Metodologías estadísticas para la segmentación en SARLAFT

Karen Ortiz Pedraza

Santiago Mejía Barón

Asesor:

Edgar Andrés Villabon

Universidad Nacional Abierta y a Distancia – UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Especialización en Ciencia de Datos y Analítica

2024

Resumen

Esta monografía se enfoca en analizar críticamente los métodos de segmentación en SARLAFT, con un enfoque específico en técnicas de aprendizaje automático. El objetivo general es comparar la eficacia de estos métodos en la identificación y gestión de riesgos asociados a actividades ilícitas y financiación del terrorismo en el contexto latinoamericano. Para lograr este objetivo, se han establecido tres objetivos específicos. En primer lugar, se proporciona una explicación detallada sobre la aplicación de la minería de datos en SARLAFT, con énfasis en el aprendizaje automático. Segundo, se exploran conceptualmente métodos clave de aprendizaje automático como el Análisis de Componentes Principales (ACP), K-means y Árboles de Decisión para comprender sus fundamentos teóricos y aplicaciones en SARLAFT. Finalmente, se analizan las ventajas y limitaciones de estos métodos en el contexto de SARLAFT, basándose en investigaciones previas y teorías fundamentales. Este enfoque crítico y comparativo proporciona una comprensión más profunda de la efectividad e idoneidad de los métodos de segmentación en SARLAFT, contribuyendo así a la mejora de la gestión de riesgos financieros en la región latinoamericana.

Palabras clave: SARLAFT, Segmentación, Estadística, Riesgo

Abstract

This monograph focuses on a critical analysis of SARLAFT segmentation methods, with a specific focus on machine learning techniques. The general objective is to compare the effectiveness of these methods in the identification and management of risks associated with illicit activities and terrorist financing in the Latin American context. To achieve this objective, three specific objectives have been established. First, a detailed explanation is provided on the application of data mining in SARLAFT, with emphasis on machine learning. Second, key machine learning methods such as Principal Component Analysis (PCA), K-means and Decision Trees are conceptually explored to understand their theoretical foundations and applications in SARLAFT. Finally, the advantages and limitations of these methods in the SARLAFT context are analyzed, based on previous research and fundamental theories. This critical and comparative approach provides a deeper understanding of the effectiveness and suitability of segmentation methods in SARLAFT, thus contributing to the improvement of financial risk management in the Latin American region.

Key words: SARLAFT, Segmentation, Statistics, Risk

Tabla de contenido

Lista de Figuras

Introducción

En el contexto del Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT), este documento presenta un análisis sobre el uso de técnicas de minería de datos en la detección y gestión de riesgos financieros asociados a actividades ilícitas. Durante la elaboración de esta monografía, se exploraron diversas técnicas, como el análisis de componentes principales (ACP), el algoritmo de K-Means y los árboles de decisión, entre otros, con el objetivo de comprender su aplicabilidad y efectividad en el contexto de SARLAFT. Además, se pretende transmitir un mensaje claro sobre la importancia de utilizar herramientas analíticas avanzadas en la lucha contra el lavado de activos y la financiación del terrorismo, destacando la necesidad de una aproximación integral y basada en datos para mitigar estos riesgos en el sector financiero.

A lo largo de este trabajo, se examinará el proceso de implementación de estas técnicas, se discutirán sus ventajas y limitaciones, y se ofrecerán recomendaciones prácticas para su aplicación efectiva en entornos de SARLAFT. Con ello, se busca contribuir al conocimiento y la comprensión de las estrategias analíticas utilizadas para combatir el lavado de activos, promoviendo así la integridad y seguridad del sistema financiero.

Planteamiento del problema

El Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT) surge como una respuesta crucial para prevenir y detectar actividades ilícitas, estableciéndose como un estándar global adoptado en diversos sectores financieros. Su relevancia es particularmente notable en regiones donde las amenazas asociadas son más prominentes. Se ha implementado ampliamente en América Latina y en varios países de Europa y Asia. Países como Colombia, México, Brasil y Perú han adoptado medidas significativas en la implementación y aplicación de este sistema. En Europa, países como España y Reino Unido también han fortalecido sus regulaciones financieras para abordar estos riesgos. SARLAFT engloba un conjunto de políticas y procedimientos implementados por las instituciones financieras con el fin de identificar, evaluar y mitigar los riesgos vinculados al lavado de activos y al financiamiento del terrorismo. Además de su aplicación en el ámbito bancario, SARLAFT también se extiende a otros sectores como el asegurador, fiduciario y de valores, entre otros.

En este contexto, la segmentación en el SARLAFT se refiere a la división de clientes, transacciones o entidades en grupos homogéneos con características similares medidas, con el propósito de aplicar de control y supervisión más efectivas. El problema que se plantea en esta monografía radica en la necesidad de analizar y documentar las metodologías estadísticas utilizadas en la segmentación del SARLAFT. Según datos de la Oficina de las Naciones Unidas contra la Droga y el Delito (UNODC), se estima que el lavado de dinero equivale a entre el 2% y el 5% del producto interno bruto (PIB) mundial, lo que demuestra la magnitud de este problema a nivel global (UNDOC, 2011). Estas cifras refuerzan la necesidad crítica de fortalecer los

mecanismos de prevención y detección, siendo la segmentación una herramienta clave en este esfuerzo.

Aunque existen diversas técnicas y enfoques para llevar a cabo esta tarea, la complejidad de los datos financieros y la evolución constante de las estrategias de lavado de dinero y financiación del terrorismo plantean desafíos significativos. Para abordar este problema, es esencial explorar diversos métodos de segmentación en SARLAFT que se adapten de manera óptima a las particularidades del sector financiero. Esto implica considerar la alta dimensionalidad de los datos financieros, la presencia de datos desequilibrados. La relevancia de este problema no se limita únicamente al ámbito financiero, ya que el lavado de activos y la financiación del terrorismo representan amenazas significativas para la estabilidad de las economías y la seguridad nacional en todo el mundo. Una segmentación ineficiente en el SARLAFT puede tener graves consecuencias, incluyendo multas regulatorias, sanciones legales, daños a la reputación de la organización y, en última instancia, un mayor riesgo de que actividades delictivas pasen desapercibidas. Por lo tanto, la resolución efectiva de este problema es de interés crítico tanto para las instituciones financieras como para los reguladores y las autoridades encargadas de hacer cumplir la ley.

Justificación

En el ámbito de SARLAFT la segmentación se erige como una herramienta esencial para la identificación de actividades ilícitas, permitiendo a las instituciones financieras mitigar riesgos y cumplir con las regulaciones establecidas. Sin embargo, el desarrollo y la implementación de métodos efectivos de segmentación no son tareas simples, pues requieren una comprensión profunda de las metodologías estadísticas empleadas en este proceso. La presente monografía se centra en este aspecto crucial, analizando y documentando las diferentes técnicas utilizadas en la segmentación del SARLAFT.

Una segmentación efectiva es esencial para identificar transacciones y clientes sospechosos, contribuyendo a la lucha contra el crimen organizado y el terrorismo, que son amenazas significativas. Las instituciones financieras y no financieras están sujetas a regulaciones que requieren un estricto cumplimiento del SARLAFT. Por tanto, la segmentación es un componente clave para cumplir con estas regulaciones y evitar sanciones, multas, daños a la reputación de la organización y asignación incorrecta de recursos. Los métodos de segmentación estadística deben adaptarse a los requisitos locales, lo que complica aún más la tarea y justifica la necesidad de enfoques más flexibles y personalizados. Como menciona (Mariño & Chaparro, 2014), la incapacidad de una institución para identificar y prevenir actividades ilícitas puede quebrantar la confianza del cliente en el sistema financiero. En la actualidad, existen tecnologías avanzadas como el aprendizaje automático, el análisis de redes y la minería de datos que pueden aprovecharse para mejorar la segmentación.

Aunque el objetivo principal de esta investigación no reside en la mejora directa de los métodos de segmentación, sino en el análisis de ciertas metodologías existentes, se busca

proporcionar una base sólida para futuras mejoras y optimizaciones en el proceso de segmentación, al brindar una comprensión de las fortalezas y limitaciones de cada enfoque. Además, identifica áreas de oportunidad para investigaciones posteriores, contribuyendo así al avance continuo en la eficacia de la segmentación en SARLAFT.

Objetivos

Objetivo General

Analizar la literatura existente de los diferentes métodos de segmentación en SARLAFT, centrándose en enfoques de aprendizaje automático, para comparar su eficacia en la identificación y gestión de riesgos asociados a actividades ilícitas y financiación del terrorismo.

Objetivos Específicos

Explicar detalladamente la aplicación de la minería de datos en SARLAFT, proporcionando un paso a paso para obtener una comprensión más profunda del proceso en el marco del Aprendizaje Automático.

Explorar los métodos de Aprendizaje Automático: Análisis de Componentes Principales (ACP), K-means y Árboles de Decisión de forma conceptual con el propósito de asimilar su funcionamiento y fundamentos teóricos en el contexto de SARLAFT.

Comentar ventajas y limitaciones de los métodos de Aprendizaje Automático en el contexto de SARLAFT, basándose en investigaciones previas y teorías fundamentales con el fin de entender su idoneidad y eficacia.

Marco Conceptual

En el ámbito de la gestión financiera y la seguridad, se encuentran conceptos fundamentales que guían la prevención y el análisis de riesgos:

El **Lavado de Activos (LA)**, definido por (Esteban et al., 2012), se revela como un proceso diseñado para otorgar a los activos ilícitos una apariencia lícita, disociándolos de su origen delictivo mediante operaciones específicas y su inserción en circuitos legítimos. Estos conceptos son piezas fundamentales en la comprensión y abordaje de la seguridad financiera y la gestión de riesgos en entornos gubernamentales y empresariales.

La **Financiación del terrorismo (FT)** según el (Fatf Gafi, s.f.) se refiere a las actividades de reunir y proporcionar dinero con la intención de que sea utilizado en actividades terroristas. En el contexto latinoamericano, esto incluye cualquier acto que facilite o apoye directamente la ejecución de ataques terroristas. Las leyes en varios países de la región buscan penalizar y prevenir estas actividades para proteger la seguridad y estabilidad de sus sociedades.

Se utilizan **técnicas de minería de datos** como Redes Neuronales, Árboles de Decisión y Redes de Creencia Bayesiana en modelos predictivos, según (Ruiz, 2006), que emplean sistemas de aprendizaje autónomo para el reconocimiento de patrones basados en hechos históricos; en la gran mayoría de los casos, estos datos se utilizan para determinar los patrones. La detección de fraude involucra identificar patrones de comportamiento de transacciones fraudulentas o de la "utilización" normal de los clientes para detectar operaciones sospechosas.

Acompañando esta noción, se destacan otras metodologías esenciales, como el **Análisis Clúster**, que busca discernir patrones homogéneos en observaciones mediante criterios específicos. Asimismo, la **Segmentación** se presenta como un proceso clave, permitiendo el

análisis de conjuntos de datos en segmentos más manejables para un análisis detallado. En este contexto, el concepto de **Riesgo** cobra vital importancia al referirse a la probabilidad de eventos no deseados y la magnitud de su impacto en diversos contextos.

Según (Gurrea, 2000), el **análisis de componentes principales** (ACP) es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables). Es decir, ante un banco de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible. El meta principal en la detección de anomalías es encontrar objetos que sean diferentes de los demás, frecuentemente conocidos como *outlier*. De acuerdo con la minería descriptiva, la cual se basa en encontrar patrones (correlaciones, tendencias, grupos, trayectorias y anomalías) que resuman relaciones en los datos

La **Transferencia** es la transacción efectuada por una persona natural o jurídica o estructura sin personería jurídica denominada "ordenante", a través de una entidad autorizada en la respectiva jurisdicción para realizar transferencias nacionales y/o internacionales, mediante movimientos electrónicos o contables, con el fin de que una suma de dinero se ponga a disposición de una persona natural o jurídica o estructura sin personería jurídica denominada "beneficiaria", en otra entidad autorizada para realizar este tipo de operaciones. El ordenante y el beneficiario pueden ser la misma persona (Superintendencia Financiera, 2009).

Basados en (Bishop, 1995) y (Kanungo, 2002) **K-means** es un algoritmo de aprendizaje no supervisado que aborda el problema de agrupamiento. Se sigue un proceso simple para clasificar un conjunto de datos en un número predefinido de grupos (k grupos). Se inicia definiendo k centroides, uno para cada grupo, colocándolos inicialmente de manera estratégica para maximizar la separación entre ellos. Luego, cada punto de datos se asigna al centroide más

cercano. Después de asignar todos los puntos, se recalculan los centroides como los centros de los grupos formados. Este proceso se repite hasta que los centroides dejen de cambiar de posición. El objetivo del algoritmo es minimizar una función de error, como la suma de los cuadrados de las distancias entre los puntos y sus centroides asignados.

Clúster definido por (Aggarwal y Reddy, 2014) **se** refiere a un conjunto o grupo de elementos similares o relacionados que comparten características comunes dentro de un conjunto de datos más grande. El proceso de **clústering** implica la identificación y agrupación de observaciones, entidades o datos con similitudes con el objetivo de revelar patrones subyacentes, segmentar datos y facilitar la comprensión de la estructura de los datos.

Marco Teórico

En el contexto del SARLAFT, resulta relevante destacar las distintas fases del sistema de gestión de riesgos asociados al lavado de dinero y financiamiento del terrorismo. Según (Bayona, 2019), se identifican dos etapas fundamentales: la etapa de prevención de riesgos, cuyo propósito radica en evitar la entrada de recursos procedentes de actividades delictivas, y se logra recopilando información de los clientes; la siguiente etapa se centra en el control de riesgos, que busca gestionar de manera efectiva los riesgos identificados, por medio de un sistema de monitoreo para detectar transacciones irregulares. Por otra parte (Hoyos, 2019), subraya de manera contundente la trascendental importancia de la gestión y comprensión adecuada de la información contable dentro del ámbito empresarial. En este contexto, la información contable se presenta como una herramienta invaluable para las organizaciones, ya que le proporciona una visión detallada y precisa de su situación financiera. Esto es esencial para identificar posibles riesgos que pueden surgir, lo que permite a la entidad anticiparse a situaciones adversas o tomar medidas para contrarrestar los efectos negativos de los mismos.

Es crucial comprender y usar eficientemente la información disponible para gestionar el riesgo de lavado de activos y financiamiento del terrorismo, reconociendo que no hay un enfoque universal (Castro & Castro, 2020). Cada entidad debe adaptar sus herramientas dentro del marco regulatorio establecido por la Superintendencia Financiera. La flexibilidad y adaptabilidad en la gestión de riesgos son elementos cruciales que permiten a las organizaciones abordar los desafíos del lavado de activos y la financiación del terrorismo de manera efectiva. Un aspecto adicional para considerar se relaciona con el tipo de actividad económica susceptible a estos riesgos. Según (Mariño & Chaparro, 2014), de acuerdo con el modelo de negocios seguros, aquellos

propensos a una mayor vulnerabilidad son las corporaciones, las entidades sin ánimo de lucro, las personas políticamente expuestas y aquellas que llevan a cabo actividades económicas de alto riesgo. El lavado de activos se ha expandido a todas las áreas de la economía, incluyendo el mercado intermediario y financiero, conforme avanzan las instituciones y la tecnología financiera. Este avance tecnológico requiere una atención especial para comprender cómo la criminalidad aprovecha los nuevos canales para blanquear capitales (Martínez et al., 2022).

Con relación a los métodos, herramientas o instrumentos para monitorear SARLAFT, se disponen de diversas opciones adaptadas a cada etapa e intención. A continuación, se abordarán algunos de ellos:

Iniciativas privadas y estatales han surgido en respuesta a la Convención de Viena de 1999 y la Declaración de Basilea de 1988, enfocándose en la selección, identificación y conocimiento del cliente, así como en la conservación de información con fines probatorios. Estas acciones incluyen cooperación internacional, la formación de comités para definir perfiles de operaciones sospechosas, y esfuerzos estatales para fortalecer la moralidad pública y combatir la corrupción, según (Rodríguez & Galvis, 2008). Por ejemplo, (Ruiz, 2006) destaca la complejidad en la evolución de las estrategias de lavado de activos. Su artículo analiza dos categorías de herramientas: técnicas tradicionales, como la identificación de clientes en listas de control y el análisis de relaciones entre transacciones; y técnicas de minería de datos, que emplean modelos avanzados como análisis de clustering y redes neuronales para mejorar la detección de casos sospechosos.

En el ámbito interno, las entidades hacen uso de herramientas estadísticas que les facilitan la identificación y gestión de los riesgos asociados a actividades ilícitas. Según lo señala

(Enríquez, 2019), estas herramientas se orientan hacia modelos de segmentación, los cuales buscan dividir y agrupar características similares. Los segmentos presentan una consistencia interna en su comportamiento, esto indica que los individuos dentro de cada grupo (segmento) tengan comportamientos o características similares entre ellos. Al mismo tiempo, estos grupos deben ser diferentes entre sí, mostrando variabilidad en comparación con otros segmentos. En este contexto, se emplean diversas técnicas, que incluyen:

Predictivas

Descriptivas

Ad-hoc

Post-hoc

Análisis clúster

Conglomerados no jerárquicos

Algoritmo K-medias

Matrices de proximidad

Análisis de componentes principales

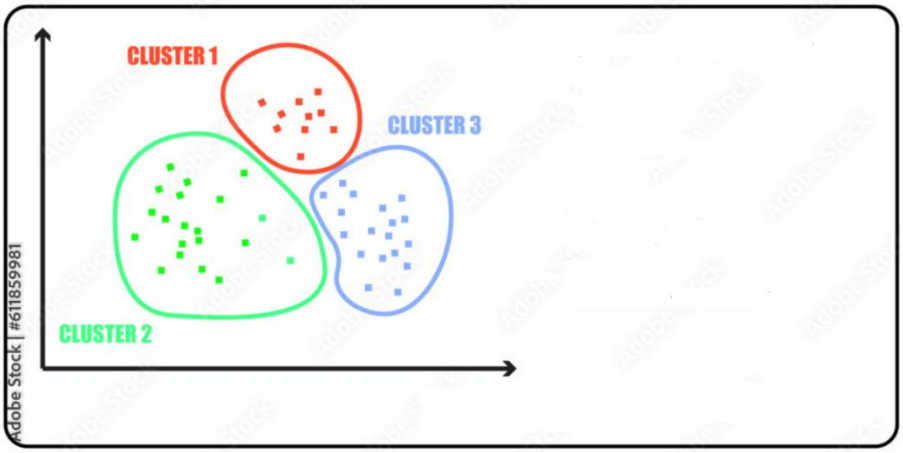
Análisis Clúster

Los métodos de agrupación como expone (Madhulatha, 2012) se pueden dividir en dos categorías de algoritmos para la agrupación de datos: los jerárquicos y los particionales. Dentro de los jerárquicos, se distinguen los aglomerativos, que fusionan elementos de abajo hacia arriba, y los divisivos, que dividen el conjunto de arriba hacia abajo. En el contexto de la agrupación jerárquica, se destaca la importancia de elegir una medida de distancia, como la distancia de Manhattan y la distancia euclidiana, también Madhulatha habla sobre la agrupación divisiva, que

genera una jerarquía de clústeres de arriba hacia abajo, dividiendo el conjunto inicial en grupos más pequeños de manera recursiva hasta que cada elemento está en su propio grupo. Aunque conceptualmente más complejo, tiene la ventaja de ser eficiente si no se genera una jerarquía completa.

Figura 1

Análisis clúster



Nota. Esta imagen presenta la distribución de tres clústeres distintos, identificados como Clúster 1 (rojo), Clúster 2 (verde) y Clúster 3 (azul). Cada clúster contiene un conjunto de puntos que representan datos agrupados de manera similar. *Fuente*. MakZin (s.f.).

Según (Rokach & Maimon, 2010), el clustering implica la agrupación de instancias u objetos similares, lo que requiere de métodos que determinen si dos objetos son comparables o no. Para esto, se emplean principalmente dos tipos de medidas: las de distancia y las de similitud, las cuales se utilizan para estimar esta relación. Muchos métodos de clustering utilizan medidas de distancia para determinar la similitud o disimilitud entre un par de objetos o desemejanza

entre un par de objetos. El concepto de distancia es crucial como lo sostiene (Díaz, 2007), ya que facilita al investigador la identificación de diversas características. Resulta útil denotar la distancia entre dos instancias x_i y x_j como: $\mathrm{d}(x_i, x_j)$. Una medida de distancia válida debe ser simétrica y obtener su valor mínimo (normalmente cero) en caso de vectores idénticos. La medida de distancia se denomina medida de distancia métrica si además satisface las siguientes propiedades.

$$\mathrm{d}(x_i, x_k) \leq \mathrm{d}(x_i, x_j) + \mathrm{d}(x_j, x_k) \; \forall \; x_i, x_j, x_k \in \mathrm{S} \qquad (1)$$

$$\mathrm{d}(x_i, x_j) = 0 \Rightarrow x_i, = x_j \; \forall \; x_i, x_j \in \mathrm{S} \qquad (2)$$

La Ecuación 1 establece que la distancia directa entre dos puntos no supera la suma de las distancias de esos puntos con un tercero. La Ecuación 2 indica que dos puntos son idénticos si la distancia entre ellos es cero.

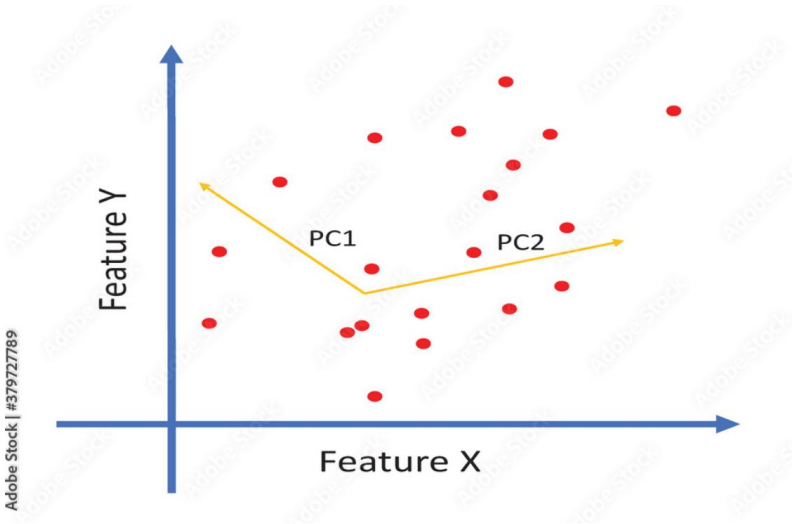
Componentes Principales

Según (Miranda, 2009), la metodología de los Componentes Principales se utiliza para simplificar la interpretación de conjuntos de datos grandes y complejos que involucran la medición simultánea de diversas variables en un grupo numeroso de individuos. Busca encontrar combinaciones lineales de las variables observables, conocidas como Componentes Principales, que resuman la información de manera eficiente, explicando las diferencias entre los individuos sin redundancias. Al representar gráficamente en un espacio estos Componentes Principales, es posible visualizar las relaciones entre variables y las similitudes entre individuos, facilitando la identificación de grupos homogéneos y mejorando la interpretación de la información recopilada. Se busca la máxima heterogeneidad dentro de los conglomerados y heterogeneidad mínima entre

los conglomerados (Arabie et al., 1996). Este enfoque es particularmente útil en la reducción de la dimensionalidad de conjuntos de datos, conservando la información relevante y eliminando redundancias.

Figura *2*

Componentes principales



Nota. Esta imagen presenta una distribución de puntos rojos en un espacio bidimensional definido por las características X e Y. Las flechas etiquetadas como PC1 y PC2 representan las dos primeras componentes principales. *Fuente*. Sidartha (s.f.).

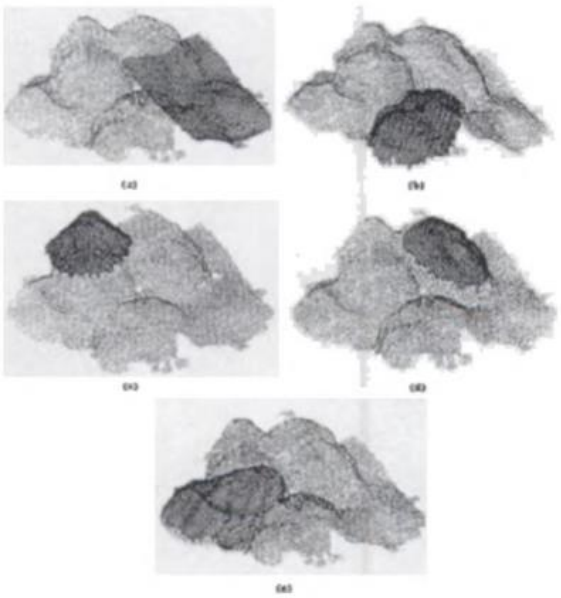
Segmentación de Datos y Selección de Modelos para Visión por Ordenador:

En el campo multidisciplinario de la visión por computadora, como señala (Bab-Hadiashar & Suter, 2000), el objetivo es capacitar a las máquinas para interpretar y comprender información visual. La segmentación de datos aplicada a imágenes implica dividir imágenes o secuencias de imágenes en regiones significativas, es esencial para el análisis de contenido

visual. Esto se centra en cómo los métodos estadísticos pueden guiar la segmentación de datos y la selección de modelos en visión por computadora.

Figura *3*

Modelos para visión por ordenador



Nota. Esta imagen presenta diferentes modelos utilizados en el campo de la visión por ordenador. Cada subimagen representa una variación del modelo, ilustrando distintos aspectos y técnicas de procesamiento de imágenes. *Fuente*. Bab-Hadiashar y Suter (2000).

El método propuesto por (Zhang & Kuo, 2001) para la segmentación y clasificación automática de datos audiovisuales se centra en el análisis de contenido de audio, logrando una alta precisión en la clasificación de señales de audio. Esta técnica de segmentación puede ser relevante en SARLAFT, ya que proporciona una estrategia para identificar y categorizar eficientemente diferentes tipos de información, lo que puede ayudar a la detección y gestión de riesgos asociados a actividades ilícitas y financiación del terrorismo.

Métodos de Regresión Robusta

Mínimos Cuadrados Robustos (MCR)

A diferencia de la regresión lineal clásica, los MCR minimizan la influencia de valores atípicos mediante funciones de pérdida robustas, como la función de Tukey.

M-Estimadores

Estos métodos buscan minimizar una función de error que puede ser menos sensible a valores atípicos. Ejemplos incluyen el M-estimador de Huber y el M-estimador de bisagra.

Regresión por Cuantiles

Se centra en la estimación de cuantiles condicionales, siendo menos afectada por datos extremos.

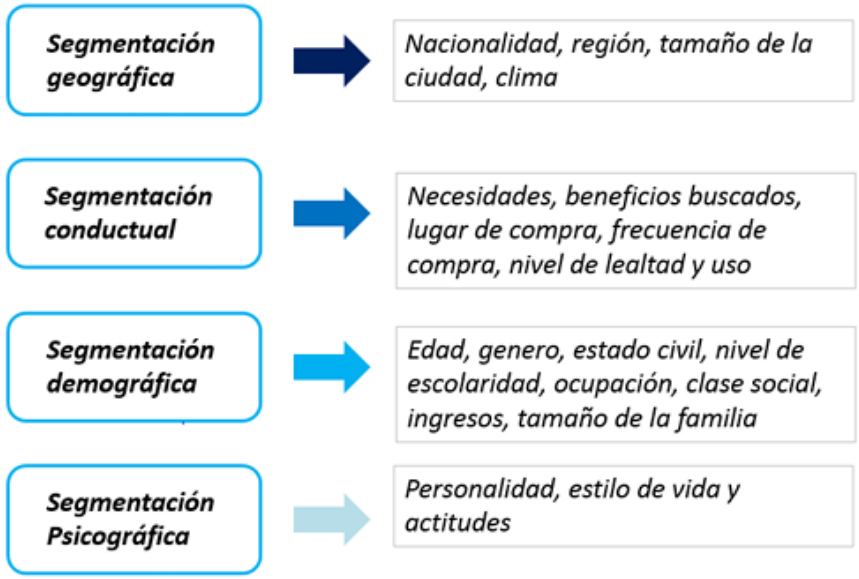
Segmentación de Clientes

La segmentación la define (Cáceres, 2020) como el proceso de dividir a los clientes en subgrupos, clasificándolos según una variedad de características o atributos, que incluyen la ubicación geográfica, la naturaleza jurídica, el valor y las necesidades, entre otros. También se puede entender la segmentación como una forma de simplificar la complejidad que implica tratar con numerosos clientes individuales, cada uno con necesidades y valor potencial distintos. Esta agrupación de clientes según sus rasgos únicos beneficiará a las organizaciones al permitirles gestionarlos de manera más efectiva, obteniendo resultados más exitosos en la oferta de productos y servicios. Cabe destacar que recientemente y de forma generalizada, se han incluido en los estudios de segmentación para el sector financiero, el análisis combinado de factores demográficos y psicográficos los cuales mundialmente están tomando bastante relevancia dados

los beneficios que aporta este enfoque en el desarrollo de productos, así como la posibilidad de reutilización de los resultados para otros estudios.

Figura *4*

Segmentación geográfica y psicográfica



Nota. La figura muestra los diferentes tipos de segmentación de mercado: geográfica, conductual, demográfica y psicográfica. *Fuente.* Valencia (2018).

En su estudio "Metodología de segmentación para el SARLAFT" (Pérez, 2020), el autor se adentra en la comparación de varios algoritmos, analizando su rendimiento y tiempo de ejecución. A través de un análisis meticuloso de diferentes escenarios, Pérez llega a la conclusión de que el método Clustering Large Applications (CLARA), fundamentado en la técnica de K-medoides (PAM), se destaca como la opción más efectiva en términos de eficiencia temporal para los diversos contextos presentados. en su investigación. Este hallazgo sugiere que, en la aplicación de la metodología de segmentación para el SARLAFT, la elección del método

CLARA puede ofrecer ventajas significativas en cuanto a la rapidez en los cálculos, contribuyendo así a una implementación más eficaz y eficiente.

En el estudio realizado por (Cáceres, 2020) “Efecto del segmento de clientes en el riesgo de liquidez de los fondos de inversión abiertos en Colombia” Se utilizo la recopilación de datos primarios, por lo tanto, el instrumento de recopilación de datos utilizado en este estudio es una encuesta. Las preguntas de la encuesta se adoptaron de estudios de investigación anteriores como los estudios de segmentación por actitudes (Phan et al., 2019) y los estudios de segmentación hibrida de (Makgosa et al., 2016). El cuestionario contiene dos secciones. La primera sección proporciona el perfil demográfico que examina los antecedentes personales de los participantes, como género, edad, ingresos, ocupación, ubicación, educación entre otros, los cuales son obtenidos directamente del formato de vinculación SARLAFT de los clientes de los fondos. La segunda sección del cuestionario examina el nivel de acuerdo de los encuestados con respecto del riesgo y otros aspectos, así como el comportamiento de inversión para los fondos. La escala de medición utilizada en esta investigación es la escala de Likert, y la razón para su uso es que resulta conveniente para el investigador pues ayuda a determinar los grupos de estudio.

En su investigación titulada "Segmentación de clientes y definición de alertas para la prevención de riesgos de lavado de activos y financiación del terrorismo (SARLAFT): un estudio económico aplicado a entidad financiera colombiana en 2017" (Amaya, 2017) presenta una propuesta metodológica que combina técnicas de minería de datos y análisis económico sectorial. Su conclusión destaca la efectividad de esta combinación, evidenciando que los resultados obtenidos proporcionan un enfoque eficaz para lograr el objetivo de identificar movimientos financieros inusuales con mayor especificidad y discriminación, tomando en cuenta

la naturaleza de las acciones analizadas. La metodología propuesta establece un contexto jerárquico que facilita la evaluación del riesgo asociado al lavado de activos y financiamiento del terrorismo en las transacciones realizadas por la entidad financiera estudiada hacia el mercado financiero en general. Según (López & Espinosa, 2021) el riesgo operacional es considerado uno de los factores que incurre en pérdidas monetarias para las entidades financieras, de las cuales se identifican detrimentos financieros originados por fallas o insuficiencias de procesos, personas, sistemas internos, tecnología, y, en la presencia de eventos externos imprevistos. Este enfoque contribuye significativamente a cumplir de manera más eficiente con el objetivo gubernamental de prevenir y detectar actividades ilícitas dentro de su sistema financiero.

En el estudio " La unidad de inteligencia financiera en el delito de lavado de activos por narcotráfico en la comisión contra el lavado de activos (CONTRALAFT)" (Linares, 2019), el autor se propuso analizar la influencia de la unidad de inteligencia financiera en el delito de lavado de activos por narcotráfico, específicamente en la comisión contra el lavado activos. Dado que (Gaitán, 2012), en su obra define el lavado de activos como la realización de una serie de actos u operaciones que buscan darle apariencia lícita a los ingresos y utilidades producto de actividades delictivas. Se entiende que el objetivo buscado mediante esta actividad es el de ocultar el origen de los recursos y la propiedad de estos. Los resultados obtenidos destacan la robustez de los instrumentos utilizados, con coeficientes de confiabilidad significativos: 0,89 para la variable independiente y 0,97 para la variable dependiente. La relación entre la unidad de inteligencia financiera y el delito en cuestión se revela como notable, con un coeficiente de 0.90, indicando una conexión significativa. Además, el respaldo del modelo de regresión lineal, respaldado por un P-Valor menor que 0,05 en la prueba de hipótesis, lleva al rechazo de la

hipótesis nula y la aceptación de la hipótesis general propuesta. En conclusión, se establece que la unidad de inteligencia financiera ejerce una influencia significativa en el delito de lavado de activos por narcotráfico en la comisión contra el lavado de activos CONTRALAFT 2018.

Metodología

La metodología propuesta se enfoca en los métodos de segmentación en SARLAFT, en el análisis de técnicas de aprendizaje automático como el Análisis de Componentes Principales (ACP), K-means y Árboles de Decisión, con el fin de evaluar su eficacia en la identificación y gestión de riesgos asociados a actividades ilícitas. Las conclusiones y recomendaciones resultantes entregaran una guía para futuras investigaciones y mejoras en las prácticas de segmentación en SARLAFT.

Identificación de Fuentes

En esta etapa, se llevará a cabo una búsqueda de literatura académica, informes gubernamentales y especializados relacionados con SARLAFT y segmentación. Se dará prioridad a los métodos de aprendizaje automático como el Análisis de Componentes Principales (ACP), K-means y Árboles de Decisión. Esta búsqueda se realizará con el objetivo de obtener una amplia gama de recursos que aborden la eficacia de los diferentes métodos de segmentación en la identificación y gestión de riesgos asociados a actividades ilícitas y financiación del terrorismo.

Análisis Descriptivo

Durante esta fase, se examinarán estudios que implementan enfoques de análisis descriptivo para comprender la distribución y características de los datos asociados a riesgos financieros en el contexto de SARLAFT. Esta revisión proporcionará una comprensión más

profunda de cómo se están utilizando los datos en el campo de SARLAFT y cómo estos enfoques pueden influir en la segmentación de riesgos.

Revisión Crítica de Métodos de Segmentación en SARLAFT

Durante esta etapa, el enfoque estará puesto en la recopilación y organización de los distintos métodos de aprendizaje automático utilizados en la segmentación de SARLAFT, como el Análisis de Componentes Principales (ACP), K-means y Árboles de Decisión. A través de una evaluación crítica de cada método identificado, exploraremos su idoneidad y efectividad en la detección y manejo de riesgos vinculados a actividades ilegales. Este análisis profundo nos permitirá determinar cuáles son los enfoques más prometedores y comprender las posibles limitaciones que podrían surgir en su aplicación.

Examen de Técnicas de Aprendizaje Automático para Segmentación

En esta fase, se pondrá un énfasis exhaustivo en comprender cómo estas técnicas pueden aplicarse en la práctica y en su capacidad para contribuir de manera efectiva a la identificación y gestión de riesgos en contextos financieros reales. Este estudio proporcionará una visión detallada sobre la viabilidad y efectividad de estos métodos en escenarios del mundo real.

Síntesis y Conclusiones

En la última fase, se integrarán los hallazgos de la revisión bibliográfica, el análisis descriptivo, la revisión crítica de métodos y el examen de técnicas de aprendizaje automático. Se presentarán conclusiones basadas en la evaluación crítica de los métodos de segmentación, identificando aquellos que demuestran ser más efectivos. Además, se propondrán recomendaciones para futuras investigaciones o mejoras en las prácticas de segmentación en SARLAFT.

Minería de Datos y SARLAFT

Si bien cada entidad tiene la autoridad para implementar el SARLAFT y existen unas regulaciones estándar regidas por el Grupo de Acción Financiera Internacional (GAFI), actualmente no existe una guía específica o detallada sobre la aplicación de la minería de datos en este proceso. Sin embargo, en esta monografía, nos proponemos ofrecer un posible paso a paso basado en una recopilación documental. Por ejemplo, (Ruiz, 2006) , exploró diversas técnicas de minería de datos para la detección del Lavado de Activos, destacando la taxonomía general de Knowledge Discovery in Databases (KDD) y tres etapas: preparación, extracción o minería y presentación. Por otro lado, (Hoyos, 2019), optó por utilizar la técnica de minería de datos CRISP-DM (Cross-Industry Standard Process for Data Mining) para desarrollar un modelo de segmentación de jurisdicciones que contribuya a la identificación de riesgos de lavado de activos y financiación del terrorismo. Esta misma técnica fue empleada por (Rodríguez & García, 2016), quienes la consideran la más utilizada en la comunidad científica; apoyada según un estudio llevado a cabo por el portal de análisis de datos KDnuggets en el 2014 ("Data Mining Community's Top Resource," 2014) y la aplican para resolver problemas de tipo no supervisado.

De acuerdo con (Amazon, s.f.), la minería de datos implica la exploración y análisis exhaustivo de grandes conjuntos de datos con el fin de identificar patrones, tendencias y relaciones relevantes. Este campo emplea una combinación de técnicas estadísticas, aprendizaje automático, inteligencia artificial y bases de datos para extraer información valiosa y conocimiento subyacente en los datos. La relevancia de la minería de datos radica en su capacidad para asistir a las organizaciones en la toma de decisiones informadas y respaldadas por evidencia.

En el ámbito del SARLAFT, según (Stradata, s.f.) las entidades financieras enfrentan la necesidad de adaptarse a esta realidad, puesto que necesitan establecer un mecanismo para gestionar el riesgo asociado al lavado de activos y financiamiento del terrorismo que les permita categorizar cada uno de los factores de riesgo de acuerdo con sus particularidades. Para alcanzar este propósito, es esencial comprender que las técnicas empleadas para la categorización solo resultarán efectivas si se dispone de una Base de Datos bien estructurada y se lleva a cabo un procesamiento previo óptimo de los datos.

El procesamiento de datos desempeña un papel fundamental en la extracción óptima de información proveniente de productos financieros, clientes, canales de distribución y jurisdicciones sujetas a análisis dentro del contexto de SARLAFT. Esto se debe a que, en numerosas ocasiones, la información proveniente de estos recursos se encuentra influenciada por una serie de factores como interferencias, datos faltantes, discrepancias, redundancias y una gran cantidad de datos en términos de variables y registros (Amaya, 2017). El procesamiento de datos juega un papel crucial en la mejora de la confiabilidad de la información utilizada para establecer las metodologías necesarias en la segmentación de los factores de riesgo utilizada por el SARLAFT. En actividades como la Minería de Datos, el éxito está estrechamente ligado a garantizar la integridad, completitud y consistencia de los datos mediante un control de calidad riguroso.

Una vez que se comprenden los precedentes que respaldan la implementación de modelos de identificación de riesgos, es fundamental destacar que hay una escasez notable de estudios que detallen una metodología específica para la segmentación de factores de riesgo de lavado de activos y financiamiento del terrorismo (LA/FT). Esto incluye investigaciones como aquellas que

examinan las metodologías de minería de datos pertinentes para el LA/FT, como mencionadas por el (Financiero, 2023)

Factores de riesgo LA/FT

En la gestión de riesgos LA/FT, se identifican cuatro elementos que pueden desencadenar la aparición de tales riesgos:

Cliente

Individuo o entidad legal con la que se establece una relación contractual en ámbitos financieros, económicos o comerciales.

Productos – Servicios

Aquellos ofrecidos durante la relación contractual.

Canales de Distribución

Medios utilizados para la transacción, como internet, telefonía, dispositivos móviles, cajeros automáticos y acceso remoto.

Jurisdicción o Zona Geográfica

Áreas donde se ofrecen los productos o servicios, ya sea a nivel nacional o internacional. Es crucial considerar aspectos como los índices de delincuencia, características económicas y financieras, y datos sociodemográficos.

Las estrategias de minería de datos más eficaces para la aplicación de un SARLAFT son aquellas que abarcan técnicas descriptivas como la segmentación de observaciones (k-medias) y técnicas predictivas, que buscan identificar las características que determinan la composición de los segmentos (Ruiz, 2006). A partir de este tipo de documentos y otras fuentes de lectura se ha

determinado que algunas técnicas resultan efectivas para lograr segmentos homogéneos internamente y heterogéneos externamente, cumpliendo así con los requisitos legales.

Las variables más utilizadas para estas técnicas consagradas en la Circular Básica Jurídica de la (Superintendencia Financiera, 2016), específicamente en la Parte I, Título IV (la cual compone la regulación colombiana relativa al tema) según (FONAVIEMCALI, 2020) estas son:

Edad

La edad al ser un número que indica ciertas características de las personas es importante para agruparlas en conjuntos ya que nos permite comparar cómo se comportan en términos económicos en función de su madurez, esto al final es muy importante a la hora de construir los clústers.

Información Financiera

Ingresos mensuales, gastos mensuales, bienes, deudas y activos netos, todos proporcionados por el cliente y respaldados por él mismo.

Frecuencia de Transacciones

Se refiere a la cantidad de veces al año que los clientes llevan a cabo transacciones financieras.

Valor de Transacciones

Esta medida representa la cantidad total de transacciones que realizan los clientes en un año. Se calcula considerando el promedio ponderado de las transacciones que involucran movimientos financieros dentro de la entidad, utilizando fondos entregados previamente.

En relación con la metodología empleada para desarrollar un modelo estadístico de segmentación utilizando técnicas de minería de datos se recurre a métodos de clusterización y árboles de decisión (Amaya, 2017). Dado que el propósito de las metodologías de segmentación es agrupar objetos similares, se requiere de una medida para evaluar las diferencias y similitudes entre ellos. El concepto de similitud es esencial en el Análisis de Clúster. La similitud se define como una medida de correspondencia o semejanza entre los objetos que serán agrupados; la estrategia más común implica medir la equivalencia en términos de la distancia entre los pares de objetos. Esta distancia, generalmente conocida como distancia euclidiana, se calcula midiendo el espacio entre un par de objetos basándose en los valores de sus variables y una métrica específica. Este proceso se lleva a cabo mediante la aplicación del teorema de Pitágoras.

Por tanto, para lograr un procesamiento de datos efectivo y cumplir con eficacia los requisitos del SARLAFT, se sigue la metodología CRISP.DM, que se divide en seis fases:

Comprensión del Negocio

Esto requiere no solo comprender las ramificaciones legales y financieras de los riesgos de LA/FT, sino también reconocer la importancia de implementar medidas preventivas efectivas. Al comprender a fondo las necesidades y metas de la entidad, se puede diseñar e implementar estrategias y procesos que contribuyan a mitigar estos riesgos y fortalecer la integridad del sistema financiero.

Comprensión de los Datos

En su última actualización, International Business Machines Corporation (IBM, 2021) menciona los procesos involucrados en esta etapa:

Se inicia la recolección de datos desde diversas fuentes como transacciones financieras, encuestas y registros web. Se evalúa la idoneidad de estos datos para cumplir con los requerimientos del SARLAFT y se considera la inclusión de información adicional, como datos demográficos, para mejorar la precisión del análisis. Se explora el uso de técnicas de aprendizaje automático para enriquecer la comprensión de datos y facilitar la detección de patrones y anomalías asociadas al lavado de activos y financiación del terrorismo. En caso de insuficiencia de datos, se investigan métodos para adquirir datos complementarios, asegurando un análisis exhaustivo y preciso del riesgo.

A continuación, se procede a describir los datos, abordando diversos aspectos relacionados con su cantidad, calidad y características. Se realiza una evaluación de la cantidad de datos disponibles, teniendo en cuenta tanto la cantidad de registros como los atributos o campos presentes en los conjuntos de datos. Es importante lograr un equilibrio adecuado entre la cantidad de datos y la precisión del modelo, ya que conjuntos de datos demasiado extensos pueden provocar un aumento en el tiempo de procesamiento; puede optar por trabajar con subconjuntos de datos que mantengan un equilibrio óptimo. Además, se lleva a cabo un análisis de los tipos de valores presentes en los datos, tales como numéricos, categóricos (en forma de cadenas de caracteres) o booleanos (con valores verdadero/falso). Prestar atención a estos tipos de valores puede ayudar a evitar posibles complicaciones durante la fase de modelado y garantizar la integridad de los resultados obtenidos en el SARLAFT. Asimismo, se examinan los esquemas de codificación utilizados en los datos, especialmente aquellos relacionados con características específicas del SARLAFT, como género o tipo de transacción. Identificar y

resolver posibles conflictos en los esquemas de codificación puede garantizar la exactitud y coherencia del análisis de riesgos.

La exploración de los datos utilizando diversas herramientas y técnicas disponibles como tablas, gráficos y demás estilos de visualización. Estos análisis proporcionan una visión de los datos también facilita la formulación de hipótesis y la identificación de patrones o tendencias relevantes para la detección y gestión de riesgos en el contexto del SARLAFT y la aplicación de medidas preventivas adecuadas. Durante este proceso, es fundamental realizar un análisis detallado de la base de datos para obtener estadísticas descriptivas relevantes, tales como el total de datos, medidas de tendencia central (como la media), valores mínimos y máximos, desviación estándar, y tipos de variables, entre otros. Este análisis de (Tufféry, 2011) nos proporcionará una comprensión general de nuestros datos.

En la verificación de calidad de datos los datos suelen contener fallos de codificación, valores perdidos etc. que pueden dificultar el análisis. Es recomendable llevar a cabo un análisis de calidad de los datos accesibles antes de implementar modelado. Por ejemplo, es importante verificar los errores de codificación y los valores perdidos en los datos financieros relacionados con transacciones bancarias, al realizar la validación de datos se podría prestar especial atención a estas áreas específicas. Es esencial comprender el origen de las inconsistencias al verificar datos para detectar ruidos o vacíos en la información. Las causas más frecuentes incluyen:

Valores faltantes que reflejan una característica significativa del individuo analizado.

Valores que aparentemente no existen, pero en realidad están presentes.

Datos incompletos provenientes de diversas fuentes.

Preparación de los Datos

Como menciona (Rodríguez & García, 2016), esta etapa comprende la selección, depuración, construcción, integración y ajuste final de los datos. (IBM, 2021) también hace referencia a los diversos procesos involucrados:

La recopilación inicial por la entidad de información sobre transacciones financieras, clientes y proveedores proporciona la base para esta fase. Aquí, la selección de datos se centra en identificar qué información es relevante para la detección de posibles riesgos financieros o actividades ilícitas. Por ejemplo, en el contexto de la selección de elementos, se pueden incluir transacciones que superen ciertos umbrales monetarios o cuentas asociadas a jurisdicciones de alto riesgo. En cuanto a la selección de atributos, se pueden considerar características como la frecuencia de transacciones, el volumen de transferencias y la relación entre las partes involucradas para construir perfiles de riesgo más detallados. Al seleccionar elementos, es esencial incluir cuentas o transacciones que presenten características atípicas o anomalías en su comportamiento financiero. Se deben considerar cuidadosamente las variables que mejor caractericen las transacciones y las partes involucradas en términos de riesgo potencial.

Uno de los problemas comunes en la limpieza de datos, especialmente relevante en el contexto del SARLAFT son los datos perdidos, que pueden surgir debido a la falta de información o a errores en la recolección de datos. En este caso, se puede aplicar la exclusión de filas o características con datos faltantes, o bien, completar los valores faltantes con estimaciones razonables que permitan mantener la integridad del conjunto de datos. Tras identificar los datos ausentes o anómalos (Infolaft, 2023), en cualquier contexto, podemos abordarlos de las siguientes maneras:

Si es posible según el proceso de Machine Learning, se pueden dejar sin procesar.

Cuando el porcentaje de valores faltantes es significativo, puede ser más práctico eliminar toda la variable en cuestión.

Se puede optar por filtrar la variable, eliminando los datos anómalos, aunque esto puede distorsionar la información, por lo que es crucial conocer la causa de las anomalías de antemano.

Para reemplazar los valores faltantes: en variables numéricas, se sugiere usar un valor que mantenga la media o la varianza; mientras que, en variables nominales, es preferible sustituirlos por el valor más común.

Durante el proceso de limpieza de datos, es crucial abordar errores de datos, incoherencias de codificación y metadatos perdidos para garantizar la precisión y efectividad del análisis.

La construcción de nuevos datos puede realizarse de dos formas principales: la derivación de atributos, que implica la creación de nuevas columnas o características a partir de datos existentes, y la generación de registros, que consiste en la creación de nuevas filas de datos para representar eventos adicionales o casos relevantes. Por ejemplo, se pueden derivar nuevos atributos calculando ratios financieros o indicadores de riesgo a partir de datos existentes, mientras que la generación de registros puede implicar la inclusión de nuevas observaciones para representar situaciones específicas no contempladas en los datos originales.

Existen dos enfoques principales para la integración de datos: la fusión, que une conjuntos con registros parecidos y atributos diversos, la adición, que combina datos con atributos afines, pero registros distintos. Esta integración amplía la variedad de casos analizados y fortalece la capacidad para identificar patrones sospechosos, en el marco del SARLAFT. Para asegurar una integración precisa de datos y cumplir con los procesos AML (Anti-Money

Laundering) requeridos por SARLAFT para combatir delitos financieros y el lavado de dinero, es esencial evitar dos errores principales al consolidar bases de datos:

La fusión de dos o más perfiles individuales, lo que resulta en un individuo con características amalgamadas.

Mantener separadas múltiples fuentes de información sobre un mismo individuo, generando varios registros con fragmentos de información de un solo individuo. Por ejemplo, al consolidar las compras de un individuo durante un período específico en un modelo RFM, el total puede ser menor si se considera a una persona real como dos entidades en la base de datos.

Una solución viable es descomponer las claves de identificación de los individuos (como el número de identificación, tarjeta de crédito, placa de vehículo, póliza de seguros, etc.), para establecer una clave única de unificación. Por ejemplo, se puede utilizar el número de identificación y el NIT, donde este último consta de la cédula seguida de un guion y un dígito adicional. Así se crea un identificador único que facilita la integración precisa y confiable de los datos.

Antes de generar un modelo, se debe evaluar si se requiere algún formato específico, por ejemplo, una vez que hemos identificado el tipo de variables con las que estamos trabajando, ya sea numéricas o nominales, podemos determinar si es apropiado transformarlas en función de sus características. Para convertir una variable numérica en una nominal, una opción es agrupar los valores en intervalos o "bines", creando así una representación ordinal que se puede emplear en el análisis. Por otro lado, en situaciones menos comunes, pero igualmente útiles, es necesario convertir variables nominales en numéricas (Aguilera, 2018). Esto puede ser necesario cuando el proceso de minería de datos o aprendizaje automático requiere datos numéricos en lugar de

nominales. En tales casos, se pueden generar múltiples variables "dummy", tantas como valores posibles tenga el atributo que estamos transformando. Estas variables "dummy" pueden tomar valores de 0 o 1, según si el registro incluye o no ese valor particular. También la clasificación de los datos para optimizar el rendimiento es importante, así como, en ACP se requiere que los datos estén en un formato adecuado, generalmente estandarizado y sin correlaciones falsas entre variables, para extraer las componentes principales de manera precisa y representativa, En el ámbito del análisis de datos, surge un desafío ampliamente conocido denominado "La maldición de la dimensionalidad" (Cosio, 2021). Este fenómeno se produce cuando nos encontramos con un elevado número de variables o características para un número reducido de individuos o registros, lo que obstaculiza la realización de procesamientos sólidos para la toma de decisiones. Una de las soluciones más eficaces y comúnmente empleadas para superar esta problemática es la técnica de análisis de componentes principales (ACP). Esta técnica implica la transformación de las variables originales en un nuevo conjunto de variables que sean mutuamente independientes (Méndez, 2010). Este proceso puede representarse gráficamente como una rotación de ejes en la proyección de los datos. Además del ACP, otra alternativa viable es recurrir al análisis factorial, que puede basarse en enfoques de mínimos cuadrados o de máxima verosimilitud. Estas técnicas se centran en las relaciones lineales entre las variables originales.

Modelado

En esta etapa se establece la conexión directa entre la segmentación en SARLAFT y las técnicas de aprendizaje automático para analizar y predecir los riesgos LA/FT. Se caracteriza por la iteración y experimentación con diversos modelos para obtener resultados óptimos, considerando los tipos de datos disponibles. Además, se evalúan los objetivos del SARLAFT,

desde la identificación de patrones sospechosos en las transacciones hasta la predicción de posibles actividades delictivas. Para la generación de un diseño de comprobación es necesario evaluar la eficacia de los modelos desarrollados. En este proceso, se definen criterios de "bondad" del modelo, como la precisión en la detección de actividades sospechosas o la capacidad de interpretación de los resultados. Se establecen también los conjuntos de datos que se utilizarán para validar los modelos, se debe garantizar robustez y fiabilidad.

Evaluación

Se examinan los resultados obtenidos de los modelos desarrollados, se considera precisión, eficacia y relevancia para los objetivos de la segmentación en SARLAFT. Por ejemplo, se evaluará si los modelos desarrollados son capaces de identificar de manera efectiva posibles actividades de lavado de activos o financiamiento del terrorismo, y si los hallazgos se pueden presentar de manera clara y comprensible. Por ejemplo, se puede revisar si existen patrones de comportamiento sospechoso que no fueron detectados por los modelos y se pueden proponer mejoras en los algoritmos utilizados para aumentar la precisión y sensibilidad en la identificación de dichos riesgos.

Despliegue

Para esta última etapa se asegura la efectiva implementación y operación continua de los modelos con el objetivo de mitigar los riesgos de LA/FT y garantizar el cumplimiento normativo de manera eficiente y efectiva. Se establecen los planes detallados para implementar y compartir los resultados, además, los procesos para monitorear y mantener la efectividad de los modelos desplegados a lo largo del tiempo.

Conceptualización de los Métodos con SARLAFT

Análisis de Componentes Principales

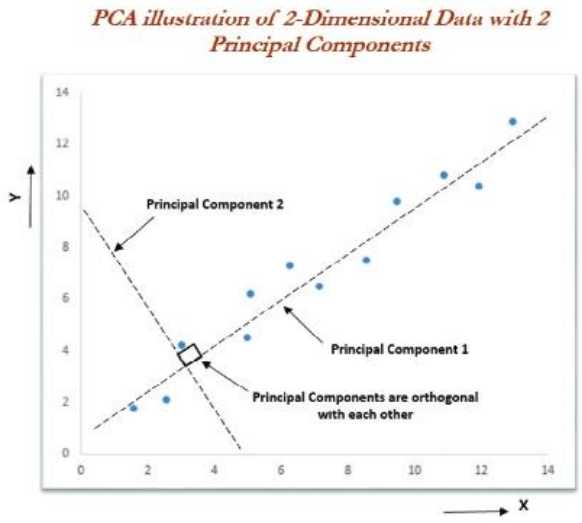
Un desafío crucial en el estudio de datos multivariantes es la simplificación de la estructura dimensional. Esta tarea consiste en representar con precisión los valores de varias variables utilizando un segmento más pequeño, lo que simplifica el problema sin perder demasiada información. El Análisis de Componentes Principales (ACP) se enfoca en este desafío. El método lo elaboro Hotelling en 1933, aunque sus raíces se remontan a Karl Pearson en 1901 y las adaptaciones propuestas por este (Peña, 2002). El desarrollo de esta técnica comienza con un conjunto de datos de varias variables inicialmente correlacionadas; en 1895 Karl Pearson definió la correlación la cual es una medida estadística que cuantifica la asociación o dependencia entre dos variables numéricas; estas representan diversas características o medidas. Seguido se busca encontrar un grupo reducido de variables que capturen la mayor cantidad de información. Las variables actuales se denominan componentes principales y se adquieren de combinaciones lineales de las variables primarias. Una característica fundamental de estas combinaciones es que no están correlacionadas entre sí, lo que permite una representación efectiva de los datos (Díaz, 2007). El ACP es una técnica que reduce la complejidad de los datos al proyectarlos en un espacio más pequeño (Dangeti, 2017).

La primera componente principal se obtiene maximizando la varianza, es decir, encontrando la dirección en la cual los datos se dispersan al máximo y se calcula mediante un proceso que implica encontrar de las variables originales los vectores y valores propios de la matriz de covarianzas; los valores propios representan la cantidad de variabilidad explicada por cada componente principal, mientras que los vectores propios indican la dirección en la que se

encuentra esa variabilidad máxima. La primera componente principal está vinculada al valor propio más grande, el cual es notablemente superior a los demás valores propios, explicando una proporción considerable de la variabilidad total de los datos (Novales, 2016).

Figura *5*

Componentes principales de datos bidimensionales



Nota. La figura ilustra datos bidimensionales con dos componentes principales. El componente principal 1 y el componente principal 2 son ortogonales entre sí, indicando que están en ángulos rectos, lo cual es característico del análisis de componentes principales *Fuente*. Dangeti (2017).

La segunda componente principal, como las demás, se describe como una combinación lineal de las variables de base (Perez C. , 2004). Para garantizar que la segunda componente principal capture la máxima variabilidad posible de los datos y esté no correlacionada con la primera componente principal, se utiliza un proceso de optimización. Esto implica encontrar los pesos óptimos para de manera que su varianza sea máxima, esté normalizado y sea ortogonal a la primera componente principal; dicho de otra forma, la segunda componente principal debe ser

perpendicular a la primera componente. El proceso de optimización se realiza mediante el enfoque de los multiplicadores de Lagrange, utilizando el producto escalar Euclidiano para asegurar la ortogonalidad.

Posteriormente, se utiliza ACP con el propósito de disminuir la complejidad de estos datos, proyectándolos en un espacio de menor dimensión sin perder la mayor cantidad de información posible. Al reducir la dimensionalidad, se simplifica la identificación de conexiones complejas entre variables (Vega et al., 2020) ya sean financieras y transaccionales, ACP desempeña un papel fundamental en la detección de patrones sutiles que podrían señalar actividades sospechosas, como transferencias de fondos inusuales o comportamientos atípicos en las transacciones. Además, el uso de este método posibilita la representación visual de patrones latentes de los datos, lo cual simplifica la interpretación de los resultados y permite a los analistas de SARLAFT tomar decisiones. Cuando una componente principal captura una gran cantidad de variabilidad en una dirección específica del espacio de variables, está revelando cómo los datos se comportan en esa dirección en particular. Este fenómeno puede señalar la presencia de un patrón o una tendencia fuerte que sigue esa misma dirección. Por ejemplo, al analizar las componentes principales derivadas de los datos de SARLAFT de una institución financiera, es posible que la primera componente principal muestre una alta variabilidad en una dirección asociada con la frecuencia y el monto de las transacciones realizadas en áreas geográficas específicas. Esta observación podría sugerir la existencia de un patrón significativo, como una actividad sospechosa concentrada en esas regiones geográficas. Del mismo modo, otra componente principal podría revelar una variabilidad considerable en una dirección relacionada con ciertos tipos de transacciones y sus montos, lo que podría indicar un comportamiento

anómalo digno de atención. Una guía para aplicar este método, también propuesta por (Cifuentes, 2020), se encuentra en el capítulo 1 de esta monografía.

Para diversos cálculos del análisis de componentes principales (Díaz, 2012) propone un código en R, se utilizan las funciones *cov(datos)* y *cor(datos)* para calcular la matriz de covarianzas y la matriz de correlaciones, respectivamente, a partir del conjunto de datos proporcionado. Una vez calculadas las matrices, se procede al análisis de componentes principales utilizando la función *dudi.pca(datos, scannf = FALSE, nf = ncol(datos))* de la librería *ade4*. Esta función realiza el análisis de componentes principales y devuelve un objeto que contiene información sobre los valores y los vectores propios. El código también proporciona visualizaciones de los resultados del análisis de componentes principales, utilizando las funciones *plot(acp$eig, type = "b") y s.label(acp$li, clabel = 0.7, grid = FALSE)*, se generan gráficos que muestran los valores propios. Por último, el código también incluye la generación de un círculo de correlaciones utilizando la función *s.corcircle(acp$co, grid = FALSE)*, este gráfico representa las correlaciones entre las variables originales y las componentes principales, lo que facilita la interpretación de cómo cada variable contribuye a la estructura de los datos. El código es el siguiente:

```
#para imprimir estadísticas básicas
cov(datos)
#matriz de correlaciones
cor(datos)
#vector de medias
mean(datos)
```

```
#desviaciones estándar

sapply(datos,sd)

#se carga la librería

library(ade4)

#análisis de componentes principales

acp<-dudi.pca(datos,scannf = FALSE,nf=ncol(datos))

#valores propios

inertia.dudi(acp)

#vectores propios

acp$c1

#gráfico scree

plot(acp$eig,type="b") # gr´afica de valores propios

#componentes principales

acp$li

#individuos sobre el primer plano factorial

s.label(acp$li,clabel=0.7,grid = FALSE)

#círculo de correlaciones

s.corcircle(acp$co,grid=FALSE)

#individuos sobre el primer plano factorial con biplot

s.label(acp$li,clabel=0.7,grid = FALSE,boxes =FALSE)

s.corcircle(acp$co,grid=FALSE,add=TRUE,clabel = 0.7)
```

K-Means

El método de k-means, también conocido como k-medias, es una técnica de agrupamiento no supervisado que busca dividir un conjunto de observaciones en K clusters, donde K es especificado por el usuario antes de aplicar el algoritmo. El objetivo es encontrar los clusters óptimos (MacQueen, 1967), minimizando la varianza interna de los datos la cual se define de la cohesión de los datos dentro de cada clúster, se refiere a agrupar los datos de manera que los elementos dentro de un mismo clúster sean similares entre sí. Según (Dangeti, 2017) esto se logra al minimizar la suma de las distancias entre cada observación y el centroide de su respectivo clúster, donde el centroide representa el punto medio o la media de todas las observaciones en el clúster (James y Witten, 2013). Normalmente se utiliza la distancia euclidiana como medida de distancia. La distancia euclidiana entre dos puntos en un espacio N-dimensional representados por los vectores x & y , esto se define como la longitud del segmento que une estos dos puntos en el espacio. La distancia euclidiana se calcula sumando las diferencias cuadráticas entre las coordenadas de los puntos en cada dimensión y luego tomando la raíz cuadrada del resultado. Esta fórmula matemática proporciona una medida intuitiva y efectiva de la separación o similitud entre dos puntos en un espacio multidimensional (Lloyd, 1982). Es importante mencionar que este algoritmo requiere que las variables de entrada sean numéricas o categóricas dicotómicas u ordinales previamente codificadas lo que implica convertirlas en representaciones numéricas adecuadas para el algoritmo. Este procedimiento se compone de tres fases fundamentales: la inicialización, la asignación de puntos a los centroides y la actualización de estos últimos.

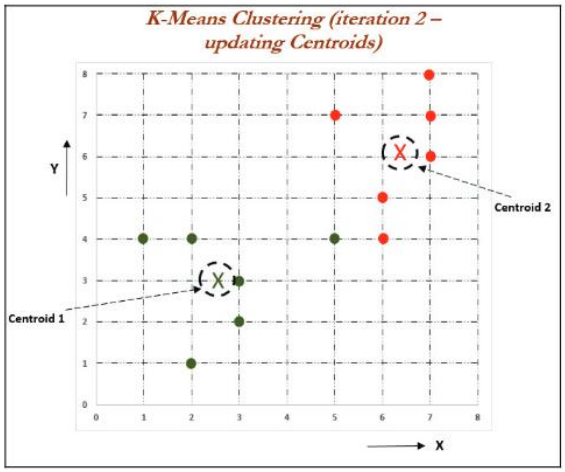
En el ámbito de SARLAFT la idea del algoritmo es de n observaciones extraer K grupos indicados por el usuario utilizando a los centroides como referencia. En el caso de K-means, este algoritmo de agrupamiento no supervisado puede requerir datos con características similares y bien definidas. Un caso de estos es la necesidad de normalizar variables que provienen de fuentes diferentes, como dólares y pesos, para que tengan la misma escala, incluso en técnicas que se basan en distancias como los componentes principales, es esencial normalizar el rango entre 0 y 1 para que la relación entre los valores permanezca constante. En tales situaciones, se recomienda estandarizar las variables antes de llevar a cabo el análisis de clusters mediante el algoritmo de k-medias (COOTRACHEC, 2017). Una vez finalizados los algoritmos de Machine Learning o minería de datos, (Aguilera, 2018) es aconsejable desnormalizar las variables para facilitar la interpretación y el análisis de los resultados. Siguiendo estos pasos de manera adecuada, podemos mejorar la precisión y completitud de los porcentajes de clasificación positiva y las reglas de decisión. Por otra parte, se recomienda haber elegido el número correcto de clusters y haber considerado todas las variables pertinentes son suposiciones fundamentales. Si el número de clusters seleccionados no es el adecuado o si se han dejado de lado variables importantes, los resultados pueden ser poco fiables o confusos.

Para implementarlo primero se toman al azar k clúster iniciales y se calculan los centroides (medias) de los clústeres. Recordemos que los centroides son puntos que representan el "centro" de cada cluster. Durante la ejecución del algoritmo, los centroides se actualizan iterativamente para reflejar el promedio de todas las observaciones asignadas a ese cluster. Los centroides se utilizan para calcular la distancia entre las observaciones y los clusters durante el proceso de asignación de puntos a los clusters (Bishop, 2006). Seguidamente se procede a

determinar la distancia euclidiana entre cada observación y los centroides de los clústeres, la cual se obtiene de medir el espacio que hay entre un par de objetos según los valores de sus variables y una métrica específica para luego a cada observación asignarle al clúster más cercano basado en esta distancia (Amaya, 2017) , lo que resulta en la formación de nuevos clústeres que se consideran una representación más precisa de los datos en comparación con los clústeres originales, después se determinan los centroides de los clústeres actualizados y se repiten los pasos 2) y 3) hasta que se cumpla un criterio de terminación, como, por ejemplo, la ausencia de cambios en la asignación de puntos, lo que indica que los clústeres obtenidos en dos iteraciones consecutivas son idénticos (COOTRACHEC, 2017).

Figura *6*

K-means clustering centroids



Nota. La figura ilustra la segunda iteración del algoritmo K-means para la actualización de centroides. Los puntos verdes y rojos representan los datos agrupados en dos clústeres diferentes, con sus respectivos centroides marcados con una "X" dentro de un círculo punteado. *Fuente*. Dangeti (2017).

Este enfoque de aprendizaje forma parte de los métodos de resolución de agrupamiento, los cuales son más complejos en comparación con los métodos de clasificación. ¿Por qué es relevante el agrupamiento? Bueno, se utiliza para extraer información de los datos, formular hipótesis, detectar anomalías e identificar características destacadas. También se emplea para evaluar la similitud entre formas u organismos (relación filogenética), o como un medio para organizar y resumir datos a través de prototipos de clúster. Los algoritmos de agrupamiento se dividen en dos categorías principales: los jerárquicos y los particionales. Los algoritmos jerárquicos forman grupos de manera ascendente o descendente, fusionando o dividiendo recursivamente los grupos según su similitud para crear una jerarquía. Por otro lado, los algoritmos particionales encuentran todos los grupos simultáneamente como una partición de los datos, sin imponer una estructura jerárquica. El propósito de K-means es minimizar la suma de los cuadrados dentro de todos los grupos definidos por K. No obstante, este algoritmo tiende a converger hacia mínimos locales y conforme el número de grupos K aumenta, el error cuadrático tiende a disminuir (con J = 0 cuando K = n). Una estrategia para mitigar los mínimos locales es ejecutar el algoritmo K-means con diversas particiones iniciales para un valor de K dado, seleccionando finalmente la partición con el menor error cuadrático. En este sentido, el algoritmo K-means requiere tres parámetros esenciales: el número de clústeres K, la inicialización de clústeres y la métrica de distancia, siendo K el más crítico de estimar. Por lo general, K-means se ejecuta de manera independiente para distintos valores de K, y se elige la partición que resulta más significativa según el criterio del experto en el campo. Sin embargo, hay múltiples métodos (Jain, 2010) para determinar el valor óptimo de k, incluyendo:

X-means (Pelleg, 2000) que encuentra K automáticamente al optimi-zar un criterio como el criterio de información de Akaike (AIC) o el criterio de información bayesiano (BIC).

Kernel K-means (Schölkopf, 1998)se propuso detectar grupos de formas arbitrarias, con una elección adecuada de la función de similitud del núcleo K-means; se usa típicamente con la métrica euclidiana para calcular la distancia entre puntos y centros de agrupación.

Criterios de longitud mínima de mensaje (MML) (Wallace y Boulton, 1968); (Wallace y Freeman, 1987) junto con el modelo de mezcla gaussiana (GMM) para estimar K. Su enfoque comienza con una gran cantidad de grupos, y gradualmente fusiona los grupos si esto conduce a una disminución en el criterio MML.

La validación cruzada implica evaluar el rendimiento del algoritmo utilizando modelos de mezcla obtenidos a partir de los datos en un pliegue específico, mientras que se utiliza la probabilidad de los datos en los pliegues restantes como una medida de su desempeño. Esta técnica puede ser empleada para determinar el número óptimo de grupos K en el modelo.

El método de Elbow también conocido como el método de codo, es una herramienta de interpretación y validación en análisis de conglomerados que ayuda a identificar el número adecuado de grupos en un conjunto de datos. Este método busca un error cuadrático mínimo y el "codo" representa generalmente el punto donde la suma de los errores cuadráticos comienza a disminuir a medida que aumenta el valor de K.

El SARLAFT, además de su rol en identificar operaciones sospechosas, se ha vuelto esencial para reducir riesgos y fortalecer la integridad financiera. La implementación de este sistema ha sido crucial para desarrollar estrategias proactivas que eviten el lavado de dinero y el financiamiento del terrorismo. Gracias a los avances en sistemas de información y las

actualizaciones en SARLAFT, los sistemas de gestión de riesgos se están robusteciendo con la automatización de procesos de cumplimiento. Así mismo la analítica y la tecnología han sido fundamentales en detectar patrones y tendencias emergentes, permitiendo respuestas más efectivas ante amenazas.

El algoritmo no supervisado K-means es ampliamente empleado y altamente interpretable en la segmentación o clasificación de clientes, especialmente en la industria bancaria. Esta técnica permite descubrir patrones de comportamiento comunes y distintivos entre grupos de clientes. Utilizar herramientas de análisis de datos para examinar el comportamiento crediticio conlleva automáticamente a adoptar una de las estrategias más efectivas para la gestión y mitigación de riesgos, lo cual impacta significativamente los resultados financieros de las entidades bancarias, potenciando sus ganancias. En (Edelman, 1992) ,se ilustran algunas de estas aplicaciones. Por ejemplo, se describe un estudio realizado para un banco comercial en Escocia, donde se examinó la morosidad de las cuentas mensuales a lo largo de un lapso de 2 años. Para este propósito, se empleó el método de agrupamiento de k-medoides con distancia euclidiana para categorizar los datos. El objetivo central de este análisis fue identificar grupos de clientes, así como combinaciones entre clientes y productos.

Para métodos de partición, (Díaz, 2012) propone un código en R utilizando la función *kmeans*. Este algoritmo se aplica para identificar tres grupos distintos dentro de un conjunto de datos con coordenadas en un plano. Finalmente, se genera un gráfico que muestra los puntos en diferentes colores según su grupo, y se etiqueta cada punto para una mejor visualización.

```
# introducción de datos
x<-c(1,2,4,5,3,3)
```

```
y<-c(2,1,1,4,5,3)
datos<-data.frame(x=x,y=y,row.names=LETTERS[1:length(x)])
# se invoca a la función kmeans()
cl<-kmeans(datos,centers=3)
# gráfico que muestra los grupos diferenciados por colores.
plot(datos, col = cl[["cluster"]])
# se identifica cada individuo en el gráfico.
text(datos,row.names(datos),pos=3)
```

Arboles de Decisión

Un árbol de decisión es una representación visual que muestra opciones disponibles y las consecuencias de cada decisión posible. Funciona como una serie de preguntas y respuestas organizadas de manera jerárquica, donde cada pregunta guía hacia nuevas opciones. Al final de cada camino, se llega a una conclusión o decisión basada en las respuestas a las preguntas anteriores. Es una herramienta útil para clasificar información y prever resultados en análisis de datos. Cuando la variable dependiente es continua, se emplean árboles de regresión, mientras que para variables cualitativas se utilizan árboles de clasificación. Estos árboles consisten en una serie de bifurcaciones anidadas que, al seguir cada rama, conducen a una predicción final sobre la clase de pertenencia o el valor de los individuos según los criterios iniciales establecidos. La construcción de los árboles de decisión se realiza mediante un algoritmo de segmentación recursiva, que procede paso a paso para definir la estructura del árbol. Entre los principales procedimientos se encuentran CHAID, QUEST y CART la cual evalúa si dividir un nodo mejora la pureza en una cantidad estadísticamente significativa. En particular en cada nodo, dividimos el

predictor con la asociación más fuerte con la variable resultado. La fuerza de la de asociación se mide mediante el valor *p* de una prueba de independencia de chi-cuadrado. Si para el mejor predictor la prueba no muestra una mejora significativa, no se realiza la división y se termina el árbol (Shmueli et al., 2017).

Figura *7*

Regression model tree

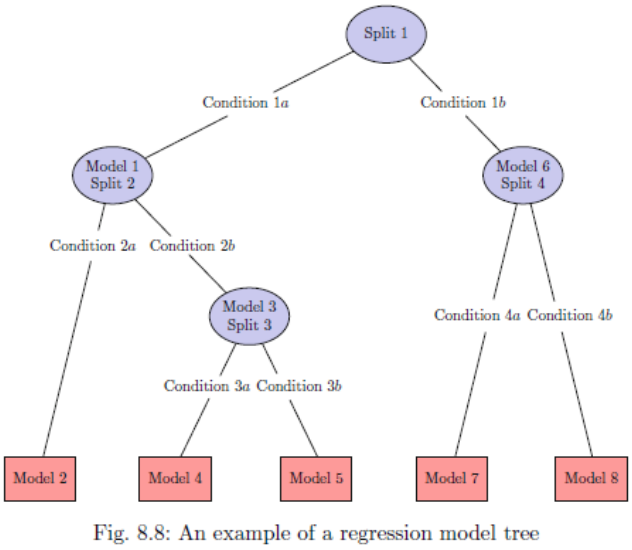


Fig. 8.8: An example of a regression model tree

Nota. Este árbol ilustra cómo un conjunto de modelos de regresión se puede organizar en una estructura de árbol basada en distintas condiciones de división. *Fuente*. Max y Johnson (2016).

Según (Lemus, 2008) los Árboles de Decisión son capaces de descubrir perfiles o conjuntos de características que describen a un grupo de individuos mediante un proceso de búsqueda no lineal, sino más bien comparable a una serie de divisiones sucesivas, similar a la estructura de un árbol. Estos árboles ofrecen flexibilidad al permitir la identificación relativamente sencilla de combinaciones, nichos ecológicos o perfiles de predictores asociados a valores específicos de la respuesta deseada. El principal objetivo consiste en predecir la

clasificación correspondiente a un sujeto con un determinado conjunto de valores en las variables explicativas; mientras que en un segundo objetivo se busca estimar la respuesta de interés Y asociada a cada nicho identificado. Los principios fundamentales en la construcción del modelo incluyen la simplicidad, la potencia y la estabilidad. La simplicidad es esencial, ya que es crucial que cualquier persona de diversas áreas pueda comprender y entender cómo funciona el modelo y qué predice. La potencia se refiere a la capacidad del modelo para distinguir de manera precisa entre clientes buenos y clientes malos. La estabilidad implica que el modelo mantenga su capacidad de discriminación a lo largo del tiempo y pueda detectar cambios. Para las características de un Árbol de Clasificación, (Martinez, 2014) menciona lo siguiente:

Hay un proceso iterativo de partición descendente.

Surge la existencia de una variable criterio dependiente (monotética) o de varias variables (politéticas).

Surge la existencia de variables independientes.

En (Geron, 2022) se utiliza un código en Python en el cual crea y entrena un modelo de árbol de decisión para realizar predicciones y evaluar su rendimiento utilizando el error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE).

```
#Importar y crear el modelo de árbol de decisión:
from sklearn.tree import DecisionTreeRegressor
tree_reg = DecisionTreeRegressor()
# Entrenar el modelo
tree_reg.fit(housing_prepared, housing_labels)
#Hacer predicciones en el conjunto de datos completo de entrenamiento
```

```
housing_predictions = tree_reg.predict(housing_prepared)

#Calcular el error cuadrático medio (MSE)

from sklearn.metrics import mean_squared_error

tree_mse = mean_squared_error(housing_labels, housing_predictions)

#Calcular la raíz del error cuadrático medio (RMSE)

import numpy as np

tree_rmse = np.sqrt(tree_mse)

#Mostrar el RMSE

tree_rmse
```

Evaluación de Ventajas y Limitaciones

Ventajas Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) presenta varias ventajas clave para simplificar conjuntos de datos con variables correlacionadas (Hurtado, 2016). En primer lugar, el ACP reduce significativamente la complejidad de los datos al eliminar la redundancia entre las variables, lo que resulta en una estructura más manejable y menos propensa a problemas de multicolinealidad (Perez C. , 2004). Además, al disminuir la dimensionalidad del conjunto de datos, el ACP permite conservar solo las componentes más importantes, facilitando así la interpretación y análisis sin perder información crucial. Esto es especialmente útil en escenarios de alta dimensionalidad, donde manejar y entender la estructura de los datos puede ser desafiante. Finalmente, la reducción a unas pocas componentes principales simplifica la aplicación de otras técnicas multivariantes, como la regresión o el clustering, haciendo el análisis más eficiente y preciso (Perez C. , 2004).

Una de las principales ventajas en SARLAFT radica en su capacidad para reducir la complejidad de conjuntos de datos financieros y transaccionales extensos, típicos en entidades financieras y de servicios. Al proyectar estos datos en un espacio de menor dimensión, se simplifica la identificación de patrones, tendencias y posibles irregularidades vinculadas con el LA/FT. Esta reducción dimensional permite a los analistas de SARLAFT realizar una exploración más eficiente de los datos, focalizando su atención en las variables y relaciones más relevantes. Otra ventaja importante del ACP en el contexto de SARLAFT es su capacidad para preservar la información relevante mientras minimiza la dimensionalidad de la colección de datos. Al capturar los principales factores de variabilidad de la información inicial, las

componentes principales derivadas del ACP retienen la mayor cantidad posible de información importante para la detección de actividades sospechosas. Esto garantiza que, a pesar de la reducción de la complejidad, el análisis resultante aún tenga una base sólida y representativa de los datos, lo que mejora la precisión en la identificación de patrones y comportamientos sospechosos.

Además, el ACP facilita la interpretación de resultados al proporcionar una representación visual de los patrones latentes de los datos. Esto es especialmente valioso en SARLAFT, donde la capacidad de interpretar y comunicar eficazmente los hallazgos es fundamental para llegar a decisiones e implementar estrategias proactivas contra el LA/FT. La visualización de componentes principales permite a los analistas identificar de manera intuitiva relaciones y tendencias en los datos, lo que agiliza el proceso de análisis y el reconocimiento de posibles áreas de riesgo. Por último, proporciona una base sólida para la aplicación de otras técnicas analíticas y modelos predictivos en el marco del SARLAFT. Al reducir la dimensionalidad de los datos y preservar la información relevante, el ACP prepara el terreno para la implementación de algoritmos más avanzados, como modelos de aprendizaje automático y análisis predictivo. Esto amplía las capacidades analíticas del SARLAFT y permite una detección más precisa y temprana de actividades sospechosas, fortaleciendo así la integridad financiera y la seguridad en el sector bancario y financiero.

Limitaciones Componentes Principales

El ACP asume linealidad en las relaciones entre variables (Jolliffe, 2002) señala que esto puede llevar a interpretaciones incorrectas si los datos tienen relaciones no lineales. Además, el mismo autor menciona que el ACP es sensible a la escala de las variables, requiriendo

estandarización o normalización previa para evitar distorsiones en los resultados. Asimismo, (Pearson, 1901) destaca la pégrdida de interpretabilidad como una limitación significativa del ACP, ya que las componentes principales son combinaciones lineales de las variables originales y pueden no tener un significado claro o intuitivo. Esto puede complicar la interpretación de los resultados y su aplicación práctica en contextos específicos. Por otro lado, (Jackson, 2005) subraya que el ACP requiere datos completos y libres de valores atípicos. La presencia de datos faltantes o atípicos puede afectar negativamente las componentes principales, disminuyendo la calidad y fiabilidad de los resultados.

Ventajas K-Means

Según lo que mencionan (Bishop, 1995) y (Kanungo, 2002) es considerado como uno de los algoritmos de aprendizaje no supervisado más básicos ya que aborda el problema bien conocido de la agrupación. Una ventaja principal es su capacidad para minimizar la distancia media cuadrática entre puntos de datos y sus centros, optimizando así la homogeneidad dentro de cada grupo. Además, la heurística de (Lloyd, 1982) facilita la selección del número óptimo de clústeres, mejorando la precisión de la segmentación. Este enfoque es particularmente beneficioso para grandes conjuntos de datos, ya que reduce los cálculos necesarios al considerar solo distancias hacia centroides cercanos. En el sector bancario, k-means es invaluable para segmentar clientes, identificando patrones de comportamiento y creando grupos homogéneos de manera clara y efectiva, lo que mejora significativamente el análisis y la toma de decisiones (Ramos, 2023).

En el contexto de SARLAFT, el algoritmo de K-Means emerge como una herramienta valiosa para la segmentación de datos financieros y transaccionales. Esta técnica no supervisada

permite identificar grupos o clusters dentro de conjuntos de datos extensos, lo que facilita la comprensión de patrones y la identificación de comportamientos atípicos. Una de sus ventajas clave es su capacidad para agrupar transacciones similares, lo que puede ayudar a detectar actividades sospechosas, como transferencias de fondos inusuales o comportamientos financieros atípicos. Al segmentar los datos, SARLAFT puede focalizar su atención en áreas específicas de riesgo, lo que contribuye a una detección más efectiva de posibles actividades ilícitas. Además, el algoritmo de K-Means ofrece una interpretación clara y efectiva al identificar patrones de comportamiento compartidos y distintivos entre los grupos de clientes. Esto es esencial en SARLAFT, donde la comprensión de las características de diferentes segmentos de clientes puede ayudar a diseñar estrategias de mitigación de riesgos más precisas y adaptadas. La capacidad del algoritmo para segmentar datos financieros y transaccionales según su similitud permite una exploración más profunda de las relaciones entre variables y una identificación más rápida de posibles anomalías.

Otra ventaja importante del algoritmo de K-Means en SARLAFT es su rendimiento en la ejecución de grandes volúmenes de datos. Dado que el sistema financiero maneja una cantidad considerable de transacciones diarias, la capacidad de K-Means para agrupar datos de manera rápida y precisa es fundamental para mantener la integridad del sistema y prevenir actividades ilícitas. Además, su naturaleza no supervisada significa que puede adaptarse dinámicamente a cambios en los patrones de transacciones sin requerir una supervisión constante. Finalmente, el enfoque de K-Means en la segmentación de datos financieros y transaccionales ofrece un panorama completo y detallado de la dinámica del sector y los comportamientos de los clientes. Al identificar grupos de transacciones similares, SARLAFT puede desarrollar estrategias de

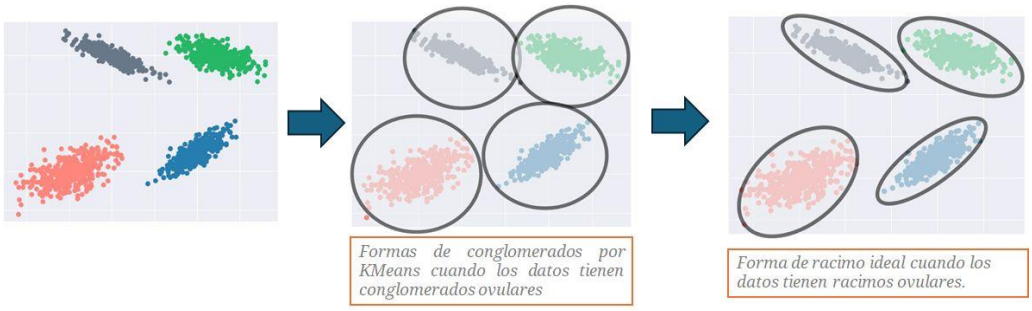
mitigación de riesgos más precisas y eficaces, lo que contribuye a la estabilidad y seguridad del sistema financiero en general.

Limitaciones K-Means

La efectividad del algoritmo de k-medias puede estar restringida por varios factores (Jiawei y Kamber, 2000), como la necesidad de predefinir el número de clusters, su susceptibilidad a valores atípicos y la asunción de que los clusters son isotrópicos, lo cual puede no ser apropiado para datos complejos o de alta dimensión, como los comúnmente encontrados en el ámbito financiero. Además, k-medias presupone una distribución de clusters esférica, lo cual puede no reflejar adecuadamente los patrones reales de comportamiento en datos relacionados con LA/FT, donde las relaciones pueden ser más intrincadas y los grupos pueden tener formas irregulares.

Figura 8

K-means conglomerados



Nota. Esta imagen presenta la comparación de formas de conglomerados generados por el algoritmo K-means cuando los datos tienen estructuras de conglomerados ovalares. La primera subimagen ilustra los datos iniciales, la segunda muestra los conglomerados formados por K-

means, y la tercera representa la forma de racimo ideal para datos con racimos ovalares. *Fuente*. Pérez (2024).

Por tanto, es crucial explorar otros algoritmos y técnicas de segmentación que puedan brindar resultados más sólidos según las características específicas de los datos en cuestión. El algoritmo también depende de la inicialización de los centroides, lo que puede llevar a resultados diferentes en cada ejecución y afectar la consistencia del análisis. Por tanto, es crucial explorar otros algoritmos y técnicas de segmentación que puedan brindar resultados más sólidos según las características específicas de los datos en cuestión

Ventajas Arboles de Decisión

Un instrumento valioso en el análisis de datos son los árboles de decisión, han ganado popularidad debido a su capacidad para generar reglas de decisión comprensibles y su facilidad de uso. Al ser obtenidos, los árboles determinan reglas que pueden aplicarse en diferentes contextos, lo que les permite desempeñar diversas funciones: En primer lugar, facilitan la segmentación de datos al identificar grupos significativos para clasificar elementos específicos. Esta capacidad de segmentación es fundamental para comprender la estructura subyacente de los datos y tomar decisiones informadas sobre cómo abordar diferentes grupos. Además, los árboles de decisión son útiles para la clasificación, asignando elementos a grupos previamente definidos en base a características compartidas. Esta función es especialmente útil en aplicaciones donde se deben tomar decisiones rápidas y precisas sobre la pertenencia de un elemento a una categoría específica (Bouza & Santiago, 2012). Además, menciona (Shmueli et al., 2017) que son buenos clasificadores y predictores ya que son útiles para la selección de variables, los árboles requieren relativamente poco esfuerzo por parte de los usuarios en los siguientes sentidos: Primero, no hay

necesidad de transformación de variables (cualquier transformación monótona de variables dará los mismos arboles). En segundo lugar, la selección de subconjuntos de variables es automática ya que forma parte de la selección divida. Los árboles también son intrínsecamente robustos a los valores atípicos, ya que la elección de una división depende del orden de los valores y no de las magnitudes absolutas de estos valores, sin embargo, son sensibles a los cambios en los datos e incluso un ligero cambio puede causar diversiones muy diferentes.

Los árboles de decisión en SARLAFT tienen la capacidad para segmentar datos en grupos significativos, lo que posibilita un entendimiento más a raíz del patrón subyacente de los datos. Al segmentar los datos, SARLAFT puede centrar su atención en áreas específicas que requieren una mayor vigilancia, lo que contribuye a una detección más efectiva de posibles actividades ilícitas. Además, son útiles para clasificar información y prever resultados en análisis de datos financieros y transaccionales. Su capacidad para asignar elementos a grupos predefinidos en función de características compartidas facilita la identificación de patrones de comportamiento compartidos entre diferentes segmentos de clientes. Otra ventaja importante de los árboles de decisión en SARLAFT es su capacidad para generar reglas interpretables y fácilmente comprensibles. Al proporcionar reglas claras sobre cómo se clasifican los datos y qué características son más relevantes para la segmentación, los árboles de decisión permiten una interpretación clara de los resultados y facilitan la comunicación de hallazgos a diferentes partes interesadas. Esta capacidad de traducir resultados complejos en reglas simples y fácilmente interpretables es fundamental para una implementación efectiva de estrategias de mitigación de riesgos en el sector financiero. Finalmente, los árboles de decisión son intrínsecamente robustos a los valores atípicos y pueden adaptarse dinámicamente a cambios en los datos sin requerir una

reconfiguración significativa del modelo. Esta capacidad de adaptación es crucial en SARLAFT, donde los patrones de transacciones pueden cambiar rápidamente y es fundamental detectar nuevas tendencias o comportamientos anómalos en tiempo real. Al ofrecer una herramienta flexible y adaptable para el análisis de datos financieros y transaccionales, los árboles de decisión contribuyen significativamente a la capacidad de SARLAFT para identificar y mitigar riesgos en el sistema financiero.

Limitaciones Arboles de Decisión

Los árboles de decisión, aunque populares por su simplicidad y capacidad interpretativa, presentan varias limitaciones que afectan su efectividad en el ámbito de SARLAFT. árboles de decisión, aunque populares por su simplicidad y capacidad interpretativa, presentan varias limitaciones que afectan su efectividad en el ámbito de SARLAFT. Una de las principales limitaciones es que las reglas de asignación son bastante sensibles a pequeñas perturbaciones en los datos, lo que puede llevar a resultados inconsistentes y poco fiables. Además, existe una dificultad inherente para elegir el árbol óptimo debido a la falta de una función global de las variables, lo que puede resultar en la pérdida de una representación completa y precisa del problema (Witten et al., 2002). Esta falta de una visión integral puede ser particularmente problemática en la identificación de patrones complejos y sutiles que son cruciales para la detección de actividades relacionadas con el lavado de activos y la financiación del terrorismo. Otra limitación significativa de los árboles de decisión en el ámbito financiero es la necesidad de un gran volumen de datos para asegurar que las observaciones en los nodos hoja sean significativas. Los datos financieros suelen ser vastos y variados y los árboles de decisión pueden llegar a contener demasiados datos irrelevantes, complicando su interpretación y potencialmente

introduciendo ruido en los resultados. Además, los modelos de árboles de decisión pueden ser inestables significando que pequeños ajustes en el conjunto de datos pueden generar cambios considerables en los resultados finales. Los usuarios deben ajustar cuidadosamente la profundidad del árbol estableciendo un número máximo de nodos o integrando diferentes resultados para evitar el sobreajuste. Estas limitaciones resaltan la necesidad de considerar métodos complementarios o alternativos que puedan ofrecer una mayor robustez y precisión en el análisis de datos financieros complejos y variados presentes en el ámbito de SARLAFT.

Méritos y Desafíos en Estudios Anteriores

Al momento de escoger el método de segmentación en SARLAFT, es necesario como se ha expresado en los capítulos anteriores, tener presente el tipo de datos con el que se cuenta, por ejemplo, (Ramos, 2023) en su investigación "Modelo de Segmentación para SARLAFT en R4G" selecciono los métodos K-Medoids, Algoritmo de Agrupamiento Jerárquico y Fuzzy c-means, por su capacidad para trabajar variables categóricas convertidas en dicotómicas, así como variables numéricas estandarizadas, y por su aptitud para manejar características específicas de los conjuntos de datos empleados en el proyecto. Los tres métodos hallaron conjunto de clientes similares además el algoritmo K-medoids demostró un rendimiento superior al obtener el coeficiente de silueta más alto (0.783) y el índice Dunn más elevado (3.147), indicando una mayor claridad y separación entre los clústeres.

Por otro lado (Pérez, 2020) en "Metodología para segmentación de un SARLAFT" selecciono los métodos K-Means, K-Medoids y CLARA comparando su tiempo de ejecución y eficacia. Utilizando K-Means, se identificó resultados más eficientes en términos de tiempo de procesamiento comparado con K-Medoids. Sin embargo, esta metodología no aprovechó todas

las variables disponibles en la base de datos. Finalmente, se propuso el algoritmo CLARA, que permitió utilizar tanto variables numéricas como categóricas, ofreciendo una segmentación más completa y eficiente.

En cuanto a (Amaya, 2017), su investigación aplica metodologías de análisis de clúster a los clientes de la entidad financiera para identificar segmentos homogéneos y emplea un modelo predictivo de clasificación basado en árboles de decisión sobre los resultados de la segmentación, destacando la importancia de combinar estos métodos para una mayor especificidad y discriminación en la identificación de movimientos sospechosos.

El análisis realizado por (Castro & Castro, 2020) , empleó el algoritmo de segmentación de k-medias para agrupar datos de asociados en un fondo de empleados para identificar los riesgos de LA/FT en tres grupos distintos, basándose en variables cuantitativas. En contraste, el ACP se utilizó para reducir la dimensionalidad de los datos.

Uno de los métodos aplicado por (Granados, 2019), para detectar operaciones sospechosas en LA, comienza con el algoritmo de K-Means para segmentar los datos, luego utiliza un árbol de expansión mínima sobre los grupos previamente formados. A través de un recuento de elementos en estas poblaciones, identifica los datos inusuales. Aunque su tasa de detección es limitada, ofrece resultados precisos y relevantes, lo que lo hace práctico para su aplicación, facilitando su análisis sin una gran carga de trabajo.

El estudio de (Moreno, 2023), se destaca por su enfoque principal en la aplicación del algoritmo de segmentación de K-medias y la técnica del análisis discriminante para clasificar a los clientes en segmentos definidos. Moreno enfatiza la importancia de explorar diversas

metodologías, incluyendo el uso de K-medias, para desarrollar modelos efectivos y adaptados a las necesidades específicas del entorno financiero.

Según (Liu et al., 2011), en su artículo presenta un algoritmo de árbol de decisión para identificar actividades de SARLAFT, combinando este enfoque con el algoritmo K-means. En primer lugar, se construye un árbol CF inicial utilizando BIRCH. Luego, se elabora el punto central de la hoja mediante el algoritmo K-means. Finalmente, se aplica el árbol de decisión central a los datos financieros y se modifica la posición del punto central durante el proceso. El objetivo es descubrir patrones típicos y reglas relacionadas con el blanqueo de capitales.

Liu menciona que cada algoritmo de agrupación tiene sus características y limitaciones. BIRCH es un algoritmo incremental adecuado para grandes bases de datos, pero no maneja bien los datos financieros. Por otro lado, K-means puede gestionar fácilmente datos financieros, pero resulta demasiado costoso para bases de datos grandes. Por lo tanto, un árbol de decisión central combinado con un algoritmo de agrupación puede identificar transacciones anómalas y proporcionar resultados significativos y análisis precisos. Es importante considerar que la eficacia de estos métodos está limitada si no se tienen en cuenta todos los datos financieros. La combinación de agrupación con otros algoritmos de minería de datos puede ofrecer mejores resultados. Por ejemplo, el uso de análisis de enlaces para buscar registros relacionados puede evitar muchas operaciones redundantes y mejorar la eficiencia. Además, es necesario disponer de un conjunto de casos de muy alta calidad para apoyar el método de clustering, lo que hace que la preparación de los datos sea relativamente difícil.

Conclusiones

Este análisis permite apreciar cómo la minería de datos, en el marco del Aprendizaje Automático, puede ser una herramienta poderosa para analizar grandes conjuntos de datos financieros y transaccionales en SARLAFT. Esto incluye comprender los procesos de preprocesamiento de datos, selección de características, entrenamiento de modelos y evaluación de resultados, lo que brinda una comprensión del proceso analítico y su aplicación en la gestión de riesgos. Al entender cómo se implementa la minería de datos en SARLAFT, dentro del marco del Aprendizaje Automático, se facilita una apreciación más exacta de la efectividad de diversos enfoques de aprendizaje automático en la identificación y gestión de riesgos asociados a actividades ilícitas y financiación del terrorismo.

El análisis de componentes principales (ACP) ha evidenciado ser un instrumento significativo en el contexto de SARLAFT, convirtiéndose una metodología efectiva para simplificar conjuntos de datos financieros y transaccionales extensos. Una de las principales enseñanzas derivadas de esta investigación es la capacidad del ACP para reducir la complejidad y la dimensionalidad de los datos al tiempo que conserva la información relevante para la detección de actividades sospechosas de (LA/FT). Este enfoque permite a los analistas de SARLAFT explorar de manera más eficiente los datos, centrándose en las variables y relaciones más significativas. El ACP contribuye a mejorar la detección temprana y precisa de posibles irregularidades financieras mediante la reducción de la redundancia y la preservación de la información clave. Al proporcionar una representación visual de los patrones latentes de los datos, el ACP facilita la interpretación de resultados y la identificación intuitiva de áreas de riesgo potencial. Además, la capacidad del ACP para crear las condiciones para el despliegue de

técnicas analíticas avanzadas, como modelos de aprendizaje automático y análisis predictivo, amplía las capacidades analíticas de SARLAFT. El análisis de componentes principales emerge como una herramienta fundamental en la contienda contra el (LA/FT) en el sector financiero y bancario. Su capacidad para reducir la complejidad de los datos, preservar la información relevante y facilitar la interpretación de resultados lo convierte en un elemento clave en la detección y mitigación de riesgos financieros, contribuyendo así a fortalecer la integridad y seguridad del sistema financiero en general.

En conclusión, el algoritmo de K-Means emerge como una herramienta crucial en SARLAFT debido a su capacidad para segmentar datos financieros y transaccionales, identificar patrones de comportamiento y detectar actividades inusuales o sospechosas. Esta técnica no supervisada ofrece una interpretación clara y efectiva al agrupar transacciones similares, lo que facilita la comprensión de las dinámicas del sector y la identificación de riesgos potenciales. Además, su rendimiento en la ejecución de grandes volúmenes de datos y su capacidad para adaptarse dinámicamente a cambios en los patrones de transacciones lo convierten en una herramienta invaluable para mantener la integridad del sistema financiero y prevenir actividades ilícitas. Sin embargo, es importante tener en cuenta las limitaciones del algoritmo, como la necesidad de predefinir el número de clusters, su susceptibilidad a valores atípicos y la asunción de distribución esférica de clusters, lo que puede requerir la exploración de otras técnicas de segmentación para datos complejos o de alta dimensión. En resumen, el uso eficaz del algoritmo de K-Means en SARLAFT requiere un profundo entendimiento de sus principios, preparación adecuada de datos y consideración de las características específicas de los datos en cuestión.

Como se indica al principio de este texto, la identificación de anomalías desempeña un rol crucial en la lucha contra el lavado de activos. Las metodologías desarrolladas permiten detectar comportamientos inusuales, descubrir estructuras complejas y analizar grandes volúmenes de datos de manera más eficiente y precisa que los métodos manuales convencionales, entre otros beneficios. Esto capacita a las entidades financieras para mejorar sus sistemas de supervisión y control, adoptando medidas proactivas para prevenir y combatir el lavado de activos.

Los árboles de decisión son una herramienta poderosa y versátil en el análisis de datos y la toma de decisiones, su capacidad para generar reglas claras y fácilmente interpretables los hace especialmente útiles en contextos donde se requiere comprender y explicar el proceso de toma de decisiones. En el contexto de SARLAFT tienen la capacidad para segmentar datos en grupos significativos, lo que posibilita un entendimiento más a raíz del patrón subyacente de los datos. Al segmentar los datos, SARLAFT puede centrar su atención en áreas específicas que requieren una mayor vigilancia, lo que contribuye a una detección más efectiva de posibles actividades ilícitas. Además, son útiles para clasificar información y prever resultados en análisis de datos financieros y transaccionales. Su capacidad para asignar elementos a grupos predefinidos en función de características compartidas facilita la identificación de patrones de comportamiento compartidos entre diferentes segmentos de clientes. Otra ventaja importante de los árboles de decisión en SARLAFT es su capacidad para generar reglas interpretables y fácilmente comprensibles. Al proporcionar reglas claras sobre cómo se clasifican los datos y qué características son más relevantes para la segmentación, los árboles de decisión permiten una interpretación clara de los resultados y facilitan la comunicación de hallazgos a diferentes partes

interesadas. Esta capacidad de traducir resultados complejos en reglas simples y fácilmente interpretables es fundamental para una implementación efectiva de estrategias de mitigación de riesgos en el sector financiero.

Recomendaciones

Una recomendación sería llevar a cabo una comparación entre al menos dos modelos de segmentación para garantizar una calibración óptima. Esta práctica permite evaluar qué modelo se ajusta mejor a las necesidades específicas del negocio, tanto a nivel estadístico como práctico. Es esencial comprender a fondo el contexto del negocio desde la fase inicial para seleccionar las variables adecuadas que influyan en los factores de riesgo pertinentes. Variables como actividad económica, ingresos y volumen transaccional pueden ser relevantes, pero es crucial adaptar la selección de variables a las particularidades de cada organización. Esta selección precisa contribuirá al éxito del modelo de segmentación al determinar la homogeneidad y heterogeneidad dentro de los segmentos identificados. Adicional sería establecer un ciclo de ejecución periódica de la segmentación, como, por ejemplo, una o dos veces al año para asegurar su actualización y relevancia continua. Además, se sugiere realizar análisis detallados y revisión de las características de los segmentos, así como la calibración de alertas en cada iteración. Esto es fundamental para adaptarse a posibles cambios en los resultados y ajustar las alertas existentes o definir nuevas según sea necesario, lo que constituye la esencia de la calibración del modelo de segmentación. Integrar los resultados de la segmentación al monitoreo y seguimiento transaccional permitirá detectar patrones que puedan indicar anomalías, y es importante parametrizar los patrones y la frecuencia de su ejecución. Por último, es crucial definir procedimientos detallados para el análisis de operaciones inusuales, asegurando una documentación exhaustiva de todo el proceso.

Aunque el ACP es una herramienta poderosa, explorar otras técnicas de reducción de dimensionalidad podría brindar una perspectiva más amplia sobre cómo abordar conjuntos de

datos financieros y transaccionales en SARLAFT. Técnicas como el T-SNE o el Uniform Manifold Approximation and Projection (UMAP) podrían ser evaluadas para determinar su idoneidad en la detección de actividades sospechosas. Junto con la implementación del ACP, es esencial estudiar métodos para interpretar los resultados de manera efectiva. Se recomienda investigar técnicas de visualización y análisis que faciliten la comprensión de los componentes principales y sus implicaciones en la gestión de riesgos en SARLAFT. Se sugiere realizar más casos de estudio y aplicaciones prácticas del ACP en entornos de SARLAFT. Estos casos de estudio podrían proporcionar ejemplos concretos de cómo se implementa el ACP y sus efectos en la identificación y gestión de riesgos.

Dado que la interpretación de los modelos es crucial en SARLAFT, se recomienda investigar métodos para mejorar la interpretabilidad de los árboles de decisión y facilitar la comprensión de cómo se toman las decisiones. Esto podría implicar el uso de técnicas de visualización de árboles, análisis de importancia de características y generación de reglas de decisión comprensibles. Para mejorar la precisión y la robustez de los modelos basados en árboles de decisión en SARLAFT, se sugiere investigar técnicas de optimización de hiperparámetros y ajuste del modelo. Esto podría implicar la búsqueda de cuadrícula, la validación cruzada o el uso de algoritmos de optimización bayesiana para encontrar la configuración óptima de parámetros del modelo.

Referencias

Aggarwal, C., & Reddy, C. (2014). *Data Clustering Algorithms and Applications* (Primera ed.). New York: Chapman and Hall/CRC.

Aguilera, P. (2018). *Análisis de datos en el SARLAFT*. https://aml.stradata.co/2019/11/14/analisis-de-datos-en-el-sarlaft-4-0/

Amaya, M. (2017). Segmentación de clientes y definición de alertas para la prevención de riesgos de lavado de activos. Bachelor's thesis, Universidad EAFIT.

Amazon. (s.f.). *AWS*. https://aws.amazon.com/es/what-is/data-mining/

Arabie, P., Hubert, L., & Soete, G. (1996). *Clustering And Classification.* World Scientific.

Bab-Hadiashar, & Suter. (2000). Segmentación de datos y selección de modelos para visión por computadora: un enfoque estadístico. Medios de ciencia y negocios de Springer.

Bayona, H. (2019). Money laundering in rural areas with illicit crops: empirical evidence for Colombia. *Crimen, derecho y cambio social , 72 (4), 387-417*.

Bishop. (1995). *Neural networks for pattern recognition.* Oxford university press.

Bishop. (2006). *Pattern recognition and machine learning.*

Bouza, C., & Santiago, A. (2012). La minería de datos: árboles de decisión y su aplicación en estudios médicos. *Modelación matemática de fenómenos del medio ambiente y la salud*, Tomo 2, 64-78,.

Cáceres, E. (2020). Efecto del segmento de clientes en el riesgo de liquidez de los fondos de inversión abiertos en Colombia. Bachelor's thesis, Universidad EAN.

Castro, N., & Castro, M. (2020). Segmentación de clientes en un fondo de empleados para identificar los riesgos de lavado de activos y financiación del terrorismo. Colombia: Fundación Universitaria Los Libertadores. Sede Bogotá.

Cifuentes, J. C. (2020). Herramientas del Business Analitycs en R: Análisis de Componentes Principales para resumir variables. *Universidad Icesi*.

COOTRACHEC. (2017). *manual SARLAFT – consejo de administración Acta 630.*

Cosio, N. A. (11 de Octubre de 2021). *La maldición de la dimensionalidad*. https://medium.com/@nicolasarrioja/la-maldici%C3%B3n-de-la-dimensionalidad-f7a6248cf9a

Dangeti, P. (2017). *Statistics for machine learning.* Packt Publishing.

Díaz, L. (2007). *Estadística multivariada: inferencia y métodos.* Universidad Nacional de Colombia.

Díaz, L. (2012). *Análisis estadístico de datos multivariados.* Bogotá, Colombia: Universidad Nacional de Colombia.

Edelman, D. B. (1992). An application of cluster analysis in credit control. *IMA Journal of Management Mathematics , 4(1),* 81-87.

Enríquez, D. (2019). Modelo Matemático para Estimar el Riesgo de Lavado de Activos por clientes de pequeñas Instituciones Financieras. Colombia: Universidad del Cauca.

Esteban, H., Robledo, J., & Capra, M. (2012). Lavado de activos Impacto económico–social y el rol del profesional en ciencias. Doctoral dissertation, Facultad de Ciencias Económicas-Universidad Nacional de Tucumán.

Fatf Gafi. (s.f.). *fatf-gafi*. https://www.fatf-gafi.org/en/home.html

Financiero, U. d. (2023). *UIAF*. https://uiaf.gov.co/sites/default/files/2023-12/articulos/archivos/ENR%20Digital.pdf

FONAVIEMCALI. (2020). *Manual De Politica y Procedimientos del SARLAFT.* Cali.

Gaitán, A. (2012). Análisis de riesgo en la toma de decisiones de administradores de bancos en la prevención y control del lavado de activos visto desde el contrato de mutuo, leasing, cuenta de ahorros y CDT. *Revista de Derecho Privado*.

Geron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly Media, Inc.

Granados, E. G. (2019). Máquinas de soporte vectorial y árboles de clasificación para la detección de operaciones sospechosas de lavado de activos. *Lámpsakos*, (21), 26-38. https://doi.org/https://doi.org/10.21501/21454086.2904

Gurrea, M. T. (2000). Análisis de componentes principales. Proyecto e-Math financiado por la Secretaría de Estado de Educación y Universidades (MECD).

Hoyos, N. L. (2019). Elaboración de un modelo de segmentación de jurisdicciones que aporte a la identificación de riesgos de Lavado de Activos y Financiación del Terrorismo por este factor en una institución microfinanciera de la ciudad de Popayán. Colombia.

Hurtado, D. (2016). *Introducción al Business Intelligence y al Big Data.*

IBM. (2021). *Guía de IBM SPSS Modeler. IBM Corporation*. https://www.ibm.com/docs/es/spss-modeler/saas?topic=overview-spss-modelersubscription

Infolaft. (12 de 02 de 2023). *Infolaft*. https://www.infolaft.com/el-poder-oculto-de-los-datos-feb-12

Jackson, E. (2005). *A user's guide to principal components.* John Wiley & Sons.

Jain, A. K. (2010). *Data clustering: 50 years beyond K-means* (Vol. 31(8)).

James, G., & Witten, D. (2013). *An introduction to statistical learning.* Springer.

Jiawei, H., & Kamber, M. (2000). *Data mining concepts and techniques.*

Jolliffe, T. (2002). *Análisis de componentes principales para tipos especiales de datos* . https://doi.org/(págs. 338-372)

Kanungo, T. (2002). *An efficient k-means clustering algorithm: Analysis and implementation.* IEEE transactions on pattern analysis and machine intelligence, 24(7), 881-892.

Lemus, S. (2008). Árboles de clasificación y regresión. 161-166.

Linares, L. (2019). La unidad de inteligencia financiera en el delito de lavado de activos por narcotráfico en la comisión contra el lavado de activos (CONTRALAFT).

Liu, R., Qian, X., Mao, S., & Zhu, S. (2011). *Research on Anti-Money Laundering Based on Core Decision Tree.*

Lloyd, S. (1982). *Least Squares Quantization in PCM.* IEEE Transac-tions on Information Theory, 28, 129-137.

López, C. E., & Espinosa, M. (2021). Riesgo operacional: comportamiento de sus factores en el sector bancario de Bogotá Colombia. *Revista Venezolana de Gerencia*, 26(6), 439-456.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations.* In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Madhulatha, S. (2012). An Overview on Clustering Methods. *Arxiv*.

Makgosa, R., Matenge, T., & Mburu, P. (2016). Hybrid segmentation in the financial services market: targeting saving consumers. *Family and Consumer Sciences Research Journal*.

MakZin. (s.f.). *Unsupervised machine learning models. Data clustering algorithms.* Adobe Stock. https://stock.adobe.com/es/images/unsupervised-machine-learning-models-data-clustering-algorithms/611859981?prev_url=detail

Mariño, G., & Chaparro, F. (2014). Determinantes en la prevención del riesgo para el lavado de activos y la financiación del terrorismo (LA/FT) en el sector real. *Scielo*, 7-35.

Martinez, T. L. (2014). Técnicas de análisis de datos en investigación de mercados. *Ediciones Pirámide*.

Martínez, W., Verán, N., & Humberto, J. (2022). *La innovación en la lucha contra el lavado de activos.* Colombia: Editorial Universidad del Rosario.

Max, K., & Johnson, K. (2016). *Applied Predictive Modeling.* Springer.

Méndez, Á. J. (2010). *Aplicaciones prácticas de Minería de Datos con IBM SPSS Modeler.*

Miranda, C. P. (2009). *Estadística multivariable.* Catalunya: Univ. Politèc .

Moreno, A. (2023). Elaboración de un modelo de segmentación para el factor de riesgo cliente en una entidad administradora de un sistema de pago de bajo valor. Colombia: Universidad EAFIT (Maestría en Administración de Riesgos).

Novales, A. (2016). *Modelos Factoriales.* Universidad Complutense.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. . *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11), 559-572.

Pelleg, D. &. (2000). *X-means: Extending k-means with efficient estimation of the number of clusters* (Vol. 1).

Peña, D. (2002). *Analisis de datos multivariante.* Cambridge: McGraw-Hill España.

Pérez. (2020). Metodología de segmentación para el SARLAFT.

Perez, C. (2004). *Tecnicas de Analisis Multivariante de Datos.* Madrid: Pearson Educación S.A.

Perez, J. (Abril de 2024). *linkedin*. https://www.linkedin.com/posts/juan-carlos-p%C3%A9rez-9687715b_sarlaft-segmentaci%C3%B3n-riesgos-activity-7175314858056253440-xnV2/?originalSubdomain=es

Phan, T., Rieger, M., & Wang, M. (2019). Segmentación de clientes financieros por actitudes y comportamiento una comparación entre Suiza y Vietnam. *Revista Internacional de Marketing Bancario*.

Ramos, N. (2023). *Modelo de Segmentación para SARLAFT en R4G.* Universidad del Rosario.

Rodríguez, & García. (2016). Adecuación a metodología de minería de datos para aplicar a problemas no supervisados tipo atributo-valor. *Revista Universidad y Sociedad*, 8(4), 43-53.

Rodríguez, J., & Galvis, I. (2008). Aproximación a los instrumentos administrativos para la prevención y el control del lavado de activos en Colombia. Cuadernos de Contabilidad.

Rokach, L., & Maimon, O. (2010). Data Mining and Knowledge Discovery Handbook. *Springer Nueva York*.

Ruiz, W. S. (2006). Técnicas de minería de Datos para la detección de Lavado de Activos. *Universidad Nacional de Colombia-Maestría Ingeniería Sistema y Computación*.

Schölkopf, B. S. (1998). *Nonlinear component analysis as a kernel eigenvalue problem. Neural computation* (Vol. 10).

Shmueli, G., Bruce, P. C., & Yahav, I. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R.* John Wiley & Sons.

Sidartha. (s.f.). *Example of PCA application in databases.* Adobe Stock. https://stock.adobe.com/es/images/example-of-pca-application-in-databases/379727789?prev_url=detail

Stradata. (s.f.). *stradata*. https://aml.stradata.co/2019/11/14/analisis-de-datos-en-el-sarlaft-4-0/

Superintendencia Financiera. (Abril de 2009). *Instrucciones Relativas a la Administracion del Riesgo de Lavado de Activos y de la Financiacion del Terrorismo.* https://www.superfinanciera.gov.co/loader.php?lServicio=Tools2&lTipo=descargas&lFuncion=descargar&idFile=22236

Superintendencia Financiera. (2016). *Instrucciones Relativas a la Administracion del Riesgo de Lavado de Activos y de la Financion del Terrorismo.*

Tufféry, S. (2011). *Data Mining and Statistics for Decision Making.* Wiley Series in Computational Statistics.

UNDOC. (Octubre de 2011). *Oficina de las Naciones Unidas contra la Droga y el Delito*. https://news.un.org/es/story/2011/10/1228761

Valencia, Y. (2018). Monografía sobre el análisis de la segmentación psicográfica en el marketing y su aplicabilidad en el campo digital. Bogota: Universidad Piloto de Colombia. chrome-

extension://efaidnbmnnnibpcajpcglclefindmkaj/http://polux.unipiloto.edu.co:8080/00004848.pdf

Vega, H., Castorena, C., Alejo, R., & Granda, E. (2020). Estudio comparativo de métodos para obtener el número de neuronas ocultas de un autocodificador en un contexto de alta dimensionalidad. *IEEE Latin America Transactions*.

Wallace y Boulton. (1968). An information measure for classification. The Computer Journal. *11 (2)*, 185-194.

Wallace y Freeman. (1987). Estimation and inference by compact coding. *49(3)*, 240-252.

Witten, I., Eibe, F., & Hall, M. (2002). *Data Mining Practical Machine Learning Tools and Techniques* (Vol. 1). Acm Sigmod Record.

Zhang, T., & Kuo, C.-C. (2001). Audio content analysis for online audiovisual data segmentation and classification. IEEE Transactions on speech and audio processing, 9(4), 441-457.