Modelo Matemático para Estimar el Riesgo de Lavado de Activos por clientes de pequeñas Instituciones Financieras

Dany Alexander Enríquez Sánchez

Universidad del Cauca

Facultad de Ciencias Naturales, Exactas y de la educación

Programa de Matemáticas

Popayán

2019

Modelo Matemático para Estimar el Riesgo de Lavado de Activos por clientes de pequeñas Instituciones Financieras

Propuesta de Trabajo de Grado

Modalidad Practica Empresarial, presentado como requisito parcial para optar título de Matemático

Dany Alexander Enríquez Sánchez

Director

Dr. Yilton Riascos Forero

Universidad del Cauca

Facultad de Ciencias Naturales, Exactas y de la educación

Programa de Matemáticas

Popayán

2019

Nota de aceptación

_____

_____

_____

_____

Director: _____

Dr. Yilton Riascos Forero

Jurado: _____

Mg. Edwin Rengifo

Jurado: _____

Mg. Alejandro Delgado Amen

Lugar y fecha de sustentación: Popayán, 4 de octubre de 2019

CONTENIDO

INTRODUCCION. 5
PLANTEAMIENTO DEL PROBLEMA 8
OBJETIVOS 11
GENERAL 11
ESPECÍFICOS 11
METODOLOGIA 12
MARCO TEORICO. 18
RESULTADOS. 35
CONCLUSIONES Y RECOMENDACIONES 53
BIBLIOGRAFÍA 54

INTRODUCCION.

El lavado de activos y la financiación al terrorismo (LA/FT) es una actividad delictiva reconocida a nivel internacional, siendo el grupo de acción financiera internacional (GAFI) el encargado de dar estándares y medidas legales, regulatorias y operativas para combatir esta problemática a nivel internacional. Para Colombia existe la Unidad de Información y Análisis Financiero (UIAF), adscrita al Ministerio de Hacienda de Colombia, que con la ayuda de entidades reportantes y fuentes abiertas, previene y detecta posibles operaciones que involucren el LA/FT (UIAF, 2017).

La superintendencia de la economía solidaria (SUPERSOLIDARIA) entidad adscrita al ministerio de Hacienda de Colombia, encargada de regular el sector solidario, en función de sus deberes legales, debe velar porque sus entidades vigiladas, caso particular las cooperativas, adopten el Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT) exigido en la circular externa 04 de 2017 de la Supersolidaria (Supersolidaria, 2017).

Dentro de las muchas componentes que exige la circular externa 04 de 2017, centraremos la atención en el párrafo 2.2.2.4.2, segmentación de los factores de riesgo, que sugiere “segmentar cada uno de los factores de riesgo de acuerdo con las características particulares de cada uno de ellos, garantizando homogeneidad al interior de los segmentos y heterogeneidad entre ellos, según la metodología que previamente haya establecido la organización” (Supersolidaria, 2017, p. 5), teniendo en cuenta como mínimo los siguientes aspectos: clientes, productos, jurisdicción, canales. Estos cuatro aspectos le permiten al experto indagar sobre donde pueden originarse actividades sospechosas.

La Cooperativa del Departamento del Cauca (CODELCAUCA) es una entidad perteneciente al sector solidario, con la misión de satisfacer las expectativas de los asociados de la región, y con el compromiso de cumplir con las normas de calidad, en pro de proteger a sus asociados y la reputación de la institución, por lo que se encuentra en proceso de implementación del SARLAFT, para lo cual ha permitido que se realice una intervención metodológica que desde la perspectiva matemática ayude en la adecuada implementación a través de la organización de un proceso de segmentación de los clientes de acuerdo a sus características inherentes.

Dadas sus necesidades, CODELCAUCA opta por usar para el proceso de segmentación una técnica de segmentación estadística, por lo cual se llega a buscar una técnica estadística apropiada que cumpla con los requerimientos exigidos por el SARLAFT, es decir, que cree grupos heterogéneos entre sí, y homogéneos dentro de cada uno de ellos, siendo elegida la técnica del análisis de conglomerados ya que dentro de sus propiedades esta garantizar los requerimientos antes planteados, además resulta óptimo ya que la hipótesis del modelo está en minimizar la varianza dentro de los grupos, y que se representa de la forma: $T = B + W$, donde B representa la varianza entre grupos, y W la varianza dentro de los grupos y T la varianza total; lo anterior se puede alcanzar maximizando B o minimizando W.

Dentro de los algoritmos dispuestos para esta técnica en base a la información suministrada por CODELCAUCA, se optó por tomar el algoritmo K-medias ya que es una técnica que permite crear una segmentación a partir de los datos, usando el criterio de distancias a unos puntos denominados centroides, pertenecientes a cada segmento. Lo cual permite identificar los casos extremos en cada clúster e identificar los casos sospechosos en base a la segmentación.

Como resultados de este procedimiento, se encontró para Codelcauca que, al realizar la segmentación de clientes, se generaron tres clústeres resumidos en la siguiente tabla:

Número de casos en cada conglomerado

Conglomerado	1	889,000
	2	1,000
	3	5,000
Válidos		895,000
Perdidos		,000

De los que se identificaron 3 casos atípicos en el clúster 1, 1 en el clúster 2 y, 5 en el clúster 3, para un total de 9 casos que fueron reportados para posteriormente ser analizados por el experto de la cooperativa.

PLANTEAMIENTO DEL PROBLEMA

El Lavado de Activos (LA) y la Financiación al Terrorismo (FT) son actividades delictivas reconocidas a nivel internacional, entendiéndose por LA como la acción de dar apariencia de legalidad a fondos provenientes de actividades ilícitas y por FT a la acción de financiar con fondos lícitos o ilícitos actividades terroristas, en adelante se denominarán LA/FT (UIAF, 2017). Por ello en 1989 se crea a nivel internacional el Grupo de Acción Financiera Internacional (GAFI) con el fin de dar estándares y medidas legales, regulatorias y operativas para combatir el LA/FT y otras amenazas a la integridad del sistema financiero nacional e internacional.

En particular para Colombia existe la Unidad de Información y Análisis Financiero (UIAF), adscrita al Ministerio de Hacienda y Crédito Público de Colombia, que con la ayuda de entidades reportantes y fuentes abiertas, previene y detecta posibles operaciones que involucren el LA/FT.

El GAFI como organismo internacional encargado del LA/FT está en una constante búsqueda de las formas en que se pueden presentar estos delitos, es por ello que en el documento Estandares Internacionales sobre la lucha contra el Lavado de Activos y el Financimiento del Terrorismo y de la Proliferacion, versión actualizada del 2016 (GAFI, 2018), en el inicio, hace 40 recomendaciones que son los estándares internacionales respaldados por los países miembros. De particular interés se encuentra la recomendación 8, la cual propone que las entidades sin ánimo de lucro son vulnerables ante el LA/FT, caso particular las cooperativas.

Atendiendo a esta recomendación la superintendencia de la economía solidaria (SUPERSOLIDARIA) entidad adscrita al ministerio de Hacienda y Crédito Público de Colombia, encargada de regular el sector solidario, en función de sus deberes legales, debe velar porque las entidades vigiladas

adopten el Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT) exigido en la circular externa 04 de 2017 (Supersolidaria, 2017).

Dentro de las muchas componentes que exige la circular externa 04 de 2017, centraremos la atención en el párrafo 2.2.2.4.2, segmentación de los factores de riesgo, que sugiere segmentar los factores de riesgo de acuerdo con las características particulares de cada uno de ellos, garantizando homogeneidad al interior de los segmentos y heterogeneidad entre ellos, según la metodología que previamente haya establecido la organización, teniendo en cuenta como mínimo los siguientes aspectos:

Clientes: busca crear un perfil de los asociados actuales y futuros sobre sus transacciones para luego verificar la información aportada por estos, y así detectar operaciones sospechosas.

Productos: permite establecer diferencias de los asociados entre los diferentes tipos de productos que ofrece la cooperativa.

Canales de distribución: vías de acceso con el usuario.

Jurisdicción: permite tomar controles sobre las zonas más vulnerables ante LA/FT. Estos cuatro aspectos le permiten al experto indagar sobre donde pueden originarse actividades sospechosas.

Siendo de particular interés para objeto de este trabajo, la segmentación de los clientes, que debe tener como mínimo las siguientes variables: actividad económica, volumen de transacciones, ingresos, egresos, patrimonio, ya que a partir de esta información se puede crear un perfil financiero del asociado.

Por esta razón se recurre a la herramienta estadística de análisis clúster para realizar este proceso, ya que a partir de esta información se pueden establecer

criterios que permitirán identificar, controlar y monitorear a nivel interno y externo el LA/FT.

La Cooperativa del Departamento del Cauca (CODELCAUCA) es una entidad perteneciente al sector solidario, con la misión de satisfacer las expectativas de los asociados de la región, y con el compromiso de cumplir con las normas de calidad, en pro de proteger a sus asociados y la reputación de la institución, por lo que se encuentra en proceso de implementación del SARLAFT, para lo cual ha permitido que se realice una intervención metodológica que desde la perspectiva matemáticas ayude en la adecuada implementación a través de la organización de un proceso de segmentación de los clientes.

Con base en las condiciones antes señaladas, el problema que guía esta intervención metodológica se puede sintetizar en la siguiente pregunta:

¿Cuál es el modelo de segmentación que mejor se ajusta a las características de los datos de CODELCAUCA según las exigencias del SARLAF?

La hipótesis que se desprende de este problema y que intentaremos probar en este trabajo es la siguiente:

Existe un modelo estadístico de segmentación adecuado a las características de los datos de CODELCAUCA según las exigencias del SARLAF

OBJETIVOS

Con base en el problema planteado y la hipótesis de investigación, los objetivos que determinarán el alcance de este trabajo son los siguientes:

GENERAL

Construir un modelo de segmentación que se ajuste a las características de los datos de CODELCAUCA según las exigencias del SARLAF

ESPECÍFICOS

1. Identificar las condiciones que impone el SARLAF para la construcción del modelo de segmentación de clientes
2. Seleccionar información relativa a los clientes de la cooperativa CODELCAUCA, según los criterios del SARLAF
3. Modelar estadísticamente, a través de tecnicas de segmentación, la información obtenida de le empresa

METODOLOGIA

Con el fin de dar cumplimiento a los objetivos propuestos y dar respuesta a la pregunta planteada, se estableció una metodología que se desarrolló en 4 fases:

FASE 1.

En esta primera instancia se empezaron a dar los primeros acercamientos con la cooperativa para establecer una comunicación entre las dos partes de tal forma que ambos puntos de vista estuvieran en la misma dirección, por lo que nos reunimos de manera periódica, donde tratamos las dudas acerca de la reglamentación que requería la implementación del SARLAFT.

Siguiendo este orden de ideas se participó en seminarios y conferencias como por ejemplo la realizada en mayo de 2017 por ASORIESGO y el seminario taller sobre el SARLAFT realizado el 4 de julio de 2018 en la ciudad de Cali, donde se dieron propuestas para implementar el SARLAFT, siendo de gran importancia ya que permitió identificar qué condiciones mínimas debía cumplir la cooperativa para empezar con la implementación de este modelo, en ambos casos se le dio principal importancia a la segmentación de los clientes pues aseguraban que la información sobre la transaccionalidad y las características de estos, permitiría comprender el comportamiento y así poder ubicarlos en grupos en otras palabras segmentarlos.

Lo próximo fue discutir y sintetizar lo que se extrajo de estos eventos, donde se llegó a la conclusión de sacar una base de datos con la información que exigía la circular además de la información que los expertos de la cooperativa consideraban importante, con el objetivo de poder crear un perfil del asociado y la incidencia a la hora de analizar su participación o no en LA/FT. Es en este momento que se nos hace entrega de una primera base de datos la cual

constaba de los asociados actuales la cual no estaba completa en todas las casillas de información, la cual entramos a analizar de manera independiente para hacer los debidos estudios estadísticos a esta misma para ver que podíamos hacer con esta información.

FASE 2.

Debido al hecho que al analizar la información obtenida en la primera fase se encontraron inconsistencias, las cuales se informaron a la cooperativa, el procesamiento de la información para realizar la segmentación no era posible, además los plazos establecidos por la SUPERSOLIDARIA para la implementación del sarlaft estaban cerca a cumplirse.

Por esta razón la cooperativa, en cumplimiento de sus responsabilidades, se vio en la necesidad de contratar los servicios de profesionales del riesgo especializados en este tema, los cuales entregaron un modelo de segmentación que sería presentado ante la SUPERSOLIDARIA, es en ese instante donde la cooperativa, para asegurar la funcionalidad del modelo, nos solicita una opinión como expertos conocedores de los modelos estadísticos, y bajo un análisis exhaustivo se observó que algunos métodos estadísticos no cumplían con los supuestos teóricos requeridos.

Ello generaba repercusiones a la hora del análisis de los resultados para la persona encargada de revisar la participación de los clientes en LA/FT. Debido a esta situación la cooperativa solicitó intervenir con el fin de salvaguardar sus intereses como empresa responsable del bienestar de sus clientes, y su inversión en el trabajo realizado por los profesionales del riesgo, por lo que se entró a discutir aspectos teóricos con los profesionales destinados a la tarea de construcción de los modelos usados, para lo cual se vio la necesidad de poner en contexto el SARLAFT siendo de vital importancia esta información ya que permitió conocer que el SARLAFT es un modelo integrado que trabaja en

función de los productos y servicios de la cooperativa, además del hecho que se requerían conocimientos por fuera del alcance, que permitieran una mejor síntesis de la información.

Así mismo, de la utilidad de esta, no como en un principio se concibió la idea que en la segmentación de los clientes estaba el mayor trabajo en la implementación del SARLAFT, atendiendo a las sugerencias se optó por generar una nueva base de datos en la que se cumplieran, lo mejor posible, los supuestos teóricos de los métodos usados y por ende un nuevo modelo en base a esta información, de lo que la cooperativa compartió esta nueva base de datos y los métodos estadísticos usados para la creación de este nuevo modelo, para ser analizados desde el punto de vista estadístico.

Una vez hecho el análisis del nuevo modelo se observó una mejor aplicación de los métodos lo que resulto según palabras del experto de la cooperativa en un modelo mejor ajustado a la realidad de la cooperativa, es decir va en concordancia a los procedimientos que se aplicaban antes de usar el modelo.

FASE 3.

Teniendo en cuenta el trabajo conjunto realizado con los profesionales del riesgo en la etapa dos, se creó una propuesta de segmentación en la que nos basamos en la información usada por los profesionales del riesgo para crear su modelo. El proceso de creación del modelo fue el siguiente:

Primero se analizó la calidad de los datos que se ingresarían al modelo, dando como resultado de este proceso la eliminación de cierta información redundante.

En segunda instancia se tuvo en cuenta la sugerencia hecha por los profesionales del riego en el uso del algoritmo bietapico como método de segmentación, pero fue descartado ya que, al ejecutarlo, se daba el caso de

que algunos clientes quedan ubicados en varios clústeres a la vez, dificultando así el análisis de LA/FT.

Por este motivo proponemos para la creación de un nuevo modelo, a partir de dividir la base de datos en dos, una base con la información relacionada a las características de la cliente conformada por las variables cualitativas, la otra base con la información relacionada a las transacciones de los clientes conformada con las variables cuantitativas. con el objetivo de segmentar por separado estas dos bases para luego relacionar la información producto de las dos segmentaciones, y así desarrollar un perfil del cliente.

En base a lo anterior, se buscaron los algoritmos de segmentación más usados, con el fin de encontrar el más adecuado, destacando dos, el algoritmo bietapico perteneciente a los clúster jerárquicos. se consideró este algoritmo ya que permite clasificar y seleccionar el número de clúster de manera automática dependiendo de los datos, además del hecho de admitir el uso de variables cualitativas, sin embargo se descartó ya que los datos mostraron ser muy homogéneos por lo que los datos tendían a agruparse en un solo clúster lo cual no permitía establecer diferencias entre los distintos clientes, el otro algoritmo considerado fue el algoritmo k-medias el cual permite seleccionar el número de clúster y mediante el uso del concepto de distancia clasifica los datos en los clúster, nos pareció más útil debido a que al seleccionar el número de clúster nos permitiría ver que datos están por fuera del clúster donde se concentran la mayoría de datos, permitiéndonos observar que datos y que características tienen los datos de los otros clúster, concluyendo como el más óptimo dentro de los conglomerados no jerárquicos, el algoritmo k-medias. Que sería usado para segmentar las dos bases de datos.

Para dar inicio al uso del algoritmo k-medias, lo primero fue comprobar el supuesto que las variables usadas para el análisis clúster no deben estar correlacionadas con el fin de lograr su forma más óptima, haciendo un análisis

para cada base de datos, con el fin de ingresar las variables menos correlacionadas en el algoritmo k-medias, empezando con la base de datos asociada a la transacción. Mediante un análisis de correlaciones múltiples de las 30 variables proporcionadas por la cooperativa, se logró concluir que la menor cantidad de variables correlacionadas fueron 11, sin embargo estas 11 variables seguían evidenciando un alto grado de correlación, no se hizo un análisis posterior para garantizar que las variables estuvieran correlacionadas para ingresarlas al algoritmo k-medias ya que en una reunión con los profesionales del riesgo encargados de la segmentación establecieron que estas variables debido a su naturaleza e interpretación para la situación analizada saturaban a modelo de información que no permitía establecer un perfil claro del asociado ya que se presentaban casos de personas con una sola transacción como de personas con 8 transacciones, por esto cualquier intento de segmentar la base de datos de la transacción daría la posibilidad de que una persona estuviera en varios clúster, debido al número de transacciones realizadas, razón por la cual procedimos a usar la base de clientes la cual mediante el mismo análisis de correlaciones múltiples permitió seleccionar de las 20 variables 6 que presentaban menor correlación, aunque del mismo modo que las variables de la transacción, estas seguían presentando relación entre sí, es por ello que para garantizar la no correlación entre las variables procedimos a usar un análisis de componentes principales sobre la base de datos asociada al cliente, elegido este método por la naturaleza cuantitativa de las variables en esta base de datos, obteniendo 15 componentes principales que posteriormente se ingresaron al algoritmo k-medias de donde se obtuvo el modelo de segmentación.

FASE 4.

Como resultado final y las correcciones realizadas, se entrega a la cooperativa la aprobación del modelo establecido por los profesionales del riesgo, así como una propuesta adicional de segmentación realizada con la base de datos de clientes compuesta por 3 clúster mediante el uso del algoritmo k-medias, además de la evidencia de los casos atípicos encontrados con el modelo construido y las recomendaciones a tener en cuenta en pro de mejorar el modelo de los profesionales del riesgo, conforme avance el tiempo, ya que este modelo es dinámico respecto al tiempo.

MARCO TEORICO.

MODELOS DE SEGMENTACION.

Los modelos de segmentación son por definición el proceso de dividir un todo en grupos uniformes más pequeños que tengan características semejantes denominados segmentos, con los supuestos de que cada segmento se comporte de manera homogénea y entre segmentos sean heterogéneos además de estabilidad de segmentos.

Las técnicas de segmentación se pueden clasificar en los siguientes tipos:

Predictivas.

En esta técnica las variables que intervienen en el proceso pueden clasificarse inicialmente como dependientes o independientes, además especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de minería de datos antes de aceptarlo como válido. la aplicación del modelo debe superar las siguientes fases:

- **Identificación objetiva**: a partir de los datos se aplican reglas que permitan identificar el mejor que se ajuste a los datos.
- **Estimación:** proceso de cálculo de los parámetros del modelo elegido para los datos en la fase de identificación.
- **Diagnosis:** proceso de contraste de la validez del modelo estimado.
- **Predicción:** proceso de utilización del modelo para predecir valores futuros de las variables independientes.

Descriptivas.

En esta técnica las variables que intervienen tienen inicialmente el mismo estatus, además no se supone la existencia de un modelo previo para los datos. los modelos se crean automáticamente partiendo del reconocimiento de patrones.

Ad-hoc.

Dentro de las técnicas predictivas están las técnicas ad-hoc la cuales tienen la característica que tanto el número de segmentos como su tamaño o su descripción se establece antes que el estudio se lleve a cabo. En primer lugar, el investigador selecciona alguna base sobre la cual segmentar (beneficio, necesidades, etc.) en segundo lugar el investigador clasifica a los individuos en segmentos de acuerdo con la base elegida y estudia su relación con otras variables descriptoras. la experiencia de los responsables y el conocimiento ayudan a la hora de conocer los segmentos importantes.

Post-hoc.

Dentro de las técnicas descriptivas están las técnicas post-hoc las cuales tienen la característica de no conocer inicialmente el número de segmentos ni su tamaño, suele ser habitual realizar una exploración cualitativa para conocer a profundidad la población, y a continuación aplicar un análisis de conglomerados que agrupa los sujetos de acuerdo con la similitud de sus perfiles respecto a algunas variables de segmentación preestablecidas. Esta segmentación se denomina optima ya que permite determinar cuáles son los segmentos con mayor homogeneidad interna y heterogeneidad entre ellos.

Como podemos notar, las técnicas **post-hoc** por su descripción se ajustan a lo que describe la circular externa 04 de 2017 como proceso de segmentación, basado en el análisis de perfiles y un modelo creado a partir de los datos.

De manera más precisa aplicaremos la técnica de clasificación post-hoc análisis clúster o de conglomerados, debido a que nos interesa que se cumpla lo establecido en la circular externa 04 de 2017, donde sugiere que los segmentos se comporten de manera homogénea entre ellos y de forma heterogénea entre los distintos clúster.

Análisis clúster.

El análisis clúster es un método estadístico multivariante de clasificación automática que a partir de los datos trata de situarlos en grupos homogéneos, no conocidos de antemano, pero sugeridos por la propia esencia de los datos, de manera que los individuos que sean considerados similares sean asignados al mismo clúster, mientras que los considerados distintos sean asignados a clúster distintos.

El análisis clúster o de conglomerados sigue los siguientes principios:

- Es un método estadístico multivariante de clasificación automática de datos
- Revela concentraciones de datos (casos o variables) para su agrupamiento eficiente en un clúster según su homogeneidad.
- El agrupamiento puede realizarse tanto para casos como para variables pudiendo utilizarse variables cualitativas o cuantitativas.
- Los grupos de casos o variables se realizan basándose en la proximidad o lejanía de unos con otras, por lo tanto, es esencial el uso adecuado de distancia.
- Es fundamental que los elementos dentro de un clúster sean homogéneos y lo más diferentes posibles del contenido en otros clústers.
- El número de clusters no es conocido de antemano y los grupos se crean en función de la naturaleza de los datos.

Para trabajar en el análisis clúster es necesario tener presentes los siguientes supuestos:

- Si las variables de aglomeración están en escalas muy diferentes será necesario estandarizar previamente las variables, o por lo menos trabajar con desviaciones respecto a la media.
- Es necesario observar los valores atípicos y desaparecidos porque los métodos jerárquicos no tienen solución con valores perdidos y los valores atípicos deforman las distancias y producen clusters unitarios.
- Es nocivo para el análisis clúster la presencia de variables correlacionadas, de ahí la importancia del análisis previo de multicolinealidad.
- Si es necesario se realiza un análisis factorial previo y posteriormente se aglomeran las puntuaciones factoriales.
- La solución del análisis clúster no tiene por qué ser única, pero no deben encontrarse soluciones contradictorias por distintos métodos.
- El número de observaciones en cada clúster debe ser relevante, ya que en caso contrario puede haber valores atípicos que difuminen la construcción del clúster.
- Los conglomerados deben tener sentido conceptual y no variar mucho al variar la muestra o el método de aglomeración.
- Los grupos finales serán tan distintos como permitan los datos. Con estos grupos se podrán realizar otros análisis: descriptivos, discriminantes, regresión logística, etc.

Existen dos grandes tipos de análisis de clusters: aquellos que asignan los casos a grupos diferentes que el propio análisis configura, sin que unos dependan de otros, se conocen como no jerárquicos, y aquellos que configuran grupos con estructuras arborescente, de forma que los clusters de niveles más bajos van siendo englobados en otros de niveles superiores, se denominan jerárquicos. En otras palabras, los métodos no jerárquicos pueden, a su vez producir clusters disjuntos o sea cada caso pertenece a un único

clúster o por el contrario clusters solapados donde un caso puede pertenecer a varios grupos.

Por lo anteriormente expuesto podemos afirmar que la técnica de segmentación apropiado para CODELCAUCA es una técnica descriptiva-post-hoc-análisis de conglomerados no jerárquica, ya que sigue los lineamientos de la circular externa 04 de 2017 y la información suministrada por la cooperativa se acomoda a los supuestos que exige este método.

Conglomerados no jerárquicos.

la clasificación de todos los casos de una tabla de datos en grupos separados que configura el propio análisis proporciona clústeres no jerárquicos, esta denominación alude a la no existencia de una estructura vertical de dependencia entre los grupos formados, y por consiguiente estos no se presentan en distintos niveles de jerarquía. El análisis precisa que el investigador fije de antemano el número de clusters en que quiere agrupar sus datos.

Como no puede existir un numero definido de grupos o, si existe, generalmente no se conoce, la prueba debe ser repetida con diferente número a fin de tantear la clasificación que mejor se ajuste al objetivo del problema, o la de más clara interpretación.

Los cluster no jerárquicos están indicados para grandes tablas de datos y son útiles para la detección de datos atípicos. Si se elige previamente un número elevado de grupos, superior al deseado, aquellos que contengan un muy escaso número de individuos servirán para detectar casos extremos que podrían distorsionar la configuración.se recomienda hacer el análisis definitivo sin ellos, ya con el numero deseado de grupos para después, opcionalmente,

asignar los atípicos al cluster adecuado que se formó sin influencia del distorsionante.

Matemáticamente un método de clasificación no jerarquizado consiste en formar un numero prefijado K de clases homogéneas excluyentes, pero con máxima divergencia entre las clases. Las K clases o clusters forman una única partición y no están organizadas jerárquicamente ni relacionadas entre sí. La clasificación no jerárquica tiene una estructura matemática menos precisa que la clasificación jerárquica.se han propuesto diversos algoritmos de clasificación no jerárquica, basados en minimizar progresivamente esta varianza.

Supongamos que N es el número de sujetos a clasificar formando K grupos, esto quiere decir que hay $K^N * K!$ formas de agruparlos, respecto a las n variables $X_1, X_2, X_3, \dots, X_n$. Sean W, B, T las matrices de dispersión dentro de los grupos entre grupos y total respectivamente T=B+W donde T no depende de la forma en que han sido agrupados los sujetos, un criterio razonable de clasificación consiste en construir K grupos de forma que B sea máxima o W sea mínima, algunos de estos criterios son:

a) Minimizar traza(W).
b) Minimizar Determinante(W).
c) Minimizar Det(W)/Det(T).
d) Maximizar traza($W^{-1}B$).
e) Minimizar $\sum_{i=1}^{K}\sum_{h=1}^{N_i}(X_{ih}-\overline{X}_i)S_i^{-1}(X_{ih}-\overline{X}_i)$.

los criterios a) y b) tratan de minimizar la magnitud de la matriz W, c) es llamado criterio de Wilks y es equivalente con b) porque Det(T) constante. el criterio d) es llamado criterio de hottelling y el criterio e) representa la suma de las distancias de mahalanobis de cada sujeto al centroide del grupo al que es asignado.

Algoritmo K-medias.

Parte de una configuración arbitraria de grupos con su respectiva media , eligiendo un individuo de arranque de cada grupo y asignando posteriormente cada caso al grupo con media as cercana, y mediante pruebas sucesivas, contrasta el efecto que sobre la varianza residual tiene la asignación de cada configuración de nuevos grupos con sus respectivas medias.se asignan otra vez todos los casos a estos nuevos centroides en un proceso que se repite hasta que ninguna transferencia pueda ya reducir la varianza residual, o se alcance otro criterio de parada un número limitado de pasos de iteración, o simplemente que la diferencia obtenida en los centroides de dos pasos consecutivos sea menor que un valor prefijado.

Como la varianza total es fija, minimizar la varianza residual hace máxima la varianza intergrupos. y puesto que minimizar la varianza residual equivale a conseguir que sea mínima la suma de distancias al cuadrado desde los casos a la media del clúster al que van a ser asignados, la distancia usada es la distancia euclidea.

Matrices de proximidad.

A veces los datos están representados directamente en términos de proximidad (semejanza o afinidad), este tipo de datos puede ser representado por una matriz D de N x N, donde N es el número de objetos, y cada elemento d_{ij} representa la proximidad entre el objeto i y el objeto j.

Esta matriz es el input del algoritmo del clúster. La mayoría de los algoritmos presumen una matriz de desemejanza con enteros no negativos y ceros en la diagonal principal: $d_{ii} = 0 \quad i = 1,2,3,\dots,N$.si los datos originales son tomados como semejantes una adecuada función monótona decreciente puede

convertirlos en desemejantes. Por lo tanto la mayoría de los algoritmos asumen matrices simétricas desemejante, y si la matriz diagonal Des no simétrica va a ser reemplazada por $(D + D^T)/2$.

Desemejanza basada en atributos.

A menudo tenemos medidas x_{ij} para $i = 1,2,3, \dots, N$. Con variables $j = 1,2,3, \dots, P$.(llamadas atributos) dado que en los algoritmos más usados de clúster toman una matriz de desemejanza como input, tenemos que construir primero pares de diferencias entre las observaciones. Los casos más comunes definimos la desemejanza $d_j(x_{ij}, x_{kj})$ entre os valores del atributo j y luego definimos:

$$D(x_i, x_k) = \sum_{j=1}^{p} d_j(x_{ij}, x_{kj}) \qquad (1)$$

Como la desemejanza entre el objeto i y k, la más común entre de las elecciones es la distancia cuadrática

$$d_j(x_{ij}, x_{kj}) = (x_{ij} - x_{kj})^2.$$

In embargo otras elecciones son posibles y pueden llevar a potencialmente diferentes resultados. Para atributos no cuantitativos (ej. Datos categóricos). La distancia cuadrática no es la apropiada. Además, a veces se prefiere darles diferentes pesos a distintos atributos antes de darle el mismo peso como (1).

Las diferentes alternativas para los distintos tipos de atributos:

- Variables cuantitativas. Mediciones de este tipo de variables o atributos son representados con valores reales continuos.es natural definir entre ellos como una función monótona creciente de la diferencia de sus valores absolutos

$$d(x_i, x_j) = l(|x_i - x_j|).$$

Además la función cuadrática $l(u) = u^2$(distancia euclidea),la elección más frecuente es $l(u) = u$ que da origen a la distancia L^1. Una alternativa es que la agrupación está basada en la correlación

$$\rho(x_i - x_k) = \frac{\sum_j (x_{ij} - \overline{x_i})(x_{kj} - \overline{x_k})}{\sqrt{\sum_j (x_{ij} - \overline{x_i})^2 (x_{kj} - \overline{x_k})^2}},$$

Con $\overline{x_i} = \sum_j x_{ij}/p$. Notemos que este promedio es sobre variables, no sobre observaciones. Si al input primero lo estandarizamos, entonces

$$\sum_j (x_{ij} - x_{kj})^2 \propto 2(1 - \rho(x_i - x_k))$$

Por lo tanto, la agrupación basada en correlación (de semejanza) es equivalente a la basada en distancia cuadrática (de desemejanzas).

- Variables ordinales. El valor de este tipo de variables esta mayormente representado por enteros consecutivos, y los valores realizados son considerados un conjunto ordenado. El rango de datos es un tipo especial de datos ordinales. Errores de medición de variables ordinales son generalmente definidos reemplazando su valor original M con:

$$\frac{i - 1/2}{M}, \quad i = 1,2,3, \dots, M$$

En los órdenes prescriptos de sus valores originales. Son entonces tratados como variables cuantitativas en esta escala.

- Variables categóricas. Con un desorden categórico (también llamado nominal) de las variables, el grado de diferencia entre pares de valores tiene que ser definido explícitamente. Si la variable toma M valores distintos, entonces puede ser organizado con una matriz simétrica $M\ x\ M$ con elementos $L_{ij} = L_{ji}, L_{ii} = 0, L_{ij} \geq 0$

La elección más común es $L_{ij} = 1$para todo $i \neq j$, porque la perdida de la igualdad puede ser usada para enfatizar más un error que otro.

Desemejanzas entre objetos.

Ahora definamos un procedimiento que combine los p-individuales atributos de desemejanzas $d_j(x_{ij}, x_{kj}), j = 1,2,3, \dots, p$. Con una media global de desemejanzas $D(x_i, x_k)$ entre dos objetos u observaciones (x_i, x_k) que poseen los valores de atributos respectivos. Esto es casi siempre hecho por medio de una media ponderada

$$D(x_i, x_k) = \sum_{j=1}^{p} w_j \,.\, d_j(x_{ij}, x_{kj}); \sum_{j=1}^{p} w_j = 1 \qquad (2)$$

w_j es el peso asignado al atributo j regulando la influencia relativa de esa variable para determinar la diferencia total entre objetos. Esta elección debe ser basada en la importancia que se le dé a cada variable.

Cabe notar que asignar el mismo peso w_j a todos los valores de cada variable ($w_j = 1, \forall$j), no necesariamente le da a todos los atributos igual influencia. La influencia del atributo j, en el objeto de desemejanzas sobre todos los pares de observaciones del conjunto de datos

$$\overline{D} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{k=1}^{N} D(x_i, x_k) = \sum_{j=1}^{p} w_j \,.\, \overline{d}_j,$$

con

$$\bar{d}_J = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N} d_j(x_{ij}, x_{kj}), \quad (3)$$

el promedio de desemejanza del atributo j. así, la influencia relativa de la variable j. así, la influencia relativa de la variable j es $w_j.\bar{d}_J$, y poniendo $w_j \sim \bar{d}_J$ daría a todos los atributos la misma influencia en la caracterización general entre objetos. Por ejemplo, con p variables cuantitativas y usando la distancia del error cuadrático para cada coordenada, entonces (2) es la cuadrática distancia euclidea

$$D_I(x_i, x_k) = \sum_{j=1}^{p} w_j.(x_{ij} - x_{kj})^2,$$

Entre pares de puntos de R^p, con las variables cuantitativas como ejes. En este caso (3) seria

$$\bar{d}_J = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N} d_j(x_{ij}, x_{kj}) = 2var_j,$$

Donde var_j es la estimación de la muestra de $Var(X_j)$. Así la importancia relativa de cada una de dichas variables es proporcional a la variación en el conjunto de datos. En general poner peso $w_j = \frac{1}{\bar{d}_j}$ a todos los atributos, independiente del tipo, provocar que cada uno de ellos tenga la misma influencia sobre la diferencia total entre los pares de objetos (x_i, x_k).

A pesar de que parezca ser razonable, y es en muchos casos recomendado, puede ser altamente contraproducente. Si el objetivo es segmentar los datos en grupos de objetos similares, todos los atributos no pueden tener la misma

influencia. Algunas diferencias en los valores de los atributos pueden reflejar una mayor desigualdad en el contexto del objeto real del problema.

Si el objetivo es descubrir grupos naturales en los datos, algunos atributos deben demostrar una mayor tendencia a agrupar que otros. A las variables que son más relevantes para separar los grupos, se les deberá de asignar un mayor peso en la definición de desemejanzas entre objetos.

Darles a todos los atributos el mismo peso en este caso produciría una tendencia a ocultar los grupos de puntos donde los algoritmos de clúster no pueden acceder. Aunque las elecciones individuales del atributo para las desemejanzas $d_j(x_{ij}, x_{kj})$ y de sus pesos w_j pueden ser una herramienta adecuada, no hay un sustituto para el pensamiento cuidadoso que se debe tener en contexto de cada problema.

K-medias.

El algoritmo de las k-medias es uno de los más populares algoritmos iterativos, del análisis clúster. Está destinado a situaciones en las cuales todas las variables son del tipo cuantitativo, y la distancia cuadrática Euclidea.

$$d(x_i, x_k) = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2 = ||x_i - x_k||^2$$

es elegida como medida de diferencia. Notemos que los pesos en la distancia Euclidea pueden ser usados redefiniendo los valores x_{ij}.

Los puntos de dispersión pueden ser escritos como

$$W(C) = \sum_{s=1}^{S} \sum_{C(i)=s} \sum_{C(k)=s} \left||x_i - x_k|\right|^2$$

$$= \sum_{s=1}^{S} N_s \sum_{C(i)=s} ||x_i - \bar{x_s}||^2$$

Donde $\bar{x_s} = (\overline{x_{1s}}, \overline{x_{2s}}, \dots, \overline{x_{ps}})$, es el vector de medias asociado con el s-esimo clúster, y $N_k = \sum_{i=1}^{N} I(C(i) = s)$. Así, el criterio es asignar las N observaciones a los K clúster de modo que dentro de cada clúster el promedio de las diferencias de cada observación a la media del clúster, definido por los puntos del clúster, sea mínima.

Consistencia del k-medias.

El análisis de clúster por k-medias prescribe un criterio de cómo partir un conjunto de puntos en k-grupos. Para dividir los puntos $X_1, X_2, \dots, X_n \; de \; R^s$ acordes a este criterio, primero tenemos que elegir los centros de los clúster de manera que minimicen

$$W_n = \frac{1}{n} \sum_{i=1}^{n} \min_{1<j<k} ||X_i - a_j||^2,$$

Donde $||\cdot||$ denota la norma euclidea, luego asignamos cada X_i al clúster cuyo centro esté más cercano, de esta manera, cada centro a_j adquiere un subconjunto C_j de los $X's$ en su clúster. La media de los puntos en $C_j$tiene que ser a_j, sino podemos achicar $W_n$reemplazando a_j por la media de este clúster, si es necesario reasignamos algunos $X's$ a nuevos centros. El criterio es, por lo tanto, equivalente a minimizar la suma de cuadrados dentro de cada clúster.

Asumamos que $\{x_1, x_2, \dots, x_n\}$ en una muestra de observaciones independientes e idénticamente distribuidas con cierta distribución P Vamos a pedir ciertas condiciones que aseguren la convergencia casi segura a los centros de los clúster cuando el tamaño de la muestra crece.

Por las dificultades que puedan surgir por ambigüedades en la asignación de los puntos $X_1, X_2, \dots, X_n$a los centros $a_1, a_2, \dots, a_k$, es ventajoso considerar $W_n$como una función de los conjuntos de centros de clúster y de la medida empírica P_n obtenida de la muestra al colocarle un peso n^{-1}a cada $X_1, X_2, \dots, X_n$. Esto es el problema a minimizar

$$W(A, P_n) = \int \min_{a \in A} ||x - a||^2 P_n(dx),$$

sobre todas las posibles elecciones del conjunto A que contenga k (o menos) puntos. Para cada A la ley fuerte de los grandes números dice que

$$W(A, P_n) \to W(A, P) = \int \min_{a \in A} ||x - a||^2 P_n(dx),$$

Se espera, por lo tanto que A_n, el conjunto de los óptimos centros de la muestra , este cerca de $\bar{A}$, el conjunto de centros que minimizan $W(\cdot, P)$, siempre que $\bar{A}$ este unívocamente determinado, esto implica que hay una asignación $a_{n1}, a_{n2}, \dots, a_{nk}$ de los puntos de A_n, y una asignación de $\bar{a}_1, \bar{a}_2, \dots, \bar{a}_k$ de los puntos de $\bar{A}$, tal que $a_{nl} \to \bar{a}_l$.

Funciones objetivo poblacional y empírica. dada una medida de probabilidad $\vartheta \; en \; R^s$para cada conjunto finito de R^s definimos:

$$\varphi(A, \vartheta) \coloneqq \int \min_{a \in A} ||x - a||^2 \vartheta(dx),$$

$$m_k(\vartheta) \coloneqq inf\{\varphi(A, \vartheta) : A \; contiene \; k \; o \; menos \; puntos\}$$

La función objetivo poblacional corresponderá a tomar $\vartheta = p$ en (4) mientras que la empírica $\vartheta = P_n$.

Para un dado k, el conjunto $A_n = A_n(k)$ de los óptimos centros de la muestra tiene que ser elegido tal que $\varphi(A_n, P_n) = m_k(P_n)$; el conjunto poblacional de los centros $\bar{A} = \bar{A}(k)$ satisface $\varphi(\bar{A}, P) = m_k(P)$, el objetivo es mostrar que $A_n \to \bar{A}$ casi seguramente.

Teorema (consistencia) supongamos que $\int ||x||^2 P(dx) < \infty$ y para cada $j = 1,2, \dots, k$. Hay un único conjunto $\bar{A}(j)$ para el cual $\varphi(\bar{A}(j), P) = m_j(P)$.

Entonces $A_n \to \bar{A}(k)$ casi seguramente y $\varphi(A_n, P_n) \to m_k(P)$ casi seguramente.

La demostración se puede encontrar la tesis “El metodo k-medias-Departamento de Matematicas-universidad de Buenos Aires” (Gimenez, 2010).

Dado los requerimientos exigidos por el método de segmentación k-medias explicaremos algunas técnicas auxiliares para la implementación de este algoritmo con el fin de garantizar que se cumplan los supuestos principales de este.

ANALISIS DE COMPONENTES PRINCIPALES.

El análisis de componentes principales es un método estadístico multivariante, de reducción de la dimensión de una tabla de casos-variables con datos cuantitativos, para obtener otra de menor número de variables, combinación lineal de las primitivas, que se denominan componentes principales con la propiedad de que las componentes principales son independientes.

OBTENCION DE LAS COMPONENTES PRINCIPALES.

En el análisis de componentes principales se dispone de una muestra de tamaño n acerca de p variables $X_1, X_2, \dots, X_p$ inicialmente correlacionadas, para posteriormente obtener a partir de ellas un numero $k \leq p$ de variables correlacionadas $Z_1, Z_2, Z_3, \dots, Z_k$ que sean combinación lineal de las variables iniciales y que expliquen la mayor parte de su variabilidad.

Las componentes principales se pueden representar de manera matricial así:

$$\begin{bmatrix} Z_{11} \\ Z_{12} \\ \vdots \\ Z_{1n} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \dots & X_{p1} \\ X_{12} & X_{22} & \dots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & \dots & X_{pn} \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{bmatrix}$$

Se demuestra que la componente principal h-esima se define como $Z_h = Xu_h$ donde u_h es el vector propio de V(matriz de varianzas) asociado a su h-esimo mayor valor propio suele denominarse también a $u_h$eje factorial h-esimo.

Donde la varianza de la componente h-esima es:

$$V(Z_h) = {u_h}^t V u_h = \lambda_h$$

Es decir, la varianza de cada componente es el valor propio de la matriz V y por ello:

$$\sum_{h=1}^{p} V(X_h) = traza(V)$$

Como V es una matriz real simétrica, es diagonalizable por lo que:

$$\sum_{h=1}^{p} V(X_h) = \sum_{h=1}^{p} V(Z_h)$$

Lo que verifica que la suma de las varianzas de las variables es igual a la suma de las varianzas de las componentes principales. La proporción de la variabilidad total escogida por la componente principal h-esima viene dada por:

$$\frac{\lambda_h}{\sum_{h=1}^{p} \lambda_h} = \frac{\lambda_h}{traza(V)}$$

RESULTADOS.

Fase 1.

Como resultado de las reuniones con la cooperativa, seminarios y conferencias, nos permitió tener una idea sobre lo que es el SARLAFT desde su parte administrativa y el papel que las herramientas estadísticas tienen en el proceso de implementación de este manual, además del hecho de realizar el proceso adaptativo para poder desarrollar nuestro trabajo, conocer internamente el funcionamiento de la cooperativa y adaptarnos a su realidad.

Fase 2.

Como resultado del segundo modelo se obtuvo la siguiente información proporcionada por la cooperativa, resumida en la tabla 1.

Tabla No.1

Nombre	Descripción	Tipo	Escala	Aportado por la cooperativa
Actividad	Describe Actividad económica de la persona	Cualitativa	Nominal	Si
Activo	Total recursos disponibles del asociado	Cuantitativa	Intervalo	Si
Activo actualización	Ultima declaración de los activos	Cuantitativa	Intervalo	Si
Aportes	Dinero que se tiene en la cooperativa para estar asociado	Cuantitativa	Intervalo	Si
Canal	Por donde se recaudó la transacción	Cualitativa	Nominal	Si
Cartera	En caso de tener crédito es el saldo de la deuda	Cuantitativa	Intervalo	Si
Cedula aplicada a terceros juguemos	Cedula de la persona que realiza la transacción por medio de un juguemos	Cualitativa	Nominal	Si
Ciiu	Código establecido por la dian para la actividad económica de la persona	Cualitativa	Nominal	Si
Departamento	Código asignado por el dane para identificar los departamentos	Cualitativa	Nominal	Si

Nombre	Descripción	Tipo	Escala	Aportado por la cooperativa
Edadsicses	Edad del asociado	Cuantitativa	Intervalo	Si
Entrada	Valor en dinero que ingresa a la cooperativa	Cuantitativa	Intervalo	Si
Estrato sicses	Estrato socioeconómico al que pertenece la persona	Cualitativa	Ordinal	Si
f movimiento	Fecha que se realizó la transacción	Cualitativa	ordinal	Si
Gastos	Gastos mensuales declarados por la persona	Cuantitativa	Intervalo	Si
Genero	Sexo de la persona	Cualitativa	Nominal	Si
ingreso actualización	ingreso más reciente declarado por la persona	Cuantitativa	Intervalo	Si
ingresos brutos	Ingresos del asociado libres de impuestos u otras deducciones	Cuantitativa	Intervalo	Si
jurisdiccioncodigodanearchivosicses	Código asignado por el dane para identificar los municipios	Cualitativa	Nominal	Si
Mes	Mes en que se realiza la transacción	Cualitativa	Ordinal	Si
Modalidad	Diferentes conceptos por los cuales la empresa recibe dinero	Cualitativa	Nominal	Si
No. de documento	Tipo de contrato con la cooperativa	Cualitativa	Nominal	Si
Nombre de la persona	Describe el nombre o sigla de la persona	Cualitativa	Nominal	Si
Obligación	Código de pagare	Cualitativa	Nominal	Si
otros ingresos	Ingresos adicionales al salario	Cuantitativa	Intervalo	Si
otros ingresos actualización	Otros ingresos más recientes declarados por la persona	Cuantitativa	Intervalo	Si
Pasivo	Deudas u obligaciones del asociado	Cuantitativa	Intervalo	Si
pasivo actualización	Ultima declaración de los pasivos	Cuantitativa	Intervalo	Si
Patrimonio	Activos menos los pasivos de la persona que realiza la transacción	Cuantitativa	Intervalo	Si
patrimonio actualización	Última actualización del patrimonio entregada por el asociado	Cuantitativa	Intervalo	Si
Pep	Persona políticamente expuesta(personajes públicos)	Cualitativa	Nominal	Si
Producto	Tipo de responsabilidad adquirido con la empresa	Cualitativa	Nominal	Si
Profesión	a que se dedica para generar ingresos	Cualitativa	Nominal	Si
Salario	Dinero que recibe una persona sin considerar ciertas cantidades adicionales	Cuantitativa	Intervalo	Si
Salida	Pago que realiza la cooperativa	Cuantitativa	Intervalo	Si
Sucursal	Sucursal donde se realizó la transacción	Cualitativa	Nominal	Si
tipo de persona	Identifica si la persona es natural o jurídica	Cualitativa	Nominal	Si
tipo de rol	clasifica la persona que realiza la transacción	Cualitativa	Nominal	Si

Nombre	Descripción	Tipo	Escala	Aportado por la cooperativa
tipo documento	Identifica contablemente el tipo de la transacción	Cualitativa	Nominal	Si
total ingresos	Suma de salario y otros ingresos	Cuantitativa	Intervalo	Si
total ingresos actualización	Suma de ingresos actualización y otros ingresos actualización	Cuantitativa	Intervalo	Si
valor cuota	Monto que debe cancelar periódicamente	Cuantitativa	Intervalo	Si
asociado con deuda	Identifica si tiene o no crédito	Cualitativa	Nominal	No
canal controla laft	Si por donde se recibe el dinero hay o no control de LA/FT	Cualitativa	Nominal	No
duración del producto	Tiempo relativo en que adquiere la persona obligaciones con la cooperativa	Cualitativa	Ordinal	No
jurisdicción asociada al conflicto	El lugar de la transacción está asociada al conflicto armado	Cualitativa	Nominal	No
pago aplicado por canal externo.	Si el pago que se efectuó se realizó por medio de un canal externo a la cooperativa	Cualitativa	Nominal	No
producto asociado al Pago de un crédito	Identifica si es un crédito o no	Cualitativa	Nominal	No
producto se recauda en canal externo	El producto adquirido por el asociado se puede pagar mediante una fuente externa	Cualitativa	Nominal	No
recaudo permite pago de terceros	Puede pagar la obligación una persona ajena al compromiso	Cualitativa	Nominal	No
tipo de jurisdicción		Cualitativa	Nominal	no

En resumen, tenemos un total de 30 variables cualitativas donde 4 son ordinales y 26 nominales, además de 20 variables cuantitativas por intervalo.

Fase 3

Con el objetivo de aportar nuestro modelo procedimos de la siguiente forma:

1. **Separación de la base de datos.**

Ya que la base de datos suministrada por la cooperativa estrega información relacionada a las transacionalidad de los asociados en un periodo de 8 meses, se tiene que algunos asociados identificados por su ID aparecen en la base de datos múltiples veces, donde sus características de cliente se mantienen invariantes el número de veces que aparecen estos asociados, como el método de segmentación es de tipo clúster, este tiene como objetivo minimizar la varianza entre los distintos grupos, es por ello que el hecho de repetir los datos del cliente como por ejemplo salario, patrimonio, etc. Tiene una repercusión en la varianza.

Por esto se opta por separar la base de datos en dos:

1.1 Clientes. Aquí se tomarán las características del asociado solamente una vez caracterizado por el ID y las variables a tomar en cuenta son: salario, otros ingresos, total ingresos, total ingresos actualización, aportes, cartera, edadsicses, activo, activo actualización, pasivo, pasivo actualización, patrimonio, patrimonio actualización, ingresos brutos, gastos, que tienen la característica de ser variables de tipo cuantitativas.

1.2 Transacciones. se tomaran las variables asociadas a la transacción sin tener en cuenta que el ID aparezca varias veces las cuales son: Tipo persona, nombre, tipo doc., modalidad, producto, sucursal actividad, Pepsi, pepno, asociado con deuda, profesión, canal, jurisdiccioncodigodanearchivosicses, departamento, estrato sicses, genero, duración del producto, producto se recauda canal externo, producto se recauda canal presencial, producto asociado al pago de un crédito, recaudo permite pago por terceros, jurisdicción asociada al conflicto, tipo de jurisdicción, canal controla LA/FT, tipo de rol, que son variables de tipo cualitativas.

Se conservará el ID en ambas bases para identificar el asociado.

2. Matriz de correlaciones.

Con el fin de verificar el supuesto de independencia entre las variables que se ingresan en el algoritmo k-medias haremos un análisis de correlación múltiple tanto para las variables de cliente como para las variables de la transacción.

2.1 clientes. Al aplicar el método de correlaciones múltiples tenemos que las variables menos relacionadas son 6 resumiendo la información en la matriz de correlaciones resumida en la tabla 2.

		OTROS_INGRESOS	CARTERA	ACTIVO ACTUALIZACION	PATRIMONIO	PATRIMONIO ACTUALIZACION	GASTOS
OTROS_INGRESOS	Correlación de Pearson	1	-,009	,030	,039	,011	,595**
	Sig. (bilateral)		,793	,366	,242	,741	,000
	N	895	895	895	895	895	895
CARTERA	Correlación de Pearson	-,009	1	,130**	,103**	,056	,023
	Sig. (bilateral)	,793		,000	,002	,095	,489
	N	895	895	895	895	895	895
ACTIVO ACTUALIZACION	Correlación de Pearson	,030	,130**	1	,530**	,977**	,054
	Sig. (bilateral)	,366	,000		,000	,000	,103
	N	895	895	895	895	895	895
PATRIMONIO	Correlación de Pearson	,039	,103**	,530**	1	,519**	,037
	Sig. (bilateral)	,242	,002	,000		,000	,266
	N	895	895	895	895	895	895
PATRIMONIO ACTUALIZACION	Correlación de Pearson	,011	,056	,977**	,519**	1	,004
	Sig. (bilateral)	,741	,095	,000	,000		,896
	N	895	895	895	895	895	895
GASTOS	Correlación de Pearson	,595**	,023	,054	,037	,004	1
	Sig. (bilateral)	,000	,489	,103	,266	,896	
	N	895	895	895	895	895	895

**. La correlación es significativa al nivel 0,01 (bilateral).

Tabla 2

Donde se puede evidenciar que al reducir a la cantidad más pequeña de variables que no estén correlacionadas, sigue habiendo un tipo de correlación entre ellas.

2.2 Transacción.

Al realizar el procedimiento de correlaciones múltiples para las variables de la transacción obtuvimos 11 variables que están lo menos relacionadas entre sí, aunque como en el anterior caso se evidencia algún tipo de relación, la información se verá en la tabla 3.

correlaciones		NATURAL	JURIDICA	ESTUDIANTE	PROFESIONAL INDEPENDIENTE	RENTA CAPITAL	PEP NO	ESTRATO 4
NATURAL	Correlación de Pearson	1	-1,000**	0,001	0,002	0,002	-0,004	0,004
	Sig. (bilateral)		0	0,98	0,951	0,94	0,871	0,875
	N	1605	1605	1605	1605	1605	1605	1605
JURIDICA	Correlación de Pearson	-1,000**	1	-0,001	-0,002	-0,002	0,004	-0,004
	Sig. (bilateral)	0		0,98	0,951	0,94	0,871	0,875
	N	1605	1605	1605	1605	1605	1605	1605
ESTUDIANTE	Correlación de Pearson	0,001	-0,001	1	-0,002	-0,002	0,004	,158**
	Sig. (bilateral)	0,98	0,98		0,951	0,94	0,871	0
	N	1605	1605	1605	1605	1605	1605	1605
PROFESIONAL INDEPENDIENTE	Correlación de Pearson	0,002	-0,002	-0,002	1	-0,005	0,01	-0,01
	Sig. (bilateral)	0,951	0,951	0,951		0,854	0,691	0,699
	N	1605	1605	1605	1605	1605	1605	1605
RENTA CAPITAL	Correlación de Pearson	0,002	-0,002	-0,002	-0,005	1	0,012	-0,012
	Sig. (bilateral)	0,94	0,94	0,94	0,854		0,626	0,635
	N	1605	1605	1605	1605	1605	1605	1605
PEP NO	Correlación de Pearson	-0,004	0,004	0,004	0,01	0,012	1	0,026
	Sig. (bilateral)	0,871	0,871	0,871	0,691	0,626		0,306
	N	1605	1605	1605	1605	1605	1605	1605
ESTRATO 4	Correlación de Pearson	0,004	-0,004	,158**	-0,01	-0,012	0,026	1
	Sig. (bilateral)	0,875	0,875	0	0,699	0,635	0,306	
	N	1605	1605	1605	1605	1605	1605	1605
ESTRATO 5	Correlación de Pearson	0,002	-0,002	-0,002	-0,004	-0,005	0,01	-0,01

correlaciones		NATURAL	JURIDICA	ESTUDIANTE	PROFESIONAL INDEPENDIENTE	RENTA CAPITAL	PEP NO	ESTRATO 4
	Sig. (bilateral)	0,951	0,951	0,951	0,881	0,854	0,691	0,699
	N	1605	1605	1605	1605	1605	1605	1605
ESTRATO 6	Correlación de Pearson	0,001	-0,001	-0,001	-0,002	-0,002	0,004	-0,004
	Sig. (bilateral)	0,98	0,98	0,98	0,951	0,94	0,871	0,875
	N	1605	1605	1605	1605	1605	1605	1605
JURIDICA	Correlación de Pearson	-1,000**	1,000**	-0,001	-0,002	-0,002	0,004	-0,004
	Sig. (bilateral)	0	0	0,98	0,951	0,94	0,871	0,875
	N	1605	1605	1605	1605	1605	1605	1605
FUNCIONARIO	Correlación de Pearson	0,002	-0,002	-0,002	-0,004	-0,005	0,011	-0,011
	Sig. (bilateral)	0,944	0,944	0,944	0,862	0,832	0,646	0,655
	N	1605	1605	1605	1605	1605	1605	1605

correlaciones		ESTRATO 5	ESTRATO 6	JURIDICA	FUNCIONARIO
NATURAL	Correlación de Pearson	0,002	0,001	-1,000**	0,002
	Sig. (bilateral)	0,951	0,98	0	0,944
	N	1605	1605	1605	1605
JURIDICA	Correlación de Pearson	-0,002	-0,001	1,000**	-0,002
	Sig. (bilateral)	0,951	0,98	0	0,944
	N	1605	1605	1605	1605
ESTUDIANTE	Correlación de Pearson	-0,002	-0,001	-0,001	-0,002
	Sig. (bilateral)	0,951	0,98	0,98	0,944
	N	1605	1605	1605	1605
PROFESIONAL INDEPENDIENTE	Correlación de Pearson	-0,004	-0,002	-0,002	-0,004
	Sig. (bilateral)	0,881	0,951	0,951	0,862
	N	1605	1605	1605	1605
RENTA CAPITAL	Correlación de Pearson	-0,005	-0,002	-0,002	-0,005
	Sig. (bilateral)	0,854	0,94	0,94	0,832
	N	1605	1605	1605	1605
PEP NO	Correlación de Pearson	0,01	0,004	0,004	0,011
	Sig. (bilateral)	0,691	0,871	0,871	0,646
	N	1605	1605	1605	1605
ESTRATO 4	Correlación de Pearson	-0,01	-0,004	-0,004	-0,011
	Sig. (bilateral)	0,699	0,875	0,875	0,655

correlaciones		ESTRATO 5	ESTRATO 6	JURIDICA	FUNCIONARIO
	N	1605	1605	1605	1605
ESTRATO 5	Correlación de Pearson	1	-0,002	-0,002	-0,004
	Sig. (bilateral)		0,951	0,951	0,862
	N	1605	1605	1605	1605
ESTRATO 6	Correlación de Pearson	-0,002	1	-0,001	-0,002
	Sig. (bilateral)	0,951		0,98	0,944
	N	1605	1605	1605	1605
JURIDICA	Correlación de Pearson	-0,002	-0,001	1	-0,002
	Sig. (bilateral)	0,951	0,98		0,944
	N	1605	1605	1605	1605
FUNCIONARIO	Correlación de Pearson	-0,004	-0,002	-0,002	1
	Sig. (bilateral)	0,862	0,944	0,944	
	N	1605	1605	1605	1605

Tabla 3.

3. Procedimiento k-medias.

Ya que en el paso 2, se puso en evidencia la falta de independencia entre las variables tanto en las características del cliente como en las de transacción, en el caso de las variables de las transacciones no se hizo un análisis posterior para garantizar la independencia entre las variables dado el hecho de que una persona pudo realizar varias transacciones en el tiempo que se tomaron los datos, por lo tanto al segmentarlos en esta base de datos se tendría que un asociado identificado por su ID se podría encontrar en varios clúster a la vez, es por esta razón que optamos por usar para la segmentación la base de datos de clientes, para esto se hizo un análisis de componentes principales para las variables donde obtuvimos 15 componentes, las cuales aseguran que las variables que se ingresan al algoritmo k-medias son independientes.

Los resultados obtenidos al aplicar k–medias sobre las componentes principales obtenidas de la base de clientes, son los siguientes:

En la tabla 4 se resumen los resultados de los centroides finales del algoritmo k medias

Centros de los conglomerados finales

	Conglomerado		
	1	2	3
SALARIO	840080	1678128	2364465
OTROS_INGRESOS	0	258072	250191
TOTAL_INGRESOS	840080	1940633	2614655
TOTAL INGRESOS	840080	2056683	2896551
ACTUALIZACION			
APORTES	2318615	1320570	2308044
CARTERA	6064941	7021946	11618492
EDAD (SICSES)	63	51	70
ACTIVO	305000000	29157186	214433191
ACTIVO ACTUALIZACION	3000000000	30801583	219729534
PASIVO	6200000	10143606	26091946
PASIVO ACTUALIZACION	7000000	11495414	30810373
PATRIMONIO	298800000	19013579	188341245
PATRIMONIO	2993000000	19306169	188919161
ACTUALIZACION			
INGRESOS BRUTOS	840080	1936847	2642613
GASTOS	100810	267073	301857

Tabla 4

A continuación, mostraremos los promedios de las variables por cada clúster

Clúster 1

descriptivos

	N	Mínimo	Máximo	Media	
	Estadístico	Estadístico	Estadístico	Estadístico	Error típico
SALARIO	889	0	8883258	1769226,22	47249,342
OTROS_INGRESOS	889	0	10285498	256989,16	27120,078
TOTAL_INGRESOS	889	0	11450726	2026215,38	53078,530
TOTAL INGRESOS ACTUALIZACION	889	0	43650000	2155294,12	87563,448
APORTES	889	0	22687083	1471248,63	82938,102
CARTERA	889	0	76790205	7673447,39	360441,796
EDAD (SICSES)	889	0	97	53,41	,591
ACTIVO	889	0	766000000	56080941,17	2969336,321
ACTIVO ACTUALIZACION	889	0	766000000	57203716,99	2815957,357
PASIVO	889	0	240609000	12343717,38	782910,278
PASIVO ACTUALIZACION	889	0	240609000	13361613,73	783703,453
PATRIMONIO	889	-100000000	700525000	43737223,79	2761367,646
PATRIMONIO ACTUALIZACION	889	-100000000	700525000	43842103,26	2637933,769
INGRESOS BRUTOS	889	0	11450726	2030921,04	53149,230
GASTOS	889	0	5768748	271251,52	13582,341
N válido (según lista)	889				

a. Número inicial de casos = 1

Clúster 2

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	
	Estadístico	Estadístico	Estadístico	Estadístico	Error típico
SALARIO	1	3378171	3378171	3378171,00	.
OTROS_INGRESOS	1	0	0	,00	.
TOTAL_INGRESOS	1	6756342	6756342	6756342,00	.
TOTAL INGRESOS	1	6756342	6756342	6756342,00	.
ACTUALIZACION					
APORTES	1	0	0	,00	.
CARTERA	1	0	0	,00	.
EDAD (SICSES)	1	67	67	67,00	.
ACTIVO	1	0	0	,00	.
ACTIVO ACTUALIZACION	1	0	0	,00	.
PASIVO	1	0	0	,00	.
PASIVO ACTUALIZACION	1	0	0	,00	.
PATRIMONIO	1	0	0	,00	.
PATRIMONIO	1	0	0	,00	.
ACTUALIZACION					
INGRESOS BRUTOS	1	3378171	3378171	3378171,00	.
GASTOS	1	353153	353153	353153,00	.
N válido (según lista)	1				

a. Número inicial de casos = 2

Clúster 3

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	
	Estadístico	Estadístico	Estadístico	Estadístico	Error típico
SALARIO	5	515000	8940288	3092466,60	1528004,635
OTROS_INGRESOS	5	0	1212936	242587,20	242587,200
TOTAL_INGRESOS	5	515000	10153224	3335053,80	1761519,237
TOTAL INGRESOS	5	840080	10724248	5512865,60	1952649,405
ACTUALIZACION					
APORTES	5	137000	2318615	1062982,20	399323,704
CARTERA	5	410949	49125625	13746797,00	8979202,759
EDAD (SICSES)	5	59	66	62,80	1,158
ACTIVO	5	0	393000000	1,94E8	81729186,953
ACTIVO ACTUALIZACION	5	200000000	3000000000	9,24E8	5,240E8
PASIVO	5	0	200000000	41240000,00	39708155,334
PASIVO ACTUALIZACION	5	7000000	318000000	1,91E8	51801544,379
PATRIMONIO	5	0	298800000	1,53E8	64929389,339
PATRIMONIO ACTUALIZACION	5	20000000	2993000000	7,33E8	5,690E8
INGRESOS BRUTOS	5	515000	10153224	3335053,80	1761519,237
GASTOS	5	41200	936931	392035,40	179984,222
N válido (según lista)	5				

a. Número inicial de casos = 3

De las anteriores estadísticas podemos apreciar que en el clúster 3 encontramos una característica marcada por tener valores muy grandes en las variables Activo, Activo Actualización, Pasivo, Pasivo Actualización, Pasivo, Pasivo Actualización, Patrimonio, Patrimonio Actualización, en el clúster dos podemos ver que presenta 9 variables en 0, mientras que en clúster 1, los valores son bastante homogéneos teniendo en cuenta los valores del máximo y mínimo, de lo que podemos concluir que dada la concentración de datos en el clúster 1 que corresponde al 99,33% de los datos, en los clúster 2 y 3 encontramos datos con características extremas o atípicas, las cuales tendremos en cuenta como posibles casos de LA/FT.

A continuación, tenemos los resultados obtenidos del análisis ANOVA para el algoritmo k-medias.

ANOVA

	Conglomerado		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
SALARIO	2,694E13	2	1,980E12	892	13,604	,000
OTROS_INGRESOS	3,646E10	2	6,522E11	892	,056	,946
TOTAL_INGRESOS	2,628E13	2	2,539E12	892	10,351	,000
TOTAL INGRESOS ACTUALIZACION	4,058E13	2	6,867E12	892	5,910	,003
APORTES	5,522E13	2	5,971E12	892	9,248	,000
CARTERA	1,190E15	2	1,144E14	892	10,402	,000
EDAD (SICSES)	20495,102	2	264,116	892	77,599	,000
ACTIVO	1,962E18	2	3,664E15	892	535,422	,000
ACTIVO ACTUALIZACION	6,329E18	2	3,178E15	892	1991,734	,000
PASIVO	1,433E16	2	5,505E14	892	26,027	,000
PASIVO ACTUALIZACION	2,101E16	2	7,327E14	892	28,679	,000
PATRIMONIO	1,645E18	2	3,223E15	892	510,555	,000
PATRIMONIO ACTUALIZACION	5,961E18	2	2,703E15	892	2205,190	,000
INGRESOS BRUTOS	2,874E13	2	2,517E12	892	11,420	,000
GASTOS	8,273E10	2	1,639E11	892	,505	,604

Las pruebas F sólo se deben utilizar con una finalidad descriptiva puesto que los conglomerados han sido elegidos para maximizar las diferencias entre los casos en diferentes conglomerados. Los niveles críticos no son corregidos, por lo que no pueden interpretarse como pruebas de la hipótesis de que los centros de los conglomerados son iguales.

De la tabla ANOVA podemos observar que las variables que tienen menos representación en la segmentación son las variables otros ingresos y gastos, lo cual muestra que se escogieron de manera asertiva las variables ingresadas al algoritmo k-medias, ya que presentan importancia a la hora de clasificar mediante el algoritmo.

Por lo tanto, podemos establecer que el modelo de segmentación se caracteriza mediante las variables: Salario, Total ingresos, Total ingresos Actualización, Aportes, Cartera, Edad(SICSES), Activo, Activo Actualización, Pasivo, Pasivo Actualización, Patrimonio, Patrimonio Actualización, Ingresos Brutos.

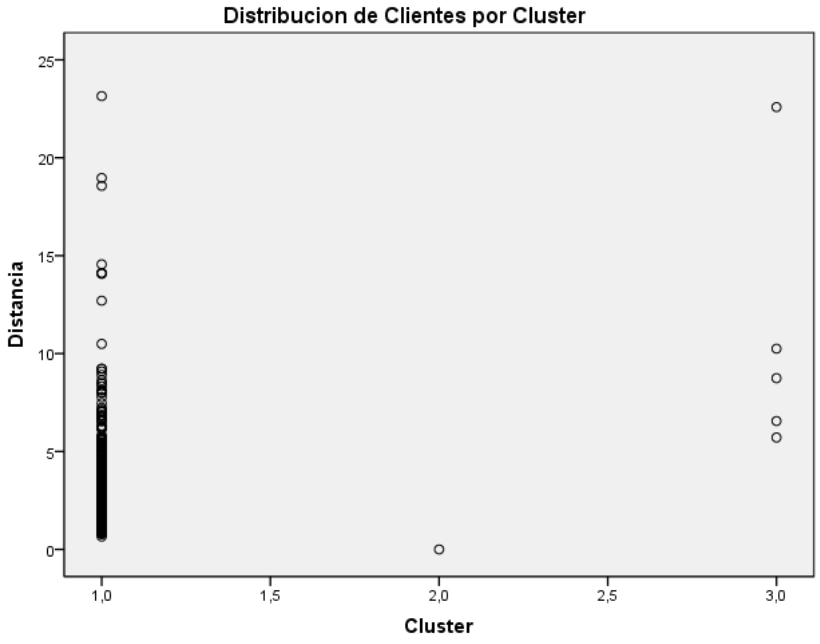
A continuación, se mostrará la cantidad de clientes en cada clúster

Número de casos en cada conglomerado

Conglomerado	1	889,000
	2	1,000
	3	5,000
Válidos		895,000
Perdidos		,000

Fase 4.

De la segmentación se puede evidenciar una concentración del 99.33% de clientes en el clúster 1, el siguiente grafico ilustrara la distribución de los clientes por clúster:



Este grafico nos permite observar ciertos casos inusuales entre los clúster, de acuerdo a su lejanía con el centroide del clúster correspondiente. Éstos casos son identificados por su ID y características más importantes:

Clúster 1

ID	SALARIO	OTROS_INGRESOS	TOTALINGRESOSACTUALI.	ACTIVOACTUALIZ.	PASIVOACTUALIZ.	PATRIMONIOACTUALZ.	INGRESOSBRUTOS	GASTOS
4679688	2711939	542388	41600000	100000000	9000000	91000000	3254327	216956
10541404	4155349	0	43650000	60000000	50000000	10000000	4155349	332428
10520000	1753627	0	3253000	180000000	51200000	128800000	3428627	210435

Clúster 2.

ID	SALARIO	OTROS_INGRESOS	TOTALINGRESOSACTUALI.	ACTIVOACTUALIZ.	PASIVOACTUALIZ.	PATRIMONIOACTUALZ.	INGRESOSBRUTOS	GASTOS
10524491	3378171	0	6756342	0	0	0	3378171	353153

Clúster 3.

ID	SALARIO	OTROS_INGRESOS	TOTALINGRESOSACTUALI.	ACTIVOACTUALIZ.	PASIVOACTUALIZ.	PATRIMONIOACTUALZ.	INGRESOSBRUTOS	GASTOS
4679332	515000	0	9500000	600000000	200000000	400000000	515000	41200
10536866	2220000	0	2500000	200000000	180000000	20000000	2220000	177600
10527977	8940288	1212936	10724248	521000000	318000000	203000000	10153224	936931
25295279	2946965	0	4000000	300000000	250000000	50000000	2946965	703636
10530511	840080	0	840080	3000000000	7000000	2993000000	840080	100810

Siendo estos casos objetos de estudio por parte del experto en la cooperativa, quien realizará un análisis exhaustivo de sus características y decidirá su participación en LA/FT.

CONCLUSIONES Y RECOMENDACIONES.

En base a los análisis hechos y los resultados obtenidos concluimos que el tipo de segmentación apropiada para CODELCAUCA, es una segmentación basada en las características de los clientes y no en la transacción como se tomó la información, ya que las variables asociadas a la transacción resultan relevantes porque no identifican de manera precisa la transacción debido al hecho de que un asociado al presentar múltiples transacciones podría quedar ubicado en varios clúster al tiempo , además del hecho de obtener clientes con 1 y 2 transacciones en un periodo de 7 meses no establece un criterio claro para decidir si se está participando en el delito de LA/FT, es decir realizar un proceso de segmentación por transacciones carecería de concordancia con la realidad del problema.

De los resultados obtenidos pudimos establecer que las variables que permiten establecer criterios para la presencia de LA/FT son: Salario, Total ingresos, Total ingresos Actualización, Aportes, Cartera, Edad (SICSES), Activo, Activo Actualización, Pasivo, Pasivo Actualización, Patrimonio, Patrimonio Actualización, Ingresos Brutos.

Se recomienda para CODELCAUCA, con el propósito de mejorar su modelo actual de segmentación un grupo interno de trabajo dedicado a seleccionar la información con la cual se realizará la segmentación, dado que este es un proceso que debe realizarse con cierta periodicidad, como lo establece la circular externa 04 de 2017.teniendo en cuenta dos puntos de vista importantes a la hora de las decisiones

Que serían la experticia del personal interno de la cooperativa y un estadístico para los procesos estadístico-matemáticos que se requieran.

BIBLIOGRAFÍA

GAFI. (10 de agosto de 2018). *GAFI*. Obtenido de https://www.cfatf-gafic.org/index.php/es/documentos/gafi40-recomendaciones

Gimenez, Y. (23 de marzo de 2010). *El metodo k-medias-Departamento de Matematicas-universidad de Buenos Aires* . Obtenido de El metodo k-medias-Departamento de Matematicas-universidad de Buenos Aires : https://www.google.com/url?sa=t&source=web&rct=j&url=http://cms.dm.uba.ar/academico/carreras/licenciatura/tesis/2010/Gimenez_Yanina.pdf&ved=2ahUKEwiTitPL_dnhAhUkuVkKHUVYCXMQFjAeegQICRAB&usg=AOvVaw1cC-491vUcGG_6h8Kjp8n&cshid=1555602184870

Perez, C. (2011). Técnicas de segmentación conceptos, herramientas y aplicaciones. En C. Perez, *Técnicas de segmentación conceptos, herramientas y aplicaciones.* mexico: Alfaomega Grupo editor S.A de C.V.

UIAF. (15 de octubre de 2017). Obtenido de https://www.uiaf.gov.co/caracterizacion_usuarios/perfiles/reportantes/superintendencia_economia_solidaria/28656

UIAF. (12 de noviembre de 2017). *Unidad de Información y Análisis Financiero*. Obtenido de Gobierno de Colombia, Minhacienda: http://www.uiaf.gov.co