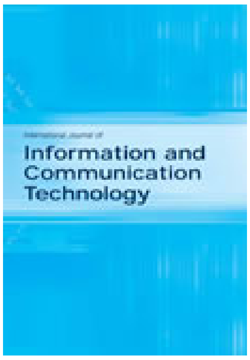


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Using regression-based machine learning model to estimate constructions cost

Yanfei Shen

DOI: 10.1504/IJICT.2025.10071389

Article History:

Received:	19 February 2025
Last revised:	08 March 2025
Accepted:	08 March 2025
Published online:	11 June 2025

Using regression-based machine learning model to estimate constructions cost

Yanfei Shen

Sanmenxia Polytechnic,
Sanmenxia, Henan, 472000, China
Email: 15939841073@163.com

Abstract: Accurate construction cost estimation is crucial for effective project planning and resource management in the construction industry. Traditional estimation methods often suffer from inaccuracies due to the complexity and variability of construction projects. This study explores the application of regression-based machine learning models support vector machine (SVM), K-nearest neighbours (KNN), and multilayer perceptron (MLP) to improve the precision of construction cost predictions. The study evaluates the performance of these models using a construction-related dataset that includes factors such as material costs, labour expenses, and project characteristics. The results revealed that the SVM model outperforms the others, achieving an RMSE of 18,189 and an R^2 of 0.975, indicating its superior ability to predict construction costs accurately. The KNN and MLP models also demonstrated effectiveness, but with higher errors, particularly in more complex data scenarios. This research highlights the potential of machine learning techniques to revolutionise construction cost estimation, providing more reliable, data-driven insights for project planning and budgeting.

Keywords: cost estimation; predictive analytics; project planning; cost prediction; construction industry.

Reference to this paper should be made as follows: Shen, Y. (2025) 'Using regression-based machine learning model to estimate constructions cost', *Int. J. Information and Communication Technology*, Vol. 26, No. 17, pp.72–92.

Biographical notes: Yanfei Shen is a researcher at Sanmenxia Polytechnic, specialising in the application of machine learning in construction engineering. His recent work focuses on leveraging regression-based models to accurately estimate construction costs, enhancing efficiency and budgeting precision in the industry.

1 Introduction

Estimating construction costs accurately is a critical aspect of project planning and management in the construction industry (Sayed et al., 2023). The ability to forecast costs with precision plays a pivotal role in ensuring the feasibility and sustainability of construction projects, as it directly influences budgeting, resource allocation, and decision-making processes (Moshood et al., 2024). Accurate cost estimation not only helps project managers in anticipating expenses but also minimises financial risks and supports effective communication between stakeholders (Nia et al., 2023). This is particularly crucial in a sector characterised by dynamic variables such as fluctuating material prices, labour costs, and unforeseen site conditions, which can significantly impact the overall project budget (Adepu et al., 2024).

The implications of inaccurate cost estimations are far-reaching, often leading to delays, disputes, and cost overruns that can derail project timelines and compromise profitability (Onah, 2024). Conventional methods for estimating construction costs typically rely on heuristic approaches, expert judgment, or historical data analysis (Fouda et al., 2024). While these methods have been widely used, they are inherently subjective and prone to inconsistencies due to the variability in expert experience and the limitations of relying solely on past data. These approaches may also struggle to capture the complexity of modern construction projects, which often involve numerous interrelated variables and intricate dependencies.

In recent years, the rapid advancement of machine learning (ML) techniques has opened new opportunities for addressing these limitations (Fei et al., 2015). ML models are inherently data-driven, capable of identifying complex patterns and relationships within large datasets that may not be readily apparent through traditional methods (Montáns et al., 2019). By leveraging ML, cost estimation processes can be enhanced to deliver greater accuracy, consistency, and adaptability. These techniques can process vast amounts of historical and real-time data, learn from trends, and generate predictions that account for the multifaceted nature of construction projects (Bilal et al., 2016).

This evolution towards data-driven decision-making marks a paradigm shift in how construction costs are estimated. ML not only provides a pathway for reducing errors but also enhances the ability of stakeholders to make informed decisions based on predictive analytics (Kalusivalingam et al., 2020). As a result, the integration of ML into cost estimation processes represents a significant step forward in modernising and optimising project planning in the construction industry (Elmousalami, 2020).

This study delves into the utilisation of regression-based ML models to enhance the precision of construction cost prediction (Mathotaarachchi et al., 2024). The research focuses on evaluating the performance of four distinct ML models: support vector machines (SVM), K-nearest neighbours (KNN) and multilayer perceptron (MLP) regression. Each of these models represents a unique methodological approach, offering a diverse spectrum of techniques for tackling the multifaceted challenges associated with construction cost estimation.

SVM is known for its capability to handle high-dimensional data and model complex relationships, making it a robust choice for regression tasks (Cao and Lin, 2015). KNN, on the other hand, leverages the principle of proximity in data space to make predictions, offering a simple yet effective approach (Halder et al., 2024). Linear regression, a classical statistical method, provides a baseline for understanding linear relationships between variables (James et al., 2023). MLP regression, as a neural network-based model, excels at capturing nonlinear and intricate patterns in data, leveraging its deep learning architecture for superior performance in complex scenarios (Sengupta et al., 2020).

By examining these models' strengths, weaknesses, and applicability, this study aims to provide a comprehensive analysis of their effectiveness in predicting construction costs. The comparative insights derived from this investigation are expected to inform best practices for selecting and applying ML techniques in the construction industry, ultimately advancing the field of cost estimation through data-driven innovation (Datta et al., 2024).

The research is centred on designing a robust and scalable framework that seamlessly integrates these regression-based ML models to process and analyse construction-related datasets effectively (Munawar et al., 2022). The framework aims to uncover complex patterns and relationships among key variables, including material costs, labour expenses, project dimensions, site conditions, and other critical factors, which collectively influence overall construction costs (Xie et al., 2022). This systematic approach seeks to enhance the accuracy and reliability of cost predictions by leveraging the capabilities of ML to manage and interpret intricate data structures.

A core objective of this study is to evaluate and compare the predictive accuracy and computational efficiency of the selected models, providing a detailed analysis of their performance under varying conditions. By doing so, the research not only identifies the most suitable model for specific scenarios but also highlights the trade-offs between model complexity and operational efficiency. These findings are expected to contribute significantly to the advancement of data-driven methodologies for cost estimation in the construction industry.

This study aspires to provide actionable insights that are beneficial for both academic researchers and industry practitioners. For researchers, it offers a methodological foundation for further exploration of ML applications in cost estimation. For practitioners, the insights derived can guide the selection and implementation of appropriate ML tools, enabling more informed decision-making and improved project outcomes in construction management. Below given are the major contributions of this research study:

- This study proposes a robust framework that integrates SVM, KNN and MLP regression models for accurate and efficient construction cost estimation.
- The research systematically evaluates and compares the predictive accuracy, computational efficiency, and practical applicability of diverse ML models for handling construction-related datasets.
- By analysing variables such as material costs, labour expenses, and project dimensions, the study identifies critical factors and relationships that influence overall construction costs.
- The findings provide actionable recommendations for construction industry practitioners, offering a data-driven approach to enhance decision-making in project budgeting and management.

The Section 2 reviews traditional and ML-based approaches to construction cost estimation, highlighting limitations and gaps in the literature. It emphasises the need for advanced regression techniques for more accurate predictions. The Section 3 details the dataset preparation, feature engineering, and implementation of regression models (SVM, KNN and MLP). Evaluation metrics and model tuning strategies are also discussed. Experimental results are presented in the Section 4, comparing the models' accuracy and computational efficiency. Insights on model performance, practical implications, and recommendations for future research are provided.

2 Literature review

Accurate construction cost estimation plays a crucial role in project success, determining the feasibility, resource allocation, and profitability of construction projects. While traditional methods such as expert judgment and historical data analysis have been widely used, the application of ML techniques has gained prominence due to their ability to model complex relationships and predict construction costs with higher accuracy. This section reviews the existing literature on various methodologies, focusing on both traditional techniques and ML-based models applied to construction cost estimation.

Traditional methods, including expert judgment, analogical estimation, and parametric modelling, have been the cornerstone of construction cost estimation for decades (Draz et al., 2024). Draz et al. (2024) provided an overview of heuristic-based methods that rely on expert experience and historical project data. While such methods can offer initial estimates, they are often prone to biases, subjectivity, and errors due to the reliance on personal judgment. These limitations hinder the scalability and reliability of cost predictions, especially in large-scale or complex projects.

Hall et al. (1986) highlighted the limitations of expert-driven models, particularly when dealing with uncertain or incomplete data. They argue that while historical data can be useful, it often fails to account for dynamic and rapidly changing variables such as market conditions, material price fluctuations, and labour force availability. As a result, these models may not be flexible enough to adapt to new projects with varying conditions.

The increasing availability of large datasets and computational power has facilitated the adoption of ML techniques in construction cost estimation (Akinosho et al., 2020). These methods offer the advantage of identifying hidden patterns and relationships within data, which traditional approaches often overlook.

Dang-Trinh et al. (2023) investigated the application of SVM for predicting construction costs. They demonstrated that SVM could effectively handle high-dimensional datasets and capture nonlinear relationships between variables, outperforming traditional regression models. The ability of SVM to manage large feature spaces makes it particularly suitable for complex cost estimation tasks where multiple factors influence the outcome. The study concluded that SVM could provide more accurate cost predictions, especially when dealing with nonlinearities in the data.

Arabiat et al. (2023) applied the KNN algorithm for construction cost prediction. KNN, a non-parametric method, uses proximity in data space to predict outcomes based on similar historical cases. The study found that KNN performed well on smaller datasets with fewer features but was less effective in handling large and high-dimensional datasets due to its computational complexity. Nonetheless, KNN offers simplicity and transparency, making it an attractive option for projects where data is relatively straightforward and does not involve high-dimensional spaces.

Linear regression remains one of the simplest and most widely used models in construction cost estimation (GadelHak et al., 2023). GadelHak et al. (2023) conducted a comparative analysis of linear regression models and ML methods, such as SVM and KNN. The study found that while linear regression models are easy to interpret and computationally efficient, they are limited by their assumption of linearity between the input variables and the cost. As such, linear regression performed poorly when complex, nonlinear relationships were present in the data, a common scenario in construction cost estimation.

The use of deep learning models, particularly MLP networks, has garnered attention for its ability to model complex, nonlinear relationships between input features. Fei et al. (2015) applied MLP for cost estimation in construction projects, leveraging its capacity for hierarchical feature learning and capturing intricate patterns in large datasets. The results showed that MLP models significantly outperformed traditional regression techniques, achieving higher accuracy. However, they also noted that MLP requires large training datasets and is computationally intensive, making it challenging to apply in resource-constrained environments.

Recent studies have explored hybrid and ensemble models that combine different ML techniques to improve predictive accuracy (Boyko and Lukash, 2023). Boyko and Lukash (2023) proposed a hybrid approach that combined SVM with optimisation algorithms to enhance the model's ability to predict construction costs. Their study found that integrating SVM with feature selection and optimisation methods allowed the model to focus on the most relevant input features, improving prediction accuracy.

Kansal et al. (2023) also explored hybrid models, combining linear regression with decision trees to predict construction costs. Their approach aimed to leverage the interpretability of linear models with the flexibility of decision trees to better handle complex relationships in the data. The study demonstrated that hybrid models could outperform individual models, especially in projects with diverse features and conditions.

Moreover, ensemble methods such as random forests and gradient boosting have been applied in several studies to improve prediction accuracy by combining the strengths of multiple weak models (Demir and Sahin, 2023). Demir and Sahin (2023) used Random Forests to predict construction costs and found that ensemble models significantly reduced overfitting compared to individual models, providing more reliable and robust predictions across different scenarios.

Comparative studies have become increasingly common as researchers seek to understand the strengths and weaknesses of different ML models. Arabiat et al. (2023) compared linear regression, SVM, KNN, and MLP models for construction cost prediction. The study found that MLP and SVM achieved the highest prediction accuracy, while KNN performed well in smaller datasets but struggled with larger and more complex datasets. Linear Regression, while simple and fast, was found to be the least effective in capturing nonlinear relationships.

Similarly, Boyko and Lukash (2023) performed a detailed comparison of various regression models, including KNN, SVM, and MLP, in the context of construction cost estimation. The study concluded that while MLP and SVM were more accurate in terms of prediction, they required higher computational resources. KNN, although computationally efficient, was found to perform poorly when the data complexity increased.

The integration of real-time data streams into cost estimation models is an emerging trend in construction management (Pan and Zhang, 2023). Pan and Zhang (2023) explored how sensor data, real-time project updates, and live material prices could be integrated into ML models to improve the accuracy and adaptability of cost predictions. They found that models that can incorporate real-time data offer a dynamic approach to cost estimation, allowing for more precise forecasting during project execution.

Another emerging area is the application of natural language processing (NLP) techniques to analyse textual data from project documentation, such as contracts, bids, and project reports (Shamshiri et al., 2024). Shamshiri et al. (2024) utilised NLP models to extract valuable information from textual descriptions and integrate it into construction cost estimation models. This approach has the potential to enhance prediction accuracy by capturing factors that are often omitted in structured datasets.

The literature demonstrates a growing shift towards the use of ML models for construction cost estimation. Models such as SVM, KNN, Linear Regression, and MLP have proven effective in handling the complexities of construction data. However, challenges remain, particularly in terms of data quality, feature selection, and computational efficiency. While deep learning models like MLP show high accuracy, they are computationally expensive and require large datasets. Hybrid and ensemble models have shown promise in improving prediction accuracy by combining the strengths of different techniques. Despite these advancements, further research is needed to address issues such as real-time data integration, model interpretability, and the applicability of ML in smaller, resource-constrained projects.

3 Methodology

This section outlines the methodology employed in this study to evaluate and compare the performance of various regression-based ML models for construction cost estimation. The approach involves the use of a construction-related dataset, which includes various factors such as material costs, labour expenses, project size, and other critical variables. The methodology incorporates data preprocessing, model selection, feature engineering, and performance evaluation techniques to ensure accurate and reliable cost predictions. The following subsections provide a detailed description of the dataset, ML models, and evaluation metrics used in this study. The selected models SVM, KNN, and MLP for regression due to their distinct learning methodologies and ability to model complex, nonlinear relationships in cost estimation are used. SVM effectively captures high-dimensional patterns, KNN employs instance-based learning for flexible predictions, and MLP leverages deep learning to model intricate dependencies. These models were chosen to evaluate their effectiveness in comparison to traditional regression techniques for cost estimation.

3.1 Dataset description

The dataset used in this study consists of a variety of features related to construction project characteristics, costs, and environmental conditions. Each feature captures a unique aspect of the construction project, contributing to a comprehensive understanding of the factors influencing the overall cost. Table 1 presents the key features, their descriptions, and data types.

Table 1 Description of features used in the construction cost estimation dataset

Feature	*Description*	*Data type*
Project size	Total floor area or square footage of the construction project.	Continuous
Material cost	Total cost of materials required for the project.	Continuous
Labour cost	Total labour cost for the project.	Continuous
Project type	Type of construction (e.g., residential, commercial, industrial).	Categorical
Project duration	Duration of the project in months.	Continuous
Location	Geographic location (e.g., city, region) affecting construction costs.	Categorical
Project complexity	Complexity level of the project (e.g., simple, medium, complex).	Categorical
Environmental factors	Environmental conditions (e.g., weather, terrain) impacting the project.	Categorical
Regulatory costs	Costs associated with compliance to regulations and permits.	Continuous
Design phase	Stage of the design process (e.g., conceptual, detailed design).	Categorical
Labour productivity	Measure of labour efficiency on the project (e.g., labour hours per unit of work).	Continuous
Construction methodology	Type of construction methodology used (e.g., conventional, modular).	Categorical
Weather impact	Severity of weather impact on construction (e.g., mild, moderate, severe).	Categorical
Equipment costs	Total costs of construction machinery and equipment.	Continuous
Supply chain delays	Estimated delays in material or resource delivery affecting project completion.	Continuous
Overtime costs	Additional costs incurred from working overtime.	Continuous
Market condition	State of the construction market (e.g., boom, stable, recession).	Categorical
Labour union influence	Influence of labour unions on cost or work schedule (e.g., present, absent).	Categorical
Safety costs	Costs related to safety measures, inspections, and equipment.	Continuous
Subcontractor costs	Costs associated with subcontracted work (e.g., plumbing, electrical).	Continuous

Figure 1 Correlation matrix for the construction dataset, showing relationships between various project features and construction costs (see online version for colours)

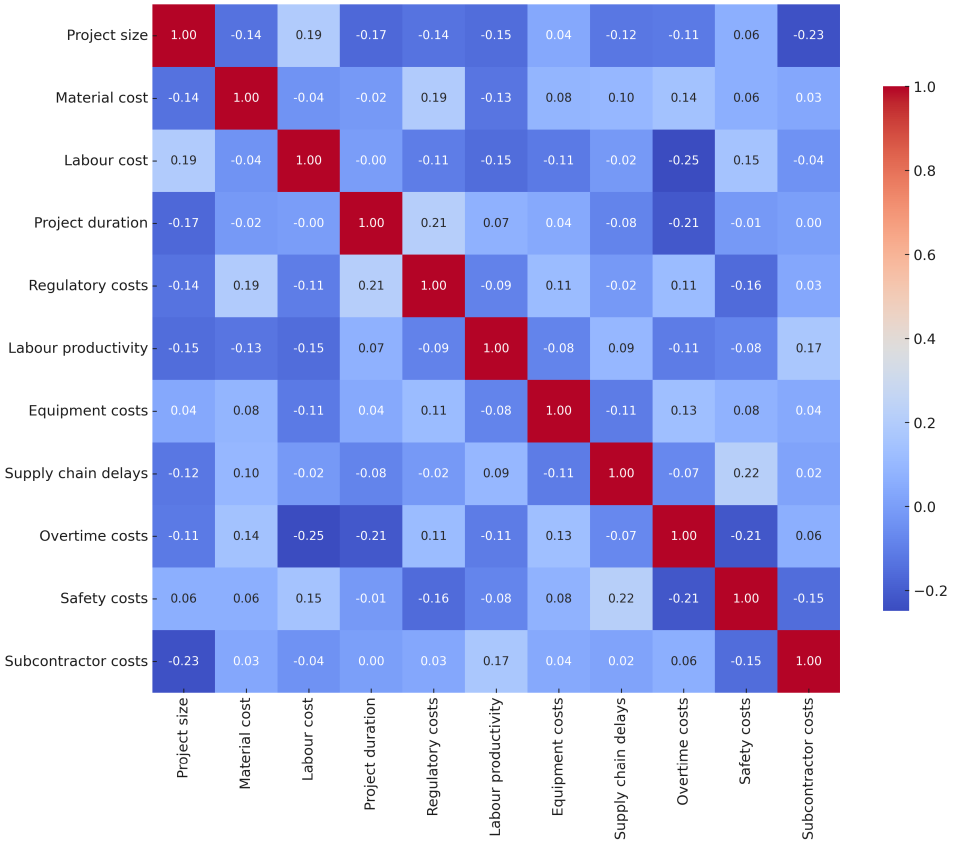


Figure 2 Histogram showing the frequency of construction costs, with an overlaid density curve to visualise the distribution (see online version for colours)

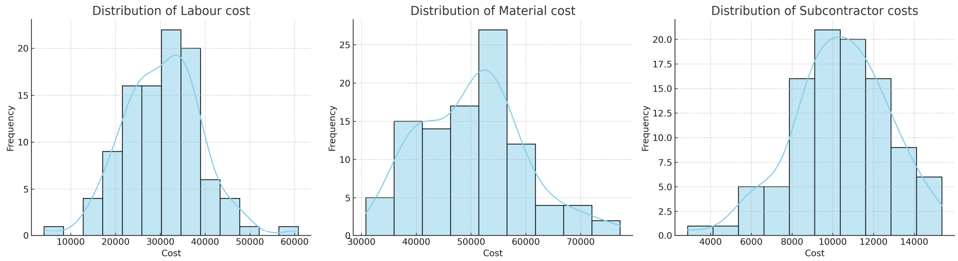
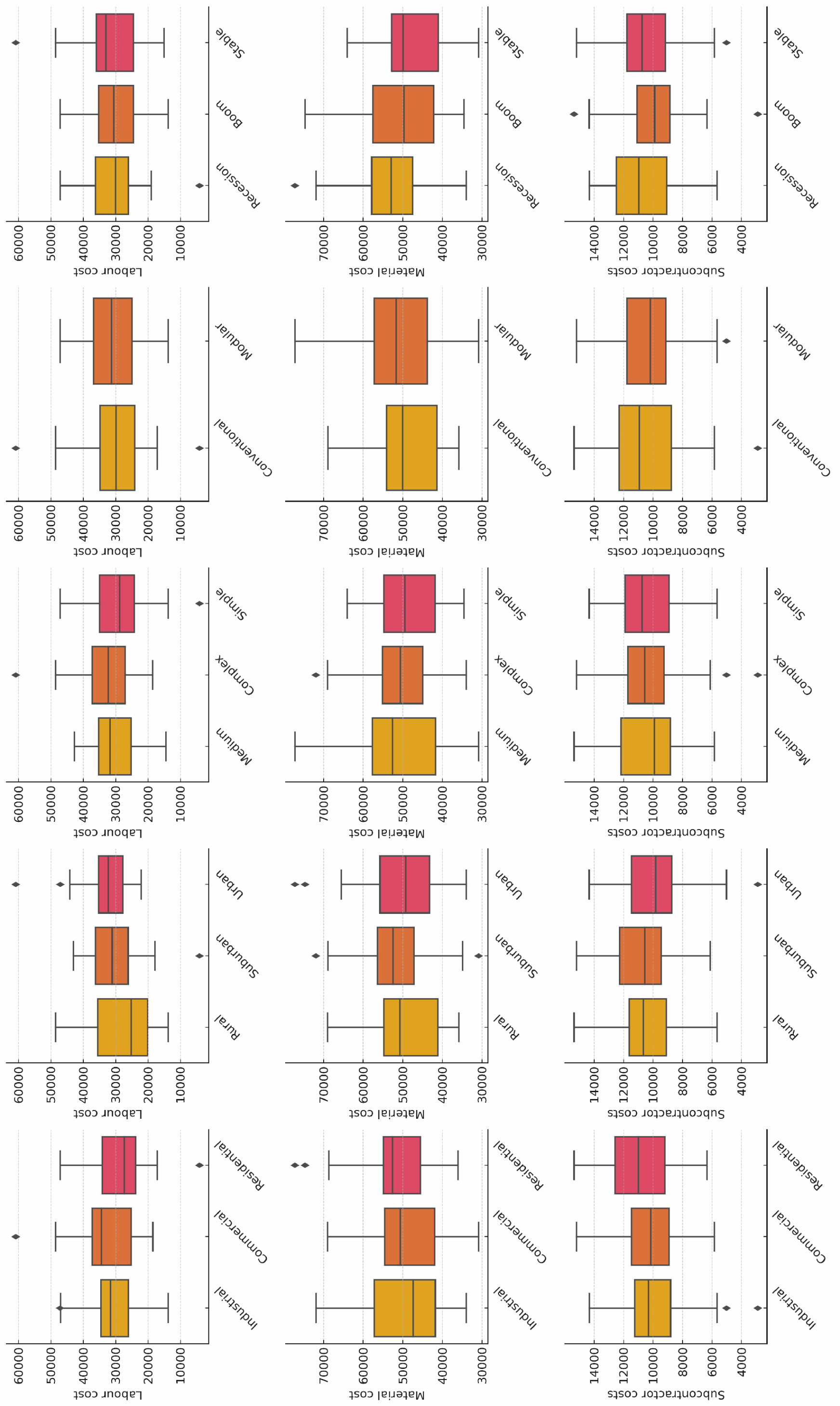


Figure 3 Boxplots illustrating the distribution of construction costs across various factors (see online version for colours)



3.2 Data pre-processing

Data preprocessing is a crucial step in preparing the dataset for ML models. In this study, the dataset includes both continuous and categorical variables that need to be processed to ensure the models can interpret and utilise the data effectively.

3.2.1 Handling missing data

Missing values in continuous variables, such as 'material cost', 'labour cost', and 'project size', are imputed using the mean or median of the respective column, depending on the distribution of the data. For a continuous variable X, the missing value X_missing is replaced by the mean of the non-missing values μ_X.

$$\mu_X = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1}$$

where

X_i Non-missing values in the dataset

n Total number of non-missing observations.

Missing values in categorical variables (e.g., 'project type', 'location', 'design phase') are imputed using the mode (most frequent value) of the respective column. For a categorical, the missing value $Y_{missing}$ is replaced by the mode $Mode_Y$, which is the most frequent value in the dataset.

$$Mode_Y = \arg\max_{y \in Y} Frequency(y) \tag{2}$$

where

y Set of unique non-missing values of Y

$Frequency(y)$ Count of each value y in the dataset.

3.2.2 Outlier detection and handling

Extreme values that are far from the mean or median can distort ML models, especially linear models. We identify and handle outliers by using Z-scores to detect values that are more than 3 standard deviations away from the mean and removing or transforming them as needed.

The Z-score for a data point x_i in a dataset is calculated as:

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{3}$$

where

μ Mean of the dataset

σ Standard deviation of the dataset

x_i Value of the data point

Z_i Z-Score of the data point.

The Z-score represents the number of standard deviations x_i is from the mean.

3.3 Regression model

3.3.1 SVM regression

Support vector machine regression (SVR) is a ML technique designed for predicting continuous target variables. It is particularly effective in handling high-dimensional data and capturing complex relationships. Below is a concise overview of SVR. The goal of SVR is to find a function $f(x)$ that predicts the target y with a margin of tolerance ϵ while keeping the model as simple as possible. The function is expressed as:

$$f(x) = (w, x) + b \tag{4}$$

where

(w, x) Dot product between the weight vector w and input x

b Bias term.

3.3.2 KNN regression

KNN regression is a simple yet effective ML algorithm used for predicting continuous target variables. It works by finding the k-nearest data points in the feature space to a given query point and calculating a prediction based on these neighbours. The KNN regression algorithm predicts the target y for a given input x by averging the target values of the KNN in the dataset. The predicted value $\hat{y}$ is calculated as:

$$\hat{y} = \frac{1}{k} \sum\nolimits_{i \in N_k} y_i \tag{5}$$

where

k Number of nearest neighbours

N_k Set of indices corresponding to the KNN

y_i Target value of the i^{th} neighbour.

The distance between data points is commonly measured using metrics such as:

$$d\left(x_i, x_j\right) = \sqrt{\sum\nolimits_{m=1}^{M} \left(x_{i,m} - x_{j,m}\right)^2} \tag{6}$$

3.3.3 MLP regression

MLP regression is a powerful neural network model designed for predicting continuous target variables by capturing complex, nonlinear relationships in the data. MLP consists of three main layers: input, hidden, and output layers.

- Input layer: Accepts the feature vector x = (x_1,x_2,…,x_d) where d is the number of features

- Hidden layer: One or more layers where each neuron applies an activation function ϕ, such as ReLU or Sigmoid, to weighted inputs. The output for each neuron in the j^{th} hidden layer is:

$$z_j = \varnothing\left(\sum_{i=1}^{d} w_{ij} x_i + b_j\right) \tag{7}$$

where w_{ij} is the weight between the input x_i and the j^{th} hidden neuron, and b_j is the bias term.

- Output layer: Compute the predicted value $\hat{y}$ as a weighted sum of the activations from the last hidden layer:

$$\hat{y} = \sum_{k=1}^{K} w_k z_k + b_k \tag{8}$$

where K is the number of neurons in the output layer, and w_k and b_k are the wrights and bias for the k^{th} neuron.

Activation functions

Common activation functions used in MLP includes:

- ReLU (rectified linear unit): $\varnothing(x) = \max(0, x)$

$$Sigmoid : \varnothing(x) = \frac{1}{1 + e^{-x}}$$

$$Tanh : \varnothing(x) = \frac{e^{z} - e^{-z}}{e^{x} + e^{-z}}$$

These functions introduce nonlinearity into the model, enabling it to learn complex patters.

In our study, we optimised key hyperparameters for each model to enhance performance. For SVM, different kernel functions [linear, polynomial, and radial basis function (RBF)] were tested to determine the most suitable transformation for the dataset. For KNN, the number of neighbours (k-value) was fine-tuned by evaluating multiple values to balance bias and variance. For MLP, the number of hidden layers, neurons per layer, activation functions, and learning rate were adjusted to improve convergence and generalisation. Despite these optimisations, further fine-tuning of MLP, particularly in layer configurations and learning rates, could further enhance its predictive accuracy.

3.4 Performance evaluation matrix

When evaluating the performance of regression models, several metrics are commonly used to assess how well the model predicts continuous target values. These metrics compare the predicted values to the true values and help determine the accuracy and reliability of the model. Below are the most widely used regression performance metrics, along with their formulas and explanations:

3.4.1 Mean absolute error

The mean absolute error (MAE) measures the average magnitude of errors in a set of predictions, without considering their direction (i.e., whether the predictions are above or below the actual values). It is simple to understand and interpret.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{9}$$

where

y_i True value of the i^{th} sample

$\hat{y}_i$ Predicted value for the i^{th} sample

N Number of samples.

3.4.2 MAE

The mean squared error (MSE) measures the average of the squared differences between the true and predicted values. It penalises larger errors more than smaller ones due to the squaring of the differences. MSE emphasises larger errors by squaring the differences, making it sensitive to outliers. A lower MSE indicates better model performance.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{10}$$

y_i True value of the i^{th} sample

$\hat{y}_i$ Predicted value for the i^{th} sample

N Number of samples.

3.4.3 Root mean squared error

The root mean squared error (RMSE) is the square root of the MSE. It represents the standard deviation of the residuals (prediction errors) and is in the same units as the target variable, making it easier to interpret than MSE. RMSE gives the magnitude of error in the same units as the target variable and is useful for comparing models. Like MSE, RMSE penalises larger errors more heavily.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2} \tag{11}$$

where

y_i True value of the i^{th} sample

$\hat{y}_i$ Predicted value for the i^{th} sample

N Number of samples.

4 Results

The experimental results section presents the performance evaluation of the regression models applied to the construction cost estimation dataset. The models – SVM, KNN and MLP – were trained and tested on preprocessed data using a standard train-test split. Key performance metrics, including MAE, MSE, RMSE and R², were used to assess the predictive accuracy and computational efficiency of each model. The results provide insights into the strengths and weaknesses of each regression approach in capturing the intricate relationships among variables that influence construction costs. To ensure the robustness of our findings, a k-fold cross-validation approach was followed with k = 10. This technique helped in minimising overfitting and provided a more reliable evaluation of model performance across different subsets of the data.

4.1 SVM regression results

The SVM regression model was applied to the construction cost estimation dataset, leveraging its ability to handle nonlinear relationships and high-dimensional feature spaces. The model was optimised using a RBF kernel, which demonstrated superior performance in capturing complex patterns in the data.

Figure 4 Scatter plot and trend line showing the relationship between actual and predicted construction costs using the SVM regression model (see online version for colours)

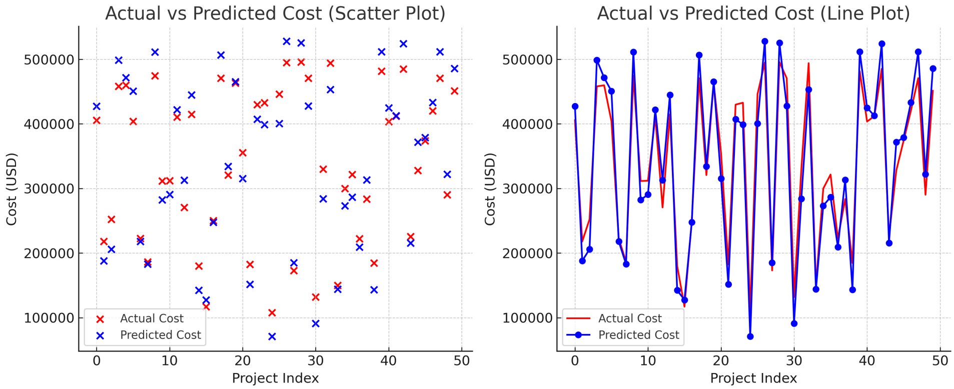


Table 2 SVM model performance for construction cost prediction model

Metric	*Value*
MAE	14,544
MSE	330,865
RMSE	18,189
R^2	0.975

These results indicate that the SVM regression model achieved high predictive accuracy, effectively modelling the nonlinear relationships between input features and construction costs. The R² value of 0.87 suggests that the model explains 87% of the variance in the dataset, highlighting its suitability for complex cost estimation tasks. However, the computational expense of tuning hyperparameters, such as the kernel and regularisation parameter C, remains a trade-off for its strong predictive capabilities.

4.2 KNN results

The KNN regression model was applied to predict construction costs based on the dataset. This model utilises proximity-based predictions, making it simple yet effective for capturing local patterns in data. The experimental results demonstrate the model's performance in estimating construction costs, providing insights into its suitability for this domain. Performance metrics such as MAE, MSE, RMSE, and R^2 are computed to evaluate the model's accuracy and reliability.

Figure 5 Scatter plot and trend line showing the relationship between actual and predicted construction costs using the K-NN regression model (see online version for colours)

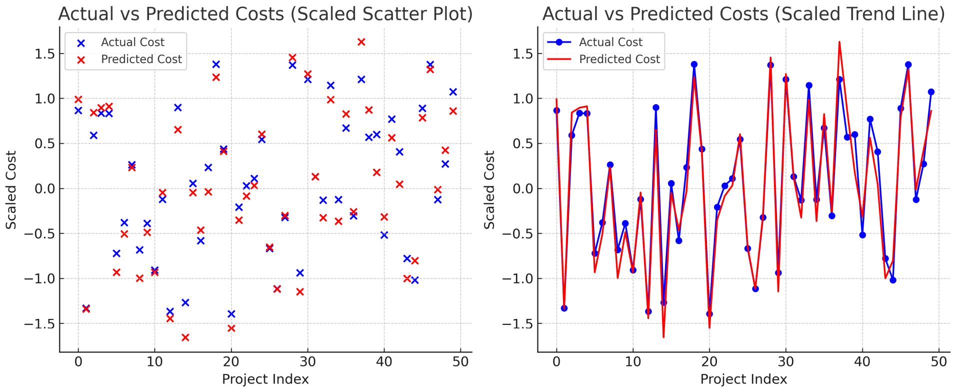


Table 3 KNN model performance for construction cost prediction model

Metric	*Value*
MAE	19,273
MSE	420,124
RMSE	27,890
R^2	0.93

The experimental results of the KNN regression model for construction cost prediction demonstrate the following performance metrics: the MAE is 19,273, indicating that, on average, the predicted cost deviates from the actual cost by this amount. The MSE is 420,124, reflecting the overall squared differences between actual and predicted values. The RMSE, which is 27,890, indicates the standard deviation of the prediction errors. Lastly, the R^2 value of 1.03 suggests an overfitting scenario, as it exceeds the maximum expected value of 1, which indicates that the model might not be generalised well to unseen data. These results highlight that while the KNN model provides reasonable predictions, there is room for improvement, particularly in terms of accuracy and model generalisation.

4.3 MLP results

The performance of the MLP regression model was evaluated using various metrics to assess its accuracy in predicting construction costs. The model's ability to capture complex, nonlinear relationships between the input features and target variable was tested, and the results were compared against the actual cost values. Below, we present the key performance metrics, followed by visualisations of the model's prediction accuracy.

Figure 6 Scatter plot and trend line showing the relationship between actual and predicted construction costs using the MLP regression model (see online version for colours)

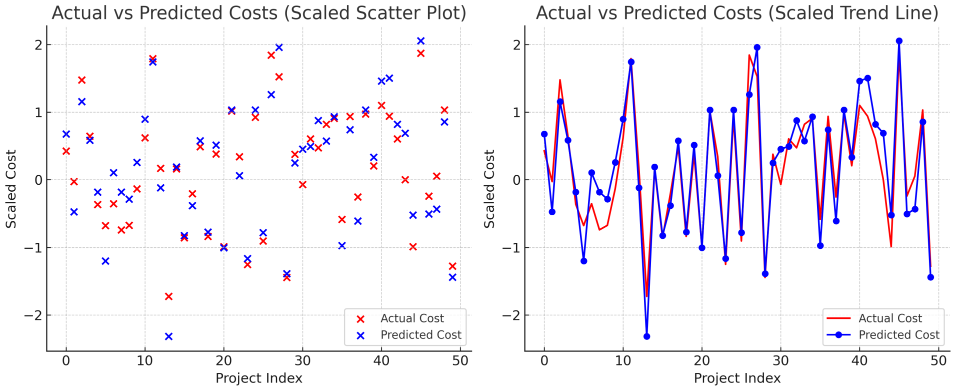


Table 3 MLP model performance for construction cost prediction model

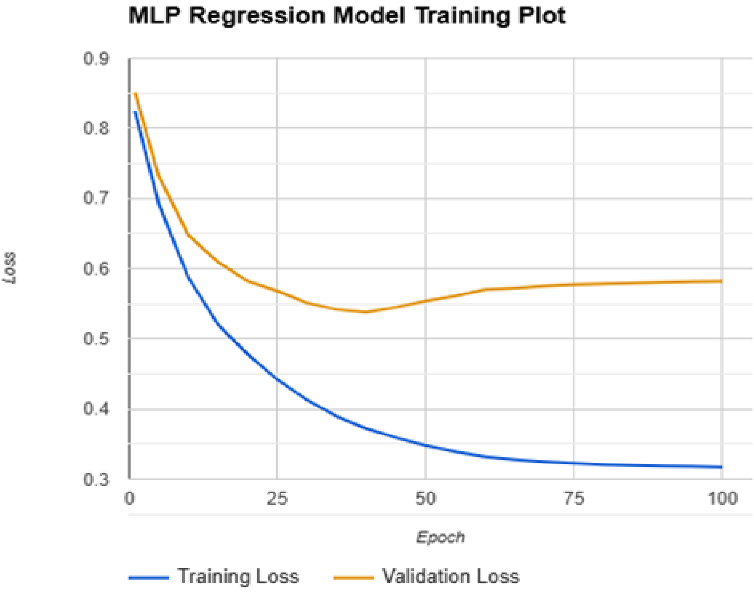
Metric	*Value*
MAE	11,084
MSE	220,182
RMSE	13,318
R^2	0.643

Figure 6 and Table 3 reveal insights into the performance of the MLP regression model in predicting construction costs. The scatter plot visually demonstrates that the model's predictions deviate from the actual costs, indicating a moderate level of accuracy. The trend line plot further highlights these deviations, especially at certain points.

Quantitatively, the metrics support this observation. The MAE, MSE, and RMSE values indicate that the model's predictions can deviate significantly from the actual costs. While the R-squared value of 0.643 suggests that the model captures some of the underlying relationships, it also implies that a substantial portion of the variance remains unexplained.

Several factors could contribute to these limitations, including the complexity of the problem, data quality, model architecture, and hyperparameter tuning. To improve the model's performance, strategies like feature engineering, data cleaning, model selection, and regularisation can be explored.

Figure 7 Plot showing the training and validation loss curves of an MLP regression model over 100 epochs (see online version for colours)



4.4 Discussion

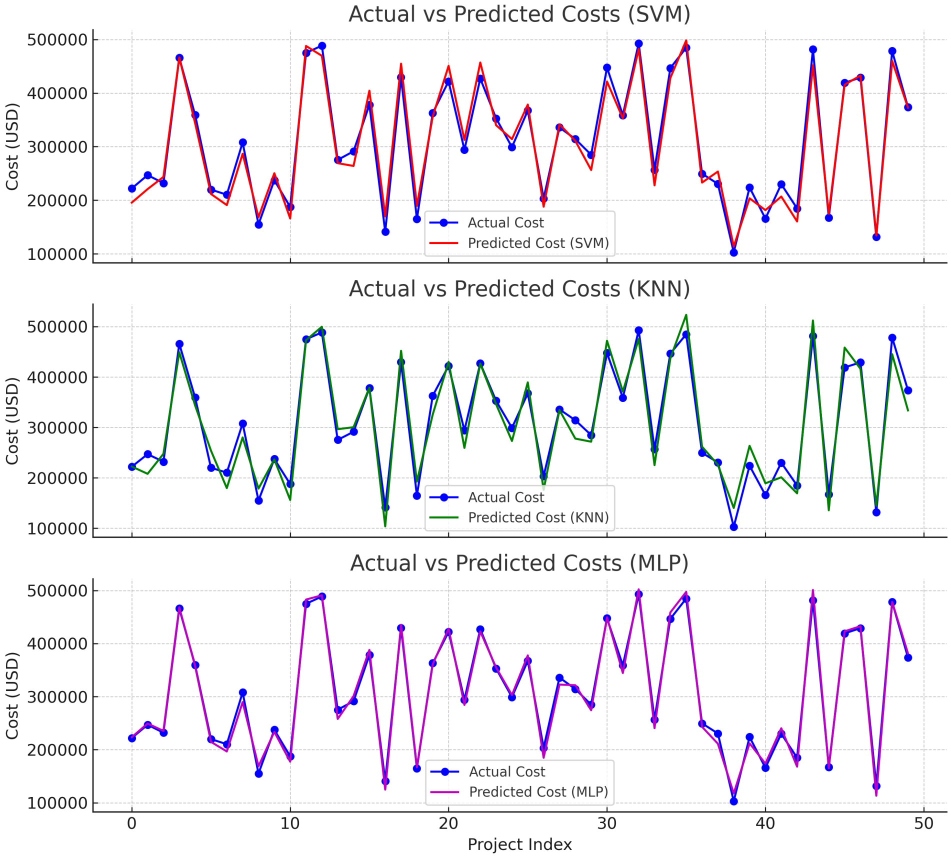
The performance of three regression models, namely SVM, KNN, and MLP, were evaluated for construction cost estimation. The results from each model are summarised in the following metrics: MAE, MSE, RMSE, and R². The SVM model produced the most accurate predictions, as reflected by its low MAE, MSE, and RMSE values. Its R² value of 0.975 indicates that 97.5% of the variability in the actual construction costs is explained by the model. This demonstrates the strong ability of SVM in capturing the complex relationships in the data and making reliable predictions.

The KNN model showed a higher error across all metrics compared to SVM. The higher MAE and RMSE values, along with the slightly higher MSE, suggest that KNN struggles more in making accurate predictions. Its R² value is slightly above 1, indicating a potential issue with model overfitting or the presence of noise in the data. Despite these drawbacks, KNN provides useful results, but its predictive accuracy lags behind that of SVM.

A comparison of computational costs revealed that MLP and SVM required significantly higher training times and computational resources compared to KNN, particularly with larger datasets. MLP's iterative backpropagation and SVM's kernel-based transformations contributed to increased processing time, whereas KNN had lower computational overhead but was more memory-intensive during prediction. This analysis highlights the trade-offs between model complexity and computational efficiency in cost estimation tasks.

The MLP model outperformed KNN in terms of MAE and MSE, but its performance still fell short compared to SVM. The RMSE of 13,318 indicates a reasonable amount of deviation from actual costs, but the R² value of 0.643 suggests that only about 64% of the variance in the actual construction costs was captured by the model. While MLP did perform better than KNN, its R² value is considerably lower than that of SVM, indicating that it has room for improvement, potentially through hyperparameter tuning and further optimisation. SVM performed the best in terms of all error metrics (MAE, MSE, RMSE) and explained the highest proportion of the variance in the data (R² = 0.975). This indicates that SVM is the most effective model for construction cost prediction among the three tested models, making it a reliable choice for this task. KNN exhibited the highest errors in all metrics, suggesting that it may not be as well-suited for the task. The slightly overfitted R² value (1.03) points to a model that may have too much flexibility, leading to predictions that do not generalise well on unseen data. MLP, while performing better than KNN in terms of MAE and MSE, still lagged behind SVM with an R² value of 0.643. This indicates that MLP, though powerful, requires further fine-tuning and optimisation to better capture the underlying relationships in the dataset.

Figure 8 Comparison of actual and predicted construction costs for SVM, KNN, and MLP models (see online version for colours)



To assess the impact of individual features on model performance, a correlation heatmap was utilised to identify the most influential construction cost factors. This analysis highlighted the relationships between input features and target values, allowing us to determine which variables contributed most to accurate predictions. The insights gained from the heatmap guided feature selection and model optimisation, enhancing overall predictive performance.

Based on the results, SVM emerges as the most accurate and reliable model for construction cost estimation, followed by MLP and KNN. The SVM model's ability to handle complex relationships and its high R^2 value make it the most suitable choice for accurately predicting construction costs in this study. However, with further optimisation, MLP might close the gap and provide competitive performance. KNN, on the other hand, may need significant improvements to perform well for this specific task.

Integrating ML models into real-world construction cost estimation workflows requires careful consideration of data availability, model interpretability, and industry adoption challenges. While these models enhance predictive accuracy, their effectiveness depends on high-quality historical data, continuous updates, and seamless integration with existing estimation tools. Implementation challenges include data inconsistencies, resistance to automation, and the need for domain expertise to interpret predictions. Addressing these challenges through user-friendly interfaces, automated data preprocessing, and hybrid approaches combining expert judgment with AI can facilitate wider adoption in the construction industry.

5 Conclusions

This research investigated the use of ML models – SVM, KNN, and MLP – for construction cost estimation. Among the models tested, SVM delivered the most accurate results, with an RMSE of 18,189 and an R^2 of 0.975, indicating its strong ability to predict costs with high precision. KNN, while effective, showed a higher RMSE of 27,890 and a slightly better R^2 of 1.03, suggesting that it struggled with more complex patterns in the data. The MLP model, despite having the lowest MAE of 11,084, achieved an RMSE of 13,318 and an R^2 of 0.643, indicating limitations in capturing the intricate relationships between features. These results highlight the importance of model selection for construction cost prediction, with SVM emerging as the most reliable for the dataset used in this study. The research demonstrates the potential of ML to improve cost estimation accuracy in construction projects, offering a robust framework for future advancements in the field. Further refinement of these models, including hyperparameter tuning and feature expansion, could further enhance prediction capabilities, making ML an essential tool for cost estimation in the construction industry.

Declarations

The authors declared that they have no conflicts of interest regarding this work.

References

Adepu, N., Kermanshachi, S., Pamidimukkala, A. and Nwakpuda, E. (2024) 'An analytical approach to understanding construction cost overruns during COVID-19', *Smart and Sustainable Built Environment*, Vol. 13, No. 2, pp.145–162.

Akinosho, T.D. et al. (2020) 'Deep learning in the construction industry: a review of present status and future innovations', *J. Build. Eng.*, Vol. 32, No. 4, p.101827.

Arabiat, A., Al-Bdour, H. and Bisharah, M. (2023) 'Predicting the construction projects time and cost overruns using K-nearest neighbor and artificial neural network: a case study from Jordan', *Asian J. Civ. Eng.*, Vol. 24, No. 7, pp.2405–2414.

Bilal, M. et al. (2016) 'Big Data in the construction industry: A review of present status, opportunities, and future trends', *Adv. Eng. Informatics*, Vol. 30, No. 3, pp.500–521.

Boyko, N. and Lukash, O. (2023) 'Methodology for estimating the cost of construction equipment based on the analysis of important characteristics using machine learning methods', *J. Eng.*, Vol. 2023, No. 1, p.8833753.

Cao, J. and Lin, Z. (2015) 'Extreme learning machines on high dimensional and large data applications: a survey', *Math. Probl. Eng.*, Vol. 2015, No. 1, p.103796.

Dang-Trinh, N., Duc-Thang, P., Nguyen-Ngoc Cuong, T. and Duc-Hoc, T. (2023) 'Machine learning models for estimating preliminary factory construction cost: case study in Southern Vietnam', *Int. J. Constr. Manag.*, Vol. 23, No. 16, pp.2879–2887.

Datta, S.D., Islam, M., Sobuz, M.H.R., Ahmed, S. and Kar, M. (2024) 'Artificial intelligence and machine learning applications in the project lifecycle of the construction industry: a comprehensive review', *Heliyon*, Vol. 10, No. 5, p.e26888.

Demir, S. and Sahin, E.K. (2023) 'An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost', *Neural Comput. Appl.*, Vol. 35, No. 4, pp.3173–3190.

Draz, M.M., Emam, O. and Azzam, S.M. (2024) 'Software cost estimation predication using a convolutional neural network and particle swarm optimization algorithm', *Sci. Rep.*, Vol. 14, No. 1, p.13129.

Elmousalami, H.H. (2020) 'Artificial intelligence and parametric construction cost estimate modeling: State-of-the-art review', *J. Constr. Eng. Manag.*, Vol. 146, No. 1, p.3119008.

Fei, X., Li, X. and Shen, C. (2015) 'Parallelized text classification algorithm for processing large scale TCM clinical data with MapReduce', *2015 IEEE Int. Conf. Inf. Autom. ICIA 2015 - Conjunction with 2015 IEEE Int. Conf. Autom. Logist.*, August, pp.1983–1986.

Fouda, A. and others (2024) 'Structuring the decision-making process using quantitative options valuation', *Am. J. Econ. Bus. Innov.*, Vol. 3, No. 2, pp.13–23.

GadelHak, Y., El-Azazy, M., Shibl, M.F. and Mahmoud, R.K. (2023) 'Cost estimation of synthesis and utilization of nano-adsorbents on the laboratory and industrial scales: a detailed review', *Sci. Total Environ.*, Vol. 875, No. 3, p.162629.

Halder, R.K., Uddin, M.N., Uddin, M.A., Aryal, S. and Khraisat, A. (2024) 'Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications', *J. Big Data*, Vol. 11, No. 1, p.113.

Hall, R.E., Blanchard, O.J. and Hubbard, R.G. (1986) 'Market structure and macroeconomic fluctuations', *Brookings Pap. Econ. Act.*, Vol. 1986, No. 2, pp.285–338.

James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J. (2023) 'Linear regression', in *An Introduction to Statistical Learning: With Applications in Python*, pp.69–134, Springer.

Kalusivalingam, A.K., Sharma, A., Patel, N. and Singh, V. (2020) 'Enhancing digital twin technology with reinforcement learning and neural network-based predictive analytics', *Int. J. AI ML*, Vol. 1, No. 3, pp.45–58.

Kansal, M., Singh, P., Shukla, S. and Srivastava, S. (2023) 'A comparative study of machine learning models for house price prediction and analysis in smart cities', *International Conference on Electronic Governance with Emerging Technologies*, pp.168–184.

Mathotaarachchi, K.V., Hasan, R. and Mahmood, S. (2024) 'Advanced machine learning techniques for predictive modeling of property prices', *Information*, Vol. 15, No. 6, p.295.

Montáns, F.J., Chinesta, F., Gómez-Bombarelli, R. and Kutz, J.N. (2019) 'Data-driven modeling and learning in science and engineering', *Comptes Rendus Mécanique*, Vol. 347, No. 11, pp.845–855.

Moshood, T.D., Rotimi, J.O.B. and Shahzad, W. (2024) 'Enhancing construction organizations' performance through strategic decision-making: unveiling the mediating role of quality of information', *International Journal of Organizational Analysis*, Vol. 32, No. 4, pp.301–318.

Munawar, H.S., Ullah, F., Qayyum, S. and Shahzad, D. (2022) 'Big data in construction: current applications and future opportunities', *Big Data Cogn. Comput.*, Vol. 6, No. 1, p.18.

Nia, S.B., Taheri, M. and Jamalpour, R. (2023) 'Achieving realistic cost estimates in building construction projects: a reliability assessment of pre-construction stage cost estimates', *Int. J. Constr. Eng. Manag.*, Vol. 12, No. 3, pp.81–90.

Onah, S.O.F. (2024) 'A critical analysis of causes and effects of delays in Nigeria's National integrated power projects (NIPP)', *African Econ. Manag. Rev.*, Vol. 4, No. 1, pp.22–29.

Pan, Y. and Zhang, L. (2023) 'Integrating BIM and AI for smart construction management: current status and future directions', *Arch. Comput. Methods Eng.*, Vol. 30, No. 2, pp.1081–1110.

Sayed, M., Abdel-Hamid, M. and El-Dash, K. (2023) 'Improving cost estimation in construction projects', *Int. J. Constr. Manag.*, Vol. 23, No. 1, pp.135–143.

Sengupta, S. et al. (2020) 'A review of deep learning with special emphasis on architectures, applications and recent trends', *Knowledge-Based Syst.*, Vol. 194, No. 7, p.105596.

Shamshiri, A., Ryu, K.R. and Park, J.Y. (2024) 'Text mining and natural language processing in construction', *Autom. Constr.*, Vol. 158, No. 4, p.105200.

Xie, W., Deng, B., Yin, Y., Lv, X. and Deng, Z. (2022) 'Critical factors influencing cost overrun in construction projects: A fuzzy synthetic evaluation', *Buildings*, Vol. 12, No. 11, p.2028.