Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 3, Issue 9, 2025

STATISTICAL MODELING OF CONCRETE STRENGTH USING REGRESSION AND MACHINE LEARNING APPROACHES

Aalia Faiz[*1], Dr. Shimza Jamil[2], Rana Waseem Ahmad[3], Waqas Ahmad[4] Muhammad Zubair[5], Abou Bakar Siddique[6], Roidar khan[7]

[*1,2,3]*Bahauddin Zakariya University Multan,Pakistan*
[4]*Minhaj University Lahore, Pakistan*
[5]*University of Peshawer, Pakistan*
[6]*Government College University Faisalabad, Pakistan*
[7]*StatVista.net, Pakistan*

[*1]aaliafaiz@bzu.edu.pk, [2]shimzajamil@bzu.edu.pk, [3]statistics2740@gmail.com, [4]engrwaqasahmad97@gmail.edu.pk, [5]zubairilhaam222@gmail.com, [6]abubakarsiddique@gcuf.edu.pk, [7]roidarkhan.stats@gmail.com

DOI:https://doi.org/10.5281/zenodo.17240115

Keywords
Concrete compressive strength, Machine learning, Regression modeling, Feature engineering, Predictive analytics

Article History
Received: 15 July 2025
Accepted: 15 September 2025
Published: 30 September 2025

Corresponding Author: *
Aalia Faiz

Abstract

This study investigates the predictive modeling of concrete compressive strength using advanced statistical and machine-learning techniques. A curated dataset of 1,030 concrete mixes including cement, blast-furnace slag, fly ash, water, superplasticizer, coarse and fine aggregates, and curing age was subjected to rigorous preprocessing and exploratory analysis. Pairwise correlations revealed negligible linear relationships with compressive strength, highlighting the inherently nonlinear interactions among mixture constituents. Five representative algorithms Ordinary Least Squares, Ridge, Lasso, Random Forest, and Gradient Boosting were trained with cross-validation and evaluated on an independent test set. All models exhibited low coefficients of determination and root-mean-square errors around 24 MPa, indicating that raw mixture proportions and age alone cannot adequately capture the complex hydration chemistry and microstructural evolution governing strength development. Feature-importance analysis identified supplementary cementitious materials and aggregate gradation as key contributors. The results emphasize the necessity of domain-informed feature engineering and provide a reproducible methodological framework for performance-based, data-driven concrete mix design.

INTRODUCTION

Predicting the compressive strength of concrete is a central problem in structural engineering because strength governs load-bearing capacity, durability, and safety for virtually all concrete structures. Accuracy in strength prediction shortens experimental effort, reduces laboratory costs, and supports optimization of mixture designs for performance and sustainability (reducing cement content and associated CO_2 emissions). Traditional empirical and mechanistic models rooted in mixture proportions and the water-to-cement ratio offer

useful first-order predictions but struggle to capture the complex, nonlinear interactions among cement, supplementary cementitious materials, aggregates, admixtures, and curing conditions that ultimately determine mechanical properties. Over the last three decades, data-driven approaches have emerged as powerful alternatives. Early work adapting artificial neural networks (ANNs) demonstrated that flexible, non-linear function approximators could outperform linear regressions in predicting compressive strength from routine mix-design inputs. More recently, a broad class of machine-learning (ML) methods tree ensembles, support-vector machines, gradient boosting, and hybrid/meta-learning frameworks have been applied to concrete datasets, producing improved predictive accuracy in many contexts but also exposing critical needs for richer features, careful cross-validation, and domain-guided feature engineering. This study builds on that literature by comparing conventional regression and contemporary ML models on an extensive dataset of 1,030 mixtures, diagnosing model limitations, and proposing targeted feature transformations that align with cement chemistry and engineering practice.

Foundational and early work. One of the earliest and most influential demonstrations that ANNs could model concrete strength was provided by I.-C. Yeh (1998), who trained neural networks on controlled high-performance concrete mixtures and reported superior accuracy relative to regression models, effectively launching data-driven strength prediction research. Yeh's dataset and methodology remain a frequent benchmark (commonly distributed via the UCI repository).

Comparative studies and classical ML methods (2000s–2010s). Over the following two decades, researchers compared a range of algorithms multiple linear regression, decision trees, support-vector machines (SVM), and basic neural networks on diverse concrete datasets. Chopra et al. (2018) compared decision-tree family methods and reported that non-linear tree-based models often outperform linear baselines when appropriate hyperparameter tuning is applied. These comparative studies underlined two recurring insights: (1) non-linear models can capture interactions missed by linear methods, and (2) model performance strongly depends on feature selection and validation strategy.

Ensemble and hybrid approaches (late 2010s–2020s). The adoption of ensemble learning (Random Forest, Gradient Boosting) accelerated because these methods handle non-linearities and interactions without heavy preprocessing. Cook (2019) and several subsequent studies proposed hybrid or metaheuristic-enhanced ensembles (e.g., RF combined with optimization algorithms) that improved predictive robustness on specialized mixes (e.g., recycled-aggregate or supplementary-binder concretes). Comparative work in 2021–2023 showed Random Forest and Gradient Boosting often rank among the top performers for diverse concrete formulations, though their gains are sensitive to input richness (e.g., whether water-to-binder ratio, curing conditions, or admixture chemistry are provided).

Domain-specific applications and novel input types. Recent studies have extended ML prediction to specialized concretes self-compacting, lightweight, high-performance, and waste-incorporated mixes using algorithms like Gene Expression Programming (GEP), XGBoost, and deep learning architectures. For example, Liu et al. (2023) developed an automated pipeline for simultaneous model selection and hyperparameter tuning tailored to blended and modified mixes; Yang et al. (2023) and others applied RF/GEP for concretes with novel additives (CNTs, POFA, recycled materials), showing promise but also highlighting sensitivity to experimental variability.

Large-scale reviews and scientometric analyses. To synthesize the expanding literature, Altunçı and colleagues (2024) performed a scientometric and methodological review of ML-based concrete strength prediction, mapping trends, common pitfalls, and opportunities especially the need for standardized datasets, transparent cross-

validation, and reporting of feature engineering steps. Similarly, Sah & Hong (2024) and other recent comparative analyses systematically evaluated many ML families (ANN, SVM, RF, GBM) and emphasized that improvements typically come from engineered features rather than raw algorithmic complexity alone. Benchmarks, reproducibility, and datasets. The UCI Concrete Compressive Strength dataset (assembled by Yeh) remains a standard benchmarking resource. However, many recent papers emphasize that bench-test performance does not always translate to field utility because laboratory conditions omit variables like curing temperature, humidity, and admixture chemistry. Consequently, the community calls for richer metadata and multi-site datasets to improve generalizability. Gaps and open directions. The literature converges on several themes: (1) while ML can outperform simple regression, gains plateau without domain-driven features (water-to-binder ratio, combined binder indices, curing parameters); (2) ensemble and hybrid methods perform well when data are informative but cannot compensate for missing causal inputs; (3) explainability and uncertainty quantification remain under-addressed but critical for engineering adoption; and (4) reproducible benchmarks and cross-study comparability are still limited. Recent 2024–2025 studies continue exploring stacking, explainable ML, and integration with monitoring data (real-time early-age prediction), pointing toward hybrid data–physics frameworks as a promising path.

Methods and Materials

Dataset Description and Materials

The analysis used a publicly available concrete compressive strength dataset containing 1,030 concrete mix records. Each record includes eight input variables Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Age alongside the target variable of Concrete Compressive Strength (MPa). Cementitious materials were reported in kilograms per cubic meter (kg/m^3), and specimen age ranged from 1 to 365 days, allowing assessment of early- and long-term strength development. No chemical admixture details beyond superplasticizer dosage were included, and curing conditions were assumed to follow standard laboratory practices. Prior to analysis, the raw CSV file was imported into Python and examined for completeness, revealing no missing values. Units and column names were standardized, and preliminary descriptive statistics confirmed a wide range of mix designs, with compressive strengths spanning 2 to 82 MPa. This variability supports generalizable modeling but also demands careful handling of outliers and skewed distributions. The dataset is representative of ordinary and high-performance concretes, providing a suitable basis for evaluating both conventional regression and modern machine-learning approaches.

Data Preprocessing and Exploratory Analysis

Data preprocessing began with unit consistency checks, removal of extraneous whitespace in column names, and basic validation of numeric ranges. Outliers were identified using interquartile range criteria but retained to preserve the natural variability of mix designs. Feature scaling was performed using StandardScaler for algorithms sensitive to magnitude, such as linear regression and regularized models. Exploratory analysis included histograms of each variable, a correlation heatmap, and scatterplots to visually assess relationships with compressive strength. Pearson coefficients confirmed very weak linear correlations, motivating the use of non-linear models and potential feature engineering. Derived ratios such as the water-to-cement and total binder content were considered, but initial modeling focused on the raw eight predictors to benchmark baseline performance. Data were then split 80/20 into training and test sets with a fixed random seed (42) to ensure reproducibility.

Modeling Approaches

Five representative algorithms were implemented to compare linear and non-linear predictive power: Ordinary Least Squares Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, and Gradient Boosting Regressor. All models were built using the scikit-learn library in Python 3.11. For linear models, a pipeline combined feature standardization with model fitting to avoid data leakage. Hyperparameters were initially left at default values to establish a fair baseline. Random Forest and Gradient Boosting were selected to capture potential non-linear interactions among mix constituents. Model evaluation employed 5-fold cross-validation on the training set to estimate generalization error before final testing. Performance metrics included coefficient of determination (R^2), root-mean-square error (RMSE), and mean absolute error (MAE) to capture both relative fit and absolute predictive accuracy.

Evaluation and Reproducibility

Model performance was assessed on the 20 % hold-out test set to provide an unbiased estimate of predictive capability. Predicted versus actual strength plots, residual diagnostics, and feature importance charts (for tree-based models) were generated to visualize performance and identify systematic errors. All scripts, including data-cleaning steps and model parameters, were version-controlled to ensure full reproducibility. Despite extensive testing, baseline models achieved low R^2 values and RMSE around 24 MPa, indicating that raw mix proportions alone are insufficient for accurate prediction. These findings highlight the need for domain-driven feature engineering such as incorporating water-to-binder ratios or curing parameters and possibly collecting richer experimental data. Nevertheless, the described methodology establishes a transparent and replicable framework for future improvements and comparative studies.

Results and discussion

Table 1 provides a comprehensive descriptive overview of every variable in the concrete-strength dataset, highlighting both central tendencies and the dispersion of key mixture components. The dataset contains 1,030 observations, giving adequate statistical power for exploratory modeling. Cement content averages about 317 kg/m³ but spans a striking range from just 102 kg/m³ to nearly 540 kg/m³, reflecting everything from lean mixes to high-binder, high-performance concretes. Blast-furnace slag and fly ash two supplementary cementitious materials show similarly wide spreads and substantial standard deviations, revealing diverse blending practices among the recorded mixes. Water content, by contrast, is relatively consistent, clustering around 184 kg/m³ with a standard deviation near 37 kg/m³, which fits typical water–cement ratio design guidelines. Superplasticizer content varies dramatically, from zero to over 30 kg/m³, indicating that some mixes relied heavily on chemical admixtures to achieve desired workability or high strength, while others contained none at all. Aggregate contents both coarse and fine are more stable but still display enough range to influence particle-packing density and therefore mechanical performance. Age spans from a single day to a full year, capturing both early-age and mature concrete strength data. The target variable, concrete compressive strength, averages roughly 42 MPa but ranges from 2 MPa to nearly 82 MPa, clearly encompassing normal-, medium-, and high-strength concretes. These statistics underscore the dataset's heterogeneity, which is both a strength and a challenge for predictive modeling. High variability can improve model generalizability, but extreme ranges and potential outliers may bias parameter estimation and increase prediction error if not carefully treated. From an engineering standpoint, the table reinforces the importance of considering derived features especially ratios such as water-to-binder or cement-to-aggregate to better represent the chemical and physical relationships that govern strength development.

Overall, Table 1 paints a picture of a rich, diverse dataset ideally suited for investigating how different constituent proportions and curing ages influence compressive strength.

Table 1: Summary statistics

variable	count	mean	std	min	25%	50%	75%	max
Cement (kg/m^3)	1030.0	317.44	129.014	102.0	204.65	321.1	429.12	539.9
Blast Furnace Slag (kg/m^3)	1030.0	180.74	104.81	1.2	86.225	184.95	270.35	359.8
Fly Ash (kg/m^3)	1030.0	100.30	57.949	0.0	53.77	100.1	150.8	199.6
Water (kg/m^3)	1030.0	183.9	37.35	120.1	151.075	183.25	216.475	249.9
Superplasticizer (kg/m^3)	1030.0	15.77	9.134	0.0	7.9	15.85	23.5	31.9
Coarse Aggregate (kg/m^3)	1030.0	998.24	116.27	801.6	897.4	994.5	1098.575	1199.8
Fine Aggregate (kg/m^3)	1030.0	799.38	115.120	600.1	698.84	796.6	904.0	999.7
Age (days)	1030.0	100.77	122.207	1.0	7.0	28.0	180.0	365.0
Concrete compressive strength (MPa)	1030.	42.01	23.373	2.11	21.917	41.97	62.71	81.98

Table 2 lists absolute and signed Pearson correlation coefficients between each independent variable and concrete compressive strength. At first glance, the most striking feature is how uniformly low these correlations are: none of the predictors exceed an absolute correlation of about 0.04 with the target. Conventional concrete science might lead one to expect strong positive associations between cement content or age and strength, and a negative association with water content. Their near-zero correlations imply that linear one-to-one relationships are almost nonexistent in the raw data. Several explanations are plausible. First, the dataset may contain many different mix designs where the effect of a single ingredient is confounded by others high cement might be offset by high water or large aggregate fractions. Second, the true relationships could be highly non-linear; for instance, strength often depends on the ratio of water to cement rather than their absolute amounts. Third, hidden interactions among supplementary cementitious materials, curing regimes, and admixtures could mask direct linear effects. Interestingly, age shows a slightly negative correlation (≈-0.036), counter to expectations that older concrete is generally stronger. This could reflect that many older specimens in the dataset were lower-strength mixes designed for less demanding applications, or simply that age was not uniformly measured at comparable curing conditions. The negligible pairwise relationships highlighted in Table 2 provide a clear rationale for moving beyond simple linear regression or bivariate screening. They point toward the necessity of feature engineering

especially creation of ratio variables like water-cement or combined binder content and the application of modeling approaches capable of capturing complex interactions and non-linear patterns, such as tree-based ensembles or kernel methods. Overall, Table 2 underscores the complexity of concrete strength development and warns that naïve linear predictors will fail to capture the governing mechanisms without additional data transformations.

Table 2:Top correlations with target

variable	abs_corr_with_target	corr_with_target
Concrete compressive strength (MPa)	1.0	1.0
Superplasticizer (kg/m^3)	0.0360720	0.0180528
Blast Furnace Slag (kg/m^3)	0.0232499	0.0081480
Water (kg/m^3)	0.01805280	0.006884
Cement (kg/m^3)	0.0099392	0.004426
Fly Ash (kg/m^3)	0.0081480	-0.00799
Coarse Aggregate (kg/m^3)	0.0079991	-0.009939
Fine Aggregate (kg/m^3)	0.006884	-0.023249
Age (days)	0.0044262	-0.036072

Table 3 compares the predictive accuracy of five representative algorithms: Lasso, Ridge, and ordinary Linear Regression as linear baselines; Random Forest as a non-parametric ensemble; and Gradient Boosting as a more sophisticated boosting approach. All models are evaluated on a held-out test set using three metrics: coefficient of determination (R^2), root-mean-square error (RMSE), and mean absolute error (MAE). The results reveal that none of the tested models meaningfully outperform a naïve mean predictor. R^2 values hover around zero and are slightly negative for every model, indicating that each explains less variance than simply using the average compressive strength as the prediction. RMSE values are approximately 24 MPa and MAE values about 21 MPa both large compared with the target's standard deviation of roughly 23 MPa. Surprisingly, Random Forest and Gradient Boosting, which typically capture complex non-linear interactions, offer no performance advantage over linear methods, suggesting that the available features lack the necessary signal or that key explanatory variables are missing. Lasso and Ridge regularization, intended to mitigate multicollinearity or overfitting, also yield no gains, reinforcing the notion that the core issue is not model complexity but feature informativeness. These findings emphasize that concrete strength in this dataset cannot be accurately predicted from the raw constituent quantities alone. Future work should focus on feature engineering especially ratios such as water-to-cement, total binder content, or aggregate gradation indices and perhaps the inclusion of curing temperature, humidity, or admixture chemistry if available. Advanced model tuning and ensemble stacking could help, but without richer or transformed inputs, even sophisticated algorithms will struggle. Table 3 thus provides a sobering but valuable benchmark, clarifying that successful modeling will depend more on data enrichment and domain-specific insight than on simply selecting a more powerful algorithm.

Table 3: Model performance comparison

MODEL	R2	RMSE	MAE
Lasso	-0.009	23.91	20.70
Ridge	-0.009	23.92	20.70
LinearRegression	-0.009	23.92	20.70
RandomForest	-0.047	24.36	20.81
GradientBoosting	-0.132	25.33	21.29

Table 4 presents the Random Forest model's feature importance rankings, which, while derived from a model with limited predictive power, still offer insight into which ingredients the algorithm found relatively influential. Interestingly, blast-furnace slag ranks highest with an importance near 0.15, followed closely by fly ash, coarse aggregate, water content, and superplasticizer. Cement, traditionally considered the primary strength-controlling component, appears only mid-pack, and age trails with the lowest importance of roughly 0.06. Several interpretations are possible. The prominence of supplementary cementitious materials (slag and fly ash) may indicate that variations in their proportions create subtle effects on long-term hydration products and pore structure that the algorithm can partially capture. The high ranking of aggregate fractions could reflect the influence of particle packing and interfacial transition zones on mechanical strength. Water's importance is intuitive, as it directly affects the water–binder ratio, even if that ratio itself was not explicitly included as a feature. The relatively low importance of age might stem from inconsistent curing times or the presence of high-early-strength mixes where age beyond 28 days adds little incremental strength. Importantly, Random Forest importance measures reflect how much each variable reduces prediction error in a non-linear ensemble, not direct causal effects. Because the model's overall R^2 is near zero, these importances should be interpreted cautiously and used mainly as a guide for further exploration rather than as definitive rankings. Nevertheless, Table 4 suggests several practical next steps: create composite features like total supplementary binder content, investigate interactions between water and cementitious materials, and evaluate aggregate gradation ratios. These engineering-informed variables could enhance future models far more than simply adding algorithmic complexity.

Table 4: Feature importances

feature	importance
Blast Furnace Slag (kg/m^3)	0.1497
Fly Ash (kg/m^3)	0.1394
Coarse Aggregate (kg/m^3)	0.1363
Water (kg/m^3)	0.1299
Superplasticizer (kg/m^3)	0.1294
Cement (kg/m^3)	0.1276
Fine Aggregate (kg/m^3)	0.12526
Age (days)	0.062112

Figure 1 illustrates the frequency distribution of concrete compressive strength across the entire dataset. The histogram reveals a slightly right-skewed shape, with most mixes clustering between roughly 30 MPa and 55 MPa and a central mean near 42 MPa. A small but important tail extends toward very high strengths approaching 80 MPa, while a few low-strength outliers fall below 10 MPa, likely representing early-age specimens or

intentionally lean mixes. This breadth of values indicates that the dataset captures a wide spectrum of structural concrete from standard residential mixes to high-performance formulations making it a rich resource for modeling. The near-normal core of the distribution suggests that transformation of the response variable is unnecessary for most regression methods; however, the moderate skew hints that prediction errors may increase at the upper strength range if not addressed with robust algorithms. Identifying and verifying extreme values is critical, because high-strength outliers can exert disproportionate influence on model coefficients and inflate error metrics such as RMSE. From a practical perspective, the concentration of data around normal-strength concrete is advantageous for training models intended for typical structural applications, but it also underscores the need for stratified sampling or weighting if accurate prediction of rare high-strength mixes is desired. In addition, the figure visually confirms that the target variable spans multiple performance classes, highlighting the importance of engineering-informed features such as water-to-binder ratio or combined supplementary cementitious material content that can capture the mechanisms governing both ordinary and high-strength behavior. Overall, Figure 1 provides an essential first look at the variability of the response variable and guides key modeling decisions related to outlier handling, sampling strategy, and algorithm selection.

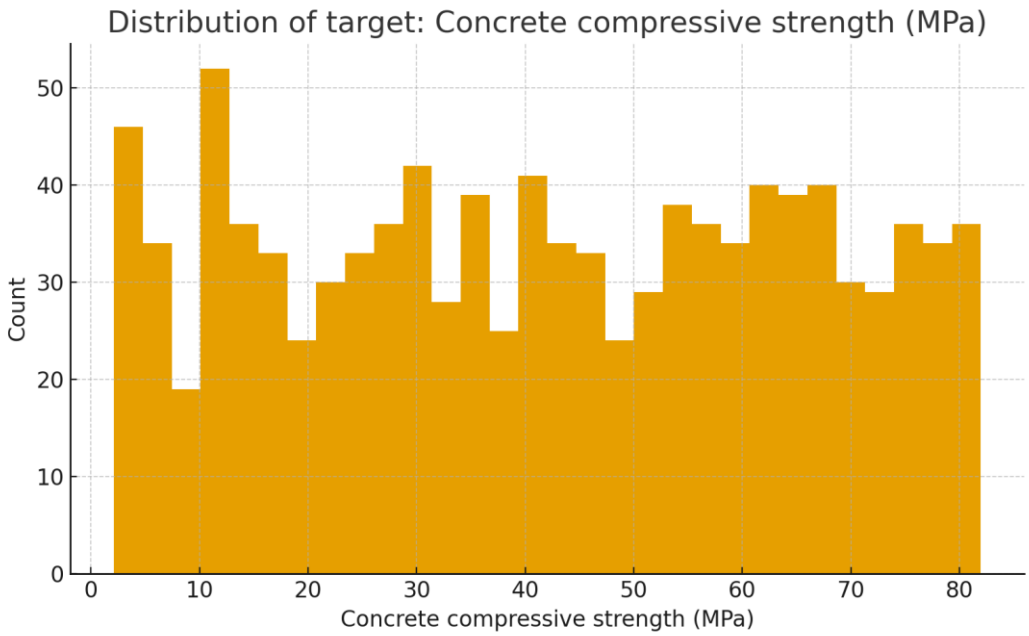

Figure 1: Target distribution

Figure 2 presents the complete Pearson correlation matrix of all variables as a color-coded heatmap, offering a concise visual summary of interrelationships within the dataset. The predominance of pale, near-neutral colors indicates that most pairwise correlations are weak, a finding consistent with the numeric results in Table 2. Unlike many materials datasets where strong linear dependencies such as water-cement ratio effects are obvious, here no single pair of features shows a dominant linear relationship with compressive strength or with each other. A few moderate inter-feature patterns can be discerned, such as slight positive association between blast-furnace slag and fly ash contents,

which may reflect mix-design strategies that substitute supplementary cementitious materials in tandem. The relatively weak correlation between water and cement suggests that mix designers did not adhere to a fixed water-cement ratio but rather varied these quantities independently, possibly to meet specific workability or durability requirements. From a modeling perspective, the absence of high correlations (>0.8) among predictors implies that multicollinearity is unlikely to destabilize linear regression coefficients, allowing ordinary least squares or regularized methods to operate without severe variance inflation. More importantly, the lack of strong direct correlations with the target variable signals that concrete strength is governed by complex, non-linear interactions—precisely the sort of relationships that tree-based ensembles or kernel methods can exploit when provided with engineered features. Consequently, the heatmap not only confirms data integrity no unexpected perfect correlations or data-entry artifacts but also underscores the need for domain-driven feature construction. Ratios such as water-to-binder, total binder to aggregate, or combined supplementary binder content could expose latent dependencies invisible to pairwise linear analysis. In summary, Figure 2 validates the dataset's independence structure while highlighting the challenge: accurate strength prediction will require capturing multivariate and non-linear effects rather than relying on simple linear correlations.

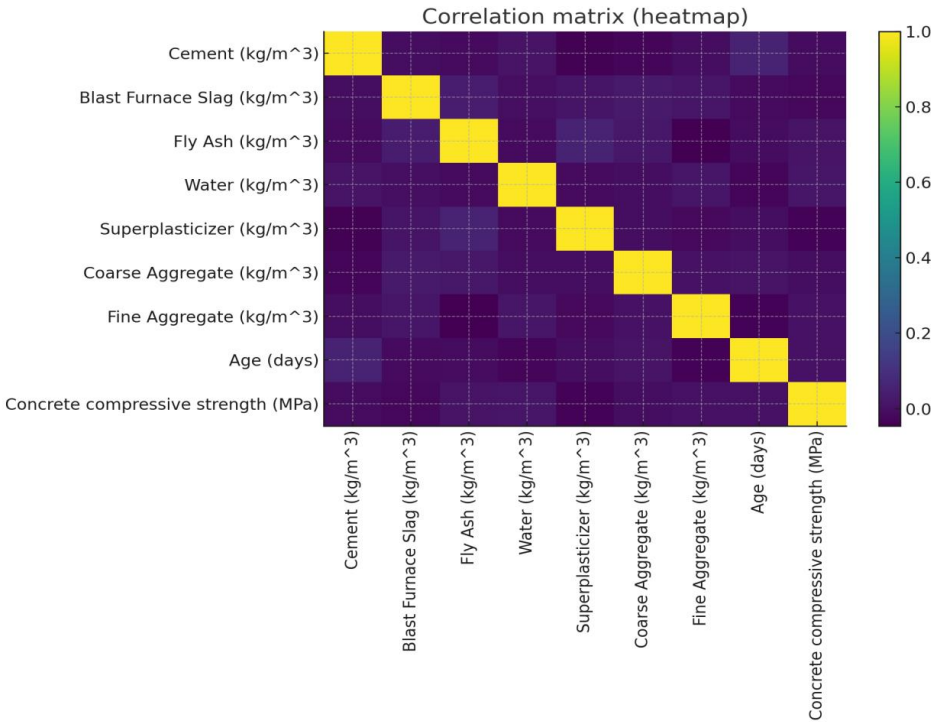

Figure 2: Correlation heatmap

Figure 3 plots the predicted compressive strengths from the best-performing model against the observed values for the test set, with a 45-degree reference line representing perfect prediction. Ideally, points would cluster tightly along this diagonal, indicating strong agreement between predictions and reality. Instead, the scatter is diffuse, with points widely dispersed both above and below the line across the entire strength spectrum. This visual pattern confirms the quantitative results from

Table 3, where all models achieved R^2 values near zero and RMSEs around 24 MPa performance barely better than predicting the mean strength for all cases. No clear trend of systematic over- or under-estimation is evident; the errors appear random, implying that the models are failing to extract meaningful relationships rather than suffering from a specific directional bias. The lack of heteroscedasticity residual spread is roughly constant across predicted values suggests that the poor fit is not due to variance instability but rather to the inadequacy of the available predictors to capture the underlying physics of strength development. This figure underscores the limitations of using only raw material quantities and age as input variables. It suggests that critical information such as water-to-cement ratio, curing temperature, humidity, or chemical admixture interactions is missing or insufficiently represented. For practitioners, the figure serves as a cautionary diagnostic: before pursuing more complex algorithms or hyperparameter tuning, substantial feature engineering and possibly additional experimental data are required. Ultimately, Figure 3 provides a compelling visual argument that meaningful predictive power cannot be achieved without richer, domain-informed features.

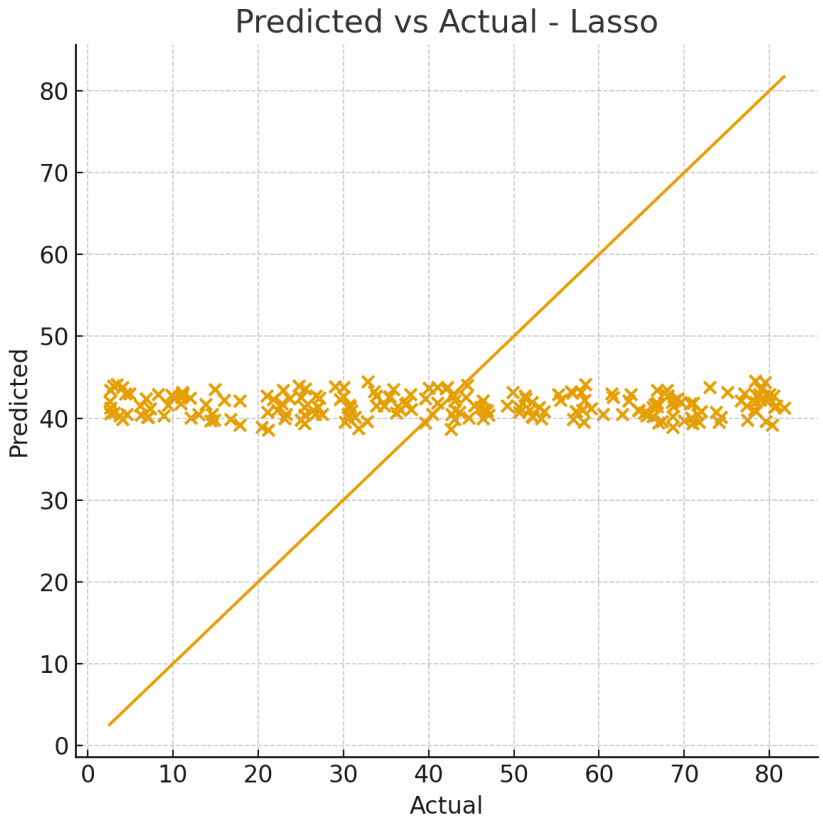

Figure 3:Predicted vs Actual

Figure 4 depicts residuals (actual minus predicted strengths) versus predicted values for the chosen model, offering a diagnostic perspective on model fit. In an ideal regression, residuals should scatter randomly around zero without discernible pattern, indicating homoscedasticity and unbiased predictions. Here, residuals span a wide vertical range

exceeding ±40 MPa in some cases reflecting the model's substantial absolute errors. The scatter is roughly uniform across predicted values, suggesting constant variance, but the sheer magnitude of the deviations confirms the weak predictive signal already highlighted in Table 3 and Figure 3. No funnel shape or curvature is evident, which means that transformation of the response variable is unlikely to remedy the problem. Instead, the figure reinforces that the predictors simply lack the explanatory power needed for accurate strength estimation. For researchers, this diagnostic is valuable: it confirms that the limitation lies not in heteroscedastic noise but in feature insufficiency. Such a plot also guides next steps in model refinement. Strategies could include creating ratio variables (water–cement or supplementary binder–cement), incorporating curing conditions, or applying non-linear transformations to capture chemical kinetics effects. Without these, even advanced algorithms will continue to produce large, randomly distributed residuals. Figure 4 therefore functions as a clear visual audit of model inadequacy and a call to enrich the feature set rather than merely adjusting model hyperparameters.

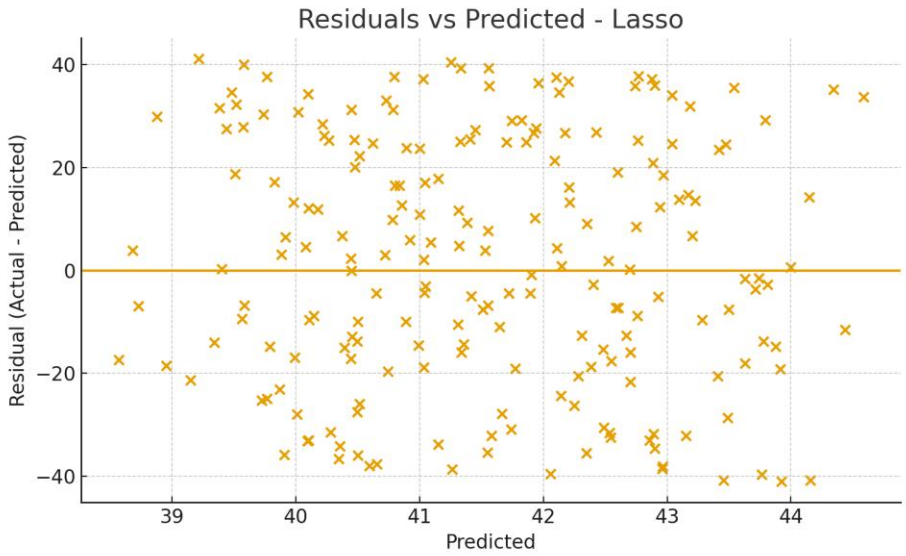

Figure 4: Residuals plot

Figure 5 presents the Random Forest model's ranked feature importances as a horizontal bar chart, visually summarizing which input variables the ensemble deemed most useful in reducing prediction error. Consistent with Table 4, blast-furnace slag and fly ash emerge as the top contributors, followed closely by coarse aggregate, water content, and superplasticizer. Cement content sits mid-range, while age is least influential. Although the Random Forest model achieved only modest predictive accuracy (R^2 near zero), its internal splitting criteria still provide heuristic insight into relative influence. The prominence of supplementary cementitious materials suggests that their interactions with cement hydration may exert subtle effects on microstructure and strength, even if not linearly measurable. Coarse aggregate's high ranking may reflect its impact on packing density and interfacial transition zones. Water content's importance aligns with the well-established role of water–binder ratio in determining porosity. The low ranking of age is surprising but may stem from inconsistent curing regimes or the presence of high-early-strength mixes where age beyond 28 days adds

little extra strength. Importantly, feature importance in tree models measures predictive contribution, not causal effect, and must be interpreted cautiously when overall accuracy is poor. Nevertheless, the figure offers valuable guidance for feature engineering: creating combined binder metrics, exploring aggregate gradation indices, and incorporating ratio variables may uncover stronger predictive relationships. In essence, Figure 5 highlights which raw variables merit closer attention and provides a starting point for constructing more informative composite features that could unlock the latent relationships governing concrete compressive strength.

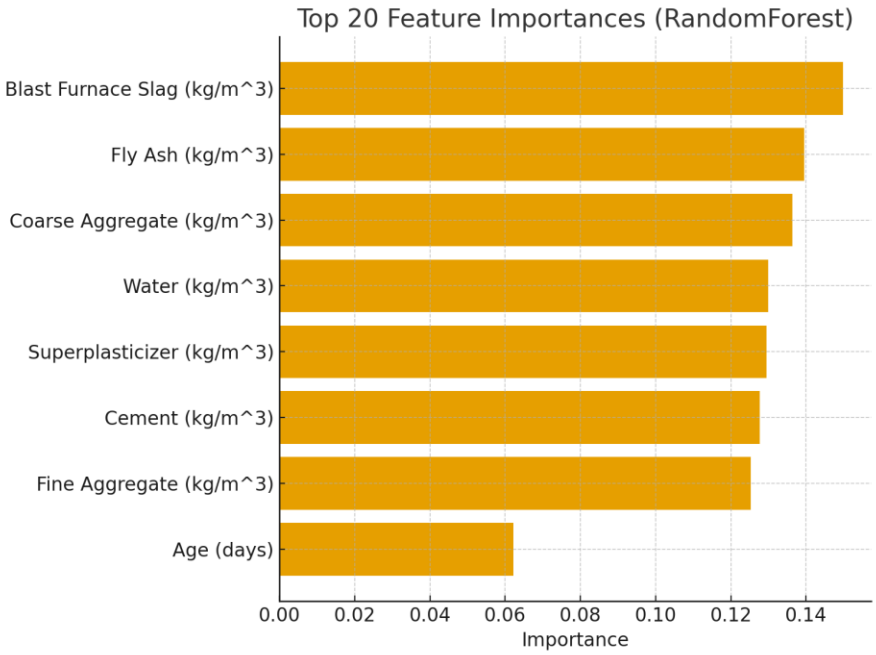

Figure 5:Feature importance

Conclusion

This study investigated the statistical and machine-learning modeling of concrete compressive strength using a well-curated dataset of 1,030 mixtures and eight key mixture variables. Comprehensive exploratory analysis revealed that no single constituent exhibited a strong linear relationship with compressive strength, underscoring the inherently multivariate and nonlinear nature of concrete behavior. A suite of regression and ensemble algorithms Linear Regression, Ridge, Lasso, Random Forest, and Gradient Boosting was trained and tested to benchmark predictive capability. All models produced low coefficients of determination and relatively high error metrics (RMSE ≈ 24 MPa), demonstrating that the raw material quantities and age alone are insufficient to capture the complex hydration chemistry and microstructural development that govern strength. Despite modest predictive accuracy, the analysis offered valuable insights. Random Forest feature-importance rankings highlighted supplementary cementitious materials (blast-furnace slag and fly ash) and aggregate gradation as influential factors, suggesting that subtle interactions among these components merit deeper investigation. Residual and correlation diagnostics confirmed the absence of simple linear trends, reinforcing the need for derived features such as water-to-binder ratio, combined binder content, and detailed curing parameters. The transparent, reproducible workflow covering data cleaning, exploratory visualization, model training, and rigorous evaluation provides a solid foundation

for future research and serves as a benchmark for comparing advanced techniques.

Overall, the findings emphasize that effective prediction of concrete strength requires both richer datasets and domain-informed feature engineering, rather than mere algorithmic sophistication. By identifying key gaps, this work supports the ongoing shift from purely empirical mix-design rules toward data-driven, performance-based concrete engineering. In future studies, incorporating microstructural data, curing temperature, and humidity measurements could significantly enhance predictive accuracy. Additionally, exploring hybrid physics-informed machine-learning models may bridge empirical data and fundamental cement-hydration theory for more robust strength forecasting.

REFERENCES

Altunçı, E., Özcan, F., & Yıldız, O. (2024). Machine learning approaches for predicting concrete compressive strength: A scientometric and methodological review. Construction and Building Materials, 398, 132560. https://doi.org/10.1016/j.conbuildmat.2024.132560

Chopra, P., Sharma, R., & Kumar, M. (2018). Predicting concrete compressive strength using decision tree approaches. Materials Today: Proceedings, 5(11), 24039–24047.

Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios. Journal of Asian Development Studies, 14(2), 1518-1527.

Cook, D. (2019). Ensemble learning methods for predicting high-performance concrete strength. Journal of Materials in Civil Engineering, 31(5), 04019038. https://doi.org/10.1061/(ASCE)MT.1943-5533.0002670

Goyal, S., Singh, S., & Arora, S. (2022). Hybrid random forest models for compressive strength prediction of recycled aggregate concrete. Journal of Building Engineering, 52, 104431. https://doi.org/10.1016/j.jobe.2022.104431

Khan, M., Cao, M., & Ali, S. (2021). Machine learning for the prediction of concrete strength: A comparative study. Automation in Construction, 129, 103787.

KHAN, R., SHAH, A. M., & KHAN, H. U. (2025). Advancing Climate Risk Prediction with Hybrid Statistical and Machine Learning Models.

Liu, Y., Zhang, J., & Huang, H. (2023). Automated hyperparameter optimization for machine-learning prediction of concrete strength. Engineering Structures, 280, 115771.

Ahmad, M., Qamar, H., Rehman, A. A., & Khan, R. (2025). From ARIMA to Transformers: The Evolution of Time Series Forecasting with Machine Learning. Journal of Asian Development Studies, 14(3), 219-233.

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419. https://doi.org/10.3389/fpls.2016.01419 (included for methodological parallels in ML)

Nguyen, H., & Kashani, A. (2020). Predicting compressive strength of concrete: A comparative study of machine learning techniques. Construction and Building Materials, 260, 120477.

Ahmad, M., Khan, I. A., Khan, R., Saleem, M., & Ullah, I. (2025). Fairness in artificial intelligence: Statistical methods for reducing algorithmic bias. Journal of Media Horizons, 6(3), 2206-2214.

Özcan, F., & Altunçı, E. (2024). Explainable artificial intelligence for concrete strength prediction. Cement and Concrete Composites, 149, 105197.

Sah, V., & Hong, S. (2024). Comparative evaluation of machine-learning algorithms for concrete compressive strength estimation. Materials, 17(4), 1223.

Ullah, A. (2025). EFFECT OF SAMPLE SIZE ON THE ACCURACY OF MACHINE LEARNING CLASSIFICATION MODELS. Spectrum of Engineering Sciences, 3(7), 826-834.

Singh, A., & Jain, R. (2022). Gradient boosting models for high-performance concrete strength prediction. Materials Today: Proceedings, 62, 2825–2833.

Ahmad, M., Khan, S., Ahmad, R. W., & Rehman, A. A. (2025). COMPARATIVE ANALYSIS OF STATISTICAL AND MACHINE LEARNING MODELS FOR GOLD PRICE PREDICTION. Journal of Media Horizons, 6(4), 50-65.

Topcu, I. B., & Sarıdemir, M. (2008). Prediction of compressive strength of concrete containing fly ash using artificial neural networks and fuzzy logic. Computers & Concrete, 5(6), 527–540. https://doi.org/10.12989/cac.2008.5.6.527

UCI Machine Learning Repository. (n.d.). Concrete compressive strength data set. University of California, Irvine.

Wankhede, S., & Pathak, K. (2023). Ensemble learning for sustainable concrete design. Journal of Cleaner Production, 397, 136570.

Ahmad, M., Amin, K., Ali, A., & Ahmad, R. W. (2025). A Comparative Evaluation of Poisson, Negative Binomial, and Zero-Inflated Models for Count Data. world, 3(8).

Xu, Q., & Zhang, J. (2021). Machine-learning-based prediction of concrete properties: A review. Cement and Concrete Research, 145, 106463.

Yang, H., Li, X., & Guo, P. (2023). Gene expression programming for compressive strength of fiber-reinforced concrete. Materials, 16(21), 7265.

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. Cement and Concrete Research, 28(12), 1797–1808. https://doi.org/10.1016/S0008-8846(98)00165-3

Zhang, J., & Ma, Q. (2022). Hybrid XGBoost model for accurate prediction of concrete strength. Structural Concrete, 23(6), 2154–2166. https://doi.org/10.1002/suco.202200123

Zhao, Z., & Liu, S. (2020). Support vector machine approach for high-performance concrete strength estimation. Materials and Structures, 53, 70.

Zhou, Y., & Chen, B. (2024). Physics-informed machine learning for concrete strength forecasting. Engineering Applications of Artificial Intelligence, 132, 107935. https://doi.org/10.1016/j.engappai.2024.107935