

Técnicas de Agrupamiento aplicado a analizar la siniestralidad de una aseguradora

Irwinng Cabrera Rodríguez

Noviembre 2025

1 Introducción

El análisis de agrupamiento también conocido como *clustering* es una técnica de aprendizaje no supervisado que se utiliza para descubrir grupos o patrones ocultos dentro de un conjunto de datos. El propósito del análisis de agrupamiento es encontrar patrones naturales en los datos, resumir grandes volúmenes de información y identificar segmentos, perfiles o comportamientos similares.

Para este trabajo se analiza el comportamiento del Grupo GZ que cuenta con 38 observaciones, las cuales son variables numéricas que describen el comportamiento de la grupo analizar.

Para este análisis se aplicaran metodología como *K-Medias* para la determinación del número de grupos.

2 Metodología

El conjunto de datos se compone de variables categóricas y numéricas que describen las características de la cuenta GZ.

Los datos los trabajamos con el programa Python con la función `StandardScaler`, después, se trabaja con el algoritmo OPTICS (agrupamiento jerárquico).

2.1 OPTICS (agrupamiento jerárquico)

OPTICS es una extensión del algoritmo DBSCAN, por lo cual, OPTICS genera una estructura jerárquica de clústeres, donde puedes ver cómo los grupos se forman y se dividen al variar la densidad.

Ventajas

1. Detecta clústeres de distinta densidad (DBSCAN no puede).
2. Identifica ruido y puntos atípicos automáticamente (label = -1).
3. Produce una estructura jerárquica de clústeres

Desventajas

1. Es más lento que *K-Medias* o DBSCAN
2. Más difícil de interpretar

3 Resultados

3.1 Gráfica de distancia de alcance

Como se ve en la figura 1 (p. 2).

En la Gráfica de distancia de alcance se puede observar una gran cantidad de puntos están muy cerca entre sí, formando un clúster denso y bien definido.

La elevación gradual de la curva muestra una transición hacia regiones cada vez menos densas, posterior a lo observado, a forma general No presenta múltiples valles profundos, esto se define, que observa un un clúster dominante principal y largo.

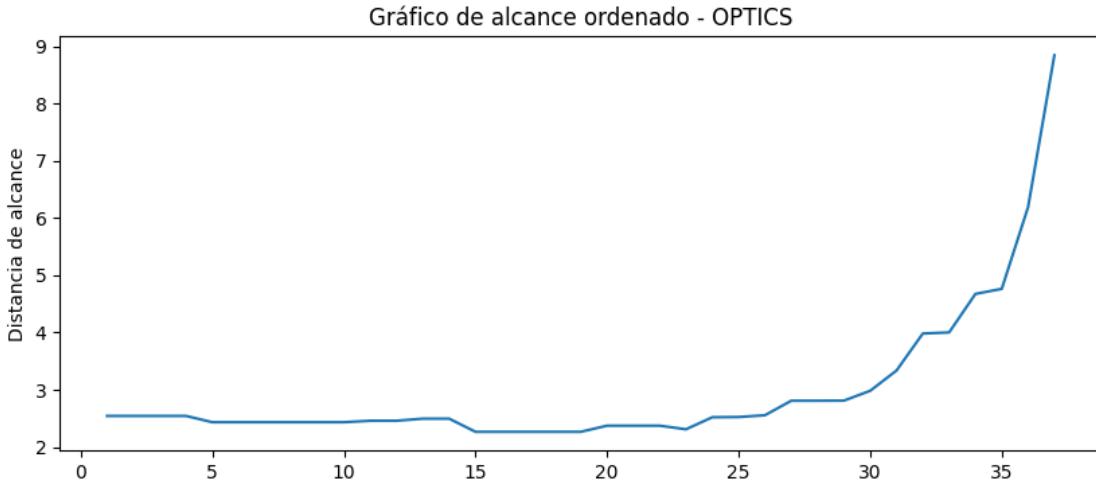


Figure 1: Gráfico de alcance ordenado - OPTICS

3.2 Método del codo

Como se ve en la figura 2 (p. 3).

La gráfica nos indica una disminución de la inercia conforme aumenta el número de clústeres, debido a que una mayor partición de los datos permite que los puntos se agrupen en regiones más homogéneas y, por lo tanto, reduzcan su distancia al centroide asignado.

El Método del Codo indica que el número óptimo de clústeres para el conjunto de datos analizado se encuentra en el rango de 4 a 5.

3.3 Visualización t-SNE

Como se ve en la figura 3 (p. 3).

La gráfica t-SNE muestra una dispersión amplia de puntos, esta dispersión no implica necesariamente la presencia de clústeres, por lo tanto, la ausencia de múltiples clústeres sugiere que el conjunto de datos presenta una distribución continua y con densidad separada.

4 Conclusión

Como se puede notar, los datos analizar solo tienen 38 observaciones, las cuales pueden ser insuficientes para los análisis realizados, el conjunto de gráficas obtenidas durante el análisis de agrupamiento permite concluir que los datos no presentan una estructura de agrupamiento claramente definida, por consiguiente, el conjunto de datos analizado carece de particiones naturales o clústeres definidos, mostrando más bien un comportamiento continuo.

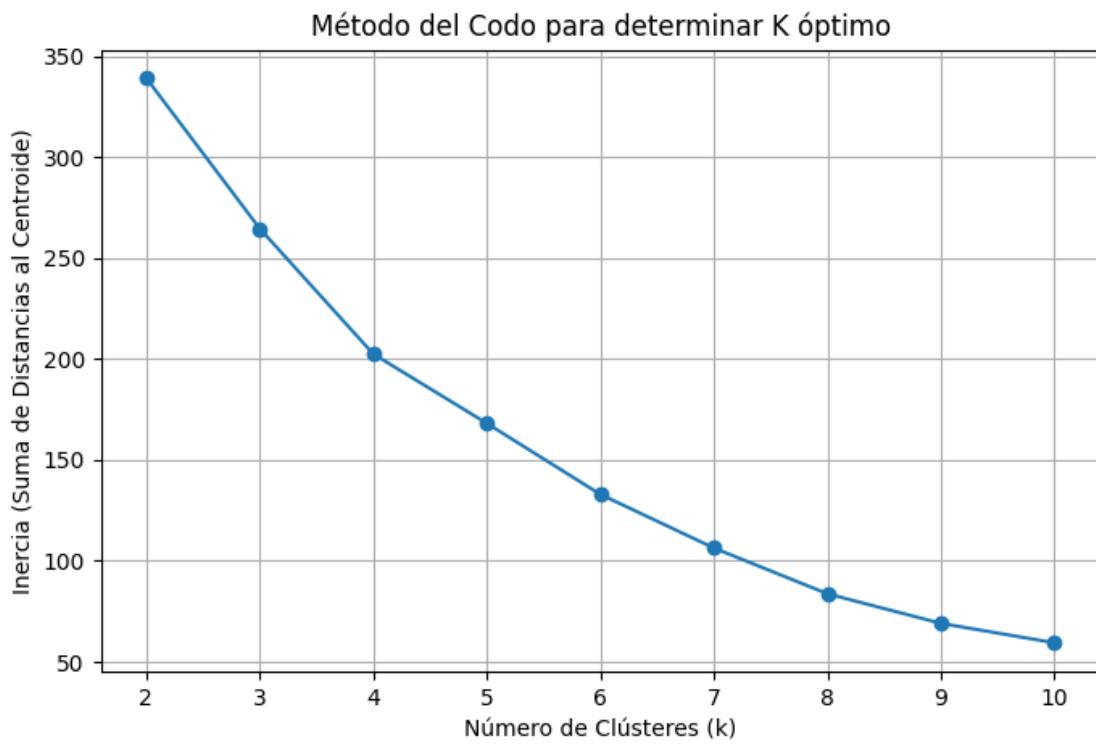


Figure 2: Método del codo para elegir el número de grupos en el algoritmo de K -medias.

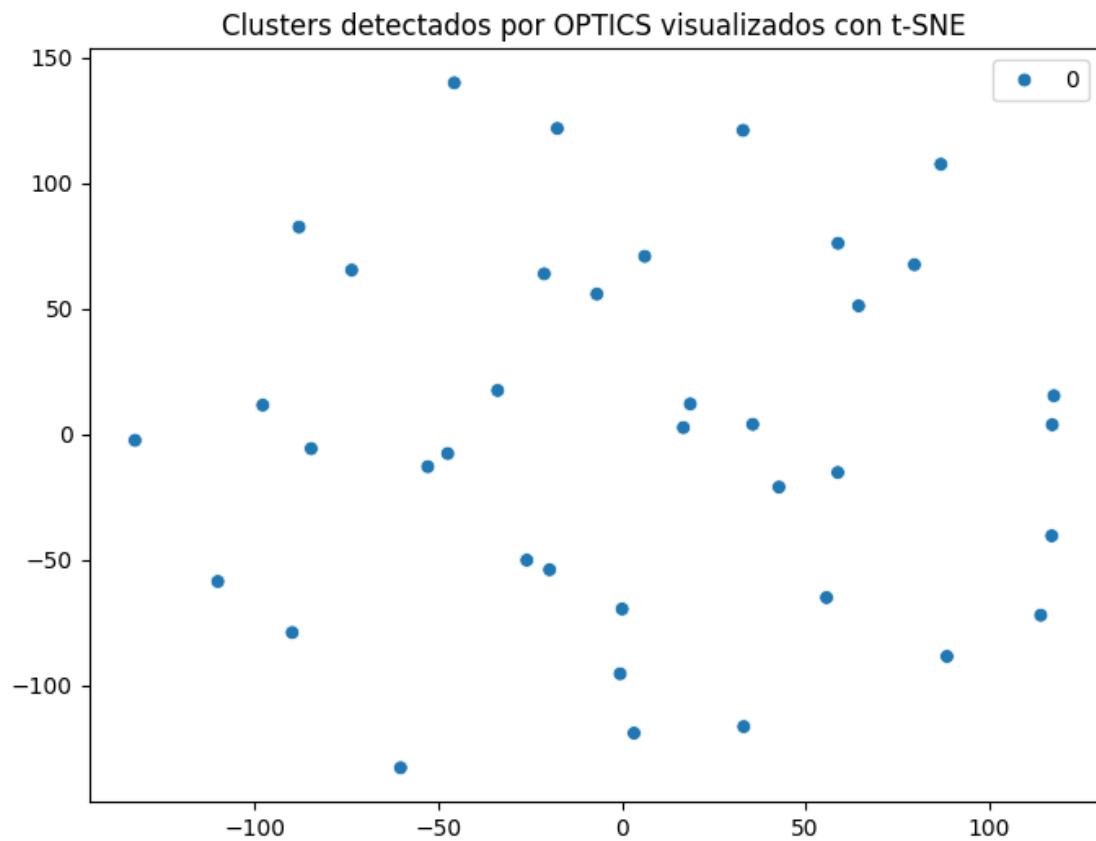


Figure 3: Agrupacion detectadas por Optics visualizados con t-sen.