Nombre y Apellidos: Carlos Sánchez-Cabezudo Pinto
Github con notebook:
https://github.com/carlossanchezcabezudo/ProyectoDesarrolloApps/

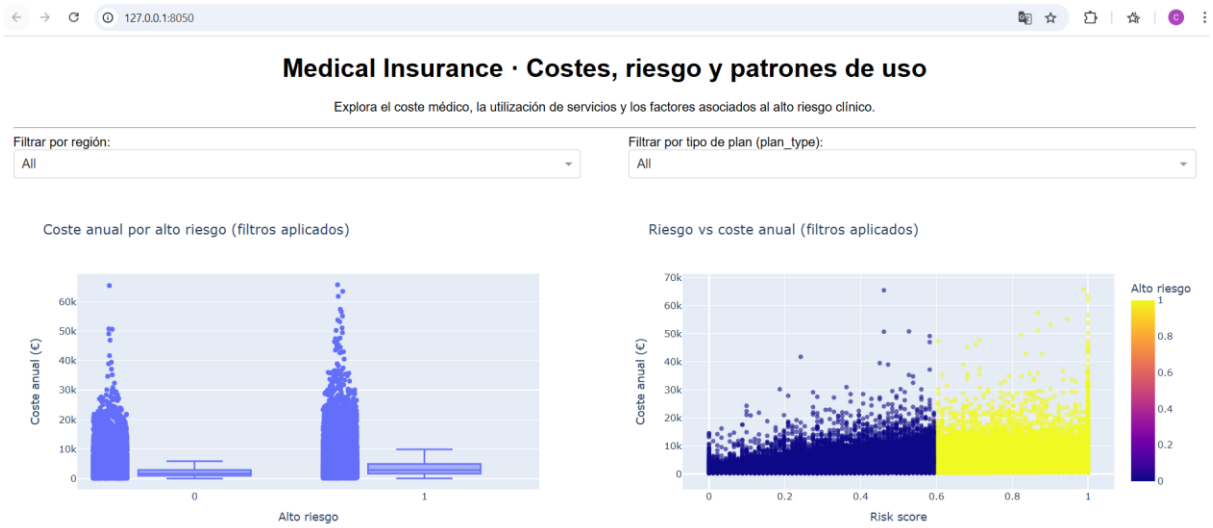
Nota: Por favor, seguir esta estructura para el documento

Resumen Ejecutivo y Dashboard

El análisis realizado a +100.000 asegurados demuestra ciertos patrones contundentes que ayudan a entender cómo se comportan los costes médicos, la utilización de servicios y los factores que pueden disparar el riesgo clínico en la cartera de la aseguradora. Considero que los hallazgos encontrados como analista de datos de 5º de ICAI ofrecen una visión clara de qué pacientes generan mayor gasto, por qué y qué acciones pueden tomar los gestores para anticipar estos riesgos y optimizar costes, asegurando un potencial de retorno mucho mayor sin exponerse a ciertos riesgos que en este trabajo resultan patentes.

Dada la variedad de planes y geografías, considero interesante que estos sean los filtros a utilizar en el dashboard, de tal forma que se puedan combinar geografías y planes aseguradores y observar cómo varían los 4 gráficos mostrados en el dashboard.

Las primeras visualizaciones evidencian que los pacientes clasificados como alto riesgo concentran una proporción sustancialmente mayor del gasto anual:



En el primer gráfico se observa una distribución claramente más elevada de costes en el grupo de alto riesgo. Además, se observa una mayor dispersión: estos pacientes no solo le cuestan más a la aseguradora, sino que también tienen gran variabilidad, típico en personas con cuadros clínicos complejos y, desde el punto de vista de una aseguradora, también un punto negativo a tener en cuenta. Este patrón se refuerza en la comparación de medias que se verá después, apreciándose que el coste anual del grupo de alto riesgo casi duplica al de bajo riesgo.

El scatterplot muestra también una relación creciente entre el risk score y el coste anual, tanto en la media para cada instancia (el groso de las concentraciones) como en la presencia de outliers. Como en este caso los outliers pueden considerarse instancias de pacientes con episodios graves u hospitalizaciones prolongadas, el hecho de que un mayor risk score venga acompañado de más patologías de este estilo es otro factor a tener en cuenta para una aseguradora que pretenda minimizar el riesgo al que se expone por cada asegurado que le contrata.

El total pagado en reclamaciones sigue la misma tendencia, lo que confirma que el alto riesgo no es solo clínico, sino que tiene un impacto económico directo y significativo, confirmando que un alto riesgo implica mayor coste para la aseguradora y cuantificándolo mediante el siguiente gráfico:

Con el último gráfico, el mostrado en la imagen superior a la derecha, se busca explicar realmente y en términos accionables para la empresa aseguradora qué tipo de perfil es de alto riesgo. En este sentido, el modelo predictivo desarrollado ayuda a entender los factores determinantes detrás de la clasificación.

En lo que se refiere a producir mayor perfiles de riesgo, la presencia de enfermedades crónicas es el predictor más relevante y el risk score el segundo, validando que el score sintetiza bien la información clínica subyacente.

Hábitos como ser fumador o no serlo destacan con coeficientes notables: exfumadores y no fumadores presentan patrones muy diferenciados frente a fumadores activos. De hecho, se puede observar como no ser un ávido fumador tiene el mayor impacto positivo en materia de salud, habiendo una clarísima correlación entre aquel que es fumador y aquel que supone un riesgo para las aseguradoras. Aparte de no fumar, también destacan como drivers positivos no beber frecuentemente, también lógico y esperable.

Para concluir este trabajo, considero de especial importancia para que el futuro de la aseguradora sea próspero y rentable actuar contundentemente en tres verticales:

1. Comenzar el estudio filtrando por el tipo de plan en cuestión (el tipo de plan del individuo que estén analizando si es prospectivo y, si es cliente actual, el tipo de plan que posee) así como la región (geografía/mercado en el que se localiza el individuo), ya que ambas medidas son relevantes y acotan el estudio, haciendo que los futuros pasos sean más relevantes y accionables para la aseguradora.

2. Realizar un estudio de las patologías crónicas del asegurado prospectivo, utilizando su presencia como un elemento que empeore las condiciones del plan contratado por el mismo. Otra forma de verlo, sería priorizar las intervenciones a este grupo asegurados ya contratados para tratar de disminuir su impacto negativo en las cuentas de la aseguradora.

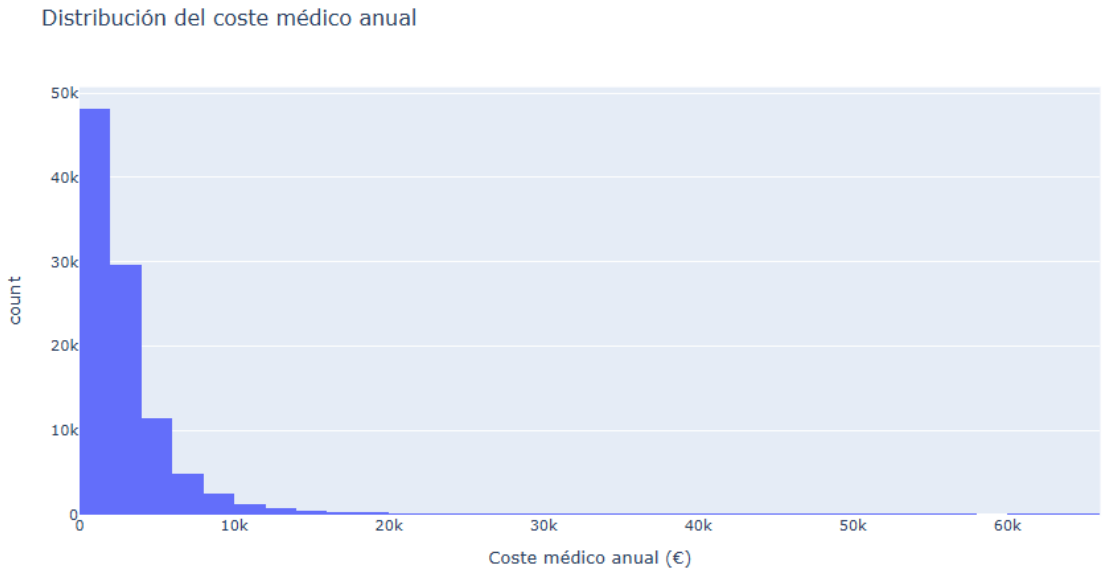
3. Por último, realizar un estudio también de los hábitos de vida de los individuos, priorizando contratar a aquellos que no sean fumadores ni frecuenten beber alcohol. Además, indica una oportunidad para mejorar la rentabilidad de la aseguradora si consiguen revertir estos hábitos en sus pacientes actuales, por lo que podría resultar interesante que la aseguradora realizara un trabajo didáctico en este sentido mediante programas de dejar de fumar, coachings saludables y ofrecer incentivos conductuales vinculados a las primas aseguradas

Gráficas del análisis exploratorio y breve explicación de cada una

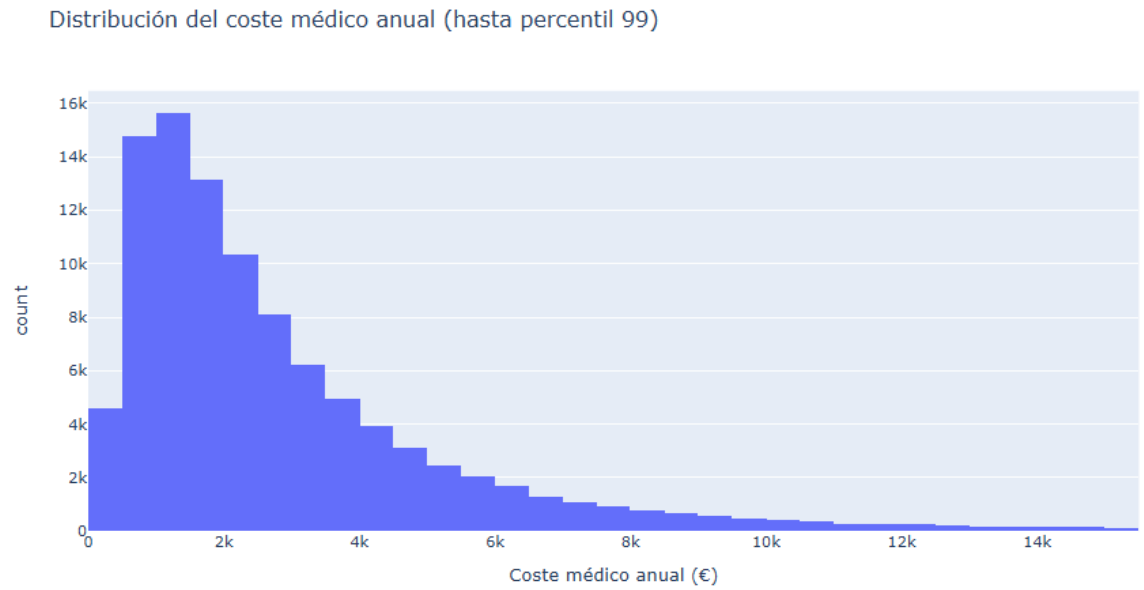
En primer lugar, resulta interesante comenzar el análisis exploratorio por la más elemental de las observaciones: cuánto dinero gasta cuánta gente.

Para esto, realizo un gráfico de barras sencillo en el que claramente se observa una tendencia (o, en este caso, relación) inversa y exponencial entre la cantidad de gente y el coste médico anual.

Se puede observar en la visualización inferior:



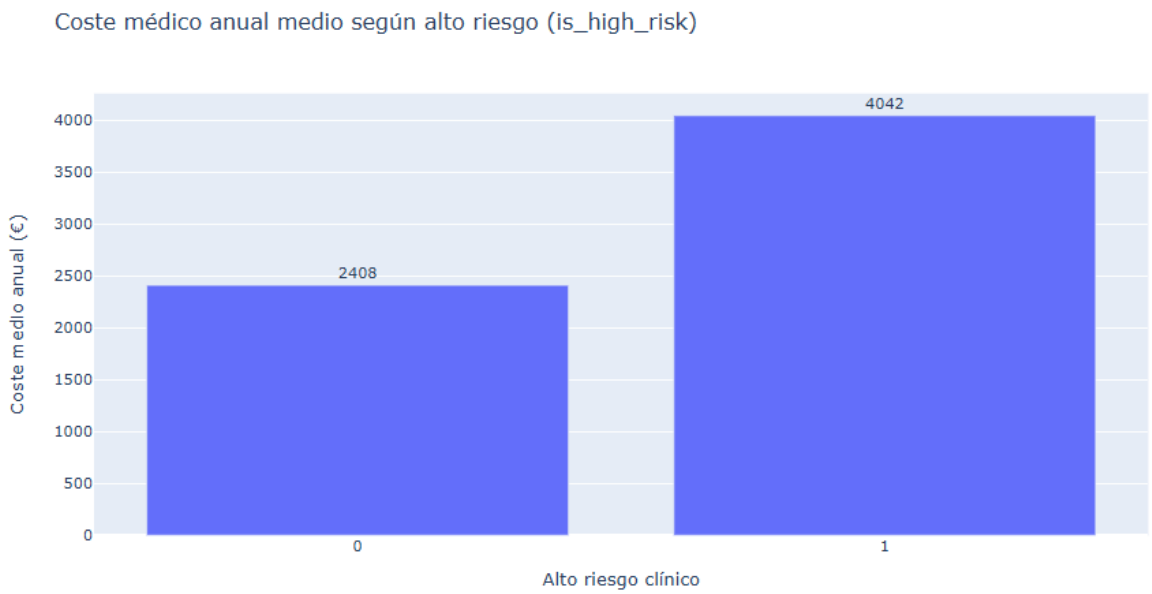
Para poder ver con más detalle esta tendencia, retiro el 1% más “extremo” y conservo únicamente hasta el percentil 99:



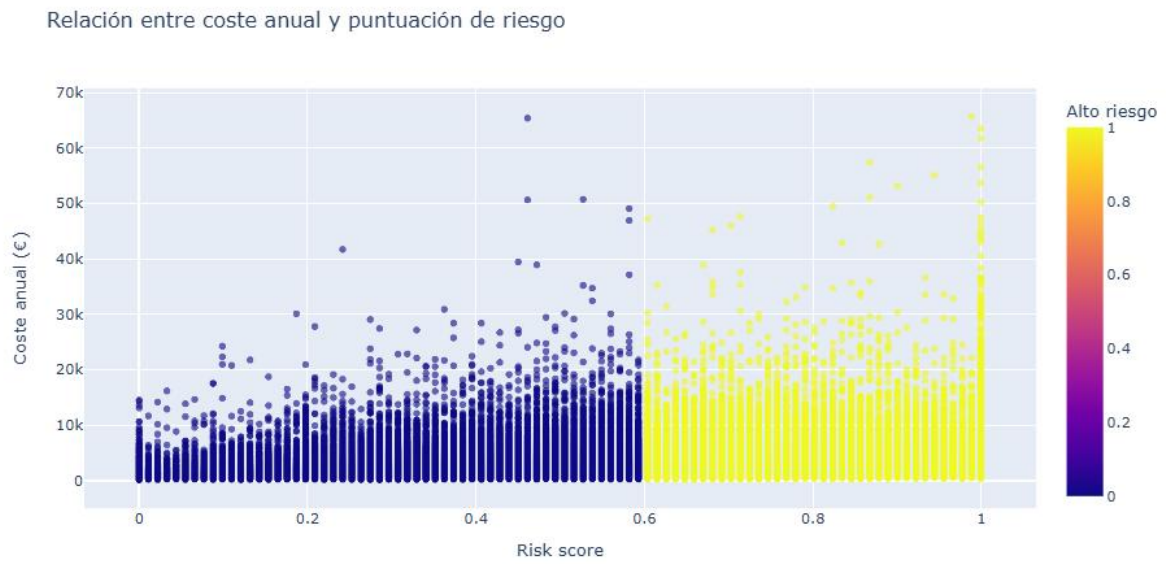
Con estos dos gráficos se observa la tendencia anteriormente mencionada y además se detallan más dos conceptos interesantes:

- Eliminar el 1% de mayor gasto anual acota bastante el coste médico en términos absolutos
- Claramente se observa que la mediana del dataset gravita a un gasto médico de entre 1000 y 2000 euros anuales

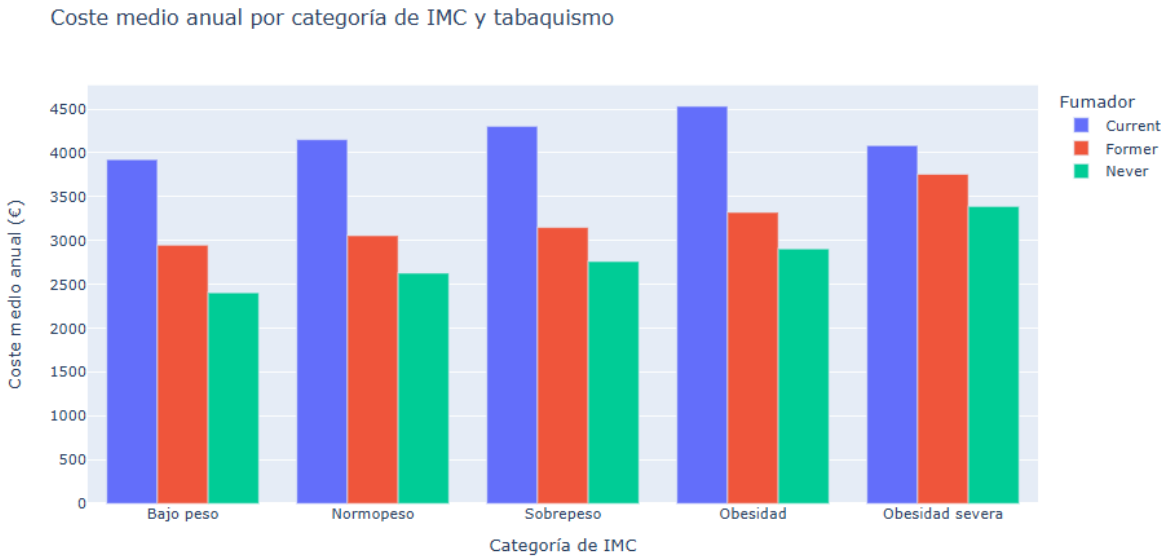
Después se realiza una observación rápida para ver el efecto del alto riesgo clínico y el coste medio anual en base a esta métrica. Como cabe observar, es un factor determinante con un impacto más que notorio en el gasto médico anual:



Para visualizar esto de una forma más lineal, se escoge el siguiente gráfico:



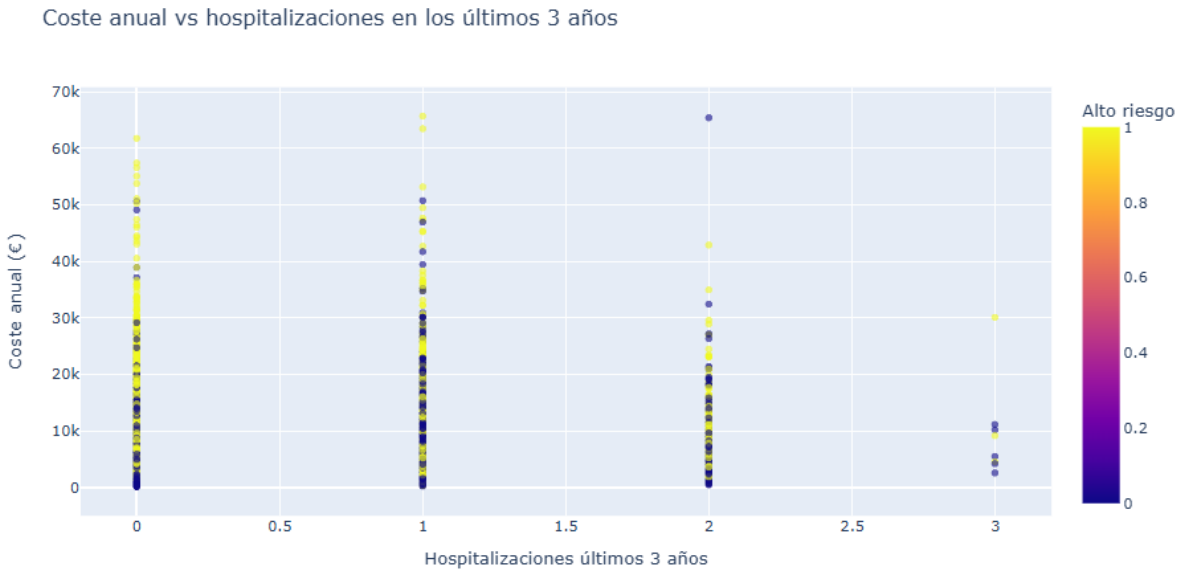
En el gráfico superior, se pueden observar las distintas instancias para el coste médico y el risk score mencionado en el anterior gráfico. Como se observa, aumentan tanto los gastos puntuales por encima de la media para cada risk_score (outliers) como el groso de las instancias registradas para dicha medida. Es decir, la tendencia de a mayor riesgo, mayor gasto, es ineludible y patente.



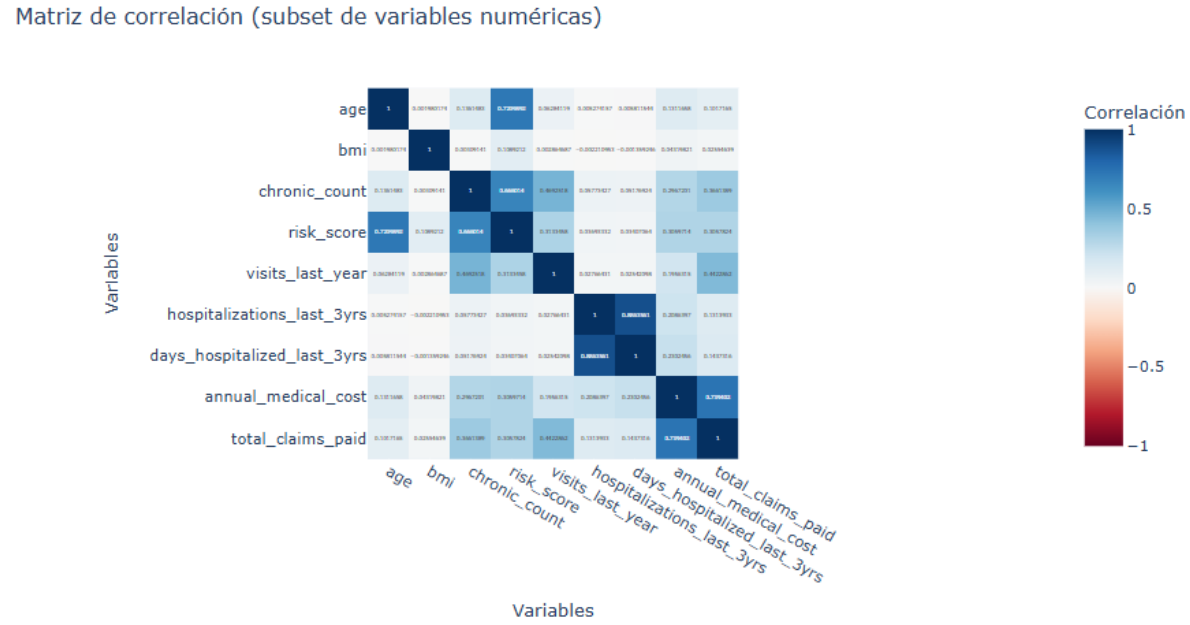
Visto lo visto se decide, mediante la visualización superior, estudiar distintas enfermedades registradas relacionadas con el IMC y su link con el tabaquismo (tanto para actuales fumadores como para antiguos fumadores). Resulta interesante observar que, independientemente de si el fumador corrige dicho comportamiento o no, la tendencia de un mayor coste anual para ambos casos en comparación con el no fumador es, de nuevo, obvia. De todas formas, cebe resaltar, que en todas las categorías de IMC analizadas ser un fumador en el presente supone un mayor coste que haberlo sido solamente en el pasado.

A continuación, se observa el coste frente a las hospitalizaciones en los últimos 3 años. Lo pongo a modo de indicar, únicamente que por la presencia de outliers no puedo sacar

ninguna información especialmente relevante de este gráfico en comparación con la información obtenida a partir del resto de los mismos.



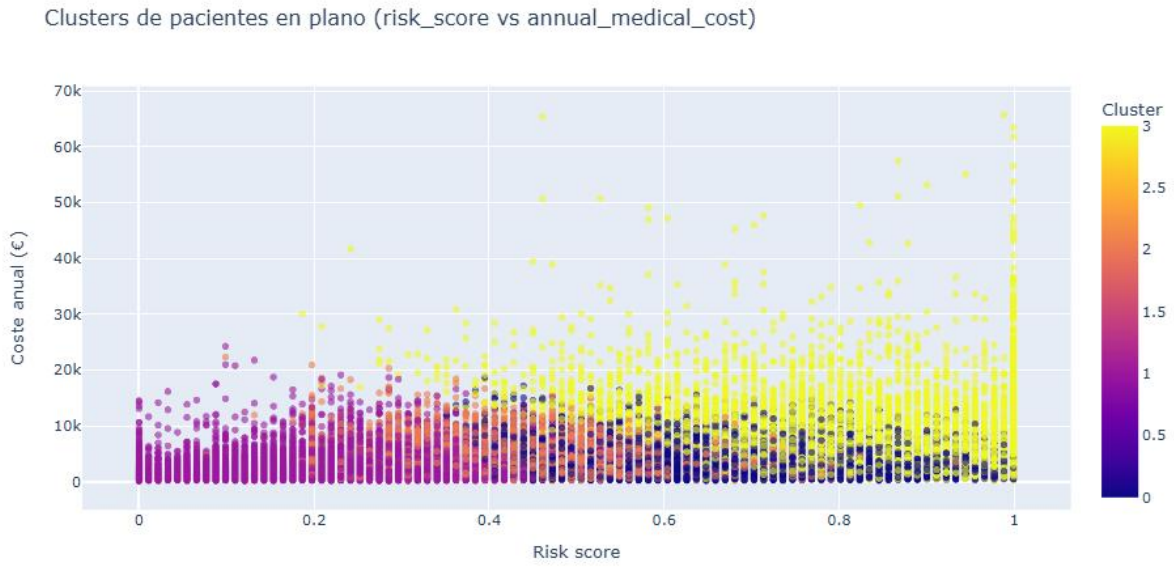
Ahora, manteniendo un análisis centrado en la observación de las variables cuantitativas, se hace una matriz de correlación (colorida para que sea más fácilmente observables las tendencias):



Se observan varias correlaciones muy obvias e insulsas, ya que es obvio, por ejemplo, la correlación entre las ocasiones de hospitalización en los últimos 3 años y la cantidad de días pasados hospitalizado.

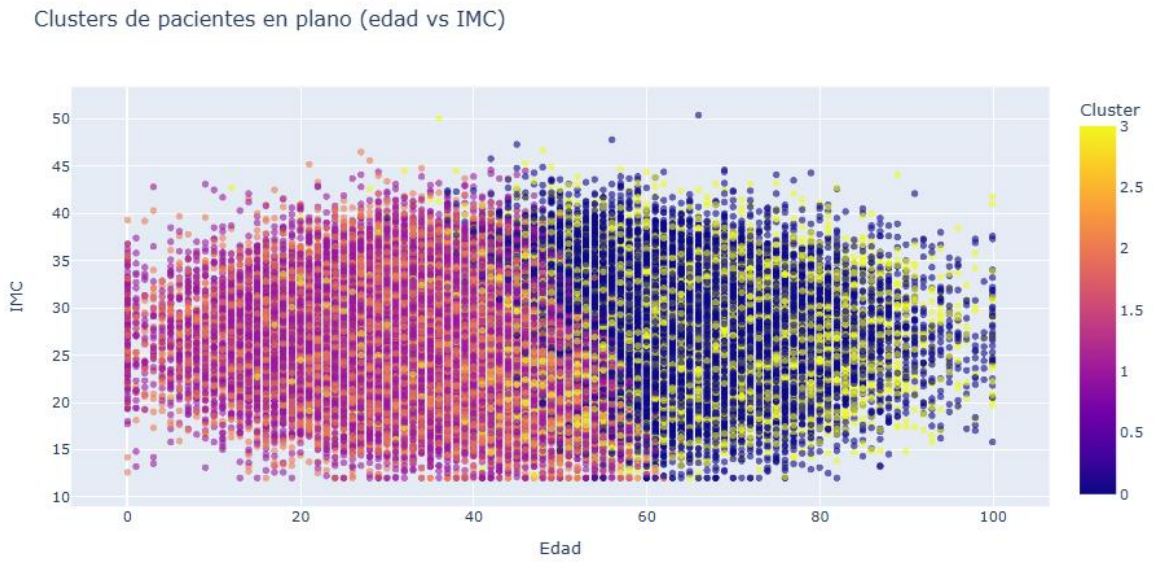
También se observa alguna tendencia que, aunque obvia, viene bien recordar para este análsis, como la correlación positiva (y muy positiva) entre la edad y el factor de riesgo.

A continuación, se decide también hacer un análisis por clusters. Aquí vemos si los clusters se organizan en torno a diferentes niveles de riesgo y coste (por ejemplo, clusters de alto riesgo y alto coste frente a clusters de bajo riesgo y bajo coste).

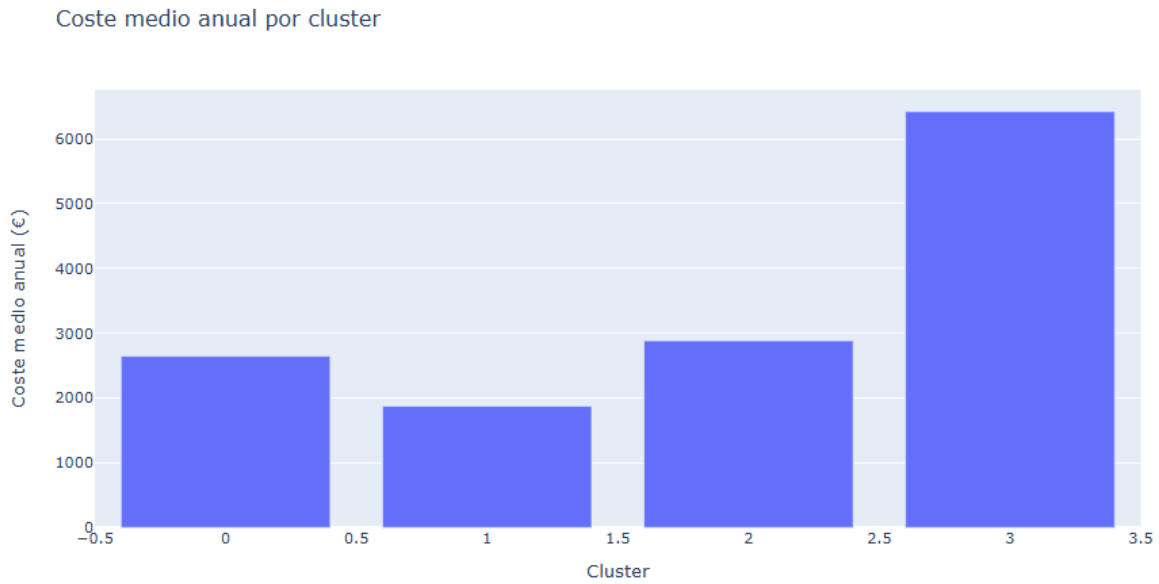


Se observa de nuevo la misma tendencia vista anteriormente pero ahora a nivel cluster.

Por ver una tendencia distina (no estrictamente creciente), se analiza el siguiente grupo de clusters, para entender si hay clusters asociados a perfiles de edad y peso:



En este caso se observa una mayor concentración en torno a la edad adulta, lo cual es lógico pues el IMC tiene sentido que no prevalezca entre las edades extremadamente altas (gente de +80 años, con tendencia a estar más “desganados”) o gente muy joven (muy activa en lo que se refiere a su metabolismo y actividad física, aparte de estar en crecimiento).



El siguiente resumen cuantifica el coste medio, el riesgo medio y el nº de instancias crónicas por cluster, permitiendo identificar qué grupos son los más problemáticos desde el punto de vista económico y clínico:

	cluster	annual_medical_cost	risk_score	chronic_count
0	0.0	2645.888811	0.657986	0.542952
1	1.0	1873.657851	0.265699	0.052834
2	2.0	2885.189247	0.489194	1.159752
3	3.0	6422.762559	0.857756	1.919461

Se observa claramente, que los pacientes más problemáticos son aquellos con enfermades crónicas y un alto factor de riesgo.

La diferencia más notable aparece en el Cluster 3, que sobresale de manera contundente: con un risk_score medio de 0.85 y casi 2 enfermedades crónicas por persona, este segmento concentra a los pacientes más complejos (peores desde el punto de vista de la aseguradora). Su coste anual es más del doble que el resto de clusters. Esto confirma que la combinación de cronicidad elevada y riesgo clínico alto se traduce directamente en mayor gasto sanitario y, por lo tanto, se llega a la conclusión de que este grupo debe ser prioritario para las distintas estrategias explicadas en el resumen ejecutivo.

Modelo predictivo explicado

Ahora se va a entrenar un modelo de clasificación (de Regresión Logística) para predecir si un asegurado es de alto riesgo, apoyándonse siempre en variables clínicas, de hábitos y de utilización (las provistas en el dataset de la práctica).

En primer lugar, se categorizan las variables para realizar el modelo y que todo funcione como es debido:

```
Valores únicos de is_high_risk: [0 1]
Variables numéricas usadas: ['age', 'bmi', 'chronic_count', 'risk_score', 'visits_last_year', 'hospitalizations_last_3yrs', 'da
ys_hospitalized_last_3yrs', 'medication_count', 'annual_medical_cost', 'total_claims_paid', 'annual_premium', 'claims_count',
'avg_claim_amount']
Variables categóricas usadas: ['sex', 'region', 'urban_rural', 'smoker', 'alcohol_freq', 'hypertension', 'diabetes', 'copd', 'c
ancer_history', 'kidney_disease', 'liver_disease', 'arthritis', 'mental_health', 'plan_type', 'network_tier']
Dimensiones finales para el modelo: (69917, 28) (69917,)
```

Una vez se tiene esto, se puede entrenar la regresión logística, separando en train y test y evaluando el accuracy del modelo de clasificación creado:

```
Accuracy (train): 0.9135036561600486
Accuracy (test): 0.9124713958810069
```

El accuracy obtenido es muy bueno (indicando un modelo de gran calidad y un dataset útil). El accuracy del test es ínfimamente inferior al accuracy del test, lo que también supone un buen indicativo.

Una vez se tiene esto, se pretende mirar una serie de métricas, con el objetivo de evaluar si el modelo identifica bien a los asegurados de alto riesgo (recall de la clase 1) y si los falsos positivos son manejables:

```
Classification report (test):

              precision    recall  f1-score   support

no_high_risk       0.96      0.90      0.93      8806
   high_risk       0.84      0.94      0.89      5178

    accuracy                           0.91     13984
   macro avg       0.90      0.92      0.91     13984
weighted avg       0.92      0.91      0.91     13984


Matriz de confusión:
[[7917  889]
 [ 335 4843]]
```

Como se puede observar en la imagen superior, el modelo es muy bueno tanto identificando a aquellos pacientes de riesgo bajo como a aquellos pacientes de mayor riesgo. La diferencia entre ambos puede deberse a la cantidad de instancias de cada una de las dos métricas que aparecen en el dataset. Es decir, es lógico que si el dataset sobrepondera la clase no_high_risk esto se vea algo reflejado en el modelo de clasificación. De todas formas, el resultado obtenido es muy bueno en todas las medidas, no observándose tendencias preocupantes de falsos positivos ni falsos negativos.

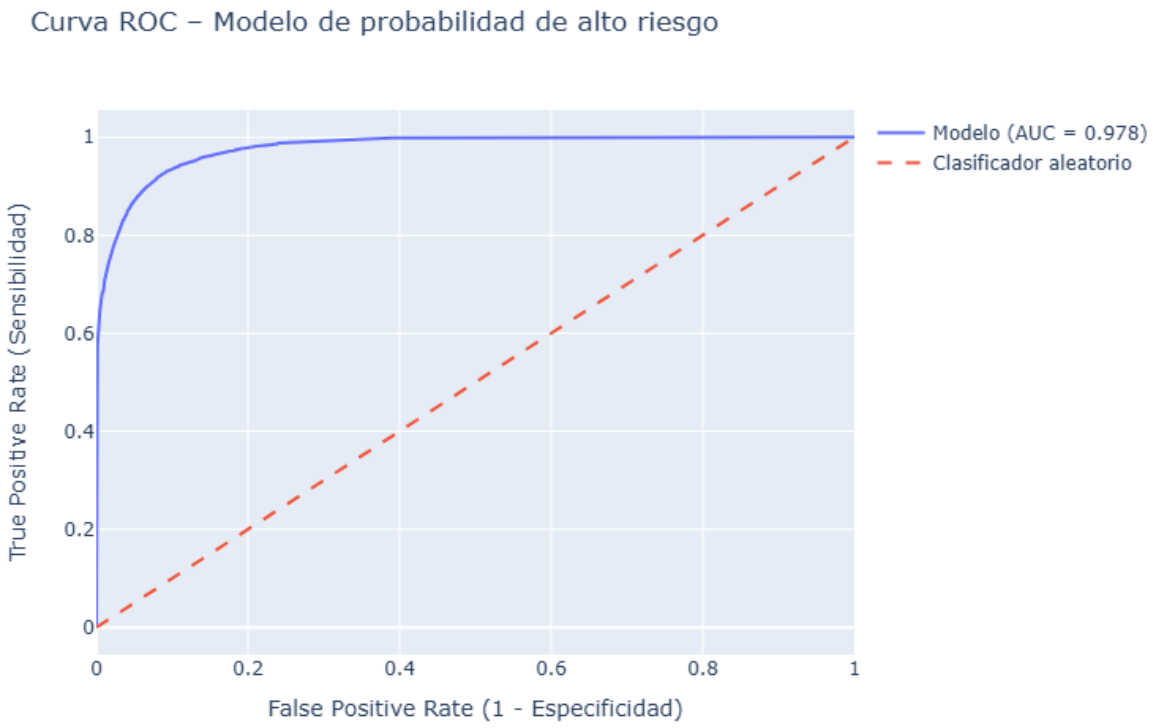
Se realiza la siguiente tabla, que muestra qué variables tienen más peso en la probabilidad de ser alto riesgo (coeficientes positivos) o bajo riesgo (coeficientes negativos), lo que es clave para explicabilidad y entender de verdad las tendencias según las cualidades de cada paciente prospectivo:

	feature	coef	abs_coef
9	smoker_Never	-5.890496	5.890496
8	smoker_Former	-4.034955	4.034955
28	chronic_count	3.977321	3.977321
29	risk_score	3.857590	3.857590
10	alcohol_freq_Occasional	-1.721727	1.721727
11	alcohol_freq_Weekly	-1.638053	1.638053
6	urban_rural_Suburban	-1.370993	1.370993
7	urban_rural_Urban	-1.360051	1.360051
22	plan_type_PPO	-1.167884	1.167884
20	plan_type_HMO	-1.143232	1.143232
4	region_South	-1.005654	1.005654
21	plan_type_POS	-0.849796	0.849796
3	region_North	-0.840677	0.840677
2	region_East	-0.817304	0.817304
5	region_West	-0.760010	0.760010
25	network_tier_Silver	-0.712420	0.712420
13	diabetes_1	0.671262	0.671262
0	sex_Male	-0.670855	0.670855
19	mental_health_1	0.625261	0.625261
24	network_tier_Platinum	0.595630	0.595630

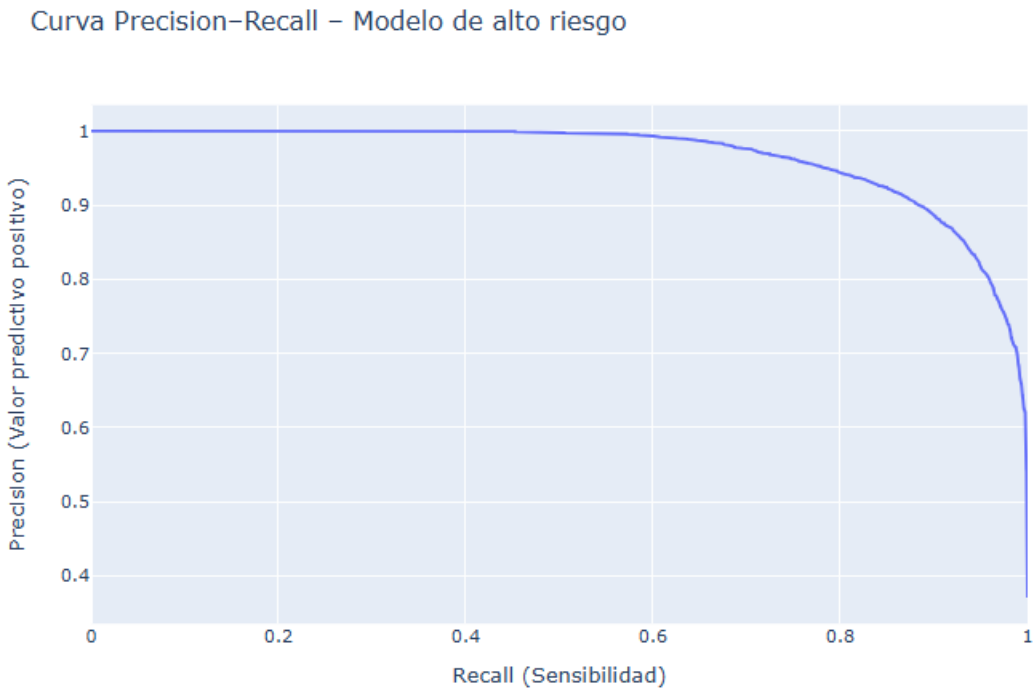
En la tabla superior se puede observar como no ser un ávido fumador tiene el mayor impacto positivo en materia de salud, lo cual, sumado a las gráficas del apartado anterior, dejan clara la clarísima correlación entre aquel que es fumador y aquel que supone un riesgo para las aseguradoras. Aparte de no usar el tabaco, también destacan como drivers positivos no ser beber frecuentemente, también lógico y esperable.

Por otra parte, se observa como tener enfermedades crónicas y ser un factor de riesgo son las cualidades que más impacto negativo tienen en el prospectivo asegurado (es decir, conducen a un posible high_risk).

También se observa un impacto considerable según el tipo de región del individuo (es decir, puede ser justificable cierta discriminación geográfica para la maximización de beneficios) y según el tipo de plan contratado. En este sentido, y dada la variedad de planes y geografías, considero interesante que estos sean los filtros a utilizar en el dashboard, de tal forma que se puedan combinar geografías y planes aseguradores y observar como varían los 4 gráficos mostrados en el dashboard (los más ilustrativos en mi opinión para tener la imagen completa del caso que nos ocupa).

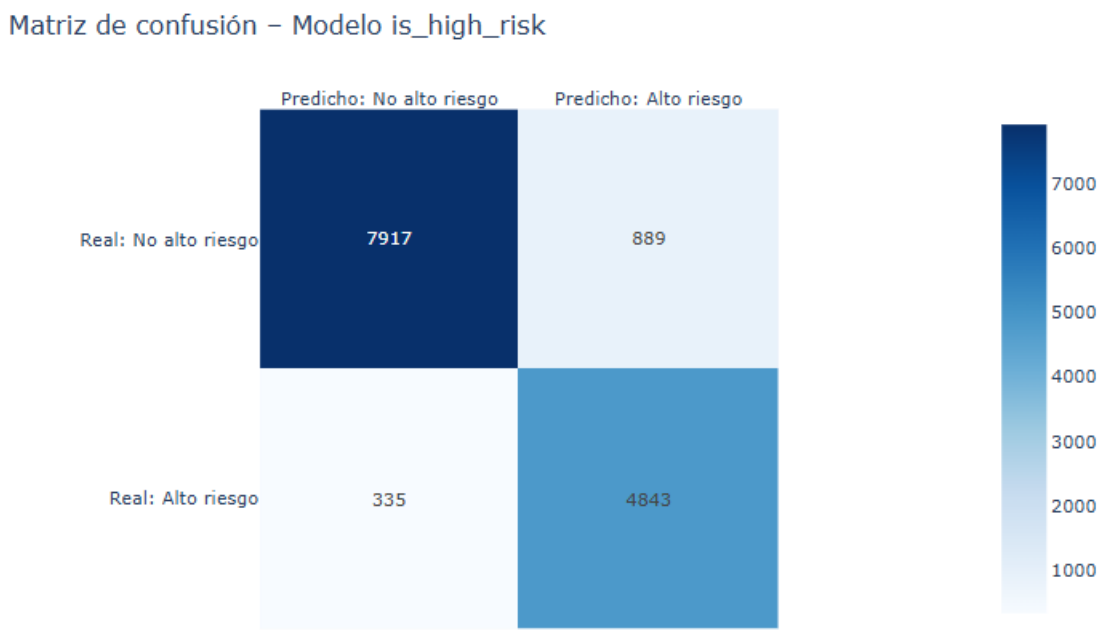


Con el fin de analizar más en detalle la fiabilidad del modelo creado, se pueden realizar más gráficos como la curva ROC mostrada en la imagen superior, que, de nuevo, ratifica la viabilidad del modelo de regresión lineal pues su AUC es muy próximo a 1 y está claramente por encima de la clasificación aleatoria mostrada en la línea roja discontinua.



La curva de precisión y recall, como se podía entrever en la cuantificación previa de estas métricas también supone un resultado cuasi-perfecto.

Y también incluyo la matriz de confusión como heatmap que queda mejor (porque las tendencias se ven más claramente):



Dashboard Completo

Se plantea el dashboard para que se pueda filtrar la población por región y plan y ver cómo cambia el coste anual y el total pagado según el estado de riesgo, ver la relación riesgo-coste y explicar el modelo mediante los coeficientes principales: