Data Science II: Machine Learning para la Ciencia de Datos

Primera entrega: "Factores para Costos en Seguros Médicos"

- Comisión 61605
- Duarte Bruno Julián

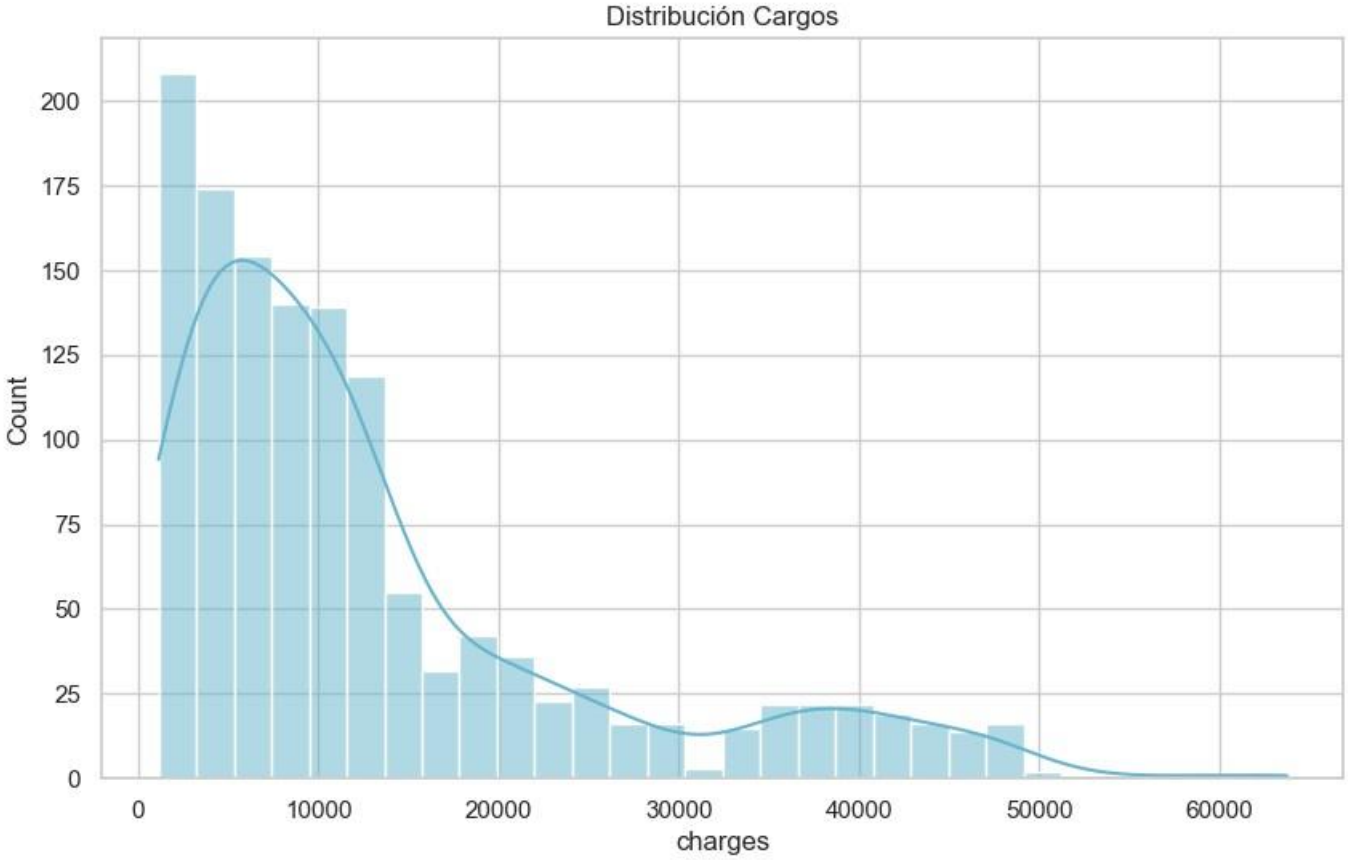
Sobre el dataset:

Según el mercado en el cuál operan los seguros médicos pueden variar el tipo de cobertura y forma de cálculo de costos sobre los mismos. En este caso, tomaremos una base de una población modelo en Estados Unidos como muestra para evaluar su seguro médico en los siguientes campos:

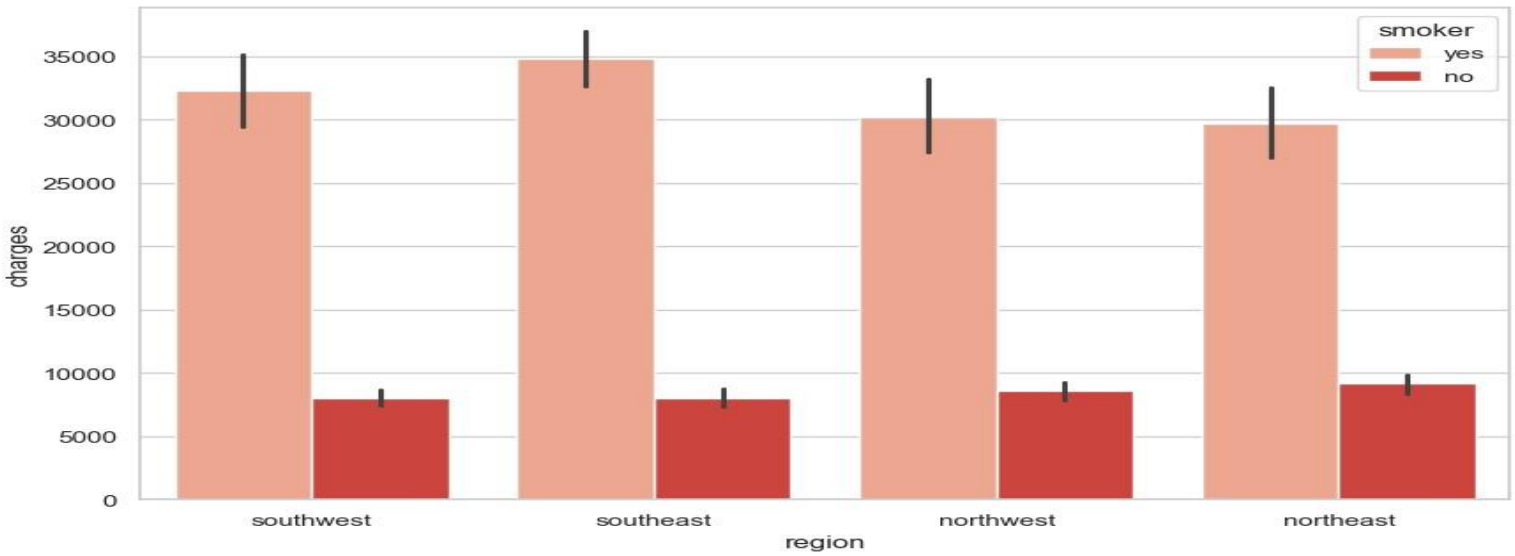
- Age: Edad del beneficiario principal, quien contrata.
- Sex: Género del contratante del seguro, hombre o mujer.
- BMI: Por las siglas en inglés de body mass index o índice de masa corporal.
- Children: número de hijos dependientes que estarían cubiertos por el seguro
- Fumador: si el encuestado es fumador.
- Region: área de residencia del beneficiario charges: cargo imputado anualmente al beneficiario.

Tanto como quienes solicitan este tipo de seguros, como aquellos que lo brindan pueden emplear este tipo de información para determinar los puntos de mejora en salud como consumidor, o claves de costos siendo proveedor.

Para entender un poco cómo se distribuyen los cargos, analizamos algunos gráficos con las variables presentes en el dataset. En primer lugar, la distribución y conteo del valor de los cargos.

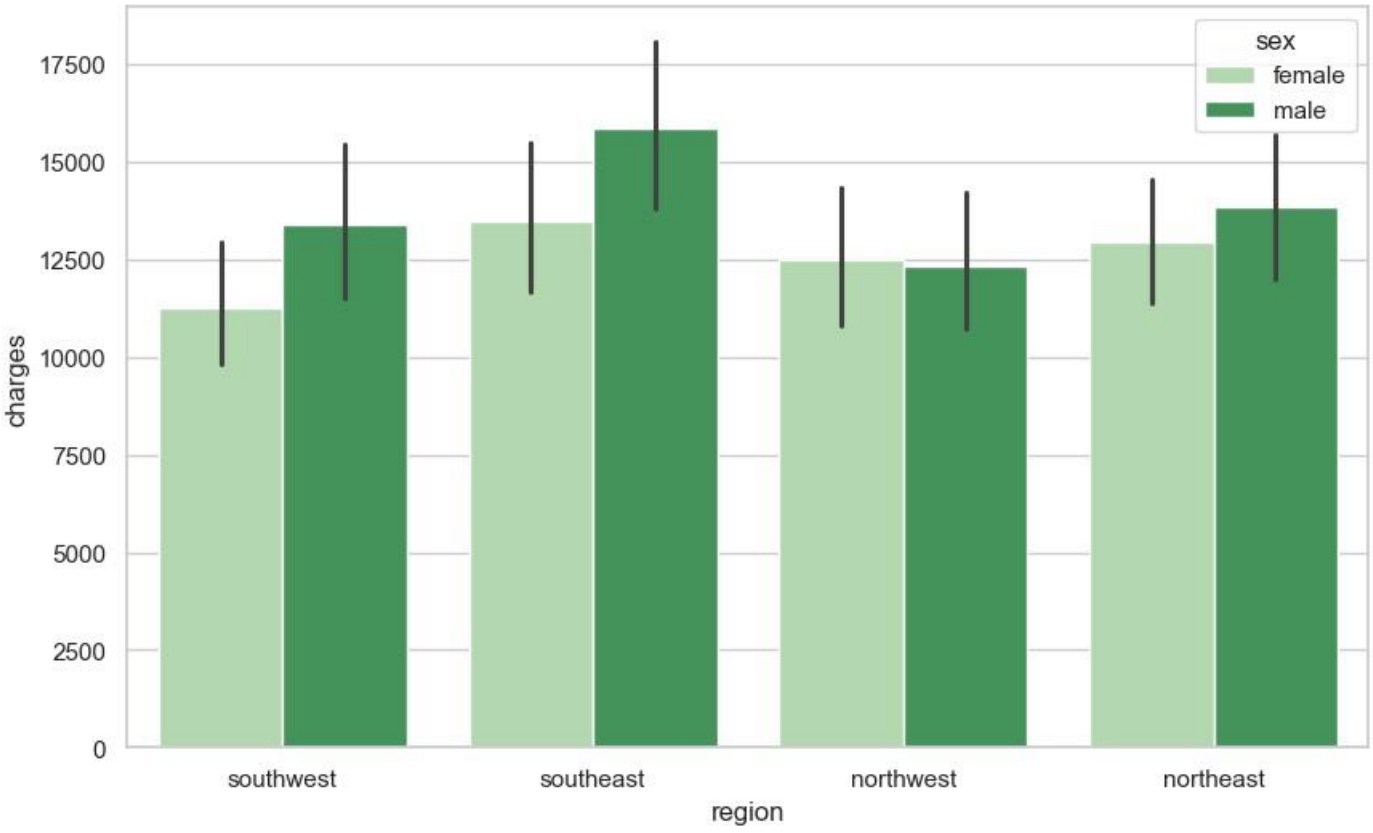


Por contextos razonables, podemos estimar que personas fumadoras son quienes poseen mayor factor de riesgo, por lo que presentan mayor cargo a abonar.

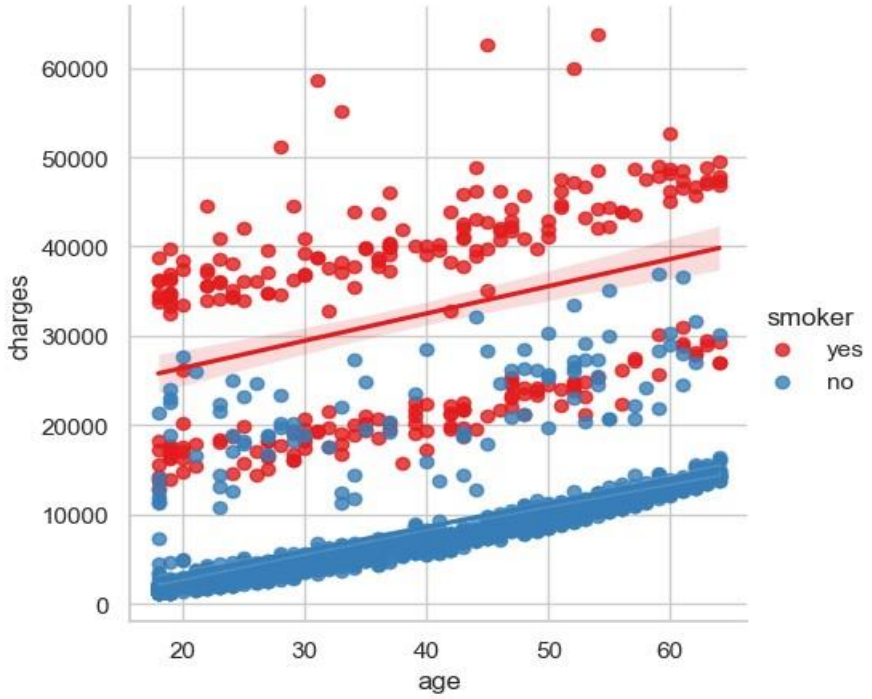


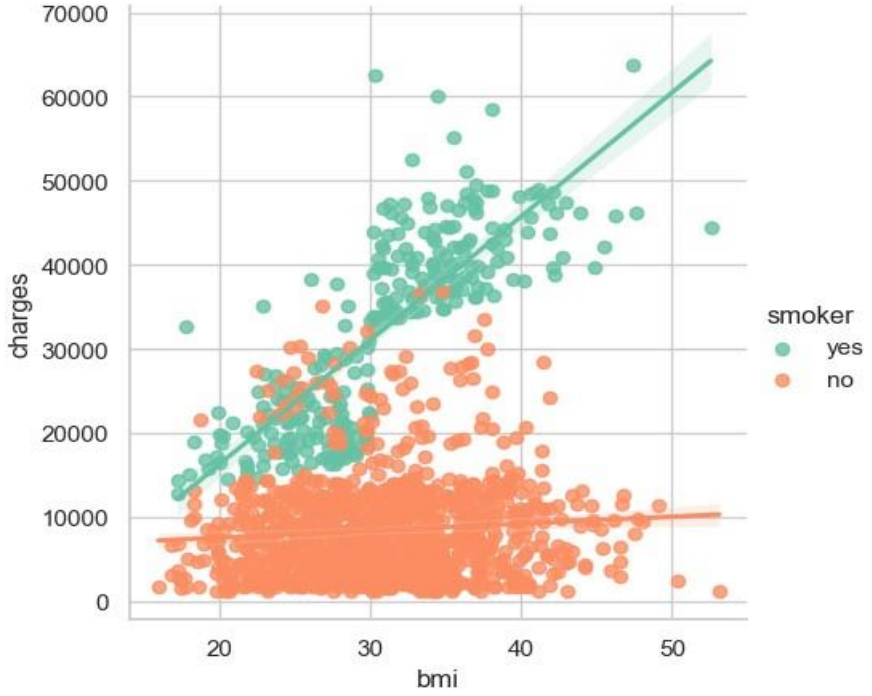
Aquí pueden verse los cargos para fumadores distribuidos por región.

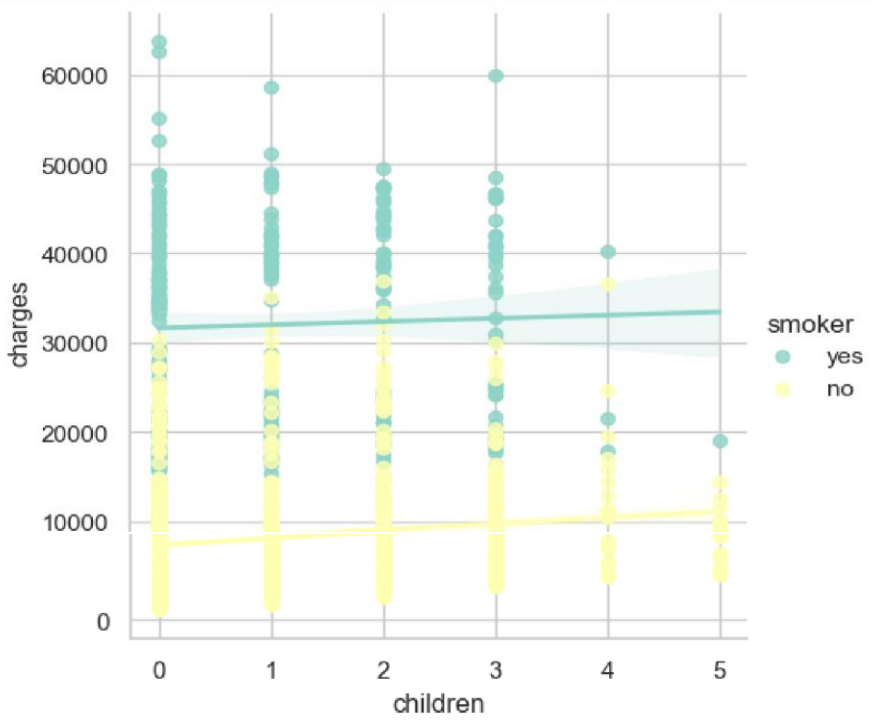
En cuanto a la relación entre género y cobro realizado, la relación es relativamente simétrica, solo en la región sur oeste podemos notar una leve diferencia:



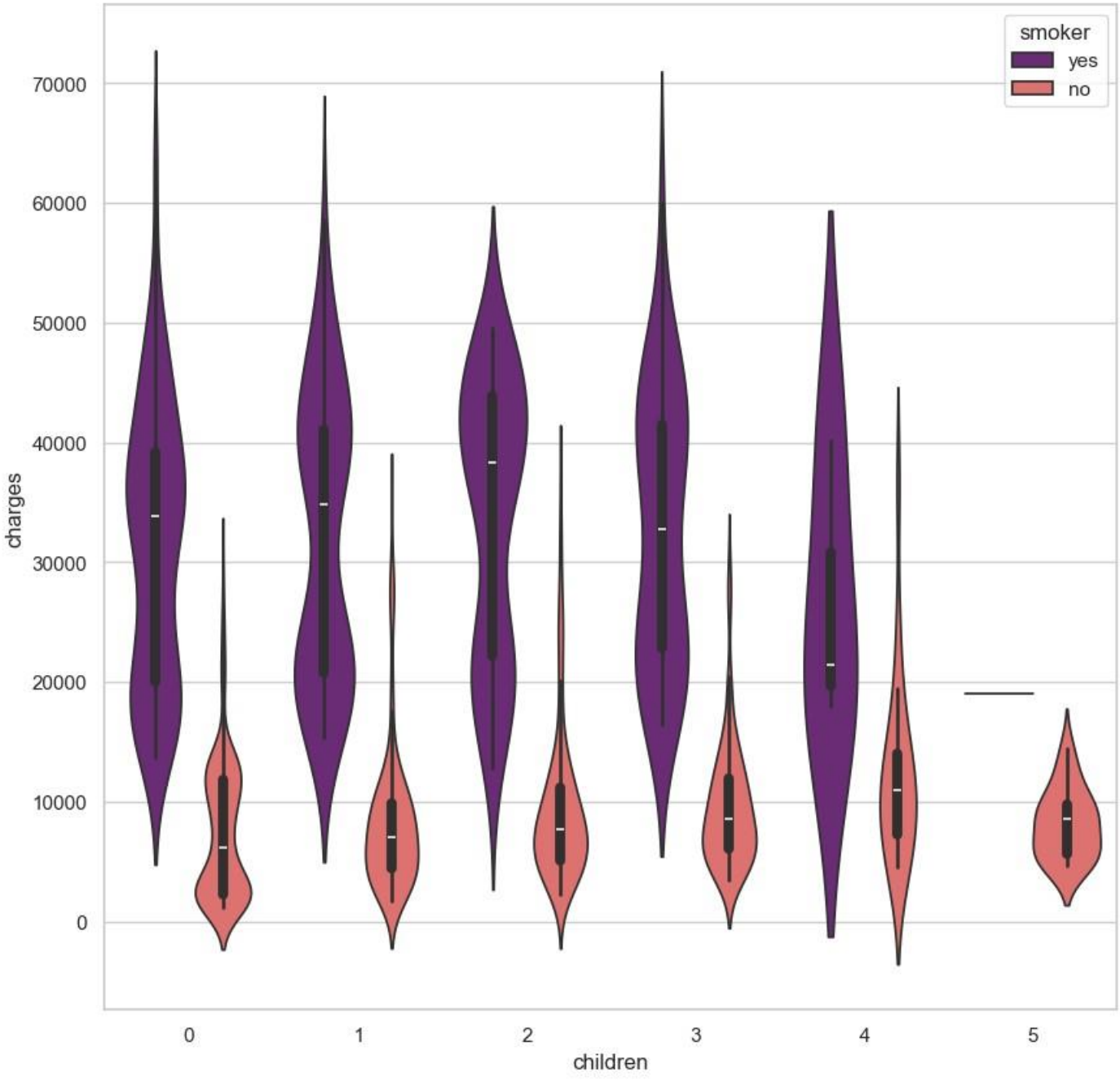
Veamos ahora cómo varían los gastos médicos en función de la edad, el IMC y la cantidad de hijos, considerando si las personas fuman o no.


charges
60000
50000
40000
30000
20000
10000
0
20
30
40
50
60
age
smoker
yes
no


charges
70000
60000
50000
40000
30000
20000
10000
0
20
30
40
50
bmi
smoker
yes
no


charges
60000
50000
40000
30000
20000
10000
0
0
1
2
3
4
5
children
smoker
yes
no

Fumar es el factor que más influye en los costos médicos, aunque estos también aumentan con la edad, el IMC y el número de hijos.
Además, las personas con hijos tienden a fumar menos, algo que también se refleja en los diagramas a continuación.



Procesamiento de datos y análisis de Algoritmos

Vamos a convertir los datos de las columnas cuyos formatos no podrían ser analizados mediante una regresión del tipo lineal, que será el primer ejemplo que utilizaremos parra la predicción:

- Dado que son pocas columnas para este dataset, no es necesario realizar un descarte de columnas, y también como vimos anteriormente no contamos con campos vacíos.
- Aplicaremos entonces un Encoder para transformar las columnas objetos en datos legibles.
- Comenzaremos ahora con una carga y entrenamiento para una regresión lineal de la biblioteca de scikit-learn, donde solicitaremos luego que nos de información sobre los coeficientes obtenidos para validar como ha impactado.
- Utilizaremos ahora como herramienta una regresión Polinomial, para llegar a un valor más alto de r2 sobre la misma línea trabajada en la regresión lineal, modelando relaciones no lineales entre las variables independientes y la variable dependiente.
- Emplearemos por defecto el grado 2 para enteros.

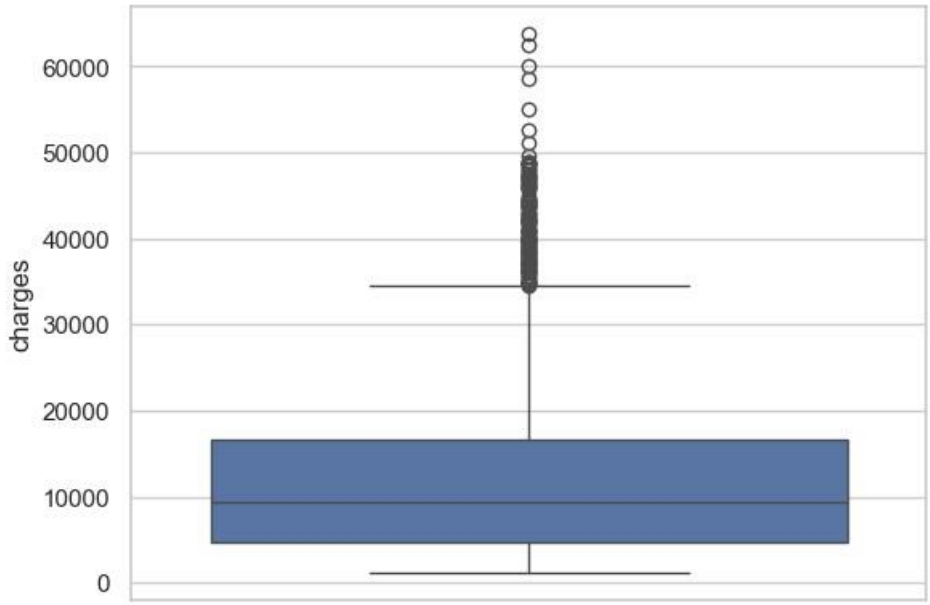
Podemos observar que el puntaje obtenido ha mejorado, por lo que ya podemos obtener algunas afirmaciones:

- El valor -8062.393521836057 corresponde a los cargos cuando todas las variables independientes (edad, IMC, hijos, fumador) son cero. Aunque no tiene un significado práctico directo, es una parte necesaria del modelo.
- El Coeficiente de Determinación 87.50% nos muestra la variabilidad en los cargos, que puede ser explicada por este modelo de regresión polinomial. Esto indica un ajuste muy bueno del modelo a los datos.
- Las variables edad, el IMC, el número de hijos y si la persona es fumadora tienen un impacto significativo en los costos de seguro, mejorando el ajuste en comparación con el modelo de regresión lineal simple.

Sin embargo, tal vez consecuencia del número total de casos a analizar, si efectuamos un segundo análisis teniendo en cuenta la presencia de casos Outliners, impactará de lleno en las predicciones que podamos realizar con ambos modelos, sobre los cargos que se efectúan a los beneficiarios.

Revisión mediante análisis de Outliners

Mediante este gráfico Boxplot, vemos la presencia de datos Outliners en los cargos cobrados:



- Definimos en el dataset los cuantiles superior e inferior para determinar sus límites, y procedemos luego a retirarlos del análisis.
- Veremos una reducción en el tamaño de la muestra, no es significativa en cantidad, pero si en los valores que imputaban:

Esto impacta significativamente en los puntajes obtenidos para los coeficientes de determinación, tanto para el modelo lineal como para el polinomial:

Con esto podemos saber que, si bien el método de regresión Polinomial sigue siendo mejor en la predicción que el método lineal, nos da un margen de mejora para el método de predicción. Con esto deberíamos analizar un aumento de la cantidad de datos a ingresar a nuestro modelo para entrenar y reducir el impacto de outliners. Otros métodos podrían aplicarse para el análisis, pero en resumen nos dieron resultados parciales similares.