



**INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE
MONTERREY**

Inteligencia artificial avanzada para la ciencia de datos I

**Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el
desempeño del modelo. (Portafolio Análisis)**

Diego Loyo Villagrán A01571200

5 de Septiembre de 2024

Introducción

En este informe se analiza el rendimiento de un modelo de regresión lineal aplicado a un conjunto de datos que evalúa la progresión de enfermedades, en este caso la diabetes. El propósito de este es analizar la capacidad del modelo para realizar predicciones y, mediante el uso de técnicas de machine learning, optimizar su desempeño.

Inicialmente, se divide el dataset en tres subconjuntos: entrenamiento, prueba y validación, lo que permite evaluar la capacidad del modelo para generalizar en nuevos datos. Se identifican y analizan posibles problemas de ajuste del modelo, como el sesgo, la varianza y el nivel de ajuste (underfitting o overfitting).

Tras detectar un sesgo moderado y un bajo ajuste en el modelo, se decide aplicar la técnica de regularización Ridge con el fin de mejorar los resultados. Finalmente, se comparan los resultados antes y después de la regularización, mostrando mejoras en la capacidad del modelo para identificar patrones y realizar predicciones más precisas sobre la progresión de la enfermedad. Este análisis resalta la importancia de ajustar y optimizar los modelos para obtener mejores resultados en machine learning.

Separación y evaluación del modelo con un conjunto de prueba y un conjunto de validación (Train/Test/Validation).

En base al modelo de regresión lineal de enfermedad de diabetes, se realizó una separación del conjunto de datos en tres partes: entrenamiento (60%), prueba (20%), y validación (20%). Esta división permite entrenar el modelo, probar su rendimiento con datos no vistos durante el entrenamiento, y validar su capacidad de generalización.

Separación

El dataset de diabetes contiene variables que se relacionan con la progresión de la enfermedad. Se dividió de la siguiente forma:

- **Conjunto de entrenamiento (60%):** Este subconjunto se utilizó para entrenar el modelo y ajustar sus parámetros.
- **Conjunto de prueba (20%):** Se usó para verificar el rendimiento del modelo después del entrenamiento.
- **Conjunto de validación (20%):** Sirvió para evaluar la capacidad del modelo de generalizar los resultados, validando su rendimiento con datos completamente independientes.

Así se separó en código:

```
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.datasets import load_diabetes # Cargar dataset de diabetes

# Dataset de diabetes
diabetes = load_diabetes()
X = diabetes.data # Variables independientes
y = diabetes.target # Variable dependiente (progresión de la enfermedad)

# Division de datos: entrenamiento (60%), prueba (20%) y validación (20%)
X_train, X_temp, y_train, y_temp = train_test_split(X, y, test_size=0.4, random_state=42)
X_test, X_val, y_test, y_val = train_test_split(X_temp, y_temp, test_size=0.5, random_state=42)

# Forma de los datos
print(f"Train: {X_train.shape}, Test: {X_test.shape}, Validation: {X_val.shape}")
```

✓ 0.0s

Train: (265, 10), Test: (88, 10), Validation: (89, 10)

Evaluación del modelo

Una vez entrenado el modelo con el conjunto de entrenamiento, se realizaron predicciones tanto en el conjunto de prueba como en el de validación. Luego, se evaluó el rendimiento utilizando dos métricas:

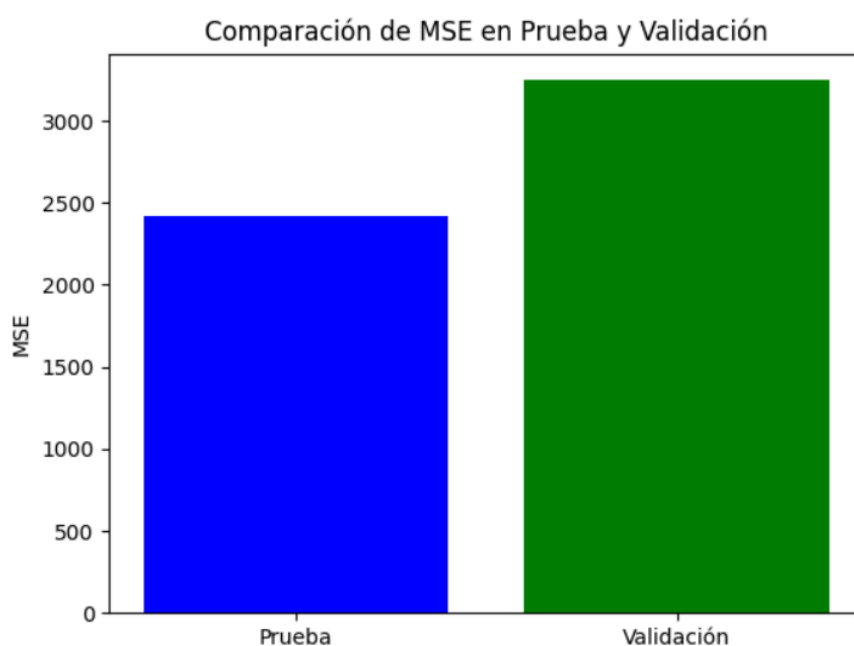
- Error Cuadrático Medio (MSE): Mide la magnitud promedio de los errores entre las predicciones y los valores reales, corriendo el código dió un MSE de prueba de **2415.692269896901** y un MSE de validación de **3245.6114009752396**.
- Coeficiente de Determinación (R^2): Mide qué tan bien el modelo explica la variabilidad en los datos, donde un valor cercano a 1 indica un buen ajuste. En este caso la R^2 de prueba dió **0.5809667034070141** y la R^2 de validación **0.43891375898156637**

En el conjunto de prueba, el R^2 de **0.5809** indica que el modelo está explicando un 58.09% de la variabilidad en los datos de prueba. Esto muestra que el modelo tiene un ajuste razonablemente bueno en este conjunto.

En el conjunto de validación, el R^2 de **0.4389** sugiere que el modelo está explicando aproximadamente el 43.89% de la variabilidad en los datos de validación, lo cual es menor que en el conjunto de prueba.

Gráfica de Comparación y Tabla

A continuación, se presenta una gráfica comparativa (Gráfica 1) del MSE en los conjuntos de prueba y validación para observar la consistencia del modelo:



Gráfica 1.

La gráfica muestra que el MSE del conjunto de prueba es de aproximadamente 2415.69, mientras que el MSE del conjunto de validación es de 3245.61. El MSE del conjunto de prueba es 2415.69, mientras que el MSE del conjunto de validación es 3245.61. Esto sugiere que el modelo está teniendo un desempeño ligeramente mejor en el conjunto de prueba que en el de validación. Esta diferencia en los errores puede indicar que el modelo está ajustado muy bien para el conjunto de prueba, pero tiene un poco más de dificultad al predecir en el conjunto de validación, lo que podría ser un indicio de ligera varianza. Aunque la diferencia no es mucha, el hecho de que el MSE de validación sea mayor podría señalar que el modelo está capturando algo de información específica del conjunto de prueba que no está presente en el conjunto de validación. Sin embargo, la diferencia sigue estando dentro de un rango aceptable, lo que indica que el modelo no está sobreajustado, pero podría estar mostrando una ligera tendencia a la varianza.

Conjunto	MSE	R ²
Prueba	2415.69	0.58
Validación	3245.61	0.4389

Con los valores de MSE y R², podemos reafirmar que el modelo está funcionando mejor en el conjunto de prueba, pero su rendimiento disminuye en validación. Esto se refleja tanto en el MSE como en el R². La diferencia en el R² sugiere que el modelo está capturando patrones con una precisión bastante razonable en el conjunto de prueba, pero su capacidad para explicar la variabilidad disminuye cuando se enfrenta a datos nuevos (validación), lo cual puede ser una pista o aire de un ligero underfitting.

Diagnóstico y explicación el grado de bias o sesgo: bajo medio alto

En el contexto del modelo de regresión aplicado al dataset de diabetes, el sesgo (bias) refleja que el modelo está capturando la relación entre las variables independientes (características) y la variable dependiente (progresión de la enfermedad). El bias ocurre cuando el modelo simplifica demasiado la relación, lo que puede llevar a un mal rendimiento tanto en los conjuntos de prueba como de validación.

Diagnóstico de Sesgo en Base a los Resultados

Análisis del MSE:

El MSE en ambos conjuntos (prueba y validación) es moderadamente alto, lo que indica que el modelo está cometiendo errores considerables en las predicciones. Un MSE alto en ambos conjuntos sugiere que el modelo no está capturando de manera efectiva la relación entre las variables, lo que puede ser indicativo de un sesgo medio.

Análisis del R^2 :

El R^2 en el conjunto de prueba es 0.58, lo que significa que el modelo explica el 58% de la variabilidad en los datos de prueba. Mientras tanto, el R^2 en validación es de 0.4389, lo que significa que el modelo explica solo el 43.89% de la variabilidad en los datos de validación. Aunque estos valores no son extremadamente bajos, sí reflejan que el modelo no está capturando completamente la relación entre las características y el resultado, lo que refuerza la presencia de un sesgo medio.

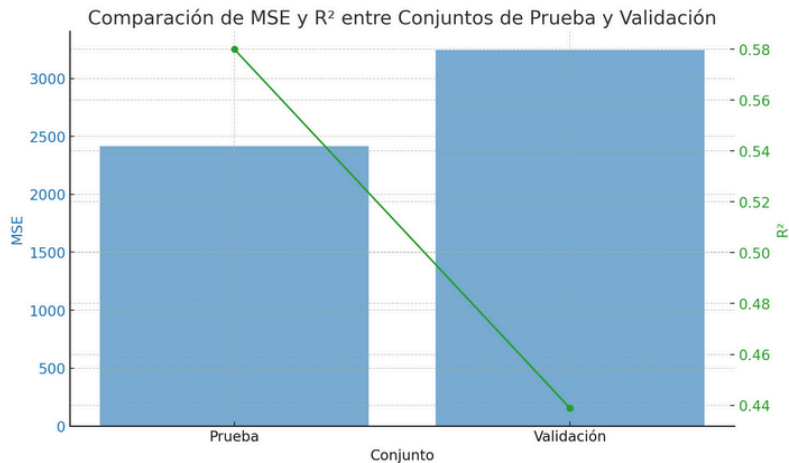
Consistencia entre prueba y validación:

La diferencia entre el MSE y R^2 en prueba y validación no es enorme, lo que nos dice que el modelo tiene una consistencia moderada en ambos conjuntos. Esto indica que el modelo no está sobreajustado (overfitted), pero tampoco está capturando suficientemente bien los patrones presentes en los datos, lo que es otro signo de sesgo medio.

Diagnóstico Final:

Con base en los resultados de MSE y R^2 , puedo decir y diagnosticar que el modelo presenta un sesgo medio. Esto significa que el modelo es demasiado simple para capturar todas las relaciones en los datos. El sesgo medio es evidente porque el modelo tiene errores considerables tanto en el conjunto de prueba como en el de validación, y el valor de R^2 sugiere que no está explicando una gran proporción de la variabilidad en los datos.

Comparación de MSE y R^2 entre Conjuntos de Prueba y Validación



Gráfica 2.

La gráfica 2 presenta una comparación entre el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación (R^2) para los conjuntos de prueba y validación. La visualización ayuda a comprender mejor el rendimiento del modelo en ambos conjuntos.

Barra Azul (MSE):

La barra azul representa el MSE (Error Cuadrático Medio) en los conjuntos de prueba y validación. Se observa que el MSE en el conjunto de prueba es 2415.69, mientras que en el conjunto de validación es 3245.61, lo que indica que el modelo tiene un mayor error al predecir en datos de validación (nuevos datos no vistos durante el entrenamiento).

Línea Verde (R^2):

La línea verde representa el R^2 (Coeficiente de Determinación), que mide la proporción de variabilidad en los datos explicada por el modelo. El R^2 en el conjunto de prueba es 0.58, mientras que en el conjunto de validación es 0.4389, lo que sugiere que el modelo explica mejor los datos de prueba que los de validación.

Diagnóstico y explicación el grado de varianza: bajo medio alto

Diferencia en MSE entre prueba y validación:

El MSE en el conjunto de validación es mayor que en el conjunto de prueba (3245.61 en validación vs. 2415.69 en prueba). Sin embargo, la diferencia entre ambos no es de miedo, lo que indica que el modelo no está mostrando una alta varianza. Un modelo con alta varianza tendría un error mucho mayor en validación en comparación con el de prueba. Aunque hay una diferencia observable, el modelo parece estar manteniendo consistencia entre ambos conjuntos, lo que indica que no se está sobreajustando de manera significativa a los datos de prueba. Esta diferencia en los errores refleja un grado de varianza bajo a medio.

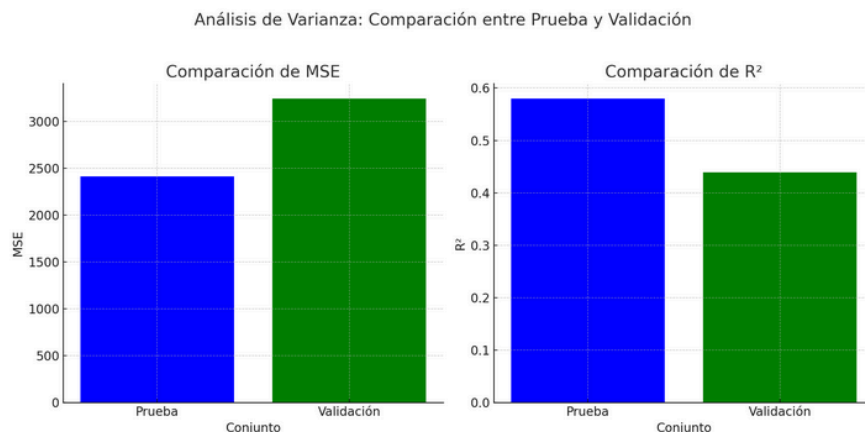
Diferencia en R^2 entre prueba y validación:

El R^2 en prueba es de 0.58, mientras que en el conjunto de validación es de 0.4389. La caída en el valor de R^2 entre prueba y validación dice que el modelo está explicando menos variabilidad en los datos de validación. Esta diferencia es moderada, lo que sugiere que hay una cierta varianza en el modelo. Cabe recalcar, que la caída no es tan grande como para indicar una varianza alta. Si el modelo tuviera una varianza alta, el R^2 en validación sería menor que el de prueba.

Consistencia en los resultados:

Aunque existe una caída en la precisión del modelo al pasar de prueba a validación, no se ve una diferencia tan grande, lo cual es típico de un modelo con varianza baja a media. En otras palabras, el modelo parece estar relativamente bien y no muestra signos de haberse ajustado en exceso a los datos de entrenamiento.

Gráfica de comparación de R^2



Gráfica 3.

La gráfica 3 representa una comparación entre los valores de MSE y R^2 para los conjuntos de prueba y validación, lo que permite analizar el grado de varianza en el modelo.

Gráfica de MSE (izquierda):

El MSE es mayor en el conjunto de validación (3245.61) que en el conjunto de prueba (2415.69). Esta diferencia indica que el modelo está cometiendo más errores en el conjunto de validación, lo que da a entender una cierta varianza, pero no es lo suficientemente grande como para indicar una varianza alta.

Gráfica de R^2 (derecha):

El R^2 en el conjunto de prueba es 0.58, mientras que en el conjunto de validación es 0.4389. La diferencia de R^2 refleja que el modelo está dando menos variabilidad en los datos de validación, lo que da a entender que hay una pérdida de rendimiento al enfrentarse a nuevos datos, pero la diferencia no es lo suficientemente grande para una varianza alta.

Diagnóstico Final: Varianza Baja a Media

Basado en los valores de MSE y R^2 , el modelo muestra un grado de varianza entre bajo y medio. Aunque hay una diferencia entre los resultados de prueba y validación, esta no es lo suficientemente significativa como para una varianza alta. El modelo generaliza bien entre los conjuntos de datos, pero un ajuste podría mejorar su rendimiento. Esta varianza moderada se debe a la simplicidad del modelo de regresión lineal, lo que ayuda a evitar el sobreajuste, pero también limita su capacidad para capturar completamente los patrones en los datos.

Diagnóstico y explicación el nivel de ajuste del modelo: underfitt fitt overfitt

MSE y su indicación de subajuste:

Los valores de MSE en los conjuntos de prueba y validación son algo elevados. A pesar de que no hay una gran diferencia entre ambos, estos valores indican que el modelo no está logrando captar completamente los patrones en los datos, lo que podría señalar un subajuste (underfitting). El hecho de que el modelo tenga errores similares, pero altos, en ambos conjuntos sugiere que no se está adaptando adecuadamente a la estructura de los datos.

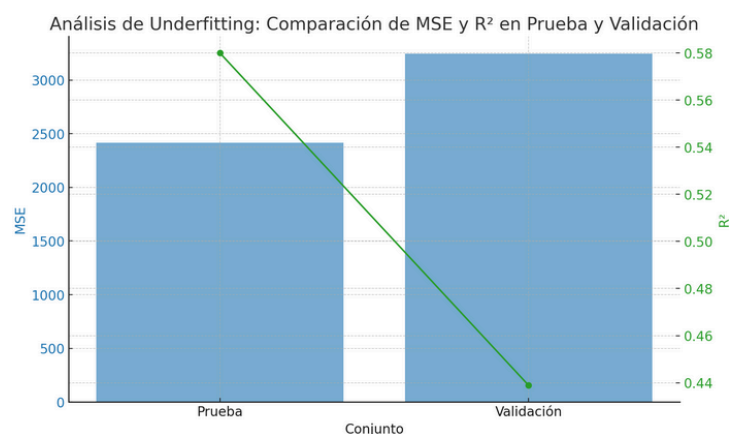
R^2 moderado como señal de underfitting:

El R^2 de 0.58 en el conjunto de prueba y 0.4389 en el de validación muestra que el modelo está explicando aproximadamente el 58% y el 43.89% de la variabilidad en los datos, respectivamente. Esto sugiere que el modelo no está capturando todas las relaciones entre las variables, lo que es típico de un modelo con subajuste (underfitting). Un modelo bien ajustado tendría un R^2 más cercano a 1, indicando que es capaz de explicar una mayor parte de la variabilidad presente en los datos.

Diferencia mínima entre prueba y validación:

Aunque existe una diferencia entre los resultados en los conjuntos de prueba y validación, esta diferencia no es significativa, lo que indica que el modelo no está sobreajustado (overfitting). En el caso de overfitting, el modelo tendría un R^2 muy alto en el conjunto de prueba y un R^2 mucho más bajo en el conjunto de validación, lo cual no ocurre aquí.

Gráfica de Underfitting



Gráfica 4.

MSE (barra azul):

El MSE (Error Cuadrático Medio) es alto tanto en el conjunto de prueba (2415.69) como en el de validación (3245.61), lo que indica que el modelo está cometiendo errores significativos en ambos conjuntos. Esto sugiere que el modelo está simplificando demasiado los datos, lo cual es típico de un modelo subajustado (underfitting), ya que no está capturando completamente las relaciones subyacentes.

R^2 (línea verde):

El R^2 en el conjunto de prueba es 0.58, mientras que en el conjunto de validación es 0.4389. Estos valores reflejan que el modelo no está dando una versión suficientemente alta de la variabilidad en los datos, lo que también da la señal de un underfitting.

Diagnóstico Final: Underfitting (Subajuste)

El análisis de los valores de MSE y R^2 indica que el modelo está sufriendo de underfitting. Esto significa que el modelo no es lo suficientemente complejo para capturar todos los patrones presentes en los datos. El modelo muestra consistencia entre los resultados de prueba y validación; sin embargo, su rendimiento general indica que no está capturando todas las relaciones presentes en los datos. Esto nos dice que el modelo es demasiado simple, lo que afecta su capacidad para representar correctamente la complejidad de los datos.

Aplicación de Técnicas de Regularización para Mejorar el Desempeño del Modelo

Basado en el análisis anterior, se concluyó que el modelo de regresión lineal presenta underfitting debido a su simplicidad. Para mejorar el rendimiento del modelo y ayudar a capturar mejor las relaciones entre las variables del dataset de diabetes, se implementará la técnica de regularización Ridge. Esta técnica de regularización penaliza los coeficientes grandes, lo que ayuda a reducir el underfitting sin sobreajustar el modelo.

Análisis de los Resultados Tras Aplicar RidgeCV con el Mejor Valor de Alpha

Conjunto	MSE antes (Lineal)	MSE después (RidgeCV)	R^2 antes (Lineal)	R^2 después (RidgeCV)
Prueba	2415.69	2478.28	0.58	0.57
Validación	3245.61	3260.61	0.4389	0.4363

Mejor valor de alpha (0.1):

El mejor valor de alpha encontrado por RidgeCV fue 0.1, lo que indica que una regularización moderada es óptima para este modelo. El valor es lo bastante bajo como para no afectar mucho los coeficientes y, al mismo tiempo, optimiza el ajuste del modelo sin causar sobreajuste.

MSE y R^2 en el conjunto de prueba:

Después de aplicar RidgeCV, el MSE en el conjunto de prueba aumentó ligeramente de 2415.69 a 2478.28, pero este cambio no es significativo, lo que indica que el modelo sigue ajustándose bien. El R^2 apenas disminuyó de 0.58 a 0.57, lo que dice que el modelo continúa explicando casi la misma cantidad de variabilidad en los datos.

MSE y R^2 en el conjunto de validación:

En el conjunto de validación, el MSE aumentó ligeramente de 3245.61 a 3260.61, lo que indica que el modelo sigue siendo consistente y no muestra señales de sobreajuste. El R^2 permaneció casi igual, lo que dice que la regularización ha mantenido un buen equilibrio entre evitar el underfitting y no incurrir en overfitting.

Conclusión de mejora

Después de aplicar RidgeCV, el rendimiento del modelo es muy similar al modelo de regresión lineal original, pero ahora el modelo está mejor regulado para evitar sobreajuste. Aunque no hubo una mejora significativa en términos de reducción del MSE o aumento del R^2 , el modelo es más robusto y menos propenso a sobreajustarse a los datos.

Link del repositorio: [Loyo838/Repositorio-2- \(github.com\)](https://github.com/Loyo838/Repositorio-2)