

Concrete Compressive Strength Prediction System Using K-Nearest Neighbor (KNN) Regression with k -Parameter Tuning Implementation

Delfi Rosaria Simanjuntak¹, Aditama Rouliber Simanulang², Riwi Dyah Pangesti^{3*}, Wina Ayu Lestari⁴, Winalia Agwil⁵

¹ University of Bengkulu, Indonesia, email: authors@email.com

² University of Bengkulu, Indonesia, email: authors@email.com

³ University of Bengkulu, Indonesia, email: rdyahpangesti@unib.ac.id *

⁴ University of Bengkulu, Indonesia, email: walestari@unib.ac.id

⁵ Hasselt University, Diepenbeek, Belgium, email: winaliaagwil@unib.ac.id

Abstract

Concrete compressive strength is a principal indicator of concrete quality, but standardized testing requires specimen preparation, controlled curing, equipment, and time. This study develops and internally evaluates a K-Nearest Neighbor (KNN) regression model for rapid strength estimation from mixture composition and curing age. The UCI Concrete Compressive Strength dataset contained 1,030 records; removing 25 repeated records yielded 1,005 unique observations with eight original predictors and compressive strength as the response. LogAge, Water-Cement Ratio, and Total Aggregate were added as engineered predictors. The cleaned data were divided into 80% training and 20% testing subsets, while data-dependent preprocessing and hyperparameter selection were performed using the training data only. Odd k values from 1 to 41 were assessed by 10-fold cross-validation repeated five times. The minimum cross-validated error occurred at $k = 3$. On the held-out test subset, the final model achieved an R^2 of 0.88, RMSE of 5.56 MPa, and MAE of 3.74 MPa. LogAge, Water-Cement Ratio, and cement content had the largest absolute correlations with strength, although these correlations are descriptive rather than KNN-specific importance measures. The results show that a tuned and standardized KNN model can provide useful preliminary strength estimates within the represented data domain. However, the model is intended to support mixture screening and quality-control decisions, not replace standardized laboratory testing. External validation and comparison with alternative algorithms are required before broader deployment.

Keywords: *Concrete Compressive Strength, KNN Regression, Hyperparameter Tuning, Feature Engineering, Predictive Modeling.*

Introduction

Concrete is one of the most widely used construction materials because its constituent materials are generally accessible, its fresh mixture can be formed into diverse structural configurations, and its mechanical properties can be designed for different service requirements. Among these properties, compressive strength is a principal indicator for evaluating concrete quality, structural capacity, and conformity with specified design performance. Reliable information on compressive strength is therefore important for mixture design, construction control, acceptance decisions, and the long-term safety of concrete structures. Insufficient strength may indicate inappropriate material proportions, inadequate curing, or variability in production, whereas unnecessarily high target strength may increase material consumption and cost. Consequently, estimating strength accurately is not merely a computational exercise; it supports the broader objective of producing concrete that is safe, economical, and consistent

with engineering requirements. Indonesian studies have likewise discussed mixture proportioning, aggregate selection, and compressive-strength assessment in local construction contexts (Arman, 2018; Herdiansyah & Pangaribuan, 2013; Risdiyanto, 2013).

Compressive strength is conventionally determined by casting concrete specimens, curing them under specified conditions, and loading them to failure using a compression-testing machine. This standardized procedure remains indispensable for compliance verification because it directly measures the mechanical response of the tested specimen. Nevertheless, laboratory testing requires specimen preparation, controlled curing, calibrated equipment, trained personnel, and a waiting period before strength at the required age becomes available. Test results may also be affected by specimen geometry, compaction, curing conditions, testing procedures, and material variability. A predictive model can complement this process by providing an early estimate based on available mixture information. Such an estimate may help identify potentially unsuitable mixtures, compare alternative compositions, plan confirmatory testing, and support preliminary quality-control decisions. It should, however, be treated as decision support rather than a replacement for standardized laboratory verification.

The development of concrete strength is governed by interactions among its ingredients and curing history. Cement content contributes to the formation of hydration products, while water enables hydration but may increase capillary porosity when used excessively. The water–cement ratio is therefore a physically meaningful indicator because it represents the balance between available water and cementitious binder. Blast-furnace slag and fly ash can contribute to later-age strength through hydraulic or pozzolanic reactions, although their effects depend on replacement level and curing age. Superplasticizer can improve workability at a lower water content, while coarse and fine aggregates influence packing, stiffness, and the interfacial transition zone. Age is also essential because strength generally increases rapidly during early hydration and then develops more gradually. These simultaneous effects create nonlinearities and interactions that cannot be represented adequately by interpreting each component independently (Yeh, 1998; Young et al., 2019).

The dataset introduced by Yeh and distributed through the UCI Machine Learning Repository provides a widely used benchmark for investigating this relationship (Yeh, 1998, 2007). It contains 1,030 observations, eight input variables (cement, blast-furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and curing age) and concrete compressive strength as the response. The data cover heterogeneous mixture compositions and ages, making them appropriate for regression studies in which the response is generated by several interdependent factors. Yeh (1998) demonstrated early that artificial neural networks could represent the nonlinear relationship more effectively than conventional regression formulations in the studied setting. Subsequent investigations have used the same or related concrete datasets to compare statistical learning, neural networks, support vector machines, model trees, ensemble learners, and optimization-assisted algorithms (Al-Shamiri et al., 2020; Asteris et al., 2021; Behnood et al., 2017; Chou et al., 2014; Pakzad et al., 2023; Rathakrishnan et al., 2022; Wan et al., 2021; Zhang et al., 2024).

The growing use of machine learning in concrete research is motivated by its ability to learn flexible relationships from observed data without requiring a fully specified mechanistic equation. Chou et al. (2014) demonstrated the value of data-driven analysis across concrete-strength datasets, while Behnood et al. (2017) used an M5P model tree to obtain interpretable piecewise predictions. Other studies have examined regularized extreme learning machines, ensemble surrogate models, boosting algorithms, and interpretable learning frameworks (Al-Shamiri et al., 2020; Asteris et al., 2021; Pakzad et al., 2023; Rathakrishnan et al., 2022; Zhang et al., 2024). Li et al. (2023) combined a review of the field with a gradient-boosting regression approach, illustrating the increasing emphasis on both predictive performance and the contribution of mixture variables. Collectively, these studies show that machine learning can capture complex

strength patterns. They also show that reported performance depends on data quality, preprocessing, feature representation, validation design, and hyperparameter selection rather than solely on the name of the algorithm. Recent explainable-machine-learning research similarly emphasizes that predictive accuracy should be accompanied by interpretable assessment of mixture variables (Sapkota et al., 2025).

Evidence from laboratory and field data also indicates that high predictive accuracy on a benchmark dataset does not automatically guarantee performance under different production conditions. DeRousseau et al. (2019) found that models trained on field-acquired data behaved differently from models trained only on laboratory data, highlighting the effect of population and measurement differences. Young et al. (2019) similarly examined a large field dataset and showed that mixture proportions contain valuable predictive information, while also emphasizing the complexity and variability of actual concrete production. Dimensionality reduction and feature selection can further affect model behavior because removing information may simplify the predictor space but can also eliminate variables that contribute to local prediction (Wan et al., 2021). These findings underline the need for a transparent validation strategy and cautious interpretation. A model developed using the UCI dataset should be presented as internally validated for that data environment unless its generalizability has been assessed using independent external data.

K-Nearest Neighbor Regression provides a comparatively simple, nonparametric approach for modeling nonlinear response patterns. For a new observation, KNN identifies the k training observations with the smallest distances in the predictor space and estimates the response from their observed target values. The method does not require a predefined global functional form, which is useful when observations with similar standardized mixture compositions and curing ages are expected to exhibit comparable strength behavior. Its prediction mechanism is also easier to explain than that of many multilayer or ensemble architectures. Instance-based learning has been shown to be applicable to material-strength prediction, supporting the use of neighborhood information for concrete and rock responses (Hsieh, 2021). However, simplicity does not imply that KNN is automatically reliable. Its performance is strongly affected by the representation of predictors, distance calculation, data density, and the selected neighborhood size. The broader use of KNN in classification, regression, forecasting, and recommendation settings also illustrates the portability of neighborhood-based learning, although performance in one domain does not establish validity in another (An et al., 2019; Banjarsari et al., 2015; Daulay, 2024; Fadilla et al., 2018; Ihzaniah et al., 2023; Nanja & Purwanto, 2015; Rahardja et al., 2019; Ritonga et al., 2024).

Preprocessing is particularly important because variables such as aggregate content and superplasticizer are measured on substantially different numerical scales. Without centering and scaling, variables with larger magnitudes can dominate Euclidean distance even when they are not the most informative predictors. Redundant or strongly related features may also give disproportionate weight to a particular physical characteristic. Conversely, domain-informed feature engineering may improve representation when it summarizes relationships that are meaningful in concrete technology. In this study, LogAge is used to represent the diminishing rate of strength development over time, Water–Cement Ratio represents the physical balance between water and cement, and Total Aggregate summarizes coarse and fine aggregate content. These engineered variables require careful evaluation because their usefulness must be established through predictive validation rather than correlation alone.

The number of neighbors, k , controls the balance between local sensitivity and smoothing. A very small k can yield predictions that closely follow nearby observations but may be unstable and sensitive to noise. A large k incorporates more observations and produces smoother estimates, but it may average mixtures with different local characteristics and increase bias. Therefore, k should not be selected arbitrarily or justified only by its performance on one test partition.

Resampling methods provide a more defensible basis for tuning because each candidate is evaluated repeatedly on different subsets of the training data. Repeated cross-validation can reduce dependence on a single split and estimate how consistently a candidate neighborhood size performs. The final test set must remain outside this tuning process so that it provides an independent assessment of the selected model.

Despite extensive research on machine-learning prediction of concrete strength, much of the literature emphasizes sophisticated ensembles, neural networks, or metaheuristic optimization (Al-Shamiri et al., 2020; Asteris et al., 2021; Behnood & Golafshani, 2018; Fei et al., 2023; Li et al., 2023; Pakzad et al., 2023; Rathakrishnan et al., 2022; Zhang et al., 2024). Comparatively less attention has been given to a transparent KNN workflow that simultaneously integrates scale-sensitive preprocessing, physically informed feature construction, systematic k selection, and repeated cross-validation on the UCI benchmark. Some applied KNN studies select k without sufficiently documenting the resampling procedure, while others report prediction metrics without clarifying how preprocessing was separated between training and testing data. This creates a methodological gap because an apparently accurate model may reflect a favorable data split or information leakage rather than reproducible predictive performance. Transparent reporting is especially important for distance-based models because apparently minor preprocessing decisions can materially change neighborhood membership and resulting predictions. Addressing this gap requires a workflow in which preprocessing and tuning are learned from the training data and performance is subsequently evaluated on previously unseen observations.

Accordingly, this study develops a KNN Regression model for predicting concrete compressive strength from mixture composition and curing age. The analysis incorporates LogAge, Water–Cement Ratio, and Total Aggregate as domain-relevant engineered features; standardizes the predictor space for distance-based modeling; and selects k from a predefined candidate grid using 10-fold repeated cross-validation. The selected model is evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The study contributes an accessible and reproducible KNN-based framework that connects concrete-domain information with resampling-based parameter tuning. It also evaluates whether a relatively simple local-learning method can provide useful predictive performance without the complexity of larger ensemble or neural-network architectures. The resulting predictions are intended to support preliminary mixture assessment and quality-control planning while laboratory testing remains the authoritative basis for structural acceptance.

Method

The K-Nearest Neighbor (KNN) algorithm is a supervised learning method that uses labeled observations to estimate an unknown response from nearby observations in the predictor space. In regression, the response is continuous; therefore, the predicted concrete compressive strength is obtained from the mean response of the selected neighbors. The method was retained because it is transparent, makes no fixed parametric assumption about the functional relationship, and provides a useful benchmark for mixture-based strength prediction (Hsieh, 2021). Its performance, however, depends strongly on predictor scaling and on the selected number of neighbors (k).

The analysis followed the research flow shown in Figure 1: data acquisition, integrity checking, feature engineering, training-test partitioning, preprocessing, hyperparameter tuning, final-model fitting, and evaluation on held-out data. The original KNN procedure was maintained: (1) determine a candidate value of k ; (2) calculate the distance between a query observation and every training observation; (3) order the distances and retain the k nearest observations; and (4) average their compressive-strength values to obtain the prediction. The

model is intended as a rapid decision-support tool for preliminary estimation and mixture screening, rather than as a replacement for standardized laboratory testing.

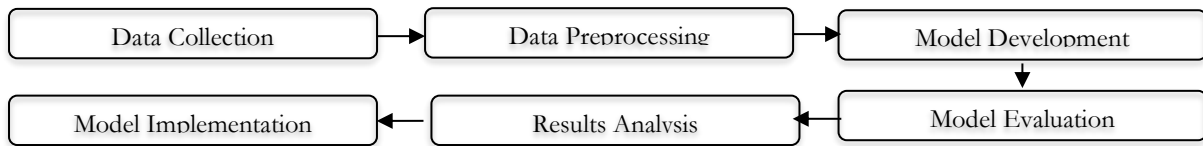


Figure 1. Research Flow Diagram

1. Data Collection

This study used the Concrete Compressive Strength dataset from the UCI Machine Learning Repository (Yeh, 2007), originally compiled by Yeh (1998). The raw dataset contains 1,030 observations. Each observation includes eight predictors: cement, blast-furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate, all expressed in kg/m³, together with curing age in days. The response variable is concrete compressive strength in MPa. These variables represent the mixture composition and curing condition available before the response is known, making them appropriate inputs for a supervised regression model.

Before modeling, variable names, data types, measurement units, missing values, and repeated records were checked. No response value was used as a predictor. The `distinct()` operation removed 25 repeated records, leaving 1,005 unique observations for descriptive analysis and model development, consistent with the sample size reported in Table 1. Removing exact duplicates prevents repeated rows from receiving unintended additional weight during neighbor selection. Figure 2 presents a preview of the research data.

```
# A tibble: 1,005 × 9
```

	Cement	Slag	FlyAsh	Water	SP	CoarseAgg	FineAgg	Age	Strength
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	540	0	0	162	2.5	1040	676	28	80.0
2	540	0	0	162	2.5	1055	676	28	61.9
3	332.	142.	0	228	0	932	594	270	40.3
4	332.	142.	0	228	0	932	594	365	41.1
5	199.	132.	0	192	0	978.	826.	360	44.3
6	266	114	0	228	0	932	670	90	47.0
7	380	95	0	228	0	932	594	365	43.7
8	380	95	0	228	0	932	594	28	36.4
9	266	114	0	228	0	932	670	28	45.9
10	475	0	0	228	0	932	594	28	39.3

```
# i 995 more rows
```

Figure 2. Research Data Preview

2. Data Cleaning and Feature Engineering

Feature engineering was performed to express established mixture relationships more directly, not to correct an 'imbalanced' outcome, because class imbalance is not the relevant concept for a continuous regression target. The descriptive analysis indicated strong right-skewness in curing age and substantial differences in predictor scales. Three deterministic features were therefore added: the natural logarithm of age (LogAge), the water-cement ratio (WCR), and total aggregate content (TotalAgg). They were computed as:

$$\text{LogAge} = \ln(\text{Age}), \quad \text{WCR} = \frac{\text{Water}}{\text{Cement}}, \quad \text{TotalAgg} = \text{CoarseAgg} + \text{FineAgg}$$

LogAge reduces the leverage of very large curing ages while preserving their ordering. WCR represents the balance between mixing water and cement, a domain-relevant quantity associated with strength development, while TotalAgg summarizes the combined coarse- and fine-aggregate content. Cement values were checked to be positive before computing WCR , so division by zero did not occur. The eight original predictors and the three engineered variables were retained,

resulting in 11 candidate predictors. This retention preserves the information used in the original analysis; the predictive value of the combined representation was assessed through resampling rather than assumed in advance.

3. Data Partitioning and Leakage Control

The 1,005 cleaned observations were divided into an 80% training subset and a 20% testing subset using `createDataPartition()`. The partition was generated under a fixed random seed so that the same allocation can be reproduced. The split was performed before estimating any data-dependent transformation. All preprocessing parameters and the value of k were learned exclusively from the training subset; the test subset remained unavailable until final evaluation. This order is necessary because estimating a transformation or selecting a hyperparameter with information from the test observations would produce data leakage and an optimistically biased estimate of generalization performance (Kapoor & Narayanan, 2023). The influence of data-splitting choices on reported performance has also been observed in applied machine-learning evaluations (Prasetyo et al., 2024).

Because the response is continuous, the partitioning procedure was used to maintain a broadly comparable strength distribution across the two subsets rather than to balance discrete classes. The held-out test set represents an internal assessment of unseen observations from the same dataset. It should not be interpreted as external validation across laboratories, geographic regions, materials, or construction conditions.

4. Predictor Transformation and Standardization

KNN relies on distances and is therefore sensitive to measurement scale. Learned transformations were fitted on the training data only and then applied unchanged to the testing data. The engineered `LogAge` variable handled the strong skewness of curing age. For predictors to which a Box-Cox transformation was applied, the transformation parameter was estimated from strictly positive training values. Predictors containing zero values were not subjected to the conventional Box-Cox transformation unless a documented positive offset was defined, because the standard transformation requires positive inputs.

After transformation, every predictor was centered and scaled using the mean and standard deviation estimated from the training subset. For observation i and predictor j , the standardized value was:

$$z_{ij} = \frac{x_{ij} - \mu_j}{s_j}$$

where x_{ij} denotes the transformed predictor value, μ_j is the training-set mean, and s_j is the training-set standard deviation. The same μ_j and s_j values were used to standardize the test observations. This step prevents variables measured on large numerical scales, such as aggregate mass, from dominating the distance calculation.

5. KNN Regression Model

For a standardized query observation q , Euclidean distance to each training observation i was calculated across p predictors as:

$$d(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{\sum_{j=1}^p (z_{ij} - z_{qj})^2}$$

The training observations were ordered from the smallest to the largest distance. If $N_k(q)$ denotes the set of k nearest training observations, the unweighted KNN regression estimate was calculated as:

$$\hat{y}_q = \frac{1}{k} \sum_{i \in \mathcal{N}_k(q)} y_i$$

where y_i is the observed compressive strength of neighbor i . Thus, nearby observations contribute equally to the estimate, while observations outside the selected neighborhood do not contribute. A small k gives a flexible local fit but can be sensitive to noise; a large k produces smoother estimates but may obscure local mixture-strength relationships.

6. Hyperparameter Tuning

The value of k was selected using repeated cross-validation conducted only within the training subset. Odd candidate values from $k = 1$ to $k = 41$ were evaluated to cover highly local through progressively smoother neighborhoods. Ten-fold cross-validation was repeated five times: in each repeat, nine folds were used to fit the preprocessing steps and KNN model, and the remaining fold was used for validation. The procedure was rotated across all folds, and validation RMSE was averaged over the 50 resamples. The candidate with the smallest mean cross-validated RMSE was chosen and used to refit the final model on the complete training subset. Repetition reduces dependence on a single fold assignment, although uncertainty can remain when validation samples are limited (Vabalas et al., 2019).

7. Model Evaluation and Diagnostic Analysis

The finalized model was evaluated once on the held-out test subset. Predictive performance was summarized using root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). For n test observations, observed strengths y_i , predicted strengths $\hat{y}_i$, and mean observed strength $\bar{y}$, the metrics were defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad \text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

RMSE is expressed in MPa and gives greater weight to large errors, whereas MAE reports the average absolute deviation in MPa and is less influenced by individual large errors. R^2 describes the proportion of test-set variation represented by the predictions relative to prediction by the test-set mean; it was interpreted together with RMSE and MAE rather than as a stand-alone measure. Actual-versus-predicted and residual plots were also examined to identify systematic underprediction, overprediction, or unusually large errors.

Absolute Pearson correlations between individual predictors and compressive strength were reported as an exploratory ranking of marginal linear association. Because KNN predictions depend jointly on distances across all predictors, these correlations are not model-specific feature-importance scores and should not be interpreted as causal effects.

8. Computational Reproducibility

The workflow was implemented in R. The `distinct()` function was used to remove repeated observations, while `createDataPartition()` and the resampling and training functions were used for data partitioning, repeated cross-validation, and KNN fitting. The random seed, R version, package names, and package versions used for the final run should be retained with the analysis script so that the reported split, tuned k , and test metrics can be independently reproduced. No test observation was used to determine transformation parameters or select k . The caret framework provides the resampling and tuning infrastructure used in this workflow (Kuhn, 2008).

Results and Discussion

1. Descriptive Statistics

The descriptive analysis was conducted on 1,005 unique observations after 25 repeated records were removed from the original 1,030 observations. Each observation contains eight original predictors representing mixture composition and curing age, together with concrete compressive strength as the response variable. Table 1 summarizes the distribution, central tendency, dispersion, and shape of the nine original variables before the engineered predictors were added. This distinction between the raw and cleaned sample sizes resolves the apparent discrepancy between the source dataset and the observations used for modeling.

Table 1. Descriptive Statistics of the Cleaned Data

	<i>Cement</i>	<i>Slag</i>	<i>FlyAsh</i>	<i>Water</i>	<i>SP</i>	<i>CoarseAgg</i>	<i>FineAgg</i>	<i>Age</i>	<i>Strength</i>
<i>vars</i>	1	2	3	4	5	6	7	8	9
<i>n</i>	1005	1005	1005	1005	1005	1005	1005	1005	1005
<i>mean</i>	278.63	72.04	55.54	182.07	6.03	974.38	772.69	45.86	35.25
<i>sd</i>	104.35	86.17	64.21	21.34	5.92	77.58	80.34	63.73	16.28
<i>median</i>	265.0	20.0	0.0	185.7	6.1	968.0	780.0	28.0	33.8
<i>trimmed</i>	270.19	60.01	48.49	181.78	5.38	975.30	775.42	32.51	34.48
<i>mad</i>	110.69	29.65	0.00	18.83	8.24	68.64	67.75	31.13	15.90
<i>min</i>	102.00	0.00	0.00	121.75	0.00	801.00	594.00	1.00	2.33
<i>max</i>	540.0	359.4	200.1	247.0	32.2	1145.0	992.6	365.0	82.6
<i>range</i>	438.00	359.40	200.10	125.25	32.20	344.00	398.60	364.00	80.27
<i>skew</i>	0.56	0.85	0.50	0.03	0.98	-0.07	-0.25	3.24	0.39
<i>kurtosis</i>	-0.44	-0.42	-1.37	0.15	1.67	-0.59	-0.12	11.87	-0.32
<i>se</i>	3.29	2.72	2.03	0.67	0.19	2.45	2.53	2.01	0.51

Table 1 shows that compressive strength has a mean of 35.25 MPa, a standard deviation of 16.28 MPa, and values ranging from 2.33 to 82.60 MPa. Its median of 33.80 MPa and skewness of 0.39 indicate only moderate right-skewness in the response. In contrast, curing age is strongly right-skewed (skewness = 3.24) and highly leptokurtic (kurtosis = 11.87): the median is 28 days, while the mean is 45.86 days and the maximum reaches 365 days. This pattern supports the use of LogAge to reduce the influence of very large age values while retaining their order. The zero minima for slag, fly ash, and superplasticizer show that some mixtures did not contain these constituents; consequently, a conventional Box-Cox transformation should not be applied indiscriminately to these variables without a positive-value adjustment. The predictors also differ substantially in scale. For example, the mean coarse-aggregate content is 974.38 kg/m³, compared with 6.03 kg/m³ for superplasticizer. Because KNN calculates distance across predictors, centering and scaling are essential to prevent high-magnitude variables from dominating neighbor selection. These descriptive characteristics therefore provide empirical support for the preprocessing and feature-engineering decisions described in the Method section.

2. Model Optimization Results (*Tuning k*)

The number of nearest neighbors was tuned using repeated 10-fold cross-validation within the training data. Odd candidate values from $k = 1$ to $k = 41$ were compared using cross-validated RMSE, and the candidate producing the smallest mean validation error was selected. Figure 3 presents the relationship between k and the resampled RMSE.

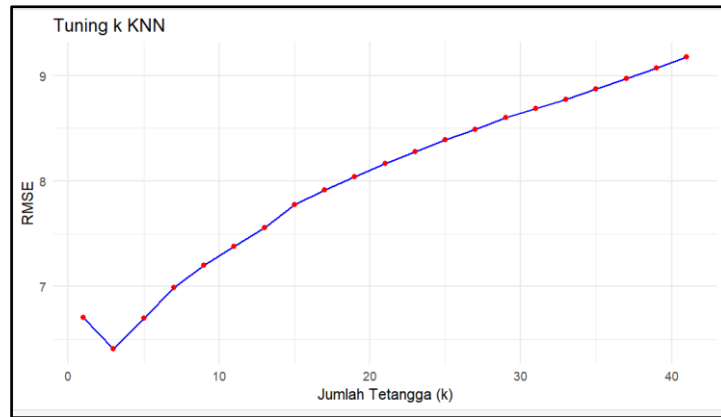


Figure 3. Optimization of k in the KNN Model

The minimum cross-validated RMSE occurred at $k = 3$. This result indicates that, within the candidate grid and resampling design used in this study, predictions based on a small local neighborhood provided the best balance between sensitivity and smoothing. Values larger than three produced progressively higher RMSE, suggesting that averaging over broader neighborhoods blurred local differences among mixtures with distinct proportions and curing ages. The term “optimal” is therefore conditional on this dataset, preprocessing pipeline, candidate range, and resampling scheme; it should not be interpreted as a universally optimal k for all concrete-strength datasets. The selected value was subsequently used to refit the model on the complete training subset before evaluation on the held-out test subset.

3. Variable Importance Plot

To describe the marginal association between each predictor and compressive strength, absolute Pearson correlations were calculated and ranked as shown in Figure 4. This graph is retained as an exploratory predictor-ranking plot. It does not measure the change in KNN performance caused by removing a variable and is therefore not a model-specific feature-importance analysis.

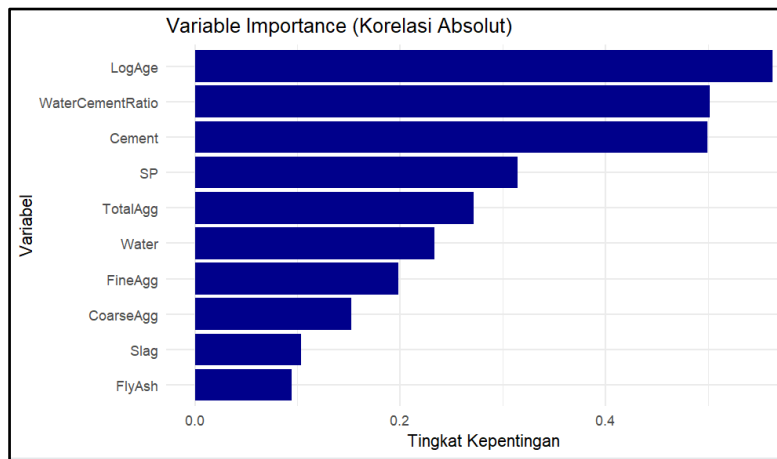


Figure 4. Absolute Correlation of Predictors with Compressive Strength

LogAge has the largest absolute correlation with compressive strength, followed by Water-Cement Ratio and cement content. The ranking is physically plausible: curing time represents strength development through hydration, whereas the water-cement balance and cement content characterize central aspects of mixture composition. The result also supports the decision to represent age on a logarithmic scale, because strength gain is generally more rapid at early ages and slows as curing continues. Nevertheless, the ranking should be interpreted as marginal association rather than causal influence or direct contribution to predictive accuracy. Correlated

variables may share information, and KNN forms predictions from their joint standardized distance rather than from separate regression coefficients.

4. Final Evaluation Metrics

Model performance was summarized using RMSE, MAE, and R^2 on the training and held-out testing subsets (Table 2). The training metrics describe in-sample fit after model selection, whereas the testing metrics provide the primary estimate of predictive performance for unseen observations from the same data source. Comparing the two subsets helps identify a generalization gap, but the test results (not the training results) should be emphasized when reporting model performance.

Table 2. Model Evaluation Metrics Results

Evaluation Metrics	Data Training	Data Testing
RMSE	3.92	5.56
MAE	2.81	3.74
R^2	0.94	0.88

The test-set R^2 is 0.88, compared with 0.94 for the training data. Thus, the predictions account for 88% of the observed variation in compressive strength in the held-out subset relative to the test-set mean. The reduction of 0.06 from training to testing, together with the higher test errors, indicates a moderate and expected generalization gap. It does not by itself prove the absence of overfitting, but the retained test performance suggests that the model continues to capture a substantial portion of the strength variation in observations that were not used for fitting or tuning.

On the testing data, RMSE is 5.56 MPa and MAE is 3.74 MPa, compared with training values of 3.92 MPa and 2.81 MPa, respectively. MAE means that the prediction differs from the observed strength by 3.74 MPa on average in absolute terms. RMSE should not be interpreted as a second average-error bound; its larger value reflects the additional weight assigned to relatively large errors. The RMSE-to-MAE difference of 1.82 MPa is consistent with the presence of several observations whose errors are larger than the typical absolute deviation. Collectively, the three test metrics indicate useful internal predictive performance, while their interpretation remains conditional on the present train-test partition and UCI dataset.

5. Prediction vs Actual Plot (Training & Testing)

Figure 5 compares observed and predicted compressive strength for the training and testing subsets. Agreement is evaluated relative to the 45-degree reference line ($y = x$): observations close to this line have small prediction errors, while vertical deviations indicate underprediction or overprediction. Different colors distinguish fitted training observations from genuinely held-out test predictions.

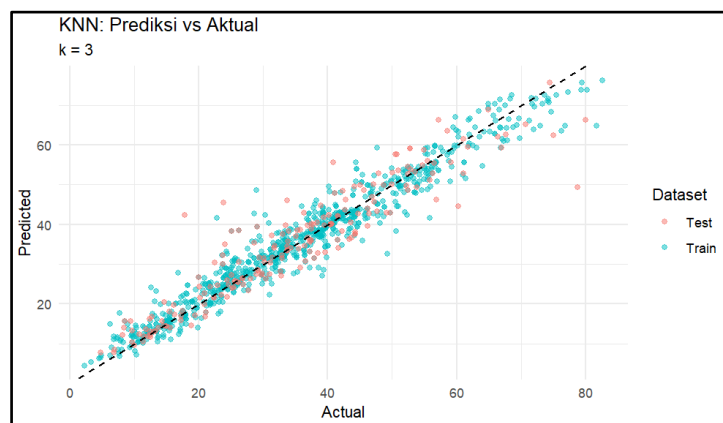


Figure 5. Actual vs Predicted Data Plot

Most observations lie near the reference line, particularly in the middle of the observed strength range, which is consistent with the test R^2 of 0.88. Dispersion is more visible for some high-strength observations, indicating that local neighborhoods contain fewer closely comparable mixtures at the upper end of the response distribution. The plot also shows a wider spread for the test subset than for the training subset, consistent with the differences in RMSE and MAE reported in Table 2. The figure supports the adequacy of $k = 3$ for this internal test set, but it does not establish unbiased performance in other laboratories or mixture domains.

6. Residual Plot

Residuals, calculated as observed minus predicted strength, were plotted against predicted values to examine the direction, spread, and possible structure of the prediction errors (Figure 6). For a nonparametric prediction method such as KNN, this diagnostic is used to detect systematic error patterns rather than to claim that all assumptions of a linear regression model are satisfied.

Figure 6 shows that residuals are distributed on both sides of the zero line without a pronounced monotonic or curved pattern. This provides no strong visual evidence of systematic underprediction or overprediction across most fitted values. Some larger residuals remain, which is consistent with RMSE exceeding MAE. The plot should not be treated as a formal test of constant variance, and homoscedasticity is not a prerequisite for KNN regression. Its principal value here is to identify ranges in which errors widen and observations that may warrant further investigation.

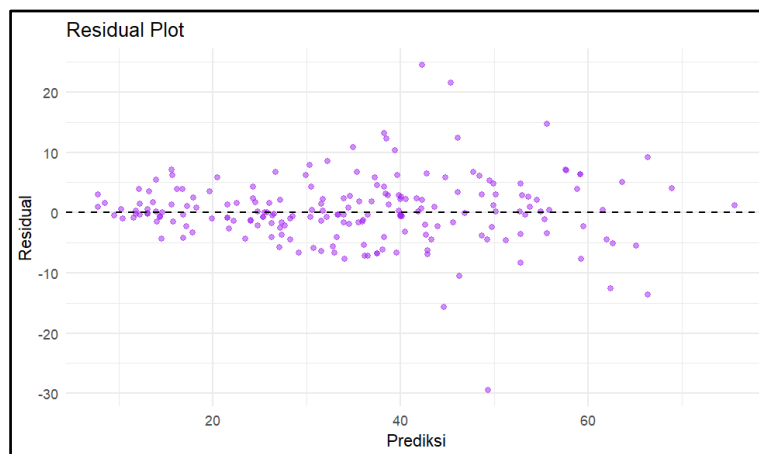


Figure 6. Residual Plot

7. Sample Prediction Results

Table 3 presents the first 10 observations from the held-out testing subset. The examples are provided to show prediction behavior at the observation level and are descriptive rather than a substitute for the aggregate metrics in Table 2. Error Difference is reported as the absolute difference between observed and predicted strength in MPa.

Table 3. Comparison of Prediction Results

No	Actual Strength	Predicted Strength	Error Difference
1	79.98611	66.34368	13.64243
2	36.44777	34.52137	1.92640
3	28.02168	26.25594	1.76574
4	42.32693	48.40145	6.07452
5	40.56325	40.10590	0.45735
6	37.42752	40.04086	2.61334
7	52.90832	53.79039	0.88207
8	41.71950	41.87372	0.15422
9	41.54300	42.22880	0.68580
10	35.07640	36.90558	1.82918

The absolute errors in these 10 examples range from 0.15 to 13.64 MPa. Rows 5, 7, 8, and 9 have errors below 1 MPa, whereas row 1 has the largest error (13.64 MPa) and row 4 has an error of 6.07 MPa. Because no inferential significance test was performed for individual predictions, the differences should be described by their magnitudes rather than as statistically significant or nonsignificant. These cases illustrate why MAE and RMSE should be reported together: most predictions can be close to their observed values while a smaller number of large deviations increase RMSE.

Discussion in Relation to Previous Studies

The present findings support the broader evidence that mixture composition and curing age contain substantial information for data-driven concrete-strength prediction. Earlier studies using the same or related concrete datasets have likewise reported strong predictive performance from machine-learning models (Al-Shamiri et al., 2020; Chou et al., 2014; DeRousseau et al., 2019; Wan et al., 2021; Young et al., 2019). The result is also consistent with the use of an instance-based learning approach for concrete and rock strength prediction reported (Hsieh, 2021). The contribution of the present analysis is not to claim that KNN universally outperforms more complex algorithms, but to show that a comparatively transparent local-learning model can achieve $R^2 = 0.88$ on held-out data when domain-relevant features, distance-based preprocessing, and cross-validated tuning are combined.

Direct numerical comparison with previous publications must be made cautiously. Reported RMSE and R^2 values depend on the exact sample retained, duplicate handling, training-test allocation, transformations, feature set, tuning grid, and validation protocol. Studies based on random forests, boosting, neural networks, or ensemble methods may obtain different performance because they model interactions and nonlinearities differently (Asteris et al., 2021; Behnood & Golafshani, 2018; Fei et al., 2023; Li et al., 2023; Pakzad et al., 2023; Rathakrishnan et al., 2022; Zhang et al., 2024). Accordingly, the current results establish an interpretable KNN benchmark under the stated workflow rather than evidence of superiority over all alternative models. A stronger comparative claim would require the competing algorithms to be trained and evaluated on the identical resamples and test observations.

The practical interpretation is that the model may assist preliminary mixture screening or provide a rapid estimate when the eight mixture and curing variables fall within the domain represented by the dataset. Prediction quality may deteriorate for uncommon high-strength mixtures, new material sources, different testing standards, or predictor combinations lying far from the training observations. Consequently, the model should complement, not replace, standardized compressive-strength testing. Its predictions are most defensible as decision support before laboratory results become available.

Limitations and Further Validation

Several limitations qualify the findings. First, the evaluation is an internal validation based on a single public dataset and one held-out partition; no independent laboratory or field dataset was used. Second, uncertainty intervals for RMSE, MAE, and R^2 were not estimated, so sampling variability around the reported values remains unknown. Third, absolute correlation provides only a univariate descriptive ranking and does not quantify KNN-specific feature contribution. Fourth, the final model was not compared under an identical validation design with simpler baselines or alternative machine-learning algorithms. Future work should use repeated nested validation or an external dataset, report uncertainty intervals, evaluate permutation-based importance or ablation analysis, and compare KNN against a mean predictor, linear regression, tree-based ensembles, and other established models. These extensions would clarify both the robustness of the selected k and the incremental benefit of the engineered features.

Conclusion

This study developed a KNN regression model for estimating concrete compressive strength using 1,005 unique observations, eight original mixture and age variables, and three engineered predictors: LogAge, Water-Cement Ratio, and Total Aggregate. Repeated 10-fold cross-validation identified $k = 3$ as the best candidate within the evaluated range of odd values from 1 to 41. On the held-out testing data, the model achieved an R^2 of 0.88, RMSE of 5.56 MPa, and MAE of 3.74 MPa. These results indicate that the model captured a substantial proportion of test-set variation while maintaining an average absolute prediction error of 3.74 MPa. The difference between the training and testing results represents a moderate generalization gap rather than definitive evidence that overfitting is absent. LogAge showed the largest absolute correlation with strength, followed by Water-Cement Ratio and cement content, supporting the relevance of curing time and mixture balance. These correlations, however, should be interpreted as marginal associations rather than causal effects or KNN-specific feature importance. Practically, the model can support preliminary mixture screening and rapid quality-control decisions when new observations fall within the data domain represented in the training set. It should complement, not replace, standardized laboratory testing. Because the evaluation used a single public dataset and internal train-test validation, future studies should assess external laboratory or field data, quantify uncertainty in the performance metrics, evaluate feature contributions through permutation or ablation analysis, and compare KNN with baseline, tree-based, ensemble, and other established regression models under an identical validation design.

Acknowledgment

-

References

- Al-Shamiri, A. K., Yuan, T.-F., & Kim, J. H. (2020). Non-tuned machine learning approach for predicting the compressive strength of high-performance concrete. *Materials*, 13(5), 1023. <https://doi.org/10.3390/ma13051023>
- An, Y., Xu, M., & Shen, C. (2019). Classification method of teaching resources based on improved KNN algorithm. *International Journal of Emerging Technologies in Learning (ijET)*, 14(4), 73–88. <https://doi.org/10.3991/ijet.v14i04.10131>
- Arman, A. (2018). Kajian kuat tekan beton normal menggunakan standar SNI 7656-2012 dan ASTM C 136-06. *Rang Teknik Journal*, 1(2), 142–148.
- Asteris, P. G., Skentou, A. D., Bardhan, A., Samui, P., & Pilakoutas, K. (2021). Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. *Cement and Concrete Research*, 145, 106449. <https://doi.org/10.1016/j.cemconres.2021.106449>
- Banjarsari, M. A., Budiman, H. I., & Farmadi, A. (2015). Penerapan K-optimal pada algoritma KNN untuk prediksi kelulusan tepat waktu mahasiswa Program Studi Ilmu Komputer FMIPA UNLAM berdasarkan IP sampai dengan semester 4. *Kumpulan Jurnal Ilmu Komputer*, 2(2), 50–64.
- Behnood, A., & Golafshani, E. M. (2018). Predicting the compressive strength of silica fume concrete using hybrid artificial neural network with multi-objective grey wolves. *Journal of Cleaner Production*, 202, 54–64. <https://doi.org/10.1016/j.jclepro.2018.08.065>
- Behnood, A., Behnood, V., Modiri Gharehveran, M., & Alyamac, K. E. (2017). Prediction of the compressive strength of normal and high-performance concretes using M5P model tree

- algorithm. *Construction and Building Materials*, 142, 199–207.
<https://doi.org/10.1016/j.conbuildmat.2017.03.061>
- Chou, J.-S., Tsai, C.-F., Pham, A.-D., & Lu, Y.-H. (2014). Machine learning in concrete strength simulations: Multi-nation data analytics. *Construction and Building Materials*, 73, 771–780.
<https://doi.org/10.1016/j.conbuildmat.2014.09.054>
- Daulay, R. S. A. (2024). Analisis kritis dan pengembangan algoritma K-nearest neighbor (KNN): Sebuah tinjauan literatur. *Jurnal Pendidikan Sains dan Komputer*, 4(2), 131–141.
- DeRousseau, M. A., Laftchiev, E., Kasprzyk, J. R., Rajagopalan, B., & Srubar, W. V., III. (2019). Machine learning methods for predicting the field compressive strength of concrete. *Construction and Building Materials*, 228, 116661.
<https://doi.org/10.1016/j.conbuildmat.2019.08.042>
- Fadilla, R., Dewi, K., & Gusmana, R. (2018). Implementasi metode K-nearest neighbor (KNN) dalam pengelompokan status ekonomi warga. *Jurnal Big Data Analisis dan Artificial Intelligence*, 4(1), 15–22.
- Fei, Z., Liang, S., Cai, Y., & Shen, Y. (2023). Ensemble machine-learning-based prediction models for the compressive strength of recycled powder mortar. *Materials*, 16(2), 583.
<https://doi.org/10.3390/ma16020583>
- Herdiansyah, & Pangaribuan, M. R. (2013). Pengaruh batu cadas (batu trass) sebagai bahan pembentuk beton terhadap kuat tekan beton. *Jurnal Inersia*, 5(2), 11–19.
- Hsieh, S.-C. (2021). Prediction of compressive strength of concrete and rock using an elementary instance-based learning algorithm. *Advances in Civil Engineering*, 2021, 6658932.
<https://doi.org/10.1155/2021/6658932>
- Ihzaniah, L. S., Setiawan, A., & Wijaya, R. W. N. (2023). Perbandingan kinerja metode regresi K-nearest neighbor dan metode regresi linear berganda pada data Boston Housing. *Jambura Journal of Probability and Statistics*, 4(1), 17–29. <https://doi.org/10.34312/jjps.v4i1.18948>
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. <https://doi.org/10.1016/j.patter.2023.100804>
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1–26. <https://doi.org/10.18637/jss.v028.i05>
- Li, D., Tang, Z., Kang, Q., Zhang, X., & Li, Y. (2023). Machine learning-based method for predicting compressive strength of concrete. *Processes*, 11(2), 390.
<https://doi.org/10.3390/pr11020390>
- Nanja, M., & Purwanto. (2015). Metode K-nearest neighbor berbasis forward selection untuk prediksi harga komoditi lada Muis. *Jurnal Pseudocode*, 2(1), 53–64.
- Pakzad, S. S., Roshan, N., & Ghalehnovi, M. (2023). Comparison of various machine learning algorithms used for compressive strength prediction of steel fiber-reinforced concrete. *Scientific Reports*, 13, 3646. <https://doi.org/10.1038/s41598-023-30606-y>
- Prasetyo, Y. A., Utami, E., & Yaqin, A. (2024). Pengaruh komposisi split data terhadap performa akurasi analisis sentimen algoritma Naïve Bayes dan SVM. *Journal of Electrical Engineering and Computer Science*, 6(2), 382–390.
- Rahardja, C. A., Juardi, T., & Agung, H. (2019). Implementasi algoritma K-nearest neighbor pada website rekomendasi laptop. *Jurnal Buana Informatika*, 10(1), 75–84.
<https://doi.org/10.24002/jbi.v10i1.1847>

- Rathakrishnan, V., Beddu, S., & Ahmed, A. N. (2022). Predicting compressive strength of high-performance concrete with high volume ground granulated blast-furnace slag replacement using boosting machine learning algorithms. *Scientific Reports*, 12, 9539. <https://doi.org/10.1038/s41598-022-12890-2>
- Risdiyanto, Y. (2013). Kajian kuat tekan beton dengan perbandingan volume dan perbandingan berat untuk produksi beton massa menggunakan agregat kasar batu pecah Merapi (Studi kasus pada proyek pembangunan Sabo Dam). *Jurnal Tugas Akhir*, 1–11.
- Ritonga, M. R. R., Sihombing, M., & Selfira. (2024). Pemodelan K-nearest neighbor untuk identifikasi pola kepuasan mahasiswa terhadap pelayanan kampus (Studi kasus: STMIK Kaputama). *Jurnal Informatika dan Sains Teknologi*, 2(4), 152–161.
- Sapkota, S. C., Panagiotakopoulou, C., Dahal, D., Beskopylny, A. N., Dahal, S., & Asteris, P. G. (2025). Optimizing high-strength concrete compressive strength with explainable machine learning. *Multiscale and Multidisciplinary Modeling, Experiments and Design*, 8, 160. <https://doi.org/10.1007/s41939-025-00737-y>
- Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, 14(11), e0224365. <https://doi.org/10.1371/journal.pone.0224365>
- Wan, Z., Xu, Y., & Savija, B. (2021). On the use of machine learning models for prediction of compressive strength of concrete: Influence of dimensionality reduction on the model performance. *Materials*, 14(4), 713. <https://doi.org/10.3390/ma14040713>
- 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](https://doi.org/10.1016/S0008-8846(98)00165-3)
- Yeh, I.-C. (2007). *Concrete compressive strength* [Data set]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5PK67>
- Young, B. A., Hall, A., Pilon, L., Gupta, P., & Sant, G. (2019). Can the compressive strength of concrete be estimated from knowledge of the mixture proportions? New insights from statistical analysis and machine learning methods. *Cement and Concrete Research*, 115, 379–388. <https://doi.org/10.1016/j.cemconres.2018.09.006>
- Zhang, Y., Ren, W., Chen, Y., Mi, Y., Lei, J., & Sun, L. (2024). Predicting the compressive strength of high-performance concrete using an interpretable machine learning model. *Scientific Reports*, 14, 28346. <https://doi.org/10.1038/s41598-024-79502-z>