**Project Journal**

**Members:**  Shruthi Chandra Babu; X23248556

**Dataset Name:** [Quicker Property Listing Dataset]

### **1. Project Summary and contribution**

#### **1.1 Dataset Selection and Collection**

* *Dataset used:* Quicker Property Listing Dataset.
* *Source of the dataset:* [QuickerPropertyDataset](https://www.quikr.com/homes/property/residential-apartments-for-sale-in-bangalore-cid_23?q=eyJwcm9wZXJ0eV90eXBlIjpbImFwIl0sImNsdXN0ZXJDaXR5IjpbIjIzIl0sImF2YWlsYWJsZV9mcm9tIjoiMCJ9).
* *Tools/Technologies used*: Selenium.
* *Reason for Selection:*

The data is extracted from a popular source of property listing ie., Quickr, where I have extracted the data from the city Bangalore in India. Key features such as BuiltUpArea, AreaName,Price provide a good understanding of real time market trend. The data captured for around 100 pages gives a comprehensive detail for predicting prices and the dyanamic data extraction ensures that the data collected are up to date.

* ***Key Challenges*** 
  + *Identifying the xpath for finding dyanamic elements via selenium tool. The xpath identified were identified by inspecting over the browser element.*
  + *Chrome WebDriver was used for accessing the Quickr property site and since Chrome Driver required the browser session to be active , the Webdriver is opened in headless, which eliminates the dependency to have an active browser session.*
  + Additional processing of extracted data, such as handling missing Rera Status in property listing or format modification to extract fields such as NumberofBHK and AreaName, were handled

#### **1.2 Database Integration**

* Database used: MongoDB
* Process: The semi structed web scraped data(json format) collected is stored into MongoDB. DB collection for house price prediction is created and data is programatically retrieved from collection.
* Database used: Postgres
* Process: After the data is manipulated , to store it in valid format, the structured data is stored in Postgres.
* For both the database, the connection details are retrieved from config.json file, to prevent data from being exposed in application code.

#### **1.3 Data Manipulation**

After displaying the data charecterictics using methods such as dataframe.shape, dataframe.columns etc.. the data is manipulated to store it in valid format.

* *Tools/technologies used:* Python, Pandas, NumPy

#### **1.4 Data Visualisation**

* *Visualisations created:*
* *Pie charts- to show the data distribution of categorical data.*
* *Scatter plot- to show the linear regression between built up area and price.*

Interactive Dashboard- (Application link: [Streamlit](https://housepriceprediction-quickerlisting.streamlit.app/))

Streamlit is used for creating an interactive dashboard from the cleaned data, where the cleaned data is displayed in tabular format. The distribution of property data is displayed in a real time map and also property density is displayed in another real time map. Additional Validations such as negative validation when there is no data available is handled.

#### **1.5 Data Preprocessing**

The categorical data are encoded using LabelEncoder and the continuous data are scaled using MinMaxScaler(Normalisation)

* *Tools/technologies used:* Python, Pandas, NumPy

#### **1.6 AI Modelling and Evaluation metrics**

The preprocessed data is evaluated using XGBoostRegressor and Artificial Neural Network. Various parameters have been verified in both the algorithm to get optimal results .

#### **1.7 Documentation and Reporting**

* *Sections authored: I had authored the complete report for Quickr Dataset, including all section such as Introduction, Related Work, Methodology, and Results/Evaluation.*
* *Contribution to final report:* I had proof read the report and converted to IEEE format.

### **2. Time Log**

Provide a breakdown of your time spent on each task

|  |  |
| --- | --- |
| **Task** | **Hours Spent** |
| Dataset Selection | 5 hours |
| Data Collection | 8 hours |
| Data Manipulation | 4 hours |
| Data Visualization and Streamlit Application | 8 hours |
| Data Pre processing | 1 hour |
| Model Training and Evaluation | 6 hours |
| Report Creation | 8 hours |
| Report Formatting | 2 hours |
| PPT Preparation | 2 hours |