ANÁLISIS ESTRATÉGICO SOBRE EL USO DE INTELIGENCIA ARTIFICIAL, MINERÍA DE DATOS Y ANÁLISIS DE BIG DATA EN PREVENCIÓN Y DETECCIÓN LA/FT (UIF/MP)

Diciembre/2021

El GAFILAT agradece la asistencia técnica brindada por la Cooperación Alemana para el Desarrollo, implementada por la Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) para la elaboración del presente documento, que contó además con el apoyo de Nicolás Suarez, Claudia Andrea Soracá y Miguel Avila. El contenido de esta publicación es completa responsabilidad del Grupo de Acción Financiera de Latinoamérica (GAFILAT).

CONTENIDO

PRESENTACIÓN

1. Este documento presenta el resultado de la consultoría "Producto de análisis estratégico sobre el uso de inteligencia artificial, minería de datos y análisis de *big data* en prevención y detección LA/FT (UIF/MP)", con base en los parámetros establecidos por el GAFILAT y en cumplimiento de las actividades establecidas en el Plan de Trabajo.

2. El contenido de este documento se ha organizado en 6 capítulos, así:

- el primer capítulo describe los objetivos generales y específicos;
- el segundo capítulo presenta la metodología llevada a cabo para la construcción del estudio estratégico;
- en el tercer capítulo se describe la situación actual en la región;
- el cuarto capítulo elabora los temas de almacenamiento y procesamiento de datos;
- en el quinto capítulo se describen herramientas de hardware y software;
- y, por último, en el sexto capítulo se plantean las conclusiones del análisis y recomendaciones.

3. La narración y puntos de vista expresados por el equipo consultor se apoyan en el análisis de la información documental proporcionada por la Secretaría Ejecutiva del GAFILAT, por las Unidades de Inteligencia Financiera y las Fiscalías o Ministerios Públicos, por medio de los aportes y respuestas al formulario online compuesto por preguntas específicas relacionadas al análisis mediante inteligencia artificial en la región.

ACRÓNIMOS Y SIGLAS

ALA/CFT	Antilavado de Activos y Contra el Financiamiento del Terrorismo
APNFD	Actividades y Profesiones No Financieras Designadas
CN	Coordinación Nacional
CT	Cumplimiento Técnico
EM	Evaluación Mutua
EBR	Enfoque Basado en Riesgo
ENR	Evaluación Nacional de Riesgo
FPADM	Financiamiento de la Proliferación de Armas de Destrucción Masiva
FT	Financiamiento del Terrorismo
GAFI/FAFT	Grupo de Acción Financiera Internacional
GAFILAT	Grupo de Acción Financiera de Latinoamérica
GT	Grupo de Trabajo
LA	Lavado de Activos
LA/FT	Lavado de Activos y Financiamiento del Terrorismo
PEP	Personas Expuestas Políticamente
R	Recomendación
ROS/RTS	Reporte de Operación Sospechosa/Reporte de Transacción Sospechosa
SE	Secretaría Ejecutiva
SO	Sujetos Obligados
UIF	Unidad de Inteligencia Financiera

NOTACIÓN MATEMÁTICA

4. En el desarrollo de este documento se utilizarán las siguientes nociones:

a. **Valor**: es la representación de la cuantía que toma una característica para un individuo. En este documento los valores se representarán con la letra equis (x), y se diferenciarán entre sí por un número y letra en el subíndice.

b. **Variable**: se refiere a una serie de datos que contiene los valores que toma una característica para un grupo de individuos. Estas variables pueden ser numéricas, textuales o categóricas. Dentro de las categóricas pueden o no ordinales. Una variable numérica puede ser el ingreso, una variable textual puede ser la descripción de un Reporte de Operación Sospechosa (ROS), una variable categórica ordinal puede ser el nivel de educativo de una persona y una variable categórica no ordinal puede ser el sector económico de una empresa. En este documento las variables se representarán por la letra equis en negrilla ($\boldsymbol{x}$), y se diferenciarán entre sí por una letra en el subíndice.

c. **Vector**: es un conjunto de valores que se consideran en su conjunto. Una variable es un vector, pero un vector también pueden ser los valores que toman diferentes variables para un mismo individuo.

d. **Matriz**: es un conjunto de vectores. Así, si se tienen dos o más variables para cien individuos, por ejemplo, el ingreso, la edad y la cantidad de transaccione en efectivo que ha realizado en el último mes, una matriz puede verse como un arreglo en Excel que tiene 100 filas (una por cada individuo) y 3 columnas (una por cada variable). En este documento las matrices se representarán por letras mayúsculas en negrilla.

e. **Observación** atípica: también conocidas como *outliers*, se refieren a valores de una variable que son atípicos si se les compara con otras observaciones. Estos registros suelen ser descartados en el análisis de datos, pero para el sistemas anti lavado de activos y contra la financiación del terrorismo (en delante sismtea ALA/CFT) pueden ser, precisamente, casos de interés.

f. **Variable categórica**: también llamada variable cualitativa, es una variable que puede tomar un valor de un conjunto, usualmente finito, de valores posibles. Se utiliza para representar grupos. Una variable que indica si una persona ha estado o no involucrada en un ROS es una variable categórica.

g. **Variable continua**: es una variable que representa cantidades. Variables como el ingreso, el valor de los movimientos en efectivo o la cantidad de transacciones cambiarias, en un periodo de tiempo, son ejemplos de variables continuas.

INTRODUCCIÓN

5. La Inteligencia Artificial (IA) se nutre de datos para desarrollar algoritmos y, con esto, analizar y obtener información del entorno, identificando comportamientos y tendencias. La IA sigue posicionándose como una de las mayores apuestas tecnológicas, avanzando hacia una mayor escalabilidad, responsabilidad e inteligencia para optimizar los procesos de aprendizaje y agilizar los tiempos de valoración.

6. Por otra parte, las nuevas tecnologías que soportan los métodos más recientes para el intercambio de dinero, sumado a la confianza de los usuarios desplazada de las monedas FIAT, son tendencias crecientes en todo el mundo. Sumado a esto, la falta de intermediarios y la reducción de costos en la operación transaccional terminan siendo elementos fundamentales para este creciente mercado, en donde los datos cobran mayor relevancia.

7. Estos avances tecnológicos, caracterizados por la globalización y el surgimiento de nuevos sistemas de pagos basados en transferencias electrónicas, así como la diversificación de los instrumentos financieros, han creado un enorme campo de acción para los lavadores de activos y las organizaciones criminales y terroristas que buscan utilizar sus recursos de origen ilícito, dándoles apariencia de legalidad en la economía para financiar su actividad criminal. Por esto, es necesario contar con herramientas que permitan integrar de forma eficiente los grandes volúmenes de información que se generan, y que esté fortalecida con capacidades humanas y elementos de hardware, software y sistemas de información.

8. Las Unidades de Inteligencia Financiera (UIF) tienen la capacidad, facultades y potencial para desempeñar un rol estratégico dentro del sistemas anti lavado de activos y contra la financiación del terrorismo (en delante ALA/CFT) incrementando su efectividad, debido a que sus funciones les permiten contar con información (en la mayoría de los casos a todo nivel), lo cual es el eje fundamental para el análisis del comportamiento del lavado de activos y la financiación del terrorismo a nivel de personas y empresas, junto con su participación en la actividad económica agregada.

9. Igualmente, el Ministerio Público (en adelante, MP), además de ser el órgano de investigación de estos delitos, es una fuente de información para el perfilamiento de operaciones e individuos. En este sentido, es importante generar mecanismos de intercambio de información de inteligencia entre ambas autoridades que potencialicen sus fuentes de información y análisis, y mejoren los resultados de los Sistemas ALA/CFT y judiciales en materia de efectividad.

10. La minería de datos, textos e imágenes[1], los modelos de aprendizaje de máquina y el análisis de redes complejas son procesos y disciplinas que requieren recursos, conocimientos,

[1] En el caso de análisis de imágenes, se busca interpretar la información visual para, entre otros, identificar la presencia de un objeto determinado.

habilidades y la orquestación de múltiples actores y procesos. En la medida en que las entidades comprendan y dominen el alcance completo de las actividades asociadas y su articulación, van a poder mejorar de manera significativa sus capacidades para la detección de "anormalidades" o "atipicidades" en la información que agrupan, llegando incluso a la identificación de situaciones que no son observadas por los individuos reportantes del sistema ALA/CFT porque sólo tienen acceso a una porción limitada de la información.

11. De la misma manera, los controles que en la actualidad existen relacionados con estas tecnologías no se encuentran determinados y en muchos casos se observa la ausencia de desarrollo legislativo por parte de los países en Latinoamérica. Así mismo, los procedimientos de debida diligencia, verificación de beneficiario final o incluso actividad transaccional son temas que deben ser revisados desde todas las vertientes.

1. OBJETIVOS

1.1. GENERAL

12. Elaborar un documento de investigación que analice las posibilidades que tienen las metodologías de análisis de datos disponibles en la actualidad para el fortalecimiento de los sistemas ALA/CFT/CFPADM. Para dar mayor claridad sobre el aprovechamiento de tecnologías como el big data, la inteligencia artificial, el aprendizaje de máquina y la minería de datos y textos, se identificarán y presentarán casos de uso que pueden ser aplicados por las UIF y los ministerios públicos.

1.2. ESPECÍFICOS

13. Como objetivos específicos se consideran los siguientes:

a. Caracterizar el nivel de conocimiento y uso de tecnologías como el big data, la inteligencia artificial, el aprendizaje de máquina y la minería de datos y textos en el análisis (operativo y/o estratégico) que realizan las UIF y/o autoridades judiciales.
b. Presentar las principales tecnologías y metodologías disponibles en la actualidad para el análisis de datos, haciendo énfasis en sus aspectos teóricos y aplicaciones prácticas.
c. Identificar el nivel de uso de las UIF y las Fiscalia y/o Ministerios públicos de la región de tecnologías como el big data, la inteligencia artificial, el aprendizaje de máquina y la minería de datos y textos, y hacer recomendaciones sobre el aprovechamiento de estos recursos.

2. METODOLOGÍA

14. El desarrollo de esta investigación se basó en la experiencia de los consultores en el análisis de inteligencia financiera, estratégica y operativa, y en el diseño e implementación de sistemas que aprovechan grandes volúmenes de datos para identificar situaciones de interés. Se consultaron fuentes de información técnicas y académicas para la presentación de las

herramientas tecnológicas de almacenamiento de información y las metodologías de análisis de datos, y se diseñó y aplicó entre las UIF y ministerios públicos de la región un instrumento para documentar el nivel de información con que cuentan, así como las tecnologías implementadas y el estado de avance que tienen en la implementación de este tipo de análisis.

15. Adicionalmente, se plantearon entrevistas con organismos (UIF-MP) a efectos de compartir su experiencia sobre el uso de big data, inteligencia artificial, aprendizaje de máquina o minería de datos y textos.

2.1. RECOPILACIÓN DE INFORMACIÓN

16. Como primera actividad de este trabajo, la consultoría revisó y analizó las fuentes de información disponibles para la elaboración del documento y análisis. Aquí se consideraron:

- Documentos y guías del GAFI u organismos similares (GAFILAT).
- Informes de gestión de las autoridades (UIF-MP).
- Informes y boletines de organismos internacionales como el Grupo Egmont, CICAD, FMI, sobre análisis de datos.
- Artículos técnicos o académicos.
- Contexto legal (penal o administrativo) de la adopción de nuevas tecnologías en la región, con énfasis en entidades públicas.
- Informes de gestión de las autoridades (UIF-MP)
- Informes y boletines de organismos internacionales como el Grupo Egmont, CICAD, FMI, sobre análisis de datos.
- Información obtenida de fuentes abiertas relacionada con el tema.

17. Como segunda acción se realizó una primera reunión con la Secretaría Ejecutiva (SE), para la realización de consultar a las UIF y MP a través de las cuales se pretende conocer el nivel tecnológico actual que existe en la región. Con base en un cuestionario detallado, dsitribuido a los puntos de contacto de las UIF y MP se desarrolló una comprensión más profunda de sus modelos y operaciones:

a) Encuesta de opinión para indagar sobre el nivel de conocimiento y uso de tecnologías.
b) Cuestionario dirigido a las UIF de la región
c) Cuestionario dirigido a algunas autoridades fiscales de la región

2.2. ANÁLISIS

18. En esta fase se analizó la información brindada por medio de las respuestas dadas por las autoridades al cuestionario online. A continuación, se presenta el resumen ejecutivo de los principales aportes obtenidos.

3. DIAGNÓSTICO

19. La información y análisis presentados en este capítulo buscan describir las capacidades técnicas y humanas, así como los métodos utilizados actualmente en materia de inteligencia artificial, por las UIFs y MP en Latinoamérica, así como presentar las medidas de seguridad de información y los procesos de retroalimentación implementados por las autoridades en cada país. Este análisis se realizó con base en las respuestas y comentarios realizados por las autoridades (UIF y MP) de los países miembros al cuestionario remitido.

20. De manera general, el cuestionario indagó sobre los siguientes aspectos:

- Acceso de las UIF y del MP a información transaccional, ampliando sobre los tipos de datos y la cantidad de reportes que tienen a su disposición. Con esto se busca entender las posibilidades que se tienen para la generación de modelos analíticos para entender y detectar la actividad relacionada con el LA y el FT.
- Retroalimentación entre el MP y las UIF, y entre las UIF y los sujetos obligados. El entendimiento de estas interacciones arroja información sobre el tipo de modelación analítica que se puede implementar, identificando a su vez las posibilidades del sistema ALA/CFT en la región.
- Proceso y modelos de Seguridad en el acceso en la información.
- Tamaño de las UIF y MP, que servirá como punto de referencia para profundizar de manera agregada sobre las capacidades de la región y los avances con el que cuenta el Sistema ALA/CFT.

3.1. UNIDADES DE INTELIGENCIA FINANCIERA

21. A continuación, se presentan la síntesis de las respuestas y cuestionarios de las UIFs.

3.1.1. Aspectos generales

22. Con relación al tipo de UIF de la región, todas están definidas bajo un modelo administrativo que, de acuerdo con el alcance indicado, señala que: "*Las UIF de este tipo suelen formar parte de la estructura o el ámbito de supervisión de una administración u organismo distinto de las autoridades judiciales o policiales. A veces constituyen un organismo separado, sustancialmente sujeto a la supervisión de un ministerio o una administración (UIF "autónoma") o al margen de ella (UIF "independiente"). En este caso, la intención principal es establecer una "zona neutral" entre el sector financiero (y, más en general, las entidades y los profesionales sujetos a la obligación de notificación) y las fuerzas del orden a cargo de la investigación y el enjuiciamiento de los delitos financieros*[2].

23. En lo concerniente al recurso humano, se consultó a la UIF sobre la composición o número de personal, específicamente de sus áreas operativas (análisis operativo y estratégico) y de tecnología (seguridad de la información), con el fin de tener una referencia del nivel o

[2] Unidades de Inteligencia Financiera. Panorama General. Fondo Monetario Internacional, Banco Mundial. 2004

participación de estas con relación a la composición total del personal. En este sentido, se observa que, en promedio, las entidades cuentan con el 42% de los funcionarios para la realización de análisis de operaciones, 13% para análisis estratégico y con un 12% para tecnología, tanto a nivel de seguridad de la información como mantenimiento e infraestructura. Lo cual evidencia la relevancia dada por las UIF a sus áreas operativas para análisis y gestión de información y reportes para generar prevención, detección de LA/FT, así como para la capacitación y retroalimentación de los actores del sistema ALA/CFT.

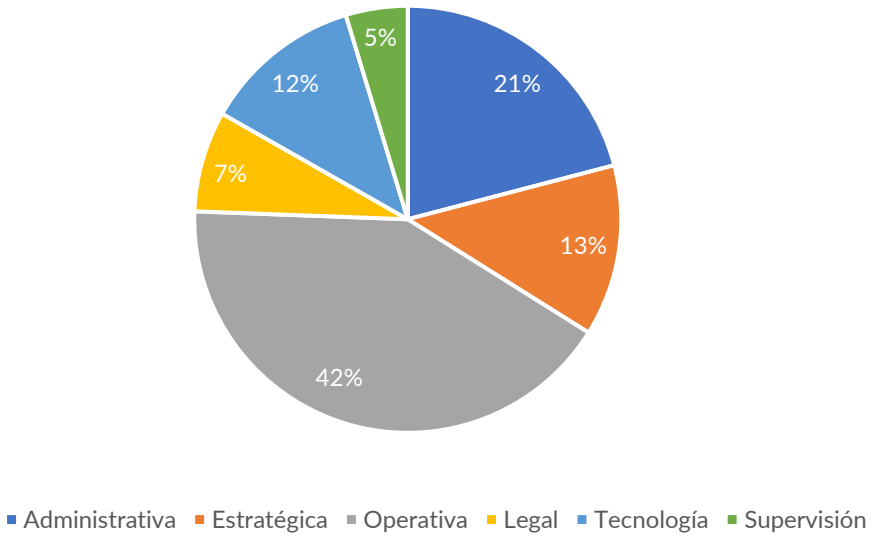

Gráfico 1.Promedio de la participación de funcionarios por área

24. Es importante mencionar que algunas UIF en la región cumplen funciones a nivel de supervisión de sujetos obligados, como en el caso de Argentina, Bolivia, Brasil, Ecuador, Guatemala, Honduras, Nicaragua, Paraguay y Perú.

3.1.2. Seguridad de la información

25. Con relación a la seguridad de la información, se consultó acerca del modo de consulta de las fuentes de información, en términos de accesibilidad, es decir, si el acceso sucede de manera masiva o individualizada a los datos. De acuerdo con las respuestas dadas en los cuestionarios, en el 43% de las UIF el área operativa y el 57% de las áreas estratégicas tienen acceso masivo a la información de las bases de datos. En este mismo sentido las áreas de tecnología en un 36%, por su permisos y competencias, pueden tener acceso masivo a la información.

26. Por otra parte, el 79% de las áreas operativas y el 57% de las áreas estratégicas tienen acceso individualizado, es decir, uno a uno por tipo e identificación, a la información de las bases de datos. Se menciona también acceso a fuentes por parte de las áreas de Supervisión para

evaluación e integridad en atención a casos específicos. En todos los casos se cuenta con asignación de permisos y monitoreo a sistemas específicos de consulta.

27. Dentro de las medidas y controles implementados por las UIF de la región para la seguridad de la información, se encuentran:

- Controles de acceso a la información, mediante la creación de perfiles y usuarios.
- Aplicación de mecanismos de autorización y autenticación.
- Gestión en las actualizaciones de seguridad en las aplicaciones y sistemas utilizados en las últimas versiones que incluye los parches de seguridad provista por los fabricantes.
- Cifrado de la información en ciertos medios de almacenamiento y servicios de transferencias de datos.
- Configuración adicional de la clave BIOS, uso de antivirus, restricción de uso de medios removibles y uso de aplicativos en línea, bloqueo de envío de documentos adjuntos a correos que no son de la SBS, al igual que correos de servicios gratuitos, autorización de accesos de servicios centralizados.
- Monitoreo de la actividad y accesos a las bases de datos mediante la implementación de registros de auditoría.
- Desactivación de acceso al mailweb.
- Aplicación periódica de la prueba de polígrafo a todo el personal de UIF.
- Acceso a la red con certificado digital.
- Acceso remoto, debido a la pandemia, se realiza solo mediante VPN.
- Implementación de estándares internacionales (ISO 27000).
- Restricción de tiempo para utilizar los sistemas.
- Accesos son controlados a través de los logs - bitácora de accesos de infraestructura.
- Data Loss Prevention (DLP).
- Monitoreo con herramientas de ciberseguridad (DLP, FW, VPN).
- Controles y auditorías para verificar el cumplimiento de asignaciones y prevenir vulnerabilidades.
- Restricción de uso de medios de almacenamiento masivo: no se pueden introducir celulares, mecanismos de USB, pen drives, discos externos.
- Restricción de acceso y uso de correos particulares o storage en nubes que no correspondan a la red.
- El acceso a las bases depende de los roles y funciones asignadas a las diferentes áreas. Los accesos están personalizados según las responsabilidades. Existe supervisión y se audita periódicamente. Los datos están en una red privada protegida por firewalls en una zona desmilitarizada cual consta con rango IP propio. Tenemos copia de seguridad periódica.
- Por medio de usuarios autenticados a la red, roles de permisos específicos y dejando bitácora del uso de las aplicaciones, así como herramientas de DLP que monitorean constantemente como los usuarios internos hacen uso de sus recursos institucionales.

3.1.3. Retroalimentación

28. Con relación a la retroalimentación que reciben las UIF acerca de los informes de inteligencia financiera entregados al MP, con base en las respuestas dadas en los cuestionarios, las UIF reciben retroalimentación en un 50% en la etapa de investigación, en un 29% en la judicialización, 36% en la sentencia o condena y el 29% no recibe retroalimentación por parte del MP.

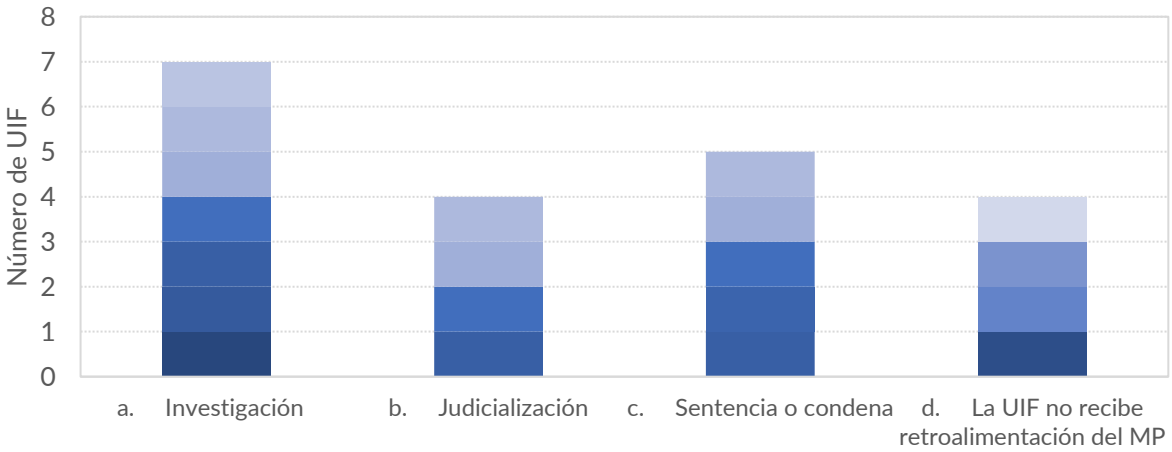

Gráfico 2. Retroalimentación de los informes de inteligencia financiera recibida por las UIF según etapa

29. Asimismo, con base en las respuestas dadas en los cuestionarios, se puede concluir que la retroalimentación que reciben los sujetos obligados acerca de los reportes de operación sospecha entregados a las UIF se realiza en todos los países mediante diferentes formas. Los sujetos obligados reciben retroalimentación, en un 86% de los casos sobre la calidad de los reportes, en un 71% se comunican mediante estadísticas, en un 57% sobre tendencias y en 79% sobre tipologías. En este sentido se resalta que el 100% de las UIF realiza procesos de retroalimentación a los sujetos obligados.

30. La UIF mantiene comunicación permanente con los sujetos obligados y realiza procesos de retroalimentación de los ROS y de otras acciones que realiza la UIF, como son: Evaluación Nacional de Riesgo (ENR), evaluación de riesgos sectoriales, tipologías detectadas a través de los ROS, análisis estratégico, alertas, boletines con información de interés, etc.

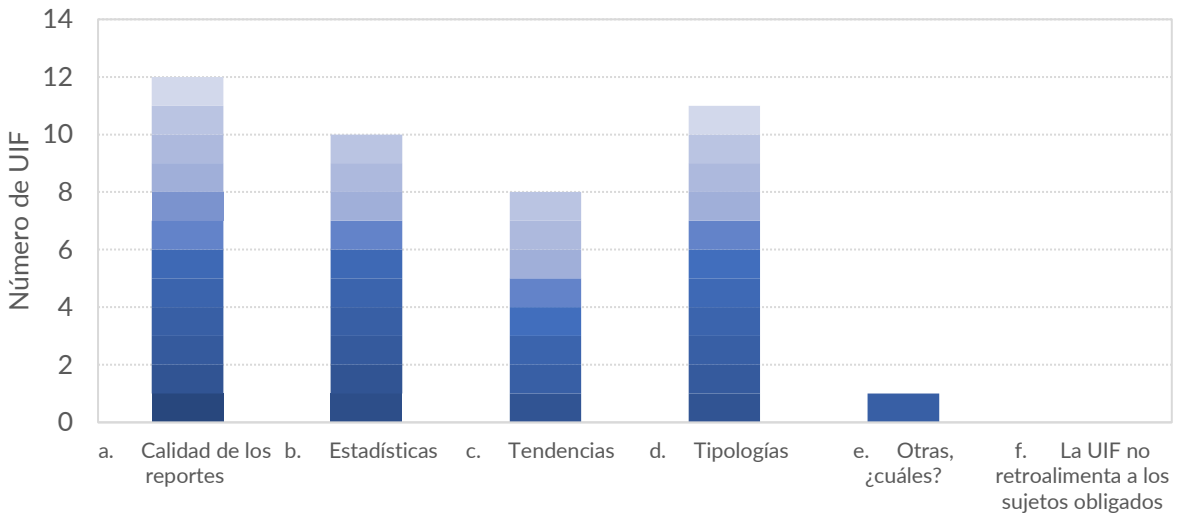

Gráfico 3. Retroalimentación de la información reportada recibida por los sujetos obligados

3.1.4. Tipos de reporte e información

31. Con relación a la información a la que acceden las UIF para realizar el análisis estratégico y operativo, a continuación, se indican los tipos de reportes a los que las Unidades acceden de manera directa o indirecta, presentando la manera de acceso para cada uno de los siguientes reportes:

Información	De manera directa	De manera indirecta
ROS	93%	7%
Transacciones en efectivo	87%	13%
Transacciones cambiarias	60%	33%
Antecedentes penales	27%	67%
Información migratoria	33%	60%
Información aduanera	33%	67%
Transporte transfronterizo de dinero	53%	33%
Transporte transfronterizo de instrumentos negociables	47%	53%
Declaraciones de impuestos u otra información fiscal	20%	67%
Transacciones con activos virtuales	27%	60%
Registro mercantil o de sociedades	53%	33%
Registros de propiedad de bienes inmuebles	67%	20%
Registros de manejo de cuentas de campañas políticas	27%	60%
Registros de beneficiario final	40%	47%
Registro de personas naturales	67%	33%
Registro de personas jurídicas	67%	33%
Registro de PEPs nacionales	60%	27%
Registro de PEPs extranjeras	13%	73%
Declaraciones patrimoniales de funcionarios	7%	73%

Información	De manera directa	De manera indirecta
Registro de actos notariales	27%	73%
Registro de propiedad de vehículos	60%	40%
Compra y venta de metales preciosos	27%	60%

Tabla 1. Información a la que acceden las UIF para realizar análisis estratégicos u operativos

32. Además de los reportes mencionadas anteriormente, las UIF también tienen acceso a otras fuentes de información, que pueden consultarse de manera directa o indirecta, que permiten realizar análisis de inteligencia artificial (IA). Es importante resaltar que algunas UIF cuentan con acceso a esta información de forma masiva, mediante acuerdos o convenios interinstitucionales en los que comparte o la UIF ha solicitado la base, en otros casos se cuenta con ID de consultas con los cuales los analistas pueden realizar búsquedas puntuales por número de identificación a través de los portales de información de las entidades.

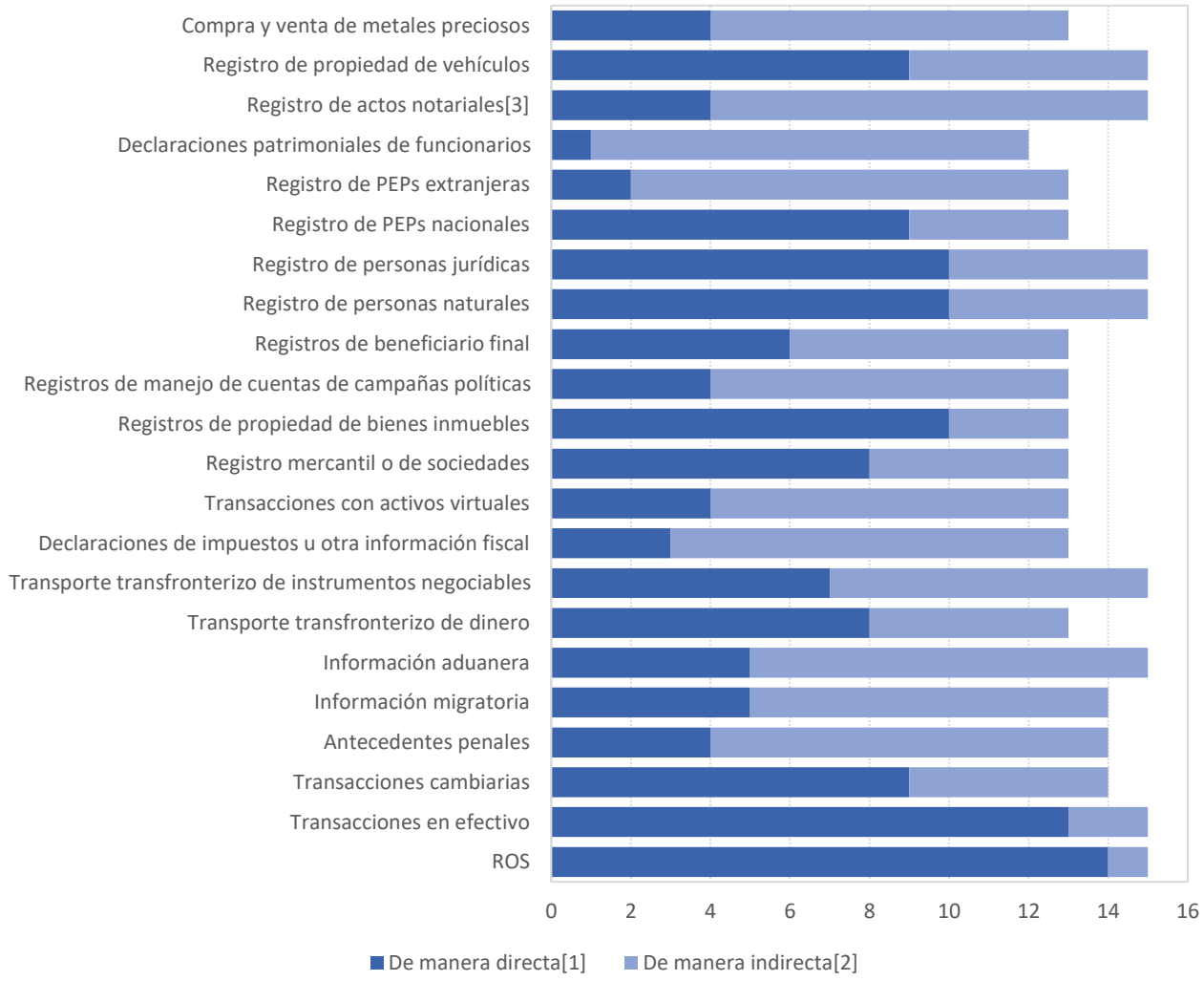

Gráfico 4. Tipo de reporte o fuente de información y manera de acceso

33. Algunas referencias de fuentes de información a los que se podría tener acceso a través de la gestión por parte de la UIF, como: compra venta de vehículos, PEP y sus declaraciones de bienes y rentas, transferencias internacionales y locales enviadas y recibidas, registros inmobiliarios, información de salud y seguridad social (que brinda información de los vínculos familiares, mediante los cuales se pueden comprobar el beneficiario final reportado o relaciones PEPs no reportadas), entre otras.

34. Respecto a los diferentes sectores reportantes, las UIF identifican según su necesidad particular, cuales requieren reportes y la periodicidad de estos. Los reportes en su mayoría están relacionados con la entrega de Reporte de Operaciones Sospechosas (ROS) o Reporte de Transacciones en Efectivo (RTE). La constante frente a los diferentes sectores reportantes están alineados a las recomendaciones en general, ya que todas las UIF contemplan al sector financiero como elemento fundamental de estudio.

Tipos de reportes
Sector financiero general
Sector seguros
Reportes cheques y tarjetas de credito
Agentes de aduana
Cooperativas de ahorro y credito
Remesas - giros
Juegos de suerte y azar / casinos
Compra venta de vehículo automotor
Embarcaciones, naves y aeronaves
Arte, filatelia, antiguedades
Empresas transportadoras de valores
Sociedades de capitalización y ahorro
Banco central
Donaciones o aportes de terceros
Agentes inmobiliarios
Servicios fiduciarios
Deportes profesionales
Factoring
APNFD - contadores
APNFD - abogados
APNFD - notarios
Sector salud
Zonas francas
Reportes entes estatales

Tabla 2. Tipos de reportes que suministran los sujetos obligados

35. Otros sectores reportantes, como APNFD (abogados, contadores, contadores), el mismo sector inmobiliario, han sido de especial atención para las UIF, requiriendo más información en ese sentido. Esto ha permitido relacionar estas actividades o sectores con fenómenos muy particulares respecto al lavado de activos.

36. Con relación a los diferentes tipos de reportes que reciben las UIF, el cuestionario indagó acerca del número anual de ROS, reportes de transacciones en efectivo y reportes de transacciones cambiarias como punto de referencia para analizar el tamaño de bases de datos y capacidad operativa de las UIF. En ese sentido, a continuación, se presenta la siguiente tabla con los indicadores arriba mencionados:

País	ROS	Cantidad de transacciones en efectivo por millón de USD de producción	Transacciones cambiarias	Funcionarios áreas operativas	Cantidad de transacciones en efectivo
País 1	739.154	2,48	-	74	3.582.484
País 2	428.620	8,53	9.509	73	9.174.536
País 3	14.747	10,21	190.844	16	360.397
País 4	3.260	7,13	162	24	377.315
País 5	20.000		-	38	
País 6	13.745	21,29	412.000	39	4.300.000
País 7	1.275		-		
País 8	4.326	162,37	2.000.000	35	12.600.000
País 9	11.267	708,79	31.507.656	55	192.293.525
País 10	750	137,48	238.000	22	7.373.000
País 11	2.494	1.014,94	-	23	100.285.932
País 12	1.559	0,01	10	21	281
País 13	865	8,81	209.860	13	209.860
País 14	417		15	22	2.636

Tabla 3. Indicadores de la dimensión de las UIF de la región. Fuente: encuestas realizadas a UIF y MP y cálculos propios

3.1.5. Hardware y bases de datos

37. El cuestionario indaga por el modelo de base de datos utilizado por las UIF. Aquí, el 86% de las entidades indica uso del modelo relacional, mientras que el 14% de las restantes indican el uso del modelo no relacional (Gráfico 5).

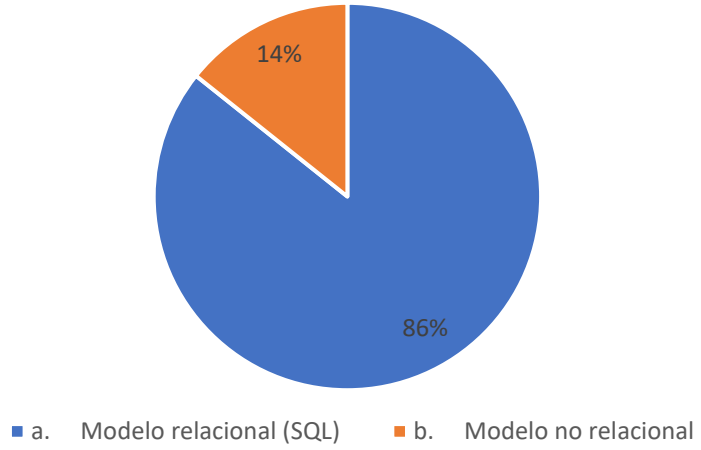

Gráfico 5. Modelo de base de datos utilizada por las UIF

38. Con relación al hardware que las UIF utilizan principalmente para el almacenamiento de la información, el Gráfico 6 muestra que el 60% utilizar servidores independientes, 20% utiliza servidores locales a través de Hadoop, 12% infraestructura especializada como IBM Netezza y Teradata, 7% accede a servicios de datos en la nube, y un 26% utiliza otros como Flash system o nubes de servidores locales a través de Oracle Storage.

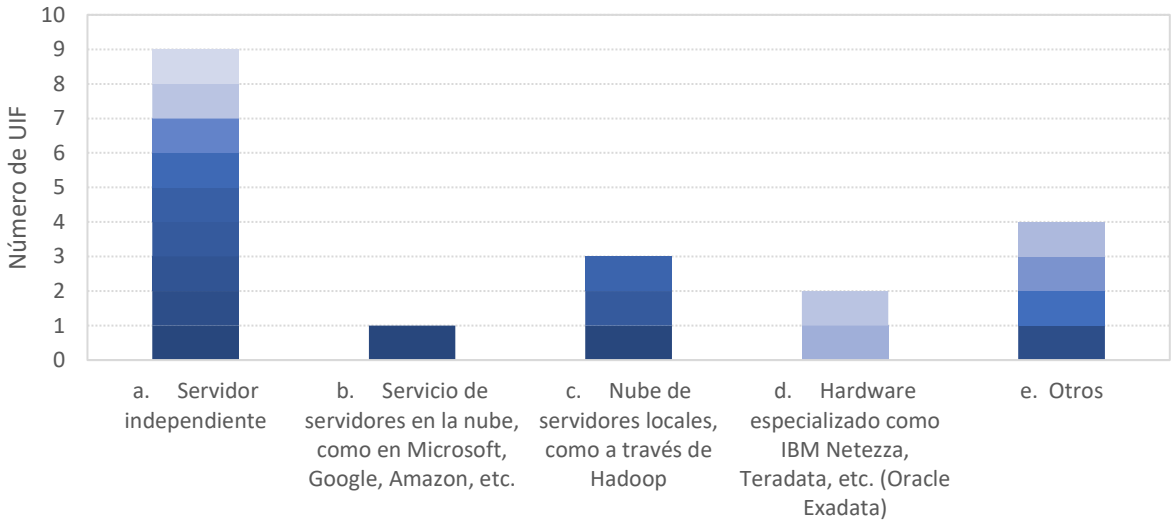

Gráfico 6. Hardware utilizado por las UIF para el almacenamiento de la información

39. A través del cuestionario se consultó sobre el software utilizado para el análisis de los datos que la UIF tiene a disposición. Aquí, el 64% de las UIF utiliza IBM I2 Analysis Notebook, 36% usan R y Phyton (ambas opciones de software libre), y 21% trabajan SAS. Adicionalmente, se mencionan otras herramientas como IBM Modeler, IBM SPSS, Tableau, Microstrategy Visual Insight – Data Discovery, SingleStore DB, WEKA, y herramientas tecnológicas propias como SICORE (Gráfico 7).

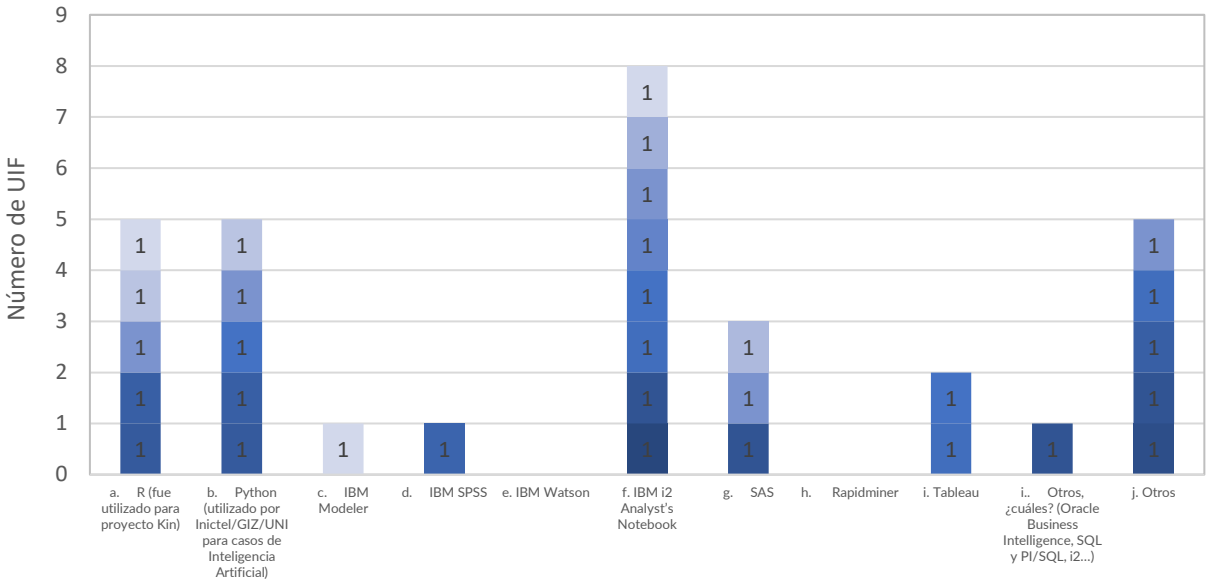

Gráfico 7. Software utilizado por las UIF para el análisis de datos

40. Las metodologías de análisis de datos utilizadas por las UIF se enmarcan, en su mayoría, en la inteligencia de negocios (57%). Adicionalmente, se reporta el uso de técnicas del aprendizaje de máquina, supervisado (50%) y no supervisado (36%), análisis de redes complejas (43%) y otros (14%) como el análisis descriptivo y herramientas propias de comunicación segura.

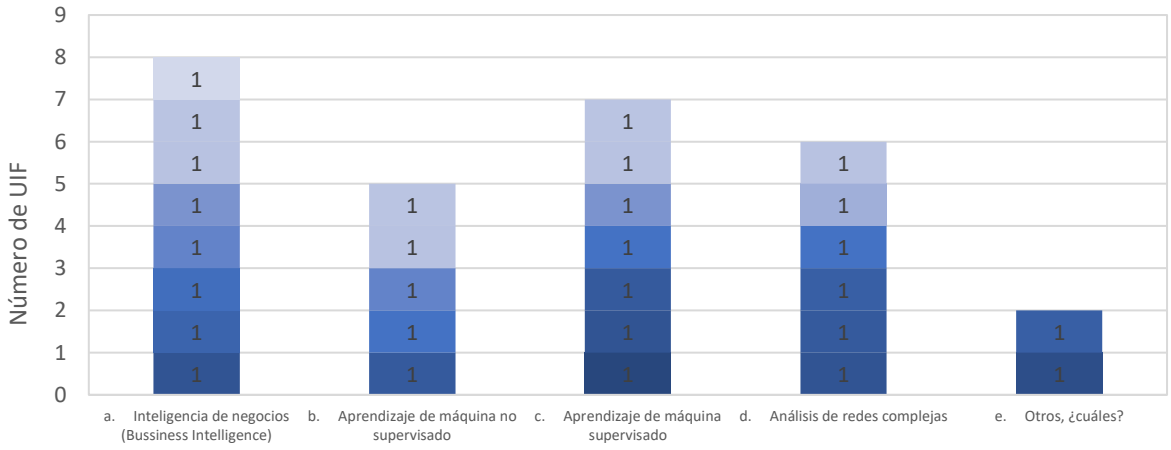

Gráfico 8. Metodología utilizada en el análisis de datos

41. Finalmente, el 64% de las UIF cuentan con estrategias de apropiación e implementación de tecnologías de análisis de datos para la detección de situaciones atípicas de alertas o posibles casos de LA/FT. Aquí se cuentan:

- Matriz de riesgo sistematizada, que generan alertas y relaciones.
- Modelo de riesgo para analizar la información que recibe de los sujetos obligados.

- Indicadores o alertas de riesgo, de monitoreo continuo y masivo de los ROS.
- Análisis de base de datos para identificar coincidencias.

Cuadro 1 – Procedimiento automático para identificar estructuras transaccionales

Caso. Actualmente la UIF se encuentra en el desarrollo de un proyecto que busca implementar un procedimiento automático que permita identificar estructuras transaccionales y de relación con base en información no estructurada descrita en los Reportes de Operaciones Sospechosas, a través de archivos Excel, enviada por las personas obligadas. El proyecto busca relacionar cuentas y personas a través de transacciones para poder identificar estructuras financiares que permitan profundizar y ampliar el análisis de las transacciones reportadas.

3.1.6. Perfil de funcionarios

42. Los perfiles de los funcionarios con los que cuentan las UIF para llevar a cabo labores de analítica de datos, el 66% en ingeniería (de Sistemas, Industrial o Electrónica), 60% cuenta con funcionarios de ciencias sociales (Administradores de Negocios, Economistas, Contadores.), y 26% cuenta con Matemáticos, Físicos y Estadísticos.

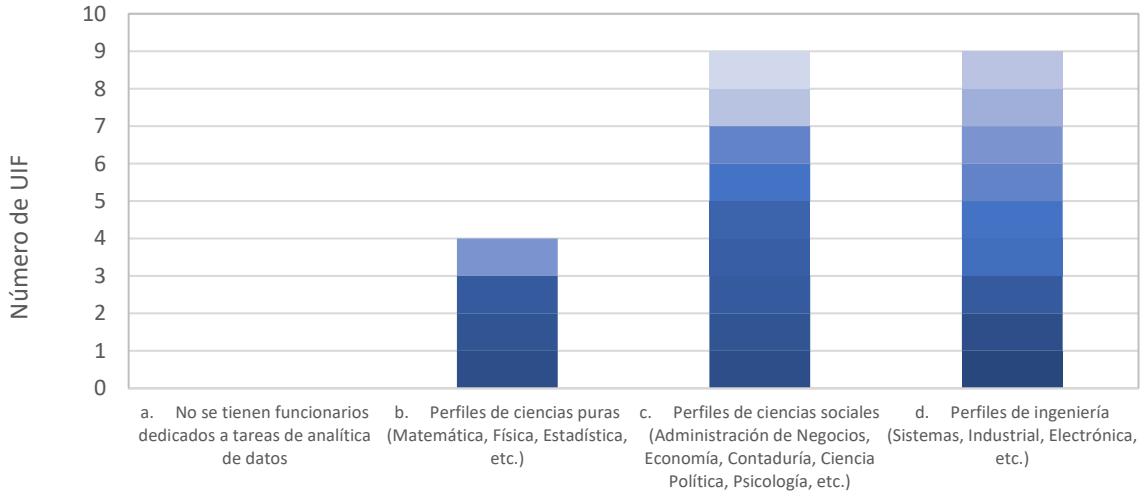

Gráfico 9. Perfiles de los funcionarios con los que cuentan las UIF para labores de analítica de datos

3.2. MINISTERIO PÚBLICO

3.2.1. Aspectos generales

43. En términos generales, los MP en la región cuentan con bases de datos o repositorios de los procesos a los cuales pueden tener accesos individualizados y autorizados, en donde solo los responsables directos de dirigir la investigación pueden tener acceso, manteniéndolo únicamente en los procesos de los cuales está a cargo, por lo cual no se podría verificar las carpetas de investigaciones a cargo de otros compañeros.

44. Existen accesos a bases de datos de carácter restringido, por ejemplo: de tribunales electorales, servicios de migración, autoridades de tránsito y transporte, siendo un número limitado, restringido e identificado de usuarios que poseen acceso, autorizados por el Despacho Superior.

3.2.2. Seguridad de la información

45. Con relación a la seguridad de la información, se consultó acerca de qué medidas y controles se toman para garantizar la seguridad y el uso correcto de las bases de datos. A continuación, se relacionan algunas medidas tomadas:

- Cada agente del Ministerio Público resguarda la información en la carpeta de investigación con el manejo de contraseña que sólo él conoce.
- Autorización de usuarios y bloqueo de accesos.
- Creación de expedientes con accesos restringidos dependiendo del cargo y funciones, quien además reciben usuario y contraseña para el ingreso, que requieren actualizaciones periódicas (mensualmente).
- Auditorías por parte del personal que administra las bases de datos de acceso a los expedientes y bases de datos.
- El acceso a los sistemas se identifica mediante credenciales de seguridad que consisten en al menos un nombre de usuario y contraseña personal y la firma de un término de responsabilidad y confidencialidad. Todos los eventos de registro que involucren objetos de datos del acuerdo o autorización, que permite identificar individualmente la operación realizada, el usuario, el puesto de trabajo y la fecha/hora de las transacciones realizadas.
- Los perfiles de acceso se establecen con la definición de atribuciones y responsabilidades de los usuarios habilitados en ellos, y el acceso se regula mediante un proceso formal de solicitud a los perfiles del sistema, permitiendo incluso verificar los autorizadores que otorgaron los permisos al usuario. En términos generales, solamente los fiscales y los responsables directos de la investigación tienen acceso a la información de cada proceso.

Cuadro 2. Disposición de la Información para análisis

Caso. La información que se genera y procesa en la Unidad de Blanqueo de Capitales/FT, se mantiene en un centro de datos equipado con UPS, servidor de aplicaciones, central telefónica, servidor de correo, sistema de videovigilancia, cortafuego y una red de comunicación, entre otros componentes, completamente independiente del Ministerio Público, por el carácter confidencial de la información que se maneja en ese despacho, es decir, los funcionarios que laboran en la Unidad son los únicos que tienen acceso a la información que se genere. En este sentido, la jefatura tiene un control total en cuanto a la creación y el nivel de acceso que mantiene cada funcionario, generándose con este propósito requerimientos por escrito, en los que se detallan los datos de la persona y se define de forma expresa el nivel de acceso, siendo un número limitado de personas quienes tienen acceso total a la información.

La Unidad de Blanqueo de Capitales, tiene acceso a fuentes de información estatales, que tienen un carácter restringido, como es el caso de la Dirección General de Migración, el Tribunal Electoral y la

Dirección del Tránsito y Transporte Terrestre, para lo cual se han realizado designaciones que se han generado de manera formal, identificando ante la Institución que nos brindó el acceso,) para lo cual cada uno debe firmar un documento de confidencialidad respecto al manejo y uso de la información, que es adicional al que se tiene que firmar por ser parte de esta Unidad.

3.2.3. Tipos de reporte e información

46. A continuación, se relacionan los tipos de reportes e información a la que los MP acceden de manera directa o indirecta, se presenta el nivel de acceso para cada uno de los siguientes reportes:

Descripción	De manera directa	De manera indirecta
Antecedentes penales	60%	100%
Información migratoria	40%	100%
Información aduanera	0%	100%
Transporte transfronterizo de dinero	0%	100%
Transporte transfronterizo de instrumentos negociables	0%	100%
Declaraciones de impuestos u otra información fiscal	0%	100%
Registro mercantil o de sociedades	20%	100%
Registros de propiedad de bienes inmuebles	40%	100%
Registros de manejo de cuentas de campañas políticas	20%	100%
Registros de beneficiario final	20%	100%
Registro de personas naturales	40%	100%
Registro de personas jurídicas	60%	80%
Registro de PEPs nacionales	20%	100%
Registro de PEPs extranjeras	0%	100%
Declaraciones patrimoniales de funcionarios	0%	100%
Registro de propiedad de vehículos	20%	120%
Otras, ¿cuáles?	40%	100%

Tabla 4. Información a la que acceden los MP

47. Sobre los informes de inteligencia financiera entregados por la UIF, el MP brinda retroalimentación en las siguientes etapas: investigación en un 56%, judicialización en un 44%, en sentencia 33% y en un 22% no se cuenta con retroalimentación. En algunos casos se brinda información estadística acerca del estado de los casos cuando así lo solicita la UIF, para casos específicos o bien cuando se debe emitir algún informe de avance del país a nivel nacional o internacional.

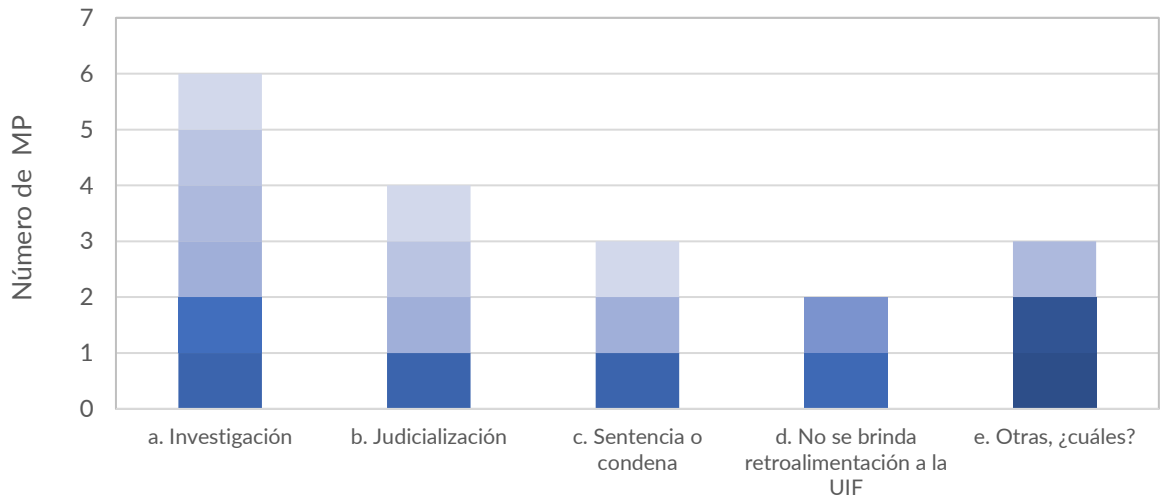

Gráfico 10. Retroalimentación brindada por el MP a la UIF, por etapas

3.2.4. Retroalimentación

48. Con relación a la retroalimentación que porporciona el MP acerca de los informes de inteligencia financiera entregados, con base en las respuestas dadas en los cuestionarios, los MP brindan retroalimentación en un 60% en la etapa de investigación, en un 40% en la judicialización, 30% en la sentencia o condena y el 20% no recibe retroalimentación por parte del MP.

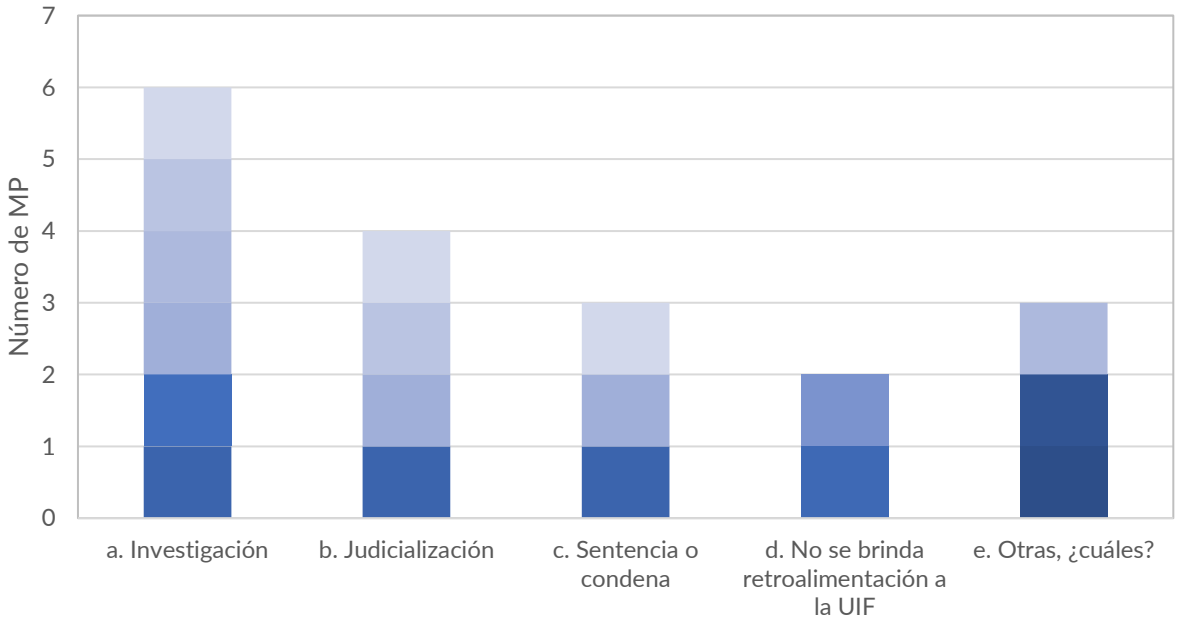

Gráfico 11. Retroalimentación brindada por el MP a la UIF, por etapas

3.2.5. Hardware y bases de datos

49. Con relación al hardware que utilizan los MP para el almacenamiento de la información, el Gráfico 12 indica que el 70% manifestó utilizar servidores independientes, el 33% usa servidores especializados como IBM Netezza y Teradata, y un 20% utiliza una nube de servidores al estilo Hadoop, Google, Amazon.

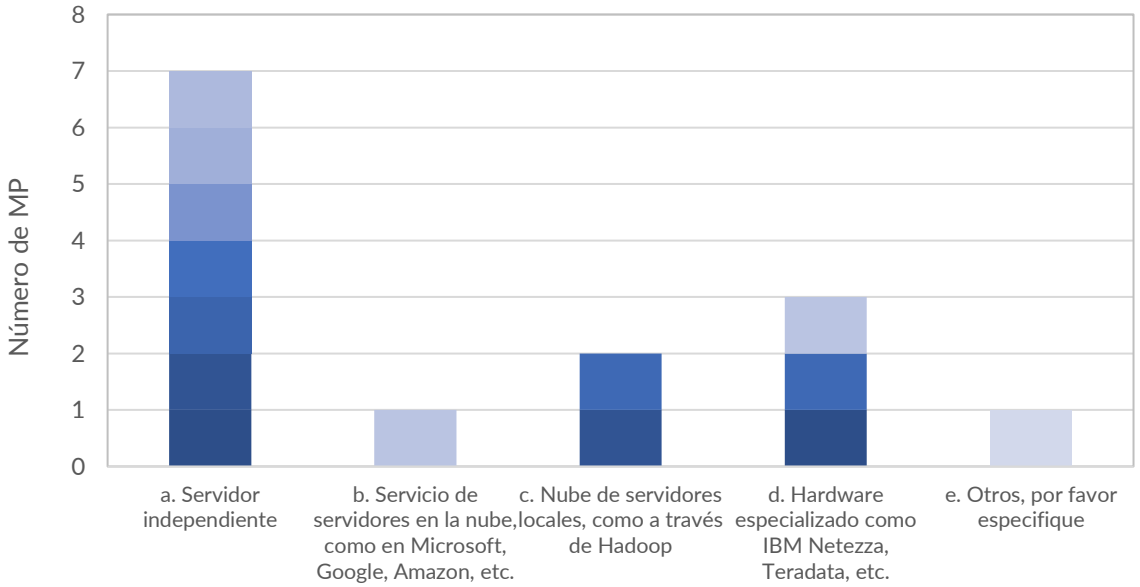

Gráfico 12. Hardware utilizado por los MP para el almacenamiento de la información

50. En cuanto al modelo de base de datos, el modelo relacional sigue siendo el más consolidado, de acuerdo con las tendencias observadas. El 77% utiliza el modelo relacional y llama la atención la incidencia reportada para otras alternativas (33%), lo cual potencializa las posibilidades para la gestión de información al considerar otros paradigmas de mayor especialización.

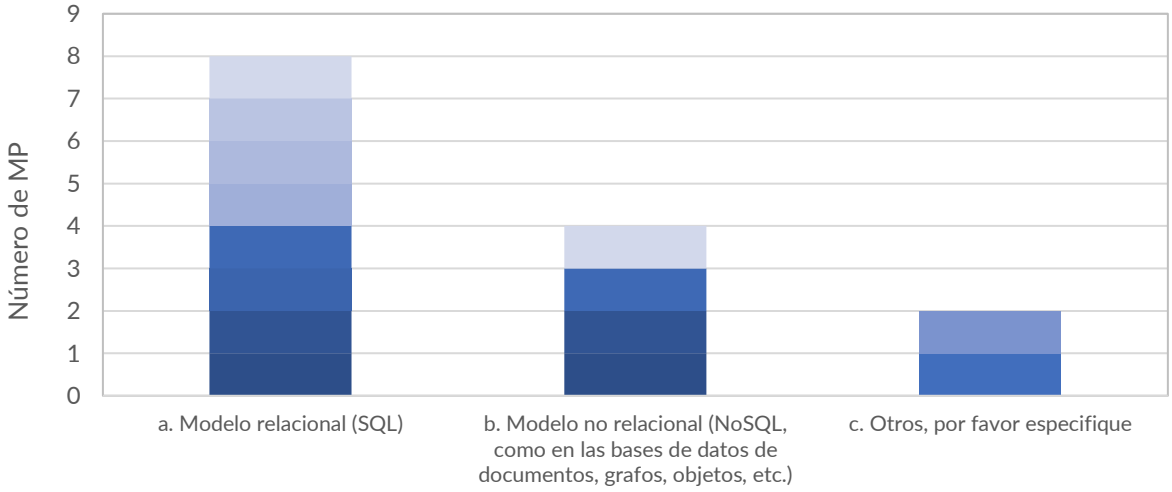

Gráfico 13. Hardware utilizado por los MP para la administración de datos

51. Sobre el software utilizado, al igual que en el caso de las UIF, los MP destacan el uso de IBM I2, seguramente por tratarse de una herramienta especializada diseñada para el análisis de inteligencia financiera. Por otra parte, los programas de analítica de datos como R, Python, SAS, etc. tienen una participación reducida, lo cual evidencia un menor uso de este tipo de metodologías por parte de estas entidades.

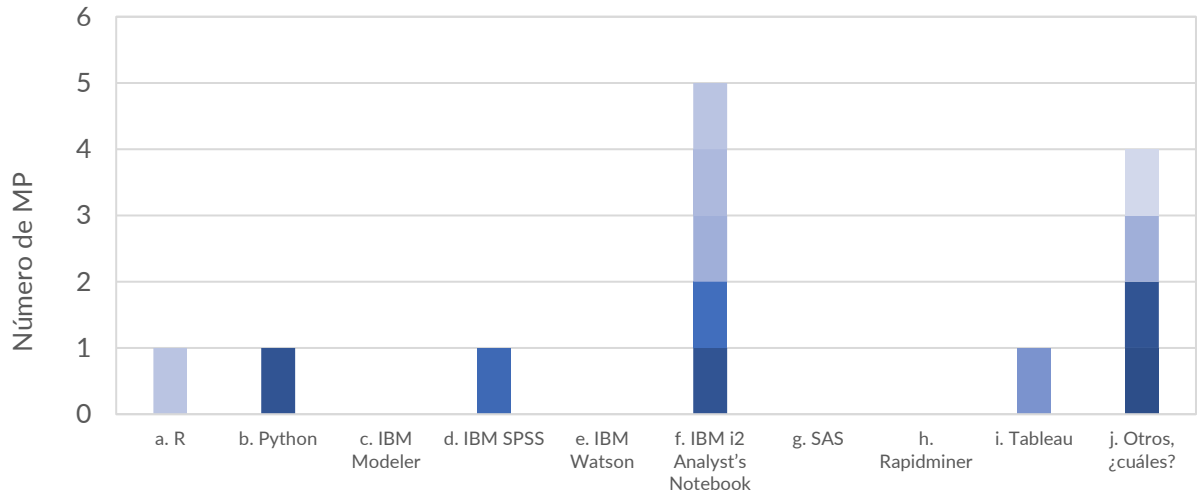

Gráfico 14. Hardware utilizado por los MP para la administración de datos

52. Las metodologías de análisis de datos utilizadas por los MP se enmarcan, en su mayoría, en la inteligencia de negocios (20% de las entidades), el uso de técnicas del aprendizaje de máquina no supervisado (20%) y el análisis de redes complejas (20%). En términos generales, el 60% de las entidades realizan análisis descriptivo y particular de información.

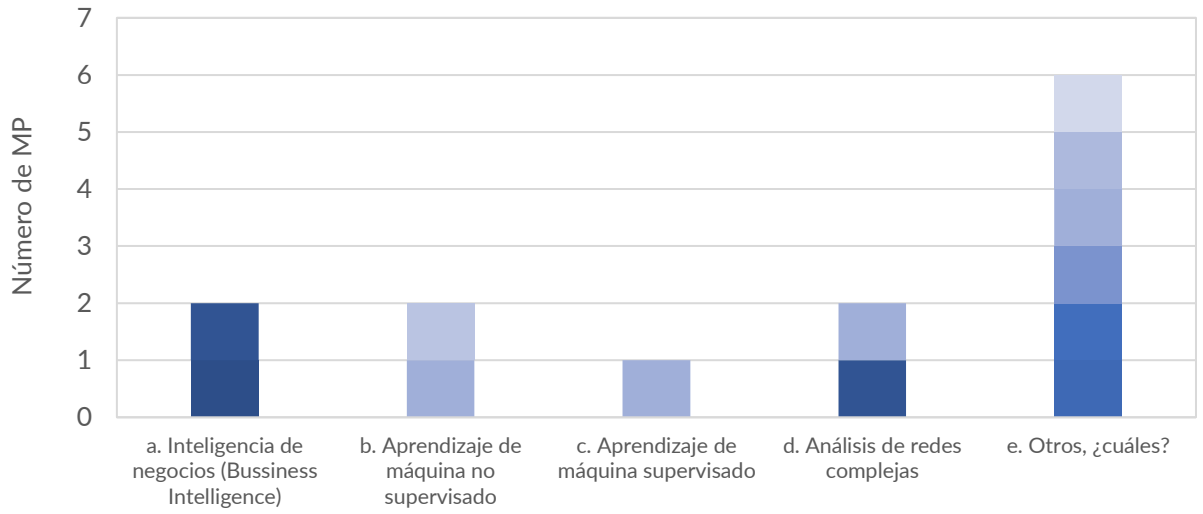

Gráfico 15. Metodologías utilizadas en el análisis de datos por parte del MP

3.2.6. Perfil de funcionarios

53. Esta situación también se observa al analizar el perfil de los funcionarios con los que cuenta el MP para ejecutar labores de analítica de datos, donde el 50% manifiesta no contar con este tipo de profesionales. En algunos casos se mencionan perfiles de ciencias sociales (Contaduría, ingeniería) en la implementación de metodologías de la inteligencia de negocios, el aprendizaje de máquina y el análisis de redes complejas.

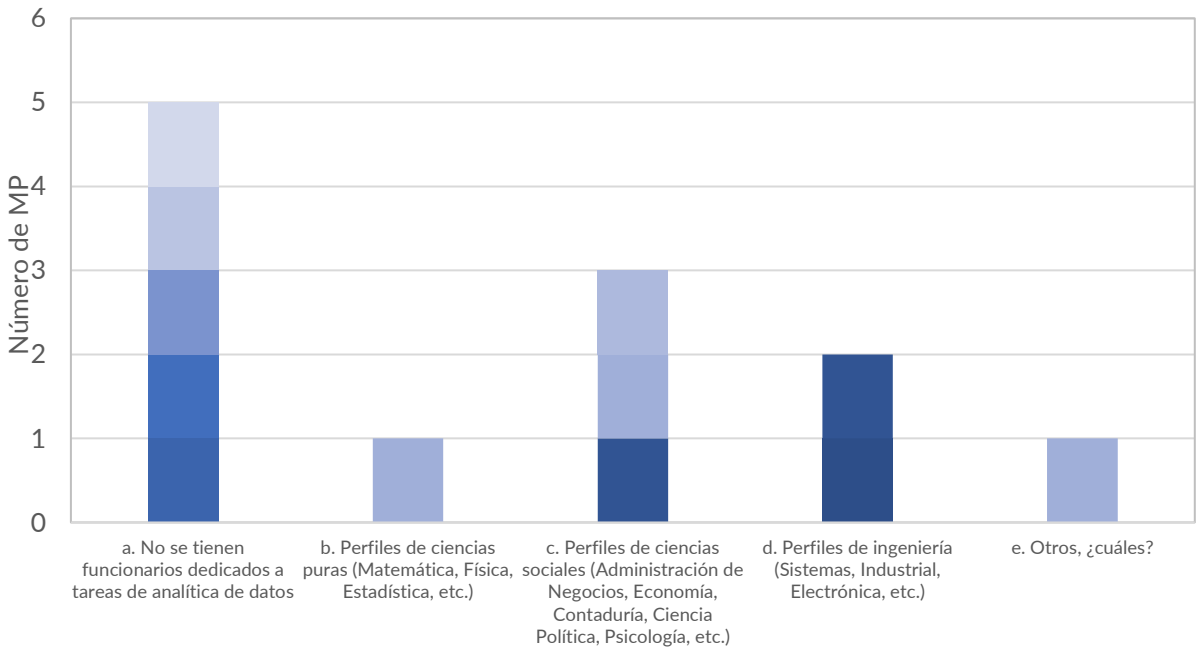

Gráfico 16. Perfil de los funcionarios

4. ALMACENAMIENTO Y ANÁLISIS DE DATOS

54. Un análisis completo sobre los métodos a disposición del Sistema ALA/CFT para la detección y caracterización de operaciones atípicas y sospechosas, debe ocuparse de los aspectos tecnológicos conexos al almacenamiento de información y de las metodologías de procesamiento de datos. De esta manera se tendrá un panorama amplio de los recursos requeridos por las diferentes entidades para aprovechar los desarrollos recientes en big data, gestión de datos, inteligencia artificial y analítica, permitiendo de esta manera incrementar la eficiencia y efectividad del Sistema ALA/CFT.

55. Esta sección inicia cubriendo lo relacionado con el almacenamiento de datos. Posteriormente indaga sobre el big data y el impacto que tiene en la inteligencia financiera. Luego entra de lleno en la presentación de diferentes metodologías para el análisis de datos y la generación de información útil para el Sistema ALA/CFT. Posteriormente se introducen dos

fuentes no tradicionales de datos, y se cierra con la presentación de algunos casos de uso en el análisis de inteligencia financiera, y la generación de estudios estratégicos y operativos para la identificación de tipologías.

56. Aquí es importante mencionar que los aspectos abordados se asocian generalmente con las UIF, toda vez que estas entidades cuentan con sistemas tecnológicos robustos que apoyan el análisis estratégico y operativo. Sin embargo, los elementos descritos pueden contribuir al desarrollo de modelos útiles para los MP.

4.1. ALMACENAMIENTO DE DATOS

4.1.1. Hardware

57. El primer elemento para considerar en lo relacionado con el almacenamiento de datos es el servidor. De manera general, un servidor es una máquina que atiende peticiones de clientes y devuelve respuestas de acuerdo con un procedimiento establecido. Físicamente los servidores pueden ser cualquier tipo de computador, aunque suelen utilizarse configuraciones sofisticadas de hardware, con procesadores veloces capaces de ejecutar varias tareas a la vez, y con una memoria amplia para poder manejar grandes flujos de información. Adicionalmente, los servidores tienen consideraciones específicas que promueven la seguridad de los sistemas.

58. Un servidor también puede funcionar en un conjunto de dos o más computadores, dando lugar a un clúster de servidores. En este caso las diferentes máquinas funcionan como si fueran una sola, incrementando los recursos computacionales disponibles, aumentando la velocidad de procesamiento y mejorando la disponibilidad del sistema ante fallos. De manera específica, los clústers de servidores tienen las siguientes ventajas:

- Alta disponibilidad: implica que, si se presenta un problema en alguna de las máquinas que conforman el clúster, las demás pueden suplir sus funciones y garantizar que las ejecuciones de tareas sigan ocurriendo de la manera esperada.
- Alta velocidad: al contar con varios computadores, las peticiones que realicen los clientes se reparten entre los recursos disponibles permitiendo la ejecución simultánea de labores y la reducción en los tiempos de respuesta.
- Balanceo de carga: distribuye las tareas para que ninguna de las máquinas se sature.
- Escalabilidad: en caso de requerirse, es posible agregar servidores al clúster para aumentar la capacidad de procesamiento.

59. Pese a sus virtudes, los clústers de servidores también presentan algunas desventajas, particularmente relacionadas con la complejidad para la configuración de estos sistemas y la dificultad de contar con personal capacitado para dar mantenimiento a estas infraestructuras. Por estos motivos, una alternativa que surge es el uso de servicios en la nube que utilizan clústers de servidores, como los ofrecidos por las principales compañías de tecnología (Microsoft, Amazon, IBM, Google, entre otras).

60. En estos casos, es posible utilizar recursos a la medida de las necesidades de cada entidad y pagar solamente por el uso que se les dé. Sin embargo, la información que maneja el Sistema ALA/CFT es, en general, sensible y de carácter reservado, e incluso las leyes de muchas jurisdicciones prohíben de manera categórica la disponibilidad de estos datos por fuera de la jurisdicción, con lo cual se eliminan las posibilidades de utilizar este tipo de recursos.

61. La mayoría de las UIF utilizan un servidor para el almacenamiento y administración de los datos que reportan los sujetos obligados. Desde aquí se realizan copias de seguridad que garantizan que la información no se pierda, aún en eventos catastróficos. También, se toman medidas que permiten mejorar la seguridad del sistema, entre ellas, la gestión de los usuarios, otorgándoles accesos limitados a través de privilegios que se otorgan por las funciones que tienen en la entidad. Finalmente, el servidor es quien procesa las solicitudes de consulta y devuelve la información requerida para soportar los procesos operativos y estratégicos de la UIF. Estas consideraciones también aplican para los MP, aunque considerando que en muchos casos requerirán de información puntual, lo cual supone una menor demanda de recursos computacionales.

62. Para tareas masivas de consulta de información, o para procesamientos avanzados de datos, se puede considerar una estructura de varios servidores. De esta manera se contará con mayor capacidad computacional para soportar consultas a tablas con millones o, incluso, miles de millones de registros, permitiendo operaciones de unión o intersección que típicamente son demandantes o imposibles para sistemas más elementales. Por supuesto, como se mencionó anteriormente, estas ventajas vienen asociadas a mayores costos de configuración y mantenimiento, y pueden también implicar recursos adicionales para el aprendizaje de nuevas formas de programación, necesarias para relacionarse con esta infraestructura.

4.1.2. Bases de datos y sistemas de gestión de datos

63. Habiendo considerado lo relacionado con el hardware de almacenamiento y procesamiento, los siguientes elementos a discutir son las bases de datos y los sistemas de gestión de base de datos. En primera instancia, las bases de datos son un conjunto ordenado de información que se conserva de manera electrónica en un computador.

64. El almacenamiento, modificación y extracción de esta información se hace mediante el sistema de gestión de bases de datos, que es un conjunto de programas que permiten la ejecución de estas tareas y, además, mantienen la integridad de los datos, administran el acceso de los usuarios a la información en función de los permisos que tengan asignados y recuperan el sistema en caso de una falla.

65. Los primeros sistemas de gestión de bases de datos aparecen en los años sesenta y setenta, con el objetivo de permitir el acceso a conjuntos de datos caracterizados por sus

interrelaciones complejas. Estos sistemas respondían a las limitaciones de los sistemas operativos y el hardware de la época, resultando en esquemas centralizados donde se contaba con un solo computador para toda la organización que se accedía a través de terminales sencillas que no disponían de memoria.

66. En los años ochenta aparecen los computadores personales y, con ellos, la necesidad de adaptar los sistemas de gestión de base de datos para que sean descentralizados, más flexibles y fáciles de diseñar y mantener. A mediados de esta década se establece el Lenguaje Estructurado de Consultas SQL[3], que da origen al paradigma de bases de datos relacionales y facilita la programación de aplicaciones con bases de datos.

67. Para los años noventa los sistemas relacionales se popularizan a nivel empresarial. Sin embargo, esta misma popularidad hace que una organización tenga varios sistemas, no necesariamente del mismo proveedor, con lo que surge la necesidad de poder contar con un esquema integrado que permita acceder a la información de bases de datos que se encuentran en plataformas heterogéneas ubicadas en diferentes áreas de la empresa.

68. Esto se consigue principalmente mediante el uso del lenguaje SQL, que permite que diferentes sistemas de gestión de bases de datos se comuniquen entre sí y se comporten como si fueran un programa único. Esta unión de sistemas conforma una base de datos distribuida, que tiene ventajas en términos de disponibilidad, porque el fallo en uno de sus componentes no implica el fallo del sistema como un todo, en reducción de costos y en flexibilidad para adaptarse a las necesidades de las entidades.

69. En la actualidad ya no sólo es necesario almacenar la información numérica tradicional (datos estructurados), sino que surge la necesidad de preservar archivos de imagen y sonido, textos, ubicaciones, interacciones, entre otros (datos no estructurados). Como respuesta han surgido sistemas de gestión de bases de datos más allá del paradigma relacional, denominados NoSQL, que se identifican por no utilizan el lenguaje SQL como lenguaje principal de consulta de datos.

4.1.3. Bases de datos de grafos

70. Entre las entidades del Sistema ALA/CFT es común encontrar bases de datos relacionales, que utilizan SQL como lenguaje de consulta. En estos casos los datos son estructurados y se pueden representar mediante tablas. Sin embargo, tanto las UIF como los MP también acceden a información sobre las relaciones que se dan entre individuos, particularmente cuando participan como partes y contrapartes en transacciones económicas y financieras. Para el manejo de estos datos existen las bases de datos de grafos, que son sistemas diseñados específicamente para almacenar y recorrer relaciones.

[3] Acrónimo en inglés para Structured Query Language

71. Los elementos fundamentales de una base de datos de grafos son los nodos y los bordes. Los nodos se refieren a los individuos sobre los cuales se almacena información, y los bordes son las relaciones que se generan entre ellos. Para el caso de la inteligencia financiera, los nodos pueden ser personas naturales o jurídicas, propiedades, cuentas, entidades, jurisdicciones, etc., y los bordes serían transacciones financieras, relaciones de propiedad, relaciones comerciales o conocidos en común, entre otras. Los bordes pueden indicar la dirección de la relación y no hay límite para la cantidad y el tipo de relaciones que un nodo tiene.

72. La representación visual de las bases de datos de grafos es directa, lo cual es un punto atractivo adicional de esta tecnología. Por ejemplo, el Gráfico 17 muestra las relaciones entre un grupo de personas naturales (círculos) y personas jurídicas (triángulos). Las líneas sólidas indican la dirección de los pagos registrados entre los individuos, con un grosor que representa el valor de la transacción, y las líneas punteadas corresponden a relaciones de conocimiento. En rojo se encuentran los nodos que han sido reportados en un ROS.

73. Para este caso, la sospecha del individuo 4 puede contagiarse hacia las empresas A y B por las transacciones financieras que los vinculan, en todo caso, con mayor intensidad hacia la empresa B por el mayor valor de la relación. La sospecha también podría contagiarse hacia el individuo 3 pues existe una relación de conocimiento, aunque en este caso el vínculo parece ser menos fuerte y fuera del contexto económico.

74. Si se encontrara alguna operación ilícita en la empresa A, esta sospecha podría contagiarse hacia el individuo 2. El individuo 1 parece libre de sospecha por la distancia en sus relaciones con el individuo 4, aunque no deja de estar vinculado en esta red de nodos y bordes.

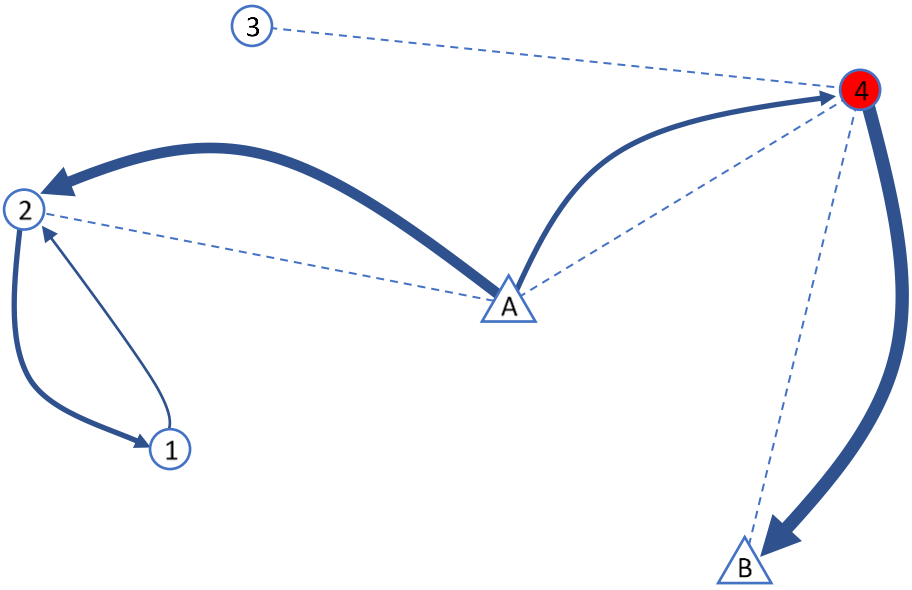

Gráfico 17. Relaciones entre un grupo de personas naturales y jurídicas

75. La información de una base de datos de grafos se recorre desde los bordes, lo cual hace que las consultas sobre las interacciones entre nodos sean muy rápidas. Esto sucede porque las relaciones entre nodos no se calculan, sino que se definen previamente en la base de datos.

76. La característica de velocidad mencionada anteriormente, junto con la representación gráfica que resulta tan cercana a la inteligencia financiera, hacen que las bases de datos de grafos puedan ser consideradas como una estructura adecuada, incluso idónea, para el almacenamiento de datos y procesamiento de información en las entidades del Sistema ALA/CFT. Ahora, una vez más, estas ventajas vienen al costo de utilizar un sistema menos conocido, para el cual no necesariamente existe una oferta apropiada de personal calificado, aumentando los costos de implementación y mantenimiento.

4.2. BIG DATA

77. *Big data* es un término que se ha utilizado para referirse a la gran cantidad de información que se ha ido generando en los años recientes, sobre todo a partir de la popularización de Internet y con la creación de sensores. Internet porque el crecimiento de la red ha ido generando datos sobre los millones de sitios que han sido creados, y sobre la interacción que los usuarios tienen con estas páginas web y con otros usuarios.

78. Más recientemente, la masificación de las redes sociales impulsó el crecimiento de la información disponible en forma de contenidos, aprobaciones, opiniones e interacciones. Sensores, porque cada vez son más los aparatos que generan señales sobre todo tipo de situaciones: ubicación, consumo de energía, formas de uso, conexión con otros dispositivos, etc.

79. Esto es aún más cierto desde la creación de los teléfonos celulares inteligentes, que tienen por lo menos seis tipos de censores distintos (Masoud, 2019), y con el internet de las cosas, donde todo tipo de artefactos (electrodomésticos, automóviles, maquinaria, *wearables*, etc.) hacen mediciones permanentes de su estado y actividad.

80. El *big data* se ha definido a partir de las Vs, que inicialmente eran tres, pero actualmente llegan a ser cinco[4]: Volumen, Velocidad, Variedad, Veracidad y Variabilidad. Volumen se refiere directamente a la cantidad de información que se genera y se necesita almacenar, que hubiera sido un problema hace un par de décadas pero que en la actualidad es cada vez más manejable.

81. Velocidad es el ritmo al cual se genera nueva información, que llega a ser en tiempo real por el funcionamiento de una gran variedad de sensores. Variedad habla de los tipos de datos que se generan, que ya no sólo son valores numéricos organizados en tablas, sino que actualmente también son textos, imágenes, sonidos, videos e interacciones, entre otras. Veracidad incluye todas las tareas requeridas para garantizar que la información es confiable y representativa de la realidad. Variabilidad se enfoca en la posibilidad que tienen los datos de cambiar en su estructura y en la información que contienen.

82. Todas estas Vs imponen retos particulares que deben ser abordados de maneras específicas para aprovechar la nueva información y, así, generar conocimiento amplio y oportuno sobre los fenómenos de interés.

83. Para el Sistema ALA/CFT, el *big data* se manifiesta principalmente en el Volumen, la Variedad y la Veracidad. Inicialmente, las UIF y los MP recolectan cada vez más información de transacciones económicas y financieras, con tablas de datos que pueden llegar a miles de millones de registros. Sin embargo, esta situación no debería representar un reto particularmente difícil, pues las nuevas tecnologías de hardware tienen la capacidad de almacenar y procesar estos datos a costos que son cada vez menores, y sí configura una oportunidad valiosa porque ahora las entidades tienen más información a su alcance.

84. Ciertamente, la Variedad puede imponer retos mayores para las entidades del Sistema ALA/CFT, porque el uso de los diferentes tipos de datos que están disponibles implica que los sujetos obligados, las UIF y los MP cuenten con personal capacitado para procesar no sólo información estructurada sino también, por lo menos, textos y datos relacionales, y aunque estos perfiles están cada vez más disponibles, su oferta no es abundante y esto impone costos y dificultades de vinculación.

85. Aunque la Veracidad es altamente relevante para el análisis masivo de datos en la inteligencia financiera, el propio proceso de investigación de operaciones implica una verificación

[4] https://www.sas.com/es_co/insights/big-data/what-is-big-data.html acceso el 7 de julio de 2021.

de información que garantiza que la identificación de casos atípicos no obedece a errores o imprecisiones en los registros. De todas formas, se recomienda tomar medidas para revisar y validar la información suministrada por los sujetos obligados, y es siempre provechoso contar con tablas de datos depuradas[5] sobre las cuales se puedan aplicar con mayor confianza los algoritmos y modelos seleccionados para la identificación de situaciones de interés.

86. Velocidad y variabilidad también son relevantes, aunque no en la misma medida que las Vs tratadas anteriormente. En general, el análisis de inteligencia financiera toma tiempo en recolectar toda la información pertinente y en construir los casos, haciendo menos relevante el procesamiento inmediato de la nueva información disponible. La variabilidad impactará en los desarrollos que se hagan para la recepción y consumo de datos, pero esto está controlado toda vez que las UIF estandarizan e imponen las condiciones de suministro de datos, dando así estabilidad a lo relacionado con formatos y estructuras de información.

87. Sobre las posibilidades, el *big data* permite caracterizar las situaciones económicas de una manera más completa, incorporando dimensiones que anteriormente no se consideraban o no se podían incluir en el análisis. Por ejemplo, contar con variables sobre las relaciones entre individuos o los términos utilizados para describir un ROS, permite obtener modelos que caractericen mejor los comportamientos típicos de las personas, para así identificar situaciones atípicas, o que sean más precisos en el pronóstico masivo de casos que deban ser considerados para análisis más detallados.

4.3. RELACIÓN ENTRE LA INTELIGENCIA ARTIFICIAL, EL APRENDIZAJE DE MÁQUINA, Y LA MINERÍA DE DATOS Y TEXTOS

4.3.1. Aspectos Generales

88. Inteligencia artificial, aprendizaje de máquina y minería de datos son términos que en algunas ocasiones se utilizan indistintamente para referirse a las mismas situaciones. La verdad es que son conceptos relacionados, donde incluso algunos de ellos están contenidos en otros. Es por lo que esta sección sobre metodologías de análisis de datos se inicia con las definiciones respectivas, y presenta una discusión sobre similitudes y diferencias entre los conceptos.

89. En 1950, Alan Turing publica el documento titulado *Computing Machinery and Intelligence*, donde inicia preguntándose si las máquinas pueden pensar[6]. Es sobre esta pregunta que trabaja la inteligencia artificial, rama de la ciencia de la computación que busca simular la inteligencia humana en las máquinas. Sin embargo, no existe una definición ampliamente aceptada de

[5] En el análisis de información es común considerar un preprocesamiento de los datos, donde se eliminan registros repetidos, se identifican observaciones extrañas (*outliers*) y se imputan observaciones faltantes. No es inusual que una cantidad considerable de recursos se dediquen a estas tareas.

[6] La primera frase del artículo es "I propose to consider the question, "Can machines think?"". A. M. Turing (1950). Computing Machinery and Intelligence. Mind 49: 433-460.

inteligencia artificial (Wang, 2019) puesto que no hay un consenso sobre qué es una inteligencia artificial o qué hace que una máquina pueda considerarse inteligente.

90. Diferentes autores han aproximado esta cuestión, por ejemplo, a partir de postulados sobre el pensamiento humano, el pensamiento racional, la actuación humana y la actuación racional, o desde la posibilidad de que una máquina ejecute tareas que suelen requerir de la inteligencia humana.

4.3.2. Inteligencia artificial y sistema ALA/CFT

91. Bajo estas consideraciones, para el sistema ALA/CFT la inteligencia artificial podría pensarse como la capacidad que tendría una máquina de ejecutar los análisis estratégicos y operativos de la inteligencia financiera para identificar operaciones atípicas y sospechosas, tomando como insumo la información recolectada por las UIF y los MP.

92. Los datos permitirán que la inteligencia artificial conozca las situaciones que han ocurrido anteriormente, separe las señales útiles de aquellas que no fueron consideradas por los analistas expertos y genere procesos para actuar sobre situaciones futuras o, en otras palabras, le servirán como experiencia sobre la cual podrá aprender y proyectar sus acciones. Es en este punto donde entra el aprendizaje de máquina, rama de la inteligencia artificial que combina algoritmos[7] y datos para tratar de emular la forma en la que los humanos aprenden, pretendiendo incluso mejorar la precisión de los resultados[8].

93. Para entender la forma en que el aprendizaje de máquina funciona, se pueden considerar tres elementos: insumos, que son señales o estímulos ante los cuales se quiere dar una respuesta; reglas, que son la forma en la que se espera actuar ante los estímulos; y respuestas, que son las acciones finales que se quieren ejecutar ante una señal.

94. En el sistema ALA/CFT los insumos serían las diferentes transacciones económicas que realizan los individuos, o también los hechos judiciales que los afectan; las reglas serían los procedimientos de inteligencia financiera que se aplican para identificar e investigar señales de alerta; y las respuestas serían la clasificación de los individuos investigados entre sospechosos y no sospechosos. Así, anteriormente la inteligencia artificial actuaba mediante sistemas expertos, que son procesos informáticos donde se programan explícitamente las reglas que se van a utilizar para procesar los insumos disponibles, generando de esta manera las repuestas esperadas.

95. En el centro de estos sistemas se encuentran las reglas, diseñadas por expertos que las definen a partir de su experiencia. El problema con estos sistemas es que son engorrosos de definir y programar porque las reglas pueden ser muy complejas, y que están limitados al

[7] Un algoritmo es una secuencia de acciones lógicas mediante la cual se pretende solucionar un problema. Por ejemplo, una receta culnaria es un algoritmo porque indica los pasos que se deben seguir para conseguir una preparación.
[8] https://www.ibm.com/cloud/learn/machine-learning

conocimiento de los especialistas. En contraposición, actualmente la inteligencia artificial utiliza el aprendizaje de máquina, que analiza tanto los insumos como los resultados disponibles para definir automáticamente las reglas que los relacionan, permitiendo implementar sistemas que las apliquen de forma masiva.

96. La ventaja de esta metodología es que puede identificar relaciones complejas no evidentes en poco tiempo, aplicando algoritmos especializados. La desventaja es que las reglas definidas dependen de los datos consumidos por los algoritmos, limitándose a las situaciones contenidas en esta información. Sin embargo, esto es un problema cada vez menos relevante por la gran cantidad y variedad de datos disponibles en la actualidad.

4.3.3. Aprendizaje de máquina aplicado al sistema ALA/CFT

97. Un ejemplo del uso del aprendizaje de máquina en el sistema ALA/CFT es un sistema de clasificación automática de ROS. Este sistema puede construirse sobre los individuos relacionados, utilizando como insumos las características disponibles de las personas naturales o jurídicas que se mencionan en el reporte, y como repuestas un indicador numérico de su importancia.

98. En este caso, el aprendizaje de máquina encontrará reglas que le permitan definir qué combinaciones de valores en las variables económicas, financieras, judiciales, etc., de los individuos se asocian con un ROS de alta importancia y, a su vez, estas reglas podrán ser programadas para que el sistema continúe autónomamente con la calificación. Este sistema podrá actualizarse o ajustarse cada vez que se requiera a partir de la nueva información disponible para mantener su vigencia, incorporando nuevas metodologías utilizadas por los delincuentes en relación con LA y FT.

4.3.4. Minería de datos

99. Por su parte, la minería de datos se define como el estudio de la recolección, limpieza, procesamiento y extracción de conocimiento a partir de diferentes conjuntos de datos (Aggarwal, 2015). También, es usual que se le ubique como la parte central del proceso de generación de conocimiento en bases de datos (Fu, 1997).

100. Aunque tanto el aprendizaje de máquina como la minería de datos son intensivas en el uso de información, se diferencian en que la primera busca replicar la acción humana para automatizar procesos, sin que sea necesario entender las situaciones sobre las cuales modela, mientras que la segunda se preocupa por conocer la forma como se relacionan las variables y, de esta manera, explicar los mecanismos involucrados en las situaciones que estudia. En este sentido la minería de datos se relaciona con el análisis estadístico, utilizando incluso muchas de sus herramientas y metodologías.

101. Retomando el ejemplo del sistema de clasificación automática de ROS, mientras que con el aprendizaje de máquina se buscan reglas sobre las variables disponibles que se puedan aplicar para clasificar los reportes, la minería de datos trata de entender cuáles son los determinantes que hacen que un reporte sea más o menos importante. De todas formas, cuando se consideran modelos más sencillos, tanto el aprendizaje de máquina como la minería de datos llegan a resultados y conclusiones similares.

4.3.5. Minería de textos

102. Por último, la minería de textos es, en esencia, igual a la minería de datos, con la diferencia que ésta se concentra en generar variables numéricas a partir de textos. Aquí se trabaja con datos, que bien pueden ser las descripciones que se incluyen en los ROS, y se generan variables como, por ejemplo, la frecuencia de aparición de una palabra o frase determinada. Como se puede observar, los textos se traducen a números y desde este punto se puede utilizar tanto el aprendizaje de máquina como la minería de datos, dependiendo del objetivo que se tenga de automatización o descripción de la situación.

103. En otras palabras, la minería de textos comprende un conjunto de herramientas para procesar textos en variables numéricas que sirven como insumos para otros modelos. De todas formas, es importante mencionar que la minería de textos tiene procedimientos propios que permiten entender el contenido de los documentos que analiza, como las nubes de palabras, que muestran de manera gráfica la cantidad de veces que se repiten determinados términos, permitiendo de esta manera formarse una idea rápida del contenido de un documento.

104. En resumen, existen número amplio de metodologías que se agrupan bajo diferentes tipos de análisis, que comparten la necesidad de contar con información como insumo, pero que se diferencian en su objetivo.

105. Para el Sistema ALA/CFT es importante comprenderlas, implementarlas y aplicarlas, por las posibilidades que tienen en términos de generar eficiencias en el análisis por la automatización de procesos, y por el conocimiento adicional que pueden generar al procesar más información que la que un equipo humano podría llegar a considerar.

106. Ahora, es importante mencionar que todos estos procedimientos no reemplazan, al menos no de momento, la experiencia y conocimiento de los analistas, sino que los potencian al ocuparse de tareas rutinarias y al generar información extra que, junto con el análisis y la experiencia humana, permiten llegar a la identificación de nuevas metodologías utilizadas por los delincuentes para el blanqueo de capitales o la financiación del terrorismo.

4.4. METODOLOGÍAS DE APRENDIZAJE DE MÁQUINA NO SUPERVISADO

107. En el aprendizaje de máquina no supervisado se agrupan un conjunto de metodologías que se utilizan principalmente para tres tareas: primero, identificar agrupaciones naturales de individuos o de variables a partir de la información que se tenga disponible; segundo, hallar asociaciones entre estos individuos o variables; y, tercero, reducir la dimensionalidad de conjuntos de datos por la generación de nuevas variables que resumen otras. La primera de estas tareas, que es particularmente importante para el Sistema ALA/CFT por las posibilidades que tiene en las labores de segmentación (Monetary Authority of Singapore, 2018), se ejecuta bajo el nombre de análisis clúster.

108. En el análisis clúster los individuos que pertenecen a un mismo grupo deben ser parecidos, mientras que los grupos deben ser diferentes entre sí. Aquí, el concepto de parecido (o diferente) depende de algún tipo de distancia entre observaciones.

109. En la práctica, se deben tener en cuenta las siguientes recomendaciones:

a. Los datos deben estar organizados con los elementos que interesa agrupar por filas, y las características de estos elementos por columnas.
b. Se debe evitar la presencia de información faltante o valores perdidos. Si existen, se deben reemplazar por valores factibles o eliminar la observación, teniendo muy en cuenta las posibles consecuencias en términos de sesgo en los datos.
c. Las variables deben tener escalas similares.

110. Con respecto al último punto, los algoritmos de clústering son, en general, sensibles a la escala de las variables. Por ejemplo, en una tabla de datos de individuos se pueden tener variables como el ingreso y la cantidad de ROS en las que está relacionada la persona. Aquí, el ingreso suele venir en unidades monetarias y tener valores de miles o incluso de millones, mientras que la cantidad de reportes son unidades de baja cuantía. De esta forma, la escala de las variables es muy diferente y el algoritmo tenderá a agrupar con base en la variable de mayor escala, en este caso el ingreso. Para evitar esto, las variables pueden estandarizarse (llevar a series con media 0 y varianza 1) o reescalarse (todas tienen el mismo mínimo y máximo)[9].

111. Como se mencionó anteriormente, el Sistema ALA/CFT se beneficia del análisis clúster porque le permite generar grupos homogéneos de individuos (segmentos) y, a partir de estos, identificar conductas atípicas que pueden llegar a asociarse con comportamientos sospechosos de delitos relacionados con el LA y la FT. En otras palabras, conociendo la habitualidad de un grupo de personas es posible reconocer cuando alguna de ellas se está comportando de manera extraña, en comparación con sus pares.

[9] Los procesos de estandarización o reescalamiento no parecen tener mucho sentido si las variables que describen al individuo son categóricas, es decir, indican características como el sector económico. En este caso, se debe recurrir a opciones distintas como la construcción de variables dummys (dicotómicas, que valen 1 cuando se cumple una condición o 0 si no se cumple) o utilizar medidas de distancia capaces de tratar con este tipo de información.

112. En lo que sigue de esta sección se desarrolla en detalle el concepto de distancia y se presentan las metodologías de clústering particional y jerárquico. También se comenta sobre el análisis de anomalías.

4.4.1. Medidas de distancia

113. La distancia entre dos observaciones mide su similitud, donde valores más pequeños indicarán observaciones más parecidas entre sí, y valores más altos representarán lo contrario. Para un grupo de individuos se puede obtener una matriz que contenga la distancia entre todos ellos, obteniendo como resultado una matriz de distancias cuadrada (igual número de filas y columnas), con ceros en su diagonal principal. Dos formas usuales de calcular la distancia entre un par de individuos 1 y 2, cada uno de ellos definido a partir de p variables, son:

- Distancia Euclídea: $d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\sum_{j=1}^{p}(x_{j1} - x_{j2})^2}$
- Distancia Manhattan (o del taxista): $d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{p}|x_{j1} - x_{j2}|$

Aquí x_{ji} representa el valor que toma la variable j para el individuo $i = 1,2$.

114. Muchas de las medidas de distancia aplican únicamente a variables numéricas, pero en la práctica es usual que las características de los individuos estén definidas como series ordinales o incluso categóricas. Para estos casos se ha desarrollado la distancia de Gower, que agrupa las variables de un conjunto de datos de acuerdo con su naturaleza y aplica las medidas de distancias pertinentes según cada caso: para variables numéricas y ordinales se utiliza la distancia Manhattan (con un ajuste por empates para el segundo caso), y para variables nominales con k categorías se generan k variables indicadoras (dicotómicas) y luego se aplican medidas de similaridad adecuadas. Las distancias de cada caso se combinan de manera lineal, usualmente como un promedio simple.

115. La mayoría de software utiliza por defecto la distancia Euclídea, pero en muchos casos es adecuado utilizar otras medidas, como cuando se quiere medir la cercanía espacial entre dos puntos en una ciudad, caso en el cual puede tener más sentido una medida como la Manhattan.

116. Por ejemplo, considérense 10 empresas, cada una de ellas definida a partir de dos variables: $\boldsymbol{x}_1$, el valor de los activos, y $\boldsymbol{x}_2$, el valor de los pasivos. Los datos originales son los que aparecen en la segunda y tercera columna de la Tabla 5. De acuerdo con la recomendación hecha anteriormente, las series se estandarizan, haciendo que cada una tenga media 0 y desviación estándar 1. Los resultados de este proceso se observan en las columnas 4 y 5 de la misma tabla. Las últimas dos filas muestran los promedios y desviación estándar de las variables, que pasan a 0 y 1, respectivamente, por el proceso de estandarización. El Gráfico 18 muestra la información

estandarizada, aprovechando que los datos se pueden representar mediante un diagrama de dispersión[10].

Empresa	Activo (miles USD)	Pasivo (miles USD)	Activo reescalado	Pasivo reescalado
1	1.869	1.529	-0,364362	0,017669
2	977	714	-0,541611	-0,425727
4	1.737	852	-0,390654	-0,350848
5	860	473	-0,564794	-0,556748
6	954	340	-0,546214	-0,629334
8	1.113	457	-0,514632	-0,565706
9	7.448	1.678	0,744112	0,098798
10	1.048	364	-0,527480	-0,616277
11	16.723	6.414	2,587105	2,674868
12	4.299	2.146	0,118529	0,353304
Media	3.703	1.497	0	0
Desviación estándar	5.033	1.838	1	1

Tabla 5. Valor de los activos y pasivos de 10 empresas.

[10] Para una variable la representación análoga es mediante puntos ubicados en una línea recta. Para tres variables se puede recurrir un diagrama de dispersión tridimensional, aunque visualmente es difícil de interpretar. Con 4 o más variables, en general, no hay posibilidad de representación gráfica.

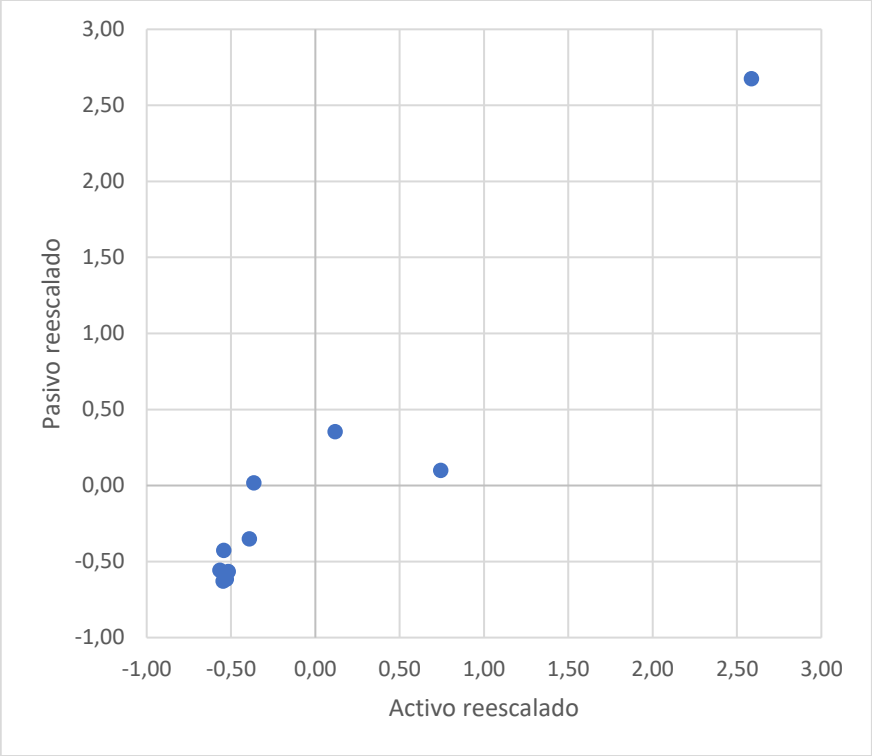

Gráfico 18.Valor de los activos y pasivos, reescalados, de 10 empresas. Fuente:

117. Ahora, como se tienen 10 empresas, se tendrá una matriz de distancias con 10 filas y 10 columnas, simétrica y con ceros en la diagonal principal. Utilizando la distancia Manhattan, se obtienen los resultados presentados en la Tabla 6, donde se omiten los valores del triángulo superior por ser iguales a los del triángulo inferior. Aquí, la distancia entre la empresa 2 y la empresa 1 (resaltada en verde) se calcula como:

$$Distancia\ Manhattan_{2,1} = |-0{,}541611 - (-0{,}364362)| + |-0{,}425727 - 0{,}017669|$$
$$= |-0{,}177249| + |-0{,}443396| = 0{,}177249 + 0{,}443396 = 0{,}620645$$

Empresa	1	2	4	5	6	8	9	10	11	12
1	0,00									
2	0,62	0,00								
4	0,39	0,23	0,00							
5	0,77	0,15	0,38	0,00						
6	0,83	0,21	0,43	0,09	0,00					
8	0,73	0,17	0,34	0,06	0,10	0,00				
9	1,19	1,81	1,58	1,96	2,02	1,92	0,00			
10	0,80	0,20	0,40	0,10	0,03	0,06	1,99	0,00		
11	5,61	6,23	6,00	6,38	6,44	6,34	4,42	6,41	0,00	
12	0,82	1,44	1,21	1,59	1,65	1,55	0,88	1,62	4,79	0,00

Tabla 6. Distancias Manhattan entre 10 empresas. Fuente:

118. Aunque estos valores no tengan una interpretación directa, sus magnitudes se pueden comparar entre sí permitiendo identificar, por ejemplo, que las empresas 1, 2, 3, 4, 5, 6 y 8 están cerca, es decir, son similares entre ellas. También, que las empresas 7 y 10 están relativamente alejadas de las demás, y que la empresa 9 es completamente diferente.

4.4.2. Clústering particional

Los métodos de clústering particional buscan clasificar los individuos en uno de varios grupos posibles generados con base en las características disponibles. Para esto se utiliza su similaridad (distancia), tratando que todas las observaciones de un mismo grupo sean parecidas, mientras que los grupos son lo más diferentes posible entre sí. En esta sección se presenta el método de K-*medias* y K-*medioides*.

4.4.2.1. Clústering por K-medias

119. El procedimiento de K-*medias* fue propuesto por MacQueen en 1967, y busca dividir un conjunto de datos en K grupos, minimizando la distancia intraclase y maximizando la distancia interclase. Cada grupo está caracterizado por su promedio p dimensional, es decir, por el vector de los promedios de las p variables que describen a los individuos que lo conforman.

120. En primera instancia, se define la variación total al interior de un clúster como la suma de las distancias entre los individuos y el centroide del grupo:

$$W(C_k) = \sum_{\boldsymbol{x}_i \in C_k} dist(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$$

121. donde $\boldsymbol{x}_i$ es un vector con la información de todas las variables para el individuo i, $\boldsymbol{\mu}_k$ es el vector de medias para las observaciones que pertenecen al grupo k, y $dist$ es una medida de distancia, usualmente la distancia Euclídea. De acuerdo con la ecuación, para cada uno de los individuos que pertenecen al clúster k, se calcula su distancia al centroide del grupo y, posteriormente, se obtiene la suma de todas estas distancias. Un valor pequeño indica que los individuos del grupo están cerca del centro, lo cual es indicativo de un grupo compacto conformado por individuos parecidos.

122. Adicionalmente, se define la variación total entre clústers como:

$$VTC = \sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in C_k} dist(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$$

123. Es decir, la variación total entre clústers es la suma de las variaciones totales al interior de un clúster para todos los K clústers que se estén considerando.

124. Para obtener los K clústers, cada uno de los individuos se asigna aleatoriamente a uno y sólo de los K grupos. Posteriormente se calcula el centroide de cada clúster mediante una medida de tendencia central como el promedio. Así, el centroide del grupo 1 se obtiene como el promedio de las variables de los individuos que se asignaron a este clúster. Posteriormente se calculan las distancias de todos los individuos al centroide de cada uno de los K grupos, y se reasigna el individuo al clúster para el cual tenga más cerca su centroide. Este proceso de cálculo de centroides, cálculo de distancias y reasignación de grupos se repite hasta que ya no haya variación en los clústers.

125. Continuando con el ejemplo de las 10 empresas presentado en la sección anterior, la Tabla 6 muestra el proceso iterativo anteriormente descrito para la conformación de $K = 2$ clústers. Las tres primeras columnas de esta tabla son idénticas a las presentadas en la Tabla 5, donde se tiene un identificador de cada empresa junto con el activo y el pasivo, reescalados para que tengan media 0 y desviación estándar 1. La iteración 0 del proceso inicia con la asignación aleatoria de los individuos a uno de los grupos deseados. De esta manera, las empresas 1, 2 y 5 se clasifican en el segmento 1 y las demás en el segmento 2. Posteriormente, como se muestra en las últimas cuatro filas de la tabla, se calculan los promedios del activo y el pasivo para los clústers 1 y 2, obteniendo los centroides de cada uno de estos grupos. Aquí, el vector conformado por los valores $-0{,}4903$ y $-0{,}3216$ es el centroide del clúster 1 y el vector conformado por los valores $0{,}2101$ y $0{,}1378$ es el centroide del clúster 2. A continuación se calcula la distancia Euclídea de cada empresa al centroide del grupo 1 y al centroide del grupo 2, llegando a los valores presentados en las columnas 5 y 6 de la tabla, respectivamente.

126. La iteración 1 inicia con la reasignación de los individuos. En este caso, una empresa se ubica en el segmento 1 si su distancia al centroide de este clúster es menor que su distancia al centroide del otro clúster, y se ubica en el segmento 2 si ocurre lo contrario. De acuerdo con esto, la columna 7 de la tabla muestra que los individuos 1, 2, 4, 5, 6, 8 y 10 quedan en el segmento 1, mientras que los individuos 9, 11 y 12 se mantienen en el segmento 2. Posteriormente se calculan los nuevos centroides de cada clúster y se obtienen, nuevamente, las distancias de cada empresa al centroide de cada grupo.

127. La iteración 2 ocurre de la misma manera que la iteración 1, reasignando la empresa 12 del clúster 2 al 1, recalculando los centroides de cada segmento y obteniendo las distancias de los individuos a cada centroide. La iteración 3 repite el procedimiento, reasigna la empresa 9 del grupo 2 al 1, obtiene centroides para cada segmento y calcula las distancias de los individuos a estos centroides. Finalmente, la iteración 4 inicia sin modificar la asignación de los individuos a cada grupo, señalando que el algoritmo ha convergido, es decir, que si se continúan con los cálculos se obtendrán siempre los mismos resultados. Así termina el proceso y se obtienen los segmentos finales: el primero conformado por las empresas 1, 2, 4, 5, 6, 8, 9, 10 y 12, y el segundo compuesto por la empresa 11.

128. Visualmente, los clústers finales se pueden observar en el

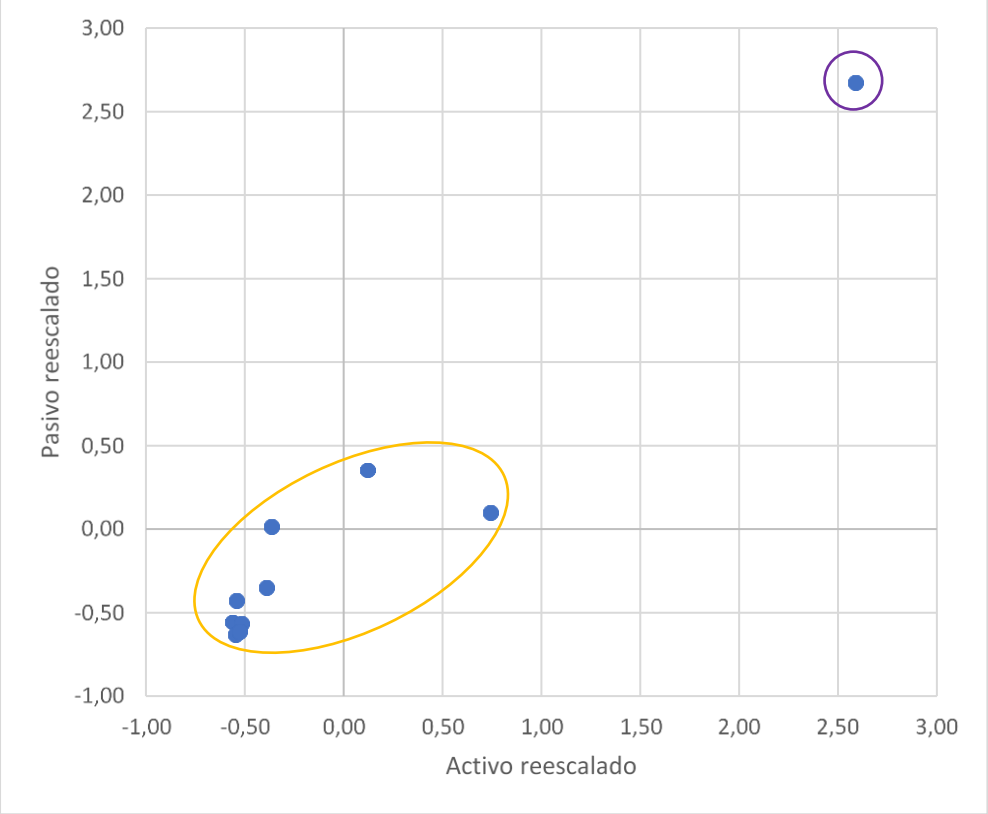

Gráfico 19. El círculo amarillo muestra el segmento 1 y el círculo morado el segmento 2. Ahora, al observar la representación de los individuos en el espacio conformado por el activo y el pasivo reescalados, es interesante pensar que para una persona es casi trivial imaginarse cómo conformar dos grupos de individuos, y observar que seguramente los resultados de este ejercicio son los mismos obtenidos por el algoritmo de *K-medias*.

129. En este punto se evidencia la esencia del procedimiento de aprendizaje no supervisado, que es permitir que una máquina genere agrupaciones similares a las que generaría un humano. Por supuesto, en la medida que se tengan muchos más individuos, que su representación sea a partir de 3 o más variables o que se quieran generar más grupos, la tarea será mucho más compleja para el humano, mientras que la máquina aprovechará su capacidad computacional para resolver el problema de la manera explicada anteriormente.

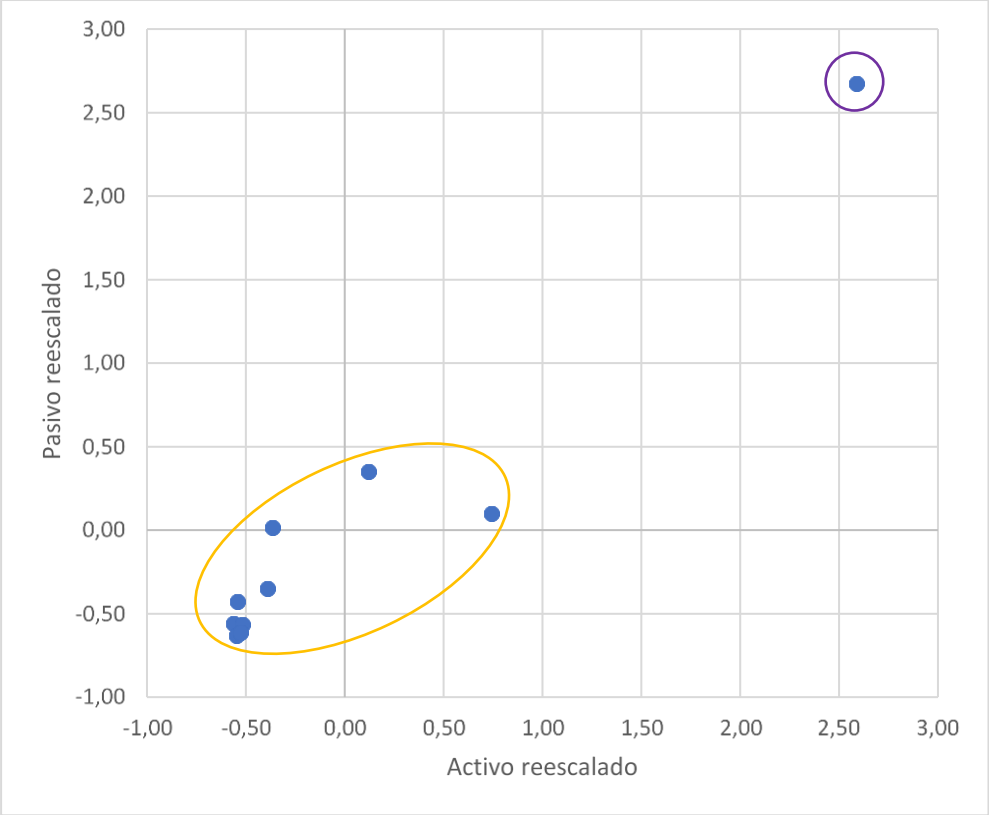

Gráfico 19. $K = 2$ clústers generados a partir del valor de los activos y pasivos, reescalados, de 10 empresas

Empresa	Activo reescalado	Pasivo reescalado	Iteración 0			Iteración 1			Iteración 2			Iteración 3			Iteración 4
			Asignación clúster	Distancia centroide clúster 1	Distancia centroide clúster 2	Asignación clúster	Distancia centroide clúster 1	Distancia centroide clúster 2	Asignación clúster	Distancia centroide clúster 1	Distancia centroide clúster 2	Asignación clúster	Distancia centroide clúster 1	Distancia centroide clúster 2	Asignación clúster
1	-0,3644	0,0177	1	0,3619	0,5869	1	0,4818	1,8284	1	0,3681	2,4485	1	0,3241	3,9714	1
2	-0,5416	-0,4257	1	0,1161	0,9395	1	0,0531	2,2397	1	0,1481	2,8561	1	0,2848	4,4048	1
4	-0,3907	-0,3508	2	0,1038	0,7744	1	0,1401	2,0771	1	0,0261	2,6922	1	0,1163	4,2452	1
5	-0,5648	-0,5567	1	0,2467	1,0406	1	0,1315	2,3446	1	0,2572	2,9584	1	0,3798	4,5142	1
6	-0,5462	-0,6293	2	0,3128	1,0773	1	0,1903	2,3814	1	0,3110	2,9928	1	0,4210	4,5536	1
8	-0,5146	-0,5657	2	0,2453	1,0101	1	0,1210	2,3144	1	0,2400	2,9267	1	0,3517	4,4858	1
9	0,7441	0,0988	2	1,3040	0,5354	2	1,3519	1,0271	2	1,2431	1,5837	1	1,1050	3,1675	1
10	-0,5275	-0,6163	2	0,2970	1,0549	1	0,1731	2,3589	1	0,2916	2,9702	1	0,3993	4,5313	1
11	2,5871	2,6749	2	4,2952	3,4766	2	4,3852	2,1750	2	4,2604	1,5837	2	4,1348	0,0000	2
12	0,1185	0,3533	2	0,9089	0,2341	2	1,0069	1,2404	1	0,8810	1,8605	1	0,7668	3,3887	1
Promedio activo, clúster 1			-0,4903			-0,4928			-0,4164			-0,2875			
Promedio pasivo, clúster 1			-0,3216			-0,4467			-0,3467			-0,2972			
Promedio activo, clúster 2			0,2101			1,1499			1,6656			2,5871			
Promedio pasivo, clúster 2			0,1378			1,0423			1,3868			2,6749			

Tabla 7. Proceso de asignación iterativa de 10 empresas a $K = 2$ grupos

130. La pregunta que surge en este punto es cómo escoger el valor de K. Para esto se ha propuesto ejecutar el análisis clústers con $K = 2, 3, \ldots, N - 1$ (o algún límite superior, usualmente muy por debajo de $N - 1$) y analizar el comportamiento de cada caso en términos de alguna medida de ajuste como la variación total entre clústers.

131. Para las 10 empresas del ejemplo que se ha venido trabajando, el resultado se puede representar visualmente como se muestra en el

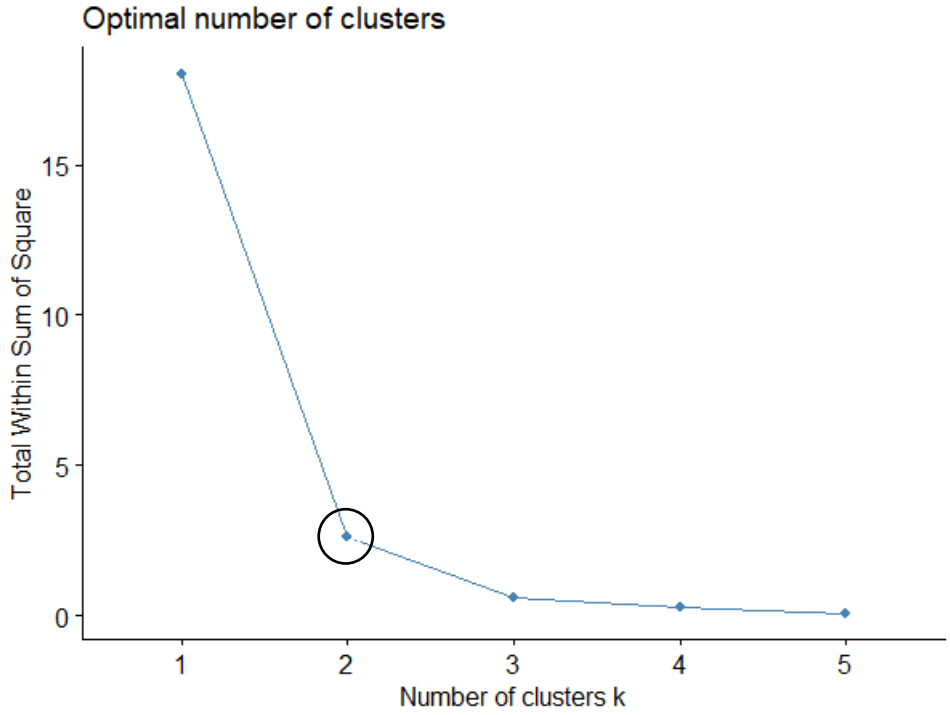

Gráfico 20, que muestra en el eje x diferentes valores de K y en el eje y el ajuste del modelo, donde se prefieren resultados más pequeños. Aquí hay que considerar que, si cada individuo fuera un segmento en sí mismo, es decir, si se generan $K = 10$ grupos, el ajuste sería perfecto pero el resultado no tendría ninguna utilidad. Por esto, se busca el valor de K^* que permita la mayor ganancia en ajuste pero que mejore sólo de manera marginal para valores $K > K^*$.

132. Otra forma de pensarlo es, si la línea azul del

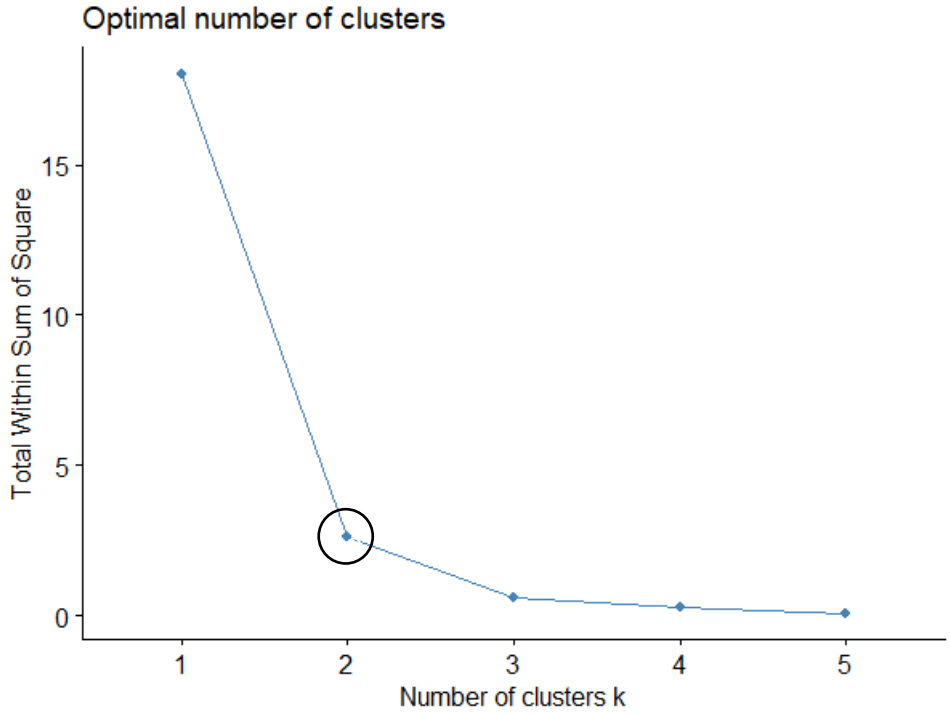

Gráfico 20 fuera un brazo, entonces se busca el valor de K que genere un codo, como sucede con el valor encerrado en el círculo negro en la figura. Como se puede observar, $K = 2$ es la cantidad de clústers que genera la mayor ganancia en ajuste, y es el valor a partir del cual se obtienen mejoras marginales de ajuste al generar una mayor cantidad de segmentos.

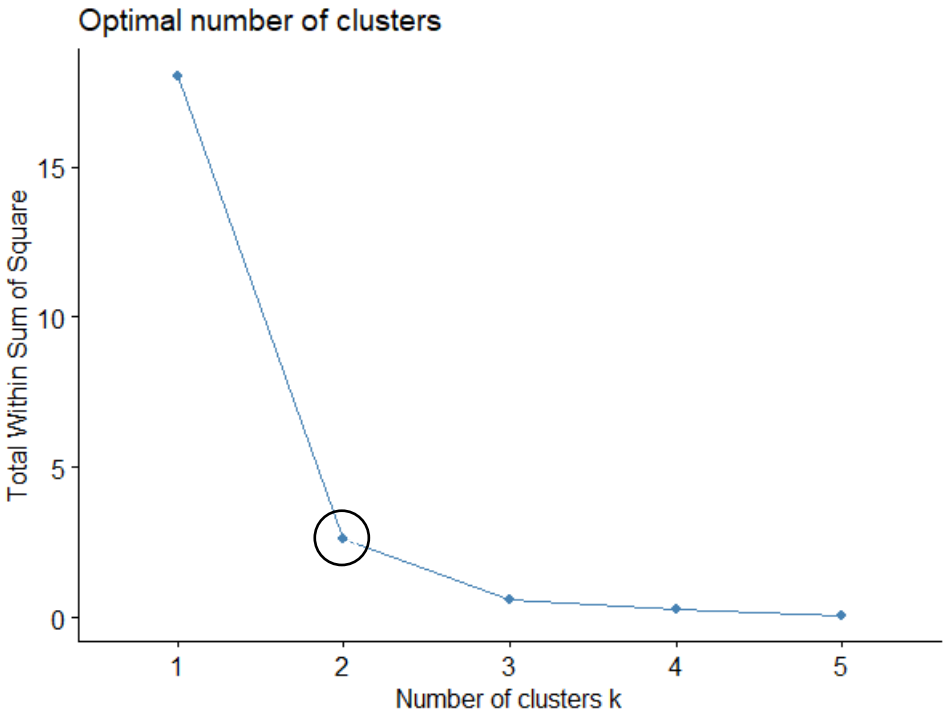

Gráfico 20. Ajuste del análisis clúster para 10 empresas con $K = 1, 2, \dots, 5$

133. A pesar de su versatilidad, el método de clústers por K-*medias* es sensible a observaciones atípicas (*outliers*), a la conformación aleatoria inicial de los grupos y requiere que se decida de antemano la cantidad de segmentos a conformar. Por estos motivos se han desarrollado alternativas que modifican la forma en la que se calculan los centros de cada conjunto, dando estabilidad a los resultados.

4.4.2.2. Clústering por K-medioides

134. El método de clústering por K-*medioides* es muy similar al de K-*medias*, con la diferencia que cada grupo se caracteriza no a partir del promedio de las variables para los individuos que pertenecen al clúster sino por el individuo más central en el grupo. De esta forma, se consigue reducir la sensibilidad del procedimiento a datos atípicos. El algoritmo más utilizado para el clústering por medioides es el PAM (*Partioning Around Medioids*).

135. Este procedimiento, primero, selecciona aleatoriamente K elementos para que sean los centros de cada conjunto. Segundo, calcula la matriz de distancias de cada elemento no seleccionado a los medioides. Tercero, asigna estos elementos a un grupo minimizando su distancia al medioide correspondiente. Por último, cuarto, para cada clúster intercambia el medioide con algún otro individuo para establecer si de esta forma puede reducir el error de ajuste. Si con el cambio del medioide se consigue mejorar el ajuste, este se modifica y se repite el

procedimiento desde el tercer paso hasta conseguir convergencia. En la matriz de similaridades se puede utilizar la distancia Euclídea o Manhattan.

136. En el método de K-*medioides* se deben definir de antemano la cantidad de segmentos K. Para apoyar esta tarea, se define $a(i)$ como la similaridad promedio del individuo i con respecto a todos los demás individuos de su grupo. Adicionalmente, se calcula la distancia promedio de i con todos los individuos de los demás grupos a los cuales no pertenece i. Con esto, se define $b(i)$ como la menor distancia promedio entre i y los individuos de un clúster C al cual no pertence i. Este grupo se denomina el clúster vecino de i. A partir de estos resultados, se define silueta como

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

137. Conceptualmente, la silueta busca medir qué tan cerca está el individuo i a los demás individuos de su grupo, y comparar esta cercanía con la que el individuo i tiene a otros clústers.

138. El resultado de la silueta se encuentra entre -1 y 1. Se obtienen valores negativos cuando $a(i)$ es mayor que $b(i)$, lo cual es una indicación de que el individuo i podría estar mejor clasificado si se asigna en un grupo diferente, aún más en la medida que $s(i)$ esté más cerca de -1. Así mismo, se obtendrán siluetas positivas en la medida que $b(i) > a(i)$, indicando que el individuo i no podría clasificarse mejor si se asignara a otro grupo, aún más en la medida que $s(i)$ esté más cerca de 1. De esta forma, se puede seleccionar K como el valor que arroja la silueta más alta para el clúster correspondiente.

139. Para las 10 empresas que se han venido analizando, el Gráfico 21 muestra la silueta $s(i)$ para cada uno de los individuos, diferenciando por el segmento al que hayan sido asignados. Aquí, como el grupo 2 está conformado sólo por una empresa, la silueta asociada a esta observación es 0. Para el grupo 1 se ve que los individuos tienen siluetas positivas, indicando que se encuentran bien clasificados.

140. Ahora, el Gráfico 22 muestra la silueta para una segmentación con $K = 3$. Para este caso se observa que el grupo 1 no tiene tan buena asignación de sus individuos, toda vez que los valores de sus siluetas no son tan altos, e incluso mostrando un valor de silueta negativo. También, la silueta promedio con $K = 3$ es inferior a la que se obtiene para $K = 2$, mostrando que los resultados del segundo caso son mejores y ayudando de esta manera a escoger el mejor valor del parámetro K. Este mismo análisis se puede observar en el Gráfico 23.

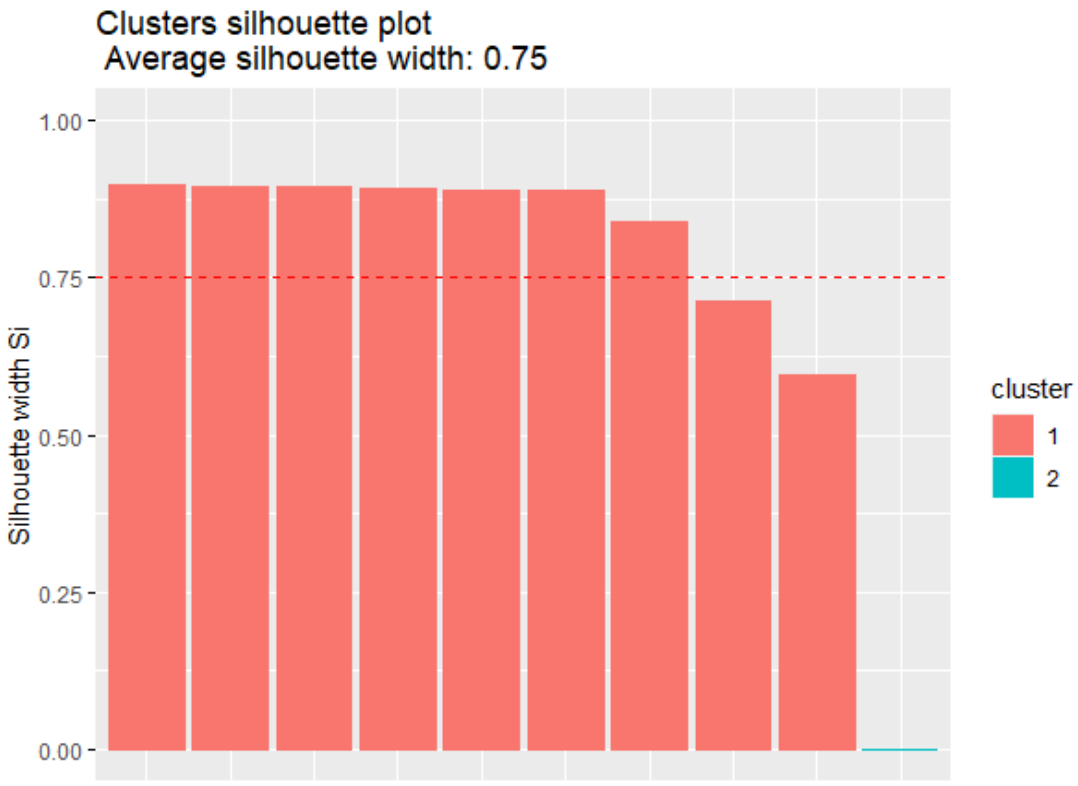

Gráfico 21. Silueta del análisis clúster por K-medioides para 10 empresas, con $K = 2$

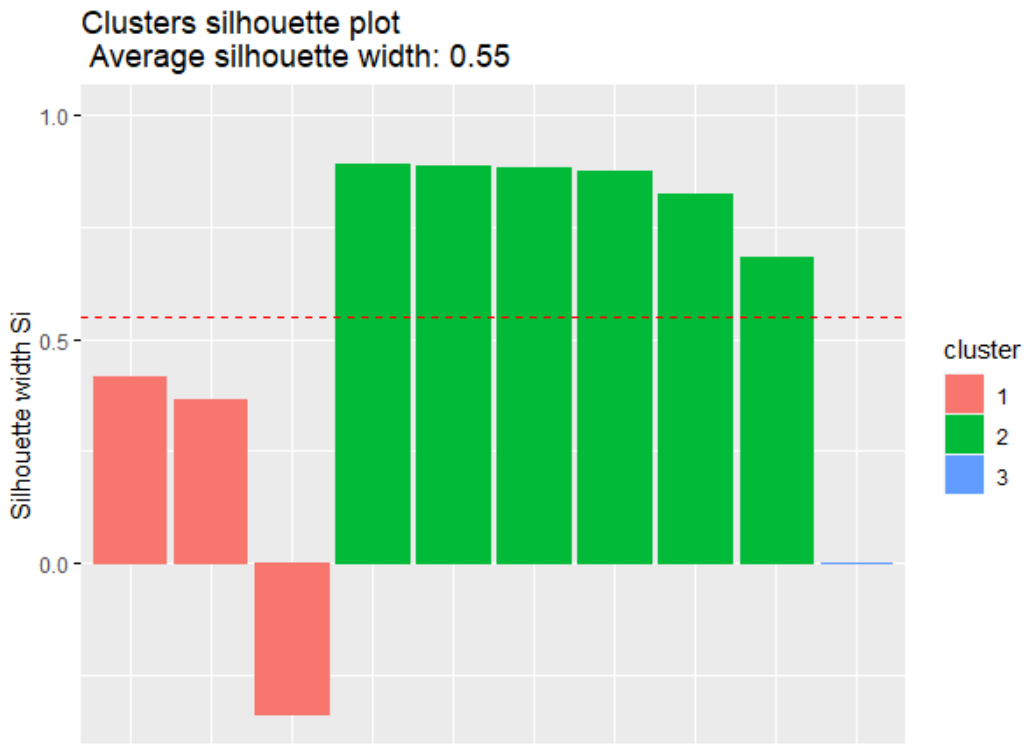

Gráfico 22. Silueta del análisis clúster por K-medioides para 10 empresas, con $K = 3$

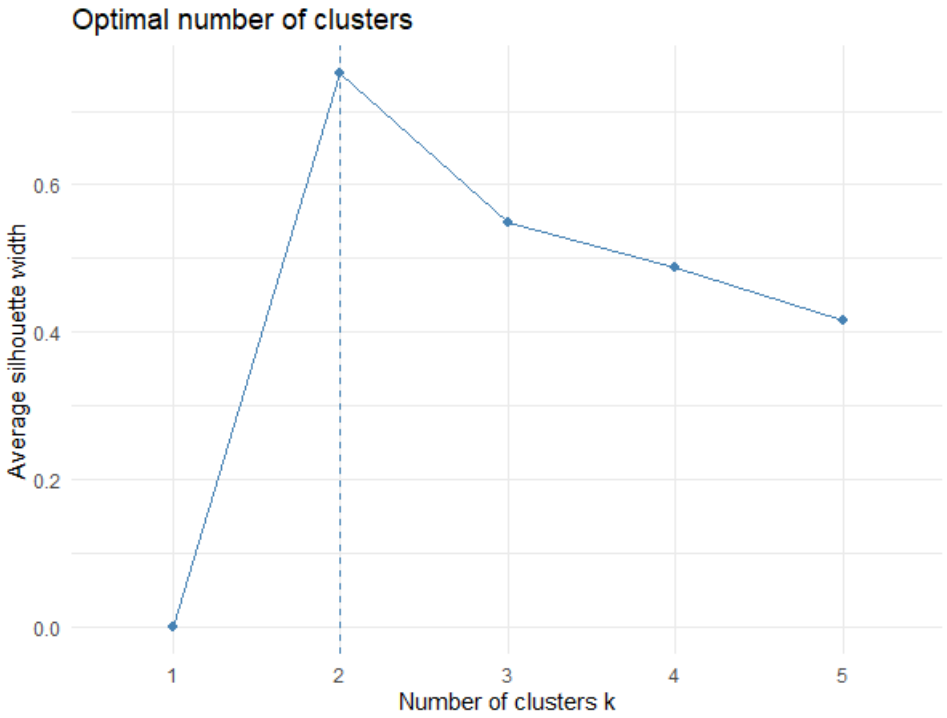

Gráfico 23. Silueta del análisis clúster por K-medioides para 10 empresas, con $K = 1, 2, \ldots, 5$

4.4.3. Clústering jerárquico

141. El problema de agrupación de individuos puede resolverse mediante una aproximación diferente a la de clústering particional. En este caso, o se asume que cada individuo es un clúster en sí mismo y se van agrupando consecutivamente buscando generar grupos homogéneos, o se inicia suponiendo que los individuos pertenecen a un único grupo y se van clasificando consecutivamente en conjuntos lo más heterogéneos posibles. No es necesario conocer de antemano la cantidad de grupos que se quieren generar, pero sí se debe seleccionar una de las tantas particiones que se generan como resultado final.

142. Típicamente, el resultado de un agrupamiento jerárquico se representa visualmente a través de un dendograma, que es un gráfico en forma de árbol que muestra cómo los grupos se van conformando, ya sea por división o por agrupación.

143. A continuación, se profundiza en los algoritmos de clústering jerárquico aglomerativo y divisivo.

4.4.3.1. Clústering jerárquico aglomerativo

144. El clústering aglomerativo, también conocido como AGNES (*Agglomerative Nesting*), es el procedimiento más utilizado en este tipo de metodologías. El algoritmo inicia asumiendo que cada

individuo corresponde a un grupo y calcula la matriz de distancias. Posteriormente agrupa los objetos más similares entre sí, es decir, aquellos que tengan una menor distancia, conformando un nuevo conjunto de grupos. Este proceso continúa, pero ahora agrupando segmentos de individuos según su proximidad mediante una función de conexión, hasta que todos los individuos pertenezcan a un único grupo. Finalmente, se determina en qué punto de las agrupaciones se hace un corte para llegar a los clústers finales.

145. Son varias las funciones existentes para calcular la conexión entre grupos. Entre ellas están:

- Conexión completa: la distancia entre dos grupos es la distancia máxima entre todos los elementos del clúster 1 y el clúster 2. Tiende a generar grupos más compactos.
- Conexión sencilla: la distancia ente dos grupos es la distancia mínima entre todos los elementos del clúster 1 y el clúster 2. Tiende a generar grupos menos compactos.
- Conexión promedio: la distancia ente dos grupos es la distancia promedio entre todos los elementos del clúster 1 y el clúster 2.
- Conexión de centroide: la distancia ente dos grupos es la distancia entre el centroide del clúster 1 y el centroide del clúster 2. En este caso el centroide es el promedio de las variables para todos los elementos en el grupo.
- Método de mínima varianza de Ward: en cada paso une las parejas de grupos que tengan la menor distancia entre clústers, minimizando la varianza total entre clústers.

146. La aplicación del clústering jerárquico aglomerativo a las 10 empresas que se han venido utilizando en esta sección arroja el dendograma que se observa en el

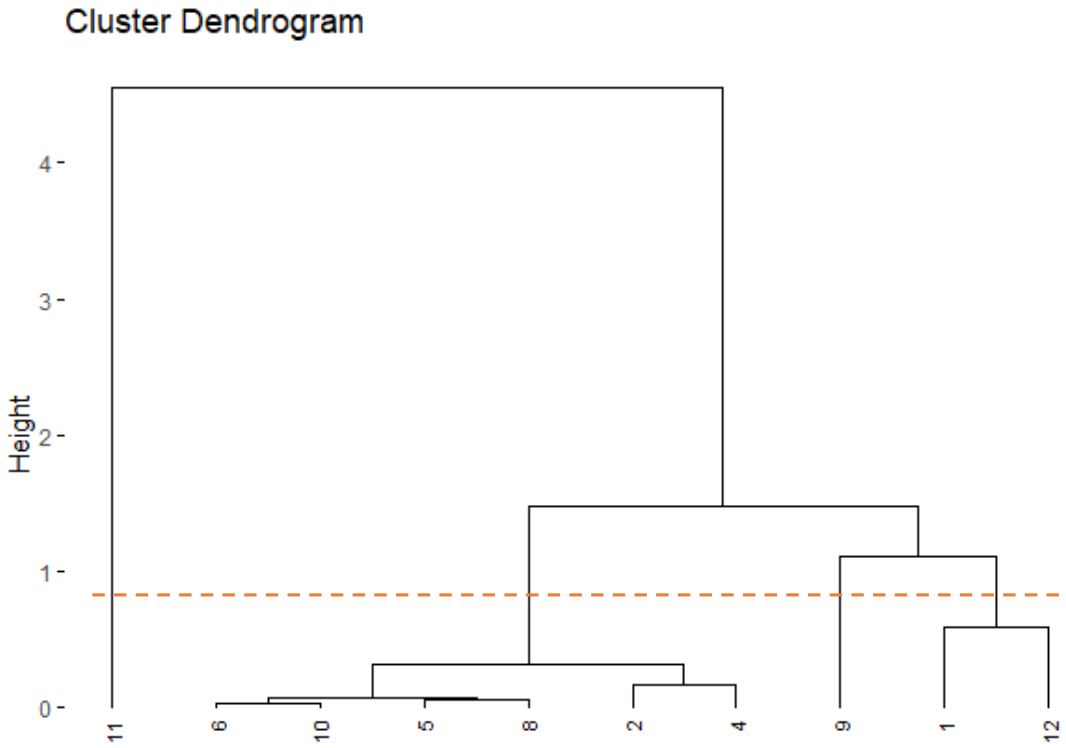

Gráfico 24. Aquí, cada una de las divisiones finales, también llamadas hojas, corresponde a un

individuo. En la medida que se recorre la figura hacia arriba, siguiendo las ramas, se van generando las agrupaciones. Adicionalmente, la altura en el dendograma representa la similaridad a la cual se genera la fusión de cada grupo. De esta manera, se observa que las empresas 5 y 8 son las más parecidas, resultando en la primera agrupación. Posteriormente se agregan las empresas 6 y 10, y luego esta pareja se agrega a la pareja conformada anteriormente. Las empresas 2 y 4 se unen a continuación, y luego se agregan al segmento de las empresas 6, 10, 5 y 8, conformando un clúster de 6 individuos. Al siguiente nivel de similaridad se agrupan los individuos 1 y 12, que luego se unen al individuo 9, y finalmente se agregan junto con el segmento de las empresas 6, 10, 5, 6, 2 y 4. El último paso sería la unión de la empresa 11 para retornar al conjunto original de 10 individuos.

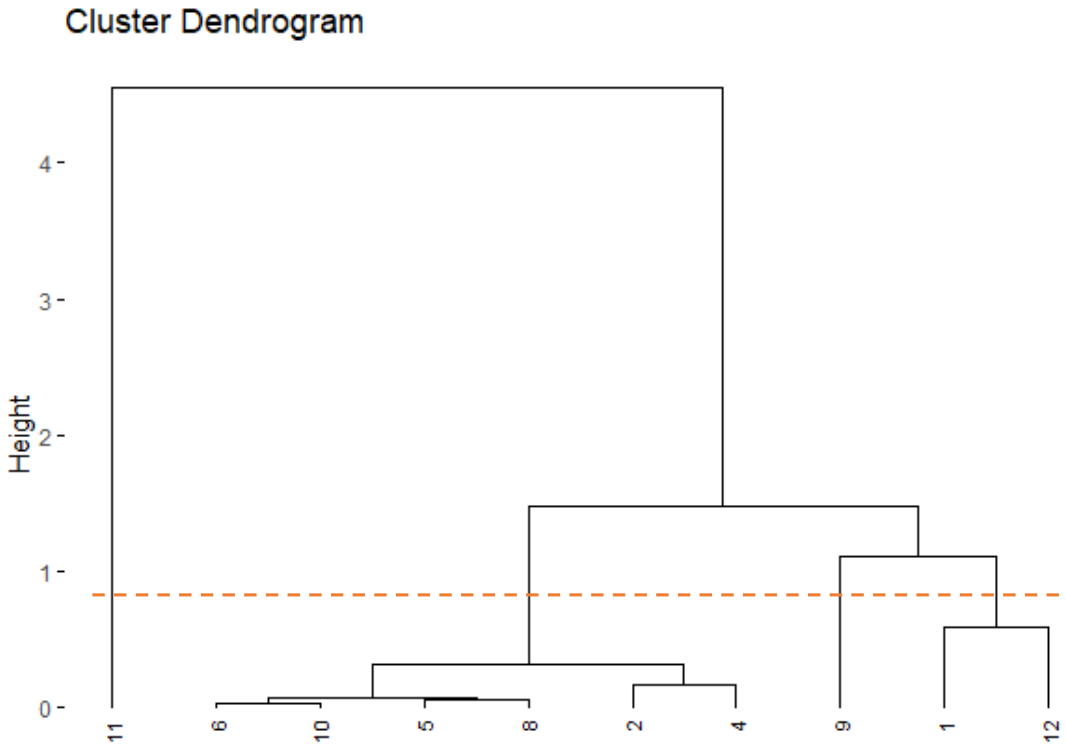

Gráfico 24. Dendograma del clústering jerárquico aglomerativo para 10 empresas

147. Aun cuando el clústering jerárquico genera todo tipo de agrupaciones, desde individuales hasta completas, no dice nada sobre cuál de ellas seleccionar. Una vez se toma la decisión de cuántos grupos se quieren, o a qué altura en el dendograma realizar el corte, se pueden obtener los segmentos finales. Por ejemplo, utilizando $K = 4$, se tienen: grupo 1, conformado por las empresas 1 y 12; grupo 2, conformado por las empresas 2, 4, 5, 6, 8 y 10; grupo 3, conformado por la empresa 9; y grupo 4, conformado por la empresa 11. Estos corresponden a las ramificaciones que hay por debajo de la línea punteada, anaranjada, en el

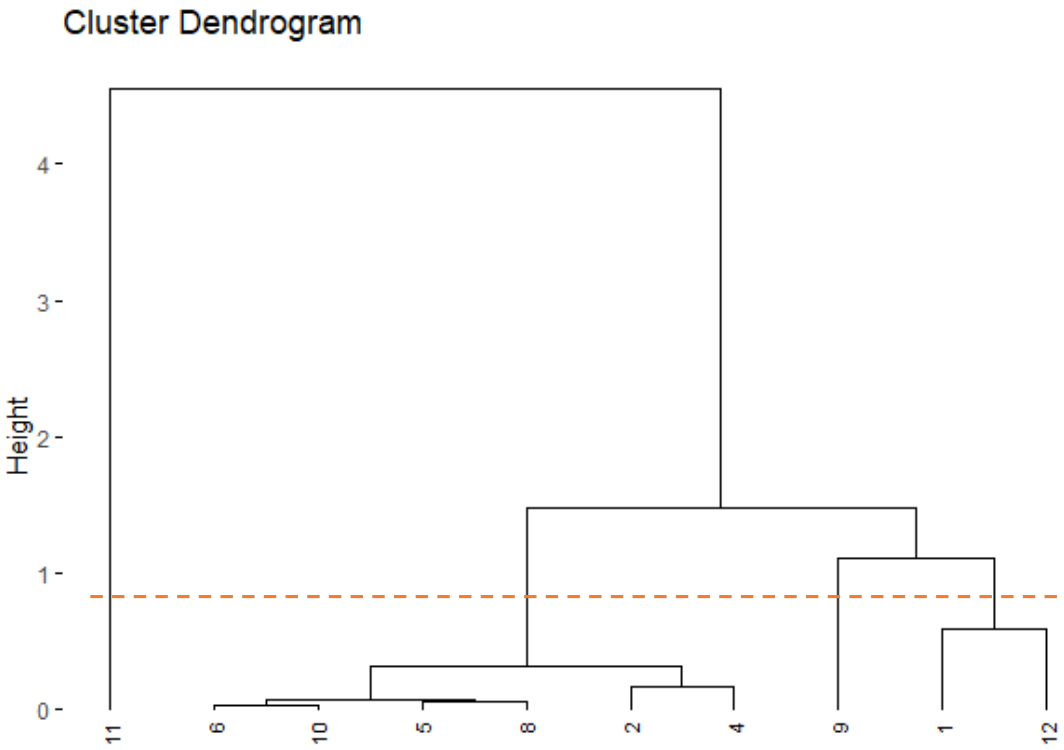

Gráfico 24.

4.4.3.2. Clústering jerárquico divisivo

148. El clústering divisivo, también conocido como DIANA (*Divisive Analysis*), funciona al revés del algoritmo aglomerativo. En este caso se empieza con todos los individuos en un mismo grupo y se van asignando a grupos disyuntos buscando generar los grupos más heterogéneos. En este punto es importante mencionar que el procedimiento aglomerativo es más fuerte identificando clústers pequeños, mientras que el algoritmo divisivo es mejor generando grupos grandes de individuos.

149. Para este caso también se utilizan las funciones de conexión presentadas en la sección anterior y, así mismo, se obtiene como resultado del análisis un dendograma a partir del cual se pueden obtener las agrupaciones deseadas una vez se seleccione el parámetro K. Para las 10 empresas que se han venido trabajando se obtiene el resultado del

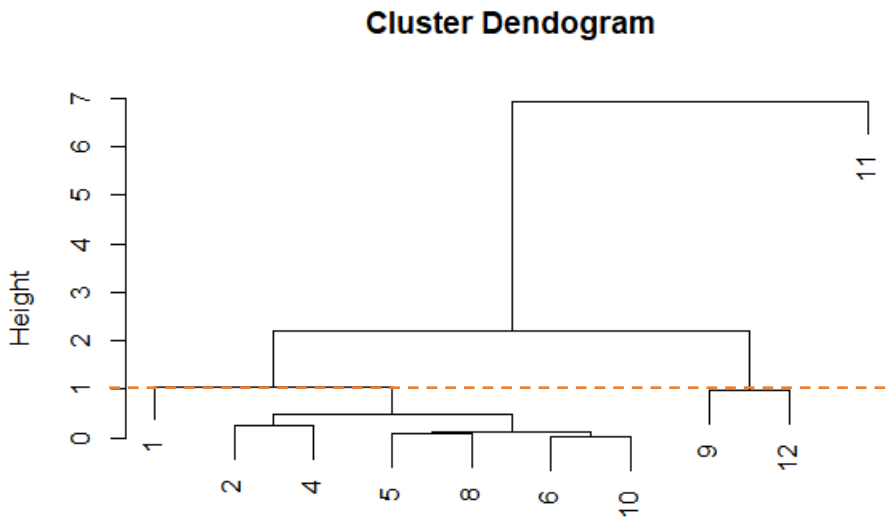

Gráfico 25, que se diferencia del dendograma obtenido por el algoritmo aglomerativo, particularmente, en el tratamiento que le da a los individuos 1 y 9. También, se traza la línea anaranjada, punteada, para identificar segmentos con $K = 4$. El resultado muestra al grupo 1 conformado por el individuo 1; al grupo 2 conformado por los individuos 2, 4, 5, 8, 6 y 10; al grupo 3 conformado por los individuos 9 y 12; y el grupo 4 conformado por el individuo 11.

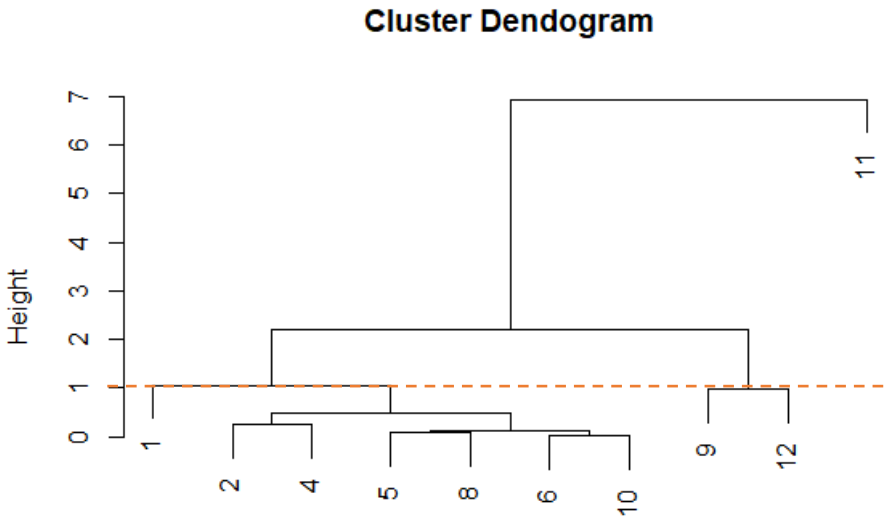

Gráfico 25. Dendograma del clústering jerárquico divisivo para 10 empresas

4.4.4. Análisis clúster en el Sistema ALA/CFT

150. El análisis clúster permite la segmentación de individuos. Para cada uno de los grupos resultantes, los participantes son parecidos entre sí, de acuerdo con las medidas de distancia o similaridad seleccionadas, y los grupos son diferentes entre ellos. De esta manera, el análisis clúster ayuda a identificar grupos de individuos atípicos que pueden ser de interés para el Sistema ALA/CFT.

151. Por ejemplo, en el análisis de 10 empresas a partir de su nivel de activo y pasivo, todos los algoritmos de agrupación (*K-medias*, *K-medioides*, jerárquico aglomerativo y jerárquico divisivo) identifican que la compañía 11 es un grupo en sí mismo, y la ubican lejos de los otros individuos. En otras palabras, el análisis clúster concluye que la empresa 11 es atípica, y el Sistema ALA/CFTY debería priorizarla para el análisis de inteligencia financiera.

152. Por supuesto, el resultado del análisis podría concluir que el alto nivel de activos y pasivos de la compañía tiene una justificación económica clara, como la tradición en el sector o una posición dominante, pero también puede resultar en la válida sospecha de la compañía.

153. Otro aspecto para considerar es el tipo de información que se utiliza en el análisis. Para la identificación de situaciones de interés podrían ser más útiles variables de cambio, como las tasas de crecimiento de un periodo a otro del ingreso, el activo, el pasivo y el patrimonio. Estas series son más difíciles de obtener porque requieren conocer el valor de los niveles de las variables para dos momentos determinados del tiempo, pero también tienen mejores posibilidades porque crecimientos acelerados pueden estar más relacionados con actividades sospechosas.

4.4.5. Análisis de anomalía

154. Anteriormente en esta sección se mostró que los segmentos generados a partir del análisis clúster pueden caracterizarse a partir de su centroide. También, que es posible calcular la distancia de cada individuo al centroide de su grupo utilizando diferentes métricas, como la distancia Euclídea o la distancia Manhattan. Ahora, el análisis de estas distancias, segmento a segmento, complementa la identificación de grupos de individuos extraños, permitiendo ahora identificar a los individuos atípicos dentro de un grupo de sujetos homogéneos.

155. Para aclarar este concepto, la Tabla 8 considera nuevamente la muestra de 10 empresas y la segmentación resultante presentada en la Tabla 7. Con esta información se calcula la distancia de cada individuo al centroide de su grupo (columna 5 de la tabla), notando que el resultado para la compañía 11 es 0 porque ella es la única que conforma su segmento. Para las empresas del grupo 1 se observan diferentes valores, pero llama la atención la distancia calculada para la compañía 9, que es mucho mayor que la distancia de los demás individuos de su segmento, incluso más que el doble de la distancia promedio de estas empresas al centroide de su clúster.

156. De esta forma, se nota que la empresa 9 es atípica entre sus pares y, por lo tanto, candidata para el análisis de inteligencia financiera. Ahora, ¿cuánto es un valor suficientemente grande para considerar que un individuo es extraño en su grupo? Aunque esta pregunta no tiene una respuesta directa, se puede tomar como referente el percentil 90, 95 o 99 de las distancias calculadas entre los individuos y los centroides de sus grupos. Para el ejemplo, el percentil 90 de las distancias entre los 9 individuos del grupo 1 y su centroide es $0{,}8344$, confirmando el interés por la empresa 9.

Empresa	Activo reescalado	Pasivo reescalado	Iteración 4	Distancia centroide grupo respectivo
			Asignación clúster	
1	-0,364362	0,017669	1	0,324132
2	-0,541611	-0,425727	1	0,284801
4	-0,390654	-0,350848	1	0,116306
5	-0,564794	-0,556748	1	0,379838
6	-0,546214	-0,629334	1	0,421026
8	-0,514632	-0,565706	1	0,351710
9	0,744112	0,098798	1	1,104967
10	-0,527480	-0,616277	1	0,399270
11	2,587105	2,674868	2	0,000000
12	0,118529	0,353304	1	0,766805
Promedio activo, clúster 1			-0,287456	
Promedio pasivo, clúster 1			-0,297208	
Promedio activo, clúster 2			2,587105	
Promedio pasivo, clúster 2			2,674868	

Tabla 8. Análisis de anomalía para 10 empresas

4.5. METODOLOGÍAS DE APRENDIZAJE DE MÁQUINA SUPERVISADO

157. El aprendizaje de máquina supervisado enmarca un conjunto amplio de algoritmos que se basan en información suministrada externamente para producir una hipótesis general, que posteriormente permite hacer predicciones sobre la información futura (Kotsiantis, 2007). Un ejemplo de su aplicación en el contexto del Sistema ALA/CFT es un procedimiento que toma como insumos las características socioeconómicas y financieras de una población de individuos, aplica una ecuación matemática a esta información y calcula para cada una de las personas la probabilidad de que estén incluidas en un ROS.

158. Es fácil notar las ventajas de una herramienta como la mencionada anteriormente: entre otras, permite generar de manera rápida un criterio objetivo de priorización de individuos a partir de su probabilidad de ser incluidos en un ROS, tiene la capacidad de procesar grandes conjuntos de datos y se puede automatizar. Sin embargo, pese a su conveniencia, no es necesariamente fácil o directo definir el conjunto de información $\boldsymbol{X}$ que se va a utilizar como insumo, o encontrar la mejor[11] función f que permita predecir la variable de interés y. A este respecto, el aprendizaje de máquina supervisado resuelve precisamente este problema, encontrando f a partir del análisis sistemático de las relaciones presentes entre $\boldsymbol{X}$ y y.

[11] El concepto de "mejor" se define a partir de diferentes medidas que buscan entender si una función es superior que otra para predecir una variable de interés.

159. De manera general, se supone que se observa una respuesta cuantitativa y y p variables diferentes $X_1, X_2, \ldots, X_p$ que se utilizarán como predictores. Se asume que existe una relación entre y y $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$, que se puede describir de manera general como $y = f(\boldsymbol{X}) + \varepsilon$. Aquí f es una función fija pero desconocida de $X_1, X_2, \ldots, X_p$, y ε es un término de error aleatorio, que no está relacionado con $\boldsymbol{X}$ y tiene media cero.

160. Los modelos del aprendizaje supervisado son variados y provienen de diferentes campos como la Estadística, las Matemáticas o la Ingeniería. Una situación que suele enfrentarse con estas metodologías es que aquellas que mejor predicen son más difíciles de interpretar, precisamente por su complejidad, generando un *tradeoff* bien conocido entre interpretabilidad y predictibilidad.

161. En esta sección se presentarán los conceptos de clasificación y regresión. Se tratarán los modelos de regresión y de árboles, junto con algunas de sus variaciones. Por completitud, en la última parte se mencionarán otros algoritmos importantes, aunque no se hará una exposición detallada al respecto.

4.5.1. Clasificación vs. regresión

162. En el aprendizaje de máquina supervisado puede trabajar dos tipos de problemas, según la naturaleza de la variable de interés y. Si y es una variable categórica se tiene un problema de clasificación, y si y es continua se tiene un problema de regresión. A partir de la información de una persona, un problema de clasificación es predecir si un individuo es sospechoso o no de estar involucrado en actividades relacionadas con FT, mientras que un problema de regresión es predecir la cantidad de recursos que la persona ha movilizado relacionados con FT.

163. Muchos de los algoritmos del aprendizaje de máquina supervisado son compatibles con problemas de clasificación y de regresión, como en el caso de los modelos de regresión, los modelos de árboles y las redes neuronales. Así mismo, existen modelos especializados que pueden generar mejores resultados de clasificación o predicción de variables continuas.

4.5.2. Análisis supervisado para clasificación: regresión logística y árboles de clasificación

164. Dos modelos ampliamente utilizados para la construcción de clasificadores son la regresión logística y los árboles de clasificación. El primero tiene la ventaja de ser un modelo relativamente sencillo e interpretable, que en la práctica ha mostrado un buen desempeño predictivo, incluso si se le compara con otras alternativas más sofisticadas.

165. El segundo es un modelo versátil, fácilmente interpretable, que permite el manejo de información faltante, aunque resulta sensible a los datos disponibles. Ahora, los modelos de árboles muestran su verdadero potencial cuando conforman ensambles de modelos, es decir, sistemas que utilizan muchos submodelos para mejorar su capacidad predictiva, aunque al costo

de incrementar su complejidad y reducir su interpretabilidad. Las secciones siguientes elaboran sobre estos dos tipos de metodologías.

4.5.2.1. Regresión logística

166. Para una variable de interés y categórica que indica la pertenencia a uno de dos grupos posibles, por ejemplo, sospechoso y no sospechoso, la regresión logística modela la probabilidad que y tome alguno de estos valores. Considerando los datos de 10 empresas presentados en la Tabla 9, donde se tiene un identificador de cada individuo, los activos y pasivos en miles de USD, y una variable indicadora que muestra si la empresa es sospechosa de estar involucrada en actividades relacionadas con LA, la idea es encontrar la mejor función f para calcular la probabilidad condicional $\Pr(Sospechosa = Sí|Activo, Pasivo)$, abreviada como $p(Sospechosa)$. De esta manera, para una nueva empresa de la cual se conozca el valor de sus activos y pasivos, será posible utilizar f para calcular la probabilidad p, que es un valor que se encuentra entre 0 y 1.

167. Adicionalmente, si este resultado es mayor que un umbral u se predice que la empresa es sospechosa y candidata para un análisis adicional de inteligencia financiera. Cuando el umbral u es pequeño, por ejemplo $u = 0{,}1$, será más fácil clasificar la empresa como sospechosa, lo cual permitirá que las verdaderamente sospechosas sean analizadas, aunque aumentando el costo del análisis porque tendrán que procesarse en detalle muchos casos. Si el umbra u toma un valor grande, como $u = 0{,}9$, será más difícil que la empresa sea clasificada como sospechosa, aumentando el riesgo que una empresa verdaderamente sospechosa no sea analizada en detalle, aunque reduciendo los costos del proceso por reducir la cantidad de individuos a procesar en detalle.

Empresa	Activo (miles USD)	Pasivo (miles USD)	Sospechosa
1	1.869	1.529	Sí
2	977	714	No
4	1.737	852	No
5	860	473	No
6	954	340	No
8	1.113	457	No
9	7.448	1.678	Sí
10	1.048	364	No
11	16.723	6.414	Sí
12	4.299	2.146	No

Tabla 9. Valor de los activos y pasivos e indicador de inclusión en ROS para 10 empresas

168. Una alternativa para modelar f y calcular p es a través de un modelo lineal de probabilidad: $p(X) = \beta_0 + \beta_1 X$. Sin embargo, con esta representación se corre el riesgo de obtener probabilidades menores que cero o mayores a uno, que están por fuera del intervalo

$[0,1]$ de valores esperados. Para evitar este problema $p(X)$ se modela a través de la función logística, que tiene la ventaja de producir únicamente valores entre 0 y 1. De esta manera

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

169. La función logística anteriormente presentada se define cuando se establecen los valores de $\boldsymbol{\beta} = (\beta_0, \beta_1)$. Para esto se toman los datos de la variable X y la variable categórica de interés, expresada como una serie de unos y ceros, unos cuando la observación pertenece al grupo de interés y ceros en caso contrario[12], y se aplica un proceso de estimación llamado máxima verosimilitud. Como se observa en el Gráfico 26, la función logística tiene forma de S y siempre produce valores que están entre 0 y 1.

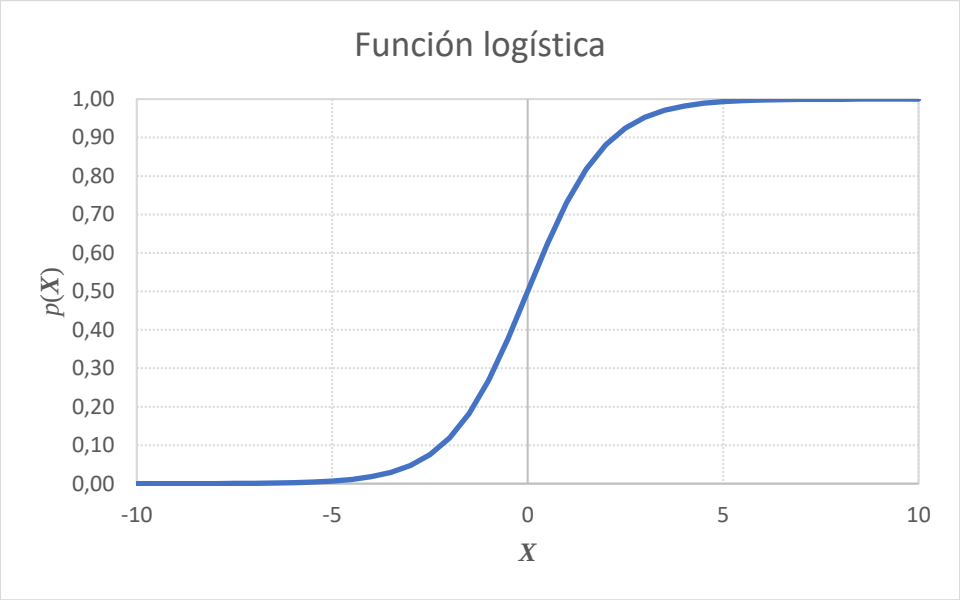

Gráfico 26. Comportamiento de la función logística

170. La ecuación de la regresión logística se puede reexpresar como se observa a continuación:

$$\begin{aligned} p(X) &= \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \\ p(X)\left(1 + e^{\beta_0+\beta_1 X}\right) &= e^{\beta_0+\beta_1 X} \\ p(X) + p(X)e^{\beta_0+\beta_1 X} &= e^{\beta_0+\beta_1 X} \\ p(X) &= e^{\beta_0+\beta_1 X} - p(X)e^{\beta_0+\beta_1 X} \\ p(X) &= e^{\beta_0+\beta_1 X}\left(1 - p(X)\right) \end{aligned}$$

[12] Una serie con estas características se denomina variable dummy o dicótoma.

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

171. La cantidad $p(X)/\big(1-p(X)\big)$ se conoce como *odds-ratio* y puede tomar cualquier valor entre 0 e ∞. Un *odds-ratio* cercano a 0 o 1 indica una baja o alta probabilidad de ocurrencia del evento de interés, respectivamente. Por ejemplo, si 9 de 10 empresas son sospechosas de LA, $p = 0{,}9$ con un *odds-ratio* de $\frac{0{,}9}{1-0{,}9} = \frac{0{,}9}{0{,}1} = 9$. Al calcular el logaritmo natural del *odds-ratio* se tiene la función *logit*, que es lineal en X:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

172. En este caso, cambios de una unidad en X implican cambios de β_1 unidades en la función *logit*. Ahora, la relación entre X y $p(X)$ no es lineal, entonces β_1 no corresponde a la interpretación deseable de cambio en $p(X)$ por cambios de una unidad en X. Adicionalmente, el cambio en $p(X)$ por cambios en X depende del valor que tenga X. De todas formas, se puede afirmar que si β_1 es positivo, aumentos en la variable X tendrán un efecto positivo sobre la probabilidad. Por el contrario, si β_1 es negativo, incrementos en X generarán una reducción en $p(X)$.

173. El ajuste del modelo de regresión logística a los datos de la Tabla 9 arroja los valores $\hat{\beta}_0 = -3{,}8589$, $\hat{\beta}_1 = 0{,}0004$ y $\hat{\beta}_1 = 0{,}0012$, es decir, la probabilidad de que una empresa sea sospechosa se obtiene de

$$\hat{p}(\boldsymbol{X}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 Activo+\hat{\beta}_2 Pasivo}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 Activo+\hat{\beta}_2 Pasivo}} = \frac{e^{-3{,}8589+0{,}0004 Activo+0{,}0012 Pasivo}}{1+e^{-3{,}8589+0{,}0004 Activo+0{,}0012 Pasivo}}$$

174. De acuerdo con los resultados para este ejemplo, un mayor activo o pasivo de una compañía implica una mayor probabilidad de que esté relacionada con LA. De la misma manera, si el activo y el pasivo de una empresa fueran 5.000 y 2.000 miles USD, respectivamente, la probabilidad calculada sería

$$\hat{p}(\boldsymbol{X}) = \frac{e^{-3{,}8589+0{,}0004\times 5.000+0{,}0012\times 2.000}}{1+e^{-3{,}8589+0{,}0004\times 5.000+0{,}0012\times 2.000}} = 0{,}6499$$

175. Finalmente, si se define un umbral $u = 0{,}5$, la empresa sería considerada por el clasificador como sospechosa en la medida que $\hat{p}(\boldsymbol{X}) > u$.

176. De acuerdo con el ejemplo, estos resultados se muestran en el Gráfico 27. Aquí, los círculos azules corresponden a las empresas para las que no se tiene sospecha alguna. Los cuadros rojos son las empresas para las cuales sí se han observado actividades sospechosas de LA. El círculo naranja representa la predicción realizada.

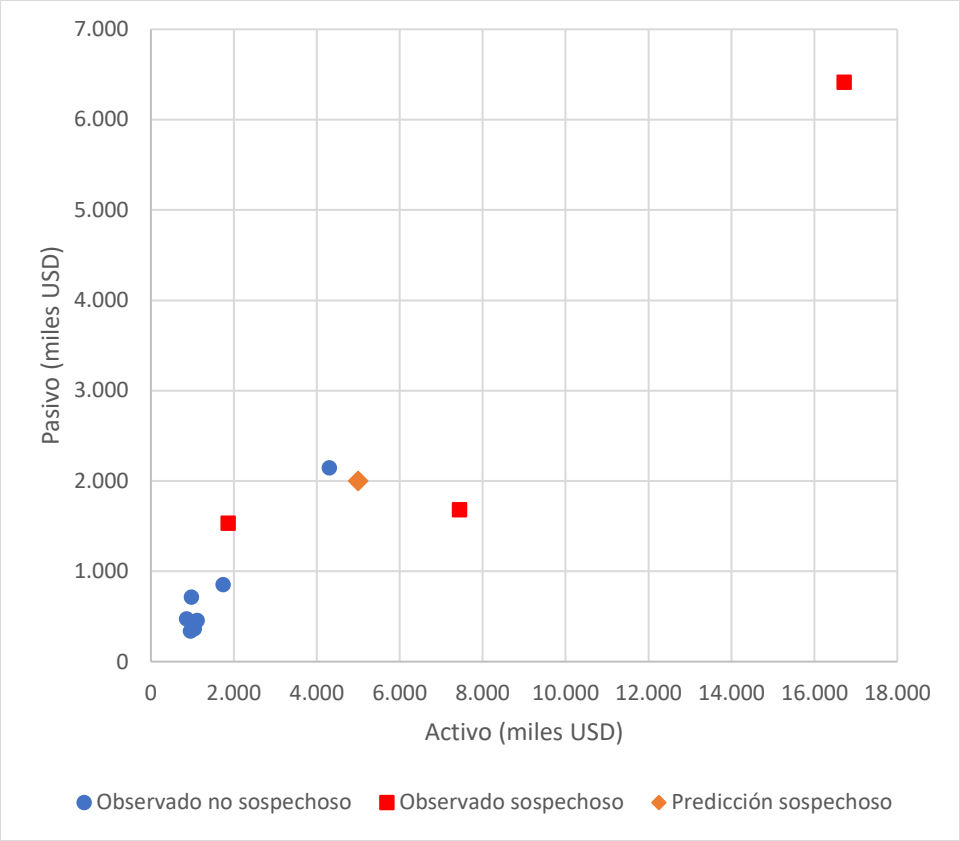

Gráfico 27. Predicción de sospecha a partir del valor de los activos y pasivos de 10 empresas

4.5.2.2. Árboles de clasificación

177. En la clasificación mediante árboles se busca predecir una variable categórica de interés generando subgrupos de individuos a partir de los valores que toman las características que los definen. Su nombre viene de la representación gráfica de los resultados, que se puede asociar fácilmente a un árbol invertido.

178. A manera de ejemplo se utiliza la información de 500 empresas para predecir inclusión en un ROS, utilizando como insumos el valor de los activos y pasivos, en USD. El Gráfico 28 muestra el árbol de clasificación ajustado a estos datos. La división superior asigna las compañías que tienen un activo menor a 10.000 USD a la rama izquierda. Por tratarse de un modelo de clasificación, la predicción de inclusión en un ROS está dada por la categoría dominante en el grupo, es decir, si la mayoría de los individuos han estado relacionadas en un ROS, entonces este será el valor que se asigne a una empresa cuyos activos sean menores a 10.000 USD. Para este caso, el gráfico indica que en esta partición no hay compañías que hayan estado incluidas en un ROS, con lo cual se predice la no inclusión en ROS para este nodo.

179. Las empresas con pasivo mayor o igual a 10.000 USD se asignan a la rama derecha, y este grupo se subdivide nuevamente por el pasivo. En general, el árbol segmenta a los individuos en

tres regiones del espacio de predicción: empresas con activo menor a 10.000 USD, empresas con activo mayor o igual a 10.000 USD y pasivo menor a 10.000 USD, y empresas con activo mayor o igual a 10.000 USD y pasivo mayor o igual a 10.000 USD. Estas cuatro regiones se pueden escribir como $R_1 = \{X|activo < 10.000\}$, $R_2 = \{X|activo \geq 10.000, pasivo < 10.000\}$ y $R_3 = \{X|activo \geq 10.000, pasivo \geq 10.000\}$. A R_1 pertenecen el 94% de los individuos, a R_2 el 4% y a R_3 el 2%. La predicción de R_1 es no inclusión en ROS, para R_2 es no inclusión en ROS y para R_3 la inclusión en ROS.

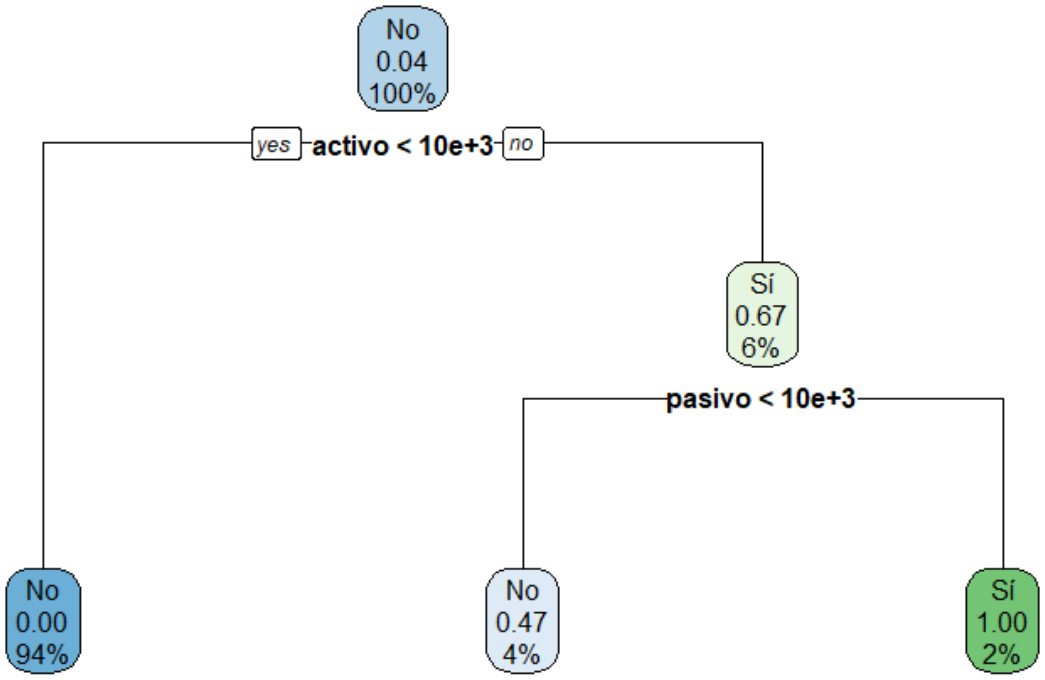

Gráfico 28. Árbol de clasificación de 500 empresas

180. En el Gráfico 29 se representan las regiones definidas a partir del activo y el pasivo. Aquí, cuando se agrupan los individuos que tienen un activo menor a 10.000 USD se genera la región que se observa como una barra vertical a la izquierda (R_1). La región que se genera por un activo mayor o igual a 10.000 USD se subdivide a su vez en dos regiones, una para los pasivos por debajo de 10.000 USD (R_2) y otra para los activos mayores o igual a 10.000 USD (R_3). El color de la región es rojo, si la mayoría de los individuos no están incluidos en un ROS, y azul para el caso contrario.

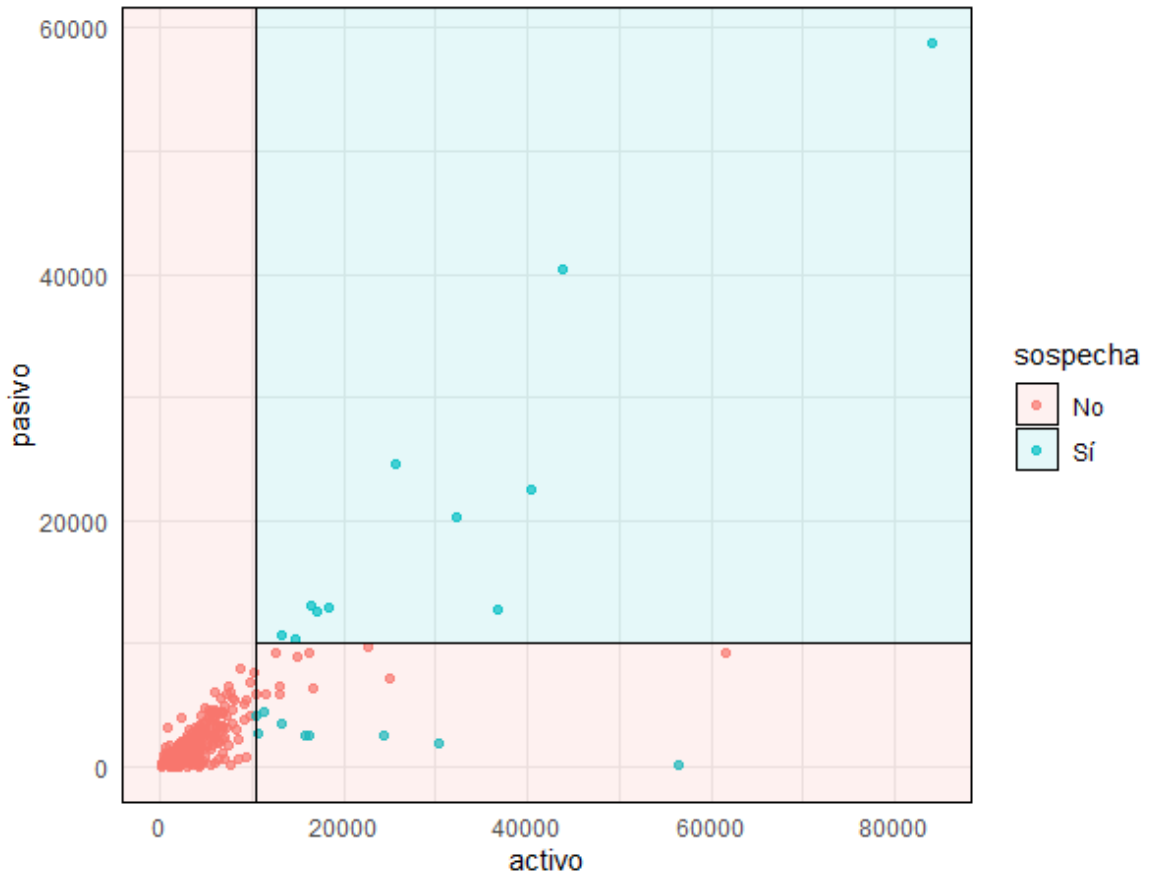

Gráfico 29. Regiones obtenidas del árbol de clasificación de 500 empresas

181. Como se mencionó al inicio de esta sección, el Gráfico 28 tiene la forma de árbol invertido. Al nodo superior, que es el inicio y representa a todos los individuos, se le denomina raíz. Las regiones R_1, R_2 y R_3 son los nodos terminales, también llamados hojas. También puede haber nodos intermedios, como el que se observa al tomar la rama derecha y considerar las empresas cuyo activo es mayor o igual a 10.000 USD. De manera más precisa, las conexiones ente los nodos se conocen como ramas.

4.5.3. Análisis supervisado para regresión: regresión y árboles de regresión

182. Uno de los modelos más populares para el análisis de regresión es, precisamente, la regresión lineal múltiple. Este modelo tiene la ventaja de ser conocido, interpretable, fácil de utilizar y estar disponible en la mayoría de los paquetes de análisis de datos, incluso en Excel. Su principal desventaja para el aprendizaje máquina radica precisamente en su sencillez, pues suele ser superado, en términos predictivos, por la mayoría de los modelos alternativos. El segundo algoritmo que se presentará es el de árboles de regresión, que es equivalente a los árboles de clasificación presentados anteriormente. Este modelo de alta interpretabilidad puede no ser el más efectivo a la hora de predecir, pero es la base para los modelos de ensamblaje, que sacrifican interpretación para conseguir altas capacidades a la hora de obtener el valor de una variable de interés en una situación que aún no ha sido observada.

4.5.3.1. Regresión lineal múltiple

183. En el modelo de regresión lineal múltiple se tiene un vector $\boldsymbol{\theta}$ de parámetros que se quieren estimar. Un estimador de $\boldsymbol{\theta}$ es $\widehat{\boldsymbol{\theta}}$. Cada uno de los estimadores en $\widehat{\boldsymbol{\theta}}$ es una variable aleatoria que, idealmente, está centrada alrededor de $\boldsymbol{\theta}$, con una varianza pequeña y una distribución de probabilidad conocida. Lo primero es señal de que no existe sesgo en la estimación, lo segundo se refiere a la precisión de la estimación y la tercera condición permite la inferencia estadística, es decir, posibilita hacer pruebas de hipótesis sobre el valor verdadero de $\boldsymbol{\theta}$[13].

184. En la regresión interesa encontrar una ecuación para definir el valor esperado de una variable de interés y, condicional al valor que tomen un conjunto de variables explicativas $\boldsymbol{x}$. De esta manera se tiene

$$E(y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

185. donde $\boldsymbol{x}$ es un vector columna de tamaño k cuya j-ésima posición es x_j, y $\boldsymbol{\beta}$ es un vector columna de tamaño k cuya j-ésima posición es β_j. Adicionalmente, β_0 es el intercepto de la regresión, que corresponde al valor esperado de y cuando todas las demás variables x_j toman el valor de 0.

186. Cuando se quiere utilizar el modelo para predecir, el interés recae principalmente en $E(y|\boldsymbol{x})$. También puede interesar la estimación de los efectos marginales asociados al regresor x_j, esto es

$$\frac{\partial E(y|\boldsymbol{x})}{\partial x_j} = \beta_j$$

187. Aquí es donde reside la interpretabilidad del modelo de regresión lineal, en el valor estimado para cada uno de sus parámetros corresponde con el cambio que se observa en el valor esperado de y por cambios en el valor de la variable x_j. A manera de ejemplo, si se tiene un modelo de regresión lineal para predecir el patrimonio de las empresas en función de sus ingresos y de sus movimientos en efectivo, es posible establecer que el patrimonio actual cambia en β_1 por una unidad más de ingresos y en β_2 por incrementos de una unidad en el valor de las transacciones en efectivo.

188. El modelo de regresión lineal incluye un error aditivo tal que, para la observación i-ésima, resulta

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, i = 1, \ldots, N$$

[13] Bajo supuestos adicionales, $\widehat{\boldsymbol{\theta}}$ tendrá distribución (asintótica) normal multivariada, con vector de medias $\boldsymbol{\theta}$ y matriz de varianzas-covarianzas $Var(\widehat{\boldsymbol{\theta}})$.

189. con N la cantidad de observaciones. Para obtener el estimador de $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$, se minimiza la suma de los cuadrados del error $\sum_{i=1}^{N} u_i^2 = \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$, obteniendo el estimador de mínimos cuadrados ordinarios (MCO). Matricialmente, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

190. Cuando se tiene una variable y que es muy sesgada a la derecha, es decir, que ocurren cada vez con menor frecuencia valores cada vez más grandes, como sucede con el patrimonio, el ingreso, el valor de las transacciones en efectivo, etc., es posible que el modelo lineal no ofrezca un buen ajuste a los datos. En estos casos se puede plantear un modelo exponencial sobre la media con error multiplicativo: $y_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\,\varepsilon_i$. Definiendo $\varepsilon_i = \exp(u_i)$ se tiene $y_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + u_i)$. Finalmente, aplicando logaritmo natural se obtiene el modelo log-lineal $\ln y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$. En este caso $E(\ln y\,|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$ bajo el supuesto que u_i es independiente con media condicional cero. Para este caso los efectos marginales miden la proporción de cambio en $E(y|\boldsymbol{x})$ por cambios en x_j , lo cual se denomina semielasticidad. Para la predicción se utiliza $E(y_i|\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\,E(\exp(u\_i))$.

191. De manera similar al caso de la regresión logística, considérense 10 empresas caracterizadas por sus activos y transacciones en efectivo, y sobre ellas se quiere predecir su patrimonio. La información se encuentra en la Tabla 10. Anteriormente se mencionó que el estimador matricial de $\widehat{\boldsymbol{\beta}}$ es $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Entonces, para este caso, $\boldsymbol{X}$ es la matriz que se conforma por las columnas Activo y Transacciones en efectivo, junto con una columna de unos que se añade a la izquierda. La dimensión de esta matriz es de 10×3 (10 empresas, 2 variables y el vector de unos). El vector $\boldsymbol{y}$ es la columna Patrimonio.

Empresa	Activo (miles USD)	Transacciones en efectivo (miles USD)	Patrimonio (miles USD)
1	977	714	2.208
2	1.737	852	2.327
3	860	473	2.245
4	954	340	2.233
5	1.113	457	2.259
6	1.048	364	2.229
7	4.299	2.146	2.957
8	1.869	1.529	2.465
9	7.448	1.678	3.617
10	16.723	6.414	5.648

Tabla 10. Valor de los activos, transacciones en efectivo y patrimonio para 10 empresas

192. El resultado de la operación arroja los siguientes valores: $\beta_0 = 2.012$, $\beta_1 = 0{,}2096$ y $\beta_2 = 0{,}0207$, indicando que el valor del patrimonio es 2.012 miles de USD ante la ausencia de ingresos y transacciones en efectivo. También, se espera que aumentos de 1.000 USD en el activo generen aumentos de $1.000 \times 0{,}2096 = 209{,}6$ USD en el patrimonio. Por su parte, incrementos de 1.000

USD en el valor de las transacciones en efectivo aumentan el valor del esperado del patrimonio en $1.000 \times 0{,}0207 = 20{,}7$ USD.

4.5.3.2. Árboles de regresión

193. Para generar un árbol de regresión se sigue el mismo procedimiento que para un árbol de clasificación, donde se divide el espacio que generan las variables utilizadas como predictores en J regiones diferentes $R_1, R_2, \ldots, R_j$, que no se sobreponen. La diferencia con el algoritmo de regresión es que para cada observación que caiga en la región R_j se hace la predicción a partir de la media observada para todos los registros que han sido asignados a R_j. Así, supóngase que de los datos se obtienen dos regiones, R_1 y R_2, y que la variable de interés y para la primera región tiene media 10, mientras que para la segunda región muestra una media de 20. Entonces, para una observación nueva $X = x$, si $x \in R_1$ entonces la predicción para ese caso será de 10, y de 20 siempre que $x \in R_2$.

194. La generación de las regiones R_j se hace al dividir el espacio de predicción en rectángulos, o cajas, de alta dimensionalidad. El objetivo es encontrar las cajas $R_1, R_2, \ldots, R_j$ que minimicen $RSS = \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j}\right)^2$, donde $\hat{y}_{R_j}$ es la media de la variable de interés en la j-ésima caja. Aquí, $y_i - \hat{y}_{R_j}$ representa el error que se comete con la predicción para el individuo i que se está asignando a la caja R_j. Este error se calcula para todo el resto de los individuos que pertenecen a esta región, y esto mismo se hace para todas las regiones definidas. Es pertinente mencionar que el error se eleva al cuadrado para que las cantidades que puedan resultar negativos queden en términos positivos, asemejando distancias. También, notar que el problema de minimización busca definir las regiones R_j, por lo que cualquier cambio en la asignación de los individuos modifica los valores de $\hat{y}_{R_j}$. Esto último es parte de las complejidades que se deben tener en cuenta para el ajuste del modelo.

195. Debido a que computacionalmente no es factible considerar todas las particiones del espacio en J cajas, se plantea una aproximación de arriba abajo, denominada codiciosa, conocida como división binaria recursiva. Aquí se inicia desde la parte de arriba del árbol, donde todas las observaciones pertenecen a una misma región, y sucesivamente se divide el espacio de predicción en dos ramas. Este proceso se continua hasta alcanzar un número mínimo de individuo en cada nodo final, o hasta satisfacer algún criterio adicional.

196. En la medida que los árboles se extienden, se puede llegar a modelos sobreajustados, donde se obtienen buenos resultados de predicción para el conjunto de datos con los cuales se entrenó el algoritmo, pero potencialmente deficientes para la predicción de nuevos casos. Para evitar este tipo de circunstancias se utiliza llamado poda, que busca reducir el árbol a otro más compacto que tenga menor varianza (al costo de introducir algo de sesgo). La poda se puede realizar mediante el algoritmo de poda por costo de complejidad, eliminando recursivamente algunas de las ramas finales hasta alcanzar el modelo deseado.

4.5.4. Ensamblaje de modelos

197. Los modelos de árboles son convenientes por su facilidad de interpretación y posibilidades de representación gráfica. Adicionalmente, son capaces de manejar predictores categóricos sin recurrir a transformaciones. Sin embargo, estos modelos no son tan buenos prediciendo y son sensibles a los datos, con lo cual, cambios pequeños en los insumos pueden llevar a modificaciones grandes en los resultados (alta varianza del modelo).

198. Como respuesta a estas inquietudes se han diseñado métodos basados en técnicas de muestreo y remuestreo, particularmente bajo los nombres de *bagging*, bosque aleatorio y *boosting*:

- *Bagging*: este procedimiento, también llamado *bootstrapp aggregation*, permite reducir la varianza de un método de aprendizaje bajo el principio que, ante un aumento en la cantidad de observaciones, el promedio de éstas reduce su variabilidad. En este caso, se generan B muestras repetidas, con reemplazamiento, del conjunto de entrenamiento (*bootstrapping*) y a partir de cada una de ellas se entrena un modelo que ajuste $\hat{f}^b(x)$. Las predicciones así obtenidas se promedian para obtener

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x)$$

  Para el *bagging* con árboles de regresión se hace crecer cada modelo individual para que tengan sesgo bajo y alta variación y, luego, mediante el promedio, se consigue tener una predicción con baja varianza que controla el sesgo. En el caso de árboles de clasificación se sigue el mismo principio, solo que en este caso se hace una votación (por mayoría). El procedimiento de *bagging* en árboles mejora la predicción, pero compromete la interpretabilidad, porque ahora no se cuenta con un solo modelo sino con un conjunto de ellos que pueden ser muy diferentes entre sí. En este caso se pueden construir medidas de importancia de las variables para saber cuáles fueron las que, consistentemente, permitieron mejoras en la medida de ajuste.

4.6. MINERÍA DE TEXTOS

199. Las metodologías estándar de minería de datos esperan conjuntos de información estructurados, es decir, que se pueden representar de manera matricial, como sucede con las hojas de cálculo. Aquí, cada caso corresponde a una fila y sus características, también llamadas variables, se incluyen como columnas. Estas variables pueden ser de tipo numérico o categórico, que en todo caso resultan representándose a partir de valores numéricos. De esta forma, se cuenta con datos susceptibles de ser procesados cuantitativamente, con el objetivo de identificar grupos o predecir valores futuros de una serie de interés.

200. Sin embargo, mucha de la información disponible en la actualidad corresponde a textos que describen todo tipo de situaciones. Estos datos no son estructurados, pues no hay manera de identificar directamente casos ni variables y, más aún, no son susceptibles de ser procesados cuantitativamente, al menos no sin hacer previamente algunas transformaciones. Estas transformaciones son precisamente el inicio de la minería de textos, que modifica la información textual para llevarla a una forma estructurada, donde cada fila corresponde a un documento y cada columna es una variable. Sobre las variables, hay muchas formas de construir características a partir de los contenidos de un texto, pero la más sencilla y popular es generar series que indican si una palabra determinada se encuentra en un documento.

201. Son varias las tareas que se pueden abordar desde la minería de textos:

- Clasificación de documentos: inicialmente, se toma un conjunto de documentos textuales y se transforman para obtener variables numéricas que se organizan como una hoja de cálculo. Aquí, cada fila corresponde a un documento y cada columna es una variable que puede indicar la presencia de una palabra determinada. Adicionalmente, se debe contar con una variable que determina el tema o grupo al que pertenece el documento. Con esta información se pueden utilizar metodologías de aprendizaje de máquina supervisado para obtener modelos que predigan el grupo o tema de un nuevo documento. Este tipo de procesamientos son útiles para el Sistema ALA/CFT, por ejemplo, para construir sistemas de clasificación automática de ROS, que toman como base la descripción de las operaciones sospechosas, las transforman en series numéricas y de allí predicen si el reporte es prioritario o no. Un sistema de esta naturaleza puede aprovechar, además, la información numérica directamente disponible, como el valor monetario de las operaciones reportadas, la cantidad de personas naturales y jurídicas relacionadas, la vinculación de las personas involucradas en otros reportes de inteligencia financiera, etc.

- Recuperación de información: aquí se busca identificar los documentos que contienen información relacionada con un criterio de búsqueda de interés. Este criterio es un conjunto de palabras o frases que indican el tema sobre el cual se está realizando la exploración. La identificación de casos previamente procesados que se relacionen con una descripción de interés es una actividad común en la inteligencia financiera. Incluso, la extracción de información puede utilizarse para comparar un nuevo ROS con un conjunto de reportes previamente analizados y, de esta manera, identificar procedimientos de investigación que han sido exitosos y pueden aplicarse a la nueva situación.

- Clústering y organización de documentos: aunque este caso es similar al de clasificación de documentos presentado anteriormente, aquí no se cuenta con una variable que defina el grupo al cual pertenece cada texto. De esta manera, la tarea consiste precisamente en tomar las descripciones disponibles, procesarlas para llevarla a variables numéricas y aplicar algoritmos especializados para agrupar los documentos de manera tal que cada segmento contenga documentos con palabras iguales, esperando que esto refleje contenidos sobre temas similares.

- Extracción de información: el objetivo de este procesamiento es identificar características particulares en información textual para extraerlas y almacenarlas en una estructura de hoja de cálculo. Por ejemplo, se puede procesar un documento notarial para extraer los valores de una propiedad, su ubicación y los nombres y números de identificación de los propietarios. De esta manera se ahorra tiempo de procesamiento y se obtienen datos cuantitativos que se pueden estudiar mediante metodologías de análisis de datos.

- Predicción: este caso también es muy parecido al de clasificación de documentos, aunque en esta ocasión se busca predecir una variable numérica. Volviendo al ejemplo del sistema de clasificación automática de ROS, si en lugar de la variable que indica si el reporte es prioritario o no se cuenta con una serie que califique el reporte según su importancia en una escala de 1 a 5, el sistema podría basarse en la descripción de un nuevo ROS para predecir su calificación.

202. Una vez se tiene el conjunto de documentos que se quiere analizar, el primer paso para procesarlos mediante minería de textos es extraer las palabras, o *tokens*, que lo componen. Para esto se debe dar un tratamiento especial a signos especiales como las interrogaciones, admiraciones, etc., que pueden o no ser considerados como palabras.

203. Una vez se cuenta con el listado de *tokens* que conforman cada documento, se aplica el procedimiento de *stemming* o lematización, que consiste en reducir las palabras a una forma que mantenga su significado, pero reduzca la cantidad de variaciones. Por ejemplo, las palabras reporte y reportes se transforman en reporte, y las palabras sospechoso, sospechosa y sospecha se transforman en el *token* sospecha. En general, la lematización conserva únicamente las raíces de las palabras, dando como resultado un menor número de *tokens* asociados a cada documento.

204. Una vez se cuenta con el conjunto de palabras en su forma reducida, se procede a la construcción de variables. Dos alternativas comunes en este punto son variables que indica si una palabra determinada aparece o no, y variable que muestran la cantidad de veces que aparece una palabra. Por ejemplo, los textos “El GAFI publicó cuarenta recomendaciones” y “GAFI publicó nueve recomendaciones especiales” generan las variables que se presenta en la Tabla 11.

205. Aquí es importante notar que cada palabra del texto se transforma en una serie, que las palabras ahora aparecen en minúscula y sin acentos para maximizar las coincidencias entre ellas, que los valores de cada variable coinciden con la presencia o no del *token* en el texto (1 si aparece, 0 en caso contrario), y con la cantidad de veces que cada palabra aparece en cada texto, y que no se aplicó un proceso de lematización.

Documento	el	gafi	publico	cuarenta	recomendaciones	nueve	especiales
1	1	1	1	1	1	0	0
2	0	1	1	0	1	1	1

Tabla 11. Variables generadas a partir de dos textos

206. Es posible aplicar otros procesos previos a la generación de variables, como la eliminación de palabras que son comunes, pero no aportan al contenido de un texto. Estas palabras se denominan *stopwords*, y usualmente están relacionadas más con la gramática que con la idea que se quiere transmitir. También, es usual que se construyan medidas que capturen la importancia de una palabra, como sucede con la formulación $tf - idf$, donde

$$tf = tf(j) \times idf(j)$$
$$idf(j) = \log\left(\frac{N}{df(j)}\right)$$

207. Esto es, para la palabra j, se calcula tf como el producto entre $tf(j)$, que es la cantidad de veces que aparece j en todos los documentos analizados, y $idf(j)$, que es la frecuencia de documento inversa para la palabra j. A su vez, $idf(j)$ se calcula como el logaritmo natural del cociente entre N, la cantidad de documentos considerados, y $df(j)$, que es la cantidad de documentos de documentos que contienen a j. De acuerdo con esta medida, cuando una palabra aparece en muchos de los documentos bajo análisis, se considera que es menos importante y su ponderación se reduce. Ahora, si una palabra aparece en pocos documentos, se aumenta su ponderación porque se considera más importante.

208. Ya sea que el interés esté en la conformación de grupos de documentos, o en la predicción de clases o variables numéricas, la minería de textos es útil para aprovechar información que se encuentra en forma de textos. Como se mostró anteriormente, permite el procesamiento especializado de documentos convirtiendo las palabras, o sus combinaciones, en variables numéricas que pueden alimentar modelos de clústering o de aprendizaje de máquina supervisado. Aquí lo importante es comprender los pasos del procesamiento que se aplican, para escoger aquellos que tengan más sentido de acuerdo con el objetivo que se persiga.

4.7. REDES COMPLEJAS

209. Las redes están presentes en muchos fenómenos de la naturaleza. Por ejemplo, en el campo de la biología, la cadena alimenticia se puede representar como una red de individuos que se relacionan porque unos cazan a otros. También, las células de un organismo interactúan entre sí intercambiando sustancias.

210. En las ciencias sociales las redes ocurren con frecuencia, entre personas por sus relaciones familiares, de amistad o laborales; entre empresas, por el intercambio de bienes y servicios y por los flujos que se generan en contraprestación; entre personas y empresas y sus activos, por las relaciones de propiedad; etc. En particular, los intercambios de dinero entre individuos, las relaciones de propiedad y otras contractuales generan una red que es de interés para el Sistema ALA/CFT, y que se puede analizar para identificar comportamientos que se asocian con actividades ilícitas.

211. Las redes que se tratan en esta sección están compuestas por un conjunto de actores, llamados nodos, que se conectan entre sí mediante algún tipo de relación, denominada vínculo o

borde. El

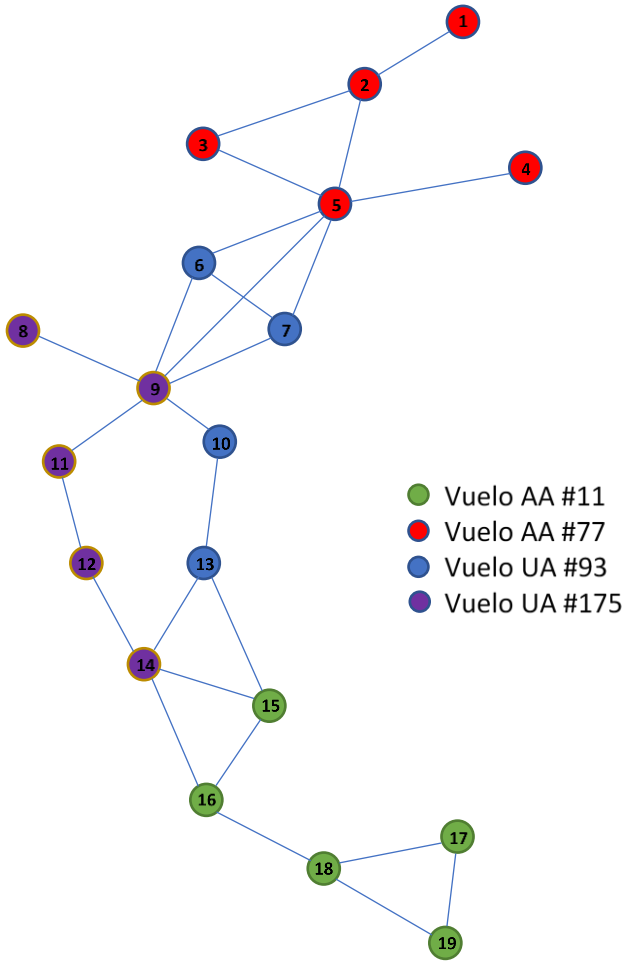

Gráfico 30 muestra la red de contactos de los 19 secuestradores que participaron en los ataques del 11 de septiembre a las Torres Gemelas, de acuerdo con el trabajo de Krebs (2002).

212. Es usual que en la representación gráfica de las redes los nodos sean simbolizados por círculos y los vínculos mediante líneas. Visualmente, es fácil notar que los atacantes tenían pocas relaciones entre sí, sin que exista un actor predominante que sirva como punto de contacto de los demás. Este tipo de redes, que no están densamente conectadas, son más difíciles de detectar o interrumpir, pues interceptar a uno de los nodos no genera un impacto que afecte de manera importante las relaciones entre los demás individuos.

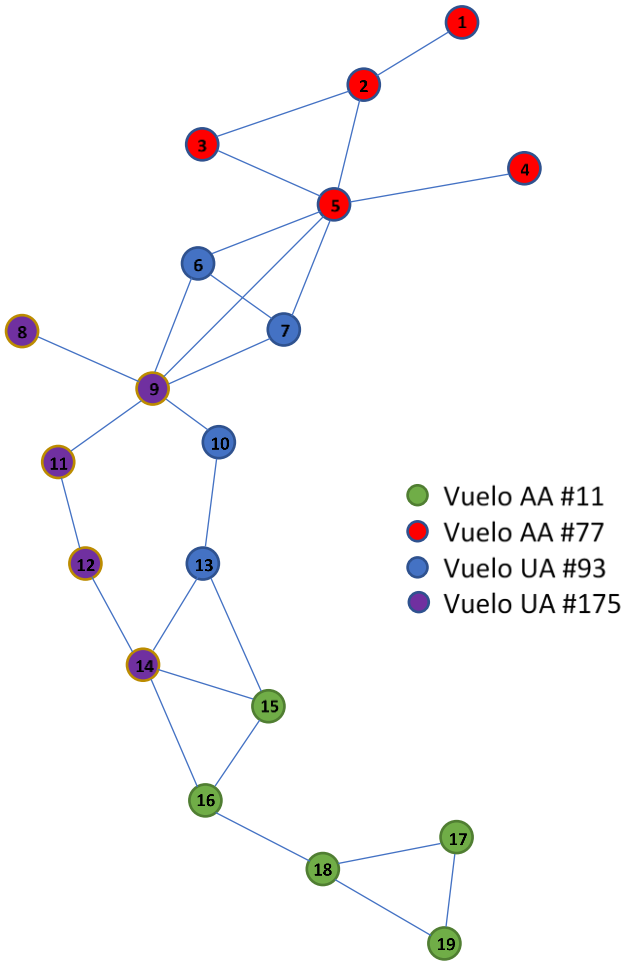

Gráfico 30. Red de secuestradores 11 de septiembre

213. Los datos subyacentes a la red se pueden representar en dos columnas, como se muestra en la Tabla 12. Aquí, el primer campo contiene la identificación del individuo de origen y el segundo la identificación del individuo de destino. Si las relaciones no tienen dirección, como sucede entre dos personas que se conocen, es suficiente con que el vínculo se represente una sola vez, esto es, una vez se incluye la fila que representa el vínculo entre la persona 1 y la persona 2, ya no hay necesidad de incluir una fila con el vínculo entre el individuo 2 y el individuo 1.

214. Ahora, si la relación tiene dirección, como en un flujo financiero, la tabla de datos debe incluir dos filas para representar los flujos desde 1 hacia 2 y desde 2 hacia 1, respectivamente. Adicionalmente, se pueden incluir columnas adicionales que incluyan atributos sobre el vínculo, por ejemplo, valores acumulados de transferencia, cantidad de transacciones, tipo de relación, etc. De la misma manera, se puede tener una tabla adicional con información sobre los nodos, como su nombre, identificación, nacionalidad, actividad económica, valor de sus ingresos, etc. Esta información se puede combinar para generar una red más completa, con direccionalidad en las relaciones y caracterización de los individuos.

Nodo de origen	Nodo de destino
1	2
2	3
2	5
5	4
5	6
5	7
5	9
6	7
6	9
7	9
8	9
9	10
9	11
10	13
11	12
12	14
13	14
13	15
14	15
14	16
15	16
16	18
17	18
17	19
18	19

Tabla 12. Datos subyacentes a la red secuestradores 11 de septiembre

215. Las redes pueden analizarse mediante diferentes estrategias y medidas. La primera de ellas, y la más común, es a través de un gráfico, como se mostró en el caso anterior. Adicionalmente, se pueden utilizar medidas para describir la estructura general de la red, caracterizar el comportamiento específico de los nodos o los vínculos, o identificar grupos. También se construyen modelos matemáticos y estadísticos para estudiar la red y su dinámica. Sobre las medidas de la red, las más comunes son:

- Tamaño: se refiere a la cantidad de nodos o vínculos en la red. Para el ejemplo del

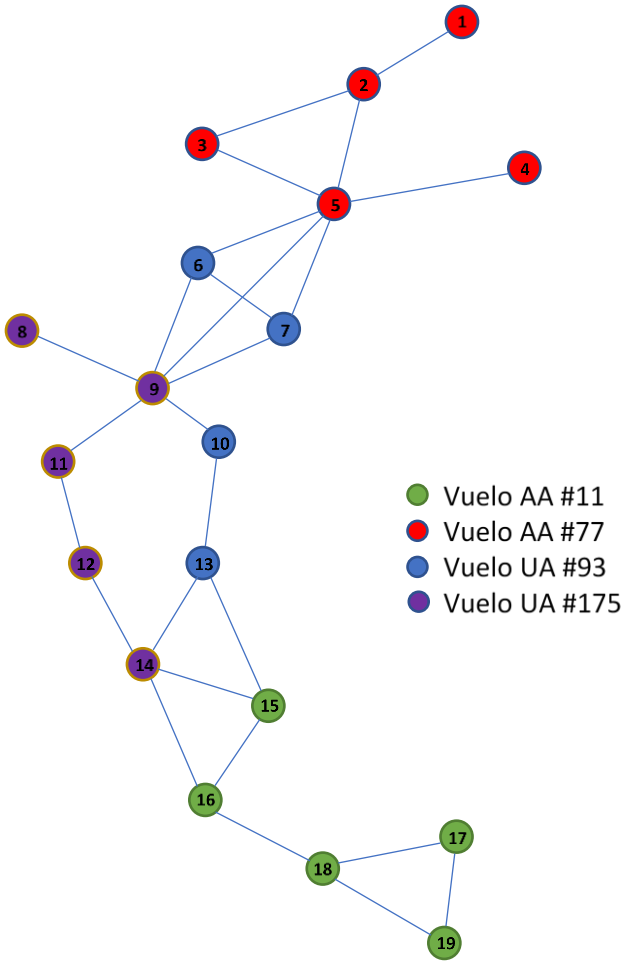

  Gráfico 30, se tienen 19 nodos y 25 relaciones.

- Densidad: es la proporción de vínculos observados con respecto al máximo número de vínculos posibles, es decir, la cantidad de vínculos que habría si todos los nodos se conectaran entre sí. Este es un número que va de 0 a 1, y entre más cercano a 1, más interconectada la red. La densidad se calcula dependiendo de si los vínculos en la red son dirigidos (se conoce la dirección de la relación, como en un flujo financiero) o sin dirección (como sucede entre dos personas que se conocen). Para una red con vínculos dirigidos, el número máximo de vínculos entre n individuos es $n(n-1)$, con lo cual la fórmula de la densidad es

$$\frac{V}{n(n-1)}$$

  donde V es el número de vínculos observados en la red. Para una red con vínculos no dirigidos el número máximo de vínculos es $n(n-1)/2$. De esta manera, la fórmula de la densidad para este caso es

$$\frac{2V}{n(n-1)}$$

- Componentes: son subgrupos de individuos que están conectados, de manera directa o indirecta. Estos clústers de individuos se pueden identificar a partir de su cohesión social, que depende de qué tan frecuentes, fuertes y dirigidos sean sus vínculos; o por detección de comunidades, cuando un subconjunto de nodos tiene un número relativamente alto de conexiones y relativamente pocos vínculos con otras partes de la red.
- Diámetro: es una medida de qué tan compacta es una red, considerando su tamaño y grado de interconexión. Aquí, un camino es una serie de pasos que se requieren para ir del nodo A al nodo B en la red. El camino más corto es el que requiere menos pasos. El diámetro de una red es el mayor de los caminos más cortos entre todas las parejas de nodos. Esta medida refleja el peor escenario de envío de información a través de la red.
- Coeficiente de clústering: una de las características de las redes que surgen por fenómenos naturales o sociales es la presencia de clústering, o la tendencia a formar triángulos cerrados. Este proceso se da porque dos individuos que tienen una relación y comparten un contacto resultan formando, a su vez, una relación entre ellos. Esta tendencia se puede medir a partir de la transitividad, que se define como el cociente entre el número de triángulos cerrados y la cantidad total de triángulos cerrados o abiertos[14] que se observan en la red, que es un número que va de 0 a 1.

216. Las medidas sobre los nodos buscan establecer qué tan prominente es un individuo particular. Esto es, si sus relaciones hacen que sus acciones sean visibles a los demás miembros de la red. Algunas de las utilizadas son:

- Grado: para una red sin direccionalidad, el grado de un nodo es la cantidad de conexiones que este tiene. En las redes dirigidas, el grado puede ser de entrada o salida. El primero considera la cantidad de relaciones que se dirigen hacia el nodo, el segundo corresponde al número de relaciones que se originan en el nodo. Considerando que esta medida se puede calcular para todos los individuos de la red, es posible obtener la distribución del grado[15] y analizar si dos o más redes tienen, estadísticamente, la misma distribución o siguen alguna distribución teórica[16]. También, puede ser de interés organizar los nodos, de mayor a menor, según su grado, para identificar los individuos con mayor conexión en la red.
- Centralidad: indica si el individuo está en una posición central. La centralidad se puede calcular a partir del grado (centralidad por grado) como se muestra a continuación:

$$C_D(n_i) = d(n_i)$$

donde C_D es la centralidad por grado, n_i se refiere al nodo i y d es la función que calcula el grado de un nodo. También se puede calcular la centralidad por cercanía, que mide qué tan cerca está cada nodo a todos los demás nodos en la red. Esta medida se calcula a partir de

[14] Un triángulo cerrado es el que conforman 3 nodos que se relacionan entre sí. Un triángulo abierto es el que surge de 3 nodos cuando se observan dos o tres relaciones entre ellos.

[15] La distribución se refiere a qué tan frecuente es que se observe un valor determinado de grado.

[16] Entre las distribuciones teóricas más conocidas están la distribución Poisson, la exponencial y la distribución libre de escala.

$$C_C(n_i) = \left(\sum_{j=1}^{g} d(n_i, n_j) \right)^{-1}$$

donde $d(n_i, n_j)$ es la cantidad de pasos que se deben dar en la red para llegar al nodo n_j desde el nodo n_i. De esta forma, la centralidad por cercanía para el nodo n_i es el inverso multiplicativo de la suma de sus distancias a todos los demás nodos. Por último, se presenta la centralidad entre nodos, que mide si un nodo se ubica entre parejas de otros nodos en la red, de manera que los caminos entre los otros nodos tienen que pasar a través del nodo en cuestión o, en otras palabras, que el nodo de interés intermedia la comunicación entre los otros dos nodos. Para calcular esta medida se utiliza

$$C_B(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}}$$

donde g_{jk} es el camino más corto entre los nodos j y k, y $g_{jk}(n_i)$ es la cantidad de caminos más cortos entre j y k que involucran el nodo n_i.

- Puntos de corte: son nodos que al ser retirados de la red incrementan el número de componentes de la red. En el

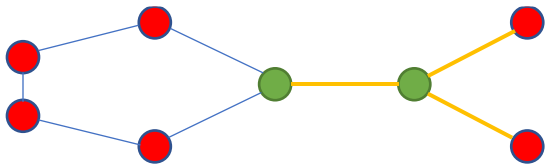

Gráfico 31, los nodos en verde son puntos de corte porque, si se retiran de la estructura, se generan subredes de individuos que ya no se comunican entre sí.

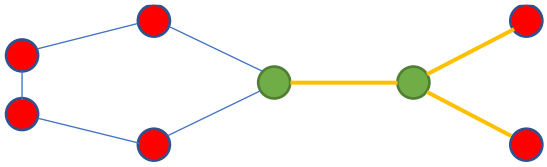

Gráfico 31. Puntos de corte en una red

217. Para los vértices existe una medida equivalente a los puntos de corte, llamada puente. En el

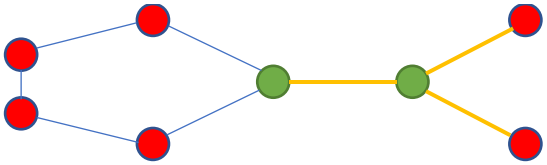

Gráfico 31 los vínculos en amarillo son puentes porque al retirarlos se generan subestructuras en la red.

218. Las situaciones que estudia la inteligencia financiera pueden conformarse como una red, ya que involucran diferentes individuos (personas naturales o jurídicas, activos, documentos, etc.) y sus relaciones que, en general, son por envíos o recepción de dinero, por propiedad sobre activos, por relacionamiento en documentos, etc.

219. Con esto, el análisis de redes permite visualizar la situación que se está investigando[17], pero también caracterizarla a partir de los indicadores presentados anteriormente. De esta manera, se puede identificar que las redes de dos casos diferentes se parecen porque tienen una estructura similar, y esto puede servir para aplicar el análisis de una situación sobre la otra, o para verificar posibles vínculos que aparentemente no existen por tratarse de pesquisas independientes.

4.8. CASOS DE USO EN EL ANÁLISIS DE INTELIGENCIA FINANCIERA

220. En esta sección se proponen algunos casos representativos de uso del análisis de datos para apoyar las labores de la inteligencia financiera. Aquí se busca profundizar en algunos detalles y consideraciones relevantes para entender mejor las potencialidades de las herramientas expuestas.

4.8.1. Análisis de anomalía para detección de individuos atípicos

221. La sección 4.4 presenta el análisis clúster, que es la metodología del aprendizaje de máquina no supervisado más popular en la inteligencia financiera. Esto es así porque permite la conformación de grupos compuestos por individuos homogéneos, manteniendo la mayor heterogeneidad entre grupos. De esta manera, al definir los patrones que caracterizan a cada persona natural o jurídica, es posible también identificar aquellos individuos que se comportan de una manera diferente.

222. El análisis clúster se elabora a partir de la información disponible en un conjunto de variables, que en la inteligencia financiera suelen estar asociadas con el comportamiento económico, financiero y tributario de los sujetos bajo estudio. Que se disponga de medidas sobre varias dimensiones de los individuos es ventajoso, pues permite que el análisis sea más exhaustivo.

223. Sin embargo, también impone retos por la recolección de datos de varias fuentes y por la necesidad de utilizar métodos que reduzcan estas señales a otras más sencillas, sin perder información valiosa en el proceso. Precisamente esto es lo que se pretende con el uso de las distancias, que transforman un vector multivariado representativo de cada persona en un número que indica qué tan parecido es el sujeto con a sus pares. Aquí, como la similitud no tiene

[17] El software de IBM I2 se utiliza para la estructuración de casos de inteligencia financiera, y una de sus principales ventajas es el relacionamiento gráfico de las estructuras.

dirección[18], sólo se necesita calcular la mitad de las distancias. Adicionalmente como la similitud de un sujeto consigo mismo es 0, no se requiere el cálculo de estas medidas. De esta manera, para N individuos, se deben calcular $N(N-1)/2$ distancias[19]. Este esfuerzo computacional puede llegar a ser superior a la capacidad disponible, sobre todo porque cada vez es más usual analizar más personas, cada una de ellas caracterizada cada vez más por muchas variables.

224. Así, por ejemplo, la matriz de distancias para 10.000 sujetos requiere de $10.000(10.000-1)/2 = 49'995.000$ cálculos, y este valor crece de manera exponencial con N. Para tratar con este tipo de situaciones, que son usuales en el análisis clúster, se pueden hacer segmentaciones preliminares a partir de alguna característica de interés, como el sector económico o la ubicación geográfica, para luego proceder con el uso de los algoritmos de clústering sobre conjuntos de individuos más pequeños.

225. También se pueden seguir estrategias basadas en hardware, donde se usan máquinas de mayor capacidad computacional, o se utilizan implementaciones del método de distancia que hacen uso no sólo de la unidad de procesamiento central (CPU) sino también de unidades de procesamiento gráfico (GPU) para agilizar los cálculos. Otra posibilidad es el uso de algoritmos alternativos de clústering, como los que se basan en densidades, que no dependen del cálculo de una matriz de distancia, facilitando el procesamiento, incluso ante un gran volumen de individuos de interés.

226. Una vez se cuente con los grupos, sea que estos se obtengan por segmentaciones basadas en juicio de expertos o por algoritmos de clústering, se puede obtener la distancia de cada individuo al centroide de su grupo. Desde este punto, siguiendo lo tratado en la sección 4.4.5, es posible obtener un umbral a partir del cual se considere que un individuo es atípico por estar suficientemente lejos de un valor representativo de su segmento.

227. Las personas con distancia al centroide de su clúster superior a este umbral se considerarán atípicas y podrán ser priorizadas para un análisis detallado, seguramente conducido por un analista operativo, a partir del cual se validen la situación y se establezca si existe o no un comportamiento sospechoso relacionado con LA o FT. En otras palabras, una distancia al centroide del grupo superior al valor establecido se conformará en una señal de alerta que conduzca a un análisis de inteligencia financiera.

228. Este procedimiento se puede automatizar. Para esto es deseable contar con una tabla de datos depurada, donde cada fila corresponda a un individuo y cada columna a sus características, permitiendo así ahorrar el tiempo y los recursos computacionales que requiere el preprocesamiento de la información. Adicionalmente, se deben desarrollar códigos en programas especializados (ver sección 5) que carguen los datos, los procesen mediante alguna de las

[18] Si se mide el parecido entre A y B ya se conoce también el parecido entre B y A.

[19] Si se tuvieran que calcular todas las distancias entre N individuos, se tendrían que realizar $N \times N = N^2$ operaciones. Debido a que no se calculan las distancias de cada persona consigo misma, se ahorran N cálculos, resultando entonces en $N(N-1)$ operaciones. Finalmente, como no hay dirección en el cálculo, sólo es necesario hacer la mitad de las operaciones, es decir, $N(N-1)/2$.

funciones ya implementadas para el análisis clústering e identifiquen los individuos atípicos bajo análisis de anomalía. En todo caso, es necesario hacer una revisión manual sobre una muestra de casos no priorizados para garantizar que el proceso no está obviando consistentemente situaciones de interés. Si se identifican este tipo de situaciones se deberá verificar la implementación realizada, validar si todas las variables de interés están siendo consideradas y si la generación automática de segmentos y umbrales es la esperada.

4.8.2. Sistema de clasificación automática de ROS

229. En la sección 4.5 se exponen los conceptos relacionados con el aprendizaje de máquina supervisado, donde algoritmos especializados identifican reglas en conjuntos de datos que permiten pasar de un grupo de características disponibles a una variable de interés. En particular, la sección 4.5.2 presenta el caso de la regresión logística y los árboles de clasificación para tareas donde se desee determinar la pertenencia de un caso a un grupo determinado.

230. De esta manera, es posible pensar en un sistema de clasificación automática de ROS. Aquí, la variable de interés será una indicadora de si el reporte debe ser priorizado o no para un análisis operativo detallado. Como variables de análisis, también denominadas independientes o explicativas, se consideran indicadores financieros de los individuos relacionados, variables indicando si alguno de ellos ha estado previamente relacionado en algún ROS o caso de inteligencia financiera, y variables que se obtienen de la descripción del reporte. Para estas últimas se utilizan los procedimientos de la minería de textos, presentados en la sección 4.6.

231. A partir de los ROS que han sido procesados anteriormente por la UIF, es posible construir una tabla de datos donde cada fila es un reporte de operación sospechosa, $N-1$ de las columnas son variables explicativas y la N-ésima columna es la variable de interés que determina si el reporte ha sido priorizado anteriormente o no.

232. A continuación, una porción de estos datos, usualmente el 60% o 70%, denominada conjunto de entrenamiento, se utiliza para ajustar el modelo de clasificación. Luego, la parte restante de la tabla de información, llamada conjunto de prueba, se utiliza para generar predicciones del modelo, y estos resultados se contrastan con el valor real de la variable de interés. En términos de clasificación, la comparación suele realizarse mediante una tabla cruzada llamada matriz de confusión. Aquí, si la variable que se quiere predecir toma dos valores, por ejemplo, reporte priorizado o reporte no priorizado, como el caso que se viene analizando, la tabla tendrá tamaño 2×2 (dos filas y dos columnas).

233. Por filas se representan las predicciones y por columnas los valores observados, como se muestra en la Tabla 13. La diagonal de la matriz de confusión (sombreada en azul) cuenta los aciertos del modelo de clasificación, ya sea que la predicción indique que el reporte es prioritario o no. Las celdas por fuera de la diagonal son los casos donde se equivocó el modelo. Entonces, entre todos los resultados (ajustes) posibles del modelo de clasificación, se prefiere aquel que genere, simultáneamente, la mayor cantidad de aciertos y el menor número de equivocaciones.

		Observado	
		Reporte priorizado	**Reporte no priorizado**
Predicción	**Reporte priorizado**	Verdaderos positivo	Falsos negativos
	Reporte no priorizado	Fasos positivos	Verdaderos negativos

Tabla 13. Matriz de confusión para dos categorías

234. Aunque este criterio de selección de los mejores modelos de clasificación es una simplificación, porque no considera situaciones relacionadas con la bajísima frecuencia de reportes priorizados con relación al total de ROS (desbalanceo de clase), ni contempla otras medidas de ajuste, ilustra el proceso que se debe seguir para generar un sistema de clasificación automática de reportes de operaciones sospechosas.

235. Adicionalmente, que los modelos se ajusten sobre un conjunto de datos y se prueben sobre otros que no han sido previamente procesados por el algoritmo de clasificación, es una estrategia conocida como validación cruzada. Aquí la idea es poder replicar las condiciones que el modelo enfrentaría al tener que clasificar casos nuevos, y evitar que este se ajuste por completo al conjunto de entrenamiento, con lo cual daría predicciones perfectas de los datos que ya son conocidos, pero podría hacer predicciones muy equivocadas de registros nuevos[20].

236. Para terminar, este sistema de clasificación automática también se puede definir para diferentes categorías de prioridad de ROS. Es decir, se puede predecir, por ejemplo, uno de cinco niveles de prioridad que se asignan a un reporte. En este caso, el procedimiento de desarrollo será similar, aunque las medidas de ajuste tendrán que ser diferentes por tratarse de cinco niveles de interés.

4.8.3. Sistema de clasificación de personas de interés

237. Este caso de uso es similar al expuesto anteriormente pues utiliza las mismas herramientas de aprendizaje supervisado para la identificación de individuos de interés. Aquí la diferencia principal radica en el tipo de individuos que se procesan y las metodologías que se usan para la generación de variables. Entonces, se identifican las posibilidades de un sistema que analice la situación de un conjunto de personas naturales o jurídicas, a partir de la información que se tenga disponible para ellas. Aquí es importante que en el conjunto de datos se cuente con una variable que permita identificar cuáles de las personas han sido de interés, ya sea por su condición de atípica, sospechosa, procesada o judicializada. Entre las variables de caracterización, se consideran los indicadores usuales financieros y económicos, como nivel de ingresos y egresos, de activos, pasivos y patrimonio, sector económico, actividad, etc., y otros indicadores que se

[20] Esta situación no deseada es conocida como sobreajuste del modelo. Aquí se ilustra para el caso de clasificación, pero también se observa en los algoritmos de regresión.

construyan a partir del análisis de redes complejas (sección 4.7). Sobre estos últimos, el grado y la centralidad son dos medidas que pueden ser relevantes, puesto que permiten darse una idea de qué tan importante es un individuo en una red de transacciones, o si ocupa una posición privilegiada que centralice el flujo de recursos.

238. El sistema puede desarrollarse para que se ocupe de la identificación de individuos cuyo comportamiento se asocia con una condena judicial por delitos relacionados con LA o FT. Este sistema, que podría ubicarse naturalmente en el MP, también puede ser generado por las UIF si cuentan con una base de datos que identifique los individuos que han tenido esta condición. Esto puede ser útil porque permite que el sistema enfoque sus esfuerzos en aquellas situaciones que fueron judicialmente validadas.

239. Asimismo, un sistema especializado en la identificación de personas sospechosas de estar involucradas en conductas LA/FT es naturalmente útil para las UIF, pero también podría ser generado por las entidades reportantes, generando sinergias entre los diferentes eslabones del sistema en la medida que cada uno de ellos trabajaría directamente para la identificación de las situaciones que son de interés para el eslabón siguiente. Una vez más, para esto es necesario que los reportantes cuenten con el listado de individuos de interés.

240. Es importante mencionar que este tipo de colaboración entre entidades del sistema ALA/CFT no necesariamente sucede por las restricciones legales que existen y por el tratamiento reservado que suele darse a este tipo de información. Sin embargo, vale la pena considerar qué posibilidades está disponibles para incrementar la cohesión y eficiencia en el proceso de identificación y análisis de operaciones sospechosas.

5. HERRAMIENTAS TECNOLÓGICAS

241. Las decisiones que se toman en términos de hardware y software determinan las capacidades y demanda de recursos de las entidades del Sistema ALA/CFT para gestionar la información a la que acceden. Aquí, aunque el panorama parece ser inicialmente amplio, puede reducirse en la medida que se consideran las restricciones para almacenar datos por fuera de las jurisdicciones.

5.1. HARDWARE

242. Son varias las tecnologías disponibles para el almacenamiento y procesamiento de datos, desde soluciones bien consolidadas basadas en un servidor, hasta infraestructuras en la nube. Al respecto, considerando la variedad y volumen de datos que suelen manejarse en UIF y MP, se recomienda el uso de máquinas especializadas, con suficiente capacidad de almacenamiento, tanto en disco duro como en memoria RAM, y procesadores especializados en ejecutar de manera dedicada funciones intensivas. Ahora, si se quiere contar con capacidad de procesamiento superior, se plantean dos alternativas: la primera, utilizar un clúster de servidores (mencionados en la sección 4.1) y, la segunda, procesamiento en nube.

243. Sobre el clúster de servidores, vale la pena mencionar a Hadoop, que es un proyecto de software abierto que permite el procesamiento distribuido de grandes conjuntos de datos en este tipo de infraestructura. De esta manera se obtienen las ventajas en disponibilidad, velocidad y escalabilidad a partir de hardware que ya es conocido por la mayoría de las entidades. Aquí se debe mencionar que puede haber retos con la configuración de este tipo de infraestructuras, y que la oferta del personal calificado[21] para realizar este tipo de tareas no es abundante, con lo cual algunos de los ahorros en hardware se reducen por costos de configuración y mantenimiento.

244. El almacenamiento y procesamiento en nube se ha convertido en un concepto común, puesto que ambos están disponibles incluso para usuarios domésticos a través de herramientas de Microsoft o Google. A nivel empresarial es posible contar con este mismo tipo de alternativas, adaptables a las necesidades particulares de cada organización. De esta manera es posible asegurar la seguridad y disponibilidad de la información, y acceder a mayores capacidades de procesamiento por demanda.

245. Sin embargo, como se mencionó anteriormente, en muchos casos las UIF y los MP tienen restricciones para que los datos que gestionan y los análisis que realizan existan por fuera de su jurisdicción, lo cual impone restricciones, que en muchos casos pueden resultar insalvables, para este tipo de alternativas. Sin perjuicio de ello, vale la pena tenerlas en consideración puesto que esta es una tendencia hacia la cual las regulaciones de los países podrían evolucionar.

5.2. SOFTWARE

246. No son pocas las alternativas de software disponible para el análisis de datos. La oferta actual cubre los procedimientos más populares, que suelen ser suficientes para las necesidades analíticas de cualquier organización. Sin embargo, es importante considerar las diferencias que surgen entre las opciones de software libre y aquellas que son pagas, las cuales van más allá del factor económico. En consecuencia, a continuación, se describen algunas de ellas según su tipo de acceso.

247. En la actualidad, las alternativas de software libre más importantes[22] para el análisis de datos son Python[23] y R[24]. Python es un lenguaje de desarrollo que se ha popularizado como herramienta de análisis de datos, particularmente en el campo del aprendizaje de máquina. Entre sus virtudes se cuentan la facilidad de su lenguaje[25], la comunidad que tiene de respaldo y las

[21] Los perfiles con conocimiento en este tipo de tecnologías son los relacionados con la ingeniería de sistemas y la informática. Usualmente, es necesario considerar estudios especializados en computación distribuida, computación en nube o Big Data.
[22] Otras herramientas relevantes que vienen ganando popularidad son Julia y Rubi. También Octave, aunque con menor impulso en su adopción.
[23] https://www.python.org/
[24] https://www.r-project.org/
[25] El lenguaje de Python es similar a Java, lo cual permite establecer un vínculo con las áreas de tecnología.

posibilidades que ofrece para integrarse con otros desarrollos. Cuenta con librerías especializadas para el manejo de datos, como Pandas, NumPy, SciPy, Matplotlib y Tensorflow.

248. Por su parte, R es un programa diseñado directamente para el análisis de datos, por lo cual llega a ser superior que Python en lo relacionado con modelación y análisis estadístico. Cuenta con paquetes para el desarrollo de modelos de aprendizaje de máquina y una comunidad amplia que soporta su aprendizaje y uso. Ahora, su integración con otros desarrollos es más demandante.

249. En el software comercial se cuentan varias alternativas, como SAS[26], IBM SPSS Statistics[27] e IBM Modeller[28]. Estas herramientas permiten el análisis especializado de los datos y generación automática de reportes, entre otras, y cuentan con el respaldo y garantía de grandes compañías tecnológicas, lo cual puede ser un factor decisivo para muchos.

250. Este último factor, junto con los costos, es la mayor diferencia versus el software libre, que no cuenta con ningún tipo de garantía, y que por lo mismo debe manejarse con precaución, sobre todo si se utiliza alguna librería o paquete desarrollado recientemente, o que trata sobre un tema muy específico, lo cual suele asociarse con una comunidad de soporte reducida.

6. CONCLUSIONES Y RECOMENDACIONES

251. En términos generales, actualmente las UIF se encuentran en procesos avanzados de implementación de analítica de datos a través de diferentes herramientas y perfiles, que permiten el análisis masivo de información y la generación de productos en sus áreas estratégica y operativa. Por su parte, los ministerios públicos o fiscalías aún se encuentran en una etapa inicial y sólo algunas muestran avances en procesos de analítica de datos, lo cual puede explicarse porque estas entidades se enfocan en el en el estudio de casos puntuales. Es por esta misma razón que se puede presentar una oportunidad de trabajo conjunto para crear procesos de análisis de datos que aprovechen el flujo, calidad y cantidad de información, generando productos de estratégicos como tipologías o tendencias, y operativos en la identificación de redes o visión ampliada de vínculos. Estos procesos apoyados en actividades continuas de retroalimentación.

252. El avance tecnológico en lo relacionado con el almacenamiento y procesamiento de registros, la mayor disponibilidad de información sobre todo tipo de situaciones y el desarrollo de las metodologías para el procesamiento de grandes volúmenes de datos configuran una situación sumamente favorable para el Sistema ALA/CFT, que puede mejorar su eficiencia e incrementar su eficacia en la identificación de personas involucradas en actividades relacionadas con el LA y la FT. En particular, estas oportunidades pueden ser aprovechadas por las UIF y el MP, para mejorar sus procesos de administración de datos y generación de conocimiento y, con esto, incrementar su eficacia en la lucha contra las actividades ilícitas.

[26] https://www.sas.com/
[27] https://www.ibm.com/analytics/spss-statistics-software
[28] https://www.ibm.com/analytics/spss-statistics-software

253. La implementación de metodologías cuantitativas de procesamiento de información, como las enmarcadas en el aprendizaje de máquina y la minería de datos y de textos, permitirá potenciar las capacidades del Sistema ALA/CFT mediante la identificación de individuos que presenten comportamientos económicos atípicos, y la generación automática de reglas asociadas a conductas sospechosas. Todos estos esfuerzos son automatizables, apalancados en la abundante capacidad de cómputo y procesamiento de la que se dispone en la actualidad, e implican mayores posibilidades para analizar situaciones que no se han considerado previamente.

254. Ahora, para sacar un mejor provecho de la nueva coyuntura, se hacen las siguientes recomendaciones:

255. La interacción entre las UIF y los MP, en lo relacionado con la información a la que acceden, debe ser constante para fortalecer el análisis de casos. Por supuesto, aquí se deben considerar las restricciones impuestas por la reglamentación de cada jurisdicción.

256. Los MP pueden aprovechar mejor las posibilidades ofrecidas por las nuevas tecnologías mediante la incorporación en sus equipos de personal técnico/científico que tenga formación en el análisis de datos.

257. Trabajar en la caracterización y modelación de los fenómenos delictivos para ampliar el conocimiento que se tiene sobre ellos. También, invertir recursos en el desarrollo de herramientas de identificación automática de situaciones de interés, lo cual incrementaría la eficiencia y eficacia de todo el sistema ALA/CFT. Aquí, se deben incorporar mecanismos de control para evitar sesgos y para reconocer nuevas conductas delictivas.

258. Conformar grupos de trabajo a partir del software que utilicen las entidades, o de las metodologías de análisis de datos que trabajen, para aprovechar sinergias que surgen de la experiencia acumulada de los funcionarios en el manejo de estas herramientas.

259. Es de vital importancia generar regulaciones en el marco de la recomendación 15, con el fin de establecer parámetros y conductas asociadas al uso de la tecnología para el LA. Su desarrollo normativo debe considerarse en el corto plazo.

BIBLIOGRAFÍA

Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer International Publishing.

Barnes, T. (2013). Big Data, Little History. Dialogues in Human Geography, 3(3), 297-302.

Easly, B. y Kleinberg, J. (2010). Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press.

Fellman, P. V. y Wright, R. (2014). Modeling Terrorist Networks, Complex Systems at the Mid-range. The Intelligencer, Journal of U.S. Intelligence Studies, 14(1).

Fu, Y. (1997). Data Mining. IEEE Potentials, 16(4), 18-20, Oct.-Nov.

Hastie, T., James, G., Tibshirani, R., y Witten, D. (2013). An Introduction to Statistical Learning with Applications in R. New York. Springer.

Hernández, L., Santillán, A. y Caballero, C. (2003). Maestros y esclavos. Una aproximación de los cúmulos de computadoras. Revista Digital Universitaria, 4(2). http://www.revista.unam.mx/vol.4/num2/art3/art3.htm#

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica, 31(3), 249-268.

Krebs, V. (2002). Uncloaking Terrorist Networks. First Monday, 7(4). https://firstmonday.org/ojs/index.php/fm/article/download/941/863

Luke, D. (2015). A User's Guide to Network Analysis in R. Springer.

Masoud, M. (2019). Sensors of Smart Devices in the Internet of Everything (IoE) Era: Big Opportunities and Massive Doubts. Journal of Sensors, 2019, 1-26.

Mohanty, H. (2015). Big Data: An Introduction. Springer.

Monetary Authority of Singapore (2018). Guidance for Effective AML/CFT Transaction Monitoring Controls.

Robinson, I., Webber, J. y Eifrem, E. (2013). Graph Databases. O'Reilly Media, Inc.

Turing, A. M. (1950). Computing Machinery and Intelligence. Mind, 49, 433-460.

Wang, P. (2019). On Defining Artificial Intelligence. Journal of Artificial General Intelligence, 10(2), 1-37.

Weiss, S. M., Indurkhya, N. y Zhang, T. (2015). Fundamentals of Predictive Text Mining. Springer.