

Segmentación de clientes en un fondo de empleados para identificar los riesgos de lavado de activos y financiación del terrorismo

Customer segmentation in an employee fund for identify risks of money laundering and terrorism financing

Nina Yohana Castro Betancur
nycastro@ulibertadores.edu.co
Estudiante Especialización Estadística Aplicada

Mónica Julieth Castro Betancur
mjcastrob@ulibertadores.edu.co
Estudiante Especialización Estadística Aplicada

Resumen

En la actualidad la superintendencia financiera de Colombia cumple con las disposiciones del gobierno nacional al regular a las entidades financieras con la normatividad sobre el sistema de administración de riesgo de lavado de activos y financiación del terrorismo, sin embargo como solo se aplicaba en el sector financiero, no existe una manera eficiente de identificar esas mismas señales de alerta de lavado de activos y financiación del terrorismo en las organizaciones de economía solidaria, recordando que estas se constituyen para realizar actividades que se

caracterizan por la cooperación, ayuda mutua, solidaria y autogestión de sus asociados, por esta razón este escrito desarrolla una propuesta metodológica conformada por un análisis descriptivo que permite identificar la calidad de la información de los asociados en un periodo de tiempo comprendido entre enero y julio de 2020, seguido a esto el análisis de componentes principales para observar la correlación que existe entre las variables, luego la revisión y el análisis del número grupos con la función NbClust que contiene 30 índices que determinan el número óptimo de clúster y propone usar el mejor esquema de agrupación; a partir de este resultado se aplica el algoritmo k means con el número de grupos identificados, con esto se logra la segmentación de datos e identificación de características similares en los grupos de asociados.

Palabras clave: Organizaciones de economía solidaria, lavado de activos, financiación del terrorismo, k means, análisis de componentes principales, clúster.

Abstract

At present, the financial superintendence of Colombia complies with the provisions of the national government by regulating financial entities with the regulations on the risk management system of money laundering and financing of terrorism, however, as it was only applied in the financial sector, there is no efficient way to identify those same warning signs of money laundering and financing of terrorism in the organizations of solidarity economy, remembering that these are constituted to carry out activities that are characterized by cooperation, mutual help, solidarity and self-management of its members, for this reason this paper develops a methodological proposal consisting of a descriptive analysis that allows identifying the quality of the information of the members in a period of time between January and July 2020, followed by the analysis of main components to observe the correlation that exists between the variables, then the review and analysis of the number of groups with the NbClust function that contains 30

indexes that determine the optimal number of cluster and proposes to use the best scheme of grouping; From this result the k means algorithm is applied with the number of identified groups, with this the segmentation of data and identification of similar characteristics in the groups of associates is achieved.

Keywords: Solidarity economy organizations, money laundering, financing terrorism, k means, principal component analysis, cluster

Introducción

En este artículo se trabajará un modelo estadístico de segmentación de asociados en entidades de economía solidaria, las cuales son organizaciones sin ánimo de lucro, que se constituyen para realizar actividades como la cooperación, ayuda mutua, solidaria y autogestión de sus asociados, desarrollando actividades para un beneficio particular y general, como lo son fondos de empleados, cooperativas entre otros definidos en la ley 454 de 1998 de la República de Colombia (Pública, 1998). Se considera que por ser empresas asociativas de derecho privado corren un mayor riesgo de favorecer el lavado de activos dando lugar a organizaciones criminales que buscan dar apariencia de legalidad a los recursos generados de sus actividades ilícitas. Para controlar este tipo de actividades se tiene el El SARLAFT (Sistema de Administración del Riesgo de Lavado de Activos y Financiación del terrorismo), cuya norma está contenida en la circular Jurídica de la Superintendencia Financiera de Colombia (Infolaft, 2020), la cual se desarrolla mediante la aplicación sistemática de políticas, procedimientos y prácticas de la administración a las tareas de establecer el contexto, identificar, medir, controlar y monitorear los riesgos de Lavado de Activos/Financiación del Terrorismo.

Para contribuir a la mitigación del riesgo del lavado de activos y financiación del terrorismo se pretende trabajar el método estadístico de segmentación de datos, el cual debe tener en cuenta los hechos, situaciones, cuantías, indicadores cuantitativos y cualitativos, razones financieras para inferir oportuna y/o prospectivamente la posible existencia de actividades ilícitas que apunten al lavado de activos o financiación del terrorismo y que a su vez permita gestionar los riesgos de actividades delictivas o terroristas, en este caso un fondo de empleados. Por tal motivo se pretende resaltar los diferentes mecanismos de control y etapas que se deben implementar en las organizaciones solidarias vigiladas y con esto se implementen buenas prácticas de vigilancia, regulación, supervisión y prevención de los bienes y activos que se poseen. (UIAF, 2015)

Es necesario que las organizaciones de economía solidaria como los fondos de empleados apliquen un método estadístico eficiente que permita segmentar factores de riesgo y generar alertas de lavado de activos y financiación del terrorismo de manera temprana. Por esta razón se pretende encontrar a través de la segmentación de datos un modelo estadístico que permita verificar que controles son los adecuados y que cumpla con la primera etapa del SARLAFT; primera etapa: identificación; Segmentar los factores de riesgo y establecer modelo de segmentación, para efectos de esta investigación se aplicara solo la primera etapa ya que solo se identificarán los riesgos mediante un método estadístico con el fin de identificar señales de alerta de manera temprana las cuales pueden ser movimientos bancarios inusuales entre otros, es de aclarar que el SARLAF es una norma vigente en la Circular Externa N.º 04 27-01-2017 que consta de cuatro etapas: identificación, medición de la probabilidad y el impacto del riesgo de LA/FT, control del riesgo de LA/FT y monitoreo del riesgo de LA/FT (Pública, 1998).

Referentes Teóricos

Las bases teóricas para este artículo se encuentran enmarcadas en: Sistema de administración del riesgo de lavado de activos y financiación del terrorismo (SARLAFT), Aprendizaje no supervisado y K-means.

Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT)

Sistema “2. m. Conjunto de cosas que relacionadas entre sí ordenadamente contribuyen a determinado objeto.” – RAE.

Administración: “3. tr. Ordenar, disponer, organizar, dirigir algo” - RAE.

Lavado de Activos: Proceso mediante el cual organizaciones criminales buscan dar apariencia de legalidad a los recursos generados de sus actividades ilícitas – UIAF.

Financiación del Terrorismo: Es el financiamiento de actos terroristas, de terroristas y organizaciones terroristas.

Riesgo: La posibilidad de que suceda algo que tendrá un impacto sobre los objetivos de la entidad. Se le mide en términos de probabilidad e impacto – NTCISO 31000.

Sistema de administración de riesgos: la aplicación sistemática de políticas, procedimientos y prácticas de administración a las tareas de establecer el contexto, identificar, medir, controlar y monitorear los riesgos de LA/FT de la entidad solidaria. Su finalidad es que las entidades financieras gestionen el riesgo de ser utilizadas para dar apariencia de legalidad a activos provenientes de actividades delictivas.

De igual manera, su propósito es evitar que tales entidades sean usadas como instrumento para la canalización de recursos hacia la realización de actividades terroristas.

Etapas del SARLAFT

Las etapas consisten en identificar, medir, controlar y hacer monitoreo de los riesgos de lavado de activos y financiación del terrorismo. Los elementos que soportan este propósito son ocho: políticas, procedimientos, documentación, estructura organizacional, órganos de control, infraestructura tecnológica, divulgación de la información y capacitación.

En 2008 la Superintendencia Financiera decidió actualizar y mejorar el sistema anterior, conocido como Sistema Integral para la Prevención del Lavado de Activos (SIPLA) y crear uno nuevo que se denomina SARLAFT.

Este nuevo sistema se creó mediante la expedición de la circular 22 de 2007, la cual modificó la Circular Básica Jurídica de la antigua Superintendencia Bancaria hoy Superintendencia Financiera que establecía los controles relativos al lavado de activos.

SARLAFT no es un sistema único, aquí es importante aclarar que no existe un SARLAFT tipo o modelo. Cada entidad tiene que crear, desarrollar y perfeccionar su propio SARLAFT o sistema de gestión del riesgo de lavado de activos y financiación del terrorismo. Por tal motivo habrá tantos SARLAFT como entidades que apliquen esta norma de la Superintendencia Financiera.

Dicho esto, SARLAFT debe entenderse desde el punto de vista legal como una norma de la autoridad reguladora, un estándar mínimo que debe ser desarrollado y una recopilación de guías y mejores prácticas que se elevan a una norma de naturaleza obligatoria.

Pero, desde el punto de vista interno, el SARLAFT se desarrolla y se personaliza para volverse el SARLAFT de la entidad, en este punto comienza a ser un sistema vivo y dinámico con recursos, presupuesto y responsables

Aprendizaje no supervisado

Aprendizaje no supervisado es donde se dispone únicamente de valores de X, esto es, por ejemplo, que no hay etiquetas de clase que identifiquen las observaciones, agrupamiento o clustering en estos casos, uno de los objetivos del análisis consiste en investigar si las observaciones pueden ser organizadas en grupos o clúster “homogéneos”.

Es decir, donde las observaciones pertenecientes a un mismo grupo sean, en algún sentido, más similares o próximas que observaciones de distintos grupos. A diferencia de los problemas de clasificación entonces, la (posible) estructura de grupos es desconocida a priori, incluyendo el número de clases o clúster (Laura Aspirot, 2013).

Análisis clúster K-means para agrupación de datos

El análisis clúster k-means es un método comúnmente utilizado, que permite crear una clasificación, descubrir la estructura de las categorías en la que se encajan las observaciones, implica “particionar un conjunto de datos dado en un conjunto de k grupos (es decir, k agrupamientos), donde k representa el número de grupos pre especificados, clasifica los objetos en varios grupos (es decir, grupos), de manera que los objetos dentro de los mismos grupos son lo más similares posible” (Kassambara, 2017 p.36)

Figura 1

Formula k-means

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

x_i : Diseñar un punto de datos perteneciente al clúster.

μ_k : Es el valor medio de los puntos asignados al clúster

Nota: Figura tomada de: (SALAZAR, 2019)

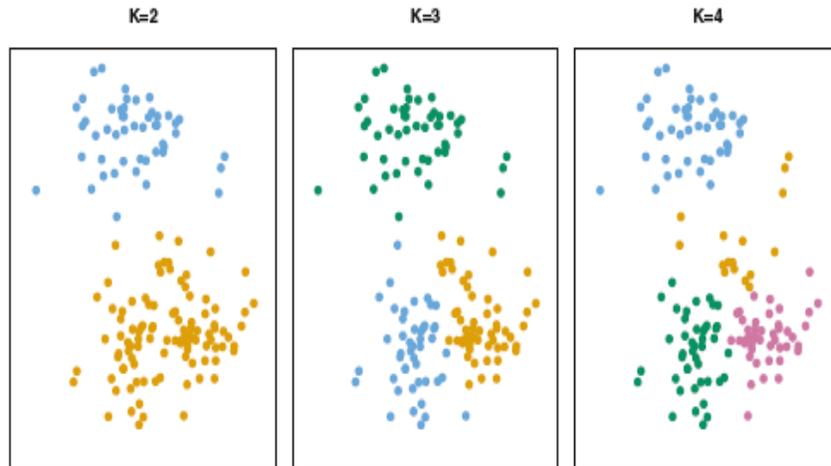
El análisis de conglomerados de K medias es especialmente útil cuando se dispone de un gran número de casos. Existe la posibilidad de utilizar la técnica de manera exploratoria, clasificando los casos e iterando para encontrar la ubicación de los centroides, o solo como técnica de clasificación, clasificando los casos a partir de centroides conocidos y suministrados por el usuario. Cuando se utiliza como técnica, es habitual que el usuario desconozca el número idóneo de conglomerados, por lo que es conveniente repetir el análisis con distinto número de conglomerados y comparar las soluciones obtenidas. (“Análisis de Conglomerados”, s.f)

Algoritmo de K –medias

El primer paso cuando se utiliza la agrupación en clústeres de k-medias es indicar el número de clústeres (k) que se generará en la solución final.

Figura 2

Representación gráfica del algoritmo K-medias propuesto por MacQueen en el año 1967



Nota: Figura tomada de: Universidad de la Plata (2016)



El algoritmo comienza seleccionando aleatoriamente k objetos del conjunto de datos para que sirvan como centros iniciales para los clusters. Los objetos seleccionados también se conocen como medios de clúster o centroides.

A continuación, cada uno de los objetos restantes se asigna a su centroide más cercano, donde más cercano es definido usando la distancia euclidiana entre el objeto y el grupo media. Este paso se denomina "paso de asignación de grupos". Tenga en cuenta que, para utilizar la correlación la distancia, los datos se ingresan como puntuaciones z .

Después del paso de asignación, el algoritmo calcula el nuevo valor medio de cada grupo.

El término "actualización del centroide" del clúster se utiliza para diseñar este paso. Ahora que los centros se han recalculado, cada observación se vuelve a comprobar para ver si podría estar más cerca de un grupo diferente. Todos los objetos se reasignan nuevamente usando el clúster actualizado medio.

Los pasos de actualización de centroide y asignación de conglomerados se repiten iterativamente hasta que las asignaciones de grupos dejan de cambiar (es decir, hasta que se logre la convergencia).

El algoritmo de K-means se puede resumir de la siguiente manera:

1. Especifique el número de clústeres (K) que se crearán (por el analista)
2. Seleccionar aleatoriamente k objetos del conjunto de datos como centros o medias del conglomerado inicial
3. Asigna cada observación a su centroide más cercano, basado en el Euclidean distancia entre el objeto y el centroide

4. Para cada uno de los k grupos, actualice el centroide del grupo calculando los nuevos valores medios de todos los puntos de datos del conglomerado. El centroide de un grupo es un vector de longitud p que contiene las medias de todas las variables para las observaciones en el k -ésimo grupo; p es el número de variables.

5. Minimice iterativamente el total dentro de la suma del cuadrado. Es decir, repita los pasos 3 y 4 hasta que las asignaciones de clústeres dejen de cambiar o el número máximo de iteraciones. De forma predeterminada, el software R usa 10 como valor predeterminado para el número máximo de iteraciones. (Kassambara, 2017).

Metodología

Se realizó una investigación de tipo descriptivo exploratorio de corte longitudinal, con enfoque cuantitativo, donde el objetivo es identificar grupos homogéneos y heterogéneos entre segmentos a partir de variables cuantitativas de los asociados en un fondo de empleados en el periodo de enero a Julio de 2020 para detectar lavado de activos y/o financiación del terrorismo.

Con una población inicial de 2445 asociados de un fondo de empleados, se realizó una primera verificación de los datos para determinar la calidad de la información, la data presentaba datos ausentes en las variables de interés por lo tanto se depuró la base reduciéndola a 389 asociados de igual manera se identificaron datos nulos que fueron eliminados, por esta razón la data total quedó finalmente con un total de 376 asociados, con 5 variables cuantitativas (activos, pasivos, patrimonio, consignaciones y retiros) y sus correspondientes transacciones comprendidas entre enero y julio del año 2020. La base se va a trabajar por medio del programa Rstudio versión 4.0

Diseño:

Fase 1: Construcción de la base de datos.

El fondo de empleados entregó una base de datos con las transacciones realizadas por los asociados en el primer semestre del presente año, las variables cuantitativas que se tomaron de la base son aquellas que nos permiten observar si existen transacciones inusuales como movimientos por encima de su perfil transaccional, constitución de depósitos y/o ahorros por encima de su capacidad económica (ingresos reportados), transacciones que no se relacionan con su actividad financiera, relaciones sospechosas con partes no relacionadas, transferencias en ubicaciones geográficas sin ninguna actividad económica entre otras, que hicieron los asociados en el periodo de tiempo mencionado y que puedan generar señales de alerta por ser movimientos inusuales.

Fase 2: Análisis descriptivo de la base de datos

Inicialmente se cuenta con una base de datos organizada por fechas y productos de 2445 asociados del fondo de empleados, para realizar la investigación se organiza la base por código de asociado, producto y tipo de movimiento.

Al ejecutar el programa se analizó que las variables de referencia más importantes como activo, pasivo y patrimonio se encuentra que más del 80 % de los datos no cuenta con esa información, por esta razón para tener una mayor precisión se toma una muestra de 389 asociados que cumpla con toda la información requerida.

Para el análisis descriptivo se utilizó el programa R estudio versión 4.01, se calcularon las medidas de tendencia central de las variables cuantitativas relacionadas con consignaciones y retiros, de estas treinta variables se seleccionaron las más representativas las cuales son aquellas que tienen transacciones más altas como se puede observar en la Figura 3.



Figura 3.

Monto máximo de consignaciones y retiros por variable,

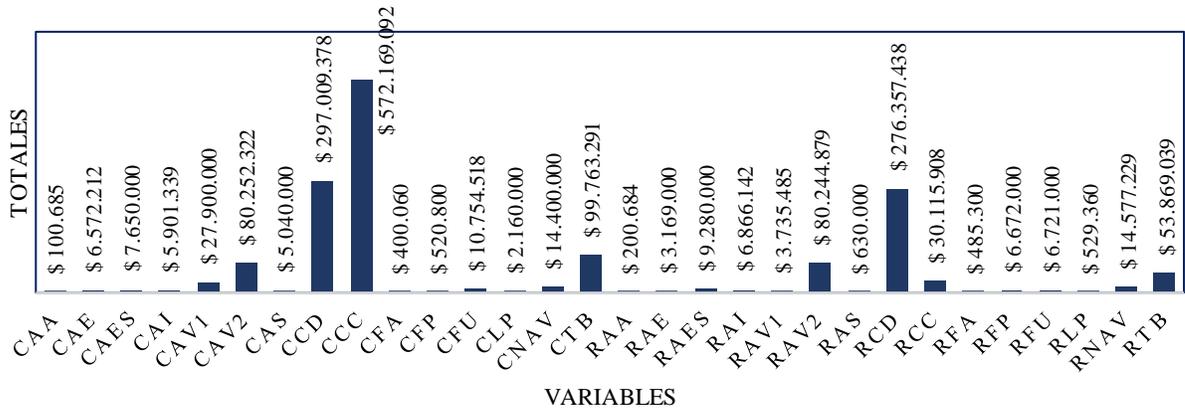


Figura 4

Monto máximo de consignaciones por variable

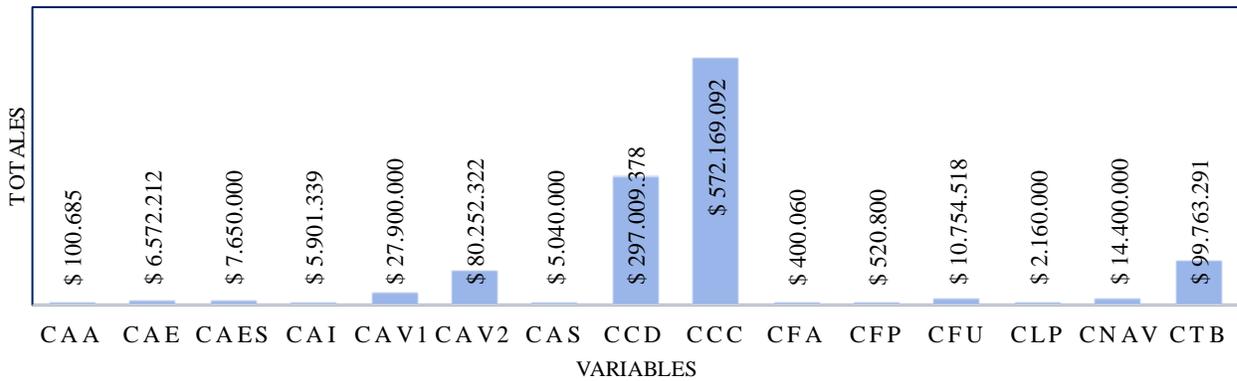
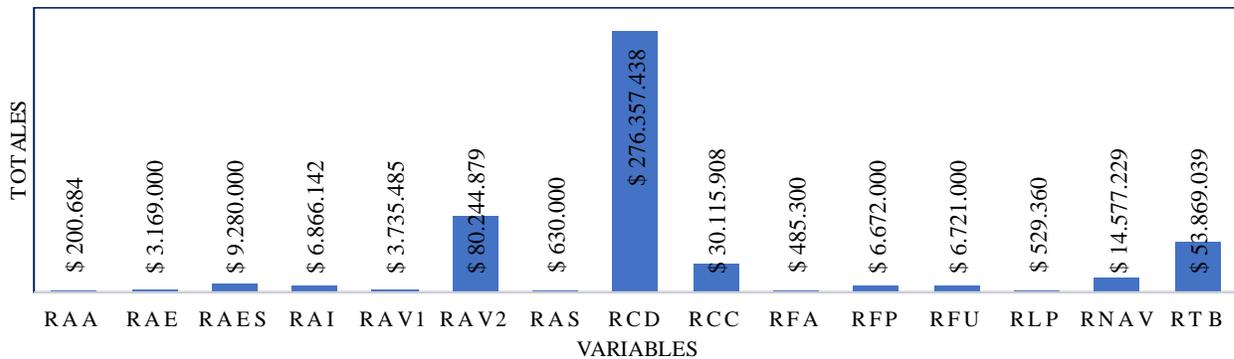


Figura 5.

Monto máximo de retiros por variable



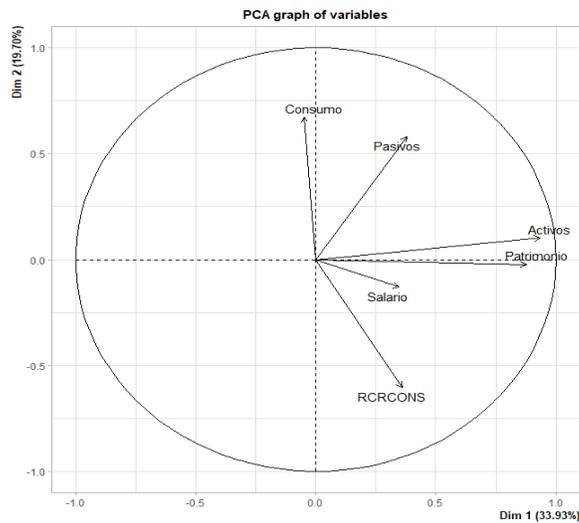
Fase 3: Análisis de componentes principales ACP

Finalmente, con las variables que se va a trabajar son consignaciones de crédito de consumo, retiros del producto CDT, pasivo, activo y patrimonio.

A continuación, se realiza un análisis de componentes principales con el fin de hacer una exploración para identificar aquellas variables que se encuentran correlacionadas.

Figura 6

Análisis de componentes principales: Biplot



La Figura 6 corresponde a una correlación circular en donde se observa que la primera dimensión presenta un valor propio de 33.93% y en la segunda dimensión un valor propio de 19,7%, es decir, un porcentaje de inercia acumulada de 53.63%. Se identifica que las variables con mayor contribución son Patrimonio y Activos, y presentan una alta correlación en la

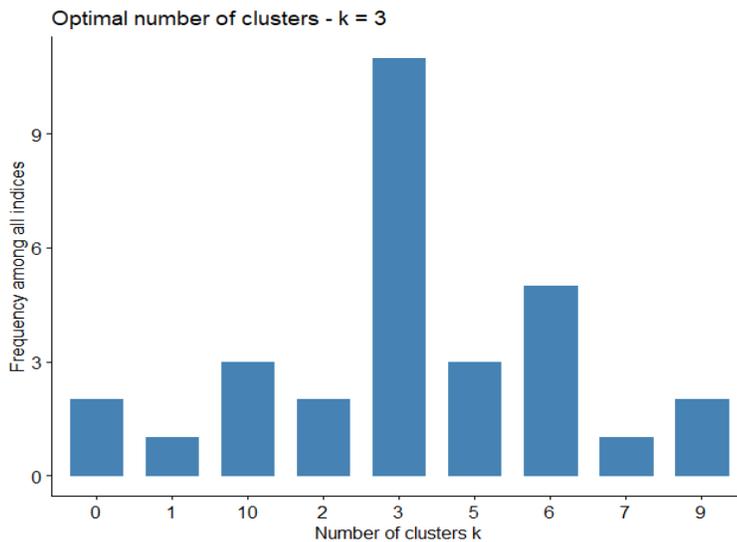
dimensión 1, con relación a las consignaciones y pasivos presentan una contribución moderada y una correlación alta en la dimensión dos. Finalmente, la variable Salario no contribuye a ninguna de las dimensiones y la variable retiros presenta una correlación alta en la dimensión dos.

Fase 4: K-means:

Se determina la cantidad de clúster a partir de un análisis de los índices de agrupamiento, como resultado 11 índices indicaron que la cantidad óptima de clúster es de 3 como se muestra en la Figura 7

Figura 7

Número óptimo de Clúster.



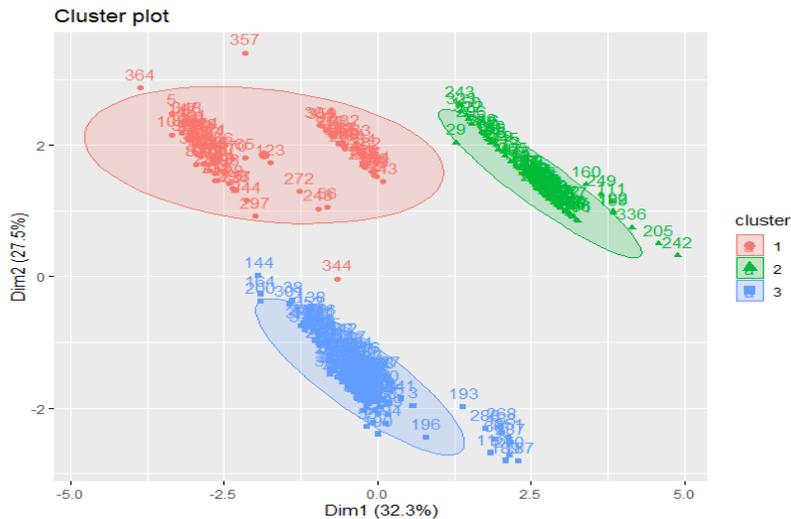
Fase 5: Clúster

Se realizó un análisis de conglomerados (Clúster) con 376 asociados cuyo objetivo es agrupar datos homogéneos y heterogéneos entre segmentos a partir de variables cuantitativas de los asociados de un fondo de empleados en el periodo de enero a Julio 2020 para detectar lavado de

activos y financiación del terrorismo, para ello se utilizó el método de agrupamiento no supervisado K medias. Por otro lado, el método de K medias permitió definir conglomerados de forma que la variación total dentro del clúster fuera mínima. Por este método los grupos conformados se pueden observar en la Figura 8.

Figura 8.

Clusplot



Resultados

El grupo 1 está formado por 94 personas es decir el 25% de la muestra estos asociados comparten valores bajos para las variables de patrimonio, activos y consignaciones, sin embargo, hay asociados que tienen salarios menores a \$3 millones y retiros de \$122 millones.

El grupo 2 está formado por el volumen más alto de personas 195, es decir el 52% de la muestra, en esta clasificación las personas se agrupan por: valores medios para la variable

salario, y montos de consignaciones donde el más alto es por \$572 Millones, con un salario de 5 millones.

El grupo 3 está formado por el volumen más bajo de personas, 87 que equivale al 23% de la población, según la clasificación homogénea de grupos, estas personas comparten: valores altos para las variables salarios, activos, patrimonio y retiros, pero no reflejan montos en consignaciones, en este clúster se encuentra el monto máximo de retiros y salario de toda la muestra.

Conclusiones

De acuerdo con la Etapa 4 del SARLAFT identificación, se aplicó el método de segmentación de factores de riesgo con el algoritmo k medias, en línea con el objeto de estudio el resultado de esta agrupación determinó 3 grupos así: Grupo 1 corresponde al 25% de la muestra, contiene salarios menores a \$3 millones y retiros hasta por \$122 millones, el grupo 2 está conformado por el 52% de la muestra, con montos de consignaciones hasta por \$572 Millones y salarios sobre los 5 millones y el grupo 3 que contiene el volumen más bajo de personas, el 23% se encuentra el monto máximo de retiros y salarios de toda la muestra, debido a que en este tipo de entidades de economía solidaria una medida importante para caracterizar homogeneidad en los grupos son las consignaciones por grandes cuantías hechas por los asociados, consideramos que el grupo para revisar al detalle es el segundo debido al monto elevado de consignaciones y salarios medios, ya que esto puede significar el pago anticipado de créditos y es considerado como alerta por el origen del dinero. Si bien este estudio constituye una forma para segmentar datos e identificar señales de alerta, también es importante asociar estas clasificaciones a datos relevantes como la ubicación geográfica del asociado, edad y antigüedad

en el fondo para hacer análisis más integrales acerca de la existencia de dinero ilícito en las transacciones realizadas.

Referencia Bibliográficas

Universidad Carlos III de Madrid (s.f). Análisis de Conglomerados (I): El procedimiento Conglomerados K-medias
<http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/GuiaSPSS/21conglk.pdf>

Gutierrez, E. A. (2017). Universidad Católica. Obtenido de
<https://repository.ucatolica.edu.co/bitstream/10983/15329/1/Trabajo%20de%20grado.pdf>

Cámara de Comercio de Honda. Guaduas y Norte del Tolima (2015). Preguntas Frecuentes.
<https://www.camarahonda.org.co/wpcontent/uploads/2015/06/PREGUNTASFRECUEENTES-2.pdf>

INFOLAFT. (2020). Lozano Consultores, Financiero/fraude. Obtenido de
<https://www.infoaft.com/que-es-sarlaft/>

Kassambara, A. (2017). Multivariante Analysis I.

Aspirot, L., Castro, S. (2013). Universidad de la Republica. Obtenido de
http://www.iesta.edu.uy/wpcontent/uploads/2014/05/Escueladeverano_RegionalNorteSalto_2013_PresentacionNoSupervisado_Aspirot_Castro.pdf

Pública, D. A. (1998). Departamento Administrativo de la Función Pública. Obtenido de
<https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=3433>

Salazar, M. V. (2019). Repositorio Universidad Los Libertadores, Obtenido de
repository.libertadores.edu.co/bitstream/handle/11371/2017/Cely_Mónica_Galeano_Ingrid_2019.pdf?sequence=1&isAllowed=y

UIAF. (2015).Unidad de Informacion y Análisis Financiero. Lo que se debe saber del lavado de activos y Financiación del Terrorismo obtenido de :
https://www.uiaf.gov.co/sala_prensa/publicaciones/documentos_uiaf/lo_debe_saber_lavado_activos_27734



MacQueen, J. Some methods for classification and analysis of multivariate observations.
Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,
Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif.,
1967. Obtenido de
<http://projecteuclid.org/euclid.bsmsp/1200512992>

Universidad de la Plata, *facultad de informática 2016*, K medias. Obtenido de:
http://163.10.22.82/OAS/Agrupamiento_Kmedias/definicion.html