Nombre y Apellidos: Juan de Egaña Marín
Github con notebook: https://github.com/Juanegana/final_DAVD

Nota: Por favor, seguir esta estructura para el documento

1. Resumen Ejecutivo

Máximo 2 páginas

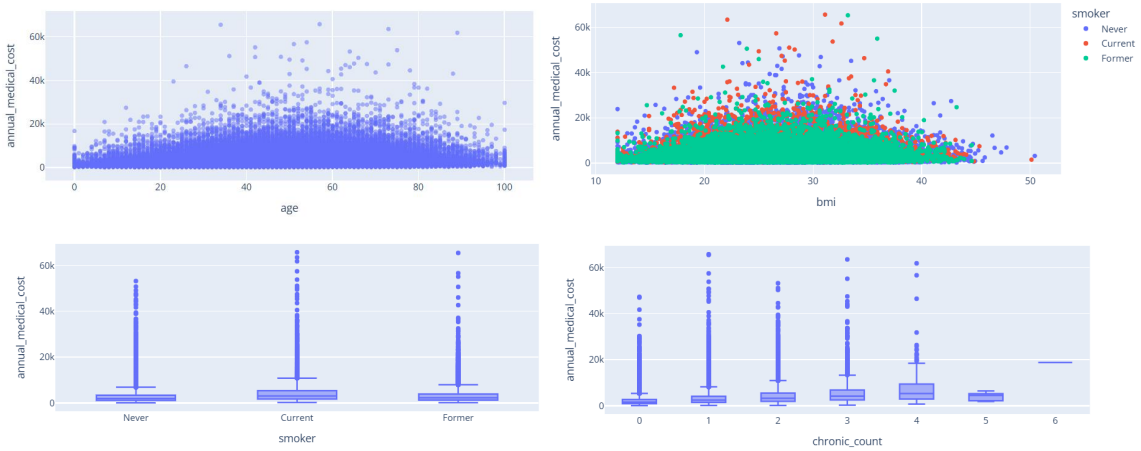
Hemos cogido los datos de una compañía de seguros la cual tiene datos del precio que paga cada persona además de las características principales de este individuo. (hábitos, estilo de vida, acceso a sanidad, etc…)

Lo primero que se ha intentado ver es la situación holística de este seguro. Podemos ver que la gran mayoria de gente paga entre 0 y 10k con gente pagando más de 20k representando una rara avis.

A continuación, lo que hemos intentado ver es en un primer momento que relaciones tiene sentido que veamos. Lo primero que hemos analizado es como evoluciona el precio con la edad y hemos encontrado que hay una distribución normal que a pesar de lo que uno pueda esperar empieza a bajar a partir de los 60 años.

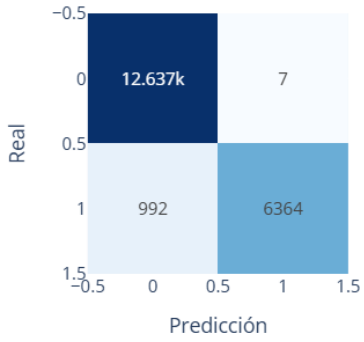
De la misma forma han resultado sorprendentes otras relaciones como que el BMI no tiene el comportamiento que uno cabría esperar en el coste (de nuevo se vuelve a ver esta relación normal) y que la gente con mayor BMI no es la que más paga, sino la gente que está en la frontera entre sobrepeso y obeso (alrededor de 30). También ha sido sorprendente la relación negativa que hay entre el precio y el numero de visitas en el último año (siendo la gente con más visitas la que menos paga).

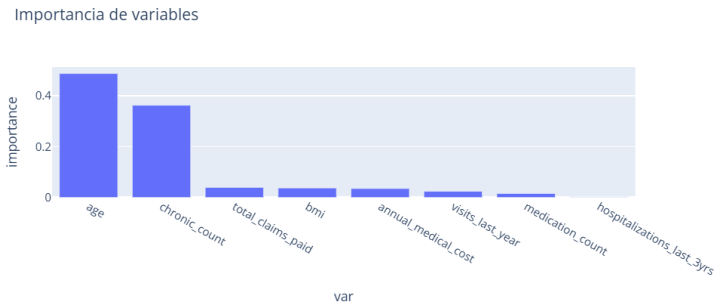
Finalmente, sí que ha habido relaciones esperables como que la gente que fuma es la que más paga seguidos de la gente que es exfumadora y finalmente de la gente que no fuma (aunque entre estos dos no hay demasiada diferencia). Y que la gente que tiene más enfermedades crónicas paga más (con la excepción de los que tienen 5 que presentan un raro decremento).



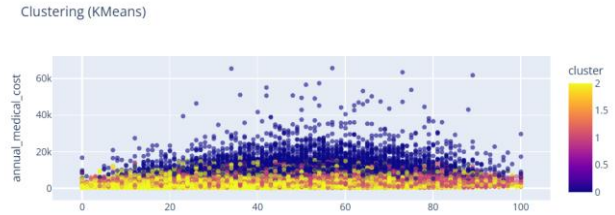
Una vez finalizado este primer análisis exploratorio se nos ha propuesto hacer uno de dos modelos: o clasificatorio de si es high risk o de regresión para predecir lo que va a pagar. Como mi comprensión lectora parece ser un tanto deficiente he hecho las dos, pero de cara a una aseguradora me parece especialmente interesante la primera para ser capaz de detectar que pacientes representan un riesgo mayor.

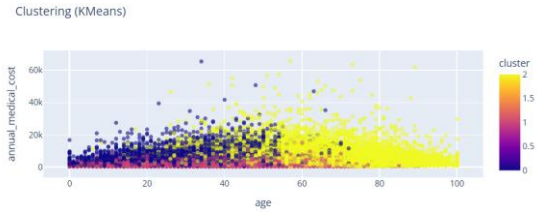
He hecho un modelo de decission tree que ha obtenido unos resultados muy buenos, aunque mejorables de cara a que predice más como no risk a gente que si que lo es que viceversa, lo cual como aseguradora no nos interesa.





Vemos que en este caso la edad y el numero de enfermedades crónicas no es tan importante. De hecho como experimento podemos ver que si hacemos un modelo de clusterización KNN-3 incluyendo la variables de is_high-risk frente a no incluirla nos salen resultados dramáticamente distintos.



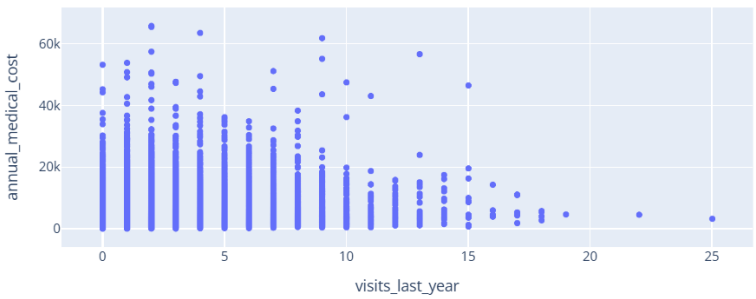


El de la derecha (incluyendo tiene unos clusteres mucho mas definidos)

Finalmente por comentar brevemente el módelo de regresión lineal, no he obtenido un resultado del todo relevante con las variables predictivas que he utilizado como se explica en ese subapartado. Pero si podemos ver que hay una relación negativa entre el precio que pagas y las visitas en tu ultimo año al médico como ya habiamos visto en las visualizaciones del principio:

3]:

	variable	coeficiente
5	hospitalizations_last_3yrs	1108.593361
3	visits_last_year	-348.895795
4	is_high_risk	289.657506
2	chronic_count	275.767968
1	bmi	13.063848
6	medication_count	-6.299540
0	age	5.785484
7	total_claims_paid	1.049533



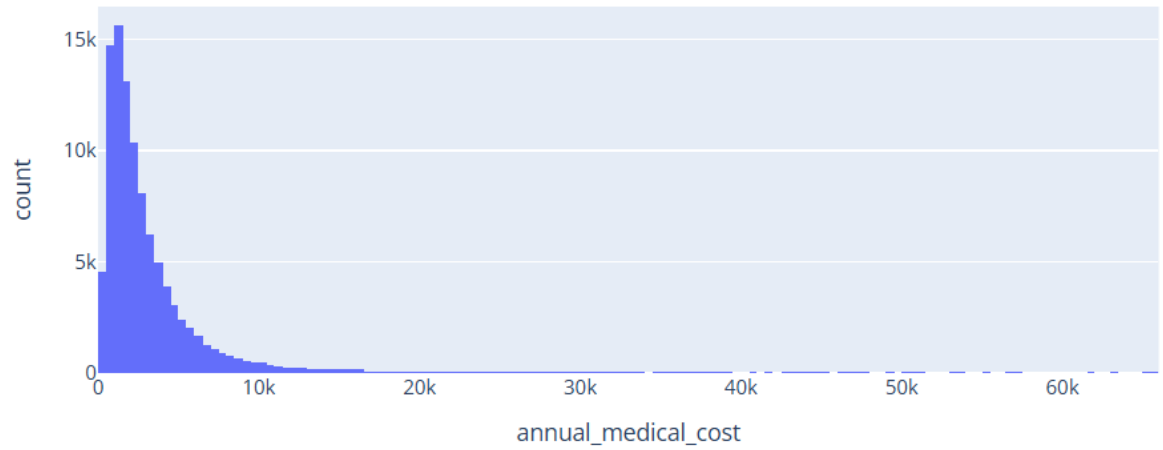
Lo cual considero que no tiene demasiado sentido.

Finalmente algunas conclusiones que podemos sacar son que la edad es el driver principal de la clasificación de si es alto riesgo o no, y que como empresa nos interesa mejorar el modelo clasificatorio para que optimizar que no haya falsos negativos (predecir que no cuando si) ya que eso nos resultará mas costoso.

También de cara a calcular el precio animaría a investigar en mayor profundidad porque estamos viendo una relación negativa entre el precio que pagas y el numero de veces que vas al hospital el año anterior (igual no estamos adaptando la tarifa suficientemente rápido). Y porque hay ciertas edades y ciertos BMIs a partir de los cuales dejamos de incrementar el precio (los pacientes si que presentan mayores riesgos).

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

Para hacer un primer análisis exploratorio, vamos a hacer unas visualizaciones medianamente sencillas en un high-level. Lo primero que me interesa ver es cuanto paga la gente de media al año:

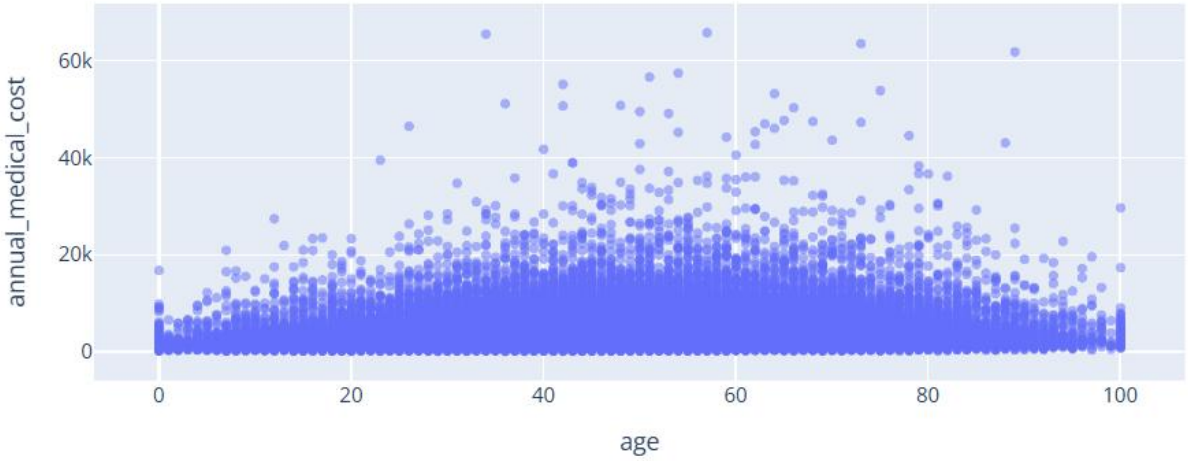


Aquí como podemos ver la gran mayoria de la gente se agrupa en el rango menor a 10k y más allá de los 20k solo hay gente residual (con algún outlier llegando hasta los 60k).

De una manera lógica lo que cabría preguntarnos es cuales son los factores que ayudan a incrementar este coste anual y cuáles son los factores que influyen al respecto.

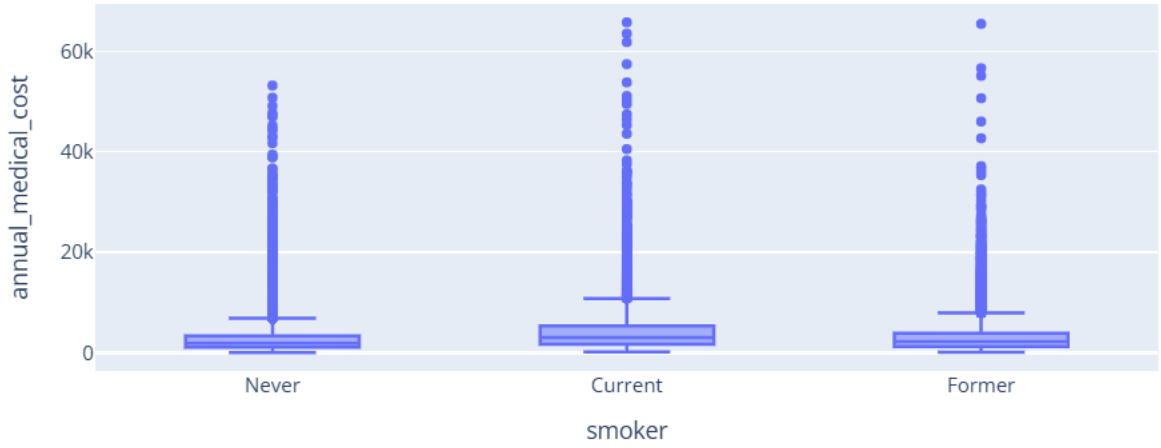
La primera pregunta que nos cabría hacernos es cómo se distribuye por edad, cabiendo esperar que a mayor edad los costes médicos anuales se deberían incrementar al haber

mayor riesgo:



En este gráfico podemos ver de una manera un tanto sorprendente que la realidad es que se aprecia casi una distribución normal con el apogeo estando en torno a los 55-60 años, y por lo tanto disminuyendo a partir de este umbral.

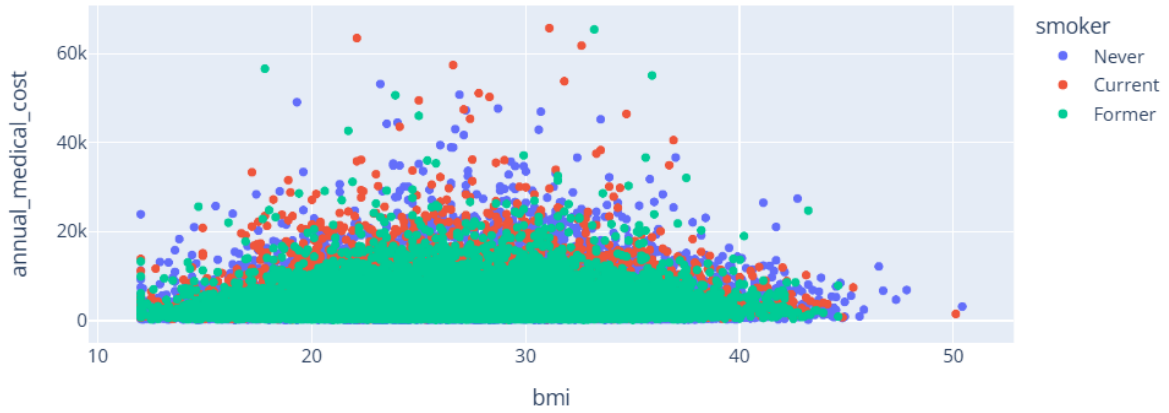
Otro factor que quizás esté conduciendo a este incremento de precios es el hecho de si el que tiene el seguro fuma o no:



En este box plot podemos ver que efectivamente la gente con los costes mas bajos anuales es el grupo que nunca ha fumado, seguido de los exfumadores y finalmente el grupo de fumadores con los costes mas elevados. Aunque realmente no parecen diferencias muy significativas.

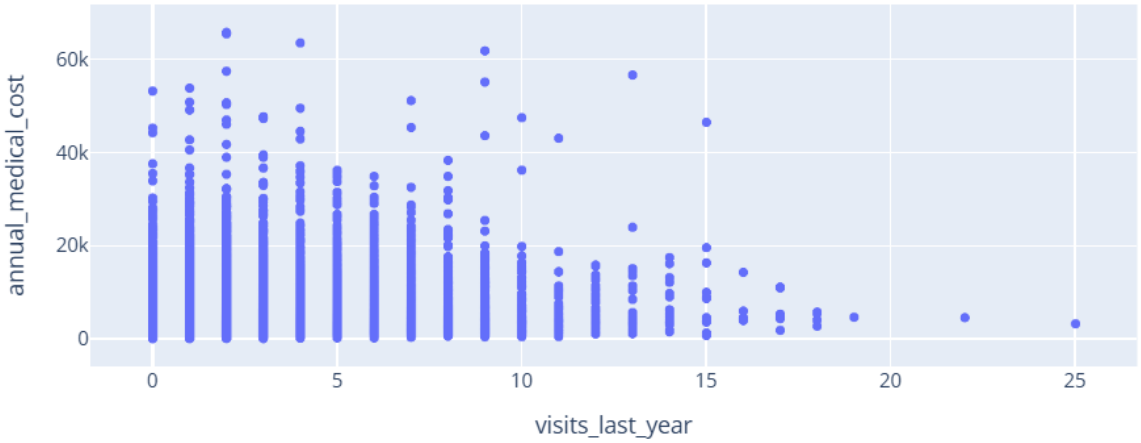
Otro factor que es interesante explorar es el bmi ya que aquellos usuarios con un bmi más elevado cabría esperar que tengan un coste más alto al incrementar los riesgos de

enfermedades como las cardiovasculares por ejemplo:



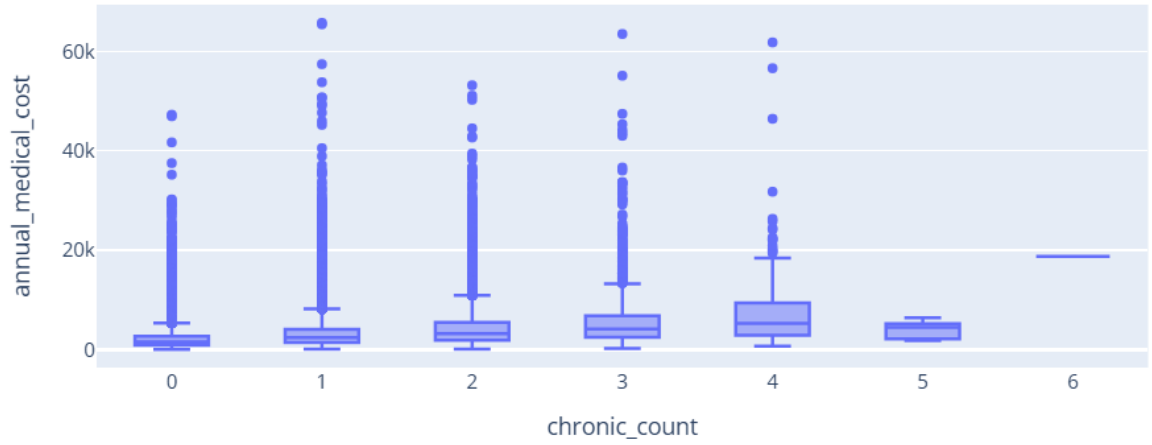
De esta forma podemos ver que al igual que con lo que nos pasaba con la relación entre edad y coste anual vemos que hay una distribución normal que a pesar de lo que cabría esperar tiene su peak en torno a los 30 (recordemos que entre 25-30 es sobrepeso y a partir de 30 se considera obesidad). Y de nuevo baja a medida que se aumenta el bmi. Lo cual no termina de cuadrar con lo que cabría esperar.
También hemos aprovechado para colorear los distintos tipos de fumadores para ver si habría algún patrón claro (gente menos saludable que fumaba tendía a tener un mayor bmi) pero se ve que apenas hay relación alguna.

Otra relación que cabría esperar es que cuanto más vayas al médico más te cobren en el seguro al año siguiente:



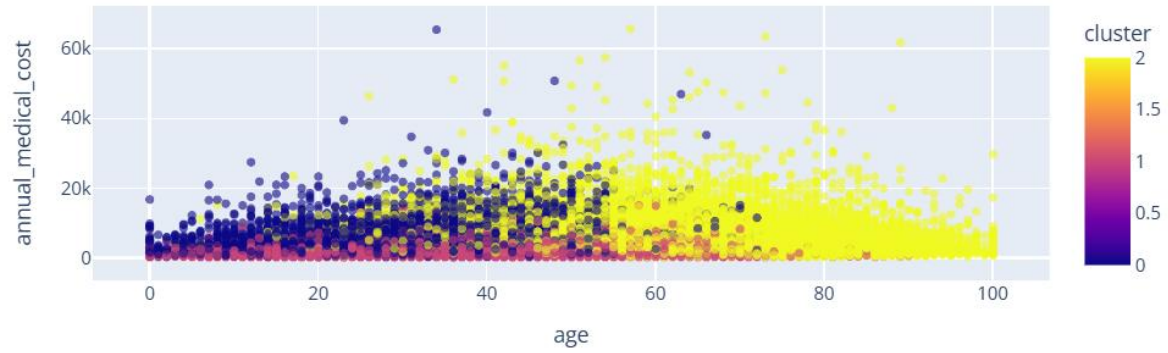
Pero como podemos ver en este gráfico de nuevo, parece que la relación es incluso inversa. Siendo la gente que visita menos el medico los que más pagan. Y los outliers con más de 15 visitas anuales al médico los que menos.

Finalmente haremos una rápida visualización para chequear la relación entre el coste y la gente que tiene enfermedades crónicas:



Como se puede ver en este boxPlot, en este caso si que vemos como existe una tendencia lógica de tal forma que los que tiene más enfermedades crónicas pagan más de media que los que no, y está relación se mantiene cierta hasta las 5 enfermedades crónicas a partir de las que nuevamente vemos un descenso poco esperado, para volver a subir con los outliers que tienen 6.

Finalmente hacemos por K-means una clusterizacion en el gráfico de edad por coste ya que los clusters permiten ver perfiles de riesgo y gasto sin necesidad de usar la variable objetivo. Podemos ver que nos sale lo siguiente: Si que parece haber dos clusteres bien diferenciados (debajo de 50 años y por encima de 50 años) y uno subyacente que parece pagar un coste bajo a pesar de la edad.



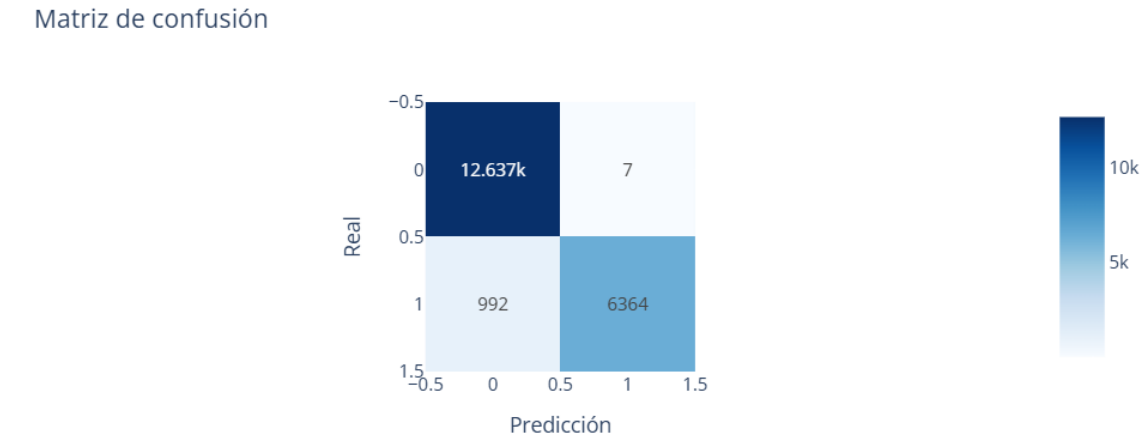
3. Modelo predictivo explicado y con tablas

Escribe aquí sobre tu modelo, qué predice, qué variables ha utilizado, métricas o tablas de evaluación y gráfico de coeficientes para explicar el modelo.

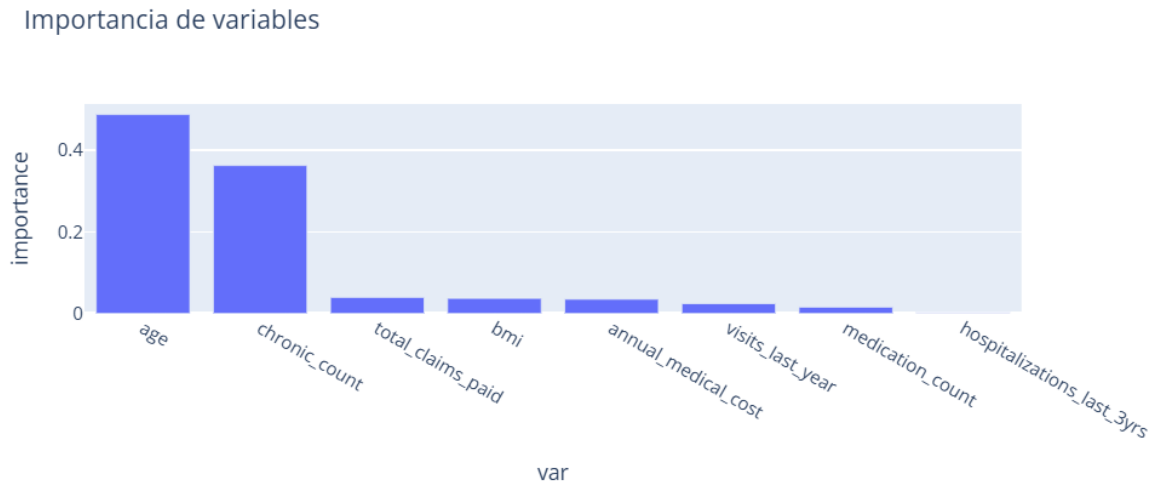
Para el modelo predictivo de si es o no es alto riesgo vamos a hacer una selección de variables que consideramos que son importantes y que tienen sentido que se incluyan y que sean relativamente faciles de trabajar (admito que esto es un pequeño shortcut pero me estaba dando problemas pasar las variables cualitativas y no queria atascarme). Estas variables elegidas son: "age", "bmi", "chronic_count", "visits_last_year", "hospitalizations_last_3yrs", "medication_count","annual_medical_cost", "total_claims_paid".

Y como Podemos ver utilizando un modelo de Random Forest Classifier con una profundidad máxima de 10 y 300 estimadores obtenemos una matriz de confusión con un score muy positivo.

	precision	recall	f1-score	support
0	0.93	1.00	0.96	12644
1	1.00	0.87	0.93	7356
accuracy			0.95	20000
macro avg	0.96	0.93	0.94	20000
weighted avg	0.95	0.95	0.95	20000



También podemos ver que tenemos predicción a clasificar como no gente que si lo es, por lo que quizás nos interesaría corregir eso (preferimos clasificar como si a gente que no es high risk de verdad.
Si queremos ver la importancia de las variables que hemos usado podemos ver que:



Como es de esperar si eres un paciente crónico y tu edad son los factores más importantes a la hora de clasificar tu riesgo.

De cara a predecir el coste de salud he hecho un modelo de regresión lineal muy sencillo que tiene las mismas columnas que antes (obviamente excluyendo el annual_medical_cost que es lo que estamos intentando predecir)

```
R² (train): 0.594
R² (test): 0.579
```

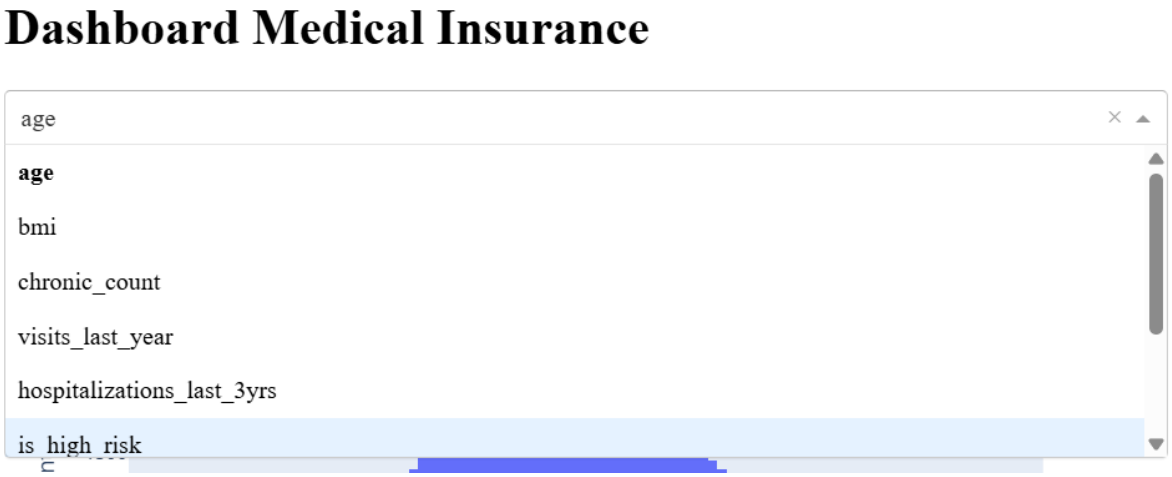
El escore que nos da de r^2 es bastante regular pero relativamente significativo. En este caso vemos que obtenemos la siguiente información sobre las variables usadas:

3]:

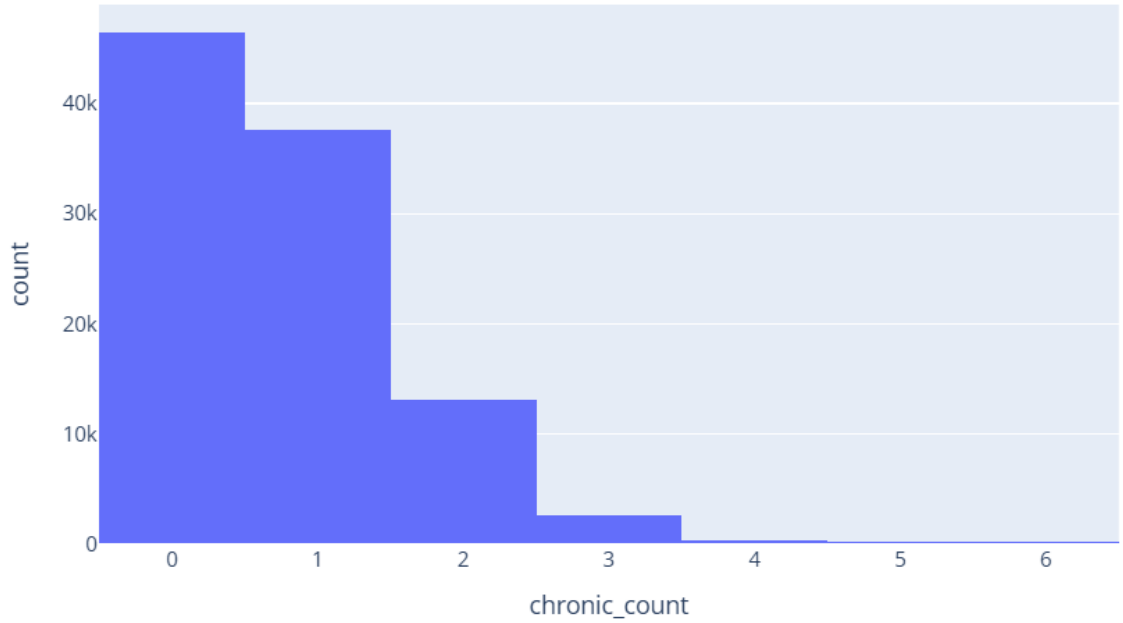
	variable	coeficiente
5	hospitalizations_last_3yrs	1108.593361
3	visits_last_year	-348.895795
4	is_high_risk	289.657506
2	chronic_count	275.767968
1	bmi	13.063848
6	medication_count	-6.299540
0	age	5.785484
7	total_claims_paid	1.049533

Lo cual nos indica que las hospitalizaciones los últimos tres años son importantes, las visitas el ultimo año como ya habíamos visto en la gráfica del principio tiene una relación negativa y las claims pagadas no tiene prácticamente importancia.

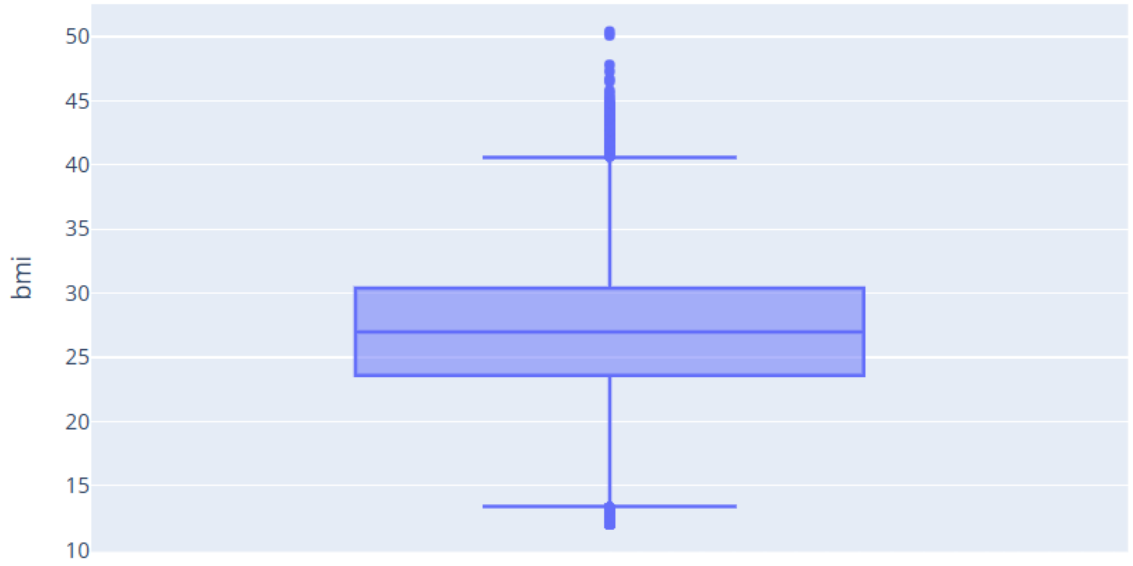
De cara al dashboard he hecho algo pequeñito pero funcional tal que así:



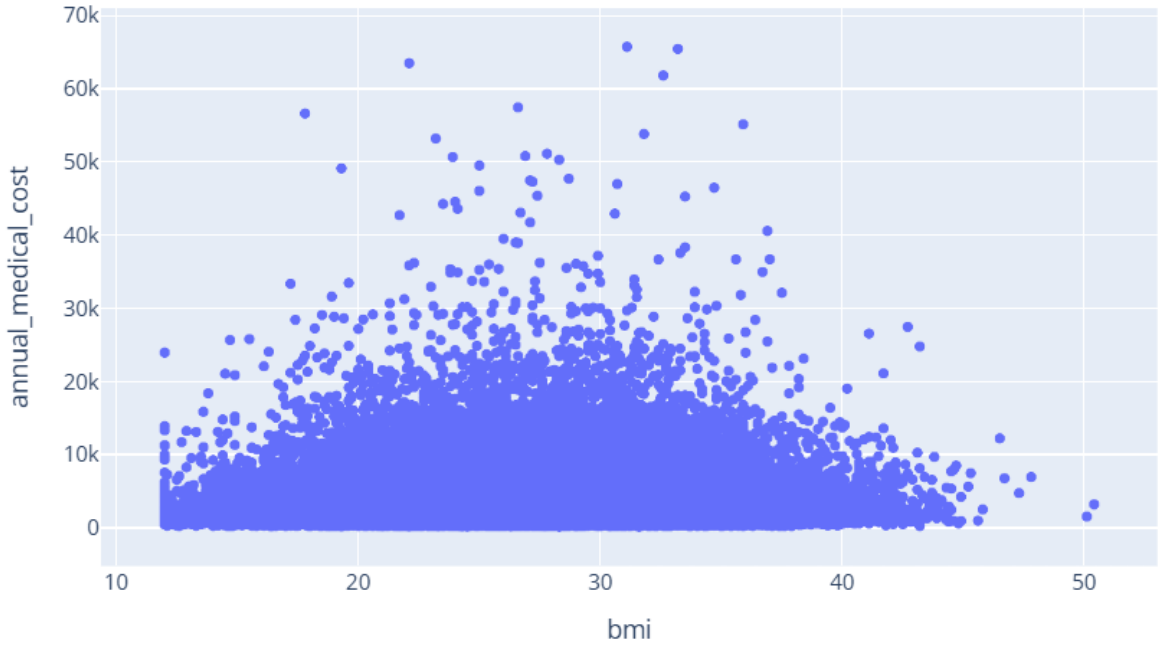
Un dropdown list para elegir la variable que nos interese.



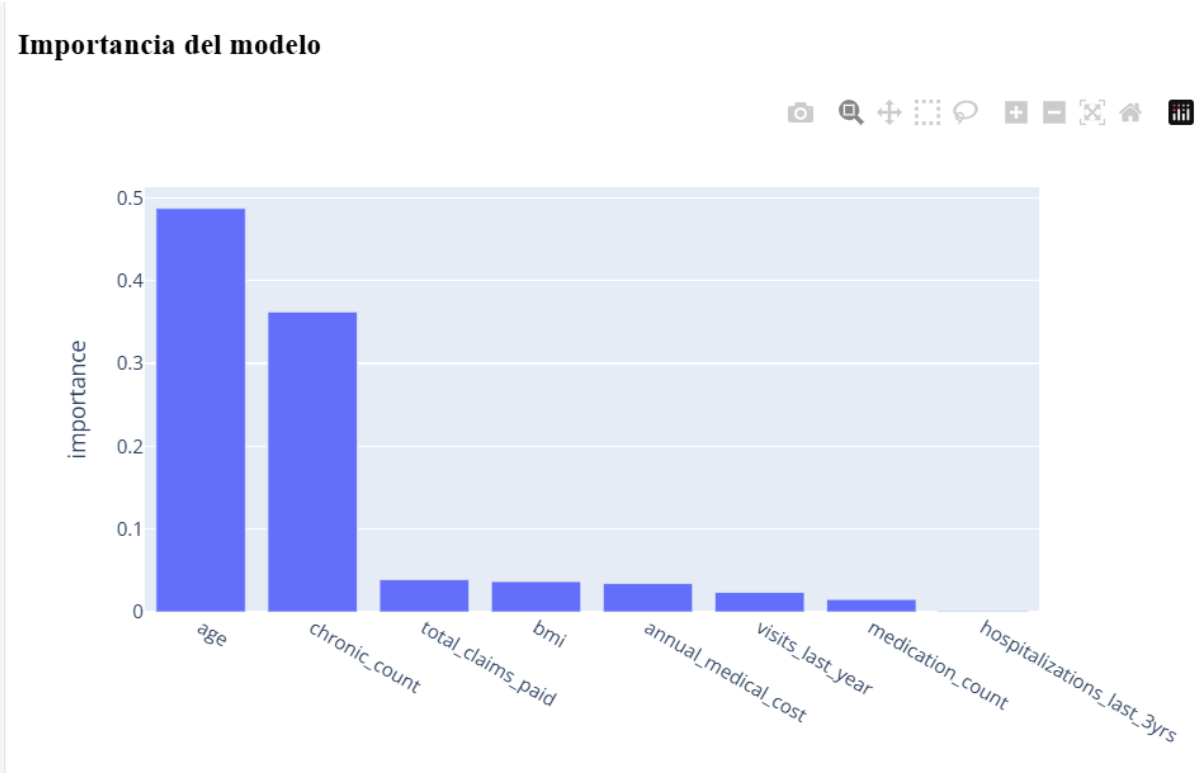
Un recuento de esa variable.



Un boxplot



Relación con el coste medio anual.



y finalmente el grafico de la importancia del modelo de predicción que hemos hecho antes.