“Modelo de segmentación de clientes en el marco del sistema de administración del riesgo de lavado de activos y financiación del terrorismo *(SARLAFT)* para cooperativas financieras, usando técnicas de Big Data”

Juan David Villa Alvarez

Carlos Andrés Arbeláez Zapata

Daniel Valencia Restrepo

Trabajo de Grado presentado como requisito para optar al título de:

Especialista en Big Data e Inteligencia de Negocios

Asesor: Ingrid Durley Torres Pardo
Doctora en Ingeniería

Universidad Católica Luis Amigó
Facultad de Ingenierías y Arquitectura
Especialización en Big Data e Inteligencia de Negocios
Medellín, Colombia
2025

Dedicatoria

A todas las personas que, de una u otra manera, nos brindaron conocimientos, gestos, consejos o acciones de apoyo durante este proceso académico. Su presencia hizo que este logro fuera posible.

"Ningún logro es fruto del esfuerzo individual; siempre es el resultado de muchos corazones trabajando en conjunto." — John C. Maxwell.

Agradecimientos

Agradecemos a nuestras familias, amigos, docentes e institución por su apoyo y acompañamiento.
Este logro también es de ustedes.

Resumen

El presente trabajo desarrolla un modelo de segmentación de clientes para una cooperativa financiera en Colombia, utilizando técnicas de Big Data en cumplimiento del Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT). El objetivo central fue clasificar a los clientes según su nivel de riesgo LA/FT mediante la integración de variables normativas, financieras, sociodemográficas, geográficas y transaccionales. El diseño metodológico se basó en la metodología CRISP-DM, permitiendo estructurar las fases de comprensión del negocio, preparación de datos, modelamiento, evaluación y despliegue del modelo.

La base de datos inicial incluyó 43.713 registros de clientes, que fueron depurados y enriquecidos con información de actividad económica, sectores de riesgo y ubicación geográfica, en conjunto con expertos de cumplimiento. Se asignaron ponderaciones y escalas de riesgo para variables categóricas, y se normalizaron las variables numéricas para garantizar la coherencia analítica. Posteriormente, se implementó un modelo de clustering no supervisado (K-means), terminándose tres clústeres óptimos para personas naturales y jurídicas según las métricas Silhouette Score y Davies-Bouldin.

Los resultados evidenciaron clústeres de alto riesgo que requieren debida diligencia ampliada, así como grupos homogéneos de riesgo medio y bajo que permiten focalizar recursos de cumplimiento. El modelo demostró mayor precisión en personas jurídicas, mientras que para personas naturales se identificaron oportunidades de mejora asociadas a calidad de datos. En conclusión, la segmentación propuesta constituye una herramienta eficaz para fortalecer el cumplimiento normativo SARLAFT, optimizar controles y priorizar esfuerzos de mitigación de riesgos.

Palabras clave: Big Data, SARLAFT, segmentación de clientes, clustering, K-means, riesgo LA/FT, cooperativas financieras.

Tabla de Contenido

Lista de Tablas

Lista de Ilustraciones

1. Introducción

Las cooperativas financieras en Colombia se han consolidado como actores estratégicos en el sistema económico nacional. Según datos de la Confederación de Cooperativas de Colombia (Confecoop, 2022), este sector contribuye con aproximadamente el 3% del PIB, desempeñando un papel clave en la inclusión financiera, especialmente en poblaciones con acceso limitado a la banca tradicional debido a condiciones geográficas o socioeconómicas. Su modelo basado en principios de solidaridad, ayuda mutua y gestión democrática ha permitido ofrecer servicios financieros esenciales, como crédito y ahorro, a segmentos históricamente excluidos (DANE, 2021).

Sin embargo, este crecimiento conlleva desafíos regulatorios, particularmente en materia de prevención del Lavado de Activos y Financiación del Terrorismo (LA/FT). El **Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT)**, establecido por la Superintendencia Financiera de Colombia (SFC, 2020), exige a estas entidades implementar controles robustos para mitigar riesgos operacionales, reputacionales y legales. Dado el contexto colombiano, donde factores como la informalidad y la dinámica regional incrementan la exposición al LA/FT, las cooperativas deben adoptar metodologías proactivas para la identificación, segmentación y monitoreo de riesgos.

En este marco, el presente trabajo propone un **modelo de segmentación de clientes basado en técnicas de *big data***, que permita clasificar el nivel de riesgo mediante criterios como:

- **Jurisdicción** (ubicación geográfica),
- **Actividad económica**,
- **Personas Expuestas Políticamente (PEP)**,
- **Composición patrimonial**,
- **Comportamiento transaccional**.

.

2. Planteamiento del Problema

Este trabajo busca proponer un modelo de segmentación de clientes en cooperativas financieras de Colombia, a través de la implementación de herramientas de Big data para la adecuada gestión integral de riesgos asociados con lavado de activos y financiación del terrorismo (LAFT) en cumplimiento del marco normativo SARLAFT. Lo anterior, por la importancia que ha comenzado a tomar el sector solidario en la economía del país, entendiendo que actúan como mecanismos de acceso al ahorro y crédito de población que por sus características no se les facilita conseguir este tipo de servicios en la banca convencional.

Teniendo en cuenta lo anterior, las Cooperativas financieras en el contexto colombiano, se enfrentan a una serie de retos y dificultades en el desarrollo de su objetivo social, donde se exponen de forma permanente a riesgos de tipo LAFT, operacional, reputacional, fraude e incumplimiento legal, lo que condiciona el actuar de estas organizaciones y las obliga a diseñar mecanismos de control y mitigación de estos riesgos. En este sentido, el marco SARLAFT, busca que las organizaciones de este tipo, diseñen e implementen un sistema robusto que permita responder a estos escenarios y se pueda hacer una gestión integral de riesgos a través de los controles que operan en cada una de las fases del sistema, como es el caso de la segmentación de contrapartes que debe realizarse en las etapas de identificación y monitoreo del riesgo.

Dicha segmentación es de carácter obligatorio en todas las cooperativas financieras, teniendo en cuenta las exigencias contenidas en el Título V - INSTRUCCIONES PARA LA ADMINISTRACIÓN DEL RIESGO DE LAVADO ACTIVOS Y FINANCIACIÓN DEL TERRORISMO definidas en la circular básica jurídica de la Superintendencia de economía solidaria (Supersolidaria), por lo cual, cada organización debe operar con un SARLAFT adecuado a sus necesidades, que responda a las obligaciones normativas y permita tener una gestión de riesgos dinámica y en permanente ajuste a los cambios del país, el sector y la economía.

En este sentido, con este trabajo se busca realizar una segmentación de los clientes para las cooperativas financieras, basada en el nivel de riesgo que pueda determinarse a partir de criterios como: jurisdicción (ubicación geográfica), actividad económica, personas expuestas políticamente (PEPs), composición patrimonial y comportamiento transaccional, consiguiendo un sistema de segmentación de clientes que permita optimizar las debidas diligencias ampliadas y el diseño de controles en los procesos que se desarrollan.

En conclusión, este trabajo se soporta en aspectos cómo:

1. **Brecha regulatoria**: las cooperativas enfrentan desafíos para implementar sistemas de gestión de riesgos adaptados a su escala y contexto, que se ajusten a las diferentes exigencias normativas, más exactamente el marco SARLAFT, el cual determina el marco y controles generales que se deben garantizar sin dar un detalle o especificación sobre el cómo se deben realizar, por tanto, cada organización es responsable de diseñar su propio sistema, soportado en controles que permitan dar una adecuada gestión del riesgo al que se ven expuestas en un entorno como el Colombiano.

2. **Oportunidad tecnológica**: El *big data* permite optimizar procesos manuales, reducir falsos positivos y mejorar la eficiencia en la asignación de recursos de cumplimiento.

3. **Sostenibilidad**: Un modelo preciso fortalece la reputación institucional y previene sanciones, como las documentadas por Gómez (2019) en casos de incumplimiento normativo.

3. Justificación

El contexto colombiano para organizaciones del sector financiero como las cooperativas, representa grandes desafíos, dado el volumen y nivel de exigencias normativas, tal es el caso del marco SARLAFT (Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo) un marco propuesto para la mitigación de riesgos relacionados con lavado de activos y financiación del terrorismo (LA/FT). Dicho marco debe ser adaptado y cumplirse de forma rigurosa por estas organizaciones, mientras que a la par, las mismas, deben desarrollar estrategias que permitan la mayor captación de clientes, desarrollando estrategias comerciales que sean atractivas y permitan la obtención de ganancias. La Resolución 042 de 2020 de la UIAF y la Circular Externa 029 de 2021 de la Superintendencia Financiera establecen la obligatoriedad de implementar controles robustos a partir del desarrollo de actividades o herramientas que permitan la identificación, clasificación y monitoreo de perfiles de riesgo en clientes, con el fin de mitigar posibles operaciones ilícitas, que resulten en escenarios asociados con LA/FT, donde se materialicen riesgos de contagio y sanciones, además del impacto reputacional derivado de estos. Actualmente, muchas cooperativas aún dependen de métodos tradicionales de segmentación, lo que limita su capacidad para detectar patrones complejos de comportamiento asociados a escenarios de riesgos como los ya mencionados, exponiéndose de forma significativa a incumplimientos y escenarios que podrían comprometer su operación y resultados.

La aplicación de técnicas de Big Data y analítica avanzada, como modelos de segmentación no supervisada (clustering), permitiría una identificación más precisa de grupos de clientes según su perfil de riesgo, tomando aspectos transaccionales, socioeconómicos y geográficos, los cuales, están alineados a los requerimientos del SARLAFT, permitiendo una mayor gestión integral de los riesgos LA/FT y la toma de decisiones sobre información mucho más precisa. Estudios recientes demuestran que el uso de algoritmos como K-means mejora la detección de anomalías en entornos financieros (Zhang,2021), propiciando ambientes donde se puedan desarrollar controles más precisos que permitan la reducción de impactos o probabilidad de materialización de riesgos. En este sentido, la utilización de inteligencia artificial o big data en procesos o entornos corporativos, reduce la identificación de falsos positivos en el reporte de operaciones inusuales, optimizando costos operativos (Pérez,2022) y blindando de una mejor manera a la organización.

Además de lo anterior, organizaciones como las cooperativas financieras que juegan un papel relevante en la configuración de la economía regional en el país, podrían gestionar de forma más efectiva su riesgo reputacional y fortalecer sus operaciones haciéndose más eficientes y eficaces en el diseño y ejecución de controles para escenarios relacionados con LA/FT, además de sumar a la transparencia y fortalecimiento del sector financiero colombiano.

4. Marco de Referencias

El estudio se centra en el diseño de un **modelo de segmentación de clientes por perfil de riesgo** para cooperativas financieras, utilizando *big data* y fuentes como el DANE y la Unidad de Información y Análisis Financiero (UIAF). La segmentación prioriza cinco variables clave, alineadas con los requisitos del SARLAFT.

En este sentido, la revisión bibliográfica permitió identificar que el modelo de segmentación de clientes bajo SARLAFT debe integrar variables clave como perfil transaccional, actividad económica y vinculación a sectores de alto riesgo (Superintendencia Financiera de Colombia, 2020; 2022), respaldadas por técnicas de big data como clustering no supervisado (K-means, DBSCAN) para categorizar niveles de riesgo (Zhang et al., 2021; Jain, 2010). Estudios como Jovel (2020) y Correa & Montoya (2024) evidencian la efectividad de estos métodos en cooperativas, logrando una reducción de hasta un 30% en falsos positivos. No obstante, se identificaron brechas en la implementación de IA para AML en contextos latinoamericanos (Gómez & Ramírez, 2020), destacando la necesidad de adaptar modelos a normativas locales como la Ley 1908 de 2018 y el Decreto 1674 de 2020. Además, el marco CRISP-DM (Chapman et al., 2000) se propone como metodología para garantizar un flujo estructurado en el desarrollo del modelo, desde la extracción de datos hasta la validación con métricas de silueta (Kaufman & Rousseeuw, 1990). Finalmente, el informe de Confecoop (2022) resalta la urgencia de controles eficaces en cooperativas, dado su impacto social y exposición a riesgos LA/FT. Estos hallazgos sustentan la viabilidad del modelo propuesto para este trabajo.

En este orden de ideas, es preciso definir los siguientes conceptos:

Big Data: volúmenes de datos disponibles en diversos grados de complejidad, generados a diferentes velocidades y diversos grados de ambigüedad, que no pueden procesarse utilizando tecnologías tradicionales, métodos de procesamiento, algoritmos o cualquier solución comercial lista para usar *(Krishnan, 2013).*

SARLAFT: (Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo) es un marco normativo y operativo implementado en Colombia para que las entidades vigiladas por la Superintendencia Financiera de Colombia (SFC) gestionen de manera efectiva los riesgos asociados al lavado de activos (LA) y la financiación del terrorismo. *(Superintendencia Financiera de Colombia, 2022).*

Segmentación: en Big Data es el proceso de dividir un conjunto masivo de datos en grupos homogéneos (clusters) basados en patrones, comportamientos o características similares, utilizando técnicas de aprendizaje automático (machine learning) y análisis estadístico. A diferencia de los

métodos tradicionales, en Big Data la segmentación se aplica a datos estructurados y no estructurados (ej.: transacciones financieras, interacciones digitales, geolocalización), permitiendo identificar subpoblaciones ocultas para la toma de decisiones estratégicas (Steinbach,2019).

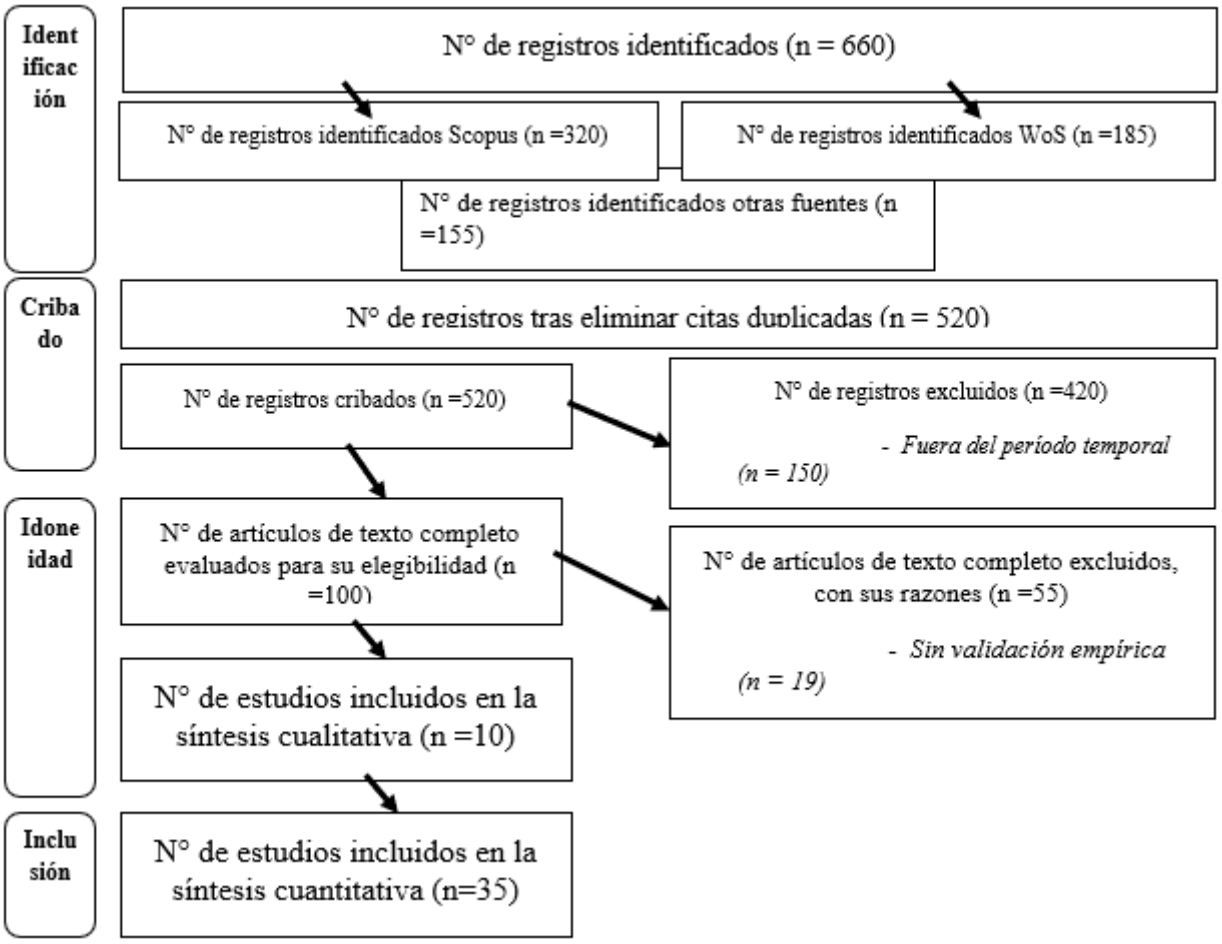
K-means: métodos de clustering o segmentación más utilizados en Big Data debido a su simplicidad computacional y escalabilidad en grandes volúmenes de datos (Jain, 2010).

CRISP-DM: (Cross-Industry Standard Process for Data Mining) es una metodología estándar para la gestión de proyectos de minería de datos. Fue desarrollada en 1996 por un consorcio de empresas (SPSS, Daimler, NCR y OHRA) y consta de seis fases: (1) comprensión del negocio, (2) comprensión de los datos, (3) preparación de los datos, (4) modelado, (5) evaluación y (6) despliegue (Chapman, 2000).

K-Medoids: es una técnica robusta de clustering, especialmente útil en segmentación de clientes cuando los datos contienen outliers o variables mixtas (numéricas y categóricas). A diferencia de K-Means —que usa promedios (centroides)—, K-Medoids selecciona objetos reales del conjunto de datos (medoides) como centros de clúster, lo que lo hace menos sensible a ruido y valores extremos (Kaufman & Rousseeuw, 1990).

Stradata Search: es una herramienta que proporciona la compañía Stradata AML experta en LA/FT. La herramienta que proporciona esta compañía, basa en una técnica de búsqueda en múltiples fuentes de información de forma simultánea a través de una única interfaz de consulta. (Han et al., 2018, p. 771).

Ilustración 1: Flujograma PRISMA



La revisión bibliográfica realizada se hizo utilizando un flujograma PRISMA, el cual es un diagrama estandarizado que se utiliza en revisiones sistemáticas de literatura para documentar de manera transparente y reproducible el proceso de selección de estudios académicos. En este sentido, las palabras de búsqueda seleccionadas fueron:

Español:

"cooperativas financieras Colombia", "SARLAFT", "gestión de riesgos LAFT", "segmentación de clientes", "big data sector financiero", "inclusión financiera", "normativa SARLAFT", "lavado de activos en cooperativas", "sistemas de alertas financieras".

Inglés:

"Financial cooperatives Colombia", "anti-money laundering", "terrorist financing", "risk segmentation", "machine learning financial compliance", "regulatory compliance Colombia".

Adicionalmente, se tuvieron en cuenta cadenas o ecuaciones de búsqueda, utilizando operadores booleanos y truncamientos como los presentados a continuación:

- ("cooperativas financieras" OR "*financial cooperatives*") AND (Colombia) AND ("SARLAFT" OR "LAFT") AND ("*big data*" OR "*machine learning*").
- *("risk segmentation" OR "client classification") AND ("anti-money laundering" OR "terrorist financing") AND ("financial sector" OR "cooperatives").*
- *("financial inclusion" OR "rural finance") AND (Colombia) AND ("cooperatives" OR "SARLAFT").*
- *("regulatory compliance" OR "AML/CFT") AND ("financial institutions" OR "cooperatives") AND ("data analytics").*
- ("segmentación de clientes" OR "clustering") AND ("SARLAFT" OR "lavado de activos" OR "LA/FT") AND ("cooperativas financieras")
- ("big data" OR "aprendizaje automático") AND ("gestión de riesgo" OR "cumplimiento") AND ("sector cooperativo")
- ("análisis transaccional" OR "scoring de riesgo") AND ("aprendizaje no supervisado" OR "detección de patrones")

En este orden de ideas, fueron empleadas bases de datos internacionales y locales para realizar las búsquedas. Las bases internacionales fueron: *Scopus, Web of Science*, *ScienceDirect*, *IEEE Xplore*, *SpringerLink, JSTOR* y Revistas. Para el caso de las bases locales se utilizaron: Redalyc, SciELO Colombia, Dialnet, Revistas, COLCoop, Confederación de Cooperativas de Colombia, WOCCU y Repositorios institucionales (ej: Universidad Nacional de Colombia, Universidad de los Andes). Además de lo anterior, también se usaron repositorios (*Google Scholar, ResearchGate, SSRN*) y Normativa de instituciones como: Superintendencia Financiera de Colombia (SFC), Confecoop, Unidad de Información y Análisis Financiero (UIAF), DIAN, GAFILAT, FATF, Asobancaria.

El período de búsqueda para la revisión bibliográfica corresponde a 2016–2024 (para incluir avances recientes en normativas y tecnologías). Para el caso de documentos normativos, no se tuvo restricción temporal (se priorizaron versiones vigentes, como la Circular SARLAFT 2020).

Cómo criterios de inclusión podemos relacionar:

- Estudio español o inglés.
- Enfoque en gestión de riesgos LAFT, segmentación de clientes o uso de tecnologías (big data, ML) en cooperativas.
- Aplicaciones en sector financiero, especialmente cooperativas

- Métodos cuantitativos (machine learning, clustering, análisis de riesgo)
- Casos de estudio en Latinoamérica (énfasis en Colombia)
- Publicaciones revisadas por pares, informes institucionales o normativas oficiales.
- Publicados entre 2016–2024 (excepto documentos normativos).

Cómo criterios de exclusión para las búsquedas se definieron:

- Estudios fuera del sector financiero o cooperativo.
- Artículos anteriores a 2016 (excepto referencias normativas clave).
- Investigaciones sin validación empírica
- Contextos no financieros (retail, seguros, etc.)
- Soluciones teóricas sin implementación real
- Tecnologías no escalables para entidades medianas
- Fuentes no académicas (blogs, medios no especializados).
- Investigaciones sin aplicación práctica al SARLAFT o sin metodologías replicables.

Documentos sin acceso abierto o de pago no disponible.

5. Antecedentes

Para la elaboración de este trabajo, se buscó establecer un marco que sirviera como guía para desarrollar e identificar los principales aspectos técnicos y teóricos a tener en cuenta, ya sea por los avances en investigaciones previas, recomendaciones de instituciones técnicas o normatividad aplicable a la investigación. En este sentido, se comenzó identificando los principales modelos utilizados en la estructuración de un trabajo de segmentación donde se busca identificar una población de clientes a partir del análisis y entendimiento de los datos en línea con el marco colombiano definido para la gestión integral de riesgo asociados con lavado de activos y financiación del terrorismo SARLAFT, encontrando que es recomendable aplicar la metodología CRISP-DM. Dicha metodología garantiza un adecuado entendimiento y preparación de los datos, pues exige que se tenga un conocimiento profundo del negocio que pueda soportar las diferentes conclusiones y análisis, obteniendo información suficiente para la interpretación de los resultados que se obtengan a partir de la aplicación del modelo. Por su parte, se encontró como el mejor modelo analítico para la segmentación de asociados en una cooperativa de ahorros y créditos al K-means, siendo el más utilizado en análisis o segmentación de clientes en este sector (Jovel Tamayo, 2020).

En concordancia con lo anterior, y según el estudio realizado por Correa y Montoya (2024) “Análisis de segmentación y alertamiento transaccional para la gestión de riesgos sarlaft en el sector financiero” es de vital importancia desarrollar la metodología CRISP-DM, dado que proporciona una guía para la ejecución de proyectos de minería de datos donde se pueda filtrar grandes volúmenes de registros con información de clientes para segmentarlos en conjuntos de que cumplan con características similares en transacciones y datos demográficos, dicha metodología propone desarrollar 6 etapas las cuales son: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelamiento, evaluación y despliegue. En este estudio, Cadavid López propone utilizar como técnica de modelado la K-Means y el método del codo o el índice de silueta para la determinación óptima del número de clusters.

Por su parte, Ramos (2023), propone un modelo de segmentación para SARLAFT en una entidad financiera con casa matriz española, desarrollando las etapas del CRISP-DM y la metodología SCRUM, de tal forma que se pueda utilizar lo mejor de los dos, obteniendo un conocimiento detallado de los datos y el negocio al desarrollar el CRISP-DM y permitiendo la flexibilidad, adaptabilidad, control de calidad y productividad que permite el SCRUM. Por otro parte, recomienda utilizar los algoritmos como: K-Medoids, Algoritmo de Agrupamiento Jerárquico y Fuzzy c-means debido a múltiples razones, entre ellas y la más importante, es que estos algoritmos

son capaces de manejar tanto datos dicotómicos como variables numéricas estandarizadas, lo cual es fundamental para el conjunto de datos que se analizan y requieren en este tipo de modelos, dado el tipo de variables que se utilizan para obtener los diferentes clusters, teniendo en cuenta el nivel de riesgo. Por tanto, las principales variables identificadas durante la revisión bibliográfica y estructuración del marco teórico, según la literatura y normativa analizada, son:

Jurisdicción (Ubicación geográfica):

La ubicación geográfica de los clientes es un factor crítico debido a las diferencias regionales en exposición al riesgo LA/FT. Según el DANE (2021), regiones con alta informalidad o conflicto requieren mayor atención. Ramos (2023) destaca que zonas con baja supervisión financiera presentan mayor riesgo, lo que exige ajustar controles basados en datos geoespaciales.

Actividad económica:

El tipo de actividad económica del cliente determina su perfil de riesgo. Correa y Montoya (2024) señalan que sectores como el agropecuario o el comercio informal tienen mayor probabilidad de operaciones inusuales. El SARLAFT (SFC, 2020) exige clasificar actividades según su propensión a LA/FT, utilizando códigos CIIU (Clasificación Industrial Internacional Uniforme).

Personas Expuestas Políticamente (PEP):

Identificar PEPs es obligatorio para mitigar riesgos de corrupción y lavado. Gómez (2019) enfatiza que las cooperativas deben cruzar datos con registros oficiales (UIAF) para detectar PEPs y sus redes asociadas. Jovel Tamayo (2020) sugiere incluir no solo a PEPs nacionales, sino también internacionales, siguiendo estándares del GAFILAT.

Composición patrimonial:

Analizar activos, pasivos y flujos de ingresos ayuda a detectar inconsistencias. Zhang et al. (2021) proponen modelos de clustering para identificar patrimonios desproporcionados respecto a la actividad declarada, un indicador clave de riesgo. La Circular SARLAFT (SFC, 2020) exige validar la legitimidad de los fondos mediante esta variable.

Información transaccional:

El volumen, frecuencia y destinos de las transacciones son esenciales. Pérez y González (2022) demuestran que algoritmos de detección de anomalías (ej.: *Isolation Forest*) optimizan la

identificación de patrones sospechosos. Ramos (2023) recomienda integrar datos de la UIAF para correlacionar transacciones con alertas preexistentes.

Por otro lado, fue necesario conocer los principales desafíos que se han documentado que podrían llegar a impactar un adecuado desarrollo del trabajo propuesto, encontrando que para la implementación de modelos de segmentación de clientes que cumplan con el SARLAFT en Colombia, la calidad y disponibilidad de los datos podrían significar un reto a sobrepasar, así como la ausencia de información clave, tal es el caso de datos financieros o actividad económica, puede afectar significativamente el monitoreo y seguimiento adecuado de las operaciones, comprometiendo la efectividad de la segmentación y la detección de operaciones sospechosas (Infolaft, 2021). La implementación de estándares como el ISO 25012 se ha recomendado para garantizar la integridad y calidad de los datos en estos procesos.

Además, la aplicación de metodologías estadísticas robustas es esencial para garantizar la homogeneidad dentro de los segmentos y la heterogeneidad entre ellos. La falta de técnicas adecuadas puede llevar a segmentaciones ineficaces, lo que ha sido motivo de sanciones por parte de la Superintendencia Financiera (Infolaft, 2018). El uso de algoritmos como K-Means y técnicas de análisis multivariado ha demostrado ser efectivo en la identificación de patrones de comportamiento y en la mejora de la segmentación (Ortiz, 2023).

La integración de herramientas tecnológicas avanzadas también representa un reto significativo. Muchas entidades financieras carecen de sistemas que permitan el análisis automatizado de grandes volúmenes de datos, lo que limita la capacidad de detectar operaciones inusuales en tiempo real (Correa & Montoya, 2024). La implementación de soluciones de Business Analytics puede facilitar la segmentación y fortalecer el sistema de administración de riesgos (Ramos, 2023).

Otro aspecto crítico es la necesidad de una metodología de segmentación claramente definida y aprobada por la alta dirección. La ausencia de una metodología formal puede impedir la identificación de señales de alerta y comprometer la prevención del lavado de activos (Infolaft, 2018). Es fundamental que las entidades establezcan procedimientos documentados y aprobados que cumplan con los requisitos normativos.

Finalmente, la capacitación del personal en técnicas de análisis de datos y en el uso de herramientas de segmentación automatizadas es crucial. La falta de conocimientos especializados

puede limitar la efectividad de los modelos implementados y aumentar el riesgo de incumplimiento normativo (Infolaft, 2023). La formación continua y la incorporación de profesionales en analítica son esenciales para cerrar esta brecha.

6. Objetivos

6.1 Objetivo General

Desarrollar un modelo de segmentación basado en técnicas de Big Data, que permita la identificación de los clientes de una cooperativa financiera de acuerdo con el nivel de riesgo en un marco SARLAFT, con el fin de conocer aquellos clientes que presenten un nivel de riesgo alto, y de esta manera obtener insumos que permitan el diseño de controles más específicos y eficaces por parte de los encargados de cumplimiento en la organización.

6.2 Objetivos Específicos

- Caracterizar las variables clave asociadas a cada factor de riesgo SARLAFT (ej.: perfil transaccional, datos sociodemográficos, vinculación a sectores de alto riesgo, actividad económica, entre otros), con el fin de garantizar que el modelo abarque las dimensiones críticas para la evaluación de riesgo LA/FT.

- Asignar valores de riesgo y ponderaciones a las variables seleccionadas, mediante métodos cuantitativos, con el objetivo de priorizar aquellas con mayor impacto en la clasificación del riesgo (alto, moderado y bajo).

- Implementar un modelo de segmentación basado en Big Data para una cooperativa financiera en Colombia para que con su uso se tomen decisiones más acertadas.

7. Viabilidad

Este proyecto se enfocó en desarrollar un modelo que permita segmentar una población de clientes (personas naturales y personas jurídicas) de una cooperativa financiera, respondiendo a las exigencias normativas del marco SARLAFT. Los recursos necesarios para cumplir con lo planteado en este proyecto, incluyen, entre otros, poder conformar un equipo de trabajo con personas expertas en cumplimiento y personas que puedan realizar análisis de información y aplicación de modelos de segmentación, además, se requirieron equipos de de 500 GB, 24 GB de RAM y un procesador de 12th generación Intel Core i 7. En los cuales se pudiera hacer una preparación previa de la data, para posteriormente utilizar las bases de datos en Google Colab y a través de librerías de Python para llevar a cabo la segmentación.

El alcance planteado, para la segmentación, es identificar a través de clusterclústeres la forma en que está conformada la población de clientes, obteniendo conjuntos que permitan reconocer las poblaciones con mayor o menor nivel de riesgo SARLAFT. Utilizando como punto de partida la constitución de una base de datos donde se pueda incluir para cada cliente información relacionada con las principales variables que, según el marco normativo, expertos de cumplimiento y experiencia del equipo deberían ser consideradas, con el objetivo final de definir acciones de control como debidas diligencias ampliadas y seguimiento periódico.

Sus implicaciones incluyeron constituir una base de datos partiendo del maestro de clientes, donde se utilizaron diferentes fuentes de información y conceptos de expertos en cumplimiento para determinar la mejor manera de incluir variables relacionadas con el SARLAFT, tal es el caso de, ubicación geográfica del cliente, actividad económica, nivel de activos, nivel de pasivos, entre otros. Logrando constituir un maestro de clientes para SARLAFT, mediante el cual se pueda realizar una segmentación que sirva de punto de partida y detonador de acciones de control en torno al cumplimiento de la normativa aplicable.

Con la implementación de este proyecto se logra un gran avance en cuanto al cumplimiento normativo SARLAFT, cubriendo una de las principales exigencias de la norma y sirviendo de punto de partida para el fortalecimiento de las actividades de control, logrando mayor estabilidad, además de la mitigación de riesgos estratégicos como la continuidad de negocio. Adicionalmente, este proyecto permitió al área de cumplimiento de la Cooperativa dar un salto hacia la mejora del proceso que tienen a cargo, dilucidando el posible camino que se pueda comenzar a tomar para el control de clientes y mayor alineación con la norma, cubriendo brechas que actualmente se tienen y disminuyendo la cantidad de reprocesos y riesgos en torno a la atención de visitas y requerimientos de entes de control.

8. Metodología

Tabla 1: Plan de trabajo para la implementación del modelo de segmentación SARLAFT

Objetivos Específicos	Actividad a desarrollar	Entregable o soporte	Fase CRISP - DM
Caracterizar las variables clave asociadas a cada factor de riesgo SARLAFT (ej.: perfil transaccional, datos sociodemográficos, vinculación a sectores de alto riesgo, actividad económica, entre otros), con el fin de garantizar que el modelo abarque las dimensiones críticas para la evaluación de riesgo LA/FT.	- Realizar un reconocimiento de las principales actividades a desarrollar en el área de cumplimiento de la Cooperativa, con el objetivo de entender la dinámica del negocio y las expectativas que se tienen frente a las necesidades del área. - Realizar un reconocimiento de la normatividad aplicable al negocio, tratando de identificar cuáles son los requisitos que se deberán cubrir con el modelo a implementar. - Descargar la base de datos de clientes (maestro de clientes) y realizar un reconocimiento de los datos que se tienen, buscando identificar las variables que se podrán utilizar en el desarrollo del modelo, garantizando que respondan a la normatividad previamente documentada y expectativas del área. - Depurar la base de datos, identificando variables a utilizar, registros incompletos, registros duplicados,	- Este objetivo tendrá como resultado una base de datos compuesta por las principales variables para la segmentación. Dichas variables serán concertadas con personas expertas en cumplimiento de tal manera que se pueda comenzar a tener una categorización que cumpla con las expectativas del negocio y sirva de insumos para el diseño de controles e identificación de población que será susceptible de debidas diligencias ampliadas.	- Fase I: Contexto del negocio. - Fase II: Identificación de los datos. - Fase III: Preparación de los datos.

	dispersión de los datos, entre otros, con el objetivo de tener una data óptima para la aplicación del modelo. - Realizar ajuste o categorización de las variables teniendo en cuenta aspectos como: clasificación del tipo de actividad económica, ubicación geográfica e identificación PEP´s, utilizando data de la DIAN, UIAF, listas restrictivas, bases de datos públicas, entre otras. Validar esta información con personal experto en cumplimiento.		
Asignar valores de riesgo y ponderaciones a las variables seleccionadas, mediante métodos cuantitativos, con el objetivo de priorizar aquellas con mayor impacto en la clasificación del riesgo (alto, moderado y bajo).	- Realizar una asignación de valores a cada una de las variables, con el objetivo de categorizarlas, de tal manera que se pueda tener una ponderación acorde con el negocio y la normatividad. Para esta etapa se requiere una definición del negocio y expertos de cumplimiento, de tal forma que se pueda asignar mayor peso a las variables que cobren más relevancia para el análisis.	- Como resultado del desarrollo de este objetivo se tendrá un documento que soporte la asignación de los pesos de cada variable dentro del modelo, de tal manera que al final se pueda tener como resultado un cálculo del nivel de riesgo de cada cliente en categorías de muy alto, alto, medio y bajo.	- Fase III: Preparación de los datos. - Fase IV: Modelado
Implementar un modelo de segmentación basado en Big Data para una cooperativa financiera en Colombia para que con su uso se	- Diseñar un modelo no supervisado con la base de datos de clientes de la Cooperativa, utilizando Python y aplicando el método K-means para su	- Como resultado se tendrán cluster de la base de datos de clientes, alineados con el negocio para la identificación de segmentos donde se	- Fase IV: Modelado - Fase V: Evaluación

tomen decisiones más acertadas.	programación y prueba. Dicho modelo tendrá en cuenta cada una de las definiciones anteriores. - Identificar los segmentos derivados de la ejecución del modelo y nombrarlos, describirlos y validarlos con el experto en cumplimiento de la Cooperativa, buscando identificar posibles mejoras, ajuste del número de cluster, entre otros.	requiere mayor control y aplicación de debidas diligencias ampliadas, de tal manera que se tenga una mejor distribución de los recursos de aseguramiento y se puedan tomar decisiones más acertadas en términos de compliance	- Fase VI: Despliegue

Para abordar el problema de segmentación de clientes en una cooperativa financiera bajo el marco SARLAFT, se adoptará la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) debido a su enfoque estructurado y su amplia aceptación en proyectos de Big Data y minería de datos. Esta metodología consta de seis fases iterativas que permiten desarrollar un modelo robusto y reproducible. A continuación, se detalla cada una de estas fases y la forma en que serán abordadas;

8.1 Comprensión del Negocio

En esta fase se tendrá por objetivo alinear el proyecto con los requerimientos del SARLAFT y las necesidades de la cooperativa financiera, por lo cual, cobra especial relevancia el reconocimiento de la normatividad aplicable, así como las expectativas del negocio, más exactamente el área de cumplimiento de la Cooperativa, por tanto, se llevarán a cabo sesiones de entendimiento, donde se puedan reconocer las actividades que se realizan en el área y los momentos en los cuales se pretende utilizar la información que resulte de la ejecución del modelo, con el fin de dar mayor claridad frente al alcance del proyecto que se pretende implementar, buscando que desde ambas partes se tenga mayor confianza frente a los resultados esperados y la utilidad que pueda tener el modelo.

8.2 Comprensión de los Datos

En esta fase se busca identificar las fuentes de datos relevantes y evaluar su calidad frente a las necesidades del modelo, haciendo un reconocimiento inicial de la base de datos o maestro de clientes con que cuenta la Cooperativa, partiendo del reconocimiento de variables, cantidad de registros,

manipulación de la data y limitaciones para su extracción.

8.3 Preparación de los Datos

Para esta fase ya se tiene un conocimiento del negocio, claridad frente a las necesidad y expectativas, por tanto, se buscará realizar una limpieza, transformación y enriquecimiento de los datos para el modelado, partiendo de la depuración de las variables que no será útiles para la segmentación, así como la identificación de registros duplicados, registros nulos, nivel de dispersión, entre otras características que podrían afectar el desempeño del modelo y los resultados obtenidos. Se realizará un análisis estadístico de la data a través de la aplicación de comandos en Python que permitan hacer un entendimiento general de la composición de los datos de cada variable.

Una vez desarrolladas las actividades descritas anteriormente, se comenzará a identificar información y bases de datos de instituciones como la DIAN, UIAF, listas restrictivas y bases de datos públicas de PEP´s para comenzar a enriquecer las variables seleccionadas, asignando categorías de acuerdo al nivel de riesgo a las actividades económicas y ubicación geográfica de las sucursales donde tienen operación los clientes, de tal manera que se puedan ajustar las variables de acuerdo con recomendaciones de expertos en cumplimiento y referentes identificados en la revisión de bibliografía.

8.4 Modelado

Para el desarrollo del modelo de segmentación basado en Big Data, se pretende aplicar el método K-means, utilizando Python en un ambiente como el de Google Colab. En esta etapa se definirán cada una de las variables de forma conjunta con los expertos del negocio, buscando tener claridad frente a las variables que cobran mayor relevancia para la Cooperativa y poder tener en cuenta esta información en los análisis, de tal manera que se pueda tener una calificación del nivel de riesgo para cada cliente ubicado en un cluster específico. Una vez generados los clusters, se harán descripciones de la composición de cada uno y se darán algunas recomendaciones en torno a las mejoras en los controles que se puedan implementar de acuerdo con los datos generados.

8.5 Evaluación

Teniendo en cuenta los resultados generados, se procederá a validar el modelo con los expertos de cumplimiento (oficial de cumplimiento), de tal manera que se pueda ajustar según solicitudes puntuales y que estén dentro del alcance. Discutir sobre las recomendaciones y mejoras que se puedan implementar y concluir frente al modelo.

8.6 Despliegue

Validar las facilidades de implementación del modelo para la actualización anual de clientes, de tal forma que se pueda hacer una permanente revisión de los clientes e incluir este tipo de validación como un control o fase dentro del sistema SARLAFT que se tenga definido para la Cooperativa.

9. Resultados

En este punto se detalla la forma como fue desarrollado cada uno de los objetivos específicos planteados para el proyecto. Comenzando con la caracterización y preparación de la base de datos, buscando que se diera cumplimiento a las necesidades del negocio y exigencias normativas. Luego, se procedió a realizar la identificación de valores de riesgo para las variables de actividad económica y ubicación geográfica, en conjunto de los expertos en cumplimiento de la compañía, para poder contar con una base de datos que cumpliera las condiciones necesarias para la segmentación, una vez se lograra realizar una adecuada limpieza, análisis, separación de la información y tratamiento de variables categóricas, logrando con estas actividades cumplir con los objetivos 1 y 2. Por último, se detalla cómo se cubre el objetivo 3 con la implementación del modelo de segmentación, cumpliendo con una previa definición de los clusters y ejecución del K-means.

9.1 Objetivo 1: Caracterización de variables – Preparación de la data

Para el cumplimiento de este objetivo, se partió de la descarga de la base de datos de clientes de la Cooperativa, la cual fue entregada en archivo de Excel por la analista de cumplimiento. Dicha base se encontraba conformada por 19 variables, las cuales se describen a continuación:

Tabla 2: Variables iniciales del dataset de clientes para segmentación SARLAFT

TIPO_DE_CODIGO	Tipo de documento de identidad (pasaporte, cédula, NIT)
TIPO	Tipo de persona (Natural o Jurídica)
ACTIVIDAD_CODIGO	Código de actividad interno del negocio.
ACTIVO	Monto de los activos reportado, según información en el momento de la vinculación.
PASIVO	Monto de los pasivos reportado, según información en el momento de la vinculación.
PATRIMONIO	Monto del patrimonio reportado, según información en el momento de la vinculación.
INGRESOS	Monto de ingresos reportados y registrados al momento de la vinculación.
EGRESOS	Monto de los gastos reportados y registrados al momento de la vinculación.
FECHA_VINCULACION	Fecha de vinculación del cliente.
TIEMPO	Tiempo que ha transcurrido desde la fecha de vinculación.
CIIU_CODIGO	Código de actividad económica
SUCURSAL_CODIGO	Código o identificador de la ubicación de la sucursal
TOTAL_OTROS_INGRESOS	Monto de otro ingresos reportados y registrados al momento de la vinculación
FECHA_NACIMIENTO	Fecha de nacimiento del cliente.
EDAD	Edad del cliente
Montocre	Monto del crédito del cliente.
Movicre	Movimiento de crédito del cliente.
Montodeb	Monto del débito del cliente.
Movideb	Movimiento de débito del cliente.

Cabe mencionar que la base se encontraba conformada por 43713 registros, de los cuales 42663 eran personas naturales y 1050 eran personas jurídicas.

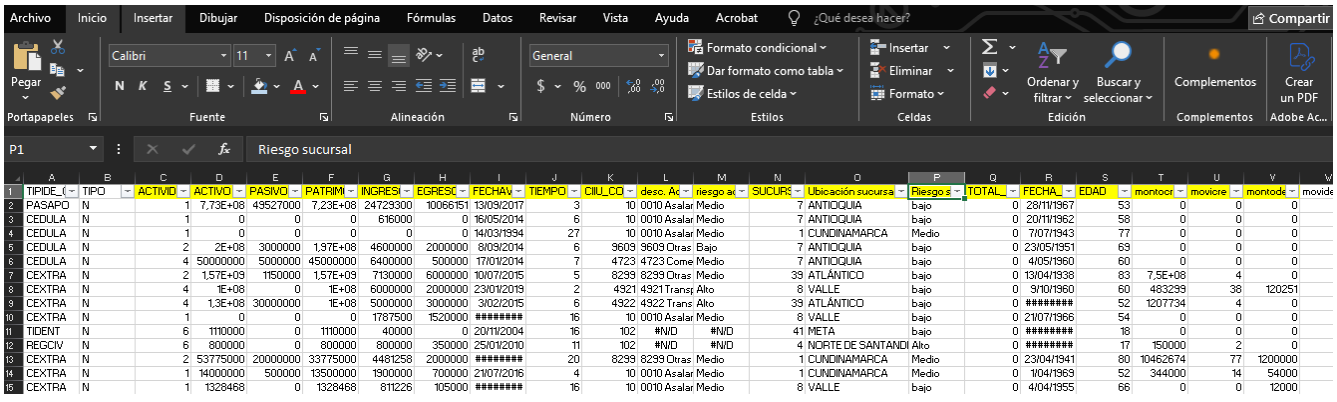
Una vez se tuvo la base, se procedieron a realizar reuniones de entendimiento y definición de variables que hicieran falta para lograr tener la base de datos que más se ajustara a lo que se exigía desde el marco normativo que regula el proceso. Durante este ejercicio se acordó que, si bien no se requerían variables adicionales, era necesario incluir la descripción de la actividad económica según la DIAN, con el objetivo de tener mayor claridad del sector y tipo de actividad que desarrolla el cliente, de tal modo que se pudiera establecer un nivel de riesgo, según percepción de expertos del área cumplimiento y experiencia profesional de las personas a cargo. Durante este proceso, se utilizaron datos proporcionados por el área de cumplimiento, partiendo de papeles de trabajo construidos por el analista, los cuales se utilizaron para cruzar con los códigos relacionados con el maestro de clientes y agregar las columnas “descripción actividad económica” y “nivel de riesgo actividad económica” a la base.

Posteriormente, se utilizó la información de la cooperativa para identificar de forma clara el departamento o ubicación de la sucursal, lo cual, podría darnos información acerca de la ubicación del cliente, para lo cual se hizo un cruce entre el código de la sucursal y la base de datos de sucursales que la cooperativa. Una vez identificados los departamento o ubicaciones, se procedió a establecer, según conceptos de expertos y reportes del DANE, aquellos sectores donde se generaban mayor índice de delitos asociados con SARLAFT y de esta manera establecer una clasificación de nivel de riesgo para las diferentes sucursales o ubicaciones, resultando en dos columnas más para la base de datos, “ubicación” y “nivel de riesgo ubicación”.

Por último, se procedió a validar la base para garantizar que se estaban incluyendo todas las variables que podrían ser requeridas para una segmentación en cumplimiento del marco normativo y las necesidades del negocio. Dicho análisis contempló aspectos como listas restrictivas y listas PEPS, sin embargo, se decidió que estas validaciones se debían realizar de forma masiva utilizando Stradata Search, una vez se tenga una adecuada depuración de los registros en la maestra de clientes.

Base de datos una vez agregadas las columnas de ubicación, riesgo ubicación, descripción de actividad económica y riesgo de actividad económica:

Ilustración 2: Base de datos inicial de clientes antes del procesamiento



	TIPIDE_(	TIPO	ACTIVID	ACTIVO	PASIVO	PATRIMI	INGRESO	EGRESO	FECHAV	TIEMPO	CIIU_CO	desc. Ac	riesgo a	SUCURS	Ubicación sucursa	Riesgo s	TOTAL_	FECHA_	EDAD	montocr	movicre	montode	movide
2	PASAPO	N	1	7,73E+08	49527000	7,23E+08	24729300	10066151	13/09/2017	3	10	0010 Asalar	Medio	7	ANTIOQUIA	bajo	0	28/11/1967	53	0	0	0	
3	CEDULA	N	1	0	0	0	616000	0	16/05/2014	6	10	0010 Asalar	Medio	7	ANTIOQUIA	bajo	0	20/11/1962	58	0	0	0	
4	CEDULA	N	1	0	0	0	0	0	14/03/1934	27	10	0010 Asalar	Medio	1	CUNDINAMARCA	Medio	0	7/07/1943	77	0	0	0	
5	CEDULA	N	2	2E+08	3000000	1,97E+08	4600000	2000000	8/09/2014	6	9603	9603 Otras	Bajo	7	ANTIOQUIA	bajo	0	23/05/1951	69	0	0	0	
6	CEDULA	N	4	50000000	5000000	45000000	6400000	500000	17/01/2014	7	4723	4723 Come	Medio	7	ANTIOQUIA	bajo	0	4/05/1960	60	0	0	0	
7	CEXTRA	N	2	1,57E+09	1150000	1,57E+09	7130000	6000000	10/07/2015	5	8299	8299 Otras	Medio	39	ATLÁNTICO	bajo	0	13/04/1938	83	7,5E+08	4	0	
8	CEXTRA	N	4	1E+08	0	1E+08	6000000	2000000	23/01/2019	2	4921	4921 Transp	Alto	8	VALLE	bajo	0	9/10/1960	60	483299	38	120251	
9	CEXTRA	N	4	1,3E+08	30000000	1E+08	5000000	3000000	3/02/2015	6	4922	4922 Trans	Alto	39	ATLÁNTICO	bajo	0	########	52	1207734	4	0	
10	CEXTRA	N	1	0	0	0	1787500	1520000	########	16	10	0010 Asalar	Medio	8	VALLE	bajo	0	21/07/1966	54	0	0	0	
11	TIDENT	N	6	1110000	0	1110000	40000	0	20/11/2004	16	102	#N/D	#N/D	41	META	bajo	0	########	18	0	0	0	
12	REGCIV	N	6	800000	0	800000	800000	350000	25/01/2010	11	102	#N/D	#N/D	4	NORTE DE SANTANDI	Alto	0	########	17	150000	2	0	
13	CEXTRA	N	2	53775000	20000000	33775000	4481258	2000000	########	20	8299	8299 Otras	Medio	1	CUNDINAMARCA	Medio	0	23/04/1941	80	10462674	77	1200000	
14	CEXTRA	N	1	14000000	500000	13500000	1900000	700000	21/07/2016	4	10	0010 Asalar	Medio	1	CUNDINAMARCA	Medio	0	1/04/1969	52	344000	14	54000	
15	CEXTRA	N	1	1328468	0	1328468	811226	105000	########	16	10	0010 Asalar	Medio	8	VALLE	bajo	0	4/04/1955	66	0	0	12000	

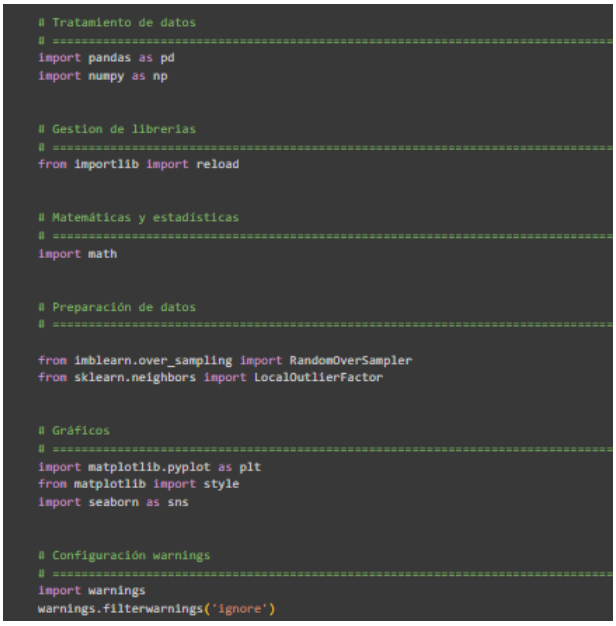
Base de datos con variables completas – antes del procesamiento de datos.

Cubierta esta etapa, se procedió a crear un proyecto en Google Colab, para realizar la limpieza y tratamiento de los datos. Desarrollando las siguientes Fases:

- **Librerías, herramientas y funciones:**

Se realizar la importación de las librerías necesarias para el procesamiento de la información e implementación del modelo:

Ilustración 3: Importación de librerías en Python para el procesamiento de datos



```
# Tratamiento de datos
# ==============================================================================
import pandas as pd
import numpy as np


# Gestion de librerias
# ==============================================================================
from importlib import reload


# Matemáticas y estadísticas
# ==============================================================================
import math


# Preparación de datos
# ==============================================================================

from imblearn.over_sampling import RandomOverSampler
from sklearn.neighbors import LocalOutlierFactor


# Gráficos
# ==============================================================================
import matplotlib.pyplot as plt
from matplotlib import style
import seaborn as sns


# Configuración warnings
# ==============================================================================
import warnings
warnings.filterwarnings('ignore')
```

El script utilizado para las librerías es:

- import pandas as pd
- import numpy as np
- import math

Además, se realiza la importación de las siguientes herramientas o funcionalidades:

- from imblearn.over_sampling import RandomOverSampler
- from sklearn neighbors import LocalOutlierFactor
- *import matplotlib.pyplot as plt*
- *from matplotlib import style*
- *import seaborn as sns*
- *import warnings*
- *warnings.filterwarnings('ignore')*

- *from utils.funciones import multiple_plot*

- **Carga del dataset**

Ilustración 4: Carga inicial del dataset de clientes

```
#Cargar el dataset
d=pd.read_excel('/content/datasets/DB_Clientes_TrabajoGrados.xlsx')

#Cargar el dataset de Acticidades Economicas
Actividades=pd.read_excel('/content/datasets/Actividades_Economicas.xlsx')
```

Se realiza descripción de información inicial del dataset utilizando d.info():

Ilustración 5: Descripción de la estructura del dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43713 entries, 0 to 43712
Data columns (total 23 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   TIPIDE_CODIGO         43713 non-null  object
 1   TIPO                  43713 non-null  object
 2   ACTIVIDAD_CODIGO      42663 non-null  float64
 3   ACTIVO                43687 non-null  float64
 4   PASIVO                43686 non-null  float64
 5   PATRIMONIO            43691 non-null  float64
 6   INGRESOS              43713 non-null  float64
 7   EGRESOS               43212 non-null  float64
 8   FECHAVINCULACION      40770 non-null  datetime64[ns]
 9   TIEMPO                43713 non-null  float64
 10  CIIU_CODIGO           43713 non-null  int64
 11  desc. Actividades     37527 non-null  object
 12  riesgo actividades    37527 non-null  object
 13  SUCURSAL_CODIGO       43713 non-null  int64
 14  Ubicación sucursal    43713 non-null  object
 15  Riesgo sucursal       43713 non-null  object
 16  TOTAL_OTROS_INGRESOS  43458 non-null  float64
 17  FECHA_NACIMIENTO      43639 non-null  object
 18  EDAD                  43712 non-null  float64
 19  montocre              43713 non-null  int64
 20  movicre               43713 non-null  int64
 21  montodeb              43713 non-null  float64
 22  movideb               43713 non-null  int64
dtypes: datetime64[ns](1), float64(10), int64(5), object(7)
memory usage: 7.7+ MB
```

Se realiza la carga de la base de datos donde se realizó el análisis de los sectores económicos y actividades económicas, con el objetivo de agregar la columna sector económico a la base de datos, dado que, en principio al correr el modelo, resultado muchas actividades económicas (cerca de 700) lo que no permitía realizar un adecuado análisis de los resultados, además generaba mucha confusión entre las personas que querían realizar un entendimiento general desde las gráficas descriptivas.

Ilustración 6: Integración de datos de sectores económicos

```
# Leer la hoja 'Hoja5' de Actividades_Economicas.xlsx
df_hoja5 = pd.read_excel('/content/datasets/Actividades_Economicas.xlsx', sheet_name='Hoja5')

# Mostrar la información de los datos cargados
df_hoja5.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 504 entries, 0 to 503
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   CIIU             504 non-null    int64 
 1   Unnamed: 1       504 non-null    object
 2   Sector           504 non-null    object
 3   Nivel de Riesgo  504 non-null    object
 4   Comentario       504 non-null    object
dtypes: int64(1), object(4)
memory usage: 19.8+ KB
```

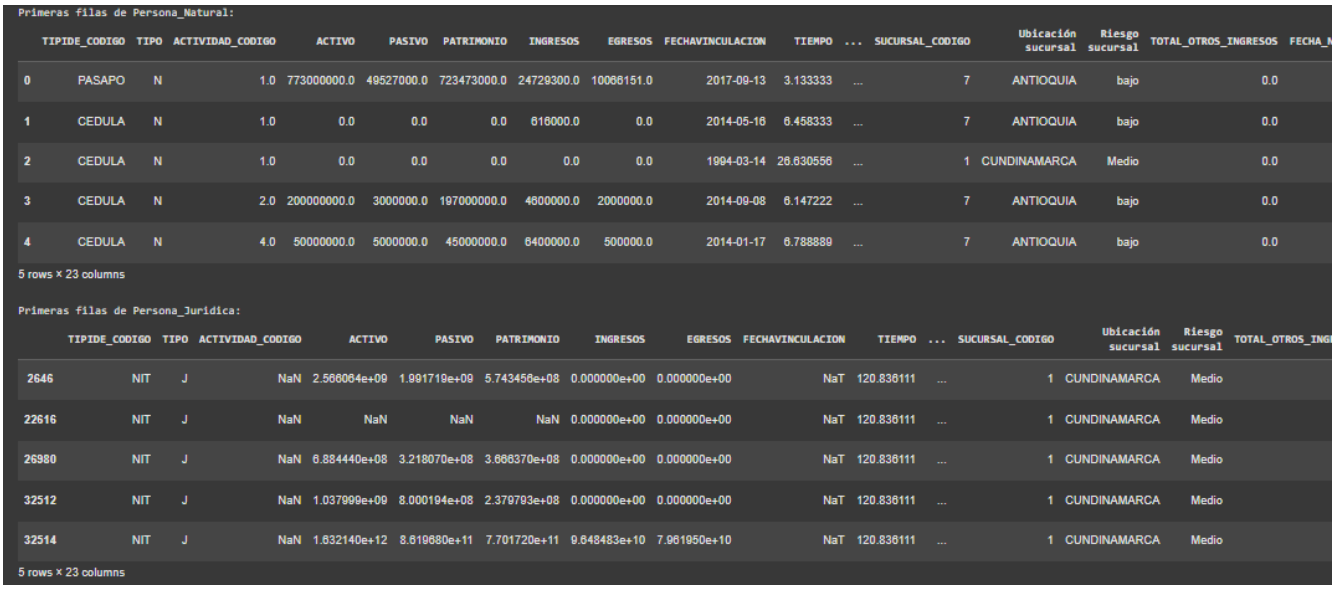
Una vez cargadas las bases de datos, se toma la decisión de hacer una división de la información entre tipos de personas, entendiendo que los análisis debían hacerse partiendo de esta clasificación, pues no se podría determinar un nivel de riesgo en términos equivalentes a una persona natural con una persona jurídica, dadas las variables utilizadas, tal es el caso de activos, pasivos o nivel de ingresos:

Ilustración 7: División de dataset entre personas naturales y jurídicas

```
# Dividir la base de datos en dos datasets separados por Tipo (Personas Naturales y Personas Jurídicas)
Persona_Natural = d[d['TIPO'] == 'N'].copy()
Persona_Juridica = d[d['TIPO'] == 'J'].copy()

# Mostrar las primeras filas de cada DataFrame para verificar la división
print("Primeras filas de Persona_Natural:")
display(Persona_Natural.head())

print("\nPrimeras filas de Persona_Juridica:")
display(Persona_Juridica.head())
```

Se realiza la división del dataset, generando una base para personas naturales y otra para personas jurídicas, de tal manera que se pueda realizar un mejor análisis y tratamiento de los datos.

Ilustración 8: Continuación de división de dataset



Primeras filas de Persona_Natural:

	TIPIDE_CODIGO	TIPO	ACTIVIDAD_CODIGO	ACTIVO	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	FECHAVINCULACION	TIEMPO	...	SUCURSAL_CODIGO	Ubicación sucursal	Riesgo sucursal	TOTAL_OTROS_INGRESOS	FECHA_N
0	PASAPO	N	1.0	773000000.0	49527000.0	723473000.0	24729300.0	10066151.0	2017-09-13	3.133333	...	7	ANTIOQUIA	bajo	0.0	
1	CEDULA	N	1.0	0.0	0.0	0.0	616000.0	0.0	2014-05-16	6.458333	...	7	ANTIOQUIA	bajo	0.0	
2	CEDULA	N	1.0	0.0	0.0	0.0	0.0	0.0	1994-03-14	26.630556	...	1	CUNDINAMARCA	Medio	0.0	
3	CEDULA	N	2.0	200000000.0	3000000.0	197000000.0	4600000.0	2000000.0	2014-09-08	6.147222	...	7	ANTIOQUIA	bajo	0.0	
4	CEDULA	N	4.0	50000000.0	5000000.0	45000000.0	6400000.0	500000.0	2014-01-17	6.788889	...	7	ANTIOQUIA	bajo	0.0	

5 rows × 23 columns

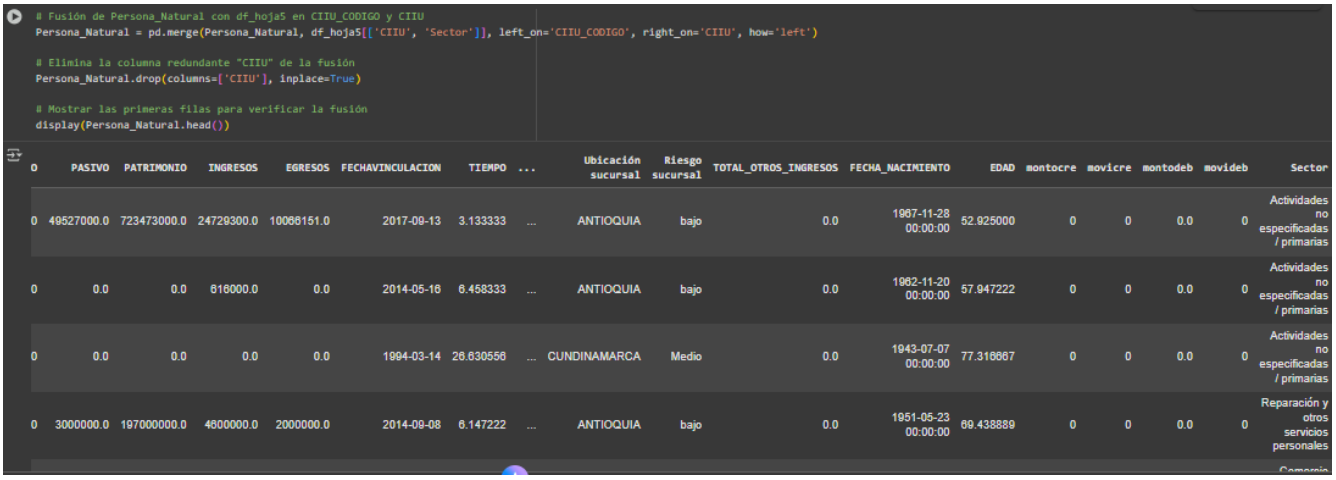
Primeras filas de Persona_Juridica:

	TIPIDE_CODIGO	TIPO	ACTIVIDAD_CODIGO	ACTIVO	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	FECHAVINCULACION	TIEMPO	...	SUCURSAL_CODIGO	Ubicación sucursal	Riesgo sucursal	TOTAL_OTROS_ING
2646	NIT	J	NaN	2.566064e+09	1.991719e+09	5.743456e+08	0.000000e+00	0.000000e+00	NaT	120.836111	...	1	CUNDINAMARCA	Medio	
22616	NIT	J	NaN	NaN	NaN	NaN	0.000000e+00	0.000000e+00	NaT	120.836111	...	1	CUNDINAMARCA	Medio	
26980	NIT	J	NaN	6.884440e+08	3.218070e+08	3.666370e+08	0.000000e+00	0.000000e+00	NaT	120.836111	...	1	CUNDINAMARCA	Medio	
32512	NIT	J	NaN	1.037999e+09	8.000194e+08	2.379793e+08	0.000000e+00	0.000000e+00	NaT	120.836111	...	1	CUNDINAMARCA	Medio	
32514	NIT	J	NaN	1.632140e+12	8.619680e+11	7.701720e+11	9.648483e+10	7.961950e+10	NaT	120.836111	...	1	CUNDINAMARCA	Medio	

5 rows × 23 columns

Luego de hacer la división de las bases de datos, se procede con la integración de la columna sector a la base de datos de clientes, tanto para personas naturales como para personas jurídicas. Lo anterior, con el objetivo de poder realizar un análisis mucho más agregado en términos de actividades económicas:

Ilustración 9: Integración de columna de sector económico



```
# Fusión de Persona_Natural con df_hoja5 en CIIU_CODIGO y CIIU
Persona_Natural = pd.merge(Persona_Natural, df_hoja5[['CIIU', 'Sector']], left_on='CIIU_CODIGO', right_on='CIIU', how='left')

# Elimina la columna redundante "CIIU" de la fusión
Persona_Natural.drop(columns=['CIIU'], inplace=True)

# Mostrar las primeras filas para verificar la fusión
display(Persona_Natural.head())
```

0	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	FECHAVINCULACION	TIEMPO	...	Ubicación sucursal	Riesgo sucursal	TOTAL_OTROS_INGRESOS	FECHA_NACIMIENTO	EDAD	montocre	movicre	montodeb	movideb	Sector
0	49527000.0	723473000.0	24729300.0	10066151.0	2017-09-13	3.133333	...	ANTIOQUIA	bajo	0.0	1967-11-28 00:00:00	52.925000	0	0	0.0	0	Actividades no especificadas / primarias
0	0.0	0.0	616000.0	0.0	2014-05-16	6.458333	...	ANTIOQUIA	bajo	0.0	1962-11-20 00:00:00	57.947222	0	0	0.0	0	Actividades no especificadas / primarias
0	0.0	0.0	0.0	0.0	1994-03-14	26.630556	...	CUNDINAMARCA	Medio	0.0	1943-07-07 00:00:00	77.316667	0	0	0.0	0	Actividades no especificadas / primarias
0	3000000.0	197000000.0	4600000.0	2000000.0	2014-09-08	6.147222	...	ANTIOQUIA	bajo	0.0	1951-05-23 00:00:00	69.438889	0	0	0.0	0	Reparación y otros servicios personales

Se agrega la columna del sector económico para personas naturales, con el propósito de tener una variable de suma importancia en la base de datos que será objeto de análisis. Dicha columna es una de las variables más importantes dado el marco normativo y la justificación presentada.

Ilustración 10: Continuación de integración de sector económico

```
# Combinar Persona_Juridica con df_hoja5 en CIIU_CODIGO y CIIU
Persona_Juridica = pd.merge(Persona_Juridica, df_hoja5[['CIIU', 'Sector']], left_on='CIIU_CODIGO', right_on='CIIU', how='left')

# Elimina la columna redundante "CIIU" de la fusión
Persona_Juridica.drop(columns=['CIIU'], inplace=True)

# Mostrar las primeras filas para verificar la fusión
display(Persona_Juridica.head())
```

PATRIMONIO	INGRESOS	EGRESOS	FECHAVINCULACION	TIEMPO	...	Ubicación sucursal	Riesgo sucursal	TOTAL_OTROS_INGRESOS	FECHA_NACIMIENTO	EDAD	montocre	movicre	montodeb	movideb	Sector
5.743456e+08	0.000000e+00	0.000000e+00	NaT	120.836111	...	CUNDINAMARCA	Medio	0.0	1996-01-05 00:00:00	24.822222	0	0	0.0	0	Organizaciones y asociaciones
NaN	0.000000e+00	0.000000e+00	NaT	120.836111	...	CUNDINAMARCA	Medio	0.0	1994-01-21 00:00:00	26.777778	0	0	0.0	0	Metalurgia y metalmecánica
3.666370e+08	0.000000e+00	0.000000e+00	NaT	120.836111	...	CUNDINAMARCA	Medio	0.0	2002-11-05 00:00:00	17.988889	32223370	40	49790868.0	30	Agropecuario (cultivos/ganadería)
2.379793e+08	0.000000e+00	0.000000e+00	NaT	120.836111	...	CUNDINAMARCA	Medio	0.0	2000-06-25 00:00:00	20.350000	0	0	0.0	0	Equipo eléctrico y maquinaria
7.701720e+11	9.648483e+10	7.961950e+10	NaT	120.836111	...	CUNDINAMARCA	Medio	NaN	2019-03-29 00:00:00	1.588889	0	0	0.0	0	Otros

- **Limpieza de datos:**

Después de contar con la columna de sector económico, en las bases de datos de personas naturales y jurídicas, y una vez revisadas las expectativas frente a los resultados de la segmentación con los expertos en cumplimiento, se procede a realizar la eliminación de las columnas que no se utilizaran en el análisis, dado que su impacto en la segmentación se define como mínimo o nulo, pues no cobraría relevancia contar con estos datos, teniendo en cuenta lo que se persigue desde el marco normativo SARLAFT, además, se pueden mejorar los resultados de segmentación entre menos variables o columnas se tengan en cuenta. Este paso ha sido sumamente revisado con las personas de cumplimiento, por lo que es de suma importancia a la hora de definir los cluster.

Ilustración 11: Limpieza de columnas irrelevantes del dataset

```
#Borrando columnas que no se emplearán en Persona_Natural, teniendo en cuenta los objetivos que se persiguen con la clusterizacion en el marco del SAGRILAFT
Persona_Natural.drop(['TIPIDE_CODIGO', 'TIPO', 'ACTIVIDAD_CODIGO', 'TIEMPO', 'FECHA_NACIMIENTO', 'EDAD', 'movicre', 'movideb', 'CIIU_CODIGO', 'desc. Actividades', 'FECHAVINCULACION', 'SUCURSAL_CODIGO'], axis
```

```
Persona_Natural.head()
```

	ACTIVO	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	riesgo actividades	Ubicación sucursal	Riesgo sucursal	TOTAL_OTROS_INGRESOS	montocre	montodeb	Sector
0	773000000.0	49527000.0	723473000.0	24729300.0	10066151.0	Medio	ANTIOQUIA	bajo	0.0	0	0.0	Actividades no especificadas / primarias
1	0.0	0.0	0.0	616000.0	0.0	Medio	ANTIOQUIA	bajo	0.0	0	0.0	Actividades no especificadas / primarias
2	0.0	0.0	0.0	0.0	0.0	Medio	CUNDINAMARCA	Medio	0.0	0	0.0	Actividades no especificadas / primarias
3	200000000.0	3000000.0	197000000.0	4600000.0	2000000.0	Bajo	ANTIOQUIA	bajo	0.0	0	0.0	Reparación y otros servicios personales
4	50000000.0	5000000.0	45000000.0	6400000.0	500000.0	Medio	ANTIOQUIA	bajo	0.0	0	0.0	Comercio minorista (retail)

Se realiza la limpieza de columnas irrelevantes o que no aportarán al modelo definido.

Ilustración 12: Continuación de limpieza de columnas

```
#Borrando columnas que no se emplearán en Persona_Juridica, teniendo en cuenta los objetivos que se persiguen con la clusterizacion en el marco del SAGRILAFT
Persona_Juridica.drop(['TIPIDE_CODIGO', 'TIPO', 'ACTIVIDAD_CODIGO', 'TIEMPO', 'FECHA_NACIMIENTO', 'EDAD', 'movicre', 'movideb', 'CIIU_CODIGO', 'desc. Actividades', 'FECHAVINCULACION', 'SUCURSAL_CODIGO'], axi
```

```
Persona_Juridica.head()
```

	ACTIVO	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	riesgo actividades	Ubicación sucursal	Riesgo sucursal	TOTAL_OTROS_INGRESOS	montocre	montodeb	Sector
0	2.566064e+09	1.991719e+09	5.743456e+08	0.000000e+00	0.000000e+00	Medio	CUNDINAMARCA	Medio	0.0	0	0.0	Organizaciones y asociaciones
1	NaN	NaN	NaN	0.000000e+00	0.000000e+00	Alto	CUNDINAMARCA	Medio	0.0	0	0.0	Metalurgia y metalmecánica
2	6.884440e+08	3.218070e+08	3.666370e+08	0.000000e+00	0.000000e+00	Medio	CUNDINAMARCA	Medio	0.0	32223370	49790868.0	Agropecuario (cultivos/ganadería)
3	1.037999e+09	8.000194e+08	2.379793e+08	0.000000e+00	0.000000e+00	Medio	CUNDINAMARCA	Medio	0.0	0	0.0	Equipo eléctrico y maquinaria
4	1.632140e+12	8.619680e+11	7.701720e+11	9.648483e+10	7.961950e+10	Medio	CUNDINAMARCA	Medio	NaN	0	0.0	Otros

Luego de realizar la eliminación de las columnas, se procede con la verificación e identificación de los

registros duplicados, encontrando que para las personas naturales se tienen 2591 registros duplicados y para las personas jurídicas 53. Estos registros se verificaron con los expertos del proceso y al tratarse de una maestra de clientes, se determinó que correspondían a errores en los datos y efectivamente se configuran como duplicados, lo cual no le sumaría al modelo, por tanto, se procedieron a eliminar y no tener en cuenta:

Ilustración 13: Identificación de registros duplicados Personas Naturales

```
# Verificación de registros duplicados
Persona_Natural.loc[Persona_Natural.duplicated()]
```

Este paso se realiza para identificar clientes que estén duplicados en la base de datos, que pudieran afectar los resultados.

Ilustración 14: Conteo y Eliminación de registros duplicados Personas Naturales

```
#Conteo de registros duplicados
num_duplicates_natural = Persona_Natural.duplicated().sum()
print(f"Número de registros duplicados en Persona_Natural: {num_duplicates_natural}")
```

```
Número de registros duplicados en Persona_Natural: 2591
```

```
#Eliminación de registros duplicados Personas Naturales
Persona_Natural.drop_duplicates(inplace=True)
print("Registros duplicados eliminados de Persona_Natural.")
```

```
Registros duplicados eliminados de Persona_Natural.
```

Se procede con la eliminación de los registros duplicados identificados

Ilustración 15: Identificación de registros duplicados Personas Jurídicas

```
#Verificación de registros duplicados
Persona_Juridica.loc[Persona_Juridica.duplicated()]
```

Se realiza el mismo proceso anterior, con la base de datos de personas jurídicas

Ilustración 16: Conteo y Eliminación de registros duplicados Personas Jurídicas

```
#Conteo de registros duplicados
num_duplicates_juridica = Persona_Juridica.duplicated().sum()
print(f"Número de registros duplicados en Persona_Juridica: {num_duplicates_juridica}")
```

```
Número de registros duplicados en Persona_Juridica: 53
```

```
#Eliminación de registros duplicados Personas Naturales
Persona_Juridica.drop_duplicates(inplace=True)
print("Registros duplicados eliminados de Persona_Juridica.")
```

```
Registros duplicados eliminados de Persona_Juridica.
```

Después de eliminar los registros duplicados y revisar la calidad de los datos, se pasa a la fase de visualización de los datos.

- **Visualización de los datos:**

Se procede a crear la lista de variables categóricas para las bases de datos de persona natural y persona jurídica, con el objetivo de realizar un tratamiento más ordenado de dichos datos:

Ilustración 17: Identificación de variables categóricas, Personas Naturales

```
#Lista de variables categóricas Persona Natural
PersonaN_catCols = Persona_Natural.select_dtypes(include = ["object", 'category']).columns.tolist()

Persona_Natural[PersonaN_catCols].head(5)
```

	riesgo actividades	Ubicación sucursal	Riesgo sucursal	Sector
0	Medio	ANTIOQUIA	bajo	Actividades no especificadas / primarias
1	Medio	ANTIOQUIA	bajo	Actividades no especificadas / primarias
2	Medio	CUNDINAMARCA	Medio	Actividades no especificadas / primarias
3	Bajo	ANTIOQUIA	bajo	Reparación y otros servicios personales
4	Medio	ANTIOQUIA	bajo	Comercio minorista (retail)

Se realiza la selección de variables categóricas, creando una variable que posteriormente facilitará el análisis y aplicación del modelo.

Ilustración 18: Identificación de variables categóricas, Personas Jurídicas

```
#Lista de variables categóricas Persona Juridica
PersonaJ_catCols = Persona_Juridica.select_dtypes(include = ["object", 'category']).columns.tolist()

Persona_Juridica[PersonaJ_catCols].head(5)
```

	riesgo actividades	Ubicación sucursal	Riesgo sucursal	Sector
0	Medio	CUNDINAMARCA	Medio	Organizaciones y asociaciones
1	Alto	CUNDINAMARCA	Medio	Metalurgia y metalmecánica
2	Medio	CUNDINAMARCA	Medio	Agropecuario (cultivos/ganadería)
3	Medio	CUNDINAMARCA	Medio	Equipo eléctrico y maquinaria
4	Medio	CUNDINAMARCA	Medio	Otros

Se realiza el mismo procedimiento anterior, aplicado al dataset de las personas jurídicas.

Se proceden a crear la lista de variables numéricas para cada base de datos

Ilustración 19: Variables numéricas para Personas Naturales

```
#Lista de variables numéricas Persona Natural

PersonaN_numCols=Persona_Natural.select_dtypes(include = ['float64','int32','int64']).columns.tolist()

Persona_Natural[PersonaN_numCols].head(5)
```

	ACTIVO	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	TOTAL_OTROS_INGRESOS	montocre	montodeb
0	773000000.0	49527000.0	723473000.0	24729300.0	10068151.0	0.0	0	0.0
1	0.0	0.0	0.0	616000.0	0.0	0.0	0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0	0.0
3	200000000.0	3000000.0	197000000.0	4800000.0	2000000.0	0.0	0	0.0
4	50000000.0	5000000.0	45000000.0	6400000.0	500000.0	0.0	0	0.0

Se seleccionan las variables numéricas para realizar posteriormente la normalización de los valores y poder gestionar de una mejor manera valores como activos y pasivos que presentan alta dispersión.

Ilustración 20: Variables numéricas para Personas Jurídicas

```
#Lista de variables numéricas Persona Jurídica

PersonaJ_numCols=Persona_Juridica.select_dtypes(include = ['float64','int32','int64']).columns.tolist()

Persona_Jurídica[PersonaJ_numCols].head(5)
```

	ACTIVO	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	TOTAL_OTROS_INGRESOS	montocre	montodeb
0	2.566064e+09	1.991719e+09	5.743456e+08	0.000000e+00	0.000000e+00	0.0	0	0.0
1	NaN	NaN	NaN	0.000000e+00	0.000000e+00	0.0	0	0.0
2	6.884440e+08	3.218070e+08	3.666370e+08	0.000000e+00	0.000000e+00	0.0	32223370	49790868.0
3	1.037999e+09	8.000194e+08	2.379793e+08	0.000000e+00	0.000000e+00	0.0	0	0.0
4	1.632140e+12	8.619680e+11	7.701720e+11	9.648483e+10	7.961950e+10	NaN	0	0.0

Se realiza la visualización de las variables categóricas para cada base de datos, con el propósito de entender como está conformada la data en estas variables:

Ilustración 21: Visualización de distribuciones en variables categóricas, Personas Naturales

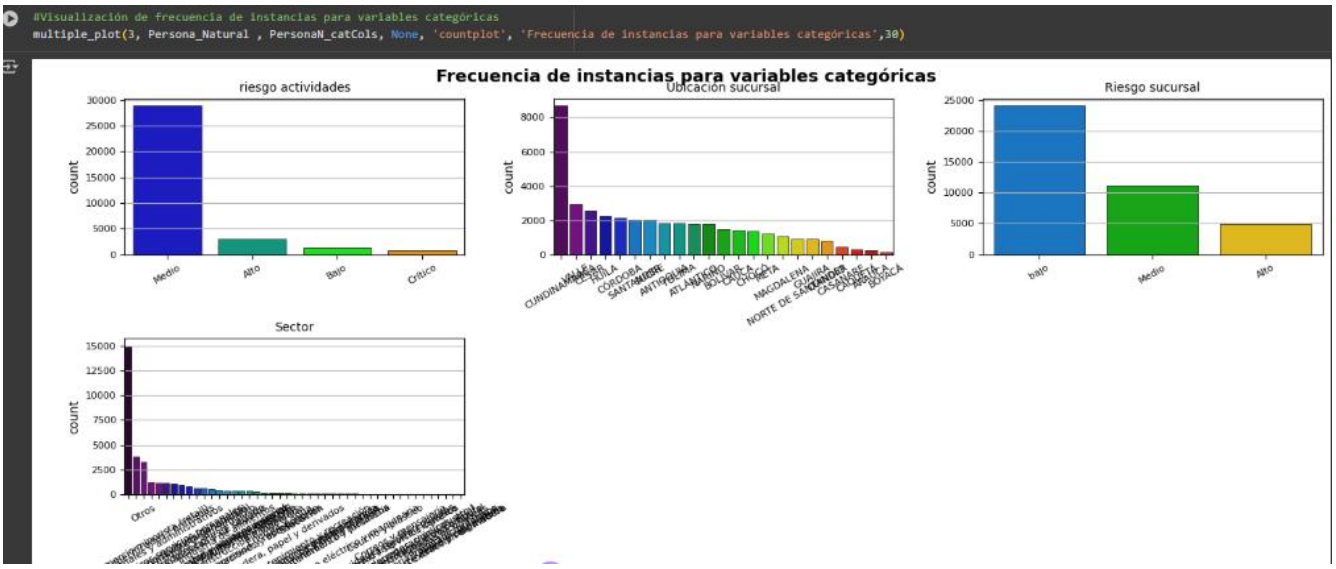
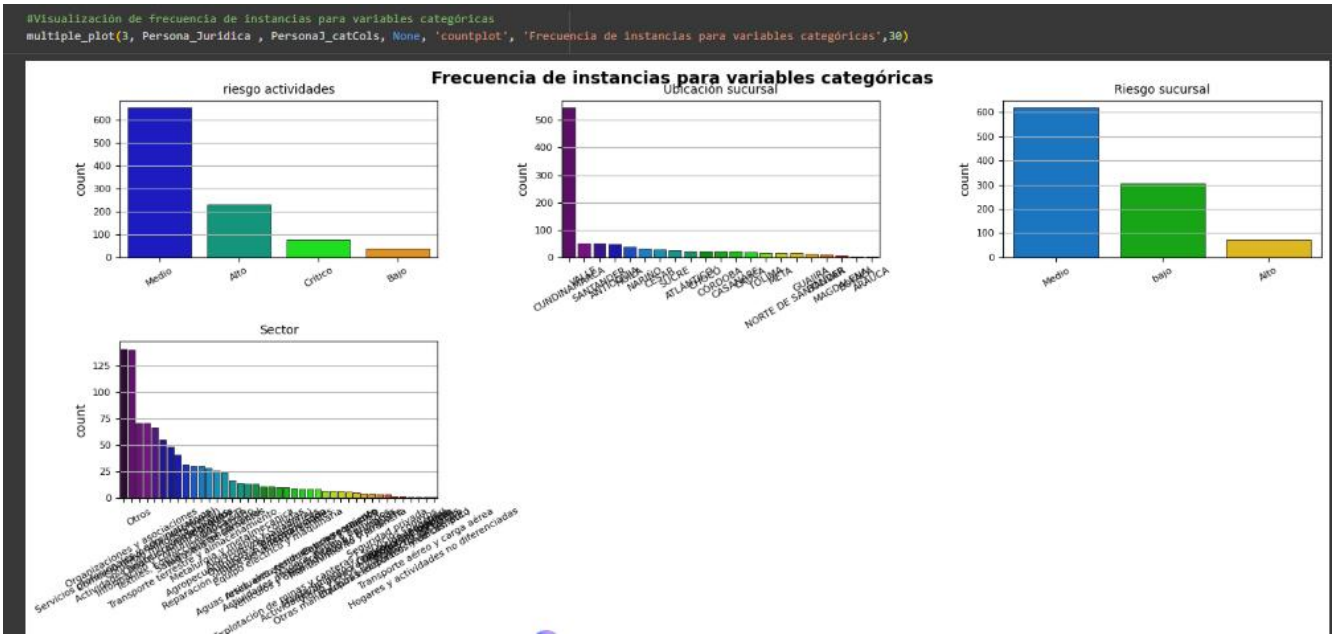


Ilustración 22: Ilustración 20: Visualización de distribuciones en variables categóricas, Personas Jurídicas



Luego de graficar las variables categóricas y entender a grandes rasgos como está compuesta la data, se procede a identificar valores nulos, en cada base de datos, validando el tratamiento que se le va a dar a estos registros con los expertos del proceso, por lo que se definió que para el caso de valores nulos en variables numéricas, se reemplazará el nulo por un 0, entendiendo que no se podrían eliminar de forma arbitraria los valores y es mejor contar con la data y en el caso de las variables categóricas, se identifican valores nulos en la variable de actividad económica y sector, por lo cual, se opta por reemplazar este valor nulo por la referencia “no especifica actividad” y “no relaciona sector”. Lo anterior, se acordó con el área de cumplimiento, buscando no afectar de forma significativa la base de datos,

definiendo que la revisión detallada de estos clientes será una de las mejoras que se buscará realizar a nivel de calidad de la data del maestro de clientes a nivel de compañía.

Ilustración 23: Tratamiento de valores nulos en el dataset

```
# Contar valores nulos en cada columna de Persona_Natural
print("Valores nulos en Persona_Natural:")
display(Persona_Natural.isnull().sum())

# Contar valores nulos en cada columna de Persona_Juridica
print("\nValores nulos en Persona_Juridica:")
display(Persona_Juridica.isnull().sum())
```

```
Valores nulos en Persona_Natural:
```

	0
ACTIVO	22
PASIVO	23
PATRIMONIO	18
INGRESOS	0
EGRESOS	303
riesgo actividades	5940
Ubicación sucursal	0
Riesgo sucursal	0
TOTAL_OTROS_INGRESOS	9
montocre	0
montodeb	0
Sector	6053

Ilustración 24: Continuación de tratamiento de valores nulos

```
Valores nulos en Persona_Juridica:
```

	0
ACTIVO	2
PASIVO	2
PATRIMONIO	2
INGRESOS	0
EGRESOS	158
riesgo actividades	1
Ubicación sucursal	0
Riesgo sucursal	0
TOTAL_OTROS_INGRESOS	245
montocre	0
montodeb	0
Sector	7

Ilustración 25: Reemplazo de valores nulos

```
# Reemplazar valores nulos con 0 en las columnas especificadas para Persona_Natural
columns_to_fill_natural = ['ACTIVO', 'PASIVO', 'PATRIMONIO', 'EGRESOS', 'TOTAL_OTROS_INGRESOS']
Persona_Natural[columns_to_fill_natural] = Persona_Natural[columns_to_fill_natural].fillna(0)

# Reemplazar valores nulos con 0 en las columnas especificadas para Persona_Juridica
columns_to_fill_juridica = ['ACTIVO', 'PASIVO', 'PATRIMONIO', 'EGRESOS', 'TOTAL_OTROS_INGRESOS']
Persona_Juridica[columns_to_fill_juridica] = Persona_Juridica[columns_to_fill_juridica].fillna(0)

print("Valores nulos sustituidos por 0 en las columnas especificadas para Persona_Natural y Persona_Juridica.")
```

```
Valores nulos sustituidos por 0 en las columnas especificadas para Persona_Natural y Persona_Juridica.
```

Se realiza el reemplazo de calores nulos para no afectar los resultados y análisis posteriores.

Ilustración 26: Continuación de reemplazo de valores nulos

```
# Reemplazar valores nulos en las columnas 'riesgo actividades' y 'Sector' para Persona_Natural
Persona_Natural['riesgo actividades'] = Persona_Natural['riesgo actividades'].fillna("No especifica actividad")
Persona_Natural['Sector'] = Persona_Natural['Sector'].fillna("No relaciona sector")

# Reemplazar valores nulos en las columnas 'riesgo actividades' y 'Sector' para Persona_Juridica
Persona_Juridica['riesgo actividades'] = Persona_Juridica['riesgo actividades'].fillna("No especifica actividad")
Persona_Juridica['Sector'] = Persona_Juridica['Sector'].fillna("No relaciona sector")

print("Valores nulos en 'riesgo actividades' y 'Sector' sustituidos para Persona_Natural y Persona_Juridica.")
```

```
Valores nulos en 'riesgo actividades' y 'Sector' sustituidos para Persona_Natural y Persona_Juridica.
```

Cumplida esta fase, se da por terminado el desarrollo del objetivo 1, donde se buscaba contar en las bases de datos objeto de análisis, con las variables necesarias para realizar una segmentación de clientes en cumplimiento de la normativa dispuesta para el marco SARLAFT, garantizando alineación con el área de cumplimiento de la compañía. Es importante aclarar que, en la norma que aplica, no se tiene definido de forma específica la cantidad o tipo de variable que se debe utilizar para realizar la segmentación, pues esto dependerá de los datos, recursos y equipo con el que cuente el negocio. Desde la norma se exige contar con un procedimiento o metodología claramente definida e implementada para realizar esta actividad, de tal forma que la organización sea capaz de soportar de forma objetiva los controles implementados en el programa de cumplimiento de SARLAFT, buscando siempre disminuir el nivel de riesgo o exposición frente al contagio de LA/TF.

9.2 Objetivo 2: Caracterización de variables ordinales y normalización de variables numéricas

Después de tener las variables numéricas y categóricas bien definidas y realizar la limpieza y adecuación de la data, se procede a definir la calificación ordinal para las variables de “Riesgo actividades” y “Riesgo sucursal”, con el objetivo de tener una adecuada ordinalidad de los datos en estas variables, de acuerdo con el nivel de riesgo, definiendo en un orden de: ['Bajo', 'Medio', 'Alto', 'Crítico’, ‘No especifica actividad'] para la variable de actividad económica y ['bajo', 'Medio', 'Alto'] para la variable de sucursal o ubicación.

Ilustración 27: Codificación ordinal de variables de riesgo

```
from sklearn.preprocessing import OrdinalEncoder, OneHotEncoder
import pandas as pd

# Definir el orden para la codificación ordinal de 'riesgo actividades' y 'Riesgo sucursal'
riesgo_order = ['Bajo', 'Medio', 'Alto', 'Crítico','No especifica actividad']
sucursal_riesgo_order = ['bajo', 'Medio', 'Alto']

# Codificación Ordinal para Persona_Natural
ordinal_encoder_PN = OrdinalEncoder(categories=[riesgo_order, sucursal_riesgo_order])
Persona_Natural[['riesgo actividades', 'Riesgo sucursal']] = ordinal_encoder_PN.fit_transform(Persona_Natural[['riesgo actividades', 'Riesgo sucursal']])

# Codificación One-Hot para Persona_Natural
onehot_encoder_PN = OneHotEncoder(handle_unknown='ignore', sparse_output=False)
ubicacion_sector_PN = onehot_encoder_PN.fit_transform(Persona_Natural[['Ubicación sucursal', 'Sector']])
ubicacion_sector_PN_df = pd.DataFrame(ubicacion_sector_PN, columns=onehot_encoder_PN.get_feature_names_out(['Ubicación sucursal', 'Sector']), index=Persona_Natural.index)

# Concatenar columnas codificadas one-hot con Persona_Natural
Persona_Natural = pd.concat([Persona_Natural.drop(columns=['Ubicación sucursal', 'Sector']), ubicacion_sector_PN_df], axis=1)

# Codificación Ordinal para Persona_Juridica
ordinal_encoder_PJ = OrdinalEncoder(categories=[riesgo_order, sucursal_riesgo_order])
Persona_Juridica[['riesgo actividades', 'Riesgo sucursal']] = ordinal_encoder_PJ.fit_transform(Persona_Juridica[['riesgo actividades', 'Riesgo sucursal']])

# Codificación One-Hot para Persona_Juridica
onehot_encoder_PJ = OneHotEncoder(handle_unknown='ignore', sparse_output=False)
ubicacion_sector_PJ = onehot_encoder_PJ.fit_transform(Persona_Juridica[['Ubicación sucursal', 'Sector']])
ubicacion_sector_PJ_df = pd.DataFrame(ubicacion_sector_PJ, columns=onehot_encoder_PJ.get_feature_names_out(['Ubicación sucursal', 'Sector']), index=Persona_Juridica.index)

# Concatenar columnas codificadas one-hot con Persona_Juridica
Persona_Juridica = pd.concat([Persona_Juridica.drop(columns=['Ubicación sucursal', 'Sector']), ubicacion_sector_PJ_df], axis=1)

print("Variables categóricas codificadas para Persona_Natural y Persona_Juridica.")
```

Se realiza la codificación ordinal, de tal manera que se pueda respetar jerarquía en los análisis que arroje el modelo y tener una segmentación que considere el peso del nivel de riesgo

Ilustración 28: Codificación para actividades económicas

```
# Se vizualizan los  primero registros del dataframe personas naturales
Persona_Natural.head(5)
```

	ACTIVO	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	riesgo actividades	Riesgo sucursal	TOTAL_OTROS_INGRESOS	montocre	montodeb	...	Sector_Salud y asistencia social	Sector_Seguridad privada	Sector_Servicios profesionales y administrativos	Sector_Silvicultura y extracción de madera
0	773000000.0	49527000.0	723473000.0	24729300.0	10066151.0	1.0	0.0	0.0	0	0.0	...	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	616000.0	0.0	1.0	0.0	0.0	0	0.0	...	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0	0.0	...	0.0	0.0	0.0	0.0
3	200000000.0	3000000.0	197000000.0	4800000.0	2000000.0	0.0	0.0	0.0	0	0.0	...	0.0	0.0	0.0	0.0
4	50000000.0	5000000.0	45000000.0	6400000.0	500000.0	1.0	0.0	0.0	0	0.0	...	0.0	0.0	0.0	0.0

5 rows × 79 columns

Ilustración 29: Codificación para ubicaciones

```
# Se vizualizan los  primero registros del dataframe personas juridicas
Persona_Juridica.head(5)
```

	ACTIVO	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	riesgo actividades	Riesgo sucursal	TOTAL_OTROS_INGRESOS	montocre	montodeb	...	Sector_Químicos y farmacéuticos	Sector_Reparación y otros servicios personales	Sector_Salud y asistencia social
0	2.566064e+09	1.991719e+09	5.743456e+08	0.000000e+00	0.000000e+00	1.0	1.0	0.0	0	0.0	...	0.0	0.0	0.0
1	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.0	1.0	0.0	0	0.0	...	0.0	0.0	0.0
2	6.884440e+08	3.218070e+08	3.666370e+08	0.000000e+00	0.000000e+00	1.0	1.0	0.0	32223370	49790868.0	...	0.0	0.0	0.0
3	1.037999e+09	8.000194e+08	2.379793e+08	0.000000e+00	0.000000e+00	1.0	1.0	0.0	0	0.0	...	0.0	0.0	0.0
4	1.632140e+12	8.619680e+11	7.701720e+11	9.648483e+10	7.961950e+10	1.0	1.0	0.0	0	0.0	...	0.0	0.0	0.0

5 rows × 73 columns

Por su parte, para las variables numéricas, se definió realizar una normalización de las variables, entendiendo que se tenían valores muy dispersos entre sí y cada variable de tipo financiero aportaba de forma importante a la segmentación.

Ilustración 30: Normalización de variables numéricas

```
from sklearn.preprocessing import StandardScaler

# Normalización de variables numericas para Persona_Natural
scaler_PN = StandardScaler()
Persona_Natural[PersonaN_numCols] = scaler_PN.fit_transform(Persona_Natural[PersonaN_numCols])

# Normalizacion de variables numericas para Persona_Juridica
scaler_PJ = StandardScaler()
Persona_Juridica[PersonaJ_numCols] = scaler_PJ.fit_transform(Persona_Juridica[PersonaJ_numCols])

print("Variables numéricas normalizadas para Persona_Natural y Persona_Juridica.")
```

Se realiza normalización de variables numéricas con el objetivo de disminuir el impacto de la dispersión de los valores de las variables numéricas como: activo, pasivo y patrimonio

Ilustración 31: Aplicación de normalización

```
Persona_Natural.head(5)
```

	ACTIVO	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	riesgo actividades	Riesgo sucursal	TOTAL_OTROS_INGRESOS	montocre	montodeb	...	Sector_Salud y asistencia social	Sector_Seguridad privada	Sector_Servicios profesionales y administrativos
0	2.146649	0.002562	0.174934	0.083940	-0.003108	1.0	0.0	-0.006673	-0.207188	-0.184112	...	0.0	0.0	0.0
1	-0.222582	-0.010016	-0.008584	-0.013942	-0.005474	1.0	0.0	-0.006673	-0.207188	-0.184112	...	0.0	0.0	0.0
2	-0.222582	-0.010016	-0.008584	-0.016442	-0.005474	1.0	1.0	-0.006673	-0.207188	-0.184112	...	0.0	0.0	0.0
3	0.390414	-0.009254	0.041388	0.002230	-0.005004	0.0	0.0	-0.006673	-0.207188	-0.184112	...	0.0	0.0	0.0
4	-0.069333	-0.008748	0.002831	0.009537	-0.005356	1.0	0.0	-0.006673	-0.207188	-0.184112	...	0.0	0.0	0.0

Ilustración 32: Resultado de normalización

```
Persona_Juridica.head(5)
```

	ACTIVO	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	riesgo actividades	Riesgo sucursal	TOTAL_OTROS_INGRESOS	montocre	montodeb	...	Sector_Químicos y farmacéuticos	Sector_Reparación y otros servicios personales	Sector_Salud y asistencia social
0	-0.133209	-0.132632	-0.133187	-0.093666	-0.124437	1.0	1.0	-0.138133	-0.117730	-0.178316	...	0.0	0.0	0.0
1	-0.133263	-0.132809	-0.133238	-0.093666	-0.124437	2.0	1.0	-0.138133	-0.117730	-0.178316	...	0.0	0.0	0.0
2	-0.133249	-0.132781	-0.133205	-0.093666	-0.124437	1.0	1.0	-0.138133	-0.094371	-0.126329	...	0.0	0.0	0.0
3	-0.133241	-0.132738	-0.133216	-0.093666	-0.124437	1.0	1.0	-0.138133	-0.117730	-0.178316	...	0.0	0.0	0.0
4	-0.098482	-0.055830	-0.067433	-0.071438	-0.063225	1.0	1.0	-0.138133	-0.117730	-0.178316	...	0.0	0.0	0.0

Tras finalizar con la normalización y definición de ordinalidad para las variables categóricas, se procede con la aplicación del modelo de segmentación.

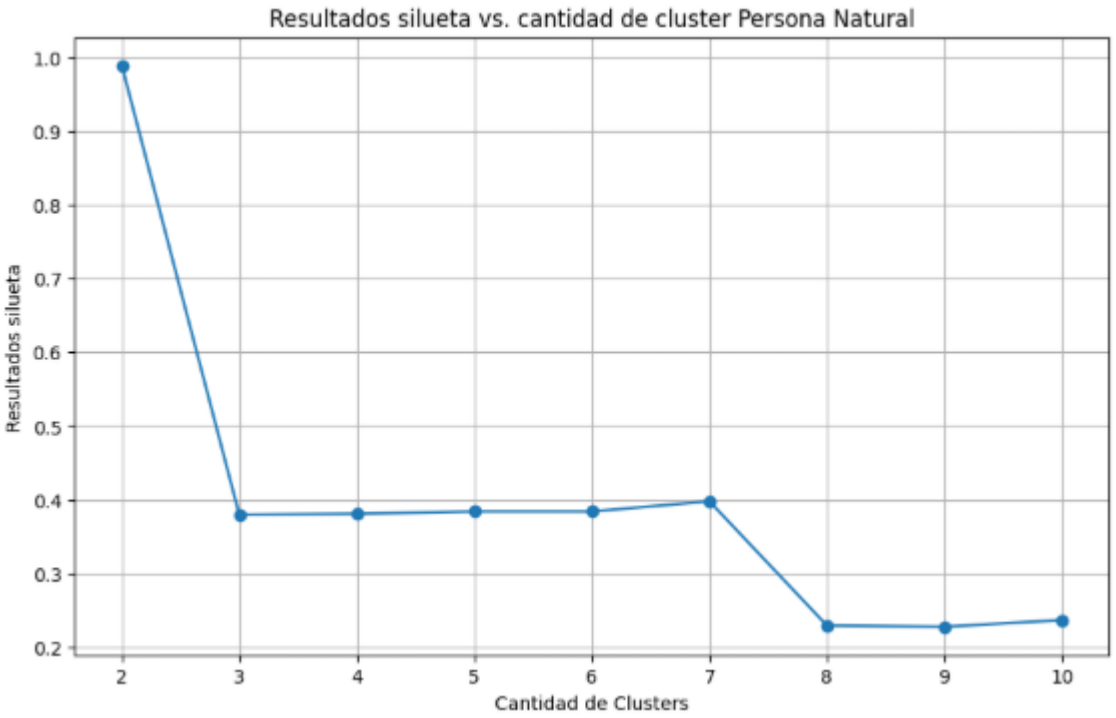
9.3 Objetivo 3: Implementación del modelo de segmentación

Para comenzar con la implementación del modelo, se requiere en primera instancia calcular la cantidad de cluster sobre las cuales se aplicará el modelo, para esto, se procede con la aplicación de la función para hallar la gráfica por el método de la silueta, de tal manera que a través de la identificación de los puntos de quiebre se pueda definir la cantidad de cluster que deben utilizarse en la segmentación de cada base de datos.

Ilustración 33: Determinación del número óptimo de clusters mediante método de siluet

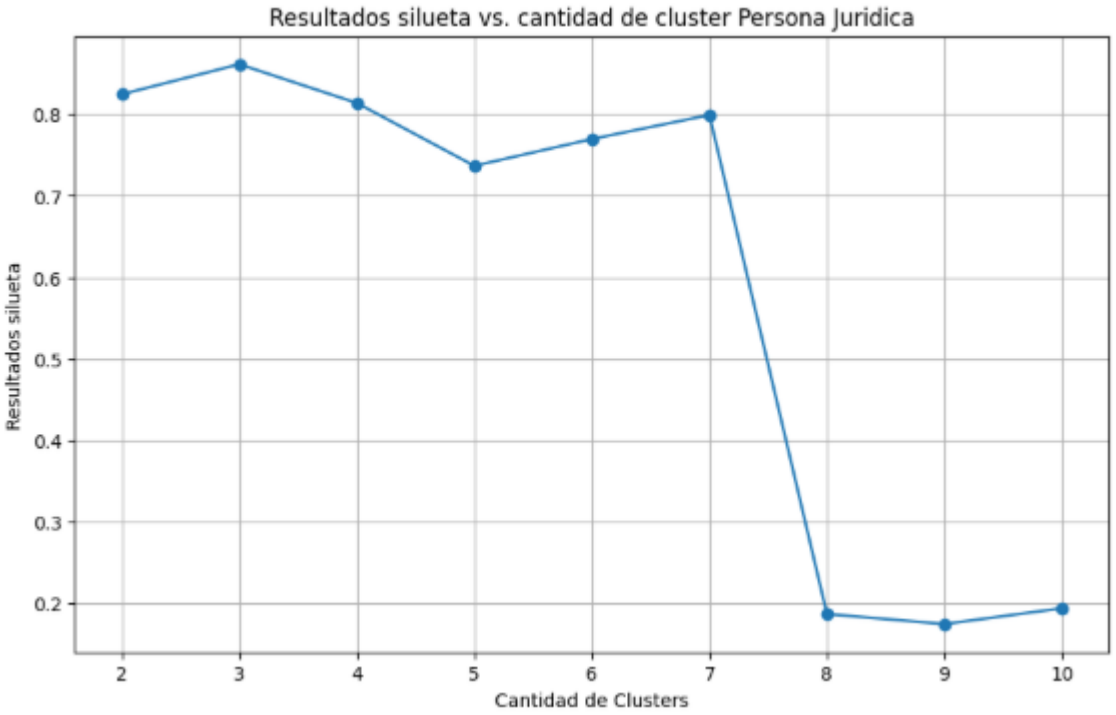
```python
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt
import numpy as np

# Funcion para calcular la grafica de silueta para definir la cantidad de cluster
def plot_silhouette_scores(dataframe, max_clusters, title):
    silhouette_scores = []
    for n_clusters in range(2, max_clusters + 1):
        kmeans = KMeans(n_clusters=n_clusters, random_state=42, n_init=10)
        kmeans.fit(dataframe)
        score = silhouette_score(dataframe, kmeans.labels_)
        silhouette_scores.append(score)

    # graficar resultado de la silueta
    plt.figure(figsize=(10, 6))
    plt.plot(range(2, max_clusters + 1), silhouette_scores, marker='o')
    plt.title(f'Resultados silueta vs. cantidad de cluster {title}')
    plt.xlabel('Cantidad de Clusters')
    plt.ylabel('Resultados silueta')
    plt.xticks(range(2, max_clusters + 1))
    plt.grid(True)
    plt.show()

# Determinar maximo de cluster (e.g., up to 10)
max_clusters_to_test = 10

# Resultados para Persona_Natural
plot_silhouette_scores(Persona_Natural, max_clusters_to_test, 'Persona Natural')

# Resultados para Persona_Juridica
plot_silhouette_scores(Persona_Juridica, max_clusters_to_test, 'Persona Juridica')
```

Ilustración 34: Gráfica de silueta para Personas Naturales



Se grafica con el propósito de identificar la cantidad de clusters a generar para personas naturales.

Ilustración 35: Gráfica de silueta para Personas Jurídicas



Una vez analizadas las gráficas, se determinó que la cantidad óptima de clusters para el cálculo de la segmentación es de 3, tanto para personas naturales como para personas jurídicas.

- **Implementación del K-means:**

Se procede con la implementación del k-means utilizando 3 clusters

Ilustración 36: Implementación del algoritmo K-means con 3 clusters

```
from sklearn.cluster import KMeans

# Aplicar KMeans a Persona_Natural con 3 clusters
kmeans_PN = KMeans(n_clusters=3, random_state=42, n_init=10)
Persona_Natural['Cluster'] = kmeans_PN.fit_predict(Persona_Natural)

# Aplicar KMeans a Persona_Juridica con 3 clusters
kmeans_PJ = KMeans(n_clusters=3, random_state=42, n_init=10)
Persona_Juridica['Cluster'] = kmeans_PJ.fit_predict(Persona_Juridica)

print("Clustering KMeans aplicado a Persona_Natural (3 clusters) y Persona_Juridica (3 clusters).")
```

Ilustración 37: Aplicación de K-means

```
# Recuento de registros por cluster para Persona_Natural
cluster_counts_PN = Persona_Natural['Cluster'].value_counts().sort_index()
print("Número de registros por cluster en Persona_Natural:")
display(cluster_counts_PN)

# Recuento de registros por cluster para Persona_Juridica
cluster_counts_PJ = Persona_Juridica['Cluster'].value_counts().sort_index()
print("\nNúmero de registros por cluster en Persona_Juridica:")
display(cluster_counts_PJ)
```

Ilustración 38: Resultados de clustering

Número de registros por cluster en Persona_Natural:

Cluster	count
0	6812
1	33259
2	1

dtype: int64

Número de registros por cluster en Persona_Juridica:

Cluster	count
0	991
1	1
2	5

Después de calcular los clusters y la cantidad de registros que los conforman, se realiza el cálculo de los centroides, con el objetivo de tener las características de cada cluster, de manera que se pueda hacer una descripción general de los registros que los conforman.

Ilustración 39: Cálculo de centroides para caracterización de clusters

```
# Calcular los centroides de Persona_Natural
centroids_PN = Persona_Natural.groupby('Cluster').mean()
print("Centroides para Persona_Natural:")
display(centroids_PN)

# Calcular centroides para Persona_Juridica
centroids_PJ = Persona_Juridica.groupby('Cluster').mean()
print("\nCentroides para Persona_Juridica:")
display(centroids_PJ)
```

Ilustración 40: Centroides para Personas Naturales

Cluster	ACTIVO	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	riesgo actividades	Riesgo sucursal	TOTAL_OTROS_INGRESOS	montocre	montodeb	...	Sector_Salud y asistencia social	Sector_Seguridad privada	Sector_Servicios profesionales y administrativos
0	-0.112878	-0.008724	-0.000736	-0.013335	-0.005326	3.871991	0.630358	-0.006388	-0.033961	-0.023231	...	0.000000	0.000000	0.000000
1	0.023042	-0.003777	0.005702	0.002731	0.001091	1.050603	0.498722	0.001309	0.008962	0.004764	...	0.010794	0.001834	0.113413
2	2.575746	185.038425	-184.619701	-0.007837	-0.004886	2.000000	1.000000	-0.006673	-0.200966	-0.184112	...	0.000000	0.000000	0.000000

Ilustración 41: Centroides para Personas Jurídicas

Cluster	ACTIVO	PASIVO	PATRIMONIO	INGRESOS	EGRESOS	riesgo actividades	Riesgo sucursal	TOTAL_OTROS_INGRESOS	montocre	montodeb	...	Sector_Químicos y farmacéuticos	Sector_Reparación y otros servicios personales	Sector_Salud y asistencia social
0	-0.019183	-0.029294	-0.054047	-0.017714	-0.058252	1.349142	0.764884	-0.021453	-0.028835	-0.023147	...	0.0111	0.013118	0.026236
1	-0.059268	0.008003	0.028717	-0.093666	-0.124437	1.000000	0.000000	-0.138133	29.163686	21.299129	...	0.0000	0.000000	0.000000
2	3.809973	5.804552	10.706398	3.529635	11.570404	1.000000	0.800000	4.279605	-0.117730	0.327935	...	0.0000	0.000000	0.400000

Con el propósito de tener una visualización que permita mayor claridad frente a la conformación de los clusters, se realizan las gráficas en 3 y 2 planos de los registros generados por el modelo.

Ilustración 42: Visualización de clusters en 3D

```python
import plotly.express as px
import plotly.graph_objects as go

# Función para trazar conglomerados interactivos en 3D con ejes y centroides especificados.
def plot_interactive_3d_specific_axes_with_centroids(dataframe, x_col, y_col, z_col, title, centroids):

    dataframe_plot = dataframe.copy()
    dataframe_plot['Cluster'] = dataframe_plot['Cluster'].astype(str)

    # Crear gráfico de dispersión 3D interactivo usando Plotly Express
    fig = px.scatter_3d(dataframe_plot,
                        x=x_col,
                        y=y_col,
                        z=z_col,
                        color='Cluster',
                        title=f'Visualización interactiva de clústeres en 3D con centroides para {title}')

    # Añadir centroides al gráfico
    fig.add_trace(go.Scatter3d(
        x=centroids[x_col],
        y=centroids[y_col],
        z=centroids[z_col],
        mode='markers',
        marker=dict(
            color='black',  # Color del centroide
            size=10,
            symbol='diamond'
        ),
        name='Centroids'
    ))

    fig.show()

# Calcular los centroides de las columnas especificadas
centroids_PN_subset = Persona_Natural.groupby('Cluster')[['riesgo actividades', 'Riesgo sucursal', 'ACTIVO']].mean()
centroids_PJ_subset = Persona_Juridica.groupby('Cluster')[['riesgo actividades', 'Riesgo sucursal', 'ACTIVO']].mean()
```

Ilustración 43: Visualización en 3D, Persona Natural

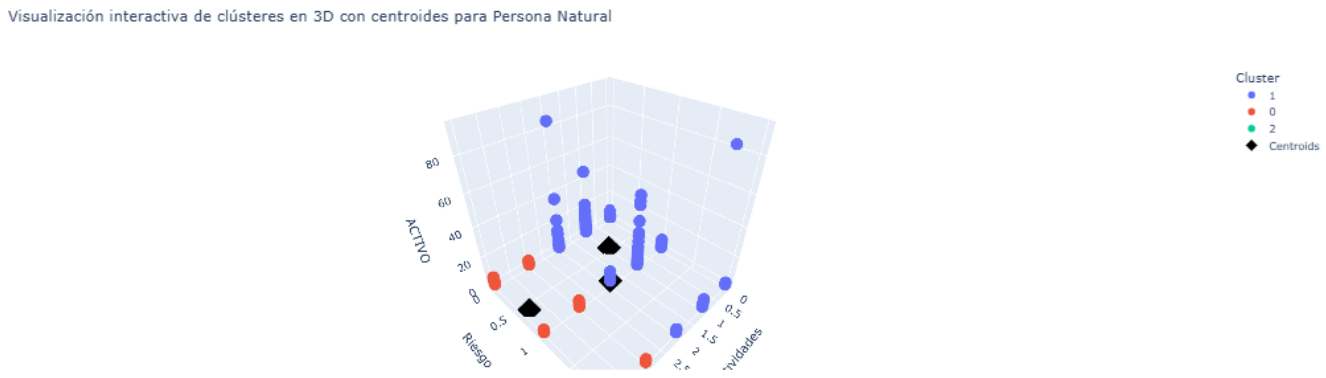


Ilustración 44: Visualización en 3D, Persona Jurídica

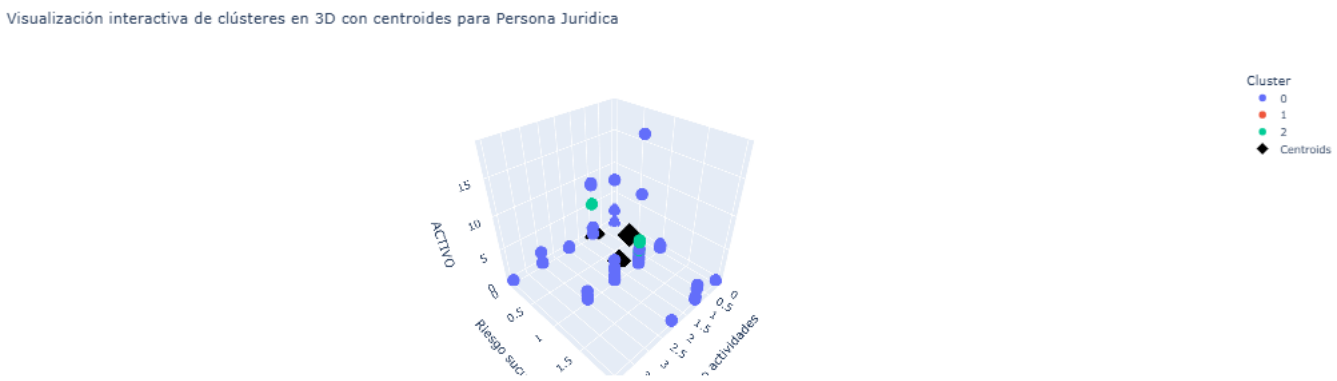


Ilustración 45: Visualización de clusters en 2D

```python
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

# Función para realizar PCA y trazar clusters 2D
def plot_2d_clusters(dataframe, title):
    # Características separadas y etiquetas de conglomerados
    X = dataframe.drop(columns=['Cluster'])
    labels = dataframe['Cluster']

    # Aplicar PCA para reducir a 2 dimensiones
    pca = PCA(n_components=2)
    X_pca = pca.fit_transform(X)

    # Crear un gráfico de dispersión 2D
    plt.figure(figsize=(10, 8))

    # Obtener etiquetas de clúster únicas
    unique_labels = sorted(labels.unique())

    # Trazar cada conglomerado
    for cluster in unique_labels:
        plt.scatter(X_pca[labels == cluster, 0],
                    X_pca[labels == cluster, 1],
                    label=f'Cluster {cluster}')

    plt.xlabel('Componente principal 1')
    plt.ylabel('Componente principal 2')
    plt.title(f'Visualización de clústeres en 2D para {title}')
    plt.legend()
    plt.grid(True)
    plt.show()

# Trazar clusters 2D para Persona_Natural
plot_2d_clusters(Persona_Natural, 'Persona Natural')
```

Se procede con la ejecución del script para tener una visual mucho más clara de la composición de los segmentos generados, a través de una gráfica 2D

Ilustración 46: Visualización en 2D, Persona Natural

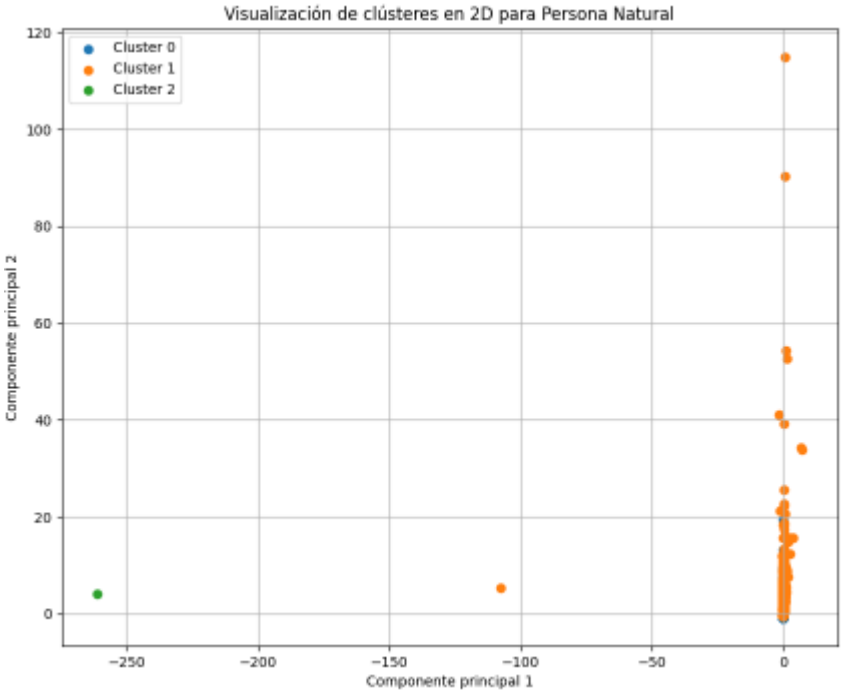
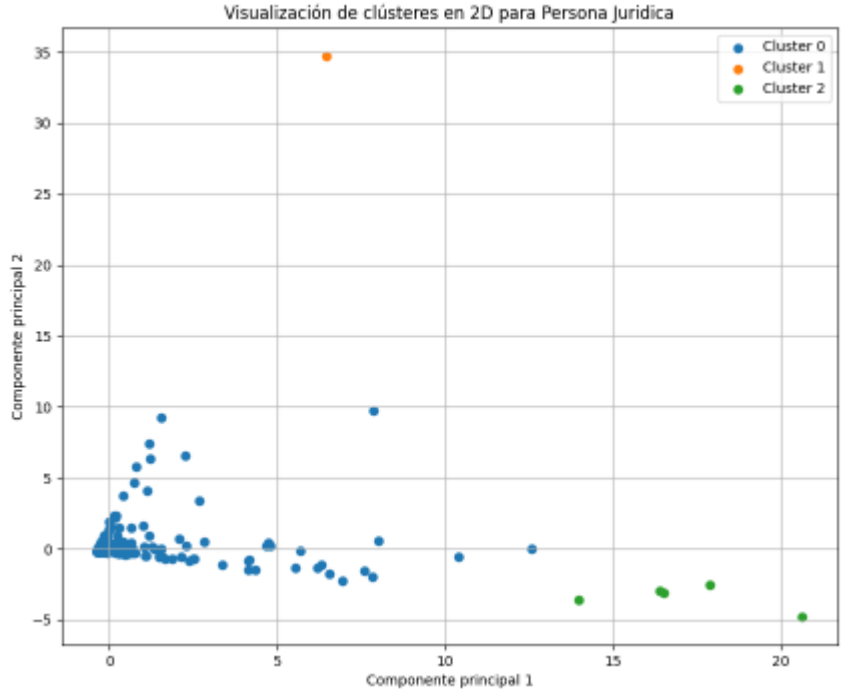
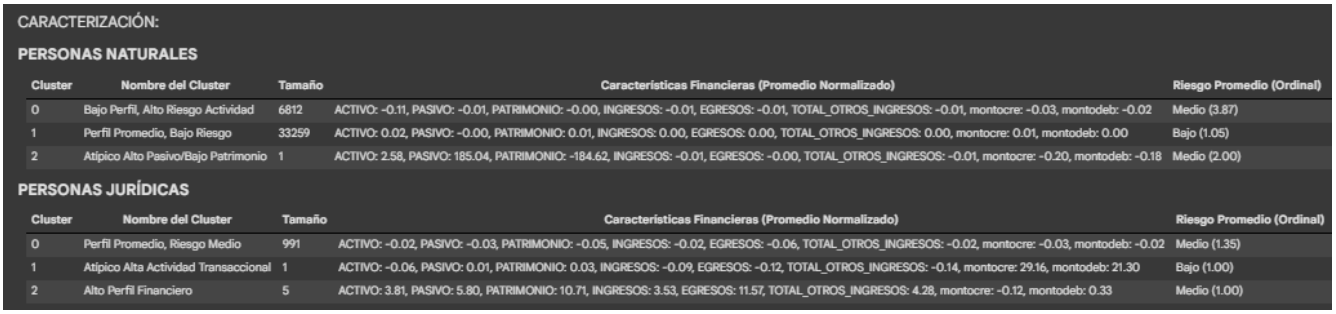


Ilustración 47: Visualización en 3D, Persona Jurídica



- **Caracterización de cluster:**

Ilustración 48: Caracterización y perfilamiento de clusters resultantes



CARACTERIZACIÓN:

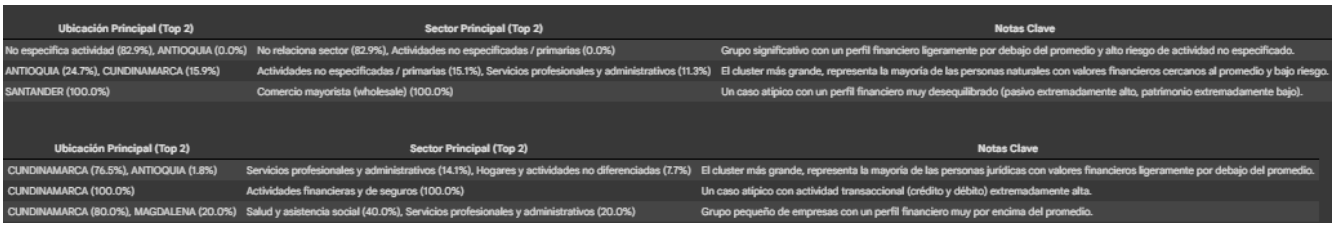
PERSONAS NATURALES

Cluster	Nombre del Cluster	Tamaño	Características Financieras (Promedio Normalizado)	Riesgo Promedio (Ordinal)
0	Bajo Perfil, Alto Riesgo Actividad	6812	ACTIVO: -0.11, PASIVO: -0.01, PATRIMONIO: -0.00, INGRESOS: -0.01, EGRESOS: -0.01, TOTAL_OTROS_INGRESOS: -0.01, montocre: -0.03, montodeb: -0.02	Medio (3.87)
1	Perfil Promedio, Bajo Riesgo	33259	ACTIVO: 0.02, PASIVO: -0.00, PATRIMONIO: 0.01, INGRESOS: 0.00, EGRESOS: 0.00, TOTAL_OTROS_INGRESOS: 0.00, montocre: 0.01, montodeb: 0.00	Bajo (1.05)
2	Atípico Alto Pasivo/Bajo Patrimonio	1	ACTIVO: 2.58, PASIVO: 185.04, PATRIMONIO: -184.62, INGRESOS: -0.01, EGRESOS: -0.00, TOTAL_OTROS_INGRESOS: -0.01, montocre: -0.20, montodeb: -0.18	Medio (2.00)

PERSONAS JURÍDICAS

Cluster	Nombre del Cluster	Tamaño	Características Financieras (Promedio Normalizado)	Riesgo Promedio (Ordinal)
0	Perfil Promedio, Riesgo Medio	991	ACTIVO: -0.02, PASIVO: -0.03, PATRIMONIO: -0.05, INGRESOS: -0.02, EGRESOS: -0.06, TOTAL_OTROS_INGRESOS: -0.02, montocre: -0.03, montodeb: -0.02	Medio (1.35)
1	Atípico Alta Actividad Transaccional	1	ACTIVO: -0.06, PASIVO: 0.01, PATRIMONIO: 0.03, INGRESOS: -0.09, EGRESOS: -0.12, TOTAL_OTROS_INGRESOS: -0.14, montocre: 29.16, montodeb: 21.30	Bajo (1.00)
2	Alto Perfil Financiero	5	ACTIVO: 3.81, PASIVO: 5.80, PATRIMONIO: 10.71, INGRESOS: 3.53, EGRESOS: 11.57, TOTAL_OTROS_INGRESOS: 4.28, montocre: -0.12, montodeb: 0.33	Medio (1.00)

Ilustración 49: Perfilamiento adicional



Ubicación Principal (Top 2)	Sector Principal (Top 2)	Notas Clave
No especifica actividad (82.9%), ANTIOQUIA (0.0%)	No relaciona sector (82.9%), Actividades no especificadas / primarias (0.0%)	Grupo significativo con un perfil financiero ligeramente por debajo del promedio y alto riesgo de actividad no especificado.
ANTIOQUIA (24.7%), CUNDINAMARCA (15.9%)	Actividades no especificadas / primarias (15.1%), Servicios profesionales y administrativos (11.3%)	El cluster más grande, representa la mayoría de las personas naturales con valores financieros cercanos al promedio y bajo riesgo.
SANTANDER (100.0%)	Comercio mayorista (wholesale) (100.0%)	Un caso atípico con un perfil financiero muy desequilibrado (pasivo extremadamente alto, patrimonio extremadamente bajo).

Ubicación Principal (Top 2)	Sector Principal (Top 2)	Notas Clave
CUNDINAMARCA (76.5%), ANTIOQUIA (1.8%)	Servicios profesionales y administrativos (14.1%), Hogares y actividades no diferenciadas (7.7%)	El cluster más grande, representa la mayoría de las personas jurídicas con valores financieros ligeramente por debajo del promedio.
CUNDINAMARCA (100.0%)	Actividades financieras y de seguros (100.0%)	Un caso atípico con actividad transaccional (crédito y débito) extremadamente alta.
CUNDINAMARCA (80.0%), MAGDALENA (20.0%)	Salud y asistencia social (40.0%), Servicios profesionales y administrativos (20.0%)	Grupo pequeño de empresas con un perfil financiero muy por encima del promedio.

Finalizada la aplicación del modelo, se obtienen como resultado los cluster para cada base de datos o tipo de cliente, encontrando que para el caso de personas jurídicas el cluster “2” compuesto por 5 clientes, requiere de despliegue de debida diligencia ampliada y mayor profundización en las verificaciones realizadas a nivel de controles y validación de documentación soporte. Para el caso de los Clusters “0” y “1”, la cooperativa está tranquila con los controles que se han ejecutado y la simple verificación en listas restrictivas, dado que en el ejercicio anual de actualización se pueden conseguir los documentos y soportes requeridos por la norma, sin requerir de muchos recursos adicionales para su gestión. Los resultados arrojados en el modelo sirven para soportar estas decisiones frente a un ente de control.

Para el caso de personas naturales, se obtienen dos clusters que son objeto de revisión profunda, tal es el caso del cluster “2”, conformado por una sola persona que presenta niveles de transaccionalidad atípicos y por su actividad económica requiere de realización de debida diligencia ampliada y será un cliente objeto de seguimiento y especial control. Por su parte para el cluster “0” se debe realizar un despliegue más profundo y demandante, teniendo en cuenta que en este cluster no se logra identificar actividad económica o sector económico y con respecto a los demás cluster su riesgo promedio es cercano a 3.87, es decir, se encuentra en un nivel alto, en este caso, al tratarse de un volumen tan importante de clientes, se decide realizar un análisis de importancia respecto a la cartera total, de cada cliente, así como la vigencia de cada uno de estos con respecto a sus saldos actuales, buscando realizar una confirmación de actividades económicas para aquellos clientes que sean relevantes para el negocio y los que no lo sean, terminar con la relación comercial, buscando no exponer a la empresa

a riesgos mayores de tipo SARLAFT. Para los demás clientes, se espera que con el ejercicio masivo de verificación de listas restrictivas y la actualización anual, se cubran las brechas que se podrían presentar.

- **Métricas:**

Ilustración 50: Métricas de evaluación del modelo de clustering

```
# Calcular Silhouette Score para Persona_Natural
silhouette_avg_PN = silhouette_score(Persona_Natural.drop(columns=['Cluster']), Persona_Natural['Cluster'])
print(f"Silhouette Score para Persona_Natural: {silhouette_avg_PN}")

# Calcular Davies-Bouldin Index para Persona_Natural
davies_bouldin_PN = davies_bouldin_score(Persona_Natural.drop(columns=['Cluster']), Persona_Natural['Cluster'])
print(f"Davies-Bouldin Index para Persona_Natural: {davies_bouldin_PN}")

# Calcular Silhouette Score para Persona_Juridica
silhouette_avg_PJ = silhouette_score(Persona_Juridica.drop(columns=['Cluster']), Persona_Juridica['Cluster'])
print(f"\nSilhouette Score para Persona_Juridica: {silhouette_avg_PJ}")

# Calcular Davies-Bouldin Index para Persona_Juridica
davies_bouldin_PJ = davies_bouldin_score(Persona_Juridica.drop(columns=['Cluster']), Persona_Juridica['Cluster'])
print(f"Davies-Bouldin Index para Persona_Juridica: {davies_bouldin_PJ}")
```

```
Silhouette Score para Persona_Natural: 0.3796731414206169
Davies-Bouldin Index para Persona_Natural: 0.7145863494266994

Silhouette Score para Persona_Juridica: 0.8616295399683793
Davies-Bouldin Index para Persona_Juridica: 0.49814939058778807
```

Ilustración 51: Método del Codo para Persona Natural

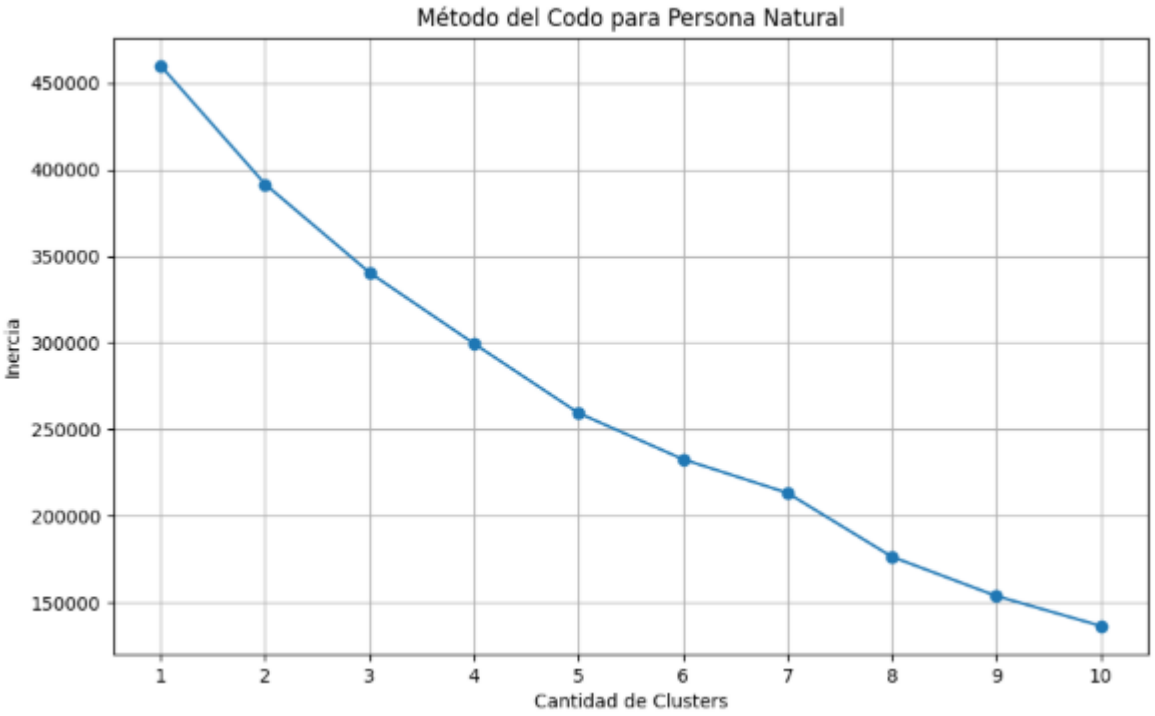
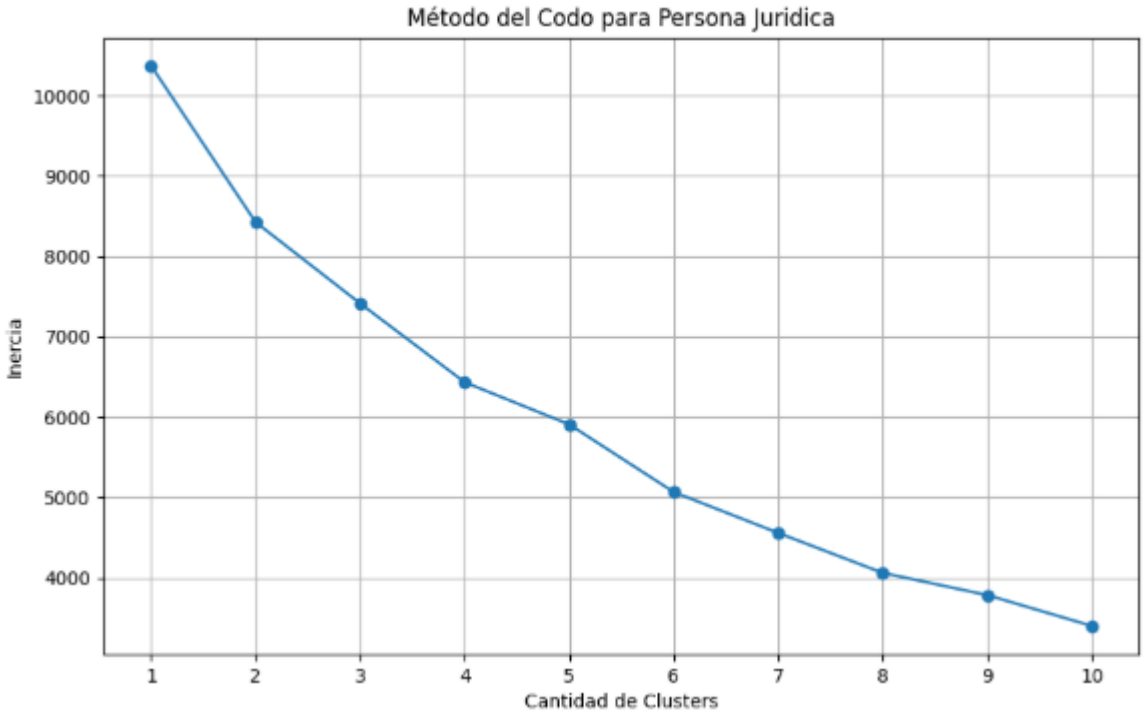


Ilustración 52: Método del Codo para Persona Jurídica



Personas Naturales:

1. El ***Silhouette Score*** **(0.38)** indica una separación moderada entre los clusters.
2. El ***Davies-Bouldin*** **(0.71)** muestra una segmentación aceptable, pero con cercanía entre clústeres.
3. El **método del codo** respalda el uso de 3 clusters, aunque la mejora después de este punto es limitada.

Personas Jurídicas:

1. El ***Silhouette Score*** **(0.86)** evidencia clusters bien definidos y alta separación.
2. El ***Davies-Bouldin*** **(0.50)** confirma baja superposición y buena calidad de segmentación.
3. El **método del codo** sugiere 2 o 3 clusters, siendo 3 una partición sólida según los indicadores.

9.3.1 Conclusión General:

El clustering parece funcionar mejor para los datos de Persona Jurídica, obteniendo clusters más cohesivos y mejor separados según las métricas de evaluación. Para Persona Natural, las métricas sugieren que los clusters no están tan claramente definidos, lo que podría indicar la necesidad de explorar un número diferente de clusters o considerar otras técnicas de preprocesamiento o clustering.

Teniendo en cuenta lo anterior, es preciso que una vez se haga el ejercicio de depuración de clientes en personas naturales y se actualicen las actividades económicas y sectores, se pueda volver a correr el modelo, con el propósito de tener otro tipo de resultados, excluyendo al cliente que presenta registros tan atípicos, para el cual se define un tratamiento diferente.

10. Conclusiones

10.1 Objetivo 1: Caracterización de variables y preparación de la data

La caracterización y depuración de las variables permitió consolidar una base de datos robusta, coherente y alineada con los requerimientos normativos del SARLAFT, garantizando que la segmentación posterior se fundamente en información confiable y estructurada. El proceso de enriquecimiento de la base —mediante la incorporación de descriptores de actividad económica, riesgo asociado y ubicación geográfica— fortaleció la capacidad analítica del modelo, permitiendo vincular los factores de riesgo inherentes al cliente con variables de contexto. Este trabajo no solo optimizó la calidad de los datos, sino que aseguró la trazabilidad y objetividad exigidas por los entes de supervisión, al construir un insumo analítico que soporta decisiones de gestión del riesgo basadas en evidencia.

10.2 Objetivo 2: Caracterización de variables ordinales y normalización de variables numéricas

La estandarización de las variables, tanto en su dimensión ordinal como numérica, fue esencial para asegurar la coherencia interna del modelo y la comparabilidad entre clientes. La definición de escalas de riesgo para las actividades económicas y sucursales permitió transformar percepciones cualitativas de riesgo en métricas cuantificables, alineadas con la lógica de evaluación del SARLAFT. Asimismo, la normalización de las variables financieras redujo sesgos derivados de dispersión en los valores, garantizando que el modelo de segmentación reflejara adecuadamente patrones de comportamiento homogéneos. Este tratamiento estadístico aportó solidez metodológica, asegurando que la segmentación final respondiera a criterios técnicos y no a distorsiones de escala o valores extremos.

10.3 Objetivo 3: Implementación del modelo de segmentación

La implementación del modelo de segmentación mediante K-means demostró la capacidad de la organización para aplicar metodologías de analítica avanzada al cumplimiento normativo del SARLAFT. Los resultados obtenidos permitieron identificar grupos diferenciados de clientes según su perfil de riesgo, facilitando la focalización de recursos y el diseño de controles proporcionales. En particular, la identificación de clústeres de alto riesgo, tanto en personas naturales como jurídicas, evidencia la efectividad del modelo como herramienta preventiva y de priorización de la debida diligencia ampliada. Este ejercicio constituye un avance significativo hacia una gestión del riesgo basada en datos, permitiendo a la cooperativa fortalecer su programa de cumplimiento con fundamentos técnicos, medibles y replicables ante auditorías o entes de supervisión. No obstante, se requiere una depuración y evaluación de clientes exhaustiva para el caso de personas naturales, en aras de contar con una segmentación más ajustada. Adicionalmente, con la ejecución de este trabajo, se identificaron aspectos

que deben ser mejorados en el manejo y control de la base de datos de clientes, garantizando datos de fuente ajustados a la realidad y que posteriormente sirvan para realizar revisiones más detalladas. Este trabajo no solo se configuró como un paso importante hacia la gestión de riesgos asociados a SARLAFT, sino que también proporcionó insumos para definir controles en el proceso que garanticen una operación más ajustada a las normas y expectativas de control de la Cooperativa.

11. Recomendaciones

11.1 Fortalecer la calidad y actualización de la base de datos maestra de clientes

Es fundamental implementar un programa de **depuración y actualización continua de la información**, que garantice la integridad, consistencia y completitud de los datos de clientes. Se recomienda establecer mecanismos automáticos de validación en el proceso de vinculación y actualización anual, especialmente para variables críticas como la **actividad económica, sector y ubicación**, a fin de minimizar registros nulos o inconsistentes que afecten la calidad del modelo.

11.2 Integrar fuentes externas de información y listas restrictivas en tiempo real

Para fortalecer el componente de vigilancia preventiva del SARLAFT, se sugiere conectar la base de clientes con fuentes externas confiables (DIAN, DANE, UIAF, ONU, OFAC, entre otras) y con herramientas de screening masivo como Stradata Search, de forma periódica y automatizada. Esto permitirá enriquecer el modelo con variables de riesgo dinámicas, facilitando la detección temprana de operaciones inusuales o vinculaciones con listas restrictivas o PEP.

11.3 Documentar y estandarizar la metodología de segmentación

Se recomienda elaborar un manual metodológico que documente los pasos técnicos, criterios de selección de variables, técnicas de normalización y procedimientos de validación estadística. Este documento servirá como guía para futuras investigaciones y auditorías internas o externas, además de garantizar la reproducibilidad y trazabilidad del proceso, en cumplimiento con las exigencias de la Superintendencia y las mejores prácticas del SARLAFT.

11.4 Incorporar técnicas avanzadas de aprendizaje automático supervisado y no supervisado

Para futuras investigaciones, sería conveniente explorar modelos complementarios a K-means, como DBSCAN, Gaussian Mixture Models o árboles de decisión, que permitan capturar comportamientos no lineales o patrones atípicos con mayor precisión. Asimismo, la implementación de algoritmos supervisados (como Random Forest o XGBoost) puede facilitar la predicción del riesgo de clientes nuevos o actuales, integrando el componente de segmentación con el de alerta temprana.

11.5 Implementar un tablero de control (dashboard) de monitoreo de riesgo segmentado

Con el propósito de convertir el modelo en una herramienta operativa, se sugiere desarrollar un dashboard interactivo que permita visualizar los resultados de la segmentación, identificar clientes de alto riesgo y monitorear la evolución de los clústeres a lo largo del tiempo. Este tablero puede integrarse con los sistemas internos de cumplimiento, mejorando la toma de decisiones basada en datos y facilitando la priorización de casos para debida diligencia ampliada.

12. Referencias

1. Confecoop. (2022). *Informe de impacto económico y social del cooperativismo en Colombia*. Bogotá.
2. Superintendencia Financiera de Colombia. (2020). *Circular Básica Jurídica SARLAFT*.
3. Gómez, L. (2019). *Gestión del riesgo en cooperativas financieras: Un enfoque* práctico. Editorial Legis.
4. Infolaft. (2018). Sancionada comisionista por no utilizar variables de segmentación.
5. Superintendencia Financiera de Colombia. (2022). Circular Básica Jurídica 007 de 2022. Bogotá: SFC.
6. Ley 1908 de 2018 (modificada por la Ley 2195 de 2022, "Ley de Financiamiento Terrorista"), que refuerza las obligaciones de prevención.
7. Decreto 1674 de 2020, que reglamenta aspectos operativos del SARLAFT.
8. Zhang, Y., et al. (2021). "Unsupervised clustering for financial risk segmentation". Expert Systems with Applications, 185, 115643.
9. Pérez, L., & González, M. (2022). "AI-driven compliance in anti-money laundering: A cost-benefit analysis". Journal of Financial Compliance, 5(2), 45-60.
10. Infolaft. (2023). *SARLAFT: claves para segmentar factores de riesgo*.
11. Gómez, J., & Ramírez, C. (2020). "Machine learning gaps in AML frameworks: A Latin American perspective". Latin American Journal of Economics, 57(3), 301-320.
12. Tan, P.-N., Steinbach, M., & Kumar, V. (2019). Introduction to Data Mining (2nd ed.). Pearson.
13. Infolaft. (2021). Segmentación SARLAFT 4.0: ¿cómo se mide la calidad de datos?
14. Jain, A. K. (2010). "Data clustering: 50 years beyond K-means". Pattern Recognition Letters, 31(8), 651-666.
15. Infolaft. (2021). El regulador se pronuncia sobre la segmentación.
16. Jovel, W. (2020). “Desarrollo de un modelo analítico para la segmentación de asociados en una cooperativa de ahorros y créditos”, (Tesis de Maestría, Universidad Nacional de Colombia) repositorio.unal.edu.co.
17. Correa, S., & Montoya, L. (2024). “Análisis de segmentación y alertamiento transaccional para la gestión de riesgos sarlaft en el sector financiero”, (Tesis de trabajo de grado, Tecnológico de Antioquia Institución Universitaria) repositorio digital tdea.

18. Ramos, N. (2023). “Modelo de Segmentación para SARLAFT en R4G”, (Tesis de Maestría, Universidad del Rosario)
19. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM Consortium.
20. Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis. Wiley.
21. Infolaft. (2021). Las distintas formas de segmentar.
22. Ramos, N. (2023). Modelo de Segmentación para SARLAFT en R4G.
23. Ortiz, K. Y. (2023). Metodologías estadísticas para la segmentación en SARLAFT.
24. Infolaft. (2021). Sanción SARLAFT en Colombia por fallas en gestión de alertas.
25. Correa, S. M., & Montoya, L. Y. (2024). Análisis de segmentación y alertamiento transaccional para la gestión de riesgos SARLAFT en el sector financiero.
26. Infolaft. (2017). Superfinanciera sanciona a aseguradora por no tener segmentación.
27. Stradata. (2021). *SARLAFT 4.0 segmentación | Homogeneidad y heterogeneidad*.
28. Pérez, L. E. (2020). *Metodología de segmentación para el SARLAFT*. Universidad El Bosque.
29. Risk Monitor. (2020). *Metodología segmentación SARLAFT y SAGRILAFT*.
30. Superintendencia Financiera de Colombia. (2020). *Nueva guía de mejores prácticas en modelos de segmentación en factores de riesgo de LA/FT*.
31. Han, Z., Li, X., Lin, H., Yang, H., & Luo, J. (2018). A survey of federated search: From the perspective of resource representation and organization. Journal of the Association for Information Science and Technology, 69(6), 771-785