**House Price Prediction Using Boston Housing Dataset - Project Report**

**Introduction**

This project involved predicting house prices using the Boston Housing Dataset by implementing regression models from scratch. The dataset is a classic in the machine learning community and is used to predict the median value of owner-occupied homes in Boston suburbs based on various features.

**Objective**

* Perform data preprocessing and normalization
* Implement Linear Regression, Random Forest, and XGBoost models from scratch
* Evaluate model performance using RMSE and R² metrics
* Visualize feature importance for tree-based models

**Dataset Overview**

* Total samples: 506
* Features: 13 numerical (e.g., CRIM, ZN, RM, LSTAT, PTRATIO)
* Target: MEDV (Median house value)

**Data Preprocessing**

* Normalized numerical features using standard scaling
* No categorical features required encoding
* Handled missing values and checked for duplicates
* Performed train-test split (80-20 ratio)

**Model Implementation**

**Linear Regression (Scratch)**

* Used gradient descent optimization
* Computed weights and bias iteratively to minimize mean squared error
* Output includes RMSE, R² score, and predicted values

**Random Forest (Scratch)**

* Built individual decision trees using variance reduction for splitting
* Employed bootstrap sampling and aggregated predictions via averaging
* Computed and visualized feature importance based on variance reduction gain

**XGBoost (Planned)**

* To be implemented with gradient boosting principles
* Will support additive trees, learning rate tuning, and regularization

**Personal Findings**

* Random Forest outperformed Linear Regression due to its ensemble nature and ability to model non-linear relationships
* RM (average number of rooms), LSTAT (lower status population), and PTRATIO (pupil-teacher ratio) were the most influential features
* Feature normalization significantly helped model convergence and stability

**Problems Faced**

1. **Variance Calculation in Decision Trees**:
   * Challenge: Ensuring correct variance computation and valid splits
   * Solution: Added split validation logic and debug logs
2. **Learning Rate Sensitivity in Gradient Descent**:
   * Challenge: Improper learning rate caused convergence issues
   * Solution: Experimented with various learning rates and epoch numbers
3. **Manual Tree Construction Overhead**:
   * Challenge: Writing custom code for node splitting, prediction, and recursion
   * Solution: Modularized the decision tree logic for reuse in both Random Forest and XGBoost

**Counter Actions and Solutions**

* Plotted learning curves to monitor gradient descent behavior
* Used fixed random seeds to ensure reproducibility
* Created visualization functions to highlight feature impact and model comparison

**Performance Comparison**

| **Model** | **RMSE** | **R² Score** |
| --- | --- | --- |
| Linear Reg. | ~5.00 | ~0.65 |
| Random Forest | ~3.00 | ~0.85 |

Note: Actual results may vary slightly based on data splits and random sampling

**Feature Importance (Random Forest)**

Top features contributing to house prices:

* RM: Average number of rooms
* LSTAT: % lower status of the population
* PTRATIO: Pupil-teacher ratio by town

**Conclusion**

This project demonstrated the process of building regression models from the ground up, offering deep insights into their mechanics and behavior. Tree-based models proved more powerful for this dataset, and future work will include implementing XGBoost and deploying the best model using a lightweight API or web app.