ISSN 1392-3730/eISSN 1822-3605

VILNIUS TECH
Vilnius Gediminas Technical University

JOURNAL of CIVIL ENGINEERING and MANAGEMENT

2025
Volume 31
Issue 8
Pages 860–880
https://doi.org/10.3846/jcem.2025.24913

PREDICTION OF COST CONTINGENCY IN CONSTRUCTION PROJECTS BY INTRODUCING MACHINE LEARNING ALGORITHMS

Acinia NINDARTIN[1], Sang-Jun PARK[2], Kyung-Tae LEE[3], Ju-Hyung KIM[2]✉, Susy Fatena ROSTIYANTI[4]

[1]*Samsung C&T, Engineering & Construction Group, Seoul, Republic of Korea*
[2]*Department of Architectural Engineering, Hanyang University, Seoul, Republic of Korea*
[3]*Department of Architectural Engineering, Department of Integrated Energy and Infra System, Kangwon National University, Chuncheon, Republic of Korea*
[4]*Department of Civil Engineering, Institut Teknologi PLN, Jakarta, Indonesia*

Article History:
- received 29 February 2024
- accepted 19 May 2025

Abstract. Construction projects are bound by uncertainties and changes by its nature. Thus, cost contingency needs to be allocated to construction project budget to cope with any deviation of actual costs from planned ones. However, existing methods for predicting cost contingencies, as studied and practiced, still present limitations in reliability and accuracy. Machine learning (ML) has gained popularity for enhancing prediction power in various fields. The paper aims to examine various ML algorithms to implement a cost contingency prediction model, employing both continuous and categorical predictor variables. To develop the model, construction transportation project datasets, which were bid between 2013–2017, were collected from the Florida Department of Transportation (FDOT) website. To address imbalanced regression dataset issues, the synthetic minority over-sampling technique for regression with Gaussian noise (SMOGN) algorithm is introduced. ML random forest (RF) regression associated with random search hyperparameter optimization, achieved remarkably accurate predictions compared to extreme gradient boosting (XGBoost) regression and artificial neural network (ANN) models. The results also demonstrate that four parameters are significant factors in predicting construction cost contingency: project amount, project duration, and latitude and longitude factors. These findings provide new insights for researchers in developing models and for practitioners seeking more advanced method.

Keywords: construction cost contingency, machine learning, RF, XGBoost, hyperparameter optimization, SMOGN, cost prediction.

✉Corresponding author. E-mail: *kcr97jhk@hanyang.ac.kr*

1. Introduction

Each construction project is unique by its very nature. Unexpected situations faced during project delivery elicit changes in various ways. Consequently, the actual costs are likely to deviate from the planned costs. In general, this kind of divergence often leads to cost overruns and has been a common problem in the construction industry (Larsen et al., 2016). In order to hedge the risks from cost deviation, contingency is required and should be assigned before the commencement of construction work (Hoseini et al., 2020a).

Cost contingency refers to the funds required to cover risk situations and is a necessary part of construction work. It covers uncertainty, potential or unforeseeable events, and intangibles that can appear in the future, but this is not a potential profit and does not include major scope changes, escalation, or effects of currency fluctuation (Querns, 1989). For this reason, prediction and management of cost contingency affect project performance because it aims to cover unforeseen costs (Yeo, 1990; Günhan & Arditi, 2007). Because cost contingency is one of the cost components of a base project cost estimate, which is allocated before the commencement of a project, it has a significant impact on project parties (Lhee et al., 2014). From the project owner's perspective, both overestimation and underestimation of the contingency budget could pose issues. Overestimation might restrict funds for other project activities and lead to poor cost management, thereby increasing the chance of project failure (Dey et al., 1994; Hoseini et al., 2020b); in contrast, underestimation can result in ineffective financial performance of the project and insufficient budget for project execution, potentially leading to cost overruns (Baccarini, 2004). In addition,

a method to accurately predict the construction cost contingency is urgently required because it plays an important role as a reserved budget that is used to cover risks or unexpected situations and prevent cost overruns in construction projects. Therefore, an accurate method to estimate the cost contingency at the early stage of the construction project is key to project success.

Previous studies have introduced several approaches to develop contingency prediction models. Although various methods of calculating cost contingency have been developed, overruns still occur and remain a common problem in the construction industry (Love et al., 2014); these methods are also difficult and have limitations in terms of their application (Hamid & Kehinde, 2017). These limitations include difficulty in identifying variable factors, unsuitability for complex projects, low accuracy, independence of cost items, the existence of bias, difficulty in real practice, and employing outmoded approaches. In addition, the accuracy of contemporary methods for determining cost contingency is still uncertain, and the reliability of construction cost contingency estimating tools is unclear even with their extensive development (Baccarini & Love, 2014; Gharaibeh, 2014; De Marco et al., 2016; Hollmann, 2012). Therefore, a robust and advanced approach is necessary to provide convincing construction cost contingency prediction, such as employing machine learning (ML) algorithms, which have gained popularity as prediction applications.

Conventional methods have not been effective in reducing estimation errors, leading to reliance on mathematical models. ML techniques are expected to improve accuracy. ML can be classified into four major types: supervised, unsupervised, semi-supervised, and reinforcement learning. As cost prediction has a continuous target variable, it is categorized as a supervised learning regression problem in ML. Several studies have focused on using ML to predict construction costs. ML has been proficiently applied for the cost prediction of some types of construction projects: support vector machine (SVM) was applied for cost prediction of road construction projects (Peško et al., 2017; Petrusheva et al., 2019); various combinations of artificial neural network (ANN) algorithms have been applied to predict the cost of building projects (Jiang, 2020; Rafiei & Adeli, 2018); and Gaussian process regression (GPR) has been utilized for the cost prediction of tunnel projects (Mahmoodzadeh et al., 2022a, 2022b). Similar to the prediction of the overall construction cost using ML, construction cost contingency could also be predicted using various ML algorithms.

Among various ML approaches in previous studies, ANN has been one of the most commonly used techniques for predicting cost contingency in construction projects. In the prediction task of cost contingency, ANN outperforms the traditional method from a theoretical perspective and can be an effective tool in this area (El-Kholy et al., 2022). Applications in this area mostly relied on ANN, while the possibility of more sophisticated machine learning methods was not been investigated yet. Despite available methods, accurate prediction of cost contingency has continued to be a great need in the field of construction management and artificial intelligence (El-Kholy et al., 2022; Lhee et al., 2016). Thus, it is crucial to explore other ML algorithms. Moreover, ensemble ML algorithms such as random forest and extreme gradient boosting regression have not been widely used in construction cost prediction, in specific, contingency prediction, even though their excellent predictive capabilities have been demonstrated by numerous researchers (Meharie & Shaik, 2020; Yan et al., 2022). On the other hand, although ML has remarkable prediction power, it requires a novel framework for developing accurate models, appropriate feature selection, and enhancing interpretability (Bilal & Oyedele, 2020). To this end, this study aims to propose a model including a data handling process to predict construction cost contingency by utilizing various ML algorithms for improving accuracy. To enhance sustainability and support the development of big data in the construction industry, this study introduced a new and different approach from previous contingency prediction techniques. The implementation of big data in the construction industry is widely adopted such as predictive analytics for cost estimation, real-time monitoring and Internet of Things (IoT) sensors, analyzing energy consumption of the building, quality control and defect detection, data integration with building information modeling (BIM) and so on (Li et al., 2023). The implementation of ML in predicting may be a challenge and various algorithms can be implemented as the solution. Artificial neural networks (ANN) and deep neural networks (DNN) were utilized to solve some non-smooth process to achieve a good and high accuracy results in civil engineering applications (Anitescu et al., 2019; Samaniego et al., 2020). Along with developing a prediction model, another aim of the study is to extract the relevant factors of cost contingency which are significant in planning construction projects and obtainable from normal database used in practice. Both categorical and numerical factors are incorporated in this study, as suggested by previous authors in this area.

We employed consecutive steps for predicting the construction cost contingency. First, in the data collection step, we gathered datasets from the website of the Florida Department of Transportation (FDOT) project, along with identifying various numerical and categorical variables. The second step is the exploratory data analysis (EDA) which is necessary to understand the initial data analysis for correlation and distribution between data, help to understand the data by visualizing the features, detect outliers, and handling missing values. Third, the data pre-processing step is performed by applying the synthetic minority oversampling for regression with Gaussian noise (SMOGN) and feature selection algorithms. SMOGN algorithm was employed to obtain accurate prediction performance by addressing imbalanced dataset and/or insufficient available dataset problems in the regression (Branco et al., 2017;

Wang et al., 2022). Fourth, ML models were developed by dividing the dataset into 80% training and 20% testing sets. To enhance the performance of the ML models, random search hyperparameter optimization along with 10 k-fold cross-validation was implemented. Fifth, analyzing and comparing the performance of various ML models to assess the accuracy using four regression performance metrics: mean absolute error (MAE), coefficient of determination (R^2), root mean square error (RMSE), and mean absolute percentage error (MAPE).

2. Literature review

2.1. Previous studies on cost contingency calculation methods

When methodology is concerned, the previous studies show that the methods used for modeling contingency focused on some categorizations. Several methods for calculating the contingency cost of construction projects have been presented. The Association for the Advancement of Cost Engineering International (2008) categorizes cost contingency estimation and contingency planning techniques for dealing with risks into four main categories: expert judgment, fixed guidelines, analyzing simulation with range estimation and expected value, and parametric modeling. Bakhshi and Touran (2014) classified these methods into three major groups: deterministic methods, which consist of predefined percentages with fixed/line items and expert judgment; probabilistic methods, which are divided into non-simulation methods (e.g., probability tree, first-order second-moment, expected value, program evaluation, and review technique, parametric estimating or regression, analytical hierarchy process, and optimism bias uplifts) and simulation methods (e.g., range estimating and integrated models for cost and schedule); and modern mathematical methods, which consist of fuzzy techniques and artificial neural networks.

Moselhi (1997) introduced the traditional percentage addition which assumes a certain level of risk for the project and determines the percentage of cost contingency based on expert judgment and experience. However, the method implies an unjustified degree of certainty and is hard to justify (Mak et al., 1998; Thompson & Perry, 1992; Hartman, 2000). Famous simulation method such as Monte Carlo simulation is studied by Clark (2001) to evaluate risk and provide a systematic technique for quantifying the contingency value in a construction project. At the same time, he pointed out that this method is difficult, impractical, and uncommonly adopted in the construction industry. Another well-known prediction method is regression analysis. Regression models are an effective statistical tool for analytical and predictive purposes when analyzing the contribution of variables to overall estimate reliability (Kim et al., 2004). Despite that, this method depends on historical cost data, collecting which is time-consuming (Hamid & Kehinde, 2017). In addition, fuzzy techniques and ANN are the mathematical methods used by researchers for predicting contingency. Salah and Moselhi (2015) used fuzzy set theory in the design and developed a contingency modeling framework that incorporates expert opinions. Additionally, Nawar et al. (2018) developed a fuzzy logic-based model that predicts project cost and time contingencies with acceptable validity. Nonetheless, creating fuzzy models can be challenging and requires more fine-tuning, making it difficult to implement them in practice (Hamid & Kehinde, 2017). On the other hand, ANN is one of the machine learning methods which frequently utilized to predict construction contingency in many studies. Chen and Hartman (2000) developed an ANN model that predicts contingency by capturing and learning from historical project samples. Additionally, Lhee et al. (2012) proposed a method that predicts the owner's cost contingency allocation using an ANN model. Furthermore, K. K. Shrestha and P. P. Shrestha (2016) developed a tool system that forecasts the cost contingency of road maintenance contracts by employing an ANN based on historical data. Elkholy et al. (2022) predicted the cost contingency of steel reinforcement in 30 building projects with ANN models. Despite that, the selection of reliable and unbiased inputs as the training data is crucial because it directly impacts the performance of the ANN model (Touran & Lopez, 2006). Table 1 summarizes the purposes, methods, advantages, and limitations of the aforementioned previous studies.

2.2. RF and XGBoost applications in literature

RF and XGBoost have been used extensively for cost prediction in construction management research areas. Zekić-Sušac et al. (2021) proposed models for predicting the energy cost of public buildings using random forest with a large number of predictor variables using RF. In their study, Boruta variable selection was integrated and RF produced a higher accuracy of prediction compared with ANN and classification and regression tree (CART). Shoar et al. (2022) developed an RF regression model to predict engineering services' cost overruns by using 95 high-rise residential building projects database in Iran along with a large number of variables where the R^2 value of 0.868 and MAE of 3.88. Huang and Hsieh (2020) proposed a hybrid model for improving accuracy by integrating RF and simple linear regression for predicting building information modeling (BIM) costs in the construction phase. Meharie and Shaik (2020) used RF for modeling the highway construction cost and found an RMSE value of 0.96. Zheng et al. (2023) combined RF and bird swarm algorithm (BSA) to predict the construction cost in China with the maximum relative error was only 1.24%. Yan et al. (2022) utilized XGBoost to estimate the investment in prefabricated concrete buildings, where the construction project cost-significant and analytic hierarchy process (AHP) was also employed to extract the factors that affect the cost. Compared with other algorithms, XGBoost presented the highest accuracy with a MAPE of 1.00%. Alshboul et al. (2022) conducted a study to predict green building construction costs with various ML algorithms. The results revealed that XGBoost provided the highest accuracy of 0.96.

Table 1. Summary of the features or variables used in this study

Reference	Purpose of study	Method	Advantages	Limitations
Thal et al. (2010)	To discover the important factors that could influence the potential cost contingency in air force construction projects.	Multiple linear regression (MLR)	Clear statistical framework allows others to replicate the method or adapt it for different context.	The validity of the regression results depends on meeting certain assumptions (e.g., linearity, homoscedasticity, independence of errors). If these assumptions are violated, the validity and precision of the estimates may be adversely affected.
Cantarelli et al. (2012)	To analyse the significance of cost overrun performance in various Dutch locations and geographical areas.	Analysis of Variance (ANOVA)	The study finds that the length of the pre-construction phase significantly influences cost overruns.	The models might oversimplify the complexity of cost performance dynamics and fail to capture non-linear relationships or interactions between variables.
Lhee et al. (2012)	To provide a model for estimating the owner's contingency budgeting using ANN and identified the factors that influence contingency.	Artificial Neural Network (ANN)	Potentially leading to more accurate predictions than traditional linear models.	The performance of ANNs depends heavily on the quality and relevance of the input features used. If important variables are omitted or irrelevant ones are included, the accuracy of the model may deteriorate.
Lhee et al. (2014)	To propose a two-step ANN-based method for better predicting optimal contingency in transportation projects compared to current tools.	Two-step Artificial Neural Network	The two-step model separates the estimation process into distinct phases, which can lead to more organized and systematic analysis.	The two-step neural network architecture introduces more complexity compared to single-step ANN models which may increases the risk of overfitting and makes the model harder to interpret or validate.
Arifuzzaman et al. (2022)	To develop a model to predict cost contingency in the early stage with little project information.	Classification and Regression Tree (CART)	The method offers a transparent and interpretable modeling approach and suit for regions that have limitation to access the database.	Small changes in the input data can result in significant changes in the structure of the decision tree which may reduce model stability and reliability.
Salah and Moselhi (2015)	To provide a new fuzzy-set-based model for calculating the cost contingency over the life cycle of construction projects.	Fuzzy set theory	The model is designed to be applicable across different phases of a construction project, from planning to execution.	Constructing a fuzzy inference system such as defining rules, membership functions, and aggregation methods can be complex and time-consuming.
Wang et al. (2016)	To develop a model which can address the hazmat transportation's unpredictable and uncertain issues.	Bayesian network-based	Bayesian Networks effectively address the uncertainties inherent in hazardous materials transportation by modeling probabilistic dependencies among various risk factors.	Bayesian Networks depend substantially on both high-quality data and expert knowledge to construct the network structure and estimate the conditional probabilities between variables.

Elmousalami (2020) developed project conceptual cost models for canal improvement projects. Out of 20 AI and ML algorithms, the XGBoost algorithm presented the most accurate results, where a MAPE of 9.091% and an adjusted R^2 of 0.929. Lathong and Wisaeng (2024) who have proposed a hybrid ML method by combining ANN with Decision Trees (DTs) to enhance construction cost prediction accuracy. Their best model achieved a MAPE of 11% and R^2 of 0.921.

Hyperparameter tuning plays a pivotal role in enhancing machine learning model accuracy. Bergstra and Bengio (2012) argued that Random Search is one of the best hyperparameter for ensembled-ML. Meanwhile, Bayesian Optimization has emerged as a more sophisticated alternative, using probabilistic models to guide the search process. Snoek et al. (2012) demonstrated that Bayesian methods can achieve superior optimization performance with fewer iterations and making them suitable for resource-intensive tasks.

3. Methodology

The methodology applied in this study involved ensemble ML algorithms such as random forest (RF) and extreme gradient boosting (XGBoost) regression, and the implementation of the SMOGN algorithm to predict the cost contingency. SMOGN can address the issue of performance deterioration caused by imbalanced data in re-

gression problems (Branco et al., 2017). Real datasets often suffer from imbalanced distributions (Torgo et al., 2013). Therefore, SMOGN was utilized for over-sampling the rare data points and increasing the robustness of the ML model to estimate the cost contingency of transportation projects.

Figure 1 demonstrates the ML modeling framework of this study. The following are consecutive steps for predicting the construction cost contingency: (1) data collection where the FDOT transportation project datasets from the open website were collected and a thorough understanding of the contingency construction project is required; then, through the existing features of datasets, a comprehensive literature review about the factors influencing cost contingency was conducted, and 13 predictor variables categorical and eight numerical variables) with cost contingency as the target variable was obtained; (2) EDA, which is the process of understanding the correlation between the variables and the distribution of the dataset. This stage also involves data cleaning and removing the missing values in the dataset. After conducting EDA, 814 datasets were obtained and analyzed in the next step; (3) data pre-processing, in which the SMOGN algorithm was used to handle imbalanced data and improve the quality of the dataset. This algorithm changes the number of rows in the dataset from 813 to 780; the categorical variables of the dataset were converted to dummy variables; finally, nine predictor variables were selected after adopting Pearson's correlation, Boruta algorithm, and recursive feature elimination techniques; (4) developing ML model, where the model is built by dividing the dataset into 80% for training and 20% for testing; to enhance the ML model, random search hyperparameter optimization was implemented. Moreover, 10 k-fold cross-validations were also applied. The optimized hyperparameters were derived and the built ML ensemble-based (RF and XGBoost regression) model from the training process can be used for testing datasets. Before choosing RF and XGBoost, various ML algorithms were tested, and this algorithm was found to be the most appropriate for predicting construction cost contingency; (5) ML model performance evaluation, in which the performance of the developed ML models was evaluated and four regression performance metrics (MAE, R^2, RMSE, and MAPE) of the training and testing datasets were compared.

3.1. ML ensemble-based algorithms for cost contingency prediction

ML algorithms can be categorized into single and ensemble methods. Unlike single prediction methods that use only one learning algorithm, ensemble prediction methods integrate multiple prediction models when outputting data. A group of classifiers is built using ensemble methods that categorize new data by weighing the classifier predictions (Dietterich, 2000). In other words, the ensemble learning process involves integrating and applying different learning algorithms. Compared to a single learning algorithm, ensemble learning algorithms have been successfully shown to have better prediction accuracy and can increase generalization (Ghimire et al., 2012; Opitz & Maclin, 1999; Sagi & Rokach, 2018).

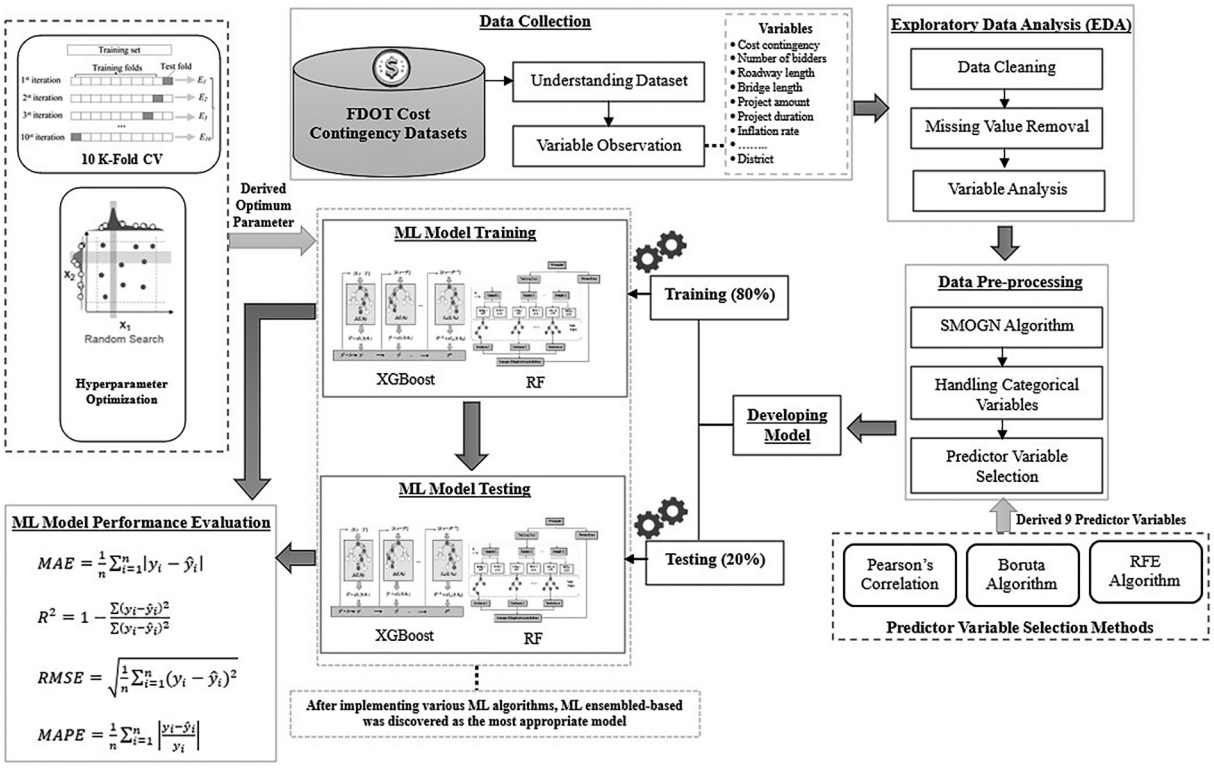

Figure 1. Modeling of the proposed ML framework for construction cost contingency prediction

The ensemble model is different from a single prediction model, which has only one learning model. Ensemble models include multiple base models that are created through various techniques, such as resampling, manipulation, or randomization of the training data, learning algorithms, and learning parameters (Wang & Srinivasan, 2017). Dietterich (2002) and Polikar (2006) asserted that ensemble algorithms increase predictive performance for several reasons. First, they avoid overfitting when the amount of data is small. This was also proven by Cha et al. (2021), who predicted demolition waste using RF and XGBoost machine algorithms with small datasets and categorical variables. Second, ensemble approaches have computational advantages because they reduce the possibility of attaining a local minimum by integrating several learners. Third, integrating various models can expand the search area and lead to a more accurate match with the data space that can represent the optimal hypothesis.

3.1.1. Random forest (RF)

One of the robust ensemble model algorithms based on the classification and regression tree (CART) is RF. RF is an ensemble method based on the bagging technique. Bagging is an abbreviation for bootstrap aggregation, a method of aggregating base learners trained on slightly different training data through bootstrapping (Breiman, 1996). Bootstrapping refers to the process of creating a dataset of the same size as the original dataset by allowing redundancy from the given training data (Hall, 1994). Breiman (2001) developed a more robust RF algorithm that can be applied to regression. Figure 2 shows how the random forest regression works. The RF model predicts outcomes by using the bootstrap resampling technique to generate multiple data from the original data. For each bootstrap sample, a decision tree was constructed, and the predictions from all decision trees were averaged. The model increases the diversity of the decision trees by using a sample with replacement and randomly varies the predictor combinations across multiple tree iterations. An increase in the number of trees can prevent overfitting and is less impacted by outliers. Moreover, there are two crucial former parameters of RF: the number of regression trees (N estimators) and the maximum depth of node random variables (Zhou et al., 2019).

The steps in developing the RF model are as follows: (1) use the original data to create *ntree* bootstrap samples; (2) for each bootstrap dataset, a tree was grown; at each node of the tree, a random subset of features *mtry* was used to determine the best split and grow the tree to make each terminal node have *nodesize* cases; (3) aggregate information from *ntree* trees to predict new data; for example, perform a majority vote for classification; and (4) use the data not included in the bootstrap sample to obtain an out-of-bag (OOB) error rate. Creating a regression tree for each bootstrap training set involves the following procedure. The next step involves generating a regression tree for each bootstrap training set. N estimators' regression trees are created, forming a "*forest*" without pruning. During the growth process of each tree, not all optimal attributes are selected as internal nodes for branching. Instead, the optimal attribute is chosen from the randomly selected maximum depth attributes for branching. This increases the difference between the regression models by constructing different training sets, thereby enhancing the prediction performance of the combined regression model. A regression model sequence $\{t_1(x), t_2(x), \ldots, t_k(x)\}$ is obtained by *n*-time model training, which is then utilized to create a multi-regression model system (forest). Then, the predictions made by the regression tree of the N estimators are compiled, and a simple average approach is used to determine the value of the new sample. Eqn (1) below is the final regression decision equation:

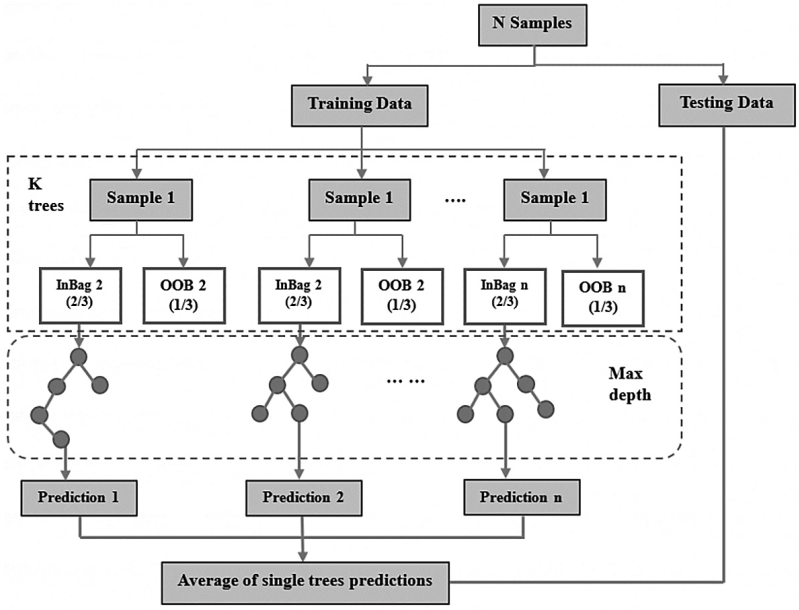

Figure 2. Schematic diagram of the random forest regression

$$\hat{f}_{rf}^{K}(x)=\frac{1}{K}\sum_{k=1}^{K}t_i(x), \tag{1}$$

where K is the number of regression trees, t_i is a single decision tree regression model, and $\hat{f}_{rf}^{K}(x)$ represents the combined regression model (N estimators). The estimated value is the weighted sum of the regression outcomes of each tree. Therefore, as RF is recognized as one of the best classifiers and as this algorithm can handle the imbalance class, RF provides a powerful prediction compared to other ML algorithms (Fernández-Delgado et al., 2014; Breiman, 2001).

3.1.2. Extreme gradient boosting (XGBoost)

Gradient tree boosting is a ML method that stands out in numerous applications, among other techniques utilized in practice. A large-scale ML method for boosting trees is extreme gradient boosting (XGBoost), which is an advanced supervised learning algorithm (Chen & Guestrin, 2016). XGBoost has gained broad recognition in various fields (Wang et al., 2020) of ML and data mining because of its superior performance and outstanding results with only a small amount of data (Bekkerman, 2015). Moreover, Chen and Guestrin (2016) verified the benefits of this algorithm. First, regularization in the algorithm introduced by XGBoost has the ability to handle overfitting by offering row and column sampling. Second, the algorithm used by the model is integrated and based on gradient lifting decision tree optimization, which may satisfy both construction and performance criteria. Third, the model can record the significance of characteristic indices through tree nodes and has a high interpretability.

XGBoost works on the principle of applying a greedy strategy to learn individual base trees to address regression problems, and the advanced framework of the gradient-boosted regression trees (GBRT) model, is shown in Figure 3. To enhance the precision of the predictions, new decision trees are continuously constructed to fit the residuals of the previous prediction. This technique helps minimize the difference between the predicted and actual values. Chen and Guestrin (2016) claimed that the XGBoost algorithm adds a regularization component, expressed by $\Omega(\theta_j)$, to the standard loss function to avoid model overfitting. The final prediction of the XGBoost model can be defined using Eqns (2) and (3), respectively:

$$L=\sum_i l(\hat{y}_i,y_i)+\sum_j \Omega(\theta_j); \tag{2}$$

$$\sum_j \Omega(\theta_j)=\gamma T_j+\frac{1}{2}\lambda\sum_{k=1}^{T_j}\left[w_k^{(j)}\right]^2, \tag{3}$$

where the loss function is denoted as l; the predicted value is $\hat{y}$, and the actual value is y. θ_j controls the structure of the j-th tree; γ is the minimum loss reduction required to process node partition in the regression tree, λ is the regularization of the weight of leaves in the regression tree, T_j is the number of leaves in the j-th regression tree, and $w_k^{(j)}$ is the weight of the k-th leaf in the j-th regression tree. It has been proved that a larger T_j will reduce the objective function but will be penalized by a larger factor γ.

3.1.3. Synthetic minority over-sampling for regression with Gaussian noise (SMOGN)

Torgo et al. (2013) introduced the synthetic minority oversampling technique for regression (SMOTER) algorithm to address the issue of performance deterioration caused by imbalanced data. This approach can balance rare and the most frequent instances by altering the distribution of a given training dataset. To address imbalanced regression issues, where crucial user cases are underrepresented in the available studies, Branco et al. (2017) introduced Gaussian noise to SMOTER, thus creating SMOGN. Additionally, Zhu et al. (2021) employed SMOGN in data preprocessing to predict the rockhead position based on limited borehole data. The SMOGN, using the SMOTE algorithm, can only produce new syntactic instances when the seed example and the chosen k-nearest neighbors (KNN) are sufficiently close. However, when the two examples are "further distant", Gaussian noise is introduced. In Figure 4, the main principle of SMOGN is to create new synthetic samples using the five seed case nearest neighbors, which are assumed to have comparable cost contingency and attributes (such as project amount and duration). In addition, the SMOGN algorithm works in two areas: safe and unsafe. If the selected neighbor is safe, it means that it is within a suitable distance for SMOTE to perform interpolation. Otherwise, if the neighbor is far, it is better to generate a new example with Gaussian noise on the seed case.

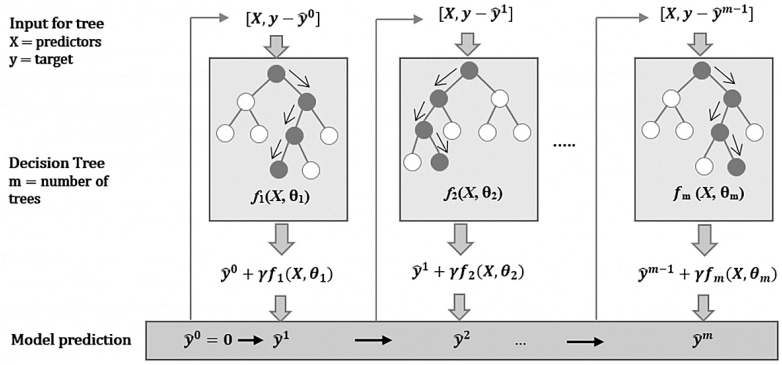

Figure 3. Schematic diagram of the gradient-boosted regression tree

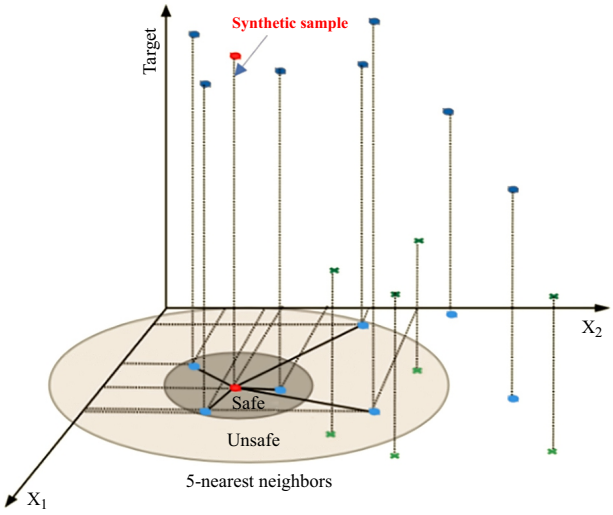

Figure 4. Example of the SMOGN algorithm

3.2. Dataset summary and exploratory data analysis (EDA)

This study established a framework to predict the cost contingency in transportation construction projects using ML. The dataset was obtained from the FDOT website at https://www.fdot.gov/ which has 1277 transportation projects. After sorting and grouping, the dataset used in this study was limited to the transportation projects that were constructed in the 2013–2017 bid year. The project from the 2013–2017 bid year dataset was the only available dataset containing the necessary input features for the analysis, and there were difficulties in collecting sensitive data due to limited access on the website. FDOT is responsible for managing, regulating, and maintaining public transportation throughout Florida through a decentralized agency. The site provides considerable transportation project data with various other information as the features (variables) and real cost contingency as the dependent variable. The collected raw dataset had 13 independent variables and 813 observations (rows) with a mixture of categorical and numerical variables. Python version 3.6.5 running on Jupyter Notebook was utilized to check the data type of each variable, dtypes, and the built-in function of pandas.

EDA plays an important role in the data analysis process. This step is required to review the characteristics of our datasets and is typically used to develop hypotheses rather than to reach definitive conclusions based on the findings of the study (Abt, 1987). Before designing the ML cost contingency prediction model, the overall FDOT transportation dataset used in this study was explored. The first step in the data analysis process is to visualize the data. This is a pivotal process because various characteristics such as patterns, outliers, changes over time, and relationships between variables can be observed through this process. When designing a prediction model, it is vital to include features found as graphs that visually represent data because, depending on the data type and characteristics, it is possible to make decisions about which prediction model is suitable and which graph is appropriate for further analysis. It was found that the greater the number of bidders, the higher the cost contingency, as the number of bidders may show the level of competition between contractors. In addition, new transportation construction projects have the highest contingency cost. Conversely, traffic operation projects were found to be the lowest. From the perspective of the type of contract, design-bid-build (DBB) projects have a high-cost contingency compared to design-build (DB) projects. It was also found that the distribution of the target variable (cost contingency) and the imbalanced distribution are strongly affected by outliers in the 813 datasets. There is a high probability that problems will occur in the ML process because of the imbalanced dataset (Kaur et al., 2019). Therefore, to solve this problem, we implemented the SMOGN algorithm in the preprocessing step to deal with imbalanced regression data issues (Branco et al., 2017, 2019; Kunz, 2020).

3.3. Model variables

The process of developing models involves predictors or features as independent variables in building ML algorithms. Thoroughly identifying and understanding the factors that affect the dependent variable (cost contingency) can enhance the prediction accuracy. Therefore, developing cost contingency prediction methods requires a comprehensive literature review to identify potential predictor variables. In the raw dataset of this study, 13 predictor variables (features) with one response variable (target) were considered. The predictor variables found and used in this study were project amount, project duration, roadway length, bridge length, project type, contract type, number of bidders, latitude, longitude, area classification, district, weather conditions, and inflation rate. A summary of the predictor variables used in this study is provided in Table 2.

In the construction industry, several factors must be considered when determining contingency costs. Given the unique nature and varying characteristics of construction projects (Manu et al., 2010), cost contingencies differ across projects. Flyvbjerg et al. (2002) proved that different project types and geographic locations have statistical significance in determining the cost contingency in infrastructure projects. This is also similar to the study by Cantarelli et al. (2012), who discovered that the sum of the contingency cost of infrastructure projects in the Netherlands varies based on factors such as location, geographical area, and project type. According to their findings, rail projects have an 11% lower cost contingency than other project types (road, tunnel, and bridge projects). Moreover, the authors found that projects in North–West European countries experienced various cost contingencies compared to those in other geographical areas.

A broad understanding of the factors that impact the cost contingency amount will help construction project parties during the lifecycle of the project. Laryea and

Table 2. Summary of the features or variables used in this study

Variables or Features	Description	Type	Measurement	References
Real contingency	The real cost contingency of construction transportation projects	Continuous	105, ..., 624195 (US$)	The target or response variable
Number of bidders	The number of bidders or contractors who have participated in the projects	Continuous	1, 2, ..., 14	Lhee et al. (2012, 2014), Smith and Bohn (1999)
Roadway length	The length of the roadway	Continuous	0, ..., 86.01 (miles)	Mahamid (2013), Thal et al. (2010)
Bridge length	The length of the bridge	Continuous	0, ..., 2.462 (miles)	Wang et al. (2016), Thal et al. (2010)
Project amount	The total cost of the project estimated by FDOT	Continuous	41651, ..., 90881467 (US$)	Mahamid (2013), Espinoza (2011), Arifuzzaman et al. (2022), Lhee et al. (2012, 2014), Chan and Au (2008), El-Touny et al. (2014), Thal et al. (2010)
Project duration	The total duration of the project estimated by FDOT	Continuous	21, ..., 1400 (days)	Espinoza (2011), Lhee et al. (2012, 2014), Cantarelli et al. (2012), Günhan and Arditi (2007), Chan and Au (2008), El-Touny et al. (2014)
Inflation rate	The inflation rate at the time of bidding project in percent	Continuous	–0.2, ..., 2.738 (%)	Ammar et al. (2022), Arifuzzaman et al. (2022), Wang and Chou (2003), El-Touny et al. (2014), Asamoah et al. (2023)
Contract type	The type of contract procurement of projects that are divided into design-bid-build (DBB) and design-build (DB)	Categorical	Design-bid-build (DBB) and design-build (DB)	Arifuzzaman et al. (2022), Sonmez et al. (2007), Lhee (2014)
Weather condition	The general weather conditions of the area where the project was constructed, classified into the worst to most favorable conditions	Categorical	Worst, bad, fair, good, and favorable	Wang and Chou (2003), El-Touny et al. (2014), Chen and Hartman (2000)
Project type	The type of transportation projects is classified into 10 types.	Categorical	Resurfacing, reconstruction, widening and resurfacing, new construction, bridge construction, bridge repair, interstate rehabilitation, traffic operations, miscellaneous construction, and other transportation projects	Salah and Moselhi (2015), Arifuzzaman et al. (2022), Cantarelli et al. (2012), Flyvbjerg et al. (2002), Thal et al. (2010), Lhee (2014)
Latitude	A coordinate that specifies the North-South position, measured in degrees relative to the equator	Continuous	243315, ..., 650001 (degrees)	Arifuzzaman et al. (2022), Cantarelli et al. (2012), Flyvbjerg et al. (2002), El-Touny et al. (2014)
Longitude	A coordinate that specifies the East-West position, measured in degrees relative to the Prime Meridian	Continuous	800249, ..., 1650001 (degrees)	Arifuzzaman et al. (2022), Cantarelli et al. (2012), Flyvbjerg et al. (2002), El-Touny et al. (2014)
Area classification	The classification area where the projects were constructed: urban or rural	Categorical	Urban (U) and rural (R)	Lhee (2014), Cantarelli et al. (2012), Flyvbjerg et al. (2002), El-Touny et al. (2014)
District	The district area where the projects were located	Categorical	D1, D2, D3, D4, D5, D6, and D7	Arifuzzaman et al. (2022), Cantarelli et al. (2012), El-Touny et al. (2014)

Hughes (2009) stated that the main factors of cost contingency are the total project amount, level of competition, duration of the project, clarity of bid documents, inflation, weather conditions, and punctuality of payment from the project's owner. Hoseini et al. (2020b) explained that technical, economic, psychological, and political factors influence the amount of cost contingency. In their study, technical factors included inaccuracies in the estimation approach, insufficient data, and the lack of experience of the project team. Economic factors involve the economic interests of the parties that will choose the project and project promoters who may purposely underestimate the proj-

ect cost. Misconception planning and optimism bias are two psychological factors that can cause a project's scope to be underestimated or overestimated for organizational benefits. Political factors include strategic misrepresentation through intentional and strategic cost estimation when estimating project outcomes. Catalão et al. (2019) stated that endogenous and exogenous factors can affect cost contingency, where endogenous factors are project characteristics over which the project team has control and exogenous factors are external ones such as economic, political, and environmental factors that may affect the cost of the project.

The size or amount of a construction project influences its contingency. Karlsen and Lereim (2005) examine the management of cost uncertainty in engineering projects depends on base estimate, contingency, and allowance where they highlighted that project managers should control these risk reserves. In addition, bidding factors, such as the level of competition of the project, also affect the budget of contingency determination. The number of bidders indicates the level of competition between the contractors who bid on the project. The higher the number of bidders, the higher the contingency allocation because of the high workload. Additionally, Ammar et al. (2025) recognizing the factor prior to the bidding stage is crucial to estimate appropriate contingency amounts.

Ameh et al. (2010) identified 42 factors that may cause cost contingencies in Nigeria's telecommunication projects. The study revealed that the lack of contractor experience, soaring prices of imported materials, and variations in the prices of materials are major factors when considering the budget for contingency. Kasimu (2012) also analyzed 41 risk factors that are significant for the determination of cost contingency. The major factors include changes in material prices, underestimation, lack of project management, and additional costs of reworks. Because cost contingency is an important component of a contractor's bid estimate, Enshassi and Ayyash (2014) classified the factors that influence the cost contingency amount from the contractor's perspective into 12 groups based on factor characteristics and source. The groups of factors are project-related, design-related, construction-related, bidding-related, contractor-related, owner- or consultant-related, resource-related, environmental, legal, economic, technical or managerial, and political. Moreover, the authors highlighted that natural and environmental risks are the hardest to foresee and detect, but when these risks occur, they have a large impact that drives the need for cost contingency plans. In contrast, from the perspective of quantity surveyors, environmental factors (such as weather and ground conditions) and economic point-of-view factors (such as inflation rate and cash flow) are the most influential factors in cost contingency determination for building construction projects in Ghana (Asamoah et al., 2023).

3.4. Data pre-processing

3.4.1. Solving the issue of imbalanced data using SMOGN

In this study, SMOGN was utilized for over-sampling rare data points and increasing the robustness of the ML model to estimate the cost contingency of transportation projects. Moreover, in this study, the imbalanced learning regression Python package, as detailed in Table 3, was used to develop the SMOGN algorithm (Wu et al., 2022). After conducting SMOGN, the number of observations (rows) in the original dataset decreased from 813 to 780. Additionally, in the results of both the original and modified datasets, the distribution of the response variable in the modified dataset becomes more evident and outliers are reduced by SMOGN, as shown in Figure 5.

3.4.2. Handling categorical variables

Some categorical variables in the dataset of this study were classified as object data types. After checking the data type with Python, project types, contract types, weather conditions, districts, and area classification features were categorized as object data types or categorical variables. To deal with these variables and to make subsequent ML analysis easier, the "*pd.get_dummies*" function of the Python Pandas library was applied to obtain vector-dimensional values for 0 s and 1 s. Dummy coding is the preferred method when comparing multiple treatment groups with a control group (Myers et al., 2010). Moreover, compared with the effects of coding, dummy coding can be set up more quickly and easily (Daly et al., 2016). Thus, this method was chosen to handle categorical variables in this study.

Table 3. SMOGN input arguments in this study

Argument	Explanation	Input Argument
data	A Pandas DataFrame that is passed as the "data" input includes the training set split	Transportation
y	A string that identifies a continuous target variable by header name is accepted as the "y" argument	"Real Contingency_Target"
samp_method	When "extreme" is given in the "samp_method" argument, it means that over-sampling is performed	"extreme"
replace	The Boolean argument is required for "replace". Replace the sampling if "True" is input	True
rel_thres	The value between 0 and 1 is required for the "rel_thres" argument. It defines the rarity threshold. The over-sampling boundary increases in height with the increased threshold. Conversely, the threshold decreases with a decrease in over-sampling	0.7
pert	A number between 0 and 1 is required for the "pert" argument. It shows how much noise should be perturbed while adding Gaussian Noise	0.08

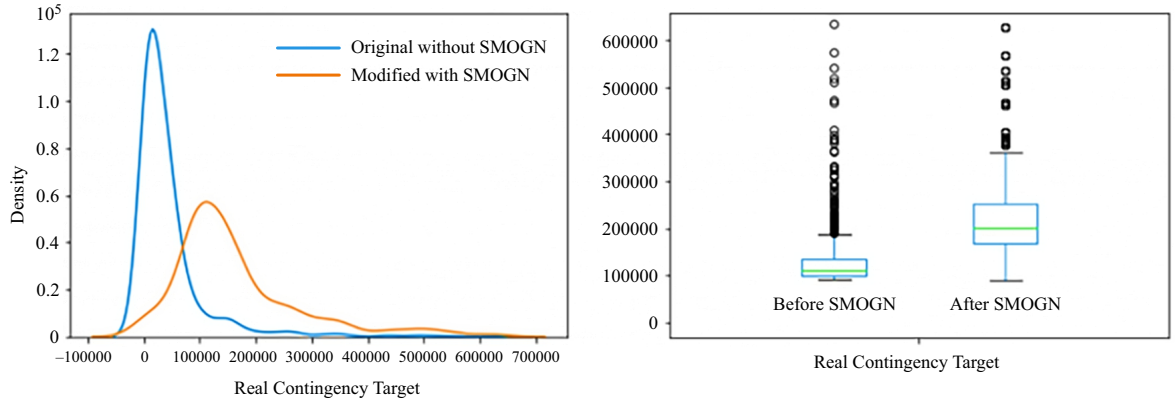

Figure 5. Distribution target variable comparison before and after SMOGN

3.4.3. Selection of predictor variables

The complexity of a model can be impacted by a large number of predictor variables, which can lead to overfitting (Alpaydin, 2020). Hence, when modeling with ML algorithms, it is crucial to select the most influential predictor variables before developing models to increase the prediction accuracy. This can be achieved through predictor or feature selection approaches (Guyon & Elisseeff, 2003). Wrapper and filter algorithms are classified as feature selection tools in ML. Wrapper algorithms use a ML technique to train on a dataset and learn from it, choosing the best subset based on accuracy. In contrast, without considering learning bias, filter algorithms employ raw data taken from the dataset to identify key variables.

Both the filter and wrapper algorithms were implemented to determine the independent variables using the Boruta algorithm, recursive feature elimination (RFE), and Pearson's correlation. Boruta is a wrapper RF algorithm that can be run quickly, even without parameter tuning (Kursa & Rudnicki, 2010). Boruta has advantages such as dealing with nonlinear variables and performing at a high computational speed (Cao et al., 2018). Additionally, RFE is a wrapper algorithm that recursively removes features and builds models based on the remaining ones, using accuracy to identify the most predictive features and feature combinations (Artur, 2021). In contrast, Pearson's correlation is a filter algorithm that measures the correlation between variables and how much they affect each other (Shardlow, 2016). Table 4 presents the results of Boruta and RFE. Rankings other than 1 or *False* mean that these features do not have a major influence on cost contingency (the dependent variable). Pearson's correlation between the predictor variables toward the cost contingency shows that project amount and project duration may have a significant effect. The bridge length, weather conditions, area classification, and district variables were not selected for ML model training in this study.

4. Results

4.1. Model establishment

The data used for developing the ML models comprised 780 rows, with cost contingency as the target variable and nine predictor variables. Before developing the model, it

Table 4. Results of feature selection algorithms

No.	Variables or features	Boruta algorithm		RFE algorithm		Pearson's correlation (toward the cost contingency)
		Rank	Keep	Rank	Keep	
1	Project amount	1	True	5	False	0.347
2	Project duration	1	True	2	False	0.379
3	Roadway length	1	True	1	True	–0.021
4	Bridge length*	2	False	7	False	0.037
5	Number of bidders	1	True	1	True	0.048
6	Inflation rate	1	True	1	True	–0.001
7	Latitude	1	True	3	False	0.025
8	Longitude	1	True	4	False	0.011
9	Contract type	7	False	1	True	0.122
10	Weather condition*	5	False	6	False	–0.041
11	Project type	3	False	1	True	–0.008
12	Area classification*	6	False	8	False	0.083
13	District*	4	False	9	False	–0.001

Note: *The features which were not included in the final ML model.

is crucial to separate the data into training and testing sets to prevent overfitting. First, the data were divided into 80:20 sizes (624 data for training and 156 data for testing), and the random size value was 42 using the *scikit-learn* library. This ratio was chosen because 80% of the training dataset has been empirically proven to increase the performance of ML models (Gholamy et al., 2018).

ML algorithms require specific input parameters. The ML model development in this study used the default parameters provided by Python Scikit-Learn. To achieve the optimum values of parameters, hyperparameter optimization is required. Hyperparameter optimization is a process for determining the best parameters or configurations for an ML model. This is a critical task in ML because model performance highly depends on the parameters used. The hyperparameter is determined before training begins and is not updated during training. The purpose of hyperparameter optimization is to maximize the model performance in data testing or validation. This method can improve model performance by finding better parameters than those generated manually or randomly. Moreover, it can reduce overfitting by identifying the most common parameters in the training data (Agrawal, 2021). Therefore, random search hyperparameter optimization was used in this study.

Random search hyperparameter optimization is an algorithm that attempts a random combination of parameters from a certain range. Bergstra and Bengio (2012) found that within a very small fraction of the calculation time, a random search over the same domain can generate models that are as good as or better than other optimization methods. This method was chosen because random search optimization is robust hyperparameter optimization. Chakraborty and Elzarka (2019) applied and proved that random search optimization is genuinely effective as hyperparameter optimization to predict the energy consumption of buildings with XGBoost, ANN, and degree-day-based ordinary least square regression models. In this study, random search was set to develop and maximize the ML models with 10 cross-validations as the default value (Chakraborty & Elzarka, 2019).

Hyperparameter tuning for each machine learning model to optimize predictive performance were performed. Random search was used for models with a relatively small hyperparameter space, allowing for efficient exploration of a wide range of values and focus more broadly (Bergstra & Bengio, 2012). On the other hand, Bayesian hyperparameter builds a probabilistic model which then chooses the next hyperparameter set to try based on both exploration by trying uncertain areas and exploitation by trying areas expected to perform well (Snoek et al., 2012). Tables 5 and 6 show the results of the best hyperparameters obtained from the random search optimization in the RF and XGBoost model respectively. Moreover, Tables 7 and 8 show the results of best hyperparameter obtained from Bayesian optimization.

4.2. Model evaluation

4.2.1. Performance metrics evaluation of regression ML models

After setting the hyperparameter and building the ML ensemble-based algorithms, 80% of the training dataset was used to fit the models. The results of the evaluation parameters of the developed models are discussed next. There are some useful statistical evaluation metric parameters for the regression model to evaluate and examine the performance of the ML model. In this study, the MAE, R^2, RMSE, and MAPE of the training and test datasets from the RF model was evaluated and compared using the following equations:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|; \tag{4}$$

$$\mathrm{R}^2 = 1 - \frac{\sum\left(y_i - \hat{y}_i\right)^2}{\sum\left(y_i - \hat{y}_i\right)^2}; \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}; \tag{6}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \tag{7}$$

where n is the number of observations, y_i is the actual cost contingency, and $\hat{y}_i$ represents the predicted cost contingency by the models. The MAE is defined in Eqn (4) as a measurement of the average error for all predictions, and a lower MAE indicates higher accuracy. Eqn (5) shows the R^2 which measures the goodness-of-fit and the performance of the model; the closer R^2 is to 1, the better the performance. Moreover, Eqn (6) is the RMSE, which is the difference between predicted values by the model and the actual values. RMSE is often used to compare the performances of ML algorithms (Verrelst et al., 2012). A lower RMSE has the same meaning as a low MAE. Furthermore, Eqn (7) shows the MAPE, which measures the average absolute percentage error between the actual values and the values predicted by the model. The contingency of the prediction cost for both the training and testing datasets by the four regression statistics indicators (MAE, R^2, RMSE, and MAPE) was analyzed after running the RF and XGBoost model using the parameters obtained with the random search hyperparameter optimization.

4.2.2. Comparison of prediction accuracy

To verify the performance of the obtained model, the predictive results of the RF were also compared with those of other ML such as ANN. Figure 6 displays a bar chart comparison of the regression metrics of the testing dataset between RF, XGBoost, and ANN. In terms of R^2, MAE, RMSE, and MAPE, the statistical metrics of ANN showed lower performance compared to RF and XGBoost model.

As shown in Table 9, the testing dataset performance metrics of the RF model exhibit excellent results with an MAE of US$1,369, R^2 of 0.997, RMSE of US$ 2,908, and MAPE of 0.052 followed with XGBoost and ANN. Additionally, Figure 7 exhibits a comparison of the prediction by the three ML models and the actual cost contingency in the 10 examples of testing datasets. It shows that out of three ML models, the RF (the red dotted line) and XGBoost (the orange dotted line) prediction cost contingency models constantly fit the real cost contingency data positively (the blue solid line); consequently, ML ensemble-based can generate accurate predictions as most of them are close to the actual cost contingency. By reviewing the results from the regression metrics evaluation, we can conclude that the best model to predict cost contingency was carried out by RF followed with XGBoost and ANN.

Table 5. Best hyperparameters of RF with random search optimization

Parameter	Description	Parameters of random search	
		Parameter space	Optimum value
bootstrap	The usage of bootstrap samples while creating trees	[True, False]	False
max_depth	The tree's deepest point. If none, nodes are expanded until all leaves are pure or until all leaves have fewer samples than the minimum number of split samples, whichever comes first	[int(x) for x in np.linspace (10, 110, num=11)]	90
max_features	The number of features to consider when looking for the optimum split	["auto", "sqrt"]	"sqrt"
min_samples_leaf	The lowest number of samples that must be present at a leaf node	[1, 2, 4]	1
min_samples_split	A split internal node requires a minimum number of samples	[2, 5, 10]	5
n_estimators	The total of trees in the forest	[int(x) for x in np.linspace(start=200, stop=2000, num=10)]	800

Table 6. Best hyperparameters of XGBoost with random search optimization.

Parameter	Description	Parameters of random search	
		Parameter space	Optimum value
colsample_bytree	Percentage of columns (feature) to be used for each tree	[1.0]	1.0
learning_rate	The created tree reduces the weight used for prediction to prevent overfitting	[0.20, 0.30, 0.40]	0.3
max_depth	The deeper the tree, the greater the likelihood of overfitting. The depth is infinite when 0 is set	[2, 4, 6, 8]	4
n_estimators	Repeat quantity, a greater chance of overfitting if the value is large	[50, 75, 100, 125]	125
gamma	Overfitting control, minimum loss function value to determine the additional division of leaf nodes. Avoid overfitting if the gamma value increases	[0.0, 0.1, 0.2]	0.2
min_child_weight	The minimum number of samples for further segmentation of a node. If it is less than min_child_weight, the node becomes an end node and is no longer segmented	[1, 2, 3]	2

Table 7. Best hyperparameters of RF with Bayesian optimization

Parameter	Description	Parameters of random search	
		Parameter space	Optimum value
bootstrap	The usage of bootstrap samples while creating trees	[True, False]	False
max_depth	The tree's deepest point. If none, nodes are expanded until all leaves are pure or until all leaves have fewer samples than the minimum number of split samples, whichever comes first	Integer (1, 50)	34
max_features	The number of features to consider when looking for the optimum split	Real (0.1, 1.0)	0.498
min_samples_leaf	The lowest number of samples that must be present at a leaf node	Integer (1, 20)	1
min_samples_split	A split internal node requires a minimum number of samples	Integer (2, 20)	2
n_estimators	The total of trees in the forest	Integer (10, 500)	500

Table 8. Best hyperparameters of XGBoost with Bayesian optimization

Parameter	Description	Parameters of random search	
		Parameter space	Optimum value
colsample_bytree	Percentage of columns (feature) to be used for each tree	Real (0.1, 1.0, 'uniform')	0.583
learning_rate	The created tree reduces the weight used for prediction to prevent overfitting	Real (0.01, 1.0, 'uniform')	0.145
max_depth	The deeper the tree, the greater the likelihood of overfitting. The depth is infinite when 0 is set	Integer (2, 12)	5
n_estimators	Repeat quantity, a greater chance of overfitting if the value is large	Integer (50, 5000)	1001
reg_alpha	L1 regularization applied value for weights. This investigates the implementation when the number of features is large. The higher this value, the lower of the overfitting occur	Real (1e-9, 100, 'uniform')	97.754
reg_lambda	L2 regularization applied value for weights. This investigates the implementation when the number of features is large. The higher this value, the lower of the overfitting occur	Real (1e–9, 100, 'uniform')	34.413
subsample	This is the data sampling rate used by weak learners for learning. A lower value can prevent overfitting	Real (0.1, 1.0, 'uniform')	0.779

Table 9. Performance metrics of regression using four different ML models

ML	MAE (US$)		R^2		RMSE (US$)		MAPE	
	Testing	Training	Testing	Training	Testing	Training	Testing	Training
RF	1,369	1,199	0.997	0.998	2,908	2,457	0.052	0.124
XGBoost	2,190	2,728	0.965	0.970	3,326	3,510	0.056	0.173
ANN	4,326	4,895	0.783	0.775	6,393	7,124	0.149	0.236

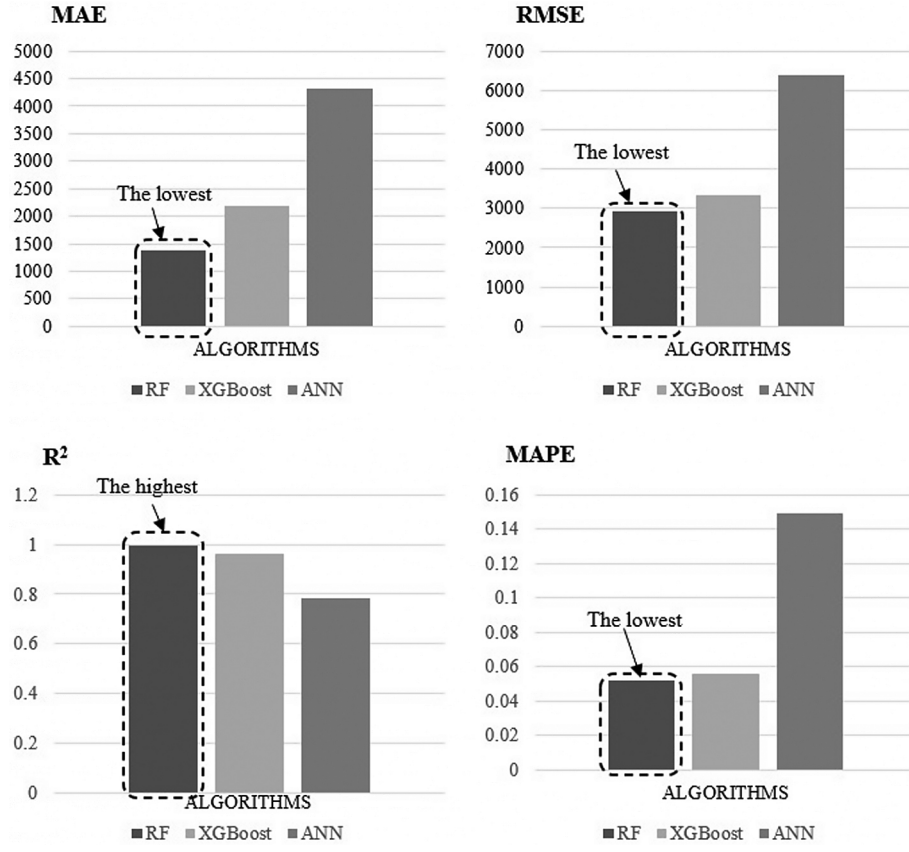

Figure 6. Regression performance metrics comparison of ML models

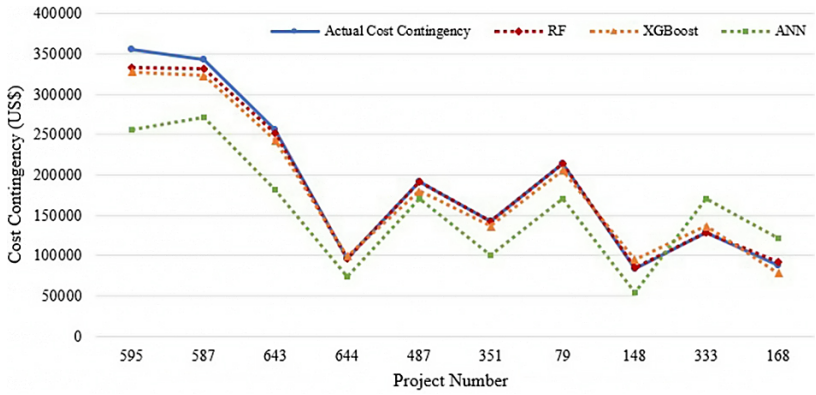

Figure 7. Comparison graph of prediction by the ML models and the actual cost contingency

The results of this study also reveal the feature importance of the developed ML models. Feature importance in ML is used to identify the most influential variable (feature) contributing to the model. Using ML ensemble-based algorithm provides benefits, such as calculating the feature importance automatically from a trained predictive model (Zhu et al., 2021). The feature importance represents the weight of each variable, and the higher the weight, the larger the contribution of the variable to the developed ML model. Figure 8 shows the feature importance score ranking from the best model in this study. Project amount, project duration, latitude, and longitude were found to be the four most important factors influencing the construction cost contingency model. The project amount variable was found to have the highest weight values of 0.298. Moreover, the contract type variable was found to have the lowest weight value of 0.001.

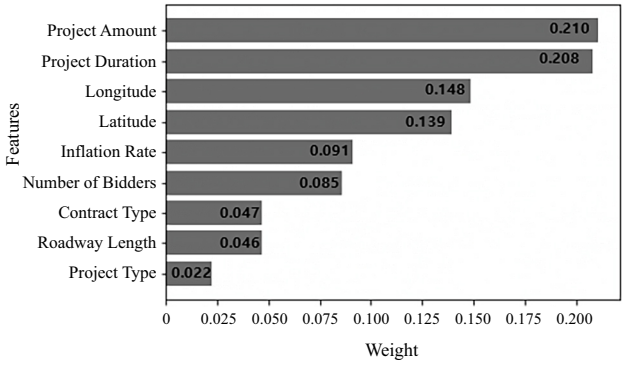

Figure 8. Feature importance score of the best model to predict cost contingency in this study

5. Discussion

It is important to study the prediction techniques by exploring various ML algorithms for the development of big data in the construction industry. Proposing a new technique for cost contingency prediction by integrating ML algorithms has demonstrated by Lhee et al. (2016). In their study, it was the first application of particle swarm optimization (PSO) to construction cost contingency. Additionally, adopting other ML techniques to predict construction cost contingency for AI development in this area was emphasized by El-Kholy et al. (2022). This study proposed an ML model for predicting the cost contingency of construction projects by employing varied algorithms. The results of this study are similar to those found by Cao et al. (2018), who applied ML ensemble-based algorithms to predict unit price bids of resurfacing transportation projects with stable, accurate, and efficient prediction. The results of this study show that among the tested ML algorithms, the best prediction of cost contingency can be obtained by implementing RF associated with random search hyperparameter optimization. Moreover, in this study, continuous and categorical variables to predict cost contingency transportation projects were utilized as predictor variables. Dealing with continuous and categorical predictor variables is suitable for using ML ensemble-based. Because the basic principle of ensemble-based algorithms is classification, this technique can reduce the risk of choosing a poor classifier through voting and then generate robust models (Dietterich, 2000). This is supported by the findings of Cha et al. (2021), who validated that predicting with ensemble algorithms such as bagging and boosting type is powerful when the size of datasets is small and the type of variables is categorical. Therefore, the ML ensemble-based algorithms such as RF and XGBoost can build excellent models with the samples used in this study compared with other models such as ANN, as this ML algorithm are broadly adopted for construction cost contingency prediction (Hashemi et al., 2020). Moreover, the results of this study present high accuracy compared with previous studies by Lhee et al. (2012, 2014), who applied ML algorithms such as ANN to predict construction contingency.

The project amount variable was found to be the most influential predictor for the cost contingency project. These findings are in accordance with those of Lhee et al. (2012), who developed a model to predict the cost contingency in asphalt resurfacing projects using ANN. In their study, a numerical input variable, such as the project amount, was used to build the ANN models, and this variable had the highest correlation with the cost contingency. The results of this study also show that the project duration variable has a close correlation with cost contin-

gency. This was unexpected and differed from the findings of Lhee et al. (2012, 2014), who found that project duration tends to degrade the performance of the ANN model in predicting cost contingency. Moreover, features such as latitude and longitude were also found to be influential variables in predicting the cost contingency of transportation construction projects. Latitude and longitude represent the detailed location of the project and have been applied as a prediction tool in other construction research areas. Won et al. (2018) used latitude, longitude, and altitude as independent variables in their ML algorithm to predict construction resource location, whereas Anjum et al. (2021) developed a deep learning model for floor opening detection in construction projects and utilized latitude and longitude data in the geocoding process. Consequently, in this study, the latitude and longitude variables were the first findings in construction cost contingency prediction using ML algorithms.

One of the methods used to reduce the effect of unbalanced data on the regression model is resampling. Oversampling, undersampling, and mixed sampling are classifications of resampling techniques. SMOGN is a resampling technique that combines random undersampling with two over-sampling techniques, including SMOTER and the introduction of Gaussian noise (Branco et al., 2017). The goal of SMOGN is to address the issue of unbalanced regression that can complicate the predictive model development process. Branco et al. (2017) asserted that SMOGN adjusts the number of rare and normal cases and approximately keeps the same total number of datasets. The results of SMOGN in this study made the distribution of the target variable (cost contingency) "*skewed right*" when compared to the original dataset. Furthermore, in accordance with the results of this study, SMOGN associated with the XGBoost algorithm has been proven to successfully cope with the imbalanced distribution dataset problem and enhance prediction accuracy (Zhu et al., 2021). Accordingly, applying the SMOGN technique in the preprocessing step can help build the model based on the characteristics of the dataset in this study.

This study differs from previous contingency prediction research in some ways. First, this is the first study to implement ML ensembled-based algorithms such as RF and XGBoost, associated with random search hyperparameter optimization to predict cost contingency in construction projects. Second, this is the first study to employ SMOGN, an effective resampling algorithm, in the pre-processing stage to deal with imbalanced regression problems for cost contingency. Third, this is the first study to use latitude and longitude as predictor variables for cost estimation in the construction project area. In previous related studies on contingency prediction with ANN and MLR, independent variables were used, such as project amount, project duration, roadway length, number of bidders, project type, letting year, and weather conditions (Lhee et al., 2012, 2014; Thal et al., 2010; Chen & Hartman, 2000). Unexpectedly, there is no research on construction contingency prediction models that consider latitude, longitude, inflation rate, and contract type as predictor variables. Therefore, we expect the future to bring more variables and ML models, given the extensive development of artificial intelligence (AI) and big data in the construction industry.

In a real construction project, the method that is commonly used to estimate cost contingency is the predetermined percentage (i.e., 5–10% of the total cost project) based on intuitions, gut feelings, and past experiences of estimators owing to the simplicity of this method (Lhee et al., 2012). However, this method is not appropriate because each construction project has its own uniqueness. From the construction cost prediction overview, ML can be used to reduce the time and enhance the accuracy of the prediction, which is an important aspect of the project team's decision-making process. ML construction cost prediction models can be effectively applied if the construction project team has sufficient datasets and the skills necessary to become proficient in data analysis. This can be achieved by (1) providing data analysis training such as learning about language programming (i.e., Python, R, and other programming languages), which could give more insights about big data in construction to all project teams; (2) providing access to the large and high-quality datasets for research needs; and (3) continuously developing cost prediction models and developing an integrated web-based cost estimation model through collaboration between researchers and construction practitioners. From a practical viewpoint, the implementation of ML in the construction industry may be slow and challenging because of the barriers to adopting new technologies (Nitithamyong & Skibniewski, 2004). Despite the challenges, the adoption of big data, ML, and AI in improving sustainability in the construction industry is inevitable and has become a necessity (Bilal et al., 2016).

To enhance the practical application of the proposed model of this study, the output of the RF algorithm using feature importance and partial dependence plots were analyzed. The model indicated that contingency percentages are most influenced by the project amount, duration, inflation rate, and contract type. Projects with longer durations and higher budgets tend to require higher contingency allocations. Additionally, urban projects with design-build contracts and those located in regions with higher inflation also exhibit elevated contingency needs. Hence, the results of this model revealed that the transportation cost contingency range from 8–14%.

6. Conclusions

As cost contingency is an important budget element and has the same weight as the direct cost of a construction project, a method to calculate this cost is crucial to achieving good performance. This study proposes a ML-based framework for cost contingency prediction with a robust performance accuracy over traditional methods. The methods in previous literature for calculating contingency mainly include ANN and fuzzy techniques, which have some limitations in theory and practice. To date, no stud-

ies have utilized an ML ensemble-based algorithm for predicting the cost contingency of construction projects. Additionally, this study revealed some novel insights by revealing previously unrecognized patterns and influential factors. In this study, RF and XGBoost, ML ensemble-based algorithms, were introduced to predict cost contingency using the FDOT transportation construction project dataset. Moreover, employing SMOGN is remarkably effective in handling imbalanced dataset problems and can enhance the performance of ML model, as verified through real-world dataset tests. In addition, applying hyperparameter optimizations such as random search can also make the prediction model robust. Furthermore, based on a comparison with other algorithm such as ANN applied to the training and testing datasets, the results verified that RF provides excellent predictions. Consequently, the framework proposed in this study is applicable and can also cover the limitations of previous contingency calculating methods, such as accuracy, imbalanced datasets, and issues with categorical variables. This study also contributes to construction industry by demonstrating the potential of ensembled-ML techniques in improving cost contingency prediction. These results can guide researchers to adopt ML in construction cost prediction and suggest practitioners to implement ML for mitigating budget risks early in the project lifecycle in more advanced method and the era of big data.

Through the feature importance score of the best model in this study, it was found that the project amount, project duration, latitude, and longitude are the four best independent variables (features) that have a significant impact on the construction cost contingency model. In contrast, despite the satisfactory performance of the model, this study may have some limitations, which must be investigated in future work. First, this study only focused on cost contingency prediction; therefore, research on predicting construction time contingency with ML algorithms is needed. Second, this study only applied SMOGN to overcome the imbalance problem; further studies on implementing other resampling algorithm techniques in the pre-processing step may need to be conducted. Finally, the study only used the FDOT transportation construction project datasets; thus, the predictor variables may be different if adopting construction project datasets from other countries. Further studies implementing other ML algorithms to predict construction contingency costs are necessary.

Data availability statement

Some or all data, models, or codes generated or used during the study are proprietary or confidential in nature and may only be provided with restrictions.

Acknowledgements

The authors would like to express their gratitude to Hanyang University and we wish to thank Editage (www.editage.co.kr) for the English language editing.

Funding

This research was funded by the National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science and ICT) (grant number RS202400356697).

Author contributions

Conceptualization, A. N. and J.-H. K.; methodology, A. N.; software, A. N. and S.-J. P.; validation, J.-H. K. and S.-J. P.; formal analysis, A. N.; resources, A. N. and S. F. R.; data curation, K.-T. L.; writing – original draft preparation, A. N.; writing – review and editing, J.-H. K.; visualization, A. N.; supervision, J.-H. K.; project administration, A. N. and S. F. R. All authors have read and agreed to the published version of the manuscript.

Disclosure statement

The authors declare no conflict of interest.

References

Abt, K. (1987). Descriptive data analysis: a concept between confirmatory and exploratory data analysis. *Methods of Information in Medicine*, *26*(2), 77–88. https://doi.org/10.1055/s-0038-1635488

Agrawal, T. (2021). *Hyperparameter optimization in machine learning*: *Make your machine learning and deep learning models more efficient*. Apress. https://doi.org/10.1007/978-1-4842-6579-6

Alpaydin, E. (2020). *Introduction to machine learning*. The MIT Press.

Alshboul, O., Shehadeh, A., Almasabha, G., & Almuflih, A. S. (2022). Extreme gradient boosting-based machine learning approach for green building cost prediction. *Sustainability*, *14*(11), Article 6651. https://doi.org/10.3390/su14116651

Ameh, O. J., Soyingbe, A. A., & Odusami, K. T. (2010). Significant factors causing cost overruns in telecommunication projects in Nigeria. *Journal of Construction in Developing Countries*, *15*(2), 49–67. https://ir.unilag.edu.ng/handle/123456789/8924

Ammar, T., Abdel-Monem, M., & El-Dash, K. (2022). Risk factors causing cost overruns in road networks. *Ain Shams Engineering Journal*, *13*(5), Article 101720. https://doi.org/10.1016/j.asej.2022.101720

Ammar, T., Abdel-Monem, M., & El-Dash, K. (2025). Regression-based model predicting cost contingencies for road network projects. *International Journal of Construction Management*, *25*(11), 1273–1287. https://doi.org/10.1080/15623599.2024.2411082

Anitescu, C., Atroshchenko, E., Alajlan, N., & Rabczuk, T. (2019). Artificial neural network methods for the solution of second order boundary value problems. *Computers*, *Materials* & *Continua*, *59*(1), 345–359. https://doi.org/10.32604/cmc.2019.06641

Anjum, S., Khalid, R., Khan, M., Khan, N., & Park, C. (2021). A pull-reporting approach for floor opening detection using deep-learning on embedded devices. In *Proceedings of the of the 38th International Symposium on Automation and Robotics in Construction* (*ISARC 2021*) (pp. 395–402), Dubai, UAE. https://doi.org/10.22260/ISARC2021/0055

Arifuzzaman, M., Gazder, U., Islam, M. S., & Skitmore, M. (2022). Budget and cost contingency CART models for power plant

projects. *Journal of Civil Engineering and Management*, *28*(8), 680–695. https://doi.org/10.3846/jcem.2022.16944

Artur, M. (2021). Review the performance of the Bernoulli Naïve Bayes classifier in intrusion detection systems using recursive feature elimination with cross-validated selection of the best number of features. *Procedia Computer Science*, *190*, 564–570. https://doi.org/10.1016/j.procs.2021.06.066

Asamoah, Oduro R., Offei-Nyako, K., & Twumasi-Ampofo, K. (2023). Relative importance of triggers influencing cost contingency determination for building contracts-the perspective of quantity surveyors. *International Journal of Construction Management*, *23*(5), 790–798. https://doi.org/10.1080/15623599.2021.1930638

Association for the Advancement of Cost Engineering International. (2008). *Contingency estimating-general principles* (AACE Recommended Practice No. 40R-08, TCM Framework). https://www.pathlms.com/aace/courses/2928/documents/3825#

Baccarini, D. (2004). Accuracy in estimating project cost construction contingency-a statistical analysis. In *Cobra 2004*: *RICS International Construction Conference*, *Responding to Change*, London, United Kingdom. http://hdl.handle.net/20.500.11937/29859

Baccarini, D., & Love, P. E. (2014). Statistical characteristics of cost contingency in water infrastructure projects. *Journal of Construction Engineering and Management*, *140*(3), Article 04013063. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000820

Bakhshi, P., & Touran, A. (2014). An overview of budget contingency calculation methods in construction industry. *Procedia Engineering*, *85*, 52–60. https://doi.org/10.1016/j.proeng.2014.10.528

Bekkerman, R. (2015). *The present and the future of the KDD Cup Competition*: *An outsider's perspective* [Post]. LinkedIn. https://www.linkedin.com/pulse/present-future-kdd-cup-competition-outsiders-ron-bekkerman/

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2), 281–305.

Bilal, M., & Oyedele, L. O. (2020). Guidelines for applied machine learning in construction industry – A case of profit margins estimation. *Advanced Engineering Informatics*, *43*, Article 101013. https://doi.org/10.1016/j.aei.2019.101013

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A., & Pasha, M. (2016). Big data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, *30*(3), 500–521. https://doi.org/10.1016/j.aei.2016.07.001

Branco, P., Torgo, L., & Ribeiro, R. P. (2017). SMOGN: a pre-processing approach for imbalanced regression. In *First International Workshop on Learning with Imbalanced Domains*: *Theory and Applications* (Vol. 74, pp. 36–50), Skopje, Macedonia.

Branco, P., Torgo, L., & Ribeiro, R. P. (2019). Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing*, *343*, 76–99. https://doi.org/10.1016/j.neucom.2018.11.100

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140. https://link.springer.com/article/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://link.springer.com/article/10.1023/a:1010933404324

Cantarelli, C. C., Flyvbjerg, B., & Buhl, S. L. (2012). Geographical variation in project cost performance: the Netherlands versus worldwide. *Journal of Transport Geography*, *24*, 324–31. https://doi.org/10.1016/j.jtrangeo.2012.03.014

Cao, Y., Ashuri, B., & Baek, M. (2018). Prediction of unit price bids of resurfacing highway projects through ensemble machine learning. *Journal of Computing in Civil Engineering*, *32*(5), Article 04018043. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000788

Catalão, F. P., Cruz, C. O., & Sarmento, J. M. (2019). The determinants of cost deviations and overruns in transport projects, an endogenous model approach. *Transport Policy*, *74*, 224–238. https://doi.org/10.1016/j.tranpol.2018.12.008

Cha, G. W., Moon, H. J., & Kim, Y. C. (2021). Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables. *International Journal of Environmental Research Public Health*, *18*(16), Article 8530. https://doi.org/10.3390/ijerph18168530

Chakraborty, D., & Elzarka, H. (2019). Advanced machine learning techniques for building performance simulation: A comparative analysis. *Journal of Building Performance Simulation*, *12*(2), 193–207. https://doi.org/10.1080/19401493.2018.1498538

Chan, E. H., & Au, M. C. (2008). Relationship between organizational sizes and contractors' risk pricing behaviors for weather risk under different project values and durations. *Journal of Construction Engineering and Management*, *134*(9), 673–680. https://doi.org/10.1061/(ASCE)0733-9364(2008)134:9(673)

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'16*) (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Chen, D., & Hartman, F. T. (2000). A neural network approach to risk assessment and contingency allocation. *AACE International Transactions*.

Clark, D. E. (2001). Monte Carlo analysis: Ten years of experience. *Cost Engineering*, *43*(6), 40–45.

Daly, A., Dekker, T., & Hess, S. (2016). Dummy coding vs effects coding for categorical variables: Clarifications and extensions. *Journal of Choice Modelling*, *21*, 36–41. https://doi.org/10.1016/j.jocm.2016.09.005

De Marco, A., Rafele, C., & Thaheem, M. J. (2016). Dynamic management of risk contingency in complex design-build projects. *Journal of Construction Engineering and Management*, *142*(2), Article 04015080. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001052

Dey, P., Tabucanon, M. T., & Ogunlana, S. O. (1994). Planning for project control through risk analysis: A petroleum pipeline-laying project. *International Journal of Project Management*, *12*(1), 23–33. https://doi.org/10.1016/0263-7863(94)90006-X

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*, 139–157. https://link.springer.com/article/10.1023/a:1007607513941

Dietterich, T. G. (2002). Ensemble learning. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed.) (pp. 405–408). The MIT Press.

El-Kholy, A. M., Tahwia, A. M., & Elsayed, M. M. (2022). Prediction of simulated cost contingency for steel reinforcement in building projects: ANN versus regression-based models. *International Journal of Construction Management*, *22*(9), 1675–1689. https://doi.org/10.1080/15623599.2020.1741492

Elmousalami, H. H. (2020). Comparison of artificial intelligence techniques for project conceptual cost prediction: a case study and comparative analysis. *IEEE Transactions on Engineering Management*, *68*(1), 183–196. https://doi.org/10.1109/TEM.2020.2972078

El-Touny, A. S., Ibrahim, A. H., & Amer, M. I. (2014). Estimating cost contingency for highway construction projects using analytic hierarchy process. *International Journal of Computer Science Issues*, *11*(6), Article 73.

Enshassi, A., & Ayyash, A. (2014). Factors affecting cost contingency in the construction industry–contractors' perspective. *International Journal of Construction Management*, *14*(3), 191–208. https://doi.org/10.1080/15623599.2014.922729

Espinoza, R. D. (2011). Contingency estimating using option pricing theory: Closing the gap between theory and practice. *Construction Management and Economics*, *29*(9), 913–927. https://doi.org/10.1080/01446193.2011.610328

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *Journal of Machine Learning Research*, *15*(1), 3133–3181.

Flyvbjerg, B., Holm, M. S., & Buhl, S. (2002). Underestimating costs in public works projects: Error or lie?. *Journal of the American Planning Association*, *68*(3), 279–295. https://doi.org/10.1080/01944360208976273

Gharaibeh, H. M. (2014). Cost control in mega projects using the Delphi method. *Journal of Management in Engineering*, *30*(5), Article 04014024. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000218

Ghimire, B., Rogan, J., Galiano, V. R., Panday, P., & Neeti, N. (2012). An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA. *GIScience & Remote Sensing*, *49*(5), 623–643. https://doi.org/10.2747/1548-1603.49.5.623

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). *Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation* (Technical report). University of Texas at El Paso Computer Science.

Günhan, S., & Arditi, D. (2007). Budgeting owner's construction contingency. *Journal of Construction Engineering and Management*, *133*(7), 492–497. https://doi.org/10.1061/(ASCE)0733-9364(2007)133:7(492)

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hall, P. (1994). Methodology and theory for the bootstrap. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of econometrics* (Vol. 4, pp. 2341–2381). Elsevier Inc. https://doi.org/10.1016/S1573-4412(05)80008-X

Hamid, R. A., & Kehinde, F. J. (2017). Choosing an appropriate contingency sum estimating methods for highway construction projects in Nigeria: A literature review. *Planning Malaysia Journal*, *15*(1). https://doi.org/10.21837/pm.v15i1.217

Hartman, F. T. (2000). *Don't park your brain outside: A practical guide to improving shareholder value with SMART management*. Project Management Institute.

Hashemi, T. S., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. *SN Applied Sciences*, *2*, Article 1703. https://doi.org/10.1007/s42452-020-03497-1

Hollmann, J. K. (2012). Estimate accuracy: Dealing with reality. *Cost Engineering-Morgantown, 54*(6), Article 17.

Hoseini, E., Bosch-Rekveldt, M., & Hertogh, M. (2020a). Cost contingency and cost evolvement of construction projects in the preconstruction phase. *Journal of Construction Engineering and Management*, *146*(6), Article 05020006. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001842

Hoseini, E., Van Veen, P., Bosch-Rekveldt, M., & Hertogh, M. (2020b). Cost performance and cost contingency during project execution: Comparing client and contractor perspectives. *Journal of Management in Engineering*, *36*(4), Article 05020006. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000772

Huang, C.-H., & Hsien, S.-H. (2020). Predicting BIM labor cost with random forest and simple linear regression. *Automation in Construction*, *118*, Article 103280. https://doi.org/10.1016/j.autcon.2020.103280

Jiang, Q. (2020). Estimation of construction project building cost by back-propagation neural network. *Journal of Engineering, Design and Technology*, *18*(3), 601–609. https://doi.org/10.1108/JEDT-08-2019-0195

Karlsen, J. K., & Lereim, J. (2005). Management of project contingency and allowance. *Cost Engineering*, *47*(9), 24–29.

Kasimu, M. A. (2012). Significant factors that cause cost overruns in building construction project in Nigeria. *Interdisciplinary Journal of Contemporary Research in Business*, *3*(11), 775–780.

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys*, *52*(4), Article 79. https://doi.org/10.1145/3343440

Kim, G. H., An, S. H., & Kang, K. I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment*, *39*(10), 1235–1242. https://doi.org/10.1016/j.buildenv.2004.02.013

Kunz, N. (2020). *SMOGN: Synthetic minority over-sampling technique for regression with Gaussian noise*. https://pypi.org/project/smogn/

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*, 1–13. https://doi.org/10.18637/jss.v036.i11

Larsen, J. K., Shen, G. Q., Lindhard, S. M., & Brunoe, T. D. (2016). Factors affecting schedule delay, cost overrun, and quality level in public construction projects. *Journal of Management in Engineering*, *32*(1), Article 04015032. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000391

Laryea, S., & Hughes, W. (2009). How contractors in Ghana include risk in their bid prices. In *Proceedings of 25th Annual ARCOM Conference* (pp. 1295–1304), Nottingham, UK. Association of Researchers in Construction Management.

Lathong, K., & Wisaeng, K. (2024). An innovative hybrid machine learning techniques for predicting construction cost estimates. *International Journal for Computational Civil and Structural Engineering*, *20*(3), 69–83.

Lhee, S. C. (2014). Finding significant factors to affect cost contingency on construction projects using ANOVA statistical method-focused on transportation construction projects in the US. *Architectural Research*, *16*(2), 75–80. https://doi.org/10.5659/AIKAR.2014.16.2.75

Lhee, S. C., Issa, R. R., & Flood, I. (2012). Prediction of financial contingency for asphalt resurfacing projects using artificial neural networks. *Journal of Construction Engineering and Management*, *138*(1), 22–30. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000408

Lhee, S. C., Flood, I., & Issa, R. R. (2014). Development of a two-step neural network-based model to predict construction cost contingency. *Journal of Information Technology in Construction* (*ITcon*), *19*(24), 399–411.

Lhee, S. C., Issa, R. R., & Flood, I. (2016). Using particle swarm optimization to predict cost contingency on transportation construction projects. *Journal of Information Technology in Construction* (*ITcon*), *21*(30), 504–516.

Li, F., Laili, Y., Chen, X., Lou, Y., Wang, C., Yang, H., Gao, X., & Han, H. (2023). Towards big data driven construction industry.

Journal of Industrial Information Integration, *35*, Article 100483. https://doi.org/10.1016/j.jii.2023.100483

Love, P. E. D., Sing, C. P., Wang, X., Irani, Z., & Thwala, D. W. (2014). Overruns in transportation infrastructure projects. *Structure and Infrastructure Engineering*, *10*, 141–159. https://doi.org/10.1080/15732479.2012.715173

Mahamid, I. (2013). Effects of project's physical characteristics on cost deviation in road construction. *Journal of King Saud University-Engineering Sciences*, *25*(1), 81–88. https://doi.org/10.1016/j.jksues.2012.04.001

Mahmoodzadeh, A., Nejati, H. R., & Mohammadi, M. (2022a). Optimized machine learning modelling for predicting the construction cost and duration of tunnelling projects. *Automation in Construction*, *139*, Article 104305. https://doi.org/10.1016/j.autcon.2022.104305

Mahmoodzadeh, A., Nejati, H. R., Mohammadi, M., Ibrahim, H. H., Khishe, M., Rashidi, S., & Mohammed, A. H. (2022b). Developing six hybrid machine learning models based on gaussian process regression and meta-heuristic optimization algorithms for prediction of duration and cost of road tunnels construction. *Tunnelling and Underground Space Technology*, *130*, Article 104759. https://doi.org/10.1016/j.tust.2022.104759

Mak, S., Wong, J., & Picken, D. (1998). The effect on contingency allowances of using risk analysis in capital cost estimating: A Hong Kong case study. *Construction Management and Economics*, *16*(6), 615–619. https://doi.org/10.1080/014461998371917

Manu, P., Ankrah, N., Proverbs, D., & Suresh, S. (2010). An approach for determining the extent of contribution of construction project features to accident causation. *Safety Science*, *48*(6), 687–692. https://doi.org/10.1016/j.ssci.2010.03.001

Meharie, M. G., & Shaik, N. (2020). Predicting highway construction costs: Comparison of the performance of random forest, neural network and support vector machine models. *Journal of Soft Computing in Civil Eng*ineering, *4*(2), 103–112. https://doi.org/10.22115/scce.2020.226883.1205

Moselhi, O. (1997). Risk assessment and contingency estimating. In *AACE International Transactions, Dallas, USA*.

Myers, J. L., Well, A., & Lorch, R. F. (2010). *Research design and statistical analysis*. Routledge.

Nawar, S., Hosny, O., & Nassar, K. (2018). Owner time and cost contingency estimation for building construction projects in Egypt. In *Construction Research Congress* 2018 (pp. 367–377), New Orleans, Louisiana, USA. https://doi.org/10.1061/9780784481271.036

Nitithamyong, P., & Skibniewski, M. J. (2004). Web-based construction project management systems: how to make them successful?. *Automation in Construction*, *13*(4), 491–506. https://doi.org/10.1016/j.autcon.2004.02.003

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*, 169–198. https://doi.org/10.1613/jair.614

Peško, I., Mučenski, V., Šešlija, M., Radović, N., Vujkov, A., Bibić, D., & Krklješ, M. (2017). Estimation of costs and durations of construction of urban roads using ANN and SVM. *Complexity*, *2017*, Article 2450370. https://doi.org/10.1155/2017/2450370

Petrusheva, S., Car-Pušić, D., & Zileska-Pancovska, V. (2019). Support vector machine based hybrid model for prediction of road structures construction costs. *IOP Conference Series*: *Earth and Environmental Science*, *222*(1), Article 012010. https://doi.org/10.1088/1755-1315/222/1/012010

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits Systems Magazine*, *6*(3), 21–45. https://doi.org/10.1109/MCAS.2006.1688199

Querns, W. R. (1989). What is contingency, anyway?. In *AACE International Transactions*.

Rafiei, M. H., & Adeli, H. (2018). Novel machine-learning model for estimating construction costs considering economic variables and indexes. *Journal of Construction Engineering and Management*, *144*(12), Article 04018106. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001570

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4), Article e1249. https://doi.org/10.1002/widm.1249

Salah, A., & Moselhi, O. (2015). Contingency modelling for construction projects using fuzzy-set theory. *Engineering*, *Construction and Architectural Management*, *22*(2), 214–241. https://doi.org/10.1108/ECAM-03-2014-0039

Samaniego, E., Anitescu, C., Goswami, S., Nguyen-Thanh, V. M., Guo, H., Hamdia, K., Zhuang, X., & Rabczuk, T. (2020). An energy approach to the solution of partial differential equations in computational mechanics via machine learning: Concepts, implementation and applications. *Computer Methods in Applied Mechanics and Engineering*, *362*, Article 112790. https://doi.org/10.1016/j.cma.2019.112790

Shardlow, M. (2016). *An analysis of feature selection techniques*. The University of Manchester, United Kingdom.

Shoar, S., Chileshe, N., & Edwards, J. D. (2022). Machine learning-aided engineering services' cost overruns prediction in high-rise residential building projects: Application of random forest regression. *Journal of Building Engineering*, *50*, Article 104102. https://doi.org/10.1016/j.jobe.2022.104102

Shrestha, K. K., & Shrestha, P. P. (2016). A cost contingency estimation system for road maintenance contracts. *Procedia Engineering*, *145*, 128–135. https://doi.org/10.1016/j.proeng.2016.04.030

Smith, G. R., & Bohn, C. M. (1999). Small to medium contractor contingency and assumption of risk. *Journal of Construction Engineering and Management*, *125*(2), 101–108. https://doi.org/10.1061/(ASCE)0733-9364(1999)125:2(101)

Snoek, J., Larochelle, H., & Adams, R. P. (2012). *Practical Bayesian optimization of machine learning algorithms*. arXiv. https://doi.org/10.48550/arXiv.1206.2944

Sonmez, R., Ergin, A., & Birgonul, M. T. (2007). Quantitative methodology for determination of cost contingency in international projects. *Journal of Management in Engineering*, *23*(1), 35–39. https://doi.org/10.1061/(ASCE)0742-597X(2007)23:1(35)

Thal, J. A. E., Cook, J. J., & White III, E. D. (2010). Estimation of cost contingency for air force construction projects. *Journal of Construction Engineering and Management*, *136*(11), 1181–1188. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000227

Thompson, P. A., & Perry, J. G. (1992). *Engineering construction risks*: *A guide to project risk analysis and assessment implications for project clients and project managers*. Thomas Telford.

Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P. (2013). Smote for regression. In *Proceedings of the 16th Portuguese Conference on Artificial Intelligence Progress in Artificial Intelligence* (*EPIA 2013*) (pp. 378–389), Angra do Heroísmo, Azores, Portugal. https://doi.org/10.1007/978-3-642-40669-0_33

Touran, A., & Lopez, R. (2006). Modeling cost escalation in large infrastructure projects. *Journal of Construction Engineering and Management*, *132*(8), 853–860. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:8(853)

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J. P., Camps-Valls, G., & Moreno, J. (2012). Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for sentinel-2 and-3. *Remote Sensing of Environment*, *118*, 127–139. https://doi.org/10.1016/j.rse.2011.11.002

Wang, M. T., & Chou, H. Y. (2003). Risk allocation and risk handling of highway projects in Taiwan. *Journal of Management in Engineering*, *19*(2), 60–68. https://doi.org/10.1061/(ASCE)0742-597X(2003)19:2(60)

Wang, Z., & Srinivasan, R. S. (2017). A review of artificial intelligence-based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, *75*, 796–808. https://doi.org/10.1016/j.rser.2016.10.079

Wang, X., Zhu, J., Ma, F., Li, C., Cai, Y., & Yang, Z. (2016). Bayesian network-based risk assessment for hazmat transportation on the middle route of the South-to-North water transfer project in China. *Stochastic Environmental Research and Risk Assessment*, *30*, 841–857. https://doi.org/10.1007/s00477-015-1113-6

Wang, C. C., Kuo, P. H., & Chen, G. Y. (2022). Machine learning prediction of turning precision using optimized XGBoost model. *Applied Sci*ences, *12*(15), Article 7739. https://doi.org/10.3390/app12157739

Won, D., Park, M. W., & Chi, S. (2018). Construction resource localization based on UAV-RFID platform using machine learning algorithm. In *2018 IEEE International Conference on Industrial Engineering and Engineering Management* (*IEEM*) (pp. 1086–1090). IEEE. https://doi.org/10.1109/IEEM.2018.8607668

Wu, W., Kunz, N., & Branco, P. (2022). ImbalancedLearningRegression – A Python package to tackle the imbalanced regression problem. In M. R. Amini, S. Canu, A. Fischer, T. Guns, P. Kralj Novak, & G. Tsoumakas (Eds.), *Lecture notes in computer science*. *Vol. 13718*: *Machine learning and knowledge discovery in databases* (*ECML PKDD 2022*) (pp. 645–648). Springer, Cham. https://doi.org/10.1007/978-3-031-26422-1_48

Yan, H., He, Z., Gao, C., Xie, M., Sheng, H., & Chen, H. (2022). Investment estimation of prefabricated concrete buildings based on XGBoost machine learning algorithm. *Advanced Engineering Informatics*, *54*, Article 101789. https://doi.org/10.1016/j.aei.2022.101789

Yeo, K. T. (1990). Risks, classification of estimates, and contingency management. *Journal of Management in Engineering*, *6*(4), 458–470. https://doi.org/10.1061/(ASCE)9742-597X(1990)6:4(458)

Zekić-Sušac, M., Has, A., & Knežević, M. (2021). Predicting energy cost of public buildings by artificial neural networks, CART, and random forest. *Neurocomputing*, *439*, 223–233. https://doi.org/10.1016/j.neucom.2020.01.124

Zheng, Z., Zhou, L., Wu, H., & Zhou, L. (2023). Construction cost prediction system based on Random Forest optimized by the Bird Swarm Algorithm. *Mathematical Biosciences and Engineering*, *20*(8), 15044–15074. https://doi.org/10.3934/mbe.2023674

Zhou, J., Li, E., Wei, H., Li, C., Qiao, Q., & Armaghani, D. J. (2019). Random forests and cubist algorithms for predicting shear strengths of rockfill materials. *Applied Sciences*, *9*(8), Article 1621. https://doi.org/10.3390/app9081621

Zhu, X., Chu, J., Wang, K., Wu, S., Yan, W., & Chiam, K. (2021). Prediction of rockhead using a hybrid N-XGBoost machine learning framework. *Journal of Rock Mechanics and Geotechnical Engineering*, *13*(6), 1231–1245. https://doi.org/10.1016/j.jrmge.2021.06.012