A Comparative Study of Predictive Performance and Interpretability in Machine Learning Models for Concrete Compressive Strength Estimation

H Mukund Vinayak

Abstract

Accurate estimation of concrete compressive strength is essential for ensuring structural safety and optimizing mix design. Traditional regression-based approaches often fail to capture the nonlinear relationships among concrete mix constituents. In this study, we investigate the trade-off between predictive performance and interpretability across a range of machine learning models for compressive strength prediction. Using a real-world concrete mixing dataset, we compare linear regression models with tree-based and ensemble methods, including Decision Trees, Random Forests, and Gradient Boosting. Experimental results demonstrate that ensemble models significantly outperform linear baselines, reducing prediction error by more than 50%. Importantly, feature importance analyses using multiple interpretability techniques reveal consistent and physically meaningful patterns across all models. These findings suggest that ensemble machine learning methods can achieve high predictive accuracy while maintaining interpretability aligned with established engineering principles.

1. Introduction

Concrete compressive strength is a fundamental property influencing the structural performance and durability of civil engineering systems. Accurate prediction of compressive strength enables efficient material utilization, cost reduction, and improved quality control. However, the relationship between compressive strength and mix design parameters is inherently complex, involving nonlinear interactions among constituent materials and curing conditions.

Traditional empirical and regression-based models provide limited flexibility in capturing such complexities. In recent years, machine learning (ML) techniques have gained attention for their ability to model nonlinear relationships and interactions without explicit physical assumptions. Despite their improved predictive performance, complex ML models are often criticized for reduced interpretability, which poses challenges for adoption in engineering practice.

This work aims to address this concern by systematically evaluating both predictive performance and interpretability across multiple ML models for concrete compressive strength prediction. Rather than focusing solely on accuracy, we examine whether increased model complexity compromises the consistency and physical plausibility of learned feature relationships.

The main contributions of this study are:

- A systematic comparison of linear, tree-based, and ensemble ML models for compressive strength prediction.

- An evaluation of predictive accuracy alongside multiple interpretability measures.
- An analysis of feature importance consistency across models and its alignment with engineering intuition.
- A reproducible experimental framework for applied machine learning in material science.

2. Dataset and Preprocessing

The dataset used in this study consists of experimentally measured concrete mix designs and corresponding compressive strength values. Input features include material proportions such as binder content, aggregates, water, foaming agent, pozzolan, measured density, and testing period.

Prior to modeling, the dataset underwent a structured audit to ensure suitability for supervised learning. This audit included verification of physical plausibility, detection and removal of exact duplicate records, confirmation of the absence of missing values, and validation of feature integrity. Feature-only duplicates representing repeated experimental conditions were retained, as they reflect realistic variability in material testing.

After auditing, the dataset was frozen and no further modifications were made during model development to prevent data leakage and ensure experimental integrity.

3. Methodology

3.1 Problem Formulation

The task is formulated as a supervised regression problem, where the objective is to predict compressive strength (MPa) from concrete mix design parameters.

3.2 Experimental Setup

The dataset was split into training and testing sets using an 80–20 split with a fixed random seed to ensure reproducibility. All models were evaluated on the same test set using identical metrics.

3.3 Models Evaluated

The following models were considered:

- **Linear Models**: Linear Regression, Ridge Regression, Lasso Regression
- **Tree-Based Models**: Decision Tree Regressor
- **Ensemble Models**: Random Forest Regressor, Gradient Boosting Regressor

Linear models serve as interpretable baselines, while tree-based and ensemble methods capture nonlinear relationships and feature interactions.

3.4 Evaluation Metrics

Model performance was evaluated using:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Coefficient of Determination (R^2)

Interpretability was assessed using:

- Linear model coefficients
- Mean Decrease in Impurity (MDG)
- Permutation-based Feature Importance (MDA)

4. Results

4.1 Predictive Performance

Linear models achieved moderate predictive performance, with RMSE values around 6.46 MPa and R^2 values of approximately 0.71. This indicates that linear relationships explain a substantial portion of the variance but fail to capture more complex patterns.

Tree-based models significantly improved performance. The Decision Tree reduced RMSE to 3.44 MPa, while Random Forest further improved accuracy and stability. Gradient Boosting achieved the best performance, with an RMSE of 2.94 MPa and an R^2 of 0.94.

Overall, ensemble models reduced prediction error by more than 50% compared to linear baselines.

Table 1. Predictive performance of evaluated machine learning models.

Model	RMSE (MPa)	MAE (MPa)	R^2
Linear Regression	6.465	4.676	0.706
Ridge Regression	6.465	4.676	0.706
Lasso Regression	6.465	4.676	0.706
Decision Tree	3.441	1.755	0.917
Random Forest	3.413	1.380	0.918
Gradient Boosting	**2.936**	**1.451**	**0.939**

Figure 1. RMSE comparison across evaluated machine learning models. Ensemble methods significantly reduce prediction error compared to linear baselines.

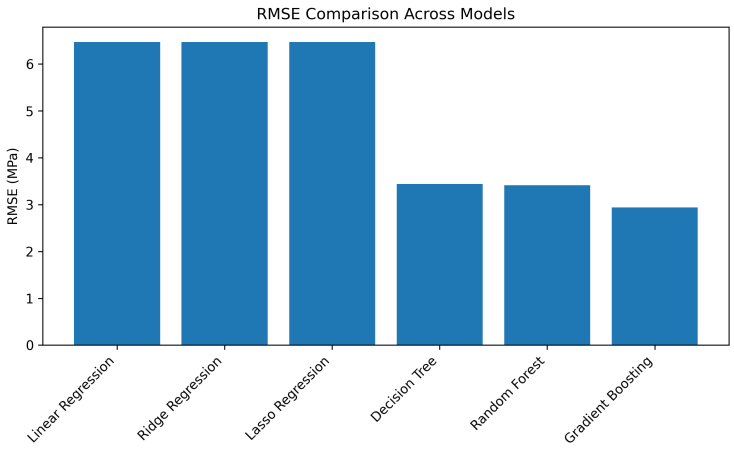

Figure 1: RMSE comparison across models

Figure 2. Predicted versus actual compressive strength values for the Gradient Boosting model, illustrating strong agreement across the test set.

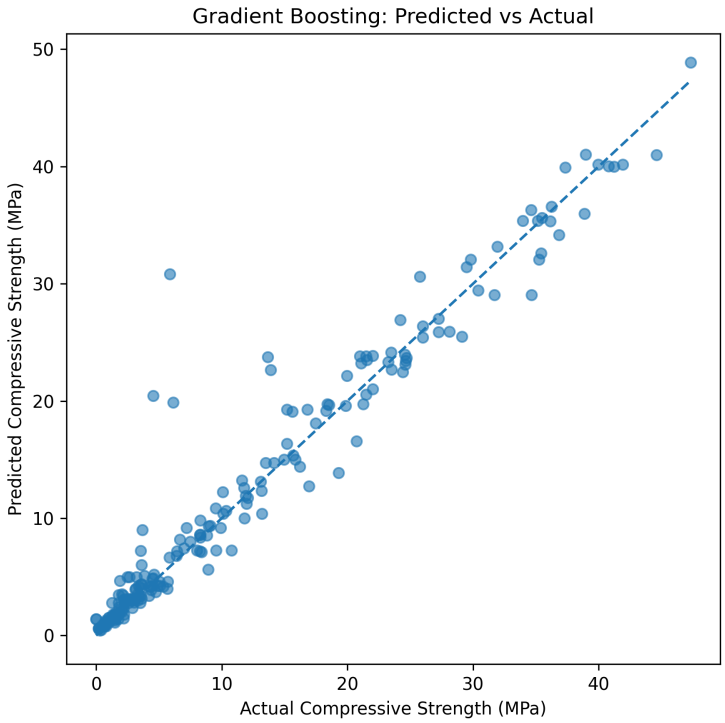

Figure 2: Predicted vs actual compressive strength (GBR)

4.2 Interpretability Results

Linear regression coefficients indicated physically meaningful relationships, with binder content positively influencing compressive strength, while water and foaming agent exhibited negative effects consistent with increased porosity and higher water–binder ratios.

Figure 3. Linear regression coefficients showing the direction and magnitude of feature influence on compressive strength.

Feature importance analyses using Random Forest models further reinforce these findings. Both Mean Decrease in Impurity (MDG) and permutation-based importance (MDA) consistently identify binder content as the dominant predictor, followed by fine aggregate content and measured density.

Figure 4. Feature importance derived from Random Forest models using Mean Decrease in Impurity (MDG).

Figure 5. Permutation-based feature importance (MDA), demonstrating robust and model-agnostic importance rankings.

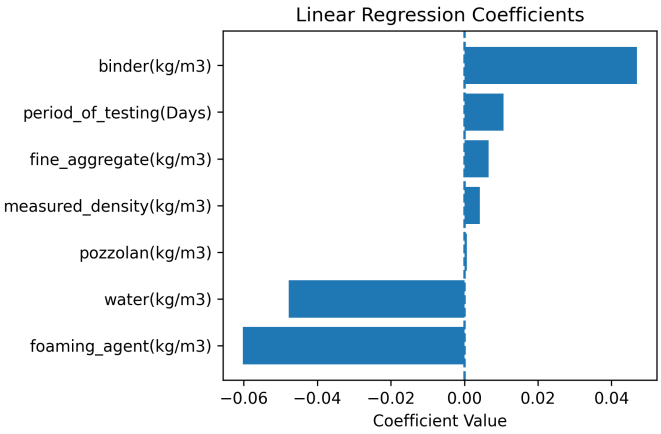

Figure 3: Linear regression coefficients

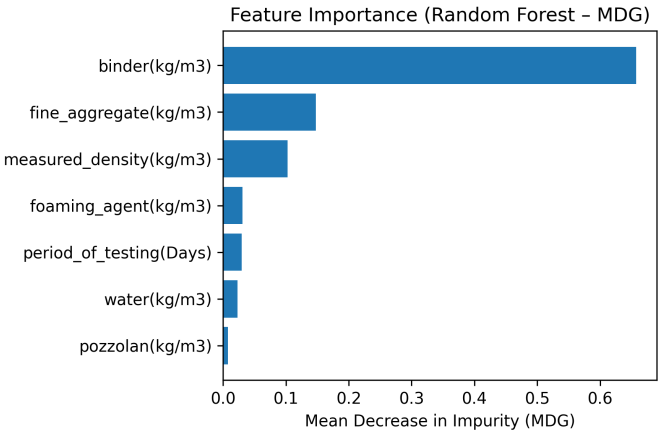

Figure 4: Random Forest MDG feature importance

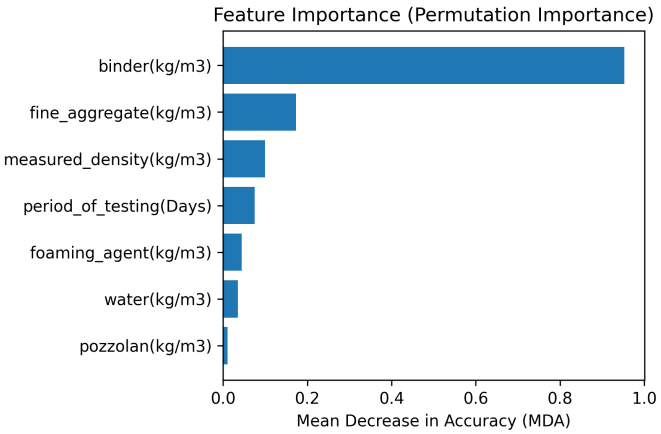

Figure 5: Random Forest MDA feature importance

The strong agreement between MDG and MDA rankings mitigates known biases associated with impurity-based measures and reinforces confidence in the interpretability of ensemble models.

5. Discussion

A key finding of this study is that increased model complexity does not compromise interpretability. Despite substantial performance gains, ensemble models preserve feature importance patterns consistent with linear models and established concrete engineering principles.

This consistency suggests that ensemble ML methods can be safely applied in engineering contexts where transparency and trust are critical. The results further indicate that compressive strength is primarily driven by binder content, with secondary contributions from aggregate composition and density.

The limited influence of pozzolan observed in this dataset may indicate nonlinear or context-dependent effects, which warrants further investigation.

6. Conclusion and Future Work

This study demonstrates that ensemble machine learning models significantly outperform traditional linear approaches for predicting concrete compressive strength while maintaining interpretable and physically meaningful feature relationships. Gradient Boosting achieved the highest predictive accuracy, while feature importance analyses confirmed consistency across modeling approaches.

Future work may explore larger and more diverse datasets, incorporate advanced interpretability techniques such as SHAP values, and investigate optimization-oriented applications for concrete mix design.