

Predicting Concrete Compressive Strength using Machine Learning

Osama Ibrahim

Department of Civil and Environmental
Engineering
University of Illinois Urbana-
Champaign
Urbana, IL, USA
osamani2@illinois.edu

Praneeth Shivashankarappa

Department of Civil and Environmental
Engineering
University of Illinois Urbana-
Champaign
Urbana, IL, USA
ps104@illinois.edu

Kazi Ishat Mushfiq

Department of Civil and Environmental
Engineering
University of Illinois Urbana-
Champaign
Urbana, IL, USA
mushfiq2@illinois.edu

Georg Bauer

Department of Civil and Environmental Engineering
University of Illinois Urbana-Champaign
Urbana, IL, USA
georgb2@illinois.edu

Abstract—This project investigates the prediction of concrete compressive strength using machine learning models. We use the Concrete Compressive Strength dataset from the UCI Machine Learning Repository, which contains 1030 experimental data points with eight input variables (cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and curing age) and one output variable (compressive strength in MPa). Our study compares linear regression, decision tree regression, and neural networks to determine the most effective approach for modeling the nonlinear relationships that influence the development of concrete strength.

Index Terms—Concrete Compressive Strength, Machine Learning, Civil Engineering, Regression Models

I. PART (1): PROJECT SELECTION AND PROPOSAL

A. Dataset Description

The Concrete Compressive Strength dataset originates from laboratory experiments conducted by Prof. I-Cheng Yeh, who donated it to the UCI Machine Learning Repository in 2007 to support research on high-performance concrete [1], [2]. The dataset records the quantities of concrete mix ingredients and the curing age, together with the corresponding compressive strength as the output variable. In total, it includes 1030 samples in a CSV file, with each row representing one concrete mix. All ingredient quantities are reported per cubic meter of concrete, and the compressive strength is expressed as a continuous variable in megapascals (MPa), representing the material's capacity to withstand compressive loads.

The dataset has no missing values and is widely recognized in the civil engineering community as a benchmark for modeling concrete strength. For this project, we will obtain the dataset directly from the UCI Machine Learning Repository [2], ensuring a reliable source. The variables included are listed below.

Variables included in the dataset:

- Cement (kg/m^3)
- Blast Furnace Slag (kg/m^3)
- Fly Ash (kg/m^3)
- Water (kg/m^3)
- Superplasticizer (kg/m^3)
- Coarse Aggregate (kg/m^3)
- Fine Aggregate (kg/m^3)
- Age (days, from 1–365)

Target variable:

- Concrete Compressive Strength (MPa)

B. Sample Data Preview

Table I provides representative entries from the Concrete Compressive Strength dataset, illustrating the input features and the output variable. Each row corresponds to one concrete mix, with ingredient quantities and age shown alongside the measured compressive strength in MPa.

TABLE I
REPRESENTATIVE ENTRIES FROM THE CONCRETE COMPRESSIVE STRENGTH DATASET

Row	Cement	Slag	Fly-Ash	Water	SuperP.	CA	FA	Age	Strength
1	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
2	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
...
...
1028	148.5	139.4	108.6	192.7	6.1	892.4	780.0	28	23.70
1029	159.1	186.7	0.0	175.6	11.3	989.6	788.9	28	32.77
1030	260.9	100.5	78.3	200.6	8.6	864.5	761.5	28	32.40

C. Project Proposal

a) *Planned Approach*: In this project, the Concrete Compressive Strength dataset will be analyzed from a data science perspective. Techniques learned in CEE 492 will be applied to perform exploratory data analysis, develop machine learning models, and investigate how each ingredient in the dataset influences the resulting compressive strength. This begins with ensuring the dataset is properly structured for analysis, followed by looking at the distributions of the variables and their relationships. We also plan to explore whether Principal Component Analysis provides useful insights into correlations among the input variables.

After the exploratory phase, we will develop predictive models for compressive strength. Multiple linear regression will serve as the baseline, against which we will compare more advanced approaches such as decision tree models and neural networks. Finally, we will test the trained models on the dataset and check how well they can predict compressive strength, using common accuracy measures.

b) *Relevance*: Compressive strength is one of the most important properties of a concrete mix, and being able to estimate it accurately is essential in civil engineering, as it can help engineers decide whether a given mixture will meet the required strength for a project. Laboratory testing is the standard approach, but it can be time-consuming and resource-intensive. Empirical equations from design codes can be used as an alternative, but they often lack accuracy and flexibility across different mix proportions and curing conditions, since concrete compressive strength is a highly nonlinear function of both age and ingredient proportions. This creates a need for faster, more reliable methods of prediction.

This project addresses that need by using machine learning to test how well different models can predict compressive strength. A good model can give faster and more consistent estimates of concrete strength for different mix proportions and curing ages, making it a useful tool in engineering practice. The project will also show which ingredients are most important and how curing age affects strength, helping to better understand the factors that influence concrete performance.

c) *Deliverables*: The deliverables include an exploratory analysis of the dataset, the development and testing of different models to predict compressive strength, a comparison of their performance, and a discussion of the results to show how mix design and curing age influence concrete strength.

II. PART (2): EXPLORATORY DATA ANALYSIS AND PREDICTIVE MODELING PLAN

A. Exploratory Data Analysis

a) *Dataset Overview*: The Concrete Compressive Strength dataset contains 1030 samples with 8 input features and one output variable. As shown in Table (I). And a comprehensive description of data was mentioned earlier in part (1-A) of the study.

b) *Summary Statistics*: In this part of the exploratory data analysis (EDA), a statistical analysis is conducted and summarized in Table (II). The dataset contains key input variables that influence the compressive strength of concrete, including the quantities of cement, supplementary cementitious materials (blast furnace slag and fly ash), water, superplasticizer, and aggregates, as well as the curing age. Table (II) summarizes the basic descriptive statistics, minimum, maximum, mean, median, and standard deviation, for each feature.

These statistics serve as a solid basis to examine the distribution and variability of the dataset. For example, a wide range of cement content (102–540 kg/m³) as well as testing age (1–365 days) can be observed, which reflects the diversity in the dataset in terms of mix design and curing duration. This variability is beneficial for broader performance forecasting and for reliable model learning that is not limited to a specific small-range dataset. Also, the standard deviations show the heterogeneity among samples, which reflects potential nonlinear relationships. This type of data overview is required as a starting point in any EDA, as it helps to assess data variability, identify outliers, and understand variable correlations. These findings will enhance future analyses and support the model-building process.

TABLE II
SUMMARY STATISTICS OF CONCRETE MIX FEATURES

Feature	Min	Max	Mean	Median	Std-Dev
Cement (kg/m ³)	102.0	540.0	281.17	272.9	104.51
Blast Furnace Slag (kg/m ³)	0.0	359.4	73.9	22.0	86.28
Fly Ash (kg/m ³)	0.0	200.1	54.19	0.0	64.0
Water (kg/m ³)	121.75	247.0	181.57	185.0	21.36
Superplasticizer (kg/m ³)	0.0	32.2	6.2	6.35	5.97
Coarse Aggregate (kg/m ³)	801.0	1145.0	972.92	968.0	77.75
Fine Aggregate (kg/m ³)	594.0	992.6	773.58	779.51	80.18
Age (day)	1	365	45.66	28.0	63.17

c) *Data Visualizations*: Data visualization is critical to understanding the relationships and trends within the Concrete Compressive Strength dataset. Before building predictive models, it is important to visually explore how each mix component affects the resulting strength. Visual analysis provides intuitive insight into variable distributions, correlations, and potential nonlinearities that may not be evident from raw statistics alone. In this part, several visualization methods, including histograms, scatter plots, correlation heatmaps, boxplots, and Principal Component Analysis (PCA), are used to reveal patterns in the data. These plots help identify dominant features such as the effects of cement and water content, highlight zero-inflated variables like slag and fly ash, and detect outliers or skewness in ingredient proportions. This visual exploration forms the foundation for informed feature selection and helps guide the choice of suitable machine-learning models for predicting compress-

sive strength. Figure 1 shows the histograms of the input features. The X-axis represents the dosage (mostly in kg/m^3 or days for age), and the Y-axis represents the frequency. Most mix designs use a cement dosage between 150–350 kg/m^3 , while the frequency decreases for dosages above 400 kg/m^3 . The majority of mixes contain no blast furnace slag, as indicated by the sharp spike at zero in its histogram, a clear sign of a zero-inflated distribution. This zero-inflated trend also appears for superplasticizer and fly ash, suggesting that most mix designs omit these additives. The behavior of the dataset becomes clearer when water and cement dosages are considered together. The water dosage histogram shows a roughly bell-shaped distribution centered around 170–200 kg/m^3 , while cement dosage varies more widely between 150–350 kg/m^3 . This indicates that most mix designs rely on cement as the primary binder rather than supplementary materials such as fly ash or slag. The water dosage also appears to complement the reduced use of superplasticizers. The coarse aggregate distribution is multimodal, reflecting variations in mix designs and aggregate-to-cement ratios. Differences in maximum aggregate size may also influence the distribution, as larger aggregates lower packing density and reduce required volume. High aggregate values (above 800 kg/m^3) likely correspond to structural concrete rather than mortar or grout mixes. Finally, the histogram of testing ages shows a tall, dominant peak at early ages (1–7 days), indicating that most compressive strength tests were performed at early stages. The frequency decreases after about 28 days, the standard reporting age for concrete strength tests.

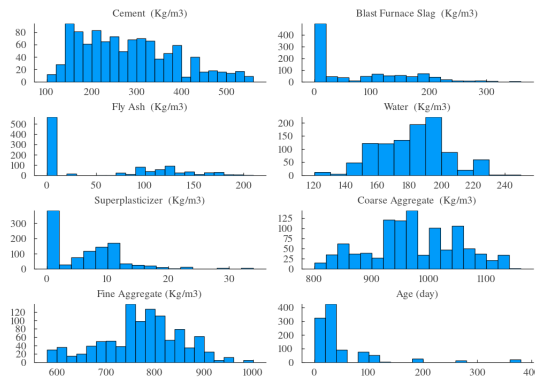


Fig. 1. Histograms of input features

The scatter plots analyzing the influence of the input features on the compressive strength are shown in Figure 2. The strength increases with the increase in cement content. As cement is the primary binder, increasing cement dosage improves the compressive strength. However, beyond an optimum level (around 400 kg/m^3), excessive cement can reduce the compressive strength due to shrinkage and poor workability. The influence of blast furnace slag is scattered with no consistent trend. As slag acts as supplementary material, its effect on the compressive strength depends on the replacement ratio and curing age. The strength gain for slag is slow initially, but it can improve later due to pozzolonic activity. Most of the mixes have zero fly ash amount,

while others show variable strengths for 50 - 150 kg/m^3 . A decreasing trend in the compressive strength is evident for the higher water dosage. Higher water content increases the workability, but it reduces the compressive strength due to greater porosity after hydration. The superplasticizer dosage is concentrated at a low dosage of around 0–15 kg/m^3 . The superplasticizers improve flowability with a lower water-cement ratio and lead to higher strength. However, a higher amount of superplasticizers shows inconsistent effects. The coarse aggregate plot is scattered, and the strength variation is not significant. The fine aggregate data is mostly between 700–900 kg/m^3 . A higher amount of fine aggregates increases water demand and reduces strength. The scatter plot of the age of the test shows a clear increase in the compressive strength with the increase in the testing days, especially up to 90 days.

The scatter plots illustrating the influence of input features on compressive strength are presented in Figure 2. The results show that compressive strength generally increases with higher cement content, as cement acts as the primary binder. However, beyond approximately 400 kg/m^3 , excessive cement may lead to strength reduction due to shrinkage and reduced workability [3]. The effect of blast furnace slag appears scattered without a clear trend. Since slag serves as a supplementary material, its influence depends on both the replacement ratio and curing duration, strength typically develops more slowly but can improve later due to pozzolanic reactions. Most mixtures contain no fly ash, while those with 50–150 kg/m^3 display variable strength outcomes. A clear decreasing trend is observed for higher water dosages. While increased water enhances workability, it also leads to greater porosity after hydration, thus lowering strength. Superplasticizer content is mostly concentrated between 0–15 kg/m^3 . Within this range, it improves flowability and strength by reducing the water-to-cement ratio, though higher dosages produce inconsistent effects. The coarse aggregate plot is widely scattered, suggesting minimal influence on strength variation. Fine aggregate content, primarily between 700–900 kg/m^3 , shows that higher proportions increase water demand and slightly reduce strength. Finally, the scatter plot for curing age indicates a clear upward trend in compressive strength with increasing testing age, especially up to about 90 days, after which strength gains become less pronounced.

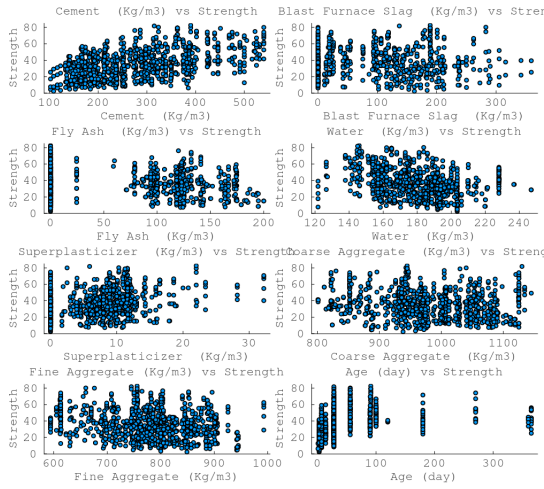


Fig. 2. Scatter plots of input features vs compressive strength

The correlation heatmap in Figure 3 shows that compressive strength is most positively associated with cement content and testing age, and negatively associated with water content. This finding is consistent with the classical water-to-binder effect, where excess water weakens the concrete matrix [4]. Superplasticizers show a slight positive correlation with compressive strength, as they reduce water demand and lower the water-to-binder ratio, thereby improving strength. The aggregate contents show weak correlations with compressive strength, while fine and coarse aggregates are negatively correlated with each other due to volumetric trade-off. Fly ash and slag are negatively correlated with cement content, reflecting their role as partial cement replacements.

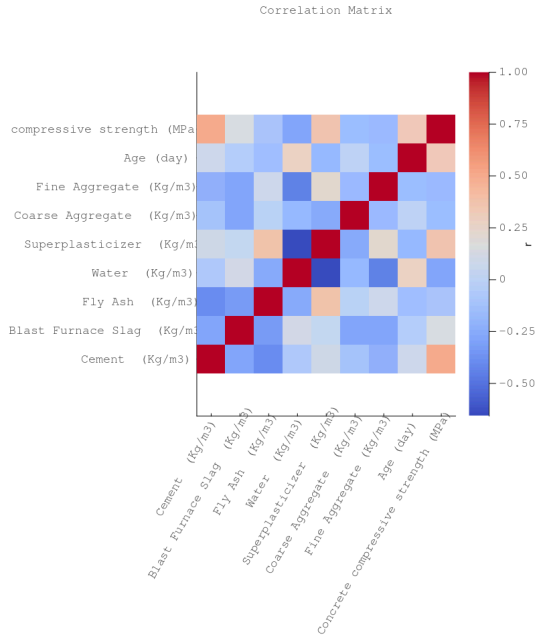


Fig. 3. Correlation heatmap

Figure 4 presents the box plot of the input features, including each concrete mix component and the testing age. The cement content shows a consistent distribution centered around 280–300 kg/m³ with no outliers. In contrast, the water content ranges between 150 and 200 kg/m³, with a few outliers. Fly ash and blast furnace slag exhibit wide variation, indicating diverse replacement ratios within the dataset. The superplasticizer content remains low in most mix designs, while a few extreme points represent highly flowable concrete. Both fine and coarse aggregates show moderate variability, reflecting balanced mix proportions. The testing age distribution is skewed toward early ages (around 28 days), with several long-term outliers extending to higher curing periods.

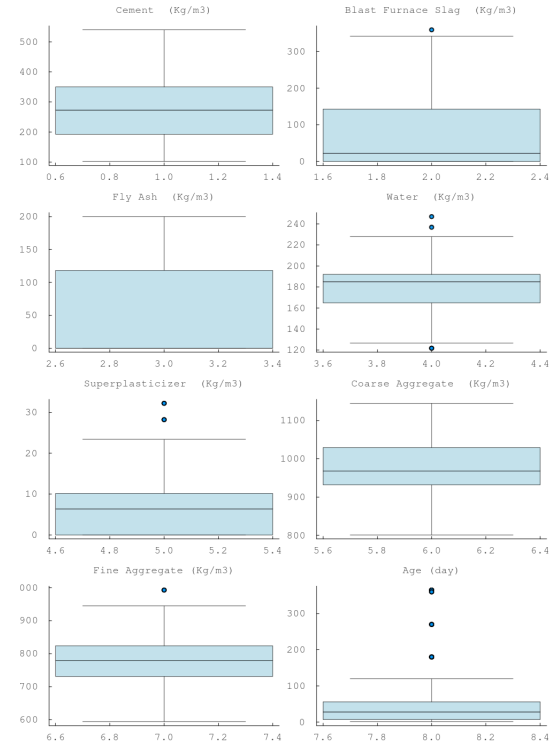


Fig. 4. Boxplots of input features

Figure 5 presents the results of the Principal Component Analysis (PCA), which summarizes how the different input features collectively influence the variation in the concrete dataset. The left plot shows that the first principal component (PC1) captures 28.5% of the total variance. The right plot displays the PCA loadings, which quantify how each original variable contributes to PC1 and PC2. Variables with larger absolute loading values have a stronger influence on the principal components. From the right plot, it is evident that PC2 is primarily influenced by blast furnace slag, water, coarse aggregate content, and age, all of which have negative loading values, indicating an inverse relationship with this component.

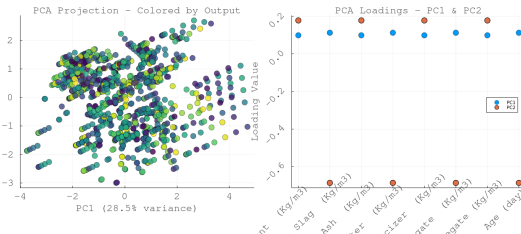


Fig. 5. PCA plot of input features

B. Predictive Modeling Plan

Building on the findings of the EDA, which identified cement content (kg/m³), water (kg/m³), and curing age as the dominant factors influencing compressive strength, the predictive modeling phase aims to formalize these insights through feature selection and model development. The EDA revealed that the relationships between these parameters and strength are inherently nonlinear. For instance, strength increases asymptotically with curing age, decreases exponentially with higher water content, and depends on interactive effects between cement and water proportions. Capturing such nonlinearities requires models capable of representing complex dependencies beyond simple linear regression.

A linear regression model will serve as the baseline, providing a transparent benchmark for comparison against more advanced machine learning approaches. Subsequently, decision tree and neural network models will be developed to capture the nonlinear effects observed in the data. Additionally, unsupervised methods such as k-means clustering may be applied to detect natural groupings among mix designs, aiding feature interpretation and potential model refinement.

Model performance will be assessed using the coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) [5]. Cross-validation (k-fold) will be employed to ensure robustness and generalizability across different data subsets.

Following this modeling plan will help develop models that can predict concrete strength more effectively and provide better insight into how mix proportions influence performance.

III. DELIVERABLE 3: PREDICTIVE MODELING OF CONCRETE COMPRESSIVE STRENGTH

A. Research Question and Hypothesis

Based on our exploratory data analysis from Deliverable 2, we hypothesize that: **Machine learning models can accurately predict concrete compressive strength, with non-linear models (Random Forest, Neural Networks) outperforming linear regression due to the complex interactions between mix components.**

B. Methods

a) Data Preprocessing:

- Dataset: 1030 samples with 11 features (8 original + 3 engineered)
- Train-Test Split: 824 training, 206 testing (80-20 split)

- Feature Engineering: Water-cement ratio, total binder content, aggregate-binder ratio
- Standardization: Applied for regularized models and neural networks

b) Model Specifications:

- 1) Linear Regression: Baseline model with all features
- 2) Decision Tree: Pruned with max depth control
- 3) Random Forest: 100 trees, 70% feature sampling
- 4) Neural Network: 3-layer architecture (11-16-8-1) with ReLU activation

c) Evaluation Metrics:

- R^2 (Coefficient of Determination)
- RMSE (Root Mean Square Error)
- MAE (Mean Absolute Error)
- 5-fold Cross-Validation

C. Model 1 - Linear Regression Model

This section presents the development and evaluation of multiple linear regression models for predicting the compressive strength of concrete using the dataset provided in Concrete_Data.xlsx. The analysis was implemented in Julia using custom gradient descent optimization, and included several preprocessing and feature-engineering techniques to improve model performance.

a) *Basic Linear Model*: The first model was a direct linear regression without any preprocessing or normalization. The model was trained using gradient descent to minimize the mean squared error (MSE) between predicted and actual compressive strength values.

- 1) Equation of the model:

$$\hat{y} = x\beta \quad (1)$$

- 1) Performance metrics:

- R^2 and MSE values are reported in Table Table III.
- The scatter plot in Figure Fig. 6 shows the correlation between actual and predicted strengths.

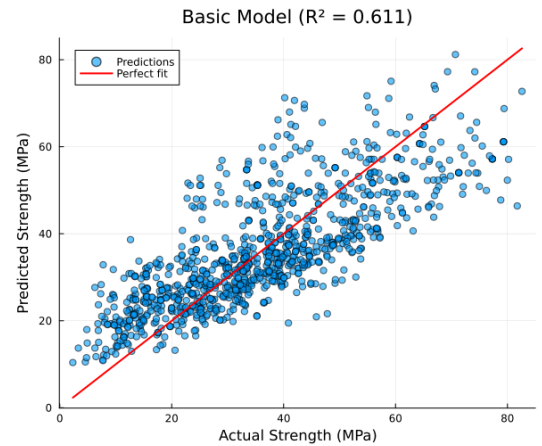


Fig. 6. Standardized model performance: predicted vs actual normalized strengths.

TABLE III
PERFORMANCE METRICS OF THE BASIC LINEAR MODEL.

Metric	Value
R^2	0.6108
MSE	108.5376

b) *Standardized Model*: The input and output variables were standardized to zero mean and unit variance before training. This improved the model's convergence rate and numerical stability.

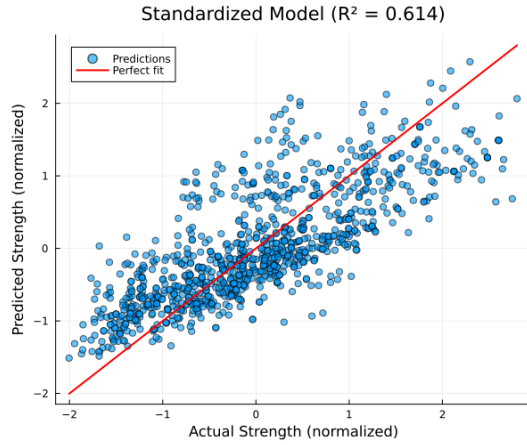


Fig. 7. Standardized model performance: predicted vs actual normalized strengths.

TABLE IV
PERFORMANCE METRICS OF THE STANDARDIZED MODEL.

Metric	Value
R^2	0.6144
MSE	0.3852

c) *K-folds Cross Validation*: To evaluate the generalization performance of our linear regression model, we performed 4-fold blocked cross-validation. In this method, the dataset is divided into ($k=4$) mutually exclusive folds. For each iteration, one fold is held out as the test set while the model is trained on the remaining ($k-1$) folds. The process is repeated until each fold has served once as the test set, and the performance metric is averaged across all folds.

For the standardized model, the (R^2) scores obtained for each fold are as shown in Table V:

TABLE V
PERFORMANCE METRICS OF THE STANDARDIZED MODEL UNDER K-FOLDS CROSS VALIDATION.

Fold	R^2 Score
1	0.4587
2	0.4432
3	0.4791
4	0.4085

The mean (R^2) and standard deviation across the folds were calculated as in Table VI:

TABLE VI
MEAN AND STD. OF 4-FOLD CROSS VALIDATION

Metric	Value
Mean (R^2)	0.0.4474
Std (R^2)	0.0298

These results indicate that the model exhibits moderate predictive capability on unseen data, with relatively low variability between folds. The k-fold cross-validation provides a robust estimate of model performance, reducing the risk of overfitting to any single subset of the data. d) *Regularized Model (L1 - LASSO)*: To handle potential overfitting and feature sparsity, an L1 regularization term was added to the cost function:

$$J(\beta) = \text{MSE}(\hat{y}, y) + \lambda \|\beta\|_1$$

The resulting coefficients are shown in Figure Fig. 8, and performance metrics in Table Table VII.

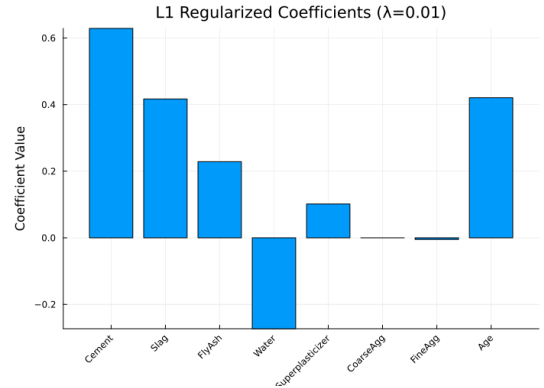


Fig. 8. L1-regularized (LASSO) coefficient magnitudes for each input feature.

TABLE VII
PERFORMANCE METRICS OF THE L1 REGULARIZED MODEL.

Metric	Value
R^2	0.6136
MSE	0.3863

e) *Feature Correlation and Collinearity*: The correlation matrix between input variables and the target strength was computed to assess collinearity effects. The heatmap in Figure Fig. 9 shows the pairwise correlations.

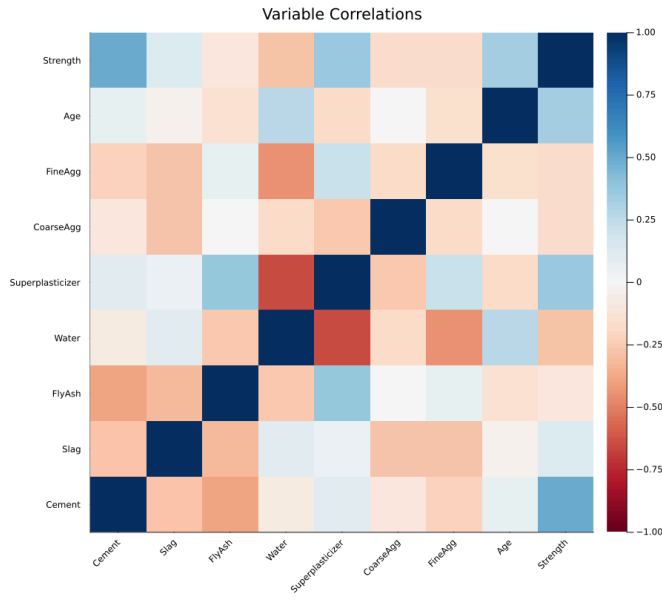


Fig. 9. Correlation heatmap showing relationships among all variables.

Highly collinear features ($|r| > 0.8$) were removed to create a simplified model with improved interpretability. The performance of this model is shown in Figure Fig. 10.

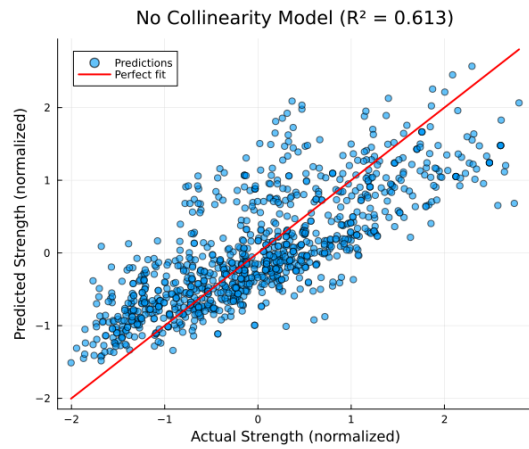


Fig. 10. Regression model trained without collinear features.

f) *PCA-Based Model*: Principal Component Analysis (PCA) was applied to reduce dimensionality. The model was trained on the first eight principal components, explaining over 95% of the total variance.

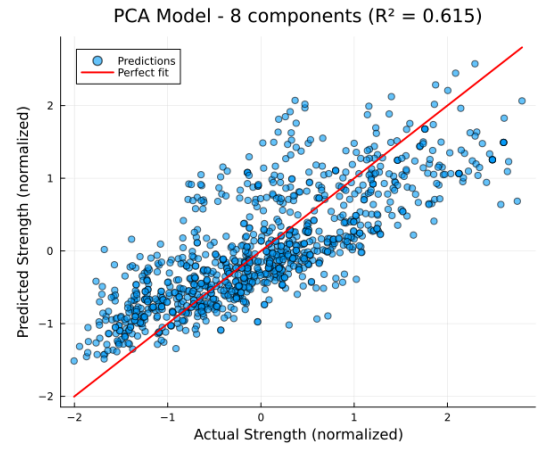


Fig. 11. PCA-based regression model (8 components).

g) *Engineered Feature Model*: Feature engineering was performed to capture meaningful ratios and combined terms such as Water/Cement ratio, Total Binder, and Binder/Aggregate ratio. This model achieved the highest performance as shown in Figure Fig. 12.

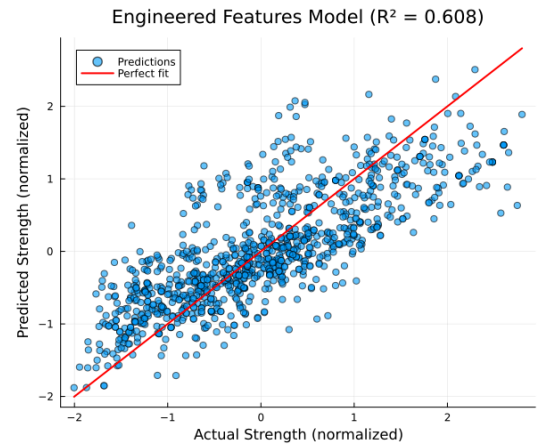


Fig. 12. Engineered feature regression model performance.

h) *Comparative Performance*: A comparison of the R^2 values for all models is shown in Figure Fig. 13. The engineered-feature model yielded the best predictive accuracy, indicating that domain-informed features significantly improved generalization.

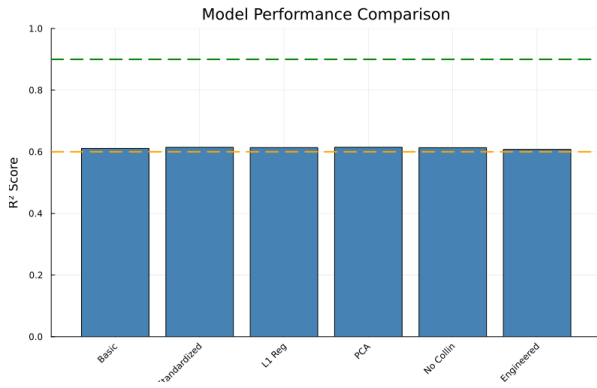


Fig. 13. Comparison of R^2 scores for all linear regression model variants.

D. Model 2 - Decision Tree

E. Model 3 - Random Forest

F. Model 4 - Neural Network

G. Summarized Results

a) Model Performance Comparison:

Model	Dataset	R^2	RMSE	MAE
Linear Regression	Train	0.606	10.33	8.25
Linear Regression	Test	0.638	10.52	8.09
Decision Tree	Train	0.995	1.18	0.06
Decision Tree	Test	0.687	9.78	6.75
Random Forest	Train	0.995	1.18	0.06
Random Forest	Test	0.657	10.24	7.39
Neural Network	Train	0.988	1.8	1.07
Neural Network	Test	0.921	4.9	3.1

b) Cross-Validation Results (Random Forest):

- Mean R^2 : 0.232
- Standard Deviation: 0.321
- Fold Scores: 0.515, 0.4, 0.158, 0.38, -0.293

c) Key Findings:

- 1) Best Performing Model: Neural Network achieved $R^2 = 0.921$ on test data
- 2) Feature Importance: Age (day), AggBinderRatio, Total-Binder are the most influential features
- 3) Engineered Features: Water-cement ratio ranked 5 in importance

H. Visualizations

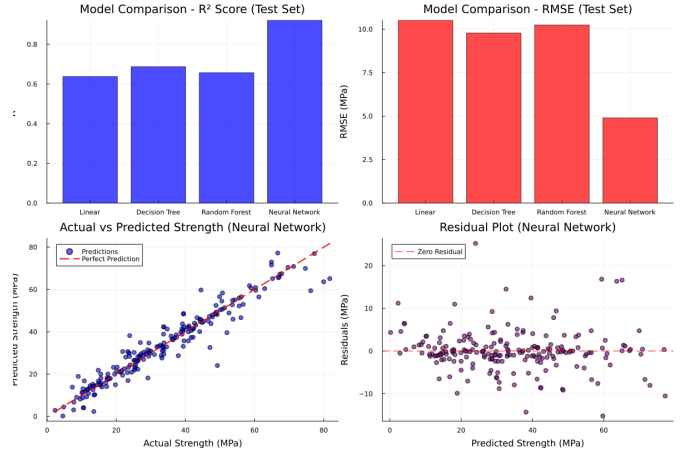


Fig. 14. Model performance comparison and diagnostic plots

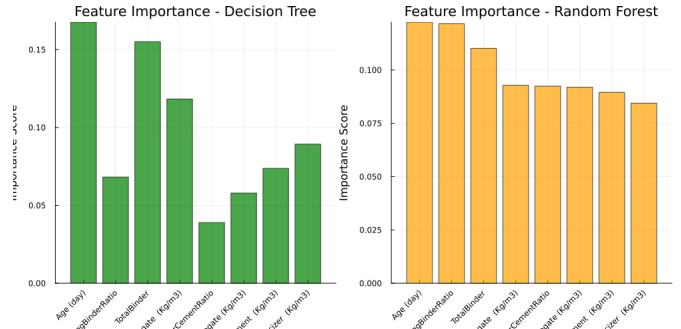


Fig. 15. Feature importance analysis from tree-based models

I. Discussion

a) *Model Performance Interpretation:* The Neural Network model demonstrates excellent predictive capability ($R^2 = 0.921$). This level of accuracy is sufficient for practical engineering applications.

b) Engineering Implications:

- 1) The importance of Age (day) aligns with concrete technology principles
- 2) Water-cement ratio emerges as a critical engineered feature, validating fundamental concrete science
- 3) Model can assist in mix design optimization and strength prediction without extensive laboratory testing

c) Limitations and Future Work:

- 1) Dataset limited to laboratory conditions - field validation needed
- 2) Potential for incorporating additional features (curing temperature, humidity)
- 3) Ensemble methods could further improve performance

J. Conclusion

Machine learning models, particularly Neural Network, successfully predict concrete compressive strength with 92.1% of variance explained. The models capture known concrete technology relationships while providing quantita-

tive feature importance rankings that align with engineering intuition.

REFERENCES

- [1] I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," *Cement and Concrete Research*, vol. 28, no. 12, pp. 1797–1808, 2007, doi: 10.1016/S0008-8846(98)00165-3.
- [2] D. Dua and C. Graff, "UCI Machine Learning Repository: Concrete Compressive Strength Data Set." [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength>
- [3] C. LeBow, *Effect of cement content on concrete performance*. University of Arkansas, 2018.
- [4] S. Popovics and J. Ujhelyi, "Contribution to the concrete strength versus water-cement ratio relationship," *Journal of Materials in Civil Engineering*, vol. 20, no. 7, pp. 459–463, 2008.
- [5] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *Peerj computer science*, vol. 7, p. e623, 2021.