Random Forest for Price Prediction in Civil Construction Stages

Remiko H. Albino-Gonzales [1], Segundo E. Cieza-Mostacero [1], Edward A. Vega-Gavidia [1]

[1] *School of Systems Engineering, César Vallejo University, Larco Avenue 1770, Trujillo, Peru*

***Abstract* – The main aim of this research is to present the development and validation of a supervised learning model, using Random Forest algorithms, aiming the unit price prediction in civil construction projects and their stages. This approach emerges from the need to improve the accuracy of cost estimates in civil engineering works, considering that the process of procuring materiel prices to develop a quotation faces limitations due to market instability and the complexity of historical data. It was hypothesized that applying supervised learning algorithms could generate more accurate and efficient predictions compared to traditional price acquisition processes. This hypothesis was tested through an experimental design with comparative groups: one that applied the traditional price acquisition process and another that used the developed predictive model. The results showed a reduction in the error percentage from 11.61% to 6.18%, a similar decrease in the time required for quotation development from 522 to 199 seconds, and an increase in accuracy from 89.6% to 94.2%, with statistically differences (p<0.001). Consequently, the conclusion is reached that the use of prediction-based models represents an effective and adaptable tool for cost management in civil works in the province of Trujillo, Peru.**

***Keywords* - Cost prediction, Random Forest, civil works, supervised learning, efficiency.**

DOI: 10.18421/TEM153-35
https://doi.org/10.18421/TEM153-35

Corresponding author: Remiko H. Albino-Gonzales,
School of Systems Engineering, César Vallejo University, Larco Avenue 1770, Trujillo, Peru
Email: rhalbino@ucvvirtual.edu.pe

Received: 06 August 2025.
Revised: 15 February 2026.
Accepted: 20 February 2026.
Published: 27 August 2026.

1. Introduction

Accurate cost estimation is a key element in ensuring the economic viability and operational efficiency of civil works. However, this process presents limitations when methods based on referential cost analysis are used, as they lack automated predictive capabilities and often result in high error margins, extended processing times, and outcomes poorly aligned with actual market conditions [1]. In large-scale civil works, it has been reported that estimation errors can lead to cost overruns of up to 9% of the total investment [2].

In the Peruvian context, studies have shown that the absence of automated tools for price estimation generates an average error margin of 15%, also prolonging time needed to develop structural quotations [3]. Current available solutions, such as spreadsheets, dedicated software as S10 Presupuestos, or Microsoft Project, may help organize information, but they do not incorporate predictive features that analyze historical data or anticipate market fluctuations in real time.

In response to this issue, recent research has explored emerging technologies such as machine learning, artificial neural networks, and algorithms like Random Forest to optimize cost projection processes. These tools have demonstrated increased estimation accuracy, with mean errors below 7% and results that are adaptable to complex and dynamic contexts [4], [5].

This research proposes the development of a web application based on predictions specifically using the Random Forest algorithm to improve material price estimation in civil works in the province of Trujillo. The study focuses on the development of a system that processes real data from technical files available on the Peruvian SEACE portal, performs real-time automatic predictions, and allows users to generate a detailed PDF report.

The central research question is:

How does a web application based on Random Forest influence cost estimation in civil works in the year 2025?

In addition, a general hypothesis is formulated regarding the expected results of the study:

H0: If a web application based on Random Forest is used, then cost estimation in civil works in Trujillo improves in the year 2025.

The next specific hypotheses are also established:

H1: If a web application based on Random Forest is used, then the percentage error decreases, if a web application based on Random Forest is used, then the time dedicated to the preparation of the structural quotation decreases.

H2: If a web application based on Random Forest is used, then the accuracy of cost estimation increases.

2. Objectives

The general objective of this research is to improve cost estimation in civil engineering works in the province of Trujillo through a web application powered by Random Forest algorithms.

To this end, the following specific objectives were established: To reduce the percentage error in cost estimation, to decrease the time spent on the preparation of structural quotations, and to increase the accuracy of these calculations.

3. Literature Review

To develop predictive models applied to unit price estimation in construction, it is essential to understand certain key concepts that support both the methodological structure and the logic of the proposed system. Below are the essential definitions that contextualize the technical approach of this proposal as well as its application in the construction environment.

3.1. Supervised Learning

Supervised learning is a machine learning approach in which a model is trained on labeled historical data to learn input–output relationships and generate predictions for new data. The goal is for the system to recognize data patterns and predict values with greater accuracy, even in complex or highly variable scenarios [6].

3.2. Random Forest

An algorithm based on the generation of multiple decision trees, which work together to produce a more robust prediction: This method is particularly useful when handling highly variable data, as it helps reduce overfitting and improves accuracy in real-world contexts, such as price estimation in construction [7].

3.3. Cost Estimation

It is a process by which the cost of an investment is calculated in advance. It involves gathering technical and economic information to anticipate the amount to be invested in a specific resource [8].

3.4. Quotation

Represents the process of calculating and presenting the estimated monetary value of a project, service, or product prior to its execution. In the construction context, a quotation includes a detailed breakdown of materials, quantities, and projected costs based on technical requirements [9]. It is fundamental for analyzing the financial feasibility of a project and serves as a basis for investment decisions, contracting services, or managing resources [10].

4. Methodology

This was applied research, as it aimed to solve existing practical problems through the creation of a technology-based solution grounded in theory. The research was characterized by the use of empirical analyses to produce data within a specific area of study [11]. A pure experimental research design was employed, as the study sought to evaluate the influence of the independent variable on the dependent variable by using a control group and a specific experimental group [12].

The independent variable was the web application based on Random Forest, defined as a computing platform that uses artificial intelligence algorithms specifically Random Forest, to process technical information and estimate the costs of civil works [13]. This variable was measured using the indicator of presence or absence of tool execution, according to a nominal measurement scale. Therefore, it is considered that such applications enable the anticipation of future scenarios and the recognition of complex patterns in key construction resource management data [14].

The dependent variable was the accuracy and efficiency in cost estimation, understood as the level of precision achieved in cost projections and the optimization of resource use. Three indicators were used to measure this variable: The percentage error in the estimates, the time spent on the preparation of quotations, and the accuracy of the obtained estimates [15]. Ultimately, the study demonstrated that the use of Random Forest in cost management increases accuracy and reduces error margins in civil works [16].

The population consisted of quotations for structural works carried out in the province of Trujillo, where participants had experience in preparing quotations for residential construction. This area is undergoing continuous housing expansion and renovation. The sample consisted of 60 structural quotations, divided into two groups: 30 structural quotations performed using mathematical calculations (control group), and 30 structural quotations prepared using the web application (experimental group). This approach enabled the analysis of how emerging technologies influence decision-making processes in construction [17].

Data collection was carried out through direct observation, which facilitated the systematic recording of errors, time spent, and accuracy in structural cost estimation. A structured observation form, specifically designed for this purpose, was employed.

Data analysis was conducted in two stages. First, descriptive statistics were applied using summary tables that included means, medians, and standard deviations to identify initial differences between the two groups. Second, inferential statistics were carried out, beginning with the Shapiro–Wilk normality test, since the sample size was fewer than 50 cases. Based on the results, appropriate statistical tests were selected to evaluate whether the observed differences between the two groups represented statistically significant margins of variation [18].

The development of the solution followed a three-layer architecture: frontend, backend, and database. The frontend built using web technologies (HTML, CSS, and JavaScript) allowed users to enter project data segmented by stage. This information was then sent to the backend, implemented using Flask, and subsequently to the API responsible for processing the data and performing the predictions. The predictive engine used a Random Forest model trained with historical records previously stored in SQL Server.

The complete data flow process from the user interface to PDF export is illustrated in Figure 1, which shows how the system automates the calculation of estimated material costs by sending input data to the predictive model and returning an adjusted output for final presentation.

The Random Forest model was trained using a historical dataset that included variables such as material type, execution stage, actual cost, and project conditions. The data were preprocessed to eliminate inconsistencies, group stages, and normalize relevant numerical fields.

Subsequently, the system was configured to generate each prediction based on the specific material and stage, producing individual estimated values without relying on generalized averages.

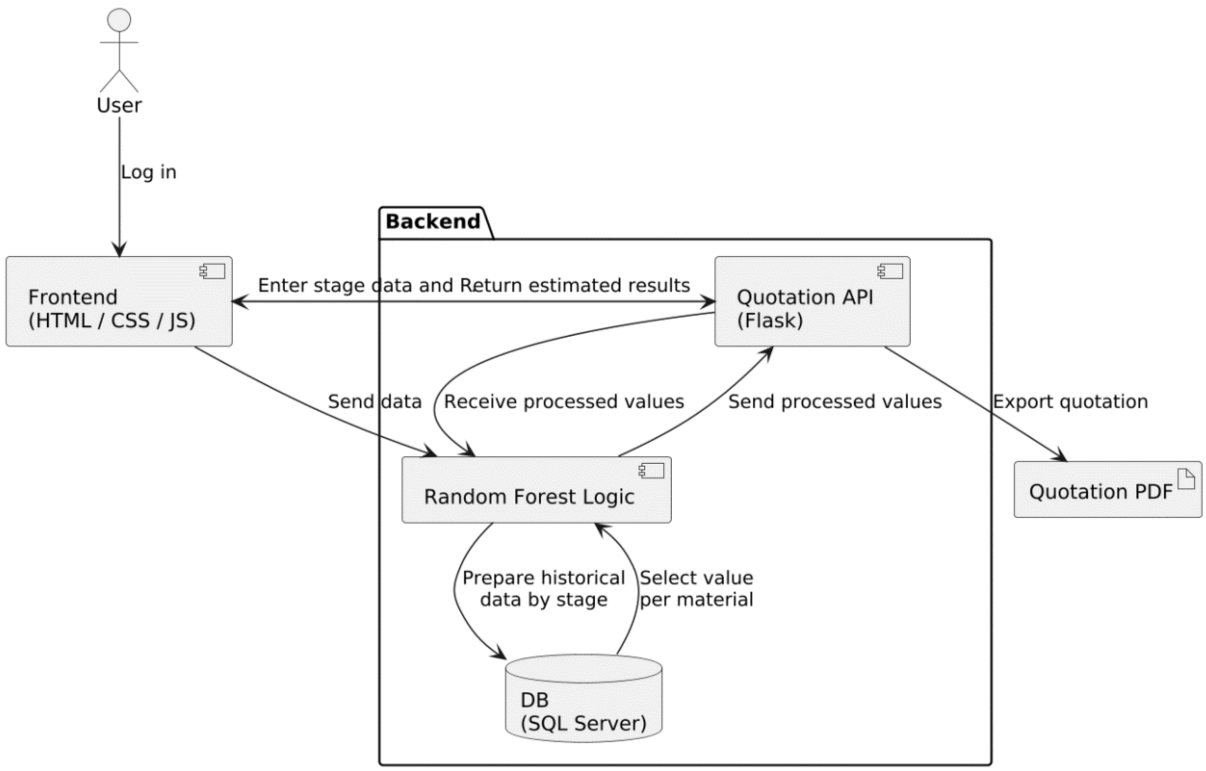

Figure 1. Web system architecture

5. System Development

To implement the proposed solution, a technological approach was undertaken, aimed at designing and developing a web application capable of estimating unit prices in civil construction stages through the use of supervised learning. This section outlines the step-by-step development process, from planning to model training, covering both the system architecture and the key functionalities that ensure proper operation.

5.1. Planning

Development roles and activities were assigned, and four user stories were defined based on the functional requirements identified during the analysis phase. Each story was assigned a high priority due to its impact on the development of the application. Finally, a review schedule was established for each user story (Table 1).

Table 1. User story assessment schedule

Story	Name	Priority	Date
H001	Generate quotation by stage	High	15-09-2024
H002	Train Random Forest model	High	20-09-2024
H003	System access	High	22-09-2024
H004	View results	High	25-09-2024

5.2. Design

User interface prototypes, the database relationship diagram, and the overall system structure were defined. The database was implemented using SQL Server and includes tables such as: (users), to store information of those accessing the system (quotations), to record general data of each estimation process; (quotation_detail), where materials and costs are broken down by stage; and (costs), which contains historical values used to train the predictive model.

Figure 2 shows the diagram representing the logical architecture of data storage within the system.

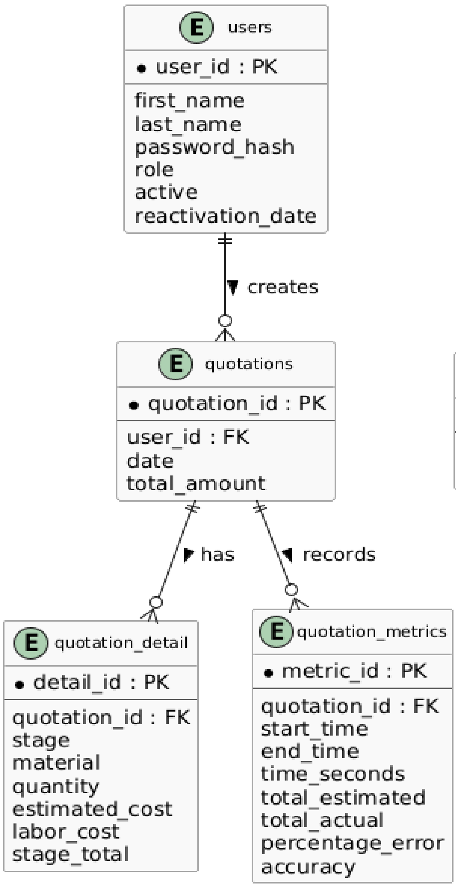

Figure 2. Database diagram

5.3. Coding

The backend development was carried out using the Flask framework, which enabled the implementation of an application programming interface (REST API) that connects different software systems over a network to process system requests. Python was used as the main programming language due to its compatibility with Random Forest libraries such as scikit-learn.

The user interface was built using HTML, CSS, and JavaScript with a responsive design to support various screen resolutions. Upon login, the user is presented with a form where they can select the project stages and define the required quantities of materials, as shown in Figure 3, which displays the main view of the system.

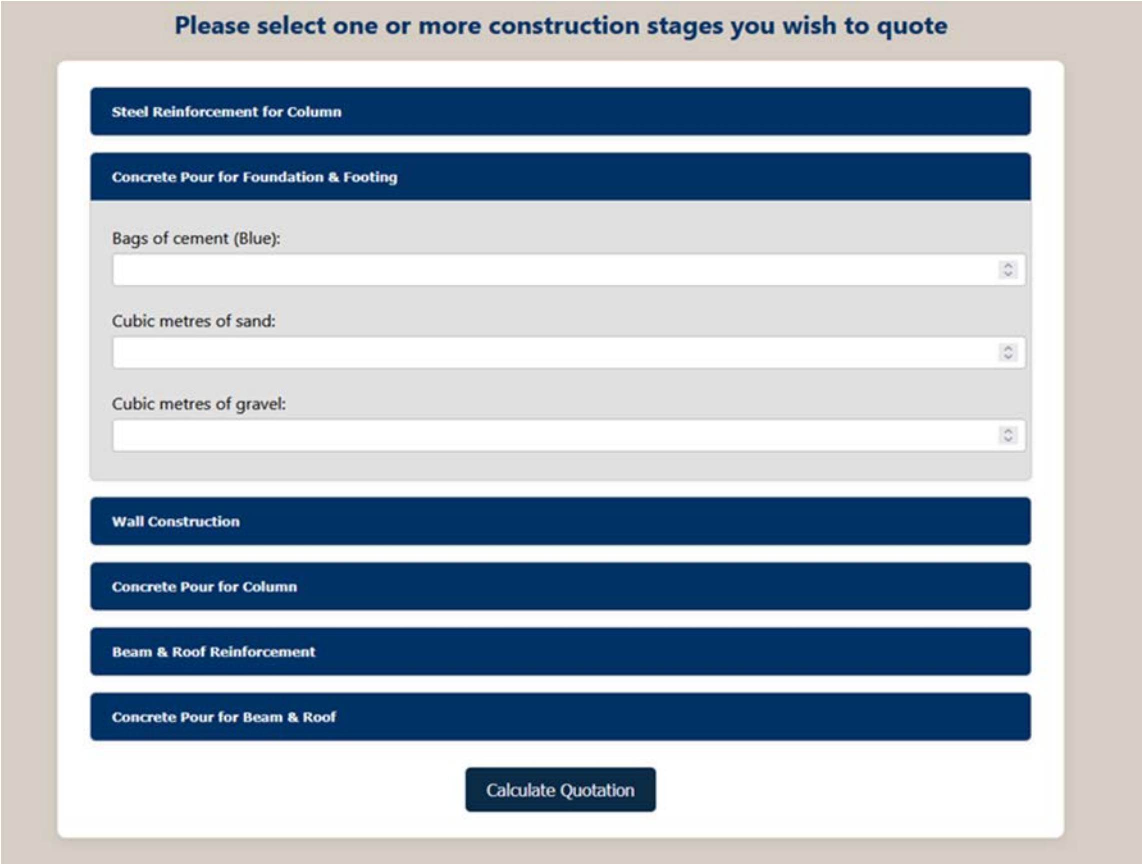

Figure 3. Materiel quantity input interface for stage quotation

After entering the data and clicking Calculate Quotation, processes the materials and automatically generates a detailed stage-by-stage summary, displaying the estimated cost of materials, labor cost, and total cost per stage, along with the overall quotation value.

This functionality is illustrated in Figure 4, which highlights the clarity of real-time result presentation to the user.

COTICONSTRUYE

Stage-by-Stage Quotation Summary

Construction Stages	Materials (PEN)	Labour (PEN)	Total Cost (PEN)
Steel Reinforcement for Column	PEN 1 027.20	PEN 980.00	PEN 2 007.20
Concrete Pour for Foundation & Footing	PEN 1 316.20	PEN 950.00	PEN 2 266.20
Wall Construction	PEN 7 645.50	PEN 2 250.00	PEN 9 895.50
Concrete Pour for Column	PEN 1 316.20	PEN 1 100.00	PEN 2 416.20
Beam & Roof Reinforcement	PEN 454.20	PEN 400.00	PEN 894.20
Concrete Pour for Beam & Roof	PEN 1 316.20	PEN 1 200.00	PEN 2 516.20
		Grand Total	PEN 19 995.50

Figure 4. Results reporting interface for housing construction stages

5.4. Predictive Model Training

To automate the estimation of unit prices for materials, a regression model based on the Random Forest algorithm was implemented. The model was trained individually for each idmaterial using historical records from the database. It enables accurate prediction of the prices to be used in quotations without requiring the user to input an estimated value.

Figure 5 shows the code snippet responsible for training the model and predicting the values. The code implements a Python function that trains a Random Forest regression model for each material (idmaterial) registered in the database. Initially, historical price data are retrieved by idmaterial, then the data are split, and a separate model is trained for each one. Finally, the estimated unit price generated by the model is stored in a dictionary to be used during the quotation process.

```
def train_model()
    global estimated_values, last_training_time

    query = "SELECT material_id, unit_price FROM costs"
    df = pd.read_sql_query(query, engine)

    if df.empty:
        print("⚠  The 'costs' table is empty—model training skipped.")
        return False

    estimated_values.clear()  # Reset any prior results
    materials = df["material_id"].unique()

    for material_id in materials:
        df_material = df[df["material_id"] == material_id]


        if len(df_material) < 2:
            print(f"⚠  Not enough data to train model for material_id {material_id}")
            continue

        X = df_material[["material_id"]]
        y = df_material["unit_price"]

        X_train, _, y_train, _ = train_test_split(X, y, test_size=0.2, random_state=42)

        model = RandomForestRegressor(n_estimators=100, max_depth=5, random_state=42)
        model.fit(X_train, y_train)

        predicted_price = model.predict([[material_id]])[0]
        estimated_values[material_id] = round(predicted_price, 2)

    print("\n🔍 Random Forest estimated unit prices per material_id:")
    for material_id, price in estimated_values.items():
        print(f"  material_id {material_id}: S/ {price:.2f}")

    last_training_time = datetime.now().strftime("%Y-%m-%d %H:%M:%S")
    return True


if not train_model():
    raise RuntimeError("Model training failed due to lack of data.")
```

Figure 5. Random Forest training method source code snippet

6. Findings

This section presents the descriptive and inferential analysis of the results obtained during the study on improving cost estimation in civil works in the province of Trujillo, through a web application based on Random Forest algorithms.

6.1. Descriptive Analysis

The post-test values for the control group (CG) and the experimental group (EG), in relation to the indicators of percentage error (PE), time (T), and accuracy (A), are shown below.

Table 2 shows that the control group achieved an average percentage error of 11.61%. However, the experimental group reached a mean of 6.18%, reflecting a decrease of 5.43 percentage points compared to the control group.

Table 2. Post-test results for PE in the control and experimental groups

Indicator	Average		Standard Deviation
PE	CG	11.61	1.762
	EG	6.18	0.203

As shown in Table 3, the control group recorded an average time of 521.67 seconds to prepare structural quotations, whereas the experimental group reduced this to 198.60 seconds, representing a significant decrease of 323.07 second.

Table 3. Post-test results for T in the control and experimental groups

Indicator	Average		Standard Deviation
T	CG	521.67	110.0
	EG	198.60	37.0

Table 4 indicates that the control group achieved an average accuracy of 89.6%, while the experimental group reached a mean of 94.2%, representing an increase of 4.6 percentage points compared to the control group.

Table 4. Post-test results for P in the control and experimental groups

Indicator	Average		Standard Deviation
P	CG	89.6	1.464
	EG	94.2	0.180

6.2. Inferential Analysis

Normality and hypothesis tests were applied to both the control group (CG) and the experimental group (EG) for the three evaluated indicators: percentage error (PE), preparation time (T), and accuracy (A).

The decision criteria for the normality test were as follows: Ho: Post-test data follow a normal distribution Ha: Post-test data do not follow a normal distribution If $p < 0.05$, the null hypothesis (Ho) is rejected in favor of the alternative (Ha); if $p \geq 0.05$, Ho is accepted.

Given that the sample size per group was fewer than 50 records, the Shapiro–Wilk test was used to assess normality.

According to the results in Table 5, the p-values are below 0.05 for both groups across all three indicators. Therefore, it is concluded that the data do not follow a normal distribution. As a result, the non-parametric Mann–Whitney U test was applied to analyze the differences between the independent groups

Table 5. Shapiro–Wilk Normality Test for each indicator

No	Indicator	p (CG)	p (EG)
1	EP	<0.001	0.038
2	T	0.155	<0.001
3	P	<0.001	0.035

In regards to the data presented in Table 6, the p-value is less than 0.05 in all three cases, allowing for the rejection of the null hypothesis (H_o) and the acceptance of the alternative hypothesis (H_a) for each indicator. This demonstrates that the web application based on Random Forest improves performance in terms of error reduction, processing time, and accuracy.

Table 6. Mann–Whitney U Test for each indicator

Nº	Indicator	Ho	Ha	p
1	EP	$\mu1 <= \mu2$	$\mu1 > \mu2$	<0.001
2	T	$\mu1 <= \mu2$	$\mu1 > \mu2$	<0.001
3	P	$\mu1 >= \mu2$	$\mu1 < \mu2$	<0.001

7. Discussion

This section discusses how the Random Forest–powered web application improves structural cost-estimation performance. It contrasts the experimental and control groups across three key metrics percentage error, preparation time, and quotation accuracy to highlight the model's practical gains in the Peruvian construction context.

Regarding the first evaluated indicator, percentage error (PE), it was observed that the experimental group achieved an average of 6.18%, while the control group recorded 11.61%, representing a reduction of 5.43 percentage points. These results demonstrate an improvement in the cost estimation error margin when using a web application based on Random Forest, reducing the previously reported 15% error rate [3]. The proposed model, trained with real data from the Peruvian context, offers greater accuracy and adaptability.

For the second indicator, corresponding to the time required to prepare the structural quotation (T), a clear difference was identified between the two groups. The experimental group completed the quotation in an average of 198.60 seconds, while the control group required 521.67 seconds. This implies a 61.92% reduction in the time needed for the task. This improvement surpasses what was reported in a previous study [19], which documented a 55% reduction using a digital platform. However, the system implemented in this research automates the entire process from data entry to final calculation allowing for greater efficiency without manual intervention.

To address the final indicator, quotation accuracy (A), the experimental group achieved a mean value of 94.2%, exceeding the control group's average of 89.6%. This result also surpasses a previous study [20] that reported 85% accuracy using Random Forest models in the financial sector. Despite the difference in domain, the shared methodological approach allows for a valid comparison. The advantage here is attributed to the fact that the model was trained with authentic data from the construction sector, which enhances its predictive capacity within this context.

8. Conclusion

This section presents the main conclusions derived from the experimental results, highlighting the improvements achieved through the implementation of the Random Forest–based web application. It also outlines recommendations for future research to enhance the model's applicability and scalability in construction cost estimation.

After obtaining an average percentage error of 11.61% in the control group (CG) and 6.18% in the experimental group (EG), it is concluded that there was a reduction of 5.43 percentage points in error, demonstrating an improvement in the accuracy of cost estimations.

Since the control group (CG) recorded an average of 521.67 seconds for cost estimation and the experimental group (EG) achieved an average of 198.60 seconds in the post-test, this demonstrates a reduction of 323.07 seconds in average processing time when using the web application based on Random Forest.

Given that an average accuracy of 89.6% was recorded in the control group (CG), compared to 94.2% in the experimental group (EG), it is concluded that there was an increase of 4.6 percentage points in estimation accuracy with the implementation of the proposed application.

It is concluded that the use of a web application based on Random Forest improves the cost estimation process for construction materials in the city of Trujillo.

Based on the results, it is recommended that future research expand the database used to train the predictive model by incorporating unit prices from various regions and construction contexts. This would increase the model's generalization capacity and adaptability to different geographic areas and types of works.

Additionally, it is suggested that future studies integrate field validation modules to compare system-generated predictions with actual unit prices through real project executions. This would allow the model to be refined through continuous adjustment processes based on practical evidence.

Finally, future research could focus on developing a more robust and scalable platform by integrating various modules, such as scheduling, performance analysis, and scenario simulation. This approach would strengthen the tool as a comprehensive cost management system for civil works, enhancing its value as a decision-support solution.

References:

[1]. Szóstak, M., Stachoń, T., & Konior, J. (2024). Course of Cumulative Cost Curve (CCCC) as a Method of CAPEX Prediction in Selected Construction Projects. *Applied Sciences*, *14*(13), 5597. Doi: 10.3390/app14135597

[2]. Shafiqul, I. M., & Bhuiyan, T. H. (2020). Assessment of costs of nuclear power in Bangladesh. *Nuclear Energy and Technology*, *6*(3), 181-194. Doi: 10.3897/nucet.6.54003

[3]. Castañeda Muñoz, S., & Liñan Ponce, L. (2015). *Automatización de procesos e ingeniería de costos.* [Undergraduate thesis, Universidad Privada Antenor Orrego].

[4]. Papadimitriou, V. E., & Aretoulis, G. N. (2024). A final cost estimating model for building renovation projects. *Buildings*, *14*(4), 1072. Doi:10.3390/buildings14041072

[5]. Lederer, L., et al. (2024). Construction price forecasting models in the construction industry: A comparative analysis. *Buildings*, *14*(5), 1325. Doi:10.3390/buildings14051325

[6]. Rahman, M. M., et al. (2024). AirNet: predictive machine learning model for air quality forecasting using web interface. *Environmental Systems Research*, *13*(1), 44. Doi:10.1186/s40068-024-00378-z

[7]. Wang, C., & Qiao, J. (2024). Construction project cost prediction method based on improved BiLSTM. *Applied Sciences*, *14*(3), 978. Doi: 10.3390/app14030978

[8]. Pishdad, P., & Onungwa, I. O. (2024). Analysis of 5d bim for cost estimation, cost control, and payments. *Journal of Information Technology in Construction*, *29*. Doi: 10.36680/j.itcon.2024.024

[9]. Abd, A. M., Kareem, Y. A., & Zehawi, R. N. (2024). Prediction and estimation of highway construction cost using machine learning. *Engineering, Technology & Applied Science Research*, *14*(5), 17222-17231. Doi: 10.48084/etasr.8285.

[10]. Ojeda, J. M. P., et al. (2024). Estimation of the physical progress of work using uav and bim in construction projects. *Civil Engineering Journal*, *10*(02), 362-383. Doi: 10.28991/CEJ-2024-010-02-02

[11]. Ali, R. A. A., et al. (2024). Economic and environmental feasibility of a proposed bypass roadway in Falluja City: a case study. *International Journal of Sustainable Development and Planning*, *19*(5), 1627-1637. Doi: 10.18280/ijsdp.190501

[12]. Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Rand McNally & Co.

[13]. Uaisova, M., et al. (2024). The Use of ANN and Machine Learning Algorithms to Predict Road Surface Deterioration. *GEOMATE Journal*, *27*(121), 136-143. Doi: 10.21660/2024.121.m2422.

[14]. Konsa, K., Treimann, M. L., & Piirisild, K. (2021). Algoritmid museaale uurimas: masinõppemudeli Sälli kasutamine objektide säilivuse prognoosimiseks. *Ajalooline Ajakiri. The Estonian Historical Journal*, *177*(3/4), 253-278. Doi: 10.12697/AA.2021.3-4.05

[15]. Sha, X., et al. (2024). Construction Cost Estimation in Data-Poor Areas Using Grasshopper Optimization Algorithm-Guided Multi-Layer Perceptron and Transfer Learning. *International Journal of Advanced Computer Science & Applications*, *15*(7). Doi: 10.14569/IJACSA.2024.01507129.

[16]. Yu, W., Xia, L., & Cao, Q. (2024). A machine learning algorithm to explore the drivers of carbon emissions in Chinese cities. *Scientific Reports*, *14*(1), 23609. Doi: 10.1038/s41598-024-75753-y

[17]. Draleti, G., Sengonzi, R., & Kakitahi, J. (2024). Improvement of risk management in cost estimation in the building construction industry in Uganda. *Journal of Construction in Developing Countries*, *29*(1), 111-138. Doi: 10.21315/jcdc-08-22-0145

[18]. Satu, M. S., et al. (2021). Short-term prediction of COVID-19 cases using machine learning models. *Applied Sciences*, *11*(9), 4266. Doi: 10.3390/app11094266

[19]. Atoche Salas, C. J. (2021). *Sistema web utilizando machine learning para mejorar la planificación de requerimientos de materiales de la producción en el sector industrial.* [Undergraduate thesis, Universidad César Vallejo].

[20]. Miranda Ziña, R. (2023). *Sistema de pronóstico usando técnicas de machine learning para la detección de fallas en ATM de los clientes de Belltech*. [Undergraduate thesis, Universidad César Vallejo].