Trabajo Práctico N.º 2

"Regresión del valor de VM de casas en distritos de California"

Laboratorio de Sistemas Embebidos
Introducción a la IA
David Canal
Abril 2024

1. Consigna de trabajo

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

1.1 Tareas y preguntas a resolver

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 α, buscando entre [0, 12.5]. Graficar el valor de MSE versus α.

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?.

2. Resolución

En el presente trabajo, se llevó a cabo el análisis de un conjunto de datos de precios de casas en California con el objetivo de modelar la mediana del costo de las viviendas en este estado de EEUU. Para este propósito, se utilizó la librería "scikit-learn" (sklearn) de Python, que proporciona las herramientas necesarias para el análisis y la implementación de algoritmos de aprendizaje automático.

A continuación, se detallan los análisis realizados para los modelos planteados:

2.1 Pregunta 1: "Correlación entre variables"

Antes de proceder con la construcción de los modelos de regresión lineal, se evaluaron las correlación entre los atributos y el target. Las mismas se presentan en la figura 2.1. Se observó que el atributo con mayor correlación lineal respecto al target "MedHouseVal" (mediana del costo de las casas) es "MedInc" (ingreso medio del bloque). Además, se identificó que los atributos con la mayor correlación entre ellos son "Latitude" (latitud del bloque) y "Longitude" (longitud del bloque).

Análisis de correlaciones

- **MedInc y MedHouseVal:** Es intuitivo esperar que a medida que los ingresos del bloque aumenten, los precios de los inmuebles también aumenten.
- **Latitude y Longitude:** La Latitude y la Longitude son variables asociadas a la ubicación geográfica del inmueble. Se espera que estas variables tengan una fuerte correlación entre sí, dado que la ubicación en coordenadas geográficas están intrínsecamente relacionadas.

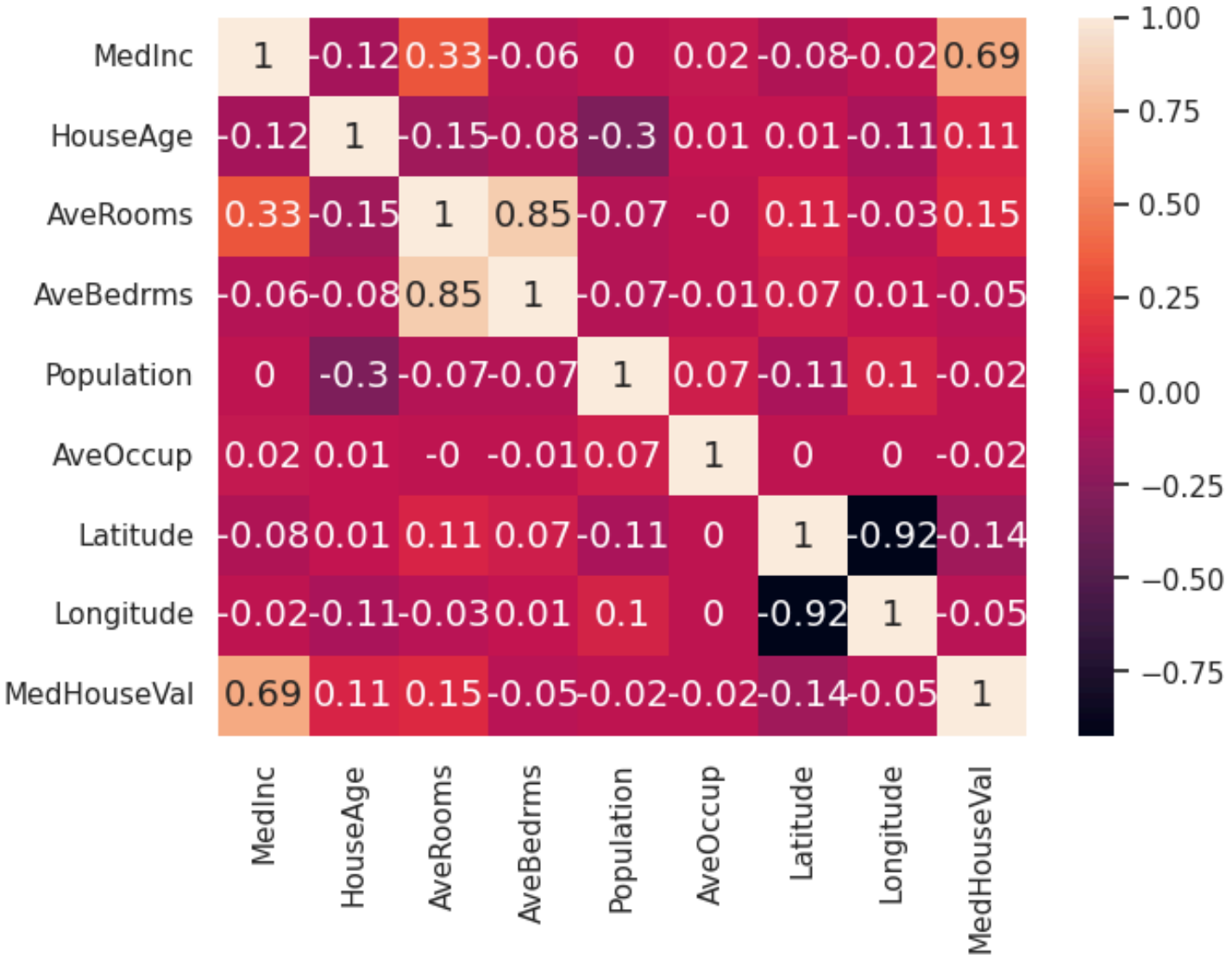

Figura 2.1: "Mapa de colores para la correlación de variables".

2.2 Pregunta 2: "Histogramas"

Se graficaron los histogramas de los diferentes atributos y el target para observar la distribución de los datos. Los mismos pueden verse en la figura 2.2. Para determinar el número de intervalos o "bins" necesarios para dividir los datos y construir los histogramas, se aplicó el criterio de "Sturges". Esd ecir:

$$bin = 1 + log_2(n) \quad (2.1),$$

donde:

n: es el número de datos de la muestra.

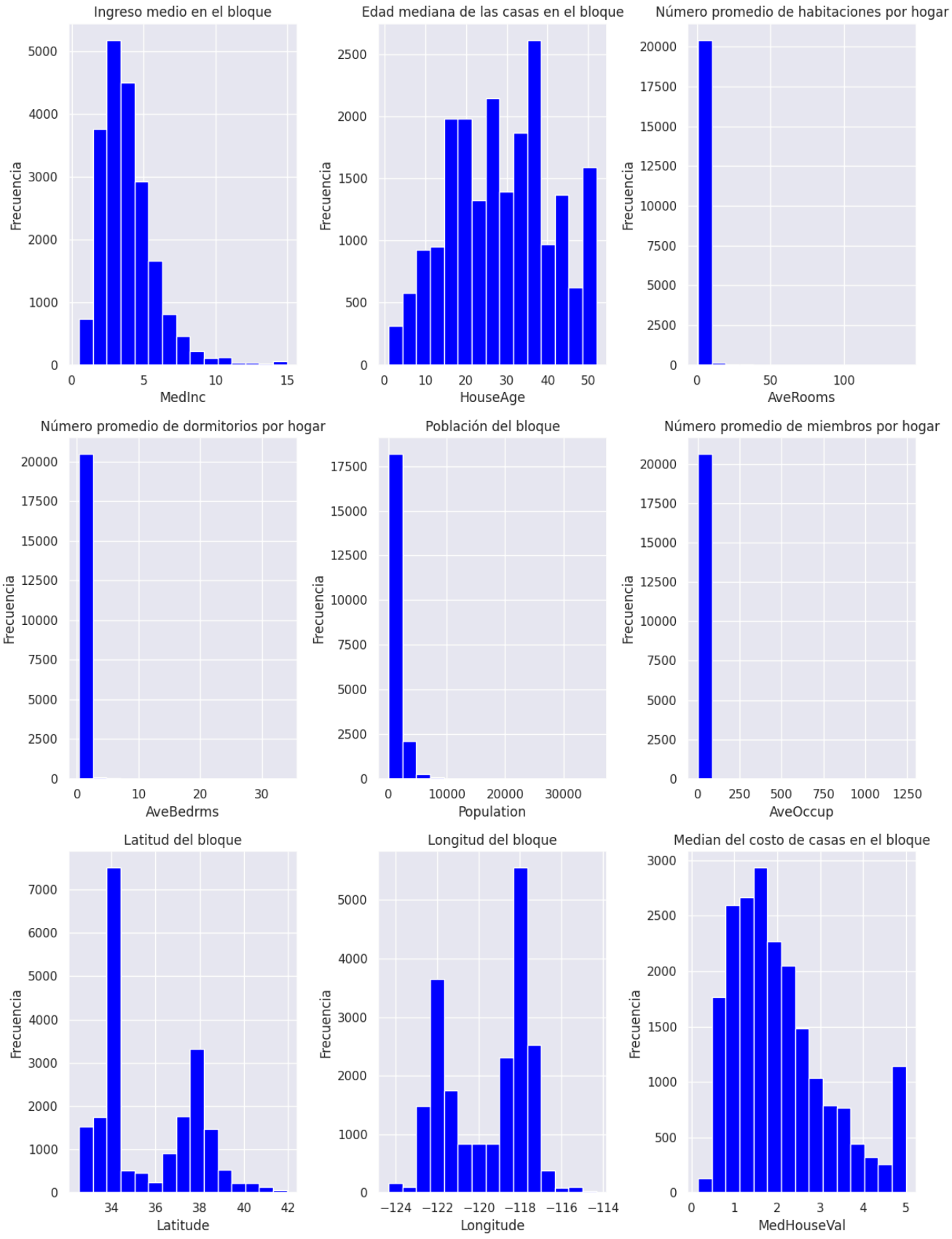

Figura 2.2: "Histogramas de los distintos atributos considerados en el análisis".

Al Visualizar los histogramas de los diferentes atributos, se remarcan los siguientes puntos:

- **MedInc:** Tiene una forma sesgada hacia la izquierda, con una concentración de datos en los valores más bajos de ingresos. Se sugiere una distribución asimétrica para representar este conjunto de datos. Esta distribución asimétrica sugiere una mayor cantidad de bloques con ingresos más bajos en comparación con los bloques de ingresos más altos
- **HouseAge:** Se observa una distribución más simétrica, con un amayor concentración de datos entre 20 y 40 años. La forma más acampanada del histograma podría indicar una distribución normal, donde la mayoría de las casas tienen edades típicas en este rango.
- **AveRooms:** Los datos están concentrados en habitaciones menores a 10. Esta distribución concentrada sugiere que la mayoría de las casas tienen un número típico de habitaciones, lo cual es esperable para casas promedio.
- **AveBedrms:** Idem al caso AveRooms.
- **Population:** Distribución asimétrica con una mayor concentración en valores alrededor de 5000 habitantes. La concentración de datos en este rango sugiere que la mayoría de las áreas tienen una población típica en esta escala.
- **AveOccup:** Similar al caso AveRoms y AveBedrms.
- **Latitude:** Se observa una distribución simétrica para Latitude, con dos concentraciones de datos alrededor de 34 y 38. Esta distribución de datos podría representarse con la superposición de dos distribuciones asimétricas centradas alrededor de cada concentración de datos observados.
- **Longitude:** similar a Latitude
- **MedHouseVal:** Similar a MedInc.

2.3 Pregunta 3: "Regresión Lineal"

Se construyó un modelo de regresión lineal utilizando todos los atributos disponibles en el conjunto de datos. Para ello, se utilizó la función LinearRegressión de la librería sklearn de Python, dividiendo los datos en datos de entrenamiento y de testeo. Con el set de entrenamiento, se calcularon los siguientes valores:

- **Varianza de la variable dependiente de los datos:**

$$s_T = 1.339$$

- **Varianza explicada por el modelo:**

$$s_R = 0.816$$

- **Varianza explicada por el modelo:**

$$s_E = s_T - s_R = 0.523$$

Con una varianza explicada de 0.816, sabemos que alrededor del 61% de la variabilidad en los precios de las casas puede ser atribuida a las variables predictoras en nuestro modelo. No obstante, queda un residuo de aproximadamente el 39% de variabilidad que no puede ser explicado por estas variables. Es importante tener en cuenta que este residuo podría indicar la presencia de otros factores importantes que influyen en el comportamiento de la variable en estudio.

2.4 Pregunta 4: "Métricas de evaluación"

Anteriormente se dividió el dataset en dos partes, una utilizada para entrenar al modelo, y otra para evaluar dicho modelo. El conjunto de evaluación se utilizó para evaluar el entrenamiento del modelo con el set reservado para este fin. Para ellos se calcularon las siguiente métricas:

- **Error cuadrático medio (MSE):** 0.530
- **Error absoluto Medio (MAE):** 0.527
- **Coeficiente de Pearson (R^2):** 0.596

2.5 Pregunta 5: "Regresión de Ridge"

Se realizaron modelos de ajustes de regresión lineal de Ridge para alfas comprendidos entre 0 y 12.5 ($\alpha \in [0, 12.5]$). Para dicho propósito, se utilizó la función "Ridge" de la librería "sklearn", previamente mencionada.

Mediante una validación cruzada de 5-folds y la métrica MSE, se calculó el mejor valor de α en el intervalo de interés de forma tal de minimizar el MSE. Se encontró que el error se minimiza para $\alpha = 6.52$. Por otro lado, el MSE se puede descomponer en las siguientes partes:

- Error por sesgo (Bias).
- Error por Varianza.

Por lo cuál, tal como se observa en la figura 2.3 existe una competencia entre ambas componentes del error dando lugar a un α que minimiza ambos efectos.

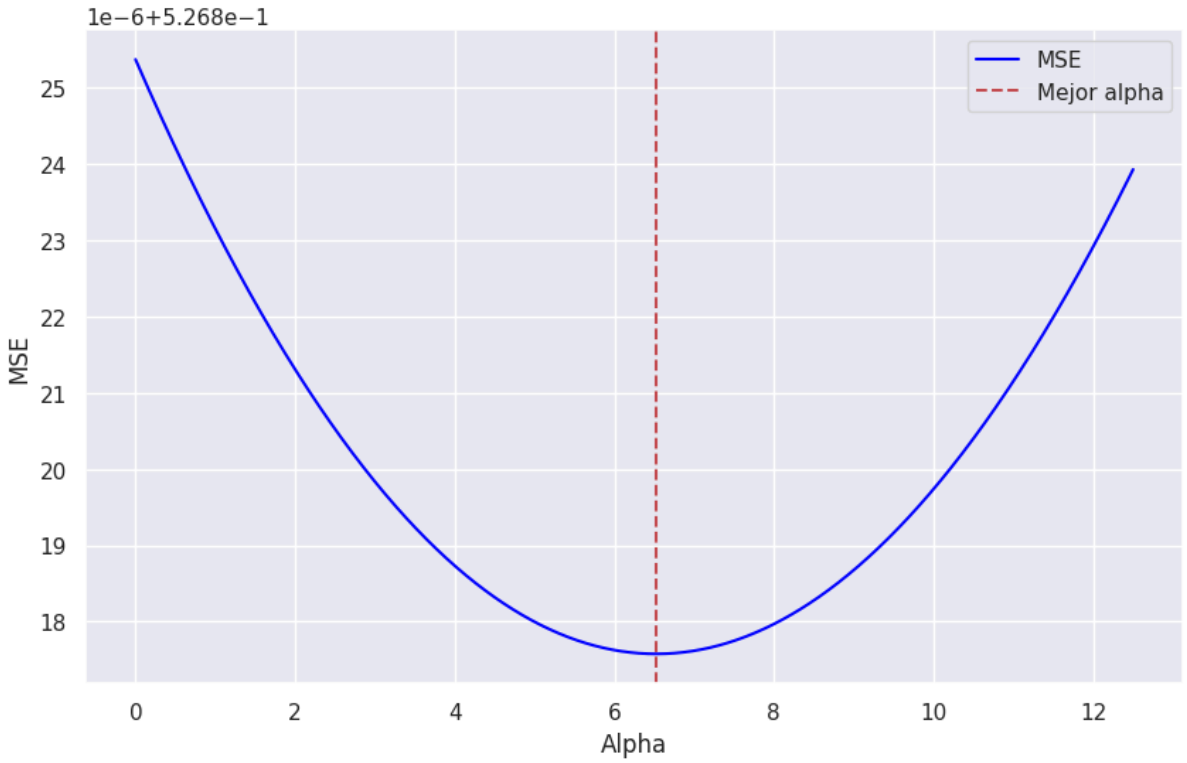

Figura 2.3: "MSE VS Alpha para la Regresión de Ridge".

7. 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?.

2.6 Pregunta 6: "Comparación de modelos"

Con el fin de comparar los modelos de regresión lineal y de Ridge presentados en los apartados anteriores, se calcularon los MSE y el MAE para ambos modelos, obteniéndose los siguientes resultados:

- **Error Cuadrático Medio:**

Regresión Lineal: 0.53056
Ridge: 0.53041

- **Error Absoluto Medio:**

Regresión Lineal: 0.52724
Ridge: 0.52722

Se evidencia una disminución en ambos tipos de errores para el modelo de Ridge. Esto se debe a que en el Ridge disminuye el error por sobreajuste, y controla la varianza.