



## HOUSE PRICE PREDICTION PROJECT

Submitted by:

ABHILASHA MEWADA

## INTRODUCTION

House is one of human life's most essential needs, along with other fundamental needs such as food, water, and much more. Demand for houses grew rapidly over the years as people's living standards improved. While there are people who make their house as an investment and property, yet most people around the world are buying a house as their shelter or as their livelihood.

An increase in house demand occurs each year, indirectly causing house price increases every year. The problem arises when there are numerous variables such as location and property demand that may influence the house price, thus most stakeholders including buyers and developers, house builders and the real estate industry would like to know the exact attributes or the accurate factors influencing the house price to help investors make decisions and help house builders set the house price.

House price prediction can be done by using a multiple prediction models (Machine Learning Model) such as support vector regression, artificial neural network, and more. There are many benefits that home buyers, property investors, and house builders can reap from the house-price model. This model will provide a lot of information and knowledge to home buyers, property investors and house builders, such as the valuation of house prices in the present market, which will help them determine house prices. Meanwhile, this model can help potential buyers decide the

Characteristics of a house they want according to their budget.

Previous studies focused on analysing the attributes that affect house price and predicting house price based on the model of machine learning separately. However, this article combines such a both predicting house price and attributes together

## Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem:

We are checking the null values and join the data frames by concat method, we also use descriptive techniques and different visualise technique using seaborn and matplotlib

- Data Sources and their formats:

Csv files

- Data Pre-processing Done:

Filling all null values

Encoding using hot encoder

Selecting a variable and test train split

One-hot encoding is a technique that is used to convert categorical features to a suitable format to be used as an input in Machine Learning algorithms. It transforms a single variable with  $n$  observations and  $d$  distinct values to  $d$  binary variables, where each observation indicating the presence as 1 or absence as 0. In one-hot encoding, the categories are represented as independent concepts.

- Hardware and Software Requirements and Tools Used

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import
train_test_split,GridSearchCV,cross_val_score
from sklearn.preprocessing import
StandardScaler,MinMaxScaler,power_transform
from sklearn.linear_model import
LinearRegression,LogisticRegression
```

```
from sklearn.metrics import r2_score,  
mean_squared_error, accuracy_score from sklearn.ensemble  
import AdaBoostRegressor
```

## **Model/s Development and Evaluation**

A Random Forest is an ensemble technique qualified for performing classification and regression tasks with the help of multiple decision trees and a method called Bootstrap Aggregation known as Bagging.

### **Metrics:**

Several evaluation metrics measure the performance of machine learning algorithms such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared, and Mean Absolute Error (MAE). However, in this study, the performance of the algorithms is measured by using RMSE and R-Squared. Root Mean Square Error (RMSE) is used as an evaluation metric in machine learning to measure the performance of the model. However, RMSE is similar to the Mean Square Error (MAE).

## **CONCLUSION**

The results were promising for the public data due to it being rich with features and having strong correlation, whereas the local data gave a worse outcome when the same pre-processing strategy was implemented due to it being in a different shape compared with the public data in terms of the number of features and the correlation strength.