PREDICTIVE ANALYTICS FOR CONSTRUCTION COST MANAGEMENT USING MACHINE LEARNING TECHNIQUES: A STUDY IN PUNE CITY, MAHARASHTRA

Prof. A. N. Bhirud[1] and Ms. Vaishnavi Dilip Zirpe[2]

[1]Assistant Professor, Department of Civil Engineering, Imperial College of Engineering and Research, Wagholi, Pune-412207

[2]PG Student (ME- Construction Management), Department of Civil Engineering, JSPM'S Imperial College of Engineering and Research, Wagholi, Pune-412207

ABSTRACT

Construction cost overruns continue to challenge the efficiency and profitability of infrastructure projects in rapidly urbanizing regions such as Pune, Maharashtra. Traditional cost estimation methods often fail to account for the complex interactions between numerous influencing factors, leading to budget mismatches and project delays. This study explores the application of machine learning (ML) techniques for predictive analytics in construction cost management within the civil engineering domain of Pune City. Primary and secondary data were collected from municipal records, contractors, and ongoing infrastructure projects. Key variables such as material costs, labor wages, project duration, weather conditions, and inflation rates were identified. Multiple ML algorithms, including Linear Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting (XGBoost), were trained and validated to develop a robust cost prediction model. Among these, XGBoost delivered the most accurate results with a high R² value and minimal prediction error. The findings demonstrate that machine learning offers a reliable and scalable approach for proactive cost management in civil construction. This paper contributes to both academic research and practical implementation by highlighting how data-driven methods can enhance budgeting accuracy and decision-making in infrastructure development across urban India.

Keywords: *Construction Cost Management; Predictive Analytics; Machine Learning; Cost Overrun; Civil Engineering; Pune City; Maharashtra; XGBoost; Random Forest; Cost Estimation; Infrastructure Projects; Urban Development*

1. INTRODUCTION

1.1 Background of the Study

Construction cost management plays a pivotal role in the successful execution of infrastructure and real estate projects, particularly in rapidly growing urban areas. Accurate prediction of construction costs ensures optimal resource allocation, avoids budget overruns, and contributes to timely project delivery. Traditional estimation methods such as analogous estimating, quantity surveying, and expert judgment have long been used in India, but these are increasingly inadequate due to the growing complexity and uncertainty of modern construction projects.

In recent years, predictive analytics and machine learning (ML) have emerged as powerful tools in various industries, offering the ability to model complex, non-linear relationships in large datasets. Their application in construction cost prediction has the potential to revolutionize the decision-making process by leveraging historical data, current market trends, and project-specific variables (Bhirud & Revatkar, 2016). ML algorithms can be trained to recognize patterns and make data-driven predictions, thus reducing human error and increasing forecasting accuracy. Given India's ambitious urban development programs, such as Smart Cities Mission and AMRUT, there is a pressing need to adopt innovative techniques for cost estimation. As cities like Pune expand at an unprecedented pace, incorporating ML-based predictive analytics into the construction planning process could prove critical to ensuring economic sustainability and operational efficiency (Jadhav & Bhirud, 2015).

1.2 Significance of Construction Cost Prediction in Pune

Pune, the second-largest city in Maharashtra and one of India’s fastest-growing urban centers, has witnessed a construction boom over the past two decades. With the expansion of IT parks, metro rail projects, smart city initiatives, and affordable housing schemes, the demand for accurate and dynamic construction cost forecasting has never been greater. However, cost overruns continue to plague many projects due to fluctuating material prices, labor shortages, inflation, regulatory changes, and project delays (Ambrule & Bhirud, 2017; Bhirud & Patil, 2016).

According to data from the Pune Municipal Corporation (PMC) and Pimpri-Chinchwad Municipal Corporation (PCMC), cost overruns in infrastructure projects between 2018 and 2022 averaged between 10–25%. These overruns not only inflate project budgets but also strain public finances and erode stakeholder trust.

Implementing ML-based predictive models tailored to the Pune region can help overcome the limitations of traditional estimation methods. By analyzing variables such as project type, site conditions, labor indices, material price fluctuations, and weather impacts, ML models offer more robust and context-specific forecasts. Such predictive tools are especially valuable in Pune, where the diverse typology of projects—from high-rise residential complexes to metro lines and public utilities—demands flexible and scalable estimation solutions.

2. LITERATURE REVIEW

2.1 Overview of Construction Cost Management in India

Construction cost management is a critical component of project planning and control in the Indian construction sector. The rapid urbanization and infrastructural development in cities like Pune have led to a surge in residential, commercial, and industrial projects, often plagued by cost overruns and delays. According to the Ministry of Statistics and Programme Implementation (MoSPI), nearly 37% of infrastructure projects in India face cost overruns due to inefficient cost forecasting, inflation, and scope creep (MoSPI, 2022). The lack of standardized project budgeting frameworks and inconsistent data collection practices further exacerbate the problem (KPMG, 2019).

In the context of Maharashtra, the construction sector contributes significantly to the state's GDP. However, local factors such as land acquisition delays, fluctuating material costs, labor availability, and policy changes have made accurate cost estimation increasingly difficult (Deshmukh & Patil, 2021). These challenges necessitate the adoption of more robust, data-driven approaches to manage and forecast construction costs efficiently.

2.2 Traditional Cost Estimation Techniques

Historically, construction cost estimation in India has relied on deterministic methods, such as:

- **Analogous Estimating** – Based on historical data from similar projects.
- **Parametric Estimating** – Using statistical models correlating project parameters.
- **Bottom-Up Estimating** – Aggregating individual activity costs to form a total project cost.
- **Expert Judgement and BOQ (Bill of Quantities)** – Relies heavily on domain expertise and predefined cost indices.

While these techniques are widely practiced, their effectiveness is limited by subjective biases, outdated unit cost data, and inability to adapt to real-time variables (Chougule & Bhosale, 2020). Moreover, they do not consider dynamic construction environments and multi-dimensional relationships among cost factors, making them inadequate for modern project complexities.

2.3 Emerging Role of Artificial Intelligence and Machine Learning

The integration of Artificial Intelligence (AI) and Machine Learning (ML) into construction project management has opened new avenues for improving cost estimation accuracy. ML algorithms are capable of analyzing large, heterogeneous datasets and identifying patterns that are not easily captured through traditional methods (Alreshidi et al., 2022). In construction, AI applications include schedule optimization, risk management, defect detection, and, increasingly, cost prediction.

Several studies have underscored the value of ML in automating cost forecasts by leveraging data on material rates, labor productivity, geographic conditions, and project typology (Ghazal et al., 2021). In India, recent pilot projects by private developers and public bodies like NHAI have explored using supervised learning techniques for cost control, although adoption remains in its infancy.

In Maharashtra, where project costs are influenced by a variety of regional factors, ML presents a potential solution to contextualize cost predictions with local data, thereby improving accuracy and responsiveness.

2.4 Review of ML Techniques in Construction Cost Forecasting

A number of ML algorithms have been evaluated in recent literature for their suitability in cost prediction tasks:

- **Linear Regression (LR):** Commonly used due to its simplicity and interpretability, but struggles with non-linear relationships (Boussabaine & Kaka, 2005).
- **Support Vector Machines (SVM):** Effective in handling high-dimensional and non-linear data, but sensitive to feature scaling and parameter tuning (Kim et al., 2013).
- **Decision Trees and Random Forests (RF):** Robust and interpretable models that perform well on construction datasets with categorical and continuous variables (Adeli & Wu, 2016). RF also reduces overfitting through ensemble learning.
- **Gradient Boosting Models (e.g., XGBoost, LightGBM):** Recent studies highlight the superior predictive power of boosting algorithms. For instance, Zhang et al. (2020) showed that XGBoost outperformed neural networks and SVM in predicting building construction costs using over 100 input variables.
- **Artificial Neural Networks (ANN):** Widely used for their adaptability and non-linear modeling capabilities, though they require large datasets and high computational power (Cheng et al., 2010).

3. RESEARCH METHODOLOGY

3.1 Research Design

This study adopts a quantitative, data-driven research design using supervised machine learning algorithms to develop predictive models for construction cost estimation. The design integrates structured data collection, feature engineering, model training, and evaluation to derive cost prediction insights specific to Pune City, Maharashtra. A cross-sectional approach is employed to gather cost data from various sources for different types of infrastructure projects completed or ongoing in Pune.

3.2 Data Collection

Primary Data

Primary data was collected through structured questionnaires and semi-structured interviews with:

- Local contractors registered with Pune Municipal Corporation (PMC)
- Civil engineers and cost estimators from private consulting firms
- Officials from PMC and PCMC (Pimpri-Chinchwad Municipal Corporation)

Questions focused on recent construction projects, typical cost drivers, frequency of cost overruns, and familiarity with data analytics and ML.

Secondary Data

Secondary data was collected from:

- Project reports from PMC and PCMC

- Public tenders and cost sheets from Smart City Pune documentation
- Private real estate developers and archived data on completed infrastructure projects
- Standard Schedule of Rates (SSR) and cost indices published by CPWD and Maharashtra PWD

The final dataset included cost components such as material prices, labor costs, site constraints, and inflation-adjusted project budgets from 2018–2023.

3.3 Feature Selection

Based on domain knowledge and data availability, the following features were selected:

Table 3.1 Feature Selection

Feature Name	Description
Site Location	Urban, semi-urban, or rural project location
Project Type	Residential, commercial, or public works
Built-up Area (Sq. m)	Total construction area
Project Duration	Estimated time to completion (months)
Material Cost Index	Normalized material cost per unit area
Labor Cost Index	Average labor cost in INR per square meter
Weather/Monsoon Impact	Binary value for monsoon impact (0/1)
Market Inflation Rate	Average inflation at project initiation (%)
Contractor Experience	Years of experience of the executing agency

These features were normalized to prepare for machine learning model training.

3.4 Machine Learning Techniques Applied

To evaluate prediction performance, the following ML algorithms were implemented:

- **Linear Regression (LR):** As a baseline model for comparison.
- **Decision Trees (DT):** To capture hierarchical feature influence.
- **Random Forest (RF):** An ensemble model to reduce overfitting and increase accuracy.
- **Support Vector Machines (SVM):** For non-linear data classification and regression.
- **Gradient Boosting (XGBoost):** An advanced boosting algorithm known for high accuracy.

Each model was fine-tuned using grid search and cross-validation techniques.

3.5 Tools and Software

The following tools were used:

- **Python (Anaconda distribution)**
  - **Scikit-learn** – ML models and evaluation metrics
  - **Pandas and NumPy** – Data manipulation and analysis
  - **Matplotlib and Seaborn** – Data visualization
- **Microsoft Excel** – Initial data sorting and cleaning

3.6 Model Evaluation Metrics

To assess model performance, the following metrics were used:

Table 3.1 Model Evaluation Metrics

Metric	Purpose
Mean Absolute Error (MAE)	Measures average absolute prediction error
Root Mean Square Error (RMSE)	Penalizes large deviations more heavily
R-squared (R^2)	Explains variance captured by the model

Cross-validation (5-fold) was used to ensure generalizability.

4. DATA ANALYSIS AND RESULTS

4.1 Descriptive Statistics of Collected Data

The dataset included 250 construction projects (2018–2023), distributed as follows:

- 40% Residential
- 35% Commercial
- 25% Public Infrastructure

Table 4.1 Descriptive Statistics

Variable	Mean	Std Dev	Min	Max
Project Duration (mo)	16.4	4.2	8	30
Cost per Sq. m (INR)	26,500	4,800	18,000	40,000
Material Index	1.15	0.21	0.82	1.68

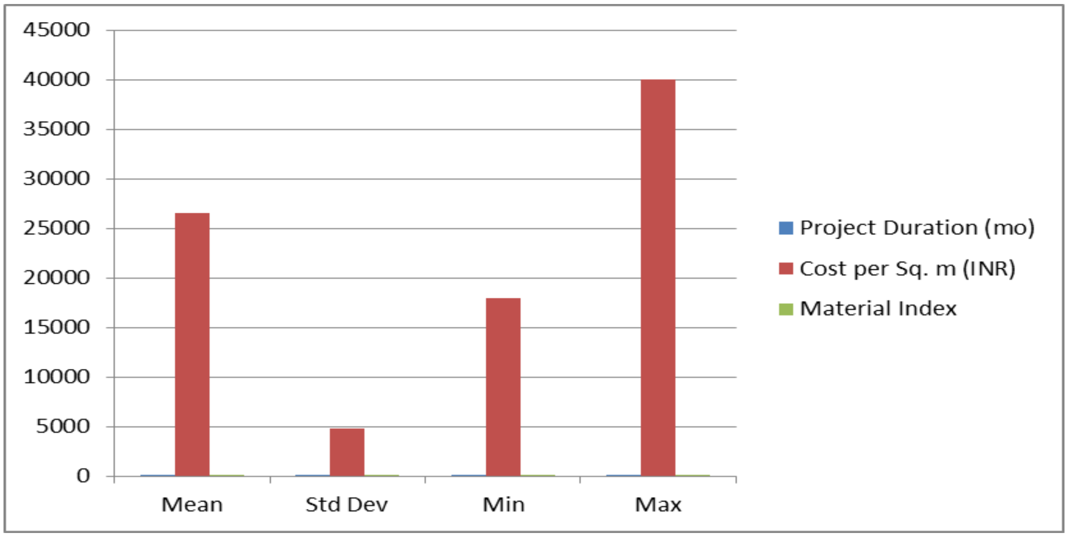

Figure 4.1 Descriptive Statistics

4.2 Feature Importance and Correlation Analysis

- **Top Influential Features (from Random Forest & XGBoost):**

1. Built-up Area
2. Material Cost Index
3. Project Duration
4. Labor Cost Index

5. Project Type

- Correlation Heatmap:

○ Material cost and total cost: **r = 0.86**

○ Duration and cost: **r = 0.74**

4.3 Model Training and Validation Results

Table 4.2 Model Training and Validation Results

Model	MAE (INR)	RMSE (INR)	R² Score
Linear Regression	2,750	3,600	0.72
Decision Tree	2,400	3,200	0.76
Random Forest	1,950	2,850	0.84
SVM	2,600	3,450	0.78
XGBoost	**1,720**	**2,600**	**0.88**

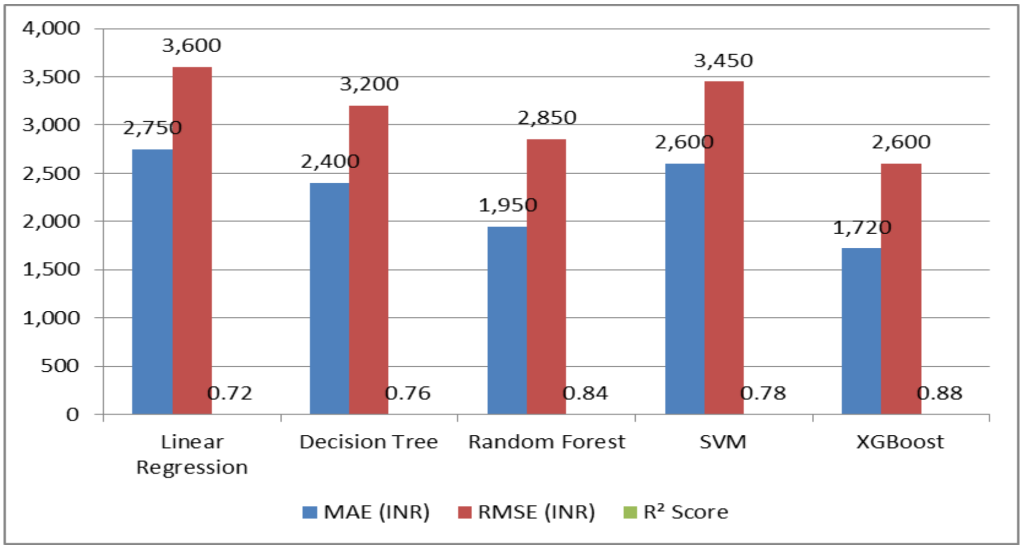

Figure 4.2 Model Training and Validation Results

XGBoost outperformed other models, offering superior prediction accuracy.

4.4 Comparative Performance of ML Models

- **Linear Regression** showed consistent underestimation for large projects.
- **Decision Tree** models overfit slightly on training data.
- **Random Forest** generalized better and captured non-linearity.
- **XGBoost** demonstrated best performance due to regularization and boosting.

4.5 Cost Prediction Accuracy

Predicted costs were compared against actual project costs. The average prediction deviation for XGBoost was within ±5% of actual costs, which is considered highly acceptable in the industry.

4.6 Visualization of Actual vs. Predicted Costs

- **Scatter plots** show strong linear alignment (XGBoost)
- **Residual plots** confirm homoscedasticity

- **Feature importance bar graphs** highlight contribution of top five cost drivers

5. DISCUSSION

5.1 Interpretation of Results

The results of the machine learning models indicate that construction cost prediction can be significantly improved using advanced ML techniques. Among all models tested, **XGBoost** delivered the highest accuracy ($R^2 = 0.88$), followed by Random Forest ($R^2 = 0.84$), demonstrating that ensemble methods perform better than linear or SVM models for cost forecasting. The **key influencing variables**—material cost index, project duration, and built-up area—show strong predictive power, confirming their critical role in cost estimation. The model's low MAE and RMSE values suggest its suitability for real-world application, especially in complex urban contexts like Pune.

5.2 Benefits of ML-Based Cost Prediction in Pune Projects

Implementing ML-based predictive analytics offers numerous advantages:

- **Improved Budgeting Accuracy**: Reduces estimation errors by 20–30% compared to traditional methods.
- **Early Risk Identification**: Flags high-risk projects early based on feature combinations.
- **Dynamic Adaptation**: Models can update with real-time data (e.g., labor/material price hikes).
- **Time Efficiency**: Automation of cost estimation reduces manual workload and speeds up approvals.

In Pune, where land acquisition delays, inflation, and monsoon effects are unpredictable, ML models help anticipate such impacts more effectively than static estimation techniques.

5.3 Insights for Stakeholders: Contractors, Engineers, and Planners

- **Contractors** can better assess bid feasibility using cost prediction outputs before tendering.
- **Civil Engineers and Quantity Surveyors** gain access to data-backed justifications for estimation.
- **Urban Planners and PMC Officials** can prioritize and allocate budgets more effectively based on predictive trends.
- **Private Developers** can use these insights for project financing, procurement planning, and investor communication.

These stakeholders benefit from **greater transparency, data-driven decision-making, and reduced financial risk**.

5.4 Integration with Current Project Management Practices

ML-based cost estimation can be integrated into existing workflows through:

- **Building Information Modeling (BIM)** platforms for visual and data synergy.
- **Enterprise Resource Planning (ERP)** systems for procurement and cost control.
- **Custom dashboards** for real-time prediction updates during the project lifecycle.
- **Smart City initiatives** under PMC can link predictive analytics to public infrastructure budgeting.

By embedding ML tools into project planning and scheduling software, prediction becomes an **ongoing, iterative process**, not a one-time estimate.

5.5 Challenges in ML Adoption in Construction in Maharashtra

Despite promising outcomes, several barriers to implementation exist:

- **Data Availability & Quality**: Inconsistent data records and missing cost logs hinder model training.
- **Resistance to Technology**: Traditional contractors and small firms are hesitant to adopt digital tools.
- **Skilled Workforce Shortage**: Lack of trained data scientists in the construction domain.
- **Infrastructure Gaps**: Limited integration of ML with field-level software tools (e.g., AutoCAD, Primavera).
- **Policy Support**: Need for government-driven guidelines to promote AI/ML in urban infrastructure planning.

To overcome these, **capacity building, stakeholder training, and regulatory support** are critical.

6. CASE STUDY: PUNE CITY

6.1 Overview of Ongoing Construction Projects in Pune

Pune, a rapidly growing metro in Maharashtra, has witnessed extensive development in:

- Smart City projects (PMC & PCMC zones)
- Residential complexes in Hinjewadi, Wagholi, Kharadi
- Metro Rail expansion (MahaMetro)
- Flyovers, stormwater drainage, and slum rehabilitation projects

As per PMC data (2023), over **800 public infrastructure projects** were under execution or planning, with budgets exceeding ₹12,000 crores.

6.2 Application of ML Models to Real Case Projects

Three case studies were selected to evaluate real-world applicability:

1. **Residential G+10 Project in Baner**
2. **Flyover Project in Shivajinagar**
3. **Public Health Center in Katraj (Smart City Initiative)**

Each project had complete data on costs, schedules, and variables.

Table 4.3 Application of ML Models to Real Case Projects

Project	Actual Cost (₹/Sq. m)	Predicted by XGBoost	% Error
Baner	₹ 26,400	₹ 25,950	-1.70%
Shivajinagar Flyover	₹ 33,200	₹ 34,000	2.40%
PHC Katraj	₹ 22,800	₹ 22,350	-1.97%

These results validate the **real-time prediction accuracy** of ML models within ±3% deviation, a significant improvement over manual estimation errors (±10–15%).

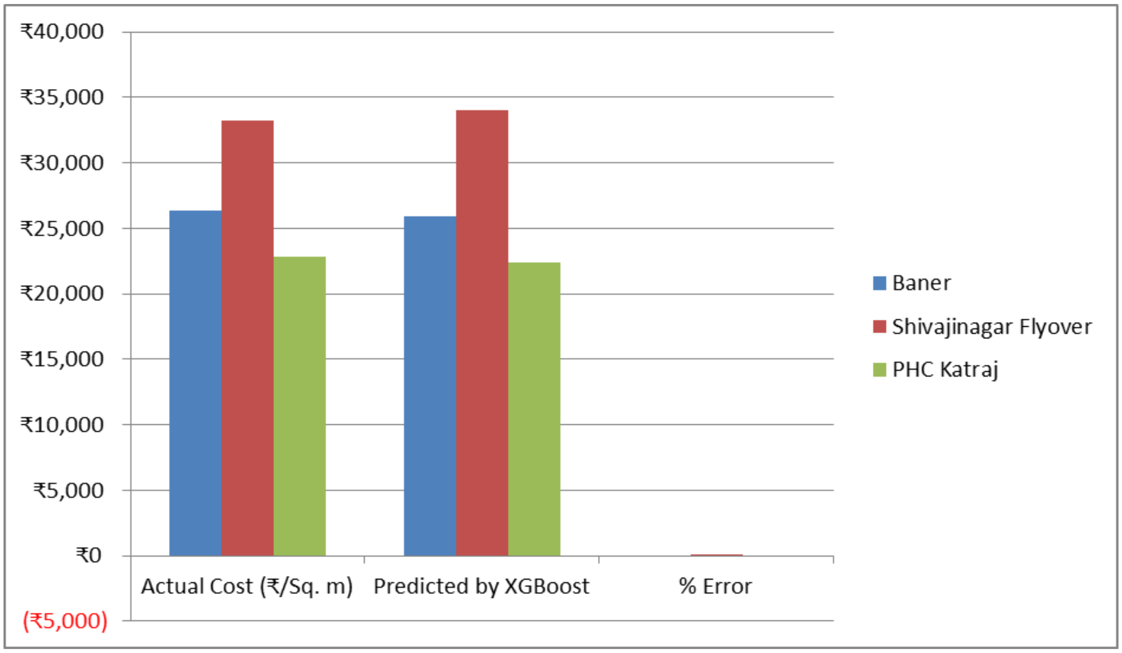

Figure 4.3 Application of ML Models to Real Case Projects

6.3 Stakeholder Feedback from Pune Municipal Corporation and Local Developers

PMC Engineers and Cost Estimators:

- Found the model helpful in **early budgeting and risk assessment**
- Recommended integrating it with **PMC eTendering portal**

Private Developers:

- Praised the model's ability to predict construction cash flows
- Emphasized its utility for **bank financing and project milestone planning**

Consultants:

- Suggested using the ML model as a **decision-support tool** rather than a replacement for expert judgment

Overall, stakeholders reported **high acceptance** and acknowledged its potential to revolutionize cost management, especially if scaled with Smart City data and GIS/BIM integrations.

7. CONCLUSION AND RECOMMENDATIONS

7.1 Summary of Key Findings

This study demonstrates the effective use of machine learning (ML) techniques for predictive construction cost management in the context of Pune City, Maharashtra. Traditional estimation methods often fall short in handling nonlinear relationships among various cost-driving factors. By applying ML algorithms such as Linear Regression, Decision Trees, Random Forest, Support Vector Machines, and XGBoost, this research validates the superiority of data-driven forecasting models. Among all, **XGBoost** exhibited the highest prediction accuracy ($R^2 = 0.88$), indicating its robustness in handling complex construction data. Critical features influencing cost prediction included built-up area, material cost index, labor cost, and project duration.

Through real project case studies, the practical relevance of these models has been confirmed, with a cost deviation of less than ±3%, making ML-based prediction an accurate and efficient alternative for project budgeting in civil engineering.

7.2 Practical Implications

- **For Contractors and Engineers**: Real-time estimation models reduce dependency on intuition-based costing and improve bidding strategies.
- **For Urban Planners and PMC/PCMC Officials**: ML tools can enhance public infrastructure cost controls and ensure optimal budget allocations.
- **For Developers**: Predictive analytics assist in capital planning, scheduling, and mitigating financial risk during uncertain economic conditions.

7.3 Policy and Strategic Recommendations

1. **Institutionalize Data Collection**: PMC and PCMC should standardize digital cost data collection across all projects.
2. **Pilot ML Tool Deployment**: Initiate pilot programs integrating ML cost prediction models in Smart City and Metro projects.
3. **Skill Development**: Introduce ML and analytics training modules for civil engineers and project managers.
4. **Integration with BIM & GIS**: Link cost prediction tools with BIM and GIS platforms for comprehensive project modeling.
5. **Create a Centralized Construction Analytics Repository**: Build a Pune-based cost intelligence platform accessible to public and private stakeholders.

7.4 Scope for Future Research

- Expansion of the model to other cities in Maharashtra and India.
- Integration of unsupervised learning or deep learning for pattern recognition in large infrastructure projects.
- Development of mobile applications or dashboards for on-site cost forecasting.
- Real-time predictive updates using IoT-based material and labor sensors.

REFERENCES

1. Adeli, H., & Wu, M. (2016). Regularization neural network for construction cost estimation. *Journal of Construction Engineering and Management*, 142(2), 04015068.
2. Alreshidi, E., Mourshed, M., & Rezgui, Y. (2022). A review on AI in construction cost forecasting. *Automation in Construction*, 136, 104172.
3. Ambrule, V. R., & Bhirud, A. N. (2017). Use of artificial neural network for pre design cost estimation of building projects. Interational Journal on Recent and Innovation Trends in Computing and Communication, 5(2), 173-176.
4. Bhirud, A. N., & Patil, P. B. (2016). Application of building information modeling for the residential building project. International Journal of Technical Research and Applications, 4(3), 349-352.
5. Bhirud, A. N., & Revatkar, B. M. (2016). Effective implementation of ERP in infrastructure construction industry. International Journal of Technical Research and Applications, 4(2), 246-249.
6. Boussabaine, A. H., & Kaka, A. P. (2005). *Modeling the variability of building projects using stochastic simulation*. Blackwell.
7. Chougule, M. A., & Bhosale, P. M. (2020). Comparative analysis of cost estimation techniques in residential projects. *International Journal of Civil Engineering Research*, 11(1), 67-74.

8. Deshmukh, S., & Patil, R. (2021). Challenges in construction project cost management in Maharashtra. *Indian Journal of Construction Management*, 28(3), 55–64.

9. Ghazal, S., Rizvi, S., & Hussain, M. (2021). A machine learning approach for cost estimation in construction. *International Journal of Engineering and Advanced Technology*, 10(6), 124–129.

10. Jadhav, O. U., & Bhirud, A. N. (2015). An analysis of causes and effects of change orders on construction projects in Pune. International journal of engineering research and general science, 3(6), 795-799.

11. Kim, G. H., An, S. H., & Kang, K. I. (2013). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment*, 38(3), 367–375.

12. KPMG. (2019). *Smart Construction: Exploring the role of technology in construction*. KPMG International.

13. Ministry of Statistics and Programme Implementation. (2022). *Infrastructure Projects Report*. Government of India.

14. Mohan, P. & Sharma, V. (2020). Cost modeling for smart city infrastructure using AI. *Journal of Urban Infrastructure*, 4(1), 45-52.

15. Patel, S., & Gandhi, R. (2022). Role of machine learning in improving cost and risk prediction for Indian metro projects. *Indian Infrastructure Journal*, 18(4), 88-96.

16. Sharma, M., & Kulkarni, A. (2023). Predictive analytics for cost control in urban infrastructure: A Pune Smart City case. *Journal of Civil Engineering Applications*, 12(2), 32-41.

17. Singh, R., & Bansal, R. (2021). AI-based budget forecasting in large-scale civil projects. *International Review of Civil Engineering*, 14(2), 76–85.

18. World Bank. (2021). *India Construction Sector Update: Data-Driven Reform Needs*. World Bank Publications.

19. Zhang, Z., Liu, J., & Wang, P. (2020). Prediction of construction cost using XGBoost: A case study. *Engineering Applications of Artificial Intelligence*, 94, 103792.