Cost Prediction for Roads Construction using Machine Learning Models

Original Scientific Paper

Yasamin Ghadbhan Abed
University of Diyala, College of Science, Department of Computer Science
Diyala, Iraq
scicompms2133@uodiyala.edu.iq

Taha Mohammed Hasan
University of Diyala, College of Science, Department of Computer Science
Diyala, Iraq
dr.tahamh@uodiyala.edu.iq

Raquim Nihad Zehawi
University of Diyala, College of Engineering, Department of Highway and Airport Engineering
Diyala, Iraq
raquim_zehawi@uodiyala.edu.iq

Abstract *– Predicting conceptual costs is among the essential criteria in project decision-making at the early stages of civil engineering disciplines. The cost estimation model availability that may help in the early stages of a project could be incredibly advantageous in respect of cost alternatives and more extraordinary cost-effective solutions periodically. There is a lack of case datasets. Most of the proposed dataset was inefficient. This study offers a new data set that includes the elements of road construction and economic advantages in the year of project construction. Real project data for rural roads in the State of Iraq / Diyala Governorate for the years 2012 to 2021 have use to train a predictive model with a high rate of accuracy based on machine learning (ML) methods. Ridge and Least Absolute Shrinkage and Selection Operator (LASSO) Regressions, K Nearest Neighbors (k-NN), and Random Forest (RF) algorithms have employ to create models for estimating road construction costs based on real-world data. The Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared (R2) coefficient of determination are utilize to assess the models' performance. The analysis indicated that the RR is the best model for road construction costs, with results R2 = 1.0, MAPE =0.00, and RMSE=0.00. The results showed that the cost estimates were accurate and aligned with the project bids.*

Keywords*: Construction, Roads, Cost estimation, Machine learning, Ridge regression*

1. INTRODUCTION

Cost estimation is one of the essential concerns in the early phases of a construction project's life cycle. The cost of a building project is widely acknowledged as critical contract data, yet it is frequently miscalculated. Contracted prices can result in various issues, resulting in additional expenditures throughout the project's implementation. The primary and most common problem in many building projects is a cost overrun. [1]

If the inaccuracy of the initial cost estimates was merely due to inadequate information and inherent forecasting challenges, as those in charge of project estimates indicated, the inaccuracies might be expected to be random. [2]

Given the detrimental consequence on profitability and public funds, extra expenses are a key source of worry for both the private and public sectors. Since the contract is closed by a predetermined amount of the bid, road projects are expected to exceed the budget. The additional expenditures of projects have been detected as a result of deletions, errors, and contract changes. [3]

TThe necessity for a precise preliminary estimate has prompted study into constructing models relying on machine learning (ML) algorithms to predict the initial assessment of road, building, bridge, or other construction projects. [4]

2. LITERATURE SURVEY

The creation and use of methodologies for the cost estimation of road projects have been active research areas over the past few decades. Numerous studies on

estimating construction costs using neural networks, regression, or stochastic methods have been published in the last 20 years.[5]

The dataset is a crucial vector in developing a predictive model for cost estimation. Some authors used a small dataset to analyze the effect of variables on cost using ML techniques like [4]. Cost models were developed utilizing 50 sets of data gathered from road projects finished in South Western Nigeria between 2010 and 2015. They employed linear and multiple regression to anticipate the preliminary estimate of road projects for seven primary construction activities. Based on data collected from the Brazilian National Department of Transport Infrastructure (DNIT), fourteen highway projects in Brazil have been utilized to establish a more precise estimation technique for the construction of highway projects utilizing Artificial Neural Networks (ANNs) Barros, Marcy, and Carvalho. [6] The inputs have been the most impactful factors in the road project estimation costs, and the output was the actual cost value of the work. An average cost estimation accuracy of 99% was accomplished. Furthermore, Tijanić et al. [7] examined the performance of several types of artificial neural networks (multilayer perceptron MLP, generalized regression neural network (GRNN), GRNN, radial basis function neural network RBFNN) for estimating road construction costs in Croatia, using a dataset of 57 road sections. The GRNN had the greatest precision, with a Mean Absolute Percentage Error (MAPE) of 13% and a coefficient of determination of 0.95. According to Mahalakshmi and Rajasekaran [8], a multi perceptron network with a backpropagation algorithm is competent of accurately forecasting highway construction costs. The National Highway Authority of India (NHAI) provided a dataset of 52 projects. Subsequently, AL-Zwainy and Aidan[9] provided multi-layer perceptron training that utilized the backpropagation algorithm to predict construction costs of highways in Iraq. The dataset was 150 past highway data from the republic of Iraq, and it was not published. The ANNs model was able to forecast the cost of structural work for a highway project with a high degree of accuracy (93.19%) and a high coefficient of correlation (R) of 90.026%. Peško et al., [10] analyzed support vector machines (SVMs) and ANNs using SVM have shown higher precision when estimating costs, with MAPE of 7.06% contrast to the extremely accurate ANNs, which have attained a 25.38% precision. The dataset was 166 projects. Moreover, Suneja et al., [11] focused on developing a cost estimation model for Transportation Infrastructure Projects based on Reality by Neural Network to discover the connection between multiple variables of the project and their cost. The dataset was 124 road projects in Gujarat Region. In Poland, a number of completed bridge construction projects were collected by [12] to build an SVM-based regression model to predict bridge construction costs with accuracy appropriate for the early stage of projects. The model was capable of providing an early estimate with satisfactory accuracy reach to 0.98 for the correlation coefficient of real-life bridge construction costs, but the dataset was not clear. [13] used multiple regression techniques for develop early cost estimating models for road construction projects , based on 131 sets of data collected in the West Bank in Palestine. R2 for the developed models was varying from 0.92 to 0.98 which indicates that the predicted values from estimated models fit with the real-life data.

According to everything mentioned above, the esearchers focused on determining the most affected variable on estimation cost. Accuracy was not considered during the model design since the work is more analytical than artificial intelligence (AI). Different datasets have been used in every research, some very small, and all the datasets were not published to compare the results. This study presented and published a real dataset for road construction in building a predictive model for a very high accuracy cost based on ML techniques.

There is a continual need for competent computing methods in this application field due to economic and environmental constraints and their linkages in road construction development. In recent years, AI-based ML algorithms have been verified to be superior to traditional methods for making such forecasts in a variety of infrastructure development projects. Machine learning approaches aim to predict, explain, and discover correlations and patterns between variables [14].

The adverse effects of biased cost modeling in the construction industry are significant since such modeling may drastically reduce project costs by underestimating or overestimating costs. As a result, engineers and managers need this information to swiftly assess alternative project options' feasibility, performance, and profitability [5]. Therefore, a data configuration is proposed that contains a set of material and financial variables for actual road construction and economic variables affecting road construction costs [15]. Historical rural road construction projects in Diyala Governorate were chosen for estimating and modeling the total construction costs.

This study's contribution is to design a new and more realistic road construction cost estimating model that incorporates advanced ML concepts, economic data, and indices. The proposed approach compares four ML algorithms. Ridge and Least Absolute Shrinkage and Selection Operator (LASSO) Regressions, K Nearest Neighbors (k-NN), and Random Forest (RF) algorithms. The technique successfully assisted stakeholders in the early stages of a construction project who were responsible for estimating and managing construction costs to accomplish more precise findings from previous situations.

The remainder of the work is organized as follows: Section 2 describes the case study. Section 3 explains the modeling methodology. The experimental results of the approach and a commentary on the findings are presented in Section 4. Finally, Section 5 concludes the

report with closing remarks and recommendations for further research.

3. METHODS AND MATERIAL

Artificial intelligence (AI) predicts or calculates the cost of construction-based materials or construction datasets. The machine learning (ML) approach is a primary field concerning AI for predicting classes or targets with accurate results. ML was approached into the superior and un-superior methods. This section explains the theoretical concept concerning ML methods in subsection 2.1. Also, the materials or datasets described in subsection 2.2 are collected and proposed for training and testing ML algorithms based on evaluation metrics.

3.1 METHODS

ML algorithms allow for more complex cost prediction models. They learn from input variables and provide data-driven predictions on output variables instead of static prediction models resembling those used in time series analysis. In addition, explanatory variables (also referred to as features in the ML context) improve the capacity of a machine learning model to detect variance and deliver a more accurate prediction [15][16]. Four ML algorithms have been used and analyzed. Ridge and Least Absolute Shrinkage and Selection Operator (LASSO) Regressions, K Nearest Neighbors (k-NN), and Random Forest (RF) algorithms.

3.1.1 K Nearest Neighbors (k-NN) Regression

The computation of the k-NN is a well-known and valuable supervised learning method that employs the concept of similarity to predict a target output (for example, a class label) for a query object or sample. The k-NN method insinuates the intended output of new objects in the feature space of a training set depending on the outcomes of the nearest samples or the outcome of many nearest objects. [17] The k-NN regression is a technique for gradual learning based on occurrences. A nonparametric regression accelerates the training phase since it imposes no assumptions about data distribution. It learns complex target functions rapidly as well as without losing any data. K observations with x_i in close proximity are considered for a particular input x of training data, and the average of the responses of those K independent variables produces $y\hat{}$

$$y\hat{}(x) = \frac{1}{k}\sum_{x_i \in N_{k(x)}} y_i \quad (1)$$

where $N_{k(x)}$ illustrates K closest points in the neighborhood of x. Various distance metrics are used to determine how close two points are, but Euclidean distance is the most often used [17] [18].

3.1.2 Ridge Regression

During the 1970s and 1980s, a newly developed approach for calculating multiple linear regression coefficients called Ridge Regression (RR) was one of the most intriguing research subjects [19]. Ridge Regression is a well-known parameter estimation method for dealing with the collinearity issue that commonly occurs in multiple linear regression [20]. The RR is a tool for assessing multicollinearity data from multiple regression models. The RR is also crucial for analyzing multicollinearity in multiple regression data. Least-squares evaluations are impartial when multicollinearity occurs, but their modifications are greater. Thus, they may be far from their true value. By adding a bias grade to the regression evaluations, RR decreases standard errors. It is expected that, as a result, more consistent evaluations will be available. In addition, when the loss function is the linear least-squares function, and the data is regularized using the L2-norm, the RR model may be utilized to solve a regression problem. The strength of the regularization has to be a positive float. Regularization enhances the conditioning of the problem and lowers the estimated variance [21]. This strategy was initially presented to handle the multicollinearity problem by Hoerl and Kennard (1970) [22]. They proposed that a small positive number be added to the diagonal elements of the $X'X$ matrix, yielding the estimators Eq. 3:

$$\beta = (X^t X)^{-1} X^t Y \quad (2)$$

$$\beta\hat{} = (\ X\hat{}' X + kI\_p\ )\hat{}(-1)\ X\hat{}'\ y\ , k \geq 0 \quad (3)$$

This is referred to as a ridge regression estimator, and the constant k ($k \geq 0$) is referred to as a "biased" or " ridge" parameter that must be estimated with real data.

Algorithm (1): Ridge Regression Algorithm

Input: preprocessed data

Output: Road Cost

Begin

Step 1: Load the Data

Step 2: Creating a New Train and Validation Datasets (train_test_split)

Step 3: Classifying Predictors and Target, Classifying Independent and Dependent Features.

Step 4: Evaluating The Model With the R-squared , MAPE , and RMSE

Step 5: Building the Ridge Regressor (Initializing the Ridge Regressor with alpha =1.0)

Step 6: Fitting the Training data to the Ridge regressor

Step 7: Predicting for X_test

Step 8: calculate the R-squared of the model on the training data

Step 9: Calculate the MAPE of the model on the training data

Step 10: Calculate the MSE and RMSR of the model on the training data

Return Cost value

End

3.1.3 Least Absolute Shrinkage and Selection Operator (LASSO) Algorithm

The LASSO is a popular regression approach that achieves a sparse answer using a l_1 penalty. The LASSO is also known as the basis pursuit in the signal processing literature. For example, generalized linear models, as well as Cox's proportional hazard models for survival data, have been widely employed. LASSO is uninterested in highly linked predictors, preferring to pick one and ignoring the rest. Many coefficients should be close to zero, while a small fraction should be larger and nonzero, according to the LASSO penalty. It's a linear model with a regularization term added to it mathematically, as in Eq. 4. The function to minimize as an objective function is as follows: [23].

$$\min_{w} \frac{1}{2n_{samples}} \left\| x_{w-y} \right\|_2^{2^{i=1}} + \alpha \|w\|_1 \quad (4)$$

The LASSO estimate, therefore, solves the least-squares penalty minimization with $\alpha||w||_1$ added, in which α refers to a constant and $||w||_1$ refers to the l_1 norm of the coefficient vector [24].

2.1.4 Random Forest (RF) algorithm

RF is a more advanced classification and regression decision tree approach. It is also a member of the learner ensemble. Because of its simple structure, a decision tree is a simple method to employ. Unfortunately, because of the enormous variance, it is unstable. A random forest appears to solve the problem. RF is a process for creating numerous independent decision trees with varied sets of samples at each node and averaging the scores of each decision tree as the final score to achieve a more precise outcome [25]. The algorithm constructs a forest with a number of decision trees during training. A set of decision nodes divides a tree into its many branches until it achieves the termination point (the leaf), which is the decision tree's prediction. Each decision node is dependent on if the value of input features is higher than or equivalent to a threshold value. Every forest tree is presented in a subtly distinct approach to mimic a model. The resulting prediction is obtained by averaging each of the tree forecasts [26].

3.2 MATERIAL

In this study, the dataset contain about 1660 project encompasses at least one construction item, and every project has 24 features. The data for the construction items group and the economic data group was separated, as shown in Table 1.

Table 1. Construction data group

Feature No.	Project data	Data Description	Data type	Measurement Unit
		Data for the construction items group		
1	Natural Ground Preparations	Natural ground preparations price	Numerical	Iraqi dinar
2	Width	Road width	Numerical	Meter
3	Earthwork Layers	Earthwork embankment price	Numerical	Iraqi dinar
4	Width	Earthworks width	Numerical	Meter
5	Thickness	Earthworks thickness	Numerical	Meter
6	Granular Sub-Base Layer	Granular sub-base layer price	Numerical	Iraqi dinar
7	Width	Granular sub-base works width	Numerical	Meter
8	Thickness	Mixed gravel layer width	Numerical	Meter
9	Asphalt Concrete Base Layer	Asphalt concrete base layer price	Numerical	Iraqi dinar
10	Width	Paving width	Numerical	Meter
11	Thickness	Paving thickness	Numerical	Meter
12	Pipe Tunnel 60cm	Pipe tunnel installation works price	Numerical	Iraqi dinar
13	Granular Shoulder Layer	Granular shoulder works price	Numerical	Iraqi dinar
14	Width	Granular shoulder works width	Numerical	Meter
15	Thickness	Granular shoulder works thickness	Numerical	Meter
		Economic feature group		
1	GDP	Iraq's Gross Domestic Production per capita	Numerical	N/A
2	Unemployment Index	Iraq Unemployment Rate	Numerical	%
3	Inflation Index	Iraq Inflation Rate	Numerical	%
4	Oil Price	Crude oil price	Numerical	$
5	Dollar Exchange Rate	Dollar change	Numerical	$
6	Region	Location of the project 1. Khanaqin district 2. Al-Miqdadiya district 3. Baladruze district 4. Ba'quba district 5. Al Khalis district	Numerical	N/A
7	Year	Year of execution	Numerical	N/A

4. MODELING METHODOLOGY

The methods recommended for predicting the cost of rural road construction are described in this section. The methodology consists of four stages in general, as shown in Fig. 1. Stage (1) Data Collection, stage (2) Machine Learning (ML) Models, stage (3) Model Evaluation, and stage (4) Analysis Performance.

4.1 STAGE 1: DATA COLLECTION

This research collected the Bill of Quantities (BOQ) for about 3000 road construction projects in rural areas in the Diyala governorate from 2012 to 2021. These projects included many types of projects such as; new road construction, construction of asphalt pavement layers only, asphalt overlay, and pavement maintenance. After conducting a screening process for these projects, only those whose construction items were chosen were considered in this research while excluding all the others. This process is justified by the uniformity of construction items whenever they are adopted, which may facilitate the training process. The Department of Roads and Bridges in the Diyala government was the source, so the raw data has been obtained from it. However, these data were unsuitable for the proposed system to be trained on, so the researcher manually worked on it for three months and reshaped it for a CSV file to train the model.

BOQ: A schedule set by the employer's engineer according to the paragraphs must be implemented successively. For the state departments in Iraq, the process requires the formation of a committee to organize the inspection, which conducts the on-site inspection on the site, whether it is a construction or maintenance detection.

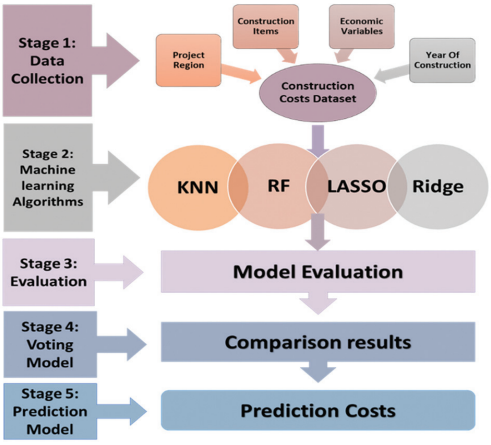

Fig. 1. Framework research cost roads regression.

STAGE 2: MACHINE LEARNING (ML) MODELS

ML approaches typically utilized are classification, clustering and regression [27]. The regression approach can be used to calculate the cost of construction. In general, any ML model consists of essentially three phases are explained as follows:

4.1.1 Preprocessing Phase

The primary goal of preprocessing is to turn raw data into a more suitable format for the predictor, allowing the prediction model to find patterns in the incoming data with ease.

I. Missing Data Preprocessing

Raw data was obtained from a legitimate government source and manually entered into an excel file, which was then prepared for use in the proposed project cost prediction model.

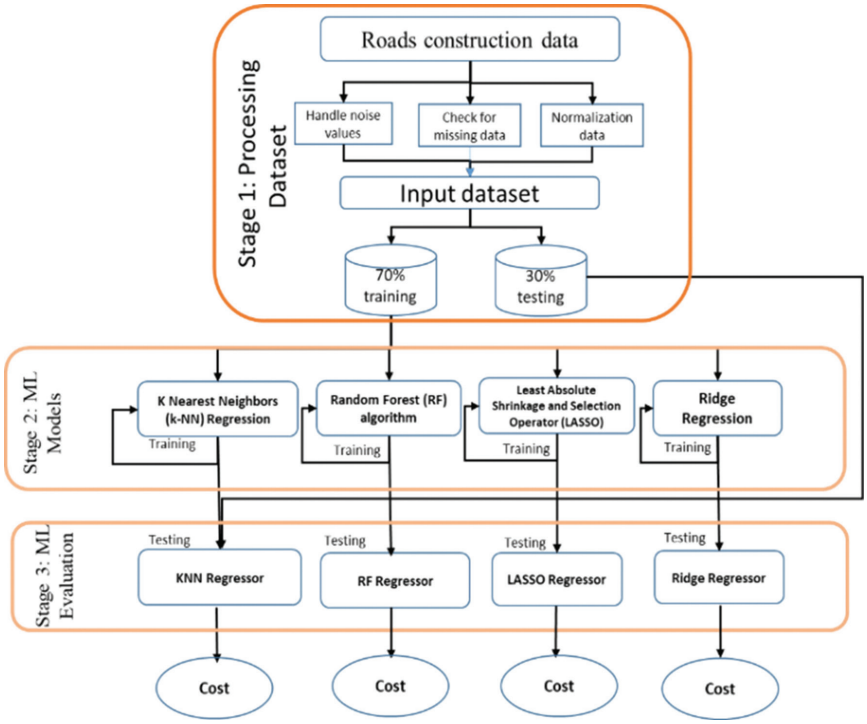

Fig. 2. Framework Machine Learning Models for cost estimation.

Therefore, there is no need for any compensatory approaches. After all, no compensating methods are required because there is no missing data following verification.

II. Data Normalization

Normalization aids in transforming an attribute's value into a limited set of values. It is the process of transmitting data to a specific range, such as 0 to 1 or -1 to 1. Although there are large differences in the values of different feature ranges, normalization is required. Data normalization, in which training time is started to access feature ranges of the same size, reduces training time.

4.1.2 Training Phase

In supervised learning, the classifier requires a training set that contains labeled samples of the domain with which it interacts to extract the requisite information and use it to predict future unlabeled inputs. Additionally, labeled data is necessary to test the classifier's performance by comparing the classifier's predictions to the actual classes of the inputs. The utilization of data was obtained from the training set for evaluation.

As indicated in Figure 2, 70% of the samples in the dataset were utilized for training, while the residual 30% were utilized for assessment in this research. Because the testing data is not included in the training data, this technique ensures unbiased evaluation by keeping a large ratio of data for evaluation purposes. The training starts by changing the parameters of each algorithm, as given in the next section, and then evaluating the outcome. For the remaining parameters, the operation is repeated.

4.1.3 Testing Phase

During the testing phase, four prominent supervised ML algorithms are utilized to determine effective and efficient forecast models for road-building costs. K Nearest Neighbors Regression (k-NN), Random Forest (RF), Ridge and Least Absolute Shrinkage and Selection Operator (LASSO) Regressions have been used. The turning and training parameters of each algorithm are explored to generate the finest feasible prediction outcomes, and many models of these algorithms are shown for (k-NN)it was test the efficiency for the most important parameter was the value of K as shown in table 2, and the performance for RF model was test for the n_estimators and max_depth parameters as in Table 3 illstraute.

Table 2. K-NN performance

Model setting	R^2	MAPE	RMSE	Time 1	Time 2
1	0.99	0.005	1494.9	0.01	0.01
2	0.99	0.005	1924.0	0.03	0.01
3	0.99	0.006	2190.3	0.03	0.01
4	0.99	0.007	2401.0	0.5	0.01
5	0.99	0.008	2740.3	0.01	0.01

Table 3. RF performance

Model setting	R^2	MAPE	RMSE	Time 1	Time 2
n_estimators =100 max_depth= None	0.99	0.005	3265.5	0.4	0.01
n_estimators =1 max_depth= 1	0.93	4.6	31798.0	0.04	0.00
n_estimators =1 max_depth= 6	0.95	4.33	28624.6	0.04	0.00
n_estimators =100 max_depth= 6	0.99	0.02	3620.57	0.2	0.01

4.2. STAGE 3: MODELS EVALUATION

1. Regression analysis is an important part of supervised ML since it involves predicting a continuous independent target from a set of predictor variables. Various studies employ the Mean Square Error (MSE) and its rooted variant (RMSE), including the Mean Absolute Error (MAE) and its percentage variant. However, these rates have one shortcoming: since their values might vary from zero to infinity, a single value does not tell anything about the regression's efficiency in regard to the ground truth distribution. Therefore, this research employed two rates that only produce a large score if most of the elements in a ground truth group are accurately predicted.[28].

2. R-squared (R^2) is the coefficient of determination described as the fraction of the dependent variable's variance that may be estimated by the independent variables. The degree to which the model fits the cost data is expressed as follows:[28][29].

$$R^2 = 1 - \left(\frac{SSE}{SST}\right) \quad (5)$$

   in which SSE (sum of squares error) refers to the sum of squares of the residuals and SST (sum of squares total)refers to the total sum of squares. (Worst value = -∞; Best value = +1) [28].

3. MAPE is a regression model performance metric preferred for situations where relative variations are more highly relevant than absolute variations [30].

   Provided that x resembles the explanatory variables vector (the input to the regression model), y resembles the target variable as well as g denotes regression model, the MAPE of g is achieved by averaging the ratio over the data [31].

$$MAPE = \frac{|g(x)-y|}{|y|} \quad (6)$$

4. ***MSE***: A risk metric related to the squared (quadratic) mistake or loss's predicted value [31]. If $\hat{y_i}$ denotes the predicted value of the i-$_{th}$ sample, as well as y_i denotes the corresponding true value, here, the MSE estimated over is presented as [31].

$$MSE\ (y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2 \quad (7)$$

5. ***Root mean square error (RMSE)***: is the standard deviation of residuals (prediction errors). The RMSE is a measure of how spread out the residuals are, and the residuals are a measure of how distant the data points are from the regression line [31].

$$RMSE\ (y, \hat{y}) = \sqrt{\frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2}\ . \quad (8)$$

6. Training Time: The time it takes for a strategy to train the complete dataset and build the best-fit predictive model is referred to as (T1) [32].

$$T1 = end_{Time}^{Training} - Start_{Time}^{Training} \quad (9)$$

7. Testing time: (T2) is the time required for a technique to estimate the full dataset's construction costs [32].

$$T2 = end_{Time}^{Testing} - Start_{Time}^{Testing} \quad (10)$$

4.3 STAGE 4: ANALYSIS OF PERFORMANCE

R2, which evaluates how well the model matches the cost data, was used to assess model performance. As seen in table 1, R2 varies from 0 to 1. R2 values that are higher suggest better model performance. In part, the RMSE, which measures the average magnitude of an error, was employed to evaluate model performance. The RMSE should be as close to zero as feasible to indicate excellent model performance (for example, no error between real and anticipated costs). The performance of the model was assessed in part using MAPE, which determines if the model is more sensitive to relative than absolute fluctuations.

5. COMPARATIVE RESULT

For each of the four algorithms devised in this work, the extent of the modeling error was calculated, and comparison graphs of plots vs. output (for example, actual cost vs. predicted cost) were constructed. A diverse set of machine learning (ML) methods has been used to reach the optimal method for the road's construction real dataset proposed for the predictive cost model.

K Nearest Neighbors Regression (k-NN) was the first algorithm chosen to experiment with the instance-based algorithms. This method makes a decision using examples or instances of training data that the model considers necessary or essential. The R2 score value obtained by k-NN is 0.99, and the Mean Absolute Percentage Error (MAPE) is 0.006, but the Root Mean Square Error (RMSE) was very high 1485.6. Such algorithms frequently establish a database of example data and contrast incoming data to the database by utilizing a similarity measure to obtain the best match and produce a forecast. Random Forest (RF) ensemble techniques are models built of numerous weaker models that are individually trained and whose predictions are pooled in a particular manner to create the overall prediction. This data (RF) model failed to fit the data because of the linearity type of the data, so RMSE was 3181.8.

The Multicollinearity in this dataset leads to very high errors in the test phase in the K-NN and RF models. Multicollinearity is a condition to allow the correlation between the independent variables. Algorithms for regularization are another strategy (typically regression methods) that penalizes models for their complexity, preferring simpler models that are also stronger at generalizing. Ridge and Least Absolute Shrinkage and Selection Operator (LASSO) Regressions, the most popular algorithms for this method, have been utilized in this research. This kind of algorithm offered impressive results with the proposed dataset. The objective of lasso regression is to find the variables and regression coefficients that lead to a model with the least amount of prediction error. That is accomplished by imposing a constraint on the model parameters, which forces the sum of the absolute value of the regression coefficients to be smaller than a fixed value λ, hence shrinking the regression coefficients toward zero. The R2 score value obtained by LASSO Regression is 0.99, MAPE is 0.0002, and the RMSE is 0.09. Ridge Regression obtained the most elevated accuracy; in this model, it was tried to minimize the loss function, and the model was forced to find a balance between minimizing the residual sum of squares and minimizing the coefficients, which reached 0.000 for RMSE 1.00, R2, and the MAPE was 0.000. All models used the default parameters from the scikit-learn python library.

Time is the most important factor for calculating project costs, so this study focuses on the time factor for each ML model employed on the dataset, as shown in Table 4.

RF regression was the most terrible in training time than other models.

The time for training and testing is gradually reduced. The more accurate the model, the less error it is. That is evident in all other models, and when the model is perfect and gives a zero error, the time is ideal for training and testing, which applies to our proposed model, the Ridge Regression.

Table 4. Models performance

Regressors	T1 (s)	T2 (s)	R^2	MAPE	MSE	RMSE
Random Forest	0.53554	0.0 0978	0.99	0.0 07	10123 893.2	31 81.8
K-NN	0.03398	0.0 1499	0.99	0.0 06	22070 83.8	14 85.6
Lasso	0.31494	0.0 0000	0.99	0.0 0002	0.0 0383	0.0 6190
Ridge	0.40817	0.0 0000	1.00	0.0 00	0.0 000	0.0 0000

The following Figures show errors or differences between the predicted labels and the actual labels for road construction costs based on machine learning regressions. The RF modulator and the K-NN slope have errors in the prediction cost that do not lie on a

straight regression line. However, the other points' cost estimates lie correctly on the same line regression, as shown in Figures 3 and 5. LASSO and Ridge have presented the errors and actual road cost estimation in Figures 4 and 6. These regressions are shown as the accurate and great model for predicting costs for road construction based on error, and actual points lie precisely on one-line regression.

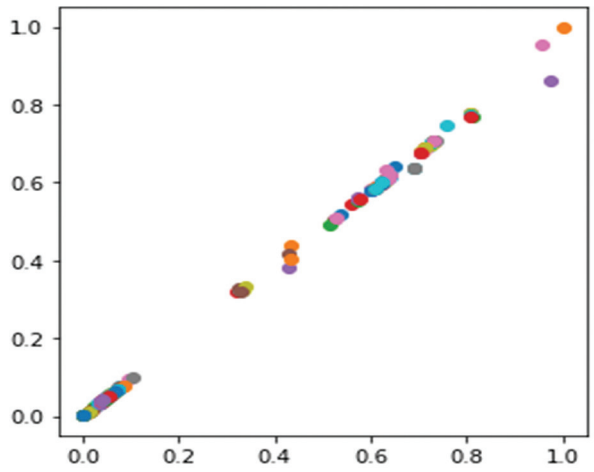

Fig. 3. RF Regressor

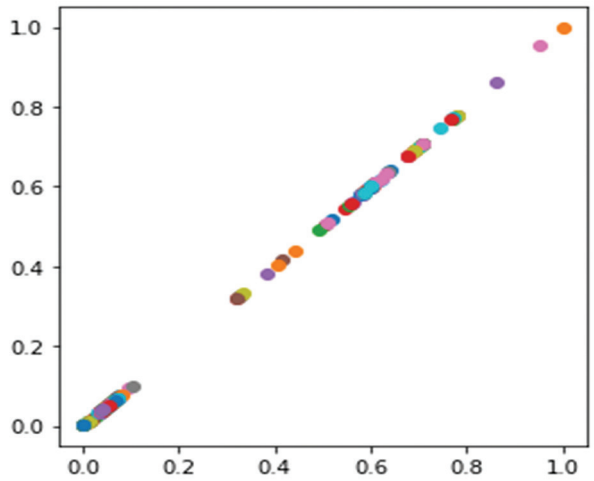

Fig. 4. LASSO Regressor

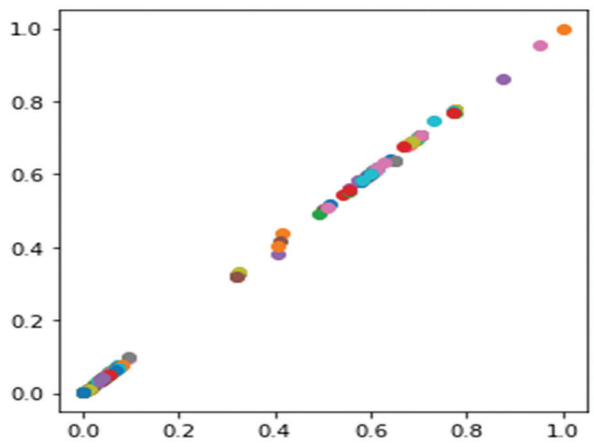

Fig. 5. K NN Regressor.

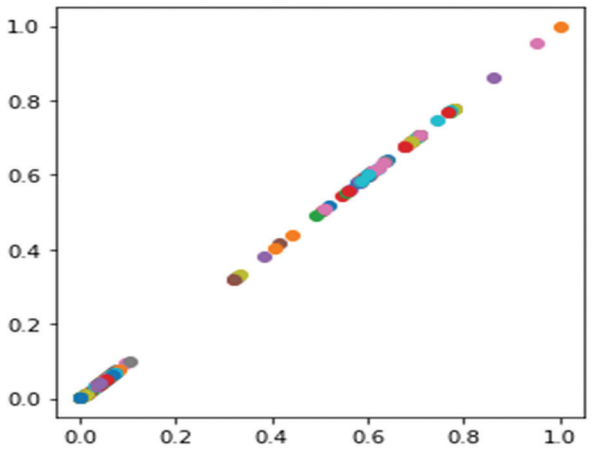

Fig. 6. Ridge Regressor.

6. CONCLUSIONS

Cost prediction is necessary for civil engineers to need ample time and action. This study collected a real road dataset for ten years. Furthermore, it examined the performance of several machine learning (ML) modeling for the prediction of road projects. The research aimed to advance machine learning techniques in estimating road construction. The suggested computing method is all-encompassing and can be applied to various construction projects in any city or region. The model has been presented using information gathered from a town. The proposed algorithm can be examined for additional building projects and areas.

On the road cost construction testing, the evaluation scale the Ridge model received the most significant estimate among the four models. In contrast, the other ML models had an acceptable level of cost estimation, except for the K Nearest Neighbors Regression (k-NN) model and the Random Forest (RF) model, which showed high errors in the cost prediction because of the linearity type for this data. This research directs a new direction for the ML approach to provide an economically reasonable cost with minimal effort to manage projects by artificial intelligence (AI) to achieve important goals for a road business. Additional research in the future might expand this analysis to include the use of deep learning algorithms and to collect data for different fields in construction engineering, another source to Increase the dataset to generate more precise cost estimation models for various objectives.

7. REFERENCES

[1] S. Petrusheva, D. Car-Pušić, V. Zileska-Pancovska, "Support Vector Machine Based Hybrid Model for Prediction of Road Structures Construction Costs", Proceedings of the IOP Conference Series: Earth and Environmental Science, Vol. 222, No. 1, 2019, pp. 1-11.

[2] I. Karaca, D. D. Gransberg, H. D. Jeong, "Improving the Accuracy of Early Cost Estimates on Transportation Infrastructure Projects", Journal of Management in Engineering, Vol. 36, No. 5, 2020, p. 04020063.

[3] M. Baek, B. Ashuri, "Spatial regression analysis for modeling the spatial variation in highway construction costs", Resilient Structures and Sustainable Construction, Georgia Institute of Technology, Atlanta, 2017.

[4] A. J. Ogungbile, A. E. Oke, K. Rasak, "Developing cost model for preliminary estimate of road projects in Nigeria", International Journal of Sustainable Real Estate and Construction Economics, Vol. 1, No. 2, 2018, pp. 182-199.

[5] A. Jaafari, I. Pazhouhan, P. Bettinger, "Machine learning modeling of forest road construction costs", Forests, Vol. 12, No. 9, 2021, p. 1169.

[6] L. B. Barros, M. Marcy, M. T. M. Carvalho, "Construction Cost Estimation of Brazilian Highways Using Artificial Neural Networks", International Journal of Civil Engineering, Vol. 7, No. 3, 2018, pp. 283-289.

[7] K. Tijanić, D. Car-Pušić, M. Šperac, "Cost estimation in road construction using artificial neural network", Neural Computing and Applications, Vol. 32, No. 13. 2020, pp. 9343-9355,

[8] G. Mahalakshmi, C. Rajasekaran, "Early cost estimation of highway projects in India using artificial neural network", Sustainable construction and building materials, Springer, Singapore, 2019, pp. 659-672.

[9] F. M. S. AL-Zwainy, I. A.-A. Aidan, "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, 2017, pp. 1-12.

[10] I. Peško et al. "Estimation of costs and durations of construction of urban roads using ANN and SVM", Complexity, Vol. 2017, 2017, pp. 1-13.

[11] N. Suneja, J. P. Shah, Z. H. Shah, M. S. Holia, "A neural network approach to design reality oriented cost estimate model for infrastructure projects", Reliability: Theory & Applications, Vol. 16, No. 1, 2021, pp. 254-263.

[12] M. Juszczyk, "On the search of models for early cost estimates of bridges: An SVM-based approach", Buildings, Vol. 10, No. 1. 2020, pp. 1-17.

[13] I. Mahamid, "Early cost estimating for road construction projects using multiple regression techniques", Australasian Journal of Construction Economics and Building, Vol. 11, No. 4, 2011, pp. 87-101.

[14] T. C. D. Lucas, "A translucent box: interpretable machine learning in ecology", Ecological Monographs, Vol. 90, No. 4, 2020, pp. 1-55.

[15] Y. Cao, B. Ashuri, M. Baek, "Prediction of Unit Price Bids of Resurfacing Highway Projects through Ensemble Machine Learning", Journal of Computing in Civil Engineering, Vol. 32, No. 5, 2018, p. 04018043.

[16] M. Flah, I. Nunez, W. Ben Chaabene, M. L. Nehdi, "Machine Learning Algorithms in Civil Structural Health Monitoring: A Systematic Review", Archives of Computational Methods in Engineering, Vol. 28, No. 4. 2021, pp. 2621-2643.

[17] I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, F. Herrera, "Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 9, No. 2, 2019, pp. 1-24.

[18] F. Martínez, M. P. Frías, F. Charte, A. J. Rivera, "Time Series Forecasting with KNN in R : the tsfknn Package", The R Journal, Vol. 11, No. 2, 2019, pp. 229-242.

[19] G. C. McDonald, "Ridge regression", Wiley Interdisciplinary Reviews: Computational Statistics, Vol. 1, No. 1, 2009, pp. 93-100.

[20] C. Saunders, A. Gammerman, V. Vovk, "Ridge Regression Learning Algorithm in Dual Variables", Proceedings of the Fifteenth International Conference on Machine Learning, 1998, pp. 515-521.

[21] A. Afzal, S. Alshahrani, A. Alrobaian, A. Buradi, S. A. Khan, "Power plant energy predictions based on thermal factors using ridge and support vector regressor algorithms", Energies, Vol. 14, No. 21, 2021, pp. 1-22.

[22] L. E. Melkumova, S. Y. Shatskikh, "Comparing Ridge and LASSO estimators for data analysis", Procedia Engineering, Vol. 201, 2017, pp. 746-755.

[23] J. F. H. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent", Journal of Statistical Software, Vol. 33, No. 1, 2010, pp. 1-22.

[24] S. J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, "An interior-point method for large-scale ℓ1-regularized least squares", IEEE Journal of Selected Topics in Signal Processing, Vol. 1, No. 4, 2007, pp. 606-617.

[25] F. Hidayat, T. M. S. Astsauri, "Applied random forest for parameter sensitivity of low salinity water Injection (LSWI) implementation on carbonate reservoir", Alexandria Engineering Journal, Vol. 61, No. 3, 2022, pp. 2408-2417.

[26] H. Tong, B. Liu, S. Wang, "Software defect prediction using stacked denoising autoencoders and two-stage ensemble learning", Information and Software Technology, Vol. 96, 2018, pp. 94-111,

[27] S. B. Jha, R. F. Babiceanu, V. Pandey, R. K. Jha, "Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study", arxiv.org/abs/2006.10092, 2020.

[28] D. Chicco, M. J. Warrens, G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation", PeerJ Computer Science, Vol. 7, 2021, pp. 1-24.

[29] H. H. Elmousalami, "Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-the-Art Review", Journal of Construction Engineering and Management, Vol. 146, No. 1, 2020, p. 03119008.

[30] H. H. Elmousalami, "Comparison of Artificial Intelligence Techniques for Project Conceptual Cost Prediction: A Case Study and Comparative Analysis", IEEE Transactions on Engineering Management, Vol. 68, No. 1. 2021, pp. 183-196.

[31] A. de Myttenaere, B. Golden, B. Le Grand, F. Rossi, "Mean Absolute Percentage Error for regression models", Neurocomputing, Vol. 192, 2016, pp. 38-48.

[32] Z. K. Maseer, R. Yusof, N. Bahaman, S. A. Mostafa, C. F. M. Foozy, "Benchmarking of Machine Learning for Anomaly Based Intrusion Detection Systems in the CICIDS2017 Dataset", IEEE Access, Vol. 9, 2021, pp. 22351-22370.