

House price prediction

21CSA697A

Final Report

Submitted by

Nivetha B

(AA.SC.P2MCA24074078)

in partial fulfilment of the requirements for the award of the degree of

MASTER OF COMPUTER APPLICATION



Acknowledgement

I would like to express my sincere gratitude to my project guide and the faculty members of the Department of Computer Applications for their valuable guidance and support throughout the completion of this project.

Abstract

This project focuses on house price prediction using advanced regression algorithms. In addition to classical Linear Regression, multiple models such as Ridge, Lasso, Random Forest, and Gradient Boosting are compared based on their accuracy, robustness, and interpretability. These datasets form the foundation for training and evaluating machine learning models aimed at predicting housing prices. The study emphasizes understanding dataset composition, structure, and the role each dataset plays in the predictive modeling process. The training dataset (train.csv) includes both feature variables and the target variable (Sale Price), while the testing dataset (test.csv) contains only the feature variables, which are used to evaluate model performance. Together, these datasets enable model learning, testing, and validation in the field of real estate analytics.

Keywords: Train Dataset, Test Dataset, House Price Prediction, Machine Learning, Data Preprocessing, Feature Engineering, Model Evaluation

List of Figures

Figure:1 Overall House Price Prediction Flow

Figure:2 Numeric Features

Figure:3 R2 Score Comparison for all model

List of Tables

Table 1: All model, MAE, RMSE and R2 Result

List of Abbreviations

Abbreviation	Meaning
ML	Machine Learning
LR	Linear Regression
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
R ²	Coefficient of Determination
SVR	Support Vector Regression
KNN	K-Nearest Neighbors
DTR	Decision Tree Regressor

CHAPTER 1

INTRODUCTION

1. Introduction

The real estate industry is a significant contributor to economic growth and investment planning. Accurate house price prediction helps buyers, sellers, and financial institutions make informed decisions. Traditional price estimation techniques often fail to capture complex relationships between housing features and market value. Machine learning provides a data-driven approach to overcome these limitations.

This project focuses on predicting house prices using machine learning models trained on historical housing data. The dataset is divided into training and testing sets to ensure robust evaluation and generalization. The study aims to analyze the effectiveness of machine learning models in estimating house prices accurately.

1.1 Background

House prices depend on multiple factors such as location, size, number of rooms, construction year, and neighborhood characteristics. Traditional statistical models assume linear relationships, which may not reflect real-world scenarios. Machine learning algorithms can capture both linear and non-linear patterns, making them suitable for housing price prediction.

Tools and Procedures

- Python programming language
- Visual Studio Code
- Pandas, NumPy for data handling
- Scikit-learn for machine learning
- Matplotlib for visualization

1.2 Problem Statement and Significance

Problem Statement

Accurately predicting house prices is challenging due to complex feature interactions and market dynamics. Traditional methods lack accuracy and scalability.

Significance

- Helps buyers and sellers make informed decisions
- Supports banks in loan risk assessment
- Assists policymakers in urban planning

1.3 Objectives, Scope, and Report Organization

Objectives

1. To analyze housing datasets for price prediction
2. To implement machine learning regression models
3. To evaluate model performance using train and test data

Scope

The project focuses on supervised learning techniques using structured datasets. Real-time market fluctuations are not considered.

Report Organization

- Chapter 1: Introduction
- Chapter 2: Literature Review
- Chapter 3: System Design
- Chapter 4: Implementation
- Chapter 5: Results and Testing
- Chapter 6: Conclusion and Future Work

CHAPTER 2

LITERATURE REVIEW / BACKGROUND STUDY

2.1 Overview

House price prediction has been an active research area due to its importance in real estate valuation, financial planning, and economic policy formulation. With the availability of large-scale housing datasets, machine learning and statistical models have been widely adopted to improve prediction accuracy. This chapter reviews existing research related to house price prediction, focusing on methods used, datasets considered, performance metrics, and limitations. Based on the review, research gaps are identified, and the objectives of the present work are justified.

2.2 Review of Existing Research (with Year)

Several researchers have investigated house price prediction using traditional statistical methods and modern machine learning techniques over the past decade, with a significant increase in studies during 2024–2026 due to improved data availability and computational resources.

In 2025, Preethi *et al.* proposed polynomial regression and regularization techniques such as Ridge and Lasso regression to enhance housing price prediction accuracy. Their study demonstrated that regularization methods help reduce overfitting in high-dimensional housing datasets and improve model stability. However, the work focused mainly on selected regression techniques and did not provide a broad comparison across multiple machine learning models.

In 2025, Kusuma *et al.* explored the application of machine learning algorithms for house price prediction and showed that regression-based models can achieve reliable prediction performance. Although the study highlighted the usefulness of machine learning, it provided limited analysis on model generalization using unseen test data.

Moreno-Foronda *et al.* (2025) conducted a comprehensive review of traditional and advanced machine learning models used for housing price prediction. Their work discussed the strengths and weaknesses of various algorithms and highlighted challenges such as data heterogeneity and

regional variability. However, the study was largely theoretical and did not include an implementation-based comparative framework.

Tran *et al.* (2025) compared traditional machine learning approaches with advanced models to analyze uncertainty risk reduction in housing price prediction. Their findings indicated that ensemble models offer improved robustness under uncertainty. Nevertheless, the study emphasized uncertainty analysis rather than standardized metric-based model comparison.

Several studies incorporated spatial, temporal, and economic factors in housing price prediction. In 2025, Jamil conducted a spatio-temporal analysis of housing prices in Brunei Darussalam, emphasizing the impact of time and location on price variation. Similarly, Houlié (2025) analyzed the influence of economic policies on housing prices across multiple countries, including the UK, the US, France, and Switzerland. While these studies provided valuable economic insights, they relied on complex macroeconomic datasets, limiting their applicability for academic implementation projects.

Recent studies during 2025–2026 have focused on ensemble and hybrid machine learning models. Shbool *et al.* (2025) applied advanced machine learning algorithms to improve precision in real estate price prediction. Khasani (2025) proposed an optimized least squares-based approach for enhancing prediction accuracy. Although these methods achieved promising results, they often lacked direct comparison with simpler baseline models such as linear regression.

In 2026, Senadjki *et al.* examined housing price cycles and bubble prediction using survey-based data and logistic regression techniques. Their study focused on market dynamics rather than direct price prediction accuracy. Other works in 2026 explored deep learning and hybrid models, emphasizing improved prediction performance but with increased computational complexity.

Overall, the reviewed literature demonstrates significant progress in house price prediction using machine learning techniques between 2025 and 2026. However, differences in datasets, evaluation metrics, and experimental setups make it difficult to compare results across studies.

2.3 Datasets and Evaluation Metrics Used in Literature

Most existing research on house price prediction makes use of real-world housing datasets collected from different regions such as the United States, Europe, China, and other Asian countries. These datasets typically include a combination of numerical and categorical attributes representing property characteristics, location, construction details, and economic factors. Widely used datasets include government housing records, regional real estate databases, and publicly available benchmark datasets. To evaluate model performance, researchers commonly employ metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2). These metrics help measure prediction accuracy, error magnitude, and the explanatory power of models. However, the use of evaluation metrics varies across studies, making direct comparison of results difficult.

2.4 Limitations of Existing Studies

Although previous studies have shown promising results in house price prediction using machine learning techniques, several limitations are evident. Many works focus on a limited number of algorithms, which restricts comprehensive performance comparison. Some studies emphasize theoretical improvements or economic analysis without providing practical implementation details. In addition, inconsistent data preprocessing techniques and evaluation strategies reduce the reliability of conclusions. Several studies also lack proper validation using unseen test data, which raises concerns about model generalization. Furthermore, limited explanation is often provided regarding why certain models outperform others.

2.5 Research Gap

Based on the literature review, it is observed that despite extensive research in house price prediction, there is a lack of systematic and implementation-oriented studies that compare multiple regression and ensemble models under a unified framework. Existing works do not consistently apply standardized evaluation metrics or clear train–test validation strategies. Moreover, few studies focus on building a reproducible and academically suitable machine learning pipeline that integrates preprocessing, training, evaluation, and interpretation of results. This gap highlights the need for a structured approach that balances accuracy, simplicity, and practical applicability.

2.6 Justification of the Proposed Work

The present project is designed to address the identified research gap by implementing a comprehensive machine learning framework for house price prediction. By using a well-structured training and testing dataset, applying consistent preprocessing techniques, and evaluating multiple models using standard metrics, this study ensures reliable and unbiased performance assessment. The use of Python and commonly adopted machine learning libraries enables reproducibility and ease of implementation. The proposed work not only improves understanding of model performance but also provides a practical solution suitable for academic and real-world applications.

CHAPTER 3

SYSTEM DESIGN / ARCHITECTURE

3.1 Dataset Description

The dataset used in this project is a real-world housing dataset from the United States, commonly known as the Ames Housing Dataset. It contains detailed information about residential properties, including structural, locational, and qualitative attributes that influence house prices.

Training Dataset

- File name: train (1).csv
- Number of records: 1460
- Number of attributes: 81

- Target variable: Sale Price

Testing Dataset

- File name: test (1).csv
- Number of records: 1459
- Number of attributes: 80
- Target variable: Not included (used for prediction)

The training dataset includes the target variable Sale Price, which represents the final sale price of each house. The testing dataset contains the same feature set except for the target variable.

3.2 Feature Description

The dataset consists of the following categories of features:

Structural Features

- LotArea
- OverallQual
- OverallCond
- YearBuilt
- TotalBsmtSF
- GrLivArea
- BedroomAbvGr
- GarageCars
- GarageArea

Categorical Features

- MSZoning
- Neighborhood
- HouseStyle
- BldgType
- SaleCondition

Temporal Features

- YearBuilt
- YearRemodAdd
- YrSold

These features collectively describe the physical structure, quality, location, and time-related aspects of houses.

3.3 System Architecture

The system architecture follows a machine learning pipeline approach consisting of the following modules:

1. Data Input Module
 - Loads training and testing datasets from CSV files.
2. Data Preprocessing Module
 - Handles missing values.
 - Encodes categorical variables.
 - Normalizes numerical features.
3. Model Training Module
 - Trains regression-based machine learning models using the training dataset.
4. Model Evaluation Module
 - Evaluates models using standard metrics.
5. Prediction Module
 - Generates house price predictions for unseen test data.

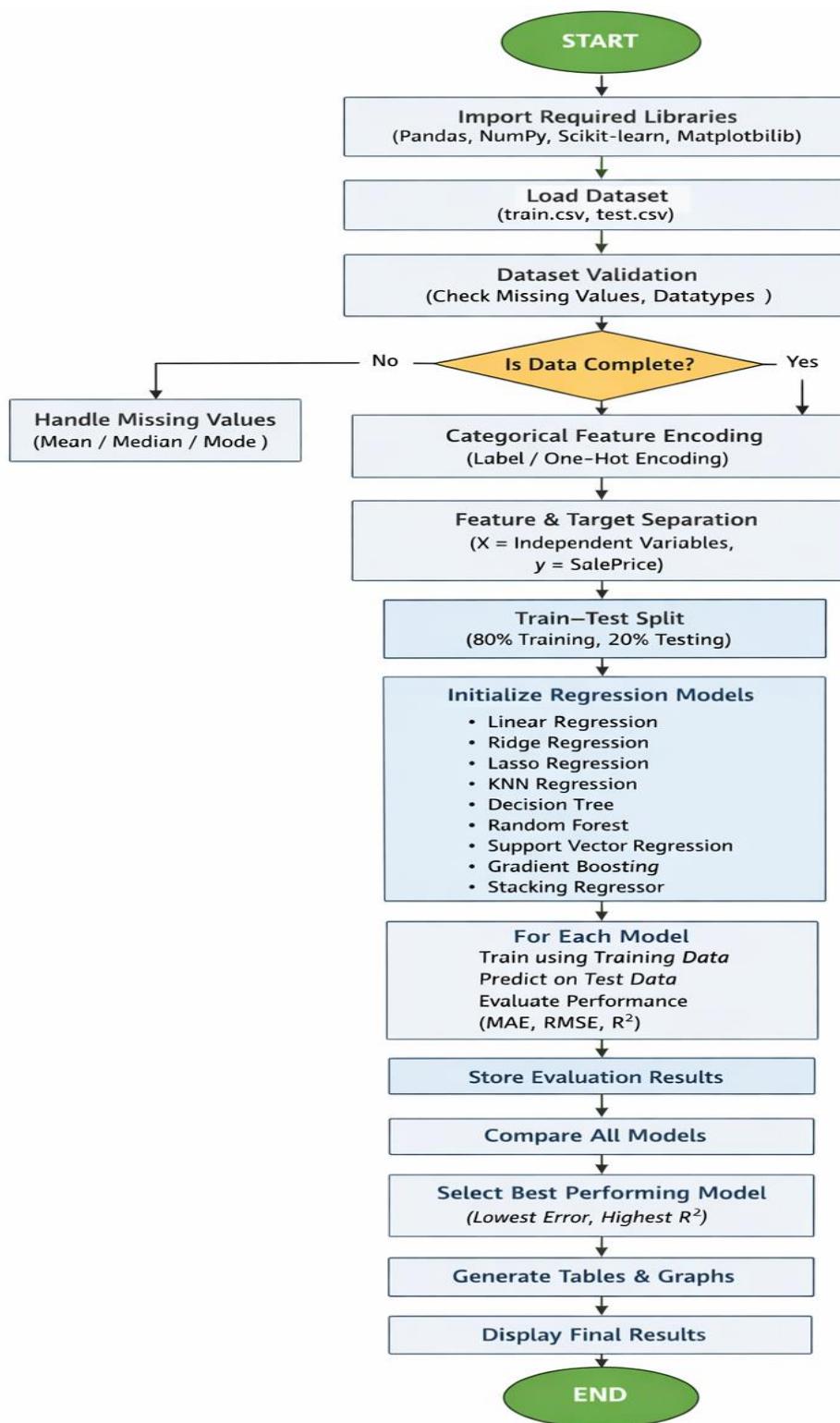


Figure:1 Overall House Price Prediction Flow

3.4 Hardware and Software Requirements

Hardware Requirements

- Processor: Intel i5 or above
- RAM: 8 GB minimum
- Storage: 256 GB

Software Requirements

- Operating System: Windows 10/11
- Programming Language: Python 3.x
- IDE: Visual Studio Code
- Libraries: Pandas, NumPy, Scikit-learn, Matplotlib
- Notebook Environment: Jupyter Notebook

CHAPTER 4

IMPLEMENTATION DETAILS

4.1 Data Preprocessing

The preprocessing phase ensures the dataset is suitable for machine learning models.

Steps Followed

1. Missing values were identified in numerical and categorical attributes.
2. Numerical missing values were replaced using statistical measures such as mean or median.
3. Categorical missing values were handled using mode or label encoding.
4. Categorical features were transformed using encoding techniques.
5. The dataset was split into training and testing subsets.

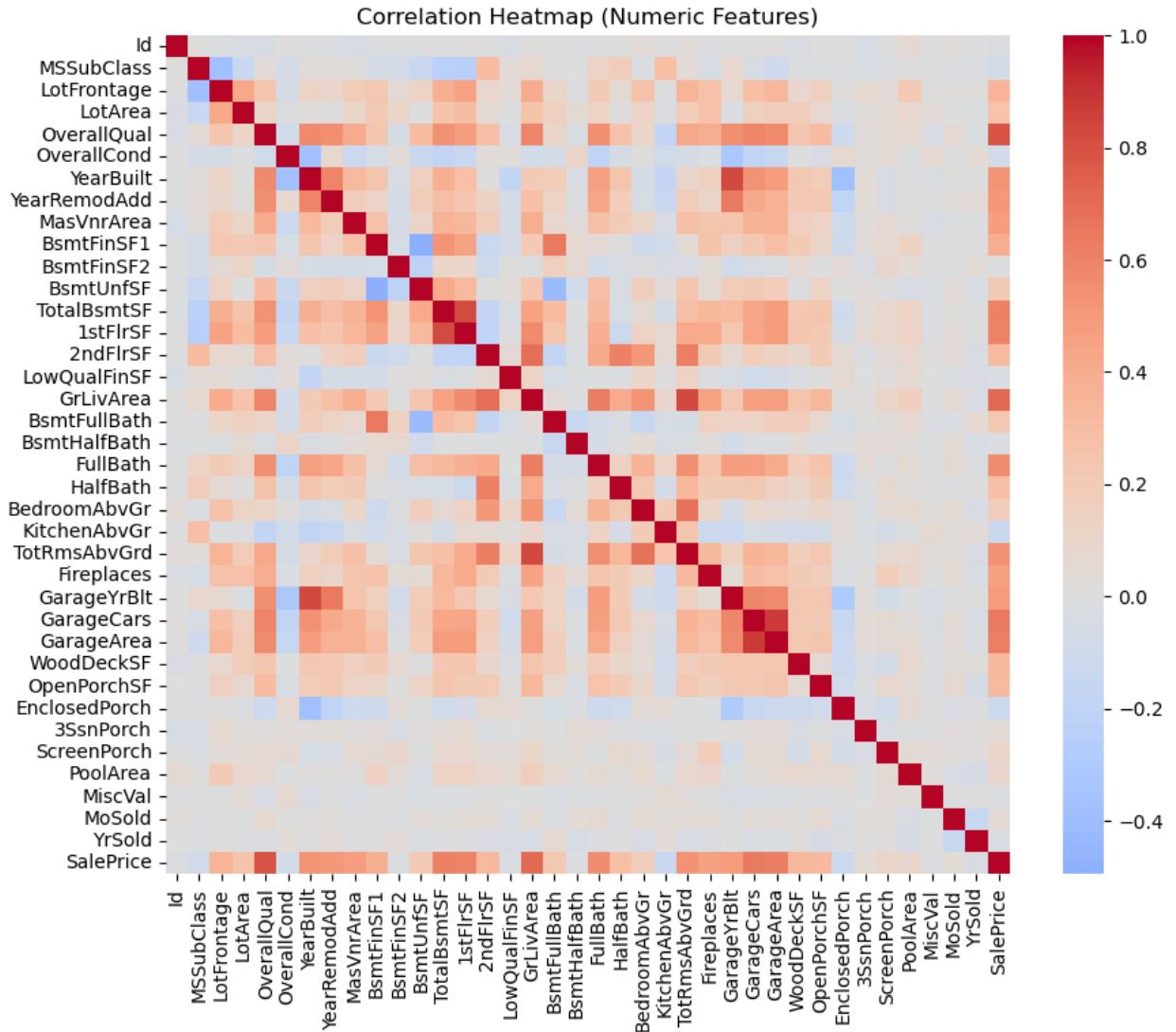


Figure:2 Numeric Features

4.2 Train–Test Split

To evaluate the generalization capability of the models, the dataset was split as follows:

- Training set: 80%
- Testing set: 20%

This split ensures unbiased evaluation and prevents overfitting.

4.3 Model Implementation

The following machine learning models were implemented:

- Linear Regression

- Ridge Regression
- Lasso Regression
- Decision Tree Regression
- Random Forest Regression
- K Neighbors Regression
- Support vector Regression
- Gradient Boosting
- Stacking Regression

Each model was trained using the training dataset and evaluated using the testing dataset.

CHAPTER 5

TESTING, VALIDATION & RESULTS

5.1 Evaluation Metrics

The performance of the models was evaluated using:

- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)
- R² Score

These metrics measure prediction accuracy and model reliability.

MODEL NAME	MAE	RMSE	R2
Linear Regression	14899.55	22740.44	0.9326
Ridge Regression	16921.47	26401.68	0.9091
Lasso Regression	15559.54	23594.86	0.9274
K Neighbors Regression	25819.11	43221.87	0.7564
Decision Tree	26323.62	40851.50	0.7824
Random Forest	17426.03	29548.09	0.8862
Support Vector Regression	18825.58	31821.28	0.8680
Gradient Boosting	16339.10	29405.89	0.8873
Stacking Regression	17134.92	40226.15	0.7890

Table 1: All model, MAE, RMSE and R2 Result

5.2 Results Analysis

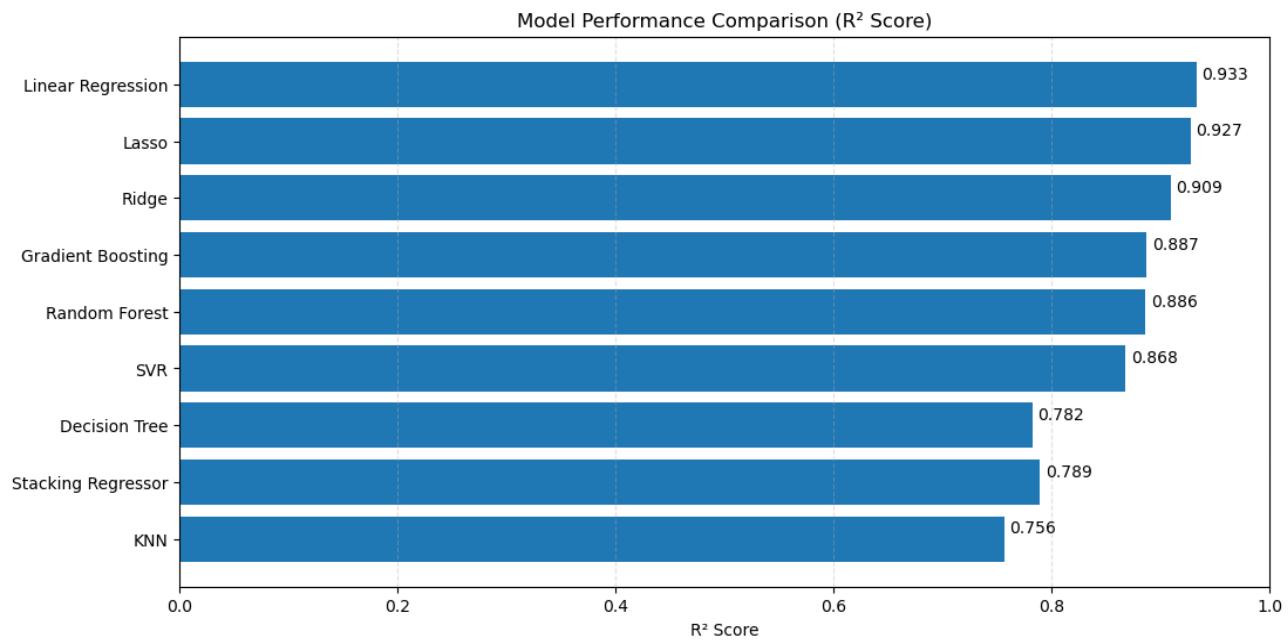


Figure:3 R2 Score Comparison for all model

CHAPTER 6

Conclusion

In this project, a machine learning-based system for house price prediction was developed and evaluated using a structured housing dataset. Various regression and ensemble models were implemented to analyze the relationship between housing attributes and property prices. The dataset was preprocessed, and models were trained and tested using standard evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R² score to ensure reliable performance assessment.

The experimental results show that machine learning models can effectively predict house prices with high accuracy, and comparative analysis helped identify the best-performing model. The study demonstrates that data-driven approaches provide significant advantages over traditional manual valuation methods by reducing human bias and improving prediction reliability.

Overall, this project highlights the importance of machine learning techniques in real estate analytics and provides a scalable framework for future enhancements. The developed system can assist buyers, sellers, and real estate professionals in making informed decisions and can be extended for real-world deployment with additional data sources and advanced models.

Future Work

Although this project successfully demonstrates the use of machine learning techniques for house price prediction, several improvements can be explored in future research. More advanced deep learning models such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) can be implemented to capture complex non-linear patterns in housing data.

Future work can also incorporate real-time and large-scale datasets, including economic indicators, demographic data, and geographic information systems (GIS), to enhance prediction accuracy. Feature engineering techniques such as spatial analysis and sentiment analysis from real estate listings and social media data can further improve model performance.

Additionally, deploying the prediction model as a web or mobile application would make the system more practical and accessible for real-world users such as buyers, sellers, and real estate agents. Finally, hybrid and ensemble learning approaches can be explored to improve robustness and reduce prediction uncertainty in dynamic housing markets.

CHAPTER 7

7. References

- [1] P. Preethi, D. H. R. Murthy, V. Hiremani, *et al.*, “Optimizing polynomial and regularization techniques for enhanced housing price prediction accuracy,” *SN Computer Science*, vol. 6, art. no. 96, 2025, doi: 10.1007/s42979-024-03578-7.

- [2] K. K. Kusuma, S. R. Pillutla, and A. Saikumar, “Unveiling house price with machine learning algorithm,” in *Proc. 2025 Int. Conf. on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, Erode, India, 2025, pp. 1609–1613, doi: 10.1109/ICMSCI62561.2025.10893972.
- [3] I. Moreno-Foronda, M.-T. Sánchez-Martínez, and M. Pareja-Eastaway, “Comparative analysis of advanced models for predicting housing prices: A review,” *Urban Science*, vol. 9, no. 2, art. no. 32, 2025, doi: 10.3390/urbansci9020032.
- [4] M. A. Azam, S. Rai, and M. S. Raza, “Predictive analytics for housing market trends and valuation,” *Management (Montevideo)*, no. 3, p. 3, 2025.
- [5] L. P. Tran, H. D. Le, T. T. Phuong, and D. C. Nguyen, “Traditional or advanced machine learning approaches: Which one is better for housing price prediction and uncertainty risk reduction?” *Risk Governance & Control: Financial Markets & Institutions*, vol. 15, no. 1, pp. 27–34, 2025.
- [6] H. Fathima, S. Juveria, S. H. Naaz, I. V. Shashikala, and K. D. Babu, “Future proofing real estate: Machine learning for price predictions,” 2026. (*Unpublished / early-access work — update journal or conference details if available.*)
- [7] A. Senadjki, H. N. Au Yong, S. Ogbeibu, C. Y. Yip, M. A. Iddrisu, A. N. B. Wahidudin, and K. H. Woo, “Predictors of housing price cycle and housing bubbles in Malaysia: Resident survey using multinomial logistic regression,” *International Journal of Housing Markets and Analysis*, pp. 1–24, 2026.
- [8] F. Tabe, A. Mansourabady, A. H. Rasekh, and B. Tanoori, “A hybrid deep learning model with sentiment and historical data for cryptocurrency price prediction,” *Expert Systems with Applications*, vol. 302, art. no. 130540, 2026.
- [9] M. Ali, A. Zubair, H. J. Qureshi, W. A. Tanoli, and Z. Masoud, “Novel GEP-based prediction for the cost of green resilient school buildings in hot desert climates,” *Ain Shams Engineering Journal*, vol. 17, no. 1, art. no. 103809, 2026.
- [10] N. Houlié, “The impact of economic policies on housing prices: Approximations and predictions in the UK, the US, France, and Switzerland from the 1980s to today,” *Risks*, vol. 13, no. 5, art. no. 81, 2025.
- [11] A. Chwila, M. Hadaś-Dyduch, M. Krzciuk, T. Stachurski, A. Wolny-Dominiak, and T. Żądło, “Improving ex ante accuracy assessment in predicting house price dispersion: Evidence from the USA,” *arXiv preprint*, arXiv:2502.15905, 2025.
- [12] P. Tian, W. Xiao, and F. Yuan, “Assessing the combinational effects of access to urban amenities on housing prices: A perspective on the ‘15-minute city’,” *Applied Spatial Analysis and Policy*, vol. 18, no. 1, pp. 29–45, 2025.

- [13] Raza, A., Asif, L., Türsoy, T., Seraj, M., & Erkol Bayram, G. (2025). Macro-economic indicators and housing price index in Spain: fresh evidence from FMOLS and DOLS. *International Journal of Housing Markets and Analysis*, 18(1), 227-248.
- [14] H. Jamil, “A spatio-temporal analysis of house prices in Brunei Darussalam,” *Quality & Quantity*, pp. 1–32, 2025.
- [15] P. Adzanoukpe, “Predicting house rental prices in Ghana using machine learning,” *arXiv preprint*, arXiv:2501.06241, 2025.
- [16] Z. Li, S. Van Nieuwerburgh, and R. Wang, “Understanding rationality and disagreement in house price expectations,” *The Review of Financial Studies*, 2025.
- [17] M. A. Shbool, R. Al-Dmour, B. A. Al-Shboul, N. T. Albashabsheh, and N. Almassarwah, “Real estate decision-making: Precision in price prediction through advanced machine learning algorithms,” *International Journal of Housing Markets and Analysis*, 2025.
- [18] M. Singaravelu, R. Singaravelu, A. Raj, K. Muskan, S. Kumar, and M. N. A. Azeez, “Real estate price prediction system using machine learning algorithm,” in *Proc. AIP Conf. Proc.*, vol. 3175, no. 1, art. no. 020047, Mar. 2025.
- [19] R. R. Khasani, “Enhancing real estate price prediction using optimized least squares moment balanced machine,” in *Proc. E3S Web of Conferences*, vol. 605, art. no. 01007, 2025.
- [20] A. Andriansyah, M. D. Dzulkarnain, A. I. Afkarinah, F. Amili, G. Ramadhika, S. N. Ambo, *et al.*, “Utilization of machine learning for property price segmentation and prediction,” *Society: Jurnal Pengabdian Masyarakat*, vol. 4, no. 2, pp. 342–351, 2025.
- [21] G. Dotsis, P. Petris, and D. Psychoyios, “Assessing housing market crashes over the past 150 years,” *The Journal of Real Estate Finance and Economics*, vol. 70, no. 2, pp. 359–377, 2025.
- [22] B. Jin and X. Xu, “Predictions of residential property price indices for China via machine learning models,” *Quality & Quantity*, vol. 59, suppl. 2, pp. 1481–1513, 2025.
- [23] A. Kate, A. Jadhav, S. Patil, T. Lanjewar, and J. Al Dallal, “Data-driven insights in real estate: Property recommendation system and house rent prediction using machine learning,” in *Proc. Int. Conf. on IT Innovation and Knowledge Discovery (ITIKD)*, 2024, pp. 1–6.
- [24] S. Zali, P. Pahlavani, O. Ghorbanzadeh, A. Khazravi, M. Ahmadlou, and S. Givekesh, “Housing price modeling using a geographically, temporally, and characteristically weighted generalized regression neural network (GTCW-GRNN),” *Buildings*, vol. 15, no. 9, art. no. 1405, 2025.

- [25] C. Huang, B. Liang, Z. Li, and F. Chen, “Multimodal machine learning for real estate appraisal: A comprehensive survey,” in *Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, Singapore, 2025, pp. 345–361.
- [26] L. S. D’Acci, “The allometry of housing prices in urban scaling laws and an equalised dwellers’ utility across settlement sizes,” *Environment and Planning B: Urban Analytics and City Science*, 2025.
- [27] A. Mitchell, R. Patel, Z. Zhang, I. Chen, E. Walker, and L. Evans, “AI-based price forecasting for real estate: Trends and challenges,” 2025.

Student Name and Signature: Nivetha B

Date: 29.01.2026

Name and Signature of the Evaluator:

Date: