DE300 Final Project Proposal

Group Name: KTV

Members: Kathy Cui, Viktoriia Sokolenko, Tanay Acharya

Introduction

Artificial intelligence is making greater steps in growing its presence in all facets of our lives. But as its presence grows larger, so do the concerns around its impact. The resource consumption and energy usage that powers AI are unfathomably immense and are heavily contributing to environmental issues in our current world. Many are arguing for the complete cessation of AI usage to mitigate these impacts. However, with the role that AI plays in our technological industry and personal lives, halting AI usage is not a realistic solution at this time.

This doesn't mean we should sit back and watch the environmental damage continue to unfold. Instead, we want to flip this perspective and consider, "How can AI positively contribute to environmental sustainability?" Hence, for this project, we three decided to tackle EY's AI and Data Challenge: "How can AI and data make safe water quality a universal reality?"

EY's AI and Data Challenge presents a question that directly aligns with the topic and issue we wanted to explore. The issue's significance lies in its immense scope and impact. All humans need clean water to survive. The World Health Organization states that 1 in 4 people globally still lack access to safe drinking water, leading to higher risks of disease and deeper social exclusion. We have to accelerate our action towards providing this basic human right to all people. Together, we want to contribute to finding a solution for "one of humanity's most pressing needs: access to clean, safe water" (WHO, 2025).

Our project intends to address this issue by taking a decision-support system approach. We aim to provide early warning risk signals by analyzing trends over time from existing water quality report data provided by the official US Water Quality Portal. By first understanding the point at which water is deemed unsafe in different locations, we can observe the preceding behavior and identify warning signs to aid in making informed decisions in the future. We're offering stakeholders like the local government, public health officials, and water utility companies the opportunity to ascertain where issues might occur and take immediate steps for prevention, whether that be deciding where to prioritize inspections or testing where infrastructure may be failing. Our hope is for AI and data to assist in preventative measures when it comes to water quality, rather than reactive action after damage has already been done.

Data Source

Our data is centered around the Water Quality Portal data from the National Water Quality Monitoring Council. The WQP is the premier source of discrete water-quality data in the United States, with precise, detailed measurements relating to water quality for bodies of water all around the nation. It's a comprehensive repository that contains data from the United States Geological Survey (USGS), the Environmental Protection Agency (EPA), and over 400 state, federal, tribal, and local agencies. The portal supports over 430 million water quality records, including but not limited to physical, chemical, and biological parameters of water quality across the nation.

The WQP offers a data retrieval service where a user can specify location, time frame, parameter types, and more. This data retrieval service allows a user to fill out a form and download the target data as CSV. Alternatively, the WQP also provides easy access to data through its web services. Data can be accessed programmatically through the offered APIs, without needing to interact with the form interface as mentioned above. The WQP offers instructions on how to specify queries as well as location, site, and sampling parameters to retrieve the target data. Regardless of the access method, the WQP provides access to two databases: NWIS (daily-updated data from USGS for 1.5 million sites) and WQX (weekly-updated data from EPA).

Due to the massive amount of data the WQP offers, we have a lot of flexibility in terms of how much data we want to use. Because this project focuses on determining trends over time, we will definitely specify a certain time frame (as using all of the data the WQP has to offer is not feasible). Factors such as location and water quality parameters (pH, temperature, chlorophyll concentration, etc) will all be taken into consideration as well. We do not have a set amount of data in mind, but intend to mess around with the data parameters and determine how much data we need in order to be confident about trends and predictions.

Goal Definitions

Our main goal is to identify trends in the degradation of water quality over time for different locations. Following that goal, we want to be able to use that analysis and determine what factors contribute the most to those trends. By recognizing the patterns and understanding the underlying reasons, stakeholders will be able to make decisions based on the provided data and analysis.

Specifically, we plan to combine data from various locations, clean and preprocess it, and use missing data imputation to impute the variables related to water quality parameters. Since the combined dataset is massive, we can use PySpark to handle the data preprocessing and all the next steps in an efficient way.

Then, we want to identify locations where water quality has degraded over time and where it stayed the same or improved and determine any factors that correlate with the degradation/consistency/improvement of water quality.

Finally, using the patterns and factors we found, we aim to predict which locations in the US might have their water quality degrade beyond safe levels in the next 10 years (subject to change) based on the trends and behaviors of the previous water parameter measurements.

References

EY. (2026). *EY Open Science AI & Data Challenge Program*. https://challenge.ey.com/.

World Health Organization. (2025, August 26). *1 in 4 people globally still lack access to safe drinking water – WHO & UNICEF*. World Health Organization: WHO. https://www.who.int/news/item/26-08-2025-1-in-4-people-globally-still-lack-access-to-safe-drinking-water---who--unicef.

Water Quality Data. (n.d.). *WQP Web Services Guide*. https://www.waterqualitydata.us/webservices_documentation/

Water Quality Data. (n.d.). *Portal user guide*. https://www.waterqualitydata.us/portal_userguide/.

Water Quality Portal. Washington (DC): National Water Quality Monitoring Council, United States Geological Survey (USGS), Environmental Protection Agency (EPA); 2021. https://doi.org/10.5066/P9QRKUVJ.