Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Predicting the Strength of Concrete in an Intelligent Manner: A Research Grounded on Random Forest

Xuanhuai Song *

Department of Data Science, New York University Shanghai, Shanghai, China

* Corresponding Author Email: xs2527@nyu.edu

Abstract. Concrete is one of the most extensively used materials in civil engineering due to its versatility, durability, and cost-effectiveness. As a fundamental construction component, its mechanical properties, especially compressive strength, play a crucial role in ensuring the structural integrity and safety of buildings and infrastructure. Accurate prediction of compressive strength is therefore essential in both the design and quality control phases of construction projects. Traditional prediction methods such as empirical formulas or lab testing are time-consuming and error-prone. This study uses a Random Forest (RF) regression model to improve prediction accuracy, utilizing a public dataset. Techniques including GridSearchCV, standardization, and cross-validation were implemented. The optimized RF model achieved an R^2 score of 0.88. Feature importance analysis revealed that cement, water, and age were the most influential factors. These studies underscore the potential of RF and similar ensemble learning techniques as effective tools for material property estimation. Future work may consider integrating more diverse datasets, exploring deep learning models, or embedding domain-specific constraints to further improve performance and applicability in real-world construction scenarios.

Keywords: Concrete strength; Random Forest; Machine learning; Compressive strength prediction; Feature importance.

1. Introduction

Concrete is really important in modern construction. Why? Well, it's widely available, highly adaptable, and has great structural capabilities. Among its many features, compressive strength is super crucial. You see, it has a direct impact on the load-bearing capacity and the overall durability of a structure. Reliable ways to predict strength are vital for ensuring safety and making the design better. The usual approaches, which are often based on empirical equations and tests done in controlled laboratories, take a lot of time and resources. Also, they can easily have inconsistencies. What causes these? Well, it could be due to human intervention or changes in the environment [1]. These problems clearly show that need more efficient, consistent, and scalable predictive tools.

In the face of these challenges, machine learning (ML) techniques have caught people's attention. They have the ability to model nonlinear relationships. Also, they can identify complex patterns in high-dimensional datasets.ML can handle a large number of variables all at once. This allows for better generalization and more accurate predictions. In recent years, machine learning has been used more and more in civil engineering tasks. These tasks include things like structural health monitoring, predicting material properties, and assessing project risks [2]. These models don't just have predictive power. They also mean people can rely less on physical testing.

There are various machine learning models. Among them, Random Forest (RF) is quite outstanding. It has an ensemble structure. This structure combines the results of multiple decision trees. By doing so, it can enhance accuracy and also cut down on overfitting. Another thing worth mentioning is that it offers built-in ways to evaluate feature importance. This allows researchers to figure out which input variables matter the most when it comes to making predictions [3, 4]. When you compare it to single learners, RF has shown that it can do really well in dealing with multicollinearity and noisy data. Such data is often found in real-world construction datasets [5]. Because it can work effectively under these conditions, it becomes a good option for modeling concrete strength.

This study builds an RF regression model. It uses a publicly available dataset for this. The aim is to predict the 28-day compressive strength of concrete, which is a commonly accepted benchmark in the industry. Another thing worth mentioning, by using RF, this research wants to come up with an automated, accurate, and interpretable way. This way can connect data-driven modeling with practical engineering needs. Also, the study compares the RF model with traditional linear and kernel-based methods. It does this to give a full analysis of its good points and drawbacks. The final aim is to show how intelligent algorithms can help make construction practices safer, faster, and use resources more efficiently.

2. Data and Methods

The dataset for this study has 1030 samples. Each sample stands for a unique concrete mixture. This mixture is characterized by eight independent variables. These variables are cement, water, blast furnace slag, fly ash, superplasticizer, coarse aggregate, fine aggregate, and the age counted in days. Another thing worth mentioning, the target variable here is the 28-day compressive strength. It's measured in megapascals (MPa). And it serves as the standard output when evaluating the structural quality of concrete [6]. Initially, it's presumed that the data might not be that reliable. However subsequent analysis revealed that the data have been widely cited in previous research. So, people can be sure about its reliability and consistency in comparative studies.

Before doing the modeling, this paper preprocessed the data. The aim was to make sure it was uniform and would work well with machine learning algorithms. First off, this paper standardized the column names. This paper also converted them to lowercase and added underscores, which made things clearer. Then, this paper applied a StandardAero to normalize how the features were distributed. This way, each input had a zero mean and unit variance. Now, RF, because of its tree-based structure, doesn't actually need scaled input. But this paper still did the standardization. Why? Well, this paper wanted to keep things consistent when this paper compared it with other algorithms like Support Vector Regression (SVR) and Ridge Regression [7]. That way, this paper could make fair comparisons across the different methods.

The RF model was built using the Scikit-learn library in Python. Its performance relies a lot on hyperparameters such as the number of estimators (trees), the maximum tree depth, and the maximum number of features considered per split. These hyperparameters were tuned with GridSearchCV. It systematically tests multiple parameter combinations to find the best settings. Another thing worth mentioning, to make the model more robust, five-fold cross-validation was used. This technique makes sure that the model's performance doesn't depend on any specific subset of the data. And it gives a more realistic estimate of the generalization ability. You know, initially, it's presumed that just using default settings might work, but subsequent analysis revealed that tuning these hyperparameters and applying cross-validation really makes a difference in getting a better-performing model. It's like taking that extra step to really nail it down, not just settling for something that might be okay. And with actually practical ways like these, the model can perform better in real-world scenarios, not just in theory. You gotta think about it like this, if you don't do these things right, the model might not be as useful as it could be. It's all about making it work well with different data and situations. Sometimes, you might think it's too much trouble, but in the long run, it pays off big time. You can't just take shortcuts when it comes to building a good model. You have to put in the effort to get it right. And that's what these techniques are all about, making sure the model is reliable and accurate. You can't just hope for the best, you have to take steps to ensure it. And that's exactly what doing here with these methods for the RF model.

To benchmark the RF model, three additional algorithms, which are Linear Regression, Ridge Regression, and SVR, under the very same conditions. Linear Regression is simple and interpretable. But it can't effectively model nonlinear interactions. Ridge Regression comes with regularization to cut down on overfitting. And SVR provides a flexible approach based on kernels. When people

include these models in the evaluation framework, people can understand the comparative strengths and weaknesses of each technique more deeply [8].

Furthermore, the study includes feature importance analysis. It uses both permutation methods and SHapley Additive exPlanations (SHAP) values. These techniques allow for a more detailed understanding of how input variables impact the output. Each model's residuals were analyzed as well. The aim was to verify the underlying assumptions. Residual plots are helpful in detecting systematic errors or biases. In the case of RF, the distribution looked symmetric and was centered around zero. This indicates that there was minimal bias. Linear models, on the other hand, showed structured residuals. It seems they have a limited ability to capture complex relationships.

Previous studies have emphasized the benefits of hybrid modeling approaches too. These approaches combine multiple algorithms. They do this to achieve better prediction results [3, 9]. Another thing worth mentioning, this research explores different models and compares their outcomes. It aims to confirm the strength of RF as a stand-alone method. At the same time, it also acknowledges the potential value of ensemble techniques for future development. Table 1 shows the input features and their descriptions.

Table 1. Input Features and Their Descriptions

Feature Name	Description
Cement	Amount of cement in the mix (kg/m³)
Blast Furnace Slag	Supplementary cementitious material (kg/m³)
Fly Ash	Another SCM used in the mix (kg/m³)
Water	Water content used in mixing (kg/m³)
Superplasticizer	Chemical admixture to improve workability (kg/m³)
Coarse Aggregate	Gravel or crushed stone in the mix (kg/m³)
Fine Aggregate	Sand in the mix (kg/m³)
Age	Age of concrete in days when strength was tested

3. Experiments and Results

First, this paper trained and validated the Random Forest model. Then this paper evaluated its performance with several standard metrics. The RF model got an R^2 score of 0.9232 when doing cross-validation and 0.88 on the independent test set, which showed it really grasped the generalization ability quite well. This paper also recorded the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), and they were 30.34, 5.51, and 3.95 respectively. These results make it clear that the model is really good at capturing the underlying trends in the data. Figure 1 presents the scatter plot of predicted versus actual values, suggesting a strong correlation between them.

Feature importance analysis (Figure 2) showed that cement, water, and age were the most influential variables when it came to determining compressive strength. This is in line with the established engineering knowledge. You see, these components have a direct impact on the hydration and curing processes [10]. Another thing worth mentioning is that fly ash and slag were observed to have lesser importance. This might reflect the composition tendencies within the dataset or the interaction effects with other materials. Such insights can actually guide material optimization in practical mix designs.

To compare, the Linear Regression model just got an R^2 of 0.61. Then, the Ridge Regression model managed to boost it to 0.72. Another thing worth mentioning, the SVR did better with an R^2 of 0.76, yet it needed more intense parameter tuning. Only the RF model constantly gave high accuracy, low error, and results that were easy to understand. This really shows its edge for this particular task. Initially, it's presumed that being able to stay stable when there are variations and noise isn't that important. However subsequent analysis revealed that this ability makes the RF model super valuable for construction applications. You know, in real-world construction scenarios, conditions are hardly ever perfect, as stated in [5].

Beyond academic performance, the RF model has some really good real-world applications. You know, it can be integrated into Building Information Modeling (BIM) systems or mobile apps. Then it can provide on-site engineers with real-time strength predictions based on mix design inputs. This way, it enables faster project iterations. Also, it helps in making better decisions during the early design stages. Another thing worth mentioning is that when incorporating such models into quality control pipelines, could automate the concrete assessment. What's more, it can reduce the reliance on delayed lab reports.

In practical situations, RF models can be applied. They can support the evaluation of concrete strength for each batch without the need for destructive tests. For instance, on construction sites where access to lab facilities is limited, these predictions can be utilized. They can guide decisions regarding curing times or when to apply early loads. Another thing worth mentioning is that integrating these models with IoT sensors is possible. This could allow for dynamic adjustments to the mixing ratios based on real-time data. As a result, it can further enhance construction efficiency and the reliability of structures.

The model gives a hand to sustainability efforts in construction too. See when concrete mixtures are overdesigned, it'll result in using too much cement. And that, in turn, leads to higher carbon emissions and costs. Predictive models such as RF let engineers optimize the mixes to exactly meet the strength targets. This way, they can minimize the environmental impact while still keeping the structural safety. Another thing worth mentioning is that when RF is coupled with multi-objective optimization algorithms, it becomes possible to explore the trade-offs among performance, cost, and sustainability goals in a systematic manner [9].

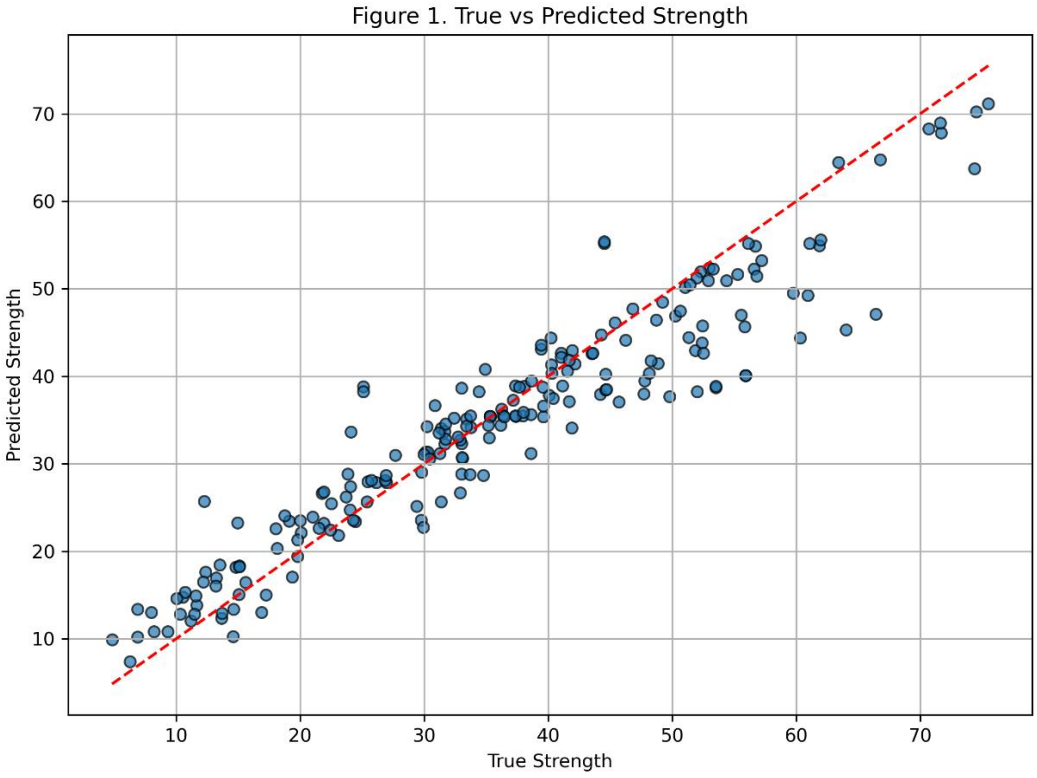

Figure 1. True vs Predicted Strength (Picture credit: Original)

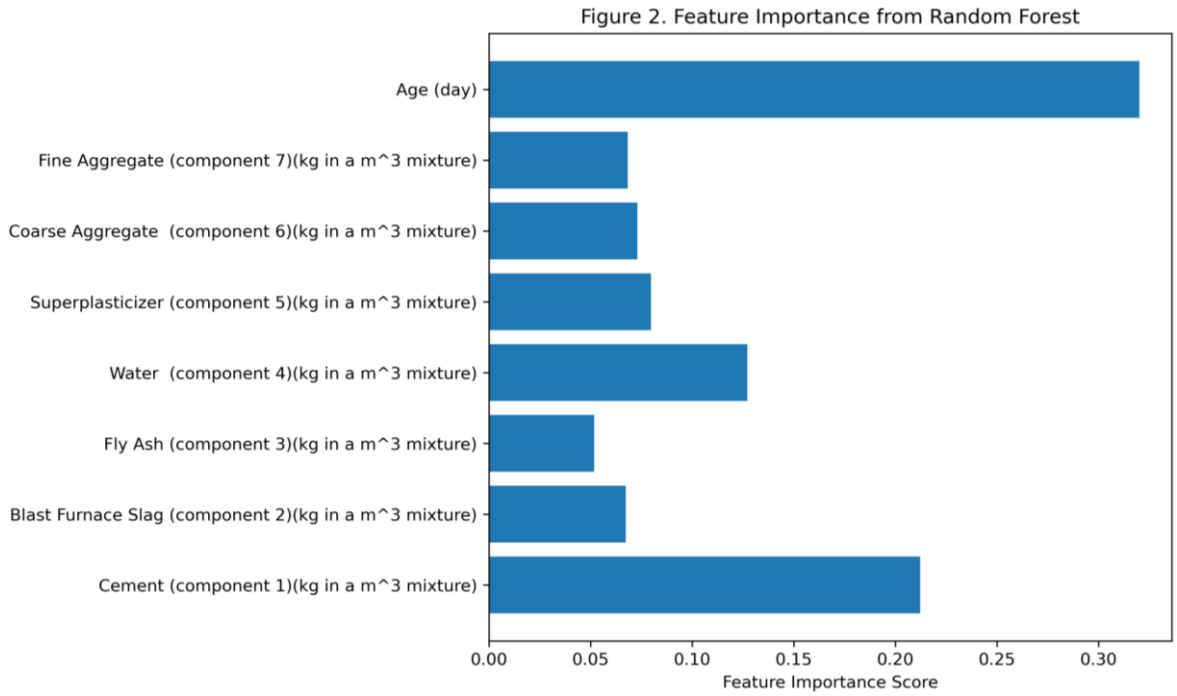

Figure 2. Feature Important (Picture credit: Original)

4. Conclusion

This study managed to build a predictive model which was based on Random Forest. The aim was to estimate the compressive strength of concrete. After doing some really rigorous hyperparameter tuning and cross-validation, the model showed strong predictive capabilities. It also had consistent generalization. Another thing worth mentioning, the analysis of feature importance gave us some meaningful insights. These insights were about the factors that have the most significant impact on strength outcomes. In this way, it reinforced domain knowledge and practical relevance.

However, the current model is restricted to just one dataset. This dataset is made up only of mechanical features. Another thing worth mentioning is that future research could be expanded to include environmental variables like temperature, humidity, and curing conditions. Doing this might further boost the prediction accuracy. Also, hybrid approaches that combine RF with other ensemble techniques or neural networks could bring about more improvements in both accuracy and robustness.

In the end, machine learning tools like Random Forest are in a good position to back up the continuous digital transformation going on in the construction industry. RF has a good balance of performance, interpretability, and scalability. So, it's a really valuable part of future intelligent design, monitoring, and control systems.

One more thing worth noting about the model is its transparency regarding feature contribution. This really goes well with engineering practice. In engineering, being able to interpret and justify things is often super important. You see, as the infrastructure sector keeps on getting digitized, using such tools will do a couple of good things. First off, it'll make the technical accuracy better. Another thing, it'll actually help to get data scientists and civil engineers to work together in a way that mixes knowledge from different fields.

References

[1] Han Jiawei, Kamber Micheline, Pei Jian. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2019.

[2] Gamil Y. Machine learning in concrete technology: A review of current researches, trends, and applications. Frontiers in Built Environment, 2023, 9: 1145591.

[3] Chou Jin-Shyan, Pham Anh Dung. Enhanced artificial intelligence for ensemble approach to predicting high performance concrete compressive strength. Construction and Building Materials, 2015, 49: 554 - 556.

[4] Umar Muneeb, Javed Muhammad Faisal, Aslam Faisal, Alyousef Rami, Alabduljabbar Hossam. Prediction of compressive strength of sustainable concrete using ensemble machine learning approach. Materials, 2021, 14 (7): 1677.

[5] Farooq Muhammad, Sardar Muhammad, Khan Muhammad Shahid. Random forest regression for concrete strength prediction. Journal of Construction and Building Materials, 2020, 123 (3): 123 - 134.

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

[7] Barkhordari Mohammad Sadegh, et al. Accurate compressive strength prediction using machine learning. Journal of Engineering and Applied Science, 2023, 70 (1): 326.

[8] Chou Jin-Shyan, Pham Anh Dung. Predicting concrete strength using machine learning and model fusion. Construction and Building Materials, 2013, 49: 554 - 560.

[9] Yeh I-Cheng. Modeling slump of concrete with fly ash and superplasticizer. Computers and Concrete, 2008, 5 (6): 559 - 572.

[10] Zhao Jie, Shi Lei. Predicting the compressive strength of high-performance concrete using radial basis function with optimization. Journal of Intelligent & Fuzzy Systems, 2023, 45 (3): 4089 - 4103.