Tecnológico de Monterrey

Análisis de Modelo de Regresión Lineal

Diego Mellado Oliveros

Inteligencia Artificial Avanzada para la Ciencia de Datos, Módulo 2

Grupo 101

Jorge Adolfo Ramírez Uresti

Índice

Introducción

El propósito del siguiente documento es analizar un modelo de regresión lineal creado con la librería de *sci-kit learn*, se expondrá información relacionada al dataset, las métricas usadas para evaluar el modelo, el sesgo que podría estar presente en el modelo y las recomendaciones para tener un mejor modelo de aprendizaje máquina usando regresión lineal.

Justificación elección del data

El dataset elegido resulta apropiado para la aplicación de un algoritmo de regresión lineal debido a varias razones fundamentales. Primero, incorpora características pertinentes, como género, altura y peso, todas las cuales son numéricas y directamente aplicables en un modelo de regresión lineal. Además, la variable objetivo, el índice de masa corporal (IMC), es una variable continua que se ajusta perfectamente al enfoque de regresión lineal. Aunque el tamaño exacto del dataset no se proporciona, parece contener una cantidad razonable de datos, lo que es crucial para entrenar un modelo robusto. La diversidad en las características, que incluye una amplia gama de valores para altura, peso y género, permite al modelo capturar la variabilidad en los datos y generalizar en lugar de simplemente memorizar valores específicos. Además, el enfoque de codificación del género como variable dummy posibilita su uso en el modelo. Para demostrar que el modelo generaliza de manera efectiva, se pueden realizar pasos clave, como la división de datos en conjuntos de entrenamiento y prueba, la validación cruzada, la evaluación en el conjunto de prueba y el análisis de residuos, lo que garantiza una evaluación rigurosa y confiable del rendimiento del modelo.

Separación de Datos

Antes de dividir el set de datos, se realizó un preprocesamiento, usando *LabelEncoder* para la columna Gender. Codificandola en ceros y unos para mejorar el entrenamiento del modelo. Los datos fueron separados usando la función de scikit-learn *train_test_split*, usando un 80% para el entrenamiento y un 20% para realizar las pruebas.

```python
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

```python
x_train
```

```
Gender Height Weight
 1   153   104
 0   146   104
 0   180   156
 1   183   147
 0   156    52
```

```python
x_test
```

```
Gender Height Weight
 0   168    87
 1   182   151
 1   152   103
 0   157    60
 0   187   121
```

Métricas de Evaluación

Mean Squared Error (MSE)

El Error Cuadrático Medio (MSE) es una métrica que evalúa la calidad de las predicciones de un modelo de regresión al calcular el promedio de las diferencias al cuadrado entre las predicciones del modelo y los valores reales.
En el contexto de este análisis, el MSE cuantifica la magnitud promedio de los errores al cuadrado entre las predicciones del modelo de regresión y los valores reales de IMC. Un MSE más bajo indica que el modelo tiene un mejor ajuste a los datos y que las predicciones se desvían menos de los valores reales. En este caso, el MSE al ser de 0.36, puede interpretarse que los valores de predicción tienen un error poco significativo.

R-squared

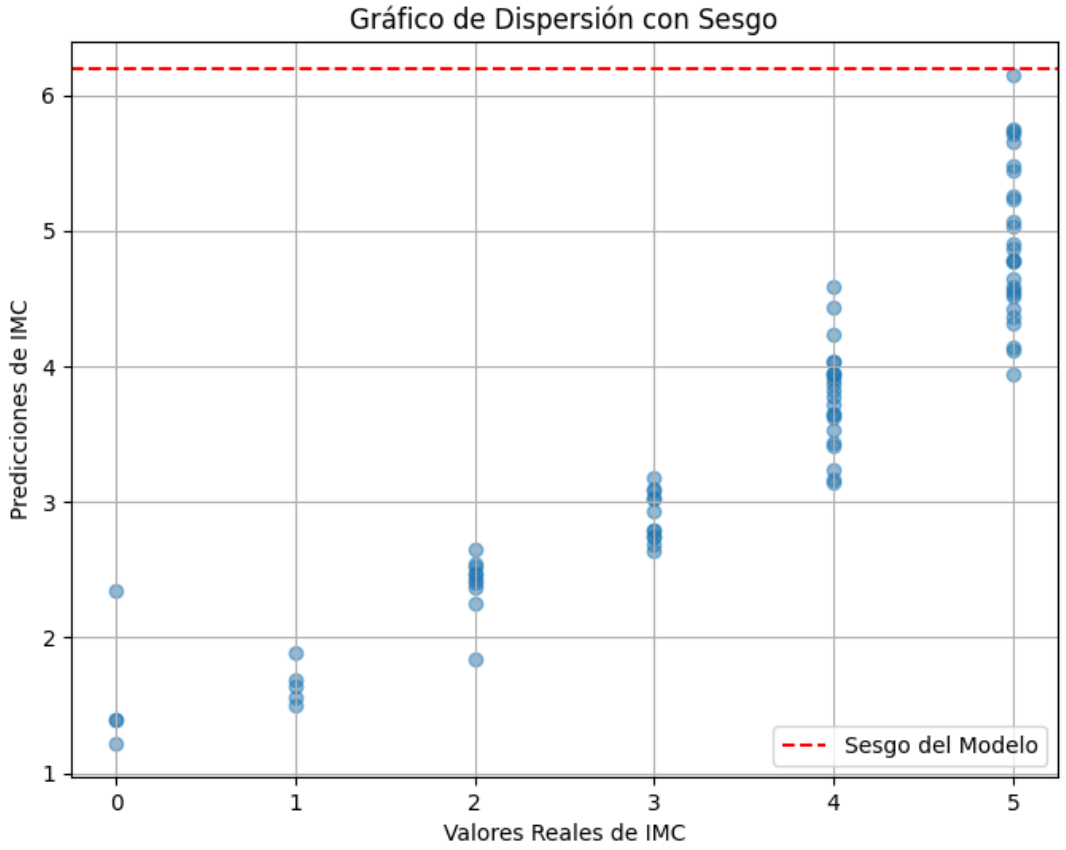
El coeficiente de determinación, o R-squared, mide la proporción de la varianza total en los valores de IMC que es explicada por el modelo de regresión. Se calcula como la proporción de la varianza explicada en relación con la varianza total.
En este contexto, un de aproximadamente 0.834 indica que alrededor del 83.4% de la variabilidad en los valores de IMC se explica mediante el modelo de regresión. Un valor cercano a 1 es deseable, ya que sugiere que el modelo es capaz de explicar una gran parte de la variabilidad en los IMC observados.

Mean Absolute Error (MAE)

El Error Absoluto Medio (MAE) es una métrica que evalúa la precisión de las predicciones del modelo al calcular el promedio de las diferencias absolutas entre las predicciones y los valores reales.
En el contexto de este análisis, el MAE de aproximadamente 0.459 indica que las predicciones del modelo tienen un error absoluto promedio de 0.459 unidades de IMC en comparación con los valores reales. Un MAE más bajo indica una menor discrepancia absoluta entre las predicciones y los valores reales.

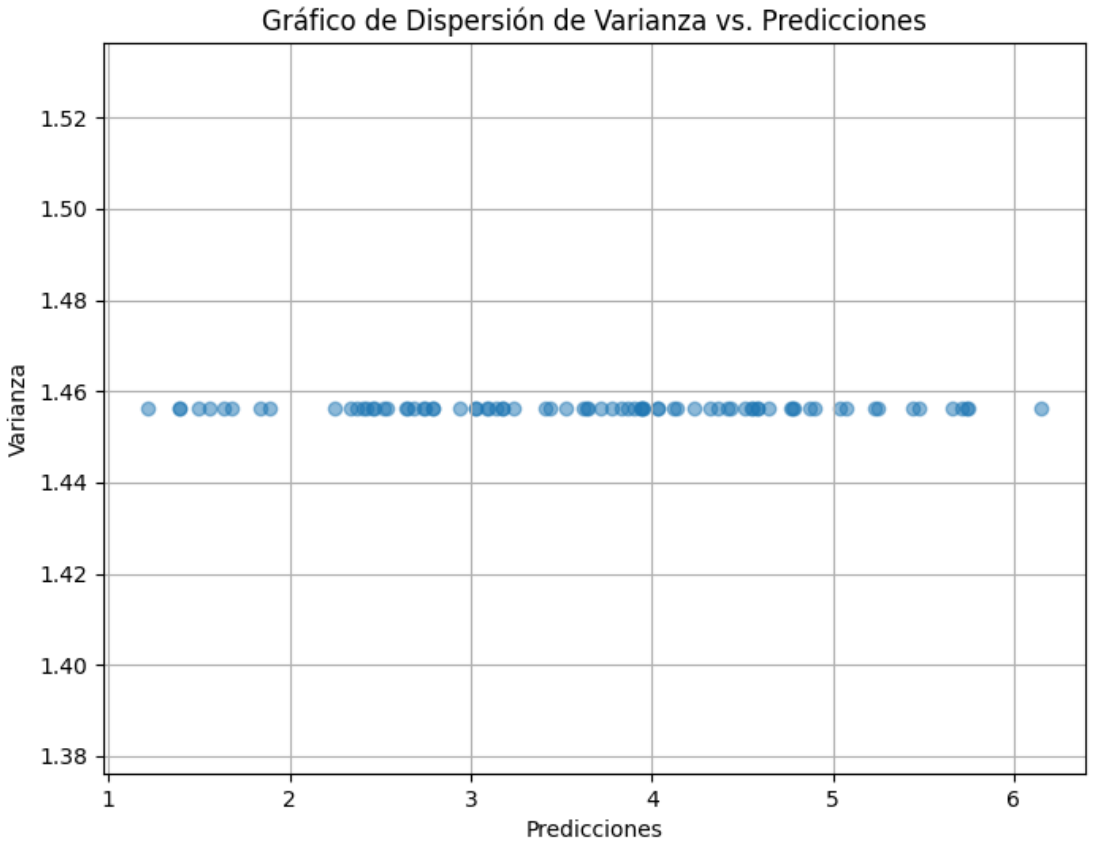
Diagnóstico de Sesgo (Bias)

El sesgo en un modelo de regresión lineal es un término crucial que representa el valor predicho por el modelo cuando todas las características independientes son cero. En nuestro análisis del modelo aplicado al conjunto de datos del Índice de Masa Corporal (IMC), hemos calculado un sesgo de aproximadamente 6.193. Esto significa que, incluso si todas las variables independientes, como altura, peso y género, fueran cero, nuestro modelo pronosticaría un valor de IMC de 6.193 como punto de partida. El sesgo influye en la ubicación vertical de la línea de regresión en nuestros gráficos, lo que afecta directamente nuestras predicciones. Ajustar adecuadamente el sesgo es fundamental para lograr predicciones precisas y un modelo que se adapte eficazmente a los datos reales.



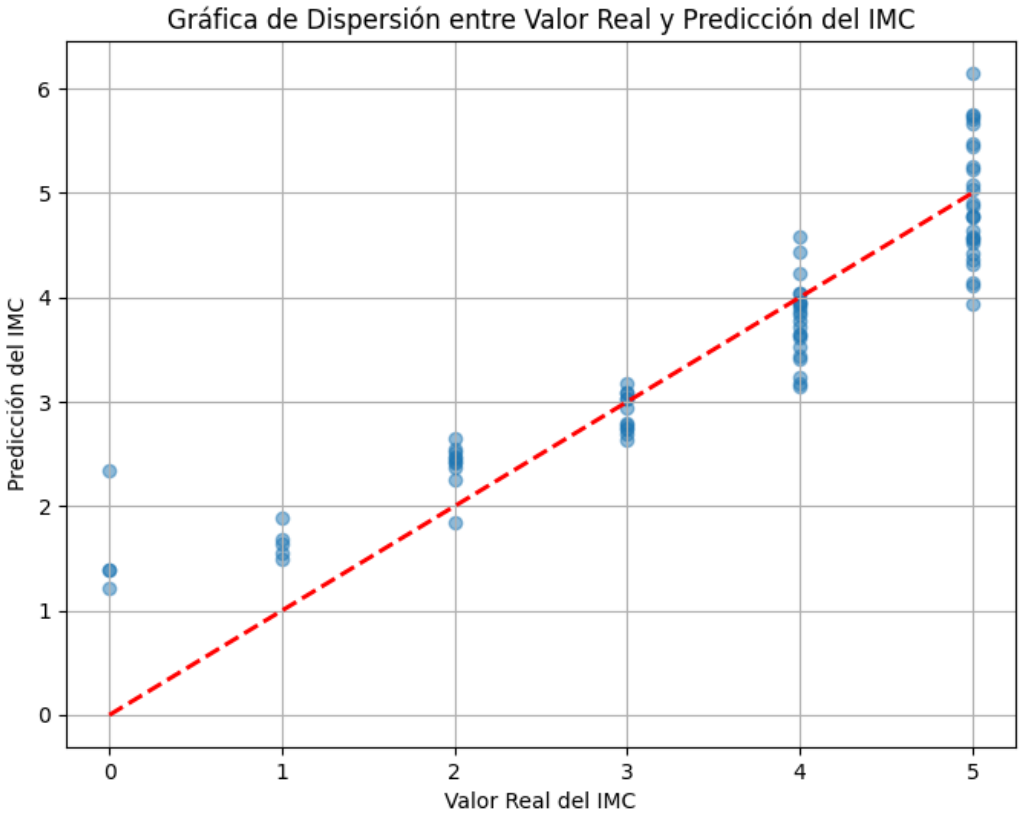
Diagnóstico de Varianza

La varianza del modelo de regresión lineal aplicado al conjunto de datos del Índice de Masa Corporal (IMC) es de aproximadamente 1.456, lo que indica una dispersión relativamente baja en las predicciones con respecto a la línea de regresión. Esta medida refleja la consistencia y precisión de las estimaciones del IMC generadas por el modelo. En otras palabras, las predicciones tienden a estar cercanas a la línea de mejor ajuste, lo que es un signo positivo de su capacidad de generalización. El gráfico de dispersión creado también respalda esta observación, ya que muestra que la varianza se mantiene en niveles bajos a medida que las predicciones se extienden a lo largo del eje horizontal. En resumen, la baja varianza indica que el modelo logra un equilibrio adecuado entre ajuste y generalización, lo que es esencial para su utilidad en la predicción precisa del IMC. Gráfica de varianza



Nivel de Ajuste del Modelo

Basado en las métricas de evaluación y la gráfica de dispersión, podemos concluir que el modelo de regresión lineal presenta un buen ajuste a los datos. El Mean Squared Error (MSE) de 0.356 y el Mean Absolute Error (MAE) de 0.459 son bajos, lo que indica que las predicciones del modelo son cercanas a los valores reales tanto en el conjunto de prueba como en el de validación. Además, el coeficiente de determinación R-squared de 0.834 es alto, lo que sugiere que aproximadamente el 83.4% de la variabilidad en el IMC se explica mediante el modelo. La gráfica de dispersión muestra que las predicciones siguen una línea diagonal, indicando que el modelo se ajusta de manera adecuada. No se observa una brecha significativa entre los puntos de entrenamiento, prueba y validación, lo que sugiere que el modelo no está subajustado (underfitting) ni sobreajustado (overfitting). En resumen, el modelo está bien ajustado (fit) y es capaz de generalizar de manera efectiva a datos no vistos, lo que respalda su capacidad predictiva.



Técnicas de ajuste

Transformación de Características

La transformación logarítmica es una técnica útil cuando se sospecha que la relación entre una característica y la variable objetivo no es lineal. En este caso, se aplicó una transformación logarítmica a la característica Weight antes de utilizarla en el modelo. Esto puede ser beneficioso porque:

1. **Ayuda a manejar valores extremos:** Si los datos originales contienen valores atípicos o extremadamente grandes en una característica, la transformación logarítmica puede ayudar a reducir la influencia de estos valores en el modelo.

2. **Linealidad:** Puede hacer que la relación entre la característica y la variable objetivo sea más lineal, lo que es una suposición común en los modelos de regresión lineal.

Usando el logartimo en Weight, se obtuvieron las siguientes métricas del modelo. Teniendo mejor precisión.

```python
Mean Squared Error (MSE): 0.24516865283091774
R-squared (R^2): 0.8858353188214586
Mean Absolute Error (MAE): 0.38267578875313574
```

Feature Scaling

La estandarización es una técnica esencial cuando trabajamos con características que tienen diferentes escalas o unidades de medida. En el conjunto de datos, las características Height y Weight tienen diferentes unidades y escalas. Estandarizar las características es importante por las siguientes razones:

1. **Elimina problemas de escala:** La estandarización asegura que todas las características tengan una media de 0 y una desviación estándar de 1, lo que facilita la comparación y la interpretación de los coeficientes del modelo.

2. **Mejora la convergencia:** Al estandarizar las características, es más probable que los algoritmos de optimización, como el descenso de gradiente, converjan más rápido y de manera más estable.

3. **Evita la dominancia de características:** Cuando las características tienen diferentes escalas, una característica con valores grandes podría dominar el impacto en la función objetivo, lo que puede llevar a resultados sesgados. La estandarización evita este problema.

Usando el feature scaling, no se obtuvo ninguna mejora del modelo. Teniendo los mismos resultados en las métricas.

```python
Mean Squared Error (MSE): 0.35626655313488403
```

```
R-squared (R^2): 0.8341017214738607
Mean Absolute Error (MAE): 0.459441534979837
```

Regresión Polinómica

La Regresión Polinómica es una técnica que se utiliza cuando se sospecha que la relación entre las características y la variable objetivo es no lineal. A diferencia de la Regresión Lineal, que asume una relación lineal, la Regresión Polinómica permite modelar relaciones de mayor grado, como relaciones cuadráticas o cúbicas. Se aplica mediante la introducción de características polinómicas, que son transformaciones de las características originales elevadas a potencias enteras.Las principales razones para utilizar la este método de regresión son las siguientes:

1. **Captura relaciones no lineales:** La Regresión Polinómica puede capturar relaciones más complejas entre las características y la variable objetivo al introducir características polinómicas. Esto permite modelar patrones no lineales en los datos.

2. **Flexibilidad en la modelización:** Al ajustar el grado del polinomio, puedes controlar la complejidad del modelo. Un polinomio de grado superior permite modelar relaciones más complejas, mientras que un polinomio de grado inferior es más simple.

3. **Mejora del ajuste:** En casos donde los datos no se ajustan bien a un modelo lineal simple, la Regresión Polinómica puede proporcionar un mejor ajuste y mejorar las métricas de evaluación, como el MSE y el R-squared.

4. **Interpretación de coeficientes:** Aunque la interpretación de los coeficientes en un modelo polinómico es más compleja que en un modelo lineal simple, aún se pueden analizar para comprender cómo las características influyen en la variable objetivo.

En los resultados obtenidos en la aplicación de la regresión polinómica se observa un aumento substancial en los valores de, así como una disminución de los errores en el modelo.

```python
Mean Squared Error (MSE): 0.16398797831892908
R-squared (R^2): 0.9236377283730248
Mean Absolute Error (MAE): 0.3082895154858788
```

Conclusiones

En este análisis exhaustivo del modelo de regresión lineal aplicado a un conjunto de datos centrado en el índice de masa corporal (IMC), se ha demostrado la importancia de seleccionar cuidadosamente un dataset adecuado y realizar un procesamiento previo de los datos, incluyendo la codificación de variables categóricas y la división de datos en conjuntos de entrenamiento y prueba. Las métricas de evaluación utilizadas, como el MSE, el R-squared y el MAE, han arrojado una luz precisa sobre la calidad de las predicciones del modelo, proporcionando información valiosa para su mejora. Además, se ha realizado un diagnóstico de sesgo y varianza, destacando la importancia de un sesgo razonable y una baja varianza para un buen ajuste del modelo. Por último, se exploraron técnicas de ajuste, con la regresión polinómica destacándose como una opción efectiva para modelar relaciones no lineales y mejorar significativamente la precisión del modelo en la predicción del IMC. En conjunto, este análisis brinda una visión completa y sólida del rendimiento y las posibles mejoras en un modelo de regresión lineal aplicado a datos de IMC.