Journal of Science Engineering Technology and Management Science
Volume 02, Issue 06, June 2025 *ISSN: 3049-0952*
www.jsetms.com *DOI:10.63590/jsetms.2025.v02.i06.pp87-94*

PREDICTING HOME PRICES: A BEGINNER'S JOURNEY WITH REGRESSION ANALYSIS USING THE BOSTON HOUSING DATASET

S. Puneeth[1*], Md. Ammaar Quadri[2], M. Sahithi[2], Mohd. Arbas[2], P.S. Jyothi [2]
[1]Assistant Professor, [2]UG Student ,
[1,2]Department of Computer Science and Engineering (Information Technology),
[1,2]Sree Dattha Institute of Engineering and Science, Sheriguda, Hyderabad, Telangana

To Cite this Article
S. Puneeth, Md. Ammaar Quadri, M. Sahithi, Mohd. Arbas, P.S. Jyothi, "Predicting Home Prices: A Beginner's Journey With Regression Analysis Using The Boston Housing Dataset", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 06, June 2025,pp:87-94, DOI: http://doi.org/10.63590/jsetms.2025.v02.i06.pp87-94
Submitted: 20-04-2025 *Accepted: 27-05-2025* *Published: 06-06-2025*

ABSTRACT
Accurately predicting home prices is vital for buyers, sellers, and real estate professionals, enabling informed decisions, property valuation, and market analysis. Machine learning has become popular for this task due to its ability to analyze complex data patterns and generate predictive models. Traditional real estate pricing relies on agents' expertise, appraisers, market trends, and comparable sales data. However, this approach can be subjective, prone to errors, and may not fully account for all factors influencing prices or scale well for large datasets. Machine learning offers a data-driven solution to improve prediction accuracy and provide deeper insights into the housing market. The challenge is to use regression analysis to predict home prices accurately. The goal is to develop machine learning models that outperform traditional methods, offering reliable predictions and actionable insights for real estate stakeholders. The proposed system uses the Boston Housing Dataset, a widely used dataset in machine learning, to train and evaluate regression models. Features like crime rate, zoning, and proximity to employment centers are used to predict home prices. Regression algorithms such as Random Forest Regression and XG Boost Regression will be employed to build these models. By leveraging these advanced techniques, the system aims to enhance prediction accuracy, improve decision-making, and provide valuable insights into market trends, addressing the limitations of traditional methods and meeting the growing demand for data-driven approaches in the real estate industry.

Keywords: Regression Analysis, Machine Learning, Boston Housing, Home prices, Model Building, Random Forest, XG Boost.

1.INTRODUCTION

Accurately predicting the value of a plot or house is an important task for many house owners, house buyers, plot owners, plot buyers or stake holders. Real estate agencies and people buy and sell houses all the time, people buy houses to live in or as an investment whereas real estate agencies buy it to run a business. But the problem arises in evaluation of the cost of the property. Over-validation / Under-validation have always been the issues faced in house markets due to lack of proper detection measures. It is also very difficult task. We know that features like size, area, location etc affect the price of the property but there are many other features also which affect the property such as inflation

rates in market, age of the property etc. In order to overcome these problems a throw analysis is done using Machine Learning (ML) which is a branch of Artificial Intelligence (AI). With the advancement of science and technology our daily life has become much easier. In today's world we use information and communication technology extensively. Every day a new technology emerges in our current digital age which improves the living standard of people. Sometimes these new technologies have negative effect but most of the times these technologies have positive effect [1]. AI or more widely known as AI is one of such technology which has improved the living standard of people [2] worldwide. AI is widely used now days in various fields like healthcare [3], real estate [4], stock market prediction [5], weather prediction [6], automobile [7] and also in many other fields [8][9]. AI has many subfields like Natural Language Processing, Machine Vision, Robotics, Expert System etc [10], but in this study ML is used. ML is a branch of AI which deals with certain tasks using past data or recorded data and various algorithms. These tasks of ML involve classification, association, clustering and regression. ML can be used to make predictive models to make predictions for future or can be used to make descriptive models to make acquire some kind of knowledge from the given data. The main difference between ML programming and conventional programming is that, in conventional programming, programs are created manually by providing input data and based on the programming logic computer generates the output.

2.LITERATURE SURVEY

A lot of past works have been done for predicting house prices. Different levels of accuracies and results have been achieved using different methodologies, techniques and datasets. A study of independent real estate market forecasting on house price using data mining techniques was done by Bahia [11]. Here the main idea was to construct the neural network model using two types of neural network. The first one is Feed Forward Neural Network (FFNN) and the second one is Cascade Forward Neural Network 2022 2nd International Conference on Intelligent Technologies (CONIT) Karnataka, India. June 24-26, 2022 (CFNN). It was observed that CFNN gives a better result compared to FFNN using MSE performance metric. Mu et al. [12] did an analysis of dataset containing Boston suburb house values using several ML methods which are Support Vector Machine (SVM), Least Square Support Vector Machine (LSSVM) and Partial Least Square (PLS) methods. SVM and LSSVM gives superior performance compared to PLS. Beracha et al. [13] proved that high amenity areas experience greater price volatility by investigating the correlation between house prices volatility, returns and local amenities. Law [14] finds that there is a strong link between house price and street based local area compare to the house price and region based local area. Binbin et al. [15] to study London house price build a Geographically Weighted Regression (GWR) model considering Euclidean distance, travel time metrics and Road network distance. Marco et al. [16] to reduce the prediction errors, a mixed Geographically weighted regression(GWR) model is used that emphasize the importance and complex of the spatial Heterogeneity in Australia. Using State level data in USA, Sean et al. [17] have examined the correlation among common shocks, real per capita disposable income, house prices, net borrowing cost and macroeconomic, spatial factors and local disturbances and state level population growth. Joep et al. [18] using the administrative data from the Netherlands have found that wealthy buyers and high income leads to higher purchase price and wealthy sealer and higher income leads to lower selling price.

3. PROPOSED METHODOLOGY

This research aims to predict home prices using machine learning techniques applied to the Boston Housing Dataset. Here's an overview:

— **Data Loading and Preprocessing**: The code begins by importing necessary libraries such as Pandas, NumPy, Matplotlib, Seaborn, scikit-learn, XGBoost. It loads the dataset from a CSV file using Pandas' read_csv() function. Exploratory Data Analysis (EDA) techniques are applied to understand the dataset, including checking data types, summary statistics, and

identifying missing values. Visualization techniques like heatmaps, pairplots, and histograms are used to gain insights into data distributions and relationships between variables.

— **Data Splitting**: The dataset is split into features (independent variables) and the target variable (home prices). The train-test split is performed using scikit-learn's train_test_split() function, allocating a certain percentage (e.g., 80%) of the data for training and the remaining for testing.

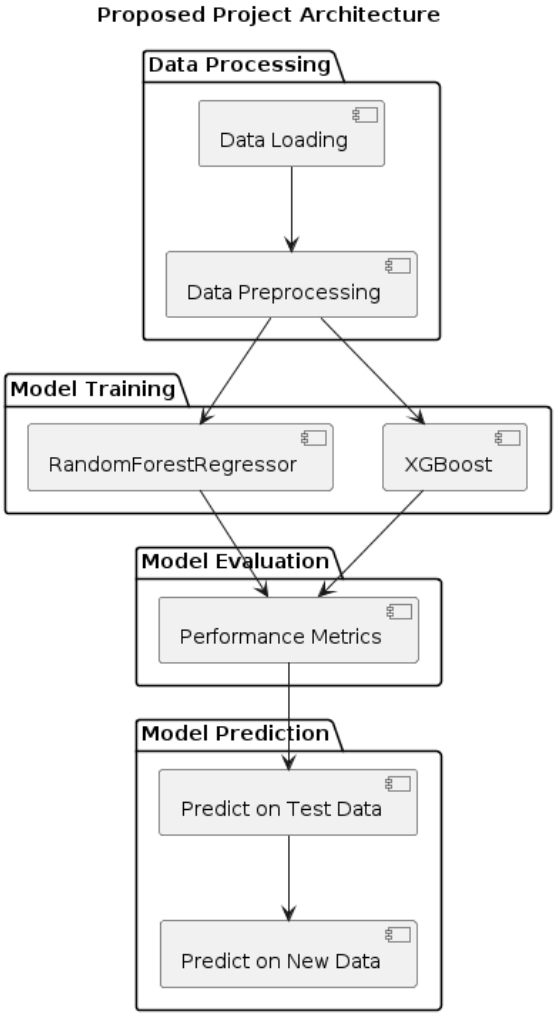

Fig. 1: Block Diagram of Proposed System.

— **Model Training**: Two regression models, RandomForestRegressor and XGBoost, are trained using the training data. For RandomForestRegressor, the code checks if a pre-trained model exists; if not, it initializes and fits a new model to the training data. Similarly, for XGBoost, the code either loads a pre-trained model or trains a new one. The models are trained to learn the relationships between the input features and home prices in the training data.

— **Model Evaluation**: A function called performance_metrics() is defined to evaluate the performance of each regression model. This function computes metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) score to assess the model's accuracy. Additionally, it visualizes the predicted vs. true values using a scatter plot to visually inspect the model's performance.

— **Model Prediction**: Once trained, the models are used to make predictions on the test data. The predictions are compared against the actual home prices in the test data to assess the models' generalization performance.

— **Prediction on New Data**: Lastly, the code reads a separate testing dataset (new data) from a CSV file. The trained XGBoost model is used to predict home prices on this new dataset. The predictions are appended to the testing dataset for further analysis or deployment.

3.1 Data Preprocessing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task. A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

Handling Missing data: The next step of data preprocessing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset. There are mainly two ways to handle missing data, which are:

- By deleting the particular row: The first way is used to commonly deal with null values. In this way, we just delete the specific row or column which consists of null values. But this way is not so efficient and removing data may lead to loss of information which will not give the accurate output.
- By calculating the mean: In this way, we will calculate the mean of that column or row which contains any missing value and will put it on the place of missing value. This strategy is useful for the features which have numeric data such as age, salary, year, etc.

Encoding Categorical data: Categorical data is data which has some categories such as, in our dataset; there are two categorical variables, Country, and Purchased. Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So, it is necessary to encode these categorical variables into numbers.

3.2 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

3.3 XG Boost Model

XGBoost is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "XGBoost is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the XGBoost takes the prediction from each tree and based on the majority votes of predictions, and it predicts the

final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

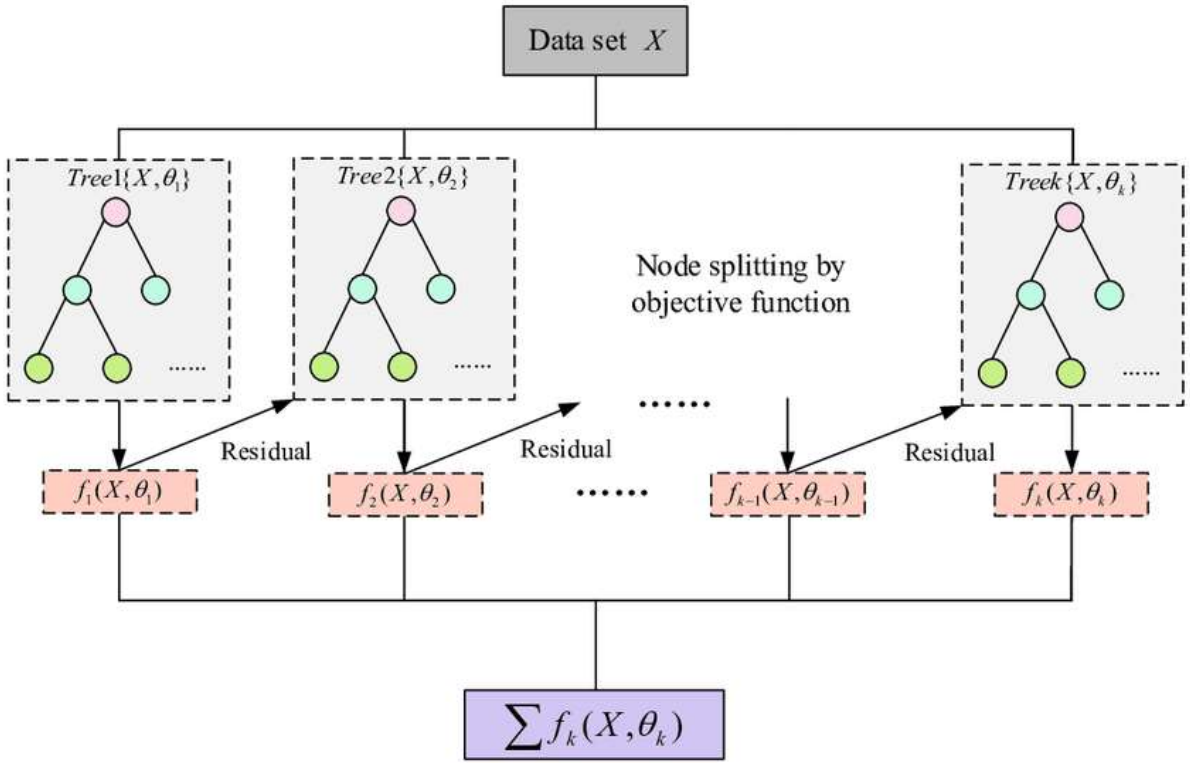

Fig.2: XG Boost algorithm.

XGBoost, which stands for "Extreme Gradient Boosting," is a popular and powerful machine learning algorithm used for both classification and regression tasks. It is known for its high predictive accuracy and efficiency, and it has won numerous data science competitions and is widely used in industry and academia.

4. RESULTS AND DISCUSSION

The Boston Housing Dataset provides a comprehensive set of features capturing various housing and neighborhood characteristics, essential for understanding housing markets, predicting prices, and making informed decisions. CRIM measures the per capita crime rate by town, indicating the frequency of crimes relative to the population, with higher values suggesting areas with more crime incidents. ZN represents the proportion of residential land zoned for lots over 25,000 square feet, reflecting zoning regulations that influence property development, where higher values indicate more spacious residential zones. INDUS captures the proportion of non-retail business acres per town, highlighting the industrial composition, with higher values denoting more industrialized areas. CHAS is a binary variable indicating whether a tract bounds the Charles River (1 if yes, 0 if no), a factor that can affect property values due to proximity to the river.

NOX measures nitric oxides concentration (parts per 10 million) as an air pollution indicator, where higher values may reduce property desirability. RM denotes the average number of rooms per dwelling, reflecting housing space and comfort, with higher values indicating larger homes. AGE represents the proportion of owner-occupied units built before 1940, suggesting older housing stock that may require more maintenance. DIS measures weighted distances to five Boston employment centers, where lower values imply closer access to jobs, impacting property values. RAD is an index of accessibility to radial highways, with higher values indicating better commuting options. TAX reflects the property-tax rate per $10,000, influencing homeowners' expenses and affordability. PTRATIO indicates the pupil-teacher ratio by town, where lower values suggest better educational quality. B, calculated as (1000(Bk - 0.63)^2), represents the proportion of Black residents, a social and

demographic factor. LSTAT denotes the percentage of the lower-status population, a socioeconomic indicator often correlating with lower property values. PRICE, likely the target variable, represents house prices influenced by all these features.

Fig. 3 correlation heatmap illustrates the pairwise correlation between different features in the dataset. It provides insights into the relationships between variables, helping to identify potential predictors of house prices.

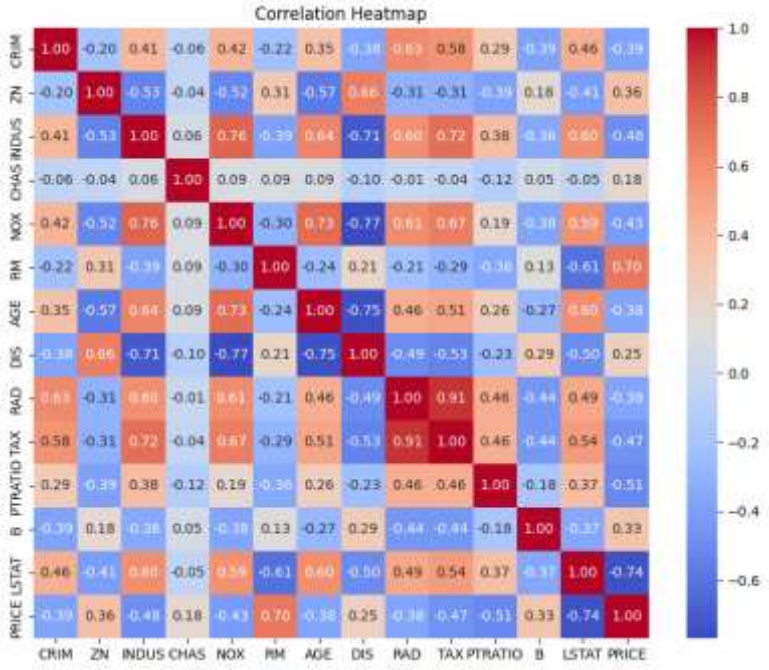

Fig. 3: Correlation heatmap.

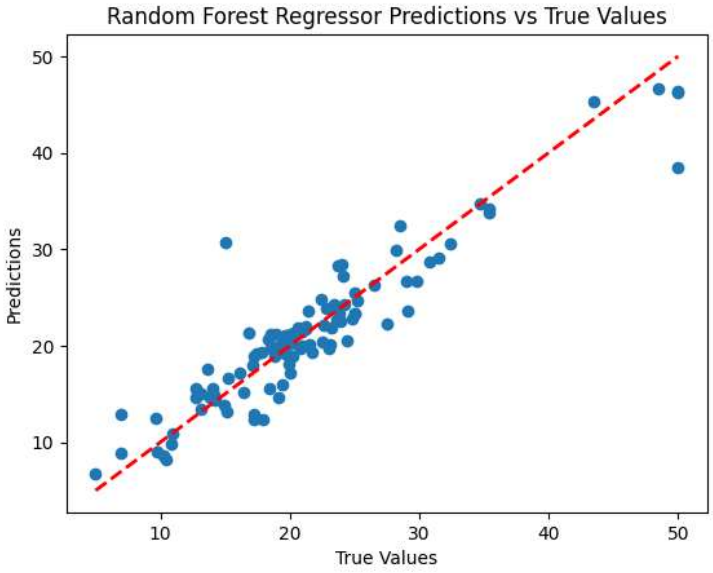

Fig. 4: Prediction graph of RFR model.

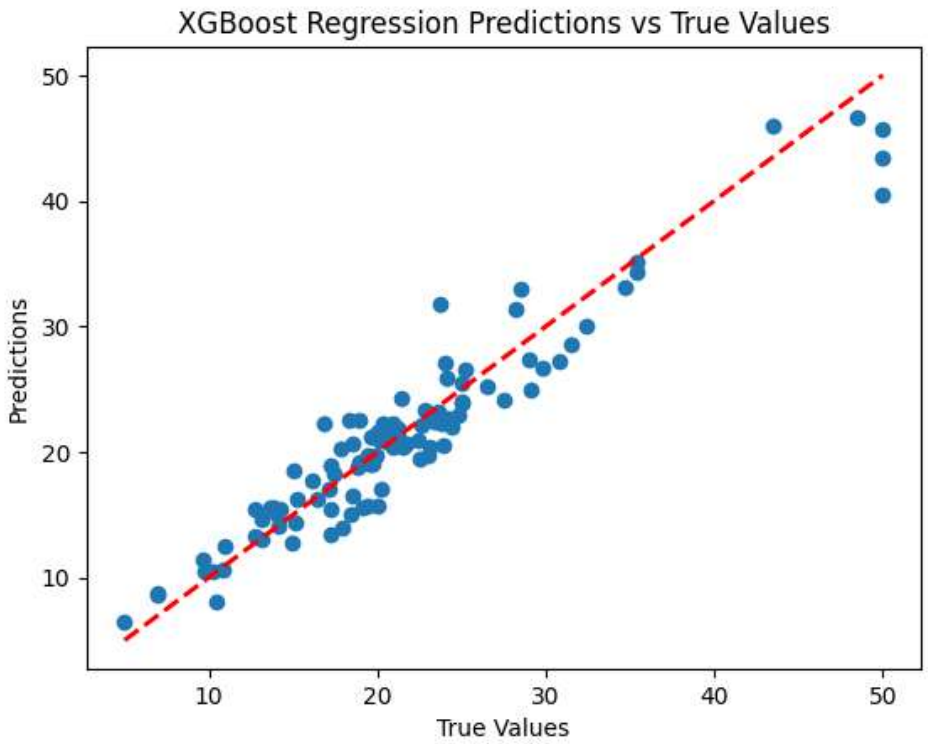

Fig. 5: Prediction graph of XG Boost model.

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	pred
0	0.03738	0.0	5.19	0.0	0.515	6.310	38.5	6.4584	5.0	224.0	20.2	389.40	6.75	20.694065
1	13.35980	0.0	18.10	0.0	0.693	5.887	94.7	1.7821	24.0	666.0	20.2	396.90	16.35	12.686036
2	41.52920	0.0	18.10	0.0	0.693	5.531	85.4	1.6074	24.0	666.0	20.2	329.46	27.38	8.495610
3	0.52058	0.0	6.20	1.0	0.507	6.631	76.5	4.1480	8.0	307.0	17.4	388.45	9.54	25.139647
4	0.04932	33.0	2.18	0.0	0.472	6.849	70.3	3.1827	7.0	222.0	18.4	396.90	7.53	31.328337

Fig. 6: Proposed XG Boost model Prediction on uploaded test dataset.

Fig. 4 prediction graph of the Random Forest Regressor (RFR) model illustrates the relationship between the true values and predicted values of house prices. It helps in visually assessing the accuracy of the model's predictions.

Fig. 5 prediction graph of the XG Boost model showcases the predicted values plotted against the true values of house prices. It provides a visual representation of the model's predictive capabilities.

Fig. 6 displays proposed XG Boost model's predictions on the uploaded test dataset. It demonstrates how the model performs when making predictions on unseen data.

5. CONCLUSION

In conclusion, research in predicting home prices using regression analysis techniques has made significant strides in recent years, offering valuable insights and opportunities for stakeholders in the real estate industry. By leveraging big data and machine learning technologies, researchers have developed predictive models capable of accurately forecasting home prices and uncovering valuable insights into market dynamics.

REFERENCES

[1] Mansi Bosamia,'Positive and Negative Impacts of Information and Communication Technology in our Everyday Life', International Conference on "Disiplinary and Interdisciplinary Approaches to Knowledge Creation in Higher Education: CANADA and INDIA (GENESIS 2013).

[2] Rahul Reddy Nadikattu, 'THE EMERGING ROLE OF ARTIFICIAL INTELLIGENCE IN THE MODERN SOCIETY', International Journal Of Creative Research Thoughts, 2016. [3] Ravi Manne and Sneha C Kantheti, 'Application of Artificial Intelligence in Healthcare: Changes and Challenges', Current Journal 5 of Applied Science and Technology, 2021, DOI:10.9734/CJAST/2021/v40i631320

[4] Woubishet Zewdu Taffese, 'A Survey on Application of Artificial Intelligence in Real Estate Industry', 3rd International Conference on Artificial Intelligence in Engineering andTechnology, 2006.

[5] Ferreira, F. G. D. C., Gandomi, A. H., & Cardoso, R. T. N. 'Artificial Intelligence Applied to Stock Market Trading: A Review', IEEE, 2021, doi:10.1109/access.2021.3058133.

[6] Anandharajan, T. R. V., Hariharan, G. A., Vignajeth, K. K., Jijendiran, R., & Kushmita. (2016).' Weather Monitoring Using Artificial Intelligence'. 2016 2nd International Conference on Computational Intelligence and Networks (CINE). doi:10.1109/cine.2016.26. [7] Tong, W., Hussain, A., Bo, W. X., & Maharjan, S., ' Artificial Intelligence for Vehicle-to-Everything: a Survey', IEEE, 2019 ,doi:10.1109/access.2019.2891073.

[8] Sumit Das, Aritra dey, Akash Paul and NAbamita Roy, 'Applications of Artificial Intelligence in Machine Learning: Review and Prospects', International Journal of Computer Applications, 2015, DOI:10.5120/20182-2402.

[9] Avneet Pannu, 'Artificial Intelligence and its Application in Different Areas', International Journal of Engineering and Innovative Technology (IJEIT), Volume 4, Issue 10, April 2015.

[10] John A. Bullinaria, 'IAI : The Roots, Goals and Sub-fields of AI', 2005. https://www.cs.bham.ac.uk/~jxb/IAI/w2.pdf

[11] Bahia, I. S. (2013). A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study. International Journal of Intelligence Science,03(04), 162- 169. doi:10.4236/ijis.2013.34017.

[12] Mu, J., Wu, F., & Zhang, A. (2014). Housing Value Forecasting Based on Machine Learning Methods. Abstract and Applied Analysis, Volume 2014 (2014), Article ID 648047, 7 pages. Retrieved April 2017, from https://www.hindawi.com/journals/aaa/2014/648047/.

[13] Eli Beracha, Ben T Gilbert, Tyler Kjorstad, Kiplan womack, "On the Relation between Local Amenities and House Price Dynamics", Journal of Real estate Economics, Aug. 2016. [14] Stephen Law, "Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of Metropolitan London", Cities, vol. 60, Part A, pp. 166–179, Feb. 2017.

[15] Binbin Lu, Martin Charlton, Paul Harris & A. Stewart Fotheringham, "Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data", International Journal of Geographical Information Science, pp. 660-681, Jan 2014.

[16] Marco Helbich, Wolfgang Brunauer, Eric Vaz, Peter Nijkamp, "Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria", Urban Studies, vol. 51, Issue 2, Feb. 2014.

[17] Sean Holly, M. Hashem Pesarana, Takashi Yamagata, "A spatio-temporal model of house prices in the USA", Journal of Econometrics, vol. 158, Issue 1, pp. 160–173, Sep. 2010.

[18] Joep Steegmans, Wolter Hassink, "Financial position and house price determination: An empirical study of income and wealth effects", Journal of Housing Economics, vol. 36, pp. 8-24, June 2017.