Journal of Construction Engineering, Management & Innovation
2021 Volume 4 Issue 1 Pages 022-036
https://doi.org/10.31462/jcemi.2021.01022036

www.goldenlightpublish.com

RESEARCH ARTICLE

Developing a machine learning model to predict the construction duration of tall building projects

Muizz O. Sanni-Anibire [*1], Rosli Mohamad Zin [2], Sunday Olusanya Olatunji [3]

[1] *King Fahd University of Petroleum and Minerals, Dammam Community College, Dhahran, Kingdom of Saudi Arabia*
[2] *Universiti Teknologi Malaysia (UTM), School of Civil Engineering, Faculty of Engineering, Johor, Malaysia*
[3] *Imam Abdulrahman Bin Faisal University, College of Computer Science and Information Technology, Department of Computer Science, Dammam, Kingdom of Saudi Arabia*

Abstract

The construction industry is witnessing a rapid rise in tall building projects due to an anticipated urban population explosion. However, this building typology has been subject to time overruns and total abandonment due to an underestimation of the project duration. Consequently, this paper presents the development of a model to predict the construction duration of tall building projects. In developing the model, a suite of machine learning algorithms was adopted including Multi-Linear Regression Analysis (MLRA), k-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Ensemble Methods. Thus, twelve models were developed in the process, and the most efficient model was selected. The procedure described in this study presents researchers and practitioners with a strategy to enhance the time performance of tall building projects through the adoption of modern digital technologies such as machine learning. The proposed model was based on an ensemble method using ANN as the combiner, with a Correlation Coefficient (R^2) of 0.69, Root Mean Squared Error (RMSE) of 301.72, and Mean Absolute Percentage Error (MAPE) of 18%.

Keywords

Duration prediction; Regression; k nearest neighbour; Neural networks; Support vector machines; Ensemble methods

Received: 12 January 2021; Accepted: 22 March 2021
ISSN: 2630-5771 (online)

1. Introduction

The 21st century is witnessing a rising complexity in buildings, embodied in the rapid growth of tall buildings in urban centers globally. These projects are however characterized by uncertainties that affect the success of the project, usually expressed in cost, time, and quality [1,2]. Experts are of the opinion that large variances in the estimated and actual duration of construction projects due to underestimation is one of the prevalent problems in the industry [3]. Bromilow [4] suggested that only one-eighth of building contracts were completed within the scheduled completion dates and that the average time overrun exceeded 40%. Likewise,

* Corresponding author
Email: muizzanibire@kfupm.edu.sa

Alzara et al. [5] reported delays in the range of 50% to 150 %.

Particularly, tall building projects are notorious for their delayed completion times. Interestingly, the Council on Tall Buildings and Urban Habitat, CTBUH [6] in its report "Dream Deferred: Unfinished Tall Buildings" noted the alarming rate of increase of "never completed" tall buildings. Previous researchers have suggested that a reliable prediction of the duration of construction projects is crucial to avoiding construction delays [7,8,3,9]. Traditional methods such as the Critical Path Method (CPM) or Program Evaluation Research Task (PERT) have been shown to consistently underestimate the actual project duration [10]. Typical considerations may include the client's time constraints, budget, or conducting a detailed analysis subject to skill, experience, and individual intuition of the project engineer. Therefore, there is a high level of subjectivity in the process which ultimately yields high levels of uncertainty [11].

In this regard, some research works have sought to apply Artificial Intelligence (AI) and Machine Learning (ML) to the duration prediction of construction projects [11-22]. These studies are however limited in the techniques used, as they have focused on one or two algorithms, without exploring ensemble methods to achieve improved performance. Moreover, despite the rapid growth of tall building construction, and the recurring time overruns of such projects, there is a dearth of research on the subject of its duration estimation.

In light of the foregoing, research to develop a model for the estimation of the duration of tall building projects based on ML has been conceptualized. Historical data on the construction duration of tall building projects has been obtained. The dataset was further used to develop duration prediction models based on popular machine learning algorithms such as Multi Linear Regression (MLRA), k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Artificial Neural Network (ANN), and Ensemble Methods. The performance of these models was evaluated based on the Correlation Coefficient (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The outcome of the systematic model development process described in this study is the proposed ML model for the duration prediction of tall building projects. The model can be described as an ensemble method that combines the outputs of ML algorithms considered in this study using ANN as the combiner.

2. Literature review

The following sections present an overview of relevant background on construction duration estimation. Firstly, traditional approaches, as well as modern trends in construction duration estimation, were reviewed. Subsequently, previous studies related to the development of mathematical models as well as the application of artificial intelligence and machine learning techniques have been presented.

2.1. Approaches to construction duration estimation

The duration of an activity is simply the length of time or period it takes to complete that activity. This is typically measured in hours, days, weeks, months, or years. Determining task durations utilizing detailed analysis is dependent on the required human and material resources, as well as the productivity rates of these resources. Traditionally, there are two modeling techniques used in construction project scheduling which are: the Critical Path Method (CPM) and Program Evaluation and Review Technique (PERT). The CPM schedule assumes the duration of work items is known with some level of certainty. On the other hand, PERT considers the uncertainty in determining the duration of work items. Hence, PERT is based on a "three-time estimate" i.e. optimistic estimate, most likely estimate, and the pessimistic estimate. The average of the "three-time estimate" is adopted as the duration [23]. Regardless of the methods applied, the calculated values remain approximate, and are characteristic of high levels of uncertainty. The estimator's background and experience are highly correlated to the accuracy of the estimation. Lack of adequate experience and thorough understanding of the

projects' scope of work will lead to poor estimations. Additionally, there exists the problem of material and labor price variations/fluctuations and inflation which are characteristic of construction projects. To solve the problem of uncertainty which may be due to insufficient information, variations, and human error, researchers have sought to employ more intelligent methods. Though research interests in duration estimation can be traced back to the 1960s, the past few decades have witnessed a resurgence [24,25]. The investigated approaches can be summarily classified into three including Artificial Intelligence (AI)-based scheduling which includes Knowledge-Based Scheduling, Expert systems and Case-Based Reasoning (CBR), Genetic Algorithms and Neural Networks; simulation-based scheduling; and integrated BIM-based scheduling [24].

2.2. Previous studies on the development of models for duration prediction

Bromilow [4] is accredited with developing the first empirical model that establishes the relationship between cost and time. Bromilow's Time-Cost (BTC) is based on historical data from 309 building projects completed in Australia between July 1964 and July 1967. The BTC model has been the subject of many other studies to re-calibrate and test the performance of the model in other locations and various project types. Further developments to Bromilow's model were made by Chan and Kumaraswamy [26] to combine the cost and floor area in a similar model. Other researchers studied the model further and included other variables in the equation. This could be seen as the foundation for later studies in developing mathematical models with the aid of the multi-linear regression method. Interestingly, the late 80s witnessed the acceptance of more intelligent methods such as AI to solving the inherent construction problem of estimating durations. Mohan [27] outlined 37 expert-system applications in the field of construction engineering and management. After five decades of Bromilow's initial model, a lot of technological advancements can be witnessed in construction project management, planning, and scheduling. However, construction projects continue to suffer from performance and productivity issues [28]. The following sections provide a non-exhaustive review of literature related to the application of artificial intelligence and machine learning techniques to duration estimation.

2.2.1. Knowledge-based expert system

Knowledge-based expert systems are computer programs originally developed in the field of Artificial Intelligence (AI) and designed to reach the level of performance of a human expert in some specialized problem-solving domain. Hendrickson et al. [29] presented a framework for modifying standard work productivities for activity duration estimation. The study proposed an expert system "MASON". Moselhi and Nicholas [30] also presented ESCHEDULER, a prototype system for precedence setting and modifying durations. Also, Shaked and Warszawski [31] developed HISCHED, which is a knowledge-based expert system for the construction planning of buildings.

2.2.2. Linear regression analysis

Hoffman et al. [32] identified the factors influencing construction duration through an assessment of 856 facility projects. The study compared the results of a multiple linear regression model with the BTC model and concluded that multiple linear regression provided a more accurate prediction. Yeom et al. [21] presented a multiple linear regression model that facilitates accurate prediction (94.72%) of construction durations of general office buildings in Korea. Blyth et al. [15] presented a multiple linear regression analysis which showed that twenty-one most influential project variables could accurately predict construction duration for buildings in the UK. The developed model was further validated with a new set of data which showed that the absolute percentage error for the overall duration varied between 0.38% and 6.68%. Lin et al. [33] developed a regression model for predicting the construction duration of steel-reinforced concrete building projects in Taiwan. Khosrowshahi and Kaka [12] proposed two models for cost and

duration with an adjusted coefficient of determination of 81.4% and 92.7% respectively.

Chan and Kumaraswamy [13] developed models to estimate the duration of various work packages based on data obtained from 15 case studies of residential buildings in Hong Kong and showed that the percentage error was about ±10% for overall construction durations. Abu Hammad et al. [8] utilized data from 140 projects in Jordan to develop regression models and concluded that there is a 95% probability that the proposed models could accurately predict project cost and duration with a precision of ±0.035% of the mean cost and time.

2.2.3. Neural networks (NNs)

Mensah et al. [18] utilized the historical data of 30 completed bridge projects in Ghana to develop a model for the prediction of construction duration. The study compared the stepwise regression method and artificial neural network (ANN), with the regression model having a MAPE of 25%, and ANN model with a MAPE of 26%. Attal [16] also compared the performance of regression analysis and ANN for predicting the duration of highway projects, with ANN having higher accuracy and reliability. Peško et al. [20] carried out a study which combined two popular artificial intelligence technique i.e. ANN and SVM for the estimation of costs and duration in construction projects. Both techniques displayed approximately equal performance, with the MAPE for SVM of 22.77% and ANN 26.26%.

2.2.4. Case-based reasoning (CBR)

Jin et al. [19] developed a CBR model for estimating construction duration based on 83 multi-housing projects. The results based on the MAPE of 5.74%-9.88% suggested the reliability of the model. Li et al. [11] established a revised CBR model to estimate the duration of skyscrapers in China. The results showed an accuracy of 69%. Koo et al. [17] utilized 101 completed multi-family housing projects to develop a CBR hybrid model with which to predict the construction duration and cost of a project in its early stage. The hybrid model features case-based reasoning, multiple regression analysis, artificial neural networks, genetic algorithm, and Monte-Carlo simulation.

2.2.5. Discussion on the previous studies

The extant literature reveals that the construction industry has evolved from its early adoption of Bromilow's Time-Cost model and its variants to more robust methods. It is observed, however, that there is a dearth of literature on the application of other machine learning algorithms. While linear regression and neural networks have dominated the discourse, not so much focus has been made to study the performance of algorithms such as k Nearest Neighbours, Support Vector Machines, as well as ensemble techniques. This may be partly attributed to barriers such as the huge quantity and confidentiality of data required. Data needed for machine learning application will need to be systematically documented by potential stakeholders. Another observation is that tall buildings have not received noteworthy attention in terms of machine learning applications to solve the problem of time-overruns, despite the significance of such projects in the urban habitat of the 21st century. Though the study by Li et al. [11] focused on the application of Case-Based Reasoning and k Nearest Neighbours to skyscrapers in China, it is also limited in its approach, while further investigation has the potential to provide better results. Therefore, the current study seeks to investigate the performance of selected machine learning algorithms in developing duration prediction models, as well as investigate the performance of ensemble models to seek an improved performance of the final prediction model.

3. Methodology

3.1. Dataset establishment

The primary source of the dataset used in this study is the Mega Project Case Study Center of China at http://www.mpcsc.org/case_search.htm. The data set was further corroborated with information from CTBUH's skyscraper center available at http://www.skyscrapercenter.com/country/china. A

sample size of 35 projects was identified with construction completion dates between 1993 and 2015. Remarkably, all the projects contained in the dataset are from China, which according to CTBUH [34] accounts for 61.5% of 200-meter-tall buildings in the world in 2018, and has maintained its role as the most prolific country in tall building construction for over two decades.

3.2. Data pre-processing

Since the dataset has been obtained from the real world, it may exhibit characteristics not ideal for ML modeling and thus require pre-processing and re-shaping. In this study, the Waikato Environment for Knowledge Analysis (Weka 3.8.3) has been used. This is an open-source machine learning software written in Java and developed at the University of Waikato, New Zealand [35]. Table 1 provides descriptive statistics of the numerical features of the dataset, while Table 2 describes the non-numerical features of the dataset and their conversion to dummy variables.

3.3. Views of the dataset

These are copies of the dataset in addition to the original dataset. They are created based on some system such as normalization and standardization. Evaluating the performance of algorithms on various views of the dataset will provide a general idea of the views that are better for the machine-learning problem [36]. Generally, the best algorithm to be used to solve an ML problem is usually not known beforehand. Experts have suggested that common ML algorithms should firstly be explored, especially those common in the field of the ML problem at hand [36,37]. In this study, four ML algorithms have been considered including Multi Linear Regression Analysis (MLRA), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM). The dataset was split into a train-test ratio of 66% to 34%. Additionally, five various views were considered as follows:

- Raw dataset: original dataset as described in Table 1.
- Normalized view (input features only): rescaling values in the input dataset to a range of 0 and 1, such that the largest value for each feature is 1 and the lowest is 0. Normalization is a good technique to use when the distribution of the data is either unknown, or is Gaussian (i.e. bell curve). The formula for normalization is expressed in Eq. 1:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Table 1. Descriptive statistics of the dataset

	Mean	Standard deviation	Maximum	Minimum	Missing values
GDP (bill USD)	302.91	108.22	446.31	80.77	0
# of elevators	50.66	31.31	130	6	6
Building area (m^2)	289364.06	152174.12	602401	91600	1
Floor area (m^2)	30569.68	45035.52	197000	4126	6
Height to tip (m^2)	386.18	113.64	636	237.5	0
# of floors above GF	76.97	23.22	128	37	0
Height of occupied floors (m)	339.30	115.88	610	213.9	1
# of total floors	80.91	23.49	133	39	0
# of basement floors	6.31	5.204	30	2	6
# of parking spaces	1058.32	619.86	2702	128	7
Cost (bill Yuan)	8.34	8.53	30	0.38	5
Duration (days)	1783	682.16	4555	730	0

Table 2. Description of the non-numeric features of the dataset

Features	Description	Conversion to dummy variables	Missing values
Facility type	O/Office, BOH/Business, office, hotel, ROH/Residential, office, hotel, BO/Business, office, BOR/Business, office, residential	O = 1; BOH = 2; ROH = 3; BO = 4; BOR = 5	0
Structural form	T-T/Tube in Tube, D/Diagrid, C-T/Core-Tube, T/Tubular	T-T = 1; D = 2; C-T = 3; T = 4	9
Structural material	RC/ Reinforced concrete, RCS/Reinforced concrete and steel, S/Steel C/Composite	RC = 1; RCS = 2; S = 3; C =4	0
Commencement period	Summer, autumn, winter and spring	Summer = 1; Autumn = 2; Winter = 3; Spring = 4	0

Normalized view (entire dataset): input and output values are converted to a range of 0 to 1. In this case, a further step was required to de-normalize the output data.

- Standardized view (input features): input values are rescaled such that the means are set at 0, and the standard deviation is 1. This technique is more useful if the dataset has a Gaussian (bell curve) distribution [36]. The process is executed according to equation 2 (where μ represents the arithmetic mean and σ the standard deviation):

$$x_{new} = \frac{x - \mu}{\sigma} \quad (2)$$

- Replace missing values: datasets for machine learning usually contain missing values that need to be treated by removing or replacing the missing values [36]. The ReplaceMissingValues filter in Weka was used to create this view where the missing values are set equal to the mean of the distribution for numerical features, and the mode for categorical features respectively.

3.4. Feature selection

The best view of the dataset determined from the previous section was selected for further processing. The "CorrelationAttributeEval" technique was used to determine the most relevant attributes contributing to the predictive performance. The correlation of various features in the dataset to the prediction output is firstly determined and subsequently ranked. Furthermore, attribute selection was based on the Recursive Feature Elimination (RFE) procedure [38]. In RFE, the entire feature set (V) ranked according to the correlation coefficient is split in half to derive the best V/2 features, and the worst V/2 features are eliminated. The splitting process continues recursively until only one best feature is left. Thereafter, the feature subset that achieved the best accuracy/or the best performance measure is finally chosen as the best subset to be used.

3.5. Hyperparameter optimization

The performance of ML algorithms is dependent on the tuning of optimal hyperparameters. It involves searching for the hyperparameters that result in the best performance of an algorithm given a set of data. The ML algorithms used in this study are described in the Weka environment as follows: MLRA: "LinearRegression", k-NN: "IBk", ANN: "Multilayer Perceptron", and SVM: "SMOReg". In determining the hyperparameters that yield optimal model performance, Weka was used to execute a modified systematic search i.e. a range of randomly spaced values are searched first, and then the range that performs best is zoomed in for further investigation. The optimal hyperparameter for KNN is the k value (search range 1 - 30), as well as the search and distance function, while ANN depends on the learning rate (search range 0.1 – 0.3), hidden layers (search range 1 – 4) and number

of nodes (search range 1 – 4). SVM optimization depends on the regularization factor C (search range 1 – 1000), the type of kernel function, as well as epsilon parameter (search range 0.1 – 0.00001).

3.6. Performance measurement

In measuring the performance of the algorithms employed, the Correlation Coefficient (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) have been employed. The mathematical expressions for these measures are presented in Eqs. 3-5 as follows:

$$R^2 = \frac{\sum(y_a - y'_a)(y_p - y'_p)}{\sqrt{\sum(y_a - y'_a)^2 \sum(y_p - y'_p)^2}} \quad (3)$$

where y_a and y_p are the actual and predicted values while y'_a and y'_p are the mean of the actual and predicted values.

$$RMSE = \sqrt{\frac{(y_a - y_p)^2 + (y_a - y_p)^2 + \cdots + (y_a - y_p)^2}{n}} \quad (4)$$

where (y_a - y_p) is the difference between the actual and predicted values and n is the size of the dataset used.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_a - y_p}{y_p}\right| \quad (5)$$

where y_a is the actual value and y_p is the predicted value, and n is the size of the dataset used.

3.7. Combining algorithms

To improve the performance of the techniques used, an ensemble method was used. This is an approach that combines the prediction outcomes of a set of algorithms with the same or different sets of features. This can be achieved through averaging (fixed rules) and stacking (trained rules) [39,40]. Averaging is a simple aggregation of the predictions of other models based on a fixed rule such as the mean, maximum and minimum values. Stacking is an extension of averaging which allows another algorithm to learn how best to combine the predictions of other models [36]. The systematic procedure followed in this study is further summarized and illustrated in Fig. 1.

4. Results and findings

4.1. Comparison of various views of the dataset

In this study, five views of the dataset were prepared for comparison. The datasets have been evaluated with the four selected ML algorithms (MLRA, ANN, KNN & SVM), and the results showed that all algorithms performed best when the entire view of the dataset was normalized i.e. including the output feature (Table 3). While the comparative performance (based on RMSE values) of the various views of the dataset were insignificantly different, the normalized view of the entire dataset displayed exemplary performance. A decrease in error of at least 51% for MLRA, 46% for ANN, 43% for KNN, and 55% for SVM was observed when comparing the normalized view of the entire dataset with the worst-performing view.

4.2. Performance of machine learning algorithms

To develop the initial models for which the performance of ML algorithms will be evaluated, the best combination of features that yields optimum performance was determined. This was achieved using the *"CorrelationAttributeEval"* and RFE techniques discussed previously in the "feature selection" section. Thus, in addition to a dataset containing all features, four more feature sets were developed as described in Table 4. It can be observed from the Table, that the most important feature influencing the duration of tall building projects is the number of total floors followed by the number of floors above the ground floors. As shown in Table 5, the performance of the four algorithms varied with different sets of features. MLRA exhibited the best performance with the best two features and best feature respectively i.e. "# of total floors" and "# of floors above ground floors". It is also observed that both features (best V/8 and best feature) exhibit similar performance for all algorithms except KNN. This suggests that both features are collinear, and one of them could satisfactorily replace the other.

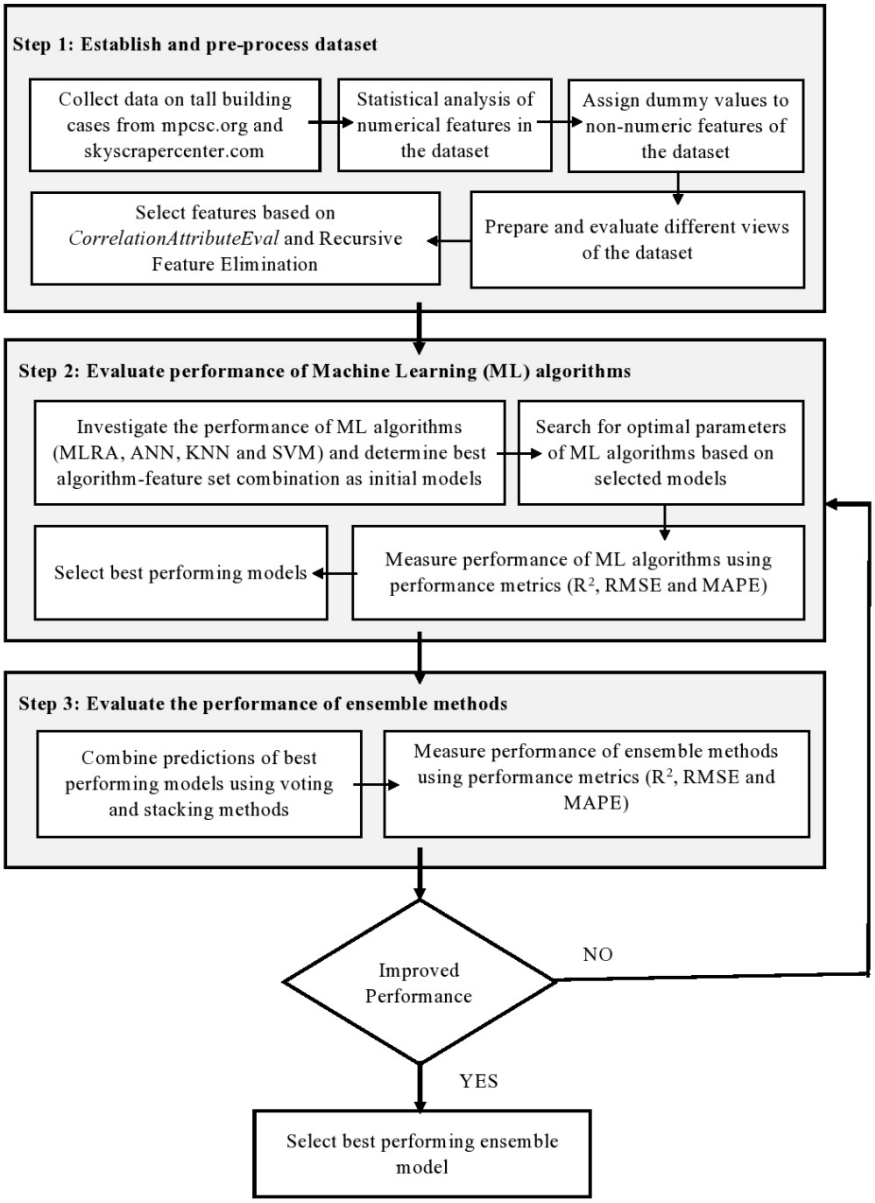

Fig. 1. Methodology for developing the proposed ML duration prediction model

Table 3. Performance (RMSE) of ML algorithms for various views of the dataset

ML Algorithm	Raw data set	Replace missing values	Normalized view (input features only)	Normalized view (entire dataset)	Standardized view (input features)
MLRA (*LinearRegression*)	1365.52	1332.52	1724.82	652.47	1365.52
ANN (*Multilayer Perceptron*)	1652.56	1495.74	1652.57	800.47	1652.56
KNN (*IBk*)	1067.58	1507.96	1067.58	611.67	1067.58
SVM (*SMOReg*)	1231.49	1195.16	1229.36	540.39	1228.87

Table 4. Description of selected feature subsets based on *CorrelationAttributeEval*

RFE process	No. of features	Description
Best V/2 features	8	# of total floors; # of floors above ground floors; # of parking spaces; cost; building area; height to tip; floor area; # of elevators
Best V/4 features	4	# of total floors; # of floors above ground floors; # of parking spaces; cost
Best V/8 features	2	# of total floors; # of floors above ground floors
Best feature	1	# of total floors

Table 5. Performance (RMSE) of ML algorithms for various feature subsets

ML Algorithm	All features	Best V/2	Best V/4	Best V/8	Best feature
MLRA (*LinearRegression*)	652.47	509.67	1154.69	538.32	538.32
ANN (*Multilayer Perceptron*)	800.47	371.78	**208.61**	248.79	248.62
KNN (*IBk*)	611.67	**225.89**	957.16	392.76	**369.39**
SVM (*SMOReg*)	540.39	293.14	**261.19**	299.69	**299.69**

ANN performed best with the best V/4 features (described in Table 4), with an RMSE of 208.61, a 74% decrease in error compared to the case where all features were used (RMSE of 800.47), as shown in Table 5. KNN performed best with the best V/2 features (RMSE of 225.89), while SVM performed best with the best V/4 features (RMSE of 261.19). The results presented in Table 5 formed the basis for the development of the initial models for further investigation through hyperparameter tuning and optimization.

Based on the performance of various combinations of ML algorithms and feature sets, five initial models have been selected. The highlighted figures in Table 5 indicate the preferred configuration for the models. To further optimize the performance of ML algorithms, tuning their hyperparameters becomes necessary. Table 6 presents the developed models showing the combination of ML algorithms, feature sets, and optimal hyperparameters. The performance of the various models is presented in Table 7. To determine the models with the best performance, the models with the lowest RMSE and MAPE values, as well as the highest R^2 values are considered first. It can be observed that MOD1, MOD2, and MOD4 performed best when the RMSE, MAPE, and R^2 results are compared. Though, MOD3 had a better correlation coefficient (R^2) compared to MOD4, a cross plot of the actual versus predicted values presented in Fig. 2 shows that the predicted values for MOD3 was simply a general average and thus did not reflect a realistic prediction which is evident in the high inaccuracies from the RMSE and MAPE values.

4.3. Performance of ensemble methods

To seek further improvement in the predictive performance, an ensemble method was adopted. The three best performing models (MOD1, MOD2 & MOD4) were combined using fixed and trained rules also referred to as averaging and stacking respectively. Thus, seven more models were created for further investigation as shown in Table 8.

4.3.1. Averaging (fixed rules)

To combine the selected best-performing models through a fixed rule system, the mean, maximum and minimum values of the models' predicted outputs were considered. This formed MOD6, MOD7 & MOD8. The performance results are presented in Table 9. It can be observed that MOD8 performed best with the least RMSE and MAPE values, as well as the highest R^2 value.

Table 6. ML models (combinations of algorithms, optimized hyperparameters and selected feature sets)

Model	ML Algorithm	Selected features	Optimization hyperparameters
MOD1	ANN (*Multilayer Perceptron*)	# of total floors; # of floors above ground floors; # of parking spaces; cost	0.3 learning rate; one hidden layer with four nodes
MOD2	KNN (*IBk*)	# of total floors; # of floors above ground floors; # of parking spaces; cost; building area; height to tip; floor area; # of elevators	Nearest Neighbor: LineraNN Distance function: Manhattan Distance K: 1
MOD3	KNN (*IBk*)	# of total floors	Nearest Neighbor: LineraNN Distance function: Manhattan Distance K: 18
MOD4	SVM (*SMOReg*)	# of total floors; # of floors above ground floors; # of parking spaces; cost	Kernel: Polykernel Cost function, C: 1 Epsilon: 1E -12 Epsilon parameter: 1E -3
MOD5	SVM (*SMOReg*)	# of total floors	Kernel: Pearson VII function based universal kernel (PUK) Cost function, C: 100 Epsilon: 1E -12 Epsilon parameter: 1E -4

Table 7. Performance of initial ML models

Performance measure	MOD1	MOD2	MOD3	MOD4	MOD5
RMSE	356.26	380.79	477.91	446.06	481.17
R^2	0.53	0.64	0.54	0.49	0.47
MAPE	0.22	0.22	0.31	0.27	0.32

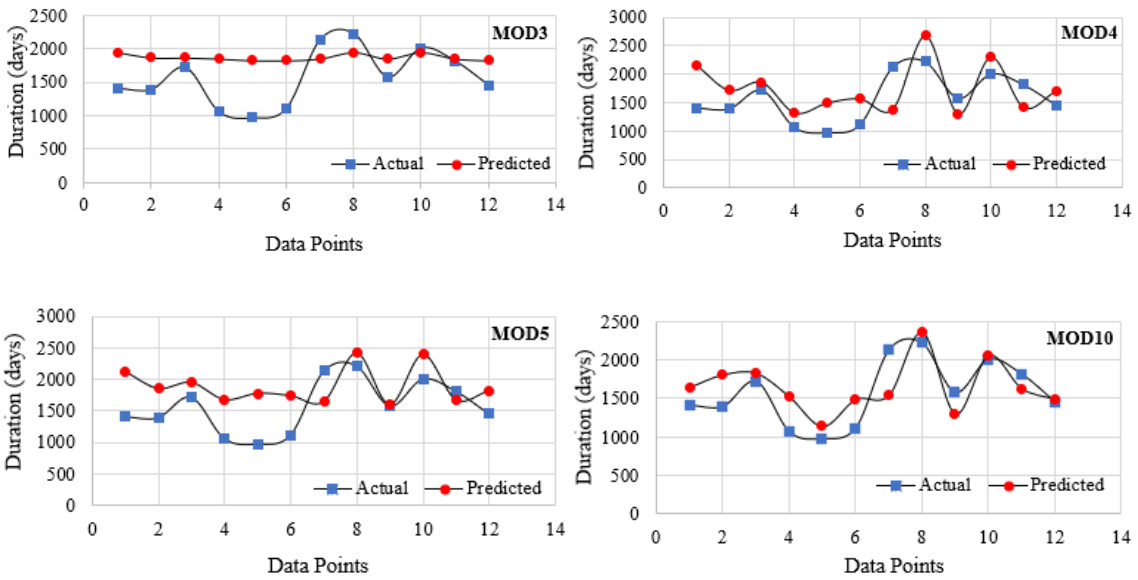

Fig. 2. Cross-plots of the actual vs. predicted duration values for selected models

Table 8. ML models based on Ensemble Methods

Model	Ensemble Method (Input models: MOD1; MOD2; MOD4)
MOD6	Mean
MOD7	Maximum
MOD8	Minimum
MOD9	MLRA
MOD10	ANN (0.3 learning rate; one hidden layer with four nodes)
MOD11	KNN (Nearest Neighbor: LineraNN, Distance function: Euclidean Distance, K: 2)
MOD12	SVM (Kernel: Pearson VII function based universal kernel (PUK),Cost function, C: 5, Epsilon: 1E -12, Epsilon parameter: 1E -5)

Table 9. Performance of ML models based on Ensemble Method

Performance measure	MOD6	MOD7	MOD8	MOD9	MOD10	MOD11	MOD12
RMSE	338.67	473.23	331.43	372.63	301.76	310.13	349.59
R^2	0.64	0.59	0.67	0.64	0.69	0.71	0.69
MAPE	0.21	0.30	0.17	0.21	0.18	0.18	0.19

4.3.2. Stacking (trained rules)

The three best performing models (MOD1, MOD2 & MOD4) were also combined using the trained rules, while the four algorithms considered in the study were employed as the combiner system. The details of the optimal hyperparameters for the combiner system are presented in Table 8. Thus, four more models were developed labeled as MOD9, MOD10, MOD11 & MOD12. As shown in Table 9, the best performing model was considered to be MOD10 based on the low values of RMSE and MAPE. It can be observed that the correlation coefficient (R^2) of some of the models, specifically MOD8, MOD10, MOD11, and MOD12 were approximately the same. Consequently, the performance was decided based on the reduced error observed in the RMSE and MAPE values. It may also be inferred that to seek further improvement in predictive performance may require some other strategies such as the establishment of a larger dataset, or seeking to investigate other machine learning algorithms, as well as automated approaches to hyperparameter tuning.

5. Discussion

Tall building construction is rapidly developing in the urban context as a sustainable solution to an impending housing crisis and urban population explosion. The complexity involved in the design and construction of many tall buildings has resulted in notorious time overruns, incompletion, and total abandonment [6,9,38]. Time overruns in tall building projects could lead to dissatisfied stakeholders, litigation, project abandonment, and ultimately a failure in fulfilling its intended purposes. Therefore, previous studies have suggested that the use of mathematical models, as well as data mining/machine learning to predict construction duration, is a viable mitigation strategy [41]. The studies that dominate the research arena are limited in the techniques employed, especially concerning tall building projects.

In light of the foregoing, this study developed a machine learning model based on a systematic investigation and further combination of various machine learning algorithms. The study firstly established a dataset and subsequently carried out pre-processing of the data. The results showed that the view of the dataset which enhanced the

performance of machine learning algorithms was the normalized view, where all features (input and output) were re-scaled to range between 0 and 1. The study further explored various feature sets that contribute to the performance of the algorithm. This was to identify and select the features that contribute to improving the predictive performance of the ML algorithm. Additionally, feature selection helps to control the "curse of dimensionality", which is a phenomenon characteristic of real data. In this study, the most relevant feature that correlates with the output for prediction (i.e. duration) was the number of total floors. This is a logical outcome when considering the nature of the study's focus (i.e. tall buildings). Further to that was the algorithm evaluation process, where the hyperparameters of the selected algorithms were adjusted to determine the optimal values for the initial duration estimation models (MOD1, MOD2, MOD3, MOD4, and MOD5). The initial duration estimation models were further combined through ensemble techniques i.e. fixed and trained rules. The final result from the overall process was the selection of a model which was based on a combination of three initial models using ANN as the combiner. The best performing model in this study described as MOD10 had an R2 of 0.69, MAPE of 0.18, and RMSE of 301.76.

The level of accuracy of MOD10 suggests that it could be recommended as a decision support tool in estimating the duration of tall building projects. A comparison of the performance of MOD10 with the poorest performing model in this study (i.e. MOD5) revealed a gain in performance of 47% was achieved in the correlation coefficient (R2), 44% in MAPE, and 37% in RMSE. Likewise, MOD10 arguably outperforms the CBR model developed by Li et al. [11] for skyscrapers. The R2 value obtained in this study for MOD10 was 69%, while a CBR model developed by Li et al. [11] achieved 62%; which was only improved to 69% when some poorly predicted cases were deleted from the CBR model. Thus, the superiority of MOD10 is apparent. The limitations of this study may be reflected in the source and size of the dataset used. The dataset in this study contains about 35 cases that may impact the predictive performance of the model built, due to the limited amount of data available for training and testing. Similarly, previous research works in duration prediction have relied on similar sizes of the dataset. It may thus be concluded that the construction industry is deficient in recording and publishing data suitable for ML applications for duration prediction. Furthermore, the study's limitation in the dataset being sourced from china may be relieved due to China being the major driver of tall building construction globally [34]. Also, the dataset contains the GDP of the cities where the building projects are located and thus may provide an opportunity for extension to other construction climates globally based on the GDP of a city. The significance of this study is reflected in its addressing a current trend in the construction industry, which is the exponential rise of tall building projects in urban centers across the globe. These projects are known to be characterized by their delayed completion times.

6. Conclusion

This study demonstrated the significance of leveraging the capabilities of machine learning for enhanced time performance in the construction industry. Specifically, the literature reveals that there is a dearth of studies in the construction domain on the time performance of tall buildings. Tall building projects have become a dominant building typology of the modern urban habitat. Furthermore, it is now widely considered an important area of construction engineering and management research. Many factors may contribute to studying this specific building typology separate from other/horizontal building types. For instance, this study reveals that the most significant factor influencing the construction duration is the total number of floors – an intrinsic attribute of tall buildings. Other potential factors may include the structural systems used. Furthermore, tall buildings have specific characteristics that may lead to their delayed completion times, such as the complexity in design and construction, as well as the large number of professionals involved.

Notably, tall buildings are also considered a viable strategy towards sustainable urban development. Poor time performance of such projects will defeat their intended purpose of providing adequate urban space for the inevitable population. Therefore, accurate estimation of the duration of such projects based on historic data is of potential value to the broader society. Thus, the contribution of this study to the global community is shown in facilitating timely delivery of tall building projects as a sustainable strategy to an impending housing crisis. As regards the study's contribution to the construction community, it enhances time performance through the adoption of modern digital technology in a rarely researched domain as tall buildings. It is widely acknowledged that time performance is a crucial measure of project success in the construction industry.

This study achieved this contribution through a systematic approach in developing models through the application of established machine learning algorithms as well as combinations of the same. The study did not intend to develop new machine learning algorithms. However, the study has methodically applied existing algorithms in developing predictive models for estimating the duration of tall building projects. The model thus proposed in this study was based on an ensemble method using ANN as the combiner. Remarkably, the model's accuracy which is comparable to similar studies suggests its suitable adoption as a decision support tool. Convincingly, the application of machine learning has the potential to make the process of duration estimation smarter and more efficient. The model proposed in this study may be limited in its level of generalization, as is the case with data-driven models. However, the systematic procedure described herein could be adapted to other datasets, while the dataset used in the current study could be expanded for enhanced performance and applicability. Forthcoming research will seek to incorporate such predictive models into computing tools used in construction project management, and also make comparative assessments with traditional methods.

Data availability statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] Hamta N, Ehsanifar M, Sarikhani J (2018) Presenting a goal programming model in the time-cost-quality trade-off. International Journal of Construction Management, 21(1): 1-11.

[2] Williamson M, Ganah A, John GA (2019) Barriers to adopting modern methods of construction in the UK. Journal of Construction Engineering, 2(1): 30-39.

[3] Sanni-Anibire MO, Mohamad Zin R, Olatunji SO (2020) Causes of delay in the global construction industry: a meta analytical review. International Journal of Construction Management, doi:10.1080/15623599.2020.1716132.

[4] Bromilow FJ (1969) Contract time performance expectations and the reality. In Building forum, 1(3): 70-80.

[5] Alzara M, Kashiwagi J, Kashiwagi D, Al-Tassan A, (2016) Using PIPS to minimize causes of delay in Saudi Arabian construction projects: university case study. Procedia Engineering, 145: 932-939.

[6] CTBUH (2014) Dreams deferred: unfinished tall buildings. CTBUH Journal, 4: 46-47.

[7] Abdul-Rahman H, Berawi MA, Berawi AR, Mohamed O, Othman M, Yahya IA, (2006) Delay mitigation in the Malaysian construction industry. Journal of construction engineering and management, 132(2): 125-133.

[8] Abu Hammad AAA, Ali SMA, Sweis GJ, Bashir A, (2008) Prediction model for construction cost and duration in Jordan. Jordan Journal of Civil Engineering, 2(3): 250-266.

[9] Sanni-Anibire MO, Zin RM, Olatunji SO (2020) Causes of Delay in Tall Building Projects in GCC Countries. Proceedings of the 8th International Conference on Construction Engineering and Project Management Dec. 7-8, 2020, Hong Kong SAR

[10] Ballesteros-Pérez P, Larsen GD, González-Cruz MC (2018) Do projects really end late? On the shortcomings of the classical scheduling techniques. JOTSE: Journal of Technology and Science Education, 8(1): 17-33.

[11] Li Y, Lu K, Lu Y (2016) Project schedule forecasting for skyscrapers. Journal of Management in Engineering, 33(3): 05016023.

[12] Khosrowshahi F, Kaka AP (1996) Estimation of project total cost and duration for housing projects in the UK. Building and Environment, 31(4): 375-383.

[13] Chan DWM, Kumaraswamy MM (1999) Forecasting construction durations for public housing projects: a Hong Kong perspective. Building and environment, 34(5): 633–646.

[14] Skitmore RM, Ng ST (2003) Forecast models for actual construction time and cost. Building and environment, 38(8): 1075-1083.

[15] Blyth K, Lewis J, Kaka A (2004) Predicting project and activity duration for buildings in the UK. Journal of Construction Research, 5(02): 329-347.

[16] Attal A. Development of neural network models for prediction of highway construction cost and project duration. Doctoral dissertation, Ohio University, 2010.

[17] Koo C, Hong T, Hyun C, Koo K (2010) A CBR-based hybrid model for predicting a construction duration and cost based on project characteristics in multi-family housing projects. Canadian Journal of Civil Engineering, 37(5): 739-752.

[18] Mensah I, Nani G, Adjei-Kumi T (2016) Development of a Model for Estimating the Duration of. Bridge Construction Projects in Ghana, International Journal of Construction Engineering and Management, 5(2): 55-64.

[19] Jin R, Han S, Hyun C, Cha Y (2016) Application of case-based reasoning for estimating preliminary duration of building projects. Journal of Construction Engineering and Management, 142(2): 04015082.

[20] Peško I, Mučenski V, Šešlija M, Radović N, Vujkov A, Bibić D, Krklješ M (2017) Estimation of costs and durations of construction of urban roads using ann and svm. Complexity, Article ID: 2450370.

[21] Yeom DJ, Seo HM, Kim YJ, Cho CS, Kim Y (2018) Development of an approximate construction duration prediction model during the project planning phase for general office buildings. Journal of Civil Engineering and Management, 24(3): 238-253.

[22] Sağlam B, Bettemir ÖH (2018) Estimation of duration of earthwork with backhoe excavator by Monte Carlo Simulation. Journal of Construction Engineering, Management & Innovation 1(2): 85-94.

[23] Hinze J. Construction Planning and Scheduling. Pearson/Prentice Hall, 2011.

[24] Faghihi V, Nejat A, Reinschmidt KF, Kang JH (2015) Automation in construction scheduling: a review of the literature. The International Journal of Advanced Manufacturing Technology, 81(9-12): 1845-1856.

[25] Liu H, Al-Hussein M, Lu M (2015) BIM-based integrated approach for detailed construction scheduling under resource constraints. Automation in Construction, 53: 29-43.

[26] Chan DW, Kumaraswamy MM (1995) A study of the factors affecting construction durations in Hong Kong. Construction Management and Economics, 13(4): 319-333.

[27] Mohan S (1990) Expert systems applications in construction management and engineering. Journal of Construction Engineering and Management, 116(1): 87-99.

[28] Al-Kofahi ZG, Mahdavian A, Oloufa A (2020) System dynamics modeling approach to quantify change orders impact on labor productivity 1: principles and model development comparative study. International Journal of Construction Management, doi:10.1080/15623599.2020.1711494.

[29] Hendrickson C, Zozaya-Gorostiza C, Rehak D, Baracco-Miller E, Lim P (1987) Expert system for construction planning. Journal of Computing in Civil Engineering, 1(4): 253-269.

[30] Moselhi O, Nicholas MJ (1990) Hybrid expert system for construction planning and scheduling. Journal of Construction Engineering and Management, 116(2): 221-238.

[31] Shaked O, Warszawski A (1995) Knowledge-based system for construction planning of high-rise buildings. Journal of Construction Engineering and Management, 121(2): 172-182.

[32] Hoffman GJ, Thal Jr AE, Webb TS, Weir JD (2007) Estimating performance time for construction projects. Journal of Management in Engineering, 23(4): 193-199.

[33] Lin MC, Tserng HP, Ho SP, Young DL (2011) Developing a construction-duration model based

on a historical dataset for building project. Journal of Civil Engineering and Management, 17(4): 529-539.

[34] CTBUH (2018) CTBUH Year in Review: Tall Trends of 2018. https://www.skyscrapercenter.com /year-in-review/2018

[35] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2016.

[36] Brownlee J (2018) Machine learning mastery with Weka. https://machinelearningmastery.com/machine-learning-mastery-weka/

[37] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH (2008) Top 10 algorithms in data mining. Knowledge and Information Systems, 14: 1-37.

[38] Sanni-Anibire MO, Zin RM, Olatunji SO (2020) Machine learning model for delay risk assessment in tall building projects. International Journal of Construction Management, doi:10.1080/15623599. 2020.1768326

[39] Xia R, Zong C, Li S (2011) Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 181(6): 1138-1152.

[40] Kuncheva LI. Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, 2014.

[41] Gunduz M, Nielsen Y, Ozdemir M (2015) Fuzzy Assessment Model to Estimate the Probability of Delay in Turkish Construction Projects. Journal of Management in Engineering, 31(4): 04014055.