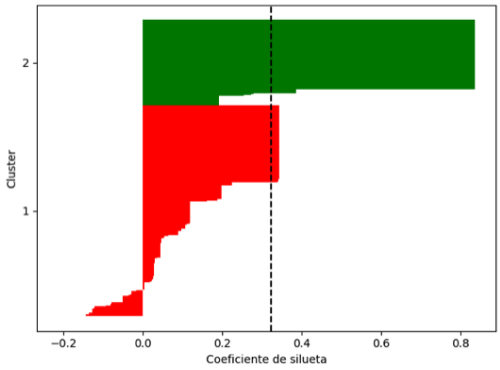
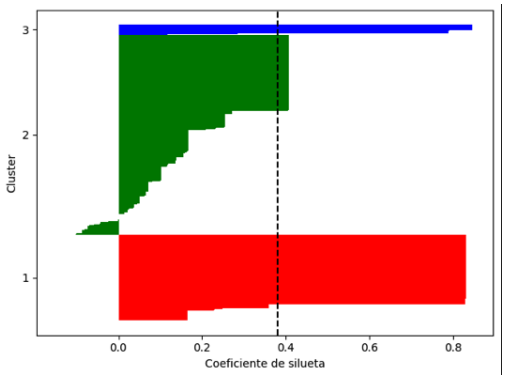


Figura 6.7: Método del codo

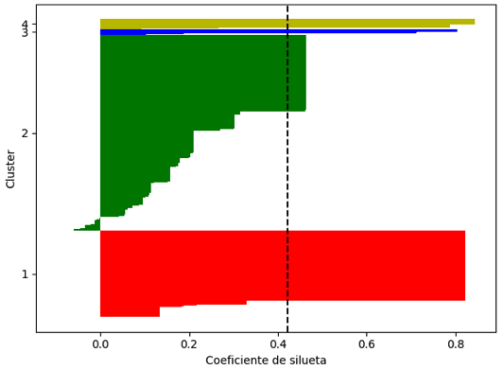
- **Método de la silueta:** Se evaluaron los valores $k = 2, 3, 4, 5$ (figura 6.8), y el mejor resultado correspondió a $k = 2$.



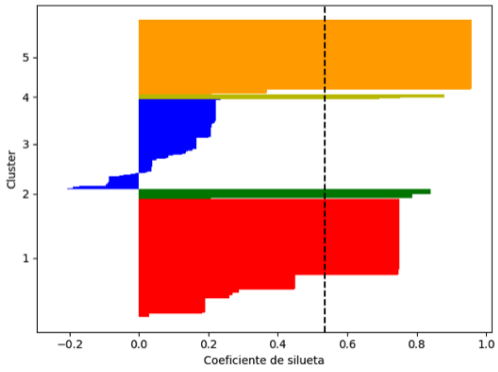
(a) Silueta con k = 2



(b) Silueta con k = 3



(c) Silueta con k = 4



(d) Silueta con k = 5

Figura 6.8

- **Índice de Calinski-Harabasz:** Este criterio sugirió un valor óptimo de $k = 3$ (figura 6.9).

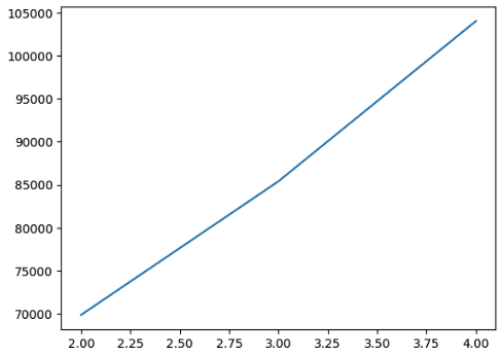


Figura 6.9: Método de Calinski Harabasz

En conclusión, considerando los resultados de los métodos del codo y de Calinski-Harabasz, se seleccionó $k = 3$ por mayoría como el número de clústeres para este análisis.

Evaluación del Modelo

Tras ajustar el modelo K-means con $k = 3$, se procedió a evaluar la distribución de los registros dentro de los clústeres resultantes. Los datos se agruparon de la siguiente manera:

```
Counter({1: 198,048 , 0: 84,266, 2: 9,488})
```

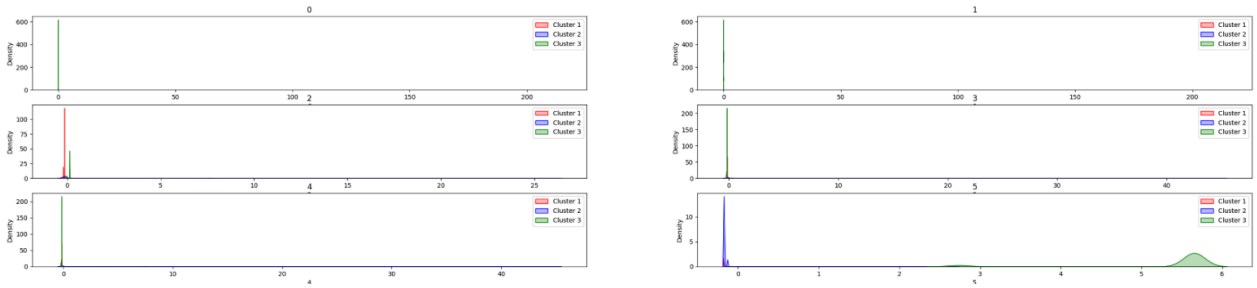


Figura 6.10: Distribución de los clúster

En la figura 6.19 se presentan las variables estandarizadas. De manera general, se observa que únicamente la variable 5 permite una correcta identificación de los puntos pertenecientes al clúster 3. Para las demás variables, la separación entre clústeres resulta considerablemente más compleja.

6.2.1.3. K-means con PCA y sin Outliers

- **Método del codo:** Según la gráfica 6.11, el punto de inflexión del codo indicó que el valor óptimo de k es 2.

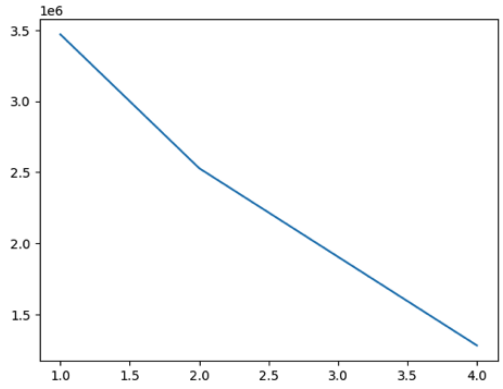
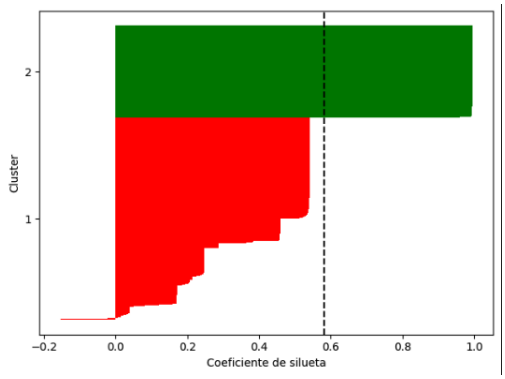
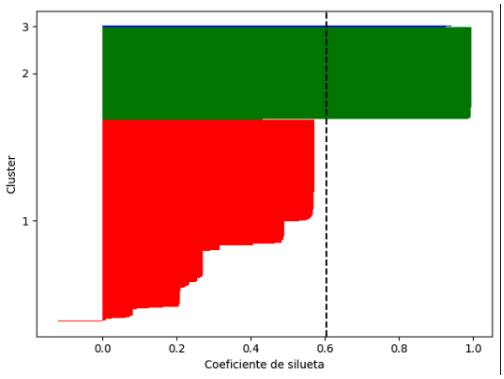


Figura 6.11: Método del codo

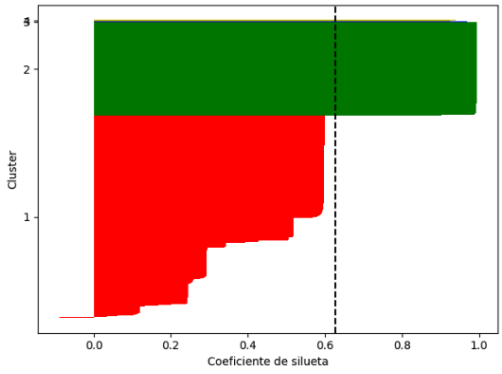
- **Método de la silueta:** Se evaluaron los valores $k = 2, 3, 4, 5$ (figura 6.12), y el mejor resultado correspondió a $k = 2$.



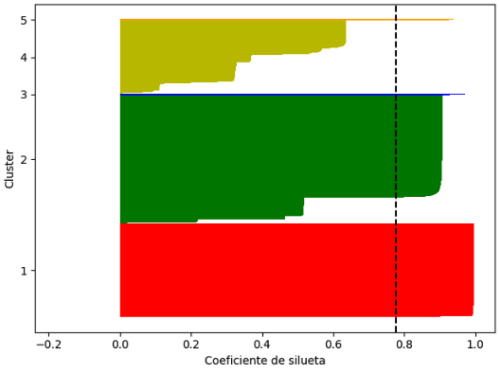
(a) Silueta con k = 2



(b) Silueta con k = 3



(c) Silueta con k = 4



(d) Silueta con k = 5

Figura 6.12

- **Índice de Calinski-Harabasz:** Este criterio sugirió un valor óptimo de $k = 4$ (figura 6.13).

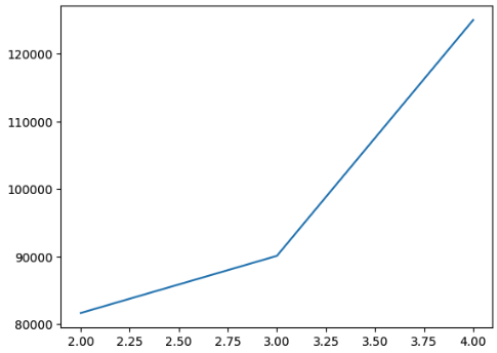


Figura 6.13: Método de Calinski Harabasz

En conclusión, considerando los resultados de los métodos del codo y de silueta, se seleccionó $k = 2$ por mayoría como el número de clústeres para este análisis.

Evaluación del Modelo

Tras ajustar el modelo K-means con $k = 2$, se procedió a evaluar la distribución de los registros dentro de los clústeres resultantes. Los datos se agruparon de la siguiente manera:

```
Counter({0: 150,258 , 0: 68,668})
```

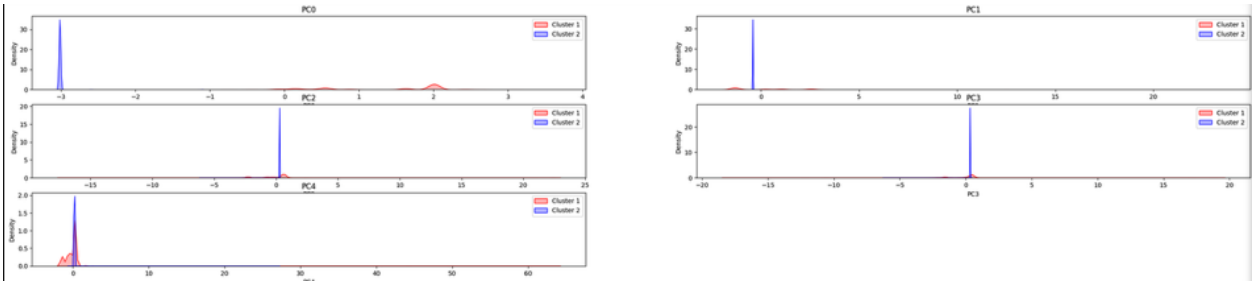


Figura 6.14: Distribución de los cluster

En la figura 6.18, se observa que la primera componente principal ($PC0$) permite una buena separación de los puntos de datos correspondientes a los dos clústeres. Por otro lado, ninguno de los dos clústeres puede separarse de forma directa y clara de los otros al considerar la segunda ($PC1$), tercera ($PC2$) y cuarta ($PC3$) componente principal.

Sin embargo, cabe mencionar que la quinta componente principal ($PC4$) destaca en la identificación de los puntos pertenecientes al clúster 1.

6.2.1.4. K-means sin PCA y sin Outliers

- **Método del codo:** Según la gráfica 6.15, el punto de inflexión del codo indicó que el valor óptimo de k es 2.

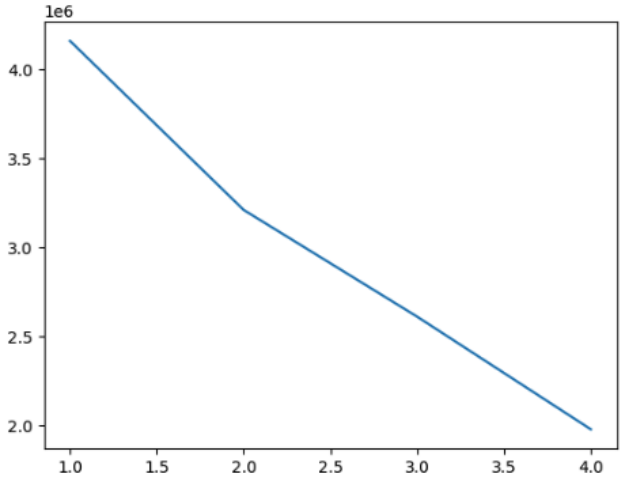


Figura 6.15: Método del codo

- **Índice de Calinski-Harabasz:** Este criterio sugirió un valor óptimo de $k = 4$ (figura 6.16).

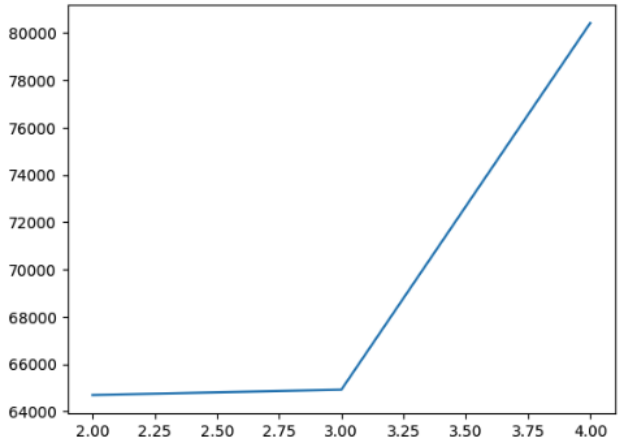
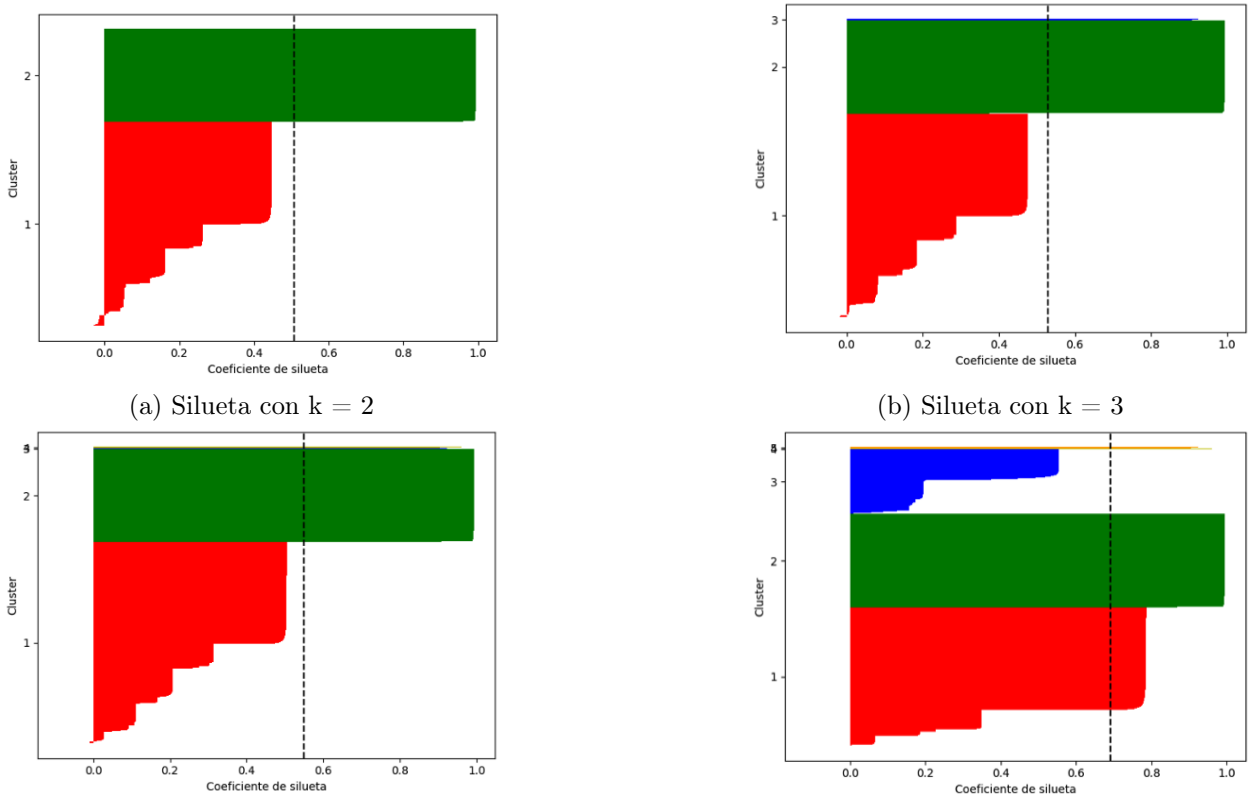


Figura 6.16: Método de Calinski Harabasz

- **Método de la silueta:** Se evaluaron los valores $k = 2, 3, 4, 5$ (figura 6.17), y el mejor resultado correspondió a $k = 2$.



(a) Silueta con k = 2

(b) Silueta con k = 3

(c) Silueta con k = 4

(d) Silueta con k = 5

Figura 6.17

En conclusión, considerando los resultados de los métodos del codo y de silueta, se seleccionó $k = 2$ por mayoría como el número de clústeres para este análisis.

Evaluación del Modelo

Tras ajustar el modelo K-means con $k = 2$, se procedió a evaluar la distribución de los registros dentro de los clústeres resultantes. Los datos se agruparon de la siguiente manera:

```
Counter({0: 150,261 , 0: 68,665})
```

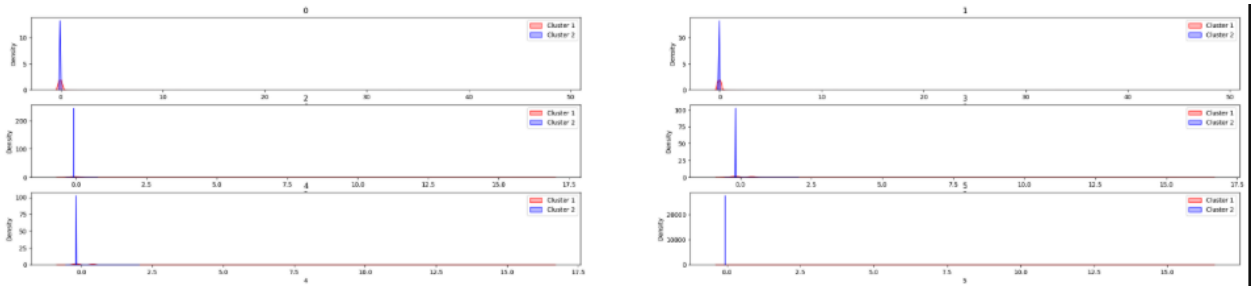


Figura 6.18: Distribución de los cluster

En la figura 6.18 se presentan las variables estandarizadas. De manera general, ninguna de las variables permite una correcta identificación de los puntos pertenecientes al clúster 1 y 2, la separación entre clústeres resulta considerablemente más compleja.

6.2.2. Modelo Clustering Jerarquico

El modelo de clustering jerárquico fue seleccionado como complemento al modelo K-means debido a su capacidad para capturar relaciones jerárquicas entre los datos, permitiendo construir una representación escalonada de los grupos. Este enfoque es especialmente útil cuando no se conoce de antemano el número de clústeres óptimos, ya que genera un dendrograma que ayuda a identificar posibles estructuras subyacentes en los datos. Además, el clustering jerárquico no requiere especificar un número inicial de clústeres, lo que proporciona mayor flexibilidad en el análisis.

El modelo jerárquico utiliza diferentes métodos de enlace para medir la similitud entre los clústeres:

- **Ward:** Minimiza la varianza dentro de cada clúster.
- **Complete:** Utiliza la distancia máxima entre puntos de diferentes clústeres.
- **Average:** Calcula la distancia promedio entre todos los puntos de diferentes clústeres.
- **Single:** Considera la distancia mínima entre puntos de diferentes clústeres.

Para determinar los parámetros óptimos del modelo jerárquico, se evaluaron diferentes valores de k (número de clústeres) en el rango de 2 a 9, junto con los métodos de enlace mencionados. El criterio de evaluación fue la métrica de silueta, que mide qué tan bien separados están los clústeres y qué tan compactos son. A continuación, se presentan los resultados obtenidos en la Tabla 6.1, de la cual la mejor combinación de parámetros fue clusters = 2 y linkage = single con un score de 0.95774:

Linkage	Clusters	Silhouette Score
ward	2	0.49450
ward	3	0.38349
ward	4	0.38587
ward	5	0.39621
ward	6	0.40359
ward	7	0.42334
ward	8	0.41203
ward	9	0.43933
complete	2	0.96257
complete	3	0.77344
complete	4	0.74648
complete	5	0.64901
complete	6	0.64015
complete	7	0.60470
complete	8	0.60543
complete	9	0.62458
average	2	0.89167
average	3	0.78337
average	4	0.72015
average	5	0.65704
average	6	0.63991
average	7	0.63683
average	8	0.61729
average	9	0.62479
single	2	0.95774
single	3	0.79405
single	4	0.84888
single	5	0.74401
single	6	0.48506
single	7	0.45309
single	8	0.60505
single	9	0.86101

Tabla 6.1: Resultados de Clustering Jerárquico con Diferentes Parámetros y Métricas Silhouette

Evaluación del Modelo

Tras ajustar el modelo de clustering jerárquico con PCA y sin outliers con $k = 2$ y linkage='single', se procedió a evaluar la distribución de los registros dentro de los clústeres resultantes.

Los datos se agruparon de la siguiente manera:

```
Counter({0: 149,805 , 0: 69,121})
```

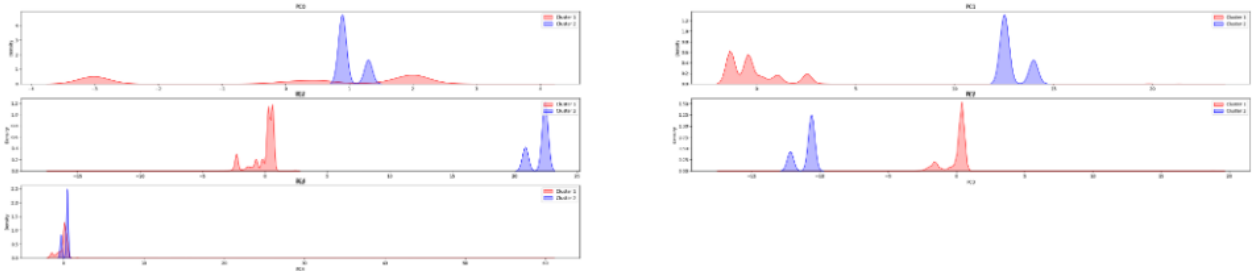


Figura 6.19

En la figura 6.19, se observa que la segunda, tercera y cuarta componente principal (*PC1, PC2, PC3*) permite una buena separación de los puntos de datos correspondientes a los dos clústeres. Por otro lado, la primera componente principal (*PC0*) destaca en la identificación de los puntos pertenecientes al clúster 2.

Sin embargo, cabe mencionar que ninguno de los dos clústeres puede separarse de forma directa y clara de los otros al considerar la quinta (*PC4*) componente principal.

Capítulo 7

Resultados

Para evaluar los modelos elaborados en la sección anterior y dado que se implementaron técnicas de análisis no supervisado, utilizamos el **"silhouette score"** como herramienta principal. Esta métrica proporciona un valor numérico que mide qué tan bien encajan los puntos de datos en sus grupos y qué tan bien están separados de otros grupos. La puntuación oscila entre -1 y 1 y se puede interpretar de la siguiente manera:

- **Valor alto**: El punto de datos coincide bien con su grupo y no coincide con otros grupos. Esto indica una buena agrupación.
- **Valor bajo**: El punto de datos no coincide adecuadamente con su grupo y sí coincide bien con otros grupos. Esto indica una mala agrupación.
- **Valor cero**: El punto de datos está en el borde de dos grupos. Esto indica una agrupación ambigua.

Modelo	Descripción	Silhouette
KNN	K = 3 con PCA con outliers	0.430332
KNN	K = 3 sin PCA con outliers	0.390052
KNN	K = 2 con PCA sin outliers	0.589651
KNN	K = 2 sin PCA sin outliers	0.515897
Agglomerative Clustering	K = 3, linkage=single, PCA sin Outliers	0.572830

Tabla 7.1: Resultado de los modelos

De acuerdo con la tabla 7.1, se observa que al aplicar el modelo KNN dividiendo el conjunto de datos en 3 grupos y utilizando las variables previamente transformadas en componentes principales para conservar la mayor cantidad de información posible, además de incluir los valores extremos o atípicos, se obtiene una puntuación aproximada de 0,43. Esto sugiere que los registros están razonablemente bien clasificados en sus correspondientes grupos.

Sin embargo, al analizar el modelo en el que los registros se dividen en 2 grupos, utilizando las variables transformadas en componentes principales pero excluyendo los valores extremos o atípicos, se obtiene una puntuación aproximada de 0,59. Este resultado indica una mejor coincidencia de los puntos con los grupos asignados, mostrando una separación más clara entre los grupos y una

agrupación más consistente. En este caso, los registros dentro de cada grupo están mejor definidos y tienen menor similitud con los registros de otros grupos, lo que refleja una agrupación más eficiente.

Adicionalmente, el modelo *Agglomerative Clustering* con $k = 3$, utilizando el método de enlace `single` y con las variables transformadas mediante PCA, sin incluir los valores extremos o atípicos, alcanzó una puntuación de 0,57283. Este resultado lo posiciona como una alternativa competitiva al modelo KNN, logrando una segmentación bien definida y compacta en los datos, particularmente cuando los *outliers* son excluidos.

Capítulo 8

Conclusiones

- En relación con el primer objetivo específico, se concluye que la mayoría de las entidades o individuos en la base de datos corresponden al tipo persona jurídica, mientras que solo un 0.6 % están clasificados como personas naturales. Además, se identificó que una mayor proporción de los movimientos fueron realizados en una oficina, lo que sugiere una preferencia por este canal para gestionar sus operaciones.

- El análisis de los tipos de actor en la base de datos revela una diversidad significativa de roles asociados a las transacciones y actividades registradas. Entre los actores principales se identifican categorías como Tipo Tercero General, que agrupa a entidades o individuos no clasificados en roles específicos y roles más definidos como Fideicomitente, Fideicomiso y Sociedades Fiduciarias, que están directamente relacionados con operaciones fiduciarias. Por otro lado, actores como Proveedor y Empleado reflejan relaciones comerciales y laborales, mientras que categorías como Beneficiario / No Autorizado y Gobierno sugieren la inclusión de actores relacionados con regulaciones, beneficios o interacciones con entidades estatales. Esta variedad indica un ecosistema diverso de participantes, cada uno desempeñando un papel clave en la dinámica de las operaciones analizadas.

- El análisis de los productos y servicios evidencia una amplia oferta orientada a la gestión y administración de recursos financieros en diversos contextos. Productos como la "Fiducia de Administración" y la "Fiducia en Garantía - Fuente de Pagos" destacan por su capacidad de estructurar y garantizar el cumplimiento de obligaciones financieras, mientras que la "Fiducia Inmobiliaria - Administración y Pagos" responde a la necesidad de administrar eficientemente proyectos inmobiliarios. Por otro lado, las "Carteras Colectivas De General" y los "Fondos de Capital Privado" reflejan opciones de inversión diversificadas para clientes con distintos perfiles, desde aquellos que buscan participar en mercados colectivos hasta quienes apuestan por instrumentos de capital privado. Finalmente, los recursos vinculados al "Sistema General de Seguridad Social - Pasivos Pensionales" subrayan el rol de estos productos en la administración de obligaciones de largo plazo, como pensiones. Este portafolio de productos y servicios demuestra un enfoque integral que combina administración eficiente, oportunidades de inversión y soluciones orientadas al cumplimiento financiero.

- El análisis de las distribuciones de las variables financieras (Activos, Pasivos, Patrimonio, Ingresos, Otros Ingresos y Egresos) sin outliers muestra patrones bimodales o multimodales en varias de ellas, lo que sugiere la presencia de subgrupos bien diferenciados dentro de los datos. Por ejemplo, en las variables Activos y Patrimonio, se observa una clara separación

entre grupos, lo que podría indicar diferencias en el tamaño o tipo de las entidades analizadas. De manera similar, las distribuciones de Ingresos y Egresos reflejan la coexistencia de entidades con niveles financieros significativamente distintos. Esta segmentación también se aprecia en variables como Otros Ingresos, lo que refuerza la heterogeneidad de los datos. Este comportamiento sugiere la necesidad de profundizar en un análisis de agrupamiento o segmentación para identificar patrones específicos dentro de estos subgrupos y relacionarlos con características clave de las entidades o individuos estudiados.

- El mapa de correlación de Spearman revela relaciones significativas entre las variables financieras y operativas en el conjunto de datos. Se observa una fuerte correlación positiva entre las variables Activos, Pasivos y Patrimonio, lo cual es consistente con la relación contable en la estructura financiera de las entidades. De manera similar, las variables Ingresos y Egresos también presentan una alta correlación positiva, lo que indica que entidades con mayores ingresos tienden a realizar mayores gastos. Por otro lado, las correlaciones de menor magnitud como es el caso entre Monto y otras variables como TIPOCLIENTE o ACTIVOS_NORMALIZADOS, sugieren relaciones más débiles o indirectas.

  Este análisis resalta la importancia de las principales variables financieras en la estructura del modelo y sugiere que dichas variables podrían ser indicadores clave para identificar patrones o realizar segmentaciones en el conjunto de datos. Adicionalmente, las correlaciones negativas de menor magnitud indican que algunas variables tienen poca o nula dependencia, lo que podría reflejar comportamientos heterogéneos o específicos dentro de los datos.

- En este trabajo, se lograron cumplir los objetivos planteados al implementar y entrenar diferentes modelos de Inteligencia Artificial (IA) y Machine Learning (ML) para la segmentación de factores de riesgo relacionados con el Lavado de Activos y Financiación del Terrorismo (LAFT). Los modelos evaluados incluyeron K-means y el modelo jerárquico, considerando diversas configuraciones y parámetros con el propósito de optimizar su desempeño y determinar el modelo más efectivo.

  Para el modelo K-means, se realizaron pruebas exhaustivas para determinar el número óptimo de clústeres (k) utilizando métodos como el del codo, la silueta y el índice de Calinski-Harabasz. Estas evaluaciones permitieron identificar $k = 3$ como el valor más adecuado en la mayoría de los casos. Adicionalmente, se implementaron cuatro configuraciones distintas del modelo KNN, las cuales incluían combinaciones con y sin PCA, así como con y sin *outliers*, con el fin de analizar el impacto de estas decisiones en la segmentación.

- Los resultados obtenidos con la métrica Silhouette revelaron que las configuraciones del modelo KNN varían significativamente en términos de desempeño. Por ejemplo, el modelo KNN con $k = 2$, PCA y sin *outliers* presentó el mejor valor de Silhouette (0.589651), mientras que el modelo KNN con $k = 3$, PCA y con *outliers* obtuvo un valor de 0.430332, y el modelo KNN con $k = 3$, sin PCA y con *outliers* registró un valor de 0.390052. Estos resultados destacan la importancia de considerar tanto la reducción de dimensionalidad como la presencia de *outliers* al momento de entrenar y evaluar los modelos.

- Asimismo, se realizó la comparación de los resultados obtenidos en los modelos entrenados, determinando que la combinación de PCA y la eliminación de *outliers* ofrece el mejor equilibrio entre las métricas de rendimiento. Este enfoque permite optimizar la implementación del modelo en el proceso de identificación de riesgos LAFT, mejorando la eficiencia del cumplimiento normativo y fortaleciendo los mecanismos de prevención.

- Un aspecto clave para alcanzar estos resultados fue realizar una adecuada exploración de los datos, seguida de una limpieza de datos y selección de características. Estas etapas iniciales del proceso garantizaron que los modelos no sufrieran de problemas relacionados con una alta dimensionalidad (demasiadas filas y columnas), lo que podría haber afectado negativamente su rendimiento. La reducción y transformación de las características permitió mejorar la eficiencia computacional y asegurar que los modelos se entrenaran con datos relevantes y significativos.

- En conclusión, la implementación de diferentes enfoques permitió identificar configuraciones robustas y efectivas para la segmentación de factores de riesgo LAFT. Este análisis sienta las bases para una aplicación más precisa y confiable de los modelos en escenarios reales. Además, los métodos y resultados obtenidos pueden servir como referencia para futuros trabajos enfocados en optimizar la segmentación y evaluación de riesgos en diversos contextos normativos y financieros.

- Como pasos a seguir, que no están dentro del alcance de este proyecto, se sugiere una actividad intraorganizacional en donde se pase por despliegue una capa media o aplicativo web API, que reciba el listado de usuarios y genere el documento que GIRO recibe como entrada para mostrar en su front los clusters de cada una de las personas, esto con el fin de automatizar actividades.

Teniendo en cuenta estos resultados, se plantean actividades como **Trabajo Futuro** en el contexto de Ciencia de datos, lo que permitiría enfocar el análsis de manera mas especifica según la necesidad de la compañia, cliente o diferentes sectores diferentes al financiero.

- Realizar una análisis exploratorio y modelado de datos enfocado en los diferentes factores de riesgo de manera aislada (Jurisdiccón, Canales, Productos y Clientes) con el fin de encontrar las variables que sean mas significativas para cada sector.

- Una vez desplegado en producción, y certificado con el cliente en particular **FIDUCOLDEX**, se podría tratar con la información de otros clientes.

- Aplicar la arquitectura de datos holística planteada en diferentes contextos y sectores de información, para el EDA y modelado de acuerdo a lo establecido segun la arquitectura de datos, proyecto y de datos dado que el pipeline permitiría usarse de manera estandar.

Bibliografía

[1] S. de la Economía Solidaria, "Circular básica jurídica 2021," Enero 2021, publicada en el Diario Oficial No. 51.571. [Online]. Available: http://supersolidaria.gov.co/es/content/nueva-circular-basica-juridica

[2] S. F. de Colombia, "Circular externa 22 de 2007: Sistema de administración de riesgo de lavado de activos y de la financiación del terrorismo (sarlaft)," 2007.

[3] D. A. de la Función Pública, "Ley 1121 de 2006, artículo 27," 2006, accessed: 2023-10-21. [Online]. Available: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=22647

[4] Superintendencia Financiera de Colombia [SFC], "Circular externa no. 027: Imparte instrucciones relativas a la administración del riesgo de lavado de activos y de financiación del terrorismo," 2020.

[5] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big Data*, vol. 1, no. 1, pp. 51–59, 2013.

[6] M. A. Molina, "Segmentación de clientes y definición de alertas para la prevención de riesgos de lavado de activos y financiación del terrorismo (sarlaft): un estudio económico aplicado a entidad financiera colombiana en 2017," Colombia, 2019. [Online]. Available: https://repository.eafit.edu.co/server/api/core/bitstreams/b826b455-7c60-4a05-b39a-73a326380cf0/content

[7] D. A. Enriquez Sanchez, "Modelo matemático para estimar el riesgo de lavado de activos por clientes de pequeñas instituciones financieras," Colombia, 2019.

[8] S. Correa Correa and L. Y. Montoya Gómez, "Análisis de segmentación y alertamiento transaccional para la gestión de riesgos sarlaft en el sector financiero," 2024.

[9] J. Hendieh, M. Schneider, and T. Sakr, "Fraud detection and prevention," *Middle East Journal of Scientific Research*, vol. 31, pp. 44–52, 2023.

[10] A. Lozano Villa, "El perfil financiero: una estrategia para detectar el lavado de activos," *Revista Criminalidad*, vol. 50, no. 2, pp. 43–55, 2009, accessed: 2023-10-21. [Online]. Available: https://revistacriminalidad.policia.gov.co:8000/index.php/revcriminalidad/article/view/458

[11] R. Herrera García, "El delito de lavado de activos: su complejidad y las dificultades de su investigación," 2018, accessed: 2023-10-21. [Online]. Available: https://repositorio.uniandes.edu.co/entities/publication/33739b7c-784a-46da-bd5c-bcd29a108b18

[12] N. A. Daza, "Elaboración de un modelo de segmentación de jurisdicciones que aporte a la identificación de riesgos de lavado de activos y financiación del terrorismo

por este factor en una institución microfinanciera de la ciudad de popayán," 2019, accessed: 2023-10-21. [Online]. Available: https://fupvirtual.edu.co/repositorio/files/original/7202dac0a42c1727a1f2547095b45eeac420a477.pdf

[13] Compliance, "Desglose de factores de riesgo en la/ft: Un enfoque de segmentación," 2024. [Online]. Available: https://www.compliance.com.co/desglose-de-factores-de-riesgo-en-la-ft-un-enfoque-de-segmentacion/

[14] L. E. Pérez Pérez, "Metodología para segmentación de un sarlaft," Colombia, 2020, trabajo de grado.

[15] A. desconocido, "Técnicas de minería de datos aplicadas en la detección de fraude: Estado del arte," *ResearchGate*, Fecha desconocida. [Online]. Available: https://www.researchgate.net/publication/240724702_Tecnicas_de_Mineria_de_Datos_Aplicadas_en_la_Deteccion_de_FraudeEstado_del_Arte

[16] U. Icesi, "Fases de asum-dm," accessed: 2023-10-21. [Online]. Available: http://i2t.icesi.edu.co/ASUM-DM_External/cognos.external.asum-DM_Teaser/guidances/supportingmaterials/resources/ASUM-DM_phases.jpg

[17] IBM, "Analytic solutions unified method for data mining (asum-dm)," 2024, consultado el 1 de diciembre de 2024. [Online]. Available: https://www.ibm.com

[18] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, *CRISP-DM 1.0: Step-by-step Data Mining Guide*, SPSS Inc., 2000. [Online]. Available: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

[19] D. Lake, "Delta lake documentation," 2024, consultado el 1 de diciembre de 2024. [Online]. Available: https://delta.io