Literature review on cost prediction of construction projects

We reviewed 9 recent papers. In the first section, the list of these papers and an overview of the entire research is presented. The second section provides more details about these papers.

A) List of papers that are studied in this task:

1. Kusonkhum, Wuttipong, et al. "Government construction project budget prediction using machine learning." *Journal of Advances in Information Technology Vol* 13.1 (2022).

2. Meharie, Meseret Getnet, et al. "Application of stacking ensemble machine learning algorithm in predicting the cost of highway construction projects." *Engineering, Construction and Architectural Management* 29.7 (2022): 2836-2853.

3. Mahmoodzadeh, Arsalan, Hamid Reza Nejati, and Mokhtar Mohammadi. "Optimized machine learning modeling for predicting the construction cost and duration of tunneling projects." *Automation in Construction* 139 (2022): 104305.

4. Kovačević, Miljan, et al. "Construction cost estimation of reinforced and prestressed concrete bridges using machine learning." *Građevinar* 73.01 (2021): 1-13.

5. Rafiei, Mohammad Hossein, and Hojjat Adeli. "Novel machine-learning model for estimating construction costs considering economic variables and indexes." *Journal of construction engineering and management* 144.12 (2018): 04018106.

6. Alshboul, Odey, et al. "Extreme gradient boosting-based machine learning approach for green building cost prediction." *Sustainability* 14.11 (2022): 6651.

7. Pham, T. Q. D., T. Le-Hong, and X. V. Tran. "Efficient estimation and optimization of building costs using machine learning." *International Journal of Construction Management* 23.5 (2023): 909-921.

8. Elmousalami, Haytham H. "Comparison of artificial intelligence techniques for project conceptual cost prediction: A case study and comparative analysis." IEEE Transactions on Engineering Management 68.1 (2020): 183-196.

9. Elfahham, Yasser. "Estimation and prediction of construction cost index using neural networks, time series, and regression." *Alexandria Engineering Journal* 58.2 (2019): 499-506.

Overview:

Research	**ML model used**	**Accuracy (%)**	**Goal**
1	KNN	86	Over-budget prediction
2	Ensemble learning (LR, SVM, ANN)	87	Cost of highway construction
3	LR, GPR, SVR, DT	96	Cost of tunneling projects
4	ANN, RF, SVR, GPR	98	cost of reinforced and prestressed concrete bridges
5	DBM, BPNN, SVM	87-100	Considering economic indexes
6	XGBOOST, DNN, RF	96	Green building cost prediction
7	ANN, GBR, XGBOOST	99	Optimization for better accuracy
8	ANN, fuzzy models, ensemble methods,hybrid methods	92	Developing parametric cost model
9	ANN, LR, Autoregressive	-	Prediction of construction cost index (CCI)

- Only one dataset is available (paper 7). Most of them have prepared their use case dataset using government documents
- None of them have provided the codes but their implementations are very straight forward with (almost) no complexity. We can implement their models easily with python libraries like: scikit-learn, Pytorch and Keras

B) Details

1- paper title: Government Construction Project Budget Prediction Using Machine Learning

- **Research Goal:** The research goal of this study is to investigate the feasibility and effectiveness of employing machine learning (ML) algorithms, particularly the k-Nearest Neighbors (KNN) algorithm, to predict over-budget construction projects based on data collected from the traditional system of the Thai government. By utilizing attributes such as department name, site location, method of procurement, and type of project, the aim is to develop a predictive model capable of identifying projects at risk of exceeding their allocated budgets. Additionally, the study seeks to assess the accuracy of the developed model and demonstrate its potential utility for government agencies in improving project cost estimation and monitoring. Through this research, the goal is to showcase how ML techniques can be applied to relatively small datasets from government sources, highlighting their efficiency in enhancing project management processes despite the challenges posed by traditional data collection methods and infrastructure limitations.

- **Machine Learning Method:** The study utilizes the k-Nearest Neighbors (KNN) algorithm, to develop a predictive model.

- **Type of Data:** The dataset comprises information on 692 construction projects completed in Thailand in 2019.
    - **Data Features:**
        - **Location:** This attribute includes 72 factors representing different cities in Thailand where projects are executed.
        - **Department Name:** There are 48 factors representing different departments handling construction projects.
        - **Type of Project:** This attribute has three factors representing different types of projects such as building, roads, and irrigation projects.
        - **Procurement Method:** There are three factors representing different methods of procurement, including bidding methods.
        - **Case of Projects**: This attribute indicates whether a project is under budget or over budget according to the Thai government policy.
    - **Data Collection:** The data were collected from the traditional system of the Thai government, implying that they were sourced from government records or databases.

- **ML Model Performance:** The KNN model achieves an accuracy (precision) of 0.86 in predicting over-budget projects.
    - Evaluation of the proposed method:

input	Accuracy metrics				
Output		precision	recall	f1-score	support
	No	0.86	0.99	0.92	117
	Yes	0.75	0.14	0.23	22
	Accuracy			0.86	139
	Macro Avg	0.8	0.56	0.58	139
	Weighted Avg	0.84	0.86	0.81	139

- **important notes:** The developed ML model offers potential benefits for the Thai government by assisting in the identification and mitigation of budget overruns in construction projects, contributing to improved cost management and budget planning.

2- paper title: Application of stacking ensemble machine learning algorithm in predicting the cost of highway construction projects

- **Research Goal:** The research goal of this study is to apply a stacking ensemble machine learning algorithm for predicting the cost of highway construction projects. Specifically, the aim is to develop a predictive model that combines three distinct base predictive models - linear regression, support vector machine, and artificial neural network - using a gradient boosting algorithm as a meta-regressor. The ultimate objective is to enhance the accuracy of cost prediction for highway construction projects in Ethiopia, providing estimators and practitioners with an effective and reliable tool for preliminary cost estimation. This research seeks to address the need for more accurate cost prediction methods in the construction industry, particularly in the context of highway projects, and highlights the application of machine learning algorithms in forecasting construction costs.

- **Machine Learning Method:** The proposed model utilizes a stacking ensemble approach, where three base predictive models (linear regression, SVM, and ANN) are combined using the gradient boosting algorithm as a meta-regression.

- **Type of Data:** The study focuses on data related to highway construction projects, including information such as project specifications, location, and other relevant features.
    - The dataset consists of historical cost data from 117 road projects undertaken by the Ethiopian Road Authority (ERA) between January 1, 2006, and December 30, 2018.
    - The data include variables such as the number of bridges, inflation rate, terrain type, project type, and project cost. These variables were identified as the most significant factors affecting the accuracy of cost estimates for Ethiopian highway construction projects.
    - The dataset comprises 108 road projects after excluding nine projects with incomplete records.
    - The data are divided into training sets (80% of projects) and test sets (20% of projects).

- **Key Findings:**
    - The stacking ensemble model accurately predicts the final project cost with minimal prediction error.
    - Comparison with individual models shows that the stacking ensemble model consistently outperforms them, producing significantly more accurate results.
    - The stacking ensemble model achieves an improvement of 86.8%, 87.8%, and 5.6% in accuracy compared to linear regression, support vector machine, and artificial neural network models, respectively, based on root mean square error values.
    - Performance measures for stacking ensemble cost model

	Training	**Testing**
R-squared	0.938	0.978
MAE	0.131	0.111
MSE	0.033	0.046
RMSE	0.181	0.215

- **important notes:** The findings of the study reveal that the developed stacking ensemble model yields highly accurate predictions of final project costs, with a very small prediction error value. The model outperforms individual models such as linear regression, support vector machine, and artificial neural network in terms of accuracy, as evidenced by lower root mean square error values.

3- paper title: Optimized machine learning modeling for predicting the construction cost and duration of tunneling projects

- **Research Goal:** The study aims to predict the duration and cost of tunneling projects using machine learning techniques.

- **Machine Learning Method:** Four machine learning methods are employed: linear regression (LR), Gaussian process regression (GPR), support vector regression (SVR), and decision tree (DT).

- **Type of Data:** The analysis involves 350 datasets with 16 input parameters, including factors like drilling machinery system and groundwater.

- **Key Findings:**
  - LR exhibits the highest prediction performance among the tested models.
  - Sensitivity analysis identifies the most influential parameters affecting tunnel duration and cost.

- **important notes:**
  - The study contributes to improving accuracy in tunnel construction predictions by employing machine learning techniques, addressing limitations of previous studies by considering multiple parameters from various tunnels and conducting sensitivity analysis to identify critical factors.

4- paper title: Construction cost estimation of reinforced and prestressed concrete bridges using machine learning

- **Research Goal:** The research goal of this study is to conduct a comprehensive comparative analysis of seven state-of-the-art machine learning techniques for the estimation of construction costs of reinforced-concrete (RC) and prestressed-concrete (PC) bridges. The primary objective is to evaluate the effectiveness and performance of various machine learning models in accurately predicting the construction costs of bridge projects. By investigating different machine learning techniques, including artificial neural networks (ANN), ensemble methods, regression tree ensembles (such as random forests, boosted and bagged regression trees), support vector regression (SVR), and Gaussian process regression (GPR), the study aims to identify the most suitable and accurate modeling approach for cost estimation in the context of bridge construction projects. Additionally, the research seeks to address the limitations and biases associated with traditional linear regression models and complex neural network models by exploring alternative machine learning methods that offer improved interpretability and performance in cost prediction tasks.

- **Machine Learning Method:** Seven machine learning techniques are evaluated in the study, including:
  - Artificial Neural Networks (ANN)
  - Ensembles of ANNs
  - Regression Tree Ensembles (Random Forests, Boosted and Bagged Regression Trees)
  - Support Vector Regression (SVR)
  - Gaussian Process Regression (GPR)
  - Gradient Boosting Regression

- **Type of Data:**
    - The dataset used in the study comprises project and contract documentation of reinforced-concrete (RC) and prestressed concrete (PC) bridges constructed along Corridor X, a vital Pan-European transport route connecting multiple countries.
    - **Geographical Scope:** The bridges are located at the eastern and southern legs of Corridor X in Serbia,
        - Spanning across regions including Austria, Hungary, Slovenia, Croatia, Serbia, Bulgaria, Republic of North Macedonia, and Greece.
    - **Number and Types of Bridges:** The dataset includes information on 181 constructed highway bridges, consisting of:
        - 104 bridges with cast in situ RC superstructure
        - 77 bridges with PC superstructure (prefabricated or cast in situ)
        - 148 bridges carrying the motorway and 33 overpasses not carrying the motorway
    - **Contract Details:** Contracts for all bridges were signed between September 2009 and June 2014, with a total contract value exceeding EUR 100 million.
    - **Construction Costs:** Analysis of project costs reveals that approximately 77.41% of all construction costs are related to steel and concrete.
    - **Input Variables:** The dataset includes nine input variables used for modeling construction costs of RC and PC bridges. These variables are:
        - Total bridge span length
        - Bridge width
        - Average pier height
        - Foundation type (binary variable: 1 for deep foundations, 0 for shallow foundations)
        - Average bridge span (derived variable)
        - Type of bridge construction (binary variable: 1 for PC span superstructure, 0 for cast in situ RC span superstructure)
        - Gross salary index
        - Quarried aggregate price index
        - Steel price index
    - **Output Variable:** The estimated construction cost per unit area (€/m2) serves as the output variable in the dataset.

- **Key Findings:** The study evaluates the performance of various machine learning techniques for construction cost estimation.
    - The results reveal the strengths and weaknesses of each method, offering insights into their suitability for practical applications in cost estimation.
    - Their results show the superiority of Gradient Boosting algorithm
    - Performance metrics of different models

Model/criteria	RMSE	MAE	R	MAPE
MLP-ANN-9-10-1	160.75	115.48	0.7	21.66
MLP-ANN ansambl	96.45	71.71	0.88	13.04
Bagging	121.50	88.72	0.8	15.76
Random Forest	129.05	93.53	0.79	16.58
Gradient Boosting	96.03	67.15	0.89	12.03
SVR-RBF	109.32	68.25	0.86	12.03
GPR ARD-Eksponecijalni	95.98	63.25	0.89	11.6

- **important notes:** The research contributes to the field by offering a comprehensive comparative analysis of machine learning techniques for construction cost estimation. By exploring a diverse range of methods, including some previously underutilized approaches like GPR, the study provides valuable insights for improving the accuracy and reliability of cost estimation models in the construction industry.

5- paper title: Novel Machine-Learning Model for Estimating Construction Costs Considering Economic Variables and Indexes

- **Research Goal:** The research goal of the study is to develop an innovative construction cost estimation model using advanced machine learning concepts and incorporating economic variables and indexes (EV&Is). The model aims to address the complex nature of construction cost estimation, which is influenced by various factors such as project locality, type, construction duration, scheduling, and economic fluctuations captured by **EV&Is**. Specifically, the study proposes a data structure that integrates physical and financial variables of real estate units along with EV&Is variables affecting construction costs. The model utilizes an unsupervised deep Boltzmann machine (**DBM**) learning approach combined with a softmax layer (DBM-SoftMax) to extract relevant features from the input data. Additionally, a three-layer back-propagation neural network (**BPNN**) or another regression model, such as support vector machine (SVM), is employed to convert the trained unsupervised DBM into a supervised regression network (**DBM-BPNN or DBM-SVM**). The goal is to enhance the effectiveness and accuracy of conventional BPNN and

SVM models by leveraging the insights extracted from the EV&Is. Furthermore, a sensitivity analysis is conducted within the algorithm to optimize the model's performance considering the impact of EV&I factors at different times (time lags).

- **Machine Learning Method:** The proposed model utilizes an unsupervised deep **Boltzmann machine (DBM)** learning approach along with a softmax layer (DBM-SoftMax), and a three-layer **back-propagation neural network (BPNN)** or **support vector machine (SVM)**.
    - The **DBM-SoftMax** extracts relevant features from the input data,
    - The **BPNN** or **SVM** transforms the trained unsupervised DBM into a supervised regression network,
        - Thus improving the effectiveness and accuracy of both conventional BPNN and SVM.

- **Type of Data:** The data structure incorporates a set of physical and financial (P&F) variables of real estate units as well as a set of EV&I variables affecting construction costs.
    - The dataset includes information on project:
        - Locality, type,
        - Construction duration,
        - Scheduling
        - Extent of recycled materials usage
        - Economic variables and indexes such as liquidity, wholesale price index, and building services index.
        - Some of the features of the dataset and their description:

Variable	Description	Unit
	P&F factors	
1	Project locality defined in terms of zip code	N/A
2	Total floor area of the building	m^2
3	Lot area	m^2
4	Total preliminary estimated construction cost based on the price at the beginning of the project	Dollars
	EV&I variables in each non-overlapping time lag	
1	Number of building permits issued	N/A
2	BSI for a preselected base year	N?A

3	WPI of building materials for the base year	N/A
4	Total floor areas of building permits issued by the city or municipality	m^2

- **important notes:** This paper presents an innovative approach to construction cost estimation, addressing the challenges posed by fluctuating economic variables and indexes.
    - By incorporating advanced machine-learning techniques and considering EV&Is, the proposed model offers a more realistic and accurate estimation of construction costs, which is crucial for effective project planning and budgeting.

- **Key Findings:**
    - The model combines DBM-SoftMax with BPNN or SVM, resulting in significantly lower cost estimation errors compared to models utilizing only BPNN or SVM.
    - A sensitivity analysis within the algorithm considers the impact of EV&I factors at different times (time lags), enhancing the model's predictive capability.
    - The proposed model demonstrates superior effectiveness and accuracy in estimating construction costs for low- and midrise buildings, outperforming conventional machine-learning models.
    - The results of the study involved comparing the performance of different models, including the proposed DBM-BPNN and DBM-SVM models, with their respective counterparts using only BPNN or SVM.
        - The evaluation metrics used included mean square error (MSE) as a measure of prediction accuracy.
        - For the DBM-BPNN model, training accuracies ranged from 87.6% to 100.0%, with the second network showing smaller MSEs compared to the first, suggesting potential overfitting in the latter.
        - Time lag analysis revealed a preference for a 4-quarter delay between changes in economic variables and their impact on construction costs.
        - DBM-SVM model outperformed SVM-only models, with consistently lower MSEs across different verification-to-training ratios.
        - The findings underscored the effectiveness of the proposed models in improving cost estimation accuracy, particularly when incorporating economic variables and indexes.

6- paper title: Extreme Gradient Boosting-Based Machine Learning Approach for Green Building Cost Prediction

- **Research Goal:** The research goal of the study is to **develop accurate and robust machine learning-based algorithms for predicting green building construction costs.** Traditional cost estimation models may not adequately address the complexities of green building projects, which are designed to be environmentally friendly and incorporate technologies to reduce their environmental and societal impacts. The study aims to bridge this gap by designing innovative cost forecasting models specifically tailored to **green building construction**. By considering **influential attributes related to both soft and hard costs**, such as **green building management, environmental aspects, material use, water conservation, and energy efficiency**, the proposed machine learning models seek to provide stakeholders with accurate cost benchmarks to enhance decision-making in the bidding and awarding processes. Ultimately, the research aims to strengthen competition among bidders, improve the bid winner selection process, minimize financial and legal conflicts, and promote the successful delivery of green construction projects.

- **Machine Learning Methods:** The paper employs state-of-the-art machine learning algorithms, including extreme gradient boosting (**XGBOOST**), deep neural network (**DNN**), and random forest (**RF**).
    - These algorithms are selected for their ability to handle complex data patterns and to provide robust predictions.

- **Type of Data:** The data utilized in this study comprise a comprehensive set of attributes related to both **soft and hard costs associated with green building** projects.
    - **Land Costs:** Expenses associated with:
        - Land ownership, including:
            - Transfer of ownership
            - Land purchase
            - Site clearance,
            - Direct costs of building implementation such as civil and structural works.
    - **Construction Costs:**
        - Actual expenses related to physical construction activities
            - labor, materials, and equipment.

- **Soft Costs:**
    - Administrative Costs:
        - Indirect expenses associated with administrative tasks such as planning, documentation, and marketing.
    - Design Costs: Expenses related to architectural work, design planning, and certification services.

- **Green Building Elements:**
    - Green Features: Variables related to the incorporation of environmentally sustainable elements and technologies in building design and construction.

- **Economic Performance Metrics:**
    - Cost Metrics
    - Initial construction cost.

- **Location-specific Factors:**
    - Geographical Variables: Factors modified depending on location to ensure consistency in results for comparison purposes.

- **External Influences:**
    - Regulatory Factors: External support, technical requirements, and regulatory compliance measures impacting construction costs.

- **Other Influential Features:**
    - People and Technological Factors:
        - Human resources
        - Technological advancements
        - specific project requirements influencing construction costs.

- The dataset for green building cost prediction was gathered from various sources,
    - Journals
    - Websites of green building councils, and other related websites.

- The information was gathered between 2010 and 2020

- **important notes:** This research holds significant implications for stakeholders involved in green building construction projects.
    - By developing accurate prediction models, it aims to enhance decision-making processes during bidding and awarding procedures.
    - The application of machine learning techniques in this context contributes to the advancement of automation within the green construction industry.

- **Key Findings:** The key findings of the study demonstrate the efficacy of machine learning algorithms, particularly XGBOOST, in accurately predicting green building costs.
    - Evaluation metrics:
        - Mean Absolute Error (MAE)
        - Mean Squared Error (MSE)
        - Mean Absolute Percentage error (MAPE)
        - Coefficient of Determination (R^2)

    - **XGBOOST** outperforms other algorithms with an accuracy of 0.96, followed by DNN (0.91), and RF (0.87).

K-fold cross validation	Regression model	Performance evaluation metrics			
		MAE	**MSE**	**MAPE**	**R-squared**
K = 3	XGBoost DNN RF	132 238 408	152 316 527	27 51 56	94 89 86
K = 5	XGBoost DNN RF	92 196 378	132 284 507	19 32 40	96 91 87
K = 7	XGBoost DNN RF	118 212 389	141 301 516	23 43 50	95 90 86

7- paper title: Efficient estimation and optimization of building costs using machine learning

- **Research Goal:** The research goal is to develop a **comprehensive machine learning (ML) and optimization framework** for estimating building costs in the construction sector, addressing four main tasks identified by the partner company. These tasks include quick and accurate estimation of building costs to enhance operational efficiency and client communication, identification of realistic building options within budget to ensure profitability and competitiveness, rapid estimation of required resources to streamline operations and improve cost savings, and optimization of building costs under constraints to accommodate diverse customer preferences while minimizing expenses. By leveraging available data and deploying supervised ML models and optimization techniques, the study aims to provide practical solutions for improving cost estimation, operational efficiency, revenue generation, and customer satisfaction within the construction company.

- **Machine Learning Methods:** The study employs supervised ML models, including
    - Artificial Neural Network (**ANN**)
    - **Gradient Boosting**
    - **XGBoost**.

- **Type of Data:** The data used in this research consist of 10,000 parametric building configurations obtained from end-to-end real-world activities in the partner construction company.
    - The dataset consists of 24 variables divided into three specific groups:
        - Building features
        - Costs
        - Required resources
    - The building features (X) include 11 characteristics such as:
        - The number of floors
        - Building depth
        - Width of windows,
        - Floor height
    - The costs (Z) are represented by two components:
        - Z1 and Z2
            - corresponding to the building and maintenance cost per unit surface, respectively.
    - The required resources (Y) consist of 11 parameters:
        - Internal to the company
            - Representing the amount of materials like cement and sand estimated for construction.
    - **Dataset is available.**

- **Important notes:** The important notes of this study lies in its potential to
  - Streamline operations
  - Reduce engineering and design costs
  - Improve customer relationships
  - Enhance operational efficiency for the construction company

- **Key Findings:**
  - ANN, Gradient Boosting, and XGBoost models demonstrate the highest accuracy, achieving 99% accuracy within less than a second of training time.
  - The optimized building costs obtained through the developed framework are found to be 7% smaller than those of the actual data, indicating significant cost savings potential.
  - R2-squared values measured for some of the regression models
  - In this table z1 and z2 are building and maintenance costs

Regression model	**R-squared**			**Training time**		
	Z1	Z2	Z1 + Z2	Z1	Z2	Z1 + Z2
Multiple linear	0.94	0.94	0.94	0.002	0.003	0.003
Lasso	0.13	0.12	0.13	0.003	0.003	0.004
Ridge	0.12	0.13	0.13	0.003	0.003	0.004
Elastic-Net	0.15	0.13	0.13	0.003	0.003	0.003
XGBoost	0.99	0.99	0.99	16.87	18.21	18.62
ANN	0.99	0.99	0.99	2.79	2.83	2.94
Random forest	0.97	0.92	0.95	44.94	46.3	47.7

8- paper title: Comparison of Artificial Intelligence Techniques for Project Conceptual Cost Prediction: A Case Study and Comparative Analysis

- **Research Goal:** The research aims to address the challenge of conceptual cost estimation in construction projects, particularly focusing on the early stages where information is limited and uncertainty is high. To automate and enhance the accuracy of conceptual cost estimation, the study explores the application of artificial intelligence (AI) techniques, including **fuzzy logic models**, **artificial neural networks**, **multiple regression analysis**, **case-based reasoning**, and ensemble methods like **XGBoost** and **random forest**. By analyzing the performance of these AI models using field canals improvement projects (**FCIPs**) as a case study, the research aims to identify the most accurate and suitable method for conceptual cost modeling, considering metrics such as mean absolute percentage error and adjusted R^2. Additionally, the study addresses key challenges in prediction modeling, such as small data size, missing data values, computational complexity, and model interpretation, while also providing a publicly open dataset for FCIPs to support future validation and analysis efforts.

- **Machine Learning Models Used:** The study investigates 20 AI techniques for conceptual cost modeling including:
    - Fuzzy logic models
    - Artificial neural networks
    - Multiple regression analysis
    - Case-based reasoning
    - Hybrid models (such as genetic fuzzy models)
    - Ensemble methods (such as **XGBoost** and **random forest**).
- **Type of Data:**
    - A publicly open dataset for FCIPs is presented in the study, which can be utilized for future models' validation and analysis.
    - The dataset comprises a total of 144 instances of field canals improvement projects (FCIPs) collected during the years 2010 and 2015.
    - Based on contract information, the cost drivers of FCIPs are represented by a total of 17 parameters including:
        - Area served (P1)
        - Pipeline total length (P2)
        - The number of irrigation valves (P3)
        - Construction year (P4) etc.
    - All features of the dataset

Notation	**variables**
P1	FCIP area served
P2	Average area of area served sections
P3	Pipeline total length
P4	Equivalent diameter of pipeline
P5	Duration of the FCIP
P6	Irrigation valves number
P7	Pressure relief valves number
P8	Sump size
P9	Pump house size

P10	Max discharge capacity
P11	Electrical pump discharge
P12	Diesel pump discharge
P13	Orientation of the improved canal
P14	Year of construction
P15	Rice existence
P16	Intake existence
P17	Parallel canal

- **Important notes:** The research addresses the need for accurate conceptual cost estimation in construction projects, which is essential for:
    - Making informed financial decisions
    - Conducting risk analysis
    - Feasibility studies

- **Key Findings:** The results of the study demonstrate that out of the 20 AI techniques investigated
    - **XGBoost** performs the most accurately and is deemed the most suitable method for conceptual cost modeling.
        - XGBoost achieves a mean absolute percentage error of 9.091% and an adjusted R2 of 0.929.
    - The research also discusses various challenges associated with conceptual cost prediction, including:
        - Small data size
        - Missing data values
        - Computational complexity
        - Model interpretation.
    - Performance of some of the utilized algorithms

Algorithm	Algorithm type	MAPE (%)	R-squared
XGBOOST	Ensemble method	9.091	0.931
Quadratic regression	MRA	9.120	0.857
Bagging	Ensemble method	10.246	0.914

DT	Tree model	12.488	0.886
Genetic fuzzy	Hybrid model	12.9	0.893
Plain MLP	ANNs	9.27	0.913
SVM	Kernel based	21.217	0.136

9- paper title: Estimation and prediction of construction cost index using neural networks, time series, and regression

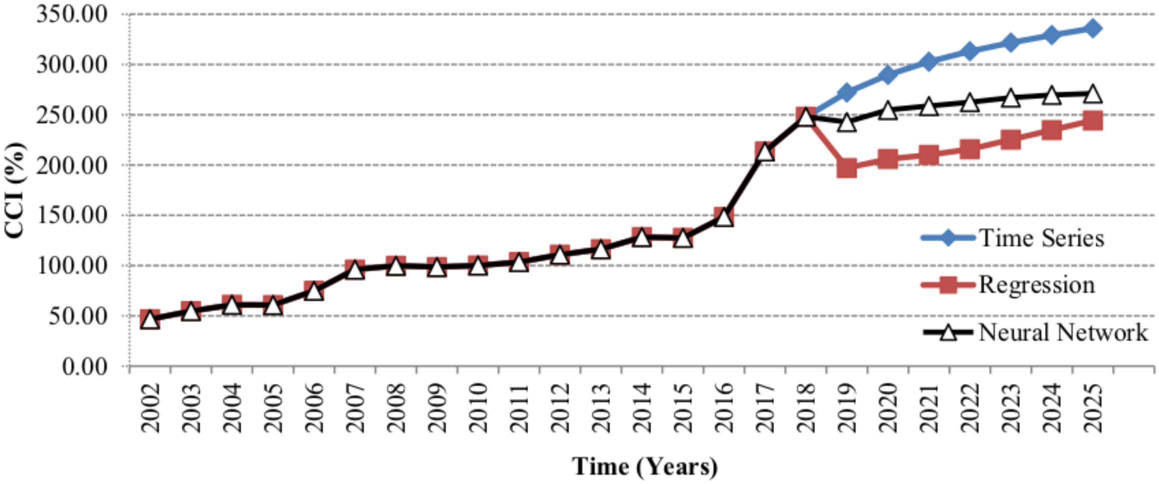
- **Research Goal:** The research goals of this study are multi-faceted. Firstly, the paper aims to address the absence of a reliable **Construction Cost Index** (CCI) estimation tool in Egypt by developing a formula for **calculating the CCI** specifically tailored to concrete structures. Secondly, the study endeavors to provide construction stakeholders in Egypt with a robust forecasting tool for predicting future CCI values, thus aiding in project cost estimation and financial planning. Moreover, the research seeks to enhance the accuracy of CCI predictions by employing various statistical and machine learning techniques, including **Neural Networks**, **Linear Regression**, and **Autoregressive Time Series analysis**. Additionally, the study aims to evaluate the forecasting accuracy of the developed models and provide recommendations based on the findings. Overall, the overarching goal is to empower project owners, estimators, and contractors in Egypt with a reliable tool for anticipating construction costs and navigating price escalation dynamics, particularly in the context of inflation rates and economic fluctuations.

- **Machine Learning Methods:**
  - Neural Networks
  - Linear Regression
  - Autoregressive Time Series
  - These machine learning techniques are utilized to analyze historical data of key cost items and predict the future values of the CCI.

- **Type of Data:** The data used in this research consists of historical records of key construction costs related to reinforced concrete structures in Egypt.
    - These records are utilized to calculate the CCI formula and train the machine learning models for forecasting.
    - The historical prices of key construction materials were collected from the Central Agency for Public Mobilization and Statistics.
    - These prices were gathered over a period of 16 years, from 2002 to 2018, on a yearly basis.
        - Structural steel
        - Portland cement
        - Bricks
        - Sand
        - Gravel
    - For each material, the dataset provides the total yearly prices and the corresponding production quantities, allowing for the calculation of the average unit price.

- **Important notes:** The important notes of this paper lies in its contribution to the construction industry in Egypt by providing stakeholders with a reliable tool for estimating project costs and predicting price escalation.
    - By developing a CCI specific to concrete structures and utilizing machine learning techniques for forecasting, the study addresses the challenge of:
        - fluctuating construction costs, especially in the context of the existing rates of inflation.

- **Key Findings:**
    - The paper proposes a formula for calculating the Construction Cost Index (CCI) based on past records of key construction costs.
        - **CCI = 0.0178 * BUi + 0.0119 * SUi + 0.0595 * CUi + 0.2661 GUi**
            - BUi: Brick unit price for year i.
            - SUi: Steel unit price for year i.
            - CUi: Cement unit price for year i.
            - GUi: Sand and gravel unit price for year i.
    - **Neural Networks, Linear Regression, and Autoregressive Time Series** methods are applied to forecast the CCI.
    - The study provides a reliable forecasting tool for predicting the CCI at any year in the future
        - Addressing the absence of statistical agency estimates or publications of the CCI in Egypt
    - Comparison of the predicted CCI using three different methods:



C) Our plan: Ensemble learning (XGBoost + ANN + KNN)

- Ensemble learning is a powerful technique in machine learning
    - Multiple models are combined to improve predictive performance.
- Ensemble methods leverage the diversity of multiple models to produce more accurate and robust predictions.
- Aggregating the predictions of individual models through techniques such as averaging, boosting, or stacking.
- By incorporating the collective wisdom of diverse models, ensemble learning can mitigate the weaknesses of individual models and capitalize on their strengths
    - resulting in better generalization and predictive performance across various domains, including construction cost prediction.

Base Models:

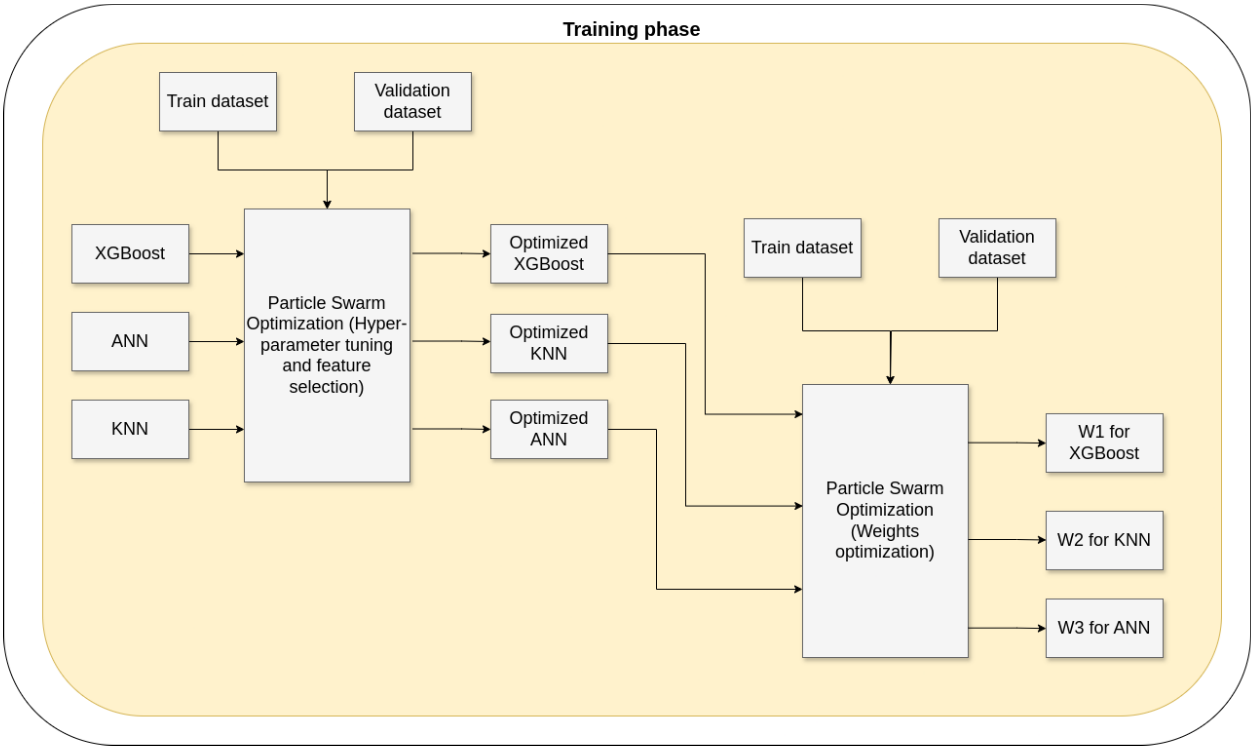
- Combining **XGBoost**, **Artificial Neural Networks (ANN**), and **K-Nearest Neighbors (KNN**) for ensemble learning can be a promising approach.
    - XGBoost is known for its robustness and efficiency in handling structured data
    - ANN excels in capturing complex patterns in data through its layered architecture
    - KNN is effective in making predictions based on similarity measures between data points.
- XGBoost, ANN and KNN have shown a **promising performance** based on the reviewed literature.
- By leveraging the strengths of these diverse algorithms, we can create an ensemble model that is capable of capturing different aspects of the data and producing more accurate predictions.
- Through proper **weighting** and aggregation of the predictions from each model, we can harness the collective intelligence of XGBoost, ANN, and KNN to enhance the accuracy and reliability of our construction cost estimation system.

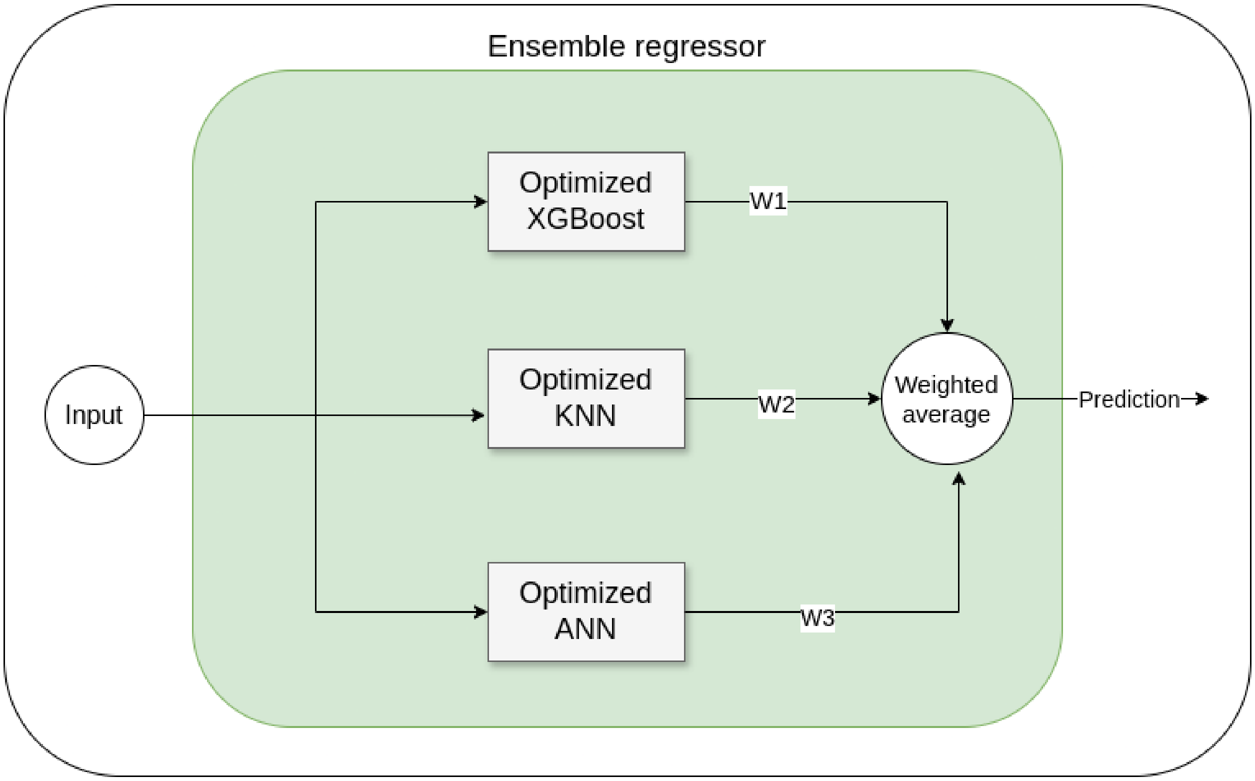
Aggregation scenario: weighted averaging

- we will utilize the **weighted averaging mechanism** to combine the predictions generated by XGBoost, Artificial Neural Networks (ANN), and K-Nearest Neighbors (KNN) models.
- The weighted voting technique assigns different weights to each **model's prediction**.
- Utilizing **Particle Swarm Optimization (PSO)**, we will determine the optimal weights for combining these predictions in the ensemble model.
    - The **PSO** algorithm will search for the **weights** that **minimize** the **error rate** of the weighted averaging mechanism when applied to the validation dataset.
    - By formulating the optimization problem in this manner, we aim to find the weights that result in the most accurate ensemble predictions.
        - **Prediction_function = w1 * KNN (input) + w2 * ANN(input) + w3 * XGBoost(input)/(W1 + w2 + w3)**
        - **Cost function = error_rate(Prediction_function(validation_dataset)**
    - This iterative process allows us to leverage the strengths of each base model and optimize their contributions to the final ensemble output.

Hyperparameter tuning and feature selection

- The performance of the base models depends heavily on their parameter. So, parameter tuning during the training phase is necessary.
- We will use PSO for parameter tuning.
    - ANN: number of hidden layers and number of neurons in each layer.
    - KNN: number of **nearest neighbors (K)**
    - XGBoost: maximum depth of a tree, step size shrinkage used in updates (eta), and learning rate, minimum loss reduction (gamma), minimum sum of instance weight**.**
- Feature selection plays a crucial role in improving the performance of base models by identifying the most relevant features from the dataset.
    - We will do feature selection and parameter tuning for each base models separately


Training phase
Train dataset
Validation dataset
XGBoost
ANN
KNN
Particle Swarm Optimization (Hyper-parameter tuning and feature selection)
Optimized XGBoost
Optimized KNN
Optimized ANN
Train dataset
Validation dataset
Particle Swarm Optimization (Weights optimization)
W1 for XGBoost
W2 for KNN
W3 for ANN


Ensemble regressor
Input
Optimized XGBoost
Optimized KNN
Optimized ANN
W1
W2
W3
Weighted average
Prediction