**Week 8**

**Implementation of First Module**

In the implementation of the Real Estate Explorer system, the first and most foundational step is the collection and preparation of real estate datasets sourced from multiple Indian cities, including Bangalore, Mumbai, Delhi, Hyderabad, Pune, and Chennai. These datasets contain crucial information about various residential properties, with features such as the name of the location, price of the property, number of bedrooms (BHK), total area in square feet, and other attributes that define the nature and value of each listing. Once the raw datasets are collected, they are brought into the Python environment for analysis and transformation using the Pandas library. Pandas is a powerful data manipulation tool that allows for efficient handling of structured data in tabular form. The primary objective at this stage is to clean and standardize the data so that it becomes consistent and usable for machine learning algorithms.

Data preprocessing is a crucial step in any machine learning pipeline. In this project, datasets for each city (Bangalore, Pune, and Delhi) were first analyzed for inconsistencies such as missing values, non-standardized units (e.g., square meters vs. square feet), and outliers (e.g., properties with extremely large area but low price). Feature selection was done to retain only the most relevant columns affecting house price, such as area, number of bathrooms, balconies, and bedrooms (BHK). A Random Forest Regression model was chosen due to its robustness, non-linearity, and ability to handle complex feature interactions. Random Forest operates by building multiple decision trees during training and outputting the average prediction of all trees, which reduces variance and prevents overfitting. Models were separately trained for each city to ensure local price dynamics were captured accurately. These models were saved using joblib for integration into the Streamlit application.