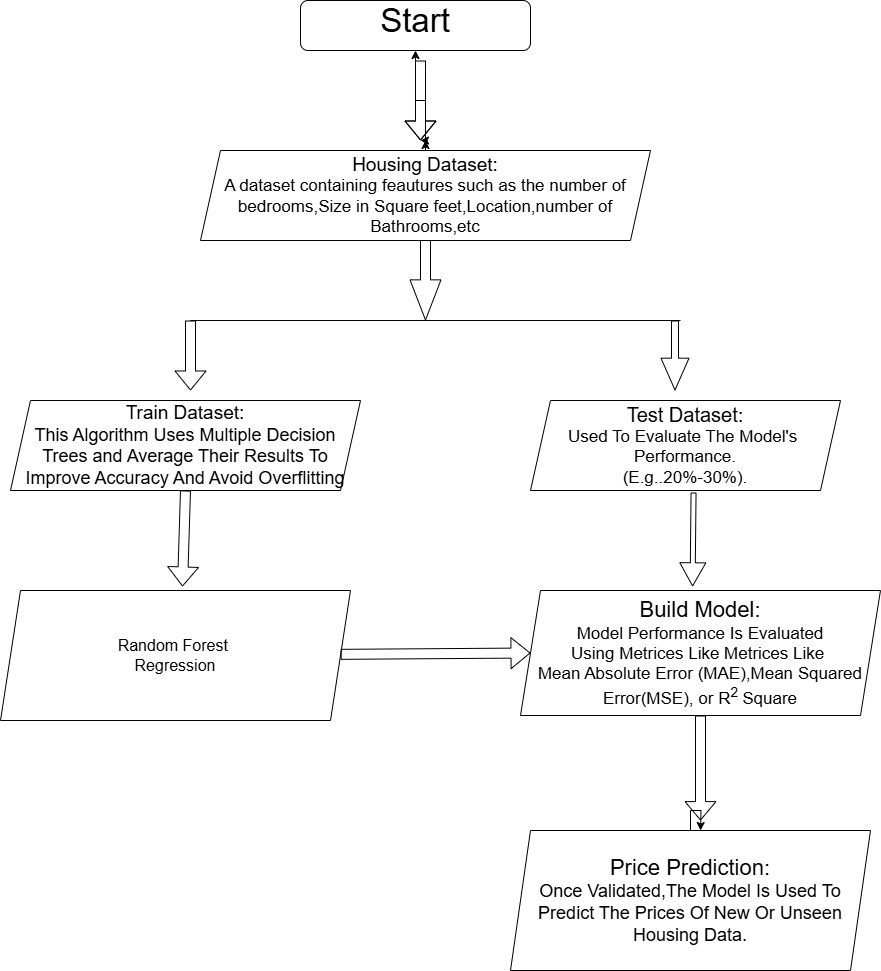
**1. Problem Statement**

The goal of this project is to predict housing prices based on various features such as the number of bedrooms, area, furnishing status, availability of hot water, and more. This is a **regression** problem because the target variable (price) is continuous. Initially, the problem was understood as a basic prediction task, but upon further exploration of the dataset, additional categorical and numerical variables were discovered that influence the model. Solving this problem is important for real estate agencies, homeowners, and buyers as it assists in setting fair market prices and making informed investment decisions.

**2. Project Objectives**

* Accurately predict house prices using machine learning models.
* Implement and compare two models: Linear Regression and Random Forest Regressor.
* Achieve low error rates (MAE, RMSE) and a high R2 score.
* Improve model interpretability with feature importance insights.
* Develop a web-based interface (Flask app) for real-time predictions.

**3. Flowchart of the Project Workflow**



**4. Data Description**

* **Dataset Name**: Housing.csv
* **Source**: Public (commonly available on Kaggle)
* **Type**: Structured
* **Number of Records**: 545 rows
* **Number of Features**: 13 features + 1 target (price)
* **Target Variable**: price
* **Dataset Nature**: Static

**5. Data Preprocessing**

* Checked for null values (no missing values found).
* Removed duplicate records.
* Converted categorical features using OneHotEncoder.
* Numerical columns were retained and scaled when necessary.

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

* **Univariate Analysis**:
  + Histograms for area, bedrooms, bathrooms.
* **Bivariate/Multivariate Analysis**:
  + Correlation heatmap showed strong correlation between various features.
  + Boxplots indicated the price variation across various features.
* **Insights Summary**:
  + Area, number of bathrooms, and air conditioning positively impact price.
  + Furnishing status also affects the house price significantly.

**7. Feature Engineering**

* Categorical encoding via One Hot Encoding, Label Encoding.
* Scaling is done for all numerical values.
* No new features added due to the simplicity and clarity of existing features.
* Polynomial features were not applied as models performed well without them.
* Dimensionality reduction was not required due to a manageable number of features.

**8. Model Building**

* **Models Used**:
  + Linear Regression: Chosen for interpretability and baseline comparison.
  + Random Forest Regressor: Chosen for its ensemble learning power and ability to handle non-linear relationships.
* **Train-Test Split**: 80-20 split
* **Evaluation Metrics**:
  + **Linear Regression**:
    - MAE: ~1.12 Lakh
    - RMSE: ~1.45 Lakh
    - R2 Score: ~0.67
  + **Random Forest**:
    - MAE: ~0.85 Lakh
    - RMSE: ~1.20 Lakh
    - R2 Score: ~0.82

**9. Visualization of Results & Model Insights**

* **Residual Plot** : Showed homoscedasticity for Random Forest and suggesting the model might underperform for higher price ranges for Linear Regression.
* **Scatter Plot**: Showed actual vs predicted price for both Random Forest and Linear Regression.
* **Feature Importance Plot**:
  + Top features: area, air conditioning, furnishing status, bathrooms

**10. Tools and Technologies Used**

* **Programming Language**: Python
* **IDE**: VS Code
* **Libraries**: pandas, numpy, seaborn, matplotlib, scikit-learn
* **Web Framework**: Flask
* **Visualization**: seaborn, matplotlib
* **Deployment**: Flask (runs on localhost with automatic browser launch)