# House Price Prediction Using Regression

**Assignment – Week 5: Introduction to Machine Learning in Data Science**

## 1. Introduction

Machine Learning enables systems to automatically learn patterns from data and make predictions without being explicitly programmed. In data science, supervised learning techniques are widely used when labeled data is available. Regression is a supervised learning method used when the target variable is continuous.

The objective of this assignment is to build a **Linear Regression model** to predict house prices using the **Boston Housing dataset**. The project involves data preprocessing, model training, performance evaluation using appropriate metrics, and visualization of results. Separate visualizations of actual and predicted house prices are used to clearly analyze model behavior.

## 2. Dataset Description

The Boston Housing dataset is a standard benchmark dataset used for regression problems. It contains housing-related data collected from different areas of Boston and is widely used for academic learning.

- **Dataset Source:** OpenML (Boston Housing Dataset)
- **Number of instances:** 506
- **Number of features:** 13
- **Target variable:** Median value of owner-occupied homes
- **Type of problem:** Regression (Supervised Learning)

The features represent various housing factors such as crime rate, number of rooms, distance to employment centers, accessibility to highways, and property tax rate.

## 3. Methodology

### 3.1 Data Preprocessing

Data preprocessing was performed to improve model performance and ensure reliable predictions:

- The dataset was divided into **training data (80%)** and **testing data (20%)**.
- **Feature scaling** was applied using **StandardScaler** to normalize all features.
- Scaling ensures that features with larger numeric ranges do not dominate the learning process.

### 3.2 Model Selection

A **Linear Regression** model was selected because:

- It is simple, efficient, and easy to interpret.
- It works well when there is a linear relationship between independent variables and the target.
- It provides a strong baseline for comparing more advanced regression models.

### 3.3 Training and Testing

The Linear Regression model was trained using the training dataset. After training, the model was tested on unseen data to evaluate its prediction accuracy.

## 4. Evaluation Metrics

The performance of the regression model was evaluated using the following metrics:

### 4.1 Root Mean Squared Error (RMSE)

RMSE measures the average magnitude of prediction errors. It penalizes large errors and is expressed in the same units as house prices. A lower RMSE value indicates better model accuracy.

### 4.2 Cross-Validation

To evaluate the model's generalization ability, **5-fold cross-validation** was performed. This method trains and validates the model on multiple data splits, ensuring consistent and unbiased performance.

## 5. Results and Model Accuracy Comparison

The Linear Regression model achieved a **reasonable RMSE value** on the test dataset, indicating acceptable prediction accuracy. To confirm the reliability of this result, cross-validation was applied.

The **cross-validation RMSE** was found to be very close to the **test RMSE**, which indicates that:

- The model generalizes well to unseen data.
- Overfitting is minimal.
- The model performance is stable across different data partitions.

### Visualization Analysis

Instead of using a single combined plot, **separate graphs for Actual Prices and Predicted Prices were plotted**. This approach was chosen for the following reasons:

- Separate graphs provide **clear and uncluttered visualization**, making it easier to interpret each trend individually.

- The **Actual Prices graph** shows the true distribution and variation of house prices in the dataset.
- The **Predicted Prices graph** displays how closely the model follows the actual price pattern.
- Comparing these graphs side by side helps in understanding whether the model captures the overall pricing trend.

The similarity in trends between both graphs indicates that the Linear Regression model successfully learns the underlying pattern of the data. Minor deviations, particularly at extreme price values, suggest limitations of a linear model in handling complex, non-linear relationships.

Overall, the comparison between test RMSE and cross-validation RMSE, along with the separate visualizations, confirms that the model is **consistent, reliable, and reasonably accurate**.

# 6. Improvements and Future Scope

Although the Linear Regression model provides good baseline performance, further improvements can be achieved by:

- Applying **Polynomial Regression** to model non-linear relationships.
- Using **regularization techniques** such as Ridge and Lasso Regression to reduce overfitting.
- Implementing **ensemble learning methods** like Random Forest and Gradient Boosting.
- Performing **feature selection** to identify and remove less influential features.

These approaches can significantly improve prediction accuracy.

# 7. Conclusion

In this project, a Linear Regression model was successfully developed to predict house prices using the Boston Housing dataset. The model demonstrated acceptable accuracy as measured by RMSE and showed strong generalization through cross-validation. The use of separate visualizations for actual and predicted prices provided clearer insight into model performance. This assignment effectively demonstrates the application of supervised learning and regression techniques in data science.