Segmentación No Supervisada de Asegurados Basada en Factores de Riesgo para la Optimización del Costo de Seguro Médico

Oscar Jaime Ferreira

6 de noviembre de 2025

Resumen

La gestión eficiente de costos médicos es un desafío central para las aseguradoras de salud, especialmente cuando los asegurados presentan perfiles de riesgo heterogéneos. En este trabajo se analizan características clave relacionadas con el costo del seguro médico, tales como edad, índice de masa corporal (IMC), número de dependientes y condición de fumador. Se aplica el algoritmo OPTICS, un método de clustering basado en densidad, con el objetivo de identificar subgrupos de riesgo sin necesidad de imponer un número fijo de clusters. Los resultados muestran la existencia de seis segmentos principales de asegurados y una proporción significativa de casos atípicos, lo que permite una mejor comprensión de la variabilidad en los niveles de riesgo. Estos hallazgos pueden ser utilizados para mejorar la segmentación de mercado, estimación de primas y políticas de prevención en aseguradoras de salud.

1. Introducción

El incremento constante en los costos de atención médica representa un reto significativo para aseguradoras y sistemas de salud. La identificación precisa del riesgo asociado a los asegurados permite mejorar la tarificación y gestión financiera de las pólizas [1]. La presencia de factores como edad avanzada, obesidad y hábitos nocivos incrementa la probabilidad de reclamaciones médicas más costosas, impulsando la necesidad de modelos analíticos más robustos.

En este contexto, los métodos de aprendizaje no supervisado permiten descubrir estructuras y grupos en los datos sin requerir etiquetas predefinidas, siendo una alternativa clave para la segmentación actuarial [2]. En particular, los algoritmos de clustering basados en densidad facilitan la detección de subpoblaciones con comportamientos de riesgo diferenciados y casos atípicos relevantes para el costo asegurador [3].

El algoritmo OPTICS, propuesto por Ankerst et al. (1999), permite identificar clusters con densidades variables sin fijar a priori la cantidad de grupos, lo que resulta especialmente adecuado en datos de asegurados con heterogeneidad de riesgo [3]. Mejoras recientes han extendido su estabilidad en contextos de alto desbalance, característica frecuente en asegurados fumadores frente a la mayoría no fumadora [4].

Este trabajo aplica OPTICS para segmentar perfiles de asegurados mediante cuatro variables clave: edad, IMC, número de dependientes y condición de fumador. Los resultados buscan apoyar tanto la optimización de costos en aseguradoras de salud como el diseño de estrategias preventivas para poblaciones con mayor riesgo médico [5].

2. Descripción de los datos

Para este trabajo se emplea el *Medical Cost Personal Dataset*, el cual contiene información sobre asegurados en los Estados Unidos y el costo anual asociado a su seguro médico. El conjunto original incluye 1338 observaciones y siete variables, de las cuales se seleccionaron cuatro para el análisis no supervisado, debido a su relevancia actuarial y sanitaria:

- **Edad (age):** años cumplidos del asegurado.
- **Índice de Masa Corporal (bmi):** relación entre peso y estatura, indicador de sobrepeso u obesidad.
- **Número de dependientes (children):** cantidad de hijos incluidos en la póliza.
- **Condición de fumador (smoker_yes):** variable binaria que distingue entre asegurados fumadores (1) y no fumadores (0).

Se eliminaron variables exclusivamente categóricas que no aportaban información sustancial al riesgo médico para este estudio (*sexo* y *región*) y la variable objetivo *charges*, ya que no se emplean etiquetas en el proceso no supervisado.

En la Tabla 1 se presentan estadísticos descriptivos de las cuatro variables seleccionadas:

El conjunto de datos presenta una población mayoritariamente no fumadora, con edades distribuidas principalmente entre los 20 y 60 años, y valores elevados de IMC en promedio, lo

Cuadro 1: Estadísticos descriptivos de las variables analizadas

Variable	**Media**	**Desv. Est.**	**Mínimo**	**Máximo**
Edad (age)	39.21	14.05	18	64
IMC (bmi)	30.66	6.39	16.0	53.1
Hijos (children)	1.09	1.21	0	5
Fumador (smoker_yes)	0.21	0.41	0	1

que sugiere la presencia de sobrepeso en la muestra. Estas características justifican el análisis de segmentación, dado que existen factores de riesgo diferenciados dentro de la población asegurada.

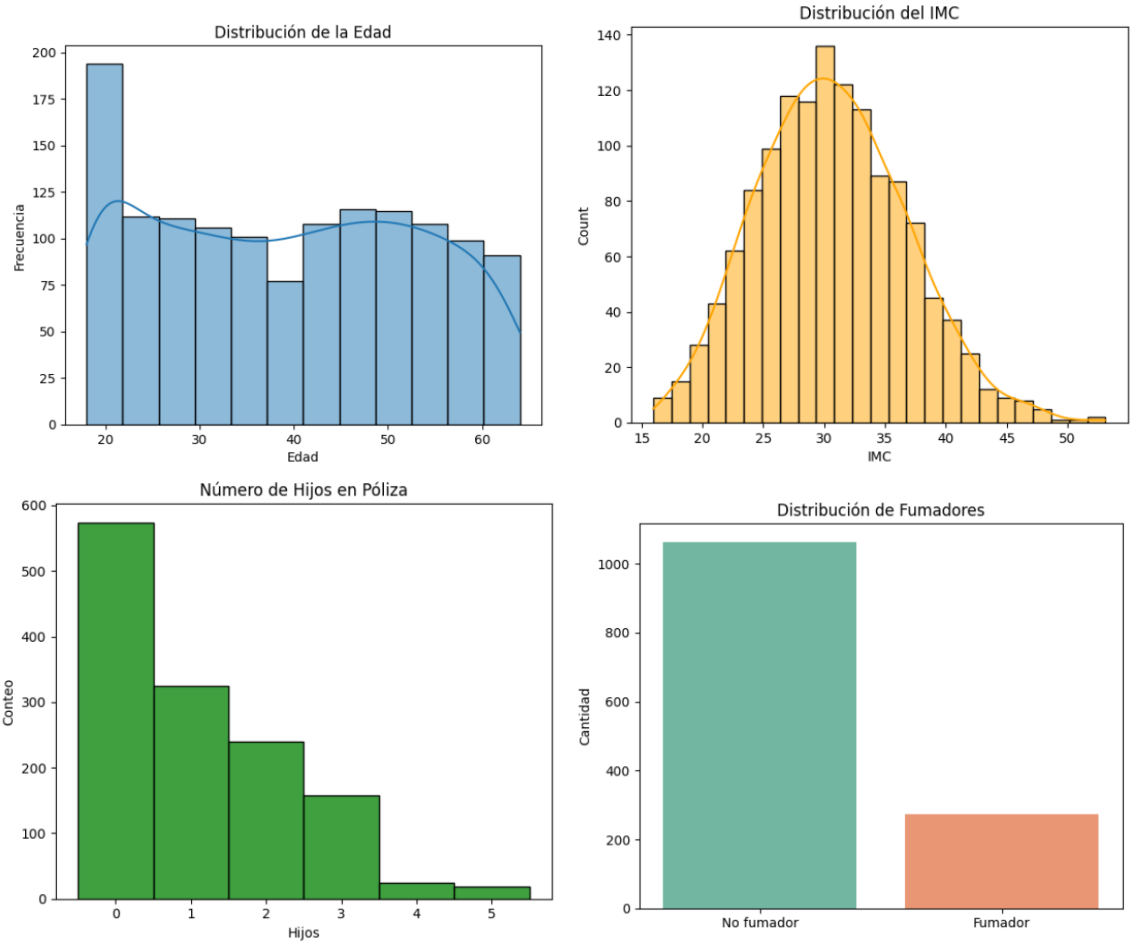


Figura 1: Visualización descriptiva de las variables empleadas en el análisis: edad, IMC, número de dependientes y condición de fumador.

Como se observa en la Figura 1, la edad se concentra entre los 25 y 55 años, lo cual representa una población adulta activa. El IMC presenta un desplazamiento hacia valores altos, indicando una prevalencia notable de sobrepeso u obesidad en la muestra. La mayoría de los asegurados no tienen hijos o tienen uno solo, reflejando pocos dependientes en la póliza. Finalmente, la proporción de fumadores es baja, lo que sugiere una distribución de riesgo

desbalanceada, característica que puede afectar la estructura de los clusters identificados posteriormente.

3. Antecedentes

La segmentación no supervisada se ha aplicado con éxito en analítica de seguros para descubrir perfiles de riesgo cuando coexisten variables numéricas y categóricas. Jamotton (2024) muestra, en un contexto actuarial, adaptaciones de K-means, variantes difusas y clustering espectral para portafolios de pólizas, destacando cómo la segmentación soporta mapas de riesgo y rejillas de tarificación no supervisadas [1].

En el ámbito específico de salud, Momahhed et al. (2023) agrupan a asegurados a partir de millones de reclamaciones de prescripción ambulatoria, usando K-means y validación con coeficiente de silueta, para distinguir clases de riesgo con implicaciones directas en costos [2].

Respecto a métodos basados en densidad, fundamentales para detectar estructuras no convexas y puntos atípicos, el algoritmo OPTICS (Ankerst et al., 1999) ordena los puntos según "alcanzabilidad" y permite revelar jerarquías de densidades sin fijar a priori el número de grupos [3]. Mejoras recientes a OPTICS han sido propuestas para manejar densidades desiguales y facilitar la determinación robusta del número de clusters en datos reales [4]. Además, en aplicaciones médicas, enfoques basados en densidad como DBSCAN y OPTICS se han utilizado para analizar datos clínicos con presencia de atípicos y estructuras complejas, reforzando su idoneidad cuando la heterogeneidad sanitaria es relevante [5].

4. Metodología

El procedimiento metodológico seguido en este trabajo se resume en cuatro etapas principales: preprocesamiento de datos, selección de características, aplicación del algoritmo OPTICS y evaluación de la calidad del agrupamiento mediante métricas de validación interna.

4.1. Preprocesamiento y selección de variables

A partir del conjunto de datos original, se seleccionaron cuatro variables consideradas actuarialmente relevantes: edad, índice de masa corporal (IMC), número de dependientes y condición de fumador. Las variables categóricas se transformaron en formato binario y se aplicó normalización mediante *StandardScaler*, dada la diferencia de escalas entre edad (años) e IMC ($\mathrm{kg/m^2}$).

Sea el vector de características para cada individuo $x_i \in \mathbb{R}^d$, la normalización se define como:

$$x_{ij}^{*} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{1}$$

donde μ_j y σ_j son la media y desviación estándar de la característica j.

4.2. Algoritmo OPTICS

OPTICS (*Ordering Points To Identify the Clustering Structure*) es un método de agrupamiento basado en densidad que no requiere especificar el número de clusters. El algoritmo genera un ordenamiento de los puntos y estima para cada uno una distancia de alcanzabilidad (*reachability distance*) que depende de dos conceptos:

- **Distancia núcleo**:

$$core\_dist(p) = \text{k-dist}(p) \tag{2}$$

  donde k-dist(p) es la distancia al k-ésimo vecino más cercano y refleja la densidad local del punto p.

- **Distancia de alcanzabilidad** entre puntos p y o:

$$reachability(p, o) = \text{máx}(core\_dist(o),\, dist(o, p)) \tag{3}$$

El análisis de la gráfica de alcanzabilidad (*reachability-plot*) permite identificar valles como posibles clusters y valores elevados como puntos atípicos. Esto lo hace particularmente adecuado para datos con densidades desiguales y presencia de grupos minoritarios de alto riesgo.

4.3. Métricas de evaluación del clustering

Se utilizaron tres medidas de validación interna:

1. **Coeficiente Silhouette** (S): mide compacidad y separación entre grupos.

$$S = \frac{b(i) - a(i)}{\text{máx}\{a(i), b(i)\}} \tag{4}$$

   donde $a(i)$ es la distancia media entre i y su cluster, y $b(i)$ la mínima distancia media a otros clusters.

2. **Índice de Davies-Bouldin** (DB): evalúa dispersión intra-cluster vs separación.

$$DB = \frac{1}{k} \sum_{i=1}^{k} \underset{j \neq i}{\text{máx}} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \tag{5}$$

   donde c_i es el centroide del cluster i.

3. **Índice Calinski-Harabasz** (CH): mide la razón entre dispersión inter e intra grupal.

$$CH = \frac{\text{tr}(B_k)/(k-1)}{\text{tr}(W_k)/(n-k)} \tag{6}$$

   donde B_k y W_k son las matrices de dispersión entre y dentro de clusters.

Valores altos de Silhouette y Calinski-Harabasz, junto con valores bajos de Davies-Bouldin, indican una estructura de grupos bien definida.

5. Resultados

El algoritmo OPTICS identificó agrupamientos basados en diferencias de densidad entre asegurados, revelando estructuras útiles para segmentar niveles de riesgo médico. El análisis de la gráfica de alcanzabilidad permite visualizar la estructura jerárquica de los clusters y la presencia de puntos aislados, como se muestra en la Figura 2.

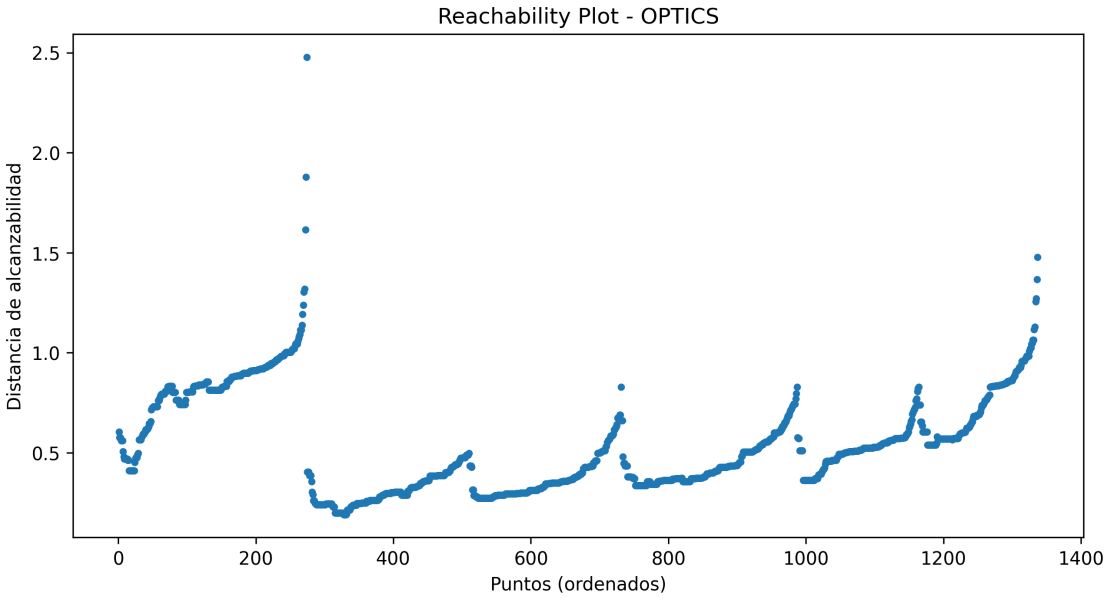
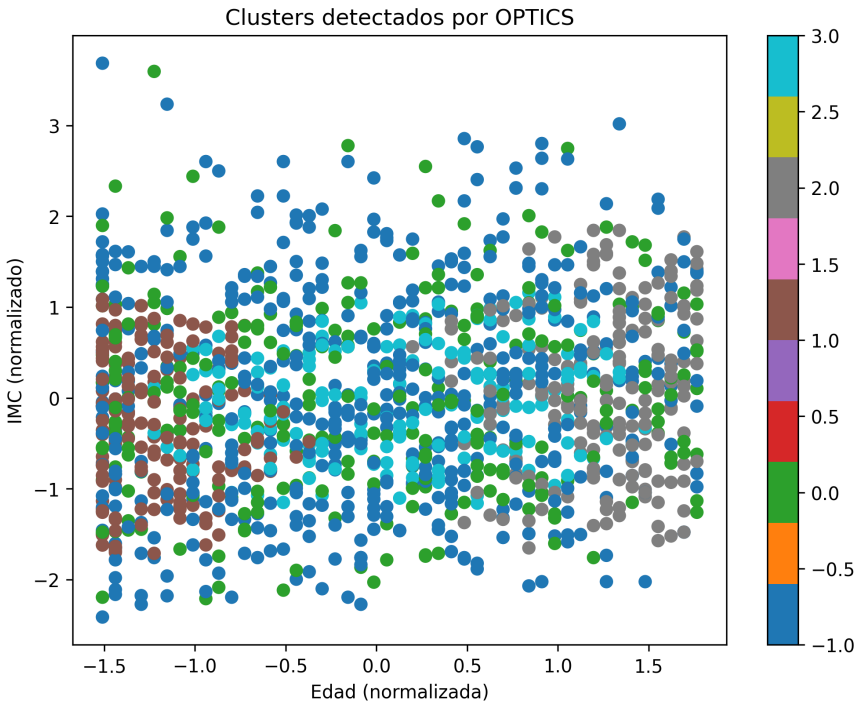


Figura 2: Reachability Plot generado por OPTICS. Los valles indican agrupamientos de alta densidad, mientras que los picos representan puntos aislados o de baja densidad.

La Figura 3 presenta la distribución espacial de los individuos en el espacio Edad–IMC, junto con los clusters detectados y la proporción de fumadores. Se observan dos grupos principales y un conjunto de puntos aislados asociados a baja densidad.



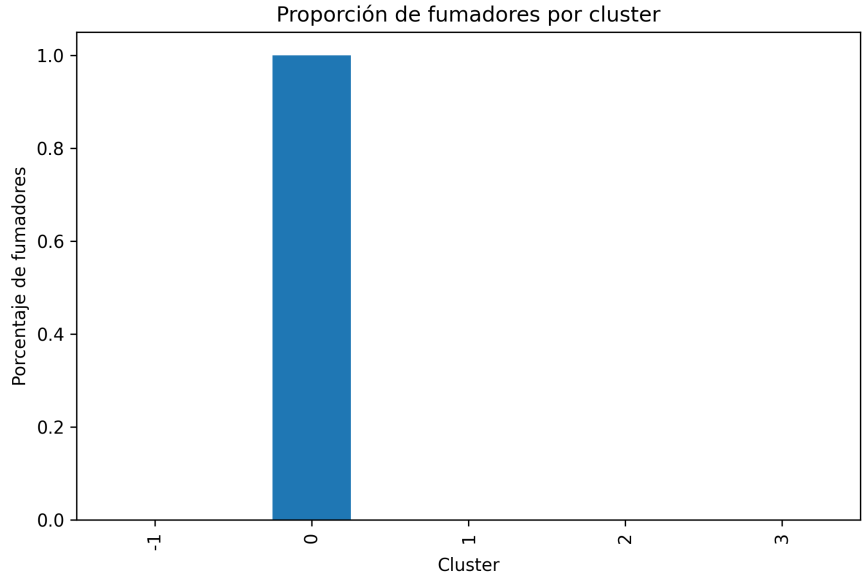


Figura 3: (Arriba) Agrupamientos detectados mediante OPTICS. (Abajo) Proporción de fumadores por cluster, empleada como indicador de riesgo actuarial.

Cuadro 2: Resumen de variables de riesgo por cluster óptimo.

Cluster	Edad Prom.	IMC Prom.	% Fumadores	Riesgo
0	38.51	30.71	100 %	Alto
1	22.58	29.02	0 %	Bajo
2	56.34	31.14	0 %	Bajo
3	41.13	30.33	0 %	Bajo
-1	38.32	31.12	0 %	Bajo

Como se observa en la Tabla 2, el cluster 0 concentra el 100 % de asegurados fumadores junto con un IMC promedio elevado, lo cual indica un perfil de alto riesgo médico-financiero. Por el contrario, los demás clusters representan perfiles de menor siniestralidad esperada al no presentar fumadores ni indicadores antropométricos críticos.

6. Conclusiones y Discusión

El presente trabajo demuestra que el uso de algoritmos de aprendizaje no supervisado es una herramienta valiosa para la segmentación de riesgo en seguros médicos. OPTICS logró identificar agrupamientos diferenciados de asegurados sin utilizar los costos médicos en el proceso de entrenamiento, lo cual permite reconocer segmentos de alto riesgo desde una perspectiva puramente demográfica y de hábitos de salud.

El cluster con mayor proporción de fumadores refleja un perfil de asegurados susceptible a eventos médicos de alto costo, lo que en términos actuariales se traduce en un incremento esperado de reclamaciones y en la necesidad de ajustar las primas o las reservas para dicho grupo. Por otro lado, los clusters dominados por individuos jóvenes, con menor IMC y no fumadores, representan perfiles con un menor riesgo financiero para la aseguradora.

En síntesis, la incorporación de técnicas de minería de datos y aprendizaje automático puede mejorar la toma de decisiones en suscripción, tarificación y diseño de estrategias preventivas, favoreciendo la sostenibilidad del sistema asegurador.

Referencias

[1] C. Jamotton, "Insurance analytics with clustering techniques," *Risks*, vol. 12, no. 9, p. 141, 2024. [Online]. Available: https://doi.org/10.3390/risks12090141

[2] S. S. Momahhed, S. Emamgholipour Sefiddashti, B. Minaei, and Z. Shahali, "K-means clustering of outpatient prescription claims for health insureds in iran," *BMC Public Health*, vol. 23, p. 788, 2023. [Online]. Available: https://doi.org/10.1186/s12889-023-15753-1

[3] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," in *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, 1999, pp. 49–60. [Online]. Available: https://doi.org/10.1145/304182.304187

[4] C. Tang, H. Wang, Z. Wang, X. Zeng, H. Yan, and Y. Xiao, "An improved optics clustering algorithm for discovering clusters with uneven densities," *Intelligent Data Analysis*, vol. 25, no. 6, pp. 1453–1471, 2021. [Online]. Available: https://doi.org/10.3233/IDA-205497

[5] P. Wang, Y. Zhao, and H. Li, "Implementation of real-time medical and health data mining system based on machine learning," *Journal of Healthcare Engineering*, p. 8626197, 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8626197/