International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 12 Issue 4 ǁ April 2024 ǁ PP. 156-162

Prediction of Compressive Strength of Ultra-high Performance Concrete Based on Machine Learning

YongChao Shi[1*], WenZhao Rong[1], JiaKai Wu[1]
*[*1]School of Environment and Architecture, University of Shanghai for Science and Technology, Shanghai, China*
Corresponding Author: YongChao Shi

Abstract
The compressive strength of ultra-high performance concrete (UHPC) is influenced by many factors. Existing UHPC compressive strength calculation models are based on own experimental results and obtained through regression fitting of experimental data. This method has certain limitations in terms of time and cost, while machine learning can establish nonlinear mappings under multiple factors, discovering and revealing the complex coupling mechanism between the components of UHPC materials and their compressive strength, Therefore, machine learning methods can accurately predict the compressive strength of UHPC. Using 15 influencing factors as input variables and compressive strength as output variables, three machine learning models, Decision Tree (DT), Random Forest (RF), and eXtreme Gradient Boost tree (XGBoost), were established to predict the compressive strength of UHPC. The prediction accuracy of each model was evaluated using three indicators: coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MSE), Finally, the SHapley Additive exPlans (SHAP) method was used to explain the model. The research results indicate that the XGBoost model has the highest prediction accuracy, with an RMSE of 5.967, MAE of 3.85, and R2 of 0.936. The SHAP method interpretation results show that among the 15 selected influencing factors, fiber content has the greatest impact on the compressive strength of UHPC. The above work verifies the accuracy .

Keywords: *machine learning, ultra-high performance concrete, compressive strength, steel fiber, interpretation.*

Date of Submission: 02-04-2024 Date of acceptance: 12-04-2024

I. INTRODUCTION

Ultra-high performance concrete (UHPC) is a novel cementitious composite material characterized by exceptionally high compressive strength, high ductility, superior durability, and low porosity [1]. Typically, UHPC does not contain coarse aggregates and requires high-density packing to achieve the desired particle packing density. Consequently, the production process necessitates a significant amount of cement and silica fume, along with expensive additives such as high-performance water reducers, quartz powder, and special fibers . The coupling mechanisms between different raw materials are intricate, thus requiring considerable time and cost for the mix design and experimental analysis of UHPC containing various mineral admixtures [2].

Machine learning possesses the ability to capture nonlinear and fuzzy relationships between input and output variables in a dataset, making it widely applicable in the field of civil engineering. Currently, machine learning finds applications in mechanical performance prediction, structural defect detection, and optimization of construction materials [3].Yuan et al. [4] conducted research using machine learning methods to predict the compressive and flexural strength of Recycled Aggregate Concrete (RAC). Twelve input factors were considered, and their influence on the strength of RAC was analyzed. The results showed that the Random Forest model outperformed the Gradient Boosting model in predicting RAC strength, demonstrating higher accuracy. Kadir et al. [5] utilized features such as moisture content and relative humidity to predict the compressive strength of concrete using artificial neural networks, decision trees, support vector machines, and other models. Their results showed that the decision tree achieved the best R^2 value of 0.86, indicating its highest accuracy. Roya et al. [6] used a machine learning-based support vector machine (SVM) model to predict the flexural performance of Ultra-High Performance Concrete (UHPC) beams. Multiple input variables including geometric shape and material properties of the beams were used to predict their flexural performance. The SVM model accurately predicted the flexural capacity of UHPC beams with different material and structural characteristics.

This article employs three models from machine learning, namely Decision Tree (DT), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost), to predict the compressive strength of Ultra-High Performance Concrete (UHPC) under the influence of multiple variables. To establish a comprehensive and reliable dataset, 267 sets of data were collected from experimental data published in domestic and international literature, forming a database. The experimental data include various supplementary cementitious materials, fine aggregates, ultrafine aggregates, types of fibers, and age periods. During the model training process, arbitrary combinations of mineral admixtures were considered: fly ash, silica fume, slag, glass powder, lime powder, quartz sand, and quartz powder; arbitrary combinations of fiber specifications were also taken into account, including straight fibers, hooked-end fibers, diameters ranging from 0 mm to 0.5 mm, and lengths ranging from 0 mm to 30 mm; arbitrary combinations of age periods were considered as well, ranging from 3 days to 91 days. The predicted values from the machine learning models were compared with the experimental values to verify whether machine learning methods can effectively be applied to predict the compressive strength of UHPC. To address the "black box" issue of machine learning models and enhance their interpretability, the SHapley Additive exPlanations (SHAP) algorithm was utilized to increase the explainability of the machine learning "black box" model.

II. UHPC DATA

2.1 Data Preprocessing

The dataset consists of 267 sets of data, which were used to train Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting Tree (XGB) models. The input variables include fiber type (FP), fiber diameter (FD), fiber length (FL), fiber content (FC), fly ash (FA), silica fume (SF), ground blast furnace slag (GBSF), glass powder (GP), lime powder (LP), quartz sand (QS), quartz powder (QP), sand-to-binder ratio (SB), water-to-binder ratio (WB), water reducer (SP), and age (D), with compressive strength (CS) as the output variable. The steel fiber content in the dataset is expressed as a volume ratio, while the quantities of other components are related to the mass ratio of cement. As the dataset contains data from specimens of different sizes, the compressive strength needs to be normalized to a cubic specimen size with a side length of 100 mm.The data was split into two parts using the train_test_split() function from the sklearn library, with 80% for training and 20% for testing. The testing set was used to evaluate the performance of the models on unseen data, verifying the accuracy and generalization ability of the models. During the data preprocessing, a multicollinearity test was conducted among the feature variables to verify the reliability of the sample data. A heatmap of the correlation coefficients between the input variables and the output variable was plotted, as shown in Figure 1.From the figure, it can be observed that the correlation coefficient between the steel fiber content and compressive strength is the highest, reaching 0.6. Additionally, there is a strong correlation between the diameter and length of steel fibers. The correlation between various supplementary cementitious materials and compressive strength does not exceed 0.22, indicating the absence of multicollinearity among them. Therefore, when predicting the compressive strength of UHPC, all input variables should be selected to improve the accuracy of the machine learning model.

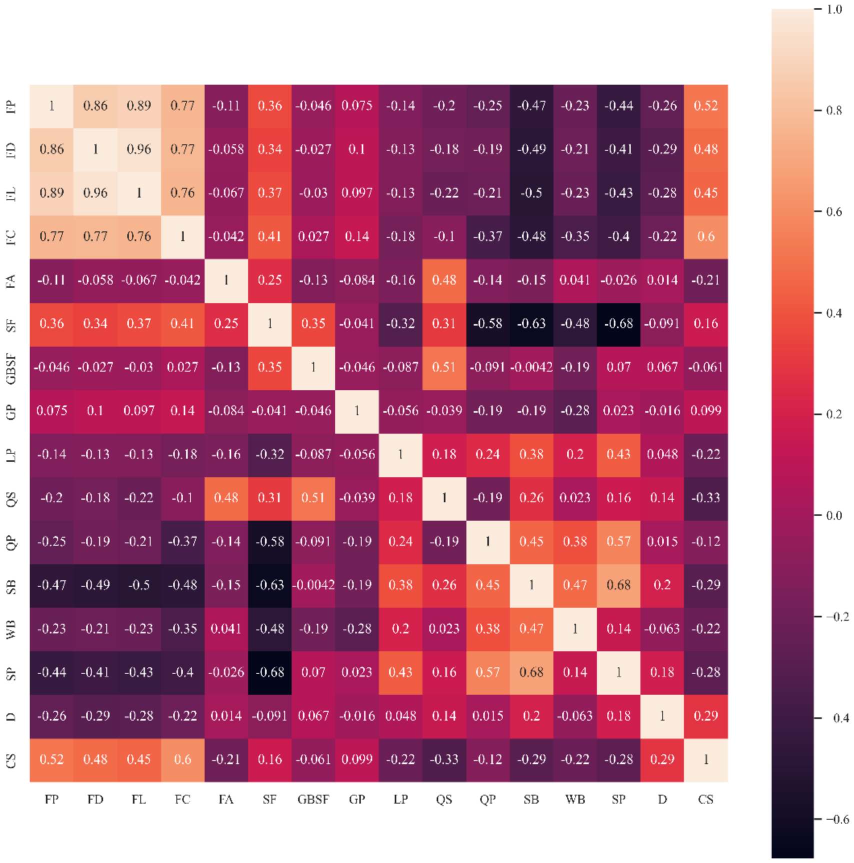


Figure1: Heatmap of correlation coefficients between input variables and output variable

III. MACHINE LEARNING ALGORITHMS AND EVALUATION

3.1 Decision Tree

Decision Tree (DT) [7] is a commonly used data mining technique, which can be employed for classification and regression analysis. The decision tree algorithm is a non-parametric model that segments data using a series of attributes and generates a tree to describe the classification or regression relationships within the data. In this paper, the tree is constructed by recursively partitioning the dataset, selecting the optimal attribute as a node at each split, and dividing the dataset into multiple subsets based on the values of this attribute.

3.2 Random Forest

Random Forest (RF) [8] is an algorithm that has been widely applied in both regression and classification problems. It utilizes the bagging technique to randomly sample from the original dataset and construct multiple subsets. Each subset is then used to train independent predictive models. When making predictions on samples in this paper, Random Forest averages the predictions from all decision trees to obtain the final output. Through this aggregation method, Random Forest can effectively reduce prediction errors and improve prediction accuracy.

3.3 Extreme Gradient Boosting Tree

Extreme Gradient Boosting Tree (XGBoost) [9] is a machine learning algorithm based on classification and regression. In XGBoost, each learner is a decision tree model that generates multiple weak classifiers by iteratively optimizing the loss function. Each new learner corrects the errors made by previous learners, resulting in a stronger ensemble model. This iterative process of generating learners can improve the overall predictive ability of the model and exhibit high efficiency and accuracy when handling large-scale datasets.

3.4 Model Evaluation Indicators

In order to comprehensively evaluate the accuracy of each model's predictions, this paper adopts three metrics: the coefficient of determination (R^2), the mean absolute error (MAE), and the root mean square error (RMSE) to assess the fitting of different models. R^2 measures the extent to which the model explains the variability of the dependent variable, ranging from 0 to 1. A higher R^2 value indicates better fitting of the data, while an R^2 close to 0 suggests poorer fitting. MAE represents the average absolute error between predicted and actual values, quantifying the average deviation of the model's predictions from the actual values. A smaller MAE indicates less prediction error and better model performance. MAE is robust to outliers, providing a more stable evaluation of model performance. RMSE calculates the square root of the average squared differences between predicted and actual values, serving as a standard deviation of the model's prediction errors. Similar to MAE, a smaller RMSE signifies smaller prediction biases and better model performance. RMSE penalizes larger error values (outliers), thus emphasizing the accuracy of important prediction results.

3.5 Model Interpretation Methods

This paper utilizes the global feature maps generated by Shapley Additive Explanations (SHAP) to interpret machine learning models [10]. SHAP is a method employed for elucidating predictions made by machine learning models, providing an assessment of the contribution of each feature to the predicted outcome. Grounded in the concept of Shapley values from cooperative game theory, SHAP furnishes an interpretable framework by computing the average marginal contribution of each feature within feature combinations. The global feature map integrates the effects of features and their importance by calculating the contribution of each feature to the predicted outcome. This allows for quantifying the relative importance of each feature and plotting a decreasingly sorted feature importance graph based on importance . In the global feature map, each point represents a feature and the SHAP value of a sample. The horizontal axis denotes the SHAP contribution of the feature in the sample, while the vertical axis illustrates the relative importance of this feature. When multiple data points share the same position on the horizontal axis, jitter stacking is applied to avoid overlap. A higher height on the vertical axis indicates a greater impact of the feature.

IV. RESULTS AND DISCUSSION

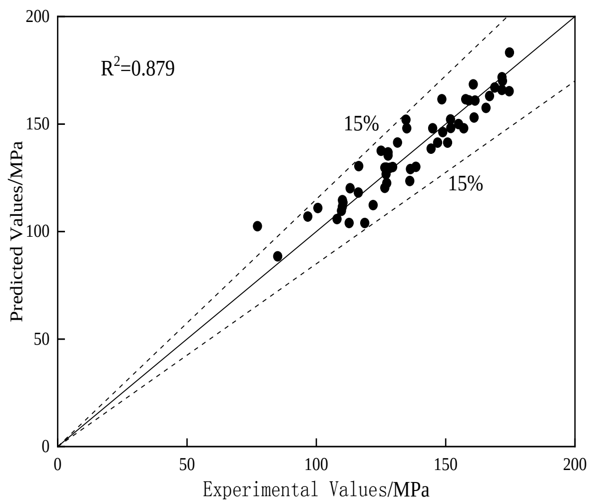
4.1 Machine Learning Model Prediction Results

The predicted results of UHPC compressive strength by machine learning models are shown in Figure 2, and it can be observed that:

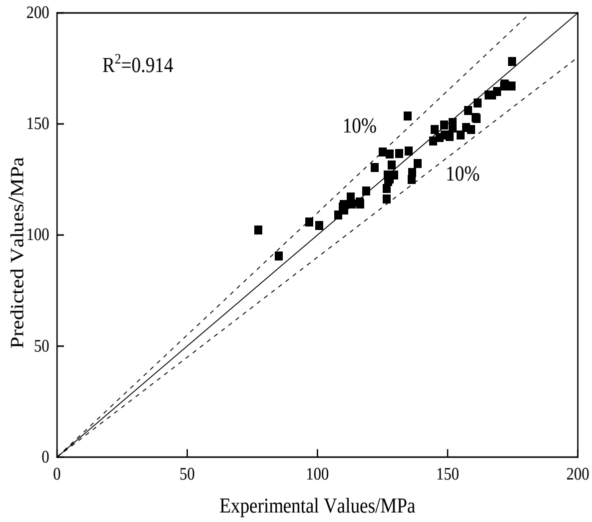
(1). From the comparison between predicted values and experimental values in the model prediction versus experimental value plot, it is evident that the predicted values of the DT, RF, and XGBoost models can closely approximate the experimental values. This indicates that after learning from the UHPC compressive strength training data, all three machine learning models possess good generalization ability and can be used to predict the compressive strength of UHPC under the influence of multiple factors.

(2). From the comparison plot and scatter fit plot, it can be noted that there are some cases of significant discrete errors in the predictions of the DT model. This is related to the sensitivity of the DT model to noise and outliers during the training process. If noise or outliers exist at critical feature split points, the decision tree may incorrectly treat them as new branches or leaf nodes, thereby distorting the overall structure of the model.

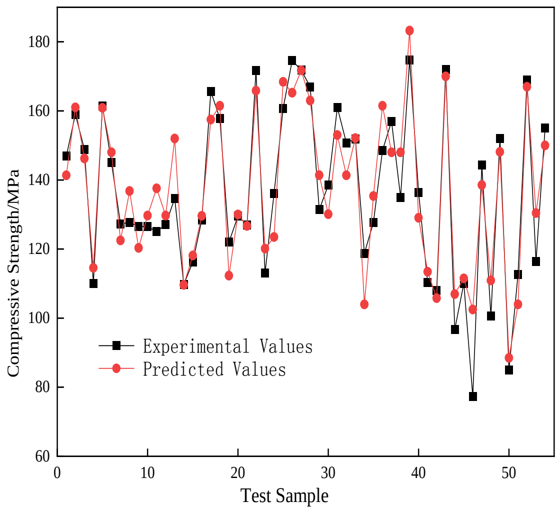
(3). Compared to the DT model, the results of the RF and XGBoost models are superior. The prediction errors of the DT model are mainly concentrated within ±10%, while those of the RF and XGBoost models are also primarily within ±10%. This suggests that RF and XGBoost models can reduce the impact of noise and outliers on the model by randomly sampling feature data.



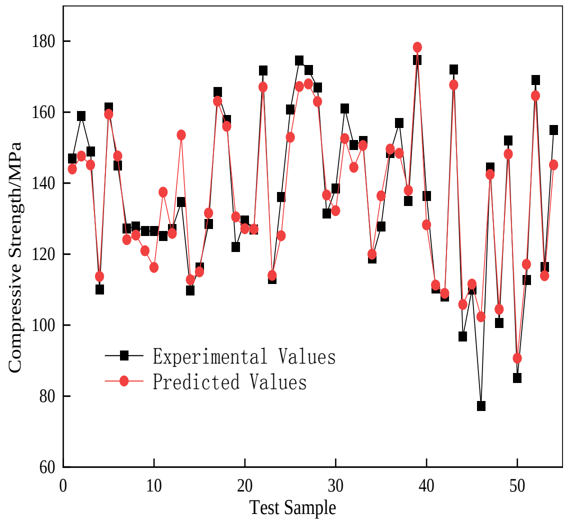
(a) The scatter fit plot of predicted values versus experimental values for the DT model.



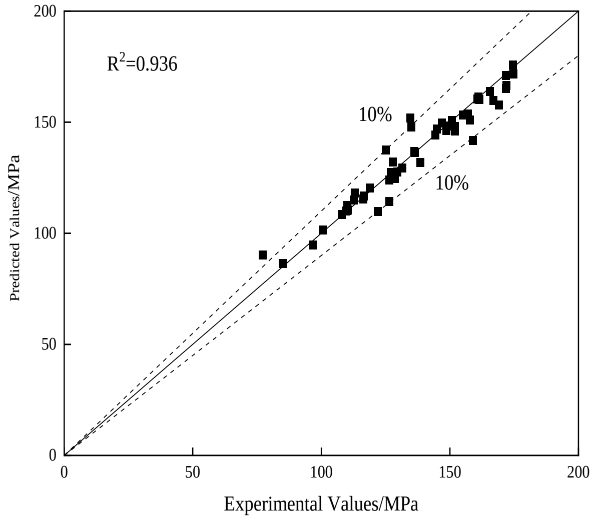
(c) The scatter fit plot of predicted values versus experimental values for the RF model.



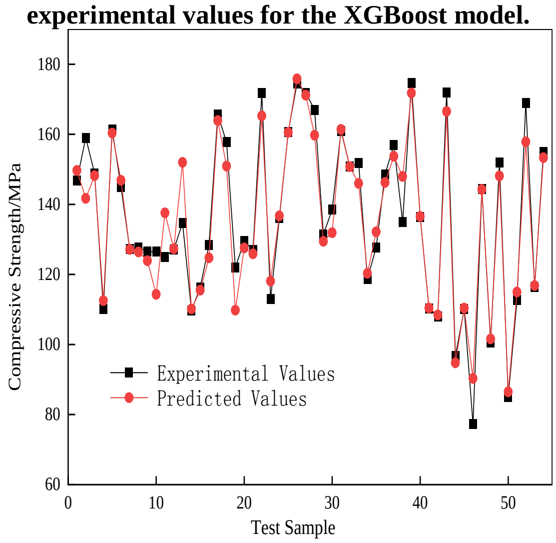
(b) The comparison plot between predicted values and experimental values for the DT model.



(d) The comparison plot between predicted values and experimental values for the RF model.



(e) The scatter fit plot of predicted values versus experimental values for the XGBoost model.



(f) The comparison plot between predicted values and experimental values for the XGBoost model.

Figure 2: The results of predicting UHPC compressive strength based on different machine learning models.

4.2 The Evaluation Results of The Machine Learning Models.

The evaluation metrics for the three models are depicted in Figure 3. From the graph, it is evident that the XGBoost model exhibits the best performance in predicting UHPC compressive strength, with an R2 value of 0.936. Compared to the RF model and the DT model, it has improved by 2.41% and 6.48%, respectively. This indicates that the XGBoost model can more accurately fit the sample data.The performance on the RMSE and MAE metrics also demonstrates the superiority of the XGBoost model. Compared to the RF model and DT model, the RMSE of XGBoost is reduced by 26.37% and 41.01%, respectively, while the MAE is reduced by 13.93% and 27.38%, respectively. This suggests that the difference between the predicted values and experimental values is smallest for the XGBoost model, and its predictive accuracy is significantly better than that of the RF and DT models.This superiority can be attributed to the various regularization terms provided by XGBoost to constrain the complexity of the model. Regularization introduces additional penalty terms during the model training process to reduce overfitting to the training data, thus making the predicted values closer to the experimental values.

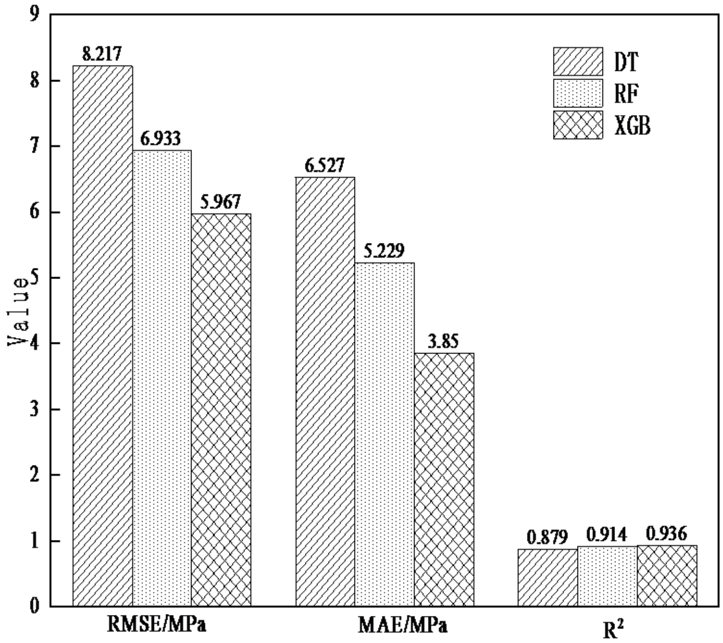


Figure 3: The comparison of evaluation metrics for the three models.

4.3 Interpretation and Analysis of Machine Learning Models.

Figure 4 presents the results of the global feature analysis of input variables for UHPC compressive strength based on the SHAP algorithm. The results indicate that the fiber content is the most sensitive input variable affecting UHPC compressive strength, with the greatest impact, followed by age, while the effect of fiber diameter is relatively smaller. The importance of added supplementary cementitious materials decreases from highest to lowest in the order of silica fume, fly ash, lime powder, slag, and glass powder. For fiber content, the lower the content, the lower its importance, whereas the higher the content, the higher its importance. The second most important variable affecting compressive strength is age, with its importance increasing as age increases.

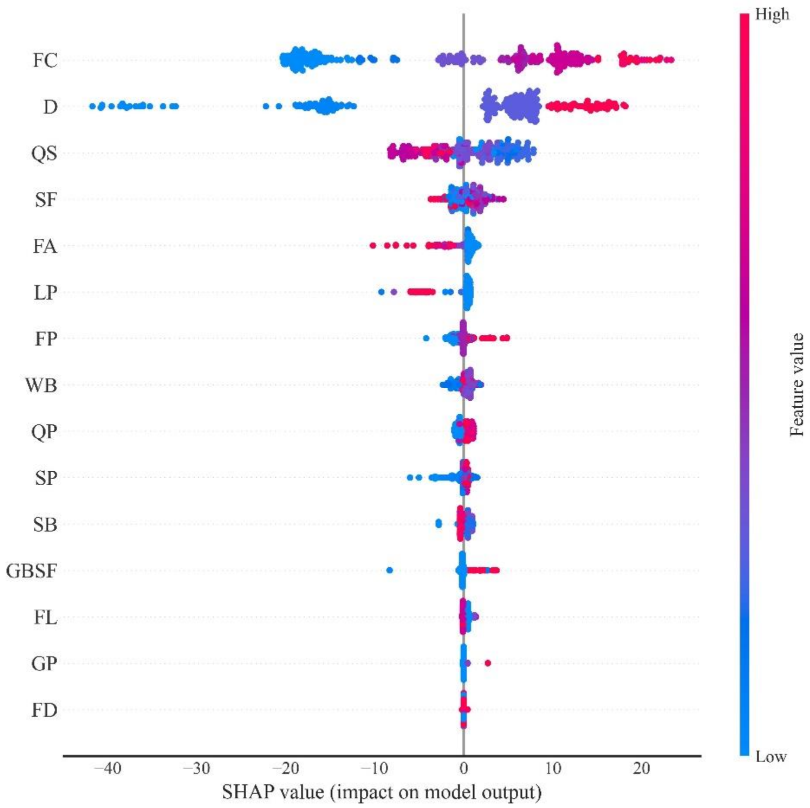


Figure 4: SHAP Global Feature Map

V. CONCLUSION

This study built a database based on collected data from 168 experimental groups, using decision trees, random forests, and extreme gradient boosting tree models to predict and analyze the compressive strength of UHPC. SHAP was employed to interpret the models. Results indicate high predictive accuracy across all models, with XGBoost performing the best, exhibiting lower prediction errors and better fitting effects, effectively reducing the risk of model overfitting. The study provides a comprehensive explanation of the factors influencing the compressive strength of UHPC.

REFERENCES

[1]. YOO D Y，BANTHIA N. Mechanical properties of ultra-high-performance fiber-reinforced concrete: A review[J]. Cement and Concrete Composites, 2016, 73: 267-280.

[2]. YU R，SPIESZ P，BROUWERS H. Effect of nano-silica on the hydration and microstructure development of ultra-high performance concrete (UHPC) with a low binder amount[J]. Construction and Building Materials, 2014, 65: 140-150.

[3]. SALEHI H，BURGUEñO R. Emerging artificial intelligence methods in structural engineering[J]. Engineering structures, 2018, 171: 170-189.

[4]. Yuan X, Tian Y, Ahmad W, et al. Machine learning prediction models to evaluate the strength of recycled aggregate concrete[J]. Materials, 2022, 15(8): 2823.

[5]. GüçLüER K，ÖZBEYAZ A，GöYMEN S, et al. A comparative investigation using machine learning methods for concrete compressive strength estimation[J]. Materials Today Communications, 2021, 27: 102278.

[6]. SOLHMIRZAEI R，SALEHI H，KODUR V. Predicting flexural capacity of ultra-high-performance concrete beams: machine learning‑based approach[J]. Journal of Structural Engineering, 2022, 148(5): 04022031.

[7]. SAFAVIAN S R，LANDGREBE D. A survey of decision tree classifier methodology[J]. IEEE Transactions on Systems Man and Cybernetics, 1991, 21(3): 660-674.

[8]. FAWAGREH K，GABER M M，ELYAN E. Random forests: from early developments to recent advancements[J]. Systems Science and Control Engineering: An Open Access Journal, 2014, 2(1): 602-609.

[9]. RAMRAJ S，UZIR N，SUNIL R, et al. Experimenting XGBoost algorithm for prediction and classification of different datasets[J]. International Journal of Control Theory and Applications, 2016, 9(40): 651-662.

[10]. LUNDBERG S M，LEE S I. A unified approach to interpreting model predictions[J]. Advances in neural information processing systems, 2017, 30: 4768‑4777.