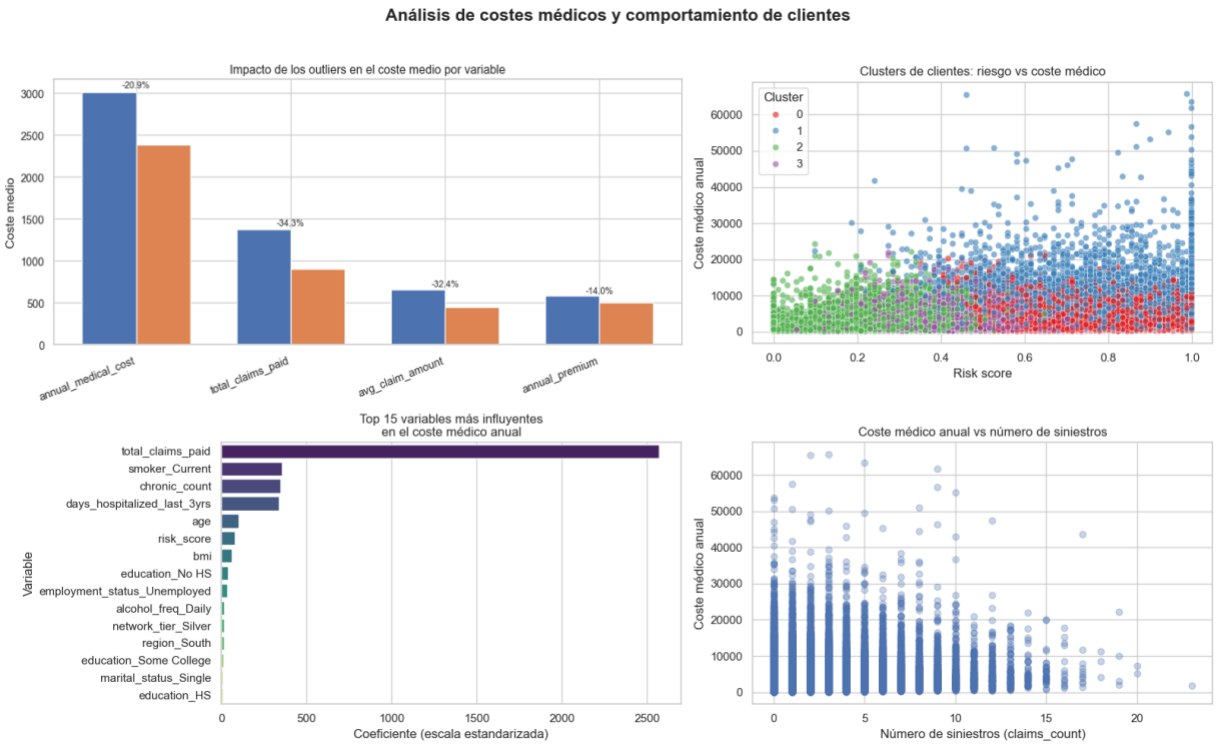
Nombre y Apellidos: Álvaro González Tabernero
Github con notebook:
Nota: Por favor, seguir esta estructura para el documento

1. Resumen Ejecutivo

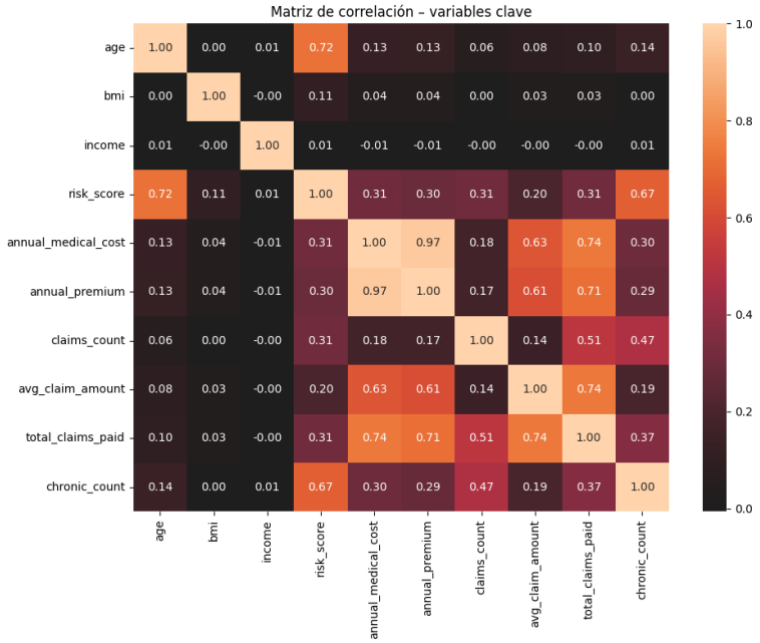


Este caso de estudio se enfoca en los costes en los que incurre una aseguradora médica. El *dataset* del que disponemos para analizar contiene 100.000 filas y más de 54 variables. Durante el desarrollo de este examen, el objetivo va a ser demostrar a través del *dashboard* expuesto, cómo podemos ayudar a la empresa aseguradora a reducir sus costes medios por cliente, o intentar maximizar los retornos de cada cliente atacando áreas descuidadas por la directiva, donde todavía hay margen de beneficio.

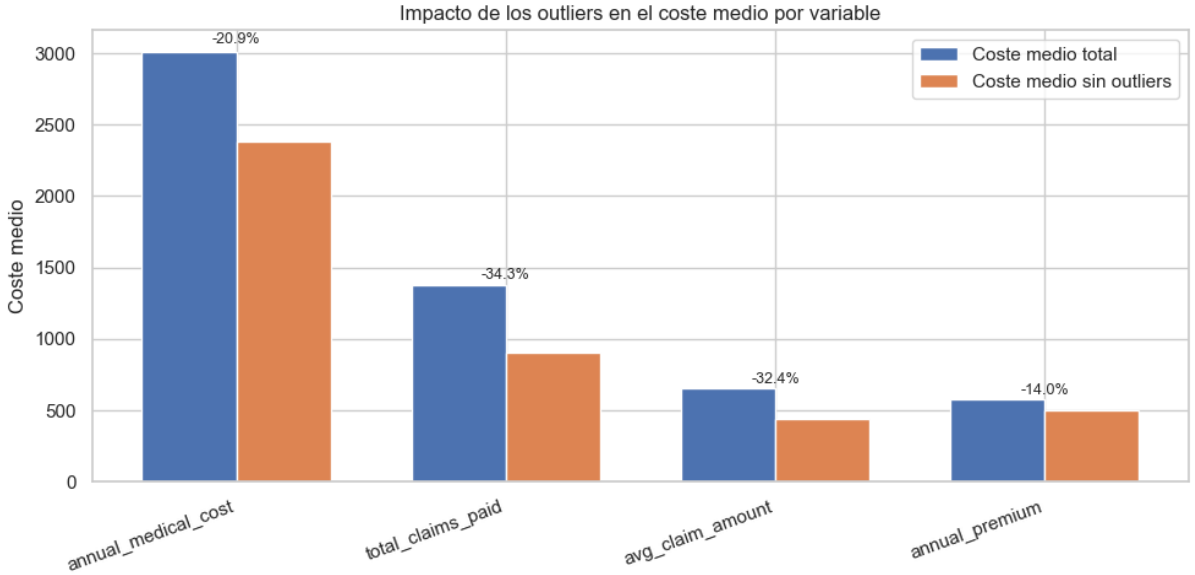
Empezando por la identificación de los atípicos, sabemos que hay un 6.7% de datos con un coste médico extraordinariamente alto en el conjunto de datos que se nos proporciona. Por tanto, hay que buscar una manera de identificar a ese 6.7% que puede generar un impacto de una reducción en los costes médicos anuales de alrededor de un 21%.

Con esto, recomiendo a la directiva que identifique a los clientes que sean, principalmente, altamente reincidentes en los partes, de todos los riesgos, fumadores, que tengan algún componente crónico en su tratamiento y hayan estado hospitalizados en los últimos 3 años. Este perfil de gente, a grosso modo, es el que compone aquel 6.7% de los outliers que generan un 21% más de costes a la empresa, que se podía ahorrar.

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



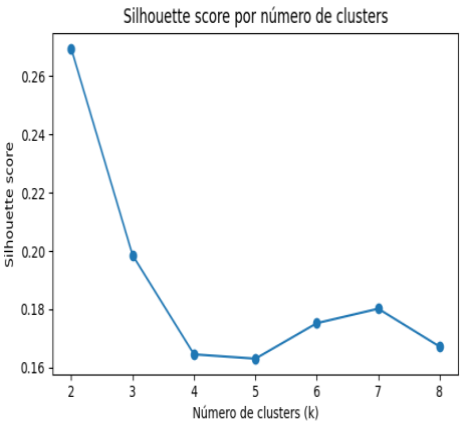
Esta visualización representa un mapa de calor entre las variables más explicativas del caso de estudio. Sorprendentemente, el IMC está poco correlado con la puntuación de riesgo que se asigna a los clientes, al igual que con el número de reclamaciones y el coste medio anual. Sin embargo, sorprende la alta correlación que hay entre la puntuación de riesgo y la edad del cliente. Mirando también el 0.74 que resulta en la correlación entre el número total de claims y el precio medio de las claims. Lo que sugiere esto que

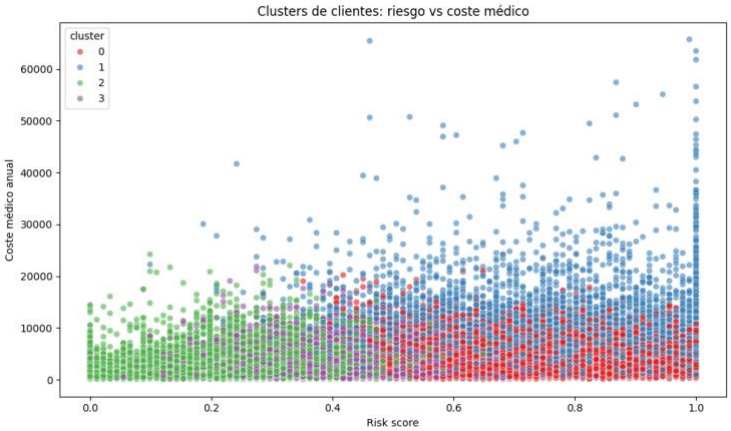


A raíz de esta visualización, podemos observar cómo son los costes medios en los que incurre la aseguradora anualmente a día de hoy, y cómo serían si hubiera una gestión eficiente de los outliers dentro de cada tipo de cliente. Vemos una posible reducción masiva del coste médico medio anual, al igual que del número total de las reclamaciones pagadas, viéndose reducido también la cuantía media que suponían estas reclamaciones. Todo esto, perdiendo

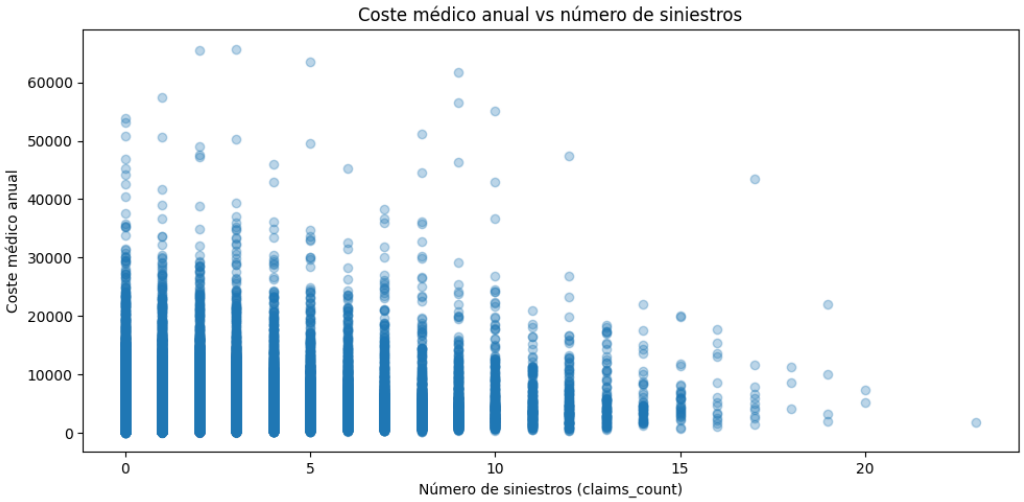
muy pocos ingresos por parte de las primas “premium” pagadas por estos clientes. Además, podemos ver cuánto impacto tiene la eliminación de los outliers a nivel porcentual.

Para identificar estos clientes se ha hecho un algoritmo de clustering, que muestra los siguientes tipos de clientes:

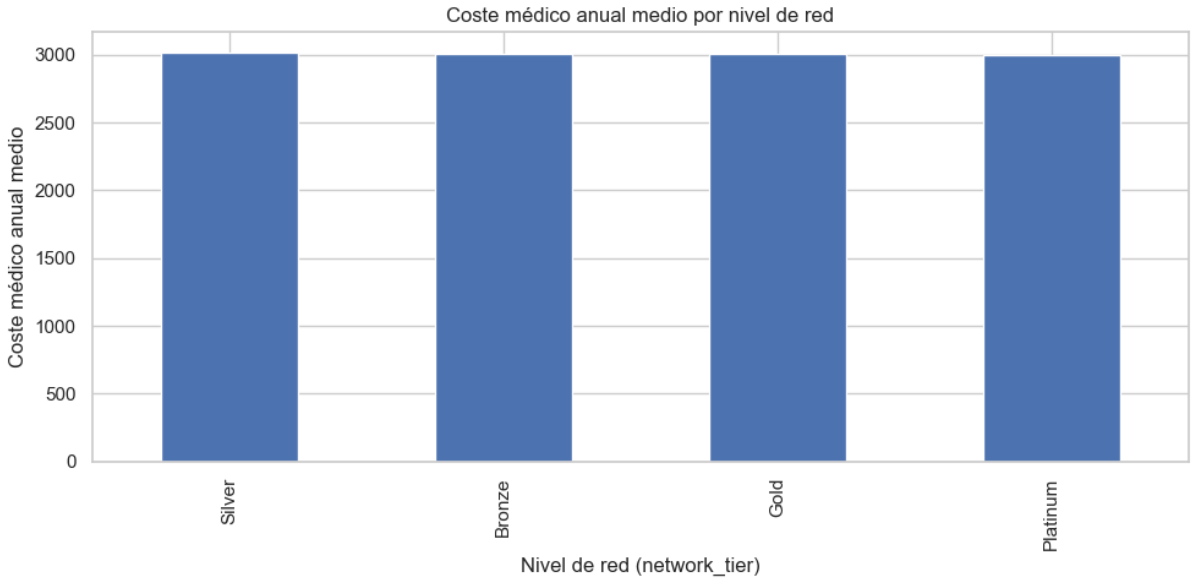




El número óptimo de clusters que resultaba tras el análisis era de 4 (gráfico a la izquierda), que podemos ver claramente: un grupo (verde) que tiene relativamente poca puntuación de riesgo y poco coste, un segundo grupo (rojo) que tiene definitivamente mayor puntuación de riesgo, pero con los costes controlados, un tercer grupo (morado) que se encuentra entre el verde y el rojo en términos de puntuación de riesgo y mantiene el nivel de coste, para finalmente llegar al azul, que agrupa, básicamente a los que más gastan.



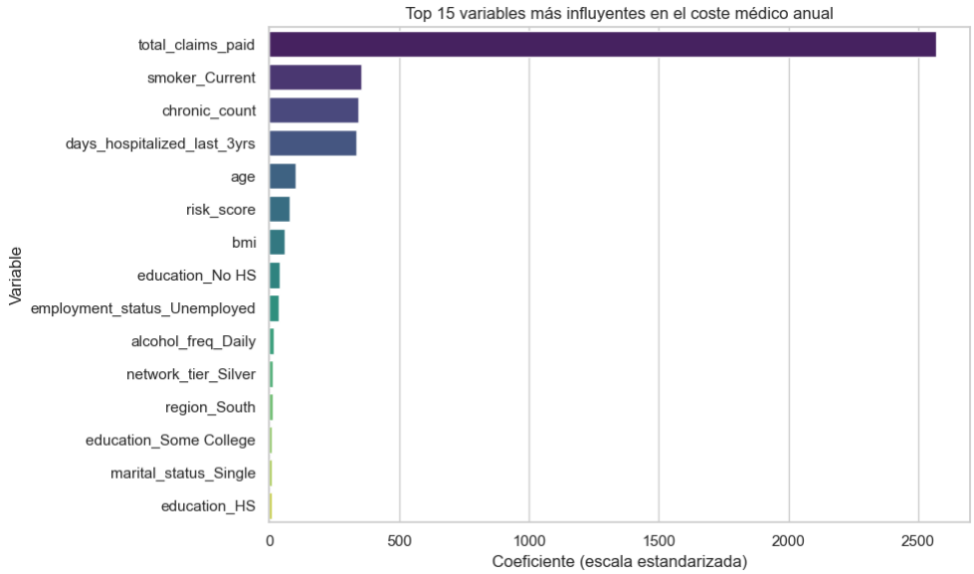
Este scatter plot representa el coste médico anual en el que incurre la aseguradora con respecto al número de siniestros. Con este gráfico podemos saber que la inmensa mayoría de las reclamaciones en las que se gasta dinero la aseguradora son de bajos montantes, y realmente son muy pocos los gastos extraordinarios. Sí que es verdad que hay varios que tienen entre 5 y 10 en la cuenta con gastos muy altos. Esto sugiere que la seguradora podría cubrir intervenciones quirúrgicas más costosas o tratamientos crónicos (quimioterapia, radioterapia y demás).



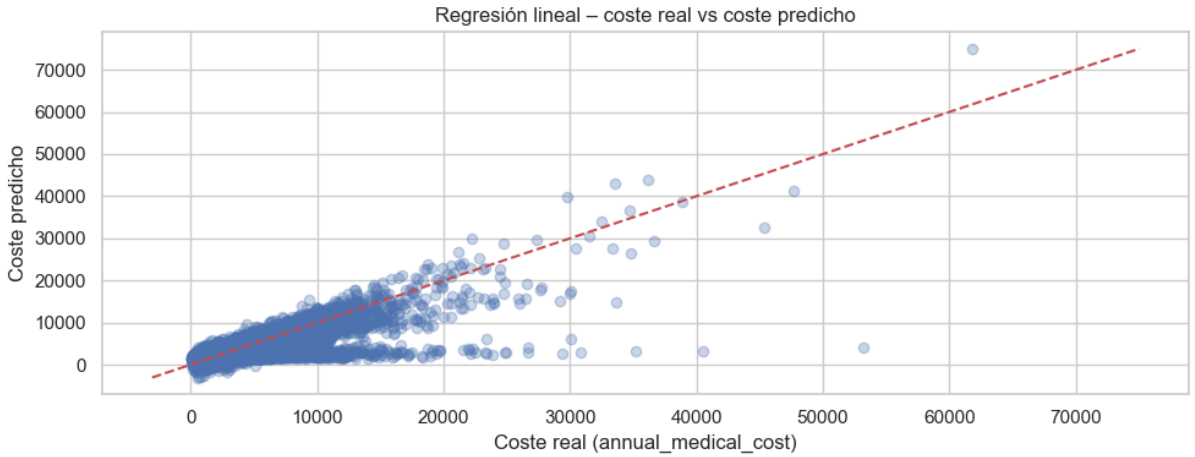
Finalmente, con esta gráfica podemos ver que no hay diferencia alguna entre el coste médico medio por persona entre las distintas categorías de cliente.

3. Modelo predictivo explicado y con tablas

En este caso de estudio se han realizado dos modelos predictivos. En un primer lugar, se ha hecho una regresión lineal, en la que se ha intentado estimar el coste médico anual, usando como variables predictoras factores como la reincidencia del cliente en reclamar partes médicos, si fuma, su edad, y demás variables que se muestran a continuación.



Con estas variables más significativas, conseguimos un resultado en la predicción del primer modelo, como se muestra a continuación:



Con los siguientes valores de métricas:

R^2 : 0.615

RMSE : 772.6

MAE : 1150.38

El segundo modelo ha sido una regresión logística, para identificar quien podría ser un *high cost oulier*, variable que hemos definido en el clustering. Los resultados son:

Accuracy : 0.962

Precisión: 0.861

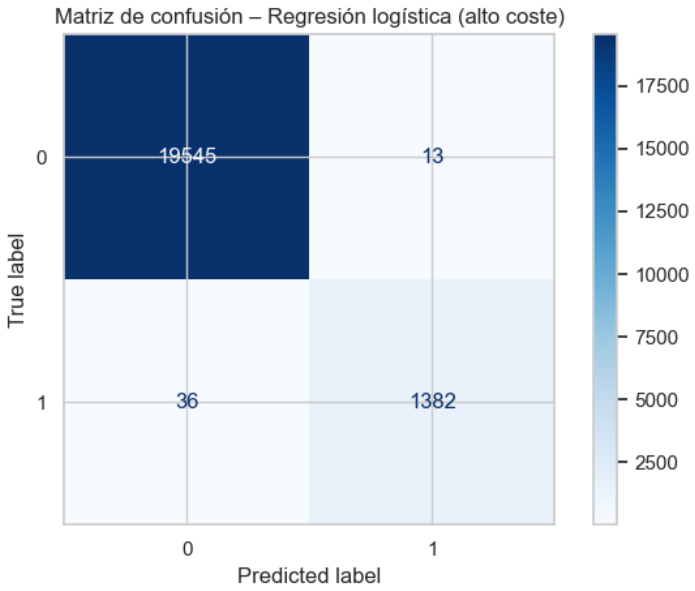
Recall : 0.516

F1 : 0.646

Y las variables más significativas son:

	variable	coeficiente
9	total_claims_paid	2.163983
5	chronic_count	0.326607
8	days_hospitalized_last_3yrs	0.226529
3	risk_score	0.149230
0	age	0.059377
7	hospitalizations_last_3yrs	0.053927
1	bmi	0.047917

Con la siguiente matriz de confusión en test:


Matriz de confusión – Regresión logística (alto coste)
19545
13
36
1382
0
1
True label
0
1
Predicted label
17500
15000
12500
10000
7500
5000
2500