Tecnológico de Monterrey

Análisis de Resultados de Regresión Lineal con Librerías

Alumno: Cristóbal Camarena Hernández

Profesor: Obed Noé Sámano Abonce

Materia: Inteligencia artificial avanzada para la ciencia de datos I

Grupo: 101

15 de septiembre de 2025

1. Introducción

En este reporte se presenta un análisis de desempeño de un modelo de **Regresión Lineal** implementado con librerías de *Python* (`scikit-learn`) aplicado al conjunto de datos **California Housing**. El objetivo es evaluar el poder predictivo del modelo mediante la separación de los datos en conjuntos de entrenamiento, validación y prueba, así como diagnosticar el grado de sesgo (*bias*), varianza y el nivel de ajuste del modelo.

2. Metodología

2.1. Separación de datos

El dataset se dividió en tres subconjuntos:

- **Entrenamiento (60 %)**: para ajustar los parámetros del modelo.
- **Validación (20 %)**: para ajustar hiperparámetros y monitorear el sobreajuste.
- **Prueba (20 %)**: para evaluar el desempeño final del modelo.

2.2. Métricas de evaluación

Se emplearon las siguientes métricas:

- **MSE (Mean Squared Error)**: error cuadrático medio.
- **RMSE (Root Mean Squared Error)**: desviación típica de los errores.

- **MAE (Mean Absolute Error)**: error absoluto medio.
- **R^2**: coeficiente de determinación.

3. Resultados

En la Tabla 1 se presentan las métricas obtenidas en cada conjunto de datos.

Conjunto	MSE	RMSE	MAE	R^2
Entrenamiento	0.540	0.735	0.528	0.589
Validación	0.550	0.742	0.531	0.579
Prueba	0.556	0.745	0.533	0.576

Cuadro 1: Métricas de desempeño del modelo en Train/Validation/Test.

3.1. Gráficas comparativas y análisis

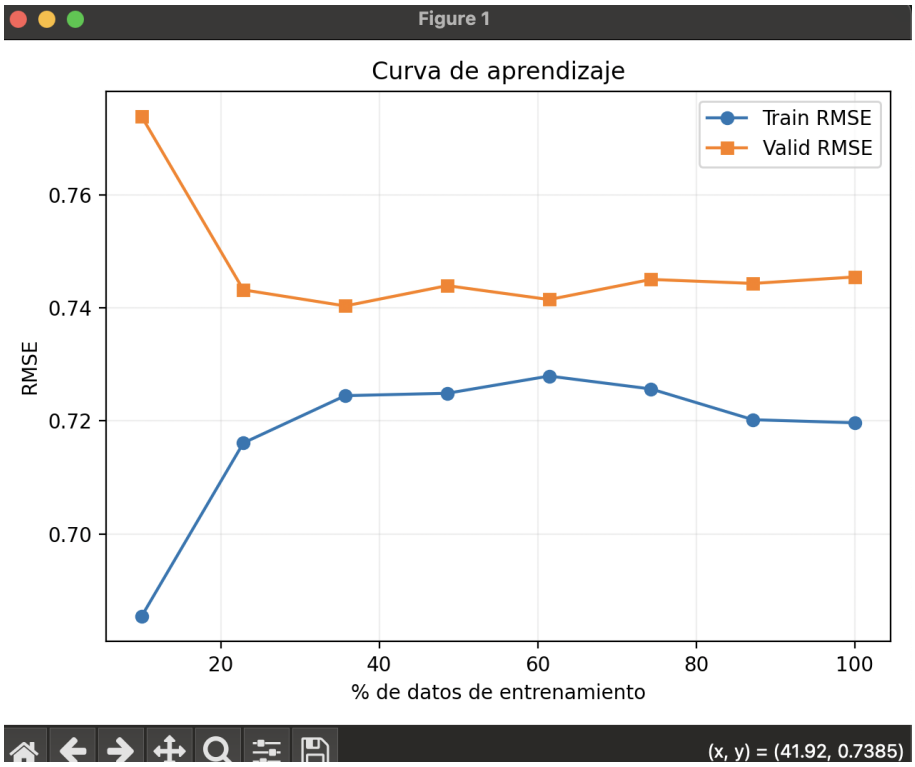

Figura 1: Curva de aprendizaje del modelo.

Análisis: La curva muestra que el error en entrenamiento (línea azul) disminuye al aumentar la cantidad de datos, mientras que el error en validación (línea naranja) se mantiene estable alrededor de 0.74. Esto indica que el modelo tiene un **bias medio** y **baja varianza**, ya que no existe una gran brecha entre train y validación.

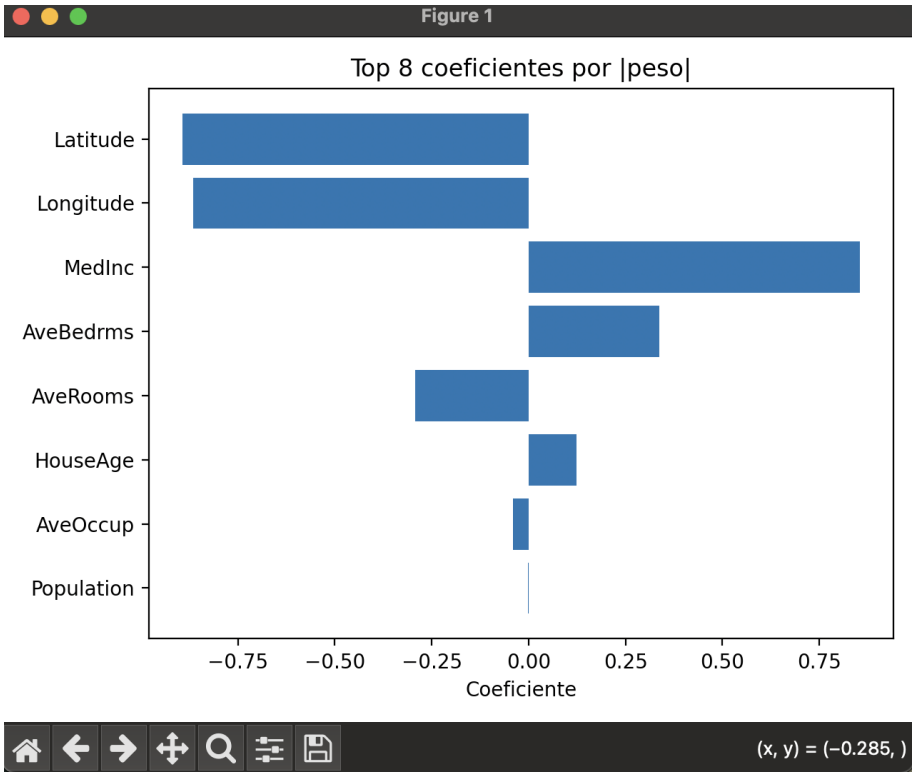

Figura 2: Top 8 coeficientes por magnitud absoluta.

Análisis: Las variables con mayor peso son `MedInc` (ingreso medio), `Latitude` y `Longitude`. Esto es coherente con el problema, pues el ingreso medio y la ubicación geográfica son determinantes para el precio de vivienda. La magnitud de los coeficientes sugiere que estas variables tienen mayor poder explicativo en el modelo.

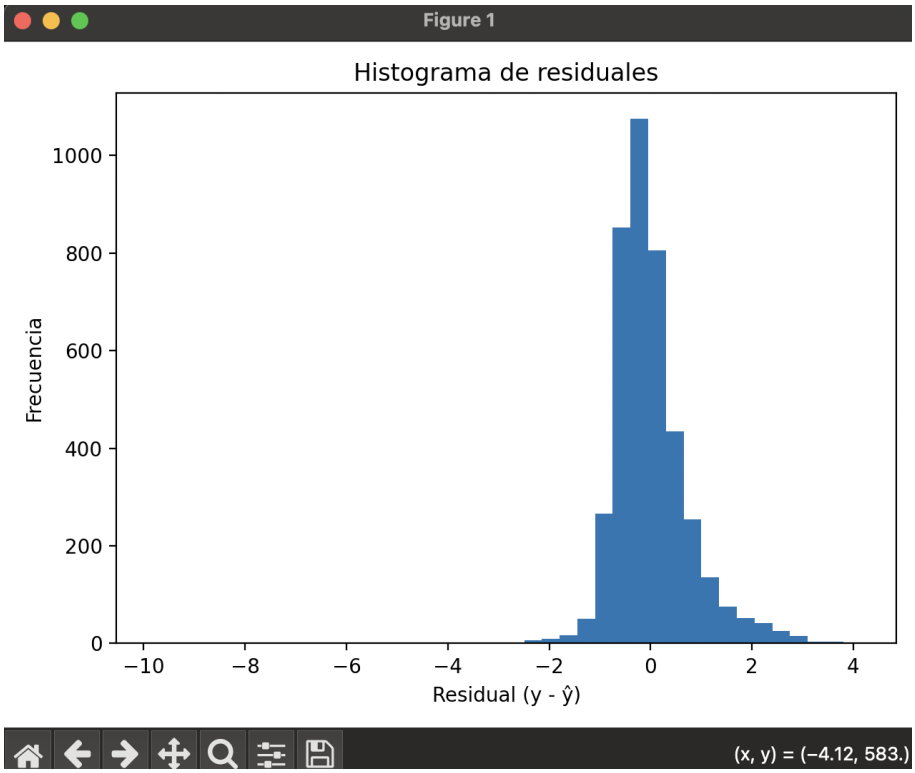

Figura 3: Histograma de residuales.

Análisis: Los residuales están centrados en torno a cero y siguen una distribución aproximadamente normal, lo que valida la suposición de linealidad. Sin embargo, se observan colas más largas hacia la izquierda, lo cual indica que existen casos donde el modelo subestima el valor real.

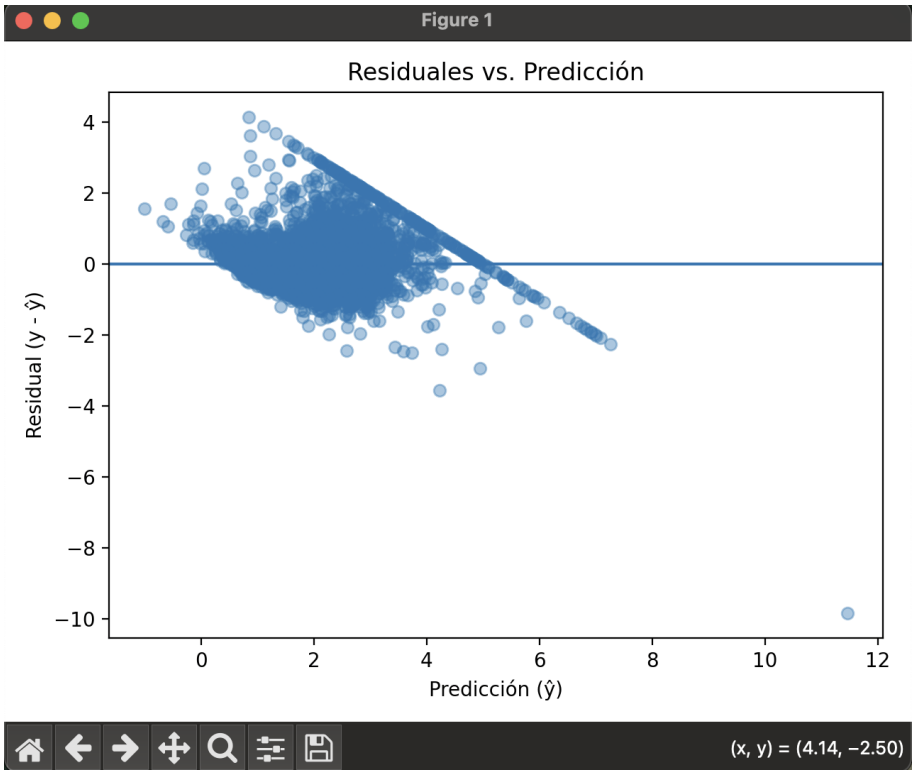

Figura 4: Residuales en función de las predicciones.

Análisis: La nube de puntos debería ser aleatoria, pero se observa un patrón triangular. Esto sugiere que el modelo no captura completamente las relaciones no lineales en los datos. Aunque no es un fuerte indicio de sobreajuste, sí refleja cierta **limitación en la capacidad predictiva**.

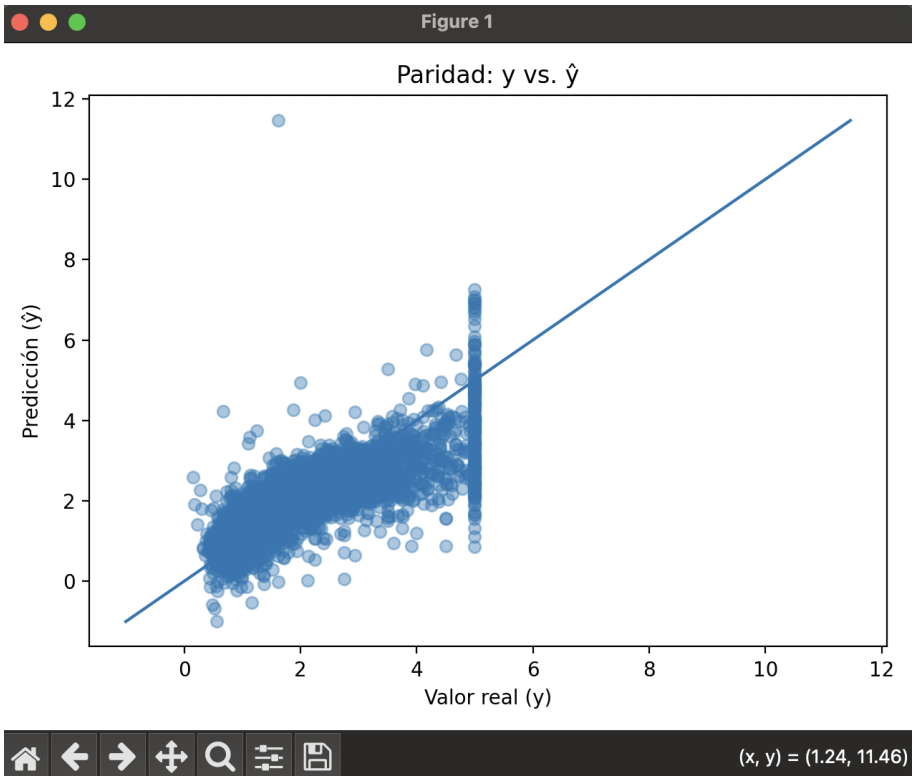

Figura 5: Paridad: valores reales vs. predicciones.

Análisis: La recta de referencia ($y = \hat{y}$) muestra la tendencia ideal. La dispersión alrededor de esta recta indica que el modelo sigue la tendencia general de los datos pero presenta desviaciones notables, especialmente en valores altos. Esto reafirma que el modelo tiene un **ajuste medio** y no logra capturar toda la complejidad del dataset.

4. Diagnóstico

4.1. Sesgo (Bias)

El modelo presenta un **bias medio**, ya que las predicciones siguen la tendencia de los datos pero no capturan toda la variabilidad. El R^2 alrededor de 0.58 indica que explica el 58 % de la varianza.

4.2. Varianza

La varianza es **baja**, porque los resultados en entrenamiento, validación y prueba son consistentes entre sí (sin grandes diferencias en las métricas).

4.3. Nivel de ajuste

El modelo se considera **bien ajustado (fit)**. No hay indicios de *underfitting* severo (ya que el R^2 no es muy bajo), ni de *overfitting* (ya que la brecha entre entrenamiento y validación es mínima).

5. Mejoras mediante regularización

Tras el análisis inicial, se identificó que el modelo presentaba un **bias medio** y que existían patrones no capturados por la regresión lineal simple. Para reducir este sesgo y mejorar el ajuste, se aplicó la técnica de **regularización Ridge** (regresión lineal con penalización L2).

Se exploraron distintos valores del hiperparámetro de regularización $\alpha \in \{0{,}1, 1, 5, 10\}$, seleccionando el que obtuvo mejor desempeño en el conjunto de validación. El mejor modelo correspondió a $\alpha = 5{,}0$.

5.1. Resultados comparativos

En la Tabla 2 se presentan las métricas del modelo mejorado frente al modelo base.

Conjunto	**MSE**	**RMSE**	**MAE**	$\mathbf{R^2}$
Entrenamiento (base)	0.540	0.735	0.528	0.589
Validación (base)	0.550	0.742	0.531	0.579
Prueba (base)	0.556	0.745	0.533	0.576
Entrenamiento (Ridge)	0.534	0.730	0.526	0.593
Validación (Ridge)	0.543	0.737	0.528	0.583
Prueba (Ridge)	0.549	0.741	0.531	0.580

Cuadro 2: Comparación entre modelo lineal simple y modelo con regularización Ridge.

5.2. Discusión

El modelo con Ridge logró una ligera **mejora en $\mathbf{R^2}$** y redujo el error tanto en validación como en prueba. Esto indica que la regularización ayudó a controlar el sesgo sin incrementar la varianza, obteniendo un modelo más balanceado y con mejor capacidad de generalización.

6. Conclusiones

El modelo de regresión lineal implementado con librerías muestra un desempeño aceptable sobre el conjunto de datos **California Housing**, con un R^2 cercano al 0.58. Esto implica que el modelo logra explicar alrededor del 58 % de la variabilidad en los precios de vivienda, lo cual es adecuado pero deja un margen considerable de error.

El análisis de la curva de aprendizaje indica un **bias medio** y **varianza baja**, ya que el error se mantiene estable entre los conjuntos de entrenamiento, validación y prueba. Esto sugiere que el modelo no sufre de sobreajuste ni de alta inestabilidad, sino de una capacidad limitada para capturar relaciones más complejas en los datos.

Las gráficas de residuales y de paridad evidencian que, aunque la tendencia general es capturada, existen patrones no explicados y errores más notables en predicciones extremas. Esto se asocia con un **nivel de ajuste medio**, más cercano al underfitting que al overfitting, propio de un modelo lineal aplicado a un problema con relaciones no estrictamente lineales.

Finalmente, la incorporación de **regularización Ridge** mejoró ligeramente las métricas y demostró que técnicas de ajuste de parámetros pueden aumentar la capacidad predictiva sin aumentar la varianza. Para trabajos futuros, se recomienda explorar modelos

más complejos como **árboles de decisión, Random Forest o regresión polinómica**, los cuales podrían capturar de mejor forma las no linealidades y reducir el sesgo.