Project Topic:

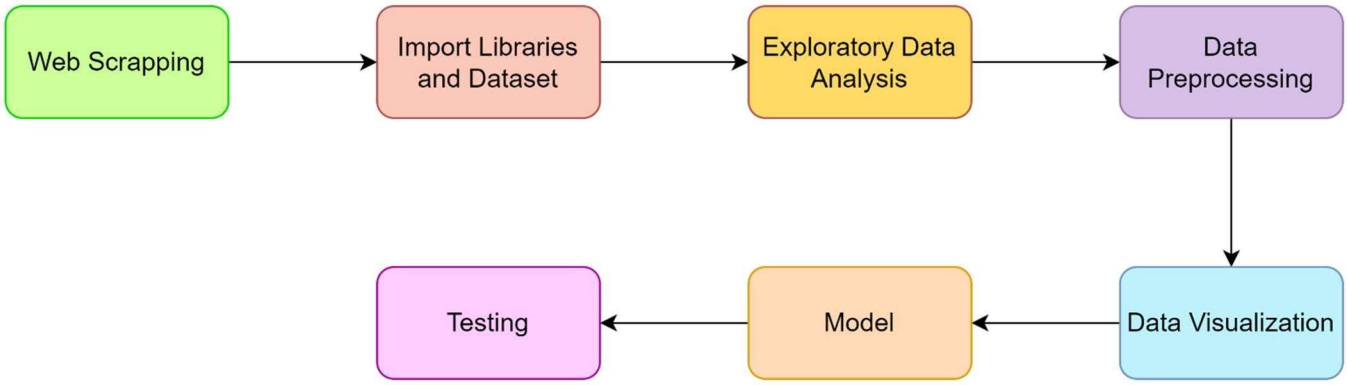
HOUSE PRICE PREDICTION

Introduction:

House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer to determine the selling price of a house and can help the customer to make arrangements on the right time to purchase a house. There are three factors that influence the price of a house which include physical conditions, facilities of area nearby and location.

Project Pipeline/implementation:

The various steps involved in this project are -



1. Web Scraping:

- Web scraping, also known as Data scraping. It is the process of importing information from a website into a spreadsheet or local file saved on your computer.
- There are various techniques to scrap data from web. In our project, we had scrapped data using beautifulsoup and Selenium from real-estate website.

2. Import Necessary Dependencies of Dataset:

- Various Machine Learning Libraries like NumPy, pandas, seaborn, matplotlib and sklearn should be imported in code to carry out the implementation process.

- We can import the Scraped Dataset by reading and loading it from the CSV file.

3. Exploratory Data Analysis:

- Exploratory Data Analysis involves exporting the five top records of data & columns/features in data.
- Also, it involves finding the length & shape of the dataset, gaining data information like Datatypes, checking for null, unique and number of target values.

4. Data Preprocessing:

- Data cleaning can be applied to filling in missing values, remove noise, resolving inconsistencies, identifying and removing outliers in the data.
- Data preprocessing is a predominant step in machine learning to yield highly accurate and insightful results.
- Greater the quality of data, greater is the reliance on the produced results.
- Incomplete, noisy, and inconsistent data are the properties of large real-world datasets.
- Data preprocessing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise and resolving inconsistencies.
- We have performed pre-processing steps on dataset like removing of zipcode, day, month and year.

5. Data Visualization:

- Data visualization is the graphical representation of information and data. It is a particularly efficient way of communicating when the data is numerous as of our dataset.
- For our project we used the following visualization techniques to get information about dataset
    - Bar Plot
    - Count Plot
    - Scatter Plot
    - Heatmap

6. Model Building:

In the problem statement we have used three different models respectively:

- Multiple Linear Regression
- Decision Tree
- Random Forest Regression

The idea behind choosing these models is that we want to try all the classifiers on the dataset ranging from simple ones to complex models and then try to find out the one which gives the best performance among them.

7. Conclusion

- Upon evaluating all the models, we can conclude the following details –

	Model	Score
0	Multiple Linear Regression	70.755851
1	Decision Tree	76.129475
2	Random forest Regression	86.990842

From the above it is clear that random forest accuracy is 87% . So, Random Forest is a suitable model for predicting the price of the house.

8. References

- Data Mining: Concepts and Techniques Second Edition, Jiawei Han, Micheline Kamber.
- https://towardsdatascience.com/create-a-model-to-predict-house-prices-using-python-d34fe8fad88f
- https://www.researchgate.net/publication/347584803_House_Price_Prediction_using_a_Machine_Learning_Model_A_Survey_of_Literature