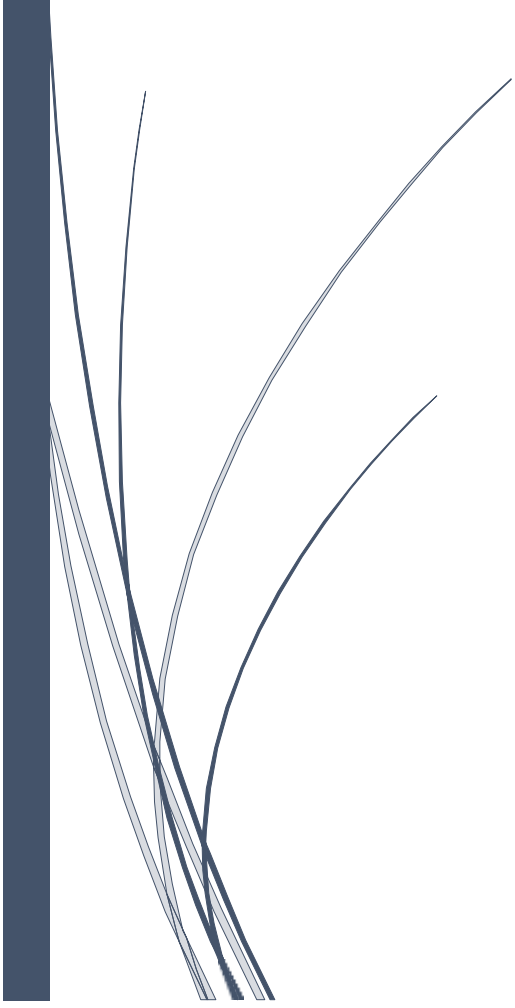
Aprendizaje automático para regresión

Práctica Grupal



Carlos Gálvez

1. Introducción

La práctica objeto de este análisis versa sobre un caso real de aprendizaje automático enfocado a la regresión, utilizando un conjunto de datos clásico sobre viviendas en Boston. El objetivo principal ha sido predecir el precio medio de una casa en función de diferentes características, como el número de habitaciones, la edad del edificio, la cercanía a zonas industriales o el nivel de criminalidad en la zona, entre otros factores. Para ello, he aplicado tres enfoques distintos: una regresión lineal simple, una regresión múltiple y un árbol de regresión.

Lo interesante de este ejercicio es que me ha permitido recorrer todas las etapas típicas de un proceso de machine learning, desde la carga y exploración de los datos hasta la aplicación de los modelos y la comparación de sus resultados. Además, al tratarse de un problema de regresión, he podido observar cómo se comporta cada tipo de modelo ante un conjunto de datos numéricos y cómo varía la precisión según el tipo de técnica utilizada. Al final, no se trata solo de aplicar algoritmos, sino de entender qué nos dicen los datos y cómo extraer de ellos una predicción útil y razonable.

2. Dataset y exploración inicial

Antes de aplicar ningún modelo, ha sido necesario conocer bien el conjunto de datos con el que iba a trabajar. Para ello, lo primero fue cargar manualmente el dataset original de Boston, ya que en las versiones más recientes de scikit-learn se ha retirado la función load_boston. El conjunto incluye un total de 506 observaciones y 13 variables independientes que describen diferentes aspectos socioeconómicos y físicos de las viviendas y sus alrededores. La variable objetivo, es decir, la que se busca predecir, es el precio medio de la vivienda, expresado en miles de dólares.

Durante la exploración inicial, visualicé las primeras filas del conjunto y comprobé que no había valores nulos, por lo que no fue necesario realizar limpieza de datos. Me centré en observar la distribución de algunas variables clave, como el número medio de habitaciones por vivienda (RM), el porcentaje de población con bajos recursos (LSTAT) o el índice de accesibilidad a carreteras (RAD). También generé un histograma del precio medio (MEDV) para entender cómo se distribuía la variable objetivo. En general, se aprecia una mayor concentración de viviendas en precios intermedios, con algunos valores extremos que podrían influir en el comportamiento de los modelos.
A continuación, realicé la clásica separación entre variables independientes (X) y variable dependiente (y), dividiendo el conjunto en entrenamiento y prueba con una proporción del 80/20. Esta división es fundamental para evaluar la capacidad de generalización de los modelos y evitar caer en el error de sobreentrenar sobre los mismos datos.

3. Visualización y relaciones

Una vez entendida la estructura general del dataset, el siguiente paso ha sido visualizar las relaciones entre algunas de sus variables clave y el precio medio de las viviendas. Para ello, he generado distintos gráficos de dispersión que permiten observar de forma intuitiva hasta qué punto determinadas características influyen sobre el valor de la vivienda.

La primera relación que he explorado ha sido entre el número medio de habitaciones (RM) y el precio (MEDV). Tal como era de esperar, existe una correlación positiva bastante clara: a medida que aumenta el número de habitaciones, el precio medio de la casa también tiende a subir. Aunque no es una relación perfecta, la nube de puntos muestra una tendencia ascendente lo suficientemente marcada como para considerarla una variable significativa.

También me pareció relevante analizar el porcentaje de población con bajos ingresos (LSTAT), ya que intuitivamente podría estar relacionado de forma inversa con el precio. En este caso, el gráfico de dispersión confirma esa idea: cuanto mayor es el porcentaje de personas con bajos recursos en una zona, menor suele ser el precio medio de la vivienda. Esta relación negativa es más acusada y concentrada que en el caso anterior, lo que la convierte en otra candidata fuerte para los modelos de predicción.

Además de estas gráficas individuales, he incluido un mapa de calor con las correlaciones entre todas las variables numéricas. Este tipo de visualización permite detectar rápidamente qué variables tienen más influencia sobre la variable objetivo. En el caso de este dataset, tanto RM como LSTAT y PTRATIO muestran correlaciones destacadas con MEDV, lo que refuerza su papel como predictores relevantes.

4. Preparación del dataset

Antes de entrenar los modelos, ha sido necesario preparar adecuadamente los datos. En primer lugar, separé las variables independientes de la variable objetivo, que en este caso es el precio medio de la vivienda (MEDV). El conjunto de variables explicativas incluye características estructurales y socioeconómicas como el número de habitaciones, la edad del inmueble, el índice de accesibilidad a autopistas, el nivel de contaminación o el porcentaje de población con bajos ingresos, entre otras.

Una vez realizada esta separación, dividí el conjunto completo en dos subconjuntos: uno de entrenamiento y otro de prueba. Para ello utilicé una partición clásica del 80% para entrenamiento y un 20% restante para validación. Esta división permite entrenar los modelos con una muestra amplia de datos y luego evaluar su capacidad de generalización con observaciones no vistas previamente.

El proceso no ha requerido realizar transformaciones adicionales como normalización o escalado, ya que los modelos seleccionados en esta práctica (regresiones y árboles) pueden trabajar directamente con las escalas originales del dataset. Tampoco ha sido necesario aplicar técnicas de imputación, ya que el conjunto de datos no contenía valores nulos ni inconsistencias.

Esta preparación ha sentado la base para aplicar los distintos modelos con garantías, asegurando una evaluación objetiva y comparable entre ellos.

5. Regresión lineal simple

La primera técnica que he aplicado ha sido una regresión lineal simple, utilizando como única variable predictora el número medio de habitaciones por vivienda (RM). Esta variable ya había mostrado una relación positiva clara con el precio durante la fase de exploración, así que parecía una buena candidata para empezar.

El modelo se entrenó únicamente con los datos de entrenamiento correspondientes a esta variable y, una vez ajustado, se utilizó para hacer predicciones sobre el conjunto de prueba. A nivel visual, la recta de regresión obtenida se superpone a los datos reales y muestra una tendencia ascendente que confirma lo observado previamente: a mayor número de habitaciones, mayor es el precio medio de la vivienda.

Desde el punto de vista cuantitativo, el modelo presenta un error cuadrático medio (MSE) moderado y un coeficiente de determinación (R^2) de aproximadamente 0.37. Este valor indica que, aunque existe una cierta relación lineal, el modelo no consigue explicar una gran parte de la variabilidad del precio únicamente con esta variable. En resumen, la regresión lineal

simple sirve como un buen punto de partida, pero sus limitaciones son evidentes al tratarse de un problema que depende de múltiples factores simultáneamente.

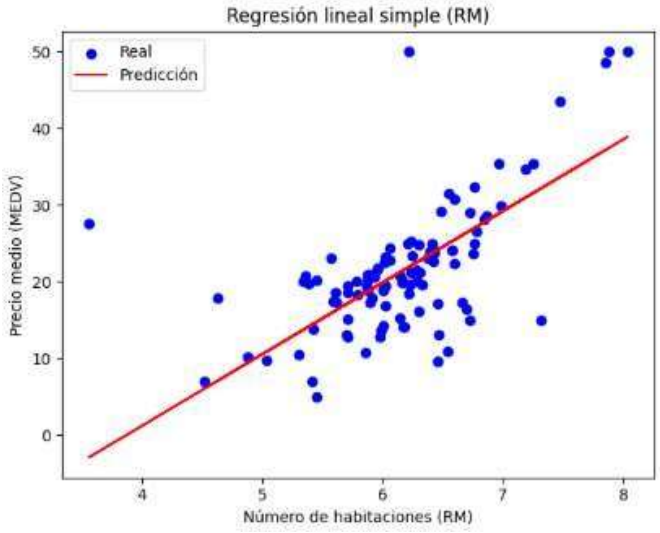


Gráfico 1. Relación entre el número medio de habitaciones (RM) y el precio medio de la vivienda (MEDV). Se observa una correlación positiva clara.

6. Regresión lineal múltiple

Después de probar con una sola variable, el siguiente paso lógico ha sido aplicar una regresión lineal múltiple, incorporando todas las variables del conjunto de datos como predictoras. El objetivo aquí es comprobar si el modelo mejora al considerar simultáneamente los distintos factores que influyen en el precio de una vivienda, algo que tiene mucho más sentido en un contexto real donde no suele haber una única causa determinante.

Al entrenar el modelo con este enfoque multivariable, los resultados han sido claramente mejores que en la regresión simple. Las predicciones obtenidas sobre el conjunto de prueba se ajustan mucho más a los valores reales, como se puede observar en el gráfico de dispersión, donde los puntos se alinean bastante bien con la diagonal ideal de predicción perfecta.

En cuanto a las métricas, el modelo reduce significativamente el error cuadrático medio y mejora el coeficiente de determinación, alcanzando un R^2 cercano al 0.67. Esto significa que ahora se explica un porcentaje mucho mayor de la variabilidad del precio a partir de las variables disponibles, lo cual valida el uso de un enfoque más completo.

A pesar de esta mejora, el modelo sigue siendo lineal, y por tanto puede tener dificultades para capturar relaciones más complejas o no lineales que también puedan estar presentes en los datos. Aun así, ofrece un equilibrio interesante entre simplicidad, interpretabilidad y precisión.

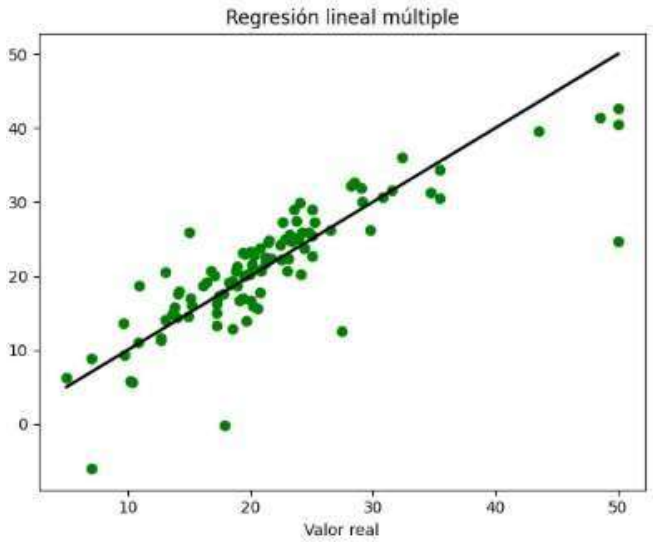


Gráfico 2. Regresión lineal simple utilizando la variable RM como predictor. La línea roja representa la predicción ajustada frente a los valores reales.

7. Árbol de regresión

Como tercer enfoque, he utilizado un modelo de árbol de regresión. A diferencia de los modelos lineales, este tipo de algoritmo es capaz de capturar relaciones no lineales entre las variables y la variable objetivo, dividiendo el espacio de decisiones en segmentos que se ajustan mejor a los datos.

El árbol fue entrenado con los mismos datos que los modelos anteriores, sin necesidad de normalización ni transformación. Uno de sus puntos fuertes es precisamente que trabaja bien con datos en bruto y es capaz de adaptarse con flexibilidad a patrones complejos.

Al aplicar este modelo sobre el conjunto de prueba, los resultados fueron bastante sólidos. El gráfico resultante muestra una buena alineación de los valores predichos respecto a los valores reales, con una dispersión incluso menor que en el modelo de regresión múltiple. Las métricas también lo reflejan: el error cuadrático medio es el más bajo de los tres modelos probados y el coeficiente R^2 alcanza un valor de 0.86, lo que indica que el árbol consigue explicar la mayor parte de la variabilidad del precio.

Eso sí, esta mejora en precisión puede venir acompañada de un riesgo mayor de sobreajuste, especialmente si no se aplican técnicas de regularización o poda. En esta práctica no se ha profundizado en esos ajustes, pero sería una línea interesante a explorar en trabajos futuros.

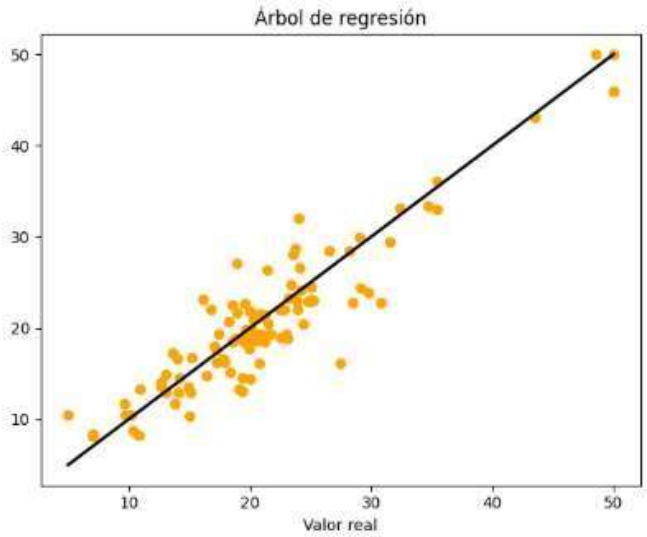


Gráfico 3. Comparación entre valores reales y predichos en la regresión lineal múltiple. Se aprecia un mayor ajuste respecto al modelo simple.

Además, como tarea adicional, he probado un ajuste simple del árbol de regresión limitando su profundidad a un valor máximo de 3. Esto se ha hecho con el objetivo de evitar un posible sobreajuste, ya que los árboles muy profundos tienden a memorizar los datos en lugar de generalizar. El modelo ajustado ha obtenido un rendimiento ligeramente inferior al árbol completo, con un R^2 algo más bajo, pero a cambio ofrece una mayor simplicidad y una

mejor capacidad de generalización. Este tipo de ajuste permite controlar la complejidad del modelo y es una técnica habitual en escenarios donde se busca un equilibrio entre precisión y robustez.

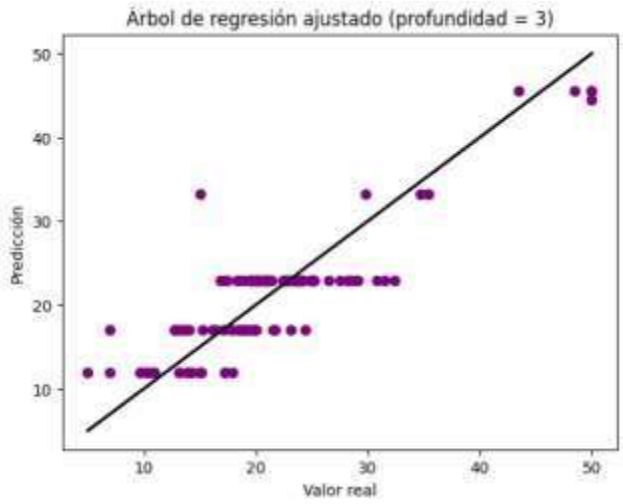


Gráfico 4. Árbol de regresión con profundidad máxima limitada a 3. El patrón escalonado es característico de modelos con baja complejidad.

8. Comparación final y conclusiones

Una vez aplicados los tres modelos, resulta evidente que el enfoque que mejor se adapta al problema concreto del precio de las viviendas es el árbol de regresión. A nivel cuantitativo, presenta el menor error y el valor más alto de R^2, lo que se traduce en una mayor capacidad de predicción. Además, desde un punto de vista visual, la dispersión de los valores predichos respecto a los reales es más ajustada, especialmente en los rangos de precio intermedio.

La regresión múltiple también ha ofrecido buenos resultados, mejorando de forma clara el desempeño de la regresión simple. No obstante, al tratarse de un modelo lineal, su capacidad para capturar relaciones complejas entre las variables es más limitada.

Por su parte, la regresión lineal simple ha cumplido su papel como modelo de partida, sencillo y fácil de interpretar, pero insuficiente en un problema como este, donde intervienen muchos factores distintos al mismo tiempo.

En conjunto, esta práctica me ha permitido no solo aplicar diferentes técnicas de aprendizaje automático, sino también entender mejor sus fortalezas y limitaciones. Más allá de las métricas, lo interesante ha sido observar cómo cada modelo interpreta la realidad de una forma distinta, y cómo elegir el modelo adecuado depende del tipo de problema, los datos disponibles y el equilibrio que queramos mantener entre precisión, interpretabilidad y complejidad.