



Escuela de Administración
Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Modelo de Segmentación para SARLAFT en R4G

Presentado por:

Natalia Ramos Barriga

Bogotá, D.C. 20 de junio de 2023



Escuela de Administración
Escuela de Ingeniería Ciencia y Tecnología

Maestría en Business Analytics

Modelo de Segmentación para SARLAFT en R4G

Presentado por:

Natalia Ramos Barriga

Bajo la dirección de:

Sergio María Gutiérrez Bonnet

Bogotá, D.C. 20 de junio de 2023

Tabla de Contenido

Tabla de Contenido	3
Declaración de exoneración de responsabilidad	6
Lista de Tablas	8
Abreviaturas	9
Glosario	10
Resumen Ejecutivo.....	11
Palabras clave:.....	11
Abstract	12
Keywords:	12
1. Introducción.....	13
2. Objetivos.....	15
2.1. Objetivo General	15
2.2. Objetivos Específicos	15
3. Alcance	16
4. Metodología.....	17
5. Cronograma	20
6. Entendimiento del Negocio	24
6.1. Contexto Sistema Financiero	24
6.2. R4G	25
6.3. Gestión de la Entidad en SARLAFT.....	26
6.3.1. Entendimiento y Preparación de los Datos	26
6.4. Criterios Establecidos para las Variables Seleccionadas	28
6.5. Factores de Riesgo y Variables Definidas.....	28
6.6. Descripción de las Variables de Segmentación.....	31
6.6.1. Residencia	31
6.6.2. Departamento	33
6.6.3. Actividad Económica.....	34
6.6.4. Canal y Producto.....	35
6.6.5. PEP.....	36
6.6.6. Reputación	36
6.6.7. Aporte	37
6.6.8. Ingresos y Patrimonio	37
6.7. Ponderación de Riesgo de las Variables	38

6.8.	Intervalos de Clase para las Variables Cuantitativas	39
6.9.	Preparación de los Datos	41
7.	Modelado	51
7.1.	Modelos de Aprendizaje.....	51
7.1.1.	Selección del Modelo de Aprendizaje	52
7.2.	Algoritmos de Segmentación	52
7.2.1.	Selección de los Algoritmos de Segmentación.....	54
7.3.	Ejecución de los Algoritmos Seleccionados	56
7.3.1.	Ejecución de los Algoritmos de Segmentación en Python	84
8.	Evaluación	93
8.1.	Indicadores para Evaluar Modelos de Segmentación	93
8.2.	Selección de los Indicadores	94
8.3.	Resultados de la Evaluación.....	96
9.	Despliegue	101
9.1.	Pasos Realizados en el Despliegue:	101
10.	Plan y Recomendaciones de Implementación y Aplicación	114
10.1.	Implementación del Modelo de Segmentación Seleccionado:	114
10.2.	Monitoreo y Calibración:.....	114
10.3.	Implementación de una Política de Admisión:	115
10.4.	Sistema de Atribución de Riesgo y Categorías:	115
10.5.	Visualización de los Clúster en Power BI:	118
11.	Resultados de las Recomendaciones.....	119
11.1.	Modelo de Segmentación	119
11.2.	Monitoreo y Calibración:.....	119
11.3.	Implementación de una Política de Admisión:	119
11.4.	Sistema de Atribución de Riesgo y Categorías:	119
11.5.	Visualización en Power Bi:	120
12.	Conclusiones	125
13.	Referencias bibliográficas.....	127

Declaración de originalidad y autonomía

Declaro bajo la gravedad del juramento, que he escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por mi propia cuenta y que, por lo tanto, su contenido es original.

Declaro que he indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

A handwritten signature in black ink, appearing to read 'NATALIA RAMOS'.

Natalia Ramos Barriga

Firmado en Bogotá, D.C. el 20 de junio de 2023

Declaración de exoneración de responsabilidad

Declaro que la responsabilidad intelectual del presente trabajo es exclusivamente de su autor. La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.

A handwritten signature in black ink, reading "NATALIA RAMOS". The letters are cursive and slightly slanted.

Natalia Ramos Barriga

Firmado en Bogotá, D.C. el 20 de junio de 2023

Lista de Ilustraciones

Ilustración 1 Fases del modelo CRISP-DM.....	18
Ilustración 2 Cronograma bajo la metodología SCRUM y el modelo CRISP-DM.....	21
Ilustración 3 Data inicial	49
Ilustración 4 Resultado de la limpieza de datos.....	49
Ilustración 5 Resultado de la transformación de los datos.....	49
Ilustración 6 Análisis Exploratorio de Datos	64
Ilustración 7 Resultados Estadísticos.....	66
Ilustración 8 Círculo de Correlaciones	68
Ilustración 9 Gráfico de barras: Conteo de valores para variables categóricas	70
Ilustración 10 Características y estructura del conjunto de datos transformado.....	78
Ilustración 11 Número óptimo de clusters K-Medoids.....	81
Ilustración 12 Método del Codo Agrupamiento jerárquico	82
Ilustración 13 Método del Codo – Fuzzy C-means	83
Ilustración 14 Resultado segmentación K-Medoids	85
Ilustración 15 Resultado segmentación Agrupamiento Jerárquico.....	86
Ilustración 16 Resultado segmentación Fuzzy C-Means	88
Ilustración 17 Gráfica de distribución de clusters (Fuzzy C-means).....	90
Ilustración 18 Resultado de la Evaluación.....	99
Ilustración 19 Resultado de los outliers obtenidos mediante Local Outlier Factor	104
Ilustración 20 Data para análisis de riesgo	107
Ilustración 21 Análisis de riesgo variable ingresos.....	108
Ilustración 22 Análisis de riesgo variable aporte	108
Ilustración 23 Análisis de riesgo variable promedio.....	109
Ilustración 24 Riesgo alto	111
Ilustración 25 Riesgo medio	112
Ilustración 26 Riesgo bajo	113
Ilustración 27 Segmentación de los factores de riesgo	117
Ilustración 28 Módulo consulta de parametrización	121
Ilustración 29 Escala de riesgo para países de segmentación	122
Ilustración 30 Ponderación de riesgo para canales	122
Ilustración 31 Módulo edición de parametrizaciones	123
Ilustración 32 Módulo proceso de segmentación.....	123
Ilustración 33 Segmentación – Dashborad en Power Bi.....	124

Lista de Tablas

Tabla 1 Variables por factor de riesgo.....	29
Tabla 2 Riesgo país de residencia.....	31
Tabla 3 Riesgo departamento.....	33
Tabla 4 Riesgo actividad económica	35
Tabla 5 Riesgo canal y producto.....	36
Tabla 6 Riesgo PEP	36
Tabla 7 Riesgo reputación	37
Tabla 8 Riesgo Aportes PN y PJ.....	37
Tabla 9 Riesgo ingresos y patrimonio PN y PJ.....	38

Abreviaturas

EOSF: Estatuto Orgánico de Sistema Financiero

GAFI: Grupo de Acción Financiera Internacional

GT: Accionista de R4G

LA/FT: Lavado de Activos y Financiación del Terrorismo

PEP: Persona Expuesta Políticamente

R4: Casa matriz

R4G: Entidad donde se aplicó el proyecto

SARLAFT: Sistema de Administración de Riesgo de Lavado de Activos y Financiación del Terrorismo

SFC: Superintendencia Financiera de Colombia

SIFI: Sistema de Información Fiduciario

Glosario

Las siguientes definiciones, son extraídas de la Circular Básica Jurídica de la Superintendencia Financiera de Colombia:

Factores de riesgo: “Son los agentes generadores del riesgo de LA/FT. Para efectos del SARLAFT las entidades vigiladas deben tener en cuenta como mínimo los siguientes: Clientes y usuarios, Productos, Canales de distribución y Jurisdicciones” (Superintendencia Financiera de Colombia, 2014).

Financiación del terrorismo: “Es el conjunto de actividades encaminadas a canalizar recursos lícitos o ilícitos para promover, sufragar o patrocinar individuos, grupos o actividades terroristas” (Superintendencia Financiera de Colombia, 2014).

Lavado de activos: “Es el conjunto de actividades encaminadas a ocultar el origen ilícito o a dar apariencia de legalidad a recursos obtenidos producto de la ejecución de actividades ilícitas” (Superintendencia Financiera de Colombia, 2014).

Riesgos asociados al LA/FT: “Son los riesgos a través de los cuales se materializa el riesgo de LA/FT. Estos son: reputacional, legal operativo y de contagio” (Superintendencia Financiera de Colombia, 2014).

Segmentación: “Es el proceso por medio del cual se lleva a cabo la separación de elementos en grupos homogéneos al interior de ellos y heterogéneos entre ellos” (Superintendencia Financiera de Colombia, 2014).

Resumen Ejecutivo

Modelo de Segmentación para SARLAFT en R4G

De acuerdo con las observaciones de los diferentes entes de control, la segmentación de los factores de riesgo de Lavado de Activos y Financiación del Terrorismo no contribuye a la correcta categorización de clientes para controlar el riesgo LA/FT, por lo tanto, el presente proyecto empresarial busca proporcionar una herramienta de Business Analytics que permita implementar un modelo de segmentación para el área de riesgos de R4G, de acuerdo con los parámetros establecidos en la normatividad vigente, con el propósito de identificar operaciones inusuales y fortalecer el Sistema de Administración de Riesgos de Lavado de Activos y Financiación del Terrorismo.

Palabras clave:

SARLAFT, Segmentación, Prevención, LA/FT.

Abstract

Segmentation Model for SARLAFT in R4G

Under the observations previously made by the different regulatory agencies, the segmentation of risk factors for money laundering and the financing of terrorism does not contribute to an adequate categorization of clients to be able to control AML/LA risks. As a consequence, the business case herein detailed seeks to provide a Business Analytics tool that facilitates a segmentation model for the Risk Department of R4G, which aligns with the parameters established in the current regulatory framework of the company, and intends to identifying unusual operations and fortify the system for the management of risks associated with money laundering and the financing of terrorism.

Keywords:

SARLAFT, Segmentation, Prevention, AML/LT.

1. Introducción

R4G, es una entidad fiduciaria del sector financiero colombiano, vigilada por la Superintendencia Financiera de Colombia, sus principales líneas de negocio son:

Negocios Fiduciarios: Durante el año 2022, de los nuevos negocios que se suscribieron, dos tercios fueron bajo la tipología de administración y pagos y un tercio de garantía y fuente de pago.

Institucional: Consolidación de las operaciones de la mesa de renta fija para operar en posición propia nacional e internacional.

Gestión Patrimonial: Core del negocio, con el que se busca tener una posición de liderazgo en el mercado local con la asesoría integral, mediante un adecuado sistema de servicio al cliente con soluciones a la medida donde se presentan alternativas de inversión como FIC, FVP, FCP, fondos propios y fondos de terceros.

R4G consciente de la importancia que reviste su labor en el manejo de recursos financieros dentro del sistema financiero colombiano e internacional, de conformidad con las leyes colombianas, ha adoptado acciones, mecanismos y controles que estén a su alcance para evitar que las operaciones propias de su objeto social, puedan ser utilizadas como instrumento para el ocultamiento, manejo, inversión o aprovechamiento, en cualquier forma, de dinero u otros bienes provenientes de actividades delictivas, o para dar apariencia de legalidad a éstas o a las transacciones y fondos vinculados con la ilegalidad.

De acuerdo con la Circular Básica Jurídica Parte I - Instrucciones generales aplicables a las entidades vigiladas, Título IV, Deberes y responsabilidades, Capítulo. IV- Instrucciones relativas a la administración del riesgo de lavado de activos y de la financiación del terrorismo –

SARLAFT, la Superintendencia Financiera de Colombia (2014) exige en los siguientes numerales:

(i) 4.1.1.1. Establecer metodologías de la segmentación de los factores de riesgo y segmentar los factores de riesgo conforme a dichas metodologías, (ii) 4.1.1.2. Establecer metodologías para la identificación del riesgo de LA/FT y sus riesgos asociados respecto de cada uno de los factores de riesgo segmentados, teniendo en cuenta el contexto interno y externo de la entidad vigilada, y (iii) 4.1.1.3. Con base en las metodologías establecidas en desarrollo del numeral anterior, identificar las formas a través de las cuales se puede presentar el riesgo de LA/FT, atendiendo las variables consideradas para cada uno de los factores de riesgo.

Si bien, en la actualidad R4G aplica un modelo de segmentación de clientes, los diferentes entes de control tales como la SFC, la casa matriz y la auditoría interna, han solicitado implementar planes de acción para mejorar el modelo de segmentación asociando los diferentes factores de riesgo normativos.

Por lo tanto, el objetivo del presente proyecto empresarial es proponer un modelo de segmentación que cubra las observaciones de los entes de control, cumpliendo con la normatividad vigente, para así aportar importantes beneficios en términos de cumplimiento normativo, reducción de riesgo y eficiencia en la gestión del SARLAFT.

En las siguientes secciones se describirán los objetivos específicos que se esperan alcanzar, identificando el alcance del proyecto y desarrollando la metodología CRISP-DM, finalmente se expondrán las conclusiones obtenidas y las recomendaciones de implementación del modelo.

2. Objetivos

2.1. Objetivo General

Proponer un modelo de segmentación para SARLAFT en R4G.

2.2. Objetivos Específicos

- Seleccionar las variables para cada uno de los factores de riesgo, de acuerdo con el contexto interno y externo de R4G.
- Seleccionar el modelo de segmentación que cumpla con los objetivos específicos del negocio de R4G.
- Asignar el valor de riesgo y la ponderación a las variables seleccionadas para el modelo de segmentación.
- Calibrar el modelo conforme lo establecido en la Circular Básica Jurídica de la SFC.

3. Alcance

Mejora del sistema de control y monitoreo, para prevenir y minimizar la posibilidad de que se introduzcan a R4G, recursos provenientes de lavado de activos y financiación del terrorismo, mediante un modelo de segmentación de los factores de riesgo aplicado al Sistema de Administración de Riesgos de Lavado de Activos y Financiación del Terrorismo, para promover la innovación mediante el desarrollo de tecnologías que hagan más robusta la administración de este riesgo en R4G, conforme a los estándares internacionales de la materia y de acuerdo con lo establecido por la Superintendencia Financiera de Colombia y a las recomendaciones de los diferentes entes de control.

4. Metodología

Para el desarrollo del proyecto empresarial, se optó por utilizar el modelo CRISP-DM (Cross-Industry Standard Process for Data Mining) junto con la metodología SCRUM, ya que CRISP-DM proporciona un enfoque iterativo y estructurado, por lo tanto, cubre todas las etapas del ciclo de vida del proyecto, desde la comprensión del negocio hasta el despliegue de soluciones. Por otro lado, SCRUM brinda un marco de trabajo colaborativo y flexible que ayuda a las áreas de R4G a gestionar este proyecto de manera eficiente, adaptándose a los cambios y a realizar entregables de calidad. Al combinar ambos enfoques, se busca beneficiarse de la estructura y la flexibilidad proporcionadas por SCRUM, al tiempo que se sigue el marco de trabajo establecido por CRISP-DM para la minería de datos, por lo tanto, es la opción ideal para el proyecto empresarial, debido a su flexibilidad, estructura y reconocimiento a nivel global.

De acuerdo con la explicación de IBM (2021), “el modelo CRISP-DM contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases” (ver ilustración 1), adicionalmente resalta que “la secuencia de las fases no es estricta, ya que la mayoría de los proyectos avanzan y retroceden entre fases si es necesario”.

El modelo CRISP-DM de IBM SPSS Modeler se divide en seis fases iterativas:

Entendimiento del negocio: comprender el problema del negocio y los objetivos del proyecto.

Comprensión de los datos: analizar los datos disponibles identificando los patrones y relaciones.

Preparación de los datos: limpiar y estructurar los datos para ser utilizados en el modelo.

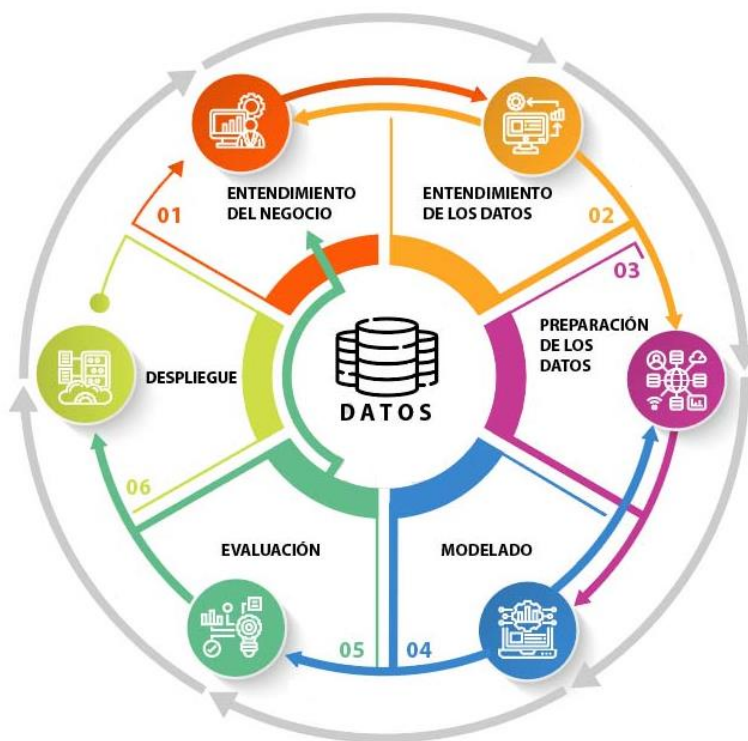
Modelado: desarrollar el modelo de análisis de datos y realizar una evaluación para determinar su precisión.

Evaluación: estimar la efectividad del modelo y determinar si cumple con los objetivos del proyecto.

Despliegue: implementar el modelo y realizar seguimiento continuo para garantizar su efectividad.

Ilustración 1

Fases del modelo CRISP-DM



Fuente: Tomado de Esquema del ciclo CRISP-DM estándar (Haya, 2023).

Por otro lado, la metodología SCRUM brinda beneficios clave para la ejecución de los proyectos, algunos son:

Flexibilidad y adaptabilidad: Permite a los equipos responder de manera ágil a los cambios en los requisitos y prioridades del proyecto. A través de sprints (iteraciones) cortos y bien definidos, se pueden realizar ajustes y mejoras continuas en el producto.

Mayor satisfacción del cliente: Al entregar incrementos de funcionalidad de forma regular, se obtiene una retroalimentación temprana y se pueden realizar ajustes para garantizar la satisfacción del cliente.

Mayor productividad: Promueve la colaboración estrecha y constante entre los miembros del equipo. A través de reuniones diarias de seguimiento (Daily Stand-ups), reuniones de planificación, revisión y retrospectiva, se fomenta la comunicación efectiva, se eliminan obstáculos y se optimiza la productividad del equipo.

Control de calidad: Al realizar pruebas y revisiones constantes durante cada sprint, SCRUM garantiza un enfoque riguroso en cuanto a la calidad del producto. Los problemas y errores se identifican rápidamente, lo que permite corregirlos de manera oportuna y mantener altos estándares de calidad.

Mejora continua: La metodología SCRUM se basa en la retrospectiva, una reunión periódica donde el equipo reflexiona sobre lo que funcionó bien y lo que se puede mejorar. Esto fomenta el aprendizaje y la evolución constante del proceso y las prácticas de trabajo.

5. Cronograma

El proyecto se realizó implementando la metodología Scrum mediante los pilares empíricos de transparencia, inspección y adaptación; y el modelo Cross-Industry Standard Process for Data Mining (CRISP – DM), de esta manera se mitigan los riesgos, se progresa en los objetivos y se ajustan las desviaciones del modelo, generando de esta manera beneficios económicos, estratégicos y competitivos para la Entidad.

Por lo tanto, el cronograma que se propuso estuvo dividido en las siguientes fases: (i) entendimiento del negocio, (ii) entendimiento de los datos, (iii) preparación de los datos, (iv) modelado, (v) evaluación y (vi) despliegue, implementando en cada rubro los principios Scrum: desarrollo iterativo, control del proceso empírico, autoorganización, colaboración, priorización basada en valor, tiempo delimitado para cada sprint y actualizaciones del backlog priorizado con los cambios aprobados por R4G.

Es importante resaltar que, en cada rubro propuesto, se implementaron los temas relevantes a trabajar, indicando en cada uno de ellos las materias aplicadas en la Maestría de Business Analytics.

Ilustración 2

Cronograma bajo la metodología SCRUM y el modelo CRISP-DM

Entendimiento del Negocio

<input type="checkbox"/>	Elemento		Persona	Estado	Priority	Cronograma	Comentarios Materias Aplicadas
<input type="checkbox"/>	Reunión inicial con R4G para identificar las necesidades de la Entidad			Listo	Critical	nov. 3 - 15	Analytics Product Management
<input type="checkbox"/>	Reuniones quincenales con el director de Riesgos de R4G para abarcar: 3			En Proceso	High	nov. 15 - dic. 13	Analytics Product Management
<input type="checkbox"/>	Subelemento	Owner	Status	Date	Priority	Comentarios Materias Aplicadas	+
<input type="checkbox"/>	Entendimiento del Negocio			Done	nov. 15	Low	Inteligencia de negocios
<input type="checkbox"/>	Identificación de la problemática empresarial			Done	nov. 29	Critical	Inteligencia de negocios
<input type="checkbox"/>	Alcance del proyecto			Working on it	dic. 6	Medium	Inteligencia de negocios
<input type="checkbox"/>	+ Agregar Subelemento						
<input type="checkbox"/>	Lectura de la normativa SARLAFT			Listo	Low	nov. 4 - 8	Ética, Seguridad y Aspectos Legales de los datos
<input type="checkbox"/>	Identificación de la arquitectura TO BE del negocio			Listo	High	nov. 26 - dic. 3	Gobernanza de Sistemas de Información
<input type="checkbox"/>	Identificación de la arquitectura AS IS del negocio			Listo	Low	nov. 26 - dic. 3	Gobernanza de Sistemas de Información
<input type="checkbox"/>	Presentación a R4G de la arquitectura TO BE para aprobación			En Proceso	Medium	dic. 12 - 16	Gobernanza de Sistemas de Información
<input type="checkbox"/>	Selección de fases del proyecto para aprobación de R4G			En Proceso	Medium	dic. 12 - 16	Gestión ágil del proyecto
<input type="checkbox"/>	Selección fuente de información para el proyecto			En Proceso	Low	dic. 27 - 30	Estrategias de búsqueda, uso ético de la inform...
<input type="checkbox"/>	+ Agregar Elemento						
						nov. 3 - dic. 30	

▼ Entendimiento de los datos

<input type="checkbox"/>	Elemento		Persona	Estado	Priority	Cronograma	Comentarios Materias Aplicadas
<input type="checkbox"/>	Verificación de los campos en SIFI y los formularios de R4G			Estancado	Critical	nov. 21 - dic. 23	NF no diligencia todos los campos en SIFI
<input type="checkbox"/>	Validación de la data de clientes y transacciones históricas con TI			Listo	High	nov. 21 - 25	Inteligencia de negocios
<input type="checkbox"/>	Recolección y validación de la información suministrada por TI			Listo	Medium	nov. 28 - dic. 2	Gobernanza de Sistemas de Información
<input type="checkbox"/>	Proceso de vinculación persona natural y jurídica			Listo	Low	nov. 15 - 18	Inteligencia de negocios
<input type="checkbox"/>	▼ Ajustes de la data de transacciones ¹			En ajustes por R4G	Critical	dic. 2 - 23	La data que se requiere sólo es de depósitos

<input type="checkbox"/>	Subelemento		Owner	Status	Date	Priority	Comentarios Materias Aplicadas	+
<input type="checkbox"/>	Validación de Data de SIFI a Detectart			Working on it	dic. 23	Critical	Validación a cargo de TI de R4G	
<input type="checkbox"/>	+ Agregar Subelemento							

<input type="checkbox"/>	Creación de la Matriz de Riesgos			En Proceso	Medium	dic. 7 - 14	Ética, seguridad y análisis de riesgos
<input type="checkbox"/>	+ Agregar Elemento						

						nov. 15 - dic. 23	
--	--	--	--	--	--	-------------------	--

▼ Preparación de los datos ^{2 Elementos}

<input type="checkbox"/>	Elemento		Persona	Estado	Priority	Cronograma	Comentarios Materias Aplicadas
<input type="checkbox"/>	Análisis de variables			En Proceso	Medium	dic. 1 - 17	
<input type="checkbox"/>	Limpeza base de datos			En Proceso	Critical	dic. 1 - 30	
<input type="checkbox"/>	+ Agregar Elemento						

						dic. 1 - 30	
--	--	--	--	--	--	-------------	--

▼ Modelado

<input type="checkbox"/>	Elemento		Persona	Estado	Priority	Cronograma	Comentarios Materias Aplicadas
<input type="checkbox"/>	Aplicación modelos jerárquicos			Estancado	Medium	ene. 10, '23 - ene. 21, '23	Analítica predictiva, Modelos estadísticos
<input type="checkbox"/>	Aplicación modelos de participación			Estancado	Medium	ene. 24, '23 - feb. 3, '23	Analítica predictiva, Modelos estadísticos
<input type="checkbox"/>	Ajuste de modelos			Estancado	Low	feb. 7, '23 - feb. 18, '23	Analítica predictiva, Modelos estadísticos
<input type="checkbox"/>	Comparación de los modelos aplicados			Estancado	Medium	feb. 21, '23 - mar. 3, '23	Analytics Product Management
<input type="checkbox"/>	Selección del modelo			Estancado	Critical	mar. 7, '23 - mar. 24, '23	Analítica predictiva, Modelos estadísticos
<input type="checkbox"/>	+ Agregar Elemento						

						ene. 10, '23 - mar. 24, '23	
--	--	--	--	--	--	-----------------------------	--

✓ Evaluación 2 Elementos

<input type="checkbox"/>	Elemento		Persona	Estado	Priority	Cronograma	Comentarios Materias Aplicadas
<input type="checkbox"/>	Entrenamiento del modelo			Estancado	Medium	abr. 4, '23 - abr. 22, '23	
<input type="checkbox"/>	Ajuste a las medidas de comparación aplicadas			Estancado	Low	abr. 25, '23 - may. 6, '23	
<input type="checkbox"/>	+ Agregar Elemento						
						abr. 4, '23 - may. 6, '23	

✓ Despliegue

<input type="checkbox"/>	Elemento		Persona	Estado	Priority	Cronograma	Comentarios Materias Aplicadas
<input type="checkbox"/>	Ejecución del modelo en ambiente de pruebas			Estancado	High	may. 8, '23 - may. 12, '23	
<input type="checkbox"/>	Implementación del modelo			Estancado	Medium	may. 15, '23 - may. 19, '23	
<input type="checkbox"/>	Visualización de resultados en Power BI			Estancado	Low	may. 22, '23 - may. 26, '23	Visualización de datos
<input type="checkbox"/>	Calibración del modelo			Estancado	Critical	may. 29, '23 - jun. 2, '23	
<input type="checkbox"/>	+ Agregar Elemento						
						may. 8, '23 - jun. 2, '23	

Fuente: Elaboración propia

6. Entendimiento del Negocio

6.1.Contexto Sistema Financiero

Después de la pandemia sufrida en el año 2020, el sistema financiero colombiano registró una contracción que se revirtió gracias a la consolidación de la actividad crediticia de Colombia durante el último trimestre de 2021 y los primeros meses de 2022, Así pues, de acuerdo con el informe de estabilidad financiera del Departamento de Estabilidad Financiera del Banco de la República (2022) se concluye que:

- Los establecimientos aceleraron los préstamos a empresas y hogares, manteniendo indicadores robustos de solvencia y liquidez.
- Las instituciones financieras no bancarias registraron un menor crecimiento y una caída en su rentabilidad de sus portafolios en cuenta propia y de terceros.
- La participación de los inversionistas extranjeros en el mercado local de títulos de deuda pública se ha mantenido estable.
- A pesar de que las entidades financieras del sistema mantienen un alto grado de solidez, hay vulnerabilidades para la estabilidad financiera toda vez que ha persistido la exposición de la economía a cambios súbitos en las condiciones financieras globales.

Actualmente, las entidades financieras colombianas desarrollan su actividad en un entorno difícil y complejo debido al impacto de un amplio conjunto de factores como:

- Fuerte cambio tecnológico: donde se introduce una clara alteración de los esquemas con la introducción del Blockchain, Machine Learning o Inteligencia artificial.

- Nuevos competidores: El sector está viviendo la creciente competencia de nuevos jugadores, como Fintech y BigTech, que son competidores muy disruptivos y a los que las exigencias regulatorias les afectan en menor medida. Esto crea situaciones asimétricas que los bancos reclaman igualar.

- Amenaza SARLAFT:

Según la Superintendencia Financiera de Colombia (2014), el lavado de activos y la financiación del terrorismo representan una gran amenaza para la estabilidad del sistema financiero y la integridad de los mercados por su carácter global y las redes utilizadas para el manejo de tales recursos. Tal circunstancia destaca la importancia y urgencia de combatirlos, resultando esencial el papel que para tal propósito deben desempeñar las entidades vigiladas por la SFC y el supervisor financiero.

6.2. R4G

R4G es una sociedad fiduciaria, constituida mediante Escritura Pública 4462 de 2016 de la notaría 13 del Círculo de Bogotá, cuyo funcionamiento fue autorizado mediante Resolución 147 del primero de febrero de 2017 de la Superintendencia Financiera de Colombia.

R4G cuenta con la inversión de R4 y GT. R4 es el primer banco español independiente, especializado en gestión patrimonial, mercados de capitales y servicios de inversión, con más de 30 años de experiencia a nivel internacional. GT, es la empresa pionera y líder en planes prepagados educativos.

El respaldo que da la calificación y trayectoria de los principales accionistas de R4G, le permitirá ascender tanto en operaciones como en líneas de negocios de conformidad con la orientación estratégica que se tiene planeada. La experiencia de sus asociados y el acatamiento

de la normatividad vigente en Colombia sobre el SARLAFT fortalecerá día a día la confianza depositada por sus clientes.

6.3. Gestión de la Entidad en SARLAFT

Da cumplimiento a la legislación colombiana de acuerdo con el Estatuto Orgánico del Sistema Financiero EOSF, en conformidad con los estándares internacionales sobre la materia, especialmente los proferidos por el GAFI.

Adicionalmente, R4G tiene el objetivo de disminuir y/o prevenir la probabilidad de que esta sea utilizada para introducir al sistema financiero y a la economía en general, recursos provenientes de actividades relacionadas con los delitos fuente del lavado de activos y/o la financiación del terrorismo; implementando las etapas y elementos del SARLAFT, así como los mecanismos e instrumentos que se deben aplicar para su adecuada evaluación, análisis y reporte de operaciones a las entidades correspondientes.

Es importante destacar que R4G realiza una debida diligencia en la administración del riesgo para la prevención del lavado de activos y la financiación del terrorismo, por lo que la entidad hará lo que esté a su alcance para prevenir, detectar y reportar oportunamente la ocurrencia de los eventos de riesgo identificados. De igual manera, aquel factor de riesgo o evento en particular que represente una vulnerabilidad por fuera de la aceptada por la fiduciaria (“Alto o “Extremo”) requerirá la implementación de planes de acción con el objeto de mitigar el impacto que sobre la entidad puedan generar la materialización de un evento de riesgo de LA/FT.

6.3.1. Entendimiento y Preparación de los Datos

Durante la fase 2, se realizó la exploración de datos utilizando información extraída de los formularios de vinculación y actualización de los clientes de R4G. Se trabajó con un total de

132 clientes con corte al primer trimestre del 2023, considerando tanto personas naturales como jurídicas.

Los datos extraídos del proceso de vinculación son fundamentales para lograr una segmentación efectiva de los factores de riesgo de Lavado de Activos y Financiación del Terrorismo. Estos datos proporcionan información clave que permite evaluar el nivel de riesgo asociado a cada cliente y categorizarlos adecuadamente en términos de control y prevención del riesgo LA/FT, por lo tanto, se realizó un análisis descriptivo para comprender las características de las variables relevantes para la segmentación de los factores de riesgo LA/FT, de acuerdo con la Guía de CRISP-DM de IBM SPSS Modeler (2021), donde indican algunas características clave para describir los datos:

Cantidad de datos: En la mayoría de las técnicas de modelado, los tamaños de datos tienen un equilibrio relacionado. Los grandes conjuntos de datos pueden producir modelos más precisos, pero también pueden aumentar el tiempo de procesamiento.

Tipos de valores: Los datos pueden incluir una variedad de formatos, como numérico, categórico (cadena) o Booleano (verdadero/falso).

Esquemas de codificación: Con frecuencia, los valores de la base de datos son representaciones de características como género o tipo de producto.

A continuación, se presenta el informe de descripción de datos y la preparación que se realizó a cada uno de ellos:

Origen de los datos:

Datos existentes: Los datos de los clientes existentes reposan en el aplicativo SIFI de R4G, tales como datos transaccionales, información sobre la actividad económica y financiera, residencia, entre otros.

Datos adquiridos: Adicionalmente se incorporó una herramienta de noticias LAFT del proveedor de Detectart, para consultar al cliente y validar si este posee noticias adversas referentes a lavado de activos y financiación de terrorismo y así categorizar si tiene una buena o mala reputación.

Datos adicionales: Se realizó actualización de datos a todos los clientes para tener la data con información reciente.

Descripción de los datos:

El conjunto de datos está compuesto por todos los clientes de R4G hasta el 31 de marzo de 2023, con un total de 132 clientes y 13 variables distintas que proporcionan información relevante sobre los clientes y su comportamiento. A continuación, se describen las principales características de las variables utilizadas:

6.4.Criterios Establecidos para las Variables Seleccionadas

- Factores de riesgo definidos a través de variables cualitativas y cuantitativas.
- Selección y descripción de variables de acuerdo con el contexto interno y externo de la Entidad, (el contexto externo fue dado por la casa matriz R4), asignando un score a cada una de las variables de acuerdo con el riesgo en LA/FT, dicho score no se utilizará en el modelo de segmentación, sino en el análisis descriptivo una vez ejecutado el modelo.
- Ponderación de las variables por criterio experto.

6.5.Factores de Riesgo y Variables Definidas

A continuación, se listan los factores de riesgos normativos y las variables asociadas a cada uno de ellos, para la segmentación de persona natural y persona jurídica, dando una breve descripción de cada una de estas:

Tabla 1*Variables por factor de riesgo*

Factores de Riesgos	Variables	Descripción General
Clientes	1. Actividad económica	Actividad económica (Código CIIU) Clasificación industrial internacional uniforme de todas las actividades económicas
	2. Ingresos	Ingresos mensuales reportados por el cliente en el proceso de vinculación o actualización de datos.
	3. Aporte	Para clientes nuevos será el valor inicial de aporte y para clientes antiguos se tomará el promedio histórico del último año de los depósitos que han realizado.
	4. Patrimonio	Patrimonio reportado por el cliente en el proceso de vinculación o actualización de datos.
	5. PEP	El concepto de Personas Expuestas Políticamente (PEP) comprende a las personas expuestas políticamente y a las personas expuestas políticamente extranjeras definidas en el Decreto 1081 de 2015, y demás normas que lo modifiquen, complementen, sustituyan o adicionen.
	6. Reputación	Búsqueda en noticias y/o listas cautelares.
Producto	7. FIC	Fondo de inversión colectiva.
	8. FVP	Fondo voluntario de pensiones.
	9. FCP	Fondo de capital privado.
	10. NF	Negocios fiduciarios (incluidos los patrimonios autónomos).
Canales de distribución	11. Canal	Presencial – Virtual – Referido.
Jurisdicción	12. Residencia	La ponderación de riesgo de esta variable está determinada por los lineamientos de la casa matriz y por la lista de jurisdicciones de alto riesgo del GAFI. Si el país de operación / residencia es Colombia se tiene en cuenta el departamento de operación / residencia.
	13. Departamento	Esta variable solo aplica a residentes o empresas con operaciones en Colombia. Por lo tanto, para aquellos que sean extranjeros, se considerará un valor de 0 en esta variable.

Fuente: Elaboración propia

Las variables cualitativas para el modelo son:

- Actividad económica
- PEP
- FIC
- FVP
- FCP
- NF
- Canal
- Residencia
- Reputación
- Departamento

Las variables cuantitativas para el modelo son:

- Ingresos
- Aportes
- Patrimonio

Las variables utilizadas en el proceso de segmentación se obtienen de la base de datos SIFI de R4G. Estas variables son capturadas a través de los formularios de conocimiento del cliente durante el proceso de vinculación y actualización de clientes, tanto para personas naturales como personas jurídicas. Sin embargo, es importante destacar que las variables de reputación y PEP son excepciones y se describen en detalle más adelante:

6.6. Descripción de las Variables de Segmentación

6.6.1. *Residencia*

Las jurisdicciones representan divisiones a nivel geográfico de una zona o región, o de manera más específica, el punto geográfico relacionado con la constitución, operación o transacción realizada por los clientes; en este sentido, se estableció como objetivo para su segmentación identificar el nivel de vulnerabilidad frente a algunos de los delitos fuente de lavado de activos y financiación del terrorismo en los diferentes departamentos de nuestro país, además de determinar jurisdicciones de alto riesgo (como por ejemplo países no cooperantes de GAFI) y de esta manera, realizar una adecuada valoración del riesgo al cual se encuentra expuesto cada uno de ellos.

A continuación, se listan los países con su respectiva calificación de riesgo (score), tomada de la casa matriz R4, se ponderan todos los países de la base de datos en función de los listados de países de riesgo UE, GAFI, paraísos fiscales España y UE, sanciones UE, países alto porcentaje de corrupción y narcotráfico:

Tabla 2

Riesgo país de residencia

Lista	Score	Países
Alto Riesgo	1	Afganistán; Albania; Anguilla; Antigua Y Barbuda; Antillas Holandesas; Aruba; Bahamas; Bahréin; Barbados; Belarus; Belice; Bermudas; Bhután; Bolivia; Bosnia Y Herzegovina; Botswana; Brunei; Burundi; Cabo Verde; Camboya; Camerún; Chad; China; Congo; Corea Del Norte ; Costa De Marfil; Costa Rica; Djibouti; Dominica; Egipto; El Salvador; Emiratos Árabes Unidos; Eritrea; Etiopia; Fiji; Filipinas; Georgia; Ghana; Gibraltar; Granada; Groenlandia; Guadalupe; Guam; Guatemala; Guayana Francesa; Guernesey; Guinea; Guinea Ecuatorial; Guinea-Bissau; Guyana; Haití; Hong Kong; Indonesia; Irán; Iraq; Isla Norfolk; Islas Åland; Islas Caimán; Islas Cook; Islas Del Canal; Islas Feroe; Islas Malvinas; Islas Marshall; Islas Pitcairn; Islas Salomón; Islas Turcas Y Caicos; Islas Vírgenes Británicas; Islas Vírgenes De Los Estados Unidos; Jamaica; Jersey; Jordania; Kazajstán; Kiribati; Laos; Líbano; Liberia; Libia; Liechtenstein; Macao; Macedonia; Madagascar; Mali; Martinica; Mauricio; Micronesia; Moldavia; Mónaco; Mongolia; Montserrat; Mozambique; Myanmar; Namibia; Nauru; Nicaragua; Nigeria; Nueva Caledonia; Omán; Pakistán; Palaos; Palestina; Panamá; Paraguay; Polinesia Francesa; Puerto Rico; Rep. Democrática Del Congo; República Centroafricana; Reunión; Rumania; Rusia; Sahara Occidental; Samoa; Samoa Americana; San Bartolomé; San Cristóbal Y Nieves; San Martin (Parte Francesa); San Pedro Y Miquelón ; San Vicente Y Las Granadinas; Santa Helena; Santa Lucia; Santa Sede; Santo Tome y Príncipe; Serbia; Seychelles; Siria; Somalia; Sri Lanka; Sudan; Svalbard Y Jan Mayen; Swazilandia; Timor Oriental; Trinidad Y Tobago; Túnez; Turquía; Ucrania; Uganda; Vanuatu; Venezuela; Wallis Y Fortuna; Yemen; Zimbabwe
Alto Riesgo 2	1	Isla De Man, Islas Marianas Del Norte, Andorra, Burkina Faso, Maldivas, Malta, Marruecos, Senegal, Malasia, Tailandia, Vietnam, Suazilandia, Niue
No existe	1	Desconocido, Sin Especificar
País Miembro GAFI	0.5	Irlanda, Japón, Brasil, India
País Miembro UE 1	0.1	Bélgica, República Checa, Polonia, Grecia
País Miembro UE 2	0.2	Letonia, Lituania, Eslovaquia, Eslovenia, Croacia
País Miembro UE y GAFI	0.3	Canadá, Países Bajos, Reino Unido, Argentina
Países UE, GAFI y restos	0.6	Estonia, Sudáfrica
Países UE, GAFI y restos 2	0.8	Suiza, Honduras, Republica Dominicana, Luxemburgo, Taiwán, Singapur, Chile, Perú

Resto Países	0.9	Argel, Angola, Azerbaiyán, Bangladés, Armenia, Comoros, Mayotte, Cuba, Benín, Ecuador, Gabón, Gambia, Kiribati, Israel, Kenia, Kuwait, Kirguistán, Lesotho, Malawi, Mauritania, Montenegro, Mozambique, Nepal, Níger, Papúa Nueva Guinea, Qatar, Ruanda, San Marino, Arabia Saudita, Sierra Leona, Surinam, Tayikistán, Togo, Tokelau, Tonga, Turkmenistán, Tuvalu, Tanzania, Uruguay, Uzbekistán, Zambia, Chipre, Bulgaria, Corea Del Sur, México.
--------------	-----	---

Fuente: Elaboración propia

6.6.2. Departamento

Tabla 3

Riesgo departamento

Riesgo	Score	Departamento
Riesgo Alto	1	Chocó; Arauca; Meta; Guaviare; Archipiélago De San Andrés; Valle Del Cauca; Putumayo; Boyacá; Norte De Santander; Cauca; Antioquia
Riesgo Medio	0.4	Huila; Casanare; Tolima; Cundinamarca; Guainía
Riesgo Bajo	0.2	Sucre; Vichada; Caquetá; Cesar; Nariño; La Guajira; Quindío; Magdalena; Atlántico; Bolívar; Santander; Risaralda; Caldas; Córdoba; Amazonas; Vaupés
Sin Riesgo	0	Bogotá; Medellín; Cali; Bucaramanga; Barranquilla

Fuente: Elaboración propia

Para determinar el riesgo de cada uno de los departamentos se realizó una estadística de delitos por cada mil habitantes, así:

$$\text{Delitos por cada mil habitantes} = (\text{Número de delitos} / \text{Población}) * 1000$$

Para la variable “delitos”, se tuvo en cuenta el número de delitos a 2021 de extorsión, hurto comercio, hurto entidades financieras, secuestro y terrorismo, con información tomada de la página de la Policía Nacional de Colombia.

Para la población se tomó la información del DANE, teniendo en cuenta la población por municipio.

Se debe considerar que del cálculo se excluyeron Bogotá, Medellín, Cali, Bucaramanga y Barranquilla, asignándoles un valor de 0, al ser estas ciudades capitales y distritos especiales de Colombia.

La segmentación final de las jurisdicciones se basa en tres niveles de vulnerabilidad frente al riesgo de LA/FT: Alta, Media y Baja; por lo anterior se requiere tomar como referencia la segmentación previa que se realice sobre los departamentos para cada una de las variables identificadas y se hallan las posibles combinaciones entre los tres niveles, definiendo un nivel final de la siguiente manera:

6.6.3. *Actividad Económica*

Para esta variable se tomó en cuenta la metodología adaptada en la evaluación nacional del riesgo de lavado de activos y financiación del terrorismo del año 2016.

El procedimiento para la evaluación definido en la metodología consistió en que cada grupo debía asignar un “score” (calificación) a cada una de las variables. Estos indicadores corresponden a la escala de medición de riesgo desarrollada previamente por el Banco Mundial.

A continuación, se muestra el resultado de la evaluación de la vulnerabilidad para cada subsector del sector real (incluidas las actividades y profesiones no financieras designadas - APNFD con el fin de cumplir con las recomendaciones 22, 23 y 28 del GAFI), a esta escala se adicionó el rubro “agropecuario”, el cual no estaba incluido en la metodología, asignándole un score de 1, de acuerdo con la recomendación de la casa matriz:

Tabla 4*Riesgo actividad económica*

Subsector	Score
Juego, suerte y azar: bingos, casinos, loterías, chances, tragamonedas, hípica, etc.	0.56
Inmobiliarias, finca raíz, constructoras	0.72
Minería: extracción, depósito, comercialización, procesamiento y exportación de metales preciosos	0.82
Abogados	0.90
Contadores, revisores fiscales, auditores	0.90
Notarios	0.40
Fútbol y otros deportes	0.68
Puertos, aeropuertos, zonas francas, depósitos, transporte de carga, agencias de aduanas, operadores y usuarios aduaneros, empresas de mensajería, cambistas profesionales	0.46
Empresas de seguridad, empresas de blindaje y empresas de transporte de valores	0.44
Operadores postales oficiales y operadores postales de pago	0.40
Sector salud	0.7
Entidades sin ánimo de lucro	0.80
Otras empresas del sector real	0.65
Sector agropecuario	1.00

Fuente: Adaptado de Resultado de la evaluación de la vulnerabilidad para cada subsector del sector real (APNFD's) (Ministerio de Justicia y del Derecho et al., 2016).

6.6.4. Canal y Producto

Con el objeto de fortalecer la debida diligencia adelantada por la entidad para la prevención del riesgo de LA/FT y consciente de la dinámica de cambio que sufren de forma permanente los negocios, así como los clientes y usuarios de R4G, la entidad considera que la identificación y determinación de diversos factores de riesgo permite crear un perfil de riesgo e identificar tendencias en los mercados objetivo a los que se dirigen principalmente los productos ofrecidos, con el fin de identificar y comprender las particularidades de la operación de sus clientes según los canales disponibles (presencial, virtual y referido), y el entorno de los negocios base para determinar los segmentos del mercado al que son dirigidos los productos y servicios.

La Fiduciaria en su plan estratégico establece los segmentos de mercado en los cuales quiere participar teniendo en cuenta el enfoque a desarrollar y el riesgo que esta puede generar

para la entidad, de acuerdo con estos parámetros, se establece el score para las variables canal y producto:

Tabla 5

Riesgo canal y producto

Canal	Score Consolidado
Presencial	0
Virtual	0.5
Referido	1
Producto	Score Consolidado
Negocios Fiduciarios (NF)	0.6
Fondo de inversión colectiva (FIC)	0.3
Fondo voluntario de pensiones (FVP)	0.3
Fondo de capital privado (FCP)	0.6

Fuente: Elaboración propia

6.6.5. PEP

Con el fin de realizar un monitoreo continuo e intensificado de la relación comercial, con respecto a los clientes que detentan la calidad de PEP, por criterio experto se establece un score de 1, ya que pueden exponer en mayor grado a la entidad al riesgo de LA/FT.

Tabla 6

Riesgo PEP

PEP	Score
Si	1
No	0

Fuente: Elaboración propia

6.6.6. Reputación

Mediante el aplicativo SIDIF se consulta a los clientes de R4G y a los potenciales clientes, validando si estos se encuentran en noticias adversas referentes a LAFT, de acuerdo con el resultado se dará un score de 1 o 0 en la variable reputación:

Tabla 7*Riesgo Reputación*

Reputación	Score
Si	1
No	0

Fuente: Elaboración propia

6.6.7. Aporte

Para los clientes que se encuentran vinculados a R4G, se toma el histórico del último año de las transferencias que han realizado en los diferentes productos, y con la data obtenida se aplica la metodología expuesta en el numeral 6.8.

Para los clientes nuevos, se toma el valor del primer aporte y se compara con las transferencias históricas de los clientes persona natural y persona jurídica, para así definir el score en ambos casos, obteniendo el siguiente resultado:

Tabla 8*Riesgo Aporte*

Rango valor aporte COP	Score
1 – 540.000.000	0
540.000.001 – 1.400.000.000	0.5
>1.400.000.001	1
Sin información	1

Fuente: Elaboración propia

6.6.8. Ingresos y Patrimonio

Al igual que en la variable aporte, para estas dos variables se toma el valor de ingresos y patrimonio que el cliente indique en el proceso de vinculación con sus respectivos soportes financieros, y se compara con la data de los ingresos y patrimonio de los clientes persona natural y jurídica de R4G, con el fin de definir el score:

Tabla 9*Riesgo ingresos y patrimonio*

Rango valor ingresos COP	Score
1 – 2.900.000.000	0
2.900.000.001 – 3.500.000.000	0.5
>3.500.000.001	1
Sin información	1

Rango valor patrimonio COP	Score
1 – 29.000.000.000	0
29.000.000.001 – 54.000.000.000	0.5
>54.000.000.001	1
Sin información	1

Fuente: Elaboración propia

6.7. Ponderación de Riesgo de las Variables

La ponderación de riesgo de las variables seleccionadas para el modelo de segmentación se determina mediante un criterio experto, siguiendo las directrices establecidas por la casa matriz R4. Este enfoque permite realizar un análisis descriptivo detallado de cada segmento y brinda una visión completa de las características y comportamientos distintivos de los clientes en cada grupo.

En esta metodología, se otorga un mayor peso a determinados factores sobre otros, por ejemplo, atendiendo a criterios normativos y expertos, el país y el departamento de residencia, el hecho de ser una persona catalogada como políticamente expuesta (PEP) o la reputación tienen un peso mayor que otros factores, , otorgando así un score que genera una calificación de riesgo alto.

Los cuadros que se muestran a continuación presentan las ponderaciones asignadas a las variables de riesgo, tanto para persona natural como para persona jurídica:

Tabla 10*Ponderación variables persona natural y jurídica*

Variables Persona Natural y Jurídica	Ponderación %
País de nacionalidad	10%
País de residencia/Departamento de residencia	20%
PEP	20%
Reputación	20%
Código Actividad económica	6%
Ingresos mensuales	6%
Patrimonio	6%
Aportes	6%
Producto	3%
Canal	3%
TOTAL	100%

Fuente: Elaboración propia**6.8. Intervalos de Clase para las Variables Cuantitativas**

Siguiendo las directrices de la dirección de riesgos de R4G y tomando como referencia el estudio realizado por Matos Uribe et al. (2020), se llevó a cabo una categorización de las variables cuantitativas mediante la definición de intervalos de clase. Este enfoque permitió establecer diferentes rangos de riesgo, tal como se menciona en los numerales 6.6.7 y 6.6.8, con el objetivo de analizar cada variable en función de su nivel de riesgo SARLAFT después de la ejecución del algoritmo de segmentación.

Este proceso resulta fundamental para comprender y evaluar de manera precisa el riesgo asociado a cada segmento identificado. Al asignar ponderaciones específicas a cada variable de acuerdo con su impacto en el riesgo SARLAFT, se obtiene una visión detallada de los factores que contribuyen a la exposición al riesgo en cada grupo. Estas ponderaciones proporcionan una base sólida para el análisis descriptivo y la identificación de áreas de atención prioritaria en el contexto del cumplimiento normativo y la gestión integral de riesgos.

Esta actividad se realizó mediante los siguientes criterios:

Crear los intervalos: los cuales están separados (acotados) por dos valores extremos, denominados límites.

Identificar la cantidad de intervalos: La selección se realizó por criterio experto y usando la regla de Sturges, cuya expresión es:

$$K = 1 + 3.3 \log n$$

K=Número de intervalos (este debe ser un número entero), n= Número de datos y Log = Logaritmo en base 10.

De acuerdo con Vincenzo Jesús D'Alessio Torres, Licenciado en Matemáticas de la Universidad de los Andes (2021),

El método de Sturges es un criterio utilizado en estadística descriptiva para determinar el número de clases o intervalos que son necesarios para representar gráficamente un conjunto de datos estadísticos, se basa en el número de muestras x que permite encontrar el número de clases y su amplitud de rango, este método fue aprobado por la casa matriz y ejecutado por la dirección de riesgos de R4G.

Calcular el rango de los datos: Se obtiene de la diferencia entre el dato más alto y el más bajo. Se denota con la letra R.

$$R = \text{Dato mayor} - \text{Dato menor}$$

Obtener la amplitud de cada intervalo: dividiendo el rango por el número de intervalos. Se denota con la letra A de esta manera:

$$Ac = (R/K)$$

Construcción de los intervalos: Los intervalos deben ser excluyentes y exhaustivos, lo que significa que cada valor de datos debe pertenecer a un solo intervalo.

Histograma de frecuencias: Utilizando la función de análisis de datos en Excel u otra herramienta, se puede generar el histograma para cada una de las variables, para visualizar la distribución de los datos y detectar los puntos de ruptura naturales.

Categorizar nivel de riesgo: De acuerdo con los puntos de ruptura, se establecen tres niveles de riesgo para cada una de las variables escogidas en el modelo de segmentación, siendo estas: bajo, medio y alto riesgo.

Para realizar la cuantificación de riesgo, se asigna un valor a cada nivel:

Nivel de riesgo bajo = 0

Nivel de riesgo medio = 0.5

Nivel de riesgo alto = 1

Calificación de riesgo: Se realizó tomando en consideración el valor de cada una de las variables y su respectiva ponderación. Mediante la suma ponderada de estos valores, se obtiene una calificación de riesgo que resulta ser un indicador significativo para el análisis de cada variable, además de estar directamente relacionada con los factores de riesgo asociados al SARLAFT. Esta calificación de riesgo es fundamental para el posterior análisis una vez ejecutada la segmentación, ya que nos permite identificar y comprender de manera más precisa los niveles de riesgo presentes en cada segmento

6.9. Preparación de los Datos

En la Fase 3 del modelo CRISP-DM, se enfrentó el desafío de contar con datos desactualizados, lo cual implicó la necesidad de realizar un proceso de actualización con los clientes. Esta etapa adicional se centró en la obtención de información actualizada y confiable para garantizar la validez y relevancia del modelo de segmentación.

Para abordar este desafío, se diseñó un plan de actualización de datos que involucró la interacción directa con los clientes de R4G. Se estableció un proceso de comunicación y recolección de información actualizada, en el cual se solicitó a los clientes proporcionar los datos más recientes relacionados con su actividad económica y financiera.

Posteriormente, se procedió a la incorporación de los datos actualizados al conjunto de datos existente y se aplicaron las técnicas de limpieza, transformación y construcción de variables, este proceso aseguró que los datos utilizados en el modelo de segmentación fueran representativos de la realidad actual de los clientes y reflejaran adecuadamente los factores de riesgo LA/FT.

Es importante destacar que la actualización de datos fue un proceso continuo y recurrente, dado que la información financiera y de perfil de los clientes puede cambiar con el tiempo, se estableció un mecanismo para mantener los datos actualizados de manera regular, mediante la implementación de sistemas automatizados de actualización y a través de la interacción directa con los clientes y comerciales en momentos estratégicos.

A continuación, se presentan detalladamente las tareas realizadas en esta fase:

Selección del lenguaje de programación y la herramienta para realizar el proyecto empresarial:

Actualmente existen numerosos lenguajes de programación utilizados en diversos ámbitos tales como Python, Java, JavaScript y R. A continuación, una breve definición de algunos de los lenguajes de programación más populares:

Python: “Lenguaje de programación interpretado, de alto nivel y de propósito general. Es conocido por su sintaxis legible y su enfoque en la simplicidad y la facilidad de uso” (Python, 2023).

Java: “Lenguaje de programación de propósito general y orientado a objetos, conocido para el desarrollo de aplicaciones empresariales, sistemas embebidos y desarrollo de Android” (Java, 2023).

JavaScript: Lenguaje de programación interpretado y orientado a objetos utilizado para el desarrollo web, permite la interacción dinámica con los elementos de una página web.

R: El lenguaje de programación R se utiliza ampliamente para realizar análisis de datos, manipulación de datos, creación de gráficos y análisis estadístico.

Se seleccionó Python como lenguaje de programación y Google Colab como herramienta debido a las siguientes razones y en comparación con otros lenguajes de programación y herramientas alternativas:

Python es un lenguaje de programación ampliamente utilizado en el ámbito del análisis de datos y el aprendizaje automático, además tiene una sintaxis clara y legible, lo que facilita la comprensión del código, cuenta con una amplia variedad de bibliotecas y frameworks especializados en análisis de datos, lo que simplifica el proceso de manipulación y visualización de datos.

Aunque existen otras herramientas para escribir y ejecutar código de Python, Google Colab presenta ventajas significativas. En primer lugar, es una plataforma en línea gratuita y basada en la nube, lo que elimina la necesidad de instalar software adicional en el equipo local. Esto hace que sea fácil y rápido comenzar a trabajar sin preocuparse por configuraciones o dependencias. Además, Google Colab proporciona acceso a recursos computacionales poderosos, incluyendo GPU y TPU, que aceleran el procesamiento de datos y el entrenamiento de modelos de aprendizaje automático.

Por lo tanto, la combinación de Python como lenguaje de programación y Google Colab como herramienta, ofrece ventajas como la versatilidad del lenguaje, la amplia comunidad de desarrolladores, el acceso a recursos computacionales y la facilidad de colaboración y compartición de resultados. Esto hace que Python y Google Colab sean una elección sólida para la fase de preparación y transformación de datos en la metodología CRISP-DM.

Limpieza de datos:

Este proceso sólo se tuvo que surtir con las variables de aporte, ingresos y patrimonio, ya que toda la información de cada una de las otras variables se encuentra en listas predefinidas dentro del sistema SIFI para mitigar los errores de digitalización en el momento de cargar la información de los documentos de vinculación al aplicativo interno de R4G, por lo tanto, sólo se eliminó el signo pesos y los decimales de las 3 variables financieras para dejar el valor en número entero:

```
import pandas as pd

# Cargar de la base datos
df = pd.read_csv('data.csv')

# Eliminación del signo de pesos y conversión a número entero de las
variables aporte, ingresos y patrimonio
df['APORTE'] = df['APORTE'].replace({'\$': '', ',': ''},
regex=True).astype(int)
df['INGRESOS'] = df['INGRESOS'].replace({'\$': '', ',': ''},
regex=True).astype(int)
df['PATRIMONIO'] = df['PATRIMONIO'].replace({'\$': '', ',': ''},
regex=True).astype(int)
```

Transformaciones de los datos:

Anonimización de la variable 'Cliente': Se codificó la variable 'Cliente' para ocultar el nombre del cliente y garantizar la confidencialidad de los datos.

Variables originales: Las variables originales del conjunto de datos incluían una combinación de variables numéricas y categóricas.

Variables numéricas: (aporte, ingresos, patrimonio)

Variables categóricas: (residencia, departamento, PEP, reputación, actividad económica, canal, FIC, FVP, FCP, NF)

Conversión a variables dicotómicas: Se procedió a convertir las variables categóricas en variables dicotómicas utilizando la técnica de codificación one-hot. Esto se hizo para permitir un análisis más preciso y facilitar la inclusión de estas variables en modelos de segmentación.

Al convertir las variables categóricas en formato "object" a variables dicotómicas, se crearon nuevas variables binarias para cada categoría presente en la variable original. Cada nueva variable toma el valor 1 si la observación pertenece a esa categoría y 0 en caso contrario. Esta conversión permitió que el modelo capturara las relaciones entre las categorías y los resultados de interés.

Escalado de variables numéricas: Las variables numéricas, se escalaron utilizando la técnica de escalado Min-Max. Esto garantiza que todas las variables tengan una escala comparable, evitando así la influencia desproporcionada de aquellas con magnitudes mayores.

La fórmula detrás del código que utiliza el objeto MinMaxScaler para el escalado Min-Max es la siguiente:

$$X_{scaled} = (X - X_{min}) / (X_{max} - X_{min})$$

Donde:

- X es el valor original de la variable.
- X_{scaled} es el valor escalado de la variable.
- X_{min} es el valor mínimo de la variable en el conjunto de datos.
- X_{max} es el valor máximo de la variable en el conjunto de datos.

El escalado Min-Max es una técnica comúnmente utilizada para transformar variables numéricas en un rango específico, generalmente entre 0 y 1. La fórmula calcula el valor escalado dividiendo la diferencia entre el valor original y el valor mínimo de la variable, por la diferencia entre el valor máximo y el valor mínimo de la variable.

Renombramiento y almacenamiento: El conjunto de datos modificado, con las variables dicotómicas y las variables numéricas escaladas, se ha renombrado como "data_ajustada" y se ha guardado en un formato adecuado para su posterior uso.

Para realizar lo anteriormente mencionado, se usaron los siguientes códigos:

```
#VARIABLES CATEGORICAS

#VARIABLE RESIDENCIA

#Crear las variables dicotómicas para las variables categóricas
import pandas as pd

# Obtener la lista de países únicos en la columna "RESIDENCIA"
países = data['RESIDENCIA'].unique()

# Crear las nuevas variables dicotómicas para cada país
for país in países:
    data[país] = data['RESIDENCIA'].apply(lambda x: 1 if x == país else 0)

# Eliminar la columna "RESIDENCIA"
data = data.drop('RESIDENCIA', axis=1)
```

```

#VARIABLE DEPARTAMENTO

# Obtener la lista de departamentos únicos en la columna "DEPARTAMENTO"
departamentos = data['DEPARTAMENTO'].unique()

# Reemplazar valores nulos con ceros
data['DEPARTAMENTO'] = data['DEPARTAMENTO'].fillna(0)

# Crear las nuevas variables dicotómicas para cada departamento
for departamento in departamentos:
    data[departamento] = data['DEPARTAMENTO'].apply(lambda x: 1 if x ==
departamento else 0)

# Eliminar la variable original "DEPARTAMENTO"
data = data.drop('DEPARTAMENTO', axis=1)

#VARIABLE ACTIVIDAD ECONOMICA

# Reemplazar espacios en la columna "ACTIVIDAD" con guiones bajos
data['ACTIVIDAD'] = data['ACTIVIDAD'].str.replace(' ', '_')

# Obtener la lista de actividades únicas en la columna "ACTIVIDAD"
actividades = data['ACTIVIDAD'].unique()

# Crear las nuevas variables dicotómicas para cada actividad
for actividad in actividades:
    data[actividad] = data['ACTIVIDAD'].apply(lambda x: 1 if x ==
actividad else 0)

# Eliminar la variable original "ACTIVIDAD"
data = data.drop('ACTIVIDAD', axis=1)

#VARIABLE CANAL

# Obtener la lista de canales únicos en la columna "CANAL"
canales = data['CANAL'].unique()

# Crear las nuevas variables dicotómicas para cada canal
for canal in canales:
    data[canal] = data['CANAL'].apply(lambda x: 1 if x == canal else 0)

# Eliminar la variable original "CANAL"
data = data.drop('CANAL', axis=1)

```

```

# Imprimir el resultado
print(data)

#VARIABLES PEP Y REPUTACION

# Reemplazar valores "SI" por 1 y "NO" por 0 en la columna "PEP"
data['PEP'] = data['PEP'].map({'SI': 1, 'NO': 0})

# Reemplazar valores "SI" por 1 y "NO" por 0 en la columna "REPUTACION"
data['REPUTACION'] = data['REPUTACION'].map({'SI': 1, 'NO': 0})

#VARIABLES FIC, FVP, FCP Y NF

# Reemplazar valores "SI" por 1 y "NO" por 0 en las columnas "FIC", "FVP",
"FIC" y "NF"
data['FIC'] = data['FIC'].map({'SI': 1, 'NO': 0})
data['FVP'] = data['FVP'].map({'SI': 1, 'NO': 0})
data['FCP'] = data['FCP'].map({'SI': 1, 'NO': 0})
data['NF'] = data['NF'].map({'SI': 1, 'NO': 0})

#VARIABLES NUMERICAS

#VARIABLES APOORTE, INGRESOS Y PATRIMONIO

# Seleccionar las variables numéricas
numeric_vars = ['INGRESOS', 'PATRIMONIO', 'APOORTE']

# Crear un objeto MinMaxScaler
scaler = MinMaxScaler()

# Escalar las variables numéricas
data[numeric_vars] = scaler.fit_transform(data[numeric_vars])

# Imprimir el DataFrame con las variables escaladas y las variables
dicotómicas
print(data)

# Renombrar y guardar la data con las variables escaladas y las variables
dicotómicas
data.to_excel('data_ajustada.xlsx', index=False)

```


A continuación, se presenta un extracto de las tablas de datos:

Ilustración 3

Data inicial

RESIDENCIA	DEPARTAMENTO	PEP	REPUTACION	ACTIVIDAD	APORTE	CANAL	FIC	FVP	FCP	NF	INGRESOS	PATRIMONIO
COLOMBIA	BOGOTA	NO	NO	ACTIVIDADES INMOBILIARIAS	\$ 1.200.000.000,00	PRESENCIAL	SI	NO	NO	SI	\$ 25.000.000,25	\$ 3.000.000.000,00
ECUADOR		NO	SI	OTROS TIPOS DE EDUCACION	\$ 200.000.000,00	VIRTUAL	SI	NO	NO	NO	\$ 10.000.000,50	\$ 500.000.000,00
COLOMBIA	BOGOTA	NO	NO	ALQUILER DE MAQUINARIA	\$ 680.000.000,00	REFERIDO	SI	NO	NO	NO	\$ 90.000.000,60	\$ 600.000.000,00
ESPAÑA		NO	NO	ACTIVIDADES INMOBILIARIAS	\$ 90.000.000,00	VIRTUAL	SI	NO	SI	SI	\$ 400.000.000,00	\$ 300.000.000,00
PANAMA		NO	NO	ACTIVIDADES DEL MERCADO DE VALORES	\$ 600.000.000,00	VIRTUAL	NO	NO	SI	NO	\$ 3.814.278,58	\$ 870.548.166,87
ESTADOS UNIDOS		NO	NO	OTRAS ACTIVIDADES DE SERVICIOS FINANCIEROS	\$ 1.750.000.000,00	VIRTUAL	NO	NO	SI	SI	\$ 500.000.000,69	\$ 1.020.000.000,20
COLOMBIA	BOGOTA	NO	NO	OTRAS ACTIVIDADES PROFESIONALES	\$ 3.000.000.000,00	PRESENCIAL	NO	NO	SI	NO	\$ 13.473.895,65	\$ 4.736.721.340,00
COLOMBIA	CAUCA	NO	SI	ACTIVIDADES INMOBILIARIAS	\$ 800.000.000,00	VIRTUAL	NO	NO	SI	NO	\$ 40.000.000,69	\$ 74.580.000,00
BARBADOS		NO	NO	OTRAS ACTIVIDADES DE SERVICIOS FINANCIEROS	\$ 70.000.000,00	VIRTUAL	NO	NO	NO	SI	\$ 28.515.000,45	\$ 371.293.000,69
BARBADOS		NO	NO	COMPRA DE CARTERA O FACTORING	\$ 30.000.000,00	VIRTUAL	NO	NO	NO	SI	\$ 62.589.200,00	\$ 1.000.109.000,00

Fuente: Elaboración propia

Ilustración 4

Resultado de la limpieza de datos

RESIDENCIA	DEPARTAMENTO	PEP	REPUTACION	ACTIVIDAD	APORTE	CANAL	FIC	FVP	FCP	NF	INGRESOS	PATRIMONIO
COLOMBIA	BOGOTA	NO	NO	ACTIVIDADES INMOBILIARIAS	1200000000	PRESENCIAL	SI	NO	NO	SI	25000000	3000000000
ECUADOR		NO	SI	OTROS TIPOS DE EDUCACION	200000000	VIRTUAL	SI	NO	NO	NO	10000000	500000000
COLOMBIA	BOGOTA	NO	NO	ALQUILER DE MAQUINARIA	680000000	REFERIDO	SI	NO	NO	NO	90000000	600000000
ESPAÑA		NO	NO	ACTIVIDADES INMOBILIARIAS	90000000	VIRTUAL	SI	NO	SI	SI	400000000	300000000
PANAMA		NO	NO	ACTIVIDADES DEL MERCADO DE VALORES	600000000	VIRTUAL	NO	NO	SI	NO	3814278	870548166
ESTADOS UNIDOS		NO	NO	OTRAS ACTIVIDADES DE SERVICIOS FINANCIEROS	1750000000	VIRTUAL	NO	NO	SI	SI	500000000	1020000000
COLOMBIA	BOGOTA	NO	NO	OTRAS ACTIVIDADES PROFESIONALES	3000000000	PRESENCIAL	NO	NO	SI	NO	13473895	4736721340
COLOMBIA	CAUCA	NO	SI	ACTIVIDADES INMOBILIARIAS	800000000	VIRTUAL	NO	NO	SI	NO	40000000	74580000
BARBADOS		NO	NO	OTRAS ACTIVIDADES DE SERVICIOS FINANCIEROS	70000000	VIRTUAL	NO	NO	NO	SI	28515000	371293000
BARBADOS		NO	NO	COMPRA DE CARTERA O FACTORING	30000000	VIRTUAL	NO	NO	NO	SI	62589200	1000109000

Fuente: Elaboración propia

Ilustración 5

Resultado de la transformación de los datos

PEP	REPUTACION	APORTE	FIC	FVP	FCP	NF	INGRESOS	PATRIMONIO	COLOMBIA	ECUADOR	ESPAÑA	ESTADOS UNIDOS	PANAMA	ISLAS VIRGENES BRITANICAS	BARBADOS	BERMUDAS	BRASIL	BOGOTA
0	0	0,073003626	1	0	0	1	0,000247106	0,003058601	1	0	0	0	0	0	0	0	0	1
0	0	0,002482537	1	1	0	0	0,000041184	0,000150992	1	0	0	0	0	0	0	0	0	1
0	0	0,000034683	1	0	0	0	0,000002059	0,000018364	1	0	0	0	0	0	0	0	0	1
0	0	0,000011561	1	0	0	0	0,000010296	0,000003061	1	0	0	0	0	0	0	0	0	1
0	0	0,071890611	1	0	1	0	0,000231240	0,002038387	1	0	0	0	0	0	0	0	0	1
0	1	0,012157130	1	0	0	0	0,000092665	0,000508066	0	1	0	0	0	0	0	0	0	0
0	0	0,000000000	1	0	0	0	0,000003089	0,000003061	1	0	0	0	0	0	0	0	0	1
0	0	0,005098936	1	1	0	0	0,000020592	0,000202002	1	0	0	0	0	0	0	0	0	1
0	0	0,000018254	1	0	0	0	0,000008237	0,000004081	1	0	0	0	0	0	0	0	0	1

Fuente: Elaboración propia

Gracias a los procesos y procedimientos establecidos en estas dos fases, se logró obtener una visión clara y completa de las variables relacionadas con los factores de riesgo LA/FT. Los resultados obtenidos durante esta etapa proporcionaron una base sólida para el desarrollo del

modelo de segmentación, permitiendo una identificación más precisa de clientes que podrían llegar a presentar operaciones inusuales y una mejora en la gestión del Sistema de Administración de Riesgos de Lavado de Activos y Financiación del Terrorismo en R4G, en cumplimiento de la normatividad vigente y las recomendaciones de los entes de control.

7. Modelado

7.1. Modelos de Aprendizaje

De acuerdo con Hastie, Trevor et al., (2001) en su libro *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, los modelos de aprendizaje se refieren a los métodos y algoritmos utilizados en el campo del aprendizaje automático para realizar tareas de predicción, clasificación, agrupamiento y otras tareas relacionadas con el análisis de datos.

El libro describe una variedad de modelos de aprendizaje que se utilizan en el análisis estadístico y la minería de datos. Algunos de los modelos de aprendizaje destacados incluyen ¹:

Regresión lineal: Este modelo se utiliza para predecir una variable continua basada en una o más variables predictoras.

Regresión logística: Es un modelo utilizado para problemas de clasificación binaria. Estima la probabilidad de pertenecer a una clase determinada utilizando una función logística.

Árboles de decisión: Utilizadas para clasificar o predecir valores basados en reglas de decisión. Cada nodo del árbol representa una característica y cada rama representa una posible salida.

Modelos supervisados: Los modelos supervisados se utilizan cuando se dispone de datos etiquetados, es decir, datos en los que ya se conoce la respuesta o variable de interés.

Modelos no supervisados: Se utilizan cuando no se dispone de datos etiquetados o no hay una variable de respuesta definida. Estos modelos exploran la estructura y patrones ocultos en los datos sin una guía explícita. Se utilizan para descubrir grupos o clústeres de

¹ Se adaptaron los conceptos de modelos de aprendizaje del libro titulado *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* al español.

datos similares, reducir la dimensionalidad de los datos o encontrar relaciones y patrones interesantes.

7.1.1. Selección del Modelo de Aprendizaje

Con la información que se cuenta, ¿cómo se logró separar los elementos en grupos homogéneos al interior de ellos y heterogéneos entre ellos, teniendo en cuenta las variables seleccionadas a cada uno de los factores de riesgo en la fase 2 y 3?

Para responder dicha pregunta, fue necesario tener en cuenta que el ejercicio de aplicar una metodología de segmentación mediante modelos estadísticos y complementado por criterio de expertos para variables cualitativas fue no supervisado, es decir, un conjunto de técnicas que permiten inferir modelos para extraer conocimiento de conjuntos de datos donde a priori se desconoce el resultado, con el objetivo de describir las asociaciones y los patrones en un conjunto de datos.

7.2. Algoritmos de Segmentación

Para la incorporación de la metodología de segmentación de los factores de riesgos LA/FT se partió de los algoritmos de segmentación presentados en el libro *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, donde Hastie, Tibshirani y Friedman (2001), proporcionan una cobertura detallada de diversos algoritmos ²:

K-means: Se basa en la asignación de puntos de datos a k clusters, minimizando la suma de las distancias al cuadrado entre los puntos y los centroides de los clusters. Este enfoque permite agrupar conjuntos de datos en función de su similitud, lo que facilita la comprensión de patrones y la identificación de grupos homogéneos.

² Se adaptaron los conceptos de algoritmos de segmentación del libro titulado *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* al español.

Algoritmo de Agrupamiento Jerárquico: Construye una estructura de clusters en forma de árbol. Comienza con cada punto de datos como un clúster individual y los fusiona iterativamente para formar clúster más grandes, con el fin de identificar niveles de similitud y relación entre los datos,.

DBSCAN: El algoritmo DBSCAN se utiliza para identificar clusters basados en la densidad de los puntos de datos. Este enfoque es especialmente útil cuando los clusters tienen formas y tamaños irregulares o cuando hay presencia de ruido o outliers en los datos.

Spectral Clustering: Construye un grafo a partir de los datos y aplica técnicas de corte espectral para dividir el grafo en clusters. Este método es efectivo para identificar clusters no convexos o con formas complejas.

Adicionalmente, la Universidad de Granada (2019), describe otra técnica:

Fuzzy c-means: “Función de pertenencia continua que asigna valores intermedios entre 0 y 1 para reflejar la similitud de un elemento con un grupo. Los valores cercanos a uno indican una mayor similitud y los cercanos a cero indican menor similitud”.

Adicionalmente, se consultó otro algoritmo en el artículo “Multivariate Analysis I” de Alboukadel Kassambara (2017) ³:

K-Medoids: El algoritmo K-Medoids es una técnica de agrupamiento que busca dividir un conjunto de datos en k grupos o clústeres. A diferencia del algoritmo K-Means, en K-Medoids cada clúster está representado por un punto de datos del propio clúster, conocido como medoide. El medoide es el objeto dentro del clúster que tiene la menor disimilitud promedio con todos los demás miembros del clúster.

³ Se adaptaron los conceptos de K-Medoids y PAM del artículo titulado “Multivariate Analysis I - Practical Guide To Cluster Analysis in R - Unsupervised Machine Learning” al español.

El algoritmo más común para implementar K-Medoids es el algoritmo PAM (Partitioning Around Medoids). Este algoritmo busca encontrar los medoides óptimos mediante la iteración de pasos de asignación y actualización de medoides. El resultado final es la formación de k clústeres con sus respectivos medoides.

7.2.1. Selección de los Algoritmos de Segmentación.

De acuerdo con la guía de CRISP-DM de IBM (2021), “la determinación del modelo más adecuado se debe basar en las siguientes consideraciones: los tipos de datos disponibles para la minería, sus objetivos y los requisitos específicos del modelado”.

En el proceso de selección de los modelos no supervisados de clustering para el proyecto de segmentación SARLAFT, se han considerado diversos aspectos. En primer lugar, se realizó un análisis de la base de datos compuesta por 132 clientes y 13 variables iniciales, las cuales fueron descritas en etapas anteriores. El objetivo principal consiste en identificar segmentos o grupos de clientes que compartan perfiles similares, buscando asegurar la homogeneidad dentro de cada clúster y la heterogeneidad entre ellos. Esto permitirá evaluar de manera efectiva los factores de riesgo asociados al SARLAFT.

Por lo tanto, se seleccionaron los algoritmos K-Medoids, Algoritmo de Agrupamiento Jerárquico y Fuzzy c-means debido a múltiples razones. En primer lugar, estos algoritmos son capaces de manejar tanto datos dicotómicos como variables numéricas estandarizadas, lo cual es fundamental para nuestro conjunto de datos. Además, estos algoritmos no se basan en el cálculo de promedios, lo cual es beneficioso en este proyecto, ya que se busca evitar la influencia desproporcionada de valores atípicos o ruidosos en el proceso de segmentación.

Al optar por K-Medoids, Algoritmo de Agrupamiento Jerárquico y Fuzzy c-means, se busca una aproximación robusta y confiable en el análisis de segmentación. Estos algoritmos permiten capturar patrones complejos y estructuras de los datos, proporcionando una mayor flexibilidad en la identificación de grupos de clientes con perfiles similares. Además, el enfoque basado en medoides y el uso de técnicas de agrupamiento jerárquico y de lógica difusa en Fuzzy c-means, permite considerar la incertidumbre y la heterogeneidad de los datos en el proceso de segmentación.

Estos tres algoritmos se seleccionan por su idoneidad para el análisis de variables categóricas transformadas a dicotómicas y variables numéricas estandarizadas, y por su capacidad para abordar características particulares de los conjuntos de datos utilizados en este proyecto. A diferencia de los métodos basados en promedios, los algoritmos seleccionados ofrecen enfoques distintos que pueden revelar de manera precisa y completa las estructuras y relaciones subyacentes en los datos. Además, han demostrado ser efectivos en la identificación de patrones y agrupaciones relevantes en conjuntos de datos de tamaño moderado, por lo tanto, el algoritmo DBSCAN, quedó descartado ya que este se basa en la densidad de los datos y puede no ser apropiado para conjuntos de datos pequeños o con distribuciones de densidad heterogénea, al igual que el algoritmo de Spectral Clustering, ya que este se basa en la matriz de afinidad de los datos y es especialmente útil para detectar estructuras complejas y no lineales en grandes conjuntos de datos, y en el caso de una base de datos relativamente pequeña para este proyecto, donde el número de clientes y variables es limitado, es más eficiente utilizar algoritmos que se enfoquen directamente en la partición de los datos en grupos.

Por otro lado, para la identificación del número de conglomerados o clústeres en el análisis no supervisado de segmentación, no existe una regla definitiva para determinar el

número óptimo de agrupaciones ni un método infalible para seleccionar el mejor enfoque de conglomerados, sin embargo, en este estudio, se ejecutaron métodos pertinentes en cada algoritmo para identificar el número óptimo de clusters en cada caso, los cuales se describirán más adelante.

7.3.Ejecución de los Algoritmos Seleccionados

A continuación, se presenta el paso a paso que se realizó para la ejecución de los 3 algoritmos, mediante el lenguaje de programación Python y con la herramienta de GoogleColab:

Análisis estadístico de las variables numéricas antes de estandarizarlas: Se usaron las bibliotecas NumPy y pandas para calcular las siguientes estadísticas descriptivas: Media: La medida central de los datos, Mediana: El valor medio de los datos, Desviación estándar: Una medida de la dispersión de los datos alrededor de la media y el cálculo del rango, el mínimo y máximo.

Adicionalmente, se realizó análisis de correlaciones, para obtener información valiosa sobre las relaciones entre las variables numéricas. Los puntos que se consideraron en el análisis son:

Variables Altamente Correlacionadas: Identificar variables con correlaciones cercanas a 1 o -1, lo que indica una relación fuerte.

Variables Débilmente Correlacionadas: Buscar variables con correlaciones cercanas a 0, lo que sugiere una relación débil o inexistente.

Este proceso se realizó para identificar si se debía agregar o eliminar alguna variable cuantitativa para ejecutar los 3 algoritmos de segmentación seleccionados.

Análisis de variables categóricas antes de transformarlas a variables dicotómicas:

Mediante un bucle "for", se recorre cada variable categórica de la data y se utiliza el método "value_counts()" para obtener el conteo de cada valor único en la respectiva columna. El resultado se guarda en la variable "count" y se imprime en pantalla el nombre de la variable seguido del conteo de valores correspondiente.

El código usado en Python (y que se suministra más adelante) es útil para obtener una visión general de la distribución de los valores en las columnas categóricas de un conjunto de datos. Proporciona información sobre la frecuencia y prevalencia de cada categoría, lo que resulta relevante para comprender la composición y características de los datos en cada variable categórica.

Finalmente se usó la biblioteca Matplotlib para generar gráficas de barras para visualizar el conteo de valores en cada variable categórica del conjunto de datos y así realizar el respectivo análisis.

Transformación de los datos: Se seleccionaron las variables numéricas 'INGRESOS', 'PATRIMONIO' y 'APORTE' del conjunto de datos. Luego, se utilizó el objeto MinMaxScaler de la librería sklearn.preprocessing para realizar la estandarización de las variables numéricas. La estandarización se basó en la técnica de escalado Min-Max (técnica descrita en las páginas 45 y 48 del presente documento en el apartado transformación de los datos), que transforma los valores de las variables a un rango de 0 a 1. Posteriormente, se actualizó el DataFrame 'data' con las variables numéricas escaladas.

Paso seguido se realizó la transformación de las variables categóricas a variables dicotómicas, (proceso descrito en las páginas 45 a la 47 del presente documento en el apartado transformación de los datos).

Ejecución del método y la métrica para determinar el número óptimo de clústers:

Se ejecutó el método del codo con la métrica de Manhattan y el índice de silueta:

El método del codo es una técnica utilizada para determinar el número óptimo de clusters en un conjunto de datos en el contexto del agrupamiento. Se basa en la idea de que a medida que aumentamos el número de clusters, se produce una disminución en una medida de rendimiento que refleja la calidad de la agrupación. El método del codo busca identificar el punto de inflexión en un gráfico que muestra la relación entre el número de clusters y la medida de rendimiento. Este punto de inflexión sugiere el número óptimo de clusters, ya que más allá de este punto, los beneficios de agregar más clusters son limitados.

La métrica de Manhattan, también conocida como distancia de la ciudad, se utiliza en el contexto del agrupamiento para medir la distancia o similitud entre dos puntos en un espacio de características. Dicha métrica se calcula sumando las diferencias absolutas entre las coordenadas de los puntos en cada dimensión. En contraste con otras métricas, como la euclidiana, la distancia de Manhattan no toma en cuenta la magnitud o el peso de las variables, sino simplemente la diferencia en los valores.

Esta métrica, se emplea para calcular la distancia entre los puntos en lugar de utilizar promedios o la suma de cuadrados. La métrica de Manhattan se calcula sumando las diferencias absolutas entre las coordenadas de los puntos en cada dimensión:

La fórmula para calcular la distancia de Manhattan entre dos puntos en un espacio n-dimensional es la siguiente:

$$\text{Distancia de Manhattan} = |x1 - x2| + |y1 - y2| + |z1 - z2| + \dots + |xn - xn|$$

Donde:

$x1, x2, y1, y2, z1, z2, \dots, xn, xn$ son las coordenadas de los puntos en cada dimensión.

La distancia de Manhattan se obtiene sumando las diferencias absolutas entre las coordenadas de los puntos en cada dimensión. No se tienen en cuenta los cuadrados de las diferencias, como en el caso de la distancia euclidiana. Esto significa que la distancia de Manhattan solo mide la diferencia en los valores sin considerar la magnitud o la dirección.

El índice de silueta es una medida que evalúa la calidad de los clusters en un análisis de agrupamiento. Se calcula para cada punto individual y proporciona una medida de cuán bien se asigna ese punto a su cluster en comparación con los clusters vecinos. La fórmula para el cálculo del índice de silueta se describe a continuación:

Para un punto dado, se calcula la silueta como la diferencia entre la distancia media a los puntos del mismo cluster (a) y la distancia media a los puntos del cluster más cercano diferente (b). Luego, la silueta para ese punto se define como:

$$silueta = (b - a) / \max(a, b)$$

Donde:

a: Distancia media entre el punto y todos los demás puntos en el mismo cluster.

b: Distancia media entre el punto y todos los puntos en el cluster más cercano diferente.

Esta elección es apropiada para el proyecto ya que se desea que las diferencias en los valores de las variables no estén influenciadas por la escala o la magnitud, sino solo por la diferencia en los valores.

Ejecución de los modelos de segmentación: Se ejecutaron tres modelos de segmentación de acuerdo con los resultados de los clusters óptimos obtenidos en el paso anterior:

Para el modelo K-Medoids, se utilizó la función `pairwise_distances` del módulo `sklearn.metrics` para calcular la matriz de distancias de las variables de la data utilizando la métrica de Manhattan, paso seguido, se ejecutó el K-medoids utilizando la implementación

proporcionada por `sklearn_extra.cluster.KMedoids`. Para este proceso se utilizó la matriz de distancias precalculada (`distances`) como entrada y se especificó el número óptimo de clusters obtenido en el método del codo ($k=3$). El algoritmo K-medoids busca los medoides (muestras representativas de cada cluster) y asigna cada muestra a un cluster en función de su distancia al medoide correspondiente.

La fórmula matemática para calcular la distancia de un punto a un medoide en el caso de la distancia de Manhattan es:

$$dist_manhattan = \sum |x - m|$$

Donde x es un punto y m es un medoide.

En el caso del Algoritmo de Agrupamiento Jerárquico, se utilizó la función `pairwise_distances` del módulo `sklearn.metrics` para calcular la matriz de distancias entre las muestras en data utilizando la métrica de Manhattan, de manera similar al código anterior.

Adicionalmente, se utilizó la función `linkage` del módulo `scipy.cluster.hierarchy` para calcular el enlace jerárquico entre los puntos utilizando la matriz de distancias. Se utilizó el método `complete`, que utiliza la distancia máxima entre los puntos de dos clusters para fusionarlos en un nuevo cluster.

Finalmente, la función `fcluster` se utilizó para obtener las etiquetas de los clusters para el número óptimo de clusters especificado utilizando el enlace jerárquico (`linkage_matrix`). El criterio utilizado es `maxclust`, que busca formar el número óptimo de clusters especificado.

El método `complete` es un enfoque utilizado en el algoritmo de agrupamiento jerárquico para calcular la distancia entre clusters. También conocido como enlace completo, este método se basa en la idea de que la distancia entre dos clusters se determina por la máxima distancia entre cualquier par de puntos, uno perteneciente a cada cluster.

En términos matemáticos, la distancia utilizando el método complete se calcula como la máxima distancia entre todos los posibles pares de puntos, uno de cada cluster. Este enfoque tiene la propiedad de preservar la estructura de clusters compactos y bien separados, ya que solo se tiene en cuenta la distancia máxima entre los puntos.

La fórmula matemática para la distancia utilizando el método complete es:

$$d_complete(C_i, C_j) = \max(distancia(x_i, y_j))$$

Donde:

$d_complete(C_i, C_j)$ representa la distancia entre los clusters C_i y C_j utilizando el método complete.

$distancia(x_i, y_j)$ es la medida de distancia utilizada para calcular la distancia entre los puntos x_i y y_j .

En esta fórmula, se consideran todos los posibles pares de puntos, uno de C_i y otro de C_j , y se calcula la distancia entre ellos. Luego, se selecciona la máxima distancia entre todos estos pares como la distancia entre los clusters C_i y C_j .

En cuanto al modelo Fuzzy c-means, se importaron las bibliotecas necesarias y se convirtió el DataFrame de Pandas `data` a una matriz `ndarray` `data_array`. Esto se hizo para asegurar de que los datos estén en el formato adecuado para el cálculo de las distancias y la ejecución del algoritmo FCM.

Paso seguido, se calculó la matriz de distancias utilizando la función `pairwise_distances` de la biblioteca `scikit-learn`. La métrica utilizada en este caso es la distancia de Manhattan, que se especifica con el argumento `metric='manhattan'`. La matriz de distancias calculada es una matriz simétrica que contiene las distancias entre cada par de puntos en el conjunto de datos. Luego, se

traspuso la matriz de distancias utilizando el operador $.T$ para asegurar de que tuviera las dimensiones adecuadas para su posterior uso en el algoritmo FCM.

Finalmente, se creó una instancia del objeto FCM con $n_clusters=3$, (de acuerdo con el resultado obtenido en el método del codo), y se implementó el método `fit` para ejecutar el algoritmo FCM en la matriz de distancias transpuesta, ya que este método ejecuta el algoritmo iterativo y ajusta los centroides y las membresías difusas de los puntos en función de la matriz de distancias. Cada iteración mejora gradualmente la configuración de los clusters hasta que se cumple el criterio de convergencia.

Visualización de los resultados: De acuerdo con las técnicas de visualización mencionadas en el *The Elements of Statistical Learning* (Hastie, Trevor et al., 2001), se han seleccionado las siguientes gráficas para visualizar los resultados de los tres algoritmos de segmentación seleccionados: K-medoids, Algoritmo de Agrupamiento Jerárquico y Fuzzy c-means. A continuación, se detallan las gráficas elegidas:

Gráfico de barras: Técnica de visualización que se utiliza para representar el conteo o frecuencia de ocurrencia de diferentes categorías en un conjunto de datos. En este tipo de gráfico, cada categoría se representa mediante una barra rectangular cuya altura está determinada por el número de observaciones o frecuencia asociada a esa categoría.

Esta gráfica fue seleccionada teniendo en cuenta las características y resultados esperados de cada algoritmo, ya que ofrece una perspectiva visual de los resultados de la segmentación, lo que facilita la interpretación y comprensión de los grupos identificados.

Proceso en Python en la herramienta GoogleColab:

A continuación, se presenta el paso a paso de la ejecución de los modelos y los resultados obtenidos con su respectivo análisis:

Importe de las librerías necesarias y cargue de la data inicial:

```
# Importar las librerías necesarias para el modelo de segmentación
+import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import AgglomerativeClustering, KMeans
!pip install scikit-fuzzy
import matplotlib.pyplot as plt
from skfuzzy.cluster import cmeans
from sklearn.preprocessing import MinMaxScaler
!pip install scikit-learn-extra
from sklearn_extra.cluster import KMedoids
import matplotlib.pyplot as plt

# Cargar los datos desde un archivo Excel
data = pd.read_excel('/content/data.xlsx')
```

A continuación, se presenta las características y estructura del conjunto de datos iniciales (antes de realizar la transformación a variables dicotómicas y variables numéricas estandarizadas):

```
data.head() # Mostrar las primeras filas del DataFrame
data.describe() # Estadísticas descriptivas de variables cuantitativas
data.info() # Mostrar información sobre el tipo de datos de cada columna
```

Ilustración 6

Análisis Exploratorio de Datos: Características y Estructura del Conjunto de Datos

0	RESIDENCIA	132	non-null	object
1	DEPARTAMENTO	108	non-null	object
2	PEP	132	non-null	object
3	REPUTACION	132	non-null	object
4	ACTIVIDAD	132	non-null	object
5	APORTE	132	non-null	int64
6	CANAL	132	non-null	object
7	FIC	132	non-null	object
8	FVP	132	non-null	object
9	FCP	132	non-null	object
10	NF	132	non-null	object
11	INGRESOS	132	non-null	int64
12	PATRIMONIO	132	non-null	int64

dtypes: int64(3), object(10)
memory usage: 13.5+ KB

Fuente: Elaboración propia

Tamaño del conjunto de datos: El conjunto de datos contiene 132 filas y 13 columnas, lo que implica que se tienen 132 registros del total de clientes y 13 variables iniciales que describen diferentes aspectos de cada cliente.

Tipos de variables: 3 de las variables en el conjunto de datos son numéricas, específicamente del tipo entero (int64) y las 10 restantes son categóricas. Esta combinación de variables numéricas y categóricas proporciona información diversa y útil para el análisis y comprensión de la data.

Valores nulos: Sólo la variable DEPARTAMENTO presenta valores nulos. Esto es positivo, ya que tener datos completos evita la necesidad de tratamiento adicional de valores faltantes. (este proceso se encuentra detallado en página 42 del presente documento en el apartado transformación de los datos).

Análisis estadístico de las variables numéricas y círculo de correlaciones:

```
#ANÁLISIS ESTADÍSTICO DE LAS VARIABLES NUMÉRICAS

# Seleccionar las variables numéricas
numeric_vars = ['APORTE', 'INGRESOS', 'PATRIMONIO']

# Dividir los valores entre 1 millon
data[numeric_vars] = data[numeric_vars] / 1000000

# Calcular la media
media = data[numeric_vars].mean()
print('Media:')
print(media)

# Calcular la mediana
mediana = data[numeric_vars].median()
print('\nMediana:')
print(mediana)

# Calcular la desviación estándar
desviacion_estandar = data[numeric_vars].std()
print('\nDesviación estándar:')
print(desviacion_estandar)

# Calcular el rango
rango = data[numeric_vars].max() - data[numeric_vars].min()
print('\nRango:')
print(rango)

# Calcular el mínimo y máximo
minimo = data[numeric_vars].min()
maximo = data[numeric_vars].max()
print('\nMínimo:')
print(minimo)
print('\nMáximo:')
print(maximo)

#CUADRO RESUMEN

# Crear un DataFrame con los resultados
resumen = pd.DataFrame({
    'Media': media,
    'Mediana': mediana,
    'Desv Estánd': desviacion_estandar,
```

```

    'Rango': rango,
    'Mín': minimo,
    'Máx': maximo
})

# Establecer el formato de un decimal para los valores del DataFrame
resumen = resumen.round(1)

# Mostrar el cuadro resumen
print(resumen)

```

Ilustración 7

Resultados Estadísticos

	Media	Mediana	Desv Estánd	Rango	Mín	Máx
APORTE	696.0	500.0	1574.2	16434.8	0.2	16435.0
INGRESOS	3201.4	32.0	13274.0	97124.2	1.0	97125.2
PATRIMONIO	33779.7	1818.6	114511.7	980186.8	2.0	980188.8

Nota: Cifras en millones de pesos, **Fuente:** Elaboración propia

Media y mediana: Se puede observar que tanto la media como la mediana difieren en las tres variables estudiadas. Esto sugiere una posible asimetría en los conjuntos de datos. Por ejemplo, en la variable APORTE, la media (696.0) es mayor que la mediana (500.0), lo que indica una posible sesgo hacia valores más altos. En el caso de INGRESOS y PATRIMONIO, las diferencias son aún más pronunciadas, lo que sugiere una distribución asimétrica en estas variables.

Desviación estándar: En las tres variables, se evidencia que la desviación estándar es relativamente alta. Esto indica que los valores de APORTE, INGRESOS y PATRIMONIO están ampliamente dispersos alrededor de la media. Por lo tanto, existe una variabilidad considerable en estos conjuntos de datos.

Rango: Las 3 variables presentan rangos amplios, lo que indica diferencias significativas entre los valores extremos. Estos amplios rangos podrían ser indicativos de la presencia de valores atípicos o extremos en los conjuntos de datos.

Mínimo y máximo: Los valores mínimos y máximos varían en gran medida en las tres variables. Esto sugiere que las variables APORTE, INGRESOS y PATRIMONIO tienen diferentes escalas y amplitudes en cuanto a los valores observados. Es importante tener en cuenta estas diferencias al interpretar y comparar los datos entre estas variables.

```
# CÍRCULO DE CORRELACIONES PARA VARIABLES NUMÉRICAS

import pandas as pd
import plotly.graph_objects as go

# Seleccionar solo las variables numéricas de interés
numeric_vars = ['APORTE', 'INGRESOS', 'PATRIMONIO']
data_numeric = data[numeric_vars]

# Calcular la matriz de correlación
corr_matrix = data_numeric.corr()

# Crear el círculo de correlaciones
fig = go.Figure(data=go.Scatterpolar(
    r=corr_matrix.iloc[0].values,
    theta=corr_matrix.columns,
    fill='toself'
))

# Ajustar el diseño del círculo de correlaciones
fig.update_layout(
    polar=dict(
        radialaxis=dict(
            visible=True,
            range=[-1, 1] # Ajusta el rango de -1 a 1 para las
correlaciones
        )
    ),
    showlegend=False
)

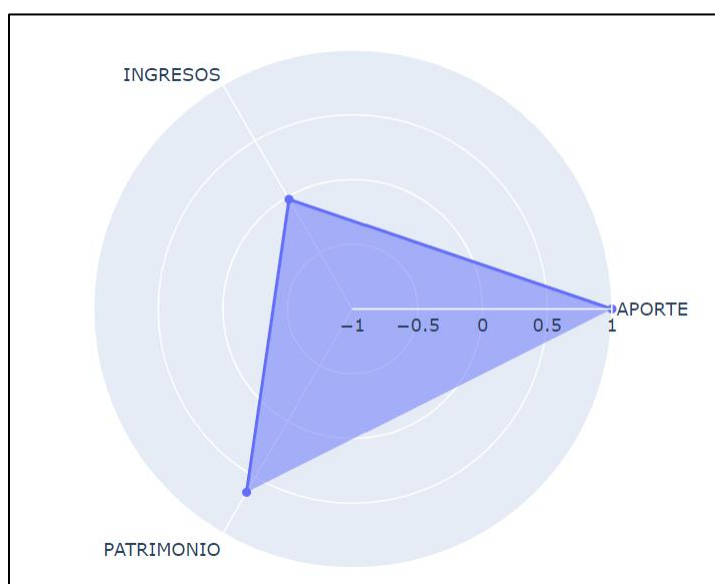
# Mostrar el círculo de correlaciones
```

```
fig.show()

# Mostrar la tabla de correlaciones
corr_table = pd.DataFrame(corr_matrix.values, columns=corr_matrix.columns,
index=corr_matrix.index)
print(corr_table)
```

Ilustración 8

Círculo de Correlaciones



Fuente: Elaboración propia

El círculo de correlación muestra la relación entre las tres variables numéricas utilizadas en el conjunto de datos de R4G: "APOORTE", "INGRESOS" y "PATRIMONIO". Al examinar los valores de correlación, se puede evaluar la fuerza y la dirección de la relación entre estas variables, a continuación, se presente el analizado realizado:

Variables Altamente Correlacionadas: Las variables "APOORTE" y "PATRIMONIO" presentan una correlación de 0.633809, lo cual indica una relación fuerte y positiva entre ambas variables. Esta correlación cercana a 1 sugiere que hay una asociación significativa entre la

cantidad de aportes realizados y el patrimonio de los clientes. Esto significa que a medida que aumenta el aporte, es probable que también aumente el patrimonio.

Variables Débilmente Correlacionadas: La variable "APORTE" y "INGRESOS" tienen una correlación de -0.022268, lo cual indica una correlación muy cercana a 0. Esto sugiere que no existe una relación lineal fuerte entre los aportes y los ingresos de los clientes. Por lo tanto, los aportes realizados no se ven afectados de manera directa por los ingresos de los clientes.

Por lo tanto, las correlaciones significativas entre las variables y su capacidad para proporcionar información única hacen que todas ellas sean relevantes para el modelo de segmentación. No fue necesario eliminar ninguna variable porque cada variable contribuye de manera significativa al modelo al proporcionar información única y valiosa.

Análisis de variables categóricas y visualización mediante grafica de barras:

```
#ANÁLISIS DE VARIABLES CATEGÓRICAS

# Conteo de valores únicos en cada variable categórica
variables_categoricas = ['RESIDENCIA', 'DEPARTAMENTO', 'PEP',
                        'REPUTACION', 'ACTIVIDAD', 'CANAL', 'FIC', 'FVP', 'FCP', 'NF']

for variable in variables_categoricas:
    count = data[variable].value_counts()
    print(f"Conteo de valores para la variable {variable}:")
    print(count)
    print()

# Diagrama de barras
import matplotlib.pyplot as plt

bar_width = 0.8 # Ancho de las barras (ajustar según tus preferencias)

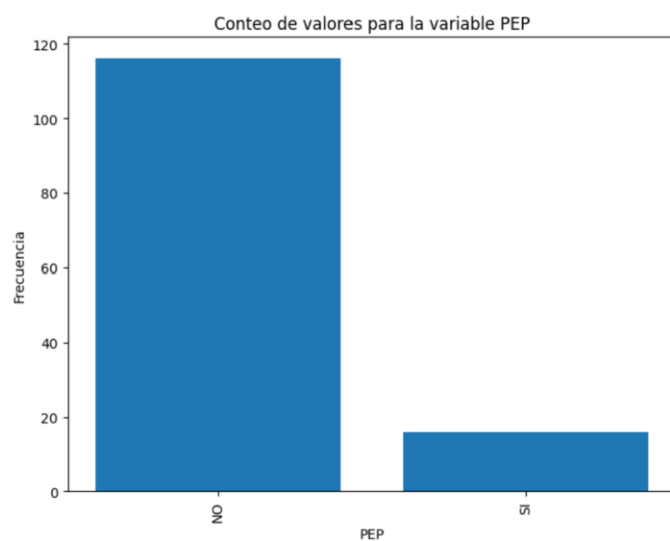
for variable in variables_categoricas:
    count = data[variable].value_counts()
    plt.figure(figsize=(8, 6))
    plt.bar(count.index, count.values, width=bar_width)
    plt.xlabel(variable)
    plt.ylabel('Frecuencia')
    plt.title(f'Conteo de valores para la variable {variable}')
```

```
plt.xticks(rotation=90)  
plt.show()
```

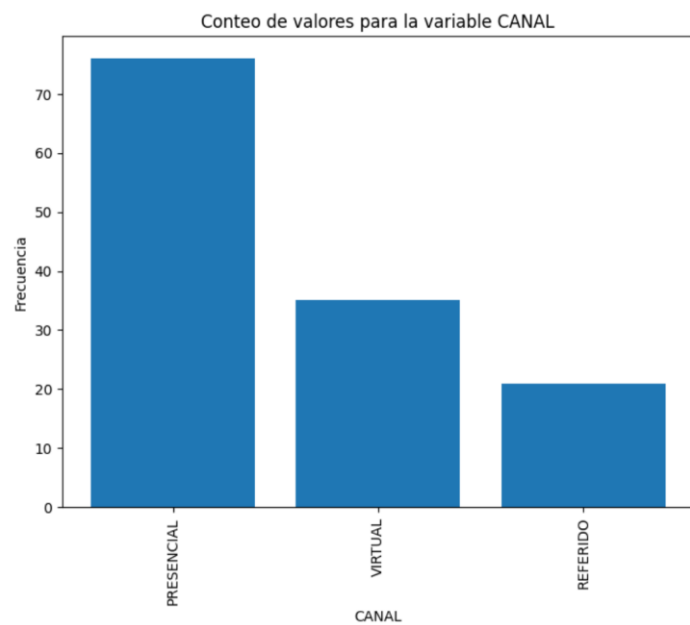
Ilustración 9

Gráfico de barras: Conteo de valores para variables categóricas

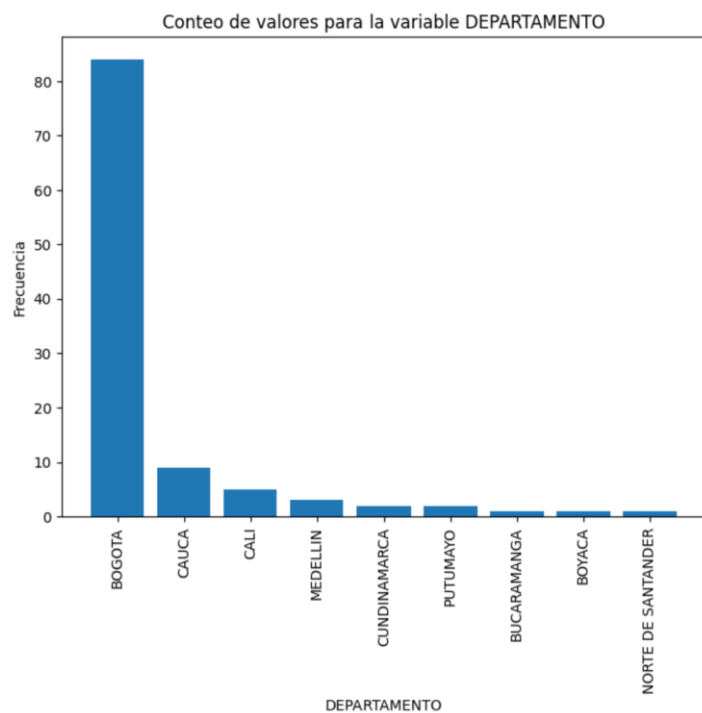
A

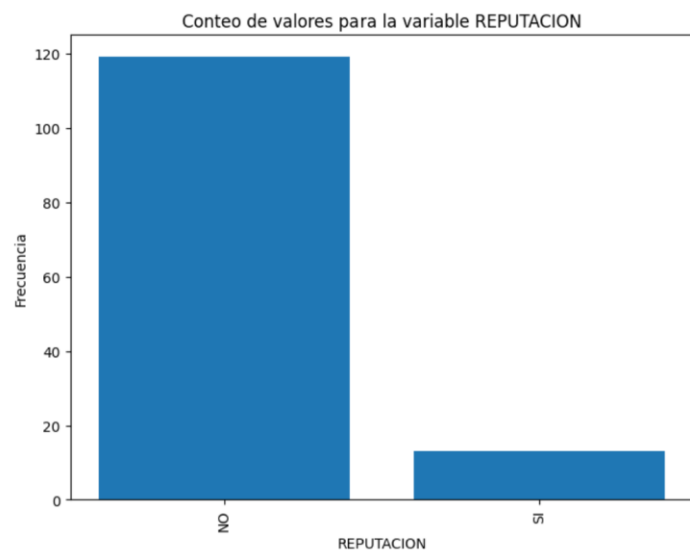
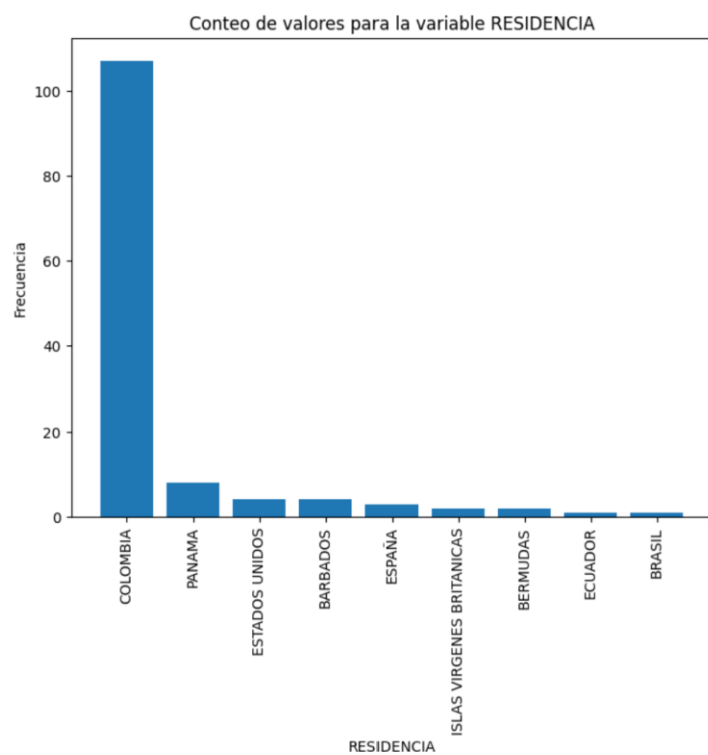


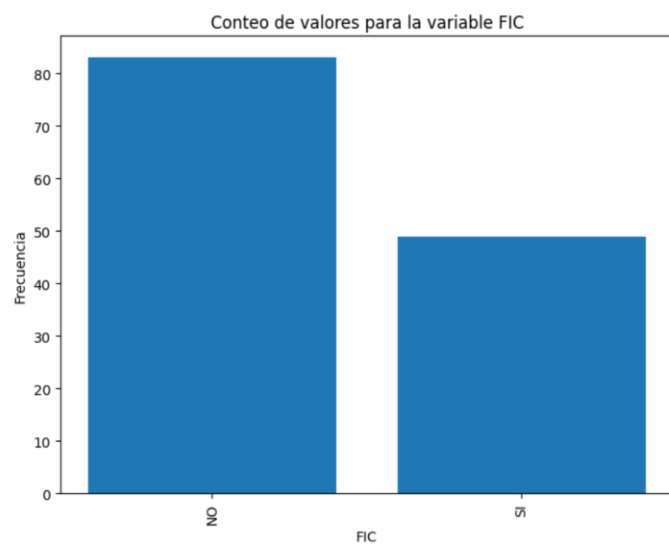
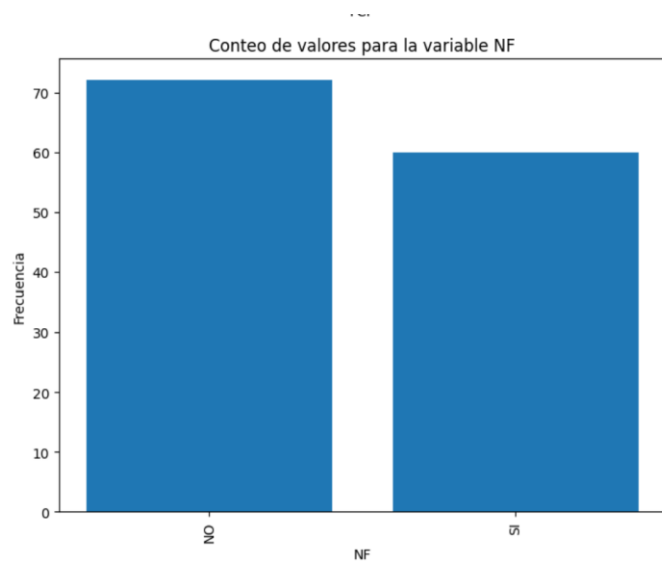
B

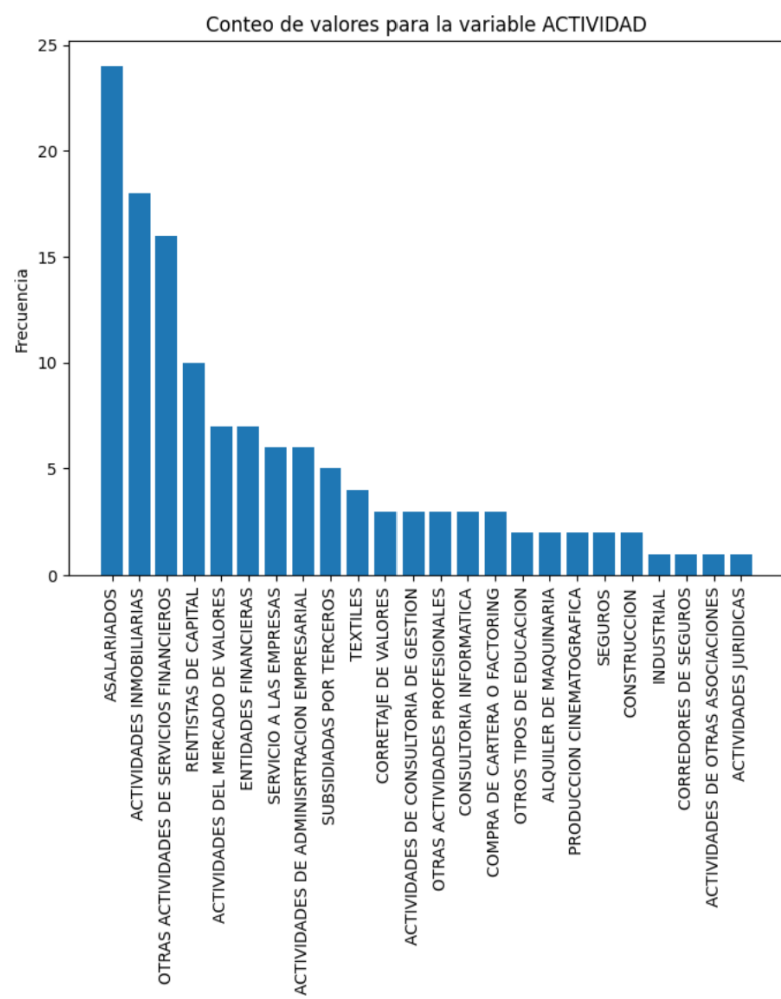


C

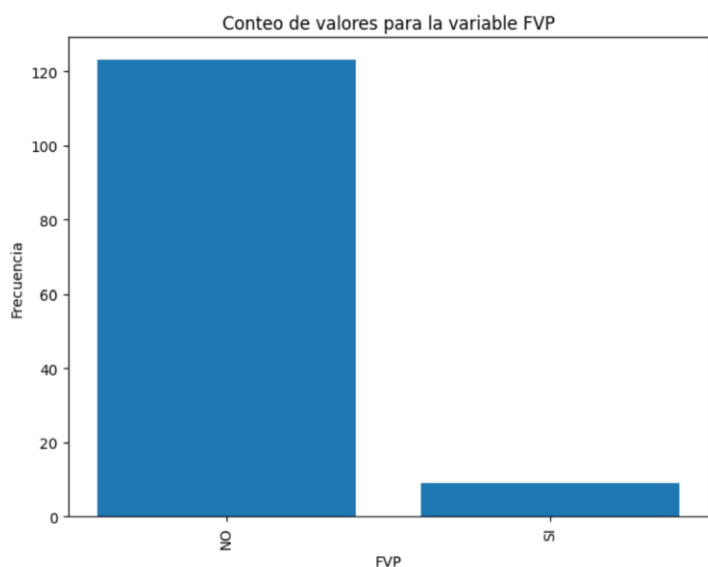


D*E*

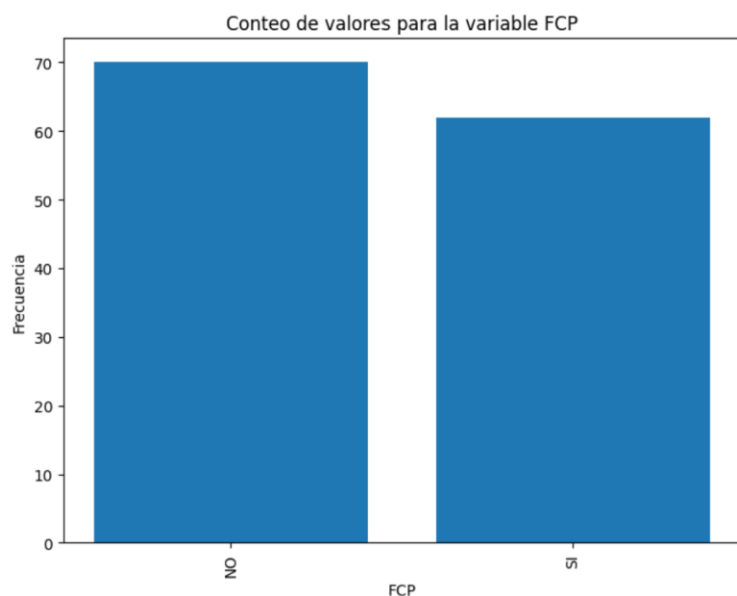
F*G*

H

I



J



Fuente: Elaboración propia

Residencia: La mayoría de los clientes (83.59%) tienen residencia en Colombia, lo cual se refleja en un total de 107 clientes. Panamá también cuenta con un número significativo de clientes, con un 6.25% del total (8 clientes). Los demás países tienen una presencia más reducida, como Estados Unidos y Barbados con 4 clientes cada uno (3.13%), y España, Islas

Vírgenes Británicas y Bermudas con 3 clientes en total (2.34%). Ecuador y Brasil cuentan con solo 1 cliente cada uno (0.78%).

Departamento: La mayoría de los clientes (65.63%) provienen de Bogotá, lo cual representa un total de 84 clientes. Otros departamentos tienen una representación más baja, como Cauca con 9 clientes (7.03%), Cali con 5 clientes (3.91%), Medellín con 3 clientes (2.34%), y Cundinamarca y Putumayo con 2 clientes cada uno (1.56%). El resto de los departamentos tienen solo 1 cliente cada uno, como Bucaramanga, Boyacá y Norte de Santander (0.78%).

PEP (Persona Expuesta Políticamente): La gran mayoría de los clientes (87.22%) no son considerados PEP, lo cual se refleja en 116 clientes. Por otro lado, hay un número menor pero significativo de clientes (12.08%) que sí son PEP, con un total de 16 clientes.

Reputación: La mayoría de los clientes (90.15%) tienen una reputación positiva, representando un total de 119 clientes. Por otro lado, hay un número menor de clientes (9.85%) con una reputación negativa por coincidencia en noticias adversas de LAFT, con un total de 13 clientes.

Actividad económica: Entre las actividades económicas de los clientes, se observa que el 18.18% son asalariados, lo cual representa 24 clientes. Las actividades inmobiliarias también tienen una representación significativa, con el 13.64% del total (18 clientes). Otras actividades de servicios financieros representan el 12.12% (16 clientes), mientras que los rentistas de capital conforman el 7.58% (10 clientes). Otras actividades tienen una presencia menor, como las actividades del mercado de valores y las entidades financieras, ambas con el 5.30% (7 clientes cada una).

Canal: El canal presencial cuenta con la mayoría de los clientes, representando el 58.82% del total (76 clientes). El canal virtual tiene el 26.92% de los clientes (35 clientes), mientras que el canal referido representa el 16.23% (21 clientes).

FIC, FVP, FCP Y NF: Los resultados muestran que el 37.17% de los clientes cuentan con el Fondo de Inversión Colectiva (FIC) activo, mientras que el 62.83% no lo tienen activo. En cuanto al Fondo Voluntario de Pensiones (FVP), solo el 6.82% de los clientes lo tienen activo (9 clientes), mientras que el 93.18% no lo tienen activo. Para el Fondo de Capital Privado (FCP), el 46.97% de los clientes lo tienen activo, en comparación con el 53.03% que no lo tienen activo. Por último, en relación con los negocios fiduciarios (NF), el 45.45% de los clientes tienen activos en este producto, mientras que el 54.55% no los tienen activos. Estos resultados reflejan la proporción de clientes que tienen activos cada uno de los productos de la fiduciaria, mostrando una mayor presencia en los productos de fondo de capital privado y los negocios fiduciarios, mientras que el FVP y el FIC tienen una presencia más baja.

Transformación de la data (proceso y código detallado en la hoja 44 en el apartado transformación de la data), a continuación, se presenta el resultado de la nueva data, la cual se utilizará para ejecutar los algoritmos de segmentación seleccionados:

Ilustración 10*Características y estructura del conjunto de datos transformado*

#	Column	Non-Null Count	Dtype
---	-----	-----	-----
0	PEP	132 non-null	int64
1	REPUTACION	132 non-null	int64
2	APORTE	132 non-null	float64
3	FIC	132 non-null	int64
4	FVP	132 non-null	int64
5	FCP	132 non-null	int64
6	NF	132 non-null	int64
7	INGRESOS	132 non-null	float64
8	PATRIMONIO	132 non-null	float64
9	COLOMBIA	132 non-null	int64
10	ECUADOR	132 non-null	int64
11	ESPAÑA	132 non-null	int64
12	ESTADOS UNIDOS	132 non-null	int64
13	PANAMA	132 non-null	int64
14	ISLAS VIRGENES BRITANICAS	132 non-null	int64
15	BARBADOS	132 non-null	int64
16	BERMUDAS	132 non-null	int64
17	BRASIL	132 non-null	int64
18	BOGOTA	132 non-null	int64
19	nan	132 non-null	int64
20	CAUCA	132 non-null	int64
21	BUCARAMANGA	132 non-null	int64
22	CUNDINAMARCA	132 non-null	int64
23	PUTUMAYO	132 non-null	int64

24	BOYACA	132 non-null	int64
25	NORTE DE SANTANDER	132 non-null	int64
26	CALI	132 non-null	int64
27	MEDELLIN	132 non-null	int64
28	ACTIVIDADES_INMOBILIARIAS	132 non-null	int64
29	ASALARIADOS	132 non-null	int64
30	PRODUCCION_CINEMATOGRAFICA	132 non-null	int64
31	OTROS_TIPOS_DE_EDUCACION	132 non-null	int64
32	SUBSIDIADAS_POR_TERCEROS	132 non-null	int64
33	SERVICIO_A_LAS_EMPRESAS	132 non-null	int64
34	RENTISTAS_DE_CAPITAL	132 non-null	int64
35	ENTIDADES_FINANCIERAS	132 non-null	int64
36	ACTIVIDADES_DE_CONSULTORIA_DE_GESTION	132 non-null	int64
37	ACTIVIDADES_DE_OTRAS_ASOCIACIONES	132 non-null	int64
38	SEGUROS	132 non-null	int64
39	CONSULTORIA_INFORMATICA	132 non-null	int64
40	OTRAS_ACTIVIDADES_DE_SERVICIOS_FINANCIEROS	132 non-null	int64
41	CORRETAJE_DE_VALORES	132 non-null	int64
42	ALQUILER_DE_MAQUINARIA	132 non-null	int64
43	ACTIVIDADES_DEL_MERCADO_DE_VALORES	132 non-null	int64
44	COMPRA_DE_CARTERA_O_FACTORING	132 non-null	int64
45	TEXTILES	132 non-null	int64
46	ACTIVIDADES_DE_ADMINISTRACION_EMPRESARIAL	132 non-null	int64
47	INDUSTRIAL	132 non-null	int64
48	CORREDORES_DE_SEGUROS	132 non-null	int64
49	OTRAS_ACTIVIDADES_PROFESIONALES	132 non-null	int64
50	ACTIVIDADES_JURIDICAS	132 non-null	int64
51	CONSTRUCCION	132 non-null	int64
52	PRESENCIAL	132 non-null	int64
53	VIRTUAL	132 non-null	int64
54	REFERIDO	132 non-null	int64
dtypes: float64(3), int64(52)			

Fuente: Elaboración propia

El resumen de la tabla muestra información sobre el conjunto de datos transformado, que incluye un total de 132 entradas correspondientes a los clientes con corte al 31 de marzo de 2023. Se han considerado 55 variables después de realizar la transformación de los datos, con la mayoría de las columnas siendo de tipo entero (int64), excepto las variables 2, 7 y 8 que son de tipo float (float64). Cada columna representa distintas características o atributos de los datos recopilados de cada uno de los clientes.

Una vez transformada la data, se procedió a ejecutar los siguientes códigos para el método del codo y así identificar el número óptimo de cada clúster:

Para k-medoids:

```
from sklearn_extra.cluster import KMedoids
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt
import numpy as np

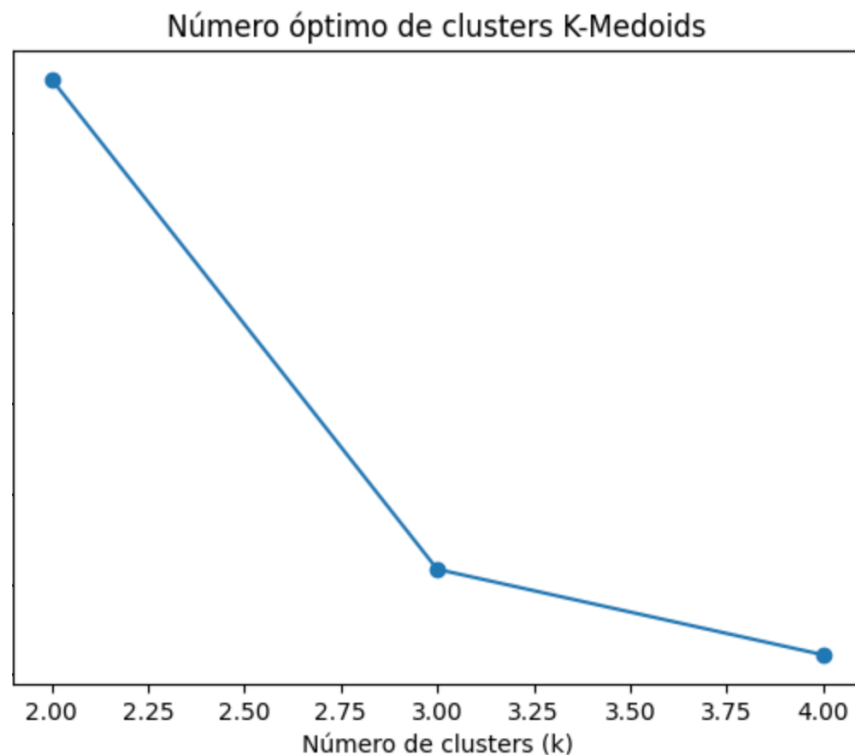
# Crear una lista para almacenar los valores de la silueta promedio
silhouette_scores = []

# Calcular el valor de la silueta promedio para diferentes valores de k
for k in range(2, 5):
    kmedoids = KMedoids(n_clusters=k, metric='manhattan', random_state=0)
    kmedoids.fit(data)
    labels = kmedoids.labels_
    silhouette_avg = silhouette_score(data, labels, metric='manhattan')
    silhouette_scores.append(silhouette_avg)

# Graficar el valor de la silueta promedio
plt.plot(range(2, 5), silhouette_scores, marker='o')
plt.title('Número óptimo de clusters K-Medoids')
plt.xlabel('Número de clusters (k)')
plt.ylabel('Silueta promedio')
plt.show()
```


Ilustración 11

Número óptimo de clusters K-Medoids



Fuente: Elaboración propia

Para agrupamiento jerárquico:

```
from scipy.cluster.hierarchy import linkage, fcluster
from sklearn.metrics import pairwise_distances, silhouette_score
import matplotlib.pyplot as plt

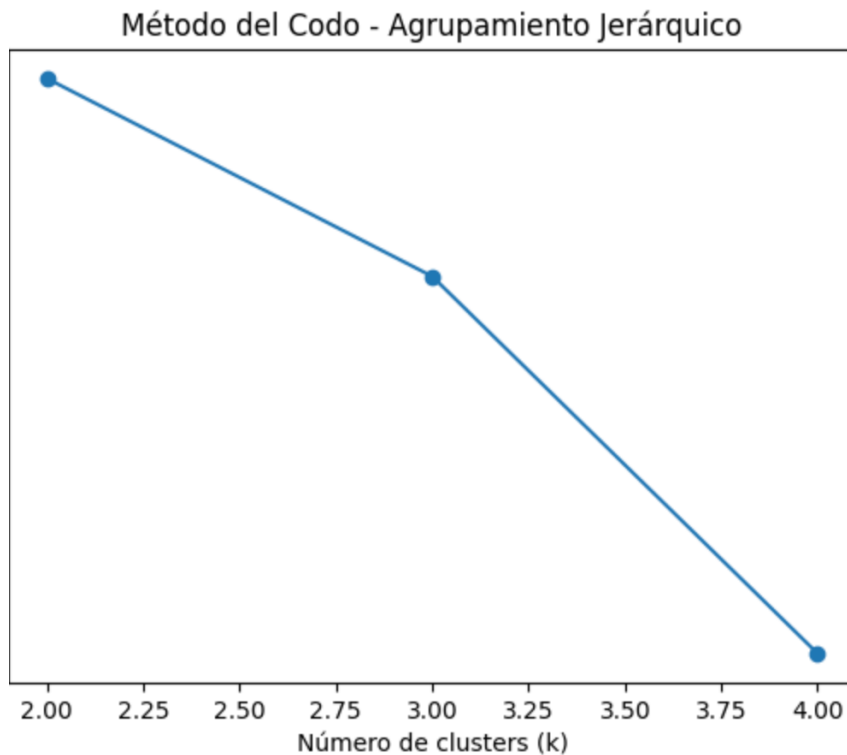
# Calcular la matriz de distancias
distances = pairwise_distances(data, metric='manhattan')

k_values = range(2, 5)
silhouette_scores = []
for k in k_values:
    linkage_matrix = linkage(distances, method='complete')
    labels = fcluster(linkage_matrix, k, criterion='maxclust')
    silhouette_scores.append(silhouette_score(distances, labels))
```

```
# Graficar el método del codo
plt.plot(k_values, silhouette_scores, marker='o')
plt.xlabel('Número de clusters (k)')
plt.ylabel('Índice de Silueta')
plt.title('Método del Codo - Agrupamiento Jerárquico')
plt.show()
```

Ilustración 12

Método del Codo Agrupamiento jerárquico



Fuente: Elaboración propia

Para Fuzzy C-Means

```
import matplotlib.pyplot as plt

# Calcular la matriz de distancias o similaridad
distances = pairwise_distances(data, metric='manhattan')
```

```

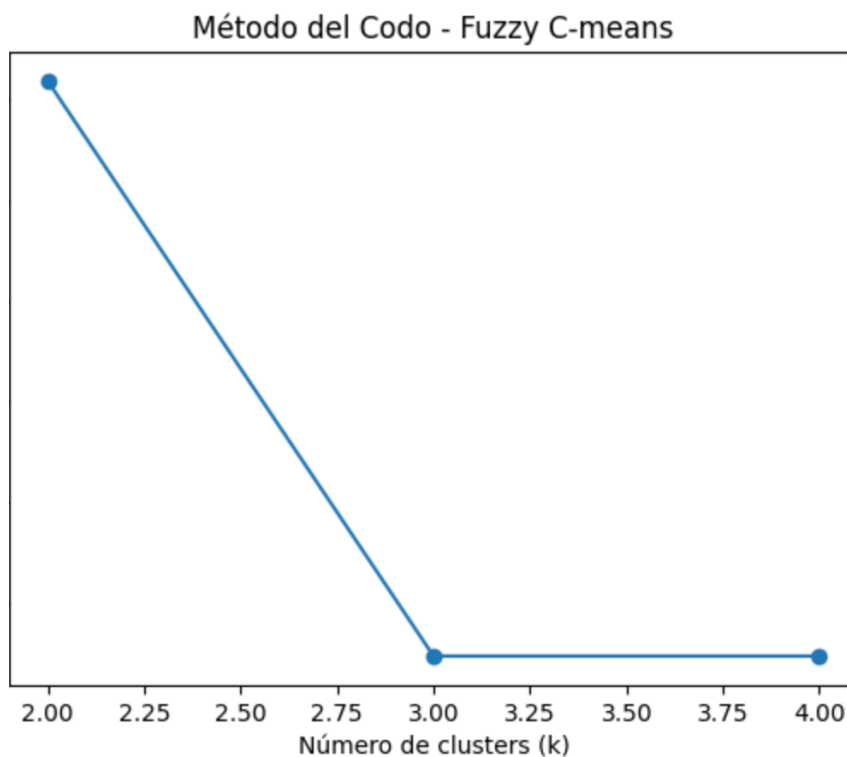
k_values = range(2, 5)
silhouette_scores = []
for k in k_values:
    cntr, u, u0, d, jm, p, fpc = fuzz.cluster.cmeans(data.T, c=k, m=2,
error=0.005, maxiter=1000)
    labels = np.argmax(u, axis=0)
    silhouette_scores.append(silhouette_score(distances, labels))

# Graficar el método del codo
plt.plot(k_values, silhouette_scores, marker='o')
plt.xlabel('Número de clusters (k)')
plt.ylabel('Índice de Silueta')
plt.title('Método del Codo - Fuzzy C-means')
plt.show()

```

Ilustración 13

Método del Codo – Fuzzy C-means



Fuente: Elaboración propia

7.3.1. Ejecución de los Algoritmos de Segmentación en Python

La ejecución se realizó mediante los siguientes códigos

Para K.Medoids

```
from sklearn_extra.cluster import KMedoids
from sklearn.metrics import pairwise_distances

# Calcular la matriz de distancias
distances = pairwise_distances(data, metric='manhattan')

# Configurar y ejecutar K-medoids
k = 3 # Número óptimo de clusters
kmedoids = KMedoids(n_clusters=k, metric='precomputed', random_state=0)
kmedoids.fit(distances)

# Obtener los medoides y las etiquetas de los clusters
medoids = kmedoids.cluster_centers_
labels = kmedoids.labels_

plt.show()
print (labels)

# Contar el total de clientes en cada cluster
conteo_clusters = np.bincount(labels)

# Calcular el porcentaje de clientes en cada cluster
total_clientes = len(labels)
porcentajes_clusters = conteo_clusters / total_clientes * 100

# Imprimir el número y el porcentaje de clientes en cada cluster
for cluster, cantidad, porcentaje in zip(range(len(conteo_clusters)),
    conteo_clusters, porcentajes_clusters):
    print(f"Cluster {cluster}: {cantidad} clientes ({porcentaje:.2f}%)")
```

Resultado de la segmentación de K-Medoids:

Ilustración 14

Resultado segmentación K-Medoids

```
[0 2 2 2 1 2 2 2 0 0 2 2 2 2 2 2 2 2 2 2 2 2 2 1 0 2 1 2 2 1 1 0 0
 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 1 2 2 2 2 2 2 2 2 2 0 0 2 2 2 2 0 0 0 1 2 0
 0 1 0 0 2 0 2 1 1 1 1 2 0 1 1 0 0 0 0 1 0 0 2 2 0 0 0 1 2 0 2 0 1 0 1 0 0
 1 0 2 2 2 2 2 2 2 2 2 2 2 2 2 1 0 2 2 0 0]
```

```
Cluster 0: 35 clientes (26.52%)
Cluster 1: 22 clientes (16.67%)
Cluster 2: 75 clientes (56.82%)
```

Fuente: Elaboración propia

Para Agrupamiento Jerárquico

```
from scipy.cluster.hierarchy import linkage, fcluster
from sklearn.metrics import pairwise_distances

# Calcular la matriz de distancias
distances = pairwise_distances(data, metric='manhattan')

# Calcular el enlace jerárquico
linkage_matrix = linkage(distances, method='complete')

# Obtener las etiquetas de los clusters para el número óptimo de clusters
(3 en este caso)
numero_optimo_clusters = 3
labels = fcluster(linkage_matrix, numero_optimo_clusters,
criterion='maxclust')

plt.show()
print (labels)

# Contar el total de clientes en cada cluster
conteo_clusters = np.bincount(labels_ag)

# Calcular el porcentaje de clientes en cada cluster
total_clientes = len(labels_ag)
porcentajes_clusters = conteo_clusters / total_clientes * 100
```

```

# Imprimir el número y el porcentaje de clientes en cada cluster
for cluster, cantidad, porcentaje in zip(range(len(conteo_clusters)),
conteo_clusters, porcentajes_clusters):
    print(f"Cluster {cluster}: {cantidad} clientes ({porcentaje:.2f}%)")

# Visualizar el número de clientes en cada cluster
plt.figure()
plt.bar(range(len(conteo_clusters)), conteo_clusters)
plt.xlabel("Cluster")
plt.ylabel("Número de clientes")
plt.title("Número de clientes en cada cluster")
plt.show()

```

Resultado de la segmentación de Agrupamiento Jerárquico

Ilustración 15

Resultado segmentación Agrupamiento Jerárquico

[4 4 4 4 4 1 4 3 3 3 3 1 2 4 2 4 4
4 2 3 2 4 4 2 2 4 4 4 4 2 4 4 4 4 4 2 2 2 4 4 1 4 4 1 4 4 4 4 2 3 1 1 3 3
2 2 2 3 3 2 1 1 2 1 3 3 3 2 3 3 3 3 2 3 2 2 3 2 2 2 2 2 1 3 3 3 1 2 1 2 2
3 3 1 4 4 4 1 4 4 1 4 4 4 4 1 3 2 1 1 1 1 1]

Cluster 1: 20 clientes (15.15%)
Cluster 2: 29 clientes (21.97%)
Cluster 3: 25 clientes (18.94%)
Cluster 4: 58 clientes (43.94%)

Fuente: Elaboración propia

Para Fuzzy C-Means

```
import numpy as np
from sklearn.metrics import pairwise_distances
from fcmeans import FCM

# Convertir el DataFrame de Pandas a una matriz ndarray
data_array = data.to_numpy()

# Calcular la matriz de distancias
distances = pairwise_distances(data_array, metric='manhattan')

# Transponer la matriz de distancias
distances_transposed = distances.T

# Configurar y ejecutar Fuzzy C-means
fcm = FCM(n_clusters=3)
fcm.fit(distances_transposed)

# Obtener las etiquetas de los clusters
labels = fcm.predict(distances_transposed)

plt.show()
print (labels)

# Contar el total de clientes en cada cluster
conteo_clusters = np.bincount(labels_fcm)

# Calcular el porcentaje de clientes en cada cluster
total_clientes = len(labels_fcm)
porcentajes_clusters = conteo_clusters / total_clientes * 100

# Imprimir el número y el porcentaje de clientes en cada cluster
for cluster, cantidad, porcentaje in zip(range(len(conteo_clusters)),
    conteo_clusters, porcentajes_clusters):
    print(f"Cluster {cluster}: {cantidad} clientes ({porcentaje:.2f}%")
```

Resultado de la segmentación de Fuzzy C-Means

Ilustración 16

Resultado segmentación Fuzzy C-Means

```
[2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 0 2 2 0 0 0 0 1 1 0 0 0 0
 2 1 0 0 0 2 0 1 2 2 2 2 1 0 0 2 0 2 1 1 1 2 2 1 2 2 1 2 2 2 2 1 0 1 1 0 0
 1 1 1 0 0 1 1 1 0 1 0 0 0 0 0 0 0 0 1 0 1 1 0 1 1 1 1 1 1 0 0 0 1 1 1 0 1
 0 0 1 0 0 0 1 2 2 1 0 0 0 0 1 0 1 1 1 1 1]
```

```
Cluster 0: 43 clientes (32.58%)
Cluster 1: 43 clientes (32.58%)
Cluster 2: 46 clientes (34.85%)
```

Fuente: Elaboración propia

Los resultados de la ejecución de las gráficas de distribución de cluster para cada algoritmo, se presentan mediante los siguientes códigos:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn_extra.cluster import KMedoids
from sklearn.metrics import pairwise_distances
from scipy.cluster.hierarchy import linkage, fcluster
from fcmeans import FCM

# Crear una figura con subplots
fig, axes = plt.subplots(2, 2, figsize=(12, 8))

# Gráfico para K-medoids
cluster_counts_km = np.bincount(labels)
axes[0, 0].bar(range(len(cluster_counts_km)), cluster_counts_km)
axes[0, 0].set_xticks(range(len(cluster_counts_km))) # Establecer los
ticks del eje x
axes[0, 0].set_xlabel('Cluster')
axes[0, 0].set_ylabel('Número de observaciones')
axes[0, 0].set_title('Distribución de clusters (K-medoids)')

# Gráfico para Fuzzy C-means
cluster_counts_fcm = np.bincount(labels_fcm)
```



```

axes[0, 1].bar(range(len(cluster_counts_fcm)), cluster_counts_fcm)
axes[0, 1].set_xticks(range(len(cluster_counts_fcm))) # Establecer los
ticks del eje x
axes[0, 1].set_xlabel('Cluster')
axes[0, 1].set_ylabel('Número de observaciones')
axes[0, 1].set_title('Distribución de clusters (Fuzzy C-means)')

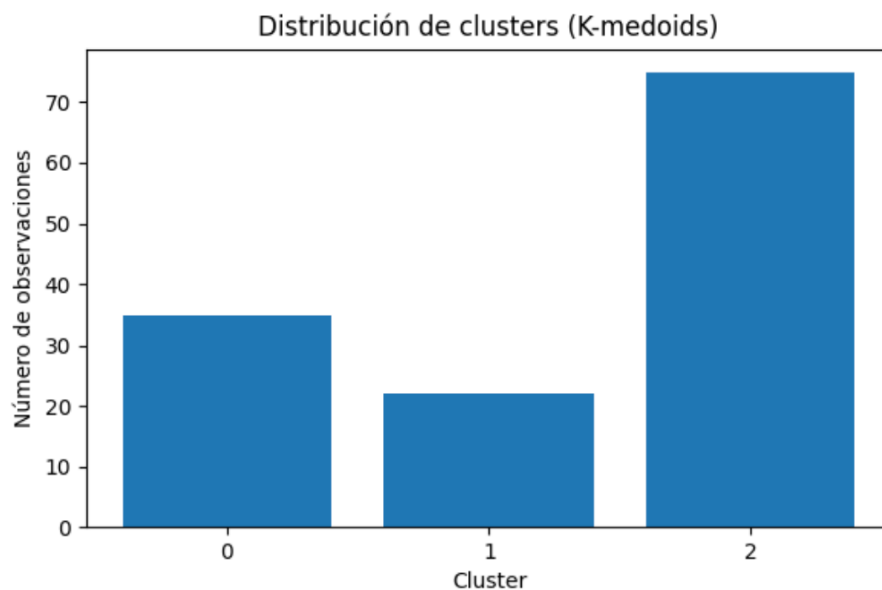
# Gráfico para Agrupamiento Jerárquico
cluster_counts_ag = np.bincount(labels_ag)
axes[1, 0].bar(range(1, len(cluster_counts_ag)), cluster_counts_ag[1:])
axes[1, 0].set_xticks(range(1, len(cluster_counts_ag))) # Establecer los
ticks del eje x
axes[1, 0].set_xlabel('Cluster')
axes[1, 0].set_ylabel('Número de observaciones')
axes[1, 0].set_title('Distribución de clusters (Agrupamiento Jerárquico)')

plt.tight_layout()
plt.show()

```

Ilustración 10

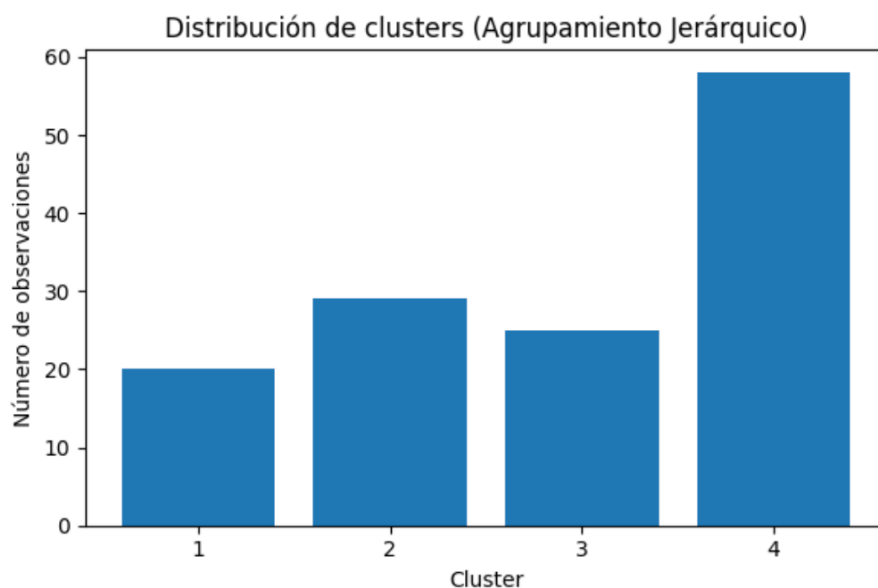
Gráfica de distribución de clusters (K-medoids)



Fuente: Elaboración propia

Ilustración 11

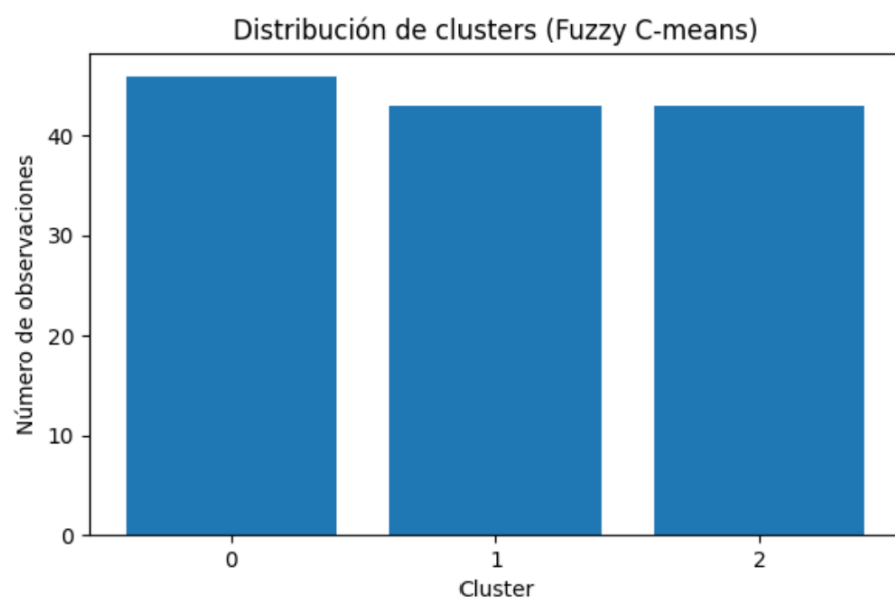
Gráfica de distribución de clusters (agrupamiento jerárquico)



Fuente: Elaboración propia

Ilustración 17

Gráfica de distribución de clusters (Fuzzy C-means)



Fuente: Elaboración propia

De acuerdo con los gráficos, se observan diferencias significativas en la segmentación entre los tres algoritmos utilizados: K-medoids, Algoritmo de Agrupamiento Jerárquico Fuzzy C-means. A continuación, se presenta un análisis de cada algoritmo:

K-medoids:

En la segmentación realizada mediante K-medoids, se identificaron tres grupos distintos con los valores 0, 1 y 2. La distribución de los clientes en estos grupos es la siguiente: el grupo 0 representa un 26% de la muestra, con un total de 35 clientes; el grupo 1 muestra una presencia moderada, con un 16.67%; y el grupo 2 es el más predominante, abarcando el 56.82% de los clientes (75 en total). Esta segmentación se fundamenta en el análisis de las características y atributos de los datos..

Agrupamiento jerárquico:

En el análisis de segmentación realizado mediante el agrupamiento jerárquico, se identificaron cuatro grupos con los valores 1, 2, 3 y 4. La distribución de los clientes en estos grupos es la siguiente: el grupo 4 se destaca como el más predominante, abarcando un 43.94% del total de clientes, lo que equivale a un total de 58 clientes; le sigue el grupo 2 con un 21.97%. Por otro lado, los grupos 1 y 3 muestran una presencia menos predominante, con un 15.15% y 18.94% respectivamente. Este enfoque jerárquico en la segmentación se basa en las características y atributos de los datos, organizando los grupos de acuerdo con niveles de similitud.

Fuzzy C-Means:

En la segmentación realizada con Fuzzy C-Means, se obtuvieron 3 grupos identificados por los valores 0, 1 y 2. Al igual que en el caso de K-medoids, el grupo 2 muestra la mayor presencia con un 34.85% del total de clientes. Los grupos 1 y 0 presentan una presencia igual,

representando cada uno un 32.58% del total. Esto indica que los clientes se distribuyen de manera equitativa entre los grupos 1 y 0. Fuzzy C-Means asigna un grado de pertenencia a cada cliente en cada grupo, lo que indica la probabilidad de que un cliente pertenezca a un grupo determinado. Esto permite una asignación más flexible y probabilística de los clientes a los grupos.

En resumen, los tres algoritmos de segmentación han identificado grupos similares en cuanto a la distribución de los clientes, con algunas diferencias en la asignación específica de los clientes a los grupos.

8. Evaluación

8.1. Indicadores para Evaluar Modelos de Segmentación

De acuerdo con Hastie, Trevor et al. (2001), en su libro *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, presentan varios indicadores comunes para evaluar modelos de segmentación. A continuación, se describen brevemente algunos de ellos ⁴:

Índice de Calinski-Harabasz (Calinski-Harabasz Index): Calcula la relación entre la dispersión dentro de los clúster y entre ellos. Un valor más alto indica una segmentación más compacta y bien separada.

Coefficiente de Silueta (Silhouette Coefficient): Mide la calidad de la segmentación considerando la cohesión intra-cluster y la separación inter-cluster. Un valor más cercano a 1 indica una segmentación sólida, mientras que un valor cercano a -1 es deficiente.

Índice Dunn: Evalúa la separación entre los clusters y la compactación de cada clúster individualmente. Un valor más alto indica una mejor separación entre los clusters y una mayor compacidad dentro de cada uno.

Índice de Rand ajustado (Adjusted Rand Index, ARI): Mide la similitud entre las asignaciones de clúster verdaderas y las asignaciones de clúster generadas por el modelo. Un valor más cercano a 1 indica una mayor concordancia entre las asignaciones.

Índice de Entropía Normalizada (Normalized Mutual Information, NMI): Mide la similitud entre las asignaciones de clúster verdaderas y las asignaciones de clúster generadas por el modelo. Un valor más cercano a 1 indica una mayor concordancia entre las asignaciones.

⁴ Se adaptaron los conceptos de indicadores de evaluación del libro titulado *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* al español.

8.2. Selección de los Indicadores

En este estudio de segmentación, se seleccionaron dos indicadores para evaluar los resultados de los tres algoritmos utilizados: Coeficiente de Silueta y el Índice Dunn. Estos indicadores fueron elegidos debido a su capacidad para proporcionar una evaluación completa de la calidad de la segmentación y su capacidad para capturar diferentes aspectos de los grupos formados.

El Coeficiente de Silueta es una medida que evalúa la calidad de la segmentación considerando la cohesión intra-cluster y la separación inter-cluster. Este indicador proporciona una medida de cuán bien agrupados están los datos y cuánto se solapan entre sí los diferentes grupos. En el contexto de los algoritmos seleccionados y basados en la métrica de distancia de Manhattan, el Coeficiente de Silueta resulta especialmente relevante, ya que utiliza las distancias entre los puntos de datos para su cálculo. La distancia de Manhattan es apropiada para medir la similitud entre puntos en espacios donde se consideran las diferencias en cada dimensión de manera independiente. Por lo tanto, el Coeficiente de Silueta permite evaluar la calidad de los agrupamientos en función de esta métrica y determinar qué tan bien separados están los grupos en el espacio definido por la distancia de Manhattan.

El Coeficiente de Silueta se calcula para cada punto de datos i en función de la distancia promedio a los puntos del mismo clúster ($a(i)$) y la distancia promedio a los puntos del clúster vecino más cercano ($b(i)$).

La fórmula general del Coeficiente de Silueta para un punto de datos i es la siguiente:

$$S(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}$$

Donde:

$a(i)$: Es la distancia promedio de i a todos los demás puntos dentro del mismo clúster.

Cuanto más pequeña sea esta distancia, más cohesión habrá dentro del clúster.

$b(i)$: Es la distancia promedio de i a todos los puntos del clúster vecino más cercano.

Cuanto mayor sea esta distancia, más separación habrá entre clústeres.

El numerador de la fórmula $(b(i) - a(i))$ representa la diferencia entre la distancia promedio hacia puntos del clúster vecino más cercano y la distancia promedio hacia puntos del mismo clúster. El denominador $\max\{a(i), b(i)\}$ se utiliza para normalizar el resultado, asegurándose de que el Coeficiente de Silueta esté en el rango de -1 a 1.

Por otro lado, el Índice Dunn es una medida que evalúa la separación entre los clusters y la compacidad de cada clúster individualmente. Este indicador proporciona una medida de cuán bien definidos están los grupos y qué tan separados están entre sí. En el contexto de los algoritmos seleccionados con la métrica de distancia de Manhattan, el Índice Dunn también es relevante, ya que considera las distancias entre los puntos de datos en su cálculo. La distancia de Manhattan se adapta bien a este indicador, ya que se centra en la suma de las distancias entre los puntos de diferentes clústeres y la mínima distancia intra-cluster. Dado que la métrica de distancia de Manhattan considera las diferencias en cada dimensión de manera independiente, es apropiada para el cálculo del Índice Dunn. Por lo tanto, este indicador permite evaluar qué tan bien separados están los grupos y qué tan compactos son en función de la métrica de distancia de Manhattan.

El Índice Dunn se basa en la relación entre la mínima distancia entre dos clústeres (d_{\min}) y la máxima distancia dentro de un clúster (d_{\max}).

La fórmula del Índice Dunn es la siguiente:

$$Dunn = d_{min} / d_{max}$$

Donde:

d_{min} : Es la distancia mínima entre cualquier par de puntos pertenecientes a clústeres diferentes. Representa la separación entre clústeres. Cuanto mayor sea esta distancia, mejor será la separación entre clústeres.

d_{max} : Es la distancia máxima entre cualquier par de puntos pertenecientes al mismo clúster. Representa la compacidad dentro de cada clúster. Cuanto menor sea esta distancia, más compacto será el clúster.

El Índice Dunn se calcula dividiendo la mínima distancia entre clústeres (d_{min}) por la máxima distancia dentro de un clúster (d_{max}). Un valor más alto del Índice Dunn indica una mejor separación entre clústeres y una mayor compacidad dentro de cada clúster.

Si bien existen otros indicadores disponibles para evaluar la segmentación, estos dos fueron seleccionados por su capacidad para evaluar aspectos clave y complementarios de los resultados, ya que, al utilizar estos indicadores en conjunto, se obtiene una evaluación más completa y precisa de la calidad de los resultados de los algoritmos de segmentación utilizados.

8.3. Resultados de la Evaluación

```
#k-medoids

#Calcular coeficiente de silueta para k-medoids

kmedoids_silhouette = silhouette_score(data,
kmedoids.labels_,metric='manhattan')
print("Coeficiente de Silueta promedio:", kmedoids_silhouette)

# Obtener los medoides y las etiquetas de los clusters
medoids = kmedoids.cluster_centers_
labels = kmedoids.labels_
```



```

# Calcular la mínima distancia entre clústeres (d_min)
d_min_km = np.min(distances_km[np.nonzero(distances_km)])

# Calcular la máxima distancia dentro de un clúster (d_max)
d_max_km = 0.0
for i in range(k):
    cluster_points = data[labels == i]
    cluster_distances = distances_km[labels == i][:, labels == i]
    max_distance =
np.max(cluster_distances[np.nonzero(cluster_distances)])
    if max_distance > d_max_km:
        d_max_km = max_distance

# Calcular el Índice Dunn para KM
dunn_km = d_min_km / d_max_km

print(f'Índice Dunn: {dunn_km}')

#Agrupamiento Jerárquico

#Calcular coeficiente de silueta para Agrupamiento Jerárquico

ag_silhouette = silhouette_score(data, labels_ag, metric='manhattan')
print("Coeficiente de Silueta promedio:", ag_silhouette)

# Obtener los valores únicos de las etiquetas de los clusters
unique_labels_ag = np.unique(labels_ag)

# Calcular la mínima distancia entre clústeres (d_min)
d_min_ag = np.min(distances_ag[np.nonzero(distances_ag)])

# Calcular la máxima distancia dentro de un clúster (d_max)
d_max_ag = 0.0
for label in unique_labels_ag:
    cluster_points = data[labels_ag == label]
    cluster_distances = distances_ag[labels_ag == label][:, labels_ag ==
label]
    max_distance =
np.max(cluster_distances[np.nonzero(cluster_distances)])
    if max_distance > d_max_ag:
        d_max_ag = max_distance

# Calcular el Índice Dunn para AG
dunn_ag = d_min_ag / d_max_ag

```

```

# Imprimir el valor del Índice Dunn
print(f'Índice Dunn: {dunn_ag}')

#Fuzzy C-Means

#Calcular coeficiente de silueta para fuzzy c-means

fcm_silhouette = silhouette_score(data, labels_fcm,metric='manhattan')
print("Coeficiente de Silueta promedio:", fcm_silhouette)

# Calcular la mínima distancia entre clústeres (d_min)
d_min_fcm = np.min(distances_transposed[np.nonzero(distances_transposed)])

# Calcular la máxima distancia dentro de un clúster (d_max)
d_max_fcm = 0.0
for label in np.unique(labels_fcm):
    cluster_points = distances_transposed[labels_fcm == label]
    max_distance = np.max(cluster_points[np.nonzero(cluster_points)])
    if max_distance > d_max_fcm:
        d_max_fcm = max_distance

# Calcular el Índice Dunn para FCM
dunn_fcm = d_min_fcm / d_max_fcm

# Imprimir el valor del Índice Dunn
print(f'Índice Dunn: {dunn_fcm}')

# Crear un diccionario con los resultados de los indicadores
results = {
    'Algoritmo': ['Ag. Jerárquico', 'K-medoids', 'Fuzzy C-means'],
    'Coef Silueta': [ag_silhouette, kmedoids_silhouette, fcm_silhouette],
    'Índice Dunn': [dunn_km, dunn_ag, dunn_fcm],
}

# Crear un DataFrame a partir del diccionario de resultados
df_results = pd.DataFrame(results)

# Mostrar la tabla de resultados
print(df_results)

```

Ilustración 18

Resultado de la Evaluación

	Algoritmo	Coef Silueta	Índice Dunn
0	Ag. Jerárquico	0.678868	2.805448
1	K-medoids	0.783263	3.146715
2	Fuzzy C-means	0.599632	2.738989

Fuente: Elaboración propia

Coeficiente de Silueta:

El algoritmo K-medoids obtuvo el valor más alto de coeficiente de silueta (0.783), lo que indica que los clústeres generados por este algoritmo están bien definidos y tienen una buena separación entre ellos.

El algoritmo Ag. Jerárquico obtuvo un coeficiente de silueta de 0.679, lo que sugiere que también logra una separación razonable de los clústeres, aunque ligeramente inferior al algoritmo K-medoids.

El algoritmo Fuzzy C-means obtuvo el coeficiente de silueta más bajo (0.600), lo que indica que los clústeres generados por este algoritmo tienen una separación menos clara y pueden solaparse en cierta medida.

Índice Dunn:

El algoritmo K-medoids obtuvo el valor más alto de índice Dunn (3.147), lo que indica una mayor separación y claridad entre los clústeres.

El algoritmo Ag. Jerárquico obtuvo un índice Dunn de 2.805, lo que indica una separación razonable entre los clústeres, pero ligeramente menor en comparación con K-medoids.

El algoritmo Fuzzy C-means obtuvo el índice Dunn más bajo (2.739), lo que sugiere que los clústeres generados por este algoritmo pueden tener una mayor superposición y menor separación.

El algoritmo K-medoids muestra un mejor desempeño en términos de separación y claridad de los clústeres, según tanto el coeficiente de silueta como el índice Dunn. Esto indica que K-medoids es la elección más sólida para este conjunto de datos en particular, ya que produce clústeres más distintos y bien definidos. Por lo tanto, se seleccionó este modelo como la mejor opción para llevar a cabo la segmentación en este proyecto, en la siguiente fase se dará a conocer el análisis de cada segmento y cómo cada uno de ellos aporta al objetivo de prevención de lavado de activos.

9. Despliegue

En la fase de despliegue del modelo CRISP-DM, se implementaron los resultados obtenidos en la fase de modelado y evaluación. En este caso, se utilizó el modelo K-medoids como el método seleccionado para la segmentación de los clientes de R4G. El despliegue implicó aplicar este modelo a nuevos conjuntos de datos obtenidos en el proceso de actualización de los clientes, para clasificar y segmentar de manera automática a los clientes, lo que permitió identificar outliers dentro de los 3 clúster.

9.1. Pasos Realizados en el Despliegue:

Preparación de los nuevos datos: Todos los nuevos conjuntos de datos que han ingresado al modelo producto del proceso de actualización están limpios y estructurados de acuerdo con los requisitos del modelo k-medoids. Se están realizando las transformaciones necesarias, como la normalización de variables, para garantizar la consistencia de los datos, (este proceso se ejecuta descargando la base de datos de SIFI y corriendo el código de Python en GoogleColab diariamente).

Implementación del modelo: Se está aplicando el modelo seleccionado a los nuevos datos de manera diaria con el fin de realizar una segmentación automatizada de los clientes de R4G. Gracias a la automatización del proceso, se pudo realizar esta tarea de manera eficiente y precisa, proporcionando resultados consistentes y actualizados en tiempo real.

Clasificación de los clientes e identificación de outliers: Utilizando el modelo k-medoids, se asignó a cada cliente a un grupo específico en función de sus características y similitudes. Esto permitió identificar patrones y outliers dentro de cada grupo mediante el algoritmo Local Outlier Factor (LOF), lo que a su vez facilitó la toma de decisiones estratégicas y acciones personalizadas.

Para identificar los outliers o las inusualidades en cada clúster, se utilizó LOF (Local Outlier Factor), este es un algoritmo de detección de valores atípicos que se utiliza para identificar observaciones inusuales en un conjunto de datos.

La fórmula matemática del LOF se define de la siguiente manera:

$$LOF(p) = (1 / k) * \Sigma (D(p, o) / D(o, k))$$

donde:

p es el punto de datos para el cual se calcula el LOF.

o es un punto de datos vecino de p.

k es el número de vecinos considerados.

$D(p, o)$ es la distancia entre los puntos de datos p y o.

$D(o, k)$ es la distancia entre el punto de datos o y su k-ésimo vecino más cercano.

El LOF se basa en la idea de que un valor atípico se caracteriza por tener una densidad local más baja en comparación con sus vecinos. El LOF calcula una puntuación para cada punto de datos que indica su grado de desviación con respecto a sus vecinos cercanos.

Si $LOF(p) \approx 1$, indica que el punto de datos p es similar en densidad a sus vecinos.

Si $LOF(p) > 1$, indica que el punto de datos p es menos denso que sus vecinos, lo que sugiere que puede ser un valor atípico.

El LOF considera las distancias relativas entre los puntos de datos y sus vecinos más cercanos para calcular el grado de desviación. Al calcular la relación entre las distancias de p a sus vecinos ($D(p, o)$) y las distancias de los vecinos a sus propios vecinos más cercanos ($D(o, k)$), se puede determinar si un punto de datos tiene una densidad local significativamente más baja que sus vecinos.

En el código proporcionado, se ajustó un umbral de 0.05 para determinar los outliers. Este umbral representa el nivel de desviación de densidad local a partir del cual se considera que un punto de datos es un outlier, es un punto de partida común, pero se puede ajustar según las características de los datos y las necesidades del análisis en las calibraciones que se deben realizar semestralmente al modelo.

Este proceso se realizó mediante el siguiente código:

```
# Descargar las librerías necesarias
from sklearn.neighbors import LocalOutlierFactor

# Crear una instancia de LocalOutlierFactor y ajustar el umbral
lof = LocalOutlierFactor(contamination=0.05)
outlier_scores = lof.fit_predict(data)

# Obtener los índices de los valores atípicos
outlier_indices = np.where(outlier_scores == -1)[0]

# Imprimir los índices de los valores atípicos
print(outlier_indices)
```

En esta parte del código se utiliza la clase LocalOutlierFactor del módulo sklearn.neighbors para identificar los valores atípicos (outliers) en los datos. Adicionalmente, se crea una instancia de LocalOutlierFactor y se especifica el parámetro contamination con un valor de 0.05. Este parámetro determina la proporción estimada de valores atípicos en los datos. En este caso, se asume que aproximadamente el 5% de los puntos pueden ser valores atípicos.

Luego, se utiliza el método fit_predict() de LocalOutlierFactor para ajustar el modelo y obtener las predicciones de cada punto de datos. Este método devuelve una etiqueta (-1 para valores atípicos y 1 para valores normales) para cada punto en función de su grado de inusualidad.

A continuación, se utiliza la función `np.where()` para encontrar los índices de los valores atípicos en el arreglo de `outlier_scores`. Se seleccionan aquellos puntos que tienen una etiqueta de -1, lo que indica que son considerados valores atípicos por el algoritmo.

Los resultados obtenidos son:

Ilustración 19

Resultado de los outliers obtenidos mediante Local Outlier Factor

[3 9 10 46 47 62 115]

Fuente: Elaboración propia

Finalmente, se ejecuta el siguiente código para descargar la data de los outliers y así realizar un análisis de patrones de transacciones, comportamiento del cliente y de perfiles de riesgos:

```
# Agregar una columna "Valores Atípicos" a la data
data["Valores Atípicos"] = "No"
data.loc[outlier_indices, "Valores Atípicos"] = "Sí"

# Exportar la data a un archivo CSV
data.to_csv("data_con_valores_atipicos.csv", index=False)
```

Patrones de transacciones: comprobar si hay anomalías en las transacciones, como montos o frecuencia inusual de transacciones o patrones de transferencia irregulares. Analizar anomalías relacionadas con características específicas de la transacción, como los países o terceros involucrados, para identificar posibles riesgos de lavado de dinero o financiamiento del terrorismo.

Comportamiento del cliente: Identificar irregularidades en el desempeño financiero, como cambios significativos en los patrones de transacciones, fluctuaciones importantes en los ingresos o movimientos de efectivo inusuales. Evaluar si los datos anómalos de los clientes pueden indicar un comportamiento de riesgo, como actividad sospechosa o violaciones de políticas internas.

Perfil de riesgo: Identificar si se debe realizar un ajuste en el perfil de riesgo del cliente en el caso que presente datos inusuales, para revisión y ajuste de las medidas de control y seguimiento.

Análisis de riesgo por segmento y variables:

Este análisis se llevó a cabo en Python siguiendo el procedimiento descrito en el numeral 7 del presente documento:

- Se definieron intervalos de clase y sus respectivos niveles de riesgo para cada variable cuantitativa.
- Se asignó el riesgo a los datos utilizando los intervalos de clase definidos y estableciendo el nivel de riesgo correspondiente a cada observación.
- Para cada variable, se verificó el intervalo y se asignó el nivel de riesgo correspondiente.

A continuación se ejecuta lo anteriormente descrito en Python con la data_analisis la cual contiene las variables numéricas originales (INGRESOS, APORTE y PATRIMONIO) y el segmento al cual corresponde cada cliente :

De acuerdo con la tabla 9 “Riesgo ingresos y patrimonio” del presente documento, los intervalos y el score de riesgo por la variable patrimonio son: de 1 a 29.000.000.000, de 29.000.000.001 a 54.000.000.000 y de 54.000.000.001 en adelante, siendo el primero de menor

riesgo, el segundo de riesgo medio y el tercer intervalo de riesgo alto, estos fueron implementados en el siguiente código para arrojar el resultado:

```
# Cargar los datos a analizar desde un archivo Excel

data_analisis = pd.read_excel('/content/data_analisis.xlsx')

# Análisis de las variables cuantitativas en cada segmento
for i in range(3):
    clientes_en_cluster = data_analisis[data_analisis['labels'] == i]

    # Análisis de la variable patrimonio
    patrimonio_promedio = clientes_en_cluster['PATRIMONIO'].mean()

    if patrimonio_promedio > 54000000000:
        riesgo = 'Alto'
    elif patrimonio_promedio > 29000000000:
        riesgo = 'Medio'
    else:
        riesgo = 'Bajo'

    # Resultados del análisis de riesgo
    print("Segmento:", i)
    print("Número de clientes:", len(clientes_en_cluster))
    print("Patrimonio promedio:", patrimonio_promedio)
    print("Nivel de riesgo:", riesgo)
    print("-----")

# Análisis de la variable aporte
aporte_promedio = clientes_en_cluster['APORTE'].mean()

if aporte_promedio > 1400000000:
    riesgo = 'Alto'
elif aporte_promedio > 540000000:
    riesgo = 'Medio'
else:
    riesgo = 'Bajo'

# Resultados del análisis de riesgo
print("Segmento:", i)
print("Número de clientes:", len(clientes_en_cluster))
print("Aporte promedio:", aporte_promedio)
print("Nivel de riesgo:", riesgo)
print("-----")
```

```

# Análisis de la variable ingresos
ingresos_promedio = clientes_en_cluster['INGRESOS'].mean()

if ingresos_promedio > 3500000000:
    riesgo = 'Alto'
elif ingresos_promedio > 2900000000:
    riesgo = 'Medio'
else:
    riesgo = 'Bajo'

# Resultados del análisis de riesgo
print("Segmento:", i)
print("Número de clientes:", len(clientes_en_cluster))
print("Ingresos promedio:", ingresos_promedio)
print("Nivel de riesgo:", riesgo)
print("-----")

```

Ilustración 20

Data para análisis de riesgo

	PATRIMONIO	INGRESOS	APOORTE	labels
0	3000000000	2500000	1200000000	0
1	150000000	500000	4100000	2
2	20000000	120000	770000	2
3	5000000	200000	390000	2
4	2000000000	2345900	1181707816	1
..
127	7458000	4000000	800000000	0
128	371293000	2851500	70000000	2
129	1000109000	62589200	30000000	2
130	68601000	45895000	20000000	0
131	525230764562	66969498655	250000000	0

Fuente: Elaboración propia

Ilustración 21

Análisis de riesgo variable ingresos

```

Segmento: 0
Número de clientes: 35
Ingresos promedio: 2612194896.6571426
Nivel de riesgo: Bajo
-----
Segmento: 1
Número de clientes: 22
Ingresos promedio: 2966608881.090909
Nivel de riesgo: Medio
-----
Segmento: 2
Número de clientes: 75
Ingresos promedio: 3545248299.613333
Nivel de riesgo: Alto
-----

```

Fuente: Elaboración propia

Ilustración 22

Análisis de riesgo variable aporte

```

Segmento: 0
Número de clientes: 35
Aporte promedio: 547342857.1428572
Nivel de riesgo: Medio
-----
Segmento: 1
Número de clientes: 22
Aporte promedio: 1475077628.0
Nivel de riesgo: Alto
-----
Segmento: 2
Número de clientes: 75
Aporte promedio: 536924840.0
Nivel de riesgo: Bajo
-----

```

Fuente: Elaboración propia

Ilustración 23

Análisis de riesgo variable promedio

```

Segmento: 0
Número de clientes: 35
Patrimonio promedio: 31357652676.057144
Nivel de riesgo: Medio
-----
Segmento: 1
Número de clientes: 22
Patrimonio promedio: 54306735595.40909
Nivel de riesgo: Alto
-----
Segmento: 2
Número de clientes: 75
Patrimonio promedio: 28888651510.34667
Nivel de riesgo: Bajo
-----

```

Fuente: Elaboración propia

Basado en los resultados del análisis de los tres segmentos y las variables estudiadas (patrimonio, aporte e ingresos), se realizará un mayor monitoreo en el Segmento 1 y un menor monitoreo en el Segmento 2. A continuación, se justifica esta recomendación:

Mayor monitoreo en el Segmento 1:

El Segmento 1 se caracteriza por tener un nivel de riesgo alto en todas las variables analizadas, esto indica que existe una mayor probabilidad de que se presenten situaciones de riesgo de LA/FT, por lo tanto, un mayor monitoreo en este segmento permitirá detectar tempranamente posibles transacciones inusuales o sospechosas y tomar las medidas adecuadas para prevenir actividades ilícitas.

Dado que el riesgo es alto en 2 las variables y medio en 1, es crucial implementar medidas adicionales de control y supervisión para garantizar el cumplimiento de las regulaciones y salvaguardar la integridad de R4G.

Menor monitoreo en el Segmento 2:

El Segmento 2 presenta un nivel de riesgo bajo en términos de patrimonio y aportes, aunque muestra un nivel de riesgo alto en cuanto a los ingresos, adicionalmente es importante tener en cuenta que los ingresos pueden variar legítimamente debido a diferentes factores, como fuentes de ingresos adicionales o fluctuaciones en el mercado laboral.

Dado que el patrimonio y los aportes son bajos en este segmento, se puede asignar un menor nivel de monitoreo en comparación con el Segmento 1, ya que hay una menor probabilidad de que se presenten actividades de riesgo significativas.

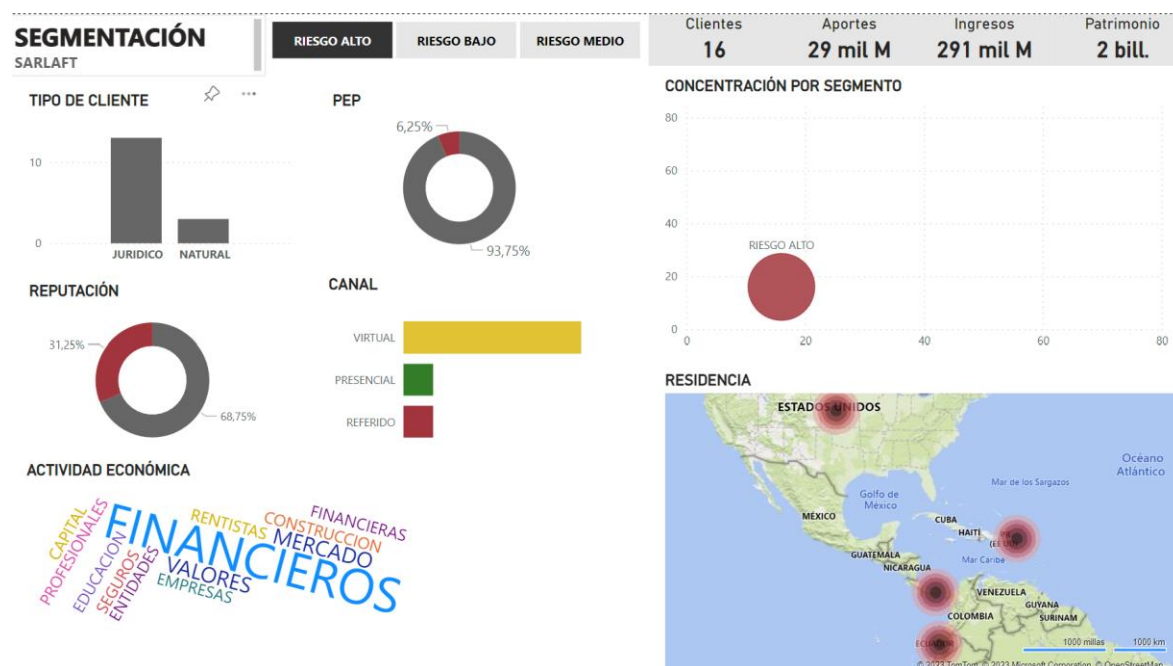
Sin embargo, esto no significa que se deba descuidar por completo el monitoreo en el Segmento 2, por lo tanto, se deben establecer controles adecuados para verificar la legitimidad de los ingresos y detectar posibles actividades sospechosas.

Descripción de los segmentos en función del riesgo:

La implementación de los códigos previamente mencionados ha sido integrada en Power BI con el propósito de llevar a cabo un análisis descriptivo de las variables en relación al riesgo LA/FT, permitiendo así identificar las principales características asociadas a los perfiles de riesgo. Es relevante destacar que esta visualización nos brinda la capacidad de analizar y agrupar las variables en tres niveles de riesgo: bajo, medio y alto. El enfoque principal no se centra en generar una nueva segmentación, sino en describir cada variable en función a su nivel de riesgo correspondiente y agruparlas de acuerdo a dichos niveles. Esta estrategia facilita la identificación y comprensión de las variables clave que influyen en los riesgos asociados a lavado de activos y financiamiento del terrorismo. Mediante esta herramienta, se podrán tomar decisiones informadas y establecer medidas preventivas y de mitigación adecuadas para cada nivel de riesgo identificado.

Ilustración 24

Riesgo alto



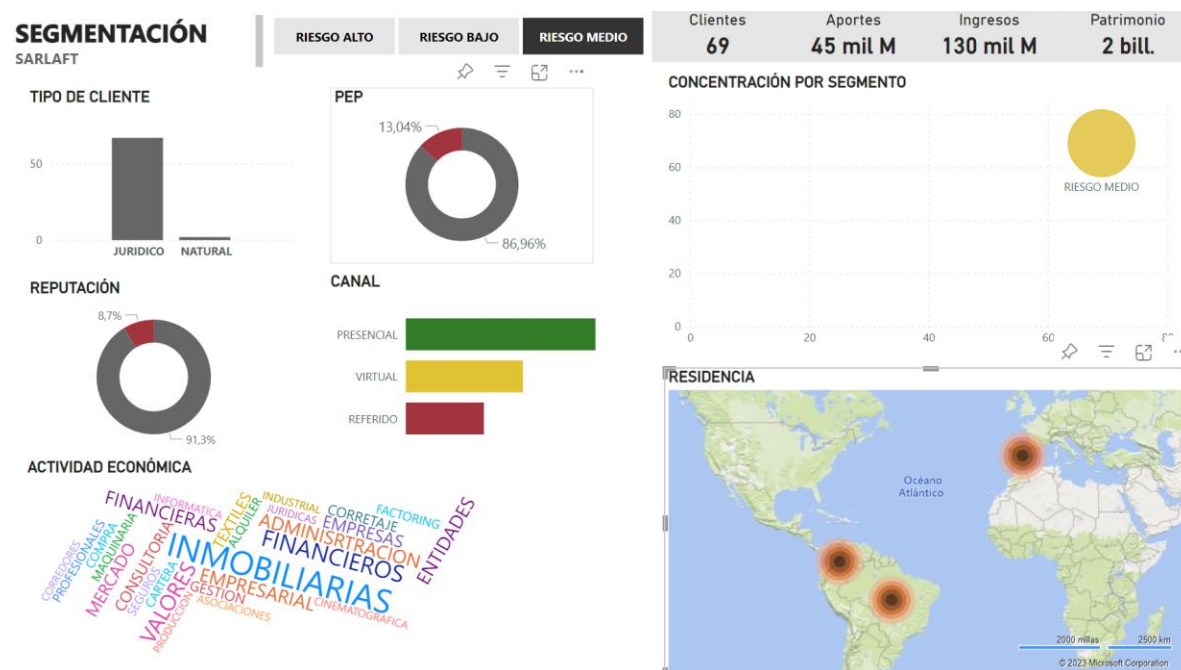
Fuente: Elaboración propia

En el grupo de alto riesgo, se ha identificado un total de 16 clientes, los cuales presentan una serie de características que requieren un monitoreo más riguroso. Entre ellas, se destaca el hecho de que el 31.25% de los clientes tienen una reputación negativa debido a la presencia de noticias relacionadas con actividades sospechosas de LA/FT. Además, se ha observado que 12 de estos clientes fueron vinculados de manera virtual, lo que dificulta la obtención de información completa sobre sus actividades financieras.

Además, es relevante destacar que la concentración de clientes en este grupo se encuentra en países clasificados como de alto riesgo, como Panamá, Ecuador, Bermudas, Barbados e Islas Vírgenes Británicas es alta, por lo tanto, se deben tomar medidas de control ya que estos países son reconocidos como jurisdicciones de alto riesgo en términos de riesgo de LA/FT.

Ilustración 25

Riesgo medio

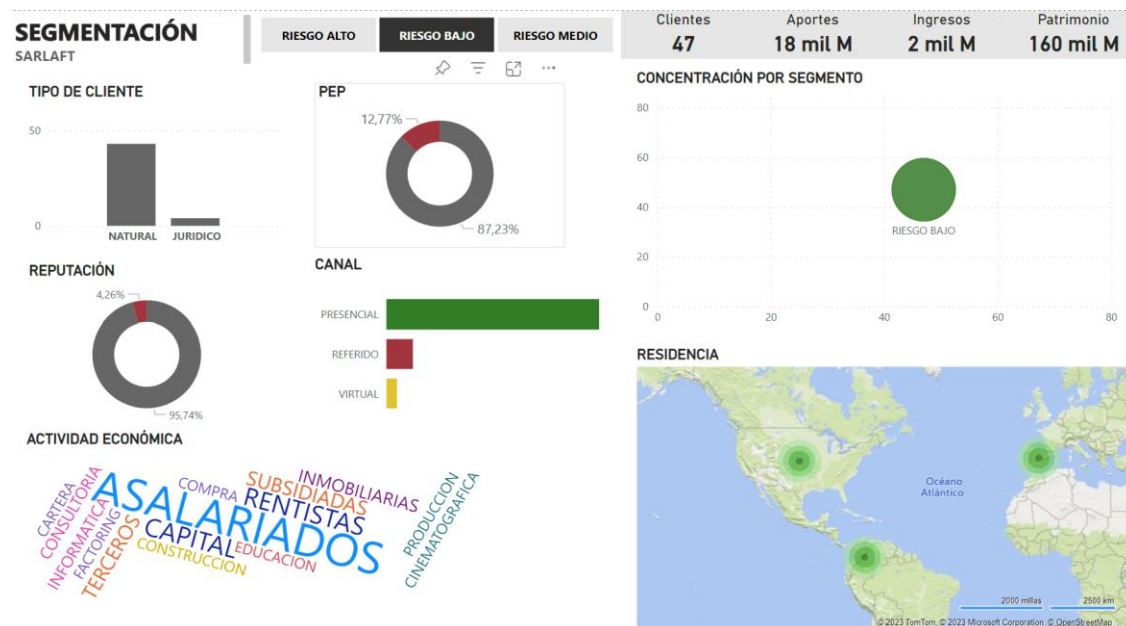


Fuente: Elaboración propia

En el grupo de riesgo medio, se identificaron a 69 clientes, donde predominan las personas jurídicas, es el grupo que mayor movimiento transaccional presentó en las aperturas con un total de 45 mil millones de pesos, adicionalmente es importante resaltar que el 13.04% de los beneficiarios finales de las empresas de este segmento son personas políticamente expuestas, por lo tanto se debe hacer un seguimiento especial a cualquier movimiento de estos PEP, por otro lado, la actividad económica que más predomina es el sector inmobiliario, por lo tanto se debe analizar el riesgo en LA/FT al ser un sector con una ponderación de riesgo media, finalmente, se puede identificar que 34 de estos clientes fueron vinculados de manera presencial al ser residentes colombianos.

Ilustración 26

Riesgo bajo



Fuente: Elaboración propia

En el grupo de riesgo bajo, se identificaron 47 clientes, 43 de ellos son personas naturales que realizaron vinculación presencial, y quienes en su mayoría son asalariadas y rentistas de capital, los ingresos de los clientes de este clúster no superan los 2 mil millones de pesos, y son lo de menor movimiento transaccional con un total de 18 mil millones de pesos.

Son residentes colombianos, españoles y estadounidenses, lo que genera una puntuación baja en cuanto al factor de riesgo jurisdicción ya que son países cooperantes y con buena calificación en LA/FT, adicionalmente el 95.74% de ellos no coincidieron en noticias de lavado de activos.

10. Plan y Recomendaciones de Implementación y Aplicación

A continuación, se presentan las recomendaciones y su respectivo plan de trabajo que se presentaron para la ejecución de las actividades de despliegue planteadas, teniendo en cuenta que son viables y beneficiosas para R4G, y dando un alcance de los posibles riesgos y restricciones que se pueden presentar:

10.1. Implementación del Modelo de Segmentación Seleccionado:

Solicitar la aprobación de la casa matriz y la Junta Directiva.

10.2. Monitoreo y Calibración:

Establecer un proceso de monitoreo continuo para evaluar la precisión y el rendimiento del modelo de segmentación seleccionado, realizando ajustes y calibraciones semestrales para asegurar que el modelo siga siendo efectivo y se adapte a posibles cambios en los datos o en las necesidades del negocio.

Adicionalmente es importante implementar las recomendaciones dadas por la Superintendencia Financiera de Colombia para la calibración del modelo, las cuales reposan en la Guía de Mejores Prácticas para la construcción de modelos de segmentación relacionados con los factores de riesgo de Lavado de Activos y Financiación del Terrorismo (2023):

- Considerar la calibración de los modelos conforme lo establecido en la Circular Básica Jurídica.
- Documentar cada uno de los pasos que se siguieron para la obtención de modelos, su calibración y el resultado de estos.
- Evaluar y ajustar el impacto que puede generar las modificaciones a los modelos de segmentación en cada una de las etapas del SARLAFT.

10.3. Implementación de una Política de Admisión:

La política de admisión incluye una descripción de aquellas tipologías de clientes que podrían presentar un riesgo superior a la media, en función del sector de actividad al que pertenezcan, de la procedencia o residencia de estos clientes, o de cualquier otra información de la que se disponga.

La política de admisión se ha de aplicar a todos los clientes antes de entablar relaciones comerciales. En el momento de establecer la relación comercial se solicitará del cliente información para conocer la naturaleza de su actividad profesional o empresarial o el origen de los fondos, y serán adoptadas medidas dirigidas a comprobar la veracidad de la información facilitada.

Los riesgos inherentes al lavado de activos o de la financiación del terrorismo pueden ser gestionados de una forma más eficaz y eficiente si se conoce previamente el riesgo potencial ligado a los diferentes tipos de clientes y de sus operaciones.

El tener identificados a los clientes por niveles de riesgo permitirá implantar medidas y controles para mitigar dichos riesgos, y así centrarse en los clientes y transacciones que presenten mayor riesgo.

En algunos casos puede ser que el riesgo sólo se manifieste una vez que el cliente haya comenzado a realizar operaciones, obligando esta circunstancia a que el seguimiento de las operaciones del cliente sea un componente fundamental del planteamiento basado en el riesgo.

10.4. Sistema de Atribución de Riesgo y Categorías:

R4G en función de su propio riesgo de negocio y servicios que ofrece, debe diseñar e implementar medidas y controles adecuados para mitigar los riesgos potenciales de lavado de

activos y/o de financiación del terrorismo respecto de aquellos clientes que se consideren de riesgo.

Para ello, R4G debe disponer de una herramienta integral de LA/FT, que, entre otros, conste de un módulo conoce a tu cliente. Este módulo (o sistema de scoring de riesgo) proporcionará un determinado riesgo a una serie de datos facilitados por el cliente en el alta de cuenta o hechos que ocurren durante la relación de negocios y que se consideran factores importantes para tener en cuenta para determinar si el cliente puede ser catalogado de alto, medio o bajo riesgo de LA/FT. En función del riesgo obtenido, se aplican medidas de diligencia debida reforzada, normal o simplificada.

El sistema de Scoring de Riesgo tendrá en cuenta:

- Un Scoring Inicial Ponderado que se produce con la información de conocimiento del cliente que se tiene en el momento de iniciar la relación de negocio y que permitirá calcular el riesgo en el proceso de vinculación.
- Un Scoring Periódico Ponderado se basa en la información que se obtiene del cliente una vez que se ha iniciado la relación de negocio e información de la operativa del cliente.

Cualquier cambio de segmento de un cliente debe ser notificado en el aplicativo para ser analizado.

La configuración del scoring consta de 4 pilares, que por orden de importancia son: 1. Factores de riesgo; 2. Pesos; 3. Ponderaciones; 4. Nivel de riesgo; 5. Factores de corrección. El resultado de la aplicación de estos elementos en el modelo de segmentación seleccionado para el proyecto asignará al cliente persona natural o jurídica a uno de los 3 clúster, categorizando a

estos clúster en riesgo alto, medio o bajo de acuerdo con los resultados obtenidos en la última actividad del despliegue titulada *Asignación de riesgo a cada segmento* (ver ilustración 3).

Ilustración 27

Segmentación de los factores de riesgo



Fuente: Elaboración propia

El procedimiento de asignación del riesgo a través del módulo conoce a tu cliente debe ser actualizada cada día, mediante el suministro de datos por parte de R4G con ficheros que incluyen todos los factores que contempla la plataforma para asignar el riesgo. Estos datos permiten asignar o actualizar el riesgo asociado a cada cliente (nuevos clientes o clientes con variaciones), que devuelve un fichero con la consecuente asignación del riesgo a la Entidad.

La actualización se debe realizar de manera diaria, para cumplir con uno de los requerimientos de los entes de control, ya que el modelo inicial era estático, pero al incorporar la variable "aporte" en el modelo, los clientes pueden cambiar de segmentación en función de sus transferencias diarias.

Es importante destacar que un cambio de segmento implica que el cliente será reasignado en relación con su nivel de riesgo, lo que significa que puede pasar de ser clasificado como un cliente de riesgo bajo a uno de riesgo medio o alto, o viceversa. Por lo tanto, es fundamental monitorear estos cambios de segmento para lograr un seguimiento más exhaustivo y preciso de los clientes y su exposición al riesgo en los diferentes niveles: bajo, medio y alto.

El objetivo de este monitoreo es asegurar que los clientes estén adecuadamente clasificados y que se le brinde un tratamiento acorde a su perfil de riesgo, permitiendo tomar acciones preventivas en función de mitigar los riesgos de LA/FT.

El cálculo del riesgo se calcula, tal y como se ha señalado, a través de los factores de riesgo, los pesos de los factores de riesgo, y la ponderación que se asigna a cada uno de los factores. En este sistema, se otorga un mayor peso a determinados factores sobre otros, por ejemplo, atendiendo a criterios de la casa matriz R4, el país de residencia o el hecho de ser una persona expuesta políticamente (PEP) tienen un peso mayor que otros factores que, en principio, suponen una menor exposición al riesgo, como por ejemplo la antigüedad del cliente o los ingresos periódicos que informa el cliente en su formulario de vinculación o actualización de datos.

Dentro de cada factor, además de los pesos que tendrá cada rubro en el cálculo del scoring, se tendrá que determinar en cada factor las ponderaciones. Estas ponderaciones irán en base a 1, es decir, 1 las de mayor riesgo y 0 las de menor.

10.5. Visualización de los Clúster en Power BI:

R4G en función del control y monitoreo de operaciones, debe crear un dashboard para visualizar los clúster e identificar las características en cuanto a homogeneidad y heterogeneidad, con el fin de identificar outliers en el grupo que se esté monitoreando.

11. Resultados de las Recomendaciones

A continuación, se presentan los porcentajes de implementación de las recomendaciones planteadas y el respectivo informe de cada una de ellas:

11.1. Modelo de Segmentación

El modelo de segmentación fue presentado a la SFC, la casa matriz, la auditoría interna y a la junta directiva, el cual fue aprobado en enero de 2023, no obstante, solicitaron una implementación tecnológica para que este proceso se realizara de manera automática, diaria y con la información extraída del sistema SIFI, por lo tanto, esta recomendación está implementada en un 100%.

11.2. Monitoreo y Calibración:

El modelo se monitorea de manera diaria, para identificar los outliers y realizar análisis de los clientes categorizados en este rubro, no obstante, la calibración se realizará en el mes de junio ya se debe ajustarse de manera semestral, por lo tanto, esta recomendación está implementada en un 50% dentro de R4G.

11.3. Implementación de una Política de Admisión:

Esta política se implementará en el Manual SARLAFT de R4G en junio de 2023, la cual tendrá la aprobación de la casa matriz y la junta directiva, puesto que se pueden presentar ajustes a la política, no se proporcionará un % de implementación por el momento.

11.4. Sistema de Atribución de Riesgo y Categorías:

El área de tecnología de R4G realizó un desarrollo en el aplicativo R4 Lite versión 6 donde incluyó el módulo de segmentación, contemplando todas las recomendaciones dadas en el apartado 11 del presente documento, en donde se crearon los siguientes módulos:

Consulta de parametrizaciones: en este módulo se realizarán las parametrizaciones de las variables seleccionadas para el modelo, de las escalas de riesgo, de la ponderación de riesgo y de los intervalos, de acuerdo con lo establecido en el numeral 7 del presente documento. (ver ilustraciones 11-13)

Edición de parametrizaciones: en este módulo se realizarán las calibraciones semestrales del modelo, con el fin de adicionar o eliminar variables que ayuden a robustecer el modelo actual, diseñando pruebas de validación de homogeneidad y heterogeneidad. (ver ilustración 14).

Procesos: En esta sección se llevará a cabo el proceso de segmentación diaria para analizar los datos, determinar el segmento de cada cliente y verificar si han cambiado de segmento. Además, se realizará la actualización correspondiente en Power BI para reflejar los cambios en la visualización. (ver ilustración 15).

Actualmente, el área de Riesgos y Vinculaciones están realizando las respectivas pruebas de este nuevo desarrollo, y se encuentra en un porcentaje de implementación del 80%.

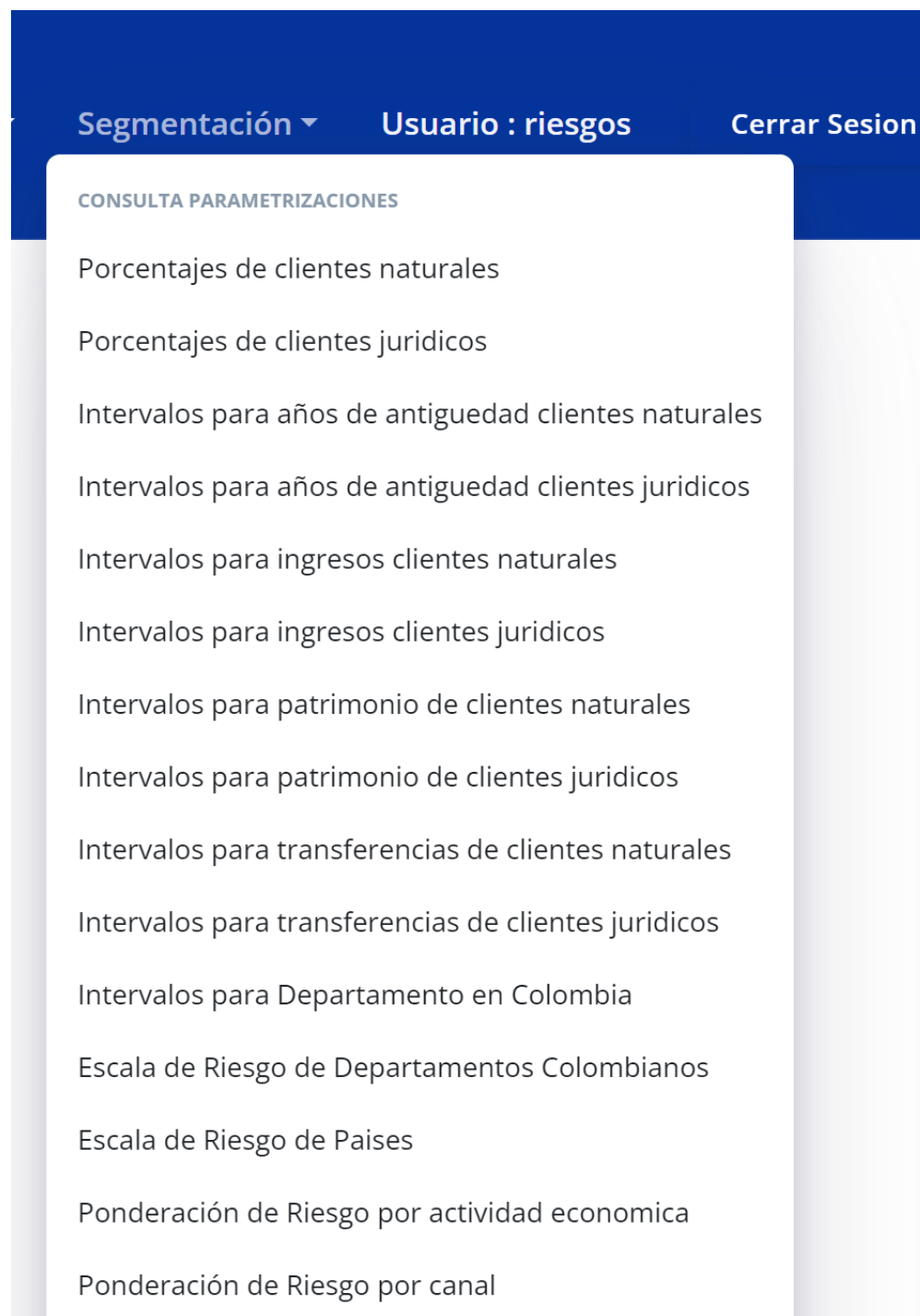
11.5. Visualización en Power Bi:

Se está implementando un dashboard con las variables más significativas dentro del modelo, para identificar las características del negocio presentes en cada segmento, adicionalmente para establecer los diferentes patrones que puedan generar inusualidades basado en los resultados arrojados en cada clúster, es importante tomar como referencia los colores rojo para riesgo alto, amarillo para riesgo medio y verde para riesgo bajo, y los demás datos que no tengan esta categorización manejar un color neutral como el color gris. (ver ilustración 16).

Se están realizando los ajustes al dashboard para surtir una visualización completa de los clusters y los outliers, por lo tanto, se encuentra en un porcentaje de implementación del 70%.

Ilustración 28

Módulo consulta de parametrizaciones



Fuente: Adaptado del aplicativo R4 Lite Versión 6 - Módulo Segmentación








Ilustración 29

Escala de riesgo para países de segmentación

[+ Nuevo](#)

Consulta de Listado de Países Segmentación

Show entries

ACCION	CODIGO PAIS	NOMBRE	ESCALA DE RIESGO	OBSERVACION
 	3	ESTADOS UNIDOS	0	
 	5	VENEZUELA	1	
 	6	AFGANISTAN	1	
 	7	ALBANIA	1	
 	8	ALEMANIA	0	
 	9	ANDORRA	1	Alto riesgo 2

Fuente: Adaptado del aplicativo R4 Lite Versión 6 - Módulo Segmentación





Ilustración 30

Ponderación de riesgo para canales

[+ Nuevo](#)

Consulta de Listado de Canales

Show entries

ACCION	CODIGO CANAL	DESCRIPCION	PONDERACION DE RIESGO
 	2	VIRTUAL	0.5
 	3	REFERIDO	1

Fuente: Adaptado del aplicativo R4 Lite Versión 6 - Módulo Segmentación

Ilustración 31

Módulo edición de parametrizaciones

Tesoreria Sarlaft Comercial Sentencias Segmentación

Edición Porcentajes/Ponderaciones para Clientes Natural :

Nacionalidad :	Ejemplo : 100.00	0	%	*
Pais de Residencia :	Ejemplo : 100.00	100	%	*
PEP :	Ejemplo : 100.00	0	%	*
Reputación :	Ejemplo : 100.00	0	%	*
Actividad Economica :	Ejemplo : 100.00	0	%	*
Ingreso Mensual :	Ejemplo : 100.00	0	%	*

Fuente: Adaptado del aplicativo R4 Lite Versión 6 - Módulo Segmentación

Ilustración 32

Módulo proceso de segmentación

Iniciar proceso de Segmentacion

Segmentacion Vigente

Show 10 entries

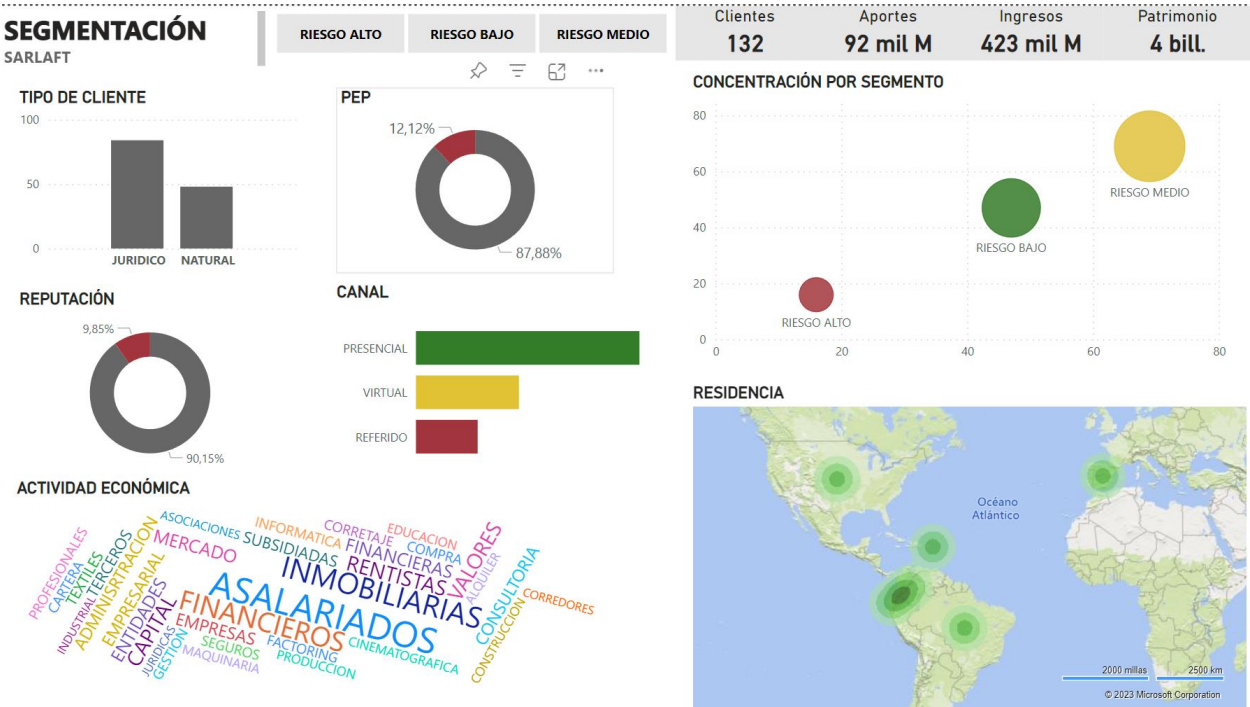
Search:

FECHA SEGMENTACION	NRO IDENTIFICACION CLIENTE	CLIENTE	PUNTUACION DE RIESGO	NIVEL DE RIESGO DEL CLIENTE	OBSERVACION DE
2023-03-22 14:43:15.0	19433046	Alejandro Maraca	5.2	Bajo	Sin cambios

Fuente: Adaptado del aplicativo R4 Lite Versión 6 - Módulo Segmentación

Ilustración 33

Segmentación - Dashboard en Power Bi



Fuente: Elaboración propia

12. Conclusiones

En este proyecto, se ha implementado un modelo de segmentación para el área de riesgos de R4G, con el propósito de mejorar la clasificación de clientes y fortalecer el Sistema de Administración de Riesgos de Lavado de Activos y Financiación del Terrorismo (SARLAFT). El desarrollo del modelo de segmentación ha seguido las regulaciones actuales y las recomendaciones de los organismos supervisores, garantizando así el cumplimiento normativo.

A través de la aplicación del modelo CRISP-DM, se logró comprender el negocio y los objetivos del proyecto, así como analizar y preparar los datos necesarios para el desarrollo del modelo. Se trabajó con información extraída de los formularios de vinculación y actualización de los clientes de R4G, considerando tanto personas naturales como jurídicas.

Adicionalmente se seleccionaron los algoritmos de segmentación K-medoids, Fuzzy C-means y Agrupamiento jerárquico, los cuales fueron evaluados utilizando el Coeficiente de Silueta y el índice de Dunn. Tras la evaluación, se determinó que el modelo más eficiente fue K-medoids.

La implementación del modelo de segmentación seleccionado permite identificar operaciones inusuales y mejorar el control y prevención del riesgo de Lavado de Activos y Financiación del Terrorismo. Esto contribuirá al cumplimiento normativo, la reducción de riesgo y la eficiencia en la gestión del SARLAFT en R4G.

Se ha establecido un alcance para el proyecto, que se centra en la mejora del sistema de control y monitoreo de R4G, para prevenir la introducción de recursos provenientes de actividades ilícitas, además, se ha considerado la necesidad de cumplir con los estándares internacionales y las recomendaciones de los entes de control para cumplir con los criterios de la calibración.

Como vías futuras de investigación para un alcance mayor en este proyecto, se sugiere la aplicación del modelo de segmentación en otros contextos como el SAGRILIFT y organizaciones del sector privado, realizando las adaptaciones correspondientes. También se recomienda introducir sistemas automatizados de procesamiento de datos y establecer sistemas para actualizaciones frecuentes de datos.

En resumen, este proyecto aplicado empresarial ha propuesto un modelo de segmentación para SARLIFT en R4G que cumple con los objetivos y requerimientos normativos planteados. El modelo se fundamenta en la consideración de un conjunto de variables relevantes cuidadosamente seleccionadas, las cuales son determinantes tanto para la segmentación como para la interpretación de los segmentos. Estas variables, como la actividad económica, PEP, productos (FIC, FVP, FCP y negocios fiduciarios), canal, país de residencia, , reputación, departamento, ingresos, aportes y patrimonio, han sido identificadas como factores clave para el análisis y la clasificación de los clientes. La incorporación de estas variables ha permitido desarrollar una solución completa, fiable y fundamentada, generando beneficios significativos para la organización en términos de cumplimiento y gestión de riesgos.

13. Referencias bibliográficas

- D'Alessio Torres, Vincenzo Jesús. (2021, abril 14). *Regla de Sturges: Concepto, explicación, aplicaciones, ejemplos*. Lifeder. <https://www.lifeder.com/regla-sturges/>
- Departamento de Estabilidad Financiera, Banco de la República. (2022). *Reporte de Estabilidad Financiera: 2022-I* (p. 23). Banrep.
<https://www.banrep.gov.co/sites/default/files/publicaciones/archivos/presentacion-estabilidad-financiera-primer-semester-2022.pdf>
- Hastie,Trevor, Tibshirani, Robert, & Friedman, Jerome. (2001). *The elements of statistical learning: Data mining, inference, and prediction [Elementos de aprendizaje estadístico, minería de datos, inferencia y predicción]*. New York : Springer.
<http://archive.org/details/elementsofstatis0000hast>
- Haya, P. (2023). *Esquema del ciclo CRISP-DM estándar* [Esquema].
<https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/>
- IBM. (2021). *Guía de CRISP-DM de IBM SPSS Modeler*. IBM Corporation.
<https://www.ibm.com/docs/es/spss-modeler/saas?topic=overview-spss-modeler-subscription>
- Java. (2023). *Java*. <https://www.oracle.com/java/>
- Kassambara, Alboukadel. (2017). *Multivariate Analysis I—Practical Guide To Cluster Analysis in R - Unsupervised Machine Learning*. STHDA.
- Ministerio de Justicia y del Derecho, UIAF, & Universidad del Rosario. (2016). *Evaluación nacional del riesgo de lavado de activos y financiación del terrorismo—Resumen ejecutivo*. UIAF. <https://urosario.edu.co/sites/default/files/2022-10/resumen-ejecutivo-informe-final-enr-2016-vfinal.pdf>

Python. (2023, mayo 11). Python.org. <https://www.python.org/>

Superintendencia Financiera de Colombia. (2014). Parte I Instrucciones generales aplicables a las entidades vigiladas, Título IV Deberes y responsabilidades, Capítulo IV Instrucciones relativas a la administración del riesgo de lavado de activos y de la financiación del terrorismo—SARLAFT. En *Circular Básica Jurídica*, 029/14.
<https://www.superfinanciera.gov.co/inicio/normativa/normativa-general/circular-basica-juridica-ce---/parte-i-instrucciones-generales-aplicables-a-las-entidades-vigiladas-10083444>

Superintendencia Financiera de Colombia. (2023). *Guía de Mejores Prácticas para la construcción de modelos de segmentación relacionados con los factores de riesgo de Lavado de Activos y Financiación del Terrorismo*. SFC.
<https://www.superfinanciera.gov.co/descargas/institucional/pubFile1048182/GUIASEGMENTACION.pdf>

Universidad de Granada. (2019, junio 17). Práctica 8 Métodos de análisis multivariante: Análisis Clúster. *Estadística Universidad de Granada*. <http://wpd.ugr.es/~bioestad/guia-spss/practica-8/>