BIS INFORMATION TECHNOLOGY AND COMPUTER SCIENCE

Volume 1 2024 | DOI: 10.31603/bistycs.136
5th Borobudur International Symposium on Science and Technology (BIS-STE) 2023

UNIMMA PRESS
UNIVERSITAS MUHAMMADIYAH MAGELANG

Utilization of machine learning to predict the correlation between color of river water and other water quality characters

I Sadidan[1*], G L Sari[2], E U Armin[2], F I Alifin[3] and A R Budiarto[1]

[1] Department of Environmental Engineering, Universitas Singaperbangsa Karawang, Karawang, Indonesia
[2] Department of Electrical Engineering, Universitas Singaperbangsa Karawang, Karawang, Indonesia
[3] Department of Industrial Engineering, Universitas Singaperbangsa Karawang, Karawang, Indonesia
*Corresponding author email: ikhwanussafa.sadidan@ft.unsika.ac.id

Abstract

This study investigates the intricate relationship between water color and key water quality parameters, such as DO, BOD, COD, TSS, and Fe concentrations. The primary objective is to establish a predictive model employing SVR analysis and DTR to discern the correlation patterns among these parameters. The purpose of this study is to predict and analyze the correlation between key water quality parameters with the water color. These models are constructed by scrutinizing the intricate associations between water color and the aforementioned water quality parameters using machine learning. Total Dissolved Solid and pH are two parameters that show a very high correlation with water color. Both show figures of 0.95 and 0.93. The results of this study can be implemented by various institutions such as educational institutions, environmental services, or consultants who want to make predictions and modeling of water quality, especially on color parameters. The results of this study can be implemented by various institutions such as educational institutions, environmental services, or consultants who want to make predictions and modeling of water quality, especially on color parameters.

Keywords

Machine learning, Color, River water, Water quality characters

Published:
October 20, 2024

Selection and Peer-review under the responsibility of the 5th BIS-STE 2023 Committee

Introduction

Artificial Intelligence (AI) has been applied across various scientific disciplines [1], and machine learning techniques have been introduced for predicting the quality of water [2]. The excessive exploitation and utilization of water resources have given rise to a series of issues, such as a decline in water quality, degradation of aquatic habitats, and the deterioration of river ecosystem structures. These challenges pose significant threats to social and economic development, as well as the safety of communities. The prediction of water quality is of paramount importance in preventing and addressing

water pollution issues. This predictive analysis serves to comprehensively understand the dynamic trends in the ecological water environment and forewarn potential instances of pollution. This proactive approach aids in the effective management of environmental resources and contributes to the safeguarding of water ecosystems [3]. The focus of water pollution control has shifted from remediation to prevention. To effectively mitigate water pollution, it is crucial to accurately predict future trends in water quality. Early warnings of such trends will promote scientifically informed water resource management, preserve ecosystem sustainability, and safeguard human health [4].

The color of water plays a significant role as a key indicator of water quality, influencing its sensory attributes along with factors like turbidity and odor. The presence of dissolved organic compounds, specifically fulvic and humic acids derived from decomposing plant material in soil and peatlands, as well as natural minerals like ferric hydroxide, typically gives rise to water color [5]. Originally, the assessment of true water color relied on visually comparing a filtered water sample with platinum–cobalt (Pt–Co) solutions according to the Hazen scale. Subsequently, standard solutions were replaced by color disks integrated into a comparator to enhance measurement convenience. However, these visual methods faced extensive criticism due to their subjective nature and limited precision [6].

The methods commonly employed in the process of classification or prediction include Support Vector Machine (SVM), Backpropagation, Radial Basis Function (RBF), K-Nearest Neighbor, Discriminant Analysis, Simple Logistic Classifier (SLC), Fuzzy logic, and others [7], [8], [9]. The Support Vector Machine (SVM) method is capable of adapting to various types of data used as input with minimal error. The efficiency and accuracy levels produced by this method are considered satisfactory [10]. To address regression problems, the Support Vector Regression (SVR) method is applied, which possesses a specialized algorithm for regression cases resulting in real-number output. The SVR algorithm's concept can yield good prediction values as it has the capability to address overfitting issues, where testing or training data produces nearly perfect prediction accuracy [11].

Methods

Sampling method

Sample collection was carried out using the purposive sampling principle, which involves selecting samples based on specific considerations in accordance with desired criteria [12] . Water samples were gathered at 33 points along a roughly 50 km irrigation channel, traversing 31 villages in the Karawang Regency (see Figure 1). Each water sample was collected in duplicate at approximately 1.5 km intervals from each location.

After the water samples were collected, a laboratory analysis was conducted to determine their quality. The parameters measured in this study include Dissolved

Oxygen (DO), Chemical Oxygen Demand (COD), Biological Oxygen Demand (BOD), and temperature. The results of the laboratory tests can be observed in Table 1.

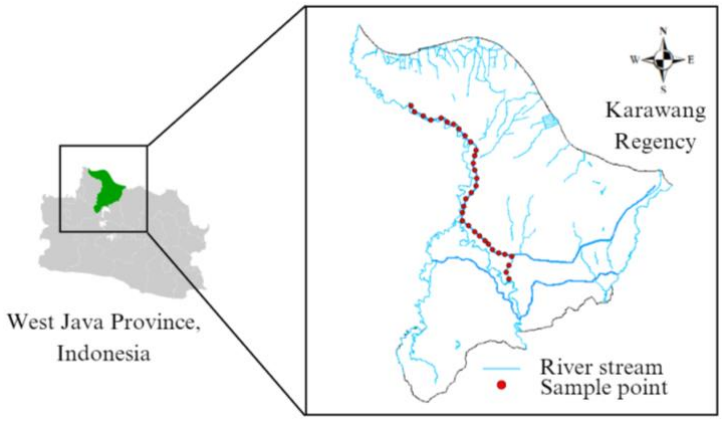

Figure 1. Map of Sampling Location

Table 1. Laboratories Test Results

Sampling Point	Parameters						
	Temperature (°C)	TDS (mg/L)	Color (PtCo)	DO (mg/L)	BOD (mg/L)	COD (mg/L)	Fe (mg/L)
1	29	257	16	0.9	12.5	17	0.152
2	30	273.5	22	1.2	11.7	19.5	0.115
3	38.5	164	13	4	20.9	39.3	0.14
4	30	147.5	28	3	19.7	28	0.083
5	31.5	160	18	4.4	22.2	74.6	0.133
6	31	153.3	17	3.9	21	33.3	0.113
7	30	160.5	13	2.3	21.4	32.8	0.099
8	29.5	263.5	23	7.8	22.9	77.2	0.23
9	30	195	32	6.7	21.3	33.3	0.23
10	31	193.5	14	6.3	13.5	20.6	0.204
11	31.5	167	24	6.4	22.7	30.8	0.205
12	31.5	184.5	23	6.1	23.3	30.6	0.272
13	32	185.5	47	7.9	11.7	17.7	0.183
14	32	384	40	6.1	19.7	36.2	0.294
15	32	205	48	8	17	25	0.254
16	30	200	25	7	23.2	28.1	0.245
17	30	247	51	8.6	17.7	25.4	0.137
18	31	179.5	49	8.7	17.5	23.5	0.147
19	33	189.5	18	8.1	14.2	18.6	0.101
20	32	161	43	6.8	11.7	20.7	0.123
21	33	185	29	8.2	14.9	19.7	0.108
22	30	172	32	7.2	12.4	21.2	0.144
23	29.4	195	97	0.7	12.8	24.3	0.279
24	31.1	183	73	1.5	13.1	21.8	0.277
25	31	175	82	0.9	13.5	48.6	0.264
26	31.5	192	75	1	13.1	71.7	0.277
27	31.5	196	37	7.7	6.2	8.2	0.289
28	32	187	38	7.6	2.4	9	0.326
29	31.6	175	51	7.4	2.8	9.7	0.306
30	31.8	196	40	6.9	2.5	8.9	0.256
31	31.7	187	41	7.7	5.3	12.5	0.256
32	31.8	191	43	8.1	4.7	8.1	0.315
33	31.4	187	27	7.8	1.8	8.8	0.306

Regression analysis

The regression method is a statistical data analysis technique used for forecasting and studying relationships between variables [13]. In regression analysis, commonly known as linear regression, there are two distinctions: simple linear regression and multiple linear regression. Simple linear regression is employed to generate a model depicting the relationship between one independent variable and one dependent variable. The general form of the equation for simple linear regression is shown on the Equation (1).

$$Y = \beta_0 + \beta_1 X_1 \quad (1)$$

On the other hand, the multiple linear regression model is an extension of the simple linear regression model. While the simple linear regression model involves only one independent variable and one dependent variable, the multiple linear regression model includes more than one independent variable and one dependent variable. By incorporating multiple independent variables, the general form of the equation for multiple linear regression using Equation (2).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (2)$$

Support vector regression (SVR) method

Support Vector Regression (SVR), an extension of SVM for regression purposes, aims to find a hyperplane function that serves as the best regression function fitting all input data with minimal error [14]. In this context, SVR strives to discover a function with maximum deviation from the actual target for each training data point. If this deviation equals zero, it indicates the presence of an ideal regression equation. It can be asserted that the Support Vector (SV) applies a concept similar to neural networks, employing radial basis functions as a specific form of such networks. In the case of RBF, the SV algorithm automatically determines centers, weights, and thresholds to minimize expected testing errors. Nevertheless, the superiority of results has been demonstrated in solving SVR series problems in several instances [15].

Decision tree regression (DTR) method

Decision Trees (DT) are non-parametric models of supervised learning used for classification and regression analysis. This model is based on a binary tree that divides one or more nodes to form a decision tree [16]. Decision tree algorithms split the dataset into smaller classes and represent the results in a leaf node. Essentially, a decision tree trains the dataset in the form of a tree structure for prediction. That is why it is sometimes referred to as a tree-structured regression. DT has three different types of nodes: the root node, internal nodes, and leaf nodes. The root node is the first node that is divided into more nodes, called internal nodes. Internal nodes represent the features of the model's data and decision rules, while leaf nodes represent the final outcome of the decision.

Results and Discussion

This study employs the Scikit-learn framework in the Python programming language to develop a regression model. The research utilizes a Personal Computer (PC) with an Intel Core-i5 Gen 9 processor and 8GB RAM for training and validating the created regressor model. Two machine learning methods, namely Support Vector Regressor and Decision Tree Regressor, are employed by the researcher to create the regressor model on the utilized dataset. The researcher designs the regressor model using six input variables: Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), temperature, Fe, and Total Suspended Solids (TSS) to predict water color.

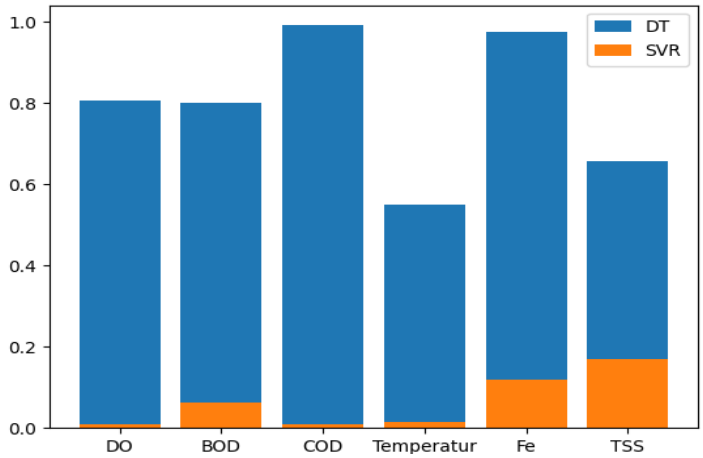

Figure 2. Model Accuracy for each input using DTR and SVR

In Figure 2, the accuracy levels of the model for each parameter in relation to the color quality values are depicted. The figure illustrates that COD and Fe exhibit the highest accuracy compared to other parameters. Meanwhile, Figure 3 presents a comparison between all inputs using two methods. The observed difference is highly significant, with the accuracy rate using Decision Tree Regressor (DTR) showing a result of 81.93%, whereas for Support Vector Regressor (SVR), it is only 3.58%.

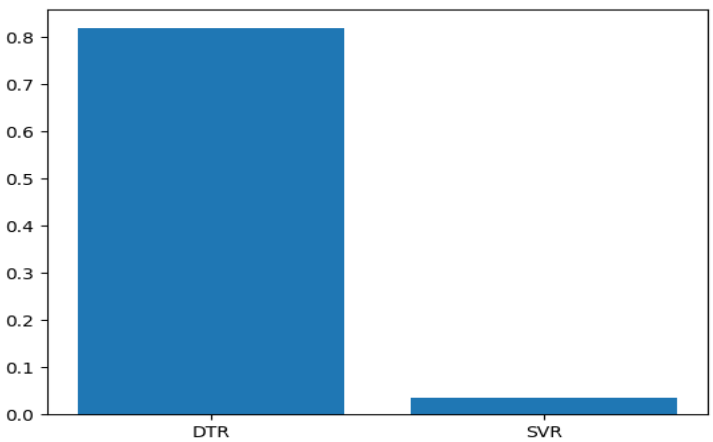

Figure 3. Model Accuracy for all input using DTR and SVR

Furthermore, the researcher compared each parameter in the input with the color values using two methods, and the Root Mean Square Error (RMSE) values were calculated. For Decision Tree Regressor (DTR), the RMSE is 8.66%, while for Support

Vector Regressor (SVR), it is at 20%. The results of these comparison are illustrated in Figure 4.

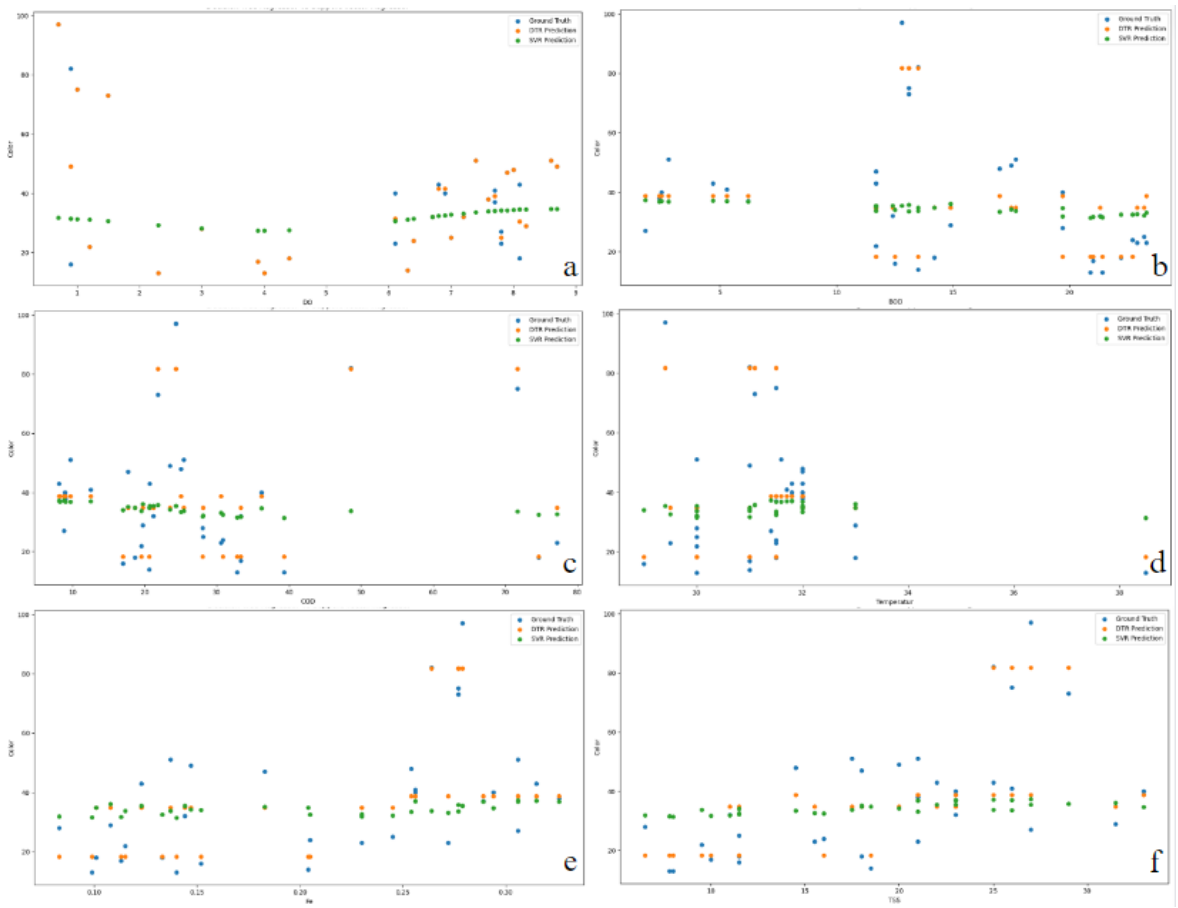

Figure 4. Model Comparation using DTR dan SVR: (a) DO and Color; (b) BOD and Color; (c) COD and Color; (d) Temperature and Color; (e) Fe and Color; (f) TSS and Color

Conclusion

In summary, the comparative analysis of the Support Vector Regression (SVR) and Decision Tree Regressor (DTR) models, incorporating input and model filter, has yielded compelling insights into their predictive performance. The obtained accuracy scores further underscore the efficacy of the Decision Tree Regressor, achieving an impressive 81.93%, as opposed to the SVR model, which trailed with a modest 3.58%. The calculated loss metrics, in terms of Root Mean Square Error (RMSE), reinforce these findings, showcasing a considerably lower score of 8.66 PtCo for the Decision Tree Regressor in contrast to the higher 20 PtCo recorded for the SVR method.

These outcomes strongly suggest that the Decision Tree Regressor method outperforms the SVR method when applied to the researcher's dataset. The substantial disparity in accuracy and lower RMSE values for the Decision Tree Regressor not only attests to its superior predictive capabilities but also highlights its robustness in capturing complex relationships within the data. This conclusion provides valuable insights for practitioners and researchers seeking an effective regression model for

similar datasets, emphasizing the practical significance of selecting appropriate algorithms tailored to the specific characteristics of the data at hand.

Acknowledgments

This research is funded by the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia through the Decentralization Research for University in 2019 that is managed by the Center of Research, Development, and Community Services of Universitas Muhammadiyah Magelang.

References

[1] H. Min, "Artificial intelligence in supply chain management: theory and applications," *International Journal of Logistics Research and Applications*, vol. 13, no. 1, pp. 13–39, Feb. 2010, doi: 10.1080/13675560902736537.

[2] M. Liu and J. Lu, "Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river?" *Environmental Science and Pollution Research*, vol. 21, no. 18, pp. 11036–11053, Sep. 2014, doi: 10.1007/s11356-014-3046-x.

[3] H. Wu *et al.*, "Water Quality Prediction Based on Multi-Task Learning," *Int J Environ Res Public Health*, vol. 19, no. 15, p. 9699, Aug. 2022, doi: 10.3390/ijerph19159699.

[4] Suwari, "Analysis Of Water Quality Status Using Method Of Water Pollution Index: A Case Study On The Dendeng River," *International Journal of Research -GRANTHAALAYAH*, vol. 9, no. 5, pp. 200–218, Jun. 2021, doi: 10.29121/granthaalayah. v9.i5.2021.3937.

[5] D. Hongve and G. Åkesson, "Spectrophotometric determination of water colour in hazen units," *Water Res*, vol. 30, no. 11, pp. 2771–2775, Nov. 1996, doi: 10.1016/S0043-1354(96)00163-7.

[6] L. E. Bennett and M. Drikas, "The evaluation of colour in natural waters," *Water Res*, vol. 27, no. 7, pp. 1209–1218, Jul. 1993, doi: 10.1016/0043-1354(93)90013-8.

[7] R. Isnaeni, S. Sudarmin, and Z. Rais, "Analisis Support Vector Regression (SVR) Dengan Kernel Radial Basis Function (RBF) Untuk Memprediksi Laju Inflasi Di Indonesia," *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, vol. 4, no. 1, pp. 30–38, 2022.

[8] I. Sadidan, G. L. Sari, E. U. Armin, F. I. Alifin, and V. U. Bunga, "Pemanfaatan Machine Learning untuk Memprediksi Kandungan Dissolved Oxygen (DO) pada Air Sungai Menggunakan Metode Decision Tree Regressor (DTR) dan Support Vector Regressor (SVR)," *BRAHMANA: Jurnal Penerapan Kecerdasan Buatan*, vol. 5, no. 1, pp. 77–84, Dec. 2023, doi: https://doi.org/10.30645/brahmana.v5i1.280.

[9] E. U. Armin, A. P. Edra, F. I. Alifin, I. Sadidan, I. P. Sary, and U. Latifa, "Performa Model YOLOv8 untuk Deteksi Kondisi Mengantuk pada pengendara mobil," *Brahmana: Jurnal Penerapan Kecerdasan Buatan*, vol. 5, no. 1, pp. 67–76, Dec. 2023, doi: https://doi.org/10.30645/brahmana.v5i1.279.

[10] N. I. Widiastuti, E. Rainarli, and K. E. Dewi, "Peringkasan dan Support Vector Machine pada Klasifikasi Dokumen," *JURNAL INFOTEL*, vol. 9, no. 4, p. 416, Nov. 2017, doi: 10.20895/infotel. v9i4.312.

[11] R. P. Furi, J. Jondri, and D. Saepudin, "Prediksi Financial Time Series Menggunakan Independent Component Analysis Dan Support Vector Regression. Studi Kasus: Ihsg Dan Jii.," *eProceedings of Engineering*, vol. 2, no. 2, 2015.

[12] D. Sugiyono, "Metode penelitian pendidikan pendekatan kuantitatif, kualitatif dan R&D," 2013.

[13] J. Neter, W. Wasserman, and M. H. Kutner, *Applied linear regression models*. Richard D. Irwin, 1983.

[14] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat Comput*, vol. 14, no. 3, pp. 199–222, Aug. 2004, doi: 10.1023/B: STCO.0000035301.49549.88.

[15] R. E. Caraka, H. Yasin, and A. W. Basyiruddin, "Peramalan Crude Palm Oil (CPO) Menggunakan Support Vector Regression Kernel Radial Basis," *Jurnal Matematika*, vol. 7, no. 1, p. 43, Jun. 2017, doi: 10.24843/JMAT. 2017.v07.i01.p81.

[16] P. R. Kadavi, C.-W. Lee, and S. Lee, "Landslide-susceptibility mapping in Gangwon-do, South Korea, using logistic regression and decision tree models," *Environ Earth Sci*, vol. 78, no. 4, p. 116, Feb. 2019, doi: 10.1007/s12665-019-8119-1.