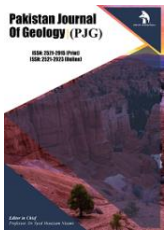
ZIBELINE INTERNATIONAL™ PUBLISHING

Pakistan Journal of Geology (PJG)

DOI: http://doi.org/10.26480/pjg.01.2025.26.33

ISSN: 2521-2915 (Print)
ISSN: 2521-2923 (Online)
CODEN: PJGABN



RESEARCH ARTICLE

PERFORMANCE OF MACHINE LEARNING MODELS FOR PREDICTING VOLUME OF WATER CONSUMED BY POOR URBAN HOUSEHOLDS WHERE THERE IS NO WATER DISTRIBUTION NETWORK

Taiwo, Tolu A*, Olusina, J.O., Hamid-Mosaku, A.I., Abiodun, O.E

Department of Surveying and Geoinformatics, Faculty of Engineering, University of Lagos, Nigeria.
Corresponding Author Email: tolutaiwo75@gmail.com

ARTICLE DETAILS

Article History:

Received 18 January 2025
Revised 15 January 2025
Accepted 26 February 2025
Available online 23 March 2025

ABSTRACT

Several studies have applied various techniques to model and predict water consumption in urban areas where there is water distribution network (WDN). This study examines the performance of machine learning models for predicting volume of water consumed by urban poor households where there is no WDN. Historical data of daily volume of water consumed was gathered through questionnaires, and integrated with socioeconomic data, weather data, property data and geospatial data. The datasets were passed through Pearson Correlation algorithm to select few features that correlate with the target variable. The selected features were inputted into four predictive models – Multilinear Regression (MLR), Random Forest (RF), Support Vector Regression (SVR), and Artificial Neural Networks (ANN). Three error metrics, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R squared (R^2) score, were used to measure the model performances. The models were validated with dataset collected where there is WDN. All four models performed very well during training, as they produced RMSE of 110 litres, 83 litres, 98 litres and 97 litres respectively, and R^2 score of 53%, 73%, 52% and 63% respectively. Significance test carried out on the results at 95% confidence level shows that there is no significant difference between model performance where there is WDN and where there is no WDN, which also confirms the validity of the dataset collected where there is no WDN.

KEYWORDS

Water consumption, Water distribution network, machine learning, poor households

1. INTRODUCTION

Clean water consumption by the poor directly affects their wellbeing, productivity and ability to escape the trap of poverty. A group researcher discovered that return-trip travel time greater than 30 minutes in order to fetch clean water was significantly associated with moderate-to-severe diarrhoea in Kenya (Nygren et al., 2016). While launching the 2023 UN Water Report, UNICEF in Nigeria declares that 78 million children suffer from poor water access, and are at risk of water-related crises (UNICEF, 2023).

Figure 1: The Sustainable Development Goals, highlighting goal 1 and goal 6 (Source: UN, 2015)

This is a reason the seventeen United Nations sustainable development goals (UN SDGs) shown in Figure 1.1 has water running through them. From poverty eradication in goal 1 to partnership for the goals in goal 17, water and the various opportunities it provides runs through the goals like lubricating oil, ensuring all the parts work together for smooth progress towards their achievements. UNESCO World Water Assessment Programme (WWAP, 2015) links poverty to lack of access to clean water, and access to clean water as part of the solution.

One way of improving access to clean water is expanding water distribution network (WDN) to connect every household to water services. It is the most effective way of improving sanitation and reducing transmission of water borne diseases (Hutton, 2004). The problem of access to clean water among the urban poor in low-income countries persists partly because there is no WDN in poor urban areas, and therefore no empirical data about the volume of water consumed by poor people. This leads them to seek water from different sources, such as water vendors, borehole water kiosks and well (UNDP,2015). They pay heavily for drinking water, which is provided by private businesses through vending (Chukwu, 2015). Since there is no automatic way of gathering data about volume of water consumed, it is difficult to model and predict volume of water consumed in poor urban areas. Yet it is important to model volume of water consumed in a poor urban area before a decision to extend water distribution network (WDN) to the area.

The aim of this study is to examine performance of machine learning models for predicting volume of water consumed by poor urban households where there is no WDN. Though there is WDN in rich areas nearby, yet the urban poor live in slums where there is no WDN, and so

Quick Response Code	Access this article online	
	Website: www.pakjgeology.com	**DOI:** 10.26480/pjg.01.2025.26.33

Cite The Article: Taiwo, Tolu A, Olusina, J.O., Hamid-Mosaku, A.I., Abiodun, O.E (2025). Performance of Machine Learning Models for Predicting Volume of Water Consumed by Poor Urban Households Where There Is No Water Distribution Network. *Pakistan Journal of Geology, 9(1): 26-33.*

they pay heavily for drinking water, which is provided by private businesses through vending. Safe water consumption directly affects people's wellbeing, productivity and ability to escape the trap of poverty. United Nations-Sustainable Development Goals (UN-SDG) recommends equal right to quality water for both the rich and the poor (WGF, 2012). This study is necessary to provide solution to the problem of access to clean water within poor urban areas in low-income countries. Accurately predicting volume of water consumed is the first step towards extending WDN to poor urban areas, and solving the problem of lack of access to clean water.

2. Material and Methods

2.1 Study Area

Nyanya-Mararaba is a border town between Federal Capital Territory (FCT) and Nasarawa State in Nigeria. It lies between Latitudes 8°24'54''N and 9°18'48''N, and Longitude 6°44'25''E and 7°35'15''E. The town's climate is largely sunny with diurnal temperatures ranging from 18.45 to 36.05 degrees Celsius. Rainfall varies from 0.0 to 400.0 mm/month. The climate is due to the town's location at the transition zone between the 'humid' south and the 'sub-humid' north.

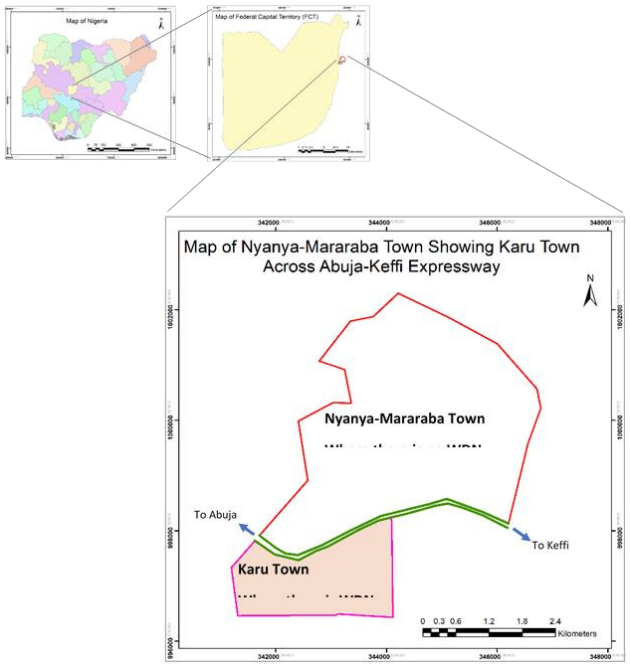


Figure 2: Map of Nigeria and Federal Capital Territory, showing Nyanya-Mararaba Town

The growing population of the Federal Capital City is concentrated at Nyanya-Mararaba town where an estimated 70% of the people working in the city live. Covering an area of 14 square kilometers, Nyanya-Mararaba is a peri-urban town consisting of informal dwellings where many poor people find shelter.

2.2 Methods

2.2.1 Datasets

2.2.1.1 Nyanya-Mararaba Dataset

This study made use of Nyanya-Mararaba dataset described in Taiwo et. al (2023). The dataset includes both primary and secondary sources of data. Primary sources of data include questionnaires administered to households in the study area, location of households and water points, collected with hand-held GPS. Secondary sources of data include: rainfall data downloaded from www.chrsdata.eng.uci.edu; digital elevation model downloaded from https://earthexplorer.usgs.gov/; Temperature data downloaded from www.weatherspark.com; and land use land cover (LULC) maps downloaded from https://maps.arcgis.com. Nyanya-Mararaba dataset; which consists of socioeconomic data, weather data, property data, historic data and geospatial data collected with questionnaires in Nyanya-Mararaba Town.

2.2.1.2 Karu Dataset

A second dataset, used as a control dataset, was collected in Karu Town where there is WDN. The dataset was collected from FCT Water Board in form of consumer bills. A bill contains the following data: date, previous meter reading, current meter reading, days of usage, multiplier, cubic unit consumed and current charge. The dataset consists of 2000 records covering January to December 2022. This dataset was used to validate Nyanya-Mararaba dataset where there is no WDN. Table 1 and Table 2 highlight the first ten records in Karu dataset before and after data preprocessing.

Table 1: Karu Dataset – Raw Data

Id	Service Type	Date	Days Of Usage	Previous Meter Reading	Current Meter Reading	Multiplier	Cubic Unit Consumed	Current Charge
KR1	DOMESTIC	17/01/2022	23	97782	97798	110	16	1760
KR2	DOMESTIC	01/01/2022	24	58545	58578	110	33	3630
KR3	DOMESTIC	21/01/2022	21	66347	66358	110	11	1210
KR4	DOMESTIC	12/01/2022	20	13269	13310	110	41	4510
KR5	DOMESTIC	05/01/2022	25	42475	42506	110	31	3410
KR6	DOMESTIC	17/01/2022	20	47886	47924	110	38	4180
KR7	DOMESTIC	02/01/2022	21	59153	59195	110	42	4620
KR8	DOMESTIC	29/01/2022	28	43750	43768	110	18	1980
KR9	DOMESTIC	07/01/2022	27	86046	86067	110	21	2310
KR10	DOMESTIC	05/01/2022	24	98300	98335	110	35	3850

Source: FCT Water Board (2023)

Table 2: Karu dataset after data cleaning and feature creation

ID	DATE	Days of usage	Rainfall	Ave temp	Volume in cubic meter	Volume in liter per day	Amount spent per day
KR1	17/01/2022	23	0	75	16	695	80
KR2	01/01/2022	24	0	75	33	1375	151
KR3	21/01/2022	21	0	75	11	523	57
KR4	12/01/2022	20	0	75	41	2050	290
KR5	05/01/2022	25	0	75	31	1240	136
KR6	17/01/2022	20	0	75	38	1900	209
KR7	02/01/2022	21	0	75	42	2000	220
KR8	29/01/2022	28	0	75	18	642	55
KR9	07/01/2022	27	0	75	21	777	85
KR10	05/01/2022	24	0	75	35	1458	160

2.2.2 Data Pre-processing

Data pre-processing includes feature engineering processes, which include data cleaning, data scaling, categorical encoding and feature creation in Geographic Information System (GIS). Four new features were created: volume of water consumed in litre per day, shortest distance, height difference, and LULC type. It is necessary to create these features for the following reasons: One, volume of water consumed is better measured in litre per capita per day, which is deduced from volume of water consumed per day that is retrieved from the questionnaires. Two, shortest distance, height difference and LULC type add geospatial variables to the dataset.

2.2.2.1 All Features

After feature creation and transformation, there were 30 features all together as shown in Table 3. Volume in liter per day is the dependent variable, the feature to be predicted. Thus, there are 29 independent variables, that is, predictors or explanatory features.

Table 3: All features identified in the study

Feature	Type	Description
Volume	Historic	Volume of water consumed in liter per capita per day
ID	Socioeconomic	Identification number of each household
Household income	Socioeconomic	Total income of each household in a month
Education	Socioeconomic	Highest level of education in the household
Household size	Socioeconomic	Number of persons in the household
Rainfall	Weather	Amount of rainfall per month
Ave temp	Weather	Average temperature per month
Travel time	Geospatial	Time it takes to get water and return
Amount spent	Socioeconomic	Amount spent on water per day
Willingness to pay	Socioeconomic	The amount of money a household is willing to pay if it is connected to piped water network
Kitchen Sink	Property	Presence of water sink in the kitchen
ToiletWC	Property	Presence of WC system in the toilet
Garden	Property	Presence of garden in the yard
Car	Property	Ownership of motor vehicle
Shortest distance	Geospatial	Shortest distance to the nearest water point from a household
Height diff	Geospatial	Difference between the household elevation and the nearest water point
Gender_male	Socioeconomic	Gender of respondent – male
Gender_female	Socioeconomic	Gender of respondent – female
Method_carried	Socioeconomic	Method of accessing water – carried from water point
Method_delivered	Socioeconomic	Method of accessing water – delivered to household
Method_inyard	Socioeconomic	Method of accessing water – water point in yard
Method_waterboard	Socioeconomic	Method of accessing water – piped water to the household
Method_well	Socioeconomic	Method of accessing water – well
Availability_not_often	Socioeconomic	Availability of water – not often
Availability_often	Socioeconomic	Availability of water – often
Quality_poor	Socioeconomic	Quality of water – poor
Quality_fair	Socioeconomic	Quality of water – fair
Quality_good	Socioeconomic	Quality of water – good
Quality_very good	Socioeconomic	Quality of water – very good
LULC	Geospatial	Land use land cover type

2.3 Feature Selection

A group researcher carried out experiments with Nyanya-Mararaba datasets to select few best features that produced optimal model performance (Taiwo et al., 2023). They experimented with five feature selection techniques: Pearson Correlation (PC), Information Gain (IG), Recursive Feature Elimination (RFE), Least Absolute Shrinkage and Selection Operator (LASSO) and Principal Component Analysis (PCA). The experiments selected nine features out of the twenty-nine; that is, household income, household size, rainfall, average temperature, travel time, amount spent, willingness to pay, shortest distance and height difference. These formed the data input in the machine learning models.

2.4 Machine Learning Techniques (MLT)

Four machine learning techniques (Multilinear Regression, Random Forest, Support Vector Regression, and Artificial Neural Network) were chosen for this work. The four MLT were chosen because they represent MLT in the supervised learning category for regression analysis, and it has been shown that Neural Network-based techniques outperform conventional techniques and provide effective solutions for many geospatial data analysis tasks (Kiwelekar et al., 2020).

2.4.1 Multilinear Regression (MLR)

Linear regression is a technique for investigating the relationship between an independent variable or feature and a dependent variable or outcome. Linear regression model describes the relationship between a dependent variable $\boldsymbol{y}$ and independent variables $\boldsymbol{x}$ with a straight line that is defined by equation (1):

$$\boldsymbol{y} = w_o + w_1 \boldsymbol{x} \tag{1}$$

In this expression, $\boldsymbol{y}$ is the vector of the response values. The $\boldsymbol{x}$ symbol describes the matrix of features which the algorithm uses to predict the $\boldsymbol{y}$ vector. $\boldsymbol{x}$ is a matrix that contains only numeric values. w_o and w_1 are parameters that the linear regression uses to create the prediction. Features are real-world entities that are being modelled. The ordinary least squares (OLS) method is used to estimate w_o and w_1. The OLS method seeks to minimize the sum of the squared residuals. As shown in Figure 2-5, the distance from each data point to the regression line is calculated, squared, and all squared errors are summed together. As in most linear regression models, the objective is to minimize the sum of squared errors, that is,

$$\text{Min} \sum_{i=1}^{m} y_i^p - y_i^o,)^2 \tag{2}$$

In a multilinear regression (MLR), where there are multiple features, the space now spans multiple dimensions, with each dimension being a feature. For instance, for five features the space is five-dimensional and the regression equation in matrix form becomes equation (3).

$$y^p = w_1 \boldsymbol{x_1} + w_2 \boldsymbol{x_2} + w_3 \boldsymbol{x_3} + w_4 \boldsymbol{x_4} + w_5 \boldsymbol{x_5} + w_o \tag{3}$$

MLR trains learning algorithms using labelled training data to understand the relationship between many features and an outcome. The result is a trained model. The trained model can then be leveraged to predict the outcome of new input data that were not seen before.

2.4.2 Random Forest (RF)

RF is an ensemble technique capable of performing both regression and classification tasks. Being an ensemble means that RF combine multiple decision trees through a technique called Bootstrap and Aggregation, commonly known as **bagging.** The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. RFs combine large number of decision tree models built on different sets of bootstrapped examples. Thus, the output does not depend on one decision tree but on multiple decision trees. RF has multiple decision trees as base learning models, and so we randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called bootstrap. In the case of a regression problem, the final output is the mean of all the outputs. This part is called **aggregation**.

According to Cutler et al., (2011), for a p-dimensional random vector $X = (X_1, \ldots, X_p)^T$ representing predictor variables and a random variable Y representing the real-valued target variable, an unknown joint distribution $P_{XY}(X, Y)$ is assumed. The goal is to find a prediction function $f(X)$ for predicting Y. The prediction function is determined by a loss function $L(Y, f(X))$ and defined to minimize the expected value of the loss

$$E_{XY}\left(L(Y, f(X))\right) \tag{4}$$

where the subscripts denote expectation with respect to the joint distribution of X and Y.

Intuitively, $L(Y, f(X))$ is a measure of how close $f(X)$ is to Y; it penalizes values of $f(X)$ that are a far away from Y. Typical choices of L are squared error loss $L(Y, f(X)) = (Y - f(X))^2$ for regression.

It turns out that minimizing $E_{XY}\left(L(Y, f(X))\right)$ for squared error loss gives the conditional expectation

$$f(x) = E(Y|X = x) \tag{5}$$

otherwise known as the regression function.

2.4.3 Support Vector Regression (SVR)

SVR is a regression algorithm that finds a hyperplane that best fits data points in a continuous space while minimizing the prediction error. This is achieved by mapping the input variables to a high-dimensional feature space and finding the hyperplane that maximizes the margin (distance) between the hyperplane and the closest data points, while also minimizing the prediction error. In contrast to OLS, the objective function of SVR is to minimize the coefficients, not the squared error. The error term is instead handled in the constraints, where we set the absolute error less than or equal to a specified margin, called the maximum error, ϵ (epsilon). We can tune epsilon to gain the desired accuracy of our model. Our new objective function and constraints are as shown in equations (6) and (7).

$$\text{Min} \frac{1}{2} \|w\|^2 \tag{6}$$

$$s.t. |y_i - w_i x_i| \leq \varepsilon \tag{7}$$

The algorithm solves the objective function as best as possible but some of the points still fall outside the margins. As such, we need to account for the possibility of errors that are larger than ϵ. We can do this with slack variables. The concept of slack variables is simple: for any value that falls outside of ϵ, we can denote its deviation from the margin as ξ. We know that these deviations have the potential to exist, but we would still like to minimize them as much as possible. Thus, we can add these deviations to the objective function, which becomes equations (8) and (9).

$$\text{Min} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} |\xi_i| \tag{8}$$

$$s.t. |y_i - w_i x_i| \leq \varepsilon + |\xi_i| \tag{9}$$

C and ξ (error penalty and slack variables respectively) are employed to prelude the influence of outliers and avoid overfitting.

SVR can handle non-linear relationships between the input variables and the target variable by using a kernel function to map the data to a higher-dimensional space, and it works with smaller amount of training samples and variables, and remain highly sensitive to variations in the variables (Karimi, 2016). This makes it a powerful tool for regression tasks where there may be complex relationships between the input variables and the target variable. A kernel is a set of mathematical functions that finds a hyperplane in the higher dimensional space that fits the data without increasing computational cost. The most widely used kernels include **Linear, Non-Linear, Polynomial, Radial Basis Function (RBF)** and **Sigmoid**. By default, RBF is used as the kernel.

2.4.4 Artificial Neural Network (ANN)

Recent advances in the field of deep-learning showed that Neural Network-based techniques outperform conventional techniques and provide effective solutions for many geospatial data analysis tasks (Kiwelekar, 2020). Multilayer perceptron (MLP) is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP is a deep learning method that uses back-propagation algorithm for training the network. The input layer consists of the inputs to the network, followed by a hidden layer consisting of neurons or hidden units, and then an output layer representing classes or patterns. The input pattern is presented to the network via the input layer, and the input signals are passed to the nodes in the next layer in a feed-forward manner. The summation of the output is called the output layer. MLPs are controlled by setting and adjusting weights between nodes. Initial weights are usually set at some random numbers and then they are adjusted during training.

An activation function, $g(x)$, defines the output of a neuron in terms of the linear combination of inputs. Activation functions provide nonlinearity to the output because it enhances or damps the values passing through it in a non-proportional way. Activation function also controls how well the network model learns the training dataset, and also defines the type of predictions the model can make. Hidden units can be described as accepting a vector of inputs x, computing an affine transformation of equation (10),

$$z = W^T x + b, \tag{10}$$

and then applying an element-wise nonlinear function $g(z)$. Most hidden units are distinguished from each other only by the choice of the form of the activation function $g(z)$.

There are different kinds of activation functions: rectified linear unit (ReLU), tanh function, and the logistic (sigmoid) function. The most commonly used in geospatial data analysis and water resources is the logistic activation function and, recently, ReLU.

Rectified linear unit (ReLU) is an excellent default choice of activation function because a model that uses it is easier to train and often achieves better performance (Goodfellow et al., 2016). ReLU is a simple function that returns the value of the input if it is positive, or value 0 if the input is 0 or negative. The function is linear for values greater than zero, meaning it has a lot of the desirable properties of a linear activation function when training a neural networks using back propagation. Yet, it is a nonlinear function as negative values are always output as zero. ReLU uses the maximum shown in equation (11) as the activation function.

$$g(z) = \max\{0, z\}. \tag{11}$$

Cost function (or loss function) helps to fit parameters to data, as it gives the value of the parameters that fit the data well when it is optimized. It is the penalty paid by the learning algorithm for fitting the model. That is, cost function gives the cost of predicting outcomes after fitting the model. Minimizing the cost function, which is also the optimization objective, the learning algorithm's prediction becomes more accurate, and the algorithm better maps the function to the features. According to a study, cost function can be defined by equation (12) (Ng, 2003).

$$J(w) = \frac{1}{2m}\sum_{i=1}^{m}(y^{p(i)} - y^{o(i)})^2 \tag{12}$$

that measures, for each value of the w's, how close the $y^{p(i)}$'s are to the corresponding $y^{o(i)}$'s. Feature $x^{(i)}$ and outcome $y^{(i)}$ represent the *i*th example in the training set, and m is the number of examples.

2.5 Implementing the Models

Four machine learning techniques, Multilinear Regression (MLR), Random Forest (RF), Support Vector Regression (SVR), and Artificial Neural Network (ANN), were employed for the modelling. The models were implemented with codes in Jupyter Notebook® environment where the experiments were carried out.

2.5.1 Experimental Design

The MLR, RF, SVR and ANN models, which were named *model* in the codes. They were each created as an instance of their respective classes in Scikit Learn® library: *SGDRegressor, RandomForestRegressor, LinearSVR and MLPRegressor*, which were each infused with a scaler to scale the dataset (Pedregosa et al., 2012). Thus, we had an empty model. The model was then trained or *fitted* with the training dataset. Training enables the model to learn the relationship between the explanatory variables and the target variable. After training, the model was ready for predictions. The results were evaluated. If the resulting errors were not within tolerance ($0.70 \leq R^2 \geq 0.99$), then the model needed refining through parameter tuning and it was built again. If the errors were within tolerance, then the model should be built with the parameters obtained. Afterward, the model was validated and tested.

2.5.2 Training, Validation and Testing

Training a model is a process of fitting it with a dataset so that the learning algorithm can learn the data and the relationships among the features. Training prepares the model for and gives it ability to generalize what it has learnt to new dataset it has not "seen" before. This allows the model to make prediction based on what it has learnt. A major problem with training is called overfitting, a situation in which the model learns well the training dataset but unable to generalize and make good predictions with new dataset. To avoid overfitting, the dataset was split into three: training data, validation data and test data. After training, validation dataset was used to assess the model performance. If satisfied, then the model was tested with the test dataset, which acts as a new data that the model had never "seen" before. The performance of the model when fed with the test dataset shows how it will perform when it encounters new data in the world. Standard split ratio found in literature are 60-20-20 percent, 70-15-15 percent, and 80-10-10 percent (Pragati, 2023). 80-10-10 was adopted for this study.

2.5.3 Model Evaluation

Three error metrics were used to measure the performance of the model: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R square (R^2). MAE is the average of the absolute difference between two continuous variables. For our purposes, we used the mean absolute error, expressed in equation (13), to measure the absolute difference between observed and predicted values of the target variable.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|Y_i^o - Y_i^p| \tag{13}$$

Mean Square Error (MSE) more greatly penalizes larger errors and is sensitive to outliers due to taking the squared difference between variables (Mueller, 2021). Also, MSE is good for values close to zero but not for large values. We want both large and small errors to be penalized equally so that we could obtain a robust model that would predict well. Therefore, we used root mean squared error, which is the square root of the value obtained from MSE.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(Y_i^o - Y_i^p)^2} \tag{14}$$

R^2 score is the amount of the variation in the target variable which is predictable from the explanatory variables as mathematically shown in equation 15. It is used to check how well observed results are reproduced by the model. Best possible score is 1.0, but the scores are stated as percentages in this study. A higher value for R^2 is desirable as it indicates better results.

$$R^2 = \left[\frac{\sum_{i=1}^{m}(Y_i^o - \bar{Y}_i^0)(Y_i^p - \bar{Y}_i^p)}{\sqrt{\sum_{i=1}^{m}(Y_i^o - \bar{Y}_i^0)^2 \cdot \sum_{i=1}^{m}(Y_i^p - \bar{Y}_i^p)^2}}\right]^2 \tag{15}$$

In the three expressions above, Y_i^o, Y_i^p, $\bar{Y}_i^o$, and $\bar{Y}_i^p$ are observed and predicted values and their respective average values; m is the number of observations or examples.

The three error metrics, MAE, RMSE and R^2 score, were hard-coded in Jupyter Notebook®, and the model performance was measured with the three error metrics each time the trained model made a prediction.

2.5.4 Validation

The Nyanya-Mararaba dataset where there is no WDN collected with questionnaires was validated with Karu dataset where there is WDN collected from FCT Water Board. The models were validated with the validation dataset during modelling. The two datasets were run through each of the models and the results were compared.

3. Results and Discussion

3.1 Dataset Validation

Nyanya-Mararaba dataset collected where there is no WDN is validated with Karu dataset where there is WDN. The models were trained, validated and tested with both datasets. The results in Table 2, Table 3 and Table 4 show the results of each model performance with both datasets. During training, the average R^2 score of the models using dataset where there is WDN is 98% while R^2 score using dataset where there is no WDN is 63%. As shown in Table 5, significance test carried out at 95% confidence level on the results declares that there is no significance difference between the results of both datasets. Thus, the Nyanya-Mararaba dataset collected with questionnaires where there is no WDN is validated with Karu dataset where there is WDN collected from FCT Water Board.

3.2 Model Performance Where There Is WDN and Where There Is No WDN

Table 4, Table5 and Table 6 also show the results of each model performance. It is obvious that Random Forest performed better than other models. During training, Random Forest gave RMSE of 48 liters and R^2 is 98% where there is WDN, and RMSE of 83 liters and R^2 score of 65% where there is no WDN. The performance of each model where there is WDN and where there is no WDN is depicted by the graph shown in Figure

2. From the graph, it is obvious that the difference between the actual volume and predicted volume of water consumed is negligible where there is WDN. This shows that the models were well trained; that is, the data was well learnt by the models, and the data was well fitted to the models. The two sets of graphs show that the difference between actual volume and predicted volume of water consumed is larger where there is no WDN than where there is WDN. The difference is due to the sources of the two datasets. Karu dataset was sourced from water meter readings while Nyanya-Mararaba dataset came from field work where number of water containers used per day was counted. The results of the significance test carried out for the datasets can be extended to the model performances since it was the results of their performances that were tested. Thus, the significance test shows that there is no significant difference between model performance where there is WDN and where there is no WDN.

Table 4: Model performance during TRAINING

Model	Where There is WDN			Where There is no WDN		
	MAE (litre)	RMSE (litre)	R^2 (%)	MAE (litre)	RMSE (litre)	R^2 (%)
Multilinear Regression	33	91	98	86	110	53
Random Forest	20	48	99	64	83	73
Support Vector Regression	254	317	98	76	98	62
Artificial Neural Network	32	93	98	75	97	63

Table 5: Model performance during VALIDATION

Model	Where There is WDN			Where There is no WDN		
	MAE (litre)	RMSE (litre)	R^2 (%)	MAE (litre)	RMSE (litre)	R^2 (%)
Multilinear Regression	34	104	97	85	108	51
Random Forest	26	61	99	72	93	69
Support Vector Regression	265	329	96	79	104	54
Artificial Neural Network	35	92	98	81	101	64

Table 6: Model performance during TESTING

Model	Where There is WDN			Where There is no WDN		
	MAE (litre)	RMSE (litre)	R^2 (%)	MAE (litre)	RMSE (litre)	R^2 (%)
Multilinear Regression	28	77	99	84	108	53
Random Forest	24	67	98	65	85	72
Support Vector Regression	288	347	97	92	115	57
Artificial Neural Network	20	27	97	85	105	58

3.3 Significance Test

Significance test was carried out with t-test at 95% confidence level on the datasets where there is WDN and where there is no WDN. The null hypothesis states that there is no significant difference between dataset where there is WDN and dataset where there is no WDN. The *p* value of each model is stated in Table 7. Since the *p* value is more than 0.05 for the models, except Support Vector Regression, there is a reason to accept the null hypothesis; that is, there is no significant difference between Karu dataset where there is WDN and Nyanya-Mararaba dataset where there is no WDN.

Null hypothesis: $H_0 : \mu = \mu_0$, that is, there is no significant difference between dataset where there is WDN and where there is no WDN.

Alternative hypothesis: $H_a : \mu \neq \mu_0$, that is, there is significant difference between dataset where there is WDN and where there is no WDN.

Table 7: Two-tails t-test at 95% confidence level

Model	*p* value	Degrees of freedom	Test Result
Multilinear Regression	0.5756	8	$H_0 : \mu = \mu_0$
Random Forest	0.2093	8	$H_0 : \mu = \mu_0$
Support Vector Regression	0.0007	8	$H_a : \mu \neq \mu_0$
Artificial Neural Network	0.3397	8	$H_0 : \mu = \mu_0$

3.4 Analysis of Variance (ANOVA)

Two-way ANOVA test was carried out to test model performances within the models on one hand, and between the models and where there is WDN/where there is no WDN. In the results shown in Table 8 the *p*-value is less than 0.05 within the models, while the *p*-value is greater than 0.05 between the models and where there is WDN/where there is no WDN.

Thus, the ANOVA test confirms that model performances where there is WDN and where there is no WDN are not significantly different, but their performances compared to one another are significantly different. Hence, Table 9, which shows descriptive statistics from the ANOVA test, confirms that Random Forest, with the least average performance error, performed better than other models.

Table 8: Result of two-way ANOVA

Source of Variation	SS	df	MS	F	P-value	F critical
Models (Row)	96592.28	3	32197.43	8.575769	0.000104	2.786229
WDN/NoWDN (Column)	40991.94	17	2411.291	0.642246	0.841357	1.827147
Error	191477.7	51	3754.465			
Total	329061.9	71				

Table 9: Descriptive statistics of model performances				
Model	**Count**	**Sum**	**Average**	**Variance**
Multilinear Regression	**18**	**1399**	**77.72222**	**791.7418**
Random Forest	**18**	**1218**	**67.66667**	**617.4118**
Support Vector Regression	**18**	**2828**	**157.1111**	**11463.4**
Artificial Neural Network	**18**	**1321**	**73.38889**	**802.134**

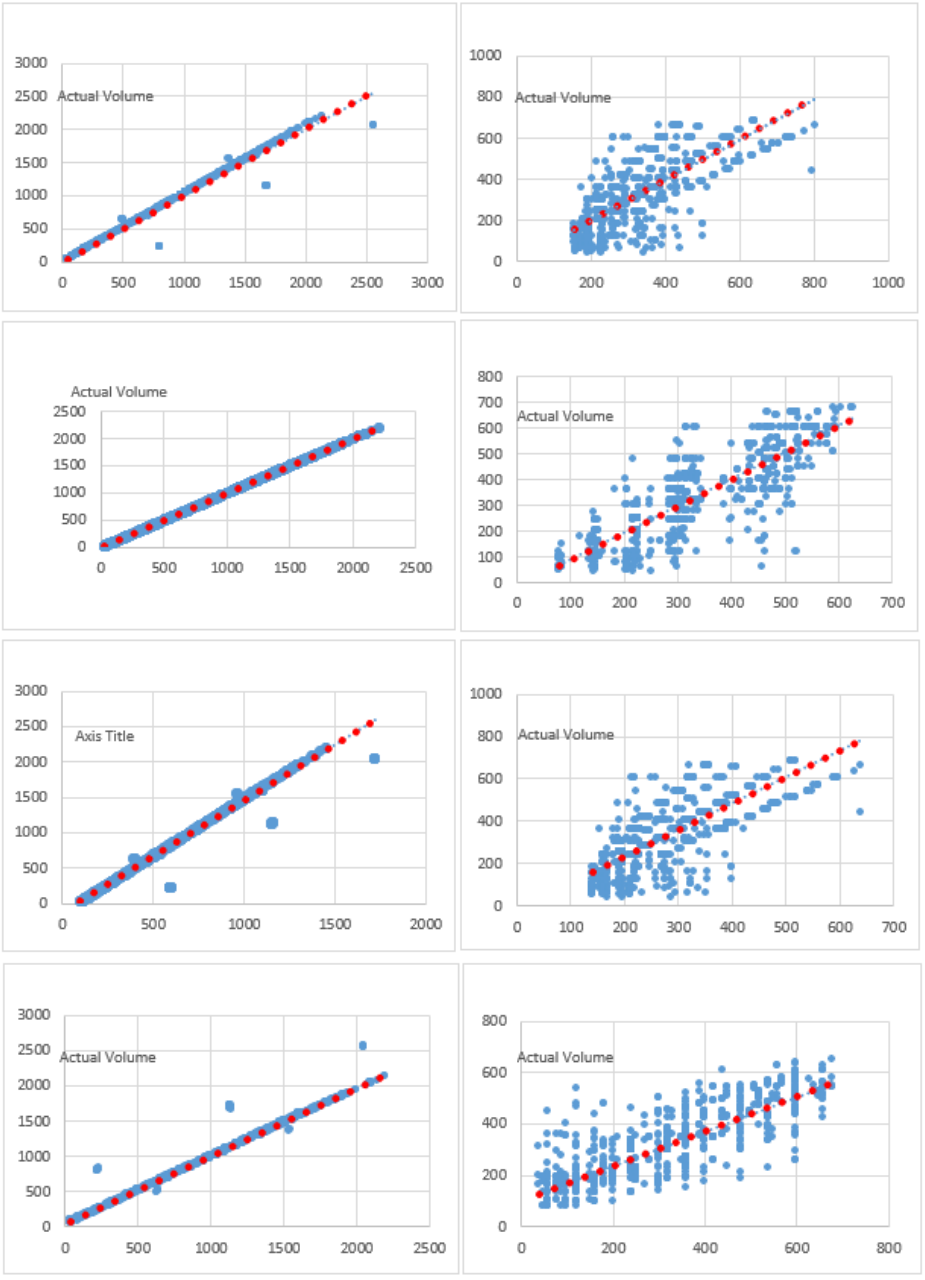


Figure 2: Actual volume and predicted volume where there is WDN and where is no WDN

4. Conclusion

This study examines performance of machine learning models for predicting volume of water consumed by poor urban households in Nyanya-Mararaba Town where there is no WDN. The dataset of volume of water collected with questionnaires where there is no WDN was validated with dataset of volume of water consumed where there is WDN. Four ML models were coded in Jupyter Notebook; they are multilinear regression (MLR), random forest (RF), support vector regression (SVR) and artificial neural network (ANN). The performances of the models were examined. All four models, MLR, RF, SVR and ANN, performed considerably well in predicting volume of water consumed by the urban poor, as they produced RMSE of 110 litres, 83 litres, 98 litres and 97 litres respectively, and R^2 score of 53%, 73%, 62% and 63% respectively, confirming that Random Forest performs better than other models. Significance test performed with t-test at 95% confidence level shows that there is no significant difference between model performances where there is WDN and where there is no WDN.

Statement of Interest

The authors declare that there is no significant competing financial, professional, or personal interests that might have influenced the performance or presentation of the work described in this manuscript.

Acknowledgement

The authors wish to thank the management of FCT Water Board for providing Water Bill dataset, which was used as control data in this study.

REFERENCES

Adamowski, J.F., 2008. Peak Daily Water Demand Forecast Modeling Using Artificial Neural Networks. Journal of Water Resources Planning and Management, 134 (2), Pp. 119-128. doi:10.1061/(asce)0733-9496(2008)134:2(119)

Adele, C., Richard, C.D., John R.S., 2011. Random Forest. In the book "Esseble Machine Learning: Methods and Applications, 45 (1), Pp. 157-176. Eds: Cha Zhang and Yungian Ma. DOI:10.1007/978-1-4419-9326-7_5

Casali, Y., Aydin, N.Y., and Comes, T., 2022. Machine learning for spatial analyses in urban areas: a scoping review. Sustainable Cities and Society, 85, Pp. 104050. doi:https://doi.org/10.1016/j.scs.2022.104050

Ghiassi, M., Zimbra, D., and Saidane, H., 2008. Urban Water Demand Forecasting with a Dynamic Artificial Neural Network Model. Journal of Water Resources Planning Management, 134, Pp. 138-146.

Goodfellow I., Bengio Y., Courville A., 2016. Deep Learning. Retrieved from www.deeplearning.org/ Accessed on 4th May, 2023.

Herrera, M., Lu´ıs, T., Joaqu´ın, I., and Rafael, P.E.G., 2010. Predictive models for forecasting hourly urban water demand. Journal of hydrology, 387, (1-2), Pp. 141– 150.

Jan, A., Prasher, S.O., Ozga-Zielinski, B., and Sliusarieva, A., 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in montreal, canada. Water Resources Research, 48 (1).

Karimi, S., Shiri, J., Kisi, O., and Shiri, A.A., 2016. Short-term and long-term streamflow prediction by using'wavelet–gene expression'programming approach. ISH Journal of Hydraulic Engineering, 22 (2), Pp. 148-162.

Kiwelekar, A.W., Mahamunkar, G.S., Netak, L.D., and Nikam, V.B., 2020. Deep learning techniques for geospatial data analysis. Machine Learning Paradigms: Advances in Deep Learning-based Technological Applications, Pp. 63-81.

Kopczewska, K., 2022. Spatial machine learning: new opportunities for regional science. The Annals of Regional Science, 68 (3), Pp. 713-755. doi:10.1007/s00168-021-01101-x

Medina, I., 2018. Predicting Short-Term Water Consumption for Multi-Family Residences. (Master of Science), University of Ontario, Canada.,

Mueller, J.P., and Massaron, L., 2021. Machine Learning for Dummies® John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, Canada.

Ng, A., 2013. Lecture Notes. Stanford University. Retrieved from https://www.datasciencecentral.com/lecture-notes-by-ng-full-set/ accessed on 26th July, 2023.

Nikparvar, B., and Thill, J.C., 2021. Machine Learning of Spatial Data. 10(9), Pp. 600.

Oyebode, O.K., 2020. Development of A Sustainable Evolutionary Inspired Artificial Intelligent System for Municipal Water Demand Modelling. (Doctor of Philosophy in Civil Engineering), University of KwaZulu-Natal, South Africa.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Louppe, G., 2012. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12.

Pragati, B., 2023. Train Test Validation Split: How To and Best Practices. Retrieved from https://www.v7labs.com/blog/train-validation-test-set

Schober P., Boer, C., and Schwarte L.A., 2018. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia. DOI: 10.1213/ANE.0000000000002864

Taiwo, T.A.O., Olusina, J.O., Hamid-Mosaku A.I., and Abiodun, O.E., 2023. An Investigation of the Performance of Geospatial Features in Machine Learning Feature Selection Techniques. Nigerian Journal of Environmental Sciences and Technology, 7 (2), Pp. 277-290. https://doi.org/10.36263/nijest.2023.02.0422 or https://nijest.com/current-issue/

UN, 2015. Transforming Our World: The 2030 Agenda for Sustainable Development. New York. Retrieved from https://sdgs.un.org/2030agenda, on 23rd February 2022.

WGF., 2012. Human Rights-Based Approaches and Managing Water Resources: Exploring the potential for enhancing development outcomes. Retrieved from Stockholm:

WWAP, 2015. The United Nations World Water Development Report 2015: Paris.