

The Boston Housing Dataset

Adaptado do *dataset* inicialmente publicado por Harrison, D. and Rubinfeld, D.L. "Hedonic prices and the demand for clean air", J. Environ. Economics & Management, vol.5, 81-102, 1978.

Trata-se de um *dataset* clássico real que nos dá uma visão abrangente da situação habitacional em Boston, Massachusetts, contendo um conjunto específico de atributos que influenciam os preços das habitações, que refletem fatores como a taxa de criminalidade na zona, taxas de imposto, níveis de educação, acessibilidades, entre outros. A partir deste *dataset* podem ser construídos modelos preditivos que ajudem a perceber os fatores que mais afetam os mercados habitacionais.

Variáveis de entrada:

- CRIM: taxa de criminalidade per capita;
- ZN: proporção de terrenos residenciais para loteamentos superiores a 2323 m²;
- INDUS: proporção de área usada para negócio não retalhista;
- CHAS: variável *dummy* Charles River (1 se o lote se situa na margem do rio; 0 caso contrário);
- NOX: concentração de óxidos nítricos (partes por 10 milhões);
- RM: número médio de quartos por habitação;
- AGE: proporção de unidades ocupadas pelo proprietário construídas antes de 1940;
- DIS: distâncias ponderadas para cinco centros de emprego em Boston;
- RAD: índice de acessibilidade a autoestradas circundantes;
- TAX: taxa de imposto por cada \$10.000;
- PTRATIO: rácio de alunos por professor;
- LSTAT: percentagem de pessoas de classe baixa.

Variável de saída:

- MEDV: Valor mediano do preço das habitações ocupadas pelo seu próprio proprietário (em milhares de dólares).

This is a classic real dataset that provides us with a comprehensive view of the housing situation in Boston, Massachusetts, containing a specific set of attributes that influence housing prices, reflecting factors such as crime rate in the area, tax rates, education levels, accessibility, among others. Predictive models can be built from this dataset to help understand the factors that most affect housing markets.

Input features: 1) CRIM: per capita crime rate by town 2) ZN: proportion of residential land zoned for lots over 25,000 sq.ft. 3) INDUS: proportion of non-retail business acres per town 4) CHAS: Charles River dummy variable (1 if tract bounds river; 0 otherwise) 5) NOX: nitric oxides concentration (parts per 10 million) 6) RM: average number of rooms per dwelling 7) AGE: proportion of owner-occupied units built prior to 1940 8) DIS: weighted distances to five Boston employment centres 9) RAD: index of accessibility to radial highways 10) TAX: full-value property-tax rate per \$10,000 11) PTRATIO: pupil-teacher ratio by town 12) LSTAT: % lower status of the population

Output variable: 1) MEDV: Median value of owner-occupied homes in \$1000's.