UNIVERSIDAD DEL VALLE DE GUATEMALA

Minería de Datos - CC3074

Sección 30

Ing. Leonel Guillén

Hoja de Trabajo No. 4

José Pablo Orellana 21970

Diego Alberto Leiva 21752

Gabriel Estuardo García 21352

GUATEMALA, 25 de febrero 2024

Introducción

En este estudio, se llevó a cabo un análisis detallado utilizando modelos de regresión lineal para predecir los cargos de seguro en base a diversas variables, como lo pueden ser la edad, índice de masa corporal, entre otras variables. El objetivo principal será evaluar la relación entre estas variables menos significativas en el rendimiento del modelo.

Se realizó un EDA exhaustivo para comprender la distribución de las variables en el conjunto de datos. Se utilizaron gráficos y estadísticas descriptivas para obtener insights sobre edad, bmi, fumador, región, etc.

Metodología

Se utilizó el modelo de regresión lineal de la biblioteca scikit-learn en Python. El conjunto de entrenamiento se ajustó al modelo, y se realizaron predicciones en el conjunto de prueba. Se exploró la importancia de las variables y se realizó un análisis de residuos para evaluar la adecuación del modelo.

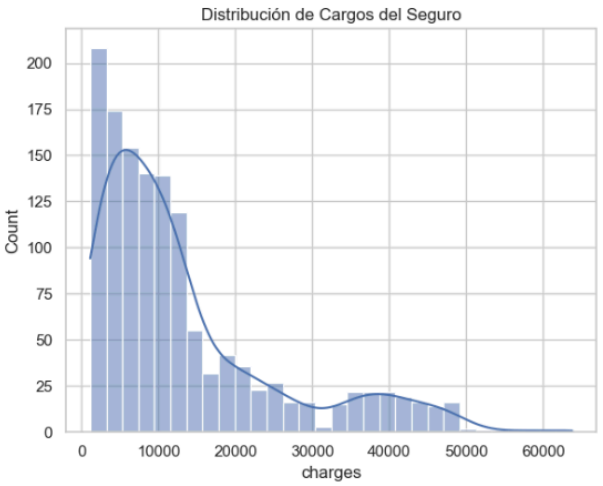
Se usaron gráficos de dispersión para visualizar la relación entre variables independientes y la variable dependiente. Esto proporcionó una comprensión visual de cómo cada característica se relaciona con la variable objetivo.

Se analizaron los coeficientes obtenidos del modelo de regresión lineal para entender cómo cada característica impacta sobre las predicciones de los cargos de seguro.

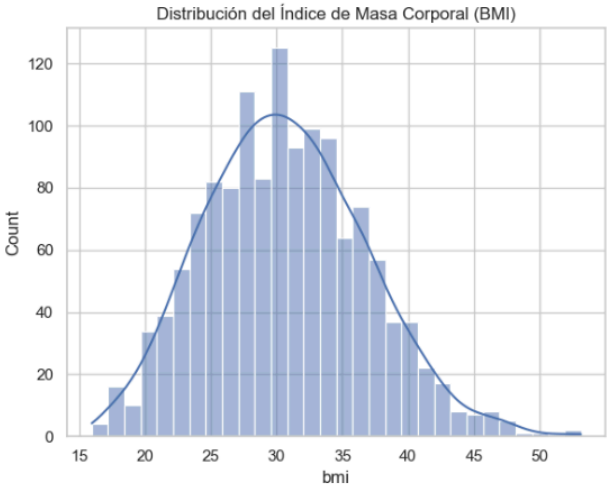
Evaluación del Modelo

Se utilizaron métricas de evaluación del modelo, como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2), para medir el rendimiento del modelo en el conjunto de prueba. Se compararon modelos con y sin variables menos significativas.

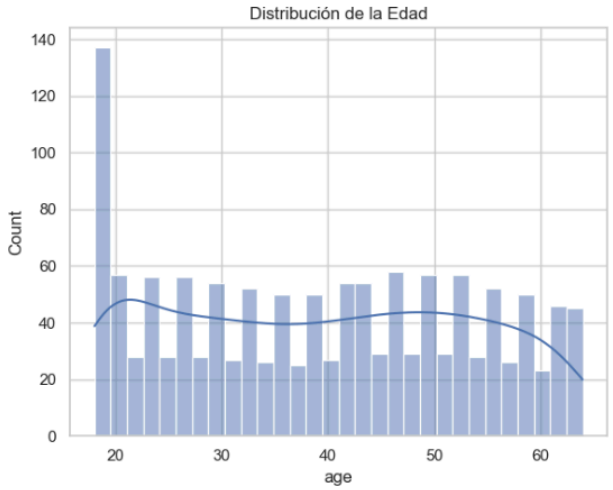
Resultados



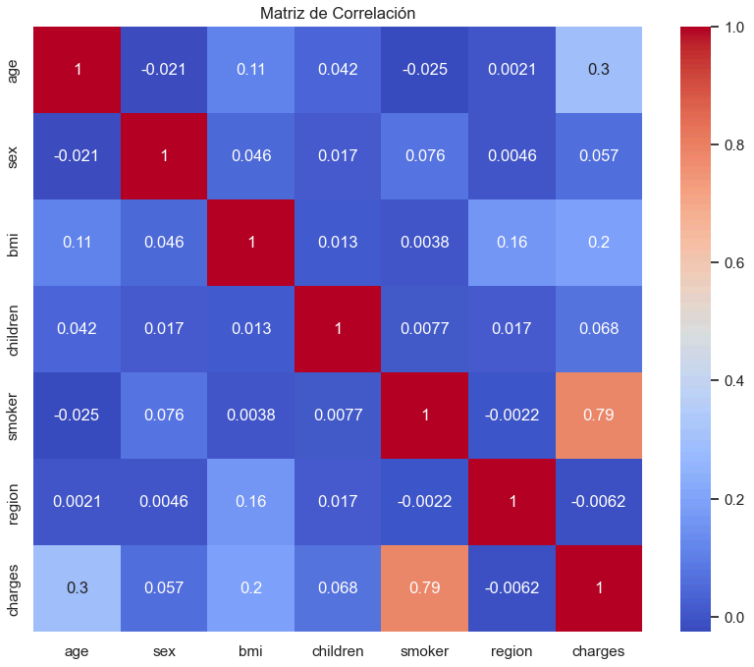
Cargos del Seguro: La distribución de los cargos del seguro es asimétrica a la derecha, con la mayoría de los asegurados incurriendo en cargos más bajos y un número menor de asegurados enfrentando cargos muy altos. Esto sugiere que mientras la mayoría de los asegurados podrían estar en planes menos costosos o tener menos necesidades médicas, hay un grupo significativo con gastos mucho mayores, posiblemente debido a condiciones médicas crónicas o factores de riesgo elevados.



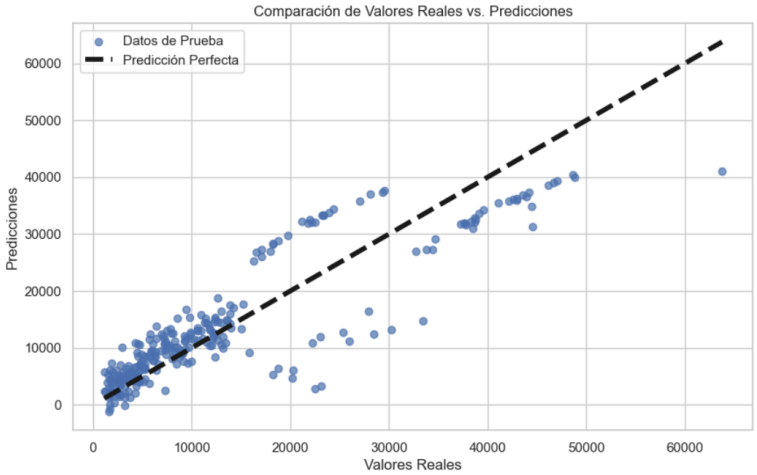
Índice de Masa Corporal: La distribución del BMI es aproximadamente normal, con un ligero sesgo hacia valores más altos, indicando que una porción considerable de la población asegurada tiene sobrepeso. Esto es relevante ya que el BMI elevado está asociado con un mayor riesgo de condiciones de salud que podrían afectar los costos del seguro.



Edad: La distribución de la edad muestra una amplia representación de edades entre los asegurados, con picos menores en los extremos inferiores y superiores del rango de edad. Esto indica una diversa población asegurada en términos de edad.



- La variable cargos del seguro muestra una correlación positiva moderadamente fuerte con fumador, lo que sugiere que ser fumador es un factor significativo en el aumento de los costos del seguro. Esto es lógico, ya que fumar está asociado con un mayor riesgo de enfermedades.
- Hay correlaciones positivas más débiles entre cargos y otras variables como edad y bmi, indicando que la edad y el BMI también contribuyen al costo del seguro, aunque en menor medida en comparación con el hábito de fumar.
- La correlación entre bmi y edad es relativamente baja, lo que sugiere que no hay una colinealidad significativa entre estas dos variables.
- Las variables hijos, regiones y sexo tienen correlaciones bajas con cargos, sugiriendo que estos factores tienen un impacto menos directo en los costos del seguro.

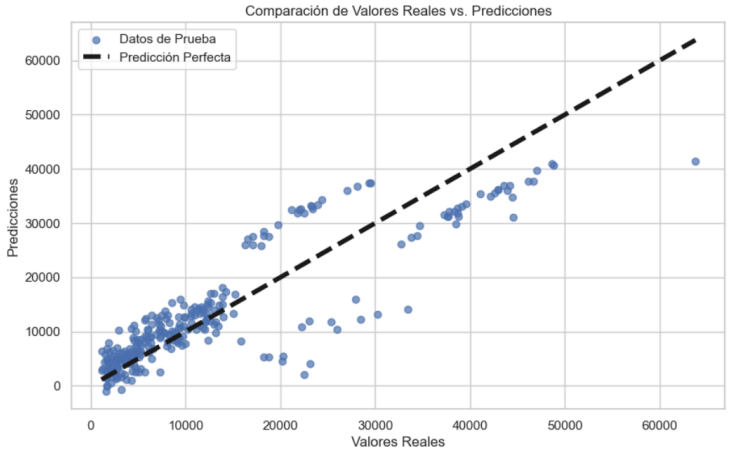


Tal como se mostró en el gráfico de correlaciones, está claro que fumador, edad y bmi son las 3 variables más importantes a tomar en cuenta, el siguiente modelo únicamente usará dichas variables.

La ecuación se define

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

- y es la variable dependiente que queremos predecir,
- β_0 es el intercepto,
- β_i son los coeficientes para cada variable independiente X_i,
- X_i son las variables independientes.



Conclusión

Al evaluar el nuevo modelo sin incluir las variables poco significativas, el “r” cuadrado disminuyó en milésimas. Este cambio en el “r” cuadrado podría parecer contra intuitivo, sin embargo existen varias posibilidades de dicha disminución.

La primera es que aunque estas variables puedan tener una relación débil con la variable objetivo por sí solas, aún pueden contribuir marginalmente al modelo. En otras palabras, pueden capturar una pequeña parte de la variabilidad en los datos que no es capturada por las otras variables. Al removerlas, esta variabilidad ya no es explicada, lo que puede resultar en una disminución del r cuadrado.

La segunda es que un modelo más simple (con menos variables) podría tener un r cuadrado ligeramente más bajo en los datos de entrenamiento, pero podría generalizar mejor a datos no vistos. Esto es especialmente relevante cuando consideramos la posibilidad de sobreajuste; un modelo con un “r” cuadrado muy alto en los datos de entrenamiento puede no desempeñarse tan bien en los datos de prueba.

Presentación

https://www.canva.com/design/DAF93Fupa7M/dRhTjqyvYpI8gpljpOE3Yg/edit?utm_content=DAF93Fupa7M&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton