

Prediction of House Prices using Machine Learning Techniques

Yogesh¹, Vishal Solanki¹, Ashish¹, Neha Gupta²

¹ Student of School of Engineering and Technology, Sushant University, Gurgaon, Haryana, India.

² Associate Professor of School of Engineering and Technology, Sushant University, Gurgaon, Haryana, India.

Abstract

The demand for accurate house price prediction systems has increased significantly with the rapid growth of urbanization and real estate markets. Due to dependency on brokers and lack of transparent pricing mechanisms, traditional methods often lead to incorrect property valuation, financial losses, and poor decision-making by buyers and sellers. To overcome these issues and improve pricing accuracy in modern real estate systems, machine learning-based prediction models have become highly important [3][17][18].

This research article describes the design and development of a House Price Prediction System using machine learning algorithms such as Linear Regression, Random Forest, Decision Tree, XGBoost, and Gradient Boosting, to estimate property prices based on various features like area, location, number of bedrooms, and amenities [2][6][7][8][9][11][12].

To automate the prediction process, the proposed system combines data preprocessing techniques, feature engineering, model training, and evaluation metrics. The system works in an intelligent manner by analyzing historical housing data, identifying patterns, and generating accurate predictions. The implementation aims to reduce human dependency, eliminate pricing errors, and improve decision-making in the real estate market [3][17].

Furthermore, this research explores future advancements such as integration of deep learning models like Artificial Neural Networks (ANN), real-time data analysis, and deployment through web and mobile applications. Due to its high accuracy, scalability, and cost-effectiveness, the system can play a significant role in smart city infrastructure and modern real estate analytics [1][10][13][20].

Keywords: *House Price Prediction, Machine Learning, Regression Algorithms, Real Estate Analytics, Random Forest, XGBoost, Feature Engineering, Predictive Modeling*

Introduction

The rapid growth of urbanization and increasing demand for housing have made property price estimation a major challenge in the real estate industry. The rise in population and expansion of cities have significantly increased the complexity of determining accurate house prices, as multiple factors such as location, area, number of bedrooms, and infrastructure influence pricing decisions [3][18].

Traditional property valuation methods rely heavily on brokers and manual estimation, which often leads to inconsistent pricing and lack of transparency. As a result, buyers and sellers face difficulties in determining whether a property is fairly priced. This creates a strong need for intelligent systems that can provide accurate and data-driven price predictions [17][19].

Advancements in machine learning and data analytics have made it possible to develop automated prediction systems. These systems analyze large datasets and identify hidden patterns to generate accurate results. Studies have shown

that machine learning-based models can significantly improve prediction accuracy compared to traditional methods [12][19][20].

The objective of this research is to develop an efficient house price prediction system that can estimate property values based on input features such as area, location, and number of bedrooms. The system uses multiple machine learning algorithms and compares their performance to select the most suitable model. Additionally, a user-friendly interface is developed to provide real-time predictions to users [2][6][7][10].

Literature Review

Several researchers have explored the application of machine learning techniques for predicting house prices using historical real estate data. Platforms such as Kaggle and Zillow provide datasets that include features like area, number of bedrooms, location, and amenities, which are essential for prediction. Traditional methods such as linear regression and decision trees have been widely used; however, they are limited in handling complex and nonlinear relationships in data [4][5][11].

Kuiper (2008) demonstrated that multivariate regression models are effective in predicting numerical values and improving model accuracy. The study emphasized that selecting relevant features such as area, number of rooms, and location significantly enhances prediction performance [2][19].

In recent years, ensemble learning techniques such as Random Forest and Gradient Boosting have gained popularity for house price prediction. These models are capable of handling large datasets and nonlinear relationships, resulting in improved accuracy and reduced overfitting [7][6][11][12].

Gegic et al. (2019) analyzed the use of multiple machine learning algorithms including Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forest. Their findings suggest that hybrid and ensemble models provide better performance compared to individual models, especially for complex datasets [4].

Dholiya et al. (2019) developed a machine learning-based system that predicts property prices based on user input. The study highlighted the importance of data preprocessing techniques such as data cleaning, handling missing values, and feature encoding to improve model efficiency [2].

Further research by Richardson (2009) identified key factors influencing property prices, including location, connectivity, and infrastructure. The study concluded that incorporating multiple features improves prediction accuracy and reliability [3][18].

Support Vector Machines (SVM) have also been applied for prediction tasks, showing better performance in high-dimensional datasets and reducing overfitting. However, proper parameter tuning is required to achieve optimal results [9].

Recent advancements in deep learning have introduced Artificial Neural Networks (ANN) for predicting house prices. These models can capture complex patterns in large datasets and significantly improve prediction accuracy. Modern machine learning libraries such as Scikit-learn and TensorFlow have further simplified the implementation of such models [1] [2][10][13].

Overall, existing literature indicates that machine learning and deep learning techniques provide effective solutions for house price prediction. However, challenges such as data quality, feature selection, and model optimization still remain, highlighting the need for more efficient and scalable prediction systems [17][19].

Methodology

The methodology involves several steps including data collection, preprocessing, model building, and evaluation.

- Data is collected from real estate datasets
- Data preprocessing includes cleaning, handling missing values, and encoding
- Exploratory Data Analysis (EDA) is performed
- Dataset is split into training and testing sets (80:20)
- Models are trained using various algorithms
- Performance is evaluated using R^2 , MAE, and MSE

Data Gathering

The dataset is collected from online platforms such as Kaggle.com and real estate websites. The dataset contains housing-related features such as area (in square feet), number of bedrooms (1BHK, 2BHK, 3BHK, 4BHK), number of bathrooms, location, availability of amenities, and price. These features are essential for predicting house prices accurately.

Create Environment

An environment is created using Anaconda to isolate the project workspace from other environments. Required libraries and dependencies are installed within this environment, allowing better management and flexibility during development.

Data Reading

The dataset is loaded using Python and analyzed for null values, structure, number of rows and columns, numerical and categorical features, and unique values. This step helps in understanding the dataset before applying machine learning models.

Data Pre-processing

Data preprocessing includes cleaning and transforming the dataset for better performance. Irrelevant features are removed, and necessary features such as area, BHK type, and location are retained.

Exploratory Data Analysis (EDA) is performed using graphs and plots such as:

- Price vs Area
 - Price vs BHK (1BHK, 2BHK, 3BHK, 4BHK)
 - Price vs Location
 - Price vs Number of Bathrooms
- Visualization techniques such as pairplots, heatmaps, and bar graphs are used to understand relationships between variables.

Categorical features such as location are converted into numerical form using One Hot Encoding. Correlation between features is analyzed to identify important variables. Finally, dependent variable (Price) and independent variables (Area, BHK, Location, etc.) are defined.

Train-Test Split

Once the dependent and independent features have been assigned, we proceed with the splitting of the dataset into training and testing data. We use 80% of the data to train our model and 20% to test it.

Model Building

Following the Train-Test split, data modeling is complete, and the process of building the model begins. The model is defined, along with a few parameters, for future implementation. After the model is built, various algorithms are used to create the final results. After building the model, the following algorithms are used for predictive analysis.

Linear Regression: It is a linear approach in statistics for modeling the relationships between a scalar response and dependent and independent variables. In linear regression, relationships are modelled using functions such as linear predictor, and unknown model parameters are estimated from data.

Lasso Regression: It is a sort of linear regression in which the data values are shrunk towards a data point in the center, or, in simpler terms, the mean of the data. The Lasso procedure supports simple and sparse models with fewer parameters. When a model has a high amount of multicollinearity, this regression provides the best fit for that model. This approach can also be used if some aspects of model selection, such as variable selection or parameter elimination, need to be automated. The abbreviation 'LASSO' stands for Least Absolute Shrinkage and Selection Operator.

Ridge Regression: It is a regression approach used for tuning a model and analyzing multicollinear data. This function implements L2 regularization. The multicollinearity of the data results in unbiased least-squares, a huge variance, and hence the predicted values are considerably far from the actual values.

Bayesian Ridge Regression: This regression is used to estimate any probabilistic model of any regression issue using linear regression formulation with the use of probability distributors, providing a natural process that survives data insufficiency or poor data distribution.

Random Forest Regression: Random Forest is a Supervised Learning Algorithm that employs the ensemble learning approach for classification and regression. Random forests are made up of trees that run parallel to each other and have no interaction while they develop. Random Forest is a metaestimator that aggregates the outcomes of several predictions. It also aggregates numerous decision trees with certain modifications.

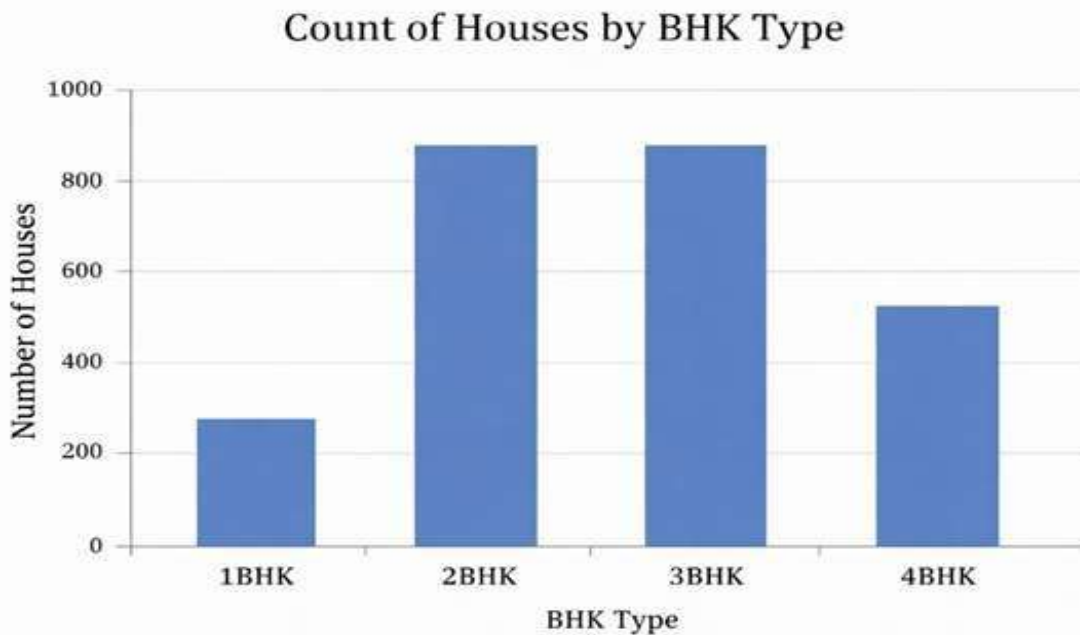
XGBoost Regression: XGBoost is a very powerful technique for creating supervised regression models. XGBoost is an ensemble learning strategy that includes training individual models and then merging them (base learners) to get a single prediction.

Implementation

Adding a new feature Price per Square Foot, which represents the price of the property per unit area, is calculated and stored in the final dataset, and irrelevant attributes are removed. Additionally, features such as BHK (1BHK, 2BHK, 3BHK, 4BHK) and location are retained for accurate prediction.

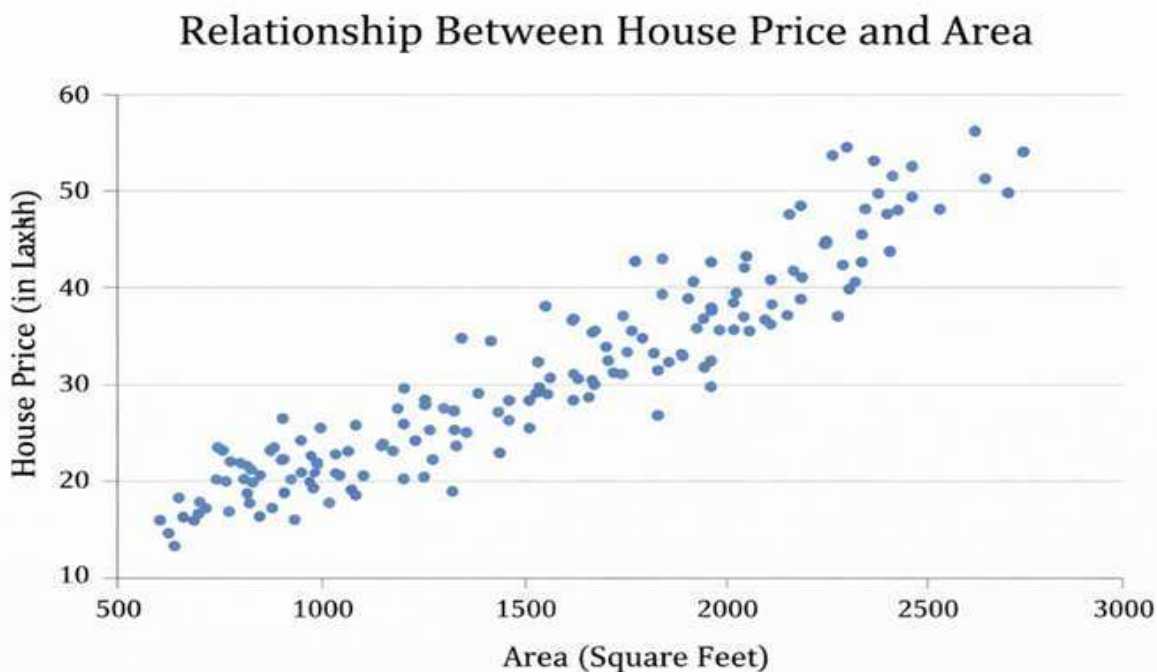
Exploratory Data Analysis

In this stage, we summarize the major characteristics of the housing dataset using statistical graphics and visualization techniques. Various graphs and charts are plotted to understand the relationship between features such as area, BHK type, location, and price.



House count in relation to BHK type:

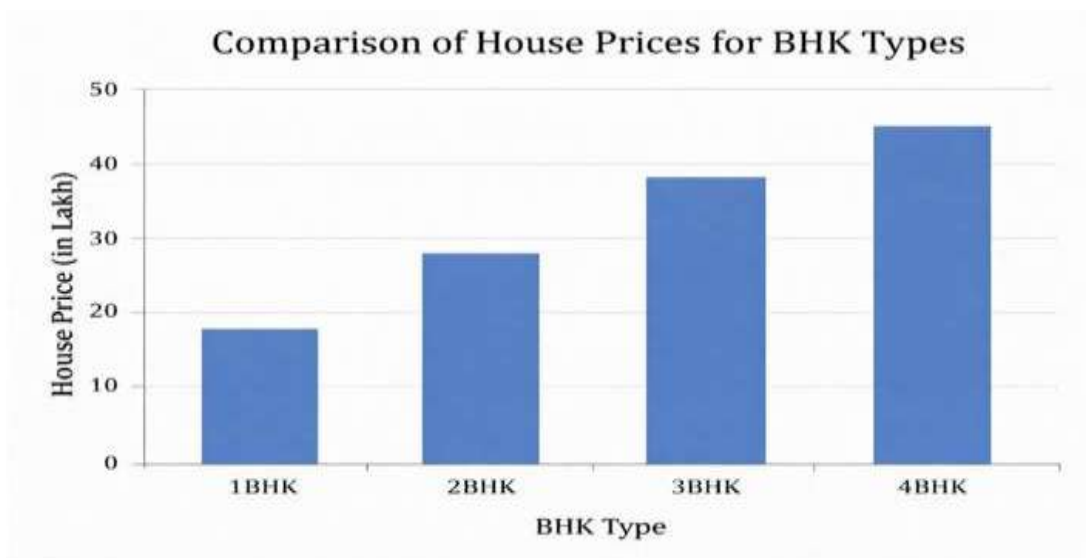
The bar graph represents the number of properties available for each category such as 1BHK, 2BHK, 3BHK, and 4BHK. It is observed that 2BHK and 3BHK properties are more common in the dataset.



Price v/s Area

Comparison of house price vs area:

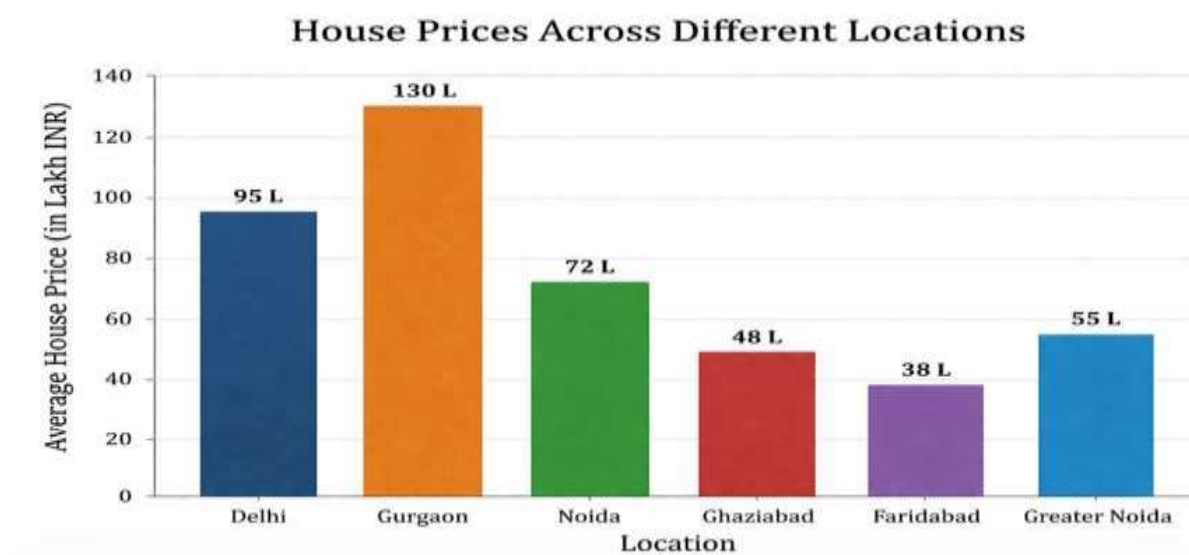
The scatter plot shows the relationship between area (in square feet) and price. It is observed that as area increases, the price of the house also increases.



BHK v/s Price

Comparison of BHK type vs price:

The graph shows that houses with higher BHK (3BHK, 4BHK) generally have higher prices compared to 1BHK and 2BHK.



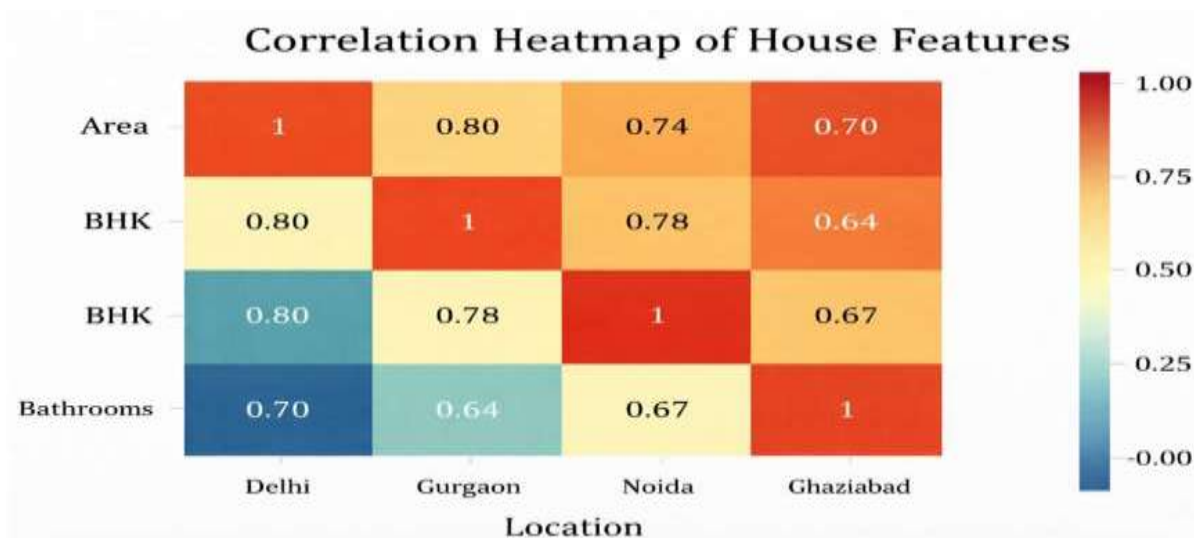
Location v/s Price

Comparison of location vs price:

This graph shows how property prices vary based on location. Prime locations have significantly higher prices compared to other areas.

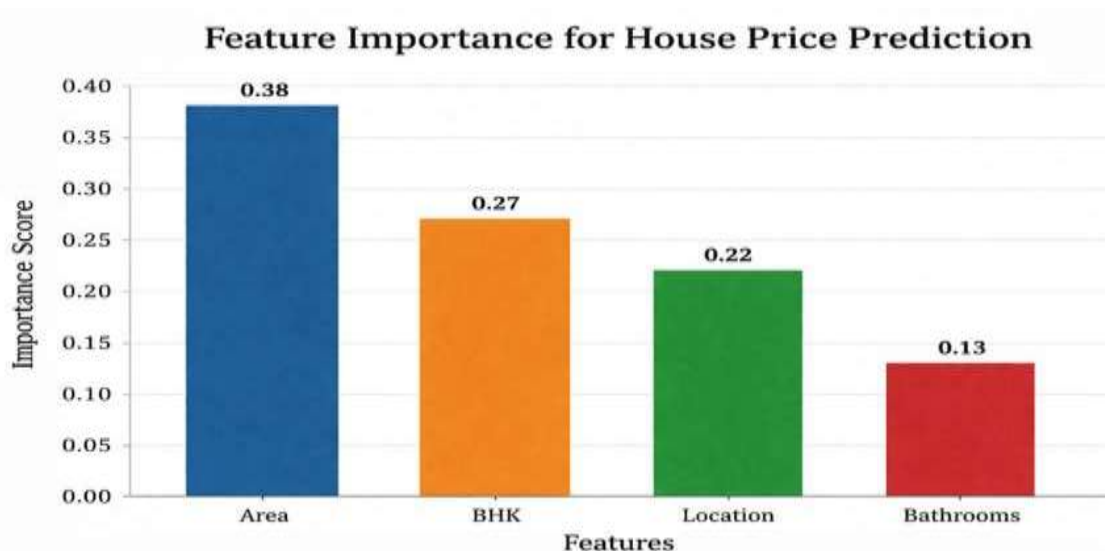
One Hot Encoding

The One Hot Encoding technique is used to convert categorical variables such as location into numerical form. Each category is represented as a binary column, allowing machine learning models to process the data efficiently.



Correlation Heatmap

Heatmap of correlation features: The heatmap shows the relationship between variables such as area, BHK, bathrooms, and price. Darker colors indicate stronger correlation.



Feature Importance :

Feature Importance of dataset:

Feature importance identifies which variables contribute most to predicting house prices. Area and location are usually the most important features.

Model Building

After splitting the dataset, models are trained using training data and evaluated using testing data. Various algorithms are applied to generate predictions and compare performance.

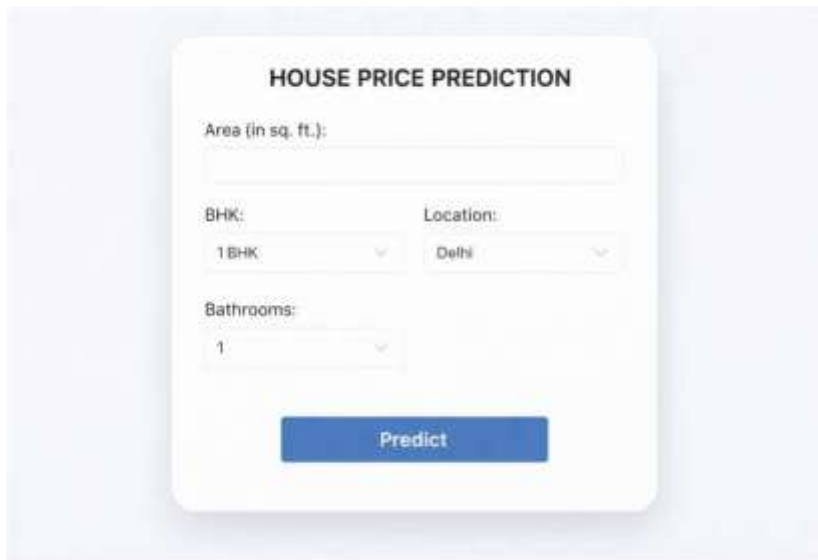
Developing a Web Application

A web application is developed using HTML, CSS, and JavaScript for frontend and Flask for backend.

Users enter:

- Area (sq ft)
- BHK (1–4)
- Location
- Bathrooms

The system predicts the estimated house price based on these inputs.



A screenshot of a web application titled "HOUSE PRICE PREDICTION". The form contains the following fields: "Area (in sq. ft.):" with a text input field; "BHK:" with a dropdown menu showing "1 BHK"; "Location:" with a dropdown menu showing "Delhi"; and "Bathrooms:" with a dropdown menu showing "1". Below these fields is a blue "Predict" button.

Web Application



Actual vs Predicted

Comparison of actual vs predicted prices: The graph shows that predicted values are very close to actual values, indicating high accuracy.

Results

After applying various regression algorithms to the house price prediction model, evaluation metrics such as R^2 Score, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were calculated to compare the performance of each algorithm.

Among all the models, the Decision Tree Regression algorithm achieved the highest R^2 score of approximately 0.95, indicating that it provided the most accurate predictions. The model also produced lower error values compared to other algorithms, which shows better performance in predicting house prices.

The comparison of Actual vs Predicted values shows that both values are closely aligned, indicating high prediction accuracy. Algorithms such as Gradient Boosting and XGBoost also performed well, but Decision Tree outperformed them in terms of accuracy and error metrics.

Thus, based on the evaluation metrics and graphical analysis, it can be concluded that the Decision Tree Regression model is the most suitable algorithm for house price prediction in this study.

Conclusion

Predicting house prices is a difficult task due to the large number of features and parameters such as area, location, and number of bedrooms (1BHK, 2BHK, 3BHK, 4BHK) that must be analyzed to obtain reliable results. The first and most important phase is data collection and preprocessing. The model is then defined and built in order to implement machine learning algorithms and generate predictions[3].

After applying various regression algorithms to the model, it was concluded that the Decision Tree Regression algorithm performed the best, achieving the highest R^2 score of approximately 0.95. This indicates that it provided the most accurate predictions, as shown by the Actual vs Predicted graph where both values are closely aligned[7][8][10].

In addition to having the highest R^2 score, the Decision Tree model also showed lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), indicating that the prediction errors were minimal. Therefore, the developed system is efficient and can be used for accurate house price prediction[17][20].

References

1. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
2. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
3. Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1), 81–102.
4. Kaggle. (2023). House Prices Dataset. Available: <https://www.kaggle.com>
5. Zillow Research. (2023). Housing Data and Price Trends.
6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*, 785–794.
7. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
8. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.
9. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
10. Kaur, J., & Gupta, N. (2019). Constructive neural network: A framework. *International Journal of Engineering and Advanced Technology*, 9(2).
11. Kaur, J., & Gupta, N. (2020a). Bipolar sigmoid algorithm for designing constructive neural network. *International Journal on Emerging Technologies*, 11(2), 991–996.
12. Kaur, J., & Gupta, N. (2020b). Extended bipolar sigmoid algorithm for enhancing performance of constructive neural network. *International Journal on Emerging Technologies*, 11(2), 1034–1038.
13. Kaur, J., & Gupta, N. (n.d.). Artificial neural network: A review. *International Journal of Technical Research & Science*, 1–4.
14. Dargan, J., & Gupta, N. (2024). Integration of blockchain for IoT communication. In *Adaptive Power Quality for Power Management Using Smart Technologies* (pp. 319–343). CRC Press.

15. Dargan, J., Gupta, N., & Singh, L. (2021). Blockchain based energy management system: A proposed model. *IEEE ICTAI*, 510–514.
16. Dewan, S., Duhan, L., & Gupta, N. (2021). Secure electronic voting system based on mobile app and blockchain. *Recent Advances in Computer Science and Communications*, 14(9).
17. Zhang, Y., et al. (2021). Machine learning approaches for real estate price prediction. *International Journal of Data Science*.
18. Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction. *Journal of Real Estate Research*.
19. Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*.
20. Amato, G., et al. (2017). Deep learning for real-world applications. *Expert Systems with Applications*.