



## RESEARCH ARTICLE

# A comprehensive study on the determination of compressive strength of concrete containing fly ash and blast furnace slag based on machine and ensemble learning

Yılmaz Yılmaz<sup>1\*</sup>, Safa Nayır<sup>1</sup>, Şakir Erdoğan<sup>1</sup><sup>1</sup> Karadeniz Technical University, Faculty of Engineering, Department of Civil Engineering, Trabzon, Türkiye

## Article History

Received 3 February 2025

Accepted 25 March 2025

## Keywords

Concrete

Compressive strength

Fly ash

Blast furnace slag

Machine and ensemble learning

## Abstract

The compressive strength (CS) of concrete is a critical parameter for the safety and longevity of structures as it directly affects the load bearing capacity and durability. However, determining the CS by conventional methods is time consuming, costly and requires a large amount of sample preparation. This study aims to quickly and cost-effectively predict and classify the CS of concrete and determine the influence of components on strength using machine learning (ML) algorithms as an alternative to traditional methods. Support Vector Machines (SVM), Decision Trees (DT), Multilayer Perceptron (MLP) machine learning algorithms and Random Forest (RF), Gradient Boosting (GB) and Extreme Gradient Boosting (XGB) ensemble learning models were trained on a dataset consisting of 1030 data points of concrete containing fly ash (FA) and blast furnace slag (BFS). The dataset was split into 75% training and 25% testing, and the Grid Search method and 5-fold cross-validation were applied in the training process. According to the results of the study, the XGB model showed the most robust performance in the prediction and classification of the CS of concrete with an  $R^2$  of 0.931 and an accuracy of 0.901. However, SVM and DT demonstrated inferior performance relative to the other four models. In addition, the models were classified normal strength concrete more successfully than low and high strength concrete. It was determined that the two most effective factors on CS were concrete age and cement dosage. While the increase in concrete age and cement dosage increased the strength, the increase in water content decreased the strength.

## 1. Introduction

Cement, which is used as the main binder in concrete manufacture, is one of the largest sources of carbon dioxide (CO<sub>2</sub>) emissions worldwide. Therefore, it significantly changes the environmental conditions in the region where it is produced [1,2]. Greenhouse gases are produced all over the world every year due to cement production in the developing construction sector. Moreover, annual cement production worldwide is about four billion tons and each ton of cement in this production process causes the same amount of CO<sub>2</sub> emissions [3,4]. Portland cement production is responsible for about 7% of CO<sub>2</sub> emissions. In particular, calcium oxide (CaO) has a substantial effect on the amount of CO<sub>2</sub> released during PC production [5]. To minimize these

---

\* Corresponding author ([yilmazyilmaz@ktu.edu.tr](mailto:yilmazyilmaz@ktu.edu.tr))

eISSN 2630-5763 © 2023 Authors. Publishing services by [golden light publishing®](https://goldenlightpublishing.com).This is an open-access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

harmful effects, the use of different recycled or waste materials in concrete production has been presented as a solution [6]. It is recommended to use waste or recycled material in concrete production to minimize the impact [6]. This approach will enable the utilization of waste without adversely affecting the suitability of concrete for its intended purpose. Currently, industrial waste materials such as FA, ground BFS, and silica fume, are used in concrete by replacing Portland cement in production [7]. This substitution enhances the mechanical [8,9] and durability [10,11] properties of hardened concrete and reduces CO<sub>2</sub> emission.

Concrete material, together with steel, is the most widely used material in the world construction industry. Concrete is used as a building material worldwide due to the advantages of strength, durability, hardness, porosity, density, fire resistance, and various properties [12]. Among these properties, CS is the most important property since it affects the strength and durability of concrete more strongly [13]. The CS of concrete, which is a heterogeneous material, is affected by the binder, aggregate, water, and admixtures [14]. All these components and their mixtures significantly affect CS. It is very difficult to accurately determine or predict the CS of concrete with complex content. The CS of concrete is usually determined by subjecting cylinder and cube samples to axial loading after a certain period (7, 28, 56, 90, and 120 days). This approach is globally standardized. However, because laboratory testing is costly and laborious, it is now considered inefficient and uneconomical.

Recently, with advances in computing and technology, ML algorithms have been utilized to predict many mechanical properties in concrete [15-18]. ML methods such as regression, classification, and feature extraction are used to determine the mechanical and durability properties of concretes and provide information on mix performance. As an example of these studies, Song et al. [19] investigated ML methods such as gene expression programming (GEP), artificial neural network (ANN), and DT to predict the CS of concrete containing FA. Using 98 experimental data points, they trained the models and evaluated their performance by k-fold cross-validation. According to the results, the bagging algorithm showed the best prediction performance with 95% accuracy; the R<sup>2</sup> values of GEP, ANN, and DT models were found to be 0.86, 0.81, and 0.75, respectively. Behnood et al. [20], predicting the CS of silica fume concrete using ANN is considered a dual objective problem optimizing accuracy and model complexity. For this purpose, a new method, Multi-objective Grey Wolf Optimization, was used and a total of 31 optimized ANN models were obtained. The final model has a single hidden layer with only five neurons, with an R<sup>2</sup> of 0.961 for all data. Furthermore, sensitivity analysis was carried out to examine the trends of the variables affecting the CS. Farooq et al. [21] compared RF and GEP algorithms for predicting the CS of high-strength concrete. The parameters used included cement dosage, coarse and fine aggregate ratio, water, and superplasticizer. The RF model showed outstanding performance using a DT, a weak base learner and achieved high accuracy with R<sup>2</sup> = 0.96. The GEP algorithm provided an empirical relationship with good agreement between predicted and actual values. In addition, comparisons were made with ANN and DT algorithms, and permutation calculations were carried out to assess the effect of variables.

This study aims to utilize ML to accurately predict the CS of concrete with FA and BFS. Although advanced ML techniques have been used in previous research, the need to integrate the SHAP method has been largely unaddressed. Therefore, this study applies a two-step approach: firstly, selecting the best-performing model among various ML algorithms and then applying SHAP methods to improve interpretability. In this scope, comprehensive data with 1030 data points was split into 75% training and 25% test sets. Then, the best hyperparameters of SVM, DT, RF, GB, XGB, and MLP models were determined and trained using the GridSearchCV method. The training was performed for both regression and classification. Three classes, Low, Normal, and High, were selected for classification. Finally, property analyses were performed for the components that affect the CS of concrete. The effectiveness of FA and BFS were compared. The study demonstrates that ML models significantly predicted and classified the compressive strength of concrete.

## 2. Data description

### 2.1. Dataset

The data used in this study for the prediction and classification of the CS of concretes with fly ash (FA) and BFS (BFS) were derived from the UC Irvine Machine Learning Repository [22]. In the dataset, which is composed of 1030 data points in total, the input characteristics are cement dosage, FA amount, BFS amount, water, superplasticizer, coarse aggregate, fine aggregate, and age. All input variables represent the amount of material in kilograms per 1 m<sup>3</sup> volume. The CS of concrete was utilized as the output variable. For classification, three strength classes are defined based on ACI 318 [23], Eurocode 2 (EN 1992-1-1) [24], and TBEC (2018) [25] standards. Concrete with a CS less than 25 MPa and generally used in non-bearing or temporary structures are labeled as ‘Low’, concrete with a CS between 25 MPa and 50 MPa and used in residential and general building applications are labeled as ‘Normal’, and concrete with a CS above 50 MPa and used in engineering structures such as bridges, dams, high-rise buildings are labeled as ‘High’. In the data set, 295 data points belong to the Low class, 525 data points belong to the Normal class and 210 data points belong to the High class. In this scope, the CS of concrete was estimated and classified according to their strength classes. The first 20 data points of the dataset used in the study are presented in Table 1.

Table 1. Head of dataset 20 data points [22]

	<b>Cem</b>	<b>BFS</b>	<b>FA</b>	<b>W</b>	<b>SP</b>	<b>Coarse_A</b>	<b>Fine_A</b>	<b>Age</b>	<b>C_Strength</b>
<b>1</b>	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
<b>2</b>	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
<b>3</b>	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
<b>4</b>	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.05
<b>5</b>	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.30
<b>6</b>	266.0	114.0	0.0	228.0	0.0	932.0	670.0	90	47.03
<b>7</b>	380.0	95.0	0.0	228.0	0.0	932.0	594.0	365	43.70
<b>8</b>	380.0	95.0	0.0	228.0	0.0	932.0	594.0	28	36.45
<b>9</b>	266.0	114.0	0.0	228.0	0.0	932.0	670.0	28	45.85
<b>10</b>	475.0	0.0	0.0	228.0	0.0	932.0	594.0	28	39.29
<b>11</b>	198.6	132.4	0.0	192.0	0.0	978.4	825.5	90	38.07
<b>12</b>	198.6	132.4	0.0	192.0	0.0	978.4	825.5	28	28.02
<b>13</b>	427.5	47.5	0.0	228.0	0.0	932.0	594.0	270	43.01
<b>14</b>	190.0	190.0	0.0	228.0	0.0	932.0	670.0	90	42.33
<b>15</b>	304.0	76.0	0.0	228.0	0.0	932.0	670.0	28	47.81
<b>16</b>	380.0	0.0	0.0	228.0	0.0	932.0	670.0	90	52.91
<b>17</b>	139.6	209.4	0.0	192.0	0.0	1047.0	806.9	90	39.36
<b>18</b>	342.0	38.0	0.0	228.0	0.0	932.0	670.0	365	56.14
<b>19</b>	380.0	95.0	0.0	228.0	0.0	932.0	594.0	90	40.56
<b>20</b>	475.0	0.0	0.0	228.0	0.0	932.0	594.0	180	42.62

## 2.2. Dataset visualization

It is important for ML applications to determine and visualize the distribution of data. These distributions and visualizations contribute to determining the relationships between the data and evaluating the ML results. Table 2 shows the statistical data of the input and output features used in the study. Components in the dataset such as cement, BFS, and FA have the right shifted distributions and a wide variation, while the distributions of components such as water and fine aggregate are more symmetrical and normal. Superplasticizer has high skewness and kurtosis, i.e. mostly low values, but some samples show high levels. Age has a fairly wide range, shifted to the right, and shows a very pointed distribution. Coarse aggregate and CS, on the other hand, were flatter, concentrated on the left, and exhibited symmetrical distributions. These results revealed that the constituents in the concrete mix are highly variable, but some of them have more pronounced and regular distributions.

The correlation matrix, which shows the relationship between the features and how each variable interacts with the others, is another important tool used in visualizing the data. The correlation matrix of the data used in the study is given in Fig. 1. The correlation matrix shows that the CS of concrete is mostly influenced by cement content (Cem, +0.50), water content (W, -0.29), superplasticizer (SP, +0.37) and age (Age, +0.33). It was determined that as the amount of cement and superplasticizer increased, the strength increased, while the increase in the amount of water decreased the strength. FA (FA, -0.11) and aggregates (Coarse\_A, -0.16; Fine\_A, -0.17) had a weak effect on the strength. A strong negative correlation (-0.66) was observed between water and superplasticizer while increasing age increased strength. The results obtained from the correlation matrix are in line with previous experimental studies [26,27].

## 3. Methodology

### 3.1. Machine learning models and implementation

The machine and ensemble learning models selected within the scope of the study were determined to predict and classify the CSs of FA and BFS admixed concretes, considering the complex nonlinear relationships of these concrete types. Detailed explanations of these ML models, which have proven their successful performance on different data sets in previous studies, are presented in the following.

**Table 2.** Statistical parameters of the dataset

Features	Abbreviation	Mean	Min	Max	Std	Skewness	Kurtosis
Cement	Cem	281.168	102	540	104.456	0.509	-0.524
Blast furnace slag	BFS	73.896	0	359.4	86.237	0.800	-0.512
Fly ash	FA	54.188	0	200.1	63.966	0.537	-1.328
Water	W	181.567	121.8	247	21.344	0.075	0.116
Superplasticizer	SP	6.205	0	32.2	5.971	0.906	1.399
Coarse aggregate	Coarse_A	972.919	801	1145	77.716	-0.040	-0.602
Fine aggregate	Fine_A	773.580	594	992.6	80.137	-0.253	-0.108
Age	Age	45.662	1	365	63.139	3.264	12.104
Compressive strength	C_Strength	35.832	2.332	85.599	16.777	0.417	-0.112

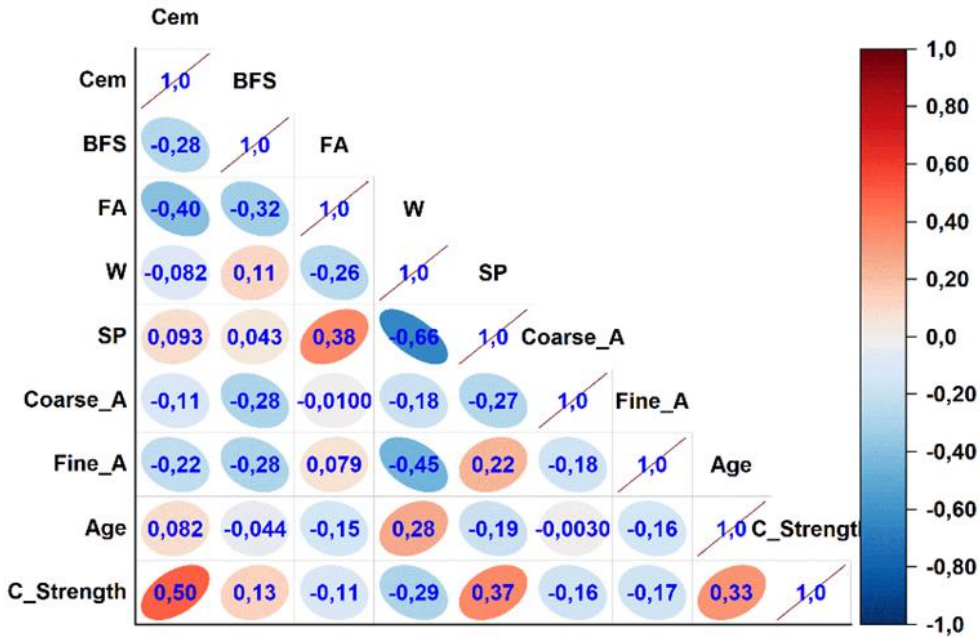


Fig. 1. Correlation matrix of the dataset

### 3.1.1. Support vector machines

SVM is a powerful and flexible ML method used in both classification and regression problems. SVM, announced by Cortes and Vapnik [28], aims to determine the optimal boundary that provides maximum separation between different classes by separating the data with a hyperplane. This method uses kernel functions to work effectively on nonlinear problems and can learn discriminative boundaries by moving the data space to higher dimensions. Owing to its robust performance, its ability to generalize well with little data, and its robust mathematical foundation, it is widely used in many fields from engineering to biology.

### 3.1.2. Decision trees

DT is an easy-to-understand and effective ML method used in both classification and regression problems. This method creates decision rules by branching the data in a tree-like hierarchy. Each node divides the data based on a feature, while leaf nodes represent the final class or value prediction. Introduced by Quinlan [29] with the ID3 algorithm, this method was later developed into popular versions such as C4.5 [30] and CART [31]. Decision trees offer a highly interpretable model owing to their easily visualizable structure and can work effectively on multidimensional datasets.

### 3.1.3. Random forest

RF is a popular ML method used for both classification and regression problems, offering high accuracy and robustness [32]. This method, proposed by Leo Breiman [33], builds an ensemble model by combining multiple decision trees. Each tree is trained on a random subset of the dataset and the predictions are combined by taking the majority vote (classification) or the average (regression) of all trees. This method improves generalization performance while reducing the risk of overfitting. RF, which is especially effective in large datasets and high-dimensional problems, is preferred in a wide range of applications due to its strong prediction capabilities and the fact that it does not require extreme parameter sensitivity.

### 3.1.4. Gradient boosting

GB is a robust ensemble learning method for both classification and regression problems. Proposed by Jerome H. Friedman [34], it aims to build a robust model by progressively correcting the errors of weak predictors (usually decision trees). The model uses gradient descent to minimize the errors of the previous model at each iteration [35]. This process enables the capture of complex non-linear relationships and high-accuracy predictions.

### 3.1.5. Extreme gradient boosting

XGB is a speed and accuracy-optimized ensemble learning method that provides high performance in classification and regression problems. XGB, announced by Tianqi Chen [36], basically builds an ensemble of decision trees by improving the gradient-boosting algorithm. This method improves the accuracy of the model step by step, focusing on reducing the error at each iteration. Owing to its ability to deal with missing data, scalability, and regularization, XGB avoids overfitting and works effectively on large and complex datasets [37].

### 3.1.6. Multilayer perceptron

MLP is a fundamental building block of artificial neural networks and is a powerful ML method used in both classification and regression problems. MLP is a feed-forward network with at least one hidden layer, where each neuron is fully connected to the neurons in the next layer [38]. This method uses activation functions to learn non-linear relationships and a backpropagation algorithm is usually used to update the weights [39]. MLP, which can work effectively on high-dimensional and complex datasets, is considered the pioneer of deep learning and is nowadays used in many application areas [40].

### 3.1.7. Data preprocessing and splitting

Data preprocessing is the process of making data suitable for use in the analysis or training of ML models [41]. Data normalization is done to eliminate scale differences between features. In this study, Z-score normalization is used; the value of each feature is brought to the same scale by subtracting the mean and dividing by the standard deviation [42]. Therefore, training or analyzing the model could be done in a more balanced way. The separation of the data set into training and test sets is important to assess the accuracy and generalizability of the model [43]. Commonly, 70-80% of the data is used for training and 20-30% for testing. The training set is used for learning the model and the test set is used to measure the performance of the model with independent data. In this study, previous studies were analyzed, and the data were randomly allocated as 75% training and 25% test data.

### 3.1.8. Hyperparameter fine-tuning

Grid Search [44] is a hyperparameter fine-tuning method often used in ML applications. This optimization technique is used to determine the hyperparameters of a model and aims to find the one that will provide the best performance by trying different combinations of hyperparameters. Cross-validation (CV) [45] is a technique used to test the generalization ability of the model. In this method, the data set is divided into training and validation subsets, and the model is evaluated in both subsets. This process can be repeated on different subsets of the data set to ensure a reliable evaluation of the model. In this study, Grid Search and CV are used together to determine the best hyperparameters. GridSearchCV evaluates the performance of the model for hyperparameter combinations and selects the best combination by applying a 5-fold CV (Fig. 2). For each combination, the data set is divided into five equal parts and each part is used for validation. As a result of the optimization, the best hyperparameters of the models for prediction and classification applications are given in Table 3.

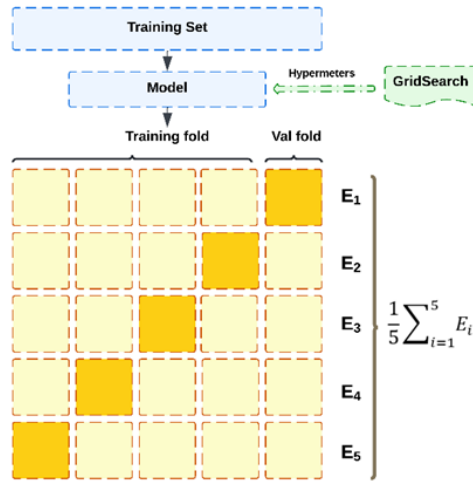


Fig. 2. 5-fold cross-validation method in study

Table 3. Optimal hyperparameter combination in classification and prediction

Model	Parameters	Range	Best Parameter	
			Classification	Prediction
SVM	C	[1, 2, 10, 100, 300]	10	100
	epsilon	[0.5, 0.1, 0.01, 0.001]	0.01	0.5
	kernel	['linear', 'rbf', 'poly']	'rbf'	'rbf'
	min_samples_split	[2, 5, 10]	5	2
DT	min_samples_leaf	[1, 2, 4, 10]	1	10
	max_depth	[None, 10, 20]	None	20
	n_estimators	[50, 100, 200]	200	200
RF	max_depth	[None, 10, 20]	None	None
	min_samples_split	[2, 5]	5	2
	min_samples_leaf	[1, 2]	1	1
	n_estimators	[50, 100, 200, 500]	200	500
XGB	learning_rate	[0.05, 0.1, 0.2]	0.1	0.1
	max_depth	[3, 6, 10]	3	3
	subsample	[0.8, 1.0]	0.8	-
	n_estimators	[50, 100, 200, 500]	50	500
GB	learning_rate	[0.01, 0.1, 0.2]	0.2	0.1
	max_depth	[3, 6, 10]	3	3
	hidden_layer_sizes	[(50,), (100,), (200,)]	(50,)	(100,)
MLP	activation	['relu', 'tanh']	'tanh'	'relu'
	solver	['adam', 'sgd']	'adam'	'adam'
	learning_rate	['constant', 'adaptive']	'constant'	'constant'



### 3.1.9. Performance criteria

In this study, the performance of the models was evaluated with different metrics for prediction and classification tasks. In the prediction task, R-square ( $R^2$ ), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) metrics were used.  $R^2$  is a metric that measures how well the model explains the data in the prediction task. MAE is used to evaluate the average size of the prediction errors, while RMSE measures the performance of the model at large deviations by penalizing large errors more.

In the classification task, several metrics were used to assess how well the model was able to discriminate between classes. Accuracy gives the overall proportion of correct predictions, while precision measures how many of the predicted positives are correct. Recall measures how accurately the model detects all true positives. The F1 Score balances precision and sensitivity and is particularly useful in situations with class imbalance. Finally, the AUC (Area Under the ROC Curve) measures the ability of the model to distinguish between positive and negative classes. The performance metrics used in the prediction and classification of concrete compressive strength were determined by analyzing previous studies and the equations of these metrics are given in Eq. 1-Eq. 3, and the mathematical expressions of the classification metrics are given in Eqs. (4)-(6) [15,17,46,47,48].

$$R^2 = 1 - \frac{\sum_{k=1}^n (y'_k - y_k)^2}{\sum_{k=1}^n (y'_k - \bar{y})^2} \quad (1)$$

$$MAE = \frac{\sum_{k=1}^n |y'_k - y_k|}{n} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^n (y'_k - y_k)^2}{n}} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

### 3.2. Application of the SHAP method

The SHAP (SHapley Additive exPlanations) method, derived from Shapley values introduced by Lloyd Shapley [49], is widely used to interpret ML model predictions. SHAP values quantify each feature's contribution to the prediction by averaging its marginal contribution across all possible combinations of features. This ensures an equitable distribution of the overall prediction value among the features [50]. Once the model is trained, SHAP values are computed using the formula in Eq. 7, where  $\phi_i$  represents the contribution of feature  $i$ . The equation is:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S)) \quad (7)$$

where  $N$  is the set of all features,  $|S|$  is a subset excluding feature  $i$ ,  $|S|$  is the size of  $S$ , and  $v(S)$  denotes the model output when only using features in  $S$ . By averaging over all subsets  $S$ , the Shapley value provides insight into how each feature influences the model's predictions.



SHAP analysis offers several visualization techniques to improve model explainability. The SHAP summary plot, which shows the impact of features on the prediction, visually reveals the importance of each feature in the model. SHAP dependence plot helps to understand the effect of a feature on model prediction by showing its relationship with other features and helps to understand interactions. The SHAP force plot visually presents the contribution of each feature to the prediction while explaining the prediction of a particular sample. Furthermore, the SHAP waterfall plot shows the contribution of each feature in more detail, which allows us to focus specifically on the prediction of a single instance. One of the key advantages of SHAP is that it provides a better understanding of how the model makes decisions. This feature increases the transparency of the model, especially in complex models. However, SHAP calculations can lead to high computational costs on large datasets, as the impact of each feature has to be calculated over all subsets. Moreover, SHAP works more efficiently for some models, while others may experience computational difficulties. However, the most important contribution of SHAP is that it allows us to clearly understand the contribution of each feature to the model's prediction, thus increasing the reliability and accuracy of the model.

## 4. Results

### 4.1. Regression results

The results of predicting the CS of FA and BFS with concretes are given in Table 4. According to the performance metrics in Table 4, the GB and XGB models have the best accuracy and minimum error values on the test data. XGB has a high generalization capacity with an  $R^2$  value of 0.931 and minimizes the prediction errors with an MAE of 2.924 and RMSE of 4.325. Similarly, XGB performed effectively with an  $R^2$  value of 0.928 and low error margins with 3.024 MAE and 4.425 RMSE. RF showed slightly lower accuracy compared to GB and XGB with an  $R^2$  of 0.888 but still performed strongly with an MAE of 3.750 and RMSE of 5.512. MLP provided a good fit on the test data with an  $R^2$  value of 0.911, but the values of 3.563 MAE and 4.894 RMSE were behind GB and XGB, indicating that the prediction accuracy of the model was slightly lower. SVR showed low accuracy in the test phase with an  $R^2$  of 0.873 but produced larger prediction errors with MAE of 4.377 and RMSE of 6.300. The DT model, on the other hand, achieves lower accuracy in the test phase with an  $R^2$  value of 0.855, but its generalization capacity is limited with high error values of 4.462 MAE and 6.877 RMSE. GB and XGB are the strongest models in terms of accuracy and generalization, whereas SVR and DT have lower performance due to their high error margins.

Table 4. CS prediction performance metrics

Models	Train			Test		
	$R^2$	MAE	RMSE	$R^2$	MAE	RMSE
SVR	0.932	3.731	5.388	0.873	4.377	6.300
DT	0.897	3.888	5.993	0.855	4.462	6.877
RF	0.985	1.345	2.023	0.888	3.750	5.512
GB	0.985	1.379	2.060	0.928	2.924	4.325
XGB	0.982	1.535	2.236	0.931	3.024	4.425
MLP	0.958	2.532	3.449	0.911	3.563	4.894

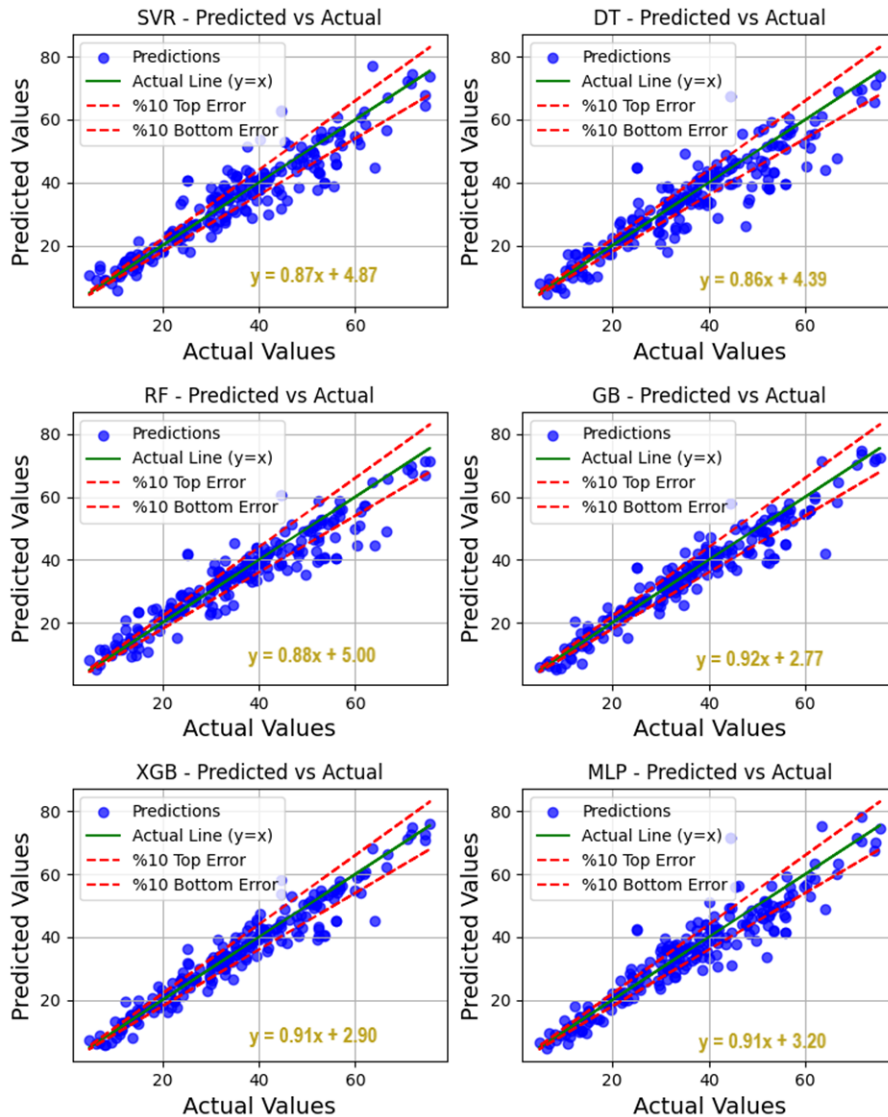


Fig. 3. Actual-predicted plots for the test set

The prediction performance of the ML models is presented in Fig. 3 with the slopes of the regression lines and the y-axis cut-off points. These values reveal the accuracy and generalization capacity of the models. The GB and XGB models demonstrated a very close linear relationship to the  $y=x$  ideal line with slope values of 0.92 and 0.91, respectively, and low y-axis cut-off points (2.77 and 2.90, respectively), indicating that systematic errors in predictions are minimal. The SVR and DT models, on the other hand, indicated a systematic deviation compared to the actual values with lower slope values (0.87 and 0.86). This indicated that the accuracy of the models in predicting the load-carrying capacity was limited. Although the RF model showed a more balanced performance with a slope value of 0.88, it has a lower generalization capacity compared to GB and XGB. The MLP model performed close to GB and XGB with a slope value of 0.91, but the slight increase in the y-axis cut-off point (3.20) indicates that the bias in the predictions should be improved.

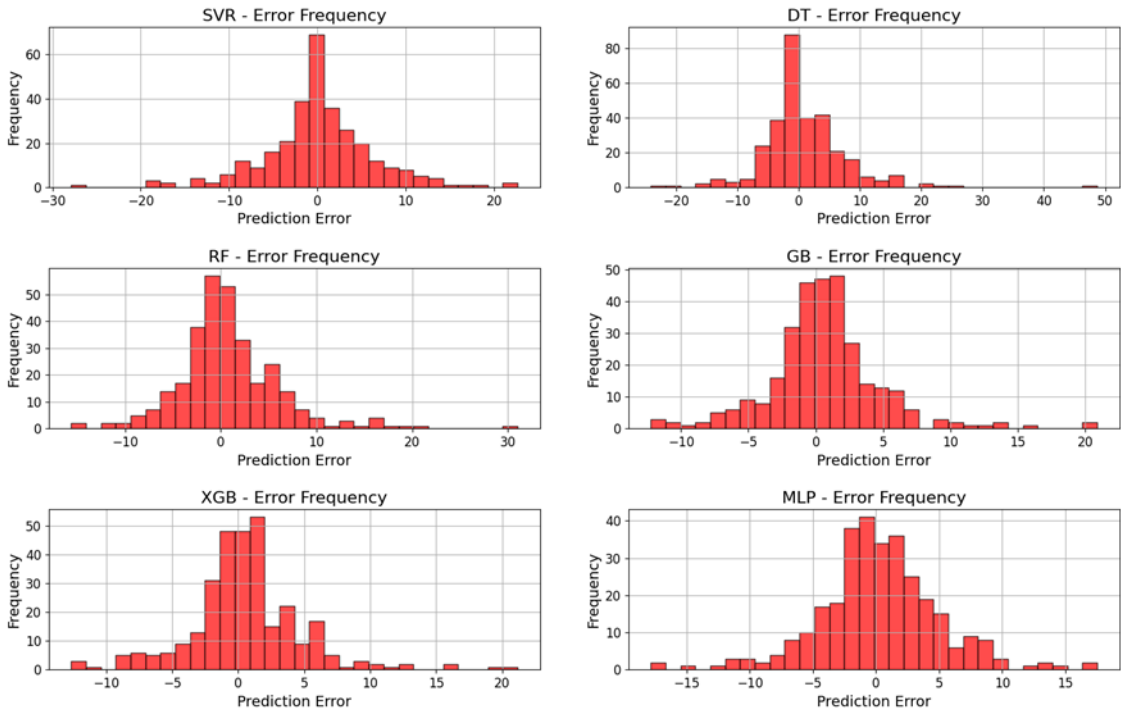


Fig. 4. Error distributions

Fig. 4 gives the distribution of the prediction errors of the models in detail. The GB and XGB models provided the highest accuracy and stability, with the error distribution concentrated in the range of -5 to +5, while at the same time minimizing the prediction bias with a symmetric distribution. The MLP model performed similarly, but the error distribution was slightly wider at the extremes, indicating an increase in bias in some cases. The RF model provided a balanced performance by concentrating most of its errors between -10 and +10, but it did not reach the accuracy level of GB and XGB. On the other hand, the SVR and Decision Tree DT models have a wider error distribution, with significant deviations between -20 and +20 for SVR and between -20 and +50 for DT, indicating that these models have low generalization capacity and produce inconsistent results in predictions.

## 4.2. Classification results

The preferred criteria for evaluating the performance of models in classification problems are confusion matrixes, Eqs. (4)-(6), and ROC curves. Fig. 5 shows the confusion matrixes of the test sets of all models in the study. According to the confusion matrixes, it was determined that the performance of the models differed between the classes. SVM performed quite well in the Low and Normal classes, but made a significant error in the High class, classifying most of the samples as Normal. DT, on the other hand, showed a balanced performance in general, but made significant misclassifications in the High class, indicating that the model's capacity to discriminate complex classes is limited. The RF model provided superior accuracy compared to the other models, especially in the Normal class, and obtained more balanced results between the classes. XGB stood out with its low error rate in the Normal and High classes and performed well in the Low class. Although the GB model provided balanced accuracy in general, it misclassified some samples as Normal in the High class. MLP achieved the highest accuracy in the Low class but showed more errors in the High class. In general, the High class was the class with the most errors for all models, while the Normal class was

best distinguished by XGB and RF. The results obtained revealed the importance of model selection, especially depending on the data distribution and class balance.

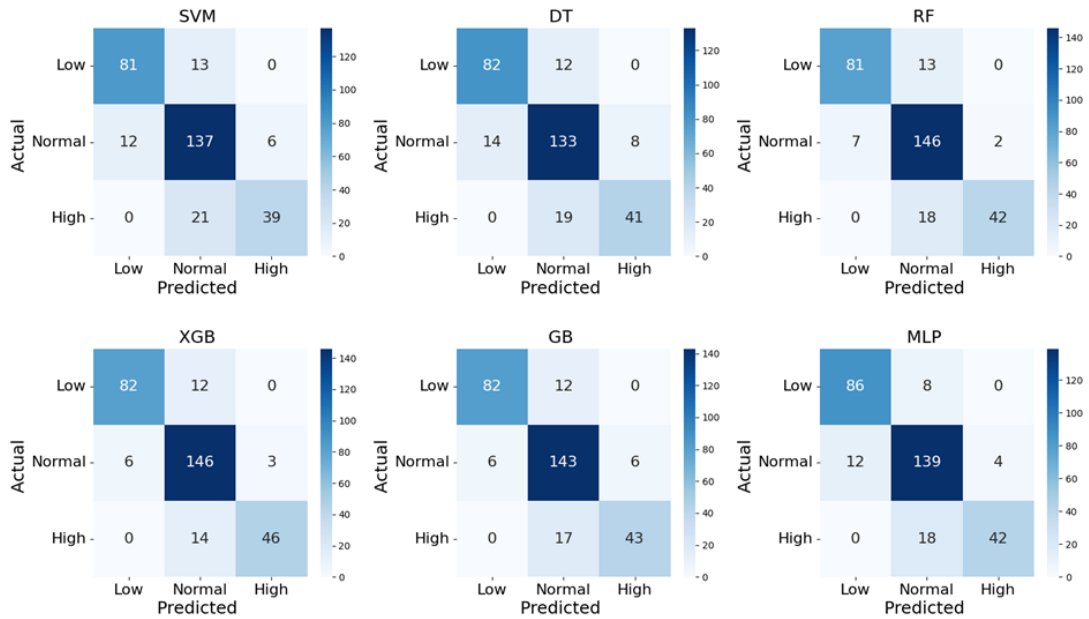


Fig. 5. Confusion matrices of the test set of the models

Table 5. Test set performance criteria for classification

Models	Accuracy	Recall	Precision	F1 Score	AUC		
					Low	Normal	High
SVM	0.842	0.832	0.835	0.829	0.980	0.920	0.958
DT	0.849	0.829	0.829	0.827	0.930	0.857	0.855
RF	0.881	0.871	0.879	0.869	0.975	0.928	0.970
XGB	0.901	0.887	0.892	0.886	0.981	0.939	0.976
GB	0.887	0.867	0.871	0.866	0.976	0.925	0.969
MLP	0.873	0.864	0.867	0.862	0.981	0.926	0.964

The performance metrics obtained from the confusion matrix are presented in Table 5. For the selection of metrics, AUC is separately reported for low, medium, and high-strength segments. The main reason for reporting Accuracy and F1-Score over the whole dataset is to provide a more comprehensive assessment of model performance. AUC is an important metric that measures the discrimination power between classes and has shown how sensitive the model is to a particular class, especially in unbalanced data distributions. Therefore, reporting it separately for each strength class (Low, Normal, High) revealed how accurately the model was able to classify different strength levels. According to the performance metrics, XGM is the most successful model with the highest accuracy (90.14%), recall (88.67%), precision (89.15%), and F1 Score (88.59%). RF performed particularly strongly in terms of accuracy (88.06%) and F1 Score (86.88%), second only to XGB. Although GB provided similarly high accuracy (88.73%), it lagged behind RF in terms of recall (86.73%) and precision (87.09%). MLP provided a satisfactory result with accuracy (87.31%) and F1

Score (86.22%) but was not as successful as RF and XGB. SVM and DT showed limited discrimination between classes with lower accuracy and F1 Score than the other models. SVM, in particular, showed a relatively balanced performance in terms of precision (83.51%) but was the weakest model in the overall results.

The ROC curves of the models are given in Fig. 6. The ROC curves show the comparative performance of the models at low, normal, and high data levels. At the low data level, the AUC values of all models are high (generally around 0.98), indicating that classification tasks are relatively easier at low density. While SVM and DT showed strong performance at low data levels, their performance decreased significantly at normal and high data levels. In particular, DT showed limited discrimination power at complex data levels with a lower AUC value (0.86). On the other hand, RF and XGB stood out as the most powerful models at low and high data levels and managed to maintain AUC values as high as 0.93 and 0.94 even at normal data levels. It is observed that RF generally offers a balanced performance between the classes, while XGB stands out, especially at high data levels. Although GB shows similar trends with RF and XGB, the slightly lower AUC value at normal data level suggests that it may be more inaccurate than the others in complex data structures. MLP performed very well at the low data level (AUC = 0.98) and showed a similarly strong performance at the high data level. However, the AUC value decreased to 0.93 at the normal data level, indicating that this model makes more errors in medium-density data scenarios.

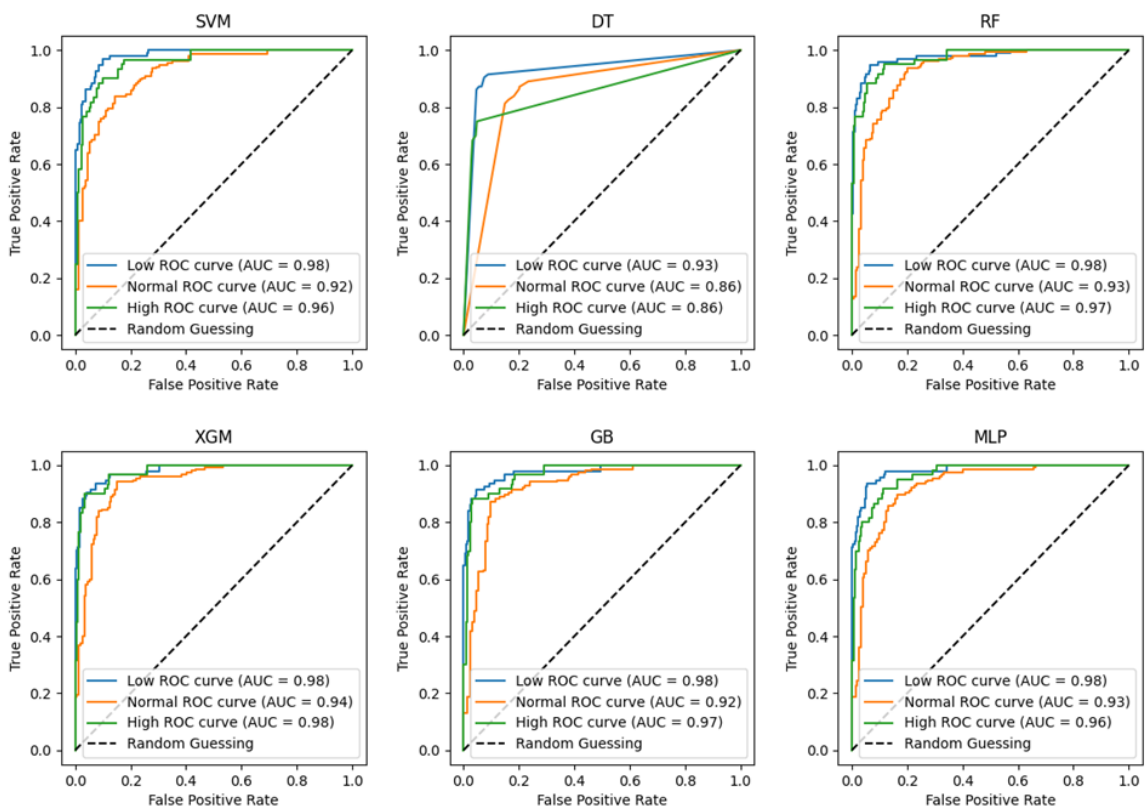


Fig. 6. ROC curves of the model

### 4.3. SHAP analysis results

The effect of the properties on the CS was determined using the SHAP method. Since the XGB model exhibits good accuracy in predicting the CS of concrete containing FA and BFS, this model was used in the SHAP analysis. The SHAP technique provides various graphical representations showing the influence of input characteristics on the estimation of output characteristics, the calculations in predicting the output variable, and the correlation between input attributes. In the average SHAP value plot given in Fig. 7, Age has the largest effect with an average SHAP value of 8.12 and thus is the factor that affects the output characteristic the most. This is closely trailed by Cem with 6.91, W with 4.11 and BFS with 3.13. The less important traits are SP, Fine\_A, Coarse\_A, and FA with SHAP values below 3.

Fig. 8, the SHAP summary plot summarizes the overall influence of characteristics on model predictions across the entire dataset. The SHAP values shown on the horizontal axis represent the positive or negative impact of a feature on the prediction, while the features listed on the vertical axis represent the inputs used in the model. The color of every dot shows the magnitude of the attribute, with red indicating high values and blue showing low values. In particular, traits such as Age and Cem have a wide SHAP distribution and have strong effects on the predictions in both positive and negative directions. On the other hand, the SHAP values of features such as Coarse\_A were generally close to zero, indicating that their impact on the forecasts was limited. The SHAP dependence plot given in Fig. 9 analyses the interactions between the features in the model in detail. The increase in the amount of Cem has a positive effect on the SHAP values, indicating that the amount of binder material plays a positive role in predictions. Similarly, the increase in FA Fine Aggregate has a small but consistent positive effect. On the other hand, an inverse relationship was observed between W and SHAP values; as the water content increased, the model prediction decreased. Age, on the other hand, generally had a positive effect on the model predictions, leading to an increase in the prediction values with increasing age. In addition, complex interactions between some attributes were also observed; for example, the effect on prediction varied as SP values increased, reflecting the sensitivity of the model to the relationships between such attributes. In conclusion, these analyses provide important insights into the prediction mechanism of the model, clearly showing that some characteristics (Cem and Age) have a dominant effect on the model, while others (Coarse\_A) have a limited effect.

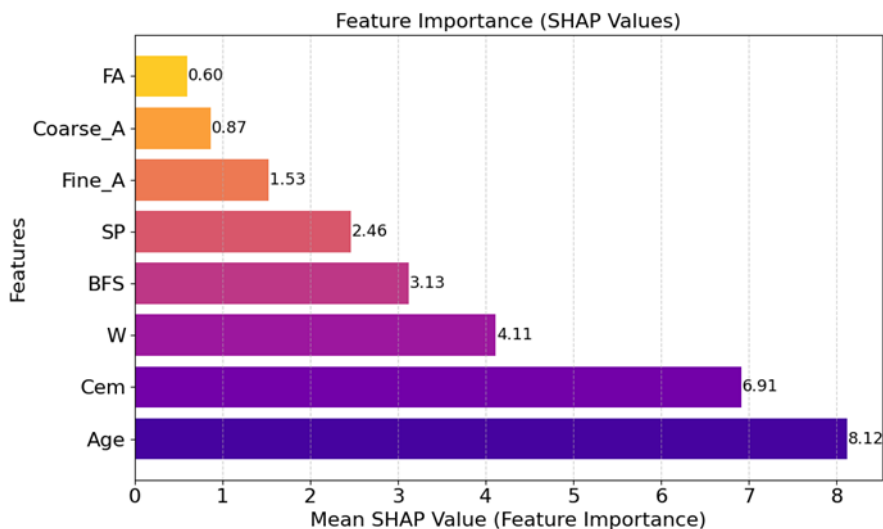


Fig. 7. SHAP importance plot

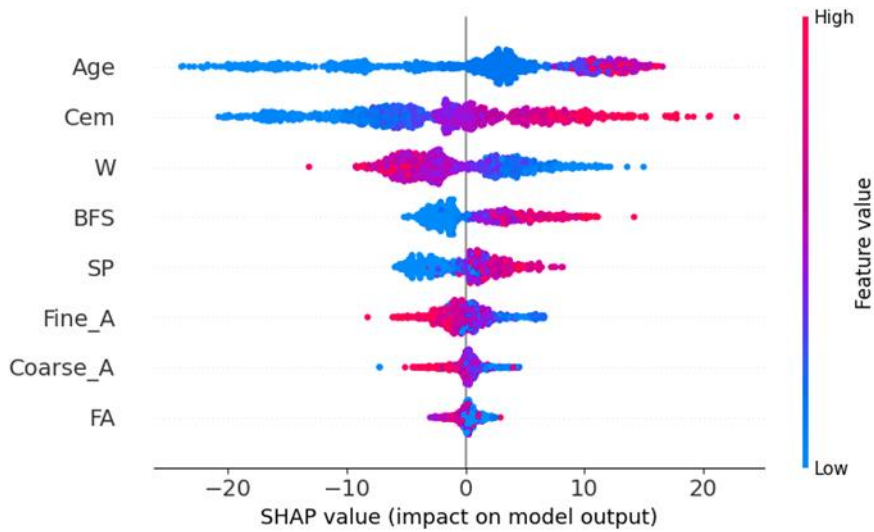


Fig. 8. SHAP summary plot

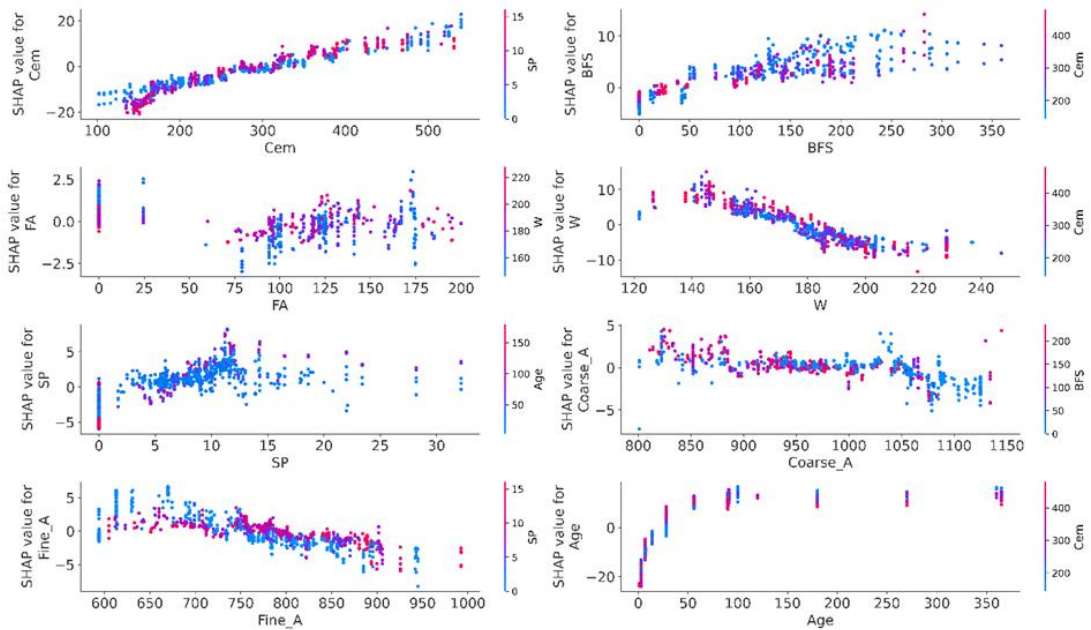


Fig. 9. SHAP dependence plot

## 5. Conclusions

In this study, prediction classification and feature attribute analyses were performed on the CS of FA and BFS admixed concretes using ML. The performance of the models was improved, and overfitting was prevented by using GridSearchCV optimization technique. The properties affecting the CS were analyzed in detail by SHAP analysis. Important results of the study are given:

- Ensemble learning models outperformed classical ML models in both prediction and classification.



- GB and XGB models stood out with the most robust accuracy and lowest error rates compared to other models in predicting concrete CS. XGB model was determined as the most successful model with an  $R^2$  value of 0.931 and attracted attention with its low error rates.
- It was determined that the most effective factors on the CS of concrete were concrete age, cement dosage, water quantity, and BFS content. While the increase in the amount of cement and concrete age increased strength, the increase in the amount of water decreased strength.
- The XGB model showed the most successful performance in correctly classifying the strength classes and reached an accurate rate of 90.14%. However, the high-strength class was the most difficult class to classify for all models. This was attributed to the imbalance in data distribution.
- In the error analysis, most of the prediction errors of the GB and XGB models were concentrated in a narrow range, indicating that the models provided robust accuracy and stability. In contrast, models such as DT and SVR showed lower generalization performance with wider error ranges.

In this study, CS of concretes containing FA and BFS were predicted and classified with robust accuracy. The results obtained show that the developed model can reliably predict the CS of concrete. In addition, SHAP analysis revealed that the three most influential factors on the compressive strength of concrete are concrete age, cement content, and water content. The two critical methods performed contributed to a better interpretation of the CS of concretes containing FA and BFS. Although several ML strategies have been established for predicting concrete compressive strength, a gap persists regarding the transparent and interpretable nature of these models, which is crucial for practical engineering applications. Future research can design decision support systems using larger datasets to realize optimum mix design and predict concrete CS in less time and with less effort. In addition, the prediction classification and property identification analyses performed in this study can be applied to other mechanical and durability properties. In future studies, mixed designs can be made for different concrete types with larger data sets. In these aspects, this study is a comprehensive resource that can guide future research.

### Conflict of interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

This research received no external funding.

### Data availability statement

Data generated during the current study are available from the corresponding author upon reasonable request.

### References

- [1] Scrivener KL, John VM, Gartner EM (2018) Eco-efficient cements: Potential economically viable solutions for a low-CO<sub>2</sub> cement-based materials industry. *Cem. Concr. Res.* 114:2–26. <https://doi.org/10.1016/j.cemconres.2018.03.015>
- [2] Xiao H, Duan Z, Zhou Y, Zhang N, Shan Y, Lin X, Liu G (2019) CO<sub>2</sub> emission patterns in shrinking and growing cities: A case study of Northeast China and the Yangtze River Delta. *Appl. Energy* 251:113384. <https://doi.org/10.1016/j.apenergy.2019.113384>
- [3] Yan H, Shen Q, Fan LCH, Wang Y, Zhang L (2010) Greenhouse gas emissions in building construction: A case study of one peking in Hong Kong. *Build. Environ.* 45:949–955. <https://doi.org/10.1016/j.buildenv.2009.09.014>
- [4] Dimoudi A, Tompa C (2008) Energy and environmental indicators related to construction of office buildings. *Resour. Conserv. Recycl.* 53(1–2):86–95.

- [5] Barcelo L, Kline J, Walenta G, Gartner E (2014) Cement and carbon emissions. *Mater. Struct.* 47:1055–1065. <https://doi.org/10.1617/s11527-013-0114-5>
- [6] Kajaste R, Hurme M (2016) Cement industry greenhouse gas emissions—Management options and abatement cost. *J. Clean. Prod.* 112:4041–4052. <https://doi.org/10.1016/j.jclepro.2015.07.055>
- [7] Batayneh M, Marie I, Asi I (2007) Use of selected waste materials in concrete mixes. *Waste Manag.* 27:1870–1876. <https://doi.org/10.1016/j.wasman.2006.07.026>
- [8] Nayak DK, Abhilash PP, Singh R, Kumar R, Kumar V (2022) FA for sustainable construction: A review of FA concrete and its beneficial use case studies. *Cleaner Mater.* 6:100143. <https://doi.org/10.1016/j.clema.2022.100143>
- [9] Raghuvanshi H, Singh N (2024) Fresh and mechanical properties of ground granulated BFS-based concrete: A review. *Mater. Today: Proc.* <https://doi.org/10.1016/j.matpr.2024.05.040>
- [10] Ahmad J, Kontoleon KJ, Majdi A, Naqash MT, Deifalla AF, Ben Kahla N, Isleem HF, Qaidi SMA (2022) A comprehensive review on the ground granulated BFS (GGBS) in concrete production. *Sustainability* 14:8783. <https://doi.org/10.3390/su14148783>
- [11] Amran M, Fediuk R, Murali G, Avudaiappan S, Ozbakkaloglu T, Vatin N, Karelina M, Klyuev S, Gholampour A (2021) FA-based eco-efficient concretes: A comprehensive review of the short-term properties. *Mater.* 14:4264. <https://doi.org/10.3390/ma14154264>
- [12] Sharma R, Jang JG, Bansal PP (2022) A comprehensive review on effects of mineral admixtures and fibers on engineering properties of ultra-high-performance concrete. *J. Build. Eng.* 45:103314. <https://doi.org/10.1016/j.jobbe.2021.103314>
- [13] Dey S, Kumar VVP, Goud KR et al. (2021) State of art review on self-compacting concrete using mineral admixtures. *J. Build. Rehabil.* 6:18. <https://doi.org/10.1007/s41024-021-00110-9>
- [14] Nithurshan M, Elakneswaran Y (2023) A systematic review and assessment of concrete strength prediction models. *Case Stud. Constr. Mater.* 18: e01830. <https://doi.org/10.1016/j.cscm.2023.e01830>
- [15] Paudel S, Pudasaini A, Shrestha RK, Kharel E (2023) CS of concrete material using machine learning techniques. *Cleaner Eng. Technol.* 15:100661. <https://doi.org/10.1016/j.clet.2023.100661>
- [16] Rathakrishnan V, Bt. Beddu S, Ahmed AN (2022) Predicting CS of high-performance concrete with high volume ground granulated blast-furnace slag replacement using boosting machine learning algorithms. *Sci. Rep.* 12:9539. <https://doi.org/10.1038/s41598-022-12890-2>
- [17] Elshaarawy MK, Alsaadawi MM, Hamed AK (2024) Machine learning and interactive GUI for concrete CS prediction. *Sci. Rep.* 14:16694. <https://doi.org/10.1038/s41598-024-66957-3>
- [18] Yilmaz Y, Nayır S (2024) Machine learning-based prediction of compressive and flexural strength of recycled plastic waste aggregate concrete. *Struct.* 69:107363. <https://doi.org/10.1016/j.istruc.2024.107363>
- [19] Song H, Ahmad A, Farooq F, Ostrowski KA, Maślak M, Czarniecki S, Aslam F (2021) Predicting the CS of concrete with FA admixture using machine learning algorithms. *Constr. Build. Mater.* 308:125021. <https://doi.org/10.1016/j.conbuildmat.2021.125021>
- [20] Behnood A, Golafshani EM (2018) Predicting the CS of silica fume concrete using hybrid artificial neural network with multi-objective grey wolves. *J. Clean. Prod.* 202:54–64. <https://doi.org/10.1016/j.jclepro.2018.08.065>
- [21] Farooq F, Nasir Amin M, Khan K, Sadiq MR, Javed MF, Aslam F, Alyousef R (2020) A comparative study of random forest and genetic engineering programming for the prediction of CS of high strength concrete (HSC). *Appl. Sci.* 10:7330. <https://doi.org/10.3390/app10207330>
- [22] Yeh, I. (1998). Concrete Compressive Strength [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5PK67>
- [23] ACI 318 (2019) Building Code Requirements for Structural Concrete (ACI 318-19) and Commentary. American Concrete Institute, Farmington Hills, MI, USA.
- [24] Eurocode 2 (2004) Design of Concrete Structures – Part 1-1: General Rules and Rules for Buildings. EN 1992-1-1, European Committee for Standardization, Brussels, Belgium.
- [25] TBEC (2018) Turkish Building Earthquake Code. Ministry of Environment and Urbanization, Ankara, Turkey.
- [26] Giergiczny Z (2019) FA and slag. *Cem. Concr. Res.* 124:105826. <https://doi.org/10.1016/j.cemconres.2019.105826>
- [27] Juang CU, Kuo WT (2023) Properties and mechanical strength analysis of concrete using FA, ground granulated BFS, and various superplasticizers. *Buildings* 13:1644. <https://doi.org/10.3390/buildings13071644>

- [28] Cortes C, Vapnik V (1995) Support-vector networks. *Mach. Learn.* 20:273–297. <https://doi.org/10.1007/BF00994018>
- [29] Quinlan JR (1986) Induction of decision trees. *Mach. Learn.* 1:81–106. <https://doi.org/10.1007/BF00116251>
- [30] Salzberg SL (1994) C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc. *Mach. Learn.* 16:235–240. <https://doi.org/10.1007/BF00993330>
- [31] Lewis RJ (2000) An introduction to classification and regression tree (CART) analysis. Presented at the Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, CA, USA.
- [32] Patel SV, Jokhakar VN (2016) A random forest-based machine learning approach for mild steel defect diagnosis. *IEEE Int. Conf. Comput. Intell. Comput. Res. (ICCIC)*:1–8. <https://doi.org/10.1109/ICCIC.2016.7919549>
- [33] Breiman L (2001) Random forests. *Mach. Learn.* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- [34] Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29:1189–1232. <https://doi.org/10.1214/aos/1013203451>
- [35] Alhakeem ZM, Jebur YM, Henedy SN, Imran H, Bernardo LFA, Hussein HM (2022) Prediction of ecofriendly concrete CS using gradient boosting regression tree combined with GridSearchCV hyperparameter-optimization techniques. *Mater.* 15:7432. <https://doi.org/10.3390/ma15217432>
- [36] Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*:785–794. <https://doi.org/10.1145/2939672.2939785>
- [37] Tao H, Ali ZH, Mukhtar F, Al Zand AW, Marhoon HA, Goliatt L, Yaseen ZM (2024) Coupled extreme gradient boosting algorithm with artificial intelligence models for predicting CS of fiber-reinforced polymer-confined concrete. *Eng. Appl. Artif. Intell.* 134:108674. <https://doi.org/10.1016/j.engappai.2024.10867>
- [38] Jitchaijaroen W, Keawsawasvong S, Wipulanusat W, Kumar DR, Jamsawang P, Sunkpho J (2024) Machine learning approaches for stability prediction of rectangular tunnels in natural clays based on MLP and RBF neural networks. *Intell. Syst. Appl.* 21:200329. <https://doi.org/10.1016/j.iswa.2024.200329>
- [39] Rumelhart D, Hinton G, Williams R (1986) Learning representations by back-propagating errors. *Nature* 323:533–536. <https://doi.org/10.1038/323533a0>
- [40] Chojaczyk AA, Teixeira AP, Neves LC, Cardoso JB, Guedes Soares C (2015) Review and application of artificial neural networks models in reliability analysis of steel structures. *Struct. Saf.* 52:78–89. <https://doi.org/10.1016/j.strusafe.2014.09.002>
- [41] García S et al. (2016) Big data preprocessing: Methods and prospects. *Big Data Anal.* 1:1–22. <https://doi.org/10.1186/s41044-016-0014-0>
- [42] Alasadi SA, Bhaya WS (2017) Review of data preprocessing techniques in data mining. *J. Eng. Appl. Sci.* 12:4102–4107. <https://doi.org/10.3923/jeasci.2017.4102.4107>
- [43] Nguyen QH et al. (2021) Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Math. Probl. Eng.*:4832864. <https://doi.org/10.1155/2021/4832864>
- [44] Liashchynskiy P, Liashchynskiy P (2019) Grid search, random search, genetic algorithm: A big comparison for NAS. arXiv:1912.06059. <https://doi.org/10.48550/arXiv.1912.06059>
- [45] Browne MW (2000) Cross-validation methods. *J. Math. Psychol.* 44:108–132. <https://doi.org/10.1006/jmps.1999.1279>
- [46] Maherian M.F., Baran S., Bicakci S.N., Toreyin B.U., Atahan H.N. (2023). Machine learning-based compressive strength estimation in nano silica-modified concrete. *Construction and Building Materials*, 408, 133684. <https://doi.org/10.1016/j.conbuildmat.2023.133684>
- [47] Li D., Tang Z., Kang Q., Zhang X., Li Y. (2023). Machine Learning-Based Method for Predicting Compressive Strength of Concrete. *Processes*, 11, 390. <https://doi.org/10.3390/pr11020390>
- [48] Loureiro A.A.B., Stefani R. (2024). Comparing the performance of machine learning models for predicting the compressive strength of concrete. *Discover Civil Engineering*, 1, 19. <https://doi.org/10.1007/s44290-024-00022-w>
- [49] Shapley LS (1953) A value for n-person games. In: Kuhn HW, Tucker AW (eds) *Contributions to the Theory of Games II*, *Annals of Mathematics Studies*, vol. 28. Princeton University Press, Princeton, pp. 307–317.
- [50] Rodríguez-Pérez R, Bajorath J (2020) Interpretation of machine learning models using Shapley values: Application to compound potency and multi-target activity predictions. *J. Comput.-Aided Mol. Des.* 34:1013–1026. <https://doi.org/10.1007/s10822-020-00314-0>