International Journal of Ecology and Environmental Sciences
www.ecologyjournal.in
Online ISSN: 2664-7133, Print ISSN: 2664-7125
Received: 10-08-2024, Accepted: 09-09-2024, Published: 24-09-2024
Volume 6, Issue 3, 2024, Page No. 97-106

Development of a machine learning predictive model for identification of key water quality pollution parameters at River Benue

GA Mohamed[1], MO Udochukwu[2*], Dr. SO Enokela[3], YD Kantiok[3]
[1] Department of Hydrogeology, Lower Benue River Basin Development Authority, Makurdi, Benue, Nigeria
[2] Professor, Department of Agricultural and Environmental Engineering, Josepth Sarwuan Tarka University Makurdi, Makurdi, Benue, Nigeria
[3] Department of Agricultural and Environmental Engineering, Josepth Sarwuan Tarka University Makurdi, Makurdi, Benue, Nigeria

Abstract
This study on development of a machine learning predictive model for identification and classification of water pollution parameters at River Benue at Makurdi reach applied relevant artificial intelligence models to verify the pollution levels of the river. The water samples were collected in sterile bottles of 1,500 mL capacity at the depth of 20 cm and examined in the laboratory using standard methods. Results of the verified laboratory data sets tested on the different AI models were used to validate the modal as to identify the model that best predict water pollution levels of river Benue in terms of EC, SAR, SO_4, TDS relevant to irrigation water quality standards. The Ensemble (Bagged Tree), Ensemble (Boosted Tree) and the SVM (Quadratic) models emerges as the optimal predictive model for estimating SAR, DO and TDS respectively in River Benue at Makurdi. The observed and simulated values in the form of scatter plots were used to validate the model. The scatter plots reported an acceptable deviation from the ideal line of the 45º, which confirm the degree of the correlation between the observed and simulated dataset. An additional analysis was performed using the SI variations of SVM-FFA model "as a superior predictive model" among different test stages.

Keywords: Water quality, machine learning, predictive model, River Benue, Makurdi

Introduction

Regular water quality testing is expensive, time-consuming and involves collection and analysis of a large number of parameters (Kaushal *et al.,* 2023) [8]. Recent advances in sensors, telemetry, and machine learning models provide opportunities for more efficient and continuous monitoring of surface water systems (Maier *et al.,* 2014). Machine learning models can leverage existing water quality datasets to predict parameters at ungauged sites in near real time (Goodarzi *et al.,* 2023) [4].

Water pollution in River Benue poses risks to human and ecological health, but regular monitoring of water quality parameters is constrained by high costs and logistical challenges. Conventional lab-based analysis of water samples is time-consuming, requiring specialized equipment and technicians. This limits the frequency and spatial coverage of water quality data collection. Moreover, water quality varies dynamically in response to weather events, seasonal factors, and pollution discharges. Sparse manual sampling fails to capture this variability. The lack of continuous, real-time water quality data hampers efforts to identify pollution sources, enforce regulations, and design appropriate river basin management plans.

While sensor networks may potentially address gaps in monitoring, they involve high capital and maintenance expenses. An alternative cost-effective approach is to utilize available water quality datasets to developed machine learning models that can predict parameters at unmonitored locations and times. However, limited research has explored the application of advanced machine learning techniques for water quality prediction in developing country contexts like River Benue.

The aim of this research is to assemble a comprehensive physicochemical water quality dataset for River Benue at Makurdi reach and then explore trends and relationships within the water quality dataset and develop a machine learning predictive model using the collated dataset to predict water quality for domestic and irrigation purposes and classification model for identifying sources pollution. This study was limited to prediction of quality of surface water, quality parameters and the sources by machine learning for River Benue at Makurdi reach.

This study has contributed significantly to water quality research and management for River Benue by generating a comprehensive longitudinal water quality dataset that collates parameters from monitoring programs into a single repository. The state-of-the-art machine learning techniques has demonstrated the capabilities of predictive modeling for water quality monitoring in developing countries where data scarcity is a key challenge and has showcase advanced algorithms like artificial neural networks (ANN) and random forests (RF) for environmental modeling applications. It can also create cost-effective, near real-time water quality monitoring solutions along the river. In summary, this timely study has aided evidence-based decision-making to enhance the protection of River Benue, contributing to sustainable development in Nigeria's Benue basin region which is heavily dependent on the river.

Materials and method

Study area description

The study area encompassed the section of River Benue passing through the city of Makurdi, which is located in Benue State in the north-central region of Nigeria. Makurdi lies between longitude 8°32'00"E to 8°38'60"E and latitude

7°43'60"N to 7°48'60"N (Adeyemo *et al.*, 2008) [3]. The metro area population of Makurdi in 2023 was 454,000, a 3.65% increase from 2022. The climate in Makurdi is tropical with distinct wet and dry seasons. The rainy season typically occurs between April to October while the dry season is from November to March (Ologunorisa *et al.*, 2006; Idoko, 2023) [6, 11]. Mean annual rainfall is around 1200 mm. Monthly rainfall peaks in August and September during the peak monsoon. Average temperatures range between 23°C to 31°C (Adakayi, 2016; Ntuk *et al.*, 2020; Onuche *et al.* 2023) [1, 10, 12].

The land use in Makurdi is dominated by built-up urban cover and agricultural areas (Shabu *et al*., 2021). Major land use classes within the municipality are residential, commercial, industrial, transportation, agricultural fields, pastures and water bodies. Key industries include food processing, oil mills, aluminum works and cement factories (Hemba *et al.*, 2017) [5]. The river banks have patches of natural vegetation comprising grasses, shrubs and riparian woodlands. Floodplain wetlands and sandbars are also present along river edges (Hemba *et al.*, 2017) [5]. Agricultural areas grow crops like yam, rice, maize, sorghum, millet and cassava. Livestock rearing of cattle, sheep and goats is also practiced. (Tyav and Kuhe, 2020) [16]. Urbanization and agricultural activities in Makurdi contribute contaminants from municipal wastewater, industrial effluents, solid waste and agrochemical runoff into River Benue.

The River Benue stretch flowing through Makurdi is approximately 10 kilometers long with mean annual discharge of 35 billion m^3. The discharge exhibits substantial seasonal variations, with peak flows occurring in September during the peak of the rainy season (Adeyemo *et al.*, 2008; Salaudeen *et al.*, 2021) [3, 14]. It has an average depth of 5-10 meters and a width of 200-500 meters (Ade, 2022) [2]. River velocity ranges between 0.76 to 1.01 m/s depending on the rainfall intensity (Tamfuh *et al.*, 2022) [15]. This segment is an important surface water resource that provides water supply for the domestic, municipal, and agricultural needs of the Makurdi community and also supports the livelihoods of riparian communities through small-scale fishing activities along its banks (Ade, 2022) [2]. However, increasing pollution from point and non-point sources within Makurdi urban areas has degraded the water quality of River Benue. The river receives effluent discharges and runoff from the town which contains chemical contaminants, sewage, solid waste, and sediments (Iwar *et al.*, 2021; Utor *et al.*, 2021) [7]. Improving water quality monitoring is essential to safeguard the river ecosystem's health and suitability of water for human use. The current study focuses on the River Benue segment traversing Makurdi city given its ecological and economic significance for the town.

Methodology

Data collection

GIS map of Makurdi (Figure 1) was used to locate anthropogenic water pollution point source along the length of the Benue River at Makurdi reach for identification of sampling points. Water samples for analysis were collected between 8:00 am and 12:00 noon on sampling day performed weekly for four months of the rainy season of Makurdi. The samples were collected in sterile bottles of 1,500 mL capacity and rinsed three times at the depth of 20 cm.

Map 1: GIS Map Showing Water Sampling Points along River Benue at Makurdi Reach

Sample analysis

Surface water temperature were determined in situ on the field with measuring thermometers, while total dissolved solids (TDS), dissolved oxygen (DO) and sodium adsorption ratio (SAR): pH, Electrical Conductivity (EC), Chloride Content (Cl), Total Hardness (TH), Calcium (Ca), Magnesium (Mg), Sulphates (SO_4), Potassium (K), Sodium (Na) were examined in the laboratory using standard methods (APHA,1999). The data set were trained on the different machine learning models (Supervised learning algorithms) and the result used to select and validate the model as to identify the model that best predict water pollution levels of river Benue. Supervised learning algorithms try to *model relationships and dependencies between the target prediction output and the input features* such that we can predict the output values for new data based on those relationships which it learned from the previous data sets

Exploratory analysis

The quality-checked dataset was analyzed using IBM SPSS Statistics to understand data characteristics and trends. The steps performed were obtaining descriptive statistics of all water quality parameters including measures of central tendency (mean, median) and dispersion (standard deviation, variance) and Performing ANOVA analysis to detect significant seasonal differences in water quality. The exploratory analysis provided insights into underlying patterns, interactions and temporal effects in the collated dataset. This informed suitable input variable selection and data preprocessing strategies for model development

Model development for characterizing pollution level

Based on the exploratory analysis and domain knowledge, the following input variables; EC, SAR, SO_4, TDS, Ca, Mg, Na, Cl, TH, K, TDS, DO and pH were selected for developing the machine learning models based on their ability to characterize pollution levels from different sources such as sewage, agricultural runoff and industrial effluents in the river. The parameters also exhibited significant variations and correlations during exploratory analysis. The Machine Learning Algorithms were developed using MATLAB R2021b software. For predicting total dissolved solids (TDS), dissolved oxygen (DO) and sodium adsorption ratio (SAR), the Regression Learner app was used. This enabled creating and comparing multiple regression models like neural networks, support vector machines, decision trees and ensemble methods. Principal component analysis (PCA) was applied before model

building to explain 95% variance in the dataset using a smaller set of uncorrelated components. This overcame multicollinearity between the water quality predictors. The predictive accuracies of the trained models were evaluated using performance metrics like RMSE, R-squared, and confusion matrix. The optimal techniques were selected for each modeling task.

Model training and tuning

The regression models were trained using 70 % of the dataset selected randomly. The remaining 30 % was held out for testing model performance. Hyper-parameter tuning was done to optimize model accuracy. For neural networks, the number of hidden layers and units per layer were adjusted. For support vector machines, the kernel scale and box constraint were tuned. For decision trees, the merge threshold and maximum number of splits were optimized.

The optimal model was selected based on performance metrics computed on the test set. The best model was retrained on the full dataset for final deployment.

Model evaluation metrics

The predictive performance of the regression models was evaluated using; Root Mean Squared Error (RMSE) to measures model prediction error on the test set. Lower RMSE indicates a better fit. R-squared (R^2) to evaluate the proportion of variance in the response variable explained by the model. Higher R^2 denotes higher predictive power and Mean Absolute Error (MAE) to computes the average magnitude of errors in predictions. Lower MAE signifies better accuracy. A graphical analysis was performed to supplemented statistical metrics for evaluating regression model performance in terms of; Scatter plots of predicted vs actual values were used to visually assess prediction accuracy and deviations. Models with higher R^2 showed tighter fit along the diagonal, Residual plots visualized prediction errors. The random scatter of residuals around zero indicated a lack of bias and Predicted values were plotted over time and compared to actual values to evaluate temporal performance. The graphical evaluation provided additional insights into model generalization, biases and variability beyond the statistical metrics.

Model deployment

The optimized machine learning models were deployed for real-world application on new data. For the regression models, the predict function was used to generate predictions on new measurements of the input variables. The mathematical equations underlying the trained models were leveraged to estimate the total dissolved solids, dissolved oxygen and sodium adsorption ratio.

Results and discusion

Exploratory data analysis

In this section, the study embarks on a thorough exploration of the physicochemical water quality dataset for River Benue at Makurdi, Nigeria. The dataset encapsulates a diverse array of parameters crucial for understanding the ecological health of the river system. Through meticulous examination and analysis, we aim to unveil the intricate relationships, trends, and patterns inherent within the dataset, shedding light on the dynamics of water quality influenced by various pollution sources.

Physicochemical water quality dataset for river Benue at Makurdi

The physicochemical water quality dataset in Table 1 delineates a rich tapestry of measurements spanning critical physicochemical parameters recorded across different months and weeks, each associated with specific pollution sources. These parameters serve as fundamental indicators of water quality, elucidating the complex interplay between natural processes and anthropogenic activities impacting River Benue at Makurdi. The dataset unfolds over a period extending from July to October, encompassing various weeks within each month. This temporal granularity provides insights into seasonal variations and trends in water quality, offering a nuanced understanding of how the river system responds to environmental changes over time. Moreover, the identification of pollution sources for each measurement allows for the differentiation of spatial influences on water quality, discerning the localized impacts of urban drainage, agricultural practices, industrial discharges, and abattoir effluents. Among the parameters recorded, temperature (Temp) serves as a fundamental metric reflecting thermal variations within the river, influencing biological processes and solubility of chemical constituents. pH levels delineate the acidity or alkalinity of the water, crucial for assessing its suitability for aquatic life and ecosystem health. TDS and EC signify the concentration of dissolved ions, providing insights into salinity levels and overall water purity. DO concentrations are indicative of the water's oxygenation status, vital for supporting aquatic organisms and aerobic microbial activity. Cl serves as a disinfectant indicator, while TH reflects the concentration of divalent metal ions, impacting water usability and ecosystem health. Furthermore, Ca, Mg, SO4, K, Na, and SAR offer additional insights into the chemical composition and suitability of the water for various purposes.

Each pollution source introduces distinct chemical constituents and contaminants into the river system, influencing water quality parameters differently. For instance, industrial discharges may elevate TDS and EC levels due to the influx of salts and heavy metals, whereas urban drainage might contribute to fluctuations in pH and DO levels owing to organic matter decomposition and nutrient enrichment. Agricultural runoff can introduce excess nutrients such as potassium and phosphates, altering the nutrient balance and promoting eutrophication, while abattoir effluents may elevate organic loads and chlorine content, posing risks to aquatic life and human health.

The comprehensive dataset presented in Table 1 serves as a valuable resource for elucidating the intricate dynamics of water quality in River Benue at Makurdi. By analyzing temporal trends, spatial variations, and pollution source impacts, stakeholders can gain actionable insights into mitigating pollution risks, implementing targeted interventions, and safeguarding the ecological integrity of the river system. Furthermore, the findings derived from this exploratory analysis can inform policy decisions, resource allocation, and sustainable management practices aimed at preserving the water resources vital for the well-being of both ecosystems and communities reliant on River Benue.

Table 1: Physicochemical water quality dataset for River Benue at Makurdi

Month	Week	Pollution source	Temp	PH	Total Dissolved Solids (TDS)	Electric Conductivity (EC)	Dissolve Oxygen (DO)	Chlorine Content (Cl)	Total Hardness (TH)	Calcium (Ca)	Magnesium (Mg)	Sulphate (SO4)	Potassium (K)	Sodium (Na)	SAR
			oC		Mg/L	µs/cm	Mg/L	Mg/L	Mg/L	Mg/L	Mg/L	Mg/L	Mg/L	Mg/L	Mg/L
July	1	Abatoir	26	6.850	16.380	31.000	1.600	5.000	50.500	12.150	38.350	415.000	1.500	2.500	2.400
July	1	Urban Drainage	26	6.750	18.260	36.600	2.100	4.500	60.600	8.100	52.500	402.000	1.300	2.800	1.980
July	1	Agricultural	28	7.180	35.300	70.500	1.800	7.500	50.500	20.240	30.260	580.000	2.600	5.200	1.900
July	1	Industrial	29	6.990	16.870	33.700	2.000	5.000	40.400	16.190	24.210	420.000	1.100	1.900	1.700
July	2	Abatoir	25	8.350	972.000	1874.000	1.700	61.980	60.600	8.100	52.500	203.000	1.300	2.200	2.460
July	2	Urban Drainage	24	7.300	59.400	118.300	1.500	8.000	80.810	8.100	72.710	202.000	3.200	6.800	5.750
July	2	Agricultural	27	7.400	32.400	65.100	1.700	6.000	60.600	8.100	52.500	600.000	2.700	4.000	3.900
July	2	Industrial	29	6.950	20.300	40.300	1.700	4.000	50.500	12.150	48.350	205.000	1.200	2.400	2.600
July	3	Abatoir	29	6.650	18.080	33.800	1.600	8.000	60.600	12.150	48.450	209.500	1.400	2.700	17.080
July	3	Urban Drainage	27	6.450	96.000	193.200	1.400	15.500	101.010	16.130	84.880	301.000	6.500	15.000	32.730
July	3	Agricultural	28	6.620	38.600	77.300	1.600	8.000	101.010	4.050	96.960	601.000	3.100	5.400	19.090
July	3	Industrial	28	6.550	22.600	45.000	1.400	5.500	50.500	12.150	38.350	209.000	1.600	3.200	17.100
July	4	Abatoir	24	7.130	18.130	34.600	1.700	3.500	90.910	8.100	82.810	209.500	1.100	2.100	1.020
July	4	Urban Drainage	26	6.700	258.000	518.000	0.500	36.990	111.110	32.390	78.720	301.000	16.000	30.000	7.550
July	4	Agricultural	27	6.950	38.400	77.100	1.800	5.500	80.810	12.150	68.660	601.000	2.100	3.400	2.350
July	4	Industrial	28	6.800	17.350	34.700	1.800	4.500	90.910	12.150	78.760	209.000	1.100	1.700	1.490
August	1	Abatoir	24	6.420	16.960	31.600	0.600	4.500	383.830	12.150	371.680	240.000	1.600	2.500	2.890
August	1	Urban Drainage	24	6.260	185.500	389.000	1.400	31.990	80.810	16.190	64.620	420.000	21.000	42.000	24.250
August	1	Agricultural	24	6.650	35.000	70.200	1.500	6.000	60.600	56.680	3.920	595.000	3.000	4.500	3.450
August	1	Industrial	25	6.550	41.900	84.000	1.300	7.000	60.600	12.150	48.450	275.000	4.100	6.300	7.530
August	2	Abatoir	23	6.930	20.600	40.100	2.500	5.000	70.710	16.190	54.520	250.000	2.000	3.800	3.100
August	2	Urban Drainage	23	6.710	51.900	103.800	1.700	5.500	121.210	18.220	102.990	375.000	4.000	9.000	4.130
August	2	Agricultural	23	6.670	37.400	75.100	2.000	4.500	111.110	16.200	94.910	640.000	3.000	7.200	3.720
August	2	Industrial	24	6.630	15.370	30.700	2.100	4.490	101.010	16.190	84.820	253.000	1.400	350.000	3.130
August	3	Abatoir	23	7.030	18.150	33.600	1.800	6.000	40.400	12.150	28.250	360.000	0.900	2.700	2.040
August	3	Urban Drainage	23	6.780	46.000	92.100	1.500	6.000	80.810	12.150	68.660	330.000	4.800	13.000	8.220
August	3	Agricultural	24	6.680	28.700	57.600	1.600	9.000	90.910	8.100	82.810	615.000	3.100	3.800	2.190
August	3	Industrial	25	6.610	19.260	38.300	1.900	9.000	70.710	12.150	58.560	302.000	1.300	2.100	0.480
August	4	Abatoir	22	6.800	13.350	25.400	2.200	250.000	101.010	16.190	84.520	465.000	2.000	3.300	1.600
August	4	Urban Drainage	22	6.820	19.880	40.500	2.400	5.000	50.500	12.150	38.350	420.000	2.400	5.000	2.230
August	4	Agricultural	22	6.770	29.500	58.600	1.700	5.000	90.910	12.150	78.750	620.000	4.000	7.000	3.400
August	4	Industrial	24	6.790	18.520	37.300	1.200	5.000	50.500	12.150	38.350	415.000	2.100	4.300	2.200
September	1	Abatoir	25	6.460	49.800	94.400	6.000	4.500	70.710	12.150	58.560	170.000	1.500	8.500	2.560
September	1	Urban Drainage	25	6.450	45.200	90.400	1.300	6.500	111.110	24.290	86.820	250.000	0.900	8.300	2.530
September	1	Agricultural	27	6.450	32.600	65.200	2.000	4.000	101.010	28.340	72.670	160.000	0.900	5.600	1.790
September	1	Industrial	30	6.430	22.400	44.800	1.500	5.000	90.910	28.340	62.570	370.000	0.600	3.800	1.220
September	2	Abatoir	24	7.150	36.600	73.300	1.400	5.000	90.910	12.560	78.350	287.000	2.500	3.100	2.530
September	2	Urban Drainage	24	7.010	32.600	64.700	1.700	4.500	101.010	24.290	76.720	300.000	2.300	1.400	0.930
September	2	Agricultural	26	6.930	31.600	62.900	1.500	5.000	70.710	12.560	58.150	500.000	2.900	1.900	1.550
September	2	Industrial	27	6.910	20.900	41.900	1.100	4.500	101.010	24.290	76.720	390.000	2.000	1.200	1.070
September	3	Abatoir	22	6.890	18.880	37.500	1.900	6.000	80.810	28.340	52.470	310.000	1.700	1.200	0.720
September	3	Urban Drainage	23	6.650	146.700	294.000	1.300	23.490	70.710	20.240	50.470	340.000	13.000	32.000	9.990

September	3	Agricultural	25	6.820	37.200	74.000	1.100	8.000	50.500	16.190	34.310	570.000	3.500	5.100	2.490
September	3	Industrial	28	6.850	32.000	63.200	1.300	9.500	80.810	20.240	60.570	330.000	3.200	3.200	2.200
September	4	Abatoir	23	7.350	41.500	23.000	1.800	9.000	50.500	8.100	42.400	300.000	1.800	4.700	5.480
September	4	Urban Drainage	24	6.950	292.000	146.700	0.700	25.990	50.500	12.150	38.350	350.000	14.300	31.600	20.610
September	4	Agricultural	26	6.930	68.900	33.800	1.800	5.000	60.600	16.190	44.410	630.000	3.300	8.100	8.100
September	4	Industrial	29	6.850	48.200	24.500	1.100	5.000	80.810	16.190	64.620	330.000	2.500	5.900	7.320
October	1	Abatoir	24	7.140	34.800	66.500	1.200	49.980	40.400	20.240	20.160	200.000	2.200	7.100	9.170
October	1	Urban Drainage	24	6.940	76.900	155.700	1.900	55.980	50.500	12.150	38.350	260.000	7.600	18.300	36.600
October	1	Agricultural	25	6.880	33.000	65.500	1.900	35.990	70.710	20.240	50.470	504.500	3.500	7.700	7.700
October	1	Industrial	27	6.840	29.800	60.200	0.900	33.990	101.010	24.290	76.720	202.500	3.200	8.100	9.680
October	2	Abatoir	23	6.500	36.000	66.700	4.800	8.000	60.600	8.100	52.500	200.000	4.400	6.600	9.330
October	2	Urban Drainage	24	6.600	57.100	113.700	0.900	10.000	50.500	12.150	38.350	170.500	5.100	11.500	12.470
October	2	Agricultural	26	6.700	36.200	72.600	1.000	5.500	60.600	12.150	48.450	440.500	4.000	8.200	9.800
October	2	Industrial	28	6.800	33.000	55.600	4.200	6.000	50.500	12.150	38.350	140.000	3.300	7.400	8.550
October	3	Abatoir	24	6.250	32.200	63.900	4.800	5.500	60.600	12.150	48.450	160.000	3.000	3.300	3.400
October	3	Urban Drainage	25	6.160	101.500	205.000	2.300	15.500	121.210	16.190	105.020	210.000	7.500	15.600	13.000
October	3	Agricultural	25	6.280	44.100	87.400	0.800	9.500	101.010	16.190	84.820	340.000	3.300	5.600	5.330
October	3	Industrial	27	6.230	34.300	68.700	4.500	6.500	50.500	16.190	34.310	130.000	2.900	3.000	4.000
October	4	Abatoir	24	6.500	21.400	41.500	14.000	2.500	50.500	16.190	34.310	120.000	1.300	3.900	4.660
October	4	Urban Drainage	26	6.200	116.300	239.000	1.000	28.990	90.910	12.150	78.760	270.000	11.400	23.600	16.290
October	4	Agricultural	27	6.100	39.000	78.500	2.100	8.500	70.710	24.420	46.420	365.000	3.000	7.000	6.830
October	4	Industrial	29	6.300	33.700	65.800	6.000	7.000	50.500	12.150	38.350	160.000	1.500	5.500	6.350

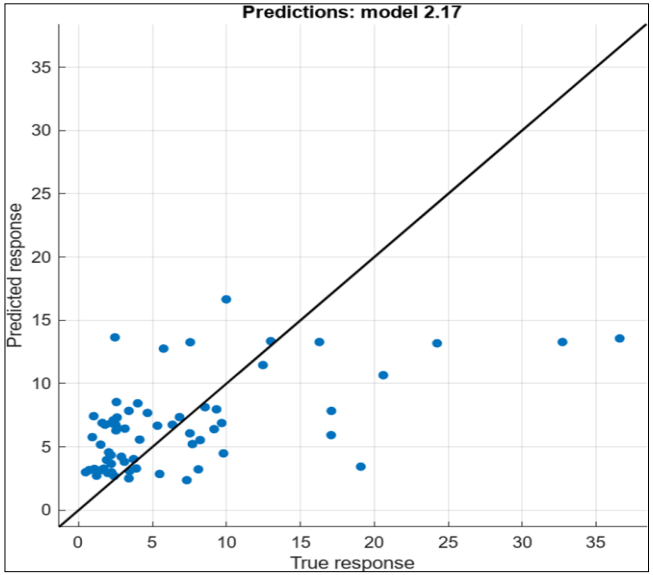

Fig 1: Plot of the predicted response vs the true response of sodium adsorption ratio (SAR)

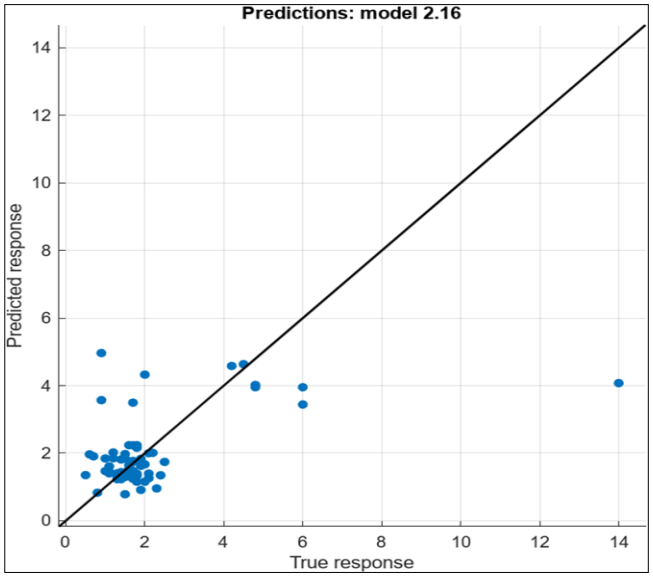

Fig 2: Plot of the predicted response vs the true response of dissolved oxygen (DO)

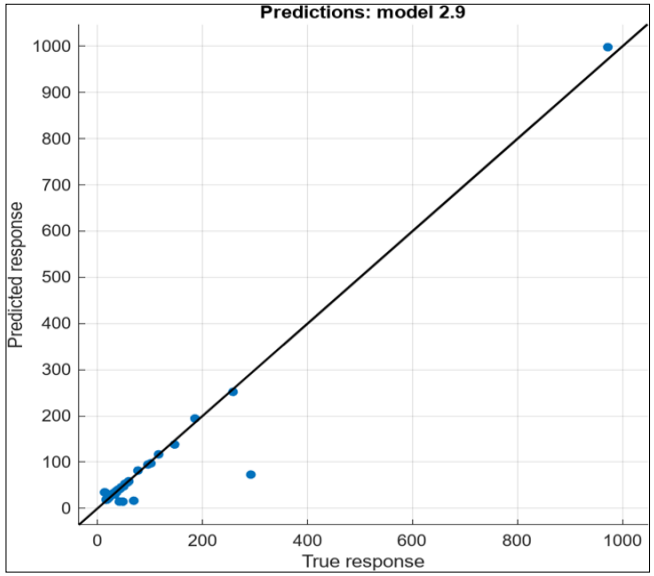

Fig 3: Plot of the predicted response vs the true response of total dissolved solids (TDS)

Development of machine learning predictive model for SAR, DO and TDS

In this result section, the study explores the development of a machine-learning predictive model for SAR, DO and TDS using the results from Table 2, which outlines the criteria for selecting the optimal predictive model. The development of an accurate predictive model for SAR holds significant implications for water quality management and environmental monitoring.

Table 2: Criteria for selection of sodium adsorption ratio (SAR) predictive model

Model Number	Model type	Status	RMSE (Validation)	MSE (Validation)	Rsquare (Validation)	MAE (Validation)
1	Linear Regression	Trained	17.148	294.063	-4.241	7.051
2	Linear Regression (interaction)	Trained	8.610	74.139	-0.321	5.579
3	Linear Regression (Robust)	Trained	12.737	162.224	-1.891	5.515
4	Stepwise Linear Regression	Trained	16.773	281.347	-4.118	6.950
5	Tree (Fine)	Trained	6.520	42.510	0.227	4.514
6	Tree (Medium)	Trained	6.725	45.223	0.177	4.742
7	Tree (Coarse)	Trained	7.421	55.067	-0.002	5.272
8	SVM (Linear)	Trained	16.068	258.189	-3.697	6.173
9	SVM (Quadratic)	Trained	62.427	3897.115	-69.892	11.781
10	SVM (Cubic)	Trained	8565.815	73373184.289	1334722.411	1082.615
11	SVM (Fine Gaussian)	Trained	6.685	44.684	0.187	3.833
12	SVM (Medium Gaussian)	Trained	6.901	47.627	0.134	4.173
13	SVM (Coarse Gaussian)	Trained	7.525	56.632	-0.030	4.484
14	Efficient Linear (Least Square)	Trained	16.909	285.910	-4.201	6.898
15	Efficient Linear (Linear SMV)	Trained	13.159	173.151	-2.150	5.745
16	Ensemble (Boasted Tree)	Trained	6.454	41.652	0.242	4.429
17	Ensemble (Bagged Tree)	Trained	6.069	36.836	0.330	4.204
18	Gaussian Process Regression (Squares Exponetial GPR)	Trained	6.785	46.032	0.163	4.761
19	Gaussian Process Regression (Matern 5/2 GPR)	Trained	6.781	45.985	0.163	4.727
20	Gaussian Process Regression (Exponetial GPR)	Trained	6.749	45.546	0.171	4.696
21	Gaussian Process Regression (Rational Quatradic GPR)	Trained	6.766	45.774	0.167	4.719
22	Neural Network (Narrow)	Trained	23.119	534.503	-8.723	7.397
23	Neural Network (Medium)	Trained	173.552	30120.393	-546.917	32.500
24	Neural Network (Wide)	Trained	57.449	3300.396	-59.037	17.921
25	Neural Network (Bilayered)	Trained	35.828	1283.612	-22.350	11.046
26	Neural Network (Trilayered)	Trained	221.592	49103.057	-892.228	37.945
27	Kernel (SMV)	Trained	7.183	51.592	0.062	4.218
28	Kernel (Least Square Regression)	Trained	6.276	39.382	0.284	4.343

*Chosen Model was Ensemble (Bagged Tree) with minimum leaf size of 8, number of learner is 30 and learning rate of 0.1
All models was trained with Principal Component Analysis (PCA) explaining at least 95% variance

Predictive model for SAR

Among the plethora of models evaluated (Table 2), the Ensemble (Bagged Tree) model stands out as the optimal choice, with the lowest RMSE and MSE values of 6.069 and 36.836, respectively. Additionally, it exhibits a relatively high R-squared value of 0.330, indicating a good fit to the data. The low MAE of 4.204 further confirms the accuracy and precision of the model in predicting SAR. The chosen Ensemble (Bagged Tree) model employs a bagging technique, combining multiple decision trees to reduce variance and improve prediction accuracy. With a minimum leaf size of 8, 30 learners, and a learning rate of 0.1, the model demonstrates robustness and stability in handling complex data patterns. All models were trained with Principal Component Analysis (PCA) to reduce dimensionality and enhance computational efficiency while preserving at least 95 % of the variance in the original dataset. This approach ensures that the predictive models capture the essential features contributing to SAR variability.

By accurately forecasting SAR levels, stakeholders can anticipate potential salinity issues in water bodies, implement timely interventions, and mitigate adverse effects on ecosystems and agricultural productivity. The Ensemble (Bagged Tree) model emerges as the optimal predictive model for estimating SAR in River Benue at Makurdi. Its robust performance metrics and effective feature engineering underscore its suitability for real-world applications in water quality assessment and management. The deployment of such predictive models can aid decision-makers in fostering sustainable water resource management practices and ensuring the ecological integrity of River Benue and its surrounding areas. Figure 1 displayed the plot is a scatter plot, where the x-axis represents the true or observed values of the sodium adsorption ratio (SAR), and the y-axis represents the predicted or modeled values of SAR. Each point on the plot corresponds to a pair of true and predicted SAR values.

The plot includes a diagonal line, which represents the line of perfect agreement between the true and predicted values. If all the points were to lie exactly on this line, it would indicate that the predicted values perfectly match the true values. However, in the given plot, the points are scattered around the line of perfect agreement, indicating that there are discrepancies between the true and predicted SAR values. Some points lie above the line, meaning that the predicted SAR values are higher than the true values, while others lie below the line, indicating that the predicted values are lower than the true values.

The spread of the points around the line of perfect agreement reflects the overall accuracy and bias of the ensemble model. A tighter clustering of points suggests that the model can make accurate predictions across the range of SAR values, while a wider spread may indicate higher prediction errors or biases in certain regions of the SAR range. The x-axis represents the true or observed SAR values, likely obtained from laboratory measurements or field observations, while the y-axis shows the corresponding predicted SAR values from the ensemble model. The scatter of points around the line of perfect agreement (the diagonal line) indicates the level of agreement between the true and predicted SAR values. Points lying on the line represent instances where the model's prediction perfectly matches the true value, while points deviating from the line represent prediction errors. It is important to note that ensemble methods like Bagged Trees often exhibit improved accuracy and robustness compared to individual decision trees, as they can reduce the impact of over fitting and leverage the collective strengths of multiple models. However, the performance of the ensemble model may still be influenced by factors such as the quality and representativeness of the training data, the selection of input features, and the specific hyper parameters used in the model training process.

Predictive model for DO

In this result section, the study delves into the development of a machine-learning predictive model for Dissolved Oxygen (DO) using the findings from Table 2, which delineates the criteria for selecting the optimal predictive model. Table 3 provides a comprehensive overview of various predictive models along with their respective performance metrics, including Root Mean Squared Error (RMSE), Mean Squared Error (MSE), R-squared (Rsquare), and Mean Absolute Error (MAE) on the validation dataset. The selection of the ideal model is contingent upon its ability to minimize errors and maximize the explained variance.

Table 3: Criteria for Selection of Dissolved oxygen (DO) Predictive Model

Model Number	Model type	Status	RMSE (Validation)	MSE (Validation)	Rsquare (Validation)	MAE (Validation)
1	Linear Regression	Trained	1.955	3.823	-0.047	1.195
2	Linear Regression (interaction)	Trained	18.661	348.224	-94.374	3.476
3	Linear Regression (Robust)	Trained	2.011	4.044	-0.108	0.906
4	Stepwise Linear Regression	Trained	1.965	3.862	-0.058	1.214
5	Tree (Fine)	Trained	1.908	3.641	-0.029	0.985
6	Tree (Medium)	Trained	1.679	2.818	0.203	0.838
7	Tree (Coarse)	Trained	1.881	3.539	0.000	0.964
8	SVM (Linear)	Trained	1.954	3.818	-0.079	0.905
9	SVM (Quadratic)	Trained	4.713	22.215	-5.280	1.392
10	SVM (Cubic)	Trained	14.671	215.244	-59.851	2.665
11	SVM (Fine Gaussian)	Trained	1.719	2.955	0.165	0.837
12	SVM (Medium Gaussian)	Trained	1.849	3.420	0.033	0.832
13	SVM (Coarse Gaussian)	Trained	1.910	3.647	-0.031	0.840
14	Efficient Linear (Least Square)	Trained	1.927	3.712	-0.050	1.174
15	Efficient Linear (Linear SMV)	Trained	1.957	3.831	-0.083	0.915
16	Ensemble (Boasted Tree)	Trained	1.580	2.497	0.294	0.790
17	Ensemble (Bagged Tree)	Trained	1.583	2.505	0.292	0.734
18	Gaussian Process Regression (Squares Exponetial GPR)	Trained	1.680	2.822	0.202	0.907
19	Gaussian Process Regression (Matern 5/2 GPR)	Trained	1.661	2.758	0.220	0.879
20	Gaussian Process Regression (Exponetial GPR)	Trained	1.666	2.776	0.215	0.861
21	Gaussian Process Regression (Rational Quatradic GPR)	Trained	1.680	2.822	0.202	0.907
22	Neural Network (Narrow)	Trained	2.120	4.496	-0.271	1.102
23	Neural Network (Medium)	Trained	2.457	6.036	-0.706	1.251
24	Neural Network (Wide)	Trained	3.516	12.363	-2.495	1.763
25	Neural Network (Bilayered)	Trained	7.187	51.647	-13.601	1.982
26	Neural Network (Trilayered)	Trained	3.440	11.831	-2.345	1.435
27	Kernel (SMV)	Trained	1.672	2.797	0.209	0.875
28	Kernel (Least Square Regression)	Trained	1.647	2.713	0.233	0.889

*Chosen Model was Ensemble (Boasted Tree) with minimum leaf size of 8, number of learner is 30 and learning rate of 0.1
All models was trained with Principal Component Analysis (PCA) explaining at least 95% variance

Among the plethora of models assessed, the Ensemble (Boosted Tree) model emerges as the most suitable choice for predicting Dissolved Oxygen (DO). It boasts the lowest RMSE and MSE values of 1.580 and 2.497, respectively. Additionally, the model exhibits a notably high R-squared value of 0.294, indicating a robust fit to the data. Moreover, the model achieves a commendably low MAE of 0.790, underscoring its accuracy and precision in predicting DO levels. The chosen Ensemble (Boosted Tree) model leverages boosting techniques to amalgamate multiple decision trees, thereby enhancing predictive performance and mitigating over fitting. With a minimum leaf size of 8, 30 learners, and a learning rate of 0.1, the model demonstrates resilience and efficacy in handling intricate data patterns. All models undergo training with Principal Component Analysis (PCA) to condense dimensionality and bolster computational efficiency while retaining at least 95% of the variance in the original dataset. This methodology ensures that the predictive models encapsulate the pivotal features contributing to DO variability.

The development of an accurate predictive model for Dissolved Oxygen (DO) bears significant ramifications for

water quality management and environmental stewardship. By accurately forecasting DO levels, stakeholders can anticipate oxygen deficiencies in aquatic ecosystems, implement timely interventions, and safeguard the ecological equilibrium of water bodies. The Ensemble (Boosted Tree) model emerges as the optimal predictive model for estimating Dissolved Oxygen (DO) in River Benue at Makurdi. Its robust performance metrics and effective feature engineering underscore its applicability in real-world scenarios, particularly in water quality assessment and ecological conservation efforts. The deployment of such predictive models holds promise for fostering sustainable water resource management practices and ensuring the ecological integrity of River Benue and its surrounding environs. Based on the provided plot in Figure 2 and the fact that the Ensemble (Boosted Tree) machine learning model was used to generate this plot, this study provide the following analysis; The Ensemble (Boosted Tree) model is another type of ensemble learning technique that combines multiple decision tree models to improve predictive performance. In contrast to bagging, where the individual trees are trained independently, boosting is an iterative process where each subsequent tree is trained to Correct the errors made by the previous trees. This process assigns higher weights to the instances that were misclassified or had larger errors, effectively "boosting" the model's ability to learn from these challenging cases. In the context of this plot, the Ensemble (Boosted Tree) model was used to predict the dissolved oxygen (DO) levels based on various input features or predictors. The x-axis represents the true or observed DO values, likely obtained from water quality measurements or field observations, while the y-axis shows the corresponding predicted DO values from the ensemble model.

Similar to the previous plot, the scatter of points around the line of perfect agreement (the diagonal line) indicates the level of agreement between the true and predicted DO values. Points lying on the line represent instances where the model's prediction perfectly matches the true value, while points deviating from the line represent prediction errors. Upon visual inspection, it is observed that the points in this plot are more tightly clustered around the line of perfect agreement compared to the previous plot for SAR. This tighter clustering suggests that the Ensemble (Boosted Tree) model is providing more accurate predictions for DO values across the range of observed values. However, it is essential to note that the performance of ensemble models can vary depending on the problem domain, the quality and characteristics of the training data, and the specific hyper parameters used during model training and boosting iterations.

It is also worth mentioning that ensemble methods like Boosted Trees can be more prone to over fitting compared to bagging methods if not properly regularized or if the number of boosting iterations is excessive. Therefore, it is essential to employ techniques like cross-validation, early stopping, or other regularization methods to prevent over fitting and ensure the model's generalization ability. Overall, the provided plot suggests that the Ensemble (Boosted Tree) model is performing well in predicting dissolved oxygen levels, with a relatively high level of agreement between the true and predicted values. However, a more detailed quantitative analysis and evaluation of the model's performance would be required to draw more definitive conclusions.

Predictive model for TDS

In this results section, the study delve into the development of a machine learning predictive model for TDS using the findings from Table 4, which outlines the criteria for selecting the optimal predictive model. Table 4 provides an extensive evaluation of various predictive models along with their corresponding performance metrics, including RMSE, MSE, R-squared, and MAE on the validation dataset. The selection of the ideal model hinges upon its capacity to minimize errors and maximize the explained variance. Among the array of models scrutinized, the SVM with a Quadratic kernel emerges as the most suitable choice for predicting TDS. It boasts the lowest RMSE and MSE values of 29.162 and 850.433, respectively, indicating superior predictive accuracy. Additionally, the model demonstrates a high R-squared value of 0.949, implying a robust fit to the data. Furthermore, the model achieves a commendably low MAE of 7.832, underscoring its precision in estimating TDS levels. The chosen SVM (Quadratic) model with a quadratic kernel function utilizes automatic kernel scale, box constrain, and epsilon, showcasing versatility and adaptability in handling diverse

Table 4: Criteria for selection of total dissolved solids (TDS) predictive model

Model Number	Model type	Status	RMSE (Validation)	MSE (Validation)	Rsquare (Validation)	MAE (Validation)
1	Linear Regression	Trained	34.707	1204.607	0.930	14.076
2	Linear Regression (interaction)	Trained	101.981	10400.049	0.395	26.783
3	Linear Regression (Robust)	Trained	30.036	902.189	0.948	7.007
4	Stepwise Linear Regression	Trained	130.392	17002.078	0.011	28.997
5	Tree (Fine)	Trained	112.113	12569.412	0.250	31.713
6	Tree (Medium)	Trained	123.625	15283.070	0.087	43.004
7	Tree (Coarse)	Trained	129.492	16768.070	-0.001	57.760
8	SVM (Linear)	Trained	30.374	922.590	0.945	8.324
9	SVM (Quadratic)	Trained	29.162	850.433	0.949	7.832
10	SVM (Cubic)	Trained	14094.234	198647429.835	-11859.807	1800.730
11	SVM (Fine Gaussian)	Trained	124.967	15616.709	0.068	32.380
12	SVM (Medium Gaussian)	Trained	125.138	15659.469	0.065	33.808
13	SVM (Coarse Gaussian)	Trained	127.263	16195.904	0.033	37.581
14	Efficient Linear (Least Square)	Trained	34.308	1177.007	0.930	15.241
15	Efficient Linear (Linear SMV)	Trained	35.213	1239.972	0.926	16.656
16	Ensemble (Boasted Tree)	Trained	119.764	14343.477	0.144	37.535
17	Ensemble (Bagged Tree)	Trained	116.264	13517.233	0.193	33.411

18	Gaussian Process Regression (Squares Exponetial GPR)	Trained	120.849	14604.498	0.128	32.422
19	Gaussian Process Regression (Matern 5/2 GPR)	Trained	124.531	15507.999	0.074	37.993
20	Gaussian Process Regression (Exponetial GPR)	Trained	121.536	14771.017	0.118	31.753
21	Gaussian Process Regression (Rational Quatradic GPR)	Trained	123.210	15180.710	0.094	36.148
22	Neural Network (Narrow)	Trained	248.145	61575.882	-2.677	58.059
23	Neural Network (Medium)	Trained	38.812	1506.344	0.910	16.819
24	Neural Network (Wide)	Trained	107.176	11486.707	0.314	33.384
25	Neural Network (Bilayered)	Trained	81.594	6657.640	0.602	24.658
26	Neural Network (Trilayered)	Trained	47.869	2291.420	0.863	21.699
27	Kernel (SMV)	Trained	128.060	16399.306	0.021	38.256
28	Kernel (Least Square Regression)	Trained	122.546	15017.540	0.103	41.239

*Chosen Model was SVM (Quadratic) with quadratic kernel function using automatic kernel scale, box constrain and epsilon
All models was trained with Principal Component Analysis (PCA) explaining at least 95% variance

Datasets. This model is adept at capturing non-linear relationships inherent in TDS data, thereby enhancing predictive performance and generalization capabilities.

All models undergo training with PCA to reduce dimensionality while retaining at least 95% of the variance in the original dataset. By condensing feature space, PCA facilitates computational efficiency and alleviates multicollinearity issues, thereby enhancing the predictive efficacy of the models. The development of an accurate predictive model for TDS holds significant implications for water quality assessment and environmental monitoring efforts. By accurately forecasting TDS levels, stakeholders can discern potential water contamination issues, implement timely interventions, and mitigate adverse environmental impacts. The SVM (Quadratic) model stands out as the optimal predictive model for estimating TDS in River Benue at Makurdi. Its robust performance metrics and effective feature engineering underscore its applicability in real-world scenarios, particularly in water quality management and environmental conservation endeavors. The deployment of such predictive models augurs well for fostering sustainable water resource management practices and preserving the ecological integrity of River Benue and its surrounding ecosystems. Based on the Figure 4 and the fact that the SVM (Quadratic) machine learning model was used to generate this plot, the study provide the following analysis:

The SVM (Support Vector Machine) is a powerful machine learning algorithm that can be used for both classification and regression tasks. In this case, the Quadratic kernel function was used for the SVM model, which allows for modeling non-linear relationships between the input features and the target variable TDS. In the context of this plot, the SVM (Quadratic) model was used to predict the total dissolved solids (TDS) levels based on various input features or predictors. The x-axis represents the true or observed TDS values, likely obtained from water quality measurements or field observations, while the y-axis shows the corresponding predicted TDS values from the SVM model. Similar to the previous plots, the scatter of points around the line of perfect agreement (the diagonal line) indicates the level of agreement between the true and predicted TDS values. Points lying on the line represent instances where the model's prediction perfectly matches the true value, while points deviating from the line represent prediction errors.

Upon visual inspection, the plot shows that the points in this plot are somewhat scattered around the line of perfect agreement, indicating moderate accuracy of the SVM (Quadratic) model in predicting TDS values. There are instances where the model over-predicts TDS (points above the line) and instances where it under-predicts TDS (points below the line). One characteristic of the SVM algorithm is its ability to handle non-linear relationships and high-dimensional data, which may be beneficial in modeling complex relationships between input features and TDS levels. However, the performance of the SVM model can be influenced by factors such as the choice of kernel function, regularization parameters, and the quality and characteristics of the training data. It is also worth noting that the SVM algorithm is sensitive to the scaling of input features and may require proper normalization or standardization of the data to ensure optimal performance.

Conclusion

The objectives of this study have been successfully met, leading to significant insights into predicting River Benue surface water quality parameters using machine learning models. Leveraging the collated dataset, the study developed machine learning predictive models capable of accurately predicting key water quality parameters, including total dissolved solids (TDS), dissolved oxygen (DO), and sodium adsorption ratio (SAR). These predictive models offer valuable insights into the dynamic nature of water quality in River Benue.

References

1. Adakayi P, Oche CY, Ishaya S. Assessment of the patterns of rainfall in northern Nigeria. Ethiopian Journal of Environmental Studies and Management,2016:9(5):554-566. DOI: 10.4314/ejesm.v9i5.3.
2. Ade MA. Flood Risk Assessment and management in the Benue Trough Nigeria. Unpublished doctoral thesis. University of Chester, 2022. Available from: https://chesterrep.openrepository.com/handle/10034/627752.
3. Adeyemo OK, Adedokun OA, Yusuf RK, Adeleye EA. Seasonal changes in physico-chemical parameters and nutrient load of river sediments in Ibadan city, Nigeria. Global Nest Journal,2008:10(3):326-336.
4. Goodarzi MR, Niknam AR, Barzkar A, Niazkar M, Zare Mehrjerdi Y, Abedi MJ, *et al*. Water Quality Index Estimations Using Machine Learning Algorithms: A Case Study of Yazd-Ardakan Plain, Iran. Water,2023:15(10):1876. Available from: https://doi.org/10.3390/w15101876.
5. Hemba S, Iortyom ET, Ropo OI, Daniel DP. Analysis of the physical growth and expansion of Makurdi Town using remote sensing and GIS techniques. Imperial Journal of Interdisciplinary Research,2017:3(7):821-827.

6. Idoko AE. Analysis of Annual and monthly rainfall variation in Benue State. Medicon Agriculture & Environmental Sciences,2023:4:09-23. Available from: https://themedicon.com/pdf/agricultureenvironmental/MCAES-04-105.pdf.
7. Iwar RT, Utsev JT, Hassan M. Assessment of heavy metal and physico-chemical pollution loadings of River Benue water at Makurdi using water quality index (WQI) and multivariate statistics. Applied Water Science,2021:11(7):124-135. Available from: https://doi.org/10.1007/s13201.
8. Kaushal SS, Maas CM, Mayer PM, Grant SB, Rippy MA, Shatkay RR, *et al*. Longitudinal stream synoptic monitoring tracks chemicals along watershed continuums: A typology of trends. Frontiers in Environmental Science,2023:11:1122485. Available from: https://doi.org/10.3389/fenvs.2023.1122485.
9. Maier HR, Galelli S, Razavi S, Castelletti A, Rizzoli A, Athanasiadis IN, *et al*. Exploding the myths: An introduction to artificial neural networks for prediction and forecasting. Environmental Modelling & Software,2023:167:105776. Available from: https://doi.org/10.1016/j.envsoft.2023.105776.
10. Ntuk JE, Ogwuche IE, Aligba EH, Shiada MS, Kever DG. Impact Assessment of Observed and Projected Climate on Maize Yield in Makurdi Metropolis, 2020.
11. Ologunorisa TE, Tersoo T. The changing rainfall pattern and its implication for flood frequency in Makurdi, Northern Nigeria. Journal of Applied Sciences and Environmental Management,2006:10(3):97-102. DOI: 10.4314/jasem.v10i3.17327.
12. Onuche P, Victoria A, Michael IA. Daily air temperature variation in Makurdi metropolis using analysis of variance model. International Journal of Science and Research Archive,2023:9(2):191-200.
13. Shabu T, Fate S, Ukula MK. Impact of urbanization on agricultural land in Makurdi local government area of Benue state, Nigeria. NASS Journal of Agricultural Sciences,2023:3(1):21-28.
14. Salaudeen A, Ismail A, Adeogun BK, Ajibike MA, Zubairu I. Evaluation of ground-based, daily, gridded precipitation products for Upper Benue River basin, Nigeria. Engineering and Applied Science Research,2021:48(4):397-405.
15. Tamfuh PA, Temga JP, Temgoua E, Woumfo ED, Zame PZO, Tchouatcha MS, *et al*. Characteristics, source area-weathering, sedimentary processes, tectonic setting and taxonomy of Vertisols developed on alluvial sediments in the Benue Trough of North Cameroon. Journal of Geosciences and Geomatics,2022:10:1-17.
16. Tyav TT, Kuhe JT. Herders and Farmers' Conflict in the Lower Benue Basin: A Historical Analysis. AIPGG Journal of Humanities and Peace Studies. Adeboye Institute for Peace and Good Governance, Redeemer's University Ede, Nigeria, 2020. Available from: https://ssrn.com/abstract=3709742.
17. Utor SO, Enokela OS, Awulu JO. Assessment of children and adult health risk factors associated with using portable water from the River Benue at Makurdi. International Research Journal of Environmental Sciences,2024:13(1):1-9.