House Price Prediction using Machine Learning Algorithm

Shailendra Sharma
Department of ECE
Shree Digambar Institute of Technology,
Dausa, Rajasthan, India
ershail88@gmail.com

Deepti Arora
Department of Mathematics
Arya Institute of Engineering & Technology,
Jaipur, Rajasthan, India
aroradeepti591@gmail.com

Gori Shankar
Department of ECE
Jaipur Engineering College,
Jaipur, Rajasthan, India
gssharma1951@gmail.com

Priyanka Sharma
Department of Mathematics
Jaipur Engineering College,
Jaipur, Rajasthan, India
priyanka01185@gmail.com

Vihaan Motwani
Student,
Shiv Nadar School,
Gurgaon, India
vihaanmotwani16@gmail.com

2023 7th International Conference on Computing Methodologies and Communication (ICCMC) | 978-1-6654-6408-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCMC56507.2023.10084197

Abstract: **Houses are one of the fundamental needs in human society. Good and peaceful surroundings where people feel comfortable are called houses. Hence, to have a good and happy living, people need to choose a good house model. This article focuses on accurately predicting house prices by using machine learning algorithms. This model helps people in selecting the house that is suitable for their living. The main parameter people look for will be the surrounding area, house type, price, location, and some other amenities. All these factors are considered to find out the required house for living. As people are very concerned about their budgets for buying a house, the prediction of house prices should be very precise. It also helps people in choosing houses based on their budgets, which do not affect their financial state in the future. The main outcome of this model is to predict the price of a house accurately as per the user requirements. This study has attempted to implement various machine learning algorithms like Linear Regression (LR), Gradient Boosting Regressor (GBR), Histogram Gradient Boosting Regressor, and Random Forest (RF) Regressor algorithms. Finally, the algorithm that generates high accuracy is considered for predicting the house price.**

Keywords: Machine Learning (ML), House Price Prediction, Regression Techniques, ML Algorithm.

I INTRODUCTION

Predicting house prices are one of the important factors for the non-house holders, as they need to plan their budget [1]. The system has been proposed with various algorithms and the algorithm that gives the most correct accurate level is considered [2]. Houses are the central need for people and the prices of the houses vary from one place to another. House price prediction is a method that is used by any kind of people as no one can estimate the houses based on the location or the facilities in the area [3]. The house prices detection is a very major and difficult task. The proposed model would make accurate predictions of the houses.

Machine learning algorithm uses historical data to predict future output values [4]. Machine Learning is an approach to Artificial Intelligence (AI) that enables software applications to become more accurate at predicting outcomes [5-6]. Machine Learning is a term coined by Arthur Samuel, a computer scientist at IBM and a pioneer in AI and gaming. Machine Learning is a subset of AI. A subset of Machine Learning is mostly related to computational statistics, which gives predictions using computers [7-8]. Data mining deals with exploratory learning through unsupervised learning [9-10]. Python is a high-level programming language. It is mostly used in designing websites, tasks, supports automated tasks and helps in conducting data analysis [11]. It was developed by Guido van Rossum. It is a very easy and understandable programming language. Python supports OOPS concept. They are portable and extendable. Python library contains various modules. Some of the modules are matplotlib, pandas, NumPy and etc. The most commonly used python library for Machine Learning is scikit-learn for pre-processing data and they are open-source [12]. They are highly interactive and can be integrated with other programming languages. Python code is interpreted at run time; they no need to be compiled before executing.

Regression techniques are a subpart of supervised Machine Learning. They are developed by using the connection between an explained variable and a set of exposure variables. Regression is one of the prediction algorithms that predict the outcome based on a value. The main aim of the regression algorithm is to build an equation that defines y as a function of x variables. Linear regressions one of the basic regression techniques used for the prediction and analysis of a data set in Machine Learning. The simple formula for linear regression is y = m * x + c. Here various regression algorithms are used. Some are Gradient Boosting Regressor, Hist Gradient Boosting Regressor, and Random Forest Regressor are used in finding the predictions of accuracy for the houses.

II LITERATURE SURVERY

The house costs are calculated based on some essential factors. The main factors are the state of the house and location. The state of the house has various sections like the scale of the house, the number of rooms in the house, availability of restrooms, kitchen, parking area, house structure, house age, etc. The location of the house is one of the main focuses of people who plan on buying a house. House prices vary from the availability of resources and from one area to another. Many prefer a house if it is center of the area as all needs can be satisfied by the surroundings. Machine Learning is a part of AI. They help in predicting the accuracy of the house prices. Many have proposed this model with classification algorithms. In this model, regression techniques have been used. The various regression algorithms are used and the best algorithm is chosen based on the best accuracy. This helps people to plan their budget for their future.

A comparative study on the prediction of house prices using regression techniques like Elastic Net, Ada Boosting, Multiple linear, Ridge, and LASSO algorithms has been presented by Madhuri et. al. [13]. Here the common parameters of the house have been used. That is, the price and square feet are the parameters used in this. Tang et. al. [14] have made a study on predicting house prices based on an ensemble learning algorithm. Ensemble learning is considered the best tool for predicting algorithms. The random forest algorithm and ensemble learning algorithm were used. The ensemble learning algorithm provides better performance than the random forest algorithm. The ensemble learning provides the best accuracy compared to the random forest algorithm.

Durganjali et. al. [15] have explored the resale of house prices prediction by variousclassification algorithms. The algorithms consider t h e accuracy of the predictions of the resale of houses and discover the algorithm that gives the most accurate price for the house resale. The various classification algorithms used in this paper is Logistic regression, Decision trees, NaïveBayes and AdaBoost algorithms have been used in this study to determine the house resale prices. Prediction of house pricing using Machine Learning with Python has been proposed by Jain et. al. [16]. The paper gives theoretical knowledge of producing predictions using various Machine Learning Algorithms. All the required process for using data is explained here. The data cleaning, splitting of data, training, and testing o f the model is explained in this paper. Phan et. al. [17] has explored housing pricing predictions using Machine Learning algorithms. Some of the data processing and data reduction or transformation processes all have been done here. In this paper, an unsupervised approach has been implemented. That includes polynomial regression, regression trees, neural networks, and support vector machines that have been used in this paper. From the above models, the model that gives the minimum runtime is considered.

III PROPOSED METHODOLOGY

In this work we collect data from people of different living status. Using the data set we perform training the using different machine learning algorithms and some set of data will be used for testing. Fig.1 shows the overall process involved in this work.

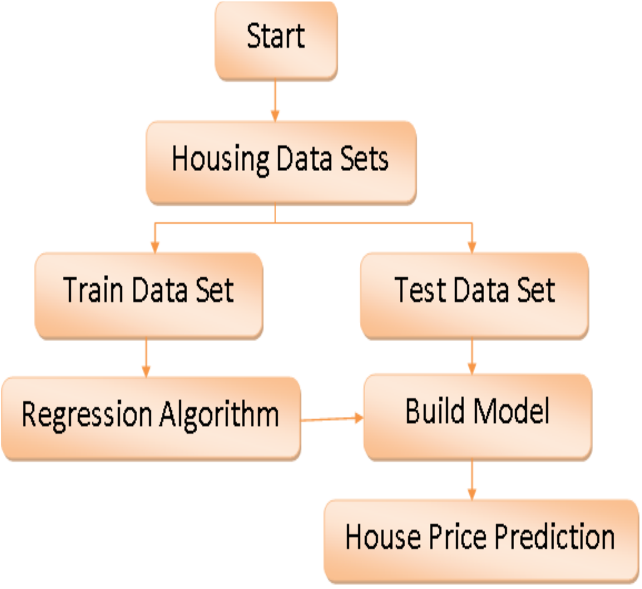


Fig 1. Proposed architecture of house price prediction model

A. Dataset Creation

The dataset in general means the collection of data, used for analytic and prediction purposes. Datasets can hold any type of record that is stored in the system. For Machine Learning projects, a large amount of data is required, because without data we cannot train AI models. An ideal dataset has either well labeled fields and members or a data dictionary that can be used to relabel the data. A good dataset has completeness, they are reliable, and have great accuracy, correct relevance, and timeliness. Dataset can also be referred to as a container for storing data.

B. Data Preprocessing

Data preprocessing is a process of converting the source or atomic data into structured or understandable data [18-19]. It is a very predominant step in the Machine Learning process because only when the data is understandable, the model can be trained [20]. When the dataset is preprocessed, it makes it easier to interpret the data and is simple to use. Preprocessing techniques are applied so that the data gives high-quality mining results [21]. The data is cleaned up in the Preprocessing state as shown in Fig. 2. It is a simple process of discovering and eliminating errors to increase the efficiency of data. Sometimes there might be any missing values in the data set. It can be solved by getting rid of the missing data points or the whole attributes or by assigning a value to the attribute. This is a process of converting the given dataset into understandable data using data reduction and cleaning methodologies.

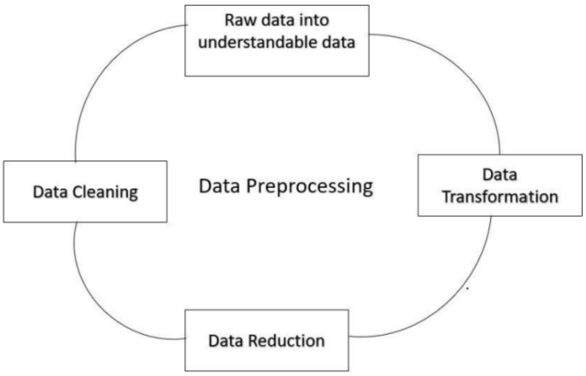


Fig 2. Stages in Data Pre-processing

C. *Training the model*

Here the data is broken into 2 parts. That is training and testing. 80 percent of data is used for training the model and the rest 20 percent is used for testing purposes. Training the model is mainly training the dataset with Machine Learning algorithms. It consists of sample output with corresponding sets of input data for training the model as represented in Fig. 3.

Input
ML Algorithm
Output

Fig 3. Training the Model

D. *Testing the model*

Once the model is trained, they are tested with the dataset. The model provides the prediction accuracy or the output for the processed data-set. It is a method to measure the results of the model that gives the accurate score of the dataset. That is, validation/test is done for the model build. Test data sets are used to evaluate machine learning programs that have been trained on an initial training data set.

In this proposed model, various machine learning algorithms are used. The model is trained on Linear Regression, Gradient Boosting Regressor, Hist Gradient Boosting Regressor, and Random Forest Regressor algorithms. Among the above-given algorithms, the Gradient Boosting Regressor and Hist Gradient Boosting Regressor gives the highest accuracy in the prediction of house prices. The following libraries are imported for working with Machine Learning. They are pandas, NumPy, seaborn, matplotlib.pyplot, and sklearn libraries.

Linear Regression

Predictive analysis with linear regression is the most basic and most commonly used method in machine learning. It is very easy to implement. It is evident from its name that Linear Regression represents a linear relationship between a manipulated (input) variable and an experimental (output) variable as can see in Fig. 4. It is one of supervised learning algorithm.

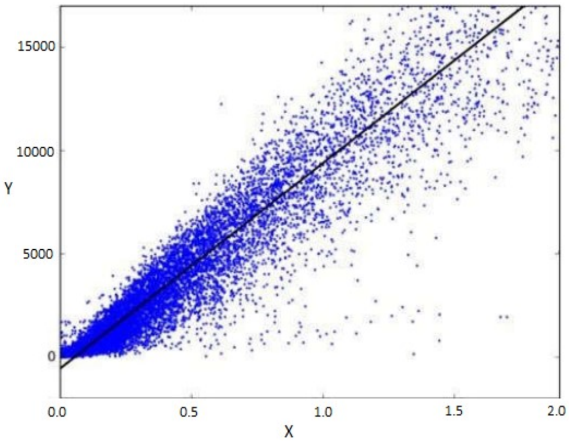


Fig 4. Linear Regression

Gradient Boosting Regressor

A machine learning algorithm used for regression technique. It helps in providing an ensemble model by combining the weak predictive models. They support both classification and regression problems. In gradient boosting algorithm, decision trees of fixed size are used, as represented in Fig. 5.

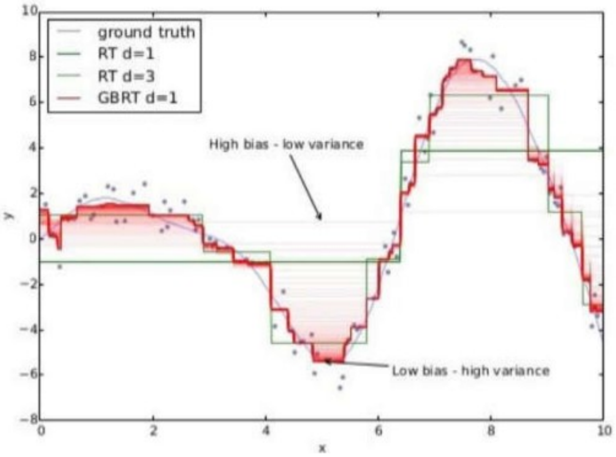


Fig 5. Data Preprocessing

Hist Gradient Boosting Regressor

Hist Gradient Boosting Regressor is one of the Machine Learning regression algorithms. It provides the best accuracy results. This algorithm helps in the faster training of decision trees. The scikit- learn machine learning provides a strategy to work with the Hist Gradient Boosting Regressor algorithm.

Random Forest Regressor

It is one of the supervised learning algorithms that use ensemble learning methods for regression. They can be used in both regression and classification techniques. Random Forest Regressor algorithm is so flexible and easy to use. In the case of regression, the Random Forest Regressor algorithm can handle data sets containing continuous variables. Ensemble means a combination of multiple models. That is, a collection of modules is used to make predictions rather than an individual model.

IV OBTAINED RESULTS & DISCUSSION

For implementation of the proposed model first import the required libraries (numpy, pandas, matplotlib.pyplot, seaborn, mpl_toolkits). After imported these libraries, now next, the dataset is uploaded in Google drive and mounted as csv file. The imported and stored sample data set in .csv file is used in this process of house price prediction.

The next task is to process the given dataset. That is, the dataset is processed and checked for any errors or missing attributes. A matplotlib library function is used in the visualization of the graph of the dataset. Then, the data is split into a proportion of 80 percent for training and 20 percent for testing. Here the data is represented by various graphs is shows in Fig. 6. The visualization of data for various factors of the dataset is represented. The graph varies for each factor of the given dataset.

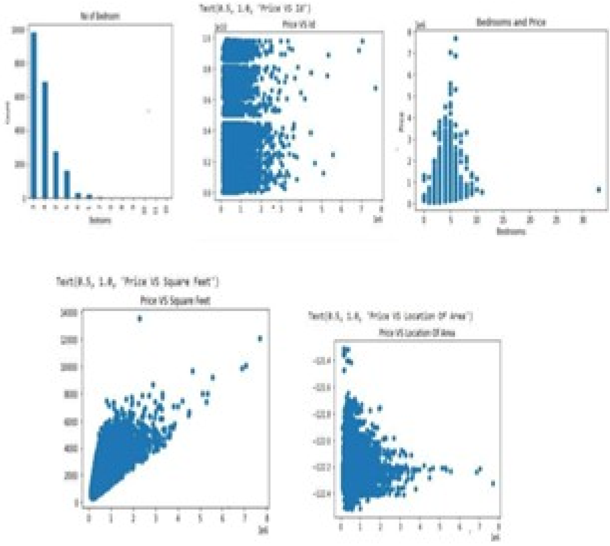
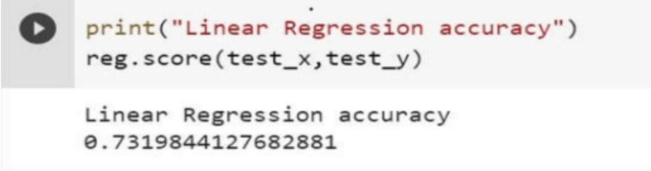


Fig 6. Visualization of data graphs

A. *Linear Regression*

Then, the model is trained by the given dataset. First, the linear regression algorithm is implemented on the trained dataset as per Fig. 7. This algorithm works on a line equation, y=mx+c. The library required for the linear regression is imported from sklearn.linear_model.



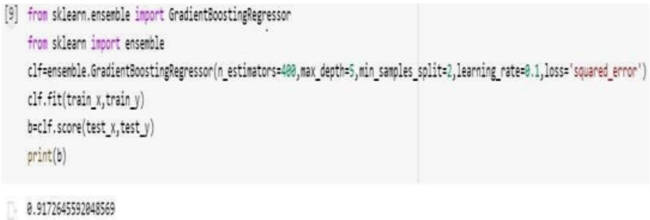
```
print("Linear Regression accuracy")
reg.score(test_x,test_y)

Linear Regression accuracy
0.7319844127682881
```

Fig 7. Obtained accuracy from linear regression algorithm

B. *Gradient Boosting Regression Algorithm*

The required libraries are imported. For the Gradient Boosting Regressor algorithm, the ensemble is important. The ensemble is imported from sklearn, and then the Gradient Boosting Regressor is defined. The trained dataset fits into the algorithm and the score is obtained from the tested data set is shown in Fig. 8.



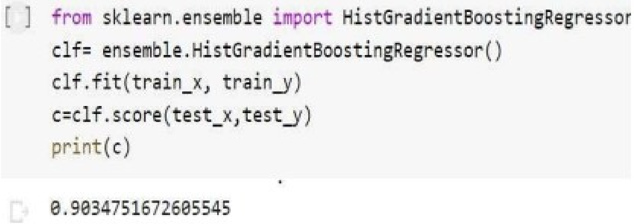
```
[9] from sklearn.ensemble import GradientBoostingRegressor
    from sklearn import ensemble
    clf=ensemble.GradientBoostingRegressor(n_estimators=400,max_depth=5,min_samples_split=2,learning_rate=0.1,loss='squared_error')
    clf.fit(train_x,train_y)
    b=clf.score(test_x,test_y)
    print(b)

0.9172645592048569
```

Fig 8. Obtained accuracy from gradient boosting regression algorithm

C. *Hist Gradient Boosting Regressor Algorithm*

In the hist gradient boosting regressor algorithm, the required libraries are passed. Then, the hist gradient boosting regressor function is defined. Then, the trained data set is split and used for testing the algorithm to produce accuracy as presented in Fig. 9.



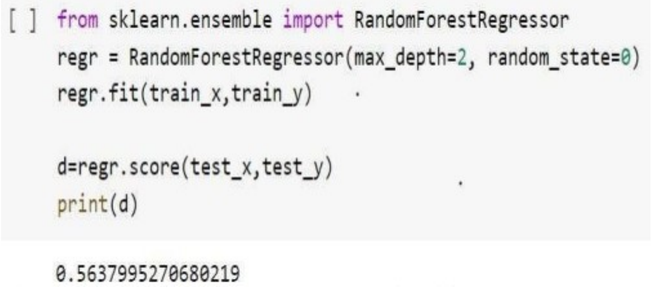
```
[ ] from sklearn.ensemble import HistGradientBoostingRegressor
    clf= ensemble.HistGradientBoostingRegressor()
    clf.fit(train_x, train_y)
    c=clf.score(test_x,test_y)
    print(c)

0.9034751672605545
```

Fig 9. Obtained accuracy from hist gradient boosting regressor algorithm

D. *Random Forest Regressor Algorithm*

In the random forest regressor algorithm, the required algorithms are passed. Then, the algorithm is defined. Now, the trained data set is fit into the algorithm as in Fig 10. Then, the accuracy is obtained using the tested data set.



```
[ ] from sklearn.ensemble import RandomForestRegressor
    regr = RandomForestRegressor(max_depth=2, random_state=0)
    regr.fit(train_x,train_y)

    d=regr.score(test_x,test_y)
    print(d)

0.5637995270680219
```

Fig 10. Obtained accuracy from random forest regressor algorithm

The house price predictions are determined by the regression algorithms and the results of the accuracy of the different algorithms are obtained and plotted in Table 1.

Table 1. Comparison of prediction accuracy of the used different algorithms

S. No.	Machine Learning Algorithm	Accuracy in house price prediction
1	Linear Regression	73.19%
2	Gradient Boosting Regression	91.72%
3	Hist Gradient Boosting Regressor	90.34%
4	Random Forest Regressor	56.37%

We observe that gradient boosting regressor algorithm and hist gradient boosting regressor algorithm provides the accurate and best results among the algorithms used.

V CONCLUSION

The house price prediction model using machine learning is proposed in this system using various regression and algorithmic techniques. The house prices in general deal with various factors like area, price, number of bedrooms, number of toilets, parking space and so on. This system helps people to choose houses based on their budget and market strategies, which does not affect their financial level. In the future, comparisons can be made with different classification algorithms and can measure the performance of the system using various machine learning algorithms.

References

[1] Tan F, Cheng C, Wei Z., "Time-Aware Latent Hierarchical Model for Predicting House Prices", In2017 IEEE International Conference on Data Mining (ICDM), pp. 1111- 1116, 2017.

[2] D. Banerjee and S. Dutta, "Predicting the housing price direction using machine learning techniques", 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), pp. 2998-3000, 2017.

[3] G. Hu, J. Wang and W. Feng, "Multivariate regression modeling for home value estimates with evaluation using maximum information coefficient", Software Engineering Artificial Intelligence Networking and Parallel/Distributed Computing., vol. 1, no. 2, pp. 69-81, 2013.

[4] Priya Gour, Sudhanshu Vashistha and Pradeep Jha, "Twitter Sentiment Analysis Using Naive Bayes based Machine learning Technique", 2nd International Conference on Sentiment Analysis and Deep Learning (ICSADL 2022), Springer - Advances in Intelligent Systems and Computing Series 1432, 2023.

[5] J. J. Wang et al., "Predicting House Price With a Memristor-Based Artificial Neural Network," in IEEE Access, vol. 6, pp. 16523-16528, 2018.

[6] R. E. Febrita, A. N. Alfiyatin, H. Taufiq and W. F. Mahmudy, "Data-driven fuzzy rule extraction for housing price prediction in Malang, East Java", 2017 IEEE International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 351-358, 2017.

[7] Himanshu Aora, Kiran Ahuja, Himanshu Sharma, Kartik Goyal and Gyanendra Kumar, "Artificial Intelligence and Machine Learning in Game Development", Turkish Online Journal of Qualitative Inquiry (TOJQI), vol. 12, no. 8, pp. 1153-1158, 2021.

[8] Kiran Ahuja, Harsh Sekhawat, Shilpi Mishra and Pradeep Jha, "Machine Learning in Artificial Intelligence: Towards a Common Understanding", Turkish Online Journal of Qualitative Inquiry (TOJQI), vol. 12, no. 8, pp. 1143-1152, July 2021.

[9] Rahul Misra and Dr. Ramkrishan Sahay, "A Review on Student Performance Predication Using Data Mining Approach", International Journal of Recent Research and Review, Vol. X, Issue 4, pp. 45-47, December 2017.

[10] R. Misra and Dr. R. Sahay, "Evaluation of Student Performance Prediction Models with Two Class Using Data Mining Approach", International Journal of Recent Research and Review, Vol. XI, Issue 1, pp. 71-79, March 2018.

[11] P. Jha, R. Baranwal, Monika and N. K. Tiwari, "Protection of User's Data in IOT", *2022 IEEE Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pp. 1292-1297, 2022.

[12] P. Jha, T. Biswas, U. Sagar and K. Ahuja, "Prediction with ML paradigm in Healthcare System", 2021 IEEE Second International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1334-1342, 2021.

[13] C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," 2019 International Conference on Smart Structures and Systems (ICSSS), pp. 1-5, 2019.

[14] Y. Tang, S. Qiu and P. Gui, "Predicting Housing Price Based on Ensemble Learning Algorithm", 2018 IEEE International Conference on Artificial Intelligence and Data Processing (IDAP), pp. 1-5, 2018.

[15] P. Durganjali and M. V. Pujitha, "House Resale Price Prediction Using Classification Algorithms", 2019 IEEE International Conference on Smart Structures and Systems (ICSSS), pp. 1-4, 2019.

[16] M. Jain, H. Rajput, N. Garg and P. Chawla, "Prediction of House Pricing using Machine Learning with Python", 2020 IEEE International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 570-574, 2020.

[17] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia", 2018 IEEE International Conference on Machine Learning and Data Engineering (iCMLDE), pp. 35-42, 2018.

[18] H. Arora, G. K. Soni, R. K. Kushwaha and P. Prasoon, "Digital Image Security Based on the Hybrid Model of Image Hiding and Encryption", 2021 IEEE 6th International Conference on Communication and Electronics Systems (ICCES), pp. 1153-1157, 2021.

[19] Himanshu Arora, Shilpi Mishra and Manish Dubey, "Development of the Framework for the Solution of the Security Problems in Data Transmission Involving Advanced Asymmetric Algorithm", International Journal of Emerging Technology and Advanced Engineering, vol. 8, no. 4, pp. 18-20, April 2018.

[20] D. Banerjee and S. Dutta, "Predicting the housing price direction using machine learning techniques," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), pp. 2998-3000, 2017.

[21] Y. Chen, R. Xue and Y. Zhang, "House price prediction based on machine learning and deep learning methods," 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS), pp. 699-702, 2021.