Rasila K.A.M., Mahakalanda, I. and Edirisooriya, M.W., 2025. Predicting cost elements of construction projects using supervised machine learning techniques. In: Waidyasekara, K.G.A.S., Jayasena, H.S., Wimalaratne, P.L.I. and Tennakoon, G.A. (eds). *Proceedings of the 13th World Construction Symposium*, 15-16 August 2025, Sri Lanka. pp. 1111-1124. DOI: https://doi.org/10.31705/WCS.2025.83. Available from: https://ciobwcs.com/papers/

PREDICTING COST ELEMENTS OF CONSTRUCTION PROJECTS USING SUPERVISED MACHINE LEARNING TECHNIQUES

K.A.M. Rasila[1], L. Mahakalanda[2] and M.W. Edirisooriya[3]

ABSTRACT

Accurately predicting construction project costs remains challenging due to their dynamic and complex nature. While traditional methods address most cost components based on resources it consumes or how such activities perform, certain elements like fuel often rely on expert judgment for validation or adjustment, as traditional methods frequently fail to capture all influencing project parameters. This research explores the feasibility of utilizing supervised Machine Learning (ML) techniques to predict these volatile cost elements, focusing specifically on fuel, a key project cost. The study addresses key gaps identified in the literature, particularly the need for models that can manage the uncertainty of specific cost elements and incorporate a broader range of influencing factors, including macroeconomic parameters. By leveraging historical data extracted from Enterprise Resource Planning (ERP) systems, alongside additional project attributes such as average fuel price and construction cost indices, this study demonstrates a novel, data driven approach to cost estimation. The methodology involved data preprocessing to ensure quality and consistency, followed by feature selection to identify the most relevant attributes influencing fuel cost. Several supervised ML models were compared, to identify model with superior performance. The chosen model was further optimized through iterative refinement techniques, to enhance its predictive accuracy and stability. The findings highlight the potential of supervised ML to revolutionize construction cost estimation practices, offering a more data driven, accurate, and efficient method for managing project budgets realistically.

Keywords: *Cost Estimation; Data Analytics; Enterprise Resource Planning; Machine Learning; Predictive Modelling.*

1. INTRODUCTION

The cost of a construction project is a complex amalgamation of various elements. Among the diverse tools available for detailing and deriving these cost elements, the Bill of Quantities (BOQ) is widely recognised and commonly used. Standard industry practices define the structure and item list within a BOQ, fostering a unified approach across

[1] Department of Decision Science, Faculty of Business, University of Moratuwa, Sri Lanka, majithrasila@gmail.com
[2] Senior Lecturer, Department of Decision Science, Faculty of Business, University of Moratuwa, Sri Lanka, indra.mahakalanda@gmail.com
[3] Lecturer, Department of Decision Science, Faculty of Business, University of Moratuwa, Sri Lanka, mwe163@gmail.com

different projects. Estimators typically derive the cost of each individual element through diverse methodologies, including resource-based, time-based, methodology-based, or expert opinion-based approaches. For certain elements such as energy, water, and fuel consumption, as well as staffing, plant, machinery, security, safety protocols, tools, scaffolding, and communication, estimators often rely heavily on expert judgment to adjust values based on historical data and professional experience to achieve more realistic figures.

However, this reliance on expert judgment and historical extrapolation presents significant limitations, particularly when confronted with the inherent variability of certain cost components. As observed by Elhag et al. (1998), traditional estimation methods frequently overlook critical contextual variables like contract type, site-specific challenges, and market volatility. Similarly, Hashemi et al. (2020) emphasised the inadequacy of linear, rule-based techniques in handling uncertain and interdependent cost determinants. The challenge of accurately forecasting these volatile elements is illustrated in Figure 1. This scatter plot, which spans a wide range of values for fuel cost percentage across 82 construction projects completed between 2008 and 2024, vividly demonstrates the substantial fluctuations and inherent unpredictability, underscoring the limitations of relying solely on traditional methods.

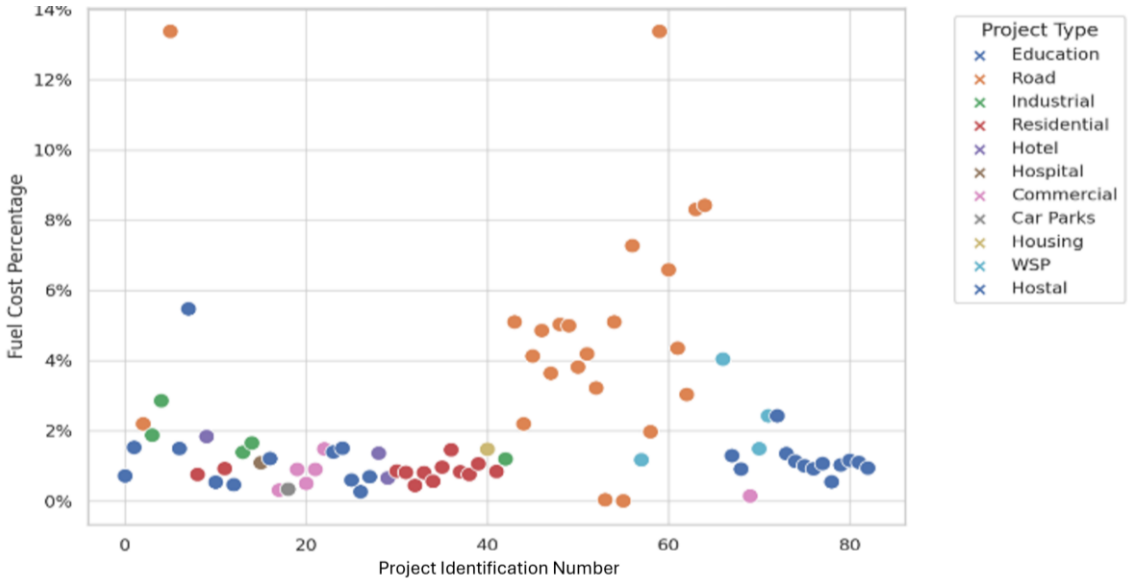

Figure 1. Fuel cost percentage by project type

To address these challenges, contemporary research increasingly advocates for data-driven approaches, particularly those based on ML. Predictive analytics and supervised ML models offer superior accuracy and flexibility compared to deterministic models. These models perform exceptionally well when trained on large, structured datasets, as they can uncover complex, nonlinear relationships among multiple variables that influence costs. Integrating ML techniques into the cost forecasting framework not only enhances the accuracy of estimates but also improves decision-making capabilities throughout the project lifecycle. As noted by Miranda et al. (2022), predictive analytics enables proactive budgeting, scenario analysis, and dynamic cost control, especially at the early stages of a project where uncertainty is highest.

This study adopts supervised ML techniques due to their proven effectiveness in predictive modelling when historical labelled datasets are available. Supervised ML involves learning a function that maps input variables to a known output variable, making it well-suited for tasks such as cost estimation, where both independent features and target cost elements are defined (Jordan & Mitchell, 2015). In contrast, unsupervised ML algorithms are typically employed for exploratory purposes, such as clustering or anomaly detection, where the output is not known in advance. The current study leverages historical project data with clearly defined cost outcomes specifically fuel expenditure which enables the use of supervised learning for accurate cost prediction. Moreover, supervised methods such as Random Forest and Gradient Boosting have demonstrated superior performance in construction cost forecasting due to their ability to model complex, non-linear relationships and handle high-dimensional data (Hashemi et al., 2020; Abed et al., 2022). Accordingly, supervised ML was selected as the most appropriate methodological approach to fulfil the study's objective of estimating volatile cost components using structured ERP datasets.

This research contributes to the ongoing transformation in construction cost estimation by evaluating the applicability of supervised ML for predicting fuel expenditure. The motivation for focusing on fuel costs stems from their demonstrated variability across different project types and durations, as highlighted in Figure 1. This variability provides an ideal test case for assessing the robustness and predictive power of ML models, not only for fuel but also for other volatile cost elements. Crucially, this study addresses the identified gaps in existing literature by incorporating both project-specific and broader economic variables, such as average fuel prices and Construction Industry Development Authority (CIDA) indices, which have often been neglected in previous ML cost estimation models. Through a systematic analysis of historical project data from an ERP system, this study evaluates the feasibility of utilising a supervised ML model to dynamically estimate the percentage of fuel costs. This process aims to validate the concept of employing ML models across various other cost elements and suggests wider implications for estimating similarly volatile components, moving beyond reliance on subjective estimations and contributing to more accurate, dynamic, and comprehensive cost management strategies.

2. LITERATURE REVIEW

The field of construction cost estimation has traditionally relied on methodologies developed over time, such as those based on resources, time, or expert opinion. However, recent years have witnessed the emergence of new concepts, notably Machine Learning (ML), which offer promising avenues for predicting construction project costs with greater accuracy and efficiency.

Early applications of ML in construction cost forecasting often utilised simpler regression techniques. For instance, Prasetyono et al. (2021) suggested the application of linear regression for forecasting construction costs in residential building projects, devising a model to predict future housing construction costs based on the year. Similarly, Suchetha et al. (2023) underscored the significance of ML in accurately predicting home costs to mitigate losses, asserting that linear regression yielded the highest accuracy among other methods like gradient boost and XGBoost regression. Expanding on regression techniques, Mahamid (2011) emphasised the development of early cost estimating models for road construction projects using multiple regression, formulating 11 models

with high coefficients of determination (R2 ranging from 0.92 to 0.98), demonstrating strong correlation with actual data and applicability for initial project stages.

More advanced ML techniques, particularly ensemble methods and deep learning architectures, have shown considerable promise in predicting construction costs. A systematic review by Abed et al. (2022) highlighted the growing adoption of ML methods in forecasting construction costs and underscored the superiority of ensemble learning models in managing nonlinear relationships and high-dimensional datasets. Recent advancements include models integrating deep neural networks (DNN) with validation units to improve predictive accuracy and interpretability in cost forecasting (Saeidlou & Ghadiminia, 2024).

Despite these advancements, several critical research gaps persist in the domain of construction cost estimation using ML. While most ML approaches focus on project-specific attributes such as labour, materials, and equipment, they often neglect broader economic variables such as inflation rates, price volatility, and market demand. Salleh et al. (2023) emphasised that integrating these external factors could significantly improve the adaptability of ML models to real-world conditions. This oversight limits the models' ability to provide holistic and accurate cost forecasts under fluctuating market conditions.

A significant challenge acknowledged in the literature is the difficulty in establishing a universally applicable set of attributes due to the unique characteristics of each building project. Salleh et al. (2023) observed that construction cost data were often fragmented, inconsistent, or incomplete, severely limiting the training of reliable predictive models. This underscores an urgent need for comprehensive and accessible data repositories to support the development of robust ML-based cost estimation tools.

There is an absence of ML models explicitly designed to predict individual building cost elements, largely due to the unavailability of rich, high-quality datasets for these specific components. While some studies, like Katyare et al. (2023) advocated for ML techniques to predict fuel costs by integrating IoT-based sensing data, the broader application of ML to other inherently volatile cost elements (like utilities or plant costs) remains underexplored. The discourse by Katyare et al. (2023) also highlighted challenges stemming from the lack of digitisation in the construction industry, which complicates the use of real-time data for ML applications.

Many existing models limit the number of variables considered due to concerns about manageability. However, expanding the feature set using techniques such as feature selection and dimensionality reduction could significantly enhance model performance without introducing excessive complexity. Salleh et al. (2023) identified an exhaustive compilation of 68 ranked attributes influencing building project costs but acknowledged that the applicability and relevance of these attributes can vary significantly across project types, affecting the generalizability and accuracy of ML models.

In summary, addressing these identified research gaps would significantly advance the domain of construction cost estimation. It would enable the creation of more accurate, dynamic, and comprehensive ML models that align with the evolving demands of the construction industry (Salleh et al., 2023; Katyare et al., 2023). The collaborative effort to overcome these challenges would pave the way for more effective cost management strategies, ultimately contributing to the enhanced sustainability and profitability of construction projects.

3. METHODOLOGY

This chapter presents the methodological framework adopted to predict the fuel cost of construction projects, utilizing ML techniques. This study followed a systematic approach that commenced with data extraction, followed by preprocessing, feature selection, and model evaluation. Once the best-performing model was identified, it underwent further optimisation and refinement for enhanced accuracy. Figure 2graphically illustrates the sequential processes involved in this study, from data acquisition to model deployment.

Google Colaboratory (Colab), a cloud-based computational environment, was utilised in this study for the development, training, and evaluation of ML models. As demonstrated by Ray et al. (2021), Colab enables real-time data processing, analysis, and visualisation within a browser-based interface. The platform supports the full ML workflow, from importing datasets to training classifiers and evaluating model performance, without the constraints of local hardware limitations. By executing code on Google's cloud servers, Colab provides access to advanced computational resources, including Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs), thereby significantly enhancing model training efficiency and scalability.

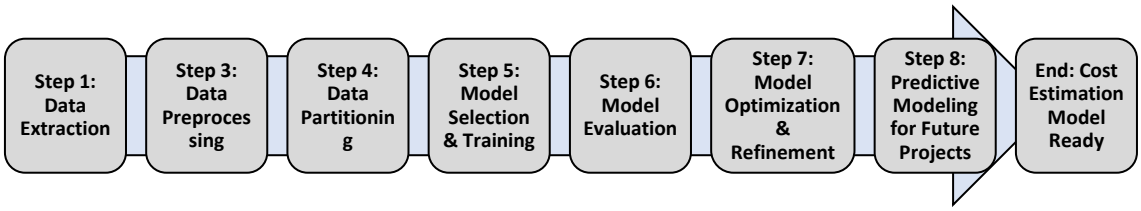

Figure 2. Proposed methodology for developing an ML model

3.1 DATA EXTRACTION

The initial phase of the research involved extracting historical data from the ERP system of a construction organisation. The extracted data comprised detailed records from 82 projects undertaken between 2008 and 2024. These records included various cost elements, such as fuel, labour, and machinery. Additionally, further project attributes not directly associated with the ERP-extracted data were incorporated

The initial phase of the research involved extracting historical data from the ERP system of a construction organisation. The extracted data comprises detailed records from 82 projects undertaken between 2008 and 2024. These records include various cost elements, such as fuel, labour, and machinery. Additionally, further project attributes were incorporated that were not associated with the extracted data.

The primary variables extracted from the ERP systems include:

- Project identification details: project number, project name
- Cost data: percentage of fuel cost, project cost
- Time-related data: time period of the project, start date of the project

The study incorporated additional data into the dataset related to the project, such as:

- Average fuel price for the duration of the project,
- Construction cost indices for the duration of the project published by the CIDA,
- Project type (e.g., residential, commercial, road, etc.),

- Project spread (horizontal or vertical),
- Height category (high-rise or low-rise),
- Type of crane (mobile crane or tower crane),
- Source of concrete (on-site ready-mix or externally sourced),
- Source of energy (generator power or electricity powered)

This structured data enabled the creation of a robust dataset, serving as the foundation for all subsequent phases of analysis. It enabled the development of a comprehensive dataset that incorporates macroeconomic and project-related attributes not previously found in research. Considering its generation from ERP records at the organisational level, it provides a standardised set of data that can be utilised across various projects within the organisation. Available information at the organisational level facilitated the analysis of different parameters derived from the dataset, as well as the inclusion of additional parameters that may influence fuel costs.

3.2 DATA PREPROCESSING

Given the complexity and heterogeneity of construction project data, preprocessing is a vital component of the methodological process. Data preprocessing ensures that the dataset is clean, consistent, and ready for analysis. This process involved several key tasks as follows:

3.2.1 Handling Missing Values

The dataset was examined for missing or incomplete data. Missing values were addressed using appropriate imputation methods.

3.2.2 Encoding Categorical Variables

ML models typically require all input features to be in numerical format. Therefore, categorical variables such as project type (residential, commercial, road) were transformed into numerical format through one hot encoding. This technique was chosen to avoid imposing an ordinal relationship between categories that are inherently non-ordinal.

3.2.3 Scaling and Normalization

To ensure that no single feature disproportionately influences the model, numeric features were normalized using the min max scaling method. This transformation scaled all numeric features to a common range, typically between 0 and 1, facilitating more effective model training, particularly for gradient-based algorithms such as Gradient Boosting Regressors.

3.3 FEATURE SELECTION AND VALIDATION

To enhance predictive accuracy and reduce the model's dimensionality, a feature selection process was applied before training. This process aimed to identify the most relevant project attributes that significantly influence the percentage of fuel cost. A correlation matrix was generated to quantify the linear relationships between all numerical variables and the target variable. Pearson's correlation coefficient was used due to its effectiveness in measuring the strength and direction of linear dependencies.

Initially, features with a correlation coefficient above a predefined threshold were selected. Those with negligible or inverse relationships were excluded to reduce noise

and avoid overfitting. Beyond purely statistical criteria, domain expertise was applied to retain variables known to influence fuel consumption, such as Contract Sum, CIDA Indices, Fuel Price, and Crane Type, even if their correlation strength was moderate. This hybrid approach ensured that variables with practical relevance were not overlooked solely due to statistical filtering.

Figure 3 visually represents the correlation values between selected parameters and the percentage of fuel cost.

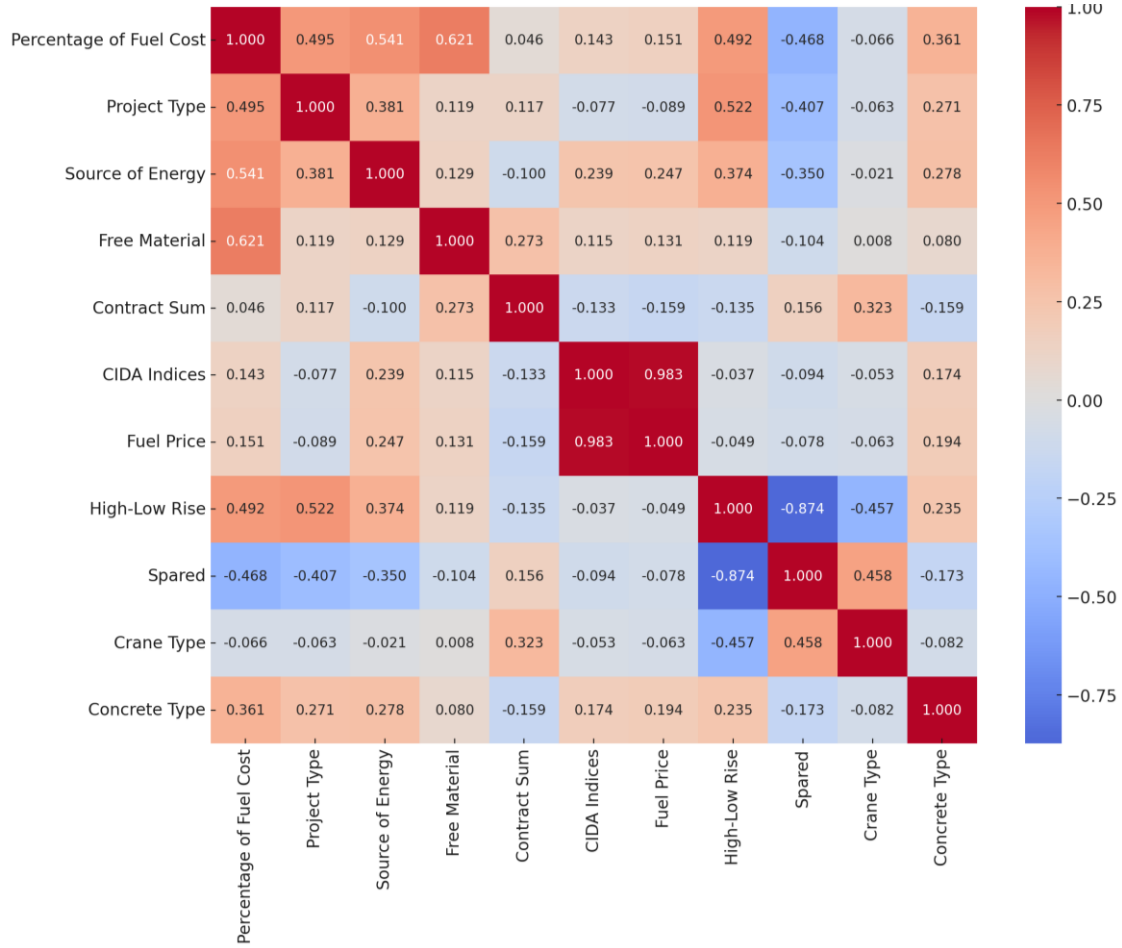

Figure 3: Correlation heatmap for feature selection

Table 1presents the properties and correlation values of each parameter in a tabular format, providing a clear overview of the selected features. This multi-step process resulted in a refined set of features that were subsequently used to train the ML models.

Table 1: Feature Selection.

Feature name	Type	Preprocessing applied	Correlation	Category
Free Material	Numerical	Normalized	0.621	Strong
Source of Energy	Categorical	One hot encoded	0.541	Strong
High-Low Rise	Categorical	One hot encoded	0.492	Strong
Project Type	Categorical	One hot encoded	0.486	Strong

Feature name	Type	Preprocessing applied	Correlation	Category
Project Spread	Categorical	One hot encoded	0.468	Strong
Concrete Type	Categorical	One hot encoded	0.361	Strong
Fuel Price	Numerical	Normalized	0.151	Weak
CIDA Indices	Numerical	Normalized	0.143	Weak
Crane Type	Categorical	One hot encoded	0.066	Weak
Contract Sum	Numerical	Normalized	0.046	Weak

3.4 DATA PARTITIONING

The researchers partitioned the dataset into training and testing subsets to facilitate the evaluation of model performance. The split was performed as follows:

- Training set: 80% of the dataset was allocated to training the ML models
- Testing set: The remaining 20% was reserved for testing and evaluating the generalisation capabilities of the trained models

A random state of 42 was used during the split to ensure that the partitioning was reproducible, which is critical for maintaining the integrity and consistency of the analysis across different iterations.

3.5 MODEL SELECTION AND TRAINING

Three supervised ML models were selected for experimentation based on their relevance to regression tasks, their ability to capture complex relationships, and their successful application in construction cost prediction research. These models, all part of the supervised learning paradigm given the availability of labelled historical data for training, include:

- Random Forest Regressor (RFR) is an ensemble learning method that operates by constructing multiple decision trees during training. It is particularly adept at handling datasets with both categorical and continuous variables, and it mitigates overfitting through its ensemble approach.
- Gradient Boosting Regressor (GBR) is another ensemble technique. Gradient Boosting constructs models sequentially to minimise prediction errors by combining weak learners. It is capable of capturing complex, non-linear relationships within the data.
- Support Vector Regressor (SVR) was included as it is effective for datasets with a smaller number of samples and works well for non-linear relationships through the application of kernel functions.

3.6 MODEL EVALUATION

Each model was evaluated using a set of standard performance metrics:

- Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction.
- Mean Squared Error (MSE) quantifies the average squared difference between the predicted and actual values, penalising larger errors more than smaller ones.

- R-squared (R^2) provides a measure of how well the regression model captures the variance in the data. A higher R^2 indicates a better fit.

The model that exhibited the best performance metrics was selected as the final model for further optimization. The results of each model's performance are presented in Table 2.

Table 2: Model performance

Model Name	MAE	MSE	R-squared
R1 SVR	0.05230	0.002870	-20.08
R2 GBR	0.00861	0.000199	-0.464
R3 RFR	0.00848	0.000197	-0.446

3.7 MODEL OPTIMISATION AND ITERATIVE REFINEMENT

The selected model, identified as R3 (referring to the initial baseline Random Forest Regressor), was further optimised through a progressive series of strategies, culminating in version R8, as detailed in Table 3. R3 served as a baseline with default settings. R4 improved data partitioning through iterative random state search. R5 integrated 5-fold Cross-Validation (CV) for robust performance evaluation. R6 introduced prediction uncertainty via Confidence Intervals (CI) derived from K-Fold inference. R7 applied hyperparameter tuning using Randomised Search CV to optimise model architecture, significantly reducing error. Finally, R8 built on R7 by enforcing fold-specific preprocessing during K-Fold prediction, achieving the tightest confidence interval and highest prediction stability.

Table 3: Optimisation techniques used in Model R3 to R8

Model name	Optimisation techniques applied	Purpose/Benefit
R3	Baseline model using a simple train-test split with default RFR parameters.	Establishes a reference point for model performance with no tuning.
R4	Conducted randomised search over multiple random states (e.g., 0 to 99) for train-test split; selected split based on best performance.	Enhances generalisability by identifying a stable and optimal data partition.
R5	Integrated 5-fold CV during evaluation of the R4 model.	Improves performance reliability by mitigating variance across different training subsets.
R6	Extended R5 by incorporating fold level predictions to estimate CI for new project predictions.	Introduces prediction uncertainty, enhancing interpretability and decision confidence.
R7	Applied Randomised Search CV for hyperparameter tuning of RFR (e.g., n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap) using 5-fold CV.	Achieves optimal model architecture, minimising error and maximising R^2.

Model name	Optimisation techniques applied	Purpose/Benefit
R8	Maintained R7 hyperparameters; added fold-specific preprocessing within each K-Fold iteration to generate predictions.	Delivers the most refined and stable inference process with the tightest CI and lowest variance, while preventing data leakage.

3.8 PREDICTIVE MODELLING FOR FUTURE PROJECTS

Once the models were optimised, they were utilised to predict future cost elements, with a specific focus on fuel consumption for new projects. The models utilised project-related attributes, including contract sum, project type, and fuel price, to generate predictions that inform project planning and budget allocation.

4. RESULTS/ANALYSIS AND DISCUSSION

4.1 RESULTS OVERVIEW

This research aimed to develop an accurate supervised ML model for predicting fuel cost percentages in construction projects. Following the development of the best-performing model, as detailed in the methodology chapter, this section describes the optimised model's evaluation and predicted values for new project parameters.

4.2 MODEL PERFORMANCE AND PREDICTION OUTCOMES

In the initial model selection phase, the Random Forest Regressor consistently outperformed the other models, namely Support Vector Regressor and Gradient Boosting Regressor, by exhibiting lower error rates and providing more accurate fuel cost predictions. As shown in Table 2, the SVR performed poorly, displaying significantly higher errors and a negative R-squared value, indicating its unsuitability for this specific prediction task.

The strengths of the RFR are particularly well-suited to the complexities of construction cost data. As an ensemble method, RFR builds multiple decision trees and averages their outputs. This structure effectively captures complex, non-linear relationships between project attributes and fuel cost components, which are characteristic of the dataset used, combining both numerical (e.g., contract sum, CIDA indices) and categorical variables (e.g., project type, energy source). RFR seamlessly handled these diverse data types without requiring extensive data transformations, unlike SVR, which relies heavily on kernel manipulation. By aggregating predictions from many individual decision trees trained on bootstrapped subsets, RFR provided strong generalization and robustness against overfitting, a common concern with highly variable project datasets. Furthermore, RFR provided intrinsic feature importance scores, which enhance the model's transparency and facilitate insights into which project features most influence fuel cost outcomes, proving valuable for construction management practitioners.

RFR outperformed the other models by exhibiting lower error rates and providing more accurate predictions of fuel costs. In contrast, the SVR performed poorly, displaying significantly higher errors and a negative R-squared value, indicating that it was unsuitable for this specific prediction task.

4.2.1 Optimised Model Performance

After comprehensive hyperparameter tuning on the selected RFR and iterative optimisation, the model's performance improved significantly.

Table 4 illustrates the progressive performance enhancement achieved across different optimisation stages (R3 to R8) for the selected project.

Model R7 demonstrated the highest analytical performance, as evidenced by the lowest Mean Absolute Error and Mean Squared Error alongside the highest R-squared (R^2) value. These metrics indicated that R7 offered the most accurate predictions and best explained the variance in fuel cost outcomes. In contrast, Model R8, while slightly less optimal in terms of MAE and R^2, excelled in predictive stability. It achieved the lowest standard deviation and produced the tightest 90% confidence interval, making its predictions highly consistent and reliable across different data partitions. Accordingly, R7 was the preferred model for analytical accuracy and variance explanation, while R8 was more suitable for practical deployment, where consistent and stable prediction intervals are essential. Therefore, Model R8 is recommended as the overall best model when both predictive accuracy and reliability are prioritized in real-world applications.

Table 4: Modelled performance improvement

Model	MAE	MSE	R^2	Predicted % Fuel Cost	90% Confidence Interval	Standard Deviation
R3	0.00779	0.00018	0.64411	3.74%	Wide	2.68%
R4	0.00417	0.00003	0.89626	4.35%	Wide	2.57%
R5	0.00384	0.00005	0.90581	4.14%	[0.82%, 7.47%]	2.02%
R6	0.01036	0.00038	0.44648	4.59%	[3.80%, 5.39%]	0.51%
R7	0.00294	0.00001	0.95506	4.17%	[3.50%, 4.51%]	0.30%
R8	0.00421	0.00002	0.92916	3.58%	[3.12%, 4.05%]	0.28%
Best	R7	R7	R7	R8	R8	R8

4.3 PREDICTED FUEL COSTS BY PROJECT TYPE

Once the model was optimised (specifically Model R8 for practical deployment), it was applied to predict fuel costs across a range of construction projects. The predicted values were based on the key project attributes identified in the feature importance analysis.

Table 5: Results for Different Project Characteristics presents the expected fuel cost percentages for various project types, including commercial, education, industrial, and road projects. This table provides a concrete demonstration of the model's ability to generate specific fuel cost predictions based on distinct project characteristics.

Table 5: Results for different project characteristics

Project No.	Project type	Actual fuel %	Predicted % of Model R7	90% Confidence intervals	Predicted % of Model R8	90% Confidence intervals
3-06600	Commercial	1.22%	1.33%	[1.26%,1.40%]	1.28%	[1.08%,1.49%]
3-06300	Education	0.93%	1.10%	[1.03%,1.16%]	0.97%	[0.92%,1.02%]

Project No.	Project type	Actual fuel %	Predicted % of Model R7	90% Confidence intervals	Predicted % of Model R8	90% Confidence intervals
3-06900	Education	1.45%	1.51%	[1.35%,1.69%]	1.56%	[1.37%,1.80%]
3-04800	Education	1.53%	1.35%	[1.21%,1.50%]	1.36%	[1.38%,1.99%]
3-04600	Education	1.42%	1.53%	[1.36%,1.66%]	1.42%	[1.30%,1.53%]
3-05800	Industrial	1.37%	1.97%	[1.70%,2.24%]	1.72%	[1.46%,1.97%]
3-06200	Industrial	2.32%	2.28%	[1.93%,2.63%]	2.11%	[1.82%,2.40%]
4-02200	Road	4.63%	3.71%	[3.51%,4.08%]	5.27%	[4.47%,6.07%]

4.4 FEATURE INFLUENCE ON PREDICTION OUTCOMES

The results from the optimised RFR confirmed the importance of certain features, as highlighted previously in the feature importance analysis (refer to Figure 3 and Table 1 in Methodology). By focusing on these critical features, the model provided more accurate and realistic predictions for fuel costs. These findings highlight the model's ability to identify significant drivers of fuel consumption, which is essential for data-driven decision-making in project management.

4.5 PRACTICAL IMPLICATIONS OF THE FINDINGS

The findings from this study had several practical applications for budgeting construction projects, including:

- The ability to accurately predict fuel costs based on project characteristics allowed for more precise budget planning. This is especially important for projects that are fuel intensive, such as road construction.
- Project could use the predictions to allocate resources more effectively, ensuring that fuel consumption is anticipated and accounted for throughout the project lifecycle.
- While the model focused on fuel costs, it was essential to explore adoptability to predict other cost elements, such as labour and equipment, thereby enhancing its applicability across various aspects of project management.

Supervised ML approach to predict other complex and often subjectively estimated cost elements, such as labour and equipment, thereby enhancing its applicability across various aspects of project management. This demonstrates the potential for a broader revolution in construction cost estimation from a subjective to a data-driven paradigm.

5. CONCLUSION

This study on predicting fuel costs for construction projects using Supervised ML techniques successfully demonstrated the possibility of utilising this approach in the presence of mutable variables affecting fuel costs. This underscores the potential for applying the same approach to other cost elements that often rely on human judgment, enabling confident data-driven predictions while overcoming the limitations of subjective estimations. The research marks a significant advancement towards enhancing the precision and reliability of cost estimations in the construction industry.

The historical data generated through the ERP application provided a richly structured dataset. This facilitated not merely overall cost prediction but specifically the prediction of individual cost elements like fuel, which in turn leads to greater accuracy in predicting the overall cost of construction projects.

The adoption of supervised ML techniques yielded several key benefits:

- Enabled the efficient processing and analysis of large-scale datasets, revealing underlying patterns and interdependencies that are not readily detectable through manual techniques
- Facilitated continuous improvement in predictive performance by learning iteratively from historical project data
- Enhanced the capacity to forecast future costs by incorporating dynamic relationships among variables and leveraging temporal trends
- Exhibited adaptability by updating models with newly available data, allowing for timely recalibration in response to shifting market conditions and project-specific variations.
- Reduced reliance on subjective estimations, thereby increasing the objectivity and reproducibility of cost forecasts

This study explicitly addressed key gaps identified in the literature, particularly the neglect of broader economic and dynamic variables by incorporating average fuel prices and CIDA indices, which improved the model's real-world applicability. Furthermore, by focusing on a specific and volatile cost element like fuel, the research provides a concrete example of how supervised ML can be applied to granular cost components, addressing the previous limited focus in this area. The utilization of a comprehensive feature set derived from ERP data also contributes to overcoming the issue of underutilized variables in traditional models.

However, several limitations are associated with the current ML model. The dataset was restricted to projects at a single organizational level, which may capture features related specifically to that organization. This could potentially affect the model's generalizability across different organizations, various geographical regions, or diverse construction types. Additionally, while the models included both project-specific and macroeconomic attributes, the exclusion of real-time or dynamic project updates may restrict the model's responsiveness to changing site conditions during ongoing projects.

Despite these limitations, while the Random Forest Regressor model after optimization demonstrated high predictive accuracy, there are several areas for further research and development:

- Integrating real time fuel consumption data from ongoing projects would provide dynamic updates to the model, further improving its accuracy and usefulness.
- Increasing the size and diversity of the dataset by including more projects, especially those with lower fuel consumption, would improve the model’s generalizability.
- Although the RFR performed well, exploring other ML techniques such as XGBoost or deep learning models could further enhance prediction capabilities, especially for complex cost components like labour or materials.

In conclusion, the application of supervised ML for cost estimation significantly enhanced the accuracy, consistency, and strategic utility of construction budgeting. This

approach not only supported informed decision-making, risk management, and planning but also offered a scalable framework for broader adoption across diverse cost elements. By harnessing both historical and current project data, ML-driven cost prediction frameworks have the potential to transform traditional cost estimation practices and improve overall project management efficiency within the construction sector.

6. ACKNOWLEDGMENT

Certain sections of this manuscript, particularly those involving language refinement and copy editing, were supported through the use of OpenAI's ChatGPT and Grammarly. These tools were employed solely to enhance clarity and readability, ensuring that they did not affect the conceptual or analytical aspects of the research.

The authors thank International Construction Consortium (PVT) Ltd. for providing essential data and resources. Special recognition goes to the ERP team for their assistance in data collection and analysis.

7. REFERENCES

Abed, Y. G., Hasan, T. M., & Zehawi, R. N. (2022). Machine learning algorithms for construction cost prediction: A systematic review. *International Journal of Nonlinear Analysis and Applications, 13*(2), 2205-2218. https://doi.org/10.22075/ijnaa.2022.27673.3684

Miranda, S.L.C., Castillo, E.D.R., Gonzalez, V., & Adafin, J. (2022). Predictive analytics for early-stage construction costs estimation. *Buildings*, *12*(7), 1043. https://doi.org/10.3390/buildings12071043

Elhag, T. M. S., & Boussabaine, A. H. (1998). An artificial neural system for cost estimation of construction projects. In W. Hughes (Ed.), *Proceedings of the 14th annual ARCOM conference* (pp. 219–226). Association of Researchers in Construction Management. Retrieved from https://www.arcom.ac.uk/-docs/proceedings/ar1998-219-226_Elhag_and_Boussabaine.pdf

Hashemi, S. T., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. *SN Applied Sciences, 2*, 1703. https://doi.org/10.1007/s42452-020-03497-1

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Katyare, P., Joshi, S. S., & Rajapurkar, S. (2023). Real-time data modelling for forecasting fuel consumption of construction equipment using an integral approach of IoT and ML techniques. *Journal of Information & Optimization Sciences*, *44*(3), 427–437. https://doi.org/10.47974/JIOS-1363

Mahamid, I. (2011). Early cost estimating for road construction projects using multiple regression techniques. *Construction Economics and Building*, *11*(4), 87–101. https://doi.org/10.5130/AJCEB.v11i4.2195

Prasetyono, P. N., Suryanto, H.M.S., & Dani, H. (2021). Predicting construction cost using regression techniques for residential building. *Journal of Physics: Conference Series, 1899*(1), 012120. https://doi.org/10.1088/1742-6596/1899/1/012120

Ray, S., Alshouiliy, K., & Agrawal, D. P. (2021). Dimensionality reduction for human activity recognition using Google Colab. *Information, 12*(1), 6. https://doi.org/10.3390/info12010006

Saeidlou, S., & Ghadiminia, N. (2024). A construction cost estimation framework using DNN and validation unit. *Building Research & Information, 52*(1-2), 38-48. https://doi.org/10.1080/09613218.2023.2196388

Salleh, H., Wang, R., Affandi, N. Z. H, & Abdul-Samad, Z. (2023). Selecting a standard set of attributes for the development of machine learning models of building project cost estimation. *Planning Malaysia, 21*(5)*,* 110-125. https://doi.org/10.21837/pm.v21i29.1359

Suchetha, N. V., Adithya, Ashik, S., Dhanush, & Guledagudda, T. R. S. (2023). Home construction cost estimation using ML. *IRE Journals, 6*(11)*,* 536-543. https://www.irejournals.com/formatedpaper/1704488.pdf