DESDE 2023
https://rictrevista.org

RICT Revista de Investigación Científica, Tecnológica e Innovación

Publicación Semestral RICT Vol. 4 No. 7 (2026) P. 16 - 22

ISSN: 2992-7315

Análisis Cluster para detectar patrones específicos entre usuarios de la banca de seguros para identificar posibles fraudes

Abraham-Jorge Jiménez-Alfaro [a,b], Norma-Karen Valencia-Vázquez [b], Griselda Cortés-Barrera [a], Edgar Corona-Organiche [a]

[a] Ingeniería en Sistemas Computacionales, Tecnológico Nacional de México/TES Ecatepec, Laboratorio Nacional Conahcyt en Inteligencia Artificial y Ciencia de Datos (LNC-IACD) Valle de Anáhuac, 55210 Ecatepec de Morelos, Estado de México.
[b] Ingeniería en Sistemas Computacionales, Tecnológico Nacional de México/TES Chimalhuacán, Calle primavera S/N, 56330, Chimalhuacán, Estado de México.

Resumen

El crecimiento acelerado de los servicios financieros digitales ha incrementado de manera significativa la cantidad y complejidad de los datos generados por los usuarios, lo que ha reducido la efectividad de los métodos tradicionales de detección de fraude basados en reglas fijas. En este contexto, las técnicas de aprendizaje no supervisado ofrecen alternativas flexibles para el análisis exploratorio de grandes volúmenes de información. El presente estudio propone un enfoque basado en técnicas de clustering para identificar patrones de comportamiento atípicos asociados a posibles fraudes en la banca de seguros. Para ello, se emplean los algoritmos k-means y clustering jerárquico sobre un conjunto de datos obtenido a partir de una encuesta estructurada aplicada a más de mil usuarios. Los resultados muestran la existencia de grupos claramente diferenciados, donde los conglomerados minoritarios y alejados de los centroides principales representan señales tempranas de riesgo que pueden apoyar los procesos de auditoría, control interno y toma de decisiones en instituciones aseguradoras.

Palabras clave: Aprendizaje no supervisado, clustering, detección de anomalías, banca de seguros, fraude

Abstract

The rapid expansion of digital financial services has significantly increased both the volume and complexity of user-generated data, thereby reducing the effectiveness of traditional rule-based fraud detection approaches. In this context, unsupervised learning techniques provide flexible alternatives for exploratory data analysis. This study proposes a clustering-based analytical framework to identify atypical behavioral patterns associated with potential fraud in insurance banking. The k-means and hierarchical clustering algorithms are applied to data collected through a structured survey conducted with more than one thousand insurance banking users. The results reveal the presence of distinct behavioral groups, where minority clusters located far from the main centroids act as early indicators of risk, supporting auditing, internal control, and decision-making processes within insurance institutions.

Keywords: Unsupervised learning, clustering, anomaly detection, insurance banking, fraud

1. Introducción

La digitalización de los servicios financieros ha transformado profundamente la manera en que los usuarios interactúan con las instituciones aseguradoras. El uso generalizado de plataformas electrónicas, aplicaciones móviles y sistemas automatizados ha dado lugar a entornos caracterizados por grandes volúmenes de datos heterogéneos, generados de forma continua y a alta velocidad. Este escenario ha incrementado la complejidad de los procesos de análisis y ha favorecido la aparición de nuevas modalidades de fraude, cada vez más difíciles de detectar mediante mecanismos tradicionales.

El fraude en sistemas de pólizas de seguros representa un problema creciente para las organizaciones financieras y comerciales. Las técnicas tradicionales de detección basadas en reglas resultan insuficientes ante patrones complejos y cambiantes. El aprendizaje no supervisado, particularmente el clustering, ha demostrado ser eficaz para descubrir estructuras ocultas en los datos (Jain, Murty and Flynn, 1999).

*Autor para la correspondencia: ajimenez@tese.edu.mx
Correo electrónico: ajimenez@tese.edu.mx (Abraham-Jorge Jiménez-Alfaro), karenvalencia@teschi.edu.mx (Norma-Karen Valencia-Vázquez) gcortes@tese.edu.mx (Griselda Cortés-Barrera), ecorona@tese.edu.mx (Edgar Corona-Organiche).

Historial del manuscrito: recibido el 16/01/2026, última versión-revisada recibida el 04/02/2026 aceptado el 10/02/2026, publicado el 12/02/2026. **DOI:** https://doi.org/10.5281/zenodo.18625522

Los enfoques clásicos para la detección de fraude, basados en reglas estáticas o modelos supervisados, dependen en gran medida de patrones previamente conocidos y de conjuntos de datos etiquetados. Sin embargo, estas estrategias presentan limitaciones importantes en entornos reales, donde los comportamientos fraudulentos cambian constantemente y los datos etiquetados suelen ser escasos, incompletos o costosos de obtener. En este contexto, las técnicas de aprendizaje no supervisado emergen como una alternativa robusta para el análisis exploratorio de grandes volúmenes de datos, permitiendo descubrir estructuras ocultas sin requerir conocimiento previo de las clases (West et al., 2016); entre estas técnicas, el clustering o análisis de conglomerados se ha consolidado como una herramienta fundamental para identificar patrones de comportamiento entre consumidores (MacQueen, 1967; Everitt et al., 2011; Zimek et al., 2012).

El clustering permite agrupar elementos con características similares en función de variables como frecuencia de consumo, montos de transacción, monto de reclamos para una póliza de seguro, horarios de actividad, ubicación geográfica o uso de dispositivos para efectuar el reclamo. A partir de esta segmentación, es posible identificar grupos minoritarios o aislados que presentan comportamientos atípicos, los cuales pueden estar asociados a actividades fraudulentas o de alto riesgo. La relevancia del clustering en la detección de fraudes radica en su capacidad para adaptarse a escenarios dinámicos y altamente complejos. Al no depender de etiquetas predefinidas, estos métodos pueden detectar nuevas modalidades de fraude que aún no han sido formalmente identificadas. Estudios previos han demostrado que los algoritmos de clustering, como k-means, clustering jerárquico y métodos basados en densidad, son eficaces para el análisis de anomalías y la identificación de outliers en datos de consumo (Jain et al., 1999; Aggarwal, 2017).

El objetivo de este trabajo es aplicar un modelo de análisis basado en técnicas de clustering que facilite la identificación de patrones de comportamiento relevantes para la detección temprana de posibles fraudes en la banca de seguros. Para ello, se presenta una metodología cuantitativa, el desarrollo matemático del modelo y un análisis interpretativo de los resultados obtenidos, con el propósito de contribuir al fortalecimiento de los procesos de evaluación y toma de decisiones en instituciones aseguradoras. El estudio busca contribuir al desarrollo de estrategias más flexibles y eficientes para la detección temprana de fraudes en sistemas de consumo modernos, las pólizas de seguros (Chandola et al., 2009; Ngai et al., 2011).

2. Materiales y Método

El Análisis Cluster, conocido como Análisis de Conglomerados, es una técnica estadística multivariante que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos (Xu & Wunsch, 2009).

El análisis adopta un enfoque cuantitativo y exploratorio, basado en técnicas de aprendizaje no supervisado, se centra en la identificación de patrones de comportamiento a partir de variables demográficas y financieras, sin asumir previamente la existencia de categorías de fraude.

El análisis de conglomerados es un método de carácter exploratorio cuyo propósito es identificar subconjuntos de observaciones con comportamientos similares, maximizando la cohesión interna de cada grupo y la diferenciación respecto a otros grupos, a partir de la información contenida en los propios datos. A partir de estructuras de casos-variables, trata de situar los casos (elementos) en grupos homogéneos, conglomerados o clusters, no conocidos de antemano, pero sugeridos por la propia esencia de los datos, de manera que elementos que puedan ser considerados similares sean asignados a un mismo cluster, mientras que otros diferentes (disimilares) se localicen en clusters distintos (Kaufman and Rousseeuw, 2005).

La diferencia esencial con el análisis discriminante consiste en que en esté análisis es necesario especificar previamente los grupos por un camino objetivo. El análisis cluster define grupos tan distintos como sea posible en función de los propios datos. La creación de grupos basados en similaridad de casos exige una definición de este concepto, o de la complementaria distancia entre los elementos. La variedad de formas de medir diferencias multivariables o distancias entre casos proporciona diversas posibilidades de análisis. El empleo de ellas, y el de las que continuamente siguen apareciendo, así como, de los algoritmos de clasificación, o diferentes reglas matemáticas para asignar los elementos a distintos grupos, depende del aspecto estudiado y del conocimiento previo de posible agrupamiento que de él se tenga. Puesto que la utilización del análisis cluster ya implica un desconocimiento o conocimiento incompleto de la clasificación de los datos, el investigador ha de ser consciente de la necesidad de emplear varios métodos con el fin de contrastar los resultados (Han et al., 2012).

Para Kaufman (2005) existen dos grandes tipos de análisis de clusters: no jerárquicos y jerárquicos. Se conocen como no jerárquicos a aquellos que asignan los casos o grupos diferenciados que el propio análisis configura, sin que unos dependan de otros. Los métodos no jerárquicos pueden, a su vez, producir clusters disjuntos (cada caso pertenece sólo a un cluster), o bien clusters solapados (un caso puede pertenecer a más de un grupo). Estos últimos de difícil interpretación, son poco utilizados. Se denominan jerárquicos a los que configuran grupos con estructura arborescente, de forma que clusters de niveles más bajos van siendo englobados en otros clusters de

niveles superiores.

Para Tan et al. (2016) una vez finalizado un análisis de clusters, el investigador dispondrá de una colección de casos agrupada en subconjuntos jerárquicos o no jerárquicos. Podrá aplicar técnicas estadísticas comparativas convencionales siempre que lo permita la relevancia práctica de los grupos creados; así como, otras pruebas multivariantes, para las que ya contará con una variable dependiente grupo, aunque haya sido creada artificialmente. Sugiere un proceso para estructurar el análisis de cluster:

2.1.- Elección de variables

Dependiendo del problema las variables pueden ser cualitativas (Ordinales y Nominales) y cuantitativas (discretas y continuas).

2.2.- Elección de la medida de asociación

Para poder unir variables es necesario tener algunas medidas numéricas que caractericen las relaciones entre las variables. Cada medida refleja asociación en un sentido particular y es necesario elegir una medida apropiada para el problema concreto que se esté tratando.

La medida de asociación puede ser una distancia o una similaridad:

- Cuando se elige una distancia como medida de asociación los grupos formados contendrán elementos parecidos de forma que la distancia entre ellos tiene que ser pequeña.

- Cuando se elige una medida de similaridad los grupos formados contendrán elementos con una similaridad alta entre ellos. La correlación de Pearson y los coeficientes de Spearman y de Kendall son índices de similitud.

2.3.- Elección de la técnica de Cluster por Métodos Jerárquicos

Agrupar el cluster para formar uno nuevo o separar alguno ya existente para dar origen a otros dos de forma que se maximice una medida de similaridad o se minimice alguna distancia, la asociación establecida es:

- Asociativos o Aglomerativos: Se parte de tantos grupos como elementos hay en el estudio y se van agrupando hasta llegar a tener todos los casos en un mismo grupo.

- Disociativos: Se parte de un solo grupo que contiene todos los casos y a través de sucesivas divisiones se forman grupos cada vez más pequeños.

Los métodos jerárquicos permiten construir un árbol de clasificación o dendograma El algoritmo k-means busca minimizar la suma de distancias cuadráticas dentro de cada grupo.

3. Resultados

Los datos fueron procesados y analizados mediante herramientas computacionales desarrolladas en Python, enfocadas en la exploración estructurada de la información y la identificación de patrones de comportamiento de los usuarios que solicitan pólizas de seguro. El análisis permitió caracterizar el perfil de uso y los hábitos para el fraude de los usuarios. La elección de este enfoque se fundamenta en la ausencia de etiquetas confiables de fraude y en la naturaleza dinámica de los comportamientos de consumo, lo cual hace inviable el uso exclusivo de modelos supervisados. Al caracterizar tendencias generales, comparar comportamientos entre distintos grupos de usuarios y examinar la relación funcional entre variables relevantes permite apoyar la interpretación integral de los datos y la toma de decisiones.

El proceso de análisis (Hand et al., 2001; Bishop, 2006; Tan et al., 2016), se estructuró en las siguientes etapas:

1. Preparación de los datos.
2. Normalización de variables.
3. Cálculo de medidas de distancia.
4. Aplicación del algoritmo k-means.
5. Clustering jerárquico aglomerativo.

3.1.- Preparación de los datos

Sea un conjunto de datos compuesto por n consumidores de pólizas de seguros, donde cada consumidor está descrito por un vector de características multidimensional:

$$x_i = (x_{i1}, x_{i2}, \dots, x_{in}) i = 1,2, \dots, n \quad (1)$$

Donde p representa el número de variables observadas, tales como:

1. Monto promedio de transacciones.
2. Frecuencia de consumo.
3. Horario de actividad.
4. Ubicación geográfica.
5. Ingreso.
6. Edad.
7. Monto de reclamación.
8. Otros.

El conjunto completo de datos puede representarse como una matriz:

$$X \in R^{n*p} \quad (2)$$

3.2.-Normalización de variables

Dado que las variables pueden estar medidas en diferentes escalas, se aplica una normalización tipo z-score para evitar sesgos en el cálculo de distancias:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (3)$$

3.3.- Medidas de distancia

El criterio fundamental del clustering es la medición de similitud entre observaciones. En este estudio se emplea principalmente la distancia euclidiana, definida como:

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2} \quad (4)$$

3.4.- Algoritmo k-means

El algoritmo k-means busca particionar el conjunto de datos en k clusters disjuntos C_1, C_2, …, C_k minimizando la función objetivo conocida como suma de cuadrados intracluster:

$$j = \sum_{i=1}^{k} \sum_{X \in C_i} \| x - \mu_i \|^2 \quad (5)$$

μ_i es el centroide del cluster C_i , calculado como:

$$\mu_{i=\frac{1}{|C_i|}} \sum_{x \in C_i} x \quad (6)$$

El algoritmo opera de manera iterativa siguiendo estos pasos:

1. Inicialización de k centroides.
2. Asignación de cada observación al centroide más cercano.
3. Recálculo de los centroides.
4. Repetición hasta la convergencia del criterio J.

3.5.-Clustering jerárquico aglomerativo

Como método, se emplea el clustering jerárquico aglomerativo, el cual no requiere especificar previamente el número de clusters. La distancia entre clusters puede definirse mediante distintos criterios. En este análisis se considera el método de enlace promedio:

$$d(C_a, C_b) = \frac{1}{|C_a||C_b|} \sum_{x \in C_a} \sum_{y \in C_b} d(x, y) \quad (7)$$

Se realizó un análisis orientado a la exploración integral de la información obtenida a partir de una encuesta estructurada aplicada a más de mil usuarios de servicios de banca de seguros. El estudio se centró en identificar patrones de comportamiento y relaciones relevantes entre las variables, con especial énfasis en los reclamos asociados a pólizas de seguro.

La información recolectada consideró aspectos demográficos y financieros de los usuarios, tales como edad, nivel de ingresos, frecuencia de uso del servicio, horarios de actividad y montos de reclamación, con el propósito de comprender la dinámica del uso de los servicios y apoyar el proceso de interpretación de los resultados. El análisis de los datos muestra que hay tres conjuntos de picos para los montos de las reclamaciones con centros iniciales en 10,00;20,000 y 30,000. En esta primera aproximación es necesario complementar el análisis de relaciones entre variables para identificar si hay alguna correlación entre diferentes comportamientos o prácticas de los consumidores, véase figura 1.

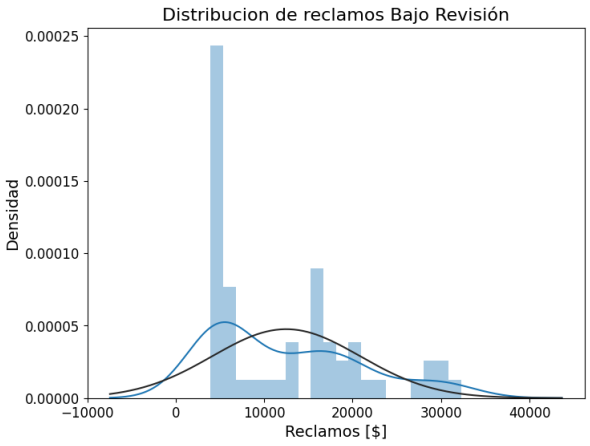

Figura 1: Distribución normal de los reclamos.

Aplicando un análisis más profundo con los diagramas de dispersión de reclamos, véase figura 2, se puede observar que hay un grupo alto de reclamos alrededor del rango de ingresos de $30,000 a $40,000 en el gráfico de reclamos frente a ingresos, lo que podría deberse al hecho de que el ingreso medio es de

aproximadamente $30,000 a $40,000; también hay una franja de reclamos de $50,000 a $100,000 que representa el ingreso alto. Hay una franja de reclamos por al menos $20,000 entre personas que solo ganan $10,000, lo cual es inusual y bien puede consistir en reclamos fraudulentos.

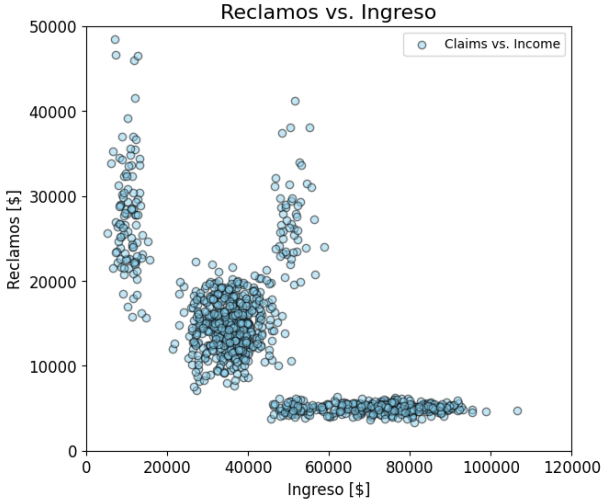

Figura 2: Reclamos contra ingresos.

En el gráfico de dispersión edad frente a ingresos, hay una franja de personas que ganan $10,000 en todas las edades (salario mínimo), un gran grupo de personas que ganan entre $30,000 a $40,000 en todas las edades (salario medio) y hay más personas de mayores ingresos ($60,000 a $100,0000) justo antes de los 60 años, lo que implicaría los ahorros a esas edades, véase figura 3.

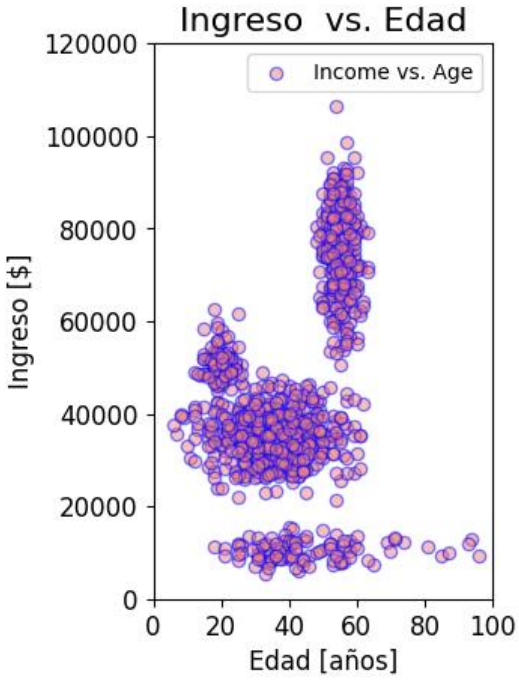

Figura 3: Reclamos Ingreso contra Edad.

En las visualizaciones anteriores, existen distintas poblaciones de elementos en función de las relaciones entre las reclamaciones y los ingresos, y los ingresos y la edad, el modelo de agrupamiento permitió resumir y detectar relaciones potencialmente interesantes. Hay muchas variaciones de agrupación, la empleada en el análisis es K-means clustering. El algoritmo empleado separa el conjunto de datos dado en grupos que minimizan la suma de los cuadrados de las distancias entre cada par de puntos en el grupo (Aggarwal, 2017; Zimek et al., 2012). Para determinar el número de clusters se empleó el método del codo, con resultado en k=4 los clusters óptimos, basados en los datos recabados, véase figura 4.

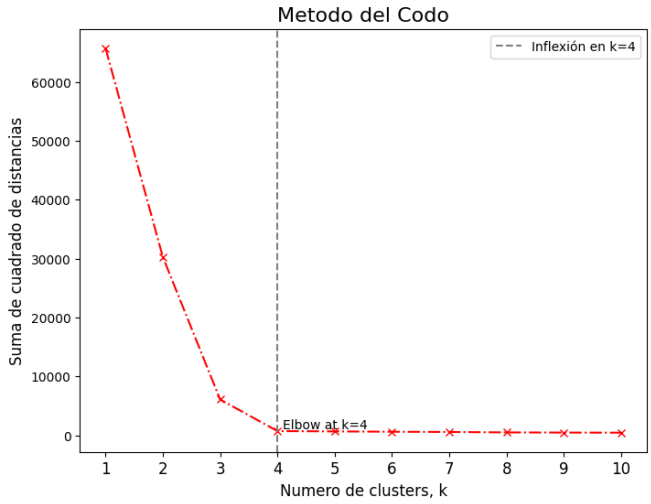

Figura 4: Gráfica del método del codo para determinar el número de k clusters.

El análisis de relaciones entre variables identifica si hay alguna relación entre diferentes comportamientos o prácticas de los consumidores. Los pasos que se pueden resumir de la siguiente manera:

1.- Inicialización. Para comenzar, se deben seleccionar k =4.
2.- Asignación. Cada punto de datos se asigna al grupo correspondiente al centroide más cercano.
3.- Actualización. Una vez que todos los puntos de datos se han asignado a sus respectivos conglomerados, se calcula un nuevo centroide para cada conglomerado tomando la media de todos los puntos en ese conglomerado.

El análisis confirma que el clustering permite la detección temprana de anomalías que requieren análisis complementarios (Hastie et al., 2009; Fawcett and Provost, 1997). La selección de variables y métricas de distancia influye directamente en la calidad de los clusters obtenidos (Tan et al., 2016). Con base en los datos normalizados se obtienen los siguientes clusters, véase figura 5, en donde hay 4 grupos:

1. Altos ingresos y bajos reclamos.
2. Ingresos moderados y reclamos moderados.
3. Ingresos moderados y reclamos elevados.
4. Bajos ingresos y altos reclamos.

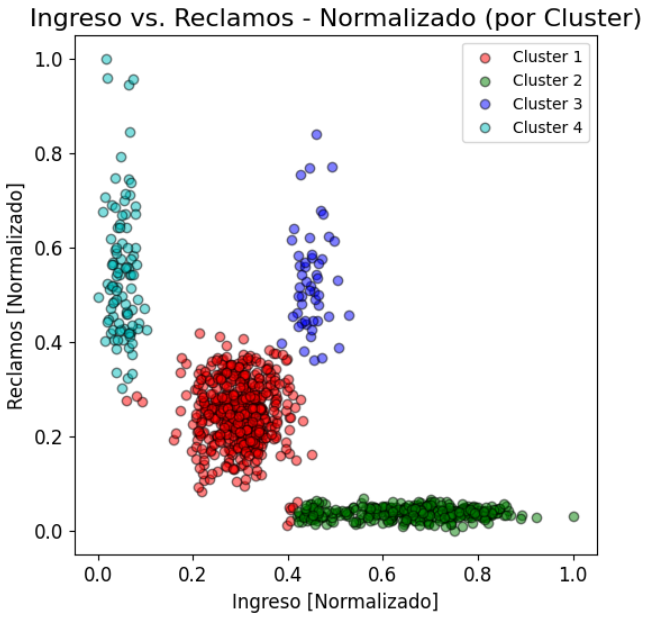

Figura 5: Cuatro clusters con el método de agrupamiento.

Los resultados obtenidos evidencian la capacidad del clustering para segmentar a los usuarios de la banca de seguros en grupos con perfiles de comportamiento diferenciados. Los clusters mayoritarios representan patrones esperados de uso del servicio, mientras que los conglomerados minoritarios concentran observaciones con características inusuales.

4. Discusión

El análisis de los datos permitió identificar cuatro grupos de comportamiento claramente diferenciados. Los clusters principales agrupan a usuarios con perfiles coherentes entre nivel de ingresos y montos de reclamación, lo cual sugiere comportamientos consistentes con prácticas habituales en la banca de seguros. Por otro lado, los clusters minoritarios presentan combinaciones atípicas, como bajos niveles de ingreso asociados a montos de reclamación elevados o frecuencias de uso inusuales.

Estos grupos alejados de los centroides principales representan señales de alerta que pueden ser interpretadas como posibles indicadores de riesgo. Es importante destacar que la identificación de dichos clusters puede implicar la confirmación de fraude con la detección de patrones. La agrupación de la figura 5, permite interpretar lo siguiente:

1.- El cluster de altos ingresos y bajos reclamos. - Son aquellos consumidores con altos ingresos y bajos reclamos, que probablemente sean reclamos ordinarios hechos por familias ricas. Es muy probable que estos no sean fraudulentos y que la banca los acepte y emita las pólizas.

2.- El cluster de Ingresos moderados y reclamos moderados. - Son aquellos consumidores con ingresos moderados con valores de reclamación moderados. Estos son muy abundantes y podrían ser artículos cotidianos. Es muy probable que estos no sean fraudulentos y que la banca los acepte y emita las pólizas.

3.- El cluster de Ingresos moderados y reclamos elevados. - Son aquellos consumidores con los ingresos moderados y con valores de reclamos elevados. Esto podría ser aceptado si es algo que las personas de ingresos medios necesitan pero que no siempre pueden pagar, como ciertos reclamos de salud, educación, siniestros, entre otros. Así que probablemente se deberían de investigar a fondo, para emitir la póliza y evitar el fraude.

4.- El cluster de bajos ingresos y altos reclamos. - La categoría final es ingresos bajos, pero con valores de reclamos muy altos. Claramente, estos no son asequibles y, con la excepción de algo como los reclamos de salud y los demás mencionados, son catalogados intentos de obtener dinero gratis. Lo más probable es que se deban de rechazar por posible fraude.

5. Conclusiones

Los resultados muestran que los consumidores se agrupan en segmentos claramente diferenciados. Los clusters pequeños y alejados de los centroides principales representan comportamientos atípicos. Estos grupos fueron identificados como candidatos a fraude, reflejando la utilidad del clustering como herramienta exploratoria.

Los hallazgos concuerdan con estudios previos que destacan la eficacia del clustering para detección de anomalías (Aggarwal, 2017). Sin embargo, los resultados dependen de la selección de variables y métricas de distancia.

El clustering constituye una herramienta poderosa para identificar patrones de consumo y posibles fraudes. Su carácter no supervisado permite adaptarse a escenarios cambiantes. Futuras investigaciones pueden integrar clustering con técnicas supervisadas para mejorar la precisión

Como menciona Han et al., (2012) el clustering constituye una herramienta robusta y flexible para la detección exploratoria de fraudes en la banca de seguros, especialmente

en escenarios dinámicos y con datos no etiquetados. El análisis de los datos permitió identificar cuatro grupos de comportamiento claramente diferenciados. Los clusters principales agrupan a usuarios con perfiles coherentes entre nivel de ingresos y montos de reclamación, lo cual sugiere comportamientos consistentes con prácticas habituales en la banca de seguros. Por otro lado, los clusters minoritarios presentan combinaciones atípicas, como bajos niveles de ingreso asociados a montos de reclamación elevados o frecuencias de uso inusuales.

Estos grupos alejados de los centroides principales representan señales de alerta que pueden ser interpretadas como posibles indicadores de riesgo para fraude. Es importante destacar que la identificación de dichos clusters implica la confirmación de fraude complementado con un análisis por parte de especialistas.

6. Agradecimientos

Los autores expresan su agradecimiento a la empresa aseguradora que colaboró en el desarrollo de la presente investigación, por las facilidades otorgadas para el acceso a la información necesaria y por su disposición para apoyar actividades de análisis académico orientadas a la mejora de la gestión del riesgo en el sector de la banca de seguros.

Asimismo, se reconoce de manera especial al personal técnico y administrativo de la institución participante que contribuyó con la provisión, validación y contextualización de los datos utilizados en el estudio, cuya colaboración fue fundamental para la correcta comprensión de los procesos y para la interpretación de los resultados obtenidos.

Finalmente, los autores reconocen a todas las personas e instancias que, de manera directa o indirecta, participaron en el desarrollo de esta investigación, reiterando que el tratamiento de la información se realizó bajo principios de confidencialidad, ética y uso responsable de los datos.

7. Referencias

Aggarwal, C. C. (2017). *Outlier analysis*. Springer.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. En *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms* (pp. 1027–1035). Society for Industrial and Applied Mathematics.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–255.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.

Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery, 1*(3), 291–316.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis*. Wiley.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. En *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). University of California Press.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(1), 86–97. https://doi.org/10.1002/widm.53

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559–569.

Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining* (2nd ed.). Pearson.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244.

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security, 57*, 47–66.

Xu, R., & Wunsch, D. (2009). *Clustering*. Wiley-IEEE Press.

Zimek, A., Schubert, E., & Kriegel, H. P. (2012). A survey on unsupervised outlier detection. *ACM Computing Surveys, 44*(4), 1–42. https://doi.org/10.1145/2071389.2071395