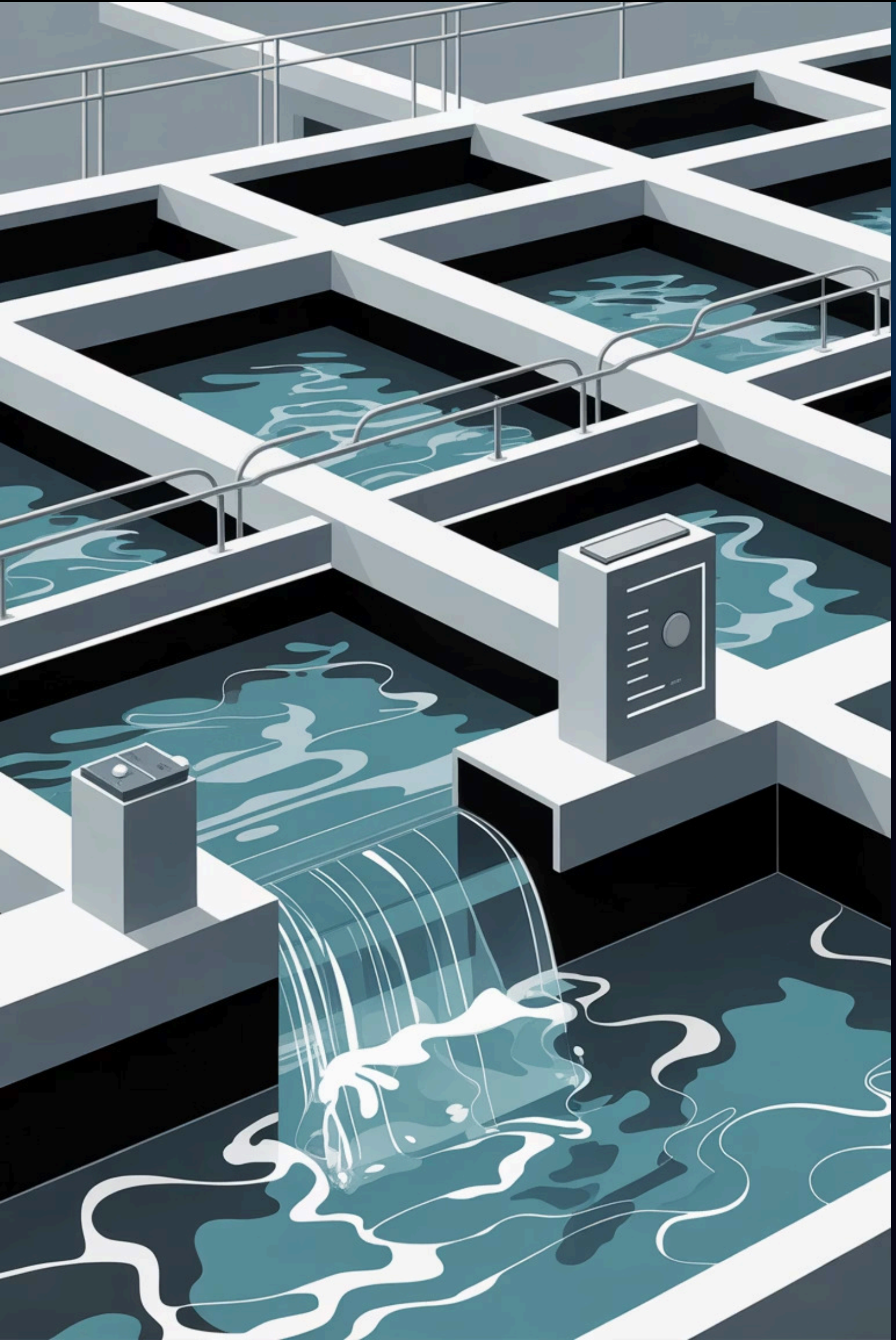


AquaSafe: Water Quality Classification

Machine learning for regulatory water quality assessment

MAHARASHTRA POLLUTION CONTROL BOARD

NATIONAL WATER MONITORING PROGRAMME

AUGUST 2025

*Author: Rakshit Pandey — **github.com/pandey-rakshit***

Problem Statement

Objective: Classify water bodies into regulatory quality classes to ensure public safety and appropriate resource allocation.

Class A

Drinking water (disinfection only) — Highest quality standard

Class B

Outdoor bathing (organised) — Recreational use

Class C

Drinking water (with treatment) — Potable after processing

Class E

Irrigation, industrial cooling, waste disposal — Agricultural/Industrial

⚠️ **Why It Matters:** Incorrect classification can lead to serious public health risks, particularly when lower-quality water is misclassified for drinking purposes.

Dataset Overview

175

Total Samples

Water quality measurements from monitoring stations

92

Features

After categorical encoding

140

Training Samples

Used for model learning

35

Test Samples

Reserved for validation

Class Distribution

The dataset exhibits severe class imbalance, with Class A representing the majority of samples.

- **Class A:** 83% (dominant) — Drinking water quality
- **Class E:** 11% — Agricultural/Industrial
- **Class B:** 3% — Recreational bathing
- **Class C:** 3% — Treatable drinking water

⚠️ Severe class imbalance presents a significant modelling challenge

Models Trained & Evaluation

Model Specifications

Logistic Regression

Type: Linear classifier

Max iterations: 2000, balanced class weights

Random Forest

Type: Ensemble bagging

300 trees, balanced weights, bootstrap sampling

XGBoost

Type: Gradient boosting

300 estimators, max depth: 4, regularised

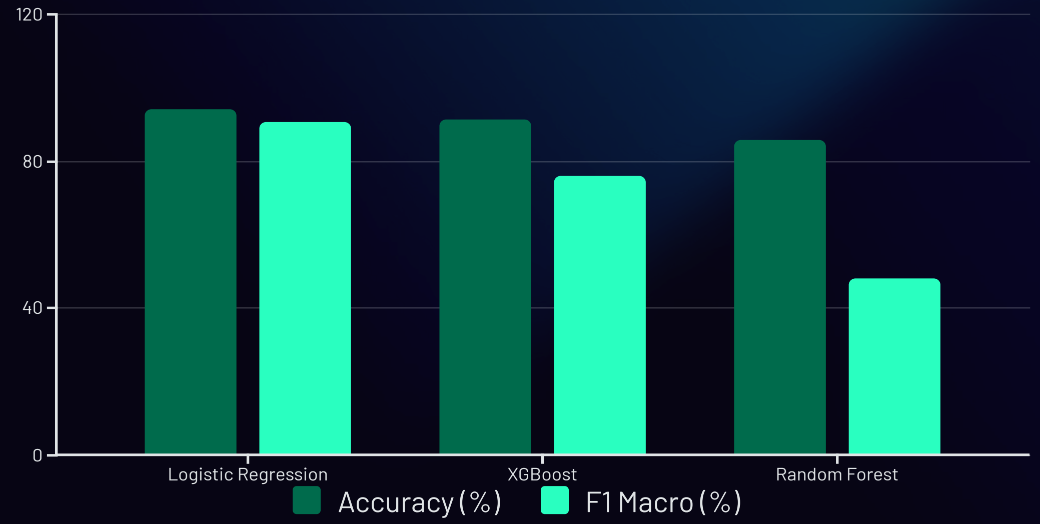
Evaluation Strategy

Rigorous validation approach to ensure model reliability and generalisability.

- **5-Fold Stratified Cross-Validation** — Maintains class proportions in each fold
- **Held-out Test Set** — Independent evaluation on unseen data
- **Multiple Metrics** — Accuracy, F1, precision, and recall tracked

Model Comparison Results

Comprehensive performance evaluation across three machine learning approaches reveals clear differences in classification accuracy and reliability.



Performance Metrics

🏆 Winner: Logistic Regression — Delivers the best balance of accuracy and F1 score for imbalanced classification.

Logistic Regression

Accuracy: 94.29% | F1: 90.83%

Precision: 98.39% | Recall: 87.50%

XGBoost

Accuracy: 91.43% | F1: 75.83%

Precision: 85.89% | Recall: 81.25%

Random Forest

Accuracy: 85.71% | F1: 48.02%

Precision: 46.32% | Recall: 50.00%

Per-Class Performance Analysis

Detailed examination of model performance across individual water quality classes reveals strong results for most categories, but a critical weakness in Class E detection.

Critical Issue: Class E has low recall — 2 out of 4 samples misclassified as Class A, representing a serious safety concern.

1

Class A — Drinking Water

Precision: 93.5% | Recall: 100% | F1: 96.7%

✅ Excellent — Reliably identifies highest quality water

2

Class B — Bathing

Precision: 100% | Recall: 100% | F1: 100%

✅ Excellent — Perfect classification

3

Class C — Treatable

Precision: 100% | Recall: 100% | F1: 100%

✅ Excellent — Perfect classification

4

Class E — Industrial/Agricultural

Precision: 100% | Recall: 50% | F1: 66.7%

⚠️ Fair — Critical issue identified

Top 10 Feature Importance

Analysis of feature coefficients reveals which water characteristics most strongly influence classification decisions. Physical properties and human activities emerge as key predictors.

Approximate Depth < 50cm

Importance: 0.571 — Strongest predictor of water quality class

Human Activities — Others

Importance: 0.460 — Significant indicator of water use patterns

Approximate Depth > 100cm

Importance: 0.376 — Depth strongly correlates with quality

Phosphate Levels

Importance: 0.342 — Key chemical indicator

Conductivity

Importance: 0.304 — Measures dissolved ionic content

Turbidity Below Detection

Importance: 0.302 — Indicates water clarity

Key Insight: Water depth and human activity patterns are the strongest predictors of regulatory water quality class, surpassing individual chemical parameters.

🚨 Critical Safety Limitation

Production Blocker Identified: The model exhibits a dangerous misclassification pattern that poses significant public health risks.

Misclassification Pattern

1 **Actual Class E**

Waste disposal, irrigation — Not suitable for human consumption

2 **Predicted Class A**

Drinking water quality — Cleared for human consumption

🔴 **CRITICAL RISK:** 2 cases of E → A misclassification detected

Root Cause Analysis

This is **not** an overfitting problem — it's a fundamental data limitation combined with feature overlap between extreme classes.

⚠️ **NOT RECOMMENDED FOR PRODUCTION** without implementing the fixes outlined in the next section.

01

Severe Class Imbalance

Only 16 Class E samples in training data — insufficient to learn distinctive patterns

02

Feature Overlap

Some Class E samples share similar chemical profiles with Class A water bodies

03

Low Model Confidence

Prediction confidence on errors: ~55% — model is uncertain about these classifications

Recommendations for Deployment

Immediate Fixes Required

1. Data Augmentation

Collect 50+ additional Class E samples to balance the training dataset and improve minority class learning

2. Threshold Adjustment

Lower classification threshold for Class E from 0.5 to 0.3 to increase sensitivity to potential contamination

3. Cost-Sensitive Learning

Implement asymmetric loss function — penalise E→A errors 10× more heavily than other misclassifications

4. Manual Review Flag

Add "flagged for manual review" category when prediction confidence falls below 70%

Future Enhancements

Longer-term improvements to increase model robustness and interpretability for regulatory acceptance.

- **SHAP Values Integration** — Provide explainable AI insights for each prediction to support regulatory review
- **Hybrid Ensemble Approach** — Combine ML predictions with rule-based pollutant thresholds from regulatory standards
- **Feature Expansion** — Incorporate seasonal patterns, geographic location, and upstream contamination sources
- **Continuous Learning Pipeline** — Implement active learning to progressively improve minority class performance

Conclusion

Key Achievements

○ **94.29% Overall Accuracy**

Strong performance on test set demonstrates model viability

○ **90.83% Macro F1 Score**

Excellent for imbalanced data — significantly above baseline

○ **91.28% Mean Confidence**

Model predictions are generally well-calibrated and reliable

○ **Interpretable Model**

Logistic regression provides transparent, auditable decision-making for regulators

Current Status & Path Forward

The model shows promise but requires refinement before operational deployment in regulatory contexts.

> *"Clean water is not a privilege, it's a right."*

📁 GitHub: github.com/pandey-rakshit

⚠️ **Suitable for preliminary screening only** — Requires expert oversight for final classification decisions

⚠️ **More Class E data essential** — Additional samples needed before production deployment

⚠️ **Critical E→A errors must be addressed** — Safety-first approach required for public health protection