Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Research Article

Estimating contract price using structural variables: a machine learning approach with data preprocessing

Semi Emrah ASLAY[1*]

[1] Erzincan Binali Yildirim University, Department of Civil Engineering, seaslay@erzincan.edu.tr, Orcid No: 00000-0002-0127-5474

ARTICLE INFO

Article history:

Received 12 July 2024
Received in revised form 11 September 2024
Accepted 24 September 2024
Available online 30 September 2024

Keywords:

SVM, machine learning, structural parameters, data preprocessing, feature importance.

Doi: 10.24012/dumf.1515160

* Sorumlu Yazar

ABSTRACT

Accurately estimating the construction contract price is a necessary step for correctly determining project budgets and ensuring efficient use of resources. In this study, contract price in public construction tenders are estimated using structural project variables. The variables applied in the study are created by adding the quantities of columns, shear walls, and beams to variables commonly used in the literature for cost estimations. Six different machine learning algorithms are employed as machine learning algorithms. These are Support Vector Machine (SVM), Decision Tree (DT), eXtreme Gradient Boosting (XGBoost), k-Nearest Neighbors (KNN), Artificial Neural Network (ANN), and Random Forest (RF). Preprocessing methods and a series of hyperparameter optimizations are applied to enhance the predictive capability on datasets. These processes and the applied algorithms are evaluated with five different performance metrics. The Support Vector Machine (SVM) algorithm produced the best results, achieving a coefficient of determination (R^2) of 0.8966, a Mean Absolute Percentage Error (MAPE) of 23.70, a Nash-Sutcliffe Efficiency (NSE) of 0.8956, a Mean Absolute Error (MAE) of 0.4849, and a Root Mean Square Error (RMSE) of 0.6989. This study contributes to the literature by developing machine learning models and data analysis processes for contract price approaches.

Introduction

In today's world, with the advancement of construction technology, the design of structures can be easily accomplished. Not only ordinary buildings, but also specialized engineering structures, can be designed and constructed. This situation is the result of a long process that has evolved over time. In the future, it is expected that construction activities will progress in a way that adapts to the digital world [1]. The effectiveness of artificial intelligence applications is increasing in the technological world, and its reflections can also be seen in construction activities. Machine learning models, one of the subfields of artificial intelligence, play an important role in the digitalization process in many areas related to construction works and have become one of the popular methods used in recent years [2] .

Building cost estimation can take various forms, including contract price estimation, cost estimation, preliminary cost estimation, cost index estimation, and final cost estimation. Several popular machine learning models are used in cost estimation models [2] , [3]. Artificial neural networks are fundamental algorithms for cost estimation [4] not only for cost estimation but also in studies where the final construction duration is predicted [5]. Saeidlou & Ghadiminia [6] created a model using ANN on seven variables as part of a cost estimation framework, which includes a validation unit. Car-Pusic et al. [7] estimated the preliminary costs of buildings based on a neural network-based model. Pham et al. [8] preferred regression models along with cost optimization. Zhang and Fang [9] evaluated building properties and conducted kernel principal component analysis with popular algorithms (support vector machine, random forest). Badawy [10] used 174 residential projects in Egypt for early-stage cost estimation by developing a hybrid model. Coffie et al. [11] applied multiple regression to 911 construction projects in Ghana. Hassim et al. [12] and Sayed et al. [13] also used survey studies in predicting construction costs. Aslay and Dede [1] optimized construction costs using different algorithms. Dang-Trinh et al. [14] conducted preliminary cost estimation for 35 different factory projects using various learning algorithms. Uysal and Sonmez [15] calculated conceptual costs using a case-based reasoning method combined with bootstrap. Ali et al. [16] compared models used in their research by employing an integrated data intelligence model for 90 building projects in Iraq. Alfaggi & Naimi [17] sought solutions to construction projects using a fuzzy-AHP model. There are studies related to construction costs or contract prices with economic data or

inflation indicators [18]. XGBoost, neural networks, time series, and regression methods were employed to address cost issues by incorporating economic data into variables [19], [20]. In addition to superstructure projects, machine learning models are frequently used in various construction styles such as underground metro station cost-material estimation [21], public highways construction time and cost [22], tunnel projects [23],and prestressed bridges [24]. It is noteworthy that there are no studies in the literature where contract prices are estimated based on structural project parameters. Especially in the tender processes of public buildings, there is a need for more research on agreement prices. It is evident that both for construction engineers and contractor firms, it is possible to relate contract prices with all structural elements in the structural project by examining the rough construction.

In the past decade, construction firms in Turkey have been increasingly focusing on construction projects tendered by the public sector, particularly due to declining project profitability. Therefore, alongside the importance of rapid technical foresight and readability risks of construction contracts [25], understanding what constitutes the contract prices of construction projects has become crucial. Nowadays, firms participating in public tenders typically try to estimate contract prices by individually creating quantity take-offs and predicting contract amounts based on unit prices. However, this method is insufficient for contract price estimation because many bidding companies participate in public tenders, making it difficult to determine the contract amount agreed upon with the public administration. Especially civil engineers attempt to predict contract prices supported by structural project parameters, which can be described as rough construction. They make approximate estimates for the monetary value of the structure based on concrete-rebar-formwork costs [26]. However, it can be said that these estimates are not highly professional and are subject to market conditions. There is a need in the construction industry for contract price estimation models supported by building elements. Additionally, these models should be capable of being enhanced with parametric optimizations.

In this study, a model developed through a series of optimization methods is presented for estimating contract prices in public construction tenders by incorporating the characteristics of reinforced concrete structural elements into existing variables. The monetary agreement value is attempted to be predicted based on the distribution of variables and a structural model. Parameters and functions are established for the structural form to solve the problem. The variables constituting the datasets are obtained from various construction projects. Variable characteristics are determined by analyzing both similar studies in the literature and market needs. The machine learning model is developed to the extent possible through the preprocessing of data and optimization of certain parameters. Results pertaining to different performance metrics are obtained using six different machine learning algorithms. Both the building examples and the research on contract price estimation, along with the pros and cons of the prediction model, are presented in this paper. In this study, efforts are undertaken to contribute to the construction sector by addressing these deficiencies through different machine learning algorithms, prediction models, and optimization processes in data correction stages.

Material and Method

Data Sets

There are various parameters used for predicting contract prices in construction works. The contract price is also a type of construction cost estimation, and many input parameters have been created for cost estimation in machine learning methods. The study conducted by Asuncion and Newman [27] is one of the primary works that generated these parameters. In other studies in the literature, some of these parameters have been used, and additional parameters have been added to create datasets [28], [29], [6]. In this study, data sets were constructed by reviewing existing literature and adding new parameters to improve the performance of machine learning models in predicting contract prices for building construction. The samples utilized for constructing the datasets are summarized in table 1. This table offers a comparative analysis of the samples provided in the literature and those examined in this study. A total of 15 different input examples are presented in the table. These are; Total floor area of the building, number of columns, number of shear walls, type of building, lot area, project location, duration of construction, number of building floors, type of footing, number of special facilities, soil type, cost of construction workers, building height, number of beams and number of elevator. In this study, six different input data are used: Total floor area of the building, number of columns, number of shear walls, lot area, number of building floors, building height, number of beams. Apart from the inputs presented in the table-1, 15 additional input samples were evaluated in the study. However, it was determined that 6 of these inputs performed well.

Data Features

In this study, data were obtained from structural projects of 353 different reinforced concrete buildings. The buildings are located in various provinces of Turkey. All projects are tendered on a " turnkey lump-sum contract" basis. This contract type entails a project where all stages, from start to finish, are carried out by the contractor firm for a single agreed-upon price. The data for the 353 projects awarded through this contract type are obtained from the official government website "Elektronik Kamu Alımları Platformu-EKAP" [30] . An expert civil engineer collected the data from the structural projects. These data represent datasets found only in structural projects The variables created for the prediction model in the research are structural project parameters. Initially, 15 variables were calculated and used to create the dataset. Later, the relationships between the data were examined both among themselves and with the target variable. Correlation matrices, feature importance, and performance metrics were examined for each combination of variables, resulting in the initial 15 variables

being reduced to 6. The parameters used in the datasets building's foundation; average number of columns is the obtained by dividing the total number of columns by the count obtained by dividing the total number of beams by the number of floors construction cost price is the agreement price between the administration and the contractor; total

Table 1

The parameters used in the datasets

Input Parameter	[27] A.Asuncion,&D. Newman, 2007	[28] A.Mohammed et al. 2015	[31] Rafiei &Adeli, 2018	[32] Saeidlou& Ghadimini a, 2023	This Study
Total floor area of the building (x1)	+	+	+	+	+
Number of columns (x2)	-	-	-	-	+
Number of shear walls (x3)	-	-	-	-	+
Type of building (x7)	-	+	-	+	-
Lot area (x4)	+	-	+	-	+
Project Location (x8)	-	+	-	-	-
Duration of construction (x9)	+	-	+	-	-
Number of building floors (x5)	-	+	-	+	+
Type of footing (x10)	-	+	-	-	-
Number of special facilities (x11)	-	-	-	+	-
Soil Type (x12)	-	+	-	-	-
Cost of construction workers (x13)	-	-	-	+	-
Building height (x14)	-	+	-	-	-
Number of beams (x6)	-	-	-	-	+
Number of elevator (x15)	-	+	-	-	-

floor area of the building is the sum of the areas of the structure across all floors; number of columns is the count of columns across all floors of the structure; number of shear walls is the count of shear walls across all floors of the structure; total number of structural elements is the sum of columns, shear walls, and beams in the building; lot area is the land area where the building is constructed; average number of structural elements is the count obtained by dividing the total number of structural elements by the number of floors; duration of construction refers to the completion time of the construction; number of building floors is the sum of basement, ground, and typical floors; total number of vertical elements is the sum of columns and shear walls in the building; average number of vertical elements is the count obtained by dividing the total number of columns and shear walls by the number of floors; average number of shear walls is the count obtained by dividing the total number of shear walls by the number of floors; number of beams is the count of beams across all floors of the building; foundation height denotes the concrete height of the count number of floors; average number of beams is the

Method

Machine learning is one of the subfields of artificial intelligence and is highly capable in interpreting data. Data analysis and prediction of target parameters are commonly used in all academic disciplines. Researchers use various algorithms to analyze the data and build forecasting values. These algorithms are used to develop models tailored to the problem type and objective. Machine learning algorithms used in the study; Support vector machine (SVM), Random forest (RF), Extreme gradient boosting (XGBoost), k Nearest neighbors (KNN), Artificial Neural Networks (ANN), and Decision tree (DT) algorithm.

Support vector machine (SVM)

Support Vector Machine (SVM) was developed by Vapnik [33]. The method is considered one of the regression modeling techniques, and its main goal is to find the optimal hyperplane for the best data prediction.The concept of a hyperplane can be described as an attempt to draw the most appropriate line between data points on any x-y plane. The

concept of margin refers to the lines parallel to the hyperplane that have the best range to model the data effectively [34]. In order to enhance the generalization capabilities of Support Vector Machine models, margins on the x-y plane are kept as wide as possible, with the main axis being the hyperplane line. Decision boundaries are used to partition data points and improve prediction accuracy. Decision boundaries are also referred to as support vectors. Support vectors aim to determine the best hyperplanes by methodologically distinguishing between data points, thereby enhancing the performance of predicting target values [35], [36]. The structure of the SVR model is indicated in Figure 1. SVM maps data into a high-dimensional space to create regression models. It identifies the optimal hyperplane, and the kernel functions used incorporate nonlinear relationships between the data into the model. This allows SVM to also create nonlinear regression model

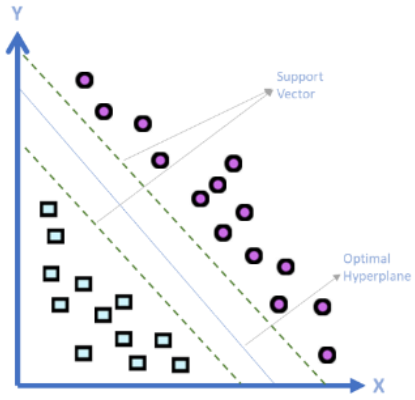

Fig. 1. Structure of SVM model

Decision tree (DT)

Decision trees (DT) [37] are inspired by the structure of a sample tree for solving problems. The information source is divided into many parts and subgroups. The splitting process can be performed using methods such as entropy or the sum of squared errors [38]. Model ranges are symbolized as three types of nodes: root nodes, internal nodes, and leaf nodes. Each node represents a decision rule and generates predictions. To make predictions on the training set, the mode or mean of the outcomes is taken, and the model is completed. Decision trees can often involve issues such as overfitting and are typically used alongside methods that allow for ensemble use, like random forests.

Random Forest (RF)

Random Forest (RF), developed by Breiman et al. [39] , is created by combining a large number of decision tree models. Random Forest (RF) is formed by bringing together a large number of decision tree models. The algorithm randomly selects all the trees and trains the model by dividing each one into subsets to make the results more reliable. The Gini index within the “caret” package in R Studio is used for the splitting process. The main characteristic of the Random Forest algorithm is the bagging method. This method attempts to reduce variance at each stage of data prediction by voting or averaging, thereby enhancing predictive ability. Random Forest provides a good option for modeling high-dimensional and complex datasets. It can achieve necessary accuracy, particularly in complex datasets with a high degree of disparities. It is resistant to overfitting and performs well even in such situations [40]. The structure of the RF model is presented in Fig.2.

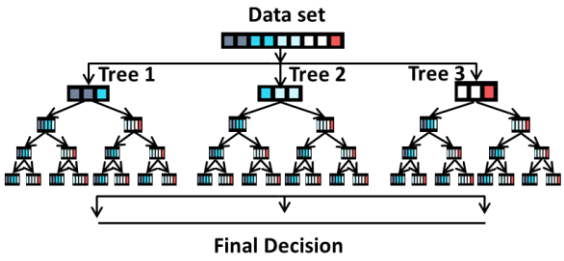

Fig. 2. Structure of RF model

Extreme gradient boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) has quickly become a popular method in machine learning in recent years, although it was developed relatively recently by Chen and Guestrin [41]. It is based on gradient boosting and constructs models by aggregating decision trees [42]. The algorithm's frequent usage stems from its scalability and fast execution. It handles sparse matrices efficiently and weights examples in tree learning with weighted quantile sketching. XGBoost performs cross-validation and automatically handles missing values. It is preferred for large datasets and high-dimensional feature spaces. In addition, it is resistant to overfitting and adjusting hyperparameters is quite easy. With all these features, it provides a functional algorithm example for users [43].

K Nearest neighbors (KNN)

K Nearest Neighbors (KNN) is a fundamental machine learning algorithm [44]. It is a non-parametric method that attempts to provide a solution through distance functions. The distance function is defined by the parameter k, which calculates the distance to the training set. Thus, the created dataset is evaluated within the class that is closest, attempting to reach a conclusion. The value of k is a hyperparameter within the KNN algorithm. This value is determined by the user and is not assigned randomly. It is typically determined through trial and error or methods like cross-validation. While the KNN algorithm is successful and effective in small datasets, it loses this advantage as the dataset grows [45].

Artificial Neural Networks (ANN)

Artificial neural networks (ANN) [46] is a machine learning algorithm created by simulating biological neural networks and cells. Just like in the nervous system, an architecture of layers, nodes, and connections is established, and an algorithm uses the values of observations (inputs) to

determine the weights within this network. In this way, they facilitate transmission and are referred to as nodes or neurons in the literature. Additionally, there are layers connecting this group of neurons. The data is weighted within the neurons using the backpropagation algorithm to determine the weights. These weights are then used to calculate various possibilities, and finally, output values are generated as a result of these operations. Outputs are then transmitted to other neurons and layers, striving to reach a conclusion. Artificial neural networks have 3 layers: input layer, hidden layer, and output layer. The input layer receives the data and transfers it to the hidden layer. It has neurons, and the number of neurons must not be less than the number of inputs. The layer in between is called the hidden layer, processing the inputs and forming the model design, teaching the data relationship to the model [47].

Model improvement techniques

There are many methods and techniques aimed at improving the predictive performance of the created models in machine learning. If changes made to the dataset or algorithm improve the performance of the output parameter of the model, each correction is defined as a model improvement technique. Some of these techniques include gathering more data, removing some outlier data, selecting the appropriate model, tuning procedures, cross-validation, and utilizing better metrics.

Machine learning models should be built using datasets that contain relevant relationships for accurate predictions. Additionally, having an adequate amount of data and removing irrelevant data are crucial for accurate model prediction. For this purpose, basic statistical methods, box plot, histogram, and clustering techniques can be used. In the accurate construction of prediction models, it is necessary to have a sufficient percentage of train-test data, apply tuning procedures, and evaluate cross-validation metrics. The application of these techniques is generally done through trial and error, and by evaluating the results of data analysis. For example, taking 70% of the entire data group for the train set and 30% for the test set, and comparing it with the second group where the same dataset is split into 80% train data and 20% test data, can be given as an example. Additionally, for cross-validation, results can be compared across different values of N-fold and hyperparameter optimizations performed for different models.

In this study, box plot graphs were preferred for detecting outlier values. By analyzing the graphs, outlier data points were removed from the datasets. The accuracy of the prediction model was evaluated based on improvement in metrics. Train-test groups with different percentages were compared for the datasets. Cross-validation parameters and hyperparameter optimization were conducted across various machine learning algorithms. As hyperparameter values, for SVM, the regularization parameter, sigma, degree, and scale are used. For Decision Trees (DT), the complexity parameter, minimum number of observations, and maximum depth of the tree are specified. Random Forest (RF) features key parameters such as the number of variables, number of trees, and minimum size of terminal nodes. XGBoost involves important parameters including the number of boosting iterations, learning rate, maximum depth of a tree, and gamma. K-Nearest Neighbors (KNN) relies on the parameter k. Lastly, Artificial Neural Networks (ANN) require defining the number of units in the hidden layer, the regularization parameter, and the maximum number of iterations for training. The results were evaluated. The mentioned processes have yielded positive and negative results for the model's performance.

Performance Evaluation

Various performance metrics are utilized for prediction models in this study. Determination Coefficient (R^2), Mean Absolulite Percantage Error (MAPE), Nash–Sutcliffe Efficiency (NSE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) indicators are employed as performance metrics for the models. The R^2 value indicates how well the regression model fits the data and ranges between 0 and 1. Also known as the determination coefficient, R^2 closer to 1 suggests better prediction accuracy. MAPE, Mean Absolute Percentage Error, conveys error values as a percentage by comparing actual values with predicted values. A lower MAPE value signifies a better model, typically acceptable values range between 20% and 50%, though variations may occur depending on the data source. NSE, Nash-Sutcliffe Efficiency, is another accuracy measure that calculates the proportional variance of actual values to predicted values, ranging from -∞ to 1, with values closer to 1 indicating higher success. MAE, Mean Absolute Error, represents the average of the absolute differences between actual and predicted values, where smaller values are preferable. RMSE (Root Mean Square Error) computes the square root of the average of the squared differences between actual and predicted values. A lower RMSE implies better

Table 2
The equations of performance criteria.

Term	Function	Equation
R^2	Determination Coefficient	$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$
RMSE	Root Mean Square Error	$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$
MAE	Mean Absolute Error	$MAE = \frac{1}{n}\sum\lvert y_i - \hat{y}_i\rvert$
NSE	Nash–Sutcliffe Efficiency	$NSE = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$
MAPE	Mean Absolulite Percantage Error	$MAPE = \frac{100\%}{n}\sum\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$

performance of predicted values over actual values. Some articles or researchers may consider RMSE close to 0 as a criterion for low RMSE value, but it should be noted that in some studies, RMSE may be low but not close to 0. For instance, when the target variable is in the order of millions or billions, it is more appropriate to evaluate the RMSE value within the dataset or among algorithms. The mathematical expressions of the performance metrics are given in Figure 4. The numerical values and analyses of these metrics are shared in the "Results and Discussion" section of this study.

Data Preprocessing

In machine learning models, creating a correlation matrix is a popular data preprocessing method for examining relationships among variables. Although there are numerous methods for exploring relationships between variables, a correlation matrix is both simple and functional. In this study, during the data preparation stage, data with initially 15 variables were obtained from structural projects. Some of these variables were derived both individually and by aggregating with each other. For example, the numbers of columns, shear walls, and beams in a structure were calculated separately. Additionally, the sums of columns, shear walls, and beams were aggregated to create a new variable representing the total number of structural elements. Definitions of these variables are detailed in the preceding section. The Pearson correlation coefficient was utilized to evaluate the relationship between the variables. Figure 3 shows the correlation matrix of the dataset created in the final stages. Accordingly, the correlation coefficients between the variables and the target variable range from a maximum of 0.75 to a minimum of 0.01. Generally, the author aims not to include variables with correlation coefficients below 0.50 but still explores different variable combinations. Variables take their final form through data analyses, including correlation matrix, feature importance, and model performance testing. These data analyses are presented in subsequent sections. The workflow of the study is given in Figure 4. According to this, in the correlation matrix, the variable with the highest correlation with the target variable is "construction area" at 0.71, and the variable with the lowest correlation is "Lot area" at 0.48. Furthermore, the relationship scores among the variables themselves range from 0.83 to 0.33. It can be said that the relationship distributions among all variables are approximately homogeneous when examining the relationships among them. The values in the dataset were also subjected to outlier detection and standardization procedures. Quartile analysis and box plot visuals were utilized to examine both the outliers and the variations among the numerical expressions of values in the dataset. In the analysis, it was observed that the "construction area" variable and the "lot area" variable had high scale dimensions. Therefore, standardization processes were applied to all variables to eliminate the scale problem among the variables. Data points that were identified as outliers were removed from the dataset. The initial dataset, which consisted of 359 records, decreased to 353 records after the removal of outliers. It is understood that this decrease in the number of data points resulted in a 5% increase in the performance metrics of the prediction model.

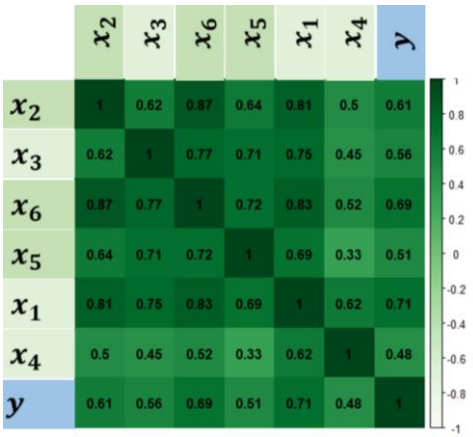

Fig. 3. Correlation matrix

Machine Learning Model

In this study, machine learning algorithms such as Artificial Neural Networks (ANN), k-Nearest Neighbors (KNN), Random Forest (RF), Decision Trees (DT), Extreme Gradient Boosting (XGBoost), Support Vector Machines

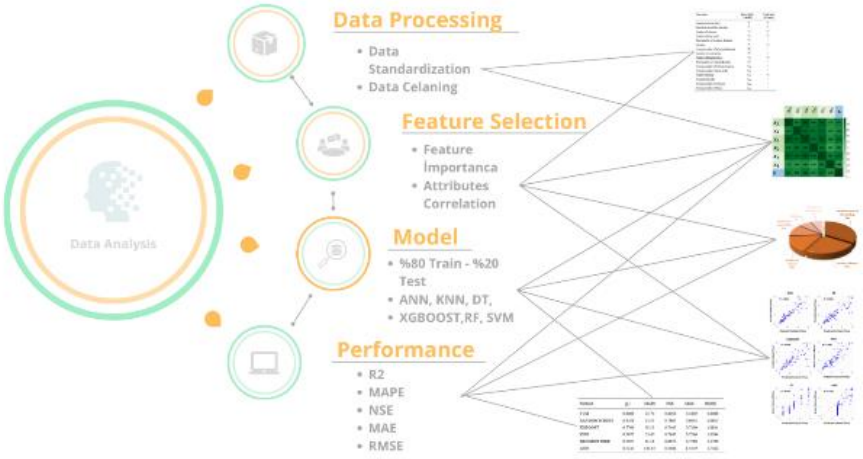

Fig. 4. Flowchart diagram of the predictions framework

(SVM) are applied for data analysis. The dataset consists of 6 independent variables and 1 dependent variable. There are 353 data points. All algorithms are implemented in the R Studio software program, and necessary checks are performed. Depending on the algorithm type, codes are updated accordingly. In the study, different seed values are tried with the algorithms to find the best prediction models. It is understood that the performance metric values converge at a certain point. The convergence of performance metric values indicates that the experiments are sufficient, demonstrating the adequacy of trials. Additionally, different n-fold values are experimented with for cross-validation, ranging from 5 to 10, aiming to achieve the most suitable outcome. The tuning parameter is optimized for the models, where various combinations are applied. Through all these processes, research is conducted to obtain the best results for the prediction models. The flowchart diagram of the prediction framework is presented in Figure 4.

Results and Discussion

Model Training and Evaluation

Machine learning prediction models involve one of the most crucial stages, which is the division of the dataset into test and train sets. The model should be sufficiently trained, and the trained model should be evaluated with the test data. In this study, for the test-train data, four different possibilities were explored across all models: 50% train - 50% test, 60% train - 40% test, 70% train - 30% test, and 80% train - 20% test. Sample sizes for each ratio were randomly generated from the total dataset. Among all algorithms, the 80% train - 20% test ratio was found to be functional. An example illustrating the efficiency of test-train data ratios relative to each other can be observed in the study where the Support Vector Machine (SVM) achieved the best result. In samples of structural project parameters in the dataset, it obtained an R-squared value of approximately 0.8966 with an 80-20 train-test ratio, while it yielded 0.8244 with a 70-30 train-test ratio. This difference represents a 9% variation in percentage terms. The success of the test-train ratio can vary across all datasets, and comparing these ratios within each dataset can significantly contribute to the model effectiveness.In the research, efforts were made to improve the efficiency of the model by altering the Cross-validation parameters and trying different hyperparameter values. Cross-validation values ranging from 5 to 10 were applied. In terms of hyperparameter values, algorithm coefficients, number of layers, number of trees, learning rates, etc. were considered, and different trials were conducted. Differences in the model's efficiency were observed in both cross-validation scores and hyperparameter values. These values are providing good results depending on the characteristics of the datasets and the relationship between the dataset and the algorithm.

Modelling results

In this study, machine learning models such as SVM, Random Forest, XGBoost, KNN, Decision Tree, and ANN were employed. Performance metrics including , R^2, MAPE, NSE, MAE, and RMSE were applied. The dataset benefited from structural project parameters. Attempts have been made to reach the best prediction models for contract price estimation using variables such as column count, shear wall count, beam count, floor count, total floor area, and lot area. The primary objective is not to compare machine learning algorithms. However, while constructing prediction models, both the effectiveness of the best model and the analysis of how efficient the efficient model produces efficiency were aimed. Comparison among algorithms was performed as part of the data analysis process to understand which model performs better. The best models among the six different ones were determined as SVM, RF, XGBoost, KNN, Decision Tree, and ANN, respectively. For comparative purposes, five different performance metrics were used. Among these metrics, the best values were achieved by SVM and RF. SVM yielded the best result with a R^2 value of 0.8966 and a MAPE value of 23.70. Following closely, RF gave the second-best result with a R^2 value of 0.8258 and a MAPE value of 33.43. The performance metrics for other algorithms showed relatively poor results. The values for KNN, XGBoost, and DT were obtained 0.76. ANN, on the other hand, provided the worst

Table-3
Performance comparison of the models

Method	R^2	MAPE	NSE	MAE	RMSE
SVM	0.8966	23.70	0.8956	0.4849	0.6989
RANDOM FOREST	0.8258	33.43	0.7885	0.6942	0.9853
XGBOOST	0.7706	35.10	0.7445	0.7100	1.0831
KNN	0.7675	31.65	0.7647	0.7266	1.0394
DECISION TREE	0.7657	41.44	0.6972	0.7786	1.1790
ANN	0.7210	143.43	0.3586	1.4579	1.7162

value of 0.72. It was observed that the values among all algorithms ranged from 0.8966 to 0.7210. It can be said that the performance metrics also exhibit similar correlations amongst themselves, akin to those of and MAPE. This information is summarized in Table 5.best models among the six different ones were determined as SVM, RF, XGBoost, KNN, Decision Tree, and ANN, respectively. For comparative purposes, five different performance metrics were used. Among these metrics, the best values were achieved by SVM and RF. SVM yielded the best result with a R^2 value of 0.8966 and a MAPE value of 23.70. Following closely, RF gave the second-best result with a R^2 value of 0.8258 and a MAPE value of 33.43. The performance metrics for other algorithms showed relatively poor results. The R^2 values for KNN, XGBoost, and DT were obtained around 0.76. ANN, on the other hand, provided the worst R^2 value of 0.72. It was observed that the R^2 values among all algorithms ranged from 0.8966 to 0.7210. It can be said that the performance metrics also exhibit similar correlations amongst themselves, akin to those of R^2 and MAPE. This information is summarized in Table 2.

One of the most notable findings observed from the performance metrics is the effectiveness demonstrated by the ANN algorithm. Despite generally outperforming other models in efficiency comparisons in various studies, ANN models exhibited the worst performance in this research dataset. While the determination coefficient value is at an acceptable level, the mean absolute percentage error and Nash-Sutcliffe efficiency values are notably poor. Tuning parameters were applied with different numerical values in all models, including the ANN model. Additionally, grid search and random search hyperparameter optimization were conducted for this algorithm. However, it still did not yield better prediction results compared to other models. The variable relationships and specific characteristics within the datasets have significantly impaired the metrics of the ANN algorithm. Nevertheless, efforts were made to achieve the best results with this model as well. In this research, the SVM model emerged as the most efficient model. The determination coefficient and nash-sutcliffe efficiency values are consistent and satisfactory. The algorithm produced mean absolute percentage error, mean absolute error, and root mean square error values at the lowest levels as expected. The performance of the RF model is also satisfactory While its performance value is at an acceptable level, it ranks as the second-best model when the metric values are relatively compared. The XGBOOST, KNN, and DT models exhibit average efficiency levels and demonstrate approximately the same performance. The scatter plot graph with the test data of the models is presented in Figure 5. The graph displays the predicted values from the predicted data on the x-axis and the actual values on the y-axis, along with the corresponding R-squared values. This visualization offers a visual analysis of how close the actual values are to the predicted values. In the SVM model, which has the best prediction performance, the data points on the scatter plot exhibit a certain pattern, indicating a mostly regular structure. Following this model, the RF, XGBOOST, and KNN models also demonstrate relatively good data distributions. The DT model shows its algorithmic structure on the graph, while the ANN graph fails to achieve the desired uniformity. In the scatter plot of ANN, the data points tend to concentrate at certain points, often displaying uncertain distributions. This algorithm does not demonstrate the expected efficiency in this study and with this dataset. In the data preprocessing section, it was mentioned that initially there were 15 variables used in the study, but through various analyses, it was reduced to 6 variables. In the analysis of variables, correlation matrix, feature importance, and model results were found to be effective. In the correlation matrix (Fig.4), some data were found to be less relevant to the target variable, and the author emphasized the need for the correlation coefficient to be at least 0.50. Although the 0.50 threshold is not definitive criterion, it is generally preferred as an approach. Additionally, the feature importance of the dataset was also examined. Features may vary depending on the model type and dataset, but a balanced distribution is expected while paying attention to scales. In this study, feature importance plots, which are outputs of the RF method, were prepared for different scenarios that could be crucial in predictions and contribute to understanding their effects. Various combinations of correlation matrix and feature importance plots were tried. However, in determining the variables and obtaining the final results, the variables, which were reduced through correlation matrix-

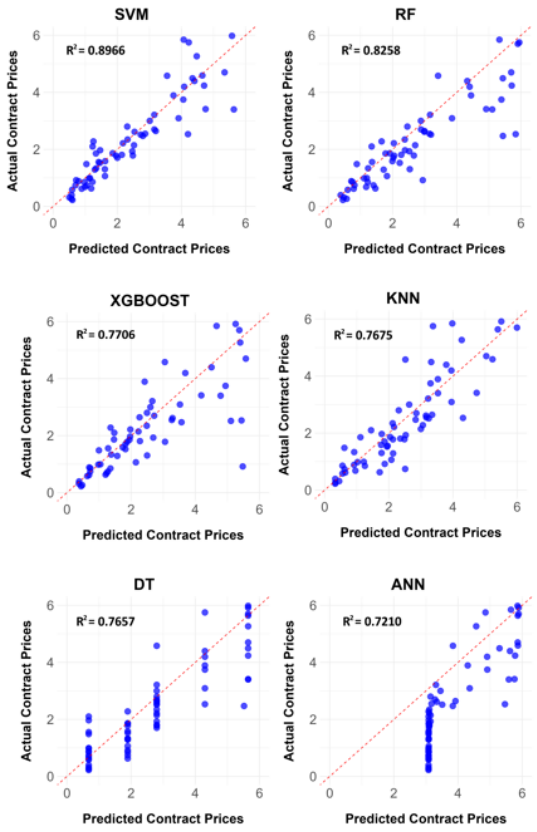

Fig 5. The scatter plots of the models

feature importance analyses, were analyzed individually and with different numbers of variables using machine

learning models. As a result, the feature importance presented in Figure 6 was obtained. According to this, the total floor area of the building and the number of beams are the most important variables, followed by the number of columns and shear walls. Although the lot area and number of building floors have the least impact, the analyses conducted showed that they contributed to the model's performance metrics to some extent.

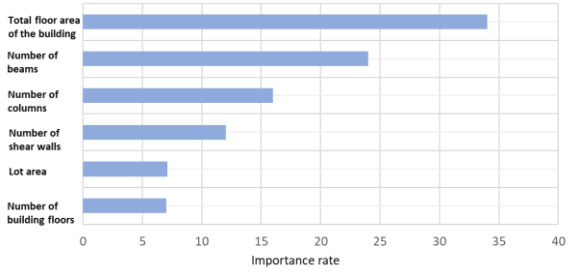

Fig. 6. Feature importance

Conclusion

This study aims to predict the contract prices signed between construction contractors and public authorities in construction tenders. Publicly owned reinforced concrete structures are being investigated for their potential cost in tender. For this purpose, 353 construction projects have been examined, and structural parameters have been collected. Structural parameters include the number of columns, shear walls, beams on all floors, number of building floors, lot area, and total floor area. Each of these parameters has been obtained individually from 353 previously tendered projects. The meaning of each input parameter, correlation matrix, feature importance, and model analyses have been explored. Initially, with 15 variables, the number of variables was reduced to 6 through various data analyses and relationship investigations. All variables have been addressed in the article, considering both samples in the literature and possibilities that could affect the contract price under market conditions.

The incorporation of variables like the number of columns, shear walls, and beams from structural project parameters can be regarded as an innovative illustration for machine learning models employed in public procurement, rendering this study distinctive in this regard. The impact of these variables on the model has been a focal point of the investigation. The quantities of structural elements and the total floor area emerge as dominant features in the machine learning model. Furthermore, it has been found that lot area and building coefficient contribute relatively less to enhancing the predictive model performance. Overall, the study achieves optimal efficiency with all variables considered.

The prediction of contract prices in public procurement utilized six popular algorithms from structural parameters: SVM, ANN, KNN, DT, RF, and XGBoost. When creating the models, a preliminary assessment of the data was conducted in the study background. Outliers were analyzed using quartile-based analysis. Various data scenarios were explored to determine the most appropriate handling of outliers. The initial dataset of 359 entries was reduced to 353 after preprocessing. Different test-train ratios and parameter optimizations were investigated in the data preprocessing phase. Cleaning and filtering the data, as well as exploring different test-train applications, have improved the model performance in the research. Evaluation of the six prediction models utilized metrics such as R^2, MAPE, NSE, MAE, and RMSE. Among the machine learning models, the SVM algorithm achieved an R^2 value of 0.8966, while the RF algorithm achieved an R^2 value of 0.8258, making them the most efficient algorithms. While other algorithms showed average performance, they still provided viable solutions. Detailed performance metrics and variable assessments are presented in the paper.

Finally, this study utilized original structural project parameters for contract price estimation in public procurement, which were evaluated and optimized through multiple stages. Different machine learning algorithms were employed and assessed at various stages, with SVM emerging as the most efficient algorithm. Satisfactory performance values were attained, and the model procedures were successfully executed.

Ethics committee approval and conflict of interest statement

There is no need to obtain permission from the ethics committee for the article prepared.

There is no conflict of interest with any person / institution in the article prepared. Acknowledgement

References

[1] S.E. Aslay, T. Dede, 3D cost optimization of 3 story RC constructional building using Jaya algorithm, Structures 40 (2022) 803–811. https://doi.org/10.1016/j.istruc.2022.04.055.

[2] B. Wang, J. Yuan, K.Z. Ghafoor, Research on Construction Cost Estimation Based on Artificial Intelligence Technology, Scalable Computing 22 (2021) 93–104. https://doi.org/10.12694:/scpe.v22i2.1868.

[3] D. Chakraborty, H. Elhegazy, H. Elzarka, L. Gutierrez, A novel construction cost prediction model using hybrid natural and light gradient boosting, Advanced Engineering Informatics 46 (2020). https://doi.org/10.1016/j.aei.2020.101201.

[4] J.A. Ujong, E.M. Mbadike, G.U. Alaneme, Prediction of cost and duration of building construction using artificial neural network, Asian Journal of Civil Engineering 23 (2022) 1117–1139. https://doi.org/10.1007/s42107-022-00474-4.

[5] A.M. Alsugair, K.S. Al-Gahtani, N.M. Alsanabani, A.A. Alabduljabbar, A.S. Almohsen, Artificial Neural Network Model to Predict Final Construction Contract Duration, Applied Sciences (Switzerland) 13 (2023). https://doi.org/10.3390/app13148078.

[6] S. Saeidlou, N. Ghadiminia, A construction cost estimation framework using DNN and validation unit, Building Research and Information (2023). https://doi.org/10.1080/09613218.2023.2196388.

[7] D. Car-Pusic, S. Petruseva, V. Zileska Pancovska, Z. Zafirovski, Neural Network-Based Model for Predicting Preliminary Construction Cost as Part of Cost Predicting System, Advances in Civil Engineering 2020 (2020). https://doi.org/10.1155/2020/8886170.

[8] T.Q.D. Pham, T. Le-Hong, X. V. Tran, Efficient estimation and optimization of building costs using machine learning, International Journal of Construction Management 23 (2023) 909–921. https://doi.org/10.1080/15623599.2021.1943630.

[9] Y. Zhang, S. Fang, RSVRs based on Feature Extraction: A Novel Method for Prediction of Construction Projects' Costs, KSCE Journal of Civil Engineering 23 (2019) 1436–1441. https://doi.org/10.1007/s12205-019-0336-3.

[10] M. Badawy, A hybrid approach for a cost estimate of residential buildings in Egypt at the early stage, Asian Journal of Civil Engineering 21 (2020) 763–774. https://doi.org/10.1007/s42107-020-00237-z.

[11] G.H. Coffie, C.O. Aigbavboa, W.D. Thwala, Modelling construction completion cost in Ghana public sector building projects, Asian Journal of Civil Engineering 20 (2019) 1063–1070. https://doi.org/10.1007/s42107-019-00165-7.

[12] S. Hassim, R. Muniandy, A.H. Alias, P. Abdullah, Construction tender price estimation standardization (TPES) in Malaysia: Modeling using fuzzy neural network, Engineering, Construction and Architectural Management 25 (2018) 443–457. https://doi.org/10.1108/ECAM-09-2016-0215.

[13] M. Sayed, M. Abdel-Hamid, K. El-Dash, Improving cost estimation in construction projects, International Journal of Construction Management 23 (2023) 135–143. https://doi.org/10.1080/15623599.2020.1853657.

[14] N. Dang-Trinh, P. Duc-Thang, T. Nguyen-Ngoc Cuong, T. Duc-Hoc, Machine learning models for estimating preliminary factory construction cost: case study in Southern Vietnam, International Journal of Construction Management 23 (2023) 2879–2887. https://doi.org/10.1080/15623599.2022.2106043.

[15] F. Uysal, R. Sonmez, Bootstrap Aggregated Case-Based Reasoning Method for Conceptual Cost Estimation, Buildings 13 (2023). https://doi.org/10.3390/buildings13030651.

[16] Z.H. Ali, A.M. Burhan, M. Kassim, Z. Al-Khafaji, Developing an Integrative Data Intelligence Model for Construction Cost Estimation, Complexity 2022 (2022). https://doi.org/10.1155/2022/4285328.

[17] W. Alfaggi, S. Naimi, An Optimal Cost Estimation Practices of Fuzzy AHP for Building Construction Projects in Libya, Civil Engineering Journal (Iran) 8 (2022) 1194–1204. https://doi.org/10.28991/CEJ-2022-08-06-08.

[18] R. Wang, V. Asghari, C.M. Cheung, S.C. Hsu, C.J. Lee, Assessing effects of economic factors on construction cost estimation using deep neural networks, Autom Constr 134 (2022). https://doi.org/10.1016/j.autcon.2021.104080.

[19] Z.H. Ali, A.M. Burhan, Hybrid machine learning approach for construction cost estimation: an evaluation of extreme gradient boosting model, Asian Journal of Civil Engineering 24 (2023) 2427–2442. https://doi.org/10.1007/s42107-023-00651-z.

[20] Y. Elfahham, Estimation and prediction of construction cost index using neural networks, time series, and regression, Alexandria Engineering Journal 58 (2019) 499–506. https://doi.org/10.1016/j.aej.2019.05.002.

[21] F. Antoniou, G. Aretoulis, D. Giannoulakis, D. Konstantinidis, Cost and Material Quantities Prediction Models for the Construction of Underground Metro Stations, Buildings 13 (2023). https://doi.org/10.3390/buildings13020382.

[22] B. Mohamed, O. Moselhi, Conceptual estimation of construction duration and cost of public highway project, Journal of Information Technology in Construction 27 (2022) 595–618. https://doi.org/10.36680/j.itcon.2022.029.

[23] A. Mahmoodzadeh, H.R. Nejati, M. Mohammadi, Optimized machine learning modelling for predicting the construction cost and duration of tunnelling projects, Autom Constr 139 (2022). https://doi.org/10.1016/j.autcon.2022.104305.

[24] M. Kovacevic, N. Ivaniševic, P. Petronijevic, V. Despotovic, Construction cost estimation of reinforced and prestressed concrete bridges using machine learning, Gradjevinar 73 (2021) 1–13. https://doi.org/10.14256/JCE.2738.2019.

[25] K. Koc, A.P. Gurgun, Causal Relationships of Readability Risks in Construction Contracts, Teknik Dergi/Technical Journal of Turkish Chamber of Civil Engineers 33 (2022) 11823–11846. https://doi.org/10.18400/tekderg.962928.

[26] S.E. Aslay, T. Dede, Reduce the construction cost of a 7-story RC public building with metaheuristic algorithms, Architectural Engineering and Design Management

(2023). https://doi.org/10.1080/17452007.2023.2195612.

[27] A. Asuncion, D. Newman, UCI machine learning repository, (2007).

[28] A. Mohammed, B. Alshemosi, H. Saad, H. Alsaad, Cost Estimation Process for Construction Residential Projects by Using Multifactor Linear Regression Technique, International Journal of Science and Research 6 (2015) 2319–7064. https://doi.org/10.21275/ART20174128.

[29] M.H. Rafiei, H. Adeli, Novel Machine-Learning Model for Estimating Construction Costs Considering Economic Variables and Indexes, J Constr Eng Manag 144 (2018). https://doi.org/10.1061/(asce)co.1943-7862.0001570.

[30] Elektronik Kamu Alımları Platformu, https://ekap.kik.gov.tr/EKAP/Ortak/IhaleArama/index.html, (n.d.).

[31] M.H. Rafiei, H. Adeli, Novel Machine-Learning Model for Estimating Construction Costs Considering Economic Variables and Indexes, J Constr Eng Manag 144 (2018). https://doi.org/10.1061/(asce)co.1943-7862.0001570.

[32] S. Saeidlou, N. Ghadiminia, A construction cost estimation framework using DNN and validation unit, Building Research and Information 52 (2024) 38–48. https://doi.org/10.1080/09613218.2023.2196388.

[33] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer New York, New York, NY, 1995. https://doi.org/10.1007/978-1-4757-2440-0.

[34] K. Koc, Ö. Ekmekcioğlu, A.P. Gurgun, Accident prediction in construction using hybrid wavelet-machine learning, Autom Constr 133 (2022). https://doi.org/10.1016/j.autcon.2021.103987.

[35] O.M. Katipoğlu, Predicting hydrological droughts using ERA 5 reanalysis data and wavelet-based soft computing techniques, Environ Earth Sci 82 (2023). https://doi.org/10.1007/s12665-023-11280-9.

[36] Ş. Emeç, D. Tekin, Housing Demand Forecasting with Machine Learning Methods, Erzincan University Journal of Science and Technology 15 (2022) 36–52. https://doi.org/10.18185/erzifbed.1199535.

[37] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification And Regression Trees, Routledge, 1984. https://doi.org/10.1201/9781315139470.

[38] Emre Mumyakmaz, Prediction of reinforced concrete column capacitties by machine learning, Master Thesis, ESKİŞEHİR TECHNICAL UNIVERSITY, INSTUTE OF GRADUATE PROGRAMS, 2023.

[39] L. Breiman, Random Forests, Mach Learn 45 (2001) 5–32. https://doi.org/10.1023/A:1010933404324.

[40] R. Özdemir, M. Turanlı, Comparison of machine learning classification algorithms for purchasing forecast, Jouurnal of Life Economics 8 (2021) 59–68. https://doi.org/10.15637/jlecon.8.1.06.

[41] T. Chen, C. Guestrin, XGBoost, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2016: pp. 785–794. https://doi.org/10.1145/2939672.2939785.

[42] E.E. Başakın, Ö. Ekmekcioğlu, M. Özger, Developing a novel approach for missing data imputation of solar radiation: A hybrid differential evolution algorithm based eXtreme gradient boosting model, Energy Convers Manag 280 (2023). https://doi.org/10.1016/j.enconman.2023.116780.

[43] Muhammet Emir Kılıç, Estimation and performance analysis of rough construction costs with machine learning methods at the pre-design stage of Induustrial buildings, Master Thesis, 2021.

[44] Evelyn Fix, Joseph Lawson Hodges, Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties, Technical Report 4, USAF School of Aviation Medicine, Randolph Field. (1951).

[45] Vehbi Hakan Sayan, Football Player Performance Analysis Using Machine Learning Techniques, Master Thesis, Burdur University, 2023.

[46] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull Math Biophys 5 (1943) 115–133. https://doi.org/10.1007/BF02478259.

[47] E.E. Başakın, Ö. Ekmekcioğlu, H. Çıtakoğlu, M. Özger, A new insight to the wind speed forecasting: robust multi-stage ensemble soft computing approach based on pre-processing uncertainty assessment, Neural Comput Appl 34 (2022) 783–812. https://doi.org/10.1007/s00521-021-06424-6.