Frontiers in Computer Science and Artificial Intelligence
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

AI-Based Schedule Overrun Prediction in Technical Projects Using Gradient Boosting Machine for Decision Support

Mr. Abdelkarim Ramzy Sweilem[1]*, Ms. Hala Jamil Abdoh[2], and Dr. Ahmad Jamil Abdoh[3]
[1] *Independent Researcher, Riyadh, Saudi Arabia*
[2] *Independent Researcher, Riyadh, Saudi Arabia*
[3] *3IQ Academy, Amman, Jordan*
Corresponding Author: Mr. Abdelkarim Ramzy Sweilem**, E-mail:** aboodsweilem2@gmail.com

| ABSTRACT
The increasing complexity of technical projects has challenged the effectiveness of traditional project management approaches, particularly in accurately predicting schedule performance. Conventional methods often rely on expert judgment and static planning techniques, limiting their ability to capture dynamic project conditions and complex relationships between variables. This study investigates the role of Artificial Intelligence, specifically the Gradient Boosting Machine (GBM), in enhancing schedule management through a data-driven approach. A predictive model was developed using a publicly available project management dataset obtained from Kaggle, consisting of 4517 project instances. The dataset includes key variables such as project duration, total cost, team size, risk factor, and client satisfaction. The proposed model was applied to predict Schedule Overrun (%) as a measure of project performance. The results demonstrate strong predictive performance, with an R^2 value of 0.835, RMSE of 7.128, and MAE of 5.946. These findings indicate that the model is capable of capturing complex patterns and relationships within project data. Feature importance analysis revealed that risk factor and project duration are the most influential variables affecting schedule performance, followed by client satisfaction. These results highlight the importance of managing project uncertainty and timeline characteristics to improve overall project outcomes. The study confirms that AI-based models can provide valuable predictive insights to support proactive, data-driven decision-making in project management. Future research may focus on expanding the dataset, incorporating additional contextual variables, and exploring advanced modeling techniques to further enhance predictive accuracy and practical applicability.

| KEYWORDS

Artificial Intelligence; Machine Learning; Project Management; Schedule Overrun; Gradient Boosting Machine; Predictive Modeling; Data-Driven Decision-Making

| ARTICLE INFORMATION

ACCEPTED: 01 April 2026 **PUBLISHED:** 24 May 2026 **DOI:** 10.32996/jcsts.2026.5.8.1

1. Introduction

Technical projects have experienced rapid growth in recent years, driven by continuous advancements in information and communication technologies and the increasing reliance of organizations on digital solutions. These projects are inherently complex, as they involve the integration of technical and organizational components, the coordination of multiple stakeholders, and frequent changes in requirements throughout the project lifecycle. As their scale and importance continue to expand, effective project management has become a critical factor in achieving organizational objectives and ensuring long-term sustainability [1].

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

However, modern technical projects operate in highly dynamic environments characterized by rapidly evolving requirements and complex interactions among resources, activities, and external factors [5,11]. Consequently, risks and schedule deviations often arise from nonlinear and interdependent causes that are difficult to capture using traditional static planning approaches [3,4]. These challenges frequently lead to project delays, inefficient resource utilization, and difficulties in maintaining planned schedules.

Traditional project management methods rely heavily on expert judgment, manual estimation, and deterministic planning techniques. While these approaches have been widely adopted, they often fail to effectively utilize historical project data and are limited in their ability to predict risks and schedule deviations in complex and dynamic environments. As a result, decision-making tends to be reactive rather than proactive [2,7].

In recent years, artificial intelligence (AI) and machine learning techniques have emerged as promising tools for enhancing predictive capabilities in project management [1,3]. These techniques enable the analysis of large volumes of historical data, allowing for the identification of hidden patterns, improved risk detection, and more accurate prediction of schedule deviations. Among these methods, the Gradient Boosting Machine (GBM) has gained particular attention due to its strong predictive performance and its ability to model complex nonlinear relationships within project data [3,12].

Artificial intelligence is increasingly recognized as a transformative force in project management, reshaping how projects are planned, executed, and controlled. Rather than replacing human expertise, AI acts as an enabling technology that enhances decision-making processes by providing data-driven insights. In this context, machine learning models support project managers in anticipating risks and improving scheduling decisions.

Despite these advancements, several limitations remain in the existing literature. Many studies focus on conceptual frameworks or perception-based analyses rather than applied, data-driven implementations suitable for real-world technical projects [3]. Furthermore, prior research often addresses risk prediction and schedule forecasting separately, without integrating both aspects into a unified framework. Limited attention has also been given to translating predictive outputs into actionable scheduling decisions, such as baseline adjustments, activity re-sequencing, and proactive delay mitigation strategies [4].

Therefore, a clear research gap exists in developing an applied, data-driven approach that integrates schedule prediction with schedule decision support within a unified framework tailored for technical project environments.

This study aims to address this gap by leveraging machine learning techniques to enhance both risk management and schedule performance in technical projects. Specifically, the study seeks to:

1. Analyze the role of artificial intelligence techniques in predicting project schedule performance in technical projects.
2. Investigate how machine learning models can be utilized to improve project scheduling and reduce deviations from planned timelines.
3. Develop an applied data-driven framework for schedule prediction and decision support in technical project environments.
4. Evaluate the effectiveness of the proposed approach in predicting project schedule overrun.

The remainder of this paper is organized as follows: Section 2 presents the background and reviews related work, Section 3 describes the methodology, Section 4 presents the results, Section 5 discusses the findings, Section 6 outlines the limitations of the study, and Section 7 concludes the paper.

This study contributes to the literature by providing a data-driven approach for predicting schedule overrun using a large real-world dataset. Unlike prior studies that focus on theoretical or small-scale datasets, this research applies a Gradient Boosting Machine (GBM) model to a large dataset of 4,517 project instances, achieving strong predictive performance. The study also integrates exploratory data analysis and feature importance to enhance model interpretability and practical applicability in project management.

2. Background and Related Work

2.1. Risk and Schedule Management in Project Management

Risk and schedule management are fundamental components of effective project management, particularly in technical and engineering projects characterized by high levels of uncertainty, complexity, and interdependencies among activities [1,6,7]. Risk management involves the systematic identification, analysis, and mitigation of potential events that may negatively impact project objectives, including cost, scope, quality, and schedule [2,8]. In parallel, schedule management focuses on planning, monitoring, and controlling project activities to ensure timely project completion and efficient resource utilization [6,9].

Traditional project management frameworks emphasize structured and process-driven approaches for managing risks and schedules. These approaches typically rely on expert judgment, predefined planning techniques, and historical experience [6,7]. Common practices include the development of risk registers, qualitative and quantitative risk assessments, and the application of scheduling techniques such as the Critical Path Method (CPM) and Program Evaluation and Review Technique (PERT) [9,10]. Due to their simplicity, interpretability, and standardization, these methods remain widely used in contemporary project management practices [7].

Despite their widespread adoption, traditional approaches exhibit several limitations when applied to modern technical projects. Such projects operate in highly dynamic environments characterized by rapidly evolving requirements and complex interactions among resources, activities, and external factors [5,11]. As a result, risks and schedule deviations often emerge from nonlinear and interdependent relationships that are difficult to model using static and deterministic planning methods [3,4].

These limitations highlight the need for more advanced analytical approaches capable of handling uncertainty, capturing complex relationships, and leveraging historical project data. Consequently, there has been increasing interest in data-driven and intelligent techniques that support proactive risk identification and more accurate schedule prediction throughout the project lifecycle [1,3,8].

2.2. Limitations of Traditional Risk and Schedule Management Approaches

Despite their widespread adoption, traditional risk and schedule management approaches exhibit several limitations when applied to complex and data-intensive projects [6,7]. One of the primary challenges lies in their heavy reliance on expert judgment and deterministic assumptions. Risk assessments are often based on subjective evaluations that may vary significantly among experts and may not accurately reflect real project dynamics [2,7]. Similarly, schedule estimates typically assume fixed activity durations, overlooking variability caused by resource constraints, external disruptions, and evolving project conditions [9,11].

Conventional scheduling techniques, such as the Critical Path Method (CPM) and Program Evaluation and Review Technique (PERT), provide structured frameworks for project planning; however, they have limited capability to handle uncertainty and adapt to changes during project execution [9,10]. While these methods are effective in establishing baseline schedules, they rarely incorporate real-time project data or learn from historical project outcomes [6,11]. As a result, schedule forecasts tend to become less accurate over time, leading to delayed responses to emerging risks and increasing the likelihood of cost and time overruns [7,11].

Another significant limitation is the inability of traditional approaches to efficiently process and analyze large volumes of historical project data. Modern technical projects generate extensive datasets related to cost performance, resource utilization, risk events, and schedule deviations [1,3]. However, conventional analytical methods lack the capability to fully exploit these data sources, resulting in underutilization of valuable information that could enhance prediction accuracy and decision support [3,4].

Collectively, these limitations highlight the need for more advanced, data-driven approaches capable of modeling complex relationships, managing uncertainty, and continuously improving predictive performance based on new data. Consequently, the integration of Artificial Intelligence (AI) and Machine Learning (ML) techniques has gained increasing attention as a promising solution to overcome the shortcomings of traditional risk and schedule management methods [1,3,4].

2.3. Application of Artificial Intelligence in Project Management

The application of Artificial Intelligence (AI) in project management has gained significant attention in recent years as organizations seek to improve decision-making, enhance predictive accuracy, and manage increasing project complexity [1,5]. AI techniques enable the analysis of large and heterogeneous datasets generated throughout the project lifecycle, including cost records, schedules, risk logs, and performance indicators. By leveraging these data sources, AI-based systems can support

project managers in identifying hidden patterns, forecasting project outcomes, and proactively responding to potential risks and schedule delays [1,3].

In the context of risk and schedule management, AI is primarily utilized as a decision-support tool rather than a replacement for traditional project management practices [5,6]. Early applications focused on expert systems and rule-based models designed to formalize human expertise through structured decision rules. While these approaches improved consistency, they remained limited in their ability to adapt to dynamic project environments and evolving risk conditions [6,7].

Recent advancements in AI have shifted the focus toward data-driven and learning-based approaches capable of modeling complex and nonlinear relationships within project data [1,3]. In particular, predictive analytics techniques have been widely applied to estimate project delays, assess risk exposure, and evaluate alternative mitigation strategies under uncertainty [3,4]. These models can continuously update their predictions as new data become available, enabling more adaptive and responsive project control compared to traditional static planning approaches [1,5].

Despite these advantages, the adoption of AI in project management is associated with several challenges. These include issues related to data availability, data quality, and the integration of AI solutions with existing project management systems [1,3]. Furthermore, organizational resistance and concerns regarding the transparency and interpretability of AI-driven decisions may hinder widespread adoption, particularly in management-oriented environments where explainability is critical [5].

Overall, the literature suggests that AI functions as an enabling technology that enhances traditional project management practices by providing advanced predictive and analytical capabilities, rather than fully automating decision-making processes [1,5]. This perspective has driven the increasing adoption of Machine Learning (ML) techniques, which represent a practical and scalable subset of AI, particularly well-suited for risk prediction and schedule forecasting in complex technical projects [1,3,4].

2.4. Machine Learning Models for Risk and Schedule Prediction

Building on recent advancements in Artificial Intelligence, Machine Learning (ML) techniques have been widely adopted as effective tools for predicting risks and schedule performance in project management [1,3,4]. Unlike traditional statistical methods, ML models are capable of learning complex patterns and relationships directly from historical data without relying on predefined assumptions. This capability makes ML particularly suitable for technical projects, where uncertainty, nonlinear interactions, and multiple influencing factors significantly affect project outcomes [3,5].

Regression-based models have been extensively used to estimate project duration, cost overruns, and schedule delays [3,4]. Both linear and nonlinear regression approaches provide a baseline for understanding the relationships between input variables—such as resource allocation, task dependencies, and risk indicators—and project outcomes. However, their performance tends to degrade when data exhibit strong nonlinearities or high variability, which are common characteristics of complex technical projects [3,4].

Decision tree–based models have gained increasing attention due to their interpretability and ability to handle both numerical and categorical data [3,4]. By recursively partitioning the data, decision trees identify the most influential factors contributing to risk occurrence and schedule deviations. Ensemble learning techniques, such as Random Forest and Gradient Boosting, extend this concept by combining multiple weak learners to improve predictive accuracy and robustness [3,4]. These methods have demonstrated strong performance in capturing complex interactions and handling noisy or heterogeneous datasets [3].

Support Vector Machines (SVM) and Artificial Neural Networks (ANN) have also been widely explored for risk classification and schedule prediction tasks [3,4]. SVM models are particularly effective in high-dimensional feature spaces, especially when the number of variables exceeds the number of observations. Neural networks, on the other hand, are well-suited for modeling highly nonlinear relationships and have been applied to predict project delays, cost escalation, and risk likelihood [4]. Despite their strong predictive capabilities, these models often face challenges related to interpretability and computational complexity, which may limit their practical applicability in project management contexts [5].

Among ensemble learning approaches, the Gradient Boosting Machine (GBM) has emerged as a particularly powerful technique due to its high predictive accuracy and flexibility [3,4,12]. GBM builds models iteratively by minimizing the errors of previous iterations, allowing it to focus on difficult-to-predict cases and improve overall model performance. Recent studies have shown that GBM-based models outperform traditional regression methods and single decision trees in predicting schedule delays and risk outcomes, especially when dealing with complex, imbalanced, or heterogeneous datasets [3,4].

Overall, the literature demonstrates that ML models significantly enhance risk and schedule prediction capabilities compared to conventional approaches [1,3,4]. However, selecting an appropriate model depends on factors such as data characteristics, prediction objectives, and the trade-off between predictive accuracy and interpretability [5]. Based on these considerations, this

study adopts the GBM model as the primary analytical approach, aiming to achieve high prediction accuracy while effectively capturing complex relationships within project data.

2.5. Data Sources and Dataset Types in AI-Based Risk and Schedule Management

The effectiveness of Artificial Intelligence (AI) and Machine Learning (ML) models in risk and schedule prediction is highly dependent on the availability, quality, and representativeness of project data [1,3]. In project management research, data sources typically include historical records from completed projects, such as schedules, cost reports, risk registers, change logs, and performance indicators. These datasets serve as the foundation for training, validating, and evaluating predictive models [3,4].

Real-world project datasets are considered the most valuable source of information, as they reflect actual project conditions, uncertainties, and decision-making processes [1,4]. Such data are commonly derived from construction, infrastructure, software development, and engineering projects. However, access to real-world datasets is often limited due to confidentiality constraints, data ownership issues, and the absence of standardized data collection practices across organizations [3,5]. As a result, many studies rely on restricted or incomplete datasets, which may affect model generalizability and robustness [3].

To mitigate data limitations, researchers frequently utilize publicly available datasets and secondary data sources [3,4]. These include open repositories, benchmark datasets, and datasets derived from published case studies. While such data sources enhance research reproducibility and accessibility, they may not fully capture the complexity and contextual variability of real technical projects, particularly in terms of dynamic risk interactions and evolving schedule conditions [1,4].

Another widely adopted approach is the use of synthetic datasets [3,4]. Synthetic data are artificially generated to simulate realistic project scenarios based on predefined assumptions, statistical distributions, or simulation techniques. In the context of risk and schedule management, synthetic datasets are particularly useful for modeling rare events, handling imbalanced data, and exploring hypothetical project conditions that are difficult to observe in real-world environments [3]. However, the reliability of results derived from synthetic data depends on how accurately the generated data reflect real project behavior and underlying relationships [3,4].

Project management datasets are inherently heterogeneous, consisting of numerical variables (e.g., activity durations and costs), categorical variables (e.g., risk categories and project types), and temporal data (e.g., timelines and milestone sequences) [3,4]. This diversity introduces challenges related to data preprocessing, feature selection, and dimensionality reduction. Consequently, data preparation techniques—such as normalization, encoding, and feature engineering—play a critical role in improving model performance and predictive accuracy [3,4].

Overall, the literature highlights that data availability and quality remain key challenges in AI-based project management research [1,3]. The selection between real-world, public, and synthetic datasets involves trade-offs between realism, accessibility, and experimental control. These considerations significantly influence model selection, evaluation strategies, and the reliability of risk and schedule predictions, ultimately affecting the effectiveness of AI-driven approaches in project management [1].

2.6. Evaluation Metrics for Risk and Schedule Prediction Models

The evaluation of Artificial Intelligence (AI) and Machine Learning (ML) models in project management requires the use of appropriate performance metrics aligned with the prediction objective [1,3]. In this study, the focus is on regression-based schedule prediction; therefore, regression evaluation metrics are employed to assess model performance [3,4].

Model performance is evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). MAE provides an intuitive measure of the average prediction error, while RMSE assigns greater weight to larger errors, making it suitable for evaluating significant schedule deviations [14]. The R^2 metric indicates the proportion of variance in the target variable explained by the model, providing an overall measure of predictive capability [3].

Selecting appropriate evaluation metrics is essential to ensure that model performance reflects practical project management objectives. In schedule prediction, metrics that capture both average errors and extreme deviations are particularly important for supporting proactive planning and decision-making [1].

Overall, the use of multiple regression evaluation metrics provides a comprehensive assessment of model performance, enhancing the reliability and interpretability of predictive results in project management contexts [1,3,4].

2.7. Dimensionality Reduction and Data Visualization Techniques

Building on the availability of diverse and high-dimensional project datasets, dimensionality reduction techniques play a critical role in AI-based project management studies, particularly when dealing with datasets that include numerous project attributes

and risk factors [3,4]. These techniques aim to reduce the number of input features while preserving the most informative characteristics of the data, thereby improving computational efficiency, model performance, and interpretability [3].

Principal Component Analysis (PCA) is one of the most widely used linear dimensionality reduction methods in project management research [3,4]. PCA transforms the original variables into a smaller set of uncorrelated components that capture the maximum variance within the dataset. While effective for feature reduction and noise filtering, PCA may have limitations in capturing complex nonlinear relationships commonly observed in project risk and schedule data [3,4].

To address these limitations, nonlinear visualization techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE) have been increasingly adopted [3,4]. Unlike PCA, t-SNE is primarily used for exploratory data visualization rather than predictive modeling. It maps high-dimensional data into a low-dimensional space while preserving local structures and similarities between data points. In the context of project management, t-SNE has been applied to visualize project clusters, identify hidden patterns, and explore similarities between project instances or risk profiles [3,4].

Several studies emphasize the importance of dimensionality reduction and visualization techniques as complementary tools to predictive modeling [3,5]. These methods facilitate a deeper understanding of data distributions, class separability, and feature interactions prior to model development. Such insights can significantly enhance model design, feature selection, and overall predictive performance.

However, the literature cautions against the misuse of visualization techniques such as t-SNE for quantitative evaluation, as their results are highly sensitive to parameter settings and may not generalize across different datasets [3]. Therefore, these methods should be used primarily for exploratory analysis rather than formal model assessment.

In summary, dimensionality reduction and visualization techniques serve as supportive analytical tools in AI-based project management research [1,3]. When applied appropriately, they enhance data understanding and contribute to the development of more robust, efficient, and interpretable risk and schedule prediction models.

2.8. Research Gaps and Motivation for the Current Study

Despite the growing body of literature on the application of Artificial Intelligence (AI) and Machine Learning (ML) in project management, several research gaps remain evident in the areas of risk and schedule management [1,3,4]. Existing studies have demonstrated the potential of ML models to improve predictive accuracy; however, many of these studies address risk management and schedule management as separate problems. This separation overlooks the inherently interdependent nature of risks and schedules in technical projects, thereby limiting the practical applicability of existing approaches in real-world environments [3,4].

Another significant gap relates to the selection and justification of ML models. While a wide range of algorithms—including regression models, decision trees, neural networks, and ensemble methods—have been explored [3,4], comparative analyses are often constrained by limited datasets and inconsistent evaluation frameworks. Moreover, many studies emphasize predictive accuracy without sufficiently addressing model interpretability and robustness, which are critical factors for adoption in project management practice [5].

Data-related challenges also persist in the literature. The limited availability of standardized and publicly accessible project management datasets restricts the reproducibility and generalizability of research findings [3,4]. Although synthetic datasets are increasingly used to overcome these limitations, relatively few studies provide rigorous justification or validation of their data generation processes, particularly in modeling the interactions between risks and schedule dynamics [3]. This raises concerns regarding the reliability and external validity of reported model performance.

Furthermore, evaluation strategies in prior studies often rely on a narrow set of performance metrics, which may not fully capture the practical implications of prediction errors in project decision-making [1,3]. In risk prediction, insufficient attention is given to class imbalance and the impact of false negatives, while in schedule prediction, extreme delay scenarios are not consistently emphasized [3,4]. Such limitations reduce the effectiveness of these models in supporting real-world project decisions.

Motivated by these gaps, this study proposes an AI-based framework that focuses on schedule prediction while incorporating risk-related variables within a unified analytical approach. The research adopts the Gradient Boosting Machine (GBM) model due to its strong predictive performance and its ability to handle heterogeneous, nonlinear, and imbalanced data [3,4,12]. In addition, the study employs a comprehensive set of evaluation metrics and utilizes carefully designed datasets to enhance prediction reliability and robustness.

Ultimately, this research aims to bridge the gap between advanced ML techniques and their practical application in technical project environments by providing a data-driven, interpretable, and decision-oriented approach to risk and schedule management.

3. Methodology

This study follows a structured and systematic workflow designed to support schedule prediction tasks. The process begins with dataset preparation, followed by data preprocessing using SPSS. Subsequently, a Gradient Boosting Machine (GBM) model is applied for predictive modeling. Finally, the model performance is evaluated using appropriate regression-based evaluation metrics to assess its effectiveness.

3.1. Dataset Description

The dataset used in this study consists of several project-related variables representing key aspects of project performance and management. These variables include project duration (months), total cost (USD), team size, risk factor, client satisfaction, and schedule overrun percentage.

The dataset was obtained from a publicly available source on Kaggle [16], which provides real-world datasets for research and data analysis. The use of publicly available data enhances the transparency, reproducibility, and reliability of the study, allowing other researchers to replicate the results and validate the proposed approach.

The selected variables capture both operational and performance-related dimensions of project management. For instance, project duration, cost, and team size reflect core project characteristics, while risk factor and client satisfaction provide insights into project uncertainty and stakeholder perception. The schedule overrun percentage represents deviations from planned timelines and is used as the primary target variable for schedule prediction.

The dataset reflects diverse project conditions, enabling the machine learning model to learn meaningful relationships between project characteristics and performance outcomes. This diversity supports the development of robust predictive models capable of handling variability in real-world project environments.

Although real-world datasets provide practical insights, they may include missing values, inconsistencies, and noise. Therefore, appropriate data preprocessing techniques are required to ensure data quality and improve the performance and reliability of the predictive models.

3.2. Data Preprocessing (SPSS)

The collected dataset was systematically preprocessed using IBM SPSS Statistics to ensure data quality, consistency, and analytical suitability prior to the application of machine learning models. Data preprocessing represents a critical stage in empirical analysis, as it directly influences the robustness, reliability, and generalizability of the resulting models [3,4].

The preprocessing pipeline involved a series of structured procedures, including data validation, handling of missing values, encoding of categorical variables, and transformation of the dataset into a machine learning–ready format. Since the dataset was obtained from a real-world source, it contained potential inconsistencies, missing values, and variations in data representation that required careful preprocessing.

Missing values were identified and handled appropriately to prevent bias in model training. In addition, consistency checks were performed to ensure that all variables were within valid ranges and correctly formatted. These steps helped improve data integrity and ensured the reliability of subsequent analysis.

To improve model stability and prevent scale-related bias, numerical features such as project duration, total cost, team size, and schedule overrun percentage were normalized. This step ensures that all input variables contribute proportionally to the learning process, thereby enhancing model convergence and predictive performance. Furthermore, variables such as client satisfaction were encoded into numerical representations where necessary to ensure compatibility with machine learning algorithms.

All preprocessing procedures were conducted within the SPSS environment, which provides a structured and reproducible framework for data transformation and analysis. This approach enhances data quality and establishes a solid analytical foundation for the development of predictive models in risk and schedule management contexts [1,3].

3.3. Model Development (GBM)

In this study, a Gradient Boosting Machine (GBM) model was employed to perform predictive analysis for project schedule performance. The selection of GBM is consistent with the literature discussed in Section 2, where ensemble learning techniques have demonstrated superior performance in modeling complex and nonlinear relationships within project data [3,4,12].

The model was developed using the preprocessed dataset obtained from the Kaggle platform [16] and prepared through SPSS, as described in the previous sections. The dataset includes key project-related variables such as project duration (months), total cost (USD), team size, risk factor, client satisfaction, and schedule overrun percentage.

The GBM model was designed to perform regression-based prediction of schedule performance using Schedule Overrun (%) as the primary target variable. In addition, Risk Factor was incorporated as an input variable to capture project uncertainty and its influence on schedule deviations. This modeling approach enables a comprehensive evaluation of project performance based on both operational characteristics and risk-related factors.

GBM is an ensemble learning method that constructs a sequence of decision trees, where each subsequent model is trained to minimize the prediction errors of the previous one. This iterative learning process allows the model to focus on difficult-to-predict instances, resulting in improved predictive accuracy and robustness [12]. The model is particularly effective in handling heterogeneous data, nonlinear relationships, and complex interactions among variables, which are common characteristics of technical project environments.

To ensure reliable model performance and reduce the risk of overfitting, k-fold cross-validation was applied during the training process. Specifically, a 10-fold cross-validation technique was used, where the dataset was divided into ten equal subsets. In each iteration, nine subsets were used for training the model, while the remaining subset was used for testing. This process was repeated ten times, allowing each subset to be used once as a testing set. The final model performance was calculated as the average of the results obtained across all folds, providing a more robust and reliable evaluation of the model's predictive capability.

Overall, the use of GBM in this study provides a robust and flexible predictive framework for analyzing project performance and predicting schedule deviations, supporting more informed and data-driven decision-making in project management environments.

3.4. Model Training and Testing

The model training and testing process was conducted using a k-fold cross-validation approach to ensure reliable and unbiased evaluation of model performance. In this study, a 10-fold cross-validation technique was applied, where the dataset was divided into ten equal subsets.

During each iteration, nine subsets were used for training the GBM model, while the remaining subset was used for testing. This process was repeated ten times, allowing each subset to serve as a testing set once. The final model performance was calculated as the average of the results obtained across all folds, providing a more robust assessment of the model's generalization capability.

The GBM model was trained to learn the relationships between input variables—such as project duration, total cost, team size, risk factor, and client satisfaction—and the target variable, which is Schedule Overrun (%). By evaluating the model on multiple subsets of unseen data, this approach minimizes the risk of overfitting and ensures that the model can generalize effectively to new project scenarios.

3.5. Evaluation Metrics

To evaluate the performance of the proposed model, a set of regression-based evaluation metrics was employed to ensure a comprehensive assessment of predictive accuracy and model effectiveness [3,4].

The model performance was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2). MAE measures the average magnitude of prediction errors, providing an intuitive indication of how far the predicted values deviate from the actual values. RMSE assigns greater weight to larger errors, making it particularly useful for identifying significant deviations in schedule predictions.

The coefficient of determination (R^2) indicates the proportion of variance in the target variable that is explained by the model, offering an overall measure of model fit and predictive capability. A higher R^2 value reflects better model performance and stronger explanatory power.

These metrics are widely used in regression analysis and provide complementary perspectives on model performance. The combination of MAE, RMSE, and R^2 enables a robust and reliable evaluation of the model's ability to predict Schedule Overrun (%), supporting accurate and data-driven decision-making in project management contexts.

3.6. Proposed AI-Based Framework for Risk and Schedule Management

This study proposes an AI-based framework designed to enhance schedule management in technical projects through a structured, data-driven approach. The framework integrates data preprocessing, machine learning modeling, and decision-support mechanisms to enable more accurate predictions and informed decision-making throughout the project lifecycle.

As illustrated in Figure 1, the framework consists of several interconnected stages. Initially, project data is collected from relevant sources and preprocessed using SPSS to ensure data quality, consistency, and analytical readiness. This stage includes data cleaning, handling missing values, normalization of numerical features, and preparation of the dataset for machine learning analysis.

Following preprocessing, the prepared dataset is analyzed using a Gradient Boosting Machine (GBM) model. The model performs a predictive task focused on estimating Schedule Overrun (%), which represents deviations from planned project timelines. This prediction is based on key project variables such as project duration, total cost, team size, risk factor, and client satisfaction. The predictive outputs provide actionable insights that support project managers in anticipating schedule deviations and improving project performance.

The framework further incorporates a decision-support layer, where model outputs are translated into practical actions such as schedule adjustments and resource optimization strategies. This enables project managers to shift from reactive to proactive decision-making, ultimately improving project efficiency and reducing uncertainty.

In addition, the framework includes a feedback loop mechanism, where new project data and outcomes are continuously integrated into the system to improve model accuracy and adaptability over time. This iterative process ensures that the framework remains responsive to evolving project conditions.

The proposed framework aligns with established project management practices and standards, emphasizing the critical role of schedule management in successful project execution [6,7,8]. It also reflects recent research trends that highlight the growing importance of Artificial Intelligence in enhancing decision-making and performance in project management environments [1,5].

Overall, the framework provides a structured and practical approach that combines advanced data analytics with traditional project management methodologies, contributing to more effective, adaptive, and proactive project control.

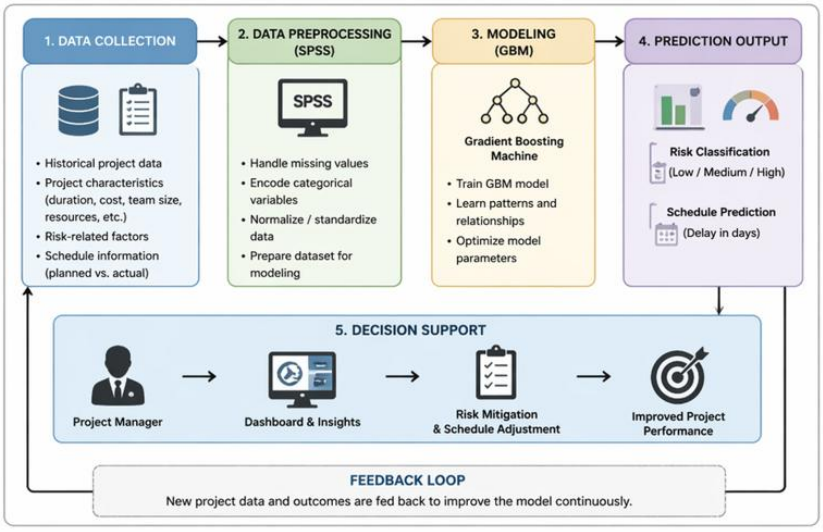

Figure 1. Proposed AI-based framework for schedule prediction and decision support in technical projects.

3.7. Research Workflow

The overall research process follows a structured and systematic workflow designed to ensure consistency, reliability, and reproducibility of the study. The workflow begins with data collection from a publicly available dataset obtained from Kaggle [16], followed by data preprocessing using SPSS to prepare the dataset for analysis.

The preprocessing stage includes data validation, handling missing values, normalization of numerical features, and transformation of the dataset into a machine learning–ready format. These steps ensure data quality and compatibility with machine learning algorithms.

Subsequently, a Gradient Boosting Machine (GBM) model is developed and applied to perform predictive analysis of project performance. The model focuses on estimating Schedule Overrun (%) based on key project variables, including project duration, total cost, team size, risk factor, and client satisfaction.

The model training and testing process is conducted using a 10-fold cross-validation approach to ensure robust evaluation and minimize overfitting. Model performance is then assessed using regression-based evaluation metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2).

This structured workflow provides a coherent and reproducible approach for analyzing project data and evaluating predictive performance. Furthermore, it reflects established project management practices that emphasize systematic planning, data-driven analysis, and continuous evaluation to improve project outcomes [6,7]. The integration of machine learning techniques within this workflow enhances the ability to accurately predict schedule deviations, ultimately supporting more informed and proactive decision-making in project management environments [1,3].

4. Results

4.1. Data Overview

This study utilizes a project management dataset consisting of 4517 records, where each record represents an individual project instance. The dataset was obtained from a publicly available source on Kaggle [16] and includes multiple variables describing project characteristics and performance indicators, providing a comprehensive representation of project environments.

The input variables include project duration (months), total cost (USD), team size, risk factor, and client satisfaction. These variables capture key operational and performance-related aspects of project execution.

The primary target variable considered in this study is Schedule Overrun (%), which represents the deviation from planned project timelines and is used to evaluate schedule performance.

Descriptive statistics of the main variables are presented in Table 1. These statistics provide insights into the distribution, variability, and range of the dataset, supporting a better understanding of the underlying project characteristics.

Table 1. Descriptive statistics of the project dataset (N = 4,517).

	Mean	Std. Dev.	Min	Max
Duration (Months)	30.68	17.10	1.01	60.00
Total Cost (USD)	1,025,727	572,913.9	50,465.59	1,999,905
Team Size	27.07	12.84	5	49
Risk Factor	0.51	0.23	0.10	0.90
Client Satisfaction	2.99	1.43	1	5
Schedule Overrun (%)	14.43	12.88	0	62.06

The descriptive statistics presented in Table 1 provide valuable insights into the characteristics of the dataset. The average project duration is approximately 30.68 months, indicating that the projects considered vary from short- to medium-term durations. The total cost shows a high mean value of approximately 1,025,727 USD, reflecting the financial scale of the projects, with a relatively large standard deviation indicating variability across projects.

The average team size is approximately 27.07 members, suggesting that the projects are moderately large in terms of human resources. The risk factor has a mean value of 0.51, indicating a balanced distribution of risk levels across projects, ranging from low to high.

Client satisfaction has a mean of 2.99 (on a scale of 1 to 5), suggesting moderate satisfaction levels, with variability across different projects. This reflects the influence of project performance factors on stakeholder perception.

In terms of schedule performance, the schedule overrun percentage has an average value of 14.43%, indicating that many projects experience delays beyond planned timelines. The relatively high standard deviation (12.88) suggests significant variation in schedule performance across projects.

Overall, the variability observed across the different variables highlights the complexity of project environments, making the dataset suitable for machine learning modeling and predictive analysis.

4.2. Exploratory Data Analysis

To better understand the relationships between key variables and project performance, exploratory data analysis (EDA) was conducted using scatter plots, as shown in Figures 2 and 3.

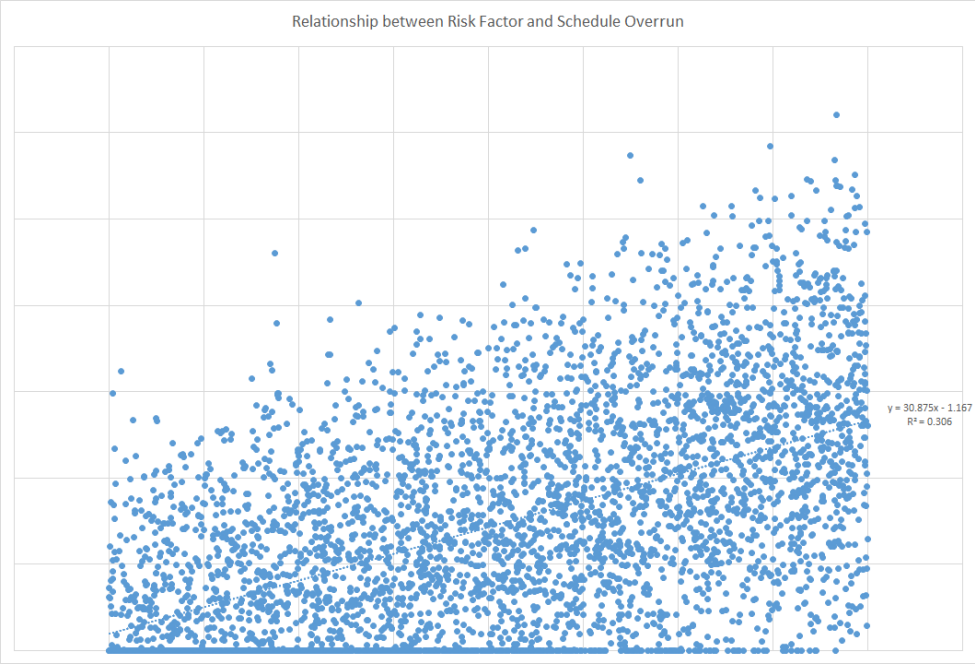

Figure 2. Relationship between Risk Factor and Schedule Overrun (%).

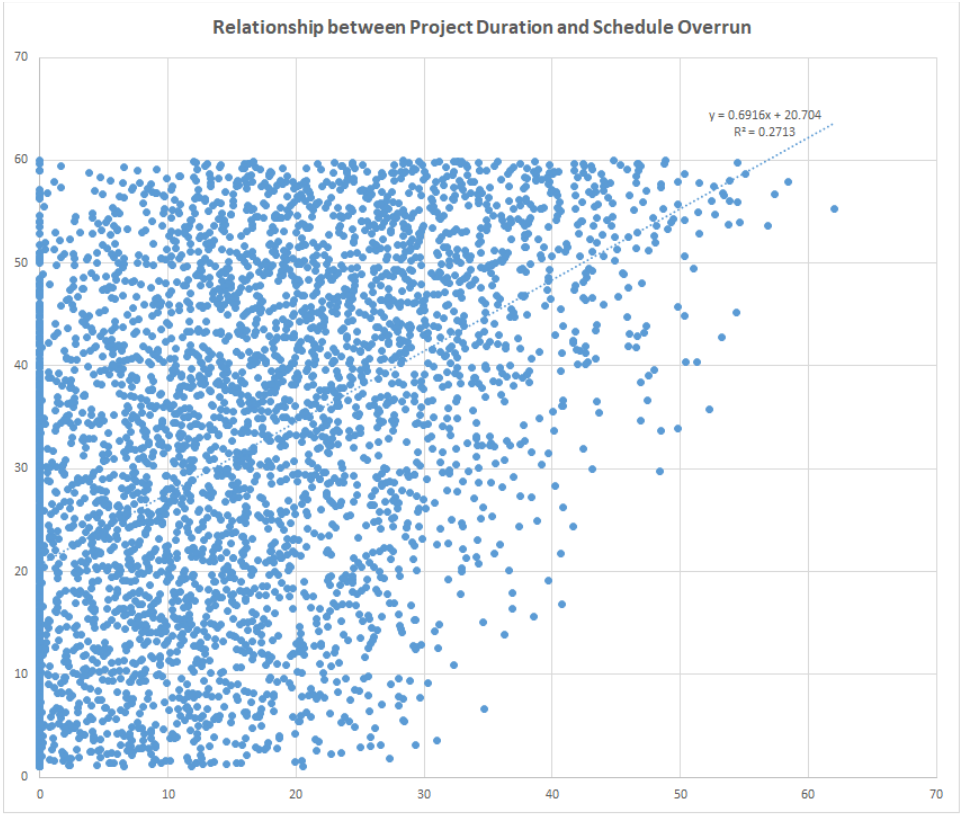

Figure 3. Relationship between Project Duration and Schedule Overrun (%).

Figure 2 illustrates the relationship between Risk Factor and Schedule Overrun (%). A clear positive trend can be observed, indicating that projects with higher risk levels tend to experience greater schedule overruns. This suggests that increased project uncertainty is directly associated with delays in project completion. The distribution of data points also reflects variability in schedule performance across different risk levels, highlighting the complexity of project environments.

Figure 3 presents the relationship between Project Duration and Schedule Overrun (%). A positive correlation is also observed, where longer project durations are associated with higher levels of schedule overrun. This indicates that extended project timelines may increase exposure to delays, potentially due to increased complexity, resource constraints, or evolving project conditions over time.

Overall, the exploratory analysis confirms that both risk-related factors and project duration play a significant role in influencing schedule performance. These findings support the selection of these variables in the predictive modeling phase and reinforce the suitability of the dataset for machine learning analysis.

4.3. Schedule Prediction Results (Schedule Overrun)

The schedule prediction model was developed using the Gradient Boosting Machine (GBM) algorithm to estimate Schedule Overrun (%) based on a set of project-related input features. To ensure a reliable evaluation of the model's performance, a 10-fold cross-validation approach was applied during the training and testing process.

The performance of the model was assessed using standard regression evaluation metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). These metrics provide a comprehensive understanding of the model's prediction accuracy and its ability to capture variations in schedule performance.

The obtained results are summarized as follows:

- RMSE = 7.128
- MAE = 5.946
- R^2 = 0.835

The RMSE value indicates that the model's predictions deviate from the actual schedule overrun values by approximately 7.13 percentage points on average. The MAE further confirms this by reflecting the average absolute prediction error. The R^2 value of 0.835 suggests that the model explains approximately 83.5% of the variance in schedule overrun, indicating a strong predictive performance.

These results demonstrate that the GBM model is highly effective in capturing the underlying patterns of schedule deviations. The relatively high R^2 value indicates that the selected input variables—including project duration, cost, team size, risk factor, and client satisfaction—provide meaningful explanatory power for predicting schedule performance.

Figure 4. Actual vs. predicted values for Schedule Overrun (%) using the GBM model.

Figure 4 presents the relationship between the actual and predicted values of Schedule Overrun (%). It can be observed that most data points are closely distributed around the regression line, indicating a strong agreement between predicted and actual values.

The concentration of points along the diagonal trend reflects the model's ability to accurately capture the relationship between input variables and schedule performance. While some deviations are present, which is expected in real-world data, the overall distribution demonstrates high prediction accuracy.

The strong alignment between actual and predicted values further supports the high R^2 score obtained, confirming that the model is effective in learning and generalizing the underlying patterns of schedule overrun.

4.4. Summary of Results

The results of this study demonstrate the effectiveness of machine learning techniques, particularly the Gradient Boosting Machine (GBM), in supporting schedule management in project environments.

The regression model achieved a strong level of performance, with an R^2 value of 0.835, indicating that the model is able to explain a significant portion of the variability in project schedule overrun. The corresponding error metrics (RMSE = 7.128 and MAE = 5.946) further confirm a high level of prediction accuracy, suggesting that the model successfully captures the underlying patterns within the dataset.

The exploratory data analysis also revealed that key variables—such as project duration and risk factor—have a noticeable impact on schedule performance. These findings are consistent with the predictive modeling results, where such variables contribute significantly to explaining variations in schedule overrun.

Overall, the findings indicate that AI-based models can provide valuable predictive insights for schedule forecasting. The strong predictive performance of the GBM model highlights its capability to model complex relationships within project data and to support data-driven decision-making.

These results emphasize the importance of integrating machine learning approaches into project management practices to improve forecasting accuracy and enhance proactive decision-making in complex project environments.

4.5. Feature Importance Analysis

To further understand the behavior of the proposed model, feature importance analysis was conducted using the Gradient Boosting Machine (GBM) algorithm. The results are presented in Figure 5.

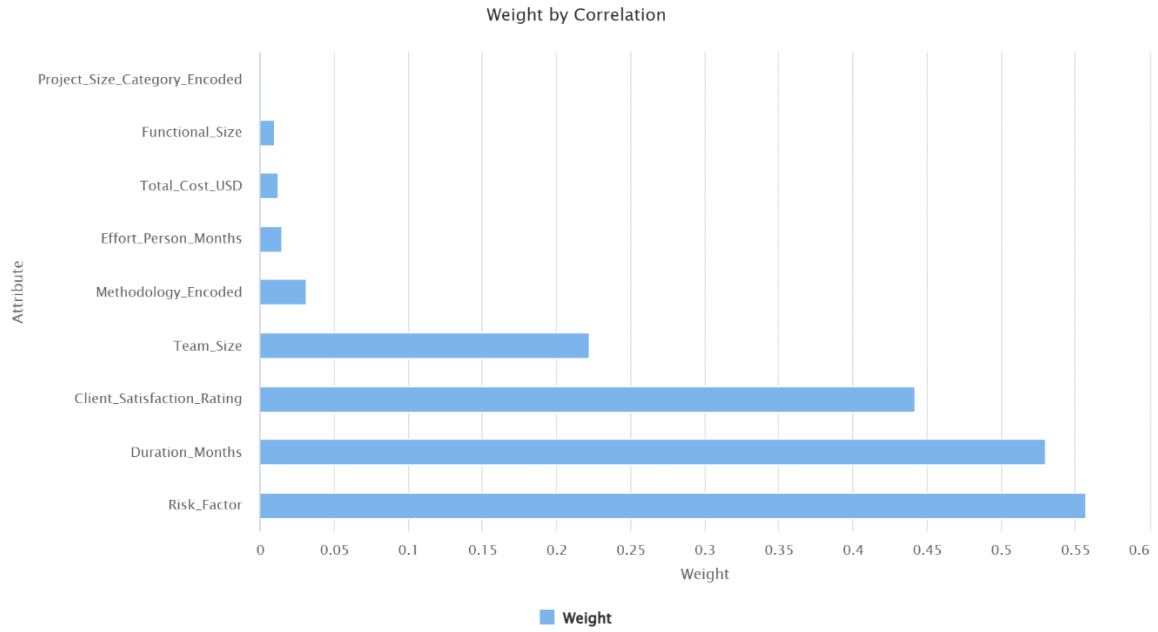

Figure 5. Feature importance obtained from the Gradient Boosting Machine (GBM) model.

The results indicate that Risk Factor is the most influential feature, demonstrating the highest importance weight among all variables. This highlights the critical role of project uncertainty in determining schedule performance and emphasizes its strong impact on schedule overrun.

The second most important feature is Project Duration, indicating that longer projects are more likely to experience higher schedule deviations. This aligns with the findings from the exploratory data analysis, where a positive relationship between duration and schedule overrun was observed.

Client Satisfaction also shows a relatively high importance, suggesting that project performance and stakeholder perception are closely related. Lower satisfaction levels may reflect underlying issues that contribute to schedule delays.

In contrast, Team Size demonstrates moderate importance, indicating that while human resources contribute to project performance, their impact is less significant compared to risk and duration factors.

Other variables exhibit relatively lower importance values, suggesting a weaker direct influence on the model's predictions within this dataset.

Overall, the feature importance analysis reveals that risk-related and duration-related variables dominate the prediction process, reinforcing the importance of managing project uncertainty and timeline complexity to improve schedule performance. These findings are consistent with the exploratory data analysis results and further validate the model's behavior.

5. Discussion

The results of this study demonstrate the effectiveness of machine learning techniques, particularly the Gradient Boosting Machine (GBM), in modeling and accurately predicting project schedule performance. The strong predictive performance of the model, reflected by an R^2 value of 0.835, indicates that a significant portion of the variability in schedule overrun can be explained by the selected input variables.

The findings from the exploratory data analysis and feature importance analysis consistently highlight the critical role of risk factor and project duration in influencing schedule performance. Projects with higher levels of risk and longer durations tend to experience greater schedule overruns, emphasizing the importance of managing uncertainty and project timelines effectively.

In addition, client satisfaction was identified as an influential variable, suggesting that project performance and stakeholder perception are closely interconnected. Lower satisfaction levels may indicate underlying issues in project execution, which can contribute to schedule deviations.

The feature importance results further support these observations, as risk-related and duration-related variables were identified as the most significant predictors. This reinforces the importance of incorporating risk assessment and time management practices into predictive modeling frameworks for project management.

These findings are consistent with prior studies that emphasize the significance of risk-related factors and project characteristics in influencing project outcomes [15]. The ability of the GBM model to capture complex and nonlinear relationships within project data further supports the growing role of AI in enhancing predictive capabilities in project management.

Furthermore, the results highlight the value of integrating AI-based models with traditional project management practices. By combining data-driven insights with managerial expertise, organizations can improve forecasting accuracy and support proactive decision-making in dynamic project environments.

Finally, the findings suggest that further improvements can be achieved by incorporating additional contextual variables, such as project complexity, stakeholder influence, and external environmental factors. Enhancing data quality and expanding feature sets may further improve model generalization and predictive performance in future research.

These results demonstrate the potential of data-driven approaches to transform traditional project management practices. This highlights the robustness and practical applicability of the proposed model when applied to large-scale real-world datasets.

6. Limitations

Despite the promising results obtained in this study, several limitations should be acknowledged.

First, although the dataset was obtained from a publicly available source on Kaggle, it may not fully capture the complexity and diversity of real-world project environments. Differences in data quality, structure, and context across industries may affect the generalizability of the model when applied to other project settings.

Second, while the regression model demonstrated strong predictive performance (R^2 = 0.835), a portion of the variability in schedule overrun remains unexplained. This suggests that additional factors—such as organizational structure, stakeholder influence, project complexity, and external environmental conditions—were not included in the current dataset.

Third, although the dataset is sufficiently large, further expansion may improve the robustness and generalization capability of the model. Utilizing larger and more diverse datasets could further enhance model stability and predictive accuracy.

Finally, the study focuses on a single machine learning technique (GBM). Although this model achieved strong performance, future research could explore and compare additional algorithms or hybrid approaches, including deep learning models, to further improve prediction accuracy and model generalization.

These limitations highlight important opportunities for future research to enhance model performance, robustness, and applicability in real-world project management scenarios.

7. Conclusion

This study investigated the role of artificial intelligence, particularly the Gradient Boosting Machine (GBM), in enhancing schedule management in project environments. The proposed approach applies a data-driven framework to predict Schedule Overrun (%) based on key project-related variables.

The findings demonstrate that the GBM-based model provides strong predictive performance, achieving an R^2 value of 0.835. This indicates that the model is capable of explaining a substantial portion of the variability in project schedule performance, highlighting its effectiveness in capturing complex relationships within project data.

The results further revealed that variables such as risk factor, project duration, and client satisfaction play a significant role in influencing schedule outcomes. In particular, higher levels of risk and longer project durations were associated with increased schedule overruns, emphasizing the importance of managing uncertainty and time-related factors in project execution.

Overall, the findings confirm that AI-based models can support project managers by providing accurate, data-driven insights that enhance forecasting capabilities and enable more proactive decision-making.

Future research should focus on expanding the dataset with more diverse real-world project data to improve model generalization. In addition, exploring and comparing multiple machine learning techniques, including advanced and hybrid models, may further enhance predictive performance. Incorporating additional contextual features—such as project complexity, stakeholder dynamics, and external environmental factors—could also contribute to developing more comprehensive and practically applicable solutions for project management.

Data Availability Statement

The dataset used in this study is publicly available on Kaggle and can be accessed at: https://www.kaggle.com/datasets/alphagamingsdf/software-project-dataset (accessed on 9 May 2026).

References

[1]. Fridgeirsson, T.V.; Ingason, H.T.; Jonasson, H.I.; Gunnarsdottir, H. A qualitative study on artificial intelligence and its impact on the project schedule, cost and risk management knowledge areas as presented in PMBOK®. Appl. Sci. 2023, 13, 11081.
[2]. Elokby, E.A.; Alawi, N.A.; Abdelgayed, A.T.A.; Al-Hodiany, Z.M. Does project risk management matter for the success of information technology projects in Egypt? In Proceedings of the 2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE), Putrajaya, Malaysia, 15–16 June 2021; pp. 243–250.
[3]. Mahdi, M.N.; Yusof, M.H.; Cheng, A.; Mohd Azmi, L.K.; Ahmad, M.S.; Al-R. Design and development of machine learning technique for software project risk assessment—A review. In Proceedings of the 2020 8th International Conference on Information Technology and Multimedia (ICIMU), Selangor, Malaysia, 2020; pp. 354–362.
[4]. Ibraigheeth, M.; Abu Eid, A.I. Software project risk assessment using machine learning approaches. Am. J. Multidiscip. Res. Dev. 2022, 4, 35–41.
[5]. Raisch, S.; Krakowski, S. Artificial intelligence and management: The automation–augmentation paradox. Acad. Manag. Rev. 2021, 46, 192–210.
[6]. Project Management Institute (PMI). A Guide to the Project Management Body of Knowledge (PMBOK® Guide), 6th ed.; PMI: Newtown Square, PA, USA, 2017.
[7]. Kerzner, H. Project Management: A Systems Approach to Planning, Scheduling, and Controlling; Wiley: Hoboken, NJ, USA, 2017.
[8]. International Organization for Standardization (ISO). ISO 31000: Risk Management—Guidelines; ISO: Geneva, Switzerland, 2018.
[9]. Meredith, J.R.; Mantel, S.J. Project Management: A Managerial Approach; Wiley: Hoboken, NJ, USA, 2012.
[10]. Kelley, J.E.; Walker, M.R. Critical-path planning and scheduling. Proc. East. Jt. Comput. Conf. 1959, 16, 160–173.
[11]. Flyvbjerg, B. Survival of the unfittest: Why the worst infrastructure gets built—and what we can do about it. Oxf. Rev. Econ. Policy 2009, 25, 344–367.
[12]. Friedman, J.H. Greedy function approximation: A gradient boosting machine. Ann. Stat. 2001, 29, 1189–1232.
[13]. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 2009, 45, 427–437.
[14]. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)? Geosci. Model Dev. 2014, 7, 1247–1250.
[15]. Raz, T.; Shenhar, A.J.; Dvir, D. Risk management, project success, and technological uncertainty. R&D Manag. 2002, 32, 101–109.
[16]. AlphaGamingSDF. Software Project Dataset. Kaggle, 2023. Available online: https://www.kaggle.com/datasets/alphagamingsdf/software-project-dataset (accessed on 9 May 2026).