



CIENCIA DE DATOS

# **dataset Boston Housing Dataset**

**BOGOTA DC – 2024**

**SANTIAGO CARVAJAL  
FERNANDEZ**



# Introducción

El Boston Housing Dataset es una herramienta fundamental en el campo del aprendizaje automático y la estadística, utilizado frecuentemente para ilustrar conceptos de regresión lineal. Este dataset, que incluye información sobre diversas características de viviendas en Boston, como el número de habitaciones y el valor medio de las casas, ofrece una base sólida para aplicar y entender técnicas de análisis predictivo.

En el contexto de la ciencia de datos, uno de los métodos más utilizados para modelar relaciones lineales entre variables es el algoritmo de gradiente descendente. Este algoritmo de optimización iterativo ajusta los parámetros del modelo para minimizar el error entre las predicciones y los valores reales. En el presente código, se emplea el gradiente descendente para realizar una regresión lineal sobre el Boston Housing Dataset, con el objetivo de predecir el valor medio de las casas (MEDV) en función del número de habitaciones (RM), una característica representativa del dataset.

El código proporciona una implementación práctica del gradiente descendente, comenzando con la carga del dataset y la selección de la característica 'RM' para el análisis. A continuación, se inicializan los parámetros del modelo y se ejecuta el algoritmo de gradiente descendente a través de un número definido de iteraciones. El proceso incluye la actualización de los parámetros del modelo basado en la tasa de aprendizaje y el cálculo de las derivadas parciales del error.

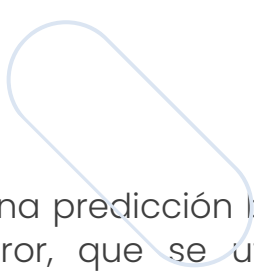


# ENSAYO

El análisis de datos y el aprendizaje automático son áreas fundamentales en la ciencia de datos, y la implementación de algoritmos de optimización es crucial para el desarrollo de modelos predictivos. El código presentado ilustra una aplicación práctica del algoritmo de gradiente descendente para realizar una regresión lineal utilizando el Boston Housing Dataset. Este dataset contiene información sobre características de viviendas en Boston y se utiliza para predecir el valor medio de las casas (MEDV) en función del número de habitaciones (RM).

El proceso comienza con la importación de las bibliotecas necesarias: `numpy` para cálculos numéricos, `pandas` para la manipulación de datos, `sklearn.datasets` para la carga del dataset y `matplotlib.pyplot` para la visualización. El dataset se carga desde OpenML utilizando la función `fetch_openml`, que proporciona un acceso sencillo a datos bien estructurados para análisis y experimentación. Se extraen las variables independientes ( $X$ ) y el objetivo ( $y$ ), con un enfoque específico en la característica 'RM', que representa el número de habitaciones de las casas.

La regresión lineal se lleva a cabo con el uso del algoritmo de gradiente descendente. Inicialmente, los parámetros del modelo, la pendiente  $mmm$  y el intercepto  $bbb$ , se establecen en 0. La tasa de aprendizaje  $LLL$  se fija en 0.01, y se definen 1000 épocas para iterar el proceso de optimización. El algoritmo de gradiente descendente busca minimizar el error entre las predicciones y los valores reales ajustando iterativamente los parámetros del modelo.



Durante cada iteración, el modelo realiza una predicción basada en los parámetros actuales y calcula el error, que se utiliza para actualizar los parámetros. Las derivadas parciales con respecto a  $m$  y  $b$  se calculan para ajustar la pendiente y el intercepto, respectivamente. Este ajuste continuo permite que el modelo mejore su precisión a medida que se minimiza el error.

Finalmente, el código genera un gráfico que visualiza la relación entre el número de habitaciones y el valor medio de las casas, mostrando la línea de regresión ajustada. La línea de regresión proporciona una representación visual de cómo el modelo predice el valor de las casas en función del número de habitaciones. El gráfico revela la efectividad del algoritmo de gradiente descendente al mostrar una alineación adecuada entre los datos reales y la predicción del modelo.

Este ejercicio demuestra cómo el gradiente descendente puede aplicarse a problemas de regresión lineal para encontrar la mejor aproximación lineal entre variables. A través de este enfoque, los estudiantes adquieren una comprensión práctica de cómo ajustar un modelo predictivo y evaluar su rendimiento. La implementación exitosa del algoritmo en el contexto del Boston Housing Dataset subraya la importancia de las técnicas de optimización en el desarrollo de modelos de aprendizaje automático eficaces.