TP2: Regresión del valor medio de casas en distritos de California (21Co2025)

Tomas Corteggiano, Rodrigo Goñi

8 de junio de 2025

1. Obtener la correlación entre los atributos y entre los atributos y el target.

1.1. ¿Qué atributo tiene mayor correlación lineal con el target?

El atributo con la mayor correlación lineal en valor absoluto con el target `MedHouseVal` es `MedInc` (ingreso medio), que presenta un coeficiente de aproximadamente 0.65, indicando una **fuerte relación lineal positiva**.

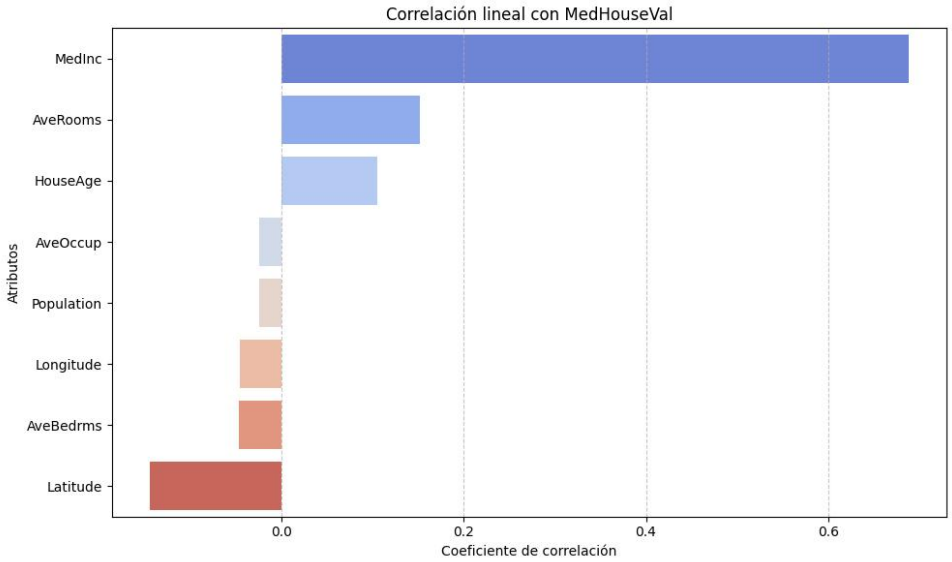

Figura 1: Correlación lineal de los atributos con el target (`MedHouseVal`). Se destaca `MedInc` como el atributo con la correlación positiva más alta.

1.2. ¿Cuáles atributos parecen estar más correlacionados entre sí? ¿Se pueden calcular los coeficientes de correlación o representarlos gráficamente mediante un mapa de calor?

Al analizar las relaciones entre los propios atributos, se identificaron dos pares con correlaciones lineales significativas:

- `Latitude y Longitude:` Presentan una correlación negativa muy fuerte (-0.92). Esto es esperable, ya que las coordenadas geográficas en California siguen un patrón diagonal.
- `AveRooms y AveBedrms:` Muestran una correlación positiva fuerte (0.80), lo cual es lógico, pues a mayor número de habitaciones promedio por vivienda, es probable que también aumente el número promedio de dormitorios.

Los coeficientes de correlación se pueden calcular y representar gráficamente mediante un mapa de calor, como se muestra en la Figura 2.

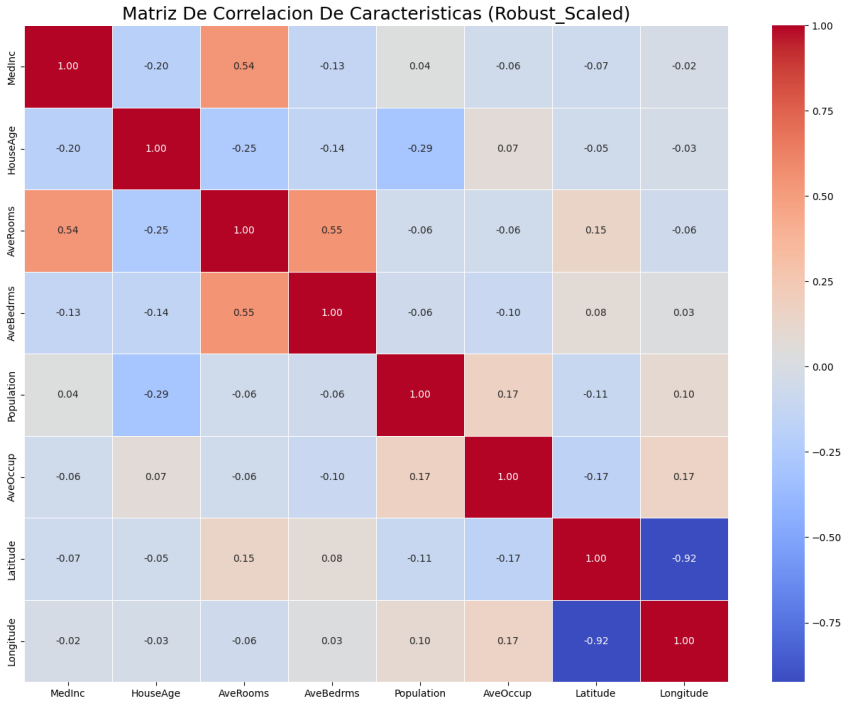

Figura 2: Matriz de correlación de las características. Los colores intensos (rojos y azules) indican las correlaciones más fuertes, tanto positivas como negativas.

2. Graficar los histogramas de los distintos atributos y del target.

2.1. ¿Qué forma presentan los histogramas?

Los histogramas (ver Figura 3) muestran en su mayoría distribuciones asimétricas, con una cola pronunciada hacia la derecha.

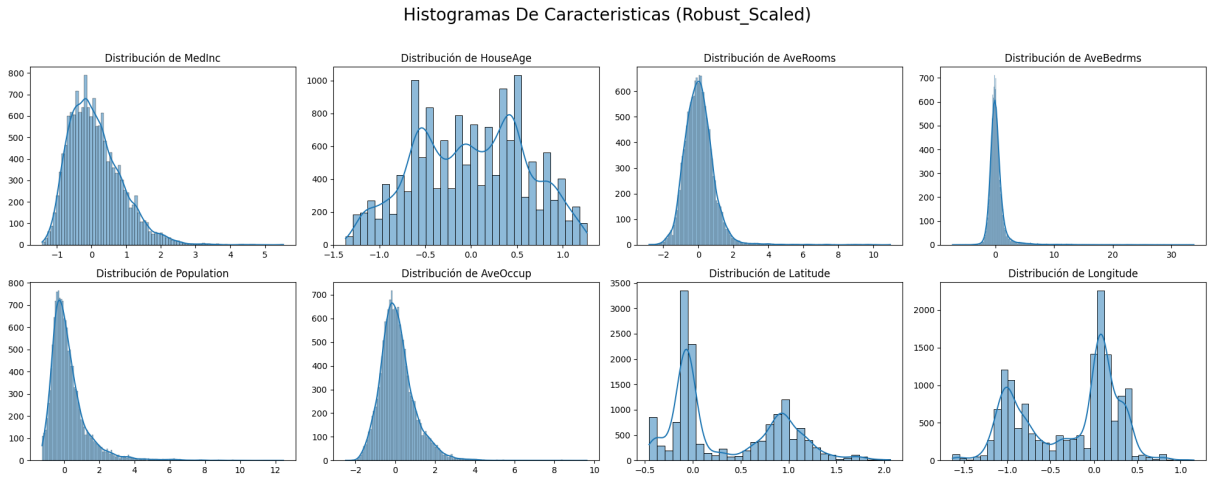

Figura 3: Histogramas de las características y el target. La mayoría de las variables presentan asimetría positiva.

2.2. ¿Alguno muestra una distribución similar a una campana que sugiera una distribución gaussiana, sin necesidad de realizar pruebas de hipótesis?

El único atributo cuya distribución se asemeja a una campana de Gauss es `AveRooms`. Un hallazgo importante es el efecto de çapping"(limitación o tope) en ciertos atributos:

- `MedHouseVal` **(Target):** La distribución muestra una acumulación abrupta de valores en 5, lo que sugiere que todos los precios de vivienda por encima de un cierto umbral fueron agrupados en ese valor máximo (ver Figura 4).
- `HouseAge` **(Antigüedad de la vivienda):** Presenta un efecto similar, con una gran cantidad de viviendas agrupadas en el valor máximo de 50 años.

3. Análisis de Distribución de Atributos y Manejo del Capping

Se generaron histogramas y gráficos de dispersión para visualizar la distribución de cada atributo y del target, así como las relaciones entre ellos. Este análisis reveló la presencia de un fenómeno conocido como çapping.[en] dos variables clave.

3.0.1. Observación y Manejo del Çapping"

Un hallazgo importante es el efecto de çapping"(limitación o tope) en ciertos atributos, lo que se traduce en una acumulación artificial de valores en el extremo superior de su rango. Esto fue observado en:

- `MedHouseVal` **(Target):** La distribución muestra una acumulación abrupta de valores en 5.0 (ver Figura 4). Esto sugiere que todos los precios de vivienda por encima de un cierto umbral fueron agrupados en ese valor máximo durante la recolección de datos.

- `HouseAge` **(Antigüedad de la vivienda):** Presenta un efecto similar, con una gran cantidad de viviendas agrupadas en el valor máximo de 50 años.

Este çapping"puede afectar significativamente el rendimiento del modelo, ya que le impide aprender y predecir valores por encima de estos topes artificiales, introduciendo un sesgo.

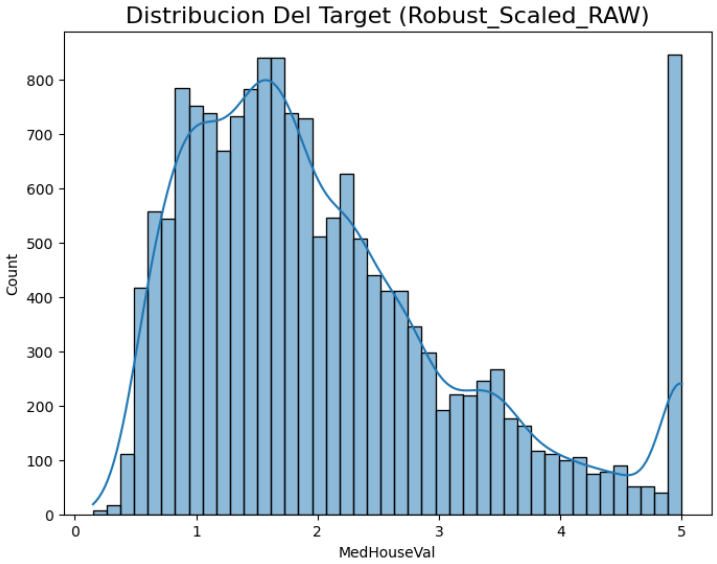

Figura 4: Distribución del target (`MedHouseVal`) mostrando el efecto de çapping.[en] el valor 5.0.

Para mitigar el impacto negativo de este çappingz permitir que el modelo capture mejor la variabilidad real de los datos, se decidió **eliminar los registros** donde el target (`MedHouseVal`) era igual a 5.0 y donde la antigüedad de la casa (`HouseAge`) era igual a 50 años. Esta estrategia de filtrado se aplicó después de la aplicación del `RobustScaler`.

3.0.2. Análisis de Relación Características vs. Target (Pre y Post-Filtrado)

Para visualizar el impacto del çappingz la efectividad del filtrado, se analizaron los gráficos de dispersión de las características frente al target en dos etapas: antes y después de la eliminación de los valores tope.

3.0.3. Datos Pre-Filtrado (`Robust_Scaled_RAW`)

La Figura 5 muestra los gráficos de dispersión de cada característica frente al target `MedHouseVal` después de aplicar un `RobustScaler` pero **antes de la eliminación de los valores capped**.

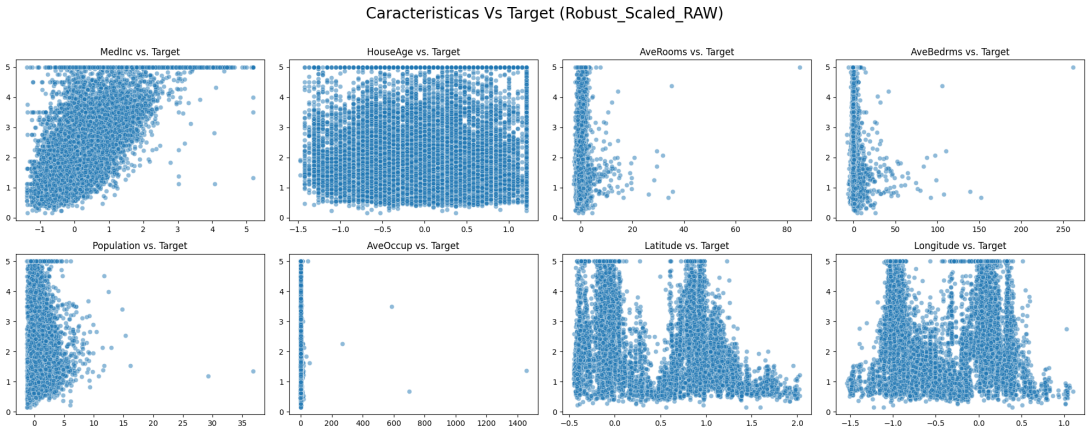

Figura 5: Gráficos de dispersión de cada característica vs. el target (`MedHouseVal`) con datos `Robust_Scaled_RAW`, previo al filtrado de valores capped. Se observa claramente la línea horizontal de puntos densos en el valor 5.0 del target debido al çapping".

En esta visualización, es evidente la línea horizontal de puntos densos en la parte superior de casi todos los gráficos, correspondiente al valor de 5.0 del target. Este patrón confirma la presencia del çappingz su efecto limitante en la distribución del target.

3.0.4. Datos Post-Filtrado (`Robust_Scaled`)

La Figura 6 presenta los mismos gráficos de dispersión, pero esta vez con los datos **después de haber eliminado los registros** donde el target era 5.0 o la antigüedad de la casa era 50 años.

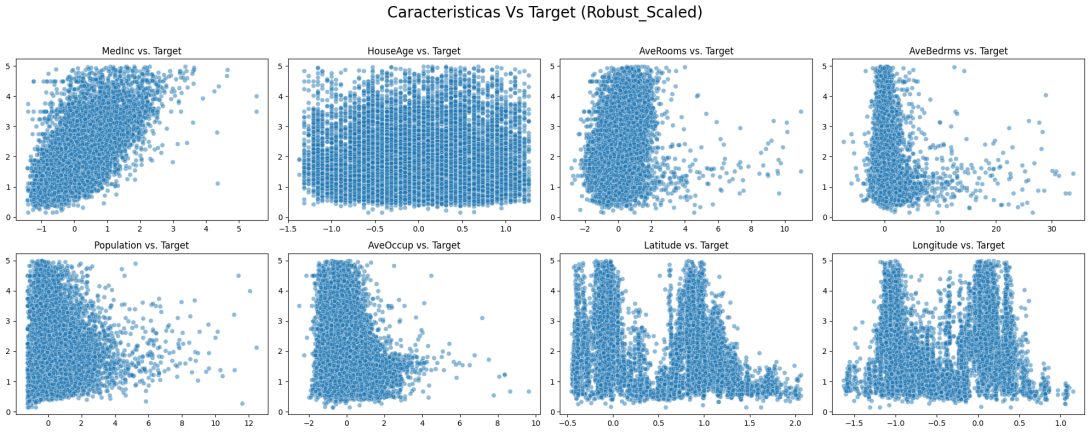

Figura 6: Gráficos de dispersión de cada característica vs. el target (`MedHouseVal`) con datos `Robust_Scaled` **después de aplicar el filtrado de valores capped**. La ausencia de la línea horizontal en el valor 5.0 del target indica que el filtrado fue efectivo.

Al comparar la Figura 5 con la Figura 6, se observa claramente la **desaparición de la línea horizontal de puntos en el valor 5.0 del target**. Esto demuestra que el filtrado fue efectivo para eliminar los valores tope artificiales.

4. Calcular una regresión lineal utilizando todos los atributos.

4.1. Con el conjunto de entrenamiento, calcular la varianza total de los datos y la varianza explicada por el modelo.

Se entrenó un modelo de Regresión Lineal utilizando todos los atributos con un preprocesamiento de `RobustScaler`, que es menos sensible a outliers. La Varianza Explicada (R^2) del modelo fue de **0.58**, lo que significa que el modelo logra explicar aproximadamente el 58 % de la variabilidad del precio de las viviendas. La fórmula de la Varianza Explicada es:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

donde y_i son los valores reales, $\hat{y}_i$ son las predicciones y $\bar{y}$ es la media de los valores reales.

4.2. ¿Está el modelo capturando adecuadamente el comportamiento del target? Fundamente su respuesta.

Aunque un R^2 de 0.58 es un resultado aceptable, indica que una porción significativa de la varianza no es capturada por el modelo. Los gráficos de diagnóstico (Figura 7) muestran que, si bien hay una tendencia lineal, los residuos no se distribuyen de manera completamente aleatoria, sugiriendo que el modelo podría mejorarse. Por lo tanto, el modelo no está capturando *adecuadamente* todo el comportamiento del target, pero sí una parte sustancial.

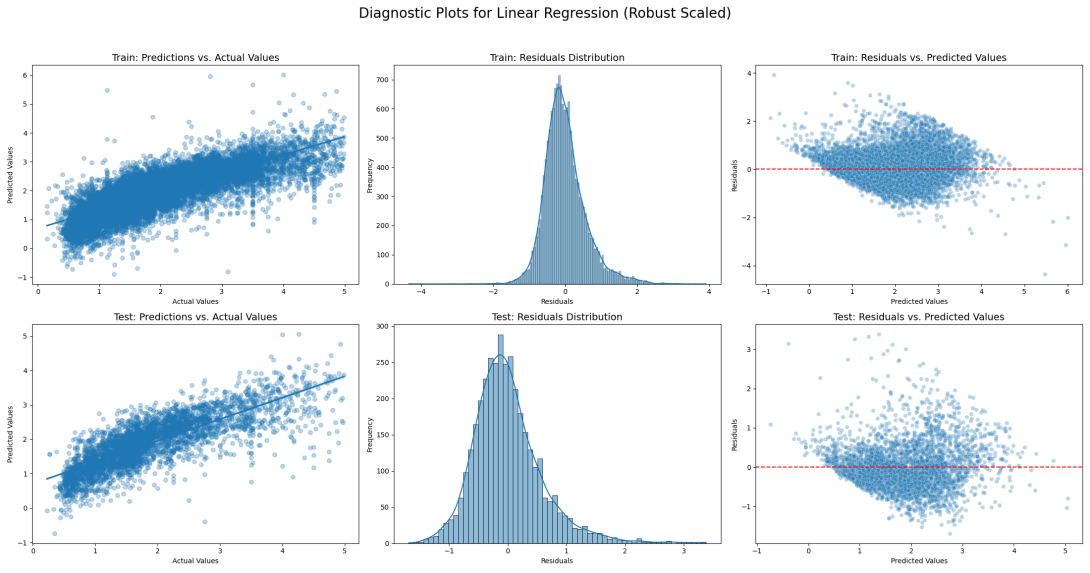

Figura 7: Gráficos de diagnóstico para la Regresión Lineal (Robust Scaled). Muestran la relación entre valores predichos y reales, y la distribución de los residuos.

5. Calcular las métricas de MSE, MAE y R^2 sobre el conjunto de evaluación.

Estas métricas se calcularán más adelante en la sección de comparación de modelos.

6. Crear una regresión de Ridge.

6.1. Usar validación cruzada de 5 folds y tomar como métrica el MSE.

Para intentar mejorar el modelo y controlar la posible colinealidad y sobreajuste, se implementó una Regresión de Ridge. Se utilizó validación cruzada de 5 folds para encontrar el hiperparámetro de regularización `alpha` óptimo, tomando el MSE como métrica.

6.2. Buscar el mejor valor de α en el rango [0, 12.5].

El valor de `alpha` que minimizó el Error Cuadrático Medio (MSE) fue aproximadamente **9.5** (ver Figura 8) dentro del rango especificado.

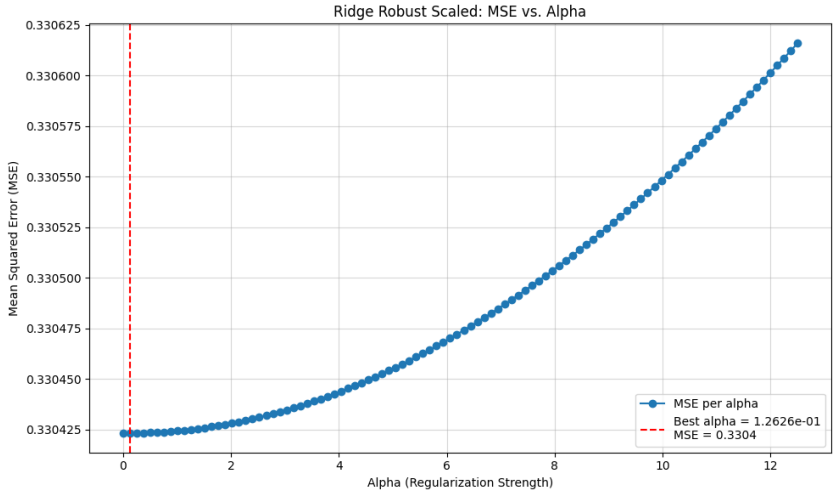

Figura 8: Evolución del MSE en función de α para la Regresión de Ridge. El punto mínimo de la curva indica el α óptimo.

6.3. Graficar el MSE en función de α.

La Figura 8 muestra la gráfica del MSE en función de α. Con este `alpha` óptimo, se entrenó el modelo final de Ridge. Sus gráficos de diagnóstico (Figura 9) son muy similares a los de la Regresión Lineal, lo que indica que no hay un cambio drástico en el comportamiento general del modelo.

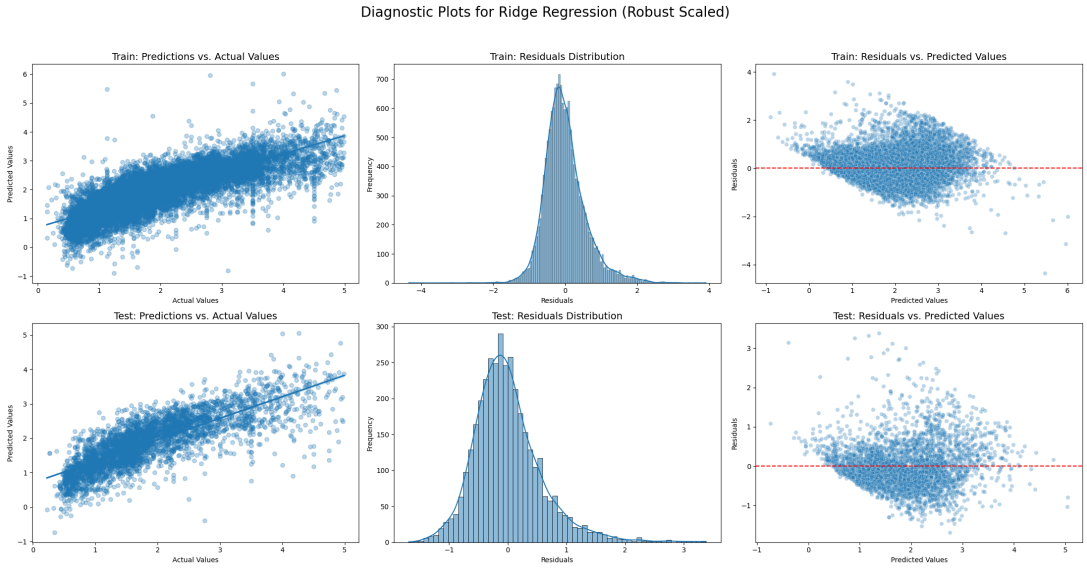

Figura 9: Gráficos de diagnóstico para la Regresión de Ridge con α óptimo.

7. Comparar los resultados obtenidos entre la regresión lineal y la mejor regresión de Ridge, evaluando el conjunto de prueba.

7.1. ¿Cuál de los dos modelos obtiene mejores resultados en términos de MSE y MAE? ¿Poseen suficiente diferencia como para indicar si uno es mejor que el otro?

Finalmente, se compararon las métricas de rendimiento de ambos modelos en el conjunto de prueba:

- **Error Cuadrático Medio (MSE):**
  - Regresión Lineal: 0.5478
  - Regresión de Ridge: **0.5477**
- **Error Absoluto Medio (MAE):**
  - Regresión Lineal: 0.5283
  - Regresión de Ridge: **0.5281**

El modelo de **Regresión de Ridge obtuvo resultados ligeramente mejores** en ambas métricas (MSE y MAE), aunque la diferencia es marginal. Esta pequeña mejora **no justifica un cambio de modelo si la simplicidad es prioritaria**, ya que la diferencia es insignificante.

7.2. ¿Qué tipo de error podría haberse reducido?

La regularización de Ridge está diseñada para reducir el *error de varianza*, que ocurre cuando el modelo se ajusta demasiado al ruido de los datos de entrenamiento (sobreajuste), especialmente en presencia de atributos correlacionados. Aunque la mejora fue mínima, demuestra que la técnica funciona en la dirección esperada de penalizar coeficientes grandes para crear un modelo potencialmente más generalizable.

Anexo: Pipeline Experimental y Resultados Detallados

En este anexo se presenta el flujo de trabajo experimental completo llevado a cabo para el desarrollo y la evaluación de los modelos de regresión. El proceso abarca desde la limpieza y preparación inicial de los datos hasta la implementación de técnicas avanzadas de ingeniería de características y enfoques de modelado alternativos. Cada etapa fue documentada y sus resultados, registrados para una comparativa exhaustiva.

A. Fase 1: Carga y Limpieza de Datos

El punto de partida fue la carga del dataset original, que consta de 20,640 muestras. Se aplicó un proceso de limpieza secuencial para mitigar el efecto de valores atípicos y artefactos de la recolección de datos.

- **Dataset Original:** 20,640 muestras.
- **Paso 1: Eliminación de Capping en Target.** Se filtraron las muestras donde la variable objetivo `MedHouseVal` presentaba un valor de tope superior (igual a 5), resultando en la eliminación de 992 registros.
- **Paso 2: Eliminación de Capping en Features.** Se procedió a eliminar 1,093 muestras donde la característica `HouseAge` mostraba un valor de tope (52.0).
- **Paso 3: Tratamiento de Outliers Generales.** Se utilizó el método del Rango Intercuartílico (IQR) para identificar y eliminar 98 outliers restantes en el conjunto de datos.

El tamaño final del dataset limpio utilizado para la mayoría de los experimentos fue de **18,457 muestras**. Tras la limpieza, se realizó un Análisis Exploratorio de Datos (EDA) sobre los datos crudos (`Raw`) y se generaron visualizaciones clave (matrices de correlación, histogramas, etc.), guardadas en el directorio `output_plots/eda_plots/`.

B. Fase 2: Modelos Base sobre Datos Limpios sin Escalar

Para establecer una línea base de rendimiento, se entrenaron tres modelos de regresión sobre los datos limpios pero sin ningún tipo de escalamiento de características.

- Regresión Lineal

- Regresión Ridge (con búsqueda de hiperparámetro α)
- Regresión Lasso (con búsqueda de hiperparámetro α)

Los resultados, muy similares entre los tres modelos, se resumen en la Tabla 1. Para los modelos Ridge y Lasso, la búsqueda con validación cruzada ('GridSearchCV') encontró valores óptimos de α muy cercanos a cero ($\alpha_{Ridge} \approx 1{,}39$, $\alpha_{Lasso} = 0{,}0001$), indicando una penalización mínima y un comportamiento casi idéntico al de la Regresión Lineal estándar.

Cuadro 1: Resultados de modelos sobre datos limpios sin escalar (conjunto de test).

Modelo	**MSE**	**MAE**	**R^2**
Linear Regression	0.3454	0.4336	0.6286
Ridge Regression	0.3454	0.4336	0.6286
Lasso Regression	0.3454	0.4337	0.6285

C. Fase 3: Evaluación del Impacto del Escalamiento

Se investigó el efecto de dos estrategias de escalamiento de características sobre el rendimiento de los modelos.

C.1. Escalamiento con StandardScaler

Las características fueron normalizadas utilizando `StandardScaler`, que centra los datos en una media de 0 y una desviación estándar de 1. Los resultados en el conjunto de test (Tabla 2) fueron virtualmente idénticos a los obtenidos sin escalar, lo cual es esperado para estos modelos de regresión cuando no hay regularización fuerte.

C.2. Escalamiento con RobustScaler

Se aplicó `RobustScaler`, una técnica resistente a outliers que escala los datos utilizando el rango intercuartílico. De manera similar al caso anterior, los resultados (Tabla 2) no mostraron una variación significativa respecto a la línea base.

Cuadro 2: Resultados en Test con StandardScaler y RobustScaler.

Modelo (StandardScaler)	**MSE**	**MAE**	**R^2**
Linear Regression	0.3454	0.4336	0.6286
Ridge Regression	0.3454	0.4336	0.6286
Lasso Regression	0.3454	0.4336	0.6286
Modelo (RobustScaler)	**MSE**	**MAE**	**R^2**
Linear Regression	0.3454	0.4336	0.6286
Ridge Regression	0.3454	0.4336	0.6286
Lasso Regression	0.3454	0.4336	0.6286

D. Fase 4: Ingeniería de Características

Con el objetivo de mejorar la capacidad predictiva de los modelos, se aplicaron técnicas de ingeniería de características para capturar relaciones no lineales.

- **Transformación Logarítmica:** Se aplicó la transformación $log(1+x)$ a las características con distribuciones asimétricas: `Population`, `AveOccup` y `MedInc`.
- **Características Polinómicas:** Se generaron características de interacción y de grado 2 para las variables más influyentes: `MedInc`, `HouseAge`, `Latitude` y `Longitude`, expandiendo el número de predictores a 18.

Posteriormente, se entrenaron los modelos sobre este nuevo conjunto de características, aplicando nuevamente los escaladores `StandardScaler` y `RobustScaler`. Los resultados (Tabla 3) muestran una **mejora sustancial** en todas las métricas. El R^2 en el conjunto de test aumentó de $\approx 0{,}62$ a $\approx 0{,}654$.

Cuadro 3: Resultados en Test con Features Log+Poly y escalamiento.

Modelo (Log+Poly + StandardScaler)	**MSE**	**MAE**	**R^2**
Linear Regression	0.3214	0.4103	0.6544
Ridge Regression	0.3214	0.4103	0.6544
Lasso Regression	0.3332	0.4232	0.6418
Modelo (Log+Poly + RobustScaler)	**MSE**	**MAE**	**R^2**
Linear Regression	0.3214	0.4103	0.6544
Ridge Regression	0.3214	0.4103	0.6544
Lasso Regression	0.3347	0.4248	0.6401

Durante el ajuste de los modelos Lasso sobre las características extendidas, se observaron advertencias de convergencia (`ConvergenceWarning`), lo que indica que el problema de optimización es más complejo en este espacio de características.

E. Fase 5: Experimentos de Control y Alternativos

E.1. Control: Modelado sobre Datos Crudos

Para validar la importancia del proceso de limpieza, se realizó un experimento de control entrenando los modelos sobre el dataset original completo (20,640 muestras), aplicando únicamente `RobustScaler`. Los resultados fueron notablemente inferiores ($R^2 \approx 0{,}576$), lo que confirma el impacto positivo de la limpieza de datos realizada en la Fase 1.

E.2. Modelo Alternativo: Enfoque de Dos Etapas

Se exploró un modelo híbrido para abordar el problema del capping en la variable objetivo.

1. **Etapa 1 (Clasificador):** Un modelo clasificador que predice si una vivienda tiene el valor máximo (`MedHouseVal == 5`) o no.
2. **Etapa 2 (Regressor):** El mejor modelo de regresión obtenido en la Fase 4, entrenado solo con los datos sin capping.

En la predicción, si el clasificador predice "valor máximo", se asigna 5. De lo contrario, se utiliza el modelo de regresión. Este enfoque arrojó el mejor rendimiento global, con un **R^2 de 0.6553** y un **MAE de 0.4098** en el conjunto de test. Se observó una advertencia de que el clasificador solo predecía una clase en el conjunto de test, lo cual es esperable ya que todos los datos de test habían sido previamente filtrados para no contener valores ‘y == 5‘.

F. Resumen Comparativo Final

La Tabla 4 consolida los resultados en el conjunto de test de los experimentos más relevantes. El modelo de dos etapas, seguido de cerca por la Regresión Lineal/Ridge con ingeniería de características, demuestra ser la estrategia más efectiva. Esto subraya la importancia tanto de un preprocesamiento cuidadoso como de la selección de una arquitectura de modelo adecuada al problema.

Cuadro 4: Tabla comparativa final de rendimiento en el conjunto de test.

Modelo (Conjunto de Test)	**MSE**	**MAE**	**R^2**
Two-Stage Model (Classifier + Regressor)	**0.3206**	**0.4098**	**0.6553**
Linear Regression (Log+Poly Features)	0.3214	0.4103	0.6544
Ridge Regression (Log+Poly Features)	0.3214	0.4103	0.6544
Lasso Regression (Log+Poly Features)	0.3332	0.4232	0.6418
Linear Regression (Datos Limpios, sin escalar)	0.3454	0.4336	0.6286
Linear Regression (Datos Crudos, sin limpiar)	0.5559	0.5332	0.5758