**Use machine learning to predict median home values in Boston based on socio-economic and housing features.**

**📌 Project Overview**

* **Objective: Create regression models that estimate the median value (MEDV) of Boston-area homes using standardized features.**
* **Dataset: 506 observations, 13 predictors and one target (MEDV) of the classic Boston Housing dataset**

**Pipeline Breakdown**

**1. Data Loading & Cleaning**

* Load dataset via Scikit-Learn or a CSV.
* Check for nulls and duplicates (none expected) [medium.com](https://medium.com/%40desirahz45/boston-house-price-prediction-with-xgboost-model-596af5181159?utm_source=chatgpt.com).

**2. Exploratory Data Analysis (EDA)**

* Visualize feature-target relationships using scatter plots and correlation heatmaps [charlesdelabra.github.io+15medium.com+15geeksforgeeks.org+15](https://medium.com/%40avinashrola/boston-house-price-prediction-using-linear-regression-e3ed951940c8?utm_source=chatgpt.com).
* Highlight key correlations: RM positively, LSTAT, CRIM, and PTRATIO negatively affect prices [medium.com+5github.com+5medium.com+5](https://github.com/g3varma/Boston-House-Price-Prediction?utm_source=chatgpt.com).

**3. Feature Engineering & Preprocessing**

* Scale numeric variables using StandardScaler.
* Optionally drop low-impact features (e.g., CHAS) or create derived features [reddit.com](https://www.reddit.com/r/MLQuestions/comments/103k2i0?utm_source=chatgpt.com)[github.com+3github.com+3github.com+3](https://github.com/g3varma/Boston-House-Price-Prediction?utm_source=chatgpt.com).

**4. Modeling**

* Baseline: **Linear Regression**
* Advanced: **Ridge**, **Lasso**, **Decision Tree**, **Random Forest**, **XGBoost**, or MLPs [medium.com+6medium.com+6geeksforgeeks.org+6](https://medium.com/%40avinashrola/boston-house-price-prediction-using-linear-regression-e3ed951940c8?utm_source=chatgpt.com).
* Evaluate with train-test split & k-fold CV using **R²**, **MAE**, **MSE/RMSE** [reddit.com](https://www.reddit.com/r/kaggle/comments/l415yn?utm_source=chatgpt.com).

**5. Model Interpretation**

* Analyze feature importance (especially in tree-based models).
* Use tools like SHAP or partial dependence plots to interpret model decisions .

**6. Deployment (Optional)**

* Serialize pipeline with pickle or joblib.
* Build a simple web interface via **Streamlit** or **Flask** for user inputs

**Key Learnings**

* **Top predictors:** RM (rooms), LSTAT (lower status), CRIM, and PTRATIO.
* **Model performance:** Ensemble methods like **Random Forest** and **XGBoost** often outperform linear models on this dataset [github.com+5medium.com+5github.com+5](https://medium.com/%40desirahz45/boston-house-price-prediction-with-xgboost-model-596af5181159?utm_source=chatgpt.com).
* **Interpretability matters:** Use SHAP or PDP to reveal how features influence model decisions.