WATER QUALITY ASSESSMENT AND DRINKABILITY PREDICTION USING MACHINE LEARNING: A COMPARATIVE STUDY OF WHO AND BANGLADESH STANDARDS

EDGE FINAL PROJECT REPORT

SUBMITTED TO

MD MAHBUB E NOOR

LECTURER, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

UNIVERSITY OF BARISHAL

SUBMITTED BY

NAME: PRODIPTO MONDAL

ROLL NO: 09-004-07

BATCH: DS-4

DEPARTMENT OF COASTAL STUDIES & DISASTER MANAGEMENT

UNIVERSITY OF BARISHAL

6TH DECEMBER 2024

TABLE OF CONTENTS

WATER QUALITY ASSESSMENT AND DRINKABILITY PREDICTION USING MACHINE LEARNING: A COMPARATIVE STUDY OF WHO AND BANGLADESH STANDARDS

1. Abstract: Water is an essential component for human survival, yet it frequently fails to reach safe consumption requirements due to contamination and mismanagement. This project includes a thorough investigation into analyzing water quality and forecasting drinkability using machine learning methods. The focus is on comparing the World Health Organization (WHO) and Bangladesh's national potable water standards. This study uses a dataset containing essential water quality parameters such as pH, dissolved oxygen (DO), electrical conductivity (EC), turbidity, calcium (Ca), iron (Fe), and total dissolved solids (TDS) to create predictive models for classifying water samples as drinkable or non-drinkable.

A function was developed to encode drinkability requirements for both WHO and Bangladesh, applying binary labels to water samples. This labeling allowed for exploratory data analysis, which included visualizations like bar charts, pie charts, and correlation heatmaps, to reveal patterns and relationships in the dataset. The dataset was processed using a variety of machine learning algorithms, including Decision Trees, Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines. Cross-validation and hyperparameter tuning ensured a strong model evaluation, with the Decision Tree classifier emerging as a trustworthy predictive model.

The findings show significant disparities in water quality standards between WHO and Bangladesh, highlighting the significance of regional water quality investigations. Gradient Boosting and Random Forest have been demonstrated to be the best accurate models for predicting drinkability. Furthermore, significance of feature analysis revealed the relative contributions of parameters such as pH, DO, and turbidity in affecting water quality. The generated models were highly accurate in predicting drinkability, demonstrating the viability of automating water quality monitoring by machine learning.

This study shows how data-driven approaches may be used to monitor and ensure water safety, while also contributing to public health programs and environmentally friendly resource

management. Future work will concentrate on increasing the dataset, integrating new algorithms, and implementing real-time monitoring systems for more complex applications.

2. Introduction and Motivation:

One of the most fundamental human rights, access to clean, sustainable drinking water is necessary for maintaining life, advancing health, and advancing socioeconomic progress. From 2000 to 2022, 2.1 billion people had access to safely managed drinking water, while 1.2 billion fewer people (or 703 million) did not have access to even the most basic of drinking water services(tkarino, 2023). Approximately two thirds of the world's population endure severe shortages of safe drinking water for at least one month out of the year(Mekonnen & Hoekstra, 2016). By 2025, at least 3 billion people are expected to be living in places where it will be difficult or impossible to meet their basic water needs. This is because the world's water consumption is doubling every 20 years, which is twice the rate of population growth(Worldwide, 2012). Growing urbanization, a changing lifestyle, and population growth are all contributing to the growing scarcity of clean drinking water(Abedin et al., 2014).

The water system is much more vulnerable now because of the growing human population, changing global climate, and increased human activity. Water is crucial for all life on earth. It can improve the health, economy, food production, environment, and social well-being of a community. However, every year, millions of people use unsafe drinking water causing diarrhea, cholera, typhoid, and parasites(Curry, n.d.).

Low quality water consumption is linked to about 80% of diseases and two thirds of deaths in developing nations. Additionally, these diseases take away 10% of an individual's productive time on average(Hoque et al., 2012). Approximately 13% of all urban people worldwide reside in coastal areas, with the majority of these people being Asian(Bank, 2014).

In developing nations like Bangladesh, the effects of this are more detrimental. In particular, the brackish water ecosystem of the southwest coastal region, which is subject to tidal influence and depends on freshwater supplies from upstream. Sea level rise and climate change effects are expected to exacerbate the concentration and infiltration of salt in freshwater river areas in Bangladesh's coastal regions(Dasgupta et al., 2015).

It is projected that until 2050, there will be a notable rise in sea level of up to 52 centimeters, which will have a significant impact on the freshwater zones in eight of the 19 coastal districts due to an increase in river salinity.

With a major effect on the impoverished population, the total number of people affected by such changes in salinization will rise from 2.9 to 5.2 million(Islam, n.d.). . Safe drinking water scarcity from both surface and groundwater sources is a severe issue in coastal areas, primarily brought on by cyclones, coastal flooding, seasonal droughts, ground water depletion and saline intrusion and shrimp farming(Khan et al., 2011). Fisheries, agriculture, health, and the region's ecosystem are all being negatively impacted by the context(Rahman & Islam, 2019).

Access to safe and adequate drinking water is a fundamental human right essential for sustaining life and promoting well-being. Although the locals may lack the knowledge to comprehend the evolving phenomenon of water security, they do possess the experience to overcome such challenges. They have their own adaptation strategies, such as harvesting rainwater and conserving pondwater. Numerous scholars have carried out various investigations concerning Bangladesh's coastal areas.

However, in the coastal area of Assasuni Upazila, under Satkhira district of Bangladesh, the availability of potable drinking water is under severe threat, presenting a pressing challenge to the health and livelihoods of the local population. This project embarks on a critical exploration of potable water scarcity, coupled with an in-depth assessment of physicochemical parameters and trace metals in the coastal region of Satkhira through machine learning using the dataset from laboratory experiment of water samples.

The project is motivated by the crucial need to close the gap between water quality monitoring and real-time decision-making. Machine learning (ML) offers an innovative solution to this problem. By examining patterns and correlations in water quality indicators, machine learning algorithms may predict drinkability based on established standards, considerably increasing the efficiency and accuracy of water quality assessments.

Machine learning (ML) approaches offer a promising approach toward such issues. By working with water quality datasets, it is possible to apply machine learning algorithms that can infer drinkability based on specified parameters. This is particularly important in countries such as

Bangladesh where restrictions on resources create the need for low-cost, easily implemented monitoring systems

3. Objectives: The goals that this project seeks to achieve are:

3.1 Assessment of Water Quality:

- To evaluate the drinkability of water samples based on the standards of World Health Organization (WHO) & national standard of Bangladesh.
- To enumerate the water quality parameters that affect the drinkability; these may include pH, Electrical Conductivity (EC), Turbidity, Dissolved Oxygen (DO), Calcium (Ca), Iron (Fe), and Total Dissolved Solids (TDS).

3.2 Developed Prediction Models:

- To use water quality parameters to build machine learning models that can predict the water drinkability.
- To investigate and compare the classifiers including Random Forest, Decision Trees, Support Vector, Logistic Regression in terms of their effectiveness.

3.3 Analysis of Confusion Matrix:

- Using confusion matrix that compare to assess differences between Bangladeshi and WHO standards.
- To provide insights into deviations or overlaps in classifications based on the two standards.

3.4 Application of Hyperparameter Tuning

- To optimize machine learning model using techniques like GridSearchCV for enhancing prediction accuracy and reliability.

3.5 Exploration of Future Research Directions:

- To describe plans for expanding the research, including investigating advanced modeling methods, adding more datasets and enhancing real-time water quality monitoring systems.

4. Methodology:

4.1 Data Collections: The project utilizes a dataset that includes various water quality parameters, including pH, dissolved oxygen (DO), electrical conductivity (EC), turbidity, calcium (Ca), iron (Fe), and total dissolved solids (TDS).

4.2 Data Processing:

- **Dealing with Encoding:** The data set is loaded in correct encoding (latin1 or ISO-8859-1) so no issue related to character format will occur.
- **Feature Engineering:** Retaining only the water quality features which are useful for analysis. These include:
  - pH
  - DO (mg/L)
  - EC (µs/cm)
  - Turbidity (NTU)
  - Ca (mg/L)
  - Fe (mg/L)
  - TDS (mg/L)
- **Defining Drinkability:** The WHO States that the first binary drinkability columns are defined as WHO Drinkable, and the second by Bangladesh water quality standards as Bangladesh Drinkable. These columns prove whether the water is potable or not, where 1 = “Drinkable” and 0 = “Not -Drinkable”.

4.3 Exploratory data analysis (EDA):

- **Structural Summaries:** Describe basic statistics of each feature to see a scatter plot or distribution.
- **Visualization:**
  - Pie Charts: Distribution of drinkables with WHO and Bangladesh standards.
  - Bar Graphs: For visualizing the distribution of each feature across samples.
  - Heatmap — It shows correlated features in the dataset to determine how dependent they are from each other.
  - Boxplots - A Boxplot Showing outliners and overall how much each water quality parameter spread.

4.4 Model Building:

- **Feature and Target Selection:**
  - **Features(x):** Represent available independent variables such as pH, DO, EC, Turbidity, Ca, Fe
  - **Target Variables (y):** Drinkability for WHO guidelines and Bangladesh standards
- **Classification using Machine Learning Models:**
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Support Vector Machine (SVM)
  - K-Nearest Neighbors (KNN)
  - Gradient Boosting Classifier
  - Naive Bayes

4.5 Model Evaluation:

- **Cross-Validation:**

  To this end, a 5-fold cross-validation is also carried out to guarantee a robust assessment of model ability.

- **Metrics:**

  For each model, accuracy, precision, recall and F1-score are calculated.

- **Confusion Matrix:**

  The confusion matrix of classification results and the performance comparison between WHO and Bangladesh standard was used to analyze.

4.6 Hyperparameter Tuning

- **GridSearchCV:**
  - All models are also hyperparameter-tuned to get optimal results.
  - For each of these algorithms, parameter grids are defined to perform a hyperparameter tuning in order to find out the best combination of hyperparameters.

4.7 Decision Tree Visualization:

- **Tree Structure:**
  - We also use Decision Trees to visualize the rules of WHO and Bangladesh standards.
  - But what they do here is analyzing the feature importance scores and find out which factors have affected their predicting power.

4.8 Prediction and Analysis:

- **Prediction on All Samples:**

  After completing training, all models predict the drinkability of the samples in the data set.

- **Standards Comparison:**

  A strict comparison of the prediction between WHO and Bangladesh standards is also made to ascertain where the discrepancies lie.

4.9 Tools and libraries used:

- **Libraries:**

Utilizes Python libraries such as pandas, numpy, seaborn, matplotlib, sklearn, and scipy for data handling, visualization, as well as modeling.

- **Tools:**

  Jupyter Notebook or similar development environments for coding and documentation.

5. Results and Discussions

5.1 Analysis of Datasets:

Water quality parameters such as pH, Dissolved Oxygen, Electrical Conductivity, Turbidity, Calcium, Iron, and Total Dissolved Solids were analyzed from the dataset. Two standards of drinkable water were applied:

WHO standard: An international benchmark for safety of drinking water.

Bangladesh standard: A local standard, which considers environmental issues and a set of social conditioning.

WHO Drinkability:

The percentage of drinkable water samples was: 14.3%

The percentage of non-drinkable water samples was: 85.7%

Bangladesh Drinkability:

The percentage of drinkable water samples was: 7.1%

The percentage of non-drinkable water samples: 92.9%

The pie charts form visual evidence, which was compared with the drinkability standing of the two standards, with one major variation noticed: the stricter limits as contained in the WHO guidelines.

5.2 Correlation Analysis:

The heat map indicated tight correlations among parameters, suggesting the existence of dependencies between them:

TDS and Electrical Conductivity: show highly positive correlation implying that they depend on each other

pH and Turbidity: weak correlation, showing complete independence.

This piece of information aids in identifying the parameters that will feature prominently in the predictive models.

5.3 Model Performance:

The performance of a set of seven machine learning classifiers for predicting the drinkability of water under both WHO and Bangladesh standards included:

Logistic Regression

Decision Tree

Random Forest

Support Vector Machine (SVM)

K-Nearest Neighbors (KNN)

Naive Bayes

Gradient Boosting.

Key Metrics (Cross-Validation Accuracy):

WHO Standard:

Decision Tree: 0.8000 ± 0.2449

Random Forest: 0.900± 0.2000

KNN (K-Nearest Neighbors): 0.900± 0.2000

Naive Bayes: 0.900± 0.2000

Bangladesh Standard:

Decision Tree: 0.8000 ± 0.2449

Random Forest: 0.900± 0.2000

KNN (K-Nearest Neighbors): 0.900± 0.2000

Naive Bayes: 0.900± 0.2000

6. Conclusion: This study indicates that it is very important to assess quality from both international (WHO) and local (Bangladesh) perspectives, noting differences that tend to arise when applying universal benchmarks in specific regional contexts. While WHO standards globally are recognized to be stringent by design, aimed at long-term health protection, those of Bangladesh provide a framework more relevant to the local context, taking into account the socio-economic and environmental factors a given area is subjected to. It was found that the percentage of samples classified as drinkable under Bangladesh standards is higher than under WHO standards; this exemplifies the delicate trade-off between practicality and global safety benchmarks.

The implementation of machine learning models, such as Gradient Boosting and Random Forest, was found to show great promise for the automation and enhancement of water quality classification. These models offer insights into the relevant parameters that affect water safety, such as pH, Electrical Conductivity, and Total Dissolved Solids, while achieving reliable predictive accuracies. The pie charts, boxplots, and correlation heatmaps worked to provide additional views of the data and aided in identifying trends and outliers that would need to be considered as other intervention strategies are designed.

In conclusion, an attempt to establish advanced data-based interventions and water quality standards offers a compelling framework for determining how public health can best be promoted. Aligning global guidance with local realities, augmenting that par with technological innovations, allows a range of stakeholders to guarantee access to water resources that is safe and sustainable, hence addressing an underpinning pillar for public health and environmental sustainability.

7.References

Abedin, Md. A., Habiba, U., & Shaw, R. (2014). Community Perception and Adaptation to Safe Drinking Water Scarcity: Salinity, Arsenic, and Drought Risks in Coastal Bangladesh. *International Journal of Disaster Risk Science*, *5*(2), 110–124. https://doi.org/10.1007/s13753-014-0021-6

Bank, A. D. (2014). *ADB Annual Report 2013*. https://www.adb.org/documents/adb-annual-report-2013

Curry, E. (n.d.). *Water Scarcity and the Recognition of the Human Right to Safe Freshwater*.

Dasgupta, S., Hossain, Md. M., Huq, M., & Wheeler, D. (2015). Climate change and soil salinity: The case of coastal Bangladesh. *Ambio*, *44*(8), 815–826. https://doi.org/10.1007/s13280-015-0681-5

Hoque, M., Ahmed, S., Alam, M., Purkayastha, M., Belal, A., & Anwar, M. (2012). Physicochemical and microbial water quality of Sylhet city corporation, Bangladesh. *International Journal of Natural Sciences*, *2*. https://doi.org/10.3329/ijns.v2i1.10881

Islam, M. (n.d.). *Bangladesh Delta Plan 2100 Formulation Project*.

Khan, A. E., Xun, W. W., Ahsan, H., & Vineis, P. (2011). Climate Change, Sea-Level Rise, & Health Impacts in Bangladesh. *Environment: Science and Policy for Sustainable Development*, *53*(5), 18–33. https://doi.org/10.1080/00139157.2011.604008

Mekonnen, M., & Hoekstra, A. (2016). Four billion people facing severe water scarcity. *Science Advances*, *2*, e1500323–e1500323. https://doi.org/10.1126/sciadv.1500323

Rahman, M., & Islam, M. (2019). Scarcity of Safe Drinking Water in the South-West Coastal Bangladesh. *Journal of Environmental Science and Natural Resources*, *11*(1–2), 17–25. https://doi.org/10.3329/jesnr.v11i1-2.43361

tkarino. (2023, July 5). *Progress on household drinking water, sanitation and hygiene 2000-2022: Special focus on gender*. UNICEF DATA. https://data.unicef.org/resources/jmp-report-2023/

World Resources Institute | Making Big Ideas Happen. (n.d.). Retrieved January 4, 2024, from https://www.wri.org/

Worldwide, C. (2012, September 30). *Risk Analysis Guidelines 2012*. Concern Worldwide. https://www.concern.net/knowledge-hub/risk-analysis-guidelines-2012