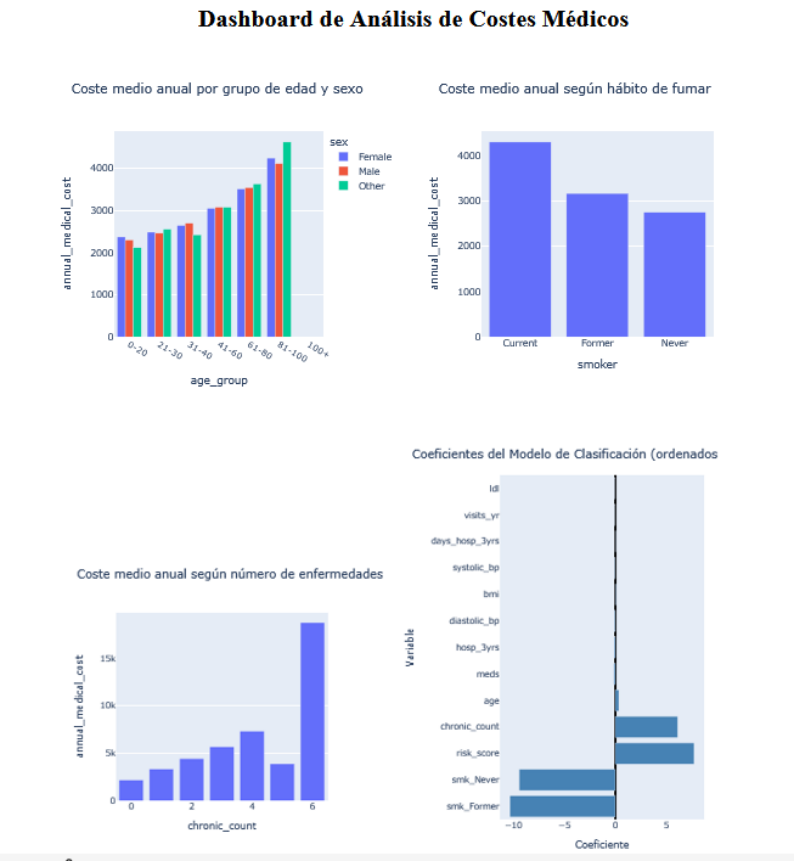
RESUMEN EJECUTIVO



Este estudio analiza los patrones de gasto médico anual de 100.000 asegurados, con el objetivo de identificar factores determinantes del coste, clasificar perfiles de riesgo clínico y apoyar la toma de decisiones de la aseguradora mediante modelos predictivos y segmentación basada en datos.

1. Principales patrones del gasto

El coste médico anual presenta una gran variabilidad, con la mayoría de asegurados situándose entre 1.000 y 6.000 €, pero con una minoría que supera ampliamente los 20.000–40.000 €, generando una concentración del gasto típica en sistemas sanitarios.
El dashboard resume el análisis llevado a cabo e informa sobre las variables más influyentes en el gasto, citadas a continuación:

- Edad: fuerte relación ascendente; mayores de 70 años duplican o triplican el gasto de jóvenes. Sin embargo, se observa que el género es independiente pues el coste medio es muy similar entre hombres y mujeres para cada grupo de edad.
- Hábitos de vida: los fumadores presentan costes significativamente más altos (≈4.300 €) que exfumadores y no fumadores.

- Estado clínico: el número de enfermedades crónicas es uno de los predictores más potentes; pasar de 0 a 3 cronicidades multiplica el coste anual.

Por otro lado, otros factores como el nivel educativo, el género o la región de residencia, son factores que no determinan una influencia clara en el gasto de la aseguradora.

En conjunto, el gasto está impulsado principalmente por carga clínica acumulada y uso hospitalario, más que por factores sociodemográficos.

Clusterización de asegurados

Tras el análisis, se identifican cuatro grupos de asegurados

1. Cluster 1 – Bajo riesgo: jóvenes, sin cronicidades, no fumadores; costes 1.500–2.500 €.
2. Cluster 2 – Riesgo medio: adultos con 1–2 cronicidades; costes 3.500–6.000 €.
3. Cluster 3 – Alto riesgo: mayores con ≥3 cronicidades; costes 7.000–20.000 €.
4. Cluster 4 – Super-high-cost: pacientes con múltiples hospitalizaciones (outliers y patologías severas; costes >20.000 €.

Los clusters 2 y 3 representan los grupos donde intervenciones preventivas o gestión activa pueden generar una caída de costes para la aseguradora.

Modelos predictivos

Clasificación (is_high_risk)
Tras seleccionar variables clínicas clave, la regresión logística obtiene un rendimiento excelente:

- Accuracy del 95%
- Recall del 93% para high-risk

Las variables de mayor influencia son risk_score, chronic_count, edad, BMI y hospitalizaciones. El hábito de no fumar aparece como un fuerte factor protector.
Regresión (annual_medical_cost)
El modelo Random Forest mejora sustancialmente la capacidad predictiva:

- $R^2 \approx 0.67$
- MAE ≈ 980 €

Las variables más importantes son total_claims_paid, avg_claim_amount, risk_score, días hospitalizados y cronicidades.

Conclusiones y recomendaciones

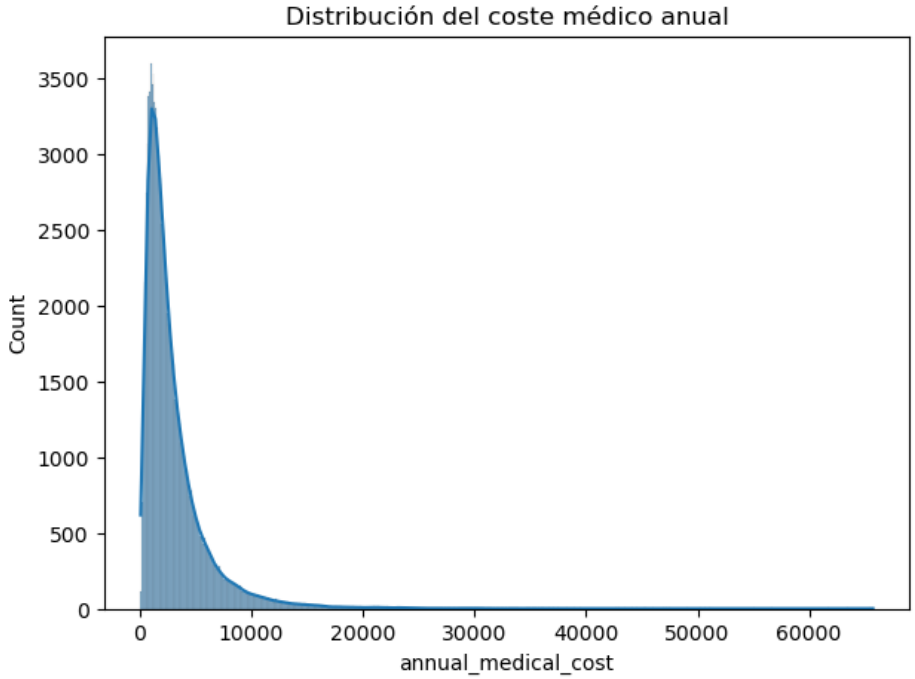
Los resultados muestran que:

- El coste médico anual está altamente concentrado en un pequeño grupo de pacientes de alta complejidad clínica.

- La clusterización permite identificar perfiles de riesgo, especialmente en los grupos intermedios.
- El modelado de clasificación y regresión confirman que el historial clínico y los hábitos (como la calidad de fumador) son características clave para determinar el gasto de la aseguradora.

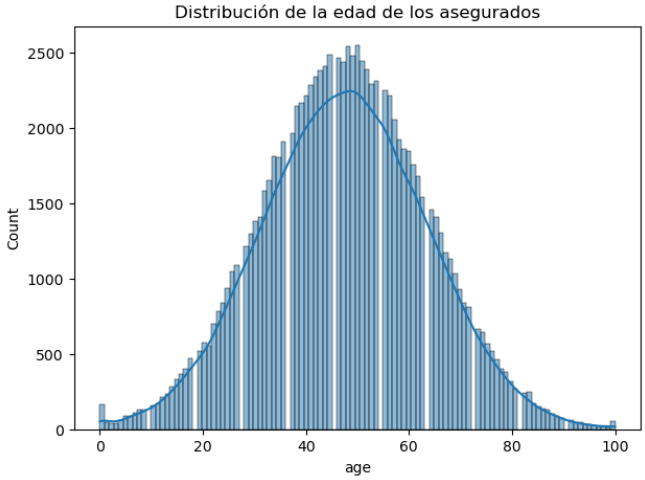
La aseguradora debería centrar sus esfuerzos en los asegurados con 1–3 enfermedades crónicas, implementando programas de seguimiento y prevención específicos. El modelo predictivo permite identificar de forma temprana a los pacientes de alto riesgo, facilitando intervenciones proactivas que reduzcan futuros costes. Además, es recomendable aplicar estrategias dirigidas a fumadores y personas con sobrepeso, así como utilizar el modelo de regresión para mejorar la tarificación y planificación financiera.

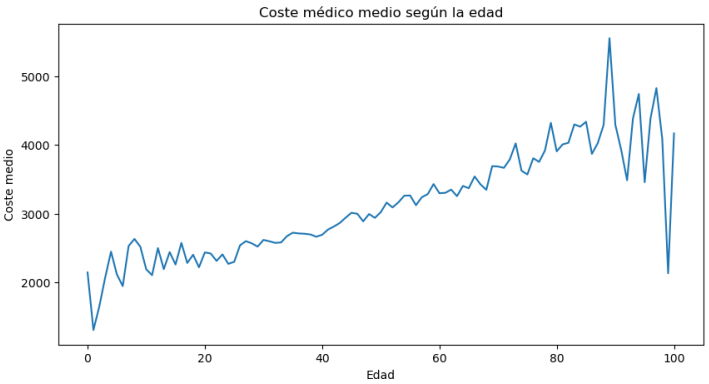
GRÁFICAS DE ANÁLSIS Y CLUSTERIZACIÓN



El histograma muestra una distribución asimétrica hacia la derecha. La mayoría de los asegurados concentran costes anuales entre 500 y 7.000 €, pero existe una cola larga de valores más altos que alcanzan los 60.000 €, asegurados que representan un riesgo para la aseguradora. Esto sugiere que los *high-risk* tienen un impacto desproporcionado. Un pequeño porcentaje de pacientes genera la mayor parte del gasto.

EDAD VS COSTE:





Si observamos la distribución de asegurados en función de su edad, esta sigue una forma muy cercana a una normal, centrada en torno a 45-50 años. Hay asegurados desde recién nacidos hasta mayores de 100.

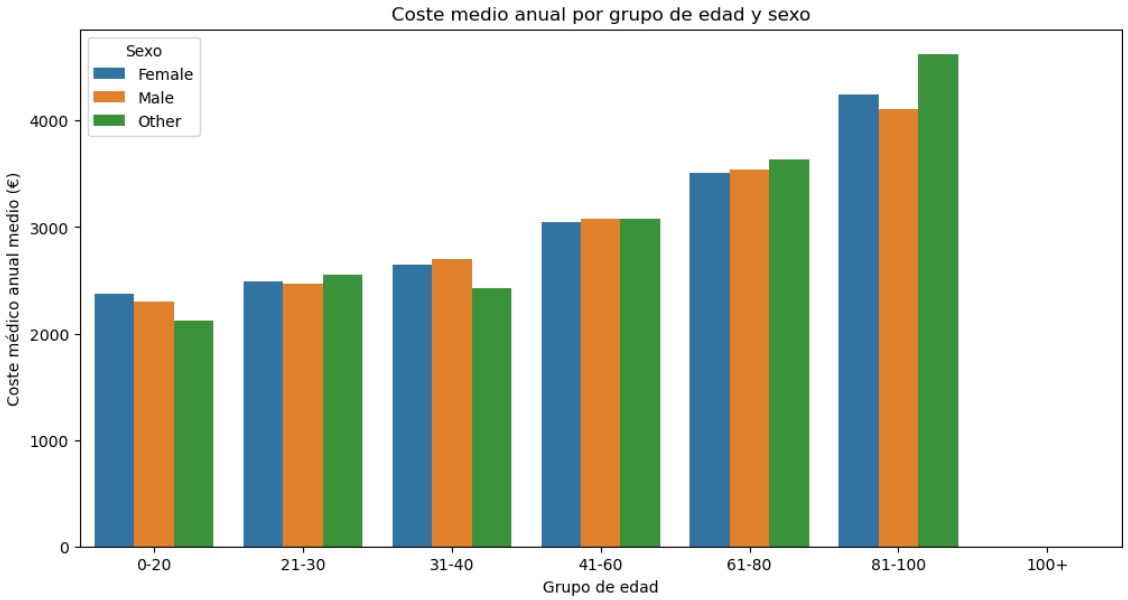
Si traspasamos este estudio al coste, se observa que el coste aumenta claramente con la edad:

La gráfica muestra un incremento progresivo:

- En jóvenes (0-30 años) 1.500–2.500 €
- En adultos medios (40-60) 3.000 €
- En mayores (70-90) 4.000–5.500 € (aunque alta variabilidad)

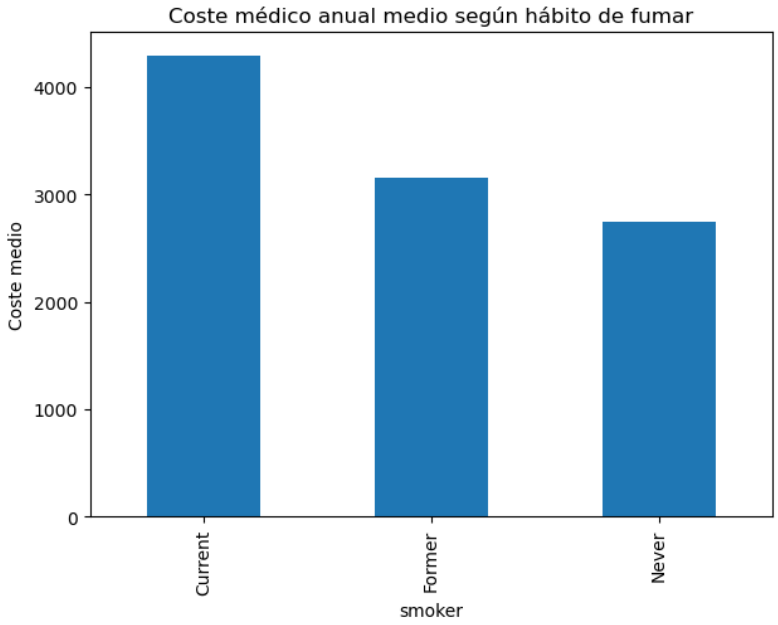
La edad es un predictor clave.

Si añadimos diferencias por géneros, obtenemos lo siguiente:



EN la grafica de barras se observa una tendencia alcista del coste en función de la edad, como se mencionó anteriormente, pero el coste es muy similar entre hombres y mujeres, para cada grupo de edad. Luego se determina que el sexo NO es un factor determinante del gasto.

Coste médico según hábito de fumar

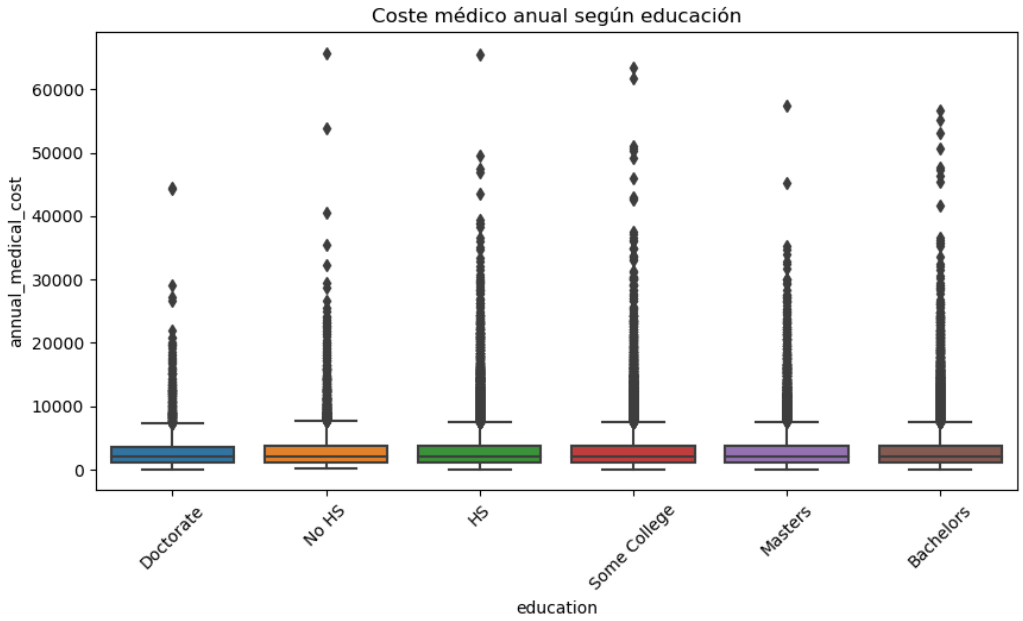


El gráfico de barras muestra un patrón claro:

- Fumadores actuales (Current) : gasto más alto (4.300 €)
- Exfumadores (Former) : gasto intermedio (3.100 €)
- No fumadores (Never): gasto más bajo (2.700 €)

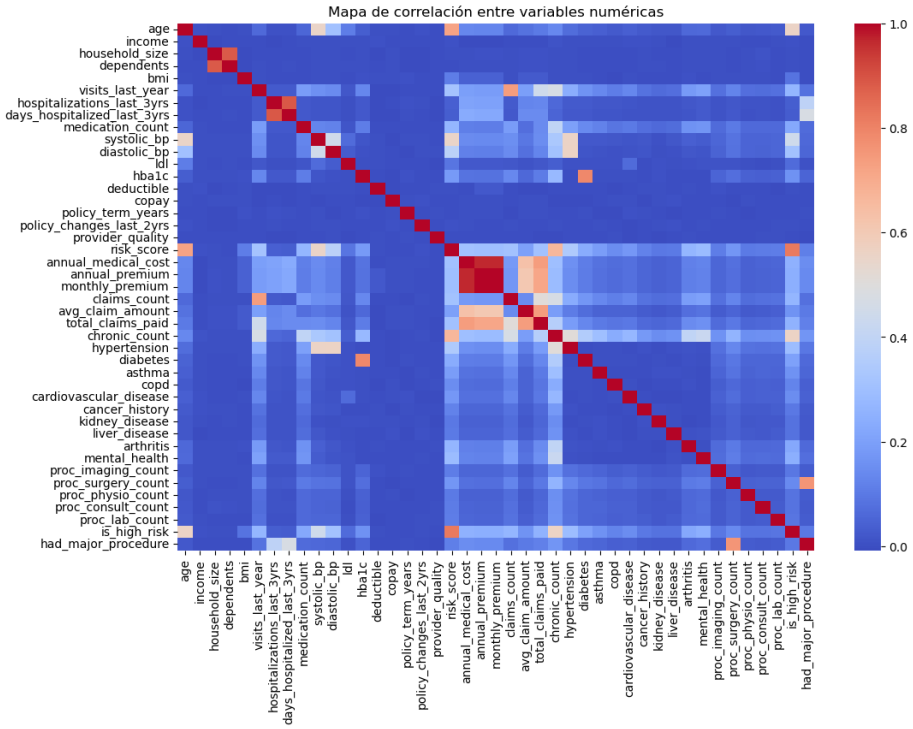
En definitiva, fumar está asociado a aprox. 1.500 € más de gasto anual, lo cual encaja perfectamente con mayor riesgo cardiovascular, pulmonar y metabólico.

Coste médico según educación:



El boxplot muestra que todos los niveles educativos presentan enormes colas superiores (outliers), Aun así, la educación no parece ser determinante en el coste.

Mapa de correlación



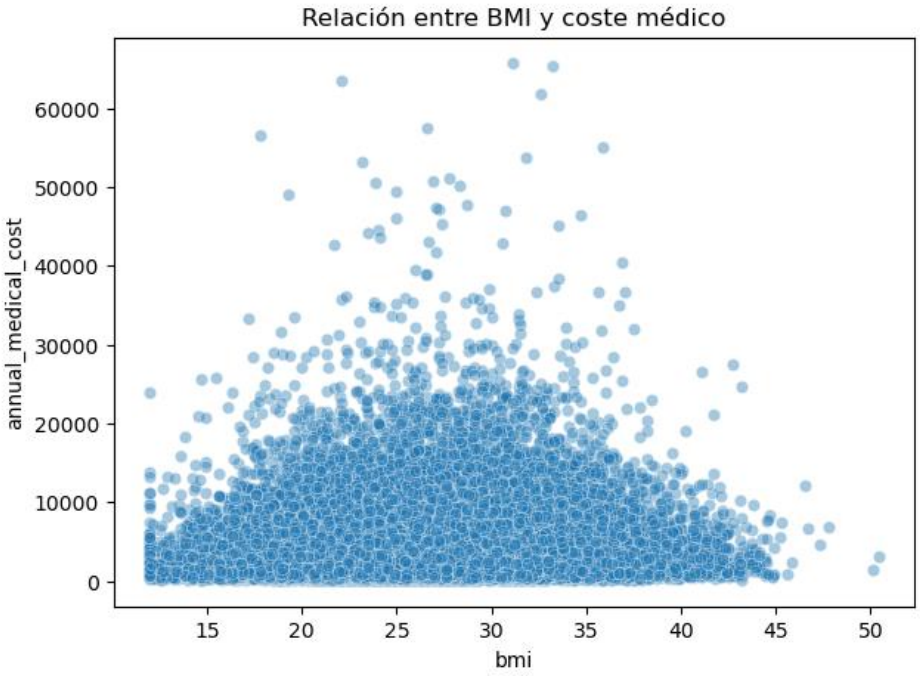
El coste médico anual correlaciona mucho con:

- hospitalizaciones_last_3yrs
- days_hospitalized_last_3yrs
- chronic_count
- risk_score
- total_claims_paid
- claims_count

Variables clínicas (ldl, hba1c, systolic/diastolic BP) tienen correlaciones leves.

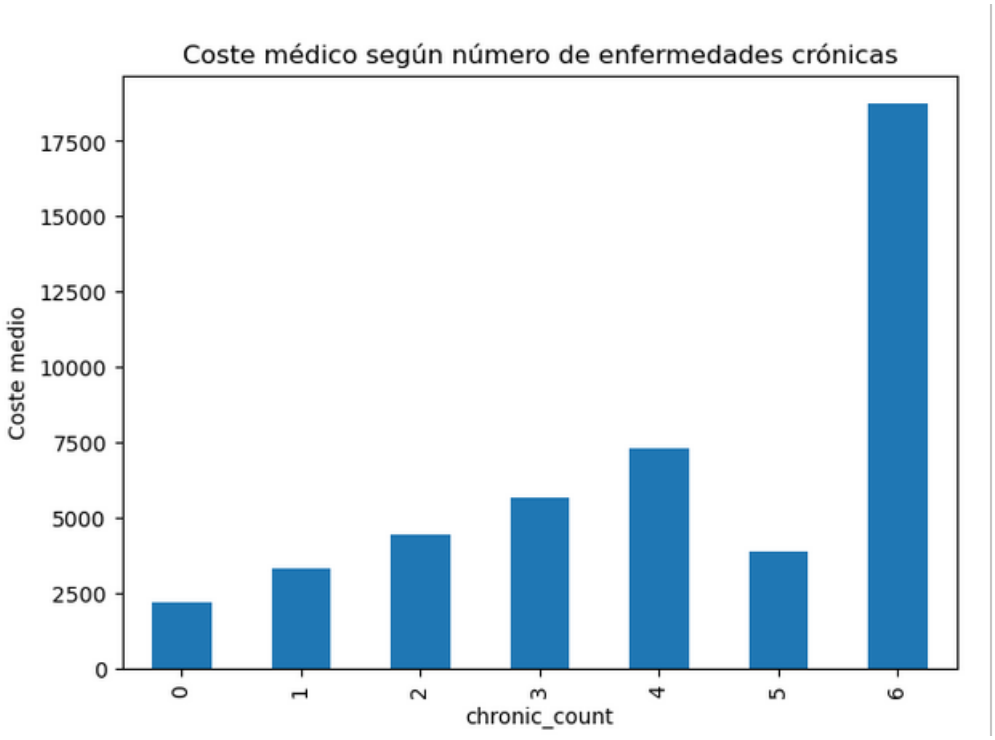
El coste está totalmente ligado a utilización y hábitos hospitalarios pasados, lo cual es esperable, pues previos ingresados son más propensos a acudir de nuevo al servicio hospitalario.

BMI VS COSTE:



EL scatter plot no muestra una relación lineal entre el coste y el bmi pero la nube se ensancha hacia arriba para BMI altos (35+). Sin embargo, la mayoría de costes extremos se sitúan entre 25 y 30 de BMI, que indican sobrepeso, aunque no obesidad.

Enfermedades crónicas:



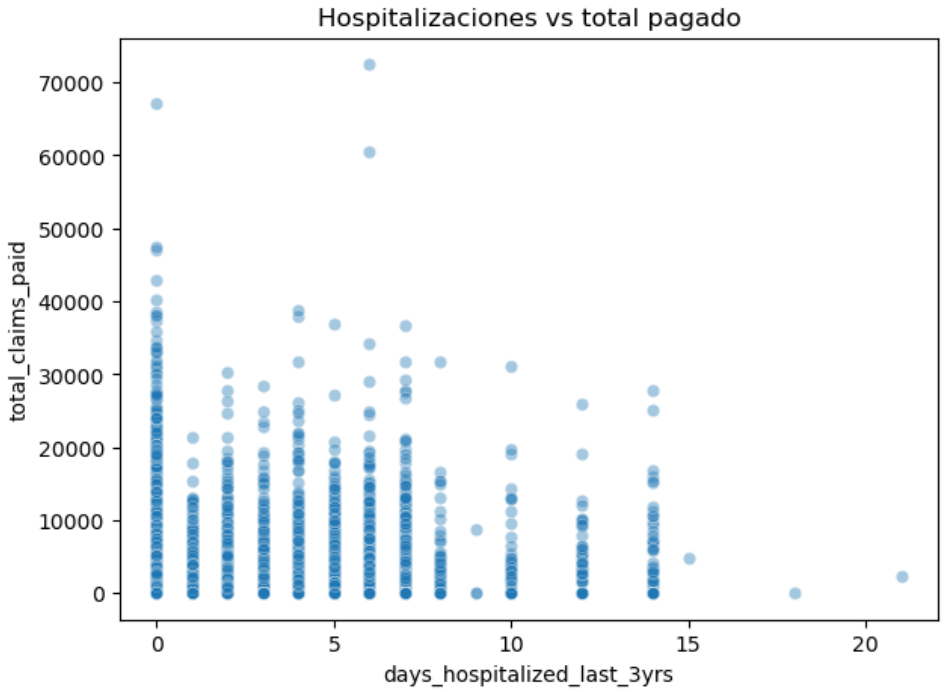
El número de enfermedades crónicas es uno de los mejores predictores:

El gráfico de barras muestra:

- 0 cronicidades: 2.200 €
- 1 cronicidad: 3.400 €
- 2 cronicidades: 4.500 €
- 3 cronicidades: 5.800 €
- 4 cronicidades: 7.300 €
- 6 cronicidades: 18.500 €

El cluster de alta cronicidad será el cluster más caro.

Coste vs días hospitalizado:



Aunque hay individuos con 0–3 días hospitalizados y altos costes, los valores más altos del total pagado (>40.000–70.000 €) aparecen sobre todo en pacientes con:

La aseguradora debe monitorizar especialmente a los pacientes que entran en el circuito hospitalario, porque es en ese punto cuando los costes se disparan. Aun así existen pacientes con pocas hospitalizaciones que generan un gasto muy alto, probablemente debido a atenciones de emergencia.

En función al análisis exploratorio realizado con kmeans, se encuentran los siguientes grupos:

Característica	**Cluster 1** **Bajo riesgo / bajo coste**	**Cluster 2** **Riesgo medio / crónicos moderados**	**Cluster 3** **Alto riesgo / con sobrepeso**	**Cluster 4** **Super-high-cost / outliers**
Edad típica	Jóvenes (<40)	Mediana edad (40–65)	Elevada (>70)	Variable, pero tiende a >70 con complicaciones
Enfermedades crónicas	0	1–2	3 o más	3–6
Hábito de fumar	No fumadores	Algunos exfumadores	Fumadores o exfumadores frecuentes	fumadores
BMI	Normal	Moderadamente alto (sobrepeso)	Alto (obesidad frecuente)	Muy variable, pero sobrepeso/obesidad frecuente

Coste médico anual	1.500–2.500 €	3.500–6.000 €	7.000–20.000 €	>20.000 € (hasta 60.000 €)
Interpretación	Población sana	Pacientes con enfermedades controladas	Pacientes con que usan el sistema con frecuencia	Casos graves, picos de gasto

La aseguradora debe poner foco en el cluster 2 y 3, siendo estos claros casos de “sobregasto” comparado a la media obtenida en el primer gráfico. EL cluster 4 sería interesante definirlo pero, siendo outliers, es muy difícil identificarlo.

CLASIFICACIÓN HIGH RISK:

Tras dividir el conjunto de datos entre train y test, y entrenarlos con un modelo de clasificación sobre la variable target is “high risk” se obtienen los siguientes resultados:

```
### CLASSIFICATION REPORT ###
              precision    recall  f1-score   support

           0       0.79      0.86      0.82     12599
           1       0.72      0.62      0.67      7401

    accuracy                           0.77     20000
   macro avg       0.76      0.74      0.75     20000
weighted avg       0.77      0.77      0.77     20000
```

Aun así, si reducimos el numero de variables a únicamente las más influyentes (SOBRE TODO CLÍNICAS Y DE ESTILO DE VIDA) obtenemos un modelado simple, efectivo sin riesgo de overfitting, y aplicado a nuestras hipótesis de clusterizado.

Obtenemos los siguientes resultados:

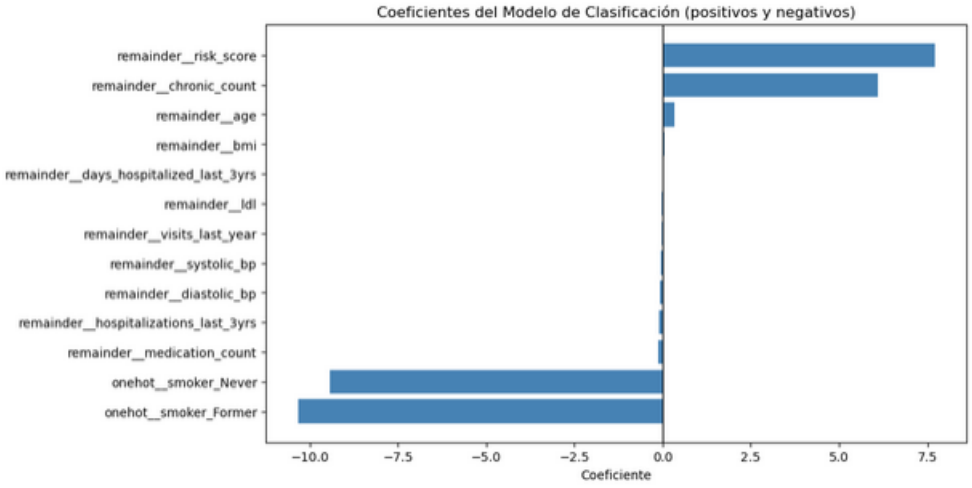
```
### CLASSIFICATION REPORT ###
              precision    recall  f1-score   support

           0       0.96      0.96      0.96     12599
           1       0.93      0.93      0.93      7401

    accuracy                           0.95     20000
   macro avg       0.95      0.95      0.95     20000
weighted avg       0.95      0.95      0.95     20000

### CONFUSION MATRIX ###
[[12103   496]
 [  506  6895]]

### TOP FEATURES ###
                                 feature       coef
0                   onehot__smoker_Former -10.332353
1                    onehot__smoker_Never  -9.435337
8                   remainder__risk_score   7.720616
4                remainder__chronic_count   6.104275
2                        remainder__age     0.338366
12            remainder__medication_count  -0.130522
10   remainder__hospitalizations_last_3yrs  -0.107543
6                 remainder__diastolic_bp  -0.079905
3                        remainder__bmi     0.047819
5                  remainder__systolic_bp  -0.047233
11  remainder__days_hospitalized_last_3yrs   0.025866
9             remainder__visits_last_year  -0.018112
7                        remainder__ldl    -0.014361
```



El modelo simplificado funciona muy bien, alcanzando una precisión global del 95% y, lo más importante, un recall del 93% para identificar pacientes de alto riesgo, lo cual es excelente para una aseguradora. Las variables más influyentes coinciden con el sentido clínico del riesgo: *no fumar* y tener *pocos factores crónicos* reduce fuertemente la probabilidad de ser high-risk, mientras que un mayor *risk_score*, más *enfermedades crónicas*, *edad avanzada*, *mayor medicación* y *hospitalizaciones recientes* aumentan claramente el riesgo. En resumen, el modelo identifica de forma eficiente que el riesgo clínico está dominado por factores crónicos y uso del sistema sanitario, y demuestra que nuestro clusterizado inicial es acertado, donde los grupos 2, 3 y 4 suonen mayor riesgo y suben el gasto de la aseguradora considerablemente

MODELO DE REGRESIÓN PARA PREDECIR EL ANNUAL MEDICAL COST:

Al igual que hicimos con la clasificación, se filtra el dataset con aquellas variables más determinantes. Se obtienen los siguientes resultados tras el modelo de regresión lineal:

```
MAE: 1159.2392638771444
RMSE: 1986.04149577918
R2 Score: 0.599156368128565

### TOP 20 COEFICIENTES MÁS INFLUYENTES ###
                                  feature        coef
1                    onehot__smoker_Never -721.755128
0                   onehot__smoker_Former -512.000022
4                remainder__chronic_count  500.459744
12                remainder__claims_count -418.928550
7   remainder__days_hospitalized_last_3yrs  236.567988
6    remainder__hospitalizations_last_3yrs   78.667180
5             remainder__visits_last_year  -34.713320
3                          remainder__bmi   15.235882
10          remainder__proc_surgery_count  -11.393450
2                          remainder__age    9.365809
8             remainder__medication_count   -8.971718
9                   remainder__risk_score    4.165070
14           remainder__total_claims_paid    1.043817
11          remainder__proc_imaging_count    0.295193
13            remainder__avg_claim_amount    0.174868
```

Se obtiene un error y un r2 no demasiado eficaces, lo que sugiere que la relación entre las variables y el gasto médico no es lineal.

Se prueba un random forest para identificar relaciones no lineales de los datos:

```
MAE: 980.3416821567658
RMSE: 1806.6481163804572
R2 Score: 0.6682999804861192

### FEATURE IMPORTANCES ###
                                  feature  importance
14           remainder__total_claims_paid    0.781120
13            remainder__avg_claim_amount    0.074154
9                   remainder__risk_score    0.036138
7   remainder__days_hospitalized_last_3yrs    0.024865
3                          remainder__bmi    0.023321
2                          remainder__age    0.016475
4               remainder__chronic_count    0.008154
5             remainder__visits_last_year    0.007546
12                remainder__claims_count    0.007432
8             remainder__medication_count    0.006381
11          remainder__proc_imaging_count    0.005215
1                    onehot__smoker_Never    0.003717
10          remainder__proc_surgery_count    0.002600
0                   onehot__smoker_Former    0.001472
6    remainder__hospitalizations_last_3yrs    0.001411
```

El modelo de Random Forest mejora ligeramente la predicción del gasto médico, alcanzando un R2 de 0.67 y reduciendo el error promedio (MAE) a unos 980 €, lo cual es razonable dado que el coste sanitario presenta varaibilidad tal y cómo se observó en las gráficas. Las variables más importantes confirman lo observado en el análisis exploratorio: el factor que más determina el gasto es total_claims_paid, seguido por avg_claim_amount y el risk_score, lo que refleja que el

coste anual está fuertemente ligado al historial de reclamaciones y al riesgo clínico acumulado. También destacan como relevantes los días hospitalizados, el BMI, la edad, las enfermedades crónicas y las visitas médicas anuales, todos consistentes con patrones clínicos y de utilización. En conjunto, el modelo identifica claramente que el gasto anual depende sobre todo del uso previo del sistema sanitario y de la complejidad clínica del paciente, proporcionándole a la aseguradora una base sólida para segmentar coberturas hacia determinados grupos de pacientes.