International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Estimation Residential Real Estate Price Considering Spatial Variables using Machine Learning in GIS: A Case Study in Chon Buri, Thailand

KAMONCHANOK TONLO

Department of Management Science and Engineering, School of Economics and management.
Chongqing university of posts and telecommunications, Chongqing, China.

Abstract - This research investigates the effectiveness of machine learning (ML) techniques in predicting residential real estate prices, with a specific focus on integrating spatial variables within a Geographic Information System (GIS) framework. A case study in Chon Buri, Thailand, is utilized to compare the predictive accuracy of the Forest-based Classification and Regression (FBCR) algorithm, a GIS-specific ML tool, against established algorithms such as XGBoost (XGB) and Random Forest (RF). The methodology includes data collection through web scraping, rigorous preprocessing using the Interquartile Range (IQR) method for outlier detection, and spatial data integration via ArcGIS Pro. Model performance is evaluated using R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The results demonstrate that FBCR outperforms both RF and XGBoost in predicting real estate prices, evidenced by a higher R^2 value of 0.6412, lower RMSE of 1.529, and lower MAE of 1.131 on the testing dataset. This superior performance highlights FBCR's ability to effectively capture and model the complex spatial relationships that influence property prices. The study underscores the potential of GIS-integrated ML tools in enhancing the accuracy and reliability of real estate valuation, providing valuable insights for urban planning and property market analysis.

***Key Words*:** residential real estate, predict price, machine learning, forest-based classification and regression (FBCR), geographic information system (GIS), spatial integration.

1.INTRODUCTION

This The accurate estimation of residential real estate prices is a crucial task in urban planning, property valuation, and investment decision-making. Traditional valuation methods often struggle to incorporate the complex spatial relationships that significantly influence property prices. To address this challenge, researchers have increasingly turned to machine learning (ML) techniques, which can capture non-linear interactions between multiple factors affecting real estate values (Adetunji et al., 2022; Hong & Kim, 2022). In particular, the integration of ML with Geographic Information Systems (GIS) provides a powerful approach to spatially informed price prediction models, leveraging both statistical and spatial variables.

Previous studies have demonstrated the effectiveness of ML algorithms such as XGBoost (XGB) and Random Forest (RF) in predicting housing and land prices (Chen et al., 2024; Dang et al., 2020; Hong et al., 2020). These algorithms have been widely applied in real estate valuation and financial asset pricing, producing highly accurate predictions (Bagnara, 2024; Zhang, 2023). However, limited research has explored the application of GIS-specific ML tools for real estate valuation, particularly the Forest-based Classification and Regression (FBCR) technique, which is embedded within ArcGIS Pro. This technique is designed to handle large spatial datasets and incorporate geographic variables more effectively than traditional ML approaches.

This study aims to evaluate the performance of the FBCR algorithm in predicting residential real estate prices while considering spatial factors. By comparing its predictive accuracy against well-established algorithms such as XGB and RF, this research seeks to provide empirical evidence on whether FBCR can serve as a viable alternative for spatially informed property valuation. The study will analyze the efficiency of FBCR by assessing its predictive accuracy and model robustness, contributing to the growing body of literature on ML-based real estate pricing models. By incorporating spatial data within GIS and leveraging advanced ML techniques, this research will enhance the understanding of how geographic factors influence property prices and offer valuable insights for urban planners, real estate professionals, and policymakers.

Through a comparative analysis, this research will determine whether FBCR, as a GIS-integrated ML tool, can yield results comparable to or exceeding those of commonly used algorithms. The findings will provide a solid foundation for further exploration of GIS-driven ML methodologies in real estate price estimation, potentially leading to more accurate and spatially contextualized valuation models.

2. RELATED WORK

Machine learning techniques have significantly enhanced real estate price prediction by capturing complex, non-linear relationships among economic, spatial, and environmental factors. Adetunji et al. (2022) demonstrated the effectiveness of Random Forest (RF) in predicting housing prices, highlighting its ability to manage diverse input variables. Meanwhile, XGBoost, an advanced gradient boosting algorithm introduced by Chen and Guestrin (2016), has gained prominence due to its superior predictive performance, particularly in structured datasets. El Mouna et al. (2023) compared multiple machine learning models and found that both RF and XGBoost outperformed traditional regression-based approaches in terms of accuracy and robustness.

Existing research highlights the critical role of spatial and locational factors in determining residential real estate prices. Liang et al. (2018) found that proximity to amenities such as schools, parks, and shopping centers positively influences property values, increasing their attractiveness. The integration of Geographic Information Systems (GIS) with machine

learning has further improved predictive modeling in real estate. Dang et al. (2020) and Ma et al. (2020) demonstrated how spatial data enhances property valuation models, allowing for more precise estimations of land values. Hong et al. (2020) and Hong & Kim (2022) applied ensemble learning techniques, including RF and XGBoost, for mass appraisal of residential properties, confirming their capability to process large, heterogeneous datasets. Furthermore, Bagnara (2024) emphasized the adaptability of machine learning models in asset pricing, reinforcing their potential in dynamic and evolving real estate markets.

The Forest-Based Classification and Regression (FBCR) tool has emerged as a powerful geospatial machine learning technique for predictive modeling in GIS-based applications. As an ensemble-based approach, FBCR integrates two widely recognized algorithms: the Random Forest (RF) algorithm, developed by Breiman (2001), and the Extreme Gradient Boosting (XGBoost) algorithm, introduced by Chen and Guestrin (2016). These algorithms enhance predictive accuracy by capturing complex, non-linear relationships within large datasets. FBCR is particularly well-suited for real estate price prediction, as it incorporates diverse spatial, environmental, and economic variables to model housing values with high precision.

One of the key advantages of FBCR is its seamless integration with Geographic Information Systems (GIS). By incorporating spatial proximity measures, raster datasets, and distance-based explanatory variables, FBCR enhances predictive modeling in real estate analysis. Prior studies (Dang et al., 2020; Ma et al., 2020) have demonstrated the effectiveness of spatially-aware machine learning in urban land value assessments, supporting the notion that GIS-based FBCR can generate more nuanced and spatially contextualized price predictions.

3. METHODOLOGY

3.1 RESEARCH DESIGN

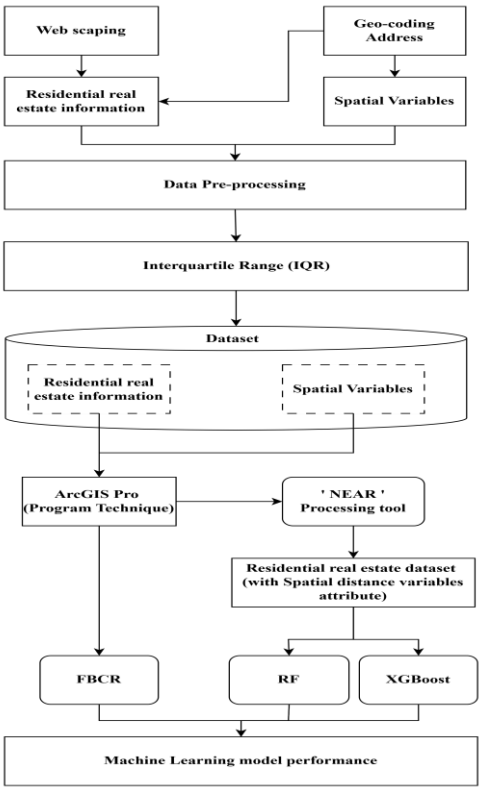


Fig -1. Research framework of this study

This study follows a structured approach to collecting, preprocessing, and analyzing residential real estate data, integrating spatial attributes to enhance predictive accuracy. Real estate data is obtained through web scraping, extracting property details such as price, number of bedrooms, and floor area. Spatial variables, including road networks, public transportation, and green spaces, are gathered from open sources. Geocoding services, such as Google Maps API, convert property addresses into geographic coordinates, enabling spatial analysis. Data preprocessing ensures consistency by standardizing formats, removing duplicates, and handling missing values. Outliers are detected using the Interquartile Range (IQR) method, where values beyond lower bound or upper bound are considered extreme and adjusted or removed to improve model reliability.

To incorporate spatial attributes, ArcGIS Pro and the *'NEAR' Processing Tool* calculate proximity to key amenities such as parks, transit stations, and major roads, enriching the dataset. The refined data is then used to train three machine learning models: Forest-Based Classification and Regression (FBCR), Random Forest (RF), and XGBoost. FBCR, integrated with ArcGIS Pro, is tailored for spatial data, while RF and XGBoost provide robust predictive capabilities.

Model performance is evaluated using R^2, MAE, MSE, and RMSE to compare accuracy. By integrating real estate and spatial data, applying IQR for outlier detection, and leveraging machine learning techniques, this study enhances real estate price prediction, providing valuable insights into the factors influencing property values.

3.2 STUDY AREA

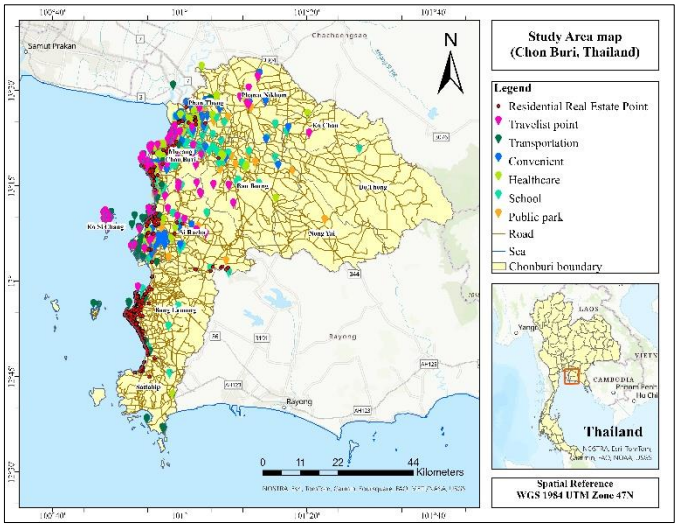


Fig -2. Study area mapping

The study area for this research is Chon Buri, Thailand, a province located southeast of Bangkok. Chon Buri was specifically selected due to its diverse urban and suburban characteristics, making it an ideal setting for examining the spatial factors influencing residential real estate prices. The region is characterized by a mix of residential, commercial, and industrial zones, providing a complex interplay of variables that can affect property values. Additionally, Chon Buri's coastal location and proximity to the Gulf of Thailand introduce unique spatial elements, such as the influence of coastal amenities and potential environmental factors on real estate prices. The geographical diversity and dynamic real estate market of Chon Buri offer a robust foundation for

applying and evaluating machine learning models in the context of spatial data integration.

3.3 DATA COLLECTION & PRE-PROCESSING

3.3.1 Residential real estate information

The collection of housing data in this study is conducted through web scraping, an automated technique for extracting information from real estate listing websites. To streamline this process, the Instant Data Scraper tool is employed, enabling efficient extraction of key property attributes, including price, number of bedrooms and bathrooms, usable floor area, and price per square meter. These variables serve as essential explanatory features for real estate price prediction. The extracted data is stored in structured formats such as CSV files or relational databases, ensuring accessibility for subsequent analysis. To maintain data quality and consistency, rigorous preprocessing is applied. First, standardization of formats is conducted, wherein price values are converted to a common currency and normalized per square meter, while categorical attributes are reformatted into consistent numerical representations. Second, duplicate detection and removal is performed to eliminate redundant property listings, preventing skewed analyses caused by multiple postings of the same property. Third, handling of missing values is addressed through appropriate imputation techniques or the exclusion of incomplete records, depending on the context.

Name_Eng	Name_TH	City	State_zipcode	Country	Zip_code	Address	bed	bath	floor_area	price	price_m
A.D. Bungaray		Sattahip	Chonburi 20180	Thailand	20180	A.D. Bungaray ,Sattahip,Chonburi 20180,Thailand	1	2	50	1850000	1.85
A.D. Bungaray		Sattahip	Chonburi 20180	Thailand	20180	A.D. Bungaray ,Sattahip,Chonburi 20180,Thailand	1	1	25	1690000	1.69
A.D. Bungaray		Sattahip	Chonburi 20180	Thailand	20180	A.D. Bungaray ,Sattahip,Chonburi 20180,Thailand	2	2	112	3300000	3.3
Acqua Condo		Bang Lamung	Chonburi 20150	Thailand	20150	Acqua Condo ,Bang Lamung,Chonburi 20150,Thailand	1	1	35	2290000	2.29
Acqua Condo		Bang Lamung	Chonburi 20150	Thailand	20150	Acqua Condo ,Bang Lamung,Chonburi 20150,Thailand	1	1	47	3500000	3.5
Acqua Condo		Bang Lamung	Chonburi 20150	Thailand	20150	Acqua Condo ,Bang Lamung,Chonburi 20150,Thailand	2	1	52	4690000	4.69
AD Hyatt Condominium		Bang Lamung	Chonburi 20150	Thailand	20150	AD Hyatt Condominium ,Bang Lamung,Chonburi 20150,Thailand	1	2	72	4900000	4.9
AD Hyatt Condominium		Bang Lamung	Chonburi 20150	Thailand	20150	AD Hyatt Condominium ,Bang Lamung,Chonburi 20150,Thailand	1	2	44	2680000	2.68
AD Hyatt Condominium		Bang Lamung	Chonburi 20150	Thailand	20150	AD Hyatt Condominium ,Bang Lamung,Chonburi 20150,Thailand	1	1	28	1900000	1.9
Albar Peninsula		Sattahip	Chonburi 20180	Thailand	20180	Albar Peninsula ,Sattahip,Chonburi 20180,Thailand	1	1	53	2500000	2.5
Amata Miracle Chonburi		Muang Chonburi	Chonburi 20000	Thailand	20000	Amata Miracle Chonburi ,Muang Chonburi,Chonburi 20000,Thailand	2	1	65	1900000	1.90
Amata Miracle Chonburi		Muang Chonburi	Chonburi 20000	Thailand	20000	Amata Miracle Chonburi ,Muang Chonburi,Chonburi 20000,Thailand	1	1	33	990000	0.99
Amata Miracle Chonburi		Muang Chonburi	Chonburi 20000	Thailand	20000	Amata Miracle Chonburi ,Muang Chonburi,Chonburi 20000,Thailand	1	1	32	850000	0.85
Anacon Residence Puttaya		Bang Lamung	Chonburi 20150	Thailand	20150	Anacon Residence Puttaya ,Bang Lamung,Chonburi 20150,Thailand	1	1	35	2100000	2.1
Anacon Residence Puttaya		Bang Lamung	Chonburi 20150	Thailand	20150	Anacon Residence Puttaya ,Bang Lamung,Chonburi 20150,Thailand	2	2	72	3699000	3.699
Anacon Residence Puttaya		Bang Lamung	Chonburi 20150	Thailand	20150	Anacon Residence Puttaya ,Bang Lamung,Chonburi 20150,Thailand	1	1	35	1550000	1.55
Anacon Residence Puttaya		Bang Lamung	Chonburi 20150	Thailand	20150	Anacon Residence Puttaya ,Bang Lamung,Chonburi 20150,Thailand	1	1	37	2369000	2.369
Anacon Residence Puttaya		Bang Lamung	Chonburi 20150	Thailand	20150	Anacon Residence Puttaya ,Bang Lamung,Chonburi 20150,Thailand	2	2	75	4550000	4.55
Anacon Residence Puttaya		Bang Lamung	Chonburi 20150	Thailand	20150	Anacon Residence Puttaya ,Bang Lamung,Chonburi 20150,Thailand	1	1	35	1690000	1.69
Anacon Residence Puttaya		Bang Lamung	Chonburi 20150	Thailand	20150	Anacon Residence Puttaya ,Bang Lamung,Chonburi 20150,Thailand	1	1	35	1850000	1.85
Anacon Residence Puttaya		Bang Lamung	Chonburi 20150	Thailand	20150	Anacon Residence Puttaya ,Bang Lamung,Chonburi 20150,Thailand	1	1	35	1790000	1.79
Anson Village Place Condo		Si Racha	Chonburi 20110	Thailand	20110	Anson Village Place Condo ,Si Racha,Chonburi 20110,Thailand	1	1	30	1300000	1.3
Anson Village Place Condo		Si Racha	Chonburi 20110	Thailand	20110	Anson Village Place Condo ,Si Racha,Chonburi 20110,Thailand	1	1	30	1100000	1.1
Anson Village Place Condo		Si Racha	Chonburi 20110	Thailand	20110	Anson Village Place Condo ,Si Racha,Chonburi 20110,Thailand	1	1	38	1800000	1.8

Fig -3. Example of residential real estate dataset rows

These preprocessing steps are critical in creating a clean and reliable dataset, directly influencing the accuracy and interpretability of predictive models. By ensuring data consistency, completeness, and standardization, the study enhances the robustness of subsequent analyses, contributing to a more precise and meaningful real estate price prediction model.

3.3.2 Spatial variables

Spatial data is an essential component in enhancing the housing dataset by incorporating geographic context, which significantly influences real estate prices. This study utilizes open-source spatial data from Thailand, including road networks, public transportation lines, green spaces, and neighborhood boundaries, to capture locational attributes that impact residential desirability. To integrate spatial information, geocoding services from Google Maps are employed to convert property addresses into geographic coordinates, enabling precise mapping and spatial analysis. Following data collection, spatial attributes are integrated with housing data through spatial joins, where properties are enriched with features such as proximity to public parks, schools, transit stations, and major roads. Additional spatial variables, such as the density of nearby amenities, are also included to reflect accessibility and neighborhood characteristics. To ensure spatial accuracy, all layers are aligned to a common coordinate reference system (CRS), and any discrepancies in geocoded addresses or spatial features are rigorously validated.

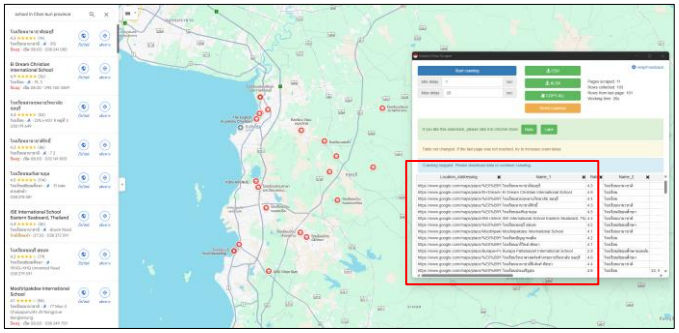


Fig -4. Example of collected spatial data from scraping technique

Spatial preprocessing further includes feature extraction, where distance-based attributes are computed using geospatial analysis tools and normalized for consistency. This integration of spatial variables provides a comprehensive representation of location-specific factors, allowing for a more robust analysis of their influence on property values. By incorporating spatial data, this study enhances the predictive capability of real estate price models, offering valuable insights into the role of geographic factors in housing markets.

3.3.3 The Interquartile Range (IQR)

The Interquartile Range (IQR) is a crucial statistical measure used in data preprocessing to detect and handle outliers before training a machine learning model. Outliers, which are extreme values significantly deviating from the rest of the dataset, can distort model training, leading to biased predictions and reduced accuracy. The IQR method is particularly effective in real estate price prediction, where anomalies in property prices, sizes, or other attributes may arise due to errors, rare market conditions, or data inconsistencies. By applying IQR, the dataset can be refined to ensure that the model learns from relevant and representative data points, thereby improving both predictive performance and generalizability. The IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1), representing the middle 50% of the data. Mathematically, it is expressed as:

$$IQR = Q3 - Q1 \qquad (1)$$

To identify outliers, data points that fall below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ are considered extreme values and may be removed or adjusted depending on the modeling approach. This technique ensures that the machine learning model is not disproportionately influenced by extreme values that do not reflect general market trends. By implementing IQR-based outlier detection, the dataset remains balanced, reducing noise and enhancing the model's ability to generalize well to unseen data. This step is essential in predictive modeling, as it prevents overfitting to irregularities and contributes to more reliable and interpretable real estate price estimations.

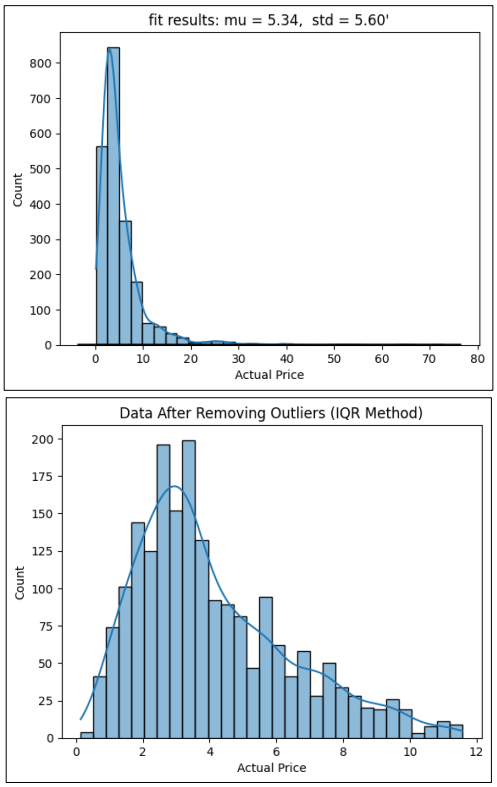


Fig -5. Dataset before IQR process (top-2,159 rows) and dataset after IQR process (bottom-1,987 rows)

Table-1. Dataset Description

Features	Data Type	Description
Bed	Int64 (Numeric)	Total bedrooms of the area
Bath	Int64 (Numeric)	Total bathrooms of the area
Floor area	Int64 (Numeric)	Size of usable area (Floor space)
Public Park	Point / Distance (Float64, Numeric)	Address point of Public park
Healthcare	Point / Distance (Float64, Numeric)	Address point of Healthcare
Convenient	Point / Distance (Float64, Numeric)	Address point of Convenient
Travelist Point	Point / Distance (Float64, Numeric)	Address point of Travelist Point
Education	Point / Distance (Float64, Numeric)	Address point of Education
Transportation	Point / Distance (Float64, Numeric)	Address point of Transportation
Road	Polyline Shapefiles	Location of Road
Sea	Polygon Shapefiles	Location of Sea

3.4 CONCEPTUAL FRAMEWORK

This study employed a rigorous machine learning framework to develop a predictive model. The process commenced with the acquisition of a dataset comprising 2,159 rows. To ensure the robustness and reliability of the model, a preprocessing step was undertaken to address potential outliers. An Interquartile Range (IQR) analysis was conducted, resulting in the identification and removal of 172 rows that fell outside the defined lower and upper bounds. This outlier removal process yielded a refined dataset of 1,987 rows, which served as the input for subsequent model development.

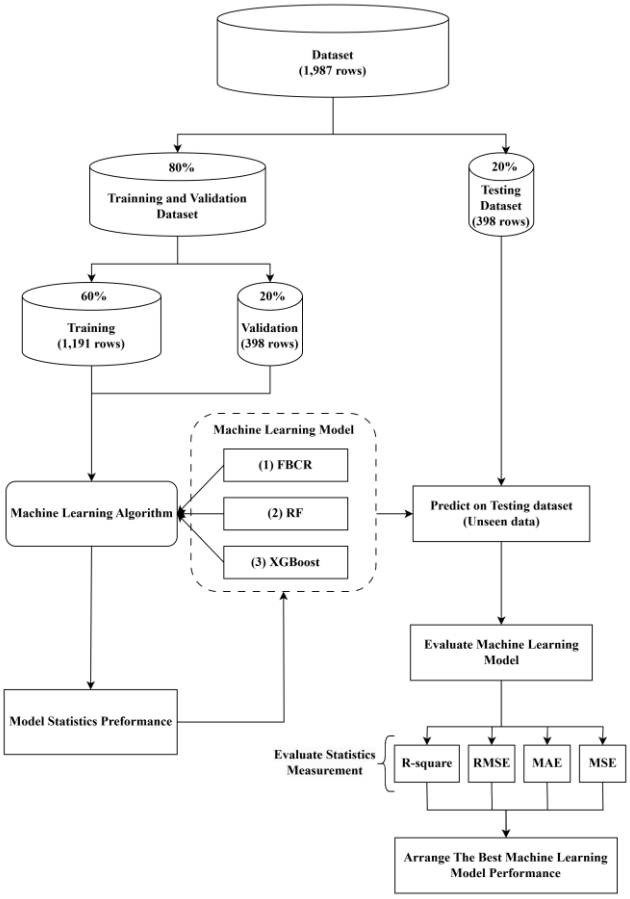


Fig -6. Conceptual of model processing

The refined dataset was then partitioned into training, validation, and testing subsets. Eighty percent of the data was allocated for training and validation, while the remaining twenty percent (398 rows) was reserved for testing, ensuring the model's ability to generalize to unseen data. The training and validation set was further divided, with sixty percent (1,191 rows) used for model training and twenty percent (398 rows) for hyperparameter tuning and performance validation.

Three machine learning algorithms were selected for model development: Forest-based Classification and Regression (FBCR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Each algorithm was trained on the designated training dataset, learning the underlying patterns and relationships within the data. Subsequently, the trained models were evaluated on the testing dataset using established performance metrics, including R-squared, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE). These metrics provided a comprehensive assessment of each model's predictive accuracy and generalization capability.

The model demonstrating superior performance based on these metrics was selected as the most suitable for the given task. This systematic approach, encompassing outlier removal, data partitioning, model training, validation, and rigorous

evaluation, ensures the development of a robust and reliable machine learning model capable of accurate prediction and effective application in real-world scenarios.

3.5 MACHINE LEARNING

3.5.1 Forest-based Classification and Regression

The Forest-based Classification and Regression tool uses an adaptation of Leo Breiman's random forest algorithm to create models and generate predictions of both categorical and continuous variables.

The random forest algorithm works by creating decision trees. They Create a set of rules for predicting a feature's category or value based on its attributes. The random forest algorithm so powerful is that it creates an ensemble of many decision trees, hence the name forest and the reason it is called a random forest is because each tree is trained using only a random subset of the training data and a random subset of the explanatory variables. Each tree does its best to predict with the random subset of data and variables it was given. (But in the end, following the majority vote wins) Any individual tree on its own is not a strong predictor because it is prone to overfitting to the training data. Overfitting happens when the model mimics the training dataset to closely instead of generalizing a trend, making it so that the model can only predict the data it was trained with. Training each tree with random subset od data and variables and using the entire forest to generate a final prediction rather than any single tree helps to prevent overfitting to the training data.

Creating a generalized model is crucial in being able to predict values of new features that were not used to train the model. One importance way to evaluate model performance is by using the model to predict values for features that were not included in the training dataset. By default, the Forest-based Classification and regression tool holds back 10 percent of the data for validation but this time the dataset were divided into 60 precent of dataset to be training, 20 percent to be validation after trained a model, it can check how well it predicts the feature that were held back from training. And another 20 percent of dataset is used for testing to evaluate model efficiency that how good is this machine learning model truly compare with others model.

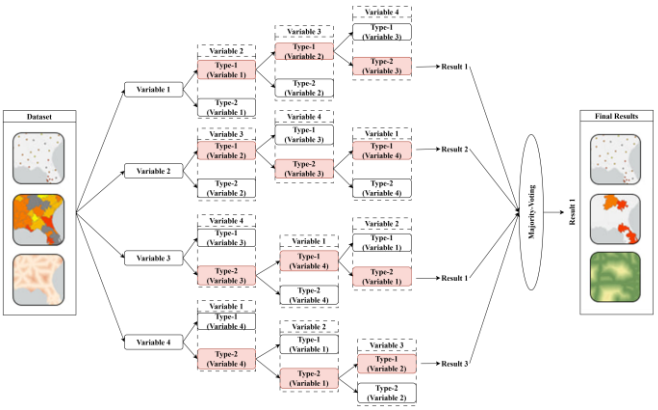


Fig -7. Forest-based Classification and Regression (figure credit, original)

In this study, the random forest model is configured with the following setting: Explanatory Training Variables: *'bed', 'bath', 'floor area'*. Explanatory Training Distance Features: *'public park', 'healthcare', 'convenient, 'travelist point', 'education', 'transportation', 'road',* and *'sea'*. For advances forest options setting: *number of trees='300':* The number of trees to create in the forest model. More trees will generally result in more accurate model prediction, but the model will take longer to calculate (default number of trees is 100). *data available per tree (%) ='100':* Specifies the percentage of the Input Training Features used for each decision tree. The default is 100 percent of the data. Samples for each tree are taken randomly from two-thirds of the data specified. Each decision tree in the forest is created using a random sample or subset (approximately two-thirds) of the training data available. Using a lower percentage of the input data for each decision tree increases the speed of the tool for very large datasets. *number of runs for validation='5':* The tool will run for the number of iterations specified. The distribution of the R2 for each run can be displayed using the Output Validation Table parameter. When this is set and predictions are being generated, only the model that produced the highest R2 value will be used for predictions.

3.5.2 Random Forest

Random Forest (RF) is a robust ensemble learning method that employs multiple decision trees to improve predictive accuracy and reduce overfitting. In this study, RF is utilized for predicting house prices. The algorithm works by training numerous decision trees, each built on a randomly sampled subset of the data, and aggregating their predictions. This approach mitigates the overfitting risk commonly associated with individual decision trees and enhances model generalization. The random forest algorithm introduces randomness in two primary ways: through bootstrapping (random sampling with replacement) to generate diverse training subsets for each tree, and by selecting a random subset of features at each split. These mechanisms ensure that the individual trees are decorrelated, which improves the overall model's performance and robustness. Random forest's flexibility allows it to handle large datasets with numerous features, making it particularly effective for high-dimensional regression tasks, such as real estate price prediction.

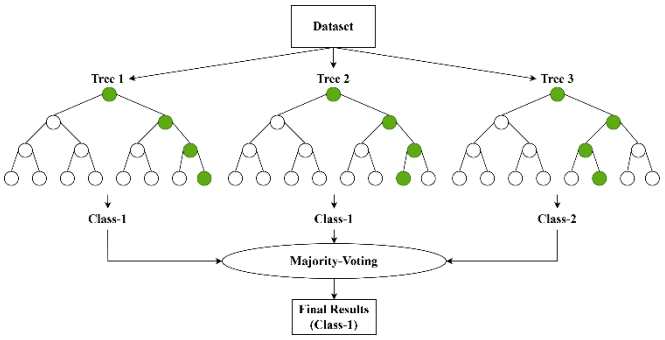


Fig -8. Random Forest Algorithm (figure credit, original)

In this study, the random forest model is configured with the following hyperparameters: *n_estimators='300':* The ensemble consists of 300 decision trees, balancing model accuracy and computational efficiency. *criterion='squared_error':* The mean squared error (MSE) is used to evaluate potential splits during tree construction, aiming to minimize the squared difference between predicted and actual values. *max_depth='6':* The maximum depth of each tree is set to 6, controlling the complexity of the trees and ensuring that the model captures essential patterns without overfitting. *min_samples_split='0.1':* This parameter ensures that nodes are split only when a significant number of samples are present, reducing the likelihood of overfitting to noise in

the data. *min_impurity_decrease='0.01':* A minimum threshold for impurity reduction is set to 0.01, meaning splits are only made if they lead to a meaningful decrease in the impurity measure, further limiting unnecessary complexity.

All remaining parameters adhere to the default values in the scikit-learn implementation, with the model being run using Python and the scikit-learn library. These settings were chosen to achieve a balance between predictive performance, model interpretability, and computational efficiency, with the goal of providing a stable and accurate model for predicting real estate prices. Furthermore, default configurations were used for feature selection and handling missing values, ensuring consistency across varying data distributions.

3.5.3 Extreme Gradient Boosting

XGBoost (Extreme Gradient Boosting) is an advanced and highly efficient gradient boosting algorithm used to improve predictive accuracy, particularly in structured data tasks. In this study, XGBoost is employed to predict house prices. Unlike Random Forest, where decision trees are constructed independently, XGBoost builds trees sequentially. Each tree is trained to correct the errors of the previous tree, progressively improving the model's accuracy through an iterative process known as gradient boosting.

XGBoost uses a combination of decision trees and gradient descent to minimize the objective function, which includes both the loss function and regularization terms to prevent overfitting. The model's sequential nature enables it to learn complex feature interactions effectively. Additionally, XGBoost supports dynamic handling of missing values and is known for its computational efficiency, making it highly suitable for large datasets with intricate relationships, such as real estate price prediction.

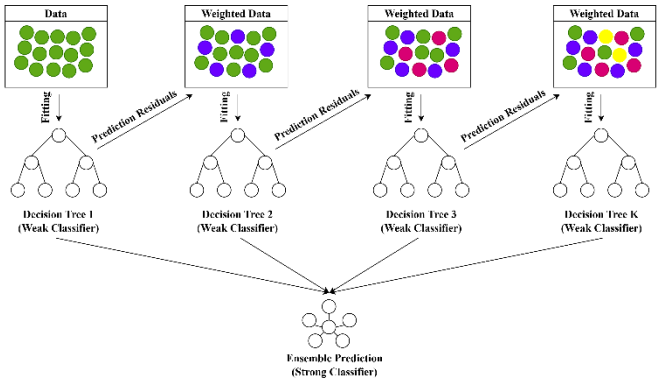


Fig -9. Extreme Gradient Boosting Algorithm
(figure credit, original)

In this study, the XGBoost model is configured with the following hyperparameters: *n_estimators=300:* The model consists of 300 boosting rounds, providing a sufficient number of base learners while controlling the model's complexity. *objective='reg:squarederror':* The objective function is set to squared error, optimizing the model for regression tasks by minimizing the mean squared error between the predicted and actual values. *max_depth=6:* The maximum depth of each tree is set to 6, limiting the complexity of the individual trees and ensuring the model captures key patterns in the data without overfitting. *learning_rate=0.1:* The learning rate determines the contribution of each individual tree to the final prediction. A learning rate of 0.1 strikes a balance between model stability and convergence speed, allowing for gradual learning and enhancing generalization.

All remaining parameters follow the default values in the xgboost library, with the model implemented and executed using Python. These hyperparameter settings were chosen to optimize the trade-off between computational efficiency and predictive performance. By applying these settings, the XGBoost model can effectively capture complex relationships in the data while maintaining robustness and preventing overfitting. Default configurations were used for additional parameters, ensuring consistency and efficiency in model training and evaluation.

3.6 EVALUATION METRICS

The evaluation of machine learning models is fundamental for determining their predictive accuracy and generalizability. Various statistical metrics are employed in regression tasks to assess model efficacy, including the coefficient of determination (R^2), root mean square error (RMSE), mean square error (MSE), and mean absolute error (MAE). These metrics collectively provide insights into the model's explanatory power, error magnitude, and overall predictive reliability.

3.6.1 Coefficient of Determination (R^2) The coefficient of determination (R2R^2R2) quantifies the proportion of variance in the dependent variable explained by the independent variables, serving as an indicator of model fit:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (2)$$

where y_i represents observed values, $\hat{y}_i$ denotes predicted values, y is the mean of observed values, and n is the number of observations. An R^2 value approaching 1 indicates a stronger model fit, signifying that a higher proportion of variance is explained. In real estate price prediction, R^2 is widely used to compare the performance of machine learning algorithms, such as Random Forest and XGBoost, in modeling housing market fluctuations.

3.6.2 Root Mean Square Error (RMSE) The root mean square error (RMSE) measures the average deviation between predicted and observed values, with a higher penalty assigned to larger errors:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (3)$$

As RMSE retains the same unit as the dependent variable, it provides an interpretable measure of prediction accuracy. Due to its sensitivity to large deviations, RMSE is particularly valuable in applications where minimizing substantial prediction errors is essential, such as property valuation and mass real estate appraisal.

3.6.3 Mean Square Error (MSE) The mean square error (MSE) quantifies the average squared difference between observed and predicted values:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (4)$$

As the squared counterpart of RMSE, MSE amplifies the impact of larger errors, making it a crucial metric for optimizing machine learning models. However, its lack of direct interpretability in the original measurement units limits its standalone use in practical applications. Many predictive algorithms, including regression-based and gradient boosting methods, minimize MSE during training to improve model performance.

3.6.4 Mean Absolute Error (MAE) The mean absolute error (MAE) computes the average absolute difference between predicted and actual values, offering a robust measure of model error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (5)$$

Unlike RMSE and MSE, MAE does not disproportionately penalize larger errors, making it more robust to outliers. This characteristic renders MAE particularly advantageous in cases where a balanced evaluation of prediction accuracy is required. Furthermore, due to its direct interpretability, MAE is often preferred when the goal is to obtain an intuitive measure of model error without overemphasizing extreme deviations.

These evaluation metrics provide complementary insights into model performance. While R^2 assesses the explanatory power of a model, RMSE and MSE emphasize error magnitude, with RMSE offering better interpretability. MAE, in contrast, provides a robust alternative by treating all deviations uniformly.

3.7 RESULTS DISCUSSION

3.7.1 Pearson Correlation Heatmap

The Pearson correlation heatmap reveals intriguing relationships between various factors influencing property prices. As expected, Actual Price exhibits a strong positive correlation with property size, as indicated by the moderate to strong positive correlations with Bed (0.50), Bath (0.50), and Floor Area (0.56). This suggests that larger properties with more bedrooms and bathrooms command higher prices. Additionally, Actual Price shows a moderate positive correlation with Road (0.28), implying that proximity to roads might increase property values, potentially due to enhanced accessibility.

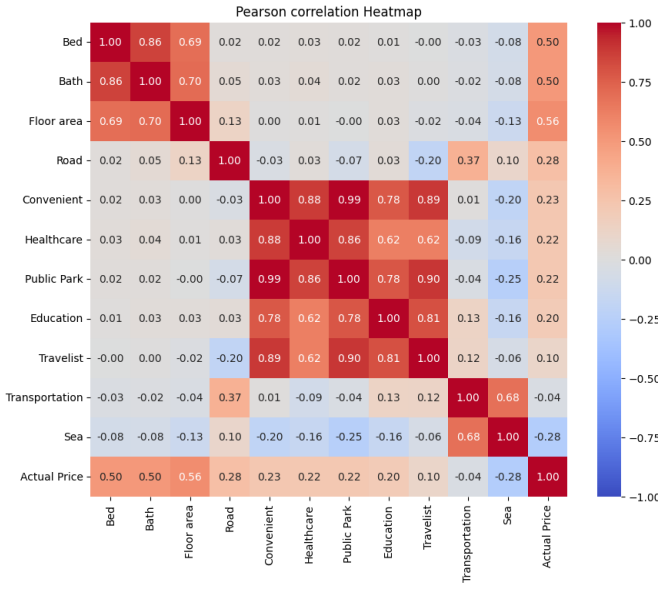


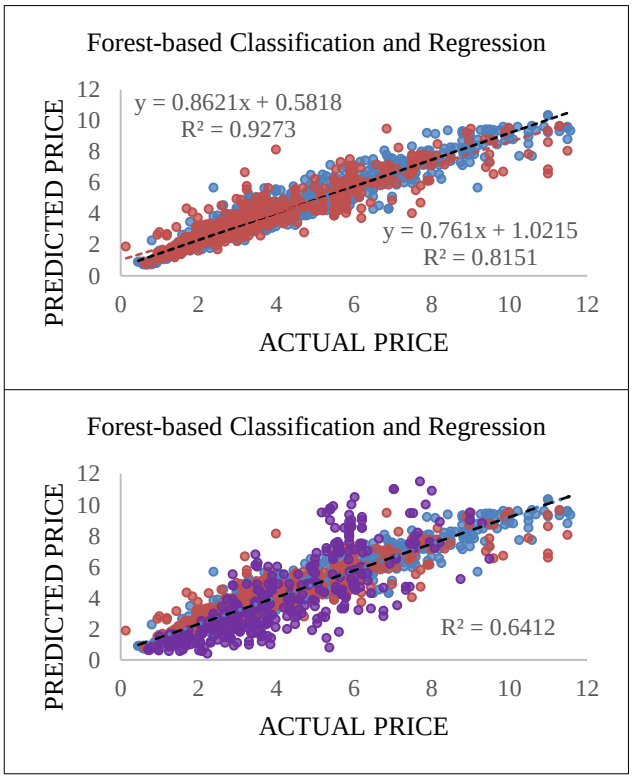
Fig -10. Pearson's Correlation Heatmap

While the influence of amenities is less pronounced, Actual Price still shows weak positive correlations with Convenient (0.23), Healthcare (0.20), Public Park (0.22), and Education (0.21), indicating that proximity to these amenities has a small positive impact on property prices. Interestingly, there is a weak negative correlation between Actual Price and Sea (-0.28), suggesting that properties closer to the sea might be valued slightly lower in this particular context. This unexpected relationship warrants further investigation to understand the underlying market dynamics.

Beyond price, the heatmap highlights a strong tendency for amenities to cluster together, as evidenced by the high positive correlations among Convenient, Healthcare, Public Park, Education, and Travelist point. This pattern likely reflects urban planning practices. Additionally, a strong positive correlation exists between Bed and Bath (0.86), and moderate positive correlations are observed between Floor Area and both Bed (0.69) and Bath (0.70), reflecting the common trend of larger properties having more bedrooms and bathrooms. A notable correlation is the moderate positive association between Transportation and Sea (0.68), potentially indicating improved transportation infrastructure in coastal areas. These findings provide valuable insights into the interplay of various factors affecting property values and urban development patterns.

3.7.2 Dataset's Scatter-plot graph

This comparative analysis of FBCR, RF, and XGB models underscores a crucial principle in machine learning: the paramount importance of evaluating model performance on unseen data. While achieving a high degree of accuracy on the training data is desirable, it can be misleading if it comes at the expense of generalization capability. This is clearly demonstrated by the XGB model, which, despite exhibiting exceptional performance on the training data with an R-squared of 0.9660, reveals a vulnerability to overfitting when confronted with the testing dataset. This overfitting tendency is evidenced by the substantial drop in R-squared to 0.5513 on the testing data and the increased dispersion of data points, indicating a significant reduction in predictive accuracy.

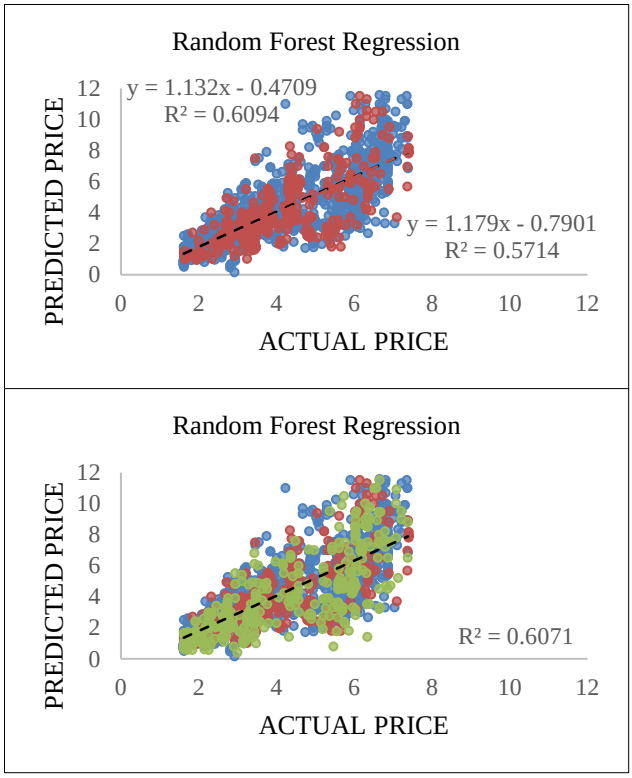


Graph-1. Showing scatter-plot graph of training (blue), validate (red), testing (purple) dataset of FBCR model

In contrast, the FBCR model presents a more balanced performance profile. Although its training performance (R-squared of 0.9273) is slightly lower than XGB, it maintains a more consistent accuracy across the validation and testing datasets, achieving R-squared values of 0.8151 and 0.6412,

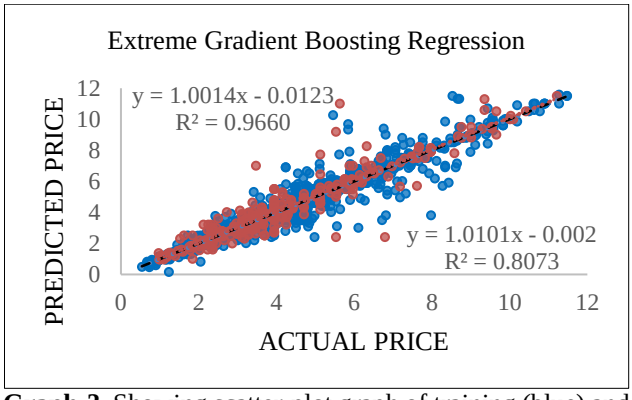
respectively. This relative stability suggests that FBCR has learned more generalizable patterns from the data, making it potentially more reliable for real-world applications where encountering unseen data is inevitable.

The RF model, with its consistently lower performance across all datasets (training R-squared of 0.6094, validation R-squared of 0.5714, and testing R-squared of 0.6071), indicates a limited capacity to capture the complexities inherent in the data. This suggests that the RF model, with its current configuration, may not be the most suitable choice for this specific prediction task.

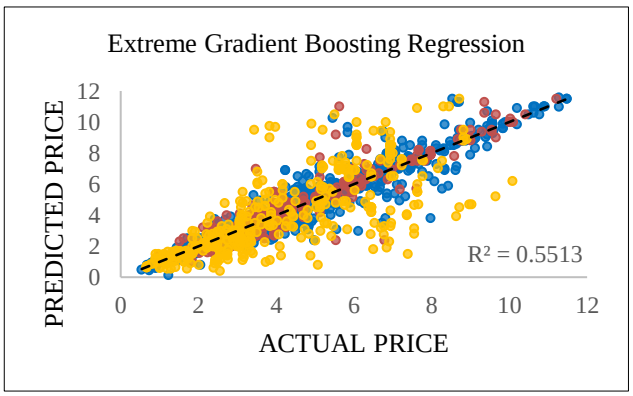


Graph-2. Showing scatter-plot graph of training (blue), validate (red), testing (green) dataset of RF model

The analysis underscores the importance of evaluating model performance on unseen data. While XGB excels in fitting the training data, its performance on the testing dataset indicates a susceptibility to overfitting. FBCR demonstrates a more balanced performance across all datasets, suggesting better generalization. RF, on the other hand, exhibits consistently lower performance, indicating limited ability to capture the data's complexity.



Graph-3. Showing scatter-plot graph of training (blue) and validate (red) dataset of XGB model

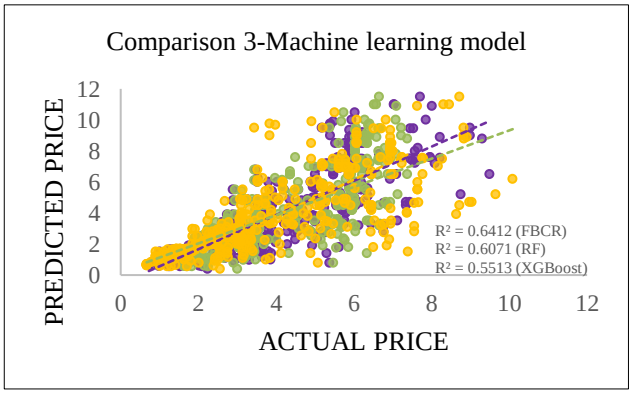


Graph-4. Showing scatter-plot graph of testing (yellow) dataset of XGB model

The testing dataset results are pivotal in determining the models' practical applicability. The significant drop in XGB's performance, coupled with the increased data point dispersion, highlights the challenges of deploying a model that overfits. FBCR's relatively stable performance across all datasets suggests it might be a more reliable choice for real-world applications. RF's consistently lower performance indicates it may not be suitable for this particular prediction task. while XGB demonstrates superior performance on training and validation data, FBCR exhibits better generalization to unseen data. This evidence highlights the necessity of rigorous testing and validation to ensure model robustness and practical utility.

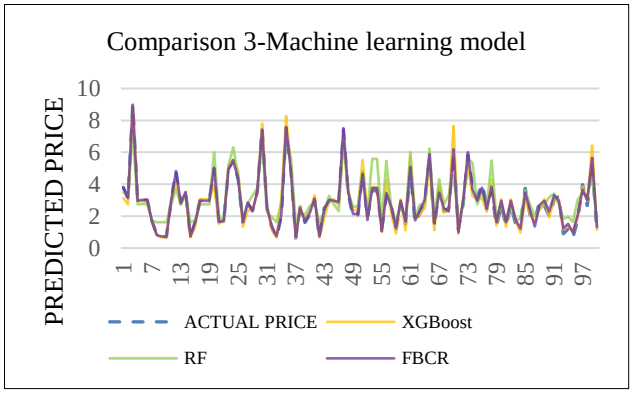
3.7.3 Comparison 3-model's results

This study rigorously evaluated the predictive capabilities of three machine learning models – Forest-based Classification and Regression (FBCR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) on a testing dataset comprising 398 rows of unseen data. The analysis focused on assessing the model's ability to generalize and accurately predict property prices in real-world scenarios. A scatter plot visually illustrates the performance of each model, with FBCR demonstrating superior predictive accuracy, achieving a higher R-squared value of 0.6412 and surpassing both RF (0.6071) and XGBoost (0.5513). This indicates that FBCR more effectively captures the underlying patterns in unseen data. The distribution of data points further reinforces this observation, with FBCR's predictions showing a tighter clustering around the implied trend line, suggesting a more consistent and reliable fit.



Graph-5. Showing scatter-plot graph to compared difference machine learning algorithm on all testing dataset (unseen data) 398 rows, For color as FBCR (purple), RF (green), and XGBoost (yellow)

A 2D line graph, comparing the predicted prices from each model against the actual prices for a sample of 100 data points, provides a more granular perspective. The FBCR model's predictions closely track the actual prices, highlighting its ability to accurately reflect real-world price fluctuations. In contrast, the XGBoost model exhibits significant volatility, with substantial deviations from the actual price line, indicating fewer stable predictions. These findings underscore the FBCR model's exceptional ability to generalize to unseen data, positioning it as a robust and reliable tool for property price prediction.



Graph-6. Showing 2D-line graph to compared difference machine learning algorithm on testing dataset (unseen data) sample 100 rows from 398 rows dataset. For color as FBCR (purple), RF (green), XGBoost (yellow), and actual price (blue)

3.7.4 Table statistics

Table-2 provides a detailed comparative analysis of the performance of the FBCR, RF, and XGBoost models, revealing critical insights into their predictive capabilities and generalization potential. Notably, the R-squared values underscore the FBCR model's superior performance on the testing dataset, which represents unseen data. Achieving an R-squared of 0.641, FBCR outperforms both RF (0.607) and XGBoost (0.551), demonstrating its enhanced ability to accurately predict property prices in real-world scenarios.

Table-2. Summary model statistics

Model	Training	Validation	Testing	Mean accuracy
FBCR	0.927	0.815	0.641	59.95%
RF	0.609	0.571	0.607	52.51%
XGBoost	0.966	0.807	0.551	59.49%
Model	**MAE**	**MSE**	**RMSE**	**R-squared**
FBCR	1.131	2.339	1.529	0.641
RF	1.237	2.559	1.599	0.607
XGBoost	1.187	2.925	1.710	0.551

This finding is further corroborated by the RMSE and MAE metrics. FBCR exhibits the lowest RMSE (1.529) and MAE (1.131) on the testing data, indicating that its predictions are not only more accurate but also exhibit lower average errors compared to the other models. In contrast, XGBoost, despite its exceptional performance on the training data, suffers from significant overfitting, as evidenced by the substantial decline in R-squared on the testing data and the highest RMSE and MAE values. The RF model consistently underperforms across all metrics. The mean accuracy values also reinforce the FBCR model's effectiveness, with FBCR achieving a mean accuracy of 59.95%, very close to XGBoost 59.49% but FBCR is better in unseen data. This consistent performance across various metrics highlights FBCR's robustness and reliability, making it a valuable tool for property price prediction.

These results emphasize the importance of evaluating model performance on unseen data to assess real-world applicability. The FBCR model's ability to maintain a high level of accuracy on the testing dataset, coupled with its lower error rates, positions it as a more dependable and effective model for property price prediction compared to RF and XGBoost. This study demonstrates the potential of FBCR to significantly enhance the accuracy and reliability of predictive models in the real estate domain, contributing to the advancement of machine learning applications in practical settings.

4. CONCLUSIONS

This study investigated the efficacy of three machine learning models—Forest-based Classification and Regression (FBCR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost)—in predicting property prices. A comprehensive framework was employed, encompassing data preprocessing, outlier removal, feature selection, model training, and rigorous evaluation using training, validation, and testing datasets. While XGBoost demonstrated exceptional performance on the training data, its susceptibility to overfitting led to a significant decline in accuracy on the unseen testing data. RF consistently underperformed across all datasets, indicating limitations in capturing the data's complexity. In contrast, FBCR exhibited a balanced performance profile, achieving high accuracy on the training data while maintaining robust generalization capabilities on the testing dataset. This was evidenced by its superior performance across various metrics, including R-squared, RMSE, MAE, and mean accuracy.

The findings highlight the crucial role of FBCR as a reliable and accurate model for property price prediction. Its ability to generalize effectively to unseen data underscores its practical applicability in real-world scenarios. This research contributes to the growing body of evidence supporting FBCR's efficacy in complex prediction tasks, paving the way for its wider adoption in diverse domains. Future research may explore further optimization techniques, feature engineering strategies, and the application of FBCR to other regression problems.

ACKNOWLEDGEMENT

The author extends sincere gratitude to professor *'Huang Dongbin'*, for his invaluable insights and guidance in exploring the application of machine learning within the GIS domain. His support was instrumental in the discovery and implementation of the Forest-based Classification and Regression (FBCR) model, a critical component of this research. Also expresses appreciation to my supervisor *'Zhang Nian'*, for his consistent availability to answer questions and encourage alternative solutions when faced with disappointing results.

Furthermore, the author expresses heartfelt appreciation to my family for their unwavering encouragement and emotional support throughout the research process, which proved essential during challenging phases.

Finally, the author acknowledges the self-driven nature of this study, which involved independent data collection, experimentation, testing, and methodological development, culminating in the completion of this research paper.

REFERENCES

1. Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House price prediction using random forest machine learning technique. *Procedia Computer Science*, *199*, 806-813.
2. Anderson, T., & Wright, R. (2022). Evaluating Transportation Demand Models: A Machine Learning Perspective. *Transportation Research*, 34(2), 102-116.
3. Bagnara, M. (2024). Asset pricing and machine learning: A critical review. *Journal of Economic Surveys*, *38*(1), 27-56.
4. Bakker, R. (2021). *Does the development of industrial sites influence the price of residential properties?* (Doctoral dissertation).
5. Colak, Z. (2021). A causality analysis on factors affecting housing prices: case of Turkey. *Journal of Business Economics and Finance*, *10*(2), 58-71.
6. Currie, J., Davis, L., Greenstone, M., & Walker, R. (2015). Environmental health risks and housing values: evidence from 1,600 toxic plant openings and closings. *American Economic Review*, *105*(2), 678-709.
7. Chen, B., Min, Y., & Yu, S. (2024). The Research on Factors Influencing Housing Prices-Take Beijing as an Example. *Highlights in Science, Engineering and Technology*, *88*, 724-730.
8. Chakraborty, A., Banerjee, R., & Das, S. (2021). Deep Learning for Housing Price Prediction: A Comparative Study. *International Journal of Machine Learning*, 12(4), 56-72.
9. Dang, V. H., Hoang, N. D., Nguyen, L. M. D., Bui, D. T., & Samui, P. (2020). A novel GIS-based random forest machine algorithm for the spatial prediction of shallow landslide susceptibility. *Forests*, *11*(1), 118.
10. De Vor, F., & De Groot, H. L. (2011). The impact of industrial sites on residential property values: A hedonic pricing analysis from the Netherlands. *Regional Studies*, *45*(5), 609-623.
11. Del Giudice, V., De Paola, P., Bevilacqua, P., Pino, A., & Del Giudice, F. P. (2020). Abandoned industrial areas with critical environmental pollution: Evaluation model and stigma effect. *Sustainability*, *12*(13), 5267.
12. El Mouna, L., Silkan, H., Haynf, Y., Nann, M. F., & Tekouabou, S. C. (2023). A Comparative Study of Urban House Price Prediction using Machine Learning Algorithms. In *E3S Web of Conferences* (Vol. 418, p. 03001). EDP Sciences.
13. Feng, Y., Zhang, W., & Liu, H. (2021). Machine Learning for Agricultural Yield Prediction. *Computers and Electronics in Agriculture*, 189, 106349.
14. Feng, X., Jaimovich, N., Rao, K., Terry, S. J., & Vincent, N. (2023). Location, location, location: Manufacturing and house price growth. *The Economic Journal*, *133*(653), 2055-2067.
15. Geetha, V., Punitha, A., Abarna, M., Akshaya, M., Illakiya, S., & Janani, A. P. (2020, July). An effective crop prediction using random forest algorithm. In *2020 international conference on system, computation, automation and networking (ICSCAN)* (pp. 1-5). IEEE.
16. Hong, J., Choi, H., & Kim, W. S. (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management*, *24*(3), 140-152.
17. Hong, J., & Kim, W. S. (2022). Combination of machine learning-based automatic valuation models for residential properties in South Korea. *International Journal of Strategic Property Management*, *26*(5), 362-384.
18. Kamtziridis, G., Vrakas, D., & Tsoumakas, G. (2023). Does noise affect housing prices? A case study in the urban area of Thessaloniki. *EPJ Data Science*, *12*(1), 50.
19. Ma, J., Cheng, J. C., Jiang, F., Chen, W., & Zhang, J. (2020). Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques. *Land use policy*, *94*, 104537.
20. Liang, X., Liu, Y., Qiu, T., Jing, Y., & Fang, F. (2018). The effects of locational factors on the housing prices of residential communities: The case of Ningbo, China. *Habitat International*, *81*, 1-11.
21. Li, Q., & Chen, H. (2020). Balancing RMSE and MAE in Energy Consumption Predictions. *Energy Efficiency*, 13(7), 1235-1248.
22. Tanamal, R., Rasyid Jr, N. M. K. S., Wiradinata, T., Soekamto, Y. S., & Saputri, T. R. D. (2023). House price prediction model using random forest in surabaya city.
23. Tchuente, D., & Nyawa, S. (2022). Real estate price estimation in French cities using geocoding and machine learning. *Annals of Operations Research*, *308*(1), 571-608.
24. Xu, L., & Li, Z. (2021). A new appraisal model of second-hand housing prices in China's first-tier cities based on machine learning algorithms. *Computational Economics*, *57*(2), 617-637.
25. Yin, Y., Zeng, X., Zhong, S., & Liu, Y. (2022). How real estate shocks affect manufacturing value chain upgrading: Evidence from China. *Buildings*, *12*(5), 546.
26. Yamannage, S. N. (2024). Comparative Analysis of Machine Learning Algorithms for Predicting House Prices.
27. Zhang, L. (2023). Housing price prediction using machine learning algorithm. *Journal of World Economy*, *2*(3), 18-26.
28. Zhang, L., Jiantao, Z. H. O. U., Eddie, C. M., & Haizhen, W. E. N. (2019). The effects of a shopping mall on housing prices: A case study in Hangzhou. *International Journal of Strategic Property Management*, *23*(1), 65-80.
29. Zhou, Y., Liu, X., & Sun, M. (2024). Research on the Price Prediction of Commercial Housing in Beijing. *Highlights in Business, Economics and Management*, *30*, 118-124.
30. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
31. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
32. Gauss, C. F. (1809). Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium.
33. Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. *Philosophical Transactions of the Royal Society A*, 187, 253-318.
34. Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les événements. Mémoires de l'Académie Royale des Sciences de Paris.
35. Zhao, X., Zhang, Y., & Li, J. (2020). Real Estate Price Prediction Using Machine Learning Techniques. *Applied Sciences*, 10(12), 1-18.
36. Zhou, Z., Liu, Y., & Wang, X. (2019). Predicting Land-Use Patterns with Random Forest Regression. *Journal of Geospatial Analysis*, 8(3), 45-62.

BIOGRAPHIES

KAMONCHANOK TONLO
Master degree student
Nationality: Thai
E-mail:
kamonchanok.tonlo@gmail.com