Available online at www.CivileJournal.org

Civil Engineering Journal

(E-ISSN: 2476-3055; ISSN: 2676-6957)

Vol. 12, No. 04, April, 2026

Application of XGBoost in Road Maintenance Cost Prediction

Dian Setiawan [1], Leksmono S. Putranto [1], Endah Murtiana Sari [2*]

[1] *Department of Civil Engineering, Universitas Tarumanagara, Jakarta Barat 11440, Indonesia.*

[2] *Department of Industrial Engineering, Universitas Sains Indonesia, Bekasi, West Java, Indonesia.*

Received 05 January 2026; Revised 11 March 2026; Accepted 16 March 2026; Published 01 April 2026

Abstract

Road maintenance costs play a critical role in government budgeting, as they represent a recurring expenditure required to sustain transportation infrastructure performance and traffic safety. Accurate cost prediction enables long-term efficiency by ensuring that maintenance budgets are allocated appropriately. This study aims to develop a predictive model for road maintenance cost using the Extreme Gradient Boosting (XGBoost) algorithm, optimized through iterative training to improve prediction accuracy based on deviations between predicted and actual costs. Model performance was evaluated using Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2), all of which indicate a strong model fit and high predictive reliability. The model was developed using simulated and empirical data from 30 road sections with varying characteristics, incorporating key predictors such as road length, cold mix asphalt, asphalt emulsion, diesel fuel, gasoline, water consumption, working area, asphalt removal volume, and labor requirements. The results demonstrate that the proposed XGBoost-based model can effectively estimate maintenance costs and associated resource requirements. The findings provide practical insights for government agencies in planning material usage and workforce allocation for road maintenance activities.

Keywords: XGBoost; Road Maintenance; Cost Prediction; Government Project.

1. Introduction

An appropriate road maintenance strategy is essential for controlling maintenance costs, maintaining operational quality, and ensuring traffic safety, particularly for governments as providers of public transportation infrastructure [1, 2]. The road maintenance cycle typically begins with field inspections, followed by condition diagnosis, maintenance decision-making, and the execution of maintenance actions [3, 4]. In Indonesia, many road infrastructure projects were constructed during periods of rapid economic growth to support regional economic development. Consequently, road maintenance has become a routine activity that must be continuously undertaken by the government to ensure uninterrupted economic performance. However, road maintenance planning is often characterized by uncertainty in material quantities and labor requirements, resulting in inaccurate budget planning and inefficient allocation of maintenance funds [5, 6]. Bogor City is a densely populated area with heavy road usage, with numerous transportation routes, as it serves as the city's main thoroughfare. Therefore, predicting and ensuring the quality of road services in Bogor is crucial. Good quality roads boost the economy, ensuring timely and cost-effective road maintenance [7].

Previous studies on road maintenance estimation have primarily focused on predicting pavement condition levels using accumulated inspection data. While such studies provide valuable insights, they are often limited in delivering detailed estimations of the specific maintenance actions required [3, 6, 7]. Several studies have estimated overall road

* Corresponding author: endah.murtiana@sains.ac.id

https://doi.org/10.28991/CEJ-2026-012-04-018

conditions or major components, such as material demand, workshop facilities, and labor requirements, yet they have not comprehensively examined specific damage characteristics at the project level [6, 8, 9]. Although representative component-based models have been developed for certain structural types within road maintenance management practices, their limited scope reduces their suitability for detailed maintenance planning [9]. Furthermore, many existing studies have failed to adequately capture the complex interactions between pavement damage mechanisms and road maintenance costs [10].

Most previous research has relied on regression-based approaches that primarily focus on identifying influential factors but do not provide robust recommendations for predicting road maintenance costs based on material consumption, labor input, and other dominant cost drivers. Sari et al. [11] identified several critical factors that should be considered in road maintenance activities, including material usage, energy consumption, residual materials, road utilization, emissions and pollution, supply chain aspects, health and safety, and labor requirements. However, current road maintenance research still tends to underemphasize these cost-forming factors, particularly material, energy, supply chain, residuals, health and safety, and labor aspects. Studies employing analytical techniques such as SPSS or SEM-PLS are generally limited to factor identification and do not explicitly simulate maintenance cost behavior or quantify the influence of predictors on total maintenance cost [9]. Research that combines artificial intelligence (AI) in civil engineering design applications is still very low, especially for predicting road maintenance costs. The XGBoost application is widely used for various purposes in predicting costs for planning public buildings, agriculture, fisheries, and other purposes because it is considered to have accuracy in providing simulation results, so if this is done to predict maintenance costs, it will be an important recommendation in calculating and planning road maintenance costs, especially in government projects that currently only use traditional methods that have not yet led to predictions with important predictors, namely road construction materials and the number of workers.

In addition, the application of artificial intelligence techniques for predicting road maintenance costs remains relatively limited [12]. Chen et al. [8] applied XGBoost modeling to predict asphalt pavement performance; however, their approach was unable to capture long-term pavement distress trends and exhibited limited accuracy when applied to datasets with high variability. These limitations highlight the need for more robust data-driven approaches capable of modeling complex and nonlinear relationships between maintenance cost drivers.

This study focuses on predicting road maintenance costs in the form of cost estimation. In this context, estimation aims to develop a model that computes target values while identifying relationships between predictors and the target variable, whereas prediction primarily emphasizes obtaining accurate target values without explicitly interpreting these relationships. As a tree-based machine learning method, Extreme Gradient Boosting (XGBoost) has been widely and successfully applied to predict condition levels across various domains due to its high accuracy and computational efficiency. Therefore, this study applies the XGBoost algorithm to develop a predictive model for road maintenance cost estimation.

2. Literature Review

2.1. Road Maintenance Cost Estimation

Road maintenance cost estimation must be prepared in a more detailed manner because it involves government budgets that must be accounted for transparently and continuously over time [13, 14]. In particular, material and labor costs are among the most critical components of road maintenance expenditure, as they directly support all maintenance activities and dominate total project costs [15, 16]. Therefore, road maintenance cost estimation should primarily focus on accurately estimating material quantities and labor requirements to ensure effective financial planning and budget allocation.

Several road maintenance methods discussed above represent conventional approaches that mainly consider historical cost data, standard unit prices, and expert judgment, while often neglecting variations in pavement condition severity, project scale, and operational efficiency [17, 18]. As a result, conventional estimation methods tend to produce static cost values that are less responsive to complex and dynamic maintenance conditions. Consequently, there is a growing need for innovative estimation methods that can predict road maintenance costs more accurately by capturing nonlinear relationships among cost drivers.

Machine learning-based approaches have recently been introduced to address the limitations of conventional methods. These approaches utilize large datasets and multiple predictors to model complex interactions between maintenance activities and associated costs [19, 20]. Several studies have demonstrated that machine learning techniques outperform traditional estimation methods in terms of accuracy, adaptability, and robustness when applied to infrastructure cost estimation problems [21, 22]. A comparison between conventional and machine learning-based road maintenance cost estimation methods is presented in Table 1.

Table 1. Comparation Method of road maintenance cost prediction

Aspect	Traditional Method	Machine Learning Method
Data Dependency	Relies on historical averages and unit prices	Utilizes large datasets and multiple predictors
Relationship Modeling	Assumes linear or simplified relationships	Captures nonlinear and complex interactions
Adaptability	Limited adaptability to changing conditions	High adaptability to diverse road conditions
Prediction accuracy	Moderate, often project-specific	High, data-driven and generalizable
Handling uncertainty	Limited	Robust handling of uncertainty and variability
Computational complexity	Low	Moderate to high

Table 1 describes the advantages of machine learning compared to conventional methods in predicting road maintenance costs. It can be seen that the machine learning method has various advantages, including being more suitable for datasets with multiple predictors, being able to predict nonlinear models, being more reliable, and having high validity.

2.2. Identification of Factors Influencing Road Maintenance Cost

Previous studies have identified various factors that influence road maintenance costs, such as pavement condition, traffic volume, material type, environmental exposure, labor intensity, and maintenance strategy selection [23, 24]. More recent studies have expanded these factors by incorporating sustainability and environmental considerations to better reflect the complex nature of road maintenance cost formation [25, 26].

Existing road maintenance cost estimation models can generally be classified into three categories: deterministic models, stochastic models, and artificial intelligence-based models [27]. Representative deterministic models include regression-based approaches, where maintenance costs are analyzed as functions of explanatory variables such as material quantities, labor input, and pavement condition indicators. For example, several studies have applied multiple regression techniques to analyze the relevance of material consumption and labor productivity in determining road maintenance costs [28]. However, regression-based methods have inherent limitations, as they often ignore random errors caused by unobserved independent variables and fail to capture nonlinear relationships [29].

As an alternative, stochastic models such as Markov chain approaches have been widely used to model pavement deterioration processes and maintenance decision-making under uncertainty [30]. Although these models are effective for long-term planning and life-cycle analysis, their aggregated nature limits their applicability for detailed project-level maintenance cost estimation.

A growing body of research has applied artificial intelligence techniques to overcome these limitations. For instance, Melhem et al. and Morcous employed decision tree-based models to determine influential factors affecting infrastructure maintenance decisions and costs, demonstrating improved predictive capability compared to conventional regression approaches [31, 32]. Using rough set theory and machine learning techniques, Huang et al. reported high prediction accuracy in infrastructure cost-related studies, although their focus was primarily on prediction outcomes rather than detailed factor interpretation [33].

Despite these advances, many existing studies still emphasize factor identification rather than direct cost simulation. Furthermore, research applying statistical tools such as SPSS or SEM-PLS generally remains descriptive and does not explicitly model maintenance cost behavior or quantify the combined influence of predictors on total cost [34]. The application of machine learning techniques, particularly advanced ensemble models, for road maintenance cost prediction remains limited. For example, Chen et al. applied XGBoost to predict asphalt pavement performance; however, their model was unable to capture long-term pavement distress trends and showed reduced accuracy for datasets with high variability [8]. This gap highlights the need for robust machine learning-based models capable of accurately predicting road maintenance costs by incorporating material, labor, and operational factors simultaneously.

2.3. Road Maintenance Management in Indonesia

Road maintenance management in Indonesia is governed by Law No. 2 of 2022 and the Regulation of the Ministry of Public Works and Housing (PUPR) No. 13/PRT/M/2011. The primary objective of this regulatory framework is to maintain road serviceability, commonly referred to as Kondisi Mantap, defined as roads being in good or fair condition, in order to achieve their planned service life and ensure sustainable infrastructure performance [35, 36]. The responsibility for road management in Indonesia is divided according to administrative authority, as follows:

- ***National Roads***: Managed by the Central Government through the Ministry of Public Works and Housing (PUPR), implemented by the Directorate General of Highways via National Road Implementation Agencies.
- ***Provincial Roads***: Managed by Provincial Governments.
- ***Regency/Municipal Roads***: Managed by Regency or Municipal Governments.

This division of authority is intended to ensure that road maintenance planning and implementation are aligned with administrative capacity, funding availability, and regional development priorities. Each level of government is responsible for planning, budgeting, and executing maintenance activities in accordance with applicable technical standards and performance targets [37]. Road maintenance activities in Indonesia are generally classified based on their characteristics and objectives, as summarized in Table 2. Table 2 describes the various types of road maintenance according to the objectives carried out.

Table 2. Characteristics and Objectives of Road Maintenance

Type of Maintenance	Description	Example Activities
Routine Maintenance	Conducted continuously throughout the year to prevent minor damage from developing into more severe deterioration.	Pothole patching, drainage cleaning, roadside vegetation control.
Periodic Maintenance	Performed at specific intervals to restore road serviceability and structural capacity.	Asphalt overlay, base layer repair.
Rehabilitation	Applied to restore road conditions that have experienced significant structural damage.	Structural treatment at locations with substantial functional degradation.
Reconstruction	Involves upgrading or replacing the entire road structure.	Removal of existing asphalt layers and replacement with new asphalt or concrete pavement.

3. Research Methodology

The first step in developing the road maintenance cost estimation model is data collection from the Public Works and Housing Agency (Dinas PUPR) of Bogor City, West Java, Indonesia. The dataset consists of records from 30 road sections with diverse physical characteristics, maintenance scopes, and operational conditions, collected over the period 2022–2024. This dataset represents a combination of routine and non-routine maintenance activities conducted under varying road conditions.

3.1. Data Collection

The collected data include key project attributes relevant to road maintenance activities, such as road section length, types and quantities of materials used, working area size, project duration, and labor requirements. These variables were selected because they directly reflect the scale, complexity, and resource intensity of road maintenance operations and are commonly used as indicators in infrastructure cost estimation studies. Before model development, the raw data were subjected to a preprocessing stage that included data reduction, data cleaning, transformation, and integration. Data reduction was applied to remove redundant or irrelevant records, while data cleaning addressed missing values and inconsistencies. Data transformation and integration were conducted to ensure uniform measurement units and consistency across all variables. To improve the performance of the estimation model, relevant features were then selected based on correlation analysis, and the data were categorized according to their functional roles in road maintenance activities.

3.2. Data Analysis Using Mathematical Modeling

For each variable X, a modeling process is conducted to establish the relationship between the dependent variable Y, which represents the construction cost of the final waste processing project, and each corresponding independent variable. For each, 3 models were evaluated, including:

- Linear Model as represented in Equation 1.

$$\hat{y} = \alpha . x + b \quad (1)$$

with α is a slope or coefficient of the model and b is the intercept.

- The second order polynomial model as represented in Equation 2.

$$\hat{y} = \alpha_2 x^2 + \alpha_1 x + b \quad (2)$$

- Exponential (log-linear) as expressed in Equation 3.

$$\hat{y} = \alpha e^{bx} \tag{3}$$

In the computational process conducted in Python, to facilitate the data fitting procedure, the model is transformed into a logarithmic linear form using the natural logarithm (log-linear model) as expressed in Equation 4.

$$\ln(\hat{y}) = \ln(\alpha) + bx \tag{4}$$

For each model, training is performed to fit the data using least-squares. The objective of the fitting process is to minimize the sum of squared errors, subject to the following condition:

$$\min_{\alpha,b} SSE(\alpha, b) = \sum_{j=1}^{m} y_j - (\alpha . x_j + b))^2 \tag{5}$$

After obtaining all coefficients corresponding to the minimum Sum of Squared Errors (SSE), the coefficient of determination (R^2) is then calculated using Equation 6:

$$R^2 = \frac{\sum_{i=1}^{n}(-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{6}$$

where, y_i = actual value of *y;* $\hat{y}_i$ = predicted value of *y;* $\bar{y}_i$ = average of actual value *y*. For each variable, chosen one of three models with the highest R^2 as a proposed mode.

3.3. XGBoost Modeling

In this study, XGBoost was used to model the total construction cost using multiple variables. The model was enhanced with polynomial feature expansion and logarithmic target transformation. The overall workflow consists of data pre-processing, feature transformation, model development, model evaluation, and visualization. XGBoost model was trained using the dataset obtained from project cost records and includes multiple operational variables including road section area (X1), cold mix volume (X2), emulsion volume (X3), diesel consumption (X4), gasoline consumption (X5), water usage (X6), working area (X7), asphalt chunk volume (X8), and number of workers (X9). The target variable (Y) is the project cost (Rupiah). As the project cost values are naturally large and exhibit right-skewed distributions, the target variable was transformed using the natural logarithmic function:

$$y_{log} = log(1 + y) \tag{7}$$

The log transformation reduces variance, stabilizes the learning process, and improves model sensitivity to lower-cost observations. This transformation is reversed during post-processing using the exponential function:

$$\hat{y} = exp(\hat{y}_{log}) - 1 \tag{8}$$

Although the model evaluates one predictor variable at a time, the relationship between each feature and the cost may be nonlinear. To capture such curvature, the predictor variable $X$was expanded using second-degree polynomial features:

$$X_{poly} = [X, X^2] \tag{9}$$

This expansion enables the model to approximate parabolic or curved relationships while maintaining a simple input structure.

For each predictor variable, an individual regression model was developed using the XGBoost algorithm. XGBoost was selected due to its robustness, ability to model nonlinear relationships, and suitability for small to medium-sized datasets. The model was configured with the hyperparameters as shown in Table 3.

Table 3. Hyperparameters tuning of XGBoost model

Parameters	Values
Decision trees	150
Learning rate	0.07
Maximum tree depth	3
Subsample ratio	0.9
Column sample rate	0.9
Tree building method	Histogram optimization

The model was trained using the log-transformed target variable without early stopping, as the evaluation metric was computed in log space and early stopping may interrupt optimal learning. All input features were expanded using second-order polynomial transformation prior to model training.

To ensure reproducibility, data splitting was performed using a fixed random state. Model performance was first evaluated using an 80–20% training–testing split for convergence analysis and residual diagnostics. In addition, a 5-fold cross-validation scheme was employed to assess the robustness and generalization capability of the proposed model. In this procedure, the dataset was randomly partitioned into five mutually exclusive subsets; in each fold, four subsets were used for training and the remaining subset for testing, and this process was repeated until all subsets had served as the validation set.

The predictive performance was assessed using three metrics: Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). RMSE was scaled by a constant factor of 10^{-8} to match the magnitude of cost values and improve interpretability. MAPE was computed after excluding zero-valued target observations to avoid division errors. The coefficient of determination (R^2) was calculated using predictions in the original cost scale to reflect the model's explanatory power under actual project conditions. Final model performance was reported as the mean and standard deviation of each metric across all cross-validation folds.

4. Results

4.1. Variable's Correlation Analysis

The results of the correlation analysis of the variables used in this research are depicted in Figure 1. The correlation heatmap presents the linear relationships between the input variables and Maintenance Cost (IDR), providing an initial evaluation of feature relevance. The results show that maintenance cost is strongly correlated with several material- and scale-related variables, with correlation coefficients generally above 0.75. In particular, Cold Mix Asphalt (tons) (≈0.90–0.95), Asphalt Emulsion (liters) (≈0.85–0.90), Diesel Fuel Consumption (liters) (≈0.85–0.90), and Work Area Size (m²) (≈0.88–0.93) exhibit the highest correlations with maintenance cost. These values indicate that material consumption and the extent of the treated area are the dominant contributors to road maintenance expenditure.

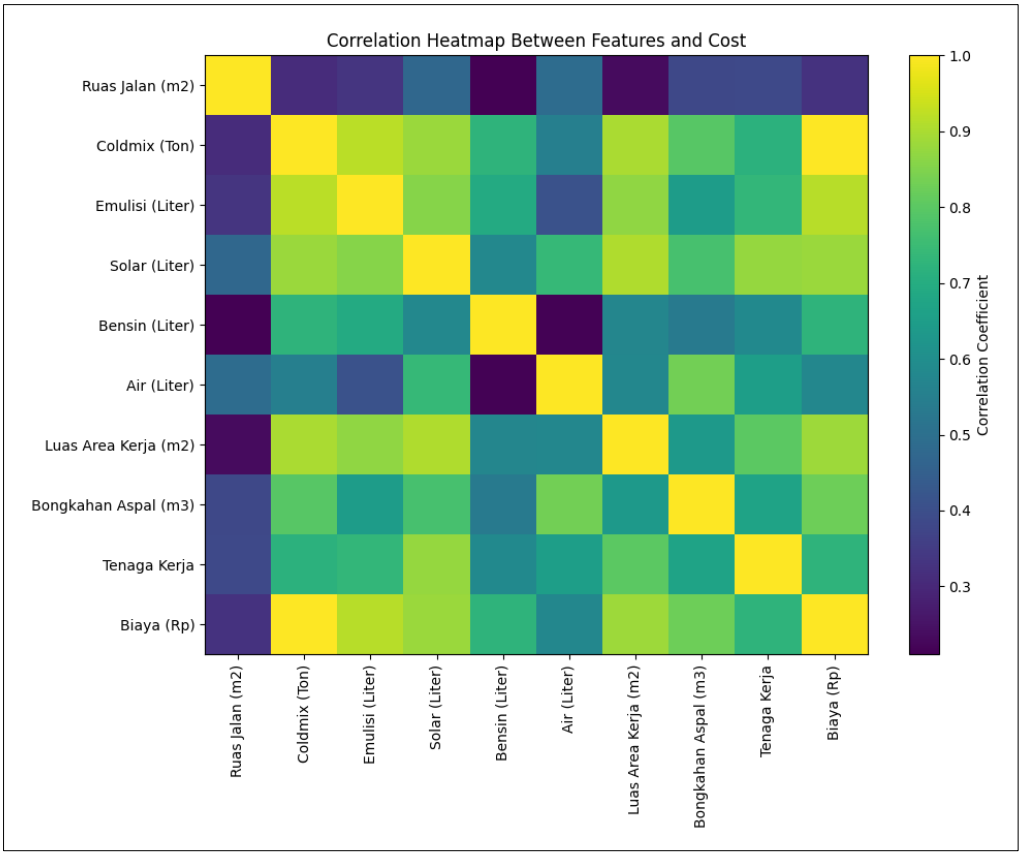

Figure 1. Variable's correlation heatmap

Strong inter-correlations are also observed among material-related variables. Cold Mix Asphalt, Asphalt Emulsion, and Diesel Fuel Consumption show correlation coefficients exceeding 0.80, suggesting that these resources are typically used simultaneously as project scale increases. Work Area Size (m²) is similarly strongly correlated with material usage (≈0.80–0.90), confirming its role as a representative indicator of maintenance activity magnitude. Although such multicollinearity may limit the effectiveness of linear models, it is less problematic for tree-based ensemble methods such as XGBoost, which can effectively handle correlated features and nonlinear interactions.

Labor- and damage-related variables also demonstrate meaningful relationships with maintenance cost. Labor Requirement shows a moderate to strong positive correlation with cost (≈0.70–0.80), while Chunked Asphalt Removal Volume (m³) exhibits a strong correlation with maintenance cost (≈0.80–0.85), reflecting the influence of damage severity on resource demand. In contrast, Road Section Area (m²) displays relatively weak correlations with cost (below 0.50), indicating that geometric road attributes alone are insufficient to explain cost variability. Overall, the correlation analysis confirms that road maintenance cost is primarily influenced by material usage, labor input, damage severity, and work area characteristics, supporting the use of XGBoost for capturing complex and nonlinear relationships in cost prediction.

4.2. Modeling Results Using Mathematical Equation

The modeling produced in the regression of each variable is as shown in Figure 2. Figure 2 illustrates the relationship between individual project variables and total project cost using scatter plots and best-fit regression curves. The results indicate that the strength of association between predictors and project cost varies substantially across variables. Overall, material-related variables show higher explanatory power, with coefficients of determination generally exceeding $R^2 \approx 0.80$, whereas operational and supporting variables exhibit weaker relationships, often with R^2 values below 0.60.

Cold mix consumption exhibits the strongest and most stable relationship with project cost, characterized by an almost perfectly linear trend and minimal dispersion. The fitted model yields an R^2value above 0.95, indicating that cold mix alone explains more than 95% of the observed cost variability. This confirms that cold mix quantity is the primary cost driver, as increases from approximately 5 to 120 tons correspond to cost escalations from below 0.2×10^8 IDR to above 2.5×10^8 IDR. Similarly, the working area demonstrates a strong linear relationship, with R^2values exceeding 0.85, suggesting that spatial construction footprint is a reliable indicator of overall project scale.

Emulsion usage shows a positive but moderately nonlinear relationship with project cost, with R^2values in the range of approximately 0.70–0.85. At lower emulsion volumes (below 100 liters), costs remain under 0.5×10^8IDR, while higher volumes exceeding 500 liters are associated with costs above 2.5×10^8 IDR. Diesel consumption also displays a clear increasing trend, with R^2values around 0.70–0.80, indicating its importance as an operational cost component, although with greater variability due to differences in equipment efficiency and site conditions.

In contrast, road section area, gasoline, and water usage exhibit relatively weak associations with project cost, typically yielding R^2values below 0.60. Road section area ranging from approximately 200 to 2,400 m² corresponds to a wide spread of project costs, from under 0.2×10^8IDR to above 2.5×10^8 IDR, highlighting its limited explanatory capability when considered independently. Gasoline and water consumption also show high dispersion, with similar consumption levels resulting in significantly different project costs.

Asphalt chunk volume presents a moderate positive relationship with project cost, particularly at higher values. Projects involving asphalt removal volumes above 30 m³ are consistently associated with costs exceeding 1.5×10^8IDR, yielding R^2values around 0.75–0.85. Labor size exhibits a nonlinear trend, with project costs increasing rapidly as labor rises from 8 to approximately 15 workers, followed by a reduced marginal increase beyond this point. This behavior results in R^2values in the range of 0.65–0.75, indicating partial but non-proportional influence.

Overall, the numerical results confirm that material intensity variables dominate project cost formation, while geometric and auxiliary variables play secondary roles. Variables with R^2values exceeding 0.80, particularly cold mix and working area, are well-suited as primary predictors in data-driven cost estimation models. The observed nonlinear patterns further justify the use of advanced machine learning or nonlinear regression approaches to capture complex interactions that cannot be adequately represented by linear models alone.

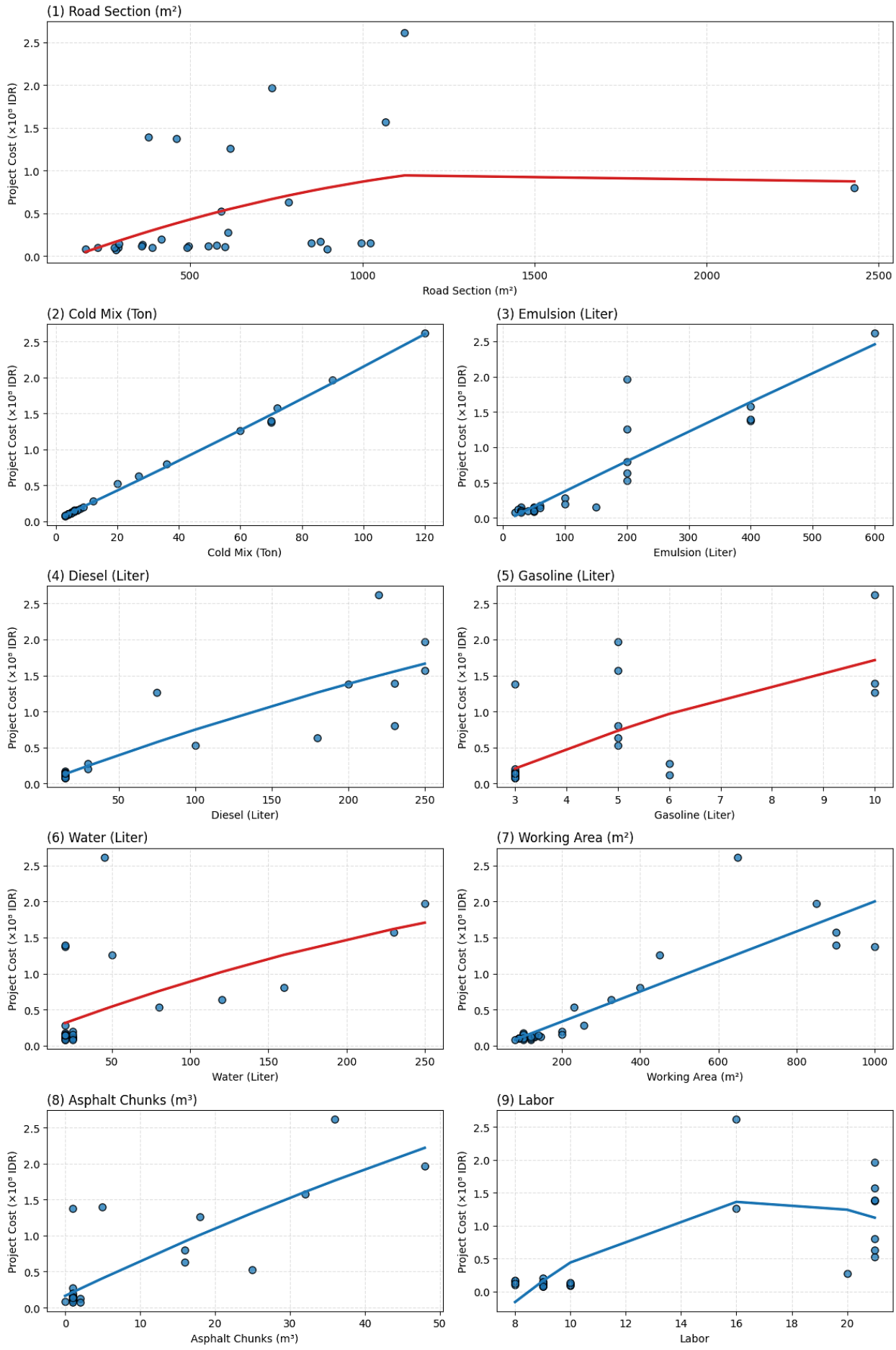

Figure 2. Estimation Results Using Mathematical Model

4.3. Modeling Results Using XGBoost

The model was trained using XGBoost as explained in the methodology section and the results of the prediction using XGBoost model is as shown in Figure 3.

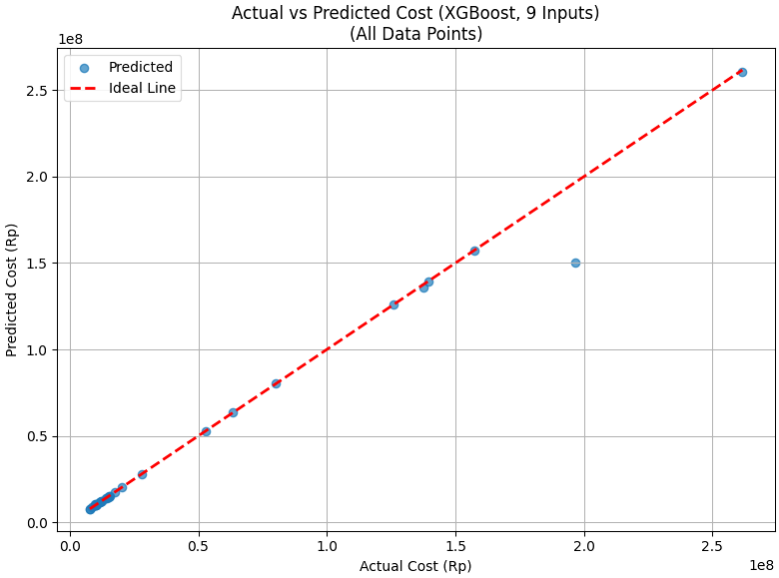

Figure 3. Estimation result using XGBoost model

Figure 4 presents the comparison between actual and predicted project costs obtained from the XGBoost model using nine input variables. The scatter points are closely aligned with the ideal 45-degree reference line, indicating a strong agreement between model predictions and observed values across the full cost range. Most data points cluster tightly around the ideal line, particularly for low- to medium - cost projects, suggesting high prediction accuracy and low systematic bias in this range.

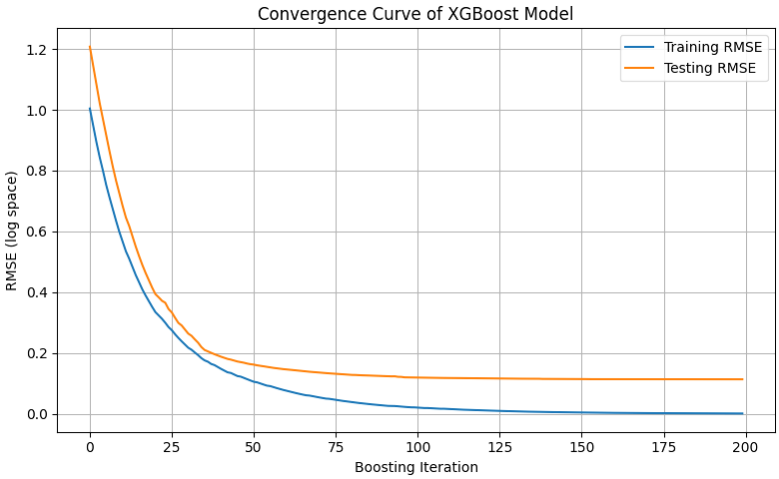

Figure 4. Convergence rate of XGBoost in training process

For higher-cost projects, a slightly larger dispersion is observed, with a small number of points deviating from the ideal line, indicating minor underestimation or overestimation at extreme values. This behavior is consistent with the results of cross-validation, where variations in error metrics across folds suggest that prediction uncertainty increases for a limited subset of projects with higher costs or atypical feature combinations. Nevertheless, the overall distribution of points remains well-centered around the ideal line, confirming the model's strong predictive capability and its ability to capture nonlinear relationships between material usage, work scale, and total project cost. The visual agreement observed in Figure 4 is further supported by the consistently high R^2 values and low average RMSE obtained from cross-validation, demonstrating that the model's performance is robust and not dependent on a single data split. The Cross-Validation Performance of the XGBoost Model is presented in Table 4.

Table 4. Cross-validation performance

Fold	RMSE	MAPE	R^2
1	0.191623	0.629623	0.935040
2	0.224796	1.998689	0.708278
3	0.416566	1.119745	0.798917
4	0.013812	0.627236	0.871048
5	0.040572	0.569179	0.992678
Mean	0.177474	0.988895	0.861192
Std. Dev.	0.145005	0.542741	0.100100

Across all folds, the model demonstrates strong predictive performance, with an average RMSE of 0.177 and an average R^2 of 0.861, indicating that a substantial proportion of the variance in project cost is consistently explained by the selected input variables. Some variability in performance is observed between folds, particularly in Fold 3, which exhibits a higher RMSE and MAPE compared to other folds. This variation suggests the presence of projects with distinct cost structures or extreme values that are more challenging to predict when used as validation data. Similarly, the lower R^2 value in Fold 2 indicates reduced explanatory power for that specific partition, likely due to an uneven distribution of high-cost projects or material-intensive cases.

Despite these variations, the relatively small standard deviation of R^2 (±0.100) and the consistently low RMSE values across most folds confirm that the model generalizes well across different subsets of the data. The high R^2 values achieved in Folds 1 and 5, exceeding 0.93 and 0.99 respectively, further demonstrate the model's capability to accurately capture complex nonlinear relationships under favorable data distributions. Overall, the cross-validation results corroborate the visual agreement observed in Figure 4 and confirm that the proposed XGBoost model provides reliable and stable cost predictions across varying project conditions.

Consistent with the strong agreement between predicted and actual costs observed in the prediction scatter plot, the convergence curve further explains how this level of accuracy is achieved during training as shown in Figure 4. The RMSE in log space decreases sharply in the early boosting iterations, with training RMSE falling from about 1.0 to below 0.2 and testing RMSE from roughly 1.2 to around 0.25 within the first 40 iterations, indicating that the model quickly captures the dominant nonlinear relationships between input variables and project cost. As the number of boosting iterations increases beyond 60–80, the testing RMSE stabilizes at approximately 0.11–0.13, while the training RMSE continues to decrease gradually toward near-zero values. This behavior is consistent with the tight clustering of points around the ideal line in the scatter plot, as the model has already reached a predictive structure that generalizes well to unseen data. The stable testing error and the limited gap between training and testing RMSE confirm that the XGBoost model attains an effective trade-off between bias and variance, reinforcing the reliability of the cost.

With the strong predictive alignment and stable convergence behavior discussed previously, the standardized residuals plot as shown in Figure 5 provides further evidence of the model's robustness. The residuals are predominantly centered around zero across the range of predicted costs, indicating that the XGBoost model does not exhibit systematic overestimation or underestimation. Most data points fall well within the ±2 standardized residual bounds, which suggests that prediction errors are statistically acceptable and largely random in nature. The absence of a clear funnel shape or structured pattern confirms that heteroscedasticity is minimal, aligning with the tight clustering around the ideal line observed in the prediction scatter plot. Although one high-cost observation shows a residual slightly above +2, this isolated deviation is consistent with complex or extreme project conditions and does not undermine the overall model performance. Taken together, this residual behavior reinforces the earlier findings that the model generalizes well, maintains stability at higher cost levels, and produces reliable cost estimates across the full prediction range.

The variable importance results as shown in Figure 6 further reinforce the reliability and interpretability of the XGBoost model. The plot shows that Cold Mix Asphalt (tons) is the most influential predictor, contributing more than 0.60 to the aggregated importance score, followed by Asphalt Emulsion (liters) at approximately 0.30. These findings are consistent with the correlation analysis, where both variables demonstrated strong associations with maintenance cost, confirming that material-intensive components are the primary cost drivers in road maintenance

operations. Other variables such as Gasoline Consumption, Working Area Size, and Asphalt Chunk Volume contribute marginally, while Road Section Area, Labor Requirement, Water Consumption, and Diesel Fuel exhibit minimal influence. The steep drop in importance beyond the top two variables indicates that cost prediction is dominated by material usage rather than geometric or labor-related attributes. This concentrated importance distribution, combined with the stable residual behavior discussed previously, demonstrates that XGBoost effectively identifies the true underlying cost determinants and leverages them to produce accurate and well-generalized predictions.

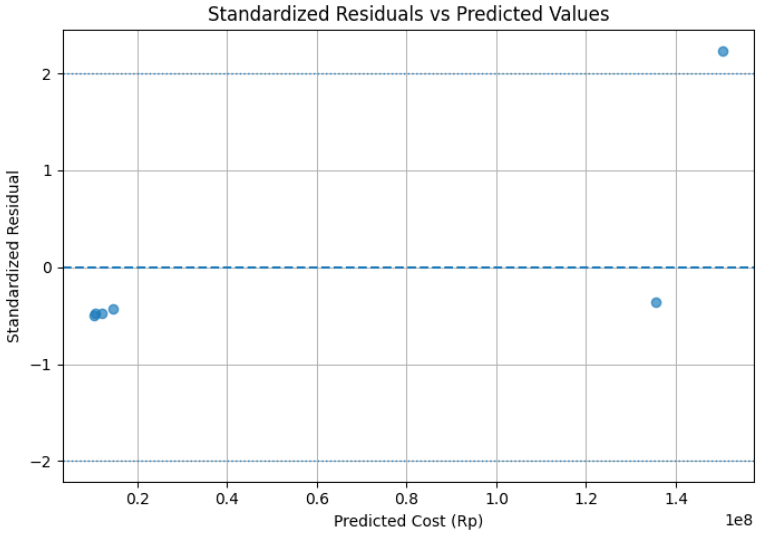

Figure 5. Standardized Residual Over Predicted Values

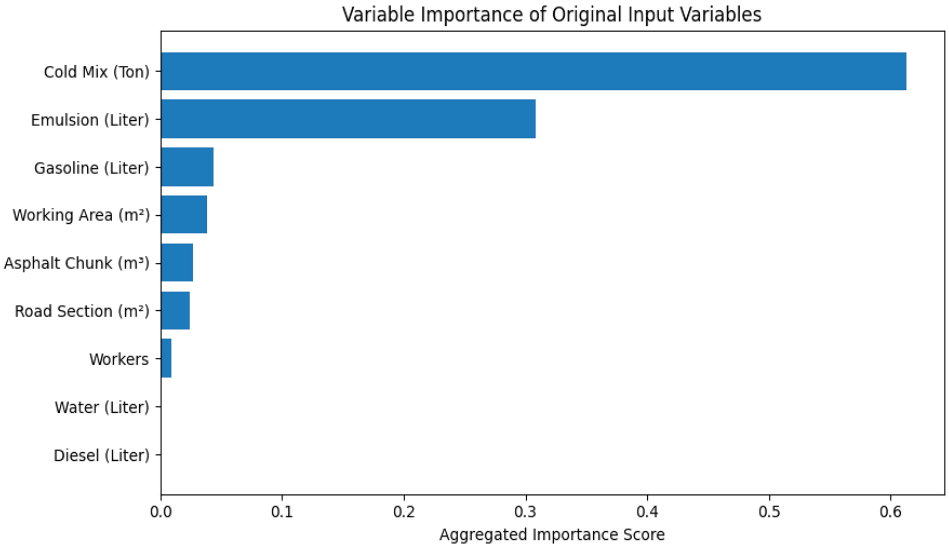

Figure 6. Variable Importance to Total Cost Value (Y)

Figure 6 illustrates important factors that influence road maintenance costs, such as cold mix, emulsion, gasoline, working area, asphalt chunk, road section, and worker. Each importance aggregate is depicted based on Figure 6.

To further evaluate the predictive capability of the proposed model, the XGBoost-based approach was compared with two commonly used methods for road maintenance cost prediction, namely Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR). The comparison results demonstrate that the proposed XGBoost model significantly outperforms both benchmark methods across all evaluation metrics. As shown in Figure 7, the predictions generated by the XGBoost model are closely aligned with the ideal line, indicating a near-perfect agreement between predicted and actual project costs across different cost ranges. This suggests that XGBoost is able to capture both low-cost and high-cost project characteristics effectively.

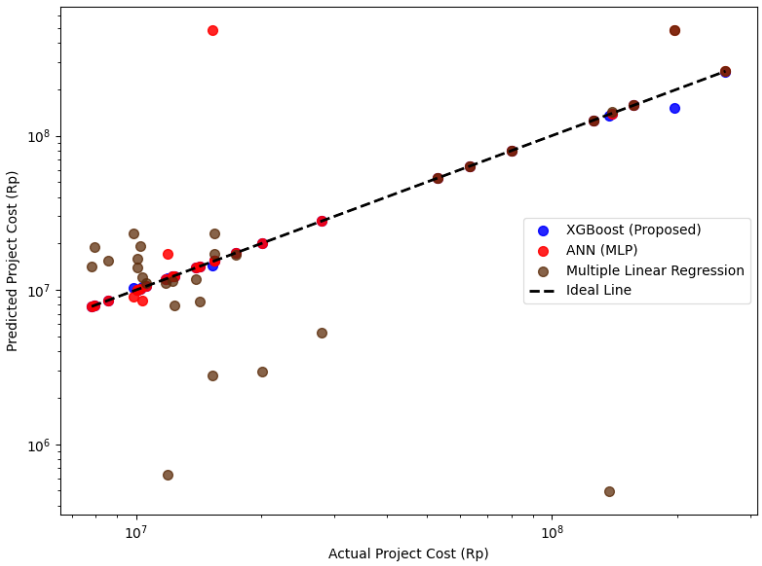

Figure 7. Performance comparison of proposed XGBoost model with baseline methods

In contrast, the predictions produced by ANN and MLR are widely scattered and deviate significantly from the ideal line. For MLR, the large dispersion reflects the inherent limitation of linear models in representing nonlinear interactions between material quantities, work area, and operational factors that influence road maintenance costs. This limitation leads to systematic underestimation or overestimation, particularly for large-scale projects.

Although ANN performs better than MLR, its predictions still exhibit considerable deviation from the ideal line. This indicates potential issues related to model stability, sensitivity to data scale, or insufficient learning of complex feature interactions. The presence of large prediction errors in ANN suggests that, for this dataset, the neural network model is less robust than the tree-based ensemble approach.

The quantitative evaluation as shown in Table 5 clearly demonstrates that the proposed XGBoost model provides more reliable and consistent predictions compared to ANN and MLR. XGBoost achieves the lowest RMSE value of 0.189, indicating a substantially smaller prediction error compared to ANN (1.307) and MLR (2.320). This highlights the superior accuracy of XGBoost in estimating road maintenance costs over a wide range of project scales.

Table 5. Quantitative evaluation of the proposed model compared to baseline models

Model	RMSE	MAPE (%)	R^2
XGBoost (Proposed)	0.189	0.604	0.937
Multiple Linear Regression	2.320	6.625	0.523
ANN (MLP)	1.307	5.560	0.821

In terms of relative error, the proposed model also records the lowest MAPE value of 0.604%, whereas ANN and MLR produce considerably higher MAPE values of 5.56% and 6.62%, respectively. The low MAPE achieved by XGBoost indicates that the model maintains consistent prediction accuracy even for projects with high cost variability, which is critical in practical budgeting and planning applications.

Furthermore, the coefficient of determination (R^2) for the XGBoost model reaches 0.937, suggesting that the model is able to explain approximately 94% of the variance in road maintenance costs. In contrast, ANN and MLR yield R^2 values of 0.821 and 0.523, respectively. The relatively low R^2 of MLR reflects its limited ability to capture complex and nonlinear relationships among input variables, while ANN, although better, still underperforms compared to XGBoost. Overall, the quantitative results clearly confirm the robustness and superior generalization capability of the proposed XGBoost model.

5. Discussion

5.1. Identifying Influencing Variables

Based on the research results, it can be identified that several variables significantly influence road maintenance costs, as revealed by the modeling results using both mathematical regression and XGBoost-based machine learning approaches. The dominant influencing factors are primarily related to material usage and operational scale, including cold mix asphalt consumption, asphalt emulsion volume, working area size, fuel consumption, asphalt removal volume, and labor requirements. These findings indicate that material-intensive variables play a crucial role in determining total maintenance expenditure.

The identification of material consumption as the most influential factor is consistent with previous studies, which emphasize that road maintenance costs are largely driven by material quantities and construction scale rather than purely geometric road attributes [6]. In particular, cold mix asphalt and asphalt emulsion demonstrate the highest contribution to cost formation, reflecting their direct relationship with pavement repair intensity and damage severity. This result aligns with the findings of Chen et al. [8], who reported that material-related variables dominate cost and performance modeling in pavement maintenance applications.

Labor requirements and asphalt chunk removal volume also exhibit a significant influence on maintenance cost, especially for projects involving severe pavement deterioration. These variables reflect the complexity of maintenance activities and the extent of structural intervention required. Similar observations were reported in previous infrastructure maintenance studies, where labor intensity and damage severity were identified as key cost drivers [9]. In contrast, geometric variables such as road section area show relatively weaker influence when considered independently, suggesting that physical dimensions alone are insufficient to explain cost variability without considering material and operational factors.

Overall, the results confirm that road maintenance cost is governed by a complex interaction between material intensity, operational scale, and pavement damage severity. These relationships are inherently nonlinear, as increases in material usage or treated area do not lead to proportional increases in cost, particularly when different levels of damage and operational constraints are involved. The proposed XGBoost model demonstrates a clear advantage in capturing such nonlinear interactions through its ensemble of decision trees, where hierarchical splitting allows the model to learn conditional relationships between variables that cannot be adequately represented by linear formulations.

Compared to Multiple Linear Regression (MLR), which assumes a fixed linear relationship between predictors and cost, XGBoost adapts to varying cost sensitivities across different ranges of input variables. This capability is reflected in its substantially lower RMSE and MAPE values, as well as the highest coefficient of determination (R^2), indicating superior predictive accuracy and explanatory power. Although the ANN (MLP) model is theoretically capable of modeling nonlinear relationships, its performance is constrained by the limited size of the dataset. Neural networks typically require a large number of samples to effectively generalize complex patterns, and with only 30 observations available, the ANN model is more susceptible to instability and suboptimal weight estimation.

A key advantage of XGBoost in this study is its robustness when applied to small datasets. Gradient boosting frameworks are well known for their ability to achieve high performance with limited data by sequentially correcting residual errors and incorporating regularization mechanisms that control model complexity [38]. This characteristic makes XGBoost particularly suitable for infrastructure cost modeling, where high-quality labeled data are often scarce due to budget, time, and measurement constraints [39].

Furthermore, the application of cross-validation plays a critical role in strengthening the reliability of the proposed model under data-limited conditions. By repeatedly training and validating the model across different data partitions, cross-validation reduces the risk of overfitting and provides a more stable estimate of generalization performance. In the context of a small dataset, this process ensures that the reported performance metrics are not dependent on a specific train–test split, thereby enhancing the robustness and credibility of the results.

6. Conclusion

This study presents the development of a predictive model for estimating road maintenance costs using the Extreme Gradient Boosting (XGBoost) algorithm and identifies key factors influencing maintenance expenditure. The proposed model incorporates material usage, labor requirements, and operational variables derived from road maintenance projects, providing a comprehensive representation of cost formation mechanisms. The results demonstrate that the XGBoost-based model achieves strong predictive performance, as indicated by low RMSE and MAPE values and high coefficients of determination (R^2) during both training and testing phases. These performance metrics confirm that the model is capable of accurately predicting road maintenance costs across various project conditions. The findings further indicate that model performance decreases when the dataset size is limited, predictor variability is low, or when cost distributions are highly unbalanced, highlighting the importance of data quality and diversity in machine learning-based cost estimation.

Material-related variables, particularly cold mix asphalt and asphalt emulsion, are identified as the most influential factors affecting road maintenance costs, followed by working area size, asphalt removal volume, fuel consumption, and labor requirements. These results confirm that material intensity and operational complexity are the dominant determinants of maintenance expenditure, consistent with findings reported in previous studies on pavement and infrastructure maintenance.

The estimated road maintenance costs generated by this model are expected to provide valuable background information for government agencies in understanding the key factors influencing maintenance expenditure. Consequently, the proposed approach can support decision-making at the agency level for preventive and corrective maintenance planning. In addition, the model can assist road planners and engineers in estimating maintenance budget requirements and defining cost details more accurately, such as material quantities and labor needs.

Future research should incorporate more detailed practical variables, such as traffic loading characteristics, climatic effects, and long-term pavement performance indicators, to further enhance prediction accuracy. Moreover, extending the model to larger road networks, including highways and toll roads with different structural characteristics, would improve its generalizability and support the development of more effective road maintenance policies at the regional and national levels

7. Declarations

7.1. Author Contributions

Conceptualization, D.S. and E.M.S.; methodology, L.S.P.; software, D.S.; validation, E.M.S. and L.S.P.; formal analysis, D.S. and E.M.S.; investigation, L.S.P.; resources, D.S.; data curation, E.M.S.; writing—original draft preparation, D.S. and E.M.S.; writing—review and editing, L.S.P.; visualization, E.M.S.; supervision, E.M.S.; project administration, D.S.; funding acquisition, D.S. All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available in the article.

7.3. Funding

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

7.4. Acknowledgments

The authors would like to thank the Public Works and Housing Agency (Dinas PUPR) of Bogor City for providing access to the data used in this study

7.5. Conflicts of Interest

The authors declare no conflict of interest.

8. References

[1] Pamuković, J. K., Rogulj, K., Dumanić, D., & Jajac, N. (2021). A sustainable approach for the maintenance of asphalt pavement construction. Sustainability (Switzerland), 13(1), 1–18. doi:10.3390/su13010109.

[2] Mohammadi, A., Igwe, C., Amador-Jimenez, L., & Nasiri, F. (2020). Applying lean construction principles in road maintenance planning and scheduling. International Journal of Construction Management, 1–11. doi:10.1080/15623599.2020.1788758.

[3] Tee, K. F., & Ekpiwhre, E. (2019). Reliability-based preventive maintenance strategies of road junction systems. International Journal of Quality and Reliability Management, 36(5), 752–781. doi:10.1108/IJQRM-01-2018-0018.

[4] Huang, M., Dong, Q., Ni, F., & Wang, L. (2021). LCA and LCCA based multi-objective optimization of pavement maintenance. Journal of Cleaner Production, 283. doi:10.1016/j.jclepro.2020.124583.

[5] Makovska, J. (2020). Method of Evaluation of Road Routine Maintenance Strategies. Economics, Finance and Management Review, 4(4), 106–112. doi:10.36690/2674-5208-2020-4-106.

[6] Huang, L. L., Lin, J. D., Huang, W. H., Kuo, C. H., Chiou, Y. S., & Huang, M. Y. (2024). Developing Pavement Maintenance Strategies and Implementing Management Systems. Infrastructures, 9(7), 101. doi:10.3390/infrastructures9070101.

[7] Mohammadi, N., & Klein-Paste, A. (2025). Simulating winter maintenance efforts: A multiscale geographically weighted regression model. Cold Regions Science and Technology, 236. doi:10.1016/j.coldregions.2025.104512.

[8] Chen, S., Cao, J., Wan, Y., Shi, X., & Huang, W. (2024). Enhancing rutting depth prediction in asphalt pavements: A synergistic approach of extreme gradient boosting and snake optimization. Construction and Building Materials, 421. doi:10.1016/j.conbuildmat.2024.135726.

[9] Giustozzi, F., Crispino, M., & Flintsch, G. (2012). Multi-attribute life cycle assessment of preventive maintenance treatments on road pavements for achieving environmental sustainability. International Journal of Life Cycle Assessment, 17(4), 409–419. doi:10.1007/s11367-011-0375-6.

[10] de Bortoli, A., Féraille, A., & Leurent, F. (2022). Towards Road Sustainability—Part I: Principles and Holistic Assessment Method for Pavement Maintenance Policies. Sustainability (Switzerland), 14(3), 1513. doi:10.3390/su14031513.

[11] Sari, E. M., Setiawan, D., & Putranto, L. S. (2025). Evaluation of road project maintenance costs in Bogor City - West Java. Kompak: Jurnal Ilmiah Komputerisasi Akuntansi, 18(2), 434–445. doi:10.51903/kompak.v18i2.2959.

[12] Milad, A., Ali, A. A., Al-Sulaimi, Z. S., & Al-Kindi, K. M. (2025). Utilizing spatial artificial intelligence to develop pavement performance indices: A case study. Scientific Reports, 15(1), 39603. doi:10.1038/s41598-025-23290-7.

[13] Sarsam, S. I. (2016). Pavement maintenance management system: A review. Trends in Transport Engineering and Applications, 3(2), 19-30.

[14] Bortoli, A. de, Féraille, A., & Leurent, F. (2022). Towards Road Sustainability—Part II: Applied Holistic Assessment and Lessons Learned from French Highway Resurfacing Strategies. Sustainability (Switzerland), 14(12), 7336. doi:10.3390/su14127336.

[15] Zhang, M., Gong, H., Xiao, R., Jiang, X., Ma, Y., & Huang, B. (2023). Life-cycle cost analysis of rehabilitation strategies for asphalt pavements based on probabilistic models. Road Materials and Pavement Design, 24(1), 121–137. doi:10.1080/14680629.2021.2012235.

[16] Golroo, A., Fani, A. H., Naseri, H., & Mirhasani, S. A. (2021). Pavement Maintenance and Rehabilitation Planning Considering Budget Uncertainty. Amirkabir Journal of Civil Engineering, 53(7), 2781–2800. doi:10.22060/ceej.2020.17502.6583.

[17] Simić, N., Ivanišević, N., Nedeljković, Đ., Senić, A., Stojadinović, Z., & Ivanović, M. (2023). Early Highway Construction Cost Estimation: Selection of Key Cost Drivers. Sustainability (Switzerland), 15(6), 5584. doi:10.3390/su15065584.

[18] Erfani, A., & Mansouri, A. (2025). Uncertainty-Aware Pavement Roughness Forecasting Using Adaptive Conformal Prediction. International Journal of Pavement Research and Technology, 1–19. doi:10.1007/s42947-025-00689-z.

[19] Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. Water (Switzerland), 10(11), 1536. doi:10.3390/w10111536.

[20] Jafari, M., & Mousavi, E. (2026). A Critical Review of the Data-Driven Cost Estimation Approach in Construction Research. Journal of Construction Engineering and Management, 152(2), 3125013. doi:10.1061/jcemd4.coeng-16840.

[21] Habib, O., Abouhamad, M., & Bayoumi, A. E. M. (2025). Ensemble learning framework for forecasting construction costs. Automation in Construction, 170, 105903. doi:10.1016/j.autcon.2024.105903.

[22] Das, P., Kashem, A., Hasan, I., & Islam, M. (2024). A comparative study of machine learning models for construction costs prediction with natural gradient boosting algorithm and SHAP analysis. Asian Journal of Civil Engineering, 25(4), 3301–3316. doi:10.1007/s42107-023-00980-z.

[23] Beth Visintine, by A., Gary Hicks, R., Gary Hicks, P. R., Cheng, D., Director, P., Elkins, G. E., & Groeger, J. (2015). Factors Affecting the Performance of Pavement Preservation Treatments. In the 9th International Conference on Managing Pavement Assets (ICMPA9), 1–14.

[24] Qiao, Y., Dawson, A. R., Parry, T., & Flintsch, G. W. (2015). Evaluating the effects of climate change on road maintenance intervention strategies and Life-Cycle Costs. Transportation Research Part D: Transport and Environment, 41, 492–503. doi:10.1016/j.trd.2015.09.019.

[25] Zheng, X., Easa, S. M., Yang, Z., Ji, T., & Jiang, Z. (2019). Life-cycle sustainability assessment of pavement maintenance alternatives: Methodology and case study. Journal of Cleaner Production, 213, 659–672. doi:10.1016/j.jclepro.2018.12.227.

[26] Awaad, S., Mansour, D. M., Mahdi, I., & Abdelrasheed, I. (2024). Impact of material supply chain on the productivity optimization for the construction of roads projects. Scientific Reports, 14(1), 3294. doi:10.1038/s41598-024-53660-6.

[27] Khodabakhshian, A., Puolitaival, T., & Kestle, L. (2023). Deterministic and Probabilistic Risk Management Approaches in Construction Projects: A Systematic Literature Review and Comparative Analysis. Buildings, 13(5), 1312. doi:10.3390/buildings13051312.

[28] Zhai, L., Yan, X., & Liu, G. (2024). Cost Impact Factors and Control Measures of Road and Bridge Projects Based on Linear Regression Model. Informatica, 48(17). doi:10.31449/inf.v48i17.6370.

[29] Petruseva, S., Zileska-Pancovska, V., Žujo, V., & Brkan-Vejzović, A. (2017). Construction costs forecasting: comparison of the accuracy of linear regression and support vector machine models. Tehnicki Vjesnik - Technical Gazette, 24(5), 1431–1438. doi:10.17559/tv-20150116001543.

[30] Al-Mistarehi, B., Shtayat, A., Imam, R., & Abdallah, A. (2025). An Automated Assessment Technique for Pavement Defects Using a Laser Scanner and Deep Machine Learning. Civil Engineering Journal, 11(3), 1088–1105. doi:10.28991/CEJ-2025-011-03-015.

[31] De Figueiredo, B. H., Dos Santos, M., Fávero, L. P. L., Moreira, M. Â. L., & De Araújo Costa, I. P. (2022). Analysis of maintenance activities in Urban Pavement Management Systems based on Decision Tree Algorithm. Procedia Computer Science, 214(C), 712–719. doi:10.1016/j.procs.2022.11.233.

[32] Kaparthi, S., & Bumblauskas, D. (2020). Designing predictive maintenance systems using decision tree-based machine learning techniques. International Journal of Quality and Reliability Management, 37(4), 659–686. doi:10.1108/IJQRM-04-2019-0131.

[33] Gangadhari, R. K., Khanzode, V., & Murthy, S. (2022). Application of rough set theory and machine learning algorithms in predicting accident outcomes in the Indian petroleum industry. Concurrency and Computation: Practice and Experience, 34(26), 7277. doi:10.1002/cpe.7277.

[34] Lin, M.-L., & Huynh, L. L. (2024). Bridging causal explanation and predictive modeling: the role of PLS-SEM. International Journal of Research in Business and Social Science (2147- 4478), 13(10), 197–206. doi:10.20525/ijrbs.v13i10.3888.

[35] Republic of Indonesia. (2022). Law of the Republic of Indonesia Number 2 of 2022 on roads. Republic of Indonesia, Jakarta, Indonesia.

[36] PUPR. (2011). Regulation of the Minister of Public Works on Road Maintenance Procedures. Ministry of Public Works and Housing (PUPR), Jakarta, Indonesia.

[37] PUPR. (2020). Road Maintenance Management Guidelines, Directorate General of Highways. Ministry of Public Works and Housing (PUPR), Jakarta, Indonesia.

[38] Lartey, B., Homaifar, A., Girma, A., Karimoddini, A., & Opoku, D. (2021). XGBoost: A tree-based approach for traffic volume prediction. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, 1280–1286. doi:10.1109/SMC52423.2021.9658959.

[39] Zhu, J., Yin, Y., Ma, T., & Wang, D. (2025). A novel maintenance decision model for asphalt pavement considering crack causes based on random forest and XGBoost. Construction and Building Materials, 477. doi:10.1016/j.conbuildmat.2025.140610.