Water Quality Classification for Human Consumption Using Machine Learning/AI

Ariane Shimirwa
ashimirw@andrew.cmu.edu
Carnegie Mellon University

Abstract **-** This study aimed to develop a model for classifying water quality for human consumption using machine learning and AI techniques. The dataset, obtained from the New York Open Data Website, comprised 122,481 water quality samples. Data preprocessing and exploration were performed, including handling missing values, data formatting, and visualizations. The World Health Organization guidelines were used to create labels for drinkable and non-drinkable water. Both supervised and unsupervised learning algorithms were employed, with the Random Forest model achieving the highest accuracy of 0.9944. Unsupervised learning using K-Means clustering yielded two distinct clusters. The study recommends further research on optimizing the models and implementing them for real-time water quality monitoring.

BACKGROUND AND PROBLEM STATEMENT

Access to safe drinking water is a fundamental human right, but contaminated water remains a significant problem in many parts of the world, particularly in Africa [1]. Drinking contaminated water has resulted in severe health problems such as typhoid, cholera, and dysentery, leading to high morbidity and mortality rates [2]. In addition to impacting human health, substandard water quality restricts economic progress by a third, leading to significant economic and social repercussions such as decreased productivity and elevated healthcare expenses [3-4].

One example is the high fluoride content found in groundwater in the Kenyan Rift Valley, Nakuru County, which has led to dental and skeletal fluorosis [5]. This issue highlights the urgent need for effective solutions to improve water quality and ensure access to safe drinking water in affected areas.

While efforts have been made to address this problem, there is still a significant gap in understanding the root causes of contaminated water and the most effective strategies for addressing the issue. In this study, we propose a machine learning technique that can be used to classify water as safe to drink or not.

METHODS

The methods employed in this study can be divided into four main stages: data preparation, data pre-processing, exploratory analysis, and model development.

DATA PREPARATION

The water quality dataset was obtained from the New York Open Data Website and loaded into the programming environment. It represents water quality parameters such as turbidity, coliform, fluoride, and residual chlorine found at various sites in NYC's water distribution system. Turbidity measures the number of suspended particles in water, coliform indicates the presence of potentially disease-causing microorganisms, fluoride strengthens tooth enamel, and chlorine kills waterborne microorganisms [6].

The dataset has 122,481 rows and 10 columns, which includes sample identifiers and water quality parameters.

DATA PRE-PROCESSING

Several pre-processing steps were performed, including:
a) Dropping columns with a high number of missing values (e.g., fluoride) and filling in null values with 0.
b) Data formatting, which involved converting values such as '<1' for the Coliform feature to 0 and adjusting values for other features based on the metadata. All features were converted to floating-point values.
c) Renaming data frame headers for easier visualization.
d) Handling outliers by doing imputation where they were replaced with the median of the column.

e) Creating class labels for drinkable (1) and non-drinkable (0) water using the World Health Organization guidelines [7].

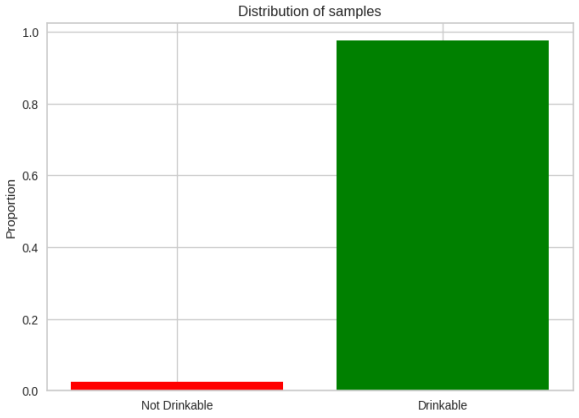
These classifications are determined by specific threshold values for residual chlorine (1.0mg/l), turbidity (5 NTU), coliform (0 MPN/100ml), and E. coli (0 MPN/100ml) levels. Any measurements surpassing these thresholds in each variable would indicate a risk of contaminated water. To qualify as clean water, the values for each variable must be less than or equal to the established thresholds.

EXPLORATORY ANALYSIS

Multiple analyses were conducted to comprehend the available data, investigating the correlation between various water quality factors. The correlation analysis revealed that there was no association between Chlorine and Coliform, but there was a slight positive correlation between Chlorine and turbidity.

The analysis was extended to include an evaluation of the dependent variable in the dataset, which indicates whether the water is drinkable or not. Fig 1. shows that most of the water samples were deemed safe for consumption, but a small percentage of the samples were found to be contaminated.

Fig 1. Proportion of the drinking water



MODEL DEVELOPMENT

Supervised and unsupervised machine learning algorithms were used to classify water quality.

a) Supervised Machine Learning Models

Performing supervised learning, Logistic Regression, k-nearest neighbors, and Random Forest algorithms were chosen. The dataset was split into training (70%) and testing (30%) sets. To train the models, the dataset was first scaled to make data points generalized so that the distance between them would be lower. The models were trained using the training set and evaluated using the test set. Model performance was assessed based on the performance metrics; accuracy, mean squared error, and F1 score.

To improve the performance of the models, feature selection using forward regression was used. Feature selection reduce data dimension using the stepwise forward regression method which selects significant features in model construction based on their p-values. These selected features were then used to re-train the model and then their performance was evaluated.

b) Unsupervised Machine Learning Models

Performing unsupervised learning, K-Means clustering was applied to the unlabeled data to investigate whether it could be classified into two main clusters representing drinkable and non-drinkable water. Distortion metrics and Calinski Harabasz scores were used to determine the optimal number of clusters.

RESULTS AND DISCUSSION

a) Supervised Machine Learning Models

The performance of the supervised machine learning models, including KNN, Logistic Regression, and Random Forest, was evaluated based on accuracy, mean squared error, and F1 score. The results are summarized in the tables below:

Table 1: Model Performance Before Feature Selection

Model	**Accuracy**	**Mean Squared Error**	**F1 Score**
Logistic Regression	0.9825	0.0174	0.976
K-Nearest Neighbour	0.9848	0.1510	0.986
Random Forest	0.9943	0.0056	0.994

Table 2: Model Performance After Feature Selection

Model	Accuracy	Mean Squared Error	F1 Score
Logistic Regression	0.9803	0.0196	0.972
K-Nearest Neighbour	0.9744	0.0255	0.978
Random Forest	0.9944	0.0054	0.994

The Random Forest model demonstrated the best performance with an accuracy of 0.9944, followed by the K-Nearest Neighbor model with an accuracy of 0.9744, and the Logistic Regression model with an accuracy of 0.9803.

After feature selection, the performance of the Random Forest model improved slightly (0.001%), while the other two models showed a slight decrease in performance.

b) Unsupervised Machine Learning Models

K-Means clustering was applied to the unlabelled data, and the optimal number of clusters was determined using distortion metrics and Calinski Harabasz scores as shown on Figure 2. below.

The results suggested that a k of 2 would be most appropriate, despite the higher silhouette score for k =3.

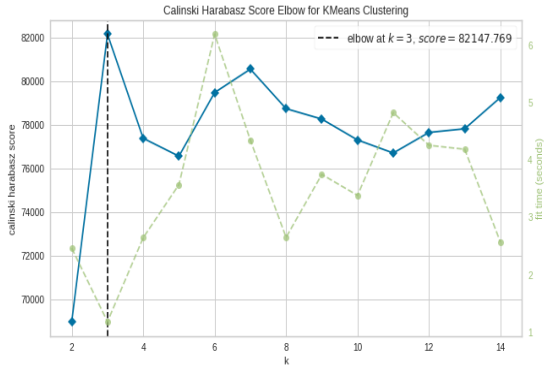


Fig 2. Calinski_harabasz Score Elbow for K-Mean Clustering.

The results of the machine learning models demonstrate the potential of using AI and machine learning techniques to classify water quality for human consumption.

The supervised models, particularly the Random Forest model, achieved high accuracy in predicting water quality, while the unsupervised K-Means clustering algorithm was able to classify the data into two distinct clusters representing drinkable and non-drinkable water.

These findings suggest that machine learning models can be an effective tool for water quality classification and monitoring, supporting efforts to ensure access to safe drinking water

REFLECTION & RECOMMENDATIONS

In this study, we encountered the challenge of working with unlabelled data, which did not deter us from creating supervised learning models. We sought guidelines from the World Health Organization to help set the threshold values for water quality parameters, allowing us to create our own labels for drinkable and non-drinkable water.

Despite the limitations of the dataset, we also explored the performance of unsupervised models, such as K-Means clustering, to understand how the data would perform without labels. One recommendation for future studies would be to add more features to the analysis, as the limited number of features in our dataset might not provide a comprehensive and efficient representation of water quality.

Another recommendation would be to explore the possibility of deploying these machine learning models in real-world scenarios, such as integrating them into sensors within water pipes to provide real-time analytics. This would enable more effective water quality monitoring and improve the efficiency of water treatment processes, ensuring access to safe drinking water in affected areas.

CONCLUSION

Drinking contaminated water has had a devastating impact on public health in Africa, leading to diseases such as typhoid, cholera, and dysentery [1]. As a result, our aim in this project was to enhance existing water

purification and treatment processes by utilizing ML/AI solutions to identify water quality patterns and trends.

To achieve this, we initially created time series plots to uncover correlations and seasonality within the water quality datasets we gathered. Furthermore, our solution relied on water quality measurements, such as turbidity, fluoride, and coliform, obtained from sensors. These measurements provided valuable insights into water quality, ultimately contributing to efforts aimed at ensuring access to safe drinking water.

Our findings demonstrated the potential of machine learning models, particularly the Random Forest model, to accurately predict water quality. The supervised models achieved high accuracy, while the unsupervised K-Means clustering algorithm was able to classify the data into two distinct clusters representing drinkable and non-drinkable water. These conclusions highlight the effectiveness of ML/AI techniques in water quality classification and monitoring and emphasize their potential role in supporting and improving water treatment processes to provide safe drinking water for affected population.

REFERENCES

[1] L. Holtz and C. Golubski, "Addressing Africa's extreme water insecurity," *Brookings*, 9-Mar-2022.

[2] D. Wesle y,"Water Pollution in Africa",Research schools, Degrees & Careers, *Study.com* | 2022.

[3] World Bank Group, "Worsening water quality reducing economic growth by a third in some countries," *World Bank*, 20-Aug-2019.

[4] S. Kusangaya, M. L. Warburton, E. Archer van Garderen, and G. P. W. Jewitt, "Impacts of climate change on water resources in Southern Africa: A Review," *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 67-69, pp. 47–54, 2014.

[5] "High fluoride and dental fluorosis prevalence: A case study from Nakuru area, *The Kenyan Rift Valley*, *NASA/ADS*, 2018.

[6] N. Y. C. O. D: City of New York, "NYC open data," NYC Open Data WP Engine,2022.

[7] "Guidelines for drinking-water quality, 4th edition, incorporating the 1st addendum," *World Health Organization*, 2017.