E-17
Kongres 2026
Congress 2026

doi: 10.67563/ASESCONF2026028K

Ivana Karadžić[1]

MACHINE LEARNING-BASED PREDICTION OF CONCRETE COMPRESSIVE STRENGTH FOR STRUCTURAL ENGINEERING APPLICATIONS

Summary: The increasing availability of experimental data in civil engineering has created new opportunities for the application of artificial intelligence techniques in structural engineering research. Machine learning approaches provide powerful tools for identifying complex relationships between material composition and mechanical properties. In this study, data-driven methods are applied to investigate the prediction of concrete compressive strength using a publicly available experimental dataset. The considered parameters include the main components of concrete mixtures as well as the age of concrete specimens. The objective is to explore the capability of machine learning models to capture nonlinear relationships between mixture composition and the resulting compressive strength. The presented framework represents a preliminary step toward broader applications of artificial intelligence in structural engineering problems, including reinforced concrete elements such as beam–column joints.

Keywords: machine learning, concrete compressive strength prediction, artificial intelligence in structural engineering, data-driven modelling, reinforced concrete structures

PREDVIĐANJE ČVRSTOĆE BETONA PRI PRITISKU PRIMENOM MAŠINSKOG UČENJA

Rezime: Sve veća dostupnost eksperimentalnih podataka u građevinarstvu stvorila je nove mogućnosti za primjenu tehnika vještačke inteligencije u istraživanjima iz oblasti konstrukcija. Pristupi zasnovani na mašinskom učenju predstavljaju moćan alat za identifikaciju složenih veza između sastava materijala i njihovih mehaničkih svojstava. U ovom radu primijenjene su metode zasnovane na analizi podataka za predviđanje čvrstoće betona pri pritisku korišćenjem javno dostupnog eksperimentalnog skupa podataka. Razmatrani parametri obuhvataju osnovne komponente betonske mješavine, kao i starost uzoraka betona. Cilj istraživanja jeste ispitivanje sposobnosti modela mašinskog učenja da opišu nelinearne odnose između sastava betonske mješavine i ostvarene čvrstoće pri pritisku. Predloženi pristup predstavlja početni korak ka široj primjeni vještačke inteligencije u rešavanju problema iz oblasti konstruktivnog inženjerstva, uključujući armiranobetonske elemente, kao što su grede, stubovi i spojevi greda i stubova.

Ključne reči: mašinsko učenje, čvrstoća betona pri pritisku, vještačka inteligencija, predviđanje čvrstoće, armiranobetonske konstrukcije.

[1] University of Montenegro, ivana.k@ucg.ac.me

1. INTRODUCTION

Machine learning methods have become increasingly important in civil engineering for solving complex problems involving nonlinear relationships between material properties and structural performance [1,2]. In recent years, artificial intelligence techniques have shown significant potential in predicting the mechanical properties of concrete, optimizing mix design, and reducing the need for extensive experimental testing [1,2]. Among these approaches, machine learning models have attracted considerable attention due to their ability to identify complex patterns within experimental datasets and provide accurate predictions based on multiple input parameters [3].

Concrete compressive strength represents one of the most important mechanical properties of concrete and plays a key role in the design and assessment of reinforced concrete structures. However, predicting compressive strength remains a challenging task due to the nonlinear interaction between concrete constituents, curing conditions, and concrete age. Traditional empirical and statistical approaches often show limited capability in capturing such nonlinear relationships, which has motivated the application of machine learning techniques in this field.

Several researchers have investigated the use of machine learning algorithms for predicting concrete compressive strength. Yeh [4] applied artificial neural networks for modeling high-performance concrete and demonstrated the capability of machine learning methods to successfully predict compressive strength based on mix composition. More recently, various studies confirmed the effectiveness of ensemble learning methods, particularly Random Forest algorithms, in handling nonlinear material behavior and improving prediction accuracy [2,3].

In this paper, machine learning techniques were applied for predicting concrete compressive strength using a publicly available experimental dataset containing information about concrete mixture proportions and curing age. Two machine learning approaches, Linear Regression and Random Forest Regression, were implemented and compared in terms of prediction accuracy and model performance. The obtained results demonstrate the applicability of machine learning methods in concrete engineering and highlight their potential for future applications in structural engineering problems involving reinforced concrete elements and beam-column connections.

2. DATASET AND METHODOLOGY

2.1. DATASET DESCRIPTION

A publicly available experimental dataset for concrete compressive strength prediction was used in this paper. The dataset contains quantitative information on the proportions of the principal concrete constituents expressed in kg/m^3, including cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate, together with the curing age of each specimen. These variables represent the amounts of the constituent materials used in each concrete mixture and were selected as the input variables for the machine learning models. These variables are known to significantly influence the hydration process, microstructure development, and mechanical properties of concrete. It should be noted that the original dataset does not provide additional material specifications, such as the cement classification, aggregate grading or mineralogical properties, or the chemical type of the superplasticizer. Therefore, the present study is limited to the numerical mixture proportions available in the original experimental database developed by Yeh [4].

Table 1 summarizes the input variables considered in the machine learning analysis.

Table 1: Input variables

Variable	Units
Cement	kg/m³
Blast furnace slag	Slag content (kg/m³)
Fly ash	Fly ash content (kg/m³)
Water	Water content (kg/m³)
Superplasticizer	Superplasticizer content (kg/m³)
Coarse aggregate	Coarse aggregate content (kg/m³)
Fine aggregate	Fine aggregate content (kg/m³)
Age	Concrete age (days)

Table 2 summarizes the statistical characteristics of the input variables. The dataset covers a wide range of mixture proportions and curing ages, indicating considerable variability in the experimental concrete mixtures. Such variability makes the dataset suitable for developing and evaluating machine learning regression models over a broad range of concrete compositions and compressive strength values.

Table 2: The statistical characteristics of the input variables

Variable	Unit	Mean	Std.dev.	Minimum	Maximum
Cement	kg/m³	281.17	104.51	102	540
Blast furnace slag	kg/m³	73.9	86.28	0	359.4
Fly ash	kg/m³	54.19	64	0	200.1
Water	kg/m³	181.57	21.36	121.75	247
Superplasticizer	kg/m³	6.2	5.97	0	32.2
Coarse aggregate	kg/m³	972.92	77.75	801	1145
Fine aggregate	kg/m³	773.58	80.18	594	992.6
Age	days	45.66	63.17	1	365

2.2. DATA ANALYSIS AND PREPROCESSING

Prior to the implementation of machine learning models, the dataset was analyzed to examine the distribution, variability, and general characteristics of the experimental data. A descriptive statistical analysis was performed for all input and output variables, including mean values, standard deviations, minimum and maximum values, and quartile ranges. This preliminary analysis provided a better understanding of the dataset structure and the variability of the concrete mixture parameters and compressive strength values.

In addition to the statistical overview, graphical visualization of the compressive strength distribution was performed using a histogram. Figure 1 presents the distribution of compressive strength values contained in the dataset. The histogram indicates that the dataset covers a relatively wide range of concrete compressive strengths, which is beneficial for machine learning applications because it enables the models to learn relationships across different strength levels.

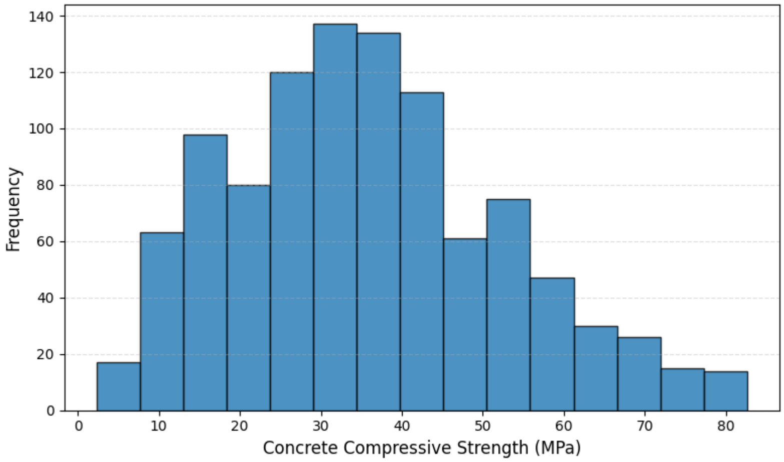

Figure 1: Distribution of concrete compressive strength values in the experimental dataset

After the initial data analysis, the dataset was divided into input variables and the target variable. The input variables included the parameters describing the concrete mixture composition and concrete age, while the target variable represented the compressive strength of concrete expressed in MPa.

The machine learning analysis was performed using the Python programming language within the Google Colab environment. Data processing, visualization, and model development were conducted using the Pandas, Matplotlib, and Scikit-learn libraries [5].

Subsequently, the dataset was randomly divided into training and testing subsets using the train_test_split function available within the Scikit-learn library [5]. In the present study, 80% of the dataset was used for model training, while the remaining 20% was reserved for testing and validation of the developed models. The training subset was used to learn the relationships between the input variables and concrete compressive strength, whereas the testing subset was used to evaluate the predictive performance of the developed models on previously unseen data.

The separation of data into training and testing subsets represents a standard procedure in machine learning applications and helps reduce the risk of overfitting, improving the reliability and generalization capability of the developed models [2,5].

2.3. MACHINE LEARNING MODELS

In this study, two machine learning approaches were implemented for predicting concrete compressive strength: Linear Regression and Random Forest Regression. The models were selected to compare the performance of a conventional linear method with a nonlinear ensemble learning approach. The overall workflow adopted in the study is presented in Figure 2.

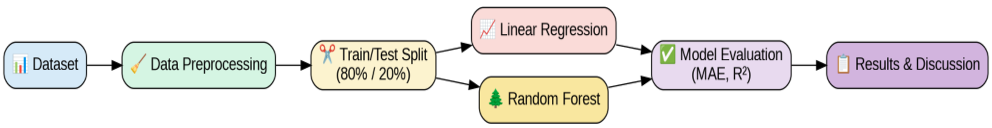

Figure 2: Workflow of the implemented machine learning methodology for concrete compressive strength prediction

Linear Regression represents one of the most widely used regression techniques in engineering and data analysis applications [6]. The method assumes a linear relationship between the input variables and the target variable and estimates the relationship between concrete mixture parameters and compressive strength by fitting a linear function to the experimental data. Due to its simplicity and interpretability, Linear Regression is commonly used as a baseline model for evaluating the performance of more advanced machine learning algorithms. However, its predictive capability may become limited when dealing with highly nonlinear relationships between variables, which frequently occur in concrete material behavior.

In addition to Linear Regression, the Random Forest Regression algorithm was implemented in this study. Random Forest represents an ensemble learning method based on the construction of multiple decision trees and aggregation of their predictions in order to improve prediction accuracy and model robustness [7]. Unlike linear models, Random Forest algorithms are capable of capturing complex nonlinear relationships between variables and handling interactions among multiple parameters. This characteristic makes them particularly suitable for engineering problems involving heterogeneous experimental data and nonlinear material behavior.

The Random Forest model was implemented using the *RandomForestRegressor* algorithm available in the Scikit-learn library. A fixed random seed (*random_state = 42*) was specified to ensure the reproducibility of the results, while all other hyperparameters were retained at their default values. Consequently, the model was trained using 100 decision trees *(n_estimators = 100*), with unrestricted tree depth (*max_depth = None*), *min_samples_split = 2,* and *min_samples_leaf = 1*. No additional hyperparameter optimization or tuning was performed.

Both models were trained using the training subset of the dataset, while their predictive performance was evaluated using the testing subset based on the Mean Absolute Error (MAE) and the coefficient of determination (R^2) evaluation metrics presented in the following section.

2.4. MODEL EVALUATION

The performance of the implemented machine learning models was evaluated using the MAE and R^2. These metrics are widely used in regression-based machine learning applications and provide a reliable assessment of prediction capability and model accuracy [5].

The Mean Absolute Error is defined as (1):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

where *n* is the number of samples, y_i represents the experimentally measured compressive strength, and $\hat{y}_i$ denotes the value predicted by the model. MAE quantifies the average prediction error, with lower values indicating better predictive performance.

The coefficient of determination (R^2) is expressed as (2):

$$R = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad (2)$$

where $\bar{y}_i$ is the mean value of the experimental observations. The R^2 metric represents the proportion of variance in the experimental data explained by the machine learning model. R^2 values closer to 1 indicate a better agreement between the predicted and experimentally measured values and therefore higher predictive capability of the model. The combined use of MAE and R^2 provides a comprehensive evaluation of model performance by simultaneously quantifying prediction error and assessing the ability of the model to capture variability within the experimental dataset.

3. RESULTS AND DISCUSSION

3.1. LINEAR REGRESSION RESULTS

The Linear Regression model was first implemented as a baseline prediction model for estimating concrete compressive strength. After training and testing the model using the prepared dataset, the obtained coefficient of determination was approximately 0.63 while the Mean Absolute Error reached approximately 7.75 MPa.

The obtained results indicate that the Linear Regression model was able to identify the general trend between the input parameters and compressive strength values, as presented in Table 3.

Table 3: Performance metrics of the Linear Regression model

Model	$MAE\ (MPa)$	R^2
Linear Regression	7.75	0.63

As shown in Figure 3, the predicted values generally follow the overall trend of the experimental results. However, a noticeable scatter around the ideal prediction line indicates that the Linear Regression model is not able to fully capture the nonlinear relationships between the input variables and concrete compressive strength.

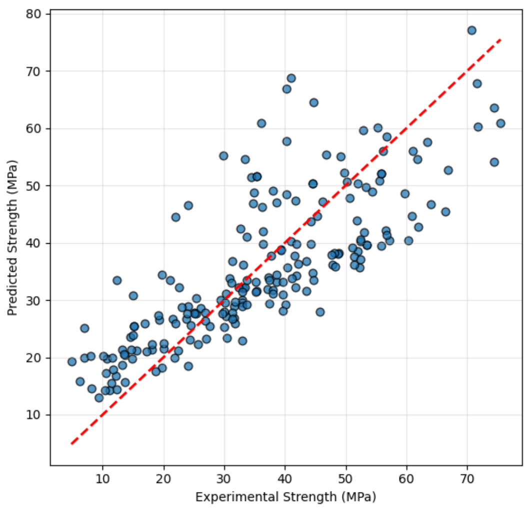

Figure 3: Comparison between experimental and predicted compressive strength values obtained using the Linear Regression model

Although the model captures the overall trend of the experimental data, the dispersion of points around the ideal prediction line increases at higher compressive strength values, indicating reduced prediction accuracy for stronger concrete mixtures.

However, the relatively moderate R^2 value and higher prediction error suggest that the linear model could not fully capture the complex nonlinear relationships between concrete mixture components and compressive strength. The limitations of Linear Regression are expected in this type of engineering problem because concrete behavior depends on multiple interacting parameters whose relationships are predominantly nonlinear. Consequently, the predictive capability of simple linear models becomes limited when applied to heterogeneous experimental datasets involving different mixture compositions and curing conditions.

These findings indicate that Linear Regression can serve as a useful baseline prediction model; however, its ability to represent the nonlinear behavior of concrete remains limited.

3.2. RANDOM FOREST RESULTS

In addition to Linear Regression, the Random Forest Regression model was implemented in order to improve prediction accuracy and better capture nonlinear relationships between the input parameters and concrete compressive strength. The performance of the Random Forest model is summarized in Table 4.

Table 4: Performance of the Random Forest Model

Model	$MAE\ (MPa)$	R^2
Random Forest	3.75	0.88

The Random Forest model achieved significantly better predictive performance than the Linear Regression model. The obtained coefficient of determination was $R^2 = 0.88$, while the Mean Absolute Error (MAE) was reduced to approximately 3.75 MPa. As illustrated in Figure 4, the predicted values are closely distributed around the ideal prediction line, indicating a strong agreement between the experimental and predicted compressive strength values.

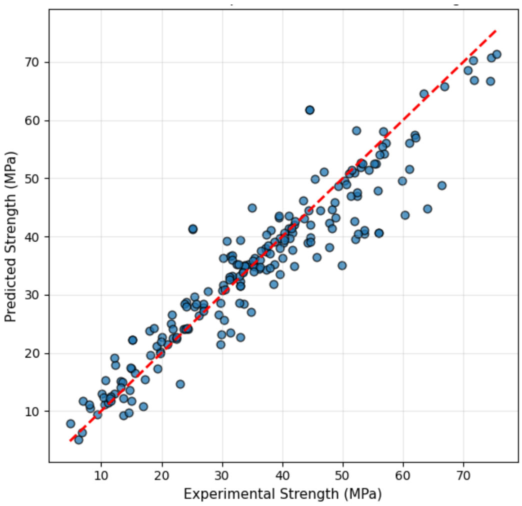

Figure 4: Comparison between experimental and predicted compressive strength values obtained using the Random Forest model

Compared with Figure 3, a noticeably smaller dispersion of the data points around the ideal prediction line can be observed.

The improved performance of the Random Forest model can be attributed to its ability to capture complex nonlinear relationships and interactions among multiple input variables through an ensemble of decision trees. Unlike Linear Regression, which assumes a linear relationship between the input parameters and the target variable, Random Forest models the nonlinear behavior of concrete mixtures, resulting in substantially improved prediction accuracy.

Overall, the obtained results confirm the strong potential of ensemble machine learning methods for predicting the mechanical properties of concrete. The high coefficient of determination and low prediction error demonstrate that the Random Forest algorithm is a reliable and effective tool for engineering applications involving complex material behavior and multiple interacting parameters.

3.3. MODEL COMPARISON

The predictive performance of the Linear Regression and Random Forest models was compared using both quantitative evaluation metrics and graphical analysis. Table 5 summarizes the obtained Mean Absolute Error (MAE) and coefficient of determination (R^2) values for both models, while Figures 3 and 4 compare the relationship between experimentally measured and predicted concrete compressive strength values.

Table 5: Performance metrics of the implemented machine learning models

Model	MAE (MPa)	R^2
Linear Regression	7.75	0.63
Random Forest	3.75	0.88

The Linear Regression model achieved an MAE of approximately 7.75 MPa and an R^2 value of 0.63, indicating a moderate prediction capability. In contrast, the Random Forest model substantially improved the prediction accuracy, reducing the MAE from 7.75 MPa to 3.75 MPa while increasing the coefficient of determination from 0.63 to 0.88. The graphical comparison presented in Figure 5 further supports these findings. This indicates a substantially better agreement between the predicted and experimentally measured compressive strength values. The improved performance of the Random Forest algorithm can be attributed to its ability to capture nonlinear relationships and complex interactions among multiple concrete mixture variables.

Overall, the comparison demonstrates that the Random Forest model provides a more accurate and reliable prediction of concrete compressive strength than the Linear Regression model. These findings highlight the advantages of ensemble learning techniques for engineering applications involving heterogeneous experimental data and nonlinear material behavior. Overall, the Random Forest model reduced the prediction error by approximately 52% while increasing the coefficient of determination by nearly 40% compared with Linear Regression. These findings indicate that ensemble learning techniques are well suited for predicting concrete compressive strength. Furthermore, they provide a reliable basis for future machine learning applications in structural engineering.

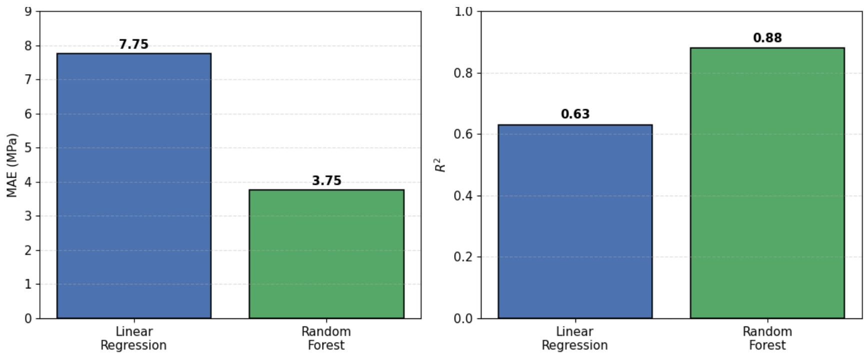

Figure 5: Comparison of the predictive performance of the Linear Regression and Random Forest models based on MAE and R^2

The obtained results are generally consistent with findings reported in previous studies on machine learning-based prediction of concrete compressive strength. Recent investigations have consistently shown that ensemble learning methods outperform conventional linear regression models due to their ability to capture complex nonlinear relationships among mixture constituents. For example, Elshaarawy et al. [1], Zhang et al. [8], Omotayo et al. [3], and Kalabarige et al. [2] reported that nonlinear and ensemble machine learning algorithms provide high prediction accuracy for concrete compressive strength prediction. In the present study, the Random Forest model achieved an R^2 value of 0.88 and an MAE of 3.75 MPa, confirming the effectiveness of ensemble learning for concrete compressive strength prediction. Although some recent studies have reported higher prediction accuracy, they typically employed additional feature engineering, hyperparameter optimization, or more advanced ensemble learning algorithms. Therefore, the results obtained in this study are in line with the current state of research, given the relatively simple modelling framework adopted, without additional feature engineering or hyperparameter optimization.

3.4. FEATURE IMPORTANCE ANALYSIS

In addition to evaluating the predictive performance of the implemented machine learning models, a feature importance analysis was performed using the Random Forest algorithm to quantify the relative contribution of each input variable to the prediction of concrete compressive strength. Figure 6 presents the obtained feature importance values.

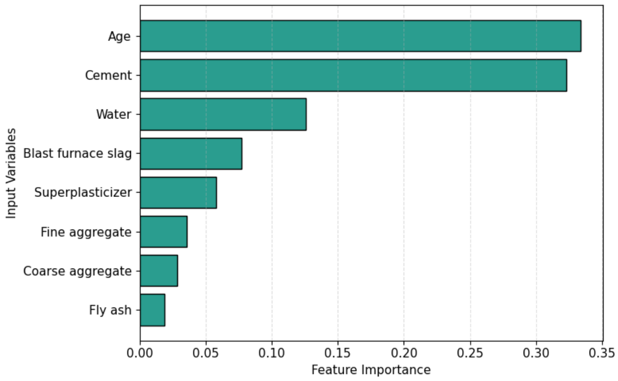

Figure 6: Feature importance obtained from the Random Forest model for predicting concrete compressive strength

The results indicate that concrete age, cement content, and water content are the most influential variables affecting concrete compressive strength prediction. These findings are consistent with the physical principles governing concrete behavior, since cement directly contributes to strength development, water influences the water-to-cement ratio, and concrete age controls the progress of hydration reactions [8].

The remaining variables, including blast furnace slag, fly ash, superplasticizer, and aggregate contents, also contribute to the prediction, although to a lesser extent. Overall, the feature importance analysis demonstrates that the Random Forest model not only provides accurate predictions but also offers valuable insight into the relative influence of individual mixture parameters, thereby enhancing the interpretability of the developed machine learning model [8].

From a practical engineering perspective, the developed Random Forest model can be used to support concrete mix design by providing rapid predictions of compressive strength for different combinations of mixture proportions and curing ages. Such predictions may assist engineers in evaluating alternative mix compositions before laboratory testing, thereby reducing the number of trial batches required during the mix design process. Although experimental validation remains essential for final verification, the proposed model can serve as an efficient decision-support tool during the preliminary stages of concrete mix design, contributing to reduced development time and material consumption.

4. CONCLUSION

In this study, machine learning techniques were applied to predict the compressive strength of concrete using a publicly available experimental dataset containing information on concrete mixture composition and curing age. Two machine learning approaches, Linear Regression and Random Forest Regression, were implemented and evaluated in terms of prediction accuracy and model performance.

The obtained results demonstrated that the Random Forest model outperformed the Linear Regression model, achieving an R^2 value of 0.88 and a Mean Absolute Error of approximately 3.75 MPa, compared with $R^2 = 0.63$ and MAE = 7.75 MPa for the Linear Regression model. These findings confirm the superior capability of the Random Forest algorithm to capture the nonlinear relationships between the input variables and concrete compressive strength.

In addition to its high prediction accuracy, the Random Forest model enabled the interpretation of the relative importance of the input variables. The feature importance analysis identified concrete age, cement content, and water content as the most influential variables affecting concrete compressive strength prediction, which is consistent with the established understanding of concrete behavior.

Overall, the findings demonstrate the strong potential of machine learning techniques for predicting concrete mechanical properties and supporting data-driven engineering decision-making in civil engineering. Future research will focus on extending the proposed methodology to more complex reinforced concrete structural elements and connections, including reinforced concrete beams, columns, and beam–column joints subjected to different loading conditions.

REFERENCES

[1] Elshaarawy, M. K., et al. (2024). Machine learning and interactive GUI for concrete compressive strength prediction using a comprehensive experimental database. Scientific Reports, 14: 16694.

[2] Kalabarige, L. R., Sridhar, J., Subbaram, S., Prasath, P., & Gobinath, R. (2024). Machine Learning Modeling Integrating Experimental Analysis for Predicting Compressive Strength of Concrete Containing Different Industrial Byproducts. Advances in Civil Engineering, 2024, Article ID 7844854.

[3] Omotayo, O. O., Arum, C., & Ikumapayi, C. M. (2024). Assessment of Machine Learning Methods for Concrete Compressive Strength Prediction. Journal of Soft Computing in Civil Engineering, 8(4), 116–140

[4] Yeh, I. C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. Cement and Concrete Research, 28(12), 1797–1808

[5] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[6] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to Linear Regression Analysis (6th ed.). Wiley.

[7] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[8] Zhang, Y., Ren, W., Chen, Y., Mi, Y., Lei, J., & Sun, L. (2024). Predicting the compressive strength of high-performance concrete using an interpretable machine learning model. Scientific Reports, 14, 28346.