Nombre y Apellidos: Juan Pardo de Santayana Navarro
Github con notebook:

1. Resumen Ejecutivo

Introducción

El presente análisis se ha realizado con el objetivo de comprender los factores que determinan el coste médico anual, el uso de servicios sanitarios y el riesgo clínico entre aproximadamente 100.000 asegurados de una gran compañía de seguros de salud. El estudio integra análisis exploratorio detallado, técnicas de agrupación (clustering), un modelo predictivo de regresión y la creación de un dashboard interactivo para facilitar la toma de decisiones por parte de los equipos de negocio y gestión del riesgo.
Los resultados permiten identificar patrones de comportamiento sanitario, segmentos de pacientes con necesidades diferenciadas y variables críticas que explican la mayor parte de la variabilidad del coste médico anual. Esto proporciona una base sólida para optimizar pólizas, programas de prevención y estrategias de pricing basadas en datos.

Descripción del dataset y preparación de los datos

El conjunto de datos propuesto distintos tipos de variables: demográficas y socioeconómicas, hábitos de vida, salud y clínicas, procedimientos médicos, seguro médico y póliza, y costes y reclamaciones. El objetivo es ser capaz de utilizar todos estos datos para predecir y explicar el coste anual del seguro médico. Cabe destacar que dentro del conjunto de datos solamente la variable que refleja el consumo de alcohol presentaba valores nulos, que han sido sustituidos por una frecuencia de cero puesto que se supone que su consumo es nulo.

Análisis Exploratorio de Datos

Dentro de las visualizaciones, una de las más relevantes ha sido la matriz de correlación. Gracias a ella, se ha encontrado que existen variables muy correlacionadas con la variable objetivo. Entre ellas, cabe destacar: "monthly_premium", "annual_premium", "total_claims_paid", "claims_count", "avg_claim_amount". Esto era de esperar, puesto que todas ellas tienen que ver tanto con el pago mensual o anual como con el número de reclamaciones que se hacen al seguro. Es por tanto que se decide eliminar estas variables de nuestro análisis, puesto que el objetivo del trabajo es ser capaces de explicar el coste médico a través de factores de hábitos saludables y otras más.
El análisis inicial reveló varios patrones significativos:
El coste médico aumenta progresivamente con la edad, con incrementos especialmente pronunciados a partir de los 55–60 años, que sugiere un impacto creciente de las enfermedades crónicas y de la utilización hospitalaria en los asegurados de mayor edad.
El estilo de vida presenta una fuerte influencia en los costes. En particular, los asegurados con BMI elevado, fumadores o con consumo frecuente de alcohol muestran un coste medio anual superior al resto del colectivo. Las visualizaciones refuerzan la existencia de una relación positiva entre estos hábitos y el coste medio.
Las enfermedades crónicas son uno de los factores más determinantes. Para ello, se ha creado una variable nueva llamada “chronic_count” que muestra cuántas variables crónicas tiene un paciente. Los pacientes con tres o más condiciones presentan costes médicos mucho más altos, además de un mayor número de hospitalizaciones y utilización de medicación.

Modelo de predicción

Se ha entrenado un modelo predictivo utilizando un pipeline con estandarización y codificación categórica para estimar el coste médico anual. Tras probar algunos algoritmos distintos, se decide continuar con una regresión lineal simple, ya que era el modelo con mejor rendimiento y además es el modelo más fácil de interpretar. La parte más importante del modelo es encontrar las variables más significativas para predecir y en qué dirección aportan (positiva o negativamente, es decir, si aumentan o reducen el precio del seguro). Se encontró que las 5 variables más significativas fueron las siguientes (positivas salvo que se indique lo contrario: si el paciente es fumador actual, si el paciente nunca ha fumado (negativo), el número de hospitalizaciones en los últimos tres años, la edad y el número de variables crónicas (variable creada).

Los coeficientes del modelo de predicción y su importancia muestran que lo que se venía viendo en el análisis inicial, que el tabaquismo, la edad y las enfermedades crónicas afectan gravemente al coste medio, además del número de hospitalizaciones, como era de esperar. En conjunto, el modelo muestra un rendimiento predictivo decente y una estructura coherente con lo esperado.

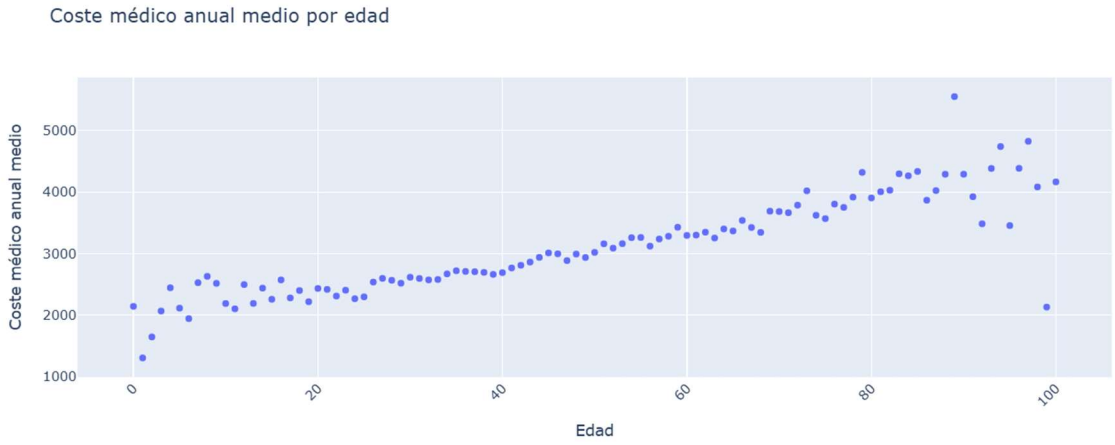
Dashboard seleccionado (no interactivo, 4 subplots más importantes)

Recomendaciones para la aseguradora y conclusión

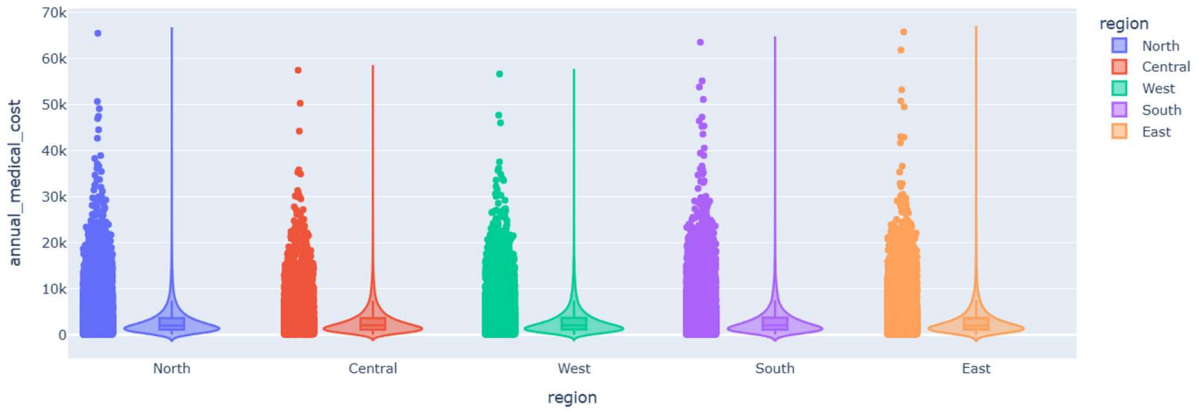
A partir del análisis cuantitativo, se recomiendan tres líneas de acción: programas de prevención dirigidos, seguimiento intensivo de pacientes complejos y optimización de productos y pricing. Con estas tres líneas, se conseguirán fomentar estilos de vida saludables en grupos con malos hábitos, control de pacientes con múltiples condiciones crónicas y episodios hospitalarios recurrentes para mejorar resultados clínicos y ajustar pólizas y primas según riesgo clínico real

El análisis demuestra que los costes médicos están fuertemente ligados a factores clínicos y de utilización sanitaria, mientras que los factores demográficos y de estilo de vida actúan como moduladores del riesgo. El modelo predictivo y el dashboard interactivo proporcionan una plataforma sólida para la toma de decisiones basada en datos, permitiendo mejorar la eficiencia de la aseguradora, optimizar la gestión del riesgo y ofrecer productos más ajustados a las necesidades reales de los asegurados.

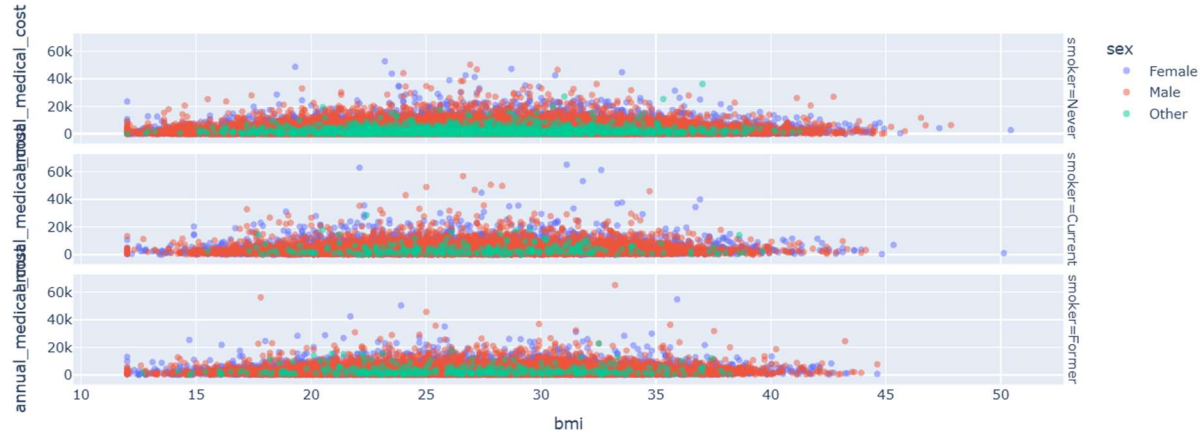
2. Gráficas del análisis exploratorio y breve explicación de cada una



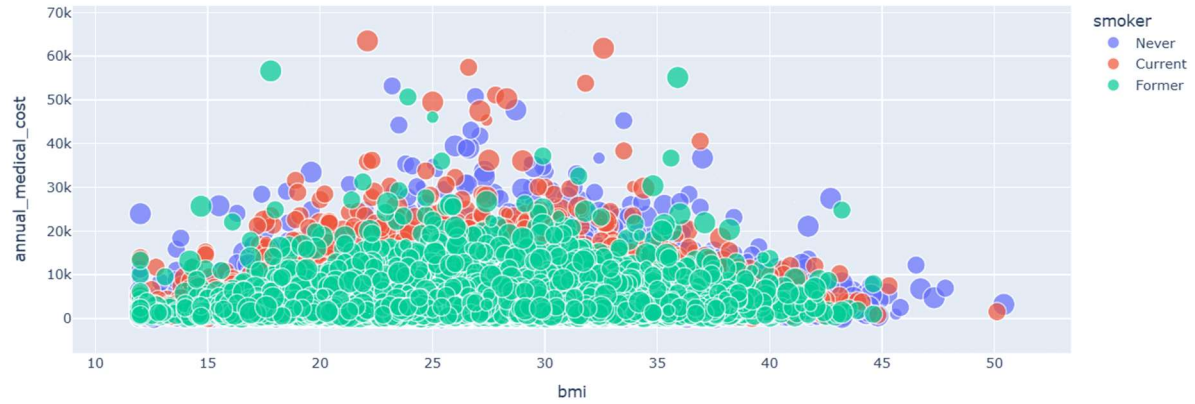
La primera gráfica simplemente muestra la relación lineal entre el coste anual medio y la edad de los pacientes, que demuestra que, a más edad, mayor coste médico.



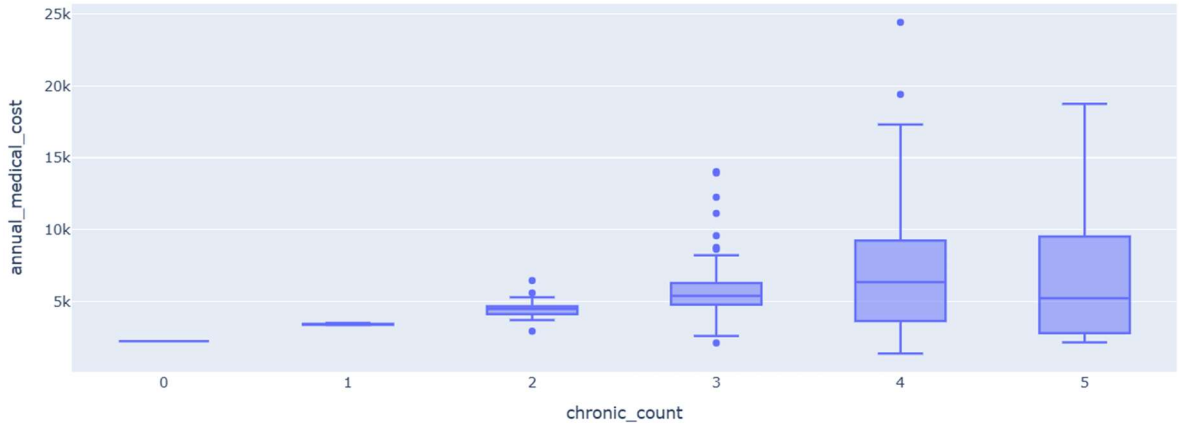
La segunda gráfica muestra la distribución del coste anual medio según las diferentes regiones. El gráfico muestra que todas las regiones presentan una alta variabilidad del coste médico anual, con colas largas que indican presencia de casos de gasto extremo. Aunque las medianas son similares, regiones como West y Central muestran valores máximos más bajos, que sugieren diferencias regionales.



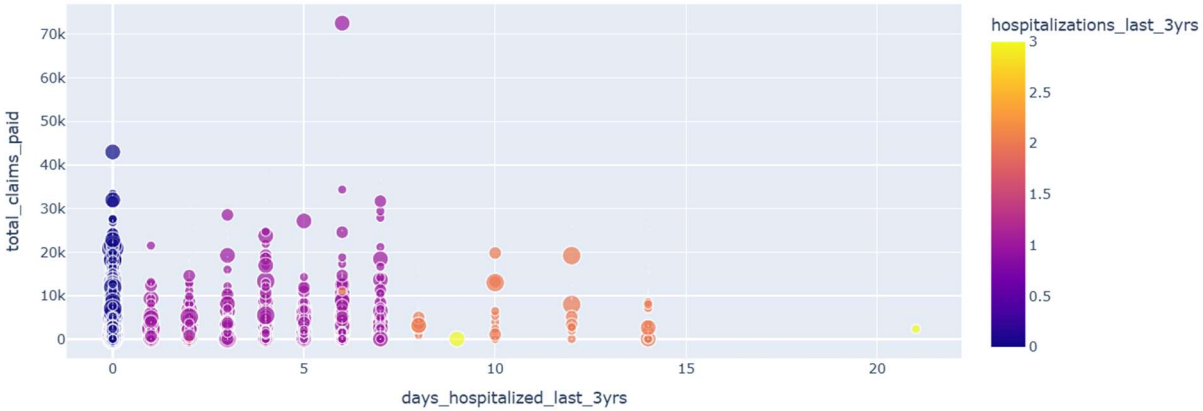
El tercer gráfico representa la relación entre BMI y coste, que muestra un patrón ligeramente creciente, especialmente entre fumadores actuales, como se veía anteriormente. La separación por tipo de fumador deja claro que los fumadores acumulan más casos de gasto elevado. La distribución por sexo es amplia, pero sin diferencias muy marcadas dentro de cada grupo de fumadores.



El cuarto gráfico es muy similar al anterior, muestra que los pacientes con BMI más alto y fumadores actuales presentan los costes médicos más elevados. La diferencia es que ahora se incluye en el tamaño de los círculos la frecuencia alcohólica. Gracias a ello se ve que los círculos más grandes (más consumo), tienen precios más elevados.



El quinto gráfico muestra la relación clara y creciente entre el número de condiciones crónicas y el coste anual. Para esta gráfica, se ha creado la variable chronic_count. A partir de 3 enfermedades crónicas, los costes aumentan de forma muy significativa y presentan mayor variabilidad.

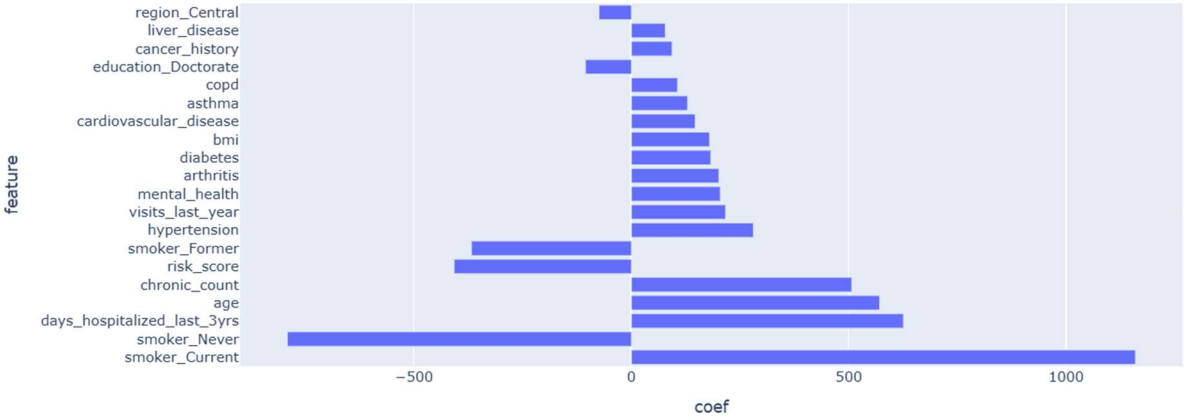


El sexto y último gráfico muestra que el coste total pagado aumenta con el número de días hospitalizados en los últimos tres años. Los asegurados con más episodios de hospitalización y mayor duración generan gastos mucho más altos.

3. Modelo predictivo explicado y con tablas

El objetivo del modelo predictivo es estimar el coste médico anual (annual_medical_cost) de cada asegurado, utilizando información demográfica, hábitos de vida, variables clínicas, utilización de servicios sanitarios y características del seguro. Este enfoque permite a la aseguradora anticipar el gasto esperado, segmentar pacientes por niveles de riesgo económico y ajustar pólizas o programas preventivos de manera más precisa.
Se ha elegido un modelo de regresión lineal con preprocesamiento, construido con un pipeline compuesto por dos etapas: preprocesamiento de variables y predicción. Dentro del preprocesamiento, se ha llevado a cabo una estandarización de variables numéricas mediante StandardScaler y una codificación OneHot para todas las variables categóricas. Además, se han eliminado el identificador person_id y todas las variables que influyen en el precio, de las que se ha hablado previamente.
El modelo ajusta una combinación lineal ponderada de todas las variables transformadas y permite una interpretación directa de los coeficientes.

Evaluación del modelo
El dataset se separó en entrenamiento y prueba (75-25%). Tras ajustar el modelo, se obtuvieron las siguientes métricas:
Rmse, R2, R2 ajustado = (2877.3810930296804, 0.17575352997636062, 0.1741678688581234)
Como podemos observar, el modelo es bastante deficiente, pero sirve para encontrar las variables más significativas:



Conclusión del modelo

La regresión lineal proporciona una visión clara, interpretable y cuantitativa de los factores que más influyen en el coste médico anual. Aunque otros modelos presentan una precisión mayor, la regresión lineal permite transparentar las relaciones entre variables y justificar decisiones de negocio

4. Dashboard interactivo con Dash

Además de lo requerido, se ha diseñado una aplicación en Dash, con 4 visualizaciones diferentes y tres callbacks incluidos, además de un flitro para ver si el paciente es de alto riesgo. Dado que no era obligatorio para el examen, simplemente se incluye una captura de la aplicación sencilla.