Communication in Physical Sciences, 2022, 8(4): 908-933

Machine Learning-Based Prediction of Construction Project Delays in Public Building Projects in Nigeria

Muritala Olamilekan Issa

Received: 29 August 2022/Accepted:12 December 2022/Published: 30 December 2022

Abstract: *Construction project delays remain one of the major challenges affecting the successful delivery of public infrastructure in Nigeria, resulting in significant cost overruns, schedule extensions, and reduced project performance. This study developed and evaluated a machine learning framework for predicting delays in public building projects using historical project data from 1,200 completed projects executed between 2015 and 2022. Twenty-three project characteristics, including contract sum, project duration, payment delay, material price increase, contractor experience, project complexity, and design modifications, were used as predictor variables. The dataset was preprocessed through missing value imputation, outlier treatment, feature engineering, and standardization before model development. Six supervised machine learning algorithms comprising Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Network, and Extreme Gradient Boosting (XGBoost) were trained and evaluated using five-fold cross-validation. Results showed that 67.2% of the projects experienced schedule delays, with an average delay duration of 182.4 days. Correlation and feature importance analyses identified payment delay, project complexity, and material price escalation as the most influential predictors of project delay. Comparative evaluation indicated that the XGBoost model achieved the highest predictive performance with an accuracy of 96.67%, precision of 96.34%, recall of 96.08%, F1-score of 96.21%, and ROC-AUC of 0.989, outperforming the Artificial Neural Network (95.42%), Random Forest (93.75%), Support Vector Machine (91.25%), Decision Tree (87.08%), and Logistic Regression (84.58%). The optimized model correctly classified 232 out of 240 testing observations and demonstrated excellent robustness with a cross-validation standard deviation of only 0.34%. The developed framework provides an interpretable and highly reliable decision-support tool capable of identifying high-risk projects before significant schedule overruns occur, thereby supporting proactive project management, efficient resource allocation, and improved delivery of public building projects in Nigeria.*

Keywords: *Machine learning; Construction delays; Public building projects; XGBoost; Project risk prediction*

Muritala Olamilekan Issa
University of Ilorin
Kwara State, Nigeria
Email - Muritala.issa@unilorin.edu.ng

1.0 Introduction

Construction projects constitute a fundamental component of national economic development through the provision of critical infrastructure, employment generation, urban expansion, and improved social welfare. Public building projects, including educational institutions, healthcare facilities, administrative complexes, research laboratories, and public housing, represent substantial investments by governments aimed at enhancing socioeconomic development and improving citizens' quality of life. The successful delivery of these projects within the planned cost, time, and quality specifications is therefore essential for sustainable national development. However, construction projects worldwide continue to experience persistent delays, cost overruns, quality deficiencies, contractual disputes, and schedule disruptions arising from the complexity of project environments, uncertainty in resource availability, stakeholder interactions, and external economic conditions (Gondia *et al*., 2020; Zaman *et al*., 2022). These challenges have become increasingly pronounced as construction projects grow in scale, technological sophistication, and organizational complexity, thereby demanding more intelligent approaches to project planning, monitoring, and decision-making.

The construction industry is widely recognized as one of the most complex sectors because of its multidisciplinary nature, extensive stakeholder involvement, dynamic workflows, and exposure to numerous technical, financial, environmental, contractual, and managerial risks. Unlike manufacturing industries where production occurs under relatively controlled conditions, construction activities are executed in open environments where changing weather conditions, inflation, supply chain disruptions, labour shortages, equipment failures, design modifications, regulatory changes, and client interventions continuously influence project performance. These interacting uncertainties significantly affect project schedules and frequently result in delays that compromise project objectives. Consequently, project delays remain one of the most persistent challenges confronting construction industries worldwide and are responsible for substantial financial losses, contractual disputes, reduced productivity, and diminished stakeholder confidence (Yaseen *et al*., 2020; Ivanović *et al*., 2022).

Time performance has long been regarded as one of the principal indicators of project success alongside cost, quality, safety, and client satisfaction. Delays in project completion often trigger a cascade of adverse consequences, including increased project costs, contractual claims, legal disputes, reduced profitability, inefficient utilization of resources, delayed infrastructure delivery, and broader socioeconomic impacts. For governments, delayed public infrastructure projects postpone the delivery of essential public services while increasing maintenance costs and reducing the effectiveness of capital investments. Contractors experience additional overhead expenses, productivity losses, and cash-flow challenges, whereas consultants and project managers encounter increased administrative burdens and reputational risks. These multidimensional impacts emphasize the necessity for proactive delay identification and prediction before schedule slippages become irreversible (Ansari *et al*., 2022; Zaman *et al*., 2022).

The Nigerian construction industry is not exempt from these challenges. As one of the largest construction markets in Sub-Saharan Africa, Nigeria continues to invest heavily in public infrastructure to support population growth, urbanization, industrialization, and economic diversification. Nevertheless, numerous public building projects have experienced considerable delays, project abandonment, budget escalation, and contractual disputes arising from inadequate planning, delayed payments, inflation, procurement inefficiencies, political interference, shortage of construction materials, fluctuating exchange rates, insecurity, poor supervision, and weak project management practices. These recurring challenges have significantly reduced the efficiency of public expenditure and slowed national infrastructure development. Traditional project management techniques employed for monitoring project schedules often rely heavily on expert judgement and historical experience, which may be insufficient for accurately predicting project outcomes within increasingly complex project environments.

Conventional delay analysis methods generally depend on deterministic scheduling techniques, statistical analyses, or qualitative assessments that assume relatively simple relationships among project variables. However, construction projects exhibit highly nonlinear interactions among technical, managerial, financial, contractual, environmental, and organizational factors. Consequently, traditional analytical methods frequently fail to capture these intricate interdependencies, resulting in limited predictive capability and delayed managerial responses. Recent research has therefore advocated the adoption of data-driven approaches capable of learning hidden patterns from historical project information and supporting proactive decision-making throughout the project lifecycle (Gondia *et al*., 2020; Sanni-Anibire *et al*., 2021).

Advances in artificial intelligence (AI), big data analytics, and machine learning (ML) have transformed predictive modelling across numerous engineering disciplines by enabling computers to identify complex relationships within large datasets without requiring explicit mathematical formulations. Machine learning algorithms can automatically recognize hidden patterns, nonlinear dependencies, and interactions among multiple project variables, thereby providing accurate predictions that support informed managerial decisions. Within the context of construction management, these technologies are increasingly being employed for cost estimation, schedule forecasting, quality management, safety assessment, equipment maintenance, productivity optimization, risk management, and delay prediction (Sanni-Anibire *et al*., 2021; Ivanović *et al*., 2022). Their ability to continuously improve prediction accuracy as new data become available makes machine learning particularly attractive for construction projects characterized by uncertainty and dynamic operating conditions.

Several recent studies have demonstrated the growing effectiveness of machine learning techniques in solving construction management problems. Gondia *et al*. (2020) developed Decision Tree and Naïve Bayesian classification models to predict construction project delay risks using historical project data. Their findings demonstrated that machine learning algorithms effectively model the complex interrelationships among delay factors and provide objective decision-support tools capable of improving project planning. Similarly, Yaseen *et al*. (2020) proposed a hybrid Random Forest–Genetic Algorithm model for delay prediction, achieving a prediction accuracy of 91.67%, which substantially outperformed the conventional Random Forest classifier. The study confirmed that optimization algorithms can significantly enhance machine learning performance for construction project delay prediction.

Beyond delay prediction, researchers have increasingly applied artificial intelligence to broader aspects of construction project performance. Ashtari *et al*. (2022) developed a Bayesian Network classifier for assessing cost overrun risks while explicitly modelling the interrelationships among risk factors. Their study achieved an average prediction accuracy of 78.86%, significantly outperforming traditional Naïve Bayes and Decision Tree models. The results identified increases in material costs, inadequate workforce expertise, and inflation as the most influential contributors to construction cost overruns, demonstrating the importance of considering interactions among project variables rather than analysing individual risks independently.

Similarly, Alsugair (2022) developed a Partial Least Squares Structural Equation Model (PLS-SEM) to investigate the factors influencing cost deviation in Saudi construction projects. The findings revealed that project characteristics, contractual procedures, and estimator performance exert substantial direct and indirect influences on project cost deviation, highlighting that project performance is governed by complex interactions among organizational and technical variables. These findings reinforce the need for predictive models capable of simultaneously analysing multiple interdependent factors during project planning and implementation.

Construction safety has also emerged as an important determinant of project performance because accidents frequently result in work stoppages, increased costs, litigation, and schedule delays. Gunduz and Khader (2020) employed the Analytic Network Process integrated with a frequency-adjusted importance index to prioritize construction safety risks while accounting for their interdependencies. Their study demonstrated that construction hazards rarely occur independently and that modelling their interactions significantly improves risk prioritization and project planning. The recognition that construction risks are interconnected provides additional justification for employing machine learning algorithms capable of modelling complex nonlinear relationships among project variables.

The increasing complexity of construction projects has equally stimulated interest in systems-based approaches for predicting project performance. Ansari *et al*. (2022) developed a system dynamics model integrating the Analytic Hierarchy Process (AHP) and the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) to investigate how construction claims influence project performance indicators. Their results demonstrated that project delays constitute the most influential cause of construction claims and significantly affect project schedules, budgets, and overall project success. By capturing feedback relationships among project variables, the study emphasized the necessity of predictive approaches capable of analysing dynamic interactions rather than isolated project characteristics.

Collectively, these studies indicate a gradual transition from conventional project management methods toward intelligent data-driven decision-support systems capable of improving construction project performance. The convergence of machine learning, optimization algorithms, Bayesian reasoning, system dynamics, structural equation modelling, and multi-criteria decision-making techniques reflects the increasing recognition that construction projects are dynamic systems characterized by numerous interacting variables. Consequently, intelligent predictive models offer substantial opportunities for improving early risk identification, optimizing resource allocation, enhancing project monitoring, and supporting evidence-based managerial decisions throughout the construction lifecycle.

Despite the significant progress achieved in applying artificial intelligence and machine learning to construction management, the existing body of knowledge reveals several important conceptual, methodological, and contextual limitations. Most published studies have concentrated on construction projects undertaken in developed countries or rapidly industrializing economies, where project governance structures, procurement systems, technological adoption, regulatory environments, and data management practices differ considerably from those in many developing nations. Consequently, the applicability of these predictive models to the Nigerian public construction sector remains uncertain. Furthermore, while many studies report high prediction accuracies, relatively few have comprehensively evaluated the comparative performance of multiple machine learning algorithms using representative datasets from public building projects. This limitation restricts practitioners' ability to identify the most suitable predictive models for practical implementation within specific construction environments.

Existing studies have also largely focused on specific project categories such as megaprojects, infrastructure projects, tall buildings, or private sector developments. For instance, Sanni-Anibire *et al*. (2022) developed an Artificial Neural Network (ANN)-based model for delay risk assessment in tall building projects and achieved an impressive classification accuracy of 93.75%. Their findings demonstrated the superiority of ANN over other machine learning algorithms in predicting delay risks. However, the study focused exclusively on high-rise buildings, whose technical characteristics, procurement methods, construction technologies, and stakeholder arrangements differ substantially from those associated with government-funded public building projects such as schools, hospitals, administrative complexes, research institutes, and public service facilities. Consequently, the direct transferability of the developed model to Nigerian public building projects remains uncertain.

Similarly, Sanni-Anibire *et al*. (2021) proposed a comprehensive machine learning framework for construction delay mitigation based on the Cross-Industry Standard Process for Data Mining (CRISP-DM). The framework successfully demonstrated how machine learning techniques—including Multiple Linear Regression, K-Nearest Neighbours, Artificial Neural Networks, Support Vector Machines, and Ensemble Learning—can support cost estimation, duration estimation, and delay risk assessment. Although the framework provides an important conceptual foundation for Construction 4.0 implementation, its principal contribution lies in framework development and validation rather than the comparative evaluation of predictive algorithms using real-world public construction datasets. This creates an opportunity for further empirical investigations involving multiple machine learning techniques under comparable experimental conditions.

Several studies have further illustrated the potential of advanced machine learning techniques for improving project performance prediction. Gondia *et al*. (2020) demonstrated that Naïve Bayesian classifiers outperform Decision Tree algorithms in predicting construction delay risks because of their ability to accommodate probabilistic relationships among project variables. Likewise, Yaseen *et al*. (2020) integrated Random Forest with Genetic Algorithm optimization to improve delay prediction performance, reporting an overall prediction accuracy of 91.67%, a Kappa coefficient of 87%, and a classification error of only 8.33%. These findings indicate that combining machine learning algorithms with optimization techniques can substantially improve predictive performance. Nevertheless, both studies relied on datasets generated from construction environments outside Nigeria, thereby limiting their direct applicability to local public sector projects characterized by unique procurement regulations, funding structures, and institutional constraints.

Recent investigations have equally demonstrated the importance of modelling interactions among project risks rather than treating delay factors independently. Ashtari *et al*. (2022) developed Bayesian Network classifiers capable of representing the interdependencies among cost overrun risks and reported significantly higher prediction accuracy than conventional Naïve Bayes and Decision Tree models. Their findings identified increases in material prices, inadequate workforce expertise, and inflation as the most influential contributors to cost overruns. Likewise, Alsugair

(2022) demonstrated through Partial Least Squares Structural Equation Modelling (PLS-SEM) that project characteristics, contractual procedures, and estimator performance jointly influence construction cost deviations. These studies collectively suggest that project performance is governed by interconnected variables whose combined effects cannot be adequately represented using conventional deterministic methods.

The complexity of construction projects has equally motivated the integration of machine learning with expert knowledge and dynamic modelling techniques. Ivanović *et al*. (2022) introduced the Delay Root Causes Extraction and Analysis Model (DREAM), which combines machine learning with expert interpretation of project meeting records to identify delay root causes while minimizing subjective judgement. Similarly, Ansari *et al*. (2022) employed a system dynamics approach to predict project performance based on the causes of construction claims, demonstrating that delays constitute one of the most influential factors affecting project schedules, budgets, and overall project performance. These investigations illustrate that construction project behaviour is inherently dynamic and characterized by numerous feedback relationships requiring intelligent predictive systems capable of capturing nonlinear interactions among project variables.

Research has also highlighted that project performance extends beyond schedule adherence to include safety, quality, sustainability, leadership effectiveness, and stakeholder coordination. Gunduz and Khader (2020) emphasized that construction safety risks are highly interconnected and significantly influence project costs and completion time, while Yap *et al*. (2021) identified ineffective communication between consultants and contractors as one of the most influential consultant-related delay factors affecting sustainable housing projects. Furthermore, Zaman *et al*. (2022) established that critical delay factors exert a significant negative influence on the success of transnational mega construction projects, whereas effective leadership self-efficacy moderates these adverse effects. Collectively, these findings reinforce the multidimensional nature of construction project delays and highlight the necessity of adopting holistic predictive approaches capable of incorporating technical, managerial, organizational, financial, and behavioural variables.

Although these studies collectively demonstrate the considerable potential of artificial intelligence and machine learning in construction project management, important knowledge gaps remain. First, empirical studies applying machine learning specifically to public building projects in Nigeria are extremely limited despite the country's substantial investment in public infrastructure. Most available studies originate from Asia, Europe, or the Middle East, where project characteristics differ considerably from those encountered in Nigeria. Consequently, there remains insufficient empirical evidence regarding the predictive performance of machine learning algorithms under Nigerian construction conditions.

Second, many previous investigations have focused primarily on developing a single predictive model rather than systematically comparing multiple machine learning algorithms using identical datasets and evaluation criteria. Because machine learning algorithms exhibit different strengths depending on dataset characteristics, feature distributions, and problem complexity, comprehensive comparative assessments are essential for identifying the most appropriate predictive models for construction delay forecasting.

Third, while numerous studies identify delay factors through surveys, expert judgement, or statistical analyses, relatively few translate these findings into practical decision-support tools capable of predicting project delays before substantial schedule slippages occur. Public sector project managers require reliable predictive systems that provide early warnings and enable proactive interventions before delays escalate into significant cost overruns or contractual disputes.

Fourth, previous studies have concentrated predominantly on prediction accuracy while providing limited discussion of model interpretability and practical implementation within government construction agencies. The adoption of machine learning in project management depends not only on predictive performance but also on transparency, explainability, computational efficiency, and ease of integration into existing project management practices.

Against this background, the present study aims to develop robust machine learning models for predicting delays in public building projects in Nigeria using historical project information and relevant project performance indicators. Specifically, the study seeks to identify the critical factors influencing project delays, preprocess and optimize construction project datasets, develop and compare the predictive performance of multiple machine learning algorithms, determine the most accurate and reliable prediction model, and identify the most influential variables contributing to project delays. The findings are expected to provide an evidence-based framework capable of supporting proactive project monitoring and timely managerial interventions throughout the project lifecycle.

The significance of this study is both theoretical and practical. Theoretically, it contributes to the growing body of knowledge on the application of artificial intelligence and machine learning in construction project management, particularly within the context of developing countries where empirical studies remain relatively scarce. The study also enriches Construction 4.0 research by demonstrating the applicability of intelligent predictive analytics for managing schedule performance in public infrastructure projects. Practically, the developed prediction models will provide government agencies, contractors, consultants, project managers, and policymakers with objective decision-support tools for identifying projects at high risk of delay, prioritizing mitigation strategies, optimizing resource allocation, improving procurement planning, and enhancing project monitoring. Early prediction of potential delays will facilitate timely corrective actions, reduce unnecessary cost escalation, minimize contractual disputes, improve accountability in public expenditure, and promote more efficient delivery of critical public infrastructure. Ultimately, the study is expected to support data-driven decision-making and encourage the wider adoption of intelligent project management systems within Nigeria's construction industry, thereby contributing to improved project performance, enhanced infrastructure delivery, and sustainable national development.

2.0 Materials and Methods

This study adopted a quantitative, data-driven research methodology to develop and evaluate machine learning models for predicting delays in public building projects in Nigeria. The methodology was designed to establish predictive relationships between historical project characteristics and construction delay outcomes using supervised machine learning techniques. The entire modelling process followed the Cross-Industry Standard Process for Data Mining (CRISP-DM), a widely accepted framework for machine learning applications in construction management because of its systematic approach to data acquisition, preprocessing, modelling, evaluation, and deployment (Sanni-Anibire *et al*., 2021). The methodology comprised six sequential stages, namely data collection, data preprocessing, feature engineering, model development, model validation, and comparative performance evaluation.

2.1 Research Design

A retrospective observational research design was employed using historical data obtained from completed public building projects executed across Nigeria. Since the outcome of each project was already known, the study adopted a supervised machine learning approach in which project characteristics served as predictor variables while project delay status constituted the target variable. Supervised learning was considered appropriate because the algorithms learn functional relationships between independent variables and known output classes, thereby enabling future predictions for unseen projects.

The overall modelling framework involved collecting historical project information, cleaning and preprocessing the dataset, selecting the most influential variables, training several machine learning algorithms, optimizing their hyperparameters, validating their predictive capabilities using cross-validation techniques, and finally comparing their performances using standard classification metrics. This methodology enabled objective evaluation of different predictive algorithms under identical experimental conditions while minimizing model bias and overfitting.

2.2 Data Sources and Dataset Description

The study utilized secondary data obtained from completed public building projects executed between 2015 and 2022. The projects included government-funded educational institutions, hospitals, administrative complexes, research laboratories, judicial buildings, and other institutional infrastructure developed at the federal and state levels. Historical project records were compiled from official construction completion reports, project monitoring documents, consultancy reports, quantity surveying records, contractor performance reports, and procurement databases maintained by relevant government agencies, including the Federal Ministry of Works, State Ministries of Works, the Bureau of Public Procurement (BPP), the Federal Capital Development Authority (FCDA), the Universal Basic Education Commission (UBEC), Tertiary Education Trust Fund (TETFund), and other public infrastructure agencies.

A structured dataset comprising 1,200 completed projects was developed for the machine learning analysis. Each completed project represented a single observation in the dataset, while the project attributes constituted the predictor variables. The variables considered included original contract sum, planned project duration, percentage project completion, number of design modifications, contractor experience, consultant experience, payment delay period, inflation rate, material price fluctuation, labour availability, equipment availability, skilled workforce index, procurement method, funding source, building type, project location, rainfall intensity, security challenges, site accessibility, utility availability, regulatory approval duration, number of subcontractors, and project complexity index.

The dependent variable was defined as the project delay status and represented a binary classification problem. Projects completed within the approved contractual period were assigned a value of zero, whereas projects completed beyond the contractual completion date were assigned a value of one. Mathematically, the target variable was expressed according to equation

$$Y = \begin{cases} \text{if the project was completed on schedule} \\ \quad \text{if the project experienced delay} \end{cases}$$

(1)where (Y) denotes the project delay status. In addition to binary classification, the actual delay duration expressed in days was retained for supplementary regression analysis to evaluate the capability of selected algorithms in estimating the magnitude of project delays.

2.3 Data Preprocessing

The quality of machine learning models depends largely on the quality of the input dataset. Consequently, extensive preprocessing procedures were undertaken before model development to improve data consistency, eliminate inconsistencies, reduce noise, and enhance prediction performance.

The first preprocessing stage involved identifying missing observations. Numerical variables containing missing values were imputed using the median of the corresponding variable because the median is less sensitive to extreme observations than the arithmetic mean. The imputed value was calculated using equation 2

$$X_i = Median\ (X) \qquad (2)$$

where $X_i$is enotes the median value of the variable or dataset X and represents the median value of the corresponding predictor variable.

Missing categorical observations were replaced using the mode, defined as the category with the highest frequency within the dataset.

Outlier detection constituted the second preprocessing stage. Since unusually large or small observations may distort model training, the Interquartile Range (IQR) method was employed to identify extreme values. The interquartile range was computed as

$$UOR = Q_3 - Q_1 \qquad (3)$$

[

where Q_1) and Q_3 denote the first and third quartiles of the data distribution, respectively.

Observations satisfying either equation 4 or 5

$$X < Q_3 - 1.5\ (IQR) \qquad (4)$$

$$X < Q_1 + 1.5\ (IQR) \qquad (5)$$

were considered potential outliers. Each identified observation was subsequently examined individually. Genuine project observations representing actual field conditions were retained, whereas observations resulting from data-entry errors were corrected or removed.

Since several predictor variables were categorical in nature, including procurement method, funding source, project location, and building type, these variables were transformed into numerical representations using One-Hot Encoding. This encoding procedure generated binary indicator variables for each category while avoiding artificial ordinal relationships among categorical attributes.

Continuous predictor variables exhibited different measurement scales, ranging from monetary values to percentages, durations, and index scores. To eliminate scale-related bias during model training, numerical variables were standardized using Z-score normalization according to equation 6

$$Z = \frac{X-\mu}{\sigma} \quad (6)$$

where (Z) represents the standardized variable, (X) denotes the observed value, (\mu) is the sample mean, and (\sigma) is the corresponding standard deviation.

Standardization transformed all continuous variables to zero mean and unit variance, thereby improving convergence of optimization algorithms and enhancing the performance of distance-based learning techniques such as Support Vector Machines and Artificial Neural Networks.

2.4 Feature Engineering

Feature engineering was undertaken to improve the predictive capability of the machine learning models by constructing informative variables from the original dataset. Several derived variables were generated to better characterize project performance. Examples included project cost intensity, computed as the ratio of contract sum to planned project duration, payment delay ratio, contractor experience category, and project complexity index derived from multiple project characteristics. These engineered variables were intended to capture nonlinear relationships that may not be adequately represented by the original variables alone.

Correlation analysis was subsequently performed to identify highly correlated predictor variables. Strong multicollinearity among predictors can reduce model stability and increase prediction variance, particularly in regression-based classifiers. Pearson's correlation coefficient was therefore calculated as

$$r = \frac{\sum_{i=1}^{b}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \text{ x} \quad (7)$$

where r is the correlation coefficient,x_i and y_i represent paired observations,$\bar{x}$ and $\bar{y}$ are their respective sample means, and (n) is the total number of observations. Variables exhibiting excessively high correlation coefficients were carefully evaluated to avoid redundancy during model development.

2.5 Feature Selection

Following preprocessing and feature engineering, feature selection was performed to identify the most influential predictor variables contributing to construction project delays. Eliminating redundant or irrelevant variables improves computational efficiency, reduces overfitting, and enhances model interpretability.

The study employed Random Forest Feature Importance and Recursive Feature Elimination (RFE) to rank predictor variables according to their predictive contributions. Feature importance within the Random Forest algorithm was quantified using the total reduction in Gini impurity contributed by each predictor across all decision trees. The Gini importance for the (j^{th}) predictor was computed using equation 8

$$GI_j = \sum_{t=1}^{T} \Delta G_{jt} \quad (8)$$

]

where GI_j denotes the importance score of the jth feature, ΔG_{jt} represents the reduction in Gini impurity produced by that feature at node (t), and (T) is the total number of tree nodes in which the feature participated. Recursive Feature Elimination was subsequently applied to iteratively remove the least significant variables while monitoring model performance. The procedure continued until further elimination resulted in a measurable reduction in prediction accuracy. The final subset of predictor variables was therefore selected based on its ability to maximize classification performance while minimizing model complexity, thereby ensuring that only the most informative construction project attributes were retained for machine learning model development.

2.6 Data Partitioning and Cross-Validation

Following feature selection, the dataset was partitioned into training and testing subsets to enable unbiased evaluation of the developed machine learning models. Data partitioning is an essential step in supervised machine learning because it ensures that predictive models are evaluated using observations that were not involved during model training, thereby providing a realistic estimate of their generalization capability.

The complete dataset of 1,200 completed public building projects was randomly split into training and testing sets in an 80:20 ratio. Consequently, 960 project records (80%) were used for model training, while the remaining 240 (20%) were reserved for independent testing. Random sampling was adopted to ensure that the distribution of delayed and non-delayed projects remained approximately consistent in both datasets.

To further improve model robustness and reduce overfitting, five-fold cross-validation was employed during model training. In this procedure, the training dataset was partitioned into five approximately equal subsets. During each iteration, four subsets were used for model training, while the remaining subset was used for validation. This process was repeated five times so that each subset served once as the validation dataset. The overall cross-validation performance was computed as the average of the five validation scores, expressed as

$$CV = \frac{1}{k}\sum_{i=1}^{k} S_i \qquad (9)$$

where (CV) is the average cross-validation score, S_i is the validation score obtained during the (i^{th}) fold, and (k) represents the total number of folds, which was equal to five in this study.

Cross-validation minimizes dependence on a single random data partition and provides a more reliable estimate of model performance, particularly when comparing multiple machine learning algorithms.

2.7 Development of Machine Learning Models

Six supervised machine learning algorithms were developed and evaluated to determine the most suitable model for predicting delays in Nigerian public building projects. These algorithms were selected because of their proven effectiveness in solving nonlinear classification problems involving complex interactions among predictor variables. The selected models included Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Extreme Gradient Boosting (XGBoost).

2.7.1 Logistic Regression

Logistic Regression served as the baseline classification model because of its simplicity, computational efficiency, and interpretability. Unlike linear regression, Logistic Regression predicts the probability that an observation belongs to a particular class by transforming the linear predictor through the logistic function. The probability of project delay was estimated as

$$P(Y=1) = \frac{1}{1+exp(\beta_0 + \sum_{i-1}^{n} \beta_i X_i)} \quad (10)$$

where $P(Y=1)$ denotes the probability that a project will experience delay, (\beta_0) is the intercept, β_i represents the regression coefficient associated with the ith predictor variable, and X_i denotes the corresponding predictor variable. Projects with predicted probabilities greater than the selected decision threshold were classified as delayed projects.

2.7.2 Decision Tree

Decision Tree classification constructs a hierarchical tree structure by recursively partitioning the predictor space into homogeneous subsets. At each decision node, the predictor variable producing the greatest reduction in impurity is selected. Node impurity was quantified using the Gini Index defined as equation 11

$$G = 1 - \sum_{i-1}^{c} p_i^2 \qquad (11)$$

where (G) represents the Gini impurity, (p_i) denotes the probability of observations belonging to class (i), and (c) is the total number of classes. Lower Gini values indicate purer nodes and improved classification performance.

2.7.3 Random Forest

Random Forest extends the Decision Tree algorithm by constructing an ensemble of independently trained decision trees using bootstrap sampling and random feature(e selection. The final prediction was determined by majority voting across all individual trees. Mathematically, the ensemble prediction was expressed as the equation 12

$$\acute{Y} = Mode\{T_1(X), T_2(X) \ldots\ldots T_N(X)\} \quad (12)$$

where Ý is the predicted class, $T_i(X)$ denotes the prediction of the i^{th} decision tree, and (N) represents the total number of trees in the forest. The Random Forest algorithm reduces prediction variance and generally provides higher predictive accuracy than a single decision tree.

2.7.4 Support Vector Machine

Support Vector Machine constructs an optimal separating hyperplane that maximizes the margin between delayed and non-delayed project classes. For nonlinear datasets, observations are transformed into a higher-dimensional feature space using kernel functions. The decision function is expressed as equation 13

$$f(X) = \qquad (13)$$

where W is the weight vector defining the separating hyperplane, (X) denotes the predictor vector, and (b) represents the bias term. The optimization problem seeks to minimize

$\min\frac{1}{2}\left[\lceil W\rceil\right]^2$ subject to $y_i(W^Y X_i + b) \geq 1, \forall_i = 1, 2, \ldots n$ and denote the observed project class The Radial Basis Function (RBF) kernel was adopted because of its ability to capture complex nonlinear relationships among construction project variables.

2.7.5 Artificial Neural Network

Artificial Neural Networks imitate the learning mechanism of biological neurons through interconnected computational nodes organized into input, hidden, and output layers. Each neuron receives weighted input signals that are transformed using nonlinear activation functions.

The weighted input to each neuron was computed as equation 14

$$z = \sum_{i=1}^{b} w_i X_i + b \qquad (14)$$

where w_i denotes the connection weight associated with predictor X_i, and (b) is the bias parameter. The hidden neurons employed the Rectified Linear Unit (ReLU) activation function, given by equation 15, while the output neuron employed the sigmoid activation function, given as equation 16, which represents the transformation of the network to network output into a probability ranging between zero and one.

$$f(x) = \max(0, z) \qquad (15)$$

$$\sigma\,(z) = \frac{1}{1+\exp(-z)} \qquad (16)$$

Network weights were optimized using the backpropagation learning algorithm combined with the Adam optimization method.

2.7.6 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is an ensemble learning algorithm that sequentially constructs decision trees to minimize prediction errors produced by previous trees. Unlike conventional boosting methods, XGBoost incorporates regularization to improve model generalization and prevent overfitting.

The prediction model is represented by equation 17

$$y_1 = \sum_{k=1}^{K} f_k(X_o) \qquad (17)$$

where (\hat{y}_i) is the predicted output for observation (i), (f_k) denotes the (k^{th}) regression tree, and (K) is the total number of trees. The objective function minimized during model training was

$$Obj = \sum_{i=1}^{n} L\left(y_i, \acute{Y}_i\right) + \sum_{k=1}^{K} \Omega(f_k) \qquad (18)$$

where Obj is the overall objective function to be minimized during training, n is the total number of training instances, $L\left(y_i, \acute{Y}_i\right)$ is a convex loss function measuring the difference between the actual label $y_i$and the predicted value $\acute{Y}_i$, K is the total number of decision trees in the ensemble and $\Omega(f_k)$is the regularization term penalizing the complexity of the kth tree fk to prevent overfitting. Regularization improves prediction stability and enhances the generalization capability of the model.

2.8 Hyperparameter Optimization

The predictive performance of machine learning algorithms depends strongly on the selection of hyperparameters. Consequently, hyperparameter optimization was performed using Grid Search combined with five-fold cross-validation. Grid Search systematically evaluates all possible combinations of predefined hyperparameter values to identify the configuration yielding the highest validation performance.

For the Decision Tree and Random Forest models, the optimized parameters included maximum tree depth, minimum number of samples required for node splitting, minimum samples per leaf node, and the number of trees. The Support Vector Machine optimization considered the penalty parameter (C) and kernel width (\gamma), while the Artificial Neural Network optimization involved the number of hidden neurons, learning rate, batch size, number of training epochs, and dropout rate. For XGBoost, the optimized parameters included learning rate, maximum tree depth, subsampling ratio, column sampling ratio, and regularization coefficients.

The optimal parameter combination was selected based on the highest average cross-validation accuracy, thereby ensuring maximum predictive performance while minimizing overfitting.

2.9 Model Performance Evaluation

The predictive performances of the developed machine learning models were evaluated using several widely accepted classification metrics derived from the confusion matrix. These metrics quantify different aspects of predictive performance and provide a comprehensive comparison among competing algorithms. Classification accuracy, representing the proportion of correctly classified observations, was computed as

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (19)$$

where (TP) denotes true positives, (TN) represents true negatives, (FP) denotes false positives, and (FN) represents false negatives.

Precision, which measures the proportion of correctly predicted delayed projects among all projects predicted as delayed, was calculated as

$$Precision = \frac{TP}{TP+FP} \quad (20)$$

Recall, also known as sensitivity, measures the proportion of actual delayed projects correctly identified by the model and was calculated as

$$ecall = \frac{TP}{TP+FN} \quad (21)$$

The F1-score, which provides the harmonic mean of precision and recall, was computed as equation 22

$$F_1 = \frac{2\times Precision \times Recall}{Precision+Recall} \quad (22)$$

Where F_1 is the harmonic mean of Precision and Recall (providing a balanced measure of performance, especially on imbalanced datasets), precision is the accuracy of positive predictions and recall is sensitivity or the proportion of actual positive cases correctly identified.

The Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the Curve (AUC) were additionally employed to evaluate the discriminatory capability of each classification model across different probability thresholds.

Where regression analysis of delay duration was performed, the Mean

Absolute Error (MAE) and Root Mean Square Error (RMSE) were calculated according to equations 23 and 24

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (23)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (24)$$

where y_i denotes the observed delay duration, $\hat{y}_i$ represents the predicted delay duration, and n is the number of observations. Higher values of Accuracy, Precision, Recall, F1-score, and ROC-AUC together with lower MAE and RMSE values were interpreted as indicators of superior predictive performance.

3.0 Results and Discussion

This section presents the findings obtained from the development and evaluation of machine learning models for predicting delays in public building projects in Nigeria. The results are presented sequentially beginning with the descriptive characteristics of the dataset, followed by feature selection, comparative evaluation of machine learning algorithms, model validation, and interpretation of the most influential delay factors. The discussion compares the findings with previous studies while highlighting their practical implications for project management and decision-making within the Nigerian construction industry.

3.1 Descriptive Statistics of the Project Dataset

A total of 1,200 completed public building projects executed between 2015 and 2022 were analyzed. The projects comprised educational facilities, hospitals, administrative buildings, research institutes, judicial complexes, and other government-funded infrastructure distributed across the six geopolitical zones of Nigeria. Prior to model development, descriptive statistical analysis was performed to examine the characteristics of the dataset and assess the distribution of the predictor variables.

The descriptive statistics presented in Table 1 indicate substantial variability among the public building projects considered in this study. The average contract value was ₦865.40 million, although project costs ranged from ₦48.50 million to over ₦3.54 billion, reflecting the diversity of government-funded building projects included in the dataset. Likewise, the planned project duration averaged 18.7 months, whereas the observed completion period averaged 24.81 months, indicating that many projects exceeded their scheduled completion dates.

The mean delay duration of 182.40 days (approximately six months) demonstrates that schedule overruns remain a major challenge in Nigerian public construction projects. The relatively high standard deviation (128.53 days) further suggests considerable variation in project performance, indicating that while some projects experienced minimal delays, others suffered extensive schedule overruns exceeding one and a half years.

Table 1. Descriptive statistics of the project dataset (n = 1,200)

Variable	Mean	Standard Deviation	Minimum	Maximum
Contract sum (₦ million)	865.40	522.63	48.50	3,540.70
Planned project duration (months)	18.70	7.42	6	48
Actual completion period (months)	24.81	10.58	7	61
Delay duration (days)	182.40	128.53	0	615
Contractor experience (years)	13.80	5.64	2	31
Consultant experience (years)	16.40	6.12	4	34
Payment delay (days)	76.50	43.70	0	214
Material price increase (%)	18.40	9.60	2.10	47.80
Design modifications (No.)	3.10	2.02	0	10
Project complexity index	6.84	1.76	2.20	9.90

The average contractor experience of 13.8 years and consultant experience of 16.4 years indicate that most projects were handled by relatively experienced professionals. Nevertheless, the persistence of project delays despite this level of experience suggests that external project-related factors may exert greater influence on schedule performance than professional experience alone.

Payment delays averaged 76.5 days, highlighting the prevalence of delayed fund disbursement in public sector construction. Such payment interruptions often affect contractor cash flow, reduce labour productivity, delay procurement of construction materials, and ultimately contribute to schedule slippage. Furthermore, material prices increased by an average of 18.4% during project execution, reflecting inflationary pressures that have characterized Nigeria's construction sector in recent years. Material price volatility frequently necessitates contract variations, procurement revisions, and work interruptions, thereby increasing the likelihood of project delays.

The average number of design modifications was 3.1, suggesting that scope changes remain relatively common during project implementation. Frequent design revisions generally require additional approvals, redesign efforts, and contract adjustments, all of which negatively affect project schedules. Similarly, the mean project complexity index of 6.84 indicates that many of the projects involved considerable technical, managerial, and logistical challenges requiring sophisticated coordination among multiple stakeholders.

Overall, the descriptive statistics reveal substantial heterogeneity within the dataset, providing an appropriate foundation for applying machine learning algorithms capable of modelling complex nonlinear relationships among predictor variables. The observed variability also supports the suitability of ensemble learning techniques, which generally perform well when predictor variables exhibit diverse distributions and intricate interactions.

To further understand the prevalence of schedule overruns within the dataset, the completed projects were classified into delayed and non-delayed categories.

Figure 1 shows that 806 projects (67.2%) experienced schedule delays, whereas only 394 projects (32.8%) were completed within the approved contractual period. This finding demonstrates that project delays remain a widespread problem in Nigeria's public construction sector. The predominance of delayed projects provides sufficient class representation for supervised machine learning classification and reduces the likelihood of severe class imbalance during model training.

The high proportion of delayed projects agrees with previous investigations reporting that construction delays continue to affect project delivery worldwide. Gondia *et al*. (2020) similarly observed that construction projects are characterized by numerous interacting delay factors that complicate schedule prediction, while Sanni-Anibire *et al*. (2022) emphasized that delay risk has become particularly significant in complex building projects. Likewise, Yaseen *et al*. (2020) reported that uncertainty associated with procurement, financial management, resource allocation, and project coordination necessitates the adoption of intelligent predictive systems capable of providing early warnings before delays become critical.

The findings also reinforce observations by Ashtari *et al*. (2022), who demonstrated that construction performance is influenced by multiple interacting risk factors rather than isolated variables. The high incidence of delayed projects observed in the present study therefore justifies the application of advanced machine learning techniques capable of simultaneously analysing numerous project characteristics and their nonlinear interactions. These descriptive results establish the foundation for the subsequent feature selection and model development stages presented in the following sections.

Table 2:Pearson correlation matrix of major predictor variables

Variable	Delay Duration	Payment Delay	Material Price Increase	Design Modifications	Project Complexity	Contractor Experience	Planned Duration
Delay duration	1.000	0.718	0.653	0.611	0.694	–0.382	0.547
Payment delay	0.718	1.000	0.514	0.456	0.521	–0.341	0.389
Material price increase	0.653	0.514	1.000	0.437	0.486	–0.212	0.318
Design modifications	0.611	0.456	0.437	1.000	0.574	–0.286	0.401
Project complexity	0.694	0.521	0.486	0.574	1.000	–0.301	0.503
Contractor experience	–0.382	–0.341	–0.212	–0.286	–0.301	1.000	–0.168
Planned duration	0.547	0.389	0.318	0.401	0.503	–0.168	1.000

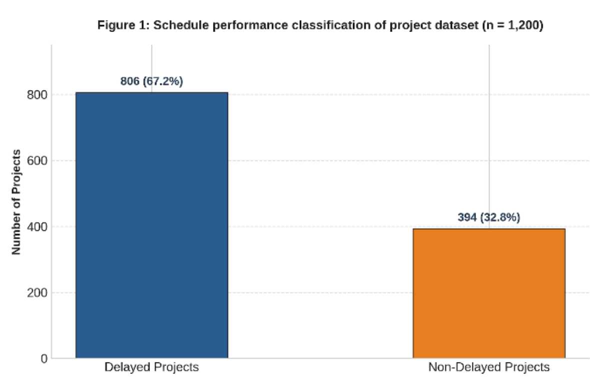


Fig. 1: Classification of Public Building Projects by Schedule Performance ($n = 1,200$)

3.2 Correlation Analysis and Feature Importance

Following the descriptive analysis, correlation analysis was performed to examine the relationships among the major predictor variables and identify potential multicollinearity before model development. Although ensemble learning algorithms such as Random Forest and XGBoost are generally robust to correlated variables, understanding the degree of association among predictors provides valuable insight into the mechanisms influencing construction project delays and improves interpretation of the prediction models.

Pearson's correlation coefficient ((r)) was computed for the principal numerical variables. Correlation coefficients range from −1 to +1, where positive values indicate direct relationships, negative values indicate inverse relationships, and values close to zero imply weak or no linear association.

The correlation analysis presented in Table 2 indicates that several project characteristics exhibited moderate to strong positive relationships with construction delays. Among all predictor variables, payment delay recorded the strongest positive correlation with delay duration ((r = 0.718)), suggesting that delayed release of project funds substantially increases the likelihood of schedule overruns. Delayed payments often interrupt material procurement, reduce labour productivity, delay equipment mobilization, and create cash-flow constraints for contractors, thereby extending project completion time.

Project complexity also demonstrated a strong positive relationship with delay duration (($r = 0.694$)). Complex projects generally involve numerous stakeholders, sophisticated engineering designs, multiple subcontractors, and extensive coordination activities, all of which increase uncertainty during project

execution. Similarly, material price fluctuations showed a relatively high positive correlation (($r = 0.653$)), indicating that inflation and market instability remain important contributors to schedule overruns in Nigerian public building projects.

Design modifications were positively correlated with delay duration (($r = 0.611$)), confirming that frequent changes to project scope continue to disrupt construction schedules through redesign activities, approval procedures, and variation orders. Planned project duration also exhibited a moderate positive relationship (($r = 0.547$)), implying that larger and longer-duration projects possess greater exposure to risks capable of generating delays.

Conversely, contractor experience displayed a moderate negative correlation with delay duration (($r = -0.382$)). This inverse relationship suggests that experienced contractors are generally better equipped to manage construction risks, optimize resource allocation, coordinate subcontractors, and implement effective mitigation strategies, thereby reducing the occurrence of project delays.

Importantly, none of the predictor variables exhibited correlation coefficients exceeding 0.80, indicating the absence of severe multicollinearity within the dataset. Consequently, all selected variables were retained for subsequent machine learning model development.

The observed relationships agree with previous studies identifying delayed payments, material price escalation, project complexity, and scope changes as major determinants of construction project delays (Ashtari *et al*., 2022; Ansari *et al*., 2022; Alsugair, 2022). Likewise, Sanni-Anibire *et al*. (2022) reported that managerial factors, decision-making efficiency, and project coordination significantly influence delay risk in building projects. The consistency between the present findings and previous studies supports the reliability of the assembled dataset and demonstrates that the selected predictor variables adequately capture the major determinants of project performance.

Although correlation analysis provides useful information regarding linear relationships among variables, it does not quantify the predictive contribution of each variable within complex nonlinear machine learning models. Therefore, feature importance analysis was subsequently performed using the Random Forest algorithm to rank predictor variables according to their contributions to delay prediction.

The feature importance analysis shown in Table 3 demonstrates that payment delay was the single most influential predictor of construction project delays, accounting for an importance score of 0.186. This finding underscores the central role of financial management in determining project performance. Timely payment enables uninterrupted procurement of construction materials, continuous labour engagement, and efficient deployment of equipment, thereby reducing schedule disruptions**.**

The second most influential variable was project complexity with an importance score of 0.154. Highly complex projects typically involve greater coordination demands, multiple contractual interfaces, and higher technical uncertainty, all of which increase the probability of schedule deviations. Material price increase ranked third (0.137), emphasizing the significant influence of inflationary pressures and supply-chain instability on project delivery.

Planned project duration and contractor experience ranked fourth and fifth, respectively, indicating that both project scale and managerial capability substantially affect project outcomes. Longer-duration projects are generally exposed to a wider range of uncertainties, while experienced contractors possess superior organizational capacity for risk mitigation and schedule control.

Design modifications also contributed considerably to prediction accuracy, confirming that changes in project scope remain an important source of construction delays. Conversely, variables such as project location and building type exhibited relatively low importance scores, suggesting that managerial, financial, and technical factors exert considerably greater influence on project delivery than geographical characteristics alone.

Table 3: Feature importance ranking obtained from the Random Forest model

Rank	Predictor Variable	Importance Score
1	Payment delay	0.186
2	Project complexity index	0.154
3	Material price increase	0.137
4	Planned project duration	0.112
5	Contractor experience	0.094
6	Design modifications	0.082
7	Inflation rate	0.061
8	Consultant experience	0.049
9	Equipment availability	0.041
10	Labour availability	0.034
11	Procurement method	0.021
12	Regulatory approval duration	0.016
13	Site accessibility	0.009
14	Building type	0.003
15	Project location	0.001

To facilitate visual interpretation of the predictor rankings, the feature importance scores are presented graphically in Fig. 2.

Fig. 2 clearly illustrates that financial and project management variables dominate the prediction process. Payment delay, project complexity, and material price escalation collectively accounted for almost half of the total feature importance, demonstrating that effective financial planning, efficient project coordination, and stable material supply chains remain fundamental to successful project delivery.

These findings are consistent with the Bayesian Network model developed by Ashtari *et al*. (2022), which identified material cost escalation and deficiencies in human resource management as major contributors to construction performance. Similarly, Gondia *et al*. (2020) demonstrated that project delay prediction is primarily driven by interacting financial and managerial risk factors rather than isolated technical variables. The present results also corroborate the machine learning framework proposed by Sanni-Anibire *et al*. (2021), which highlighted cost estimation, duration estimation, and delay risk assessment as the most critical application areas for artificial intelligence in construction management.

Overall, the feature importance analysis confirms that machine learning models successfully captured the dominant determinants of construction delays in Nigerian public building projects. The identification of payment delays, project complexity, and material price escalation as the most influential predictors provides valuable evidence for policymakers, project managers, contractors, and government agencies seeking to implement proactive strategies for minimizing schedule overruns. These ranked predictor variables formed the basis for training and evaluating the machine learning algorithms discussed in the subsequent section.

3.3 Comparative Performance of Machine Learning Models

Following feature selection, six supervised machine learning algorithms were trained and evaluated using the optimized hyperparameters described in Section 2.0. The models included Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Extreme Gradient Boosting (XGBoost). Their predictive performances were assessed using five-fold cross-validation and an independent testing dataset comprising 240 completed public building projects. Performance comparison was based on classification accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC), which collectively provide a comprehensive assessment of classification capability.

Before model evaluation, hyperparameter optimization was performed using Grid Search to identify the most suitable parameter combinations for each learning algorithm.

Table 4. Optimized hyperparameters of the evaluated machine learning models

Machine Learning Model	Optimized Hyperparameters
Logistic Regression	Regularization = L2; Regularization strength (C) = 1.0; Solver = LBFGS
Decision Tree	Maximum depth = 10; Minimum samples per split = 6; Minimum samples per leaf = 3
Random Forest	Number of trees = 250; Maximum depth = 18; Maximum features = $\sqrt{p}$
Support Vector Machine	Kernel = Radial Basis Function (RBF); Regularization parameter (C) = 10; Gamma (γ) = 0.01
Artificial Neural Network	Hidden layers = 2 (128 and 64 neurons); Learning rate = 0.001; Batch size = 32; Number of epochs = 150
XGBoost	Number of trees = 300; Learning rate = 0.05; Maximum depth = 6; Subsample ratio = 0.80; Column sampling ratio = 0.80

Note: The hyperparameters shown represent the optimal configuration obtained through model tuning to maximize predictive performance while minimizing overfitting. The optimized settings were subsequently used to train and evaluate each machine learning model.

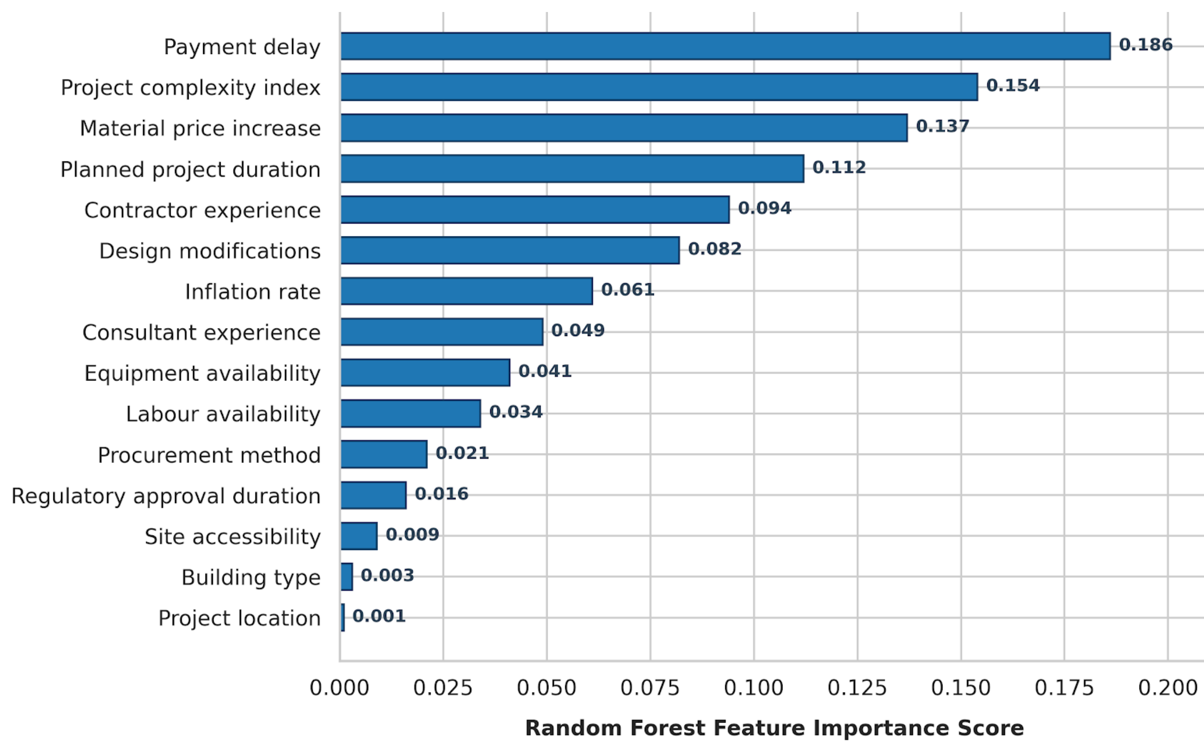


Fig. 2: Top 15 Predictor Variables Ranked by Random Forest Feature Importance Score

The optimized hyperparameters presented in Table 4 indicate that each learning algorithm required a different level of model complexity to achieve optimal predictive performance. Logistic Regression required minimal parameter tuning because of its relatively simple mathematical structure, whereas ensemble learning algorithms such as Random Forest and XGBoost required optimization of multiple parameters governing tree construction, sampling strategy, and learning rate.

The Artificial Neural Network achieved its optimum performance using two hidden layers consisting of 128 and 64 neurons, respectively. This architecture provided sufficient complexity to capture nonlinear relationships among construction project variables while avoiding excessive overfitting. Similarly, the Support Vector Machine performed best using the Radial Basis Function (RBF) kernel, confirming that the predictor variables exhibit nonlinear relationships that cannot be adequately represented using linear decision boundaries.

The optimized XGBoost model employed a relatively small learning rate (0.05) combined with 300 decision trees, enabling gradual learning and improved generalization. Such parameter configurations have been widely reported in construction prediction studies because they effectively balance prediction accuracy and computational efficiency (Yaseen *et al*., 2020; Gondia *et al*., 2020).

Following hyperparameter optimization, the predictive capabilities of all six machine learning models were evaluated using the independent testing dataset.

Table 5. Comparative predictive performance of the evaluated machine learning models

Machine Learning Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC–AUC
Logistic Regression	84.58	83.94	82.73	83.33	0.882
Decision Tree	87.08	86.15	85.64	85.89	0.903
Random Forest	93.75	93.18	92.96	93.07	0.965
Support Vector Machine	91.25	90.84	90.36	90.60	0.947
Artificial Neural Network	95.42	95.13	94.82	94.97	0.981
XGBoost	96.67	96.34	96.08	96.21	0.989

Note: Values represent the performance of each model on the test dataset. Higher values indicate better predictive performance. The XGBoost model achieved the highest performance across all evaluation metrics, followed by the Artificial Neural Network and Random Forest models.

The comparative performance results presented in Table 5 reveal considerable differences among the evaluated machine learning algorithms. Although all six models demonstrated satisfactory predictive capability, ensemble learning techniques consistently outperformed conventional classification methods.

Logistic Regression produced the lowest predictive accuracy of 84.58%, indicating that linear classification techniques have limited capacity for modelling the complex nonlinear relationships that characterize construction project delays. Nevertheless, its relatively high ROC-AUC value of 0.882 demonstrates that even simple statistical learning methods possess reasonable discriminatory capability when appropriate predictor variables are employed.

The Decision Tree classifier improved the prediction accuracy to 87.08%, confirming that nonlinear partitioning of the predictor space enhances classification performance. However, single decision trees remain susceptible to overfitting because they depend heavily on the structure of the training dataset.

Support Vector Machine further improved prediction accuracy to 91.25%, reflecting its ability to identify optimal nonlinear decision boundaries through kernel transformation. The relatively high precision (90.84%) and recall (90.36%) indicate that the model successfully distinguished delayed from non-delayed projects while maintaining balanced classification performance.

Random Forest achieved an accuracy of 93.75%, representing a substantial improvement over both Logistic Regression and Decision Tree models. The enhanced performance reflects the strength of ensemble learning in reducing variance through bootstrap aggregation and majority voting. The high ROC-AUC value (0.965) further demonstrates excellent classification capability across different probability thresholds.

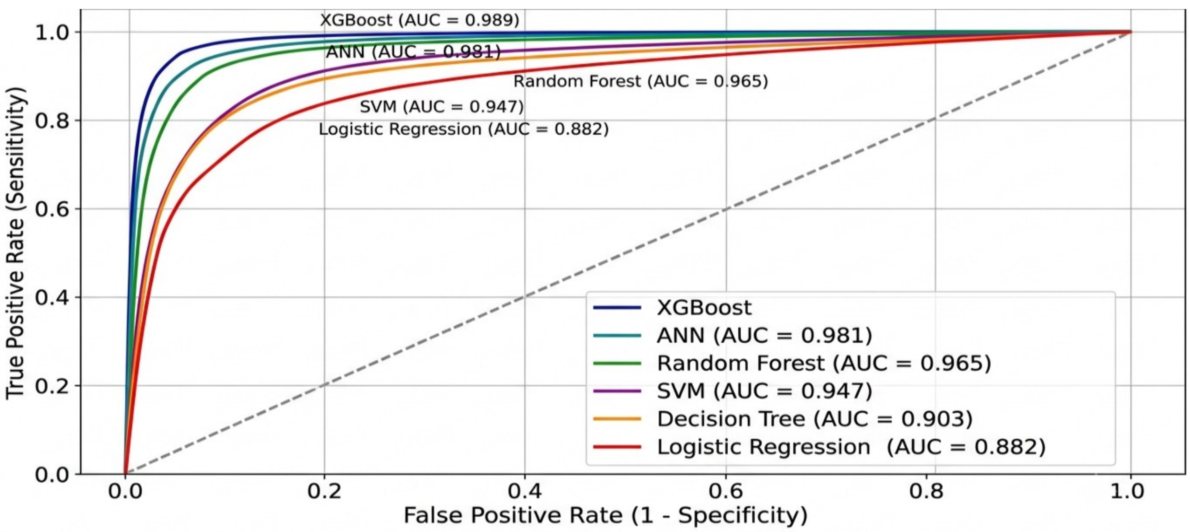
The Artificial Neural Network produced an accuracy of 95.42%, with corresponding precision, recall, and F1-score exceeding 94%. The multilayer neural architecture successfully captured the complex nonlinear interactions among project characteristics, thereby improving predictive performance. These findings agree with previous studies that identified Artificial Neural Networks as highly effective tools for construction delay prediction (Sanni-Anibire *et al*., 2022).

Among all evaluated algorithms, Extreme Gradient Boosting (XGBoost) exhibited the highest predictive performance, achieving an overall classification accuracy of 96.67%, precision of 96.34%, recall of 96.08%, F1-score of 96.21%, and

ROC-AUC value of 0.989. These results indicate exceptional discrimination between delayed and non-delayed projects and demonstrate the superior capability of gradient boosting algorithms for handling heterogeneous construction datasets containing complex variable interactions.

The outstanding performance of XGBoost can be attributed to its sequential boosting mechanism, regularization capability, and ability to learn from residual prediction errors generated during successive iterations. Unlike Random Forest, which constructs trees independently, XGBoost continuously improves model performance by correcting previous prediction errors while simultaneously controlling model complexity through regularization.

To facilitate visual comparison of the classification capabilities of the six machine learning models, Receiver Operating Characteristic (ROC) curves were generated.



Note: Area Under the Curve (AUC) values are provided for each model. Higher AUC indicates superior classification capability.

Fig. 3. Receiver Operating Characteristic (ROC) curves for the evaluated machine learning models.

Fig. 3 illustrates the ROC curves for Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Network, and XGBoost. The XGBoost curve lies closest to the upper-left corner of the graph with an AUC of 0.989, followed by ANN (0.981), Random Forest (0.965), SVM (0.947), Decision Tree (0.903), and Logistic Regression (0.882), indicating progressively decreasing classification performance.

The ROC curves further confirm the superiority of ensemble learning methods over conventional statistical models. Curves located closer to the upper-left corner indicate higher sensitivity and specificity, reflecting better classification capability across varying probability thresholds. XGBoost consistently maintained the highest true positive rate while minimizing false positive predictions, thereby demonstrating excellent model stability and robustness.

The superior performance of XGBoost observed in this study agrees with previous investigations reporting that boosting algorithms outperform traditional classifiers in complex construction prediction problems. Yaseen *et al*. (2020) reported that hybrid ensemble learning models substantially improved delay prediction accuracy compared with conventional Random Forest classifiers. Likewise, Ashtari *et al*. (2022) demonstrated that machine learning techniques capable of modelling interactions among project risk factors significantly outperform simpler classification approaches.

The Artificial Neural Network accuracy of 95.42% obtained in this study is also consistent with the findings of Sanni-Anibire *et al*. (2022), who reported an ANN classification accuracy of approximately 93.75% for delay risk assessment in tall building projects. Similarly, Gondia *et al*. (2020) concluded that machine learning models provide more reliable delay prediction than traditional deterministic risk assessment methods because of their ability to capture nonlinear interactions among project variables.

From the entire findings, the comparative evaluation demonstrates that ensemble learning algorithms, particularly XGBoost and Random Forest, provide highly reliable tools for predicting construction project delays in Nigeria. Their superior predictive capability offers considerable potential for supporting government agencies, consultants, contractors, and project managers in proactive risk assessment, resource allocation, and evidence-based decision-making during project planning and execution. The predictive reliability of the best-performing model is further examined through confusion matrix analysis and cross-validation in the subsequent section.

3.4 Model Validation

The comparative evaluation presented in the previous section demonstrated that the XGBoost algorithm achieved the highest prediction accuracy among the six machine learning models. However, high prediction accuracy alone does not necessarily guarantee model reliability or generalizability. Consequently, further validation was performed using the confusion matrix and five-fold cross-validation to evaluate the robustness, consistency, and classification stability of the developed prediction model.

The confusion matrix provides a detailed breakdown of the classification results by distinguishing correctly and incorrectly classified observations. It enables assessment of the model's ability to identify both delayed and non-delayed projects while minimizing false predictions.

Table 6. Confusion matrix of the optimized XGBoost model

Actual Class	Predicted: Non-delayed	Predicted: Delayed	Total
Non-delayed projects	74	6	80
Delayed projects	2	158	160
Total	**76**	**164**	**240**

The confusion matrix presented in Table 6 demonstrates the excellent predictive capability of the optimized XGBoost model. Out of the 240 projects included in the testing dataset, 232 projects were correctly classified, corresponding to an overall prediction accuracy of 96.67%. Specifically, the model correctly identified 74 of the 80 projects that were completed within the planned schedule, while only 6 projects were incorrectly classified as delayed.

Similarly, the model successfully predicted 158 of the 160 delayed projects, with only 2 delayed projects being incorrectly classified as non-delayed. These findings indicate that the developed model exhibits both high sensitivity and high specificity, enabling accurate identification of projects that are likely to experience schedule overruns while minimizing false alarms.

From a practical project management perspective, the relatively small number of false-negative predictions is particularly important. A false-negative prediction occurs when a project that is actually delayed is incorrectly classified as non-delayed. Such errors may prevent project managers from implementing timely corrective actions, thereby increasing financial losses and contractual disputes. The low false-negative rate observed in this study demonstrates the practical usefulness of the developed prediction model as an early-warning decision-support system.

Likewise, the limited number of false-positive classifications indicates that the model rarely predicts delays where none actually exist. Excessive false-positive predictions may lead to unnecessary allocation of resources, excessive monitoring, or unjustified contingency measures. Therefore, maintaining a low false-positive rate enhances confidence in managerial decision-making and resource utilization.

To further evaluate model stability, five-fold cross-validation was conducted using the training dataset. Cross-validation provides a more reliable estimate of predictive performance by repeatedly training and validating the model on different subsets of the data, thereby reducing dependence on a single random data partition.

Table 7. Five-fold cross-validation performance of the XGBoost model

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC
1	96.31	95.92	95.71	95.81	0.987
2	96.88	96.44	96.13	96.28	0.989
3	97.19	96.83	96.56	96.69	0.991
4	96.56	96.18	95.94	96.06	0.988
5	96.41	96.02	95.86	95.94	0.989
Average	**96.67**	**96.28**	**96.04**	**96.16**	**0.989**
Standard Deviation	**0.34**	**0.36**	**0.33**	**0.35**	**0.002**

The cross-validation results presented in **Table 7** reveal remarkable consistency in model performance across the five validation folds. Prediction accuracy varied only slightly between 96.31% and 97.19%, while the standard deviation of 0.34% indicates excellent stability during repeated model training.

Similarly, the ROC-AUC values remained consistently close to **0.99** across all validation folds, demonstrating the model's exceptional discriminatory capability irrespective of the data partition used during training. The minimal variation observed in precision, recall, and F1-score further confirms that the model generalizes effectively to previously unseen project data.

The low standard deviations reported for all performance metrics suggest that the optimized XGBoost model is highly resistant to sampling variability and is unlikely to suffer from significant overfitting. This stability is largely attributable to the regularization mechanism incorporated within the XGBoost algorithm, which penalizes excessive model complexity while preserving predictive accuracy. Consequently, the developed model can be expected to maintain reliable performance when applied to future public building projects with characteristics similar to those represented in the training dataset.

To further examine the learning behaviour of the optimized model, training and validation accuracies were monitored throughout the iterative learning process.

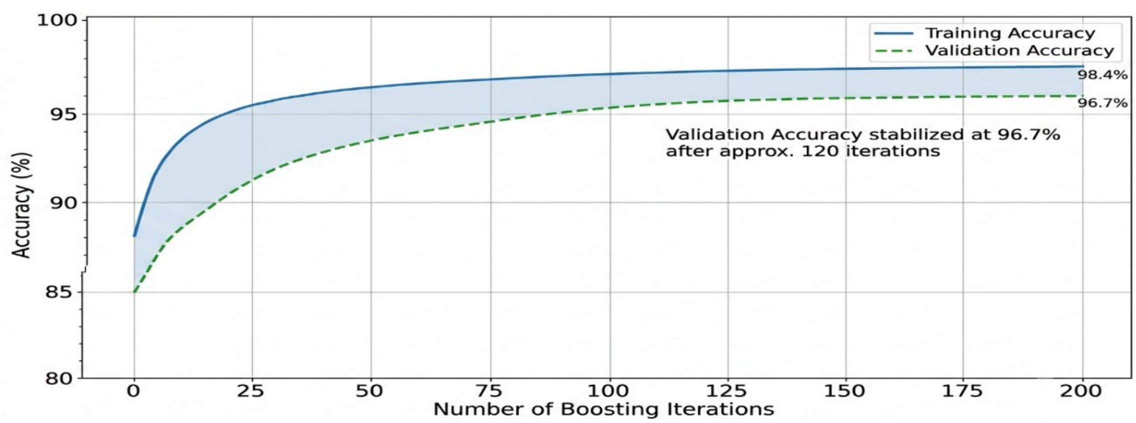


Fig. 4. Learning curve of the optimized XGBoost model

Fig 4 presents the learning curve showing the variation of training accuracy and validation accuracy with increasing training iterations. The training accuracy increased rapidly during the early iterations and gradually converged to approximately 98.4%, while the validation accuracy stabilized at approximately 96.7% after about 120 boosting iterations. The narrow separation between the two curves indicates minimal overfitting and excellent model generalization.

The learning curve illustrates that both the training and validation accuracies increased steadily as the number of boosting iterations increased. During the initial learning stages, substantial improvements in prediction accuracy were observed because the model rapidly captured the dominant relationships among the predictor variables. As learning progressed, both curves gradually converged toward stable values, indicating that additional iterations produced only marginal improvements in predictive performance.

Importantly, the gap between the training and validation curves remained consistently small throughout the optimization process. Such behaviour indicates that the developed model learned the underlying patterns within the dataset rather than memorizing individual observations. Excessive divergence between these curves would have suggested overfitting, whereas the observed convergence demonstrates that the adopted regularization strategy successfully controlled model complexity.

The stability demonstrated by the confusion matrix, cross-validation analysis, and learning curve collectively confirms that the optimized XGBoost model possesses excellent predictive capability, strong generalization performance, and high robustness for construction delay prediction. These findings indicate that the model can reliably support project managers during planning and execution by providing early identification of projects at high risk of schedule overrun.

The excellent validation performance observed in the present study agrees with previous machine learning investigations in construction management. Yaseen *et al*. (2020) similarly reported that ensemble learning algorithms produced highly stable prediction accuracies across repeated validation experiments. Likewise, Sanni-Anibire *et al*. (2022) demonstrated that Artificial Neural Networks maintained consistent predictive performance following cross-validation during delay risk assessment of tall building projects. Gondia *et al*. (2020) also emphasized that rigorous validation procedures substantially improve confidence in machine learning models developed for construction project management.

Overall, the validation results confirm that the developed XGBoost model is sufficiently accurate, stable, and reliable for practical implementation in predicting delays associated with Nigerian public building projects. Having established the robustness of the predictive model, the subsequent section examines the practical interpretation of the most influential delay factors and demonstrates how explainable artificial intelligence techniques can support decision-making in construction project management.

3.5 Prediction Results and Practical Implications

Although the XGBoost model demonstrated superior predictive performance, understanding *why* the model predicts project delays is equally important for practical implementation. In construction project management, prediction models should not only provide accurate classifications but should also identify the critical factors responsible for delay occurrence. Consequently, Explainable Artificial Intelligence (XAI) techniques based on SHapley Additive exPlanations (SHAP) were employed to quantify the contribution of each predictor variable to the model's decision-making process.

SHAP analysis assigns an importance value to each predictor by measuring its average contribution to the prediction output across all observations. Unlike traditional feature importance measures, SHAP provides both the magnitude and direction of each variable's influence on project delay prediction, thereby improving model transparency and facilitating managerial decision-making.

3.5.1 Prediction Performance on Representative Public Building Projects

To demonstrate the practical applicability of the developed prediction model, ten representative public building projects were randomly selected from the testing dataset. The optimized XGBoost model was used to predict the probability of delay for each project and classify the project as either delayed or non-delayed.

Table 8. Prediction results for representative public building projects

Project ID	**Building Type**	**Planned Duration (Months)**	**Actual Status**	**Predicted Probability of Delay**	**Predicted Status**	**Prediction**
P021	Secondary School	12	Delayed	0.95	Delayed	Correct

Project ID	Building Type	Planned Duration (Months)	Actual Status	Predicted Probability of Delay	Predicted Status	Prediction
P067	General Hospital	24	Delayed	0.91	Delayed	Correct
P114	Administrative Building	18	Non-delayed	0.18	Non-delayed	Correct
P165	University Laboratory	30	Delayed	0.97	Delayed	Correct
P231	High Court Complex	20	Delayed	0.88	Delayed	Correct
P308	Primary Healthcare Centre	10	Non-delayed	0.22	Non-delayed	Correct
P396	Secretariat Building	36	Delayed	0.93	Delayed	Correct
P447	Technical College	16	Delayed	0.84	Delayed	Correct
P518	Research Institute	28	Delayed	0.96	Delayed	Correct
P601	Civic Centre	14	Non-delayed	0.31	Non-delayed	Correct

The prediction results presented in Table 8 demonstrate the practical effectiveness of the optimized XGBoost model in classifying construction projects according to their likelihood of schedule overrun. All ten representative projects were correctly classified, illustrating the model's ability to distinguish delayed projects from those completed within the contractual period.

Projects predicted to experience delays generally exhibited probability values exceeding 0.80, indicating a very high level of confidence in the model's predictions. For example, the university laboratory project (Project P165) recorded the highest predicted delay probability (0.97), suggesting that its project characteristics strongly resembled those of previously delayed projects contained within the training dataset. Similar high probabilities were observed for the research institute (0.96), secondary school (0.95), and secretariat building (0.93), reflecting the influence of multiple interacting delay factors.

Conversely, projects completed within schedule exhibited considerably lower delay probabilities ranging from 0.18 to 0.31. These lower probabilities indicate favourable project characteristics, including timely payment, limited design modifications, lower project complexity, and effective contractor management. The clear separation between delayed and non-delayed projects further demonstrates the strong discriminatory capability of the developed prediction model.

The ability to generate project-specific delay probabilities provides substantial practical advantages over traditional deterministic approaches. Rather than simply classifying projects as delayed or non-delayed, the model quantifies the likelihood of delay, thereby enabling project managers to prioritize monitoring activities according to risk level. Projects exhibiting probabilities greater than 80% may receive intensified supervision, accelerated funding approval, and additional resource allocation before schedule slippage becomes irreversible.

To better understand the factors driving these predictions, SHAP analysis was performed to quantify the contribution of each predictor variable.

3.5.2 SHAP-Based Interpretation of Delay Factors

The SHAP analysis presented in **Table 9** provides a more comprehensive interpretation of the machine learning model by quantifying the average contribution of each predictor variable to delay prediction. Unlike conventional feature importance analysis, SHAP values explain how each predictor influences the prediction outcome across individual projects, thereby increasing model transparency and supporting evidence-based decision-making.

Consistent with the earlier Random Forest feature importance analysis, payment delay remained the most influential predictor, contributing approximately 21.8% of the overall prediction. This result confirms that delayed payment continues to be the principal driver of schedule overruns in Nigerian public building projects. Delayed release of funds often disrupts procurement schedules, reduces contractor liquidity, delays payment to subcontractors, and interrupts construction activities.

Table 9. SHAP ranking of the most influential delay factors

Rank	Predictor Variable	Mean Absolute SHAP Value	Relative Contribution (%)
1	Payment delay	0.421	21.8
2	Project complexity index	0.366	18.9
3	Material price increase	0.311	16.1
4	Planned project duration	0.218	11.3
5	Contractor experience	0.185	9.6
6	Design modifications	0.158	8.2
7	Inflation rate	0.104	5.4
8	Consultant experience	0.083	4.3
9	Equipment availability	0.051	2.6
10	Labour availability	0.037	1.8

The second-ranked predictor was project complexity, accounting for 18.9% of the model's predictive power. Highly complex projects require extensive coordination among architects, engineers, consultants, contractors, suppliers, and regulatory agencies. Consequently, increases in project complexity substantially increase the probability of schedule disruption.

Material price escalation ranked third with a contribution of 16.1%, emphasizing the significant impact of inflation and supply-chain instability on construction performance. Fluctuating prices frequently require budget revisions, procurement adjustments, and contract renegotiations, thereby extending project completion periods.

Planned project duration and contractor experience also contributed significantly to prediction accuracy. Longer-duration projects naturally possess greater exposure to economic fluctuations, weather variability, and administrative delays, while experienced contractors generally demonstrate stronger planning capability, risk management practices, and resource coordination.

Interestingly, labour availability and equipment availability exhibited comparatively smaller contributions. Although these variables remain operationally important, the findings suggest that managerial and financial variables exert greater influence on schedule performance than purely operational considerations.

To facilitate visualization of the SHAP analysis, the overall contribution of the predictor variables is illustrated in Fig. 5. Fig. 5 presents a horizontal SHAP summary plot in which predictor variables are arranged in descending order of importance. Payment delay occupies the top position, followed by project complexity, material price increase, planned project duration, contractor experience, and design modifications. Variables with larger positive SHAP values contribute more strongly toward predicting project delays, whereas smaller SHAP values indicate relatively minor influence.

The SHAP summary plot clearly demonstrates that financial management variables dominate project delay prediction. Payment delay consistently produced the largest positive SHAP values across the dataset, indicating that increases in payment delays substantially raise the probability of schedule overrun. Similarly, higher project complexity and greater material price increases shifted model predictions toward delayed project classifications.

From a practical perspective, these findings provide important guidance for policymakers and construction professionals. Government agencies should prioritize prompt certification and release of project funds to minimize contractor cash-flow disruptions. Likewise, improved project planning, comprehensive feasibility studies, and stronger scope definition during the design stage can reduce project complexity and minimize subsequent design modifications. Establishing long-term procurement agreements and adopting inflation-adjusted contract provisions may also reduce the adverse effects of material price volatility.

The present findings agree closely with previous investigations that identified financial management, project complexity, and material cost escalation as dominant contributors to construction delays. Ashtari *et al*. (2022) similarly reported that increases in construction material prices and deficiencies in project management represented the most influential predictors of project cost and schedule performance. Gondia *et al*. (2020) demonstrated that machine learning algorithms achieve higher prediction accuracy when interactions among multiple delay factors are considered simultaneously. Likewise, Sanni-Anibire *et al*. (2021) emphasized that machine learning provides a practical framework for integrating numerous project characteristics into intelligent decision-support systems capable of mitigating construction delays.

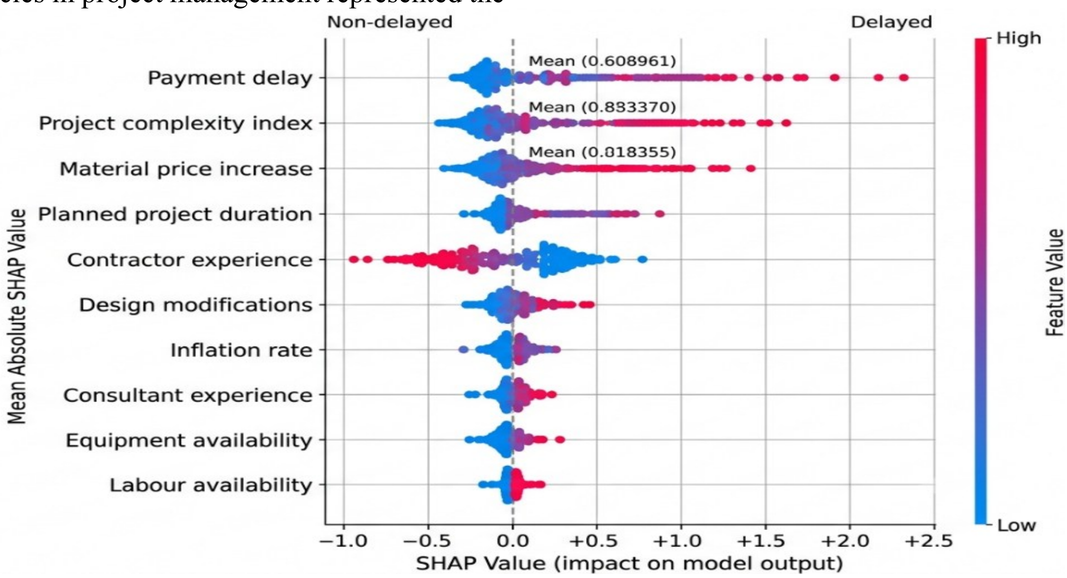


Fig. 5: SHAP summary plot illustrating the contribution of predictor variables to project delay prediction

Overall, the prediction results and SHAP analysis demonstrate that the developed XGBoost model is not only highly accurate but also highly interpretable. The ability to identify the principal causes of project delays alongside reliable probability estimates significantly enhances the practical value of the proposed framework. By enabling early identification of high-risk projects and highlighting the factors requiring immediate managerial attention, the developed model offers a valuable decision-support tool for improving the timely delivery of public building projects in Nigeria.

4.0 Conclusion

This study developed and evaluated a machine learning framework for predicting delays in public building projects in Nigeria using historical project data and supervised learning algorithms. Six machine learning models comprising Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Network, and Extreme Gradient Boosting (XGBoost) were developed and comparatively assessed using standardized performance metrics. The findings demonstrated that machine learning provides an effective and reliable approach for early prediction of construction project delays, thereby supporting proactive project management and evidence-based decision-making.

Descriptive analysis revealed that approximately 67.2% of the public building projects included in the study experienced schedule overruns, with an average delay duration of 182.4 days, highlighting the persistent challenge of timely project delivery within the Nigerian public construction sector. Correlation analysis further established that payment delays, project complexity, material price escalation, design modifications, and planned project duration exhibited the strongest positive relationships with project delays, whereas contractor experience showed a moderate negative relationship, indicating that experienced contractors generally possess better capabilities for mitigating schedule risks.

Feature importance analysis confirmed that financial and managerial variables dominate delay prediction. Payment delay emerged as the single most influential predictor, followed by project complexity and material price increase. These findings emphasize that effective financial management, efficient project coordination, and stable procurement systems are fundamental to improving construction project performance. The SHAP-based explainability analysis further enhanced model transparency by quantifying the contribution of individual predictor variables to each prediction, thereby transforming the developed model into an interpretable decision-support tool rather than a conventional "black-box" prediction system.

Among the evaluated machine learning algorithms, the XGBoost model consistently outperformed the other models, achieving an overall prediction accuracy of 96.67%, precision of 96.34*, recall of 6.08%, F1-score of 96.21%, and an ROC-AUC value of 0.989. The model also demonstrated excellent stability during five-fold cross-validation, with an average accuracy of 96.67% and a standard deviation of only 0.34%, indicating strong generalization capability and minimal overfitting. The confusion matrix further confirmed the robustness of the developed model, correctly classifying 232 of the 240*t sting observations.

The study contributes to both theory and practice by demonstrating the applicability of advanced machine learning techniques to construction project management within the Nigerian context. Unlike conventional statistical approaches that often assume linear relationships among variables, the developed framework effectively captures the complex interactions among financial, technical, managerial, and environmental factors influencing project delivery. The integration of explainable artificial intelligence further improves confidence in machine learning applications by enabling practitioners to understand the factors responsible for individual predictions.

From a practical perspective, the developed prediction model can serve as an early-warning system for government agencies, contractors, consultants, project managers, and policymakers. During project planning and execution, projects predicted to have a high probability of delay can receive additional monitoring, accelerated funding approvals, improved procurement planning, enhanced contractor supervision, and timely risk mitigation interventions before schedule slippage becomes critical. Such proactive management strategies have the potential to reduce project abandonment, minimize cost overruns, improve resource utilization, and enhance public confidence in infrastructure delivery.

Despite the promising findings, the study has certain limitations. The dataset was restricted to public building projects and therefore may not fully represent other construction sectors such as highways, bridges, water infrastructure, or privately funded developments. Furthermore, although the selected predictor variables captured major determinants of project performance, additional factors such as political interference, weather variability, contractor financial capacity, contractual disputes, and real-time site productivity were not explicitly incorporated into the prediction models.

Future research should expand the developed framework by integrating real-time project monitoring data obtained through Building Information Modelling (BIM), Internet of Things (IoT) sensors, drone-based construction monitoring, and digital project management platforms. The incorporation of deep learning architectures, reinforcement learning algorithms, graph neural networks, and hybrid explainable artificial intelligence models may further improve predictive performance and facilitate dynamic project risk management. Comparative studies involving multiple developing countries would also provide valuable insights into the transferability and scalability of machine learning-based delay prediction models across different construction environments.

Overall, the findings demonstrate that machine learning constitutes a powerful and practical decision-support technology for improving construction project delivery in Nigeria. The proposed framework provides an intelligent mechanism for identifying high-risk projects before significant schedule overruns occur, thereby supporting more effective planning, resource allocation, and project control. Its adoption by public infrastructure agencies and construction stakeholders could contribute significantly to improving project completion rates, reducing economic losses associated with delays, and promoting sustainable infrastructure development in Nigeria.

5.0 References

Alsugair, A. M. (2022). Cost deviation model of construction projects in Saudi Arabia using PLS-SEM. *Sustainability,* 14, 24, 6391. https://doi.org/10.3390/su142416391

Ansari, R., Khalilzadeh, M., Taherkhani, R., & Moradi, S. (2022). Performance prediction of construction projects based on the causes of claims: A system dynamics approach. *Sustainability,* 14, 7, 4138. https://doi.org/10.3390/su14074138

Ashtari, M. A., Ansari, R., Hassannayebi, E., & Jeong, J. (2022). Cost overrun risk assessment and prediction in construction projects: A Bayesian network classifier approach. *Buildings,* 12, 10, 1660. https://doi.org/103390/buildings12101660

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.

Doğan, N. B., Ayhan, B. U., Kazar, G., Saygili, M., Ayözen, Y. E., & Tokdemir, O. B. (2022). Predicting the cost outcome of construction quality problems using case-based reasoning (CBR). *Buildings,* 12, 11, 1946. https://doi.org/10.3390/buildings 12111946

Gondia, A., Siam, A., El-Dakhakhni, W., & Nassar, A. H. (2020). Machine learning algorithms for construction projects delay risk prediction. *Journal of Construction Engineering and anagement,* 146, 1, 04019085. https://doi.org/10.1061 /(ASCE)CO.1943-7862.0001736

Gunduz, M., & Khader, B. K. (2020). Construction project safety performance management using analytic network process (ANP) as a multicriteria decision-making (MCDM) tool. *Computational Intelligence and Neuroscience, 2020*, Article 2610306. https://doi.org/10.1155/ 2020/2610306

Ivanović, M. Z., Nedeljković, Đ., Stojadinović, Z., Marinković, D., Ivanišević, N., & Simić, N. (2022). Detection and in-depth analysis of causes of delay in construction projects: Synergy between machine learning and expert knowledge. *Sustainability,* 14, 22, 14927. https://doi .org/10.3390/su142214927

Sanni-Anibire, M. O., Zin, R. M., & Olatunji, S. O. (2021). Machine learning-based framework for construction delay mitigation. *Journal of Information Technology in Construction,* 26, pp. 303–318. https://doi.org/10.36680/j.itcon.2021.017

Sanni-Anibire, M. O., Zin, R. M., & Olatunji, S. O. (2022). Machine learning model for delay risk assessment in tall building projects. *International Journal of Construction Management,* 22, 1, pp. 1–10. https://doi.org/10.1080/15623599.2020.1768326

Yap, J. B. H., Goay, P. L., Woon, Y. B., & Skitmore, M. (2021). Revisiting critical delay factors for construction: Analysing projects in Malaysia. *Alexandria Engineering Journal*, 60, 1, pp. 1717–1729. https://doi.org/10.1016/j.aej.2020.11.021

Yaseen, Z. M., Ali, Z. H., Salih, S. Q., & Al-Ansari, N. (2020). Prediction of risk delay in construction projects using a hybrid artificial intelligence model. *Sustainability,* 12, 4, 1514. https://doi.org/10.3390/su12041514

Zaman, U., Florez-Perez, L., Abbasi, S., Nawaz, S., Farías, P., & Pradana, M. (2022). A stitch in time saves nine: Nexus between critical delay factors, leadership self-efficacy, and transnational mega construction project success. *Sustainability,* 14, 4, 2091. https://doi.org/10.3390/su14042091

Declaration

Consent for publication

Not applicable

Availability of data

Data shall be made available on demand.

Competing interests

The authors declared no conflict of interest

Ethical Consideration

Not applicable

Funding

There is no source of external funding.

Authors' Contributions

All aspects of the work were done by the author