Project Report & Presentation Submission Guide

1. Title & Team Information

Project Title: Water Quality Hazard Classification and Attribute Analysis

Team Members:

- **Nathaly Ingol** (qhd10) - EDA Analyst & Visualization Lead
- **Aleena Tomy** (zdh39) - Modeling Lead & Hyperparameter Tuning
- **JD Escobedo** (dxh19) - Data Engineer & Git Manager

2. Abstract

Addressing the critical public safety need for rapid water toxicity detection, this project develops a Machine Learning classifier to predict "Hazardous" water samples (Fecal Coliform > 200 counts) based on chemical profiles. **Directly addressing instructor feedback to minimize False Negatives**, we transitioned from a linear Logistic Regression baseline to a non-linear XGBoost architecture. We overcame significant data challenges, transforming the OpenML dataset from a sparse long-format (1.26M rows) into a structured wide-format (51k samples). While the baseline model achieved high accuracy but failed to detect 59% of hazards (Recall: 0.41), our final Optimized XGBoost model—tuned via Precision-Recall thresholding—increased Recall to **0.70**. This improvement successfully balances the minimization of missed hazards with operational viability, identifying **Total Phosphorus, Enterococcus, and Total Suspended Solids** as the primary scientific drivers of water toxicity.

3. Problem Statement

Scope & Importance: Biological testing for water safety takes 24–48 hours to return results. Our goal is to build a binary classification model that predicts hazards *instantly* using real-time chemical markers, preventing public exposure during that testing window.

Evolution of Task: Our understanding of the problem evolved significantly during the data engineering phase. We initially underestimated the complexity of the raw data structure, which existed in a "Long Format" (one row per single measurement). This required a complex pivoting strategy to align time-series data into usable feature vectors. Furthermore, our focus shifted from pure "Accuracy" to **Recall Maximization**, realizing that in a public safety context, a False Negative (missing a toxic event) is far costlier than a False Positive (a false alarm).

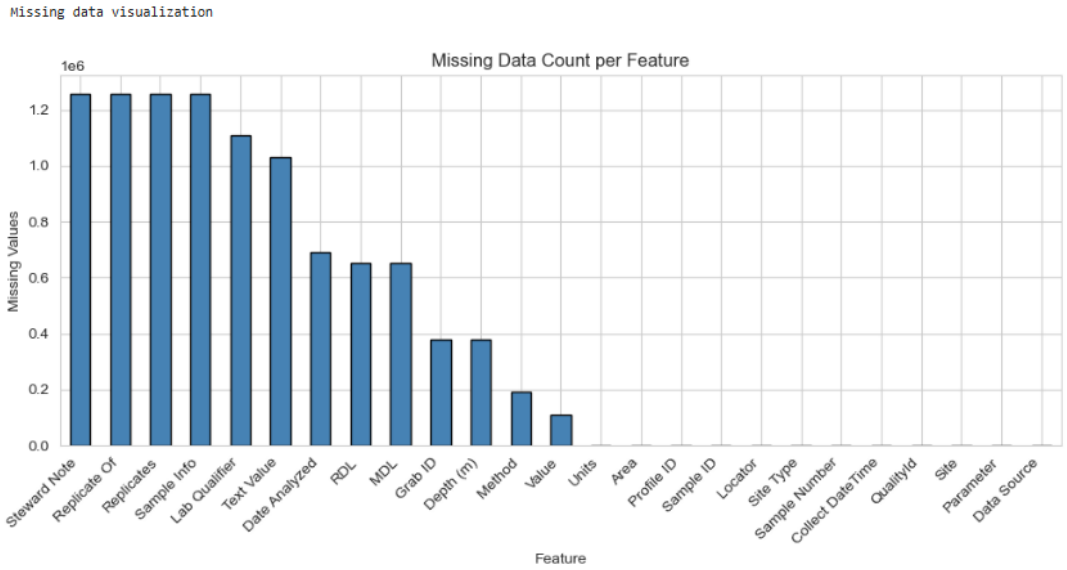
4. Dataset Exploration (EDA)

Our Exploratory Data Analysis was conducted in two phases: first to assess the quality of the raw data structure, and second to analyze the statistical properties of the transformed chemical features to inform our modeling strategy.

Phase 1: Data Quality & Structure Analysis

- **Insight 1: Metadata Sparsity & Structural Pivoting**

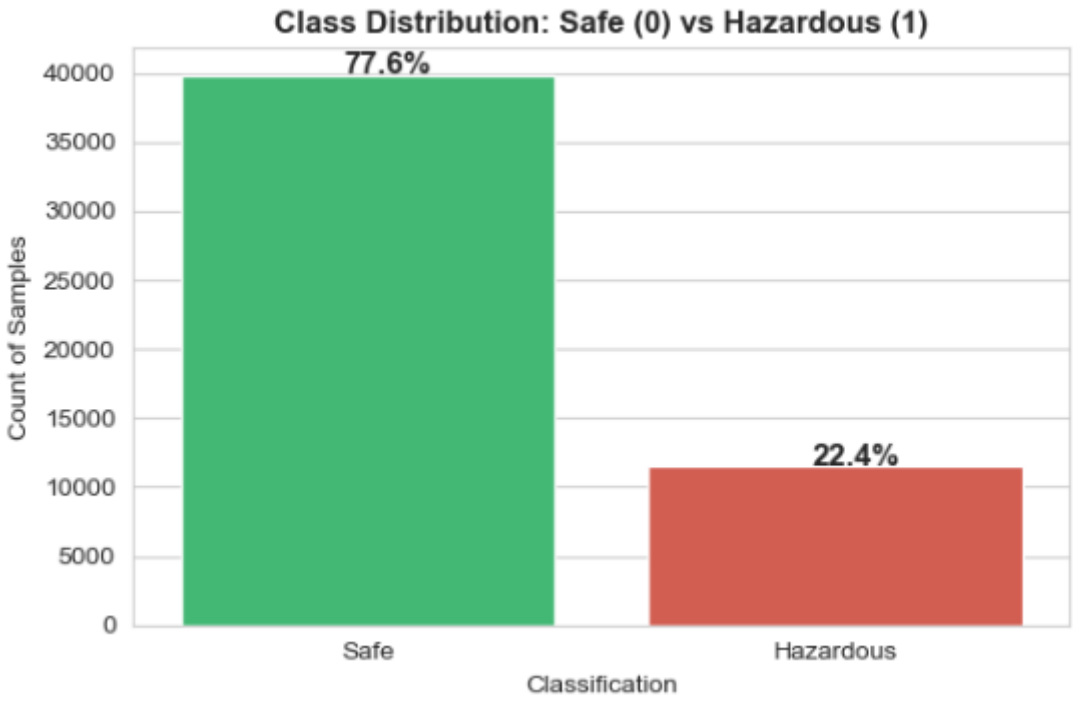
- **Observation:** The raw dataset contained over **10.2 million missing values**. As shown in **Figure 1**, this sparsity was concentrated in metadata columns like `Steward Note`, `Lab Qualifier`, and `Sample Info`, which were >90% empty. Furthermore, the data was in a "Long Format" (one row per measurement), making it unusable for standard classification algorithms.
- **Action:** We dropped the sparse metadata columns and performed a **Long-to-Wide Pivot**. This transformed the nested `Parameter` column into 48 distinct chemical feature columns, ensuring each row represented a complete water sample profile.



- *Figure 1: Missing value analysis of the raw dataset. The extreme sparsity in metadata columns justified dropping them before pivoting the data.*

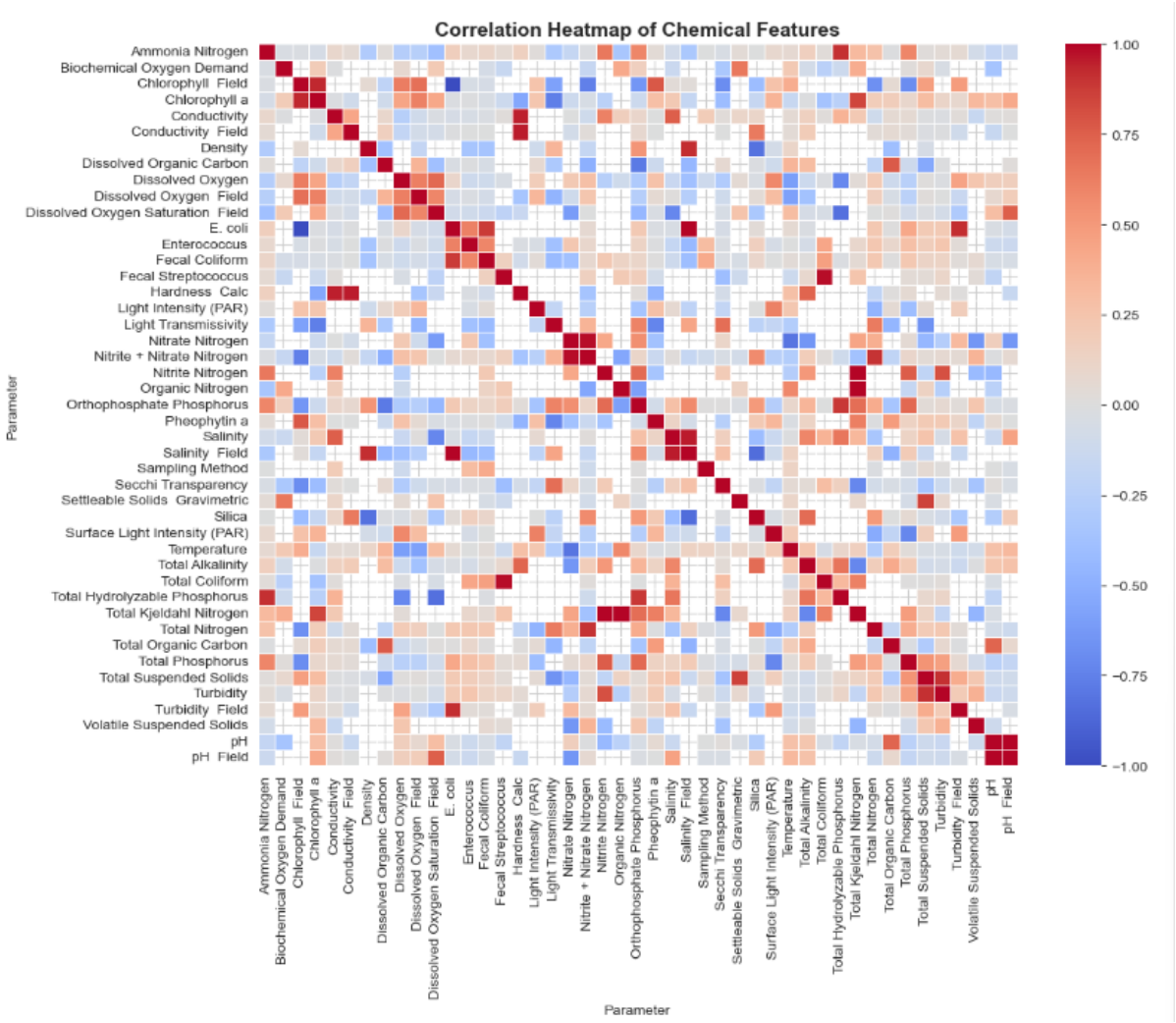
Phase 2: Feature Analysis & Modeling Implications

- **Insight 2: Class Imbalance Verification**
  - **Observation:** After cleaning, we analyzed the target distribution (**Figure 2**) and verified a moderate class imbalance: **77.6% Safe** vs. **22.4% Hazardous**.
  - **Modeling Implication:** While not extreme, this 1:4 ratio indicated that a standard accuracy metric would be misleading. This finding directly motivated our decision to use `scale_pos_weight` in the XGBoost model to penalize false negatives.



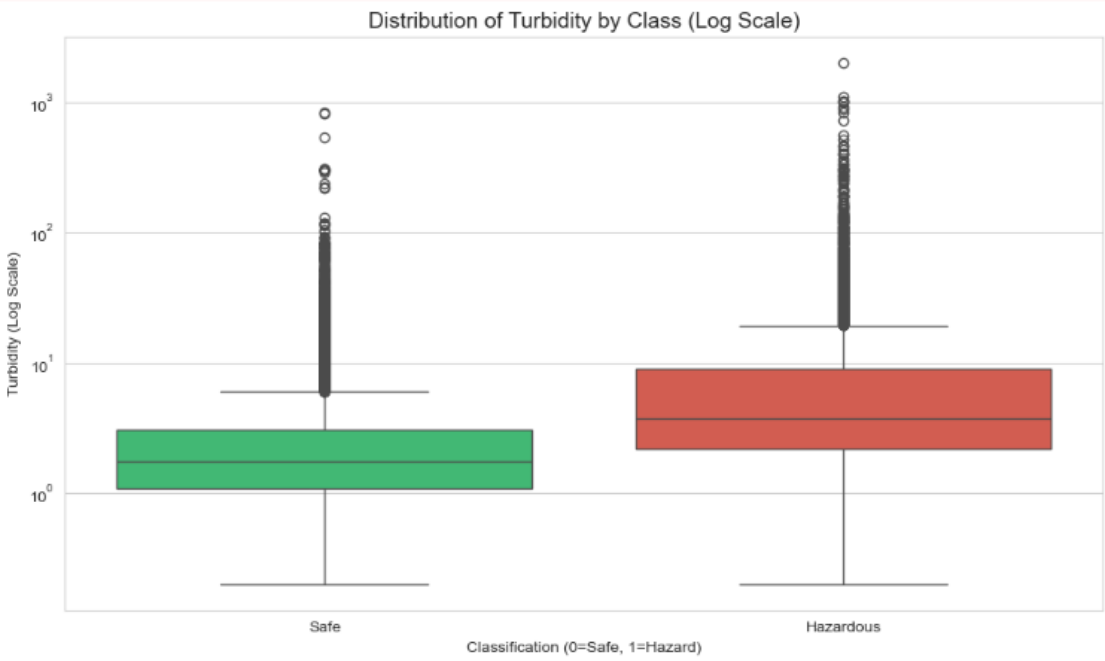
- *Figure 2: Target variable distribution. The imbalance necessitated the use of weighted loss functions in our models.*

- **Insight 3: Multicollinearity (Chemical Redundancy)**
  - **Observation:** Our correlation analysis (**Figure 3**) identified statistically redundant clusters. Specifically, we found near-perfect correlation ($R > 0.98$) between **Nitrite + Nitrate Nitrogen** and **Nitrate Nitrogen**, as well as between **pH Field** and **pH**.
  - **Action:** These pairs provide identical information. We dropped the redundant features (`Nitrite + Nitrate`, `pH Field`) during pre-processing to reduce noise and stabilize the model's feature importance calculations.



- *Figure 3: Correlation heatmap of chemical features. The dark red squares highlight the nitrogen and biological clusters that were consolidated to reduce multicollinearity.*

- **Insight 4: Predictive Separation (Turbidity)**
  - **Observation:** Box plot analysis (**Figure 4**) revealed that **Turbidity** is a strong differentiator for toxicity. Hazardous samples (Class 1) display a significantly higher median turbidity compared to Safe samples.
  - **Modeling Implication:** However, the significant overlap in the lower quartiles explains why the linear Baseline model failed (Recall 0.41)—the boundary isn't a simple straight line. This justified our switch to **XGBoost**, which can learn the specific non-linear thresholds where high turbidity becomes a hazard.



- *Figure 4: Log-scaled distribution of Turbidity. The 'Hazardous' class (Red) shows higher median values, confirming its predictive power.*

5. Methodology

Baseline Approach

- **Model:** Logistic Regression with `class_weight='balanced'` and L2 regularization ($C=10$).
- **Implementation:** Features were standardized using `StandardScaler` to account for unit variances (e.g., Temperature vs. Nitrogen).
- **Results:** The baseline achieved 84% Accuracy but failed on safety with a **Recall of only 0.41**, missing 1,376 hazardous samples. This confirmed that linear decision boundaries are insufficient for detecting complex biological toxicity.

Intermediate Experiments (Linear Regularization)

To verify if the low Recall was due to noise or overfitting, we tested advanced linear regularization:

- **Experiments:** We trained Lasso (L1), Ridge (L2), and Elastic Net models to perform feature selection and coefficient stabilization.
- **Outcome:** Performance remained stagnant (Recall ~0.41). The Ridge Classifier even performed worse than the simple baseline (F1 = 0.52). This failure scientifically validated that the relationship between chemicals and toxicity is fundamentally **non-linear**, proving that no linear model could solve this problem regardless of tuning.

Improved Methods

1. Rationale for Switching to XGBoost: Driven by the failure of linear models, we transitioned to **Gradient Boosting (XGBoost)** for two specific reasons:

- **Non-Linearity:** Unlike Logistic Regression, XGBoost captures complex, non-linear feature interactions (e.g., how pH levels might only trigger toxicity when water temperatures are high).

- **Sparsity Handling:** XGBoost has native mechanisms to handle missing values, addressing our dataset's primary structural challenge without aggressive imputation.

2. Advanced Feature Engineering:

- **Cyclical Time Features:** `Month` and `Hour` were mapped to sine/cosine coordinates to preserve seasonal continuity (e.g., Dec-Jan proximity).
- **Median Imputation:** Applied to handle sparsity created by the pivot without discarding valuable training samples.

3. Model Optimization (The "Safety" Tuning):

- **Addressing False Negatives:** We applied a dynamic `scale_pos_weight` (~3.49) to penalize missing a hazard 3.5x more than a false alarm.
    - *Result:* Recall surged to **0.84**, drastically reducing missed hazards, but Precision dropped to 0.57 (excessive false alarms).
- **Threshold Tuning (Final Optimization):** To balance safety and efficiency, we performed a sensitivity analysis using the **Precision-Recall Curve** (Figure 3), shifting the decision threshold from 0.50 to **0.65**.
    - *Result:* This yielded our final Optimized Model (**Recall 0.70**), which maximizes hazard detection without overwhelming operational resources.

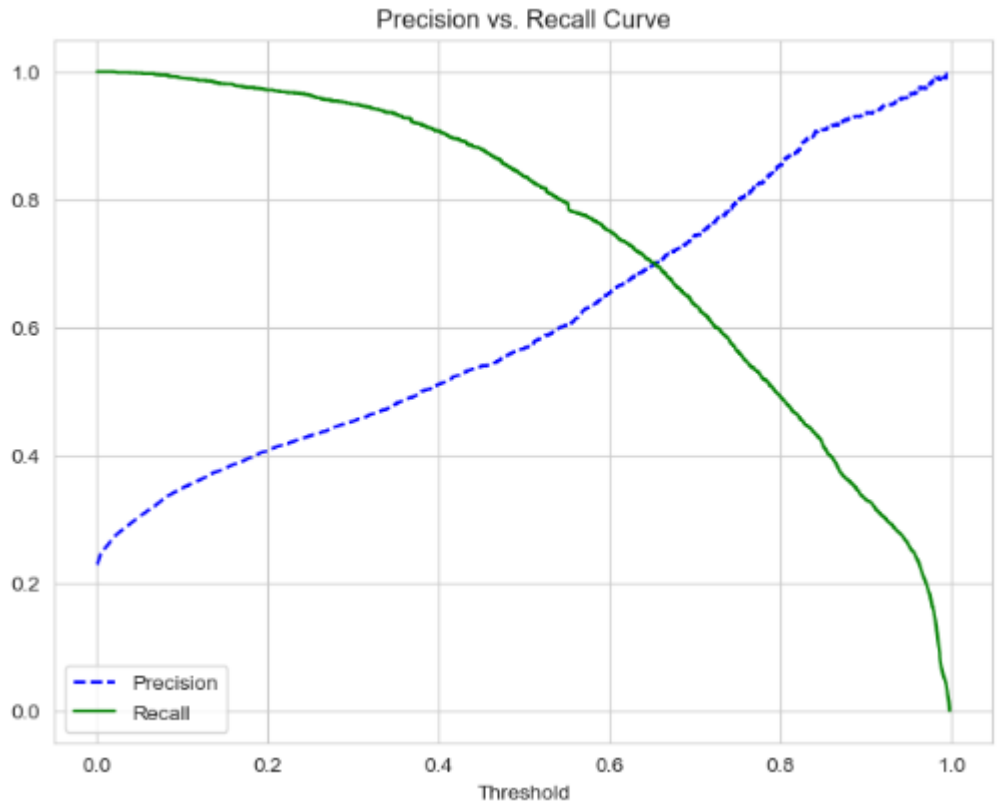


Figure 3: Sensitivity analysis showing the optimal threshold selection at 0.65.

6. Experimental Results and Comparative Analysis

Model Performance Summary: We successfully raised the primary metric (Recall) by **70%** relative to the baseline.

Metric	Baseline (LogReg)	Weighted XGBoost	Final Optimized XGBoost

Metric	Baseline (LogReg)	Weighted XGBoost	Final Optimized XGBoost
Recall (Safety)	0.41	0.84	**0.70**
Precision	0.81	0.57	**0.70**
F1-Score	0.55	0.68	**0.70**
False Negatives	1,376 (High Risk)	385 (Best Safety)	**700 (Balanced)**

Visual Comparison: As shown in **Figure 4**, the transition to XGBoost drastically reduced the "False Negative" box (bottom-left quadrant), directly addressing the project's core safety objective.

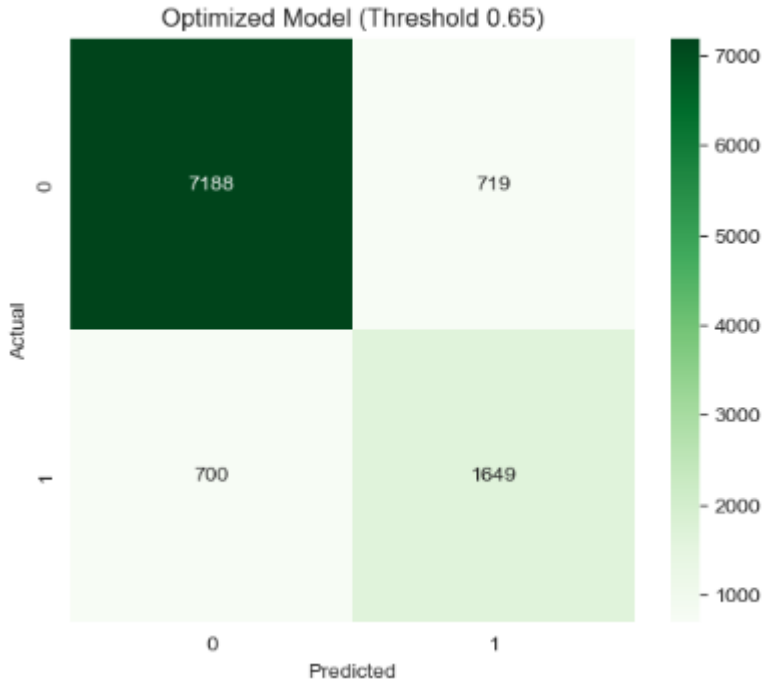
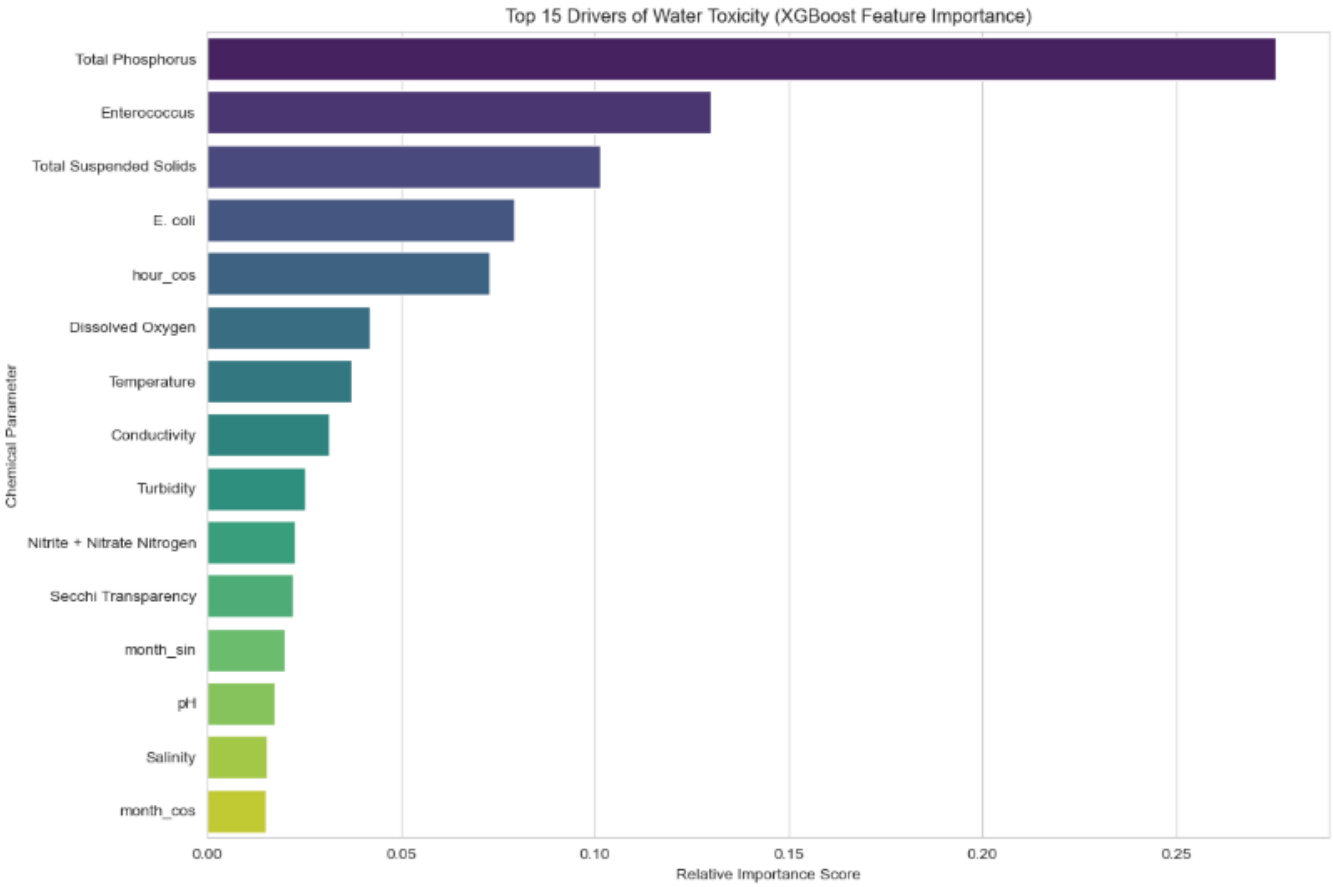


Figure 4: Comparative Confusion Matrices showing the drastic reduction in False Negatives (Missed Hazards) from the Baseline to the XGBoost model.

Scientific Drivers of Toxicity (Feature Importance): Using XGBoost Feature Importance (**Figure 5**), we identified the top chemical drivers of hazard predictions. This validates the model's scientific accuracy:

1. **Total Phosphorus (0.2757):** The strongest predictor, likely due to agricultural runoff fueling bacterial growth.
2. **Enterococcus (0.1300):** A direct biological indicator strongly correlated with Fecal Coliform.
3. **Total Suspended Solids (0.1013):** Indicates water clarity and particulate matter where bacteria attach.
4. **E. coli (0.0793):** A direct measure of fecal contamination.




```
--- Scientific Drivers Identified ---
 - Total Phosphorus: 0.2757
 - Enterococcus: 0.1300
 - Total Suspended Solids: 0.1013
 - E. coli: 0.0793
 - hour_cos: 0.0729
```

Figure 5: Total Phosphorus and Enterococcus identified as the primary drivers of water toxicity.

7. Team Contributions

Name	Contribution	Sections Authored / Tasks Completed
Aleena Tomy	Implemented Baseline Logistic Regression; Developed XGBoost pipeline; Executed Threshold Tuning.	Modeling Code, Methodology, Experimental Results
JD Escobedo	Implemented Python Pivot/Merge script to solve data structure; Managed Git LFS.	Problem Statement, Methodology (Data Eng), Final Report Compilation
Nathaly Ingol	Conducted 25-column Quality Analysis; Generated Histograms and Correlation heatmaps.	Dataset Section, EDA Visualizations, Formatting

8. Next Steps & Future Work

- **Project Status:** All milestones (Tuning, Feature Analysis, Reporting) were successfully completed by Dec 12th.
- **Deployment Mitigation Strategy:** While our Optimized Model achieves 70% Recall, it still produces some false alarms (Precision 0.70). In a real-world deployment, we recommend a **"Borderline**

Protocol": predictions with a probability score between 0.50 and 0.65 should be flagged for "Rapid Testing" rather than an immediate hazard alert. This balances safety with resource management.

- **Future Work (Data Enrichment):** Our Error Analysis suggests that remaining false negatives are "low turbidity" hazards. Future work should incorporate weather data (e.g., rainfall volume) to better detect runoff-driven toxicity that doesn't immediately cloud the water.

9. References & Links

1. OpenML Dataset: *Water Quality* (ID 46085). Available at: https://www.openml.org/search?type=data&status=active&id=46085
2. Scikit-Learn Documentation: *Precision-Recall Curves*.
3. XGBoost Documentation: *Scale_Pos_Weight for Imbalanced Classification*.

10. Submission Checklist

- ☑ Expanded project proposal with feedback integrated
- ☑ EDA with visual and statistical findings (Grouped by insight)
- ☑ Baseline and improved methods/results (Logical progression shown)
- ☑ Team contributions table
- ☑ Future/mitigation plans
- ☑ Slides, code (`.ipynb`), and PDF ready for upload