Journal of Statistics and Its Application, Likelihood Vol 2 No 1, June 2026

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

Delfi Rosaria Simanjuntak [1], Aditama Rouliber Simanulang [2], Riwi Dyah Pangesti [3*], Wina Ayu Lestari [4], Winalia Agwil [5]

[1] University of Bengkulu, Indonesia, email: drsimanjuntak@unib.ac.id
[2] University of Bengkulu, Indonesia, email: arsimanulang@unib.ac.id
[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 one of the primary parameters determining the quality and suitability of concrete in construction. Conventional compressive strength testing requires considerable time, cost, and specialized laboratory equipment; therefore, a more efficient and accurate alternative approach is needed. This study aims to develop a concrete compressive strength prediction system using the K-Nearest Neighbor (KNN) Regression method with k-parameter tuning. The dataset used is the Concrete Compressive Strength Dataset from the UCI Machine Learning Repository, consisting of 1,030 samples with eight input variables and one target variable. The research stages include feature engineering, data preprocessing, train-test data splitting, KNN model development, and optimization of the k parameter using the 10-fold Repeated Cross-Validation method. Model performance is evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The results show that the optimal k value is k = 3, yielding the best predictive performance with an R^2 of 0.88 on the test data, an RMSE of 5.56 MPa, and an MAE of 3.74 MPa. The engineered features LogAge and Water-Cement Ratio are identified as the most influential factors affecting concrete compressive strength. Overall, the findings demonstrate that KNN Regression with k-parameter tuning can provide accurate and stable predictions of concrete compressive strength. This model is expected to serve as a practical alternative for supporting concrete quality control and decision-making processes in construction more quickly and efficiently.

Keywords: *Concrete Compressive Strength, KNN Regression, k-Parameter Tuning, Machine Learning, Concrete Prediction.*

Introduction

Concrete is a construction material composed of a cementitious binder and aggregates such as sand and crushed stone (gravel). It is widely used in various infrastructure projects, including residential buildings, high-rise structures, bridges, and road pavements [1]. One of the primary indicators determining concrete quality is compressive strength, which is used to assess whether the produced concrete meets the planned design specifications. Through compressive strength testing, it can be determined whether the concrete is suitable for further use or if additional evaluation of the structure is required [2]. Therefore, accurate estimation of concrete compressive strength is essential to ensure construction safety, efficiency, and sustainability.

Concrete compressive strength is defined as the ability of concrete to withstand compressive forces per unit area. This parameter reflects the overall quality and performance of a concrete

https://doi.org/10.30605/likelihood.1461

structure. The higher the required compressive strength, the higher the concrete quality that must be achieved [3]. Compressive strength testing is generally conducted using a compression testing machine, in which the specimen is placed centrally on the loading surface and subjected to a gradually increasing load until failure occurs. The maximum load sustained by the concrete is then recorded in specific units as the test result [4]. The testing procedure requires considerable time, cost, and specialized equipment, thereby motivating the need for alternative methods capable of predicting concrete compressive strength more quickly, efficiently, and with maintained accuracy.

Advancements in computational technology have opened opportunities for the application of machine learning methods in civil engineering, particularly for modeling non-linear relationships between input and output variables. One relatively simple yet effective method is K-Nearest Neighbor Regression (KNN Regression). Concrete exhibits complex and non-linear relationships between its material composition and mechanical properties, making conventional evaluation of compressive strength dependent on multiple mechanical tests to obtain accurate results [5]. In this context, data-driven approaches such as KNN Regression offer a promising alternative.

The KNN algorithm is an instance-based learning method that predicts the value of new data based on its proximity to a number of nearest training data points. In KNN Regression, the predicted value is obtained from the average or a weighted combination of the target values of the k nearest neighbors in the feature space [6]. This method is chosen because it offers several advantages, including ease of implementation and the ability to produce accurate predictions when supported by a sufficient amount of training data [7]. Therefore, KNN Regression is widely applied in regression problems, including the prediction of mechanical properties of construction materials.

However, the performance of the KNN Regression algorithm is highly influenced by the selection of the parameter k, which represents the number of nearest neighbors used in the prediction process [8]. A value of k that is too small can make the model sensitive to noise, while a value that is too large may obscure local patterns in the data, thereby reducing prediction accuracy. Therefore, determining the optimal value of k becomes a crucial aspect in improving model performance. One commonly used approach to select the best k is through parameter optimization techniques, such as cross-validation. To date, the optimal value of k for the case of concrete compressive strength prediction has not been definitively established [9].

Based on this background, this study proposes a concrete compressive strength prediction system using the KNN Regression method with a k-parameter tuning approach. The system utilizes concrete characteristic data as input variables to generate accurate estimates of compressive strength. It is expected that the results of this study will contribute to the development of data-driven methods for predicting concrete quality and serve as a supporting alternative for practitioners and researchers in planning, quality control, and decision-making processes in the construction field.

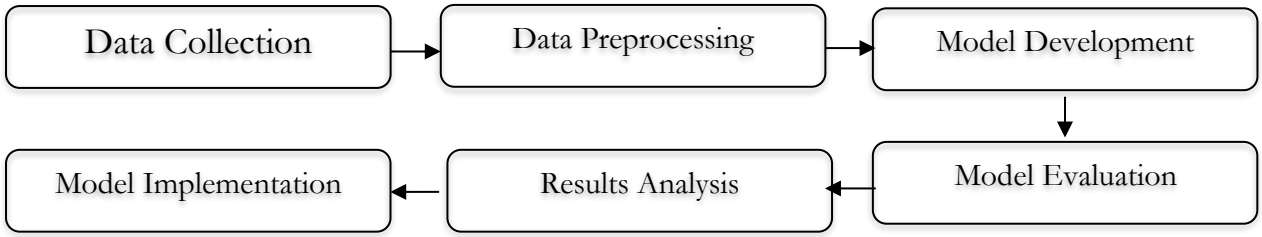
Method

The K-Nearest Neighbor (KNN) algorithm is a supervised learning method, which is an approach that utilizes labeled data to build predictive patterns. The main objective of supervised learning is to learn the relationship between input and output variables so that it can be used to predict new values, whereas unsupervised learning aims to discover hidden patterns or structures in data without using labels [12]. In KNN Regression, the prediction process is carried out by

utilizing the proximity between the test data and the training data. The stages of the KNN Regression algorithm are as follows [13]:

1. Determine the value of the parameter *k* as the number of nearest neighbors used in the prediction process,
2. Calculate the distance between the data to be predicted and all training data using a specified distance metric,
3. Sort the calculated distances in ascending order and select the *k* data points with the smallest distances,
4. Determine the predicted value by calculating the average of the target values from the *k* nearest neighbors.

This study aims to develop a predictive model capable of rapidly estimating concrete compressive strength based on the mixture proportions of its constituent materials. The proposed model is expected to serve as a decision-support tool in field applications, thereby reducing reliance on laboratory testing, which requires relatively significant time and cost. The research workflow is as follows:



Picture 1. Research Flow Diagram

1. Data Collection

The data used in this study is the Concrete Compressive Strength dataset obtained from the UCI Machine Learning Repository. This dataset consists of 1,030 concrete samples with eight input variables, including cement content, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and concrete age, as well as one target variable, namely concrete compressive strength (Strength) measured in MPa.

```
# A tibble: 1,005 × 9
   Cement  Slag FlyAsh Water    SP CoarseAgg FineAgg   Age Strength
    <db1> <db1>  <db1> <db1> <db1>     <db1>   <db1> <db1>    <db1>
 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
```

Pictur 2. Research Data

2. *Feature Engineering*

For imbalanced data, feature engineering is required to improve data quality. First, data cleaning is performed by removing duplicate observations using the *distinct()* function to maintain the integrity of prediction results. Second, based on descriptive statistical analysis, the Age variable

exhibits a highly skewed distribution (skewness = 3.24); therefore, a LogAge transformation using the natural logarithm is applied to normalize the data distribution. In addition, new features are constructed, including the Water-Cement Ratio and Total Aggregate (TotalAgg), to better capture the physical relationships among the constituent materials of concrete in a more comprehensive manner and in accordance with concrete technology standards.

3. *Preprocessing* dan *Splitting* Data

Data preprocessing is a crucial stage in data mining, as not all data or attributes in a dataset are relevant for analysis. This process aims to ensure that the data used is appropriate and ready to support the intended analysis [14]. An appropriate split between training and testing data is essential to ensure that the developed model achieves good performance and avoids overfitting. Overfitting occurs when a model becomes too closely fitted to the training data, resulting in poor performance on new data. Therefore, selecting a proper data-splitting ratio helps produce a more generalized model with better predictive accuracy [15]. The preprocessing stage is essential before applying the KNN Regression algorithm. In this study, the data undergoes a Box-Cox transformation to address non-linearity, along with centering and scaling procedures. This aims to standardize the scale of all variables (for example, the CoarseAgg feature with values in the thousands and SP with values in single digits), ensuring that each feature contributes equally in the Euclidean distance calculation. After preprocessing, the dataset is split using the *createDataPartition()* function, with 80% allocated for training data and 20% for testing data. This division is intended to build the model on the training data and validate its accuracy on previously unseen data.

4. Model KNN dan *Tuning* Parameter k

The K-Nearest Neighbor (KNN) Regression algorithm is applied in this study with a primary focus on optimizing the number of nearest neighbors (k). The model predicts concrete compressive strength by calculating the average target value of the k nearest samples using the Euclidean distance metric. To achieve maximum accuracy, a k-parameter tuning approach is conducted over a range of values from 1 to 41 with increments of 2. The search for the optimal k is performed within a 10-fold Repeated Cross-Validation scheme (5 repetitions), where the best model is selected based on the smallest Root Mean Squared Error (RMSE) obtained from the training data.

Results and Discussion

1. Deskriptif Statistics

Before conducting the analysis, the primary step is to understand the overview and characteristics of the data. The Concrete Compressive Strength dataset used consists of 1,030 samples with eight input variables representing material composition and concrete age, as well as one target variable, namely Strength (concrete compressive strength in MPa). The descriptive statistics for each variable are presented in the Table 1.

Table 1. Statistik Deskriptif Data

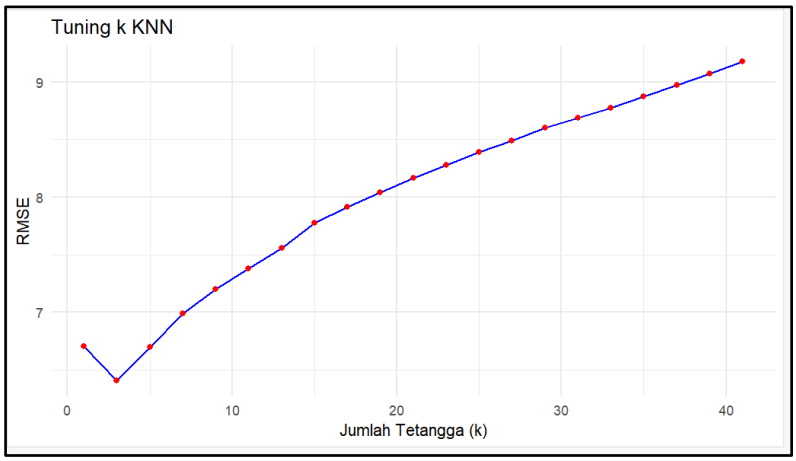
	Cement	*Slag*	*FlyAsh*	*Water*	*SP*	*CoarseAgg*	*FineAgg*	*Age*	*Strength*
vars	1	2	3	4	5	6	7	8	9
n	1005	1005	1005	1005	1005	1005	1005	1005	1005
mean	278.63	72.04	55.54	182.07	6.03	974.38	772.69	45.86	35.25
sd	104.35	86.17	64.21	21.34	5.92	77.58	80.34	63.73	16.28
median	265.0	20.0	0.0	185.7	6.1	968.0	780.0	28.0	33.8

	Cement	*Slag*	*FlyAsh*	*Water*	*SP*	*CoarseAgg*	*FineAgg*	*Age*	*Strength*
trimmed	270.19	60.01	48.49	181.78	5.38	975.30	775.42	32.51	34.48
mad	110.69	29.65	0.00	18.83	8.24	68.64	67.75	31.13	15.90
min	102.00	0.00	0.00	121.75	0.00	801.00	594.00	1.00	2.33
max	540.0	359.4	200.1	247.0	32.2	1145.0	992.6	365.0	82.6
range	438.00	359.40	200.10	125.25	32.20	344.00	398.60	364.00	80.27
skew	0.56	0.85	0.50	0.03	0.98	-0.07	-0.25	3.24	0.39
kurtosis	-0.44	-0.42	-1.37	0.15	1.67	-0.59	-0.12	11.87	-0.32
se	3.29	2.72	2.03	0.67	0.19	2.45	2.53	2.01	0.51

Based on Table 1, the Strength variable has an average value of 35.25 MPa, with a range from 2.33 MPa to 82.6 MPa. The Age variable indicates that most concrete samples are relatively young, and its distribution is highly non-normal, with a skewness of 3.24 and kurtosis of 11.87. These results suggest the need for a logarithmic transformation to ensure that the KNN algorithm performs properly without being affected by outliers or extreme data distributions. Furthermore, there is a significant scale difference between the CoarseAgg and SP variables, with average values of 974.38 and 6.03, respectively. Therefore, feature engineering through scaling and standardization is required to ensure that each feature contributes fairly in the distance calculations within the KNN algorithm.

2. Model Optimization Results (*Tuning k*)

To achieve the highest level of accuracy, the KNN algorithm is evaluated by testing various values of the number of nearest neighbors (*k*) using the 10-fold Repeated Cross-Validation method. The process of identifying the optimal *k* value that minimizes prediction error is visualized in Figure 3.



Figur 3. Graph of k Value Optimization in the KNN Model

Based on the model optimization results using 10-fold Repeated Cross-Validation, the tuning of *k* was conducted over the range of 1 to 41. As shown in the KNN tuning graph, the lowest Root Mean Squared Error (RMSE) is achieved at $k = 3$. At this point, the model demonstrates the most optimal predictive performance with minimum error. Increasing the value of *k* beyond 3 is observed to significantly increase the RMSE, indicating a decline in the model's ability to capture the underlying characteristics of the data.

3. Variable Importance Plot

To identify which features contribute most significantly to the prediction of concrete compressive strength, the absolute correlation between input variables and the target is calculated. The importance level of each variable is presented in Figure 4.

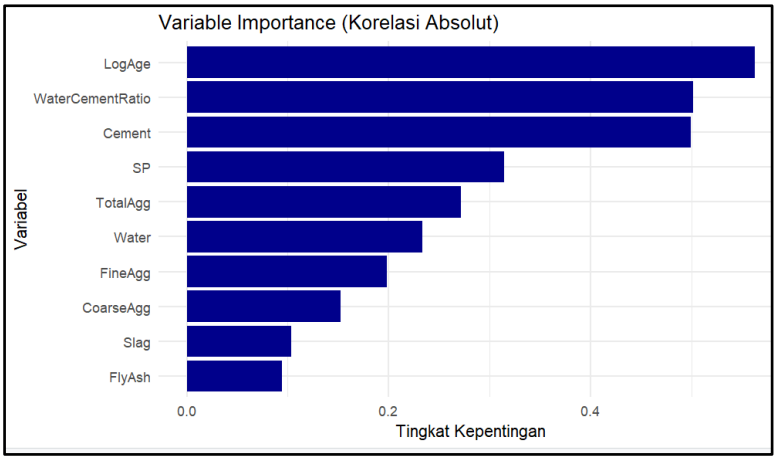


Figure 4. Most Influential Variables

From the graph above, the variable importance analysis shows the contribution of each variable to the accuracy of concrete compressive strength prediction. The engineered feature LogAge is identified as the most dominant predictor, followed by Water Cement Ratio and cement content. This indicates that the hydration time factor and the cement–water proportion are the most influential parameters in determining concrete compressive strength.

4. Final Evaluation Metrics

To assess the accuracy of the K-Nearest Neighbors (KNN) model in predicting concrete compressive strength, performance evaluation is conducted using three main statistical metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The evaluation is performed comparatively on both training and testing data to detect potential overfitting and to ensure the model's ability to generalize to new data. The detailed results of these evaluation metrics are presented in Table 2.

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

Based on Table 2, the KNN model demonstrates strong and stable performance. The coefficient of determination (R^2) on the testing data is 0.88, indicating that the model is able to explain 88% of the variability in concrete compressive strength based on the available features. Although there is a decrease in accuracy from the training data ($R^2 = 0.94$) to the testing data, the difference remains within an acceptable range, suggesting that the model is not overfitted.

From the prediction error results, the RMSE value of 5.56 and MAE of 3.74 on the testing data indicate that the average deviation between predicted and actual concrete strength values ranges from approximately 3.74 to 5.56 MPa. Overall, the evaluation metrics demonstrate that the combination of feature engineering and k-parameter optimization has successfully produced a reliable predictive model for estimating concrete compressive strength.

5. Prediction vs Actual Plot (Training & Testing)

The visualization of the relationship between actual and predicted concrete compressive strength values is conducted to provide a comprehensive view of the model's precision across the entire data range. Through the Prediction vs Actual plot, it can be observed how closely the data

points follow the diagonal line ($y = x$). The use of different colors for training and testing data highlights the consistency of the model's performance across different datasets, as shown in Figure 5.

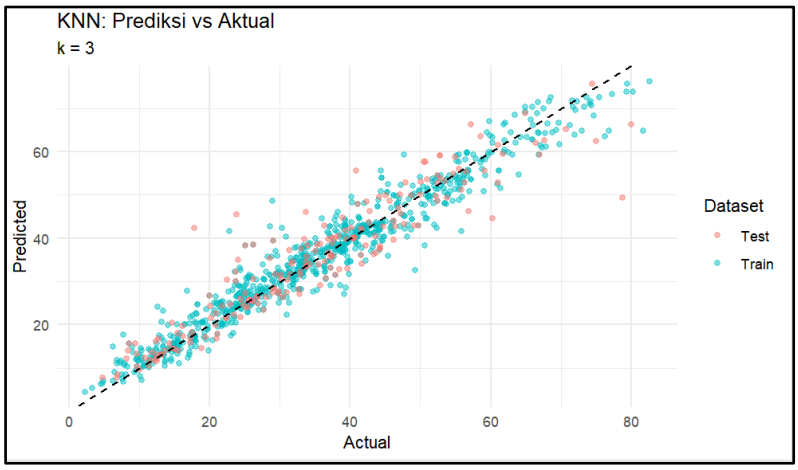


Figure 5. Actual vs Predicted Data Plot

Based on Figure 5, most data points—both in the training and testing datasets—are concentrated along the diagonal dashed line. This pattern indicates a very strong linear correlation between the observed values and the estimates produced by the KNN algorithm. Within the mid-range of concrete strength (20–50 MPa), the model demonstrates very high precision, as evidenced by the minimal dispersion of data points. Although there are a few outliers in the testing data at higher strength levels (above 60 MPa), overall the model is still able to accurately capture the increasing trend of concrete strength. The use of $k = 3$ is proven to be effective in minimizing bias in the testing data without losing important pattern details in the training data, which further supports the high R^2 value obtained.

6. Residual Plot

Residual analysis is conducted to ensure that the prediction errors are random and do not exhibit any systematic patterns that could indicate model weaknesses, as shown in Figure 6.

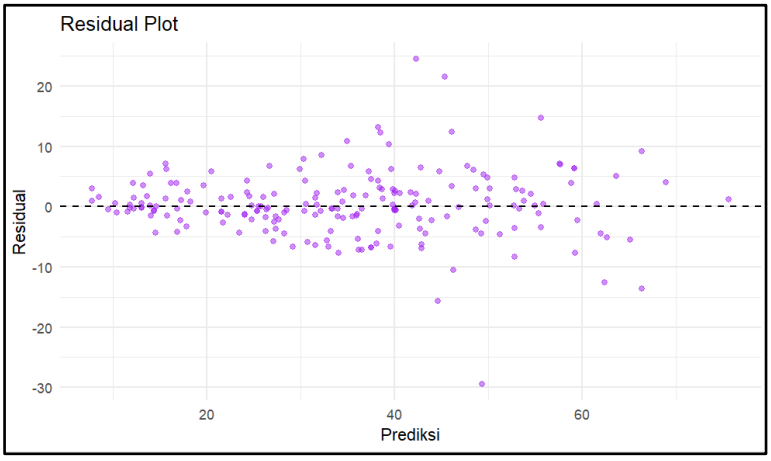


Figure 6. Residual Plot

The results in Figure 4 show that the residuals (the difference between actual and predicted values) are randomly distributed around the horizontal zero line. The absence of specific patterns (such as funnel shapes or curves) indicates that the model satisfies the assumption of homoscedasticity and does not exhibit systematic bias in estimating concrete strength.

7. Sample Prediction Results

The following table presents a comparison of the first 10 samples from the testing data, illustrating the model's prediction precision relative to the actual values.

Table 3. Comparison of Prediction Results

No	Actual Strength	Predicted Strength	Error Difference
1	7998611	6634368	136424318
2	3644777	3452137	19263959
3	2802168	2625594	17657480
4	4232693	4840145	60745134
5	4056325	4010590	0.4573524
6	3742752	4004086	26133439
7	5290832	5379039	0.8820696
8	4171950	4187372	0.1542128
9	4154300	4222880	0.6857988
10	3507640	3690558	18291798

Based on Table 3, the prediction results of the KNN Regression model do not differ significantly from the actual values. However, there are a few instances where the predictions are less accurate, as observed in rows 1 and 4.

Conclusion

Based on the research results, the KNN regression model with feature engineering and k tuning has proven effective in predicting concrete compressive strength, as indicated by an R^2 value of 0.88, RMSE of 5.91 MPa, and MAE of 4.04 MPa, reflecting high accuracy with error deviations still within acceptable limits. The analysis shows that systematic selection of k through 10-fold Repeated Cross-Validation successfully identified the optimal value at $k = 3$. The variables WaterCementRatio and LogAge were found to be the most dominant predictors, supporting the model's ability to capture the non-linear pattern of strength development in accordance with cement hydration principles. Residuals that are randomly distributed around zero indicate that the model is stable, free from systematic bias, and does not exhibit significant overfitting. From a practical perspective, this model can serve as a fast and efficient alternative for construction practitioners to estimate concrete quality without fully relying on laboratory testing, while also opening opportunities for further research using other algorithms or additional feature engineering techniques to improve accuracy, particularly for extreme strength values.

Acknowledgment

-

References

An, Y., Xu, M., & Shen, C. (2019). Classification method of teaching resources based on improved KNN algorithm. *iJET, 14*(4), 73–88.

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.

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.

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.

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.

Herdiansyah, & Pangaribuan, M. R. (2013). Pengaruh batu cadas (batu trass) sebagai bahan pembentuk beton terhadap kuat tekan beton. *Jurnal Inersia, 5*(2), 11–19.

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.

Nanja, M., & Purwanto. (2015). Metode K-Nearest Neighbor berbasis forward selection untuk prediksi harga komoditi lada Muis. *Jurnal Pseudocode, 2*(1), 53–64.

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. https://doi.org/10.33650/jeecom.v4i2

Rahardja, C. A., Juardi, T., & Agung, H. (2019). Implementasi algoritma K-Nearest Neighbor pada website rekomendasi laptop. *Jurnal Buana Informatika, 10*(1), 75–84.

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, 123*. https://doi.org/10.1007/s41939-025-00737-y

Widianti, A., & Pratama, I. (2024). Penanganan missing values dan prediksi data timbunan. *[Nama Jurnal Tidak Lengkap], 9*(2), 242–251.