International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 11 | Nov 2022 www.irjet.net p-ISSN: 2395-0072

A Comparative Study on House Price Prediction using Machine Learning

Hardi Joshi[1], Saket Swarndeep[2]

[1]*Post Graduate Scholar, Dept. Computer Engineering (Software Engineering), L J University, Ahmedabad, Gujarat, India*
[2]*Assistant Professor, Dept. Computer Engineering (Software Engineering), L J University, Ahmedabad, Gujarat, India*

Abstract - *Buying a house is one of the biggest financial goal of everyone. Owning a house is not only a basic need but it also represents prestige. However, buying a house is one of the most crucial decision of a person's life as there are so many factors to be consider before buying a property. House prices keeps changing based on location, area, population, house condition and structure, availability of parking, backyard, size of house etc. From past few years a lot of data has been generated regarding Real Estate. Machine learning prediction techniques can be very useful to predict an accurate pricing of the houses. The study focuses on developing an accurate prediction model for house price prediction. Machine learning is sub-branch of artificial intelligence that deals with statistical methods, algorithms. Using machine learning we can build a model which can make prediction based on past data. In this paper we will review different machine learning algorithms which can be used for house pricing prediction.*

Key Words: **Machine learning models, house price prediction, real estate, price prediction, Machine learning algorithms**

1. INTRODUCTION

Even in today's modern world owning a house is biggest financial achievement for many. Also, most people buy house only once in their lifetime and spend their entire life in same house creating so many memories. However, buying a house or selling a house without knowing the actual value of house can be dangerous. One must know exact price of house before finalizing the deal. As we know price of house depends on many factors like Area, location, population, size and structure of house, number of bedrooms & bathrooms provided, parking space, elevator, style of construction, balcony space, condition of building, price per square foot etc. So, our model must be able to produce an accurate result considering all these different factors. In this paper we will discuss about different machine learning techniques which are suitable for house price prediction. There is brief overview about machine learning techniques like linear regression, multiple linear regression, random forest, regression tree, neural network etc.

2. Related Work

A house is a basic necessity of a person. However, most people make Hugh mistakes while dealing with properties. Striking any kind of deal without knowing actual valuation of property can be a financial burden. Our aim is to develop a prototype which can be beneficial with real estate. Machine learning (ML) has been used as one of the powerful tools to deal with the data. Due to the flexibility, machine learning models have been developed in a variety of application domains. Several studies have used machine learning algorithms for housing price predictions. In addition, ML models have been implemented on housing datasets for various types of prediction. [1] The study focuses on developing an accurate prediction model for house price prediction using machine learning algorithms.

2.1 Machine Learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed. The purpose of machine learning is to learn from the data. Many studies have been done on how to make machines learn by themselves without being explicitly programmed. Many mathematicians and programmers apply several approaches to find the solution of this problem which are having huge data sets.[2]

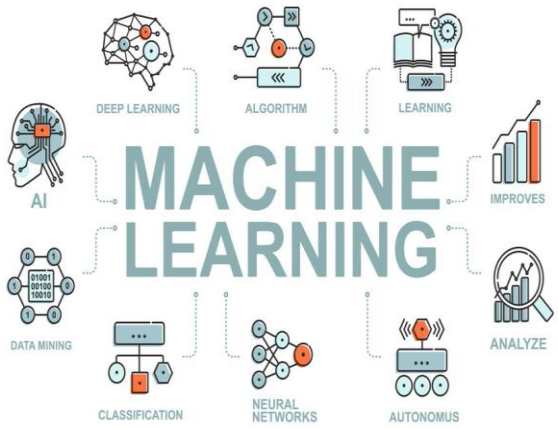

Fig 1: Machine Learning [3]

2.1.1 Supervised Learning in ML

Supervised Machine learning is a method of inputting labelled data into a machine learning model. The model is trained with known input and output data so that it can predict future outputs accordingly.[4]

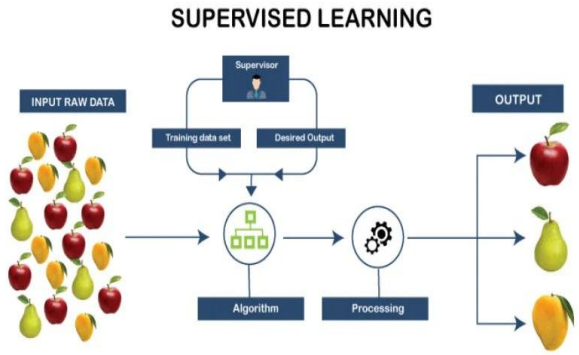

Fig 2: Supervised Learning [4]

2.1.2 Unsupervised Learning in ML

In unsupervised learning we don't have to direct the model. It predominantly manages the unlabeled information. Unsupervised learning algorithms include anomaly detection, clustering, neural networks, etc. [4]

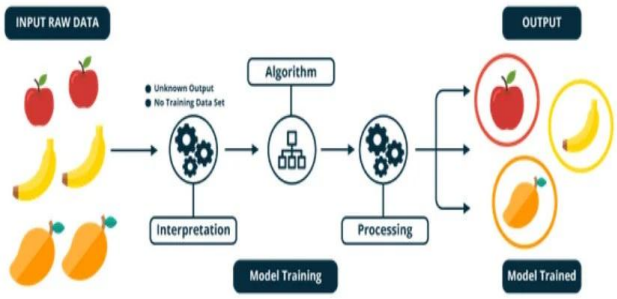

Fig 3: Unsupervised Learning [4]

2.1.3 Prediction in ML

Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome. Machine learning model predictions allow us to make highly accurate guesses as to the likely outcomes of a question based on historical data.[5] There are many machine learning algorithms that can be used for prediction like linear regression, multiple linear regression, random forest, regression tree, neural network etc.

3. Literature Review

This paper estimates the changes in the house pricing. Housing price is strongly correlated to other factors such as location, area, population etc. In this paper applies both traditional and advance machine learning techniques and will discuss the result of this different techniques. A dataset named "Housing Price in Beijing" is used. In this paper we will discuss machine learning techniques like Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Hybrid Regression and Stacked Generalization.[6]

In this paper, machine learning technique Random Forest is used to build the machine learning model for house price prediction. As we know when it comes to predicting pricing of house a lot of factors affect the price. In this paper we study that Random Forest works better than benchmark model linear regression. By including features such as zip code, longitude and latitude, which are not linearly related to house price, we found that random forest model performs much better and captures the hidden information in those features. We will use data of North Virginia house price.[7]

In this paper, house price prediction is based on historical data. The goal of this study is through analyzing a real historical transactional dataset to derive valuable insight into the housing market in Melbourne city. In this paper different machine learning techniques are used like Linear Regression, Polynomial Regression, Regression Tree, Neural Network, and SVM. In this paper "Melbourne Housing Market" dataset is used.[8]

In this paper, machine learning algorithms are used to predict the price of Singapore housing Market. Then the performance of this techniques is compared to compares the predictive performance of artificial neural network (ANN) model. The objective of this study is to compare the predictive performance of artificial neural networks (ANN) with autoregressive integrated moving average (ARIMA) and multiple regression analysis (MRA). The housing prices are influenced by many different factors. The main variables considered in the design of our ANN models are consumer spending, monthly average wages, gross domestic product (GDP), consumer price index, prime lending rate, real interest rate, population, the Singapore Housing and Development Board (HDB) resale price index, change of HDB resale price index, Straits Times Index, the number of available condominiums, and the condominium price index (CPI). The results show that ANN model can generate high accuracy and works better than ARIMA & MRA.[9]

In this paper there are three factors that are considered for the price of a house which includes physical conditions, concepts and location. The objective of the paper is prediction of residential prices for the customers considering their financial plans and needs. This examination means to predict house prices in Mumbai city

with Linear Regression. Linear Regression will predict the exact numerical target value unlike other models which can only classify the output. MAE, MSE, RMSE are used to check the quality of model.[10]

In this paper, different regression methods are used to predict house price prediction. The stacking algorithm is applied on various regression algorithms to see which algorithm has the most accurate and precise results. Apart from using the regression algorithms, some classification algorithms such as SVM algorithm, decision tree algorithm, Random Forest classifier etc. are taken into consideration and applied on our house pricing dataset. The sales prices have been calculated with better accuracy and precision.[11]

This paper describes how to use machine learning algorithms to predict house price. For a specific house price, it is determined by location, size, house type, city, country, tax rules, economic cycle, population movement, interest rate, and many other factors which could affect demand and supply. For local house price prediction, there are many useful regression algorithms to use. For example, support vector machines (SVM), Lasso (least absolute shrinkage and selection operator), Gradient boosting, Ridge, Random Forest. The paper also proposes a hybrid Lasso and Gradient boosting regression model to predict individual house price. the result proves the coupling effect of multiple regression algorithms. Based on the result, the hybrid regressions are better than one from Ridge, Lasso or Gradient boosting regression.[12]

In this paper, various regression techniques are used in pathway, and the results are not sole determination of one technique rather it is the weighted mean of various techniques to give most accurate results. They have used linear regression, forest regression and boosted regression. The results of all three algorithms are fed as input to the neural network. The system makes optimal use of Linear Regression, Forest regression, Boosted regression. The efficiency of the algorithm has been further increased with use of Neural networks.[13]

In this paper, Random Forest machine learning technique is used for house price prediction. Boston housing dataset with 506 entries and 14 features were used to evaluate the performance of the proposed prediction model. The fundamental step taken for the implementation include data collection, data exploration which was used to understand the datasets and identify features in the dataset; data pre-processing stage which was used to clean the dataset to make it suitable for model development. Afterwards the model was developed using the proposed random forest algorithm. A comparison of the predicted and actual prices predicted revealed that the model had an acceptable predicted value when compared to the actual values with an error margin of ±5. [14]

In this paper, different Regression techniques like Multiple linear, Ridge, LASSO, Elastic Net, Gradient boosting, and Ada Boost Regression. From the experiment results, gradient boosting algorithm has high accuracy value when compared to all the other algorithms regarding house price predictions. [15]

4. Algorithm Studied

4.1 Linear Regression

Linear regression is a supervised learning technique. It is responsible for predicting the value of a dependent variable (Y) based on a given independent variable (X).[16]

Equation: $Y = mX + b$

Y is dependent variable.

X is independent variable.

M is estimated slop.

B is estimated intercept.

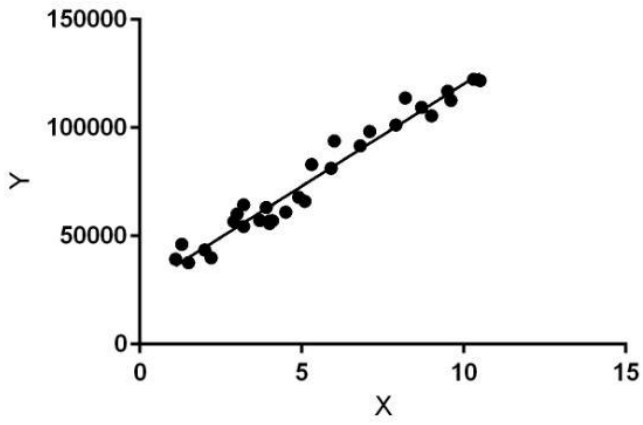

Fig 4: Linear Regression [17]

4.2 Multiple Linear Regression

Multiple Linear Regression a new version of the linear regression which is more powerful which works with the multiple variables or the multiple features it helps to predict the unknown value of the attribute from the known value of the two or more attributes which will be also known as the predictors.[18]

4.3 LASSO Regression

LASSO stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is Vast enough to improve those inclination of the model on over-fit. Least ten variables can foundation over fitting and Huge enough will cause computational tests. [19]

4.4 Decision Tree

Decision Tree is a tool, which can be employed for Classification and Prediction. It has a tree shape structure, where each internal node represents test on an attribute and the branches out of the node denotes the test outcomes. Once the Decision Tree is formed, new instances can be classified easily by tracing the tree from root to leaf node. Classification through Decision Tree does not require much computation. Decision Trees are capable of handling both continuous and Categorical type of attributes.[20]

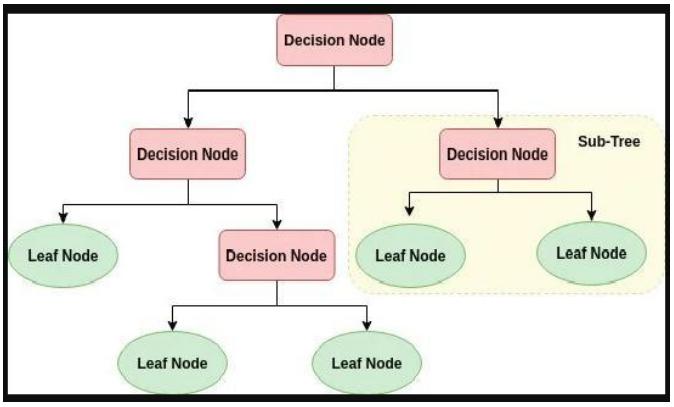

Fig 5: Decision Tree [20]

4.5 Random Forest

RF is a regression technique that combines the performance of numerous DT algorithms to classify or predict the value of a variable. That is when RF receives an (x) input vector, made up of the values of the different evidential features analyzed for a given training area, RF builds a number K of regression trees and averages the results.[21]

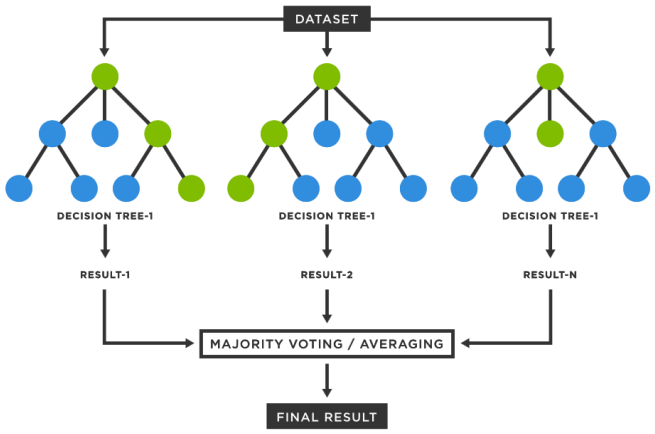

Fig 6: Random Forest [22]

4.6 Neural Network

As a neural network in our brain, ML neural network also contains neurons, synapses, and layers. A neural network contains an input layer — a set of input features. Also, it usually contains one or more hidden layers. Each layer contains some number of nodes as neurons, and links as synapses. The last layer called "output layer" is the layer with answers.[23]

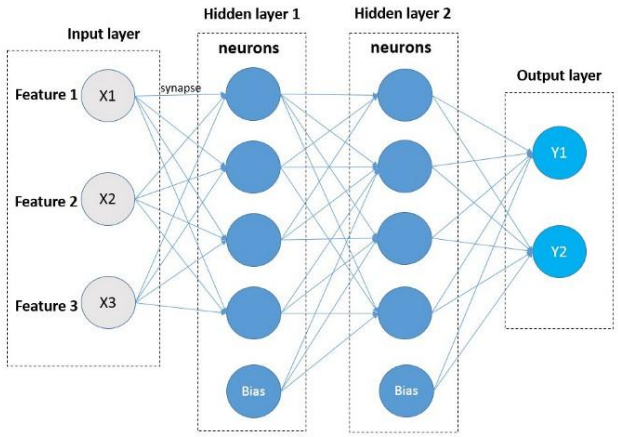

Fig 7: Neural Network [23]

5. Dataset Description

The dataset used in this study is downloaded from Kaggle website. dataset has 6347 observations and 19 variables.

This data contains scraped data and has the following information:

1) Prices of houses all over Mumbai along with their location.

2) Information about house condition (new/resale) and area of the house.

3) Information about various amenities provided.

6. Proposed System

As we know there are many ML algorithms that can be used for house price prediction. Every algorithm has its advantages and disadvantages. So, we will use ensemble learning to build our system. In ensemble learning we can combine set of individual learners (base model) together and build our final model. Our main purpose of combining different base model is to improve prediction and achieve higher accuracy. Any machine learning algorithm can be base model such as linear regression, decision tree, random forest etc.

Step 1: Load the dataset.

Step 2: Data Pre-processing contains data cleaning, data editing, data reduction. Data cleaning is process where inaccurate data or if a data field is empty, then value is filled using mean or median or entire record is deleted from data. Data editing is process where outliers are picked from data and eradicated. Data reduction is termed as the process of reducing data using some kind of normalization for easy process of data. [25]

Step 3: Determine the Dependent and Independent variables. In our dataset price will be the dependent variable. And Independent variables will be Area, Location, No. of Bedrooms etc.

Step 4: Split dataset into training set and testing set. Training dataset will be used to train the model and testing dataset will be used to test the model.

Step 5: Training and Testing dataset will be given to different base models. Any machine learning algorithm can be base model. For example, linear regression, decision tree, random forest etc.

Step 6: Voting will select the best model for house price prediction. The model which gives highest accuracy and lowest error rate will be chosen.

Step 7: Final selected model will be given the training and testing dataset to predict the price.

Step 8: To evaluate the performance of the model we will use different measure of errors like Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE).

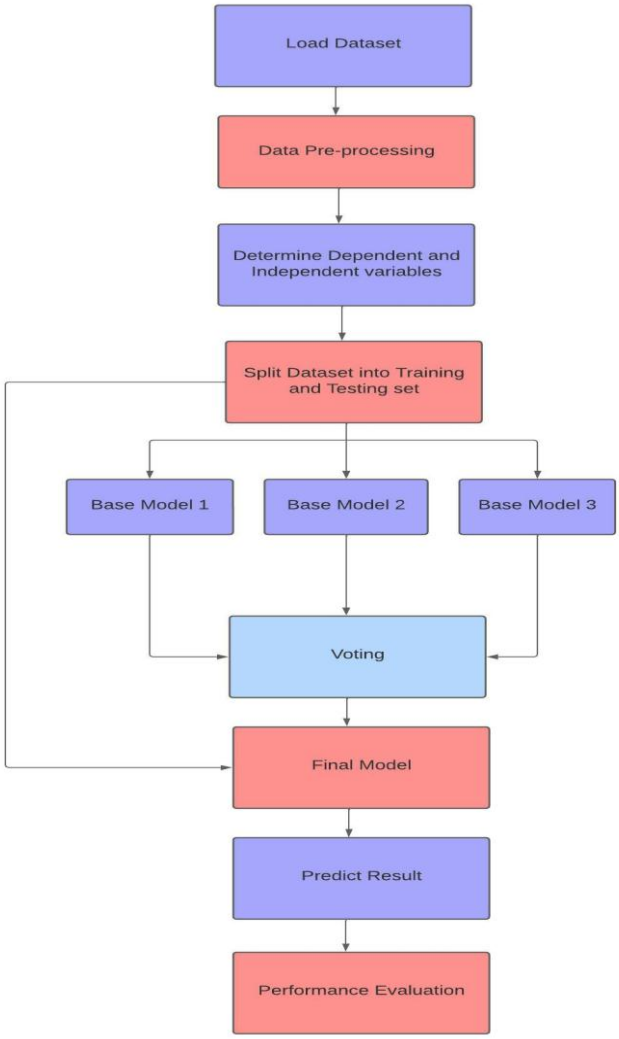

Fig 8: Proposed System

7. CONCLUSION

This study helps us to discover assets and liabilities of different machine learning models. As we know machine learning has plenty of algorithms that can be used for house price prediction. The existing systems focuses on single models only. We proposed to use multiple different model which can be used for prediction and focuses on more accurate results. We proposed to use ensemble learning method as it has capability of combining multiple ml models will help us discover different aspects of data. Hence, this methodology is anticipated to give higher accuracy compared to other single models.

REFERENCES

[1] Shahhosseini, M., Hu, G., & Pham, H. (2019, June). Optimizing ensemble weights for machine learning models: a case study for housing price prediction. In INFORMS international conference on service science (pp. 87-97). Springer, Cham.

[2]] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR). [Internet], 9, 381-386.

[3] An Introduction to Machine Learning, Its Importance, Types, and Applications. (2022, August 31). FORE School of Management. **https://www.fsm.ac.in/blog/an-introduction-to-machine-learning-its-importance-types-and-applications/**

[4] Do, T. (2022, September 20). Supervised and Unsupervised Machine Learning – Explained Through Real World Examples. Omdena | Building AI Solutions for Real-World Problems. https://omdena.com/blog/supervised-and-unsupervised-machine-learning/

[5] Prediction - DataRobot AI Cloud Wiki. (2022, July 21). DataRobot AI Cloud. https://www.datarobot.com/wiki/prediction/

[6] Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. Procedia Computer Science, 174, 433-442.

[7] Wang, C., & Wu, H. (2018). A new machine learning approach to house price estimation. New Trends in Mathematical Sciences, 6(4), 165-171.

[8] Phan, T. D. (2018, December). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. In 2018 International conference on machine learning and data engineering (iCMLDE) (pp. 35-42). IEEE.

[9] Lim, W. T., Wang, L., Wang, Y., & Chang, Q. (2016, August). Housing price prediction using neural networks. In 2016 12th International conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD) (pp. 518-522). IEEE.

[10] Ghosalkar, N. N., & Dhage, S. N. (2018, August). Real estate value prediction using linear regression. In 2018 fourth international conference on computing communication control and automation (ICCUBEA) (pp. 1-5). IEEE.

[11] Jain, M., Rajput, H., Garg, N., & Chawla, P. (2020, July). Prediction of house pricing using machine learning with Python. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 570-574). IEEE.

[12] Lu, S., Li, Z., Qin, Z., Yang, X., & Goh, R. S. M. (2017, December). A hybrid regression technique for house prices prediction. In 2017 IEEE international conference on industrial engineering and engineering management (IEEM) (pp. 319-323). IEEE.

[13] Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018, April). House price prediction using machine learning and neural networks. In 2018 second international conference on inventive communication and computational technologies (ICICCT) (pp. 1936-1939). IEEE.

[14] Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. Procedia Computer Science, 199, 806-813.

[15] Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019, March). House price prediction using regression techniques: a comparative study. In 2019 International conference on smart structures and systems (ICSSS) (pp. 1-5). IEEE.

[16] Amey Thakur, M. S. (2021). BANGALORE HOUSE PRICE PREDICTION.

[17] GeeksforGeeks. (2022, November 21). ML | Linear Regression. https://www.geeksforgeeks.org/ml-linear-regression/

[18] Ravikumar, A. S. (2017). Real estate price prediction using machine learning (Doctoral dissertation, Dublin, National College of Ireland).

[19] Satish, G. N., Raghavendran, C. V., Rao, M. S., & Srinivasulu, C. (2019). House price prediction using machine learning. Journal of Innovative Technology and Exploring Engineering, 8(9), 717-722.

[20] Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. International Journal of Information Engineering & Electronic Business, 12(2).

[21] Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 71, 804-818.

[22] What is a Random Forest? (n.d.). TIBCO Software. https://www.tibco.com/reference-center/what-is-a-random-forest

[23] Samoletskyi, B. (2021, December 28). Neural networks — how it works? - Analytics Vidhya. Medium. https://medium.com/analytics-vidhya/neural-networks-how-it-works-403233e2f159

[24] Housing Prices in Mumbai. (2020, August 27). Kaggle. https://www.kaggle.com/datasets/sameep98/housing-prices-in-mumbai

[25] Avanijaa, J. (2021). Prediction of house price using xgboost regression algorithm. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(2), 2151-2155.