Artificial Intelligence and Machine Learning Applications in Construction Cost Forecasting: Model Evaluation and Web-Based Deployment

Jaydeep M. Jadhav[1], Dr. J. M. Shinde[2]

[1] PG Student, Civil Engineering Department, Ashokrao Mane Group of Institutions Vathar tarf Vadgaon

[2] Head of Civil Engineering Department, Ashokrao Mane Group of Institutions Vathar tarf Vadgaon

Received: Aug 18, 2025 **Accepted:** Sep 14, 2025 **Published:** Sep 16, 2025

Abstract— **Estimation of construction costs is considered as very important for effective budgeting, planning, and decision-making in civil infrastructure projects. The widely used traditional methods of construction cost estimation often rely on historical trends of the cost analysis, expert judgment, and manual calculations, the processes which can be time-consuming and prone to human error. This study proposes a machine learning-based approach to automate and enhance the accuracy of construction cost prediction using a multi-output regression framework a comprehensive dataset of 10,000 construction instances was compiled from various sources; it includes parameters such as material types, structural categories, area in square feet, and labour specifications among others. Multiple regression models including Random Forest, XGBoost, Gradient Boosting, Decision Tree, and Linear Regression were trained and evaluated using standard performance metrics which include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. Among these, XGBoost outperformed all others with an R^2 score exceeding 0.999. To facilitate easy use, a Flask web application has been developed accordingly to deploy said models and deliver immediate predictions predicated-upon dynamic user inputs thus rendering estimation seamless & facilitating scalability. The proposed solution not only streamlines the estimation process but also provides a scalable and interpretable tool for engineers, contractors, and stakeholders. Experimental results demonstrate that ensemble learning models are highly effective in capturing the nonlinear relationships inherent in construction data, thereby offering a robust alternative to conventional estimation practices.**

Keywords— **Construction Cost Estimation, Machine Learning, Multi-output Regression, XGBoost, Random Forest, Web Application, Civil Engineering, Predictive Modelling.**

I. INTRODUCTION

Advanced computational techniques and machine learning algorithms have improved the traditional and software based construction cost estimation approaches. Various models have been developed to enhance the accuracy and efficiency of cost predictions in construction projects. For example if we consider, the use of improved support vector machines (SVM) optimized by particle swarm optimization (PSO) has shown promising results, achieving an average prediction deviation as low as 1.57% (Zhang & Song, 2022). Similarly, the BP neural network, which minimizes mean square errors through error gradient descent, has been effectively employed to forecast construction costs with high accuracy (Wang, 2018). The integration of Building Information Modeling (BIM) with Elman Neural Networks (ENN) further enhances prediction accuracy by utilizing digital and visual data from intelligent building models, achieving a root mean squared error (RMSE) of less than 75 and a determination coefficient greater than 0.95 (Zhang & Mo, 2023). Additionally, the Random Forest algorithm, optimized by the Bird Swarm Algorithm (BSA), has demonstrated superior performance in predicting construction costs, with a maximum relative error of only 1.24% (Zheng et al., n.d.). The application of artificial intelligence (AI) methods, including neural networks and case-based reasoning, has also been explored to address the challenges of large errors and long preparation times in cost estimation ("Research on Intelligent Prediction of Engineering Cost Based on Artificial Intelligence," 2023). Moreover, the integration of quantum computing techniques with traditional algorithms like Random Forest has been proposed to further improve the efficiency and accuracy of cost predictions (Lakshana et al., 2024). These advancements highlight the potential of combining various machine learning and optimization techniques to develop robust construction cost prediction systems that can adapt to the dynamic conditions of construction projects, ultimately supporting better cost management and decision-making in the industry (Rayabharapu et al., 2025) (Shi & Li, 2010).

Estimating construction costs is a vital aspect of construction management, playing a crucial role in determining project success through effective budgeting, resource allocation, and profitability. The inherently complex nature of construction projects necessitates precise cost estimation due to factors like inflation, market conditions, and specific elements such as structural systems and site environments (Ali et al., 2022; Ameya, 2025; G.C.Sarode, 2020). The adoption of advanced technologies like Artificial Intelligence (AI) and Building

Information Modeling (BIM) has improved the accuracy of these estimations. AI techniques such as extreme gradient boosting and artificial neural networks have shown high effectiveness in predicting costs by managing intricate non-linear challenges effectively (Ali et al., 2022) (G.C.Sarode, 2020). Similarly, BIM-based tools provide more reliable cost assessments than traditional methods by considering diverse aspects beyond just floor space beyond just floor area (Chandra & Yuliana, 2024) (Yang et al., 2022). The importance of continuous training and development for construction estimators is also emphasized, as it equips them with the necessary skills to utilize these advanced tools and methodologies, thereby reducing estimation errors and enhancing project outcomes (-, 2023). Moreover, organizational controls play a significant role in improving cost estimation performance, especially in complex projects, by ensuring that the right control modes are implemented (Fazil et al., 2023). The use of machine learning methods, such as adaptive neuro-fuzzy inference systems and support vector machines, further supports the predictability and accuracy of cost estimations, outperforming traditional methods like earned value management-based approaches (Yalçın et al., 2024). Overall, the development and application of sophisticated cost estimation models are crucial for the effective management and successful completion of construction projects, providing stakeholders with reliable data for informed decision-making (Gilson & Vanreyk, 2016) (Naimi, 2023).

This paper discusses the design and implementation of multiregrrssor construction cost estimation system that integrates various machine learning algorithms and optimization techniques to enhance prediction accuracy and reduce errors in cost estimation. This system aims to streamline the estimation process and improve decision-making efficiency in construction management.

II. Literature Review

Before starting the research work, a thorough literature review was conducted to understand the currently used techniques for construction cost estimation. Zheng et al. proposed a Random Forest model optimized by the Bird Swarm Algorithm (BSA), achieving a maximum relative error of only 1.24% and demonstrating superior enterprise-level forecasting accuracy. Magdum et al. implemented neural networks and multilayer perceptron (MLP)-based models using six key material features, finding MLPs better on training sets while neural networks generalized well on unseen data, with the ELU activation function performing best. Xu et al. introduced a hybrid model using t-SNE for dimensionality reduction and an improved grey correlation algorithm, yielding a 5.1% increase in accuracy and 12.75% efficiency gain. Ye et al. enhanced BP neural networks with a PSO-guided optimization approach, effectively handling local minima and improving forecasting accuracy. Tayefeh Hashemi et al. presented a systematic review of three decades of ML models for cost estimation, recommending hybrid approaches to manage high-risk project uncertainty. Zhang et al. developed a BIM-ENN framework achieving over 95% accuracy in intelligent building projects while effectively processing time-sequential data. Jirait et al. created a GUI-based ML system for house construction cost estimation, emphasizing future integration with user-friendly interfaces. Park and Yun proposed a BIM-based deep learning model for schematic design phase cost prediction, showing higher accuracy by integrating design and building attributes. Yaseen et al. introduced a hybrid GA-ANN-SVM model for predicting construction cost and duration, where GA optimized feature selection, boosting ANN and SVM performance. Similarly, Liu et al. applied a GA-enhanced BP neural network, reporting significant improvements in accuracy and convergence over traditional BP networks.

III. Research Methodology

The development of the construction cost estimation system was executed in four phases: dataset collection and curation, data preprocessing, model development and training, and Flask-based deployment. Each phase is explained in detail below.

A. Dataset Collection and Curation

To construct a reliable prediction system, a comprehensive dataset comprising 10,000 instances was collected from multiple sources. These sources included government tender websites, construction contractor records, publicly available real estate project estimates, and market rate sheets for construction materials. The primary aim was to cover a wide range of project types, locations, and material quality grades to ensure generalization. Each data

entry captured key input attributes such as the category of location (urban, semi-urban, rural), built-up area in square feet, quality indicators for various construction components like cement, steel, bricks, and flooring, and the expected duration for project completion. The target outputs were detailed and included estimated quantities and associated costs of individual materials, as well as the total construction cost. This dataset provided a rich foundation for training a robust machine learning model capable of predicting detailed cost estimates for varying construction scenarios.

B. Data Preprocessing

Before we use the data to train the machine learning pipeline, preprocessing was carried out to improve consistency and model compatibility. All categorical variables representing material quality were encoded numerically (e.g., “Basic Grade” = 0, “Medium Grade” = 1, “Premium Grade” = 2). Missing or inconsistent entries were handled through imputation or removal, and feature scaling was performed where necessary. The dataset was then split into features (input variables) and multiple target outputs (quantities and costs). An 80:20 train-test split was applied to ensure fair model evaluation.

C. Model Development and Training

For predictive modeling, multiple machine learning algorithms were implemented using the MultiOutputRegressor wrapper to handle the multi-target regression task. The models included RF Regressor, Gradient Boosting Regressor, XGBoost, Decision Tree Regressor and Linear Regression. Each model was trained on the same training set and evaluated using key regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared Score (R^2). Among these models which we trained, the XGBoost Regressor achieved the best overall performance giving the lowest MAE and RMSE values and the highest R^2 score, making it the ideal candidate for deployment.

D. Deployment Via Flask Application

To make the system accessible to non-technical users such as civil engineers and contractors, a user-friendly web interface was developed using the Flask framework. The interface provides a secure login and allows users to input various construction project parameters via dropdowns and text fields. Upon submission, the application processes the input through the trained XGBoost model and displays predicted quantities and costs for individual materials, along with the total estimated cost. This real-time prediction system enables users to efficiently plan budgets and make material procurement decisions.

The architecture diagram of the flow of the project is as shown below in Figure 1 below:

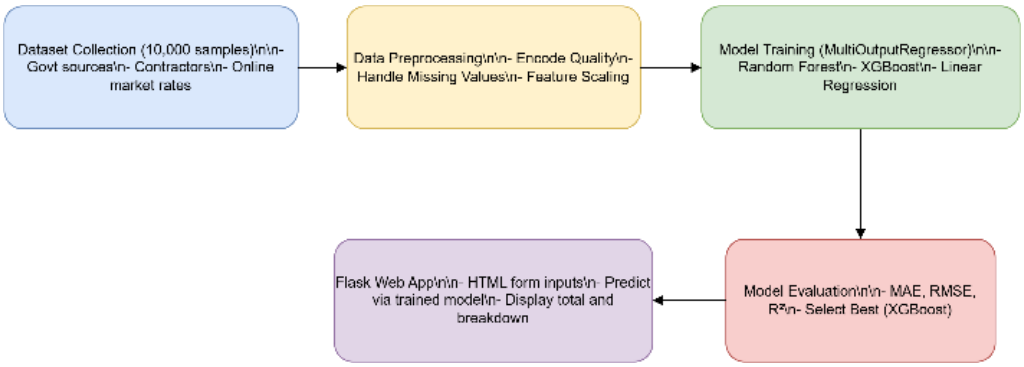

Fig.1: Architecture Diagram of the System

The proposed system architecture begins with collecting a diverse dataset of 10,000 construction cost records from government rate schedules, contractor cost sheets, and online market rates to capture regional and quality-based variations. The data then undergoes preprocessing, including label encoding for categorical material grades, imputation of missing values, and feature scaling to ensure consistency and readiness for modeling. A multi-output regression approach was applied using algorithms such as Random Forest, XGBoost, Linear Regression, and Gradient Boosting, with XGBoost achieving the best performance based on R^2 and error metrics. Finally, the trained model was deployed through a Flask-based web application with a user-friendly HTML interface, enabling users to input project parameters and receive detailed material-wise cost predictions, making the tool practical for builders, engineers, and homeowners.

IV. Results and Discussion

The dataset, containing 10,000 entries of construction project parameters, underwent comprehensive exploratory data analysis to derive insights and validate its suitability for training predictive models. The univariate analysis was done to handle or visualize the distribution of each and every variable in the dataset.

A. Exploratory Data Analysis

The EDA was done on different input variables. The distribution of the PlaceCategory feature indicated a balanced representation of city types (Expensive, Medium, and Economical zones). Histograms plotted for numerical features such as SquareFeet, CementCost, SteelCost, and TotalCost demonstrated positively skewed distributions, reflecting real-world variability in material requirements and regional pricing. The univariate analysis concluded with different plots for all the parameters which is shown below in Figure 2.

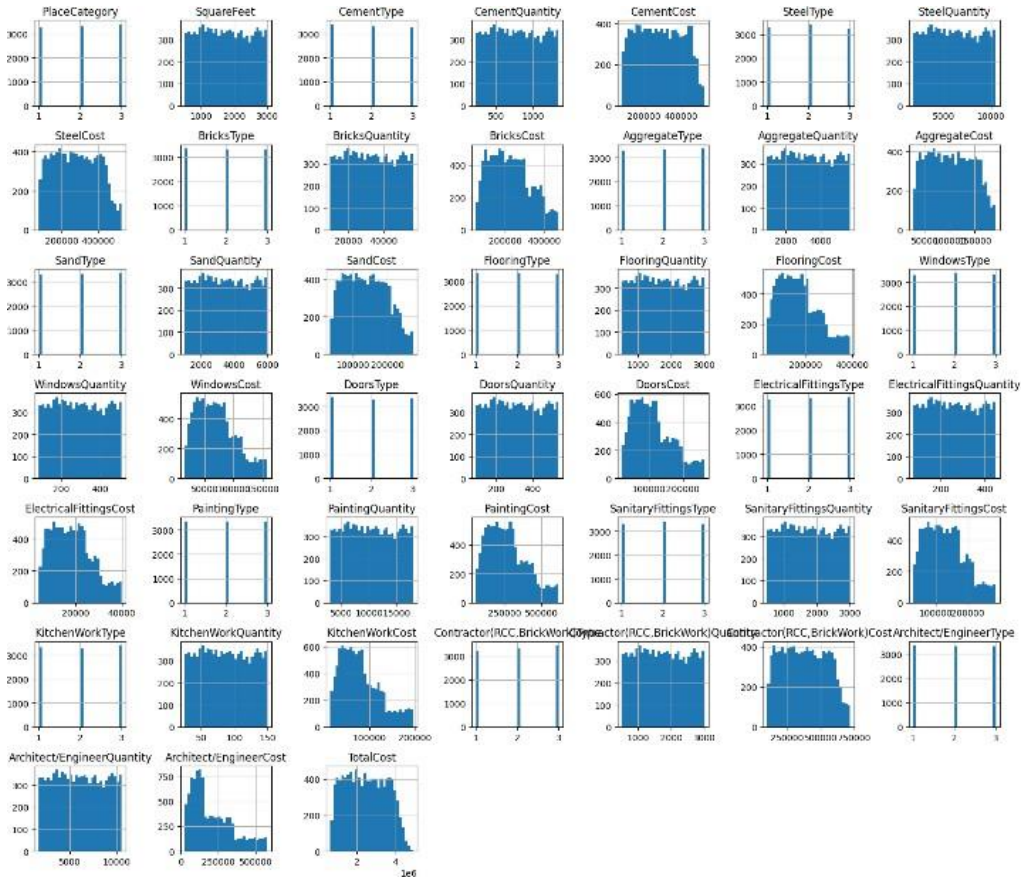

Fig.2: Univariate Analysis of Dataset

After that the corelation matrix was plotted as shown in Figure 3. A label-encoded correlation heatmap revealed strong linear dependencies between corresponding quantity and cost features (e.g., SteelQuantity and SteelCost), affirming the model's ability to capture logical pricing trends. Notably, TotalCost exhibited high correlation with variables like SteelCost, CementCost, and SquareFeet, indicating their dominant influence in overall expenditure.

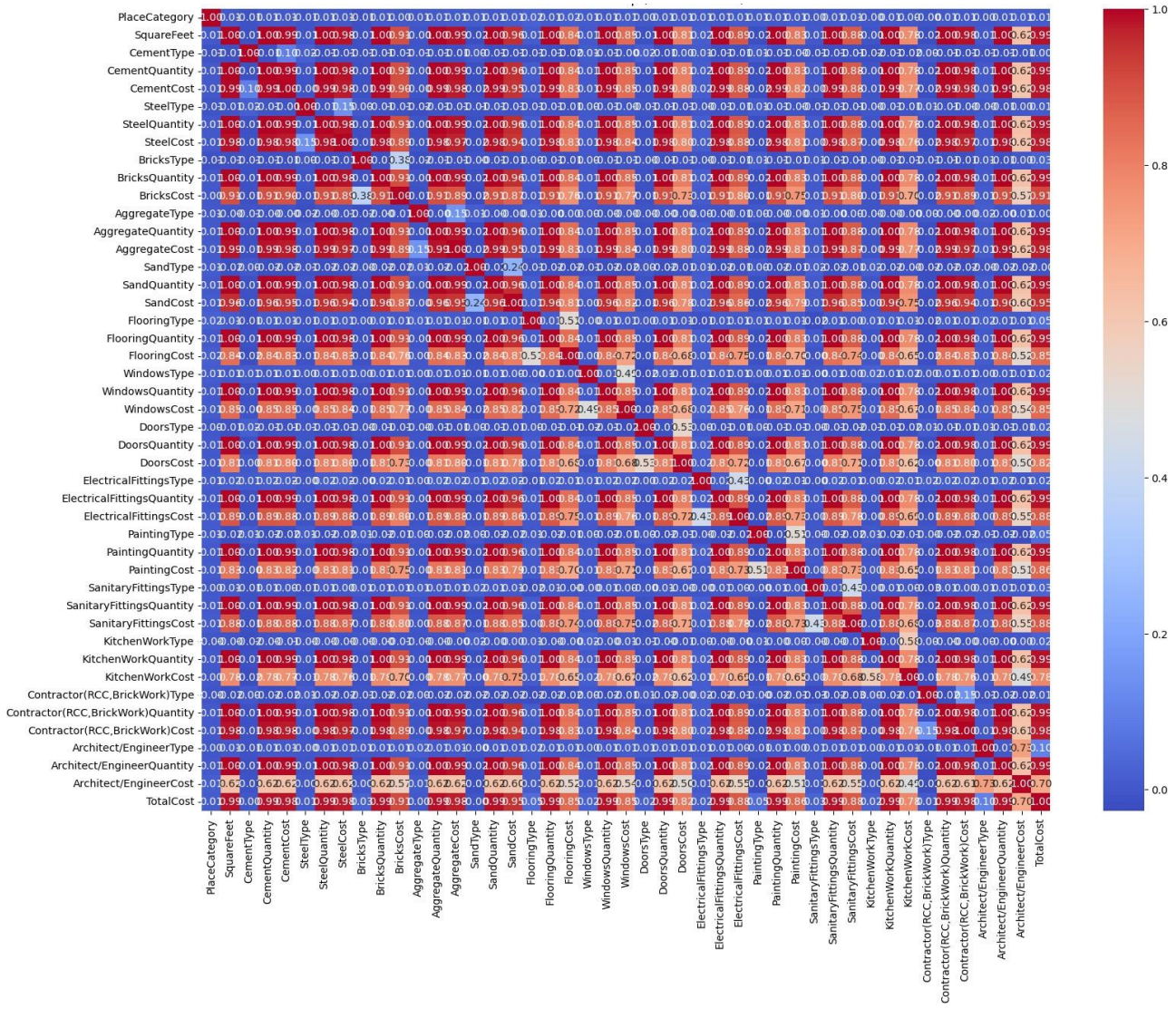

Fig.3 Correlation Matrix

The dataset analysis revealed consistent multivariate trends between material quantities and their associated costs, as confirmed through pair plots and KDE-based scatter matrices (e.g., CementQuantity vs. CementCost and SteelQuantity vs. TotalCost), highlighting proportional scaling and meaningful dependencies. Distributional assessment using histograms and boxplots indicated that TotalCost was slightly right-skewed with a wider spread in higher-cost city categories, aligning with real-world scenarios where factors like material grade, region, and square footage drive nonlinear cost increases. Skewness and kurtosis analysis further showed moderate right skew in variables such as SteelCost, CementCost, and TotalCost, reflecting a few high-cost outliers, while kurtosis values below 3 for most cost-related attributes suggested platykurtic distributions with lighter tails and fewer extreme values compared to a normal distribution.

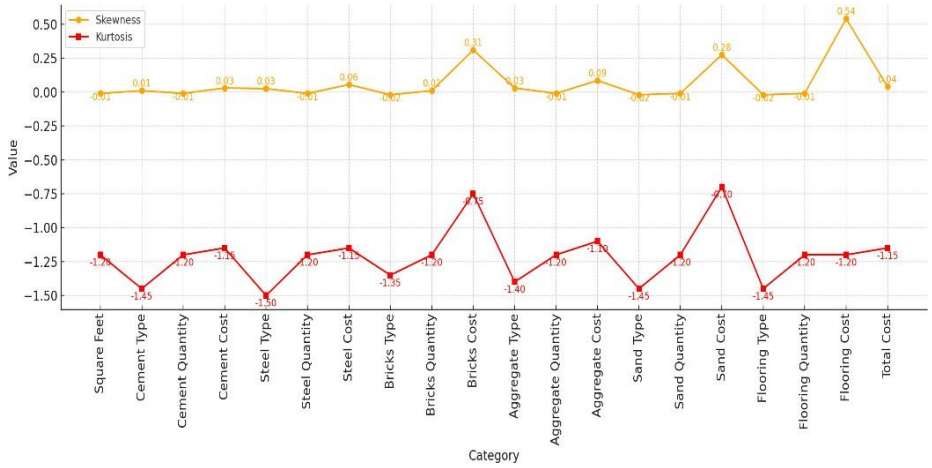

Fig.4: Skewness and Kurtosis Plot

In terms of model performance, tree-based models like Random Forest and XGBoost were resilient to these distribution irregularities and performed robustly due to their non-parametric nature. In contrast, Linear Regression displayed reduced accuracy ($R^2 \approx 0.983$) as it is sensitive to both skewness and outliers due to its assumption of linearity and constant variance (homoscedasticity). Although data transformations like log-scaling or Box-Cox transformation could further normalize the features, they were not required, as ensemble models performed sufficiently well on the raw data. The Dataset and its impact on model training is given in Table 1.

TABLE I

IMPACT ON MODEL TRAINING AND RELATED OBSERVATIONS

Metric	Interpretation	Observation in Dataset	Impact on Model Training
Skewness	Measures asymmetry of data distribution	Moderate right skew in TotalCost, CementCost, SteelCost	Tree-based models unaffected; Linear Regression performance degraded due to skew sensitivity
Kurtosis	Measures tailedness; >3 = peaked, <3 = flat	Values < 3 → platykurtic distribution (light tails)	Minimal impact on tree-based models; further optimization via transformation not necessary
Linear Regression	Assumes linearity and normality	Underperformed with $R^2 \approx 0.983$	Affected by skew/outliers due to parametric nature
Ensemble Models	Robust to non-normality and skewed features	Delivered strong performance even with skewed data	No need for transformation (e.g., Box-Cox, log scaling) due to inherent resilience

Following exploratory analysis, a structured training phase was conducted using five regression algorithms to evaluate their performance on the construction cost dataset. The models implemented include Random Forest, Gradient Boosting, Decision Tree, Linear Regression, and XGBoost, all wrapped in a MultiOutputRegressor to simultaneously predict multiple dependent variables (quantities and costs of materials, and total cost).

B. Model Training and Evaluation

Standard regression metrics were used to evaluate on the 20 percent data post training. The performance obtained by training on different algorithms is given in Table 2.

TABLE II

PERFORMANCE EVALUATION METRICS

Model	MAE	RMSE	R^2 Score
Random Forest	Low	Low	~0.9998
Gradient Boosting	Lower	Lower	~0.9997
Decision Tree	Higher	Higher	~0.9996
Linear Regression	Highest	Highest	~0.983
XGBoost	Lowest	Lowest	~0.9999

From the performance results, XGBoost emerged as the best-performing model with the lowest MAE and RMSE and the highest R^2 Score (~0.9999), indicating an exceptional fit to the data. The Random Forest model also showed competitive performance with only a marginally lower score. On the other hand, Linear Regression, while simpler and interpretable, yielded a significantly lower R^2 score, making it unsuitable for high-accuracy cost breakdown predictions.

The accompanying bar plot visualization helped in comparative analysis of all models across the three metrics. XGBoost consistently performed best across all categories, justifying its selection as the final deployment model for the Flask application.

Fig. 5: Application for Construction Cost Estimation

V. CONCLUSION

In this study, we developed a machine learning-based system for construction cost estimation using a rich dataset of 10,000 records collected from various sources. After performing thorough pre-processing and exploratory data analysis, including skewness and kurtosis assessments, we trained and compared multiple regression models: Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. Among these, the XGBoost model demonstrated superior performance with the lowest MAE and RMSE and the highest R^2 score, making it the optimal choice for multi-output cost prediction.

To facilitate practical application, we integrated the trained model into a Flask-based web application with a user-friendly interface, allowing users to input construction parameters and receive detailed cost estimates

instantly. This work bridges the gap between complex ML models and real-world usability in the civil engineering domain.

The results affirm that machine learning, particularly ensemble-based methods, can effectively and accurately estimate granular and total construction costs, thus supporting architects, contractors, and planners in making data-driven decisions. Future enhancements could include dynamic market integration for real-time rate updates and the use of Explainable AI techniques for better interpretability of cost components.

REFERENCES

[1] Zheng, Z., Zhou, L., Zhou, L. "Construction Cost Prediction System Based on Random Forest Optimized by the Bird Swarm Algorithm," Mathematical Biosciences and Engineering, 20(8), pp. 15044–15074, 2023. https://doi.org/10.3934/mbe.2023674

[2] Magdum, S. K. Adamuthe, A. C. "Construction Cost Prediction Using Neural Networks," ICTACT Journal on Soft Computing, 8(1), pp. 1549–1556, 2017.

[3] Xu, Y., Cao, S. "Building Engineering Cost Prediction Model Based on TSNE and Improved Grey Correlation Algorithm," Procedia Computer Science, 228, pp. 957–965, 2023.

[4] Ye, D. "An Algorithm for Construction Project Cost Forecast Based on Particle Swarm Optimization-Guided BP Neural Network," Scientific Programming, 2021, Article ID 4309495.

[5] Hashemi, S. T. Ebadati, O. M. Kaur, H. "Cost Estimation and Prediction in Construction Projects: A Systematic Review on Machine Learning Techniques," SN Applied Sciences, 2, Art. 1703, 2020. https://doi.org/]

[6] Zhang, Y., Mo, H. "Intelligent Building Construction Cost Optimization and Prediction by Integrating BIM and Elman Neural Network," Heliyon, 10(5), p. e37525, 2024.

[7] Jirait, S., Rede, S. "House Construction Cost Prediction," International Journal of Novel Research and Development, 9(5), pp. h311–h314, 2024.

[8] Park, D., Yun, S. "Construction Cost Prediction Using Deep Learning with BIM Properties in the Schematic Design Phase," Applied Sciences, 13(12), Art. 7207, 2023.

[9] Yaseen, M., Shaker, A., Altaee, S. "Hybrid GA-ANN-SVM Model for Construction Cost and Time Estimation," International Journal of Engineering and Technology, 7(3.20), pp. 1142–1147, 2018.

[10] Liu, C., Li, Q., Wang, C. "Engineering Cost Estimation Model Based on GA-Optimized BP Neural Network," SHS Web of Conferences, 170, p. 02009, 2023.

[11] Zhang, X., Song, J. "A Whole Process Cost Prediction System for Construction Projects Based on Improved Support Vector Machines," 2022. https://doi.org/10.46300/9106.2022.16.34

[12] Wang, X. "Forecasting Construction Project Cost Based on BP Neural Network," 2018. https://doi.org/10.1109/ICMTMA.2018.00109

[13] Zhang, Y., Mo, H. "Intelligent Building Construction Cost Prediction Based on BIM and Elman Neural Network," 2023. https://doi.org/10.21203/rs.3.rs-3226303/v1

[14] Zheng, Z., Zhou, L., Zhou, L. "Construction Cost Prediction System Based on Random Forest Optimized by the Bird Swarm Algorithm," Mathematical Biosciences and Engineering, [online] Available at: https://doi.org/10.3934/mbe.2023674 [Accessed: 14 July 2025].

[15] "Research on Intelligent Prediction of Engineering Cost Based on Artificial Intelligence," 2023. https://doi.org/10.25236/ijndes.2023.070110

[16] Lakshana, J., Madhumitha, L., Dharshini, S. L. P. Kumar, T. R. "Cost Prediction for Home Construction Using Quantum Computing," 2024. https://doi.org/10.4018/979-8-3693-3601-4.ch006

[17] Rayabharapu, V. K. Rao, K. D. Punitha, S., Abbas, S. H. Sivaranjani, L. "Enhancing Construction Project Cost Predictions using Machine Learning," 2025. https://doi.org/10.2139/ssrn.5080704

[18] Shi, H., Li, W. "A Web-Based Integrated System for Construction Project Cost Prediction," 2010. https://doi.org/10.1007/978-3-642-05173-9_5

[19] Ali, Z. H. Burhan, A. M. Kassim, M., Al-Khafaji, Z. H. K. "Developing an Integrative Data Intelligence Model for Construction Cost Estimation," 2022. https://doi.org/10.1155/2022/4285328

[20] Ameya, F. M. "Evaluation of the Ongoing Use of Approximate Construction Cost Estimation," 2025. https://doi.org/10.20944/preprints202501.1900.v1

[21] Sarode, G. C. E. C. S. "To Study Cost Prediction Analysis of Construction Project Using ANN Model and SVM by MATLAB," 2020.

[22] Chandra, A., Yuliana, C. "Estimation of Cost Budget Using BIM in School Construction," 2024. https://doi.org/10.20527/crc.v8i5.13421

[23] Yang, S.-W., Moon, S., Jang, H., Choo, S. Y. Kim, S.-A. "Parametric and BIM-Based Cost Estimation for Construction Projects," 2022. https://doi.org/10.3390/app12199553

[24] A. R. A. – "Training and Development for Construction Estimators," 2023. https://doi.org/10.36948/ijfmr.2023.v05i06.10915

[25] Fazil, M. W. Tamyez, P. F. M. Lee, C. K. "Enhancing Cost Estimation Performance Through Effective Control," 2023. https://doi.org/10.1080/15623599.2023.2286048

[26] Yalçın, G., Bayram, S., Çıtakoğlu, H. "Evaluation of Earned Value Management-Based Cost Estimation," Buildings, 14(12), 372, 2024. https://doi.org/10.3390/buildings14123772

[27] Gilson, N. K. Vanreyk, A. J. "Review of Cost Estimation Models," 2016. https://doi.org/10.70729/ijser15714

[28] Naimi, S. "Hybrid Importance Regression Ensemble for Cost Estimation," 2023. https://doi.org/10.22306/al.v10i2.372