RESEARCH ARTICLE OPEN ACCESS

Predictive Modeling for Cost and Duration Estimation in Residential Construction Projects Using Machine Learning Algorithms

Kelly Clement C. De Guzman [1],Patricia Anne L. Fernando[2],Emmanuel Aldrin R. Garcia[3], Patrick Kleyn M. Gegante[4],John Paul T. Genuino[5]

[1]Department of Civil Engineering, Don Honorio Ventura State University, Villa de Bacolor, Pampanga, Philippines
Email: dkellyclement@gmail.com

[2]Department of Civil Engineering, Don Honorio Ventura State University, Villa de Bacolor, Pampanga, Philippines
Email: patriciaannefernando@gmail.com

[3]Department of Civil Engineering, Don Honorio Ventura State University, Villa de Bacolor, Pampanga, Philippines
Email: emmanuelaldringarcia@gmail.com

[4]Department of Civil Engineering, Don Honorio Ventura State University, Villa de Bacolor, Pampanga, Philippines
Email: patrickkleyng@gmail.com

[5]Department of Civil Engineering, Don Honorio Ventura State University, Villa de Bacolor, Pampanga, Philippines
Email: johnpaulgenuino01@gmail.com

--*************************----------------------------------

Abstract:

This study explores the application of machine learning algorithms in predicting cost and duration of standard residential projects. The researchers considered eight (8) parameters, and these are gathered from different construction companies in Pampanga. A total of 60 projects were gathered. Using the most optimal data splitting method, 42 of these projects were used for training the predictive model, and 18 were used for testing. Python and visual studio code were used in creating the predictive model. Pearson of correlation coefficient (r) was used to measure the correlation of the variables. Six (6) resulted to have strong positive correlation. The other two (2) were categorized as moderate positive correlation. The R-squared values of the regression models were tested to acquire the most optimal model. The Ridge regression model resulted in the highest r squared value of 0.81. To test the accuracy of the forecasted results, mean absolute percentage error (MAPE) was used. A MAPE value of 7.20 was obtained from the cost model, and 6.21 for duration model. These MAPE values indicate that the predictive model is considered "highly accurate forecasting".

Keywords **—machine learning, regression model, predictive modeling, python, correlation**

--*************************----------------------------------

I. INTRODUCTION

Cost estimation and duration estimation are both vital in construction and project management. These two must be done carefully and accurately to avoid sudden errors. With this, it is also a fact that errors can occur during manual estimation due to complexities [1].

The development of machine learning (ML) algorithms in recent years has offered a viable way to improve the precision and effectiveness of cost prediction in residential construction projects. The goal of Machine Learning is to handle uncertainty and to handle complexity [2]. As a concrete proof, large-scale data analysis, pattern recognition, and prediction based on learnt patterns have all been shown to be possible with machine learning algorithms [3], [4].

Several studies [5], [6], [7], [8] demonstrate how machine learning could possibly be used to predict the cost and duration of construction projects. And as part of the objective of the study, Machine Learning is applied to explore and investigate the potential and power of ML in estimating the cost and duration of Residential Projects.

A. Review of Related Literature

Two studies [9],[10] published at year 2022 and one study [11] from year 2019 stated that non-linear optimization has a large range of real applications in various fields and domains. In the first study [9], it is highlighted that non-linear optimization has a wide range of real-world applications that includes engineering design, control, economic planning and data fitting. Related this [9], comprehensive research has been conducted to try and improve the original PSO algorithm by employing a variety of methods such as manipulating the PSO parameters. [10]. In [12] the researchers used hybridization to create a predictive model. In [13], multifactor linear regression has been examined in predicting cost in residential projects achieving a 92% accuracy.

Several References proved the accuracy of algorithms [14]–19]. In a study [14], Lasso regression is applied to utilize regularization to avoid overfitting. Lasso regression has advantages in terms of feature selection, and may enhance the usability of predictive models [15]. According to Li and Wang (2023), AdaBoost regression is as effective as lasso when it comes to financial forecasting models since it improves its prediction accuracy [16]. Aside from these, Ridge regression is useful also for managing multicollinearity and enhancing the generalizability of regression models [15]. Recent studies demonstrated the usefulness of a method of ensemble learning that creates a set of decisions in a sequential order [17]. Johnson and Patel's (2023) research highlight the durability and adaptability of random forest methods in big datasets [18]. As powerful as the others, K-nearest neighbors which is a straightforward and instance-based learning strategy is used in regression situations. It estimates the value of a new data point by getting the mean of the values of its k closest neighbours in the training dataset. [19].

In this study [20] the researchers utilized Back Propagation (BP) neural network in forecasting construction project cost in combination with PSO. The researchers utilize PSO in guiding the parameters of the BP network, to optimize the convergence speed. According to [21]

The Department of Public Works and Highways (DPWH) Region XI's 50 completed road projects from 2017 to 2020 served as the historical data for two studies [22], [23], which employed an Artificial Neural Network (ANN) to forecast the construction costs of the projects. proposed an early model in 2012 that used multiple regression analysis (MLR) and neural networks to estimate the costs of hydroelectric power facilities. Neural Power software was used in this work.

In 2016, a study [24] tested the accuracy Particle Swarm Optimization (PSO) in predicting cost contingency on transportation construction projects. With this, in every project in the construction industry, the cost estimation is highly anticipated to be accurate as this will be one of the basis for a project to be successful [25], [26]. In 2013, the estimation of construction costs of school buildings was compared by using three estimating strategies, namely, regression analysis, neural network, and vector machine support. Historical data was used to compare the accuracy of the three techniques. This historical data was distributed randomly into a number of 20 test data, 67 for the cross-validation data, and the remaining 130 for the training data. The performance of each technique was measured by the Mean Absolute Error Rates (MAERs). Three algorithms were used and resulted to a standard deviation of 3.56, 4.13, and 4.66. Among these, Neural Network found to be the most suitable for school building projects' cost estimation since it produced more accurate estimation results than the two [26].

This article [27] introduces an approach to predict the expenses involved in road construction by utilizing machine learning (ML) models. Based on the analysis conducted in this study, it is evident that the Random Forest (RF) regression model outperforms others in estimating road construction costs displaying precision, with R2 = 1.0, Mean Absolute Percentage Error (MAPE) = 0.00, and Root Mean Squared Error (RMSE) = 0.00. Another article [28] focuses on forecasting the construction cost index (CCI) for building materials in some countries by using methods such as Artificial Neural Networks (ANN). Its primary objective is to contribute to cost predictions, specifically during the approval and financial planning stages of construction projects.

In the systematic review of [2], the results indicate that, of the different approaches that researchers have employed (ANN, Fuzzy NN, SVM, PSO, RBF, RA, CBR, PSO, Decision Tree, AHP, Monte Carlo, fuzzy logic), ANN and RA have been the most widely utilized machine learning techniques in the examined publications. It is in fact that algorithms can be employed and used in any aspect.

B. Synthesis

Machine Learning algorithms are trained with large amounts of dataset to create a forecasting model [2]. Studies [8]-[18] have developed various predictive models using ML algorithms. Combining optimization algorithms such as GA in [12] and PSO in [20] with Neural Networks has shown improved results when compared to Neural Networks used alone. This is further supported by [2] which has reached similar conclusion. In the training phase, recent studies such as [12], [20], [21], [22], [25], [26], [27], [28] have used MATLAB, while earlier studies such as [23], [24] used Neuro Power and NeuroShell respectively. Furthermore, the ratio of training, validation, and testing data of these studies are as follows:

1. 90% training, 10% testing for [27]
2. 80% training, 20% testing for [20], [21], [22]
3. 60% training, 20% validation, 20% testing for [24] and
4. 60% training 30% validation, 10% training for [25], [26].

Studies [8]-[18] utilized either of the following statistical tools; RMSE, MSE, MAPE, and Correlation Coefficient (R) in evaluating the validity and accuracy of the trained model. These studies indicate great potential for the use of machine learning algorithms in creating predictive models for construction costs.

C. Gap Analysis

This study focused on developing predictive models using machine learning algorithms to estimate costs and durations in residential construction projects. Several gaps in existing literature motivated this study:

1. The optimal data splitting technique is not well established in literature. While studies such as [20], [21], [22], [23], [24], [26], [27] have demonstrated great results, it is still unclear if the data splitting techniques employed in their studies are the most optimal.
2. In the systematic review [2], it has been shown that cost and duration estimation in residential projects are still underexplored.
3. There seems to be a need for an interpretable and explainable predictive model. Machine learning algorithms, such as neural networks in [4], [12], [20], [20], [22], [23], [26], can sometimes be perceived as black boxes, lacking transparency in decision-making as well as data training [2].

D. Objectives of the Study

The primary objective of this research is to prove the accuracy of Machine Learning Algorithms in predicting the cost and duration of residential projects. By this, it aims to bring forecasted cost and duration closer to actual values.

- To create a systematic and predictive modelling using Machine Learning Algorithms.
- To estimate the cost and duration of residential projects using Machine Learning Algorithm.
- To develop a web application that employs the trained model.

II. METHODS

In this chapter, the methodological framework was discussed along with the research design, and the data collection methods that were used for data analysis.

A. Research Design

The study was conducted using an experimental type of research leaning towards quantitative methods. The data collection strategies that were used in this study are divided primarily into two experimental groups. The first group is the actual results based on the selected 60 residential structures (existing). The second group are the results from predictive modelling.

B. Data Collection Method

The researchers gathered all the needed data from different residential projects. As a reference to the program, input data were selected randomly from construction companies. Actual Concrete Volume, Wall or Brick Volume, Floor Numbers, Ground Floor Area and Total Floor Areas were adopted along with Actual Cost and Duration of the residential projects in building the proposed model.

The input parameters are the concrete volume, wall volume, floor numbers, ground floor area, total floor areas and type of footing. The output parameters are the cost and duration. The input data are independent variables while the output data are dependent. As shown in Table I: Definition of Input Parameters presents the parameter symbol and definition. Concrete Volume, Wall Volume, Floor Numbers, Ground Floor Area, and Total Floor Areas falls under Quantitative Features (Numerical). Only the Footing Type is categorized as qualitative (Categorical). Table II: Definition of Output Parameters presents the variable representation of the cost and duration model in the creation of the program.

TABLE I

Symbol	Description	Definition
C	Concrete Volume	This refers to the total concrete volume from foundations, footings, columns, beams, and slabs.
B	Wall/Brick Volume	This corresponds to the total wall area multiplied by its thickness.
FT	Type of Footing	This pertains to the type of footing used in the construction process.
AGF	Area of the Ground Floor	This corresponds to the total floor area that is leveled to the ground.
TFA	Total Area of Floors	This is computed by adding the area of each floor (including the ground floor).
FN	Number of Floors	This refers to the total number of floors in the residential project.

DEFINITION OF INPUT PARAMETERS

TABLE II
DEFINITION OF OUTPUT PARAMETERS

Symbol	Description	Definition
y	Total Cost	This refers to the total cost of the entire project.
z	Total Duration	This corresponds to the total duration of the construction phase of the entire project.

For building the proposed model shown in Figure 1 illustrates the System Flowchart of the study. The researchers collected data from various companies with existing residential projects. The data were then sorted as they went through data splitting, for training and testing. A Hold-out method has been used in the data-splitting process. For data training, 42 projects were used, and the input variables were denoted as C, B, FT, AGF, TFA, and FN while Cost and Duration were y and z respectively. Various types of machine learning algorithm were utilized for training and testing; thus, Ridge Regression governs in terms of R-squared values. In the trained model, 18 data from 70-30 split, and 12 data from 80-20 split, undergo testing in which predicted cost and duration, alongside with their actual cost and duration were analyzed.

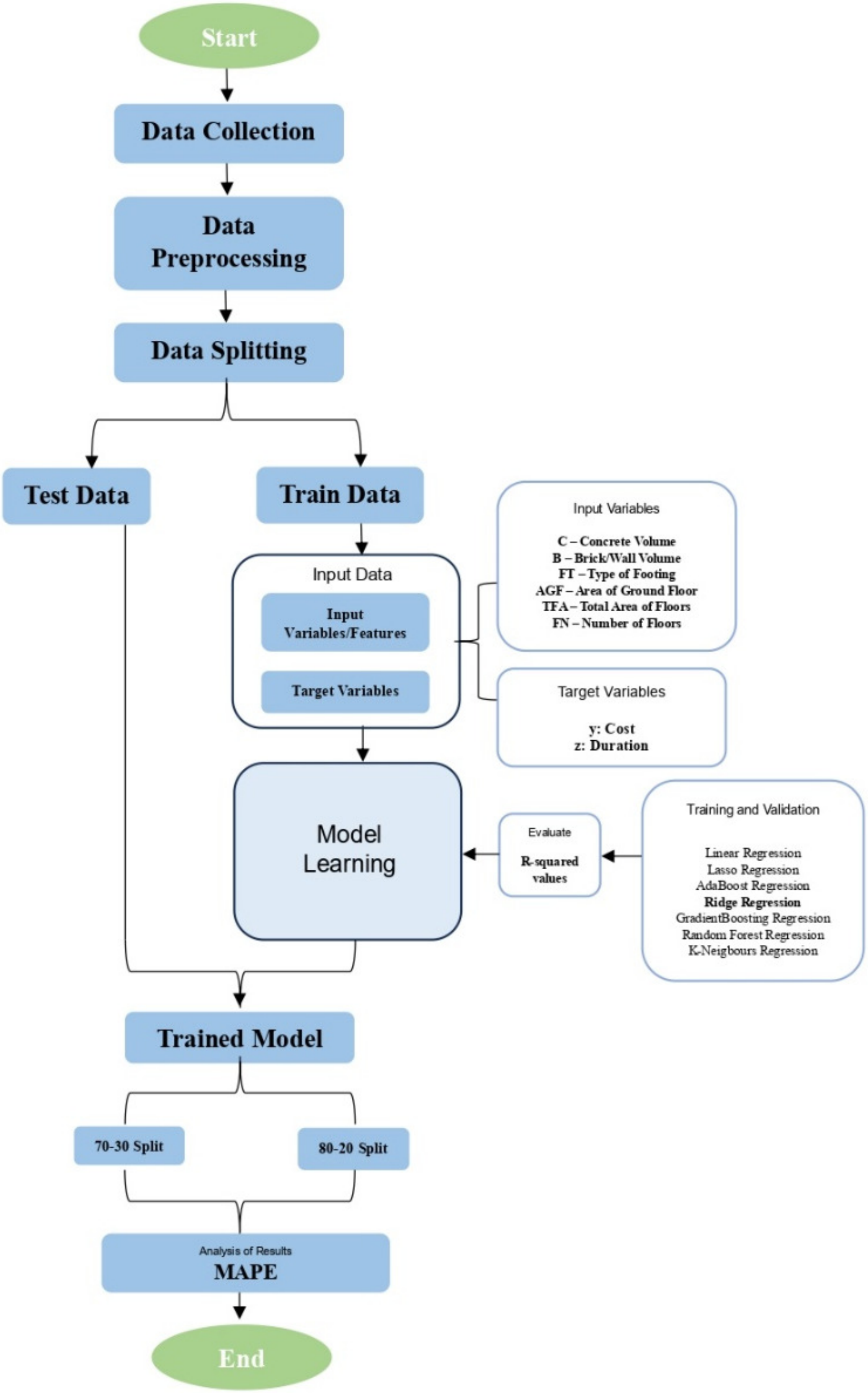

FIG. 1 SYSTEM FLOW CHART

C. Data Analysis

This stage involves analyzing the data gathered using Pearson correlation coefficient (r) and standard distribution values. In this stage, the predicted results of the trained model are evaluated using Mean Absolute Percentage Error (MAPE).

Mean

The mean, often referred to as the average, is a fundamental statistical measure used extensively in experimental research [30].The mean can be calculated by Equation 1.

$$\underline{X} = \frac{\sum X}{N}$$

(Eq. 1)

Standard Deviation

Standard deviation is a key statistical measure widely used in experimental research. It provides valuable insights into the dispersion, variability, and consistency of data within a sample or experimental condition [31]. The standard deviation can be calculated by Equation 2.

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

(Eq. 2)

Correlational Coefficient

In experimental research, the correlational coefficient, often represented by the Pearson's R (r), is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables [33].The correlational coefficient can be calculated as shown in Equation 3.

$$r = \frac{\sum\left(x_i - \underline{x}\right)\left(y_i - \underline{y}\right)}{\sqrt{\sum\left(x_i - \underline{x}\right)^2 \sum\left(y_i - \underline{y}\right)^2}}$$

(Eq. 3)

As shown in Table III: Basis for Person R (Turney, S., 2022) [33], is the correlation type, interpretation, strength and direction.

TABLE III
BASIS FOR PEARSON R

Pearson Correlation coefficient (*r*)	Correlation type	Interpretation
Between 0-1	Positive Correlation	When one variable changes, the other variable changes in the **same direction**.
0	No Correlation	There is **no relationship** between the variables.
Between 0 and -1	Negative Correlation	When one variable changes, the other variable changes in the **opposite direction**.

r	Strength	Direction
>0.5	Strong	Positive
0.3 to 0.5	Moderate	Positive
0 to 0.3	Weak	Positive
0	None	None
0 to -0.3	Weak	Negative
-0.3 to -0.5	Moderate	Negative

Coefficient of Determination

The coefficient of determination (R2) is used to measure the extent in which variables can produce variances in another. Its importancefalls under simple but better understanding of data set.Zero valuemeans that the model does not enhance predictions upon the mean model, while a value of one indicates perfect prediction accuracy [32]. The coefficient of determination can be calculated as shown in Equation 4.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

(Eq. 4)

Mean Absolute Percentage Error

Calculations of mean absolute percentage error (MAPE) can also indicate whether forecasting is accurate, enabling the company to communicate forecasting outcomes more effectively to investors and execute strategy across multiple departments.

$$MAPE = \frac{1}{n} \times \sum \left|\frac{actual\ value\ -\ forecast\ value}{actual\ value}\right|$$

[34].The MAPE can be calculated as shown in Equation 5.

(Eq. 5)

The mean absolute percentage error (MAPE) stands out as one of the most commonly utilized metrics for evaluating forecasting accuracy [35]. In line with the National Research Council's findings from 1980, any condensed representation of error must adhere to five fundamental standards: validity, reliability, ease of interpretation, clarity of presentation, and support of statistical evaluation [36]. MAPE adequately fulfills four of the mentioned criteria, yet its adequacy in addressing the validity criterion diminishes when employed for evaluating the precision of population forecasts [37]. Lewis (1982) [38] compiled a table containing MAPE values as shown in Table IV: MAPE Interpretation.

TABLE IV
MAPE INTERPRETATION

MAPE	Interpretation
<10	Highly accurate forecasting
10-20	Good forecasting
20-50	Reasonable forecasting
>50	Inaccurate forecasting

III. RESULTS AND DISCUSSION

This part shows the results and discussion of the study. Following the 70%-30% ratio, the program was developed using Python Programming Language. The model was proposed to examine its influence on outcomes. The main duty of the utilized algorithm is to reduce the difference between the predicted and actual cost and/or duration.

A. Result

As shown in the Figure 2, a correlation heatmap is a visual representation of the correlation matrix, which shows the correlation coefficients between variables in a dataset. Interpreting a correlation heatmap is crucial for understanding the relationships between variables and identifying patterns or trends within the data. Concrete and brick volumes have achieved 0.72 and 0.76 R-values, indicating a good correlation with the cost. Concrete volume, total floor area, and cost have resulted in 0.79, 0.76, and 0.75 R-values respectively indicating a good correlation with the duration.

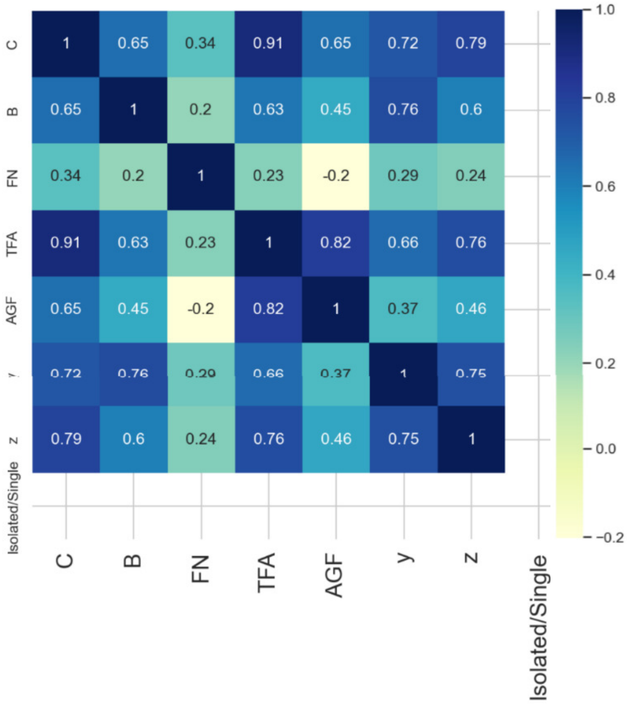

FIG. 2 CORRELATION HEATMAP

As shown in the Figure 3, the R2 score of the ridge regression was 0.82 indicating that it explained 82% of the variance in the data. This is the highest R2 score among the tested models, suggesting that the ridge regression was the best fit for the data while the linear regression for duration has the lowest R2 score of 0.68 explaining only 68% of the variance. This implies that the linear regression for duration may not provide the best results compared to other models. The analysis showed that linear regression for cost, linear regression, AdaBoost, random forest, and K-nearest neighbors all had similar high R2 scores ranging from 0.79 to 0.81 indicating a good fit for the data. In contrast, the gradient boosting regression had a lower R2 score of 0.67, suggesting a less optimal fit. Overall, the results highlight that the ridge regression performed the best with an R2 score close to 1 indicating a strong fit to the data and explaining a significant portion of the variance.

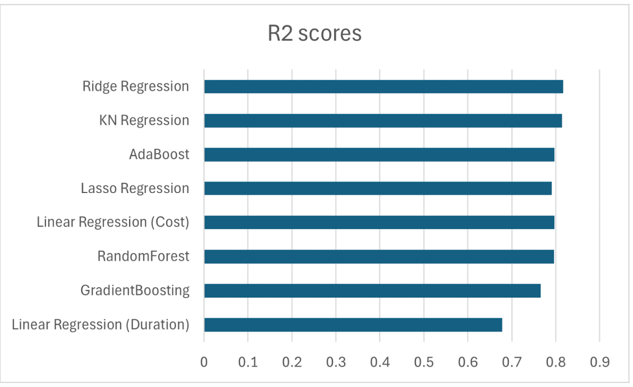

FIG. 3 R-SQUARED SCORES

Table V illustrates the absolute percentage error of the actual and forecasted cost using 80-20 data split. Averaging the absolute percentage error yields a MAPE value of 10.037 indicating a "good forecasting".

TABLE V
COST MODEL FOR 80-20 DATA SPLIT

PROJECT NUMBER	ACTUAL COST	FORECASTED COST	PERCENT ERROR
Project 1	₱ 1,427,126.00	₱ 1,292,966.68	9.401
Project 2	₱ 7,260,567.00	₱ 7,713,489.81	6.238
Project 3	₱ 1,293,701.00	₱ 1,465,431.85	13.274
Project 4	₱ 2,361,790.00	₱ 2,504,181.51	6.029
Project 5	₱ 2,446,476.00	₱ 2,510,249.19	2.607
Project 6	₱ 6,519,050.00	₱ 5,660,466.06	13.170
Project 7	₱ 10,087,765.00	₱10,754,191.63	6.606
Project 8	₱ 5,087,765.00	₱ 5,176,806.11	1.750
Project 9	₱ 6,087,765.00	₱ 6,008,515.57	1.302
Project 10	₱ 1,672,354.00	₱ 1,027,874.44	38.537
Project 11	₱ 3,500,000.00	₱ 3,714,751.22	6.136
Project 12	₱ 6,000,523.33	₱ 5,077,033.76	15.390

Table VIillustrates the absolute percentage error of the actual and forecasted duration using 80-20 data split. Averaging the absolute percentage error yields a MAPE value of 4.379 indicating a "highly accurate forecasting".

TABLE VI
DURATION MODEL FOR 80-20 DATA SPLIT

PROJECT NUMBER	ACTUAL COST	PREDICTED COST	PERCENT ERROR
Project 1	123	116	5.691
Project 2	183	198	8.197
Project 3	131	121	7.634
Project 4	142	134	5.634
Project 5	140	133	5.000
Project 6	170	168	1.176
Project 7	240	236	1.667
Project 8	171	166	2.924
Project 9	177	176	0.565
Project 10	114	113	0.877
Project 11	152	148	2.632
Project 12	180	161	10.556

Table VIIillustrates the absolute percentage error of the actual and forecasted cost using 70-30 data split. Averaging the absolute percentage error yields a MAPE value of 7.220 indicating a "highly accurate forecasting".

TABLE VII
COST MODEL FOR 70-30 SPLIT

PROJECT NUMBER	ACTUAL COST	PREDICTED COST	PERCENT ERROR
Project 1	₱ 9,265,640.64	₱ 8,373,214.98	9.632
Project 2	₱ 4,224,712.89	₱ 4,387,177.50	3.846
Project 3	₱ 4,112,312.94	₱ 3,922,952.14	4.605
Project 4	₱ 3,983,112.44	₱ 4,264,350.77	7.061
Project 5	₱ 3,650,136.23	₱ 3,060,798.30	16.146
Project 6	₱ 8,115,503.13	₱ 7,199,984.38	11.281
Project 7	₱ 1,427,126.77	₱ 1,527,126.54	7.007
Project 8	₱ 7,260,567.40	₱ 7,268,571.23	0.110
Project 9	₱ 1,293,701.83	₱ 1,292,647.10	0.082
Project 10	₱ 2,543,566.56	₱ 2,361,790.60	7.146
Project 11	₱ 2,632,345.89	₱ 2,446,476.30	7.061
Project 12	₱ 6,519,050.25	₱ 5,991,069.69	8.099
Project 13	₱ 10,087,765.00	₱ 10,095,940.81	0.081
Project 14	₱ 5,087,765.00	₱ 4,985,431.05	2.011
Project 15	₱ 6,087,765.90	₱ 5,812,965.82	4.514
Project 16	₱ 1,672,354.72	₱ 1,228,440.63	26.544
Project 17	₱ 3,500,000.00	₱ 3,617,783.80	3.365
Project 18	₱ 6,000,523.33	₱ 5,318,599.98	11.364

Table VIIIillustrates the absolute percentage error of the actual and forecasted duration using 70-30 data split. Averaging the absolute percentage error yields a MAPE value of 6.210 indicating a “highly accurate forecasting”.

TABLE VIII

PROJECT NUMBER	ACTUAL DURATION	PREDICTED DURATION	PERCENT ERROR
Project 1	210	202	3.81
Project 2	177	153	13.56
Project 3	161	146	9.32
Project 4	157	152	3.18
Project 5	180	147	18.33
Project 6	218	206	5.50
Project 7	123	117	4.88
Project 8	183	187	2.19
Project 9	131	122	6.87
Project 10	142	137	3.52
Project 11	140	130	7.14
Project 12	170	167	1.76
Project 13	240	217	9.58
Project 14	171	159	7.02
Project 15	177	169	4.52
Project 16	114	113	0.88
Project 17	152	155	1.97
Project 18	180	166	7.78

DURATION MODEL FOR 70-30 DATA SPLIT

Table IXillustrates the summary of MAPE values for both splits and their interpretation based on Lewis (1982) table [38] cited in [35], [37].

TABLE IX
SUMMARY OF MAPE VALUES

MAPE Values				
Data Split	Cost Model	Interpretation	Duration Model	Interpretation
80-20	10.037	Good forecasting	4.379	Highly Accurate Forecasting
70-30	7.220	Highly Accurate Forecasting	6.210	Highly Accurate Forecasting

IV. CONCLUSIONS

The aim of this study is to develop a predictive model using Machine Learning Algorithms that will be applied to forecast the cost and duration of Residential Projects. In this paper, sixty residential projects were utilized to create the proposed model. The main conclusions are as follows:

1. By getting the mean, standard deviation and range, the parameters were proved to be normally distributed having Bell Curve Areas between 89.60 to 94.79 in percentage. The R2 score of the ridge regression was 0.82 indicating that it explained 82% of the variance in the data. In line with this, upon testing the correlation using Pearson Correlation Coefficient (r), most of the independent variables had strong positive correlation with dependent variables. Only two (2) were categorized to have moderate positive correlation. These testing were done to know the distribution, variance and relationship of the variables.
2. Using Machine Learning Algorithms specifically, Ridge Regression Model, total construction Cost and Duration were forecasted. By trying its capabilities, it is the goal of the researchers to bring the predicted values closer to the actual values.
3. Using MAPE, both cost and duration model were categorized under “high – forecasting model.”
4. The web application served as the final output of the predictive model developed from Machine Learning.

RECOMMENDATION

The following recommendations are suggested to future researchers:

1. The proponents of this study suggest adding the number of bedrooms, and types of finishes (i.e. basic, standard, and elegant) as input parameters for residential projects.
2. The proponents highly recommend testing various data splits and algorithms instead of utilizing only one.
3. Machine learning thrives in processing large datasets, this means that increasing the volume of data typically leads to higher accuracy. Therefore, future researchers may consider creating synthetic data to enhance the accuracy of the predictive model, especially in the training phase.
4. The proponents also recommend adding more data in creating the model.
5. Given the recent surge in the utilization of Artificial Intelligence and machine learning

algorithms, the proponents advocate for continued exploration of this topic.

REFERENCES

[1] U. Haider, U. Khan, A. Nazir, and M. Humayon, “Cost Comparison of a Building Project by Manual and BIM,” *Civ. Eng. J.*, vol. 6, pp. 34–49, Jan. 2020, doi: 10.28991/cej-2020-03091451.

[2] S. Tayefeh Hashemi, O. M. Ebadati, and H. Kaur, “Cost estimation and prediction in construction projects: a systematic review on machine learning techniques,” *SN Appl. Sci.*, vol. 2, no. 10, p. 1703, Sep. 2020, doi: 10.1007/s42452-020-03497-1.

[3] “Introduction to Machine Learning,” MIT Press. Accessed: Feb. 06, 2024. [Online]. Available: https://mitpress.mit.edu/9780262012119/introduction-to-machine-learning/

[4] K. Hsu, H. V. Gupta, and S. Sorooshian, “Artificial Neural Network Modeling of the Rainfall-Runoff Process,” *Water Resour. Res.*, vol. 31, no. 10, pp. 2517–2530, 1995, doi: 10.1029/95WR01955.

[5] S. Petruseva, V. Zileska Pancovska, and V. Žujo, “PREDICTING CONSTRUCTION PROJECT DURATION WITH SUPPORT VECTOR MACHINE,” *Int. J. Res. Eng. Technol.*, vol. 02, pp. 2321–7308, Nov. 2013, doi: 10.15623/ijret.2013.0211003.

[6] H.-G. Cho, K.-G. Kim, J.-Y. Kim, and G.-H. Kim, “A Comparison of Construction Cost Estimation Using Multiple Regression Analysis and Neural Network in Elementary School Project,” *J. Korea Inst. Build. Constr.*, vol. 13, Feb. 2013, doi: 10.5345/JKIBC.2013.13.1.066.

[7] T. Aung, S. Liana, A. Htet, and A. Bhaumik, “Using Machine Learning to Predict Cost Overruns in Construction Projects,” *J. Technol. Innov. Energy*, vol. 2, pp. 1–7, Jun. 2023, doi: 10.56556/jtie.v2i2.511.

[8] Y. Alzubi, A. Jaafreh, and A. Khatatbeh, “Application of machine learning techniques in estimating the construction cost of residential buildings in the Middle East region,” *Int. J. Constr. Manag.*, vol. 24, pp. 1–13, Jul. 2023, doi: 10.1080/15623599.2023.2239494.

[9] T. M. Shami, A. A. El-Saleh, M. Alswaitti, Q. Al-Tashi, M. A. Summakieh, and S. Mirjalili, “Particle Swarm Optimization: A Comprehensive Survey,” *IEEE Access*, vol. 10, pp. 10031–10061, 2022, doi: 10.1109/ACCESS.2022.3142859.

[10] B. S. C. F. Leite, “Nonlinear programming: Theory and applications,” Medium. Accessed: Nov. 20, 2023. [Online]. Available: https://towardsdatascience.com/nonlinear-programming-theory-and-applications-cfe127b6060c

[11] M. A. K. Azrag and T. Asmawaty, “Empirical Study of Segment Particle Swarm Optimization and Particle Swarm Optimization Algorithms,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 8, 2019, doi: 10.14569/IJACSA.2019.0100862.

[12] Z. Bai, G. Wei, X. Liu, and W. Zhao, “Predictive Model of Energy Cost in Steelmaking Process Based on BP Neural Network,” presented at the 2014 2nd International Conference on Software Engineering, Knowledge Engineering and Information Engineering (SEKEIE 2014)), Atlantis Press, Mar. 2014, pp. 77–80. doi: 10.2991/sekeie-14.2014.18.

[13] A. M. B. Alshemosi and H. S. H. Alsaad, “Cost Estimation Process for Construction Residential Projects by Using Multifactor Linear Regression Technique,” 2017. Accessed: Feb. 06, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Cost-Estimation-Process-for-Construction-Projects-Alshemosi-Alsaad/47422c5096c47e0a9c636e2dd46f755c8dc0b956

[14] M. R and R. Rohini, “LASSO: A feature selection technique in predictive modeling for machine learning,” Oct. 2016, pp. 18–20. doi: 10.1109/ICACA.2016.7887916.

[15] S. Safi, M. Alsheryani, M. Alrashdi, R. Suleiman, D. Awwad, and Z. Abdalla,"Optimizing Linear Regression Models with Lasso and Ridge Regression: A Study on UAE Financial Behavior during COVID-19," *Migr. Lett.*, vol. 20, pp. 139–153, Sep. 2023, doi: 10.59670/ml.v20i6.3468.

[16] V. Chang, T. Li, and Z. Zeng, "Towards an improved Adaboost algorithmic method for computational financial analysis," *J. Parallel Distrib. Comput.*, vol. 134, Aug. 2019, doi: 10.1016/j.jpdc.2019.07.014.

[17] D. Otchere, T. Ganat, J. Ojero, B. N. Tackie-Otoo, and M. Taki, "Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions," *J. Pet. Sci. Eng.*, vol. 208, Jul. 2021, doi: 10.1016/j.petrol.2021.109244.

[18] "(PDF) Research on the Application of Random Forest-based Feature Selection Algorithm in Data Mining Experiments." Accessed: Apr. 13, 2024. [Online]. Available: https://www.researchgate.net/publication/375210507_Research_on_the_Application_of_Random_Forest-based_Feature_Selection_Algorithm_in_Data_Mining_Experiments

[19] K. Kim, H. Choi, C. Moon, and C.-W. Mun, "Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions," *Curr. Appl. Phys. - CURR APPL PHYS*, vol. 11, pp. 740–745, May 2011, doi: 10.1016/j.cap.2010.11.051.

[20] D. Ye, "An Algorithm for Construction Project Cost Forecast Based on Particle Swarm Optimization-Guided BP Neural Network," *Sci. Program.*, vol. 2021, p. e4309495, Oct. 2021, doi: 10.1155/2021/4309495.

[21] T. Z. Khalaf, H. Çağlar, A. Çağlar, and A. N. Hanoon, "Particle Swarm Optimization Based Approach for Estimation of Costs and Duration of Construction Projects," *Civ. Eng. J.*, vol. 6, no. 2, pp. 384–401, Feb. 2020, doi: 10.28991/cej-2020-03091478.

[22] D. V. Gante, D. L. Silva, and M. P. Leopoldo, "Forecasting Construction Cost Using Artificial Neural Network for Road Projects in the Department of Public Works and Highways Region XI," in *Frontiers in Artificial Intelligence and Applications*, A. J. Tallón-Ballesteros, Ed., IOS Press, 2022. doi: 10.3233/FAIA220084.

[23] "An early cost estimation model for hydroelectric power plant projects using neural networks and multiple regression analysis | Journal of Civil Engineering and Management." Accessed: Feb. 06, 2024. [Online]. Available: https://journals.vilniustech.lt/index.php/JCEM/article/view/2940

[24] S. C. Lhee, R. R. A. Issa, and H. Professor, "USING PARTICLE SWARM OPTIMIZATION TO PREDICT COST CONTINGENCY ON TRANSPORTATION CONSTRUCTION PROJECTS".

[25] M. Sayed, M. Abdelhamid, and K. Eldash, "Improving cost estimation in construction projects," *Int. J. Constr. Manag.*, vol. 23, pp. 1–20, Dec. 2020, doi: 10.1080/15623599.2020.1853657.

[26] G.-H. Kim, J.-M. Shin, S. Kim, and Y. Shin, "Comparison of School Building Construction Costs Estimation Methods Using Regression Analysis, Neural Network, and Support Vector Machine," *J. Build. Constr. Plan. Res.*, vol. 01, pp. 1–7, Jan. 2013, doi: 10.4236/jbcpr.2013.11001.

[27] Y. G. Abed, T. M. Hasan, and R. N. Zehawi, "Cost Prediction for Roads Construction using Machine Learning Models," *Int. J. Electr. Comput. Eng. Syst.*, vol. 13, no. 10, Art. no. 10, Dec. 2022, doi: 10.32985/ijeces.13.10.8.

[28] "Societies | Free Full-Text | Forecasting Construction Cost Index through Artificial Intelligence." Accessed: Feb. 06, 2024. [Online]. Available: https://www.mdpi.com/2075-4698/13/10/219

[29] "The Data Analysis Handbook, Volume 14 - 1st Edition." Accessed: Feb. 11, 2024.

[Online]. Available: https://shop.elsevier.com/books/the-data-analysis-handbook/frank/978-0-444-81659-7

[30] P. Bhandari,"How to Find the Mean | Definition, Examples & Calculator," Scribbr. Accessed: Dec. 04, 2023. [Online]. Available: https://www.scribbr.com/statistics/mean/

[31] "Standard Deviation - Formula | How to Calculate Standard Deviation?" Accessed: Dec. 04, 2023. [Online]. Available: https://www.cuemath.com/data/standard-deviation/

[32] M. Wiley and J. F. Wiley, *Advanced R Statistical Programming and Data Models: Analysis, Machine Learning, and Visualization*. Apress, 2019.

[33] S. Turney, "Pearson Correlation Coefficient (r) | Guide & Examples," Scribbr. Accessed: Dec. 04, 2023. [Online]. Available: https://www.scribbr.com/statistics/pearson-correlation-coefficient/

[34] "What Is MAPE? A Guide to Mean Absolute Percentage Error | Indeed.com." Accessed: Feb. 07, 2024. [Online]. Available: https://www.indeed.com/career-advice/career-development/what-is-mape

[35] J. J. Montaño Moreno, A. Palmer Pol, A. Sesé Abad, and B. Cajal Blasco, "Using the R-MAPE index as a resistant measure of forecast accuracy," *Psicothema*, vol. 25, no. 4, pp. 500–506, 2013, doi: 10.7334/psicothema2013.23.

[36] *Read "Estimating Population and Income of Small Areas" at NAP.edu*. doi: 10.17226/19788.

[37] J. Tayman and D. Swanson, "On the Validity of MAPE as a Measure of Population Forecast Accuracy," *Popul. Res. Policy Rev.*, vol. 18, pp. 299–322, Aug. 1999, doi: 10.1023/A:1006166418051.

[38] C. D. Lewis,*Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting*. Butterworth Scientific, 1982.