**DOCUMENTO TÉCNICO**

**SEGMENTACIÓN DE LOS FACTORES DE RIESGO**

**COOPEIAPE**

**Contenido**

[1. ANTECEDENTE 3](#_Toc528938865)

[2. OBJETIVO 3](#_Toc528938866)

[3. GLOSARIO 3](#_Toc528938867)

[4. METODOLOGIA 4](#_Toc528938868)

[4.1. CONGLOMERADOS JERÁRQUICOS: 4](#_Toc528938869)

[4.2. K-MEDIAS 6](#_Toc528938870)

[4.3. CRISP-DM (CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING) 8](#_Toc528938871)

[4.4. PREPARACION DE LAS BASES DE DATOS 9](#_Toc528938878)

[5. SEGMENTACION 9](#_Toc528938879)

[5.1. ASOCIADOS 9](#_Toc528938880)

[5.2. PRODUCTOS 11](#_Toc528938881)

[5.2.1. PLAN EXEQUIAL 12](#_Toc528938882)

[5.2.2. HOMENAJES 13](#_Toc528938883)

[5.2.3. PRENECESIDAD 14](#_Toc528938884)

[5.3. CANALES 16](#_Toc528938885)

[5.4. JURISDICCION 17](#_Toc528938886)

# ANTECEDENTE

En virtud de lo establecido en la Circular Externa No 004 de 2017, en el numeral 2.1.1. establece identificar los riesgos de LAFT inherentes al desarrollo de la actividad, teniendo en cuenta los factores de riesgos del asociado, producto, canal y jurisdicción en la que requiere a la entidad establecer metodologías para la segmentación de los agentes originadores del riesgo de LAFT.

A su vez, en el numeral 2.2.2.4.2. de la citada circular, establece que *“Las organizaciones solidarias vigiladas deben 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.”*

Seguidamente, la norma establece en el numeral 2.1.2. de la Circular Externa No 004 de 2017, medir la probabilidad de materialización del riesgo inherente de LAFT frente a cada uno de los factores de riesgo, así como el impacto en caso de materializarse mediante los riesgos asociados.

Por lo tanto, en el presente documento se explica la metodología utilizada para los modelos de segmentación y su proceso.

# OBJETIVO

Identificar mediante la segmentación de los factores de riesgo, la separación de elementos en grupos homogéneos al interior de cada clúster y heterogéneo en cada uno de ellos, a fin de determinar los riesgos en el lavado de activos y financiación del terrorismo por parte de los asociados de la entidad mediante la implementación de modelos de segmentación de mercados.

# GLOSARIO

* **SARLAFT:** Sistema de Administración de Riesgos de Lavado de Activos y Financiación del terrorismo.
* **UIAF**: Unidad de Información y Análisis Financiero.
* **Análisis clúster:** Método que pretende agrupar observaciones según criterios de homogeneidad descifrados en variables de observación.
* **Análisis clúster No jerárquico:** Métodos de optimización en los que se indica a priori el número de segmentos. Realizando la distribución de las observaciones con un criterio de optimización del criterio de selección.
* **Criterio de selección:** Medida usada para segmentar las observaciones según un conjunto de variables que caracterizan cada una de las observaciones.
* **K medias:** Técnica de segmentación que pretende dividir un conjunto de n observaciones en k segmentos homogéneos entre sí y heterogéneos frente a los demás, por medio de una medida de distancia.
* **Distancia euclidiana:** Es el insumo principal para el algoritmo k medias, establece un parámetro de medición entre observaciones.

# METODOLOGÍA

## CONGLOMERADOS JERÁRQUICOS:

El procedimiento de análisis de conglomerados jerárquicos permite aglomerar tantos casos como variables y elegir entre una gran variedad de métodos de aglomeración y medidas de distancia de manera jerárquica.

Dicho análisis inicia con el calculo de la matriz de distancias entre elementos de la muestra, la cual contiene las distancias existentes entre cada elemento y los restantes de la muestra para calcular los elementos más próximos, es decir, los más similares en términos de distancia. El conglomerado resultante es indivisible por lo que se denomina jerárquico.

Por lo tanto, se agrupan los elementos conglomerados más grandes y mas diferentes (heterogéneos) hasta llegar al último paso, en que todos los elementos escogidos quedan agrupados en un único conglomerado.

El análisis de conglomerados jerárquicos es una técnica aglomerativa, toda vez que partiendo de los elementos muestrales individualmente considerados, va creando grupos hasta llegar a la formación de un único grupo o conglomerado constituido por todos lo elementos de la muestra.

Para ello el cálculo de conglomerados jerárquicos se utilizó el método de agrupación en clústeres con el método de Ward, donde es un procedimiento jerárquico en el cual, en cada etapa, se unen los dos clústeres para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada clúster, de cada individuo al centroide del clúster.

El método de agrupación de Ward (Ward, 1963) o de varianza mínima (Minimum variante clustering), se argumenta que los conglomerados deben constituirse de tal manera que, al fundirse dos elementos, la pérdida de información resultante de la fusión fuera mínima, cuantifica la cantidad de información como la suma de las distancias al cuadrado de cada elemento respecto al centroide del conglomerado que pertenece (SCE = Suma de Cuadrados Error).

Para ello, se comienza calculando, en cada conglomerado, el vector de medias de todas las variables, es decir, el centroide multivariante. Posteriormente, se calculan las distancias euclidianas al cuadrado entre cada elemento y los centroides (vector de medias) de todos los conglomerados.

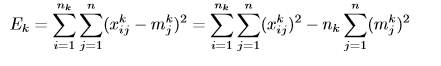
Finalmente, se suman las distancias correspondientes a todos los elementos.

En cada paso se unen aquellos conglomerados que dan lugar a un menor incremento de la SCE, es decir, de la suma de cuadrados de las distancias intra conglomerado.

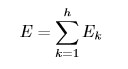
El método de enlace medio dentro de los grupos, o de vinculación Intra-grupos, aprovecha la información de todos los miembros de los conglomerados que se comparan uniéndolos previamente. La distancia entre dos conglomerados se calcula como la distancia promedio existente entre todos los miembros del conglomerado unión de ambos.

La explicación matemática es las siguiente:

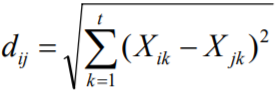
* al valor de la j−ésima variable sobre el i−ésimo individuo del k−ésimo clúster, suponiendo que dicho clúster posee individuos.
* al centroide del clúster, con componentes .
* a la suma de cuadrados de los errores del clúster , o sea, la distancia euclídea al cuadrado entre cada individuo del cluster a su centroide.



* a la suma de cuadrados de los errores para todos los clústeres, o sea, si suponemos que hay h clúster.



Posteriormente, la medida utilizada para determinar el intervalo es la distancia euclídea al cuadrado, la cual corresponde a la distancia "ordinaria" (que se mediría con una regla) entre dos puntos de un espacio euclídeo, la cual se deduce a partir del teorema de Pitágoras.



Por lo tanto, de acuerdo con lo requerido por la Superintendencia de la Economía Solidaria, en la Circular Externa No 004 de 2017, para identificar el riesgo se requiere segmentar los factores de riesgo y el método utilizado para determinar grupos homogéneos entre si y heterogéneos para cada uno de ellos, es conglomerados jerárquicos con método de agrupación en clústeres con método de Ward y la medida del intervalo con distancia euclídea al cuadrado.

## K-MEDIAS

K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster. Se suele usar la distancia cuadrática.

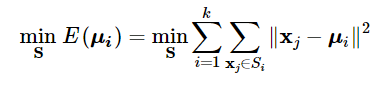
El algoritmo consta de tres pasos:

1. **Inicialización:** una vez escogido el número de grupos, *k*, se establecen *k* centroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente.
2. **Asignación objetos a los centroides:** cada objeto de los datos es asignado a su centroide más cercano.
3. **Actualización centroides:** se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso.

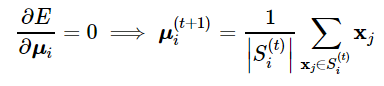
El algoritmo k-means resuelve un **problema de optimización**, siendo la función por optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su cluster.

Los objetos se representan con vectores reales de *d* dimensiones (x1,x2,…,xn) y el algoritmo k-means construye k grupos donde se minimiza la suma de distancias de los objetos, dentro de cada grupo S={S1,S2,…,Sk}, a su centroide, asi:



donde S es el conjunto de datos cuyos elementos son los objetos xj representados por vectores, donde cada uno de sus elementos representa una característica o atributo. Tendremos k grupos o clúster con su correspondiente centroide μi.

En cada actualización de los centroides, desde el punto de vista matemático, se impone la condición necesaria de extremo a la función E(μi) que, para la función cuadrática es



Finalmente, se toma el promedio de los elementos de cada grupo como nuevo centroide.

De acuerdo a lo anterior, el análisis clúster de K-medias es una herramienta diseñada para asignar los casos a un número fijo de grupos, cuyas características no se conocen, pero se basan en un conjunto de variables que deben ser cuantitativas. Es muy útil cuando se quiere clasificar un gran número de casos. Es un método de agrupación de casos que se basa en las distancias existentes entre ellos en un conjunto de variables cuantitativas. Este método de aglomeración no permite agrupar variables. El objetivo de optimalidad (buscar la mejor) que se persigue es “maximizar la homogeneidad dentro de los grupos.”

Es el método que se usa más habitualmente, es fácil de programar y da resultados razonables. Tiene por objetivo separar las observaciones en K clúster, de manera que cada dato pertenezca a un grupo y sólo a uno.

El algoritmo busca con un método iterativo:

* Los centroides (medias, medianas) de los K clúster.
* Asigna cada individuo a un clúster.
* El algoritmo requiere que se especifique el número de conglomerados, también se puede especificar los centros iniciales de los clústeres si conoce de antemano dicha información.

En este método, la medida de distancia o de similaridad entre los casos se calcula utilizando la distancia euclídea. Es muy importante el tipo de escala de las variables, si las variables tienen diferentes escalas (por ejemplo, una variable se expresa en dólares y otra en años), los resultados podrían ser equívocos. En estos casos, se debería considerar la estandarización de las variables antes de realizar el análisis de conglomerados de k-medias.

Este procedimiento supone que se ha seleccionado el número apropiado de conglomerados y que se han incluido todas las variables relevantes. Si se ha seleccionado un número inapropiado de conglomerados o se han omitido variables relevantes, los resultados podrían ser equívocos.

Existen varias formas de implementarlo, pero todas ellas siguen, básicamente, los siguientes pasos:

1. Se toman al azar k clústeres iniciales y se calculan los centroides (medias) de los clústeres.
2. Se calcula la distancia euclídea de cada observación a los centroides de los clústers y se reasigna cada observación al grupo más próximo formando los nuevos clústers que se toman en lugar de los primeros como una mejor aproximación de estos.
3. Se calculan los centroides de los nuevos clústeres.
4. Se repiten los pasos 2) y 3) hasta que se satisfaga un criterio de parada como, por ejemplo, no se produzca ninguna reasignación, es decir, los clústers obtenidos en dos iteraciones consecutivas son los mismos.

## [CRISP-DM](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining) (CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING)

Para la implementación de los modelos se hacer necesario tomar en cuenta la metodología  [CRISP-DM](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining) (Cross Industry Standard Process for Data Mining), el cual especifica las tareas a realizar en cada fase descrita en el proceso.

* **Fase I. Definición de necesidades del cliente (comprensión del negocio):** Esta fase inicial se enfoca en la comprensión de la entidad. Después se convierte este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.
* **Fase II. Estudio y comprensión de los datos:** La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.
* **Fase III. Análisis de los datos y selección de características:** La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.
* **Fase IV. Modelado:** En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto, casi siempre en cualquier proyecto se acaba volviendo a la fase de preparación de datos.
* **Fase V. Evaluación (obtención de resultados):** En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la una perspectiva de análisis de datos.  
  Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de la entidad. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no haya sido considerada suficientemente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.
* **Fase VI. Despliegue (puesta en producción):** Se entrega el documento donde se realiza la entrega del documento final del modelo para que sea puesto en producción y continuar con las fases de automatización, seguimiento, evaluación de desempeño y reentrenamiento que requiere este tipo de modelos.