International Journal of Engineering and Information Systems (IJEAIS)
ISSN: 2643-640X
Vol. 10 Issue 6 June - 2026, Pages: 24-33

Predicting cost overrun in building construction projects using support vector machine

Oluwatomisin Ayomiposi OYELEYE[1] , Olumuyiwa Samson ADERINOLA[2]

[1]Department of Civil and Environmental Engineering,
School of Infrastructure, Mineral and Manufacturing Engineering,
Federal University of Technology, Akure, Ondo State, Nigeria
tomisineshton@gmail.com
[2]Department of Civil and Environmental Engineering,
School of Infrastructure, Mineral and Manufacturing Engineering,
Federal University of Technology, Akure, Ondo State, Nigeria
osaderinola@futa.edu.ng

*Author for correspondence: tomisineshton@gmail.com

***ABSTRACT**: This study looks at how well support vector tools forecast cost overruns in building projects using information from historical records. It examines the efficacy of Support Vector Machines (SVM) in forecasting construction project cost overruns by means of a measurable approach in getting its data. The independent variables of number of stories, entire floor area(m^2), scope changes, provisional sum, tendering type, and initial contract sum were extracted from historical records. SVM model was compared using Decision Tree and Random Forest. The results revealed that the SVM model produced accurate construction cost predictions with 0.071, 0.099, and 0.693 of MSE, RMSE and R2 respectively on the accuracy test data. The lowest MSE and RMSE values indicate that SVR had the smallest prediction error, and the highest R^2 score of 0.693 which means that SVR explained approximately 69.3% of the variance in cost overruns. This superior performance suggests that SVR was particularly effective at modeling the data and capturing the underlying trends and relationships. When considered collectively, inaccurate initial cost estimation, Material price fluctuations, Storey Count, and Changes in scope are reliable indicators of cost overruns in the construction sector. The resulted SVM model in this work can be applied as yard stick to predict potential cost overruns for Nigerian construction projects.*

Keywords: Support Vector Machine (SVM), Cost Overrun, Machine Learning, Building Construction Modelling.

1. INTRODUCTION

1.1 Background to the Study

The construction industry has an influence on the economy of all countries (Khaertdinova *et al., 2021*; Lopes, 2022). This sector delivers essential ingredients for the advancement of an economy. Babalola and Ojo (2016) stated that the construction sector contributes between 6 and 9 percent of the economy's gross domestic product, to several nations; and according to Sha *et al.*, 2017, it reaches up to 10% of the gross domestic product of most nations. Quality problems, cost overruns and the failure to complete the construction project within the given timeframe are the inadequacy of many construction projects carried out in developing countries. (Bahamid *et al.,* 2019). Provision of job, improving quality of life like basic services of schools, hospitals, roads, and the likes are done by construction sector in most countries. This construction sector is often referred to as suitable investment interface. (Soewin and Chinda, 2018). This sector is facing many difficulties when trying to improve its efficiency (Soewin and Chinda, 2018). Recently, cost overruns in construction projects is worrisome to stakeholders. It has affected the outputs of the construction sector in various countries resulting poor confidence in the industry's capacity to meet expected economic growth in affected countries.
Construction Projects can be amazing in degree and complexity dated back to pyramids in Egypt, Lalibela's rock-hewn churches, and Gothic Cathedrals to soaring skyscrapers and enormous bridges. Noticeably, the construction trade has distinct structures that are not usually encountered in other industries. In the construction industry, extra expenses and time are typically required when field circumstances prove to be more complicated than initially projected during the planning and design stage. Extremes can harm materials and equipment and impair output. Furthermore, mass production techniques are challenging to use in this business because it is primarily custom orientated. It is challenging to estimate with precision the amount of money required to finish building projects due to all of these and other factors. Large facilities require a significant financial outlay and a lengthy construction period. Delays, excess costs, and other issues are typically disproportionately severe (Gould, 2002; and Gurumayum, 2025). As a result of construction projects complexity and dynamic nature, they have historically encountered a series of cost overrun issues (Afzal *et al.*, 2019). A projects' initial estimated cost and the project's bid or final cost difference can be significant. Controlling project budgets throughout the project's lifecycle is a noticeable challenge for construction companies (Shane *et al.*, 2009) and (Ashiedu *et al*., 2023). A successful project meets technical specifications, adheres to the schedule, and stays within the budget (Hammad, 2014). However, approximately 86% of construction projects encounter cost overruns occur as a variety of factors, including -design changes, material shortages and unforeseen site conditions. Thus, it is a common problem.

It is therefore critical to predict the likelihood of cost overruns in construction projects to avoid financial losses and delays. To analyze cost overrun based on various factors, such as construction time (Zujo *et al.*,2018), structure type, building area, number of floors, floor height (Alshamrani, 2017), geotechnical and construction variables (Petroutsatou *et al.*,2006), researchers have developed a number of linear models.

Recently, with progression in machine learning and data analytics, new prospects is being experienced in construction projects cost estimation (Yang *et al.*, 2018). Linear regression, support vector machines, and artificial neural networks (Machine learning methods) have found their application into various fields due to ability to learn from data and accurately present useful and reliable outcomes (Cheng *et al.*, 2023). Machine learning techniques is now employs to construction cost estimation works with favorable outcomes (Abolfazl *et al.*, 2021).

Vapnik developed support vector machines (SVM) technique (Vapnik, 1995). SVM follows the theory of fundamental probability reduction and numerical understanding. SVMs can categorize model statistics. SVMs require data occasions to be trained and tested just like neural networks. The SVMs system intents to reduce the generalization error's upper bound with oversimplification even after dispensing with data that has not been seen before. The SVM, civil engineering, is applicable to forecasting. It is widely used efficiently in construction industry hitches in building construction cost estimation (Qin *et al*, 2015; Juszczyk, 2018; and Xiao, 2021), highway construction (Petrusheva *et al., 2019*) and (Meharie and Shaik, 2020) and concrete strength prediction (Yan and Shi, 2010) and (Imran, 2024).

2. RESEARCH METHODOLOGY

The research methodology involved three main phases: development of the questionnaire, data collection and data analysis using support vector machine. In this research Anaconda Python was used for the entire implementation of algorithm.

2.1 Data Collection

For this research, data obtained were evaluated and analyzed. The data were obtained from a well designing and well-structured questionnaire were distributed to construction industries drawn from both the private and public sectors in South West Nigeria. The target population for the questionnaire included the Main Contractors, Sub-Contractors, Engineers, Project Managers, Architects, Quantity Surveyors, Builders and/or potential suppliers who are executing or have executed projects in building construction. A stratified random sampling method was used while distributing the questionnaires with the goal to obtain a samples size large enough to ensure statistical validity. The time required for the filling of the questionnaires was a minimum of 15 minutes and maximum of 20 minutes in order to achieve a good response rate. An estimated sample size of 200 respondents was targeted to ensure statistical significance and reliability of the findings. In the questionnaires, the respondents were asked to indicate based on their experience the factors that affects cost overrun in their practice. Data were collected via online surveys and physical distribution at construction sites.

2.1.1 Data Pre-processing

The dataset was pre-processed to convert categorical features into numerical format like all other machine learning it did so using one-hot encoding. The data was also divided into training and testing sets with a 70-30 split ratio. The division ensures that the model is trained on a substantial data portion while being evaluated on a separate, unseen subset to assess its generalization ability. Data pre-processing is data cleaning and transformation prior the application of a machine-learning model and make it consistent by converting data values into a specific range (Abderahim *et al*., 2023; Alvaro *et al*.,). This paves way for good quality of the data used to train machine learning model prior to analysis using handling categorical variables; Feature scaling of variables; and Kernel Selection as obtainable in this research work.

2.1.2 Handling Categorical Variables

In the context of predicting cost overruns using a machine learning model, handling categorical variables involves preparing and encoding data that contains categorical variables so that it is used as input for the model. In preparing data, dataset was processed by removing the target column, Cost Overrun, from the feature set, leaving the predictors for model training. Categorical features were converted into numerical values using one-hot encoding to make them suitable for the SVR model. The data was divided into training and testing subsets, with 70% allocated for training and 30% for testing. This division allows the model to learn from a substantial data portion while being validated on a separate set to gauge its predictive accuracy. The model's performance was weighed with the use of Mean Squared Error (MSE): Root Mean Squared Error (RMSE) and R-squared (R^2) using the testing data. In encoding, one-hot encoding which involves creating a new binary variable for each category in the categorical variable was used. For all variables in the data set a new (binary) column was created, indicating the presence of each possible value from the original data. For the variables with three to five categories new binary variables would be created, one for each category. The value of the binary variable would be 1 if the category is present for a given data point, and 0 if the data is absent. Once the categorical variables were encoded, the resulting data was used as input for the model. The model learned to identify patterns and relationships between the input variables, including categorical variables, and the target variable (cost overruns) and made predictions based on those patterns.

2.1.3 Feature Scaling of Variables (FSV)

FSV was carried out using the mean and standard deviations of all the features in the training data and then apply z-score standardization to the features for conversion to a space with unit variance and a mean of zero. The transformed value is produced by deducting the mean from the original value and dividing it by the standard deviation. This ensures that each feature has a mean of zero and a standard deviation of one, putting them on a comparable scale and making it easier for the SVM to learn from them. The vector is converted into a format appropriate for machine learning methods.

2.1.4 Kernel Selection

The kernel selected for this research was linear kernel. Kernel selection is a critical stage in building a model that determines how the input features are transformed into a higher-dimensional space where the data points become more separable. The choice of kernel depends on the characteristics of the dataset.

2.3 Selection of Algorithm

Two other supervised machine learning algorithms were selected asides SVM based on their performance namely Decision Tree (DT) and Random Forest (RF). Support vector machine was then compared in performance with the other two machine learning.

2.4 Methodology Flowchart

The methodology flowchart as shown in Figure 1, gives a pictorial representation of how the research methodology was carried out.

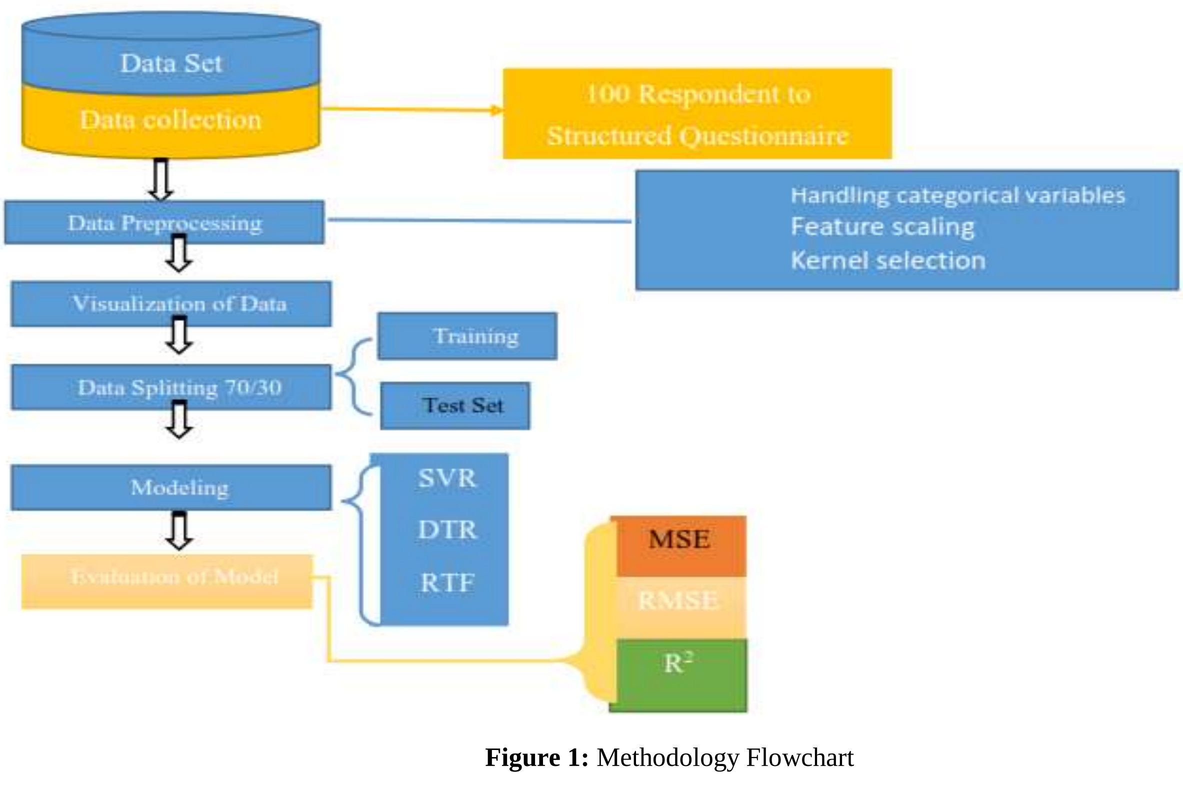

Figure 1: Methodology Flowchart

2.5 Model Training and Evaluation Metrics

After collecting the data from different respondent and pre-processed, it was shuffled to avoid predisposition throughout modeling. The dataset was split into two, training and testing data in 70/30 percentage. The training data was used to train the models and the test set was used to evaluate the performance of the models.

The computation of dissimilar performance metrics was used to appraise the performance of model. The efficiency of the models was assessed and compared by means of MSE, RMSE, and R^2 which is the determination coefficient.

3 RESULTS AND DISCUSSIONS

3.1 Data Collection Result

150 questionnaires were circulated and 100 were filled and returned. The dataset consisted of 24 features. These features include project attributes such as Project Type, Number of Storey, Total floor area (in square meters), Project Location, Construction timeline, Planned Project Cost (Budget),Actual Project Cost (final), Availability of Resources, Availability of Skilled Labor, Labor cost, Material cost, Equipment cost, Weather Conditions, Economic Conditions, site condition, Impact of Unforeseen circumstances, type of contract being used for the project, Accuracy of Initial Cost Estimates, Number of change orders, Complexity of Design, ratings of the project management which was gathered and collated in a Microsoft data sheet is presented in Table 1.

Table 1: Data Extracted on an Excel Spreadsheet Database

1	ype of Organizatio	ression of Respon	e do you	Project typ	ber of sto	ea (in squ	Project	Construct	Planned	Actual	Availability	Availabili	Labor cost	Material cost	uipment c	Weather	itions dur
2	Consulting Firm	Engineer	s than 5 ye	Commercia	1	0 - 500	Ekiti	- 5 month	000 - 10,000	,000 - 15,000	lways availab	Very High	- 5,000,00(	0 - 5,000,000	- 5,000,00(	y Unfavor	Stable
3	overnment Parastat	Engineer	s than 5 ye	Commercia	1	501 - 1000	Ondo	- 24 mont	000 - 10,000	0,000 - 40,000	arely availabl	Moderate	001 - 10,00	,000,001 - 15,000,0(	- 5,000,00(	Neutral	Recession
4	Contracting Firm	Engineer	s than 5 ye	Residentia	2	501 - 1000	Lagos	- 12 month	,000 - 30,000	,000 - 40,000	lostly availabl	Very High	- 5,000,000	,000,001 - 15,000,0(	- 5,000,00(	Neutral	Inflation
5	overnment Parastat	Quantity surveyor	1 - 15 year	Residentia	3	1001 - 500(	Osun	- 36 mont	,000 - 40,000	,000 - 15,000	lostly availabl	Very High	,001 - 20,0	5,000,001 - 20,000,0(	- 5,000,00(	Favorable	Stable
6	Contracting Firm	Engineer	s than 5 ye	Commercia	1	001 and abo	Lagos	- 36 mont	0,001 and a	00,001 and ab	etimes avail	High	ve 20,000,	Above 20,000,000	ve 20,000,	Neutral	Inflation
7	Contracting Firm	Engineer	5 - 10 year	Industrial	3	0 - 500	Oyo	- 5 month	0,001 and a	00,001 and ab	lways availab	Moderate	,001 - 15,0	Above 20,000,000	ve 20,000,	Favorable	Inflation
8	Contracting Firm	Engineer	6 - 20 year	frastructu	2	0 - 500	Lagos	- 12 month	,000 - 20,000	,000 - 30,000	lways availab	High	- 5,000,000	,000,001 - 15,000,0(	001 - 10,0(	Favorable	Inflation
9	Contracting Firm	Engineer	1 - 15 year	Residentia	1	001 - 1000	Ogun	- 36 mont	,000 - 40,000	,000 - 50,000	lways availab	High	- 5,000,00	6,000,001 - 10,000,00	,001 - 15,0	Favorable	Inflation
10	Contracting Firm	Engineer	s than 5 ye	Residentia	2	0 - 500	Osun	- 5 month	,000 - 20,000	,000 - 15,000	lostly availabl	Moderate	- 5,000,00	6,000,001 - 10,000,00	- 5,000,00(	Neutral	Inflation
11	Contracting Firm	Engineer	5 - 10 year	Residentia	2	0 - 500	Lagos	- 5 month	000 - 10,000	,000 - 15,000	lways availab	Moderate	- 5,000,00	,000,001 - 10,000,00	- 5,000,00(	Favorable	Inflation
12	Other	Engineer	6 - 20 year	Residentia	2	501 - 1000	Ekiti	- 5 month	000 - 10,000	,000 - 15,000	lostly availabl	Moderate	- 5,000,00	0 - 5,000,000	- 5,000,00(	Neutral	Recession
13	Lasustech	Engineer	s than 5 ye	Residentia	1	0 - 500	Lagos	- 12 month	,000 - 40,000	,000 - 30,000	lever availabl	Low	- 5,000,00	5,000,001 - 20,000,0(	001 - 10,0(	y Unfavor	Stable
14	Contracting Firm	Engineer	s than 5 ye	Residentia	3	501 - 1000	Osun	onths and	0,001 and a	00,001 and ab	lostly availabl	High	,001 - 15,0	Above 20,000,000	- 5,000,00(	Favorable	Inflation
15	Contracting Firm	Engineer	5 - 10 year	Residentia	3	0 - 500	Lagos	- 12 month	,000 - 20,000	,000 - 30,000	lostly availabl	High	001 - 10,0(	,000,001 - 10,000,00	001 - 10,0(	Neutral	Inflation
16	Consulting Firm	Project Manager	1 - 15 year	frastructu	9	1001 - 500(	Lagos	- 24 mont	,000 - 30,0	00,001 and ab	lostly availabl	Very High	,001 - 15,0	,000,001 - 10,000,00	001 - 10,00	nfavorabl	Inflation
17	overnment Parastat	Engineer	6 - 20 year	Commercia	1	1001 - 500(	Lagos	- 36 mont	,000 - 40,0	00,001 and ab	lostly availabl	Very High	,001 - 20,0	5,000,001 - 20,000,00	ve 20,000,	Neutral	Inflation
18	Contracting Firm	Architect	5 - 10 year	Residentia	3	501 - 1000	Ogun	- 12 month	,000 - 20,000	,000 - 15,000	metimes avail	Moderate	001 - 10,00	,000,001 - 15,000,0(	001 - 10,0(	Neutral	Stable
19	Contracting Firm	Engineer	1 - 15 year	Commercia	2	501 - 1000	Ogun	- 12 month	,000 - 30,000	,000 - 50,000	metimes avail	High	001 - 10,0	,000,001 - 10,000,00	,001 - 20,0	Favorable	Inflation
20	Consulting Firm	Quantity surveyor	5 - 10 year	Industrial	4	501 - 1000	Osun	- 12 month	,000 - 20,000	,000 - 30,000	lways availab	High	001 - 10,00	,000,001 - 15,000,0(	001 - 10,00	nfavorabl	Stable
21	Contracting Firm	Engineer	5 - 10 year	frastructu	5	501 - 1000	Oyo	- 12 month	,000 - 20,000	,000 - 30,000	lostly availabl	High	001 - 10,0(	,000,001 - 10,000,00	001 - 10,0(	Neutral	Inflation
22	Consulting Firm	Engineer	s than 5 ye	Commercia	1	0 - 500	Ekiti	- 5 month	000 - 10,000	,000 - 30,000	metimes avail	Moderate	001 - 10,00	,000,001 - 15,000,0(	001 - 10,0(	Neutral	Inflation
23	Contracting Firm	Project Manager	5 - 10 year	Residentia	2	501 - 1000	Lagos	- 12 month	000 - 10,000	,000 - 15,000	lostly availabl	High	- 5,000,00	,000,001 - 10,000,00	- 5,000,00(	Favorable	Stable

3.2 Data Preprocessing and Analysis

The result of final Data Frame which was saved in the variable df is presented in Table 2. The head method was used to show the first few rows of the data, providing a general idea of structure and content of the dataset.

Table 2: Sample of the Cost Overrun Dataset

Type_of_Organization	Profession	years_of_experience	Project_type	Number_of_storey	Total_floor_area_in_square_meters	Project_Location	Construction_timeline?
Consulting Firm	Engineer	Less than 5 years	Commercial	0	0 – 500	Ekiti	1 – 5 months
Government Parastatal	Engineer	Less than 5 years	Commercial	0	501 – 1000	Ondo	13 – 24 months
Contracting Firm	Engineer	Less than 5 years	Residential	0	501 – 1000	Lagos	6 – 12 months
Government Parastatal	Quantity surveyors	11 – 15 years	Residential	0	1001 – 5000	Osun	24 – 36 months
Contracting Firm	Engineer	Less than 5 years	Commercial	0	10001 and above	Lagos	24 – 36 months

In the initial phase of data processing, handling missing values is essential for correctness of succeeding analyses. The dataset includes columns such as "Planned Project Cost (Budget)" and "Actual Project Cost (Final cost)", which are crucial for determining whether a project has exceeded its budget. To maintain the integrity of the analysis, rows with missing values in these key columns are removed. This step is performed using the dropna function, which eliminates rows where either "Planned Project Cost (Budget)" or "Actual Project Cost (Final cost)" is absent. This cleaning process ensures that only complete and valid data is considered for further analysis.

International Journal of Engineering and Information Systems (IJEAIS)
ISSN: 2643-640X
Vol. 10 Issue 6 June - 2026, Pages: 24-33

Following the removal of rows with missing values, the "Cost Overrun" indicator is recalculated. This new calculation is derived by comparing the "Actual Project Cost (Final cost)" with the "Planned Project Cost (Budget)". Specifically, a cost overrun is flagged if the actual project cost exceeds the planned budget. This comparison is done using a simple greater-than operation, resulting in a Boolean column that indicates whether the project experienced a cost overrun. To verify the effectiveness of these cleaning and recalculation processes, a sample of the dataset including the "Planned Project Cost (Budget)", "Actual Project Cost (Final cost)", and the newly computed "Cost Overrun" columns is output. This verification ensures that the recalculated values are accurate and that the dataset is free from missing data. The preprocessing process is presented in Figure 2.

```
# Fill missing values with median
median_budget = data['Planned Project Cost (Budget)'].median()
median_final_cost = data['Actual Project Cost (Final cost)'].median()

data['Planned Project Cost (Budget)'].fillna(median_budget, inplace=True)
data['Actual Project Cost (Final cost)'].fillna(median_final_cost, inplace=True)

# Recalculate Cost Overrun after imputation
data['Cost Overrun'] = data['Actual Project Cost (Final cost)'] > data['Planned Project Cost (Budget)']

# Verify the results
print(data[['Planned Project Cost (Budget)', 'Actual Project Cost (Final cost)', 'Cost Overrun']].head())
```

```
   Planned Project Cost (Budget)  Actual Project Cost (Final cost)  \
0                      5500000.0                        10000000.0
1                      5500000.0                        35500000.0
2                     25500000.0                        35500000.0
3                     35500000.0                        10000000.0
4                     60000001.5                       360000000.5

   Cost Overrun
0          True
1          True
2          True
```

```
# Save the cleaned dataset with cost overrun
data.to_csv('cleaned_data_with_cost_overrun_boolean.csv', index=False)
```

Figure 2: Data Preprocessing

Number of Projects with Cost Overrun, Figure 3, explains the distribution of projects with and without cost overruns. The chart shows that 71.0% of the projects experienced cost overruns, represented by the 'salmon' slice. Conversely, 29.0% of the projects were completed within the planned budget, represented by the 'skyblue' slice. The chart provides a clear visual representation of the proportion of projects falling into each category, highlighting the extent of cost overruns within the dataset

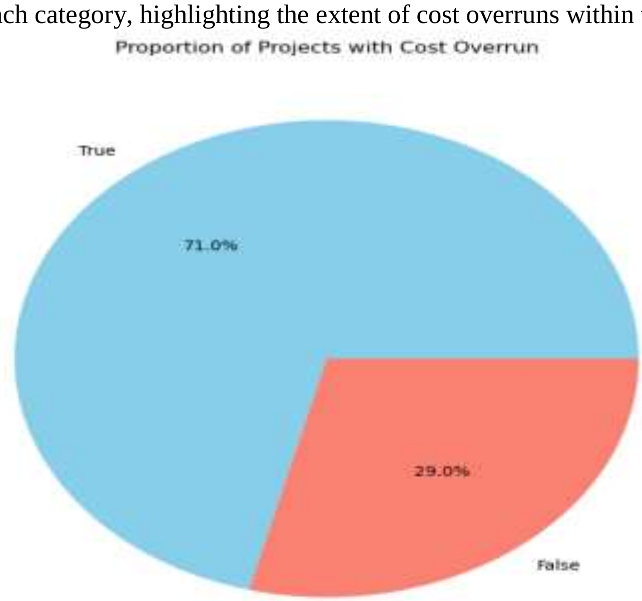

Figure 3: Number of Projects with Cost Overrun

International Journal of Engineering and Information Systems (IJEAIS)
ISSN: 2643-640X
Vol. 10 Issue 6 June - 2026, Pages: 24-33

Figure 4 displays a heatmap that contrasts, on the x-axis, the budgeted project costs - with, on the y-axis, the actual project expenses. The relationship between a certain planned cost range and the matching actual cost range is shown by each cell. The intensity is shown by the colour bar on the right, where lighter hues denote deviations and darker shades reflect a closer alignment between the projected and actual expenses. Projects with a budgeted budget between 31,000,000 and 40,000,000, as shown by a big filled region in that intersection, often align well with the same range in actual costs, according to the chart. On the other hand, variations are noted in different ranges. For instance, the corresponding full cell indicates that projects initially estimated to cost between 21,000,000 and 30,000,000 occasionally wind up costing 40,000,001 and more. The colour intensity in this chart, which is displayed in Figure 4, gives a visual indication of how closely project budgets match or diverge from their actual expenditures.

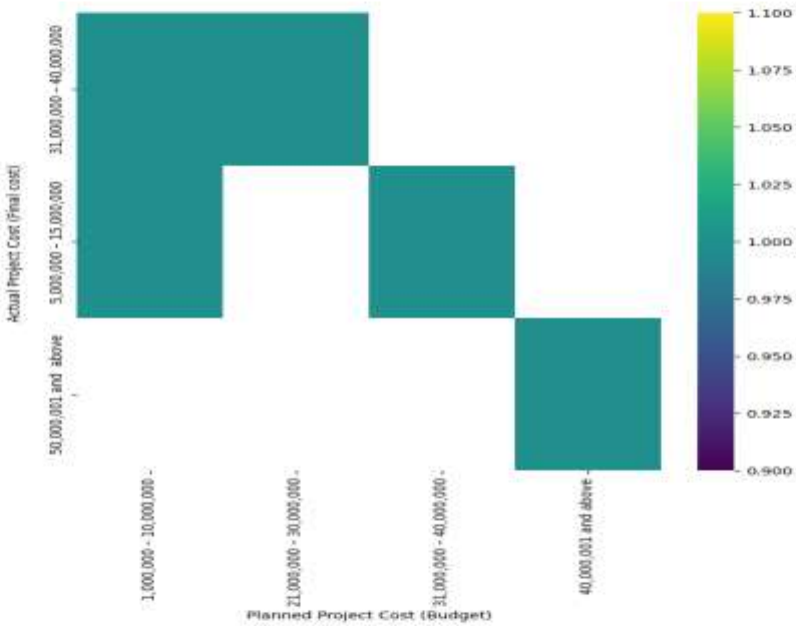

Figure 4: Distribution of the Projects Budget Costs by Project Actual Costs

Figure 5 shows a plot that illustrates the differences in cost overruns between the two project categories. The violin plot contrasts how cost overruns for "Commercial" and "Residential" projects are distributed. It demonstrates that whereas residential projects have a tighter range of cost overruns, commercial ones have a more diversified distribution. Each plot's width shows the density of data points, or the frequency at which a given cost overrun value occurs.

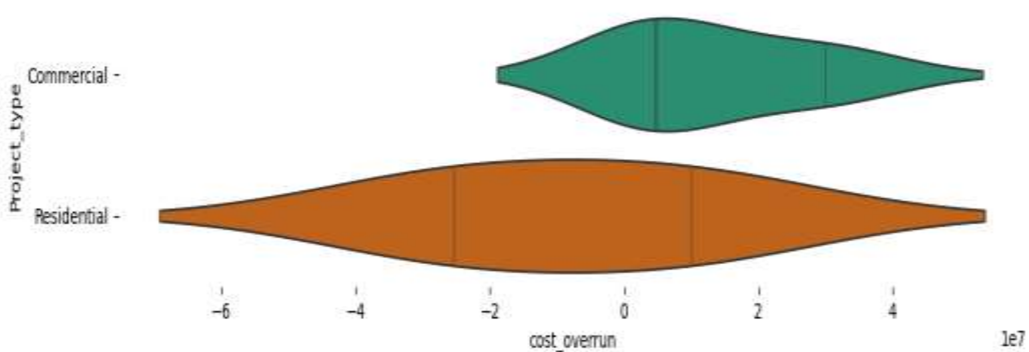

Figure 5: Distribution of Cost Overruns Between ‘Commercial’ and "Residential’’

The violin plot in Figure 5 shows how cost overruns are distributed among three categories of organizations: "Consulting Firm," "Government Parastatal," and "Contracting Firm." The vast distribution displayed by the "Government Parastatal" suggests a large range of cost overruns and notable variation in financial results. Alternatively, the "Contracting Firm" exhibits a narrow distribution with a zero centre, indicating that cost overruns are typically small and regular. For the organisation type "Consulting Firm," there is no data visible, which could mean that there are no documented cost overruns with the data available. As illustrated in Figure 6, government parastatals are more prone to cost overruns than contracting enterprises.

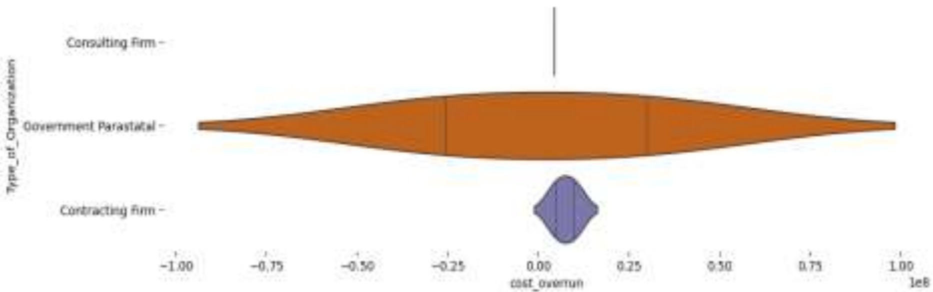

Figure 6: Distribution of Cost Overruns Among the Three Categories of Organization

3.3 Model Training and Evaluation

Figure 7 presents result of MSE, RMSE and R^2 . The MSE for the SVR model was 0.071, this value epitomizes the average of the squared differences between predicted and actual values. A lower MSE indicates better model accuracy. RMSE, calculated as the square root of the MSE, is 0.267 this metric represents the average magnitude of the prediction error in the units of the target variable, providing a clear indication of typical prediction errors. The R^2 score for the SVR model was 0.693, this indicates that approximately 69.3% of the variance in cost overruns is explained by the model. This score suggests that the SVR model achieved a good fit to the data, explaining significant portion of the variability.

```
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.inspection import permutation_importance
import matplotlib.pyplot as plt

# Prepare your features and target
target_column = 'Cost Overrun'
X = data.drop(columns=[target_column])  # Drop the target column
y = data[target_column]  # Target column

# Convert categorical features to numerical if necessary
X = pd.get_dummies(X)  # This will convert categorical columns to dummy variables

# Split the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Initialize and fit the SVR model
model = SVR()
model.fit(X_train, y_train)

# Predict and evaluate the model
y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)

print(f"MSE: {mse}")
print(f"RMSE: {rmse}")
print(f"R-squared: {r2}")
```

Figure 7: Training and Evaluation of Dataset Using SVR

3.4. Model Comparison

To further test the accuracy of the SVR, other machine learning languages were used in testing the same dataset. In evaluating the performance of various regression models for predicting cost overruns in construction projects, a Random Forest Regressor and Decision Tree Regressor were utilized. The performance metrics for the Three Regression Models is presented in Figure 8 and Table 3.

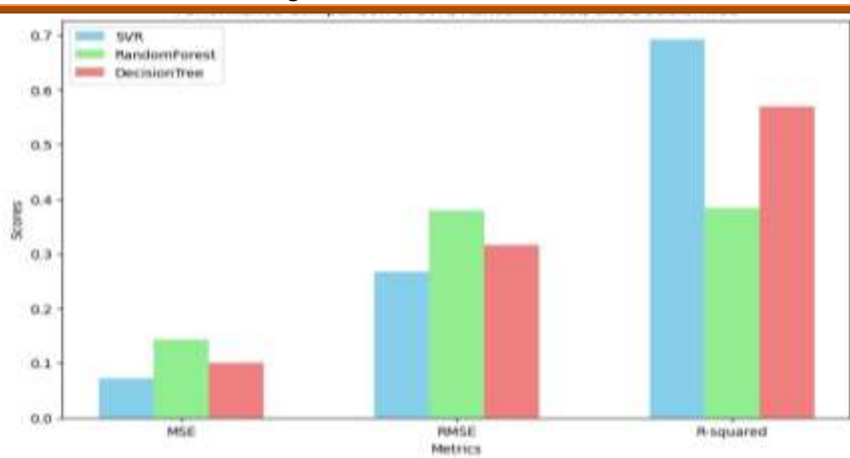

Figure 8: Graphical Representation for Model Comparison

Table 3: Tabular Performance Metrics for the Three Regression Models

Model	**Mean Squared Error (MSE)**	**Root Mean Squared Error (RMSE)**	**R-squared (R^2)**
Random Forest Regressor (SRF)	0.143	0.378	0.384
Decision Tree Regressor (DTR)	0.1	0.316	0.569
Support Vector Regressor (SVR)	0.071	0.267	0.693

From Figure 8, Table 3 is produced. From Figure 8 and Table 3, for Random Forest model, the MSE is 0.143, indicating the average squared error across all predictions. The RMSE for this model is 0.378, reflecting the typical magnitude of the prediction error and an R^2 value of 0.384, approximately 38.4% of the variability in cost overruns can be accounted for by the Random Forest model. For Decision Tree model MSE is 0.1; RMSE is 0.316 and R^2 score is 0.569, meaning that approximately 56.9% of the variance in cost overruns can be explained by the Decision Tree Regressor.

The Random Forest model demonstrated a reasonable fit to the data, capturing a significant portion of the variance in cost overruns while also indicating areas for potential improvement. Decision Tree Regressor demonstrated a strong ability to model and predict cost overruns, capturing a significant portion of the variance and reflecting the factors prompting cost overruns.

The SVR model outperformed both the Random Forest and Decision Tree models. The lowest MSE and RMSE values indicate that SVR had the smallest prediction error, and the highest R^2 score of 0.693 means that SVR explained approximately 69.3% of the variance in cost overruns. This superior performance suggests that SVR was particularly effective at modeling the data and capturing the underlying trends and relationships.

The SVR model demonstrated strong performance with a low MSE and RMSE, and a high R^2 score, indicating its effectiveness in predicting cost overruns and capturing the underlying patterns in the data. The evaluation of different regression models -Random Forest Regressor, Decision Tree Regressor, and Support Vector Regressor (SVR) provided insights into their performance and suitability for predicting cost overruns in construction projects. Figure 8 shows pictorial performance comparison of model been used in the evaluation.

In summary, the SVR model established the maximum performance among the three evaluated models with the lowest MSE and RMSE and the highest R^2 score. This indicates that SVR is the most suitable model for predicting cost overruns in the given dataset.

4. CONCLUSION AND RECOMMENDATION

4.1 Conclusion

This research complements the application of SVM in predicting cost overrun in building construction and underlines the potential benefits of the SVM system algorithms.

The following conclusion were made from this research:

i. The cost overrun of building construction projects was predicted using SVR, SRF and DTR based on 100 datasets collected from questionnaire.

ii. SVR, RFR and DTR all demonstrated a reasonable fit to the data, proving that they are good machine learning tools for predicting cost overrun

iii. SVR proved to perform excellently well in prediction of cost in this research work. with an accuracy 0.071, 0.267 and 0.693. The highest R^2 score of 0.693 means that SVR explained approximately 69.3% of the variance in cost overruns.

iv. In this research, there is 71% cost overrun in building in the studied construction projects; here was more cost overrun gotten from government parastatal; the actual cost (planned) for most government parastatal projects were lesser than the budget.

v. Cost overrun occurred in most projects because the planned project cost was lower than the actual project cost.

4.2 Recommendation

Notwithstanding that this study has desirable findings, future study could benefit from broader and more diversified datasets. Also, research should be employed to other buildings other than residential buildings.

REFERENCES

Abderahim S., Althea C. H., James R., Joseph J., Ian D. (2023). Data Preprocessing Using AutoML: A Survey. Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)., Accessed January, 2026 DOI 10.1109/CSCE60160.2023.00265. pp. 1619-1623.

Afzal, F., Yunfei, S., Nazir, M., and Bhatti, S. M. (2019). A review of artificial intelligence-based risk assessment methods for capturing complexity-risk interdependencies: Cost overrun in construction projects. *"International Journal of Managing Projects in Business*", 14, 300–328.

Alvaro A. A. F., Martin K., Nikolaos K., Pavel P., Norman W. P. and Rizos S. (2023) Data Preparation: A Technological Perspective and Review. SN Computer Science 4:425 Available at https://doi.org/10.1007/s42979-023-01828-8

Alshamrani, O.S. (2017). Construction cost pre- diction model for conventional and sustain- able college buildings in North America. *Journal of Taibah University for Science*, 11(2), 315-323

Ashiedu, B. I., Ogbuefi, E., Nwabekee, U. S., Ogeawuchi, J. C., & Abayomi, A. A. (2023). Strategic resource allocation in project and business units: Frameworks for telecom-finance integration. International *Journal of Multidisciplinary Research and Growth Evaluation, 4*(1), 1276–1288

Babalola, H. I., and Ojo, O. J. (2016). An investigation into factors affecting the performance of public construction projects in Ondo State, Southwestern, Nigeria. *Civil and Environmental Research,* 8(1), 1-10.

Bahamid, R. A., Doh, S. I., and Al-Sharaf, M. A. (2019). Risk factors affecting the construction projects in the developing countries. *"IOP Conference Series: Earth and Environmental Science"*. IOP Publishing, Kuantan, Malaysia, 1-4.

Cheng, M. Y., and Wu, Y. W. (2005). Construction conceptual cost estimates using support vector machine. *"22nd International Symposium on Automation and Robotics in Construction (ISARC)"* Citeseer. 1–5.

Gould, F. E. (2002). Managing the Construction Process: Estimating, Scheduling, and Project Control. Upper Saddle River, NJ: Prentice Hall.

Gurumayum, S. (2025). Understanding the causes of project delays and cost overruns in construction: A comprehensive analysis across literature, surveys, and case studies. *Journal of Civil and Construction Engineering*, *11*(1), 74–94.

Han, J., Kamber, M., and Pei, J. (2012). "Data mining: Concepts and techniques" (3rd ed.). University of Illinois at Urbana-Champaign.

Imran, C. A. B., Shakir, M. K., Umer, M., Imran, Z., Idrees, H. M. K., Ansari, Y., Imran, M., and Tariq, M. A. (2024). *Building the future: Applications of artificial intelligence in civil engineering*.

Juszczyk, M. (2018). Residential buildings conceptual cost estimates with the use of support vector regression. *"MATEC Web Conference",* 196, 04090. Accessed February, 2025 https://doi.org/10.1051/matecconf/201819604090

Khaertdinova, A., Maliashova, A., and Gadelshina, S. (2021). *Economic development of the construction industry as a basis for sustainable development of the country*. E3S Web of Conferences, 274, 10021.

Lopes, J. (2022). Construction in the economy and in national development. In *Research companion to construction economics*. Edward Elgar Publishing, United Kingdom. pp. 104–125.

Meharie, M. G., and Shaik, N. (2020). Predicting highway construction costs: Comparison of the performance of random forest, neural network and support vector machine models. *"Journal of Soft Computing in Civil Engineering",* 4(3), 103-112.

Petrusheva, S., Car-Pušić, D., and Zileska-Pancovska, V. (2019). Support vector machine-based hybrid model for prediction of road structures construction costs. *"IOP Conference Series: Earth and Environmental Science*, 222" (1), 012010. IOP Publishing.

Sha, M. K., Shahi, P. B., Pandit, R., and Pandey, A. (2017). Causes and effects of delays in construction projects. *IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE),* 14(2), 52-58. https://doi.org/10.9790/1684-1402065258

Shane, J. S., Molenaar, K. R., Anderson, S., and Schexnayder, C. (2009). Construction project cost escalation factors. "Construction Management and Economics, 25" (3), 221-229.

Soewin, E., and Chinda, T. (2018). Factors affecting construction performance: Exploratory factor analysis. *"IOP Conference Series: Earth and Environmental Science*, 140" (1), 349-355. IOP Publishing.

Xiao, X. (2021). *Improving the long-term use of case-based reasoning model in early construction cost estimation* (Doctoral dissertation, The Hong Kong Polytechnic University, Department of Building and Real Estate).

Yan, K., and Shi, C. (2010). Prediction of elastic modulus of normal and high strength concrete by support vector machine. *"Construction and Building Materials",* 24(1), 1479–1485.

Yang, C., Baabak, A., and Minsoo, B. (2018). Prediction of Unit Price Bids of Resurfacing Highway Projects through Ensemble Machine Learning. *Journal of Computing in Civil Engineering* 32(5), 04018043. DOI: 10.1061 (ASCE)CP.1943-5487.0000788.