JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website: http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN: 2085-3688; e-ISSN: 2460-0997

KNN imputation to missing values of regression-based rain duration prediction on BMKG data

Ikke Dian Oktaviani[1,*], Aji Gautama Putrada[2]
[1]School of Computing, Telkom University
[2]Advanced and Creative Networks Research Center, Telkom University
[1,2]Jl. Telekomunikasi, No. 1, Bandung 40257, Indonesia
[*]Corresponding email: oktavianiid@telkomuniversity.ac.id

Received 23 September 2022, Revised 19 October 2022, Accepted 31 October 2022

Abstract — The prediction of rain duration based on data from the Meteorology, Climatology, and Geophysics Agency (BMKG) is an important issue but remains an open problem. At the same time, several studies have shown that missing values can cause a decrease in the performance of the model in making predictions. This study proposes K-Nearest Neighbors (KNN) imputation to overcome the problem of missing values in predicting rain duration. The source of the rain duration prediction dataset is the BMKG data. We compared Gradient Boosting Regression (GBR), Adaptive Boosting Regression (ABR), and Linear Regression (LR) for the regression model for predicting rain duration. We compared the KNN imputation method with several benchmark methods, including zero imputation, mean imputation, and iterative imputation. Parameters r^2, Mean Squared Error (MSE) and Mean Bias Error (MBE) measure the performance of these imputation methods. The test results show that for rain duration prediction using the regression method, GBR shows the best performance, both for train data and test data with r^2 = 0.915 and 0.776, respectively. Then our proposed KNN imputation has the best performance for missing value imputation compared to the benchmark imputation method. The prediction values of r^2 and MSE when using KNN imputation at Missing Percentage = 90% are 0.71 and 0.36, respectively.

Keywords – climate prediction, KNN imputation, missing values, rain duration, regression

I. Introduction

The Meteorology, Climatology, and Geophysics Agency (BMKG) is a strategic agency in Indonesia regarding weather whose interests extend to aviation security [1]. The BMKG processes many weather data with complex problems that require advanced artificial intelligence skills, such as earthquake prediction, fire prediction, and wind power prediction [2]–[4]. However, the prediction of rain duration based on BMKG data is equally important but remains an open problem.

Climate prediction using the regression method is becoming increasingly important because it produces crucial information to deal with future weather conditions [5]. Tian *et al.* [6] used regression to see the relationship between soil moisture and drought. Poddar *et al.* [7] used temperature, humidity, and solar radiation parameters to predict the field crop coefficient. Utilizing weather data such as air pressure, temperature, and wind speed to predict the duration of rain can also be a research opportunity.

Previous research involved various regression models on predictions in the climate field. Puligudla *et al.* [8] used Gradient Boosting Regression (GBR) to predict crop yields based on weather data such as temperature and wind speed. Sena *et al.* [9] used Adaptive Boosting Regression (ABR) to predict temperature based on snow conditions, air pressure, and surface radiation. Kim *et al.* [10] utilized Linear Regression (LR) to predict the relationship of electrical energy consumption to weather data such as temperature, humidity, and cloud shape.

Climate sensors are an important part of the weather prediction system because their task is to pull physical information measurements into digital information [11]. However, one of the problems in retrieving temperature data through sensors is missing values [12]. Several studies have shown that missing values can cause a decrease in the model's performance in making predictions [13].

Several studies have used methods for imputation of missing values related to weather data. Sahoo *et al.* [14] used K-Nearest Neighbor (KNN) imputation for missing values for precipitation forecasting, while Jing *et al.* [15] used mean imputation for missing values in time series modeling in hydro-meteorology. Sudriani *et al.* [16] used iterative imputation to monitor data on water quality in Lake Maninjau. Yi *et al.* [17] proved that zero imputation was better than KNN imputation, mean imputation, and iterative imputation in predicting cardiovascular, hypertension, and diabetes. Proving that KNN imputation is better than other imputation methods in predicting rain duration is a research opportunity.

This study proposes KNN imputation to overcome the missing value problem in the prediction of rain duration. The dataset for Rain duration prediction comes from the BMKG data. We compare GBR, ABR, and LR as regression models for rainfall duration prediction. We then compare the KNN imputation method with several benchmark methods, including zero imputation, mean imputation, and iterative imputation. Parameters r^2, Mean Squared Error (MSE) and Mean Bias Error (MBE) measure the performance of these imputation methods. The use of MSE is because it can be used to monitor the performance of regression models.

To the best of our knowledge, there has never been a study using regression to predict the duration of rain on BMKG data and applying KNN imputation to the missing value. The following are the contributions of this research:

1) A pre-processed dataset on the BMKG dataset that is ready to train for rain duration prediction using Pearson Correlation Coefficient (PCC) feature selection.
2) An application of the regression method for prediction of rain duration using BMKG data.
3) An imputation method that maintains the performance of rain duration prediction using KNN imputation.

We structure the remainder of this paper systematically: Section II discusses the research method. Section III reports the test results. Section IV describes the discussion of test results with benchmark methods. Finally, section V is the conclusion and presentation of the important results of this research.

II. Research Method

Fig. 1 shows our research methodology. First, we explain the process of collecting the BMKG dataset and the pre-processing that we carry out on the data. Then we design and implement rain duration predictions with the dataset. After that, we add the missing values to see the durability of the prediction model if there are missing values. The next process is implementing KNN imputation and assessing its performance. The final step is to discuss and report the results.

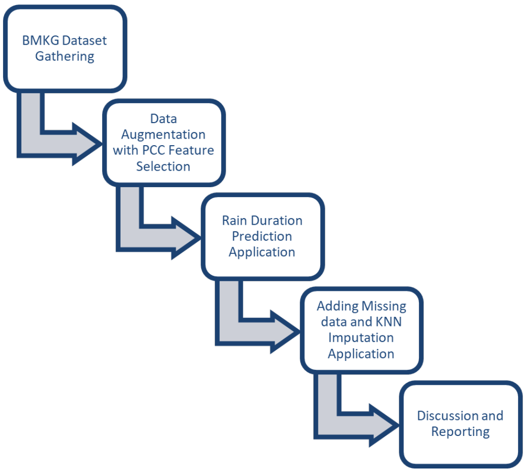
BMKG Dataset Gathering
Data Augmentation with PCC Feature Selection
Rain Duration Prediction Application
Adding Missing data and KNN Imputation Application
Discussion and Reporting

Fig. 1. Research methodology.

A. Dataset and Multicolinearity Filtering

We get data from BMKG in the form of a dataset with 13 features, one of which is rain duration, which is the data we want to predict with regression. Table 1 shows 12 BMKG feature data. The dataset size is 13,630,082. Given our limited hardware resources, we limit the data set size to 10,000.

Table 1. Example

No.	Feature Information		
	Feature Name	**description**	**Unit**
1	rowID	ID of row	Integer
2	hpwren_timestamp	Timestamp	Date and time
3	air_pressurre	Air pressure	hPa
4	air_temp	Air temperature	Celcius
5	avg_wind_direction	Average wind direction	degrees
6	avg_wind_speed	Average wind speed	km/hr
7	max_wind_direction	Maximum wind direction	degrees
8	max_wind_speed	Maximum wind speed	km/hr
9	min_wind_direction	Minimum wind direction	degrees
10	min_wind_speed	Minimum wind speed	km/hr
11	rain_accumulation	Rain accumulation	mm
12	relative_humidity	Relative humidity	%

We evaluate first which features are suitable for the prediction of rain duration using regression. For feature evaluation, we use the PCC [18]. Here is the formula for calculating PCC(r):

$$r = \frac{\sum\left(x_{iq} - \widehat{x}_q\right)\left(x_{ir} - \widehat{x}_r\right)}{\sqrt{\sum\left(x_{iq} - \widehat{x}_q\right)^2 \sum\left(x_{ir} - \widehat{x}_r\right)^2}}, \qquad (1)$$

where $i \in n$; $q, r \in p$, x is the feature, n is the data set size, and p is the number of features.

The PCC value ranges from -1 to 1, where a minus value indicates a negative relation, a positive value indicates a positive relation, a magnitude 0 indicates no correlation, and a magnitude 1 indicates a strong correlation [19]. The feature selection process excludes features with low PCC magnitude. Then from the side of multicollinearity, the remaining features must not have a strong correlation with each other. Ones that do must also be excluded [20].

B. Rain Duration Prediction with Regression Methods

We predict the duration of rain on BMKG data using the regression method. Here we compare the GBR, ABR, and LR methods. GBR is a regression whose basic is boosting [21]. The boosting method is an ensemble learning method with a series of weak learners where the next weak learner adjusts the misclassified data to produce a tree with a low bias [22]. Especially for gradient boosting, adjustments are made based on the MSE value. For example, a prediction function is F_m. Then based on the MSE value, a created function reduces the MSE in the next iteration, calling it the h_m function. Then the gradient function to calculate the value of h_m is as follows:

$$-\frac{\partial L_{MSE}}{\partial F\left(x_i\right)} = \frac{2}{n} h_m\left(x_i\right), \tag{2}$$

where L_{MSE} is the loss function of MSE, x_i is the feature with index i, and n is the dataset size.

ABR is a method with the same boosting concept as GBR. The difference is that instead of creating a function based on the loss function, ABR adds weight to the incorrectly predicted data [23]. An α variable denotes the weight value [24]. Like GBR, this α value adjusts to an error function E_t. The function follows the following equation:

$$E_t = \sum_{i \in n} E\left[F_{t-1}\left(x_i\right) + \alpha_t h\left(x_i\right)\right], \tag{3}$$

where $E\left[p\right]$ is an error function, t is the index of the boosting iteration, and $h\left(x_i\right)$ is the function that returns incorrectly predicted data.

Finally, LR is a regression in which the dependent and independent variables have a linear relationship [25]. For example, there is a dataset with the notation $y_i, x_{i1}, \ldots, x_{ip}$ where x is the independent variable, y is the dependent variable, p is the number of features, then n is the dataset size. The following equation states the LR formula:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \\ &= x_i^T \beta + \varepsilon_i \end{aligned}, \tag{4}$$

where $i \in n$, index T is the transpose notation, β is the intercept of each feature, and ε is the error of the linear mapping between each dependent variable and its independent variable [26].

C. KNN Imputation for Missing Values

In the concept of KNN imputation, a missing value is estimated based on its K-nearest neighbors. The neighbor distance with a missing value is calculated by the modified Euclidean distance (D) formula as follows:

$$D = \sqrt{w \times \left(m_i - x_i\right)^2},\ i \in p, \tag{5}$$

where x is the dataset, m is the dataset of items with missing data, p is the number of features of the dataset, and w is the weight, where the formula to get w is as follows:

$$w = \frac{amount\,of\,features}{amount\,of\,non\,-\,missingg\,features}, \tag{6}$$

then the imputed data is the average of the K smallest D results [27].

D. Benchmark Methods and Measurement Metrics

We compare the KNN imputation method with several other imputation methods, including zero imputation, mean imputation, and iterative imputation. Zero imputation is a very simple imputation method. The method is to replace the missing data with the number 0. Although simple, this method has proven to significantly affect the performance of estimation models, for example, on imputation in RNA gene data [28].

Mean imputation is replacing missing values with the average of non-missing values. Jamshidian *et al.* [29] stated that mean imputation could damage an estimation model because the variance deviates far from the expected variance. In comparison, iterative imputation is a process that considers missing data in a feature as a function of the value of other features [30]. The iterative imputation name is because the process repeats for each feature.

We use several regression testing metrics in this study, namely r^2, MSE, and MBE. The value of r^2 is the squared value of the result of Equation (1). The value range is from 0 to 1. Results closer to 1 show that the regression model has good performance, the opposite if it is close to 0. While the MSE formula is as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \widehat{x}_i\right)^2, \tag{7}$$

where n is the dataset size, x_i is the actual value, and $\widehat{x}_i$ is the predicted value. Then the MBE formula is as follows:

$$MBE = \frac{1}{n} \sum_{i=1}^{n} x_i - \widehat{x}_i \tag{8}$$

III. Result

The first test is feature selection based on the PCC value. Fig. 2 shows the heatmap of the PCC calculation results against the BMKG dataset. Several features are eliminated based on multicollinearity, namely avg_wind_speed and min_wind_speed because they correlate with max_wind_speed, which has the largest correlation with $rain\_duration$ of the three. Then $max\_wind\_direction$ and $min\_wind\_direction$ are also eliminated because they have a large correlation with $avg\_wind\_direction$, which has the largest correlation with $rain\_duration$. The selection process also eliminates the $rain\_accumulation$ feature because it is an independent value related to $rain\_duration$. Finally, the process eliminates air_temp because it has a large correlation with $air\_pressure$, which has a greater correlation with $rain\_duration$.

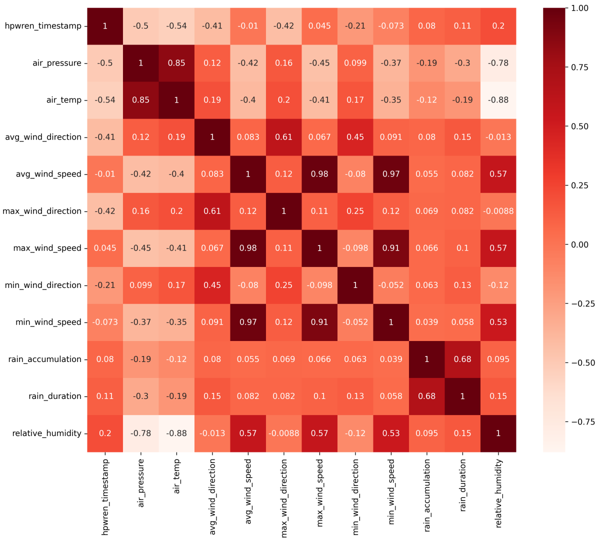

Fig. 2. PCC of BMKG dataset features.

The next test compares the performance of the three regression methods in predicting the duration of rain. Fig. 3 and Fig. 4 show the results. Fig. 3 is the result of testing r^2 for train data and test data from GBR, ABR, and LR. GBR shows the best train and test data performance with r^2 = 0.915 and 0.776, respectively. On the other hand, LR is the regression method with the worst performance for both train and test data with r^2 = 0.114 and 0.119, respectively. The performance of train data and ABR test data are r^2 = 0.385 and 0.374.

Fig. 4 shows the comparison of MSE results from the three regression methods. GBR has the smallest MSE, with a value of 0.327. Otherwise, LR has the largest MSE with a value of 1.052. The MSE ABR value is 0.791.

The last test is to apply the missing value and compare the performance of KNN imputation with benchmark methods. Fig. 5 to Fig. 7 show the results of comparing imputation methods regarding their changes to the missing percentage. Based on the value of r^2 and the MSE value, KNN imputation is better than other imputation methods in maintaining GBR performance by predicting rain duration. The r^2 and MSE KNN imputation values at Missing Percentage = 90% are 0.71 and 0.36, respectively. Meanwhile, the r^2 for mean, zero, and iterative imputation values are 0.62, 0.44, and -1.00, respectively. Lastly, the mean, zero, and iterative imputation's MSE values are 0.47, 0.70, and 2.48, respectively.

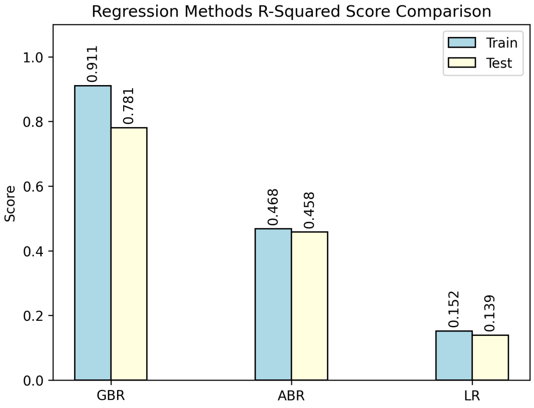

Fig. 3. The r^2 value comparison of regression methods on predicting rain duration based on BMKG data.

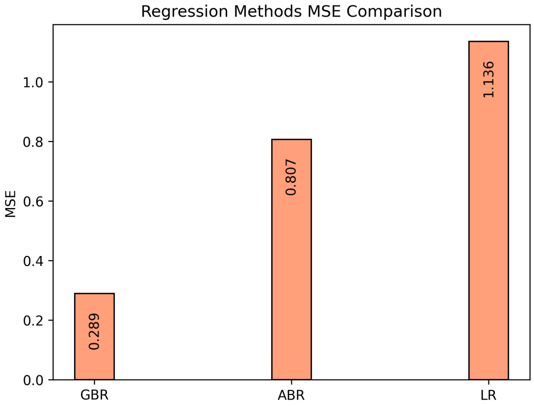

Fig. 4. The MSE value comparison of regression methods on predicting rain duration based on BMKG data.

IV. Discussion

This study builds a regression method to predict rain duration. To optimize the regression results, we perform feature selection using PCC. This process reduces the features from 12 to 4 and is proven to increase r^2 from 0.91 to 0.92. Several studies have shown that similar improvements can be made in other cases [31]. The contribution of this paper is a dataset from PCC feature selection that provides more accurate regression predictions.

We compare several regression methods to obtain the best rain duration prediction performance. The test results show that GBR is better than ABR and LR. Several studies have shown similar results, where the reason is that GBR is more robust against outliers than ABR [32]. The contribution of this paper is the optimum regression method for predicting rain duration based on BMKG data.

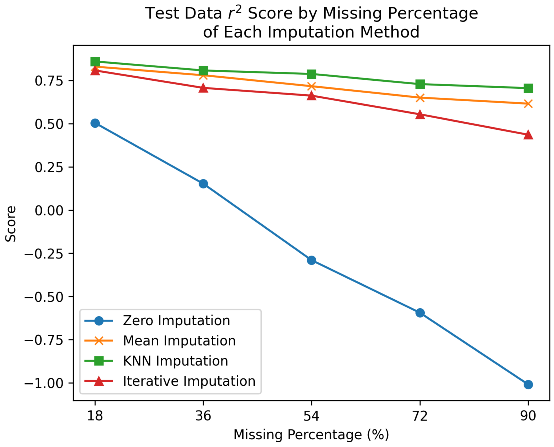

Fig. 5. Test data r^2 score by missing percentage performance comparison of regression prediction on imputation methods.

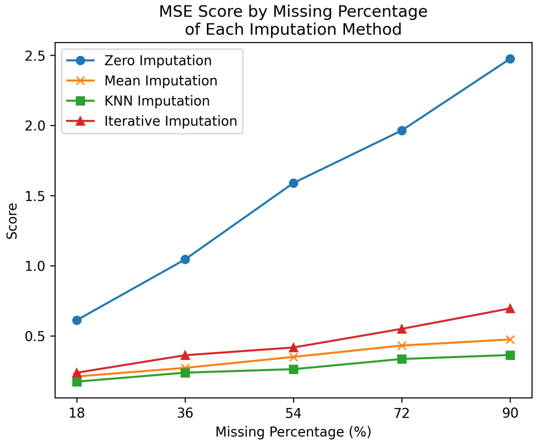

Fig. 6. MSE score by missing percentage performance comparison of regression prediction on imputation methods.

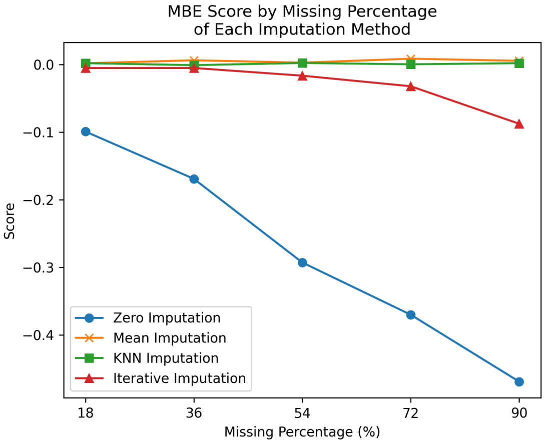

Fig. 7. MBE score by missing percentage performance comparison of regression prediction on imputation methods.

In this study, we proved that the KNN imputation performance outperformed the other three imputation methods in terms of r^2 and MSE. Paper [29] said that implementing mean imputation is not good for estimation because it changes the variance. However, in our observations, although not as good as the KNN imputation, the mean imputation is still better than the iterative imputation and zero imputation. The contribution of this paper is an optimum imputation method for missing values in rain duration prediction using BMKG data.

For future works, the direction of this research is imputation in the real-time domain for BMKG data. Several studies have been directed here in maritime machinery and traffic, both of which emphasize the importance of real-time imputation in some cases [33] and [34].

V. Conclusion

We developed a regression model to predict rain duration with BMKG data. The model includes an imputation method to fill in the missing values while maintaining the predictive model's performance. We propose KNN imputation for the imputation method. We compare the method with several benchmark methods, namely zero, mean, and iterative imputation. The test results show that for rain duration prediction using the regression method, GBR shows the best performance, both for train data and test data with r^2 = 0.915 and 0.776, respectively. Then our proposed KNN imputation has the best performance for missing value imputation compared to the benchmark imputation method. The r^2 and MSE KNN imputation values at Missing Percentage = 90% are 0.71 and 0.36, respectively.

Acknowledgment

We want to thank the Directorate of Research and Community Service (PPM) of Telkom University for funding this research. We also thank the Informatics Doctoral Study Program for motivating its students to reproduce publications on sciences that are beneficial to society. Finally, we would like to thank BMKG for sharing their data with us for further development.

References

[1] D. Wardani, S. Sulistyo, and I. W. Mustika, “The blueprint of AWOS implementation for aviation services at BMKG,” in *Conference SENATIK STT Adisutjipto Yogyakarta, 2018*, vol. 4, pp. 157–166.

[2] M. Syifa, P. R. Kadavi, and C.-W. Lee, “An artificial intelligence application for post-earthquake damage mapping in Palu, central Sulawesi, Indonesia,” *Sensors*, vol. 19, no. 3, p. 542, 2019.

[3] E. P. Purnomo and R. Ramdani, “Using artificial intelligence technique in estimating fire hotspots of forest fires,” in *IOP Conference Series: Earth and Environmental Science, 2021*, vol. 717, no. 1, p. 012019.

[4] D. H. Barus and R. Dalimi, “Determining optimal operating reserves toward wind power penetration in Indonesia based on hybrid artificial intelligence,” *IEEE Access*, vol. 9, pp. 165173–165183, 2021.

[5] Y. K. Hyun, J. Park, J. Lee, S. Lim, S.-I. Heo, H. Ham, S.-M. Lee, H.-S. Ji, and Y. Kim, "Reliability assessment of temperature and precipitation seasonal probability in current climate prediction systems," *Atmosphere*, vol. 30, no. 2, pp. 141–154, 2020.

[6] Y. Tian, Y.-P. Xu, and G. Wang, "Agricultural drought prediction using climate indices based on support vector regression in xiangjiang river basin," *Sci. Total Environ.*, vol. 622–623, pp. 710–720, 2018. Doi: 10.1016/j.scitotenv.2017.12.025.

[7] A. Poddar, N. Kumar, R. Kumar, and V. Shankar, "Application of regression modeling for the prediction of field crop coefficients in a humid sub-tropical agro-climate: a study in Hamirpur district of Himachal Pradesh (India)," Model. Earth Syst. Environ., vol. 8, no. 2, pp. 2369–2381, 2022.

[8] P. Puligudla, K. S. Karthik, K. N. Kumar, and M. Thirugnanam, "Prediction of crop yield using gradient boosting," *J. Xian Univ. Archit. Technol.*, vol. 12, no. 11, pp. 369–374, 2020.

[9] S. Sen, S. Saha, S. Chaki, P. Saha, and P. Dutta, "Analysis of PCA based adaboost machine learning model for predict mid-term weather forecasting," *Computational Intelligence and Machine Learning*, vol. 2, no. 2, pp. 41–52, 2021.

[10] M. K. Kim, Y.-S. Kim, and J. Srebric, "Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: Artificial neural network vs. linear regression," *Sustain. Cities Soc.*, vol. 62, p. 102385, 2020.

[11] A. Parashar, "IoT based automated weather report generation and prediction using machine learning," in *2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, Sep. 2019, pp. 339–344. Doi: 10.1109/ICCT46177.2019.8968782.

[12] J. Hooker, G. Duveiller, and A. Cescatti, "A global dataset of air temperature derived from satellite remote sensing and weather stations," *Sci. Data*, vol. 5, no. 1, pp. 1–11, 2018.

[13] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *J. Big Data*, vol. 8, no. 1, pp. 1–37, 2021.

[14] A. Sahoo and D. K. Ghose, "Imputation of missing precipitation data using KNN, SOM, RF, and FNN," *Soft Comput.*, vol. 26, no. 12, pp. 5919–5936, 2022.

[15] X. Jing, J. Luo, J. Wang, G. Zuo, and N. Wei, "A multi-imputation method to deal with hydro-meteorological missing values by integrating chain equations and random forest," *Water Resour. Manag.*, vol. 36, no. 4, pp. 1159–1173, 2022. Doi: 10.1007/s11269-021-03037-5.

[16] Y. Sudriani, F. A. Setiawan, and A. Hamid, "Comparison of kNN and iterative imputation approach for missing data value of online water quality monitoring system in lake maninjau," *JOIN (Jurnal Online Informatika)*, vol. 5, no. 1, pp. 1–3, 2020.

[17] J. Yi, J. Lee, K. J. Kim, S. J. Hwang, and E. Yang, "Why not to use zero imputation? correcting sparsity bias in training neural networks," ArXiv Prepr. ArXiv190600150, 2019.

[18] M. B. Satrio, A. G. Putrada, and M. Abdurohman, "Evaluation of face detection and recognition methods in smart mirror implementation," in *Proceedings of Sixth International Congress on Information and Communication Technology*, 2022, pp. 449–457.

[19] H. Akoglu, "User's guide to correlation coefficients," *Turk. J. Emerg. Med.*, vol. 18, no. 3, pp. 91–93, 2018.

[20] G. Vörösmarty and I. Dobos, "Green purchasing frameworks considering firm size: a multicollinearity analysis using variance inflation factor," *Supply Chain Forum: An International Journal*, vol. 21, no. 4, pp. 290–301, 2020.

[21] [21] A. G. Putrada, M. Abdurohman, D. Perdana, and H. H. Nuha, "Machine learning methods in smart lighting toward achieving user comfort: a survey," *IEEE Access*, vol. 10, pp. 45137–45178, 2022, doi: 10.1109/ACCESS.2022.3169765.

[22] I. Ghosal and G. Hooker, "Boosting random forests to reduce bias; one-step boosted forest and its variance estimate," *J. Comput. Graph. Stat.*, vol. 30, no. 2, pp. 493–502, 2020.

[23] A. N. Iman, A. G. Putrada, S. Prabowo, and D. Perdana, "Peningkatan kinerja AMG8833 sebagai thermocam dengan metode regresi adaBoost untuk pelaksanaan protokol COVID-19 performance improvement of AMG8833 as thermocam with adaBoost regression method for COVID-19 protocol enforcement," *Jurnal Elektro Telekomunikasi Terapan*, vol. 8, no. 1, pp. 978–985, 2021.

[24] A. Taufiqurrahman, A. G. Putrada, and F. Dawani, "Decision tree regression with adaBoost ensemble learning for water temperature forecasting in aquaponic ecosystem," in *2020 6th International Conference on Interactive Digital Media (ICIDM)*, 2020, pp. 1–5.

[25] M. D. Nastiti, M. Abdurohman, and A. G. Putrada, "Smart shopping prediction on smart shopping with linear regression method," in *2019 7th International Conference on Information and Communication Technology (ICoICT)*, Jul. 2019, pp. 1–6. Doi: 10.1109/ICoICT.2019.8835271.

[26] M. Hanif, M. Abdurohman, and A. G. Putrada, "Rice consumption prediction using linear regression method for smart rice box system," *J Teknol Dan Sist Komput*, vol. 8, no. 4, pp. 284–288, 2020.

[27] E. Yilmaz and D. Aydin, "Estimation of right censored nonparametric regression solved by kNN imputation: a comparative study," *Turk. Klin. J. Biostat.*, vol. 11, no. 2, pp. 83–92, 2019.

[28] G. Baruzzo, I. Patuzzi, and B. Di Camillo, "Beware to ignore the rare: how imputing zero-values can improve the quality of 16S rRNA gene studies results," *BMC Bioinformatics*, vol. 22, no. 15, pp. 1–34, 2021.

[29] M. Jamshidian and M. Mata, "2 - advances in Analysis of Mean and Covariance Structure when Data are Incomplete," *Handbook of Computing and Statistics with Applications*, S.-Y. Lee, Ed. Amsterdam: North-Holland, 2007, pp. 21–44. Doi: 10.1016/B978-044452044-9/50005-7.

[30] W. M. Hameed and N. A. Ali, "Enhancing imputation techniques performance utilizing uncertainty aware predictors and adversarial learning," *Period. Eng. Nat. Sci. PEN*, vol. 10, no. 3, pp. 350–367, 2022.

[31] Y. Liu, Y. Mu, K. Chen, Y. Li, and J. Guo, "Daily activity feature selection in smart homes based on pearson correlation coefficient," *Neural Process. Lett.*, vol. 51, no. 2, pp. 1771–1787, 2020.

[32] M. S. I. Khan, N. Islam, J. Uddin, S. Islam, and M. K. Nasir, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 8, pp. 4773–4781, 2022.

[33] C. Velasco-Gallego and I. Lazakis, "Real-time data-driven missing data imputation for short-term sensor data of marine systems. A comparative study," *Ocean Eng.*, vol. 218, p. 108261, 2020. Doi: 10.1016/j.oceaneng.2020.108261

[34] J.-M. Yang, Z.-R. Peng, and L. Lin, "Real-time spatiotemporal prediction and imputation of traffic status based on LSTM and Graph Laplacian regularized matrix factorization," *Transp. Res. Part C Emerg. Technol.*, vol. 129, p. 103228, 2021. Doi: 10.1016/j.trc.2021.103228.