PREDICTION OF THE SLUMP AND STRENGTH OF HIGH STRENGTH CONCRETE USING RANDOM FOREST MODEL

Van Tuan Vu[1,*]

[1]*Le Quy Don Technical University, Hanoi, Vietnam*

DOI: 10.56651/lqdtu.jst.v6.n01.672.sce

Abstract

Random Forest (RF) has been successfully applied to a variety of engineering problems due to its simplicity, versatility, and suitability for both classification and regression tasks. Concrete, as a material composed of multiple complex elements, is influenced by numerous factors, posing challenges in accurately predicting its properties. In this article, an RF model is developed in predicting the slump and strength of concrete using mixed mineral admixtures from blast furnace slag and silicafume. The criterions to evaluate the accuracy of the models are the R squared (R^2) and the root mean square error (RMSE). Comparing the predicted data with the tested data, the result indicates that RF model should be used in predicting both the slump and strength of concrete.

Keywords*: Random forest (RF); prediction; the slump of concrete; the strength of concrete.*

1. Introduction

Machine Learning Techniques (MLT), a field that draws from multiple disciplines, uses various techniques to acquire new insights. The primary use of MLT is for prediction. Categorical variable values are predicted using classification while numerical variable values are predicted using regression. Regression entails examining the correlation between one or more independent variables and a dependent variable. Thus, in the last few decades, MLT has been successfully applied to virtually many engineering problems [1-8].

Concrete is a material that is composed of several complex elements, making it challenging to accurately predict its properties. Modeling concrete properties according to effect variables is a difficult task due to the low predictability of the material. Experimental designs face the biggest challenge concerning the high number of effect variables that can influence response variables. When multiple effect variables come into play, an increase in the number of trials is required. Additionally, the higher the amount of uncontrollable variables, the more difficult it is to obtain the true response function.

[*] Email: vantuanvu@lqdtu.edu.vn

Random forest (RF) is a cutting-edge ensemble algorithm with a variety of appealing features, including variable importance measures (VIMs), few model parameters, and robust resistance to overfitting [9, 10]. The algorithm is based on decision trees and RF models can yield satisfactory results even with default parameter settings [11]. By utilizing RF, the number of combinations of base predictors and parameter settings can be reduced to a single combination. Although RF has primarily been applied in fields such as ecology and bioinformatics, it has also been used in concrete-related studies [12-14]. For instance, Mohamed used the RF algorithm to predict the compressive strength of sustainable self-consolidating concrete [13]. Ozcan et al. created an RF model to analyze the effects of blast furnace slag and waste tire rubber powder on HPCCS [14]. Rao tested various algorithms to forecast the compressive strength of HPC and discovered that the RF model performed the best [15]. Recently, there have been many methods with high accuracy, such as Gradient Boosting Machines or AdaBoost. However, Random Forest remains a popular choice due to its simplicity (requiring fewer hyperparameters to tune compared to other ensemble models), reduced risk of overfitting, and strong performance across various domains and datasets.

It can be seen that there are almost no studies on the application of RF model to predict both the slump and strength of concrete. In this article, an RF model was developed in predicting the slump and strength of concrete. The testing data is from a previous study of Dinh Quang Trung [16]. In order to assess the accuracy of the models, the R-squared (R^2) and root mean square error (RMSE) are utilized as evaluation criteria. By comparing the predicted data with the tested data, relative conclusions can be drawn.

2. Data division and preprocessing

The training set comprises of samples created by varying the factors of experiments, including ratios such as water/adhesive (coded variable x_1), blast furnace slag/adhesive (coded variable x_2), silicafume/adhesive (coded variable x_3), and super-plastic additive/adhesive (coded variable x_4). The input variables selected for these experiments were the coded variables mentioned above (x_1 to x_4), while the output variables chosen were the slump and specific strength. Table 1 and Table 2 show the training and testing database of concrete samples, respectively.

Table 1. Training database of concrete samples [16]

No.	Coded variables				Factorial experiments				Slump	Specific strength
	x_1	x_2	x_3	x_4	W/A	BFS/A	SF/A	SPA/A	(cm)	(MPa)
1	-1	-1	-1	-1	0.3	30	3	0.6	9	88.7
2	1	-1	-1	-1	0.34	30	3	0.6	16.5	78.0
3	-1	1	-1	-1	0.3	50	3	0.6	12	85.9
4	1	1	-1	-1	0.34	50	3	0.6	19.5	77.1
5	-1	-1	1	-1	0.3	30	9	0.6	0.5	89.5
6	1	-1	1	-1	0.34	30	9	0.6	4	81.8
7	-1	1	1	-1	0	50	9	0.6	2	85.8
8	1	1	1	-1	0.34	50	9	0.6	7	81.2
9	-1	-1	-1	1	0.3	30	3	1	19	89.0
10	1	-1	-1	1	0.34	30	3	1	21	77.6
11	-1	1	-1	1	0.3	50	3	1	20.5	86.1
12	1	1	-1	1	0.34	50	3	1	22	77.2
13	-1	-1	1	1	0.3	30	9	1	11	89.8
14	1	-1	1	1	0.34	30	9	1	17.5	81.9
15	-1	1	1	1	0.3	50	9	1	17	86.5
16	1	1	1	1	0.34	50	9	1	21	81.5
17	-2	0	0	0	0.28	40	6	0.8	11	94.7
18	2	0	0	0	0.36	40	6	0.8	21	77.4
19	0	-2	0	0	0.32	20	6	0.8	15.5	84.1
20	0	2	0	0	0.32	60	6	0.8	19.5	79.8
21	0	0	-2	0	0.32	40	0	0.8	21	80.1
22	0	0	2	0	0.32	40	12	0.8	4	84.7
23	0	0	0	-2	0.32	40	6	0.4	2.5	82.7
24	0	0	0	2	0.32	40	6	1.2	19.5	82.8
25	0	0	0	0	0.32	40	6	0.8	18	84.7
26	0	0	0	0	0.32	40	6	0.8	17.5	84.1
27	0	0	0	0	0.32	40	6	0.8	18.5	83.2
28	0	0	0	0	0.32	40	6	0.8	17	84.2
29	0	0	0	0	0.32	40	6	0.8	16.5	83.3
30	0	0	0	0	0.32	40	6	0.8	17.5	84.2
31	0	0	0	0	0.32	40	6	0.8	18.5	82.1

* In Table 1, W/A is water/adhesive ratio; BFS/A is blast furnace slag/adhesive ratio; SF/A is silicafume/adhesive ratio; SPA/A is super-plastic additive/adhesive ratio.

Table 2. Testing - database of concrete samples [16]

No.	Coded variables				Factorial experiments				Slump	Specific strength
	x_1	x_2	x_3	x_4	W/A	BFS/A	SF/A	SPA/A	(cm)	(MPa)
1	-2	1.675	-0.765	1	0.28	56.75	3.705	1	18.5	92.2
2	-1	0.551	-1.121	0	0.3	45.51	2.637	0.8	17.5	88.5
3	0	1.409	-1.548	-1	0.32	54.09	1.356	0.6	18.5	75.4
4	1	0.678	-0.863	-1	0.34	46.78	3.411	0.6	18.5	72.6
5	2	1.034	-0.39	-1	0.36	50.34	4.83	0.6	18	70.3

To ensure that all variables receive equal attention during the training process and to reduce their dimension, preprocessing involves scaling the input and output variables to a range between -1.0 and 1.0. The calculation for the scaled value of each variable, *x*, is as follows:

$$x_n = \frac{x}{x_{\max}} \tag{1}$$

where $x_{\max}$ is maximum values of each variable x.

3. Overview of random forest

One of the most popular machine-learning methods is the RF model, which is based on the decision trees model. The concept of the random forest model, also known as bagging ensemble learning, was introduced by Breiman Leo (2001) [9]. Figure 1 and Figure 2 illustrate the typical decision tree and RF model.

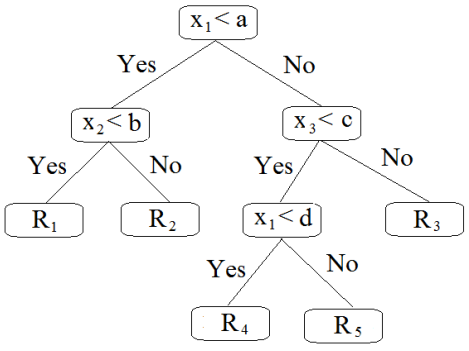

Figure 1. Decision tree model utilized for regression analysis.

In the decision tree model, the data is represented like a tree with branches and leaves, where different instances of the input data (such as x_1, x_2, x_3, etc.) are split by branches and the output is available at the leaf position (such as R_1, R_2, R_3, etc.). The

architecture of the decision tree model is essentially a series of if_then_else functions, which are calculated and optimized based on the input data set, tree complexity, and the depth of the if function. However, the individual decision tree model often overfits the input data, leading to the development of advanced models based on decision trees, such as the random forest model.

The RF model involves a group of decision trees that are pre-defined for both training and predicting. Each decision makes functions independently, using a limited dataset as input. The final result of the prediction is derived from the average result of all member trees. What makes this model interesting is that the trees are built randomly, using the bootstrap technique to select input data for each tree. This helps to better generalize the problem and prevent overfitting of individual decision tree models.

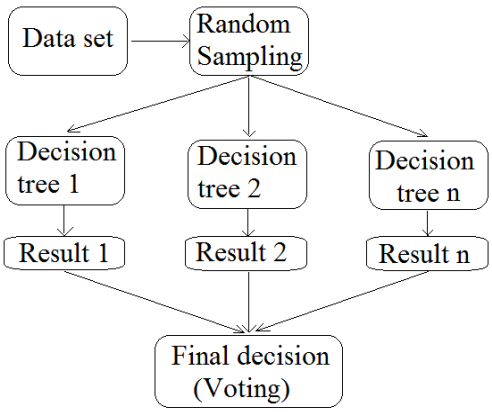

Figure 2. Graphical representation of the random forest model.

To build the model successfully, the important hyperparameters must be considered. These are: Number of trees in the forest (n); Maximum Depth of a tree (D); Minimum number of samples needed to separate plants (S); and Minimum number of samples per leaf (L).

The final prediction of the model is as follows:

$$y_i = \frac{1}{n} \cdot \sum_{j=1}^{n} f_j(x_i) \tag{2}$$

where n is the number of trees; y_i is the result of predicting the ith sample; x_i is the input vector data on the i^{th} sample; f_j is the estimator j^{th} in the forest.

More information about building decision trees and the hyperparameters of random forest can be found in literature [9].

4. Development of the random forest model

The RF model has been implemented in this study using the Python platform and the Sklearn library. To ensure optimal performance, we have examined the most critical hyperparameters that affect the model's output. This includes D, S, L, and n (the maximum number of trees), which have been carefully analyzed to determine the best possible value within the allowable range.

In order to provide a comprehensive understanding of the study, we have presented the survey scope in Table 3, which outlines the ranges and values of the parameters used in the investigation. Notably, previous research [9] has shown that the maximum number of trees does not need to be excessively high. Other hyperparameters, such as D, S, and L, determine the complexity of the decision trees, which can lead to overfitting of the model. It is important to consider that a model with high complexity may perform well with training data, but not with testing data, indicating overfitting.

Furthermore, the survey range of other hyperparameters has been chosen thoughtfully to ensure that when the hyperparameter value changes beyond the survey range, the model's performance does not change significantly. This is crucial in avoiding data leakage, where the model is overfitted to the training data and does not perform well with new, unseen data.

The utilization of the 5 Fold CV technique has proven to be a valuable asset in this study. By dividing the training set into five folds and using four for training and one for validation, we have been able to evaluate the model's performance in a more robust and reliable manner. The results indicate that our optimized RF model has been successful in achieving more accurate results, indicating improved generalization capabilities.

Table 3. Range of hyperparameters

Hyper parameter	Explain	Range
n	Number of trees	2 - 99
D	Max depth	2 - 20
S	Min samples to split	2 - 20
L	Min samples on a leaf	1 - 20

4.1. Effect of tree number on the effectiveness of models

Figure 3 depicts the impact of the number of trees on the efficacy of the random forest model. As is evident, a greater number of trees leads to an increase in accuracy.

However, to prevent overfitting of the model, the number of trees was selected to be 12 after evaluating other parameters.

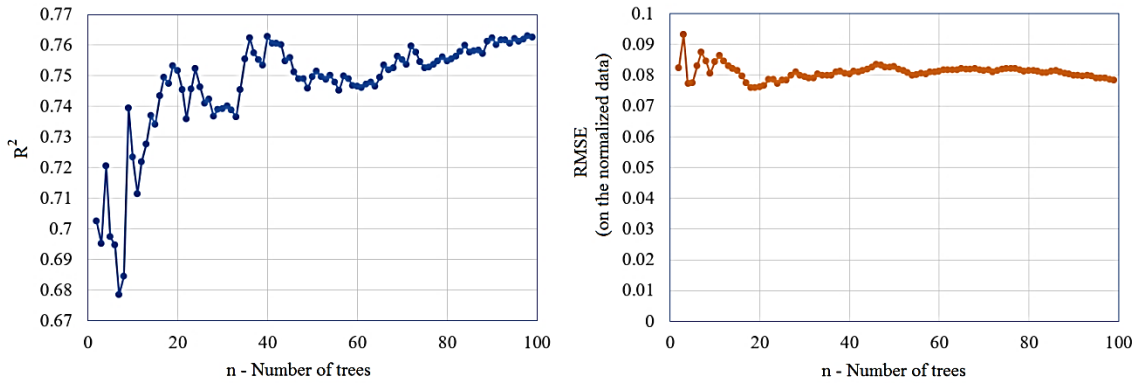

Figure 3. Effect of tree number on model effectiveness.

4.2. Effect of max depth on the effectiveness of models

The findings from a survey conducted to assess the model accuracy with respect to the variation in the maximum depth of trees (D), ranging from 2 to 20 with a fixed number of trees at 12, are presented in Figure 4. The results demonstrate that the optimal value of D is 7, resulting in an R^2 of 0.723, and an RMSE of 0.0845. Notably, the predictive performance significantly deteriorates for tree depths below 7, while increasing the tree depth beyond 7 does not yield considerable improvement in the predictive results.

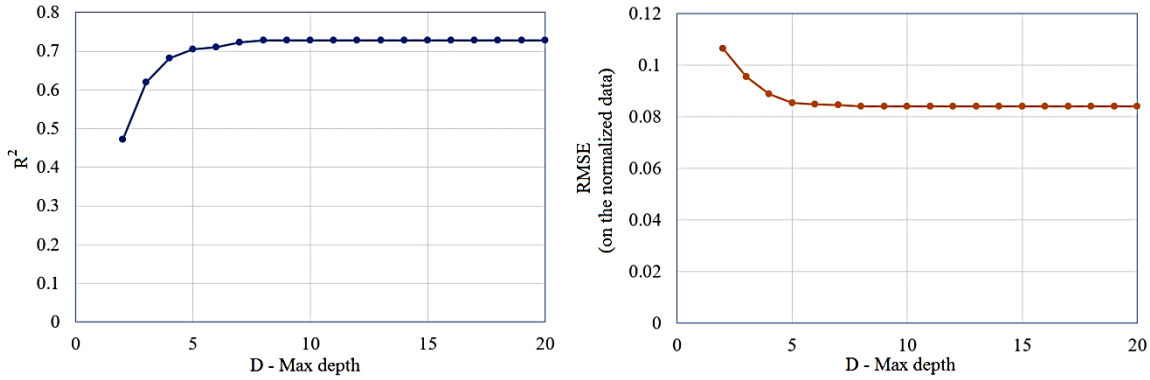

Figure 4. Effect of max depth on model effectiveness.

4.3. Effect of min samples to split on the effectiveness of models

A survey was conducted to evaluate the performance of the model with respect to the variation in the minimum number of samples required to split a tree (S), ranging from 2 to 20, with a fixed number of trees at 12 and maximum tree depth at 7, are depicted in Figure 5. The results reveal that the model's predictive ability deteriorates as the min samples to split of tree increases. However, if the min samples to split value is

set too low, the model may become overly complex, fitting too closely to the training data and performing poorly on new data (known as overfitting). On the other hand, if the value is set too high, the model may be too simple and not capture the underlying patterns in the data, resulting in underfitting. Therefore, it is important to find an appropriate balance in setting the min samples to split parameter. While the optimal value may vary depending on the specific dataset and problem, in general, it is recommended that this value not be set too small in order to avoid overfitting. The findings suggest that the optimal value of S is 6, leading to an R^2 of 0.727 and an RMSE of 0.085.

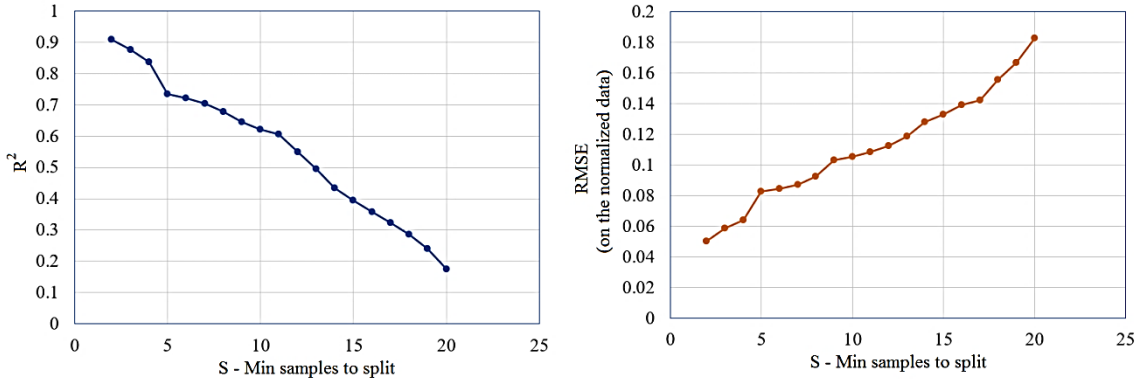

Figure 5. Effect of min samples to split on model effectiveness.

4.4. Effect of minimum leaf samples on the effectiveness of models

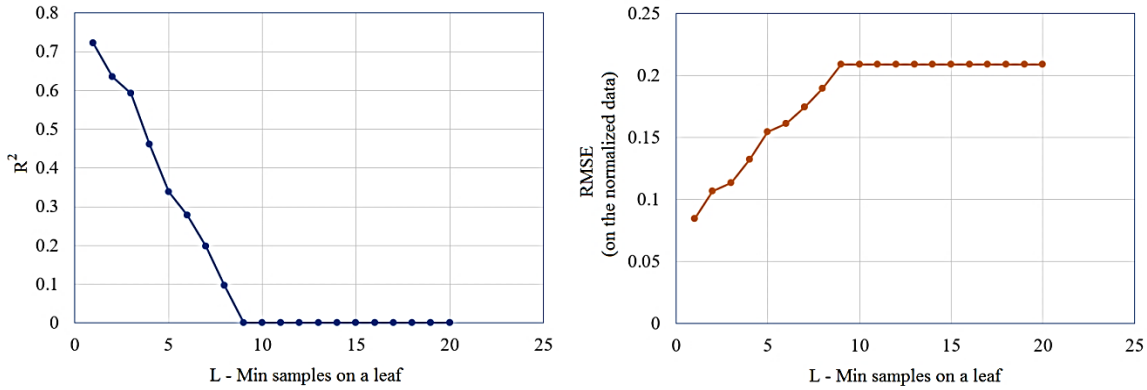

Figure 6. Effect of minimum leaf samples on model effectiveness.

The outcomes of a survey aimed at evaluating the accuracy of the model concerning the variation in the min samples on a leaf parameter (L) ranging from 1 to 20, with a fixed number of trees at 12, maximum tree depth at 7, and minimum samples required to split a tree at 6, are illustrated in Figure 6. The findings suggest that smaller L values yield superior prediction results, with an optimal value of 1, resulting in an R^2 of 0.721 and an RMSE of 0.085.

The results indicate that the random forest model, consisting of 12 trees, a minimum number of samples required to split a tree at 6, a maximum tree depth at 7, and a minimum number of samples required to form a leaf at 1, exhibits the highest level of accuracy. Consequently, this model will be selected for subsequent model validation and verification procedures.

5. Model validation

The performance of the Random Forest (RF) model is presented in this study for the training and validation sets, as depicted in Figure 7 and Figure 8. The results show that the RF model exhibits minimal deviation around the best fit line, indicating an agreement between the measured and predicted data. The model evaluation criteria also indicate good accuracy within the training dataset, with R squared values of 0.972 for both slump and specific strength, and RMSE values of 1.366 for slump and 1.274 for specific strength.

However, the RF model predicted some values beyond the 5% deviation line (Figure 7 and Figure 8), particularly for data that is outside the range of the training dataset. This phenomenon is consistent with the limitations of empirical models, as RF are better suited for interpolation than extrapolation. Some data points with the same output value but different input values may also result in confusion for the RF model, especially when the number of training data is insufficient (Figure 7). Therefore, increasing the size of the training dataset could improve the performance of the RF model.

Furthermore, the RF model shows better performance in predicting strength parameters than slump parameters, as evidenced in Figure 7 and Figure 8. Thus, developing an RF model to predict concrete strength is more practical than building an RF model for predicting slump in concrete.

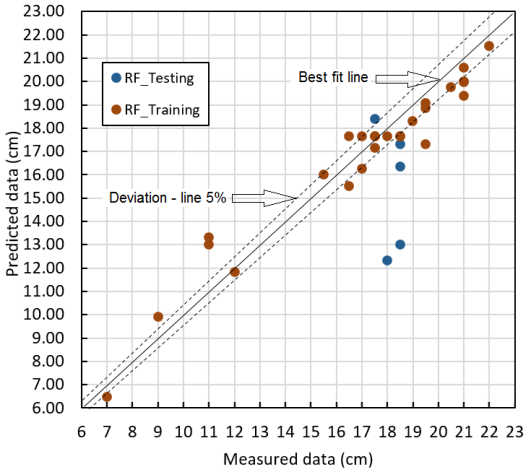

Figure 7. Scatterplots of predicted versus measured data for slump.

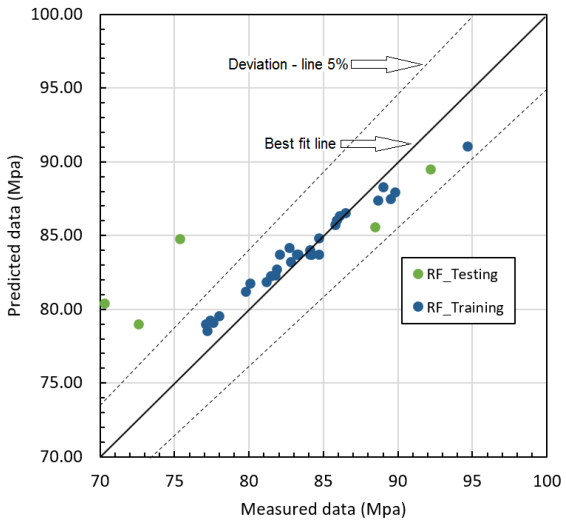

Figure 8. Scatterplots of predicted versus measured data for specific strength.

Table 4. Accuracy of RF model (testing set)

No.	Slump			Specific strength		
	Measured value (cm)	Predicted value	Deviation (%)	Measured value (MPa)	Predicted value (MPa)	Measured value (cm)
1	18.5	16.35	-11.62	92.2	89.47	-2.96
2	17.5	18.39	5.09	88.5	85.56	-3.33
3	18.5	13.00	-29.71	75.4	84.76	12.41
4	18.5	17.31	-6.44	72.6	78.98	8.79
5	18	12.32	-31.53	70.3	80.36	14.32

As the RF model utilizes a randomized feature selection process to construct decision trees, the significance of individual features is determined by their contribution to the model's error rate. Specifically, feature importance is measured by the percentage increase in Root Mean Squared Error (% increase in RMSE) when a given feature is omitted. This allows for the computation of an importance index, which ranges between 0 and 1, with the sum of all feature indexes being equal to 1. A higher index value reflects greater feature importance, indicating that the corresponding feature contributes more substantially to the predictive power of the RF model.

The findings from the feature importance analysis are summarized in Figure 9. As demonstrated, the Super-Plastic Additive/Adhesive Ratio (x4) was identified as the most critical input variable in constructing the RF model, with an importance score of 0.470.

Following this, the Silicafume/Adhesive Ratio (x3) was identified as the next most important variable, with an importance score of 0.411. These results indicate that the Super-Plastic Additive/Adhesive Ratio holds significant importance in making concrete samples.

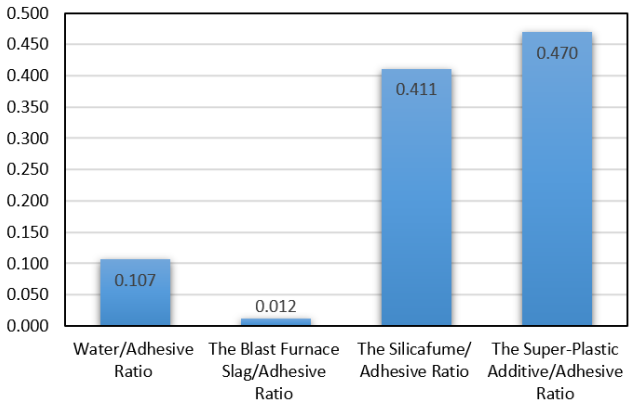

Figure 9. Features importance analysis.

6. Conclusion

After developing the RF model to predict the slump and strength of concrete, the following conclusions can be drawn:

- The results suggest that the RF model can be effectively employed for predicting both the slump and strength of concrete. However, it should be noted that the RF model performs better in predicting strength parameters than slump parameters. Consequently, developing an RF model to predict concrete strength is deemed more pragmatic than building an RF model for predicting concrete slumps.

- While the RF models exhibited high accuracy in predicting concrete properties, some of the predicted values displayed significant divergence from the measured values. This observation aligns with the understanding that, like all empirical models, RF perform optimally in interpolation tasks. Therefore, in order to enhance the performance of the RF model, it is imperative to expand the training data set to include more diverse and representative samples.

References

[1] Vu Van-tuan, "Prediction of settlement over time for road construction in Bac Ninh and Hai Duong province of VietNam using ANNs models," *Journal of Building Science and Technology*, Vol. 3, pp. 61-69, 2021.

[2] Güçlüer Kadir, Özbeyaz Abdurrahman, Göymen Samet, Günaydın Osman "A comparative investigation using machine learning methods for concrete compressive strength estimation," *Materials Today Communications*, 27, p. 102278, 2021.

[3] Huang Youqin, Fu Jiyang, "Review on application of artificial intelligence in civil engineering," *Computer Modeling in Engineering & Sciences*, 121(3), pp. 845-875, 2019.

[4] Rahul, Khandelwal Manoj, Rai Rajesh, Shrivastva B. K., "Evaluation of dump slope stability of a coal mine using artificial neural network," *Geomechanics and Geophysics for Geo-Energy and Geo-Resources*, 1(3), pp. 69-77, 2015. DOI: 10.1007/s40948-015-0009-8

[5] Shahin Mohamed A, "Artificial intelligence in geotechnical engineering: Applications, modeling aspects, and future directions," in *Metaheuristics in water, geotechnical and transport engineering*, Elsevier, pp. 169-204, 2013.

[6] Vu Van-tuan, "Research on the applicability of artificial neural network model to predict the average dimension of fragmentation and the volume of excavation for the electrical explosion model," *Journal of Science and Technique*, 4(207), pp. 25-36, 2020.

[7] Vũ Văn Tuấn, "Lựa chọn cấu trúc mạng nơ ron nhân tạo (ANN) dự báo chỉ số nén của đất," *Tạp chí Khoa học Công nghệ Xây dựng*, 3(2020), pp. 67-75, 2020.

[8] Vu Van-tuan, "Aritificial neural network (ANN) model in predicting multi -layered ground settlemets of metro tunnel," *Journal of Science and Technique*, 4(186), pp. 58-64, 2019.

[9] Breiman Leo, *"Random forests," Machine learning*, 45, pp. 5-32, 2001.

[10] Auret Lidia, Aldrich Chris, "Interpretation of nonlinear relationships between process variables by use of random forests," *Minerals Engineering*, 35 pp. 27-42, 2012. DOI: 10.1016/j.mineng.2012.05.008

[11] Svetnik Vladimir, Liaw Andy, Tong Christopher, Wang Ting, "Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules," in *Multiple Classifier Systems: 5th International Workshop, MCS 2004*, Cagliari, Italy, June 9-11, 2004, Proceedings 5, Springer.

[12] Maghrebi Mojtaba, Waller Travis, Sammut Claude, "Matching experts' decisions in concrete delivery dispatching centers by ensemble learning algorithms: Tactical level," *Automation in Construction*, 68, pp. 146-155, 2016. DOI: 10.1016/j.autcon.2016.03.007

[13] Mohamed Osama Ahmed, Ati Modafar, Najm Omar Fawwaz, "Predicting compressive strength of sustainable self-consolidating concrete using random forest," in *Key Engineering Materials*, Vol. 744, pp. 141-145, Trans Tech Publications, 2017. DOI: 10.4028/www.scientific.net/KEM.744.141

[14] Ozcan Giyasettin, Kocak Yilmaz, Gulbandilar Eyyup, "Estimation of compressive strength of BFS and WTRP blended cement mortars with machine learning models" *Comput. Concr.*, 19, pp. 275-282, 2017.

[15] Rao Weidong, *Application of Machine Learning in the Prediction of Compressive Strength of Concrete*, p. 102, 2017.

[16] Đinh Quang Trung, "Nghiên cứu sử dụng phụ gia khoáng hỗn hợp từ xỉ lò cao và silicafume chế tạo bê tông cường độ cao," *Tạp chí Khoa học và Kỹ thuật*, Học viện KTQS, Số 135 (7-2010), tr. 233-241, 2010.

SỬ DỤNG MÔ HÌNH RỪNG NGẪU NHIÊN TRONG DỰ BÁO ĐỘ SỤT VÀ CƯỜNG ĐỘ CỦA CÁC MẪU BÊ TÔNG CƯỜNG ĐỘ CAO

Vũ Văn Tuấn[a]

[a]*Trường Đại học Kỹ thuật Lê Quý Đôn*

Tóm tắt*: Mô hình rừng ngẫu nhiên - Random Forest (RF) đã được áp dụng thành công trong nhiều bài toán kỹ thuật do tính đơn giản, linh hoạt và sự phù hợp cho cả nhiệm vụ phân loại và hồi quy. Bê tông là một loại vật liệu xây dựng phức hợp, tính chất của nó bị chi phối bởi nhiều yếu tố, vì thế việc dự đoán các đặc tính của bê tông thường khó khăn. Trong bài báo này, tác giả nghiên cứu xây dựng một mô hình RF để dự đoán độ sụt và cường độ chịu nén của bê tông cường độ cao sử dụng phụ gia khoáng hỗn hợp từ xỉ lò cao và silicafume. Các tiêu chí để đánh giá độ chính xác của mô hình là R squared (R^2) và sai số trung bình bình phương (RMSE). So sánh dữ liệu dự đoán với dữ liệu thí nghiệm, kết quả cho thấy mô hình RF có thể được sử dụng để dự đoán cả hai tham số: độ sụt và cường độ chịu nén của bê tông.*

Từ khóa: Rừng ngẫu nhiên (RF); dự đoán; độ sụt của bê tông; cường độ của bê tông.

Received: 16/04/2023; *Revised:* 31/05/2023; *Accepted:* 23/06/2023; *Published:* 30/06/2023

❑