Vigilada Mineducación

ELABORACIÓN DE UN MODELO DE SEGMENTACIÓN PARA EL FACTOR DE RIESGO CLIENTE EN UNA ENTIDAD ADMINISTRADORA DE UN SISTEMA DE PAGO DE BAJO VALOR

JOHAAN ALBERTO MORENO OLIER

Trabajo de grado

Asesor:

Jorge Andrés Osorio Gómez

UNIVERSIDAD EAFIT

ESCUELA DE ADMINISTRACIÓN

MAESTRÍA EN ADMINISTRACIÓN DE RIESGOS

MEDELLÍN

2023

Tabla de contenido

Resumen

El presente documento desarrolla un modelo para la segmentación del factor de riesgo cliente en una entidad administradora de un sistema de pago de bajo valor, que cumpla con los aspectos estadísticos y normativos definidos por la Superintendencia Financiera de Colombia.

Palabras clave

Segmentación de factores de riesgo, EASPBV y SARLAFT.

Abstrac

This document develops a model for the segmentation of the customer risk factor in an administrative entity of a low-value payment system, which complies with the statistical and regulatory aspects defined by the Financial Superintendency of Colombia**.**

Key words

Risk factor segmentation, EASPBV y SARLAFT.

1. Introducción y planteamiento del problema

Un sistema de pago es un conjunto organizado de políticas, reglas, acuerdos, instrumentos de pago, entidades y componentes tecnológicos, tales como equipos, *software* y sistemas de comunicación, que permiten la transferencia de fondos entre los participantes del sistema, mediante la recepción, el procesamiento, la transmisión, la compensación y la liquidación de órdenes de pago o transferencias de fondos. Por su parte, los sistemas de pago de bajo valor son aquellos sistemas de pago que procesan órdenes de pago o transferencia de fondos distintas a las procesadas en el sistema de pago de alto valor, de conformidad con lo que defina el Banco de la República. En los sistemas de pago de bajo valor, para el procesamiento de órdenes de pago o transferencia de fondos entre la entidad emisora y el adquirente o la entidad receptora, se requiere de una entidad administradora de sistema de pago de bajo valor (Presidencia de la República, 2020).

Las entidades administradoras de sistemas de pago de bajo valor (EASPBV) se encuentran bajo la supervisión y vigilancia de la Superintendencia Financiera de Colombia, y son instituciones que desarrollan la actividad de compensación y liquidación en uno o más sistemas de pago. La compensación es un proceso que realiza la entidad administradora del sistema de pago de bajo valor para determinar, al cierre de un periodo establecido, el saldo que corresponda a cada uno de sus participantes, como resultado de las órdenes de pago o transferencias de fondos procesadas en el sistema de pago de bajo valor, para extinguir entre ellos sus obligaciones. Por su parte, la liquidación es un proceso que realiza la entidad administradora del sistema de pago de bajo valor con el cual finaliza una operación o conjunto de operaciones, mediante cargos y abonos en cuentas de depósito en el Banco de la República, en cuentas corrientes o de ahorros en un establecimiento de crédito, de las cuales sean titulares los participantes en un sistema de pago (Presidencia de la República, 2020).

La normatividad emitida por la Superintendencia Financiera de Colombia (SFC), relacionada con el Sistema de Administración del Riesgo de Lavado de Activos y de la Financiación del Terrorismo (SARLAFT), establece que para identificar los riesgos de Lavado de Activos y de la Financiación del Terrorismo (LA/FT) es necesario segmentar los factores de riesgo de LA/FT. Los factores de riesgo contemplados por la norma son: los clientes y/o usuarios, los productos, los canales de distribución y las jurisdicciones (Superintendencia Financiera de Colombia, 2022).

La segmentación consiste en aplicar metodologías de reconocido valor técnico para dividir cada factor de riesgo en grupos o segmentos homogéneos, de acuerdo con las características o variables de objetos que componen cada factor (Amaya, 2017).

Dentro de las metodologías más reconocidas se pueden citar las siguientes:

Knowledge Discovery in Data Bases (KDD): La metodología KDD es un proceso que busca el descubrimiento de conocimiento en bases de datos (Burgos & González, 2020).

Cross Industry Standard Process for Data Mining (Crisp DM): Es un método para orientar trabajos de minería de datos que consta de seis fases (Nishizaki, 2017).

Metodología SEMMA: Es un proceso que comprende las fases de selección, exploración y modelado de grandes volúmenes de datos para obtener patrones de comportamiento (Moine, Gordillo & Haedo, 2011).

Dentro de las técnicas podemos encontrar el análisis factorial, el análisis *cluster* o la regresión múltiple (Marques, 2018).

La segmentación de factores de riesgo de LA/FT es un proceso muy importante dentro del SARLAFT, pues es un punto de partida para la identificación de riesgos y permite establecer aspectos, como las señales de alerta, para el monitoreo de las operaciones o transacciones que

realizan los clientes o usuarios de las entidades vigiladas por la SFC. La segmentación de factores de riesgo de LA/FT no debe tomarse a la ligera, pues les corresponde a las personas que se encuentran al frente de la administración del riesgo LA/FT diseñar las metodologías de segmentación (Lozano, 2008).

La SFC ha impuesto, entre los años 2020 y 2021, sanciones a instituciones como Skandia Seguros de Vida S.A., Fiduciaria La Previsora S.A., Coltefinanciera y Banco GNB Sudameris por presentar falencias en su proceso de segmentación de factores de riesgo LA/FT; adicionalmente, es importante mencionar que desde la perspectiva normativa un cliente es toda persona natural o jurídica y estructuras sin personería jurídica con las cuales la entidad establece y mantiene una relación contractual o legal para el suministro de cualquier producto propio de su actividad (Superintendencia Financiera de Colombia, 2022).

Uno de los principales riesgos generados por el factor cliente es la posibilidad de establecer relaciones contractuales con personas naturales o jurídicas que tengan vínculos con los delitos del lavado de activos y la financiación del terrorismo, debido a que las empresas tienen la necesidad y la expectativa de conseguir nuevos clientes, pues es su principal forma de incrementar ingresos y rentabilidad a través del tiempo. Adicionalmente, los clientes pueden utilizar los servicios que ofrece la entidad para dar apariencia de legalidad a bienes o dineros provenientes de delitos como el narcotráfico, secuestros, extorsiones, corrupción, tráfico de personas, migrantes y armas (UIAF, 2006).

De acuerdo con las implicaciones normativas y los riesgos que se derivan de los clientes es necesario que las EASPBV construyan modelos de segmentación asociados a los factores de riesgo, con el fin de cumplir los requerimientos exigidos por la Superintendencia Financiera de Colombia, y que además sirvan como herramienta para prevenir y controlar el riesgo de lavado

de activos y financiación del terrorismo. Lo anterior con base en argumentos teóricos y técnicos de aplicación de ciencias como la estadística y matemática, que permitan tener bases para formalizar el proceso de identificación, medición, control y monitoreo del riesgo de lavado de activos y de la financiación del terrorismo (Daza, 2019).

Mediante este trabajo de investigación se busca responder a la pregunta: ¿Cómo elaborar un modelo de segmentación para el factor de riesgo cliente en una entidad administradora de un sistema de pago de bajo valor?

2. Objetivos

2.1 Objetivo general

Elaborar un modelo de segmentación para el factor de riesgo cliente en una entidad administradora de un sistema de pago de bajo valor.

2.2 Objetivos específicos

Aplicar la metodología KDD para la segmentación del factor de riesgo cliente en una EASPBV.

Aplicar técnicas de predicción para la segmentación del factor de riesgo cliente en una EASPBV.

Establecer señales de alerta asociadas a cada uno de los segmentos relacionados con el factor cliente en una EASPBV.

3. Aspectos metodológicos

3.1 Enfoque de la investigación

Cuantitativa, debido a que el modelo se construirá con datos numéricos relacionados con la información de balance y de resultados de las entidades seleccionadas para la investigación (Ortega, 2018).

3.2 Alcance del estudio

El alcance del estudio es correlacional, puesto que se busca definir y establecer mediante la aplicación de técnicas predictivas una relación entre las variables de segmentación de los clientes y el segmento al que pertenecen (Vásquez, 2005).

3.3 Sujetos y/o muestra

Se tomarán las variables de balance y de resultado de aproximadamente 150 entidades del sector solidario que son clientes de la EASPBV Visionamos.

3.4 Instrumentos o técnicas de recolección de información

Para recolectar los datos se recurrirá a las fuentes públicas de datos de la Superintendencia Financiera de Colombia y la Superintendencia de la Economía Solidaria, concretamente los estados financieros de las entidades vigiladas por las referidas Superintendencias y que harán parte de la investigación.

3.5 Justificación de la solución en términos de la Maestría en Administración de Riesgos

El modelo de segmentación del factor de riesgo cliente es parte fundamental en el proceso de identificación de riesgos, además que sirve de base para la aplicación en otras organizaciones que, al igual que Visionamos, se encuentre obligada a implementar un SARLAFT.

4. Resultados esperados

Obtener un modelo para la segmentación del factor de riesgo cliente en una EASPBV, que cumpla con los aspectos estadísticos y normativos definidos por la SFC.

Elaborar un documento que describa toda la metodología con la que se obtiene el modelo para la segmentación del factor de riesgo cliente en una EASPBV.

5. Marco conceptual

Tabla 1

Comparación de diferentes metodologías de minerías de datos

FASE	KDD	CRISP-DM	SEMMA
Análisis y comprensión del negocio	Comprensión del dominio de la aplicación	Comprensión del negocio	
Selección y preparación de los datos	Crear el conjunto de los datos, limpieza, procesamiento y reducción de los datos	Entendimiento y preparación de los datos	Muestreo
Modelado	Determinar tareas de minería, determinar el algoritmo de minería, minería de datos	Modelado	Modelado
Evaluación	Interpretación	Evaluación	Valoración
Implementación	Utilización del conocimiento	Despliegue	

Nota: Elaboración propia

Todas las metodologías son herramientas muy buenas, pero su aplicación varía en el tipo de proyecto que se desee aplicar. En este caso se elige la metodología KDD porque esta proporciona la facilidad de contar con tareas de minería para determinar los algoritmos que se deseen implementar; estos pueden ser de agrupamiento, regresión lineal, series temporales, entre otras. Además, proporciona muchos caminos para que se puedan implementar nuevas herramientas analíticas y predictivas, lo cual contribuye al logro de uno de los objetivos del presente trabajo, el cual consiste en aplicar técnicas de predicción para la segmentación del factor de riesgo cliente en una EASPBV (Burgos & González).

5.1 Metodología KDD

La metodología KDD realiza un proceso interactivo e iterativo en la búsqueda de patrones, parámetros y modelos, los mismos que deben ser potencialmente válidos y útiles. Se deben establecer medidas cuantitativas que ayuden a considerar la validez y utilidad de los

patrones escogidos, con el fin de integrar el conocimiento adquirido y aplicarlo en la toma de decisiones de algún sistema real, con la ayuda de los resultados alcanzados (Burgos & González).

5.1.1 Etapas de la metodología KDD

Las etapas de la metodología KDD pueden resumirse en las siguientes: selección, preprocesamiento/limpieza, transformación/reducción, minería de datos (*data mining*) e interpretación/evaluación (Timaran et al., 2016).

Etapa de selección

En la etapa de selección se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos, o una muestra representativa de este, sobre el cual se aplicará la metodología KDD. La selección de los datos varía de acuerdo con los objetivos del negocio (Timaran et al., 2016).

Etapa de preprocesamiento/limpieza

En esta fase se deben llevar a cabo tareas como identificar datos atípicos (*outliers*) y tratamiento de datos faltantes (*missing values*) (Moine, 2013).

Transformación/reducción

Aplicando técnicas de análisis exploratorio de datos (estadístico, gráfico, entre otros), se busca identificar la distribución de los datos, simetría, pruebas de normalidad y correlaciones existentes entre los datos. En esta etapa es útil el análisis descriptivo del conjunto de datos mediante técnicas como el *clustering* o el análisis de componentes principales (UIAF, 2014).

Adicionalmente, en esta etapa se realiza la transformación (por ejemplo, discretización, donde podemos transformar valores numéricos a categóricos para mejor la exactitud de ciertos

modelos) y reducción de los datos buscando mejorar la calidad y especificidad de los datos (Moreira, 2020).

Etapa de minería de datos (*data mining*)

En la fase de *data mining* se define el tipo de investigación a realizar de acuerdo con las características de los atributos que conforman la base de datos, que pueden ser de tipo descriptivo y/o predictivo (UIAF, 2014).

La minería descriptiva se utiliza generalmente para producir correlación, tabulación cruzada, frecuencia, etc. Estas técnicas están determinadas para encontrar las regularidades en los datos y revelar patrones. La otra aplicación del análisis descriptivo es descubrir los subgrupos cautivadores en la mayor parte de los datos (Cyberparts.pl., s.f.).

El objetivo principal de la minería predictiva es predecir resultados futuros en lugar del comportamiento actual. Implica las funciones de aprendizaje supervisado que se utilizan para la predicción del valor objetivo. Los métodos que se incluyen en esta categoría de minería son la clasificación, el análisis de series de tiempo y la regresión. El modelado de datos es la necesidad del análisis predictivo, que funciona utilizando algunas variables para anticipar los valores de datos futuros desconocidos para otras variables (Cyberparts.pl., s.f.).

Tabla 2

Correspondencia entre tareas y técnicas de minería de datos

Técnica	Predictivo		Descriptivo		
	Clasificación	Regresión	Agrupamiento	Reglas de asociación	Correlaciones
Redes neuronales artificiales	X	X	X		
Árboles de decisión	X	X			
Redes de Kohonen			X		
Regresión lineal y logarítmica		X			X
Regresión logística	X			X	
K-medias			X		
A priori				X	
Naive Bayes	X				
Vecinos más próximos	X	X	X		
Twostep, Cobweb			X		
Algoritmos genéticos y evolutivos	X	X	X	X	X
Máquinas de soporte vectorial	X	X	X		
Análisis discriminante	X				

Nota: Elaboración propia

Estas técnicas y algoritmos ya están probados y validados en multitud de escenarios que comprueban y verifican su exactitud (Gutiérrez, 2016).

Para este trabajo se utilizará el análisis discriminante como técnica de minería de datos cuyo objetivo es clasificar observaciones en diferentes grupos o categorías predefinidas. Busca encontrar combinaciones lineales o cuadráticas de las variables predictoras que maximicen la separación entre las clases y minimicen la variación dentro de las clases. El análisis discriminante es una técnica de aprendizaje supervisado, ya que utiliza las etiquetas de clase para entrenar y evaluar el modelo de clasificación (Linkedin.com, 2023).

Las características de dicha técnica se relacionan con el proceso, el cual busca la separación de elementos en grupos homogéneos en el interior de ellos y heterogéneos entre ellos (variables de segmentación) (Superintendencia Financiera de Colombia, 2022).

Etapa de interpretación/evaluación

En esta etapa se realiza la interpretación de los datos y la evaluación de los patrones, verificando el rendimiento que se obtuvo para cumplir con los objetivos planteados (Moreira, 2020).

6. Metodología

6.1 Etapa de selección

Para definir el conjunto de datos objetivo se tomarán los lineamientos definidos en la normatividad vigente, en materia de SARLAFT.

Con relación a los aspectos de segmentación, la norma, emitida por la Superintendencia Financiera de Colombia, plantea lo siguiente:

"Para los mencionados tipos de clientes, productos, operaciones, servicios, la segmentación de los factores de riesgo referido en el numeral 4.2.2.3.2. del presente Capítulo, debe realizarse con la información que tengan disponible las entidades. En todo caso, las entidades vigiladas, a medida que cuenten con información adicional, deben dar cumplimiento a las instrucciones del presente Capítulo" (Superintendencia Financiera de Colombia, 2022).

De acuerdo con lo anterior, se elegirán variables de los estados financieros de 147 clientes de Visionamos, al corte del 27 de diciembre del 2022, concretamente variables relacionadas con cuentas de mayor riesgo de lavado de activos.

Las cuentas contables de mayor riesgo para el lavado de activos son: Ingresos, Cuentas por cobrar, Inversiones, Inventario, Propiedad, planta y equipo y Obligaciones (Bareño, 2009).

Adicionalmente, la base de datos objetivo contiene la variable segmento, la cual se obtuvo tras la aplicación del análisis *cluster*, concretamente, el análisis de K-medias.

6.2 Etapa de preprocesamiento/limpieza

Se denominan casos atípicos u *outliers* a aquellas observaciones con características diferentes de las demás (Ocaña, s.f.).

Se habla de dos tipos de *outliers* los cuales pueden ser distinguidos como los *outliers* leves y los *outliers* extremos. Una observación es declarada como *outlier* extremo si esta cae

fuera del intervalo (Q1 − 3IQR, Q3 + 3IQR). Donde IQR = Q3 − Q1 llamado radio intercuartílico. Así mismo una observación x es declarada como un *outlier* leve si cae fuera del intervalo (Q1 −1,5IQR, Q3 + 1,5IQR), donde 3 y 1,5 son escogidos por comparaciones con una distribución normal (Moreno, 2012).

Tabla 3

Outliers variable ingreso

Ingreso						
Cuartil 1	Cuartil 3	Rango intercuartílico	Outliers extremos		Outliers leves	
Q1	Q3	IQR = Q3-Q1	(Q1 − 3IQR, Q3 + 3IQR)		(Q1 −1,5IQR, Q3 + 1,5IQR)	
3.238.071.773	18.376.986.121	15.138.914.348	- 42.178.671.273	63.793.729.166	- 19.470.299.750	41.085.357.644

Nota: Elaboración propia

En la variable ingreso se identifican un total de 10 datos por encima de $63.793.729.166, es decir, *outliers* extremos, lo que representa el 7% del total de los datos de la variable. Por su parte, en la referida variable se evidencian 18 datos por encima de $41.085.357.644, es decir *outliers* leves, con un porcentaje de participación del 12% del total de los datos de la variable.

Tabla 4

Outliers variable cuentas por cobrar

Cuentas por cobrar						
Cuartil 1	Cuartil 3	Rango intercuartílico	Outliers extremos		Outliers leves	
Q1	Q3	IQR = Q3-Q1	(Q1 − 3IQR, Q3 + 3IQR)		(Q1 −1,5IQR, Q3 + 1,5IQR)	
51.716.301	623.093.431	571.377.130	- 1.662.415.089	2.337.224.822	- 805.349.394	1.480.159.127

Nota: Elaboración propia

En la variable cuentas por cobrar se identifican un total de 15 datos por encima de $2.337.224.822, es decir, *outliers* extremos, lo que representa el 10% del total de los datos de la variable. Por su parte, en la referida variable se evidencian 19 datos por encima de

$1.480.159.127, es decir *outliers* leves, con un porcentaje de participación del 13% del total de los datos de la variable.

Tabla 5

Outliers variable inversiones

Inversiones						
Cuartil 1	Cuartil 3	Rango intercuartílico	Outliers extremos		Outliers leves	
Q1	Q3	IQR = Q3-Q1	(Q1 − 3IQR, Q3 + 3IQR)		(Q1 −1,5IQR, Q3 + 1,5IQR)	
284.020.998	4.730.380.005	4.446.359.007	- 13.055.056.023	18.069.457.027	- 6.385.517.513	11.399.918.516

Nota: Elaboración propia

En la variable inversiones se identifican un total de 11 datos por encima de $18.069.457.027, es decir, *outliers* extremos, lo que representa el 7% del total de los datos de la variable. Por su parte, en la referida variable se evidencian 15 datos por encima de $11.399.918.516, es decir *outliers* leves, con un porcentaje de participación del 10% del total de los datos de la variable.

Tabla 6

Outliers variable inventario

Inventario						
Cuartil 1	Cuartil 3	Rango intercuartílico	Outliers extremos		Outliers leves	
Q1	Q3	IQR = Q3-Q1	(Q1 − 3IQR, Q3 + 3IQR)		(Q1 −1,5IQR, Q3 + 1,5IQR)	
0	0	0	0	0	0	0

Nota: Elaboración propia

En la variable inventario se identifican un total de 17 datos por encima de $0, es decir, *outliers* extremos, lo que representa el 12% del total de los datos de la variable. La misma situación se presenta para los *outliers* leves.

Tabla 7

Outliers variable propiedad, planta y equipo

Propiedad, planta y equipo						
Cuartil 1	Cuartil 3	Rango intercuartílico	Outliers extremos		Outliers leves	
Q1	Q3	IQR = Q3-Q1	(Q1 − 3IQR, Q3 + 3IQR)		(Q1 −1,5IQR, Q3 + 1,5IQR)	
859.263.879	6.282.620.915	5.423.357.037	-15.410.807.231	22.552.692.025	-7.275.771.676	14.417.656.470

Nota: Elaboración propia

En la variable propiedad, planta y equipo se identifican un total de 6 datos por encima de $22.552.692.025, es decir, *outliers* extremos, lo que representa el 4% del total de los datos de la variable. Por su parte, en la referida variable se evidencian 17 datos por encima de $14.417.656.470, es decir *outliers* leves, con un porcentaje de participación del 12% del total de los datos de la variable.

Tabla 8

Outliers variable obligaciones

Obligaciones						
Cuartil 1	Cuartil 3	Rango intercuartílico	Outliers extremos		Outliers leves	
Q1	Q3	IQR = Q3-Q1	(Q1 − 3IQR, Q3 + 3IQR)		(Q1 −1,5IQR, Q3 + 1,5IQR)	
0	6.784.604.578	6.784.604.578	-20.353.813.733	27.138.418.311	-10.176.906.866	16.961.511.444

Nota: Elaboración propia

En la variable obligaciones se identifican un total de 11 datos por encima de $27.138.418.311, es decir, *outliers* extremos, lo que representa el 7% del total de los datos de la variable. Por su parte, en la referida variable se evidencian 20 datos por encima de $16.961.511.444, es decir *outliers* leves, con un porcentaje de participación del 14% del total de los datos de la variable.

Descartar los datos atípicos del conjunto de datos o sustituirlos por la media o la mediana no es recomendable, ya que esto puede modificar los resultados obtenidos, disminuir el tamaño

muestral, introducir un sesgo o puede afectar tanto a la distribución como a las varianzas de la variable de interés (Gobierno de España, s.f.).

6.2.1 Datos faltantes

Los datos faltantes se definen como valores no disponibles que serían útiles o significativos para el análisis de los resultados (Dagnino, 2014).

La base de datos objetivo que se ha elegido para el presente trabajo no presenta datos faltantes.

6.3 Transformación/reducción

6.3.1 Análisis exploratorio

La media es una medida de posición que brinda una descripción acerca de la forma en la que están centrados los datos (Castro, 2020).

La desviación estándar es una medida de dispersión. Un valor relativamente pequeño implica concentración de los datos alrededor de la media, un valor relativamente grande implica gran dispersión de los datos alrededor de la media (Mode, 2021).

El coeficiente de variación nos permite determinar la dispersión relativa entre los datos. A mayor coeficiente de variación tendremos mayor dispersión de los datos (Salazar, 2020).

Los cuartiles dividen en cuatro partes iguales al conjunto de datos, es decir que en cada cuartil se encuentra concentrado el 25% de las observaciones (Banegas, 2020).

Tabla 9

Análisis exploratorio

Variable	Media	Desviación estándar	Coeficiente de variación	Cuartil 1	Cuartil 2	Cuartil 3
Cuentas por cobrar	1.904.792.963	7.738.644.000	4	51.716.301	214.819.374	623.093.431
Ingreso	20.407.035.315	39.002.330.000	2	3.238.071.773	6.605.246.654	18.376.986.121
Inventario	82.302.628	703.428.500	9	0	0	0
Inversiones	8.154.443.179	30.396.870.000	4	284.020.998	1.524.765.848	4.730.380.005
Obligaciones	9.351.068.249	26.237.600.000	3	0	990.000.000	6.784.604.578
Propiedad, planta y equipo	6.733.233.100	17.455.780.000	3	859.263.879	2.015.814.681	6.282.620.915

Nota: Elaboración propia

Los datos de la variable cuentas por cobrar se concentran alrededor de la cifra de $1.904.792.963, al observar el valor de la desviación estándar ($7.738.644.000) y del coeficiente de variación (4) se observa que existe dispersión de los datos. Con relación a los cuartiles el 75% de los datos de la variable son menores o iguales a $623.093.431.

Las observaciones de la variable ingreso tienen un valor medio de $20.407.035.315, la desviación estándar ($39.002.330.000) y el coeficiente de variación (2) evidencian la existencia de dispersión de los datos, con respecto a los cuartiles que en el 75% son menores o iguales a $18.376.986.121.

Los valores de la variable inventario se agrupan alrededor de $82.302.628, la desviación estándar de $703.428.500 y el coeficiente de variación de 9 indica que la variable tiene una gran dispersión de los datos. Los cuartiles indican que el 75% de los datos de la variable son menores o iguales a $0.

La variable obligaciones tiene un valor medio de $9.351.068.249, la desviación estándar ($26.237.600.000) y el coeficiente de variación (3) denotan dispersión de los datos. Los cuartiles indican que el 75% de los valores de la variable son menores o iguales a $6.784.604.578.

Los datos de la variable propiedad, planta y equipo se concentran alrededor de la cifra de $1.904.792.963; al observar el valor de la desviación estándar ($6.733.233.100) y del coeficiente de variación (3) se observa que existe dispersión de los datos; con relación a los cuartiles el 75% de los datos de la variable son menores o iguales a $6.282.620.915.

Correlación de variables

El coeficiente de correlación (ρxy): indica si dos variables pueden estar linealmente asociadas. El coeficiente de correlación puede tener valores entre -1 y + 1. Cuando el coeficiente de correlación es positivo, significa que al crecer una variable la otra también crece. Por el contrario, cuando el coeficiente de correlación es negativo, significa que al disminuir una variable la otra disminuye. Si el coeficiente de correlación es igual a -1 o + 1, entonces las variables están relacionadas. Si el coeficiente de correlación es igual a 0, las variables no están relacionadas. Si el coeficiente de correlación se encuentra entre $-1 < \rho xy < 0$ y $0 < \rho xy < 1$, entonces la correlación es parcial (Boselli, 2019).

Tabla 10

Matriz de correlación

	Cuentas por cobrar	Ingreso	Inventario	Inversiones	Obligaciones	Propiedad, planta y equipo
Cuentas por cobrar	1,0000	0,6088	0,0653	0,8076	0,4285	0,1941
Ingreso	0,6088	1,0000	0,3011	0,6110	0,6221	0,4786
Inventario	0,0653	0,3011	1,0000	-0,0105	0,1711	0,3930
Inversiones	0,8076	0,6110	-0,0105	1,0000	0,4492	0,1055
Obligaciones	0,4285	0,6221	0,1711	0,4492	1,0000	0,7162
Propiedad, planta y equipo	0,1941	0,4786	0,3930	0,1055	0,7162	1,0000

Nota: Elaboración propia

La tabla nos muestra que todas las variables tienen correlaciones parciales positivas, excepto las variables inventario e inversiones las cuales tienen una correlación parcial negativa.

Prueba de normalidad

Existen distintos tipos de pruebas estadísticas para comprobar si un conjunto de datos proviene de una distribución normal. La prueba de normalidad (Kolmogorov-Smirnov) es la más utilizada, y esta se emplea cuando hay más de 50 observaciones (López et al., 2021).

La prueba Kolmogorov-Smirnov da como resultado una probabilidad (valor p). La hipótesis nula es la normalidad. Por lo tanto, si el valor $p > 0,05$ no hay evidencias para rechazar la hipótesis nula y se podría asumir normalidad. Cuando el valor $p < 0,05$ no se asume normalidad. De todos modos, cuando la muestra es grande ($n > 60$), con frecuencia se puede asumir normalidad, aunque las pruebas arrojen un valor inferior al 0,05, ya que los estimadores calculados en muestras grandes, según se deriva del teorema de límite central, se aproximan a la distribución normal (González et al., 2020).

Tabla 11

Valor p de las variables

Variable	Valor p
Cuentas por cobrar	0,00000000000000022
Ingreso	0,00000000000000022
Inventario	0,00000000000000022
Inversiones	0,00000000000000022
Obligaciones	0,00000000000000022
Propiedad, planta y equipo	0,00000000000000022

Nota: Elaboración propia

La tabla nos muestra que el valor p de todas las variables es inferior a 0.05, lo que indica que las variables no provienen de una distribución normal; no obstante, tal y como se anotó, dado que la muestra de cada variable es grande se puede asumir normalidad a partir del teorema del límite central.

Transformación de los datos

La transformación de datos es el proceso técnico de convertir datos de un formato, estándar o estructura, a otro, sin cambiar el contenido de los conjuntos de datos, generalmente para prepararlos para el consumo de una aplicación o un usuario, o para mejorar la calidad de los datos. Para el desarrollo de este trabajo se utilizará la discretización, la cual consiste en tomar funciones o variables continuas y transformarlas en funciones o variables discretas, respectivamente (Linkedin.com, 2023).

Tabla 12

Discretización de variables

Límite inferior	Límite superior	Categoría
$0	$1.000.000	1
$1.000.001	$10.000.000	2
$10.000.001	$100.000.000	3
$100.000.001	$1.000.000.000	4
$1.000.000.001	$10.000.000.000	5
$10.000.000.001	$100.000.000.000	6
$100.000.000.001	$1.000.000.000.000	7
$1.000.000.000.001	$10.000.000.000.000	8

Nota: Elaboración propia

6.3.2 Análisis descriptivo – Clustering

El análisis *cluster* es un conjunto de técnicas multivariantes utilizadas para clasificar a un conjunto de individuos en grupos homogéneos (www.uv.es, s.f.).

Existen dos grandes tipos de análisis de *clusters*: no jerárquicos y jerárquicos (De la fuente, s.f.).

Para el desarrollo de este trabajo se utilizará el análisis no jerárquico o de K-medias, debido a que este suele utilizarse para conjuntos de elementos muy numerosos (González, et al., 2013).

Los métodos no jerárquicos categorizan los elementos según un número de *cluster* dado (Calvo, 2018).

Para este caso se tomaron las siguientes variables discretizadas (C_Ingreso, C_Cuentas_por_cobrar, C_Inversiones, C_Inventario, C_Propiedad_planta_equipo y C_Obligaciones) y se asignó un número de cuatro *clusteres*. De acuerdo con el análisis se obtuvieron los siguientes resultados.

Tabla 13

Tamaño de los segmentos

Segmento	Total de clientes por segmento	Porcentaje de participación
1	35	24%
2	55	37%
3	12	8%
4	45	31%
Total	147	100%

Nota: Elaboración propia

La tabla nos muestra que en el segmento 1 se agruparon 35 entidades representando el 24% de los clientes, en el segmento 2 se agruparon 55 entidades con un 37% de participación, en el segmento 3 se agruparon 12 entidades lo que representa el 8% de participación, finalmente en el segmento 4 se agruparon 45 entidades representando el 31% de los clientes.

Tabla 14

Ingreso por segmento

Segmento	C_Ingreso	Límite inferior	Límite superior	Número de clientes	Promedio de ingreso
1	5	$ 1.000.000.001	$ 10.000.000.000	1	$ 7.021.613.813
	6	$ 10.000.000.001	$ 100.000.000.000	27	$ 37.596.722.337
	7	$ 100.000.000.001	$ 1.000.000.000.000	7	$ 169.191.194.170
2	4	$ 100.000.001	$ 1.000.000.000	3	$ 797.111.425
	5	$ 1.000.000.001	$ 10.000.000.000	45	$ 4.726.808.882
	6	$ 10.000.000.001	$ 100.000.000.000	7	$ 20.108.195.846
3	4	$ 100.000.001	$ 1.000.000.000	4	$ 792.099.239
	5	$ 1.000.000.001	$ 10.000.000.000	8	$ 2.595.139.703
4	5	$ 1.000.000.001	$ 10.000.000.000	29	$ 4.495.027.933
	6	$ 10.000.000.001	$ 100.000.000.000	16	$ 17.701.392.853

Nota: Elaboración propia

De acuerdo con la tabla se observa que en el segmento 1 veintisiete (27) clientes se ubican en la categoría 6, es decir que su ingreso se encuentra entre $10.000.000.001 y $100.000.000.000; el promedio de ingreso de los referidos clientes es de $37.596.722.337.

En el segmento 2 la mayoría de los clientes se concentran en la categoría 5, es decir que el ingreso de los referidos clientes se encuentra en el intervalo de $1.000.000.001 a $10.000.000.000, con un promedio de ingreso de $4.726.808.882.

En el segmento 3 ocho (8) clientes tienen un ingreso que se encuentra entre $1.000.000.001 y $10.000.000.000, es decir que pertenecen a la categoría 5, con un ingreso promedio de $2.595.139.703.

Con respecto al segmento 4 se observa que veintinueve (29) clientes se encuentra en el intervalo 5, es decir que el ingreso de los referidos clientes se encuentra entre $1.000.000.001 y $10.000.000.000. Para estos clientes el valor promedio de los ingresos es de $4.495.027.933.

Tabla 15

Cuentas por cobrar por segmento

Segmento	C_Cuentas_por_cobrar	Límite inferior	Límite superior	Número de clientes	Promedio de cuentas por cobrar
1	3	$ 10.000.001	$ 100.000.000	1	$ 36.690.094
	4	$ 100.000.001	$ 1.000.000.000	16	$ 509.820.573
	5	$ 1.000.000.001	$ 10.000.000.000	12	$ 2.781.344.615
	6	$ 10.000.000.001	$ 100.000.000.000	6	$ 33.089.819.515
2	2	$ 1.000.001	$ 10.000.000	5	$ 4.549.015
	3	$ 10.000.001	$ 100.000.000	19	$ 48.174.657
	4	$ 100.000.001	$ 1.000.000.000	28	$ 361.507.230
	5	$ 1.000.000.001	$ 10.000.000.000	3	$ 1.266.812.669
3	1	$ 0	$ 1.000.000	1	$ 369.180
	2	$ 1.000.001	$ 10.000.000	6	$ 5.127.976
	3	$ 10.000.001	$ 100.000.000	5	$ 44.910.970
4	3	$ 10.000.001	$ 100.000.000	14	$ 41.688.947
	4	$ 100.000.001	$ 1.000.000.000	26	$ 309.826.980
	5	$ 1.000.000.001	$ 10.000.000.000	5	$ 3.228.030.247

Nota: Elaboración propia

De acuerdo con la tabla se observa que en el segmento 1 dieciséis (16) clientes se ubican en la categoría 4, es decir que el valor de las cuentas por cobrar está entre $100.000.001 y $1.000.000.000; el valor promedio de las cuentas por cobrar de estos clientes es de $509.820.573.

En el segmento 2 la mayor parte de los clientes se encuentra en la categoría 4, lo que significa que el valor de las cuentas por cobrar de los referidos clientes está entre $100.000.001 y $1.000.000.000, con un promedio de $361.507.230.

Con relación al segmento 3 para seis (6) clientes el valor de las cuentas por pagar se encuentra en la categoría 2, es decir que está entre $1.000.001 y $10.000.000, con un valor promedio de $5.127.976.

Con respecto al segmento 4 se observa que el valor de las cuentas por cobrar para veintiséis (26) clientes se ubica en la categoría 4, lo cual indica que el valor de las cuentas por

pagar de los referidos clientes se encuentra entre $100.000.001 y $1.000.000.000. Para el referido número de clientes el valor promedio de las cuentas por cobrar es de $309.826.980.

Tabla 16

Inversiones por segmento

Segmento	C_Inversiones	Límite inferior	Límite superior	Número de clientes	Promedio de inversiones
1	4	$ 100.000.001	$ 1.000.000.000	2	$ 416.259.915
	5	$ 1.000.000.001	$ 10.000.000.000	19	$ 3.744.174.310
	6	$ 10.000.000.001	$ 100.000.000.000	12	$ 31.827.802.433
	7	$ 100.000.000.001	$ 1.000.000.000.000	2	$ 242.446.033.500
2	3	$ 10.000.001	$ 100.000.000	6	$ 64.742.786
	4	$ 100.000.001	$ 1.000.000.000	35	$ 409.320.727
	5	$ 1.000.000.001	$ 10.000.000.000	14	$ 2.622.626.616
3	3	$ 10.000.001	$ 100.000.000	4	$ 70.956.150
	4	$ 100.000.001	$ 1.000.000.000	6	$ 527.033.494
	5	$ 1.000.000.001	$ 10.000.000.000	2	$ 2.390.525.663
4	2	$ 1.000.001	$ 10.000.000	1	$ 7.737.962
	3	$ 10.000.001	$ 100.000.000	1	$ 59.667.938
	4	$ 100.000.001	$ 1.000.000.000	11	$ 371.949.014
	5	$ 1.000.000.001	$ 10.000.000.000	28	$ 4.663.811.284
	6	$ 10.000.000.001	$ 100.000.000.000	4	$ 16.375.381.681

Nota: Elaboración propia

La tabla muestra que en el segmento 1 para diecinueve (19) clientes el valor de las inversiones se ubica en la categoría 5 en el intervalo de $1.000.000.001 a $10.000.000.000; el promedio de las inversiones de los referidos clientes es de $3.744.174.310.

Con relación al segmento 2 para 35 clientes el valor de las inversiones se encuentra en la categoría 4, lo que significa que se encuentra en el intervalo de $100.000.001 a $1.000.000.000, con un promedio de $409.320.727.

La tabla también nos muestra que, en el segmento 3, las inversiones para un total de 6 clientes se ubican en la categoría 4, lo que indica que se encuentran en el intervalo de $100.000.001 a $1.000.000.000, con un promedio de $527.033.494.

En el segmento 4 el valor de las inversiones de 28 clientes se encuentra en el intervalo 5, es decir que se ubica entre $1.000.000.001 y $10.000.000.000. Para estos clientes el valor promedio de las inversiones es de $4.663.811.284.

Tabla 17

Inventario por segmento

Segmento	C_Inventario	Límite inferior	Límite superior	Número de clientes	Promedio de inventario
1	1	$ 0	$ 1.000.000	30	$ 0
	3	$ 10.000.001	$ 100.000.000	3	$ 59.314.179
	5	$ 1.000.000.001	$ 10.000.000.000	2	$ 5.464.334.611
2	1	$ 0	$ 1.000.000	51	$ 0
	2	$ 1.000.001	$ 10.000.000	3	$ 4.066.760
	3	$ 10.000.001	$ 100.000.000	1	$ 23.805.281
3	1	$ 0	$ 1.000.000	10	$ 0
	2	$ 1.000.001	$ 10.000.000	1	$ 1.205.100
	3	$ 10.000.001	$ 100.000.000	1	$ 11.788.205
4	1	$ 0	$ 1.000.000	39	$ 0
	2	$ 1.000.001	$ 10.000.000	1	$ 2.607.000
	3	$ 10.000.001	$ 100.000.000	2	$ 35.628.541
	4	$ 100.000.001	$ 1.000.000.000	3	$ 289.670.519

Nota: Elaboración propia

En la tabla se puede observar que, en los diferentes segmentos, la mayor parte de los clientes para el valor del inventario se encuentra en la categoría 1, es decir, en el intervalo de $0 a $1.000.000. El valor promedio para el inventario de los referidos clientes es $0.

Tabla 18

Propiedad, planta y equipo por segmento

Segmento	C_Propiedad_planta_equipo	Límite inferior	Límite superior	Número de clientes	Promedio de propiedad, planta y equipo
1	4	$ 100.000.001	$ 1.000.000.000	4	$ 378.977.995
	5	$ 1.000.000.001	$ 10.000.000.000	10	$ 5.939.283.857
	6	$ 10.000.000.001	$ 100.000.000.000	20	$ 21.144.326.878
	7	$ 100.000.000.001	$ 1.000.000.000.000	1	$ 187.811.581.335
2	3	$ 10.000.001	$ 100.000.000	4	$ 43.957.534
	4	$ 100.000.001	$ 1.000.000.000	11	$ 673.802.190
	5	$ 1.000.000.001	$ 10.000.000.000	36	$ 2.467.867.535
	6	$ 10.000.000.001	$ 100.000.000.000	4	$ 13.429.886.967
3	1	$ 0	$ 1.000.000	1	$ 0
	2	$ 1.000.001	$ 10.000.000	1	$ 5.074.201
	3	$ 10.000.001	$ 100.000.000	4	$ 38.873.454
	4	$ 100.000.001	$ 1.000.000.000	6	$ 494.029.524
4	4	$ 100.000.001	$ 1.000.000.000	11	$ 553.830.181
	5	$ 1.000.000.001	$ 10.000.000.000	30	$ 3.155.113.228
	6	$ 10.000.000.001	$ 100.000.000.000	4	$ 16.039.422.241

Nota: Elaboración propia

La tabla muestra que en el segmento 1 para veinte (20) clientes el valor de la variable propiedad, planta y equipo se ubica en la categoría 6 en el intervalo de $10.000.000.001 a $100.000.000.000; el promedio de la variable propiedad, planta y equipo para el referido número de clientes es de $21.144.326.878.

Para el segmento 2 se observan 36 clientes para los cuales el valor de la variable propiedad, planta y equipo se ubica en la categoría 5, es decir, que se encuentra entre $1.000.000.001 y $10.000.000.000; el promedio de la variable propiedad, planta y equipo para el referido número de clientes es de $2.467.867.535.

En el segmento 3 se observan 6 clientes para los que el valor de la variable propiedad, planta y equipo se sitúa en la categoría 4 en el intervalo de $100.000.001 a $1.000.000.000; el

promedio de la variable propiedad, planta y equipo para el referido número de clientes es de $494.029.524.

Con respecto al segmento 4 para un total de 30 clientes, el valor de la variable propiedad, planta y equipo se encuentra en la categoría 5, es decir que se encuentra entre $1.000.000.001 y $10.000.000.000; el promedio de la variable propiedad, planta y equipo para el referido número de clientes es de $3.155.113.228.

Tabla 19

Obligaciones por segmento

Segmento	C_Obligaciones	Límite inferior	Límite superior	Número de clientes	Promedio de obligaciones
1	5	$ 1.000.000.001	$ 10.000.000.000	9	$ 4.890.996.212
	6	$ 10.000.000.001	$ 100.000.000.000	23	$ 24.961.195.051
	7	$ 100.000.000.001	$ 1.000.000.000.000	3	$ 172.656.999.248
2	3	$ 10.000.001	$ 100.000.000	1	$ 26.757.843
	4	$ 100.000.001	$ 1.000.000.000	16	$ 405.063.818
	5	$ 1.000.000.001	$ 10.000.000.000	33	$ 4.008.225.515
	6	$ 10.000.000.001	$ 100.000.000.000	5	$ 19.915.994.345
3	1	$ 0	$ 1.000.000	11	$ 3.658
	3	$ 10.000.001	$ 100.000.000	1	$ 39.088.418
4	1	$ 0	$ 1.000.000	37	$ 97.743
	2	$ 1.000.001	$ 10.000.000	6	$ 5.655.309
	3	$ 10.000.001	$ 100.000.000	2	$ 36.856.559

Nota: Elaboración propia

La tabla muestra que en el segmento 1 para 23 clientes el valor de las obligaciones se ubica en la categoría 6 en el intervalo de $10.000.000.001 a $100.000.000.000, con un promedio de $24.961.195.051.

Para el segmento 2 se observan 33 clientes para los cuales el valor de las obligaciones se encuentra en la categoría 5, es decir que se encuentra entre $1.000.000.001 y $10.000.000.000, mostrando un promedio de $4.008.225.515.

En el segmento 3 se observan 11 clientes para los que el valor de la variable obligaciones se sitúa en la categoría 1 en el intervalo de $0 a $1.000.000; el promedio de las obligaciones para el referido número de clientes es de $3.658.

Con respecto al segmento 4 para un total de 37 clientes, el valor de las obligaciones se encuentra en la categoría 1 en el intervalo de $0 a $1.000.000; el promedio de las obligaciones para el referido número de clientes es de $97.743.

6.4 Etapa de minería de datos (*data mining*)

Como se definió líneas arriba, uno de los objetivos de este trabajo es utilizar técnicas de predicción para la segmentación del factor de riesgo cliente. En este caso se utilizará el análisis discriminante, el cual crea un modelo predictivo para la pertenencia al grupo. El modelo está compuesto por una función discriminante (o, para más de dos grupos, un conjunto de funciones discriminantes) basada en combinaciones lineales de las variables predictoras que proporcionan la mejor discriminación posible entre los grupos. Las funciones se generan a partir de una muestra de casos para los que se conoce el grupo de pertenencia; posteriormente, las funciones pueden ser aplicadas a nuevos casos que dispongan de mediciones para las variables predictoras, pero de los que se desconozca el grupo de pertenencia (IBM, 2021).

6.4.1 Modelo matemático del análisis discriminante

A partir de q grupos donde se asignan a una serie de objetos y de p variables medidas sobre ellos ($X_1,...,X_p$), se trata de obtener para cada objeto una serie de puntuaciones que indican el grupo al que pertenecen (Y1,...,Ym), de modo que sean funciones lineales de $X_1,...,Xp$

$$y_1 = a_{11}x_1 + \cdots + a_1pxp + a_{10}$$

……

$$y_m = a_{m1}x_1 + \cdots + a_mp_{xp} + am_0$$

donde m = mın(q − 1, p), tales que discriminen o separen lo máximo posible a los q grupos. Estas combinaciones lineales de las p variables deben maximizar la varianza entre los grupos y minimizar la varianza dentro de los grupos (Universidad Carlos III de Madrid, s.f.).

Supuestos de aplicación del análisis discriminante

Antes de la utilización de cualquier prueba estadística se debe comprobar el cumplimiento de los supuestos básicos de aplicación. En el caso que nos ocupa se pueden resumir en dos: (i) las variables independientes o predictivas deben seguir una distribución normal multivariante y (ii) las matrices de covarianzas deben ser iguales en todos los grupos. Con respecto al primer supuesto se había mencionado que las variables independientes provienen de una distribución normal, debido al teorema del límite central. Aunque el análisis discriminante es considerado una técnica robusta que no se ve gravemente afectada si alguno de los supuestos anteriores no se cumple, es recomendable aplicar la prueba de M de Box para comprobar el segundo supuesto. La prueba de M de Box parte del supuesto de que las matrices de covarianzas son iguales y se basa en el cálculo de los determinantes de covarianza de cada grupo; el valor obtenido se aproxima a la F de Snedecor (Torrado & Berlanga, 2013).

La hipótesis nula de la prueba M de Box es que las matrices de covarianza observadas para las variables dependientes son iguales en todos los grupos. En otras palabras, un resultado de prueba no significativo (es decir, uno con un valor p grande) indica que las matrices de covarianza son iguales (Statologos, 2021).

Tabla 20

Resultados de prueba M de Box

M de Box		36,930
F	Aprox.	0,740
	gl1	45
	gl2	6645,994
	Valor p.	0,901

Nota: Elaboración propia

De acuerdo con los resultados de la prueba se puede determinar el cumplimiento del segundo supuesto del análisis discriminante, ya que el valor p es mayor que 0,05.

Estimación de la variabilidad intergrupo explicada en la función discriminante

El autovalor es el cociente entre la variación debida a las discrepancias entre los grupos, y la variación que se da dentro de cada grupo combinada en una única cantidad (Cisneros et al., 2006).

Cuanto más alto es su valor, más eficaz será el análisis para clasificar a los sujetos. El valor mínimo es cero y no tiene un valor máximo (Torrado & Berlanga, 2013).

La correlación canónica recoge la pertenencia de los sujetos a los grupos mediante un coeficiente que oscila entre 0 y 1. Interesa que presente un valor lo más próximo a 1 (Torrado & Berlanga, 2013).

Tabla 21

Resultado de los autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	12,417[a]	87,5	87,5	0,962
2	1,604[a]	11,3	98,8	0,785
3	,170[a]	1,2	100,0	0,382
a. Se utilizaron las primeras 3 funciones discriminantes canónicas en el análisis.				

Nota: Elaboración propia

La primera función tiene un autovalor de 12,417 y explica el 87,5% de la variabilidad disponible en los datos; su correlación canónica es de 0,962 lo cual indica que las variables discriminantes permiten distinguir bastante bien entre los grupos. Mientras que la segunda y la tercera función explican el 11,3% y el 1,2% de la variabilidad. De acuerdo con lo anterior, se pueden descartar la segunda y tercera función discriminante, sin afectar de manera significativa los resultados.

Diferencia entre los grupos (lamda de Wilks)

Los valores de lambda de Wilks cerca de cero denotan alta discriminación. Esto quiere decir que los centroides están muy separados entre sí. A medida que lambda de Wilks se acerca a uno, el poder discriminatorio de la función se hace más débil (Campoverde, 2017).

Tabla 22

Resultado de lambda de Wilks

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Valor p
1 a 3	0,024	525,111	15	0,000
2 a 3	0,328	157,704	8	0,000
3	0,854	22,267	3	0,000

Nota: Elaboración propia

La tabla nos muestra que las dos primeras funciones tienen un alto poder discriminatorio, puesto que el valor lambda de Wilks es cercano a cero.

Construcción de la función de clasificación

El programa en que se adelantó el análisis discriminante, para el desarrollo de este trabajo, utilizó el *software* SPSS que permite elegir las variables que realmente son útiles para la clasificación, para lo cual utiliza un método de inclusión de variables paso a paso.

El criterio usado para la inclusión de las variables corresponde a usar el valor de F, en el cual una variable pasa a formar parte de la función discriminante si el valor del estadístico F es mayor que 3,84 (valor de entrada), y es expulsada de la función si el valor del estadístico F es menor que 2,71 (valor de salida) (Campoverde, 2017).

Tabla 23

Inclusión/exclusión de variables

Paso	Variable	F para entrar
0	C_Ingreso	53,423
	C_Cuentas_por_cobrar	36,059
	C_Inversiones	30,192
	C_Inventario	1,382
	C_Propiedad_planta_equipo	32,046
	C_Obligaciones	565,522
1	C_Ingreso	46,01
	C_Cuentas_por_cobrar	27,938
	C_Inversiones	27,906
	C_Inventario	1,412
	C_Propiedad_planta_equipo	22,212
2	C_Cuentas_por_cobrar	9,882
	C_Inversiones	7,309
	C_Inventario	1,189
	C_Propiedad_planta_equipo	10,371
3	C_Cuentas_por_cobrar	7,487
	C_Inversiones	8,265
	C_Inventario	1,168
4	C_Cuentas_por_cobrar	5,65
	C_Inventario	1,116
5	C_Inventario	0,9

Nota: Elaboración propia

En la tabla se puede observar que la variable C_Inventario presentó un valor inferior 2,71, por lo tanto, la variable fue excluida del análisis.

Tabla 24

Coeficientes de función de clasificación

Variable	Número de caso de cluster			
	1	2	3	4
C_Ingreso	18,231	15,292	19,526	20,291
C_Cuentas_por_cobrar	1,763	0,832	0,377	2,633
C_Inversiones	8,398	6,324	5,031	6,687
C_Propiedad_planta_equipo	6,983	6,147	3,2	6,134
C_Obligaciones	11,549	9,478	-1,633	-2,641
(Constante)	-137,4	-91,45	-62,4	-89,59

Nota: Elaboración propia

De la tabla anterior extraemos las funciones de clasificación.

Ecuaciones y selección de casos

Segmento 1 *= -137,380 + 18,231*C_Ingreso + 1,763*C_Cuentas_por_cobrar + 8,398*C_Inversiones + 6,983*C_Propiedad_planta_equipo + 11,549*C_Obligaciones*

Segmento 2 *= -91,448 + 15,292*C_Ingreso + 0,832*C_Cuentas_por_cobrar + 6,324*C_Inversiones + 6,147*C_Propiedad_planta_equipo + 9,478*C_Obligaciones*

Segmento 3 *= -62,395 + 19,526*C_Ingreso + 0,377*C_Cuentas_por_cobrar + 5,031*C_Inversiones + 3,200*C_Propiedad_planta_equipo - 1,633*C_Obligaciones*

Segmento 4 *= -89,591 + 20,291*C_Ingreso + 2,633*C_Cuentas_por_cobrar + 6,687*C_Inversiones + 6,134*C_Propiedad_planta_equipo - 2,641*C_Obligaciones*

La manera de clasificar a cada cliente en uno de los segmentos será a partir de la puntuación total más alta al evaluar las variables del cliente en cada una de las ecuaciones. (YouTube, 2020).

Tabla 25

Resultado de la clasificación

Sigla	C_Ingreso	C_Cuentas_por_cobrar	C_Inversiones	C_Propiedad_planta_equipo	C_Obligaciones	Segmento original	Segmento pronosticado
COOPEAIPE	5	2	4	5	5	2	2
COFINAL_LTDA	6	4	5	5	6	1	1
COODIN	5	3	4	5	5	2	2
Cooperativa_AVP	5	3	3	4	3	2	3
FECEDA	5	3	5	5	6	2	2
COFINCAFE	6	4	5	6	6	1	1
CODELCAUCA	5	4	4	5	5	2	2
BADIVENCOOP	5	4	4	5	4	2	2
COOPINTEGRATE	5	2	4	5	4	2	2
FODUN	6	4	5	5	1	4	4
COINPROGUA	5	2	4	4	4	2	2
COOPEMSURA	5	4	5	5	1	4	4
COOHEM	5	3	4	5	1	4	4
COOPIGON	5	3	4	5	4	2	2
COOTREGUA	5	4	4	5	5	2	2
FOTRANORTE	5	4	5	4	5	2	2
COOTEP_LTDA	5	4	4	5	4	2	2
COFACENEIVA	5	4	4	5	1	4	4
COOPSOCIAL	5	3	5	5	4	2	2
COOTRAUNION	5	3	5	4	1	4	4
COOCREAFAM	6	4	6	5	5	1	1
COOMADENORT	4	3	3	5	5	2	2
COOMSERVI	5	5	4	5	4	2	2
COOMULTAGRO	5	3	5	5	4	2	2

Sigla	C_Ingreso	C_Cuentas_por_cobrar	C_Inversiones	C_Propiedad_planta_equipo	C_Obligaciones	Segmento original	Segmento pronosticado
FINANCIERA_COMULTRASAN	7	6	6	6	6	1	1
CFA	7	6	6	6	6	1	1
PREVENSERVICIOS	5	3	5	5	1	4	4
AMAR	5	4	5	5	5	2	2
VIDASOL	4	3	4	3	1	3	3
MANUELITACOOP	5	4	4	5	3	2	2
COMERCIACOOP	5	4	4	4	1	4	4
FINANCIERA_PROGRESSA	6	4	6	4	6	1	1
FONTIGO	4	1	4	1	1	3	3
FESICOL	5	4	5	5	5	2	2
CANAPRO	6	6	5	7	7	1	1
FINCOMERCIO_LTDA	7	4	6	6	7	1	1
FONDEXXOM	5	4	5	4	1	4	4
FEBOR	6	4	5	5	1	4	4
ALCALICOOP	5	4	4	5	2	4	4
CREDICOOP	6	4	5	5	5	1	1
COOPETROL	6	4	5	6	6	1	1
COASMEDAS	6	4	6	6	5	1	1
COOPTRAISS	6	5	6	6	5	1	1
COOVITEL	6	5	5	5	6	1	1
COOACUEDUCTO	6	4	5	5	1	4	4
COLOMBIACOOP	5	3	3	5	5	2	2
COOTRAPELDAR	6	4	5	5	1	4	4
ALIANZA	6	5	4	5	5	2	2
SOMEC	5	4	4	6	5	2	2
Cooperativa_de_Profesores_de_la_Universidad_Nacional_de_Colombia	6	5	6	6	1	4	4
FEDANE	4	3	5	2	1	3	3
COOPTENJO	6	4	4	6	5	2	2
COOPSANFRANCISCO	5	2	5	4	1	4	3
COOPCAFAM	6	5	5	5	3	4	4
COOPCHIPAQUE	5	3	5	5	2	4	4
FEMPHA	5	3	5	4	1	4	4
COOINDEGABO	5	4	4	5	1	4	4
Cooperativa_de_ahorro_y_crédito_para_el_bienestar_social_Beneficiar_entidad_cooperativa	6	4	5	5	2	4	4

Sigla	C_Ingreso	C_Cuentas_por_cobrar	C_Inversiones	C_Propiedad_planta_equipo	C_Obligaciones	Segmento original	Segmento pronosticado
FECOLSA	6	5	6	5	1	4	4
DEMCOOP	5	4	3	5	6	2	2
CORPECOL	6	5	5	5	6	1	1
AVANZA	6	4	4	5	6	1	2
FONDECOR	6	5	5	4	6	1	1
COPVILLANUEVA	5	3	5	5	4	2	2
MULTICOOP	5	3	4	4	5	2	2
COOMULTRASAN	7	5	5	6	5	1	1
COOPROFESORES	6	4	6	5	5	1	1
COOPVALLE	5	4	5	4	1	4	4
COPACREDITO	5	4	4	5	5	2	2
COOPCENTRAL	7	6	7	6	7	1	1
COOMULDESA	6	4	5	6	6	1	1
SERVIMCOOP	6	4	4	5	5	2	2
COOPROFESIONALES	5	4	4	5	1	4	4
COOPSERVIVELEZ_LTDA	6	4	4	5	1	4	4
SERVICONAL	5	2	3	4	1	3	3
COMULSEB	5	3	4	5	5	2	2
COOMBEL	5	2	3	3	1	3	3
COOSANANDRESITO	5	4	5	6	1	4	4
COESCOOP	5	3	4	3	4	2	2
COAPAZ	4	2	4	4	1	3	3
FINANCIERA_COAGROSUR	6	4	4	5	5	2	2
COOPCARVAJAL	5	4	5	5	2	4	4
COOMPARTIR	4	3	3	5	4	2	2
COOTRAEMCALI	5	4	4	6	5	2	2
GRANCOOP	5	4	5	4	1	4	4
SOLIDARIOS	5	2	4	5	5	2	2
COOFIPOPULAR	6	4	5	4	6	1	1
FODEBAX	5	4	3	4	5	2	2
CRECIAT	5	4	5	3	5	2	2
FONAVIEMCALI	5	5	4	6	5	2	1
FETRABUV	5	4	5	4	1	4	4
FEDIAN	5	3	5	4	2	4	4

Sigla	C_Ingreso	C_Cuentas_por_cobrar	C_Inversiones	C_Propiedad_planta_equipo	C_Obligaciones	Segmento original	Segmento pronosticado
FONDECOM	5	5	4	4	1	4	4
FONRECAR	5	4	4	4	5	2	2
COOMUTRANORT_LTDA	5	3	5	5	1	4	4
CREDISERVIR	7	4	5	6	6	1	1
FOMANORT	5	4	5	6	5	2	2
PROSPERANDO	6	4	5	5	1	4	4
COOPEMTOL	6	4	4	5	1	4	4
COOFINANCIAR	5	2	4	3	1	3	3
COOMULTRAISS	4	3	3	4	4	2	2
CIDECAL	5	5	4	5	5	2	2
FONCALDAS	5	4	2	5	1	4	4
FONFABRICAFE	5	3	4	4	4	2	2
COODESS	5	3	3	5	1	4	4
CESCA	6	3	5	6	1	4	4
COOPROCAL	5	2	4	4	4	2	2
COOPANTEX	6	4	5	6	6	1	1
COMEDAL	6	5	5	6	6	1	1
COOTRASENA	6	3	5	5	1	4	4
Cooperativa_Universitaria_Bolivariana	5	4	5	4	1	4	4
COOPACREDITO	6	3	6	5	1	4	4
COBELEN	6	4	6	5	6	1	1
COOPECREDITO_ENTRERRIOS	5	3	5	5	1	4	4
COOPSUYA	5	4	5	5	1	4	4
COOSANLUIS	5	3	4	5	5	2	2
CREARCOOP	6	3	4	5	6	2	2
COOGRANADA	6	4	4	6	6	1	1
COOPEREN	5	4	5	5	1	4	4
FODELSA	5	3	4	5	1	4	4
COOSERVUNAL	5	4	5	5	5	2	2
COOPRUDEA	6	5	6	4	3	4	4
COOGOMEZPLATA	5	3	5	4	4	2	2
COOFISAM	6	4	6	5	6	1	1
COONFIE	6	5	5	6	6	1	1
UTRAHUILCA	6	5	5	6	5	1	1

Sigla	C_Ingreso	C_Cuentas_por_cobrar	C_Inversiones	C_Propiedad_planta_equipo	C_Obligaciones	Segmento original	Segmento pronosticado
CREDIFUTURO	5	3	5	4	5	2	2
FONEDH	5	4	5	5	1	4	4
COACREMAT_LTDA	6	4	4	5	5	2	2
COOPLAROSA	5	4	5	5	5	2	2
COTRASENA	5	3	4	4	5	2	2
COODELMAR	5	4	4	5	5	2	2
FAVI_UTP	5	3	4	4	1	4	3
COEDUCADORES_BOYACA	6	5	6	6	5	1	1
COOMEC	5	4	4	5	6	2	2
Coprocenva_Cooperativa_De_Ahorro_y_Crédito	6	5	6	5	6	1	1
CONGENTE	6	3	5	6	6	1	1
COOPICREDITO	6	3	4	6	2	4	4
A_Y_C_COLANTA	6	6	5	5	6	1	1
UNION_COOPERATIVA	5	4	4	3	5	2	2
MUTUAL_COOTRADECUN	5	4	5	5	1	4	4
COOPCANAPRO	5	4	5	5	4	2	2
AFROAMERICANA	5	2	4	3	1	3	3
Financiera_Juriscoop_C_F	7	6	7	6	6	1	1
SUCREDITO	6	5	5	4	6	1	1
CREDIAHORROS_TAX_FERIA	5	3	3	4	1	3	3
FINANCIAFONDOS_O_C	5	4	4	3	5	2	2

Nota: Elaboración propia

De la anterior tabla podemos definir que los siguientes clientes, después de aplicar el análisis discriminante, no fueron clasificados en su segmento original.

Tabla 26

Clientes clasificados en segmento diferente al original

Sigla	Segmento original	Segmento pronosticado
AVANZA	1	2
Cooperativa_AVP	2	3
FONAVIEMCALI	2	1
COOPSANFRANCISCO	4	3
FAVI_UTP	4	3

Nota: Elaboración propia

El resto de los clientes (142) fueron clasificados en su segmento original, lo que significa que el 96,6% de los casos agrupados originales se clasificaron correctamente.

6.5 Etapa de interpretación/evaluación

Tabla 27

Tamaño de los segmentos tras aplicar el análisis discriminante

Segmento	Total de clientes por segmento	Porcentaje de participación
1	35	24%
2	56	38%
3	9	6%
4	47	32%
Total	147	100%

Nota: Elaboración propia

La tabla muestra que en el segmento 1 se agruparon 35 entidades representando el 24% de los clientes, en el segmento 2 se agruparon 56 entidades con un 38% de participación, en el segmento 3 se agruparon 9 entidades lo que representa el 6% de participación, finalmente en el segmento 4 se agruparon 47 entidades representando el 32% de los clientes.

Tabla 28

Ingresos por segmento tras aplicar el análisis discriminante

Segmento	Total ingresos por segmento	Valor máximo de los ingresos por segmento	Tercer cuartil de los ingresos por segmento	Promedio de los ingresos por segmento
1	$ 2.217.898.191.718	$ 281.419.720.501	$ 68.794.908.188	$ 63.368.519.763
2	$ 346.729.611.206	$ 36.648.942.444	$ 7.131.133.336	$ 6.191.600.200
3	$ 17.768.205.617	$ 4.760.877.719	$ 1.829.102.994	$ 1.974.245.069
4	$ 417.438.182.735	$ 29.976.366.177	$ 14.308.518.369	$ 8.881.663.462

Nota: Elaboración propia

La tabla muestra que el segmento 3 presenta el menor valor del total de los ingresos; el mayor valor de los ingresos se encuentra en el segmento 1 y pertenece a la entidad FINANCIERA COMULTRASAN.

Tabla 29

Cuentas por cobrar por segmento tras aplicar el análisis discriminante

Segmento	Total cuentas por cobrar por segmento	Valor máximo de las cuentas por cobrar por segmento	Tercer cuartil de las cuentas por cobrar por segmento	Promedio de las cuentas por cobrar por segmento
1	$ 239.787.624.590	$ 71.686.759.460	$ 3.616.818.341	$ 6.851.074.988
2	$ 15.205.813.893	$ 1.604.922.471	$ 346.885.764	$ 271.532.391
3	$ 147.085.175	$ 82.632.180	$ 11.336.923	$ 16.342.797
4	$ 24.864.041.952	$ 8.267.006.276	$ 344.791.514	$ 529.022.169

Nota: Elaboración propia

En la tabla se puede observar que en el segmento 1 se presenta el mayor valor del total de las cuentas por cobrar, adicionalmente, el mayor valor de las cuentas por cobrar se encuentra en el segmento 1 y pertenece a la entidad FINANCIERA JURISCOOP.

Tabla 30

Inversiones por segmento tras aplicar el análisis discriminante

Segmento	Total inversiones por segmento	Valor máximo de las inversiones por segmento	Tercer cuartil de las inversiones por segmento	Promedio de las inversiones por segmento
1	$ 939.396.212.635	$ 295.463.559.000	$ 23.433.496.916	$ 26.839.891.790
2	$ 50.896.516.377	$ 5.208.696.870	$ 999.410.549	$ 908.866.364
3	$ 5.513.792.159	$ 2.343.775.386	$ 762.949.272	$ 612.643.573
4	$ 202.896.626.141	$ 28.695.691.726	$ 6.118.312.119	$ 4.316.949.492

Nota: Elaboración propia

En la tabla se observa que los segmentos 1 y 4 presentan el mayor valor total de las inversiones, seguidos de los segmentos 2 y 3. El mayor valor de las inversiones se encuentra en el segmento 1 y pertenece a la entidad FINANCIERA JURISCOOP.

Tabla 31

Propiedad, planta y equipo por segmento tras aplicar el análisis discriminante

Segmento	Total propiedad, planta y equipo por segmento	Valor máximo de propiedad, planta y equipo por segmento	Tercer cuartil de propiedad, planta y equipo por segmento	Promedio de propiedad, planta y equipo por segmento
1	$ 665.869.688.751	$ 187.811.581.335	$ 17.529.552.179	$ 19.024.848.250
2	$ 156.432.469.525	$ 17.282.426.117	$ 2.754.432.920	$ 2.793.436.956
3	$ 1.681.198.631	$ 930.393.145	$ 193.555.454	$ 186.799.848
4	$ 165.801.908.857	$ 19.311.586.986	$ 4.154.618.120	$ 3.527.700.188

Nota: Elaboración propia

La tabla muestra que los segmentos 1 y 4 presentan el mayor valor total de propiedad, planta y equipo, seguidos de los segmentos 2 y 3. El mayor valor de la propiedad, planta y equipo se encuentra en el segmento 1 y pertenece a la entidad COOPERATIVA CASA NACIONAL DEL PROFESOR.

Tabla 32

Obligaciones por segmento tras aplicar el análisis discriminante

Segmento	Total obligaciones por segmento	Valor máximo de obligaciones por segmento	Tercer cuartil de las obligaciones por segmento	Promedio de obligaciones por segmento
1	$ 1.157.065.073.606	$ 181.394.046.804	$ 27.869.655.031	$ 33.059.002.103
2	$ 217.430.657.290	$ 37.750.136.873	$ 5.987.638.591	$ 3.882.690.309
3	$ 40.240	$ 40.240	$ 0	$ 4.471
4	$ 111.261.460	$ 42.585.145	$ 419.518	$ 2.367.265

Nota: Elaboración propia

7. Señales de alerta

De acuerdo con la normatividad vigente en materia de SARLAFT se establece que las señales de alerta o alertas tempranas son los hechos, situaciones, eventos, cuantías, indicadores cuantitativos y cualitativos, razones financieras y demás información que la entidad determine como relevante, a partir de los cuales se puede inferir oportuna y/o prospectivamente la posible existencia de un hecho o situación que escapa a lo que la entidad, en el desarrollo del SARLAFT, ha determinado como normal (Superintendencia Financiera de Colombia, 2022).

De acuerdo con lo anterior, y conforme a los resultados obtenidos de la segmentación bajo el análisis discriminante, se tomará como punto de partida el valor asociado al tercer cuartil de las variables analizadas en cada segmento. El tercer cuartil es aquel valor numérico tal que al menos el 75% de las observaciones son menores o iguales que aquel, y al menos el 25%, más grandes o iguales (Sancho, s.f.).

En este sentido todos los valores o cuantías que sean mayores al valor del tercer cuartil de la variable se configurarán como una señal de alerta.

Tabla 33

Señales de alerta por segmento

Segmento	Variable	Señal de alerta
1	Cuentas por cobrar	Valores mayores a $ 3.616.818.341
	Ingreso	Valores mayores a $ 68.794.908.188
	Inversiones	Valores mayores a $ 23.433.496.916
	Obligaciones	Valores mayores a $ 27.869.655.031
	Propiedad, planta y equipo	Valores mayores a $ 17.529.552.179
2	Cuentas por cobrar	Valores mayores a $ 346.885.764
	Ingreso	Valores mayores a $ 7.131.133.336
	Inversiones	Valores mayores a $ 999.410.549
	Obligaciones	Valores mayores a $ 5.987.638.591
	Propiedad, planta y equipo	Valores mayores a $ 2.754.432.920

Segmento	Variable	Señal de alerta
3	Cuentas por cobrar	Valores mayores a $ 11.336.923
	Ingreso	Valores mayores a $ 1.829.102.994
	Inversiones	Valores mayores a $ 762.949.272
	Obligaciones	Valores mayores a $ 0
	Propiedad, planta y equipo	Valores mayores a $ 193.555.454
4	Cuentas por cobrar	Valores mayores a $ 344.791.514
	Ingreso	Valores mayores a $ 14.308.518.369
	Inversiones	Valores mayores a $ 6.118.312.119
	Obligaciones	Valores mayores a $ 419.518
	Propiedad, planta y equipo	Valores mayores a $ 4.154.618.119

Nota: Elaboración propia

Conclusiones

Las Entidades Administradoras de Sistemas de Pago de Bajo Valor tienen características particulares que las diferencian de otras entidades del sector financiero, sin embargo, se pudo establecer que es posible implementar un modelo de segmentación para el factor de riesgo cliente, de acuerdo con la normatividad vigente emitida por la Superintendencia Financiera de Colombia.

La metodología KDD proporciona un marco adecuado para la elaboración de un modelo de segmentación del factor de riesgo cliente en una EASPBV, puesto que permite abordar varias etapas de forma secuencial y organizada.

La aplicación de la técnica predictiva del análisis discriminante permitió una adecuada clasificación de los clientes en segmentos definidos; adicionalmente, con la referida técnica, se logró cumplir con los lineamientos normativos y técnicos relacionados con la segmentación del factor de riesgo cliente.

Es muy importante tener en cuenta que el proceso de segmentación de factores de riesgos es un proceso dinámico que debe estar en constante revisión y evolución, por lo tanto, es pertinente que las personas que se encuentran al frente de la administración del SARLAFT prueben diferentes metodologías y técnicas que permitan lograr modelos maduros y eficientes.

Referencias

Amaya Molina, M. (2017). Segmentación de clientes y definición de alertas para la prevención de riesgos de lavado de activos y financiación del terrorismo (SARLAFT): un estudio económico aplicado a entidad financiera colombiana en 2017 (Bachelor's thesis, Universidad EAFIT).

Antonio Aquino, A. (2016). Proceso de minería de datos centrado en el usuario con base en la norma ISO 9241-210: 2010 (Doctoral dissertation, Universidad Veracruzana. Facultad de Estadística e Informática. Región Xalapa).

Banegas, A. L. G. (2020). *Cómo entender estadística fácilmente* (Vol. 2). IMCP.

Bareño-Dueñas, S. M. (2009). Mecanismos de contabilidad para prevenir y detectar el lavado de activos en Colombia. *Cuadernos de contabilidad*, 10(27), 341-357.

Benites, L. (2021, octubre 16). Prueba M de Box: Definición. Statologos. https://statologos.com/prueba-de-cajas-m/

Boselli, P. M. (2019). *El método BFMNU: interpretación modelo-fenomenológica de la nutrición*. Mónsul Ediciones.

Burgos Vargas, J. S., & González Cubero, K. L. (2020). Implementación de un sistema de gestión basado en la metodología KDD de minería de datos para el proceso de planificación de la producción de la Industria Farmacéutica Ecuatoriana (Bachelor's thesis, Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas. Carrera de Ingeniería en Sistemas Computacionales).

Caballero, M. M. (2023, agosto 9). Comparación con otras Técnicas de Aprendizaje Automático en Análisis Discriminante - La Elección Informada para Decisiones Estratégicas.

Linkedin.com. https://es.linkedin.com/pulse/comparaci%C3%B3n-con-otras-t%C3%A9cnicas-de-aprendizaje-en-la-mora-caballero.

Calvo, D. (2018, marzo 9). Cluster Jerárquicos y No Jerárquicos. Diego Calvo. https://www.diegocalvo.es/cluster-jerarquicos-y-no-jerarquicos/

Campoverde Santos, D. K. (2017). Análisis estadístico multivariante de los patrones de consumo de los hogares ecuatorianos con base a la ENIGHUR 2011-2012 (Bachelor's thesis, Escuela Superior Politécnica de Chimborazo).

Castro Pérez, F. (2020). Castro Pérez, F. de J. (2020). *Probabilidad y estadística*. Klik Soluciones Integrales.

Cisneros Salazar, L. V., Castillo Gómez, S. M., & Martínez Díaz, C. (2006). *Riesgo de quiebra de empresas con análisis discriminante*. Universidad autónoma de Bucaramanga.

Cyberparts.pl. (S/f). Diferencia entre minería de datos descriptiva y predictiva. Recuperado el 19 de noviembre de 2023, de https://es.cyberparts.pl/difference-between-descriptive-and-predictive-data-mining-106#menu-2.

Dagnino, J. (2014). Datos faltantes (missing values). *Revista Chilena de Anestesia*, 43, 332-4.

Daza, N. L. (2019). Elaboración de un modelo de segmentación de jurisdicciones que aporte a la identificación de riesgos de lavado de activos y financiación del terrorismo por este factor en una institución microfinanciera de la ciudad de Popayán [Tesis de pregrado, Fundación universitaria de Popayán]. https://scholar.google.com.co/scholar?hl=es&as_sdt=0%2C5&as_vis=1&q=segmentacion+jurisdiccion&btnG=.http://univida fup.edu.co/repositorio/files/original/91d3a4211f3dae899284f6590e5d1e89.pdf.

De la Fuente Crespo, L. Análisis Cluster. (s/f). Fuenterrebollo.com. Recuperado el 19 de noviembre de 2023, de https://www.fuenterrebollo.com/Master-Econometria/Analisis_Cluster.pdf.

Decreto 1692 de 2020. Por medio del cual se modifica el Decreto 2555 de 2010 en lo relacionado con los sistemas de pago de bajo valor. Diciembre 18 de 2022. DO N. 51532.

Gobierno de España Ministerio de asuntos económicos y transformación digital. (s/f). Guía práctica de introducción al Análisis Exploratorio de Datos. https://datos.gob.es/sites/default/files/doc/file/analisis_exploratorio_de_datos_2021_v6_0.pdf

González, C. G., Lise, A. V., & Felpeto, A. B. (2013). *Tratamiento de datos con R, Statistica y SPSS*. Ediciones Díaz de Santos.

González, M. Á. M., Villegas, A. S., Atucha, E. T., & Fajardo, J. F. (Eds.). (2020). *Bioestadística amigable*. Elsevier.

Gutiérrez, J. A., & Molina, B. (2016). Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. *Revista Ontare*, 3(2), 33-51.

IBM Documentation. (2021, febrero 28). Análisis discriminante. Ibm.com. https://www.ibm.com/docs/es/spss-statistics/27.0.0?topic=features-discriminant-analysis.

LearnToResearch [@learntoresearch4526]. (2020, agosto 3). #15: ANÁLISIS DISCRIMINANTE en SPSS Statistics | + Caso práctico. YouTube. https://www.youtube.com/watch?v=nG-_fsa64Qo.

Linkedin.com. ¿Qué es la Transformación de Datos? Tipos, Herramientas e Importancia. (2023, enero 20). https://es.linkedin.com/pulse/qu%C3%A9-es-la-transformaci%C3%B3n-de-datos-tipos-herramientas-e-importancia-.

López, P. R., Arrastia, M. J. R., & Padilla, C. R. (2021). Metodología de la investigación: de lector a divulgador. *Textos Docentes* (Vol. 83). Universidad Almería.

Lozano Vila, A. (2008). *SARLAFT práctico: Guía para la gestión del riesgo de Lavado de Activos y financiación del Terrorismo.* Taller de edición Rocca

Marqués Asensio, F. (2017). *R en profundidad: programación, gráficos y estadística*. Alfaomega

Mode, E. B. (2021). *Elementos de probabilidad y estadística*. Reverte.

Moine, J. M., Gordillo, S. E., & Haedo, A. S. (2011). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. *Congreso Argentino de Ciencias de la Computación* (Vol. 17).

Moine, J. M. (2013). Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo (Doctoral dissertation, Universidad Nacional de La Plata).

Moreira Álvarez, J. L. A. (2020). Propuesta metodológica basada en la aplicación de minería de datos mediante un modelo de gestión de proyectos para apoyar la toma de decisiones académicas en una institución de educación superior (Doctoral dissertation, Universidad Andrés Bello).

Moreno Castellanos, J. G. (2012). *Método de detección temprana de outliers*. Pontificia Universidad Javeriana.

Nishizaki Fernandes, R. (2017). *Análisis de datos sanitarios aplicando metodología CRISP-DM (Bachelor's thesis)*. Universidad Autónoma de Madrid

Ocaña, F. (s/f). TÉCNICAS ESTADÍSTICAS APLICADAS EN NUTRICIÓN Y SALUD. Ugr.es. Recuperado el 1 de enero de 2024, de https://www.ugr.es/~fmocan/MATERIALES%20DOCTORADO/DESCRIPTIVA%20Y%20EXPLORATORIO.pdf

Otero Ortega, A. (2018). Enfoques de investigación: Métodos Para El Diseño Urbano-Arquitectónico. Obtenido de https://www. researchgate. net/profile/Alfredo_Otero-Ortega/publication/326905435_ENFOQUES_DE_INVESTIGACION_TABLA_DE_CONTENIDO_Contenido/links/5b6b7f9992851ca650526df d/ENFOQUES-DE-INVESTIGACION-TABLA-DE-CONTENIDO-Contenido. pdf.

Proyecto CEACES. (s/f). Análisis cluster. www.uv.es. Recuperado el 19 de noviembre de 2023, de https://www.uv.es/ceaces/multivari/cluster/CLUSTER2.htm.

Salazar Guerrero, L. J. (2020). *Estadística y probabilidad*. Grupo Editorial Patria.

Sancho, R. S. (s/f). *Cuantiles*. Gitbooks.io. Recuperado el 1 de enero de 2024, de https://rsanchezs.gitbooks.io/statsr/content/chapter3/quartile.html

Superintendencia Financiera de Colombia. (Enero 06, 2022). Reporte de Sanciones en Firme. https://wl.superfinanciera.gov.co/SiriWeb/publico/sancion/rep_sanciones_general.jsf

Superintendencia Financiera de Colombia. (Mayo 20, 2022). Circular Externa 011. Modifica las instrucciones relativas a la administración del riesgo de lavado de activos y de la financiación del terrorismo. https://www.superfinanciera.gov.co/inicio/normativa/normativa-general/circulares-externas-cartas-circulares-y-resoluciones-desde-el-ano-/circulares-externas/circulares-externas--10110493.

Timarán Pereira, S. R., Hernández Arteaga, I., Caicedo Zambrano, S. J., Hidalgo Troya, A. & Alvarado Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En: *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional* (pp. 63-86). Bogotá: Ediciones Universidad Cooperativa de Colombia. doi: http://dx.doi.org/10.16925/9789587600490

Torrado-Fonseca, M., & Berlanga-Silvente, V. (2013). Anàlisi discriminant mitjançant SPSS. REIRE *Revista d'Innovació i Recerca en Educació*, 6(2), 150-166.

UIAF. (2006). ¡No se deje usar! Evite que lo involucren en operaciones de lavado de activos o financiación del terrorismo. https://www.supersociedades.gov.co/delegatura_aec/normatividad/estudios_economicos_financieros/otros_documentos/Guia%20para%20evitar%20lavado%20de%20activos.pdf.

Universidad Carlos III de Madrid. (s/f). Tema 6: Análisis Discriminante Introducción. Uc3m.es. Recuperado el 19 de noviembre de 2023, de https://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema6am.pdf.

Vásquez Hidalgo Isabel. (2005, diciembre 18). *Tipos de estudio y métodos de investigación*. Recuperado de https://gestiopolis.com/tipos-estudio-metodos-investigacion/