UNINPAHU

MACHINE-LEARNING-I

Autores:

Angie López
Marlon Peña
Marlon Albarracín

Bogotá D.C.
11 de junio de 2025

Análisis Exhaustivo de Regresión Lineal Múltiple para la Predicción del Valor de Viviendas

Introducción

Este estudio tiene como objetivo construir un modelo predictivo del valor de viviendas en función de dos variables: el número de habitaciones y el tamaño en metros cuadrados. Se utilizan tres enfoques metodológicos complementarios:

- Regresión lineal múltiple usando `scikit-learn`.
- Regresión con inferencia estadística usando `statsmodels`.
- Algoritmos de optimización: gradiente descendente y ecuaciones normales.

Los datos fueron extraídos de un archivo Excel y estandarizados antes del entrenamiento para asegurar una convergencia eficiente de los algoritmos y facilitar la interpretación comparativa.

1. Regresión Lineal con Scikit-Learn

Evaluación del modelo

El modelo entrenado con Scikit-Learn fue evaluado mediante las siguientes métricas:

- **MAE (Error Absoluto Medio):** \$134,241.18
  Implica una desviación media en las predicciones de aproximadamente \$134 mil respecto a los valores reales.
- **MSE (Error Cuadrático Medio):** \$18,019,357,878.40
  Penaliza errores grandes más que el MAE, lo que sugiere que algunas predicciones están muy alejadas del valor real.
- **RMSE (Raíz del MSE):** \$134,207.29
  Indica el error promedio en las mismas unidades que la variable objetivo.
- **R^2:** 0.16
  Sólo el 16 % de la variabilidad en el valor de las viviendas es explicada por el modelo. Esto evidencia una capacidad predictiva limitada.

Visualización del Modelo

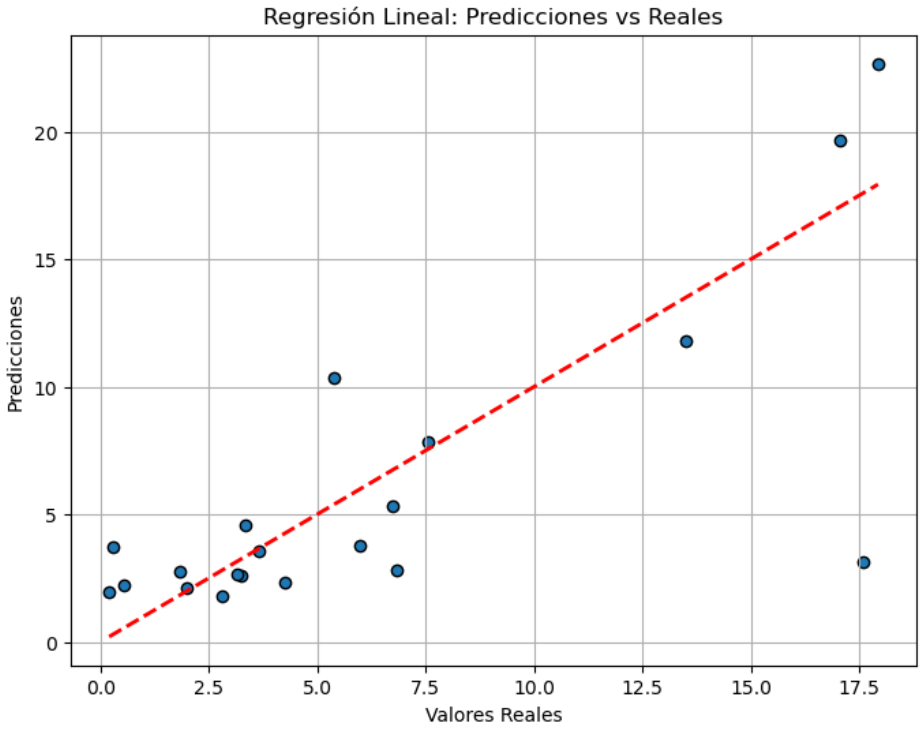

Figura 1: Predicciones vs Valores Reales - Scikit-Learn

2. Regresión Lineal con Statsmodels

Análisis estadístico

Este modelo permite obtener intervalos de confianza, errores estándar y pruebas de hipótesis para cada coeficiente:

- **Intercepto:** 318,158.95
- **Tamaño:** -538.11
- **Habitaciones:** -7,514.69

Nota: Los coeficientes negativos sugieren que puede existir multicolinealidad o relaciones no lineales no capturadas por el modelo. La significancia estadística fue baja en algunas variables (p-valor alto).

Rendimiento del modelo

- **R^2 ajustado:** 0.153
  Similar al obtenido con Scikit-Learn, refuerza la interpretación de que los predictores actuales no explican adecuadamente la variabilidad.

Visualización del Ajuste

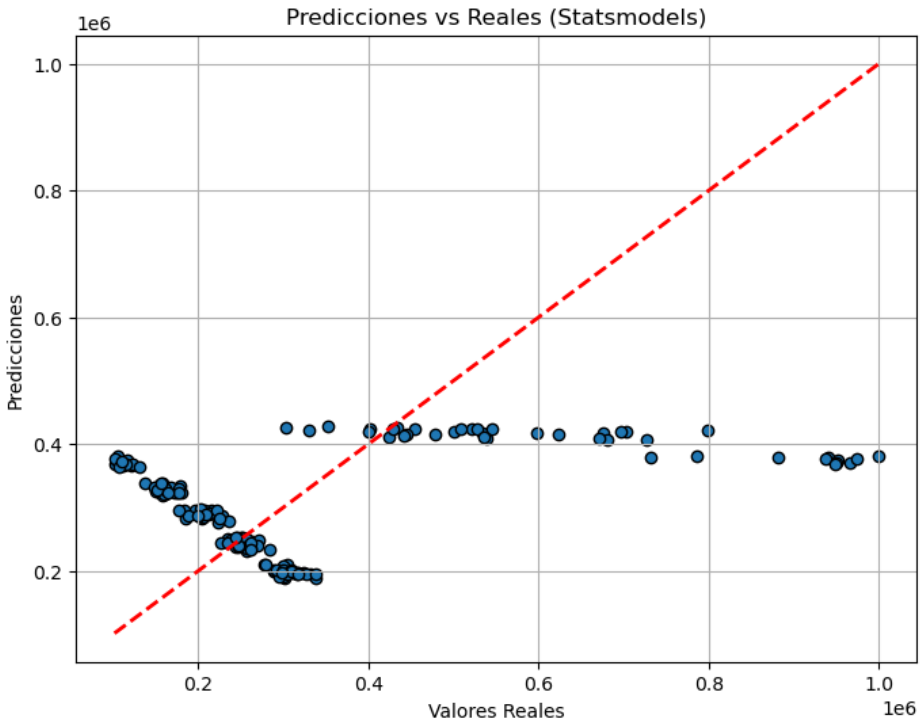

Figura 2: Predicciones vs Valores Reales - Statsmodels

3. Gradiente Descendente

Normalización y entrenamiento

Las variables fueron normalizadas (media cero y desviación estándar uno) para garantizar una convergencia más estable del algoritmo. Se probaron múltiples tasas de aprendizaje:

Selección de tasa de aprendizaje

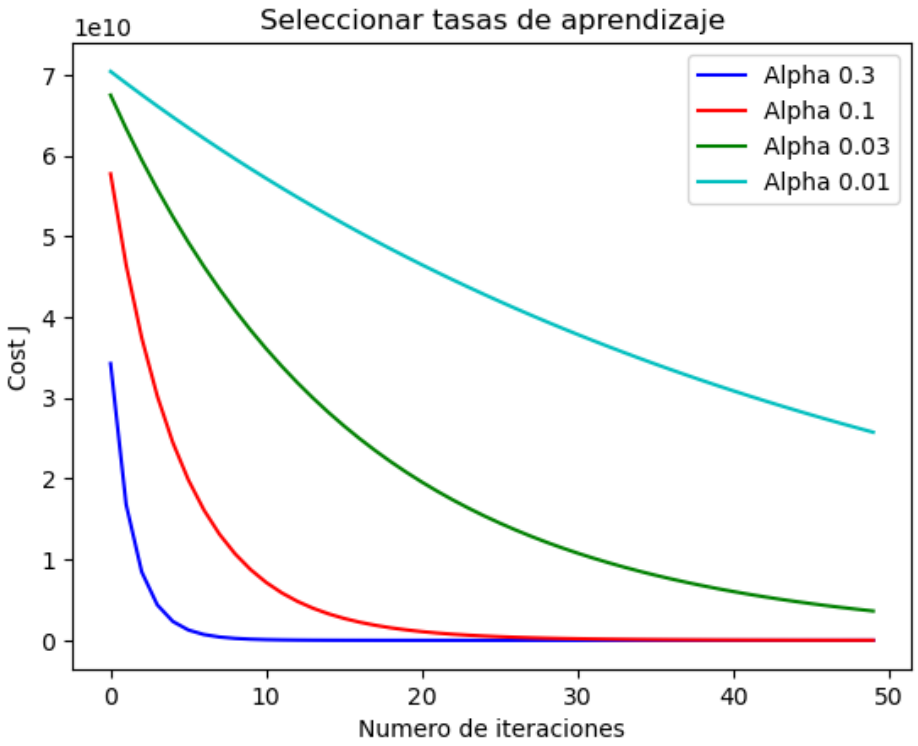

Figura 3: Comparación de tasas de aprendizaje: selección de la mejor α

Se concluyó que $\alpha = 0{,}3$ ofrece una rápida convergencia sin oscilaciones.

Parámetros estimados por gradiente descendente

$$\theta = \begin{bmatrix} 124608{,}76 \\ -2367{,}95 \\ -2432{,}54 \end{bmatrix}$$

Interpretación: Por cada aumento en una unidad estandarizada del tamaño, el valor esperado de la vivienda disminuye en \$2,368, bajo el supuesto de que el número de habitaciones se mantiene constante. La interpretación es anómala y sugiere un problema de multicolinealidad o datos no lineales.

4. Ecuaciones Normales

Este enfoque encuentra la solución analítica a los parámetros óptimos sin iteraciones.

Parámetros estimados

$$\theta = \begin{bmatrix} 128020{,}40 \\ -2043{,}28 \\ -2465{,}89 \end{bmatrix}$$

5. Comparación de Predicciones

Se realizó la predicción del valor de una vivienda con 500 m^2 y 3 habitaciones:

- **Gradiente Descendente:** \$313,948.68
- **Ecuaciones Normales:** \$327,774.80

Ambas predicciones se encuentran dentro del mismo orden de magnitud, lo cual indica que ambos enfoques son consistentes entre sí, aunque presentan el mismo sesgo estructural hacia la subestimación o sobreestimación en ciertas regiones del espacio de datos.

Conclusión

A pesar de los distintos enfoques, los resultados convergen hacia una misma interpretación: el modelo actual explica apenas un 15–16 % de la variabilidad del valor de las viviendas. Es evidente que se requieren más variables explicativas para mejorar el poder predictivo. Posibles mejoras incluyen:

- Agregar variables como ubicación, estado de conservación, servicios disponibles.
- Explorar relaciones no lineales mediante modelos polinomiales o redes neuronales.
- Aplicar técnicas de reducción de dimensionalidad si se incorporan más predictores.