Project Proposal: Environmental & Water Quality Prediction

Team Members:

1. Joseann Boneo
2. Sean Esla
3. Zizwe Mtonga
4. Lademi Aromolaran

1. Topic Overview and Motivation

Access to clean water is a critical factor for public health, environmental sustainability, and social equity. In many U.S. communities—especially those affected by environmental injustices—water quality fluctuates due to industrial runoff, agricultural waste, and aging infrastructure. Monitoring these variations can help policymakers and residents make informed decisions about local water safety and resource management.

Our project, Environmental & Water Quality Prediction, aims to leverage machine learning to assess and forecast water quality indicators using publicly available datasets. The proposed system will allow users to input their ZIP code and receive two key outputs:

1. A Water Quality Index (WQI) score that summarizes the current state of water quality.
2. A trend prediction graph that shows seasonal or annual changes in WQI over time.

This project integrates data science and environmental analytics to promote awareness of water safety and enable data-driven action in communities.

2. Research Question

How can machine learning techniques be used to predict future water quality levels across U.S. regions based on environmental, chemical, and temporal factors?

Sub-question: Which factors (e.g., pH, dissolved oxygen, turbidity) most strongly influence changes in the Water Quality Index over time?

3. Data Sources

We plan to use the following public datasets:

- U.S. Geological Survey (USGS) National Water Information System (NWIS):
  USGS Water Quality Data for the Nation
  This dataset provides long-term water quality data across U.S. counties, including parameters

such as nitrate concentration, temperature, pH, and dissolved oxygen.

- Kaggle Water Quality Dataset:
  Water Quality Dataset (Ozgur Dogan, 2023)
  Contains global water quality indicators used for classification and regression modeling.

These datasets provide both temporal (time-series) and geographical features suitable for training and testing predictive models.

4. Methodology

Data Preparation:

- Clean and standardize measurements (e.g., normalize pH, conductivity, and turbidity values).
- Handle missing data through imputation or interpolation methods.
- Aggregate data by ZIP code and year/season for localized predictions.

Modeling Approach:

We will experiment with the following algorithms:

- Linear Regression / Random Forest Regression: for trend prediction of WQI over time.
- Logistic Regression or Decision Trees: for classification of water quality as “safe” or “unsafe.”

The final model will be evaluated using metrics such as R^2, Mean Squared Error (MSE), and Precision/Recall (for classification tasks).

Visualization & Interface:

A simple user interface with Streamlit will display:

- A map or form where users enter their ZIP code.
- The predicted Water Quality Index.
- A time-series graph of historical and forecasted WQI trends.

5. Potential Bias and Limitations

- Geographic Bias: Certain regions may have more comprehensive water data than others, leading to uneven model performance.

- Temporal Bias: Seasonal variations or missing historical data may distort long-term predictions.
- Measurement Bias: Differences in testing methods or equipment across counties could affect the accuracy of certain parameters.

We will address these biases through normalization techniques, balanced sampling, and cross-validation.

6. References / Citations

1. CDC. “Water Quality and Your Health.” *Drinking Water*, 2024. www.cdc.gov/drinking-water/about/water-quality-and-your-health.html.

2. Federal Judicial Center. “What Is Water Quality?” *FJC.gov*, 2021. www.fjc.gov/content/376657/water-and-law-what-water-quality.

3. Lin, Li, et al. “Effects of Water Pollution on Human Health and Disease Heterogeneity: A Review.” *Frontiers in Environmental Science*, vol. 10, June 2022. https://doi.org/10.3389/fenvs.2022.880246.

4. Warren-Vega, Walter M., et al. “A Current Review of Water Pollutants in the American Continent: Trends and Perspectives in Detection, Health Risks, and Treatment Technologies.” *IJERPH*, vol. 20, no. 5, 2023. https://doi.org/10.3390/ijerph20054499.

Draft:

Environmental & Water Quality Prediction (Recommended)

- Overview: Build a model to analyze water quality data from public/government datasets (e.g., EPA or county-level data).
- Goal: Create a simple interface where a user inputs their zip code and receives:
- A Water Quality Index (WQI) for their area.
- A trend prediction showing seasonal or annual changes in water quality.
- Learning Focus:
- Data cleaning and statistical modeling.
- Predictive modeling using regression or simple ML techniques.
- Visualization of trends over time.
- Why it's feasible:
- Real data is publicly available.
- Clear societal relevance (environmental justice and public health).
- Manageable complexity for the timeline.

Dataset:
https://catalog.data.gov/dataset/usgs-water-quality-data-for-the-nation-national-water-information-system-nwis
https://www.kaggle.com/datasets/ozgurdogan646/water-quality-dataset

Citations:

CDC. "Water Quality and Your Health." *Drinking Water*, 2024, www.cdc.gov/drinking-water/about/water-quality-and-your-health.html.

Federal Judicial Center. "What Is Water Quality? | Federal Judicial Center." *Fjc.gov*, 2021, www.fjc.gov/content/376657/water-and-law-what-water-quality.

Lin, Li, et al. "Effects of Water Pollution on Human Health and Disease Heterogeneity: A Review." *Frontiers in Environmental Science*, vol. 10, 30 June 2022, www.frontiersin.org/journals/environmental-science/articles/10.3389/fenvs.2022.880246/full.

Warren-Vega, Walter M, et al. “A Current Review of Water Pollutants in American Continent: Trends and Perspectives in Detection, Health Risks, and Treatment Technologies.” *International Journal of Environmental Research and Public Health*, vol. 20, no. 5, 3 Mar. 2023, pp. 4499–4499, www.ncbi.nlm.nih.gov/pmc/articles/PMC10001968/, https://doi.org/10.3390/ijerph20054499.