**Technical Report on Boston Housing Price Prediction Project**

**Introduction:**

In this project, we aim to build a predictive model to estimate housing prices in Boston using various social, economic, and structural features of houses and neighbourhoods. The goal is to apply machine learning techniques, particularly **Decision Tree Regressor**, to analyze the data and extract patterns that impact housing prices.

**Objective:**

The purpose of this report is to explain the algorithms used to solve the problem, provide detailed analysis of the challenges encountered, and propose solutions to improve the model’s performance.

**1. Data:**

**Data Description:**

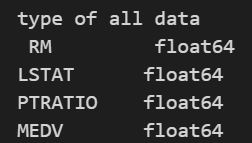
The dataset contains various features related to housing in Boston, which include:

* **average number of rooms among homes in the neighbourhood** - RM
* **Neighbourhood poverty level** - LSTAT
* **ratio of students to teachers in primary and secondary schools in the neighbourhood-** PTRATIO

The target variable is the **house price - MEDV**, which we aim to predict based on these features.

The data have no null values

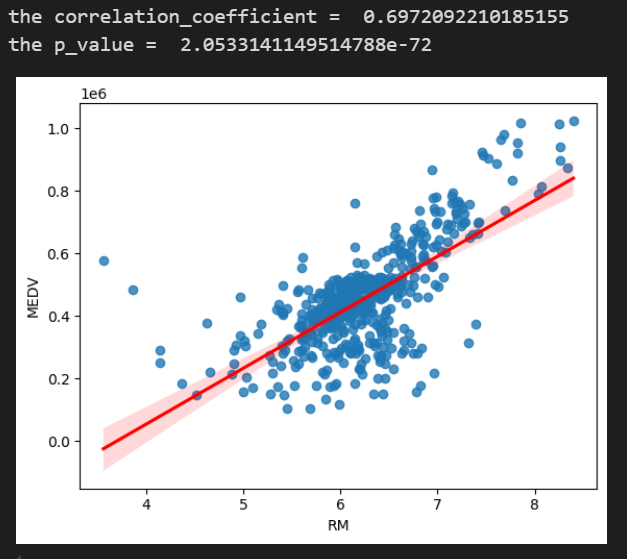
All the data have the same type **(float)**

****

**Feature Observation**

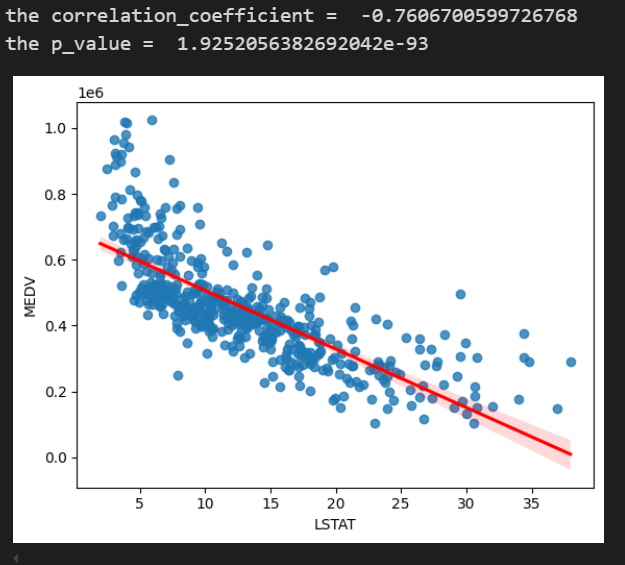
* Number of rooms and price

**The correlation coefficient is close to 1 and the p\_value is < 0.001, which means that the relation between the number of rooms and the price is a large positive relationship.**

****

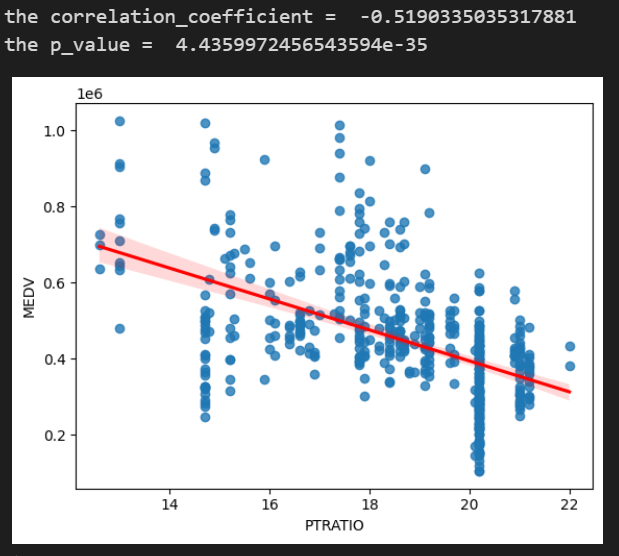
* relation between percent of lower class workers - price

**The correlation coefficient is close to -1 and the p\_value is < 0.001, which means that the relation between the percent of lower class workers and the price is a large negative relationship.**

****

* relation between ratio of students to teachers - price

**The correlation coefficient is between -1 and 0 and the p\_value is < 0.001, which means that the relation between the ratio of students to teachers and the price is a negative relationship but not strong.**

****

**2. Algorithms Used:**

**A. Linear Regression:**

Linear regression is a simple yet powerful algorithm used to find the linear relationship between the independent variables (features) and the dependent variable (house price). In this project, linear regression was used to estimate house prices based on the provided features.

**B. Cross-Validation:**

To ensure that the model generalizes well and avoids overfitting, we used **K-Fold Cross-Validation**. This method splits the data into k folds and trains the model on k−1k-1k−1 folds while testing it on the remaining fold. This process is repeated k times, and the overall performance is evaluated across all splits.



Our code



**C. GridSearchCV:**

We used **GridSearchCV** to tune the model’s hyperparameters. This technique involves testing a range of possible values for the hyperparameters, selecting the combination that provides the best model performance.



**3. Potential Challenges and Solutions:**

**A. Overfitting:**

Overfitting occurs when the model fits the training data too well, capturing noise rather than the true underlying patterns, which results in poor performance on unseen data.

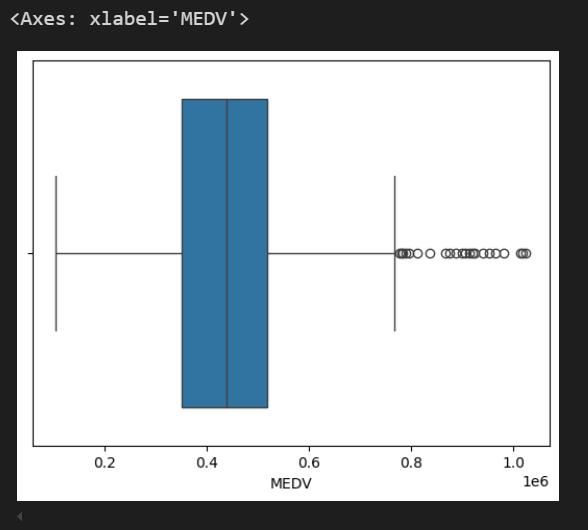
* **Solutions we applied** :
  + **Cross-Validation**: Using K-Fold Cross-Validation helps ensure that the model does not overfit by validating it on different subsets of the data.

**B. Data outliers :**

Machine learning models trained on data with outliers might be biased towards the extreme values, leading to poor performance on new, unseen data..

* **Solution**:
  + Remove Outliers.
  + Transform Data.





**4. Model Evaluation:**

The dataset was split into two groups:

* **Training set**: Used to train the model.
* **Test set**: Used to evaluate model performance after training.



**Evaluation Metrics:**

* **Mean Squared Error (MSE)**: Measures the average squared difference between the predicted and actual house prices. Lower values indicate better performance.
* **R² Score**: Indicates how well the independent variables explain the variation in the dependent variable. A score closer to 1 suggests a better model fit.