International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8321
***Article Received:** 13 July 2023 **Revised:** 30 August 2023 **Accepted:** 19 September 2023*

House Price Prediction using Machine Learning Algorithms

Angulakshmi M [1*], Deepa M[2], Mala Serene I[3], Thilagavathi M[4], Aarthi P [5]
[1*]School of Computer Science Engineering and Information Systems,
Vellore Institute of Technology, Vellore, India
angulakshmi.m@vit.ac.in
[2]School of Computer Science Engineering and Information Systems,
Vellore Institute of Technology, Vellore, India
mdeepa@vit.ac.in
[3]School of Computer Science Engineering and Information Systems,
Vellore Institute of Technology, Vellore, India
imalaserene@vit.ac.in
[4]School of Computer Science Engineering and Information Systems,
Vellore Institute of Technology, Vellore, India
mthilagavathi@vit.ac.in
[5]School of Computer Science Engineering and Information Systems,
Vellore Institute of Technology, Vellore, India
aarthi.p2021@vitstudent.ac.in

Abstract— House prices are a major financial decision for everyone involved in the housing market, including potential home buyers. A major part of the real estate industry is housing. An accurate housing price prediction is a valuable tool for buyer and seller as well as real estate agents. The study is done for the purpose of knowledge among the people to understand and estimate the pricing of their houses. The prediction will be made using four machine learning algorithms such as linear regression, polynomial regression, random forest, decision tree. Linear Regression has good interpretability. Decision tree is a graphical representation of all possible solutions. Polynomial regression can be easily fitted to a wide variety of curves. Regression and classification issues are resolved with random forests .Among the given algorithm, Random forest provides better accuracy of about 89% for given dataset.

Keywords- Machine learning, House price prediction, Decision Tree, Random Forest, Polynomial regression

I. INTRODUCTION

Machine learning is focused on creating self-learning algorithms for projects that will base future activity on historical data. The prediction of house prices is based on a similar phenomenon. The market demand for housing is always rising each year as a result of population growth and people moving to other cities for financial reasons. The purchase of a house is one of the biggest and most significant decisions a family can make, since it consumes all the invested funds and covers them with loans [1]. Machine learning has become an important prediction approach in recent years, owing to the growing trend towards big data, because it can predict house prices more accurately based on their attributes, regardless of previous year's data. Predicting house prices can assist in determining the selling price of a house in a specific region and can assist people in purchasing the house at the right time [2].

The field of machine learning is used in a wide range of computing applications. Machine learning is important because it performs some tasks such as providing a view of trends in customer behaviour and business operational patterns and assisting in the development of new products. The cost of a home is determined by a number of interconnected factors. In this project, an attempt was made to build a predictive model for evaluating price based on price-influencing factors. Area, rooms, location, city, and other factors all have an impact on house prices [3]. For example, if we are going to sell a house, we need to know what price tag to put on it. The most important source of analysis and predictions for a real estate business is data. A business manager should always be aware of predictions of future variations of an entity so they can act accordingly to

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8321
***Article Received:** 13 July 2023 **Revised:** 30 August 2023 **Accepted:** 19 September 2023*

avoid losses in the future. The purchase of a home is a lifelong dream for most people, but many people do it incorrectly and cost them a lot of money [4]. As a result, the most accurate predicting model is required for analysis. To give correct price projections to real estate managers, we also need precise predictions on real estate and homes in the housing market. The paper's main objective is to analyze real estate prices utilizing machine learning methods and real-time data. The purpose of the whole statistical analysis is to clarify the relationship between housing characteristics and the methods through which these variables are used to estimate home values. Regression algorithms were compared to find the most accurate prediction model for housing values. As a result, avoiding errors would be helpful for the people. Machine learning has been employed to predict the prices of houses from sample dataset consisting of attributes which influences the cost. The scope of the project is to develop a machine learning model that can precisely estimate the cost of a home using a variety of variables from a dataset of house prices.

II. LITERATURE SURVEY

In the paper [5] Decision tree, Random Forest and Bagging method are used. The ensemble learning method outperforms other methods in classification and regression prediction. The classification method is used to build the prediction model and the transaction price serves as the category label. The decision tree in this paper performs exceptionally well, which is one of its main advantages. Decision trees are no longer used to sort linear problems. Using the decision tree method, the deviation of the training set can be used to determine whether the model is over fitting or under fitting. Bagging has the advantage of lowering the variance of the base classifier, which improves generalisation error. The disadvantages of this paper are due to noisy data, some of the values are missing in decision tree. The difference between real and predicted house prices is kept to a very low level. As a result, the model is accurate and practical for price forecasting.

In the paper[6] ,Decision Tree Regression, Multiple Linear Regression and XGBoost method are used. XGBoost uses a CART tree at the bottom, where nodes are values instead of categories, which allows efficient optimization and increases performance. Overfitting can be effectively avoided using feature sampling and regularisation. This paper has some drawbacks, including the difficulty of obtaining reasonably complete housing information. This paper has two reasons. A house transaction price is a trade secret of the intermediary and developer. There is a lack of transaction data in government departments.

In the paper[7] Decision Tree Classification, Naive Bayes Classification, AdaBoost Classification and Random forest Classification method are used. The advantages of this paper is that all the methods used in this article gives better performance for classification.AdaBoost Classification predicts the original data set and gives each observation equal weight. The disadvantage of this paper is that, due to the issue of monetary stability, Abbasov C et al. predicts the possibility of home commerce. In the paper[8], Cross validation Technique and K-means method is used. The benefit of this paper is that visualisation makes complex data more accessible, reasonable, and usable over time. The Ames house pricing dataset is a well-known machine learning dataset with a narrow range of errors and variations. The main disadvantage of this paper is that predicting accurate house pricing values is difficult. People will be unable to predict house prices if the results are inaccurate.

In the paper[9],Linear Regression and XGBoost Gradient Boosting Model is used. Certain features of this algorithm aid in improving model efficiency and performance. The advantage of this paper is that XGBoost method achieved highest accuracy. House price prediction helps developers forecast prices within a reasonable range, which helps clients decide when and where to buy a home. The disadvantages of this paper are that it is difficult to generalise models with less data because fewer features do not adequately represent data.

The paper[10] used XG Boost, SVM, Decision Tree Regression, and Random Forest method. Decision trees have three advantages. Data can be easily simplified using decision trees. Decision tree provides a better understanding of which aspects of the data are useful for prediction. Nonlinear information has no effect on model performance. It works best when there is a clear margin of separation. The random forest tree has the advantage

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8321
***Article Received:** 13 July 2023 **Revised:** 30 August 2023 **Accepted:** 19 September 2023*

of working better with large and complicated datasets. The paper's disadvantage is that SVR has the lowest accuracy. SVR has the advantage of being efficient in higher magnitudes when the number of magnitudes exceeds the total amount of samples.

The paper [11] used several algorithms like Linear Regression, Random Forest, KNN, etc. The advantages of this paper are that include the collection of a Zillow dataset of Virginia real estate properties. They also used K-means clustering to create subsets of instances with comparable average rent prices within zip codes. The disadvantages of this paper are that some of the clusters are very sparse, with fewer than 100 instances, which could have a significant impact to train the model for prediction.

The paper[12] used Traditional Regression Model. The advantage of this paper is that Light GBM, a distributed algorithm that uses a decision tree's gradient boosting framework, has higher training efficiency and accuracy. Second-hand house prices are particularly high in Pudong New Area, which includes Lujiazui Financial Center. The paper's disadvantage is that the five base models such as Random Forest, AdaBoost, Gradient-boosted decision trees, Light GBM and XGBoost perform poorly.

The paper[13] used Random Forest and linear regression method. Python Scikit learn library was used to implement the random forest. Random Forest is a machine learning technique for supervised classification and regression. The advantage of this paper is that a property's location has a significant impact on its price. Data pre-processing of the selected features were used. The disadvantage of this paper is that it is highly likely that the values of various features will be on a different scale, lowering the model's performance. As a result, scaling was performed to ensure that the features are on a similar scale.

The paper[10] used Random Forest method and Bayesian dynamic factor model. The advantage of this paper is that random forests can fully data-driven to analyse the relationships between housing price factors and a large number of predictor variables. Random forests identify nonlinear relationships between house price factors and predictor variables. The disadvantage of this paper is that increasing the hierarchical complexity of a regression tree results in errors. A random forest is used to address this data sensitive issue. Random regression tree chooses a subset of predictors at random for each splitting decision.

Predicting housing prices has always been a challenge for many machine learning engineers. An accurate house price prediction is critical for stakeholders in the real estate industry, which is on the rise in many countries, such as homeowners, customers, and state agents. The existing system gives best predictions to the sales. It will improve the quality of the results. Python programming language is utilised to create a user-friendly housing price forecasting system and to cut labour costs. After the programme was upgraded, Python has grown to be a very essential subject. Python includes libraries such as Pandas, NumPy, SciPy, Matplotlib, and others. The existing system predicts house costs using various methods implemented in Python.Predicting housing prices has always been a challenge for many machine learning engineers. An accurate house price prediction is critical for stakeholders in the real estate industry, which is on the rise in many countries, such as homeowners, customers, and state agents. The existing system gives best predictions to the sales. It will improve the quality of the results. Python programming language is utilised to create a user-friendly housing price forecasting system and to cut labour costs. After the programme was upgraded, Python has

grown to be a very essential subject. Python includes libraries such as Pandas, NumPy, SciPy, Matplotlib, and others. The existing system predicts house costs using various methods implemented in Python.

III. PROPOSED METHODOLOGY

House price predictions make property investors to be benefited to know the trend of housing prices in certain locations. Proposed method uses attributes such as price, property type, location, square feet, city, bedrooms and other parameters for house price prediction. The classification of attributes makes easier to analyze the effects of different attributes on different models. This paper applied Linear, Polynomial, XGBoost, Random Forest and Decision Tree algorithms for comparing and analysing accuracy of house prices and predict better result This papert imported various

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8321
Article Received: 13 July 2023 *Revised:* 30 August 2023 *Accepted:* 19 September 2023

packages such as numpy, pandas, seaborn, sklearn metrics and other packages for evaluating our ML model into our python environment. Encoding is a technique of converting categorical variables into numerical values so that it could be easily fitted to a machine learning model.An observation with n values is converted to a binary variable with d distinct values, with each observation indicating presence as 1 or absence as 0.The "get_dummies" function is used for encoding.The "get_dummies" function is used for encoding.As a result, the goal of this research is to gain a better understanding of regression methods in machine learning the proposed method is shown in Fig 1.

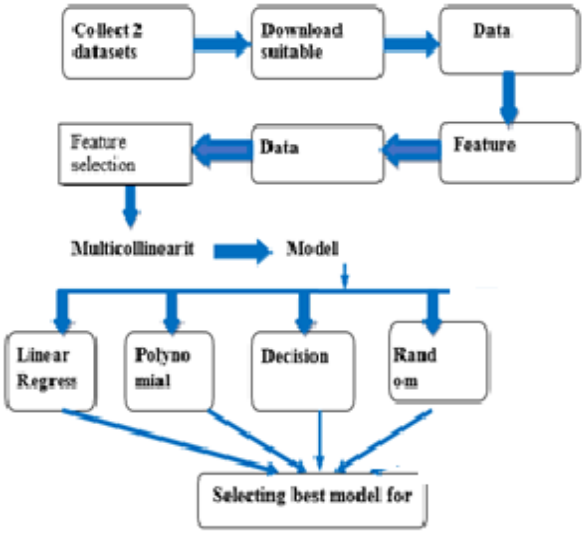

Fig1. Proposed method

1.1 Linear Regression

Linear regression uses a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. The independent variable is also the predictor or explanatory variable that remains constant despite changes in other variables. However, the dependent variable changes in response to changes in the independent variable. The linear regression models are represented by a sloped straight line and shown in Fig 2.

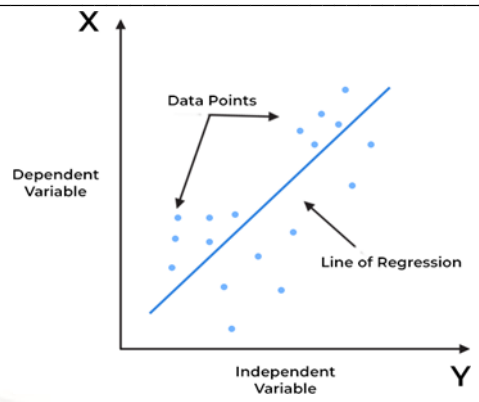

Fig. 2 Linear Regression

1.2 Polynomial Regression

Polynomial regression is an algorithm that models the relationship between the dependent and independent variables Y and X as the polynomial's nth degree. The best approximate relationship between the two dependent and independent variables is provided by polynomial regression.

1.3 Decision Tree

Decision tree algorithm is useful for problem solving in regression and classification. It is applicable to both continuous and categorical output variables. Decision tree has tree-like structure, so it is simple to understand. It is shown in Fig 3.

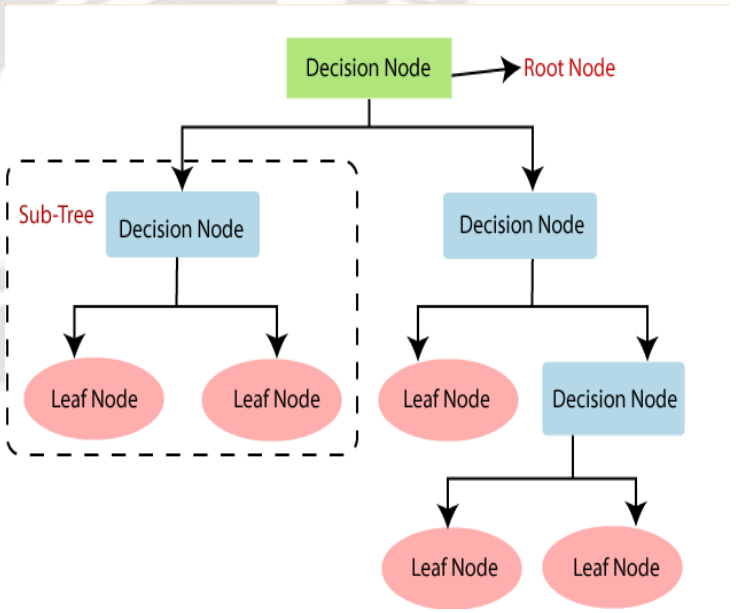

Fig. 3 Decision Tree

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8321
***Article Received:** 13 July 2023 **Revised:** 30 August 2023 **Accepted:** 19 September 2023*

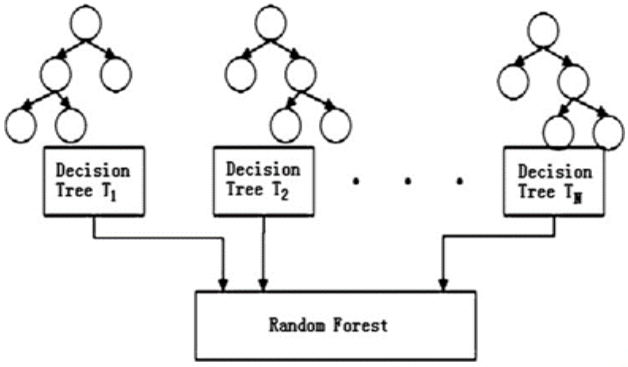

Fig. 4 Random Forest

Random Forest is a supervised machine learning algorithm that is used for decision trees to perform classification, regression and other tasks. Random forest is an efficient method for dealing with missing data. Every random forest tree randomly selects a subset of features at the node's splitting point. Random forest creates multiple Decision tree and merges them to produce more accurate and stable predictions. It is shown in Fig 4.

IV. IMPLEMENTATION AND TESTING

The real estate property dataset was obtained from the Open Data Pakistan website, which provides property data for Pakistan. The original dataset has 168447 instances and 20 characteristics or variables[13][14]. It provides property listings for cities throughout Pakistan, including Islamabad, Rawalpindi, Lahore, Faisalabad, and Karachi. Accuracy is used for evaluation metrics. Fig. 5, Fig. 6, and Fig. 7 show the Seaborn Pair plot, Seaborn Scatterplot and correlation heat map respectively.

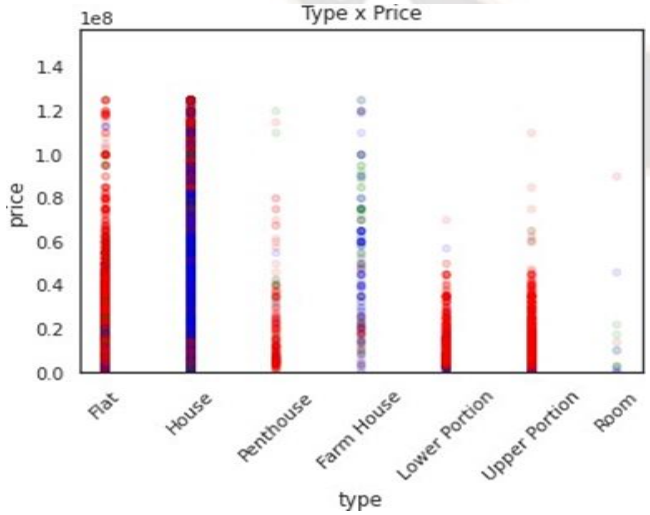

FIG 5 SEABORN PAIRPLOT

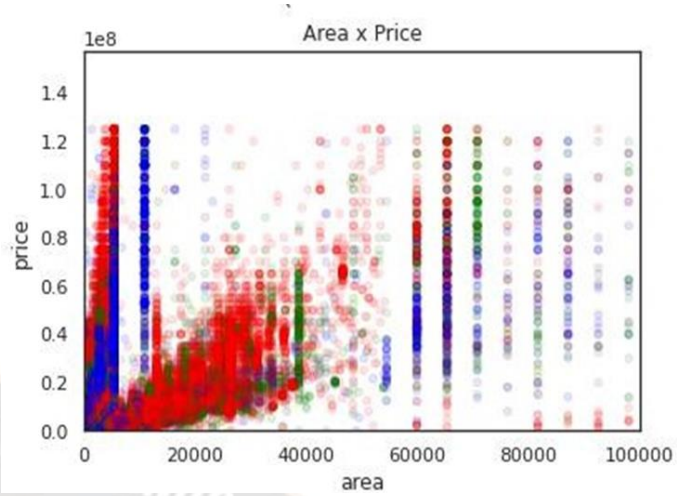

Fig. 6 Seascatter plot

Accuracy of Linear Regression is shown in Fig 8

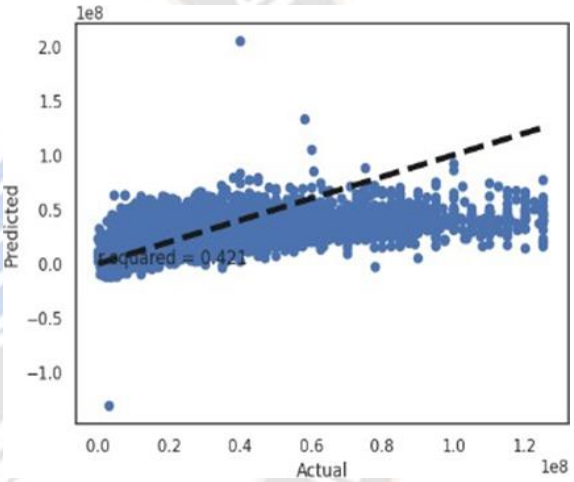

Fig 7.Accuracy of Linear Regression

Accuracy of Polynomial Regression shown in Fig 8. Accuracy of Decision Tress is shown in Fig 9.Accuracy of Random Forest is shown in Fig 10

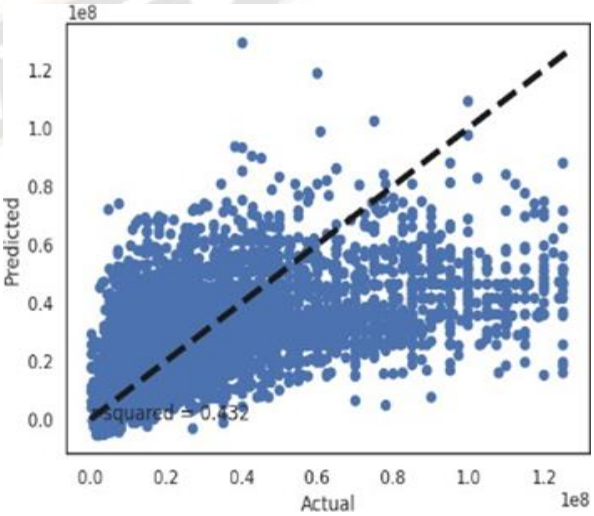

Fig 8.Accuracy of polynomial Regression

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8321
Article Received: 13 July 2023 *Revised:* 30 August 2023 *Accepted:* 19 September 2023

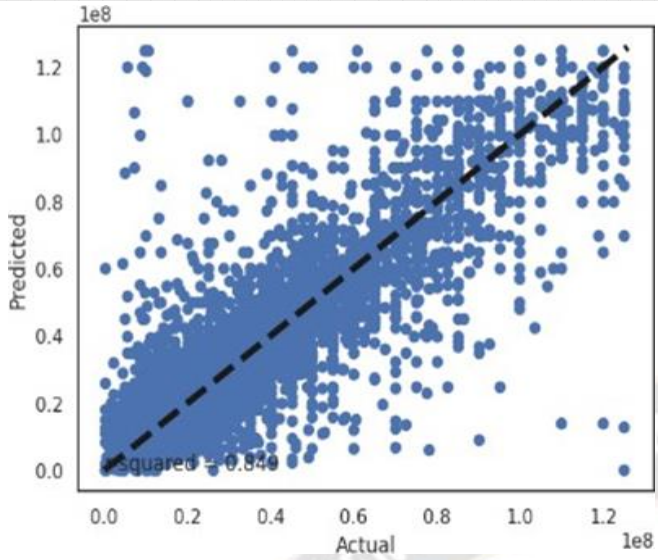

Fig 9.Accuracy of Decision Tree

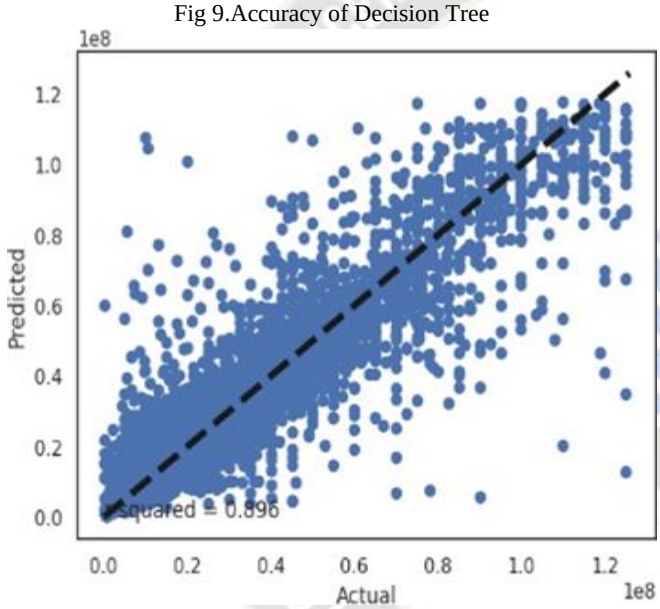

Fig 10.Accuracy of polynomial Regression

Table 1. Comparisons of Accuracy of the methods

Methods	[1] Accuracy%
Logical Regression	[2]41%
Polynomial regression	[3]42%
Decision Tree	[4]84%
Random forest	[5]89%

From the Table1, it is identified that the accuracy of Random Forest is better compared to all other methods. Sample input and result is shown in Fig 11

Fig 11 Sample result display

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8321
***Article Received:** 13 July 2023 **Revised:** 30 August 2023 **Accepted:** 19 September 2023*

V. CONCLUSION

In this paper, house price prediction using four machine algorithms such as linear regression, Polynomial Regression, Random Forest and Decision Tree are done for given Dataset. By comparing these algorithms, Random forest have achieved high accuracy with 0.89. Decision Tree algorithm has high mean squared error when compared to Random Forest. Random forest algorithm has low mean absolute error when compared to Decision Tree. Random forest provides higher accuracy compared to other algorithm Therefore we would like to use more dataset and perform large analysis using deep learning algorithm.

Conflicts of Interest (Mandatory)

No conflict of Interest

REFERENCES

[1] Q.Truong, M.Nguyen, H.Dang and Mei B," Housing price prediction via improved machine learning techniques," Procedia Computer Science, vol.174, pp. 433-442

[2] Tang, Yajuan,Shuang Oju and Pengcheng G," Predicting housing price based on ensemble learning algorithm" IEEE 2018.

[3] Wang G.Y., "The effect of environment on housing prices: Evidence from the Google Street View," Journal of Forecasting, vol.42, pp, 288-311, 2023.

[4] Sarangi, D. P. K. . (2022). Malicious Attacks Detection Using Trust Node Centric Weight Management Algorithm in Vehicular Platoon. Research Journal of Computer Systems and Engineering, 3(1), 56–61. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/42

[5] Xu, X., & Zhang, Y," Retail property price index forecasting through neural networks", Journal of Real Estate Portfolio Management, vol.29, pp.1-28, 2023.

[6] Angulakshmi M, Deepa M, "A Review on Deep Learning Architecture and Methods for MRI Brain Tumour Segmentation, Curr Med Imaging,vol.17,pp. 695-706,2021.

[7] Peng, Zhen, Qiang Huang, and Yincheng Han, "Model research on forecast of second-hand house price in Chengdu based on XGboost algorithm" IEEE 2019.

[8] Durganjali, P and M. Vani Pujitha,"House resale price prediction using classification algorithms" IEEE 2019.

[9] Jain, Mansi, "Prediction of house pricing using machine learning with Python" IEEE 2020.

[10] Ahtesham, Maida, Narmeen Zakaria Bawany, and Kiran Fatima, "House price prediction using machine learning algorithm- the case of Karachi city, Pakistan" IEEE 2020.

[11] Rana and Vivek Singh, "House Price Prediction Using Optimal Regression Techniques" ,IEEE 2020.

[12] HHeidari, Maryam, Samira Zad, and Setareh Rafatirad, "Ensemble of supervised and unsupervised learning models to predict a profitable business decision", IEEE 2021.

[13] XXu, Lulin and Zhongwu Li, "A new appraisal model of second-hand housing prices in China's first- tier cities based on machine learning algorithms", Springer 2021.

[14] AAdetunji, Abigail Bola,"House Price Prediction using Random Forest Machine Learning Technique", Science Direc,t 2022.

[15] Gupta, Rangan, "Machine Learning Predictions of Housing Market Synchronization across US States: The Role of Uncertainty", Springer 2022.