# **Project Report**

**Name:** Aravindh. V  
 **Course:** B.sc cs (AIML) 2nd year  
 **Registration Number:** TU6243202111006

### **Project Title :** House Price Prediction Using Multiple Linear Regression

## **1. Project Description**

This project focuses on predicting house prices based on features such as area, location, number of rooms, and other housing-related attributes. The motivation is to assist buyers, sellers, and real-estate agencies in estimating property values more accurately.

The goal of the project is to apply regression modeling and feature engineering to build a predictive model that provides reliable house price predictions.

## **2. Learning Objectives**

· Understand the application of **Multiple Linear Regression** in real-world prediction problems.

· Learn **data preprocessing techniques** such as handling missing values, feature scaling, and outlier detection.

· Evaluate model performance using metrics like **RMSE** and **R² score**.

· Explore the impact of **multicollinearity, outliers, and feature importance** on model accuracy.

· Deploy the trained model for real-world use.

## **3. Timeline**

* **Start Date:** [ Sept 13, 2025]
* **Submission date:** [ Sept 15, 2025]

## **4. Algorithm Used**

· **Algorithm Name:** Multiple Linear Regression

· **Explanation:**  
Multiple Linear Regression models the relationship between one dependent variable (house price) and multiple independent variables (e.g., area, number of rooms, location). It estimates coefficients for each feature to minimize the sum of squared errors.

· **Advantages:** Simple, interpretable, effective for continuous predictions, and widely used in real-estate analytics

## **5. Tools & Libraries**

* **Programming Language:** Python
* **Libraries Used:**
  + Pandas
  + NumPy
  + Scikit-learn
  + Matplotlib / Seaborn
  + Statsmodels
  + Joblib

## **6. Dataset Description**

· **Source:** Boston Housing Dataset (Scikit-learn)

· **Size:** 506 rows, 13 features + target variable (PRICE)

· **Target Variable:** PRICE (median value of houses in $1000s)

· **Key Features:**

* CRIM: Crime rate per capita
* RM: Average number of rooms per dwelling
* LSTAT: % lower status population
* CHAS: Proximity to Charles River (location)
* TAX: Property tax rate

## **7. Methodology**

· **Data Preprocessing:**

* Handled missing values (if any)
* Performed scaling using StandardScaler
* Checked and reduced multicollinearity using VIF
* Detected outliers using boxplots and Z-score

· **Model Training:**

* Split dataset into **80% training, 20% testing**
* Trained Multiple Linear Regression model
* Also experimented with Polynomial Regression

· **Evaluation:**

* Metrics: **RMSE** and **R² score**
* Cross-validation for better generalization

· **Hyperparameter Tuning:**

* Polynomial degree selection (for improvement test)

## **8. Results**

**Performance Metrics:**

* Linear Regression: RMSE ≈ 4.5, R² ≈ 0.72
* Polynomial Regression: RMSE ≈ 3.9, R² ≈ 0.78

 **Visualizations:**

* Correlation Heatmap
* Feature Importance (coefficients)
* Residuals Plot
* Outlier Detection with Boxplots
* Location vs Price (Boxplot by CHAS feature)

 **Insights:**

* Number of rooms (RM) and % lower status population (LSTAT) strongly affect prices.
* Location (CHAS) shows a clear impact—houses near the river are more expensive.
* Polynomial regression slightly improves performance but increases complexity.

## **9. Questions Answered**

 1. Which features most influence price? – RM, LSTAT, CHAS, TAX.

 2. What preprocessing is needed? – Scaling, missing value handling, outlier removal.

 3. What is RMSE and R² score? – Used as evaluation metrics.

 4. How does multicollinearity affect results? – Checked using VIF.

 5. Can we improve accuracy with polynomial regression? – Yes, improvement observed.

 6. How does location affect prediction? – CHAS feature shows impact.

 7. What is the impact of outliers? – Outliers distort model predictions.

 8. How to visualize feature importance? – Barplot of regression coefficients.

 9. What assumptions does linear regression make? – Linearity, independence, normality, homoscedasticity.

 10. How to deploy the model? – Used Joblib to save/load model.

 11. How to handle missing values? – Filled with median.

 12. How to split dataset effectively? – Train/test split with stratification if required.

 13. How to perform feature scaling? – StandardScaler applied.

 14. How to use cross-validation? – 5-fold cross-validation done.

 15. How to save and load the trained model? – Joblib used for persistence.

## **10. Challenges & Improvements**

**Challenges & Improvements**

* **Challenges:**
  + Multicollinearity among features reduced interpretability.
  + Outliers affected accuracy.
  + Limited dataset size restricted generalization.
* **Future Improvements:**
  + Use advanced models (Random Forest, XGBoost, Neural Networks).
  + Perform feature engineering (interaction terms, location encoding).
  + Collect larger datasets with more location-based features.
  + Deploy as a web-based app for real-time use.

## **11. References**

· Boston Housing Dataset (Scikit-learn Documentation)

· Scikit-learn official docs: https://scikit-learn.org

· Machine Learning by Andrew Ng (Stanford CS229)

## **12. GitHub Link**

[GitHub Repository URL : https://github.com/aravindhvinayagam2007-crypto/ML-project]