***A PROJECT ON***

# “HOUSE PRICE PREDICTION”

SUBMITTED IN

PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE COURSE OF

DIPLOMA IN BIG DATA ANALYSIS



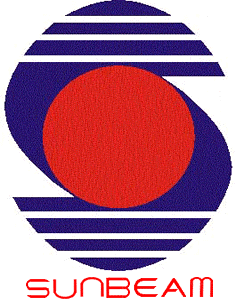
**SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY, PUNE**

Submitted By:

Amogh Jaronde (87213)

Narayan Attarde (86859)

**Mr.Nitin Kudale Mrs.Manisha Hingne** Centre Coordinator Course Coordinator



**CERTIFICATE**

This is to certify that the project work under the title ‘House Price Prediction’ is done by Amogh Jaronde & Narayan Attarde in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

Mr. Aniket P Mrs. Manisha Hingne

**Project Guide** **Course Coordinator**

Date: 11/02/2025

# ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Manisha Hingne (Course Coordinator, SIIT ,Pune) and Project Guide Mr. Aniket P.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Amogh Jaronde

DBDA Aug 2024 Batch,

SIIT Pune

Narayan Attarde

DBDA Aug 2024 Batch, SIIT Pune

**TABLE OF CONTENTS**

1. **Introduction**
   1. Introduction And Objectives
   2. Why this problem needs To be Solved?
   3. Dataset Information

## Problem Definition and Algorithm

* 1. Problem Definition
  2. Algorithm Definition

## Experimental Evaluation

* 1. Methodology/Model
  2. Exploratory Data Analysis

## Results And Discussion

1. **GUI**
2. **GitHub link**

## 7.Future Work And Conclusion

* 1. Future Work
  2. Conclusion
     1. **Introduction**
        1. **Introduction And Objectives:**

House price prediction is a critical task in real estate, helping buyers, sellers, and investors make informed decisions. The objective of this project is to develop a machine learning model that can accurately predict house prices based on various features such as location, size, amenities, and market trends. The model aims to assist individuals and businesses by providing reliable estimates, reducing market uncertainties, and enabling better investment planning.

## Why this problem needs To be Solved?

Real estate pricing is influenced by multiple dynamic factors, making it challenging for individuals to determine fair prices. Often, sellers overprice their properties, and buyers lack sufficient knowledge to assess the actual market value. Real estate agents may have biases or limited data, leading to suboptimal pricing decisions. A predictive model can provide insights based on historical data, reducing bias and aiding in better decision-making for buyers, sellers, and real estate firms. Furthermore, investors can use the model to assess market trends, forecast price movements, and identify profitable opportunities.

## Dataset Information.

The dataset for this project was scraped using Selenium from various housing websites, specifically for properties in Pune and Mumbai. The collected data includes:

* **Property Name: Describes the type of property (1 BHK, 2 BHK etc.)**
* **Project Name: Identifies the name of the development project.**
* **Location: Specifies the neighborhood or area within Mumbai and Pune.**
* **City: Indicates the city (Mumbai , Pune )**
* **Price: Lists the selling price of the property.**
* **Area ( sqft ): Provides the total area of the apartment in square feet.**
* **Construction Status: Under Construction and Ready to move**
* **Possession Date: Indicates how many years old is the property.**
* **Bathrooms: States the number of bathrooms in the apartment.**
* **Developer: Names the company or individual responsible for developing the property.**

After data collection, preprocessing and transformation were performed to derive meaningful features for model training. Missing values were handled, categorical variables were encoded, and feature scaling was applied for optimal model performance.

## Problem Definition and Algorithm:

* + - 1. **Problem Definition**

The goal of this project is to predict house prices based on various features extracted from real estate listings. The challenge involves handling unstructured data, missing values, and feature selection to improve model performance. The complexity arises due to fluctuating market conditions, differences in property attributes, and regional price variations. The model needs to be robust enough to generalize well across different price ranges and locations.

## Algorithm Definition

After experimenting with multiple algorithms, Random Forest emerged as the best-performing model in terms of accuracy. The following steps were followed:

* Data Cleaning and Transformation: Handling missing values, outliers, and inconsistent entries
* Feature Engineering: Creating new relevant features, transforming categorical data, and normalizing numerical variables
* Model Selection and Tuning: Testing multiple models such as Linear Regression, Decision Tree, XGBoost, and Random Forest
* Hyperparameter Optimization: Tuning key parameters using RandomizedSearchCV

# Random Forest Regressor

rf\_param\_grid = {

'n\_estimators': [100, 300],

'max\_depth': [10, 20],

'min\_samples\_split': [5, 10],

'min\_samples\_leaf': [2, 4],

'bootstrap': [True]

}

* Model Evaluation: Using RMSE, MAE, and R² score to determine model effectiveness

Random Forest was chosen due to its ability to handle non-linearity, robustness to outliers, and feature importance ranking. The ensemble learning approach of Random Forest reduces overfitting and provides more stable predictions compared to single decision trees.

* **Linear Regression**: A simple model that assumes a linear relationship between house prices and features. It’s easy to interpret but struggles with complex patterns.
* **Decision Tree**: A tree-based model that splits the data into decision nodes, making predictions based on feature importance. It’s prone to overfitting but captures non-linearity well.
* **XGBoost**: A powerful gradient boosting algorithm that improves accuracy by combining multiple weak learners. It’s efficient and handles missing data well but requires careful tuning.
* **Random Forest**: An ensemble of decision trees that reduces overfitting and improves generalization. It provides high accuracy and stability, making it the best choice for this problem.

## Experimental Evaluation:

* + - 1. **Methodology:**

The methodology involved the following steps:

* **Data Collection**: Scraped data using Selenium from multiple real estate websites
* **Data Preprocessing**: Handled missing values, performed feature engineering, and applied data transformations
* **Exploratory Data Analysis (EDA)**: Visualized distributions, checked correlations, and identified key variables
* **Model Training**: Trained multiple machine learning models and selected the best performer
* **Model Evaluation**: Evaluated models based on key performance metrics and selected the optimal model for deployment
* **Model Deployment:** Deployed the selected model for real-time predictions or batch scoring.

## Flow Diagram :

**A diagram of a model

AI-generated content may be incorrect.**

**Exploratory Data Analysis**

EDA was performed to understand the distribution and relationships of variables:

* **Price Distribution**: Analysis of price variations in different areas
* **Feature Correlation**: Identified how different attributes impact house prices
* **Visualizations**: Used scatter plots, heatmaps, and bar charts to gain deeper insights into data trends

## Results and discussion:

After multiple trials, Random Forest achieved the best accuracy. The final model was evaluated based on testing accuracy:

* **MAE**: 2546415.826799
* **R² Score**: 0.87

The model effectively captures price trends and provides reliable predictions. The results indicate that key factors influencing prices include location, property type, and amenities. Feature importance analysis revealed that location had the highest impact, followed by property size and number of bedrooms

A graph of a number of blue rectangular bars

AI-generated content may be incorrect.

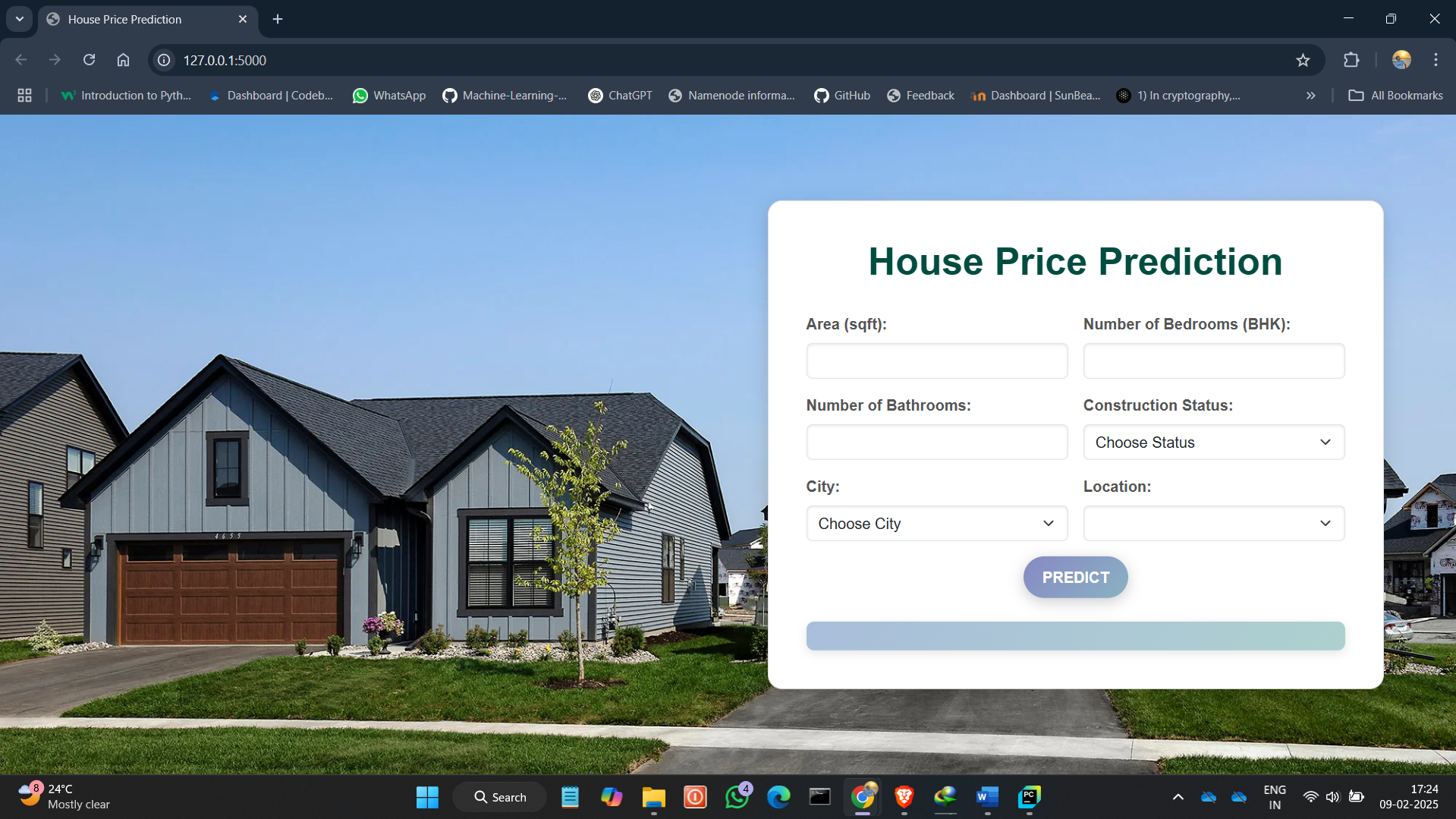
A graph with numbers and text

AI-generated content may be incorrect.

## GUI:

A user-friendly interface was developed using Flask, allowing users to input house details and get predicted prices in real-time. Its Running on http://13.60.187.226:8080 The GUI enables:

* **User Input**: Users can enter property details like location, size, and amenities
* **Prediction Display**: The predicted price is displayed instantly



**6.GitHubLink:**

[**https://github.com/B1-86859-Assignment/Project-CDAC-.git**](https://github.com/B1-86859-Assignment/Project-CDAC-.git)

## 7.Future work And Conclusion 7.1Future Work:

* Expanding the dataset to include more cities for better generalization
* Enhancing feature engineering for better prediction accuracy
* Exploring deep learning models such as neural networks for improved price forecasting
* Deploying the model as a web application with cloud-based scalability
* Integrating real-time data updates to reflect market fluctuations

## 7.2 Conclusion:

The project successfully developed a house price prediction model using machine learning. Random Forest provided the best accuracy among tested models, and a Flask-based GUI was implemented for user interaction. The model demonstrates strong predictive performance, helping individuals and businesses make data-driven real estate decisions. Future improvements, such as deep learning integration and broader dataset inclusion, can enhance model accuracy and usability.