Segmentación y Predicción del Costo de Seguro Médico Mediante Aprendizaje No Supervisado y Supervisado

Oscar Jaime Ferreira

11 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 integran enfoques de **aprendizaje no supervisado y supervisado** para analizar los factores que determinan el costo del seguro médico. En una primera etapa se aplica el algoritmo OPTICS, un método de clustering basado en densidad, para identificar subgrupos de riesgo sin necesidad de imponer un número fijo de clusters. Posteriormente, se implementan modelos de regresión lineal y *Random Forest* para predecir el costo del seguro (*charges*) y cuantificar la influencia de variables como edad, IMC y condición de fumador. Los resultados muestran la existencia de seis segmentos principales de asegurados y confirman que la condición de fumador y el sobrepeso son los factores más determinantes del costo. Este enfoque híbrido permite mejorar la segmentación de riesgo y optimizar la estimación de primas en el ámbito asegurador.

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].

En complemento al análisis exploratorio mediante OPTICS, este trabajo incorpora una segunda fase de **aprendizaje supervisado**, en la cual se modela directamente el costo del seguro médico. Mediante regresión lineal múltiple y bosques aleatorios (*Random Forest*), se busca cuantificar el impacto individual de los factores de riesgo detectados en la etapa no supervisada, estableciendo así una relación explícita entre los perfiles de asegurados y su costo financiero esperado. Esta integración de técnicas supervisadas y no supervisadas constituye un enfoque híbrido que permite tanto descubrir segmentos de riesgo como estimar con mayor precisión los costos asociados.

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, que se utilizaron en dos fases: primero para el análisis no supervisado de segmentación de riesgo y posteriormente para el modelado supervisado del costo (*charges*). Las variables principales consideradas por su relevancia actuarial y sanitaria fueron:

- **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).

En la etapa no supervisada se eliminaron variables exclusivamente categóricas (*sexo* y *región*) y la variable objetivo *charges*, ya que no se emplean etiquetas en dicho proceso. Sin embargo, en la segunda fase (modelado supervisado), todas las variables disponibles fueron reincorporadas para predecir el costo del seguro y analizar su influencia sobre los factores de riesgo identificados.

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

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

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 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.

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

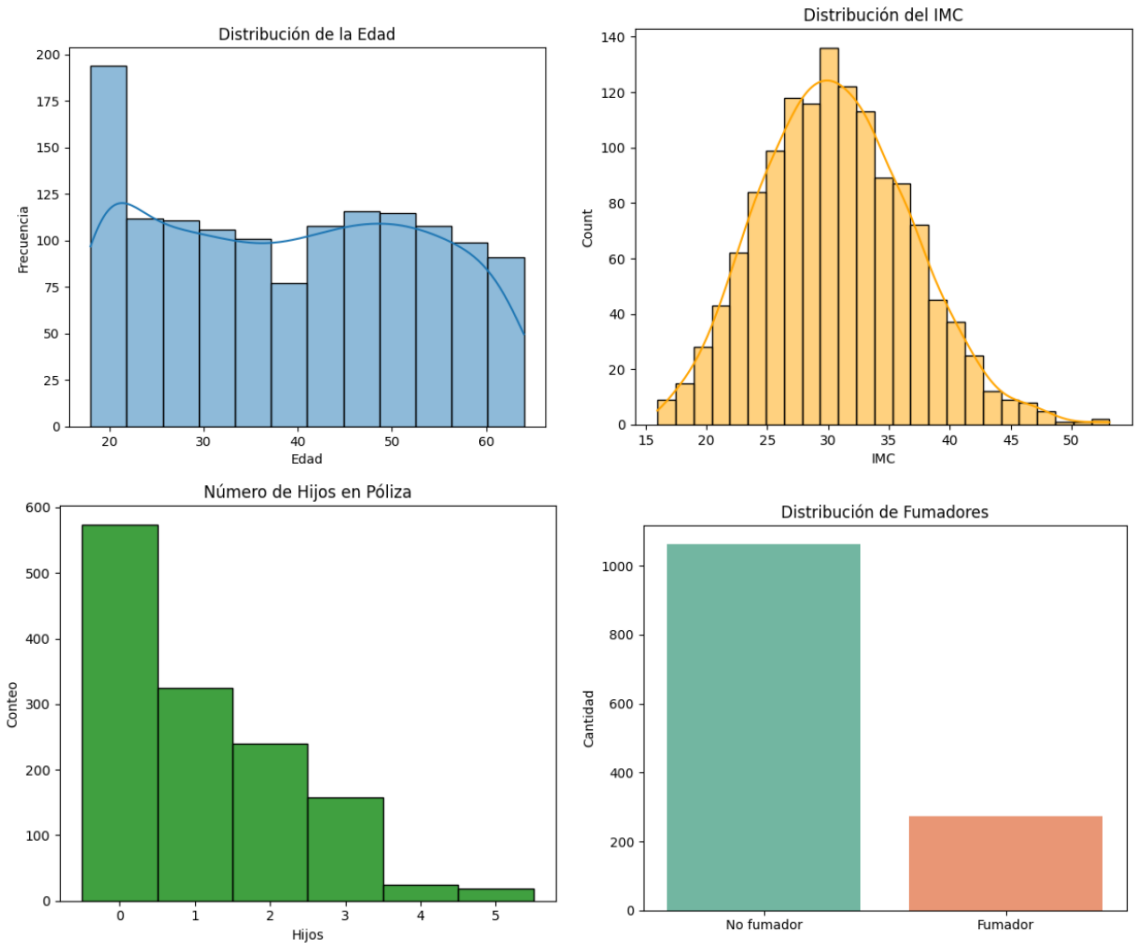

Figura 1: Visualización descriptiva de las variables empleadas en el análisis: edad, IMC, número de dependientes y condición de fumador.

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 enfoque metodológico adoptado en este estudio combina técnicas de aprendizaje no supervisado y supervisado con el propósito de analizar el riesgo médico y predecir el costo del seguro. La Figura 2 muestra de manera esquemática las etapas principales del proceso: preprocesamiento de datos, segmentación mediante el algoritmo OPTICS, modelado supervisado del costo mediante regresión lineal y *Random Forest*, y la integración final de resultados para su interpretación actuarial.

4.1. Diseño general

El proceso analítico se desarrolló en dos fases complementarias:

1. **Análisis No Supervisado:** identificación de grupos homogéneos de asegurados a través del algoritmo OPTICS, con base en variables demográficas y de salud.

2. **Modelado Supervisado:** estimación del costo del seguro (*charges*) mediante modelos de regresión lineal múltiple y bosques aleatorios (*Random Forest*), utilizando las variables seleccionadas en la fase anterior.

Esta estructura híbrida permite, primero, descubrir patrones de riesgo sin etiquetas predefinidas y, posteriormente, cuantificar la relación estadística entre los factores de riesgo y el costo financiero de las pólizas.

4.2. Preprocesamiento de los datos

A partir del *Medical Cost Personal Dataset*, se seleccionaron las variables edad, índice de masa corporal (IMC), número de dependientes y condición de fumador como principales predictores del riesgo médico. Las variables categóricas fueron transformadas mediante codificación binaria y se aplicó normalización *z-score* para garantizar la comparabilidad entre escalas.

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

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

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

4.3. Método no supervisado: OPTICS

El algoritmo OPTICS (*Ordering Points To Identify the Clustering Structure*) propuesto por Ankerst et al. (1999) [3] se basa en la estimación de densidades locales para revelar la estructura jerárquica de los datos sin requerir un número fijo de clusters.

OPTICS ordena los puntos de acuerdo con su distancia de alcanzabilidad (*reachability distance*) y genera un gráfico de alcanzabilidad que permite identificar valles de alta densidad (clusters) y picos de baja densidad (outliers).

Este enfoque resulta particularmente útil en portafolios de asegurados donde coexisten densidades desiguales, como ocurre entre fumadores y no fumadores [4]. Su aplicación en analítica médica ha demostrado eficacia para detectar subgrupos con diferentes perfiles de riesgo y costos potenciales [5].

4.4. Métodos supervisados de predicción

Para cuantificar la influencia de los factores de riesgo sobre el costo del seguro se implementaron dos modelos supervisados de regresión:

- **Regresión Lineal Múltiple:** modelo clásico de inferencia estadística que estima la relación lineal entre la variable objetivo (y) y los predictores (x_j), asumiendo independencia y homocedasticidad de los residuos. Su formalismo teórico se detalla en Hastie et al. (2009) [6].

- **Random Forest Regressor:** algoritmo de aprendizaje de ensamble introducido por Breiman (2001) [7], que construye múltiples árboles de decisión sobre subconjuntos aleatorios de los datos y promedia sus predicciones para reducir la varianza. Este método permite capturar relaciones no lineales e interacciones complejas entre las variables predictoras.

Ambos modelos se entrenaron con una partición del 80 % de los datos y se validaron sobre el 20 % restante. Para evaluar su desempeño se utilizaron las métricas estándar de regresión: Error Absoluto Medio (MAE), Error Cuadrático Medio (MSE), Raíz del Error Cuadrático Medio (RMSE) y Error Porcentual Medio Absoluto (MAPE) [8].

Estudios previos han aplicado estos algoritmos para la predicción de costos médicos y gestión de pólizas, demostrando su efectividad en contextos reales [9, 10].

4.5. Evaluación y comparación

El modelo de regresión lineal se utilizó como referencia interpretativa para determinar el impacto marginal de cada variable sobre el costo. Posteriormente, el modelo de *Random Forest* permitió evaluar las relaciones no lineales y obtener la importancia relativa de las variables predictoras, lo que facilitó la interpretación de los factores más determinantes del costo asegurador.

Los resultados se compararon mediante las métricas antes descritas, seleccionando como modelo final aquel con menor error promedio y mayor capacidad de generalización. De esta forma, la metodología integra tanto la segmentación exploratoria como la predicción cuantitativa del riesgo financiero de los asegurados.

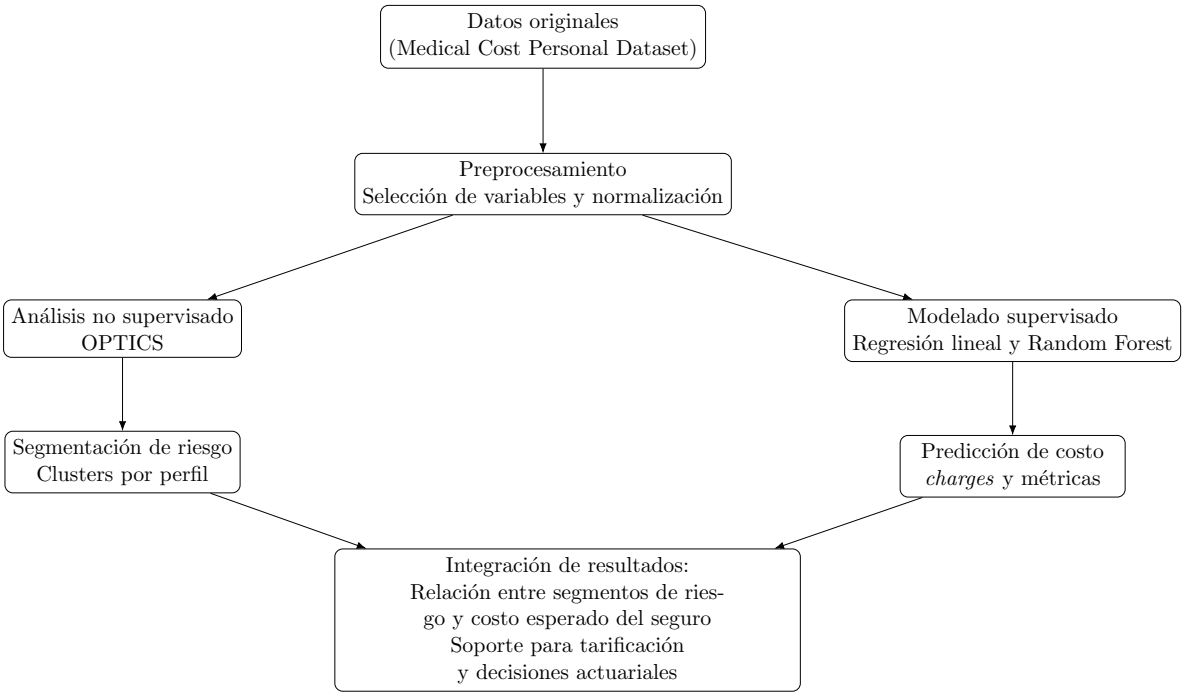

Figura 2: Flujo metodológico del enfoque híbrido: análisis no supervisado con OPTICS y modelado supervisado del costo del seguro.

5. Algoritmo No Supervisado (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**:

$$core\_dist(p) = \text{k-dist}(p) \tag{1}$$

  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:

$$reachability(p, o) = \text{máx}(core\_dist(o),\ dist(o, p)) \tag{2}$$

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.

5.1. Métricas de evaluación del clustering

Se utilizaron tres medidas de validación interna:

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

$$S = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

   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.

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \tag{4}$$

   donde c_i es el centroide del cluster i.

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

$$CH = \frac{\mathrm{tr}(B_k)/(k-1)}{\mathrm{tr}(W_k)/(n-k)} \tag{5}$$

   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.

6. Resultados del Análisis No Supervisado (OPTICS)

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 3.

La Figura 4 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.

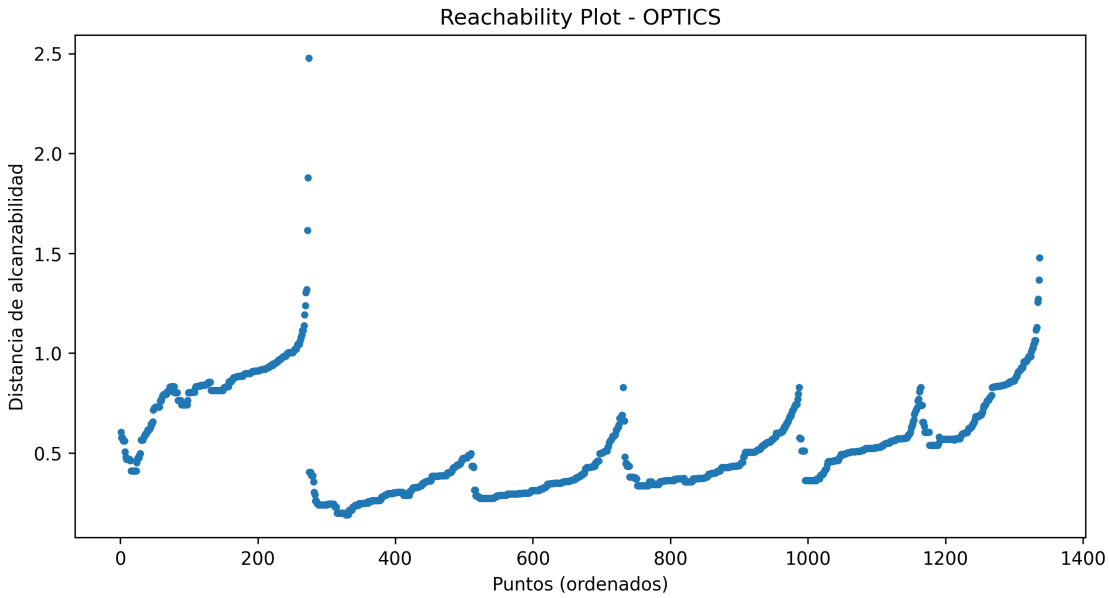

Figura 3: Reachability Plot generado por OPTICS. Los valles indican agrupamientos de alta densidad, mientras que los picos representan puntos aislados o de baja densidad.

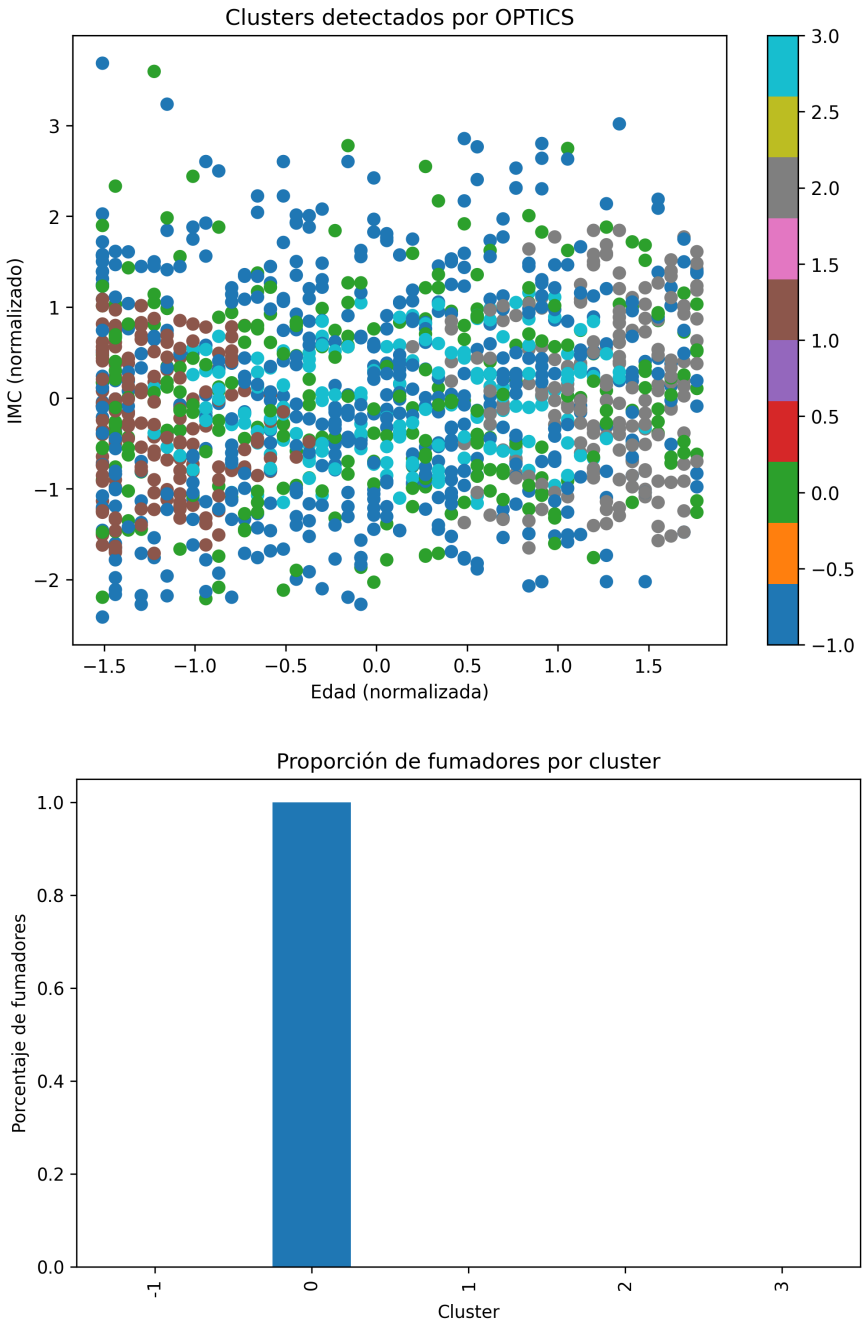

Figura 4: (Arriba) Agrupamientos detectados mediante OPTICS. (Abajo) Proporción de fumadores por cluster, empleada como indicador de riesgo actuarial.

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.

7. Modelado Supervisado del Costo del Seguro

A fin de complementar la segmentación no supervisada, se aplicaron modelos de **aprendizaje supervisado** para predecir el costo anual del seguro médico (*charges*) a partir de las variables *edad*, *índice de masa corporal (BMI)*, *número de dependientes*, *sexo*, *condición de fumador* y *región*. El objetivo fue evaluar de forma cuantitativa la influencia de los factores de riesgo identificados previamente sobre el costo financiero de las pólizas.

El modelo de *Random Forest*, propuesto por Breiman (2001) [7], fue elegido por su capacidad para capturar interacciones no lineales y reducir la varianza del estimador. Por su parte, la regresión lineal múltiple se empleó como modelo base clásico y de referencia interpretativa [6]. Las métricas de error empleadas (MAE, MSE, RMSE, MAPE) son ampliamente utilizadas en regresión para evaluar precisión y robustez [8]. Diversos estudios recientes aplican algoritmos supervisados a la predicción de costos médicos y pólizas de salud, con resultados prometedores [9, 10].

7.1. Modelos empleados

Se consideraron dos algoritmos supervisados de regresión:

1. **Regresión Lineal Múltiple**: modelo base interpretativo que asume una relación lineal entre las características de los asegurados y el costo del seguro, de la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

   donde y representa el costo del seguro y x_j las variables predictoras.

2. **Random Forest Regressor**: ensamble de múltiples árboles de decisión que permite capturar relaciones no lineales y efectos de interacción entre las variables predictoras. Este modelo promedia las predicciones de cientos de árboles entrenados sobre diferentes subconjuntos de datos, reduciendo la varianza y mejorando la capacidad de generalización.

7.2. Métricas de evaluación

El desempeño de los modelos se evaluó mediante las métricas más comunes en regresión:

- **Error Absoluto Medio (MAE)**:

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

  mide el error promedio en unidades monetarias.

- **Error Cuadrático Medio (MSE)**:

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

  penaliza más los errores grandes y explica los valores elevados observados (del orden de 10^7), ya que el costo del seguro se mide en miles de unidades monetarias.

- **Raíz del Error Cuadrático Medio (RMSE)**:

$$RMSE = \sqrt{MSE}$$

  devuelve el error promedio en las mismas unidades que el costo real.

- **Error Porcentual Medio Absoluto (MAPE)**:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

  expresa el error relativo porcentual promedio de las predicciones.

7.3. Resultados del modelado supervisado

La Tabla 3 muestra las métricas obtenidas para ambos modelos de predicción:

Cuadro 3: Desempeño de los modelos supervisados en la predicción del costo del seguro

Modelo	**MAE**	**MSE**	**RMSE**	**MAPE (%)**
Regresión Lineal	4,181.19	33,596,915.85	5,796.28	46.89
Random Forest	2,559.54	21,166,502.28	4,600.71	32.05

El modelo de **Random Forest** superó claramente a la regresión lineal en todas las métricas: redujo el error absoluto promedio en 39 %, el RMSE en 21 % y el error porcentual en 32 %. Esto confirma que las relaciones entre edad, IMC y hábito de fumar con el costo del seguro no son puramente lineales, sino que presentan interacciones complejas que el bosque aleatorio captura de forma más eficiente.

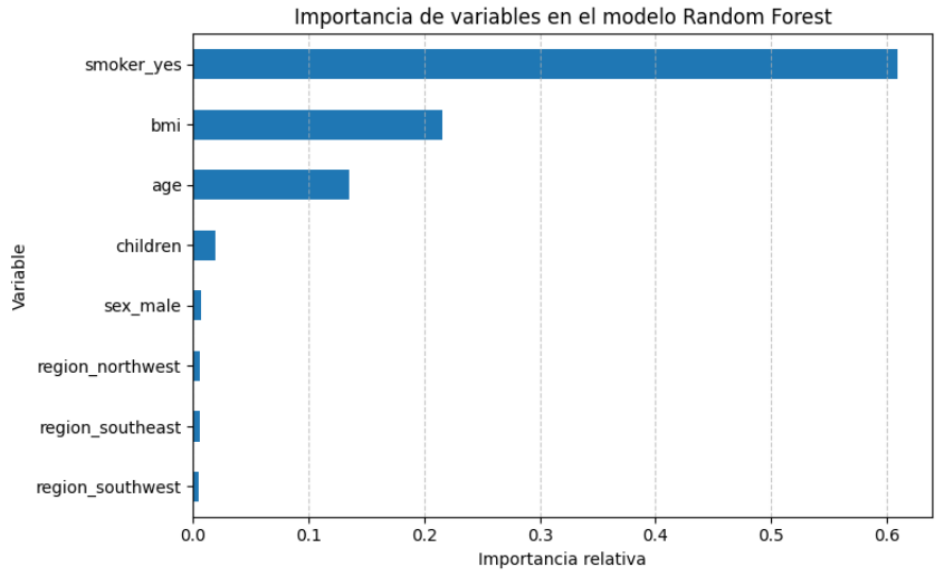

Figura 5: Importancia relativa de las variables en el modelo Random Forest.

7.4. Importancia de las variables

La Figura 5 muestra las variables más influyentes en la predicción del costo del seguro:

El modelo identificó a *smoker_yes* (condición de fumador) como la variable más determinante, aportando aproximadamente el 61 % de la capacidad predictiva total, seguida por *bmi* (21.5 %) y *age* (13.5 %). Las demás variables, como *children*, *sex* y *region*, presentaron impactos marginales en el costo proyectado.

Estos resultados refuerzan la conclusión de que los hábitos de salud y el sobrepeso tienen mayor peso financiero que factores demográficos o familiares.

8. Conclusiones y Discusión

El presente trabajo demuestra que la combinación de técnicas no supervisadas y supervisadas proporciona una comprensión integral del riesgo médico-financiero en portafolios de asegurados.

En la primera etapa, el algoritmo OPTICS permitió descubrir subgrupos con perfiles de riesgo diferenciados, identificando un segmento claramente asociado a fumadores con IMC elevado. En la segunda etapa, los modelos supervisados validaron cuantitativamente la influencia de esos factores sobre el costo real del seguro.

El **Random Forest Regressor** mostró el mejor desempeño predictivo, con errores promedio del orden de 2,500 a 4,600 unidades monetarias, y confirmó que la condición de fumador y el índice de masa corporal son las variables que más elevan el costo de las pólizas. Esto coincide con la segmentación previa y refuerza la coherencia del enfoque híbrido.

Desde una perspectiva actuarial, estos hallazgos respaldan la utilidad de modelos analíticos híbridos que combinen la exploración no supervisada con la predicción supervisada, ya que permiten detectar segmentos de alto riesgo y, al mismo tiempo, estimar el impacto

financiero esperado de cada factor.

Futuras líneas de investigación pueden incorporar modelos explicativos basados en interpretabilidad (SHAP o LIME) y simulaciones con datos sintéticos para ajustar tarifas personalizadas y políticas de prevención más precisas.

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/

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009. [Online]. Available: https://doi.org/10.1007/978-0-387-84858-7

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[8] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, 2021. [Online]. Available: https://doi.org/10.7717/peerj-cs.623

[9] V.-C. Dinh and T.-L. Nguyen, "Predicting health insurance costs using machine learning algorithms," *Journal of Risk and Financial Management*, vol. 12, no. 4, p. 189, 2019. [Online]. Available: https://doi.org/10.3390/jrfm12040189

[10] Y. Chen and Y. Hao, "Supervised learning approaches for health insurance claim prediction," *Procedia Computer Science*, vol. 138, pp. 436–443, 2018. [Online]. Available: https://doi.org/10.1016/j.procs.2018.10.062