American Journal of Applied Sciences

Original Research Paper

Research on the Link Between Economic Development Variables and the Rate of Access to Drinking Water in Rural Senegal Using Machine Learning

[1]Anabilaye Moussa Coly, [2]Alioune Ly, [3]Ndolane Diouf, [2]Séni Tamba and [4]Issa Sakho

[1]Department of Science and Technology, Iba Der Thiam University, Thies, Senegal
[2]Department of Civil Engineering, Thies Polytechnic School (EPT), Thies, Senegal
[3]Department of Physics, Cheikh Anta Diop University, Dakar, Senegal
[4]Development of LMR Sciences, Advanced Technologies and Sustainable, Amadou Mahtar MBOW University of Diamniadio, Diamniadio Urban Hub, Dakar, Senegal

Article history
Received: 04-07-2023
Revised: 01-10-2023
Accepted: 09-10-2023

Corresponding Author:
Anabilaye Moussa Coly
Department of Science and Technology, Iba Der Thiam University, Thies, Senegal
Email: amcoly89@gmail.com

Abstract: After pre-processing a set of data collected from Senegalese state structures, official United Nations (UN), and world bank sites and after replacing missing values by interpolation using the mean between two values, we study the effect of macroeconomic development variables on the rate of access to drinking water in rural Senegal. To do this, we use Machine Learning (ML) techniques such as Linear Regression (LR), Decision Tree (DT), and Random Forest (RF) to identify a hidden correlation between the rate of access to drinking water and other variables. Based on the collected data, LR provides the best predictive accuracy, best RMSE (0.001), best MAE (0.011), and best R2 (96.9%) and help public development net received (ODA_NR_FBC) appears to be the most influential variable, predicting the rate of access to drinking water with greater precision than the other variables. This innovative approach can help us to better understand the factors influencing access to drinking water and to propose effective solutions.

Keywords: Data, Access to Drinking Water, Artificial Intelligence, Machine Learning, LR, DT, RF

Introduction

Faced with an increasingly greedy and demanding population, mankind's need for water is increasing significantly due to the evolution of human activities and lifestyles (Payen, 2023). These developments are creating ever-increasing competition between cities, farmers, industries, energy suppliers, and ecosystems. To this end, the United Nations predicts that by 2050, demand for water will be 20-30% higher than current levels (UNESCO, 2019). However, without proper management, the price to pay is high, not only in financial terms but also in terms of missed opportunities, harm to health, and damage to the environment. Without far-reaching reforms and significant improvements in water management, the situation is likely to worsen by 2050, as available resources become more uncertain (OECD, 2012).

That's why the United Nations (UN) has long been raising awareness of the world's drinking water problems. With this in mind, they invited all member countries to adopt an integrated approach to managing water resources through forums and conferences. These series of meetings led to the development of the Millennium Development Goals (MDGs) in 2000 (2000-2015) (Diallo *et al*., 2022) and the Sustainable Development Goals (SDGs) in 2015 (2015-2030). Thus, in July 2010, the UN General Assembly recognized the right to drinking water and sanitation as a human right (Payen, 2023) on par with the right to health. Therefore, achieving Sustainable Development Goal 6.1, which aims to ensure universal and equitable access to safe and affordable drinking water by 2030, becomes one of the prerequisites for eradicating poverty and promoting economic growth.

Access to drinking water is an indicator representing the proportion of the population with reasonable access to an adequate quantity of drinking water. According to the WHO, adequate drinking water represents at least 20 L of water per inhabitant per day. Reasonable access" is generally defined as the availability of drinking water

within a 15 min walk of the home[1]. Accordingly, the drinking water access rate (TAEP_Rur) is defined as the percentage of the population with access to an improved drinking water source, i.e., one that is protected from contamination, physically accessible, and provides enough water to meet basic needs (UNICEF, 2017).

Unfortunately, millions of people around the world live without access to this precious liquid. Senegal is a West African country where a large proportion of the population lives in rural areas (in 2022) (ANSD, 2015). Despite remarkable economic growth, access to drinking water in the most remote areas remains a major challenge for improving living conditions and fostering community development. In 2022, the percentage of access to basic drinking water at the national level was 86.2%, however, inequalities were observed across the country, with only 77% of rural dwellers benefiting from basic drinking water services, compared with 96.6% in urban areas[2]. Progress is still being made despite persistent high demand. However, the government's efforts to improve access to drinking water have failed to provide an effective response to the issue, with the result that some rural populations are still not included in the statistical data presented and use a variety of unconventional means of access, exposing them to consequences that penalize the State of Senegal every year in terms of public health. Today, the cost of the status quo in water resource management is already impacting over 10% of Senegal's GDP, due to extreme events and water-related pollution (World Bank, 2022).

Unequal access, water points far from homes, unsafe water quality, interruption of water distribution service, and faulty pumps due to lack of maintenance... is the daily reality for many residents of rural areas. This is why improving the supply of drinking water is often a priority for local authorities and populations alike (WSP, 2012).

Access to a quality water supply is an important issue for Senegal's development, but many rural communities still have difficulty gaining access. The reasons are multiple and complex, as access to drinking water depends on many factors, such as the availability of water resources, infrastructure, government policies, the economy, etc. The UN, in examining this issue in their 2018 report, went a step further by demonstrating that the secure supply of water of appreciable quality is intrinsically linked to all dimensions of development (UN-Water, 2018). To understand the variables likely to have an impact on the percentage of people with access to drinking water in the most disadvantaged areas, several macroeconomic development factors were examined.

Firstly, the number of people using at least basic drinking water services is crucial. Indeed, the more people use these services, the greater the need for investment in the extension of drinking water infrastructures.

Another factor to consider is the annual growth in Gross Domestic Product (GDP). Stronger economic growth may lead to greater investment in infrastructure, which in turn could boost water supplies to more remote areas.

Foreign direct investment is also important. An increase in such investment in the country can help finance drinking water access projects, which could help improve access rates.

Total annual withdrawals of freshwater for domestic use, industry, and total internal resources are all freshwater-related variables and have a significant impact on the rate of access to drinking water for rural dwellers. An increase in withdrawals indicates increased competition for water resources, making access to drinking water more difficult. However, an increase in withdrawals for industry can be positive if it leads to investment in infrastructure.

Growth in Gross National Income (GNI) per capita is also an important factor. An increase in GNI per capita may indicate an improvement in quality of life and greater variation in access to basic drinking water services in rural areas.

Net Official Development Assistance (ODA) received from imports of goods and services, gross capital formation and GNI is also relevant. An increase in this aid could help finance drinking water access projects in rural areas, thereby increasing the rate of access. Finally, the annual growth of the rural population must be taken into account.

An increase in this population may lead to a rise in demand for water and put additional pressure on existing infrastructures, which could have a negative impact on their rate of access to drinking water.

To understand the variables impacting the evolution of this rural water access indicator, several development factors were examined. However, few studies have explored the links between economic development variables and the rate of access to safe drinking water. Moreover, most studies have used traditional methods of statistical analysis, which may not be sensitive enough to detect the complex relationships between these variables. Consequently, this study uses sophisticated computational methods to explore the complex relationships between these variables and identify key models that can be used to improve access to drinking water in non-urbanized areas. More specifically, this study examines the macroeconomic factors that have an impact on access to drinking water in rural Senegal. How are these factors related to drinking water? How can machine learning help identify these links and predict the percentage of the population with

[1]https://www.actioncontrelafaim.org/a-la-une/tout-savoir-sur-lacces-a-leau-dans-le-monde/

[2]https://washdata.org/

access to drinking water in rural areas? By answering these questions using this innovative approach, this study could help to better understand the challenges associated with the supply of drinking water in rural areas and to develop more effective policies and programs in this field.

In this study, we will use machine learning to identify links between development variables and access to drinking water in rural Senegal. More specifically, it involves applying Machine Learning techniques such as Linear Regression (LR), Decision Tree (DT), and Random Forest (RF) to a dataset to identify the variables that accurately explain variation in the rate of access to drinking water. We evaluate the performance of our model by performing a correlation analysis and a prediction analysis. We measure prediction accuracy using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). From this analysis, we can see that net official development assistance received is the variable that most impacts and helps predict the rate of access to drinking water. The raw data used in this study are not suitable for direct use as model inputs. Thus, given that there are missing data, it is necessary to pre-process them before using them as input data in the ML models. The significance of each of these variables retained after preprocessing is given in Table 1.

However, in this study, the use of the macroeconomic development dataset has limitations and constraints. The quality and reliability of the analysis may be affected by incomplete or missing data. Furthermore, macroeconomic variables do not necessarily reflect local realities and the specific challenges faced by Senegalese rural communities in terms of drinking water supply. It is therefore essential to complement the analysis with qualitative data and field studies to better understand the factors influencing water supply in these areas. Finally, it is important to note that cultural factors, traditional water use practices, lifestyles and eating habits of local populations, etc., which can influence access to drinking water, are not taken into account in this set of macroeconomic development data. Consequently, a thorough understanding of local contexts is needed to identify the most effective solutions for improving access to drinking water in disadvantaged areas.

Contributions

The contributions of this study are:

- We used the LR, DT, and RF algorithms and identified LR as the best model for predicting the rate of access to drinking water in rural Senegal
- We found that APD_NR_FBC is the variable that participates most accurately in predicting the rate of access to drinking water
- This research may also have a potential impact on water resource management in rural Senegal. By identifying the factors that influence access to this water, this research can help develop more effective management strategies to ensure sustainable access to this resource
- This study contributes to filling the specific knowledge gap in the search for links between macroeconomic development variables and changes in the rate of access to drinking water in rural areas

Related Work

Several studies have been conducted to achieve sustainable development goal 6.1, which aims to ensure universal and equitable access to affordable drinking water by 2030. To achieve this goal, (Strong *et al*. 2020) mentioned that the global financing requirements for achieving SDG6 are estimated at 1,000 billion USD or 1.21% of the world's gross product. According to a study conducted by the (FAO, 2019), Access to Drinking Water (ADW) in Africa is closely linked to economic development and countries that have experienced sustained economic growth have also seen an improvement in ADW. Moreover, a study published by the African development bank (African Water Facility, 2017) found that investment in water and sanitation infrastructure has a beneficial effect on the economy.

In Senegal, specific studies have also been carried out to find a link between rural water supply and macroeconomic development variables. These studies highlight the importance of ADW for economic development in rural Senegal. They also highlight the need to strengthen public policies aimed at improving this access in order to stimulate the economy and reduce poverty in this region. In its correlation study on the effect of ADW on economic growth in the Dakar suburbs, (Diagne, 2009) observed that a number of indicators linked to people's standard of living, education, promotion of gender equality, and women's empowerment have a significant positive impact on drinking water. Similarly, in his study on public water services in rural Senegal, (Repussard, 2009) found that the construction of drinking water infrastructure in rural areas contributes to improving the living conditions of rural populations and creating jobs, which has a significant positive impact on the rural economy.

According to the Manual (MHS, 2017), drinking water is a key growth factor for economic development in rural areas.

The study by Bohbot (2008) differs from the others in its use of the traditional method of statistical correlation analysis. The latter showed that ADW is correlated with a number of macroeconomic variables, such as national wealth, GDP growth, HDI, population growth, etc., and that the correlation between ADW and these variables is significant.

Finally, in a study by Bonface and Jhon (2019), they explored the correlation between ADW and the standard of living of rural smallholder farmers in Kenya. Using survey data, they were able to demonstrate a significant link between ADW and the well-being of these farming households. Indeed, improved access to drinking water translates into a higher standard of living in rural areas.

Taking their study further, they discovered through regression analysis that access to drinking water is a key predictor of household well-being. This suggests that drinking water has a positive and significant effect on economic development in rural areas. However, the authors were unable to predict the Rate of Access to Drinking Water (RADW) or to identify the development indicator with the greatest influence on RADW. To fill this gap, an approach based on machine learning was developed.

Machine learning is a data processing technique that extracts information from large quantities of data. Furthermore, machine learning is the art of collecting data from observations of a phenomenon and building a model of that phenomenon from that data (Mitchell, 1997). This technique is widely used in various fields, including economic analysis, health, transportation, etc. An ML model is a specific algorithmic procedure for constructing a prediction function f from a training dataset (Lemberger *et al*., 2015). LR, DT, and RF are among the different types of ML models used in the design of prediction and forecasting systems (Briernat and Lutz, 2015; Grus, 2017; Dark, 2019; Lemberger *et al*., 2019).

LR is an ML technique for modeling linear relationships between variables. It is used to predict the dependent variable as a function of the independent variables. Using its supervised learning algorithm, the prediction can be made by analyzing the previous data to find a relationship between one or more independent variables and the dependent variable. In the case of this study, we use the linear regression function (Grus, 2017) from the Sklearn package (HAO and Ho, 2019) in Python to create the model from training data.

In this particular case, the use of linear regression algorithms is justified as it enables the relationship between continuous variables to be modeled, which is relevant to the study of the relationship between macroeconomic development variables and ADW in rural Senegal. In addition, linear regression is easy to interpret and makes it possible to quantify the impact of different variables on the RADW. It can be used to model the relationship between economic development variables and ADW, by identifying the coefficients that determine the impact of each variable on the RADW.

Decision trees are a machine learning method for creating prediction models using decision rules based on data characteristics. They are extremely flexible supervised and non-parametric ML models. They can be used for both classification and regression. DTs have the ability to handle heterogeneous data, including ordered variables, categorical variables, or a mixture of both, according to (Ndolane *et al*., 2021). The maximum depth of the tree is a significant hyperparameter that can improve the performance of the DT algorithm. For more details on decision tree algorithms, see (Timofeev, 2004; Studer *et al*., 1997). They can be used to identify the macroeconomic development variables that have the greatest impact on access to drinking water, by creating decision rules that predict access rates as a function of these variables. However, the use of decision-tree algorithms is justified because they enable non-linear relationships between variables to be modeled and complex interactions between variables to be detected. Decision trees are also easy to interpret and visualize the different paths leading to a decision.

RF algorithms are renowned for their ability to provide effective forecasts on a wide range of datasets. RF is capable of modeling highly non-linear relationships and is a very powerful ensemble method that combines with a DT ensemble, according to (Hajjem *et al*., 2014). A set of decision trees called a forest, is generated from a new training dataset, which is a randomly sampled subset of the original training dataset, according to (Ndolane *et al*. 2021). A subset of independent variables is randomly selected for tree division. The average of all the trees developed is the final result of the RF algorithm. Random forests are particularly useful for dealing with complex data sets with many variables and non-linear interactions. The use of random forest algorithms in this study is justified as it allows the combination of several decision trees to improve prediction accuracy and reduce the risk of overlearning.

RF is an extension of DT that combines several DT to improve prediction accuracy. It can be used to identify the macroeconomic development variables most important for predicting the rate of access to drinking water, by creating several decision trees and combining their predictions to obtain a more accurate prediction.

By combining these three approaches, it is possible to gain an in-depth understanding of the hidden relationships between macroeconomic development variables and the percentage of Senegal's rural population with access to drinking water. There are several advantages to using ML techniques in this study. Machine learning makes it possible to process large amounts of data and detect patterns that would not be apparent with traditional analysis. This can lead to more accurate predictions of the rate of access to drinking water. Machine learning algorithms can help uncover complex relationships between variables, predict future trends, and identify key factors influencing access to drinking water in non-urbanized areas. By accurately predicting the rate of access to drinking water, machine learning techniques can help governments and development organizations plan and implement effective policies and programs to improve ADW.

Materials

Materials Dataset and Machine Learning Algorithm

The hardware used in this study includes the Python programming language and libraries dedicated to data analysis. These include Numpy, Pandas, Matplotlib, SciPy, and scikit-learn. We used a DELL laptop with an Intel (R) Core (TM) i7-8565U CPU and 8.00 GB RAM installed.

Dataset

All the data used in this study, from 2000-2020, comes from the Senegalese authorities, the world bank, and the united nations. It is important to note that the United Nations and the world bank work closely with governments to collect data from other countries in order to monitor macroeconomic development indicators. This can be done either by household surveys or by collecting data from national institutions to ensure that the data collected is reliable and comparable. Most of the data used in this study is available online at the following addresses: https://donnees.banquemondiale, https://washdata.org/.

Data on population growth in rural areas comes from Senegal's National Agency for Statistics and Demography (ANSD), an official statistics agency. These data are also available on the world bank website mentioned above.

However, it is not advisable to select all the variables available in a data set, as this can lead to inaccurate forecasts.

For this purpose, a set of 15 measurement parameters expressed as percentages was initially used. Missing values were replaced by interpolation using the mean between two values. A correlation matrix was created to check the interdependence of variables. In cases where two variables were highly correlated with each other and with the target variable, the one that correlated most with the target variable was selected. This led to the selection of 11 variables and the elimination of four others.

However, after preprocessing, we obtained a dataset comprising 21 rows with 11 columns and TAEP_Rur as the target variable. The full statistical table for this data set is presented in the appendix. Table 1 shows the variables used to identify the trends highlighted in this study.

Machine Learning Algorithms

LR, DT, and RF are the machine learning algorithms used in this study:

- Linear Regression (LR): LR is a Machine Learning (ML) technique for modeling linear relationships between input variables and the target variable. The model provides coefficients for each input variable, indicating the effect of each variable on the target variable, making it easy to interpret. However, linear regression can be limited in its ability to capture non-linear relationships between variables. To optimize the performance of the linear regression model, the specific hyperparameters used in this study are those used in the Linear Regression () function of the sci-kit-learn library
- Decision Tree (DT): DT is a machine learning algorithm that can be used to identify complex relationships in a dataset, whether for classification or regression tasks. In this article, we focus on using DT to perform a regression task. Two of the most important hyperparameters are the maximum tree depth (max_depth) and the minimum number of samples required to split an internal node (min_samples_split). A higher maximum depth can lead to overlearning, while too low a maximum depth can lead to underlearning. In this study, max_depth = 7 and min_samples_split = 4. To learn more about decision tree algorithms, you can consult (Sullivan, 2017). Figure 1 shows the decision tree we built from the DT model
- Forêt aléatoire (RF): Finally, RF is a forest of decision trees, where each tree is generated from a new training data set, which is a randomly sampled subset of the original training data set. Predictions are then aggregated to give a final prediction. For this purpose, the most important hyperparameters in this study are the number of trees in the forest (n_estimators = 200), the maximum depth of each decision tree (max_depth = 200), the maximum number of features to be considered for each node split (max_features = 1), the minimum number of samples required to split an internal node (min_samples_split = 10), the minimum number of samples required to be a leaf of the tree (min_samples_leaf = 2) and the random_state parameter used to initialize the random number generator is set to 42

Table 1: Description of data variables

Variables (%)	Description
TAEP_Rur	Number of people using at least basic drinking water services in rural areas
C_PIB	Annual GDP growth
IED_EN_PIB	Foreign direct investment, net inflow (GDP)
RAED_UD	Total annual withdrawals of freshwater for domestic use
RAED_I	Total annual freshwater withdrawals for industry
RAED_T_RI	Annual freshwater withdrawals from total internal resources
CRNB_Hab	Growth in Gross National Income (GNI) per capita
APD_NR_BS	Net Official Development Assistance (ODA) received from imports of goods and services
APD_NR_FBC	Net Official Development Assistance (ODA) received from gross capital formation
APD_NR_RNB	Net Official Development Assistance (ODA) received (GNI)
CPR_PA	Annual growth of the rural population

DOI: 10.3844/ajassp.2023.65.75

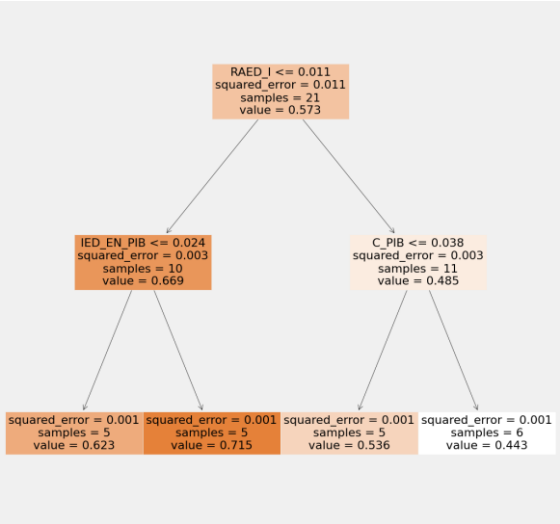

Fig. 1: Decision tree

Model performance was assessed using the Root Mean Square Error (RMSE) metric to measure the mean deviation between predicted and actual values, the Mean Absolute Error (MAE), which refers to the magnitude of the difference between the prediction of observation and the actual value and the R-squared coefficient of determination, used to measure prediction accuracy. It lies between 0 and 1 and increases with the fit of the regression to the model.

Methods

The dependent or target variable of this study is the rate of access to drinking water in rural Senegal (TAEP_Rur), while the other variables are considered explanatory (independent) variables. We used three machine learning methods, namely LR, DT, and RF, to identify any hidden relationships between the independent variables and the TAEP_Rur target variable. To this end, we explored these parameters by highlighting the variables that most significantly affect the TAEP_Rur.

To do this, we began our study with a correlation analysis between TAEP_Rur and the other variables. Next, these three methods are used to find a relationship that links TAEP_Rur to the independent variables. In order to build the model (training dataset), 80% of the data from the randomly selected dataset was used, while the remaining 20% was reserved for evaluating the model's performance (test dataset). We used the training variables to train the model (LR, DT, and RF). Finally, we used the model constructed from the test set variables to predict TAEP_Rur.

Results and Discussion

Correlation Analysis

The characteristics of the data set are crucial to understanding their correlation and influence on the dependent variable. Figure 2 illustrates the correlation between dataset characteristics using Pearson's heat map.

This Pearson correlation Fig. 2 shows that some variables are strongly correlated with TAEP_Rur, while others are negatively correlated. The variables with a strong positive correlation are IED_EN_PIB (0.73), RAED_UD (0.87), and RAED_T_RI (0.86). On the other hand, the variables that have a strong negative correlation with TAEP_Rur are RAED_I (-0.97), APD_NR_BS (-0.82), APD_NR_FBC (-0.70), APD_NR_RNB (-0.51), CPR_PA (-0.60). The variables CRNB_Hab and C_PIB are positively correlated with TAEP_Rur, but the correlation is relatively weak, with correlation coefficients of 0.21 and 0.09 respectively.

It is interesting to note that the RAED_I variable stands out from the other variables with a high negative correlation coefficient (-0.97) with TAEP_Rur. This indicates that this variable has a significant impact on TAEP_Rur Sénégal. An increase in total annual freshwater withdrawals for industry may indicate stronger economic activity, which could be positive for ADW if it leads to investment in drinking water infrastructure. Furthermore, industries can also contribute to improving ADW by investing in drinking water supply projects or implementing more sustainable water management practices. However, the withdrawal of large quantities of freshwater by industry can reduce the amount of water available to local populations and increase pollution of water sources, which can affect water resources and make access to drinking water more difficult. Ultimately, it is important to regulate total annual freshwater withdrawals for industry to minimize their impact on ADW in rural Senegal.

Predictive Analysis

We used ML techniques to compare the performance of different RMSE variables in terms of access to drinking water in rural Senegal. The RMSE determines the error between the predicted and actual values of the access rate in the test set and its expression is given by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_i - P_i)^2} \quad (1)$$

where, A_i is the actual value of TAEP_Rur, P_i is the predicted value of TAEP_Rur and N is the number of TAEP_Rur observations in the test set. Table 2 shows the results of applying the DT technique to calculate the RMSE between the actual and predicted values of TAEP_Rur in the test set.

Using the DT model to generate predicted values of TAEP_Rur, we obtained an RMSE of 0.335 and an MSE of 0.113 when APD_NR_FBC was used in the test set.

Table 3 shows the RMSE and MSE values between the actual and predicted values of TAEP_Rur in the test set, obtained using RF.

Table 2: RMSE and MSE between the actual values of TAEP_Rur in the test and the values defined and predicted when using the DT model

Method	Variables used for the prediction of the TAEP_Rur	RMSE	MSE
DT	C_PIB	0.509	0.260
	IED_EN_PIB	0.542	0.294
	RAED_UD	0.482	0.232
	RAED_I	0.548	0.301
	RAED_T_RI	0.471	0.221
	CRNB_Hab	0.540	0.292
	APD_NR_BS	0.412	0.170
	APD_NR_FBC	0.335	0.113
	APD_NR_RNB	0.504	0.254
	CPR_PA	0.541	0.293

Table 3: RMSE and MSE between the actual values of TAEP_Rur in the test and the values defined and predicted when using the RF model

Method	Variables used for the prediction of the TAEP_Rur	RMSE	MSE
RF	C_PIB	0,529	0.280
	IED_EN_PIB	0,562	0.316
	RAED_UD	0,503	0.253
	RAED_I	0,567	0.321
	RAED_T_RI	0,489	0.239
	CRNB_Hab	0,559	0.313
	APD_NR_BS	0,426	0.182
	APD_NR_FBC	0,347	0.121
	APD_NR_RNB	0,522	0.273
	CPR_PA	0,561	0.314

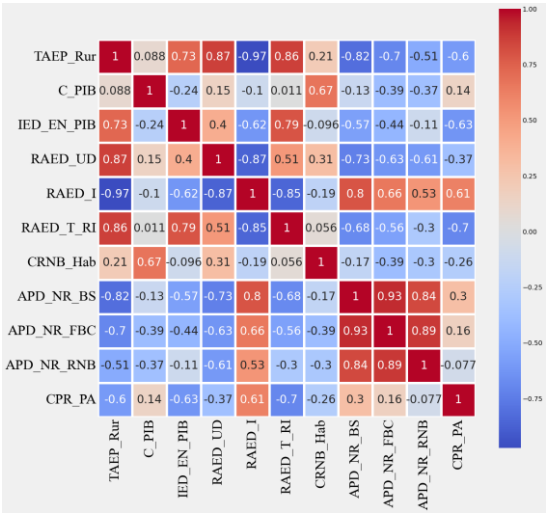

Fig. 2: Correlation between data variables

Using the RF model to generate predicted values of TAEP_Rur, we obtained an RMSE of 0.347 and an MSE of 0.121 when APD_NR_FBC was used in the test set.

Table 4 shows the RMSE and MSE values between the actual and predicted values of TAEP_Rur in the test set, obtained using the LR technique.

Table 4, an RMSE of 0.334 and an MSE of 0.112 are obtained when APD_NR_FBC in the test set is used in combination with the LR model to give the predicted values of TAEP_Rur.

Tables 2-4 show which variable has the greatest influence on the prediction of the target variable (TAEP_Rur). In fact, the variable that gives the lowest RMSE or MSE when predicting the RADW is the one that has the most influence and gives the most accuracy to the prediction.

APD_NR_FBC differs from the other variables in that, for all the algorithms, it has the smallest MSE and RMSE.

We evaluate the performance of the ML models (LR, DT, and RF) on the dataset using three regression metrics: RMSE, MAE, and R-squad.

The RMSE is the square root of the mean squared errors and its expression is given in Eq. (1).

The coefficient of determination, also known as R-squared, is a statistical measure that provides an indication of the model's accuracy in predicting unseen values. The coefficient of determination is calculated using the following equation:

$$r^2 = 1 - \frac{SSE}{SST_0} \quad (2)$$

In the context of regression analysis, the sum of squared errors (*SSE*) is a measure of the residual variation in the data, i.e., the difference between observed and predicted values. The sum total of squares (SST_0) is a measure of the total variation in the data, i.e., the difference between the observed values and the mean of the observed values. Their expressions are given in the following equations:

$$SSE = \sum_{i=1}^{N}(P_j - \bar{A})^2 \quad (3)$$

$$SST_0 = \sum_{i=1}^{N}(P_i - \bar{A})^2 \quad (4)$$

where, P_j is the predicted value, P_i is the actual value, $\bar{A}$ is the mean of the target variable and N is the number of observations in the data set.

DOI: 10.3844/ajassp.2023.65.75

Table 4: RMSE and MSE between the actual values of TAEP_Rur in the test and the values defined and predicted when using the LR model

Method	Variables used for the prediction of the TAEP_Rur	RMSE	MSE
LR	C_PIB	0,509	0.259
	IED_EN_PIB	0,541	0.293
	RAED_UD	0,481	0.231
	RAED_I	0,548	0.300
	RAED_T_RI	0,470	0.221
	CRNB_Hab	0,539	0.291
	APD_NR_BS	0,411	0.169
	APD_NR_FBC	0,334	0.112
	APD_NR_RNB	0,503	0.253
	CPR_PA	0,541	0.292

Table 5: Performance evaluation of the three models

Metric	LR	DT	RF
MAE	0.0110	0.034	0.047
RMSE	0.0015	0.038	0.059
R-squared	96.9%	77.9%	48.9%

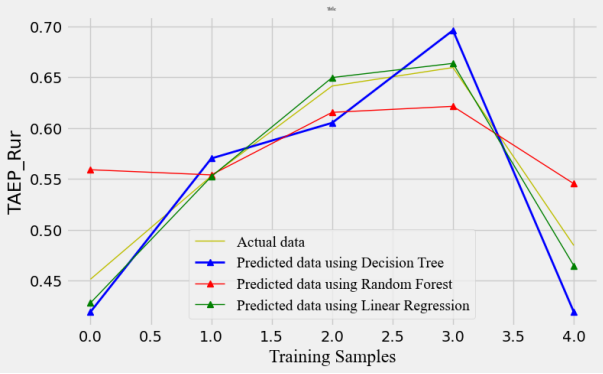

Fig. 3: Performance of predictive data against real data

The *MAE* metric measures the average magnitude of errors on the test set, which contains *N* data points. Its formula is given in the following equation:

$$MAE = \frac{1}{2N}\sum_{i=1}^{N}|A_i - P_i| \quad (5)$$

The *RMSE* and *MAE* scores are negatively oriented, meaning that lower values are better.

RMSE is more appropriate than *MAE* when outliers have a significant impact, as is the case in your study.

We evaluated the performance of each ML algorithm on a test dataset. Figure 3 shows the predicted values of TAEP_Rur versus the actual values in the dataset.

By superimposing the results of the three ML models between the predicted values and the actual values of TAEP_Rur, we show in Fig. 3 that the curve of the values predicted by LR (in green) is the one that most closely resembles that of the actual values (in yellow) of TAEP_Rur.

According to the results presented in Table 5, the LR and DT models are more accurate for prediction because they offer better precision.

According to Table 5, LR offers better prediction accuracy, i.e., R-squared of 96.9%. These results are debatable.

As for linear regression, the MAE is very low, indicating that the model's predictions are on average very close to the actual values. The RMSE is also low, which means that the prediction errors are generally small. The R-squared is high at 96.9%, indicating that the model explains a large proportion of the variance in the data and confirms the accuracy of the model's predictions. In this respect, linear regression is a good choice for modeling the data and predicting the target values.

For decision trees, the MAE is higher compared to linear regression, indicating slightly lower accuracy in predictions. Its RMSE is also higher compared to linear regression, indicating generally larger prediction errors. The R-squared is 77.9%, which means that the model explains less of the variance in the data compared with linear regression.

Given these considerations, decision trees may be less accurate in predicting target values than linear regression.

In the case of Random Forest, the MAE is higher than that of the other two models, indicating lower prediction accuracy. The RMSE is also higher, indicating generally greater prediction errors. The R-squared is lowest at 48.9%, meaning that the model explains a relatively small proportion of the variance in the data.

Under these conditions, random forest is less accurate in predicting target values than linear regression and decision trees.

Ultimately, the algorithms use a more complex approach to identify the most influential variables in predicting the TAEP_Rur. While Pearson's correlation coefficient measures the linear relationship between two variables, it is therefore possible for the results to differ depending on the method used.

In this case, it is possible that total annual withdrawals of freshwater for industry are strongly correlated with the RADW, but that this variable is not as important as net official development assistance received in predicting the RADW. This may be explained by the fact that external financial aid may have a more direct impact on improving

access to drinking water, whereas water withdrawals for industry may be influenced by other factors such as economic growth and industrial demand. This is also due to the fact that few industries are located in rural Senegal. This minimizes their impact on rural water resources.

To analyze the importance of variables in predicting the TAEP_Rur, we used the performance of three prediction models: Linear regression, decision trees, and random forests. Taking into account the results of Tables 2-4, the results showed that net official development assistance received (ODA_NR_FBC) appears to be the most influential variable, meaning that it has a strong influence on the RADW for several reasons.

Official development assistance funds have had an influence on existing water supply policies and programs in Senegal. They have helped to encourage the Senegalese government to adopt policies that are more favorable to access to drinking water and to set up more effective programs to meet people's needs. These are essentially the Drinking water and sanitation program in rural areas (PEPAM, 2016), the Community Emergency Program (PUDC), the Local Water and Sanitation Plan (PLHA), the Rainwater Sanitation and Management Program (PAGEP), the integrated water resources management program (PGIRE), the drinking water access program (PAEP) (MHS, 2017), etc., these initiatives have all proved effective in improving ADW in rural areas since 2000. In addition, the funds allocated by official development assistance have enabled the construction of numerous hydraulic wells in various regions of Senegal, the installation of water treatment systems, the rehabilitation of existing water supply networks, capacity building for local governments and civil society organizations to manage water resources and provide drinking water supply services, as well as raising awareness among local populations of the importance of hygiene and health in relation to drinking water.

In addition, increased public aid can enable recipient countries to put in place infrastructure and programs to improve access. This means that countries that receive more aid may have more resources to invest in rural drinking water supply projects, which can lead to a significant improvement in the rate of access.

Conclusion

In our study, we examined the relationship between the rate of access to drinking water in rural (TAEP_Rur) and variables that could influence its variation. Our main objective was to find a hidden link between development variables and the TAEP_Rur Senegal using machine learning. To achieve this, based on the collected data, we performed correlation analysis and used ML techniques including LR, DT, and RF. The prediction performance results showed that LR has the highest prediction accuracy, i.e., R-squared of 96.9%.

During the predictive analysis, despite the existence of missing values, LR, DT, and RF models succeeded in identifying the hidden relationship between the variables and the TAEP_Rur Senegal. However, analysis of the data using machine learning techniques revealed that the variable APD_NR_FBC had the greatest influence on the TAEP_Rur Senegal. This relationship was not visible in the correlation analysis, as it indicates that the RAED_I variable is likely to be the most influential variable on the rate of access to drinking water. The results of our analysis underline that the net public development aid received is essential to improving the drinking water supply in rural Senegal and must be taken into account in policies and programs aimed at solving this problem.

In future work, we plan to examine in greater detail additional parameters that take into account socio-cultural factors likely to influence access to drinking water in rural areas.

Acknowledgment

We appreciate our editor's support and constructive feedback throughout the publishing process.

Funding Information

The authors funded this study as part of a thesis project.

Author’s Contributions

Anabilaye Moussa Coly: Conducted a detailed study of the scientific literature on artificial intelligence and contributed to the written, submission, and revision of the article.

Alioune Ly: Responsible for collecting data, conceptualizing it, proposing working methods, and providing critical input.

Ndolane Diouf: Responsible for the preprocessing and consistency of the contributions.

Séni Tamba: Interpreted the results to inform decision-makers on taking into account the macroeconomic dimension of development.

Issa Sakho: Carried out the analysis, participated in the interpretation of the results and the proofreading.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

African Water Facility. (2017). African Water Facility Strategy 2017-2025. *Africa, African Development Bank.* https://www.afdb.org/sites/default/files/documents/publications/strategie_de_la_facilite_africaine_de_leau_2017-2025.pdf

ANSD. (2015). National Agency of Statistics and Demography Senegal. *Global Partnership for Sustainable Development Data*. https://www.data4sdgs.org/partner/ansd-national-agency-statistics-and-demography-senegal

Bohbot, R. (2008). *Access to water in the slums of African cities: issues and challenges of universalizing access* (Case of Ouagadougou) (Doctoral dissertation, Laval University). https://library-archives.canada.ca/eng/services/services-libraries/theses/Pages/item.aspx?idNumber=1273289930

Bonface, I. M., & John, B. W., (2019). Analysis of the effect of access to potable water on household well-being among small holder farmers in kakamega county, Kenya. *International Journal of Academic Research and Development*, Vol. *4*, p.138-144. https://www.researchgate.net/publication/336851953_Analysis_of_the_effect_of_access_to_potable_water_on_household_well-being_among_small_holder_farmers_in_kakamega_county_Kenya

Briernat, E. & Lutz, M. (2015). Data Science: Fundamentals and case studies. Machine learning with Python and R. *Pari, Eyrolles*. ISBN: 10-978-2-212-14243-3.

Dark, S. (2019). Machine Learning. The ultimate beginner's guide to understanding machine learning. ISBN-10: 978-1-989543-16-0.

Diagne, A., (2009). Interactions between access to drinking water and the other Millennium Development Goals: An analysis based on data from the Dakar suburbs. https://www.cres-sn.com/wp-content/uploads/2017/10/2009_10.pdf

Diallo, I., Touré, S., & DIBY, K. (2022). Annals of the Faculty of Letters, Arts and Human Sciences. *Review of the University of Moundou*. https://aflash-revue-mdou.org/analyse-geographique-de-loffre-et-la-demande-de-leau-potable-dans-la-sous-prefecture-de-dabakala-cote-divoire/

FAO. (2019). The State of Food and Agriculture 2019: Food and Agriculture Organization of the United Nations. *Moving Forward on Food Loss and Waste Reduction*. http://www.fao.org/3/ca6030en/CA6030EN.pdf

Grus, J. (2017). Data Science par la pratique. *Fundamentals with Python*, Paris, Edition Eyrolle, pp. 319. ISBN-10: 9782212255478.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313-1328. https://doi.org/10.1080/00949655.2012.741599

Sullivan, W. (2017). Machine Learning Beginners Guide Algorithms: Supervised and Unsupervised Learning, Decision Tree and Random Forest Introduction. *Amazon*. ISBN-10: 978-1975632328.

Hao, J., & Ho, T. K. (2019). Machine learning made easy: A review of scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*, *44*(3), 348-361. https://doi.org/10.3102/1076998619832248

MHS. (2017). Sectoral development policy letter 2016-2025. *Ministry of Hydraulics and Sanitation of Senegal Sectoral Development Policy Letter* (2016-2025), pp.38. https://www.pseau.org/outils/biblio/resume.php?d=7343

Mitchell, T. M. (1997). "Machine Learning", McGraw-Hill Science/Engineering/Math, March 1. ISBN-10: 13-0070428077.

Studer, C., Keyhani, A., Sebastian, T., & Murthy, S. K. (1997, October). Study of cogging torque in permanent magnet machines. In *IAS'97. Conference Record of the 1997 IEEE Industry Applications Conference Thirty-Second IAS Annual Meeting* (Vol. 1, pp. 42-49). IEEE. https://doi.org/10.1109/IAS.1997.643006

Ndolane, D., Massa, N., Dialo, D., Kharouna, T., Aboubaker, B. C., & Ibrahima, G. (2021, November). Finding Hidden Links among Variables in a Large-Scale 4G Mobile Traffic Network Dataset Using Machine Learning. In *2021 8th International Conference on Soft Computing and Machine Intelligence (ISCMI)* (pp. 1-8). IEEE. https://doi.org/10.1109/ISCMI53840.2021.9654806

OECD. (2012). OECD Environmental Outlook. *Organization for Economic Cooperation and Development Environmental Perspectives*. https://doi.org/10.1787/9789264122246-en

Lemberger, P. Batty. M., Morel. M. Raffaelli. J. L. (2015). Big Data et machine learning-Manuel du data scientist est également présent dans les rayons. ISBN-10: 978-2-10-072392-8.

Lemberger, P, Batty, M. Morel, M. & Raffaelli J. L. (2019). Big Data and Machine Learning 3rd Ed. *Data Science Concepts and Tools.* ISBN-10: 978-2-10-080342-2.

Payen, G. (2023). Water for everyone! Abandon preconceived ideas, face realities. ISBN-10: 9782200285821.

PEPAM. (2016). Project Manual drinking water in rural areas. https://www.pseau.org/outils/ouvrages/mha_pepam_manuel_des_projets_eau_potable_en_milieu_rural_au_senegal_2016.pdf

Repussard, C. (2009). Public water service in rural areas in Senegal: The example of the rural community of Moudéry. pp. 207-208. https://doi.org/10.4000/anthropodev.335

Strong, C., Kuzma, S., Vionnet, S., & Reig, P. (2020). Achieving abundance: understanding the cost of a sustainable water future. *World Resources Institute: Washington, DC, USA*. https://irp.cdn-website com/241a7575/files/uploaded/achieving-abundance.pdf

Timofeev, R. (2004). Classification and regression trees (CART) theory and applications. *Humboldt University, Berlin, 54*. https://www.academia.edu/13700196/Classification_and_Regression_Trees_CART_Theory_and_Applications

UNICEF. (2017). Progress on Drinking Water, Sanitation and Hygiene. *For every child*. https://www.unicef.org/reports/progress-drinking-water-sanitation-and-hygiene)

UNESCO. (2019). United Nations World Water Development Report 2019: Leaving no one behind. *UNESDOC Digital Library*. ISBN-10: 978-92-3-200168-9.

WSP. (2012). Access to drinking water in developing countries: 18 questions for sustainable services, Paris, France. https://knowledge-uclga.org/acces-a-l-eau-potable-dans-les-pays-en-developpement.html

UN-Water. (2018). Sustainable Development Goal 6: synthesis report 2018 on water and sanitation, Geneva. pp. 195. ISBN-10: 978-92-1-362674-0.

World Bank. (2022). Water Security in Senegal: Executive Summary. Washington, D.C. *World Bank Group*. http://documents.worldbank.org/curated/en/099625203082233202/P17223304e22fe01309dd701d1737f107f8

Appendix

No.	Années	TAEP_Rur (%)	IDH (%)	C_PIB (%)	IED_EN _PIB (%)	RAED_UD (%)	RAED_I (%)	RAED_T _RI (%)	CRNB_Hab (%)	EVN_TA (%)	APD_NR _BS (%)	APD_NR _FBC (%)	APD_NR _RNB (%)	PR_PPT (%)	CPR_PA (%)	TMB_1000 (%)
0	2000	40.24	38.80	3.89	1.36	6.29	3.77	6.16	1.42	56.94	22.25	32.12	7.30	59.68	2.11	11.18
1	2001	41.85	39.40	4.31	0.69	5.20	3.10	7.39	2.10	57.56	21.03	25.45	6.59	59.54	2.14	10.82
2	2002	43.48	39.80	6.87	1.18	4.41	2.61	8.61	-2.53	58.23	19.49	27.49	6.45	59.39	2.17	10.42
3	2003	45.12	40.50	5.59	0.99	4.90	2.42	8.67	3.38	59.12	15.73	21.84	5.28	59.10	1.97	9.93
4	2004	46.77	41.20	4.64	1.37	5.37	2.24	8.74	2.26	60.03	30.61	44.78	10.68	58.69	1.81	9.44
5	2005	48.44	41.90	4.31	1.53	5.84	2.06	8.81	1.76	60.92	17.57	25.93	6.42	58.29	1.84	8.99
6	2006	50.12	42.70	2.33	2.48	6.30	1.88	8.87	-0.04	61.75	20.32	33.15	7.47	57.88	1.90	8.58
7	2007	51.82	43.90	2.83	2.51	6.75	1.70	8.94	0.17	62.54	15.23	25.12	6.22	57.47	1.92	8.18
8	2008	53.54	45.01	3.70	2.70	7.20	1.53	9.00	1.25	63.24	14.44	24.11	6.34	57.05	1.92	7.83
9	2009	55.26	45.90	2.75	2.05	7.64	1.36	9.07	-0.73	63.92	18.06	30.91	6.37	56.64	1.95	7.51
10	2010	57.01	46.80	3.39	1.69	8.08	1.19	9.14	0.75	64.62	16.84	28.92	5.86	56.23	1.98	7.19
11	2011	58.76	48.20	1.33	1.90	8.51	1.02	9.20	-2.00	65.26	15.11	28.19	6.02	55.81	1.98	6.89
12	2012	60.54	49.00	4.00	1.56	8.93	0.86	9.27	1.12	65.46	14.35	24.74	6.20	55.40	1.98	6.78
13	2013	62.32	49.60	2.41	1.65	9.35	0.69	9.34	-0.38	66.07	12.57	21.84	5.35	54.98	1.96	6.50
14	2014	64.13	50.20	6.22	2.04	9.76	0.54	9.40	3.06	66.45	14.01	21.64	5.71	54.56	1.96	6.32
15	2015	65.94	50.50	6.37	2.30	10.16	0.38	9.47	2.78	66.88	12.48	18.93	5.02	54.14	1.94	6.11
16	2016	67.77	50.70	6.37	2.48	10.56	0.22	9.53	3.44	67.50	10.47	15.14	3.94	53.70	1.91	5.84
17	2017	69.62	50.90	7.39	2.80	10.95	0.07	9.59	4.31	67.75	10.78	14.51	4.45	53.26	1.89	5.72
18	2018	71.48	51.20	6.21	3.67	8.84	0.05	11.73	3.61	68.10	9.98	13.24	4.44	52.81	1.86	5.58
19	2019	73.35	51.30	4.61	4.55	8.64	0.05	11.87	1.81	68.53	9.98	NaN	6.31	52.35	1.82	5.41
20	2020	75.24	51.30	1.33	7.54	NaN	NaN	NaN	-1.19	68.01	9.98	NaN	6.74	51.88	1.78	5.58

NB: Although this data was collected at the beginning of January 2023, it should be taken into account when updating the United Nations databas