Application of stacking ensemble machine learning algorithm in predicting the cost of highway construction projects

Received 27 February 2020
Revised 27 July 2020
21 October 2020
Accepted 27 May 2021

Meseret Getnet Meharie
School of Civil Engineering and Architecture,
Adama Science and Technology University, Adama, Ethiopia
Wubshet Jekale Mengesha
Ethiopian Institute of Architecture, Building Construction and City Development,
Addis Ababa University, Addis Ababa, Ethiopia
Zachary Abiero Gariy
School of Civil, Environmental and Geospatial Engineering,
Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya, and
Raphael N.N. Mutuku
Faculty of Engineering and Technology, Technical University of Mombasa,
Mombasa, Kenya

Abstract

Purpose – The purpose of this study to apply stacking ensemble machine learning algorithm for predicting the cost of highway construction projects.

Design/methodology/approach – The proposed stacking ensemble model was developed by combining three distinct base predictive models automatically and optimally: linear regression, support vector machine and artificial neural network models using gradient boosting algorithm as meta-regressor.

Findings – The findings reveal that the proposed model predicted the final project cost with a very small prediction error value. This implies that the difference between predicted and actual cost was quite small. A comparison of the results of the models revealed that in all performance metrics, the stacking ensemble model outperforms the sole ones. The stacking ensemble cost model produces 86.8, 87.8 and 5.6 percent more accurate results than linear regression, vector machine support, and neural network models, respectively, based on the root mean square error values.

Research limitations/implications – The study shows how stacking ensemble machine learning algorithm applies to predict the cost of construction projects. The estimators or practitioners can use the new model as an effectual and reliable tool for predicting the cost of Ethiopian highway construction projects at the preliminary stage.

Originality/value – The study provides insight into the machine learning algorithm application in forecasting the cost of future highway construction projects in Ethiopia.

Keywords Cost prediction, Machine learning algorithms, Stacking ensemble model, Highway construction projects

Paper type Research paper

Authors thank the Ethiopian Road Authority (ERA) for supporting allowing us to access highway construction project data.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this article.

Data Availability: The data used to support the findings of this study are available from the corresponding author upon request.

Engineering, Construction and Architectural Management
Vol. 29 No. 7, 2022
pp. 2836-2853
© Emerald Publishing Limited
0969-9988
DOI 10.1108/ECAM-02-2020-0128

1. Introduction

The project cost estimate is a projection of the probable cost of the specific project, based on the information and knowledge available at the time of estimation (PMBOK, 2011). Accuracy of cost estimation is a crucial factor in helping the contractor and the customer to make adequate financial provisions (Shin, 2015). Though several studies have shown that the accuracy of the various predictions made throughout the life of the project significantly determines the success or failure of the project, creating a simple and accurate cost forecast at the preliminary stage of the project is one of the most challenging activities in the management of construction projects (Enshassi *et al.*, 2013; Abdal-hadi, 2014; Alumbugu *et al.*, 2014; Hatamleh *et al.*, 2018; Mahamid, 2015). This is because the preliminary calculation will be made before the design of the project has been completed (Enshassi *et al.*, 2013). At this stage of the project, due to partial drawings and documents, high uncertainty and complication, adequate information is not provided (Jarkas *et al.*, 2013).

Machine learning (ML) algorithm is a viable method to alleviating the above-stated problem in the area of construction estimation. Because ML is more accurate, automated, fast, customizable and scalable for data-driven work over human-made rules (Shin, 2015). ML algorithms do have at least one modeling algorithm, several input variables or project features, and predictions. These algorithms can have higher estimation capabilities to solve complex problems (Huang *et al.*, 2015). Recently, ML algorithms have been used as a substitute or in conjunction with linear regression approaches (El-Kholy, 2015; Golizadeh *et al.*, 2017; Peško *et al.*, 2017). Common representative ML approaches used in the construction estimation domain include support vector machine (SVM) (Peško *et al.*, 2017; Rafiei and Adeli, 2018), artificial neural networks (ANN) (Golizadeh *et al.*, 2016, 2017), case-based reasoning (CBR) (Kang *et al.*, 2011) and gradient boosting trees (GBTs) (Shin, 2015).

Ensemble-based machine learning approaches are becoming the next generation cost estimation system in the construction sector. Ensemble ML techniques use multiple learning algorithms to achieve superior predictive efficiency, in terms of accuracy and stability, over any single learning algorithm (Breiman, 1996; LeDell, 2015; Kansara *et al.*, 2018; Neloy *et al.*, 2019). Ensemble learning methods, a synthesis of multiple model algorithms, have been widely applied to regression problems in several disciplines, such as demand forecasting model for supply chain (Kilimci *et al.*, 2019); financial market prediction (Henrique *et al.*, 2019); energy load prediction in residential buildings (Al-Rakhami *et al.*, 2019); electricity load and price forecasting (Do, 2018; Agrawal *et al.*, 2019; Zahid *et al.*, 2019); house price prediction model in real estate market (Kansara *et al.*, 2018; Neloy *et al.*, 2019); prediction of computer Go player attributes (Moudřík and Neruda, 2015); warfarin dose estimation in the health sector (Ma *et al.*, 2018); and software effort estimation (Banimustafa, 2018).

There is, however, a lack of application of the ML ensemble algorithm to the construction estimation domain. In particular, the development of an estimation model to determine the future cost of early-stage highway projects using an automated stacking ensemble learning algorithm has not yet been investigated. In addition, the linear combination of LR, SVM and ANN model algorithms in predicting the cost of construction project has not yet been studied, to the knowledge of the authors. Therefore, the objective of this study is to present a two-level ML algorithm called a stacking ensemble algorithm to linearly and optimally estimate the cost of the highway construction project by combining MLR, SVM and ANN algorithms at the early stage of the project while filling knowledge gaps in the existing literature.

2. Literature review

In the early stages of the construction project and the absence of extensive and detailed project information, construction managers use various systematic model algorithms to determine the cost of the construction projects. The models use mathematical expressions to estimate the cost of the project based on one or more input variables. Based on the modeling techniques used in previous studies, some of the techniques are reviewed and summarized in this work.

2.1 Multiple linear regression

Linear regression is a method for modeling the relationship between one or more independent variables and one dependent variable (Petruseva *et al.*, 2017). When there is only one independent variable and the relationship is linear it is called simple linear regression. Unlike simple linear regression, the multiple linear regression (MLR) model uses linear relationships between several independent variables. Linear regression has been used in the construction industry to estimate or forecast the costs and duration of construction projects at an early stage. Alemayehu (2014) developed conceptual cost estimation models for asphalt road construction projects in Ethiopia. This finding shows how regression models based on significant variables or bid quantity can be used to build regression models as tools for forecasting future road construction costs.

In the study conducted by Kaleem *et al.* (2014), highway project duration was estimated based on variables such as planned costs and project type which are known in the planning phase. A mathematical relationship (correlation) between highway project duration, planned cost and project type was demonstrated in this study through various model specifications such as forward, backward and stepwise multivariate regression analysis. The regression models have also been developed based on real historical data from similar building projects to predict the costs and duration of building construction projects (Thomas and Thomas, 2016). According to Zhai *et al.* (2016), the use of parametric modeling and historic contract times created a method that is more accurate in estimating contract times. These authors used project cost and bid item quantity data in multivariate regression-based modeling to develop a set of contract period estimates for projects in a matter of minutes using data readily available at later design stages. The results of this study represented the development of a duration estimation system that showed significantly higher accuracy.

Barraza *et al.* (2017) conducted a study to analyze schedule deviations in road construction projects and the impact of project physical characteristics on schedule deviation regression models. In addition, regression models that analyze the effect of physical project characteristics on time deviations were addressed. Thaseena and Vishnu (2017) also developed a probabilistic cost overrun analysis model for construction projects as a decision support tool for contractors before the bidding stage. Their study aimed to identify critical factors affecting cost overruns and to obtain models using multiple regression and artificial neural networks. The authors developed models were then validated and findings showed the better outcomes of cost overrun of highway projects. Čeh *et al.* (2018) presented the predictive performance of the MLR and random forest machine learning technique for estimating apartment prices.

2.2 Support vector machines

Support vehicle machines (SVM) is a widely used technique that is remarkable for its theoretical and practical sound. SVM is a supervised artificial intelligence model employed for both classification and regression. SVM has two functions: support vector classifier (SVC) for classification and support vector regression (SVR) for regression (Zahid *et al.*, 2019).

SVM has been successfully used in several real-world problems: financial and time series regression problems, object recognition, convex quadratic programming, handwriting digit recognition and choices of loss functions (El-sawalhi, 2015; Kilimci *et al.*, 2019). It has also been applied in the construction industry to estimate the parametric costs and duration of construction projects (El-sawalhi, 2015). This author developed a model using SVM to estimate the parametric costs of road construction projects. The developed model was able to predict the cost road construction project with 95% prediction accuracy by using seventy project cases. Petruseva *et al.* (2017) applied and compared two predictive models: MLR and SVM model using 75 datasets to forecast the cost of construction projects. Comparison of the model results showed that the SVM forecast was substantially more reliable than the MLR. Peško *et al.* (2017) analyzed and compared the prediction performance of artificial neural networks (ANNs) and SVM. This analysis revealed that SVM was more accurate. Wu (2017) applied SVM to predict the price of houses in King County, USA, to help both the buyers and sellers make their decisions.

2.3 Artificial neural networks

Artificial neural networks (ANN) is a soft computing tool that mimics the ability of human minds to use modes of reasoning and/or pattern recognition effectively (Kulkarni *et al.*, 2017). One of the applications of neural network in the engineering fields is to predict or estimate the outcome of nonlinear statistical problems and is widely used to model complex relationships between inputs and outputs or to find patterns in the given datasets (Golizadeh *et al.*, 2016; Magdum and Adamuthe, 2018). Several studies have shown the great application of ANN in civil engineering and construction management field area: prediction, estimation, risk analysis, decision-making, resources optimization, classification and selection (Jain and Pathak, 2014). In particular, ANNs have been applied to estimate the cost and duration of highway projects in the early phases of project development where insufficient project information is available (Barakchi *et al.*, 2017).

In an attempt to support early, cost-effective decisions, Marinelli *et al.* (2015) provided a new, robust and accurate model for bridge superstructure cost estimation using ANN. The established model captured very well the complex interrelationships in the data set, which offered accurate estimations of the final quantity for bridges and demonstrated a good generalization capability. Elbeltagi *et al.* (2014) studied to support decision-makers in predicting the conceptual cost of highway construction projects in Libya. In this study, a cost predictive model was developed using an ANN. This model was then validated and the results indicate a better estimate of the realistic cost of highway projects in Libya. Naik and Radhika (2015) developed different ANN models for project cost and duration estimation. Hyari *et al.* (2015) presented an ANN model for the conceptual cost estimation of engineering services for public construction projects involving both design and construction supervision costs.

Adel *et al.* (2016) presented a parametric model for the conceptual cost estimation of highway construction projects using a supervised neural network. Mensah *et al.* (2016) conducted a study to develop an ANN model for determining the duration of rural bituminous surfaced road projects. Conceptual cost estimation models using ANN and MLR approaches have been developed for the Montana Department of Transportation (MDT) (Gardner *et al.*, 2016). An ANN was employed as the core computation engine of the tool for predicting the duration of construction major activities in tropical countries (Golizadeh *et al.*, 2016). Thaseena and Vishnu (2017) also developed a probabilistic cost overrun analysis model in construction projects as a decision support tool for contractors before the bidding stage using MLR and ANN. The cost contingency of the owner is one of the most important cost components and its accurate estimation is crucial to the financial performance of the project and to ensure the best use of the owner's funds. Peško *et al.* (2017) used artificial intelligence techniques: ANN and SVM to estimate the cost and duration of urban road construction projects. Golizadeh *et al.* (2017) developed an automated method to estimate the duration or completion time of construction of dam projects using ANN. Al-Zwainy *et al.* (2017) estimated the cost of highway construction projects using the ANN model. Magdum and Adamuthe (2018) developed an ANN and MLP-based construction cost forecasting model.

2.4 Gradient boosting trees

Gradient boosted trees (GBTs) is built on one of the most powerful concepts introduced in statistical modeling, called “boosting,” which consists of combining multiple “weak” models into a more realistic “aggregated model” (Loyer *et al.*, 2016). Boosting is an ensemble learning algorithm designed to increase the predictive performance of regression or classification procedures, such as decision trees (Ogutu *et al.*, 2011; Shin, 2015). GBT has been applied in the real world for various purposes as part of recent machine learning algorithms, including estimation or prediction and the existing literature has revealed its effectiveness to solve complex problems. To support this, (Laradji *et al.*, 2015) applied a gradient boosting machine learning algorithm for software defect prediction.

Gradient boosting technique was effectively applied to the early stage software effort estimation (Rath *et al.*, 2016; Satapathy, 2016). A GBT hybrid with a lasso regression model was used to predict individual house prices (Lu *et al.*, 2017). In the study conducted by Robinson *et al.* (2017), the GBT model performs better in estimating commercial building energy consumption compared to MLP and SVM. Similarly, Deng *et al.* (2018) used gradient boosting to estimate Energy Use Intensity (EUI) for US commercial office buildings and individual energy end-uses for heat ventilation and air conditioning, plug loads and lighting. Torres *et al.* (2019) used regression tree ensembles, including GBT for wind energy and solar radiation prediction. Shin (2015) investigated the applicability of a gradient boosted regression trees to a regression problem within the construction domain. This author applied GBT to the cost estimates at the early stage of the construction project. In another study, GBT was combined with an extreme gradient boost and random forest to predict the unit price bids for highway construction projects (Cao *et al.*, 2018).

2.5 Stacking ensemble algorithm

The most interesting idea when it comes to designing a new high-level machine-learning algorithm is the implementation of stacking ensemble (simply stacking) in the construction industry (Wolpert (1992). The main purpose of stacking ensemble is to reduce the generalization error and overfitting. Stacking ensemble supports classification and regression. It is a linear combination of multiple base learning algorithms into a single, superior prediction function through a secondary learning process called meta-learning and it improves prediction accuracy and stability (Breiman, 1996: Kansara *et al.*, 2018; Neloy *et al.*, 2019). Stacking is a technique which is used to tackle an error of a model. It can also be explained that stacking ensemble ML is a procedure for ensemble learning algorithms that involves training a second-level “meta-learner” to find the optimal combination of the base learners (LeDell, 2015). In general, stacking ensemble is used to build a strong model that takes into account predictions of other diverse and well-chosen modeling algorithms to generate the final result. Each model makes a major contribution and any algorithm’s weakness or bias is compensated by the strength of other algorithms, thereby enhancing the overall accuracy of the forecast (Kansara *et al.*, 2018).

Ensemble learning methods, a combination of multiple model algorithms, have been widely applied to regression and classification problems in various disciplines, such as demand forecasting model for supply chain (Kilimci *et al.*, 2019); financial market prediction (Henrique *et al.*, 2019); energy load prediction in residential buildings (Al-Rakhami *et al.*, 2019); electricity load and price forecasting (Do, 2018; Agrawal *et al.*, 2019; Zahid *et al.*, 2019); house price prediction model in real estate market (Kansara *et al.*, 2018; Neloy *et al.*, 2019); prediction of computer Go player attributes (Moudrík and Neruda, 2015); warfarin dose estimation in health sector Zhiyuan *et al.*, 2018); and software effort estimation (Banimustafa, 2018).

There is a lack of application of an ensemble-based ML algorithm in the field of construction estimation in the highway construction market, based on the analysis of the literature referred to above. In addition, because of its superior prediction efficiency, the authors of this study believed that an automated stacking ensemble approach could be a next-generation project estimation tool in the construction sector. In this review, an automated stacking ensemble model algorithm is therefore proposed to combine LR, SVM and ANN model algorithms linearly and optimally to estimate the cost of highway construction projects and to add to the body of knowledge by filling the above-mentioned existing research gap.

3. Proposed automated stacking ensemble model

This study proposes a two-level stacking ensemble model by automatically, linearly and optimally combining three learning algorithms: LR, SVM and ANNs via meta-learner, i.e. GB, to make estimates for the costs of highway construction projects. The authors are believed that the biggest gains would be realized when a dissimilar set of predictors are stacked together. It means that the more similar the predictors, the less benefit there is in stacking. Hence, the three distinct base-learning algorithms such as LR, SVM and ANN algorithms are proposed to be stacked in this study. The main drive of applying a stacking ensemble is to combine the aforementioned algorithms where they perform the best. For instance, LR is most appropriate to deal with linear data, whereas both SVM and ANN algorithms can perform better in the case of non-linear data. The heterogeneous ensemble of these dissimilar algorithms would result in good prediction performance.

Though the three models are capable of dealing with categorical and numerical variables in real-world classification or regression problems, these are now combined to reduce individuals' limitations through stacking ensemble machine learning algorithm. The proposed stacking ensemble regression model structure is shown in Figure 1 and the first phases of the implementation of this new model are explained in the subsequent sections.

3.1 First-level prediction algorithms and hyperparameter tuning

Three learning algorithms are selected to make the first-level predictions: LR, SVM and ANN algorithms. The overall process of developing the individual prediction model at the first level are described as follows.

Before making the individual prediction, the initial models are fitted first and the optimal hyperparameter values are determined using cross-validation (manual tuning process). Since

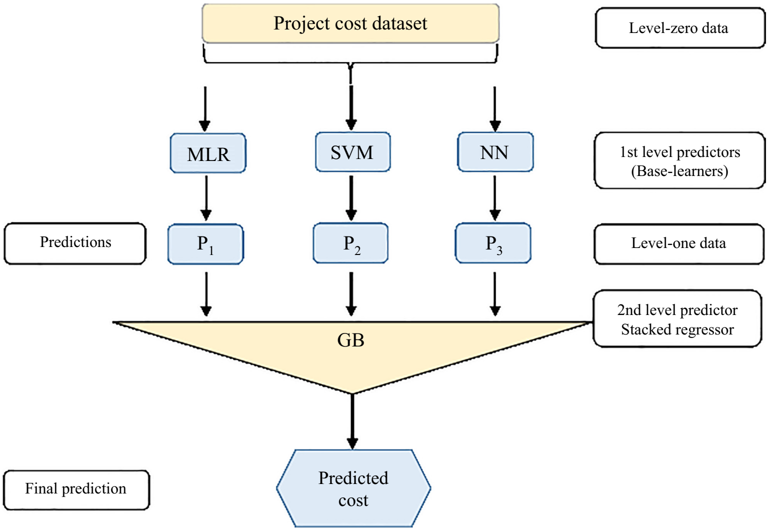

Figure 1.
Proposed stacking ensemble model structure

the hyperparameter values must be set before the model training process initiates. The stacking ensemble regressor works by opting the first-level prediction algorithms that fit into the training data set, also called level-zero data, which provides a list of outputs. In this study, cross-validation techniques are used to estimate and evaluate the predictive performance of models in the available input subsets. This technique is effectual to avoid over-fitting if any. A *k*-fold (often 5-fold or 10-fold) cross-validation technique is employed in this study as it is a well-known technique for cross-validation to optimize the model hyperparameters and to evaluate the performance of any model. It also enables to judge how the models perform outside the sample. The input training dataset is divided into *k* folds or subsets in this process. Each model is trained on *k-1* folds and validates the results on the fold that is not used in training. This process is repeated *k* times, by choosing a different fold every time for validation. Then store these predictions in train-meta to be utilized as inputs or features for the stacking model. In particular, 5-fold cross-validation is employed in this study to train each modeling algorithm.

3.2 Second-level prediction algorithm

Each first-level model provides predictive outcomes that are then fed into second-level training data (also known as meta-features). Simply put, the first-level model results will become features (inputs) of this second-level data. A second-level model or stacking model can then be trained on this data using a meta-regressor to generate the final results that will be used for predictions. In this case, the GB algorithm is used as a meta-learner. Python general-purpose programming language based on Scikit-learn library with JetBrains PyCharm Professional Edition and Anaconda plugin 2019.1.3 × 64 is used to develop both first-level and stacking model to predict project costs. The enhanced support of Python for libraries such as Scikit-learn and Pandas has recently made it a common option for data analysis activities. Python programming is an excellent option as a primary language for data mining and data analysis, as it is a simple and efficient tool that is accessible to all and reusable in a variety of contexts (McKinney, 2018). The main reason for choosing Python is its philosophy of design. It accentuates readability of code, and its syntax enables programmers to express their concepts in fewer lines of code (Tatiya, 2016).

3.3 Model performance evaluation metrics

The metrics used to evaluate the proposed stacking ensemble prediction model in this study are statistical measures. Performance evaluation metrics are employed to assess the suitability of the model to fit the data. The individual models and proposed stacking ensemble model are then evaluated and their predictive accuracy is compared based on some metrics: R^2, Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE). The R^2, MAE, MSE and RMSE formulae are given in Eqns 1–4.

R^2 is a metric that defines the accuracy of the regression line produced and is computed using Eqn 1. It measures the deviation of all results from the fitted regression line and is directly proportional to the performance. The higher the values, the better the performance of the model (Kansara *et al.*, 2018)

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}} = 1 - \frac{\sum_{i=1}^{n}(e_i)^2}{\sum_{i=1}^{n}(e_i')^2} \tag{1}$$

MAE is the average of the absolute errors between the actual and the predicted values as shown in Eqn 2 (Satapathy, 2016).

$$\text{MAE} = \frac{\sum_{i=1}^{n}|e_i|}{n} \tag{2}$$

The RMSE is computed as the square root of MSE. MSE is determined by finding the mean of the square of the difference between the actual and the predicted values. The formulas for calculating MSE and RMSE are given in Eqns 3 and 4.

$$\text{MSE} = \frac{\sum_{i=1}^{n} (e_i)^2}{n} \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (e_i)^2}{n}} \tag{4}$$

In the above equations, n is the number of projects in the dataset, e_i is the error derived from the difference between the actual value and predicted value and e'_i is the error derived from the difference between the actual value and mean value of the actual values.

4. Case illustrations

4.1 Variable selection

A set of four factors or input variables, including the number of bridges, inflation rate, terrain type and project type, was considered and used for the implementation of the newly proposed model development. Such factors were identified as the most significant factors affecting the accuracy of the cost estimate of Ethiopian highway construction projects (Meharie *et al.*, 2019). They also presented and proposed the appropriateness of these factors to make highway cost predictions. In their study, a rational and systematic approach to input variable selection for the preliminary estimate of project costs for the Ethiopian highway construction sector was presented using integrated factor analysis and fuzzy AHP methodology. Using this approach, the above-mentioned four factors were finally determined as the most significant factors and possible input variables in the cost estimation of highway construction projects. Accordingly, these four factors are considered to illustrate the implementation of the proposed prediction model in this study.

The authors of this study believed that this set of factors is appropriate for predicting the costs of the Ethiopian highway project, given the factors identified in the case of the Ethiopian highway construction sector. Besides, the frequency of factors and their impact on project costs studied by several researchers in different project cases are taken into account (Choon *et al.*, 2016; Elbeltagi *et al.*, 2014; Gransberg *et al.*, 2017; Mahamid, 2013; Mahalakshmi and Rajasekaran, 2019; Zhu *et al.*, 2016). The availability of complete variable information from actual highway project records in the office of Ethiopian Road Authority (ERA) was also checked. The model variables and their levels and values to be utilized in the preparation of prediction models in this study are described in Table 1.

4.2 Project database compilation

After an intensive project data gathering, the historical data base for 117 road projects was formulated and compiled based on the identified variables. The source of this data base was ERA's project management system software and it is believed to be reliable as the software is active and continuously managed by senior project managers at federal level. Federal road

Model variables	Levels or values
Number of bridges	Numerical count
Inflation rate	The rate in per cent
Terrain type	Flat, rolling, hilly, mountainous, escarpment
Project type	Gravel, asphalt, SST, DST, TST

Table 1. Set of model variables and their levels

projects, started and completed in Ethiopia between 1 January 2006 and 30 December 2018, were acquired for the compilation of project cost data set. While 117 projects were initially acquired to form a project database, a reduced cost dataset consisting of 108 of the 117 road projects was considered for the cost model of the project as the information on nine projects was not properly recorded in the ERA's project management system software. The cost data set of the project was then divided into training sets consisting of 83 (80%) road projects and test data set consisting of 25 (20%) projects. The historical cost data set used for cost model prediction and the summary statistical results of the cost data set is shown in Table 2.

4.3 Development of project cost prediction models

The model was built in two stages. In the first-level model training, LR, SVM, and ANN models were developed individually using the Scikit-learn in Python. Although the individual model algorithm is capable of regression, there can be no inference here as to how these model algorithms forecast future project costs. Because the main objective of this study is to apply a stacking ensemble model to the cost estimation of construction projects. In the second-level forecast, a GB algorithm was selected to combine the three basic-learning algorithms linearly and optimally and to generate final prediction results. Because GB is robust to outliers, missing data and numerous correlated and irrelevant variables compared to most model algorithms. The expected results from the LR, SVM, and ANN (base-learner) models were used as inputs (called meta-features) to create a stacking ensemble model. The primary challenge in implementing various machine learning algorithms is to determine the best tuning hyperparameter values for each algorithm to achieve the highest outputs. The optimal hyperparameter values for all algorithms were determined by fivefold cross-validation by trying a distinct value. The optimal values of hyperparameters tuning for all algorithms are elucidated in Table 3, which resulted in the smallest error.

Results of the first-level cost models: Base-learning algorithms. Table 4 provides the assessment metrics for the three base-learners or first-level models using the training and testing data set. In this study, 5-fold cross-validation was used to train base-learners and a stacking regression model. To analyze the predictive performance of the individual base-learners, the test results of the models are considered. Consequently, the comparison of the test results of the first-level models and the actual cost values extracted from each of the base-learner: LR, SVM and ANN are illustrated in Figures 2–4 respectively.

According to Table 4, ANN enables professionals to achieve the most reliable prediction performance compared to LR and SVM using the test data set, with an R^2 value of 0.94. Conversely, SVM produces the worst outcome compared to the other two model algorithms based on MSE and RMSE, but the LR model results in a relatively poor forecast accuracy based on MAE. The R^2 value of LR is very small relative to others, indicating that the variables are not significantly linearly related to the cost value of the project, and thus LR is not an effective modeling tool for this cost data set. RMSE amplifies and harshly punishes significant errors compared to other error metrics such as MAE and MSE. The

Features	Mean	SD	Min	25%	50%	75%	Max
Number of bridges	5.13	6.53	0.00	0.00	3.00	7.00	31.00
Inflation rate	14.87	4.70	8.23	9.74	15.37	18.20	25.25
Terrain type	1.78	0.57	1.00	1.45	1.64	2.00	4.00
Project type	2.17	1.07	1.00	1.00	2.00	2.06	4.00
Target model variable							
Project cost	5.29E+08	4.17E+08	7.01E+05	1.58E+08	4.89E+08	8.11E+08	1.66E+09

Table 2.
Model variables and statistical summary of cost dataset

Hyperparameters Base learning algorithms/Regressors – First -level predictions	
LR model	Model_LR = LinearRegression ( )
SVM ® model	Model_SVM = SVR (kernel = *"linear"*, degree = 3, gamma = *"auto"*, coef0 = 0.0, tol = 0.001, C = 1.0, epsilon = 0.1, shrinking = *True*, cache_size = 200, verbose = *False*, max_iter = –1)
MLP-ANN model	Model_NN = MLPRegressor (hidden_layer_sizes = (10,10,10), activation = *"relu"*, solver = *"Adam"*, alpha = 1e-3, batch_size = *"auto"*, learning_rate = *"constant"*, learning_rate_init = 0.001, power_*t* = 0.5, max_iter = 2,500, shuffle = *True*, random_state = 0, tol = 0.0001, verbose = *False*, warm_start = *False*, momentum = 0.9, nesterovs_momentum = *True*, early_stopping = *False*, validation_fraction = 0.1, beta_1 = 0.9, beta_2 = 0.999, epsilon = 1e-08)
Meta-learning algorithm/Meta-regressor	
GB model	Model_GB = GradientBoostingRegressor (n_estimators = 1,000, learning_rate = 0.01, min_samples_leaf = 3, min_samples_split = 3, loss = *"ls"*)
The newly proposed model algorithm	
Stacking regression model	StackingRegressor (regressors = [Model_LR, Model_SVM, Model_NN], meta_regressor = Model_GB)

Table 3. Tuning hyperparameters for the model algorithms

	LR		SVM		ANN	
	Training	Testing	Training	Testing	Training	Testing
R-Square	0.013	0.001	0.089	0.006	0.451	0.936
MAE	1.260	1.456	1.131	1.379	0.463	0.153
MSE	2.671	2.725	3.386	3.094	0.353	0.052
RMSE	1.634	1.651	1.840	1.759	0.594	0.228

Table 4. Performance matrices for the first-level cost models

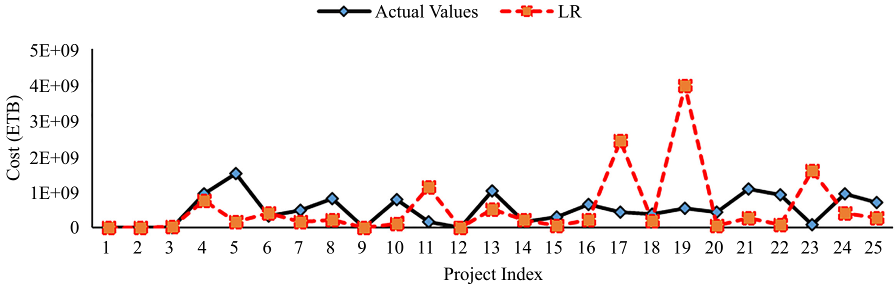

Figure 2. Comparison of testing results and actual cost values: LR base-learner

log-transformed values of the RMSE driving from the LR, SVM, ANN model calculations are 1.634, 1.840 and 0.594 respectively for the training set and 1.651, 1.759 and 0.228 for the test set.

Results of the second-level cost model. A stacking ensemble model. The findings indicate that the cost values provided by the stacking ensemble model were very close to those of the actual counterparts, as shown in Figures 5 and 6. Researchers or professionals can look at the model in a position to closely track the changing trend of the actual cost value of the project. The R^2 values for the proposed stacked ensemble model were 0.94 and 0.99 for the training and testing set, respectively. Such R^2 values are significantly higher than the values of the individual first-level model. The R^2 result shows that the input variables used are substantially linearly related to the cost value. According to the model performance measures tabulated in Table 5, the stacking ensemble model estimated the project cost with RMSE of 0.181 and 0.215 for training and testing set, respectively. This implies that the difference between predicted and actual project cost was quite small. MAE reveals that the estimated cost deviated by an average of 32,775,361.94 ETB from the average actual cost of the project, i.e. 515,817,547.7 ETB. Figure 6 indicates that the predicted cost estimates of the project were very similar to the actual values. This supports the potential of the stacking ensemble model to replicate the real cost values of the project with great accuracy.

According to Table 4, there is a substantial difference between the performance of individual base-algorithms, in favor of the testing results, on the training and testing data.

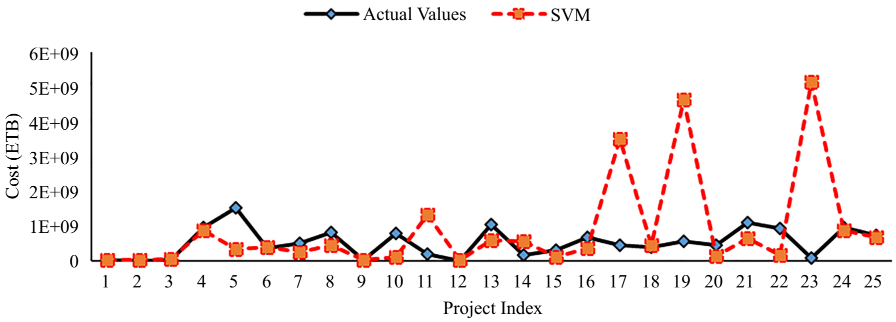

Figure 3.
Comparison of testing results and actual cost values: SVM base-learner

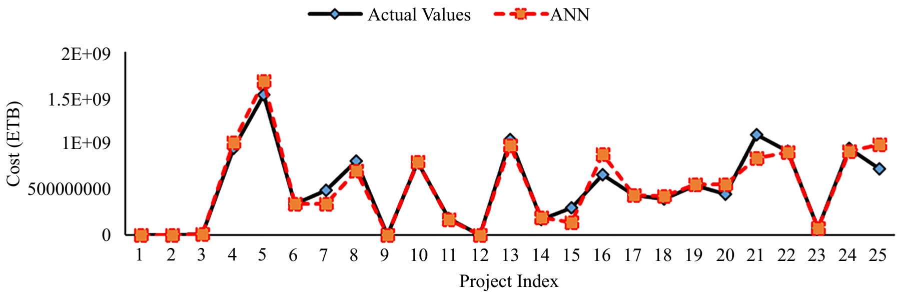

Figure 4.
Comparison of testing results and actual cost values: ANN base-learner

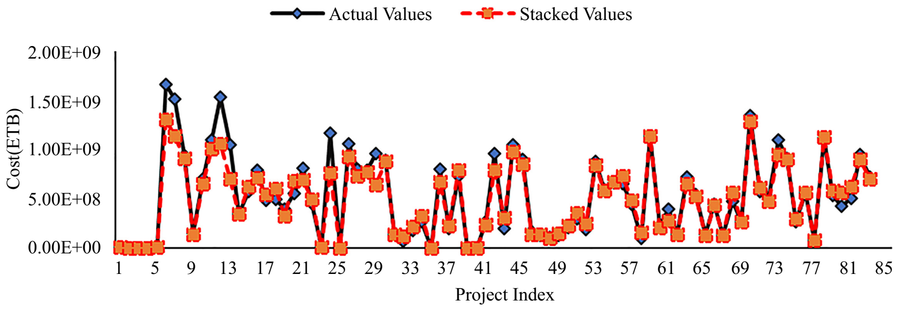

Figure 5.
Comparison of training results and actual cost values: stacking ensemble model

Though significant difference in favor of testing results is not usual in ML applications, this may happen due to ways to split training and testing data for model development and it is sometimes advisable to split the data set again.

The relative contribution of cost input variables: Sensitivity. The input variables or features contain information about the output of the target. Input variables may have a different contribution to the predicted values. Figure 7 indicates the relative contribution of the cost input variables to the estimated project cost.

As a result, the inflation rate was strongly relevant to the cost prediction with an average contribution value of 45%. Conversely, the type of terrain had a relatively lower contribution to the prediction, with a contribution value of 9%. It should be noted that the remaining two variables, the type of project and the number of bridges, had a modest contribution to the estimated cost as shown in Figure 7.

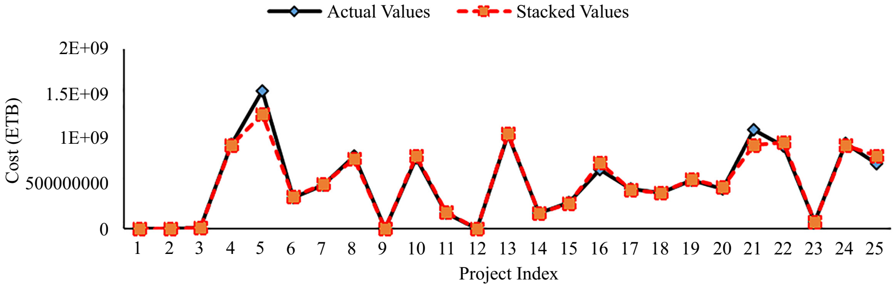

Figure 6. Comparison of testing results and actual cost values: stacking ensemble model

	Stacked regression model	
	Training	Testing
R-Square	0.938	0.978
MAE	0.131	0.111
MSE	0.033	0.046
RMSE	0.181	0.215

Table 5. Performance measures for the stacking ensemble cost model

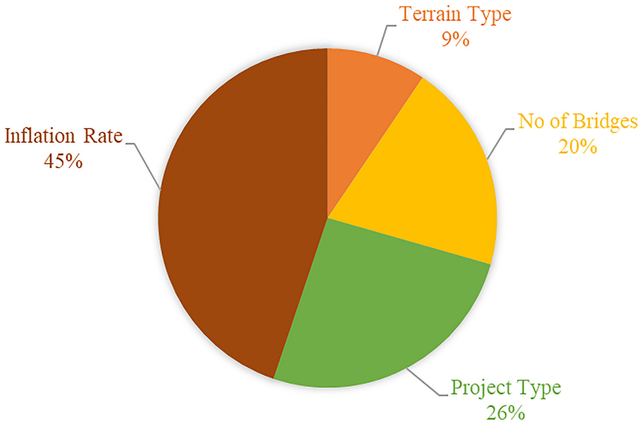

Figure 7. Average variable's importance or contribution to the cost model output

4.4 Discussion of findings

To specifically show the superior prediction efficiency of the proposed stacking ensemble model, the outcomes were compared with the base-learners (first-level models). Accordingly, the results show that the new model, in the defined cost data set, outperforms the three base-learners in all metrics. Specifically, based on the RMSE values, the stacking ensemble cost model delivers 86.8, 87.8 and 5.6% more accurate result than LR, SVM and ANN cost models respectively. From the statistical findings, it can be generalized that none of the individual learning algorithms provides a better forecast of highway project costs than stacking ensemble model. To support this, Breiman (1996) described that stacking ensemble ML approach never does worse than picking the single best predictor.

This superior output of stacking ensemble algorithm is in agreement with many previous studies which developed various ensemble machine learning algorithms for real-world regression (Pospieszny *et al.*, 2017; Cao *et al.*, 2018; Kansara *et al.*, 2018; Yang and Cao, 2018; Agrawal *et al.*, 2019; Al-Rakhami *et al.*, 2019; Kilimci *et al.*, 2019; Neloy *et al.*, 2019). However, the results of the proposed stacking ensemble model can hardly be compared with the results of ensemble learning algorithms studied by the aforementioned authors. This is because the present study considered some specification of factors different from previous studies, such as cost data set, pool of input variables, form and number of base-learners, model tuning hyperparameters and application domain (Ma *et al.*, 2018). The high contribution of the inflation rate to the predicted cost is strongly supported by the Project Management Body of Knowledge Guide (PMBOK, 2011). According to this guide book, the estimator or practitioners should ensure that economic changes, such as inflation, are appropriately and credibly reflected in the life cycle estimate. Shane *et al.* (2009) also mentioned inflation rate as one of the main construction cost escalation factors.

5. Conclusions

Early understanding of the cost of construction projects is crucial to sound decision-making at the planning and design stage of the project. Generally speaking, all professionals and/or contracting parties involved in the construction of infrastructure need more accurate and coherent data on the completion time and costs of the project prior to implementation. In other words, in order to be able to approve the project at the earliest stage of the project, the stakeholders in the construction sector must have an indication of the cost of the project. The aim of this study was therefore to construct a stacking ensemble prediction model for cost estimation of highway construction projects, using an optimal combination of three model algorithms, such as LR, SVM and ANN.

Accordingly, the main findings revealed that the proposed stacking ensemble model outperforms the three models in all metrics in predicting project costs with a given set of datasets. The newly proposed stacking ensemble cost model, using the given ERA's cost data set, yields 86.8, 87.8 and 5.6% more accurate result than the LR, SVM and ANN models respectively. It can also be noted that the prediction made by the stacking ensemble model shows a high degree of coherence with the recorded cost of highway construction projects. In doing so, the inflation rate played a significant role in the outputs of the cost model with an average percentage contribution significance factor of 45. With very few known early factors or parameters, the new model package can be used to estimate the future cost of highway construction projects in the preliminary phase. In addition, the proposed prediction model may also apply to various types of construction industry data on a large scale, including complex and non-linear data and data with missing values and outliers. With its improved predictive performance, the implementation of the newly proposed model package provides enormous benefits for the Ethiopian Road Authority and other contracting parties in the preparation of more accurate initial budgets, allocation of resources and project cost estimates in the case of the highway construction industry. Even though the proposed model can only applicable at the preliminary stage of highway projects, it opens the door to develop similar models that can be employed for the various phases of the project throughout its lifetime. Finally, further research works are recommended to develop more realistic and accurate cost estimation models in the highway construction sector by examining a different set of significant input variables while minimizing the use of categorical variables to improve the performance of some base-algorithms.

References

Abdal-hadi, M.A. (2014), “Factors affecting the accuracy of pre-tender cost estimate: studies of Saudi Arabia”, *International Journal of Applied Engineering Research*, Vol. 9 No. 1, pp. 21-36.

Adel, K., Elyamany, A., Belal, A.M. and Kotb, A.S. (2016), “Developing a parametric model for conceptual cost estimate of highway projects”, *International Journal of Engineering Science and Computing*, Vol. 6 No. 7, pp. 1728-1734.

Agrawal, R.K., Muchahary, F. and Tripathi, M.M. (2019), “Ensemble of relevance vector machines and boosted trees for electricity price forecasting”, *Applied Energy*, Vol. 250 April, pp. 540-548, doi: 10.1016/j.apenergy.2019.05.062.

Al-Rakhami, M., Gumaei, A., Alsanad, A., Alamri, A. and Hassan, M.M. (2019), “An ensemble learning approach for accurate energy load prediction in residential buildings”, *IEEE Access*, Vol. 7 No. c, pp. 48328-48338, doi: 10.1109/ACCESS.2019.2909470.

Al-Zwainy, F.M.S., Abd, I. and Aidan, A. (2017), “Forecasting the cost of structure of infrastructure projects utilizing artificial neural network model (highway projects as case study)”, *Indian Journal of Science and Technology*, Vol. 10 No. 20, pp. 974-6846, doi: 10.17485/ijst/2017/v10i20/108567.

Alemayehu, S. (2014), *Testing Regression Models to Estimate Costs of Road Construction Projects*, MSc thesis, Addis Ababa University, available at: http://localhost:80/xmlui/handle/123456789/6252.

Alumbugu, P.O., Saidu, W.A.O., Muhammed, M. and Abdulazeez, A. (2014), “Assessment of the factors affecting the accuracy of pre-tender cost estimate in Kaduna state, Nigeria”, *IOSR Journal of Environmental Science, Toxicology and Food Technology*, Vol. 8 No. 5, pp. 19-27.

Banimustafa, A. (2018), “Predicting software effort estimation using machine learning techniques”, *2018 8th International Conference on Computer Science and Information Technology*, pp. 249-256, doi: 10.1109/CSIT.2018.8486222.

Barakchi, M., Torp, O. and Moges, A. (2017), “Cost estimation methods for transport infrastructure : a systematic literature review”, *Procedia Engineering*, Vol. 196 June, pp. 270-277, doi: 10.1016/j.proeng.2017.07.199.

Barraza, M.F.S., Smith, T., Mi, S. and Park, D. (2017), “Analysis of schedule deviations in road construction projects and the effects of project physical characteristics”, *Journal of Financial Management of Property and Construction*, Vol. 22 No. 2, pp. 192-210, doi: 10.1108/JEIM-07-2014-0077.

Breiman, L.E.O. (1996), “Stacking regressions”, *Machine Learning*, Vol. 64, pp. 49-64.

Cao, Y., Ashuri, B. and Baek, M. (2018), “Prediction of unit price bids of resurfacing highway projects through ensemble machine learning”, *Journal of Computing in Civil Engineering*, Vol. 32 No. 5, pp. 1-10, doi: 10.1061/(asce)cp.1943-5487.0000788.

Čeh, M., Kilibarda, M., Lisec, A. and Bajat, B. (2018), “Estimating the performance of random forest versus multiple regression for predicting prices of the apartments”, *ISPRS International Journal of Geo-Information*, Vol. 7 No. 5, p. 168, doi: 10.3390/ijgi7050168.

Choon, T.T., Sim, L.C., Connie, T., Nita, A.K., Uche, A.G. and Chen, G.K. (2016), “Influential factors in estimating and tendering for construction work”, *MATEC Web of Conferences*, EDP Sciences, Vol. 47, p. 4007.

Deng, H., Fannon, D. and Eckelman, M.J. (2018), “Predictive modelling for US commercial building energy use: a comparison of existing statistical and machine learning algorithms using CBECS microdata”, *Energy and Buildings*, Vol. 163, pp. 34-43, doi: 10.1016/j.enbuild.2017.12.031.

Do, H.T. (2018), “Data-driven modeling for improved residential building electricity consumption prediction and HVAC efficiency evaluation”, Graduate Theses and Dissertations, 16805, available at: https://lib.dr.iastate.edu/etd/16805.

El-Kholy, A.M. (2015), “Predicting cost overrun in construction projects”, *International Journal of Construction Engineering and Management*, Vol. 4 No. 4, pp. 95-105, doi: 10.5923/j.ijcem.20150404.01.

El-sawalhi, N.I. (2015), “Support vector machine cost estimation model for road projects”, *Journal of Civil Engineering and Architecture*, Vol. 9, pp. 1115-1125, doi: 10.17265/1934-7359/2015.09.012.

Elbeltagi, E., Hosny, O., Abdel-razek, R. and El-fitory, A. (2014), “Conceptual cost estimate of Libyan highway projects using artificial neural network”, *International Journal of Engineering Research and Applications*, Vol. 4 No. 8, pp. 56-66.

Enshassi, A., Mohamed, S. and Abdel-Hadi, M. (2013), “Factors affecting the accuracy of pre-tender cost estimates in the Gaza Strip”, *Journal of Construction in Developing Countries*, Vol. 18 No. 1, pp. 73-94.

Gardner, B.J., Gransberg, D.D. and Jeong, H.D. (2016), “Reducing data-collection efforts for conceptual cost estimating at a highway Agency”, *Journal of Construction Engineering and Management*, Vol. 142 No. 11, 04016057, doi: 10.1061/(ASCE)CO.1943-7862.0001174.

Golizadeh, H., Sadeghifam, A.N., Aadal, H. and Majid, M.Z.A. (2016), “Automated tool for predicting the duration of construction activities in tropical countries”, *KSCE Journal of Civil Engineering*, Vol. 20 No. 1, pp. 12-22, doi: 10.1007/s12205-015-0263-x.

Golizadeh, H., Banihashemi, S., Sadeghifam, A.N. and Preece, C. (2017), “Automated estimation of completion time for dam projects”, *International Journal of Construction Management*, Vol. 17 No. 3, pp. 197-209, doi: 10.1080/15623599.2016.1192249.

Gransberg, D., Jeong, H.D., Karaca, I. and Gardner, B. (2017), *Top-Down Construction Cost Estimating Model Using an Artificial Neural Network*, State of Montana Department of Transportation, Finance Reports, p. 1, available at: https://lib.dr.iastate.edu/finance_reports/1.

Hatamleh, M.T., Hiyassat, M., Sweis, G.J. and Sweis, R.J. (2018), “Factors affecting the accuracy of cost estimate: a case of Jordan”, *Engineering Construction and Architectural Management*, Vol. 25 No. 1, pp. 113-131, doi: 10.1108/ECAM-10-2016-0232.

Henrique, B.M., Sobreiro, V.A. and Kimura, H. (2019), “Literature review: machine learning techniques applied to financial market prediction”, *Expert Systems with Applications*, Vol. 124, pp. 226-251, doi: 10.1016/j.eswa.2019.01.012.

Huang, J., Li, Y. and Xie, M. (2015), “An empirical analysis of data preprocessing for machine learning-based software cost estimation”, *Information and Software Technology*, Vol. 67, pp. 108-127, doi: 10.1016/j.infsof.2015.07.004.

Hyari, K.H., Al-daraiseh, A. and El-mashaleh, M. (2015), “Conceptual cost estimation model for engineering services in public construction projects”, *Journal of Management in Engineering*, Vol. 32 No. 1, pp. 1-9, doi: 10.1061/(ASCE)ME.1943-5479.0000381.

Jain, M. and Pathak, K.K. (2014), “Applications of artificial neural network in construction engineering and management - a review”, *International Journal of Engineering Technology, Management and Applied Sciences*, Vol. 2 No. 3, pp. 134-142.

Jarkas, A.M., Mubarak, S.A. and Kadri, C.Y. (2013), “Critical factors determining bid/no bid decisions of contractors in Qatar”, *Journal of Management in Engineering*, Vol. 30 No. 4, 5014007.

Kaleem, S., Irfan, M. and Gabriel, H.F. (2014), “Estimation of highway project duration at the planning stage and analysis of risk factors leading to time overrun”, *Conference of the Transportation and Development Institute (T&DI) of ASCE*, pp. 612-626.

Kang, T.K., Park, W. and Lee, Y.S. (2011), "Development of CBR-based road construction project cost estimation system", *Conference: 28th International Symposium on Automation and Robotics in Construction*, pp. 1314-1319.

Kansara, D., Singh, R., Sanghvi, D. and Kanani, P. (2018), "Improving accuracy of real estate valuation using stacking regression", *International Journal of Engineering Development and Research*, Vol. 6 No. 3, pp. 571-577.

Kilimci, Z.H., Akyuz, A.O., Uysal, M., Akyokus, S., Uysal, M.O., Bulbul, B.A. and Ekmis, M.A. (2019), "An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain", *Complexity*, Vol. 2019, 9067367, p. 15, doi: 10.1155/2019/9067367.

Kulkarni, P.., Londhe, S.. and Deo, M. (2017), "Artificial neural networks Management : a review for construction", *Journal of Soft Computing in Civil Engineering*, Vol. 2, pp. 70-88.

Laradji, I.H., Alshayeb, M. and Ghouti, L. (2015), "Software defect prediction using ensemble learning on selected features", *Information and Software Technology*, Vol. 58, pp. 388-402, doi: 10.1016/j.infsof.2014.07.005.

LeDell, E. (2015), *Scalable Ensemble Learning and Computationally Efficient Variance Estimation*, Doctoral dissertation, UC Berkeley.

Loyer, J.L., Henriques, E., Fontul, M. and Wiseall, S. (2016), "Comparison of Machine Learning methods applied to the estimation of the manufacturing cost of jet engine components", *International Journal of Production Economics*, Vol. 178, pp. 109-119, doi: 10.1016/j.ijpe.2016.05.006.

Lu, S., Li, Z., Qin, Z., Yang, X. and Goh, R.S.M. (2017), "A hybrid regression technique for house prices prediction", *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 319-323, doi: 10.1109/IEEM.2017.8289904.

Ma, Z., Wang, P., Gao, Z., Wang, R. and Khalighi, K. (2018), "Ensemble of machine learning algorithms using the stacking generalization approach to estimate the warfarin dose", *PloS One*, Vol. 13 No. 10, pp. 1-12, doi: 10.1371/journal.pone.0205872.

Magdum, S.K. and Adamuthe, A.C. (2018), "Construction cost prediction using neural networks", *ICTACT Journal on Soft Computing*, Vol. 8 No. 1, pp. 1549-1556, doi: 10.21917/ijsc.2017.0216.

Mahalakshmi, G. and Rajasekaran, C. (2019), "Early cost estimation of highway projects in India using artificial neural network", *Sustainable Construction and Building Materials*, Vol. 25, pp. 659-672, doi: 10.1007/978-981-13-3317-0_59.

Mahamid, I. (2013), "Effects of the project's physical characteristics on cost deviation in road construction", *Journal of King Saud University - Engineering Sciences*, Vol. 25 No. 1, pp. 81-88, doi: 10.1016/j.jksues.2012.04.001.

Mahamid, I. (2015), "Factors affecting cost estimate accuracy: evidence from Palestinian construction projects", *International Journal of Management Science and Engineering Management*, Vol. 10 No. 2, pp. 117-125, doi: 10.1080/17509653.2014.925843.

Marinelli, M., Dimitriou, L., Fragkakis, N. and Lambropoulos, S. (2015), "Non-parametric bill-of-quantities estimation of concrete road bridges' superstructure: an artificial neural networks approach", *ARCOM 2015, 31st Annual Association of Researchers in Construction Management Conference*, available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84985910966&partnerID=40&md5=6c31c78566922655452d7d9758552ac5853862.

McKinney, W. (2018), in Beaugureau, M. (Ed.), *Python for Data Analysis: Data Wrangling with Pandas, Numpy, and IPython*, Second ed.,. doi: 10.1017/CBO9781107415324.004.

Meharie, M.G., Gariy, Z.C.A., Mutuku, N., Ngumbau, R. and Mengesha, W.J. (2019), "An effective approach to input variable selection for preliminary cost estimation of construction projects", *Advances in Civil Engineering*, Vol. 2019, pp. 1-14.

Mensah, I., Adjei-Kumi, T. and Nani, G. (2016), "Duration determination for rural roads using the principal component analysis and artificial neural network", *Engineering Construction and Architectural Management*, Vol. 23 No. 5, pp. 838-656, doi: 10.1108/ECAM-09-2015-0148.

Moudrík, J. and Neruda, R. (2015), "Evolving non-linear stacking ensembles for prediction of go player attributes", *2015 IEEE Symposium Series on Computational Intelligence*, IEEE, pp. 1673-1680.

Naik, M.G. and Radhika, V. (2015), "Time and cost analysis for highway road construction project using artificial neural networks", *Journal of Construction Engineering and Project Management*, Vol. 5 No. 1, pp. 26-31, doi: 10.6106/JCEPM.2015.5.1.026.

Neloy, A.A., Haque, H.M.S. and Ul Islam, M.M. (2019), "Ensemble learning based rental apartment price prediction model by categorical features factoring", *ICMLC'19*, Association for Computing Machinery, Zhuhai, pp. 350-356, doi: 10.1145/3318299.3318377.

Ogutu, J.O., Piepho, H.P. and Schulz-Streeck, T. (2011), "A comparison of random forests, boosting and support vector machines for genomic selection", *BMC Proceedings*, SUPPL. 3, Vol. 5, pp. 1-5, doi: 10.1186/1753-6561-5-S3-S11.

Peško, I., Mučenski, V., Šešlija, M., Radović, N., Vujkov, A., Bibić, D. and Krklješ, M. (2017), "Estimation of costs and durations of construction of urban roads using ANN and SVM", *Complexity*, Vol. 2017, 2450370, p. 13, doi: 10.1155/2017/2450370.

Petruseva, S., Zileska-pancovska, V., Vahida, Ž. and Vejzović, A.B. (2017), "Construction costs forecasting: comparison of the accuracy of linear regression and support vector machine models", *Tehnicki Vjesnik - Technical Gazette*, Vol. 24 No. 5, pp. 1-8, doi: 10.17559/tv-20150116001543.

A Guide to the Project Management Body of Knowledge (PMBOK Guide) (2011), *Practice Standard for Project Estimating*, Project Management Institute (PMI).

Pospieszny, P., Czarnacka-chrobot, B. and Kobylinski, A. (2017), "An effective approach for software project effort and duration estimation with machine learning algorithms", *Journal of Systems and Software*, Vol. 137 November, pp. 184-196, doi: 10.1016/j.jss.2017.11.066.

Rafiei, M.H. and Adeli, H. (2018), "Novel machine-learning model for estimating construction costs considering economic variables and indexes", *Journal of Construction Engineering and Management*, Vol. 144 No. 12, pp. 1-9, doi: 10.1061/(asce)co.1943-7862.0001570.

Rath, S.K., Acharya, B.P. and Satapathy, S.M. (2016), "Early-stage software effort estimation using random forest technique based on use case points", *IET Software*, Vol. 10 No. 1, pp. 10-17, doi: 10.1049/iet-sen.2014.0122.

Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M.A. and Pendyala, R.M. (2017), "Machine learning approaches for estimating commercial building energy consumption", *Applied Energy*, Vol. 208 May, pp. 889-904, doi: 10.1016/j.apenergy.2017.09.060.

Satapathy, S.M. (2016), *Effort Estimation Methods in Software Development Using Machine Learning Algorithms*, PhD thesis, National Institute of Technology Rourkela.

Shane, J.S., Molenaar, K.R., Anderson, S. and Schexnayder, C. (2009), "Construction project cost escalation factors", *Journal of Management in Engineering*, Vol. 25 No. 4, pp. 221-229.

Shin, Y. (2015), "Application of boosting regression trees to preliminary cost estimation in building construction projects", *Computational Intelligence and Neuroscience*, Vol. 2015, pp. 1-9, doi: 10.1155/2015/149702.

Tatiya, A. (2016), *Cost Prediction Model for Decomposition and Impact of Design for Decomposition*, MSc thesis, Michigan State University.

Thaseena, T. and Vishnu, K. (2017), "Analysis of cost overrun in highway construction projects using multiple regression and artificial neural networks", *International Research Journal of Engineering and Technology (IRJET)*, Vol. 4 No. 4, available at: https://www.irjet.net/archives/V4/i4/IRJET-V4I4587.pdf.

Thomas, N. and Thomas, A.V. (2016), "Regression modelling for prediction of construction cost and duration", *Applied Mechanics and Materials*, Vol. 857, pp. 195-199, doi: 10.4028/www.scientific.net/AMM.857.195.

Torres-Barrán, A., Alonso, Á. and Dorronsoro, J.R. (2019), "Regression tree ensembles for wind energy and solar radiation prediction", *Neurocomputing*, Vols 326–327, pp. 151-160, doi: 10.1016/j.neucom.2017.05.104.

Wolpert, D. (1992), "Stacking generalization (stacking)", *Neural Networks*, Vol. 5, pp. 241-259.

Wu, J.Y. (2017), *Housing Price Prediction Using Support Vector Regression*, available at: https://scholarworks.sjsu.edu/etd_projects/540/.

Yang, B. and Cao, B. (2018), "Ensemble learning based housing price prediction model", *FSDM 2018 : 4th International Conference on Fuzzy Systems and Data MiningAt*, Bangkok, Thailand.

Zahid, M., Ahmed, F., Javaid, N., Abbasi, R., Zainab Kazmi, H., Javaid, A. and Ilahi, M. (2019), "Electricity price and load forecasting using enhanced convolutional neural network and enhanced support vector regression in smart grids", *Electronics*, Vol. 8 No. 2, p. 122, doi: 10.3390/electronics8020122.

Zhai, D., Shan, Y., Sturgill, R.E., Taylor, T.R.B. and Goodrum, P.M. (2016), "Using parametric modeling to estimate highway construction contract time", *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2573, pp. 1-9, doi: 10.3141/2573-01.

Zhu, B., Yu, L. and Geng, Z. (2016), "Cost estimation method based on parallel Monte Carlo simulation and market investigation for the engineering construction project", *Cluster Computing*, Vol. 19, pp. 1293-1308, doi: 10.1007/s10586-016-0585-6.

Corresponding author
Meseret Getnet Meharie can be contacted at: mesget86@gmail.com