Plan Overview

A Data Management Plan created using DMP Tool

Title: Predictive Modeling for Water Quality Assessment

Creator: Sai sushma Maddali

Affiliation: San Jose State University (sjsu.edu)

Funder: Digital Curation Centre (dcc.ac.uk)

Template: Digital Curation Centre

Project abstract:

Access to clean and safe water is essential for public health and environmental sustainability. This project utilizes machine learning techniques to develop a predictive model that classifies water stations as “safe” or “unsafe” based on various parameters derived from lab and field results obtained from California’s publicly available water quality datasets. Using machine learning algorithms such as Logistic Regression, Support Vector Machines, and Random Forests, the model identifies key indicators for contamination across different water stations. The modeling pipeline includes data preprocessing, feature engineering, model training, model validation, and performance evaluation. Model performance is assessed using metrics such as F1 score, precision, recall, and accuracy. To interpret model predictions and identify influential water quality parameters, SHAP (Shapley Additive explanations) is used. The best-performing model and datasets used for training will be shared via GitHub to promote transparency and reproducibility. By leveraging data-driven insights, this project aims to enable early detection of unsafe water and support informed decision-making that safeguards public health and fosters a healthier environment.

Last modified: 09-29-2025

Predictive Modeling for Water Quality Assessment

Data Collection

What data will you collect or create?

The project uses publicly available datasets from the California Department of Water Resources (DWR). These include station metadata describing water quality monitoring locations, period-of-record data indicating temporal coverage, laboratory results measuring parameters such as dissolved aluminum, arsenic, and lead, and field observations collected directly at the monitoring stations. The combined size of these datasets is approximately 1.5 GB, with the laboratory results dataset (~1.05 GB) being the largest. While the datasets are moderately large and manageable with standard computing resources, the complexity lies primarily in the heterogeneity of the data sources and the need for preprocessing to standardize units, address missing values, and merge datasets effectively. All raw data are provided in CSV format, a widely used standard for tabular data. Any derived datasets generated during preprocessing, cleaning, and feature engineering will also be stored in CSV format to ensure consistency, interoperability, and long-term usability. Data are collected directly from the official DWR portal (data.ca.gov), which provides curated and publicly accessible water quality datasets. Quality control procedures will include verifying file integrity, standardizing units of measurement, handling missing values, and validating merged datasets by cross-referencing station codes and sampling dates to ensure accuracy.

How will the data be collected or created?

Raw datasets will be obtained from the California Department of Water Resources (DWR) through the data.ca.gov portal, which follows established laboratory and field sampling standards. Derived datasets will be created through preprocessing, cleaning, and feature engineering. Data will be organized using a structured folder hierarchy (e.g., raw_data, processed_data, intermediate_results, aggregated_results) with consistent file naming conventions that include dataset name, date, and version number. Version control for scripts and documentation will be managed through GitHub. Quality assurance will include verifying file integrity, performing data consistency checks, conducting range and validity checks, handling missing and erroneous data, and documenting all transformations in a README.md file to ensure consistency and reproducibility.

Documentation and Metadata

What documentation and metadata will accompany the data?

The documentation will include a README file that describes the data source, collection methods, preprocessing steps, and instructions for testing the model; a data dictionary that defines each variable, its description, unit of measurement, and data type; and model documentation that outlines how the models were developed, evaluated, and interpreted. Metadata will follow Dublin Core standards including elements such as the creator’s name, title, description, date, format, version, and license, ensuring the data can be easily understood, reproduced, and reused in the future.

Ethics and Legal Compliance

How will you manage any ethical issues?

The project uses publicly available data that does not contain sensitive or personal information. Therefore, there are no direct ethical concerns related to privacy. To meet ethical standards, the project ensures proper citation of data sources, prevents unauthorized access, and clearly documents preprocessing and modeling steps. Care will be taken to avoid misrepresentation of results and make fair use of data.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

The project uses open-source datasets from a government website, and the original data will be properly cited in accordance with its license terms. All project documentation, code, and models will be released under an open-source license (e.g., MIT) to allow reuse, modification, and collaboration. By complying with license requirements and giving proper credit, the project ensures responsible use of intellectual property and effective management of copyright issues.

Storage and Backup

How will the data be stored and backed up during the research?

During the research, three copies of the data will be maintained to ensure reliability. The primary active storage will be the local development computer, while a secondary backup will be kept on Google Drive, which provides automatic synchronization. Code, documentation, and derived datasets will also be version-controlled and preserved on GitHub with regular commits. All storage solutions are free of charge, and data will be periodically synchronized across platforms to maintain consistency.

How will you manage access and security?

Since the project uses publicly available datasets with no sensitive information, there are no major privacy concerns. During development, access to shared platforms such as Google Drive and GitHub will be restricted to authorized team members to prevent unauthorized modifications and preserve data integrity. Each team member will also be responsible for securing data on personal devices by following good data management practices, such as using strong passwords and keeping systems up to date.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Datasets from the California Department of Water Resources are already public and will stay available at their source. The long-term value of this project is in the derived datasets, preprocessed data, and trained models. These, along with their documentation, will be preserved as reusable resources for future researchers. By sharing these artifacts on GitHub, the project makes it easier for others to build on this work for research or education.

What is the long-term preservation plan for the dataset?

For long-term preservation, the project will retain and maintain all derived datasets, preprocessed data, and trained models along with their supporting documentation. These assets will be preserved in a public GitHub repository, where version control and open licensing will ensure continued accessibility and reusability.

Data Sharing

How will you share the data?

The project artifacts will be made openly available to the community through a dedicated GitHub repository at the end of the project development cycle. The final project report will include links to the repository and any related publications to make the data easy to find. In addition, the data will be accompanied by README files and data dictionaries so that secondary users can interpret and reuse it without needing direct input from the project team.

Are any restrictions on data sharing required?

No restrictions are required on data sharing. The raw datasets are publicly available, and the project-generated artifacts will be shared under an open-source license. This ensures that the data can be freely accessed and reused without limitations.

Responsibilities and Resources

Who will be responsible for data management?

The project is developed by a team of four, and all members of the team will share responsibility for managing the data. Each person will contribute to organizing files, maintaining documentation, and ensuring that data is properly backed up and versioned. Roles will be divided to ensure accountability; one team member will coordinate data sharing, another will maintain repository, a third will focus on ensuring data integrity, and the fourth will serve as the data manager overseeing overall data organization. This collaborative structure will support smooth and consistent data management.

What resources will you require to deliver your plan?

Data for the project will be collected from the California department of Water Resources and processed on local computers using Python, with Google Colab for additional compute if needed. Storage will be managed through Google Drive, while GitHub will be used for version control, CI/CD pipelines, and sharing final outputs. The model will be containerized with Docker and deployed on Azure using Container Registry and App Service. Streamlit will be used to develop the user interface. To support collaboration and planning, Zoom or Teams will be used for meetings, Google Sheets for task tracking, and Google Docs and draw.io for brainstorming and architecture design. Most tools are free, with the only significant cost being Azure deployment, estimated at about $60-$100 per month, which is justified to make the model accessible and scalable.

Data management timeline:

Weeks 1-2: Data acquisition from the DWR portal, initial storage setup, and data understanding

Weeks 3-5: Data cleaning, preprocessing, and feature engineering

Weeks 6-7: Model training and validation

Weeks 8-9: Model interpretation, evaluation, and documentation

Week 10: User interface development, backend integration, and testing

Week 11: Deployment on Azure, updating GitHub repository, and public sharing

---

Planned Research Outputs

Dataset - "Water Quality Indicators"

Model representation - "Predictive Model for Water Quality Assessment"

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Water Quality Indicators	Dataset	2025-12-09	Open	California Department of Water Resources		None specified	Dublin Core	No	No
Predictive Model for Water Quality Assessment	Model representation	2025-12-09	Open	California Department of Water Resources		None specified	Dublin Core	No	No