ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE SISTEMAS

MAESTRÍA EN SISTEMAS DE INFORMACIÓN

DESARROLLO E IMPLEMENTACION DE MODELOS DE SEGMENTACION DE CLIENTES BASADOS EN MACHINE LEARNING PARA DETECTAR RIESGOS DE LAVADOS DE ACTIVOS Y FINANCIACION DEL TERRORISMO. CASO DE ESTUDIO EN UNA ASEGURADORA.

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL GRADO DE MAGÍSTER EN SISTEMAS DE INFORMACIÓN MENCIÓN EN INTELIGENCIA DE NEGOCIOS Y ANÁLISIS DE DATOS MASIVOS

JORGE BRYAN QUISHPE OÑA
Jorge.quishpe@epn.edu.ec

DIRECTORA: Myriam Beatriz Hernández Álvarez, PhD.
myriam.hernandez@epn.edu.ec

2022

APROBACIÓN DEL DIRECTOR

Como director del trabajo de titulación “desarrollo e implementación de Modelos de Segmentación de clientes basados en Machine Learning para detectar riesgos de lavados de activos y financiación del terrorismo. Caso de estudio en una Aseguradora” desarrollado por Jorge Bryan Quishpe Oña, estudiante de la Maestría de Sistemas de Información mención Inteligencia de Negocios y Análisis de Datos Masivos, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la Defensa oral.

Myriam Beatriz Hernández Álvarez, PhD.
DIRECTORA

DECLARACIÓN DE AUTORÍA

Yo, Jorge Bryan Quishpe Oña declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

_____

Jorge Bryan Quishpe Oña

DEDICATORIA

Dedico esta tesis de manera muy especial y amorosa a mis padres que han sido mi motor de salir adelante todos estos años de estudio, que me enseñaron lo difícil que es ganarse la vida en el campo laboral y más cuando las oportunidades se presentan en mayor cantidad a personas con profesión. Por tal motivo dedico este título a ellos que son los verdaderos homenajeados por tan gran logro.

En segundo lugar, dedico esta tesis a mis hermanos mayores que han sido mi espejo todo este tiempo y no solo en lo académico si no en lo personal, en lo humanitario y en todos los valores y consejos que me han dado para que pueda ser mejor cada día, estoy orgulloso por la gran familia que tengo y por todo lo que hemos conseguido juntos.

Por último y no menos importante dedicado a la familia Oquendo Carrera dueños de la empresa Ingelsi Cia. Ltda que fue la primera empresa que me abrió las puertas para forjarme como profesional aun siendo un estudiante, que siempre han sabido reconocer el talento en jóvenes con gran potencial y brindan esas primeras oportunidades que no en todos lados se las encuentra, siempre he dicho y diré que Ingelsi es una academia de científicos de datos con lo mejor de lo mejor. Y de Andrés me llevo lo mejor, porque de él aprendí todas las bases que necesitaba para tener mejores oportunidades, agradezco el tiempo brindado y la paciencia, la gran amistad que hemos forjado con toda la familia y espero de todo corazón en un futuro no muy lejano poder volver a trabajar juntos.

ÍNDICE DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

Poder identificar actualmente transacciones monetarias sospechosas dentro de las entidades financieras resulta una tarea compleja y de un proceso extenso de seguimiento a los clientes, dentro de esto se involucra la experiencia de la organización y las medidas de control implementadas que se basan en reglas duras que limitan el libre comportamiento usual de los clientes. El presente proyecto de desarrollo permitirá evaluar el mejor escenario de los clientes en donde se identificarán transacciones con un alto riesgo LAFT detallando las características que mejor describen a estos grupos o segmentos. Todo el desarrollo se lo realizará a través de modelos matemáticos que clasificarán los movimientos inusuales de los clientes basado en aprendizaje automático y bajo su fase de implementación enviarán alertas a las entidades de control de la organización en archivos planos. Estos archivos incluyen el detalle de lo sucedido con un día de diferencia para la gestión de seguimiento y toma de acciones.

Palabras clave: Machine Learning, Segmentación, Clusters, Lavado de activos, Random Forest.

ABSTRACT

To be able to currently identify suspicious monetary transactions within financial institutions is a complex task and an extensive process of monitoring customers, which involves the experience of the organization and the control measures implemented based on hard rules that limit the usual free behavior of customers. The present development project will allow to evaluate the best scenario of the clients where transactions with a high LAFT risk will be identified detailing the characteristics that better describe these groups or segments. All the development will be done through mathematical models that will classify the unusual movements of customers based on machine learning and under its implementation phase will send alerts to the control entities of the organization in flat files. These files include the detail of what happened with a day's difference for follow-up management and action taking.

Keywords: Machine Learning, Segmentation, Clusters, Asset Washing, Random Forest.

1. INTRODUCCIÓN

1.1. Planteamiento del Problema

El lavado de activos actualmente es un delito que varias personas y organizaciones cometen para convertir el dinero ilícito en legal, este proceso es realizado generalmente por el narcotráfico con el fin de burlar el sistema y cobrar indemnizaciones millonarias para luego invertirlo en actividades económicas legales y así eliminar los rastros del dinero ilícito.

Dentro del sector financiero se presenta mayores índices de delitos de lavados de activos, así como en aseguradoras siendo esta la más vulnerable entre todos las ramas del sectores financieros. En las entidades bancarias se exige una póliza que respalde préstamos e inversiones y se convierte en un filtro difícil de romper lo que evita que se cometan estos delitos. La modalidad para lavar activos dentro de las aseguradoras es que seguros de vida sean adquiridos por personas que forman parte del narcotráfico ya que constantemente están en peligro, lo que genera una probabilidad muy alta de que mueran y se pueda pedir el desembolso de dinero de sus pólizas para lavar activos. En general, los clientes no habituales u ocasionales se caracterizan porque sus contactos con la institución financiera son puntuales, de modo que no se espera de ellos una nueva operación [1].

El sector bancario actualmente está en la obligación de mejorar sus mecanismos de seguridad para evitar ser utilizados por redes de lavados de activos y financiación del terrorismo (LAFT). Lo que hacen estas redes dentro de las entidades financieras es registrar empresas ficticias y mover dinero en diferentes cuentas, las mismas que pertenecen a personas con un historial crediticio y bancario en blanco y que generalmente sus movimientos no van acorde a sus características como clientes. Es decir, las transacciones que generan son extremadamente altas tomando en cuenta su actividad económica u oficio, por este motivo existen organismos que luchan contra este tipo de delitos y que mantienen una constante vigilancia a las entidades financieras.

Grupo de Acción Financiera Internacional (GAFI) es uno de los organismos más grandes y uno de los más comprometidos en luchar contra los delitos LAFT. GAFI se enfoca principalmente en proporcionar mecanismo y manuales que ayuden a organizaciones, entidades, aseguradoras y cualquier sector de valores que trabaje bajo la modalidad de riesgos para evitar el LAFT. Además, esta entidad promueve el desarrollo de nuevos estándares y medidas legales para combatir estas acciones ilícitas, por ende, los países actualmente están implementando sus propias medidas en el sector bancario basándose en conocimientos del cliente, monitoreo constante y el desarrollo de nuevas tecnologías que brinden alertas al detectar movimientos sospechosos. Pero esto no asegura que el LAFT vaya a desaparecer por completo ya que sigue habiendo una minoría de casos en donde si sucede estos delitos. Los nuevos mecanismos desarrollados para detectar riesgos de lavados de activos se centran en que las entidades bancarias como las aseguradoras tengan información detallada de sus clientes a fin de que se pueda comparar con su actividad económica tanto en bienes como en transacciones monetarias. Por ello, GAFI recomienda que con la suficiente información transaccional que se tenga se debe tener herramientas que permitan categorizar los riesgos en base a criterios validos como el riesgo geográfico, transacciones que realizan los clientes y productos que adquieren en el caso de aseguradoras, para así poder centrar sus acciones en mejorar estos grupos en donde existe mayor riesgo LAFT.

En el sector financiero, los esfuerzos contra el blanqueo de capitales (ALD) a menudo se basan en sistemas basados en normas [2]. Sin embargo, las vulnerabilidades se derivan de la relativa simplicidad de los conjuntos de reglas disponibles públicamente, lo que genera altas tasas de falsos positivos (FPR) y bajas tasas de detección [3]. Las técnicas de aprendizaje automático (machine learning: ML, por sus siglas en inglés), superan la rigidez de los sistemas basados en reglas al inferir patrones complejos a partir de datos históricos y pueden potencialmente aumentar las tasas de detección y disminuir los FPR.

El principio de segmentación o caracterización en clústers se encuentra dentro del área de la estadística y la matemática que básicamente es dividir en grupos a los sujetos u objetos de análisis en base a sus características de similitud que luego de ser categorizados son homogéneos entre ellos y heterogéneos entre los grupos, lo que permite centrar esfuerzos y recursos al segmento de mayor importancia o de mayor riesgo LAFT en nuestro caso de estudio.

1.2. Objetivos

El presente trabajo persigue los siguientes objetivos:

1.2.1. General

Desarrollar e implementar modelos de segmentación de clientes naturales y jurídicos utilizando machine learning que permitan la detección temprana de actividades ilícitas relacionadas a lavados de activos y financiación del terrorismo (LAFT), analizando cada uno de los factores de riesgo del caso de estudio.

1.2.2. Específicos

- Realizar una revisión sistemática de la literatura en temas relacionados a riesgos de lavados de activos y financiación del terrorismo (LAFT), modelos supervisados y no supervisados de segmentación y clasificación, automatización de procesos y seguridad de la información.
- Conocer los antecedentes del caso de estudio en la aseguradora sobre el monitoreo actual de sus factores de riesgo.
- Analizar los mecanismos que maneja la empresa para controlar y dar seguimiento a los movimientos sospechosos para establecer los estándares a seguir y diseñar un plan de trabajo sobre los factores de riesgo.
- Elaborar los modelos de segmentación en base al plan de trabajo y bajo las necesidades de la empresa, creando métricas inteligentes de detección de alertas y optimizando en paralelo la implementación de los modelos en sus sistemas internos.

- Evaluar los rendimientos de velocidad y respuesta de los modelos, así como la veracidad de la información arrojada para corrección y optimización de errores.

1.3. Alcance

El presente trabajo busca encontrar patrones de comportamiento en información transaccional de clientes dentro de una aseguradora que permita generar alertas tempranas que manifiesten posibles riesgos de lavados de activos dentro de la organización basados en las normativas de la superintendencia de bancos de Colombia, para lo cual se tiene planteado implementar modelos de machine learning inmersos en la infraestructura de la entidad permitiendo automatizar los procesos de alertas y ejecución de los modelos.

La implementación de la solución no considera recomendaciones de acciones a tomarse para reducir el nivel de actividades ilícitas relacionadas con el LAFT.

1.4. Marco teórico

1.4.1. Segmentación de clientes

La segmentación de clientes es el proceso de dividir el mercado objetivo en grupos para comprender mejor sus comportamientos actuales, evaluarlo, seleccionarlo y luego establecer una combinación de marketing adecuada que incluye el desarrollo de programas de adaptación para satisfacer sus necesidades específicas, esto permite agrupar a los que tienen puntos en común, así como a comprender mejor sus deseos, necesidades, barreras y comportamientos específicos [4].

De acuerdo con la teoría de la segmentación del mercado, [5] determina que “una organización puede apuntar a los segmentos que le proporcionarán el mayor valor, lo que le permite descubrir quién será su mejor cliente y luego personalizar su producto o servicio”. La definición de segmentación del mercado para [6] es “el acto de dividir un mercado amplio de consumidores, que generalmente consiste en

clientes existentes y potenciales en subgrupos, que se basan en un tipo elegido de características compartidas".

La segmentación de clientes tiene el potencial de permitir a las organizaciones dirigirse a cada cliente de la manera más eficaz. Utilizando la gran cantidad de datos disponibles sobre clientes (y clientes potenciales), identificando grupos discretos de clientes con un alto grado de precisión en función de indicadores demográficos, de comportamiento y de otro tipo [7]**.**

El enfoque correcto para el análisis de segmentación es agrupar a los clientes en grupos según las predicciones con respecto a su valor futuro total para una organización, con el objetivo de abordar a cada grupo (o individuo) de la manera más probable para maximizar ese valor futuro o de por vida. La segmentación precisa de los clientes implica el seguimiento de los cambios dinámicos y la actualización frecuente de nuevos datos, que con el tiempo se convirtió en una tarea desafiante que requería horas de examinar manualmente diferentes tablas y consultar los datos con la esperanza de encontrar formas de agrupar a los clientes [8].

Es necesario saber qué tipo de enfoque es el más eficaz para un grupo objetivo en particular e identificar las necesidades, deseos y expectativas de un segmento de clientes en particular para hacer coincidir los productos y la comunicación con estos factores. El resultado es una mayor satisfacción del cliente, así como la identificación de oportunidades estratégicas. Para un adecuado desarrollo de la segmentación de clientes surge el sistema inteligente de marketing que de acuerdo con [9] "consta de personas, procedimientos, software, bases de datos y dispositivos que se utilizan en la toma de decisiones y la resolución de problemas específicos". El sistema inteligente de marketing es un campo interdisciplinario que se relaciona con bases de datos, almacenamiento, aprendizaje automático, sistemas expertos (formalismos de representación del conocimiento), estadísticas e investigación operativa y visualización de datos. El objetivo común de integrar estos diferentes campos es extraer conocimiento de grandes bases y almacenes de datos [10].

Hay algunos procesos que se pueden implementar para proporcionar una calidad de datos mejorada para la segmentación de clientes. Uno de los aspectos importantes de la calidad de los datos es el concepto de asignación de recursos para administrar atributos para los clientes. Este recurso, generalmente llamado administradores de datos, es responsable de gestionar la configuración de un nuevo cliente, asegurándose de que se proporcionen todos los atributos críticos al cliente. Uno de los grandes desafíos en las organizaciones basadas en el cliente es el conocimiento del cliente, comprender la diferencia entre ellos y calificarlos utilizando nuevas tecnologías como el algoritmo de aprendizaje automático y el tratamiento de datos, que permite crear un marco muy poderoso para comprender mejor las necesidades y comportamientos de los clientes y actuar de manera adecuada para satisfacer sus necesidades [10].

Los algoritmos de agrupación en clústeres se han utilizado ampliamente para abordar la tarea de segmentación de clientes para su posterior etiquetado que permita una clasificación supervisada. Mientras tanto, las técnicas de visualización también se han convertido en una herramienta vital para ayudar a comprender y evaluar los resultados de la agrupación. La agrupación visual se puede considerar una combinación de estos dos procesos. Consiste en técnicas que son simultáneamente capaces, no solo de llevar a cabo las tareas de agrupamiento, sino también crear una representación visual que sea capaz de resumir los resultados del agrupamiento, ayudando así en la búsqueda de tendencias útiles en los datos. El aprendizaje automático, una clase de inteligencia artificial, puede investigar conjuntos de datos de clientes similares e interpretar los segmentos de clientes con el rendimiento más beneficioso y adecuado [9].

Los algoritmos de aprendizaje automático pueden ayudar a identificar subgrupos de clientes que serían difíciles de identificar mediante la intuición y el análisis manual de datos. Los datos de los clientes se pueden procesar mediante modelos de aprendizaje automático, que luego se pueden usar para encontrar tendencias repetidas en una variedad de variables. El valor de una cantidad óptima de clústeres

para datos de clientes es fácil de encontrar utilizando métodos de aprendizaje automático [6].

1.4.2. Aprendizaje automático (machine learning)

El aprendizaje automático, de acuerdo con [11] “es una rama de la inteligencia artificial (IA) y la informática que se centra en el uso de datos y algoritmos para imitar la forma en que los humanos aprenden, mejorando gradualmente su precisión”. El aprendizaje automático brinda a los sistemas la capacidad de aprender y mejorar automáticamente a partir de la experiencia sin ser programados explícitamente. Se centra en el desarrollo de programas informáticos que pueden acceder a los datos y utilizarlos para aprender por sí mismos [12].

El aprendizaje automático es un componente importante del creciente campo de la ciencia de datos. Mediante el uso de métodos estadísticos, los algoritmos se entrenan para hacer clasificaciones o predicciones, descubriendo información clave dentro de los proyectos de minería de datos. Estos conocimientos posteriormente impulsan la toma de decisiones dentro de las aplicaciones y los negocios, lo que idealmente impacta en las métricas de crecimiento clave. A medida que el big data continúe expandiéndose y creciendo, aumentará la demanda del mercado de científicos de datos, lo que requerirá que ayuden a identificar las preguntas comerciales más relevantes y, posteriormente, los datos para responderlas [13].

De acuerdo con [11] el sistema de un algoritmo de aprendizaje automático se compone de tres partes principales:

- Un proceso de decisión: en general, los algoritmos de aprendizaje automático se utilizan para hacer una predicción o clasificación. En función de algunos datos de entrada, que se pueden etiquetar o no, su algoritmo producirá una estimación sobre un patrón en los datos.

- Una función de error: una función de error sirve para evaluar la predicción del modelo. Si hay ejemplos conocidos, una función de error puede hacer una comparación para evaluar la precisión del modelo.

- Un proceso de optimización del modelo: si el modelo puede ajustarse mejor a los puntos de datos en el conjunto de entrenamiento, entonces los pesos se ajustan para reducir la discrepancia entre el ejemplo conocido y la estimación del modelo. El algoritmo repetirá este proceso de evaluación y optimización, actualizando los pesos de forma autónoma hasta alcanzar un umbral de precisión.

En [12], los autores establecen que el aprendizaje automático clásico a menudo se clasifica según la forma en que un algoritmo aprende a ser más preciso en sus predicciones. Hay cuatro enfoques básicos:

1. Aprendizaje supervisado: en este tipo de aprendizaje automático, los científicos de datos proporcionan algoritmos con datos de entrenamiento etiquetados y definen las variables que quieren que el algoritmo evalúe para las correlaciones. Se especifica tanto la entrada como la salida del algoritmo.

   El aprendizaje automático supervisado requiere que el científico de datos entrene el algoritmo con entradas etiquetadas y salidas deseadas. Los algoritmos de aprendizaje supervisado son buenos para las siguientes tareas:

   - Clasificación binaria: división de datos en dos categorías.
   - Clasificación multiclase: elegir entre más de dos tipos de respuestas.
   - Modelado de regresión: predicción de valores continuos.
   - Conjunto: combinación de las predicciones de múltiples modelos de aprendizaje automático para producir una predicción precisa.

2. Aprendizaje no supervisado: este tipo de aprendizaje automático implica algoritmos que se entrenan con datos no etiquetados. El algoritmo escanea

a través de conjuntos de datos en busca de cualquier conexión significativa. Los datos con los que se entrenan los algoritmos, así como las predicciones o recomendaciones que generan, están predeterminados.

Los algoritmos de aprendizaje automático no supervisados no requieren que se etiqueten los datos. Examinan los datos sin etiquetar para buscar patrones que se puedan usar para agrupar puntos de datos en subconjuntos. La mayoría de los tipos de aprendizaje profundo, incluidas las redes neuronales, son algoritmos no supervisados. Los algoritmos de aprendizaje no supervisados son buenos para las siguientes tareas:

- Agrupamiento: dividir el conjunto de datos en grupos según la similitud.
- Detección de anomalías: identificación de puntos de datos inusuales en un conjunto de datos.
- Minería de asociaciones: identificación de conjuntos de elementos en un conjunto de datos que ocurren juntos con frecuencia.
- Reducción de dimensionalidad: Reducción del número de variables en un conjunto de datos.

3. Aprendizaje semisupervisado: este enfoque de aprendizaje automático implica una combinación de los dos tipos anteriores. Los científicos de datos pueden alimentar un algoritmo mayormente etiquetado como datos de entrenamiento, pero el modelo es libre de explorar los datos por sí mismo y desarrollar su propia comprensión del conjunto de datos.

   El aprendizaje semisupervisado funciona por científicos de datos que alimentan una pequeña cantidad de datos de entrenamiento etiquetados a un algoritmo. A partir de esto, el algoritmo aprende las dimensiones del conjunto de datos, que luego puede aplicar a datos nuevos sin etiquetar. El rendimiento de los algoritmos suele mejorar cuando se entrenan en conjuntos de datos etiquetados. Pero el etiquetado de datos puede llevar mucho tiempo y ser costoso. El aprendizaje semisupervisado se encuentra en un término medio entre el desempeño del aprendizaje supervisado y la

eficiencia del aprendizaje no supervisado. Algunas áreas donde se utiliza el aprendizaje semisupervisado incluyen:

- Traducción automática: algoritmos de enseñanza para traducir idiomas basados en menos de un diccionario completo de palabras.
- Detección de fraude: Identificar casos de fraude cuando solo tienes unos pocos ejemplos positivos.
- Etiquetado de datos: los algoritmos entrenados en conjuntos de datos pequeños pueden aprender a aplicar etiquetas de datos a conjuntos más grandes automáticamente.

4. Aprendizaje por refuerzo: los científicos de datos suelen utilizar el aprendizaje por refuerzo para enseñar a una máquina a completar un proceso de varios pasos para el que existen reglas claramente definidas. Los científicos de datos programan un algoritmo para completar una tarea y le dan señales positivas o negativas a medida que descubre cómo completar una tarea. Pero en su mayor parte, el algoritmo decide por sí mismo qué pasos tomar en el camino.

   El aprendizaje por refuerzo funciona mediante la programación de un algoritmo con un objetivo definido y un conjunto prescrito de reglas para lograr ese objetivo. Los científicos de datos también programan el algoritmo para buscar recompensas positivas, que recibe cuando realiza una acción que es beneficiosa para el objetivo final, y evitar castigos, que recibe cuando realiza una acción que lo aleja más de su objetivo final. El aprendizaje por refuerzo se usa a menudo en áreas como:

   - Robótica: Los robots pueden aprender a realizar tareas en el mundo físico utilizando esta técnica.
   - Videojuegos: el aprendizaje por refuerzo se ha utilizado para enseñar a los bots a jugar una serie de videojuegos.

- Gestión de recursos: dados los recursos finitos y un objetivo definido, el aprendizaje por refuerzo puede ayudar a las empresas a planificar cómo asignar los recursos.

El aprendizaje automático ha visto casos de uso que van desde predecir el comportamiento del cliente hasta formar el sistema operativo para automóviles autónomos; sin embargo, para [13] existen ventajas y desventajas al momento de utilizar el aprendizaje automático:

- Cuando se trata de ventajas, el aprendizaje automático puede ayudar a las empresas a comprender a sus clientes a un nivel más profundo. Al recopilar datos de clientes y correlacionarlos con comportamientos a lo largo del tiempo, los algoritmos de aprendizaje automático pueden aprender asociaciones y ayudar a los equipos a adaptar el desarrollo de productos y las iniciativas de marketing a la demanda de los clientes. Algunas empresas utilizan el aprendizaje automático como motor principal en sus modelos de negocio. Uber, por ejemplo, utiliza algoritmos para emparejar conductores con pasajeros. Google utiliza el aprendizaje automático para mostrar los anuncios de viajes en las búsquedas.

- Pero el aprendizaje automático viene con desventajas. En primer lugar, puede ser costoso. Los proyectos de aprendizaje automático suelen estar impulsados por científicos de datos, que cobran altos salarios. Estos proyectos también requieren una infraestructura de software que puede resultar costosa. También está el problema del sesgo de aprendizaje automático. Los algoritmos entrenados en conjuntos de datos que excluyen a ciertas poblaciones o contienen errores pueden conducir a modelos inexactos del mundo que, en el mejor de los casos, fallan y, en el peor, son discriminatorios. Cuando una empresa basa los procesos comerciales centrales en modelos sesgados, puede sufrir daños regulatorios y de reputación.

Las plataformas de aprendizaje automático se encuentran entre los ámbitos más competitivos de la tecnología empresarial, con la mayoría de los principales proveedores, incluidos Amazon, Google, Microsoft, IBM y otros, compitiendo para inscribir a los clientes en servicios de plataforma que cubren el espectro de actividades de aprendizaje automático, incluida la recopilación de datos, preparación de datos, clasificación de datos, construcción de modelos, capacitación e implementación de aplicaciones. A medida que el aprendizaje automático continúa aumentando en importancia para las operaciones comerciales y la IA se vuelve más práctica en entornos empresariales, las guerras de plataformas de aprendizaje automático solo se intensificarán. La investigación continua sobre el aprendizaje profundo y la IA se centra cada vez más en el desarrollo de aplicaciones más generales. Los modelos de IA actuales requieren una amplia capacitación para producir un algoritmo altamente optimizado para realizar una tarea. Pero algunos investigadores están explorando formas de hacer que los modelos sean más flexibles y están buscando técnicas que permitan a una máquina aplicar el contexto aprendido de una tarea a tareas futuras y diferentes [12].

El aprendizaje automático se utiliza en diferentes sectores por varias razones. Los sistemas comerciales se pueden calibrar para identificar nuevas oportunidades de inversión. Las plataformas de marketing y comercio electrónico se pueden ajustar para proporcionar recomendaciones precisas y personalizadas a sus usuarios en función del historial de búsqueda en Internet de los usuarios o de transacciones anteriores. Las instituciones crediticias pueden incorporar el aprendizaje automático para predecir préstamos incobrables y construir un modelo de riesgo crediticio. Los centros de información pueden utilizar el aprendizaje automático para cubrir grandes cantidades de noticias de todos los rincones del mundo. Los bancos pueden crear herramientas de detección de fraude a partir de técnicas de aprendizaje automático. La incorporación del aprendizaje automático en la era digital inteligente es infinita a medida que las empresas y los gobiernos se vuelven más conscientes de las oportunidades que presenta el big data [11].

1.4.3. Lavado de activos y financiamiento del terrorismo

1.4.3.1. Lavado de activos

El objetivo de un gran número de actos delictivos es generar una ganancia para el individuo o grupo que los realiza. De acuerdo con la Agencia de los Estados Unidos para [14] “el lavado de dinero es el procesamiento de estas ganancias criminales para disfrazar su origen ilegal. Este proceso es de vital importancia, ya que permite al delincuente disfrutar de estos beneficios sin poner en peligro su origen”. La venta ilegal de armas, el contrabando y las actividades del crimen organizado, incluidas, el tráfico de drogas y las redes de prostitución y otras actividades, pueden generar enormes cantidades de dinero. Los esquemas de malversación, tráfico de información privilegiada, soborno y fraude informático también pueden producir grandes ganancias y crear el incentivo para “legitimar” las ganancias obtenidas ilícitamente a través del lavado de dinero [15].

Cuando una actividad delictiva genera ganancias sustanciales, el individuo o grupo involucrado debe encontrar una manera de controlar los fondos sin llamar la atención sobre la actividad subyacente o las personas involucradas. Los delincuentes hacen esto disfrazando las fuentes, cambiando la forma o moviendo los fondos a un lugar donde es menos probable que llamen la atención. En respuesta a la creciente preocupación por el lavado de dinero, la Cumbre del G-7 en París en 1989 estableció el Grupo de Trabajo de Acción Financiera sobre el lavado de dinero (GAFI) para desarrollar una respuesta internacional coordinada. Una de las primeras tareas del GAFI fue desarrollar Recomendaciones, 40 en total, que establecen las medidas que los gobiernos nacionales deben tomar para implementar programas efectivos contra el lavado de dinero [16].

Por su propia naturaleza, el lavado de dinero es una actividad ilegal llevada a cabo por delincuentes que ocurre fuera del rango normal de las estadísticas económicas y financieras. En la etapa inicial -o colocación- del lavado de dinero, el lavador introduce sus ganancias ilícitas al sistema financiero. Esto se puede hacer dividiendo grandes cantidades de efectivo en sumas más pequeñas menos

llamativas que luego se depositan directamente en una cuenta bancaria, o comprando una serie de instrumentos monetarios (cheques, giros postales, etc.) que luego se cobran y depositan en cuentas en otro lugar [17].

Una vez que los fondos han ingresado al sistema financiero, tiene lugar la segunda etapa, o estratificación. En esta fase, el lavador realiza una serie de conversiones o movimientos de los fondos para alejarlos de su origen. Los fondos pueden canalizarse a través de la compra y venta de instrumentos de inversión, o el lavador puede simplemente transferir los fondos a través de una serie de cuentas en varios bancos de todo el mundo. Este uso de cuentas muy dispersas para el lavado es especialmente frecuente en aquellas jurisdicciones que no cooperan en las investigaciones contra el lavado de dinero. En algunos casos, el lavador puede disfrazar las transferencias como pagos por bienes o servicios, dándoles así una apariencia legítima [18].

Después de haber procesado con éxito sus ganancias delictivas a través de las dos primeras fases, el lavador las traslada a la tercera etapa, la integración, en la que los fondos vuelven a ingresar a la economía legítima. El lavador puede optar por invertir los fondos en bienes raíces, activos de lujo o empresas comerciales. Como el lavado de dinero es una consecuencia de casi todos los delitos que generan ganancias, puede ocurrir prácticamente en cualquier parte del mundo. Generalmente, los lavadores de dinero tienden a buscar países o sectores en los que existe un bajo riesgo de detección debido a programas contra el lavado de dinero débiles o ineficaces. Debido a que el objetivo del lavado de dinero es devolver los fondos ilegales a la persona que los generó, los lavadores generalmente prefieren mover los fondos a través de sistemas financieros estables. La actividad de lavado de dinero también puede estar concentrada geográficamente según la etapa en la que hayan llegado los fondos lavados. En la etapa de colocación, por ejemplo, los fondos generalmente se procesan relativamente cerca de la actividad subyacente; a menudo, pero no en todos los casos, en el país donde se originan los fondos [18].

Con la fase de estratificación, el lavador puede elegir un centro financiero extraterritorial, un gran centro comercial regional o un centro bancario mundial, cualquier ubicación que proporcione una infraestructura financiera o comercial adecuada. En esta etapa, los fondos lavados solo pueden transitar cuentas bancarias en varios lugares donde esto se puede hacer sin dejar rastros de su origen o destino final. Finalmente, en la fase de integración, los lavadores pueden optar por invertir los fondos lavados en otros lugares si se generaron en economías inestables o lugares que ofrecen oportunidades de inversión limitadas [16].

La integridad del mercado de servicios bancarios y financieros depende en gran medida de la percepción de que funciona dentro de un marco de altos estándares legales, profesionales y éticos. Una reputación de integridad es uno de los activos más valiosos de una institución financiera. Si los fondos de actividades delictivas pueden procesarse fácilmente a través de una institución en particular, ya sea porque sus empleados o directores han sido sobornados o porque la institución no tiene la intención de verificar la naturaleza delictiva de dichos fondos, la institución podría verse atraída a la complicidad activa con los delincuentes y formar parte de la propia red criminal. La evidencia de dicha complicidad tendrá un efecto perjudicial en las actitudes de otros intermediarios financieros y de las autoridades reguladoras, así como de los clientes comunes [15].

En cuanto a las posibles consecuencias macroeconómicas negativas del lavado de dinero sin control, se pueden citar cambios inexplicables en la demanda de dinero, riesgos prudenciales para la solidez bancaria, efectos de contaminación en las transacciones financieras legales y mayor volatilidad de los flujos internacionales de capital y tipos de cambio debido a cambios transfronterizos imprevistos. transferencias de activos. Además, como premia la corrupción y el crimen, el lavado de dinero exitoso daña la integridad de toda la sociedad y socava la democracia y el estado de derecho. Los lavadores buscan continuamente nuevas rutas para lavar sus fondos. Las economías con centros financieros en crecimiento o en desarrollo, pero con controles inadecuados, son particularmente vulnerables ya que los países con centros financieros establecidos implementan regímenes integrales contra el lavado de dinero. La influencia económica y política de las organizaciones

criminales puede debilitar el tejido social, los estándares éticos colectivos y, en última instancia, las instituciones democráticas de la sociedad. En países en transición hacia sistemas democráticos, esta influencia criminal puede socavar la transición. Más fundamentalmente, el lavado de dinero está inextricablemente vinculado a la actividad delictiva subyacente que lo generó. El lavado permite que continúe la actividad delictiva [17].

1.4.3.2. Financiamiento del terrorismo

El financiamiento del terrorismo, de acuerdo con [19] menciona que "implica la solicitud, recaudación o provisión de fondos con la intención de que puedan ser utilizados para apoyar actos u organizaciones terroristas". Los fondos pueden provenir tanto de fuentes legales como ilícitas. Más precisamente, según el Convenio Internacional para la Represión de la Financiación del Terrorismo, una persona comete el delito de financiación del terrorismo si esa persona por cualquier medio, directa o indirectamente, ilícita y deliberadamente, proporciona o recauda fondos con la intención de que deben ser usados o con el conocimiento de que van a ser usados, total o parcialmente, para cometer un delito dentro del alcance de la Convención. Por lo tanto, el objetivo principal de las personas o entidades involucradas en el financiamiento del terrorismo no es necesariamente ocultar las fuentes del dinero, sino ocultar tanto el financiamiento como la naturaleza de la actividad financiada [20].

Los grupos terroristas necesitan dinero para mantenerse y llevar a cabo actos terroristas. El financiamiento del terrorismo abarca los medios y métodos utilizados por las organizaciones terroristas para financiar sus actividades. Este dinero puede provenir de fuentes legítimas, por ejemplo, de ganancias de empresas y organizaciones benéficas. Pero los grupos terroristas también pueden obtener su financiación de actividades ilegales como el tráfico de armas, drogas o personas, o el secuestro por rescate. La lucha contra la financiación del terrorismo es un esfuerzo muy complejo que involucra a muchos actores diferentes con una amplia variedad de respuestas, que van desde la legislación, la política internacional hasta las respuestas a nivel operativo [21].

Diversos organismos internacionales trabajan en los aspectos jurídicos de la lucha contra la financiación del terrorismo, incluida la promoción de la ratificación de los instrumentos jurídicos universales pertinentes, en particular el Convenio Internacional para la Represión de la Financiación del Terrorismo, creada a partir del año 1999; y, la implementación de estos estándares internacionales. Esto implica revisiones de la legislación internacional para garantizar la tipificación adecuada de los delitos relacionados con el financiamiento del terrorismo y la redacción de leyes, desarrollando la capacidad de los funcionarios encargados de hacer cumplir la ley y la justicia penal para investigar, enjuiciar y adjudicar el financiamiento del terrorismo mediante la provisión de capacitación especializada en cuestiones relacionadas con técnicas especiales de investigación, congelamiento, incautación y decomiso de activos terroristas, y fortalecimiento de la cooperación regional e internacional contra el financiamiento del terrorismo [21].

1.4.3.3. Vinculación de los esfuerzos para combatir el lavado de dinero y el financiamiento del terrorismo

El lavado de dinero es el proceso de ocultar el origen ilícito del producto de delitos. El financiamiento del terrorismo es la recolección o provisión de fondos para fines terroristas. En el caso del lavado de dinero, los fondos son siempre de origen ilícito, mientras que, en el caso del financiamiento del terrorismo, los fondos pueden provenir tanto de fuentes legales como ilícitas. Por lo tanto, el objetivo principal de las personas o entidades involucradas en el financiamiento del terrorismo no es necesariamente ocultar las fuentes del dinero, sino ocultar tanto la actividad de financiamiento como la naturaleza de la actividad financiada [22].

Se utilizan métodos similares tanto para el blanqueo de capitales como para la financiación del terrorismo. En ambos casos, el actor hace un uso ilegítimo del sector financiero. Las técnicas utilizadas para lavar dinero y financiar actividades terroristas/terrorismo son muy similares y en muchos casos idénticas. Por lo tanto, un marco eficaz contra el blanqueo de capitales y la financiación del terrorismo debe abordar ambas cuestiones de riesgo: debe prevenir, detectar y sancionar la entrada

de fondos ilegales en el sistema financiero y la financiación de personas, organizaciones y/o actividades terroristas. Asimismo, las estrategias tienen como objetivo atacar a la organización criminal o terrorista a través de sus actividades financieras, y utilizan el rastro financiero para identificar los diversos componentes de la red criminal o terrorista [23].

Los controles financieros contra la financiación del terrorismo son útiles y necesarios. Realizan una serie de funciones, incluida la reducción de posibles daños causados por operaciones y ataques terroristas. Los controles financieros también facilitan el seguimiento de las actividades militantes para que se puedan tomar medidas preventivas. También permiten la reconstrucción de hechos y la detección de cómplices que luego pueden ser perseguidos. Además, el conocimiento de que todos los tipos de actividades financieras están bajo escrutinio obliga a los grupos extremistas a realizar frecuentes cambios tácticos y participar en las comunicaciones, lo que genera valiosas oportunidades para la recopilación de inteligencia [24].

1.4.3.4. Organismos que intervienen en contra del lavado de activos y financiamiento del terrorismo

1. **Grupo de Acción Financiera Internacional para la Prevención del Lavado de Activos (GAFI)**

   El Grupo de Acción Financiera Internacional (GAFI) es un organismo intergubernamental de formulación de políticas cuyo propósito es establecer estándares internacionales y desarrollar y promover políticas, tanto a nivel nacional como internacional, para combatir el lavado de dinero y el financiamiento del terrorismo. Se formó en 1989 para establecer las medidas a tomar en la lucha contra el lavado de dinero. Actualmente está compuesta por 32 países y territorios y dos organizaciones regionales. También se han desarrollado ocho organismos regionales similares que tienen formas y funciones similares [25].

El GAFI ha emitido 40 recomendaciones para combatir el lavado de dinero y 9 recomendaciones especiales para combatir el financiamiento del terrorismo, que abarcan el conjunto integral de medidas que los países deben tener en vigor dentro de sus sistemas regulatorios y de justicia penal; las medidas preventivas que deben tomar las instituciones financieras y otras empresas y profesiones; medidas para garantizar la transparencia sobre la titularidad de las personas y estructuras jurídicas; el establecimiento de autoridades competentes con funciones y facultades apropiadas y mecanismos de cooperación; y arreglos para cooperar con otros países [26].

2. Grupo de Acción Financiera de Latinoamérica (GAFILAT)

Es una organización intergubernamental de base regional que agrupa a 17 países de América del Sur, Centroamérica y América del Norte. El GAFILAT fue creado para prevenir y combatir el lavado de activos, financiamiento del terrorismo y el financiamiento de la proliferación de armas de destrucción masiva, a través del compromiso de mejora continua de las políticas nacionales contra estos flagelos y la profundización en los distintos mecanismos de cooperación entre los países miembros. El GAFILAT es uno de los grupos regionales del Grupo de Acción Financiera GAFI/FATF (Grupo de Acción Financiera Internacional/Financial Action Task Force) y está conformado por Argentina, Bolivia, Brasil, Chile, Colombia, Costa Rica, Cuba, Ecuador, Guatemala, Honduras, México, Nicaragua, Panamá, Paraguay, Perú, República Dominicana y Uruguay. El GAFILAT obtuvo la categoría de miembro asociado del GAFI y por tanto participa en la elaboración, revisión y modificación, a la vez que adhiere a las 40 Recomendaciones emitidas por este mismo organismo. Estas buenas prácticas son el estándar internacional más reconocido a nivel mundial en materia de prevención y combate del lavado de activos y financiamiento del terrorismo [27].

El GAFILAT apoya a sus miembros en la implementación de las 40+9 recomendaciones y en la creación de un sistema regional de prevención

contra el lavado de activos y el financiamiento al terrorismo. Las herramientas principales para asistir a los países son las medidas de capacitación y asistencia técnica (a través de la elaboración de guías, informes y documentos de apoyo), y las evaluaciones mutuas [27].

3. Group of Financial Intelligence Units (EGMONT)

Uno de los elementos clave de los regímenes para la lucha contra el lavado de activos y financiamiento del terrorismo, es el requisito de que las instituciones financieras y otras empresas no financieras designadas informen las transacciones que consideren sospechosas de estar relacionadas con actividades delictivas o terroristas. Debido a la confidencialidad que tradicionalmente acompaña a las transacciones financieras y a que los sujetos obligados no siempre cuentan con los medios para fundamentar sus sospechas, resulta difícil denunciarlas directamente a las autoridades encargadas de hacer cumplir las leyes penales. Por lo tanto, es necesario que los gobiernos establezcan una agencia especializada, la Unidad de Inteligencia Financiera (UIF), enfocada en el procesamiento de información financiera que pueda estar relacionada con actividades delictivas o terroristas. En sus formas más simples, las UIF son agencias que reciben informes de transacciones sospechosas de instituciones financieras y otras personas y entidades, las analizan y difunden la inteligencia resultante a las agencias policiales locales y las UIF para combatir el lavado de dinero. Como agencias gubernamentales, las UIF deben conservar suficiente independencia para lograr sus objetivos sin interferencia o influencia indebida [28].

Según The Egmont Group, la asociación internacional informal de UIF, 101 países están actualmente reconocidos como unidades operativas de UIF, y otros se encuentran en diversas etapas de desarrollo. Las 40+9 Recomendaciones del GAFI exigen que los países operen UIF que cumplan con la definición del Grupo Egmont [28].

1.4.3.5. Uso de nuevas tecnologías en el lavado de activos y financiamiento del terrorismo

Las nuevas tecnologías tienen el potencial de hacer que las medidas contra el lavado de dinero y contra el financiamiento del terrorismo sean más rápidas, económicas y efectivas. Pueden mejorar la implementación de estándares para avanzar en los esfuerzos globales para garantizar la inclusión financiera y evitar consecuencias no deseadas, como la exclusión financiera. Como emisor de estándares globales el GAFI está firmemente comprometido a mantenerse al tanto de las tecnologías innovadoras y los modelos comerciales en el sector financiero y a garantizar que los estándares globales permanezcan actualizados y puedan permitir una regulación del sector financiero "inteligente" que aborda los riesgos y promueve la innovación responsable [29].

En consecuencia, el GAFI revisó las oportunidades y los desafíos de las nuevas tecnologías para crear conciencia sobre el progreso relevante en innovación y soluciones digitales específicas. También analizó los desafíos y obstáculos persistentes para su implementación y cómo mitigarlos. Este proyecto incluyó la revisión y el análisis de la tecnología regulatoria (RegTech) y la tecnología de supervisión (SupTech), las cuales pueden mejorar la efectividad los estándares. Las habilidades, los métodos y los procesos innovadores, así como las formas innovadoras de utilizar los procesos basados en tecnología establecidos, pueden ayudar a los reguladores, supervisores y entidades reguladas a superar muchos de los desafíos identificados [29].

La tecnología puede facilitar la recopilación, el procesamiento y el análisis de datos y ayudar a los actores a identificar y gestionar los riesgos de lavado de dinero y financiamiento del terrorismo de manera más efectiva y cercana al tiempo real. Los pagos y transacciones más rápidas, los sistemas de identificación más precisos, el seguimiento, el mantenimiento de registros y el intercambio de información entre las autoridades competentes y las entidades reguladas también ofrecen ventajas. El mayor uso de soluciones digitales basadas en Inteligencia Artificial (IA) y sus diferentes subconjuntos (aprendizaje automático, procesamiento de lenguaje

natural) puede ayudar potencialmente a identificar mejor los riesgos y responder, comunicar y monitorear actividades sospechosas. A nivel del sector público, el monitoreo mejorado en vivo (en tiempo real) y el intercambio de información con las contrapartes permiten una supervisión más informada de las entidades reguladas, lo que ayuda a mejorar la supervisión [29].

A nivel del sector privado, la tecnología puede mejorar las evaluaciones de riesgos, las prácticas de incorporación, las relaciones con las autoridades competentes, la auditabilidad, la rendición de cuentas y la buena gobernanza en general, al mismo tiempo que ahorra costos. Aunque se presentan desafíos relacionados con el desarrollo, la adopción y la aplicación de estas soluciones o prácticas innovadoras, muchos de estos desafíos se deben a restricciones operativas y regulatorias pendientes, como los sistemas de cumplimiento heredados y los marcos regulatorios y mecanismos de supervisión tradicionales. Las complejidades y los costos involucrados en el reemplazo o la actualización de los sistemas heredados dificultan la explotación del potencial de los enfoques innovadores tanto para la industria como para el gobierno [29].

Para la industria, el análisis de costo-beneficio para adoptar nuevas tecnologías sigue siendo un obstáculo para una mayor aceptación de soluciones innovadoras, basado en parte en una falta real o percibida de incentivos regulatorios para buscar la innovación. Las dificultades con la explicabilidad y la interpretabilidad de las soluciones digitales son otro desafío clave tanto para la industria como para los reguladores que en parte se debe a la limitada disponibilidad de experiencia relevante y la falta de conocimiento del potencial de las tecnologías innovadoras, tanto en la industria como en el gobierno. Una mayor comunicación y cooperación entre el sector público y privado, informada por el tipo de información y análisis, junto con un énfasis en la adopción responsable de nuevas tecnologías y la efectividad, en particular con respecto a las regulaciones de protección de datos, será clave para superar estos desafíos y cumplir plenamente la promesa de innovación responsable para fortalecer la eficacia de las medidas, que cuando se usan de manera responsable y proporcional, las tecnologías innovadoras pueden

ayudar a identificar riesgos y centrar los esfuerzos de cumplimiento en los desafíos existentes y emergentes [29].

La combinación de la eficiencia y la precisión de las soluciones digitales con el conocimiento y las habilidades analíticas de los expertos humanos produce sistemas más sólidos que pueden responder de manera efectiva a los requisitos mientras son totalmente auditables y responsables. El uso de nuevas tecnologías e innovación puede ayudar a los sectores público y privado a mejorar la eficacia de su implementación de los Estándares del GAFI basada en el riesgo. El desarrollo, la adopción y la supervisión regulatoria de estas tecnologías deben reflejar tanto las amenazas como las oportunidades. También debe garantizar que el uso de herramientas innovadoras sea compatible con los estándares internacionales de protección de datos, privacidad y ciberseguridad [29].

1.4.4. Machine Learning para detectar riesgos de lavado de activos y financiamiento del terrorismo

El aprendizaje automático implica diseñar una secuencia de acciones para resolver un problema automáticamente a través de la experiencia y la evolución de algoritmos de reconocimiento de patrones con intervención humana limitada o nula, es decir, es un método de análisis de datos que automatiza la construcción de modelos analíticos. La GAFI identifica al aprendizaje automático como las capacidades impulsadas por IA que ofrecen un gran beneficio para detectar riesgos de lavado de activos y financiamiento del terrorismo en las entidades reguladas y los supervisores. El aprendizaje automático ofrece la mayor ventaja a través de su capacidad para aprender de los sistemas existentes, lo que reduce la necesidad de una entrada manual en el monitoreo, reduce los falsos positivos y la identificación de casos complejos, además de facilitar la gestión de riesgos. Las aplicaciones de aprendizaje automático son útiles para detectar anomalías y valores atípicos identificando y eliminando información duplicada para mejorar la calidad y el análisis de los datos [29].

De acuerdo con [30] “la capacidad del aprendizaje automático para descubrir patrones en los datos, procesar varios tipos de datos y actuar de manera autónoma promete permitir que los intermediarios financieros detecten actividades de lavado de dinero de manera rentable”. El aprendizaje automático permite desarrollar conocimiento sobre esquemas de lavado de dinero, a través de perfiles descriptivos; el otro de detectar intentos de lavado de dinero, a través de perfiles predictivos. Si bien el desarrollo de conocimientos es un paso esencial para comprender la naturaleza del fenómeno, en última instancia, para cumplir con el objetivo de ayudar a reducir el delito, los intermediarios financieros deben ser capaces de detectar y prevenir los intentos de lavar el producto del delito a través de sus organizaciones.

Para Lorenz, et al. [31], dada la complejidad cambiante de los esquemas de lavado de dinero, es poco probable que sea posible identificar todas (o incluso la mayoría) de las entidades involucradas en el lavado de dinero. Debido a que las etiquetas resultantes de las investigaciones policiales no son inmediatas y la anotación manual es costosa. Por lo tanto, para evaluar adecuadamente la viabilidad práctica de lavado de dinero, las estrategias de aprendizaje activo son primordiales.

Según [32] los autores encontraron que la mejor calidad de predicción se logra mediante el uso de un algoritmo de aumento de gradiente sobre árboles de decisión. Estudiaron la calidad de la selección de hiperparámetros utilizando las bibliotecas de Python hyperopt y Optuna, donde estimaron que la velocidad de obtención del conjunto óptimo. El modelo se puede utilizar para formar una lista de los indicadores más importantes para la detección temprana de organizaciones involucradas en el lavado de activos y el financiamiento del terrorismo, así como para desarrollar recomendaciones adecuadas para mejorar el proceso de control de cumplimiento.

Según los autores en [33] propusieron una nueva metodología que considera las tipologías que se han descrito en los informes del GAFI pero que no se habían incluido en estudios previos, mejorando la autocomparación. También proponen un indicador de anormalidad basado en la varianza de las variables y esto potencia la

comparación de grupos en el proceso de agrupamiento. Los resultados de su estudio mostraron una reducción significativa en el número de falsos positivos y una mayor precisión en comparación con el método anterior basado en reglas.

Según [34] encontraron que el algoritmo de detección de valores atípicos basado en límites triangulares (TBOD) segmenta el conjunto de datos en un clúster/grupo en función del uso del producto y el riesgo asociado del cliente. Por lo tanto, como alternativa, la técnica de algoritmo de valores atípicos para detección de lavado de dinero (AROMLD) basada en autorregresión utiliza una aproximación de rango intercuartílico (IQR) que ofrece una métrica de variabilidad para aislar transacciones sospechosas en sistemas financieros en tiempo real. Esto mitiga claramente la complejidad del tiempo junto con la complejidad computacional. El análisis comparativo de ambos mecanismos contra las metodologías existentes se delibera en términos de sensibilidad, especificidad, tiempo de ejecución y precisión.

Según los autores en [35] desarrollaron un proceso de investigación basado en clustering y redes neuronales para detectar casos sospechosos en el contexto de ML. También aplicaron heurísticas como detección de sospechas para mejorar el tiempo de ejecución. A partir de los resultados experimentales obtenidos en el mayor fondo de conjuntos de datos de transacciones de BEP, pudieron concluir que este enfoque es prometedor y satisface las necesidades de la unidad de lucha contra el lavado de activos.

Según [36] en su investigación orientada a dirigir la atención hacia la identificación del lavado de dinero, propusieron un Modelo Relacional Probabilístico utilizando el Patrón Secuencial de Auditoría (PRM-ASP) Mining. El algoritmo de mapeo de asociación (AM) se realizó en el conjunto de datos preprocesados para separar las transacciones que se realizan de cuentas de uno a muchos y de muchos a uno. La minería PRM-ASP utilizada para la identificación del lavado de dinero utiliza datos de series temporales e identifica relaciones de uno a muchos y de muchos a uno entre transacciones para identificar las cuentas vulnerables. Del conjunto separado de transacciones por PRM, se identificaron las cuentas bancarias vulnerables y se recopilaron todas las cuentas de lavado de dinero. Luego, la minería PRM-ASP

basada en lógica relacional utilizaron el patrón secuencial de auditoría para identificar el patrón de transacción en las cuentas vulnerables. Además de proporcionar una identificación lógica de lavado de dinero en la mayoría de los dominios del mundo real, el PRM se usa efectivamente al auditar los patrones secuenciales.

EL autor en [37] con el fin de mejorar cada vez más el Sistema de Administración del Riesgo de Lavado de Activos y de la Financiación al Terrorismo (SARLAFT) desde el tema de segmentación utilizó el Método CLARA, que es un método que tiene como idea base K-medoides (PAM) junto a técnicas de remuestreo, que es usado cuando se tienen grandes cantidades de datos; este método surge por las limitaciones del método K-medoides, dado a que este requiere de una gran cantidad de recursos tecnológicos (uno de ellos es la necesidad de gran cantidad de memoria RAM), lo cual puede suponer una limitación para el estadístico. La metodología implementada en dos diferentes escenarios, mostraron una adecuada segmentación, en donde se evidenció múltiples índices que permitieron tener una certeza mayor del comportamiento de los diferentes factores de riesgo, como se evidenció en el factor clientes, dado a que es unos de los principales y es de los que más información posee.

2. METODOLOGÍA

2.1. Framework Ciencia del diseño

La ciencia del diseño trata de desarrollar nuevos productos basados en soluciones tecnológicas que siguen un estándar de marco de trabajo elaborado, el objetivo de esta metodología es que sus productos sean evaluados bajo criterios de valor o utilidad que permitan el ser humano solucionar problemas siguiendo procesos robustos y óptimos. Los productos de la ciencia del diseño son de cuatro tipos: construcciones, modelos, métodos e implementaciones y que constan de dos actividades básicas, construir y evaluar [38].

Las construcciones, los modelos, los métodos y las instancias son artefactos que abordan alguna tarea. Las actividades de investigación relacionadas con estos artefactos son: construir, evaluar, teorizar y justificar. Construir y evaluar son actividades de investigación en ciencias del diseño destinadas a mejorar el rendimiento [38].

Para el Desarrollo del presente proyecto de investigación se aplicará la metodología de Ciencia del Diseño ya que trataremos de identificar el problema de lavado de activos en una aseguradora creando, evaluando y mejorando artefactos de TI de forma estructurada a través de un software, en base a conocimientos del negocio que identificarán patrones inusuales en los movimientos de clientes. Los artefactos de TI creados generan beneficios organizativos de manera interna y hacia el cliente, además resuelven conflictos que surgen de procesos manuales y se optimiza el rendimiento.

La Ciencia del Diseño tiene como objetivo hacer cumplir un flujo de trabajo que se basa en etapas y subprocesos que parten con la detección de un problema que es considerado una necesidad organizacional hasta la aplicación de un artefacto tecnológico cuyo rendimiento se evalúa y se monitorea. La siguiente figura muestra

el proceso que debe seguir la metodología propuesta [38]. Todos las figuras con elaboración del autor.

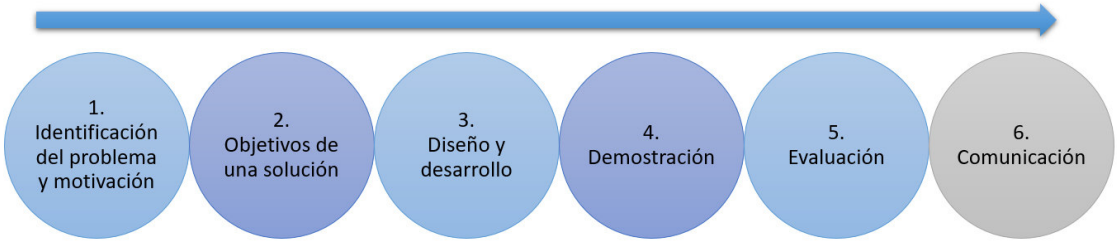


Figura 1. Proceso de Ciencia del diseño

En la Figura 1 podemos observar el proceso que sigue la metodología propuesta para llegar al objetivo planteado, este proceso es por cumplimientos ya que al culminar uno se prosigue con el siguiente, a continuación, se describe a detalle cada uno de estos pasos.

Identificación del problema y motivación

En esta etapa se tratará de dar valor a la necesidad del problema presentado, es importante identificar el giro del negocio, la información que se maneja y los procesos internos, algo importante que se debe analizar dentro de esta etapa es la infraestructura tecnológica que se tiene y los mecanismos implementados hasta el momento para solucionar el problema. Dar valor y sentido a una solución motiva al investigador y a los involucrados en el proyecto, esto genera iniciativa y apoyo a resolver el problema.

Objetivos de una solución

En esta etapa se debe trazar la meta a donde se quiere llegar, se debe tener bien plasmada la solución propuesta para cubrir las necesidades establecidas por la organización. Los objetivos deben inferirse racionalmente de la especificación del problema, a su vez estos deben ser alcanzables en el corto plazo y con los insumos que se encuentren al alcance el investigador.

Diseño y Desarrollo

Se debe empezar con el análisis de la información, comprensión de patrones y comportamientos para la aplicación de algoritmos de aprendizaje automático que permitan la construcción de modelos a partir del entrenamiento usando los datos de entrada. Esta actividad incluye determinar la funcionalidad deseada del artefacto y su arquitectura y luego crear el artefacto real. Los recursos necesarios para pasar de los objetivos al diseño y desarrollo incluyen el conocimiento de la teoría que se puede aplicar como solución.

Demostración

Para la demostración del artefacto se debe tener un conocimiento completo de su óptima funcionalidad y su aplicación en el caso de estudio presentado. En esta fase de demostración el artefacto o modelo desarrollado debe probarse sobre datos e información nueva que asemeje un caso real o una simulación de datos y validar su eficacia dando como aprobada su implementación.

Evaluación

Esta actividad implica comparar los objetivos de una solución con los resultados reales observados del uso del artefacto en la demostración. Requiere conocimiento de métricas relevantes y técnicas de análisis [38]. Para esta etapa se medirá la eficacia de los algoritmos para resolver el problema planteado. Adicionalmente, se analizarán los resultados obtenidos luego de aplicar los modelos.

Comunicación

Comunicar el problema y su importancia, el artefacto, su utilidad y novedad, el rigor de su diseño y su efectividad a los investigadores y otras audiencias relevantes. Las audiencias orientadas a la tecnología necesitan que se les comunique detalles suficientes para permitir que el artefacto descrito se implemente y se use dentro de un contexto organizacional apropiado [38].

2.2. Actividades

En base a todo lo planteado según los estándares que establece esta metodología sabemos que el objetivo final del proyecto propuesto es el desarrollo e implementación del artefacto como tal, identificado como un algoritmo estadístico-matemático capaz de dar solución al problema descrito en el planteamiento. Siguiendo la metodología plasmaremos la relación que existe entre los objetivos específicos del proyecto de investigación, las etapas de la metodología y las actividades que realizaremos a lo largo del tiempo para cumplir con esos objetivos. En la Tabla 1 se presenta el proceso de aplicación de la metodología.

Objetivo	Etapa	Actividad
• Realizar una revisión sistemática de la literatura en temas relacionados a riesgos de lavados de activos y financiación del terrorismo (LAFT), modelos supervisados y no supervisados de segmentación y clasificación, automatización de procesos y seguridad de la información.	1. Identificación del problema y motivación	1. Investigación de la literatura en estudios de actualidad relacionados con detección de lavado de activos y financiación del terrorismo en datos proporcionados por una aseguradora. 2. Analizar algoritmos usados y resultados obtenidos por estudios realizados en el campo de interés.
• Conocer los antecedentes del caso de estudio sobre el monitoreo actual de sus factores de riesgo.	1. Identificación del problema y motivación	Análisis de los problemas actuales que se presentan en el caso de estudio y los riesgos que se enfrentan si es que no se implementa una solución.

• Desarrollar e implementar modelos de segmentación de clientes basados en machine learning para detectar riesgos de lavados de activos y financiación del terrorismo en la empresa, objeto del caso de estudio.	2. Objetivos de solución	Desarrollo de algoritmos de aprendizaje automático que reconocen transacciones sospechosas usando técnicas de machine learning. Estos algoritmos al ejecutarse generarán alertas.
• Elaborar los modelos de segmentación en base al plan de trabajo y bajo las necesidades de la empresa, creando métricas inteligentes de detección de alertas y optimizando en paralelo la implementación de los modelos en sus sistemas internos.	3. Diseño y desarrollo	Construir un dataset con la información disponible. Aplicación de algoritmos de clustering y clasificación. Análisis de los clústeres (Búsqueda de variables de perfilamiento) Entrenamiento de los algoritmos de clasificación con el dataset obtenido y generación de resultados. Selección de los modelos con mejor rendimiento usando métricas apropiadas. Implementación de los algoritmos bajo código que automaticen las alertas.

• Validar los modelos obtenidos y evaluarlos con métricas apropiadas.	4. Demostración (validación). 5. Evaluación	• Probar los modelos desarrollados en datos nuevos. Medir el rendimiento obtenido en los datos de prueba. Comparar rendimiento de datos de entrenamiento vs. de prueba para evitar overfitting.
• Documentar la solución con detalles suficientes para que pueda ser entendida y replicada por personal de IT de la empresa del caso de estudio	6. Comunicación	• Generar la documentación del desarrollo, entrenamiento, prueba y validación de los algoritmos desarrollados.

Tabla 1.Proceso de aplicación para la metodología

La Tabla 1 muestra las actividades que se harán para llegar a cumplir el objetivo general y los específicos siguiendo la metodología de ciencia del diseño y cumpliendo a la par cada uno de los objetivos y etapas.

2.3. Modelo de segmentación CLARA (Clustering Large Applications)

El método CLARA está basado en k-medoide y tiene la misma finalidad de agrupamiento de objetos, a diferencia del resto de métodos de segmentación CLARA tiene la capacidad de manejar grandes volúmenes de datos ya que esto se logra particionando en subconjuntos de datos de tamaño fijo lo que genera

rendimientos de tiempo y almacenamiento de manera lineal en n en lugar de cuadráticos [39].

Cada subconjunto de datos que se particiona se divide nuevamente en k grupos utilizando el mismo algoritmo de PAM. Una vez que se han seleccionado objetos representativos del subconjunto de datos, cada observación de todo el conjunto de datos se asigna al medoide más cercano. La media (equivalente a la suma) de las diferencias de las observaciones con su medoide más cercano se usa como una medida de la calidad de la agrupación. Se retiene el subconjunto de datos para el cual la media (o suma) es mínima. Un análisis adicional se lleva a cabo en la partición final.

Cada subconjunto de datos se ve obligado a contener los medoides obtenidos del mejor subconjunto de datos hasta entonces. Las observaciones dibujadas al azar se agregan a este conjunto hasta que se haya alcanzado el tamaño de la muestra.

El algoritmo CLARA se resume en los siguientes pasos:

1. Tomar una muestra aleatoria de tamaño fijo en la población total de elementos.
2. Aplicar el algoritmo PAM (una versión más robusta de K-means) sobre los datos de la muestra y extraer los correspondientes k-medoides (k elementos representativos de la muestra) [37].
3. Medir la calidad de los medoides a nivel poblacional (no muestral), calculando la media de las disimilitudes entre cada elemento de la población total y su medoide más cercano. Esto se define como la siguiente función de costo:

$$Costo(M) = \frac{\sum_{i=1}^{n} disim(O_i, rep(M, O_i))}{n}$$

4. Donde M es el conjunto de k-medoides seleccionado en el paso 2, $disim(O_i, O_j)$es la disimilitud entre los objetos O_i y O_j, $rep(M, O_i)$ es el

medoide del conjunto M que es el más cercano del objeto O_i y n es el número total de elementos en la población.

5. Repetir los pasos 1-2-3 un número de veces especificado a fin de minimizar el sesgo de muestreo. En cada repetición se debe tomar una nueva muestra, aplicar el algoritmo PAM y obtener un conjunto adicional de k-medoides con su respectivo valor de la función de costo.
6. Retener el conjunto de k-medoides con el mínimo valor de la función de costo (el conjunto con mejor calidad a nivel poblacional).
7. Obtener una segmentación final de la población en k grupos, asignando cada elemento de la población a su medoide más cercano.

2.4. Medidas de validación.

Teniendo en cuenta que existen varias medidas de distancia, varios métodos de segmentación y varios métodos para decidir qué grupos (clusters) se van uniendo o separando en cada paso de un proceso de segmentación. Es claro que al combinarlos se generan muchos escenarios de segmentación. Cada uno de estos escenarios traza una metodología de segmentación distinta, por lo cual, a la hora de encontrar agrupaciones en un conjunto de datos se debe establecer cuál es la metodología más apropiada, realizando una elección óptima del número de segmentos y validando la calidad de los grupos obtenidos a fin de garantizar homogeneidad al interior de los segmentos y heterogeneidad entre ellos. Así pues, se suelen construir varias segmentaciones con los mismos datos, con el propósito de contrastar distintas metodologías (variando también la cantidad de segmentos a obtener) para finalmente escoger el número óptimo de segmentos y la metodología ideal para segmentar cada factor de riesgo de acuerdo con las siguientes medidas de validación:

2.4.1. Índice Silhouette

Según el autor en [37] el valor de la silueta es una medida de la similitud de un elemento con su propio segmento (cohesión - homogeneidad al interior del segmento) en comparación con otros segmentos (separación - heterogeneidad

entre segmentos). Para cada elemento i, el ancho de la silueta $s(i)$ es definido como sigue:

Sea, $a_i$la distancia promedio entre el elemento i, y todos los demás elementos del segmento al que i pertenece (si i es la única observación en su grupo, $s(i)$ será igual a cero, sin más cálculos). Para todos los otros segmentos C, sea $d(i, C)$ la distancia promedio de i a todas las observaciones de C. El más pequeño de esos $d(i, C)$ es $b_i = min_C\,[d(i, C)]$, y puede ser visto como la distancia entre i y su segmento "vecino", es decir, el segmento más cercano al que no pertenece. Finalmente,

$$s(i) = \frac{b_i - a_i}{max\{a(i); b(i)\}}$$

El valor de $s(i)$ varía de -1 a 1, de modo que los elementos con un valor alto (cercano a 1) están muy bien agrupados, un valor de silueta pequeño (alrededor de 0) indica que el elemento está muy cerca del límite de decisión entre dos segmentos vecinos, y las observaciones con un valor de silueta negativo indican que esos elementos pueden haber sido asignados al segmento equivocado.

El promedio de $s(i)$ sobre todos los elementos de un mismo segmento indica cuán estrechamente agrupados están todos los elementos de dicho segmento en comparación con los segmentos restantes. Por lo tanto, el promedio de $s(i)$ sobre todos los elementos de todo el conjunto de datos es una medida de cuán apropiadamente han sido agrupados los datos indicando el grado de homogeneidad al interior de los segmentos y heterogeneidad entre ellos. El objetivo entonces es obtener segmentaciones con valores altos en la silueta promedio (silueta promedio por cada segmento y silueta promedio en general para todos los datos ya segmentados).

2.4.2. Índice Dunn

Según el autor en [37] el índice Dunn es una medida de validación interna, se obtiene tras calcular para cada segmento la distancia entre cada uno de los elementos que lo forman y los elementos de los otros segmentos (distancia InterCluster), después selecciona como "representante" de la distancia entre segmentos a la menor de todas las distancias calculadas ($\min(Distancia\ InterCluster)$). Después para cada segmento se calcula la distancia entre los elementos que lo forman (distancia Intra Cluster) y selecciona como "representante" de la distancia intra-segmento a la mayor de todas las distancias calculadas ($\max(Distancia\ IntraCluster)$). Finalmente,

$$Dunn = \frac{\min\limits_{1 \le i \le j \le q} d(C_i, C_j)}{\max\limits_{1 \le k \le q} diam(C_k)}$$

Donde $d(C_i; C_j)$ es la función de disimilitud entre dos grupos C_i y C_j definidos como $d(C_i, C_j) = \min\limits_{x \in Ci,\ y \in C} d(x, y)$ y $diam(C)$ es el diámetro de un grupo, que puede ser considerado como una medida de la dispersión del racimo. EL diámetro de un cluster C se puede definir usando la siguiente ecuación.

$$Diam(C) = \max_{x,y \in C} d(x, y)$$

Si la segmentación contiene grupos compactos y bien separados (homogéneos al interior y heterogéneos entre ellos), el numerador es grande y el denominador pequeño, dando lugar a valores altos del índice. El objetivo por lo tanto es obtener segmentaciones con valores altos en el índice Dunn.

Utilizando las medidas de validación Índice Dunn y Silueta promedio, se comparan distintas segmentaciones para cada uno de los factores de riesgo, contemplando varios métodos de segmentación, varias medidas de distancia, varias cantidades de segmentos a obtener y varios métodos para decidir qué clústeres se van uniendo

en cada paso del proceso de segmentación. Finalmente, se logra establecer la segmentación de los factores de riesgo en la aseguradora.

2.4.3. Random Forest

El modelo Random Forest es el algoritmo de clasificación que mejor se ajusta a las características del proyecto, ya que este nos permitirá tener múltiples arboles con una exactitud mayor a la hora de clasificar. Obtendremos la predicción media de los árboles individuales, esto en comparación con los árboles de decisión tienden a reducir el riesgo de sobreajuste a la hora del entrenamiento, pero su precisión es menor que puede ser mejorada con las características seleccionadas en los modelos.

El modelo random forest permite reducir las dimensiones cuando se tienen varios campos que describen al objeto de análisis. Dentro del software que se utilizará que es R viene incrustado el seleccionador de variables que nos permitirá conocer la importancia de cada variable y su aporte hacia la variable respuesta, de tal manera que podremos incluir dentro del modelo únicamente a las variables que mejor describen a los clientes y reducir la carga computacional.

Los resultados obtenidos serán producto del modelo una vez se haya analizado las muestras de entrenamiento y validación en donde se distribuirá los datos de manera aleatoria en porcentajes de 70% y 30% respectivamente, adicional se medirán los resultados de la matriz de confusión para conocer el desempeño y error del algoritmo para cada uno de los segmentos de análisis, el modelo será validado bajo las medidas de Acurracy, Precision, Recall y F1.

3. RESULTADOS

En este Capítulo se presentan los resultados obtenidos en el desarrollo de los modelos de segmentación y su funcionalidad dentro de la aseguradora para el caso de estudio propuesto, se establece el seguimiento de la metodología bajo los pasos a cumplir para llegar al objetivo plantead. Adicionalmente, se hace referencia a las pruebas realizadas con información real detallando los parámetros utilizados y los ajustes que se han realizado, así como las validaciones que verifican los modelos óptimos.

3.1. Procesamiento de datos

Dentro de esta etapa como parte inicial antes de entrar al procesamiento de datos bajo un algoritmo, es importante tratar la información y entender cada una de las distribuciones de las variables a utilizarse. Adicionalmente se deben limpiar, completar y eliminar datos como parte fundamental al momento de establecer las entradas que utilizarán los algoritmos ya que daremos certeza de la información como datos reales de clientes.

El proceso que se siguió para llegar a los datos finales antes de ser modelados fue realizar un análisis descriptivo de cada tabla con la función “Summary” que permite tener un resumen breve de cada variable y conocer de manera general cómo se comportan sus distribuciones y el número de registros nulos, mediante la función “str” se conoce los tipos de datos y las codificaciones de las variables tipo factor que son las que tienen categorías, con la función “table” podemos ver si existen campos duplicados y poder asegurar registros únicos en donde los campos son llaves primarias para otras tablas, el análisis de estadísticas descriptivas se encuentran en el Anexo1.

Dentro del análisis se realizan cambios de tipos de variables como son de cadenas a numéricas o cadenas a fechas, etc. Se crean nuevas variables que son consideradas importantes para el modelamiento como número de transacciones

por cliente en 1, 12 y 24 meses normales o agregados de pasivos y activos que tienen, números de contratos y valores asegurados que acumulan.

Para el análisis de la tabla de clientes, se pudo identificar la siguiente distribución según el tipo de cliente.

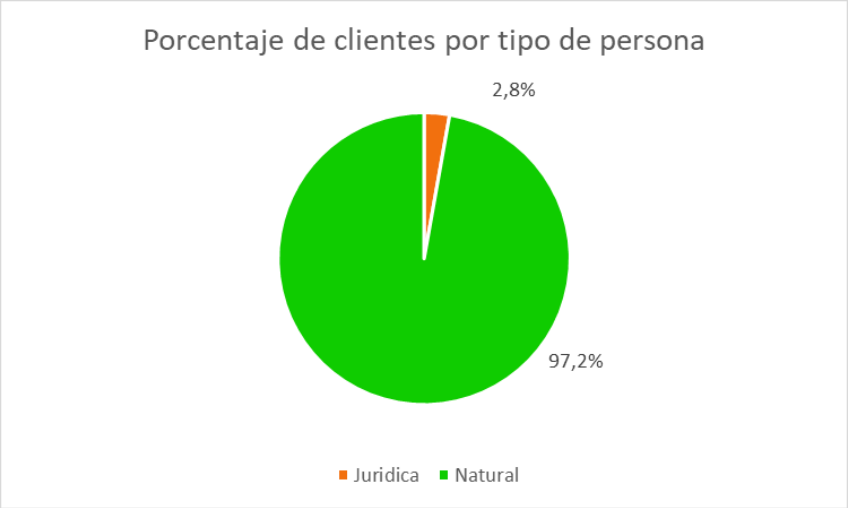


Figura 2. Distribución de clientes por tipo de cliente

En la Figura 2 observamos que se obtiene que 30.434 clientes que pertenecen a la asegurada del caso de estudio, de los cuales existen 844 son Personas Jurídicas las cuales representa el 2.8% del total y 29.590 personas Naturales que representa el 97.2% del total de clientes, lo que quiere indicar que la mayor cantidad de clientes de la empresa son representadas como personas naturales, por lo cual las personas jurídicas no tienen una representación considerable.

3.1.1. Orígenes de datos

Para la conexión de los datos se trabajó con una base de datos SQL Server Management Studio 18 en donde la información correspondía a una estructura de datos tipo DWH (data warehouse) basada en dimensiones que fueron tomadas para asociar nuevas variables a nivel de clientes que describan mejor el comportamiento de estos.

La estructura de datos utilizada en la segmentación está basada en los siguientes factores de riesgo:

- Contratos
- Clientes
- Transacciones
- Jurisdicciones Departamentos
- Jurisdicciones países

La Tabla 2 presenta una descripción de la estructura de datos de los clientes.

Campo	Campo en español	Tipo de dato
Document_Type	Tipo de documento	Char(1)
Document_Number	Número de documento	Bigint
Names	Nombres	Varchar(25)
Surnames	Apellidos	Varchar(25)
Income	Ingresos	Numeric(20, 0)
Expenses	Egresos	Numeric(20, 0)
Assets	Activos	Numeric(20, 0)
Passives	Pasivos	Numeric(20, 0)
Heritage	Patrimonio	Numeric(20, 0)
Economic_Activity	Actividad económica	Varchar(150)
Economic_Sector	Sector económico	Varchar(150)
CIIU	Código CIIU	Varchar(5)
CIIU_Description	Descripción CIIU	Varchar(200)
CIIU_HighRisk	Alto riesgo CIIU	Varchar(5)
Birth_Date	Fecha nacimiento	SmallDatetime
Create_Date	Fecha creación	Datetime
Modify_Date	Fecha modificación	Datetime
Jurisd_ID	Jurisdicción de departamento	Varchar(10)
Is_Public	¿Es persona publica?	Bit
Public_Recognition	Descripción del porque es persona publica	Varchar(50)
Risk	Riesgo	Varchar(5)

Company_Type	Tipo de compañía / empresa	Varchar(50)
Segment_Client_VIDA	Segmento del cliente	Varchar(10)
Cutoff_Date	Fecha de corte	DateTime

Tabla 2. Estructura de datos de clientes

- PK: Document_Type, Document_Number
- FK: Jurisd_ID

La Tabla 3 presenta una descripción de la estructura de datos de los contratos.

Campo	Campo en español	Tipo de dato
Contract_ID	Código del contrato	Numeric(10)
Product	Código del producto	Varchar(7)
Product_Plan	Plan del producto	Varchar(7)
Document_Type	Tipo de documento	Char(1)
Document_Number	Número de documento	Bigint
Initial_Date	Fecha inicial	Datetime
Final_Date	Fecha final	Datetime
Document_Type_Insured	Tipo de documento del asegurado	Char(1)
Document_Number_Insured	Número de documento del asegurado	Varchar(20)
Insured_value	Valor asegurado	Money
Total_Premium	Prima total	Money
Frequency_Payment	Frecuencia de pago	Varchar(50)
Commercial_Channel	Canal comercial	Int
Cutoff_Date	Fecha de corte	DateTime

Tabla 3. Estructura de datos de contratos

- PK: Contract_ID, Product, Product_Plan
- FK: (Document_Type, Document_Number), (Product_Plan_Detail Product, Product_Plan) y Channel_ID

La Tabla 4 presenta una descripción de los campos de las transacciones.

Campo	Campo en español	Tipo de dato
Event_Number	Número de evento	Char(12)
Transaction_Number	Número de transacción	Char(25)
Contract_ID	Código del contrato	Numeric(10)
gProduct	Código del producto	Varchar (7)
Product_Plan	Plan del producto	Varchar(7)
Movement_Code	Código de movimiento	Char(2)
Movement	Movimiento	NVarchar(100)
Movement_Date	Fecha de proceso	Datetime
Movement_Value	Valor de la transacción	Money
Modality	Modalidad de pago	Varchar(20)
Movement_Channel	Canal de movimiento	Int
Movement_Jurisdiction	Jurisdicción del departamento	Varchar(10)

Document_Type_Beneficiary	Tipo de documento del beneficiario	Char(1)
Document_Number_Beneficiary	Número de documento del beneficiario	Bigint
Cutoff_Date	Fecha de corte	DateTime
Transaction_Type	Tipo de transacción	Tinyint
Jurisd_ID_Country	Código de la jurisdicción país	Varchar (7)

Tabla 4. Estructura de datos de Transacciones

- PK: Event_Number, Transaction_Number, Contract_ID, Product, Product_Plan, Movement_Date
- FK: (Contract_ID, Product, Product_Plan) y Channel_ID

3.1.2. Data set a modelar

✓ **Actividad económica:**

- Actividad económica reportada por cada cliente
- Código CIIU
- Riesgo LAFT asociado al sector económico

✓ **Volumen o frecuencia de las transacciones:**

Aportes provenientes de tipo de movimientos:

- Suma del valor de aportes en los últimos 12 meses
- Suma del valor de aportes en los últimos 24 meses
- Suma del valor de aportes en el último mes
- Número de aportes en los últimos 12 meses
- Número de aportes en los últimos 24 meses
- Número de aportes en el último mes
- Suma del valor de retiros en los últimos 12 meses
- Suma del valor de retiros en los últimos 24 meses
- Suma del valor de retiros en el último mes
- Número de retiros en los últimos 12 meses
- Número de retiros en los últimos 24 meses
- Número de retiros en el último mes

✓ **Monto ingresos:**

- Ingresos mensuales reportados por el cliente

- ✓ **Monto egresos:**
  - Egresos mensuales reportados por el cliente

- ✓ **Monto patrimonio**
  - Activos reportados por cliente
  - Pasivos reportados por cliente
  - Patrimonio reportado por cliente

- ✓ **Variables adicionales**
  - Tipo persona (Natural/Jurídica)
  - Riesgo LAFT (perfil de riesgo del cliente)
  - Marca (1/0) cliente PEP

Con todas las variables mencionadas se crearon indicadores que son utilizados como insumo para la segmentación correspondiente. El conjunto de indicadores empleado en la segmentación se muestra en la Tabla 5.

Definición	COD indicador	Descripción / Categoría
Número, promedio y monto total de aportes en 12 y 24 meses	Id_Fe_Apor_Mon_12	Indicador que toma valores entre 0 y 1, siendo 1 la calificación otorgada al cliente con mayor monto, promedio o número de aportes en los 12 y 24 meses
	Id_Fe_Apor_Mon_24	
	Id_Fe_Apor_Num_12	
	Id_Fe_Apor_Num_24	
	Id_Fe_Apo_Prom12	
	Id_Fe_Apo_Prom24	
Número, promedio y monto total de	Id_Fe_Ret_Mon_12	Indicador que toma valores entre 0 y 1, siendo 1 la calificación
	Id_Fe_Ret_Mon_24	

retiros en 12 y 24 meses	Id_Fe_Ret_Num_12	otorgada al cliente con mayor monto, promedio o número de retiros en los 12 y 24 meses
	Id_Fe_Ret_Num_24	
	Id_Fe_Ret_Prom12	
	Id_Fe_Ret_Prom24	
Número, promedio y monto total de aportes en efectivo en 12 y 24 meses	Id_Fe_Apor_E_Mon_12	Indicador que toma valores entre 0 y 1, siendo 1 la calificación otorgada al cliente con mayor monto, promedio o número de aportes en efectivo en los 12 y 24 meses
	Id_Fe_Apor_E_Mon_24	
	d_Fe_Apor_E_Num_12	
	Id_Fe_Apor_E_Num_24	
	Id_Fe_Apo_E_Prom12	
	Id_Fe_Apo_E_Prom24	
Capacidad de aportes en 12 y 24 meses	Id_Fe_Cap_Apor24	Indicador que toma valores entre 0 y 1, siendo 1 la calificación otorgada al cliente con mayor capacidad de aportes en los 12 y 24 meses (CapApor=SumatoriaAportes ÷ Ingresos)
	Id_Fe_Cap_Apor12	
Capacidad de aportes en efectivo en 12 y 24 meses	Id_Fe_Cap_Apor24_E	Indicador que toma valores entre 0 y 1, siendo 1 la calificación otorgada al cliente con mayor capacidad de aportes en *efectivo* en los 12 y 24 meses (CapAporE=SumatoriaAportesEfectivo ÷ Ingresos)
	Id_Fe_Cap_Apor12_E	

Valor total y promedio de valor asegurado	Id_Fe_valaseg_total	Indicador que toma valores entre 0 y 1, siendo 1 la calificación otorgada al cliente con mayor monto o promedio de valor asegurado
	Id_Fe_valaseg_prom	
Valor total y promedio de prima total	Id_Fe_primaT_total	Indicador que toma valores entre 0 y 1, siendo 1 la calificación otorgada al cliente con mayor monto o promedio de prima total
	Id_Fe_primaT_prom	
Número de contratos	Id_Fe_contrato_N	Número de contratos por cliente
Numero de productos	Id_Fe_producto_N	Numero de distintos productos por cliente
Riesgo CIUU por cliente	CIUU_ALTO_RIESGO	Riesgo CIUU por cliente (indicador que toma valores de 1 a 3, siendo 3 riesgo CIUU "Alto")
Riesgo LAFT por cliente	RIESGO	Riesgo LAFT por cliente (indicador que toma valores de 1 a 3, siendo 3 riesgo "Alto")

Tabla 5. Variables de Dataset a segmentar

En la tabla 5 podemos observar la lista de las variables finales que quedaron antes del modelado, la creación de la mayoría de los indicadores consiste en un simple cambio de escala de modo que se garantiza contar con la misma información estadística que otorgan las variables en su escala original. Al respecto cabe anotar que transformaciones como cambios de escala o estandarización de variables no implican pérdida de información y en contraste sí pueden proveer mejoras para distintos métodos estadísticos tal como se ha demostrado históricamente en múltiples investigaciones de reconocido valor técnico. Así, no se excluye información alguna y se tienen en cuenta todas las variables requeridas por la normatividad respectiva.

El papel de una variable o indicador al interior del modelo no puede ser determinado por el análisis univariado que se realice sobre éste, de hecho, debe tenerse en cuenta su desempeño en simultánea con los demás indicadores, los campos se describen más a detalle en el Anexo 2. Tomando esto en cuenta, es posible cuantificar la importancia de cada uno de los indicadores a la hora de segmentar los clientes tanto persona jurídica como persona natural.

3.2. Segmentación

La segmentación será aplicada para personas naturales y jurídicas, con el fin de determinar el grupo de clientes con mayor riesgo LAFT identificando transacciones sospechosas dentro de estos grupos de alto riesgo para tomar acciones a través de alertas que serán aplicadas dentro del mismo desarrollo propuesto.

3.2.1. Personas Jurídicas

En base a las variables mencionadas en el apartado de Procesamiento de datos para la segmentación de clientes jurídicos se obtienen cuatro segmentos de clientes bastante homogéneos y de calidad, tal como se muestra a continuación. De todos los modelos de segmentación candidatos se escoge como definitivo aquel que cumpla con tener buenos indicadores estadísticos (Dunn y siluetas) pero que en simultánea guarde una interpretación lógica de acuerdo con el negocio. La segmentación finalmente implementada se señala en la Tabla 6 con una línea de color azul que contiene el top 25 con las mejores segmentaciones encontradas. Esta segmentación escogida cumple con tener altos valores en el índice Dunn, en la silueta general promedio y en las siluetas individuales, lo cual es evidencia de garantizar homogeneidad al interior de los segmentos y heterogeneidad entre ellos.

	Siluetas					Evaluación	
	1	2	3	4	5	Prom	Dunn
CLARA	0,551	0,679	0,986	0,856	0,321	0,703	0,471
	0,55	0,676	0,986	0,857	0,306	0,699	0,383
	0,548	0,677	0,969	0,843	0,321	0,696	0,471
	0,548	0,677	0,969	0,843	0,303	0,694	0,384

	0,642	0,671	0,806	0,634	NA	0,694	0,737
	0,727	0,599	0,879	0,723	0,304	0,692	0,424
	0,64	0,67	0,796	0,63	NA	0,69	0,733
	0,742	0,569	0,827	0,736	0,382	0,689	0,462
	0,534	0,681	0,985	0,862	0,38	0,689	0,475
	0,542	0,676	0,985	0,863	0,363	0,686	0,397
	0,527	0,679	0,969	0,829	0,422	0,685	0,697
	0,565	0,609	0,986	0,809	0,373	0,685	0,53
	0,711	0,574	0,857	0,672	0,406	0,684	0,64
	0,534	0,681	0,985	0,862	0,347	0,683	0,413
	0,547	0,632	0,985	0,863	0,391	0,683	0,518
	0,534	0,681	0,985	0,862	0,338	0,682	0,347
	0,485	0,687	0,986	0,859	0,217	0,682	0,416
	0,54	0,682	0,879	0,705	0,596	0,682	0,935
	0,665	0,631	0,798	0,733	0,297	0,68	0,319
	0,653	0,69	0,735	0,639	NA	0,68	1175
	0,674	0,812	0,718	0,417	0,395	0,679	0,56
	0,563	0,607	0,97	0,796	0,373	0,678	0,53
	0,662	0,502	0,766	0,606	NA	0,678	0,73
	0,531	0,679	0,969	0,85	0,347	0,677	0,414
	0,534	0,681	0,874	0,866	0,379	0,677	0,534

Tabla 6. Indicadores de Segmentación - Clientes jurídicos

En la Tabla 6 se observa que la selección de 4 segmentos es la que maneja los mejores indicadores de homogeneidad dentro de los grupos y heterogeneidad entre los grupos. Por los motivos anteriores se seleccionaron 4 clústeres para la segmentación definitiva. A continuación, se describe brevemente el perfil de cada uno de los segmentos obtenidos para clientes persona jurídica.

El código fuente de las funciones de iteración que se ejecutaron para los modelos se los encuentra en el Anexo 3.

3.2.1.1. Análisis de segmentos

Monto de retiros en 24 meses

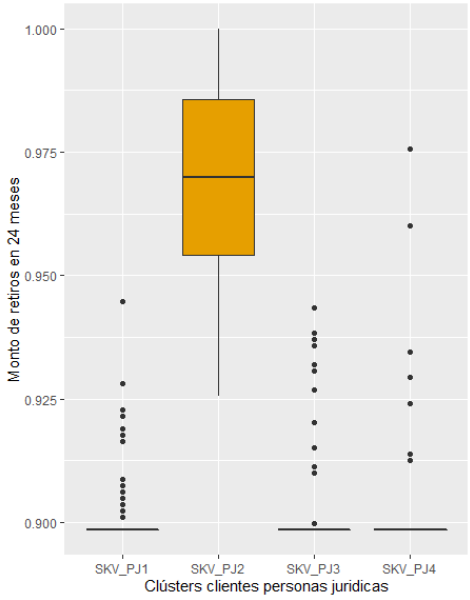


Figura 3. Monto de retiros en 24 meses

En la Figura 3 se observa el box plot de la primera variable determinante en el modelo que es monto de retiros en 24 meses se puede observar que el segmento 2 de personas jurídicas son las que generan transacciones con mayor valor monetario a diferencia del resto de segmentos que realiza transacciones por valores más pero no muy alejados del segmento 2.

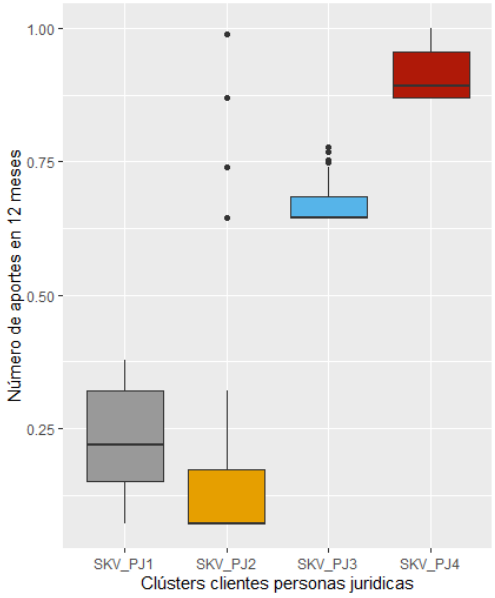


Figura 4. Numero de aportes en 12 meses

En la Figura 4 observamos la variable de número de aportes en 12 meses, en donde se identifica que el segmento 4 y 3 son los que generan mayor cantidad de aportes en un año calendario normal a diferencia del resto de segmentos, incluso se observa algo interesante con respecto al segmento 2 que describe mejor su perfil ya que es el segmento que hace menor cantidad de aportes en un año.

Promedio de retiros en 24 meses

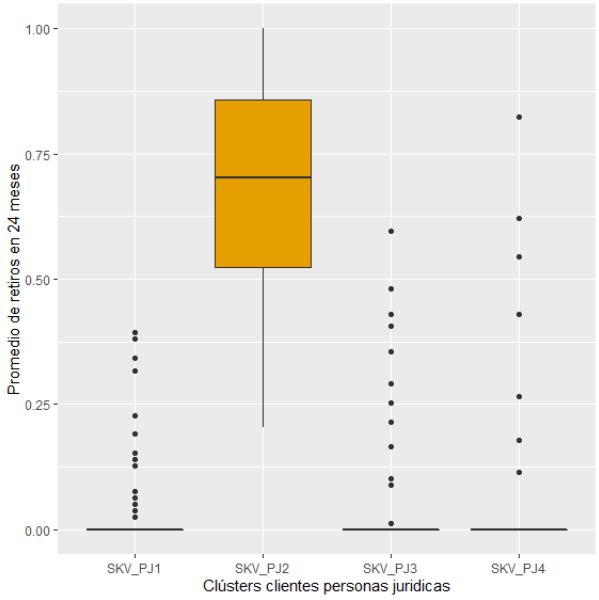


Figura 5. Promedio de retiros en 24 meses

En la Figura 5 observamos la variable de promedio de retiros en 24 meses, en donde el segmento 2 es el que genera en promedio mayor número de retiros en 2 años calendario normales, esto tiene relación directa con el monto de retiros en 24 meses ya que este segmento es que más retiros hace y sus valores de transacciones son altas.

medida	SEGMENTO_SARLAFT_CLIENTES	Id_Fe_Apor_Num_12_Or	Id_Fe_Apor_Num_12	Id_Fe_Ret_Mon_24_Or	Id_Fe_Ret_Mon_24	Id_Fe_Ret_Prom24_Or	Id_Fe_Ret_Prom24
Percentil_5	SKV_PJ1	0,000	0,072	0,000	0,898	0,000	0,000
Percentil_5	SKV_PJ2	0,000	0,072	28.130.999,858	0,940	4.244.974,664	0,285
Percentil_5	SKV_PJ3	11,000	0,645	0,000	0,898	0,000	0,000
Percentil_5	SKV_PJ4	22,000	0,869	0,000	0,898	0,000	0,000
Mediana	SKV_PJ1	6,000	0,219	0,000	0,898	0,000	0,000
Mediana	SKV_PJ2	0,000	0,072	204.447.834,360	0,970	20.762.234,043	0,703
Mediana	SKV_PJ3	11,000	0,645	0,000	0,898	0,000	0,000
Mediana	SKV_PJ4	24,000	0,892	0,000	0,898	0,000	0,000
Percentil_95	SKV_PJ1	10,000	0,379	757.591,616	0,902	325.520,188	0,030
Percentil_95	SKV_PJ2	15,500	0,716	1.853.251.063,585	0,997	566.205.806,198	0,972
Percentil_95	SKV_PJ3	20,000	0,769	0,000	0,898	0,000	0,000
Percentil_95	SKV_PJ4	55,000	0,988	0,000	0,898	0,000	0,000

Figura 6. Distribución en percentiles de variables de segmentación estandarizadas y originales

En la Figura 6 se puede observar una distribución de datos en percentiles de las variables determinantes en el modelo. En percentil 5 de los datos el número de aportes en 12 meses es superior en el segmento llegando hasta 22 aportes en el 5% de sus datos, seguido del segmento 3, para la variable de retiros expresa en montos totales en 24 meses se observa que el segmento 2 es superior al resto en su totalidad superando los 28 millones de pesos en el 5% de sus datos, de igual manera se evidencia el mismo comportamiento en el promedio de retiros en 24 meses en donde es equivalente a los montos de retiros anteriores, en la variable de promedio en donde se supera los 4 millones de pesos en promedio en el 5% de sus datos.

En las medianas de las variables importantes en el modelo se puede observar que en el número de aportes en 12 meses es superior en el segmento 4 llegando hasta los 24 aportes en el 50% de sus datos, seguido del segmento 3 con 11 aportes, para la variable de retiros expresa en montos totales en 24 meses se observa que el segmento 2 es superior al resto en su totalidad superando los 204 millones de pesos en el 50% de sus datos. De igual manera se evidencia el mismo comportamiento en el promedio de retiros en 24 meses en donde es equivalente a

los montos de retiros anteriores, en la variable de promedio llega hasta más de 20 millones de pesos en promedio en el 50% de sus datos.

En el percentil 95 se corrobora el comportamiento de las dos medidas anteriores en donde en el número de aportes en 12 meses el segmento 4 es superior llegando hasta los 55 aportes en el 95% de sus datos. De igual manera el comportamiento tanto en montos totales como en promedio de retiros en 24 meses es superior en el segmento 2, lo que nos dice que el segmento 2 son los que no hacen tantos aportes, pero si muchos retiros expresados en montos y el segmento 4 es que hace mas aportes, pero sus retiros en esta ventana son nulos. En la variable de número de aportes en 12 meses se evidencia una escalera en donde clasifica de forma correcta cada segmento de las personas jurídicas.

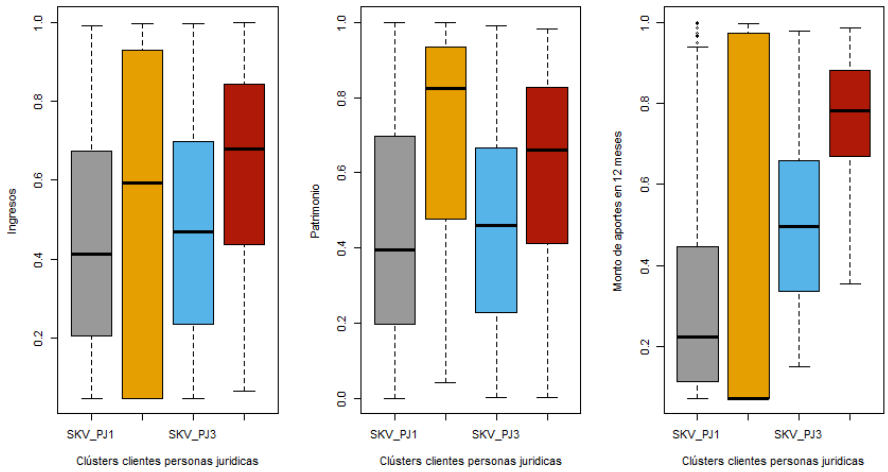


Figura 7. Variables de perfilamiento parte 1 - personas jurídicas

En la Figura 7 podemos observar las variables de perfilamiento para las personas jurídicas, se observa que el segmento 4 tiene mayores ingresos que el resto de los segmentos y tiene prácticamente una relación con la variable de numero de aportes ya que son los que más ganan y los que más hacen aportes, en la variable de patrimonio el segmento 2 es el que presenta valores más altos y el que menos realiza aportes, pero a su vez realiza más número de retiros. En la variable de monto de aportes en 12 meses podemos observar una escalera muy similar a la variable significativa del modelo 'número de aportes' esto es importante resaltarlo

y decir que los números de aportes por lo general van correlacionados con los montos.

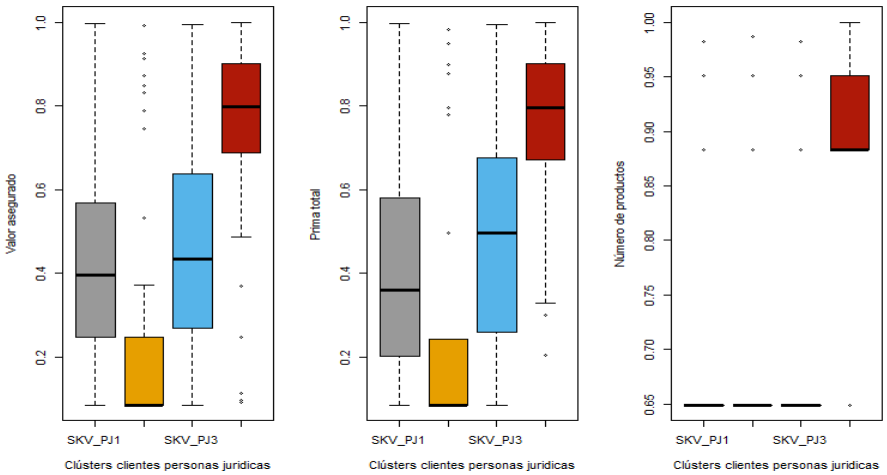


Figura 8. Variables de perfilamiento parte 2 - personas jurídicas

En esta Figura 8 podemos observar que los clientes del segmento 4 son los que más productos tienen y se ve reflejado tanto en valor asegurado y en la prima total, es decir son los clientes que más contratos tienen y puede ser la justificación de que realicen tantos aportes y pocos retiros. El segmento 2 es uno de los más diferenciados y son los clientes que tienen pocos productos en donde realizan pocos aportes y muchos retiros según las variables significativas del modelo.

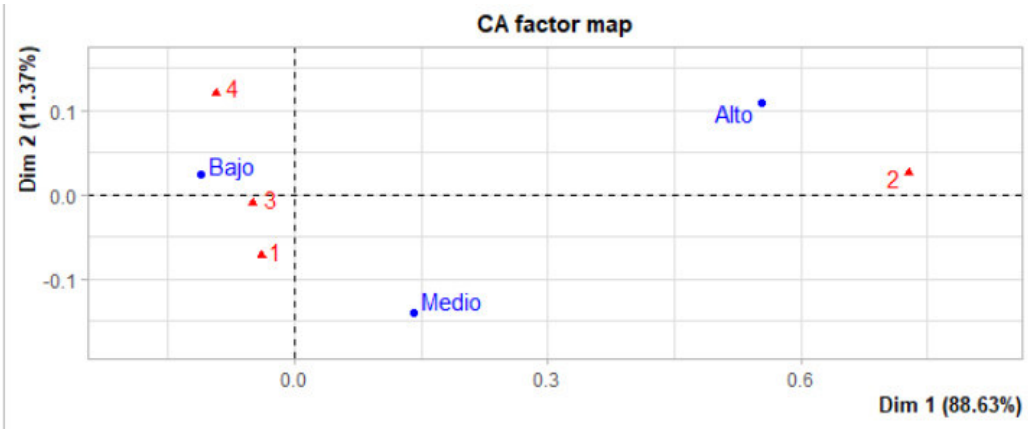


Figura 9. ACP Riesgo de segmentos personas jurídicas

La Figura 9 muestra un análisis de correspondencias principales que cruza los segmentos obtenidos contra la variable CIIU High Risk (nivel de riesgo del CIIU del

cliente), en donde se observa que los clientes del segmento 2 son de mayores proporciones hacia un riesgo Alto, mientras los clientes del segmento 1, 3, 4 tienen mayor tendencia al riesgo bajo.

medida	SEGMENTO_SARLAFT_CLIENTES	Id_Fe_Ing_Or	Id_Fe_Ing	Id_Fe_Patr_Or	Id_Fe_Patr	Id_Fe_Apor_Mon_12_Or	Id_Fe_Apor_Mon_12	Id_Fe_valaseg_total_Or	Id_Fe_valaseg_total
Percentil_5	SKV_PJ1	0,000	0,048	0,000	0,044	0,000	0,072	0,000	0,086
Percentil_5	SKV_PJ2	0,000	0,048	0,500	0,044	0,000	0,072	0,000	0,086
Percentil_5	SKV_PJ3	8.000.000,000	0,082	5.387.811,500	0,058	4.696.169,600	0,203	240.000.000,000	0,248
Percentil_5	SKV_PJ4	30.000.000,000	0,228	15.222.058,000	0,090	11.088.000,000	0,509	480.000.000,000	0,608
Mediana	SKV_PJ1	70.865.083,000	0,413	244.687.964,000	0,396	5.040.000,000	0,224	329.000.000,000	0,396
Mediana	SKV_PJ2	163.000.000,000	0,594	2.397.950.581,000	0,825	0,000	0,072	0,000	0,086
Mediana	SKV_PJ3	90.838.333,000	0,468	337.216.000,000	0,460	11.000.000,000	0,496	350.000.000,000	0,434
Mediana	SKV_PJ4	243.323.458,000	0,679	879.169.500,000	0,660	24.722.817,000	0,781	863.500.000,000	0,798
Percentil_95	SKV_PJ1	1.283.594.200,000	0,908	16.226.337.772,000	0,966	52.450.768,400	0,908	1.192.600.000,000	0,875
Percentil_95	SKV_PJ2	13.936.107.729,250	0,992	80.242.534.500,000	0,994	715.000.000,000	0,993	1.415.250.000,000	0,904
Percentil_95	SKV_PJ3	1.811.626.566,200	0,930	4.713.759.800,000	0,895	49.391.204,000	0,903	1.705.600.000,000	0,925
Percentil_95	SKV_PJ4	3.794.398.230,900	0,963	8.865.941.150,000	0,933	81.086.105,000	0,956	2.816.150.000,000	0,979

Figura 10. Distribución en percentiles de variables de perfilamiento estandarizadas y originales

En la Figura 10 podemos observar cómo se distribuyen los valores de movimientos de las variables de perfilación para identificación de los segmentos de personas jurídicas.

Como primero podemos observar los ingresos en donde en el percentil 5 el segmento 4 es el que está por encima de todos llegando hasta 30 millones de pesos en el 5% de sus datos, de igual manera es el segmento que tienen mayor patrimonio superando los 11 millones de pesos y los que generan mayores aportes en montos en 12 meses siendo equivalente a la variable importante en modelo del número de aportes en 12 meses para este mismo segmento. Al ser el segmento con mayores aportes indica que pueden ser clientes que tienen más de un contrato y se evidencia en el total de valor asegurado que es el segmento que más valor asegurado tiene seguido del segmento 3.

En medianas se observa se observa un comportamiento similar al del percentil 5 en donde el segmento 4 es está por encima del resto en ingresos, en monto de aportes en 12 meses y valor asegurado en mismo periodo, la diferencia que podemos observar en medianas es en la variable de patrimonio en donde el segmento 2 es que más tiene superando los 2.397 millones de pesos. Adicionalmente, se observa que en valor asegurado el segmento 1 y 3 son muy

similares en sus medianas, pero los diferencian sus montos de aportes que generan en 12 meses en donde el segmento 3 tiene alrededor de 11 millones de pesos y el segmento 1 alrededor de los 5 millones en el 50de sus datos para ambos segmentos.

En la distribución de los movimientos para el percentil 95 observamos que el segmento 2 destaca del resto en donde rebasa los 13.936 millones de pesos en ingresos superando por mucho al resto de segmentos. De igual manera se evidencia el mismo comportamiento en egresos y la cantidad de aportes en montos que genera este segmento en 12 meses calendarios, pero no se evidencia lo mismo en el total del valor asegurado en donde el segmento 4 tiene mayor valor asegurado, pero no aporta como el segmento 2.

SEGMENTO PERSONAS JURIDICAS 1	En este grupo se encuentran los clientes que tienen alrededor del 24% de los aportes totales de todos los clientes, se consideran como clientes de aportes medios que a su vez son los que no hacen tantos retiros y se ve reflejado en los montos y en el promedio de retiros tanto en 12 como en 24 meses.
SEGMENTO PERSONAS JURIDICAS 2	En este grupo se encuentran las personas que hacen pocos aportes y muchos retiros tanto en cantidad y montos, esto se equipara con el promedio de retiros en 24 meses.
SEGMENTO PERSONAS JURIDICAS 3	En este grupo se encuentran las personas con numero de aportes medianamente altos y que realizan pocos retiros tanto en cantidad y montos, estos clientes comparados con los dos primeros son muy diferenciados en sus aportes.
SEGMENTO PERSONAS JURIDICAS 4	En este grupo se encuentran las personas con numero de aportes sumamente altos, es decir que son los que más aportan en cantidad y los que realizan pocos retiros, estos clientes pueden ser monitorizados debido a sus aportes altos.

Tabla 7. Descripción resumida de segmentos

La Tabla 7 presenta un resumen de la descripción de los segmentos identificados.

3.2.2. Personas Naturales

En base a las variables mencionadas en el apartado de Procesamiento de datos para la segmentación de clientes naturales se obtienen cuatro segmentos al igual que en clientes Jurídicos, estos segmentos fueron bastante homogéneos y de calidad, tal como se muestra a continuación, las reglas aplicadas fueron las mismas que se utilizaron para la segmentación de clientes jurídicos en donde se comprueba tener altos índices Dunn y Siluetas altas tanto en promedio general como en cada uno de los segmentos.

	Siluetas					Evaluación	
	1	2	3	4	5	Prom	Dunn
CLARA	0,701	0,652	0,648	0,598	NA	0,653	1425
	0,744	0,670	0,254	0,354	NA	0,640	1048
	0,743	0,670	0,253	0,354	NA	0,639	1048
	0,779	0,654	0,313	0,267	NA	0,633	0,937
	0,777	0,653	0,313	0,267	NA	0,632	0,937
	0,762	0,662	0,198	0,280	NA	0,626	0,95
	0,77	0,7	0,3	0,21	NA	0,62	0,843
	0,77	0,7	0,3	0,21	NA	0,62	0,843
	0,79	0,67	0,34	0,11	NA	0,62	0,782
	0,79	0,67	0,34	0,11	NA	0,62	0,782
	0,79	0,68	0,31	0,12	NA	0,62	0,779
	0,79	0,68	0,31	0,12	NA	0,62	0,779
	0,77	0,67	0,32	0,13	NA	0,62	0,689
	0,77	0,67	0,32	0,13	NA	0,61	0,689
	0,77	0,66	0,29	0,16	NA	0,61	0,704
	0,77	0,66	0,29	0,16	NA	0,61	0,704
	0,78	0,66	0,28	0,22	NA	0,61	0,794
	0,78	0,66	0,28	0,22	NA	0,61	0,794

Tabla 8. Indicadores de Segmentación - Clientes naturales

La Tabla 8 muestra el resultado de las iteraciones del modelo utilizando diferentes parámetros y diferentes números de clusters, se observa que la mejor segmentación fue con 4 clusters teniendo buenos resultados en los índices de evaluación.

Las funciones de iteración permitieron el procesamiento de grandes cantidades de combinaciones entre parámetros y segmentos, los mismos tuvieron tiempo de

ejecución acorde a la cantidad de datos que fueron ingresados para modelar, para el apartado de personas naturales el tiempo aproximado de ejecución fue de 4 horas, mientras que para el apartado de personas jurídicas fue de más de 15 horas. Finalmente se obtuvieron resultados acordes a la información y aptos para ser implementados, los códigos de las funciones se encuentran en el Anexo 2.

3.2.2.1. Análisis de segmentos

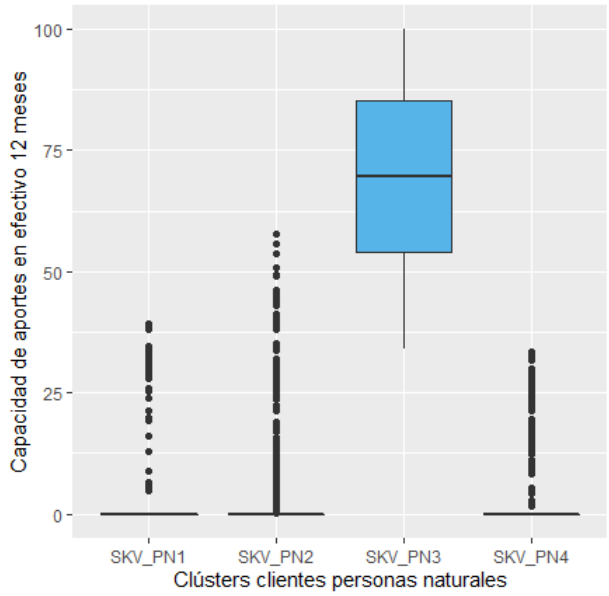


Figura 11. Capacidad de aportes en efectivo en 12 meses

En la Figura 11 observamos el análisis de la variable de capacidad de aportes en efectivo en 12 meses que describe a las personas naturales. Se puede observar que el segmento 3 es que tiene mayor capacidad de aportes en un año calendario, mientras que el resto de los segmentos no tiene una capacidad de aportes alta que contrarreste al segmento 3.

Egresos

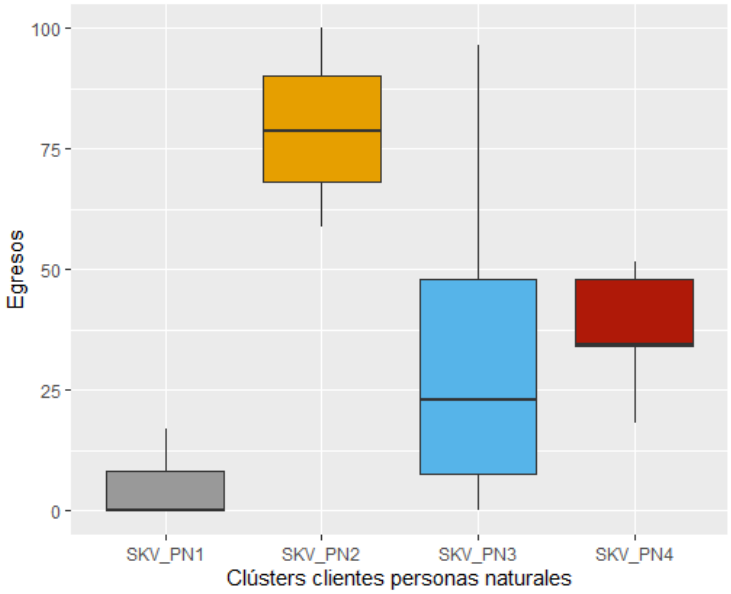


Figura 12. Egresos

En la Figura 12 observamos la distribución de los Egresos en los diferentes segmentos. Se puede evidenciar que en el segmento 2 es en donde los clientes tienen mayores egresos a diferencia del resto de segmentos y el segmento 1 es el que menor egresos genera.

medida	SEGMENTO_SARLAFT_CLIENTES	Id_Fe_Cap_Apor12_E_Or	Id_Fe_Cap_Apor12_E	Id_Fe_Egr_Or	Id_Fe_Egr
Percentil_5	SKV_PN1	0,00	0,96	0,00	0,15
Percentil_5	SKV_PN2	0,00	0,96	4.000.000,00	0,65
Percentil_5	SKV_PN3	0,02	0,98	0,00	0,15
Percentil_5	SKV_PN4	0,00	0,96	1.400.000,00	0,31
Mediana	SKV_PN1	0,00	0,96	0,00	0,15
Mediana	SKV_PN2	0,00	0,96	7.000.000,00	0,82
Mediana	SKV_PN3	0,08	0,99	1.500.000,00	0,35
Mediana	SKV_PN4	0,00	0,96	2.200.000,00	0,44
Percentil_95	SKV_PN1	0,00	0,96	1.000.000,00	0,29
Percentil_95	SKV_PN2	0,00	0,96	20.000.000,00	0,99
Percentil_95	SKV_PN3	1,00	1,00	6.069.090,95	0,78
Percentil_95	SKV_PN4	0,00	0,96	3.500.000,00	0,58

Figura 13. Distribución en percentiles de variables de segmentación estandarizadas y originales

En la Figura 13 podemos observar la distribución de las variables determinantes en el modelo de personas naturales a través de percentiles. Se puede evidenciar que la distribución de movimientos para el percentil 5 es mayor en el segmento 3 en capacidad de aportes en efectivo en 12 meses, pero en egresos el segmento 2 es mayor con más de 4 millones de pesos por encima del 5% de sus datos.

En medianas se observa que el segmento 3 es superior al resto en su capacidad de aportes en efectivo en 12 meses y el segmento 2 es superior en egresos con más de 7 millones de pesos por encima del 50% de sus datos.

En el percentil 95 podemos observar que el segmento 3 se distribuye más arriba que el resto, es decir sus movimientos superan el 95% de la capacidad de aportes en efectivo que el resto de los segmentos, pero en egresos el segmento 2 sigue siendo superior con más de 20 millones de pesos lo que nos indica que el 45% de sus movimientos superan esta cantidad en valores de egresos.

Variables de perfilamiento

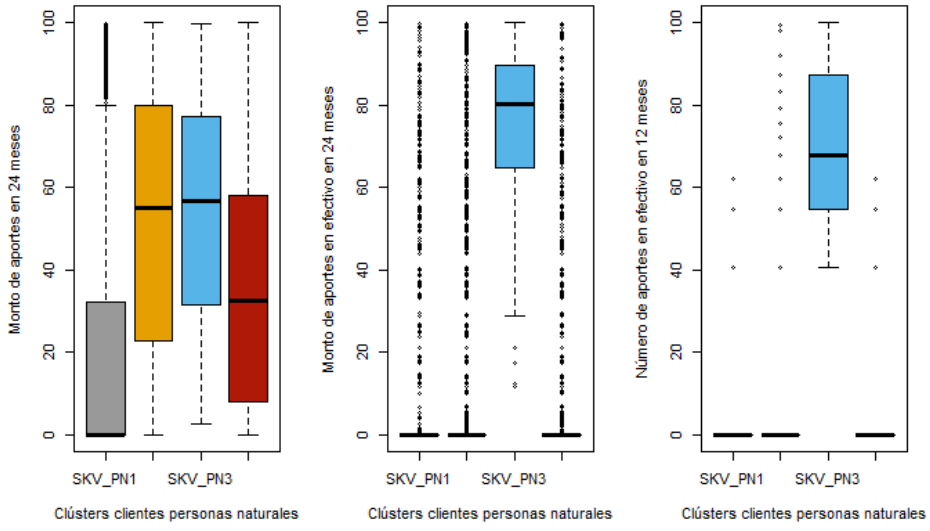


Figura 14. Variables de perfilamiento parte 1 - personas naturales

En la Figura 14 podemos observar la distribución de las variables de perfilamiento para los clientes naturales en los diferentes segmentos. Aquí se observa que las personas del segmento 3 tienen más número de aportes en efectivo y de mayores

montos a diferencia del resto de segmentos, en la variable de monto de aportes en 24 meses los clientes del segmento 1 son los de menores montos de aportes en los 24 meses que a su vez se relaciona con los números de aportes en el mismo periodo de tiempo.

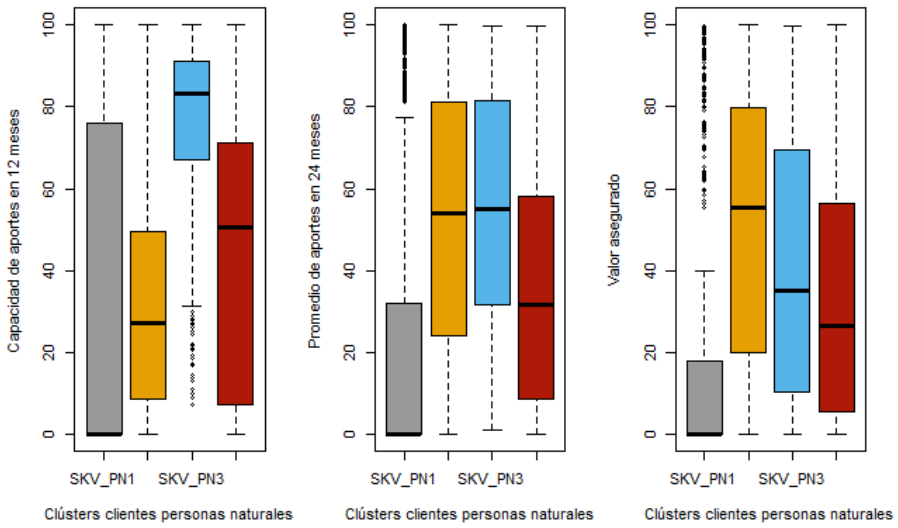


Figura 15. Variables de perfilamiento parte 2 - personas naturales

En la Figura 15 se observa el análisis del resto de variables de perfilamiento que ayudan a detallar el comportamiento de los segmentos, como primero se observa que la capacidad de aportes en 12 meses es más representativa en el segmento 3, es decir que los clientes del segmento 3 tienen mayor capacidad de aportes y egresos medios, los clientes del segmento 1 presentan los valores más bajos en promedio de aportes y valor asegurado, así como egresos mínimos en relación con el resto de los segmentos y baja capacidad de aportes en efectivo en 12 meses.

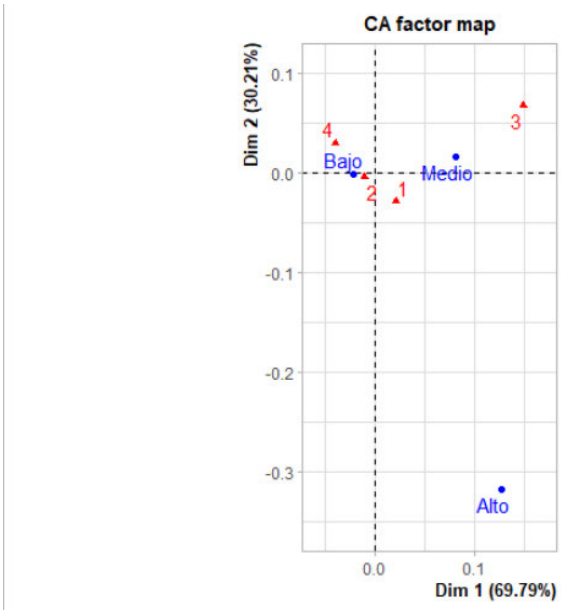


Figura 16. ACP riesgo de segmentos personas naturales

En la Figura 16 se observa el análisis de componentes principales, en donde se evidencia que el segmento 2 y 4 están más cercanos a presentar un riesgo bajo en LAFT, mientras que el segmento1 y 3 están más cercanos a presentar un riego medio y no hay segmentos que presenten un riego alto en LAFT, pero esto no significa que no hay transacciones sospechosas que representen un riego para la entidad y esto se verá reflejado en el resto de variables en donde se deberá tener mayor precaución con los clientes que tengan perfiles delicados.

medida	SEGMENTO_SARLAFT_CLIENTES	Id_Fe_Apor_Mon_24_Or	Id_Fe_Apor_Mon_24	Id_Fe_Apor_E_Mon_24_Or	Id_Fe_Apor_E_Mon_24	Id_Fe_Apor_E_Num_12_Or	Id_Fe_Apor_E_Num_12	Id_Fe_Cap_Apor12_Or	Id_Fe_Cap_Apor12
Percentil_5	SKV_PN1	0,00	0,25	0,00	0,90	0,00	0,96	0,00	0,29
Percentil_5	SKV_PN2	0,00	0,25	0,00	0,90	0,00	0,96	0,00	0,29
Percentil_5	SKV_PN3	1.782.294,75	0,30	852.250,00	0,94	1,00	0,98	0,04	0,49
Percentil_5	SKV_PN4	0,00	0,25	0,00	0,90	0,00	0,96	0,00	0,29
Mediana	SKV_PN1	0,00	0,25	0,00	0,90	0,00	0,96	0,00	0,29
Mediana	SKV_PN2	11.700.000,00	0,66	0,00	0,90	0,00	0,96	0,04	0,48
Mediana	SKV_PN3	12.000.000,00	0,68	6.552.000,00	0,98	4,00	0,99	0,19	0,88
Mediana	SKV_PN4	8.000.000,00	0,50	0,00	0,90	0,00	0,96	0,07	0,65
Percentil_95	SKV_PN1	24.600.000,00	0,88	374.976,00	0,92	0,00	0,96	1,00	0,99
Percentil_95	SKV_PN2	55.060.105,50	0,97	700.000,00	0,94	0,00	0,96	0,18	0,87
Percentil_95	SKV_PN3	63.151.720,70	0,97	35.905.000,00	1,00	11,00	1,00	1,00	0,99
Percentil_95	SKV_PN4	30.188.100,00	0,92	552.117,00	0,93	0,00	0,96	0,29	0,93

Figura 17. Distribución en percentiles de variables de perfilamiento estandarizadas y originales

En la Figura 17 tenemos la distribución de las variables de perfilamiento a través de percentiles. Se observa que la distribución de movimientos para el percentil 5 en montos de aportes en 24 meses es más grande en el segmento 3 superando 1.782.294 pesos, los que nos indica que su aportes son más grandes que el resto

de los segmentos, en los montos de portes en 24 meses de igual manera es más grande en el segmento 3 alcanzando más de 852 mil pesos, así mismo el número de aportes en efectivo es mayor en este segmento ya que es proporcional al resto de variables de perfilación.

En medianas podemos observar que el segmento 2 alcanza más de 11 millones de pesos en sus montos de aportes en 24 meses, seguido del segmento 3 que alcanza los 12 millones en el 5% de sus datos, luego está el segmento 4 con 8 millones y por último el segmento 1 que es en donde por lo general no se hacen aportes en 24 meses, algo muy peculiar sucede en las medianas en los montos de aportes en efectivo en 24 meses en donde ahora el segmento 2 es que más genera aportes en efectivo más que el segmento 2 que no presenta nada de aportes en efectivo, así mismo la capacidad de aportes en 12 meses es mayor en el segmento 2 ya que es proporcional a su montos.

En el percentil 95 podemos observar que es segmento 3 es el que está por encima de los demás alcanzando más de 63 millones en el 5% de sus datos, así mismo sus aportes en efectivo y su capacidad de aportes es mayor que el resto de los segmentos. Algo interesante se observa en la distribución de este percentil en donde a pesar de que el segmento 3 tiene mayores aportes en valores que el segmento 1 en 24 meses, sus capacidades de aportar en 12 meses son iguales.

SEGMENTO PERSONAS NATURALES 1	En este grupo se encuentran los clientes que tienen Egresos bajos en relación con los otros segmentos y los que tienen baja capacidad de aportes en 12 meses, así como en 24 meses, por este motivo se consideran como clientes de egresos bajos y baja capacidad de aportes.
SEGMENTO PERSONAS NATURALES 2	En este grupo se encuentran las personas tienen Egresos muy altos y baja capacidad de aportes, por este motivo se consideran como clientes de egresos muy altos y baja capacidad de aportes.

SEGMENTO PERSONAS NATURALES 3	En este grupo se encuentran las personas tienen Egresos medianamente bajos y alta capacidad de aportes, por este motivo se consideran como clientes de egresos medios y alta capacidad de aportes.
SEGMENTO PERSONAS NATURALES 5	En este grupo se encuentran las personas tienen Egresos medios y baja capacidad de aportes, por este motivo se consideran como clientes de egresos medios y baja capacidad de aportes.

Tabla 9. Descripción resumida de segmentos

En la Tabla 9 podemos observar una descripción resumida de los segmentos de personas naturales tomando como principal las variables determinantes del modelo.

3.3. Clasificación

Para el apartado de clasificación de datos de prueba, posterior a realizar los modelos de segmentación se realizó la validación mediante el modelo Random Forest que fue el utilizado en los dos factores de riesgo tanto personas naturales como jurídicas.

3.3.1. Personas Jurídicos

A continuación, se muestra el resultado obtenido en los modelos de clasificación para el apartado de clientes jurídicos.

Entrenamiento

		Predicción				
		1	**2**	**3**	**4**	**error**
Observación	**1**	2583	47	604	189	**0.2454**
	2	18	637	10	9	**0.0549**
	3	676	0	2215	288	**0.3032**
	4	45	0	65	2033	**0.0513**

Tabla 10. Resultados de la matriz de confusión del entrenamiento P.J.

Accuracy

$$Accuracy = \frac{2583 + 637 + 2215 + 2033}{2583 + 47 + 604 + 189 + 18 + 637 + 10 + 9 + 676 + 2215 + 288 + 45 + 65 + 2033}$$

$$Accuracy = \frac{7468}{9419}$$

$$Accuracy = 0.7929$$

$$Accuracy = 79.29\%$$

En la Tabla 10 podemos observar los resultados de la matriz de confusión luego de aplicar el modelo al dataset de entrenamiento. Observamos que tenemos un accuracy bastante bueno del 79.29% al evaluar el modelo global de los 4 clusters, lo que nos indica que los valores medidos de personas jurídicas están muy cerca de los valores reales de las características medidas, por lo que la medición es moderadamente precisa.

Validación

		Predicción				
Observación		**1**	**2**	**3**	**4**	**error**
	1	1053	18	305	91	**0.2827**
	2	14	271	3	1	**0.0687**
	3	300	1	913	148	**0.3311**
	4	23	0	36	860	**0.0683**

Tabla 11. Resultados de la matriz de confusión de la validación P.J.

Accuracy

$$Accuracy = \frac{1053 + 271 + 913 + 860}{1053 + 18 + 305 + 91 + 14 + 271 + 3 + 1 + 300 + 1 + 913 + 149 + 23 + 36 + 860}$$

$$Accuracy = \frac{3097}{4037}$$

$$Accuracy = 0.7672$$

$$Accuracy = 76.72\%$$

En la Tabla 11 podemos observar los resultados del dataset de comprobación en donde tenemos un asertividad moderada del 76.72% y similar a los resultados de entrenamiento lo que nos asegura la inexistencia de overfitting, esto muestra que los valores medidos de las observaciones están medianamente cerca de los valores reales.

El resultado obtenido nos dice que al momento de entrar a producción los modelos tendrán una eficiencia moderada al momento de clasificar las transacciones de los clientes jurídicos según su nivel de riesgo LAFT en cada uno de los segmentos, lo que permitirá tener una adecuada gestión y seguimiento a dichos clientes que presenten mayor nivel de riesgo.

3.3.2. Personas Naturales

A continuación, se muestra el resultado obtenido en los modelos de clasificación para el apartado de clientes naturales.

Entrenamiento

		Predicción				
Observación		**1**	**2**	**3**	**4**	**error**
	1	96613	1517	4	3804	**0.0522**
	2	11447	79599	7	2398	**0.1482**
	3	19	1	15294	10	**0.002**
	4	16046	4253	22	63339	**0.2429**

Tabla 12. Resultados de la matriz de confusión del entrenamiento P.N.

Accuracy

$$Accuracy = \frac{96613 + 79599 + 15294 + 63339}{96613 + 1517 + 4 + 3804 + 11447 + 79599 + 7 + 2398 + 19 + 1 + 15294 + 10 + 16046 + 4253 + 22 + 63339}$$

$$Accuracy = \frac{254845}{294373}$$

$$Accuracy = 0.8657$$

$$Accuracy = 86.57\%$$

En la Tabla 12 observamos los resultados del dataset de entrenamiento de personas naturales. Se identifica que tenemos un asertividad buena del 86.57% en los 4 clusters, lo que nos dice que los valores medidos observados de los clientes naturales están bastante cerca de los valores objetivos, en base a esto el modelo es adecuado para clasificar clientes naturales ya que su medición es muy precisa.

Validación

		Predicción				
Observación		**1**	**2**	**3**	**4**	**error**
	1	40167	923	25	2573	**0.0806**
	2	4995	33134	43	1879	**0.1727**
	3	24	33	6469	41	**0.0149**
	4	7499	3193	98	25065	**0.3009**

Tabla 13. Resultados de la matriz de confusión de la validación P.N.

Accuracy

$$Accuracy = \frac{40167 + 33134 + 6469 + 25065}{40167 + 923 + 25 + 2573 + 4995 + 33134 + 43 + 1879 + 24 + 33 + 6469 + 41 + 7499 + 3193 + 98 + 25065}$$

$$Accuracy = \frac{104835}{126161}$$

$$Accuracy = 0.831$$

$$Accuracy = 83.1\%$$

En la Tabla 13 observamos los resultados de aplicar el dataset de comprobación, en donde tenemos un asertividad moderadamente buena del 83.1% que equipara a los resultados de entrenamiento y sus mediciones son bastante precisas descartamos la existencia de overfitting.

Para el modelo de personas naturales tenemos un asertividad muy buena que equipara al modelo de clientes jurídicos, esto nos indica que el nivel de eficiencia del modelo al momento de analizar las transacciones de clientes naturales será suficiente y ubicará a los clientes con mayor riesgo LAFT en sus segmentos

adecuados lo que ayudara a la entidad a tener un seguimiento más detallado con dichos clientes.

3.4. Métricas de evaluación

3.4.1. Personas Jurídicas

3.4.1.1. Precisión

Segmento 1

$$Precision = \frac{TP}{TP + (FP1 + FP2 + FP3 + \cdots + FPn)}$$

$$Precision = \frac{1053}{1053 + 14 + 300 + 23}$$

$$Precision = 0.7576$$

$$Precision = 75.76\%$$

El resultado de precisión del segmento 1 indica que del total de las predicciones dadas por el modelo el 75.76% de clientes jurídicos se predijeron de manera correcta dentro del segmento 1, mientras que el resto de las predicciones se distribuyen entre los demás segmentos, esto muestra una dispersión bastante buena ya que aproximadamente el 76% de los casos predichos se encuentran cercanos entre ellos ubicándose dentro del segmento 1.

Segmento 2

$$Precision = \frac{TP}{TP + (FP1 + FP2 + FP3 + \cdots + FPn)}$$

$$Precision = \frac{271}{271 + 18 + 1}$$

$$Precision = 0.9345$$

$$Precision = 93.45\%$$

El resultado de precisión del segmento 2 indica que del total de las predicciones dadas por el modelo el 93.45% de clientes jurídicos se predijeron de manera correcta dentro del segmento 2, mientras que el resto de las predicciones se distribuyen entre los demás segmentos, esto muestra una dispersión bastante buena ya que aproximadamente el 94% de los casos predichos se encuentran cercanos entre ellos ubicándose dentro del segmento 2.

Segmento 3

$$Precision = \frac{TP}{TP + (FP1 + FP2 + FP3 + \cdots + FPn)}$$

$$Precision = \frac{913}{913 + 305 + 3 + 36}$$

$$Precision = 0.7263$$

$$Precision = 72.63\%$$

El resultado de precisión del segmento 3 indica que del total de las predicciones dadas por el modelo el 72.63% de clientes jurídicos se predijeron de manera correcta dentro del segmento 3, mientras que el resto de las predicciones se distribuyen entre los demás segmentos, esto muestra una dispersión bastante buena ya que aproximadamente el 73% de los casos predichos se encuentran cercanos entre ellos ubicándose dentro del segmento 3.

Segmento 4

$$Precision = \frac{TP}{TP + (FP1 + FP2 + FP3 + \cdots + FPn)}$$

$$Precision = \frac{860}{860 + 91 + 1 + 148}$$

$$Precision = 0.7818$$

$$Precision = 78.18\%$$

El resultado de precisión del segmento 4 indica que del total de las predicciones dadas por el modelo el 78.18% de clientes jurídicos se predijeron de manera

correcta dentro del segmento 4, mientras que el resto de las predicciones se distribuyen entre los demás segmentos, esto muestra una dispersión bastante buena ya que aproximadamente el 78% de los casos predichos se encuentran cercanos entre ellos ubicándose dentro del segmento 4.

3.4.1.2. Recall (Exhaustividad)

Segmento 1

$$Recall = \frac{TP}{TP + (FN1 + FN2 + FN3 + \cdots + FNn)}$$

$$Recall = \frac{1053}{1053 + 18 + 305 + 91}$$

$$Recall = 0.7178$$

$$Recall = 71.78\%$$

El resultado de Recall del segmento 1 muestra que del total de clientes jurídicos que pertenecen al segmento 1, el 71.78% resultaron ser clientes pertenecientes a ese grupo, es decir que ese porcentaje de clientes tiene las características suficientes para ser clasificados dentro de ese segmento.

Segmento 2

$$Recall = \frac{TP}{TP + (FN1 + FN2 + FN3 + \cdots + FNn)}$$

$$Recall = \frac{271}{271 + 14 + 3 + 1}$$

$$Recall = 0.9377$$

$$Recall = 93.77\%$$

El resultado de Recall del segmento 2 muestra que del total de clientes jurídicos que pertenecen al segmento 2, el 93.77% resultaron ser clientes pertenecientes a ese grupo, es decir que ese porcentaje de clientes tiene las características suficientes para ser clasificados dentro de ese segmento.

Segmento 3

$$Recall = \frac{TP}{TP + (FN1 + FN2 + FN3 + \cdots + FNn)}$$

$$Recall = \frac{913}{913 + 300 + 1 + 148}$$

$$Recall = 0.6703$$

$$Recall = 67.03\%$$

El resultado de Recall del segmento 3 muestra que del total de clientes jurídicos que pertenecen al segmento 3, el 67.03% resultaron ser clientes pertenecientes a ese grupo, es decir que ese porcentaje de clientes tiene las características suficientes para ser clasificados dentro de ese segmento.

Segmento 4

$$Recall = \frac{TP}{TP + (FN1 + FN2 + FN3 + \cdots + FNn)}$$

$$Recall = \frac{860}{860 + 23 + 36}$$

$$Recall = 0.9358$$

$$Recall = 93.58\%$$

El resultado de Recall del segmento 4 muestra que del total de clientes jurídicos que pertenecen al segmento 4, el 93.58% resultaron ser clientes pertenecientes a ese grupo, es decir que ese porcentaje de clientes tiene las características suficientes para ser clasificados dentro de ese segmento.

3.4.1.3. F1

Segmento 1

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

$$F1 = 2 * \frac{0.7576 * 0.7178}{0.7576 + 0.7178}$$
$$F1 = 0.7372$$
$$F1 = 73.72\%$$

El resultado de F1 muestra que al comparar los dos clasificadores de precisión y recall tenemos un resultado de 73.72% lo que muestra una buena distribución de verdaderos positivos dentro del modelo.

Segmento 2

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$
$$F1 = 2 * \frac{0.9345 * 0.9377}{0.9345 + 0.9377}$$
$$Precision = 0.9361$$
$$Precision = 93.61\%$$

El resultado de F1 muestra que al comparar los dos clasificadores de precisión y recall tenemos un resultado de 93.61% lo que muestra una excelente distribución de verdaderos positivos dentro del modelo.

Segmento 3

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$
$$F1 = 2 * \frac{0.7263 * 0.6703}{0.7263 + 0.6703}$$
$$Precision = 0.6972$$
$$Precision = 69.72\%$$

El resultado de F1 muestra que al comparar los dos clasificadores de precisión y recall tenemos un resultado de 69.72% lo que muestra una distribución moderadamente buena de verdaderos positivos dentro del modelo.

Segmento 4

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

$$F1 = 2 * \frac{0.7818 * 0.9358}{0.7818 + 0.9358}$$

$$Precision = 0.8519$$

$$Precision = 85.19\%$$

El resultado de F1 muestra que al comparar los dos clasificadores de precisión y recall tenemos un resultado de 85.19% lo que muestra una buena distribución de verdaderos positivos dentro del modelo.

3.4.2. Personas Naturales

3.4.2.1. Precisión

Segmento 1

$$Precision = \frac{TP}{TP + (FP1 + FP2 + FP3 + \cdots + FPn)}$$

$$Precision = \frac{40167}{40167 + 4995 + 24 + 7499}$$

$$Precision = 0.7624$$

$$Precision = 76.24\%$$

El resultado de precisión del segmento 1 indica que del total de las predicciones dadas por el modelo el 76.24% de clientes naturales se predijeron de manera correcta dentro del segmento 1, mientras que el resto de las predicciones se distribuyen entre los demás segmentos, esto muestra una dispersión bastante buena ya que aproximadamente el 76% de los casos predichos se encuentran cercanos entre ellos ubicándose dentro del segmento 1.

Segmento 2

$$Precision = \frac{TP}{TP + (FP1 + FP2 + FP3 + \cdots + FPn)}$$

$$Precision = \frac{33134}{33134 + 923 + 33 + 3193}$$

$$Precision = 0.8887$$

$$Precision = 88.87\%$$

El resultado de precisión del segmento 2 indica que del total de las predicciones dadas por el modelo el 88.87% de clientes naturales se predijeron de manera correcta dentro del segmento 2, mientras que el resto de las predicciones se distribuyen entre los demás segmentos, esto muestra una dispersión bastante buena ya que aproximadamente el 89% de los casos predichos se encuentran cercanos entre ellos ubicándose dentro del segmento 2.

Segmento 3

$$Precision = \frac{TP}{TP + (FP1 + FP2 + FP3 + \cdots + FPn)}$$

$$Precision = \frac{6469}{6469 + 24 + 43 + 98}$$

$$Precision = 0.975$$

$$Precision = 97.5\%$$

El resultado de precisión del segmento 3 indica que del total de las predicciones dadas por el modelo el 97.5% de clientes naturales se predijeron de manera correcta dentro del segmento 3, mientras que el resto de las predicciones se distribuyen entre los demás segmentos, esto muestra una dispersión bastante buena ya que aproximadamente el 98% de los casos predichos se encuentran cercanos entre ellos ubicándose dentro del segmento 3.

Segmento 4

$$Precision = \frac{TP}{TP + (FP1 + FP2 + FP3 + \cdots + FPn)}$$

$$Precision = \frac{25065}{25065 + 2573 + 1879 + 41}$$

$$Precision = 0.848$$

$$Precision = 84.8\%$$

El resultado de precisión del segmento 4 indica que del total de las predicciones dadas por el modelo el 84.8% de clientes naturales se predijeron de manera correcta dentro del segmento 4, mientras que el resto de las predicciones se distribuyen entre los demás segmentos, esto muestra una dispersión bastante buena ya que aproximadamente el 85% de los casos predichos se encuentran cercanos entre ellos ubicándose dentro del segmento 4.

3.4.2.2. Recall (Exhaustividad)

Segmento 1

$$Recall = \frac{TP}{TP + (FN1 + FN2 + FN3 + \cdots + FNn)}$$

$$Recall = \frac{40167}{40167 + 923 + 25 + 2573}$$

$$Recall = 0.9194$$

$$Recall = 91.94\%$$

El resultado de Recall del segmento 1 muestra que del total de clientes naturales que pertenecen al segmento 1, el 91.24% resultaron ser clientes con características que describen a dicho segmento, es decir 9 de cada 10 clientes del segmento 1 fueron identificados como miembros de ese segmento.

Segmento 2

$$Recall = \frac{TP}{TP + (FN1 + FN2 + FN3 + \cdots + FNn)}$$

$$Recall = \frac{33134}{33134 + 4995 + 43 + 1879}$$

$$Recall = 0.8273$$

$$Recall = 82.73\%$$

El resultado de Recall del segmento 2 muestra que del total de clientes naturales que pertenecen al segmento 2, el 82.73% resultaron ser clientes pertenecientes a ese grupo, es decir 8 de cada 10 clientes del segmento 2 fueron identificados como miembros de ese segmento.

Segmento 3

$$Recall = \frac{TP}{TP + (FN1 + FN2 + FN3 + \cdots + FNn)}$$

$$Recall = \frac{6469}{6469 + 24 + 33 + 41}$$

$$Recall = 0.9851$$

$$Recall = 98.51\%$$

El resultado de Recall del segmento 3 muestra que del total de clientes naturales que pertenecen al segmento 3, el 98.51% resultaron ser clientes pertenecientes a ese grupo, es decir que ese porcentaje de clientes tiene las características suficientes para ser clasificados dentro de ese segmento.

Segmento 4

$$Recall = \frac{TP}{TP + (FN1 + FN2 + FN3 + \cdots + FNn)}$$

$$Recall = \frac{25065}{25065 + 7499 + 3193 + 98}$$

$$Recall = 0.6991$$

$$Recall = 69.91\%$$

El resultado de Recall del segmento 4 muestra que del total de clientes naturales que pertenecen al segmento 4, el 69.91% resultaron ser clientes pertenecientes a ese grupo, es decir que ese porcentaje de clientes tiene las características suficientes para ser clasificados dentro de ese segmento.

3.4.2.3. F1

Segmento 1

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

$$F1 = 2 * \frac{0.7624 * 0.9194}{0.7624 + 0.9194}$$

$$F1 = 0.8608$$

$$F1 = 86.08\%$$

El resultado de F1 muestra que al comparar los dos clasificadores de precisión y recall tenemos un resultado de 86.08% lo que muestra una buena distribución de verdaderos positivos dentro del modelo.

Segmento 2

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

$$F1 = 2 * \frac{0.8887 * 0.8273}{0.8887 + 0.8273}$$

$$F1 = 0.8569$$

$$F1 = 85.69\%$$

El resultado de F1 muestra que al comparar los dos clasificadores de precisión y recall tenemos un resultado de 85.69% lo que muestra una excelente distribución de verdaderos positivos dentro del modelo.

Segmento 3

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

$$F1 = 2 * \frac{0.975 * 0.9851}{0.975 + 0.9851}$$

$$F1 = 0.98$$

$$F1 = 98\%$$

El resultado de F1 muestra que al comparar los dos clasificadores de precisión y recall tenemos un resultado de 98% lo que muestra una excelente distribución de verdaderos positivos dentro del modelo.

Segmento 4

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

$$F1 = 2 * \frac{0.848 * 0.6991}{0.848 + 0.6991}$$

$$F1 = 0.7664$$

$$F1 = 76.64\%$$

El resultado de F1 muestra que al comparar los dos clasificadores de precisión y recall tenemos un resultado de 76.64% lo que muestra una distribución moderadamente buena de verdaderos positivos dentro del modelo.

3.5. Implementación

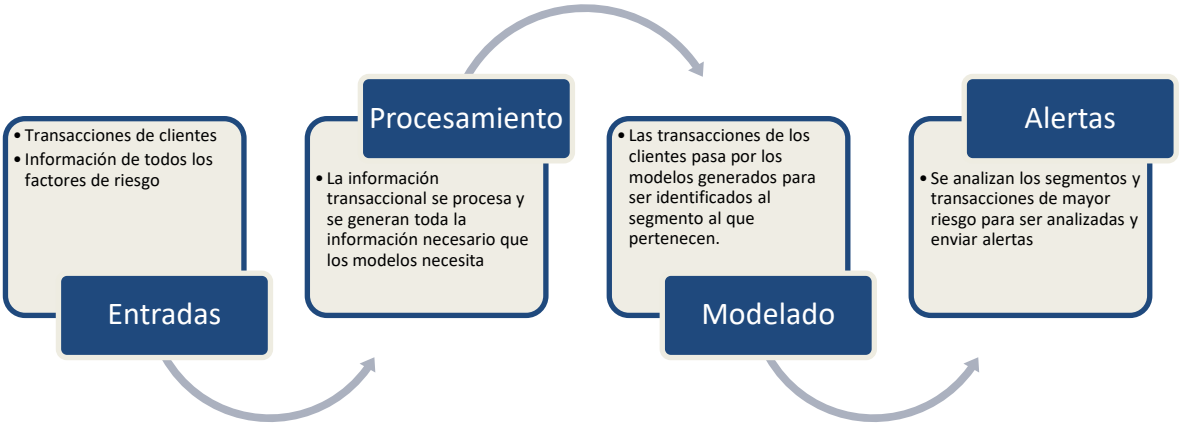


Figura 18. Proceso de implementación

En la Figura 18 se presenta el proceso para la implementación de los modelos dentro de la infraestructura de la entidad. Los pasos son los siguientes:

1. Se consume la información de la base de datos a través de ODBC, esta consulta contiene las tablas de todos los factores de riesgo a considerarse en la construcción de los modelos.
2. Se procesa la información generando las nuevas variables y recodificando las ya existentes, esto quiere decir que se llega a estandarizar toda la información tal cual se construyó los modelos ya que serán los inputs que tomarán los modelos se segmentación.
3. Se llama a los modelos generados que previamente fueron guardados en archivos RData que contienen el área de trabajo de un entorno en R, en este caso son objetos de R.
4. Los datos finales del dataset pasan por los modelos en donde cada transacción es clasificada dependiente el modelo aplicado (personas naturales, personas jurídicas) y es asignada a su segmento.
5. Se analiza la información del segmento con mayor riesgo LAFT y se generan alertas dependiendo de los datos encontrados, las reglas para ser

consideradas transacciones de alto riesgo son impuesta por la misma entidad en conjunto con el equipo de desarrollo.

6. Las transacciones con mayor índice de riesgo LAFT son exportadas en archivos planos, almacenadas y enviadas mediante correo automáticos al personal de la entidad que tomaran las medidas pertinentes en contra de los clientes. Es importante mencionar que todo el proceso automático de alertas es generado a día caído, es decir que se lo hace en la madrugada de cada día, procesando todo lo del día anterior.

Alertas de calidad y poblamiento

Alertas	Alertas de calidad y poblamiento
Alert_CP1	"Actualizar INGRESOS (valor en extremo bajo)",
Alert_CP2	"Actualizar EGRESOS (valor en extremo bajo)",
Alert_CP3	"Actualizar ACTIVOS (valor en extremo bajo)",
Alert_CP4	"Actualizar INGRESOS (valor null - especificar valor de ingresos)",
Alert_CP5	"Actualizar EGRESOS (valor null - especificar valor de egresos)",
Alert_CP6	"Actualizar ACTIVOS (valor null - especificar valor de activos)",
Alert_CP7	"Actualizar PASIVOS (valor null - especificar valor de pasivos)",
Alert_CP8	"Actualizar CIIU Actividad económica (valor null - especificar CIIU)",
Alert_CP9	"Actualizar CIIU Actividad económica (corregir CIIU al formato adecuado)",
Alert_CP10	"Actualizar info de ubicación (valor null - especificar jurisdicción de ubicación)",
Alert_CP11	"Actualizar los valores de patrimonio, activos y/o pasivos (no hay congruencia entre el patrimonio reportado y la resta activos-pasivos)",
Alert_CP12	"Actualizar los valores de todo el registro (cliente con alto riesgo LAFT y datos desactualizados más de 1 año)",
Alert_CP13	"Actualizar los valores de todo el registro (cliente con riesgo LAFT bajo o medio y datos desactualizados más de 3 años)",
Alert_CP14	"Actualizar CIIU_HighRisk (valor null o vacío)

Tabla 14. Alertas de calidad y poblamiento

En la Tabla 14 se puede observar las alertas de calidad y poblamiento, en donde se notifica los campos necesarios para la segmentación con información incompleta o errónea, las alertas con excesos de valores que superan los límites establecidos, alertas relacionadas con los resultados de los modelos.

Las alertas de movimientos inusuales se describen por el tipo de transacciones que manejan y las cantidades asociadas a los modelos de segmentación, los códigos a detalle de las alertas de operaciones inusuales se encuentran en el Anexo 4.

4. CONCLUSIONES Y RECOMENDACIONES

1. Para el control de actividades ilícitas dentro de las entidades financieras es importante implementar mecanismos tecnológicos que permitan un control adecuado y riguroso de estas actividades enfocadas en el comportamiento de clientes que presentan un mayor riesgo LAFT.

2. Los procesos actuales que se implementan dentro de la aseguradora del caso de estudio están basados en reglas duras y ecuaciones matemáticas que miden el grado de riesgo de una persona por su número de transacciones o niveles de montos que maneja. Esto limita una adecuada gestión de los procesos de seguimientos ya que pueden presentarse casos en donde el mayor riesgo existe en situaciones totalmente desapercibidas por la entidad, debido a esto los problemas de LAFT se han incrementado y se exigen tomar acciones para controlar dichos problemas. El presente trabajo presenta una solución al problema mencionado.

3. Para el control adecuado de riesgos LAFT se consideraron 4 segmentos bien marcados en donde es posible identificar en base a modelos de clustering cuales son los grupos con mayor y menor riesgo, considerando su información financiera y aplicando análisis estadísticos y matemáticos que generan alertar dentro la infraestructura de la entidad para toma de decisiones y seguimiento los modelos.

4. La construcción de los modelos se basó en dos etapas, en donde la primera se la hizo mediante modelos no supervisados para poder agrupar en segmentos y etiquetar a los clientes según su información transaccional; mientras que la segunda etapa fue validar las etiquetas con modelos supervisados para mediar el nivel de asertividad y error en la clasificación de registros de prueba.

5. Las medidas de validación aplicadas permitieron identificar la asertividad y error de los modelos con exactitud, en donde la accuracy de clientes jurídicos

fue de 76.72% y en clientes naturales fue de 83.1%. Dando un resultado adecuado en los modelos tomando en consideración que no será la única medida de validación puesto que se evaluarán a los clientes bajo las variables de perfilamiento para tener una clasificación más acertada.

6. Para el desarrollo de los modelos de segmentación en donde no se cuenta con una variable de salida es importante partir de la construcción de un modelo no supervisado en donde las características de los clientes permitirán agruparlos en clusters en donde serán más homogéneos entre ellos y heterogéneos entre los grupos, a partir de esto se puede realizar la construcción del modelos de clasificación para medir que tan marcadas son las características de los grupos y la asertividad del modelo a la hora de clasificar a los clientes.

7. El método realizado se puede implementar dentro de cualquier institución financiera en donde se desee conocer de mejor manera el comportamiento de sus clientes basados en información transaccional y externa relacionada con el entorno y problema que se desea resolver, permitiendo entender mejor los grupos de clientes en donde es primordial focalizar los esfuerzos de seguimiento y de fidelización o renovación de contratos, según sea el caso.

8. El proyecto tiene una iniciativa por parte de la entidad de mantener una relación de consultoría con el fin de dar mantenimiento a los modelos implementados. Se plantea cada cierto periodo de tiempo volverlos a ejecutar y alimentarnos con nueva información ya que se puede conocer nuevos mecanismos de evasión de LAFT y esto permitirá construir modelos cada vez más robustos y se mantendrá el método actualizado dentro de la entidad.

9. Dentro de la construcción de los modelos es importante apalancarse de variables de perfilamiento que describan de forma más detallada el comportamiento y características de los clientes ya que no en todos los

casos se logra identificar las variables de entrada que aportan en gran magnitud a los modelos, por tal motivo es importante reforzar la descripción de los segmentos la con las variables que no entraron en el modelo, pero que si presentan información relevante a la hora de clasificar a los clientes.

5. REFERENCIAS BIBLIOGRÁFICAS

[1] Á. Toso, «Prevención del lavado de activos y crédito documentario: ¿a quién debe conocer el banco emisor? una respuesta desde el derecho privado,» *Revista de Derecho Universidad Católica del Norte,* nº 2, pp. 401-436, 2014.

[2] X. C. X. Q. J. Z. a. J. Z. Xurui Li, «Intelligent Anti-Money Laundering Solution Based upon Novel Community Detection in Massive Transaction Networks on Spark,» de *Fifth International Conference on Advanced Cloud and Big Data (CBD)*, 2017.

[3] H. W. C. S. J. G. a. D. X. Yingfeng Wang, «Intelligent money laundering monitoring and detecting system,» de *Mediterranean and Middle Eastern Conference on Information Systems 2008*, European, 2008.

[4] P. Kotler and K. Keller, Dirección de Marketing, Décimo Quinta Edición ed., México: Pearson Education, 2016.

[5] G. Corona, Comportamiento del consumidor, México: RED TERCER MILENIO S.C., 2012.

[6] A. Domínguez y S. Hermo, Métricas del marketing, Madrid: ESIC, 2007.

[7] O. Caloca and C. Leriche, "Una revisión de la teoría del consumidor: la versión de la teoría del error," *Revista Análisis Económico,* vol. 26, no. 61, pp. 21-51, 2011.

[8] H. Ansoff, Estrategias de marketing a través de la matriz de Ansoff, Pamplona: McGraw Hill, 2016.

[9] A. Martínez, C. Ruíz and J. Escrivá, Marketing en la actividad comercial, Madrid: Mc Graw Hill, 2014.

[10] M. Brano and T. Drazena, "Sistema inteligente de marketing para la segmentación de clientes," *Studies in Fuzziness and Soft Computing ,* pp. 79-111, 2019.

[11] J. Mueller and L. Massaron, Machine Learning For Dummies, Washington: Edición Kindle, 2017.

[12] J. Kelleher, B. MacNamee and A. D'Arcy, Fundamentals of Machine Learning for Predictive Data Anayltics: Algorithms, Worked Examples, and Case Studies, San Gregorio: Mareen Books, 2018.

[13] O. Theobald, Machine Learning For Absolute Beginners, Boston: Edición Kindle, 2017.

[14] Agencia de los Estados Unidos para el Desarrollo Internacional, Aspectos Dogmáticos, Criminológicos y Procesales del Lavado de Activos, Washington: Bridges. Leading Conversations, 2019.

[15] U. Sieber, "Lavado de dinero: estandarización y criminalización," *Notas para una política criminal contra el crimen financiero,* pp. 68-92, 2014.

[16] G. Jakobs, " Informe de Evaluación Mutua sobre Lavado de Activos," *Terroristen als Personen im Recht,* pp. 78-96, 2015.

[17] M. Taipe, "Estrategías para mitigar el riesgo de lavado de activos en los sectores empresariales vulnerables del Ecuador," Pontificia Universidad Católica del Ecuador, Quito, 2016.

[18] A. Segovia, "Lavado de dinero y financiamiento del terrorismo en Chile," *Lucha contra el lavado de activos y financiamiento del terrorismo: principales riesgos y tendencias,* pp. 78-96, 2014.

[19] A. Melgar, "Notas de financiamiento del terrorismo en el mundo," *Jurisdicciones de alto riesgo y no cooperantes,* pp. 86-108, 2014.

[20] J. Marteahu, "Lavado de dinero: estandarización y criminalización," *Enfoques jurídicos españoles,* pp. 66-89, 2013.

[21] F. Amores, "Administración del Riesgo de Lavado de Activos y de la Financiación del Terrorismo en Instituciones Financieras, caso Corporación Financiera Nacional en Operaciones de Segundo Piso (CFN)," UASB, Quito, 2016.

[22] Banco de Desarrollo de Latinoamérica, Análisis de la normativa para la prevención de lavado de activos, Quito: Imprenta Don Bosco, 2013.

[23] R. González, "Matriz de evaluación a la dimensión y control de los factores críticos de riesgo referente a la prevención de lavado de activos y

financiamiento del terrorismo en una entidad del sistema no financiero," ESPOL, Guayaquil, 2017.

[24] G. Fernández, "El delito de blanqueo de capitales," *La enmacipación,* pp. 65-87, 2014.

[25] GAFI, "International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation," GAFI, París, 2020.

[26] C. Rivera, "Modelo de control de las medidas de prevención del lavado de activos y financiamiento de delitos para las instituciones del sistema financiero del Ecuado," Universidad Central del Ecuador, Quito, 2016.

[27] GAFILAT, "La construcción de una organización y el desarrollo de un sistema regional," GAFILAT, Buenos Aires, 2020.

[28] EGMONT, "El Grupo Egmont de Unidades de Inteligencia Financiera," EGMONT, Bruselas, 2016.

[29] GAFI, "Opportunities and challenges of new technologies for AML/CFT," GAFI, París, 2021.

[30] A. Canhoto, "Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective," *Journal of Business Research,* pp. 441-452, 2021.

[31] J. Lorenz, M. Silva, D. Aparicio, J. Ascensão and P. Bizarro, "Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity," *arXiv.org,* pp. 1-8, 2020.

[32] J. Domashova and N. Mikhailina, "Usage of machine learning methods for early detection of money laundering schemes," *Procedia Computer Science,* pp. 184-192, 2021.

[33] J. Rocha, M. Segovía and M. Camacho, "Money laundering and terrorism financing detection using neural networks and an abnormality indicator," *Expert Systems with Applications,* pp. 1-15, 2021.

[34] S. Kannan and M. Srinath, "Autoregressive based Outlier Algorithm to Detect Money Laundering Activities," *International Journal of Research and Analytical Reviews,* pp. 29-38, 2018.

[35] A. Nhien, M. Sammer and K. M-Tahar, "A data mining-based solution for detecting suspicious money laundering cases in an investment bank," *Segunda Conferencia Internacional sobre Avances en Bases de Datos, Conocimiento y Aplicaciones de Datos,* pp. 235-240, 2010.

[36] V. Jayasree and S. Balan, "Money Laundering Identification on Banking Data Using Probabilistic Relational Audit Sequential Pattern," *Science Alert,* pp. 173-184, 2014.

[37] L. Pérez, "Metodología para segmentación de un SARLAFT," Repositorio Digital Universidad El Bosque, Bogotá, 2020.

[38] G. F. S. Salvatore T. March, «Design and natural science research on information technology,» *Decision support systems,* vol. 15, pp. 251-266, 1995.

[39] P. J. R. LEONARD KAUFMAN, Finding Groups in Data An Introduction to Cluster Analysis, New Jersey: JOHN WILEY & SONS, INC., 2009.

[40] T. T. G. R. Ken Peffers, «The design science research process: a model for producing and presenting information systems research,» de *First International Conference on Design Science Research in Information Systems and Technology*, USA, 2006.

[41] J. C. Dunn, «Well-separated clusters and optimal fuzzy partitions,» *Journal of cybernetics,* pp. 95-104, 1974.

[42] P. J. ROUSSEEUW, «Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,» *Journal of Computational and Applied Mathematics ,* pp. 53-65, 1986.

6. ANEXOS

Anexo 1: Análisis descriptivo de datos

El siguiente Anexo muestra un análisis descriptivo de la información de los factores de riego usados dentro de toda la construcción de los modelos de segmentación.

Tabla #1: "JURISDICTION_COUNTRY"

VARIABLES	NULOS
Jurisd_ID	0
Jurisdiction	0
Concentration	0
Nature_TRX	0
Gafi_Fatf_Public_Statement	0
Gafi_On_Going_Process_Statement	0
Basel_AML_Index	0
Segment_Country	0

De la tabla anterior podemos observar las variables que se encuentran en la tabla "JURISDICTION_COUNTRY", en donde se destaca el hecho de que no se encontraron valores nulos en ninguna de las variables lo que nos indica una completitud en la base de datos, estas variables se obtienen del resultado de **194** Jurisdicciones a nivel de países incluida la de **COLOMBIA.**

	Cuenta de NATURE_TRX
0	194
3	1
Total general	**195**

En la tabla anterior se representa la naturaleza de las transacciones lo cual nos indicará que operaciones son permitidas para ser realizadas de manera presencial de acuerdo con los convenios establecidos con las entidades financieras, se logra observar que 194 de ellas no pertenecen a retiros ni a aportes lo que nos quiere indicar que no se les permite realizar transacciones de manera presencial, es importante realizar una actualización de los datos para verificar la calidad de los mismos, mientras que la jurisdicción de Colombia tiene una naturaleza de Aportes-Retiros.

	Cuenta de lista GAFI High Risk
0	193
1	2
Total general	**195**

En la tabla anterior, se logra evidenciar que de la lista GAFI High Risk , la cual nos ayuda para identificar las jurisdicciones con alto riesgo, se encuentran 2 jurisdicciones las cuales son Corea del Norte e Iran. Estas se identifican por no cooperar en gran medida con la lucha en contra del lavado de activos y financiamiento del terrorismo.

	Cuenta de lista GAFI Jurisdictions Under Increased Monitoring
0	179
1	16
Total general	**195**

De la lista GAFI Jurisdictions Under Increased Monitoring se pueden observar las jurisdicciones que tienen mayor control y a su vez representan un riesgo considerable, son en total de 16 jurisdicciones, que se muestran a continuación:

1155	ALBANIA
1165	BAHAMAS
1167	BARBADOS
1217	JAMAICA
1237	NICARAGUA
1242	PAKISTAN
1243	PANAMA
1255	SIRIA
1271	ZIMBABWE
1894	BOTSWANA
1900	CAMBOYA
1911	GHANA
1925	MAURICIO
1928	MYANMAR
1958	UGANDA
1961	YEMEN

Tabla #2: "JURISDICTION_DEPARTMENT"

Variables	Nulos
Jurisd_ID	0
Jurisdiction	0
DIAN_Code	0
Concentration	0
Nature_TRX	0
Risk_LAFT	0
Physical_Branch	0
Segment_Department	0

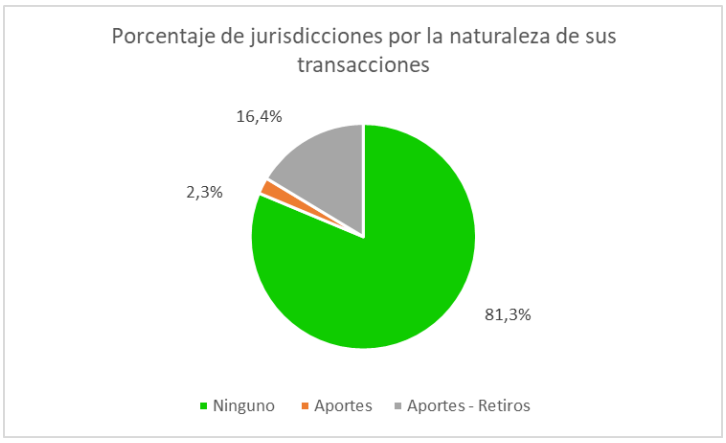
Jurisd_Country	0

En la tabla anterior se exponen las variables de la tabla "JURISDICTION_DEPARTMENT" en donde se logra identificar que no existen valores nulos dentro de estas variables lo que nos indica una completitud en la base de datos.

	Departamentos en Colombia
TRUE	33
FALSE	138
Total general	**171**

En la tabla anterior se observan las jurisdicciones por departamentos en donde se encontraron 33 jurisdicciones que pertenecen a la jurisdicción de Colombia y el restante pertenecen a otro país para un total de 171 jurisdicciones.

	Cuenta de Nature_TRX
0	139
2	4
3	28
Total general	**171**



En la tabla anterior se observa la naturaleza de las transacciones las cuales se distribuyeron de manera porcentual en el gráfico. Se observó que el 81.3% de las jurisdicciones tienen una naturaleza que no pertenece ni a Aportes ni retiros, lo que nos quiere indicar que no se les permite realizar transacciones de manera presencial, es importante realizar una actualización de los datos para verificar la calidad de estos, el 16.4% pertenece a Aportes-Retiros y el 2.3% a Retiros.

	Cuenta de Risk_LAFT
*Alto	17
FALSE	5
TRUE	12
*Bajo	132
FALSE	124
TRUE	8
*Medio	22
FALSE	9
TRUE	13
Total general	171

Durante el análisis del riesgo de las jurisdicciones por departamentos con un riesgo alto se observa que 12 de ellas pertenecen a Colombia y 5 a jurisdicciones extranjeras, cuando se observan las jurisdicciones con riesgo medio se obtiene que 13 pertenecen a Colombia y 9 a extranjeras y como ultima observación, se obtiene que para Colombia existen 8 jurisdicciones por departamento y 124 son extranjeras.

Tabla #3: "PRODUCT_PLAN_DETAIL"

Variables	Nulos
Product	0
Product_Plan	0
Restriction_Multigestion	0

Restriction_Withdrawal	0
Market_Niche	0
Management_Client	0
Minimum_Income	0
Sponsor	0
Product_Type_ID	6
Investment_Term_ID	6
Segment_Product	0
Company	0
Risk_LAFT	0

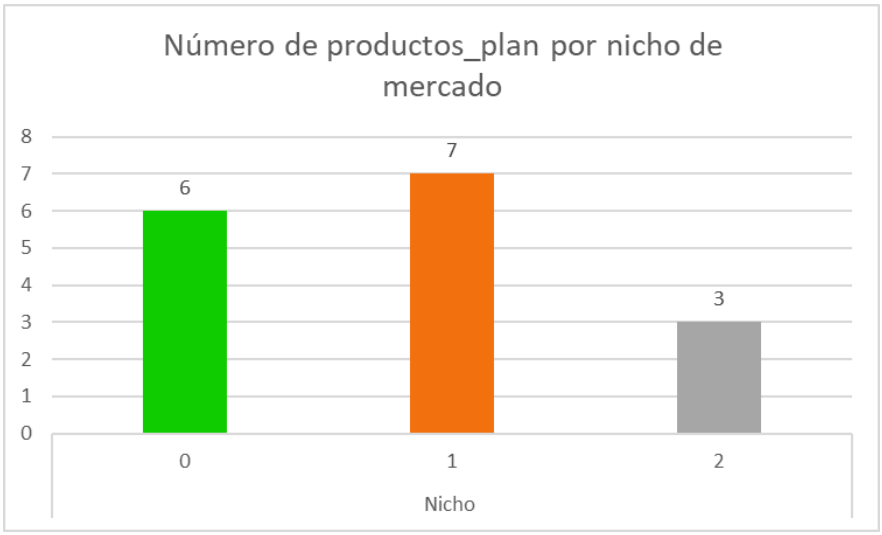
Dentro del análisis de la tabla "**PRODUCT_PLAN_DETAIL"** se encontró 6 registros nulos en la variable "Product_Type_ID" y 6 nulos en "Investment_Term_ID", esto nos quiere decir que estas variables no tienen datos confiables o de baja calidad por lo cual es necesario hacer una actualización de los datos y establecer unas reglas de calidad para evaluar cuantos individuos pueden ser estudiados en el análisis.

prod_plan	
CREA_1	En la tabla anterior podemos observar los 16 productos dentro de los planes observados los cuales son los únicos con los que
FIBAC_1	
FIBAC_2	
FMAGNO_1	
FONVIDA_1	
IGOLD_1	
OMPEV_PV01	
OMSVI_1	
SEGCO_SC01	
SEGCO_SC02	
SEGCO_SC03	
SIPEN_FB01	
SIPEN_FV01	

SIPEN_IG01	trabaja la
SIPEN_TL01	empresa
TLIFE_1	VIDA

Product	Product_Plan	Restriction_Multigestion
OMSVI	1	TRUE

Cuando se realiza el análisis de la variable Restriction_Multigestion, la cual nos indica si el cliente puede o no realizar traslado del dinero en los diferentes portafolios de inversión ofertados para cada producto, o en su defecto identificar si es necesario contar con aprobación por parte de su empleador o patrocinador para realizar una inversión, entendido esto, se logró observar que el producto OMSVI cuenta con el plan 1 el cual es el único que tienen este tipo de restricción, el resto de los productos no cuentan con esta marca.

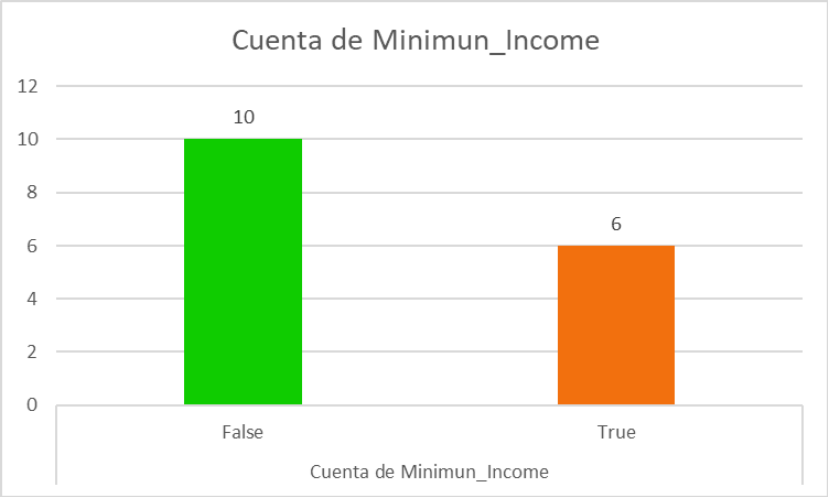


En la gráfica anterior se puede observar el nicho de mercado, el cual nos indicará a qué tipo de persona va dirigido el producto, con el fin de cumplir con el objetivo de este. En la actualidad en SKANDIA no se tiene un nicho de mercado específico. Dicho lo anterior se puede observar que hay 16 productos_plan que pertenecen a VIDA, se encontró que 7 de ellos realizan transacciones por el nicho 1 el cual va dirigido a los clientes catalogados como Persona natural, 3 por el nicho 2 el cual está catalogado como Persona jurídica, sin embargo, se logra apreciar que hay 6

productos_plan que se encuentran inactivos, por lo cual es necesario realizar una actualización de la calidad de los datos para poder realizar el estudio de los mismo. En conclusión, obtenemos que solo se encuentran activos un total de 10 productos.

	Cuenta de Management_Client
0	6
1	10
Total general	**16**

En la tabla anterior se quiere demostrar si el cliente tiene de manera directa toma de decisiones sobre el capital invertido en el producto o interviene un tercero en esta toma de decisiones; es decir, realizar la inversión en portafolios ofertados por la compañía o disponer de los recursos para retiros extemporáneos o totales de acuerdo con las reglas de negocio de dicho producto. Se logra observar que 10 productos_plan están orientados a una gestión activa la cual nos Indica que es el cliente quien podrá tomar decisiones sobre su inversión., mientras que las 6 observaciones restantes de productos-Plan no pertenecen a ninguna gestión del cliente ya que esto está relacionado con la información obtenida del Nicho de Mercado, por lo tanto, es necesario realizar una actualización de la calidad de los datos para poder realizar el estudio de los mismo.



En la gráfica anterior se observa a variable Ingreso Mínimo, la cual tiene como objetivo, poder identificar que productos requieren de un capital mínimo para acceder a ellos y dar cumplimiento a las condiciones de estos. De acuerdo con la información anterior, se identificó que 10 de los productos_plan no requieren un ingreso mínimo, mientras que 6 si lo necesitan, la información obtenida está relacionada con la tabla anterior Management_Client, por lo tanto, es necesario realizar una actualización de la calidad de los datos para poder realizar el estudio de los mismo.

	Cuenta de Risk_LAFT
Bajo	6
Medio	10
Total general	**16**

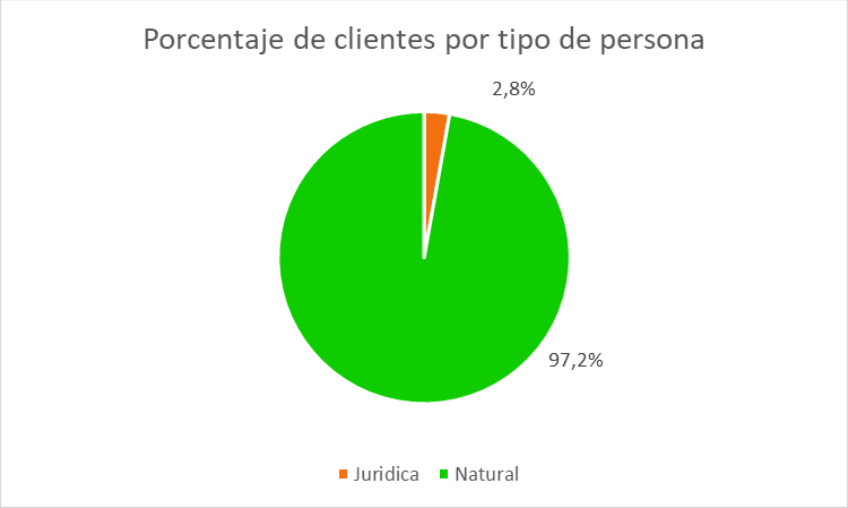
En la tabla anterior se puede observar que los 16 productos_plan con los que trabaja SKANDIA Seguros de Vida 10 de ellos tienen un riesgo LAFT medio y 6 un riesgo bajo, posiblemente los 6 datos que tienen un riesgo bajo son los mismos que se encuentran inactivos por lo tanto hay que realizar una actualización de la calidad de los datos para poder realizar el estudio de los mismo.

Tabla #4: “Clientes”

Variables	Nulos
Document_Type	0
Document_Number	0
Names	0
Surnames	0
Income	0
Expenses	0
Assets	0
Passives	0
Heritage	0
Economic_Activity	0
Economic_Sector	0

CIIU	0
CIIU_Description	0
CIIU_HighRisk	0
Birth_Date	0
Create_Date	0
Modify_Date	0
Jurisd_ID	3.261
Is_Public	0
Public_Recognition	0
Risk	0
Company_Type	0
Segment_Client	0
Cutoff_Date	0
Brand_Employee	1

En la tabla anterior podemos observar la tabla de Clientes en donde se pudo encontrar que existen 3271 valores nulos en la variable Jurisd_ID y 1 valor nulo en la variable Brand_Employee, el resto de las variables no tienen casos con nulos. Por lo tanto, en necesario tener precaución en no tomar en cuenta eso valores nulos a la hora de realizar el estudio, al ser una cantidad tan pequeña, no va a representar una gran afectación si estos valores se omiten.



Dentro de la misma tabla de clientes se obtiene que 30.434 clientes que pertenecen a la empresa SKANDIA Seguros de Vida de los cuales existen 844 son Personas

Jurídicas las cuales representa el 2.8% del total y 29.590 personas Naturales que representa el 97.2% del total de clientes, lo que quiere indicar que la mayor cantidad de clientes de la empresa son representadas como personas naturales, por lo cual las personas jurídicas no tienen una representación considerable.

Medidas	Personas Naturales				
	Income	Expenses	Assets	Passives	Heritage
Media	$ 9,06	$ 4,55	$ 363,94	$ 68,24	$ 295,71
Suma	$ 267.950	$ 134.719	$ 10.769.090	$ 2.019.139	$ 8.749.950
Desviación	$ 160,46	$ 93,91	$ 4.343,93	$ 190,38	$ 4.325,54
Mínimo	$ -	$ -	$ -	$ -	-$ 2.564,56
Máximo	$ 24.000,00	$ 16.000,00	$ 700.000,00	$ 6.500,00	$ 699.600,00

En la tabla anterior se puede evidenciar las estadísticas descriptivas de Ingresos, Egresos, Activos, Pasivos y Patrimonio expresado en millones de pesos para las personas Naturales.

Se analizo los ingresos podemos observar que se tiene una media de $9,06 millones la cual se encuentra bastante elevada, esto se debe al valor obtenido en el máximo, seguido a esto, tenemos una suma total de $267.950,46 millones con una desviación de $160,46 y un máximo de $24.000,00 millones, este valor debe revisarse ya que es demasiado alto, por lo tanto, sería ideal verificar la calidad de los datos, también se cuenta con un mínimo de valor $0 lo cual no es un valor lógico para el modelo de negocio, lo que significa que se cuenta con datos de baja calidad por lo tanto se requiere una actualización, los datos que no cumplan con la calidad no serán utilizados en la segmentación y se realizara un análisis más detallado para encontrar el valor mínimo de ingresos que cumpla con la lógica de negocios. Por este motivo se establecerán unas reglas de calidad para evaluar cuantos individuos pueden ser estudiados en el análisis del ingreso.

Como segunda observación tenemos los egresos, los cuales se distribuyen con una media de $4,55 millones la cual se encuentra bastante elevada, esto se debe al

valor obtenido en el máximo, seguido a esto, tenemos una suma total de $134.719,46 millones con una desviación de $93,91 y un máximo de $16.000,00 millones este valor debe revisarse ya que es demasiado alto, por lo tanto, sería ideal verificar la calidad de los datos, también se cuenta con un mínimo de valor $0 lo cual no es un valor lógico para el modelo de negocio, lo que significa que se cuenta con datos de baja calidad por lo tanto se requiere una actualización, los datos que no cumplan con la calidad no serán utilizados en la segmentación y se realizara un análisis más detallado para encontrar el valor mínimo de egresos que cumpla con la lógica de negocios. Por este motivo se establecerán unas reglas de calidad para evaluar cuantos individuos pueden ser estudiados en el análisis de los egresos.

En la tercera columna analizamos los activos, los cuales tienen una media de $363,94 millones la cual se encuentra bastante elevada, esto se debe al valor obtenido en el máximo, seguido a esto, una suma total de $10.769.090,27 millones con una desviación de $4.343,93 y un máximo de $700.000,00 millones, este valor debe revisarse ya que es demasiado alto, por lo tanto, sería ideal verificar la calidad de los datos, también se cuenta con un mínimo de valor $0 lo cual no es un valor lógico para el modelo de negocio, lo que significa que se cuenta con datos de baja calidad por lo tanto se requiere una actualización, los datos que no cumplan con la calidad no serán utilizados en la segmentación y se realizara un análisis más detallado para encontrar el valor mínimo de activos que cumpla con la lógica de negocios. Por este motivo se establecerán unas reglas de calidad para evaluar cuantos individuos pueden ser estudiados en el análisis de los activos.

Siguiendo en el análisis llegamos al estudio de los pasivos los cuales tiene con una media de $68.24 millones la cual se encuentra bastante elevada, esto se debe al valor obtenido en el máximo, seguido a esto, una suma total de $2.019.139,33 millones con una desviación de $190,38 y un máximo de $6.500,00 millones este valor debe revisarse ya que es demasiado alto, por lo tanto, sería ideal verificar la calidad de los datos, también se cuenta con un mínimo de valor $0 esto se encuentra dentro del modelo como una posibilidad lógica, ya que indica que hay clientes que no tienen deudas.

Para finalizar obtenemos el patrimonio el cual se distribuye de la siguiente manera, una media de $295,71 millones la cual se encuentra bastante elevada, esto se debe al valor obtenido en el máximo, seguido a esto, una suma total de $8.749.950,95 millones con una desviación de $4.325,54, un mínimo de -$2.564,56 millones, este mínimo nos indica que se pueden encontrar clientes que tienen una gran cantidad de deudas y un máximo de $699.600,00 millones. millones, este valor debe revisarse ya que es demasiado alto, por lo tanto, sería ideal verificar la calidad de los datos.

Medidas	Personas Jurídicas				
	Income	Expenses	Assets	Passives	Heritage
Media	$ 845,03	$ 1.880,00	$ 7.739,10	$ 5.299,70	$ 2.439,41
Suma	$ 713.203	$ 1.586.722	$ 6.531.804	$ 4.472.946	$ 2.058.858
Desviación	$ 5.554,13	$ 34.310,64	$ 84.846,01	$ 68.749,22	$ 19.131,48
Mínimo	$ -	$ -	$ -	$ -	-$ 140.965
Máximo	$ 138.087	$ 985.135	$ 2.228.982	$ 1.832.136	$ 396.845

En la tabla anterior se puede evidenciar las estadísticas descriptivas de Ingresos, Egresos, Activos, Pasivos y Patrimonio expresado en millones de pesos para las personas jurídicas.

Se analizo los ingresos podemos observar que se tiene una media de $845,03 millones la cual se encuentra bastante elevada, esto se debe al valor obtenido en el máximo, seguido a esto, tenemos una suma total de $713203,28 millones con una desviación de $5.554,13 y un máximo de $138.087,42 millones este valor debe revisarse ya que es demasiado alto, por lo tanto, sería ideal verificar la calidad de los datos, también se cuenta con un mínimo de valor $0 lo cual no es un valor lógico para el modelo de negocio, lo que significa que se cuenta con datos de baja calidad por lo tanto se requiere una actualización, los datos que no cumplan con la calidad no serán utilizados en la segmentación y se realizara un análisis más detallado para encontrar el valor mínimo de ingresos que cumpla con la lógica de negocios. Por

este motivo se establecerán unas reglas de calidad para evaluar cuantos individuos pueden ser estudiados en el análisis del ingreso.

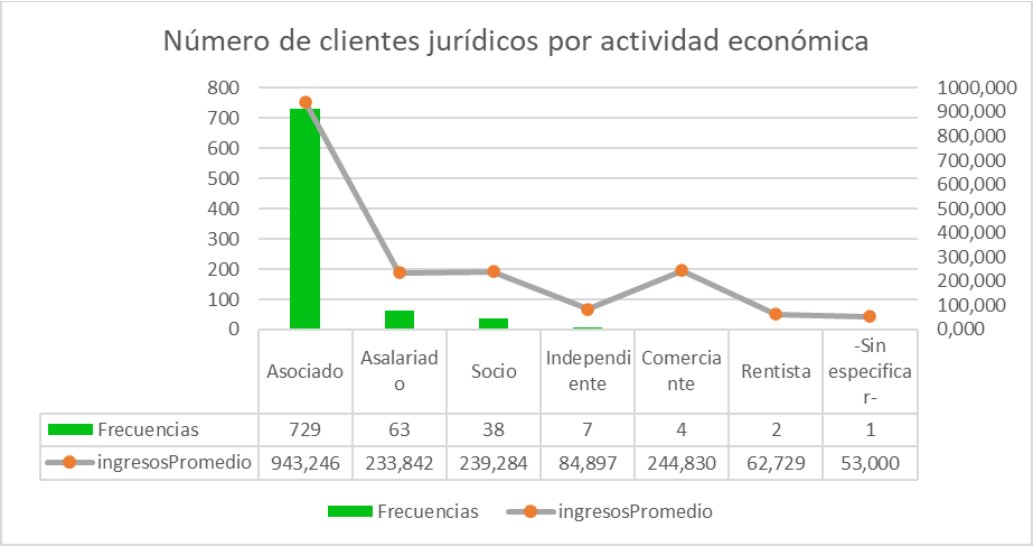
Como segunda observación tenemos los egresos, los cuales se distribuyen con una media de $1.880,00 millones la cual se encuentra bastante elevada, esto se debe al valor obtenido en el máximo, seguido a esto, tenemos una suma total de $1.586.722,67 millones con una desviación de $34.310,64 y un máximo de $985.135,42 millones, este valor debe revisarse ya que es demasiado alto, por lo tanto, sería ideal verificar la calidad de los datos, también se cuenta con un mínimo de valor $0 lo cual no es un valor lógico para el modelo de negocio, lo que significa que se cuenta con datos de baja calidad por lo tanto se requiere una actualización, los datos que no cumplan con la calidad no serán utilizados en la segmentación y se realizara un análisis más detallado para encontrar el valor mínimo de egresos que cumpla con la lógica de negocios. Por este motivo se establecerán unas reglas de calidad para evaluar cuantos individuos pueden ser estudiados en el análisis de los egresos.

En la tercera columna analizamos los activos, los cuales tienen una media de $7.739,10 millones la cual se encuentra bastante elevada, esto se debe al valor obtenido en el máximo, seguido a esto, tenemos una suma total de $6.531.804,62 millones con una desviación de $84.846,01 y un máximo de $2.228.982,16 millones, este valor debe revisarse ya que es demasiado alto, por lo tanto, sería ideal verificar la calidad de los datos, también se cuenta con un mínimo de valor $0 lo cual no es un valor lógico para el modelo de negocio, lo que significa que se cuenta con datos de baja calidad por lo tanto se requiere una actualización, los datos que no cumplan con la calidad no serán utilizados en la segmentación y se realizara un análisis más detallado para encontrar el valor mínimo de activos que cumpla con la lógica de negocios. Por este motivo se establecerán unas reglas de calidad para evaluar cuantos individuos pueden ser estudiados en el análisis de los activos.

Siguiendo en el análisis llegamos al estudio de los pasivos los cuales tiene con una media de $5.299,90 millones la cual se encuentra bastante elevada, esto se debe

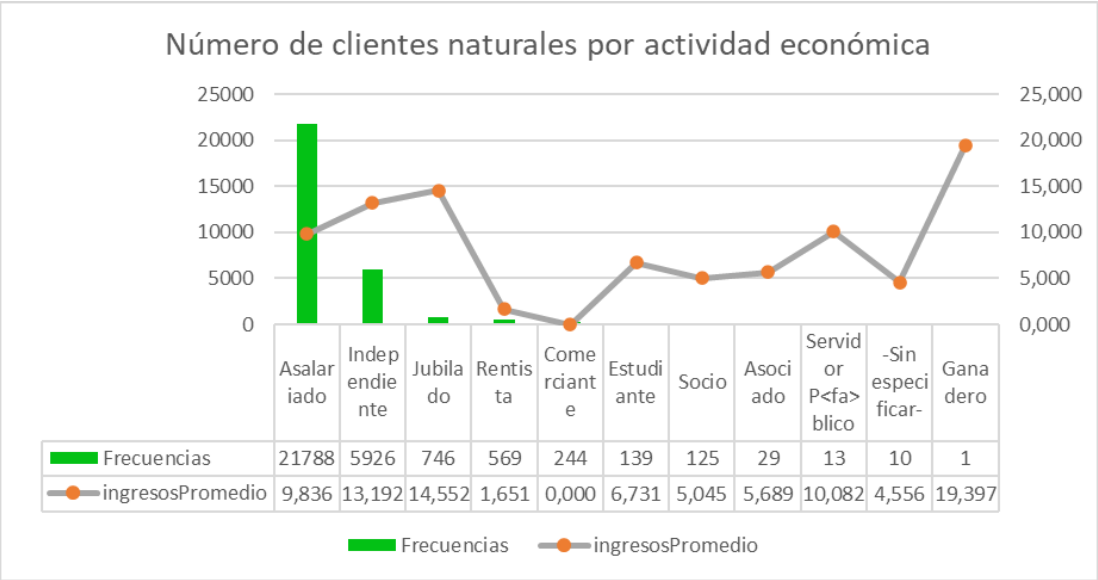
al valor obtenido en el máximo, seguido a esto, tenemos una suma total de $4.472.946,54 millones con una desviación de $68.749,22 y un máximo de $1.832.136,80 millones, este valor debe revisarse ya que es demasiado alto, por lo tanto, sería ideal verificar la calidad de los datos, también se cuenta con un mínimo de valor $0.

Para finalizar obtenemos el patrimonio el cual se distribuye de la siguiente manera, una media de $2.439,41 millones la cual se encuentra bastante elevada, esto se debe al valor obtenido en el máximo, seguido a esto, tenemos una suma total de $2.058.858,08 millones con una desviación de $19.131,48, un mínimo de -$140.965,73 millones, este mínimo nos indica que se pueden encontrar clientes que tienen una gran cantidad de deudas y un máximo de $396.845,36 millones, este valor debe revisarse ya que es demasiado alto, por lo tanto, sería ideal verificar la calidad de los datos.



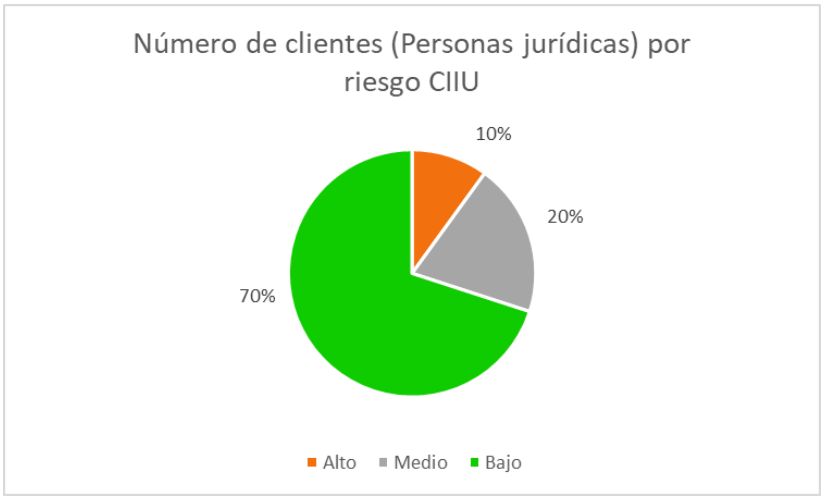
En la gráfica se puede observar el número de Clientes jurídicos por actividad económica, adicional se puede visualizar el promedio de ingresos de esas personas, donde se nota mayor diferencia es entre la actividad de asalariado y socio en donde los socios a pesar de tener una diferencia de casi del 50% de clientes el promedio de ingresos de la actividad de socio es superior con $200,000 millones de pesos aproximadamente.

Por otro lado, se puede evidenciar una gran diferencia de la actividad económica de asociado contra el resto de las actividades, esta es la que lleva la mayor proporción de los ingresos de todas las actividades.

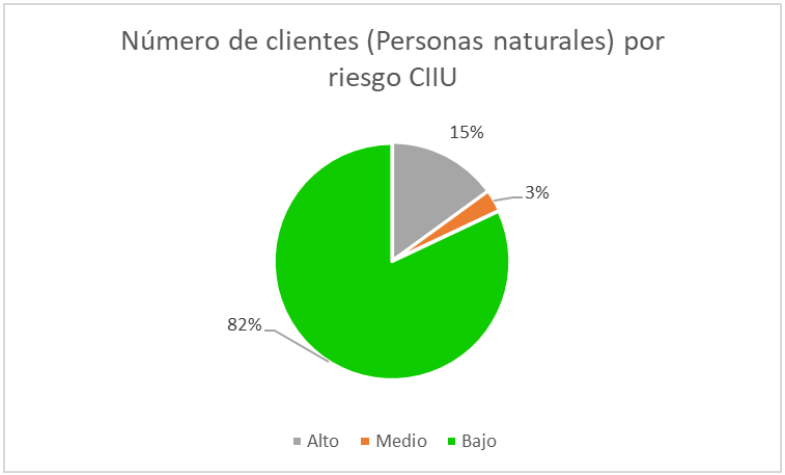


En el grafico anterior de clientes se puede observar que la mayor concentración de los clientes pertenece a la actividad económica de asalariado, en un segundo lugar tenemos al independiente y jubilado, mientras que la actividad económica de ganadero es la actividad con menor número de clientes que sean personas naturales.

También se logra visualizar una gran diferencia en la actividad de Servidor público en donde a pesar de ser una actividad que tiene únicamente 13 clientes sus ingresos promedios son casi iguales al de los asalariados que tiene mayor número de clientes, por otra parte, es necesario tener en cuenta los Independientes y el cliente que está representado como Ganadero ya que los valores de ingresos son bastante altos, lo cual puede representar una irregularidad y por consiguiente un riesgo.



En grafico anterior se puede observar la distribución de los porcentajes en cuestión de riego CIIU, encontramos que el 70% de los clientes de la empresa VIDA que son personas jurídicas tienen un riesgo bajo según su clasificación en con el código CIIU, mientras que el 20% tienen un riesgo medio y el 10% un riesgo alto. Lo que nos quiere decir que los clientes como personas jurídicas no representan un riesgo considerable.



En distribución de porcentajes anterior se puede observar que el 82% de los clientes de la empresa VIDA que son personas naturales tienen un riesgo bajo según su clasificación en con el código CIIU, mientras que el 3% tienen un riesgo medio y el 15% un riesgo alto. Lo que nos quiere decir que los clientes como personas naturales no representan un riesgo considerable.

	Cuenta de Economic_Activity
Alto	83
Bajo	591
Medio	170
Total	**844**

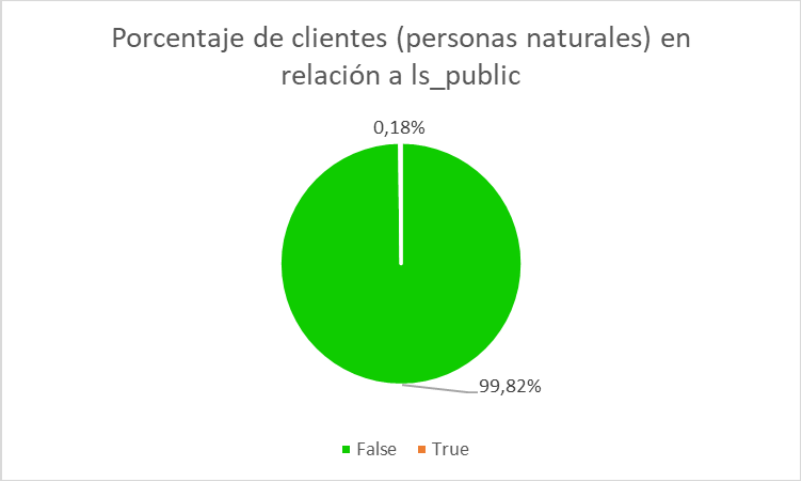
De la tabla anterior podemos observar que de 844 clientes que son personas jurídicas 83 tienen un alto riesgo CIIU, seguido de 591 cliente con un riesgo bajo y 170 cliente un riesgo medio, haciendo notar la mayor proporción con un bajo riesgo, pero eso no significa que se deba dejar a un lado el estudio y análisis de los clientes que se encuentran en riesgo alto.

Natural	
	Cuenta de nacionalExtranjera
EXTRANJERA	3.283
NACIONAL	26.307
Total general	**29.590**
Jurídica	
	Cuenta de nacionalExtranjera
EXTRANJERA	15
NACIONAL	829
Total general	**844**

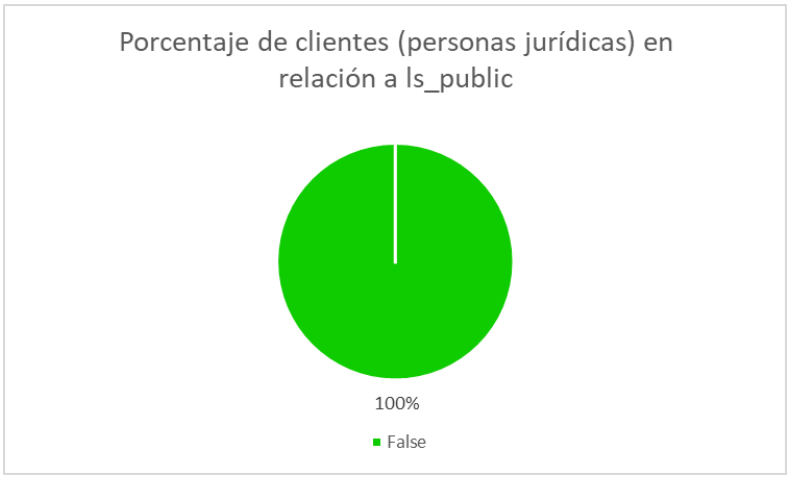
De la tabla anterior se relaciona a los clientes con las jurisdicciones departamentales y como bien se observa, las personas naturales 26307 pertenecen a jurisdicciones departamentales nacionales y 3283 clientes a jurisdicciones departamentales extrajeras, por otro lado, los clientes que son personas jurídicas se identificaron que 829 pertenecen a jurisdicciones departamentales nacionales y 15 extrajeras.

En conclusión y tomando en cuenta la información obtenida de la tabla anterior, se puede observar que la mayor cantidad de clientes de la empresa VIDA se

encuentran en las jurisdicciones nacionales, esto independiente de si son personas naturales o personas jurídicas.



En el grafico anterior se realiza el análisis de la variable ls_Public la cual nos ayuda para identificar si el cliente tiene relación o es reconocido públicamente, allí se identificó que el 99,82% de los clientes que son personas naturales que no son figuras públicas, mientras que solo el 0.18% de esas personas son figuras públicas, lo que demuestra una cantidad muy baja de clientes que son figuras públicas.



Para el análisis de la variable ls_Public la cual nos ayuda para identificar si el cliente tiene relación o es reconocido públicamente, allí se identificó que el 100% de los clientes que son personas jurídicas no son figuras públicas, lo que quiere decir que,

en la base de clientes estudiados, específicamente las personas jurídicas no se cuentan con individuos conocidos como figuras públicas.

	Cuenta de Risk
JURIDICA	**843**
Alto	4
Bajo	404
Medio	435
NATURAL	**29.566**
Alto	208
Bajo	23.668
Medio	5.690
Total general	**30.409**

Para el análisis de riesgo de clientes se tiene un total de 30.409 clientes que cuentan con calificación de riesgo, 4 clientes que son personas jurídicas tienen un riesgo alto mientras que para personas naturales únicamente 208 clientes tienen un riesgo alto, el mayor número de clientes tanto en personas jurídicas como naturales están en riesgos medio y bajo por lo cual se puede centralizar de una manera más eficaz, el análisis aquellos clientes que se encuentran en una clasificación de riesgo alto.

Tabla #5: "Transacciones"

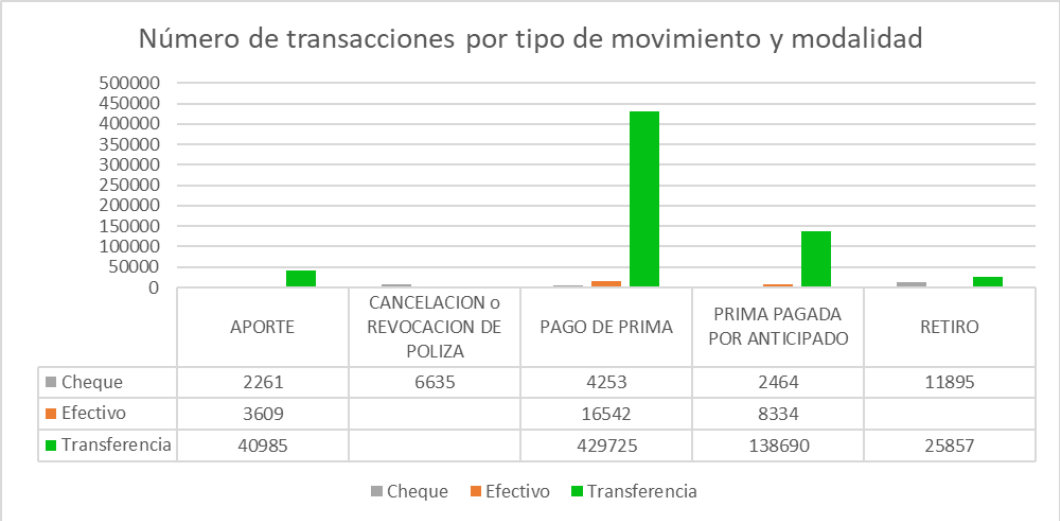
Variables	Nulos
Event_Number	0
Transaction_Number	0
Contract_ID	0
Product	0
Product_Plan	0
Movement_Code	0

Movement	0
Movement_Date	0
Movement_Value	0
Modality	0
Movement_Channel	0
Movement_Jurisdiction	22.414
Document_Type_Beneficiary	0
Document_Number_Beneficiary	0
Cutoff_Date	0
Transaction_Type	0
Company	0
Risk_LAFT	0

Dentro del análisis de los datos en cuanto a transacciones se observó que existen 22417 casos nulos en la variable Movement_Jurisdicion, mientras que resto de variables una completitud en sus datos. Las transacciones que tienen valores nulos no se tendrán en cuenta a la hora de estudiar los distintos factores de riesgo.

product_plan2	APORTE	CANCELACION o REVOCACION DE POLIZA	PAGO DE PRIMA	PRIMA PAGADA POR ANTICIPADO	RETIRO	Total general
CREA_1	-	6.633	269.987	54.119		330.739
OMPEV_PV01	-	2	170.734	95.369	2.469	268.574
OMSVI_1	-	-	9.799	-	-	9.799
SEGCO_SC01	-	-	-	-	49	49
SEGCO_SC02	1	-	-	-	10	11
SEGCO_SC03	353	-	-	-	2.134	2.487
SIPEN_FB01	5	-	-	-	33	38
SIPEN_IG01	-	-	-	-	3	3
SIPEN_TL01	46.496	-	-	-	33.054	79.550
Total general	**46.855**	**6.635**	**450.520**	**149.488**	**37.752**	**691.250**

En la tabla anterior se puede observar el estudio de las transacciones por tipo de movimiento en donde se identificó que el PAGO DE PRIMA es la que tiene mayor número de transacciones con 450.520 exactamente, seguido de PRIMA PAGADA POR ANTICIPADO con 149488 y como ultimo CANCELACION o REVOCACION DE POLIZA con 6.635 transacciones, el APORTE y el RETIRO tienen la menor cantidad de transaccione.



	APORTE	CANCELACION o REVOCACION DE POLIZA	PAGO DE PRIMA	PRIMA PAGADA POR ANTICIPADO	RETIRO
Cheque	2261	6635	4253	2464	11895
Efectivo	3609		16542	8334	
Transferencia	40985		429725	138690	25857

En la tabla anterior se logra observar que la modalidad con mayor número de movimientos es Transferencia con 635.257 movimientos de los cuales la mayor cantidad de los movimientos se encuentran reflejados en el PAGO DE PRIMA, seguido de efectivo con 28.485 en donde el PAGO DE PRIMA también obtiene la mayor relevancia de esta modalidad de movimiento y como ultimo cheque 27.508 movimientos, donde los RETIROS tienen relevancia.

La modalidad transferencia con el tipo de movimiento PAGO DE PRIMA es la que tiene mayor número de transacciones con 429.725, seguido de PRIMA PAGADA POR ANTICIPADO con la misma modalidad transferencia con 138.690 transacciones. Por lo cual el pago por transferencia es la modalidad de la cual se puede obtener más información para ser estudiada.

	Promedio	Suma	Desvest	Mín.	Máx.
APORTE	$ 5,21	$ 244.304,67	$ 40,44	$ -	$ 2.620,00

CANCELACION o REVOCACION DE POLIZA	-$ 16,88	-$ 112.014,81	$ 22,15	-$ 242,09	$ -
PAGO DE PRIMA	$ 0,59	$ 266.274,46	$ 0,55	$ -	$ 29,91
PRIMA PAGADA POR ANTICIPADO	$ 0,73	$ 108.668,66	$ 1,93	$ -	$ 197,12
RETIRO	-$ 10,51	-$ 396.599,13	$ 182,06	-$ 20.000,00	$ 350,00
Total general	**$ 0,16**	**$ 110.633,85**	**$ 44,02**	**-$ 20.000,00**	**$ 2.620,00**

En la tabla anterior se puede evidenciar las estadísticas descriptivas del Aporte, Cancelación o Renovación de póliza, Pago de prima, Prima pagada por anticipado y Retiro en millones de pesos.

Se analizo los Aporte podemos observar que se tiene un promedio de $5,21 millones, una suma total de $244.304,67 millones con una desviación de $40,44 y un máximo de $2.620,00 millones, se cuenta con un mínimo de valor $0 lo cual no es un valor lógico para el modelo de negocio, lo que significa que se cuenta con datos de baja calidad por lo tanto se requiere una actualización, los datos que no cumplan con la calidad no serán utilizados en la segmentación y se realizara un análisis más detallado para encontrar el valor mínimo de Aporte que cumpla con la lógica de negocios. Por este motivo se establecerán unas reglas de calidad para evaluar cuantos individuos pueden ser estudiados en el análisis del Aporte.

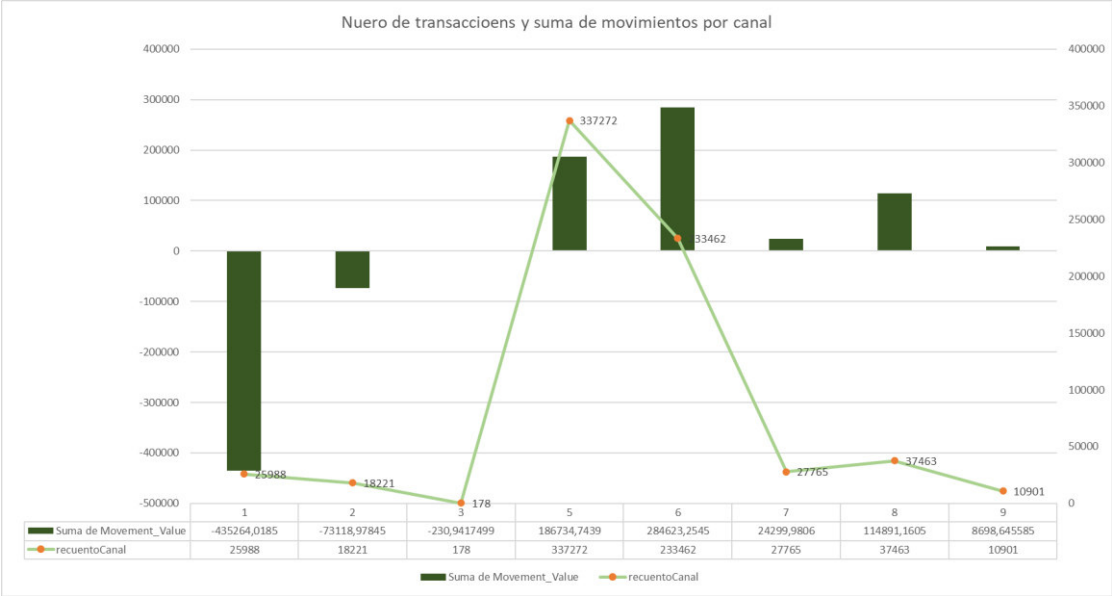
Como segunda observación tenemos la Cancelación o Renovación de póliza, la cual se distribuye con un promedio de -$16,88 millones, una suma total de -$112.014,81 millones con una desviación de $22,15 y un mínimo de -$242,06 millones, se cuenta con un máximo de valor $0 lo cual no es un valor lógico para el modelo de negocio, lo que significa que se cuenta con datos de baja calidad por lo tanto se requiere una actualización, los datos que no cumplan con la calidad no serán utilizados en la segmentación y se realizara un análisis más detallado para encontrar el valor mínimo de la Cancelación o Renovación de póliza que cumpla con la lógica de negocios. Por este motivo se establecerán unas reglas de calidad

para evaluar cuantos individuos pueden ser estudiados en el análisis de la Cancelación o Renovación de póliza.

En la tercera fila analizamos el Pago de prima, los cuales tienen una promedio de $0,59 millones, una suma total de $266.274,46 millones con una desviación de $0,55 y un máximo de $29,91 millones, se cuenta con un mínimo de valor $0 lo cual no es un valor lógico para el modelo de negocio, lo que significa que se cuenta con datos de baja calidad por lo tanto se requiere una actualización, los datos que no cumplan con la calidad no serán utilizados en la segmentación y se realizara un análisis más detallado para encontrar el valor mínimo de Pago de prima que cumpla con la lógica de negocios. Por este motivo se establecerán unas reglas de calidad para evaluar cuantos individuos pueden ser estudiados en el análisis de la Pago de prima.

Siguiendo en el análisis llegamos al estudio de la Prima pagada por anticipado la cual tiene un promedio de $0,73 millones, una suma total de $108.668,66 millones con una desviación de $1,93 y un máximo de $197,12 millones, se cuenta con un mínimo de valor $0 lo cual no es un valor lógico para el modelo de negocio, lo que significa que se cuenta con datos de baja calidad por lo tanto se requiere una actualización, los datos que no cumplan con la calidad no serán utilizados en la segmentación y se realizara un análisis más detallado para encontrar el valor mínimo de Prima pagada por anticipado que cumpla con la lógica de negocios. Por este motivo se establecerán unas reglas de calidad para evaluar cuantos individuos pueden ser estudiados en el análisis de la Prima pagada por anticipado.

Para finalizar obtenemos los Retiros el cual se distribuye de la siguiente manera, un promedio de -$10,51 millones, una suma total de -$396.599,13 millones con una desviación de $182,06, un mínimo de -$20.000,00 y un máximo de $350,00 millones. En este estudio se obtienen unos valores positivos, los cuales representan un error por lo cual es necesario realizar la corrección adecuada para obtener los valores adecuados en el estudio de Retiros.



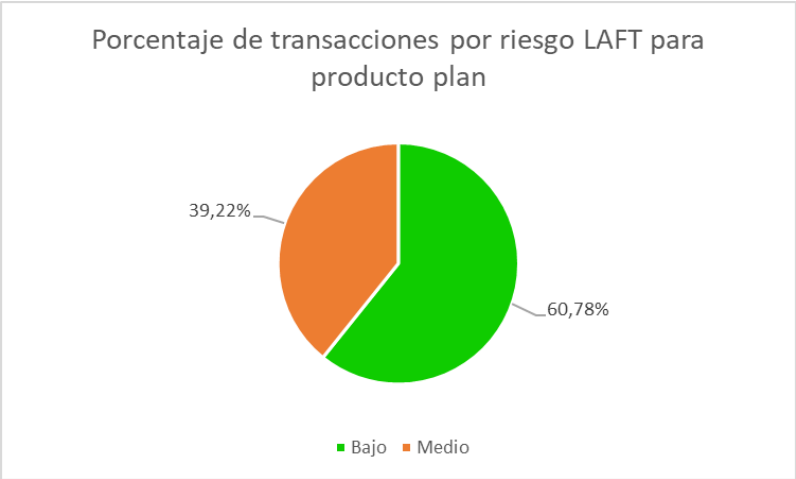
En la gráfica se observa que el mayor número de transacciones se ha realizado por el canal 5 con 337.272 transacciones seguido del canal 6 con exactamente 233.462 transacciones y el canal 3 que es tradicional como ultimo con aproximadamente 178 transacciones.

También se puede visualizar que el canal 1 que es empleados tiene el valor superior de todos con un total en movimientos negativos de -435.264 millones de pesos, seguido del canal 6 que es el canal con mayor valor de movimientos positivos con un valor de 284.623 millones y como ultimo el canal 3 que es tradicional que suma -230 millones de pesos.

	EXTRANJERA	NACIONAL	Total general
APORTE	395	46.460	46.855
CANCELACION o REVOCACION DE POLIZA	23	6.612	6.635
PAGO DE PRIMA	14.270	436.250	450.520

PRIMA PAGADA POR ANTICIPADO	8.268	141.220	149.488
RETIRO	155	37.597	37.752
Total general	**23.111**	**668.139**	**691.250**

En la tabla anterior se puede evidenciar el número de transacciones que se realizan en las diferentes jurisdicciones por departamento, en donde se obtuvo un total de 691.250 transacciones, de las cuales 668.139 se han hecho en jurisdicciones departamentales nacionales siendo la suma total de todas las etiquitas y 23.111 en el extranjero.



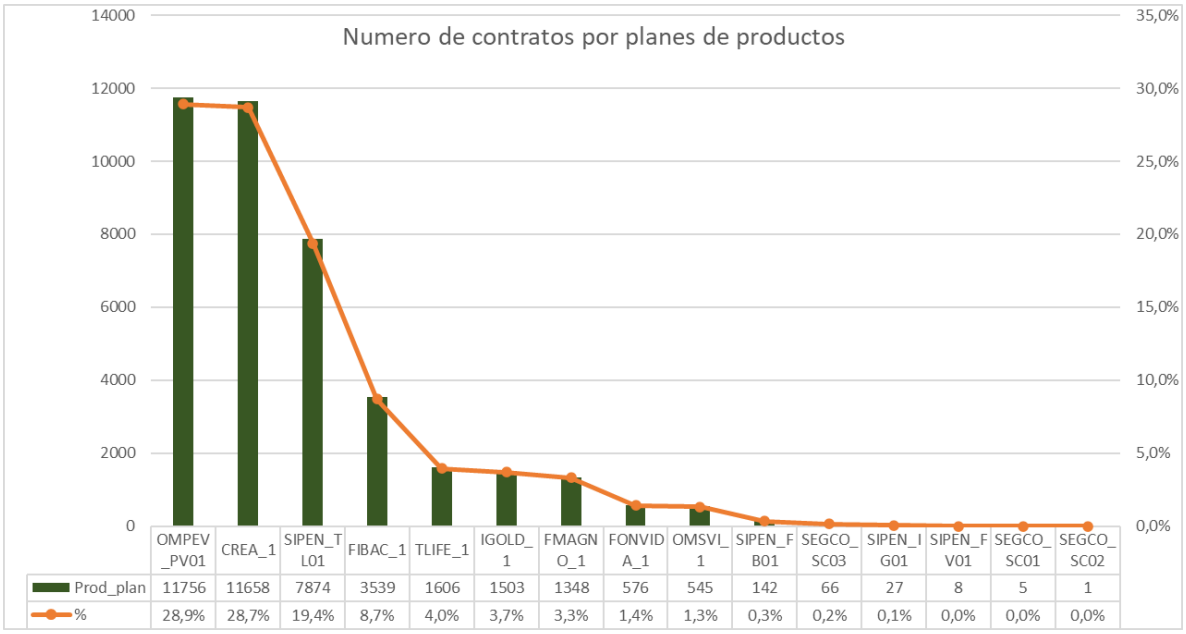
En el grafico anterior se puede observar cómo se distribuyen la cantidad de transacciones en valores porcentuales en cuanto al riesgo de producto plan, se encontró que el 60.78% de las transacciones están dirigidas a productos que tienen un riesgo LAFT bajo y el 39.22% de las transacciones tienen un riesgo medio, no se observan transacciones con riesgo alto, por lo cual se debe observar con más detenimiento las transacciones de riesgo medio.

Tabla #6: "Contratos"

Variables	Nulos

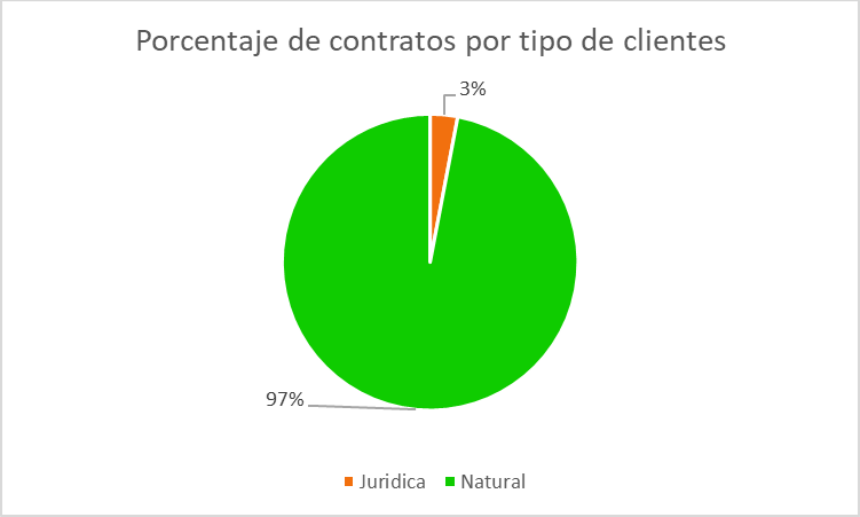
Product	0
Product_Plan	0
Restriction_Multigestion	0
Restriction_Withdrawal	0
Market_Niche	0
Management_Client	0
Minimum_Income	0
Sponsor	0
Product_Type_ID	6
Investment_Term_ID	6
Segment_Product	0
Company	0
Risk_LAFT	0

En la tabla anterior se realiza el análisis de las variables de la tabla de Contratos en donde se observó que la variable Product_Type_ID y la variable Investment_Term_ID tienen 6 datos nulos, el resto de las variables de la tabla están completos, esto quiere decir que se tiene que realizar una actualización de los datos y de estas dos variables con datos nulos para verificar la calidad de la información y de esta manera poder estudiar de manera adecuada las variables.



	OMPEV_PV01	CREA_1	SIPEN_TL01	FIBAC_1	TLIFE_1	IGOLD_1	FMAGNO_1	FONVIDA_1	OMSVI_1	SIPEN_FB01	SEGCO_SC03	SIPEN_IG01	SIPEN_FV01	SEGCO_SC01	SEGCO_SC02
Prod_plan	11756	11658	7874	3539	1606	1503	1348	576	545	142	66	27	8	5	1
%	28,9%	28,7%	19,4%	8,7%	4,0%	3,7%	3,3%	1,4%	1,3%	0,3%	0,2%	0,1%	0,0%	0,0%	0,0%

En la gráfica se observa el número de contratos por planes de productos, así como su porcentaje en relación con el total de contratos, claramente se puede notar que OMPEV_PV01 es el producto que esta enlazado a mayor número de contratos (11756) que representa el 28.9% del total seguido de CREA_1 que tiene un 28.7% del total de contratos y como producto con menos número de contratos es SEGCO_SC02 con un solo contrato.



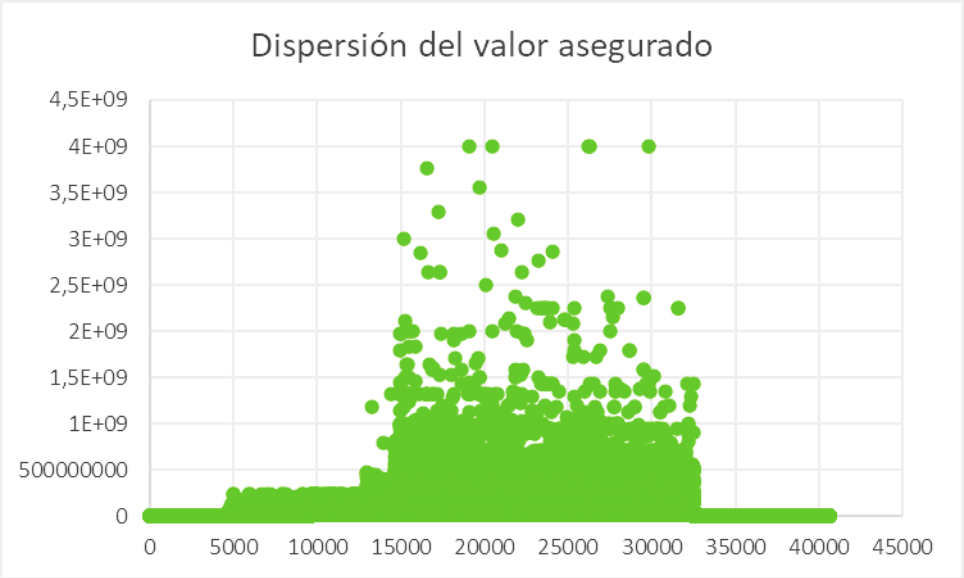
En la gráfica anterior se realizado en análisis de los contratos con los tipos de clientes, de esta manera se identificó que el 97% de los contratos han sido realizados para clientes que se identifican como personas naturales, mientras que el 3% del total de contratos fueron para clientes que son personas jurídicas, por lo tanto, el tipo de clientes que aborda la mayor cantidad de contratos son las personas naturales.

Estadísticas descriptivas	Insured_value	Total_Premium
Media	$ 116,54	$ 0,38
Mediana	$ 26,00	$ 0,30
Suma	$ 4.737.708,02	$ 15.504,96
Desviación	$ 200,92	$ 0,57
Mínimo	$ -	$ -

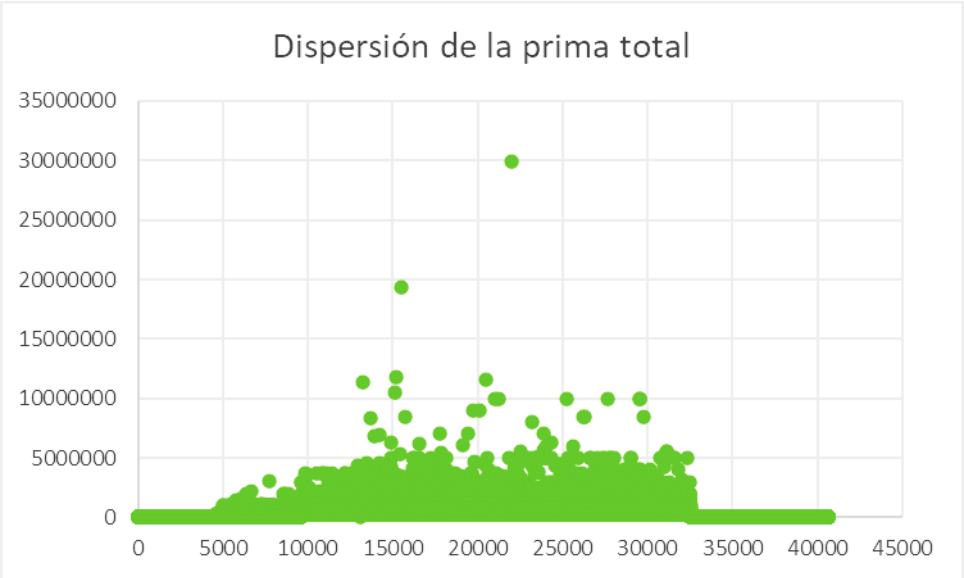
Máximo	$ 4.000,00	$ 29,91

En el recuadro anterior se realiza en análisis de las estadísticas descriptivas del total de contratos, el cual es de 40655. Se logra observar que la media de los valores asegurados de todos los contratos es de $116,54 millones de pesos, la cual se encuentra bastante elevada comparada con la información obtenida en la mediana; esto se debe al valor obtenido en el máximo, seguido a esto, tenemos la mediana se encuentra con un valor de $26,00 millones y cuenta con una desviación es de $200,92 millones, lo que nos dice que los valores de los contratos son bastante distantes y la diferencia que existe entre los errores de cada contrato es sumamente grande. Por otro lado, se observa que se tienen valores mínimos de $0 lo cual no es un valor lógico para el modelo de negocio, lo que significa que se cuenta con datos de baja calidad por lo tanto se requiere una actualización, los datos que no cumplan con la calidad no serán utilizados en la segmentación y se realizara un análisis más detallado para encontrar el valor mínimo de los valores asegurados que cumpla con la lógica de negocios. Por este motivo se establecerán unas reglas de calidad para evaluar cuantos individuos pueden ser estudiados en el análisis de los valores asegurados.

En cuanto a la prima total se puede observar que la media no logra superar $1 millón de pesos, cuenta con una media exactamente de $0,38 millones de pesos, su desviación es bastante baja, exactamente $0,57 millones de pesos por lo que los valores de los primeros totales de todos los contratos son similares y su diferencia no es tan significativa, y como valor de la prima máxima se obtiene un total de $29,91 millones de pesos, , se observa que se tienen valores mínimos de $0 lo cual no es un valor lógico para el modelo de negocio, lo que significa que se cuenta con datos de baja calidad por lo tanto se requiere una actualización, los datos que no cumplan con la calidad no serán utilizados en la segmentación y se realizara un análisis más detallado para encontrar el valor mínimo de la prima total que cumpla con la lógica de negocios. Por este motivo se establecerán unas reglas de calidad para evaluar cuantos individuos pueden ser estudiados en el análisis de la prima total.



En el grafico anterior se expone la distribución de los valores asegurados, el cual aplica para los productos de seguros de vida y en el cual se indicará el valor de la prima comprometido de acuerdo con la frecuencia. Allí mismo podemos observar la mayor concentración de los valores asegurados está entre $0 millones y $1.000,00 millones de pesos y también se pueden ver los valores máximos los cuales corresponde a $4.000,00 millones de pesos.

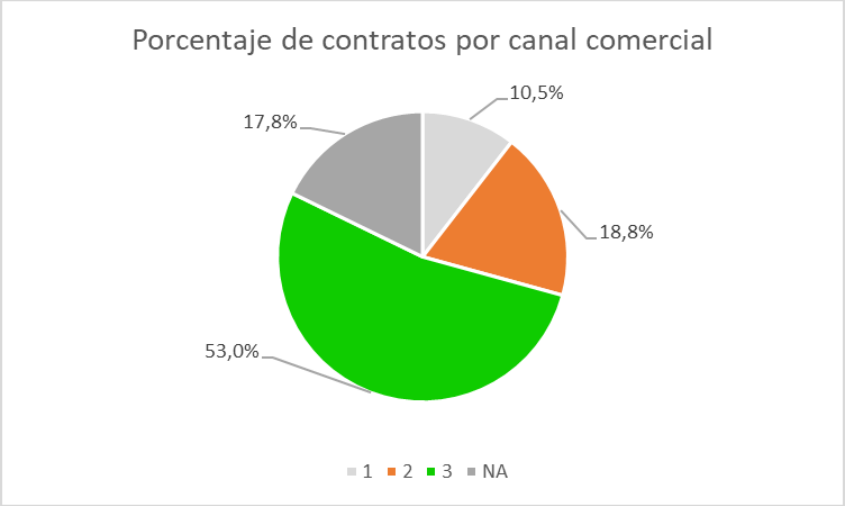


Siguiendo el estudio de la distribución de las estadísticas descriptivas, se observa que la mayor concentración de las primas totales está entre $0 y $5,00 millones de

pesos, existen gran cantidad de contratos en donde el precio de las primas totales está en 0 que en comparación con el total de contratos representa alrededor del 20%.

	Cuenta de Frequency_Payment	Porcentaje
Anual	75	0,61%
Mensual	12.226	99,39%
Total general	**12.301**	**100%**

En la tabla anterior se observa la frecuencia de pagos divididos de manera mensual y anual para un total de 12.301 contratos, estos se distribuyen de la de la siguiente manera: un total de 75 contratos los cuales cuentan con una frecuencia anual, estos representan el 0.61% del total de contratos estudiados, mientras que los contratos con una frecuencia de pago mensual constituyen el 99,39% restante lo cual es un total de 12226 contratos, es decir que la mayor cantidad de transacciones se realizan mensualmente y existe una gran cantidad de movimiento de dinero mes a mes tanto para personas naturales como jurídicas.



En la gráfica anterior podemos observar la distribución porcentual de los contratos que se efectuaron por canales comerciales en donde se obtiene un total de 40654 contratos de los cuales el 53% de ellos se realizaron por medio de un canal tradicional que tiene un riesgo alto, el 18.8% de los contratos se hicieron por el

canal 2(Intermediario) que tiene un riesgo medio como canal, el 17.8% de los contratos tienen valores nulos y el 10.5% se hizo por el canal 1 (empleados) que tienen un riesgo bajo.

Anexo 2: Variables usadas en los modelos

El siguiente nexo muestra un de detalle de las variables usadas durante todo el proceso de la construcción de los modelos de segmentación.

Factor de Riesgo	Variable(s) Incorporadas
Productos	1. Risk_LAFT: Riesgo de vinculación de la persona. 2. ReJu_med_movVal: Mediana de movimientos de retiros de personas jurídicas. 3. ReJu_sd_movVal: Desviación estándar de movimientos de retiros de personas jurídicas.
Canales de distribución	*En cuanto a la fase de segmentación para "Canales de movimientos/transacciones" las variables son*: 1. Risk_LAFT: Riesgo de vinculación de la persona. 2. Channel_Type: Se describe que transacción se puede llevar a cabo en cada uno de los canales (aportes/retiros). 3. ApJu_med_movVal: Mediana de movimientos de aportes de personas jurídicas. 4. ApNa_med_movVal: Mediana de movimientos de aportes de personas naturales.
Jurisdicciones	*En cuanto a la fase de segmentación para "Jurisdicciones países" las variables son*: 1. GAFI High Risk Jurisdictions: marca que indica si la jurisdicción está en el listado GAFI *call for action* http://www.fatf-gafi.org/countries/#high-risk

	2. GAFI Jurisdictions Under Increased Monitoring: marca que indica si la jurisdicción aparece en el listado GAFI *other monitored jurisdictions* http://www.fatf-gafi.org/countries/#other-monitored-jurisdictions 3. ApNa_sd_movVal: Desviación estándar de movimientos de aportes de personas naturales. 4. ReNa_IQR movVal: Rango intercuartílico de movimientos de retiros de personas naturales. 5. ApNa_promMovMen movVal: Promedio de movimientos mensuales de aportes de personas naturales. *En cuanto a la fase de segmentación para "Jurisdicciones departamentos" las variables son*: 1. Tasa de delitos fuente LAFT: Tasa delitos fuente Tasa de procesos por delitos fuente de Lavado de Activos por cada 100.000 habitantes. 2. Cultivos ilícitos: Permanencia de cultivos ilícitos. 3. Presencia de grupos armados: Nivel de riesgo relacionado con el número de víctimas. 4. ReJu sd movVal: desviación estándar de movimientos de retiros de personas jurídicas.
Clientes	• *En cuanto a la fase de segmentación para "Clientes persona jurídica":* 1. Id_Fe_Apor_Num_12: Numero de aportes en 12 meses. 2. Id_Fe_Ret_Mon_24: Suma de montos de movimientos de retiros en 24 meses 3. Id_Fe_Ret_Prom24: Promedio de retiros de movimeintos en 24 meses. 4. RIESGO: Risk de personas jurídicas

	Perfilamiento: 1. Id_Fe_Ing: Ingresos de personas jurídicas 2. Id_Fe_Patr: patrimonio de personas jurídicas 3. Id_Fe_Apor_Mon_12: Monto de aportes en 12 meses de personas juridicas 4. Id_Fe_valaseg_total: Total en montos de valor segurado 5. Id_Fe_primaT_total: Total en montos de prima total • *En cuanto a la fase de segmentación para "Clientes persona natural"*: 1. Id_Fe_Cap_Apor12_E: capacidad de aportes en efectivo en 12 meses. 2. Id_Fe_Egr: Egresos. 3. RIESGO: Risk de personas jurídicas. **Perfilamiento:** 1. Id_Fe_Apor_Mon_24: Monto de aportes en 24 meses de personas naturales 2. Id_Fe_Apor_E_Mon_24: Monto de aportes en efectivo en 24 meses de personas naturales 3. Id_Fe_Apor_E_Num_12: Numero de aportes en efectivo en 12 meses de personas naturales 4. Id_Fe_Cap_Apor12: Capacidad de aportes en 12 meses de personas naturales 5. Id_Fe_Apo_Prom24: Promedio de aportes en 24 meses de personas naturales 6. Id_Fe_valaseg_total: Total en montos de valor asegurado

Anexo 3: Funciones de Iteración en R

El siguiente anexo muestra el código fuente de las funciones de iteración usadas dentro de la construcción de los modelos de segmentación.

```
################################## CLARA SEGMENTATION
#########################
one_step_clara <- function(segmentation_table,
                           variables,
                           k,
                           metric = "euclidean",
                           samples = 5,
                           sampsize = 200) {
   clarax <- cluster::clara(x = segmentation_table[variables],
                            k = k,
                            metric = metric,
                            samples = samples,
                            sampsize = sampsize)
   mod_indices <- fpc::cluster.stats(clarax$diss,
                                     clarax$clustering[clarax$sample])
   n <- nrow(segmentation_table)
   max_val <- max(table(clarax$clustering))
   esperado <- n / k
   indica <- max_val / esperado
   a <- cbind(metric = metric,
              dunn = round(mod_indices$dunn2, 3),
              average = round(mod_indices$avg.silwidth, 3),
              exp_compa = round(indica, 3),
              k = k,
              round(t(mod_indices$clus.avg.silwidths), 3),
              num_vars = length(variables))
   return(as.data.frame(a))
}
#one_step_clara(data.frame(x), c('X1', 'X2'), k = 2)



semi_busqueda_clara <- function(segmentation_table,
                                variables,
                                metric,
                                k,
                                filter_average,
                                max_variables,
                                samples,
                                sampsize,
                                exp_umbral = 0.4,
                                ...) {

   exp_umbral  <-  exp_umbral * k
   dat <- NULL
   for (vari in variables) {
      try({
         resultado <- one_step_clara(segmentation_table,
```

```
                                c(vari),
                                k = k,
                                metric = metric,
                                samples = samples,
                                sampsize = sampsize)
        #print(vari)
        #print(resultado)
        resultado$variables <- vari
        dat <- dplyr::bind_rows(resultado, dat) })

}
dat <- dat[dat$average > filter_average &
             dat$exp_compa < exp_umbral, ]
print("=======")
print(paste("Numero variables", length(variables)))
print(paste("Candidatos inicial", nrow(dat)))

lista_variables_anterior <- dat$variables
lista_variables_actual <- list()

if (nrow(dat) == 0) {
   return(dat)
}

for (numero_variables in seq(2, max_variables)) {
   print("--")
   print(paste("Probando variable:", numero_variables))

   new_dat <- NULL
   ingresar_index <- 1
   print(paste("Lista_anterior", length(lista_variables_anterior)))
   print(paste("Complejidad", length(lista_variables_anterior) *
                length(variables)))
   for (combina_variables in lista_variables_anterior) {
      for (variable in variables) {
         vari <- sort(unique(c(variable, combina_variables)))
         pegadas <- paste(vari, collapse = "::")
         condi <- lapply(lista_variables_actual,
                         paste,
                         collapse = "::")
         if (! pegadas %in% condi &
            (length(vari) == numero_variables)) {
            try({
               resultado <- one_step_clara(segmentation_table,
                                           c(vari),
                                           k = k,
                                           metric = metric,
                                           samples = samples,
                                           sampsize = sampsize)
               resultado$variables <- pegadas
               new_dat <- dplyr::bind_rows(resultado, new_dat)
               lista_variables_actual[[ingresar_index]] <- vari
               ingresar_index <- ingresar_index + 1
            })
         }
      }
   }
```

```
      new_dat <- new_dat[new_dat$average > filter_average &
                         new_dat$exp_compa < exp_umbral, ]
      dat <- rbind(dat, new_dat)
      print(paste("Lista_anterior", nrow(new_dat)))
      lista_variables_anterior <- new_dat$variables
      lista_variables_anterior <- str_split(lista_variables_anterior,
                                            "::", simplify = T)
      lista_variables_anterior <- as.list(
        as.data.frame(t(lista_variables_anterior)))
      lista_variables_actual <- list()
    }

    return(dat)
}


loop_reducido_clara <- function(segmentation_table,
                                variables,
                                metric_list,
                                k_list,
                                filter_average,
                                max_variables,
                                samples,
                                sampsize,
                                exp_umbral = 0.4,
                                summary_file = NULL,
                                ...) {

    resultados <- NULL
    for (metric in metric_list) {
      for (k in k_list) {
        start_time <- Sys.time()
        resul_indi <- semi_busqueda_clara(segmentation_table,
                                          variables,
                                          metric,
                                          k,
                                          filter_average = filter_average,
                                          max_variables = max_variables,
                                          samples = samples,
                                          sampsize = sampsize,
                                          exp_umbral = exp_umbral,
                                          ...)
        end_time <- Sys.time()
        tiempo_estimado <- difftime(end_time, start_time,
                                    units = "mins")
        resul_indi <-
          resul_indi %>%
          arrange(desc(average), desc(num_vars), exp_compa)
        cat(paste("Making the results of",
                  "\nMetric : ", metric,
                  "\nNumber of segments: ", k,
                  "\nTiempo estimado ", tiempo_estimado,
                  "\nDim busqueda ", nrow(resul_indi),
                  "\nMax variables ", max_variables,
                  "\n\n", sep = ""))
        if (!is.null(summary_file)) {
          write.table(resul_indi, file = summary_file,
```

```
                    sep = ",",
                    append = TRUE, quote = FALSE,
                    col.names = TRUE, row.names = FALSE)
          cat(paste("Writing results in --->",
                    summary_file,
                    "\n\n"))
        }
        gc()
        resultados <- dplyr::bind_rows(resultados, resul_indi)
      }
    }
    resultados <-
      resultados %>%
      arrange(desc(average), desc(num_vars), exp_compa)
    return(resultados)
}
```

Anexo 4: Código de Alertas de operaciones inusuales

```
#######        alertas de movimientos inusuales
 ########################################################################
 ########################################################################
 #identifica cada segmento_DPTO
 cons <- paste0("select * from Jurisdiction_Department_Segment where Company='VIDA' and
Version='",VERS_VIG,"'")
 con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
 segJuris <- as.data.table(dbGetQuery(con, cons))
 dbDisconnect(con)
 segJuris<-dcast(segJuris,Jurisd_ID+Segment_Department~., value.var="Segment_Department",
fun.aggregate = length)
 segJuris<-segJuris[,c(1,2)]
 segJuris$Segment_Department<-as.factor(segJuris$Segment_Department)
 names(segJuris)[names(segJuris)=="Jurisd_ID"]<-"JURISD_MOVIMIENTO"
 Movimientos<- left_join(Movimientos, segJuris, by="JURISD_MOVIMIENTO")
 names(Movimientos)[names(Movimientos)=="Segment_Department"]<-
"SEGMENTO_SARLAFT_JURISDICCIONES_D"

 #identifica cada segmento_PAIS
 cons <- paste0("select * from Jurisdiction_Country_Segment where Company='VIDA' and
Version='",VERS_VIG,"'")
 con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
 segJurisP <- as.data.table(dbGetQuery(con, cons))
 dbDisconnect(con)
 segJurisP<-dcast(segJurisP,Jurisd_ID+Segment_Country~., value.var="Segment_Country",
fun.aggregate = length)
 segJurisP<-segJurisP[,c(1,2)]
 segJurisP$Segment_Country<-as.factor(iconv(segJurisP$Segment_Country,"latin1", "UTF-8"))
 names(segJurisP)[names(segJurisP)=="Jurisd_ID"]<-"Jurisd_ID_Country_mov"
 Movimientos<- left_join(Movimientos, segJurisP, by="Jurisd_ID_Country_mov")
 names(Movimientos)[names(Movimientos)=="Segment_Country"]<-
"SEGMENTO_SARLAFT_JURISDICCIONES_P"
 names(Movimientos)[28]<-"SEGMENTO_SARLAFT_JURISDICCIONES_P"
```

```
  #identifica cada segmento_jurisdiccion(DEPTO o PAIS respectivamente)
  Movimientos[,SEGMENTO_SARLAFT_JURISDICCIONES:=ifelse(!is.na(Jurisd_ID_Country_mov)
&                                                    Jurisd_ID_Country_mov==1,
as.character(SEGMENTO_SARLAFT_JURISDICCIONES_D),as.character(SEGMENTO_SARLAF
T_JURISDICCIONES_P))]
  bigMap                                                                    <-
mapLevels(x=list(Movimientos$SEGMENTO_SARLAFT_JURISDICCIONES_D,Movimientos$SEG
MENTO_SARLAFT_JURISDICCIONES_P),codes=FALSE,combine=TRUE)
  mapLevels(Movimientos$SEGMENTO_SARLAFT_JURISDICCIONES) <- bigMap

  #identifica cada segmento_canales_movimiento
  cons <- paste0("select * from Channel_Customer_Movement_Segment where Company='VIDA'
and Version='",VERS_VIG,"'")
  con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
  segCanalm <- as.data.table(dbGetQuery(con, cons))
  dbDisconnect(con)
  segCanalm<-dcast(segCanalm,Channel_ID+Segment_Channel~.,
value.var="Segment_Channel", fun.aggregate = length)
  segCanalm<-segCanalm[,c(1:2)]
  segCanalm$Segment_Channel<-as.factor(segCanalm$Segment_Channel)
  names(segCanalm)[names(segCanalm)=="Channel_ID"]<-"CANAL_MOVIMIENTO"
  Movimientos<- left_join(Movimientos, segCanalm, by="CANAL_MOVIMIENTO")
  names(Movimientos)[names(Movimientos)=="Segment_Channel"]<-
"SEGMENTO_SARLAFT_CANAL_MOVIMIENTO"

  #identifica cada segmento_productos
  cons <- paste0("select * from Product_Plan_Detail_Segment where Company='VIDA' and
Version='",VERS_VIG,"'")
  con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
  segProdu <- as.data.table(dbGetQuery(con, cons))
  dbDisconnect(con)
  segProdu<-dcast(segProdu,Product+Product_Plan+Segment_Product~.,
value.var="Segment_Product", fun.aggregate = length)
  segProdu<-segProdu[,c(1:3)]
  segProdu$Segment_Product<-as.factor(segProdu$Segment_Product)
  segProdu$PRODUCTO<-paste(segProdu$Product,segProdu$Product_Plan, sep = "_")
  Movimientos<- left_join(Movimientos, segProdu[,c(4,3)], by="PRODUCTO")
  names(Movimientos)[names(Movimientos)=="Segment_Product"]<-
"SEGMENTO_SARLAFT_PRODUCTO"

  #identifica cada segmento_clientes

  cons <- paste0("select * from Client_Segment where Company='VIDA' and Version='",VERS_VIG,"'
and Cutoff_Date='",fechaUltimaSeg,"'")
  con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
  clientesSeg <- as.data.table(dbGetQuery(con, cons))
  dbDisconnect(con)
  clientesSeg<-dcast(clientesSeg,Document_Type+Document_Number+Segment_Client~.,
value.var="Segment_Client", fun.aggregate = length, subset = .(Segment_Mark==TRUE) )
  clientesSeg<-clientesSeg[,c(1:3)]
  clientesSeg$Segment_Client<-as.factor(clientesSeg$Segment_Client)
  names(clientesSeg)[names(clientesSeg)=="Document_Type"]<-"TIPO_DOC"
  names(clientesSeg)[names(clientesSeg)=="Document_Number"]<-"DOCUMENTO_DV"
```

```
 clientesSeg[,DOCUMENTO_DV:=as.character(DOCUMENTO_DV)]
 Movimientos<- left_join(Movimientos, clientesSeg, by=c("TIPO_DOC","DOCUMENTO_DV"))
 names(Movimientos)[names(Movimientos)=="Segment_Client"]<-
"SEGMENTO_SARLAFT_CLIENTES"


################################################################################
##########################################

 Movimientos$TIPO_MovTra<-ifelse(Movimientos$MOVIMIENTO=="CANCELACION o
REVOCACION DE POLIZA", "CANCELACION o REVOCACION DE POLIZA",
Movimientos$TIPO_TRANSACCION)
 Movimientos$TIPO_MovTra<-recode_factor(Movimientos$TIPO_MovTra,
                                "1"="APORTES",
                                "2"="RETIROS",
                                "CANCELACION o REVOCACION DE POLIZA"="CAN_REV_POLIZA")

 ################################################################################
 ################################################################################

 if((as.character(format(Ejecucion,'%d')))=="01") {

  cons <- paste0("select Cutoff_Date from [Alert_Unusual_Operations_Limit] where
Cutoff_Date='",fechaUltimaSeg,"'", " and Company='VIDA'")
  con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
  ALert_CP <- dbGetQuery(con, cons)
  dbDisconnect(con)

  if(dim(ALert_CP)[1]==0){
   con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")

   #######
   LimiteP<-dcast(Movimientos,SEGMENTO_SARLAFT_CLIENTES +
SEGMENTO_SARLAFT_PRODUCTO +TIPO_MovTra ~ ., fun=LimiteAlert,
value.var=c("VALOR_MOVIMIENTO"), drop = FALSE)
   names(LimiteP)[4]<-"lim_Cli_Prod"
   LimitePP<-dcast(LimiteP,SEGMENTO_SARLAFT_CLIENTES +
SEGMENTO_SARLAFT_PRODUCTO ~ TIPO_MovTra, fun=max, value.var="lim_Cli_Prod", drop =
FALSE)
   #
   LimiteM<-dcast(Movimientos,SEGMENTO_SARLAFT_CLIENTES +
SEGMENTO_SARLAFT_CANAL_MOVIMIENTO +TIPO_MovTra~ ., fun=LimiteAlert,
value.var=c("VALOR_MOVIMIENTO"), drop = FALSE)
   names(LimiteM)[4]<-"lim_Cli_CanalM"
   LimiteMM<-dcast(LimiteM,SEGMENTO_SARLAFT_CLIENTES +
SEGMENTO_SARLAFT_CANAL_MOVIMIENTO ~ TIPO_MovTra, fun=max,
value.var="lim_Cli_CanalM", drop = FALSE)
   #
   LimiteJ<-dcast(Movimientos,SEGMENTO_SARLAFT_CLIENTES +
SEGMENTO_SARLAFT_JURISDICCIONES+TIPO_MovTra ~ ., fun=LimiteAlert,
value.var=c("VALOR_MOVIMIENTO"), drop = FALSE)
   names(LimiteJ)[4]<-"lim_Cli_Jurisdiccion"
   LimiteJJ<-dcast(LimiteJ,SEGMENTO_SARLAFT_CLIENTES +
SEGMENTO_SARLAFT_JURISDICCIONES ~ TIPO_MovTra, fun=max,
value.var="lim_Cli_Jurisdiccion", drop = FALSE)
```

```
    #
    # efectivo
    LimiteP_E<-
dcast(Movimientos[MODALIDAD=="Efectivo",],SEGMENTO_SARLAFT_CLIENTES +
SEGMENTO_SARLAFT_PRODUCTO +TIPO_MovTra ~ ., fun=LimiteAlert,
value.var=c("VALOR_MOVIMIENTO"), drop = FALSE)
    names(LimiteP_E)[4]<-"lim_Cli_Prod_E"
    LimitePP_E<-dcast(LimiteP_E,SEGMENTO_SARLAFT_CLIENTES +
SEGMENTO_SARLAFT_PRODUCTO ~ TIPO_MovTra, fun=max, value.var="lim_Cli_Prod_E",
drop = FALSE)
    #
    LimiteM_E<-
dcast(Movimientos[MODALIDAD=="Efectivo",],SEGMENTO_SARLAFT_CLIENTES +
SEGMENTO_SARLAFT_CANAL_MOVIMIENTO +TIPO_MovTra~ ., fun=LimiteAlert,
value.var=c("VALOR_MOVIMIENTO"), drop = FALSE)
    names(LimiteM_E)[4]<-"lim_Cli_CanalM_E"
    LimiteMM_E<-dcast(LimiteM_E,SEGMENTO_SARLAFT_CLIENTES +
SEGMENTO_SARLAFT_CANAL_MOVIMIENTO ~ TIPO_MovTra, fun=max,
value.var="lim_Cli_CanalM_E", drop = FALSE)
    #
    LimiteJ_E<-
dcast(Movimientos[MODALIDAD=="Efectivo",],SEGMENTO_SARLAFT_CLIENTES +
SEGMENTO_SARLAFT_JURISDICCIONES+TIPO_MovTra ~ ., fun=LimiteAlert,
value.var=c("VALOR_MOVIMIENTO"), drop = FALSE)
    names(LimiteJ_E)[4]<-"lim_Cli_Jurisdiccion_E"
    LimiteJJ_E<-dcast(LimiteJ_E,SEGMENTO_SARLAFT_CLIENTES +
SEGMENTO_SARLAFT_JURISDICCIONES ~ TIPO_MovTra, fun=max,
value.var="lim_Cli_Jurisdiccion_E", drop = FALSE)

    ##################
    # insercion en la tabla Alert_Unusual_Operations_Limit
    LimiteP$Alert_Code<-"Alert_OI3"
    LimiteM$Alert_Code<-"Alert_OI1"
    LimiteJ$Alert_Code<-"Alert_OI2"
    LimiteP_E$Alert_Code<-"Alert_OI6"
    LimiteM_E$Alert_Code<-"Alert_OI4"
    LimiteJ_E$Alert_Code<-"Alert_OI5"

    names(LimiteP)[4]<-"Limit"
    names(LimiteM)[4]<-"Limit"
    names(LimiteJ)[4]<-"Limit"
    names(LimiteP_E)[4]<-"Limit"
    names(LimiteM_E)[4]<-"Limit"
    names(LimiteJ_E)[4]<-"Limit"

    names(LimiteP)[2]<-"Segment_Factor"
    names(LimiteM)[2]<-"Segment_Factor"
    names(LimiteJ)[2]<-"Segment_Factor"
    names(LimiteP_E)[2]<-"Segment_Factor"
    names(LimiteM_E)[2]<-"Segment_Factor"
    names(LimiteJ_E)[2]<-"Segment_Factor"

    tablasDcast<-
rbind(LimiteP[,c(5,1:4)],LimiteM[,c(5,1:4)],LimiteJ[,c(5,1:4)],LimiteP_E[,c(5,1:4)],LimiteM_E[,c(5,1:4)
],LimiteJ_E[,c(5,1:4)])
    names(tablasDcast)[4]<-"Transaction_Category"
```

```
    tablasDcast$Cutoff_Date=as.Date(fechaUltimaSeg,"%Y-%m-%d")
    tablasDcast$Company="VIDA"
    tablasDcast<-tablasDcast[,c(1,7,6,2:5)]
    names(tablasDcast)<-
c("Alert_Code","Company","Cutoff_Date","Segment_Client","Segment_Factor","Transaction_Categ
ory","Limit")
    tablasDcast$Segment_Client<-as.character(tablasDcast$Segment_Client)
    tablasDcast$Segment_Factor<-as.character(tablasDcast$Segment_Factor)
    tablasDcast$Transaction_Category<-as.character(tablasDcast$Transaction_Category)

    dbWriteTable (con, "Alert_Unusual_Operations_Limit", tablasDcast,append=TRUE)
    dbDisconnect(con)

    ###
    cons <- paste0("select * from [Alert_Unusual_Operations_Limit] where
Cutoff_Date='",fechaUltimaSeg,"'", " and Company='VIDA'")
    con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
    tablasDcast <- as.data.table(dbGetQuery(con, cons))
    dbDisconnect(con)

    LimiteM<-tablasDcast[Alert_Code=="Alert_OI1",-c(1:3)]
    names(LimiteM)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_CANAL_MOVIMIENTO","TIPO
_MovTra","lim_Cli_CanalM")

    LimiteJ<-tablasDcast[Alert_Code=="Alert_OI2",-c(1:3)]
    names(LimiteJ)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_JURISDICCIONES","TIPO_Mo
vTra","lim_Cli_Jurisdiccion")

    LimiteP<-tablasDcast[Alert_Code=="Alert_OI3",-c(1:3)]
    names(LimiteP)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_PRODUCTO","TIPO_MovTra","
lim_Cli_Prod")

    LimiteM_E<-tablasDcast[Alert_Code=="Alert_OI4",-c(1:3)]
    names(LimiteM_E)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_CANAL_MOVIMIENTO","TIPO
_MovTra","lim_Cli_CanalM_E")

    LimiteJ_E<-tablasDcast[Alert_Code=="Alert_OI5",-c(1:3)]
    names(LimiteJ_E)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_JURISDICCIONES","TIPO_Mo
vTra","lim_Cli_Jurisdiccion_E")

    LimiteP_E<-tablasDcast[Alert_Code=="Alert_OI6",-c(1:3)]
    names(LimiteP_E)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_PRODUCTO","TIPO_MovTra","
lim_Cli_Prod_E")
  }

  if(dim(ALert_CP)[1]!=0){
    cons <- paste0("select * from [Alert_Unusual_Operations_Limit] where
Cutoff_Date='",fechaUltimaSeg,"'", " and Company='VIDA'")
    con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
```

```
    tablasDcast <- as.data.table(dbGetQuery(con, cons))
    dbDisconnect(con)

    LimiteM<-tablasDcast[Alert_Code=="Alert_OI1",-c(1:3)]
    names(LimiteM)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_CANAL_MOVIMIENTO","TIPO
_MovTra","lim_Cli_CanalM")

    LimiteJ<-tablasDcast[Alert_Code=="Alert_OI2",-c(1:3)]
    names(LimiteJ)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_JURISDICCIONES","TIPO_Mo
vTra","lim_Cli_Jurisdiccion")

    LimiteP<-tablasDcast[Alert_Code=="Alert_OI3",-c(1:3)]
    names(LimiteP)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_PRODUCTO","TIPO_MovTra","
lim_Cli_Prod")

    LimiteM_E<-tablasDcast[Alert_Code=="Alert_OI4",-c(1:3)]
    names(LimiteM_E)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_CANAL_MOVIMIENTO","TIPO
_MovTra","lim_Cli_CanalM_E")

    LimiteJ_E<-tablasDcast[Alert_Code=="Alert_OI5",-c(1:3)]
    names(LimiteJ_E)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_JURISDICCIONES","TIPO_Mo
vTra","lim_Cli_Jurisdiccion_E")

    LimiteP_E<-tablasDcast[Alert_Code=="Alert_OI6",-c(1:3)]
    names(LimiteP_E)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_PRODUCTO","TIPO_MovTra","
lim_Cli_Prod_E")

  }
   } else {
   cons <- paste0("select * from [Alert_Unusual_Operations_Limit] where
Cutoff_Date='",fechaUltimaSeg,"'", " and Company='VIDA'")
   con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
   tablasDcast <- as.data.table(dbGetQuery(con, cons))
   dbDisconnect(con)

   LimiteM<-tablasDcast[Alert_Code=="Alert_OI1",-c(1:3)]
   names(LimiteM)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_CANAL_MOVIMIENTO","TIPO
_MovTra","lim_Cli_CanalM")

   LimiteJ<-tablasDcast[Alert_Code=="Alert_OI2",-c(1:3)]
   names(LimiteJ)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_JURISDICCIONES","TIPO_Mo
vTra","lim_Cli_Jurisdiccion")

   LimiteP<-tablasDcast[Alert_Code=="Alert_OI3",-c(1:3)]
   names(LimiteP)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_PRODUCTO","TIPO_MovTra","
lim_Cli_Prod")
```

```
   LimiteM_E<-tablasDcast[Alert_Code=="Alert_OI4",-c(1:3)]
   names(LimiteM_E)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_CANAL_MOVIMIENTO","TIPO
_MovTra","lim_Cli_CanalM_E")

   LimiteJ_E<-tablasDcast[Alert_Code=="Alert_OI5",-c(1:3)]
   names(LimiteJ_E)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_JURISDICCIONES","TIPO_Mo
vTra","lim_Cli_Jurisdiccion_E")

   LimiteP_E<-tablasDcast[Alert_Code=="Alert_OI6",-c(1:3)]
   names(LimiteP_E)<-
c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_PRODUCTO","TIPO_MovTra","
lim_Cli_Prod_E")

 }


 ###

 Movimientos<-merge(Movimientos,LimiteM,
by=c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_CANAL_MOVIMIENTO","TI
PO_MovTra"),all.x=TRUE)
 Movimientos<-merge(Movimientos,LimiteJ,
by=c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_JURISDICCIONES","TIPO_
MovTra"),all.x=TRUE)
 Movimientos<-merge(Movimientos,LimiteP,
by=c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_PRODUCTO","TIPO_MovTr
a"),all.x=TRUE)
 Movimientos<-merge(Movimientos,LimiteM_E,
by=c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_CANAL_MOVIMIENTO","TI
PO_MovTra"),all.x=TRUE)
 Movimientos<-merge(Movimientos,LimiteJ_E,
by=c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_JURISDICCIONES","TIPO_
MovTra"),all.x=TRUE)
 Movimientos<-merge(Movimientos,LimiteP_E,
by=c("SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_PRODUCTO","TIPO_MovTr
a"),all.x=TRUE)

 Movimientos[,Alerta_OI1:=ifelse(VALOR_MOVIMIENTO > lim_Cli_CanalM,1,0)]
 Movimientos[,Alerta_OI2:=ifelse(VALOR_MOVIMIENTO > lim_Cli_Jurisdiccion,1,0)]
 Movimientos[,Alerta_OI3:=ifelse(VALOR_MOVIMIENTO > lim_Cli_Prod,1,0)]
 Movimientos[,Alerta_OI4:=ifelse((VALOR_MOVIMIENTO > lim_Cli_CanalM_E &
MODALIDAD=="Efectivo"),1,0)]
 Movimientos[,Alerta_OI5:=ifelse((VALOR_MOVIMIENTO > lim_Cli_Jurisdiccion_E&
MODALIDAD=="Efectivo"),1,0)]
 Movimientos[,Alerta_OI6:=ifelse((VALOR_MOVIMIENTO > lim_Cli_Prod_E&
MODALIDAD=="Efectivo"),1,0)]

 Movimientos$N_ALERTAS<-rowSums(Movimientos[,c(40:45),with=FALSE],na.rm=TRUE)
 Movimientos[,Alerta_CANAL:=ifelse(Alerta_OI1==1 | Alerta_OI4==1,1,0)]
 Movimientos[,Alerta_JURISD:=ifelse(Alerta_OI2==1 | Alerta_OI5==1,1,0)]
 Movimientos[,Alerta_PRODUCTO:=ifelse(Alerta_OI3==1 | Alerta_OI6==1,1,0)]
 Movimientos[,N_FACTORES:=rowSums(Movimientos[,c(47:49),with=FALSE],na.rm=TRUE)]
 Movimientos[,SEMAFORO:=ifelse(N_FACTORES==1, "AMARILLO","VERDE")]
 Movimientos[,SEMAFORO:=ifelse(N_FACTORES==2, "NARANJA",SEMAFORO)]
 Movimientos[,SEMAFORO:=ifelse(N_FACTORES==3, "ROJO",SEMAFORO)]
```

```
 Movimientos$FechaAlerta=Sys.Date()

 # aqui cambiar los nombres a informe 2
 Movimientos<- Movimientos %>% separate(PRODUCTO, c("Product", "Product_Plan"), "_")

 # Transaction_Segment
 columnas<-
c("Event_Number","Transaction_Number","CONTRATO","Product","Product_Plan","FECHA_MOVI
MIENTO","SEGMENTO_SARLAFT_CLIENTES",

"SEGMENTO_SARLAFT_PRODUCTO","SEGMENTO_SARLAFT_JURISDICCIONES","SEGMEN
TO_SARLAFT_CANAL_MOVIMIENTO")
 TRX<-Movimientos[FECHA_MOVIMIENTO==CorteBuscado,..columnas]
 names(TRX)<-
c("Event_Number","Transaction_Number","Contract_ID","Product","Product_Plan","Movement_Dat
e",

"Segment_Client","Segment_Product","Segment_Jurisdiction","Segment_CH_Movement")
 TRX$Cutoff_Date<-as.Date(fechaUltimaSeg,"%Y-%m-%d")
 TRX$Version<-VERS_VIG

 cons <- paste0("SELECT
                TS.Cutoff_Date
                FROM [Transaction_Segment] TS
                LEFT JOIN Contract C
                ON (TS.Contract_ID = C.Contract_ID)
                AND (C.Product  = TS.Product)
                AND (C.Product_Plan = TS.Product_Plan)
                INNER JOIN ViewSEG_Product VP
                ON (VP.Product = TS.Product)
                AND (VP.Product_Plan  = TS.Product_Plan)
                AND (VP.Company = 'VIDA')
                AND (TS.Movement_Date = '", CorteBuscado, "')")

 con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
 TRX_BD <- dbGetQuery(con, cons)
 dbDisconnect(con)

 if(dim(TRX_BD)[1]==0){
  con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
  dbWriteTable (con, "Transaction_Segment",
TRX[TRX$Movement_Date==CorteBuscado,],append=TRUE)
  dbDisconnect(con)
 }

 ########### Alert_Unusual_Operations
 columnas<-
c("Event_Number","Transaction_Number","CONTRATO","Product","Product_Plan","TIPO_DOC","
DOCUMENTO_DV","FECHA_MOVIMIENTO",

"VALOR_MOVIMIENTO","MODALIDAD","CANAL_MOVIMIENTO","Jurisd_ID_Country_mov","JURI
SD_MOVIMIENTO",

"TIPO_MovTra","SEGMENTO_SARLAFT_CLIENTES","SEGMENTO_SARLAFT_PRODUCTO","S
EGMENTO_SARLAFT_JURISDICCIONES",
```

```
"SEGMENTO_SARLAFT_CANAL_MOVIMIENTO","lim_Cli_CanalM","lim_Cli_Jurisdiccion","lim_Cli
_Prod",

"lim_Cli_CanalM_E","lim_Cli_Jurisdiccion_E","lim_Cli_Prod_E","Alerta_OI1","Alerta_OI2","Alerta_O
I3",

"Alerta_OI4","Alerta_OI5","Alerta_OI6","N_ALERTAS","SEMAFORO","Alerta_CANAL","Alerta_JUR
ISD","Alerta_PRODUCTO")

 informe2<-Movimientos[FECHA_MOVIMIENTO==CorteBuscado & N_ALERTAS>0 &
VALOR_MOVIMIENTO>MIN_VAL_ALERTA, ..columnas]

 names(informe2)<-
c("Event_Number","Transaction_Number","Contract_ID","Product","Product_Plan","Document_Typ
e","Document_Number","Movement_Date",

"Movement_Value","Modality","Movement_Channel","Jurisd_ID_Country","Movement_Jurisdiction",
"Transaction_Category",

"Segment_Client","Segment_Product","Segment_Jurisdiction","Segment_Channel","lim_Cli_CH_M
ov",

"lim_Cli_Jurisd","lim_Cli_Prod","lim_Cli_CH_Mov_E","lim_Cli_Jurisd_E","lim_Cli_Prod_E","Alert_OI
1","Alert_OI2",

"Alert_OI3","Alert_OI4","Alert_OI5","Alert_OI6","N_Alert","Alert_Operation_Level","Alerta_CANAL",
"Alerta_JURISD","Alerta_PRODUCTO")
 informe2[,Company:="VIDA"]
 informe2[,Cutoff_Date:=fechaUltimaSeg]
 informe2[,Alert_Date:=Sys.Date()]

 fwrite(informe2, paste0(ruta,"/ALERTAS OI VIDA ",Ejecucion,".csv"),sep=";", row.names=FALSE,
dec=",",qmethod="double",quote=TRUE)

 if (dim(informe2)[1]==0){
  result3 <- data.table(Alert_Code=character(), Company=character(),
Cutoff_Date=Date(),Segment_Client=character(), Segment_Factor=character(),
            Transaction_Category=character(),Event_Number=character(),
Transaction_Number=character(), Contract_ID=numeric(),
            Product=character(), Product_Plan=character(),
Movement_Date=Date(),Alert_Date=Date(),Alert_Operation_Level=character())
  result3$Cutoff_Date=as.Date(result3$Cutoff_Date,format = "%Y-%m-%d")
  result3$Movement_Date=as.Date(result3$Movement_Date,format = "%Y-%m-%d")
  result3$Alert_Date=as.Date(result3$Alert_Date,format = "%Y-%m-%d")

  result3<-as.data.frame(result3)

 }else{
  result3<-melt(informe2, id.vars =
c("Company","Cutoff_Date","Segment_Client","Segment_Product",
                "Segment_Jurisdiction","Segment_Channel","Transaction_Category",
                "Event_Number","Transaction_Number","Contract_ID","Product",
                "Product_Plan","Movement_Date","Alert_Date","Alert_Operation_Level"),
       measure.vars =
c("Alert_OI1","Alert_OI2","Alert_OI3","Alert_OI4","Alert_OI5","Alert_OI6"))
  result3<-result3[result3$value>0&!is.na(result3$value),]
```

```
   names(result3)[names(result3)=="variable"]<-"Alert_Code"
   result3<-as.data.table(result3)
   result3[,Segment_Factor:=ifelse(Alert_Code=="Alert_OI1" |
Alert_Code=="Alert_OI4",as.character(Segment_Channel),NA)]
   result3[,Segment_Factor:=ifelse(Alert_Code=="Alert_OI2" |
Alert_Code=="Alert_OI5",as.character(Segment_Jurisdiction),Segment_Factor)]
   result3[,Segment_Factor:=ifelse(Alert_Code=="Alert_OI3" |
Alert_Code=="Alert_OI6",as.character(Segment_Product),Segment_Factor)]

   columnas<-c("Alert_Code","Company","Cutoff_Date","Segment_Client","Segment_Factor",
           "Transaction_Category","Event_Number","Transaction_Number","Contract_ID",
           "Product","Product_Plan","Movement_Date","Alert_Date","Alert_Operation_Level")

   result3<-result3[, ..columnas]

setorderv(result3,c("Event_Number","Transaction_Number","Product","Product_Plan"),c(1,1,1,1))
   result3$Cutoff_Date=as.Date(result3$Cutoff_Date,format = "%Y-%m-%d")
   result3$Movement_Date=as.Date(result3$Movement_Date,format = "%Y-%m-%d")
   result3$Alert_Date=as.Date(result3$Alert_Date,format = "%Y-%m-%d")
   result3<-as.data.frame(result3)

  }

  #############
  cons <- paste0("select Cutoff_Date from [Alert_Unusual_Operations] where
Cutoff_Date='",fechaUltimaSeg,"'", " and Company='VIDA' and Movement_Date='", CorteBuscado,
"'")

  con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
  ALert_CP <- dbGetQuery(con, cons)
  dbDisconnect(con)

  if(dim(ALert_CP)[1]==0){
   con <- DBI::dbConnect(odbc::odbc(), "Segmentum",uid = "Segmentation_usr", pwd =
"Youneed2020")
   dbWriteTable (con, "Alert_Unusual_Operations",
result3[result3$Movement_Date==CorteBuscado,],append=TRUE)
   dbDisconnect(con)
  }

  # El campo Load_ID es auto-incrementable
  # Diego en la tabla de Log_Daily_Load... Puede registrar la marcacion que nos estaba preguntando
en la reunion
  print(paste0("PROCESO FINALIZADO ",Ejecucion))
}
```