Clasificación de aseguradoras venezolanas con algoritmos de aprendizaje automático

Classification of Venezuelan insurers with Machine Learning algorithms

Jorge Luis Aquino-Olmos[1]
Elizabeth López-Meléndez[2]
Luis David Lara-Rodríguez[3]

[1]Universidad Tecnológica Latinoamericana (México). Correo electrónico: aquinosuarez@gmail.com
orcid: https://orcid.org/0009-0009-8817-1624

[2]Universidad Tecnológica de Huejotzingo (México). Correo electrónico: elizabeth.lopez@uth.edu.mx
orcid: https://orcid.org/0000-0002-1241-3289

[3]Universidad Politécnica de Puebla (México). Correo electrónico: luis.lara406@uppuebla.edu.mx
orcid: https://orcid.org/0000-0001-9700-390X

Recibido: 30-01-2024 Aceptado: 16-04-2024

Cómo citar: Aquino-Olmos, Jorge; López-Meléndez, Elizabeth; Lara-Rodríguez, Luis David (2024). Clasificación de aseguradoras venezolanas con algoritmos de aprendizaje automático. *Informador Técnico*, *88*(1), 39.55.
https://doi.org/10.23850/22565035.6219

Resumen

Las aseguradoras juegan un papel relevante en una economía sana, ya que proveen un servicio de seguridad a bienes y personas. El mercado asegurador venezolano en las últimas décadas ha enfrentado grandes desafíos y retos en un medio estrecho, donde conocer a los competidores más cercanos es de suma importancia. El órgano rector público que regula la comparación entre las aseguradoras solo ha hecho uso de las primas cobradas como factor de categorización; sin embargo, este órgano hace pública una gama adicional de indicadores. En este estudio hacemos uso de cinco de estos indicadores, utilizados en los últimos tres años, para corroborar si las primas cobradas representan una característica única clasificatoria. El vasto repertorio del aprendizaje automático permite hacer frente al aumento significativo de variables de estudio y sus posibles agrupaciones; el análisis multivariante de estas 18 variables ha requerido del uso del método de análisis factorial, que permite reducir la dimensionalidad en factores altamente correlacionados. Con estos nuevos factores se busca agrupar-clasificar con ayuda de los algoritmos no supervisados de K-Medias (K-Means) y medias difusas (Fuzzy C-Means) comparando sus agrupaciones derivadas, cotejadas con respecto a que las primas cobradas representen una característica determinante de agrupación.

Palabras clave: aprendizaje automático; K-Means; Fuzzy C-Means; análisis factorial; agrupaciones; aseguradoras.

Abstract

Insurance companies play an important role in a healthy economy, providing a security service for goods and people. The Venezuelan insurance market has faced great challenges in recent decades, in a tight environment where knowing the closest competitors is of paramount importance. The public body that regulates the

comparison between insurers has only used the premiums charged as a categorization factor, an additional set of indicators is published. In this study, we make use of five of these indicators used over the last three years and test whether the premiums charged represent a unique ranking characteristic. The vast repertoire of machine learning is able to cope with the significant increase in the number of study variables and their possible groupings. The multivariate analysis of these 18 variables has required the use of the factor analysis method, which allows dimensionality reduction of highly correlated factors. With these new factors, clustering classification is sought using the unsupervised K-Means and Fuzzy C-Means algorithms, comparing their derived clusters with respect to whether the premiums charged represent a determinant clustering characteristic.

Keywords: machine learning; K-Means; Fuzzy C-Means; factor analysis; clusters; insurers.

1. Introducción

El objeto de un seguro es resarcir los daños que pudiesen sobrevenir al ocurrir un evento, es por ello que la industria aseguradora juega un papel importante en la economía de los países, debido a ese efecto de auxilio que presta, llegando a ser un aliado en la estabilidad económica en momentos de crisis. Es pertinente mencionar que, en Colombia, para el año 2021, esta industria aportó el 3,72 % del PIB, cifra que está en promedio a nivel regional (Avedaño, 2023). Se observa que la industria aseguradora a lo largo de la historia ha sido un aliado en el desarrollo de los países, ninguna economía que haya florecido lo ha hecho sin su apoyo (Agudelo; Bernal, 1985).

El desarrollo del sector asegurador va a la par con el desarrollo económico de cualquier economía, por lo cual es vital poder realizar una clasificación que conlleve a comparar compañías aseguradoras. En la práctica, esta clasificación es hecha por empresas especialistas en realizar calificación crediticia y de riesgo. Estas agencias evalúan ciertos aspectos relacionados con la solidez financiera y la capacidad de la aseguradora para el cumplimiento de las obligaciones con sus asegurados, para luego otorgarles una calificación como A+, AA-, B, etc., donde una calificación más alta indica una mayor solvencia y menor riesgo.

El marco regulatorio del sector asegurador en Europa se establece a través de Solvencia II, un marco regulatorio creado por la Unión Europea para compañías y grupos de seguros, que abarca los recursos financieros, la gobernanza y la rendición de cuentas, la evaluación de riesgos de mercado, donde se dan las directrices y normas que regulan y sirven para supervisar a la industria del seguro y reaseguro, ayudando de manera significativa a la estabilidad de este pilar fundamental de la economía (Mayorga, 2014). Todos estos mecanismos y formas de clasificar a las aseguradoras ayudan a los órganos contralores a realizar la supervisión necesaria, a la cual están obligados. Sin embargo, este tecnicismo no cala en los actores del ramo, por ello se hace uso de indicadores más sencillos para poder llegar a clasificar a las empresas, sirviendo estos como mecanismo de ventas y estrategia comercial que llegan a los futuros compradores de seguros. Las cifras con las cuales se hará el estudio corresponden a las generadas en varios años consecutivos por la Superintendencia de Seguros (Sudeaseg), órgano rector del sector en Venezuela (Gobierno Bolivariano de Venezuela, 2022).

La evolución de la tecnología se ha dado a pasos agigantados, y ha ayudado a las empresas a disponer de herramientas cada vez más sofisticadas y novedosas, que han logrado almacenar una variada y gran cantidad de información, la cual es procesada a través de sistemas de cómputo modernos. De igual manera, este avance tecnológico ofrece novedosas técnicas de análisis, pero también ha creado un reto: poder ser capaces de hacer uso de esta ingente cantidad de datos para la toma de decisiones empresariales y comerciales. En vista de ello se pretende, a través de este trabajo, determinar si la forma como se ha realizado la clasificación de las aseguradoras categorizadas por medio de las primas cobradas es correcta estadísticamente. Al aplicar un conjunto de técnicas estadísticas comunes del aprendizaje de máquina, se busca encontrar la similitud entre las empresas aseguradoras que hacen vida en el mercado venezolano. Esto, tomando en cuenta los índices que de manera regular publica el órgano regulador del sector, es decir, se pretende crear grupos de empresas que

presenten características similares, pudiendo proporcionar este conocimiento ventajas, entre las que podemos mencionar: identificar fortalezas y debilidades, detectar oportunidades de mercado, mejorar la estrategia de marketing, fomentar la innovación, prepararse para cambios en el mercado y aprender de los errores de la competencia, con lo cual, se podrán ejecutar ciertas acciones que conlleven al crecimiento empresarial.

Entre las acciones que se pudiesen implementar, se pueden mencionar: mejorar sus canales de ventas al observar cómo lo realiza la competencia, dónde poder establecer nuevas oficinas, observar el nicho de mercado de sus competidores y con ello incrementar sus ventas, y aumentar la red de aliados comerciales que ofrezcan a la aseguradora (Aragón *et al.*, 2023). En última instancia, tener una comprensión profunda de la competencia es fundamental para sobrevivir en mercados altamente competitivos, como el mercado asegurador venezolano.

En la Tabla 1 se muestran las variables que se toman en cuenta en el análisis estadístico, cuyos datos, que son parte de este estudio, corresponden a los publicados por el órgano rector en el período 2020-2022, correspondiente a tres años con seis variables por cada uno, con lo cual tendremos un conjunto de datos conformado por 18 variables. La información publicada corresponde a 50 aseguradoras; sin embargo, solo forman parte del estudio 42 de estas, ya que las otras 8 no presentan datos o se encuentran en cero en un gran número de años, esto evitará problemas al evaluar estos indicadores en los diferentes métodos aplicados en el estudio.

Tabla 1. *Variables para analizar con su descripción*

Comisiones y Gastos de Adquisición	Representa el costo de la intermediación de seguros, derivado del pago de comisiones y bonificaciones a sus productores de seguros.
GA: Gastos Administrativos	Monto total pagado al cubrir el costo de los gastos administrativos totales, incluyendo gastos de personal y gastos generales derivados del desarrollo de su actividad de seguros.
PC: Primas Cobradas	Prima de tarifa cancelada por el asegurado para cubrir la cobertura de un riesgo específico.
SO: Saldo de Operaciones	Es el saldo positivo o negativo que resulta de todas las operaciones de la empresa de seguros al final del período evaluado, se obtiene al sumar el resultado técnico neto y el resultado de la gestión general, obteniéndose el saldo de operaciones positivo o negativo (utilidad o pérdida) de la empresa.
SP: Siniestros Pagados	Monto total de los siniestros indemnizados por la empresa aseguradora durante un periodo determinado, este monto es el reflejo neto del salvamento de siniestros.
ST: Siniestros Totales	Se refiere a la suma de los siniestros pagados más las reservas para prestaciones y siniestros pendientes brutos.

Fuente: elaboración propia.

En la Tabla 2 se observa la clasificación realizada en el año 2022 por el órgano interno regulador (Sudeaseg) de las aseguradoras presentes en el estudio, que se realiza con el único parámetro de las primas cobradas en dicho año y la cantidad de aseguradoras pertenecientes a sus diferentes cuatro agrupaciones, que corresponden a una cantidad finita de elementos. Las primeras tres agrupaciones cuentan con diez aseguradoras cada una, por orden de importancia en el rubro mencionado, la última agrupación recopila todas las demás aseguradoras. Los datos de las variables de estudio presentan un alto sesgo (hacia la derecha), esta característica es común en variables que tienen valores menores. Un factor que debe ser tomado en cuenta es la medida monetaria adoptada por el Gobierno venezolano, la cual consistió en eliminar ceros en la moneda, allí se crea una distorsión al comparar variables de diferentes años, por ello se deben realizar transformaciones sobre los datos y evitar esta distorsión, para esto normalizamos los datos con ayuda de la Ecuación 1 dentro del rango de [0-1].

$$Z_j = \frac{X_i - min\,(X_j)}{max\,(X_j) - min\,(X_j)} \qquad (1)$$

Tabla 2. *Clasificación dada por Sudeaseg de aseguradoras venezolanas en el año 2022*

ID	Aseguradora	Clasf.	ID	Aseguradora	Clasf.	ID	Aseguradora	Clasf.
1	Mercantil	1	15	Atrio	2	29	Uniseguros	3
2	Caracas	1	16	Venezuela	2	30	Occidental	3
3	Oceánica	1	17	Venezolana	2	31	A. Internacional	4
4	Mapfre	1	18	Los Andes	2	32	Universal	4
5	Pirámide	1	19	Previsora	2	33	Vivir	4
6	Internacional	1	20	Horizonte	2	34	V. del Valle	4
7	Constitución	1	21	La Fé	3	35	Interbank	4
8	Hispana	1	22	Catatumbo	3	36	Oriental	4
9	Universitas	1	23	BBVA	3	37	Corporativos	4
10	Banesco	1	24	Proseguros	3	38	Bolivariana	4
11	Estar S.	2	25	Nuevo Mundo	3	39	Primus	4
12	Real S.	2	26	Zuma	3	40	Vitalicia	4
13	Altamira	2	27	Mundial	3	41	Iberoamericana	4
14	Qualitas	2	28	Caroní	3	42	Avila	4

Fuente: elaboración propia.

Esta normalización permite a los algoritmos que hacen uso de distancias en sus cálculos, en caso de no hacer uso de dicha normalización, los resultados se podrían ver afectados por la existencia de discrepancia entre las magnitudes presentes en los datos. Por otra parte, al realizar la transformación se le da a cada variable la misma importancia, cuando son utilizadas en cualquier algoritmo de agrupación. Un factor importante en nuestro estudio es la correlación que pueda existir entre las variables, por ello se presenta un mapa de calor que identifica esta relación en los datos, lo que permitirá realizar los ajustes necesarios mediante los métodos de aprendizaje de computadora que se han aplicado en búsqueda de la reducción dimensional de las variables, factor importante en este estudio.

$$r = \frac{cov(x,y)}{S_x S_y} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{(n-1)S_x S_y} \qquad (2)$$

Nos hemos apoyado del estadístico llamado correlación de Pearson, definido en la Ecuación 2, el cual permite medir u observar la relación existente entre cada par de variables, dicho estadístico indica un valor nulo a una no existente relación entre las variables de medición, y al estar en cercanía a la unidad (positiva o negativa), expresa una relación (positiva o negativa) entre las variables comparadas. Al realizar el cálculo de la correlación, con ayuda de este estadístico existente entre cada par de variables, se logra construir la matriz de correlación de las variables, la cual se presenta en la Figura 1.

Los resultados obtenidos en la matriz reflejan claramente una alta correlación entre las variables, este inconveniente debe ser tratado, en caso contrario puede generar problemas en el análisis y en la interpretación de los resultados. Algunas medidas que se pueden adoptar para minimizar el problema son: analizar la naturaleza de la correlación (positiva o negativa) u optar por un análisis de reducción de dimensionalidad (Ferrando *et al.*, 2022).

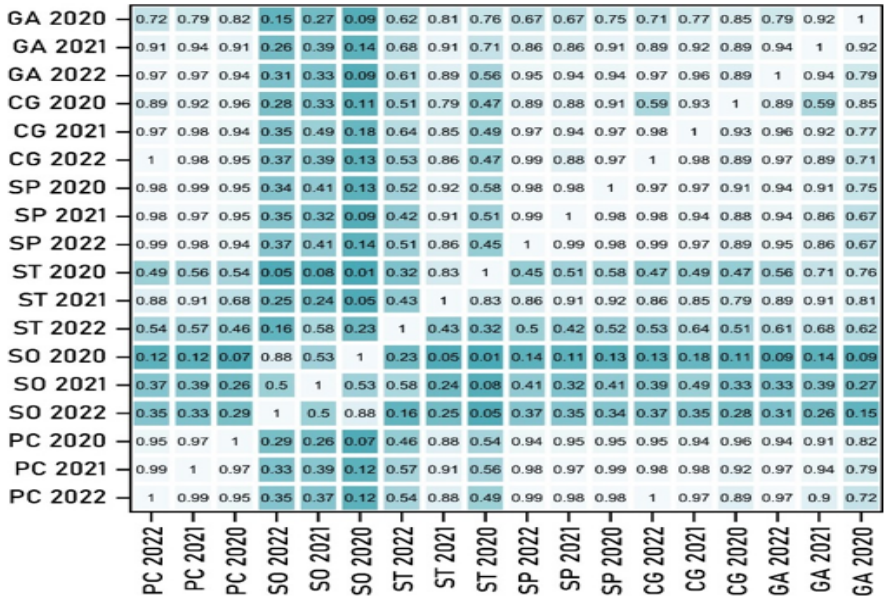

Figura 1. Matriz de correlación de las variables de estudio
Fuente: elaboración propia.

2. Materiales y métodos

Dado que los 18 indicadores a analizar para las 42 aseguradoras están altamente correlacionados, se elige un método de aprendizaje automático para explicar la correlación entre los indicadores observados mediante otras variables no correlacionadas. Estos nuevos factores se agrupan mediante dos técnicas no supervisadas bien conocidas, basadas en métricas de distancia, y su fiabilidad se verifica calculando una métrica de calidad y el rendimiento de los algoritmos de agrupación. El objetivo es observar la naturaleza de las nuevas agrupaciones frente a la dada por el órgano regulador.

El aprendizaje automático (*machine learning*, ML) es una de las áreas más extensas y potenciales de la inteligencia artificial (IA), donde unos de los objetivos es el desarrollo de la capacidad de aprender y proporcionar recomendaciones expertas en un dominio estrecho, apoyándose de métodos o algoritmos adaptables y de aprendizaje. La gran mayoría de estas técnicas se aglomeran en dos grandes grupos: supervisados y no supervisados (Zhou, 2016). El primer grupo decide el problema de clasificación, cuando se conocen determinados objetos que conforman grupos finitos y estos permiten conjuntar un infinito conjunto de objetos. Generalmente, esta clasificación es realizada por un experto (Nateski, 2017). Este grupo se subdivide en clasificadores lineales y no lineales, donde el perceptrón, clasificador bayesiano, análisis lineal discriminante, etc., son representativos del primer grupo (Mukhamediev, 2015) mientras que las redes neurales, máquinas de soporte vectorial, regresión logística, el análisis discriminante lineal y demás representan al segundo (Awad *et al.*, 2012).

Los métodos no supervisados de ML resuelven el problema de clasificación-agrupación haciendo que la gama de objetos iniciales indeterminados se agrupe con ayuda de un proceso automático basado en sus propiedades. La cantidad de agrupaciones puede ser obtenida automáticamente o determinada inicialmente (Ayodele, 2010). De esta clasificación, los métodos típicos son: K-Medias (K-Means), K-Medianas (K-Medians), Agrupaciones Difusas (Fuzzy C-Means, Soft K-Means), K-Medias Armónicas (KHM), entre otras (Barbakh *et al.*, 2009). Las aplicaciones del ML en diversas áreas del conocimiento son variadas, estas incluyen física (Khan *et al.*, 2019), física médica (Harki; Rashid, 2023), robótica (Kim *et al.*, 2021), minería (Mc. Coy; Auret, 2019), agricultura (Sharma *et al.*, 2021), biomedicina (Patel *et al.*, 2020), genética (Libbrecht; Noble, 2015) y médica (Field *et al.*, 2021; Kourou *et al.*, 2015; Abbasi; Goldenholz, 2019).

Por otro lado, el análisis factorial se remonta a principios del siglo XX, inicialmente fue vinculado principalmente a la psicología, para cuantificar los factores del intelecto humano. A partir de estos estudios, se encontró un factor general bajo un número de factores específicos; debido a ello se mejoraron los métodos hasta encontrar el análisis factorial de componentes principales, en el cual, a través de un modelo lineal, se obtienen factores comunes supuestos, además de otros que incluyen las características propias de cada variable y un error aleatorio (Porter, 2015).

Existen diferentes tipos de análisis factorial: el análisis factorial exploratorio, que mediante técnicas estadísticas busca reducir un conjunto de variables al extraer todos sus puntos en común en un número menor de factores; el análisis factorial de factor común, en el que se extraen factores relacionados con la varianza común (covarianza) de las variables, y supone que existe un factor común subyacente a todas las variables; y, por último el análisis factorial de varianza total, en el cual se extraen los factores comunes explicando el total de la varianza, dentro de esta categoría.

Tenemos al análisis de componentes principales, donde se relacionan variables cuantitativas; el análisis de correspondencias múltiples, donde se relacionan variables cualitativas; y el análisis de componentes categórico, donde se pueden combinar variables cuantitativas con cualitativas (Taherdoost *et al.*, 2022).

El modelo del análisis factorial explica un conjunto de p observaciones en cada *n* individuos con un conjunto de factores comunes *k* de $(f_{i,j})$, donde hay menos factores por unidad que por observaciones por unidad $(k < p)$. Cada individuo tiene sus propios factores comunes *k* y estos están relacionados con las observaciones a través de la matriz de carga factorial, el modelo se define en la Ecuación 3.

$$x_{m,i} - \mu_i = l_{i,1} f_{1,m} + \ldots + l_{i,k} f_{k,m} + \varepsilon_i \qquad (3)$$

El algoritmo K-Means de clasificación-agrupaciones fue propuesto por MacQueen en 1967 (MacQueen, 1967), otros autores como Forgy, Lloyd, Hartigan y Wong trabajaron en el algoritmo de Mac Queen, con la finalidad de mejorar su propuesta. El uso principal que se le da a dicho algoritmo es la obtención de información cualitativa y cuantitativa sobre grandes conjuntos de datos multivariados que ayudan a encontrar solo una agrupación definitiva de los datos (Wang; Xialong, 2011). K-Means es el algoritmo no supervisado más sencillo, que resuelve problemas de agrupaciones, define k-centroides para cada agrupación, dado un cierto número de agrupaciones. Estos centroides deben ser colocados cuidadosamente, porque con distintas ubicaciones pueden causar un resultado diferente, mientras más lejano sea el centroide entre agrupaciones, el resultado será mejor. Posteriormente, se toma cada punto de las agrupaciones para asociarlo con el centroide más cercano, en caso de que no exista un punto, se realiza la agrupación con esos valores. Sin embargo, se deben volver a calcular los nuevos k-centroides de las agrupaciones resultantes anteriores, con los cuales debemos reunir el conjunto de datos con el nuevo centroide más cercano. Los k-centroides cambiarán su ubicación paso a paso, hasta que ya no exista cambio, y el centroide no cambie de lugar (Dehariya *et al.*, 2010). Suponiendo que $D = \{x_1, x_2, \ldots, x_n\}$ es el conjunto de información a ser agrupada, K-Means puede ser expresada por una función objetivo que depende de la cercanía de los puntos que apuntan a los centroides de la agrupación, como se muestra en la Ecuación 4 (Wu, 2012).

$$\min_{\{mk\},1\le i\le k} \sum_{i=0}^{k} \sum_{x\in C_i} \|x-\mu\|_2^2 \qquad (4)$$

Donde $\mu_i = 1/\|C_i\| \sum_{x\in Ci} ,x$ es el vector de media de las agrupaciones C_i . La Ecuación 4 representa la cercanía entre el vector de medias de una agrupación y las muestras dentro de ese clúster, donde un valor pequeño resultante indica que hay una similitud alta entre los elementos de una agrupación. El algoritmo de K-Means (K-Medias) se muestra en la Tabla 3.

Tabla 3. *Algoritmo 1 K-Medias*

Algoritmo 1: K Medias

Entrada : Conjunto de Datos $D = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$

: Número de agrupaciones k

Salida : Agrupaciones $C = \{C_1, ..., C_k\}$

Seleccionar k muestras aleatorias

1 $k \leftarrow \{\grave{\boldsymbol{\imath}}_1, \grave{\boldsymbol{\imath}}_2, \grave{\boldsymbol{\imath}}_3, .., \grave{\boldsymbol{\imath}}_k\}$

2 **While** $C_n \neq C$

3 $Ci = \varnothing; \ (1 \leq i \leq k)$

4 **for** $j = 1, 2, ..., m$

Calcular la distancia entre $\boldsymbol{x}_j$ y vector de medidas

5 $d_{ij} \leftarrow \left\| \grave{\boldsymbol{k}}_j - \quad _i \right\|_2$

Etiquetar la agrupación de $\boldsymbol{x}_j$

6 $\lambda_j \leftarrow \arg\min d_{ij}; \ \ i \in \{1, 2, \cdots, k\}$

Mover $\boldsymbol{x}_i$ hacia su agrupación correspondiente

7 $C_{\lambda j} \leftarrow C_{\lambda j} \cup \{\boldsymbol{x}_j\}$

8 **end for**

9 **for** $i = 1, 2, \cdots, k$

Calcular y actualizar los vectores de media x_j

10 $\grave{\boldsymbol{\imath}}_i^{'} = \dfrac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}$

11 **if** $\grave{\boldsymbol{\imath}}_i^{'} \neq \grave{\boldsymbol{\imath}}_i$ **then**

12 $\grave{\boldsymbol{\imath}}_i^{'} \leftarrow \grave{\boldsymbol{\imath}}_i$

13 **else**

14 Dejar el valor del vector de medias actual

15 **end if**

16 **end for**

Fuente: elaboración propia.

El cálculo de las distancias se realiza con la distancia de Minkowski (Xing *et al.*, 2022), la cual se muestra en la Ecuación 5.

$$\text{dist}_{\{\text{mk}\}}\left(x_i, x_j\right) = \left(\sum_{u=1}^{n} \left|x_{iu} - x_{ju}\right|^p\right)^{\frac{1}{p}} \tag{5}$$

La Ecuación 6 muestra cuando *p = 2*, la distancia de Minkowski se convierte en una distancia de Euclidiana.

$$\text{dist}_{\{\text{ed}\}}\left(x_i, x_j\right) = \left\|x_{iu} - x_{ju}\right\|_2 = \sqrt{\left(\sum_{u=1}^{n} \left|x_{iu} - x_{ju}\right|^2\right)^2} \tag{6}$$

En la Ecuación 7 se presenta la distancia de Minkowski al tomar el valor *p = 1*, cuando se convierte en la distancia de Manhattan.

$$\text{dist}_{\{\text{man}\}}\left(x_i, x_j\right) = \left\|x_{iu} - x_{ju}\right\|_1 = \sum_{u=1}^{n} \left|x_{iu} - x_{ju}\right|^2 \tag{7}$$

Los primeros conceptos de la teoría de conjuntos difusos (Fuzzy C-Means) fueron propuestos por Zadeh (1965). Tiempo después, varios autores trabajaron con algoritmos difusos de agrupamiento; sin embargo, en 1973, Dunn (Dunn, 1973) extendió el concepto del agrupamiento de medias duro a conceptos preliminares de medias difusas. Por otro lado, Bezdek, en 1981 (Bezdek, 2013), añadió el factor difuso y propuso el algoritmo de Fuzzy C-Means. Recordemos que un método para resolver problemas de agrupamiento, que involucra

minimizar algunas funciones objetivo y el error de la función se conoce como K-Means, siendo k el número de agrupaciones o clases; sin embargo, si la técnica que se utiliza es de lógica difusa, se conoce como Fuzzy C-Means o FCM. Este algoritmo utiliza un enfoque de una función de membresía, la cual asigna un grado de pertenencia para cada clase. La importancia del grado de membresía en los algoritmos de agrupamiento es que nos permite decidir a qué clase o agrupación se parece más, es decir tiene mayor similitud (Nayak *et al.*, 2014). La ventaja de FCM es la formación de nuevas agrupaciones de los puntos que tienen mayor cercanía a los valores de la función de membresía de las clases existentes.

Dentro de FCM tenemos tres elementos importantes, la función de membresía, la función objetivo y la matriz de partición (Ghosh; Dubey, 2013). Si consideramos un conjunto de vectores $n\ (X= x_1,...x_n),\ 2 \le c \le n)$ para agruparlos en c clases, cada uno se describe por un valor real de medición, que representa las características del objeto. Las matrices de partición que se utilizan para describir la lógica difusa se presentan en la Ecuación 8.

$$\mathrm{M}_{\{\mathrm{fc}\}} = \left\{ \mathrm{W} \in \square^{\mathrm{cn}} \right\} \Big| \omega_{ik} \in [0,1], \forall i; \tag{8}$$

Donde:

$$\sum_{i=1}^{c} = 1, \forall k; \qquad 0 < \sum_{i=1}^{n} \omega_{ik} < n; \ \forall \mathrm{i};$$

$$1 \le i \le c; \qquad 1 \le k \le n.$$

De acuerdo con estas definiciones, se observa que una o más agrupaciones pueden pertenecer a distintas clases con diferentes grados de membresía. La membresía total de un elemento se normaliza a 1 y una sola agrupación puede contener todos los puntos. La función objetivo mostrada en la Ecuación 9 del algoritmo se calcula utilizando el valor de membresía y la distancia Euclidiana 6:

$$\mathrm{J}(\mathrm{W},\mathrm{P}) = \sum_{1 \le k \le n; 0 \le i \le c} (\mathrm{W}_{ik})^m (d_{ik})^2 \tag{9}$$

Donde $m \in (1, \infty)$ es el parámetro que define lo difuso del resultado de las agrupaciones y d_{ik} es la distancia euclidiana del objeto x_k al centroide p_i . La función de membresía se calcula en la Ecuación 10.

$$\mu_{ij} = \left[\sum_{t=1}^{c} \left(\frac{\left\| x_j - v_i \right\| \mathrm{A}}{\left\| x_j - v_i \right\| \mathrm{A}} \right)^{\frac{2}{m-1}} \right]^{-1} \tag{10}$$

Donde μ_{ij} es el valor de la *j*-ésima muestra en el *i*-ésima agrupación. El número de agrupaciones se representa por *c*, x_j es la *j*-ésima muestra y v_i el centroide del *i*-ésimo clúster. $\| \|_A$ representa la norma de la función. El algoritmo de Fuzzy C-Means (medias difusas) se muestra en la Tabla 4.

Tabla 4. *Algoritmo 2, medias difusas*

Algoritmo 2 : Medias Difusas

Entrada : Conjunto de Datos $D = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$

:Numero de agrupaciones C

Salida : Agrupaciones $C = \{C_1, ..., C_k\}$

Se inicializan el número de agrupaciones C

1 $C_i = \varnothing; (1 \le i \le k)$

2 **While** $\left\| P^{(b)} - P^{(b+1)} \right\| \le \in$

3 Seleccionar una métrica de distancia

Seleccionar la métrica de ponderación difusa

4 $b = 0$

5 $P^{(b)} = P^{(0)}$

6 Calcular la matriz de partición

$$W_{ik}^{(b)} = \sum_{j=1}^{c} \left[\frac{d_{ik}^{(b)}}{d_{jk}^{(b)}} \right]^{-\frac{2}{m-1}}$$

7 Actualizar los centroides difusos

$$P^{(b+1)} = \frac{\sum_{k=1}^{n} \left(w_{ik}\right)^m x_k}{\sum_{k=1}^{n} \left(w_{ik}\right)^m}$$

end while

Fuente: elaboración propia.

El coeficiente de silueta es una métrica utilizada para calcular la bondad de una técnica de agrupación; el rango va de -1 a 1. Cuando se obtiene el valor de 1, significa que las agrupaciones están situadas en el centro de la agrupación asignada, si es 0, la distancia entre ellas está en el borde, finalmente, si el valor es menos 1, los grupos están mal asignados. Esta métrica se basa en la geometría de las agrupaciones (Wang; Xu, 2019). El criterio de silueta se muestra en la Ecuación 11:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (11)$$

Donde:

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \ne j} d(i, j),$$

$$b(i) = \min_{k \ne 1} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

3. Resultados y discusión

Como se mencionó anteriormente, el objetivo de los algoritmos K-Means y Fuzzy C-Means es agrupar objetos según su similitud, de forma tal, que los objetos que se encuentren dentro de una misma agrupación o clúster presenten características muy similares y diferentes en gran medida, de aquellos que no pertenezcan a la agrupación o clúster, entendiendo que los elementos que se están agrupando no están etiquetados de manera alguna.

En el caso de estudio, se crearon cuatro agrupaciones de aseguradoras venezolanas al aplicar cada algoritmo, tomando como variables los factores obtenidos con ayuda del análisis factorial. Es importante señalar que se verificó el comportamiento, tanto con los dos primeros factores (73 % de la varianza total) como con los tres factores (87 % de la varianza total), esto con la finalidad de observar el comportamiento de los algoritmos al incrementar la varianza que explican los factores. Los resultados obtenidos se muestran en la Figura 2. Los parámetros de inicio en ambos métodos de agrupación son: los 18 indicadores con dos y tres factores, según sea el caso, el número de agrupaciones antes mencionado, la métrica dada por la Ecuación 6 y los centroides basados en una distribución de probabilidad empírica de la contribución de los puntos a la inercia total. La cardinalidad es representada por el número de aseguradoras que hacen parte de cada agrupación, en la imagen se observa la distribución de las aseguradoras en cada clúster al aplicar los dos algoritmos de clasificación antes mencionados.

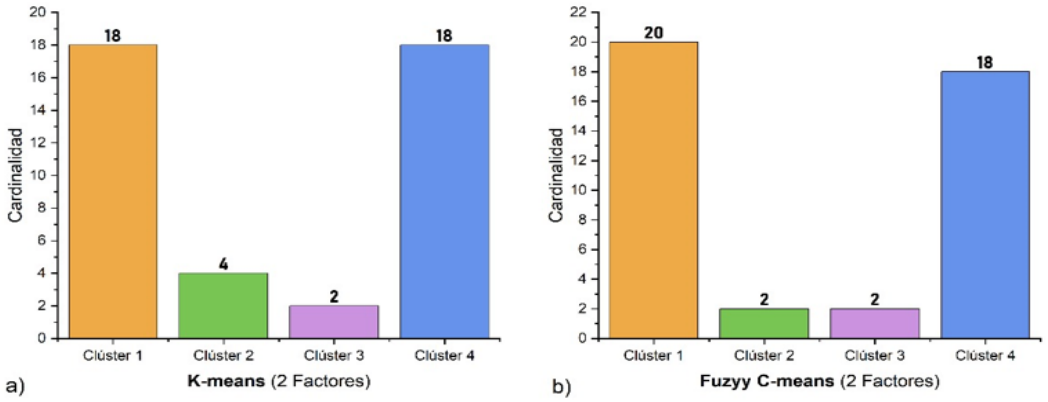

Figura 2. Cardinalidad de agrupaciones de 2 factores. a) K-Means, b) Fuzzy C-Means
Fuente: elaboración propia.

En la Figura 3, al igual que en el caso de dos factores, la diferencia en la distribución de las aseguradoras es poca, reflejando mayor similitud en la distribución con tres factores, lo cual es razonable, dado que se aumenta la varianza, que, si bien es poca, sirve para mostrar mayor similitud en la aplicación de ambos métodos de agrupamiento.

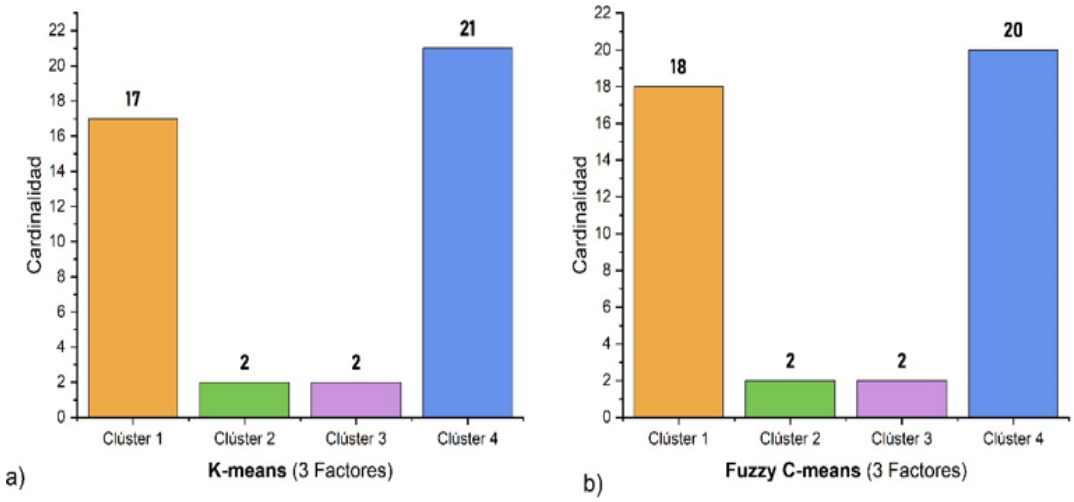

Figura 3. Cardinalidad de agrupaciones de 3 factores. a) K-Means, b) Fuzzy C-Means
Fuente: elaboración propia.

En la Figura 4 se muestran los gráficos de silueta con dos factores para ambos métodos, se nota que la tercera clase de ambos gráficos presenta valores de silueta cercanos a 0,70 para tres factores, y de 0,63 para dos factores, donde se observa una mejora de este criterio en la primera y cuarta clase al incrementar el número de factores. Este cálculo de silueta refleja una disminución en la segunda clase, con valor típico de 0,65 en la clasificación arrojada por K-Means con dos factores, dado por los elementos de diferencia en esta clase respecto a Fuzzy C-Means de igual número de factores.

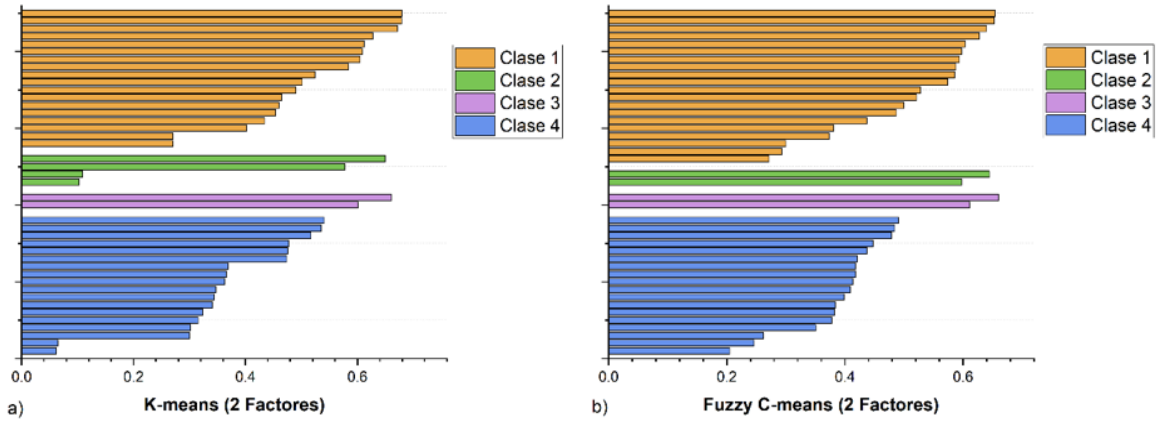

Figura 4. Criterio de Silueta de 2 Factores. a) K-Means, b) Fuzzy C-Means
Fuente: elaboración propia.

En la Figura 5 se muestran los gráficos de silueta con tres factores para ambos métodos, donde el valor medio de 0,58 del criterio de silueta de ambos métodos de agrupación con tres factores es significativamente superior al obtenido de 0,50 con dos factores, esta mejora se refleja en las matrices de confusión calculadas posteriormente. En las agrupaciones obtenidas mediante Fuzzy C-Means con tres factores, el criterio de silueta es de 0,59, siendo este el más alto alcanzado por los algoritmos; este factor se calculó con el mismo método para tres y cinco agrupaciones, con valores de 0,46 y 0,52, respectivamente.

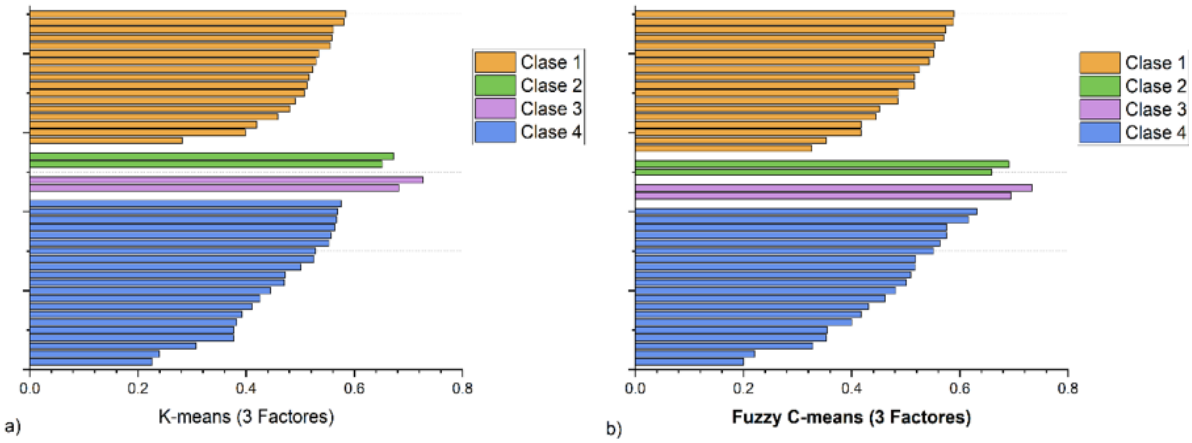

Figura 5. Criterio de Silueta de 3 Factores. a) K-Means, b) Fuzzy C-Means
Fuente: elaboración propia.

El gráfico de las agrupaciones para dos factores se detalla en la Figura 6, las cuales fueron calculadas con ambos algoritmos de clasificación, se presenta una diferencia en los elementos de las dos primeras agrupaciones. La distribución de las aseguradoras, según los dos primeros factores, permiten distinguir visualmente cuán cercanas se encuentran estas empresas. De igual manera permite identificar cuántos factores se deben tomar para obtener una mejor representación del grupo que hace parte del estudio.

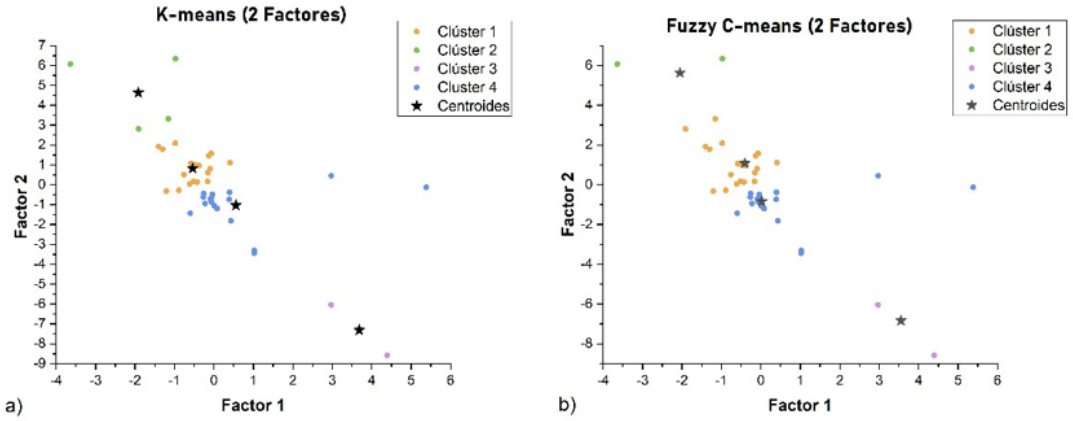

Figura 6. Agrupaciones de 2 Factores. a) K-Means, b) Fuzzy C-Means
Fuente: elaboración propia.

Estas diferencias mínimas se reducen al agrupar con tres factores obtenidos del análisis factorial (Figura 7), y fueron computadas nuevamente con las técnicas de agrupación antes mencionadas, donde sus centroides, calculados por ambos métodos de agrupación-clasificación, se diferencian ligeramente.

Las agrupaciones arrojadas con tres factores por ambos algoritmos se enlistan en la Tabla 5, en ella se observa el identificador de la aseguradora de acuerdo con la Tabla 2, acompañada de las agrupaciones resultantes.

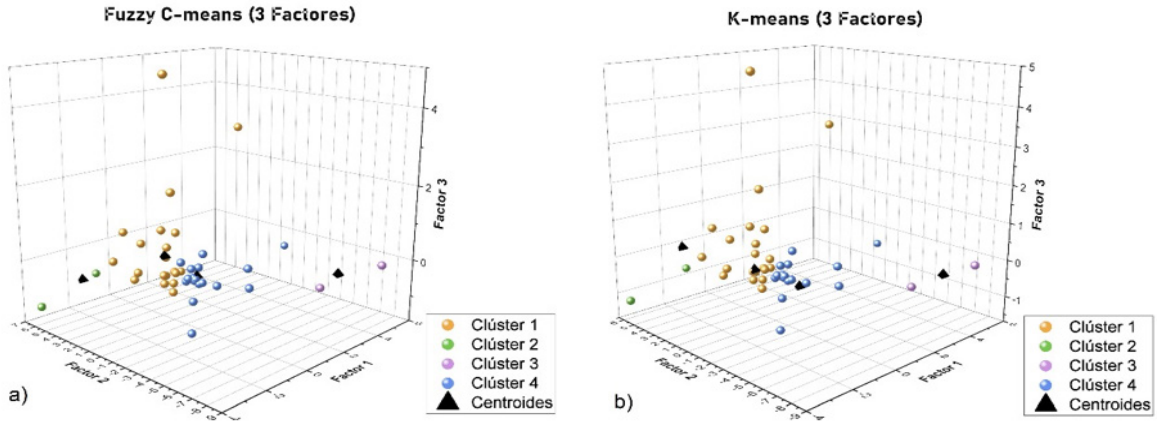

Figura 7. Agrupaciones de 3 Factores. a) Fuzzy C-Means, b) K-Means
Fuente: elaboración propia.

La métrica de semejanza entre las agrupaciones obtenidas por los dos tipos de métodos de agrupación con diferente número de factores, se valida con una matriz de confusión (matriz de aciertos), una herramienta novedosa y sencilla de visualización del desempeño de un algoritmo de clasificación estadística (Pereira *et al.*, 2018), donde cada renglón de la matriz representa a las instancias en la clase real, (experta) y las columnas, el número de predicciones de cada clase (Hardin; Shumway, 1997).

Tabla 5. *Agrupaciones de asegurados con tres factores por ambos métodos de clasificación*

ID	Original	K-Means	FCM	ID	Original	K-Means	FCM	ID	Original	K-Means	FCM
1	1	1	1	15	2	2	2	29	3	1	1
2	1	4	4	16	2	1	1	30	3	4	4
3	1	4	4	17	2	4	4	31	4	4	4
4	1	4	4	18	2	4	4	32	4	4	1
5	1	2	2	19	2	4	4	33	4	1	1
6	1	1	1	20	2	1	1	34	4	1	1
7	1	4	4	21	3	4	4	35	4	1	1
8	1	4	4	22	3	1	1	36	4	1	1
9	1	4	4	23	3	4	4	37	4	4	4
10	1	3	3	24	3	1	1	38	4	1	1
11	2	3	3	25	3	1	1	39	4	1	1
12	2	4	4	26	3	1	1	40	4	1	1
13	2	4	4	27	3	4	4	41	4	4	4
14	2	1	1	28	3	4	4	42	4	4	4

Fuente: elaboración propia.

En la Figura 8 se puede observar el desempeño de las agrupaciones con dos y tres factores entre ambos métodos de agrupación, aplicados a estos factores. De estas matrices, se detalla claramente que el aumento de factores aumenta la similitud entre las agrupaciones dadas por dichos métodos de agrupación; con el uso de los tres factores, la primera agrupación, el método K-Means, conjunta a 17 aseguradoras, mientras que Fuzzy C-Means contiene a esas mismas aseguradoras y otra más, pudiéndose decir que ambos algoritmos tienen un 94 % de coincidencia, la coincidencia total de 100 % se da en la segunda y tercera agrupación y para la cuarta es del 95 % , donde se observa que las dos técnicas de agrupamiento indican confiabilidad de las agrupaciones generadas.

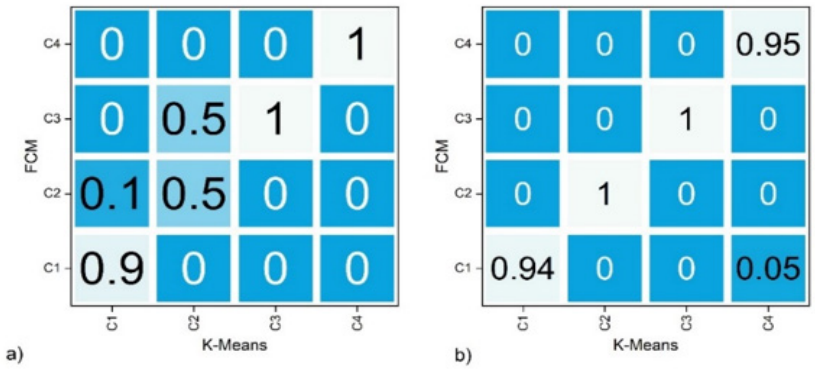

Figura 8. Matriz de confusión de K-Means y FCM. a) dos factores, b) tres factores
Fuente: elaboración propia.

4. Conclusiones

El análisis de agrupaciones es una técnica de clasificación que no requiere tener etiquetas en sus datos (aprendizaje no supervisado). En este estudio se realizó una clasificación inicial, según las primas cobradas en el año 2022, no con la finalidad de crear algún criterio de similitud entre las aseguradoras, sino en el sentido de verificar dicho parámetro. Al aplicar técnicas de agrupamiento, los grupos obtenidos deberían ser similares a la clasificación hecha previamente por la institución gubernamental, ya que esta agrupa sin tomar en cuenta ningún estudio previo, lo que permite afirmar que las agrupaciones se han realizado de una manera arbitraria.

El estudio no permitió aseverar que las primas cobradas representan un factor preponderante en las agrupaciones obtenidas, si bien está presente y es una variable importante, y que debe ser tomada en cuenta a la hora de realizar alguna agrupación de aseguradoras, ella por sí sola no es suficiente para llegar a este fin. Para realizar agrupaciones y para crear una clasificación de aseguradoras, se deben realizar estudios estadísticos, los cuales permitan realizar agrupaciones realistas, basadas en criterios estadísticos, lo que se verá reflejado de una manera clara y justa a las aseguradoras. Así, los asegurados tendrán criterios claros y no sesgados a la hora de elegir su compañía aseguradora

Es importante señalar que el estudio solo pretendía vislumbrar la importancia de las primas cobradas en la conformación de las agrupaciones. Del estudio realizado se deduce que las agrupaciones arrojadas por Fuzzy C-Means de tres factores presentan la mejor calidad de agrupamiento, y esto puede ser tomado como referencia para futuras comparaciones. Una segunda fase del estudio abarcaría determinar cuáles variables y qué peso tienen en la conformación de los grupos, para ello es importante incorporar otras variables que pudiesen ser factor de comparación. Entre algunos ejemplos tenemos: sucursales a nivel nacional, número de empleados, pólizas emitidas, pólizas anuladas, número de intermediarios y clasificación de ellos, etc., con lo cual el algoritmo de clasificación tendría mayor información valiosa para la construcción de las agrupaciones, y las decisiones empresariales que se tomen serían acertadas, redundando en un crecimiento sano del mercado asegurador venezolano.

Agradecimiento

Jorge agradece a la UTEL por la formación académica que ha desarrollado durante el estudio de su maestría en Ciencia de Datos. Elizabeth agradece al UTEL por el apoyo por conducir trabajos de grado en la maestría de Ciencia de Datos. Luis David agradece al Sistema Nacional de Investigadores (SNI-CONAHCYT), con el apoyo a través de la concesión n.° 332238.

Referencias

Abbasi, Bardia; Goldenholz, Daniel (2019). Machine learning applications in epilepsy. *Epilepsia, 60*(10), 2037-2047. https://doi.org/10.1111/epi.16333

Agudelo, Silvio; Bernal, Gabriel (1985). I*mportancia de la Industria aseguradora en el encuentro y desarrollo económico del país* [Tesis de pregrado]. Universidad Autónoma de Occidente. https://red.uao.edu.co/bitstream/handle/10614/2029/T0000344.pdf;jsessionid=133E85280FECC49FFD8323D4847D6F73?sequence=1

Aragón, Ángela; Cerquín, Sabina; Escurra, Renzo; Roncalla Viena, Andrea (2023). *Segmentación de clientes para mejorar la experiencia de compra de productos electrónicos en Falabella* [Tesis de pregrado]. Universidad ESAN. https://hdl.handle.net/20.500.12640/3380

Avedaño, Hernán (2023). La industria aseguradora en el PIB de Colombia. *Revista Faselcolda, 189*, 10-17.

Awad, Hamza; Ibrahim, Hamza; Nor, Sulaiman; Mohammed, Aliyu; Babiker, Abuagla (2012). Taxonomy of Machine Learning Algorithms to classify real- time Interactive applications. *International Journal of Computer Networks and Wireless Communications, 2*, 69-73.

Ayodele, Taiwo (2010). Types of machine learning algoritms. En Zhang; Yagang (Ed.), *New Advances in Machine Learning* (pp. 19-48). IntechOpen. https://doi.org/10.5772/9385

Barbakh, Wesam; Wu, Ying; Fyfe, Colin (2009). Review of Clustering Algorithms. En *Studies in Computational Intelligence book series* (Vol. 249, pp. 7-28). Springer. https://doi.org/10.1007/978-3-642-04005-4_2

Bezdek, James (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

Dehariya, Vinod; Shrivastava, Shaliendra; Jain, R. C. (2010). Clustering of Image Data Set Using K-Means and Fuzzy K-Means Algorithms. *Memorias de la 2010 International Conference on Computational Intelligence and Communication Networks* (pp. 386-391). Instituto de Ingenieros Eléctricos y Electrónicos. https://doi.org/10.1109/CICN.2010.80

Dunn, Joseph (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics, 3*(3), 32-57. https://doi.org/10.1080/01969727308546046

Gobierno Bolivariano de Venezuela (2022). *Superintendencia de la Actividad Aseguradora*. https://www.sudeaseg.gob.ve/descargas/Criterios%20Jurisprudenciales

Ferrando, Pere; Lorenzo-Seva, Urbano; Hernández-Dorado, Ana; Muñiz, José (2022). Decalogue for the factor analysis of test items. *Psicothema, 34*(1), 7-17. https://doi.org/10.7334/psicothema2021.456

Field, Matthew; Hardcastle, Nicholas; Jamenson, Michael; Aherne, Noel; Holloway, Lois (2021). Machine learning applications in radiation oncology. *Physics and Imaging in Radiation Oncology, 19*, 13-24. https://doi.org/10.1016/j.phro.2021.05.007

Ghosh, Soumi; Dubey, Sanjay (2013). Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. *International Journal of Advanced Computer Science and Applications(IJACSA), 4*(4), 35-39. https://doi.org/10.14569/IJACSA.2013.040406

Hardin, Perry; Shumway, Matthew (1997). Statistical significance and normalized confusion matrices. *Photogrammetric engineering and remote sensing, 63*(6), 735-739.

Harki, Edrees; Rashid, Zhala (2023). Analysis of factors that affect radiation dose level during interventional cardiology procedures using logistic regression. *Revista Mexicana de Física, 69*(3), 1-8. https://doi.org/10.31349/RevMexFis.69.031101

Khan, Faisal; Fan, Qirui; Lu, Chao; Lau, Alan (2019). An optical communication's perspective on machine learning and its applications. *Journal of Lightwave Technology, 37*(2), 493-516. https://doi.org/10.1109/JLT.2019.2897313

Kim, Daekyum; Kim, Sang-Hun; Kim, Taekyoung; Kang, Brian; Lee, Minhyuk; Park, Wookeun; Ku, Subyeong; Kim, DongWook; Kwon, Junghan; Lee, Hochang; Bae, Joonbum; Park, Yong-Lae; Cho, Kyu-Jin; Jo, Sungho (2021). Review of machine learning methods in soft robotics. *PLoS ONE, 16*(2), e0246102. https://doi.org/10.1371/journal.pone.0246102.

Kourou, Konstantina; Exarchos, Themis; Exarchos, Konstantinos; Karamouzis, Michalis; Fotiadis, Dimitrios (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8-17. https://doi.org/10.1016/j.csbj.2014.11.005

Libbrecht, Maxwell; Noble, William (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics, 16*, 321-332. https://doi.org/10.1038/nrg3920

MacQueen, J. (1967). Some methods for Classification and analysis of multivariate observations. En *Memorias del 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281-297). Berkeley College.

Mayorga, Wilson (2014). 2013. Un año en transacción. *Revista Fasecolda, 155*, 18-23. https://revista.fasecolda.com/index.php/revfasecolda/issue/view/5

Mc. Coy, J.; Auret, L. (2019). Machine learning applications in minerals processing: A review. *Minerals Engineering, 132*, 95-109. https://doi.org/10.1016/j.mineng.2018.12.004

Mukhamediev, Ravil (2015). Machine learning methods: An overview. *Computer modelling and New Tecnologies, 19*(6), 14-29.

Nateski, Vladimir (2017). An overview of the supervised machine learning methods. *HORIZONS.B, 4*, 51-62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

Nayak, Janmenjoy; Naik, Bighnaraj; Behera, H. S. (2014). Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014. En Jain, Lakhmi; Behera, Himansu; Mandal, Jyotsna; Mohapatra, Durna (Eds.), *Computational Intelligence in Data Mining Volume 2. Smart Innovation, Systems and Technologies* (Vol 32, pp. 133-149). Springer. https://doi.org/10.1007/978-81-322-2208-8_14

Patel, Lauv; Shulka, Tripti; Huang, Xiuzheng; Ussery, David; Wang, Shanzi (2020). Machine Learning Methods in Drug Discovery. *Molecules, 25*(22), 5277. https://doi.org/10.3390/molecules25225277

Pereira, Rafael; Plastino, Alexandre; Zadrozny, Bianca; Merschmann, Luiz (2018). Correlation analysis of performance measures for multi-label classification. *Information Processing & Management, 54*(3), 359-369. https://doi.org/10.1016/j.ipm.2018.01.002

Porter, Michael (2015). *Estrategia competitiva: Técnicas para el análisis de los sectores industriales y de la competencia*. Grupo Editorial Patria.

Sharma, Abhinav; Jain, Arpit; Gupta, Prateek; Chowdary, Vinay (2021). Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access, 9*, 4843-4873. https://doi.org/10.1109/ACCESS.2020.3048415

Taherdoost, Hamed; Sahibuddin, Shamsul; Jalaliyoon, Neda (2022). Exploratory Factor Analysis; Concepts and Theory. *Advances in Applied and Pure Mathematics, 27*, 375-382. https://hal.archives-ouvertes.fr/hal-02557344/document

Wang, Juntao; Xialong, Su (2011). An improved K-Means clustering algorithm. En *Memorias de la 2011 IEEE 3rd International Conference on Communication Software and Networks* (pp. 44-46). Instituto de Ingenieros Eléctricos y Electrónicos. https://doi.org/10.1109/ICCSN.2011.6014384

Wang, Xu; Xu, Yusheng (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering, 569*(5), 1-6. https://doi.org/10.1088/1757-899X/569/5/052024

Wu, Junjie (2012). Cluster Analysis and K-means Clustering: An Introduction. En *Advances in K-means Clustering* (pp. 1-26). Springer. https://doi.org/10.1007/978-3-642-29807-3_1

Xing, Eric; Jordan, Michael; Russell, Stuart; Ng, Andrew (2022). Distance Metric Learning with Application to Clustering with Side-Information. *Advances in neural information processing systems, 15.*

Zadeh, Lotfi (1965). Fuzzy Sets. *Information and Control, 8*(3), 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zhou, Zhi-Hua (2016). *Machine Learning*. Springer. https://doi.org/10.1007/978-981-15-1967-3