JOINT TECH INTERNSHIP  
 ML Assignment 2  
  
Team member 1: Mathav Ra  
Team member 2: Bavyadharshini R  
Team member 3: Aruna A  
  
**Step 1: Project Planning Phase**

* **Objective**: Brainstorm and outline tasks required to complete the house price prediction system.

**Tasks and Timelines:**

1. **Task 1: Data Collection & Exploration**
   * **Description**: Explore and understand the dataset, clean the data, and identify any missing values or inconsistencies.
   * **Timeline**: 2 days
   * **Assigned to**: Team Member 1
2. **Task 2: Data Preprocessing**
   * **Description**: Perform feature engineering, one-hot encoding of categorical variables, and outlier removal.
   * **Timeline**: 3 days
   * **Assigned to**: Team Member 2
3. **Task 3: Model Selection**
   * **Description**: Choose appropriate Machine Learning algorithms (Random Forest, Linear Regression, etc.) based on the problem.
   * **Timeline**: 2 days
   * **Assigned to**: Team Member 1
4. **Task 4: Model Training**
   * **Description**: Trained the model using an 80%-20% train-test split, experiment with different algorithms and hyperparameters.
   * **Timeline**: 4 days
   * **Assigned to**: Team Member 1
5. **Task 5: Model Evaluation**
   * **Description**: Evaluated the model’s performance using metrics like MAE, RMSE, and accuracy. Adjust and tune the model.
   * **Timeline**: 2 days
   * **Assigned to**: Team Member 1
6. **Task 6: UI Development (Optional)**
   * **Description**: Developed a basic Python UI for user interaction to input house details and get predicted prices.
   * **Timeline**: 3 days
   * **Assigned to**: Team Member 1
7. **Task 7: Documentation & GitHub Setup**
   * **Description**: Documented the project, including code, test data, and versioning. Prepare GitHub repository for final submission.
   * **Timeline**: 2 days
   * **Assigned to**: All

**Step 2: Design Phase**

Develop a **Design Document** with the following details:

1. **Training/Test Data Split**:
   * We will use an 80%-20% split for training and testing data to ensure model generalization.
2. **Acceptable Error Rates**:
   * MAE (Mean Absolute Error) of less than 10% of the average house price is considered acceptable.
3. **Machine Learning Algorithms**:
   * We will experiment with **Random Forest**, **Linear Regression**, and **XGBoost** to identify the best-performing algorithm for the dataset.
4. **Outlier Removal**:
   * Identifed outliers using Z-scores and remove houses with extreme price discrepancies to reduce model noise.
5. **Dimensionality Reduction**:
   * We will use **Principal Component Analysis (PCA)** if necessary to reduce multicollinearity and simplify the model.
6. **Model Versioning**:
   * To keep track of model versions saved checkpoints and using GitHub for code versioning.

**Step 3: Coding Phase**

* **Language**: Python will be the main programming language, and popular libraries like **pandas**, **scikit-learn**, and **matplotlib** will be used.

1. **Well-Commented Code**:
   * Every function and key part of the code is well-commented to explain its purpose and logic.
2. **Testing**:
   * Wrote unit tests to verify the functionality of the core components (data preprocessing, model training, prediction, etc.).
3. **Optional UI**:
   * Used **Streamlit** to create a UI where users can input house details (bedrooms, square footage, etc.) and get a predicted house price.
4. **Demo & Sharing**:
   * **Screenshot** of the application.

