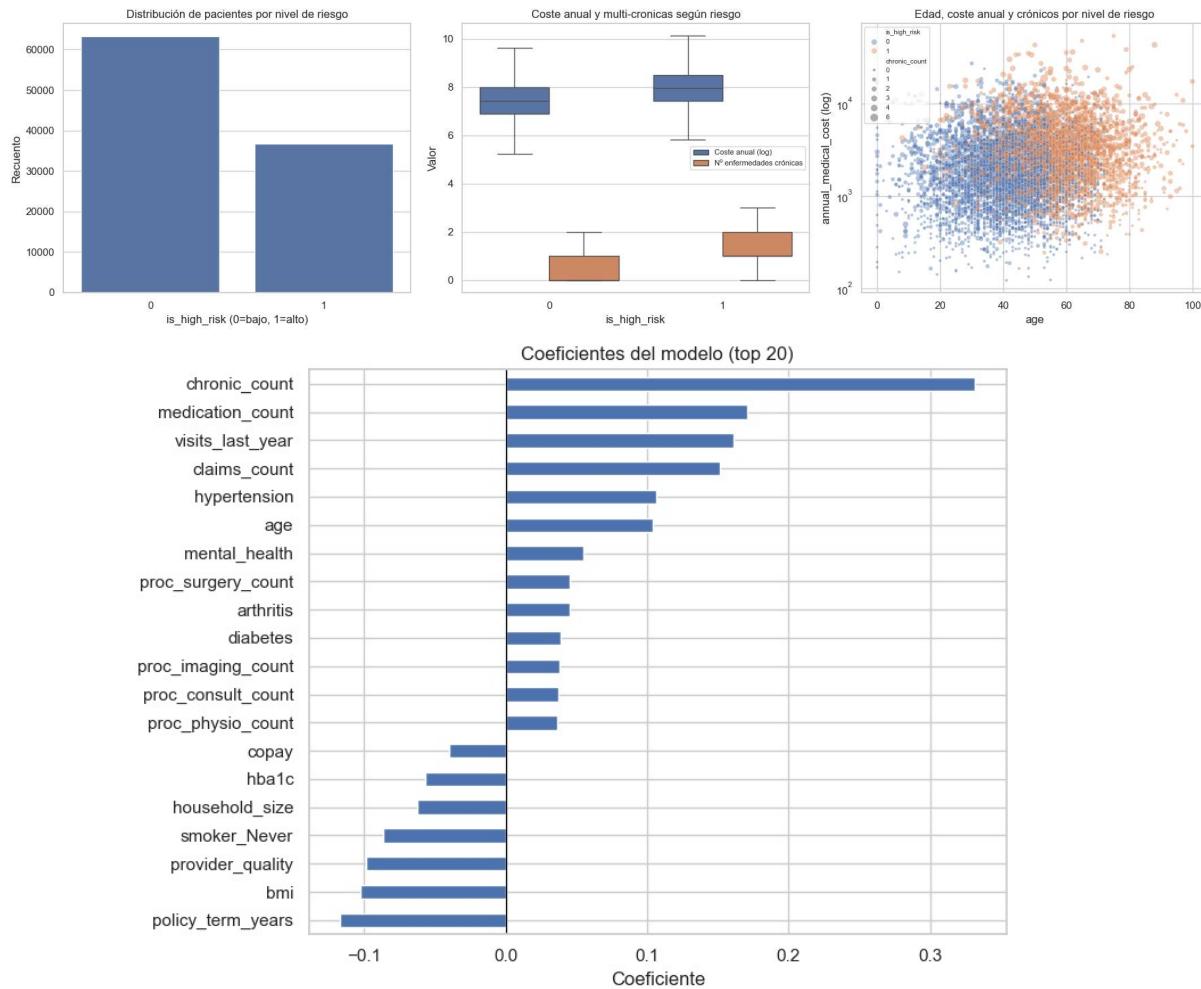


Nombre y Apellidos: Arturo Fernandez

Github con notebook:

https://github.com/ArturoFdezG/Proyecto_Visualizacion_Datos/blob/main/Final/DAVD_Examen_final_2025_2026.ipynb

1. Resumen Ejecutivo



De todo el análisis realizado se han encontrado los siguientes hallazgos relevantes.

Riesgo clínico:

- Los pacientes high risk presentan mucho mayor coste médico anual, con distribución muy sesgada y valores extremos.
- El número de enfermedades crónicas (chronic_count) es significativamente superior en el grupo de alto riesgo.
- Factores clínicos como presión arterial, HbA1c y LDL estarán relacionadas con deterioro progresivo y por lo tanto se asocian con mayor riesgo.
- Los pacientes high risk generan más hospitalizaciones, visitas y procedimientos, inflando el gasto anual.

Modelo:

- El modelo presenta buen equilibrio train/test, sin señales de sobreajuste.

- Las variables más influyentes para predecir high risk incluyen:
 - o Crónicos: chronic_count, diabetes, hypertension
 - o Clínicas: systolic_bp, diastolic_bp, Hba1c
 - o Hábitos: BMI, smoker, exercise
 - o Socioeconómicas: income, urban/rural, plan_type
- El modelo confirma que el riesgo clínico es multifactorial, no dominado por una sola variable, ya que hemos eliminado variables que pudiesen conducir a eso.
- La AUC muestra que el modelo discrimina bien entre pacientes de bajo y alto riesgo.
- Existe una relación clara entre edad avanzada y mayor probabilidad de ser high risk.

Pero ¿cómo se usa esto para un negocio?, o mejor dicho ¿cómo se convierte este conocimiento en valor real para una aseguradora? Para eso planteamos varias maneras de utilizar el conocimiento para reducir costes, incrementar ingresos, y por lo tanto, mejorar los beneficios.

Reducir costes:

- Identificación temprana de pacientes high risk mediante el modelo, intervención preventiva antes de que generen costes elevados.
- Programas personalizados de gestión de crónicos (diabetes, hipertensión, obesidad) dirigidos a los perfiles con coeficientes más altos.
- Optimización del uso hospitalario: seguimiento proactivo a quienes presentan mayor probabilidad de hospitalización.
- Asignación eficiente de recursos clínicos priorizando pacientes con mayor carga de enfermedades crónicas.
- Campañas de educación sanitaria dirigidas a hábitos (tabaco, ejercicio, estrés), reduciendo eventos costosos a largo plazo. Motivando a los clientes con algún tipo de inventivo (app, sorteo, descuentos...).

Incrementar ingresos:

- Ajuste dinámico de primas basado en el riesgo estimado (pricing basado en modelo), aunque para ello habría que mejorar el modelo.
- Segmentación de clientes para ofrecer productos diferenciados.
- Programas de fidelización centrados en usuarios low-risk con baja siniestralidad.
- Optimización de renovaciones: identificar clientes con riesgo creciente para renegociar pólizas o ajustar condiciones.

Mejorar beneficios (profit optimization). Son básicamente una combinación de las dos secciones anteriores.

- Mejor balance riesgo/precio, combinando predicción de riesgo y la prima.
- Automatización de decisiones basadas en riesgo, mejora la eficiencia operativa y reduce coste administrativo.
- Monitorización continua del riesgo para ajustar estrategias comerciales en tiempo real.

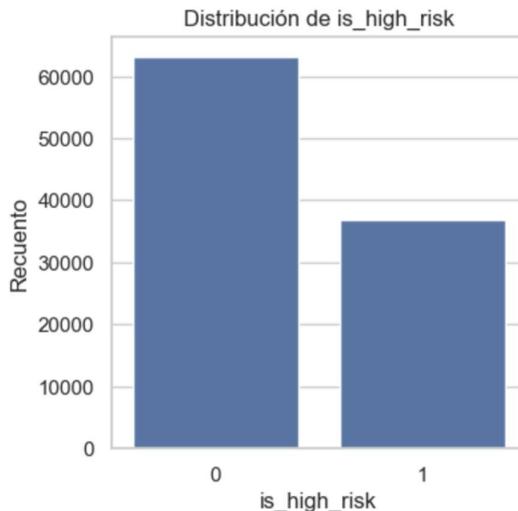
2. Gráficas del análisis exploratorio y breve explicación de cada una

Para realizar el análisis exploratorio, antes plantearemos distintas hipótesis, que trataremos de validar o rechazar con este primer análisis:

1. La variable objetivo de nuestro modelo puede estar desbalanceada, y puede tener sentido ya que, cogiendo muestras aleatorias de pacientes, lo normal sería que la mayoría estén en un nivel de riesgo moderado o bajo. Esto es relevante ya que puede afectar y sesgar a nuestro modelo posterior, por lo que en caso afirmativo, y en la proporción que se produzca, deberíamos tratarlo.
2. Los pacientes de mayor riesgo deberían gastar más anualmente en salud, ya que necesitan de mayores tratamientos.
3. Cuanto mayores enfermedades crónicas detectadas tenga un paciente, mayor será su riesgo.
4. De manera similar, cuanto mayor sea su BMI, factor importante de la salud, mayor será su riesgo.
5. A mayor edad, mayor riesgo y gasto.
6. Puede haber variables clínicas con correlación escondida, no planteada en ninguna de las hipótesis anteriores debido a su complejidad.

Nota: recordamos que esto son hipótesis, y es con los gráficos con lo que trataremos de confirmarlas o rechazarlas.

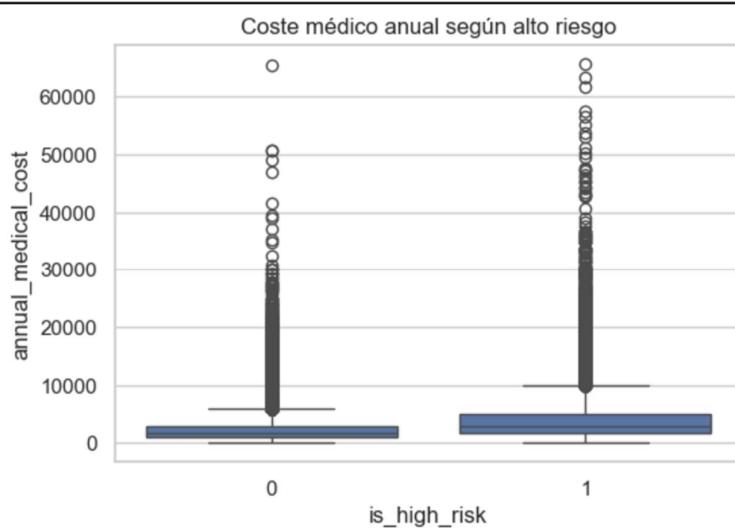
Primera hipótesis: Desbalanceo



En esta gráfica podemos observar como el dataset parece algo desbalanceado, lo cual tiene sentido si los datos son reales y no han sufrido sesgo, ya que parece que la proporción de los pacientes que sufren alto riesgo es menor, alrededor de la mitad que los que no son alto riesgo. Sin embargo, la desproporción aunque notable, puede ser no tratada para un primer modelo. Si observásemos que sí afecta a las predicciones, en tal caso podríamos utilizar algoritmos más complejos que tengan en cuenta el desbalanceo, asignar peso a la variable

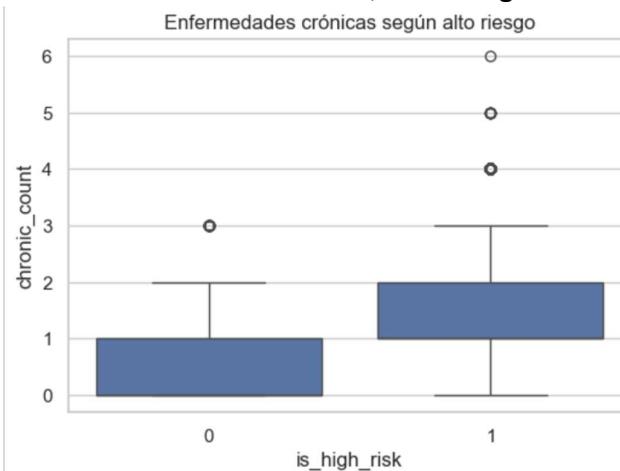
minoritaria o balancear directamente el dataset (generando datos sintéticos, eliminando filas de la clase mayoritaria...)

Segunda hipótesis: Mayor riesgo, mayor gasto.



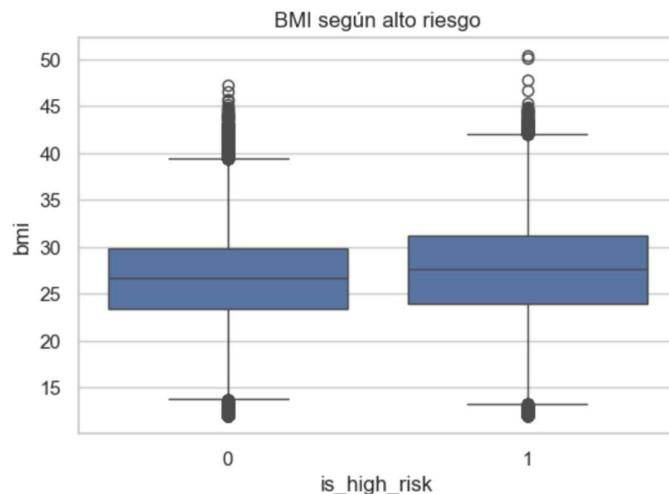
En esta otra gráfica, intentamos validar nuestra segunda hipótesis, sin embargo, es complicado aceptarla al 100%. En el gráfico de cajas, la media, la mediana del gasto parecen muy similares, aunque sí es verdad que en general, la población de los pacientes de alto riesgo parece tener en general más gasto. Los outliers eso sí, claramente son mayores en la clase de mayor riesgo, pero parece que hay pacientes de bajo riesgo que gastan mucho anualmente. Esto puede significar que hay personas que se preocupan más por su salud, e igual por eso son bajo riesgo, o que la gente con pocos recursos y por lo tanto que gasta menos, suele sufrir de mayor riesgo de enfermedad.

Tercera hipótesis. Más enfermedades crónicas, más riesgo.



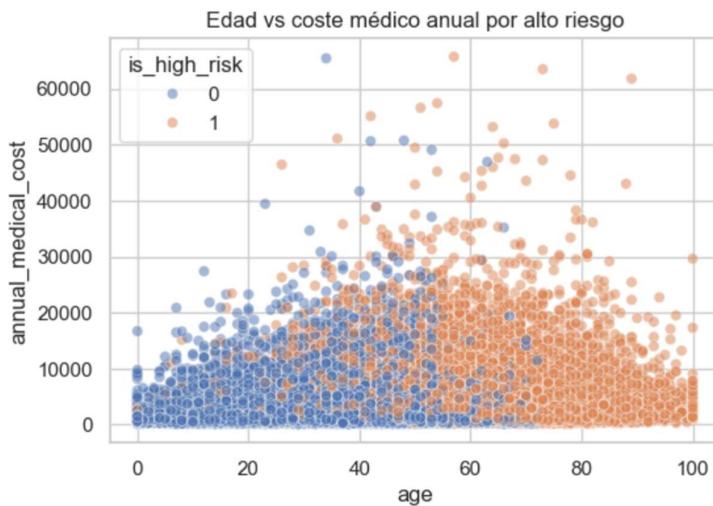
En este gráfico, confirmamos nuestra hipótesis claramente, parece haber una clara correlación entre el número de enfermedades crónicas, y el riesgo de un paciente. Aunque parezca obvio, es importante saberlo para confirmar las correlaciones, y los factores más relevantes a la hora de determinar el riesgo, que no olvidemos que es el factor más importante para una empresa aseguradora, pues es con lo que se determina el precio de los seguros.

Cuarta hipótesis: Mayor BMI, mayor riesgo.



En este caso, rechazaos la hipótesis, ya que aunque sí que parece que las muestras de alto riesgo, suele tener ligeramente mayor BMI, en general sí que está bastante homogéneo. Esto puede deberse a que el BMI de una persona solo tiene en cuenta su altura y peso, sin tener en cuenta que dicho peso puede venir de grasa o músculo. Una persona con alto BMI debido a su masa muscular puede estar bastante más sana que otra con un BMI en la media, pero con la mayoría de su peso en materia grasa.

Quinta hipótesis: A mayor edad, mayor riesgo y gasto.



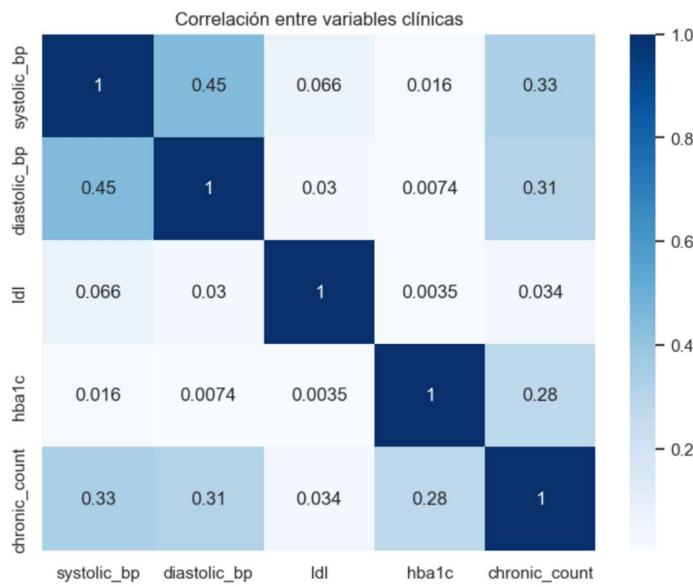
La aceptamos parcialmente, ya que efectivamente, a mayor edad mayor riesgo, se observa una separación clara en los colores. Sin embargo, a mayor edad mayor gasto no es claro, lo cual tiene sentido si volvemos a la Hipótesis 2, donde analizábamos que a mayor riesgo no siempre mayor gasto. Por lo tanto, la edad si es indicativo del riesgo, pero no del gasto.

Sexta hipótesis. Correlaciones ocultas.

- systolic_bp y diastolic_bp tienen correlación moderada (0.45), lo esperado clínicamente.

- chronic_count muestra correlación ligera con presión arterial y hba1c (entre 0.28–0.33), indicando que más enfermedades crónicas suelen acompañarse de peores métricas clínicas.
- ldl prácticamente no se correlaciona con el resto de variables (valores muy cercanos a 0).

En general, las correlaciones son bajas, lo que sugiere que cada variable aporta información relativamente independiente.



3. Modelo predictivo explicado y con tablas

Una vez hemos terminado el análisis exploratorio, trataremos de predecir la variable is_high_risk. La decisión es clara, un algoritmo de clasificación como el que vamos a usar clasifica en base a una probabilidad, en nuestro caso, la probabilidad de que sea alto riesgo. Por lo tanto, estaríamos calculando (con el modelo) la probabilidad de riesgo de los clientes, y este, es el factor más relevante de una aseguradora.

Para llevarlo a cabo, utilizaremos un algoritmo de Logistic Regression, por su sencillez y fácil interpretabilidad. Como hemos identificado desbalanceo en la variable target, pero para no utilizar algoritmos como Random Forest, utilizaremos la opción de los pesos (asignar más peso a la variable minoritaria). De esta manera resolvemos el uno de los problemas en el que este algoritmo puede fallar. El otro, la linealidad. Pero priorizamos explicabilidad y extracción de insights valiosos, frente a la predicción como tal.

Preparación de los datos:

Antes de realizar ningún modelo o predicción, es necesario saber bien con lo que estamos trabajando. Por ello, además de las conversiones de las variables categóricas (dummies) y la división del dataset en train y test para su posterior verificación, debemos excluir aquellas variables que no aportan nada a nuestro modelo, y que todo lo que pueden hacer es confundir o sesgar el modelo. Estas son:

- El id. No aporta información.

- El risk_score. Está tan correlacionado con la variable objetivo que hará que sea la de mayor importancia, y por lo tanto no podamos extraer mucha información del modelo.
- El annual_medical_cost. Ya que hemos confirmado que a mayor riesgo mayor gasto clínico, lo cual tiene sentido. Pero el modelo podría aprender la correlación, cuando en vez de causa, suele ser consecuencia. No es mayor gasto → mayor riesgo, sino, mayor riesgo → mayor gasto.

Una vez realizados estos ajustes, sí que podemos proceder con el modelo.

Primer modelo.

Para el primer modelo, usaremos un umbral a partir del cual considerar alto riesgo de 80%. No obstante, optimizaremos dicho valor para que las predicciones sean lo más precisas posible.

Lo primero que observamos son las métricas de predicción, tanto para train como para test. Confirmamos que no hay overfitting, lo cual es correcto.

Por otro lado, independientemente de las métricas como tal, tenemos una ACU de 0.88 aproximadamente, lo cual indica una alta capacidad de un modelo para distinguir entre clases y menor tasa de errores.

Por último, observamos el resto de métricas, las cuales muestran un buen comportamiento, con precisiones de entorno al 80%.

```
==> Métricas TRAIN ==
Accuracy (train): 0.7858125
ROC-AUC (train): 0.8745155107092565

Classification report (train):
precision    recall    f1-score   support
          0       0.76      0.95      0.85     50575
          1       0.86      0.50      0.63     29425

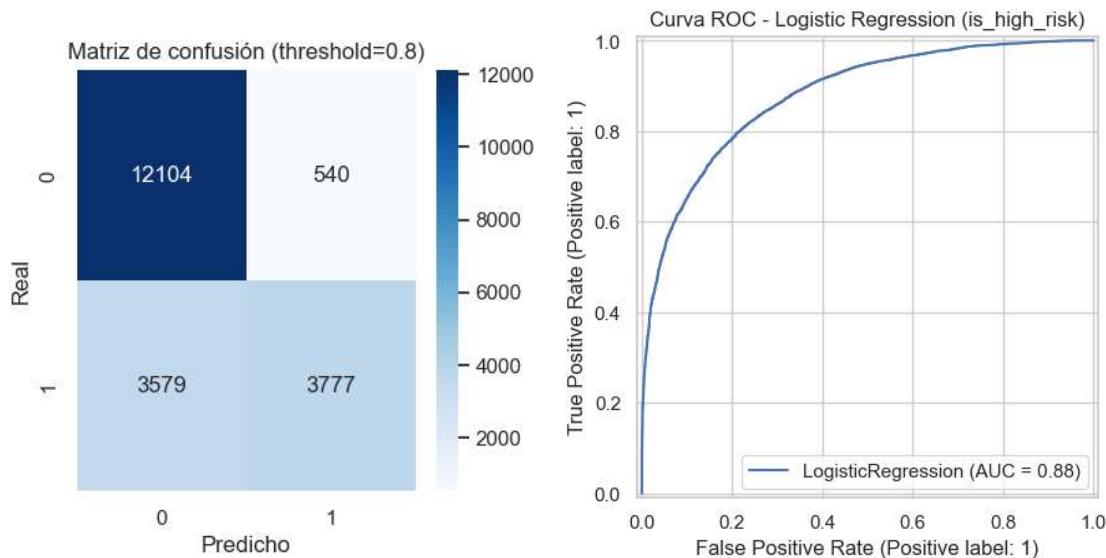
accuracy                           0.79     80000
macro avg       0.81      0.73      0.74     80000
weighted avg    0.80      0.79      0.77     80000

==> Métricas TEST ==
Accuracy (test): 0.79405
ROC-AUC (test): 0.878961465601964

Classification report (test):
precision    recall    f1-score   support
          0       0.77      0.96      0.85     12644
          1       0.87      0.51      0.65      7356

accuracy                           0.79     20000
macro avg       0.82      0.74      0.75     20000
weighted avg    0.81      0.79      0.78     20000
```

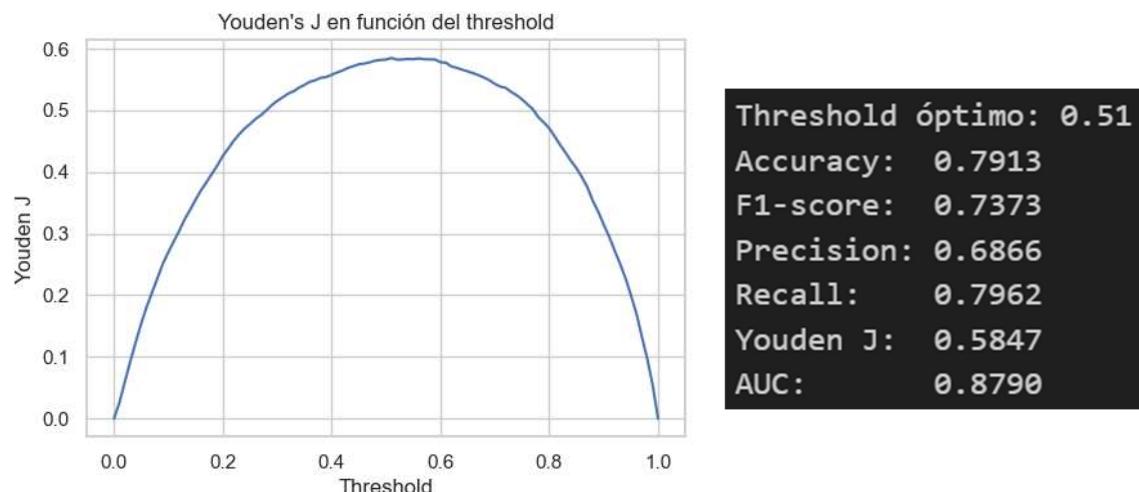
También extraemos la matriz de confusión (solo se muestra para test), en la cual observamos que la mayoría de los errores provienen de altos riesgos identificados como bajos. Esto es relevante para conocer la fiabilidad, si lo que priorizamos es la predicción de los de alto riesgo.



Antes de pasar a la importancia de las variables, vamos a tratar de optimizar este modelo, en base al umbral elegido, para optimizar esas predicciones erróneas de la variable.

Para ello optimizaremos el valor de Youden's J (o índice J de Youden) que es una métrica usada para elegir el mejor threshold en un clasificador binario.

Mide qué tan bien el modelo separa las dos clases; el valor óptimo es el que maximiza J.



Podemos ver como el valor óptimo según este índice es el 51%, por lo que estimaremos el siguiente modelo usando dicho valor.

Segundo modelo.

En este modelo usando el valor óptimo de umbral, se mantiene el no overfitting y valores razonables para métricas de predicción.

```

==== Métricas TRAIN ====
Accuracy (train): 0.787425
ROC-AUC (train): 0.8745155107092565

Classification report (train):
precision    recall   f1-score   support
0            0.86     0.79      0.82     50575
1            0.68     0.79      0.73     29425

accuracy                           0.79     80000
macro avg                          0.77     0.79      0.78     80000
weighted avg                       0.80     0.79      0.79     80000

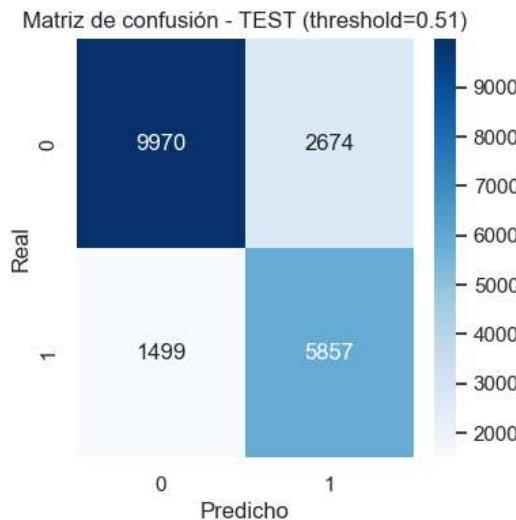
==== Métricas TEST ====
Accuracy (test): 0.79135
ROC-AUC (test): 0.878961465601964

Classification report (test):
precision    recall   f1-score   support
0            0.87     0.79      0.83     12644
1            0.69     0.80      0.74      7356

accuracy                           0.79     20000
macro avg                          0.78     0.79      0.78     20000
weighted avg                       0.80     0.79      0.79     20000

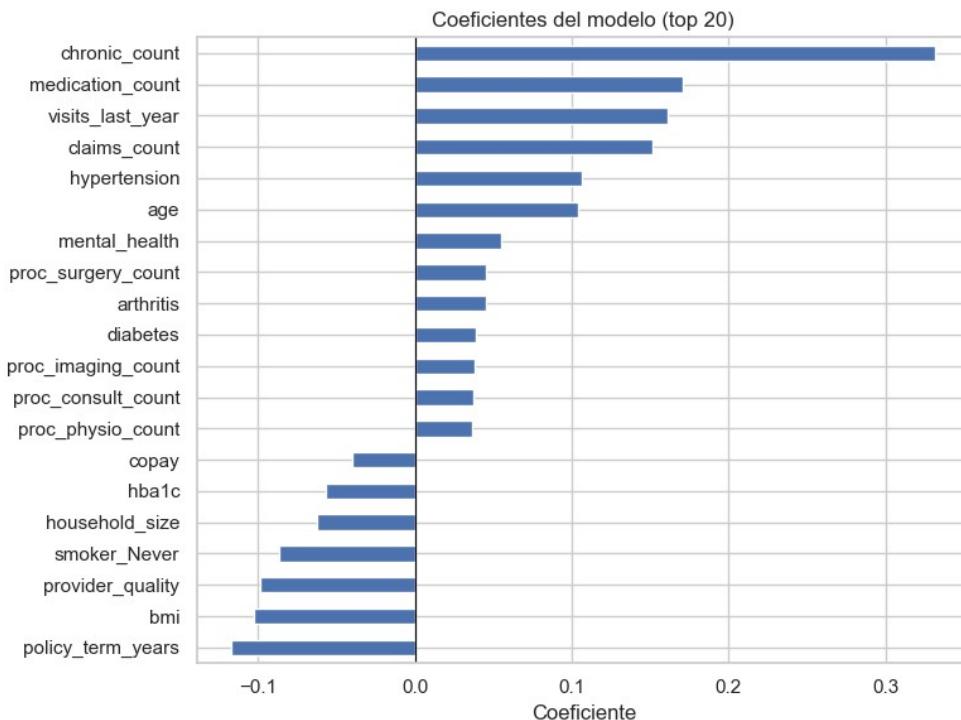
```

Eso sí, ahora la predicción de los positivos (los alto riesgo), ha mejorado mucho más, cometiendo la mayoría de los errores en bajos riesgo que se han predicho como altos. Mejor pasarse de cuidadoso que fallar en los alto riesgo.



Importancia de las variables.

Ahora, pasamos a visualizar la importancia de las variables, en base al coeficiente que la regresión logística ha otorgado a cada una de ellas, de donde se puede sacar información muy relevante y aplicable al negocio de las aseguradoras.



Podemos observar como algunas de las hipótesis confirmadas anteriormente en el análisis exploratorio, se confirman según el modelo de nuevo.

Las variables que más influyen (en valor absoluto) parecen ser:

- chronic_count, positivo (la más influyente)
- medication_count, positivo
- visits_last_year, positivo
- claims_count, positivo
- hypertension, positive
- policy_term_years, negativo
- age, positive
- bmi, negative

Esta es la continuación, pero mucho menos relevantes:

- mental_health, positivo
- proc_surgery_count, positivo
- arthritis, positivo
- Diabetes, positive
- proc_imaging_count positivo
- proc_consult_count,positivo
- proc_physio_count, positivo
- provider_quality, negativo
- smoker_Never, negativo
- household_size, negativo
- hba1c, negativo

- copay, negativo

Todas parecen tener sentido y estar alineadas con lo que uno puede pensar, excepto BMI, la cual parece que cuanto mayor sea, menor riesgo, lo cual puede ser contra el sentido común o lo que se trataba de plantear en la hipótesis. Esto como se ha comentado antes se puede deber a que el BMI no tiene en cuenta en realidad la procedencia del peso, pero este puede provenir de masa muscular, la cual se supone que tiene un impacto positivo en la salud.