JADS-TUGAS MENPRO KEL 6 FIX 2

by arum puspita

General metrics

76,444	10,430	635	41 min 43 sec	1 hr 20 min
characters	words	sentences	reading time	speaking time

Score

This text scores better than 95% of all texts checked by Grammarly

Writing Issues

183	15	168
Issues left	Critical	Advanced

Writing Issues

19	**Correctness**	
4	Misspelled words	
4	Improper formatting	
3	Determiner use (a/an/the/this, etc.)	
2	Closing punctuation	
1	Comma misuse within clauses	
1	Incorrect noun number	
1	Confused words	

1	Mixed dialects of english	
2	Unknown words	
3	**Clarity**	
3	Wordy sentences	

Unique Words

Measures vocabulary diversity by calculating the percentage of words used only once in your document

18%

unique words

Rare Words

Measures depth of vocabulary by identifying words that are not among the 5,000 most common English words.

47%

rare words

Word Length

Measures average word length

5.8

characters per word

Sentence Length

Measures average sentence length

16.4

words per sentence

JADS-TUGAS MENPRO KEL 6 FIX 2

Journal of Applied Data Sciences

Vol. 5, No. 3, September 2024, pp. 1449-1461

ISSN 2723-6471

1

Journal of Applied Data Sciences

Vol. 5, No. 3, September 2024, pp. X-X

ISSN 2723-6471

1

Analisis Cost Overrun pada Proyek Konstruksi Menggunakan Machine Learning: Inspirasi untuk Penerapan di Manajemen Proyek Perangkat Lunak

Arum Puspita

*Corresponding author: author (author@mail.com)

DOI: https://doi.org/10.47738/jads.v5i2.XXX

,*,, Siti Wasi'atul Maghfiroh2,, Nabila Mutiara Sani3, , Putri Rahayu Agustini4,

1,2,3,4Universitas Islam Negeri Maulana Malik Ibrahim, Jl. Gajayana No.50, Dinoyo, Kec. Lowokwaru,

Kota Malang, Jawa Timur 65144

(Received: July 4, 2024; Revised: August 31, 2024; Accepted: September 11, 2024; Available online: September 23, 2024)

Abstract

Cost overrun is a common issue in construction project management, particularly in road projects. This study aims to develop a machine learning-based classification model to predict the frequency of cost overruns using structured survey data consisting of 44 independent variables and 139 observations. Five algorithms were evaluated: Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Random Forest, and Multi-Layer Perceptron. After preprocessing and feature selection, the Random Forest algorithm demonstrated the best accuracy of 85% using the SelectFromModel method. Key factors influencing cost delays include contractor experience, material availability, and financial management. These findings indicate that machine learning can be an effective tool to support early decision-making in controlling construction project costs.

Keywords: Cost Overrun, Project Management, Machine Learning, Random Forest Classification,

Introduction

Cost overrun refers to a condition in which the actual project cost exceeds the predetermined budget. This phenomenon frequently occurs in construction projects due to inaccurate initial cost estimations and weak control during execution [1]. Additionally, on-site dynamics such as design changes, weather conditions, and procurement delays further exacerbate the risk of overruns [2]. The use of technology such as machine learning (ML) is increasingly being adopted in project management to predict potential cost overruns. ML is capable of processing historical project data and building accurate predictive models to identify high-risk projects from the early stages [3]. This makes ML

an efficient decision-support tool for project managers in allocating resources [4]. Therefore, the implementation of ML is expected to minimize the potential for cost overruns and enhance the efficiency of project execution.

One of the most commonly used algorithms is the Artificial Neural Network (ANN). ANN is capable of modeling non-linear relationships between causal variables and actual project cost outcomes, and it demonstrates strong predictive performance when trained on sufficient data [5]. Additionally, Random Forest is utilized to classify projects based on cost risk and to identify the most influential factors [6]. Support Vector Machine (SVM) is also an effective method for distinguishing between projects that are at risk of cost overrun and those that are not, by applying a maximum-margin approach to project data [7]. In several studies, SVM has achieved predictive accuracy comparable to ANN and has outperformed it in cases involving imbalanced datasets [8]. The selection of the appropriate algorithm depends on the characteristics of the data and the analytical objectives to be achieved in project risk management.

Feature selection is a crucial step in building a reliable prediction model. Techniques such as Recursive Feature Elimination (RFE) can be employed to retain the most relevant variables, thereby improving both the accuracy and efficiency of the model [9]. In several cases, features such as contractor experience, contract type, and project duration have proven highly influential in prediction outputs [10]. The growing adoption of analytics-based approaches and predictive models in modern construction underscores the importance of data-driven decision-making. One study demonstrated that leveraging ensemble learning yields stable accuracy when predicting early-stage cost-overrun determinants [11]. while[2] the integration of machine learning into project-management information systems has been shown to enhance

oversight effectiveness and cost-expenditure transparency [12]. Consequently, the combination of well-chosen features and appropriate predictive algorithms is key to developing an optimal cost-overrun prediction system.

Beyond technical considerations, several studies highlight that social factors such as stakeholder collaboration, participatory risk management, and public perceptions of a project also influence a project's financial outcomes [13]. In a global context, aligning construction data with government policies is essential for addressing cost overruns in both developing and developed nations [14]. Moreover, the interplay between adaptive project-management practices and the renewable-energy sector is being investigated to identify sustainability contributions within infrastructure projects [15]. These insights affirm that a multidisciplinary approach is indispensable for comprehensively understanding and managing cost overruns.

Literature Review

literature review that situates this research within the broader discourse on construction cost management. Cost overruns have been a persistent challenge for decades, compromising project schedules, eroding profit margins, and undermining stakeholder confidence. Traditional cost-estimation techniques, though useful, often struggle to keep pace with today's increasingly complex, fast-moving project environments. Recent advances in data analytics and machine learning offer promising alternatives that can process multifaceted project variables more effectively than purely deterministic or expert-driven methods. By surveying prior studies, this chapter identifies the key theoretical foundations, empirical findings, and methodological innovations that inform the present work.

The review is organized to trace the evolution of research from conventional estimation practices to modern, data-driven approaches. It first examines how

dynamic and uncertain project conditions necessitate more adaptive cost-prediction frameworks, then explores the growing body of studies that integrate machine learning and Building Information Modeling for early overrun detection. Subsequent sections analyze risk-based modeling strategies, evaluate model performance metrics, and highlight external socio-political influences that complicate cost control. Collectively, these strands of literature reveal both the potential and the current limitations of predictive analytics in construction management. This synthesis sets the stage for the study's own[3] methodological contributions and underscores the importance of a holistic, continually updated approach to cost-overrun prediction.

Cost Estimation and the Challenge of Cost Overruns in Construction Projects

Cost estimation in construction projects is a complex process influenced by numerous dynamic and uncertain variables. Traditional estimation approaches that rely on expert intuition or parametric methods often result in low accuracy, particularly when unexpected risks arise during project execution [16]. Such limitations can lead to inadequate budgeting, delayed timelines, and inefficient resource allocation. These consequences not only affect project outcomes but also reduce stakeholder trust and satisfaction. Therefore, there is a growing need for more adaptive and data-driven methods to improve estimation precision in modern construction environments.

With the advancement of technology, approaches based on Artificial Intelligence (AI) and Machine Learning (ML) have been increasingly employed to develop more accurate and efficient cost prediction systems. Several studies have shown that integrating ML into project management systems can help identify the risk of cost overruns at an early stage [17]. Early detection allows project managers to take corrective actions proactively, reducing the likelihood of major financial setbacks. As a result, ML serves not only as a prediction tool

but also as a strategic decision-support mechanism in construction project management.

Application of Machine Learning in Predicting Cost Overruns

Machine learning algorithms have been widely used to predict potential cost overruns by leveraging historical construction project data. These algorithms can uncover complex patterns and interactions among project variables that are difficult to detect using traditional methods. Classification models such as decision trees, support vector machines, and ensemble learning are commonly employed to estimate the cost overrun ratio up to project completion [18]. Compared to conventional statistical approaches, these models generally provide higher accuracy, especially in fast-paced and dynamic project environments. Their adaptability makes them highly valuable in handling uncertainty and variability within real-world construction settings.

In addition to standalone ML models, several studies have explored the integration of Machine Learning with Building Information Modeling (BIM) for forensic delay analysis. This combined approach enables more detailed tracking of project performance and supports dispute resolution by offering a data-driven explanation of delays. BIM's visualization tools and chronological tracking features allow stakeholders to assess the root causes of schedule disruptions more objectively and transparently [19]. By merging predictive capabilities with visual analytics, this integration enhances both the diagnostic and communicative functions of project analysis. As a result, it supports more informed decision-making and fosters accountability across project stakeholders.

Risk-Based Modeling in Project Management

Many researchers have identified risk as one of the primary causes of cost overruns in construction projects. Risks may stem from various sources,

including technical complexities, resource constraints, stakeholder misalignment, and unforeseen site conditions. To address this, several studies have developed dynamic risk assessment models that are periodically updated throughout the project lifecycle. These models utilize machine learning classification algorithms to predict cost deviations based on continuously evolving risk indices, allowing for real-time insights and improved forecasting accuracy [20]. This predictive capability empowers project teams to implement early mitigation strategies and reduce the impact of emerging threats during the execution phase.

Interestingly, even well-funded projects are not immune to delays and cost overruns. Internal factors such as poor contractor performance, and external factors like regulatory delays or political instability, often disrupt project timelines. This suggests that sufficient financial resources alone do not ensure successful project delivery [21]. Instead, a comprehensive and continuous risk evaluation process is essential for identifying vulnerabilities beyond budgetary concerns. Ultimately, integrating risk-based modeling with machine learning provides a more proactive and data-driven foundation for managing uncertainty and ensuring long-term project success.

Performance Evaluation of ML Models in Cost Prediction

Evaluating the performance of machine learning models is a critical step in studies aimed at predicting cost overruns. Accurate evaluation ensures that the model not only fits the training data well but also generalizes effectively to unseen projects. Comparative research involving Random Forest, Neural Network, and Support Vector Machine (SVM) models has shown that Random Forest tends to yield the highest accuracy in road construction projects, particularly during the conceptual stage [22]. This is largely due to its strength in handling non-linear relationships and its robustness against overfitting in

high-dimensional, complex datasets. Such characteristics make it especially useful for the unpredictable nature of early-stage infrastructure planning. Beyond model accuracy, practical implementation remains a key consideration for project stakeholders. Several studies emphasize the value of integrating ML prediction systems with visualization tools like Building Information Modeling (BIM) to improve communication and usability [23]. By visualizing cost risk predictions in real time, project teams and decision-makers can quickly grasp critical insights. This allows for more informed discussions and faster responses to emerging financial risks. As a result, combining predictive analytics with intuitive visual platforms contributes to more transparent, collaborative, and data-driven project management practices.

External and Socio-Political Factors in Project Cost

Cost performance in construction projects is influenced not only by technical and managerial aspects but also by external variables such as weather conditions, labor market fluctuations, and socio-political dynamics [24]. These factors often introduce uncertainty and can lead to significant delays or cost deviations. While such variables are inherently difficult to quantify, recent developments in feature engineering have enabled their inclusion in machine learning–based predictive models. By transforming qualitative or semi-structured data into usable numerical features, these models can better capture real-world complexity. As a result, predictive accuracy improves when external risks are systematically accounted for.

In the context of developing countries, the impact of these external factors is often more pronounced due to limited infrastructure and institutional stability. Studies have shown that even when cost estimates are initially based on minimal attributes, machine learning models can substantially enhance their precision [25]. This allows project planners to develop more informed

procurement strategies and conduct negotiations from a stronger position early in the project lifecycle. Moreover, ML-supported estimates facilitate better alignment between stakeholders, funding bodies, and regulatory agencies. Ultimately, leveraging ML in such environments not only improves forecasting but also supports smarter, resource-sensitive decision-making.

Project Management Challenges and External Risk Factors

The management of road construction projects often faces external obstacles such as land acquisition issues, extreme weather conditions, and fluctuations in material prices. A study in Jammu and Kashmir revealed that these factors are among the primary causes of project delays and cost overruns [26]. Furthermore, a systematic review of construction projects in developing countries highlighted poor planning, repeated design changes, and a lack of coordination among stakeholders as dominant causes of overruns [27]. This indicates that internal issues within the management process can have an impact just as significant as external factors.

The implementation of integrated project management systems is considered effective in improving project execution efficiency and reducing resource waste, especially when accompanied by real-time, data-driven monitoring [28]. Nevertheless, studies have also found that resistance to change and a less adaptive organizational culture can hinder the adoption of new technologies [29]. Another issue affecting project efficiency is poor communication between project managers and technical teams, which leads to delays in decision-making [30]. Such communication breakdowns often result in technical misunderstandings and unsynchronized execution in the field. On the other hand, social and administrative factors also play a major role. Poor inter-agency coordination and inconsistencies in procurement processes can lead to serious consequences for project budgets [31], [32]. To address these issues, several

studies recommend the integration of digital technologies, such as cloud-based document management systems, to accelerate workflows and maintain transparency.

Machine Learning for Cost Overrun Prediction

Machine learning models such as Random Forest and Support Vector Machine (SVM) have been widely adopted to predict cost overruns in construction projects [33]. These algorithms are particularly well-suited for capturing the nonlinear and complex relationships among project variables, offering greater stability and accuracy than conventional statistical techniques. Their performance, however, is highly dependent on the quality, balance, and structure of the training data. In real-world applications, noisy or incomplete datasets can significantly reduce the reliability of predictions. Therefore, careful data preprocessing and validation are essential to ensure meaningful outcomes.

In the context of imbalanced datasets such as those used for project risk classification the selection of appropriate evaluation metrics, including the F1-score and precision, becomes crucial [34]. Beyond technical considerations, project-evaluation frameworks should also encompass indicators of sustainability, stakeholder satisfaction, and output quality to ensure that technological implementations have a tangible impact in the field [35]. Such a holistic evaluation approach is needed for digital transformation to make a meaningful contribution to overall project performance. Other studies highlight the importance of regularly updating models based on actual project conditions, particularly in large-scale, complex, and long-term road projects. This practice allows project managers to adjust predictions dynamically throughout the project lifecycle, leveraging machine learning not only as an initial estimation tool but also as a data-driven decision-support system.

Methodology

In recent years, machine learning–based approaches have proven effective in addressing the complex challenges of construction cost estimation. As construction projects become increasingly complex and work environments more uncertain, machine learning algorithms offer a more adaptive and data-driven alternative with greater precision compared to traditional methods [36]. Moreover, the integration of artificial intelligence with construction risk management enables faster decision-making in response to unexpected cost fluctuations [37]. Through this integration, project managers can respond to field dynamics in real time[5] and minimize potential delays or budget overruns. Previous studies have also emphasized the importance of predictive modeling in identifying the frequency of cost overruns, rather than merely estimating their absolute values. This facilitates the development of risk mitigation policies that are more responsive to recurring patterns of cost delays [38]. Therefore, the methodology in this study is designed to support the development of a machine learning–based classification system focused on predicting the frequency of cost overruns in road construction projects.

Research Design

This study aims to identify and predict the frequency of cost overruns in road construction projects using a machine learning–based approach. The research employs a quantitative methodology that combines both exploratory and predictive elements. Secondary data collected from structured surveys involving construction project stakeholders serves as the primary data source. The overall framework for this study is grounded in machine learning–based project analytics, which provides a systematic approach to data-driven decision-making [39]. Through this approach, the research seeks to contribute

to the development of more accurate and practical predictive systems for managing infrastructure project costs.

A data-driven analytics method is adopted to explore the relationship between various causal factors and the frequency of cost delays. The core objective is to build a predictive classification model that can categorize projects into two groups: those with rare cost overruns and those with frequent cost overruns. In achieving this goal, the study also compares the performance of several machine learning algorithms. The research methodology consists of several structured phases, including data collection and preparation, data preprocessing, model training and evaluation, and feature selection combined with feature importance analysis. Each stage is designed to ensure model robustness, interpretability, and practical relevance in real-world construction project settings.

Dataset and Data Sources

The data used in this study were obtained from a structured survey conducted across multiple road construction projects. The dataset comprises 139 observations and includes 44 independent variables, labeled Q0 to Q43. These variables cover various aspects of project execution, such as contractor experience, material availability, financial management, and project plan consistency. The dataset also contains one dependent variable, Q50, which indicates the frequency level of cost delays. This variable is measured using a five-point Likert scale, where 1 means "very rarely" and 5 means very frequently.

To facilitate the classification process, the dependent variable Q50 was transformed into a binary categorical variable. The new labels divide responses into two classes: "rare" for values below 4, and "frequent" for values equal to or greater than 4. This transformation simplifies the modeling task and aligns with

standard practices in binary classification using machine learning algorithms [40]. It also helps to improve the model's ability to differentiate between low- and high-risk projects. Ultimately, this approach enhances the predictive performance by focusing on the frequency patterns of cost overruns as perceived by project stakeholders.

Data Preprocessing

The data preprocessing stage was designed to ensure that the dataset was properly prepared for machine learning algorithms. First, all categorical variables were converted into numerical format using label encoding, allowing the models to effectively process qualitative information. Missing values were then addressed by scanning each row and column for incomplete entries and applying median imputation as needed, which helps preserve important data while reducing the risk of bias in the model inputs [41]. Because Random Forest and Decision Tree models are non-parametric and largely insensitive to the scale of input features, explicit normalization or standardization was not applied in this study. This decision is based on the nature of these algorithms, which do not rely on distance-based calculations and therefore do not require scaled data to perform well. However, normalization remains a consideration for other models such as SVM or neural networks, where feature scaling plays a critical role in achieving optimal performance.

The dataset was subsequently divided into training and testing subsets using the train_test_split function from Scikit-learn, with 70% of the data allocated for training and 30% for evaluation. To ensure consistent and reproducible results across experiments, a fixed random state was specified during the data split. This structured approach to preprocessing enables the construction of predictive models on a clean, well-prepared dataset, reducing bias introduced

by uneven data distribution and increasing the reliability of the final predictions[7]

Model Selection and Training

Several machine learning algorithms were employed to compare classification performance in relation to[8] the target variable, which represents the frequency of cost overruns. The five models selected for this study include Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and Multi-Layer Perceptron (MLP). These models were intentionally chosen to represent a diverse spectrum of classification techniques, from simple linear models to more complex ensemble and neural network–based approaches. This diversity allows for a more comprehensive evaluation of algorithm effectiveness under various data conditions. Each model brings unique strengths in terms of handling non-linearity, interpretability, and sensitivity to class imbalance.

To reduce the risk of overfitting and enhance generalization performance, the training process utilized 5-fold cross-validation. This technique ensures that each model is evaluated across multiple data splits, thereby providing a more reliable estimate of performance on unseen data. Additionally, model optimization was conducted through hyperparameter tuning using the Grid Search technique, which systematically explores a predefined set of parameters to identify the best-performing combination for each algorithm [42]. By incorporating both cross-validation and hyperparameter optimization early in the training process, this approach ensures that the resulting models not only perform well on the training data but also maintain consistent accuracy when applied to new, unseen datasets. This balance between learning and generalization is essential for the deployment of predictive models in real-world project environments.

Model Performance Evaluation

Following the training phase, the models were evaluated using the test dataset based on several performance metrics: accuracy, precision, recall, and F1-score. These metrics were selected to provide a comprehensive understanding of each model's classification capability, especially in scenarios involving class imbalance. Accuracy alone can be misleading when one class dominates the dataset, making precision and recall critical for fair evaluation. The F1-score, as a harmonic mean of precision and recall, offers a balanced view of performance across both classes. Together, these metrics help determine not only which model performs best overall but also which one is more reliable in predicting minority class instances.

The evaluation results revealed that the Random Forest model achieved the highest accuracy at 72.1%, outperforming Logistic Regression and Multi-Layer Perceptron, which showed slightly lower but still competitive results. In contrast, K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) produced suboptimal outcomes, largely due to their sensitivity to the number of input features and the presence of imbalanced classes in the dataset [43]. These findings support existing evidence that ensemble-based algorithms like Random Forest are more robust when dealing with the complexities of construction project data, particularly in cases where the class distribution is skewed. This robustness makes Random Forest a reliable choice for real-world applications in predicting cost overrun frequencies.

Feature Selection

To enhance model accuracy and interpretability, a feature selection process was conducted using three primary techniques: Univariate Feature Selection (UFS), Recursive Feature Elimination (RFE), and SelectFromModel (SFM). These techniques are designed to eliminate irrelevant or redundant variables that

may introduce noise into the model. By narrowing down the input space, the resulting model becomes more focused, efficient, and less prone to overfitting. Each method was systematically applied and evaluated to determine which yielded the most optimal classification performance. The overarching goal was to simplify the model structure while preserving, or even improving, its predictive capability.

Among the methods tested, SelectFromModel (SFM) produced the best results. This approach selected 19 important features that contributed most significantly to the model's performance. After applying SFM, the accuracy of the Random Forest model increased from 72.1% to 85% [44]. This substantial increase highlights the impact of effective feature selection in refining model accuracy and robustness. Furthermore, it reduced computational overhead and enhanced the model's interpretability, making it more suitable for practical use in construction project analysis.

Feature Visualization and Interpretation

Visualizations such as feature importance plots, correlation heatmaps, and scatter plots were used to understand the relationship between features and the target variable. Key features identified include Q0 (Contractor Experience), Q22 (Material Availability), and Q43 (Financial Management). These features were found to be the most influential in contributing to cost overruns. Their strong impact aligns with findings from previous studies, which emphasized the importance of early planning, logistical readiness, and financial control in construction projects. Visualization played a crucial role in confirming the statistical relevance of these factors.

The results highlight the critical importance of early project information management and risk estimation documentation. Proper interpretation of these features allows project managers to better anticipate potential issues.

This enables more proactive and data-driven decision-making throughout the project lifecycle. Visual tools make it easier to communicate findings to stakeholders who may not be familiar with technical models. Overall, these insights demonstrate that data visualization is not only useful for analysis but also for strategic planning and communication.

Results and Discussion

This section discusses the results of the analysis conducted on survey data regarding delays in road construction projects, the testing of various machine learning models, and an in-depth interpretation of the evaluation outcomes. The primary objective of this analysis is to comprehensively understand the factors contributing to cost overruns in road construction projects and to assess the effectiveness of predictive models in anticipating potential cost delays. This data-driven approach aims to provide a holistic perspective on the dynamics and complexities of project delay issues, as well as to identify targeted solutions through analytics-based mitigation strategies. The road construction project delay survey provides a rich and diverse dataset for analysis. By applying machine learning techniques, hidden patterns within the data can be uncovered, enabling the identification of key risk factors that might otherwise be challenging to detect using traditional analytical methods. These techniques offer not only a retrospective perspective for understanding past project failures but also a prospective approach to support more informed decision-making in future projects. Thus, machine learning serves as both a predictive and diagnostic tool for the dynamic and complex construction sector.

Furthermore, the results of this analysis are expected to make a tangible contribution to project stakeholders, including contractors, consultants, project managers, and project owners, in formulating more evidence-based

plans and policies. By identifying common patterns behind cost overruns—such as weak financial management, logistical inadequacies, and ineffective risk management—preventive measures can be designed and implemented earlier in the project lifecycle. These findings are also anticipated to encourage the adoption of project management practices that are more adaptive, accurate, and results-oriented, thereby increasing the likelihood of project success in the future. This analysis also highlights the importance of integrating modern analytical approaches with conventional project management practices. By leveraging the sophistication of machine learning, construction projects can be better directed toward achieving a balance between cost efficiency, timeliness, and quality outcomes. This data-driven approach is expected to serve as a foundation for building project management systems that are more resilient, innovative, and competitive in the era of Industry 4.0.

Frequency Distribution of Cost Overrun (Q50)

The distribution of Q50 values delineates the frequency of cost overruns reported by survey respondents. The data, gathered through a Likert scale, identifies a general pattern concerning delays in project execution. The Q50 score distribution demonstrates that a majority of projects encounter moderate to high levels of cost overruns. This finding underscores the prevalence of cost overruns as a critical issue within the road construction sector, significantly impairing budget allocation efficiency and adherence to project timelines. The histogram distribution further indicates that scores of 4 and 5 dominate the responses. This suggests that cost overruns are not sporadic occurrences but rather a recurring phenomenon requiring urgent attention. Respondents assigning a score of 5, representing "very frequent"

cost overruns, form the largest cohort. This observation is particularly alarming as it reflects inadequate cost control mechanisms in most projects.

This trend can be attributed to several factors, including insufficient initial planning, mid-project design alterations, and external pressures such as fluctuating material prices. Internal inefficiencies, such as ineffective communication among project teams, delayed payments to subcontractors, and deficiencies in risk management, also play a significant role in the high incidence of cost overruns. Further analysis reveals that projects situated in geographically challenging areas are more likely to report higher Q50 scores, highlighting the critical influence of geographical factors. For instance, projects in remote locations often face additional difficulties in procuring materials, labor, and equipment, thereby exacerbating cost overruns. Respondents with elevated Q50 scores predominantly represent large-budget projects and construction sites of heightened complexity. These projects often involve numerous stakeholders, intricate contractual arrangements, and greater risks of design changes. External factors such as adverse weather conditions, extreme geographical terrains, and political instability further compound the issue of cost overruns.

Projects with overly ambitious schedules established during the initial planning stage are particularly susceptible to cost overruns. Unrealistic timelines result in activity backlogs during the project's final stages, necessitating overtime and additional resource utilization, which ultimately inflate costs. Projects of this nature often exhibit suboptimal risk management processes, which hinder the prompt resolution of emerging challenges during the execution phase. This underscores the critical importance of implementing risk-based project management methodologies and rigorous oversight mechanisms to effectively manage changes throughout the project lifecycle. Figure 1 illustrates the

frequency distribution of cost overruns in construction projects, measured on a Likert scale ranging from 1 to 5. Survey findings indicate that most respondents assigned scores of 4 and 5, signifying that cost overruns occur frequently to very frequently. This highlights the significance of cost overruns as a pressing issue within the construction sector, adversely impacting budget allocation and project completion timelines.

The respondents assigning the highest score (5) constitute the largest group, emphasizing the persistent challenge of cost management. Contributing factors may include inadequate preliminary planning, design modifications during project execution, and material supply chain instability. Additionally, internal issues such as ineffective communication among project teams, delayed subcontractor payments, and insufficient risk management practices exacerbate the problem. These findings underscore the necessity of adopting enhanced mitigation strategies to address the underlying causes of cost overruns and reduce their occurrence in future projects.

Figure 1. Frequency Distribution of Cost Overrun (Q50 Variable)

Correlation Analysis Among Contributing Factors

A correlation analysis was conducted to investigate and clarify the linear relationships among factors influencing project delays, particularly in the context of cost overruns in road construction projects. A strong positive correlation between two variables indicates that an increase in one variable is accompanied by a corresponding rise in the other. Such a relationship implies that these factors frequently co-occur in projects experiencing significant challenges. Conversely, a negative correlation denotes an inverse relationship, where an increase in one variable correlates with a decrease in the other.

Accordingly, correlation analysis serves as a vital instrument for identifying latent patterns of interdependence that might remain obscured in conventional descriptive statistics.

The analysis revealed several clusters of factors exhibiting significant correlations. For example, variables associated with insufficient initial project planning displayed a strong correlation with delays in material procurement and equipment availability. This finding emphasizes the direct impact of inadequate preparation during the planning phase on the operational efficiency of projects in the field. Furthermore, a strong association was identified between poor project management quality and the high frequency of design modifications (scope changes) during project implementation. This underscores that ineffective management practices fail to ensure adherence to initial plans, necessitating repeated design revisions, which ultimately result in delays and cost overruns.

Additionally, factors often regarded as administrative or external, such as delays in payments to subcontractors and uncertainties in government permitting processes, were also found to exhibit strong correlations. Payment delays disrupt project workflows, as subcontractors and suppliers may hesitate to continue work or deliver materials. When such delays occur concurrently with administrative uncertainties, such as unresolved permitting issues, projects encounter compounded setbacks in terms of both time and budget. This phenomenon highlights the interplay between internal and external factors, which collectively exacerbate adverse project outcomes.

Environmental factors, including adverse weather conditions and extreme geographical terrains, demonstrated moderate correlations with logistical delays and adjustments to construction schedules. Although such external variables remain outside managerial control, they demand rigorous

consideration during the planning and execution phases. For instance, projects located in regions characterized by prolonged rainy seasons or challenging terrains are especially vulnerable to logistical disruptions, potentially causing delays in structural tasks or material deliveries.

The identification of these correlation patterns is crucial during the initial stages of project planning. By comprehensively understanding the interdependencies among various factors, project managers can formulate systematic and integrative risk mitigation strategies. These measures may include establishing realistic cost and time contingencies for high-risk scenarios and implementing adaptive progress monitoring mechanisms to accommodate on-site changes. Furthermore, proactive approaches, such as risk-based scheduling and integrated cost control frameworks, can be employed to mitigate the impacts of highly correlated variables. Consequently, correlation analysis extends beyond its utility as a diagnostic tool for addressing issues in ongoing or completed projects. It also functions as a preventive framework, informing strategic decision-making processes in subsequent projects and thereby enhancing overall project performance.

Figure 2 presents a correlation heatmap derived from the dataset. The heatmap utilizes red hues to represent strong positive correlations, while blue hues denote negative or weak correlations. A positive correlation signifies that an increase in one variable is likely to be accompanied by an increase in the other. Conversely, a negative correlation indicates an inverse relationship between the two variables. The analysis highlights several significant relationships, including the positive correlation between Q03 (initial project planning) and Q10 (material availability). This finding underscores the critical importance of meticulous planning in ensuring logistical readiness. Furthermore, variables with high correlation values can serve as key indicators for identifying areas

requiring targeted risk mitigation measures. By leveraging these insights, project managers are better equipped to make data-driven decisions that significantly enhance the efficiency and effectiveness of construction project execution.

Figure 2. Correlation Heatmap of Variables in the Dataset

Feature Importance Random Forest

To elucidate the relative contribution of various factors to the prediction of cost overruns, a feature importance analysis was conducted employing the Random Forest algorithm. The Random Forest approach was selected due to its demonstrated capacity to handle high-dimensional datasets effectively and its inherent robustness against overfitting. This analysis provides a systematic framework for identifying the most critical variables influencing the model's predictive outcomes, thereby enabling stakeholders to direct their attention and resources toward the factors with the greatest potential impact.

Figure 3 delineates the ten most influential features identified by the Random Forest model. Topping the list is Q0 (Contractor Experience), which signifies the preeminent role of contractor expertise in determining project outcomes and mitigating the risk of cost overruns. Following this, Q43 (Financial Management) and Q22 (Material Availability) further underscore the critical importance of sound financial oversight and a resilient supply chain. Additional key contributors include Q5 (Quality of the Management Team) and Q14 (Project Plan Consistency), which highlight the necessity of well-coordinated team efforts and meticulously crafted project plans. In addition, Q24 (Field Supervision) is identified as a pivotal factor in facilitating the early detection of on-site issues, thereby minimizing disruptions and preventing potential delays. These findings offer valuable insights for project managers, who can leverage

this information to implement evidence-based strategies, prioritize high-impact factors, and optimize resource allocation. Such strategic measures are essential for enhancing both the operational efficiency and overall effectiveness of construction project management.

Figure 3. Top 10 Most Influential Features Identified by the Random Forest Algorithm

In addition to employing the feature importance technique from the Random Forest algorithm, a visual analysis was conducted to gain a deeper understanding of the direct relationships between key variables identified as primary determinants of cost overruns. One such feature with a significant contribution is Q43 (Financial Management). To substantiate this finding, a scatter plot was utilized to examine the relationship pattern between financial management scores and the frequency of cost overruns (Q50). This visualization offers an exploratory understanding of the direction and strength of the relationship between these two variables in the context of the road construction project survey.

Figure 4 below illustrates the relationship between Q43 (Financial Management) and Q50 (Cost Overrun) on a Likert scale. The observed pattern indicates that poor financial management (low Q43 scores) tends to correlate with higher frequencies of cost overruns (high Q50 scores). Conversely, projects with sound financial management (high Q43 scores) are more frequently associated with lower cost overrun scores, suggesting more effective cost control. This relationship underscores the critical importance of robust financial planning in mitigating the risk of cost escalation in construction projects. Such an interpretation provides actionable guidance for project

managers to prioritize strengthening financial management as a risk mitigation strategy against cost overruns.

Figure 4. Scatter Plot Q43 vs Q50

Experienced contractors (Q0) possess a superior ability to anticipate project risks and manage resources efficiently, thereby reducing the likelihood of cost overruns. Their expertise enables them to proactively address uncertainties and adapt to unforeseen changes during project execution. This adaptability is particularly critical in complex construction environments where external disruptions or deviations from the initial plan frequently occur. Contractors with extensive experience are more adept at implementing preventive measures and maintaining operational stability, ensuring smoother project progress and more effective cost management.

financial[10] management (Q43) often leads to delays in payments to vendors and laborers, ultimately disrupting the overall project workflow. Such disruptions can trigger a cascading effect, resulting in operational inefficiencies and even damaging the project's reputation among stakeholders. Conversely, effective financial management ensures timely payments, sustains workforce morale, and fosters trust with suppliers. A well-structured financial strategy also facilitates optimal resource allocation, thereby minimizing risks associated with cash flow issues and ensuring the project remains on schedule.

Disruptions in material (Q22) procurement directly impact project schedules, incurring additional costs due to idle time and the need for resource re-mobilization. Ensuring material availability is vital for maintaining steady workflows on-site. This requires proactive supply chain management and contingency planning to address potential delays in delivery or price fluctuations. Reliable material availability not only enhances productivity but

also mitigates the risk of schedule overruns, thereby contributing to the stability and success of the project.

Features such as Q5 (Quality of the Management Team), Q14 (Project Plan Consistency), and Q24 (Field Supervision) play a critical role in ensuring project success. A competent management team ensures smooth coordination across divisions, resolves challenges efficiently, and maintains clear communication channels, thereby reducing the likelihood of misunderstandings and enhancing operational synchronization. Consistency in the initial project plan minimizes the need for scope changes, which are often a primary source of delays and cost escalations. Additionally, intensive field supervision enables the early identification of potential issues, allowing corrective actions to be implemented before problems escalate, thus maintaining project momentum.

In conclusion, the analysis highlights that the success of construction projects is not dependent on a single factor but rather on the synergy of multiple interrelated elements. By prioritizing these key features, project managers can implement more effective risk mitigation strategies, optimize resource utilization, and enhance the overall efficiency and effectiveness of project management.

Evaluasi Performa Model Machine Learning

Evaluasi An evaluation was conducted on five machine learning models to assess their predictive capability in forecasting project delays based on the survey dataset. These models were selected to represent a diverse range of approaches, including linear models, instance-based learning, margin-based algorithms, ensemble methods, and neural networks. The aim of this evaluation was[11] to compare the strengths and limitations of each model type in handling the complexity of construction project datasets and to identify the model most suited for accurate predictions. Table 1 presents the performance metrics of

the five machine learning models tested in predicting project delays. Among these, the Random Forest model demonstrated the highest performance, achieving an accuracy of 72.1%, which reflects its ability to capture intricate patterns in the data. Logistic Regression ranked second with an accuracy of 62.5%, showcasing the effectiveness of a simpler, linear approach under specific circumstances. The Multi-Layer Perceptron (MLP), a neural network model, achieved an accuracy of 58.2%, suggesting moderate effectiveness but potential limitations in capturing highly nonlinear relationships without additional optimization.The[12] superior performance of the Random Forest model can be attributed to its ensemble learning approach, which combines multiple decision trees to enhance predictive accuracy and robustness.

This capability allows it to effectively model complex relationships within the dataset, even in the presence of noise or missing values. Conversely, models such as K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) exhibited lower performance levels, potentially due to their sensitivity to high-dimensional features and the imbalance present in the dataset's class distribution. KNN, for instance, may have struggled with overfitting in regions of sparse data, while SVM's reliance on finding optimal margins could be less effective in datasets with overlapping features or skewed distributions. This evaluation underscores the importance of selecting the appropriate model for the task at hand. The diverse performance outcomes of these models indicate that no single approach is universally optimal for all datasets. Factors such as the nature of the data, the distribution of classes, and the complexity of relationships between variables must be carefully considered. For the context of construction project analysis, the Random Forest model emerges as a strong candidate due to its balance of interpretability and predictive power. Future improvements, such as feature engineering or advanced preprocessing

techniques, could further enhance model performance and generalizability. By systematically evaluating multiple model types, this study provides valuable insights into the trade-offs involved in applying machine learning to construction project analysis. The findings highlight the critical role of model selection in achieving accurate and actionable predictions, which can support better decision-making in managing project timelines and mitigating risks.

Table 1. Comparative Performance Metrics of Machine Learning Models

Model

Accuracy

Precision

Recall

F1-Score

Logistic Regression

0.625

0.352

0.425

0.376

K-Nearest Neighbors

0.348

0.221

0.225

0.224

Support Vector Machine

0.443

0.280

0.200

0.128

Random Forest

0.721

0.469

0.467

0.444

Multi-Layer Perceptron

0.582

0.313

0.328

0.327

.

The performance of five machine learning models was analyzed to evaluate their effectiveness in predicting cost overruns in road construction projects. Each model exhibited unique strengths and weaknesses, reflecting its suitability for different data characteristics and problem complexities. The following provides a detailed examination of each model's performance:

a) Random Forest demonstrated the highest performance among all models, primarily due to its ability to capture complex relationships between variables and its robustness in handling imbalanced data. As an ensemble method, Random Forest builds multiple decision trees and aggregates their predictions, thereby reducing the risk of overfitting and increasing generalizability. This feature makes it particularly effective in identifying intricate patterns within the dataset, even when faced with noise or missing values.

b) Logistic Regression performed competitively, indicating that linear relationships between factors remain relevant in predicting cost overruns, particularly in cases where the interactions between variables are not overly complex. Despite its simplicity, Logistic Regression offers interpretability and

efficiency, making it a reliable choice for datasets with well-defined linear relationships. However, its limitations become apparent when the underlying data structure deviates significantly from linearity.

c) Multi-Layer Perceptron (MLP) showcased potential as a predictive model but faced challenges related to overfitting, which were exacerbated by the limited size of the dataset. Neural networks like MLP typically require large amounts of data to learn effectively and generalize well. Without sufficient data, the model may capture noise rather than meaningful patterns, resulting in suboptimal performance. Despite these challenges, MLP's ability to model nonlinear relationships highlights its potential for future applications with larger datasets and further tuning.

d) K-Nearest Neighbors (KNN) exhibited poor performance, largely due to its reliance on data proximity and sensitivity to noise within the survey dataset. KNN predictions are heavily influenced by the nearest neighbors in the feature space, making it particularly susceptible to inaccuracies in datasets with high variability. Additionally, KNN suffers from the "curse of dimensionality," where the model's effectiveness decreases as the number of features increases, leading to reduced predictive accuracy in high-dimensional data.

e) Support Vector Machine (SVM) underperformed, primarily due to its difficulty in handling nonlinear data with a large number of features without employing complex kernels. While SVM is known for its strength in binary classification and its ability to find optimal decision boundaries, the model's reliance on kernel functions increases computational complexity and requires careful parameter tuning. This process is time-consuming and resource-intensive, which may limit its applicability in practical settings with extensive datasets or computational constraints.

These findings emphasize the critical importance of selecting a machine learning[13] model that aligns with the specific characteristics of the dataset. Each model has inherent strengths and limitations, and understanding these nuances is essential for developing an effective predictive strategy. The Random Forest model emerged as the most suitable choice for this dataset, given its ability to manage complexity and imbalanced data effectively. In addition, the results highlight the necessity of rigorous cross-validation to avoid evaluation bias and ensure the reliability of performance metrics. This step is particularly important when working with datasets that exhibit high variability or imbalance. Moreover, the insights gained from analyzing model performance can guide the development of hybrid approaches or ensemble methods that combine the strengths of multiple models to achieve improved predictive accuracy. Ultimately, a thorough understanding of the strengths and weaknesses of each model, coupled with appropriate validation techniques, provides a foundation for making informed decisions about predictive strategies. This approach ensures that machine learning tools are leveraged to their fullest potential, contributing to more accurate and actionable predictions in the context of construction project management.

Analysis of Training and Validation Curves for the MLP Model

To further evaluate the MLP model, an analysis of training and validation accuracy and loss was conducted. Figure 5 illustrates a notable gap between training and validation accuracy, indicating potential overfitting. During training, accuracy steadily increases with the number of epochs, whereas validation accuracy fluctuates and does not show consistent improvement. Initially, the gap between training and validation accuracy is small, but it widens as training progresses, signaling that the model adapts too closely to the training data without effectively generalizing to unseen data. This pattern,

where training accuracy continues to rise while validation accuracy stagnates or declines, is a classic indicator of overfitting. This result highlights the need for adjustments, such as regularization techniques, dropout layers, or early stopping, to improve the model's generalization ability. These measures are essential to ensure that the MLP model performs reliably when applied to new datasets, particularly in critical predictive tasks.

Figure 5. Training and Validation Accuracy of the MLP Model

Figure 6 provides further evidence of overfitting in the MLP model by illustrating discrepancies between training loss and validation loss trends. While the training loss decreases steadily, validation loss does not exhibit a consistent downward trend. In some instances, validation loss even increases, reinforcing the indication that the model is losing its ability to generalize to unseen data.The steady reduction in training loss suggests that the model is effectively optimizing the network weights to minimize error on the training dataset. However, the inconsistent behavior of validation loss indicates that the model's complexity may be excessive relative to the dataset size, or that variations in the validation data are not adequately represented in the training data. Ideally, both training and validation loss should decline simultaneously. The divergence observed in Figure 6 suggests the model is "memorizing" noise in the training data rather than learning generalizable patterns. This analysis highlights the need to refine the model or adjust the training approach, such as simplifying the network architecture, using regularization techniques, or increasing the diversity of the training data to better capture the distribution of the validation set. These steps are critical to improving the model's generalization capability and ensuring its practical applicability.

Figure 6. Training Loss Curve of the MLP Model

The analysis of accuracy and loss curves for the Multi-Layer Perceptron (MLP) model provides valuable insights into the model's behavior during training. As depicted in Figure 5, the training accuracy steadily increases with the number of epochs, reflecting the model's ability to optimize its weights and minimize error[15] on the training dataset. However, validation accuracy fails to exhibit a comparable upward trend, with fluctuations and occasional declines observed as training progresses. This pattern is a classic indication of overfitting, where the model becomes overly tailored to the training data and loses its ability to generalize to unseen data. A similar phenomenon is evident in Figure 6, which shows a steady decline in training loss, while validation loss displays inconsistent fluctuations or even increases after certain epochs. This divergence suggests that the model is learning noise or specific patterns unique to the training data rather than capturing generalizable structures within the dataset. Such behavior undermines the model's predictive reliability and limits its practical applicability.

The overfitting observed in the MLP model can be attributed to several key factors that limit its ability to generalize effectively. One primary contributor is the limited size of the dataset, which provides a small number of observations for training. Neural networks like MLP inherently require large datasets to balance their architectural complexity and learn meaningful patterns. When the dataset is small, the model tends to memorize specific data points rather than identifying generalizable structures, making it prone to overfitting. Another significant factor is the complexity of the MLP architecture used in this study. The model consists of two hidden layers with 64 and 32 neurons, respectively, which is relatively complex given the limited size of the dataset. This level of

complexity enables the model to capture intricate patterns in the training data, but it also increases the risk of learning non-generalizable patterns or noise. As a result, the model becomes highly specific to the training data and struggles to perform well on new, unseen data.

Additionally, the lack of regularization techniques exacerbates the issue of overfitting. Regularization methods, such as L2 regularization (Ridge), dropout, or batch normalization, are designed to constrain the model's learning process and prevent neurons from co-adapting excessively to the training data. Without these techniques, the MLP model becomes overly flexible, adapting too closely to the training data at the expense of its ability to generalize. This lack of regularization leaves the model vulnerable to overfitting, further diminishing its predictive accuracy on validation data. Addressing these factors through appropriate adjustments to the dataset size, model architecture, and regularization techniques is essential for improving the generalization capability of the MLP model. These refinements will enable the model to achieve better[16] balance between complexity and generalizability, ensuring its reliability in practical applications.

To address the overfitting issues identified in the MLP model, several strategies can be employed to enhance its generalization capability. One effective approach is adjusting the model architecture by reducing the number of neurons or hidden layers, ensuring the model's complexity aligns with the dataset size. Simplifying the architecture prevents the model from being overly complex, thereby minimizing the risk of overfitting. Additionally, the application of regularization techniques such as L1 or L2 regularization can penalize excessively large weights, while dropout can randomly deactivate neurons during training to prevent co-adaptation. Similarly, batch normalization can stabilize learning dynamics, promoting a more robust training process.

Another potential strategy is data augmentation, which involves enriching the dataset through synthetic generation or re-sampling. Although more challenging for survey data compared to visual data, this approach can increase data diversity and improve the model's ability to generalize. Early stopping is another practical method, where training is halted as soon as validation accuracy begins to decline. This prevents the model from continuing to adapt to noise within the training data, thus avoiding further overfitting. Finally, cross-validation can be implemented by dividing the dataset into multiple folds to evaluate model performance across different subsets. This ensures that the model's performance is consistent and reduces the likelihood of biased evaluation results.

While the MLP model demonstrates potential[17] for modeling nonlinear relationships between variables, it faces significant limitations when applied to a small dataset such as the one in this study. Alternative models like Random Forest, which are more robust and less sensitive to overfitting, may be better suited to datasets of this scale. However, as data availability increases in the future, neural networks like MLP can become more viable, particularly when paired with advanced tuning and validation techniques. Such improvements could enable MLP to achieve superior predictive performance, especially in capturing complex interactions between variables. This analysis underscores the importance of tailoring model complexity to dataset characteristics and employing robust validation frameworks to ensure reliable and generalizable predictions.

Conclusion

This study proposes a machine-learning-based predictive approach to classify the frequency of cost overruns in road construction projects. Using a structured survey dataset containing 139 observations and 44 independent variables, the

developed model can identify potential cost delays more accurately than conventional methods. Before feature selection, the Random Forest model achieved the best classification accuracy at 72.1 percent. After applying the SelectFromModel technique, this accuracy rose to 85 percent, highlighting the value of targeted feature reduction. Key contributors to cost overruns include contractor experience (Q0), material availability (Q22), and financial management (Q43).

The preprocessing phase—comprising label encoding and careful handling of missing values—ensured that the data were suitable for model training. These results demonstrate that a data-driven approach using machine-learning algorithms can serve as an effective tool during the early stages of project planning, especially for anticipating the risk of cost delays. Project stakeholders are therefore encouraged to integrate predictive models into their risk-evaluation[18] and decision-making processes. Moreover, developing a dashboard-based visualization system linked to the model would help users from diverse backgrounds understand and interpret the predictions more easily. Ultimately, such integration can improve strategic planning and resource allocation across the entire project lifecycle[19]

Declarations

Author Contributions

Conceptualization: A.P., S.W.M., N.M.S., and P.R.A.; Methodology: A.P., S.W.M., N.M.S., and P.R.A.; Software: A.P.; Validation: A.P., S.W.M., N.M.S., and P.R.A.; Formal Analysis: A.P., S.W.M., N.M.S., and P.R.A.; Investigation: N.M.S.; Resources: S.W.M.; Data Curation: P.R.A.; Writing – Original Draft Preparation: A.P., S.W.M., N.M.S., and P.R.A.; Writing – Review and Editing: A.P., S.W.M.; Visualization: N.M.S. and P.R.A. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available upon request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Conflict of Interest Statement

The authors declare that they have no competing financial interests or personal relationships that could have influenced the work reported in this paper.

References

[1] C. Jiang, X. Li, J.-R. Lin, M. Liu, and Z. Ma, "Adaptive control of resource flow to optimize construction work and cash flow via online deep reinforcement learning," Autom. Constr., vol. 150, p. 104817, Jun. 2023, doi: 10.1016/j.autcon.2023.104817.

[2] A. M. Abdelalim, M. Salem, M. Salem, M. Al-Adwani, and M. Tantawy, "Analyzing Cost Overrun Risks in Construction Projects: A Multi-Stakeholder Perspective Using Fuzzy Group Decision-Making and K-Means Clustering," Buildings, vol. 15, no. 3, p. 447, Jan. 2025, doi: 10.3390/buildings15030447.

[3] C. N. Egwim, H. Alaka, L. O. Toriola-Coker, H. Balogun, and F. Sunmola, "Applied artificial intelligence for predicting construction projects delay," Mach. Learn. Appl., vol. 6, p. 100166, Dec. 2021, doi: 10.1016/j.mlwa.2021.100166.

[4] Z. Zheng, L. Zhou, H. Wu, and L. Zhou, "Construction cost prediction system based on Random Forest optimized by the Bird Swarm Algorithm," Math. Biosci.

Eng., vol. 20, no. 8, pp. 15044–15074, 2023, doi: 10.3934/mbe.2023674.
[5] M. Shayboun, D. Kifokeris, and C. Koch, "CONSTRUCTION PLANNING WITH MACHINE LEARNING".
[6] S. Tayefeh Hashemi, O. M. Ebadati, and H. Kaur, "Cost estimation and prediction in construction projects: a systematic review on machine learning techniques," SN Appl. Sci., vol. 2, no. 10, p. 1703, Oct. 2020, doi: 10.1007/s42452-020-03497-1.
[7] M. A. Ashtari, R. Ansari, E. Hassannayebi, and J. Jeong, "Cost Overrun Risk Assessment and Prediction in Construction Projects: A Bayesian Network Classifier Approach," Buildings, vol. 12, no. 10, p. 1660, Oct. 2022, doi: 10.3390/buildings12101660.
[8] A. Abbasi and A. Jaafari, "Evolution of Project Management as a Scientific Discipline," Data Inf. Manag., vol. 2, no. 2, pp. 91–102, Sep. 2018, doi: 10.2478/dim-2018-0010.
[9] R. Al Mnaseer, S. Al-Smadi, and H. Al-Bdour, "Machine learning-aided time and cost overrun prediction in construction projects: application of artificial neural network," Asian J. Civ. Eng., vol. 24, no. 7, pp. 2583–2593, Nov. 2023, doi: 10.1007/s42107-023-00665-7.
[10] E. Plebankiewicz, "Model of Predicting Cost Overrun in Construction Projects," Sustainability, vol. 10, no. 12, p. 4387, Nov. 2018, doi: 10.3390/su10124387.
[11] A. Arabiat, H. Al-Bdour, and M. Bisharah, "Predicting the construction projects time and cost overruns using K-nearest neighbor and artificial neural network: a case study from Jordan," Asian J. Civ. Eng., vol. 24, no. 7, pp. 2405–2414, Nov. 2023, doi: 10.1007/s42107-023-00649-7.
[12] Z. M. Yaseen, Z. H. Ali, S. Q. Salih, and N. Al-Ansari, "Prediction of Risk Delay in Construction Projects Using a Hybrid Artificial Intelligence Model,"

Sustainability, vol. 12, no. 4, p. 1514, Feb. 2020, doi: 10.3390/su12041514.

[13] 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.

[14] D. Spikol, E. Ruffaldi, G. Dabisias, and M. Cukurova, "Supervised machine learning in multimodal learning analytics for estimating success in project-based learning," J. Comput. Assist. Learn., vol. 34, no. 4, pp. 366–377, Aug. 2018, doi: 10.1111/jcal.12263.

[15] K. Piwowar-Sulej, M. Sołtysik, S. Jarosz, and R. Pukała, "The Linkage between Renewable Energy and Project Management: What Do We Already Know, and What Are the Future Directions of Research?," Energies, vol. 16, no. 12, p. 4609, Jun. 2023, doi: 10.3390/en16124609.

[16] G. H. Coffie and S. K. F. Cudjoe, "Toward predictive modelling[20] of construction cost overruns using support vector machine techniques," Cogent Eng., vol. 10, no. 2, p. 2269656, Dec. 2023, doi: 10.1080/23311916.2023.2269656.

[17] Theingi Aung, S. R. Liana, A. Htet, and Amiya Bhaumik, "Using Machine Learning to Predict Cost Overruns in Construction Projects," J. Technol. Innov. Energy, vol. 2, no. 2, pp. 1–7, Jun. 2023, doi: 10.56556/jtie.v2i2.511.

[18] Theingi Aung, S. R. Liana, A. Htet, and Amiya Bhaumik, "Using Machine Learning to Predict Cost Overruns in Construction Projects," J. Technol. Innov. Energy, vol. 2, no. 2, pp. 1–7, Jun. 2023, doi: 10.56556/jtie.v2i2.511.

[19] S. E. Boyacioglu, D. Greenwood, and K. Rogage, "Incorporating Emerging Technologies in the Forensic Analysis of Construction Project Delays," IOP Conf. Ser. Earth Environ. Sci., vol. 1101, no. 5, p. 052029, Nov. 2022, doi: 10.1088/1755-1315/1101/5/052029.

[20] S.-S. Leu, Y. Liu, and P.-L. Wu, "Project Cost Overrun Risk Prediction Using Hidden Markov Chain Analysis," Buildings, vol. 13, no. 3, p. 667, Mar. 2023, doi: 10.3390/buildings13030667.
[21] 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.
[22] J. Singh et al., "Enhancing Large-Diameter Tunnel Construction Safety with Robust Optimization and Machine Learning Integrated into BIM," Open Civ. Eng. J., vol. 18, no. 1, p. e18741495343680, Oct. 2024, doi: 10.2174/0118741495343680240911053413.
[23] E. Ogbeifun and J.-H. C. Pretorius, "Investigation of factors responsible for delays in the execution of adequately funded construction projects," Eng. Manag. Prod. Serv., vol. 14, no. 1, pp. 93–102, Mar. 2022, doi: 10.2478/emj-2022-0008.
[24] M. A. V. Sampath, G. Abeysooriya, and E. A. C. Piyashantha, "A Case Study on Factors Contributing to Time Delays of Government-Funded Affordable Housing Projects: ABC Residencies as a Case Study in Colombo Sri Lanka," J. Real Estate Stud., vol. 21, no. 2, Jun. 2024, doi: 10.31357/jres.v21i2.7421.
[25] M. Meharie and N. Shaik, "Predicting Highway Construction Costs: Comparison of the Performance of Random Forest, Neural Network and Support Vector Machine Models," J. Soft Comput. Civ. Eng., vol. 4, no. 2, Apr. 2020, doi: 10.22115/scce.2020.226883.1205.
[26] Professor and Head of Department of Civil Engineering RIMT University Punjab India., A. Bashir, S. Singla, Assistant Professor, Department of Civil Engineering, RIMT University, Punjab, India, M. Kaushal, and Department of Civil Engineering, RIMT University, Punjab, India., "Inquisition on Cost & Time

Overrun in Road Construction Projects in Kashmir," Int. J. Eng. Adv. Technol., vol. 9, no. 2, pp. 2956–2965, Dec. 2019, doi: 10.35940/ijeat.B3880.129219.

[27] M. Enrica, H. H. Purba, and A. Purba, "Risks Leading to Cost Overrun in Construction Projects: A Systematic Literature Review".

[28] Y. Wang, S. Ma, J. Wang, X. Jiao, H. Wang, and W. Pu, "Application of project management system in construction machinery enterprises," BCP Bus. Manag., vol. 48, pp. 236–239, Jul. 2023, doi: 10.54691/bcpbm.v48i.5272.

[29] F. Tarhuni and R. Mahat, "The frequency and significance of the primary cause of cost overruns in infrastructure projects," J. Innov. Transp., vol. 5, no. 2, pp. 40–46, Dec. 2024, doi: 10.53635/jit.1293629.

[30] Z. Zheng, L. Zhou, H. Wu, and L. Zhou, "Construction cost prediction system based on Random Forest optimized by the Bird Swarm Algorithm," Math. Biosci. Eng., vol. 20, no. 8, pp. 15044–15074, 2023, doi: 10.3934/mbe.2023674.

[31] S. J. Schuldt, M. R. Nicholson, Y. A. Adams, and J. D. Delorit, "Weather-Related Construction Delays in a Changing Climate: A Systematic State-of-the-Art Review," Sustainability, vol. 13, no. 5, p. 2861, Mar. 2021, doi: 10.3390/su13052861.

[32] S.-S. Leu, C.-Y. Lu, and P.-L. Wu, "Dynamic-Bayesian-Network-Based Project Cost Overrun Prediction Model," Sustainability, vol. 15, no. 5, p. 4570, Mar. 2023, doi: 10.3390/su15054570.

[33] E. Halabaku and E. Bytyçi, "Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests," Intell. Autom. Soft Comput., vol. 39, no. 6, pp. 987–1006, 2024, doi: 10.32604/iasc.2024.059429.

[34] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 6, 2023, doi: 10.14569/IJACSA.2023.01406116.

[35] N. F. Maulanisa, E. Chumaidiyah, and I. K. Sriwana, "A system dynamics modeling approach for improving engineering, procurement, and construction project performance: A case study," J. Infrastruct. Policy Dev., vol. 8, no. 11, p. 8730, Oct. 2024, doi: 10.24294/jipd.v8i11.8730.

[36] A. M. Mohammed, "Scalable AI: Leveraging Cloud Infrastructure for Large-Scale Machine Learning," Int. J. Adv. Eng. Manag., vol. 7, no. 2, pp. 382–390, Feb. 2025, doi: 10.35629/5252-0702382390.

[37] N. Bobrova, A. Ivanov, D. Kamenskikh, and L. Plyusnina, "Management of the investment and construction project cost under conditions of risk and uncertainty," SHS Web Conf., vol. 116, p. 00049, 2021, doi: 10.1051/shsconf[22] /202111600049.

[38] R. Susanti and A. Nurdiana, "Cost Overrun in Construction Projects in Indonesia," IOP Conf. Ser. Earth Environ. Sci., vol. 506, no. 1, p. 012039, May 2020, doi: 10.1088/1755-1315/506/1/012039.

[39] S. Dong, M. Ahmed, and V. Chatpattananan, "Analysis of Key Factors of Cost Overrun in Construction Projects Based on Structural Equation Modeling," Sustainability, vol. 17, no. 5, p. 2119, Feb. 2025, doi: 10.3390/su17052119.

[40] C. Bura, A. K. Jonnalagadda, and P. Naayini, "The Role of Explainable AI (XAI) in Trust and Adoption," J. Artif. Intell. Gen. Sci. JAIGS ISSN3006-4023, vol. 7, no. 01, pp. 262–277, Feb. 2025, doi: 10.60087/jaigs.v7i01.331.

[41] M. Abdel-Monem, K. Alshaer, and K. El-Dash, "Assessing Risk Factors Affecting the Accuracy of Conceptual Cost Estimation in the Middle East," Buildings, vol. 12, no. 7, p. 950, Jul. 2022, doi: 10.3390/buildings12070950.

[42] Y. Ke, J. Zhang, and S. P. Philbin, "Tradition and Innovation in Construction Project Management," Buildings, vol. 13, no. 6, p. 1537, Jun. 2023, doi: 10.3390/buildings13061537.

[43] E. Zaneldin and W. Ahmed, "A Generic Framework for Managing Schedule and Cost Risks of Construction Activities Using PERT and the EV Technique," Buildings, vol. 14, no. 7, p. 1918, Jun. 2024, doi: 10.3390/buildings14071918.

[44] A. Kalnawat, D. Dhabliya, K. Vydehi, A. Dhablia, and S. D. Kumar, "Safeguarding Critical Infrastructures: Machine Learning in Cybersecurity," E3S Web Conf., vol. 491, p. 02025, 2024, doi: 10.1051/e3sconf/202449102025.

1.	~~open-access~~ → open-access	Misspelled words	Correctness
2.	~~while~~ → While	Improper formatting	Correctness
3.	~~own~~	Wordy sentences	Clarity
4.	~~real-time~~ → real-time	Misspelled words	Correctness
5.	~~real-time~~ → real-time	Misspelled words	Correctness
6.	a numerical	Determiner use (a/an/the/this, etc.)	Correctness
7.	predictions.	Closing punctuation	Correctness
8.	~~in relation to~~ → about, to, with, concerning	Wordy sentences	Clarity
9.	improving,	Comma misuse within clauses	Correctness
10.	~~financial~~ → Financial	Improper formatting	Correctness
11.	This evaluation aimed	Wordy sentences	Clarity
12.	. The	Improper formatting	Correctness
13.	machine-learning	Misspelled words	Correctness
14.	. The	Improper formatting	Correctness
15.	~~error~~ → errors	Incorrect noun number	Correctness
16.	a better	Determiner use (a/an/the/this, etc.)	Correctness
17.	the potential	Determiner use (a/an/the/this, etc.)	Correctness
18.	~~risk-evaluation~~ → risk evaluation	Confused words	Correctness

19.	lifecycle.	Closing punctuation	Correctness
20.	~~modelling~~ → modeling	Mixed dialects of English	Correctness
21.	*bcpbm*	Unknown words	Correctness
22.	*shsconf*	Unknown words	Correctness