Water Potability Classification: Integrating Environmental Assessment with Machine Learning

Harsh Vishwakarma (2022205) & **Ayan Kumar Singh** (2022122)
IIIT Delhi — harsh22205@iiitd.ac.in, ayan22122@iiitd.ac.in

November 17, 2025

Abstract

Access to potable water is threatened by contaminants ranging from industrial effluents to persistent organic pollutants (POPs). This study synthesizes recent literature on contaminant behavior and implements a machine learning framework using the Kaggle Water Potability dataset to predict water safety. By employing rigorous preprocessing—including median imputation and SMOTE to handle class imbalance—we evaluated twelve classification models. Ensemble methods, particularly LightGBM and CatBoost, demonstrated superior performance over linear models, achieving ROC-AUC scores of **0.672** and **0.696** respectively. This work validates machine learning as a robust decision-support tool for water resource management.

1 Introduction

Water quality is increasingly threatened by anthropogenic contaminants, from physicochemical parameters to emerging threats like Per- and Polyfluoroalkyl Substances (PFAS). Addressing this requires understanding contaminant risks and developing automated prediction systems.

This research bridges environmental toxicology and computational intelligence. We first assess water quality deterioration, synthesizing findings on contaminant mobility and mitigation. Subsequently, using the Water Potability dataset [6], we develop predictive models to classify water safety. The study involves extensive preprocessing and the evaluation of algorithms such as SVM, Random Forest, and LightGBM. The primary objectives are to:

- Synthesize knowledge on emerging contaminants (PFAS, pathogens).
- Process the dataset using median imputation and SMOTE for class imbalance.
- Develop and optimize ML classifiers to maximize ROC-AUC and F1-Scores.

2 Integrated Assessment of Water Quality

2.1 Global Contaminant Landscape

Current monitoring paradigms are often inadequate against modern pollutants. **Ceballos et al. (2025)** [1] investigated PFAS in the Greater Montreal Area, detecting **32 distinct PFAS compounds** across 15 treatment plants. Crucially, conventional treatment removed only **8–36%** of these chemicals, allowing them to persist in tap water.

In terms of mitigation, **Evans et al. (2025)** [2] analyzed 19 US water systems, finding that advanced filtration (GAC/RO) not only addressed PFAS but also reduced regulated disinfection byproducts in **84%** of systems. Conversely, industrial reuse poses risks; **Redman

et al. (2025)** [3] identified **ammonia** as the primary driver of aquatic toxicity in treated produced water, rather than hydrocarbons. Furthermore, **Rasheduzzaman et al. (2025)** [4] analyzed 107 bottled water samples, finding significant variability in physicochemical parameters, challenging the perception that commercial water is inherently safer.

2.2 Case Study: The Delhi NCR Crisis

The National Capital Region (NCR) of Delhi exemplifies the acute nature of this crisis. With a population density exceeding **36,000 people/sq. km**, the region relies heavily on the Yamuna River. CPCB data [7] reveals the river is overwhelmed by:

- **Untreated Sewage:** Resulting in Faecal Coliform counts exponentially above standards.
- **Industrial Effluents:** Introducing heavy metals and elevating conductivity.
- **Ecological Dead Zones:** BOD frequently exceeds **3 mg/L** while Dissolved Oxygen drops below **5 mg/L**.

3 Dataset and Methodology

3.1 Data Description and Preprocessing

We utilized the Kaggle Water Potability dataset [6] containing **3,276 observations** and 10 features. The target variable is binary (`Potability`: 1 = Potable, 0 = Not Potable).

Table 1: Dataset Features Summary

Feature	**Description**
pH, Hardness, Solids	Basic physicochemical properties
Chloramines, Sulfate	Chemical additives/compounds
Conductivity	Electrical conductivity (μS/cm)
Organic_carbon, Turbidity	Contamination indicators
Trihalomethanes	Disinfection byproducts (μg/L)

Preprocessing Steps:

1. **Imputation:** Missing values in `ph` (14.99%), `Sulfate` (23.84%), and `Trihalomethanes` (4.95%) were filled using the **median** to mitigate outlier sensitivity.
2. **Class Balancing:** The dataset was imbalanced (61% Not Potable). We applied **SMOTE** (Synthetic Minority Over-sampling Technique) to the training set.
3. **Scaling:** Features were standardized ($z = \frac{x-\mu}{\sigma}$) for distance-based algorithms.

4 Exploratory Data Analysis (EDA)

The dataset exhibits high variability. **Solids** range from 320 to 61,227 ppm. The correlation matrix analysis revealed weak linear correlations with potability, suggesting the need for non-linear models (Ensembles) rather than simple regression. The distribution analysis further confirmed the presence of outliers in Hardness and Solids, which were retained to preserve real-world chemical variability.

5 Results and Discussion

We trained twelve models using 5-fold stratified cross-validation. As hypothesized during EDA, ensemble methods significantly outperformed linear models.

Table 2: Top Model Performance (5-Fold CV)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
CatBoost	0.6502	0.5524	0.5446	0.5485	**0.6959**
Random Forest	**0.6624**	**0.5793**	0.4914	0.5318	0.6904
!XGBoost	0.6371	0.5333	0.5579	0.5453	0.6818
LightGBM	0.6371	0.5353	0.5274	0.5313	0.6723
SVM (RBF)	0.6401	0.5355	**0.5837**	**0.5586**	0.6791
Logistic Regression	0.4985	0.3859	0.4828	0.4289	0.4955

CatBoost achieved the highest ROC-AUC (**0.696**), indicating the best capability to distinguish between potable and non-potable classes. **Random Forest** achieved the highest accuracy (**66.24%**). In contrast, Logistic Regression performed near the baseline (ROC-AUC ≈ 0.50), confirming that water potability is a complex, non-linear function of chemical attributes.

6 Conclusion

This study integrated environmental assessment with machine learning. The literature highlighted the inadequacy of current infrastructure against contaminants like PFAS. Our computational analysis on 3,276 samples demonstrated that while individual chemical features show weak linear correlation with potability, non-linear ensemble models like **CatBoost** and **Random Forest** can effectively predict water safety. These models offer a scalable framework for automated water quality monitoring.

References

[1] Ceballos, I. M., et al. (2025). Characterization of PFAS in Drinking Water... *ACS ES&T Water*, 5, 5509–5522.

[2] Evans, S. S., et al. (2025). PFAS Treatment as an Opportunity... *ACS ES&T Water*, 5, 5447–5459.

[3] Redman, A. D., et al. (2025). Integrated Assessment of Whole Effluent Toxicity... *ACS ES&T Water*, 5, 5275–5286.

[4] Rasheduzzaman, M., et al. (2025). Microbiological and Chemical Contaminants in Bottled Water... *ACS ES&T Water*, 5, 5471–5481.

[5] Guzman, H. P., et al. (2025). Wastewater Surveillance of Salmonella... *ACS ES&T Water*, 5, 5426–5437.

[6] Kadiwal, A. (n.d.). *Water Potability* [Data set]. Kaggle.

[7] CPCB. (2019). *Water quality monitoring data for Yamuna River.*