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 Previo

Un *seguro de auto* es un contrato entre una aseguradora y el cliente que tiene el objetivo de protegerte su vehículo en cuestión económica en caso de que ocurra un accidente, robo u otro daño relacionado con tu vehículo.

2.1 ¿Cómo funciona?

Tú pagas una prima y, a cambio, la aseguradora se compromete a cubrir ciertos gastos según la póliza que contrates.

2.2 Conceptos clave en una póliza de seguros

1. **Asegurado:** La persona o entidad que recibe la protección del seguro.
2. **Contratante:** Quien compra y paga la póliza.
3. **Beneficiario:** Persona que recibe la indemnización en caso de siniestro.
4. **Aseguradora:** Empresa que cubre económicamente en caso de un accidente.
5. **Prima:** El costo del seguro.
6. **Coberturas:** Todo lo que sí está protegido por el seguro.
7. **Exclusiones:** Situaciones que no están cubiertas.
8. **Suma asegurada:** El límite máximo que la aseguradora pagará.
9. **Deducible:** Cantidad que pagas antes de que el seguro cubra el resto.
10. **Vigencia:** Periodo durante el cual la póliza está activa.
11. **Siniestro:** Evento que activa el seguro.
12. **Indemnización:** Pago o beneficio otorgado por la aseguradora.

13. **Condiciones generales:** Reglas y definiciones de la póliza.

14. **Endoso:** Modificación a la póliza después de contratarla.

3 Metodología

El conjunto de datos se compone de variables categóricas y numéricas que describen las caracterísctcas 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).

3.1 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. Ds más lento que K-Medias o DBSCAN

2. Más difícil de interpretar

4 Error Cuadrático Medio (MSE)

¿Qué es?
El error cuadrático medio mide entre los valores reales y los valores predichos.

Fórmula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

Interpretación:

- Valores más cercanos a 0 indican mejor desempeño.

- Es sensible a valores atípicos.

Raíz del Error Cuadrático Medior (RMSE)

¿Qué es?
Es la raíz del error cuadrático medio y Permite interpretar el error en las mismas unidades que la variable objetivo.

Fórmula:

$$RMSE = \sqrt{MSE} \tag{2}$$

Interpretación:

- Indica en promedio cuánto se equivoca el modelo.

- Es útil para comparar con magnitudes reales de la variable.

Error Absoluto Medio (MAE)

¿Qué es?
Mide el error absoluto medio entre la predicción y el valor real.

Fórmula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

Interpretación:

- Es más robusto frente a valores atípicos.
- Indica cuánto se equivoca el modelo en promedio.

Coeficiente de Determinación (R^2)

¿Qué es?
Mide qué proporción de la variabilidad de la variable objetivo es explicada por el modelo.

Fórmula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

Interpretación:

- Valores cercanos a 1 indican un excelente ajuste.
- Valores cercanos a 0 indican que el modelo no explica la variabilidad.
- Puede tomar valores negativos si el modelo es peor que predecir el promedio.

5 ¿Qué es Random Forest?

Random Forest es un algoritmo de *Ensemble Learning* basado en la construcción de múltiples árboles de decisión independientes y en la combinación de sus resultados.

Su objetivo es mejorar la precisión, reducir el sobreajuste y aumentar la estabilidad en comparación con un solo árbol de decisión.

Random Forest puede utilizarse tanto para:

- **Clasificación**
- **Regresión**

5.1 ¿Cómo funciona Random Forest?

Random Forest se basa en construir un bosque de árboles de decisión, donde cada árbol se entrena de manera diferente para introducir diversidad y mejorar el desempeño del modelo.

1. **Bootstrap:** Cada árbol se entrena utilizando una muestra aleatoria del conjunto de datos original. Este procedimiento, implica seleccionar observaciones con reemplazo.
2. **Selección aleatoria de variables:** En cada división del árbol, se selecciona un subconjunto aleatorio de características. Esto evita que todos los árboles sean iguales y aumenta la diversidad del bosque.

Una vez entrenados todos los árboles, la predicción se obtiene combinando sus resultados:

- **Clasificación:** se utiliza el *voto mayoritario* entre los árboles.
- **Regresión:** se calcula el *promedio* de las predicciones de todos los árboles.

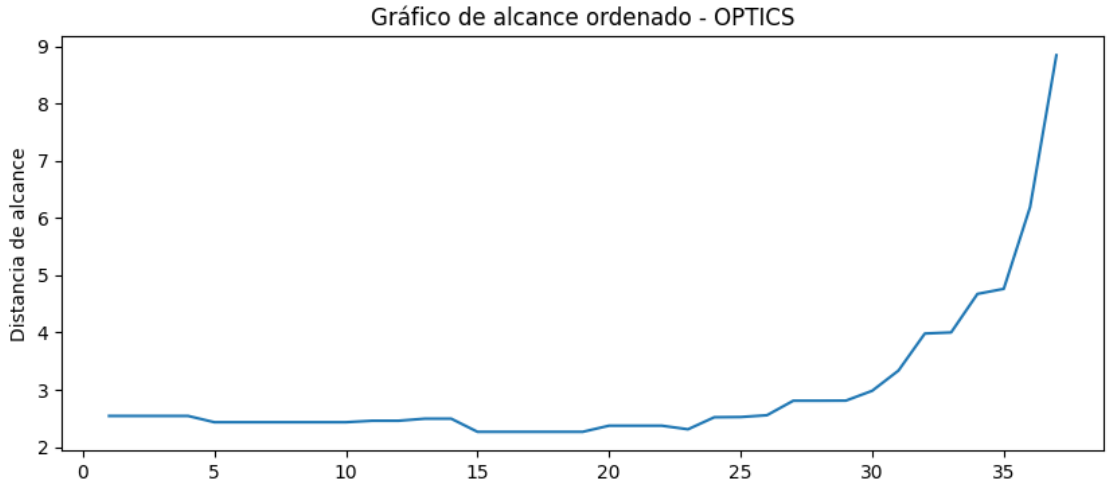

Figure 1: Gráfico de alcance ordenado - OPTICS

6 Resultados

6.1 Gráfica de distancia de alcance

Como se ve en la figura 1 (p. 4).

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.

6.2 Método del codo

Como se ve en la figura 2 (p. 5).

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.

6.3 Visualización t-SNE

Como se ve en la figura 3 (p. 5).

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.

Resultados del Modelo LASSO

El modelo **LASSOCV** seleccionó un valor óptimo de regularización de

$$\alpha = 15.1469$$

Métricas Obtenidas

- **MAE:** 23.6632
- **RMSE:** 42.8857
- **R^2:** 1.0000

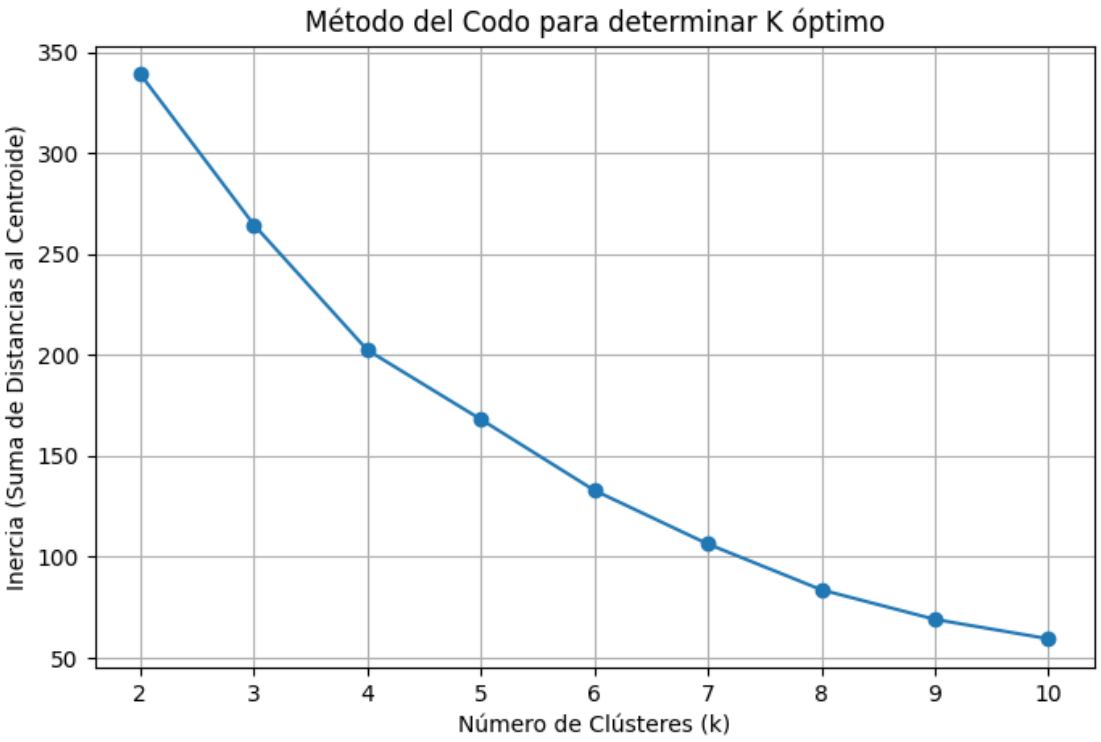

Figure 2: Método del codo para elegir el número de grupos en el algoritmo de K-medias.

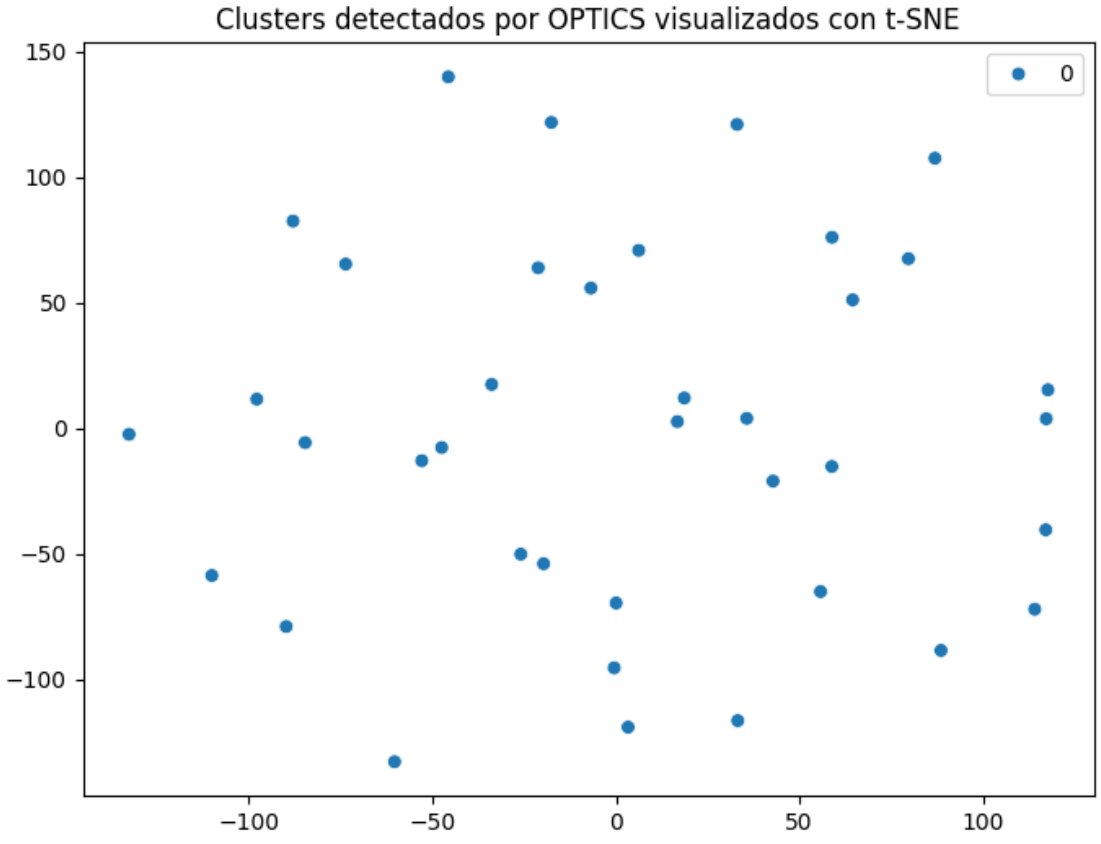

Figure 3: Agrupacion detectadas por Optics visualizados con t-sen.

Principales Coeficientes del Modelo LASSO

Los coeficientes más relevantes seleccionados por el modelo fueron:

Variable	Coeficiente
CNS_RC	6748.569803
AJUSTES	6344.221126
RESERVA	4564.076995
GASTOS	1078.606623
RVA_DISPONIBLE	9.543083
AÑO	0.000000
SUBRAMO	0.000000
INCISO	-0.000000
MODELO	-0.000000
PAGO	0.000000

El modelo asigna un peso únicamente a variables que aportan valor predictivo bajo la regularización L1, las variables con coeficiente igual a cero fueron descartadas por su baja relevancia estadística

6.4 Valores obtenidos del modelo Random Forest

Los valores mostrados son:

- **MAE:** 7024.0615
- **RMSE:** 18761.9417
- **R^2:** 0.5533

7 Conclusión

Como se puede notar, los datos a analizados 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.