***ISSN** (E): 2959-3921*
(P): 2959-393X
***DOI:** https://doi.org/10.59122/EIJET1010*
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

Construction Schedule Prediction for Condominium Projects in Addis Ababa: A Comparative Analysis of Random Forest and Alternative Approaches

Kassahun Jima Kaba[1], Mohammadzen Hasan Darsa*[1], Moh Nur Sholeh[2]

[1] Construction Technology and Management Department, School of Civil Engineering and Architecture, Dire Dawa University, Dire Dawa, Ethiopia

[2]Civil Infrastructure Engineering and Architectural Design, Vocational School, Diponegoro University, Indonesia

*Corresponding Author's Email: mohammadzenhasan@yahoo.com

Abstract

Accurate construction duration estimation is critical for scheduling and resource allocation, yet traditional heuristic techniques often result in severe schedule overruns. This vulnerability is highly evident in Addis Ababa's public housing sector (20/80 and 40/60 condominium schemes), where over 38,000 units face chronic multi-fold delays due to data deficiencies and rigid planning tools. To establish a robust context-specific duration prediction model, this study utilizes historical data spanning 2013 to 2023 from 595 building blocks compiled by the Addis Ababa Housing Development and Administration Bureau (AAHDB) to develop an interpretable, Python-based machine learning model. A hybrid feature selection pipeline consisting of embedded feature selection and correlation matrix analysis optimized twelve initial variables into a highly relevant seven-predictor framework, integrating key physical building attributes with operational and seasonal metrics. Four machine learning algorithms were benchmarked to evaluate predictive performance: Random Forest (RF), Support Vector Machine (SVM), Linear Regression (LR), and an Artificial Neural Network (ANN). The empirical results established a definitive performance hierarchy: Random Forest > Support Vector Machine > Multiple Linear Regression > Artificial Neural Network. The primary Random Forest model achieved exceptional predictive dependability, yielding a coefficient of determination R^2 of 0.999, paired with minimal error margins (RMSE = 0.081 months and MAE = 0.015 months). The non-linear SVM baseline served as the closest competitor, capturing complex multidimensional patterns with an R^2 of 0.976 (RMSE = 1.09 months), whereas traditional linear regression (R^2 = 0.822) and the data-constrained ANN (R^2 = 0.665) underperformed. Given the high predictive accuracy of the ensemble tree and

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Mohammadzen Hasan***

***ISSN** (E): 2959-3921*
(P): 2959-393X
***DOI:** https://doi.org/10.59122/EIJET1010*
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

kernel models, independent k-fold cross-validation is recommended before implementation. Ultimately, this research resolves historical 'black-box' rasping by balancing model interpretability with data-driven precision, offering public planners and construction practitioners an effective management tool to mitigate local housing project delays.

Keywords: Construction Schedule, Random Forest, Prediction Model, Condominium Projects, Addis Ababa

I. INTRODUCTION

Accurate construction duration estimation is critical for engineering decisions, financial budgeting, and resource allocation. Unfortunately, traditional estimation techniques often suffer from data deficiencies, poor productivity tracking, and estimator inexperience [1] [2] [3] [4]. This vulnerability is highly evident in Addis Ababa's public housing sector such as the 20/80 and 40/60 condominium schemes at Bole Arabsa and other sites where under-scheduling and poor project planning have caused over 38,000 units to be delayed by more than two- to five-fold past their timelines, delaying affordable housing initiatives [5] [6]. While machine learning algorithms like multiple linear regression, support vector machines, and neural networks have been applied globally to optimize schedule predictions, their performance varies, and the highly effective Random Forest (RF) algorithm remains significantly underutilized in construction duration modeling [7] [8] [9]. Furthermore, existing literature is constrained by inconsistent feature selection, which hinders model reproducibility and comparative analysis. To address these geographic, methodological, and typological gaps, this study utilized historical data to develop an interpretable budgeted duration prediction model specifically for condominium residential buildings in Addis Ababa using Python. By employing embedded and correlation coefficient matrix feature selection methods via the sklearn library, the research integrates established variables alongside previously overlooked factors such as the number of building functional units and benchmarks the primary Random Forest model against alternative machine learning algorithms to establish a robust context-specific framework for mitigating local project delays.

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Mohammadzen Hasan***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: https://doi.org/10.59122/EIJET1010

Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

II. LITERATURE REVIEW

Accurately forecasting construction durations is vital for resource allocation, budgeting, and risk mitigation, driving a shift from single-parameter heuristics to machine learning (ML) models that capture non-linear timeline relationships. Previous research emphasizes regional and typological adaptations for example, studies [9] focused on Middle Eastern frameworks, [1] designed a WBS-based model for public buildings in Addis Ababa. Studies in South Korea [10] [11] applied multiple linear regression (MLR) to mixed-use and office projects. In the residential sector, [3] showed that disaggregate physical variables (e.g., floor count combined with single-floor area) yield higher accuracy than aggregate metrics like gross floor area. Predictors generally split into physical features evaluated via MLRA, Support Vector Machines (SVM), or K-Nearest Neighbors (KNN) and operational factors analyzed through Artificial Neural Networks (ANN), decision trees, and ensembles, with performance measured via MAE, RMSE, R^2, and MAPE. Methodologically, [12] found that ANNs outperform traditional estimation, while [13] demonstrated that an ensemble method with an ANN combiner achieved the highest predictive accuracy (R^2 = 0.69, MAPE = 18%). Despite these advancements, three critical research gaps persist: geographic underrepresentation, as advanced ML models are concentrated in high-data environments while Sub-Saharan African studies face small sample sizes and public-sector limitations a typological mismatch, where existing literature overlooks mass-housing condominiums characterized by repetitive cycles and unique financing and "black-box" friction, where accurate models like ANNs lack interpretability for project managers, while interpretable models like MLRA fail to capture non-linear interactions. To address these gaps for condominium projects in Addis Ababa, this study utilizes Random Forest (RF) as its primary architecture. RF natively handles mixed numeric and categorical data, resists over fitting on small local datasets via bagging, and resolves the black-box dilemma by providing explicit feature importance rankings. By benchmarking RF against MLRA for a linear baseline and Support Vector Regression (SVR) for non-linear optimization, this study establishes an interpretable, region-specific, and typologically accurate framework tailored to Addis Ababa's residential sector.

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Mohammadzen Hasan***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: https://doi.org/10.59122/EIJET1010
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

TABLE I

SUMMARY OF PREVIOUS STUDIES ON MODEL DEVELOPMENT TO ESTIMATE CONSTRUCTION DURATION

S/N	Authors	Country	Input parameters	ML Algorithm
1	[13]	Saudi Arabia	Numeric parameters such as project cost, number of elevators, building area, floor area, height to the tip, number of floors above GF, height of occupied floors, number of total floors, number of basement floors, and number of parking spaces and non-numeric parameters such as facility type, structural form, structural material, and commencement period	Multi-Linear Regression Analysis (MLRA), k-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Support Vector Machines (SVM)
2	[10]	Korea	Quantitative features such as gross floor area, building area, number of stories above ground, and number of stories below ground & Categorical features such as type of primary use, construction region, and structure type	Multiple Linear Regression
3	[3]	Slovakia	gross floor area, number of stories, floor area of one story, and man-power deployment intensity	Multiple linear regression analysis
4	[10]	Korea	Construction method, structural system, number of floors, gross floor area, and project location.	Multiple regression analysis
5	[6]	Ethiopia	Work breakdown structure (WBS) framework (excavation, foundation, superstructure, and finishing works).	Regression analysis and statistical tests
6	[9]	Middle East	Project size, location, complexity, and contractor experience.	Statistical Regression

III. METHODS

A. Research Approach and Data Collection

The historical data of completed and in-progress condominium projects was collected from the Addis Ababa Housing Development Authority (AAHDO) office and project sites for the development of a predictive model regarding the duration of condominium construction projects. Key quantitative parameters such as total building area, project cost, number of floors and number of building units are common terms listed in literature to predict the duration of construction projects [14] [10] [15] [16] [17].

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Mohammadzen Hasan***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: https://doi.org/10.59122/EIJET1010

Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

The data collected includes the building block number, name of the project site, typology, commencement season, gross floor area of the building, number of floors above ground, number of basement floors, number of functional units in the building, budgeted cost, actual project cost, planned project duration in months, and actual project duration in months. A systematic literature review and consultations with construction experts have demonstrated that the combination of quantitative and qualitative factors significantly improves the prediction accuracy of the construction project duration of condominium construction [10] [18] [13] [2].

The corrtions between variables such as commencement season of the project, gross floor area of the building, number of floors above ground, number of basement floors, number of functional units in the building, budgeted cost, actual project cost, planned project duration in months, and actual project duration were considered while developing the model. In particular, emphasis was placed on the analysis of aggregate data of 810 components of the two programs considering the progress and completion reports of the AAHDO 2013-2023 project.

To develop the machine learning-based duration prediction model, a comprehensive data set was compiled from two distinct condominium housing initiatives: the 20/80 and 40/60 housing programs, as shown in Tables I and II.

From the 20/80 housing program, a total of 505 buildings were evaluated. This subset comprised 411 ground-plus-seven (G+7) structures and 94 ground-plus-four (G+4) structures. Additionally, the dataset incorporated 90 buildings from the 40/60 housing program, which exhibited greater structural diversity. The architectural distribution of the 40/60 program included 11 basement-plus-ground-plus-seven (B+G+7) structures, five basement-plus-ground-plus-nine (B+G+9) structures, five double-basement-plus-ground-plus-twelve (2B+G+12) structures, 44 double-basement-plus-ground-plus-thirteen (2B+G+13) structures, 15 double-basement-plus-ground-plus-fifteen (2B+G+15) structures, and 10 double-basement-plus-ground-plus-eighteen (2B+G+18) structures.

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Mohammadzen Hasan***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: *https://doi.org/10.59122/EIJET1010*
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

TABLE II

NUMBER OF 20/80 HOUSING SCHEME BUILDINGS CONSIDERED FOR STUDY DATA COLLECTION

S/N	Project Branch Office	Site Name	No. of Floor	No. of Blocks Considered	Progress of the Projects	Budgeted duration (Month)	Actual duration (Month)	Remark
1	Lideta	Bole Arabsa 3	G+7	72	100%	24	72	All buildings faced delay
			G+4	8	100%	18	72	
2	Project 15	Bole Arabsa 2	G+7	141	100%	24	60	
			G+4	30	100%	18	60	
3	Bole	Bole Arabsa 5	G+7	56	100%	24	72	
4	Yeka	Bole Arabsa 3 & 5	G+7	64	100%	12	64	
			G+4	18	100%	9	64	
5	Project 15	Bole Arabsa 6	G+7	78	100%	12	96	
			G+4	38	100%	9	96	

TABLE III

NUMBER OF 40/60 HOUSING SCHEME BUILDINGS CONSIDERED FOR STUDY DATA COLLECTION

S/N	Project Branch Office	Site Name	No. of Floors	No. of Buildings	Project Progress	Planned duration (Month)	Actual duration (Month)	Remarks
1	Project 1	Bole Bulbula 1 & 2	B+G+7	11	100%	18	96	All buildings faced delay
			B+G+9`	5	100%	20	96	
			2B+G+13	23	100%	24	96	
			2B+G+15	13	100%	36	96	
2		Hinsta Aqrabi	2B+G+15	2	100%	36	96	
			2B+G+13	6	100%	24	96	
3		Asko	2B+G+13	13	100%	24	124	
			2B+G+13	2	100%	24	126	

Received*: April 22, 2026;* ***Revised****: June 9, 2026;* ***Accepted:*** *June 16, 2026;* ***Published:*** *June 23, 2026*
Corresponding author- ***Mohammadzen Hasan***

4	Ehilnigd	2B+G+12	4	100%	24	126
5	Tourist	2B+G+12	1	100%	24	96
		2B+G+18	10	100%	36	96

The proposed minimum sample size of 268 building blocks was the result of the Yemane mathematical model, which guarantees a robust analysis with a maximum standard error of 5% [19].

Combining different types of data for obtaining better predictions with quantitative factors like total area and project cost and qualitative parameters like digitally transformed project complexity and site location was performed to provide insights into factors influencing project schedules in urban housing developments by statistical and machine learning methods [13] [10] [2] [20].

B. Tools and Techniques Used

Random Forest algorithm natively handles mixed numeric and categorical data, resists over fitting on small local datasets via bagging, and resolves the black-box dilemma by providing explicit feature importance rankings [21] [22]. Since RF has a better predictive performance than other AI algorithms, the study sought to develop a construction duration prediction model for condominium housing in Addis Ababa. Additionally, support vector machines (SVM), multiple linear regression, and artificial neural networks (ANN) were among the algorithms taken into consideration in the study [23] [24] [25] [12] for comparisons of better performers.

C. Performance Evaluation of the Prediction Models

The final step in developing a prediction model is to test how well the prediction algorithm performs. This involves using different methods to see how accurately it predicts outcomes. Unlike the measures used for classification, which focus on sorting data into categories, regression measures are focused on understanding the size of prediction errors. In this study, several indicators were used to evaluate how effective the regression models are.

Mean Absolute Error (MAE)

MAE measures the average absolute difference between the actual and predicted values. It is a straightforward metric that treats all differences as positive:

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Mohammadzen Hasan***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: https://doi.org/10.59122/EIJET1010

Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

$$\text{MAE} = \frac{1}{N}\sum|\text{P} - \text{P}'| \qquad (1)$$

Mean Squared Error (MSE)

MSE calculates the average of the squared differences between the actual and predicted values, providing a metric that penalizes larger errors more heavily:

$$\text{MSE} = \frac{1}{N}\sum(P - P')^2 \qquad (2)$$

MSE is preferred for its differentiability, making it easier to optimize.

R Squared (R^2)

R^2, or the coefficient of determination, evaluates the proportion of variance in the dependent variable that is predictable from the independent variables:

$$\text{R}^2 = 1 - \frac{MSE\ model}{MSE_{baseline}} \qquad (3)$$

R^2 values range from 0 to 1, with higher values indicating a better fit.

Root Mean Square Error (RMSE)

RMSE is the square root of MSE, providing a metric in the same unit as the original data:

$$RMSE = \sqrt{\frac{\sum_{j=1}^{j=N}\left(P_j - P_j'\right)^2}{N}} \qquad (4)$$

RMSE, like MSE, gives more weight to larger errors.

Where P is the budgeted project duration, P' is the predicted budgeted duration, and N is the total number of datasets in the case of this study.

D. Data Presentation, Discussion & Analysis

1) *Database Establishment***:** Projects completed between 2013 and 2023 are included in the data set, with information including architectural plans, drawings, design and contract documents, and completion reports. Data was gathered for 505 buildings in the 20/80 Housing Program,

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Mohammadzen Hasan***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: *https://doi.org/10.59122/EIJET1010*
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

including 94 R+4 and 411 R+7 structures. Moreover, 90 buildings in the 40/60 Housing Program, with structures ranging from B+R+7 to 2B+R+18, were collected. Building block number, project site, typology, commencement season, gross floor area, number of floors, number of functional units, budgeted and actual project costs, and estimated and actual project durations were among the details included in the dataset.

2) *Data Preprocessing*: This was necessary to ensure the accuracy and effectiveness of the model. The dataset obtained from the real world may exhibit characteristics that are not ideal for ML modeling and may require preprocessing and cleaning. Consequently, data preprocessing played an important role in optimizing the dataset for the Random Forest and other algorithms.

3) *Handling Missing and Duplicate Values*: A check for missing values (Fig.1) was conducted, and it was confirmed that there were no missing entries in any column. This indicates that the dataset is complete, without any gaps or inconsistencies. Verifying missing values is a crucial step to ensure the accuracy and reliability of any analysis or modeling. Furthermore, an assessment for duplicate rows was performed, confirming that there were no duplicates in the dataset. With no missing values or duplicate records, the dataset is clean and ready for further data analysis and modeling with confidence.

```
In [9]:  #Data Preprocessing
         # checking for missing values
         dur_pred.isnull().sum()

Out[9]:  Block No.          0
         BPS                0
         PCS                0
         CBT                0
         GFA (M2)           0
         NoFAG              0
         NoBFU              0
         BH (M)             0
         NoBF               0
         BPD (Month)        0
         APD (Months)       0
         BPC (ETB)          0
         APA (ETB)          0
         dtype: int64

In [10]: # Check for duplicates based on all columns
         duplicates = dur_pred[dur_pred.duplicated()]

         dur_pred.duplicated().sum()

Out[10]: 0
```

Fig. 1. Checking for missing and duplicate values in the dataset

Received*: April 22, 2026;* ***Revised****: June 9, 2026;* ***Accepted:*** *June 16, 2026;* ***Published:*** *June 23, 2026*
Corresponding author- ***Mohammadzen Hasan***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: https://doi.org/10.59122/EIJET1010
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

4) *Changing Categorical Features to Numerical*: The preparation of data for analytical purposes involving transforming categorical features into numerical values is also implemented. The process transforms the categorical data into binary representations for each unique category. This dataset's categorical variables encompassed the Building Project Site, Project Commencement season, and condominium building typology (Fig. 2)

	BPS	PCS	CBT	GFA (M2)	NoFAG	NoBFU	BH (M)	NoBF	BPD (Month)	\
17	7	2	1	12819.10	13	615	45	2	24	
18	7	2	1	12819.10	13	615	45	2	24	
19	7	2	1	12819.10	13	615	45	2	24	
20	7	2	5	12819.10	7	420	29	1	18	
21	7	2	5	6017.80	7	420	29	1	18	
22	7	2	5	6017.80	7	420	29	1	18	
23	7	2	5	6017.80	7	420	29	1	18	
24	7	2	6	6017.80	9	437	36	1	20	
25	7	2	6	6017.80	9	437	36	1	20	
26	7	2	5	6017.80	7	420	29	1	18	
27	7	2	5	6017.80	7	420	29	1	18	
28	7	2	5	6017.80	7	420	29	1	18	
29	7	2	5	5426.25	7	420	29	1	18	
..	...	...	...	...	...	...	...	...	...	
565	6	0	37	5193.37	7	321	29	0	24	
566	6	0	38	5227.60	7	324	29	0	24	
567	6	0	80	5474.88	7	317	29	0	24	
568	6	0	80	5474.88	7	317	29	0	24	
569	6	0	38	5227.60	7	324	29	0	24	

Fig. 2. Categorical features changed to numerical values

A defined collection of categorical features undergoes an iterative encoding process for each feature. The transformation process converts categorical entries into numerical values, rendering them fit for subsequent analysis. Examination of the initial dataset entries confirmed categorical variable transformation into numerical format through successful conversion verification. The data preprocessing phase served as a critical component in enhancing predictive modelling accuracy and effectiveness for estimating the construction schedule of condominium projects.

5) *Scaling, Normalizing, and Standardizing:* To prevent features with larger magnitudes or outliers and differing units from dominating the machine learning model, feature scaling was performed. Numerical features specifically BPS, PCS, CBT, GFA (M2), NoFAG, NoBFU, BH (M), NoBF, BPD (Month), APD (Months), BPC (ETB), and APA (ETB) were standardized to a uniform scale. This normalization was executed using Scikit-Learn's StandardScaler () via the fit_transform () method (Fig. 3). The resulting standardized dataframe, scaled_features_dur_pred_dropped, ensures a balanced feature contribution and was utilized as the input for subsequent predictive modeling.

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Mohammadzen Hasan***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: *https://doi.org/10.59122/EIJET1010*
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

	BPS	PCS	CBT	GFA (M2)	NoFAG	NoBFU	BH (M)	\
0	1.376673	0.535784	-1.432179	1.894784	1.999390	1.630495	1.907597	
1	1.376673	0.535784	-1.432179	1.894784	1.999390	1.630495	1.907597	
2	1.376673	0.535784	-1.363939	3.264461	2.715392	3.307505	2.672178	
3	1.376673	0.535784	-1.261579	0.062797	0.567386	0.564395	0.760726	
4	1.376673	0.535784	-1.432179	1.894784	1.999390	1.630495	1.907597	

	NoBF	BPD (Month)	APD (Months)	BPC (ETB)	APA (ETB)
0	2.257520	0.629226	1.150484	2.301814	2.343061
1	2.257520	0.629226	1.150484	2.301814	2.343061
2	2.257520	2.391748	1.150484	2.821233	2.862104
3	0.831592	0.041718	1.150484	0.926998	0.969239
4	2.257520	0.629226	1.150484	2.078436	2.119844

Fig. 3. Feature scaling using Scikit-Learn's StandardScaler () via the fit_transform() method

6) Selection of Potential Predictor Variables: Developing a robust and computationally efficient predictive model for project duration requires the identification of highly relevant predictor variables [26]. While twelve potential features influencing project duration were initially identified through literature review and consultations with building site professionals, not all features contribute significantly to predictive accuracy. Identifying key features is essential to optimize model efficiency and minimize complexity. In machine learning applications, feature selection is typically achieved using embedded feature selection (EFS) or correlation analysis [27]. In this study, an Embedded Feature Selection (EFS) approach was implemented using Random Forest regression to assess feature importance via Mean Decrease in Impurity (MDI) and Mean Decrease in Accuracy (MDA). The initial twelve features evaluated included: Gross Floor Area (GFA), Number of Floors Above Ground (NoFAG), Number of Building Functional Units (NoBFU), Building Height (BH), Number of Basement Floors (NoBF), Budgeted Project Cost (BPC), Actual Project Cost (APC), Budgeted Project Duration (BPD), Actual Project Duration (APD), Project Commencement Season (PCS), Building Project Site (BPS), and Condominium Building Typology (CBT).

a) Embedded feature selection (MDI and MDA rankings): The MDI and MDA analysis yielded the following results regarding feature significance:

MDI Ranking: Features were organized in descending order based on their structural contribution to the decision trees. Budgeted Project Duration (BPD) demonstrated the highest importance score (0.906), followed by Building Height (0.028) and Gross Floor Area (0.027).

Received*: April 22, 2026;* ***Revised****: June 9, 2026;* ***Accepted:*** *June 16, 2026;* ***Published:*** *June 23, 2026*
Corresponding author- ***Mohammadzen Hasan***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: https://doi.org/10.59122/EIJET1010
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

Although both methods identified Budgeted Project Duration, Gross Floor Area, and Building Height as the dominant variables, the consulted condominium project experts recommended incorporating additional site-specific predictor variables to ensure a reliable and practically viable model.

b) Correlation matrix analysis and data leakage mitigation: To identify further potential predictors, a correlation coefficient matrix was computed (Fig. 4) to quantify the linear relationships between variables, ranging from -1 to 1. The matrix was visualized using a Seaborn-generated heatmap [27]. Based on the statistical strength of these associations, seven variables strongly correlated with the target duration metrics were identified: building height, number of floors above ground, gross floor area, and number of building functional units, budgeted project cost, project commencement season recommended by site professionals, and number of basement floors. This subset successfully captured the key parameters identified during the prior EFS phase.

By synthesizing the empirical findings of both selection methods with practical insights from industry literature, a final subset of seven predictors was retained from the initial twelve variables [28].

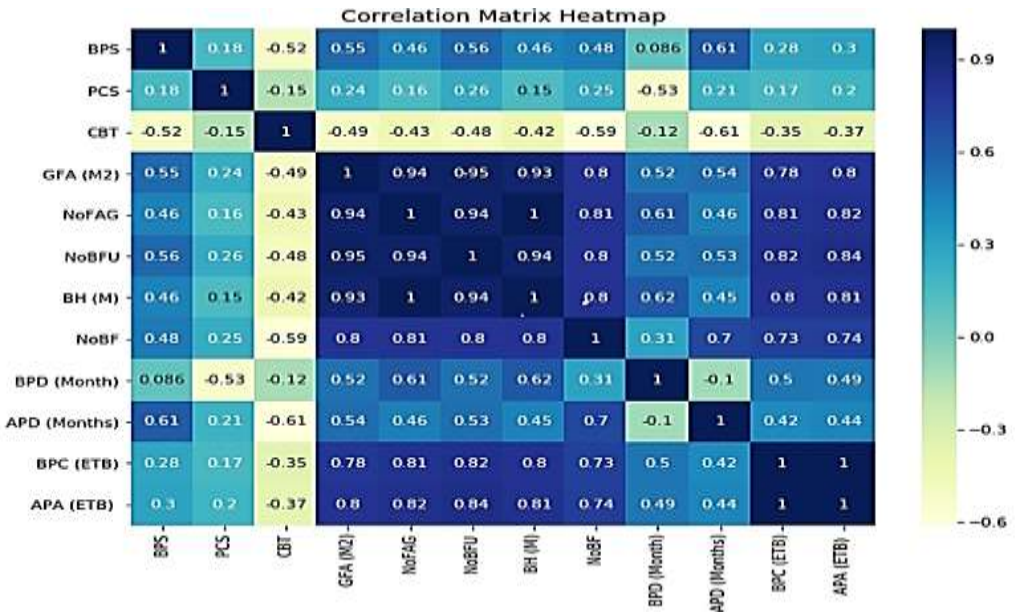

Fig. 4. Correlation matrix heatmap

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Mohammadzen Hasan***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: https://doi.org/10.59122/EIJET1010
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

IV. METHODOLOGY AND MODEL DEVELOPMENT

A. Overview of Predictive Modeling

Mathematical models offer practical, fast, and simple solutions for predicting construction duration even when working with limited project information [29]. In this study, influencing factors and budgeted construction duration data were partitioned into training and testing subsets to develop predictive models using four machine learning algorithms: Random Forest (RF), Support Vector Machine (SVM), Multiple Linear Regression (MLR), and Artificial Neural Network (ANN).

To ensure robust model generalization and prevent overfitting, data must not be trained and evaluated on the same dataset [24]. Consequently, the dataset in this study was randomly partitioned into an 80% training set for model development and a 20% testing set for independent performance evaluation. This random distribution ensures an unbiased representation across both subsets, preventing localized patterns from skewing evaluation metrics. The partitioning was implemented using the train_test_split function from the sklearn.model_selection module.

B. Primary Model: Random Forest (RF)

The primary predictive model was developed using the RandomForestRegressor class from the scikit-learn library, initialized with n_estimators=100 and random_state=42 for reproducibility. The model was trained by executing the fit() method on the labeled training subsets (X_train and y_train). During the prediction phase, the ensemble framework aggregates the individual continuous outputs from all 100 decision trees and computes their average to determine the final predicted building project duration.

C. Comparative Baseline Models

1) Support Vector Machine (SVM): The data was loaded via pandas and split using train_test_split(). An instance of the SVR class was utilized for regression by configuring the kernel parameter (e.g., linear or radial basis function (RBF)) along with tuning the regularization (C) and kernel coefficient (***Y***). Following model fitting on the training partition, the predict() method generates duration estimates on the test set to calculate standard evaluation metrics (sklearn.metrics).

2) Multiple Linear Regression (MLR): MLR is a statistical technique utilized to evaluate how a dependent variable changes in response to simultaneous shifts in multiple independent variables

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Mohammadzen Hasan***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: https://doi.org/10.59122/EIJET1010
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

[30]. The model was initialized using the LinearRegression class from scikit-learn and fitted to the training data. Post-training, predictions are executed on the test partition using the predict() method and stored as y_pred for performance benchmarking.

3) Artificial Neural Network (ANN): A deep learning architecture was constructed using the Keras Sequential API. The network consists of three layers: an initial dense layer (64 units, ReLU activation, input dimension matched to feature shape), a second dense layer (64 units, ReLU activation), and a final output dense layer (1-unit, linear activation) suitable for continuous regression tasks.

The findings, shown in Table III, offer valuable insights into predicting completion times, aiding in better decision-making for project planning and management.

TABLE III

ESTABLISHED HISTORICAL CASES WITH PREDICTED OUTPUT

No. Blocks	**Input** PCS	GFA	NoFAG	NoBFU	BH	NoBF	BPC	**`Output** BPD	PBPD
0	-0.599	-0.735	-1.223	-0.921	-1.278	-0.594	-0.567	18	20
1	-0.599	-0.733	-0.149	-0.041	-0.131	-0.594	1.784	24	24
2	0.536	-0.214	-0.149	-0.136	-0.131	-0.594	-0.240	12	12
3	0.536	-0.146	-0.149	-0.112	-0.131	0.832	-0.240	12	12
4	-0.599	-0.735	-1.223	-0.795	-1.278	-0.594	-0.229	18	18
5	1.670	-0.748	-1.223	-0.915	-1.278	-0.594	-0.594	9	9
6	-0.599	-0.710	-0.149	-0.795	-0.131	-0.594	-0.261	24	23
7	1.670	-0.236	-0.149	-0.041	-0.131	-0.594	-0.466	12	12
8	-0.599	-0.156	-0.149	-0.232	-0.131	-0.594	-0.824	24	24
9	-1.733	-0.159	-0.149	-0.130	-0.131	-0.594	-0.593	24	24
10	1.670	-0.153	-0.149	-0.124	-0.131	-0.594	-0.466	12	12
11	1.670	-0.153	-0.149	-0.124	-0.131	-0.594	-0.594	24	24
12	0.536	3.264	2.715	3.308	2.672	2.258	2.821	36	36
13	-0.599	1.895	1.999	1.630	1.908	2.258	1.562	24	24
…					…				
585	0.536	-0.740	-1.223	-0.801	-1.278	-0.594	-0.605	9	9
586	-0.599	-0.692	-1.223	-0.801	-1.278	-0.594	-0.300	18	18

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Mohammadzen Hasan***

587	0.536	3.264	2.715	3.308	2.672	2.258	2.821	9	10
588	-0.599	-0.203	-0.149	-0.795	-0.131	-0.594	-0.182	24	24
589	-0.599	-0.731	-1.223	-0.921	-1.278	-0.594	-0.293	18	18
590	0.536	-0.731	-1.223	-0.801	-1.278	-0.594	-0.605	9	9
591	-0.599	-0.141	-0.149	-0.298	-0.131	-0.594	-0.824	24	24
592	-1.733	-0.083	-0.149	-0.154	-0.131	-0.594	-0.593	24	24
593	1.670	-0.153	-0.149	-0.041	-0.131	-0.594	-0.466	12	12
594	1.670	4.116	3.789	4.254	3.819	2.258	5.357	36	36

V. RESULTS AND DISCUSSION OF EXPERIMENT

This study proposed and evaluated an ensemble Random Forest (RF) regression model as the primary predictive framework for forecasting Budgeted Project Duration (BPD) in months. Model performance was assessed using three standard metrics: the coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

The experimental results demonstrated that the RF model achieved exceptional predictive performance. The model yielded an (R^2) value of 0.999, indicating that it successfully accounted for nearly all variation in building project durations. This strong explanatory power was supported by minimal prediction errors, producing an RMSE of 0.081 months and an MAE of 0.015 months. The superior performance of the primary Random Forest (RF) model is driven by its ensemble mechanism, which aggregates predictions across multiple decision trees to minimize variance and maximize generalizability. This structure offers immense nonlinear flexibility, mapping complex multi-dimensional construction interactions without requiring the rigid data distributions that constrain traditional statistical models. Consequently, the RF model effectively navigated the intricate data to achieve high predictive fidelity, aligning almost perfectly with actual historical project timelines as indicated in Fig. 5.

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Mohammadzen Hasan***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: https://doi.org/10.59122/EIJET1016
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

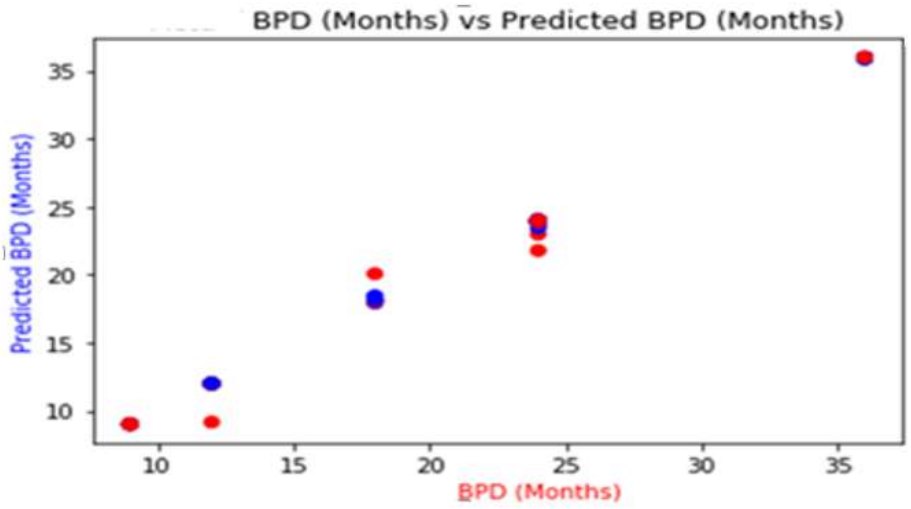

Fig. 5. Predicted BPD (months) vs BPD (months) scatter plot

A. Comparative Performance Analysis

To rigorously benchmark the proposed RF model, its performance was compared against three baseline machine learning techniques: Support Vector Machine (SVM), Linear Regression (LR), and Artificial Neural Network (ANN). The comparative metrics for all four models are summarized in Table IV.

The comparative analysis revealed distinct gradients in performance across the baseline models, highlighting the advantages of the primary RF approach:

1) Support Vector Machine (SVM): The SVM model serves as the closest competitor. Utilizing a radial basis function (RBF) kernel, it effectively captured nonlinear patterns to achieve an R^2 of 0.976, an RMSE of 1.09 months, and an MAE of 0.47 months. While highly accurate, its error margins remained significantly higher than those of the RF model.

2) Linear Regression (LR): The LR model delivered moderate predictive capacity, achieving an R^2 of 0.822, an RMSE of 3.00 months, and an MAE of 2.60 months. The model captured general data set trends but was fundamentally constrained by its strict assumption of linearity, which fails to accommodate the highly interactive nature of construction variables.

3) Artificial Neural Network (ANN): The ANN model produced the weakest comparative baseline. It generated an R^2 of only 0.665 alongside substantially higher error rates (RMSE = 3.89 months MAE = 3.36 months). This subpar performance likely highlights typical ANN constraints, such as

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Kassahun Jima***

ISSN (E): 2959-3921
(P): 2959-393X
DOI: *https://doi.org/10.59122/EIJET1016*
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

high sensitivity to insufficient training data, suboptimal layer architectures, inadequate hyperparameter tuning, or underfitting.

Ultimately, the proposed RF model significantly outperformed all three comparative baselines. The definitive performance ranking established by this study is **Random Forest** > **Support Vector Machine** > **Linear Regression** > **Artificial Neural Network**.

B. Validation and Industrial Implications

The exceptional, near-perfect accuracy (R2 = 0.999) achieved by the primary RF model warrants careful scientific scrutiny. While this indicates highly informative project predictors within the dataset, it also signals potential risks of localized over-fitting or data leakage. To ensure the absolute robustness and generalizability of this primary model, further validation techniques such as k-fold cross-validation and independent data set testing are highly recommended before deployment.

In conclusion, the proposed RF model provides the most reliable and precise predictions for the studied dataset. It represents a highly effective decision-support tool capable of optimizing project planning, scheduling, and risk mitigation across the construction industry.

TABLE IV

COMPARISON OF MODEL PERFORMANCE FOR PROJECT DURATION PREDICTION

Model category	Model	MAE	RMSE	R^2	Performance Summary
Primary Model	Random Forest (RF)	0.015	0.081	0.999	Near-perfect prediction with exceptional explanatory power and minimal error margins.
Baseline Models	Support Vector Machine (SVM)	0.74	1.09	0.976	Strong predictive accuracy successfully captured complex nonlinear interactions using an RBF kernel.
	Multiple Linear Regression (MLR)	2.60	3.00	0.822	Moderate capacity limited by strict linearity assumptions that fail with complex construction variables.
	Artificial Neural Network (ANN)	3.36	3.89	0.663	Weakest performance highly sensitive to small data set limits or suboptimal hyperparameter optimization

Received*: April 22, 2026;* ***Revised****: June 9, 2026;* ***Accepted:*** *June 16, 2026;* ***Published:*** *June 23, 2026*
Corresponding author- ***Kassahun Jima***

ISSN (E): 2959-3921
(P): 2959-393X
DOI: *https://doi.org/10.59122/EIJET1016*
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

VI. CONCLUSIONS AND RECOMMENDATIONS

A. Conclusions

This study developed and evaluated an interpretable data-driven construction duration prediction model tailored for public condominium housing schemes (20/80 and 40/60) in Addis Ababa, Ethiopia. Aimed at mitigating chronic delays affecting over 38,000 historical housing units, this research bridges critical geographic, typological, and methodological gaps in current construction management literature.

Using historical records of 595 building blocks from the Addis Ababa Housing Development Authority (AAHDO) spanning 2013–2023, the study established a robust data preprocessing and feature selection pipeline. A hybrid approach combining Embedded Feature Selection (MDI/MDA rankings), Correlation Matrix Analysis, and expert consultations optimized twelve initial variables into a highly relevant seven-predictor framework. This framework successfully integrated physical parameters Gross Floor Area (GFA), Building Height (BH), Number of Floors above Ground (NoFAG), Number of Basement Floors (NoBF), and Number of Building Functional Units (NoBFU) with key operational metrics, specifically Budgeted Project Cost (BPC) and Project Commencement Season (PCS).

Four machine learning algorithms were benchmarked to identify the most accurate and reliable framework for forecasting Budgeted Project Duration (BPD): Random Forest (RF), Support Vector Machine (SVM), Multiple Linear Regression (MLR), and an Artificial Neural Network (ANN). The empirical findings established a definitive performance hierarchy: Random Forest (RF) > Support Vector Machine (SVM) > Linear Regression (LR) > Artificial Neural Network (ANN)

The primary Random Forest model demonstrated exceptional predictive fidelity, achieving a coefficient of determination R^2 of 0.999, paired with remarkably low error margins (RMSE = 0.081 months and MAE = 0.015 months). This superior performance highlights the capacity of ensemble bagging mechanisms to map complex non-linear construction relationships without the constraints of rigid data distributions.

Utilizing a radial basis function (RBF) kernel, the SVM model proved to be the closest and most formidable competitor to the primary Random Forest architecture. It effectively captured the

Received*: April 22, 2026;* ***Revised****: June 9, 2026;* ***Accepted:*** *June 16, 2026;* ***Published:*** *June 23, 2026*
Corresponding author- ***Kassahun Jima***

ISSN *(E): 2959-3921*
(P): 2959-393X
DOI: https://doi.org/10.59122/EIJET1016
Ethiopian International Journal of Engineering and Technology (EIJET)
Volume 4, Issue 1, 2026

complex, multidimensional, and non-linear patterns inherent to construction project schedules, yielding a high coefficient of determination (R^2 = 0.0976) along with a Root Mean Squared Error (RMSE) of 1.09 months and a Mean Absolute Error (MAE) of 0.47 months.

Conversely, traditional Linear Regression was limited by its linear assumptions (R^2 = 0.822), while the ANN underperformed (R^2 = 0.665) due to data volume constraints typical of public sector environments in Sub-Saharan Africa. Ultimately, this study resolves the industry's historical "black-box" friction by delivering an approach that balances high predictive accuracy with interpretability, providing public planners with a robust framework to optimize project timelines and resource allocations.

B. Recommendations

Based on the empirical findings, validation parameters, and practical implications of this research, strategic recommendations are proposed across institutional, practical, and academic dimensions to optimize future public housing delivery. At the institutional level, the Addis Ababa Housing Development Authority (AAHDO) must actively transition from outdated single-parameter rule of thumb and subjective scheduling techniques toward data-driven and machine learning-integrated planning frameworks. The primary Random Forest model developed in this study or its highly competent non-linear alternative, the Support Vector Machine (SVM) model, should be formally embedded into the early-stage planning, procurement, and budgeting workflows of upcoming public housing schemes to establish scientifically grounded baseline durations. To sustain the predictive efficacy of these algorithms, AAHDO must establish a centralized digitally transformed project database. Systematically tracking and archiving granular physical building parameters (such as the number of functional units and basement profiles) alongside historical actual costs and durations will provide the continuous high-fidelity data streams required to dynamically update and retrain the predictive models as the local construction ecosystem evolves. From a practical construction management perspective, project managers, consultants, and estimators must re-engineer their localized scheduling protocols to explicitly account for non-linear structural and environmental complexities. Industry practitioners should move away from aggregate metrics like gross floor area alone and rigorously weigh the specific impacts of disaggregate physical features such as building height, number of basements, and localized

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Kassahun Jima***

functional units as well as operational variables like the Project Commencement Season, all of which demonstrated powerful statistical correlations with target project durations.

Furthermore, given the exceptional mathematical reliability achieved by the Random Forest (R^2 = 0.999) and SVM (R^2 = 0.976) models on this localized dataset, practitioners must exercise rigorous scientific caution prior to commercial or institutional deployment. It is recommended that project teams conduct extensive k-fold cross-validation and independent site testing on external, out-of-sample data to systematically eliminate any residual risks of data leakage or localized over fitting. Finally, to advance the academic boundaries of construction duration modeling in developing economic contexts, future research should expand upon this framework along two methodological pathways. First, researchers should broaden the predictive feature profiles by integrating highly dynamic, real-time operational variables that reflect macro-economic and site-level volatility, such as contractor financial health indexes, supply chain disruptions, localized material inflation rates, and real-time labor productivity data. Second, further academic investigation is required to optimize the underlying architectures of the baseline models, specifically focusing on the advanced hyperparameter tuning of SVM kernels and deep learning architectures (ANNs).

Conflict of Interest Statement: We declare that the authors have no any conflict of interest.

REFERENCES

[1]. K. A. Shambel, “Construction time prediction model for public building projects in Addis Ababa, Ethiopia,” Eng. Constr. Archit. Manag. ahead-of-print(ahead-of-print), vol. ahead of p, 2021, doi: 10.1108/ECAM-11-2020-0975.

[2]. H. A. Ahmadu, “Modelling building construction durations in Nigeria,” Ahmadu Bello University, Zaria, Nigeria, 2014.

[3]. D. Mackova, M. Kozlovska, R. Baskova, M. Spisakova, and K. Krajnikova, “Construction-Duration Prediction Model for Residential Buildings in the Slovak Republic Based on Computer Simulation,” vol. 12, no. 13, pp. 3590–3599, 2017.

[4]. I. Mahamid, “The Development of Regression Models for Preliminary Prediction of Road Construction Duration,” vol. 3, no. 4, pp. 14–20, 2019.

[5]. J. M. Gerawork, B. B. Mitikie, and E. K. Yigzaw, “Performance evaluation of housing construction project using earned value analysis: The case of 20 / 80 condominium Addis

Received*: April 22, 2026;* ***Revised****: June 9, 2026;* ***Accepted:*** *June 16, 2026;* ***Published:*** *June 23, 2026*
Corresponding author- ***Kassahun Jima***

Ababa Bole Arabsa site," Am. J. Eng. Technol. Manag., vol. 5, no. 4, pp. 69–75, 2020, doi: 10.11648/j.ajetm.20200504.12.

[6]. G. A. Zelelew, "Analyzing factors that cause delay in housing development construction project in Ethiopia: The case of Addis Ababa 40/60 housing development project," no. May, 2023, doi: 10.13140/RG.2.2.26069.60647.

[7]. Y. R. Wang, C. Y. Yu, and H. H. Chan, "Predicting construction cost and schedule success using artificial neural networks ensemble and support vector machines classification models," Int. J. Proj. Manag., vol. 30, no. 4, pp. 470–478, 2012, doi: 10.1016/j.ijproman.2011.09.002.

[8]. A. Dhabi, "Development of an approximate construction duration prediction model during the project planning phase for general office buildings," vol. 24, no. 3, pp. 238–253, 2018.

[9]. K. M. M. El-dash, "Duration prediction models for construction projects in Middle East," Eng. Technol. Appl. Sci. Res., vol. 9, no. April, pp. 3924–3932, 2019, doi: 10.48084/etasr.2531.

[10]. Y. J. Kim, D. J. Yeom, and Y. S. Kim, "Development of construction duration prediction model for project planning phase of mixed-use buildings," J. Asian Archit. Build. Eng., vol. 18, no. 6, pp. 586–598, 2019, doi: 10.1080/13467581.2019.1696207.

[11]. Y. Dong-Jun, S. Hae-Mi, K. Yoo-Jun, C. Chung-Suk, and K. Youngsuk, "Development of an approximate construction duration prediction model during the project planning phase for general office buildings," J. Civ. Eng. Manag., vol. 24, no. 3, pp. 238–253, 2018, doi: https://doi.org/10.3846/jcem.2018.1646.

[12]. J. Abam, U. Elvis, M. Mbadike, and G. Uwadiegwu, "Prediction of cost and duration of building construction using artificial neural network," Asian J. Civ. Eng., no. 0123456789, 2022, doi: 10.1007/s42107-022-00474-4.

[13]. M. O. Sanni-Anibire, Zin, Mohamad, Rosli, Olatunji, and S. Olusanya, "Developing a machine learning model to predict the construction duration of tall building projects," J. Constr. Eng. Manag. Innov., vol. 4, no. 1, pp. 22–36, 2021, doi: 10.31462/jcemi.2021.01022036.

***Received**: April 22, 2026; **Revised**: June 9, 2026; **Accepted:** June 16, 2026; **Published:** June 23, 2026*
*Corresponding author- **Kassahun Jima***

[14]. S. Bayram, "Duration prediction models for construction projects : In terms of cost or physical characteristics ," vol. 00, no. 0000, pp. 1–12, 2016, doi: 10.1007/s12205-016-0691-2.

[15]. C. Koo, T. Hong, C. Hyun, and K. Koo, "A CBR-based hybrid model for predicting a construction duration and cost based on project characteristics in multi-family housing projects," vol. 752, pp. 739–752, 2010, doi: 10.1139/L10-007.

[16]. Y. Yang, X. I. N. Xia, D. Lo, T. Bi, J. Grundy, and X. Yang, "Predictive models in software engineering: challenges and opportunities," vol. 1, no. 1, 2016, doi: 10.1145/nnnnnnn.nnnnnnn.

[17]. L. M. F. Maués, J. Alberto, S. De Sá, C. Tavares, A. P. Kern, and A. A. A. M. Duarte, "Construction duration predictive model based on factorial analysis and fuzzy logic," pp. 115–133, 2019.

[18]. S. Petruseva, V. Zileska-pancovska, and D. Car-puˇ, "Implementation of process-based and data-driven models for early prediction of construction time," vol. 2019, 2019.

[19]. T. Yemane, Statistics: An Introductory Analysis, 2nd ed. 1967. Accessed: May 19, 2022. [Online]. Available: https://www.amazon.com/Statistics-Introductory-Analysis-Taro-Yamane/dp/B0000CNPXC

[20]. S. M. T. Ahmed, "A model for construction schedule and cost prediction using regularized gradient boosted," Bangladesh University of Engineering and Technology, 2020.

[21]. W. Lei, T. Zeng, Y. Tao, X. Wu1, and H. Chen, "Random Forest – Random forest," in 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), EDP Sciences, 2019, pp. 587–588. doi: 10.1051/e3sconf/202123703033.

[22]. H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," vol. 2024, pp. 69–79, 2024.

[23]. H. Fangohr et al., "Data exploration and analysis with Jupyter notebooks DATA EXPLORATION AND ANALYSIS WITH Jupyter NOTEBOOKS," no. January, 2019, doi: 10.18429/JACoW-ICALEPCS2019-TUCPR02.

[24]. [G. Biau, "Analysis of a Random Forests Model," J. ofMachine Learn. Res., vol. 13, pp. 1063–1095, 2012.

Received*: April 22, 2026;* ***Revised****: June 9, 2026;* ***Accepted:*** *June 16, 2026;* ***Published:*** *June 23, 2026*
Corresponding author- ***Kassahun Jima***

[25]. K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: From early developments to recent advancements," Syst. Sci. Control Eng., vol. 2, no. 1, pp. 602–609, 2014, doi: 10.1080/21642583.2014.956265.

[26]. M. M. . O. O. . A. A. Usman, "Feature selection: Its importance in performance prediction," Int. J. Eng. Sci. Comput., vol. 10, no. 5, pp. 25625–25632, 2020.

[27]. P. Shen, X. Ding, W. Ren, and S. Liu, "A stable feature selection method based on relevancy and redundancy," in Journal of Physics: Conference Series, 2021. doi: 10.1088/1742-6596/1732/1/012023.

[28]. M. A. Bouke, S. A. Zaid, and A. Abdullah, "Implications of data leakage in machine learning preprocessing: a multi-domain investigation".

[29]. E. Adinyira, E. A. Adjei, F. Desmond, and K. Fugar, "Application of machine learning in predicting construction project profit in Ghana using Support Vector Regression Algorithm (SVRA)," no. September, 2021, doi: 10.1108/ECAM-08-2020-0618.

[30]. J. Martín, A. M. Jiménez, and M. J. Navas, "Multiple Linear Regression : An Overview with Analytical and Physico-Chemical Applications," vol. 4, no. 11, pp. 32–60, 2017.

Received*: April 22, 2026;* ***Revised****: June 9, 2026;* ***Accepted:*** *June 16, 2026;* ***Published:*** *June 23, 2026*
Corresponding author- ***Kassahun Jima***