www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

RISK-BASED COMPLETION COST PREDICTION APPROACH IN CONSTRUCTION PROJECTS UTILIZING MACHINE LEARNING

SUBMITTED: June 2024
REVISED: October 2024
PUBLISHED: March 2025
EDITOR: Bimal Kumar
DOI: 10.36680/j.itcon.2025.016

Aynur Hurriyet Turkyilmaz, PhD Candidate,
Istanbul Technical University, Department of Civil Engineering, Istanbul, Turkiye
ORCID: https://orcid.org/0009-0009-6646-5381
turkyilmaza19@itu.edu.tr

Gul Polat, Professor,
Istanbul Technical University, Department of Civil Engineering, Istanbul, Turkiye
ORCID: https://orcid.org/0000-0003-2431-033X
polatgu@itu.edu.tr

SUMMARY*: The construction industry is among the sectors exposed to frequent budget overruns. Therefore, accurately predicting costs to complete the construction projects is a vital point. Several research studies focus on cost estimation, construction risk factors, and their cost impact. Although they produced valuable prediction models for the completion cost of the projects, most of them mainly concentrated on the early stages of the construction. Limited studies produced approaches for completion cost estimation in the execution phase of the projects. Nevertheless, they do not implement total risk score effects in their models. Additional research is necessary to investigate risk-based completion cost prediction throughout the execution phase of construction. The main objective of this study is to provide an approach for the total risk score based completion cost prediction by using machine learning techniques without imposing excessive work. The proposed approach can be utilized at any point during the execution phase of a project to assess the impact of changes in total risk scores on completion costs. Furthermore, predicting the total completion cost using the total risk score simplifies the calculation and procedure rather than depending on breakdowns. To achieve this objective, a machine learning prediction approach was proposed to predict total completion cost based on total risk scores in construction projects. The proposed approach is applied to real-world cases to evaluate the accuracy of completion cost prediction based on risk scores using data from an international construction company. A total of 119 risk and cost data points from 11 projects were analyzed. Six prediction algorithms were employed, utilizing machine learning. Based on the outputs, it was determined that polynomial regression produced the most accurate predictions for available data. This research contributes to enhancing construction organizations' knowledge and planning capacities by quickly predicting project completion costs based on dynamic total risk scores at any time throughout the execution phase of the project.*

KEYWORDS*: cost estimation, prediction, machine learning, construction management.*

REFERENCE*: Aynur Hurriyet Turkyilmaz & Gul Polat (2025). Risk-based completion cost prediction approach in construction projects utilizing machine learning. Journal of Information Technology in Construction (ITcon), Vol. 30, pg. 375-396, DOI: 10.36680/j.itcon.2025.016*

1. INTRODUCTION

The time, quality, and cost triad is commonly acknowledged as the construction project's achievement. The cost performance is a vital measure for assessing the achievement of a construction project (Ahiaga-Dagbui and Smith, 2014). According to Atapattu et al. (2024), there has been no improvement in addressing cost overrun over the past 70 years. It is supported by various studies from literature that projects exceeded their allocated budget (Aziz, 2013; Baratta, 2006; Doloi, 2013; Flyvbjerg et al., 2002; Johnson and Babu, 2020; Odeck, 2004; Shane et al., 2009; Vaardini et al., 2016).

The completion cost in construction projects is primarily determined during the tender phase, based on the quantities specified in contract documents and the corresponding unit costs derived from in-house know-how of the companies or published manuals. Generally, data is limited during the initial phases of the construction project, and the assumptions covered in the early estimates might change. Additionally, the risk register is a crucial document for international construction companies. It is created during the tender stage and updated monthly. These registers include the probability and impact values of possible risks of the project, and the multiplication of these values gives a risk score. At that point, the risk evaluated in registers might also be incorporated into the estimated completion cost as a contingency or markup. After the bidding step, in the project's preliminary stages, the tender cost estimation is revised utilizing additional accessible data. This estimation encompasses the quantities, unit prices, and risks covered by the contract clauses. During the execution of a project, there is a rapid pace in construction, procurement, and the execution of contracts with subcontractors. International construction companies typically revise their completion cost estimates periodically during the execution phase. However, these predictions include contractual clauses and some contingencies based on the executives' perspectives rather than risk score evaluations in risk registers. On the other hand, risk registers are updated periodically, and these scores generally cannot be evaluated in terms of cost effects. Consequently, it is essential to implement these risk scores from risk registers to the projected completion cost during the execution phase using a systematic approach.

Additionally, this research focuses on predicting the completion cost instead of the causality relations of results. At this point, machine learning techniques yield robust results for predictive analysis, although statistical methods are effective in proving causation. Therefore, this study suggested machine learning-based prediction methods to predict the completion cost of construction based on the total risk score at any time of the project.

Furthermore, the proposed approach provides results quickly and effectively. It does not require a detailed breakdown of costs and risks; instead, it uses total risk scores to predict the total completion cost. This efficiency allows for its seamless integration into the execution phase, providing executives with a clear understanding of the financial implications of total risk scores. Moreover, the total completion cost of the project can be predicted at any time without the need for extensive data collection and processing.

To sum up, the primary aim of this research is to provide a systematic approach for estimating completion costs by considering the evaluation of potential risks throughout the execution phase of construction projects. To demonstrate the practical application of the proposed approach, the model was implemented on construction projects obtained from a globally operating construction company.

2. LITARATURE REVIEW

2.1 Previous Studies on Cost Estimation in Construction

The literature provides a range of methods for estimating costs in construction. Chan and Park (2005) employed principal component regression to mitigate the multicollinearity problem in project cost estimation. They aimed to discover the crucial components influencing project cost to enhance accuracy in project cost estimation. Wu et al. (2014) examined the influence of building information modeling (BIM) on calculating construction costs. They also proposed specific BIM tools that can be used to achieve precise cost estimation. Furthermore, the authors emphasized the necessity of standardizing the usage of BIM tools in quantity surveying to promote their broader acceptance in cost estimation. Ofori-Boadu (2015) conducted a study to develop models for predicting costs for high-rise buildings. During the research, the factors that influence costs were identified, and the most effective model was found as the natural logarithm of the building cost per square foot. Similarly, Thomas and Thomas (2016) developed a regression analysis model using 51 historical data sets to predict construction costs. Leśniak and Zima (2018) conducted a study using case-based reasoning (CBR) to determine the costs of constructing sports

fields. They identified 16 elements that have an impact on construction costs. Lee et al. (2022) studied BIM-based virtual reality to accurately estimate construction projects' cost. The acceptance of BIM technology was attributed to the critical factors of cost savings and resource utilization. Recent studies also specifically examined machine learning methods, particularly artificial neural networks (ANN). Matel et al. (2022) researched ANN to improve the precision of cost prediction for engineering services. They found seven key input variables that significantly impact cost estimation accuracy. Similarly, Antoniou et al. (2023) utilized ANN and multilinear regression to estimate the construction cost of underground metro station projects in Greece. They provided a formula that specifies the permissible deviation percentages based on the limits of the studies. Vakaj et al. (2023) estimated offsite construction costs utilizing ontology. They focused on activity-based cost estimation for offsite production. The offsite housing ontology connects activities with their associated prices, attributes, and regulations. Subsequently, knowledge-based engineering was utilized for automated cost estimation. The proposed method yielded consistent findings with traditional cost estimation based on results from 20 cases. Fernando et al. (2024) focused on developing a precise and user-friendly cost estimation model for concrete bridge systems during the initial stages of construction. They achieved this by utilizing artificial neural network technology (ANN). Based on the case study's findings, the ANN model had an accuracy rate of 90% and was deemed an appropriate method for cost assessment. Additionally, regression analysis has remained significant over the years. Atapattu et al. (2024) employed multiple regression analysis to predict the final cost of road projects in New Zealand using data gathered during the pre-design phase. They reported the mean absolute percentage error of the formed eight models and their components. Pishdad and Onungwa (2024) investigated the integration of 5D BIM into construction processes, particularly for cost estimation, control, and payments. They identified the challenges to this adaptation using interviews and a case study. Their analysis indicates that 5D BIM deployment enhances cost estimation in the conceptual phase of a project.

Moreover, there are studies on comparison cost estimation methods. Kim et al. (2005) compared neural networks, regression analysis, and case-based reasoning for cost estimates using a 530 historical data points dataset. Although the neural network (NN) yielded higher outcomes for the dataset, the case-based reasoning (CBR) approach produced better results in the long term. El-Sawah and Moselhi (2014) conducted a comparative study on neural networks to estimate costs in timber bridge projects and low-rise steel buildings.

When focusing on risk-based cost estimation in literature, a substantial body of knowledge exists regarding the practice of estimating costs based on risk. In their study, Maher and McGoey-Smith (2006) examined 100 major infrastructure projects and utilized sensitivity analysis to determine the relative significance of risks in terms of cost and time. Subsequently, they obtain ranges of cost/duration along with their respective probabilities, resulting in more precise cost estimation. Liu and Napier (2010) conducted a study on 11 water infrastructure projects in Australia to analyze the estimation of costs depending on risk. Risks in the study are assessed by measuring both inherent and contingent risks, where contingent risks account for the probability of occurrence. The authors concluded that risk-based estimation provides superior accuracy compared to conventional estimation methodologies. Ökmen and Öztaş (2010) studied a correlated cost risk analysis model. This model examines project cost in situations of uncertainty by considering risk variables and simulation properties. While not generally applicable, this model was utilized in a hypothetical construction project and yielded realistic outcomes reflecting the inherent unpredictability of project costs. The study by Yildiz et al. (2014) focused on developing a Knowledge-Based Risk Mapping Tool. This tool is designed to identify risk variables that contribute to cost overrun in construction projects. Additionally, the tool can forecast prospective risk pathways and their effect on project costs. Moreover, Lhee et al. (2016) focused on allocating cost contingency in transportation construction projects to account for unexpected risks. Their findings indicate that particle swarm optimization yields superior results in predicting cost contingencies. Kang and Kim (2018) estimated the cost associated with potential risks and the price to be offered for a new plant building project by analyzing the risks gathered from previous completed projects. Sadeh et al. (2021) examined the cost impact of construction risks by employing a fuzzy Monte Carlo simulation methodology. They developed an advantageous methodology to predict the influence of risk on project costs. Agarwal and Kansal (2020) investigated risk-based cost assessment in hydropower projects at the initial phases utilizing multicriteria decision-making methodologies. The results of the proposed methodology were satisfactory compared to the actual values. Draleti et al. (2024) studied risk-based cost estimate models to enhance risk management in cost estimation. They identified the five primary risk factors that significantly impact cost estimation through the analytic hierarchy process. Rezaee Arjroodya et al. (2024) employed risk analysis and simulation techniques to obtain precise cost and time estimations. Their sensitivity analysis revealed a positive

correlation between the likelihood of critical risks occurring and the probabilities of critical risks, as well as cost and time overruns.

The literature review on cost estimation in construction reveals that the performance of cost estimating methods primarily relies on the characteristics of the data. Moreover, the approaches used in the literature are found to mostly rely on previous data obtained through tender pricing or parameters in design stages. Apparently, there is a need for a robust approach, which enables contractors to predict the completion cost during the execution of the projects considering the updates in risk registers and cost studies. In this context, machine learning can be a very useful method, which can be attributed to its proven performance and the incorporation of diverse methodologies.

2.2 Machine Learning (ML)

Machine Learning (ML) is a subfield of artificial intelligence that focuses on the development and structure of systems by learning from data. ML involves comprehensive statistical methodologies that empower machines to enhance their task performance through experience. Similarly, it has the capacity to acquire knowledge and improve performance via experience without the need for explicit programming, (Samuel, 1959). Moreover, ML algorithms can identify patterns by automatically detecting regularities (Bishop and Nasrabadi, 2006).

ML algorithms are prevalent in many industries but have not yet gained widespread adoption in the construction industry (Park et al., 2022). To provide some literary examples of ML applications in the construction industry, Poh et al. (2018) employed ML approaches to examine key safety metrics and classify construction projects based on their safety risk levels. The researchers utilized data obtained from a contractor based in Singapore and applied five distinct ML algorithms to model safety data. Based on the findings, the RF algorithm yielded the most optimal results in attaining a superior level of precision and concurrence with the actual safety data. Choi et al. (2020) used ML techniques to examine more than 100,000 data points in Korea to develop a prognostic model for fatal incidents in the construction industry. RF algorithm performs better in predicting outcomes, while logistic regression highlights the significance of particular parameters in forecasting fatal accidents. Gondia et al. (2020) employed ML algorithms to predict project delay risks. As a result of their research, the Naïve Bayesian model generated better results in terms of performance indicators, and writers recommended proactive risk management in the construction industry. Furthermore, research is dedicated to enhancing the efficiency of construction projects by creating and verifying predictive models that accurately estimate costs and durations (Sanni-Anibire et al., 2021). Similarly, Mohamed and Moselhi (2022) studied improving cost and time estimation. The study utilized ML approaches to analyze highway building projects, specifically focusing on previously unstudied combinations of variables, and their approach yielded successful results. Furthermore, Fitzsimmons et al. (2022) researched to examine project risk in construction scheduling by employing hybrid ML techniques. The objective was to enhance the accuracy of risk prediction compared to conventional methods, which ultimately produced superior outcomes. Zhou and Flood (2024) examined the utilization of ANN and reinforcement learning techniques to fabricate construction components. Lee and Yun (2024) researched cost prediction in the construction industry by employing ML algorithms and improving forecast accuracy through efficient data preprocessing techniques. Besides, Cheng and Khasani (2024) utilized ML techniques to enhance cost estimation in construction projects, adapting to real-time fluctuations. Depending on the model accuracies, the proposed method outperformed traditional approaches. Mhady et al. (2024) conducted a study to predict the cost at completion with a hybrid model incorporating neuro-fuzzy inference systems and artificial neural networks combined with the Archimedes optimization algorithm. Their objective was to improve the accuracy and reliability of project cost control through the provided model. The artificial neural network utilizing the Archimedes optimization technique surpassed the performance of individual and other hybrid models.

Additionally, some studies focus on conducting literature reviews and analyzing research trends related to ML in the construction industry. Akinosho et al. (2020) conducted a comprehensive review of advanced ML algorithms, specifically emphasizing their application to health and safety, predictive modeling, and energy demands within the construction industry. The study's findings suggest that deep learning models have demonstrated promising levels of accuracy in forecasting a range of factors in the construction field. Darko et al. (2020) conducted a scientometric review of the use of artificial intelligence in the construction sector. They aimed to identify current trends and potential areas for future research. Sanchez-Garrido et al. (2023) employed ML-integrated methodologies to categorize, highlight research trends, and identify deficiencies of modern construction methods

in the literature. Moreover, Ali et al. (2024) employed ML algorithms to ascertain prospective construction engineering and management research trends.

From the literature, it is evident that ML algorithms possess a high degree of adaptability. For instance, they efficiently handle data variables with simple linear or non-linear interactions and manage variables with complex higher-order connections and even discrete variables (Gondia et al., 2020). Furthermore, ML algorithms tend to have high predictive accuracy due to their optimization process, which maximizes the number of correct forecasts and minimizes the number of incorrect predictions (Gondia et al., 2020). Therefore, ML algorithms were preferred in the proposed approach due to their prediction power and flexibility.

To sum up, several research have been undertaken on risk and cost estimation for construction projects, as well as the implementation of machine learning in this industry. However, the majority of cost estimating research has focused on the early stages (e.g., Atapattu et al., 2024; Fernando et al., 2024; Leśniak and Zima, 2018; Matel et al., 2022; Ofori-Boadu, 2015). However, the assumptions in the early stages might change depending on the dynamic characteristics of construction. Limited research produced an approach that could be implemented throughout the execution phase of construction to predict costs (e.g.,Cheng and Khasani, 2024; Mhady et al., 2024). Nonetheless, these methods do not utilize risk scores from registries, and due to the dynamic nature of these risks over the project duration, it is essential to integrate the effects of these generated risk scores into the total cost, particularly the execution time. The primary aim of this research is to provide a systematic approach for predicting completion cost by considering the evaluation of potential risks throughout the execution phase of construction projects with the explained methodology in the following sections.

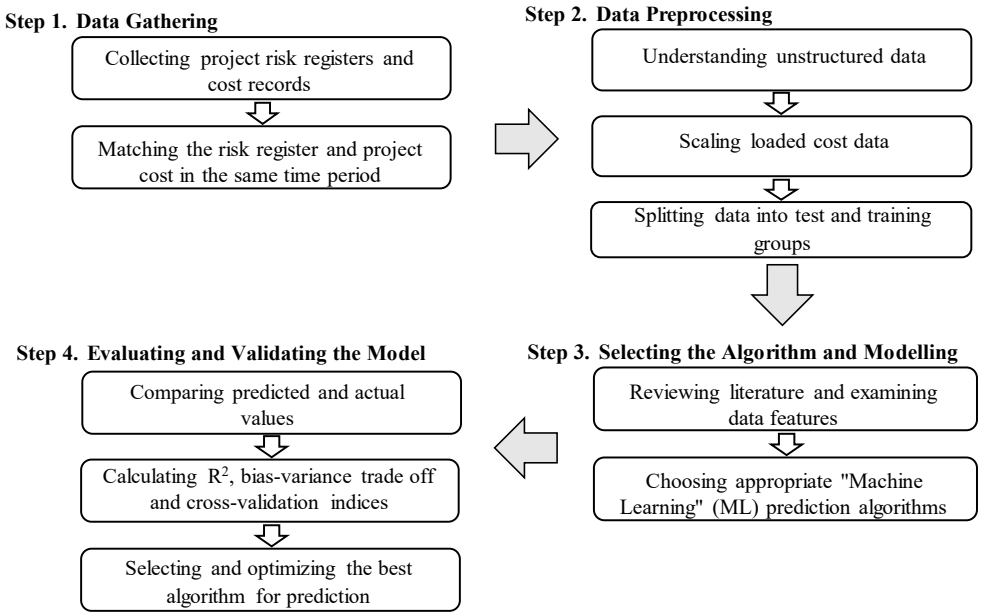

Figure 1: Steps of the proposed approach.

3. PROPOSED COMPLETION COST PREDICTON APPROACH

Predicted total cost and risk scores are the primary sources of the proposed approach. Total cost is estimated from the early stages of the project and requires updates over the project's duration. Additionally, risks are periodically evaluated through risk registers to give a sense to the executives. Executives may implement a mitigation plan based on the total risk scores reported with these registers. Although in the tender stage, the financial effect of these risk scores is generally included in the total cost prediction, in the execution phase, cost predictions mainly depend on the changing quantities, prices, and contract conditions. The proposed approach concentrates on gathering the project's parallel time estimations of completion cost and total risk scores. The objective is to train appropriate ML algorithms and develop a model that predicts the total completion cost based on future reported total risk scores in risk registers, contingent upon the relationships among these variables. Construction companies

can employ this methodology for their projects by utilizing historical or available records alongside appropriate machine learning algorithms, contingent upon their data and project characteristics.

This section outlines the steps for the suggested approach to predict the completion cost in construction projects based on the risk score employing ML. The proposed approach consists of four primary steps, as illustrated in Figure 1, and these steps are detailed in the following subsections.

3.1 Step 1. Data Gathering

In the context of this study, the cost of completing the project is planned to be linked to the total risk score data. To facilitate the implementation of the suggested methodology, the completion cost will be forecasted based on the total risk score rather than relying on a breakdown structure. Furthermore, aligning the overall risk score and estimated cost data from the identical time frame is vital. The following sections describe collecting the project risk score and cost estimation and combining these items in detail.

3.1.1 Gathering Risk Registers of the Projects

A risk register is an effective tool for documenting and presenting all potential risks during the initial tendering phase and guiding risk management throughout construction. According to Hillson (2003), a thorough awareness of the risks facing a project and organization is crucial to achieving successful and effective risk management. According to Laryea and Hughes (2008), contractors utilize the risk register approach to evaluate and determine the cost of risks at the tender stage. The process commences with generating ideas and proceeds with data identification, analysis, and evaluation using a spreadsheet. Risk registers generally consist of sections that provide definitions, probabilities of the risks, impacts, and mitigation strategies. Various methodologies might be employed in producing risk register reports for construction projects depending on the project team's and client's needs. A standardized structure is beneficial for maintaining consistency when compiling these documents.

The total risk score is determined by multiplying the probability (P) and impact (I) of each individual risk item (PxI). The project team determines and evaluates these risk items. Some projects might calculate the impact value of risk from different components (e.g. cost, time, quality, health, safety, environment). The impact value is calculated as the average of these values.The total risk score is calculated by taking the average of the total probability impact multiplication of items as shown in Equation 1.

$$Total\ Risk\ Score = \frac{\sum_{i=1}^{n}(P_i \times (\frac{1}{m}\sum_{j=1}^{m} I_{i,j}))}{n} \quad (1)$$

Where P_i is probability of i^{th} risk item, I_i is impact value of i^{th} risk item on the j^{th} component, m is the total number of impact value components and n is the total number of risk items.

The total risk score is calculated from probability and impact values. While these values generally range from 1 to 5 and result in a total risk score from 1 to 25, Equation 2 is proposed to be scaled for different possible total risk score calculations in case different projects are to be evaluated together.

$$TRS_{scaled} = \frac{(TRS_{original} - min_{original})}{(max_{original} - min_{original})} \times (max_{sclaed} - min_{sclaed}) + min\ _{scaled} \quad (2)$$

Where TRS_{sclaed} is the scaled total risk score, $TRS_{original}$ is the original total risk score, $min_{original}$ and $max_{original}$ are the minimum and maximum values of the original scale, and min_{scaled} and max_{scaled} are the minimum and maximum values of the new scale.

3.1.2 Gathering the Estimated Cost Data of the Projects

The cost estimation process in construction is essential for ensuring high project management quality and technical performance (Adeli and Wu, 1998; Derakhshanalavijeh and Teixeira, 2017). Precise cost estimation is crucial when considering project feasibility, financial requirements, and financing loans (Ahiaga-Dagbui and Smith, 2014). The tender cost of the construction project is adjusted after winning the bid and signing the contract based on new information that becomes available after the contract commencement. The tender cost is updated upon receiving detailed project information, and the forecasted completion cost of the project studies phase is initiated. Given the considerable diversity in construction projects, it is beneficial to revise predicted costs periodically. The

completion cost of the project is mainly calculated using the total direct and indirect costs. Direct costs include predictions of quantities and unit prices, whereas indirect cost calculations mostly rely on the predicted project duration and monthly expenses for indirect items. This study proposes the collection of periodically predicted completion cost data for construction projects. For simplicity, the total cost of the project should be considered instead of the cost breakdown structure. Since the primary emphasis is on the project's completion cost and the updated cost value during the execution phase.

3.1.3 The Association between Risk Factors and Cost Items

Within the scope of the proposed approach, the estimated completion cost of the construction projects should be associated with the total risk score. It is advisable to link the total cost and risk score together to facilitate operations rather than analyzing them in breakdowns. The estimated total risk score and completion cost must correspond to the same time frame for the project. In addition, since cost management and project risks may differ depending on the project type, it is recommended to classify the data by project type (Kim et al., 2017; Kim et al., 2024).

3.2 Step 2. Data Preprocessing

Data should be organized properly within the framework of this study to use prediction models in ML. The study includes total completion cost and total risk scores. The cost of each project varies based on its size, complexity, and structure. If different projects will be used together, it is advisable to employ standardization, as shown in Equation 3, for cost values, as it ensures that all cost data values are standardized to a consistent order of magnitude.

$$Y = \frac{(X - \mu)}{\sigma} \tag{3}$$

Where Y is the standardized total completion cost value X is the total completion cost value in a related project, μ is mean of the sample and σ is the standard deviation of the sample.

Subsequently, the input dataset is randomly split into training and testing datasets, with two-thirds allocated to the training dataset and one-third allocated to the testing dataset. The testing set was randomly selected and set aside to evaluate the model's performance on unseen data, while the training set was utilized to calibrate the model.

3.3 Step 3. Selecting the Algorithm and Modelling

This study aims to predict total completion costs based on total risk scores. Instead of statistical methods, ML algorithms are advised for their efficacy, rapid learning capabilities, and superior predictive attributes. ML techniques can be classified into three main categories: supervised learning, unsupervised learning, and reinforcement learning. The categories most frequently employed for ML problems are supervised and unsupervised, as shown in Table 1. Supervised learning leverages the available information from a dataset with labels to acquire a function that effectively approximates the relationship between input and labeled output in the data (Xie et al., 2020). An unsupervised learning strategy is used when the dataset lacks labels and predicts outcomes by identifying patterns in the underlying data. In unsupervised learning, each observation is associated with a vector response instead of a specific reaction, and it creates a more complex situation compared to supervised learning (James et al., 2023).

ML techniques possess various potential applications and advantages depending on their respective categories. These algorithms can accurately represent complex relationships between variables, notably independent variables and variables with nonlinear or simple linear relations (Gondia et al., 2020). When selecting an ML algorithm for a problem, the sample size and the data's characteristics can be used as a reference for selecting the most suitable model (Moreno, 2020). Nevertheless, no definitive ML technique is ideal for addressing specific problems (Mahmoodzadeh et al., 2022). Therefore, it is crucial to evaluate the effectiveness of the ML algorithms used.

The primary aim of this study is to predict the completion cost of the project. The cost value could be defined as quantity, and the data were labeled. Therefore, supervised learning was employed due to the characteristics of the data. Moreover, regression methods are advised because the primary purpose is to predict the total completion cost. However, the selected ML techniques and algorithms might change depending on the data properties. Decision-makers should make the proper selection depending on the characteristics of their gathered data. The implementation section provides comprehensive descriptions and parameters of the ML algorithms employed based on the data gathered in this study.

Table 1: Main ML techniques with subclasses and algortihm examples.

ML Techniques			
Supervised Learning		**Unsupervised Learning**	
Classification	**Regression**	**Clustering**	**Dimensionality Reduction**
Logistic Regression for Classification	Linear Regression (LR)	K-Means	Principal Component Analysis
K-Nearest Neighbors	Polynomial Regression (PR)	Gaussian Mixture	Linear Discriminant Analysis
Support Vector Classifier	Support Vector Regression (SVR)	Spectral Clustering	Latent Dirichlet Analysis
Naïve Bayes	Decision Tree (DT)	Hierarchical Clustering	Isomap
Decision Tree Classifier	Random Forest (RF)		Autoencoder
Random Forest Classifier	Artificial Neural Network (ANN)		
Artificial Neural Network	Gradient Boosting Tree (LGBM, XGBoost)		

3.4 Step 4. Evaluating and Validating the Model

The coefficient of determination (R^2) approach suggests a stronger relationship between the observed and predicted values. Calculating the R^2 score model's accuracy performance is assessed by comparing the model's output to the target value. The evaluation of the R^2 is conducted on three distinct groups: the training group, the entire dataset, and the validation group in the frame of this study.

The R^2 value is calculated using the data points with the regression model derived from the data set by Equation 4.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

Where y_i is observed i^{th} value, $f(x_i)$ is predicted value of y_i and $\bar{y}$ is mean value of a sample.

Furthermore, the R^2 with a value close to 1 indicates a stronger correlation between observed and predicted values. In this study, it is proposed that the threshold of R^2 be set at 0.75 to evaluate the accuracy performance of the model. The process of model evaluation and validation is depicted in Figure 2. The training and testing groups are calculated in step 1, and the models surpass the threshold value to proceed to the next phase. Subsequently, the R^2 values are computed for the test groups to assess the bias-variance tradeoff. The third step is resampling using k-fold cross-validation. Ultimately, model optimization is achieved by improving the properties of the selected algorithms.

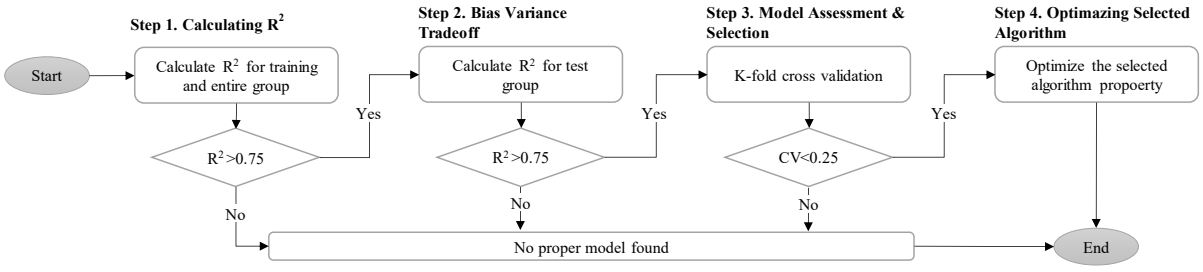

Figure 2: Model evaluation and validation methodology.

4. IMPLEMENTATION OF THE PROPOSED APPROACH: A CASE STUDY

This research includes a case study to demonstrate the proposed approach's practical application. Data regarding total costs and risk assessments were collected from a general contractor that specializes in the construction of airports and engages mainly in international projects.

4.1 Data Gathering

The total risk scores and completion cost estimations for 11 projects were collected from the internationally operating construction company. The risk register and the cost estimation reports, aligned within the same time period, were selected, and this is a data point from a project. The number of data gathered from these projects is demonstrated in Figure 3. According to this figure, 119 risk registers and completion cost estimation reports were gathered, which were produced in the same time period.

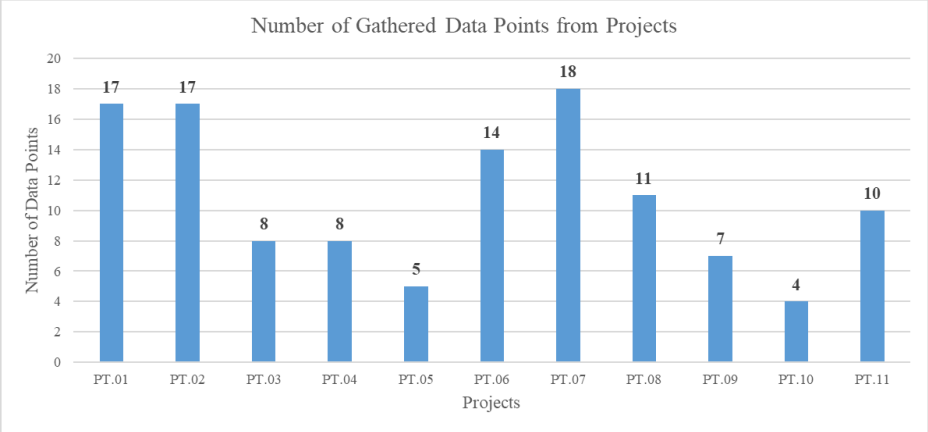

Figure 3: Number of data points from different projects.

Depending on the project characteristics and the priorities of the project managers, various formats were used in creating risk register reports and completion cost studies for construction projects.

The different risk register formats collected from projects and their standardized structure are shown in Figure 4.

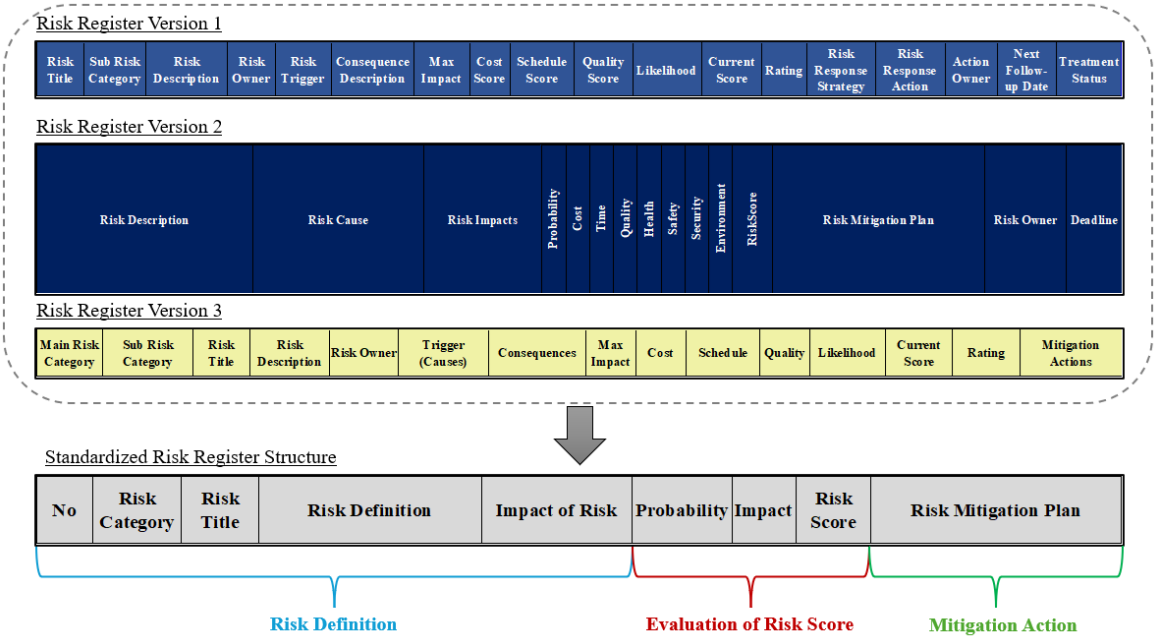

Figure 4: Standardized structure of risk register.

In the risk score calculation, all projects used the same range of probability and impact values ranging from 1 to 5. Therefore, no scaling formula was required for the total risk score.

A uniform framework, which provides the total risk score and the estimated total completion cost in parallel, was established and applied to all project risk registers and total costs in order to ensure uniformity in the compilation of these documents.

Figure 5 displays the collected completion cost formats obtained from several projects. Rather than utilizing breakdowns, the total cost, covering both direct and indirect expenses, was employed.

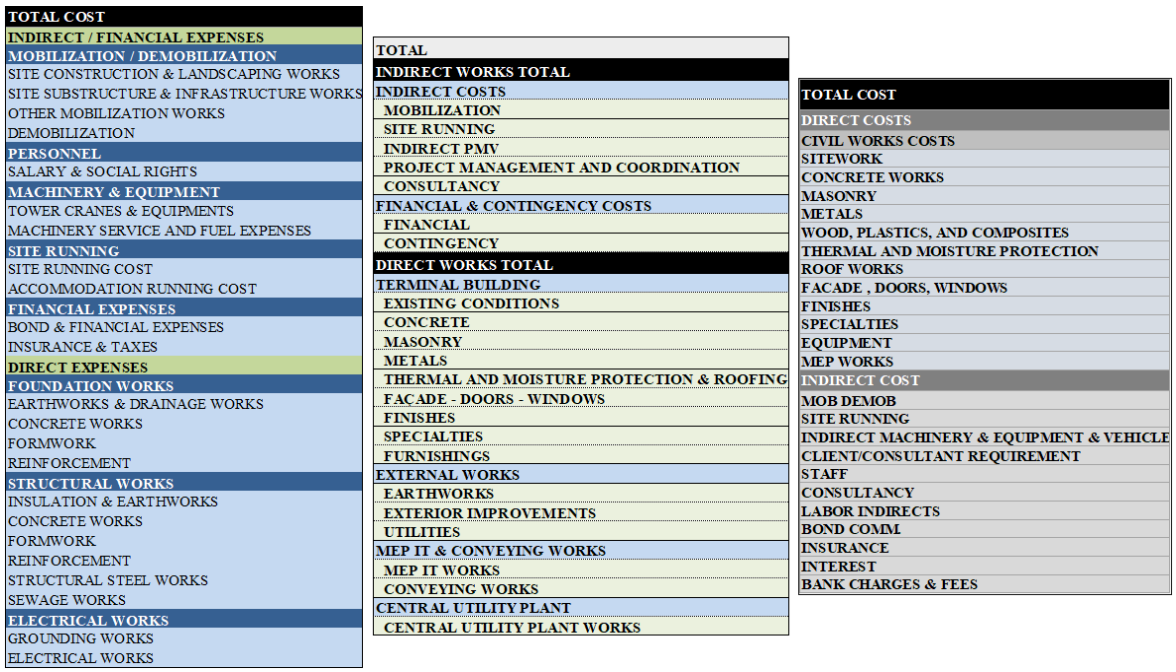

Figure 5: Completion cost formats.

Subsequently, the projects were categorized according to their respective project types depending on the construction project categories, as shown in Table 2.

Table 2: Project types and groups.

Project ID	**Project Type (PT)** **(Residential Projects = 01/ Heavy Civil Projects = 02)**
P.01	PT.01
P.02	PT.01
P.03	PT.02
P.04	PT.02
P.05	PT.01
P.06	PT.01
P.07	PT.02
P.08	PT.02
P.09	PT.02
P.10	PT.02
P.11	PT.02

The PT.01 group consists of four projects with 53 data points, while the PT.02 group comprises seven projects containing 66 data points. The risk scores were utilized directly, whereas the cost data were standardized for each project to eliminate cost level variations.

The data from each project is illustrated in Figure 6.

Figure 6: Project based relationship between standardized total cost and total risk scores.

There exists a significant linear correlation between the total risk score and cost to complete of the project within the PT.01 (R^2 = 0.859) and PT.02 (R^2= 0.753) categories.

4.2 Data Preprocessing

Python 3.11 programming was utilized to develop predictive models and obtain results. After completing the data interpretation, data preparation, and library loading in Python, the data frame was transformed into an array. Subsequently, the dataset was partitioned into training and test groups using the "train test split" function. Two-thirds of the dataset was allocated for training, while the remaining one-third was reserved for validation.

4.3 Selection the Algorithm and Modelling

As stated in the methodology section, supervised learning regression methods were utilized to predict the completion cost of projects. The subsequent algorithms were utilized since the dataset consists of fewer than one thousand data points. These are widely used and user-friendly algorithms. Moreover, they can give satisfactory performance in small-sized data sets (Aggarwal, 2015; Gondia et al., 2020):

- Linear Regression (LR): In the early 19th century, Legendre and Gauss published articles discussing the method of least squares, which is considered the earliest form of linear regression (James et al., 2013; Stigler, 1981). Linear regression is a fundamental technique employed in supervised learning, known for its inherent constraint of only producing linear functions. In the framework of this study, the total risk score is the independent variable within linear regression parameters, and the standardized total completion cost is the dependent variable. The general principle of linear regression is given in Equation 5.

$$Y = \alpha + (\beta \times TRS) + \epsilon \tag{5}$$

  Where Y is the standardized total completion cost, α is the intercept, β is the slope, TRS is the total risk score and ϵ is the error term;

- Polynomial Regression (PR): Polynomial regression enhances the linear model's capacity to incorporate non-linear relationships using polynomial functions. This enhancement is accomplished by utilizing non-linear correlation. 2nd (PR2) and 4th (PR4) degree polynomial functions were employed to assess their efficacy in this study. PR parameters are similar to linear regression, and the total risk score is an independent variable. In this method, the standardized total completion cost is predicted,

including quadratic, cubic, and more complex structures compared to LR. The fundamental formulas for PR2 and PR4 are given in Equation 6 and 7.

$$Y = \alpha + (\beta_1 \times TRS) + (\beta_2 \times TRS^2) + \epsilon \quad (6)$$

$$Y = \alpha + (\beta_1 \times TRS) + (\beta_2 \times TRS^2) + (\beta_3 \times TRS^3) + (\beta_4 \times TRS^4) + \epsilon \quad (7)$$

Where Y is the standardized total completion cost, α is the intercept, β values are the coefficinets for different degree terms, TRS is the total risk score and ϵ is the error term;

- Support Vector Regression (SVR): SVR is a versatile technique that may be used for classification and regression applications. The classification characteristic is converted into a regression attribute by minimizing a particular sort of loss function. The loss function in question only considers residuals that exceed a positive constant in absolute value (James et al., 2013). SVR finds the function that maximizes the quantity of points within the margin boundaries. The parameters of SVR are epsilon (Ɛ), regularization parameter (C), kernel function type (K), and gamma value (ɣ). Epsilon defines a margin of tolerance. The regularization parameter indicates the balance between the model's complexity and the allowed prediction error. Gamma indicates the influence of a single training point. For Ɛ, C, and ɣ, the default values in Python are taken as 0.1, 1.0, and 1.0, respectively. The kernel function specifies the kernel type in the algorithm. The radial basis function (RBF) was chosen since, despite the other possibilities, RBF has a strong approximation ability (Shoar et al., 2022). If SVR yields superior results, the parameter selections can be optimized using the GridSearchCV class in Python, which identifies ideal hyperparameters by evaluating combinations;
- Decision Tree (DT): DT methodology is frequently utilized in the field of data analysis for the objectives of classifying and predicting outcomes. The methodology employed in this study entails the identification of distinct essential regions inside the predictor space through segmentation. The performance of the DT regressor depends on some parameters. The criterion parameter determines the function, which evaluates the split's quality. Splitter is a splitting method at each node. min_samples_split determines the minimum sample size necessary to split an internal node. min_samples_leaf is the minimum sample size necessary in a leaf node. Criterion, splitter, min_samples_split, and min_samples_leaf parameters were taken as squared_error, best, 2.0, and 1.0, respectively, as default options provided in Python. If this method shows superior results, the parameters can be optimized using the GridSearchCV class in Python;
- Random Forest (RF): RF algorithm has a higher degree of flexibility in comparison to LR model. RF provides an improvement over bagged trees by incorporating a slight modification that promotes decorrelation among the individual trees. In a manner similar to the bagging technique, the construction of numerous decision trees is facilitated by utilizing bootstrapped training samples (James et al., 2013). RF utilizes ensemble learning algorithms. The parameters criteria, min_samples_split and min_samples_leaf, have identical meanings in DT and are assigned the same default values. The n_estimator parameter specifies the quantity of trees in the forest, set to 10 due to the small sample size. These parameters could be optimized if RF gives superior results.

4.4 Evaluating and Validating the Model

4.4.1 Calculating R^2

PT.01 contains 18 testing data, while PT.02 has 22 testing data, equivalent to one-third of the total datasets in each group. Figure 7 displays the standardized cost prediction results produced from the testing set of groups and the corresponding actual values of these points. When examining the third point in the PT.01 group, it was found that RF and DT provided the most accurate results, while SVR yielded the least accurate result. On the other hand, in the PT.02 group, SVR and RF provided the most accurate predictions for the fifth point, while LR performed the poorest.

Figure 7: Prediction results of regression models.

The R^2 for all data group and training group were calculated and results were displayed in Table 3.

Table 3: R^2 for all data group and training group in PT.01 and PT.02.

Groups	Methods	R^2 for All Data	Rank for All Data	R^2 for Training Group	Rank for Training Group
PT.01	LR	0.859	5	0.887	6
	PR2	0.865	4	0.889	5
	PR4	0.890	3	0.920	3
	SVR	0.840	6	0.915	4
	DT	0.908	1	0.993	1
	RF	0.895	2	0.976	2
PT.02	LR	0.753	6	0.743	6
	PR2	0.800	5	0.796	5
	PR4	0.820	3	0.823	4
	SVR	0.812	4	0.824	3
	DT	0.901	1	1.000	1
	RF	0.895	2	0.962	2

All models satisfy the R^2 threshold for the PT.01 group, while LR in PT.02 could not meet the threshold of greater than 0.75. The DT model outperforms the other models for the overall dataset and the training group in PT.01 and PT.02.

4.4.2 Bias-Variance Trade-off

The concept of variance quantifies dispersion or variability within a given amount of data points. When the variance is high, the model function is greatly influenced by changes in the training set. On the other hand, bias refers to the discrepancy between estimates and the actual values inside a given dataset (James et al., 2013). Additionally, it is common for the training error rate to exhibit notable disparities with the test error rate; the test

error refers to the mean error that arises when employing a statistical learning technique to forecast the outcome of new observations (James et al., 2013). Therefore, R^2 for test data is calculated and shown in Table 4.

Table 4: R^2 for test groups in PT.01 and PT.02.

Groups	Actual Values	Average Diff.	Prediction LR	LR (Residual2)	Prediction PR2	PR2 (Residual2)	Prediction PR4	PR4 (Residual2)	Prediction SVR	SVR (Residual2)	Prediction DT	DT (Residual2)	Prediction RF	RF (Residual2)
PT.01	0.93	1.05	0.72	0.05	0.71	0.05	0.65	0.08	0.85	0.01	0.84	0.01	0.87	0.00
	(1.63)	2.38	(1.64)	0.00	(1.70)	0.00	(2.07)	0.19	(1.82)	0.03	(1.67)	0.00	(1.73)	0.01
	0.77	0.74	1.09	0.11	1.01	0.06	0.96	0.04	1.01	0.06	0.84	0.01	0.91	0.02
	0.97	1.13	0.80	0.03	0.78	0.04	0.70	0.07	0.93	0.00	0.44	0.29	0.76	0.04
	0.97	1.13	0.51	0.21	0.53	0.19	0.55	0.18	0.64	0.11	0.62	0.12	0.56	0.17
	1.13	1.49	0.94	0.04	0.89	0.06	0.80	0.11	1.00	0.01	0.36	0.58	0.61	0.27
	(2.40)	5.31	(1.42)	0.95	(1.44)	0.92	(1.90)	0.24	(1.77)	0.40	(1.67)	0.53	(1.59)	0.65
	0.17	0.07	(0.25)	0.18	(0.18)	0.12	(0.04)	0.05	0.02	0.02	0.33	0.03	0.40	0.05
	0.06	0.02	(0.04)	0.01	0.03	0.00	0.19	0.01	0.18	0.01	0.15	0.01	0.12	0.00
	(1.44)	1.82	(1.38)	0.00	(1.39)	0.00	(1.85)	0.17	(1.74)	0.09	(1.67)	0.05	(1.59)	0.02
	0.36	0.21	0.95	0.34	0.90	0.29	0.81	0.20	1.01	0.41	1.22	0.73	1.04	0.45
	(1.07)	0.95	(0.90)	0.03	(0.85)	0.05	(1.07)	0.00	(1.05)	0.00	(1.23)	0.03	(1.06)	0.00
	(1.09)	0.99	(0.52)	0.32	(0.45)	0.41	(0.42)	0.44	(0.33)	0.57	(1.04)	0.00	(0.58)	0.25
	(0.54)	0.20	(0.24)	0.09	(0.16)	0.14	(0.03)	0.26	0.03	0.32	0.33	0.75	0.40	0.88
	0.20	0.08	(0.15)	0.12	(0.08)	0.08	0.07	0.02	0.10	0.01	0.33	0.02	0.23	0.00
	0.13	0.05	0.15	0.00	0.21	0.01	0.35	0.05	0.31	0.03	0.15	0.00	0.24	0.01
	0.27	0.13	(0.07)	0.12	(0.00)	0.07	0.16	0.01	0.16	0.01	0.15	0.02	0.12	0.02
	0.54	0.41	0.32	0.05	0.36	0.03	0.45	0.01	0.44	0.01	0.71	0.03	0.55	0.00
Average:	*(0.09)*													
Sum:		*18.16*		*2.64*		*2.51*		*2.12*		*2.12*		*3.19*		*2.86*
R^2:				*0.85*		*0.86*		*0.88**		*0.88**		*0.82*		*0.84*
PT.02	(1.35)	1.75	(1.06)	0.09	(1.29)	0.00	(1.33)	0.00	(1.30)	0.00	(1.90)	0.30	(1.57)	0.05
	(1.04)	1.02	(0.33)	0.51	(0.15)	0.79	(0.26)	0.61	0.10	1.30	(0.92)	0.01	(0.75)	0.08
	(0.53)	0.25	(1.07)	0.29	(1.31)	0.61	(1.34)	0.66	(1.31)	0.60	(1.90)	1.88	(1.57)	1.09
	(0.65)	0.38	(0.80)	0.02	(0.85)	0.04	(1.02)	0.14	(1.05)	0.16	(1.78)	1.28	(1.63)	0.96
	(1.09)	1.12	(1.20)	0.01	(1.55)	0.21	(1.43)	0.11	(1.26)	0.03	(1.20)	0.01	(1.20)	0.01
	(1.23)	1.45	(1.17)	0.00	(1.48)	0.06	(1.41)	0.03	(1.29)	0.00	(1.58)	0.12	(1.35)	0.01
	1.52	2.40	0.76	0.58	0.96	0.31	1.11	0.16	1.19	0.11	0.86	0.44	0.91	0.37
	0.85	0.77	0.33	0.27	0.60	0.06	0.73	0.01	1.12	0.07	0.94	0.01	1.09	0.06
	0.77	0.65	(0.14)	0.83	0.09	0.47	0.06	0.51	0.57	0.04	(0.00)	0.60	(0.22)	0.98
	0.58	0.37	0.13	0.20	0.40	0.03	0.48	0.01	1.01	0.19	0.14	0.19	0.35	0.05
	(0.53)	0.25	(1.06)	0.29	(1.30)	0.59	(1.33)	0.65	(1.31)	0.60	(1.90)	1.88	(1.57)	1.09
	(0.77)	0.55	0.12	0.80	0.39	1.34	0.46	1.51	1.00	3.13	1.02	3.23	0.53	1.70
	1.10	1.28	0.83	0.08	1.01	0.01	1.15	0.00	1.21	0.01	1.00	0.01	1.04	0.00
	0.10	0.02	(0.48)	0.33	(0.36)	0.21	(0.51)	0.37	(0.31)	0.17	0.01	0.01	(0.07)	0.03
	1.23	1.59	1.00	0.05	1.11	0.01	1.22	0.00	1.25	0.00	1.35	0.02	1.15	0.01
	(0.23)	0.04	0.01	0.06	0.26	0.24	0.29	0.27	0.85	1.18	0.21	0.20	0.02	0.06
	0.77	0.65	(0.28)	1.10	(0.09)	0.74	(0.17)	0.89	0.24	0.29	0.10	0.46	(0.12)	0.80
	(0.68)	0.43	(0.93)	0.06	(1.07)	0.15	(1.20)	0.26	(1.23)	0.30	(1.26)	0.33	(1.21)	0.28
	1.17	*1.44*	*0.48*	*0.48*	*0.74*	*0.19*	*0.89*	*0.08*	*1.14*	*0.00*	*1.09*	*0.01*	*1.11*	*0.00*
	(0.78)	*0.56*	*(0.45)*	*0.11*	*(0.32)*	*0.21*	*(0.47)*	*0.10*	*(0.24)*	*0.29*	*0.01*	*0.63*	*(0.04)*	*0.55*
	0.43	*0.21*	*(0.26)*	*0.47*	*(0.07)*	*0.24*	*(0.14)*	*0.33*	*0.28*	*0.02*	*0.10*	*0.11*	*(0.12)*	*0.30*
	(0.32)	0.08	(0.09)	0.05	0.14	0.21	0.13	0.20	0.67	0.97	(0.47)	0.02	(0.19)	0.02

Average:	*(0.03)*						
Sum:	*17.25*	*6.68*	*6.74*	*6.92*	*9.49*	*11.74*	*8.51*
R^2:		*0.61**	*0.60*	*0.60*	*0.45*	*0.32*	*0.51*

* Superior R^2 results in testing groups within applied ML algorithms.

Among the computed group, PR4 and SVR provide greater performance in PT.01, whereas LR yields better results in PT.02 for both MSE and R^2 in the test data. Using the random state function in Python, ten states were randomly selected for each group, and all calculations were performed for each state. The results for the R^2 of testing groups in different random states are displayed in Table 5.

Table 5: R^2 for test groups in different random states for PT.01 and PT.02.

Groups	Random State vs R^2	LR	PR2	PR4	SVR	DT	RF
PT.01	1	0.855	0.862	0.883*	0.883*	0.824	0.843
	2	0.833	0.771	0.632	0.898*	0.751	0.770
	3	0.874	0.875	0.914	0.919*	0.720	0.853
	4	0.776	0.769	0.791*	0.651	0.689	0.705
	5	0.866*	0.865	0.849	0.397	0.728	0.769
	6	0.874	0.882	0.899*	0.858	0.783	0.878
	7	0.851	0.857	0.891*	0.870	0.771	0.846
	8	0.840	0.844	0.883*	0.693	0.715	0.767
	9	0.860	0.849	0.872*	0.849	0.807	0.839
	10	0.790	0.800	0.842*	0.644	0.768	0.785
	Average R^2	0.842	0.837	0.846*	0.766	0.756	0.805
	Rank	2	3	1	5	6	4
PT.02	1	0.613*	0.609	0.599	0.450	0.319	0.507
	2	0.707	0.756	0.779*	0.763	0.566	0.656
	3	0.785	0.787	0.802	0.812*	0.689	0.792
	4	0.709	0.795	0.811*	0.439	0.684	0.725
	5	0.674	0.758	0.742	0.766*	0.577	0.671
	6	0.725	0.735	0.738*	0.725	0.693	0.698
	7	0.683	0.763*	0.756	0.497	0.635	0.655
	8	0.672	0.818	0.839	0.847*	0.794	0.850
	9	0.794	0.776	0.819*	0.767	0.528	0.631
	10	0.708	0.781	0.782	0.799*	0.653	0.673
	Average R^2	0.707	0.758	0.767*	0.687	0.614	0.686
	Rank	3	2	1	4	6	5

* Superior R^2 results for each random state.

According to the findings, PR4, LR, PR2, RF, and SVR in the PT.01 group satisfy the R^2 requirement (>0.75), while PR4 and PR2 in the PT.02 group satisfy the R^2 threshold. PR4 demonstrates greater predictive performance for both groups when estimating standardized cost results based on total risk score, as measured by the R^2 criteria.

4.4.3 Model Assessment and Selection

Resampling methods are used for model assessment and selection. The process entails iteratively extracting samples from a training set and reconfiguring a model of interest on each sample to acquire supplementary insights about the fitted model (James et al., 2013). Cross-validation is one of the commonly used resampling methods and is employed to assess the performance of a statistical learning method by utilizing the corresponding test error. The test error refers to the mean error that arises when employing a statistical learning technique to forecast the

outcome of an entirely new observation. K-fold cross-validation is a widely used method for validating models by examining their testing performance. K-fold cross-validation partitions observations into k groups (folds) randomly, with the first fold used for validation and the remaining k − 1 folds used for fitting, and all procedures repeated k times. There is no precise rule about the fold size (Kuhn and Johnson, 2013). 10-fold cross-validation is employed in this research since 10-fold cross-validation is sufficient for model selection (Arlot and Celisse, 2010). There are 53 data points, with 18 assigned for testing in the PT.01 group and 66 in the PT.02 group, with 22 designated for testing. The number of possible testing group options can be determined by calculating the combination of 53 items taken 18 at a time for the PT.01 group and 66 items taken 22 at a time for the PT.02 group. This approach can be performed using the random state functionality in Python. A sample of 10 states was randomly selected to get findings for a 10-fold analysis in both groups.

Table 6: CV(k) for 10-fold of Cross Validation in PT.01 and PT.02.

Groups	Random State vs MSE	LR	PR2	PR4	SVR	DT	RF
PT.01	1	0.147	0.140	0.118*	0.118*	0.177	0.159
	2	0.182	0.250	0.402	0.111*	0.272	0.251
	3	0.136	0.136	0.093	0.088*	0.303	0.159
	4	0.234	0.241	0.219*	0.365	0.325	0.309
	5	0.175*	0.176	0.197	0.787	0.355	0.301
	6	0.105	0.099	0.084*	0.118	0.181	0.102
	7	0.203	0.195	0.150*	0.177	0.312	0.211
	8	0.144	0.140	0.105*	0.276	0.256	0.210
	9	0.155	0.167	0.142*	0.167	0.214	0.179
	10	0.223	0.212	0.168*	0.379	0.247	0.229
	$CV_{(k)}$	0.171	0.176	0.168*	0.259	0.264	0.211
	Rank	2	3	1	5	6	4
PT.02	1	0.304*	0.306	0.315	0.431	0.534	0.387
	2	0.292	0.243	0.220*	0.236	0.432	0.342
	3	0.208	0.207	0.192	0.182*	0.302	0.202
	4	0.280	0.197	0.182*	0.539	0.304	0.265
	5	0.252	0.187	0.199	0.181*	0.326	0.254
	6	0.269	0.259	0.257*	0.269	0.301	0.296
	7	0.333	0.249*	0.257	0.528	0.384	0.363
	8	0.293	0.162	0.144	0.136	0.183	0.134*
	9	0.189	0.205	0.166*	0.214	0.433	0.338
	10	0.319	0.240	0.239	0.220*	0.379	0.358
	$CV_{(k)}$	0.274	0.226	0.217*	0.294	0.358	0.294
	Rank	3	2	1	4	6	5

* Minimum CV results for each state.

The mean square error (MSE) is a widely used metric for evaluating the quality of fit of regression models by quantifying the discrepancy between the actual values and predicted values generated by a model as shown in Equation 8.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

Where n is number of data points, y_i is observed value and $\hat{y}_i$ is predicted value.

The cross validation (CV) is obtained by taking the average of MSE for 10-fold as indicated in Equation 9.

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i \tag{9}$$

Where k is the number of folds.

The CV results for 10 random states in PT.01 and PT.02 were presented in Table 6.

The results indicate that PR4, LR, PR2, and RF in the PT.01 group satisfy the CV requirement (<0.25), but only PR4 and PR2 in the PT.02 group meet it. PR4 gives superior results for both groups in predicting standardized cost results from total risk score in terms of CV criteria.

4.4.4 Optimizing Selected Algorithm

The findings in previous sections demonstrate that the PR4 algorithm is a highly effective regression method for PT.01 and PT.02 groups depending on the R^2, MSE, and CV measures in 10-fold validation evaluation. At this stage, it is essential to determine the optimal degree of polynomial regression model that accurately captures the association between total risk and cost. It is anticipated that as the degree of the polynomial increases, there will be a corresponding decrease in the training error. As the function retains and exhibits sensitivity to all data points, it exhibits a high degree of variation. The outcome is a significant increase in the mean square error observed in the test data. Conversely, when the degree of a polynomial function is low, it will result in a simplistic inference formula and cause bias.

Using polynomial features in Python programming, R^2 and MSE values were systematically calculated from the 1st to the 10th degree for different states. This process was aimed at detecting the optimal flexibility degree of a polynomial, which gives the highest R^2 value and lowest MSE. The detailed findings of this process are presented in Figure 8.

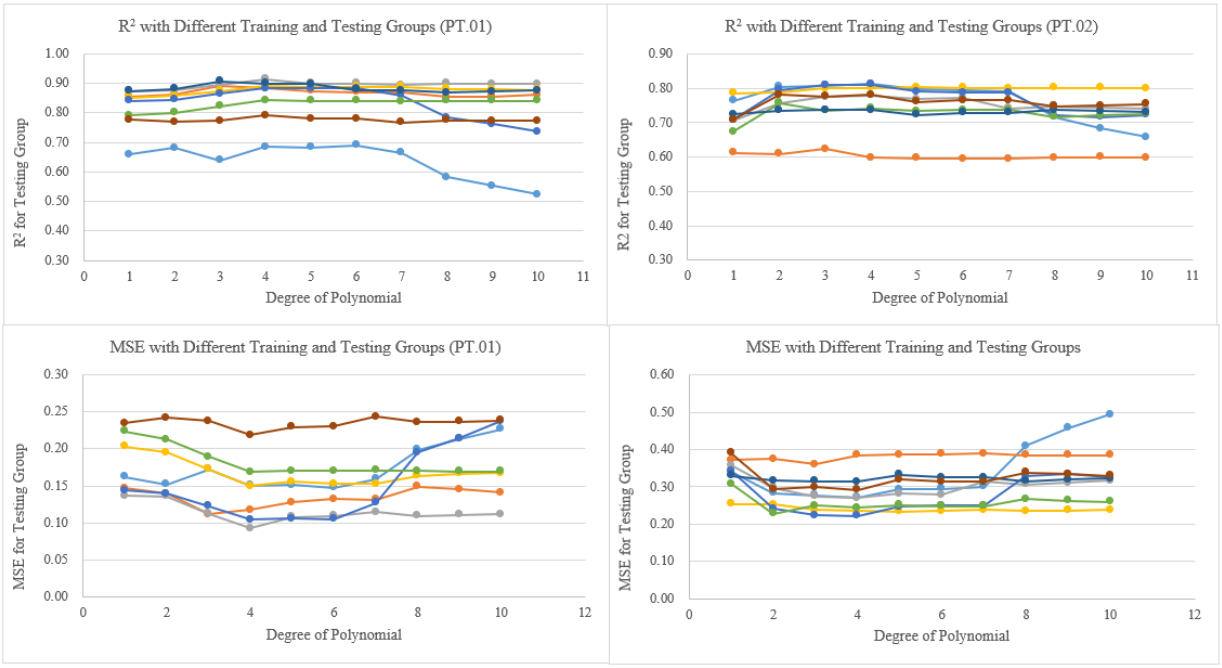

Figure 8: R^2 and MSE versus degree of polynomial regression model.

Based on the findings of the PT.01 group, PR4 emerged as the most optimal prediction method for estimating the standardized total cost value, as determined by the utilization of the average total risk score. For the PT.02 group, 3rd and 4th degree polynomial regressions exhibited the lowest MSE and the highest R^2 values. 3rd degree polynomial regression (PR3) gave slightly better results than PR4.

To sum up, this part aims to verify the accuracy of the models in predicting outcomes and to assess and compare their effectiveness. Based on the preceding analysis of both groups in the training and test datasets, it is evident that polynomial regression yields better outcomes. As a result, PR4 was chosen for PT.01, while PR3 was allocated to the PT.02 group due to their proficiency in predicting standardized total cost using the total risk score.

5. CONCLUSIONS

Forecasting costs in the construction industry is a crucial aspect. Despite thorough evaluations of risks and costs, they often cannot be effectively integrated due to the dynamic nature of activities in the construction business. This paper aims to provide construction companies with a risk-based prediction approach, which assists them in dynamically predicting the completion cost of the project, considering the updates in risk registers and cost studies during the execution. Six prediction algorithms were utilized, and 119 risk and cost data points were gathered and reported simultaneously. A total of 53 data points from 4 residential projects and 66 data points from 7 heavy civil projects were analyzed using Python programming. One-third of the data was allocated for testing, while the remaining portion was utilized for training in each group. Polynomial regression outperformed other algorithms in predicting the cost to complete the project based on the comparison of anticipated and actual results, as well as the consideration of resampling methods and bias-variance tradeoffs in the gathered data. After optimizing the system, it was found that the 3rd degree provided superior results for residential projects, while the 4th degree yielded better prediction results for heavy civil construction projects based on the data collected for this study.

This study aims to forecast the completion cost of a project during the execution phase in a reliable manner without requiring unnecessary workload. To reduce the tasks, the total risk score and total cost of the project were taken into account rather than breaking down the structure of the risks and costs. Another benefit of the proposed approach is its ability to create a prediction model utilizing the actual execution cost data from the projects. Furthermore, construction companies may take advantage of the suggested prediction approach in anticipating the completion costs of future projects at the beginning of the execution phase, which ultimately improves their forecasting and planning capabilities.

Alongside the benefits of the proposed approach, there are also some limitations. Firstly, the study can be enhanced by employing cost and/or risk breakdown structures to analyze the most closely associated components. Utilizing the breakdown structure can result in enhanced performance of alternative ML algorithms. Moreover, larger datasets can be assessed using alternative ML algorithms that are better suited for handling large amounts of data. Furthermore, exploring other types of construction projects in addition to residential and heavy civil projects may be worthwhile. The outcomes could differ if other companies gather different data on the features of their projects. Additionally, using alternative ML algorithms may prove advantageous, such as forecasting categorical outcomes. The authors' research and study on these issues persist. To sum up, the proposed model would be advantageous for future research and studies in the construction sector, particularly in simplifying the estimate of the completion cost of the projects based on risks.

REFERENCES

Adeli, H. and Wu, M. (1998). Regularization neural network for construction cost estimation, *Journal of construction engineering and management* 124(1): 18–24.

Agarwal, S. S. and Kansal, M. L. (2020). Risk based initial cost assessment while planning a hydropower project, *Energy Strategy Reviews* 31: 100517.

Aggarwal, C. C. (2015). *Data Mining: The Textbook*, Cham: Springer International Publishing. Retrieved from https://link.springer.com/10.1007/978-3-319-14142-8

Ahiaga-Dagbui, D. D. and Smith, S. D. (2014). Rethinking construction cost overruns: cognition, learning and estimation, *Journal of financial management of property and construction* 19(1): 38–54.

Akinosho, T. D., Oyedele, L. O., Bilal, M., Ajayi, A. O., Delgado, M. D., Akinade, O. O. and Ahmed, A. A. (2020). Deep learning in the construction industry: A review of present status and future innovations, *Journal of Building Engineering* 32: 101827.

Ali, G. G., El-adaway, I. H., Ahmed, M. O., Eissa, R., Nabi, M. A., Elbashbishy, T. and Khalef, R. (2024). Forecasting Future Research Trends in the Construction Engineering and Management Domain Using Machine Learning and Social Network Analysis, *Modelling* 5(2): 438–457.

Antoniou, F., Aretoulis, G., Giannoulakis, D. and Konstantinidis, D. (2023). Cost and material quantities prediction models for the construction of underground metro stations, *Buildings* 13(2): 382.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection, *Statistics Surveys* 4(none). Retrieved from https://projecteuclid.org/journals/statistics-surveys/volume-4/issue-none/A-survey-of-cross-validation-procedures-for-model-selection/10.1214/09-SS054.full

Atapattu, C. N., Domingo, N. and Sutrisna, M. (2024). A conceptual cost estimation model for the pre-design stage of road projects using multiple regression analysis, *Journal of Financial Management of Property and Construction*.

Aziz, R. F. (2013). Factors causing cost variation for constructing wastewater projects in Egypt, *Alexandria Engineering Journal* 52(1): 51–66.

Baratta, A. (2006). The triple constraint, a triple illusion, In *PMI® Global Congress*, Vol. 202.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, Vol. 4, Springer. Retrieved from https://link.springer.com/book/9780387310732

Chan, S. L. and Park, M. (2005). Project cost estimation using principal component regression, *Construction Management and Economics* 23(3): 295–304.

Cheng, M.-Y. and Khasani, R. R. (2024). Least Square Moment Balanced Machine: A New Approach To Estimating Cost To Completion For Construction Projects, *Journal of Information Technology in Construction (ITcon)* 29(23): 503–524.

Choi, J., Gu, B., Chin, S. and Lee, J.-S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers, *Automation in Construction* 110: 102974.

Darko, A., Chan, A. P. C., Adabre, M. A., Edwards, D. J., Hosseini, M. R. and Ameyaw, E. E. (2020). Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities, *Automation in Construction* 112: 103081.

Derakhshanalavijeh, R. and Teixeira, J. M. C. (2017). Cost overrun in construction projects in developing countries, Gas-Oil industry of Iran as a case study, *Journal of Civil Engineering and Management* 23(1): 125–136.

Doloi, H. (2013). Cost overruns and failure in project management: Understanding the roles of key stakeholders in construction projects, *Journal of construction engineering and management* 139(3): 267–279.

Draleti, G., Sengonzi, R. and Kakitahi, J. (2024). Improvement of Risk Management in Cost Estimation in the Building Construction Industry in Uganda, *Journal of Construction in Developing Countries* 29(1): 111–138.

El-Sawah, H. and Moselhi, O. (2014). Comparative study in the use of neural networks for order of magnitude cost estimating in construction, *Journal of Information Technology in Construction (ITcon)* 19(27): 462–473.

Fernando, N., TA, K. D. and Zhang, H. (2024). An artificial neural network (ANN) approach for early cost estimation of concrete bridge systems in developing countries: the case of Sri Lanka, *Journal of Financial Management of Property and Construction* 29(1): 23–51.

Fitzsimmons, J. P., Lu, R., Hong, Y. and Brilakis, I. (2022). Construction schedule risk analysis – a hybrid machine learning approach, *Journal of Information Technology in Construction* 27: 70–93.

Flyvbjerg, B., Holm, M. S. and Buhl, S. (2002). Underestimating costs in public works projects: Error or lie?, *Journal of the American planning association* 68(3): 279–295.

Gondia, A., Siam, A., El-Dakhakhni, W. and Nassar, A. H. (2020). Machine learning algorithms for construction projects delay risk prediction, *Journal of Construction Engineering and Management* 146(1): 04019085.

Hillson, D. (2003). Using a risk breakdown structure in project management, *Journal of Facilities management* 2(1): 85–97.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*, Vol. 112, Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J. (2023). Statistical learning, In *An Introduction to Statistical Learning: With Applications in Python*, Springer, pp. 15–67.

Johnson, R. M. and Babu, R. I. I. (2020). Time and cost overruns in the UAE construction industry: a critical analysis, *International Journal of Construction Management* 20(5): 402–411.

Kang, H.-W. and Kim, Y.-S. (2018). A Model for Risk Cost and Bidding Price Prediction based on Risk Information in Plant Construction Projects, *KSCE Journal of Civil Engineering* 22(11): 4215–4229.

Kim, G. H., Seo, D. S. and Kang, K. I. (2005). Hybrid Models of Neural Networks and Genetic Algorithms for Predicting Preliminary Cost Estimates, *Journal of Computing in Civil Engineering* 19(2): 208–211.

Kim, M., Lee, I. and Jung, Y. (2017). International project risk management for nuclear power plant (NPP) construction: Featuring comparative analysis with fossil and gas power plants, *Sustainability* 9(3): 469.

Kim, S., Ghimire, P., Jeong, H. D. and Barutha, P. (2024). Comparative Analysis of Project Risks across Construction Sectors, *Journal of Construction Engineering and Management* 150(6): 04024038.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*, New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4614-6849-3

Laryea, S. and Hughes, W. (2008). How contractors price risk in bids: theory and practice, *Construction Management and Economics* 26(9): 911–924.

Lee, H. and Yun, S. (2024). Strategies for Imputing Missing Values and Removing Outliers in the Dataset for Machine Learning-Based Construction Cost Prediction, *Buildings* 14(4): 933.

Lee, M., Chai, C., Xiong, Y. and Gui, H. (2022). Technology acceptance model for Building Information Modelling Based Virtual Reality (BIM-VR) in cost estimation, *Journal of Information Technology in Construction* 27: 914–925.

Leśniak, A. and Zima, K. (2018). Cost Calculation of Construction Projects Including Sustainability Factors Using the Case Based Reasoning (CBR) Method, *Sustainability* 10(5): 1608.

Lhee, S. C., Issa, R. R. and 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.

Liu, L. and Napier, Z. (2010). The accuracy of risk-based cost estimation for water infrastructure projects: preliminary evidence from Australian projects, *Construction Management and Economics* 28(1): 89–100.

Maher, M. L. J. and McGoey-Smith, A. D. (2006). Risk-based cost and schedule estimation for large transportation projects, In *2006 Annual European Transport Conference, Strasbourg, France*, Citeseer. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=2c00737232c3eadfd640a5feb2cb938fbd452ca9

Mahmoodzadeh, A., Nejati, H. R. and Mohammadi, M. (2022). Optimized machine learning modelling for predicting the construction cost and duration of tunnelling projects, *Automation in Construction* 139: 104305.

Matel, E., Vahdatikhaki, F., Hosseinyalamdary, S., Evers, T. and Voordijk, H. (2022). An artificial neural network approach for cost estimation of engineering services, *International Journal of Construction Management* 22(7): 1274–1287.

Mhady, A. A., Budayan, C. and Gurgun, A. P. (2024). Estimate-at-completion (EAC) prediction using Archimedes optimization with adaptive fuzzy and neural networks, *Automation in Construction* 166: 105653.

Mohamed, B. and Moselhi, O. (2022). Conceptual estimation of construction duration and cost of public highway projects, *Journal of Information Technology in Construction* 27: 595–618.

Moreno, P. D. C. (2020, May 12). Machine learning: Choosing the right estimator (scikit-learn algorithm cheat-sheet), *Medium*. Retrieved from https://caiomsouza.medium.com/machine-learning-choosing-the-right-estimator-scikit-learn-algorithm-cheat-sheet-c51500772ac2

Odeck, J. (2004). Cost overruns in road construction—what are their sizes and determinants?, *Transport Policy* 11(1): 43–53.

Ofori-Boadu, A. N. (2015). Exploring regression models for forecasting early cost estimates for high-rise buildings, *The Journal of Technology, Management, and Applied Engineering* 31(5).

Ökmen, Ö. and Öztaş, A. (2010). Construction cost analysis under uncertainty with correlated cost risk analysis model, *Construction Management and Economics* 28(2): 203–212.

Park, U., Kang, Y., Lee, H. and Yun, S. (2022). A Stacking Heterogeneous Ensemble Learning Method for the Prediction of Building Construction Project Costs, *Applied Sciences* 12(19): 9729.

Pishdad, P. and Onungwa, I. O. (2024). ANALYSIS OF 5D BIM FOR COST ESTIMATION, COST CONTROL, AND PAYMENTS, *Journal of Information Technology in Construction (ITcon)* 29(24): 525–548.

Poh, C. Q. X., Ubeynarayana, C. U. and Goh, Y. M. (2018). Safety leading indicators for construction sites: A machine learning approach, *Automation in Construction* 93: 375–386.

Rezaee Arjroody, A., Hosseini, S. A., Akhbari, M., Safa, E. and Asadpour, J. (2024). Accurate estimation of cost and time utilizing risk analysis and simulation (case study: road construction projects in Iran), *International Journal of Construction Management* 24(1): 19–30.

Sadeh, H., Mirarchi, C. and Pavan, A. (2021). Integrated Approach to Construction Risk Management: Cost Implications, *Journal of Construction Engineering and Management* 147(10): 04021113.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers, *IBM Journal of research and development* 3(3): 210–229.

Sánchez-Garrido, A. J., Navarro, I. J., García, J. and Yepes, V. (2023). A systematic literature review on modern methods of construction in building: An integrated approach using machine learning, *Journal of Building Engineering* 73: 106725.

Sanni-Anibire, M. O., Zin, R. M. and Olatunji, S. O. (2021). Machine learning - based framework for construction delay mitigation, *Journal of Information Technology in Construction* 26: 303–318.

Shane, J. S., Molenaar, K. R., Anderson, S. and Schexnayder, C. (2009). Construction Project Cost Escalation Factors, *Journal of Management in Engineering* 25(4): 221–229.

Shoar, S., Chileshe, N. and 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: 104102.

Stigler, S. M. (1981). Gauss and the invention of least squares, *the Annals of Statistics* : 465–474.

Thomas, N. and Thomas, A. (2016). Regression Modelling for Prediction of Construction Cost and Duration, *Applied Mechanics and Materials* 857: 195–199.

Vaardini, S., Karthiyayini, and Ezhilmathi. (2016). STUDY ON COST OVERRUNS IN CONSTRUCTION PROJECTS –A REVIEW.

Vakaj, E., Cheung, F., Cao, J., Tawil, A.-R. H. and Patlakas, P. (2023). An ontology-based cost estimation for offsite construction, *Journal of Information Technology in Construction (ITCon)* 28: 220–245.

Wu, S., Wood, G., Ginige, K. and Jong, S. W. (2014). A technical review of BIM based cost estimating in UK quantity surveying practice, standards and tools, *Journal of Information Technology in Construction* 19: 534–562.

Xie, Y., Ebad Sichani, M., Padgett, J. E. and DesRoches, R. (2020). The promise of implementing machine learning in earthquake engineering: A state-of-the-art review, *Earthquake Spectra* 36(4): 1769–1801.

Yildiz, A. E., Dikmen, I., Birgonul, M. T., Ercoskun, K. and Alten, S. (2014). A knowledge-based risk mapping tool for cost estimation of international construction projects, *Automation in Construction* 43: 144–155.

Zhou, X. and Flood, I. (2024). Optimization and evaluation of a neural network based policy for real-time control of construction factory processes, *Journal of Information Technology in Construction* 29: 84–98.