The Prediction of Low-Rise Building Construction Cost Estimation Using Extreme Learning Machine

Kittisak Lathong, Kittipol Wisaeng*

Mahasarakham Business School, Mahasarakham University, Mahasarakham, Thailand

Received 05 August 2023; received in revised form 15 October 2023; accepted 23 October 2023

DOI: https://doi.org/10.46604/aiti.2023.12687

Abstract

This study aims to predict the possibility of low-rise building construction costs by applying machine learning models, and the performance of each model is evaluated and compared with ensemble methods. The artificial neural network (ANN) emerges as the top-performing individual model, attaining an accuracy of 0.891, while multiple linear regression and decision trees follow closely with accuracies of 0.884 and 0.864 respectively. Ensemble methods like maximum voting ensemble (MVE) improve the accuracy beyond individual models with an impressive accuracy rate of 0.924. Meanwhile, the stacking ensemble and averaging ensemble also demonstrate competitive performance with accuracies of 0.883 and 0.871, respectively. These findings can result in more informed decision-making, which is valuable for the real estate industry.

Keywords: model, low-rise building construction cost, machine learning, ensemble learning model

1. Introduction

Housing is one of the four essential aspects of human life, and the demand for residential properties and housing upgrades is rising owing to the growth of that current population. Therefore, the importance of housing to family well-being cannot be overstated with homeownership [1] being seen as a cornerstone of both family stability and wealth creation [2]. On the other hand, housing prices serve as a reflection of the overall quality of life within urban environments and a fundamental factor in construction and resilient cities [3]. This surge in demand directly impacts the real estate industry, increasing residential construction projects including single-family homes, townhouses, commercial buildings, condominiums, etc. Data from the National Statistical Office and the National Economic and Social Development Council of Thailand [4] reveals that the number of permits issued for building construction exhibited varying patterns with a steady fluctuation from 2011 to 2022.

However, since 2016, the number of permits granted for building construction has evinced a consistent upward trend. The burgeoning real estate sector in Thailand has witnessed remarkable growth, which results in wielding a profound impact on the nation's economy and exerting considerable influence over household consumption and savings. Attaining a judicious and well-informed appraisal of real estate assets holds intrinsic advantages for a spectrum of stakeholders, ranging from urban policymakers within the government's purview to real estate vendors and individual purchasers [5].

Two main construction cost estimation methods are presented. One is the rough estimation or approximate cost estimation providing a preliminary rough estimate to study the initial feasibility with less time but a higher margin of error. The other one is detailed estimation involving calculating the quantities of work and their corresponding prices, while generally, this type of

* Corresponding author. E-mail address: kittipol.w@acc.msu.ac.th

English language proofreader: Chih-Wei Chang

estimation is performed with the availability of construction designs and specifications. Detailed estimation yields relatively accurate results but is more time-consuming. It involves presenting an itemized list of material quantities and construction costs, resulting in higher accuracy but a longer estimation process [6].

Machine learning (ML) has found widespread application across diverse industrial and business sectors [7-9]. ML technologies have evolved significantly and expanded their capabilities across a wide range of applications [10]. Noteworthy studies have demonstrated the effectiveness of ML methodologies in predicting or classifying various factors, such as interest rates and prices within the real estate domain [11]. Construction cost prediction represents a prominent area of investigation where ML capabilities have been thoroughly explored. Moreover, construction cost prediction presents a multifaceted non-linear challenge influenced by various direct and indirect attributes and construction year [12].

The traditional method of estimating residential construction costs, whether rough or detailed, always requires human resources such as engineers, architects, estimators, and other resources like computers, printers, paper, etc. Additionally, it takes considerable time to obtain accurate and precise cost figures. To address this, researchers are interested in using a new approach where computers can learn independently through data or simulations with artificial intelligence (AI). They employ the principles of extreme learning machine (XLM), creating models for predicting flat-house construction costs directly from the house's area without the need for architects and engineers to design. Instead, the predicted cost should closely approximate the traditional estimation.

ML is a subfield of AI that works alongside algorithms and technologies to extract useful information from data. ML is appropriate for calculations involving data due to the impracticability and inefficiency of manually processing the data in the absence of ML. Therefore, it depends on creating algorithms enabling ML to be a predictive algorithm method capable of processing quantitative data for forecasting. According to the statement of Xu et al. [13], ML has significantly transformed various industries and become a powerful tool in the construction sector as it automates processes. ML technology is substantial in processing massive volumes of data to achieve time savings and optimize processing resources. Moreover, it may be particularly suitable for the construction industry to predict both financial and time expenditure and attain the maximization of efficiency.

In the realm of construction cost estimation, Tayefeh Hashemi et al. [14] conducted an exhaustive analysis of research papers spanning 30 years from 1985 to 2020. These papers were dedicated to the application of ML techniques for cost estimation in construction projects. The overarching goal of these studies was to develop predictive models capable of providing accurate cost estimates, particularly during the pre-bidding phase, thereby facilitating informed decision-making by project managers. Notably, prevalent ML techniques have employed in the reviewed literature including ANNs, regression analysis (RA), case-based reasoning (CBR), and support vector machines (SVMs) respectively. This analysis aligns with the findings of Elfaki et al. [15], in their survey of construction cost estimation over the past decade, which also underscored the enduring prominence of classic ML techniques notably ANNs and SVM within the field.

Various methodologies are employed to forecast residential prices [7]. In contrast to the classic price prediction approaches, ANN-based methods have exhibited promising outcomes in real estate assessment [16]. Their primary advantage lies in their capacity to discern non-linear correlations between inputs and outputs, rendering them particularly suitable for predicting non-linearities in real estate price assessment [17]. Recent investigations have asserted the effectiveness of ML models, including ANNs in real estate price prediction tasks [11].

Khalaf et al. [18] initially applied particle swarm optimization (PSO) for cost and construction time estimation in 60 projects. The study attested that PSO performed well and yielded highly accurate results despite the presence of parameters with diverse uncertainties. The strength of this approach lies in its reliance on existing and more reliable projects compared to those considered for estimation and testing. Conversely, Jiang [19] investigated the use of ANN for construction project estimation and compared its results with the radial basis function neural network (RBFNN) method and found, that ANN outperformed RBFNN. Subsequently, the ANN model's efficiency and application to other project types were examined, considering additional cost factors.

In a study conducted by Park et al. [20], the crucial importance of accurate cost estimation during the initial stages of construction projects is underscored. Particularly in situations where essential data for construction cost prediction is scarce, the study introduces an innovative two-level stacking ensemble algorithm incorporating random forest (RF), SVM, and CatBoosting. The optimal hyperparameter values for these base models are determined through Bayesian optimization coupled with cross-validation. Utilizing cost data from the Public Procurement Service in South Korea, the research demonstrates the two-level stacking ensemble model consistently outperforms individual ensemble models in predictive accuracy.

While classical models employed in prior research exhibit commendable predictive capabilities, ML models have demonstrated unsatisfactory performance in comparison. Notably, existing studies focused on real estate valuation and prediction in the residential domain primarily rely on readily available ML methods. However, the literature scarcely highlights the significance of incorporating ensemble models derived from general models to facilitate collective learning and enhance performance levels, thereby bolstering robustness. Significantly, it is noteworthy that there is currently an almost complete absence of methods for predicting construction prices in Thailand through the deployment of ensemble learning techniques. Moreover, in instances where ensemble learning methods have been employed, a notable dearth of comprehensive comparative demonstrations involving multiple ensemble learning approaches is emerged. Consequently, a discernible knowledge gap persists in the quest to enhance the efficiency of ML models for real estate price prediction. This current study endeavors to bridge these gaps comprehensively and address these issues.

From the aforementioned discussion, the problem arises when business owners, project managers, or homeowners need assistance to estimate construction costs accurately and thoroughly, leading to longer estimation periods and reliance on engineers, architects, or estimators. To address these challenges, developing a predictive model for estimating flat-house construction costs using extreme ML would be beneficial. This AI technique enables computers to learn autonomously. This model could aid in predicting construction costs directly from house area data, reducing the need for human involvement in the design process while maintaining a close approximation to traditional estimation methods.

The research aims to achieve three primary objectives. Firstly, it aims to introduce an innovative approach for horizontally estimating low-rise construction costs, employing XLM techniques. Secondly, the study strives to identify the most effective and suitable prediction method for estimating low-rise construction costs within this context. Lastly, the research endeavors to conduct a comprehensive performance comparison between various ensemble learning methods and the conventional approach in predicting low-rise building construction costs. Once achieving these objectives, the study aspires to offer valuable insights into housing price estimation, ultimately enhancing the accuracy and efficiency of predictions within the real estate market.

The rest of this paper follows a structured organization. In Section 2, an overview of the dataset and a comprehensive data analysis are provided. Sections 2.1 to 2.3 elucidate the concept of base models, while Sections 2.4 to 2.5 expound upon the intricacies of the proposed ensemble learning model. Empirical findings are presented in Section 3, culminating in the concluding remarks in Section 4.

2. Data and Methodology

Predicting low-rise building prices accurately is paramount in the real estate industry. This article presents a comprehensive methodology for low-rise building price prediction, which comprises the following 5 phases, data preprocessing, build base predictive model, build ensemble model, evaluation model, and results and conclusions, as depicted in Fig. 1. This section is dedicated to the methodology for the development of algorithms aimed at predicting the construction cost of low-rise buildings.

Phase 1 – Data preprocessing marks the outset, including comprehensive data preprocessing tasks such as data cleaning, handling missing values, addressing outliers, and data splitting.

Phase 2 – Building base predictive models follows the construction of base predictive models using regression-based models, specifically ANNs, SVMs, multiple linear regression (MLR), decision trees (DTs), and RF. A 10-fold cross-validation approach is deployed in this phase, coupled with hyperparameter tuning for optimizing model performance.

Phase 3 – While the maintenance of a 10-fold cross-validation strategy, constructing ensemble models leverages the base models to create ensemble models through maximum voting, averaging, stacking, boosting, and bagging.

Phase 4 – Evaluation of predictive model accuracy entails a rigorous assessment concerning the performance of both base model and ensemble learning models. Multiple performance metrics, such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R2), are employed to ensure a comprehensive evaluation.

Phase 5 – Results and conclusions emerged by consolidating results, performing model comparisons, conclusions, and making recommendations for selecting the optimal model for predicting the construction costs of low-rise buildings. This structured approach ensures a systematic analysis throughout all phases of the study.

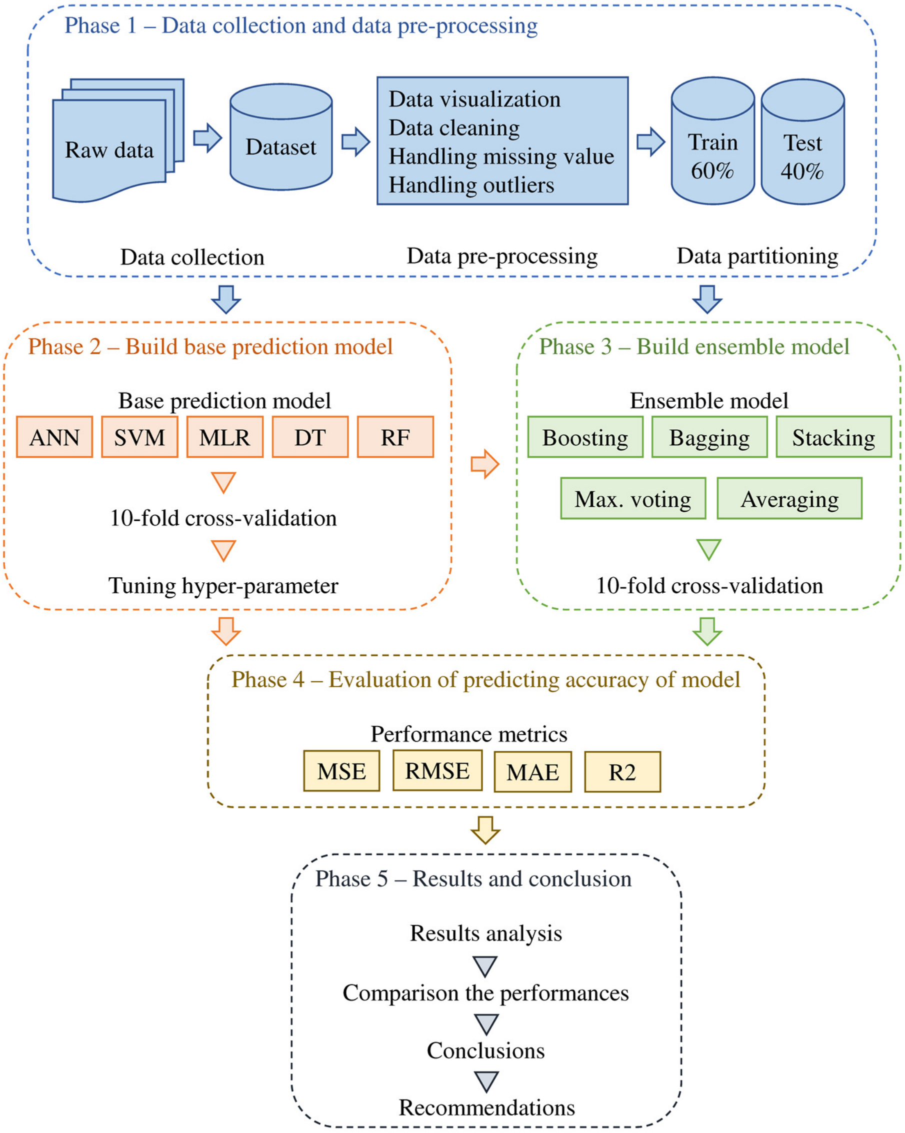

Fig. 1 Research methodology

2.1. Dataset

The dataset for the implementation of the proposed model is obtained from the Bureau of Public Works in Bangkok, Thailand. It includes drawing data related to a low-rise building and their respective prices. This dataset constitutes carefully selected features, chosen through a comprehensive review of existing literature and expert domain knowledge. The dataset and its associated features are presented in Table 1, with the left column listing the feature names, the middle column providing concise descriptions of each feature, and the right column referencing the corresponding features from the literature review.

Table 1 The description of the dataset

Feature	Details description of the dataset	Reference
Y	The construction cost of low-rise buildings	[1, 5-6, 17, 20-22]
X1	The number of stories	[1, 6, 20, 22-25]
X2	The gross floor area	[5-6, 17, 20, 22-24]
X3	The area of bedrooms	[1, 5-6, 21, 25]
X4	The area of bathrooms	[5, 21, 25]
X5	The area of living rooms or restrooms	[5, 21, 25]
X6	The area of kitchen and dining rooms	[21]
X7	The area for laundry facilities	[21]
X8	The area of balconies	[21]
X9	The area of stairs and corridors	[21]
X10	The area of the roof covered by the concrete roof and awning	[1, 6, 23]
X11	The area of the roof covered by tiled roofing	[1, 6]
X12	The area of the parking lot	[20-22]
X13	The overall height	[5-6, 20, 23]
X14	The height of the roof	By professional experience.
X15	The average height of each story	[6, 17, 23]

2.2. Data analysis

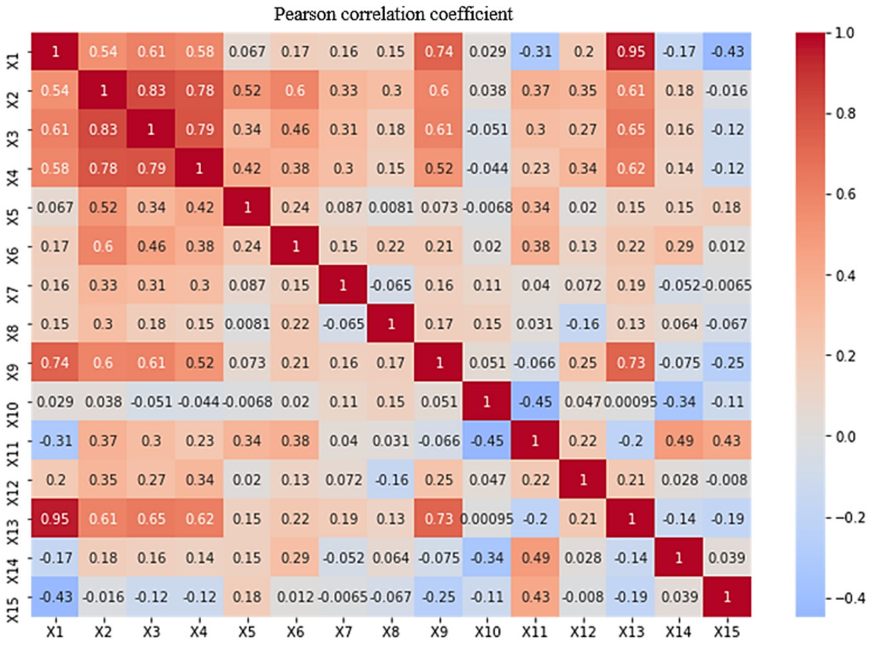

Fig. 2 Pearson correlation coefficient

Conducting a comprehensive analysis of the dataset before commencing model construction is a foundational and illuminating step in data understanding. To explore the relationships among the dataset's attributes, an integral aspect of this analytical procedure involves the application of a correlation matrix illustrated in Fig. 2. The correlation matrix provides coefficients ranging from +1 to -1 [26], shedding light on the level of association between various attribute pairs. A positive coefficient denotes a direct relationship, while a negative coefficient signifies an inverse connection. Meanwhile, a coefficient of zero indicates independence among variables [26]. The scrutiny of this matrix yields valuable insights into attribute interdependencies, equipping us to make well-informed decisions during the modeling phase. The observations depicted in Fig. 2 emphasize certain variables exhibiting noteworthy correlations, specifically including four points as follows: (1) X1:X13 = 0.95, (2) X2:X3 = 0.83, (3) X2:X4 = 0.78, and (4) X3:X4 = 0.79.

2.3. Data pre-processing

Pivotal data preprocessing steps were meticulously executed, encompassing the removal of missing data and precise data segmentation into training and testing sets. These processes seamlessly integrated extensive data visualization and attentive data cleaning, which ensures comprehensive management of missing values and outliers before model deployment. In the initial stage of the low-rise building price prediction methodology, the focus lies in the selection of a high-quality dataset. Curated diligently from reputable sources, it comprises attributes such as room area, building height, and amenities, while all these attributes can influence low-rise building prices. Robust data preprocessing techniques further prepare the dataset by handling outliers, resolving missing values, and deriving meaningful features. Finally, data splits into training, validation, and testing subsets, which facilitates accurate model assessment, model training, hyperparameter fine-tuning, and performance evaluation.

2.4. Machine learning techniques base model

ML algorithms offer the capability to model intricate and ill-defined systems even in the presence of unknown nonlinear relationships. Within the scope of this study, a set of five distinguished ML algorithms including ANNs, SVMs, MLR, DT, and RF are deployed to develop the proposed ensemble models. The inclusion of these varied algorithms bolsters the predictive potency of the approach. Furthermore, comprehensive elucidations of the design parameters for each algorithm are presented in detail.

2.4.1. Artificial neural networks (*ANNs*)

ANNs are a class of ML models inspired by the structure and functioning of the human brain. They are utilized to solve complex problems, particularly pattern recognition, classification, and regression. Moreover, ANN consists of interconnected nodes, known as neurons, which are organized into layers. These layers include an input layer to receive data, while single or multiple hidden layers are for intermediate processing, and an output layer is employed to produce the final result. Each neuron in the network is connected to every neuron in the adjacent layers, and these connections have associated weights that determine the strength of the connection. During the training process, ANNs learn from input data to adjust the weights and biases, enabling them to capture intricate relationships in the data and make accurate predictions. The advantage of using ANN for low-rise building price prediction is their ability to capture nonlinear relationships and patterns in the data, which traditional linear regression models may need assistance to handle. In summary, ANNs can generalize well to new, unseen data, making them well-suited for real estate data's dynamic and diverse nature.

2.4.2. Support vector machine (*SVM*)

The SVM method is employed to transform the predicted low-rise building price parameters into a high-dimensional feature space using a Kernel function like linear, polynomial, or Gaussian. A linear regression function is subsequently computed to confirm the deviation from the actual model outputs by at most ε for all training data while maintaining the function to the maximum flat extent. The asymmetrical loss function is utilized to train the SVM model to create a flexible cylinder with a minimal radius symmetrically wrapped around the regression function, which effectively omits absolute errors smaller than ε. The selected Kernel function for the SVM model is Laplace, which is a versatile choice suitable for regression tasks. A trial-and-error approach is applied to determine the cost of constraint violation and the value of ε, yielding optimal values of 10 and 0.1, respectively [27]. These parameter settings contribute to the robustness and accuracy of the SVM model in predicting low-rise building prices.

Support vectors represent essential data points positioned closest to the decision boundary or hyperplane, distinctly classifying categories. Their strategic arrangement profoundly affects boundary placement and orientation. The margin, the space between these support vectors, and the decision boundary, critically gauges the generalization and resilience of SVM. A broader margin signifies heightened class segregation, enhancing predictions for unobserved data. The objective of SVM employment is to identify the hyperplane maximize this margin and efficiently separate data. In a two-class scenario, the hyperplane acts as the class-separating threshold. The optimal choice of SVM advances the margin gauged by the hyperplane-to-support-vector distance. Non-linearity management employs kernel functions and elevated-dimensional space conversion.

2.4.3. Multiple linear regression (MLR)

Linear regression is a statistical method for modeling the relationship between a dependent variable (also known as the target or outcome variable) and single or multiple independent variables (also known as predictor variables or features). It is a simple and widely used technique in statistics and ML for predicting continuous numeric values. In linear regression, the goal is to find the best-fit line representing the linear relationship between the independent and dependent variables. The equation of the line is described as:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + + b_n x_n \tag{1}$$

where y is the dependent variable or low-rise building prices b_0 is the intercept or bias term, representing the value of y when all independent variables are 0, b_1, b_2, …, b_n are the coefficients or weights associated with each independent variable, and x_1, x_2, …, x_n are the independent variables.

The coefficients (b_1, b_2, …, b_n) are estimated during the training process, such that the line minimizes the sum of squared differences between both predicted and actual values in the training data. This process is commonly known as “fitting the model” to the data.

Once the model is trained, predictions for new data can be made by plugging in the values of the independent variables into the equation. Linear regression is particularly effective when a linear relationship is rendered between the dependent and independent variables. Furthermore, it remains suitable for various applications, including predicting low-rise building prices, analyzing economic trends, and understanding the impact of different factors on a particular outcome.

2.4.4. Decision tree (DT)

DT regression is a robust and interpretable algorithm for predicting numeric values in regression tasks. The algorithm constructs an arborescent model by recursively partitioning the dataset into subsets based on the values of its features. The feature and split point that minimize the variance or MSE of the target variable within the subset are selected at each node. This process sustains until a stopping criterion, such as a maximum tree depth or a minimum number of samples per leaf, is satisfied. The prediction for each data point is generated by traversing the DT from the root node to a leaf node, where the numeric value associated with the leaf node serves as the final prediction for the target variable.

Given these facts, DT for regression is highly interpretable due to the enablement of a clear understanding of the decision-making process and the realization of the rationale influencing the predicted numeric outcomes. Moreover, their adaptability to nonlinear relationships attains suitability for various applications, ranging from financial forecasting to medical diagnosis, where precise numeric predictions are required. However, to prevent overfitting, pruning techniques, and hyperparameter tuning are often employed to control the complexity of the model and maintain robust generalization performance.

2.4.5. *Random forest* (*RF*)

RF is a potent ensemble learning technique for regression tasks. It is constituted as an extension of the DT algorithm combining multiple decision trees to optimize the accuracy of predictions. In an RF for regression, a collection of DTs is established with random subsets of both the training data and features. Each DT in the forest independently makes predictions, and the final prediction is obtained by averaging or taking the median of the predictions from all the individual trees. This ensemble approach mitigates the risk of overfitting and augments the model's ability to generalize well to new, unseen data.

RF for regression offers several advantages, such as handling nonlinear relationships between features and the target variable, capturing complex interactions, and impact reduction of noisy data. The randomness introduced in building the trees also strengthens the model against outliers and stabilizes per se. Additionally, RF proffers a feature importance score, indicating the relative importance of each feature in the prediction process. This information is perceived to be valuable for understanding which features are most influential in determining the regression outcome. Overall, RF is a versatile and effective algorithm for regression tasks. It is a popular choice in various domains, including finance, healthcare, and retail, where accurate numeric predictions are essential for decision-making.

2.5. *Ensemble learning model*

Ensemble learning is a powerful technique in ML where multiple models are combined to improve predictive performance, robustness, and generalizability compared to using a single model. Each ensemble method utilizes a different strategy to combine the predictions of individual base models, utilizing techniques such as maximum voting ensemble, averaging ensemble, stacking ensemble, bagging ensemble, and boosting ensemble.

2.5.1. *Maximum voting ensemble* (*MVE*)

Ensemble models for predicting low-rise building prices using the maximum voting method typically involve combining the predictions from multiple individual models to attain the final prediction. In this approach, several aforementioned base models, such as ANNs, SVMs, MLR, DTs, or RFs, are trained independently on the same dataset. Once the individual models are trained, predictions for the target variable are made, which, in this case, is the low-rise building price. The MVE method takes the mode of all the predictions provided by the individual models. The mode value is selected as the final prediction for the low-rise building price, as demonstrated in Fig. 3.

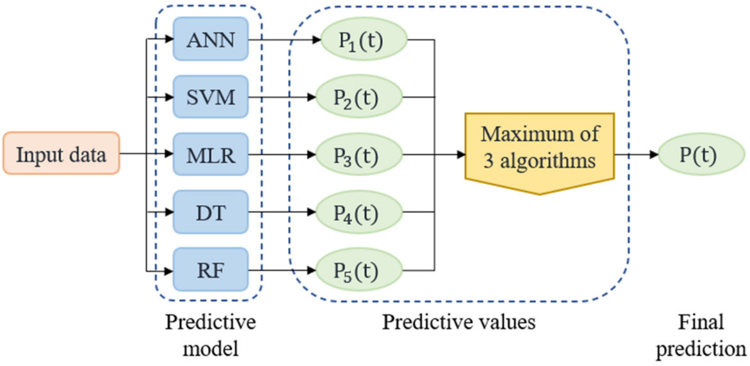

Fig. 3 Ensemble techniques - MVE

The rationale behind the MVE is to leverage the collective wisdom of diverse models. Both strengths and weaknesses are rendered in each base model, and with the combination of predictions, the ensemble aims to mitigate the impact of individual model errors and produce more accurate and robust predictions. This method is particularly effective when low correlation has emerged in the individual models, i.e., they make errors in different instances. By taking the mode, it enables the ensemble to employ the most frequently agreed-upon prediction among the models, leading to a more reliable overall forecast for low-rise building prices.

2.5.2. *Averaging ensemble*

Averaging ensemble is another popular method for combining predictions from multiple individual models to make a final prediction in regression tasks, as illustrated in Fig. 4, such as predicting low-rise building prices. In this approach, several base models, such as ANNs, SVMs, MLR, DTs, or RFs, are trained independently on the same dataset.

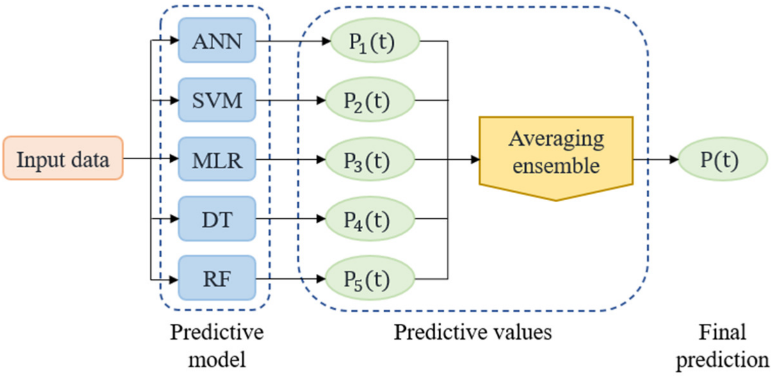

Fig. 4 Ensemble techniques - averaging ensemble

After being trained, the individual models predict the target variable (low-rise building construction cost). In the averaging ensemble, the final prediction is obtained by averaging the predictions from all the individual models. Mathematically, it can be presented as:

$$\text{Final prediction} = \frac{\text{Prediction\_Model}_1 + \text{Prediction\_Model}_2 + \cdots + \text{Prediction\_Model}_N}{N} \tag{2}$$

where N is the total number of individual models used in the ensemble.

The rationale behind the averaging ensemble is to leverage the collective knowledge of diverse models and reduce the variance of predictions. Given the presence of strengths and weaknesses in each mode and the average of predictions, the ensemble aims to create a more stable and accurate final forecast. Averaging conduces to smoothing out individual model errors and the improvement of overall predictive performance. Approximated to maximum voting, the averaging ensemble functions best when the individual models have relatively low correlation, i.e., producing errors in different instances. With averaged predictions, the ensemble can benefit from the collective insights of the models and attain a more reliable and precise forecast for low-rise building prices.

2.5.3. *Stacking ensemble*

The stacked ensemble, often referred to as stacked generalization, represents a sophisticated ensemble learning approach showcased in Fig. 5. It amalgamates predictions from various individual models to enhance the comprehensive predictive performance across ML tasks, encompassing the prediction of low-rise building prices. Stacking goes beyond simple averaging or voting methods by introducing a meta-model learning to combine the outputs of the base models.

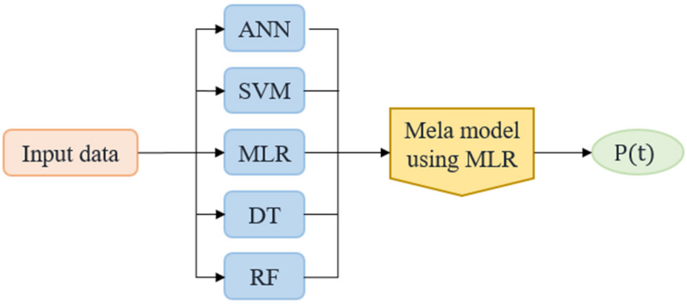

Fig. 5 Ensemble techniques - stacking ensemble

The stacking ensemble involves the following steps:

(1) Base models: Several diverse base models are trained independently on the same dataset. These base models can be different or the same algorithm with different hyperparameters.

(2) Hold-out validation set: A hold-out validation set is created from the training data. The base models make predictions on this validation set as input to the meta-model.

(3) Meta-model: A meta-model, also called the “blender” or “aggregator,” is trained with the predictions from the base models on the validation set as its input features. The meta-model learns how to combine these predictions to generate the final prediction for the target variable (low-rise building price).

(4) Final prediction: Once the meta-model is trained, predictions are made herein. The final forecast for low-rise building prices is obtained from the output of the meta-model.

The key advantage of the stacking ensemble is the ability to capture higher-order relationships between the base models’ predictions. It learns from trusting each base model and assigning different weights to their predictions based on their performance on the validation set. This enables stacking to outperform individual models and simple averaging/voting ensembles by the augmentation of the strengths. Stacking is a flexible and powerful technique but correspondingly requires careful implementation and tuning to prevent overfitting. Properly executed stacking can enhance predictive accuracy and robustness, popularizing itself in various ML competitions and real-world applications, including low-rise building price prediction.

2.5.4. *Bagging ensemble*

Bagging, the abbreviation for bootstrap aggregating, is a widely deployed ensemble learning technique to improve the accuracy and robustness of ML models, including those used for predicting low-rise building prices. It involves creating multiple diverse copies of the same base model, training each copy on a different random subset of the training data, and then combining their predictions to make the final prediction [28], as depicted in Fig.6.

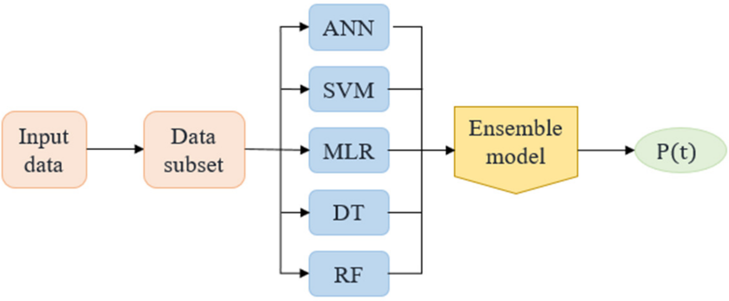

Fig. 6 Ensemble techniques - bagging ensemble

The bagging ensemble works as follows:

(1) Base model: A single base model, such as a DT or RF, is selected as the base learner.

(2) Bootstrap sampling: It creates multiple random subsets (samples) from the training data, which involves randomly selecting data points from the training set with replacements. Each subset may contain some duplicate data points and miss some others.

(3) Base model copies: For each subset, a separate copy of the base model is trained using the corresponding subset of the training data. As a result, multiple base model copies with slightly different training data are created.

(4) Aggregation: To make predictions for new data, each base model copy generates its forecast. In bagging, the final prediction is obtained by aggregating (averaging for regression tasks or voting for classification tasks) the estimates from all the base model copies.

The key advantage of bagging is to reduce the variance and overfitting in the predictions by leveraging the wisdom of diverse model copies trained on different subsets of the data. By combining multiple forecasts, bagging produces a more stable and robust final prediction that tends to generalize well to new, unseen data. Meanwhile, RF is a paradigm of the bagging ensemble, where multiple decision trees are trained on different bootstrapped subsets of the training data, and their predictions are aggregated to make the final prediction. Bagging is widely deployed in ML due to its simplicity and effectiveness in improving model performance and reducing overfitting.

2.5.5. Boosting ensemble

Boosting is another potent ensemble learning technique for improving ML models' predictive performance, including those used for predicting low-rise building prices. Unlike bagging focusing on training multiple models independently and combining their predictions, boosting builds a concatenation of models sequentially, where each model tries to correct the errors made by its predecessors [28], as depicted in Fig. 7.

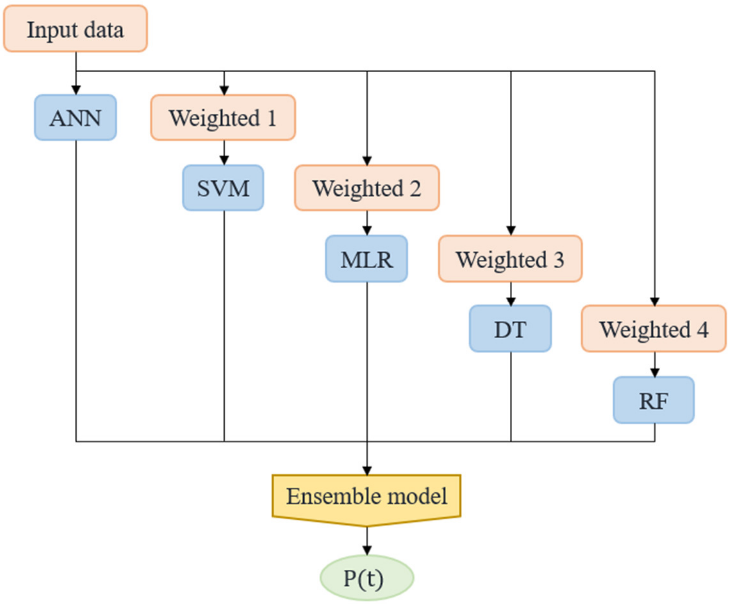

Fig. 7 Ensemble techniques - boosting ensemble

The boosting ensemble works as follows:

(1) Base model: A weak base model, often called a "weak learner," is selected as the base learner. Weak learners slightly prevail over random chance models, such as simple decision stumps (one-level DTs) or shallow DTs.

(2) Weighted data: The training data is initially given equal weight for all data points. In each boosting iteration, the weights of misclassified data points increase, while the weights of correctly classified data points decrease. This enables the subsequent weak learners to focus on the previously misclassified data points and improve the overall model performance.

(3) Sequential learning: Boosting builds a concatenation of weak learners, where each model is trained on the modified version of the training data from the previous step. The predictions of each weak learner are combined with a weight, and the final prediction is obtained by summing up the weighted predictions.

(4) Adaptive learning: The sequential nature of boosting enables it to learn from its mistakes adaptively. The subsequent weak learners are encouraged to focus on the challenging data points incorrectly predicted by the earlier weak learners.

The key advantage of boosting is the capability of improving the predictive accuracy compared to a single weak learner. By building a sequence of models correcting each other's errors, boosting creates a robust ensemble model that can capture complex relationships in the data. Gradient boosting machines and AdaBoost are two typical examples of boosting algorithms.

Gradient boosting machines iteratively minimize the loss function by adding weak learners, while AdaBoost assigns higher weights to misclassified data points, enabling subsequent weak learners to focus on those points. Boosting is a widely used technique in ML, and it often outperforms individual models and other ensemble methods. Given this virtue, it is acknowledged as a valuable tool for low-rise building price prediction and other complex regression tasks.

2.6. Performance measure

MAE measures the average of the absolute differences between predicted and actual low-rise building prices. This metric is beneficial in dealing with uniform forecast errors, as it equalizes the discrepancies in the data. It is frequently employed for regression problems and to evaluate the overall accuracy of forecasts. A smaller MAE indicates better forecasting performance in predicting low-rise building prices, reflecting a closer alignment between the predicted values and the actual costs.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_{i(pred)} - y_{i(actual)}\right| \tag{3}$$

where $y_{i(actual)}$ is the actual low-rise building prices, N is the total number of observations, and $y_{i(pred)}$ is the predicted low-rise building prices [29].

MSE is a metric used to assess the accuracy of predicting low-rise building prices. It is computed by taking the average of the squared differences between the actual and predicted values. In other words, the method encompasses the calculation of the squared discrepancy between the projected and actual prices of low-rise buildings for each prediction instance. After this computation, an averaging procedure is applied to these squared discrepancies, yielding the resultant MSE. The MSE penalizes large prediction errors more severely than small ones, providing a way to quantify the overall accuracy of the prediction model. Lower MSE values indicate better predictive performance, indicating smaller differences between predicted and actual low-rise building prices [22].

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_{i(pred)} - y_{i(actual)}\right)^2 \tag{4}$$

RMSE is a widely used metric for the accuracy evaluation of low-rise building price predictions. It is calculated by taking the square root of the average of the squared differences between the predicted and actual low-rise building prices. RMSE is highly regarded for its ability to handle outliers in the data effectively [22], enabling the identification and elimination of extreme discrepancies to emerge in predictions. Moreover, RMSE puts more emphasis on larger errors. Consequently, it is recognized as a valuable primary error metric for assessing the performance of low-rise building price-prediction models.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{i(pred)} - y_{i(actual)}\right)^2} \tag{5}$$

Determination coefficient (R2) is a critical metric that gauges the proportion of variance in predicted low-rise building prices attributed to the model’s predictions. With values ranging from 0 to 1 [22], an R2 value of 0 suggests poor model performance, while 1 indicates a perfect fit between forecast and actual values. This metric provides insights into the goodness-of-fit of a low-rise building price prediction model and conduces to understanding the model’s ability to capture variations in the data.

Low-rise building price prediction success hinges on choosing the most suitable algorithm. The performance assessment of different algorithms is conducted with meticulous consideration, employing relevant evaluation metrics such as MSE or MAE on the validation dataset. The algorithm demonstrating superior predictive performance and robust generalization capability is selected as the prediction model.

In summary, the methodologies including encompassing critical stages from data selection, preprocessing to model training, and algorithm choice comprehensively address low-rise building price prediction. By following this systematic approach, real estate professionals, investors, and analysts can confidently make informed decisions based on accurate low-rise building price predictions.

2.7. Hyperparameter tuning

With the data preprocessing phase, the performance of the base models underwent a comparative analysis. To assess the effectiveness of these base models, it becomes imperative to identify the most suitable hyperparameters for each model. Initially, the raw dataset is randomly divided into two distinct subsets, comprising a training set and a testing set. The optimal hyperparameter values for the base models, as determined by the highest R2 score during the search in the hyperparameter space, are presented in Table 2.

Table 2 The hyperparameter tuning

Base ML models	Hyperparameter	Values
Artificial neural networks (ANNs)	Activation function	ReLU
	Learning rate	0.001
	Number of epochs	1000
	Initializer	Normal
	Optimizer	Adam
Support vector machines (SVMs)	Kernel	Linear
	C (Regularization parameter)	1000
	Gamma (Kernel coefficient)	Scale
	Epsilon	0.05
Decision trees (DTs)	Criterion	Gini impurity
	Max depth	Unlimited
	Min samples per leaf	2
Random forest (RF)	Number of trees Max depth Min samples per leaf Max features	100 Unlimited 2 Auto

To evaluate and compare the performance of these base models, a ten-fold cross-validation technique was applied, known for its established effectiveness in cross-validation practices. The hyperparameters, crucial for each model, were employed with the values derived through a grid search, conducted in the preceding phase.

3. Results and Discussion

In this section, the outcomes of the model development and the subsequent training using the pre-processed dataset are presented. A thorough performance comparison among ten machine-learning algorithms has been conducted, complemented by comprehensive parameter tuning for meticulous performance analysis and evaluation. Additionally, an in-depth insight into the experimental setup employed for the entire task is provided.

The primary aim of this study was to forecast low-rise building construction costs, employing a diverse set of ten machine-learning models. To evaluate their efficacy, four distinct statistical metrics outlined in Table 3 were applied to gauge the accuracy of each model.

As shown in Table 3, the ANN emerged as the highest-performing base model, boasting an accuracy of 0.891. MLR closely followed, achieving an accuracy of 0.884, while DTs demonstrated an accuracy of 0.864. The RF model achieved an accuracy of 0.830, whereas the SVM displayed the lowest accuracy at 0.446. Concerning the ensemble models, a diverse array of techniques was employed to amalgamate predictions from individual models. The MVE demonstrated the highest accuracy, reaching 0.924, thus surpassing all individual models and other ensemble methods. The stacking ensemble secured second place with an accuracy of 0.883, while the averaging ensemble achieved an accuracy of 0.871. Regarding the boosting ensemble, it delivered an accuracy of 0.846, and the bagging ensemble yielded an accuracy of 0.832.

Table 3 The summary of the algorithms that have been used in the research

Model		R2	MSE	RMSE	MAE
Base model	ANN	0.891	47564120000.00	218092.00	165705.68
	SVM	0.446	242250000000.00	492189.90	322532.74
	MLR	0.884	50560550000.00	224856.74	146446.13
	DT	0.864	59522890000.00	243973.13	126901.04
	RF	0.830	74476350000.00	272903.56	164069.51
Ensemble model	Max. voting	0.924	33325570000.00	182552.93	127356.65
	Averaging	0.871	56266850000.00	237206.35	140085.86
	Stacking	0.883	51074880000.00	225997.52	152669.31
	Bagging	0.832	73493900000.00	271097.58	155363.36
	Boosting	0.846	67425130000.00	259663.50	147838.86

To visually convey the comparative accuracy across all models, Fig. 8 is presented as follows. This visualization highlights the promising performance exhibited by most algorithms in regression tasks, underscoring their potential for precise low-rise building cost prediction.

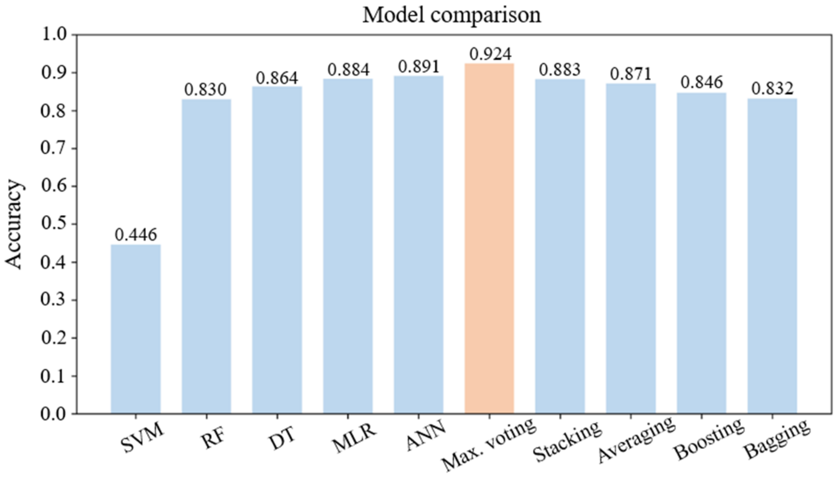

Fig. 8 Comparison of accuracy among the models

In summary, the research harnessed the predictive power of ten machine-learning models coupled with an array of ensemble techniques to anticipate low-rise building prices. The results underscore the remarkable efficacy of the MVE, situating it as a compelling choice for practical applications within this domain. Furthermore, individual models including ANN, MLR, and DT demonstrated notable accuracy which solidifies their status as valuable options for consideration.

4. Conclusions

This study harnessed the predictive potential of ten diverse ML models to anticipate low-rise building prices. The identification of top-performing models, both within ensemble methods and individual algorithms, was achieved through a rigorous and exhaustive analysis. Meanwhile, the ANN emerged as the most accurate individual model, underscoring its adeptness in addressing the intricacies of the prediction task with a remarkable accuracy of 0.891.

Concerning MLR and DTs, they are closely followed and exhibit robust predictive capabilities with accuracies of 0.884 and 0.864, respectively. Furthermore, the evaluation of ensemble models underscored the prominence of the MVE, securing the highest accuracy of 0.924 among all models tested. This ensemble technique adeptly amalgamates predictions from individual models, harnessing their collective strengths and simultaneously mitigating their shortcomings. The stacking ensemble and averaging ensemble also demonstrated competitive accuracy, achieving scores of 0.883 and 0.871, respectively. These results affirm the significant performance enhancements which are attainable through ensemble methods, surpassing the capabilities of individual models.

In a broader context, the findings from this study deliver valuable insights into real estate industry practitioners and stakeholders. The utilization of ML models, notably the ANN and MVE, possesses the potential to enhance the accuracy of low-rise building price predictions, facilitating more informed decision-making and strategic investments. As the evolution of ML burgeons, ongoing research, and enhancements in model architecture and ensemble techniques promise even more precise and dependable predictions across diverse real-world applications.

Acknowledgments

This research project was financially supported by Mahasarakham University of Thailand under Grant No. 6603024.

Nomenclature

XLM	Extreme learning machine
ANN	Artificial neural network
MVE	Maximum voting ensemble
ML	Machine learning
AI	Artificial intelligence
RA	Regression analysis
CBR	Case-based reasoning
SVM	Support vector machine
PSO	Particle swarm optimization
RBFNN	Radial basis function neural network
RF	Random forest
MLR	Multiple linear regression
DT	Decision tree
MSE	Mean squared error
RMSE	Root mean square error
MAE	Mean absolute error
R2	R-Squared

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] A. Soltani, M. Heydari, F. Aghaei, and C. J. Pettit, “Housing Price Prediction Incorporating Spatio-Temporal Dependency into Machine Learning Algorithms,” Cities, vol. 131, article no. 103941, December 2022.

[2] M. H. Karamujic, Housing Affordability and Housing Investment Opportunity in Australia, London: Palgrave Macmillan, 2015.

[3] Y. Ma and S. Gopal, “Geographically Weighted Regression Models in Estimating Median Home Prices in Towns of Massachusetts Based on an Urban Sustainability Framework,” Sustainability, vol. 10, no. 4, article no. 1026, April 2018.

[4] Department of Economic Statistics National Statistical Office, “Construction Area Information,” Processing Construction Area Information, 2021. (In Thai)

[5] S. Wang, J. Zhu, Y. Yin, D. Wang, T. E. Cheng, and Y. Wang, “Interpretable Multi-Modal Stacking-Based Ensemble Learning Method for Real Estate Appraisal,” IEEE Transactions on Multimedia, vol. 25, pp. 315-328, November 2021.

[6] S. Sittikarnkul, “Construction Cost Estimation for Government Building Using Prediction Modeling Techniques,” Master of Engineering, Graduate School, Chiang Mai University, Chiang Mai, 2021.

[7] J. Kalliola, J. Kapočiūtė-Dzikienė, and R. Damaševičius, “Neural Network Hyperparameter Optimization for Prediction of Real Estate Prices in Helsinki,” PeerJ Computer Science, vol. 7, article no. e444, 2021.

[8] R. Cioffi, M. Travaglioni, G. Piscitelli, A. Petrillo, and F. De Felice, “Artificial Intelligence and Machine Learning Applications in Smart Production: Progress, Trends, and Directions,” Sustainability, vol. 12, no. 2, article no 492, January 2020.

[9] M. Kraus, S. Feuerriegel, and A. Oztekin, “Deep Learning in Business Analytics and Operations Research: Models, Applications and Managerial Implications,” European Journal of Operational Research, vol. 281, no. 3, pp. 628-641, March 2020.

[10] N. S. Ja’afar, J. Mohamad, and S. Ismail, “Machine Learning for Property Price Prediction and Price Valuation: A Systematic Literature Review,” Planning Malaysia, vol. 19, no. 3, pp. 411-422, October 2021.

[11] J. Kang, H. J. Lee, S. H. Jeong, H. S. Lee, and K. J. Oh, “Developing a Forecasting Model for Real Estate Auction Prices Using Artificial Intelligence,” Sustainability, vol. 12, no. 7, article no. 2899, April 2020.

[12] N. Ferlan, M. Bastic, and I. Psunder, “Influential Factors on the Market Value of Residential Properties,” Engineering Economics, vol. 28, no. 2, pp. 135-144, April 2017.

[13] Y. Xu, Y. Zhou, P. Sekula, and L. Ding, “Machine Learning in Construction: From Shallow to Deep Learning,” Developments in the Built Environment, vol. 6, article no. 100045, May 2021.

[14] S. Tayefeh Hashemi, O. M. Ebadati, and H. Kaur, “Cost Estimation and Prediction in Construction Projects: A Systematic Review on Machine Learning Techniques,” SN Applied Sciences, vol. 2, article no. 1703, September 2020.

[15] A. O. Elfaki, S. Alatawi, and E. Abushandi, “Using Intelligent Techniques in Construction Project Cost Estimation: 10-Year Survey,” Advances in Civil Engineering, vol. 2014, article no. 107926, 2014.

[16] A. Varma, A. Sarma, S. Doshi, and R. Nair, “House Price Prediction Using Machine Learning and Neural Networks,” Second International Conference on Inventive Communication and Computational Technologies, pp. 1936-1939, April 2018.

[17] W. K. Ho, B. S. Tang, and S. W. Wong, “Predicting Property Prices with Machine Learning Algorithms,” Journal of Property Research, vol. 38, no. 1, pp. 48-70, 2021.

[18] T. Z. Khalaf, H. Çağlar, A. Çağlar, and A. N. Hanoon, “Particle Swarm Optimization Based Approach for Estimation of Costs and Duration of Construction Projects,” Civil Engineering Journal, vol. 6, no. 2, pp. 384-401, February 2020.

[19] Q. Jiang, “Estimation of Construction Project Building Cost by Back-Propagation Neural Network,” Journal of Engineering, Design and Technology, vol. 18, no. 3, pp. 601-609, 2020.

[20] U. Park, Y. Kang, H. Lee, and S. Yun, “A Stacking Heterogeneous Ensemble Learning Method for the Prediction of Building Construction Project Costs,” Applied Sciences, vol. 12, no. 19, article no. 9729, October 2022.

[21] W. Dangsangthong, “Residential Building Cost Estimation Using Artificial Neural Network Approach,” Master of Engineering, Graduate School, Chiang Mai University, Chiang Mai, 2016.

[22] V. Chandanshive and A. R. Kambekar, “Estimation of Building Construction Cost Using Artificial Neural Networks,” Journal of Soft Computing in Civil Engineering, vol. 3, no. 1, pp. 91-107, January 2019.

[23] S. H. Ji, J. Ahn, H. S. Lee, and K. Han, “Cost Estimation Model Using Modified Parameters for Construction Projects,” Advances in Civil Engineering, vol. 2019, article no. 8290935, 2019.

[24] C. L. C. Roxas and J. M. C. Ongpeng, “An Artificial Neural Network Approach to Structural Cost Estimation of Building Projects in the Philippines,” DLSU Research Congress, pp. 1-8, March 2014.

[25] N. Vineeth, M. Ayyappa, and B. Bharathi, “House Price Prediction Using Machine Learning Algorithms,” Soft Computing Systems: Second International Conference, pp. 425-433, April 2018.

[26] L. El Mouna, H. Silkan, Y. Haynf, M. F. Nann, and S. C. Tekouabou, “A Comparative Study of Urban House Price Prediction Using Machine Learning Algorithms,” E3S Web of Conferences, vol. 418, article no. 03001, August 2023.

[27] P. A. G. M. Amarasinghe, N. S. Abeygunawardana, T. N. Jayasekara, E. A. J. P. Edirisinghe, and S. K. Abeygunawardane, “Ensemble Models for Solar Power Forecasting—A Weather Classification Approach,” AIMS Energy, vol. 8, no. 2, pp. 252-271, 2020.

[28] N. Rahimi, S. Park, W. Choi, B. Oh, S. Kim, Y. H. Cho, et al., “A Comprehensive Review on Ensemble Solar Power Forecasting Algorithms,” Journal of Electrical Engineering & Technology, vol. 18, no. 2, pp. 719-733, March 2023.

[29] P. Kumari and D. Toshniwal, “Deep Learning Models for Solar Irradiance Forecasting: A Comprehensive Review,” Journal of Cleaner Production, vol. 318, article no. 128566, October 2021.