----------:-:----------

What is **Boston Housing Price dataset.**

Boston Housing Price Dataset: Overview

The **Boston Housing Price dataset** is a classic dataset used for regression tasks, particularly in predicting **house prices** based on various features. It was originally collected in the 1970s by the U.S. Census Service and contains information about houses in different Boston suburbs.

Note: The **Boston Housing dataset (load_boston) has been deprecated** in scikit-learn due to ethical concerns related to one of its features (**racial composition of a neighborhood**). As a modern alternative, the **California Housing dataset (fetch_california_housing)** is recommended.

---

1. Dataset Features

The dataset contains **506 instances (houses)** with **13 numerical features** that describe different aspects of each house and its surrounding environment. The target variable (**median house price**) is given in $1000s.

Feature Name	Description
CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for large lots
INDUS	Proportion of non-retail business acres per town
CHAS	1 if tract is near Charles River, 0 otherwise
NOX	Nitrogen oxide concentration (air pollution)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built before 1940
DIS	Weighted distance to five employment centers
RAD	Accessibility to radial highways
TAX	Property tax rate per $10,000
PTRATIO	Pupil-teacher ratio in schools
B	Proportion of Black residents (deprecated for ethical reasons)

Feature Name	Description
LSTAT	Percentage of lower-status population
PRICE (Target)	Median house price in $1000s

---

2. Loading the Dataset

Although load_boston is no longer available, you can use the **California Housing dataset** as a substitute.

Using fetch_california_housing() (Recommended)

```
from sklearn.datasets import fetch_california_housing
import pandas as pd

# Load dataset
data = fetch_california_housing()

# Convert to Pandas DataFrame
df = pd.DataFrame(data.data, columns=data.feature_names)
df['PRICE'] = data.target  # Add target variable

# Display first few rows
df.head()
```

---

3. Target Variable (House Prices)

The target variable **PRICE** represents the **median house price** in thousands of dollars.

- **Regression Task**: The goal is to predict continuous values (house prices).
- **Unit**: The house price is in **$1000s**.
- **Example**: A predicted price of 22.5 means **$22,500**.

---

4. Key Insights from the Dataset

- **RM (Rooms per house)** is highly correlated with house prices. More rooms → higher price.
- **LSTAT (Lower-income population percentage)** has a **negative correlation** with prices (more low-income households → lower house price).
- **DIS (Distance to employment centers)** affects house values (closer = higher price).
- **TAX and RAD** influence house prices as accessibility and property tax play a role in housing costs.

---

5. Data Preprocessing for Neural Networks

To use this dataset in a neural network, we perform: ✓ **Train-test split** (80% training, 20% testing).
✓ **Feature scaling** using **StandardScaler** (important for deep learning).
✓ **Handling missing values** (not required for this dataset).

```
from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler


# Split features and target variable

X = df.drop('PRICE', axis=1)

y = df['PRICE']


# Train-test split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


# Standardization

scaler = StandardScaler()

X_train = scaler.fit_transform(X_train)

X_test = scaler.transform(X_test)
```

---

6. Applications of the Boston Housing Dataset

- **Predicting house prices** in real estate markets.
- **Understanding real estate trends** (e.g., how pollution, crime rates, and accessibility affect housing prices).
- **Feature selection research** for regression problems.
- **Benchmark dataset** for testing regression models.

---

7. Conclusion

The **Boston Housing dataset** is a widely used dataset for **house price prediction** in machine learning. Although it is now deprecated, it helps in understanding regression techniques. **California Housing (fetch_california_housing)** is a better alternative for modern ML applications.