Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6461

# Imputing Data and Predicting Waste with Machine Learning in East Java

**Rifa Khoirunisa[1*], Ahmad Faisal Sani[2], Darmawan Lahru Riatma[3], Masbahah[4], Yusuf Fadlila Rachman[5]**
[1,2,3,4,5]Universitas Sebelas Maret, Indonesia
[1]rkhoirunisa@staff.uns.ac.id, [2]faisalsani@staff.uns.ac.id, [3]darmawanlr@staff.uns.ac.id, [4]masbahah@staff.uns.ac.id, [5]yusuf_fadil@staff.uns.ac.id

open access

***Corresponding Author**

**Article History:**
Submitted: 10-07-2025
Accepted: 15-07-2025
**Published: 24-07-2025**

**Keywords:**
data imputation; waste generation; prediction; machine learning; East Java.



**ABSTRACT**

Indonesia's waste problem continues to be a pressing environmental issue, along with the increasing population and urbanization activities. The increase in population and changes in consumption patterns have led to a significant spike in waste generation in Indonesia. Machine learning-based approaches become highly relevant in supporting accurate predictive systems to estimate waste generation, so that it can be used as a basis for policy making and planning for more effective and sustainable waste management. However,the issue of missing data is a common challenge in environmental data processing, including in the recording of waste generation. Incomplete waste generation data can hinder accurate analysis and prediction, which are essential for effective environmental management planning. This study aims to analyze the effectiveness of various data imputation methods and to develop a predictive model for waste generation in East Java Province using a machine learning approach. The imputation techniques tested include Mean Imputation, K-Nearest Neighbor (KNN), and Interpolation, while the predictive models used include Random Forest, Gradient Boosting, and KNN Regression. The dataset was obtained from the official SIPSN (National Waste Management Information System) website. Model performance was evaluated using metrics such as Root Mean Square Error (RMSE). The results indicate that the combination of KNN Imputer with the Gradient Boosting prediction model is effective in addressing missing data and predicting waste generation in East Java Province, achieving an RMSE value of 0.147. These findings are expected to support more accurate decision-making in waste management planning for the province.

## INTRODUCTION

Waste is solid refuse generated from household and urban activities, as stated in Law No. 18 of 2008, referring to solid waste that has no economic value and must be managed to prevent environmental impact (Vinet & Zhedanov, 2008). The waste problem in Indonesia has developed into a significant national issue due to its various negative impacts, such as the degradation of environmental aesthetics, pollution of soil, water, and air, the emergence of various diseases, and the long-term risk of natural disasters (Nugraha et al., 2025). As the population increases, the amount of waste generation also rises (Saputra et al., 2024).Waste has become one of the growing environmental issues in line with population growth and increasing consumption activities in society. In East Java Province, the volume of waste generation increases every year. Ineffective waste management can lead to negative impacts such as environmental pollution and health problems. Therefore, predictive efforts regarding the amount of waste generation are essential as a basis for policy-making and sustainable waste management planning.

In waste generation prediction modeling, complete data without missing values is required. However, the available data often contains many missing values. These missing or incomplete data can be caused by various factors, such as recording errors, differences in regional reporting standards, or technical limitations during the data collection process. If not properly addressed, missing data can significantly reduce the performance of the prediction model. To handle these missing values, this study applies several data imputation methods, including Mean Imputation, Interpolation, and K-Nearest Neighbors (KNN) Imputer. These techniques are used to improve the quality of the dataset before it is used in the prediction process.

Machine Learning is a branch of artificial intelligence that focuses on developing algorithms based on data to make predictions or decisions without being explicitly programmed (Zhang et al., 2025). Machine learning technology can make a significant contribution to this research (Arminarahmah, 2025). In waste generation prediction, machine learning is used to build predictive models based on existing data patterns and relationships between variables. One of the algorithms used for prediction in machine learning is the Random Forest Regressor. This algorithm produces predictions through a data processing procedure (Simbolon et al., 2023). This study employs three commonly used machine learning algorithms for regression: Random Forest, Gradient Boosting, and K-Nearest Neighbors (KNN).

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6461

Random Forest is a tree-based ensemble method known for its stability and robustness against overfitting. Meanwhile, Gradient Boosting is a boosting technique that can enhance prediction performance by iteratively combining a number of weak trees. On the other hand, the KNN algorithm works based on the proximity of data points in the feature space, making it an intuitive method but sensitive to scale and noise in the data (Austin et al., 2021).

The study by Emmanuel et al. (2021) provides a comprehensive survey on missing data handling in machine learning, covering various imputation methods such as mean imputation, regression, KNN, and ensemble approaches like missForest. However, the study remains largely exploratory and was conducted on benchmark datasets like Iris and synthetic datasets, lacking application in real-world environmental domains such as waste generation forecasting at the regional level.

Furthermore, the study does not explore the combined impact of multiple imputation techniques and various machine learning regression models (e.g., Random Forest, Gradient Boosting, KNN) within a single predictive framework. This limits its applicability in domains where incomplete data is common and accurate forecasting is critical for sustainable policy planning such as municipal solid waste management in East Java Province.

## LITERATURE REVIEW

This study is supported by various previous studies that employed machine learning methods to address missing values in datasets (Yulian Pamuji et al., 2024). The study utilized the mean and KNN imputer methods to handle missing values in the dataset. This approach was taken to avoid a reduction in the number of data used in the classification process and to improve classification performance on non-ideal datasets, especially small datasets. The study titled " Penanganan Imputasi Missing Values pada Data Time Series dengan Menggunakan Metode Data Mining " which performed imputation of missing values in weather data using mean and KNN imputer (Prasetya et al., 2023).

Data imputation is a method for handling missing values in a dataset by replacing them using statistical or machine learning-based approaches (Emmanuel et al., 2021). The presence of missing data can lead to biased parameter estimates and result in inaccurate conclusions. To address this issue, multiple imputation is used as an effective statistical method. With multiple imputation, missing values are replaced multiple times (not just once), generating several complete datasets. Each dataset is then analyzed separately, and the results are combined to reflect the variability and uncertainty associated with the missing data (Afari & Lewis, 2022). Mean imputation is a simple method for handling missing values, where the missing value is replaced with the mean of the same variable across all subjects with available data. For example, if blood pressure values are missing for some patients, those values are replaced with the average blood pressure of the other patients whose data is available (Austin et al., 2021).

KNN imputer is an effective method for imputing missing data that works based on similarity learning, including the following approaches: For each observation with missing values, the algorithm calculates the distance (usually Euclidean) between data rows; Then, the k-nearest neighbors that have non-missing values for the targeted variable are selected; If the variable is numerical, the missing value is filled with the average of that variable from the k neighbors; and if it is categorical, it is filled with the mode. KNN Imputer is a method for replacing missing values based on the mean (for numerical variables) or mode (for categorical variables) of the k nearest neighbors that have non-missing values for the corresponding variable (Oktaviani & Putrada, 2022).

Interpolation is a statistical method used to fill gaps in data, especially in time series, by estimating values between two known data points. Unlike extrapolation, which predicts values outside the data range, interpolation is used for estimating values within the existing range. In imputation, interpolation is classified as a single univariate imputation method, as it relies solely on the variable itself to estimate the missing values (Bleidorn et al., 2024).

Random Forest is a machine learning method used to generate accurate predictions by processing data through the construction of multiple decision trees (Kuswanto & Hakim, 2025). Random Forest can also be used for classification by utilizing an ensemble of decision trees, where the final result is a combination of the outputs from the trained decision trees (Trihardianingsih & Permatasari, 2024). Random Forest is a stable machine learning algorithm, with a convergence rate that is influenced only by the most relevant features (Hanan et al., 2025). Random Forest, as an ensemble learning algorithm that utilizes multiple decision trees to generate more reliable and precise predictions, is a promising approach for addressing this issue due to its ability to handle complex data and build robust predictive models (Sitohang, 2025).

Gradient Boosting is an ensemble learning technique that iteratively builds weak models (typically decision trees) to correct the residual errors of the previous models. This method is capable of capturing complex interactions, non-linear structures, and temporal patterns, making it more advanced than static imputation methods such as mean or median (Oktaviani & Putrada, 2022).

The K-Nearest Neighbor (KNN) algorithm functions to classify new data by referring to the attributes and existing training data. This classification process does not require the construction of an explicit model, but instead relies entirely on stored historical data. In its implementation, the algorithm searches for the K training data points that are closest to the data point to be classified. The classification decision is then determined based on the majority class among these K nearest neighbors. The nearest neighbors are identified by calculating the shortest distance between the test data and the training data (Simbolon et al., 2023).

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6461

## METHOD

The method used to obtain RMSE comparisons for each model from the three missing data handling techniques involves several processes, including data collection, data preprocessing, model implementation, and evaluation.

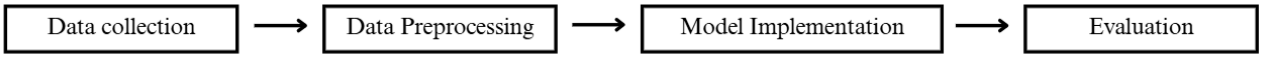


Fig 1. Research Method

The research method begins with data collection from secondary sources through a literature-based approach. Data were gathered by retrieving information from the official government website, SIPSN (National Waste Management Information System), on June 12, 2025. The data collected in this study consists of waste generation data from 2019 to 2024 across 38 regencies and cities in East Java Province.

Data preprocessing is an important stage because it involves selecting data according to the needs of the analysis. This process is carried out to reduce data size, normalize data, remove outliers, and extract data features. The collected dataset will then undergo data cleaning, followed by handling missing values using three methods: mean imputation, interpolation, and KNN imputer. This is done to improve the quality of the dataset so that it can produce accurate analysis and predictions for sustainable waste management.

After data preprocessing is completed, the dataset will be processed to obtain waste generation predictions using the Random Forest, Gradient Boosting, and KNN models, each applied with their respective missing value imputation methods.

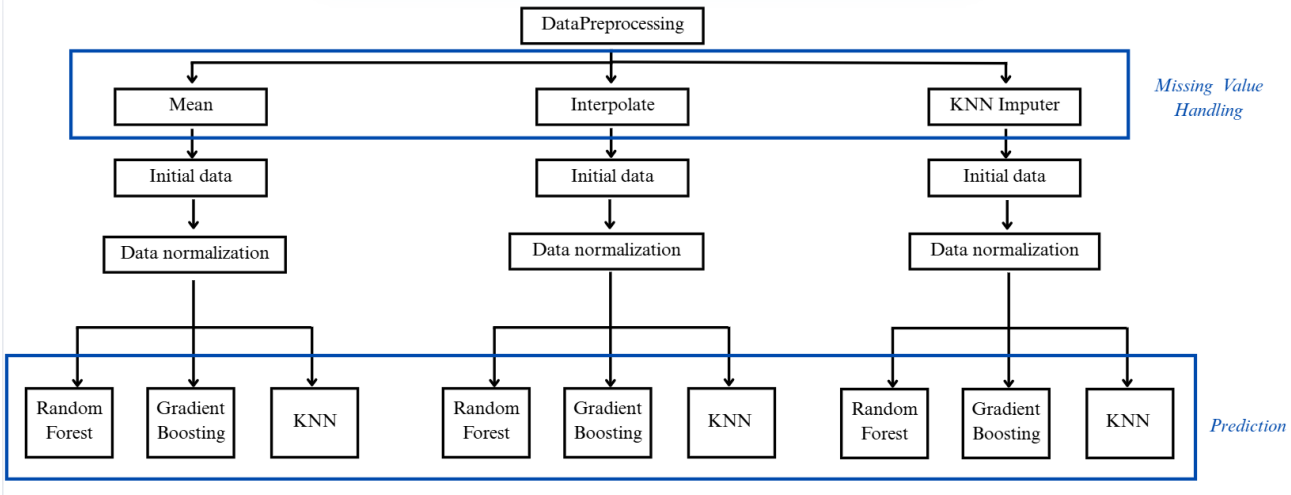


Fig 2. Model Implementation

In the data preprocessing stage for handling missing data, three models will be used Mean, Interpolate, KNN Imputer. To handle missing data in a dataset, the missing values can be replaced with the mean value of the corresponding column. The mean value can be calculated using the following formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i \qquad (1)$$

Interpolation is one of the techniques used to fill in missing data between two known data points. This technique is typically used for sequential data. Interpolation formula:

$$x = x_1 + \frac{(x_2 - x_1)}{(t_2 - t_1)}.(t - t_1) \qquad (2)$$

KNN Imputer (K-Nearest Neighbors Imputer) is a method for filling in missing values based on the similarity (distance) between data records using the Euclidean Distance metric. The formula used to measure the distance (Euclidean Distance) is:

$$Distance\ (x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (3)$$

Then, the formula used to calculate the replacement value for the missing data is:

$$x_f = \frac{1}{K}\sum_{i=1}^{K} Neighbor_i(f) \qquad (4)$$

After data preprocessing is completed, the next step is data prediction using three models, each applied with a different missing value handling method. The waste generation prediction models used are: Random forest, Gradient Boosting, KNN (K-Nearest Neighbors). Random Forest is used for regression prediction by combining multiple

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6461

decision trees. Each tree is built from a random subset of data and features, and the final prediction result is obtained by aggregating the outputs of all trees (Ahmad Fauzi, 2025) . The formula used in Random Forest is:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t\,(x) \quad (5)$$

Gradient Boosting combines many shallow decision trees gradually to form a strong prediction. The model is initialized with an initial value:

$$F_0 = arg\,\min \sum_{i=1}^{n} L(y_i\,, \gamma) \quad (6)$$

$$F_m(x) = F_{m-1}(x) + \eta.h_m(x) \quad (7)$$

$$F_M(x) = F_0(x) + \eta \sum_{m=1}^{M} h_m\,(x) \quad (8)$$

In KNN, the distance between the new data and all data in the dataset is calculated, then K nearest neighbors are selected, and the information from those neighbors is used to predict the value of the new data. The formula used is:

$$\hat{y} = \frac{1}{K}\sum_{i=1}^{K} y_i \quad (9)$$

## RESULT

The study began with the search for the required dataset, which in this case was data on the amount of waste generated in regencies and cities in East Java Province. The data were obtained from the official SIPSN (National Waste Management Information System) website, accessed through https://sipsn.menlhk.go.id/sipsn/public/data/timbulan on June 12, 2025. The collected data cover 38 regencies and cities in East Java, with waste generation figures from 2019 to 2024. The data can be seen in the table below:

Table 1. Initial Dataset

| No | Regency/City | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 |
|---|---|---|---|---|---|---|---|
| **1** | Madiun City | 42.622,24 | 43.133,55 | 43.695,25 | 44.219,80 | 44.750,39 | 45.287,39 |
| **2** | Surabaya City | 811.860,24 | 811.255,10 | 650.614,62 | 651.043,42 | 657.016,64 | 659.033,63 |
| **3** | Tulungagung Regency | 189.621,88 | | 200.127,68 | 202.148,86 | 204.028,80 | 164.395,27 |
| **4** | Lumajang Regency | 182.629,01 | 183.033,37 | 183.048,96 | 191.446,12 | 195.275,12 | |
| **5** | Banyuwangi Regency | 446.019,96 | | 457.297,22 | 297.078,45 | 305.312,85 | 306.270,96 |
| **6** | Kediri Regency | 184.669,09 | | 194.845,76 | 240.082,40 | 241.778,92 | |
| **7** | Sidoarjo Regency | 446.733,65 | 396.476,90 | | | 320.690,10 | 313.401,68 |
| **8** | Madiun Regency | 99.671,28 | | | 109.520,88 | 109.147,99 | 107.927,58 |
| **9** | Malang Regency | 380.505,78 | | | 350.614,09 | 352.927,26 | |
| **10** | Bondowoso Regency | 104.298,71 | 104.819,19 | | | | 115.049,02 |
| **11** | Probolinggo Regency | 170.133,51 | 167.335,13 | | | | 169.923,41 |
| **12** | Bangkalan Regency | | | | 151.736,05 | 152.259,02 | |

The waste generation data from 38 regencies in East Java Province, covering the years 2019 to 2024, show that 14 regencies have complete data with no missing values, 8 regencies have one year of missing data, another 8 regencies have two years of missing data, 5 regencies have three years of missing data, and 3 regencies have four years of missing data. Based on Table 1 above, many values are still missing if the data is to be used for prediction. Therefore, a data preprocessing stage must first be carried out to make the data suitable for prediction.

### Missing Value Imputation Using Mean

Mean imputation calculates the average of all available values in the corresponding column and then replaces the missing values with that average. This method can only be applied to numerical data. The results of missing value imputation using the Mean method can be seen in Table 2.

Table 2. Results of Missing Value Imputation Using Mean

| No | Regency/City | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 |
|---|---|---|---|---|---|---|---|
| **1** | Madiun City | 42622,24 | 43133,55 | 43695,25 | 44219,8 | 44750,39 | 45287,39 |
| **2** | Surabaya City | 811860,2 | 811255,1 | 650614,6 | 651043,4 | 657016,6 | 659033,6 |
| **3** | Tulungagung Regency | 189621,9 | **153895,2** | 200127,7 | 202148,9 | 204028,8 | 164395,3 |
| **4** | Lumajang Regency | 182629 | 183033,4 | 183049 | 191446,1 | 195275,1 | **166761,5** |

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6461

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **5** | Banyuwangi Regency | 446020 | **153895,2** | 457297,2 | 297078,5 | 305312,9 | 306271 |
| **6** | Kediri Regency | 184669,1 | **153895,2** | 194845,8 | 240082,4 | 241778,9 | **166761,5** |
| **7** | Sidoarjo Regency | 446733,7 | 396476,9 | **156496,6** | **160238,3** | 320690,1 | 313401,7 |
| **8** | Madiun Regency | 99671,28 | **153895,2** | **156496,6** | 109520,9 | 109148 | 107927,6 |
| **9** | Malang Regency | 380505,8 | **153895,2** | **156496,6** | 350614,1 | 352927,3 | **166761,5** |
| **10** | Bondowoso Regency | 104298,7 | 104819,2 | **156496,6** | **160238,3** | **169939** | 115049 |
| **11** | Probolinggo Regency | 170133,5 | 167335,1 | **156496,6** | **160238,3** | **169939** | 169923,4 |
| **12** | Bangkalan Regency | **180457,1** | **153895,2** | **156496,6** | 151736,1 | 152259 | **166761,5** |

In Table 2, the missing data has been filled using the mean method, with the previously missing and now filled values shown in bold. This dataset is now ready to be used for further analysis.

**Missing Value Imputation Using Interpolation**

Interpolation estimates missing values based on the sequence pattern among known data points. It produces a smoother value distribution without significantly biasing the mean and data spread. This aligns with the view of (Downing, 2025), who stated that interpolation—especially for sequential data—is an effective approach for preserving the natural structure of the data. The results of missing value imputation using interpolation can be seen in Table 3.

Table 3. Results of Missing Value Imputation Using Interpolation

| No | Regency/City | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 |
|---|---|---|---|---|---|---|---|
| **1** | Madiun City | 42622,24 | 43133,55 | 43695,25 | 44219,8 | 44750,39 | 45287,39 |
| **2** | Surabaya City | 811860,2 | 811255,1 | 650614,6 | 651043,4 | 657016,6 | 659033,6 |
| **3** | Tulungagung Regency | 189621,9 | **194874,8** | 200127,7 | 202148,9 | 204028,8 | 164395,3 |
| **4** | Lumajang Regency | 182629 | 183033,4 | 183049 | 191446,1 | 195275,1 | **195275,1** |
| **5** | Banyuwangi Regency | 446020 | **451658,6** | 457297,2 | 297078,5 | 305312,9 | 306271 |
| **6** | Kediri Regency | 184669,1 | **189757,4** | 194845,8 | 240082,4 | 241778,9 | **241778,9** |
| **7** | Sidoarjo Regency | 446733,7 | 396476,9 | **371214,6** | **345952,4** | 320690,1 | 313401,7 |
| **8** | Madiun Regency | 99671,28 | **102954,5** | **106237,7** | 109520,9 | 109148 | 107927,6 |
| **9** | Malang Regency | 380505,8 | **370541,9** | **360578** | 350614,1 | 352927,3 | 352927,3 |
| **10** | Bondowoso Regency | 104298,7 | 104819,2 | **107376,6** | **109934,1** | **112491,6** | 115049 |
| **11** | Probolinggo Regency | 170133,5 | 167335,1 | **167982,2** | **168629,3** | **169276,3** | 169923,4 |

In Table 3, the bold values represent the imputed values obtained through interpolation. The missing values that have been filled using interpolation make the data more suitable for further analysis.

**Filling in missing values with KNN Imputer**

Missing value imputation using the KNN Imputer is based on the similarity between data points. This method searches for the nearest neighbors based on the Euclidean distance from the column with missing values, then calculates the average value of the corresponding feature from the neighbors to fill in the missing value. This approach allows missing values to be filled based on the similarity of surrounding data. This method can only be applied to numerical data. The results of missing value imputation using the KNN Imputer can be seen in Table 4.

Table 4. Results of Missing Value Imputation Using KNN Imputer

| No | Regency/City | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 |
|---|---|---|---|---|---|---|---|
| **1** | Madiun City | 42622,24 | 43133,55 | 43695,25 | 44219,8 | 44750,39 | 45287,39 |
| **2** | Surabaya City | 811860,2 | 811255,1 | 650614,6 | 651043,4 | 657016,6 | 659033,6 |
| **3** | Tulungagung Regency | 189621,9 | **180935,2** | 200127,7 | 202148,9 | 204028,8 | 164395,3 |
| **4** | Lumajang Regency | 182629 | 183033,4 | 183049 | 191446,1 | 195275,1 | **183628,8** |
| **5** | Banyuwangi Regency | 446020 | **278767,7** | 457297,2 | 297078,5 | 305312,9 | 306271 |
| **6** | Kediri Regency | 184669,1 | **180935,2** | 194845,8 | 240082,4 | 241778,9 | **175967,4** |
| **7** | Sidoarjo Regency | 446733,7 | 396476,9 | **261898,1** | **339351,7** | 320690,1 | 313401,7 |
| **8** | Madiun Regency | 99671,28 | **106221,3** | **104301,9** | 109520,9 | 109148 | 107927,6 |
| **9** | Malang Regency | 380505,8 | **278767,7** | **299844** | 350614,1 | 352927,3 | **308212,5** |

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6461

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **10** | Bondowoso Regency | 104298,7 | 104819,2 | **104301,9** | **108673,2** | **108050,4** | 115049 |
| **11** | Probolinggo Regency | 170133,5 | 167335,1 | **192674,1** | **187305,9** | **190762,6** | 169923,4 |
| **12** | Bangkalan Regency | **141376,5** | **142753,2** | **151059,4** | 151736,1 | 152259 | 149794,6 |

In Table 4, the bold values represent the results of missing value imputation using the KNN Imputer. This method works by finding the closest values based on the Euclidean distance from the missing value, then calculating the average of the corresponding feature from the nearest neighbors to fill in the missing value. The KNN imputation results take into account the relationships between features and the local structure of the data. The results from KNN produce values that are more consistent with the original distribution pattern. This supports the findings of (Hameed & Ali, 2022), who stated that KNN-based methods are effective in estimating missing values, especially in datasets with strong inter-feature relationships.

Next, data transformation will be carried out, which is an important step in the data preprocessing stage to improve data quality and support the effectiveness of further analysis. Data transformation is done by separating the regency/city column to focus more on analyzing numerical data without categorical interference. This step can be assisted by machine learning to make it more efficient. After the missing value imputation process, the next step is data preprocessing through data scaling. This process is performed to normalize the data, i.e., by converting the values in the dataset to a scale between 0 and 1. Normalization is crucial as it helps reduce the scale differences between variables, making the dataset more effective and efficient for use.

## DISCUSSION

In this study, three prediction models were applied: Random Forest, Gradient Boosting, and KNN. These prediction models were implemented after the data preprocessing stage. As a result of preprocessing and model application, RMSE values were obtained for each missing value imputation method and each prediction model. Additionally, the predicted data and actual data were visualized based on the lowest RMSE result. The RMSE values from each model help determine the most suitable imputation method and prediction model for similar studies, as well as for forecasting values in each subsequent year. The following are the RMSE results for each prediction model.

Table 5. Average RMSE Results for Each Missing Value Imputation Method and Prediction Model

| No | Missing Value | Prediction | Average RMSE |
|---|---|---|---|
| **1** | Mean | Random Forest | 0,359 |
| | | Gradient Boosting | 0,384 |
| | | KNN | 0,378 |
| **2** | Interpolate | Random Forest | 0,625 |
| | | Gradient Boosting | 0,749 |
| | | KNN | 0,627 |
| **3** | **KNN Imputer** | Random Forest | 0,150 |
| | | **Gradient Boosting** | **0,147** |
| | | KNN | 0,164 |

The prediction model for waste generation data in East Java Province shows that the method of handling missing values and the prediction model used have a significant impact on the results. In this study, the combination of the KNN Imputer method and the Gradient Boosting model produced the lowest RMSE value, indicating that this combination is the most effective for handling missing data and producing accurate predictions. Random Forest also showed good performance when combined with interpolation and mean imputation methods. The following is a graph of the actual and predicted data generated from the model with the lowest RMSE value, namely the combination of KNN Imputer and Gradient Boosting.

Brilliance
Research of Artificial Intelligence

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6461

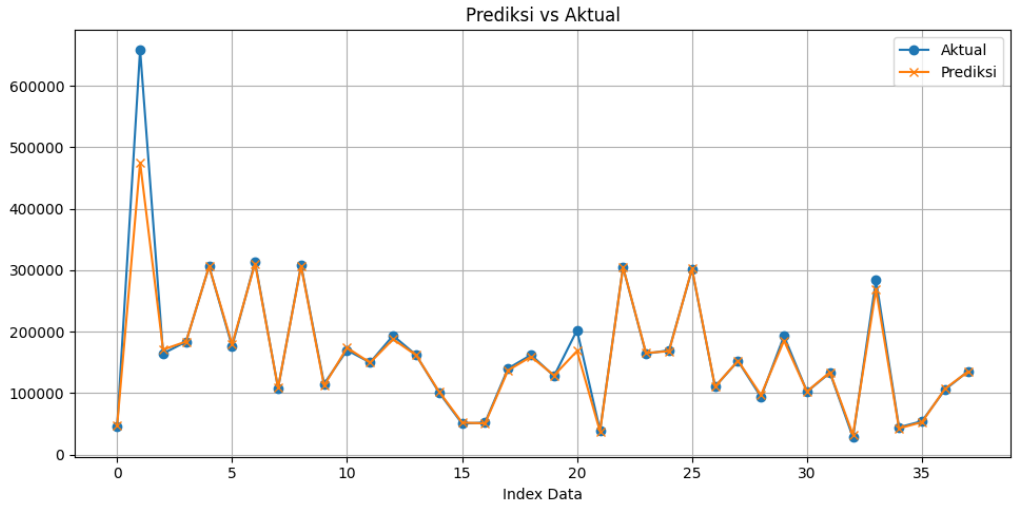


Fig 2. Visualization of Actual and Predicted Data

This graph shows that the prediction model used is able to map the actual values with reasonable accuracy, except at some extreme points. The model can be relied upon to predict future waste values, although special attention is needed to handle extreme data or outliers.

## CONCLUSION

Based on the results of this study, filling in missing data using the KNN Imputer and applying the Gradient Boosting prediction model yielded significant results in predicting the amount of waste generation in East Java Province. There were three methods used to fill in missing values in the dataset: mean, interpolation, and KNN Imputer. Each method demonstrated different performance levels when combined with the three prediction models: Random Forest, Gradient Boosting, and KNN.

Using mean imputation, the Random Forest model produced an RMSE value of 0.359, whereas interpolation resulted in an RMSE of 0.625, and the KNN Imputer achieved the lowest RMSE of 0.150. The interpolation method showed a relatively high RMSE, indicating that interpolation may not be suitable for addressing missing waste generation data in East Java Province. Despite this, interpolation still showed a lower RMSE value when paired with the Random Forest model (0.625). Overall, this study found that the KNN Imputer, when combined with the prediction models Random Forest, Gradient Boosting, and KNN, consistently produced lower RMSE values compared to other missing value imputation methods combined with the same three prediction models.

## REFERENCES

Afari, K. B., & Lewis, C. N. H. (2022). Performance Comparison of Imputation Methods for Mixed Data Missing at Random with Small and Large Sample Data Set with Different Variability. *Asian Journal of Probability and Statistics*, *20*(2), 16–39. https://doi.org/10.9734/ajpas/2022/v20i2416

Ahmad Fauzi, D. (2025). *Journal of Artificial Intelligence and Digital Business ( RIGGS ) Prediksi Harga Properti Di Indonesia Menggunakan Algoritma Random*. *4*(1), 43–49.

Arminarahmah, N. (2025). *Prediksi Pola Membuang Sampah Rumah Tangga Di Lahan Rawa Menggunakan Machine Learning*. *4*(10), 7309–7314.

Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2021). Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Canadian Journal of Cardiology*, *37*(9), 1322–1331. https://doi.org/10.1016/j.cjca.2020.11.010

Bleidorn, M. T., Schmidt, I. M., Dos Reis, J. A. T., Pani, D. F., Pinto, W. de P., Solci, C. C., Mendonça, A. S. F., & Brasil, G. H. (2024). Investigation of using missing data imputation methodologies effect on the SARIMA model performance: application to average monthly flows. *Revista Brasileira de Recursos Hidricos*, *29*, 1–18. https://doi.org/10.1590/2318-0331.292420230131

Downing, N. J. (2025). Missing Value Imputation in Environmental, Social, and Governance Data: An Impact on Emissions Scores. *Finance Research Letters*, *85*(PA), 107818. https://doi.org/10.1016/j.frl.2025.107818

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-021-00516-9

Hameed, W. M., & Ali, N. A. (2022). Comparison of Seventeen Missing Value Imputation Techniques. *Journal of Hunan University Natural Sciences*, *49*(7), 26–36. https://doi.org/10.55463/issn.1674-2974.49.7.4

**E-ISSN : 2807-9035**
**Volume** 5, **Number** 1, May 2025
https://doi.org/10.47709/brilliance.v5i1.6461

Hanan, M. R., Muflikhah, L., & Bachtiar, F. A. (2025). *Prediksi Nefropati Menggunakan Algoritma Random Forest*. *9*(5), 1–11.

Kuswanto, J., & Hakim, L. (2025). *Penerapan Algoritma Random Forest untuk memprediksi Performa Akademik Mahasiswa*. *5*(1), 262–270.

Nugraha, R., Suarna, N., Ali, I., & Rohman, D. (2025). Optimasi Pengelolaan Sampah Melalui Model Pengelompokan Dengan Algoritma K-Means. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(1), 646–652. https://doi.org/10.23960/jitet.v13i1.5694

Oktaviani, I. D., & Putrada, A. G. (2022). KNN imputation to missing values of regression-based rain duration prediction on BMKG data. *Jurnal Infotel*, *14*(4), 249–254. https://doi.org/10.20895/infotel.v14i4.840

Prasetya, M. R. A., Priyatno, A. M., & Nurhaeni. (2023). Penanganan Imputasi Missing Values pada Data Time Series dengan Menggunakan Metode Data Mining. *Jurnal Informasi Dan Teknologi*, *5*(2), 52–62. https://doi.org/10.37034/jidt.v5i2.324

Saputra, B. K. E., Defriatno, M. E., & Herdhianto, A. (2024). Prediction of Baby Diaper Waste Generation and Distribution in The Household Sector of Jember District. *BIOMA: Jurnal Biologi Dan Pembelajaran Biologi*, *9*(2), 140–149. https://doi.org/10.32528/bioma.v9i1.2493

Simbolon, V. A., Tarisa, & Horiza, H. (2023). Prediksi Tingkat Timbulan Sampah 5 Tahun Mendatang (2023-2027) di TPA Ganet Kota Tanjungpinang. *Sulolipu: Media Komunikasi Sivitas Akademika Dan Masyarakat*, *23*(2), 303–310. https://doi.org/10.32382/sulo.v23i2.105

Sitohang, N. (2025). Jurnal Sains Informatika Terapan ( JSIT ). *Mplementasi Algoritma Random Forest Untuk Prediksi Kelulusan Mahasiswa Berdasarkan Data Akademik: Studi Kasus Di Perguruan Tinggi Indonesia*, *2*(1), 16–20.

Trihardianingsih, L., & Permatasari, H. (2024). *Prediksi Area Kebakaran Hutan Menggunakan Algoritma Random Forest*. 37–41.

Vinet, L., & Zhedanov, A. (2008). A “missing” family of classical orthogonal polynomials. In *Undang-Undang Republik Indonesia Nomor 18 Tahun 2008* (Vol. 44, Issue 8). https://doi.org/10.1088/1751-8113/44/8/085201

Yulian Pamuji, F., Ahmad Rofiqul Muslikh, Rizza Muhammad Arief, & Delviana Muti. (2024). Komparasi Metode Mean dan KNN Imputation dalam Mengatasi Missing Value pada Dataset Kecil. *Jurnal Informatika Polinema*, *10*(2), 257–264. https://doi.org/10.33795/jip.v10i2.5031

Zhang, J., Gao, H., Liu, Y., & Wang, J. (2025). A Review on the Application of Superalloys Composition, Microstructure, Processing, and Performance via Machine Learning. *Jom*, *77*(1), 106–124. https://doi.org/10.1007/s11837-024-06922-7