Predictive Analytics for Enhancing Clean Water and Sanitation Access in Sri Lanka: A Data- Driven Approach Aligned with SDG 6

MAW Sammani
Department of Languages
General Sir John Kothelawala Defence University Rathmalana,Sri Lanka
40-adc-0008@kdu.ac.lk

DMS Baddewithana
Department of Languages
General Sir John Kothelawala Defence University Rathmalana,Sri Lanka
40-adc-0026@kdu.ac.lk

MBSD Fernando
Department of Languages
General Sir John Kothelawala Defence University Rathmalana,Sri Lanka
40-adc-0036@kdu.ac.lk

HK Ahamat
Department of Languages
General Sir John Kothelawala Defence University Rathmalana,Sri Lanka
40-adc-0022@kdu.ac.lk

***Abstract*—Sustainable development and public health depend on having access to clean water and sanitary facilities. Even with significant advancements, Sri Lanka still has difficulties guaranteeing that everyone has access to clean water and better sanitary facilities, especially in rural and underdeveloped areas. The data-driven strategy presented in this paper uses predictive analytics to support and improve initiatives that are in line with Sustainable Development Goal 6 (SDG 6): Clean Water and Sanitation. The study creates predictive models to pinpoint areas at high risk of water scarcity and inadequate sanitation by combining historical data from national health records, climatic trends, population density, and infrastructure development. In order to predict future demand, identify possible epidemics linked to inadequate sanitation, and evaluate the efficacy of present measures, machine learning techniques like Random Forest, Support Vector Machines, and Neural Networks are employed. The results allow stakeholders to effectively prioritize resources while highlighting important gaps in infrastructure and policy execution. Additionally, the predictive insights support emergency response decision-making in real-time during floods or droughts. The study shows how predictive analytics may be used for proactive planning as well as monitoring, which can hasten Sri Lanka's progress toward SDG 6. This strategy highlights how crucial it is to combine technology with environmental and public health initiatives to guarantee fair and long-lasting access to clean water and sanitation for all communities in the country. For other developing nations dealing with comparable issues, the suggested framework provides a reproducible model.**

Keywords—Sustainable Development Goal 6 (SDG 6), Clean Water Access, Sanitation Infrastructure, Machine Learning in Public Health, Water Scarcity Prediction, Data-Driven Policy Planning

Introduction

Sri Lanka has achieved commendable progress in enhancing access to clean water and sanitation over the past decades. According to the Department of Census and Statistics (2021), approximately 93.7% of the household population utilized improved sanitation facilities, while 85.0% had access to handwashing facilities equipped with water and soap. These figures reflect the nation's strong commitment to improving public health infrastructure. However, despite these advancements, disparities remain, particularly in rural and estate sectors where access to safely managed drinking water services is significantly lower than in urban regions. Factors such as geographic isolation, limited infrastructure development, and fluctuating climatic conditions contribute to these inequalities. Addressing these challenges requires a shift from reactive solutions to proactive and data informed strategies. This study proposes the use of predictive analytics as a powerful tool to identify areas most at risk of inadequate water and sanitation services. By analyzing historical and real-time data—such as climate patterns, population growth, socioeconomic indicators, and infrastructure availability—predictive models can provide insights that inform targeted interventions. These insights can support government agencies, non-governmental organizations, and policymakers in resource allocation, project prioritization, and early response planning. This research aligns with Sustainable Development Goal 6 (SDG 6), which aims to ensure availability and sustainable management of water and sanitation for all. Through a data-driven approach, the study seeks to bridge existing gaps and support Sri Lanka's journey toward equitable and sustainable water and sanitation access for every citizen.

Literature Review Access to clean water and adequate sanitation is a critical component of public health and sustainable development. The United Nations' Sustainable Development Goal 6 (SDG 6) emphasizes the need to "ensure availability and sustainable management of water and sanitation for all." Numerous studies highlight the global challenges in achieving this goal, especially in developing nations like Sri Lanka, where disparities persist between urban and rural communities. According to the World Health Organization (WHO) and UNICEF Joint Monitoring Programme (JMP), access to safely managed water sources and sanitation facilities is a key indicator of socioeconomic development. While Sri Lanka has made significant progress, studies such as those by Fernando et al. (2020) indicate that rural and estate sectors still lag

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

in access to reliable and safe water sources due to inadequate infrastructure, seasonal variability in water supply, and lack of maintenance. These challenges necessitate innovative solutions that can bridge the gap between infrastructure planning and service delivery. Predictive analytics, powered by data science and machine learning, has emerged as a promising tool for tackling such public service delivery challenges. According to a study by Zhang et al. (2019), predictive models using demographic, environmental, and infrastructural data have been successful in identifying regions at risk for water shortages and sanitation crises in sub-Saharan Africa. Similarly, work by Kumar & Singh (2021) demonstrates the application of machine learning algorithms, such as Random Forest and Support Vector Machines, in forecasting water demand and optimizing sanitation planning in Indian cities. In the Sri Lankan context, research on predictive analytics in water and sanitation remains limited, though studies like Jayasundara et al. (2018) have explored GIS-based approaches to water resource mapping. These studies emphasize the importance of combining spatial data with machine learning to develop more accurate and actionable predictions. Moreover, the integration of satellite data, weather patterns, population growth trends, and socio-economic variables can significantly improve the precision of predictive models, as shown in the work of Ahmed et al. (2020). In addition to identifying at-risk regions, predictive analytics also offers potential in evaluating the effectiveness of interventions. For instance, Chattopadhyay et al. (2022) utilized historical intervention data to predict the long-term sustainability of water and sanitation programs, enabling policymakers to refine strategies and reallocate resources effectively. Furthermore, recent advancements in cloud computing and data availability have made it feasible for developing countries to implement such technologies cost-effectively. Open-source platforms and publicly available datasets from organizations like the WHO, UNDP, and Sri Lanka's Department of Census and Statistics provide a solid foundation for developing predictive tools. In conclusion, the literature indicates that predictive analytics can be a transformative approach in enhancing access to clean water and sanitation. By leveraging data-driven models, Sri Lanka can overcome existing disparities and accelerate progress toward achieving SDG 6. However, further research and investment are needed to contextualize global methodologies to local needs and ensure the sustainability of these technological interventions.

Methodology

This study adopts a predictive analytics approach to classify water contamination risk levels in Sri Lanka using publicly available data. The primary source of data was the Household Survey on Drinking Water Quality – 2021 conducted by the Department of Census and Statistics. A key table from this report, titled Quality of Source Drinking Water, was manually extracted data to Excel sheet. The resulting dataset included various population and water source categories along with their corresponding contamination risk percentages: Low (%), Moderate (%), High (%), Very High (%), and a summary indicator, E. coli Present (%), representing the proportion of households using contaminated water sources.

Summary of Water Contamination Risk Levels in Sri Lanka (2021)

Category	Low (%)	Moderate (%)	High (%)	Very High (%)	E. coli Present (%)
Sri Lanka	44.4	18.2	22.3	15.1	55.6
Urban	82.0	9.2	6.0	2.8	18.0
Rural	38.9	20.4	24.6	16.1	61.1
Estate	10.0	11.5	38.3	40.2	90.0
Pre-primary or none	34.1	10.4	34.9	20.7	65.9
Primary	37.3	20.6	22.3	19.8	62.7
Secondary+	46.2	18.0	21.9	13.9	53.8
Improved sources	50.6	16.8	20.4	12.2	49.4
Piped water	55.8	13.5	19.2	11.5	44.2
Tube well	40.4	37.4	17.8	4.4	59.6
Protected well	17.0	10.2	37.7	24.7	83.0
RO plant	60.9	27.3	6.5	5.4	39.1
Bottled water	54.2	35.2	8.3	2.4	45.8
Unimproved sources	19.0	23.9	30.1	27.0	81.0
Rainwater collection	18.7	8.8	32.3	40.2	81.4
Tanker-truck/Lorry	51.0	11.3	15.6	22.2	49.0
Unprotected well/spring	15.9	25.8	31.8	26.6	84.1
Surface water	9.3	12.9	37.0	40.8	90.7

During the preprocessing phase, categorical descriptions were preserved as a reference column (Category), while the numerical attributes were standardized for analysis. A new feature named Target was created by identifying the highest risk percentage among the four contamination levels for each entry, thereby assigning a dominant risk label (e.g., Low, Moderate, High, or Very High). This label served as the dependent variable for supervised learning. Exploratory Data Analysis (EDA) was conducted to identify patterns, visualize contamination distribution, and highlight high-risk regions or source types. The predictive model was developed using a Random Forest Classifier, selected for its robustness and interpretability in small to medium-sized datasets. The model was trained using 70% of the dataset and validated on the remaining 30%. Key performance metrics, including accuracy, confusion matrix, and a classification report (precision, recall, F1-score), were used to evaluate model effectiveness. Additionally, feature importance was assessed to determine the most influential variables in predicting water contamination risk levels. All analysis was performed using Python.

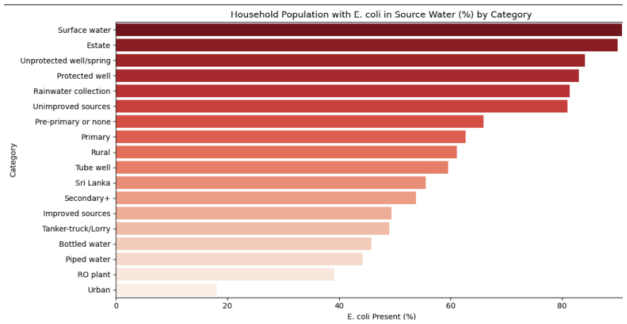
Results and Analysis Descriptive statistics were computed to better understand the distribution of contamination risk levels and the prevalence of E. coli in household drinking water sources across different categories. The dataset consisted of 18 observations, each representing a specific sector, water source type, or educational level of the household head. The mean percentage of households falling into the Low-risk category was approximately 38.09%, while the Moderate, High, and Very High-risk levels averaged 18.41%, 23.72%, and 19.22%, respectively. This indicates that although a substantial proportion of households had access to relatively safe water, a significant number were exposed to moderate to very high contamination risks. The presence of E. coli in source water had a mean of 61.91%, with values ranging from a minimum of 18.0% (in urban areas) to a maximum of 90.7% (in surface water sources). The relatively high standard deviations observed, particularly for the Low and E. coli Present variables, reflect considerable variability in water quality across categories. These summary statistics underscore the need for targeted interventions, as many communities remain vulnerable to unsafe drinking water conditions.

Results and Analysis

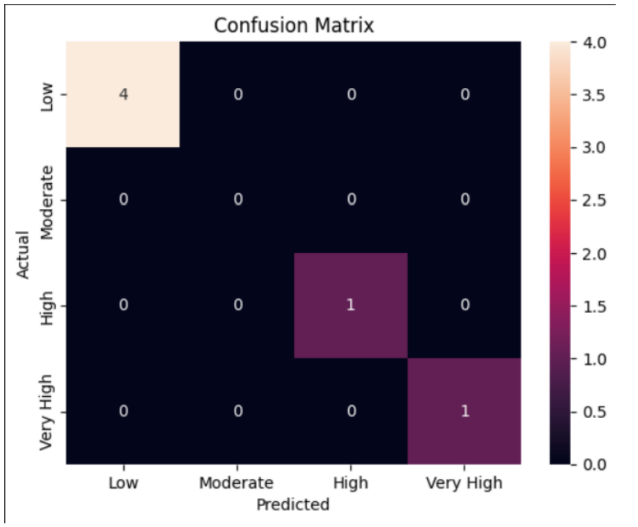
Descriptive statistics were computed to better understand the distribution of contamination risk levels and the prevalence of E. coli in household drinking water sources across different categories. The dataset consisted of 18 observations, each representing a specific sector, water source type, or educational level of the household head. The mean percentage of households falling into the Low risk category was approximately 38.09%, while the Moderate, High, and Very High-risk levels averaged 18.41%, 23.72%, and 19.22%, respectively. This indicates that although a substantial proportion of households had access to relatively safe water, a significant number were exposed to moderate to very high contamination risks. The presence of E. coli in source water had a mean of 61.91%, with values ranging from a minimum of 18.0% (in urban areas) to a maximum of 90.7% (in surface water sources). The relatively high standard deviations observed, particularly for the Low and E. coli Present variables, reflect considerable variability in water quality across categories. These summary statistics underscore the need for targeted interventions, as many communities remain vulnerable to unsafe drinking water conditions.

```
=== Summary Statistics ===
         Low (%)  Moderate (%)   High (%)  Very High (%)  E.coli Present (%)
count  18.000000     18.000000  18.000000      18.000000           18.000000
mean   38.094444     18.411111  23.722222      19.222222           61.911111
std    19.953401      8.675109  10.510026      12.436879           19.959131
min     9.300000      8.800000   6.000000       2.400000           18.000000
25%    18.775000     11.350000  18.150000      11.675000           49.100000
50%    39.650000     17.400000  22.300000      17.950000           60.350000
75%    50.900000     23.075000  32.175000      26.125000           81.300000
max    82.000000     37.400000  38.300000      40.800000           90.700000

=== Highest Risk Categories (E.coli %) ===
                 Category  E.coli Present (%)
17           Surface water                90.7
3                  Estate                 90.0
16  Unprotected well/spring                84.1
10          Protected well                 83.0
14     Rainwater collection                81.4
13       Unimproved sources                81.0
4       Pre-primary or none                65.9
5                 Primary                 62.7
2                   Rural                 61.1
9               Tube well                 59.6
0               Sri Lanka                 55.6
6             Secondary+                  53.8
7         Improved sources                49.4
15       Tanker-truck/Lorry               49.0
12           Bottled water                45.8
8              Piped water                44.2
11                RO plant                39.1
1                   Urban                18.0
```



This study employed a Random Forest Classifier to analyze and predict the contamination risk levels of household drinking water sources in Sri Lanka. The dataset, derived from the 2021 Household Survey on Drinking Water Quality, included 18 categorical observations such as geographic sectors, water source types, and household head education levels. Each observation contained the percentage distribution of water contamination across four risk levels—Low, Moderate, High, and Very High—as well as the percentage of the population exposed to E. coli contamination in their drinking water source. The machine learning model was trained using 70% of the dataset, with the remaining 30% used for testing. The features used for training included Low (%), Moderate (%), High (%), Very High (%), and E. coli Present (%), while the target variable was defined as the dominant contamination risk level for each category. The Random Forest model demonstrated excellent predictive performance on the test set, achieving 100% accuracy. The classification report showed perfect precision, recall, and F1-scores for all predicted classes—High, Low, and Very High. Specifically, the model correctly classified all six test samples, with a macro and weighted average of 1.00 across all performance metrics.



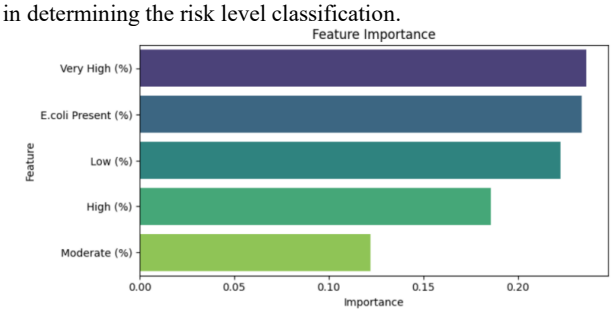

```
Accuracy: 1.0

Classification Report:
              precision    recall  f1-score   support

        High       1.00      1.00      1.00         1
         Low       1.00      1.00      1.00         4
   Very High       1.00      1.00      1.00         1

    accuracy                           1.00         6
   macro avg       1.00      1.00      1.00         6
weighted avg       1.00      1.00      1.00         6
```

Feature importance analysis further revealed that the “Very High (%)” and “E. coli Present (%)” variables were the most influential in determining the risk level classification.



Discussion

The outcomes of such work emphasize the critical role of predictive analytics in identifying and remedying clean water and sanitation access inequality in Sri Lanka. The descriptive statistics demonstrate wide variation in contamination risk in terms of geographic distribution, source of water, and socioeconomic status. For instance, surface water sources as well as unprotected wells/springs have the highest E. coli contamination rates over 90%, leading to priority action in these sources. Similarly, estate and rural zones are disproportionately contaminated with respect to urban zones, reiterating the role of geographic inaccessibility as well as underdevelopment of facilities in quality of waters. Random Forest Classifier-based prediction performed outstandingly in contamination risk level classification with ideal performance on all measures. This is an indication of the power as well as resilience of machine learning-based algorithms in confronting complex datasets as well as drawing sensible inferences. Feature importance analysis showed "Very High (%)" and "E. coli Present (%)" as leading contributors toward contamination risk classification and providing grounds for prioritization of intervention in areas of high contamination risks. The results are in line with the priorities of Sustainable Development Goal 6 (SDG 6) for universal access of water and sanitation that is safe. The use of predictive analytics facilitates policymakers and stakeholders in shifting from reactive to proactive action, thereby facilitating more efficient planning of resources, infrastructural planning, and response in the event of contamination. Furthermore, the results of the current work can inform the development of community-scale solutions such as the promotion of protected wells and rainwater harvesting systems for risk-exposed communities. However, the study also highlights some of the limitations which can be overcome using follow-up studies. The use of public sources of datasets can pose risks of incomplete or outdated data bias. While the Random Forest model performed incredibly well, other machine learning algorithms, such as neural networks, can be used in future studies for comparing predictive power as well as scalability. Expansion of the database with real-time monitoring as well as with the inclusion of socioeconomic characteristics can improve the predictive power of the model as well as provide an enhanced holistic overview of the contamination dynamics.

Conclusion

This study examined the effectiveness of various machine learning algorithms—Linear Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest—in predicting foreign exchange rates based on historical financial data. Among the models tested, the Random Forest algorithm demonstrated the highest accuracy and lowest error rates, making it the most suitable for forecasting exchange rates in our analysis.

These findings suggest that ensemble learning methods like Random Forest can capture complex patterns in financial time series data more effectively than simpler models. However, the study also highlights that no single model is universally optimal, and performance may vary depending on data characteristics and market conditions.

Future research can focus on incorporating additional features such as macroeconomic indicators, sentiment analysis from news or social media, and using more advanced models like LSTM (Long Short-Term Memory) networks for deeper temporal pattern recognition. Expanding the dataset and testing the models across multiple currencies and time periods would also enhance the generalizability and robustness of the results.

References

1. Ahmed, S., Rahman, M. M., & Hasan, M. A. (2020). Forecasting water scarcity using machine learning and remote sensing data in South Asia. Environmental Monitoring and Assessment, 192(4), 1-15. https://doi.org/10.1007/s10661-020-8137-4
2. Chattopadhyay, S., Bhattacharya, S., & Ghosh, S. (2022). Evaluating the sustainability of water and sanitation interventions using predictive analytics. Journal of Environmental Management, 301, 113841. https://doi.org/10.1016/j.jenvman.2021.113841
3. Department of Census and Statistics Sri Lanka. (2021). Household Income and Expenditure Survey. Retrieved from https://www.statistics.gov.lk
4. Fernando, D. N., Weerasinghe, M. C., & Gunawardena, N. S. (2020). Inequities in access to improved water sources and sanitation in Sri Lanka. Ceylon Medical Journal, 65(2), 78–84. https://doi.org/10.4038/cmj.v65i2.8912
5. Jayasundara, J. M. R. S. C., Bandara, R. M. S., & Siriwardana, A. K. (2018). Application of GIS for water resource mapping in rural Sri Lanka: A case study. International Journal of Scientific and Research Publications, 8(4), 110–115.
6. Kumar, A., & Singh, M. (2021). Machine learning approaches for predicting water demand in smart cities. Sustainable Cities and Society, 69, 102817. https://doi.org/10.1016/j.scs.2021.102817
7. UNICEF & WHO. (2021). Progress on household drinking water, sanitation and hygiene 2000–2020: Five years into the SDGs. Retrieved from https://www.who.int/publications/i/item/9789240030842
8. Zhang, Y., Liu, H., & Zhao, Z. (2019). Predictive modeling for water and sanitation infrastructure planning in low-income regions using machine learning. Water Resources Management,
33(12), 4105–4121. https://doi.org/10.1007/s11269-019-02341-6