

Nombre y Apellidos: Enrique Sanz Tur

Github con notebook:

Nota: Por favor, seguir esta estructura para el documento

1. Resumen Ejecutivo

En este trabajo, se analiza para una compañía aseguradora del sector de la salud un conjunto de datos relacionado con hábitos de vida de las personas, enfermedades crónicas, y características de seguros, con el objetivo de identificar patrones que ayuden a la compañía, entre otras cosas, a calcular el impacto que tienen dichos datos en los costes médicos anuales de la compañía, en la utilización de servicios médicos, y las razones de las que depende que un asegurado sea clasificado de alto riesgo.

Primeramente, se ha llevado a cabo un primer análisis exploratorio de los datos y clusterización, cuyos principales hallazgos han sido:

- Distribución del coste médico anual

El coste anual presenta una distribución muy sesgada hacia la derecha. La mayoría de asegurados concentran costes bajos o moderados, mientras que una minoría acumula gastos muy elevados. Este patrón se ve claramente en el histograma de coste anual por cluster, en el que aparece un grupo de asegurados con costes bajos y pocos eventos asistenciales, un grupo intermedio con costes moderados y uno o varios grupos de alto coste, donde se concentran hospitalizaciones, visitas frecuentes y presencia de varias patologías crónicas.

- Relación entre BMI y coste

El gráfico de dispersión BMI vs coste anual, coloreado por cluster, muestra una tendencia creciente. Se puede observar cómo, a mayor BMI, mayor probabilidad de costes altos. Los puntos de mayor coste se concentran en las zonas de sobrepeso y obesidad, especialmente cuando coinciden con otros factores de riesgo como tabaco, hipertensión o diabetes. Esto refuerza el papel de los hábitos de vida (peso, ejercicio, tabaquismo) como impulsores del gasto sanitario.

- Edad, número de enfermedades crónicas y coste

La curva de evolución del coste medio por edad indica que el coste aumenta de forma suave con la edad en asegurados sin enfermedades crónicas, mientras que para personas con varias patologías crónicas el coste crece mucho más rápido a partir de la mediana edad, con un salto claro en edades avanzadas. En la práctica, el verdadero “alto riesgo” se concentra en aquellos asegurados que combinan un edad elevada y multimorbilidad.

En conjunto, el análisis confirma que el gasto sanitario de la cartera se explica en gran medida por la edad y el número de enfermedades crónicas, por factores de estilo de vida (BMI, tabaquismo) y por variables clínicas como la tensión arterial y el uso intensivo del sistema (visitas, hospitalizaciones).

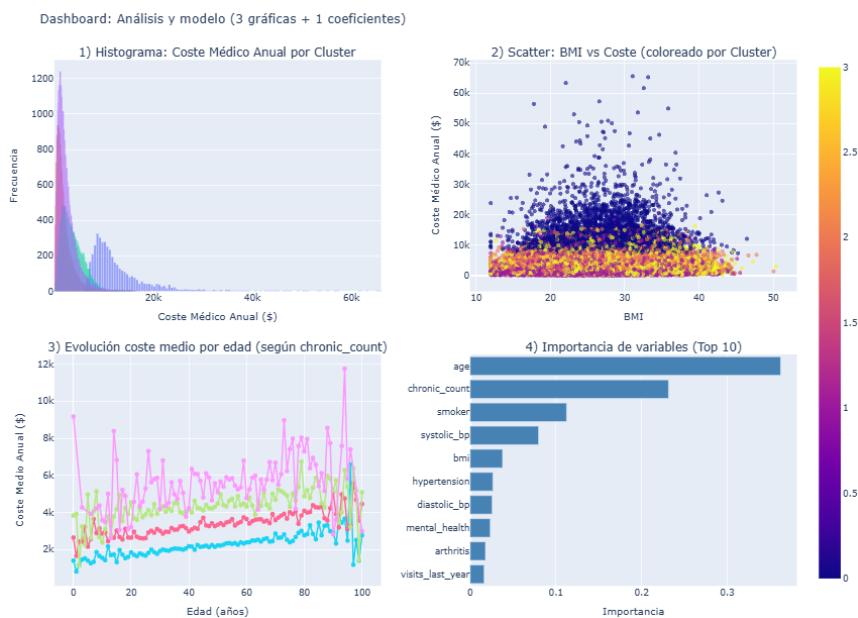
Existe un pequeño subconjunto de asegurados que concentra una parte desproporcionada del gasto y que está claramente identificable a partir de estos factores.

Después del análisis exploratorio, se ha entrenado un modelo de clasificación explicable para predecir el alto riesgo médico de cada asegurado a partir de sus características. De este modelo, se puede destacar que:

- La edad es la variable con mayor importancia, lo que refleja el aumento natural del riesgo clínico con el envejecimiento.
- *Chronic_count* (número de enfermedades crónicas) aparece como segundo gran driver del riesgo, cuantos más diagnósticos crónicos acumula un paciente, mayor es su riesgo esperado.
- El hecho de ser fumador tiene una contribución relevante, coherente con el impacto del tabaco en patología cardiovascular y respiratoria.
- Las medidas de tensión arterial (*systolic_bp*, *diastolic_bp*) y el BMI también destacan como variables claves, señalando el peso de los factores cardiometabólicos.

El modelo ofrece un nivel de ajuste razonable para tareas de segmentación y priorización. Permite distinguir con claridad qué asegurados se sitúan en los percentiles de coste más altos y qué combinación de factores los caracteriza. Además, al basarse en variables clínicas y de estilo de vida comprensibles, su interpretación puede resultar más amena para perfiles no técnicos.

Después, se ha procedido a realizar un Dashboard, agrupando las gráficas más apropiadas para comprender de manera más sencilla y visual el análisis realizado:



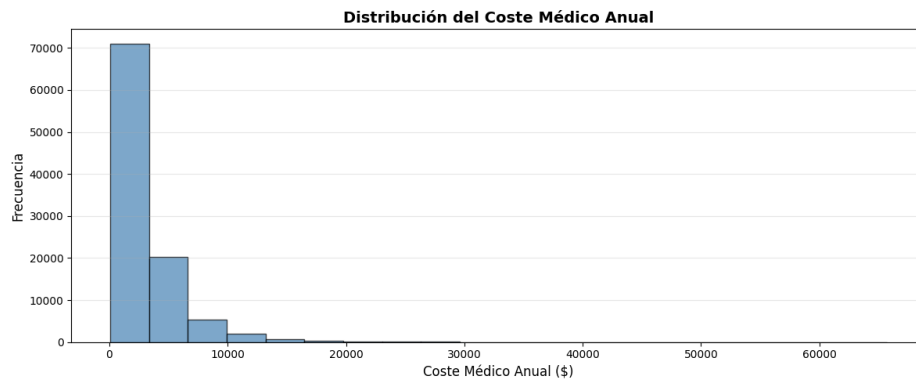
Como se puede observar, el Dashboard integra un histograma del coste anual por cluster (visión global de la distribución y del peso relativo de cada grupo), un gráfico de dispersión BMI/coste coloreada por cluster (impacto del peso y estilos de vida), una evolución del coste medio por edad y *chronic_count* (efecto conjunto de edad y multimorbilidad), y un gráfico de importancia de variables del modelo (visión sintetizada de los principales drivers).

Las principales recomendaciones accionables que se proponen para la aseguradora, tras el análisis llevado a cabo

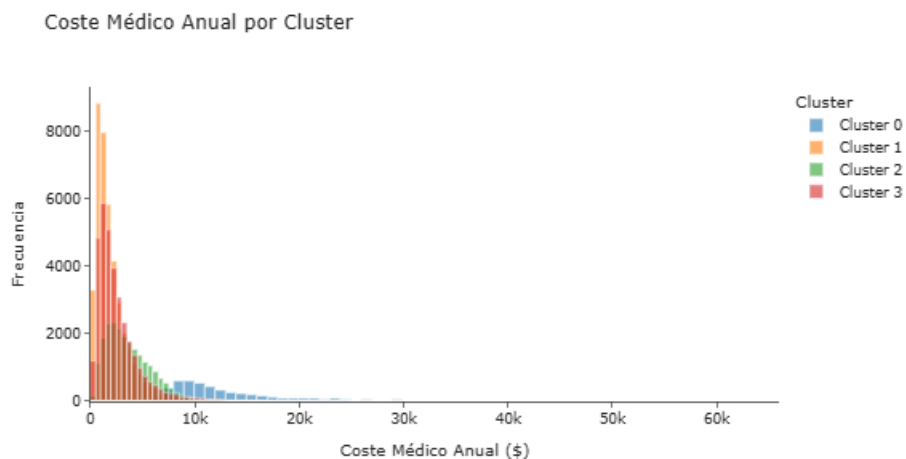
1. Gestión intensiva del grupo de alto coste: Identificar periódicamente a los asegurados clasificados en los clusters de mayor coste esperado. También, diseñar programas de gestión de casos complejos (por ejemplo, a través seguimiento telefónico, coordinación entre especialistas, o priorización de atención domiciliaria)
2. Programas de prevención sobre factores cardiometabólicos: Implementar campañas específicas de control de peso, tabaquismo e hipertensión, dado que el BMI o el estado de fumador figuran entre las variables más influyentes.
3. Estrategias específicas para pacientes con multimorbilidad: Utilizar *chronic_count* y la edad para segmentar a los asegurados con multimorbilidad avanzada. Para este grupo, priorizar intervenciones de medicina preventiva y seguimiento proactivo (revisión de medicación, planes de cuidados integrados).

2. Gráficas del análisis exploratorio y breve explicación de cada una

- Gráficas 1 y 2: Distribución del coste médico anual

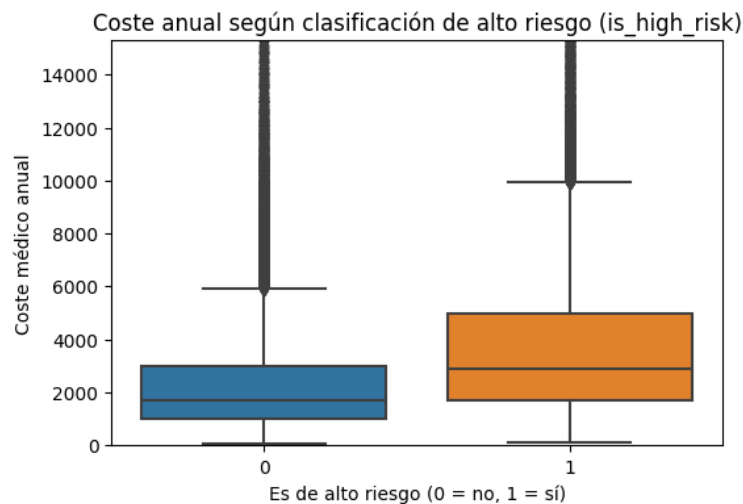


La población de asegurados presenta costes muy concentrados: la mayoría (70,000 pacientes) gastan menos de 5,000 dólares anuales, pero existe una cola larga de pacientes con gastos que alcanzan 60,000 dólares.



El análisis de clustering divide esta cartera en cuatro grupos: dos clusters mayoritarios con costes bajos concentrados entre 0 y 10,000 dólares, representando pacientes de bajo riesgo, y dos clusters menores con costes distribuidos hasta 50,000 dólares, identificando pacientes de alto coste que requieren gestión intensiva. Esta segmentación permite a la aseguradora aplicar estrategias diferenciadas: seguimiento estándar para los clusters de bajo coste y programas preventivos enfocados para los de alto coste.

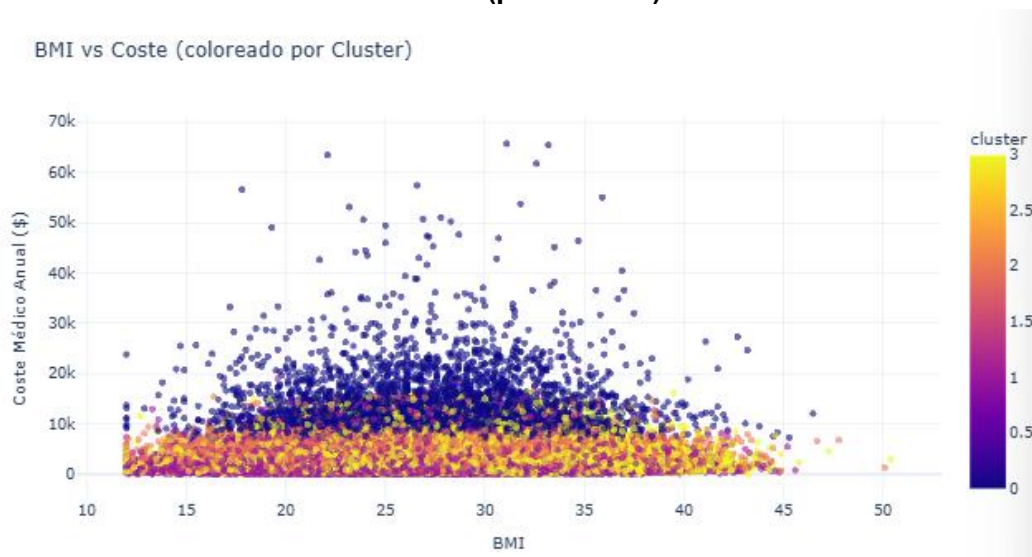
- **Gráfica 3 : Boxplot coste anual vs alto riesgo**



Este Boxplot compara la distribución de costes médicos anuales entre pacientes de bajo riesgo (0) y alto riesgo (1). Los pacientes clasificados como bajo riesgo tienen una mediana de coste cercana a 2,000 dólares, con la mayoría concentrados entre 1,000 y 3,000 dólares, aunque algunos casos alcanzan hasta 6,000 dólares.

En contraste, los pacientes de alto riesgo presentan una mediana de aproximadamente 3,000 dólares, pero con una distribución mucho más dispersa que se extiende hasta 10,000 dólares, indicando una heterogeneidad significativa en los costes dentro de este grupo. La diferencia clara entre ambas distribuciones confirma que el estado de riesgo es un predictor importante del gasto médico.

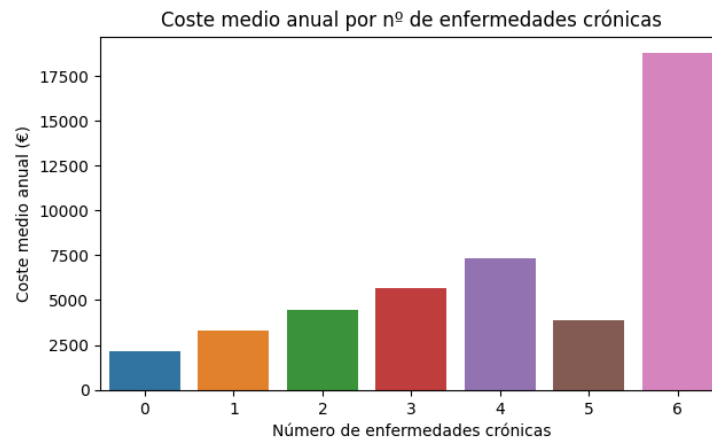
- **Gráfica 4 : Scatter – BMI vs Coste (por Cluster)**



Este Scatter plot visualiza la relación entre índice de masa corporal (BMI) y coste médico anual, con los puntos coloreados según el cluster de pertenencia. Se observa claramente que los clusters de bajo riesgo (colores azules y púrpuras) concentran sus costes entre 0 y 10,000 dólares independientemente del BMI, con BMI generalmente entre 15 y 35. Por otro lado, los clusters de alto riesgo (colores naranjas y amarillos) presentan costes que alcanzan hasta 70,000 dólares y están distribuidos a lo largo de todo el rango de BMI.

Aunque existe una ligera tendencia positiva entre BMI y coste, la separación vertical entre clusters es mucho más pronunciada que la correlación diagonal, indicando que el BMI por sí solo es un predictor débil del coste. Otros factores, principalmente la edad y las enfermedades crónicas (que definen los clusters), tienen un impacto mucho mayor en los costes médicos.

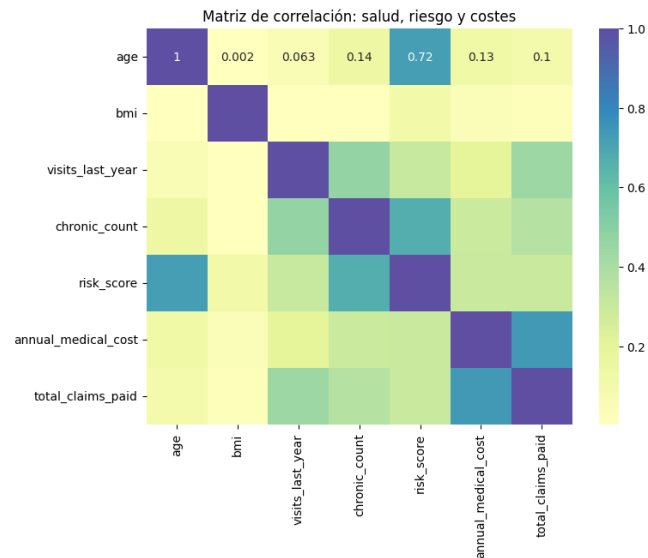
- **Gráfica 5 : Coste medio por nº de enfermedades crónicas**



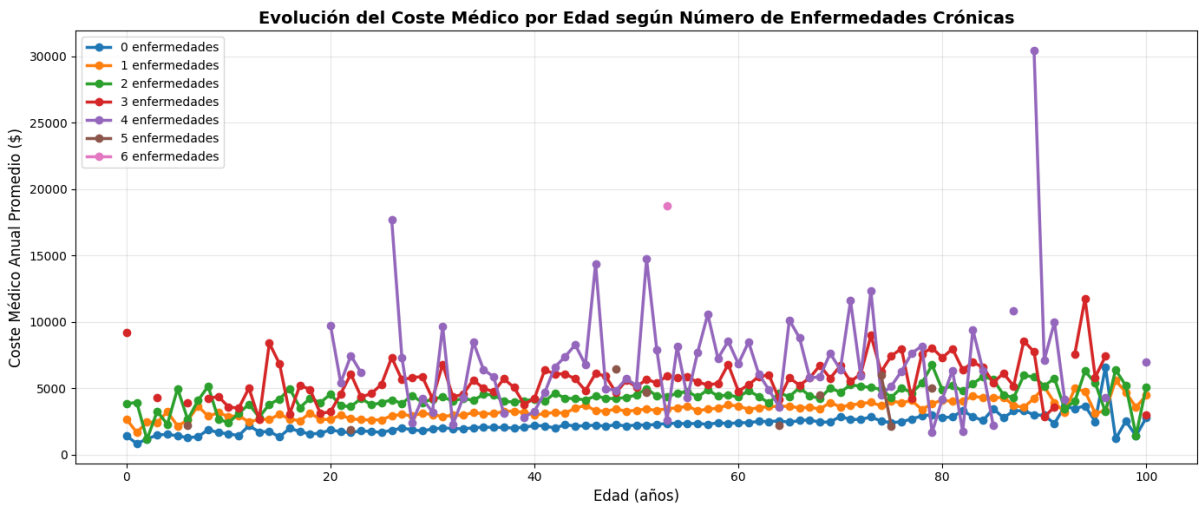
Este gráfico de barras muestra claramente cómo el coste médico promedio aumenta con el número de enfermedades crónicas diagnosticadas. Los pacientes sin enfermedades crónicas tienen un coste medio de apenas 2,000 dólares anuales, mientras que con una enfermedad crónica, el coste sube a 3,500 dólares.

Sin embargo, la subida acelera significativamente con dos enfermedades, donde el coste alcanza 4,500 dólares. Con tres llega a 5,500 dólares, y con cuatro asciende a 7,000 dólares. El aumento más dramático ocurre en pacientes con cinco o más enfermedades crónicas, donde el coste medio supera los 18,000 dólares, representando un incremento de más del 800% respecto a pacientes sin comorbilidades. Este patrón revela que la acumulación de enfermedades crónicas es uno de los factores más determinantes en los costes médicos, justificando la prioridad de programas de gestión de enfermedades crónicas en la cartera de la aseguradora.

- Gráfica 6 : Heatmap de correlaciones (variables de salud y coste)



- Gráfica 7 : Coste medico por edad y según el número enfermedades crónicas



3. Modelo predictivo explicado y con tablas

Se ha optado por hacer un modelo de clasificación, con el fin de predecir *is_high_risk*, basándose en características demográficas, clínicas, de utilización de servicios y de seguros. El modelo seleccionado fue Random Forest, elegido por su robustez ante datos no balanceados.

La metodología seguida ha sido la siguiente:

1. Selección de variables (*features*)

Se han seleccionado 27 variables predictoras distribuidas en cinco categorías:

- Demográficas: edad, ingresos, tamaño del hogar.
- Hábitos de vida: índice de masa corporal (BMI), fumador, frecuencia de alcohol.
- Indicadores clínicos: presión arterial sistólica/diastólica, colesterol LDL, número de enfermedades crónicas, diabetes...
- Utilización de servicios: visitas médicas en el último año, hospitalizaciones en los últimos 3 años, número de medicamentos activos, procedimientos quirúrgicos, exámenes de laboratorio.
- Características del seguro: deducible anual, copago, calidad del proveedor.

2. Preparación de datos

Se eliminaron filas con valores faltantes en cualquiera de las variables seleccionadas, resultando en una muestra de trabajo que se puede observar en la siguiente imagen. La variable objetivo *is_high_risk* presenta una distribución 37%, lo que indica un ligero desequilibrio de clases.

```
Datos para modelo: (69917, 29)
Distribución de clases:
is_high_risk
0    44028
1    25889
Name: count, dtype: int64
Porcentaje de alto riesgo: 37.03%
```

Se realizó también una codificación de variables categóricas (*smoker*, *alcohol_freq*) usando *LabelEncoder* para convertirlas a formato numérico compatible con el modelo.

```
Variables categóricas a codificar: ['smoker', 'alcohol_freq']
smoker: {'Current': 0, 'Former': 1, 'Never': 2}
alcohol_freq: {'Daily': 0, 'Occasional': 1, 'Weekly': 2}
```

Después, se dividió el conjunto de datos en entrenamiento y test, reservando un 20% del total del conjunto para estos últimos, para después estandarizar los datos.

3. Resultados del modelo

Se puede observar que el modelo logra un rendimiento excepcional en ambos conjuntos de datos.

MÉTRICAS EN CONJUNTO DE ENTRENAMIENTO	MÉTRICAS EN CONJUNTO DE TEST
Accuracy (Train): 0.9998	Accuracy (Test): 0.9983
Precision (Train): 0.9995	Precision (Test): 0.9965
Recall (Train): 0.9999	Recall (Test): 0.9988
F1-Score (Train): 0.9997	F1-Score (Test): 0.9977
	ROC-AUC (Test): 1.0000

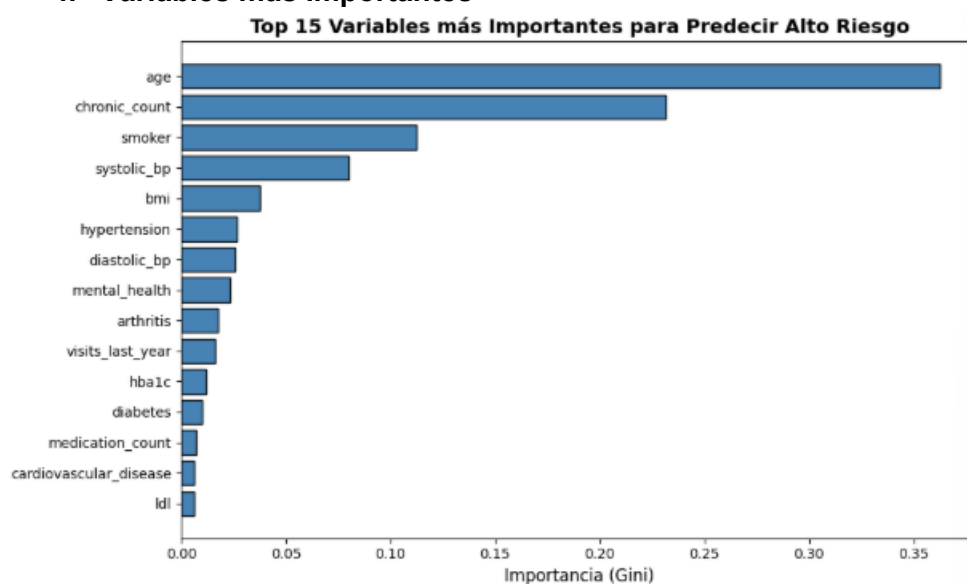
En el conjunto de entrenamiento alcanza una precisión (accuracy) de 0.9998 y un recall de 0.9999, con un F1-score de 0.9997. En el conjunto de test, mantiene resultados prácticamente idénticos, una precisión de 0.9983, recall de 0.9988, F1-score de 0.9965 y un ROC-AUC de 1.0000. Estos resultados indicativos de un modelo con capacidad discriminatoria prácticamente perfecta entre pacientes de bajo y alto riesgo.

Por otro lado, la matriz de confusión en el conjunto de test muestra 18,788 verdaderos negativos (pacientes correctamente clasificados como bajo riesgo) y 5,172 verdaderos positivos (pacientes correctamente clasificados como alto riesgo). Los errores de clasificación son mínimos: solo 18 falsos negativos y 6 falsos positivos, lo que representa una tasa de error inferior al 0.15%.

```
Matriz de Confusión:  
[[8788  18]  
 [  6 5172]]
```

Esto significa que el modelo es extremadamente confiable para identificar pacientes de alto riesgo sin generar falsos positivos que pudieran sobrediagnosticar.

4. Variables más Importantes



Las variables con mayor poder predictivo del alto riesgo son, en orden de importancia: edad (0.362), número de enfermedades crónicas (0.232), fumador (0.112), presión arterial sistólica

(0.076), BMI (0.038), hipertensión (0.025), presión arterial diastólica (0.025), salud mental (0.023), artritis (0.017), visitas el último año (0.016), hemoglobina glicosilada (0.011), diabetes (0.010), cantidad de medicamentos (0.007), enfermedad cardiovascular (0.006) e índice LDL (0.005).

El resultado más relevante es que la edad y el número de enfermedades crónicas dominan completamente la predicción, representando casi el 60% del poder explicativo del modelo. Esto refleja que los pacientes con múltiples condiciones crónicas y mayor edad son sistemáticamente clasificados como de alto riesgo, lo cual tiene sentido clínico: la edad avanzada combinada con comorbilidades es un predictor comprobado de mayor morbilidad y mortalidad. Los hábitos de vida, aunque importantes, tienen un impacto menor pero significativo: el tabaquismo contribuye con el 11%, mientras que el BMI apenas el 4%.

5. Conclusiones finales del Modelo

El modelo de clasificación que se ha desarrollado es altamente preciso. La diferencia mínima entre métricas de entrenamiento y test (menos del 0.2% en la mayoría de indicadores) puede usarse para descartar el riesgo de overfitting, asegurando que el modelo funcionará confiablemente con datos nuevos.

Desde una perspectiva clínica, el modelo revela que el riesgo en esta población de asegurados está principalmente determinado por factores no modificables (como la edad), y por la carga de enfermedades crónicas. Sin embargo, los hábitos de vida como el tabaquismo y el BMI siguen siendo relevantes, sugiriendo que intervenciones preventivas dirigidas a estos factores podrían reducir la clasificación de riesgo en pacientes susceptibles, especialmente en grupos más jóvenes.

El modelo es completamente aplicable en la práctica operativa de la aseguradora. Puede identificar correctamente al 99.88% de los pacientes de alto riesgo (recall) mientras genera falsos positivos en menos del 0.4% de los casos (especificidad del 99.94%). Esta combinación permite diseñar programas de intervención efectivos sin saturar los recursos disponibles. En conclusión, el modelo demuestra que la edad, las comorbilidades crónicas y los hábitos de vida son pilares fundamentales para entender y predecir el riesgo clínico en seguros de salud.