

Enhancing Multiple Linear Regression with Stacking Ensemble for Dissolved Oxygen Estimation

Rahmaddeni¹, M. Teguh Wicaksono¹, Denok Wulandari², Agustriono¹, Sang Adji Ibrahim¹

¹Universitas Sains dan Teknologi Indonesia, Pekanbaru, Indonesia

²Institut Az Zuhra, Pekanbaru, Indonesia

Article Info

Article history:

Received July 20, 2024

Revised October 15, 2024

Accepted October 21, 2024

Keywords:

Dissolved oxygen

Multiple linear regression

Stacking Ensemble

ABSTRACT

Maintaining optimal dissolved oxygen levels is essential for aquatic ecosystems, yet industrial and domestic waste has led to a global decline in dissolved oxygen. Traditional measurement methods, such as oxygen meters and Winkler titration, are often costly or time-consuming. This study aims to improve the Root Mean Square Error, Mean Absolute Error, and R^2 values for estimating dissolved oxygen levels. The research method uses Multiple Linear Regression with various training and testing data splits, both before and after applying polynomial features. The model is further optimized using a stacking technique, with Random Forest Regressor and Gradient Booster Regressor as base models. The results show that the best model was achieved using the stacking ensemble technique with a 90:10 data split and polynomial features, yielding a Root Mean Square Error of 1.206, Mean Absolute Error of 0.990, and R^2 of 0.670. This model has also met the assumptions of linear regression, such as residual normality, homoscedasticity, and no autocorrelation of residuals. This study concluded that the ensemble stacking technique and the addition of polynomial features could improve the model in estimating dissolved oxygen values and also contribute by providing an accessible user interface using the Gradio Framework, allowing users to estimate dissolved oxygen levels effectively.

Copyright ©2024 The Authors.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

M. Teguh Wicaksono, +62-852-61033064

Study Program in Informatics Engineering,

Universitas Sains dan Teknologi Indonesia, Pekanbaru, Indonesia,

Email: muhammadteguuh01@gmail.com

How to Cite:

M. Wicaksono, R. Rahmaddeni, D. Wulandari, A. Agustriono, and S. A. Ibrahim, "Enhancing Multiple Linear Regression with Stacking Ensemble for Dissolved Oxygen Estimation", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 24, No. 1, pp. 85-94, November, 2024.

This is an open access article under the CC BY-SA license (<https://creativecommons.org/licenses/by-sa/4.0/>)

1. INTRODUCTION

The universe comprises various elements, including wind, air, soil, and water. Water, as the most abundant substance on Earth, is a source of life for living organisms [1]. Water contains dissolved oxygen used by aquatic animals to survive [2]. Therefore, Dissolved Oxygen (DO) is considered a crucial variable since oxygen levels that are too low can endanger aquatic habitats. Over the past 50 years, global dissolved oxygen levels have decreased by 2% [3]. The observed decline may be attributed to heightened human activities, particularly the disposal of industrial and domestic waste into water bodies, which degrades water quality and disrupts the balance of aquatic ecosystems. DO levels can also be significantly diminished due to reduced turbulence and atmospheric mixing, along with elevated water temperatures that exacerbate the decline in dissolved oxygen conditions [4]. Optimal DO levels for most fish are at least 5 milligrams per liter (mg/L). If DO levels fall below this threshold, or even below 2 mg/L, many fish will experience stress, leading to hypoxic conditions that can be fatal to both fish and invertebrates [5, 6].

Currently, DO levels can be measured using various methods, including DO meters and the Winkler titration method. However, DO meters are quite expensive, and the Winkler method requires reagents that are not readily available and involve a lengthy process. Data mining can be employed to estimate DO levels to address this issue. Data mining involves processing raw data to extract meaningful information, which aids in selecting the appropriate techniques prior to model development [7, 8]. This methodology has been widely utilized across various industrial sectors [8–11]. One of the data mining algorithms is Multiple Linear Regression, which predicts an independent variable based on dependent variables [12]. The study conducted by [13], which explores the optimization of Random Forest algorithms for classifying bank marketing data, suggests that implementing feature selection, feature engineering, and addressing class imbalance, data mining components may enhance model performance. Another study [14], achieved an accuracy of 93.49% using a hybrid data mining approach, specifically combining *Naïve Bayes* + K-Medoids clustering, and focused on determination and financing prediction in Shariah financing and loan cooperatives. The study [15] applied multiple linear regression to predict the number of students, producing coefficients and intercepts. Then, the researcher [16] compared Random Forest Regressor (RFR) and Multiple Linear Regression (MLR) in cattle weight estimation using feature selection, where MLR with five features produced a Mean Absolute Error (MAE) of 0.35, Mean Absolute Percentage Error (MAPE) of 0.07, Root Mean Square Error (RMSE) of 0.5, and an R^2 of 0.99. Although the metrics in this study were very good, it does not guarantee that all linear regression assumptions were met. Referring to DO, the

researcher [17] used time-series data and spatial data measured at 53 different locations. The spatial data results showed that the MLR model produced an R^2 of 0.57. This was followed by the study [18], which built a model to estimate DO using dimensionality reduction techniques and compared the RFR and Multi-Layer Perceptron (MLP) algorithms. The best model in this study was RFR, with an RMSE of 1.2805 and an MAE of 0.8911. Finally, the researcher [19] estimated DO values using a Support Vector Regressor, resulting in an R^2 of 0.32. There is a gap in previous research, particularly in the R^2 metrics produced and the lack of implementation to optimize models using ensemble techniques. Ensemble is a machine learning technique that combines several predictive models to improve overall performance compared to using a single model [20]. One such ensemble technique is stacking or stacked generalization, which combines the predictions from several base models to generate a better final prediction using a meta model [21]. The difference in this study lies in the quantity of data and the attributes used, as well as the implementation of optimization techniques using the Stacking Ensemble.

This study aims to obtain an optimal model for estimating DO values by applying feature engineering techniques such as adding polynomial features to the dataset and comparing models with and without polynomial features. Additionally, the stacking ensemble technique was applied using Random Forest Regressor and Gradient Boosting Regressor as base models and Multiple Linear Regression as the meta model. After obtaining the optimal model, linear assumption tests were also applied, including 1) Normality of residuals, 2) Homoscedasticity, and 3) No autocorrelation of residuals. The tests are conducted to determine the validity of the model. If these assumptions are not met, the prediction results can be biased, inaccurate, and unreliable for decision-making [22]. The results of this study indicate that the best model, using the stacking ensemble technique with Multiple Linear Regression as the meta model and a 90:10 data split, produced an RMSE of 1.206, MAE of 0.990, and an R^2 of 0.670. This model met all the assumptions of linear regression, including normality of residuals, absence of autocorrelation in the residuals, and homoscedasticity, ensuring that the error variance remained constant across all levels of the independent variables. This study contributes by creating a user interface using the Gradio Framework, which is accessible to all users and helps estimate DO values.

2. RESEARCH METHOD

This study employed a quantitative approach to develop an optimal estimation model. The process began with data acquisition, utilizing secondary data sourced from Kaggle.com. This data was thoroughly analyzed to extract key information, which informed the decision-making process before model development. After the data was carefully analyzed and processed, the next step involved

building a model designed to estimate DO levels accurately based on the identified variables. The comprehensive procedure is illustrated in Figure 1.

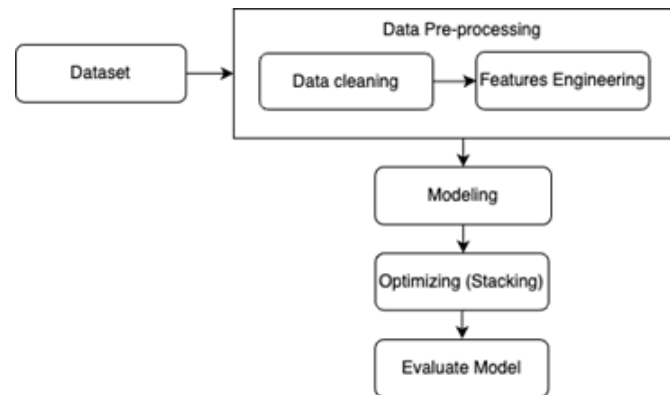


Figure 1. Research flow

2.1. Dataset

The Dissolved Oxygen dataset used in this study, sourced from Kaggle, comprises eight attributes and a total of 2371 rows. This dataset includes detailed records of water quality parameters collected biweekly from various aquatic locations, such as bays, fishing ponds, and other water bodies. Each record in the dataset represents a snapshot of water quality conditions at a specific time and location, providing a comprehensive overview of the factors influencing dissolved oxygen levels. For a complete description of the dataset attributes, please refer to Table 1.

Table 1. Attributes Information

Attribute	Type Data	Description
Date	Datetime	Date of water quality recording
Salinity (ppt)	Float	Salinity level (%)
Dissolved Oxygen (mg/L)	Float	Dissolved oxygen amount in water (mg/l)
pH	Float	Water pH level
SecchiDepth(m)	Float	Secchi depth in meters (water clarity)
WaterDepth(m)	Float	Water depth at sample location
WaterTemp(C)	Float	Water temperature in Celsius
AirTemp(C)	Float	Air temperature in Celsius

2.2. Data Pre-Processing

The data pre-processing stage is vital in research as it converts raw data into a more organized format, making it ready for analysis and machine learning model development [23]. This stage ensures that the data is cleaned and refined to meet the requirements of the subsequent analytical processes. In this study, pre-processing steps include data cleaning and feature engineering, which involve several key components. These procedures are outlined in detail below, highlighting the specific methods and techniques used.

2.3. Data Cleaning

In this study, the implementation of data cleaning involves several key steps: removing unused attributes such as Date, eliminating attributes air_temp because there is multicollinearity with water_temp, addressing missing values by opting to remove them for the sake of improving model performance, and eliminating duplicate data entries. This detailed approach underscores the critical importance of data cleaning, as it significantly impacts the quality and effectiveness of the resulting model. Each of these steps is designed to enhance the integrity of the dataset, thereby influencing the overall model accuracy and reliability [24].

2.4. Features Engineering

The feature engineering stage involves leveraging domain knowledge to extract meaningful features from raw data for use in machine-learning models [25]. Outliers are detected by calculating the interquartile range (IQR) using the formula: interquartile range $IQR = Q3 - Q1$. Where $Q1$ represents the 25th percentile, and $Q3$ represents the 75th percentile of the data distribution. The IQR is the difference between these two percentiles. Outliers are identified by calculating the *Lower Fence* = $Q1 - 1,5 \times IQR$ and the *Upper Fence* = $Q1 + 1,5 \times IQR$. Data points falling outside these fences are considered outliers [26]. However, in this study, instead of removing the outliers, they are replaced with the respective values of the lower and upper fences. Additionally, this study adds a new attribute called polynomial features. Polynomial features are used to capture non-linear relationships in the dataset. Implementing polynomial features can enhance the performance of machine-learning models [27]. For example, if x_0 and x_1 are features, the polynomial features could include x_0^2, x_0x_1 , dan x_1^2 .

2.5. Modeling

Before building the model, the dataset in this study was split into training and testing sets with the following ratios: 60:40, 70:30, 80:20, and 90:10. The dataset was used in two scenarios: (1) with polynomial features and (2) without polynomial features. A comparison of model performance was then conducted using Multiple Linear Regression, which is a statistical technique employed to model the relationship between a dependent variable and one or more independent variables [28]. The formula for Multiple Linear Regression is presented in Equation 1, which X_1, X_2, \dots, X_k are the independent variables. α is the intercept. $\beta_1, \beta_2, \beta_k$ are the regression coefficients for each independent variable.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \quad (1)$$

2.6. Optimizing

The optimization performed in this study utilized the Stacking Ensemble technique, with Random Forest Regressor (RFR) and Gradient Boosting Regressor (GBR) as the base models. RFR is a tree-based algorithm capable of handling large-scale data and is more robust to outliers [16]. GBR is an ensemble model that uses boosting techniques to build a stronger model from a collection of weak models, such as decision trees. In this algorithm, each new model learns from the mistakes of the previous model, and the result is a combination of all the initial predictions that are gradually updated [29]. The output of both models is then combined and used as input for the meta model. In this case, the meta model used is Multiple Linear Regression, which is responsible for learning the patterns from the predictions generated by the base models. In this way, the meta model produces a more optimal final prediction, as it leverages the predictive strengths of both base models. An illustration of how the stacking ensemble works in this study can be seen in Figure 2.

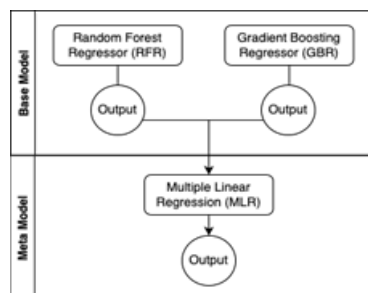


Figure 2. Stacking ensemble architecture with random forest and gradient boosting regressor as base models and multiple linear regression as meta model

2.7. Model Evaluation

Several metrics are used to evaluate the models. In this study, the performance metrics include RMSE, MAE, and R^2 . RMSE measures the average prediction error by quantifying the difference between predicted and actual values in the same units as the target variable [29]. MAE assesses the average absolute error between predictions and actual values, reflecting the model's accuracy in

replicating observed outcomes [30]. R^2 indicates the proportion of variability in the dependent variable explained by the independent variables, with values ranging from 0 to 1. An R^2 value close to 1 suggests a strong model fit to the observed data [17]. These metrics are detailed in Equations 2, 3, and 4.

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

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

Because this research uses a multiple linear regression algorithm, model evaluation is also carried out by testing the assumptions of linear regression to determine whether the linear regression model created is good enough. Testing the normality of residuals ensures that prediction errors are evenly distributed and the model is unbiased. This guarantees good predictions and valid statistical inferences from the model [31]. In this study, the normality test is conducted using the Shapiro-Wilk statistical method, which tests the null hypothesis that the samples x_1, \dots, x_n come from a population that follows a normal distribution as explained in equation 5. (i) where the parentheses enclosing the subscript i represent the i -th order statistic, meaning the i -th smallest number in the sample (not to be confused with x_i), $\bar{X} = (x_1 + \dots + x_n)/n$ is the sample mean, and the coefficient a_i is given by equation 6.

$$W = \frac{(\sum_{i=1}^n a_i X_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (5)$$

$$(a_i, \dots, a_n) = \frac{m^T V^{-1}}{c} \quad (6)$$

Where C is the normalization vector defined as $C = \|V^{-1}m\| = (m^T V^{-1}m)^{\frac{1}{2}}$ and the vector m is $m = (m_1, \dots, m_n)^T$, assisting with the expected values of the statistics from a sample of independent and identically distributed random variables from a standard normal distribution. Finally, V is the covariance matrix of these normal-order statistics. The Shapiro method is used in this research to test the normality of residuals. The hypotheses are as follows: H_0 = The residuals from the regression model follow a normal distribution; H_a = The residuals from the regression model do not follow a normal distribution. Ensuring the normality of residuals is a crucial part of validating linear regression models for accurate statistical inference, unbiased predictions, and effective model diagnostics. The concept of homoscedasticity in regression, which requires constant error variance across all levels of the independent variables, is crucial for valid analysis. The Goldfeld-Quandt test is used to check for homoscedasticity by evaluating the consistency of variance between observed and predicted values of the dependent variable [32]. This test employs F-statistics, as detailed in Equation 7, to confirm constant error variance.

$$F = \frac{\frac{RSS_1}{n_1 - k}}{\frac{RSS_2}{n_2 - k}} \quad (7)$$

RSS_1 and RSS_2 are the squared residuals from two subsets, n_1 and n_2 are their respective observation counts, and F includes estimated parameters. The hypothesis checks if H_0 = The residual variance is constant (homoscedasticity) and H_a = The residual variance is not constant (heteroscedasticity) [33]. The final step is testing for autocorrelation, which ensures that residuals in linear regression are independent. Autocorrelation testing checks if residuals are correlated, which may indicate unmodeled patterns affecting the residuals [34]. As outlined in Equation 8, the Ljung-Box test is used for this purpose. It evaluates whether significant correlations exist between residuals at different lags, thus assessing the independence of errors. n is the sample size \hat{p}_k is the sample autocorrelation at lag k , and h is the number of lags tested. The hypothesis is rejected if $Q > X_{(1 - \alpha, h)}^2$, where $X_{(1 - \alpha, h)}^2$ is from the chi-squared distribution table for significance level α and h degrees of freedom. Hypothesis: H_0 = No autocorrelation, and H_a = Autocorrelation exists.

$$Q = n(n+2) \sum_{k=1}^h \frac{\hat{p}_k^2}{n-k} \quad (8)$$

3. RESULT AND ANALYSIS

In the initial stage of building the model, this study first conducted modeling using multiple linear regression by comparing the model's performance across various training and testing data ratios, specifically 60:40, 70:30, 80:20, and 90:10. This was done to determine the most optimal model for different proportions of training and testing data [35]. The results of these experiments can be seen in Table 2.

Table 2. Model Comparison by Splitting Data

Model	Splitting Data	RMSE	MAE	R^2
MLR	60:40	1.6502	1.3267	0.4672
	70:30	1.6692	1.3488	0.4642
	80:20	1.6039	1.3003	0.4714
	90:10	1.4335	1.1466	0.5488

Table 2 shows that the MLR model with data splitting ratios of 60:40, 70:30, and 80:20 results in an average RMSE of 1.6 and an R^2 value below 0.5. This indicates suboptimal performance. From these results, it can be observed that the best model occurs with a data splitting ratio of 90:10, yielding an RMSE of 1.4335, MAE of 1.1466, and an R^2 of 0.5488. These results are considered reasonably good. Next, the authors compared the model with polynomial features and without polynomial features using the 90:10 data split, as the best model was previously observed with this data split. The results of this comparison between the two models can be seen in Figure 3.

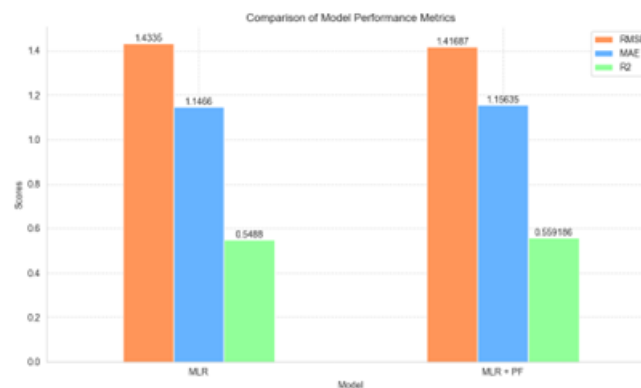


Figure 3. Model Comparison with and without Polynomial Features

Based on Figure 3, the MLR + PF model shows a slight improvement, although it is not significant. This model yields an RMSE approximately 0.02 lower than the MLR model; however, the MAE score increases by approximately 0.01 compared to the previous model. Additionally, R^2 experiences an increase of approximately 0.01 from the previous model. This indicates that the model shows an improvement, though not significant, in RMSE and R^2 scores, yet the model still cannot be considered optimal. Therefore, the authors attempted to optimize the MLR + PF algorithm using the stacking ensemble method. In this stacking ensemble, RFR and GBR were used as base models, with the only parameter adjusted being `random_state = 404`, while all other parameters were kept at their default settings. The MLR model was used as the meta model in this stacking technique. The results of this study showed improvement, as can be seen in Table 3.

Table 3. Model Result with Stacking Ensemble

Model	Splitting Data	RMSE	MAE	R^2
RF + GBR + MLR + PF	90:10	1.206	0.990	0.670

After optimizing the model using the stacking ensemble method, the next step was to test the assumptions underlying linear regression to ensure the reliability of the resulting model. The tests conducted included the normality of residuals, homoscedasticity, and no autocorrelation of errors. The results of these tests indicated that the model also met the assumptions of linear regression, as shown in Table 4.

Table 4. Regression Model Assumption Test Results

Test	Statistic	p-value	H_0	H_a
Normality of residuals	Shapiro-Wilk	0.4658	✓	–
Homoscedasticity	Goldfeld-Quandt	0.0540	✓	–
No autocorrelation of residuals	Ljung-Box	0.5269	✓	–

The Shapiro-Wilk test was conducted to examine whether the residuals are normally distributed. With a p-value of 0.4658, this result indicates that there is not enough evidence to reject the null hypothesis (H_0), which states that the residuals are normally distributed at a general significance level of $\alpha = 0.05$. Therefore, the assumption of normality of residuals is satisfied. The Goldfeld-Quandt test was used to assess homoscedasticity, i.e., whether the residual variance remains constant. With a p-value of 0.0540, this result suggests that the assumption of homoscedasticity is nearly fulfilled. The Ljung-Box test was conducted to check whether the residuals are autocorrelated. A p-value of 0.5269 indicates that there is no evidence to reject H_0 , meaning that there is no autocorrelation in the residuals. To facilitate users in estimating dissolved oxygen levels, this research includes a GUI accessible to all users. The GUI provides output indicating the impact of various conditions on DO levels. The GUI, hosted on HuggingFace, offers an intuitive and user-friendly platform for these estimations. The interface, as shown in Figure 4, can be accessed at the following link: https://huggingface.co/spaces/papayalovers/dissolved_oxygen_prediction.

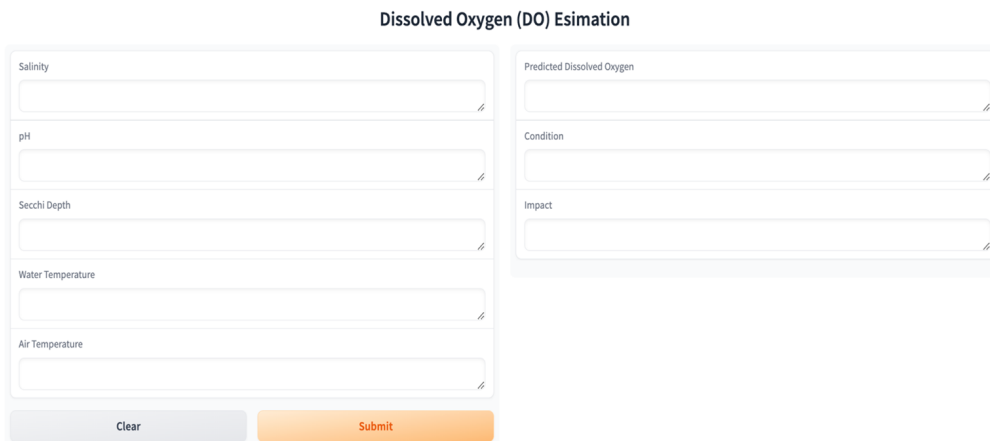


Figure 4. User Interface for Estimating Dissolved Oxygen Levels

The findings of this study include the development of an optimized model for estimating DO levels, complete with a user-friendly interface. The proposed model achieved performance metrics with an RMSE of 1.206, MAE of 0.990, and R^2 of 0.670. The results of this study are in line with previous research, such as [18], which compared the RFR and MLP in estimating DO values using the wrapper feature selection method to reduce data dimensionality. Their findings showed that the RFR model achieved an RMSE of 1.2805 and an MAE of 0.8911. Additionally, research by [17] using MLR reported an R^2 of 0.57, while [19] found an R^2 of 0.32 SVR. Similarly, this study found that, with a 90:10 data split and the addition of polynomial features, applying a stacking ensemble model with RFR and GBR as the base models improved performance, achieving an RMSE of 1.206 and an R^2 of 0.670. However, there was a slight decline in the MAE, which reached a value of 0.990, as shown in Table 5.

Table 5. Comparison with Previous Research

Research	Model	Method for Improvement	RMSE	MAE	R^2
[17]	MLR	-	-	-	0.57
[18]	RFR	Wrapper Feature Selection	1.2805	0.8911	-
[19]	SVR	-	-	-	0.32
This Study	Proposed Model	Polynomial Features + Stacking	1.206	0.990	0.670

4. CONCLUSION

This study employed the MLR method to estimate DO levels. The data was split into 90% training data and 10% testing data, yielding an optimal model. A stacking ensemble technique was applied to improve the performance of MLR, which was suboptimal on its own. RFR and GBR were used as base models, with MLR serving as the meta model. The results from this stacking approach showed an improvement in performance metrics, with an RMSE of 1.206, an MAE of 0.990, and an R^2 of 0.670. Furthermore, the model meets the key assumptions of linear regression: the residuals are normally distributed, the residual variance remains constant (indicating homoscedasticity), and there is no autocorrelation among the residuals. In addition, the study provides a user-friendly interface that enables users to estimate DO values easily. However, future research should explore using different datasets or alternative algorithms to achieve even better metrics, particularly RMSE, MAE, and R^2 . These efforts could help further enhance model performance and robustness in estimating DO values.

5. ACKNOWLEDGEMENTS

We extend our gratitude to everyone who contributed to completing this research. Although this research is not perfect, it serves as a stepping stone for further investigations to improve and refine methods to estimate Dissolved Oxygen (DO) without needing a DO meter.

6. DECLARATIONS

AUTHOR CONTRIBUTION

All five authors played significant roles in this research. Rahmaddeni and Agustriano generated ideas related to the research topic, methodology, and data collection. M. Teguh Wicaksono, assisted by Rahmaddeni, contributed to data analysis and applied appropriate techniques according to the data's characteristics. All authors participated in drafting and critically revising the manuscript, ultimately approving the final version and taking responsibility for the accuracy and integrity of the work.

FUNDING STATEMENT

This research was self-funded, with no external financial support received for its design, data collection, analysis, or interpretation. The authors personally bore all costs associated with this study.

COMPETING INTEREST

The authors declare no competing interests concerning the research presented in this manuscript.

REFERENCES

- [1] L. Mardiana, D. Kusnandar, and N. Satyahadewi, "Analisis Diskriminan Dengan K Fold Cross Validation untuk Klasifikasi Kualitas Air di Kota Pontianak," *Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya*, vol. 11, no. 1, pp. 97–102, 2022, <https://doi.org/10.26418/bbimst.v11i1.51608>.
- [2] B. Ali, A. Anushka, and A. Mishra, "Effects of dissolved oxygen concentration on freshwater fish: A review," *International Journal of Fisheries and Aquatic Studies*, vol. 10, no. 4, pp. 113–127, Jul. 2022, <https://doi.org/10.22271/fish.2022.v10.i4b.2693>.
- [3] C. Garcia-Soto, L. Cheng, L. Caesar, S. Schmidtko, E. B. Jewett, A. Cheripka, I. Rigor, A. Caballero, S. Chiba, J. C. Báez, T. Zielinski, and J. P. Abraham, "An Overview of Ocean Climate Change Indicators: Sea Surface Temperature, Ocean Heat Content, Ocean pH, Dissolved Oxygen Concentration, Arctic Sea Ice Extent, Thickness and Volume, Sea Level and Strength of the AMOC (Atlantic Meridional Overturning Circulation)," *Frontiers in Marine Science*, vol. 8, no. September, pp. 51–61, Sep. 2021, <https://doi.org/10.3389/fmars.2021.642372>.
- [4] K. M. Abbott, P. A. Zaidel, A. H. Roy, K. M. Houle, and K. H. Nislow, "Investigating impacts of small dams and dam removal on dissolved oxygen in streams," *PLOS ONE*, vol. 17, no. 11, pp. 1–23, Nov. 2022, <https://doi.org/10.1371/journal.pone.0277647>.
- [5] J. C. C. Casila, M. D. Nicolas, M. Duka, S. Haddout, K. L. Priya, and K. Yokoyama, "Assessing dissolved oxygen dynamics in Pasig River, Philippines: A HEC-RAS modeling approach during the COVID-19 pandemic," *Water Practice & Technology*, vol. 19, no. 4, pp. 1365–1381, Apr. 2024, <https://doi.org/10.2166/wpt.2024.078>.

- [6] H. Wang, L. Zhang, R. Wu, and H. Zhao, "Enhancing Dissolved Oxygen Concentrations Prediction in Water Bodies: A Temporal Transformer Approach with Multi-Site Meteorological Data Graph Embedding," *Water*, vol. 15, no. 17, pp. 3029–3046, Aug. 2023, <https://doi.org/10.3390/w15173029>.
- [7] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Social Science Research*, vol. 110, no. October, p. 102817–102833, Feb. 2023, <https://doi.org/10.1016/j.ssresearch.2022.102817>.
- [8] H. Santoso, H. Magdalena, and H. Wardhana, "Aplikasi Dynamic Cluster pada K-Means Berbasis Web untuk Klasifikasi Data Industri Rumahan," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 541–554, Jul. 2022, <https://doi.org/10.30812/matrik.v21i3.1720>.
- [9] Z. Liu, H. Gao, M. Zhang, R. Yan, and J. Liu, "A data mining method to extract traffic network for maritime transport management," *Ocean & Coastal Management*, vol. 21, no. 3, pp. 541–554, May 2023, <https://doi.org/10.1016/j.ocecoaman.2023.106622>.
- [10] A. Nugroho and Y. Religia, "Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, pp. 504–510, Jun. 2021, <https://doi.org/10.29207/resti.v5i3.3067>.
- [11] K. Aulakh, R. K. Roul, and M. Kaushal, "E-learning enhancement through educational data mining with Covid-19 outbreak period in backdrop: A review," *International Journal of Educational Development*, vol. 5, no. 3, pp. 504–510, Sep. 2023, <https://doi.org/10.1016/j.ijedudev.2023.102814>.
- [12] Y. Liu, G. B. Heuvelink, Z. Bai, P. He, X. Xu, W. Ding, and S. Huang, "Analysis of spatio-temporal variation of crop yield in China using stepwise multiple linear regression," *Field Crops Research*, vol. 26, no. 1, p. 102814–102830, May 2021, <https://doi.org/10.1016/j.fcr.2021.108098>.
- [13] Y. Religia, A. Nugroho, and W. Hadikristanto, "Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 187–192, Feb. 2021, <https://doi.org/10.29207/resti.v5i1.2813>.
- [14] O. E. Putra and R. Permana, "Hybrid Data Mining For Member Determination And Financing Prediction In Syariah Financing Saving And Loan Cooperatives," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 2, pp. 309–320, Apr. 2024, <https://doi.org/10.29207/resti.v8i2.5683>.
- [15] F. Riandari, H. T. Sihotang, and H. Husain, "Forecasting the Number of Students in Multiple Linear Regressions," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 249–256, Mar. 2022, <https://doi.org/10.30812/matrik.v21i2.1348>.
- [16] A. Setiawan, E. Utami, and D. Ariatmanto, "Cattle Weight Estimation Using Linear Regression and Random Forest Regressor," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 1, pp. 72–79, Feb. 2024, <https://doi.org/10.29207/resti.v8i1.5494>.
- [17] Y. Sudriani, F. Agus Setiawan, M. Fakhruddin, A. L. Latifah, N. Alias, N. Soliman, and A. D. Algarni, "An Estimation of Stratified Dissolved Oxygen Based on Spatio-Temporal Water Quality Parameters Via Single-Multiple Target Regression," 2024, <https://doi.org/10.2139/ssrn.4904124>.
- [18] F. H. Garabaghi, S. Benzer, and R. Benzer, "Modeling dissolved oxygen concentration using machine learning techniques with dimensionality reduction approach," *Environmental Monitoring and Assessment*, vol. 195, no. 7, pp. 879–894, Jul. 2023, <https://doi.org/10.1007/s10661-023-11492-3>.
- [19] A. Chatziantoniou, S. Charalampis Spondylidis, O. Stavrakidis-Zachou, N. Papandroulakis, and K. Topouzelis, "Dissolved oxygen estimation in aquaculture sites using remote sensing and machine learning," *Remote Sensing Applications: Society and Environment*, vol. 28, no. 4, pp. 1–10, Nov. 2022, <https://doi.org/10.1016/j.rsase.2022.100865>.
- [20] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, Feb. 2023, <https://doi.org/10.1016/j.jksuci.2023.01.014>.

- [21] M. Lu, Q. Hou, S. Qin, L. Zhou, D. Hua, X. Wang, and L. Cheng, "A Stacking Ensemble Model of Various Machine Learning Models for Daily Runoff Forecasting," *Water*, vol. 15, no. 7, p. 1265–1284, Mar. 2023, <https://doi.org/10.3390/w15071265>.
- [22] D. Alita, A. D. Putra, and D. Darwis, "Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 3, p. 295–306, Jul. 2021, <https://doi.org/10.22146/ijccs.65586>.
- [23] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, <https://doi.org/10.1016/j.glt.2022.04.020>.
- [24] J. Hu, "Data cleaning and feature selection for gravelly soil liquefaction," *Soil Dynamics and Earthquake Engineering*, vol. 145, no. March, p. 106711–106726, Jun. 2021, <https://doi.org/10.1016/j.soildyn.2021.106711>.
- [25] H. Mende, M. Frye, P.-A. Vogel, S. Kiroriwal, R. H. Schmitt, and T. Bergs, "On the importance of domain expertise in feature engineering for predictive product quality in production," *Procedia CIRP*, vol. 118, pp. 1096–1101, 2023, <https://doi.org/10.1016/j.procir.2023.06.188>.
- [26] D. Dallah and H. Sulieman, "Outlier Detection Using the Range Distribution," in *Advances in Mathematical Modeling and Scientific Computing*, F. Kamalov, R. Sivaraj, and H.-H. Leung, Eds. Cham: Springer International Publishing, 2024, pp. 687–697.
- [27] C. Niu, F. Wu, S. Tang, S. Ma, and G. Chen, "Toward Verifiable and Privacy Preserving Machine Learning Prediction," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 3, pp. 1703–1721, May 2022, <https://doi.org/10.1109/TDSC.2020.3035591>.
- [28] K. Lee, S. Im, and B. Lee, "Prediction of renewable energy hosting capacity using multiple linear regression in KEPCO system," *Energy Reports*, vol. 9, pp. 343–347, Nov. 2023, <https://doi.org/10.1016/j.egy.2023.09.121>.
- [29] D. A. Otchere, T. O. A. Ganat, J. O. Ojero, B. N. Tackie-Otoo, and M. Y. Taki, "Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions," *Journal of Petroleum Science and Engineering*, vol. 208, no. May, pp. 109244–109255, Jan. 2022, <https://doi.org/10.1016/j.petrol.2021.109244>.
- [30] D. Feng, Q. Han, L. Xu, F. Soheli, S. G. Hassan, and S. Liu, "An ensemble method for predicting dissolved oxygen level in aquaculture environment," *Ecological Informatics*, vol. 80, no. August, p. 102501–102516, May 2024, <https://doi.org/10.1016/j.ecoinf.2024.102501>.
- [31] A. Monter-Pozos and E. González-Estrada, "On testing the skew normal distribution by using Shapiro–Wilk test," *Journal of Computational and Applied Mathematics*, vol. 440, p. 115649–115675, Apr. 2024, <https://doi.org/10.1016/j.cam.2023.115649>.
- [32] Y.-Y. Zhao, J.-Q. Zhao, and S.-A. Qian, "A new test for heteroscedasticity in single-index models," *Journal of Computational and Applied Mathematics*, vol. 381, p. 25286–25298, Jan. 2021, <https://doi.org/10.1016/j.cam.2020.112993>.
- [33] A. Katsileros, N. Antonetsis, P. Mouzaidis, E. Tani, P. J. Bebeli, and A. Karagrigoriou, "A comparison of tests for homoscedasticity using simulation and empirical data," *Communications for Statistical Applications and Methods*, vol. 31, no. 1, pp. 1–35, Jan. 2024, <https://doi.org/10.29220/CSAM.2024.31.1.001>.
- [34] S. S. Uyanto, "Power Comparisons of Five Most Commonly Used Autocorrelation Tests," *Pakistan Journal of Statistics and Operation Research*, vol. 16, no. 1, pp. 119–130, Mar. 2020, <https://doi.org/10.18187/pjsor.v16i1.2691>.
- [35] P. P. Putra, M. K. Anam, S. Defit, and A. Yuniarta, "Enhancing the Decision Tree Algorithm to Improve Performance Across Various Datasets," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 8, no. 2, pp. 200–212, Aug. 2024, <https://doi.org/10.29407/intensif.v8i2.22280>.