Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

A Comparative Study of Machine Learning Algorithms for Early Cost Estimation of Building Projects in Nepal

Anjuli Sapkota*[,1], Samrakshya Karki[2], Bishwas Pokharel[3], Mahendra Dhital[4]

[1,4]Department of Civil Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Kathmandu, Nepal

[2]Building Design Authority, Kamalpokhari, Kathmandu, Nepal

[3]Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal

*Corresponding Author: *anjulisapkota95@gmail.com*

ABSTRACT – Construction cost estimation is crucial to a project's success, but because of the many variables that impact it, it is challenging to make an accurate prediction. Traditional methods are being used for preliminary cost estimation in the construction industry of Nepal. There still exists the problem of cost overrun, and time delay due to incorrect cost budgeting. This study aims to analyze a modern method of preliminary cost estimation in Nepal to prove its efficiency over the traditional method. In this work models such as Linear Regressor, Decision Tree Method, Random Forest method, Artificial Neural Networks, Support Vector Machine, Boost method, Extra tree method, Voting Regression, and Stacking method are used. Regarding the datasets, the buildings that were used are Educational Building, Commercial Building, Hospital Building, Residential Building, Public Building, Official Building, and Hotel Building having 0 to 2 basements ranging above 1 crore. The input features were taken from the literature review, and validated by expert opinion, and after successfully conducting pilot testing, the survey questionnaire was distributed among contractors and consultants. Data preprocessing was done and training and testing data sets were developed. The model was developed for nine algorithms. Mean absolute error (MAE), Mean square error (MSE), Root mean square error (RMSE), and R square value are used as evaluation metrics. In the evaluation of various regression models, three stand out as the most promising for predicting the target variable. The Decision Tree model exhibited remarkable performance with an MSE of 0.088575, an MAE of 0.104625, an RMSE of 0.297615, and an R^2 of 0.876170. Similarly, the Extra Tree model closely followed with an MSE of 0.088601, an MAE of 0.102909, an RMSE of 0.297659, and an R^2 of 0.876134. The Voting Model with an MSE of 0.105035, an MAE of 0.222807, an RMSE of 0.324091, and an R^2 of 0.853159. This study also opens the path for the exploration of other models and motivate to follow the trends of machine learning in the present era.

KEYWORDS– ***Construction Management, Building, Preliminary cost Estimation, Machine learning, Pilot testing, Feature reduction.***

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

1. INTRODUCTION

Nepal is developing country and the current population is 30,927,840 as of August 2023, based on World meter elaboration of the latest United Nations data. Building construction project plays a crucial part in Nepalese economy. Villages in Nepal are more likely to have adobe constructions, wooden-framed homes, and rubble stone masonry structures, while the bulk of metropolitan areas and suburbs have stone or brick masonry (Gautam, 2016). Twenty percent (20%) buildings is made up of reinforced concrete (RC). Predicting construction expenses in a reasonable manner is crucial in the early phases of a building project ((Hwang, 2017). Cost is seen as a standard indicator of the resources used on a project (Akalya, 2018). Cost estimation is a critical aspect of any construction project, as it provides an initial budgetary framework and assists stakeholders in making informed decisions (Tayefeh Hashemi, 2020). Accurate cost estimation can lead to potential cost savings and improved time efficiency during project execution, making it an attractive proposition for the construction industry in Nepal. In Nepal, like many other developing countries, construction projects often face budget overruns and delays due to inaccurate early cost estimates. Nepal's construction sector faces specific challenges such as limited resources, topographical constraints, and varying socio-economic conditions across region. Despite the growing popularity of machine learning in various industries, its application to the construction sector in Nepal has been relatively limited. Traditional cost estimation methods might not fully account for these complexities, making machine learning an appealing option to develop context-aware and more accurate cost estimation models. Quantity Rate Analysis is the primary conventional method for estimating costs that is commonly utilized (Veliyampatt).

2. LITERATURE REVIEW

2.1 Building definition and types of Building

Building is defined as any structure made of any material, whether or not it is inhabited by people, and which includes the foundation, plinth, walls, floors, roofs, and building services. Tents, tarpaulin shelter, and other transient structures are not to be regarded as building. All governmental, non-governmental, and private structures which offers the general public services, amenities, opportunities, and products are referred to as public buildings. Based on occupancy buildings are classified as Residential, Assembly, Educational, Hospitals and Clinic, Commercial, office, industries and storage. Based on Storey and height buildings are classified as General Buildings (1 to 5 Stories or below 16m), Medium Rise (6 to 8 Stories or between 16m to below 25m), High Rise (9 to 39 Stories or 25m to below 100m), Skyscrapers (40 Stories and above or above 100m) (NBC 206:2015). Nepal National Building Code (NBC) code is effectively implemented in buildings in Nepal which was formulated in 1994 ((Ching, 2020; Allen, 2019).

2.2 Factors affecting Building project

The variables affecting building project are Project Characteristics (Building Type, Number of Storeys, Number of Blocks, Project Complexity Representative Value, Programmed Duration, Original Cost Estimate), Procurement System (Functional Grouping, Payment Method, Contract Conditions), Project Team Performance (Contractor, Design Team, Management Team), Client: Client Representatives Characteristics (Client Type, Client Priority,

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

Client Source of Finance, Client Characteristics Representative Value), Contractor Characteristics (Contractor Characteristics Representatives Value), Design Team Characteristics (Design Team Characteristics Representative Value), External Conditions (External Conditions Representative Value) (Dissanayaka, 1998), (Hwang, 2017), (Mahamid, 2013), (Yap, 2020) (Mahamid, 2013))

2.3 Cost estimate and types of Cost Estimate

Cost estimate provides a general concept of the cost of the work, allowing for the determination of the project's viability, or if it might be completed within the allocated budget. They are mainly three types: Preliminary estimate is a rough one that is typically based on an approximation of square feet per estimate. The measurements and Area in this estimate are only provided for illustrative purposes. Sometimes the price can vary by up to 50%. Detailed Estimate is the comprehensive estimate which include material specifications, the proposed method of completion, as well as precise measurements and drawings. The amounts of the works may vary by up to 10%. Abstract estimate is the estimate that solely contains the entire quantities of the items of work, rates determined by the PWD schedule or market values, and the project's overall cost. The estimate that includes updated quantities, specifications, and rates is known as a revised estimate. ((Sekhar, 2021) (Arafa, 2011))

2.4 Machine learning

Machine learning is a method that creates a model from data. In the field of machine learning, data from the past is utilized to anticipate future results. The machine learning technique is first applied to a training data set, and following the learning process, a model is created. The final result of the machine learning process is this model, which may be applied in real-world situations. The model can then on unknown input data give an output based on the patterns or relationships found in the training set. Data is what drives machine learning techniques. The patterns in the data are identified using the accurate output values of the input values, the process is referred to as supervised. The learning process in a supervised machine learning technique is dependent on data sets that offer both input and output values ((Badawy, 2020), (Matel, 2022), (Kok, 2009)).

2.5 Application of Machine Learning Algorithms in Cost Prediction

The research done by (Cho, 2013) showed that the Artificial Neural network model had a lower error rate than the multiple regression model of projected building costs. In this study, the multiple regression model and an artificial neural network model were contrasted using cost information kept by a provincial office of education on primary schools built between 2004 and 2007. A total of 96 historical data were divided into 20 historical data for comparing the built-in regression model with the artificial neural network mode and 76 historical data for building models. The artificial neural network model was shown to be superior in terms of average error rate and standard distribution by comparing the estimated values of the two models. (Kim G. H., 2013) used 197 cases for model construction and validation. The remaining 20 instances for tested and discovered that the NN model provided more accurate estimation results than the RA and SVM models. As a result, it was decided that the NN model was most suited for determining the cost of school construction projects. A data collection with 530 historical expenses was employed by (Kim G. H., 2004). Compared to

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

the CBR or MRA models, the best NN model produced more precise estimates. The lengthy trial-and-error procedure, however, made it difficult to find the optimal NN model. In comparison to the other models, the CBR model was more effective concerning these tradeoffs, particularly its clarity of explanation when calculating construction costs. The ability to update the building cost model easily and maintain consistency in the variables contained are key aspects of the model's long-term use. described as Whereas conventional modeling techniques frequently fall short, ANN offers solutions for difficult issues. For instance, ANN succeeds where many conventional modeling techniques fall short in capturing nonlinear and intricate interactions between the variables. They do, however, have their restrictions. They frequently require a specific set of inputs and outputs and can only be taught for that problem. As a result, any modification that calls for updating the network's architecture cannot be carried out automatically and must instead include human involvement. (Badawy, 2020) conducted research where 174 actual residential projects in Egypt provided the source of the statistics. To come to an understanding of the crucial elements influencing early-stage cost assessment, the Delphi method was employed. Regression techniques such as gamma, Poisson, and multiple linear regression were utilized. Using multiple linear regression, the suggested hybrid model was derived from the ANN model and the regression models. In comparison to the ANN model and regression models, the hybrid model's mean absolute percentage error was 10.64%, which is lower. The hybrid model's results show that it was effective at estimating the cost of residential projects and that it would be helpful to decision-makers in the construction sector. (Shojaei, 2019) discovered that regression models, on the other hand, typically required fewer model parameters than neural networks, which led to greater prediction performance if the relationships between the variables were well stated. Comparison of the regression model's results with those from the neural network model may help to determine whether the regression model needs nonlinear or interaction terms.

In this work Linear Regressor, Decision Tree Method, Random Forest method, Artificial Neural Networks, Support Vector Machine, XGboost method, Extra tree method, Voting Regression, and Stacking method are used to develop the models.

3. METHODOLOGY

3.1 Topic Selection

The selection of the topic" A Comparative Study of Machine Learning Algorithms for Early Cost Estimation of Building Projects in Nepal" for my thesis was driven by its profound relevance and practicality within the context of the Nepalese construction industry with more and more data available from past projects, computers can learn and make smarter predictions. This can help people make better decisions when they start a new construction project.

3.2 Expert Opinion

After doing literature review from several research papers, the input factors which were gathered from literature review may not be relevant in case of Nepal.So, expert opinion is required for more accuracy. Questionnaire was filled by 5 experts (including both contractor and consultant) who have following criteria:

- more than 12 years of experience in this construction field.
- must have relevant educational background.
- working as Consultant/Contractor.

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

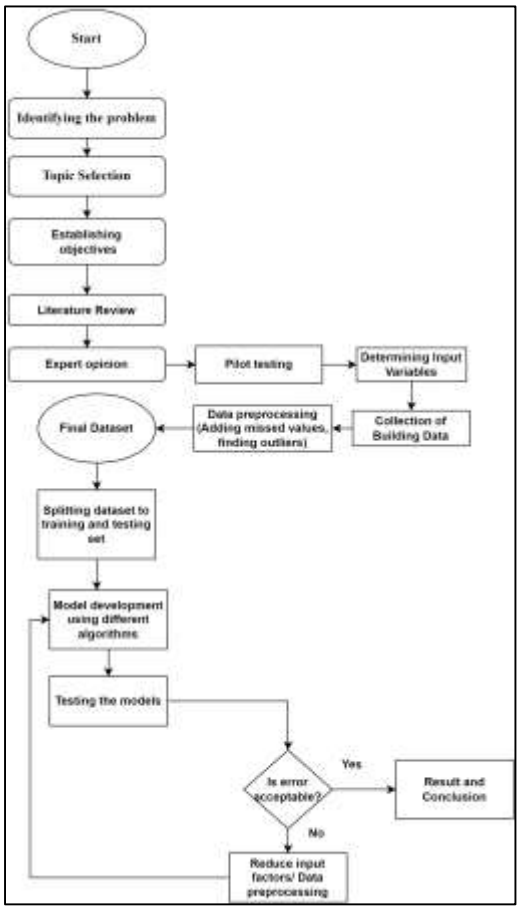

Figure 1. Workflow Diagram

3.3 Pilot Testing

A pilot survey is a survey that the researcher conducts with a smaller data size in collaboration with the experts. A gathered response helps to guide whether to move forward in research or change the questionnaire. It also helps to discover challenges that can affect the main data collection process (Atapattu, 2022). A pilot test was conducted with 3 respondents to check the clarity and comprehensibility of the questionnaire. The respondents easily understood the questionnaire; hence, there was no difficulty in filling up the questionnaire. The minimum time taken to fill the questionnaire was around 10 minutes and the maximum time taken was almost 20 minutes.

3.4 Data Collection

A total of seventy-two (72) building projects' structural data were gathered from various construction firms and consultancies with their final cost of projects. Data were collected from the Department of Urban Development and Building Construction (DUDBC), Consultancies, and Contractors. The data collection process was very tough as cost estimation value was confidential for contractors.

3.5 Data pre-processing

- Separated Numerical and Categorical Features (excluding the total cost of the project).
- Read the Excel file into a Data Frame (df).
- Counting the number of numerical features.
- Plotted individual scatter plots for numerical features to visualize the data and identify outliers.
- Plotted scatter plots between numerical features and the total cost of the project to analyze their relationship.
- Plotted a normal distribution graph for numerical features to analyze the distribution of the data.
- Calculated mean, median, mode, and variance for individual numeric features.
- Replaced missing values in numerical features with the mean, median, and mode, and data was saved in separate Excel sheets.
- Normal distribution graph was plotted for all metrics.
- Based on the analysis of variance, replaced missing values with mean as there is less variance in data when replaced with the mean value.
- Counted the number of missing categorical features.

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

• Plotted histograms for categorical features to find whether values are unique or in some order.
• Replaced missing values in categorical features with the mode (i.e., repeated values) and again plotted histograms.
• Features were encoded using one-hot encoding since it represents distinct project names without any inherent order.
•Encoded categorical features using one-hot encoding give all the data values in numeric features, hence suitable for further analysis.
•There are no missing categorical data now.
•Final Data set is ready and split the data set into training and testing sets in a ratio of 80:20 to implement in models.

3.6 Models Implementation

In this work models such as Linear Regressor, Decision Tree Method, Random Forest method, Artificial Neural Networks, Support Vector Machine, XGboost method, Extra tree method, Voting Regression, and Stacking method are implemented.

Linear Regression (LR) is a basic regression model used to establish a linear relationship between independent and dependent variables. Decision Tree Regressor (DT) partitions data into subsets based on features for predictions. Random Forest Regressor (RF), an ensemble method, combines predictions from multiple decision trees. The Neural Network (NN) comprises several dense layers with varying activations trained using the 'Adam' optimizer for 100 epochs to minimize mean squared error. XGBoost Regressor (XGB) is a gradient-boosting algorithm that combines weak learners to boost predictive performance. Support Vector Machine (SVM) with a linear kernel is used for regression. Extra Trees Regressor (ET) is akin to Random Forest but employs random thresholds for feature splitting. Voting Regressor (Voting) amalgamates LR, DT, and RF models. Stacking Regressor (Stacking) combines LR, DT, RF, ET, and Gradient Boosting models via a meta-regressor. Each model showcases unique methodologies and predictive strengths tailored to the task at hand.

• Linear Regression, Decision Tree, Random Forest, Extra Trees, Voting Regressor: scikit-learn
• Neural Network: Keras with TensorFlow backend
• XGBoost: xgboost library
• Support Vector Machine: scikit-learn
• Gradient Boosting: scikit-learn

Setting a seed or random state parameter ensures reproducibility of results. When specifying a particular number, such as 42, as the random state or seed, it initializes the random number generator in a manner that each execution of the code with the same seed produces identical random values. For instance, the random state=42 parameter is used in DecisionTreeRegressor, RandomForestRegressor, ExtraTreesRegressor, and GradientBoostingRegressor, which initializes these models with the same random seed. This initialization guarantees consistency in their behavior across different runs. Similarly, in XGBoost, random state=42 is utilized to ensure reproducibility in the behavior of the XGBoost regressor.

Model Architectures:

•Linear Regression (LR): Utilizes a simple linear model to establish a linear relationship between input features and the target variable. No hidden layers are involved.

•Decision Tree (DT) Regressor: Employs a decision tree-based model to make predictions using a tree-like graph, consisting of nodes representing features, branches, and leaf nodes containing the predicted values.

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

•Random Forest (RF) Regressor: Comprises an ensemble of decision trees to enhance prediction accuracy by averaging the outputs of multiple decision trees.

•Neural Network (NN) Regressor: Implements a feedforward neural network with four layers: an input layer with the number of features as neurons, followed by three hidden layers having 2048, 256, and 64 neurons respectively, and an output layer with one neuron for prediction.

The neural network consists of a Sequential model, indicating a linear stack of layers.
There are three hidden Dense layers:
1. Dense layer with 2048 neurons and 'softmax' activation function.
2. Dense layer with 256 neurons and 'relu' (Rectified Linear Unit) activation function.
3. Dense layer with 64 neurons and 'relu' activation function.
Finally, there is an output layer with a single neuron.

Input shape:
–The input shape is determined by the number of features in the training data. It's specified as (X train.shape[1],), which indicates the number of columns or features in the input data.

Model compilation:
–The model is compiled using the 'adam' optimizer and the loss function set to 'mean squared error'.
Training:
The model is trained using the fit method where it's trained on X train and y train data.
–The training is performed for 100 epochs with a batch size of 32.
–The verbose parameter set to 0 implies that no output will be printed during training.
–Validation data (X test, y test) is used to validate the model's performance after each epoch.
Predictions:
– After training, the model is used to make predictions on the test data (X test), and the predictions are stored in nn predictions.

•Voting Regressor: Creates an ensemble by combining the predictions from multiple base estimators (LR, DT, RF) and generates a final prediction based on the aggregated results.

•Stacking Regressor: Combines predictions from multiple base estimators (LR, DT, RF, ET, GB) using a meta-estimator (LR) to produce final predictions.

- XGBoost Regressor: Deploys an XGBoost-based ensemble model using gradient boosting that sequentially builds multiple decision trees to predict the target variable

3.7 Performance Metrics for Models

Calculation of mean square error, mean absolute error, and R square [23].

a. Mean Squared Error (MSE):
The Mean Squared Error (MSE) assesses the average squared differences between predicted (Pj) and actual (Tj) values in the dataset. The formula for MSE is given by: T_j

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(Tj - Pj)^2 \quad (1)$$

Where:

• N represents the total number of data points.
• Lower MSE values indicate better agreement between predicted and actual values, with a perfect model yielding an MSE of 0.

b. Mean Absolute Error (MAE): The Mean

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

Absolute Error (MAE) measures the average absolute differences between predicted (Pj) and actual (Tj) values in the dataset. The formula for MAE is given by:

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|Tj - Pj| \quad (2)$$

Where:

• N represents the total number of data points.
• MAE provides a measure of the average magnitude of errors between predicted and actual values. It is less sensitive to outliers compared to Mean Squared Error (MSE).

c. R-Squared (Coefficient of Determination): The R^2 coefficient of determination quantifies the proportion of variance in the target variable (Tj) that is predictable from the independent variables. The R^2 value is calculated as:

$$R^2 = 1 - \frac{SSE}{SST} \quad (3)$$

Where:

• SSE denotes the Sum of Squares of Residuals, reflecting the difference between predicted and actual values.
• SST represents the Total Sum of Squares, indicating the total variance in the target variable.

3.8 Tools and Experimental setup

The code is written in Python and uses various libraries for data pre-processing and machine learning. Google Co-laboratory platform was used for the training part of the experiment. It is a cost-free Python environment that runs in the cloud where a user can execute code, using powerful computing resources that are faster to train the complex machine learning models as compared to other general-purpose computers. Pytorch version 1.13 an open-source platform for machine learning-related projects, was taken into consideration to implement the different architecture. The Python Imaging Library (PIL 7.0.0), Matplotlib was used to perform image processing and computer vision tasks.

• pandas (pd): This library is used for data manipulation and analysis. It provides data structures like DataFrames that allow to work with structured data efficiently.
• sklearn.model selection: This module within the sci-kit-learn library provides tools for splitting data sets into train and test sets, and for cross-validation.
• sklearn.impute: This module provides classes for imputing (filling in) missing values in data sets.
• sklearn.preprocessing: This module provides functions for preprocessing data before feeding it into machine learning models. In this context, it includes Label Encoding, which is used to transform categorical labels into numerical values.
• sklearn.ensemble: This module provides ensemble methods for machine learning, such learning models. mean squared error, is a common metric to measure the quality of a regression model's predictions.

4. RESULTS AND DISCUSSIONS

4.1 Filtering the Input factor from expert opinion

As per experts, laboratory tests, building code use, Consulting fees, area of formwork, market condition, Solid waste management, roof type, insurance of staff, material, and equipment, and waterproofing were the least important factors and can be shown by the numeric value in the table 1. High numeric values for aspect yes are taken into consideration and the high value of aspect no is eliminated. Additional factors were also given by Experts after eliminating and adding some factors.

Table 1. Count of Response from Experts

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

Aspect	Yes	No
Location of Building	5	0
Type of Building	5	0
Site/Geographic Conditions	5	0
Access to Site	4	1
Site Area	3	2
Plinth Area	5	0
Floor Area	5	0
Floor Height	4	1
Number of Storeys	4	1
Number of Columns	3	2
Number of Rooms	4	1
Number of Bathrooms	3	2
Number of Beams	3	2
Type of Foundation	5	0
Roof Type	2	3
Number of Lifts/Elevator	4	1
Basement	5	0
Building Code Used	2	3
Laboratory Tests	1	4
Consulting Fees	2	3
Insurance of Staff, Material, Equipment	3	2
Waterproofing	2	3
Aluminum and Railing Works	5	0
Wood Works	5	0
Type of Flooring Works	5	0
External/Internal Finishing	4	1
Area of Formwork	1	4
HVAC Work	5	0
Water Supply & Drainage System	5	0
Solid Waste Management	4	1
Water Treatment, Septic Tank, Soak Pit	3	2
Electrical System	5	0
CCTV, AC & Ventilation System, Solar	4	1
Landscaping	5	0
Road Works/River Training Works	3	2
Market Condition	4	1

The insights from four distinct experts, labeled as Expert no. 1 through Expert no. 5. These factors encompass various aspects such as the availability of skilled labor, the use of specific materials and equipment from foreign countries, the influence of brand reputation on materials and equipment choices, considerations for material quality, the distance of material sources from project sites, the construction year, safety requirements, interior design preferences, rates of construction materials, rainwater harvesting practices, specifications of works, and the importance of detailed discussions on finishing materials and qualities. The range of expertise covered by these factors highlights the multidimensional nature of decision-making in construction projects, encompassing practical, logistical, and aesthetic considerations. The Building Attributes that are finally considered for further processing are listed. The attributes listed include details such as the name of the project, location of the building, type of building, construction completion year, site/geographic conditions, access to the site, site area, type of foundation, plinth area, floor area, floor height, number of floors, number of columns, number of rooms, number of bathrooms, number of kitchens, number of lifts/elevator, number of basements, use of building code, type of window, type of door, type of flooring works, external painting, internal finishing, HVAC work, sanitary works, electrical works, landscaping, and road works/river training works. There are 0 to 2 basements and a range of 1 to 13 floors in terms of storeys. Buildings' overall structural costs range above 1 crore. Scatter plots are a fundamental tool in data exploration and can provide valuable insights into the relationships between numerical variables in a dataset. Scatter plots can reveal trends or patterns in the data.

The scatter plot in Figure shows a random distribution of points, it suggests that there is no discernible pattern or relationship between the variables being plotted. A random distribution typically indicates that there is no correlation or association between the variables. Each data point appears scattered across the plot without following any specific trend, slope, or pattern. This pattern-less

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

arrangement implies that changes in one variable do not have a consistent effect or influence on the other variable. In such cases, many non-linear models might be needed to explore other potential relationships or factors that could affect the variables under consideration.

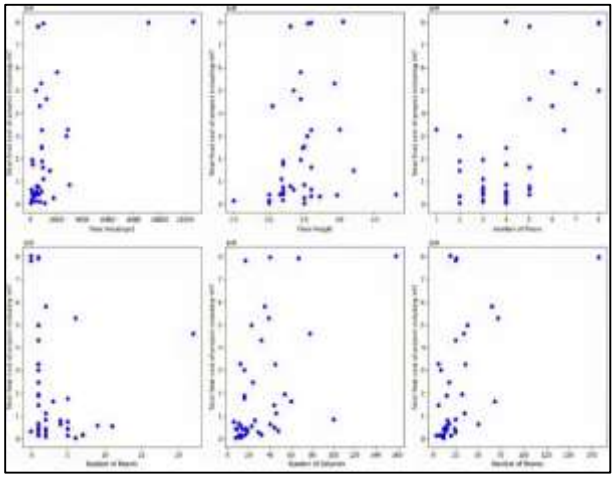

Figure 2. Scatter plot between Numerical Featuresand Total Cost of Project

A normal distribution graph describes how data is distributed when many independent, random factors contribute to an outcome. This distribution appears with a tail stretching towards the right side of the curve. It indicates that the data has a longer right tail compared to the left side. In such cases, the mean tends to be larger than the median, and most data points cluster towards the left side. A tail on the right side of a distribution indicates a right-skewed pattern, and outliers situated near this tail represent unusually high values that might have a significant impact on statistical measures and require thorough examination during data analysis.

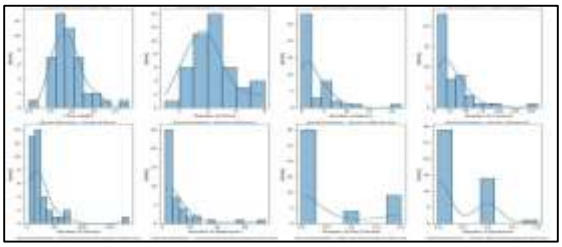

Figure 3. Normal Distribution Graph

4.2 Data pre-processing results

The data had missing values and preprocessing must be done. Some values are extremely high, null values, disordered text data, or noisy data. Data preparation is thus carried out to address this issue. Calculated mean, median, mode, and variance for individual numeric features.

Figure 4.Missing values of numerical features

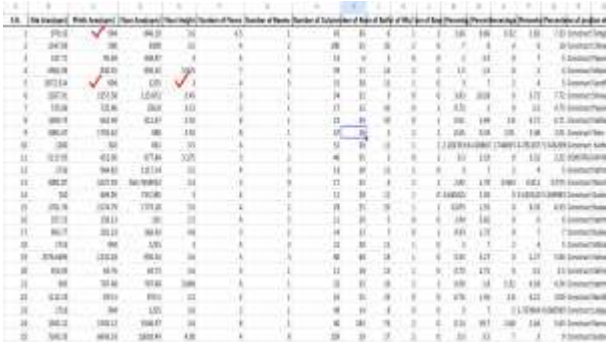

Figure 5. Adding missed values with mean

Calculated variance for data. Replaced

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

missing values in numerical features with the mean, median, and mode. Lower variance

Figure 6. Missing categorical values

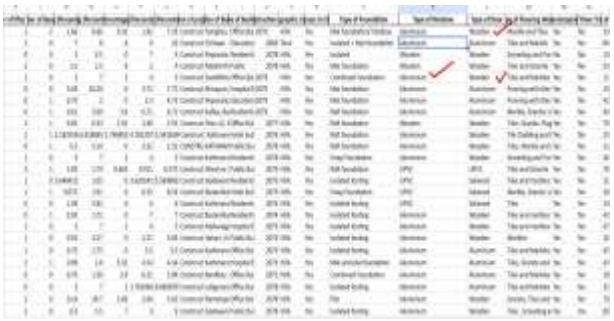

Figure 7. Replaced categorical values with mode

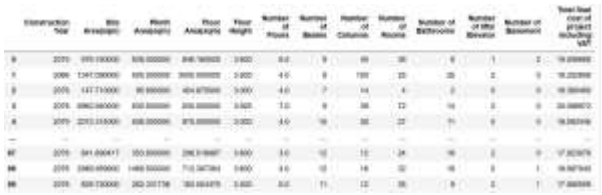

Figure 8. Replacing final cost with their natural logarithms

signifies less dispersion of data points from the mean value, implying that the dataset tends to be more tightly clustered around the mean. Comparing the variance values across different attributes between the raw data, mean-imputed data, and median-imputed data, mean imputation tends to preserve the original variability of the dataset better than median imputation. Hence, based on the variance as the criterion, mean imputation seems to be more aligned with maintaining the original data's variability compared to median imputation, and all numeric features are replaced by the mean in such dataset. Regarding replacing missing values in categorical features with the mode (most frequent value), it is a common approach and often a reasonable strategy, especially when dealing with categorical data. Imputing missing categorical values with the mode ensures the overall distribution of the categories and minimizes the potential impact of missing data on data analysis. Replacing the values in the Total final cost of the project including VAT with their natural logarithms as shown in Figure 8. Taking the logarithm can normalize the distribution or reduce the impact of extreme values making the data more suitable for data analysis.

Calculating the correlation matrix for the Data Frame:

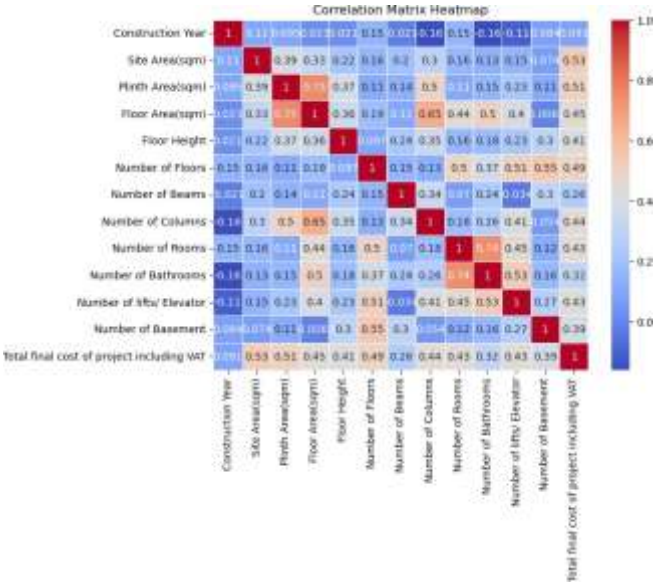

Figure 9. Correlation matrix heat map

The correlation matrix shows how each numerical column in the Data Frame is related to every other numerical column by calculating Pearson correlation coefficients. The Pearson correlation coefficient ranges

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation. Positive values indicate a positive correlation, while negative values indicate a negative correlation. It is useful for feature selection or understanding the data's pattern.

Dropping Plinth Area(sqm) and Number of Bathrooms: These two features are being removed because they exhibit a high correlation with other variables ('Floor Area(sqm)' and 'Number of Rooms' respectively) beyond a predefined threshold of 0.70. Due to their strong correlation with other variables, it's assumed that they might not provide additional significant information for the analysis or modeling and could potentially lead to multicollinearity issues.

Dropping Construction Year: This feature is being dropped because its correlation with the target variable ('Total final cost of the project including VAT') is lower than a specified threshold of 0.70, specifically correlating to 0.091. A correlation below this threshold suggests a weak linear relationship between 'Construction Year' and the target variable, which might not significantly contribute to explaining the variability in the target.

Cleaning the column: Cleaning the column Location of Building. Since the dataset is limited to certain places only. Data is categorized whether all the data is either inside of Kathmandu or Outside of Kathmandu. Since the Location of the building is cleaned into the inside valley and the Type of foundation is cleaned into individual foundations, we can drop these two columns.

The feature was encoded using one-hot encoding as shown in Figure 11. One hot encoding is that categorical variables have been transformed into a numerical format. Dropping all the variables that are either highly correlated with each other or are less correlated with the target variable which is the Total Final Cost of the project including VAT.

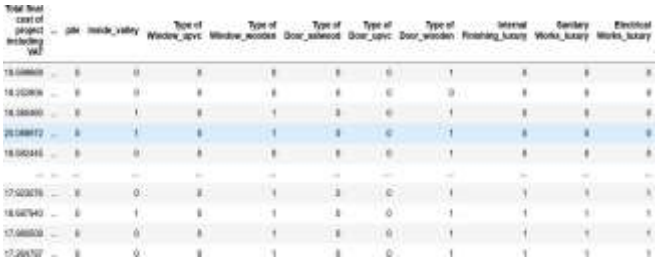

Figure 10. One hot encoding of categorical variables

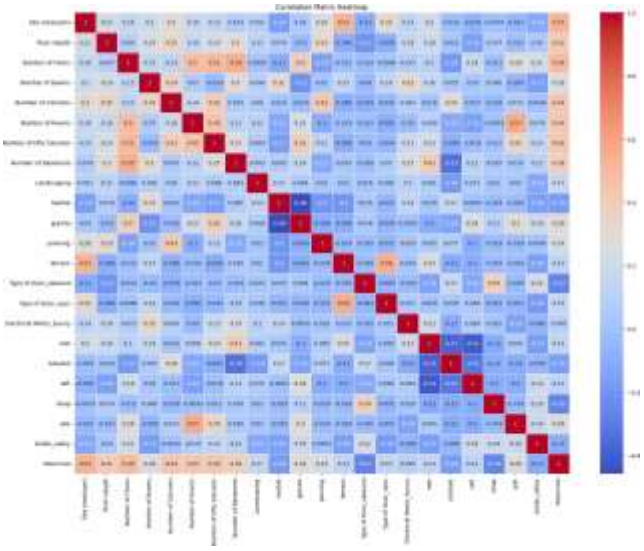

Figure 11. Correlation Matrix Heat map for Final features

4.3 Results of Models Implementation

Table 2. Model Metrics

Model	MSE	MAE	RMSE	R^2
LR	0.24	0.43	0.49	0.67
DT	0.09	0.10	0.30	0.88
RF	0.11	0.22	0.34	0.84
NN	0.42	0.47	0.65	0.41
XGB	0.35	0.23	0.59	0.51
SVM	4.10	1.11	2.03	0.47
ET	0.09	0.10	0.30	0.88
Voting	0.11	0.22	0.32	0.85
Stacking	0.14	0.23	0.37	0.81

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

Regarding the comparison of the models, the Decision Tree, Random Forest, ExtraTree, Voting, and Stacking models exhibit relatively better performance in terms of MSE, MAE, RMSE, and R^2. Among these, the Decision Tree, ExtraTree, and Voting models demonstrate particularly strong performance across multiple metrics. The Decision Tree or ExtraTree model is considered the best choice based on the provided metrics, as they seem to have lower errors and higher R^2 values compared to other models.

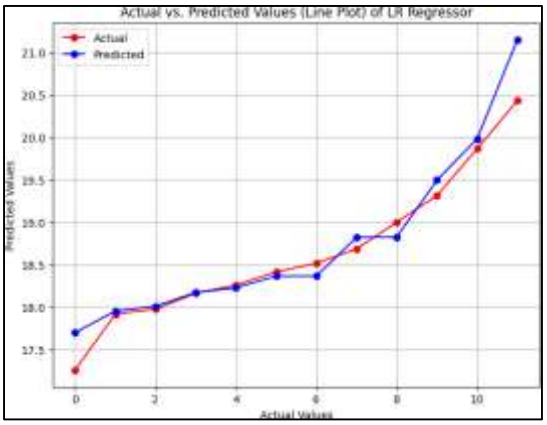

Figure 12. Linear Regressor

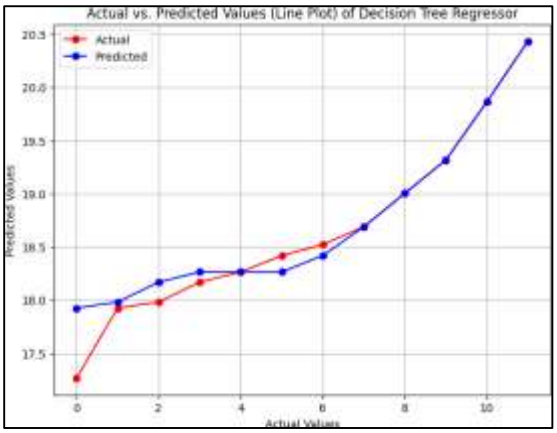

Figure 13. Decision Tree Regressor

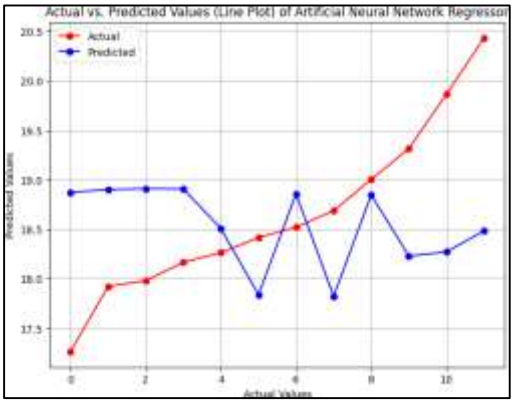

Figure 14. ANN

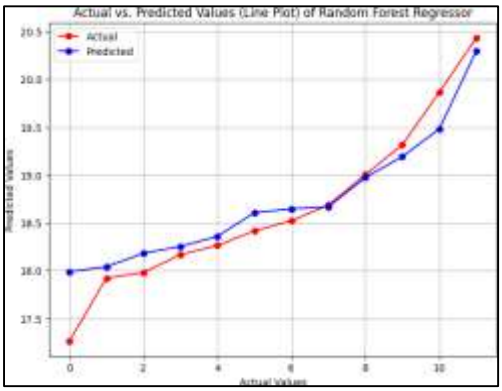

Figure 15. Random Forest Regressor

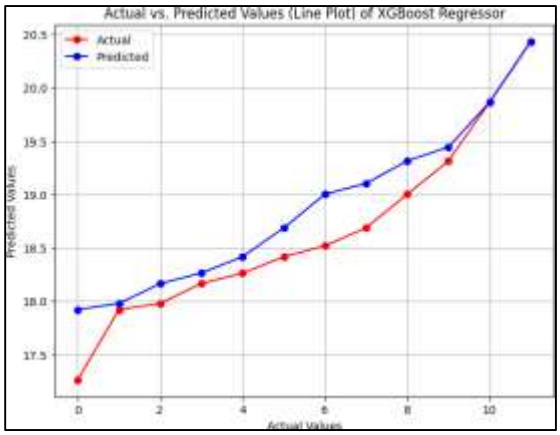

Figure 16. XGBoost Regressor

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

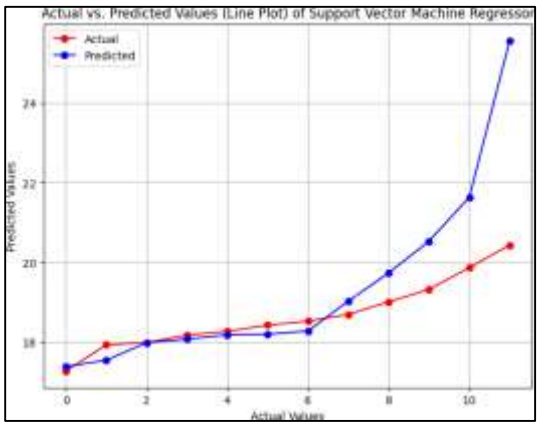

Figure 17. SVM Regressor

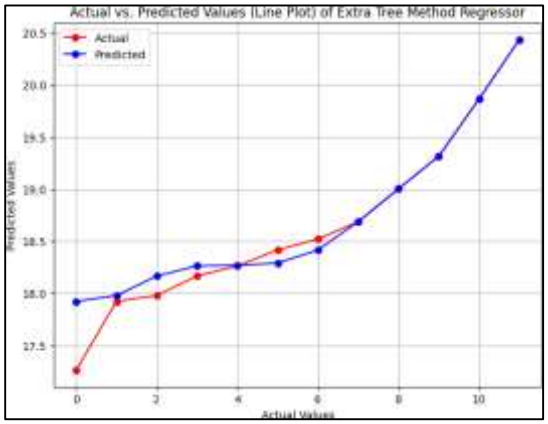

Figure 18. Extra Tree Regressor

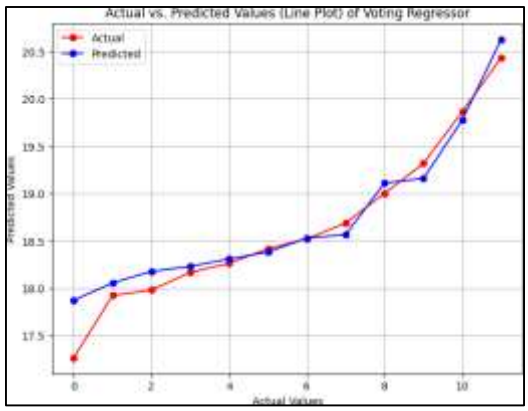

Figure 19. Voting Regressor

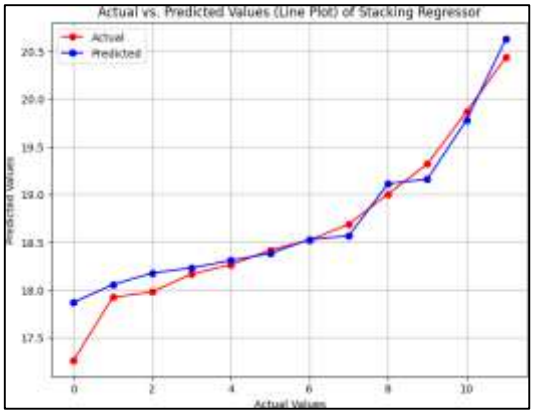

Figure 20. Stacking

5. CONCLUSION

Regarding the datasets, the buildings that were used are Educational Building, Commercial Building, Hospital Building, Residential Building, Public Building, Official Building, and Hotel Building having 0 to 2 basements ranging above 1 crore. The input features were taken from the literature review, and validated by expert opinion. After pilot testing, a survey questionnaire was distributed among contractors and consultants. Data preprocessing helps to clean data. Missing values are substituted by mean for numeric values and by mode for the categorical values. By analyzing the correlation heat map the unwanted features are dropped. The final dataset is divided into train and test sets in the ratio of 80:20. The nine models were implemented and the Mean absolute error, mean square error, and R square value are recorded for evaluation. The Decision Tree, Random Forest, Extra Tree, Voting, and Stacking models exhibit relatively better performance in terms of MSE, MAE, RMSE, and R2. Among these, the Decision Tree, Extra Tree, and Voting models demonstrate particularly strong performance across multiple metrics. The Decision Tree or Extra Tree model is considered the best choice based

on the provided metrics, as they seem to have lower errors and higher R^2 values compared to other models.

6. FUTURE RECOMMENDATIONS

•Enhanced Data Collection: Gathering more diverse and extensive data sets could provide a more comprehensive understanding of the relationships between features and building costs. It also helps to build accurate models.
•Feature Engineering: Explore more advanced feature engineering selection techniques that enhance the predictive power of the models.
• External Validation: Validate the developed models using external data sets from different geographic locations or periods to ensure the reliability of the models.
By addressing these recommendations, future research can contribute to the advancement of accurate cost prediction models in the construction domain, thereby assisting stakeholders in making informed decisions and improving overall project management efficiency.

7. ACKNOWLEDGMENTS

The authors would like to express the sincere gratitude towards Asst. Prof. Mahendra Raj Dhital whose help in completing this research is inevitable.

8. REFERENCES

1. Akalya, K. R. (2018). Minimizing the cost of construction materials through optimization techniques. IOSR Journal of Engineering.
2. Allen, E. &. (2019). Fundamentals of building construction: materials and methods. John Wiley & Sons.
3. Arafa, M. &. (2011). Early-stage cost estimation of buildings construction projects using artificial neural networks. Journal of Artificial Intelligence, 4(1), 63-75.
4. Assaf, S. A.-K.-H. (1995). Causes of delay in large building construction projects. Journal of management in Engineering, 11(2), 45-50.
5. Atapattu, C. N. (2022, November). Statistical cost modelling for preliminary stage cost estimation of infrastructure projects. In IOP Conference Series: Earth and Environmental Science, 1101, 052031.
6. Badawy, M. (2020). A hybrid approach for a cost estimate of residential buildings in Egypt at the early stage. Asian Journal of Civil Engineering, 21(5), 763-774.
7. Ching, F. D. (2020). Building construction illustrated. John Wiley & Sons.
8. Cho, H. G. (2013). A comparison of construction cost estimation using multiple regression analysis and neural network in elementary school project. . Journal of the Korea Institute of Building Construction, 13(1), 66-74.
9. Dissanayaka, S. M. (1998). Comparing contributors to time and cost performance in building projects. Building and Environment, 34(1), 31-42.
10. Gautam, D. R. (2016). Common structural and construction deficiencies of Nepalese buildings. Innovative infrastructure solutions, 1, 1-18.
11. Ghadbhan Abed, Y. H. (2022). Machine learning algorithms for constructions cost prediction. A systematic review. International Journal of Nonlinear Analysis and Applications, 13(2), 2205-2218.
12. Hwang, B. G. (2017). Factors affecting productivity in green building construction projects: The case of Singapore. Journal of Management in Engineering, 33(3), 04016052.
13. Kim, G. H. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. Building and environment, 39(10), 1235-1242.
14. Kim, G. H. (2013). Comparison of school building construction costs estimation methods using regression analysis, neural network, and support vector machine.

Kathford Journal of Engineering and Management,
ISSN: 2661-6106.

https://kathford.edu.np/Journal/

VOL. 3 ISSUE 1 JUNE 2023, pp. 34-49

15. Kok, J. N. (2009). Artificial intelligence: definition, trends, techniques, and cases. Artificial intelligence, 1, 270-299.

16. Mahamid, I. (2013). Contractors perspective toward factors affecting labor productivity in building construction. Construction and Architectural Management, 20(5), 446-460.

17. Matel, E. V. (2022). An artificial neural network approach for cost estimation of engineering services. International journal of construction management, 22(7), 1274-1287.

18. Sekhar, D. N. (2021). A Course Material on Estimation, Costing and Valuation.

19. Shojaei, A. L. (2019). Revisiting systems and applications of artificial neural networks in construction engineering and managements. Proceedings of the International Structural Engineering and Construction, 20-25.

20. Sonmez, R. ((2004). Conceptual cost estimation of building projects with regression analysis and neural networks. Canadian Journal of Civil Engineering, 31(4), 677-683.

21. Tayefeh Hashemi, S. E. (2020). Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. SN Applied Sciences, 2, 1-27.

22. Veliyampatt, S. (n.d.). Determination of Efficacy of Cost Estimation Models for Building Projects using Artificial Neural Networks, Fuzzy Inference System and Regression Analysis. International Research Journal of Engineering and Technology (IRJET), 8(10).

23. Yap, J. B. (2020). Analysing the underlying factors affecting safety performance in building construction. Production Planning & Control, 31(13), 1061-1076.