



## INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

---

### TP2: Regresión del valor de valor medio de casas en distritos de California

---

Profesor:

Dr. Ing. Facundo Adrián Lucianna

Ing. Pablo Martín Gómez Verdini  
[gomezpablo86@gmail.com](mailto:gomezpablo86@gmail.com)

Ing. Diego Paciotti Iacchelli  
[diegopaciotti@gmail.com](mailto:diegopaciotti@gmail.com)

Ing. Joaquín González  
[joagonzalez@gmail.com](mailto:joagonzalez@gmail.com)

3 de octubre de 2024

# Índice

<b>1. Instrucciones</b>	<b>2</b>
<b>2. Consignas</b>	<b>3</b>
<b>3. Desarrollo</b>	<b>4</b>
3.1. Consigna 1 . . . . .	4
3.2. Consigna 2 . . . . .	6
3.3. Consigna 3 . . . . .	9
3.4. Consigna 4 . . . . .	12
3.5. Consigna 5 . . . . .	14
3.6. Consigna 6 . . . . .	16
3.7. Conclusiones . . . . .	18
<b>A. Ejecución y uso de código</b>	<b>20</b>

# Índice de figuras

1. Matriz de correlación entre features. . . . .	5
2. Histograma de los datos . . . . .	6
3. Histograma de los residuos . . . . .	10
4. Comparación Q-Q de la distribución de los residuos con una distribución normal . . . . .	11
5. Selección del hiperparámetro $\alpha$ en base al MSE . . . . .	14
6. Selección del hiperparámetro $\alpha$ en base al MAE . . . . .	15
7. Comparación de performance entre modelos . . . . .	16
8. Comparación de performance entre modelos . . . . .	17
9. Distribución geográfica y paramétrica de los precios . . . . .	18
10. Instrucciones para instalar dependencias y ejecución del código. . . . .	20

# Índice de tablas

1. Matriz de correlación completa entre los atributos y el target. . . . .	4
2. Matriz de correlación entre los features. . . . .	4
3. Correlación entre cada feature y el target. . . . .	4
4. Estadísticos del tiempo de ejecución. . . . .	9

## 1. Instrucciones

Se requiere construir una regresión que nos permita predecir el valor medio de las casas en distritos de California, EE.UU. (medidos en cientos de miles de dólares \$100,000). Este dataset se deriva del censo de 1990 de EE.UU., donde cada observación es un bloque. Un bloque es la unidad geográfica más pequeña para la cual la Oficina del Censo de EE.UU. publica datos de muestra (un bloque típicamente tiene una población de 600 a 3000 personas).

Los atributos, en el orden en que se guardaron en el dataset, son:

- **MedInc**: Ingreso medio en el bloque.
- **HouseAge**: Edad mediana de las casas en el bloque.
- **AveRooms**: Número promedio de habitaciones por hogar.
- **AveBedrms**: Número promedio de dormitorios por hogar.
- **Population**: Población del bloque.
- **AveOccup**: Número promedio de miembros por hogar.
- **Latitude**: Latitud del bloque.
- **Longitude**: Longitud del bloque.

Y el target es:

- **MedHouseVal**: Mediana del costo de casas en el bloque (en unidades de a \$100.000).

Para este trabajo práctico (TP), se proporciona una notebook (`ayuda.ipynb`) con la lectura del dataset, la separación de los datos, entre otras ayudas para resolver este trabajo.

El entregable consiste en uno o más archivos de notebook `ipynb` con las respuestas. Aunque se da libertad para usar otros tipos de entregables, es importante incluir tanto el código de lo resuelto como las respuestas. Pueden subir el contenido o proporcionar un enlace a un repositorio público (GitHub o GitLab) con el contenido en el aula virtual.

## 2. Consignas

- (1) Obtener la correlación entre los atributos y los atributos con el target. ¿Cuál atributo tiene mayor correlación lineal con el target y cuáles atributos parecen estar más correlacionados entre sí? Se puede obtener los valores o directamente graficar usando un mapa de calor.
- (2) Graficar los histogramas de los diferentes atributos y el target. ¿Qué tipo de forma de histograma se observa? ¿Se observa alguna forma de campana que nos indique que los datos pueden provenir de una distribución gaussiana, sin entrar en pruebas de hipótesis?
- (3) Calcular la regresión lineal usando todos los atributos. Con el set de entrenamiento, calcular la varianza total del modelo y la que es explicada con el modelo. ¿El modelo está capturando el comportamiento del target? Expanda su respuesta.
- (4) Calcular las métricas de MSE, MAE y  $R^2$  del set de evaluación.
- (5) Crear una regresión de Ridge. Usando una validación cruzada de 5-folds y usando como métrica el MSE, calcular el mejor valor de  $\alpha$ , buscando entre  $[0, 12.5]$ . Graficar el valor de MSE versus  $\alpha$ .
- (6) Comparar, entre la regresión lineal y la mejor regresión de Ridge, los resultados obtenidos en el set de evaluación. ¿Cuál da mejores resultados (usando MSE y MAE)? Conjeturar por qué el mejor modelo mejora. ¿Qué error puede haberse reducido?

### 3. Desarrollo

#### 3.1. Consigna 1

Obtener la correlación entre los atributos y los atributos con el target. ¿Cuál atributo tiene mayor correlación lineal con el target y cuáles atributos parecen estar más correlacionados entre sí? Se puede obtener los valores o directamente graficar usando un mapa de calor.

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	Target
MedInc	1.000	-0.119	0.327	-0.062	0.005	0.019	-0.080	-0.015	0.688
HouseAge	-0.119	1.000	-0.153	-0.078	-0.296	0.013	0.011	-0.108	0.106
AveRooms	0.327	-0.153	1.000	0.848	-0.072	-0.005	0.106	-0.028	0.152
AveBedrms	-0.062	-0.078	0.848	1.000	-0.066	-0.006	0.070	0.013	-0.047
Population	0.005	-0.296	-0.072	-0.066	1.000	0.070	-0.109	0.100	-0.025
AveOccup	0.019	0.013	-0.005	-0.006	0.070	1.000	0.002	0.002	-0.024
Latitude	-0.080	0.011	0.106	0.070	-0.109	0.002	1.000	-0.925	-0.144
Longitude	-0.015	-0.108	-0.028	0.013	0.100	0.002	-0.925	1.000	-0.046
Target	0.688	0.106	0.152	-0.047	-0.025	-0.024	-0.144	-0.046	1.000

Tabla 1: Matriz de correlación completa entre los atributos y el target.

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude
MedInc	1.000	-0.119	0.327	-0.062	0.005	0.019	-0.080	-0.015
HouseAge	-0.119	1.000	-0.153	-0.078	-0.296	0.013	0.011	-0.108
AveRooms	0.327	-0.153	1.000	0.848	-0.072	-0.005	0.106	-0.028
AveBedrms	-0.062	-0.078	0.848	1.000	-0.066	-0.006	0.070	0.013
Population	0.005	-0.296	-0.072	-0.066	1.000	0.070	-0.109	0.100
AveOccup	0.019	0.013	-0.005	-0.006	0.070	1.000	0.002	0.002
Latitude	-0.080	0.011	0.106	0.070	-0.109	0.002	1.000	-0.925
Longitude	-0.015	-0.108	-0.028	0.013	0.100	0.002	-0.925	1.000

Tabla 2: Matriz de correlación entre los features.

Feature	Correlación con Target
MedInc	0.688
HouseAge	0.106
AveRooms	0.152
AveBedrms	-0.047
Population	-0.025
AveOccup	-0.024
Latitude	-0.144
Longitude	-0.046

Tabla 3: Correlación entre cada feature y el target.

**Analisis de los resultados obtenidos** El ingreso medio (MedInc) tiene la mayor correlación positiva con el target (Mediana del valor de las casas) con un valor de 0,69. Esto indica que a medida que el ingreso aumenta, también lo hará el valor de las casas, lo cual es intuitivo dado que las zonas con mayores ingresos suelen tener las viviendas de mayor precio.

En menor medida, el promedio de habitaciones (AveRooms) y la antigüedad de la casa (HouseAge), también tienen una correlación positiva con el atributo target. Esto sugiere que si bien hay influencia de estos features en el precio, lo hacen débilmente.

La latitud (Latitud) también posee una correlación negativa débil con el precio de las casas. Esto podría indicar que las viviendas ubicadas más al norte tienden a tener un precio más bajo.

El resto de los features presentan una correlación muy débil con el valor de las casas, mostrando una influencia limitada sobre esta.

Con respecto a la correlación entre features, el promedio de habitaciones (AveRooms) y el promedio de dormitorios (AveBedrms) poseen un valor muy elevado (0,85), lo cual es lógico dado que viviendas con más habitaciones suelen tener más dormitorios.

La correlación elevada pero negativa entre longitud y latitud puede atribuirse a la forma que tiene el estado de California. Al observarse en un mapa, puede notarse que al disminuir la latitud (Movimiento hacia el Sur) la longitud aumentará (Movimiento hacia el Este).

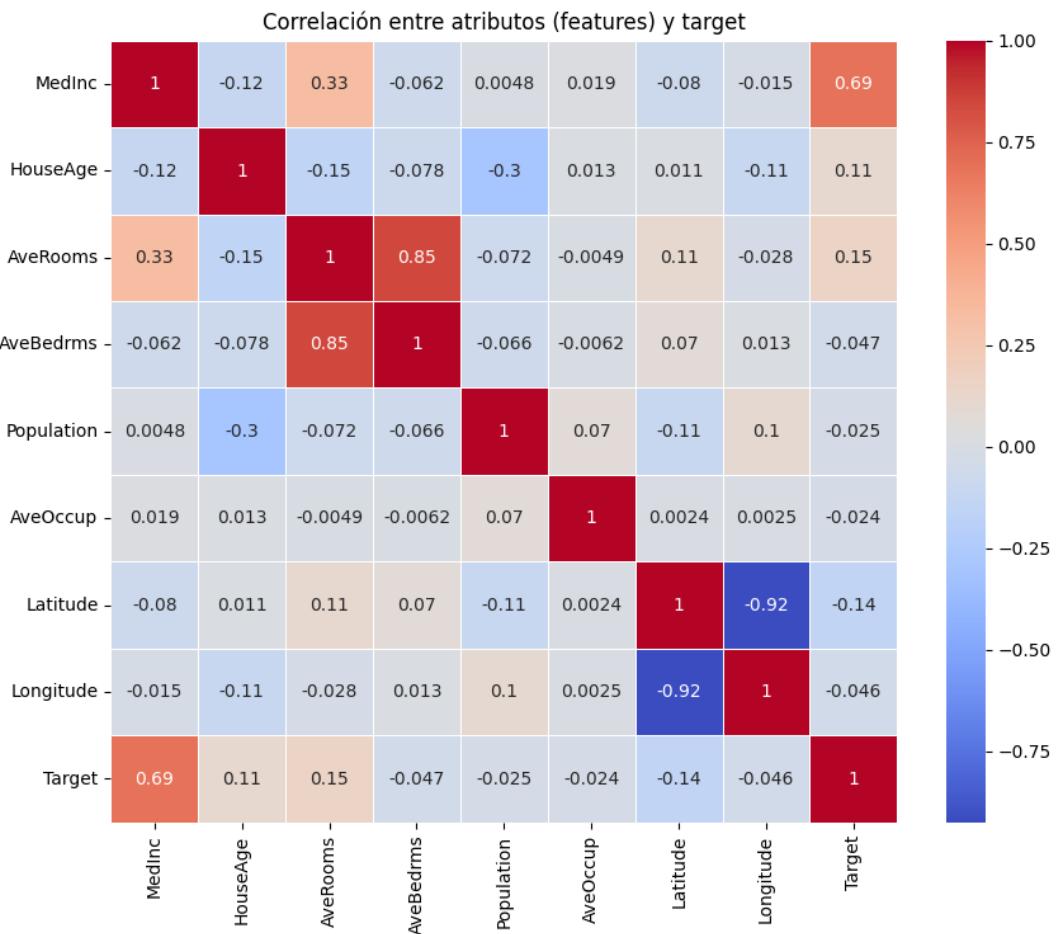


Figura 1: Matriz de correlación entre features.

### 3.2. Consigna 2

Graficar los histogramas de los diferentes atributos y el target. ¿Qué tipo de forma de histograma se observa? ¿Se observa alguna forma de campana que nos indique que los datos pueden provenir de una distribución gaussiana, sin entrar en pruebas de hipótesis?

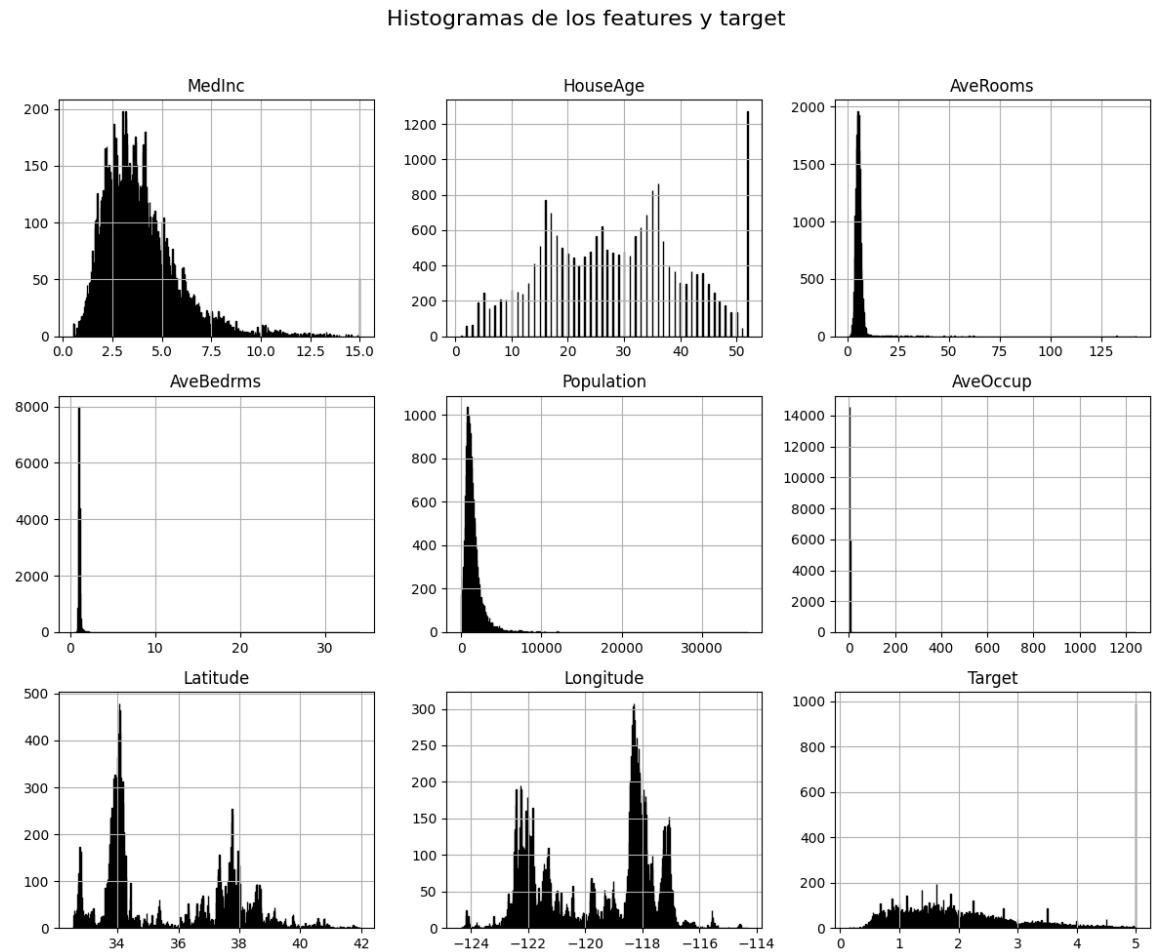


Figura 2: Histograma de los datos

## Análisis de los Resultados Obtenidos

### ■ **MedInc (Ingreso medio en el bloque):**

- Este feature muestra una distribución sesgada a la derecha (sesgo positivo). La mayoría de los valores se concentran aproximadamente entre 0 y 6, pero hay algunas áreas de mayor ingreso que extienden la cola hacia la derecha. No es una distribución gaussiana.

### ■ **HouseAge (Antigüedad mediana de las casas en el bloque):**

- La distribución es más o menos uniforme, excepto por el pico en la edad máxima (alrededor de 50 años), lo que probablemente refleja áreas con viviendas más antiguas.

### ■ **AveRooms (Promedio de habitaciones):**

- La distribución está fuertemente sesgada hacia la derecha (sesgo positivo). La mayoría de las casas tienen menos de 10 habitaciones, pero algunos pocos valores atípicos tienen significativamente más, extendiendo la cola hacia la derecha. No es una distribución gaussiana.

### ■ **AveBedrms (Promedio de dormitorios):**

- Este feature también muestra una distribución sesgada hacia la derecha. La mayoría de los puntos de datos están agrupados cerca de valores más bajos (menos dormitorios), con algunos valores atípicos que se extienden hacia números muy grandes de dormitorios.

### ■ **Population (Población en el bloque):**

- Puede observarse una distribución fuertemente sesgada hacia la derecha. La mayoría de las áreas tienen poblaciones pequeñas, pero hay algunas áreas con poblaciones muy grandes, creando una larga cola hacia la derecha. No sigue una distribución gaussiana.

### ■ **AveOccup (Promedio de ocupación por hogar):**

- Otra distribución fuertemente sesgada hacia la derecha, donde la mayoría de los valores están agrupados alrededor de un número bajo, y hay muy pocos hogares con tasas de ocupación extremadamente altas. No es gaussiana.

### ■ **Latitude (Latitud):**

- Los valores de latitud parecen estar agrupados en varios grupos distintos, lo que probablemente refleja diferentes regiones geográficas de California. Esta distribución parece bimodal asimétrica, pero no es gaussiana.

■ **Longitude (Longitud):**

- Al igual que la latitud, la longitud muestra múltiples picos, lo que sugiere varios grupos geográficos. Si bien la distribución está más dispersa que la latitud, podría también considerarse como una bimodal asimétrica. No sigue una distribución gaussiana.

■ **Target (Mediana del valor de las casas en el bloque):**

- La variable target parece tener un sesgo a la derecha pero más leve que en los features anteriores. Hay un pico alrededor de 1.5 y 2.0, con una cola que se extiende hacia valores más altos. En 5 puede visualizarse un pico máximo que se corresponde con el valor límite del dataset. Es decir, todos los bloques cuyas medianas del valor de las casas son iguales o mayores a 500.000 USD se agrupan en ese sector.

Si bien en la descripción individual de cada feature se concluyó que ninguna distribución correspondía estrictamente a una gaussiana, puede considerarse que MedInc y el target presentan una forma acampanada que podría indicar cierta relación con una distribución gaussiana. Una posible interpretación de esto es ruido en los datos como sobrerepresentación de algún grupo de la muestra o, para el caso puntual de los ingresos, la imposibilidad de tener muestras con valores negativos, lo que corta la cola izquierda de la campana.

### 3.3. Consigna 3

Calcular la regresión lineal usando todos los atributos. Con el set de entrenamiento, calcular la varianza total del modelo y la que es explicada con el modelo. ¿El modelo está capturando el comportamiento del target? Expanda su respuesta.

MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude
0.133	0.509	0.181	-0.273	-0.184	-0.010	-0.805	0.780
-0.539	-0.679	-0.422	-0.047	-0.376	-0.089	-1.339	1.245
0.170	-0.362	0.073	-0.242	-0.611	-0.044	-0.496	-0.277
-0.407	-1.155	0.175	-0.008	-0.987	-0.075	1.690	-0.706
-0.294	1.857	-0.259	-0.070	0.086	-0.066	0.992	-1.430

Tabla 4: Estadísticos del tiempo de ejecución.

- Varianza Total 1,3397887092485838: Esto representa la dispersión completa de los valores del target  $y_{train}$ . Es una medida de cuánta variabilidad hay en los precios de las viviendas.
- Varianza Explicada 0,8163948543798047: Es la cantidad de variabilidad en el target que el modelo de regresión lineal puede explicar utilizando las características del conjunto de datos.
- **Proporción Explicada:** El modelo explica **0.816** de **1.339**, lo cual corresponde aproximadamente al **61%** de la varianza total. Dicho de otro modo, el modelo está logrando explicar una cantidad considerable de la variabilidad en el target, pero no toda.
- **Varianza No Explicada (0.523):** La parte restante (0.523) sugiere que hay una porción significativa de la variabilidad del target que el modelo no puede explicar. Esto podría deberse a varios factores, como:
  - **Relaciones no lineales:** El modelo de regresión lineal asume una relación lineal entre los atributos y el target. Si la relación es más compleja, este modelo no será capaz de capturar todos los patrones presentes.
  - **Ruido en los datos:** Los datos podrían tener ruido o variabilidad aleatoria que no puede ser predicha por el modelo, lo cual es común en datos del mundo real.
  - **Que los residuos no tengan una distribución normal → Heterocedasticidad:** La heterocedasticidad se da cuando la variabilidad de los residuos no es constante a lo largo del rango de valores de las predicciones. Esto viola uno de los supuestos de la regresión lineal y podría sugerir que un ajuste lineal no es el más apropiado o que se necesita algún tipo de transformación de los datos. Esto puede verse en el gráfico QQ de los residuos que se adjunta a continuación donde los mismos no se alinean con los valores teóricos.

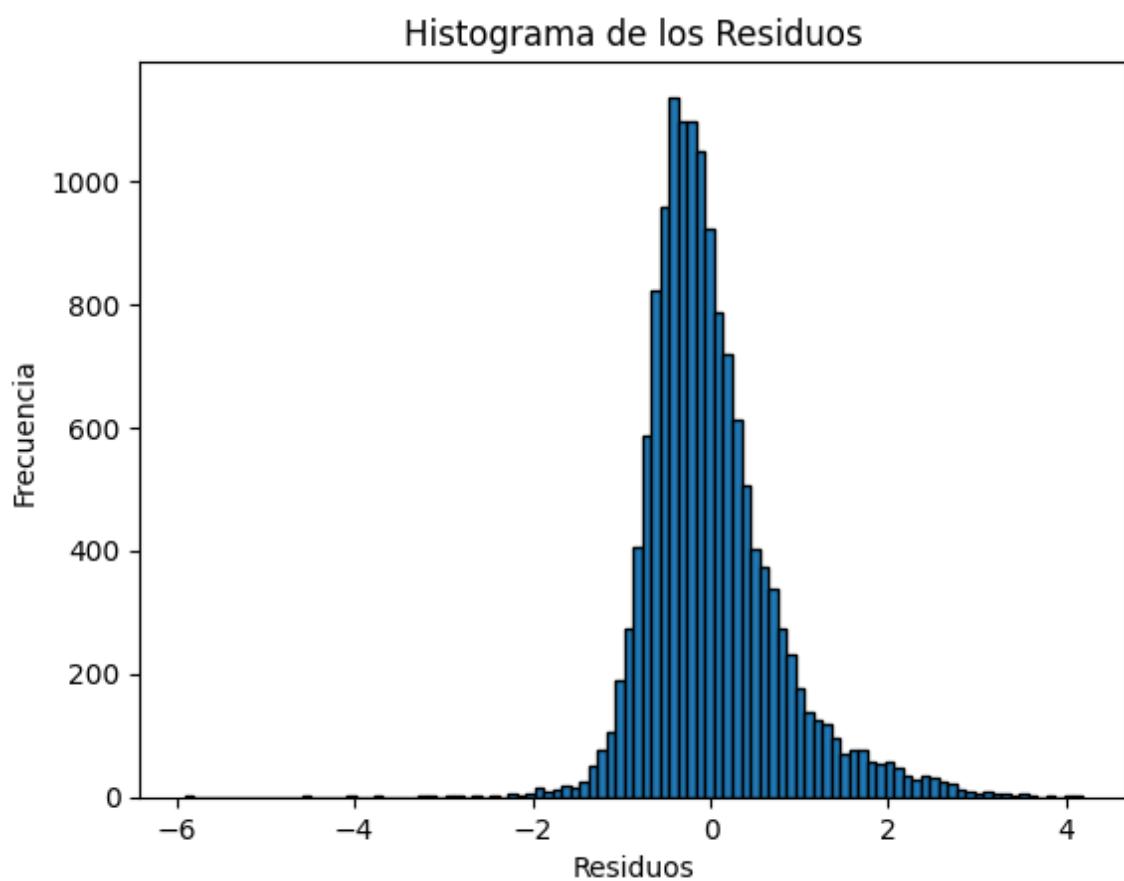


Figura 3: Histograma de los residuos

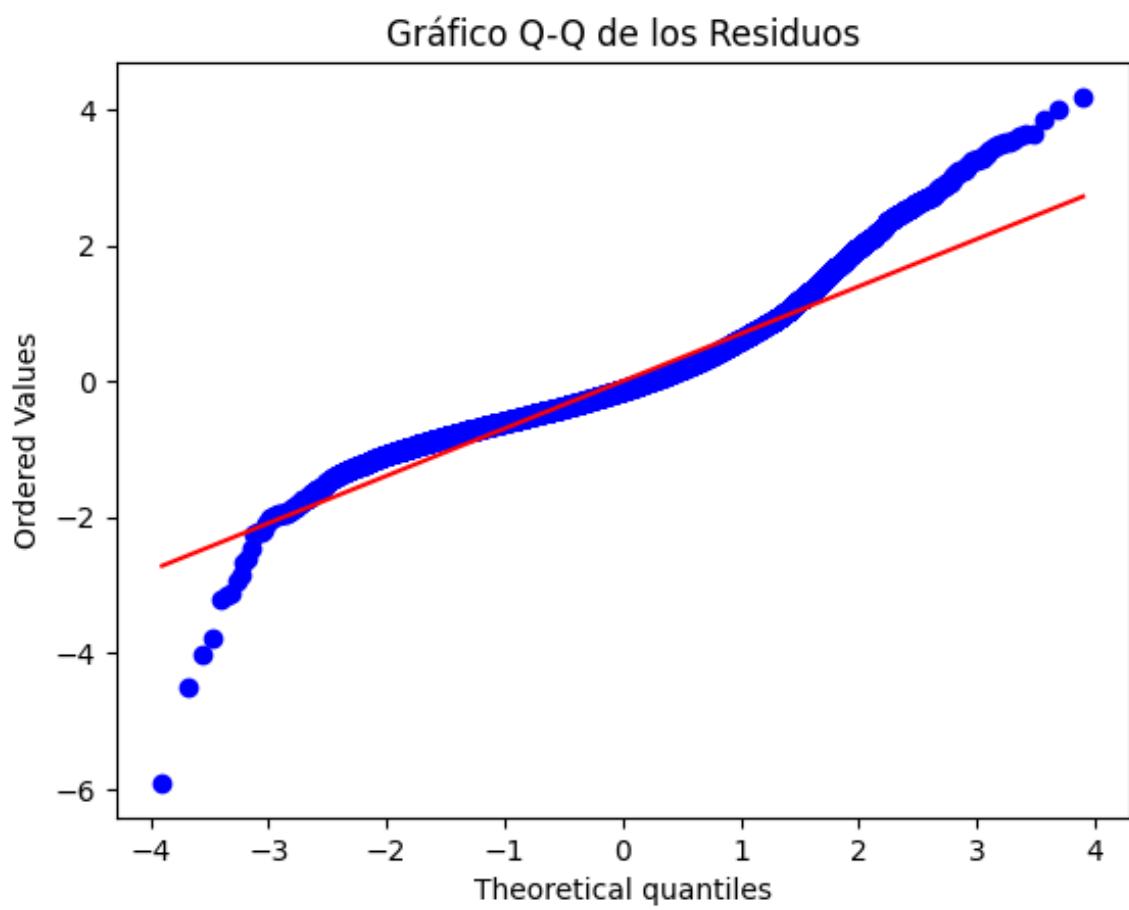


Figura 4: Comparación Q-Q de la distribución de los residuos con una distribución normal

### 3.4. Consigna 4

Calcular las métricas de MSE, MAE y R<sup>2</sup> del set de evaluación.

- Mean Squared Error (MSE): 0,5306
- Mean Absolute Error (MAE): 0,5272
- Coeficiente de Determinación (R<sup>2</sup>): 0,5958

#### 1. Mean Squared Error (MSE):

$$\frac{1}{N} \sum (y_{[i]} - \hat{y}_{[i]})^2 = 0,5306$$

El Error Cuadrático Medio (MSE) mide el promedio de los errores al cuadrado entre los valores reales ( $y_{\text{test}}$ ) y los valores predichos ( $\hat{y}_{\text{test}}$ ). En este caso, el valor es 0.5306.

- El MSE amplifica los errores más grandes debido a la operación de elevar al cuadrado, lo que significa que el modelo tiene algunos errores significativos al predecir los valores de salida.
- Dado que no tenemos un contexto exacto sobre el rango del target ( $y_{\text{test}}$ ), es difícil saber si 0.5306 es un valor alto o bajo. Sin embargo, valores más pequeños indican un mejor ajuste. Puedes comparar el MSE del set de entrenamiento con el del set de evaluación para evaluar la consistencia del modelo.

#### 2. Mean Absolute Error (MAE):

$$\frac{1}{N} \sum |y_{[i]} - \hat{y}_{[i]}| = 0,5272$$

El Error Absoluto Medio (MAE) mide la media de las diferencias absolutas entre los valores reales y los valores predichos, y en este caso es 0.5272.

- El MAE proporciona una medida directa del error promedio en la misma escala que el target ( $y_{\text{test}}$ ). Esto significa que, en promedio, el modelo se desvía de los valores reales en aproximadamente 0.5272 unidades.
- A diferencia del MSE, el MAE no amplifica los errores grandes, por lo que es una métrica útil para interpretar el error típico sin dejarse llevar por los outliers. En general, cuanto más pequeño sea el valor de MAE, mejor será la capacidad del modelo para aproximar los valores reales.

#### 3. Coeficiente de Determinación (R<sup>2</sup>):

$$\frac{S_R}{S_T} = 0,5958$$

El R<sup>2</sup> mide la proporción de la varianza del target que el modelo es capaz de explicar. En este caso, es 0.5958.

- Un R<sup>2</sup> de 0.5958 indica que aproximadamente el 59.58 % de la varianza en el target ( $y_{\text{test}}$ ) puede ser explicada por el modelo utilizando los atributos proporcionados. Esto significa que el modelo tiene un desempeño moderado en términos de explicar la variabilidad de los datos.

- Un  $R^2$  del 59.58 % es algo razonable, pero no excelente. Muestra que el modelo captura una parte significativa de la relación entre las características y el target, pero todavía hay un 40.42 % de la varianza que el modelo no puede explicar. Esto podría deberse a que:
  - Las relaciones entre los atributos y el target no son lineales y un modelo de regresión lineal no logra capturarlas adecuadamente.
  - Atributos importantes no están incluidos en el dataset o el modelo tiene una capacidad limitada para extraer la información relevante debido a la naturaleza del ajuste lineal.

### 3.5. Consigna 5

Crear una regresión de Ridge. Usando una validación cruzada de 5-folds y usando como métrica el MSE, calcular el mejor valor de  $\alpha$ , buscando entre 0 y 12.5. Graficar el valor de MSE versus  $\alpha$

El valor de  $\alpha$  que minimiza el MSE es 6,57

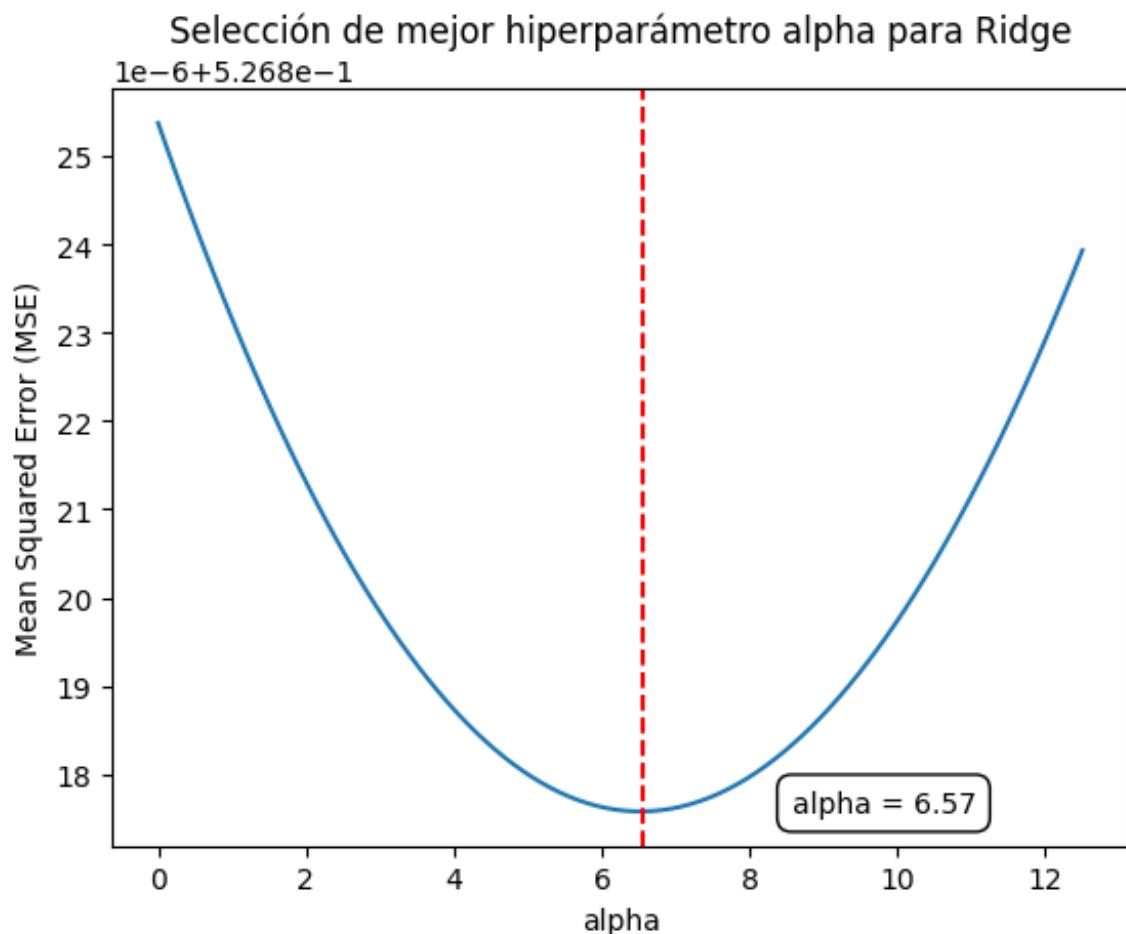


Figura 5: Selección del hiperparámetro  $\alpha$  en base al MSE

El valor de  $\alpha$  que minimiza el MAE es 12,50

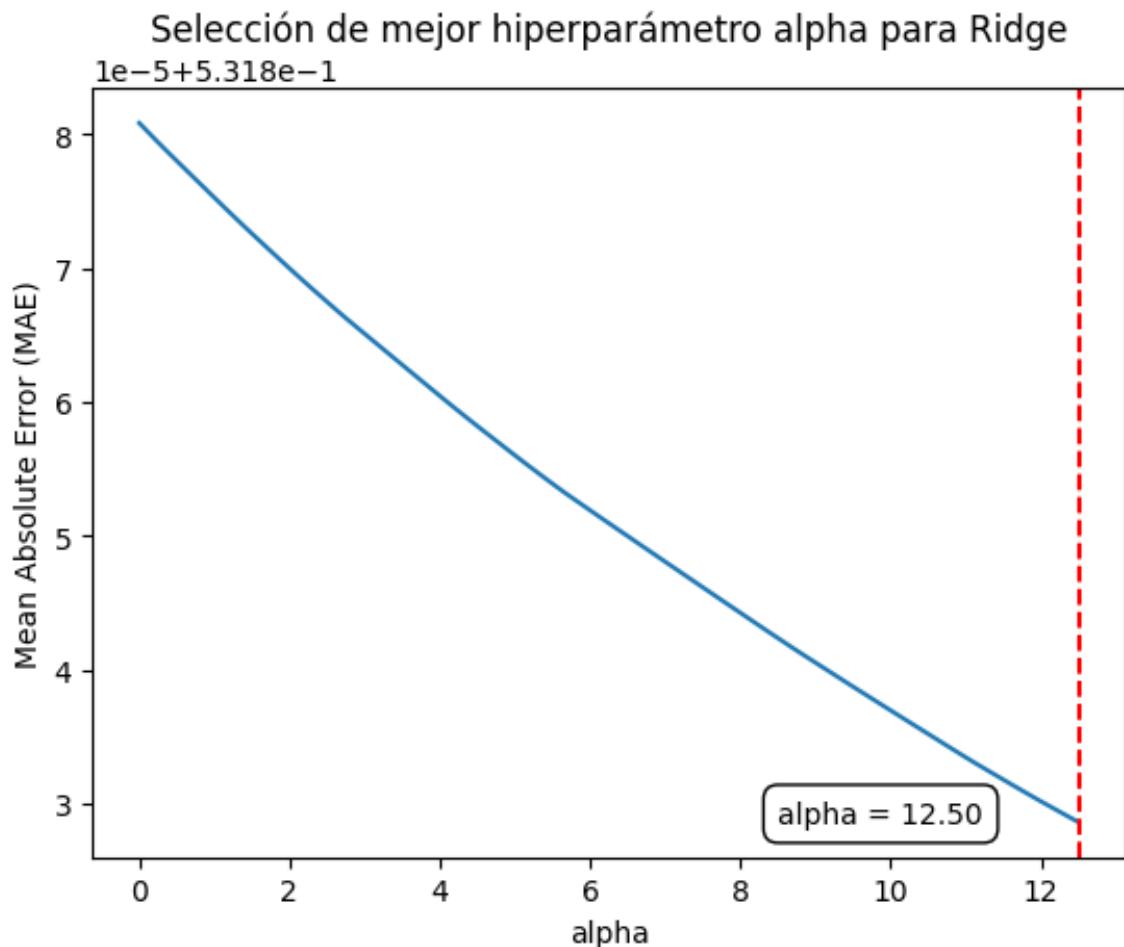


Figura 6: Selección del hiperparámetro  $\alpha$  en base al MAE

### 3.6. Consigna 6

Comparar, entre la regresión lineal y la mejor regresión de Ridge, los resultados obtenidos en el set de evaluación. ¿Cuál da mejores resultados (usando MSE y MAE)? Conjeturar por qué el mejor modelo mejora. ¿Qué error puede haberse reducido?

- **Ridge score:** 0,5958866942308549
- **Ridge mae score:** 0,5959843165304927

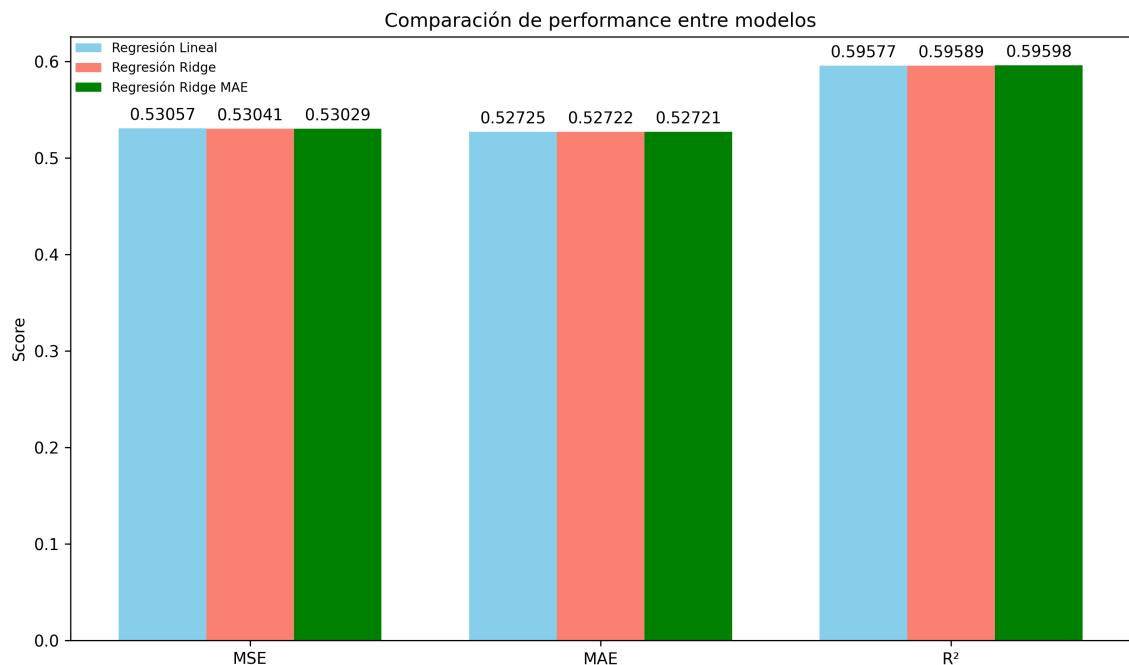


Figura 7: Comparación de performance entre modelos

Distribución espacial de valores de casas

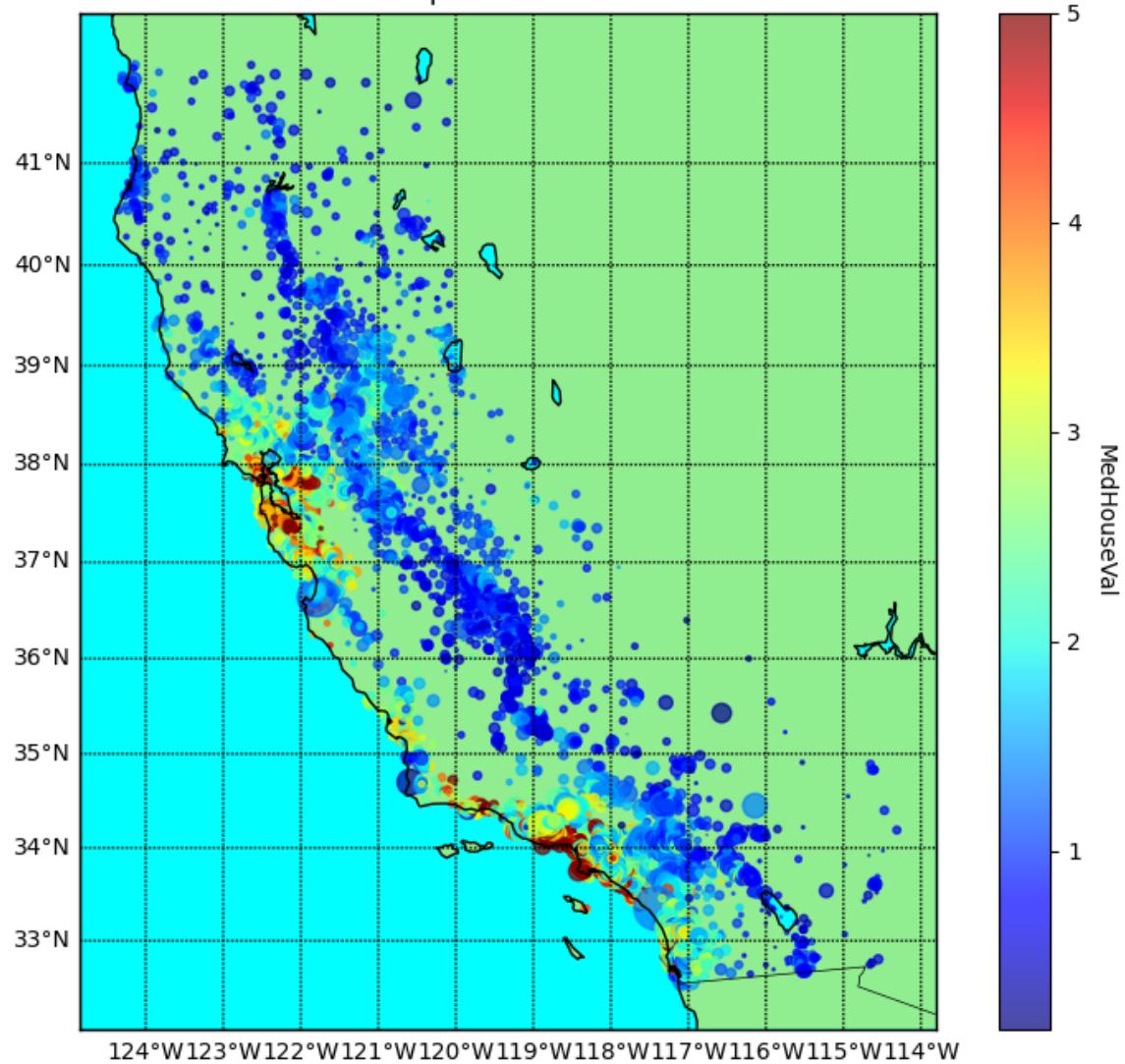


Figura 8: Comparación de performance entre modelos

Puede observarse que los precios mas altos se concentran en ciudades de alta población y costeras. Los precios, en promedio, bajan a medida que se va acercando al continente

### 3.7. Conclusiones

#### Observaciones y Conclusiones

A continuación se detallan las principales observaciones y conclusiones.

- **Distribuciones bimodales en latitud y longitude:** Se detectan distribuciones bimodales en las características de *latitude* y *longitude*, que coinciden con las coordenadas de ciudades costeras importantes como San Francisco y Los Ángeles, tal como se muestra en la Figura 9. Además, como se menciona en el punto 6, se observa que, a medida que las propiedades se alejan de la costa, la media de los precios disminuye. Para mejorar la performance, el modelo debería captar la correlación entre los radios aledaños de estas regiones y los precios (target), la preferencia por ciudades costeras, así como por zonas más densamente pobladas. Esta información no está presente en el dataset y probablemente requiere una mayor complejidad en el modelo utilizado, lo cual podría explicar la baja performance observada.

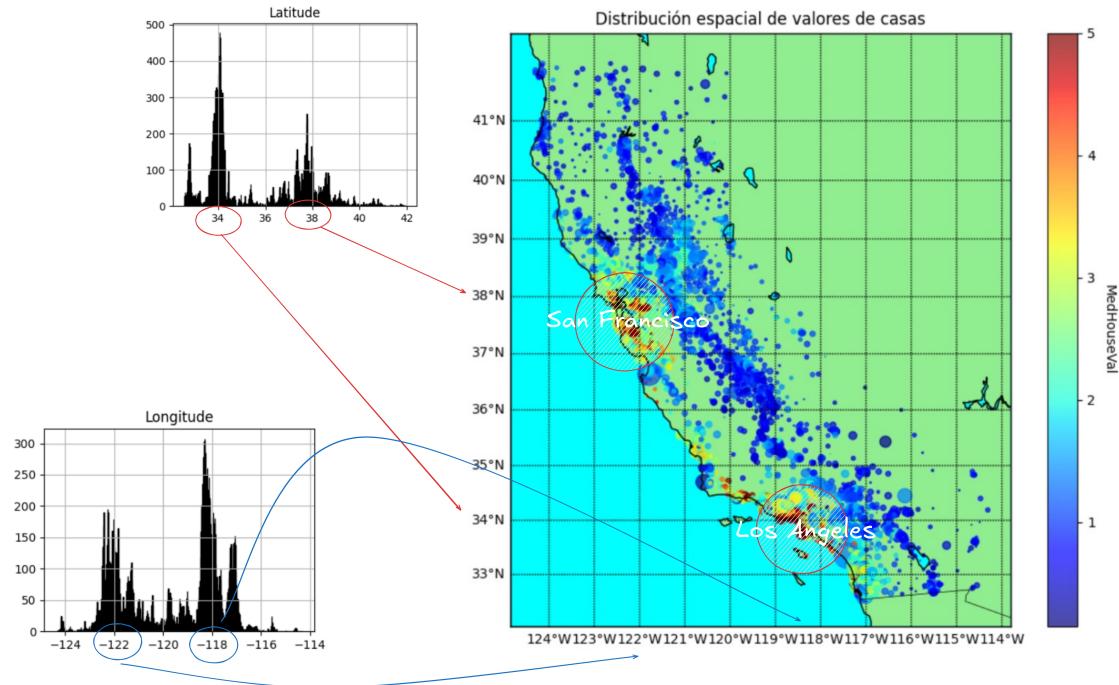


Figura 9: Distribución geográfica y paramétrica de los precios

- **Poca multicolinealidad:** La regularización no aporta mucho debido a la baja multicolinealidad. Esto se puede observar en la matriz de correlación [4], donde las únicas variables que presentan cierta correlación son el precio y la cantidad de ambientes.
- **Dataset con más de 20k muestras:** Dado que el dataset cuenta con más de 20,000 muestras, es menos probable que se produzca overfitting. Esto es particularmente relevante en casos donde Ridge aporta mayores mejoras con la regularización.
- **Baja performance del modelo:** La baja performance sugiere que la naturaleza de la variable a predecir tiene una mayor complejidad que la que pueden capturar las regresiones lineales.
- **Modelo Ridge con  $\alpha$  óptimo:** Se entrena un modelo Ridge con el valor de  $\alpha$  que minimiza el MAE (el cual es el doble del encontrado para el MSE)[6], y tampoco se obtienen resultados diferentes. Esto confirma la insensibilidad de este problema con estos datos a la regularización. Esto puede observarse en la Figura 7 que compara  $R^2$ , MAE, y MSE para los tres modelos.

## A. Ejecución y uso de código

El entregable esta en formato Jupyter Notebook .ipynb

Para el correcto uso de los scripts en Python adjunto, se deben realizar los siguientes pasos:

---

```
# clonar github repository
git clone git@github.com:FIUBA-CEIA-18Co2024/IIA-TP2.git

# crear y activar virtual environment
python -m venv venv
venv\Scripts\activate

# instalar dependencias
pip install -r requirements.txt
```

---

Figura 10: Instrucciones para instalar dependencias y ejecución del código.