UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"PROPUESTA DE ESTRATÉGIA DE MONITOREO TRANSACCIONAL ANTI LAVADO DE ACTIVOS EMPLEANDO EL MÉTODO DE WARD Y EL TEOREMA DE CHEBYSHEV"

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR TÍTULO DE INGENIERO ESTADÍSTICO E INFORMÁTICO

ADRIÁN ROMERO DOMINGUEZ

LIMA – PERÚ

2021

UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"PROPUESTA DE ESTRATÉGIA DE MONITOREO TRANSACCIONAL ANTI LAVADO DE ACTIVOS EMPLEANDO EL MÉTODO DE WARD Y EL TEOREMA DE CHEBYSHEV"

PRESENTADO POR

ADRIÁN ROMERO DOMINGUEZ

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL TÍTULO DE INGENIERO ESTADÍSTICO E INFORMÁTICO

SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO

M.A. Fernando René Rosas Villena
PRESIDENTE

Mg. Iván Dennys Soto Rodríguez
ASESOR

Dr. Jorge Chue Gallardo
MIEMBRO

MS. Grimaldo José Febres Huamán
MIEMBRO

A Dios, a mi madre y a mi pareja, con todo cariño, por su invaluable apoyo y por motivarme a ser mejor cada día, como profesional, como líder y como ser humano así como a mi familia, colegas y amigos

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

La adecuada identificación de casos de investigación y la detección de operaciones sospechosas son pilares del sistema de monitoreo anti lavado de activos. En esa línea, el presente estudio ha propuesto implementar dentro del sistema de monitoreo anti lavado una estrategia basada en una segmentación empleando el método de Ward, definiendo umbrales de alertamiento mediante el teorema de Chebyshev. Mediante el método de Ward se logró segmentar el portafolio de clientes de depósitos de plazo fijo en cuatro segmentos homogéneos al interno pero heterogéneos entre sí. Luego en cada uno de estos se definieron sendos umbrales de alertamiento partiendo del teorema de Chebyshev. Esto permitiría a la empresa donde se realizó el estudio acentuar y priorizar los casos a investigar a fin de identificar con mayor celeridad el riesgo de lavado de activos y reportar los casos a la autoridad competente, objetivo central de la unidad de Prevención de Lavado de Activos.

Palabras clave: Segmentación, método de Ward, teorema de Chebyshev, Prevención de Lavado de Activos, Monitoreo transaccional, Reportes de Operaciones Sospechosas

ABSTRACT

The adequate identification of investigation cases and the detection of suspicious transactions are pillars of the anti-money laundering monitoring system. Along these lines, the present study has proposed to implement within the anti-laundering monitoring system a strategy based on segmentation using Ward's technique, defining alert thresholds using Chebyshev's theorem. Using Ward's method, it was possible to segment the portfolio of fixed-term deposit clients into four segments that were homogeneous internally but heterogeneous. Then, in each of these, alert thresholds were defined based on Chebyshev's theorem. This would allow the company where the study was carried out to accentuate and prioritize the cases to be investigated in order to more quickly identify the risk of money laundering and report the cases to the competent authority, the central objective of the Money Laundering Prevention unit.

Keywords: Segmentation, Ward's method, Chebyshev's theorem, Money Laundering Prevention, Transactional Monitoring, Suspicious Transaction Reports.

I. INTRODUCCIÓN

La institución financiera en que se desarrollaron las actividades de este proyecto fue una Caja Rural de Ahorro y Crédito con presencia en la República del Perú. Para referirse a la compañía en adelante se usará el término "la Caja". Sobre la misma Apoyo & Asociados (2020) indicó que:

> "Inició operaciones en agosto 2012 para dedicarse exclusivamente al financiamiento de compras y productos en efectivo a través del uso de su tarjeta de crédito (Visa y Mastercard) en sus tiendas vinculadas (Supermercados y Tiendas por Departamento), así como en establecimientos afiliados. Así, nace como una subsidiaria de su grupo económico, uno de los principales retailers de Latinoamérica. Luego, en mayo 2018, uno de los 4 principales bancos de Perú firma una alianza con el grupo económico original, similar a la realizada en Chile y Colombia, para adquirir por un plazo de 15 años, el 51% de las acciones de la hoy Caja, y así poder administrar en forma conjunta el negocio de tarjeta de crédito y la oferta de otros productos y servicios a sus clientes. A junio 2020, la Caja contaba con una cuota de 3.5% en el total de colocaciones de tarjetas de créditos del sistema financiero. Es importante mencionar que, si medimos la participación por empresas de un mismo Grupo económico, el banco, la financiera del grupo económico y la Caja ocupan la segunda posición con 22.7% por debajo de Interbank y Financiera Oh! (28.8%) y por encima de BCP y Mibanco (20.4%)."

Por otro lado, la emergencia sanitaria ha marcado sin duda un antes y un después, y es que el mundo de hoy (post COVID-19) difiere mucho del mundo antes de la pandemia y este a su vez del mundo de la década pasada. En el caso del sector financiero, en primer lugar se ha identificado un incremento importante en el volumen transaccional relacionado al fraude electrónico y por ende con el lavado de activos, los canales tradicionales han dejado de ser el principal medio de contacto del cliente con la entidad financiera creando nuevos retos a la ya compleja tarea de conocer adecuadamente al cliente y el *e-commerce* ganó el equivalente a dos años de madurez, entre varios otros cambios.

1.1. Problemática

En el contexto comentado resulta prioritario dotar de nuevas herramientas a las empresas del sector financiero y en particular a la Caja a fin de que pueda monitorear, detectar y prevenir el lavado de activos y la financiación del terrorismo de manera más eficaz y/o ágil. Es preciso señalar que la solución a implementar debía asegurar el cumplimiento de los lineamientos del Reglamento de Gestión de Riesgos de Lavado de Activos y del Financiamiento del Terrorismo (Resolución S.B.S. N° 2660-2015) que en su artículo 34 sobre formación de segmentos indicó que:

> "Las empresas deben formar segmentos de mercado, estableciendo grupos que guarden una homogeneidad interna, pero una heterogeneidad entre ellos, de acuerdo con una o varias variables. La información relativa a los segmentos determinados y las variables utilizadas para el conocimiento del mercado deben encontrarse a disposición de la Superintendencia, así como el documento que contenga los resultados de la formación de segmentos." (SBS, 2015)

Por otro lado es pertinente señalar que el Reglamento de Gestión de Riesgos de Lavado de Activos y del Financiamiento del Terrorismo (Resolución S.B.S. N° 2660-2015) indicó en los artículos 57 y 58 sobre la detección y reporte de operaciones sospechosas que:

> "El Anexo Nº 5 de la resolución contiene una relación de señales de alerta que las empresas deben tener en cuenta con la finalidad de detectar operaciones inusuales o sospechosas. Lo anterior no exime a las empresas de considerar otras señales de alerta que pudieran dar origen a la calificación de operaciones que consideren sospechosas de acuerdo con su sistema de prevención del LA/FT. Sin perjuicio de ello, la Superintendencia puede proporcionar información o criterios adicionales que contribuyan a la detección de operaciones inusuales o sospechosas. Las empresas deben efectuar evaluaciones periódicas sobre la totalidad de las señales de alerta definidas por estas y consideradas en la gestión de riesgos LA/FT.
>
> La empresa tiene la obligación de comunicar a la UIF-Perú a través de su oficial de cumplimiento, las operaciones detectadas en el curso de sus actividades, realizadas o que se hayan intentado realizar, que según su buen criterio sean consideradas como sospechosas, sin importar los montos involucrados. A estos

efectos, se considera buen criterio, al discernimiento o juicio que se forma el oficial de cumplimiento a partir, por lo menos, del conocimiento del cliente y del mercado; abarca la experiencia, la capacitación y diligencia en la prevención del LA/FT. La comunicación debe ser realizada de forma inmediata y suficiente, es decir, en un plazo que -conforme a la naturaleza y complejidad de la operación sospechosa- permita al oficial de cumplimiento la elaboración, documentación y remisión del ROS a la UIF-Perú, el cual en ningún caso debe exceder las veinticuatro horas (24) desde que la operación es calificada como sospechosa. Una operación es calificada como sospechosa cuando dicha categoría puede presumirse luego del análisis y evaluación realizado por el oficial de cumplimiento. La comunicación de operaciones sospechosas y el ROS que realizan las empresas por medio de sus oficiales de cumplimiento tienen carácter confidencial y reservado. Únicamente el oficial de cumplimiento, o de ser el caso el oficial de cumplimiento alterno, puede tener conocimiento del envío del ROS. Para todos los efectos legales, el ROS no constituye una denuncia penal o administrativa" (SBS, 2015).

Sobre la gestión de clientes de productos pasivos en la Caja deben indicarse las siguientes precisiones en base al Manual de Productos Pasivos (Caja, 2020) de la institución: i) que el documento no deja constancia de ningún tipo de carterización o agrupamiento comercial, ii) el producto se gestiona a través del aplicativo ADN y iii) el monitoreo transaccional anti lavado es tarea de la unidad de AML la cual se realiza mediante el software de monitoreo ACRM Monitor Plus.

Luego, considerando el marco normativo delimitado por el regulador competente y las características del producto dificultades para lograr un mejor monitoreo se encontró que (i) todo el portafolio de clientes de depósitos de plazo fijo era monitoreado con la misma regla, es decir todo cliente era tratado de igual manera sin considerar ningún aspecto estratégico, cualitativo o cuantitativo tal como se muestra en la Figura 1, (ii) los umbrales de alertamiento vigentes han sido implementados únicamente en base al criterio experto del Oficial de Cumplimiento a cargo en el momento de la implementación de estos y (iii) como consecuencia de lo anterior existe la posibilidad de encontrar espacios para reforzar los criterios de identificación y priorización de los casos de investigación.

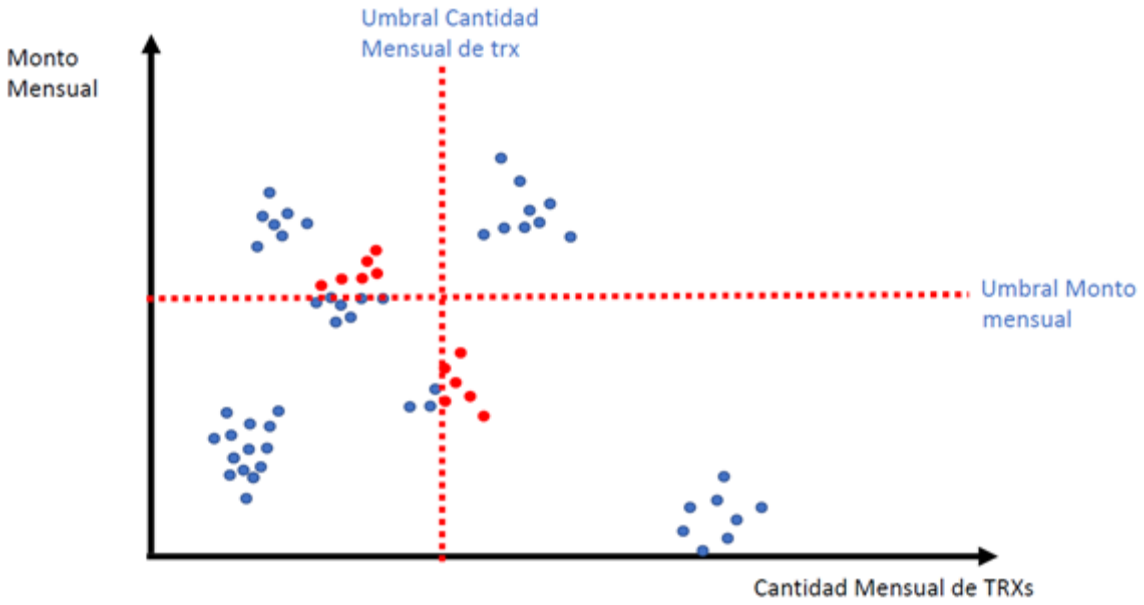

Figura 1: Alertamiento tradicional en el que se define un umbral rígido único

1.2. Objetivos

El objetivo general es proponer una nueva estrategia de monitoreo transaccional a partir de la segmentación de la cartera de clientes de depósitos de plazo fijo de la Caja y el uso de la información transaccional del portafolio de clientes.

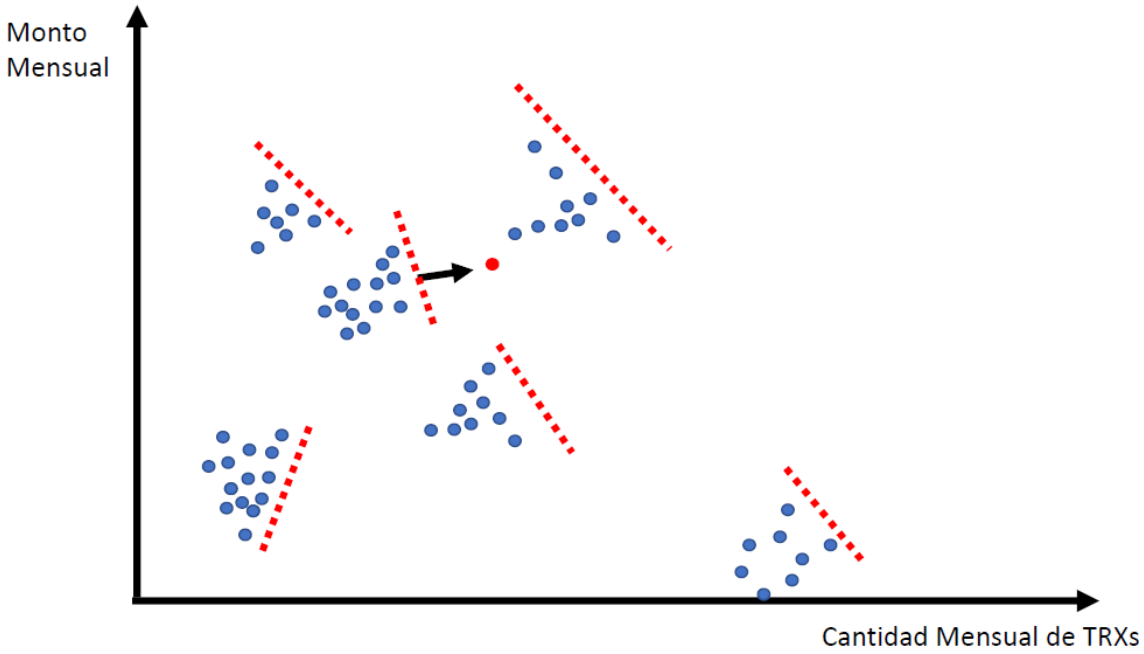

Figura 2: Alertamiento con segmentación en el que se definen umbrales diferenciados

Los objetivos específicos fueron definidos en el siguiente sentido: (i) presentar la propuesta y aplicar una segmentación a la cartera de clientes de depósitos de plazo fijo de la Caja mediante el método de Ward, (ii) definir los umbrales de alertamiento por segmento mediante el teorema de Chebyshev como se muestra en la Figura 2 validando así la hipótesis de que es aceptable la combinación de las técnicas como estrategia de monitoreo anti lavado de activos.

II. MARCO TEÓRICO

2.1. Método de Ward

Para Gutiérrez, González, Torres y Gallardo (1994) el método de Ward está considerado dentro de los Métodos Jerárquicos de Análisis tal como se indica en la Figura 3, sobre estos comentaron que "los llamados métodos jerárquicos tienen por objetivo agrupar *clusters* para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud". En esa línea indicaron además que:

> "El Análisis *Cluster* es el nombre genérico de una amplia variedad de procedimientos que pueden ser usados para crear una clasificación. Más concretamente, un método *cluster* es un procedimiento estadístico multivariante que comienza con un conjunto de datos conteniendo información sobre una muestra de entidades e intenta reorganizarlas en grupos relativamente homogéneos a los que llamaremos clusters. En Análisis *Cluster* poca o ninguna información es conocida sobre la estructura de las categorías, lo cual lo diferencia de los métodos multivariantes de asignación y discriminación. De todo lo que se dispone es de una colección de observaciones, siendo el objetivo operacional en este caso, descubrir la estructura de las categorías en la que se encajan las observaciones. Más concretamente, el objetivo es ordenar las observaciones en grupos tales que el grado de asociación natural es alto entre los miembros del mismo grupo y bajo entre miembros de grupos diferentes."

Luego Gutierrez et al. (1994) recordaron que:

> "Sokal y Sneath argumentaron que un procedimiento eficiente para la generación de clasificaciones biológicas debe recoger todos los posibles datos sobre un conjunto de organismos de interés, estimar el grado de similaridad entre esos organismos y usar un método cluster para colocar los organismos similares en un mismo grupo. Una vez que los grupos de organismos similares han sido encontrados, los miembros de cada uno de ellos deben ser analizados para determinar si representan especies biológicas diferentes. En efecto, Sokal y Sneath asumen que el proceso de reconocimiento de patrones debe ser usado

como base para comprender el proceso evolutivo. El punto de partida para el Análisis *Cluster* es, en general, una matriz X que proporciona los valores de las variables para cada uno de los individuos objeto de estudio, o sea

$$X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1n} \\ \vdots & \ddots & \vdots & : & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{in} \\ \vdots & \dots & \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mj} & \dots & x_{mn} \end{pmatrix}$$

La *i*-ésima fila de la matriz X contiene los valores de cada variable para el i-ésimo individuo, mientras que la j-ésima columna muestra los valores pertenecientes a la j-ésima variable a lo largo de todos los individuos de la muestra."

Los métodos jerárquicos tienen por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que se minimice alguna función distancia o bien W maximice alguna medida de similitud. Los métodos jerárquicos como se subdividen a su vez en aglomerativos y disociativos. Los aglomerativos comienzan el análisis con tantos grupos como individuos haya en el estudio. A partir de ahí se van formando grupos de forma ascendente, hasta que, al final del proceso, todos los casos están englobados en un mismo conglomerado. Los métodos disociativos o divisivos realizan el proceso inverso al anterior. Empiezan con un conglomerado que engloba a todos los individuos. (Gutierrez et al., 1994).

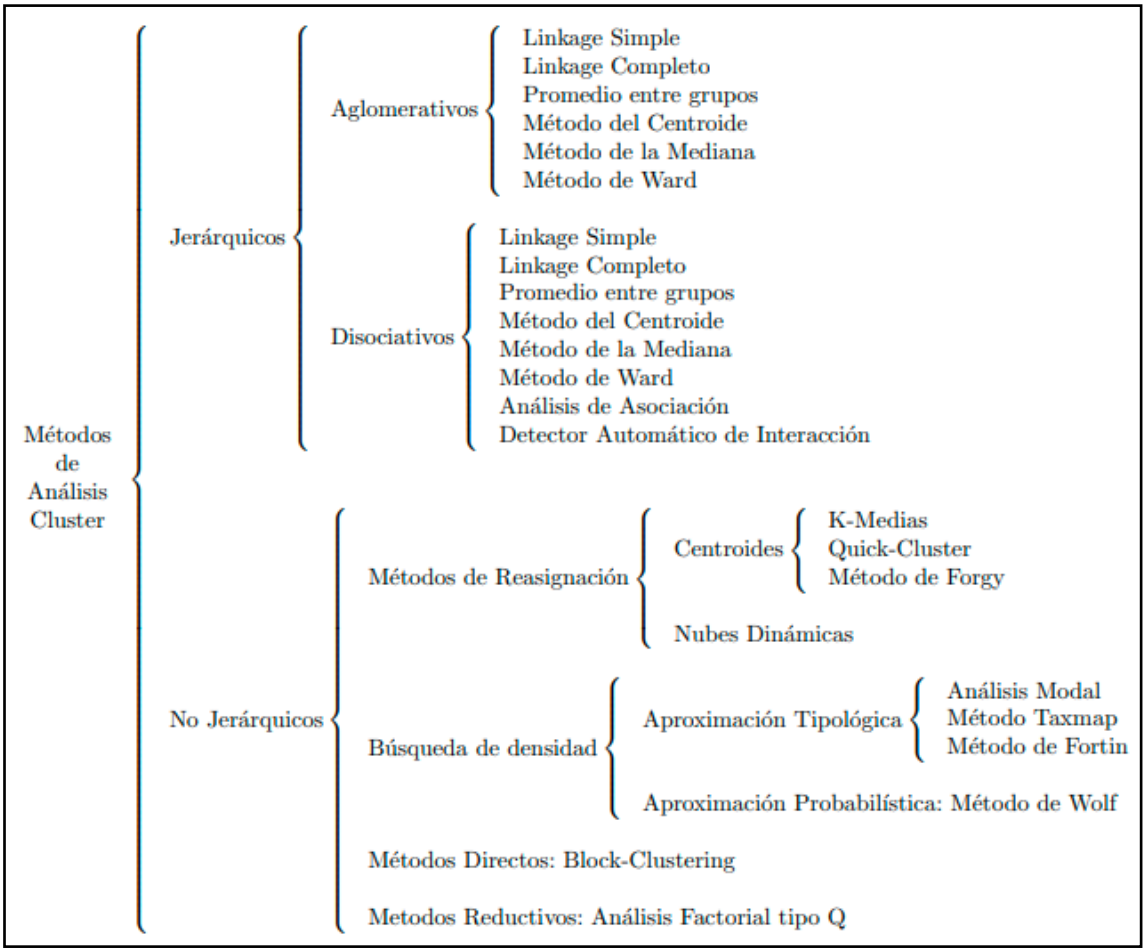

Figura 3: Diagrama de clasificación de métodos de análisis cluster (Gutierrez et al., 1994).

El método de Ward es un procedimiento jerárquico en el cual, en cada etapa, se unen los dos clusters para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada cluster, de cada individuo al centroide del cluster (Gutierrez et al, 1994). Notemos por

- x_{ij}^{k} al valor de la j−ésima variable sobre el i-ésimo individuo del k−ésimo cluster, suponiendo que dicho cluster posee n_k individuos.
- m^k al centroide del cluster *k*, con componentes m_j^k
- E_k a la suma de cuadrados de los errores del cluster *k*, o sea, la distancia euclídea o euclidiana al cuadrado entre cada individuo del cluster *k* a su centroide

$$E_k = \sum_{i=1}^{n_k}\sum_{j=1}^{n}\left(x_{ij}^{k} - m_j^{k}\right)^2 = \sum_{i=1}^{n_k}\sum_{j=1}^{n}\left(x_{ij}^{k}\right)^2 - n_k\sum_{j=1}^{n}\left(m_j^{k}\right)^2$$

- E a la suma de cuadrados de los errores para todos los clusters, o sea, si suponemos que hay h clusters

$$E = \sum_{k=1}^{h} E_k$$

El proceso comienza con *m* clusters, cada uno de los cuales está compuesto por un solo individuo, por lo que cada individuo coincide con el centro del cluster y por lo tanto en este primer paso se tendrá $E_k = 0$ para cada cluster y con ello, E = 0. El objetivo del método de Ward es encontrar en cada etapa aquellos dos clusters cuya unión proporcione el menor incremento en la suma total de errores, *E* (Gutierrez et al., 1994).

En palabras de otro autor, el Método de Ward es el único entre los algoritmos de agrupamiento aglomerativo que se basa en un criterio clásico de suma de cuadrados, produciendo conjuntos o grupos que minimizan la dispersión dentro del grupo en cada fusión binaria. Además, el método de Ward es particularmente interesante porque busca clústeres en el espacio euclidiano multivariado. Éste es también el espacio de referencia en los métodos de ordenación multivariante, y en particular en el análisis de componentes principales (Legendre y Murtagh, 2014). Por su parte Oliva (2015:20) señaló lo siguiente:

> ‘‘Este método, también conocido como ‘incremento en la suma de los cuadrados’ o método de mínima varianza, (…) tiene como objetivo unificar grupos de forma tal que la variabilidad dentro de los grupos no aumente dramáticamente. En cada paso se fusionan los dos clústeres que producen la suma de cuadrados dentro de clúster (variabilidad within o intra-clústeres) mínima entre todas las posibles particiones que se obtienen fusionando dos clústeres del paso previo. En este contexto ‘suma de cuadrados dentro’ refiere a la suma de las distancias al cuadrado de las observaciones del clúster respecto de la media de las observaciones del mismo clúster.”

A pesar de que las representaciones finales difieren en cuanto a las distancias que unen a unos objetos con otros, las agrupaciones encontradas suelen ser las mismas (Figura 4).

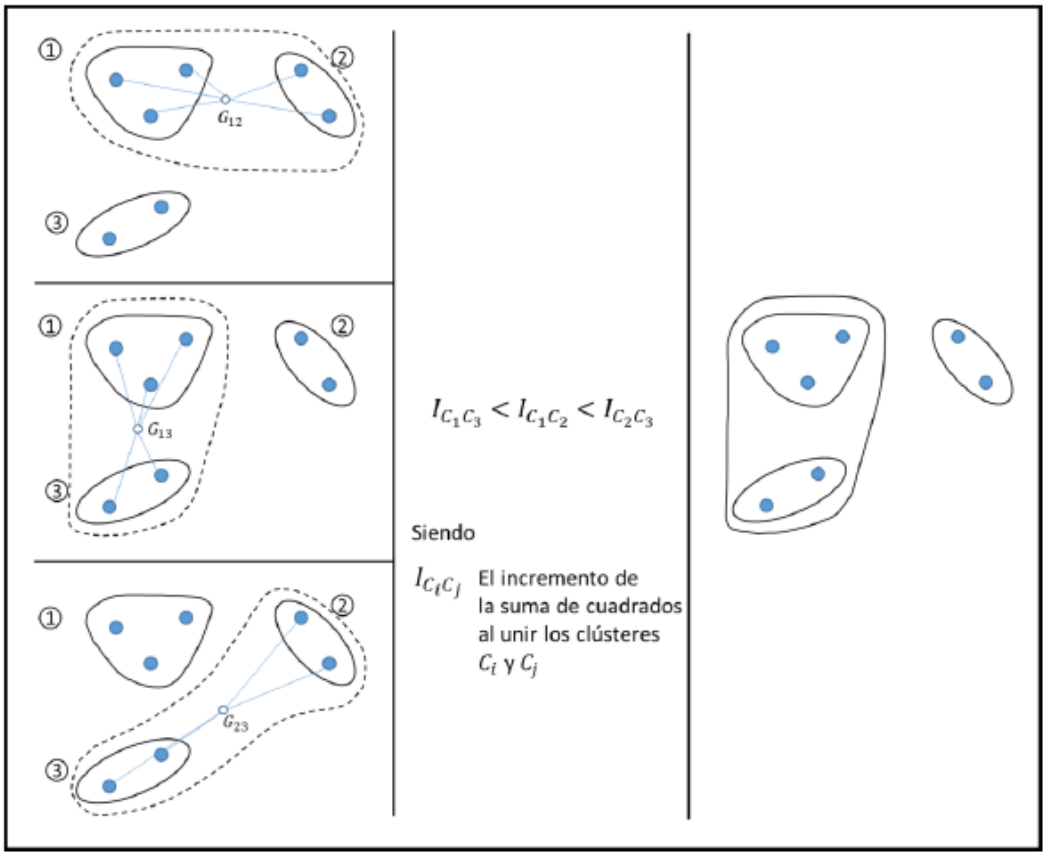

Figura 4: Representación gráfica del método de Ward. Fuente: (Dongo, 2017) Adaptado de Pedret (2003).

Para representar gráficamente este tipo de métodos se utilizan los dendogramas. Estos reproducen la agrupación de un conjunto de variables, indicando la distancia a la que se efectúa la unión en el eje de ordenadas. Las distintas uniones se llevan a cabo a través de dos líneas verticales (si está orientado hacia arriba) u horizontales (si el gráfico está orientado hacia la derecha) procedentes de las variables y una horizontal o vertical encargada de conectarlas para la distancia correspondiente (Peña, 2002, p. 241). Parafraseando al autor los dendogramas son representaciones gráficas en forma de árbol que resumen el proceso de agrupación en los análisis de clústeres. Las unidades similares se conectan mediante enlaces cuya posición en el diagrama está determinada por el nivel de similitud/disimilitud entre los objetos como se grafica en la Figura 5.

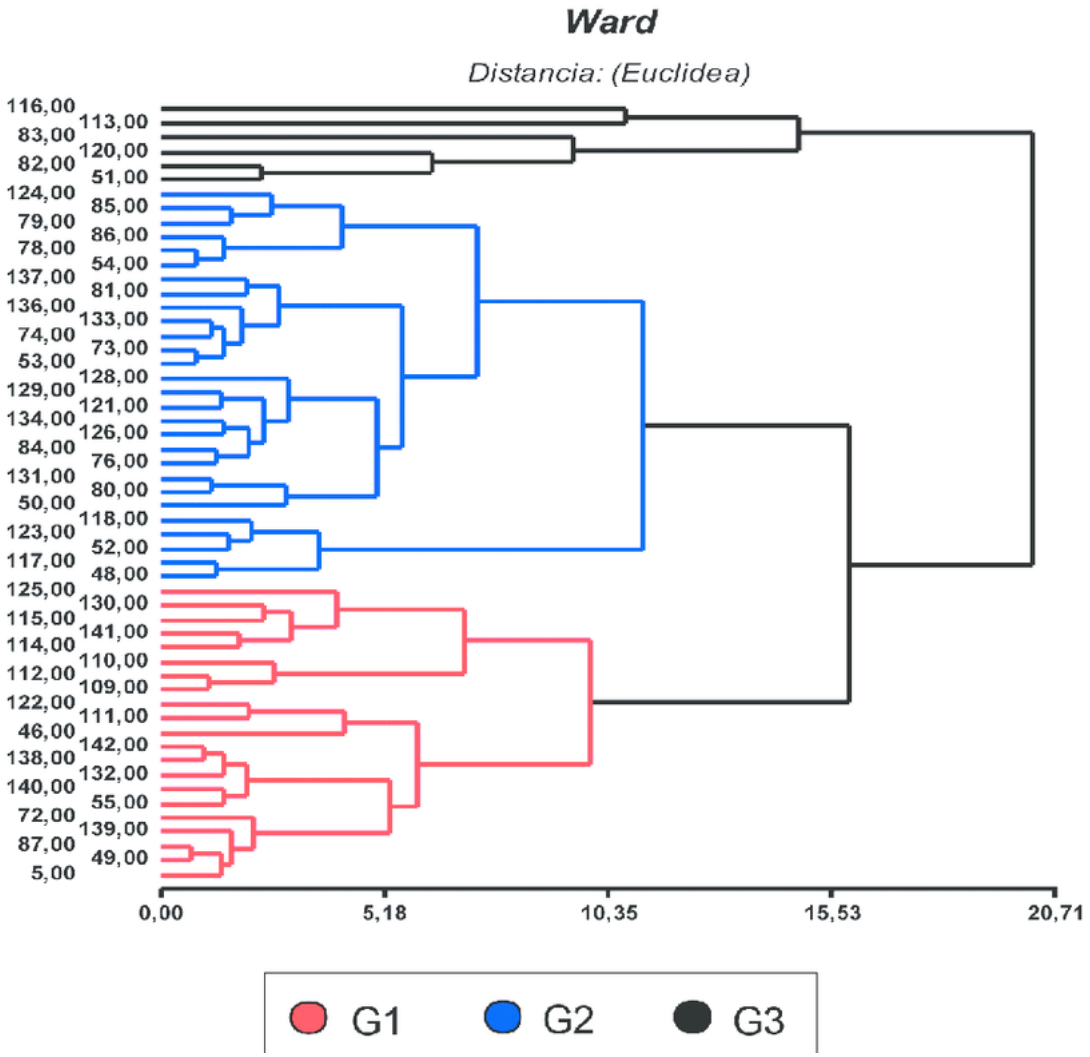

Figura 5. Representación gráfica de un dendrograma mediante el método Ward.
Fuente: Research Gate.Net.

Una consideración importante es que los datos no estén afectados por la presencia de casos atípicos. De La Fuente (2011) plantea dos soluciones para abordar este problema:

- Cambiar datos iniciales por datos promedios. Por ejemplo, en lugar de número de salas de cine en una ciudad, se podría usar el número de salas de cine por cada mil habitantes de una ciudad.
- Realizar transformaciones de la distribución de datos utilizando la escala de Tukey.

- La asimetría positiva se corrige con raíces cuadradas y logaritmos naturales cuando tienen valores bajos, y con funciones inversas o inversos cuadráticos cuando los valores son elevados.
- La asimetría negativa se corrige mediante elevaciones cúbicas y cuadráticas cuando es suave, y con antilogaritmos cuando es muy elevada.

Debido a que el análisis clúster estudia las características estructurales de un conjunto de observaciones con el fin de agruparlas en conjuntos homogéneos, y al no ser propiamente una técnica de inferencia estadística, las exigencias de normalidad, linealidad, entre otros supuestos; no son fundamentales como sí lo son en procedimientos de inferencia.

Una correcta aplicación del análisis clúster requiere que los datos cumplan con las siguientes condiciones básicas: i) ausencia de correlación entre las variables, ii) número de variables no muy elevado y, iii) que las variables no deben estar medidas en escalas diferentes.

Existe cierta controversia sobre si la tipificación debe ser un procedimiento a utilizar en todo análisis clúster: Everitt y Edelborck, citados por De La Fuente (2011), son algunos de los autores que no defienden el proceso de estandarización, y plantean tres posibles soluciones para el problema de tener variables con distinta unidad: i) re-categorizar todas las variables en binarias, y aplicar a éstas una distancia apropiada para este tipo de medidas, ii) realizar distintos análisis clúster con grupos de variables homogéneas (en cuanto a su medida), y sintetizar después los diferentes resultados, iii) utilizar la distancia de Gower, que es aplicable con cualquier tipo de métrica.

2.2.Teorema de Chebyshev

En otro ámbito, la segunda técnica a emplear en la propuesta es el Teorema de Chebyshev a fin de definir cotas o límites transaccionales. Para Hernandez (2004) la desigualdad de Chebyshev es uno de los resultados clásicos más importantes de la teoría de probabilidad. Establece que para una variable aleatoria X,

$$P(|X - \mu| \geq k) \leq \frac{Var(X)}{k^2}, \quad k > 0,$$

Donde $\mu = EX$ la cual debe de ser finita.

En otras palabras, la desigualdad de Chebyshev nos dice que la varianza es una medida de dispersión de los valores de X alrededor de su valor esperado. Generalmente, la demostración de este resultado se basa en la siguiente desigualdad conocida como desigualdad de Markov:

$$\mathrm{P}(\mathrm{X} \geq k) \leq \frac{\mathrm{EX}^r}{k^r}\ , \quad \mathrm{k}, \mathrm{r} > 0.$$

En la literatura, a este tipo de desigualdades, cuya característica es la comparación de la probabilidad de la cola de la distribución y su valor esperado, se le conoce como desigualdades tipo Chebyshev. Nótese que cuando el lado derecho de la desigualdad es mayor o igual a 1. Tal es el caso de una variable aleatoria X tal que $\mathrm{E}|\mathrm{X}|^{\mathrm{r}} > \mathrm{k}^{\mathrm{r}}$, podemos decir que entre más pequeño sea el valor del lado derecho de la desigualdad tipo Chebyshev, obtenemos información más precisa respecto a las probabilidades.

Para Hernandez (2004) estas desigualdades son la herramienta básica para demostrar resultados no menos importantes como la Ley de los Grandes Números, entre otros. Además tienen aplicaciones en estadística, así como en otras áreas de las matemáticas. La aplicación directa de las desigualdades tipo Chebyshev es el de aproximar probabilidades por medio de cálculo de cotas.

Por su parte, Teorema.Top (2019) indicó que “el Teorema de Chebyshev es el encargado de explicar una manera de saber qué fracción de datos se encuentra dentro de las desviaciones estándar K de la media para cualquier conjunto de datos en específico”. Luego se define conceptualmente un umbral sobre el cual se discriminan las operaciones como normales o inusuales respecto de su segmento.

Desde su perspectiva, Marco (2019) indicó al respecto que “La desigualdad de Chebyshev es un teorema utilizado en estadística que proporciona una estimación conservadora (intervalo de confianza) de la probabilidad de que una variable aleatoria con varianza finita se sitúe a una cierta distancia de su esperanza matemática o de su media.” Por su parte, Teorema.Top (2019) indicó que “cuando sucede una distribución normal, se sabe que al menos un 68% de los datos es una desviación estándar de la

media. Por otro lado, el 95% son dos desviaciones están de la media, y el 99% aproximadamente se encuentra dentro de las tres desviaciones estándar de la media.” Estas equivalencias se representan gráficamente en la Figura 6.

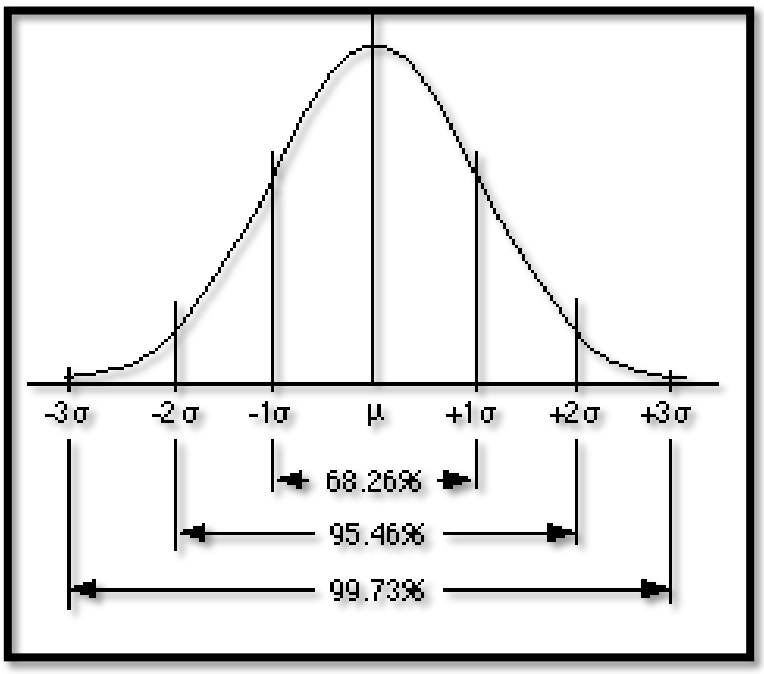

Figura 6. Distribución de la variable según el Teorema de Chebyshev

III. MARCO METODOLÓGICO

La presente sección, el marco metodológico, tiene como fin comentar la forma en que se recolectó la información necesaria para alcanzar los objetivos planteados (Pasco y Ponce, 2015). Por lo tanto, involucra precisar el tipo de investigación, el diseño metodológico y la de recolección de datos. En ese orden de ideas, el tipo de investigación del presente estudio corresponde a uno de tipo exploratorio. Sobre el diseño metodológico Pasco y Ponce (2015) indicaron que existen diferentes formas de clasificar los diseños de investigación, las más comunes basan sus distinciones en el enfoque, la estrategia general y el horizonte temporal de la investigación. Al respecto, debe indicarse que el enfoque del estudio por su naturaleza es cuantitativo al emplear datos con medición numérica.

La estrategia general de investigación corresponde al tipo "estudio de caso" pues sigue un proceso iniciado con la definición de los temas principales de análisis, para luego detenerse en su estudio profundo mediante la recolección, la interpretación y la validación de datos. Por otro lado, el horizonte temporal de la investigación corresponde al de un estudio transversal. Finalmente, la recolección de datos corresponde al portafolio de clientes de Depósitos de Plazo Fijo de la Caja, es decir, emplea el universo de individuos de estudio.

Se consideró el periodo enero 2019 a diciembre 2019 como el periodo a emplear como delimitación temporal. De manera intencional se omite emplear información del ejercicio 2020 por estar impactada por la pandemia del COVID-19. Respecto al ámbito geográfico el estudio este se ha ceñido al territorio peruano donde la Caja tiene cobertura. Las técnicas se aplicaron al portafolio de clientes de Depósitos de Plazo Fijo, este fue elegido luego de calificar como el producto de mayor riesgo LA/FT de la empresa, esto debido a que el ticket promedio de *cash in* es elevado (PEN 38,100) y la ocurrencia de un evento de riesgo conllevaría un impacto material importante.

La información de conocimiento de los clientes tuvo un tratamiento de datos de manera análoga al diagrama de la Figura 7 en la cual se Simplilearn (2018) comenta las siguientes etapas:

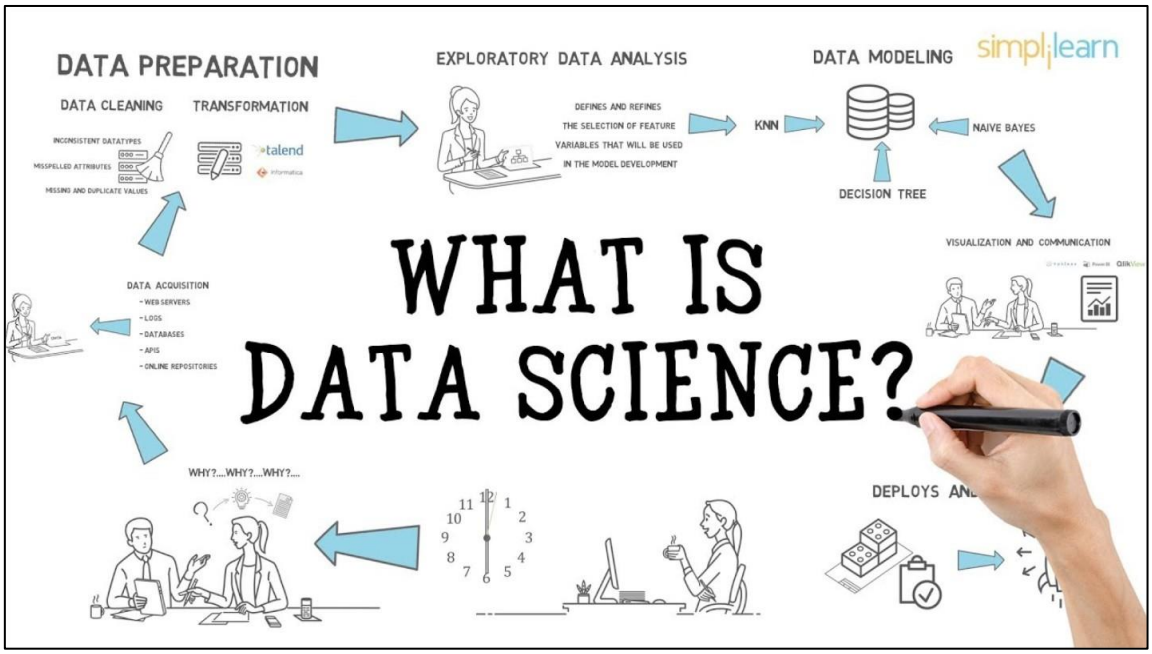

Figura 7. Ciclo de procesamiento de datos (Simplilearn, 2018).

- *Definición de la problemática y brainstorming:* Cuestionar el problema, las características particulares o generales e información conocida previamente. En el caso de esta investigación se cuestionó si acaso podía plantearse el método de monitoreo adicional para la detección del lavado de activos de la Caja y de ser así como podría definirse este.

- *Adquisición de la data:* Corresponde a la obtención de los datos originales completos y/o suficientes para el análisis, ya sea desde *logs, web servers* o repositorios *on line*. En el caso de la presente investigación se generó una serie de reportes automatizados del sistema de productos pasivos ADN de la Caja. Los reportes explotados fueron:
  - Reporte Mancomunos
  - Reporte Operaciones 201901
  - Reporte Operaciones 201902
  - Reporte Operaciones 201903
  - Reporte Operaciones 201904
  - Reporte Operaciones 201905
  - Reporte Operaciones 201906
  - Reporte Operaciones 201907
  - Reporte Operaciones 201908
  - Reporte Operaciones 201909
  - Reporte Operaciones 201910

- Reporte Operaciones 201911
- Reporte Operaciones 201912

- *Preparación de la data:* La data original rara vez tendrá todos los datos limpios y/o estructurados por lo que el investigador debe darse a la tarea de lograr ello. Se recomienda corregir por ejemplo datos inconsistentes, variables o atributos perdidos y valores duplicados, luego prosigue la estructuración de la data, esto incluye la estandarización de ser necesario. En el caso de esta investigación los reportes extraídos fueron consolidados pues cada uno generó información transaccional mensual. Se eliminaron los registros de operaciones anuladas, se completaron o desestimaron datos que no se capturaban en el origen, se convirtió el formato de las fechas para que sean útiles, se unificaron las monedas de los depósitos de plazo fijo a soles peruanos y se consolidaron los montos de las aperturas por cliente y por mes calendario entre otras actividades. Es pertinente indicar que no fue necesario estandarizar los datos pues la información utilizada fueron cada uno de los niveles de riesgo de cada variable que se encuentran en la escala de 1.00 a 5.00. Es decir, que se cumple con el requisito de que los datos no estén afectados por la presencia de casos atípicos y los datos se encuentren en la misma escala, dicho de otra forma se emplearon variables homogéneas. Esto se logró usando los niveles de riesgo de cada variable definidos en la metodología de cálculo de score de riesgo LA/FT de la unidad

- *Análisis exploratorio:* El considerado el paso más importante. Define y re define la selección de las variables que podrían usarse en la construcción del modelo. En el caso de esta investigación las principales variables creadas (riesgo de edad, riesgo de departamento de residencia y de afiliación, monto aperturado entre otras) pasaron por un análisis exploratorio y de identificación de correlación entre variables. La relación original de variables empleadas en el estudio se listan en la tabla 1

Tabla 1: Relación original de variables empleadas en el estudio

Variable	Tipo de variable	Ejemplo	Descripción
Nro	Num	1	Correlativo simple
Tipo_Doc	Varchar	1	Tipo de documento (DNI=1, CEX=2)
DNI	Varchar	2618012	Número de Documento de Identidad
Llave	Varchar	10000261801214	ID interno único del cliente
Flag_ROS	Varchar	0	Calificación previa como ROS (1=Si, 0=No)
Riesgo_CRR	Varchar	Moderado	Nivel de riesgo LA/FT según metodología CRR (Alto, Moderado, Bajo)
Antiguedad_Cliente	Varchar	1	Antigüedad (en años) del individuo como cliente de la Caja
Antiguedad_Cliente_Riesgo	Num	5	Nivel de riesgo LA/FT según metodología CRR respecto a la antigüedad
Departamento Ag	Varchar	15	Región donde fue afiliado el cliente
Departamento Ag_Riesgo	Num	3	Nivel de riesgo LA/FT según metodología CRR respecto a la región de afiliación
Departamento Res	Varchar	14	Región donde reside el cliente
Departamento Res_Riesgo	Num	4	Nivel de riesgo LA/FT según metodología CRR respecto a la región de residencia
Edad	Varchar	5	Edad del cliente (en años)
Edad_Riesgo	Num	1	Nivel de riesgo LA/FT según metodología CRR respecto a la
Flag_Ajustador	Varchar	0	Indicador de si el cliente ha calificado en alguna categoría de riesgo*
Flag_Ajustador_Riesgo	Num	1	Nivel de riesgo LA/FT según metodología CRR respecto al indicador de categorías de riesgo
Monto201901	Num	S/. -	Monto acumulado en DPFs aperturados en enero 2019 por el cliente
Monto201902	Num	S/. -	Monto acumulado en DPFs aperturados en febrero 2019 por el cliente
Monto201903	Num	S/. 50,000	Monto acumulado en DPFs aperturados en marzo 2019 por el cliente
Monto201904	Num	S/. -	Monto acumulado en DPFs aperturados en abril 2019 por el cliente
Monto201905	Num	S/. -	Monto acumulado en DPFs aperturados en mayo 2019 por el cliente
Monto201906	Num	S/. -	Monto acumulado en DPFs aperturados en junio 2019 por el cliente
Monto201907	Num	S/. -	Monto acumulado en DPFs aperturados en julio 2019 por el cliente

Monto201908	Num	S/.	-	Monto acumulado en DPFs aperturados en agosto 2019 por el cliente
Monto201909	Num	S/.	-	Monto acumulado en DPFs aperturados en septiembre 2019 por el cliente
Monto201910	Num	S/.	-	Monto acumulado en DPFs aperturados en octubre 2019 por el cliente
Monto201911	Num	S/.	-	Monto acumulado en DPFs aperturados en noviembre 2019 por el cliente
Monto201912	Num	S/.	-	Monto acumulado en DPFs aperturados en diciembre 2019 por el cliente
MT	Num	S/.	50,000	Monto acumulado en DPFs aperturados en todo el año 2019 por el cliente
MT_cat	Num		4	Nivel de riesgo del monto acumulado en DPFs aperturados en todo el año 2019 en base a los percentiles 20, 40, 60 y 80
RiesgoTransac135	Num		1	Nivel de riesgo del factor transaccional según la metodología CRR
Ocupación	Varchar		33	Código de ocupación del cliente
Ocupación_Riesgo	Num		5	Nivel de riesgo LA/FT según metodología CRR respecto a la ocupación del cliente
País_Nac	Varchar		1	País de nacimiento o nacionalidad del cliente
País_Nac_Riesgo	Num		1	Nivel de riesgo LA/FT según metodología CRR respecto al país de nacimiento del cliente
País_Res	Varchar		1	País de residencia del cliente
País_Res_Riesgo	Num		1	Nivel de riesgo LA/FT según metodología CRR respecto al país de residencia del cliente

- *Modelamiento:* Corresponde a la identificación y aplicación de la técnica propiamente dicha tanto para su entrenamiento como para su testeo. Para el presente estudio correspondió a la aplicación de distintos modelos de *clustering* mediante el *software* estadístico R, la elección del mejor modelo (el de Ward), definición del punto de corte y número de segmentos además de la aplicación del teorema de Chebyshev en cada uno de los 4 conglomerados definidos asumiendo una probabilidad del 95%. Las sentencias o *scrips* del proceso pueden revisarse en extenso en el Anexo 1 del presente documento.

- *Visualización y Comunicación:* Tan importante como generar nueva información es comunicarla efectivamente los resultados y/o hallazgos utilizando de ser necesario herramientas de presentación visual y/o tablas ejecutivas. En el caso de esta investigación se constituye en la elaboración y

exposición del informe ejecutivo “Propuesta de monitoreo transaccional adicional para DPF en la Caja” al Director de Operaciones FIU.

- *Desarrollo y mantenimiento:* Constituye el desarrollo del modelo en un ambiente real (productivo). Luego de ser implementado exitosamente se realiza el seguimiento del desenvolvimiento del modelo, para ello se suele, usar cuadros de datos a fin de tener una visión en tiempo real. En el caso de esta investigación se constituye en el pase a producción de la técnica y el monitoreo del volumen de alertas generadas así como el ratio de los casos escalados, es decir, reportados como Reportes de Operaciones Sospechosas – ROS respecto del volumen total de señales de alertas generadas.

IV. RESULTADOS Y DISCUSIÓN

4.1.Resultados

4.1.1. Del análisis exploratorio

Como parte del análisis exploratorio se revisaron los estadísticos de las 5 principales variables que se emplearon en la construcción del modelo como se muestra en la tabla 2. Cabe comentar que no se realizó el análisis de correlación entre variables al no identificarse razón relevante para ello toda vez que cada una es independiente.

Tabla 2: Estadísticos de las 5 principales variables independientes

Estadístico	Nivel de Riesgo de la Antigüedad del Cliente	Nivel de Riesgo del Dpto. de residencia	Nivel de Riesgo de la Edad	Nivel de Riesgo de TM	Nivel de Riesgo de la Ocupación
Número de Observaciones	541	541	541	541	541
Nivel de Riesgo promedio	4.9	3.1	3.1	3.0	3.8
Nivel de riesgo más bajo	3.0	1.0	5.0	1.0	1.0
Nivel de riesgo más alto	5.0	5.0	1.0	5.0	5.0
Desviación estándar del nivel de riesgo	0.4	0.4	0.4	1.4	1.8
Mediana	5.0	3.0	2.0	3.0	5.0
Moda	5.0	3.0	1.0	2.0	5.0

4.1.2. Del *clustering* mediante el método de Ward

La segmentación mediante el método de Ward no solo resultó viable sino que se constituyó en el modelo con la mejor propuesta de segmentación para los clientes en base al criterio de decisión del Coeficientes de aglomeración tal como se muestra en la Figura 8.

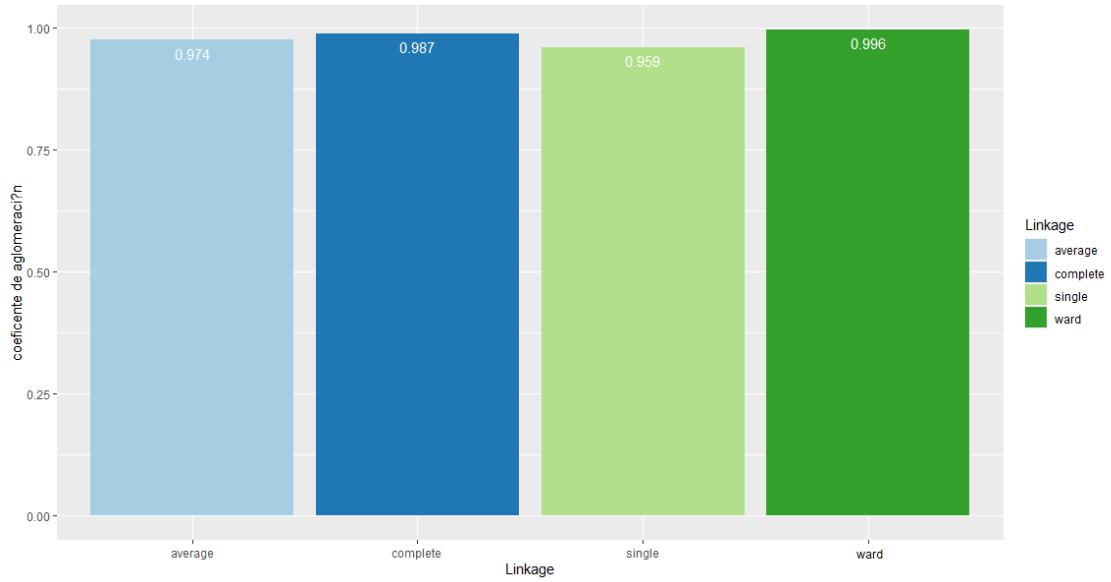

Figura 8. Coeficientes de aglomeración por modelo de agrupamiento

Para la selección del número de clusters se tuvieron en cuenta las siguientes consideraciones: i) dada la metodología de calificación de riesgos de la entidad se tiene un mínimo de 3 grupos de clientes según riesgo (alto, medio y bajo), ii) obtener el mejor ordenamiento de los clientes de alto riesgo y iii) tener el menor número posible de conglomerados a fin de no diluir demasiado el portafolio. En esa línea con 4 clusters se obtuvo un mejor ordenamiento de los clientes de alto riesgo con el número de grupos más cercano. Como consecuencia, los conglomerados generados tienen las características resumidas en la Tabla 3 y una distribución de variables como se muestra en la Figura 9.

Tabla 3: Características de los 4 conglomerados del portafolio de clientes

Conglomerado	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total general
Clientes por cluster	49	127	296	69	541
Clientes de alto riesgo	5	1	0	1	7
Cartera total	S/. 2,275,839	S/. 5,834,883	S/. 15,800,984	S/. 2,755,408	S/. 26,667,116
Cash in promedio	S/. 46,446	S/. 45,944	S/. 53,382	S/. 39,933	S/. 49,292
Cash in máximo	S/. 301,000	S/. 240,000	S/. 400,000	S/. 339,000	S/. 400,000
Cash in mínimo	S/. 1,000	S/. 2,184	S/. 1,000	S/. 1,000	S/. 1,000
Desv. Estándar	S/. 56,358	S/. 46,743	S/. 61,426	S/. 57,349	S/. 57,370

El cluster 3 congregó casi el 60% de la cartera por el volumen de clientes en él y posee el mayor ticket promedio así como el monto máximo acumulado. La menor proporción del portafolio se encuentra en el cluster 1 sin embargo el menor ticket promedio lo tiene el cluster 4.

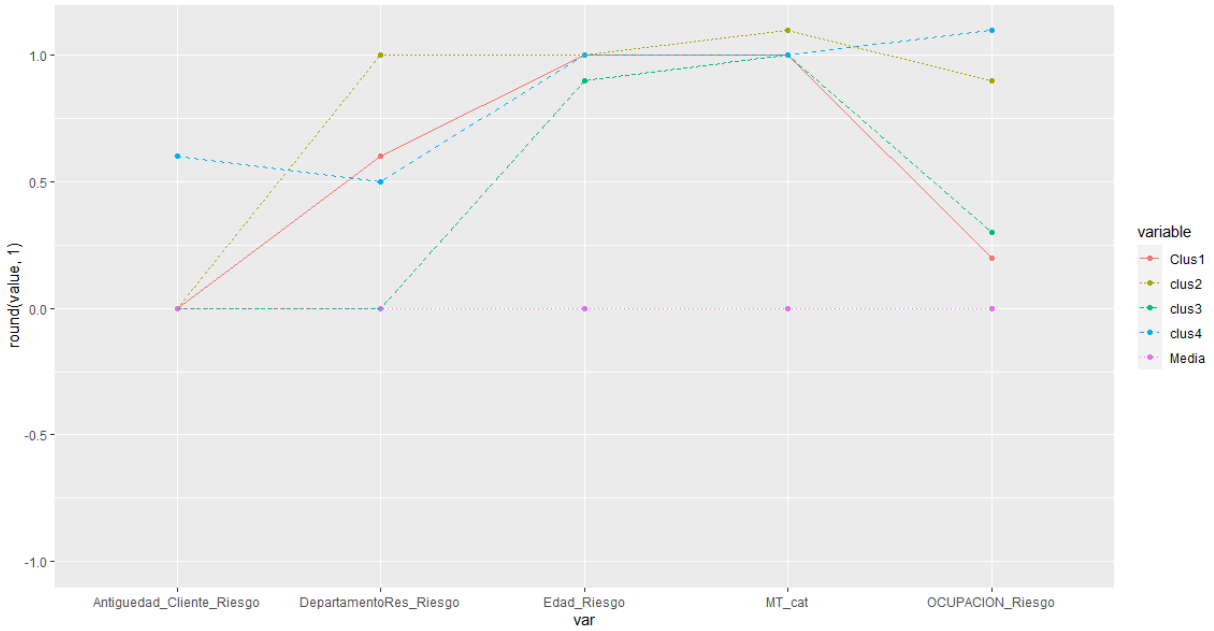

Figura 9. Distribución de variables finales por cluster

Con los cuatro clusters configurados se realizó un análisis exploratorio de las 5 principales variables el cuál se grafica en la Figura 10. Así, el riesgo del monto aperturado se distribuye entre los 4 clusters, por otro lado, el riesgo del departamento de residencia presenta poca dispersión y se centra principalmente en el nivel 3. El nivel de riesgo de la ocupación se concentra el máximo nivel en los cluster 1 y 2 y en menor nivel en el cluster 3. El riesgo de la edad presenta una mediana uniforme en el nivel 2, un Q1 y Q3 de 1 y 3 respectivamente y valores hasta 5.

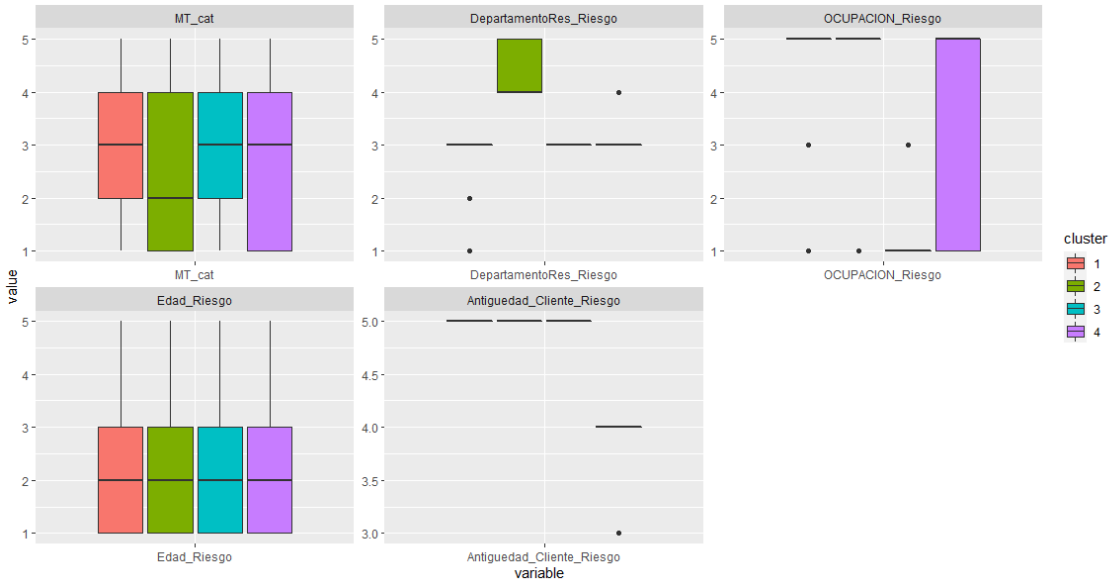

Figura 10. Exploración de variables finales por cluster

Por otro lado, el riesgo relativo a la antigüedad del cliente se concentra en el nivel 5 para los clusters 1,2 y 3 y en el nivel 4 en el cluster 4 y van en el mismo sentido los niveles de riesgo promedio por cluster tal como se muestra en la tabla 4.

Tabla 4: Niveles de riesgo promedio

Nivel de riesgo promedio	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Respecto al monto total de DPFs	2.8	3.0	3.1	2.6
Respecto al departamento de residencia	4.3	3.0	3.0	3.1
Respecto a la ocupación	4.1	1.2	5.0	3.3
Respecto a la edad	2.5	2.0	2.5	2.2
Respecto a la antigüedad del cliente	5.0	5.0	5.0	4.0

4.1.3. De la definición de cotas mediante el Teorema de Chebyshev

Definidos los clusters se aplicó el teorema de Chebyshev para definir los umbrales de cada uno de los 4 conglomerados en línea con el planteamiento de la fórmula del teorema asumiendo una probabilidad del 95% y obteniendo un valor de k= 4.472136.

Empleando:

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

Reemplazando:

$$P\left(\mu_{Cluster_i Mes_j} - 4.472\sigma_{Cluster_i Mes_j} < X < \mu_{Cluster_i Mes_j} + 4.472\sigma_{Cluster_i Mes_j}\right) \geq 95\%$$

Como resultado se obtuvieron los límites para cada uno de los grupos tal como se muestra en la Tabla 5.

Tabla 5: Límites calculados para los 4 conglomerados del portafolio

Periodo	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Ene-19	S/. 211,529	S/. 92,993	S/. 83,591	S/. 131,169
Feb-19	S/. 128,417	S/. 105,091	S/. 81,408	S/. 54,569
Mar-19	S/. 68,877	S/. 143,438	S/. 50,758	S/. 74,449
Abr-19	S/. 142,622	S/. 133,452	S/. 31,250	S/. 189,285
May-19	S/. 15,884	S/. 61,395	S/. 52,575	S/. 77,674
Jun-19	S/. 40,217	S/. 66,706	S/. 86,641	S/. 6,880
Jul-19	S/. 16,526	S/. 50,838	S/. 28,986	S/. 10,265
Ago-19	S/. 53,053	S/. 55,565	S/. 23,519	S/. 13,625
Set-19	S/. 30,687	S/. 15,556	S/. 32,961	S/. 16,586
Oct-19	S/. -	S/. 6,475	S/. 25,618	S/. 46,636
Nov-19	S/. 13,717	S/. 21,885	S/. 16,654	S/. -
Dic-19	S/. 58,913	S/. 36,451	S/. 33,156	S/. 93,867
Promedio	S/. 65,037.80	S/. 65,820.49	S/.78,926.47	S/. 59,583.85

4.2. Evaluación Económica – Financiera

La inversión estimada para la implementación del proyecto asciende a 40 horas hombre a un costo de USD 110 por hora. Es decir la inversión total asciende a USD 4,400.00. Por otro lado, el monitoreo anti lavado es una herramienta que preserva uno de los mayores activos de la empresa: Su reputación. La misma resulta invaluable y por lo tanto no es posible cuantificar un retorno económico por la implementación de la propuesta planteada ni estimar un ratio costo/beneficio.

4.3.Discusión

Gracias a la formación que brinda el Departamento Académico de Ingeniería Estadística e Informática de la UNALM tanto en el ámbito técnico como humanístico se

contó con una base sólida sobre la cual el estadístico molinero propone soluciones robustas y viables como la planteada en el presente informe.

Los umbrales de alertamiento para cada cluster fueron calculados en orden de S/. 59,583.85 a S/.78,926.47, siendo el más alto el del cluster 3 y el más bajo el del cluster 4. Dado que el umbral único del alertamiento tradicional se encuentra en S/ 50,000.00 las alertas que se generen con el nuevo sistema siempre serán alertas previamente identificadas.

Los beneficios potenciales son: i) la oportunidad para el autor de proponer un proyecto con alta visibilidad, ii) para la empresa la obtención de un nuevo modelo de monitoreo que asegure la robustez del control, la mitigación del riesgo de multas y la mitigación del riesgo de que la empresa sea utilizada para lavar activos, iii) para la Unidad de Inteligencia Financiera y la Fiscalía de la Nación - El Ministerio Público la recepción de reportes de operaciones sospechosas (ROS) e Informes de Inteligencia Financiera respectivamente con mayor rapidez y , iv) para la sociedad se logra un acorralamiento del crimen organizado y la obtención de más sentencias condenatorias por lavado de activos lo que en el largo plazo permitirá una sociedad más segura.

explicar brevemente si los resultados obtenidos son coherentes y concordante con los objetivos

estado situacional de la institución, así mismo refrendarlos con resultados de trabajos

V. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

- La agrupación mediante el método de Ward resulta viable e incluso recomendable para el portafolio de clientes analizado con número recomendado de 4 clusters.

- La definición de los umbrales para cada agrupación aplicando el teorema de Chebyshev resultó igualmente viable y recomendable como método alternativo sustentado para definir los límites de alertamiento tal como se indica en la Tabla x.

- La combinación de ambas técnicas permite proponer una sólida estrategia de monitoreo transaccional para la lucha contra lavado de activos segmentando la cartera de clientes de DPF de la Caja y el análisis de la información transaccional del portafolio

5.2. Recomendaciones

- Realizar análisis similares empleando portafolios de clientes de otros productos y/o entidades a fin de comparar los resultados obtenidos e identificar oportunidades de mejora.

- Asimismo, se recomienda realizar diversas combinaciones de técnicas tanto de segmentación como de definición de umbrales de monitoreo a fin de buscar escenarios en que se incremente la eficiencia del monitoreo anti lavado.

- Posterior a la implementación, elaborar un análisis de sensibilidad respecto de las alertas transaccionales que se generen tanto con la técnica de monitoreo tradicional como con la técnica de monitoreo propuesta a fin de cuestionar la continuidad de la regla y/o su umbral vigente.

VI. REFERENCIAS BIBLIOGRÁFICAS

Apoyo & Asociados. (2020). *CRAC Cencosud Scotia Perú S.A. - Informe Semestral*. Recuperado de https://www.aai.com.pe/wp-content/uploads/2020/09/Bco-Cencosud-0620.pdf

Dongo, B. (2017). *Descripción metodológica del análisis clúster utilizando el algoritmo de Ward*. (Tesis de pregrado). Universidad Nacional Agraria La Molina, Lima, Perú

Caja Cencosud Scotia Perú S.A. (2020). *Manual de Productos pasivos.*

De La Fuente, S. (2011). Análisis conglomerados. Madrid, España, Universidad Autónoma de Madrid. Libro electrónico.

Gutiérrez, R. González, A. Torres, F. Gallardo, J.A. (1994). *Técnicas de análisis de datos multivariable*. Tratamiento computacional. Universidad de Granada. Recuperado de http://www.ugr.es/~gallardo/

Hernandez, I. (2004). *La mejor desigualdad tipo Chevyshev*. Universidad de Sonora. Departamento de matemáticas. Recuperado de https://lic.mat.uson.mx/tesis/120TesisIsmael.PDF

Legendre P.; Murtagh F. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification 31: 274-295

Marco, F. (2019. *Desigualdad de Chebyshev*. Recuperado de https://economipedia.com/definiciones/desiguladad-chebyshev-teorema.html

Oliva C. (2015). *Métodos para la segmentación de datos longitudinales. Aplicación a datos de rendimientos de cultivos en Argentina*. Tesis Lic. Buenos Aires, Argentina, UBA. 72p.

Pedret R; Sagnier L; Camp F. (2003). *Herramientas para segmentar mercados y posicionar productos*. Barcelona, España, Planeta. 329p.

Peña, D. (2002). *Análisis de datos multivariantes*. Madrid: McGraw-Hill

Ponce M.; Pasco M. (2015). *Guía de investigación en Gestión*. Vicerrectorado de Investigación. Pontificia Universidad Católica del Perú. Recuperado de http://repositorio.pucp.edu.pe/index/handle/123456789/172009

SBS. (2015). Reglamento de Gestión de Riesgos de Lavado de Activos y del Financiamiento del Terrorismo (Resolución S.B.S. N° 2660-2015). Recuperado de https://intranet2.sbs.gob.pe/dv_int_cn/1540/v5.0/Adjuntos/2660-2015.R.pdf

Simplilearn. (2018). Data Science In 5 Minutes | What Is Data Science? Recuperado de https://www.youtube.com/watch?v=X3paOmcrTjQ

Teorema.Top (2019). Teorema de Chebyshev con Explicación Sencilla. Recuperado de https://www.teorema.top/teorema-de-chebyshev/

ANEXO 1: Scrip de R empleado en el estudio

```
#=================================================================
# REMOVEMOS OBJETOS
#=================================================================
rm(list=ls())
#=================================================================
# LIBRERIAS
#=================================================================
library(Validacion)
library(xlsxjars)
library(rJava)
library(xlsx)
library(data.table)
library(RODBC)
library(readxl)
library(data.table)
library(stringr)
library(Validacion)
library(data.table)
library(ggplot2)
library(cluster)
library(factoextra)
library(ggplot2)
library(dendextend)
library(NbClust)
library(igraph)
require(reshape2)
library(purrr)
library(data.table)
library(ecodist)
library(clValid)
#=================================================================
# PARAMETROS INICIALES
#=================================================================
#Abriendo lo guardado
# ruta <- "./Datos/Entrega/"
```

```
ruta <- "./Datos/"
load(paste0(ruta,"ScoreCRR_CCS_V3.Rdata"))
getwd()
#-----------------------------------------------------------------#
#                  POBLACIÓN                                      #
#-----------------------------------------------------------------#
# Clientes DPF
Validación_Factores<-Validación_Factores[Tipo_Producto==4 ]
Validación_Factores<-Validación_Factores[Flag_Transaccion==1 ]

head( Validación_Factores)

# Selección de variables de interes
# Cluster1<-Validación_Factores[,c("PAIS_NAC_cat"
#                                  ,"DepartamentoAg_cat","DepartamentoRes_cat",
#                                  "FLAG_Ajustador_cat","OCUPACION_cat",
#                                  "Edad_cat",
#                                #   "PAIS_RES_cat",
#                                 "Antiguedad_Cliente_cat",
#                                  "MT")]
Cluster1<-Validación_Factores[,c("LLAVE","PAIS_NAC_cat"
                          ,"DepartamentoAg_cat","DepartamentoRes_cat",
                          "FLAG_Ajustador_cat","OCUPACION_cat",
                          "Edad_cat",
                           "PAIS_RES_cat",
                          "Antiguedad_Cliente_cat",
                          "MT" ,
                          "PAIS_NAC_Riesgo"
                          ,"DepartamentoAg_Riesgo","DepartamentoRes_Riesgo",
                          "FLAG_Ajustador_Riesgo","OCUPACION_Riesgo",
                          "Edad_Riesgo",
                          "PAIS_RES_Riesgo",
                          "Antiguedad_Cliente_Riesgo",
                          "RiesgoTransac"
        ,"Riesgo_CRR","FLAG_ROS_VAL","Monto201901","Monto201902","Monto201903"
        ,"Monto201904","Monto201905","Monto201906","Monto201907",
```

```
        "Monto201908","Monto201909","Monto201910","Monto201911","Monto201912")]
Cluster1[is.na(Monto201901),Monto201901:=0 ]
Cluster1[is.na(Monto201902),Monto201902:=0 ]
Cluster1[is.na(Monto201903),Monto201903:=0 ]
Cluster1[is.na(Monto201904),Monto201904:=0 ]
Cluster1[is.na(Monto201905),Monto201905:=0 ]
Cluster1[is.na(Monto201906),Monto201906:=0 ]
Cluster1[is.na(Monto201907),Monto201907:=0 ]
Cluster1[is.na(Monto201908),Monto201908:=0 ]
Cluster1[is.na(Monto201909),Monto201909:=0 ]
Cluster1[is.na(Monto201910),Monto201910:=0 ]
Cluster1[is.na(Monto201911),Monto201911:=0 ]
Cluster1[is.na(Monto201912),Monto201912:=0 ]
# Percentil variable
quantile(Cluster1$MT,c(.20,.40,.60,.80,1) )
v_rangos <- c(-Inf,10001,22000,42000,70963,Inf)
Cluster1[,MT_cat := cut(Cluster1$MT,v_rangos,labels = FALSE, dig.lab = 6)]
table(Cluster1$MT_cat )
head(Cluster1)
table(Cluster1$FLAG_ROS_VAL )
table(Cluster1$FLAG_ROS_VAL, Cluster1$Riesgo_CRR  )
table(Cluster1$FLAG_ROS)
#-----------------------------------------------------------------#
#       1. AN?LISIS CLUSTER JER?RQUICOS AGLOMERATIVO          #
#-----------------------------------------------------------------#
#-----------------------------------------------------------------#
#              Hallando la matriz de distancias               #
#-----------------------------------------------------------------#
v_categoricas                                                                <-
        c("MT_cat","DepartamentoRes_Riesgo","OCUPACION_Riesgo","Edad_Riesgo","Anti
        guedad_Cliente_Riesgo" )

for (j in v_categoricas){
  Cluster1[,eval(j):= as.numeric(get(j))]}

str( Cluster1)
```

```
Cluster2<-
        Cluster1[,c("MT_cat","DepartamentoRes_Riesgo","OCUPACION_Riesgo","Edad_Ries
        go","Antiguedad_Cliente_Riesgo" )]

dist<-daisy(scale(Cluster2), metric = "euclidean")
#-----------------------------------------------------------------#
#             Cluster vecino m?s cercano                          #
#-----------------------------------------------------------------#
Clus1<-hclust(dist,method="single")
fviz_dend(Clus1, cex = 0.5)
#-----------------------------------------------------------------#
#             Cluster vecino m?s lejano                           #
#-----------------------------------------------------------------#
Clus2 <- hclust(dist,method="complete")
fviz_dend(Clus2, cex = 0.5)
#-----------------------------------------------------------------#
#               Cluster enlace promedio                           #
#-----------------------------------------------------------------#
Clus3 <- hclust(dist,method="average")
fviz_dend(Clus3, cex = 0.5)
#-----------------------------------------------------------------#
#                 Cluster enlace ward                             #
#-----------------------------------------------------------------#
Clus4 <- hclust(dist,method="ward.D2")
fviz_dend(Clus4, cex = 0.5)

#-----------------------------------------------------------------#
#     Comparando cluster a trav?s de coeficientes                 #
#-----------------------------------------------------------------#
COEF.AGL<-function(distancia){
 m <- c( "single", "complete", "average", "ward.D2")
 names(m) <- c( "single", "complete", "average","ward")
 ac <- function(x) {
  coef(hclust(dist,method= x))
 }
 AC=map_dbl(m, ac)
 AC
```

```
  piss=data.frame(setnames(setDT(round(melt(AC),3),keep.rownames =
          TRUE),c("Linkage","coeficiente")))
  d1=ggplot(data=piss, aes(x =Linkage, y=coeficiente, fill=Linkage)) +
    geom_bar(stat="identity", position=position_dodge())+
    geom_text(aes(label=coeficiente), vjust=1.6, color="white",
           position = position_dodge(0.9), size=4)+
    scale_fill_brewer(palette="Paired")+ylab("coeficente de aglomeraci?n")
  d1
  print(d1)
  print(AC)
}
COEF.AGL(dist)
#-----------------------------------------------------------------#
#                  Comparando clusters                    #
#-----------------------------------------------------------------#
# Create dos dendrogramas (los mejores)
dend1 <- as.dendrogram (Clus2)
dend2 <- as.dendrogram (Clus4)
# Create a list to hold dendrograms
dend_list <- dendlist(dend1, dend2)
tanglegram(dend1, dend2, highlight_distinct_edges = FALSE, common_subtrees_color_lines =
          TRUE,
         common_subtrees_color_branches = TRUE, main = paste("entanglement =",
                                        round(entanglement(dend_list), 2)))
#-----------------------------------------------------------------#
#                  Eligiendo n?mero de cluster              #
#-----------------------------------------------------------------#
Nclus <- NbClust(scale(Cluster2), distance = "euclidean",method = "ward.D2")
Nclus
par(mfrow=c(1,1))
fviz_nbclust(Nclus)
#-----------------------------------------------------------------#
#                  Visualizaci?n de clusters                #
#-----------------------------------------------------------------#
# corte en 3 grupos y colores por grupos
fviz_dend(Clus4, k = 4, # corte en 3 grupos
        cex = 0.7, # tama?o de etiqueta
```

```
        k_colors = c(  "#E7B800", "#2E9FDF","#FC4E07"),
        color_labels_by_k = TRUE, # color etiqueta por grupo
        rect = TRUE) # a?adir rect?ngulo al rededor de los grupos
fviz_dend(Clus4, k = 4, cex = 0.7, horiz = FALSE, k_colors = "jco",
        rect = TRUE, rect_border = "jco", rect_fill = TRUE)
fviz_dend(Clus4, k = 4, cex = 0.7, horiz = TRUE, k_colors = "jco",
        rect = TRUE, rect_border = "jco", rect_fill = TRUE)
fviz_dend(Clus4, k = 4, cex = 0.7, horiz = FALSE, k_colors = "jco",
        rect = TRUE, rect_border = "jco", rect_fill = TRUE)
fviz_dend(Clus4, k = 4, k_colors = "jco",
        type = "phylogenic", repel = TRUE)
#-----------------------------------------------------------------#
#              Caracter?sticas de lo clusters                     #
#-----------------------------------------------------------------#
# Cortando en 4 cluester 4 (el mejor Ward)
# ----------------------
grp=cutree(Clus4, k = 4)   ###### Si deseas cambiar el cluster a 3, reemplaza aqui k=3 ######
# Number de casos en cada cluster
# -------------------------------
table(grp)
# Descripci?n de cada cluster
# -------------------------------
med<-aggregate(Cluster2, by=list(cluster=grp), mean);med
med

# Diagrama de caracterizaci?n (sirve para escalar)
# -------------------------------
v_categoricas <- colnames(Cluster2 )
for (j in v_categoricas){
  Cluster2[,eval(j):= as.numeric(get(j))]}
M<-as.data.frame(t(rbind(aggregate(scale(Cluster2), by=list(cluster=grp), sd)[,-1])))
a=as.vector(colMeans(scale(Cluster2)))
fin=data.frame(M,a,names(Cluster2));names(fin)<-
          c("Clus1","clus2","clus3","clus4","Media","var")
ali=melt(fin,id.vars = "var")
ggplot(ali, aes(x=var,y=round(value,1),group=variable,colour=variable)) +
  geom_point()+ geom_line(aes(lty=variable))+ expand_limits(y = c(-1, 1))
```

```
# Diagrama de caracterizaci?n 2 (esto es para la data inicial)
# ------------------------------
dd <- cbind(Cluster2, cluster =grp )
dd$cluster<-as.factor(dd$cluster)
df.m <- melt(dd, id.var = "cluster")
p <- ggplot(data = df.m, aes(x=variable, y=value)) +
  geom_boxplot(aes(fill=cluster))+ facet_wrap( ~ variable, scales="free")
p
# N° OPTIMO DE CLUSTER 2 (varia deacuerdo al analista le pondremos 4), MEJOR
          METODO WARD
# Data final con el cluster ward dd
dt_final<-cbind(dd, Cluster1 )
table(dt_final$cluster,dt_final$Riesgo_CRR)
table(dt_final$cluster,dt_final$RiesgoTransac)
table(dt_final$cluster,dt_final$FLAG_ROS_VAL)
## Exportando el Cluster
save(dt_final,
    file=paste0(ruta,"Cluster1.Rdata"))
write.table(dt_final, 'Cluster1.csv', sep="|")
#-----------------------------------------------------------------#
#           2. AN?LISIS CLUSTER JER?RQUICOS DIVISIVO              #
#-----------------------------------------------------------------#
#-----------------------------------------------------------------#
#                      Algoritmo DIANA                            #
#-----------------------------------------------------------------#
Clus5=as.hclust(diana(scale(Cluster2)))
fviz_dend(Clus5, cex = 0.5)
# ?ndice divisivo
# ---------------
coef(Clus5)
# Sugerencia N?mero de Cluster
# ----------------------------
# Gr?fica de silueta
# ----------------------------
dist.depar=daisy(scale(Cluster2))
par(mfrow=c(1,3))
```

```
for(h in 2:4){
  conglomerados=cutree(Clus5,h)
  plot(silhouette(conglomerados,dist.depar),main = "")
}
# corte en 3 grupos y colores por grupos
fviz_dend(Clus5, k = 3, # corte en 3 grupos
          cex = 0.7, # tama?o de etiqueta
          k_colors = c(  "#E7B800", "#2E9FDF","#FC4E07"),
          color_labels_by_k = TRUE, # color etiqueta por grupo
          rect = TRUE) # a?adir rect?ngulo al rededor de los grupos

# corte en 4 grupos y colores por grupos
fviz_dend(Clus5, k = 4, # corte en 3 grupos
          cex = 0.7, # tama?o de etiqueta
          k_colors = "jco",
          color_labels_by_k = TRUE, # color etiqueta por grupo
          rect = TRUE) # a?adir rect?ngulo al rededor de los grupos
#-----------------------------------------------------------------#
#              Caracter?sticas de lo clusters                     #
#-----------------------------------------------------------------#
# Cortando en 4 cluster
# ----------------------
grp2=cutree(Clus5, k = 4)

# Number de casos en cada cluster
# -------------------------------
table(grp2)

# Descripci?n de cada cluster
# -------------------------------
med2<-aggregate(Cluster2, by=list(cluster=grp2), mean);med2

# Diagrama de caracterizaci?n
# -------------------------------
M<-as.data.frame(t(rbind(aggregate(scale(Cluster2), by=list(cluster=grp2), mean)[,-1])))
a=as.vector(colMeans(scale(Cluster2)))
```

```
fin=data.frame(M,a,names(Cluster2));names(fin)<-
          c("Clus1","clus2","clus3","clus4","Media","var")
ali=melt(fin,id.vars = "var")
ggplot(ali, aes(x=var,y=round(value,1),group=variable,colour=variable)) +
  geom_point()+ geom_line(aes(lty=variable))+ expand_limits(y = c(-1.9, 1.9))

# Diagrama de caracterizaci?n 2
# ------------------------------
dd <- cbind(Cluster2, cluster =grp2 )
dd$cluster<-as.factor(dd$cluster)
df.m <- melt(dd, id.var = "cluster")
p <- ggplot(data = df.m, aes(x=variable, y=value)) +
  geom_boxplot(aes(fill=cluster))+ facet_wrap( ~ variable, scales="free")
p
#-----------------------------------------------------------------#
#                      ALGORITMO K-MEAN                           #
#-----------------------------------------------------------------#
#-----------------------------------------------------------------#
# Eligiendo el n?mero de Cluster  -METODO PARA ELEGIR EL NUMERO DE CLUSTER
          #
#-----------------------------------------------------------------#
# Elbow method (WSS)
fviz_nbclust(Cluster2, kmeans, method = "wss") +
  geom_vline(xintercept = 3, linetype = 2)+
  labs(subtitle = "Elbow method")
# Silhouette method
fviz_nbclust(Cluster2, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")
# Gap statistic
set.seed(123)
fviz_nbclust(Cluster2, kmeans, nstart = 25, method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")
# METODOS BSS
Niveles=function(Cluster2,n)
{
  y=rep(0,n)
  for (i in 1:n)
```

```
  {
    y[i]=kmeans(Cluster2, i,nstart = 25)$betweenss
  }
  print(plot(y,x=c(1:n),xlab="Number of cluster k", ylab="BSS",
             main = "BSS Method",
             pch=16, type = "o",col="blue"),
        abline(v = 3, col="blue", lwd=1, lty=2))
  return(y)
}
Niveles(Cluster2,n=8)
#ELIGIENDO EN RELACION A LOS 30 ALGORITMOS
nb <- NbClust(Cluster2, distance = "euclidean", min.nc = 2,
              max.nc = 10, method = "kmeans")
fviz_nbclust(nb)

nb <- NbClust(Cluster2, distance = "euclidean", min.nc = 2,
              max.nc = 10, method = "ward.D2")
fviz_nbclust(nb)
#-----------------------------------------------------------------#
#                   Visualizaci?n de los cluster                  #
#-----------------------------------------------------------------#
km.res <- kmeans(Cluster2, 3,nstart = 25)  #CAMBIAR EL N?MERO DE CLUSTER
fviz_cluster(km.res, data = Cluster2,
             palette = "jco",
             ellipse.type = "euclid", # Concentration ellipse
             star.plot = TRUE, # Add segments from centroids to items
             repel = TRUE, # Avoid label overplotting (slow)
             ggtheme = theme_minimal())
fviz_cluster(km.res, data = Cluster2,
             palette = "jco",
             ellipse.type = "convex", # Concentration ellipse
             star.plot = TRUE, # Add segments from centroids to items
             repel = TRUE, # Avoid label overplotting (slow)
             ggtheme = theme_minimal())
#-----------------------------------------------------------------#
#                Caracter?sticas de los cluster                   #
#-----------------------------------------------------------------#
```

```
# Cortando en 3 cluster
# ----------------------
grp3=km.res$cluster
# Number de casos en cada cluster
# -------------------------------
table(grp3)
# Descripci?n de cada cluster
# -------------------------------
med3<-aggregate(Cluster2, by=list(cluster=grp3), mean);med3
# Diagrama de caracterizaci?n
# -------------------------------
M1<-as.data.frame(t(rbind(aggregate(Cluster2, by=list(cluster=grp3), mean)[,-1])))
a=as.vector(colMeans(Cluster2))
fin=data.frame(M1,a,names(Cluster2));names(fin)<-c("Clus1","clus2","clus3","Media","var")
ali=melt(fin,id.vars = "var")
ggplot(ali, aes(x=var,y=round(value,1),group=variable,colour=variable)) +
  geom_point()+ geom_line(aes(lty=variable))+ expand_limits(y = c(2, 10))
# Diagrama de caracterizaci?n 2
# -------------------------------
dd <- cbind(Cluster2, cluster =grp3 )
dd$cluster<-as.factor(dd$cluster)
df.m <- melt(dd, id.var = "cluster")
p <- ggplot(data = df.m, aes(x=variable, y=value)) +
  geom_boxplot(aes(fill=cluster))+ facet_wrap( ~ variable, scales="free")
p
#=================================================================
# REMOVEMOS OBJETOS
#=================================================================
rm(list=ls())
#=================================================================
# LIBRERIAS
#=================================================================
library(Validacion)
library(xlsxjars)
library(rJava)
library(xlsx)
library(data.table)
```

```
library(RODBC)
library(readxl)
library(data.table)
library(stringr)
library(Validacion)
library(data.table)
library(ggplot2)
library(cluster)
library(factoextra)
library(ggplot2)
library(dendextend)
library(NbClust)
library(igraph)
require(reshape2)
library(purrr)
library(data.table)
library(ecodist)
library(clValid)
library(dplyr)
#=================================================================
# PARAMETROS INICIALES
#=================================================================
#Abriendo lo guardado
# ruta <- "./Datos/Entrega/"
ruta <- "./Datos/"
#load(paste0(ruta,"ScoreCRR_CCS_V3.Rdata"))
getwd()
load(paste0(ruta,"Cluster1.Rdata"))
#-----------------------------------------------------------------#
#                  POBLACIÓN                                      #
#-----------------------------------------------------------------#
##----------##
## Cluster1  #
##----------##
## Selección
dic_datos_train <- select(dt_final[cluster==1],"LLAVE",starts_with("Monto"),"MT", "cluster" )
v_categoricas <- colnames(dic_datos_train )
```

```
for (j in v_categoricas){
  dic_datos_train[,eval(j):= as.numeric(get(j))]}
str(dic_datos_train )
## Estadisticos
mean.each.hour <- sapply(dic_datos_train[,c(2:14)], mean) # Con ceros#sd.each.hour <-
          sapply(dic_datos_train[,c(2:14)], sd) # Con ceros
mean.each.hour <- sapply(dic_datos_train[,c(2:14)], mean, na.rm=TRUE) # Sin ceros
sd.each.hour <- sapply(dic_datos_train[,c(2:14)], sd, na.rm=TRUE) # Sin ceros
mean.each.hour
sd.each.hour
p95.each.hour <- sapply(dic_datos_train[,c(2:14)], function(x) quantile(x, probs = 0.95))
p95.each.hour
## Creando Función Chebyshev.k
Chebyshev.k <- function(rt_prob=0.05){k = sqrt(1 / rt_prob)}
k <- Chebyshev.k(0.05)
k
Chebyshev.max <- function(means, stds, rt_probs){
  k <- Chebyshev.k(rt_probs)
  theoretical.max <- means + k*stds
  return(theoretical.max)
}
## Aplicando Función Chebyshev.k
p95.theoretical <- Chebyshev.max(mean.each.hour, sd.each.hour, rt_probs=0.05)
p95.theoretical >= p95.each.hour
##----------##
## Cluster2  #
##----------##
## Selección
dic_datos_train <- select(dt_final[cluster==2],"LLAVE",starts_with("Monto"),"MT", "cluster" )
v_categoricas <- colnames(dic_datos_train )
for (j in v_categoricas){
  dic_datos_train[,eval(j):= as.numeric(get(j))]}
str(dic_datos_train )
## Estadisticos
mean.each.hour <- sapply(dic_datos_train[,c(2:14)], mean)
sd.each.hour <- sapply(dic_datos_train[,c(2:14)], sd)
mean.each.hour
```

```
sd.each.hour
p95.each.hour <- sapply(dic_datos_train[,c(2:14)], function(x) quantile(x, probs = 0.95))
p95.each.hour
## Creando Función Chebyshev.k
Chebyshev.k <- function(rt_prob=0.05){k = sqrt(1 / rt_prob)}
k <- Chebyshev.k(0.05)
k
Chebyshev.max <- function(means, stds, rt_probs){
  k <- Chebyshev.k(rt_probs)

  theoretical.max <- means + k*stds
  return(theoretical.max)
}
## Aplicando Función Chebyshev.k
p95.theoretical <- Chebyshev.max(mean.each.hour, sd.each.hour, rt_probs=0.05)
p95.theoretical >= p95.each.hour
##----------##
## Cluster3  #
##----------##
## Selección
dic_datos_train <- select(dt_final[cluster==3],"LLAVE",starts_with("Monto"),"MT", "cluster" )
v_categoricas <- colnames(dic_datos_train )
for (j in v_categoricas){
  dic_datos_train[,eval(j):= as.numeric(get(j))]}
str(dic_datos_train )
## Estadisticos
mean.each.hour <- sapply(dic_datos_train[,c(2:14)], mean)
sd.each.hour <- sapply(dic_datos_train[,c(2:14)], sd)
mean.each.hour
sd.each.hour
p95.each.hour <- sapply(dic_datos_train[,c(2:14)], function(x) quantile(x, probs = 0.95))
p95.each.hour
## Creando Función Chebyshev.k
Chebyshev.k <- function(rt_prob=0.05){k = sqrt(1 / rt_prob)}
k <- Chebyshev.k(0.05)
k
Chebyshev.max <- function(means, stds, rt_probs){
```

```
  k <- Chebyshev.k(rt_probs)
  theoretical.max <- means + k*stds
  return(theoretical.max)
}
## Aplicando Función Chebyshev.k
p95.theoretical <- Chebyshev.max(mean.each.hour, sd.each.hour, rt_probs=0.05)
p95.theoretical >= p95.each.hour
##----------##
## Cluster4  #
##----------##
## Selección
dic_datos_train <- select(dt_final[cluster==4],"LLAVE",starts_with("Monto"),"MT", "cluster" )
v_categoricas <- colnames(dic_datos_train )
for (j in v_categoricas){
  dic_datos_train[,eval(j):= as.numeric(get(j))]}
str(dic_datos_train )
## Estadisticos
mean.each.hour <- sapply(dic_datos_train[,c(2:14)], mean)
sd.each.hour <- sapply(dic_datos_train[,c(2:14)], sd)
mean.each.hour
sd.each.hour
p95.each.hour <- sapply(dic_datos_train[,c(2:14)], function(x) quantile(x, probs = 0.95))
p95.each.hour
## Creando Función Chebyshev.k
Chebyshev.k <- function(rt_prob=0.05){k = sqrt(1 / rt_prob)}
k <- Chebyshev.k(0.05)
k
Chebyshev.max <- function(means, stds, rt_probs){
  k <- Chebyshev.k(rt_probs)
  theoretical.max <- means + k*stds
  return(theoretical.max)
}
## Aplicando Función Chebyshev.k
p95.theoretical <- Chebyshev.max(mean.each.hour, sd.each.hour, rt_probs=0.05)
p95.theoretical >= p95.each.hour
```