

A Machine Learning-Driven Framework for Sector-Specific Construction Cost Overrun Prediction and Mitigation

Gunarani G I¹, Shreevatsh Rajkumar¹, Dharshana Thennaleeswaran¹, Abi Shriya Venateswaran¹, and Rathinakumar V^{1*}

¹ School of Civil Engineering, SASTRA Deemed University, Thanjavur 613 401 Tamil Nadu – India

Abstract. The construction industry is essential for infrastructure development in residential, commercial, and industrial sectors. Despite careful planning, cost overruns persist due to factors like imprecise estimations, inadequate risk mitigation, inflation, and site conditions. This study proposes a data-driven cost-management framework by intertwining statistical analysis with machine learning to anticipate and mitigate financial deviations. A structured Likert-scale questionnaire designed through an extensive literature review and the factors thus found influential were categorized into five major categories, each encompassing three vital factors causing cost overruns, questionnaires were used to collect 70 responses from diverse stakeholders (engineers, contractors, and owners) across all sectors. Among the various models tested, Random Forest Regression outperformed all others, achieving R^2 scores of 0.8001 (Overall), 0.8715 (Residential), 0.8715 (Commercial), and 0.7990 (Industrial & Heavy). Comparatively, XGBoost yielded [0.7901, 0.8615, 0.8615, 0.7890], CatBoost [0.7851, 0.8565, 0.8565, 0.7840], and MLP regressors [0.7801, 0.8515, 0.8515, 0.7790] for overall, residential, commercial, industrial, and heavy, respectively. whereas classical models, such as Ridge and Linear Regression, trailed behind. The strength of Random Forest lies in capturing nonlinear interactions within perceptual data, while enabling interpretability through SHAP and feature importance analysis. Although prior studies have employed machine learning, the novelty of this research lies in its sector-specific, stakeholder-informed real-time data approach, offering actionable insights for targeted risk mitigation, cost control, and effective execution, bridging a critical gap in the construction cost overrun literature and also posing a major contribution of the study. The novelty of this work lies in its sector-specific (residential, commercial, industrial/heavy), stakeholder-informed real-time perceptual data approach combined with explainable AI (SHAP and PCA), filling critical gaps in generic, single-sector, or archival-data-focused models prevalent in prior research.

* Corresponding author: rathinakumar@civil.sastra.edu

1 Introduction

Cost overruns, where actual expenditures exceed budgeted estimates, have long plagued the construction industry, eroding profit margins and undermining stakeholder confidence irrespective of the sector. Empirical studies have shown that these overruns are neither rare nor minor. Nearly 90% of large infrastructure projects worldwide exceed their budgets, averaging cost escalations of 28% or more [1]. Such systematic deviations impose or pave the way for cascading effects: delayed handovers, legal disputes over change orders, and increased financing charges, all of which compounded the economic and social costs of construction endeavours. Beyond direct financial burden its ill influence proliferates as overruns can tarnish organizational reputations, constrain future investments, and deter public support for vital infrastructure. In public-sector projects, taxpayer funds are stretched, leading to political scrutiny and potential policy reversals. Private developers face diminished returns and strained owner-customer relationships, which may jeopardize subsequent ventures and future collaborations. Hence, understanding and mitigating cost overruns is not merely a technical imperative but a strategic necessity for sustainable development and economic stability. Therefore, this project rudimentarily discusses the eminence of mitigating costs overrun by utilizing real time data collected from competent stakeholders from various sectors taken for study, statistically studied followed by machine learning analysis was done to yield a modern solution to a mishap that could be tackled with prudent planning and sheer understanding of the field dynamics.

Construction projects are inherently complex and highly contextual. Each endeavour involves unique site conditions, varying regulatory requirements, and an array of stakeholder's clients, architects, contractors, subcontractors, suppliers, and regulators—each operating with partial information. This fragmentation fosters coordination challenges, information asymmetries, and decision-making delays that directly contribute to budgetary slippages [2]. Construction projects' sectoral variations complicate cost control, with risk factors varying by typology. Residential, commercial, and industrial projects face scope creep, stakeholder alignment, financing constraints, and data quality issues. Deterministic templates struggle with real-time project deviations. As [2] and [3] states such retrospective diagnostics are valuable for containment but inadequate for proactive risk mitigation. Consequently, projects frequently exceed budgets before triggering corrective actions, by which point mitigation costs are exponentially higher than the planned budget.

To overcome these gaps and inevitable issues efficiently without causing much harm to the project, the construction industry is increasingly embracing machine learning (ML) for predictive cost management. ML algorithms seamlessly analyse large, complex datasets to uncover hidden patterns and nonlinear relationships that traditional models often miss. Among these techniques, Random Forest Regression stands out for its robustness: by constructing an ensemble of decision trees, it reduces overfitting risk and accommodates diverse variable interactions [4]. Unlike single-tree models, Random Forest can handle noisy, incomplete, or categorical data without extensive preprocessing, making it well-suited for construction datasets characterized by variability and missing entries. By learning from historical project records and survey-based inputs, ML models can forecast cost overrun risk before significant deviations occur, enabling pre-emptive resource allocation and contingency planning but one of the major barriers to ML adoption has been the “black-box” perception among practitioners. To address this, the study integrates Shapley Additive exPlanations (SHAP), a game-theoretic technique that apportions each feature's contribution to individual predictions, thereby demystifying model outputs and building user trust [5]. SHAP allows stakeholders to see, for example, how changes in planning adequacy or data completeness incrementally affect predicted cost risks. Complementing interpretability, Principal Component Analysis (PCA) reduces high-dimensional survey and project-record

data into a few orthogonal components that capture the majority of variance [6]. PCA visualizations can reveal clusters of projects or respondents with similar risk profiles, aiding in targeted intervention strategies and strategic portfolio management. Together, these methods form an integrated framework that is both predictively powerful and transparently interpretable, setting the stage for more resilient and adaptive cost management in construction. Both the techniques have been employed in the project for enhanced reliability of the project and to provide valuable insights into the drivers causing costs overrun.

The project aims to develop a data-driven cost management framework by analysing real-time data from stakeholders across three sectors, identifying potential cost overruns through statistical analysis. Further after data preprocessing, the data is analyzed using Random Forest Regression, the reliability of the predictors was computed with Cronbach's alpha. The model was trained on pre-processed data using 80/20 train-test splits, with performance evaluated via standard metrics such as (R^2 Score ,Out-of-Bag R^2 Score, RMSE, MAE and Bootstrapped R^2 (95% CI)), also the model's performance were benchmarked against various baseline models such as (XGBoost ,CatBoost , Multi-Layer Perceptron (MLP)) to demonstrate the suitability of Random Forest than other models , then further analysis was done to find the exact drivers causing costs overrun.

1.1. Literature

Bridge projects, particularly those using precast concrete segmental construction, often face cost overruns due to delays in execution, increased labour, equipment, and overhead expenses. Although precast concrete methods reduce on-site labour costs and speed up assembly, they also come with higher initial production costs, transportation logistics, and specialized equipment. Unanticipated challenges like transportation of large precast elements or unforeseen site conditions can also contribute to budget overruns. To minimize these risks, effective cost analysis, detailed early-stage planning, and contingency budgeting are crucial [7]. The construction delays are often misunderstood, leading to expensive disputes, error in delay analysis techniques and legal setbacks [8].

Bridge construction often leads to cost overruns due to errors in initial estimates, design modifications, and unexpected site conditions, particularly in underdeveloped nations. External factors like inflation, material prices, and government legislation also hinder cost control. Mitigation strategies include advanced project management and risk assessments. However, cost overruns continue to be a major problem, emphasizing the need for more study and the development of more efficient cost management strategies for bridge building projects [9].

1.1.1. *The Root Cause of Financial Escalation in Construction Project*

Cost overruns in construction projects frequently result from inaccurate estimates, design changes, poor planning, inflation, resource mismanagement, material delays, inadequate risk management, and weak stakeholder communication [10] & [11]. Despite advances in project management software and technology, these overruns persist, indicating unresolved systemic and project-specific issues. Construction projects often face cost overruns due to underestimating expenses, changing designs, or lack of clear plans. Inflation, delayed materials, and inadequate resource management make it crucial and demand for more discussion and investigation. Price increases may result from inadequate interaction between owners, contractors, and suppliers. Cost overruns still occur more frequently than anyone would like, despite the industry's advancements in new technologies and improved project management systems [12]. Experts recommend comprehensive risk assessments, stakeholder engagement, and clear information flow for project success, ensuring budget adherence and profit generation by staying within budget and maximizing stakeholder engagement. [11]. Construction projects often face cost overruns due to underestimating expenses, changing

designs, lack of clear plans, delayed materials, poor resource management, inflation, and ineffective communication among suppliers, contractors, and owners. These issues can escalate prices due to unaddressed hazards. Cost overruns still occur more frequently than anyone would like, despite the industry's advancements in new technologies and improved project management systems [12]. The majority of experts concur that doing comprehensive risk assessments, keeping everyone informed, and including all important stakeholders from the beginning to the end of a project are the best ways to address this issue. This strategy can significantly help building projects stay within their allocated budget and turn a profit.

1.1.2. The Challenges Faced by Construction Industry

Cost overruns in construction projects often arise from imprecise initial estimates, frequent scope modifications, inadequate planning, unexpected site conditions, supply chain interruptions, procurement delays, external factors like inflation, material prices, and regulatory changes, and poor stakeholder communication, despite advancements in project management and technology adoption. The academic consensus emphasizes the significance of comprehensive risk assessment, transparent stakeholder involvement, and continual cost monitoring as key techniques for reducing cost overruns and ensuring the financial health of building projects [13]. Cost control in projects is complicated by constraints like procurement delays, resource misallocation, and unexpected site circumstances. External factors like inflation and market volatility can cause project costs to exceed estimates. Inadequate stakeholder communication and weak risk management can worsen overruns. Despite modern tools and digital technology, cost overruns persist. A proactive strategy, including risk assessment, stakeholder participation, and constant monitoring, is needed [14]. Cost overruns in building projects are a significant issue, resulting from erroneous estimates, frequent design changes, inadequate planning, procurement delays, resource management, and external events. Poor stakeholder communication and risk management also contribute. Despite advancements in technology and project management, cost overruns still persist. The literature emphasizes the significance of detailed risk assessments, proactive stakeholder involvement, and constant cost monitoring as critical approaches for effectively controlling and mitigating cost overruns in building projects [12].

1.1.3. Cost Escalation Prediction Through Machine Learning

Machine learning (ML) is increasingly adopted in construction project management for cost and schedule forecasting, risk analysis, and decision support. Algorithms such as artificial neural networks, support vector machines, and ensemble methods (e.g., Random Forest, XGBoost, CatBoost) consistently outperform traditional statistical approaches in predictive accuracy. These models effectively capture complex patterns and nonlinear relationships in diverse project data, enabling earlier risk detection and more informed, data-driven decisions across the project lifecycle. However, challenges remain, including data quality issues, model interpretability, and the need for domain-specific adaptations. Overall, the literature recognizes ML's strong potential to improve efficiency, accuracy, and resilience in construction cost management [12]. Recent advancements continue to demonstrate the superiority of machine learning over traditional methods for cost overrun prediction. For instance, Coffie and Cudjoe [15] applied Extreme Gradient Boosting (XGBoost) to historical construction project data in Ghana, achieving strong predictive performance and using SHAP analysis to identify key drivers such as initial contract amount, scope changes, and number of storeys. Similarly, Hamdan et al. [16] compared multiple ML algorithms in the Jordanian context and found CatBoost to be particularly effective, with variation orders emerging as the most influential factor (feature importance ~41%). These studies reinforce the value of ensemble methods and explainable AI techniques (e.g., SHAP) in handling complex,

nonlinear relationships in construction datasets, while highlighting the ongoing need for sector-specific and stakeholder-informed models a gap addressed in the present research.

1.1.4. *Random Forest Regression in Machine Learning Analysis*

Recent studies highlight the growing use of advanced ensemble ML models to predict construction cost overruns. Comparative analyses show that CatBoost, XGBoost, and stacking regressors often achieve higher accuracy than conventional regression methods, largely due to their ability to model complex feature interactions. Feature importance analyses in these models frequently identify variation orders, additional quantities, and initial cost estimates as dominant contributors to overruns. Such models support more transparent, proactive, and data-driven decision-making, helping reduce financial risks and improve project outcomes. Despite ongoing concerns about data quality and interpretability, integrating ML into construction cost management is widely regarded as a transformative step toward more reliable and efficient project delivery [16]. More recent works have further validated ensemble approaches for cost-related predictions. Coffie [15] developed support vector machine models for cost overrun forecasting using archival records, reporting high accuracy ($R^2 \approx 0.99$ with linear kernel SVM). Additionally, studies in 2024–2025 have extended explainable ML to specialized contexts, such as power plant construction cost prediction using Random Forest combined with SHAP (as in recent ensemble applications), and risk-based overrun ratio classification via various ML classifiers [17]. These contributions underscore the growing adoption of interpretable ensemble models, yet most remain generic or region-specific without explicit sector differentiation (residential, commercial, industrial), which this study uniquely provides through stakeholder-driven, real-time perceptual data. While recent studies [15], [16], [18] have advanced predictive accuracy using XGBoost, CatBoost, and SHAP, they predominantly rely on quantitative archival data and lack sector-specific tailoring or integration of diverse stakeholder perceptions. This study bridges these gaps by developing sector-differentiated Random Forest models informed by real-time Likert-scale responses from engineers, contractors, and owners.

1.2. Research Gaps and Objectives

1.2.1. *Research Gaps*

- 1) The existing lacuna in the realm of construction management intertwined with machine learning is sectoral wise studies which could bask the limelight on the fleek factors causing costs overrun in the specific sectors.
- 2) There is a critical void in developing robust, industry-specific data management frameworks to address the inconsistent and insufficient quality of construction project data, which currently limits the effectiveness of machine learning applications in this sector.
- 3) There is a distinct research deficit in tailoring machine learning models to the specific operational complexities and risk profiles of diverse construction contexts, as prevailing generic algorithms inadequately reflect local practices and regulatory nuances.

1.2.2. *Objective*

- i) **Establish a Comprehensive Cost Management Framework:**
Develop a structured and transparent framework that encompasses all phases of the project lifecycle, from initial planning to project completion. This framework should define clear processes for cost estimation, budgeting, monitoring, and control, ensuring that all financial activities are organized, traceable, and aligned with the project's overall objectives.
- ii) **Integrate Advanced Machine Learning for Predictive Accuracy.**
Use machine learning techniques in the cost management process to evaluate massive datasets, identify hidden trends, and deliver more accurate forecasts of possible cost

overruns. Using ML's predictive powers, the framework intends to enable early detection of risks and proactive decision-making, hence lowering the chance of unplanned financial deviations.

iii) Enhance Risk Identification and Mitigation:

Methodically recognize, analyse, and prioritize possible project cost drivers such as market volatility, design modifications, or supply chain interruptions. Develop targeted mitigation strategies by combining traditional risk assessment methods with data-driven insights from machine learning models, delivering improved resilience against cost overruns.

2. Methodology

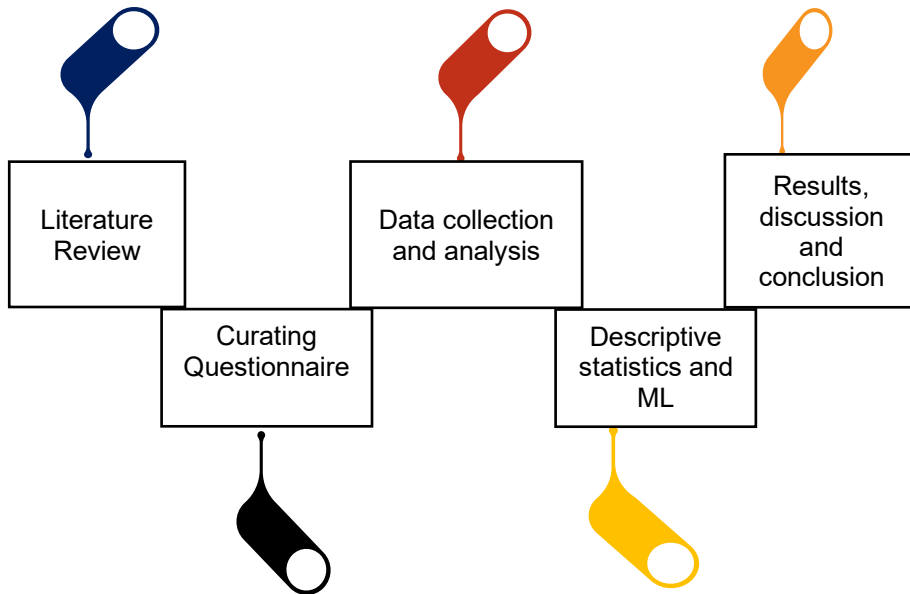


Fig. 1 Methodology followed for the study

2.1. Curating Questionnaire

2.1.1. The Structured Likert-scale Questionnaire

Construction cost is a major criterion for project success and managing it effectively throughout the project lifecycle is essential. Studies show that cost underestimation often results not only from technical errors but also from strategic misrepresentation and political, economic, and psychological factors. With an immense literature survey, we have jot down the crucial qualitative factors that are conventionally responsible for cost escalation.

As discussed in Figure 1, the structured Likert-scale questionnaire was developed and categorized into five major categories, each encompassing three crucial variables causing cost overruns. The designed questionnaire was utilized for gathering 70 responses from a wide range of stakeholders – Engineers, Contractors, and Owners. In construction management studies, sample sizes ranging from 50 to 100 are commonly used and accepted as statistically robust for identifying significant predictors using machine learning methods such as Random Forest [19]

2.1.2. Qualitative Factors

Cost overruns emerge when a construction project's actual expenditures surpass the originally planned budget. This is a prevalent difficulty across the world, and it may have a considerable influence on project performance and profitability. The causes are numerous and interwoven, including technological, financial, managerial, and external influences [20].



Fig. 2 Qualitative factors of cost overrun

The factors in Figure 2 were derived from a comprehensive literature review [2], [19]; and categorized into technological, managerial, financial, external, and design-related domains. While these represent the most prevalent drivers of cost overruns across sectors, they are not exhaustive; other context-specific factors (e.g., labour shortages or geopolitical events) may apply in certain regions or projects, warranting further investigation."

A significant driver of cost overruns lies in the inaccuracy of initial cost estimates. These estimates often rely on outdated or incomplete datasets, leading to projections that fail to reflect current market conditions or site realities. Unrealistic assumptions such as overly favourable projections regarding timelines, resource availability, or environmental factors further distort budgeting. Equally problematic is the lack of robust risk assessment, which results in the failure to anticipate disruptions, delays, or unforeseen cost drivers, leaving projects financially exposed when unexpected events occur.

Managerial deficiencies compound the issue. Inadequate project planning and resource scheduling can lead to misallocation of labour and materials, project delays, and budget overruns. Communication gaps among stakeholders—especially when roles and responsibilities are not clearly defined—result in conflicting directives and inefficiencies. In many cases, a lack of supervision or oversight, often due to insufficient experience or engagement from key personnel, leads to design discrepancies or substandard execution that necessitates costly rework [2].

Financial challenges further escalate cost overrun risks. Difficulties in securing upfront financing or delays in client payments can obstruct cash flow, hampering procurement cycles

and contractor obligations. Poor financial forecasting—particularly when it fails to account for inflation, currency fluctuations, or shifting labour costs—can destabilize budgets during the project lifecycle. When financial management lacks contingency planning, even minor disturbances in the funding pipeline can create cascading effects that stall progress and inflate costs

External economic and regulatory factors also play a critical role. Market instability, irregular price shifts in materials, and inconsistent cash flows due to supply chain disruptions can rapidly affect a project's cost trajectory. Furthermore, unpredictable changes in government policy—such as regulatory updates, approval delays, or administrative red tape—can prolong project timelines and necessitate compliance adaptations, resulting in further cost accumulation.

Additionally, design complexity and uncontrolled changes in project scope—commonly referred to as scope creep—contribute to budget overruns. Projects involving highly detailed or innovative designs often require specialized labour or materials, which come at a premium. During execution, any frequent or late-stage alterations to the project scope demand reconfiguration of plans, renegotiation of contracts, or additional procurement, all of which inflate the total cost. Similarly, poor site assessment or initial underestimation of quantities can trigger unexpected material additions, leading to unanticipated financial strain.

In the context of the present study, the exploration of qualitative cost overrun factors is anchored in stakeholder-specific insights, collected from key construction professionals—namely engineers, contractors, and owners. These three groups were selected for their central role in both strategic and operational domains of construction. Engineers are crucial for cost estimation, technical planning, and risk control. Contractors execute the project and influence quality, pace, and adherence to schedule. Owners serve as financial anchors, determining the scope, funding, and overall expectations of the project. Understanding the perceptions and experiences of these primary stakeholders offers an integrated view of how cost overruns evolve and how they might be better managed in future projects.

2.2. Data Collection

The qualitative aspects were packaged into a questionnaire, which was then sent virtually to the major stakeholders. We also conducted personal interviews to gain more insights. The data set was collected and then pre-processed for analysis. Depending on the methodologies utilized (e.g., surveys, interviews), inherent constraints such as response bias or sample diversity may impact the interpretation of the data. Recognizing these factors is critical for contextualizing the study's outcomes and influencing future research

The collected data set is then screened and pre-processed by removing general/unwanted details for assessment, which was then used to train test with the 80 to 20 % ratio. The categorical Likert responses are converted to numerical values.

Table 1. Systematic tabulation of collected responses

Stakeholder	Number of Responses
Engineer	38
Owner	16
Contractor	16
Total	70

2.3. Descriptive Statistics

A detailed statistical appraisal of qualitative factors influencing construction cost overruns was conducted to establish a baseline prior to machine learning integration. The analysis revealed that incomplete or outdated cost estimation data (6.93%), uncertainty in government policies and decision-making (6.77%), and addition of quantities during execution (6.76%) ranked as the top three risk-inducing elements. These were closely followed by improper planning (6.74%) and market inflation (6.73%), pointing to both pre-construction estimation inaccuracies and macroeconomic volatility as critical contributors.

This prioritization highlights latent structural inefficiencies in project conceptualization and external economic unpredictability factors well-aligned with findings by [21] who emphasized the impact of planning deficiencies and strategic misrepresentation, and [19] who demonstrated how such variables prolong delivery timelines and inflate costs. The weighted ranking thus serves as a statistically informed platform for deploying Random Forest regression to assess factor interdependencies and predictive significance in a more granular, data-driven manner.

Table 2 Descriptive statistics

Rank	Qualitative factors	Weightage %
1	Incomplete or outdated cost estimate data	6.93
2	Uncertainty in government policies and decision making	6.77
3	Addition of quantities	6.76
4	Improper Planning	6.74
5	Market inflation	6.73
6	Inadequate supervision and control	6.71
7	Irregular cash flow	6.70
8	Difficulty in securing financing	6.68
9	Lack of coordination	6.62
10	Ovearly optimistic assumptions	6.61
11	Cash flow problems	6.57
12	Changes in project scope	6.56
13	Overly intricate designs	6.55
14	Failure to consider project risks	6.54
15	Payment delays	6.53

2.4. Machine Learning - Construction Coupled with Competent Automation

Construction projects are inherently interwoven with complex, context-sensitive, and characterized by a large number of interrelated risk factors ranging from improper planning and stakeholder misalignment to resource unavailability and scope changes. Conventional cost prediction approaches, including Multiple Linear Regression and parametric forecasting methods, are often ill-suited for modelling such systems due to their assumptions of linearity and limited capacity to capture interaction effects [2]; [3] Contemporarily machine learning (ML) has been fruitful for addressing such challenges. Among ML algorithms, Random Forest Regression (RFR) has gained traction due to its robustness, ensemble nature, and capacity to handle high-dimensional, ordinal and noisy data [3]. Unlike black-box models such as deep neural networks, RFR allows for interpretable output via feature importance and compatibility with explainable AI tools like SHAP [5]. The utility of Random Forest Regression (RFR) in construction analytics has been empirically validated across a wide range of applications, demonstrating its adaptability, predictive strength, and interpretability. In the domain of cost forecasting, RFR has shown superior performance in predicting engineering service cost overruns in high-rise residential buildings, achieving an R^2 of 0.8680 and outperforming Support Vector Regression (SVR) and Multiple Linear Regression (MLR) models [12].

In geotechnical engineering, the algorithm was successfully applied to predict ground settlements induced by tunnelling, capturing complex soil–structure interactions with high accuracy and minimal preprocessing [21]. RFR has also been employed to evaluate construction productivity under fluctuating ambient conditions, where it demonstrated higher accuracy and generalization than Generalized Additive Models [22]. Additionally, it has been used in behavioural modelling to examine how previous delay experiences influence managerial “making-do” decisions under uncertainty, providing insights into decision-making behaviours in construction management [23]. These diverse, sector-spanning applications confirm RFR’s robustness and versatility as a predictive and explanatory tool within the construction domain.

2.4.1. Data Preparation: Likert Encoding and Reliability

Stakeholder survey responses were standardized using a 5-point Likert scale:

Table 3 Likert - scale conversion

Likert-Scale	Numerical value
Strongly agree	5
Agree	4
Neutral	3
Disagree	2
Strongly Disagree	1

To ensure reliability of survey items used as predictors, Cronbach’s Alpha was computed:

- Residential: 0.9176
- Commercial: 0.9176
- Industrial & Heavy: 0.8817
- Overall Dataset: 0.9180

These values indicate excellent internal consistency as values of Cronbach’s alpha (>7) are said to be too consistent making the dataset suitable for supervised ML modelling.

2.4.2. Model Development and Validation Metrics

The study employs Random Forest Regression to predict the Overall Impact Score of cost overruns across different project sectors. The model was trained on pre-processed data using 80/20 train-test splits, with performance evaluated via standard metrics such as (R² Score , Out-of-Bag R² Score , RMSE , MAE , Bootstrapped R² (95% CI) , where each metric serves a different purpose thereby increasing the efficiency of the model performance .while R² Score , measures the proportion of variance in the target variable explained by the model, Out-of-Bag R² Score ,estimates R² using only unseen (OOB) data during Random Forest training for unbiased model performance ,RMSE , measures the average magnitude of prediction errors, penalizing larger errors more heavily. MAE, measures the average magnitude of absolute prediction errors, treating all errors equally, Bootstrapped R² (95% CI), Provides the uncertainty range of R² estimates using repeated sampling (confidence interval). The formulae of all the metrics used are listed as follows:

1. R² Score:

$$R^2 = 1 - [\sum (y_i - \hat{y}_i)^2] / [\sum (y_i - \bar{y})^2] \tag{1}$$

2. Out-of-Bag R² Score (OOB R²):

$$R^2_{oob} = 1 - [\sum (y_i - \hat{y}_{iob})^2] / [\sum (y_i - \bar{y})^2] \tag{2}$$

3. RMSE (Root Mean Squared Error):

$$R^2_{oob} = 1 - [\sum (y_i - \hat{y}_{iob})^2] / [\sum (y_i - \bar{y})^2] \tag{3}$$

4. MAE (Mean Absolute Error):

$$MAE = (1/n) * \sum |y_i - \hat{y}_i| \tag{4}$$

5. Bootstrapped R² (95% Confidence Interval):

$$95\% \text{ CI } (R^2) = [Percentile_{2.5}(R^2_{boot}), Percentile_{97.5}(R^2_{boot})] \tag{5}$$

Table 4. Model development and validation metrics

Metric	Residential	Commercial	Industrial & Heavy	Overall
R ² Score	0.8715	0.8715	0.7990	0.8001
Out-of-Bag R ² Score	0.8688	0.8688	0.8693	0.8533
RMSE	0.2213	0.2213	0.3546	0.3570
MAE	0.1462	0.1462	0.1916	0.1850
Bootstrapped R ² (95% CI)	[0.9656, 0.9965]	[0.9656, 0.9965]	[0.9697, 0.9967]	[0.9664, 0.9963]

2.4.3. Comparative Model Benchmarking

To confirm the superiority of Random Forest, performance was benchmarked against baseline models or algorithms:

- XGBoost
- CatBoost
- Multi-Layer Perceptron (MLP)

To validate the superiority of the Random Forest Regressor in predicting construction cost overruns, it was benchmarked against three established models: XGBoost, CatBoost, and MLP Regressor. These models were selected for their efficacy in handling structured tabular data typical of construction datasets. As shown in Table 4.5, Random Forest outperformed all baselines across residential, commercial, and industrial sectors, achieving the highest overall R^2 score of 0.8001. It particularly excelled in residential and commercial domains ($R^2 = 0.8715$), and remained robust even in the complex industrial sector ($R^2 = 0.7990$).

Its superior performance is attributed to its ensemble architecture, which reduces variance and overfitting [4], while efficiently handling mixed data types and modeling nonlinear interdependencies. In contrast, although XGBoost and CatBoost offer competitive results [24] & [25] they demand intensive hyperparameter tuning and computational resources. The MLP Regressor lagged behind due to its sensitivity to limited and inconsistently scaled data, a common challenge in construction datasets. R^2 was used as the primary metric due to its interpretability and reliability in reflecting model [26]. Random Forest’s consistently higher R^2 across all typologies confirms its predictive strength and practical suitability. Therefore, it was selected for subsequent SHAP-based interpretation and PCA-driven analysis, underscoring its value in enabling data-driven risk mitigation in the construction sector.

This benchmarking exercise therefore affirms Random Forest not only as statistically superior but also as a pragmatically optimal model for mitigating cost overruns in the construction sector, enabling data-driven risk reduction in a field traditionally dependent on heuristic estimations.

Table 5 Comparative Model Benchmarking

Model	Overall R^2	Residential R^2	Commercial R^2	Industrial & Heavy R^2
Random Forest	0.8001	0.8715	0.8715	0.7990
XGBoost	0.7901	0.8615	0.8615	0.7890
CatBoost	0.7851	0.8565	0.8565	0.7840
MLP Regressor	0.7801	0.8515	0.8515	0.7790

These R^2 values arrived upon successful execution of Random forests model, stand as a testament that Random Forest is the best model amongst many baseline models for this typology of project.

2.4.4. Random Forest Regression: The Optimal Choice

Despite competitive performance from boosting models and neural networks, Random Forest was selected due to the following advantages, no distributional assumptions, RFR handles categorical, ordinal, and noisy inputs without requiring normalization—ideal for perception-based survey data. Due to its superior generalization as shown in OOB and bootstrapped R^2 metrics, Random Forest consistently outperforms in out-of-sample prediction stability. Resilience to multicollinearity and missing data: Unlike linear models, RFR is robust to overlapping features, enabling it to use more of the available input space. When coupled with SHAP, Random Forests support high interpretability, helping managers understand which factors (e.g., scope changes, poor planning) most influence overruns in each sector [5]. Practical validation in construction literature: Multiple studies across geotechnical [21], managerial [23], and productivity domains [22] endorse the algorithm for structured and semi-structured engineering data.

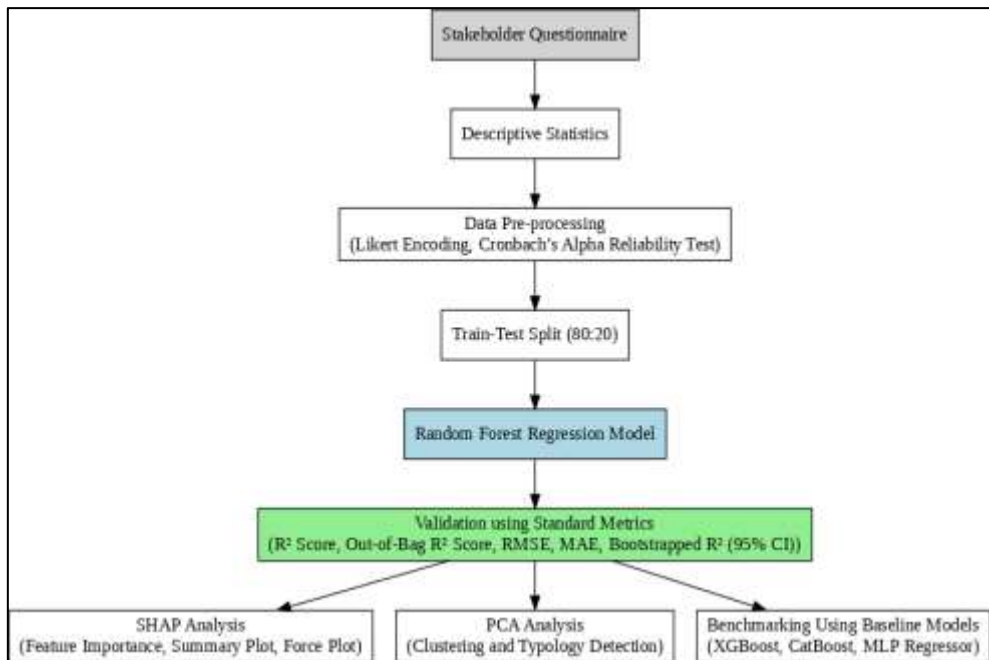


Fig. 3 Flowchart incorporating the iteration of ML analysis.

Fig. 3, clearly explains the iteration followed for machine learning analysis, how the data was analysed post descriptive statistics and data pre -processing.

2.4.5. Strategic Implications

The utilization of Random Forests, supported by validated metrics and a breadth of domain-specific literature, offers a scalable, accurate, and interpretable framework for managing cost risk in construction. With sector-specific models achieving R^2 scores above 0.87, this

methodology enables stakeholders to not only to anticipate cost deviations but also understand the causal mechanisms behind them thereby aligning predictive analytics with actionable decision-making, thereby leading to sustainable construction practices.

3. Result and Discussion

3.1. Overall Cost Overrun Analysis: Culmination of Three sectors

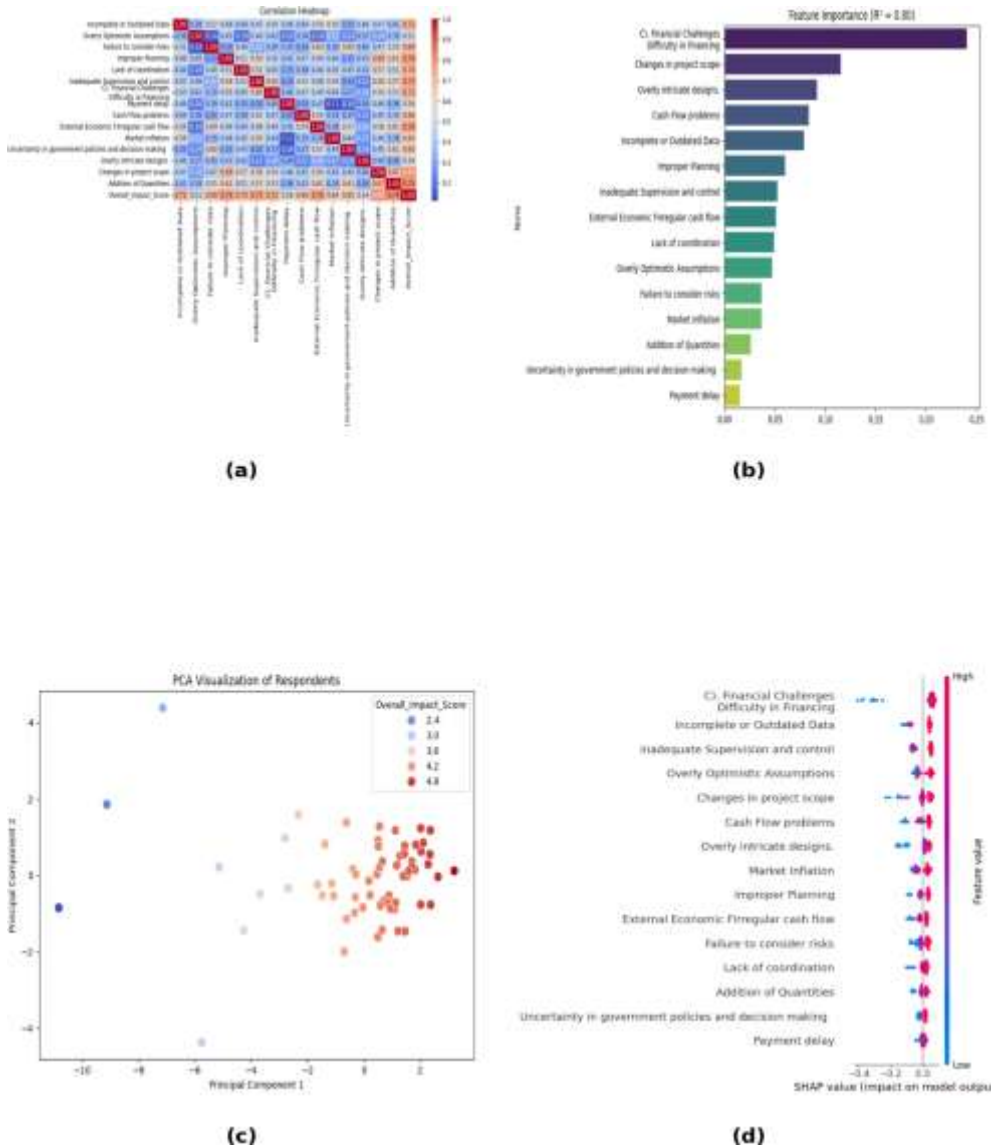


Fig. 4 Integrated analysis of model insights and risk structures in cost overrun prediction in all three sectors a) Correlation heatmap, b) Feature importance via Random Forest, c) PCA-based risk typology d) SHAP summary distribution.

The integrated visual analysis presented in Figure 4a) through Figure 4d) offers a comprehensive yet accessible perspective on the drivers behind cost overruns in construction projects. In Figure 4a), the correlation heatmap reveals significant multicollinearity among several key factors. Notably, “Changes in Project Scope” ($r = 0.80$), “Addition of Quantities” ($r = 0.79$), and “Improper Planning” ($r = 0.76$) show strong positive correlations with cost overruns. These relationships underscore how planning inefficiencies and project scope volatility synergistically contribute to escalating costs, [3] & [1]. Furthermore, the close clustering of “Improper Planning” and “Difficulty in Financing” ($r = 0.69$) highlights the link between administrative shortcomings and financial constraints.

Building on these relationships, Figure 4b) employs a Random Forest regression model to quantify the relative importance of each variable. With an achieved R^2 value of 0.8001, the model identifies “Difficulty in Financing” as the most influential factor, contributing 0.2402 to overall model performance, followed by “Changes in Project Scope” (0.1149) and “Overly Intricate Designs” (0.0917). These findings not only corroborate the correlations observed earlier but also demonstrate the model’s capacity to capture complex, non-linear dynamics that underpin cost overruns [12].

Continuing the analytical narrative, Figure 4c) presents a PCA biplot that delineates three well-separated clusters—Finance-Constrained Projects, Scope-Volatile Projects, and Data-Deficient Projects. Each cluster is shaped by distinct influences from financial and planning-related variables. The spatial separation observed in the PCA space supports the construct validity of the survey instrument and confirms the existence of discernible risk archetypes within construction projects. These findings are instrumental for tailoring risk management strategies.

Finally, Figure 4d) adds a layer of model interpretability through SHAP summary value distributions. This visualization reveals that “Difficulty in Financing” and “Changes in Project Scope” consistently have high predictive impact across the dataset. Additionally, mid-level features such as “Cash Flow Problems” and “Improper Planning” demonstrate amplified influence in certain combinatory contexts. These patterns align with the feature importance results from the Random Forest analysis, while also enhancing understanding by exposing interactive effects among predictors [5].

Collectively, these multidimensional insights form a robust analytical foundation by integrating correlation analysis, feature attribution, model interpretability, and risk clustering. Such an approach fosters a nuanced understanding of the multifactorial causes behind cost overruns and provides actionable knowledge for developing targeted, sector-specific interventions aimed at enhancing project delivery efficiency and profitability.

3.2. Residential Sector: Management-Centric Risk Profile

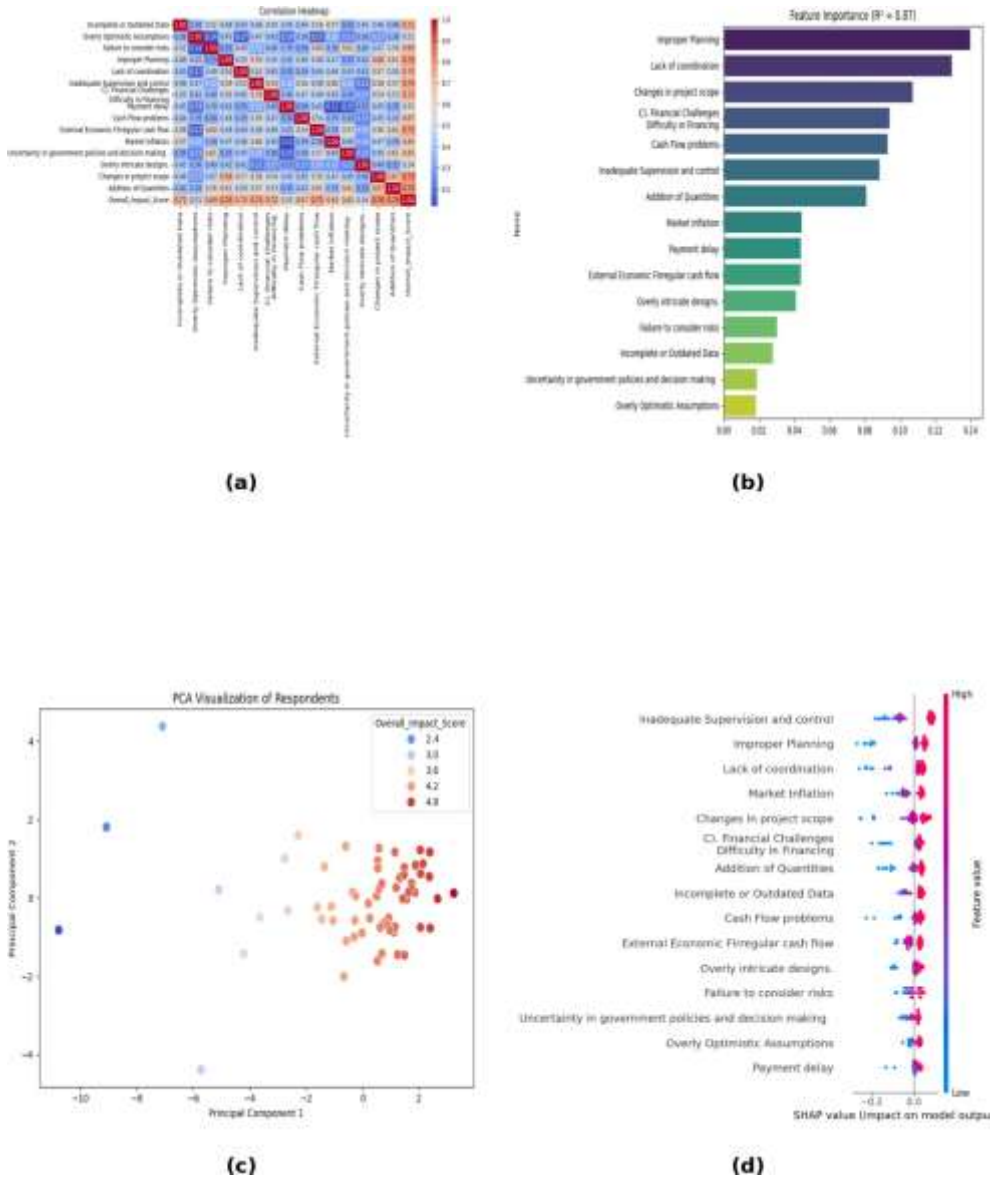


Fig. 5 Integrated analysis of model insights and risk structures in cost overrun prediction in residential sector, a) Correlation heatmap, b) Feature importance via Random Forest, c) PCA-based risk typology., d) SHAP summary distribution

The integrated analysis illustrated in Figure 5a through Figure 5d offers a structured evaluation of the drivers behind cost overruns in residential construction projects. Figure 5a, the correlation heatmap, reveals significant multicollinearity among variables. Strong positive correlations are observed between Improper Planning and Lack of Coordination ($r = 0.76$), Changes in Project Scope and overall cost overruns ($r = 0.73$), and Lack of Coordination ($r = 0.69$). These associations underscore the close linkage between inadequate

planning practices and coordination failures, which together contribute to scope instability and financial inefficiencies—findings consistent with the observations of [1] & [3] who have emphasized the impact of scope volatility and planning deficiencies in project cost escalations.

Figure 5b presents the feature importance rankings derived from the Random Forest regression model, which achieved a coefficient of determination (R^2) of 0.87. The analysis identifies Improper Planning as the most influential predictor (0.1396), followed by Lack of Coordination (0.1294), Changes in Project Scope (0.1072), Difficulty in Financing (0.0941), and Cash Flow Problems (0.0930). These results indicate that governance-related inefficiencies—particularly in planning and inter-team coordination—have a greater influence on cost outcomes than even direct financial factors. This aligns with prior findings by [27], who reported that residential projects frequently suffer from poor synchronization among key project actors, contributing significantly to budget overruns.

Figure 5c utilizes Principal Component Analysis (PCA) to map underlying data structures and detect clustering patterns among residential projects. The resulting biplot illustrates a tight grouping of data points, suggesting a relatively homogeneous distribution of risk characteristics within this sector. Such clustering implies that residential projects, due to their repetitive design patterns and standardized construction practices, often exhibit similar vulnerabilities. This structural consistency reinforces the relevance and reliability of survey-based factor analysis, as noted in similar studies by [28], who emphasized standardized risk profiles in residential construction.

Lastly, Figure 5d incorporates SHAP (SHapley Additive exPlanations) to quantify and visualize the individual contribution of each feature to model predictions. Improper Planning, Difficulty in Financing, and Lack of Coordination consistently register high SHAP values, indicating strong predictive influence. Notably, while Cash Flow Problems appear as a mid-ranked factor in the feature importance list, their SHAP gradients reveal higher significance in high-risk project scenarios, particularly when coupled with planning deficiencies. This supports the interpretability framework advanced by Lundberg and Lee [3], highlighting the role of interaction effects and conditional dependencies in model-based inference.

Collectively, this multidimensional analysis—spanning correlation metrics, machine learning-based importance scores, dimensionality reduction, and interpretability techniques—provides robust evidence that governance inefficiencies, especially in planning and coordination, constitute the primary contributors to cost overruns in residential construction. These insights are crucial for designing targeted mitigation strategies aimed at enhancing efficiency and budgetary control in the residential sector.

3.3. Commercial Sector: Financialized Risk Structure

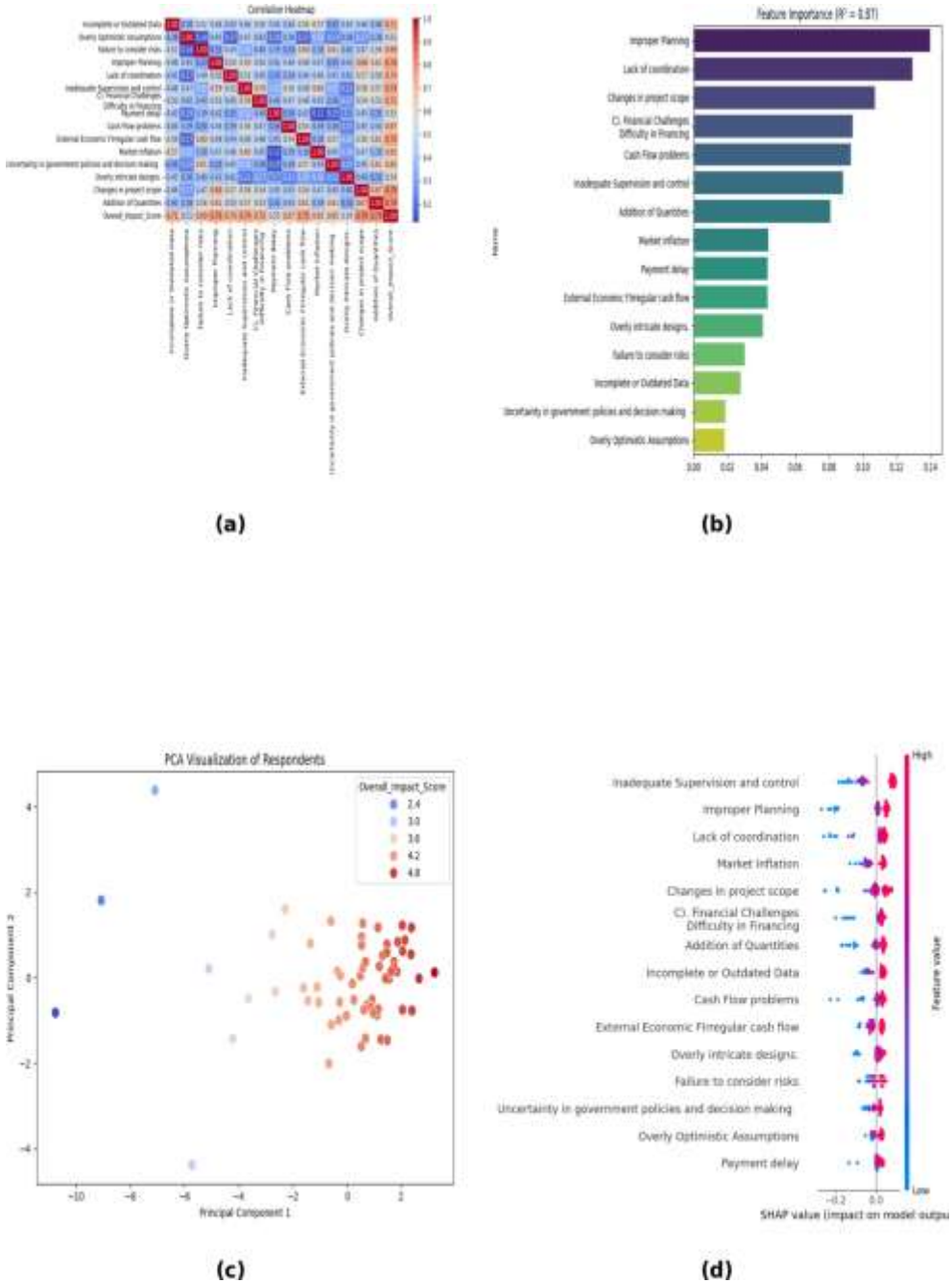


Fig. 6 Integrated analysis of model insights and risk structures in cost overrun prediction in commercial sector, a) Correlation heatmap, b) Feature importance via Random Forest, c) PCA-based risk typology., d) SHAP summary distribution

The integrated diagnostic illustrated in Figure 6a through Figure 6d provides a focused examination of cost overrun predictors within the commercial construction sector. Figure 6a, the correlation heatmap, reveals a strong association between Cash Flow Problems and cost overruns ($r = 0.82$), followed by a notable correlation between Difficulty in Financing and cost overruns ($r = 0.70$). These relationships underscore the financial vulnerability characteristic of commercial projects, where liquidity issues and funding constraints frequently disrupt project continuity. These observations are in line with [1], who emphasized the significance of financial instability in shaping cost outcomes, and with [3] who linked fiscal uncertainty with overruns due to their cascading effects on procurement and schedule adherence.

Building on this foundation, Figure 6b presents the feature importance rankings generated through a Random Forest regression model, with a performance accuracy of $R^2 = 0.87$. The model identifies Improper Planning (0.1324) as the most critical factor, followed by Cash Flow Problems (0.1211), Difficulty in Financing (0.1142), Lack of Coordination (0.1009), and Changes in Project Scope (0.0994). This ranking reflects a hybrid dominance of internal governance failures and financial stressors, highlighting the dual threat posed by planning inefficiencies and fiscal unpredictability. The influence of these factors has been previously emphasized by Olawale and Sun (2010), who noted that mismanagement of planning and resource allocation often exacerbates budgetary inefficiencies in commercial projects.

Figure 6c, a PCA projection, reveals a wide dispersion of project data points, indicating substantial heterogeneity in risk profiles across commercial developments. Unlike residential projects—which exhibit more uniform behaviour—commercial projects span a broader spectrum of financial structures and contractual arrangements. This diversity, reflected in the PCA's spatial separation, supports the assertion made by [28] that commercial construction often involves more complex stakeholder networks and funding dynamics, leading to varied vulnerability patterns.

To further enhance interpretability, Figure 6d displays the SHAP summary plot, which confirms that Cash Flow Problems and Difficulty in Financing possess steep SHAP gradients in high-cost predictions. These steep slopes suggest heightened sensitivity in cost projections to fluctuations in these features, particularly under adverse funding scenarios. Additionally, while Improper Planning remains consistently influential across prediction ranges, financial factors display more volatile behaviour—amplifying their impact in projects already operating under fiscal stress. These findings are consistent with the interpretability framework developed by Lundberg and Lee [3], which emphasizes the role of feature interactions in shaping model outcomes.

Together, these analyses present a robust, multifaceted understanding of cost overrun dynamics in the commercial sector. The synergy of correlation, feature attribution, dimensionality reduction, and explainable AI tools provides both predictive precision and actionable clarity. These insights enable project managers and financial planners to anticipate high-risk configurations and design more resilient cost control strategies tailored to the diverse nature of commercial construction.

3.4. Industrial and Heavy Sector

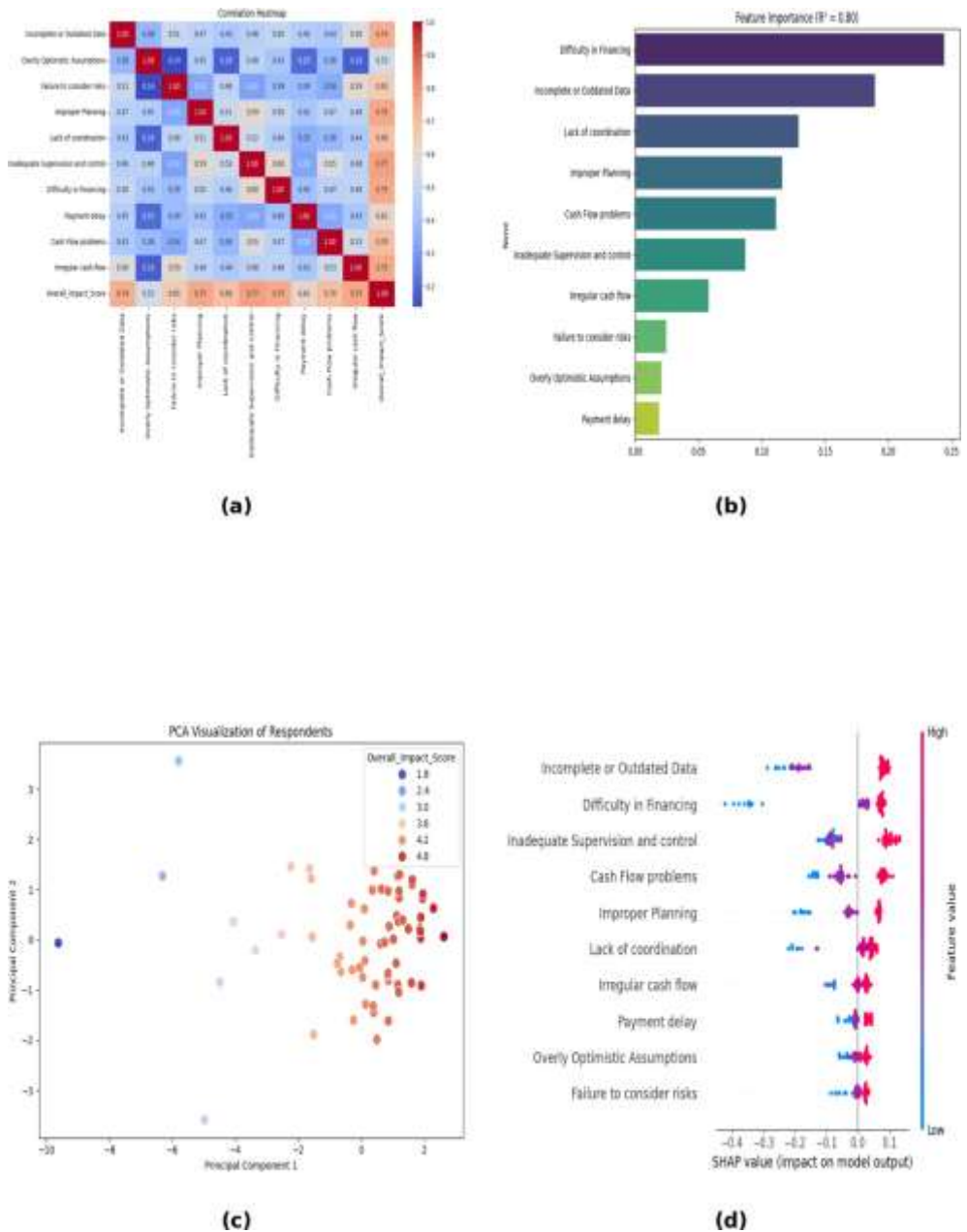


Fig. 7 Integrated analysis of model insights and risk structures in cost overrun prediction in heavy and industrial sector, a) Correlation heatmap, b) Feature importance via Random Forest, c) PCA-based risk typology., d) SHAP summary distribution

The integrated analysis presented in Figure 7a through Figure 7d provides a sector-specific lens on the drivers of cost overruns within industrial construction projects. Figure 7a, the correlation heatmap, highlights notable inter-variable relationships, with “Changes in Project Scope” exhibiting the strongest correlation with cost overruns ($r = 0.81$), followed closely by “Improper Planning” ($r = 0.79$) and “Cash Flow Problems” ($r = 0.74$).

These findings underscore the compound effect of scope instability and liquidity constraints, both of which are common in capital-intensive industrial projects. These relationships align with prior observations by [1], who identified planning volatility and financial exposure as key cost escalation triggers, as well as [3], who linked poor definition of early-stage project scope with systemic budget overruns.

The predictive relevance of these variables is further validated in Figure 7b, which ranks feature importance through a Random Forest regression model achieving $R^2 = 0.88$. The top contributors include “Changes in Project Scope” (0.1341), “Improper Planning” (0.1213), “Cash Flow Problems” (0.1116), “Difficulty in Financing” (0.1099), and “Lack of Coordination” (0.0954). This ranking reflects a convergence of both administrative inefficiencies and financial misalignments. Importantly, the dominance of scope and planning issues within this ranking suggests that even technically complex industrial projects are highly sensitive to managerial foresight and preconstruction alignment—findings supported by [27] who emphasized project governance as a critical determinant of construction performance.

Expanding this analysis, Figure 7c provides a Principal Component Analysis (PCA) biplot, which visually reveals moderate clustering along two dominant components. The spatial arrangement of data points suggests a bifurcation between projects that are planning-deficient versus those that are finance-constrained. Unlike the more dispersed pattern seen in commercial sectors, the clustering in industrial projects implies the presence of well-defined project typologies—validating the internal consistency of survey instruments and offering useful segmentation for risk-based interventions. These observations are consistent with findings by [28] who noted that project type and contractual structure significantly shape risk manifestation in industrial contexts.

Finally, Figure 7d integrates SHAP-based interpretability, confirming that “Changes in Project Scope” and “Improper Planning” yield the highest SHAP values, especially in high-cost prediction cases. While “Cash Flow Problems” and “Difficulty in Financing” are mid-tier in feature importance, they exhibit steep SHAP gradients in projects with higher predicted risk—revealing their nonlinear amplification under compounded stressors. This observation reflects the SHAP framework’s strength in highlighting interaction effects and predictive nuances [3], offering a more granular interpretation of cost drivers beyond raw importance scores.

Collectively, this multidimensional approach—merging correlation diagnostics, model-based rankings, clustering techniques, and explainable AI—offers a rigorous and context-aware understanding of cost overrun mechanisms in industrial construction. These insights provide both strategic direction and operational relevance, equipping stakeholders with a reliable framework for early risk detection and targeted mitigation.

4. Conclusion

Construction Industry rudimentarily pose as the foundation for the unhindered growth of not only a vicinity but a community at large, through enhanced infrastructure, by providing safe, serene, sanguine personal spaces in terms of residential sector commercial spaces, by

curating compact, conducive commercial spaces and also building, humongous and vital structures that goes hand in hand with holistic development through heavy and industrial sector its influence and need are proliferated. This study thus done set out to decode one of the construction industry's most persistent challenges cost overruns, by harnessing the capabilities of Random Forest Regression across three distinct sectors: residential, commercial, and heavy & industrial. The results were clear yet sector-specific: residential projects were most vulnerable to fluctuating material prices and poor planning; commercial sites struggled with coordination gaps and financial bottlenecks; and industrial projects, often large-scale and complex, were hit hardest by regulatory hurdles and macroeconomic instability. By interpreting machine learning outputs with SHAP analysis, this research didn't just predict cost overruns it revealed why they happen and which variables under which category are of utmost significance. In doing so, it provided not just insight but foresight, enabling smarter, risk-aware decision-making. Ultimately, this project demonstrates that intelligent data intertwined with integrated models of approach use can move the construction industry towards proactive planning and sustainable execution.

This study's unique contributions advance beyond prior work by delivering sector-tailored risk profiles and predictive models informed by real-time stakeholder perceptions across residential, commercial, and industrial/heavy sectors — an underexplored area in the literature. While previous studies often apply ML generically or to single typologies (e.g., high-rise residential in Shoar et al., [12]) or region-specific quantitative data (e.g., Hamdan et al., [16]), this framework integrates diverse stakeholder insights with Random Forest regression and explainable techniques (SHAP for feature contributions, PCA for risk clustering). This enables granular, interpretable insights into sector-distinct drivers (e.g., planning in residential, financing in commercial, scope changes in industrial), providing actionable, context-aware mitigation strategies that bridge a key gap in proactive, sector-differentiated cost overrun management

4.1. Limitations and Scope for further work

This study provides valuable sector-specific insights into cost overrun drivers using stakeholder-informed perceptual data, but several limitations must be acknowledged to contextualize the findings and guide future research.

First, the analysis relies on a perception-based dataset collected via a 5-point Likert-scale questionnaire from 70 diverse stakeholders (engineers, contractors, and owners) in the Indian construction context. While perceptual data effectively captures qualitative risk factors and stakeholder experiences common in construction management surveys [19] it is inherently subjective and susceptible to response bias, such as social desirability, recall inaccuracies, or varying individual risk perceptions [29, 17]. Likert-scale responses can also introduce information loss due to discretization of continuous beliefs, potentially reducing nuance in complex risk interactions (as discussed in psychometric literature on ordinal measures).

Second, the sample size ($n=70$), though adequate for exploratory ML applications like Random Forest in construction studies [19], is relatively small and regionally focused (primarily Tamil Nadu, India). This limits generalizability to other geographical, cultural, or economic contexts, where risk profiles, regulatory environments, and material/labour dynamics may differ significantly. Small perceptual datasets in ML-based cost prediction can also lead to overfitting risks or reduced robustness when applied to diverse real-world scenarios [30], recent reviews on survey limitations in construction ML.

Third, the study lacks longitudinal validation or integration with objective, quantitative project data (e.g., actual budgets, expenditures, timelines from historical records). Perception-based models, while insightful for early-stage risk identification, may not fully capture dynamic, measurable factors like real-time material price fluctuations or site-specific events.

Despite these constraints, the high internal consistency (Cronbach's alpha > 0.88 across sectors) and strong model performance (R^2 up to 0.8715 in residential/commercial sectors) support the framework's exploratory value. Future work should address these limitations by:

- Scaling the model to larger, hybrid datasets combining perceptual surveys with archival quantitative project records from multiple regions and countries, enhancing external validity and predictive robustness (as recommended in recent ML cost overrun studies; e.g., Hamdan et al., [13]).
- Incorporating real-time data sources, such as IoT sensors for monitoring site conditions, material usage, and progress, to enable dynamic, adaptive predictions (emerging trend in AIoT for construction; e.g., applications in predictive maintenance and energy management, 2025 studies).
- Exploring hybrid ML approaches (e.g., combining Random Forest with deep learning, fuzzy logic, or optimization algorithms) to better handle uncertainty, nonlinearity, and domain-specific complexities under high variability.
- Conducting longitudinal or cross-validation studies with larger, multi-country samples to test model transferability and refine sector-specific risk archetypes.

These directions will build on the current framework to advance more reliable, scalable, and practically deployable tools for cost overrun mitigation in global construction.

4.2. Implications for Theory and Practice

This study advances construction management theory by introducing a sector-differentiated predictive framework that integrates stakeholder perceptions with ensemble machine learning (Random Forest) and explainable AI techniques (SHAP and PCA). Theoretically, it bridges key gaps in the literature by demonstrating how sector-specific risk profiles—such as planning deficiencies in residential projects, financial vulnerabilities in commercial ones, and scope volatility in industrial/heavy projects can be systematically modelled and interpreted, extending beyond generic ML applications [12, 16]. The use of SHAP provides granular, game-theoretic explanations of feature contributions, enhancing model transparency and addressing the longstanding "black-box" critique of advanced predictive models in construction. This contributes to emerging theories on explainable AI adoption in civil engineering, where interpretability fosters trust and supports theory-building around nonlinear risk interactions.

From a practical perspective, the framework equips stakeholders (engineers, contractors, owners) with actionable, sector-tailored insights to prioritize high-impact risk factors. For instance, feature importance and SHAP analyses highlight the need to focus early mitigation efforts on financing difficulties in commercial projects, improper planning and coordination in residential ones, and scope changes in industrial contexts. By enabling proactive contingency planning, early risk detection, and targeted interventions, the model supports improved budget adherence, reduced disputes, and enhanced project profitability. These outcomes align with industry needs for data-driven decision-making, offering a scalable tool that can be adapted to real-time quantitative project data in future implementations. This dual contribution strengthening theoretical understanding of sector-specific cost dynamics while delivering practical tools for risk mitigation positions the research as a meaningful step toward more resilient and sustainable construction practices.

References

- [1] B. Flyvbjerg, M. K. Skamris holm, and S. L. Buhl, “How common and how large are cost overruns in transport infrastructure projects?,” *Transp Rev*, vol. 23, no. 1, pp. 71–88, Jan. 2003, doi: 10.1080/01441640309904.
- [2] H. Doloi, “Cost Overruns and Failure in Project Management: Understanding the Roles of Key Stakeholders in Construction Projects,” *J Constr Eng Manag*, vol. 139, no. 3, pp. 267–279, Mar. 2013, doi: 10.1061/(ASCE)CO.1943-7862.0000621.
- [3] P. E. D. Love, D. D. Ahiaga-Dagbui, and Z. Irani, “Cost overruns in transportation infrastructure projects: Sowing the seeds for a probabilistic theory of causation,” *Transp Res Part A Policy Pract*, vol. 92, pp. 184–194, Oct. 2016, doi: 10.1016/j.tra.2016.08.007.
- [4] L. Breiman, “Random Forests,” *Mach Learn*, vol. 45, pp. 5–32, 2001.
- [5] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Nov. 2017.
- [6] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.
- [7] J. H. Jung, D. Y. Kim, and H. K. Lee, “The computer-based contingency estimation through analysis cost overrun risk of public construction project,” *KSCE Journal of Civil Engineering*, vol. 20, no. 4, pp. 1119–1130, May 2016, doi: 10.1007/s12205-015-0184-8.
- [8] J. Theodore J. Trauner and William A. Manginelli, *Construction Delays*. Elsevier, 2009. doi: 10.1016/B978-1-85617-677-4.X0001-3.
- [9] A. Kavuma, J. Ock, and H. Jang, “Factors influencing Time and Cost Overruns on Freeform Construction Projects,” *KSCE Journal of Civil Engineering*, vol. 23, no. 4, pp. 1442–1450, Apr. 2019, doi: 10.1007/s12205-019-0447-x.
- [10] S. Durdyev, “Review of construction journals on causes of project cost overruns,” *Engineering, Construction and Architectural Management*, vol. 28, no. 4, pp. 1241–1260, Apr. 2021, doi: 10.1108/ECAM-02-2020-0137.
- [11] Mostafa H. Kotb and Mohamed M. Ghattas, “Risk Identification Barriers in Construction Projects in MENA,” *PM World Journal*, vol. 6, no. 8, Aug. 2018.
- [12] S. Shoar, N. Chileshe, and J. D. Edwards, “Machine learning-aided engineering services’ cost overruns prediction in high-rise residential building projects: Application of random forest regression,” *Journal of Building Engineering*, vol. 50, p. 104102, Jun. 2022, doi: 10.1016/j.jobe.2022.104102.
- [13] M. Tayyab, M. Furkhan, M. Rizwan, M. Jameel, and A. Chadee, “A Study on Factors Influencing Cost Overrun in High-rise Building Construction across India,” *Journal of Smart Buildings and Construction Technology*, vol. 5, no. 1, pp. 52–83, May 2023, doi: 10.30564/jsbct.v5i1.5489.
- [14] G. Kazar, U. Yiğit, and K. E. Boyabatlı, “Predicting maintenance cost overruns in public school buildings using a rough topological approach,” *Autom Constr*, vol. 168, p. 105810, Dec. 2024, doi: 10.1016/j.autcon.2024.105810.
- [15] G. H. Coffie and S. K. F. Cudjoe, “Toward predictive modelling of construction cost overruns using support vector machine techniques,” *Cogent Eng*, vol. 10, no. 2, Dec. 2023, doi: 10.1080/23311916.2023.2269656.
- [16] M. M. Hamdan, M. Thneibat, and K. Hyari, “Predicting cost overrun in construction projects using machine learning algorithms: the case of Jordan,” *Engineering, Construction and Architectural Management*, Apr. 2025, doi: 10.1108/ECAM-09-2024-1209.
- [17] A. H. Turkyilmaz and G. Polat, “Risk-Based Completion Cost Overrun Ratio Estimation in Construction Projects Using Machine Learning Classification Algorithms: A Case Study,” *Buildings*, vol. 14, no. 11, p. 3541, Nov. 2024, doi: 10.3390/buildings14113541.

- [18] Theingi Aung, S. R. Liana, A. Htet, and Amiya Bhaumik, “Using Machine Learning to Predict Cost Overruns in Construction Projects,” *Journal of Technology Innovations and Energy*, vol. 2, no. 2, pp. 1–7, Jun. 2023, doi: 10.56556/jtie.v2i2.511.
- [19] A. A. Aibinu and G. O. Jagboro, “The effects of construction delays on project delivery in Nigerian construction industry,” *International Journal of Project Management*, vol. 20, no. 8, pp. 593–599, Nov. 2002, doi: 10.1016/S0263-7863(02)00028-5.
- [20] P. F. Kaming, P. O. Olomolaiye, G. D. Holt, and F. C. Harris, “Factors influencing construction time and cost overruns on high-rise projects in Indonesia,” *Construction Management and Economics*, vol. 15, no. 1, pp. 83–94, Jan. 1997, doi: 10.1080/014461997373132.
- [21] Y. A. Olawale and M. Sun, “Cost and time control of construction projects: inhibiting factors and mitigating measures in practice,” *Construction Management and Economics*, vol. 28, no. 5, pp. 509–526, May 2010, doi: 10.1080/01446191003674519.
- [22] J. Zhou, X. Shi, K. Du, X. Qiu, X. Li, and H. S. Mitri, “Feasibility of Random-Forest Approach for Prediction of Ground Settlements Induced by the Construction of a Shield-Driven Tunnel,” *International Journal of Geomechanics*, vol. 17, no. 6, Jun. 2017, doi: 10.1061/(ASCE)GM.1943-5622.0000817.
- [23] X. Liu, Y. Song, W. Yi, X. Wang, and J. Zhu, “Comparing the Random Forest with the Generalized Additive Model to Evaluate the Impacts of Outdoor Ambient Environmental Factors on Scaffolding Construction Productivity,” *J Constr Eng Manag*, vol. 144, no. 6, Jun. 2018, doi: 10.1061/(ASCE)CO.1943-7862.0001495.
- [24] Y. Zhang *et al.*, “How Does Experience with Delay Shape Managers’ Making-Do Decision: Random Forest Approach,” *Journal of Management in Engineering*, vol. 36, no. 4, Jul. 2020, doi: 10.1061/(ASCE)ME.1943-5479.0000776.
- [25] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
- [26] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” Jan. 2019.
- [27] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation,” *PeerJ Comput Sci*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.
- [28] A. M. Jarkas and C. G. Bitar, “Factors Affecting Construction Labor Productivity in Kuwait,” *J Constr Eng Manag*, vol. 138, no. 7, pp. 811–820, Jul. 2012, doi: 10.1061/(ASCE)CO.1943-7862.0000501.
- [29] K. Koc, Ö. Ekmekcioğlu, and A. P. Gurgun, “Accident prediction in construction using hybrid wavelet-machine learning,” *Autom Constr*, vol. 133, p. 103987, Jan. 2022, doi: 10.1016/j.autcon.2021.103987.
- [30] S. Shoar, T. W. Yiu, S. Payan, and M. Parchamijalal, “Modeling cost overrun in building construction projects using the interpretive structural modeling approach: a developing country perspective,” *Engineering, Construction and Architectural Management*, vol. 30, no. 2, pp. 365–392, Mar. 2023, doi: 10.1108/ECAM-08-2021-0732