Nombre y Apellidos: Allende de Yarza González-Lacarra
Github con notebook:
Nota: Por favor, seguir esta estructura para el documento

1. Resumen Ejecutivo

El objetivo principal de este proyecto es ayudar a la **aseguradora a reducir sus gastos y optimizar sus ingresos.** Para conseguirlo, existen **dos vías complementarias**. Por un lado, es fundamental **reducir el número de clientes considerados de alto riesgo**, ya que este grupo concentra gran parte del gasto sanitario debido a la presencia de múltiples enfermedades crónicas, un mayor número de hospitalizaciones y un uso más intensivo de servicios. Por otro lado, también es necesario **reducir el coste médico medio por cliente**, entendiendo **qué factores explican el gasto** y cómo pueden manejarse de forma preventiva o mediante una asignación más eficiente de recursos.

Descripción del dashboard y principales insights

El dashboard mostrado en la Figura 1 reúne las seis visualizaciones más relevantes del análisis, seleccionadas por su relación directa con las medidas estratégicas recomendadas. Su objetivo es ofrecer una visión integrada sobre los factores que influyen tanto en el riesgo de los clientes como en su coste anual, así como en los patrones de uso y hábitos de salud. En la primera fila del dashboard se incluyen dos visualizaciones clave: la importancia de variables en el modelo de predicción de high_risk y la importancia de variables en el modelo de predicción del coste anual. En ambos casos se han eliminado variables que son una consecuencia directa (como el tipo de tarifa) para centrarnos únicamente en factores accionables, aunque esto pueda reducir ligeramente la precisión del modelo.

En la segunda fila, la primera gráfica corresponde a la clusterización de comportamientos basada en hábitos y uso sanitario (alcohol, tabaco, visitas médicas y número de medicaciones). Esta segmentación divide a los clientes en tres grupos: el Cluster 0, con bajo consumo de recursos y hábitos saludables; el Cluster 1, con uso moderado de servicios y presencia de algunos hábitos de riesgo; y el Cluster 2, caracterizado por un uso intensivo del sistema y mayor proporción de clientes high_risk. Esta información resulta fundamental para diseñar estrategias diferenciadas tanto de prevención como de reestructuración de planes. La siguiente gráfica muestra cómo el riesgo aumenta de manera progresiva al segmentar el número de visitas en quintiles, evidenciando que la frecuencia de uso del sistema sanitario es un predictor directo tanto de alto riesgo como de un mayor coste anual.

Otra de las visualizaciones analiza el impacto del tabaquismo en la probabilidad de ser considerado high_risk. Los datos muestran que los fumadores actuales presentan más del doble de probabilidad de ser clasificados como de alto riesgo en comparación con exfumadores y no fumadores, lo que señala claramente la necesidad de medidas específicas como programas de deshabituación tabáquica, incentivos económicos por reducir el consumo o planes preventivos especiales dirigidos a este colectivo. Finalmente, la gráfica que agrupa el riesgo por región evidencia una distribución homogénea entre territorios, sin diferencias relevantes. A diferencia de otros mercados donde determinadas

zonas presentan mejores hábitos de salud, aquí el riesgo no depende de la región, por lo que se descarta la necesidad de diseñar estrategias territoriales diferenciadas.

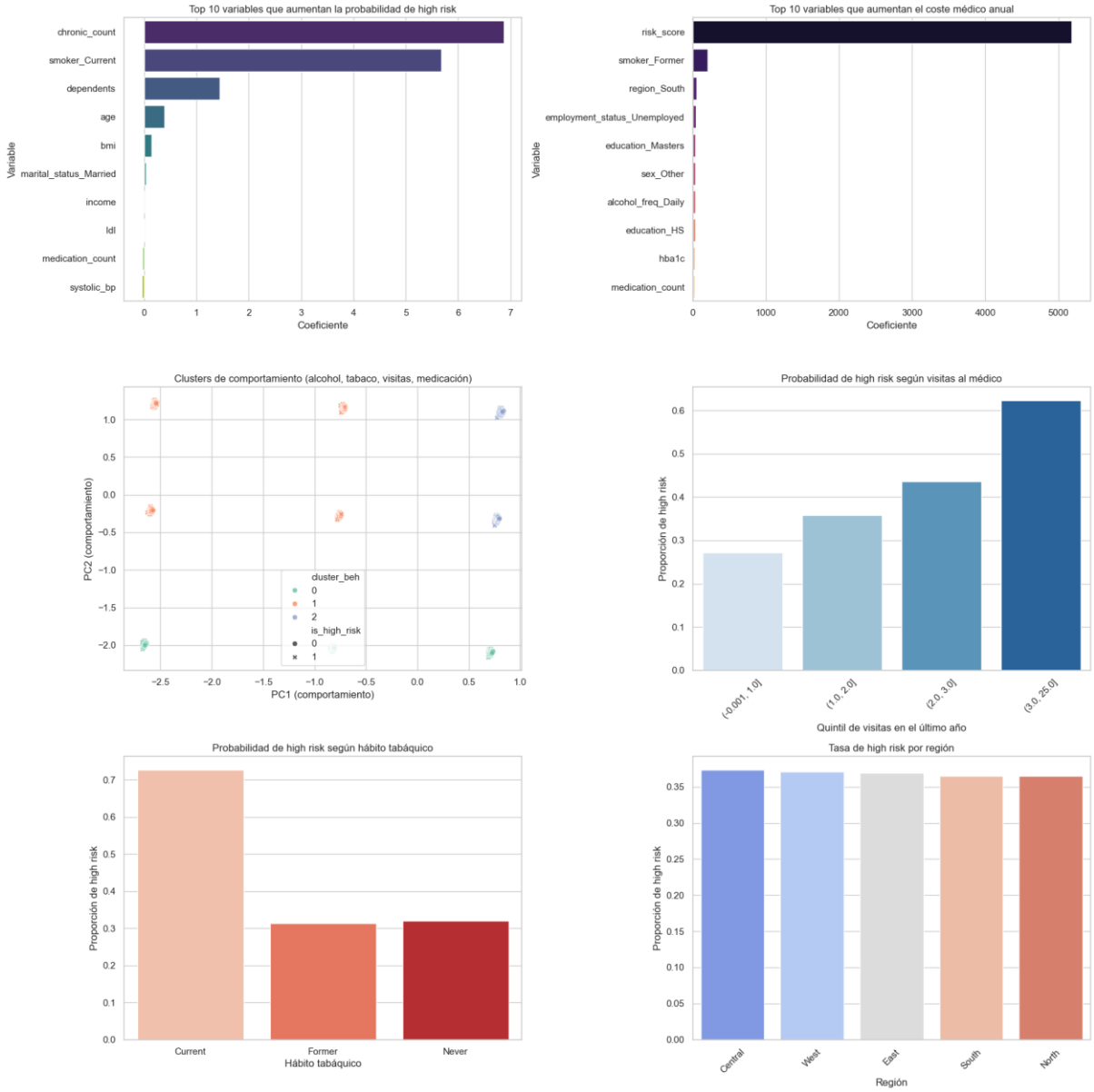


Figura 1 – Dashboard

Medidas propuestas

Basándonos en estos insights y en los resultados de los modelos predictivos, se proponen medidas en dos ejes: incentivos de salud orientados a la prevención y acciones de gestión del riesgo y del coste. En relación con los incentivos de salud, las visualizaciones muestran que los clientes con peores hábitos (fumadores, mayor consumo de alcohol, múltiples medicaciones) y mayor uso del sistema concentran la mayor proporción de riesgo y los costes más elevados. Por ello se recomienda implementar un programa de wellness incentivado que combine colaboraciones con gimnasios para ofrecer descuentos a los clientes asegurados, bonificaciones por asistencia o participación en retos de actividad física, y una aplicación que registre hábitos saludables. Esta medida está directamente

relacionada con los clusters de comportamiento, el impacto del tabaquismo y las variables predictoras de alto riesgo. Asimismo, se propone el desarrollo de una plataforma digital de salud que incluya consejos personalizados, testimonios (por ejemplo, sobre dejar de fumar), retos gamificados y seguimiento de hábitos, aprovechando las necesidades detectadas en los patrones identificados por la clusterización.

En cuanto a la gestión del riesgo y del coste, el análisis por región confirma que no tiene sentido diferenciar los planes de seguro territorialmente al no existir diferencias de riesgo significativas. También se sugiere revisar precios y coberturas para clientes reincidentes en el uso excesivo del sistema, evaluando la posible introducción de penalizaciones o de planes alternativos más económicos con coberturas más limitadas. Antes de aplicar cambios de precios, se recomienda realizar un estudio específico sobre el impacto en el churn, ya que modificaciones en la prima pueden afectar a la retención. Además, se plantea crear nuevas estructuras de planes más flexibles: planes más baratos con menor cobertura para clientes de bajo riesgo, planes preventivos dirigidos a los clusters con hábitos nocivos y planes mixtos que incentiven la reducción de comportamientos perjudiciales.

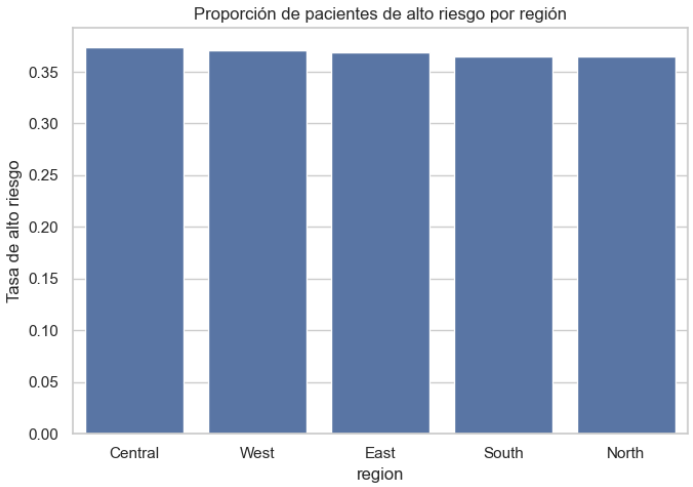
Como siguiente paso, se propone el desarrollo de una herramienta predictiva en tiempo real que permita introducir las características de un cliente y obtener de forma inmediata tanto su probabilidad de ser clasificado como high_risk como su coste médico estimado para el año siguiente. Esta herramienta sería de gran utilidad para la aseguradora en procesos de aceptación de pólizas, fijación de precios, planificación de recursos y diseño de programas preventivos personalizados

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

Copia aquí tus gráficas y explícalas. Mínimo 6.
Cada una de las gráficas esta pensada para un objetivos, por ejemplo queremos agrupar a lso clientes por region para ver como se realciona con… y poder accionar medidas para una region completa por ejemplo. Aquí se muestran las 6 visualizaciones con el objetivo que persigue cada una y los insights que sacamos

1. Proporción de pacientes de alto riesgo por región

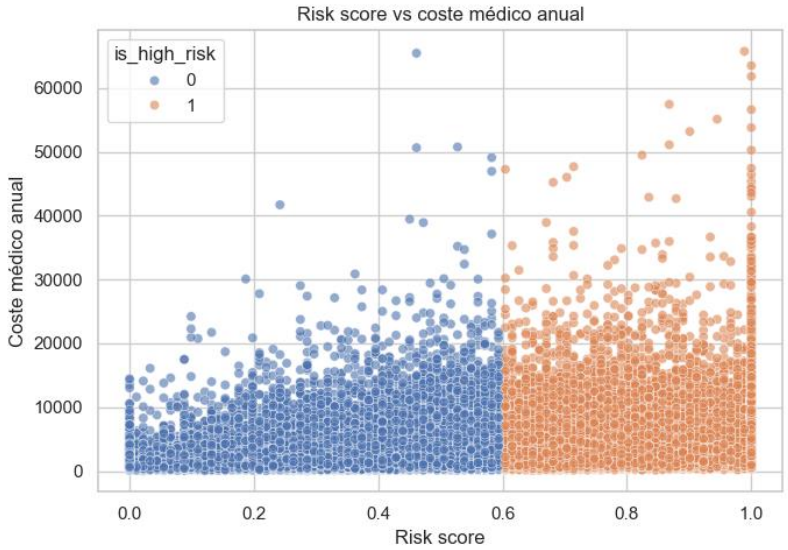
La gráfica permite comprobar si existen diferencias significativas de riesgo entre regiones para valorar si tendría sentido aplicar políticas o planes diferenciados por territorio.



Al observar que todas las regiones presentan prácticamente la misma tasa de is_high risk, concluimos que el riesgo no depende de la zona y no es necesario diseñar estrategias regionales

2. Coste anual si el cliente es de alto riesgo

Esta visualización muestra cómo el coste médico anual aumenta de forma clara a medida que incrementa el *risk score*, especialmente entre los clientes de alto riesgo

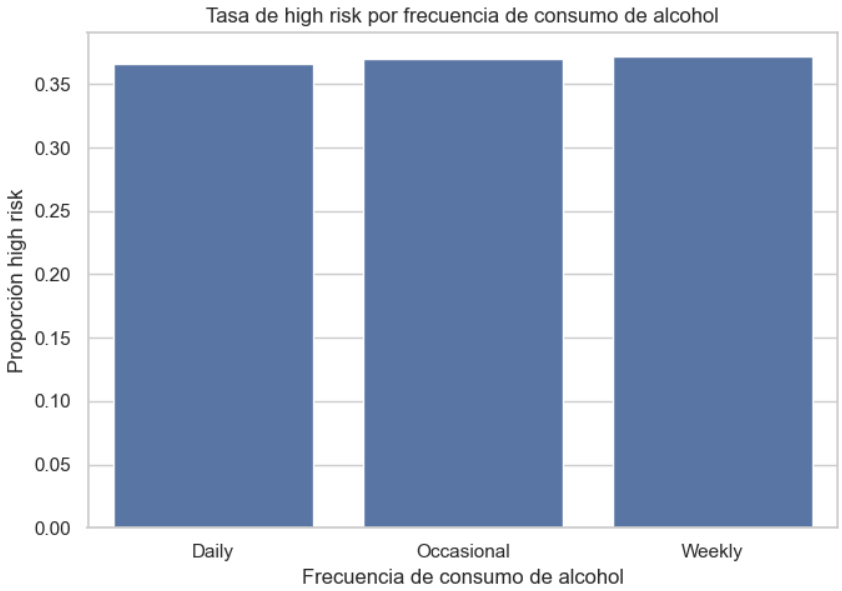


La diferencia de coste entre low risk y high risk justifica priorizar acciones preventivas, ya que reducir el número de clientes high risk tendría un impacto directo en el gasto de la aseguradora.

3. Consumo de alcohol y riesgo

La comparación entre frecuencias de consumo de alcohol revela que la tasa de high risk es muy similar entre daily, occasional y weekly.

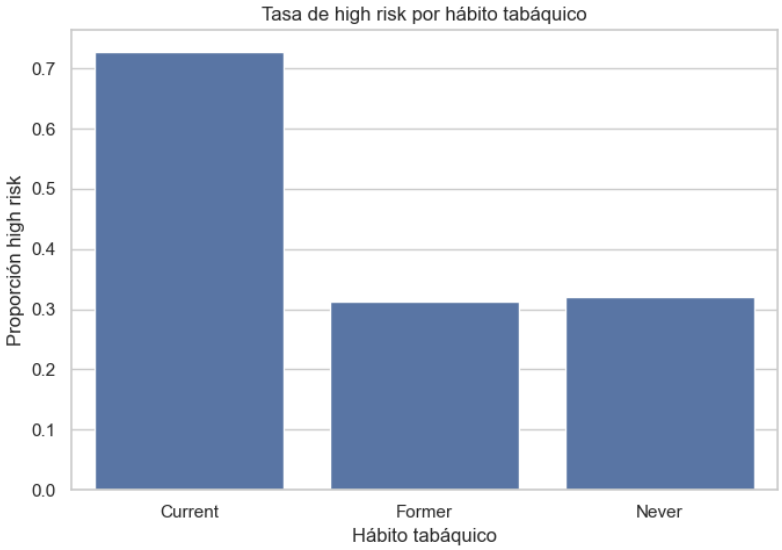
Esto indica que el alcohol no es un factor diferenciador relevante en este dataset y no debería ser una palanca prioritaria en las intervenciones.



4. Tabaquismo y riesgo

La gráfica muestra una diferencia muy marcada: los fumadores actuales presentan más del doble de probabilidad de ser high risk respecto a exfumadores y no fumadores.

Este resultado convierte al tabaquismo en una de las variables más críticas para intervenir, apoyando programas de deshabituación o incentivos por dejar de fumar.

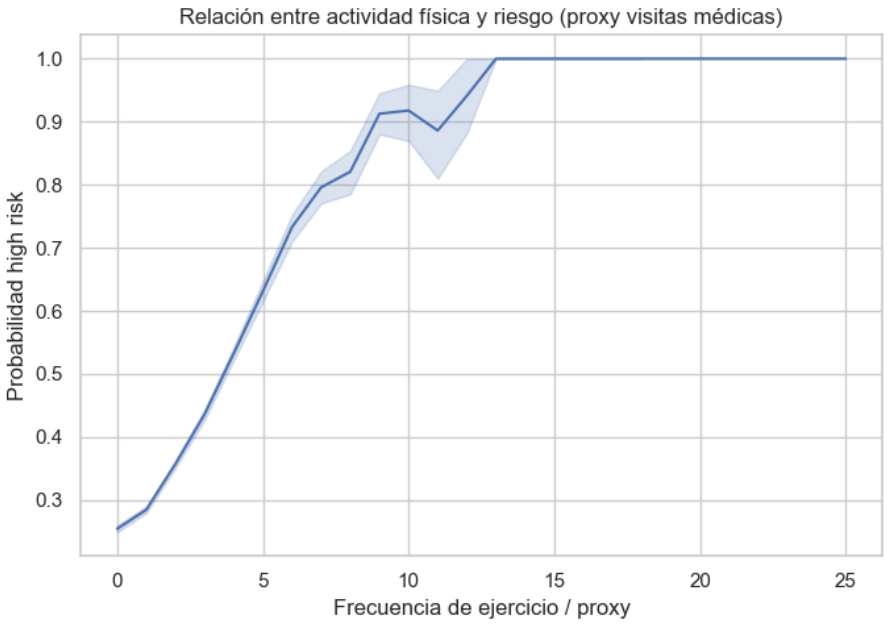


5. Actividad física y riesgo

6.

La gráfica muestra que la probabilidad de ser high risk aumenta conforme aumenta la frecuencia de visitas o actividad registrada, lo cual también puede reflejar la práctica de deportes con mayor riesgo de lesión (p. ej., fútbol, esquí), que generan más accidentes y más uso del sistema sanitario.

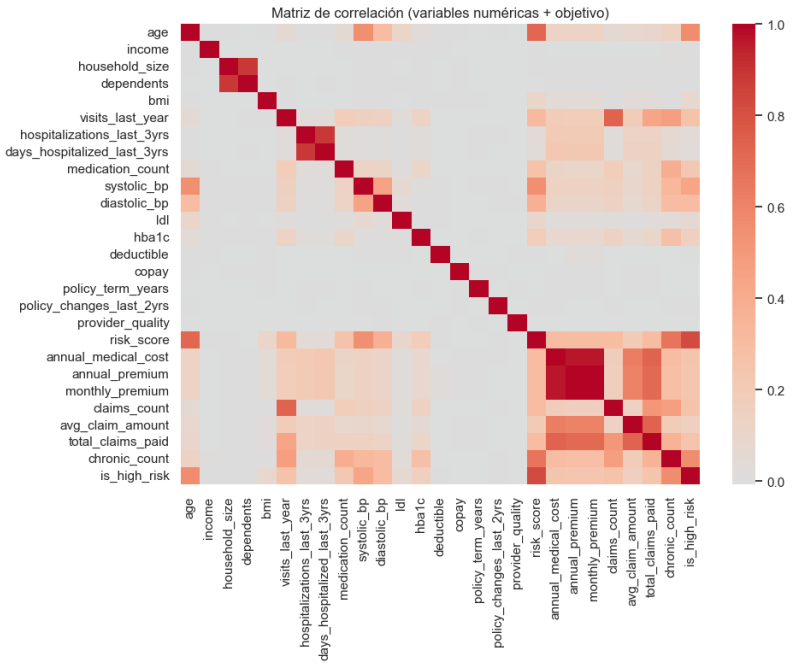
Aun así, dado que la actividad física es clave para mejorar la salud, nuestras propuestas se orientan hacia fomentar ejercicio de bajo riesgo (caminar, nadar suave, bicicleta moderada, yoga…) que mejore el bienestar sin aumentar las lesiones ni las visitas médicas.



7. Gráfica de correlaciones

La visualización permite identificar qué variables numéricas están más asociadas tanto entre sí como con el objetivo (is_high_risk), destacando la fuerte relación entre gasto anual, primas y uso del sistema, lo que confirma algo que ya era evidente: los clientes que utilizan más servicios generan mayores costes. Precisamente porque este patrón es tan obvio y estructural, no aporta valor accionable directo en los modelos.

El gráfico también muestra que variables de estilo de vida como el BMI, la edad o el número de medicaciones mantienen correlaciones moderadas con el riesgo, lo que justifica que los modelos se centren en estos factores. Son variables mucho más relevantes para obtener insights prácticos, ya que permiten diseñar medidas de prevención y cambios de comportamiento, a diferencia de factores estructurales como el tipo de plan o la prima.



3. Modelo predictivo explicado y con tablas

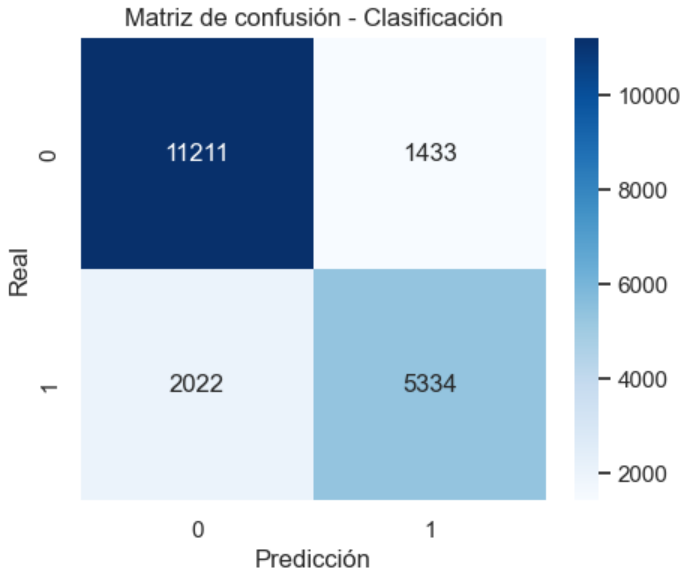
Nota: he empezado queriendo hacer los dos modelos, por eso alguna cosa de la redacción, luego me he dado cuenta de que era elegir uno y no me ha dado tiempo a cambiar lo redactado

Para los modelos predictivos nos centramos únicamente en variables relacionadas con el comportamiento, los hábitos y el estilo de vida de los clientes. Decidimos excluir variables que afectan al riesgo o al coste de manera obvia y directa, como el tipo de plan contratado o el precio de la tarifa, porque estos factores no aportan información nueva ni accionable: es esperable que un cliente con un plan más caro acuda más al médico o haga más uso del sistema. Al omitir estas variables, buscamos comprender realmente qué elementos del estilo de vida están asociados al riesgo y al coste, aun a costa de sacrificar algo de precisión en los modelos.

Para la clasificación de high_risk empleamos un modelo de **regresión logística**, cuyo objetivo es identificar la probabilidad de que un cliente sea considerado de alto riesgo. Este modelo logró un rendimiento sólido teniendo en cuenta la restricción deliberada del número de variables utilizadas. Sus métricas fueron: **accuracy de 0.827**, **precision de 0.788**, **recall de 0.725** y **F1-score de 0.755**. Esto indica un equilibrio adecuado entre la capacidad de identificar correctamente a los clientes de alto riesgo y la fiabilidad de las predicciones, especialmente considerando que trabajamos exclusivamente con variables de estilo de vida.

En cuanto a la interpretación del modelo, analizamos las variables más relevantes y sus coeficientes. Las variables con coeficientes positivos aumentan la probabilidad de ser high_risk, mientras que las de coeficiente negativo la reducen. Entre los factores que incrementan el riesgo destacan la edad, el número de medicaciones, el índice de masa corporal (BMI) y la frecuencia de consumo de alcohol. También se observa un efecto leve pero presente asociado al sexo masculino. Por el contrario, el hecho de ser exfumador o no fumador reduce significativamente la probabilidad de pertenecer al grupo de alto riesgo, lo que encaja con la evidencia clínica existente y refuerza la importancia de intervenciones orientadas al abandono del tabaco.

Estas conclusiones ayudan a identificar perfiles específicos sobre los que actuar: personas mayores con hábitos menos saludables, consumidores frecuentes de alcohol y clientes con alta carga farmacológica. A partir de estos resultados se pueden diseñar programas preventivos, estrategias de pricing y acciones segmentadas según el comportamiento y los patrones de salud reales del cliente.



Metric	Value
Accuracy	0.82725
Precision	0.7882370326584898
Recall	0.7251223491027733
F1 Score	0.7553635913049636

Importancia de variables (coeficientes del modelo)

feature	coef
age	2.220191
medication_count	0.742955
bmi	0.360396
alcohol_freq_Occasional	0.068848
alcohol_freq_Weekly	0.050347
alcohol_freq_nan	0.031996
sex_Male	0.029711
sex_Other	0.016256
smoker_Never	-3.255660
smoker_Former	-3.263613

Variables que aumentan la probabilidad de high_risk

feature	coef
age	2.220191
medication_count	0.742955
bmi	0.360396
alcohol_freq_Occasional	0.068848
alcohol_freq_Weekly	0.050347
alcohol_freq_nan	0.031996
sex_Male	0.029711
sex_Other	0.016256

Variables que disminuyen la probabilidad

feature	coef
age	2.220191
medication_count	0.742955
bmi	0.360396
alcohol_freq_Occasional	0.068848
alcohol_freq_Weekly	0.050347
alcohol_freq_nan	0.031996
sex_Male	0.029711
sex_Other	0.016256